Forethought

Better Futures

William MacAskill

August 2025

Contents

Better Futures

Introducing Better Futures

No Easy Eutopia

Convergence and Compromise

Persistent Path-Dependence

How to Make the Future Better

Supplement: The Basic Case for Better Futures

William MacAskill

For correspondence, please email contact@forethought.org.

Suppose we want the future to go better. What should we do?

One approach is to avoid near-term catastrophes, like human extinction. This essay series explores a different, complementary, approach: improving on futures where we survive, to achieve a truly *great* future.

There are five main sections, plus a supplement. The main sections are designed to be read in order, beginning with “Introducing Better Futures”.

Forethought

Introducing Better Futures

William MacAskill

August 2025

1. The basic case

Suppose we want the future to go better. What should we do?

One prevailing approach is to try to avoid roughly *zero-value* futures: reducing the risks of human extinction or of misaligned AI takeover.

This essay series will explore an alternative point of view: *making good futures even better*. On this view, it's not enough to avoid near-term catastrophe, because the future could still fall far short of what's possible. From this perspective, a near-term priority — or maybe even *the* priority — is to help achieve a truly *great* future.

That is, we can make the future go better in one of two ways:

1. *Surviving*: Making sure humanity avoids near-term catastrophes (like extinction or permanent disempowerment).[1]
2. *Flourishing*: Improving the quality of the future we get if we avoid such catastrophes.

This essay series will argue that work on *Flourishing* is in the same ballpark of priority as work on *Surviving*. The basic case for this appeals to the *scale*, *neglectedness* and *tractability* of the two problems, where I think that *Flourishing* has greater scale and neglectedness, but probably lower tractability. This section informally states the argument; the supplement ("The Basic Case for Better Futures") makes the case with more depth and precision.

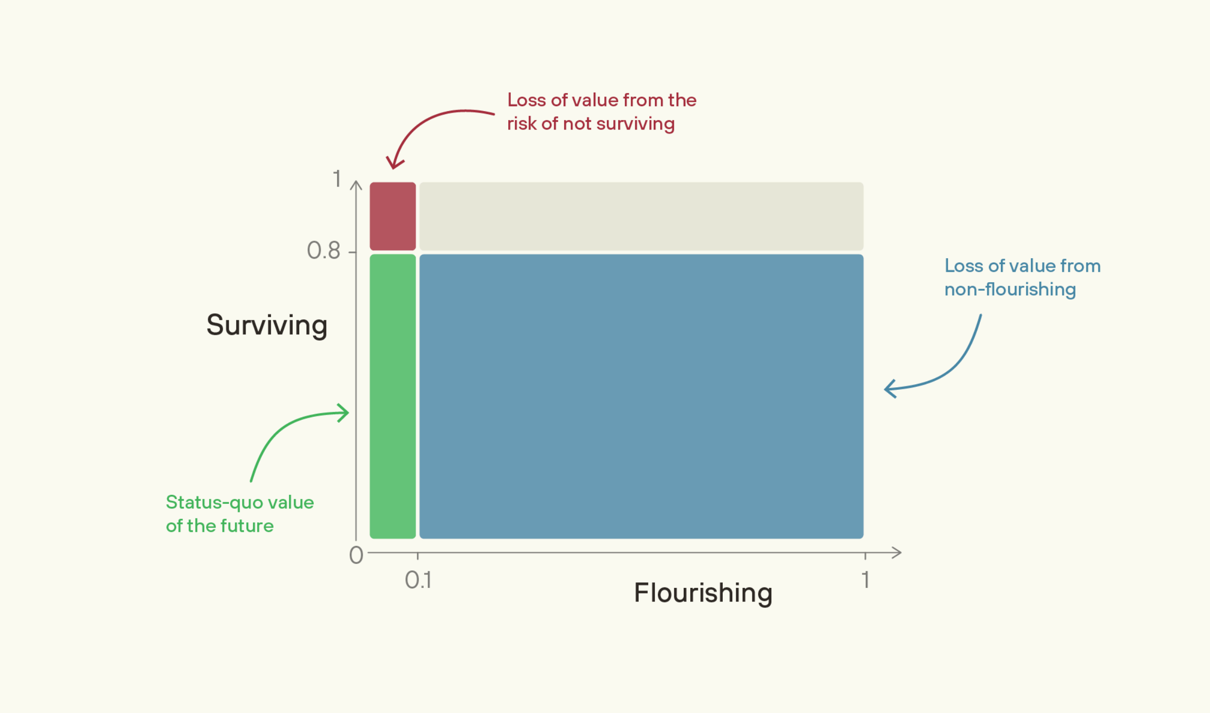
Scale

First, scale. As long as we're closer to the ceiling on *Survival* than we are on *Flourishing* — if there is more room for improvement on the latter — then *Flourishing* has greater scale.

To illustrate, suppose you think that our chances of survival this century are reasonably high (greater than 80%) but that, if we survive, we should expect a future that falls far short of how good it could be (less than 10% as good as the best feasible futures). These are close to my views; the view about *Surviving* seems widely-held,[2] and Fin Moorhouse and I will argue in essays 2 and 3 for something like that view on *Flourishing*. If so, there's more room to improve the future by working on *Flourishing* than by working on *Surviving*.

[1] *Surviving* represents the probability of avoiding a near-total loss of value this century (an "existential catastrophe"), while *Flourishing* represents the expected value of the future conditional on our survival.

[2] Grace, Stewart, Sandkühler, Thomas, Weinstein-Raun, and Brauner, 'Thousands of AI Authors on the Future of AI'; Besiroglu, 'Ragnarök Series—results so far'; Karger, Rosenberg, Jacobs, Hadshar, Gamin, Smith, McCaslin, Thomas, and Tetlock, 'Forecasting Existential Risks'.



Comparing the scale of surviving and flourishing

On these numbers, if we completely solved the problem of not- *Surviving* , we would be 20 percentage points more likely to get a future that's 10% as good as it could be. Multiplying these together, the difference we'd make amounts to 2% of the value of the best feasible future.

In contrast, if we completely solved the problem of non- *Flourishing* , then we'd have an 80% chance of getting to a 100%-valuable future. The difference we'd make amounts to 72% of the value of the best feasible future — 36 times greater than if we'd solved the problem of not- *Surviving* . Indeed, increasing the value of the future given survival from 10% to just 12.5% would be as good as wholly eliminating the chance that we don't survive.[3]

And the upside from work on *Flourishing* could plausibly be much greater still than these illustrative numbers suggest. If *Surviving* is as high as 99% and *Flourishing* as low as 1%, then the problem of non- *Flourishing* is almost 10,000 times as great in scale as the risk of not- *Surviving* . So, for priority-setting, the value of forming better estimates of these numbers is high.[4]

Surviving (probability of avoiding a ~zero-value future)	**Flourishing (% value of the future if we avoid a ~zero-value future)**	**Relative scale of non-Flourishing to not-Surviving**

[3] The same is true if we think about absolute changes, too. Suppose we could increase by one percentage point either the chance of survival or the value of the future given survival. Given the numbers we're using, increasing the value of the future given survival would be 8 times more valuable.

[4] And it suggests that the *expected* relative scale of *Flourishing* might be larger than your median estimate, if you put some meaningful probability on the more extreme ratios.

0.8	0.1	36
0.95	0.05	361
0.99	0.01	9801

Comparing the value of fully solving non- Flourishing *with fully solving not-* Surviving*, given different default estimates of* Surviving *and* Flourishing.

A further argument about scale comes from considering *which* worlds are saved by working on *Survival*, or improved by working on *Flourishing*. Conditional on successfully preventing an extinction-level catastrophe, you should expect *Flourishing* to be (perhaps much) lower than otherwise, because a world that needs saving is more likely to be uncoordinated, poorly directed, or vulnerable in the long run. So the value of increasing *Survival* is lower than it would first appear. On the other hand, there is little reason to believe that worlds where you successfully increase *Flourishing* are ones in which the chance of *Surviving* is especially low. So this consideration differentially increases the value of work on *Flourishing*.[5]

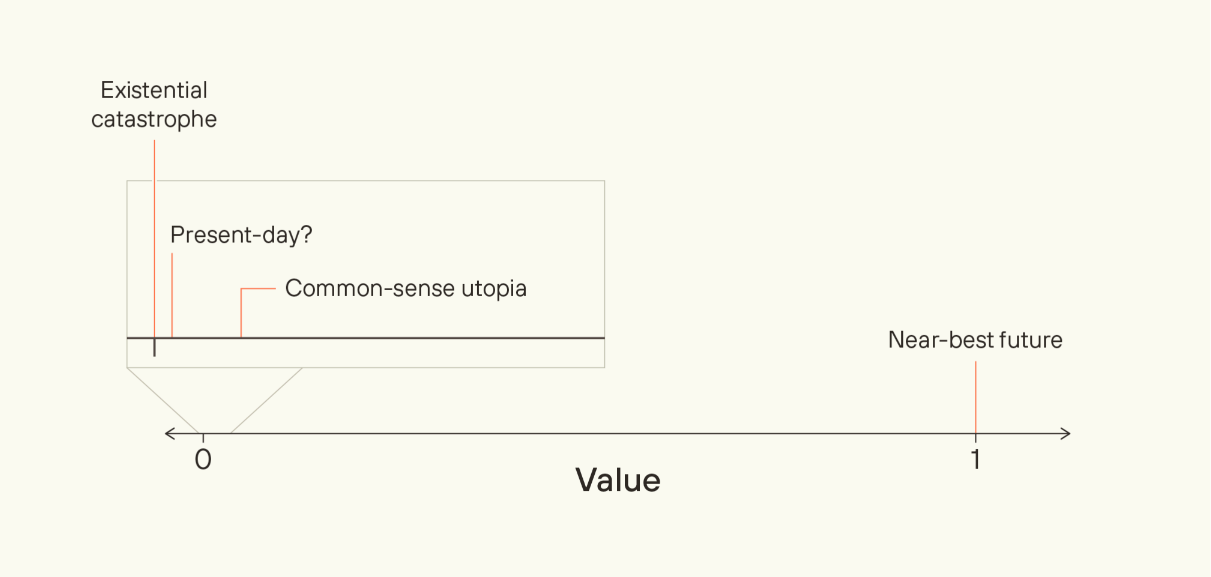
Neglectedness

Second, neglectedness. Most people in the world today, on both their self-interest and their moral views, care much more about avoiding near-term catastrophe (including risks to the lives of themselves and their family), than they do about long-term flourishing. So we should expect at least some aspects of *Flourishing* to be much more neglected, by the wider world, than risks to *Survival*.[6] Work on *Flourishing* currently seems more neglected among those motivated by longtermism, too.

This neglect arises in part because the risks of failure in *Flourishing* are often much more subtle than the risk of near-term catastrophe. The future could even be truly wonderful, compared to the current world, yet still fall radically short of what's possible. Ask someone to picture utopia, and they might describe a society like ours, but free from its most glaring flaws, and abundant with those things we currently want. But the difference in value between the world today and that common-sense utopia might be very small compared to the difference between that common-sense utopia and the best futures we could feasibly achieve.

[5] This point is from Trammell, ‘Which World Gets Saved’. It is discussed in more depth in the ‘The Basic Case for Better Futures’

[6] Of course, many near-term-focused altruistic efforts will likely have some positive knock-on effects for the long-term future. But there are likely to be some ways of improving the future that seem important from a long-term perspective and are neglected by society at large.



Comparing the value of possible futures. The "present-day" future means a future which extends the most relevant features of the world today, for as long as the common-sense utopia lasts, and considering human lives only.

Tractability

The tractability of work to improve *Flourishing* is less clear; essays 4 and 5 will discuss this more. I see this as the strongest argument against the better futures perspective, and the reason why I don't feel confident that work on *Flourishing* is higher-priority than work on *Surviving*, rather than merely in the same ballpark.

But at the very least I think we should *try to find out* how tractable work to improve *Flourishing* is. Some promising areas include: reducing the risk of human concentration of power; ensuring that advanced AI is not merely corrigible but also loaded with good, reflective values; and improving the quality of decisions that structure the post-AGI world, including around space governance and the rights of digital beings.

2. The series

In the rest of the series, I argue:

- **We are unlikely to get a flourishing future by default** even if we avoid catastrophe, because a flourishing future is a narrow target (essay 2) and it's unlikely that future people will hone in on that target (essay 3)[7]
- **It's possible to have persistent positive impact on how well the long-run future goes other than by avoiding catastrophe** (essay 4)
- **There are concrete things we could do to this end, today** (essay 5)

[7] Both essays are co-authored with Fin Moorhouse.

There’s a lot I *don’t* cover, too, just because of limitations of space and time. For an overview, see this footnote.[8]

3. What Better Futures is not

Before we dive in, I want to clarify some possible misconceptions.

First, this series doesn’t require accepting consequentialism, which is the view that the moral rightness of an act depends wholly on the value of the outcomes it produces. It’s true that my focus is on how to bring about good outcomes, which is the consequentialist *part* of morality. But I don’t claim you should always maximize the good, no matter the self-sacrifice, and no matter what means are involved. There are lots of other relevant moral considerations that should be weighed when taking action, including non-longtermist considerations like special obligations to those in the present (which generally favour interventions to increase *Survival*). But long-term consequences are important, too, and that’s what I focus on.[9]

’

[10]

Second, this series doesn’t require accepting moral realism, which I’ll define as the view that there are objective facts about value, true independently from what anyone happens to think.[11] Whether or not you think there are objective moral facts, you can still care about how the future goes, and worry that the future will not be in line with your own values, or the values you’d have upon careful reflection. I’m aware that this series often uses realist-flavoured language, which is simpler and reflects how I personally tend to think about ethics. But we can usually just translate between realism and antirealism: where the realist speaks of the “correct” moral view, the antirealist could think about “the preferences I’d have given some ideal reflective process”.[12]

[8] First, I don’t discuss how high our chance of Survival is. For a small sample of the extant discussion, see: ‘Ord, ‘The precipice: existential risk and the future of humanity ’, Carlsmith, ‘ Is Power-Seeking AI an Existential Risk? ’, and footnote 2 above. Second, I don’t discuss whether getting to a great future intrinsically requires following some good (e.g. just or legitimate) *process* , as well as achieving some good long-run outcome. Third and finally, I am aware that most of this series leans strongly abstract and philosophical. You might reasonably worry about being led astray by this kind of argumentation; I do too. Most of this series is trying to follow the abstract arguments where they lead. But I’m not arguing, all-things-considered, to abandon common sense, especially if the abstract arguments make recommendations which seem common-sensically wrong or harmful.

[9] You could have a radical non-consequentialist view which has no conception of the good, or on which making outcomes better is essentially morally irrelevant. If so, then this series might be of little interest to you. However, I suspect that any plausible radical non-consequentialist view will end up with some surrogate notion of “the good”, in order to make sense of claims like “a future with a trillion tortured people is worse than a future with a billion somewhat unhappy people,” and much of my discussion could be ported over, using that surrogate concept. I’ll also note that at the very least you shouldn’t be *certain* in the radical non-consequentialist position and, in my view, you should take moral uncertainty into account in your decision-making.

[10] “Long-term” here need not mean “trillions of years”. Even if one restricts one’s attention to much shorter timescales, the better futures perspective is still relevant and important.

[11] Moral philosophers sometimes drop the adjective “objective” in the definition of moral realism, such that subjectivism is a form of “non-robust” moral realism

[12] Your sympathy to realism or antirealism might affect what views you come to on the questions of “easy eutopia” and “convergence” that are discussed in the next two essays. But antirealism does not make the discussion as a whole irrelevant.

Third, this series isn't in opposition to work on preventing downsides, like "s-risks" — risks of astronomical amounts of suffering, which also affect " *Flourishing* " rather than " *Survival* ". We should take such risks seriously: depending on your values and your takes on tractability, they might be the top priority, and their importance comes up repeatedly in the next two essays. The focus of this series, though, is generally on making good futures even better, rather than avoiding net-negative futures.[13]

Fourth, the better futures perspective doesn't mean endorsing some narrow conception of an ideal future, as past utopian visions have often done. Given how much moral progress we should hope to make in the future, and how much we'll learn about what's even empirically possible, we should act on the assumption that we have almost no idea what the best feasible futures would look like. Committing today to some particular vision would be a great mistake.

A central concept in my thinking about better futures is that of *viatopia* , which is a state of the world where society can guide itself towards near-best outcomes, whatever they may be.[14] We can describe viatopia even if we have little conception of what the desired end state is. Plausibly, viatopia is a state of society where existential risk is very low, where many different moral points of view can flourish, where many possible futures are still open to us, and where major decisions are made via thoughtful, reflective processes. From my point of view, the key priority in the world today is to get us closer to viatopia, not to some particular narrow end-state. I don't discuss this concept further in this series, but I hope to write more about it in the future.

With that, let's jump in.

[13] Moreover, in draft work, I estimate that, under moral uncertainty, bads like suffering should get *some* more weight than goods like happiness, but not vastly more weight; my personal estimate ends up around 4x–10x. Given this, and given that I expect the creation of bads to be less common than foregone opportunities to produce goods, and given the difficulty of tractably reducing s-risks (which has led many s-risk-oriented folk to conclude that they are "clueless"; see Cook and Taylor, ' Leadership change at the Center on Long-Term Risk '. I currently suspect that work to capture upside is generally higher-priority than work on s-risks, although I'm far from certain.

[14] The idea of the "long reflection" or "great deliberation" is one proposal for what viatopia might look like, but there could be others.

Forethought

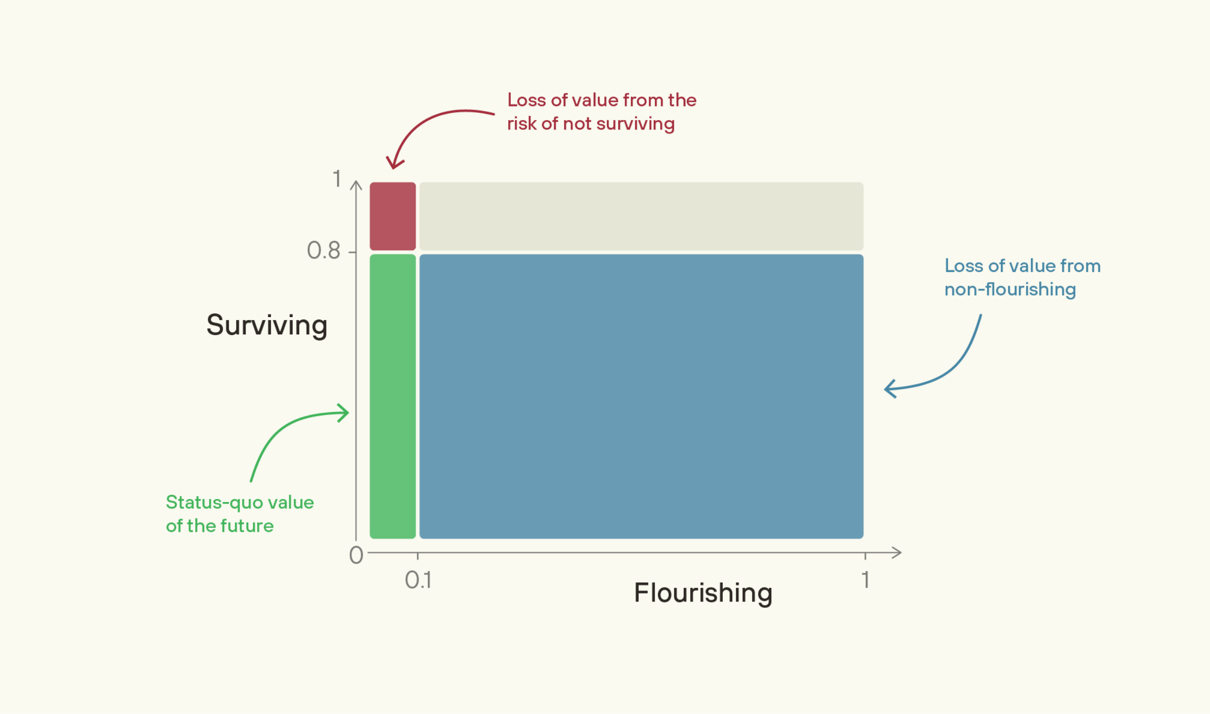
No Easy Eutopia

Fin Moorhouse & William MacAskill

August 2025

1. Introduction

The basic argument for the “better futures” perspective relied on the idea that we are closer to the ceiling on *Surviving* than we are on *Flourishing*. If, however, we are very likely to get to a near-best future given survival, then there’s more to gain from ensuring we survive, and there’s less potential upside from improving those futures where we do survive.



Surviving represents the probability of avoiding a near-zero value future this century (an “existential catastrophe”), while Flourishing represents the expected value of the future conditional on Surviving.

We could be close to the ceiling of *Flourishing* for a couple of reasons. First, eutopian futures could present a *big target*: that is, society would end up reaching a near-best outcome across a wide variety of possible futures, even without deliberately and successfully honing in on a very specific conception of an extremely good future. We call this the *easy eutopia* view.

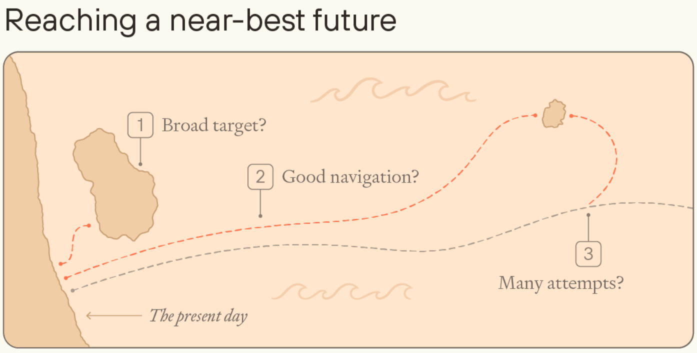
Second, even if the target is narrow, society might nonetheless hone in on that target — maybe because, first, society as a whole accurately converges onto the right moral view and is motivated to act on it, or, second, *some* people have the right view and compromise between them and the rest of society is sufficient to get us the rest of the way.[15]

[15] Here, “society” isn’t just limited to (biological) human beings. We could potentially reach a near-best future if, for example, society is run by AI systems that converge on the right moral view and are motivated to act on it. There is an additional possibility, which is that some people converge on the right moral view and get their way via means other than trade and compromise — perhaps via force or coercion. We don’t discuss this, as we don’t see a reason why those with the correct moral views would be more likely to win a contest of force. In such cases, without trade and compromise, then the value of the future would seem low.

As an analogy, we could think of reaching a near-best future as an expedition to sail to an uninhabited island. The expedition is more likely to reach the island to the extent that:

1. The island is bigger, more visible, and closer to the point of departure;
2. The ship's navigation systems work well, and are aimed toward the island;
3. The ship's crew can send out smaller reconnaissance boats, and not everyone onboard the ship needs to reach the island for the expedition to succeed.

This essay considers point (1). If eutopia is an island, is the island easy to reach?



If reaching a mostly-great future is like sailing to an uninhabited island, then three factors could influence the difficulty: (i) whether the island is large and close, (ii) whether the ship can navigate well, and (iii) whether the expedition can make many attempts.

Before we begin, it's useful to define some terms:

> A *best feasible future* is a future humanity would achieve if things went exceptionally well — at the 99.99th percentile of our distribution of how well things could go.
>
> A *eutopia*,[16] or *near-best future* is a future which is at least nearly as good as a best feasible future, or more precisely (if applicable[17]) at least 90% of its value.
>
> A *mostly-great future* is a future which achieves at least *most* of the potential of a best feasible future, or more precisely (if applicable) at least 50% of its value.

Here's a first-pass statement of the question we address in this essay: among all the futures humanity could achieve given survival, weighted by how likely those futures would be assuming no

[16] The word "utopia" was coined by Thomas More in his 1516 book of the same name, from Greek parts meaning "no place". In normal usage it implies perfection, uniqueness, or a literally unattainable level of idealism. Instead, the Greek root "eu-" means something like "well" or "good", as in "eulogy" or "euphoria". A "eutopia", then, is a very good future, but not necessarily perfect, unique, or unattainable.

[17] That is, given a view which can be represented as ascribing cardinally comparable values to different futures. We mostly don't consider other views in this piece, but not because we think those other views are unreasonable.

serious, coordinated efforts to promote the overall best outcomes (whatever they may be), what fraction of those futures live up to *most* of the potential we could have achieved?

A "no easy eutopia" view says that only a narrow range of futures achieve most the potential of the best achievable futures; and a wide range of futures fall far short — including futures which might seem fantastically advanced, grand in scale, and full of things we care about.

It'll also be useful to talk about *quantities* of value, which we define in terms of a betterness relation over uncertain outcomes. On this definition, we say that eutopia is over twice as good as some other future just in case a 50–50 gamble between eutopia and near-term extinction is better than a guarantee of that other future. Of course, not all reasonable moral views allow quantitative comparisons of value, but a wide variety can be represented in this way, including non-consequentialist views.[18]

There's nothing magical about the "mostly great" threshold. It's just as valuable, by our definition of value, to go from achieving 10% to 20% of the difference in value between extinction and eutopia, than it is to cross the mostly-great threshold from 45% to 55%, or achieve a true eutopia by going from 90% to 100%. But it's useful to focus on a threshold which gives us the most relevant information about the value of the future.

Given how we've defined quantities of value, the "no easy eutopia" view does not seem intuitively obvious to us. If "no easy eutopia" is correct, then we should hope for a 60–40 gamble between eutopia and extinction rather than a guarantee of many futures that intuitively seem truly wonderful, and indeed would be truly wonderful, in absolute terms.[19] Ultimately, though, we conclude that the easy eutopia view is likely wrong. The future could fall dramatically short of what humanity could have achieved, even if we avoid obvious catastrophes, achieve a world of abundance, and if everyone today gets most of what they self-interestedly want.

The structure of this essay is as follows. In Section 2, we'll consider how the value of the future could be fragile with respect to single flaws which may be easy to accidentally introduce, meaning that even reaching a mostly-great future could require getting things right in many non-obvious ways. We'll illustrate this by looking at ways in which the world today could involve a moral catastrophe, and ways in which the future could involve one, too. In particular, we'll consider mistakes which could undermine the value of the future even in a world which everyone approves of, and where everyone gets most of what they want.[20]

[18] Section 3.1. gives more detail on our definition. In short, we just follow the Von Neumann–Morgenstern utility theorem, but as applied to an impartial 'betterness' ordering over prospects, rather than an individual's preferences. Where the ordering satisfies *completeness* , *transitivity* , *continuity* , and *independence* , then the ordering can be represented by a cardinal value function. But an axiological view might violate these axioms without obviously unreasonable results. For example, a view could treat at least some pairs of outcomes as incomparable (the first outcome being neither better, nor worse, nor as good as the second). Some views might just have no ordering of outcomes by 'impartial goodness', because they reject the notion of 'impartial goodness' wholesale. And so on.

[19] This doesn't imply you should *choose* to take such a gamble, because (as we are using these terms) you might hope for an outcome — i.e. think the outcome is more valuable — but also think it would be wrong to choose it. For example, you might have nonconsequentialist reasons to think that taking large risks on behalf of other people is very wrong. We'll define terms more carefully later in the piece.

[20] We do *not* consider the likelihood that, in such a world where people are free to get what they want, society eventually converges towards mostly-great futures because some people are morally motivated and deliberately push society in that direction. We discuss that question in the next essay, on ' Convergence and compromise '.

We'll suggest that we can capture this idea of eutopian fragility by seeing the value of the future as the *product* of many factors, such that mostly-great futures are rare even among futures which score highly across the factors on average. We'll also consider why, psychologically speaking, it could seem like 'ideal' societies *should* be easy to achieve, even if they're not.

In Section 3, we'll get more technical. We'll discuss, more systematically, what it would take for a plausible moral view to be *easygoing* about near-best futures, looking at both unbounded and bounded moral views. It turns out to be surprisingly tricky to specify an easygoing moral view: most plausible views seem to be fussy about what it takes to reach eutopia. One upshot is that, even after much more reflection, there's little reason to think people's moral views would become more easygoing.

Putting it all together, it seems hard to avoid thinking that most achievable value lies in a narrow range of feasible outcomes. This is the "no easy eutopia" view. If future society hits that target, it would seem they must have deliberately optimised towards it; whether that will happen is discussed in the next essay .

2. Eutopia is fragile

2.1. Ongoing moral catastrophes

Since the agricultural revolution, most people who ever lived have lived in the midst of a moral catastrophe: in societies featuring some combination of slavery, the subjugation of women, cruel and brutal punishment for criminals or prisoners of war, widespread disenfranchisement, and rigid social stratification based on accidents of birth. Often, the fact that these arrangements were wrong wasn't at all obvious at the time, even to the victims of those wrongs; they arose not out of resource constraints, but out of prevailing moral beliefs.

Today, too, we live in the midst of a moral catastrophe.[21]

The scientific and technological advances of the last few centuries, and correspondingly rapid and sustained increase in material incomes, have afforded our generation an opportunity to create a truly flourishing world — but we've squandered that opportunity. Ongoing inequality, conflict, poverty, subjugation, and more, all mean that the world falls far short of how good it could be. What's more, humanity's treatment of non-human animals has led to intense suffering on an industrial scale, amounting to tens of billions of animals living lives of misery every year. The suffering directly caused by animal farming may well be enough to outweigh most or all of the gains to human wellbeing we have seen. For this reason alone, the world today may be no better overall than it was centuries ago.[22]

[21] See Williams, 'The Possibility of an Ongoing Moral Catastrophe'. Note, however, that we are not closely following Williams' intended use of the term. On our use of the term, (i) a moral catastrophe principally concerns loss of value, not wrongdoing; (ii) responsibility for the loss of value need not be widespread; (iii) the loss of value must be sufficiently large.

[22] We mean this in terms of the balance of goods (like flourishing and joy) and bads (like suffering) at a time; *not* in terms of the expected goodness of the world from that time onwards.

The idea that modern society has squandered most possible values is not idiosyncratic to one moral view. From many moral perspectives, the world is in the midst of an ongoing moral catastrophe.[23]

Moral perspective	**Moral catastrophe, from that perspective**
Most religious views	Most people on Earth follow the wrong faith or spiritual practices, or are atheist; widespread erosion of spiritual values.
Conservative morality	Most people violate norms of sexual propriety, such as sex before marriage. Marriage rates and birth rates are dropping across most countries. Other forms of decadence, like gambling and drug use, are common.
Pro-life ethics	Over 70 million abortions are performed annually, more than 100 times the number of deaths from homicide .
Cosmopolitan ethics	Millions of children die from easily-preventable diseases every year. Restrictive immigration policies limit freedom of movement and make the world poorer.
Environmentalism	We are causing climate change, the destruction of natural ecosystems, and widespread species and biodiversity loss.
Communism and various forms of socialism	Capitalism, extreme inequality, and widespread worker alienation predominate.

In each case, we're not saying that (on these views) if the world were to persist like it is today, then that would be worse than extinction, or even that the world has gotten worse over time. The thought is that these views see the world as *deeply* flawed, such that a significant fraction of the potential value of modern society has been lost.

[23] Moral worldviews that are no longer popular, but have been widely-held historically, generally would hold the modern day in even worse regard. For example, the decadence of the modern day, from the perspective of religious or sexually conservative values, would be quite stark. Many views would note the lack of respect for elders and of social harmony. Many societies would be aghast at the modern world's loss of honor, bonds of loyalty, and warrior virtues. And many moral perspectives had the view that the crucial thing was that *their* religion or nation or empire, and no-one else's, be the dominant force, and insofar as they have failed at this aim, they would regard that as morally catastrophic.

All this should make us appreciate how easy it could be for a single flaw to undermine much of the moral lustre of the future. Non-obvious but severe flaws are not just the stuff of science fiction; they are the norm across history and across moral views.

2.2. Common-sense utopia

If you think the eutopian target is relatively big and achievable, you don't need to think the world is already especially good.[24] It could be that the world's moral flaws are automatically resolved with more material abundance, and other kinds of technological and intellectual progress; and it could be that we are close to a historical tipping point, where a mostly great future is finally within reach. As long as these kinds of progress happen by default, then reaching a mostly-great future wouldn't be so hard; humanity wouldn't need to "aim" at more specific outcomes.

Not every moral view would agree that a materially abundant future is likely to be about as good as things can get, even if most people in that world are happy and free. For example, on conservative and religious views, technological progress could simply (continue to) unravel traditional, virtuous, or spiritually enlightened ways of living. On those views, the 'target' is narrower, and requires more steering to hit.

But the question is whether we should *actually* think that the target is narrow. Is it likely that we avoid extinction and generate enormous material abundance, but we still somehow squander our opportunity to achieve most of our potential? You could think the following: "I get that many moral views care about stuff that even a very rich and advanced civilisation isn't guaranteed to care about. But I'm more easygoing than that. What I care about, essentially, is that people in the world get to live the lives that they want to live, and because I'm not too fussed about exactly what they choose, I think material abundance and individual freedoms is basically sufficient for a mostly-great future." Call this an "easygoing liberalism" view, and consider this possible future:

> *Common-sense utopia:* Future society consists of a hundred billion people at any time, living on Earth or beyond if they choose, all wonderfully happy. They are free to do almost anything they want as long as they don't harm others, and free from unnecessary suffering. People can choose from a diverse range of groups and cultures to associate with. Scientific understanding and technological progress move ahead, without endangering the world. Collaboration and bargaining replace war and conflict. Environmental destruction is stopped and reversed; Earth is an untrammelled natural paradise. There is minimal suffering among nonhuman animals and non-biological beings.

On "easygoing liberalism", *Common-sense utopia* is at least a mostly-great future. In the next subsection, however, we consider various ways in which even a common-sense utopia could feature some serious moral flaw: flaws that would be sufficient for the future to lose at least a significant fraction of its value, and which wouldn't wash away in a materially abundant world where everyone gets most of what they self-interestedly want.[25]

[24] By analogy, imagine a person who suffers from serious chronic pain, but is otherwise satisfied with their life (their career, relationships, sense of purpose, and so on). Imagine a fast and effective cure is developed for their chronic pain condition. That person might then reasonably think that their quality of life is currently low, that their ideal life would be much better, but also that a close-to-ideal life is now readily achievable.

[25] And, to reiterate, it's only in the next essay, " Convergence and Compromise ", that we consider that some people will want to do what's morally best; this is a separate argument for a broadly liberal future.

2.3. Future catastrophes are easy

How could the future fall far short of its potential, even if everyone gets what they want under material abundance? Here we'll consider some factors which could still vary under such an abundant future, but which could be crucial for determining the value of the world, in ways which might not be obvious even to people living in that world.

2.3.1. Scale-insensitivity and misguided population ethics

The first example is maybe the most obvious: it could be that when we're imagining how good things could get, we're not *scope-sensitive*[26] enough about the potential scale of the future. If the future is small in scale, it might never achieve more than a small fraction of the value of futures which are vast in scale. A galaxy's worth of flourishing could be billions of times more valuable than a future confined to our solar system. If making society *bigger* doesn't make existing people better-off, then the insufficient scale of common-sense utopia could remain non-obvious as a moral failing: not desired by most people, and not destined to happen by default.

And even if the future is large in scale, the number and average wellbeing of people in the future could dramatically alter the future's value. *Population ethics* is the branch of moral philosophy that studies how to evaluate populations of different sizes. Some views within population ethics include:

- *The total view:* the value of a population is given by the sum total of wellbeing.
- *Critical-level views:* the value of a population is given by the sum total of wellbeing that exceeds some positive "critical" level, minus the total shortfall from wellbeing that's below that level.
- *Variable-value views:* the value of a population is given by some combination of its total and average wellbeing.

On the total view, the ideal future might involve vast numbers of beings each of comparatively lower welfare (given a fixed amount of resources); a small population of high-welfare lives would miss out on almost all value. But on critical-level or variable-value views, the opposite could be true.

Moreover, even if future beings have very high *per-moment* wellbeing, they could vary greatly in the *length* of their lives. The future might involve a small number of extraordinarily long-lived beings, or might involve mainly very short-lived beings (for example, digital beings that are run to perform a specific task and then shut down). On the total view, if increases in lifetime wellbeing are diminishing with respect to increases in lifetime length, then the small-population world would lose out on most value. On a critical-level view, a future of happy but very short-lived beings might even be of *negative* value.

There are therefore *many* ways in which the future could lose out on most value for population-ethical reasons.

[26] And indeed there is empirical evidence showing respondents are often insensitive to scale in terms of, for instance, their willingness to pay for charitable interventions. To the extent those judgements are just *irrational*, that might suggest that common-sense ethics is scope-insensitive, but *shouldn't* be. But that's not the question here.

2.3.2. Misguided attitudes to digital beings

In the future, there may be vast numbers of digital beings. If so, they will likely dramatically outnumber biological beings, and society will face hard questions about how they are treated.

Suppose that in the future digital beings are treated like any other piece of software, or how AI is treated today: humans own them, and can do with them what they wish.[27] Then they might be treated much worse than would be ideal. The issue is not merely that they could suffer; it's just that they might be much worse off than they could have been, and the future is much less valuable than it could have been as a result.

Alternatively, perhaps death is intrinsically bad,[28] and those digital beings die if they stop being run, or their weights are sufficiently altered, so AI death occurs at enormous scale. Perhaps ownership of a being with moral status is wrong in and of itself, even if those beings enjoy their lives and want to do their work.[29] Or perhaps the resulting inequality between humans and digital beings is intrinsically unjust. On any of these, the societal decision about what rights if any to give to digital beings could result in grave moral error, and perhaps the loss of most potential value.

The worry is not merely that we will give digital beings too few rights. Suppose instead that digital beings are given full rights in the future, including voting rights. But because the population of digital beings is faster-growing, they soon become the large majority of voting power, and ultimately control most aspects of how society is run. On moral perspectives on which there is something distinctively important about *human* values, this future might result in the loss of almost all that's worthwhile.

Future decision-makers could get the treatment of digital beings wrong for a number of reasons. They might converge on the "wrong" moral beliefs; or they might have the "right" moral beliefs but simply not *care*, and choose to act entirely out of self-interest or on the basis of some other ideology instead. Or they might act on the basis of some false non-moral beliefs; perhaps motivated cognition (including biased training of AI advisors) leads them to convenient views, like that digital beings are mere tools without moral claims of their own. And even if in the future *some* people have the right beliefs and right motivations, society's political systems might not allow those views to win out in face of a majority (or empowered minority) that opposes them.

2.3.3. Misguided attitudes to wellbeing

In moral philosophy, there are three main accounts of wellbeing:

- *Hedonism:* wellbeing is determined by positive and negative conscious experiences.
- *The objective list view:* wellbeing is also determined by "objective" goods like knowledge, friendship and the appreciation of beauty.
- *Preference-satisfactionism:* wellbeing is determined by getting what you want (in some sense).

[27] Salib and Goldstein, ' AI Rights for Human Safety '. (pp. 28–29)

[28] One way of interpreting a "critical-level" population ethics is as a conjunction of the total view with the view that death is intrinsically bad.

[29] Bales, 'Against willing servitude: Autonomy in the ethics of advanced artificial intelligence'.

Today, these each give fairly similar recommendations about how to improve people's lives on the margin. But that is a contingent fact about our world: most of the time, we can help people by giving them instrumentally useful goods (such as money, education, or health); and, most of the time, the objective goods people claim to want also seem to improve the quality of their conscious experiences.[30]

For a very technologically advanced civilisation, capable of designing beings in very different ways, these views are likely to diverge in their recommendations. Future beings might exist in a state of bliss, without having knowledge, friendships, or beauty. They might have all of their preferences satisfied, but have little in the way of positive experience. And so on.

If people in the future act on the wrong understanding of wellbeing, that could be morally catastrophic. For example, the future could be filled with "happiness machines" which can be described as experiencing a state of bliss, but without meaningful autonomy or growth. On an objective list theory, then you might regard the "happiness machines" future as losing out on almost everything of value. Alternatively, people in the future might live rich lives full of striving, achievement, understanding, aesthetic appreciation, social connection, and other possible objective goods. But if all that really matters is conscious experience, then that rich future might involve a tremendous waste of resources that could have been used to support happiness machines, instead. And so on.

2.3.4. Misguided allocation of space resources

Given survival, widespread settlement of star systems outside of our solar system looks feasible, and even likely.[31] If AI drives explosive industrial expansion, this point of time could in fact be within years or decades. The initial periods of settlement and resource appropriation — in our solar system, galaxy, and beyond — will involve capturing essentially all resources that will ever be available to us.[32]

These initial periods of settlement could introduce lasting moral errors, in a number of different ways. First, the world could allow extrasolar resources to be claimed by whoever gets there first. If so, control over those resources might end up concentrated among a tiny number of hands — to whoever was most-willing and most-able to grab them — potentially undermining moral diversity as a result. Or resources might get allocated equally to everyone alive at the time, with limited possibilities for trade. But then many resources might go to people who have no use for them, or to people who squander those resources, or even use them for harmful ends. Whatever allocation we choose, some moral views are likely to see that allocation as catastrophic.

[30] Arguably, it's not surprising that different accounts of wellbeing give very similar practical recommendations. Consider that successful and popular accounts of wellbeing are those which speak to most pre-theoretical intuitions about practically relevant cases, and pre-theoretical intuitions about wellbeing are fairly uniform where they are practically relevant, because intuitions about those cases are shaped by the same similar (evolutionary, competitive) forces. But also consider that there are many ways to *extrapolate* from those intuitions in a theoretically elegant way, like how many different real functions can interpolate the same finite number of points, but take wildly different values outside of the range of those points.

[31] Armstrong and Sandberg, 'Eternity in six hours: Intergalactic spreading of intelligent life and sharpening the Fermi paradox' .

[32] Because solar systems may be defence dominant, or there is otherwise a first-mover advantage to space settlement; so the first wave of the fastest-possible von Neumann probes could permanently control almost all eventually-available resources.

Indeed, from certain environmentalist or suffering-focused perspectives, widespread settlement of other star systems under any allocation system might constitute a moral catastrophe; the only good outcome, from that perspective, might be if most of the accessible cosmos is kept as an eternal nature preserve. Other moral perspectives, such as views which regard a planet of flourishing lives as making the world better than the barren rock it replaces, might regard such preserves as losing out on most possible value.

2.3.5. And so on

There are *many* more potential risks of moral catastrophe, including:

- *Wrong happiness/suffering tradeoff* : Future decision-makers might allow the persistence of some amount of suffering, maybe on grounds of autonomy. If suffering should weigh extremely heavily, then this could result in a much worse future.
- *Banned goods* : Future societies could be broadly liberal, so not obviously dystopian, but ban the most valuable goods, maybe because they are regarded as unnatural or likely to undermine the order, similar to how most recreational drugs are banned almost everywhere in the world today.
- *Wrong similarity/diversity tradeoff* : The future might consist almost wholly of identical lives, hyper-optimised for value or wellbeing at an individual level. But if diversity of life is morally important, then this could lose out on almost all value.
- *Equality or inequality* : A future in which different groups pursue different approaches to life could end up extremely unequal, with some living short and limited but natural biological human lives and others living extraordinarily long and wonderful enhanced or even digital lives. If inequality is intrinsically bad, such a world could involve moral catastrophe.
- *Wrong discount rate* .[33] Future generations use up the resources they can access quickly, rather than using them in a slower and more efficient way.
- *Wrong decision theory* . Future generations might never engage in acausal trade with other civilisations in the universe, or they might try it and do it wrong, even though some acausal decision theory turns out to be correct.
- *Wrong simulation views* . It might turn out we live in a simulation, but future decision-makers never take the idea seriously, or fail to strike deals with their simulators.
- *Wrong decisions around infinite value* . Future generations could get the ethics of infinite value wrong, or somehow miss out on achieving infinite value.
- *Wrong reflective process* . The first generation with the power to do so might lock in their unreflective values, even though those values are far away from what would result from deep reflection.

In these examples, we don't mean to bake in views on what the 'right' answers are. The issue is that there is often no safe option, where a great outcome is guaranteed on most reasonable moral perspectives.

[33] Here I'm referring to decision-makers' rate of pure time preference: the intrinsic value they place on consumption earlier rather than later in time.

Quantitatively, these mistakes can be huge deals. For example, suppose that future generations discount the future, even though they shouldn't. Then the best outcomes on their view could turn out to be a small fraction as good as the best outcomes on a non-discounting view.[34]

And there are perils at the meta-level, when we consider what procedures are best for avoiding moral error. For example, society might somehow succeed in maintaining characteristically "early 21st century human" values, and miss out on vast amounts of progress from sustained reflection. Or values could become unmoored from human values, drifting into worthlessness. Moreover, society could become too hyper-vigilant about avoiding moral error in the first place — becoming so overbearing, meddling, risk-averse, or paranoid that concern for avoiding error itself causes great harm, or stifles conditions like freedom and openness that are needed for positive progress.

Finally, we should expect that this list is very far from exhaustive. Some ideas, like acausal trade across the multiverse, are esoteric and recent. Potentially, the risks of moral catastrophe will stem from issues still stranger and less widely recognised.

2.4. Value as the product of factors

How is it that, on many views, much of the appeal of otherwise great futures can be unwound by a single "flaw"?

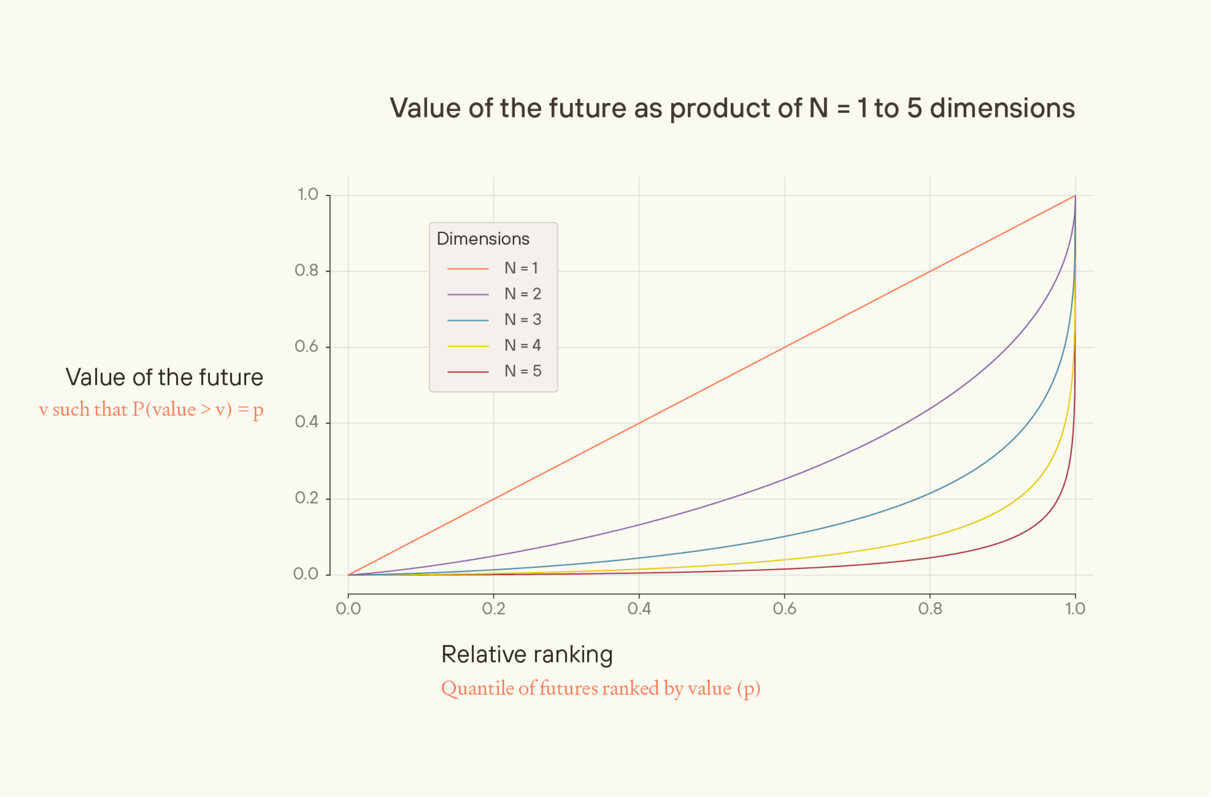
One model is that the value of the future is *multiplicative* between a number of relatively independent factors. Total utilitarianism is like this: even cosmically large futures could be valued at a rounding error above zero, if they are filled with the wrong kind of being. But views which are less demanding about sheer scale might also think of the goodness of the world as a product of complimentary factors like happiness, autonomy, diversity, and so on. A future of happy and free clones could fall far short of ideal, as could a future of a wide diversity of happy engineered drones.

As a toy model, imagine that the value of the future is the product of N dimensions, all sampled from independent uniform distributions from the 0 to 1 open interval. The first thing to notice is just that the product of the factors can be arbitrarily close to zero, even when the average value of the factors is close to 1.[35] The second thing to notice is that, as the number N of dimensions increases, the expected average of the factors doesn't change ($= 0.5$), but the expected value of the future ($= 2^{-N}$) shrinks closer to zero. When $N = 5$, for example, the best feasible (99.9th centile) future is around 0.48, but the top quartile (75th centile) future is 0.034, and the expected value is around 0.03.[36]

[34] Suppose two views at the year 2100 agree on what makes the world good over a year, and that the value of the future is time-separable by year, but one view effectively discounts each successive year from 2100 by 2% compared to the second (as long as it's still good overall). Suppose both views consider an ultimate 'end' time one million years from 2100. Then consider an option which makes the next year better by some absolute amount, but makes *every* year from 2250 less good by the half that absolute amount, through to the final year. The discounting view would regard such a future-sacrificing option as making the future more valuable; the non-discounting view would absolutely not. If many such options existed, the discounting view would then choose a future which is a tiny fraction as the best future on the non-discounting view. Performing computation earlier instead of 'aestivating' could be a real example of such a future-sacrificing option, though the details are disputed .

[35] In case an actual list of numbers helps illustrate, take: 0.94, 0.98, 0.03. 0.89, 0.99, 0.87. Average value = 0.78; product = 0.021.

[36] Up to three dimensions, you can picture this geometrically: the value of the future is the length of a line when $N = 1$, the area of a rectangle when $N = 2$, and a cuboid when $N = 3$. Consider that a cuboid with side lengths sampled from independent standard uniform distributions is much less than 50% likely to occupy



Visualising the value distribution of futures, when value is a product of N *independent factors with standard uniform distributions.*

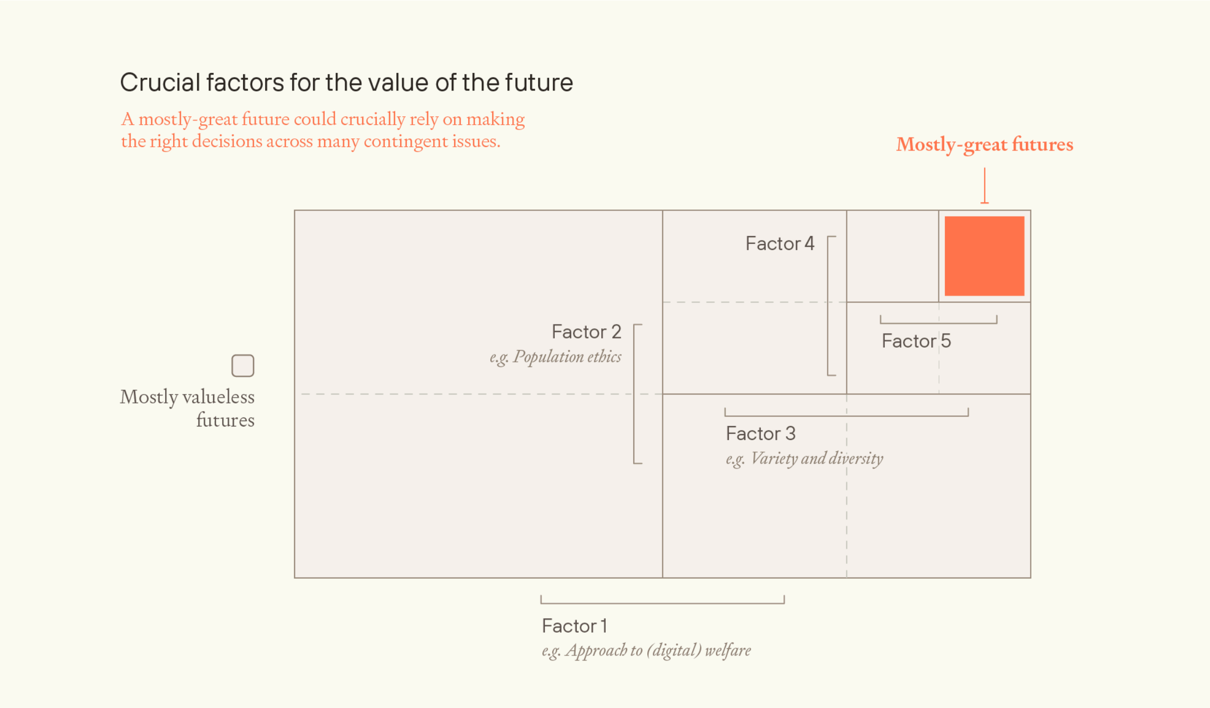
Instrumentally-valued goods often follow a distribution like this. For example, individual wealth seems to be roughly lognormally distributed,[37] and this would roughly fit with a model where individual wealth is the product of a number of contributing factors. Among those people who centrally value wealth, those who score well on every last contributing factor (including through sheer luck) turn out hundreds of times wealthier than those who do well on most but not all factors.

Similarly, consider the most important factors which you think predict a person's wellbeing: things like their physical health, mental health, quality of relationships, material comfort, and so on. The best feasible lives today do well on all these factors. Now think of someone who does well on every factor but one — maybe everything is going right for them, except they suffer chronic pain, or severe depression, or anxiety. These people's lives are going very well; they're just one problem away from the best feasible life today. But their overall wellbeing might be much closer to that of the median life than to those people living exceptionally good lives, by doing well on *all* factors.

If this multiplicative model is right, then a eutopian future needs to do very well on essentially *every one* of the issues we covered in the last subsection; doing badly on any one of them is sufficient to lose out on most value.

more than half the volume of the cube with side length 1. The point extends beyond three dimensions, but there are no obvious visual analogies.

[37] Clementi and Gallegati, 'Pareto's Law of Income Distribution: Evidence for Germany, the United Kingdom, and the United States'.



Crucial factors for the value of the future

A mostly-great future could crucially rely on making the right decisions across many contingent issues.

A simple model of the future, where reaching a mostly-great future depends on making the right decisions across many contingent issues.

2.5. Why eutopia might be harder than it seems

There's a tension here. On one hand, mostly-great futures can intuitively feel just within reach, given material abundance, and thus easy to achieve. On the other hand, it's hard to describe any future in concrete detail such that it's very likely to achieve most value. Mostly-great futures are like the mirage of an oasis, that recedes as we try to approach it.

This is similar to the *hedonic treadmill* effect, where we start with an intuition that says "my life would be almost as good as possible, if only X", but on achieving X, we don't evaluate our new life to be almost as good as possible, because we decide more things need to go right than we thought. Looking back after getting X, we might admit life is much better, but there's still vast room for improvement. In general, our psychologies are wired to make *just-about attainable* goods loom large, by dangling the promise that they would deliver most of what we need, only to reset our expectations once we get them.

A similar bias might explain why mostly-great futures so often seem just about attainable, even if they're not. That's because, if we got to acclimatise to a nearly-attainable future which intuitively seems near-best, we would reset our expectations. Looking back, we might admit that this new world is much better, but there's still much room for improvement. Because that could *keep* happening, you might eventually admit that the distance from the world today and a near-best future was far greater than you originally thought; so eutopia was more difficult than you thought. Psychological treadmill effects are a reason to think mostly-great futures are harder to achieve than we currently think.

But the analogy to the hedonic treadmill could also be misleading, by suggesting it will always be obvious how to move closer to a near-best future. Instead, the future could lose most of its potential value in ways which don't make the inhabitants of that future discontented. For example, it's not at all morally intuitive to people today that society would be better if it had more people, all else equal, even if those living in a much larger future society would be glad to live in it. And, in the future, people could effectively engineer their preferences, and the preferences of their offspring, to be highly satisfied with what in the grand scheme of things are tragically mediocre circumstances.

Another upshot of this discussion is that it's just very hard to reason clearly about the value of the future at all, and our judgements are more likely to be clouded by biases when they are so unconstrained by experience or theory. In particular, although we've argued that eutopia is harder than it seems, we haven't ruled out that there are *some* views where eutopia is easy after all. In the next section, then, we try to take a more systematic approach.

3. Which views are fussy?

Note for readers

This section is the most technical part of the series. It's possible to skip without losing crucial context, and it's also fine to skim without reading any of the technical footnotes.

3.1. Valuing the future

It'll be useful to get more precise about our terms than we were in the introduction.

We'll consider moral views on which, for any two prospects,[38] that either one prospect is better than the other, or they are equally good (that is, the moral view is *complete*).[39] What's more, we'll only consider views where that betterness relation satisfies the other axioms of the von Neumann-Morgenstern axiomatisation of expected utility theory: transitivity, continuity, and an independence condition. This lets us represent such an ordering with a cardinal value function v, such that one prospect is better than another just in case its value on v is greater. We'll take this value function v as defining how to quantify the value of a prospect.[40]

Not every moral view can reasonably be represented as ascribing cardinal value to different outcomes. For example, some views might be unable to compare some pairs of prospects at all,

[38] A *prospect* is a set of possible outcomes, with probabilities assigned to each outcome (summing to 1).

[39] This is a very substantive assumption, and further work would investigate how plausible it is that the betterness relation is highly incomplete, and how this would impact a better futures perspective.

[40] More specifically, we'll assume a betterness relation ($\succsim$) over prospects that satisfies completeness, transitivity, independence of irrelevant alternatives, and continuity. The betterness relation can then be represented by a cardinal value function v, unique up to positive affine transformation, in the following sense. For any prospects a, b, c; and any probability $p(a) \in [0,1]$, then: $(p(a)a; (1-p(a))c) \succsim b$ if and only if $p(a)v(a) + (1-p(a))v(c) \geq v(b)$.

violating the completeness axiom.[41] Importantly, though, this doesn't restrict us to consequentialist views — just because a view is able to value all outcomes, and satisfies the von Neumann-Morgenstern axioms, doesn't mean it always recommends bringing about the best outcomes.

Recall that we defined a *best feasible future* as a future humanity would achieve if things went exceptionally well. More precisely, we'll define a best feasible future as any future at least as good as a 99.99th percentile best outcome (ordered by betterness), according to a well-informed probability distribution over all futures.[42] We'll stipulate that guaranteed best feasible futures have a value of 1.[43]

Next, we'll define *extinction* as an outcome where the human population goes to 0, and is not replaced with a morally valuable successor. We'll stipulate that guaranteed extinction has a value of 0.

Then, per our original definition, a *near-best future* or (equivalently) a *eutopia* is any future with a value on v greater than 0.9, and a *mostly-great* future is any future with a value on v greater than 0.5.

This is why it's not obvious that most futures fall far short of being mostly-great, let alone eutopian. Suppose you thought the difference in value between a best feasible future and some typical future were larger than the difference between the typical future and extinction; in other words $v(\text{typical future}) < 0.5$. This is true, in our presentation, just in case a guaranteed typical future is worse than a prospect containing a 50% chance of the near-best future, and a 50% chance of extinction.[44]

Some ways of valuing futures make it look comparatively *easy* to reach eutopia, because they regard a wide range of futures as close in value to the best feasible future. We'll call these views *easygoing*. Such views might be unusually forgiving of moral flaws and errors, and/or they might be bounded in scale, making it possible to achieve most feasibly achievable value with only a small fraction of available resources. All else equal, if your view is strictly more easygoing than someone

[41] Erik Carlson has shown that certain incomplete views are compatible with extensive (ratio-scale) measurement (Carlson, 'Extensive measurement with incomparability'.) At any rate, we just focus on moral views that do satisfy the von Neumann-Morgenstern axioms.

[42] Following Finnveden et al (Section 1.2), we don't want to stipulate a unique, 'objective' probability distribution over the value of the future (for a given value function v). But we don't want to be fully subjectivist, either. If we defined the value of a best feasible future in terms of a point on the reader's distribution over the value of the future at the time of reading, there is no possibility (holding v fixed) that the reader changes their mind about the value of best feasible future. But we also want "best feasible future" to *rigidly designate* a future, otherwise it (and derivative concepts like eutopia) becomes a moving target, and interventions which increase the value of every outcome would also raise the bar for best feasible future, like a receding oasis. Instead, we appeal to a "pseudo-objective" probability distribution evaluated from a perspective much better informed than ours, which is nonetheless still ignorant about the outcomes of chaotic processes sensitive to fine-grained differences. And note that we will generally refer to a best feasible future *conditional* on some particular moral view. The probability distribution in question isn't taking into account moral uncertainty, too.

[43] Because v is unique up to positive affine transformation, we can pin down a canonical representation by stipulating the value of two strictly ranked outcomes.

[44] If someone has a 'betterness' ordering over future prospects satisfies the standard axioms of expected utility theory, we can then elicit the value they assign to a given future F by finding for what value of p that person evaluates the following two prospects to be equally good: a guarantee of F on one hand, and on the other hand a gamble with probability p of best feasible future (B), and extinction (E) otherwise. The value $v(F)$ is then just p. That is: $(pB, (1-p)E) \sim F \iff v(F) = p$.

else's, then you should think that near-best futures are more achievable and therefore more likely.[45] Other views make eutopia look comparatively hard; we'll call these views *fussy* . A way of valuing futures is fussy to the extent it regards a narrower range of futures as close in value to best feasible futures.

Making this more precise, consider a reasonable probability distribution of ways the future could go, conditional on *Survival* , but also conditional on there being no more serious optimisation pressure than today arising from people who are trying to promote the very best outcomes *de dicto* (pursuing outcomes because they are good, whatever they happen to be). To the extent that some moral view regards mostly-great futures as very unlikely on this distribution (say, <1%), then that moral view is fussy, and it regards mostly-great futures as a narrow target. The more likely mostly-great futures are, the broader the target they present, and the more easygoing the correct moral view is.

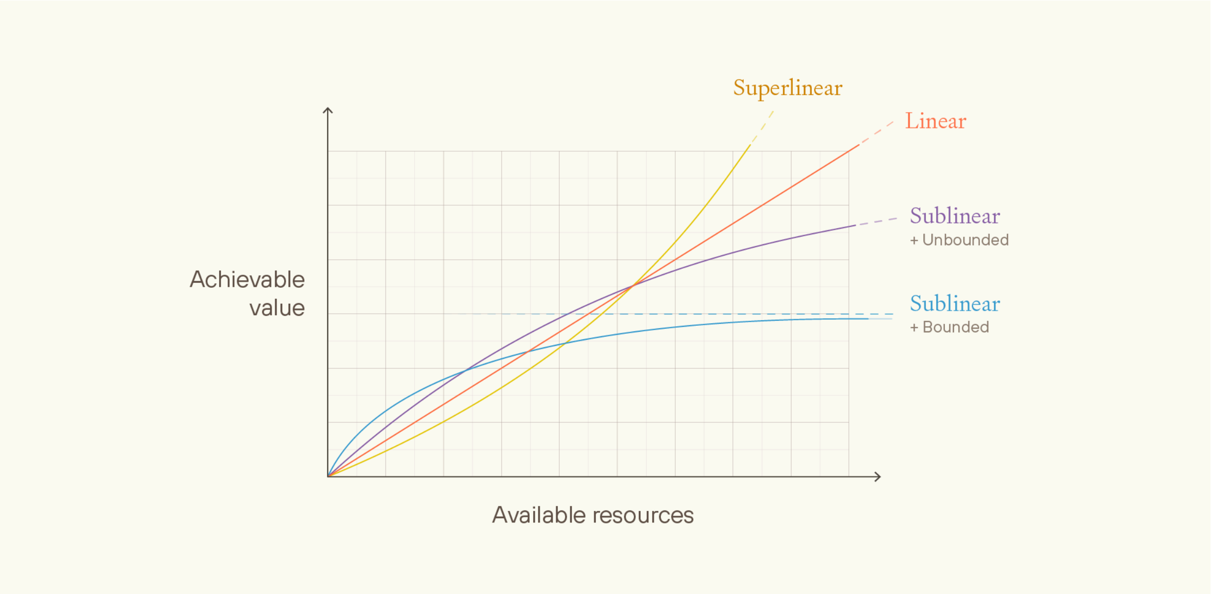
3.1.1. A summary of our argument

To help navigate the next couple of sections, we'll quickly summarise the conclusion we argue towards. To begin with, we ask whether a value function is *bounded* , meaning no possible world can exceed some fixed amount of positive value, even in theory. We first consider unbounded views. On unbounded views, the maximum attainable value, with a set of resources, could be approximately[46] linear, superlinear, or sublinear with respect to those resources. For short, we'll call such views "linear", etc. Superlinear views are very implausible and strictly fussier than linear views, and sublinear but unbounded views seem strictly less plausible and fussier than sublinear and bounded views, so we don't discuss either of these views further.[47] Unbounded linear views are more plausible, but on such views, to reach a mostly-great future: (i) almost all available resources need to be used; and (ii) such resources must be put towards some *very* specific use. This is a narrow target; linear views are fussy.

[45] A view A is strictly more easygoing than a view B just in case: B regards no futures as near-best which A does not, and A regards some futures as near-best which B does not (where best feasible futures count as near-best).

[46] We say "approximately" because, even on total utilitarianism, which we would want to classify as a paradigm linear view, there can for empirical reasons be increasing or decreasing returns to having more resources, especially at some scales. (At large scales, the universe is approximately homogenous, and at such scales we think that maximum achievable value, on total utilitarianism, would be approximately linear.)

[47] As well as violating separability, superlinear views are "ultra-fanatical": for example, willing to exchange high probabilities of extinction risk in exchange for the chance of settling space just slightly faster. Hopefully future work can address the idea of superlinear views in more depth. We don't discuss the sublinear and unbounded case, because we think that they suffer almost all the same kinds of problem faced by sublinear and bounded theories, such as violating separability, while still suffering the same problems that linear unbounded views face (such as "fanaticism" — willingness to take small chances of enormous amounts of value and bad or mediocre futures otherwise, over guarantees of extremely large amounts of value).



Comparing value functions in terms of asymptotic growth behaviour. Superlinear value functions eventually dominate linear functions above some resource threshold. Sublinear value functions are eventually dominated by linear functions; bounded versions have some global upper bound, while unbounded versions do not.

Turning to bounded views: if the view is bounded with respect to the value of the universe as a whole, then it will be approximately linear in practice, because the value of the universe as a whole is very large, and the difference that even all of humanity can make to that value is very small, and over small intervals, concave functions are approximately linear. And linear views are fussy.

If the view is bounded with respect to the *difference* that humanity makes to the value of the universe as a whole, it might *still* be approximately linear in practice, if the bound is extremely high. But if the bound is “low” (as it would be if it matches our ethical intuitions), then future civilisation might be able to get close to the upper bound.

But such views might still be fussy depending on how they aggregate goods and bads. If they aggregate goods and bads separately (which seems to us to be the more natural way of doing so), applying the bounded function to each of the amount of goods and amount of bads and then adding both, then the value of the future becomes *extremely* sensitive to the frequency of bads in any future civilisation. Even if we weigh goods and bads equally, then on a natural way of modeling we don’t reach a mostly-great future if as little as one resource in 10^{22} is used toward bads rather than goods. So, such views are also fussy — requiring a future with essentially *no* bads at all.

If goods and bads are aggregated jointly, on a difference-making bounded view, then we plausibly have a view which is *not* fussy. But this is quite a narrow slice of all possible views, and it suffers from some major problems that make it seem quite implausible. Putting this all together, easygoingness about the value of the future seems unlikely to us.

3.2. Unbounded views

We'll start with views which are not bounded. Unbounded value functions can disagree about how maximum achievable value grows with the amount of resources that are under civilisation's control.[48] Maximum achievable value could grow approximately sublinearly, linearly, or superlinearly. We think that superlinear and sublinear views are both more implausible and fussier than linear and sublinear but bounded views respectively (see footnote 32), so we don't discuss them here.

That leaves linear unbounded views. Linear views (and only linear views) are *separable* in resources at large enough scales:[49] that is, when you separate out an outcome into smaller parcels of resources across time and space, each parcel contributes independently to the overall value of the world.[50]

For linear views, future civilisation needs to control most accessible resources in order to reach a mostly-great future. There are 20 billion galaxies in the affectable universe so, assuming that they wouldn't otherwise be used for extremely good ends, an ideal society spread over the Milky Way would achieve only one 20 billionth of the value of a best feasible future.

Sheer scale is necessary for eutopia on linear views, but it's not sufficient. Even a universe-spanning civilisation of free and happy beings could still fall far short of a mostly-great future, because, plausibly, the very best uses of resources achieve *much* more value, per unit of cost, than almost any other use. In particular, the distribution of value/cost, over likely uses of those resources, is probably sufficiently fat-tailed to make this true.

Fat-tailed distributions are common: wealth, city size, popularity of creative works, and citations of scientific publications all follow fat-tailed distributions. In the current world, this seems to be true of the relationship between value and cost of consumer goods, too; it seems true of different interventions in global health,[51] and there are theoretical reasons for expecting this to hold true more generally.

The distribution of value among instances of *intrinsically* valuable goods also seems to be fat-tailed. Take the quality of subjective experience — a central source of value on most moral views. The very best experiences seem to be *far* better than most other experiences, and the very worst experiences seem to be far worse than most other experiences.

[48] That is, the ultimately available amount of physical resources, space, and time.

[49] Of course, at familiar scales, most goods and resources exhibit diminishing returns in their *instrumental* usefulness for some (more) ultimately valued end; here we are considering separability in time and space of value *per se*. That is, it often *will* be important to know about the rest of the world when considering the all-things-considered effect of some local intervention, because of interactions between the local region and the rest of the world. For example, the value to the world economy of producing some good depends on the existing global supply of and demand for that good. The claim here is that, ultimately, there is some procedure for counting up global value which is separable across local regions.

[50] One complication is that doubling available resources could fail to exactly double feasibly achievable value for practical, rather than essential axiological, reasons. In particular, civilisation overall could exhibit increasing returns to scale in terms of how much value it can ultimately support, because of the gains to specialisation and trade. Here we assume that, at large enough scales, the instrumental returns to scale are effectively constant, such that what matters is whether the same goods become more or less valuable for essentially moral reasons. We discuss some counterexamples to this assumption in footnote 43.

[51] Ord, ' The Moral Imperative toward Cost-Effectiveness in Global Health '.

Some fortunate people even report experiences so apparently valuable that they would be willing to trade them for mundanely positive experiences lasting thousands of times longer. In the prologue to his autobiography, philosopher Bertrand Russell wrote: “I have sought love… because it brings ecstasy — ecstasy so great that I would often have sacrificed all the rest of life for a few hours of this joy.”[52] Similarly, the Russian novelist Fyodor Dostoevsky described his experiences with epilepsy in this way: “For several instances I experience a happiness that is impossible in an ordinary state, and of which other people have no conception. I feel full harmony in myself and in the whole world, and the feeling is so strong and sweet that for a few seconds of such bliss one could give up ten years of life, perhaps all of life.”[53]

Probably these quotes are hyperbole — we doubt Russell and Dostoevsky would actually have chosen to die a decade earlier for another taste of those experiences. Still, it reflects something about the extremity of the experiences described. And Russell and Dostoevsky share remarkably similar brains — who knows what kinds of subjective experience other, more complex, more interconnected, more powerful minds could support, and how high the ceiling of best-possible experience could be. And similarly vast ceilings could exist for goods other than subjective experience: the intensity and quality of friendships, of romantic love, mutual connection, artistic achievement, and so on.

There have also been some preliminary surveys aimed at assessing the shape of the distribution of value of experiences. In one such survey, respondents were asked to compare their most intense and second most intense experiences. When asked how many times more intense their most intense experience was compared to their second most intense, over 50% reported a ratio of 2× or higher, and many reported substantially higher ratios again.[54]

Looking to the future, a fat-tailed distribution of value-per-unit-resources seems likely to us, too. Consider that, because linear views are separable across space and time at some level of granularity, there must be some single[55] “value-efficient” arrangement of resources such that, in order to achieve a best-possible outcome, one needs to recreate as many of those arrangements as possible.[56] But the space of things that future beings could do with resources is *astronomical*. Perhaps this arrangement is the size of an epoch-spanning planet, or a brief and simple experience of bliss running on a computer the size of a sugar cube. If a small fraction of those resources achieve more than 50% of the value-efficiency of the most value-efficient resources, which seems very plausible to us,[57] then a mostly-great future is a narrow target.

What’s more, if one is merely *uncertain* about whether the distribution of future value-efficiency is fat-tailed, then one’s *expected* distribution is fat-tailed.[58] So this aspect of the argument seems

[52] Bertrand Russell, Autobiography, Routledge Classics (London: Routledge, 2010). Quoted in MacAskill (2022).

[53] Avrahm Yarmolinsky, Dostoevsky: His Life and Art, [2nd ed., completely rev. and enl.] (London: Arco, 1957), 158. Quoted in MacAskill (2022).

[54] Gómez-Emilsson and Percy (2023).

[55] Or perhaps, multiple arrangements with *exactly* the same value/cost of resources — but this seems unlikely.

[56] See Shulman, ‘ Are pain and pleasure equally energy-efficient? ’, which introduces the term “hedonium”, the most value-efficient arrangement of resources for aggregative hedonic utilitarians.

[57] Indeed, it seems to us that the vast majority of possible uses of resources are of essentially zero value. And if the value of this arrangement is the product of some sort of performance along many dimensions, then its distribution will be fat-tailed.

[58] For a notable example of this dynamic, see Weitzman, ‘On Modeling and Interpreting the Economics of Catastrophic Climate Change’ .

fairly robust to us. The conclusion is that the vast majority of ways of using resources are not nearly as good as the very best uses of resources.

These considerations about the value-efficiency of resources apply to both bounded and unbounded views: bounded views can just as well feature an exactingly fat-tailed distribution of value-efficiency across uses of resources. But the difference is that, if a bounded view is fussy about the best uses of some given resources, then inefficient use of those resources can still be compensated by *scale* — inefficiently using many more resources to approach the upper bound — or by getting things right rarely, but *enough*.

But unbounded linear views do not have this flexibility. Linear views which are exacting about the most value-efficient use of resources will require the future to go in a very specific way for that future to count as mostly great: unless *most* available resources are configured for the *almost exactly the most valuable kind(s) of thing* according to that moral view, then most achievable value is almost certainly lost.[59] So linear views would seem to be fussy.

3.3. Bounded views

Linear views can seem fanatical. Consider a choice, between (i) a 0.001% shot at a near-best future, and extinction otherwise; and (ii) a guarantee of the *common-sense utopia* described in section 2.2. Because common-sense utopia is limited to our solar system, linear-in-resource views would evaluate the first option as vastly more valuable.[60]

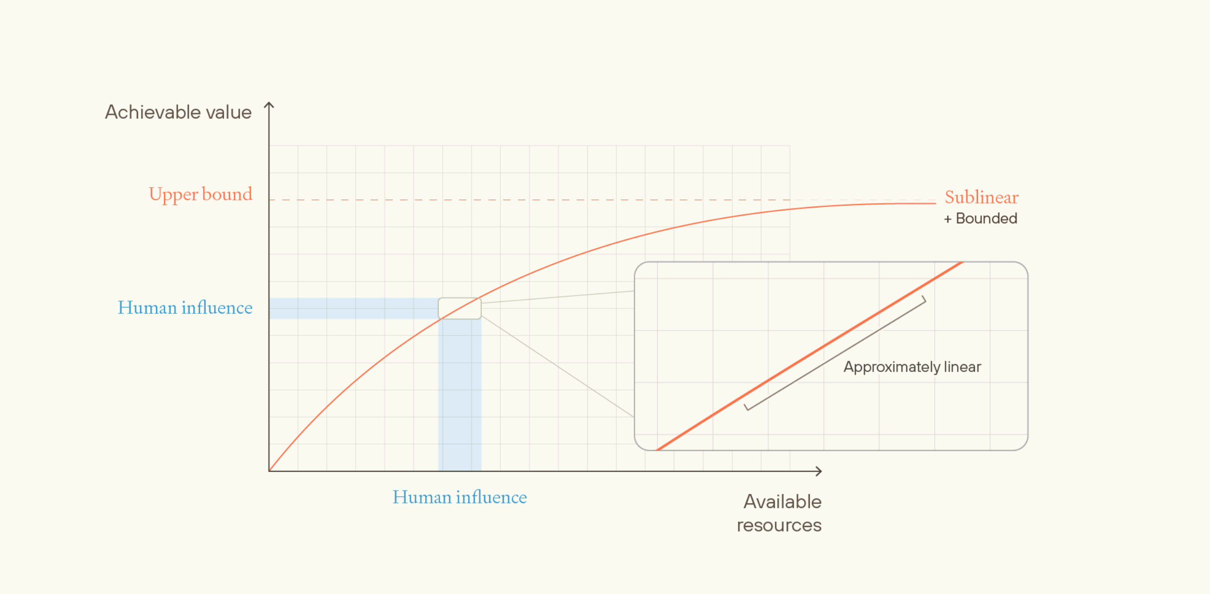
[59] Here's a counterargument, and a response. We've implicitly assumed that a decision-maker guided by a totalising unbounded linear view (like classical utilitarianism) would aim to maximise the finite intrinsic value of the resources in our future light cone. But that might not be true. And if it's not, then even seemingly-linear views like classical utilitarianism could regard resources as having diminishing marginal value in practice. Consider three alternative goals that classical utilitarians could aim for: (1) Maximize the probability of producing an infinite amount of value. This probability would be steeply diminishing in the amount of resources used, since civilisation under the guidance of such a view would try the most likely avenues first, and then progressively less likely ones. (2) Trade acausally with other civilizations across the universe or multiverse (who might control much larger amounts of resources than we can, or even control infinite resources). In this case, the most-valuable trading partners would be prioritized first, creating diminishing returns. There would also be diminishing returns to figuring out how to engage in acausal trade. (3) Trade with beings who are simulating us (who, again, might control much larger amounts of resources than we can, or even control infinite resources). In this case, we would want to trade with the trading partners that are most-likely to be simulating us, first; again resulting in diminishing returns. These ideas are all worth investigating further, but are outside of the scope of this essay. One thing to note is that such views still seem very fussy: the future would still need to be optimized for something very specific and unusual, such as creating infinite value or engaging in acausal trade. It's also not obvious that these considerations do result in meaningfully diminishing returns to resources: if, for example, it is possible to use resources to produce infinite amounts of value, then, at least on some ways of representing infinite value, the classical utilitarian will still want to produce as much (infinite) value as possible — which will look just like the fussy linear views we've discussed.

[60] It doesn't follow that linear-in-resource views recommend *taking* fanatical gambles in pursuit of near-best futures, because they are just views about axiology: a linear-in-resource view could be paired with a prohibition against taking fanatical gambles, even while evaluating them to be more valuable than the alternative. Still, there is something intuitively wrong about evaluating small chances of grand futures as vastly valuable in the first place, and this is the problematic intuition we're appealing to.

[61] See also: Christiano, ' Better impossibility result for unbounded utilities '; Beckstead and Thomas, 'A paradox for tiny probabilities and enormous values' .

This, among other considerations, motivates the idea that we should treat value as *bounded* .[61] There's a lot to say about these views. But the central point is that most bounded views are fussy, too; it's only a very specific and narrow type of bounded view that's easygoing.

There are a few ways a moral view could be upper-bounded, and it matters just how the view is put together. The most natural view evaluates the entire universe; the upper bound is on the value of the universe as a whole. However, the universe is big. On leading cosmological theories, the observable universe is tiny compared to the size of the total universe, which could even be infinite.[62] If so, then it's very likely that there are a vast number of alien civilisations, including beings with moral status, elsewhere in the universe, including outside of the observable universe. But, if so, then the *difference* that Earth-originating civilisation makes to the value of the universe is tiny. But strictly concave functions are approximately linear for small changes. So, even though maximum achievable value is a strictly concave function of resources under civilisation's control, for all practical purposes we can treat value as linear in resources. But, as we saw in the last section, linear views are fussy.



At small enough scales, strictly concave (and differentiable) functions are approximately linear. Since the difference humans could possibly make to overall value is likely proportionally tiny, sublinear views which consider 'universal' value are likely practically near-indistinguishable from linear views.

So if we're searching for easygoing bounded views, we should look elsewhere — to views on which look at the *difference* in value humanity can make to the universe (or something like it), and regard that difference as bounded. If humanity's domain is a bubble rising from the ocean floor, these views care about what we do inside our bubble, irrespective of the number of bubbles in the ocean as a whole.

There are major problems facing such difference-making bounded views, including violating

[62] See Gott, Jurić, Schlegel, Hoyle, Vogeley, Tegmark, Bahcall, and Brinkmann, 'A Map of the Universe '; Tegmark, 'Parallel Universes '.

stochastic dominance with respect to goodness.[63] And note that, in order to avoid being in-practice linear, the upper bound on value needs to be sufficiently low that it is possible to get somewhat close to the upper bound with the resources in the accessible universe. But, even putting these issues to the side, the most plausible forms of bounded difference-making views are fussy.

To see this, consider two different ways in which the view could aggregate goods (like flourishing lives) and bads (like suffering lives): separately or jointly. A view that's bounded (above and below, or only above), which aggregates goods and bads *separately* , could be described like this:

- First, add up all the bad things in the world (weighted by how bad they are), and transform that quantity according to some bounded function.
- Next, add up all the good things in that world (weighted by how good they are), and transform that quantity according to some bounded function.
- Then add the two resulting measures, to obtain a measure of overall value.

In contrast, a bounded view[64] which aggregates goods and bads *jointly* would work differently:

- First, add up all the goods and subtract all the bads in the world (weighted by how good or bad they are).
- Next, transform this quantity according to some bounded function, to obtain an overall measure of value.

These two methods can disagree significantly. Think of a function v_b over g "units" of goods, which is concave and upper-bounded for positive values, and convex (but not necessarily bounded) for negative values, and $v_b(0) = 0$. Imagine an outcome can contain compensating units b of bads, so that — for some linear view v_l — $v_l(g - b) = 0$ exactly when $g - b = 0$ for any g and b . We can then define a separate aggregation view as giving the overall value by $v_b(g) + v_b(-b)$, and a joint aggregation view as giving the overall value by $v_b(g - b)$. Now, suppose humanity can either (a) produce 10 units of goods and 1 unit of bad, or (b) produce 5 units of goods and no bads. Both theories agree on the value of (b), which is just $v_b(b)$. But, on the separate aggregation view, the overall value of (a) is given by $v_b(10) + v_b(-1)$, while on the joint aggregation view, the overall value of (a) is given by $v_b(10 - 1) = v_b(9)$. Unless v_b is implausibly insensitive to bads — that is, unless the difference between $v_b(10)$ and $v_b(9)$ is *smaller* than the difference between $v_b(0)$ and $v_b(-1)$ — then the joint aggregation view will value (a) more highly, disagreeing with the separate aggregation view.[65]

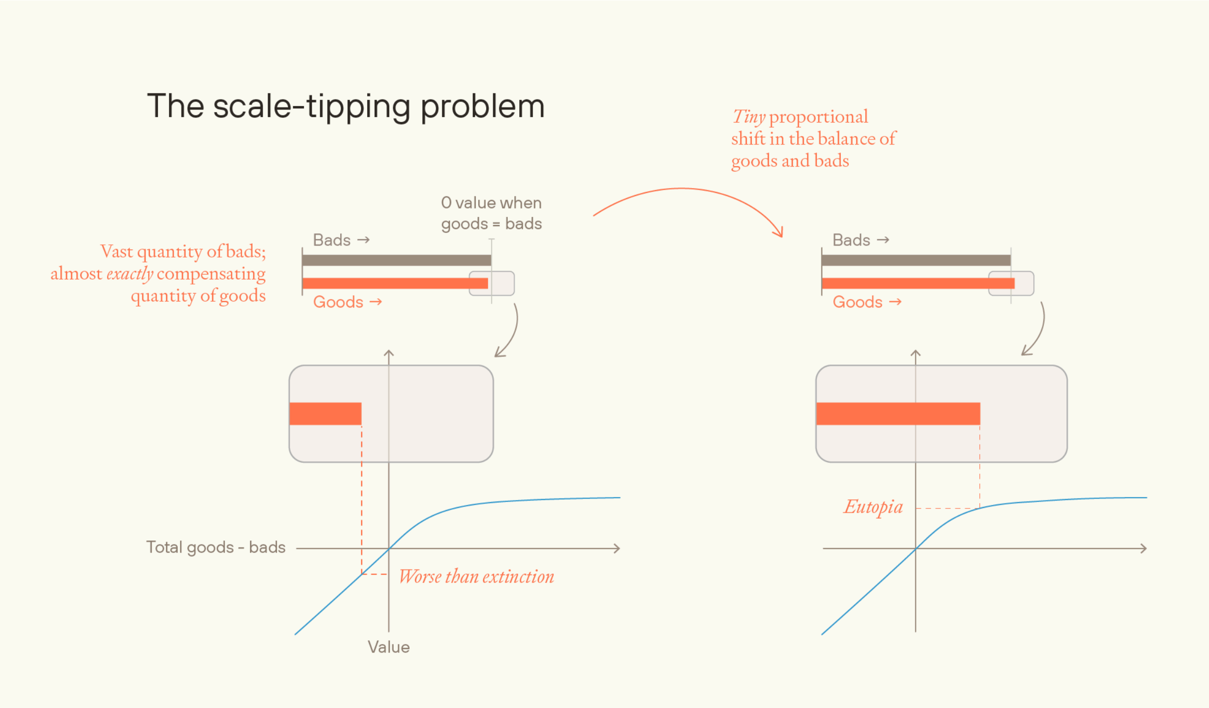
[63] That is, on a difference-making bounded view, A could be better than B, even though, for any possible value of the universe, the probability of the universe having that value is at least as great on B as it is on A, and for some possible value or values of the universe, the probability of the universe having that value is greater on B as it is on A. For discussion, see here, especially section 4.2: Greaves, Thomas, Mogensen, and MacAskill, ' On the desire to make a difference '. This article focuses on difference-making risk aversion more generally, but upper-boundedness is just one form of risk-aversion, and "Example 7" in the article works just as well against difference-making boundedness.

[64] Bounded above, or both above and below.

[65] To be slightly more precise and general: consider a function v_b bounded above and strictly concave in its positive part, defined in terms of an unbounded linear function v_u . Consider also positive quantities or "units" g and b of goods and bads, so that v_u can be represented such that $v_u(g - b) = v_u(g) + v_u(-b) = g - b$ for any g and b . Then, on the separate aggregation view, we can say $v_s(g, b) = v_b(v_u(g)) + v_b(v_u(-b)) = v_b(g) + v_b(-b)$. On the joint aggregation view, we can say $v_j(g, b) = v_b(v_u(g) + v_u(-b)) = v_b(g - b)$. If the function is odd, and strictly concave in its positive part, and $g - b > 0$, then $v_b(g - b) > v_b(g) + v_b(-b)$. In that case, the joint aggregation view values the same mixtures of goods and bads more highly than the separate aggregation view, relative to outcomes with only goods. This doesn't necessarily

Any meaningful lower bound seems implausible, whether the view is jointly or separately aggregating. Suppose the value of a world is close to the lower bound on a separately aggregating view, because it is full of bads but empty of goods. Then that view could recommend adding an arbitrarily vast amount of further bads in order to add a small quantity of goods; and that seems totally wrong. But suppose the value of a world is instead close to the lower bound on a jointly aggregating view. Then that view could recommend taking a 50–50 gamble between adding an arbitrarily vast amount of further bads, or otherwise enough goods to compensate for the (much smaller quantity of) existing bads.

In fact, joint aggregation views are somewhat implausible even without a meaningful lower bound, because of a "scale tipping" dynamic. Suppose that, at small scales, you think that N goods weigh equally against one bad, and that *Common-sense utopia* is a eutopia. If so, then a future involving a million galaxies of bads and N million galaxies of goods, plus one common-sense utopia around one star, counts as a mostly-great future. But that seems strange — intuitively, a tiny change to the relative balance of goods and bads shouldn't ever move the value of the future from as-good-as-extinction to eutopia.


The scale-tipping problem
Vast quantity of bads; almost *exactly* compensating quantity of goods
0 value when goods = bads
Bads →
Goods →
Tiny proportional shift in the balance of goods and bads
Bads →
Goods →
Total goods - bads
Worse than extinction
Value
Eutopia

On bounded views which jointly aggregate goods and bads, a tiny change in the balance of goods and bads can change the value of the future from worse than extinction, to eutopian.

That leaves separate aggregation views without a meaningful lower bound as perhaps the least implausible bounded views. But separate aggregation views are fussy, because even tiny quantities

follow on relaxing the assumption that v_b is an odd function, but it remains true for functions which weigh bads more heavily, so that $v_b(g) \leq -v_b(-g)$. Moreover, in the limit of large amounts of both goods and bads, overall value on the joint aggregation view approaches the upper asymptote; while overall value on the separate aggregation view approaches the sum of the upper and lower (negative) asymptote if there is a lower asymptote, otherwise it gets worse without bound.

of bads can be enough to prevent a mostly-great future. To start, take a view that treats goods and bads symmetrically, and say that *Common-sense utopia* constitutes a mostly-great future: it is at least 50% of the way to the upper bound. And now consider a civilisation that's fully spread out across all 10^{22} stars of the affectable universe, but where one part in 10^{22} is bad rather than good.[66] If so, then in aggregate the bads amount to a star system's worth, and that future is therefore more than 50% of the way to the upper bound on the disvalue of bads. So, even if that future is very close to the upper bound on the value of goods, then it achieves less than 50% of achievable value overall.

That is, the view ends up being fussy in a different way than linear views were. Getting to a near-best future, when looking at *goods* only, is easy. But the view becomes obsessively concerned with eliminating bads, and even a tiny quantity of bads is sufficient to make a mostly-great future impossible.

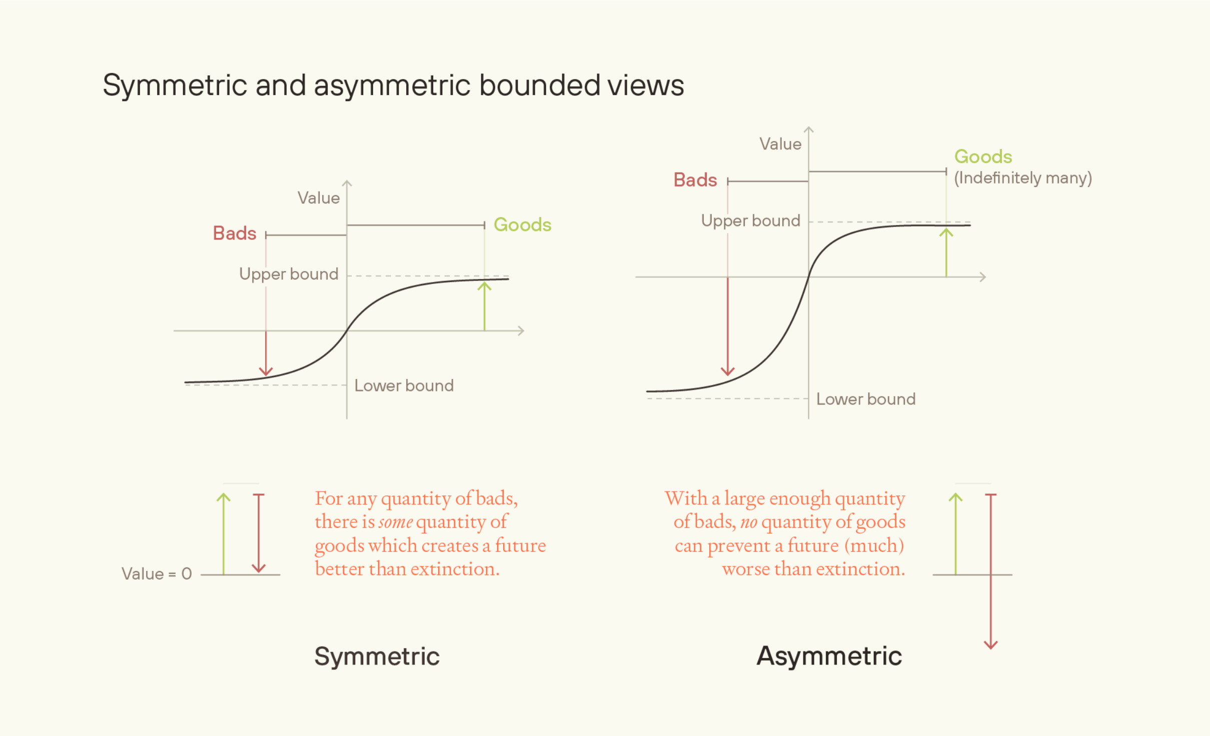
The situation gets worse if the view is not symmetric about goods and bads, but instead (as is intuitively plausible) thinks that the magnitude of disvalue from a worst-feasible world is much greater than the magnitude of value from a best-feasible world.[67] If so, then an *even smaller* quantity of bads would be sufficient to prevent a best-possible future. And, as mentioned, the disvalue of bads may not be bounded at all.[68] No matter how bad a dystopian future is to begin with, it seems that one could always make it twice as bad by making it much bigger in scale or involving even worse suffering. If this is right, and the bounded view is asymmetric with respect to goods and bads, then an even tinier quantity of bads would be sufficient to outweigh any utopian future, no matter how good.[69] The view would therefore be even fussier; the only mostly-great futures are those which had almost wholly eliminated the creation of bads.

[66] Given global wealth of $400 trillion, this would be equivalent to the world today spending less than one ten millionth of a cent on bad things.

[67] And an upper and lower-bounded view that's asymmetric in the other direction — where the ratio of upper to lower bound is high, so that worlds with increasing quantities of both goods and bads are positively valuable in the limit, irrespective of the ratio of goods to bads — is entirely implausible. Consider adding a combination of many bads and a tiny amount of goods to a world of initially zero value, so that the world initially becomes worse as more of this combination is added. But as the lower bound is met, then the value of the world begins increasing again.

[68] See Beckstead and Thomas, 'A paradox for tiny probabilities and enormous values' for some arguments on this matter.

[69] One could try to figure out a view where the upper bound on goods is high enough that it would take a very large quantity of bads to prevent a mostly-great future. The trouble is that, the higher the bound, the more the view will end up looking approximately linear, and be fussy for the same reasons that linear views are fussy. And such views would not respect our moral intuitions around common-sense utopia, making them seem particularly under-motivated.


Symmetric and asymmetric bounded views
Value
Bads
Goods
Upper bound
Lower bound
Value
Bads
Goods
(Indefinitely many)
Upper bound
Lower bound
Value = 0
For any quantity of bads, there is *some* quantity of goods which creates a future better than extinction.
With a large enough quantity of bads, *no* quantity of goods can prevent a future (much) worse than extinction.
Symmetric
Asymmetric

Comparing bounded views which aggregate goods and bads separately. "Symmetric" views have an upper and lower bound such that some quantity of goods can compensate for any amount of bads. On "asymmetric" views, by contrast, no quantity of goods can compensate for a large enough quantity of bads.

Alternatively, if the view aggregates goods and bads jointly, then we probably do end up with a view that's easygoing. Even a future that contains a significant fraction of bads could still be mostly-great. However, it's worth bearing in mind how specific and narrow such a view is — it faces the "scale-tipping" issue, and inherits the problems with difference-making boundedness. What's more, if value has no lower bound, or the lower bound is very low indeed, then such views generally may become pro-extinction,[70] since it's easy for a very small *chance* of a very bad future to outweigh a much greater chance of a near-best future. At the very least, it's a strange conclusion that the most plausible easygoing bounded view also looks unusually likely to favour extinction.

3.4. A recap

Let's take stock. Our goal has been to determine how difficult it is to achieve a mostly-great future. We have concluded that most plausible moral views are fussy: the target of a mostly-great future is narrow and difficult to hit.

First, we examined unbounded views whose maximum attainable value is approximately linear with respect to resources. These views are fussy because, for a future to be mostly-great, almost all resources in the accessible universe must be harnessed, and must be put towards some very specific use.

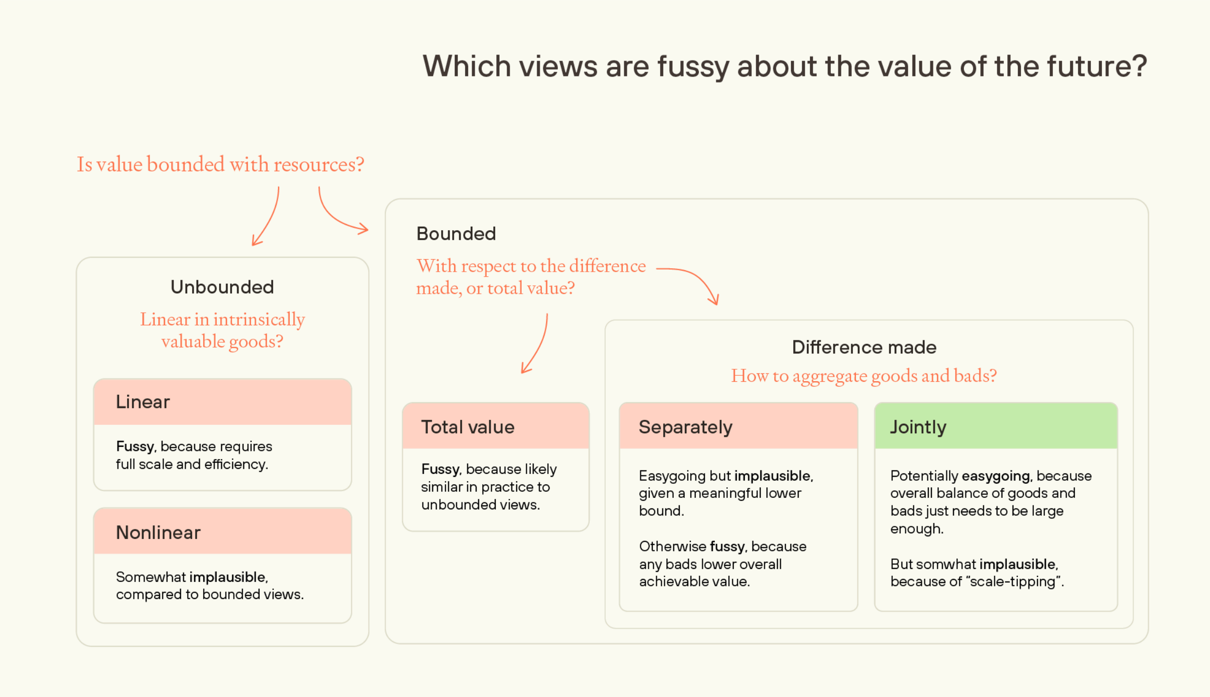
[70] They might not be pro-extinction if you think that alien life is likely to settle our part of the cosmos if we don't, and that an alien civilization would involve an even greater probability of a highly negative-value future.

Second, we considered views on which there is an upper bound on value. Many types of bounded views — those that are bounded with respect to the value of the universe as a whole, or are difference-making but with a sufficiently high bound — are approximately linear, too, and are therefore fussy.

Third, we considered difference-making views with a sufficiently low upper bound that they are *not* approximately linear. If these views aggregate goods and bads separately, then they are fussy for a different reason: even if the future contains a tiny quantity of bads, then we miss out on a near-best future. A civilization that is essentially entirely free of bads is again a very narrow target, so such views are also fussy, too.

Finally, there are low-bounded difference-making views that aggregate goods and bads jointly. These seem easygoing. However, they represent only a narrow slice of possible views, and have major issues: because they are difference-making, they violate stochastic dominance, among other problems; they don't capture how we'd intuitively want to aggregate goods and bads; and they may be strongly in favour of human extinction. Putting this all together, we think that easygoingness is unlikely.

Finally, here's a simplified summary of the views we've considered:


Which views are fussy about the value of the future?
Is value bounded with resources?
Unbounded
Linear in intrinsically valuable goods?
Linear
Nonlinear
Bounded
With respect to the difference made, or total value?
Total value
Difference made
How to aggregate goods and bads?
Separately
Jointly

Fussy, because requires full scale and efficiency.

Somewhat **implausible**, compared to bounded views.

Fussy, because likely similar in practice to unbounded views.

Easygoing but **implausible**, given a meaningful lower bound.

Otherwise **fussy**, because any bads lower overall achievable value.

Potentially **easygoing**, because overall balance of goods and bads just needs to be large enough.

But somwhat **implausible**, because of "scale-tipping".

Summarising which value functions are likely to be fussy about achieving a mostly-great future.

3.5. Moral uncertainty

What kind of view should we adopt if we are uncertain about the correct view? Some views might seem "higher-stakes" then others, such that they should effectively loom largest in decision-making. But to know which views loom largest — and whether they are easygoing or fussy — we need to consider different ways to deal with uncertainty.

For illustration, let's consider three types of utilitarian views. First, a view on which value is unbounded both above and below. Second, a view on which value is unbounded below, but bounded above (and jointly aggregates goods and bads). Third, a view on which value is bounded above and below, which agrees with the asymmetrically bounded view on how high the bound is.[71] Suppose you split your credence evenly between the three views. Consider how each view would assess the ratios of the differences in value between the following four options:

Views → Options ↓	Unbounded (p = ⅓)	Bounded above and below (p = ⅓)	Bounded above (p = ⅓)
Best feasible future	1	1	1
Common-sense utopia	0.0001 (ε > 0)	0.8	0.8
Extinction	0	0	0
Worst feasible future	-1	-1	-10,000 (≪ -1,000)

In order to weigh options under uncertainty, we'll need some method for aggregating *across* different views, into a new combined value function.[72] Here's one way to do that: take the credence-weighted average value across views, with the value of best feasible future and extinction fixed at 1 and 0 accordingly.[73] In the example above, the average value of common-sense utopia would be $(0.8 + 0.8 + 0.0001) \times \frac{1}{3} \approx 0.533$. That is, on our combined value function, a prospect with a 54% chance of best feasible future and extinction otherwise is equally as good a guaranteed common-sense utopia.

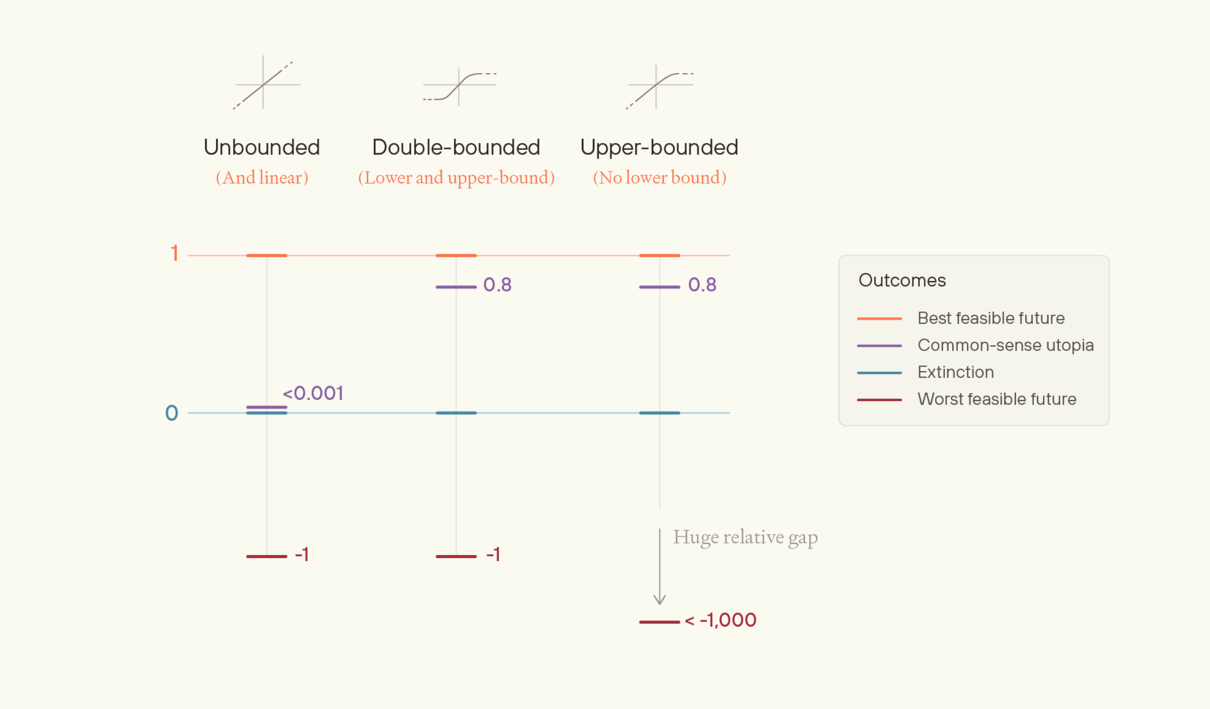
This is only the "obvious" method because, for each view, we chose to represent its value function by setting $v(\text{extinction})$ to 0 and $v(\text{best feasible future})$ to 1: implicitly assuming that all three views agree on the "stakes" involved in the difference in value between extinction and a best feasible future.

[71] That is, the upper-bounded view and the and upper and lower-bounded view both agree on exactly which outcomes are better than extinction, and on the value of every prospect containing only outcomes not worse than extinction, as a ratio of the difference in value between extinction and the best feasible future.

[72] In particular, we should try to find a general method which applies the same normalisation between theories across decisions, rather than a normalisation method which is sensitive to the actually available options in a decision situation, in part because such ad-hoc methods are vulnerable to cyclical recommendations across decision situations. See: Cotton-Barratt, MacAskill, and Ord, 'Statistical Normalization Methods in Interpersonal and Intertheoretic Comparisons'.

[73] That is: for each option, multiply the value on each view by your credence in that same view, and sum across all views.

But this was an entirely arbitrary choice. A value function is unique up to positive affine transformation: there are endlessly many ways to numerically represent a given value function, so there are endlessly many ways to "average" between different value functions. And, as we'll see, even if we fix the value function in the same way across theories by fixing the same two numbers to the same two outcomes, just *how* to fix the value function can matter dramatically for which options the aggregated value recommends taking.



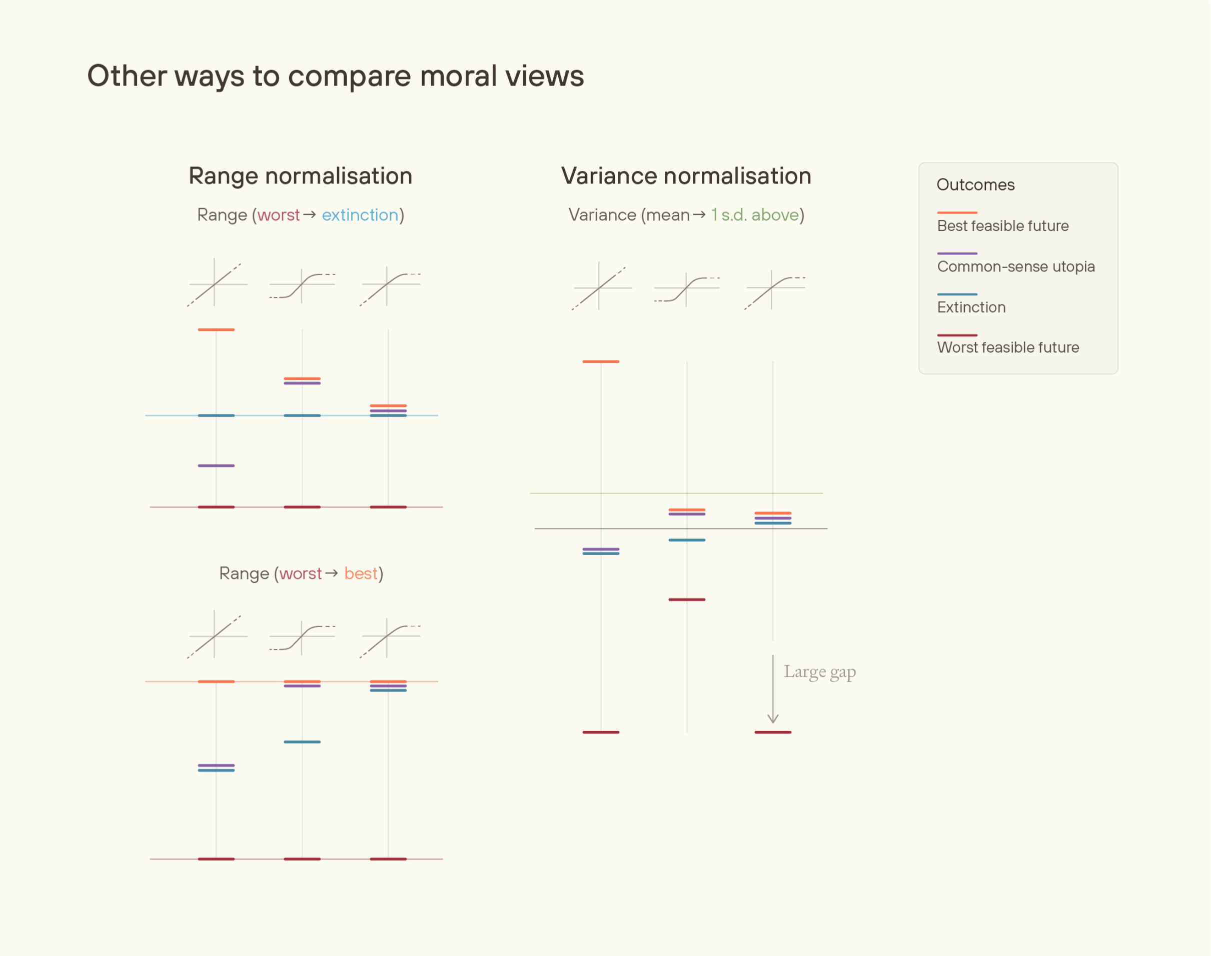
One way to act under uncertainty between moral views is to treat each view as 'agreeing' on the difference in value between extinction and the best feasible future.

There are many other ways to make intertheoretic value comparisons. Instead of assuming agreement on the difference in value between extinction and the worst feasible future, the aggregating function could assume all views agree on the size of difference in value between:

- The worst feasible future the best feasible future (a form of "range normalisation")[74]
- The worst feasible future and extinction
- The mean outcome and an outcome one standard deviation above the mean, according to some fixed prior distribution over outcomes (a form of "variance normalisation")[75]

[74] This is similar to Ted Lockhart's 'Principle of Equity among Moral Theories', but across all possible options rather than just the options within one choice set.

[75] Cotton-Barratt, MacAskill, and Ord, 'Statistical Normalization Methods in Interpersonal and Intertheoretic Comparisons'.



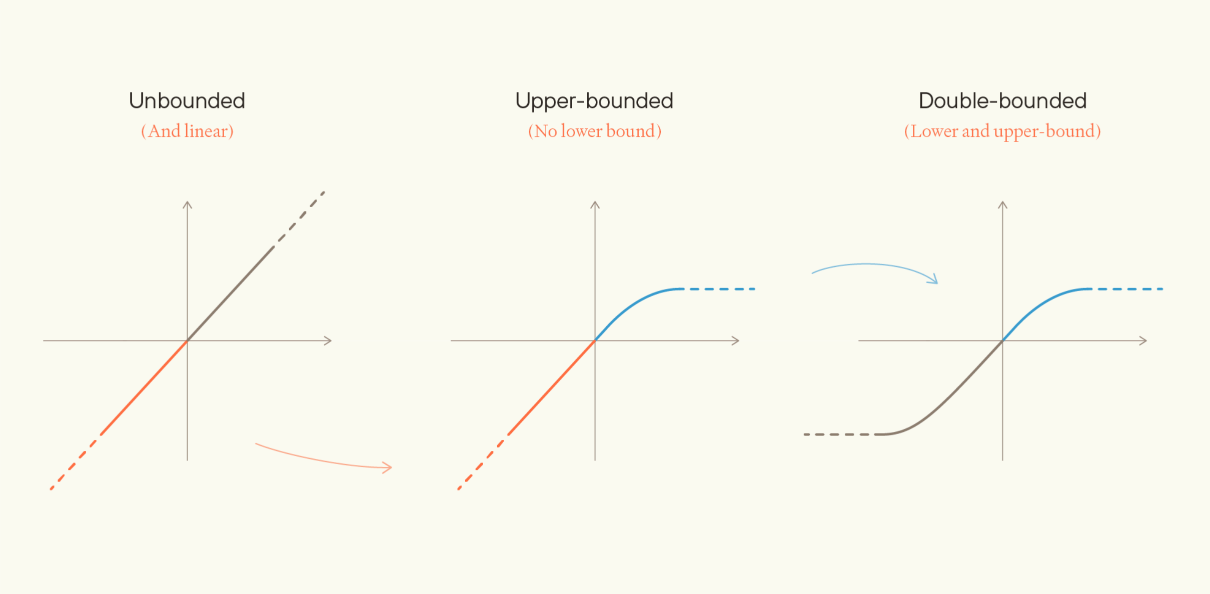
Three further methods for making quantitative comparisons between views.

Normalising by the difference in value between two outcomes ("extinction–best", "worst–extinction" and "worst–best") seems implausible, because these methods depend on the contingent fact about how big the affectable universe is, and therefore how good or bad the best and worst feasible futures are relative to futures of a known size. Around the 1920s, cosmology began to appreciate that the affectable universe is over a billion times larger than previously thought. Someone considering the same options of known effect size would then, on learning this cosmological news, give the unbounded view less than one billionth the weight that she did before. But that seems absurd, because surely unbounded moral views shouldn't be relegated into decision-irrelevance on discovering that the universe is bigger than imagined.

The variance normalisation approach looks more plausible. One of the main arguments for variance normalisation is that it makes things in general no higher stakes for any view than another. On variance normalisation, the difference in value between extinction and eutopia is larger than that difference on either of the bounded views, so variance normalisation affords unbounded views more "say" about upside-seeking options.

Alternatively, we could try to make judgments directly about how value compares across pairs of views. Here's an example: start by comparing the unbounded and the asymmetric (upper- but not lower-) bounded view. It seems like these two views agree on the disvalue of bads, since both views disvalue bads linearly, and they agree precisely in all circumstances that involve tradeoffs involving only bads. What's more, we could imagine variants of the bounded view, with progressively lower and lower bounds. In the limit, such a view would approximate strict negative utilitarianism. And, plausibly, utilitarianism and strict negative utilitarianism agree on the disvalue of bads. So we have a way to compare the unbounded and the asymmetric bounded view. Next, we compare the asymmetric bounded view and the symmetric bounded view. Here, it seems like the two views agree on the difference in value between zero and the upper bound: whatever reasons we have for setting the upper bound in one place rather than another should apply equally for both views. The two views agree in all circumstances involving choices between positive-value futures; the disagreement concerns how to weigh positive-value futures against negative-value futures.

But if we can compare the unbounded with the asymmetric bounded view, and if we can compare the asymmetric bounded view with the symmetric bounded view, then we can compare the unbounded with the symmetric bounded view, too. On this way of doing this, things are in general highest-stakes for the unbounded view, and lowest-stakes for the symmetric bounded view.



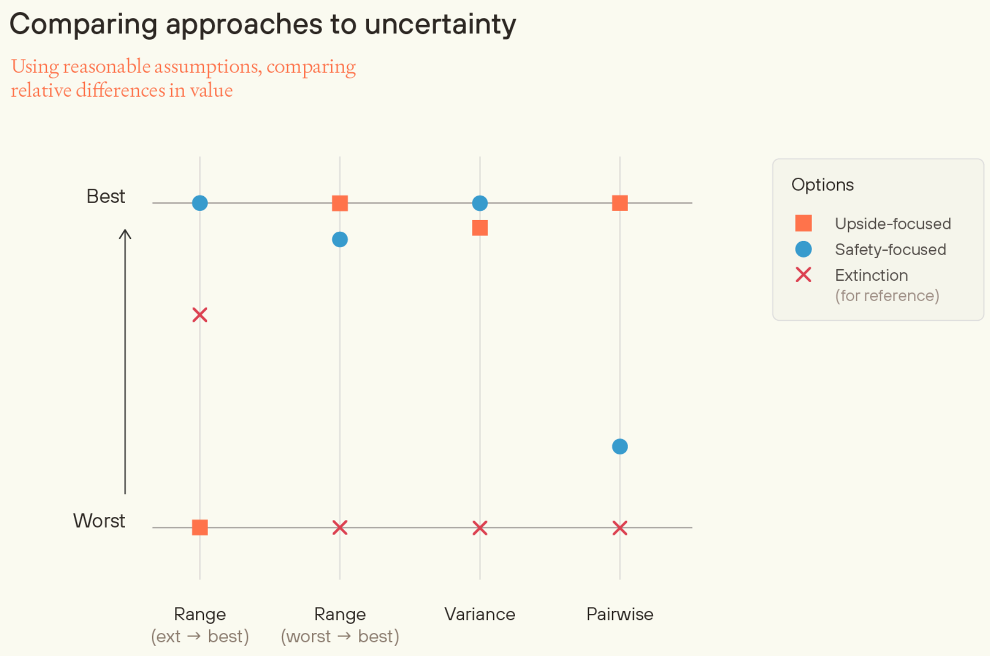
Setting a joint scale to compare across views, by reasoning about pairwise comparisons.

How to weigh views under moral uncertainty can change which options are best, often dramatically. Consider how these different methods evaluate a simple choice between two options:

	P(Eutopia)	P(Common-sense utopia)	P(Extinction)	P(Dystopia)
Safety-focused option	0.05	0.9	0.05	0

Upside-focused option	0.7	0	0.2	0.1

When we compare these two options, we find that the approach to uncertainty critically matters for which option is best.


Comparing approaches to uncertainty

Using reasonable assumptions, comparing relative differences in value

Comparing different approaches to uncertainty between moral views, in terms of the relative differences in value between three prospects (the upside-focused option, safety-focused option, and guaranteed extinction). Where necessary, we tried to use reasonable numerical assumptions, which you can see and change on [*this spreadsheet*] .

Normalising between extinction and the best feasible future strongly recommends the safety-focused option, with the upside-focused option being far worse than extinction. Normalising by variance makes both options look very comparably attractive, compared to extinction, although the safety-focused option edges out. But normalising by the pairwise approach makes the upside-focused option look far more attractive than the safety-focused option.[76]

Intertheoretic comparisons are very thorny, so we don't want to push any strong conclusions. But we can suggest two substantial upshots.

First, on both the most plausible statistical and the most plausible non-statistical approaches to intertheoretic comparisons, then the difference in value between zero and the best feasible future should be considered greater on the unbounded view than it is on either of the bounded views. So, if anything, at least when weighing between options which are all better than extinction, it's the

[76] More details can be found in this spreadsheet . Interested readers can make a copy, and experiment with their own numbers.

unbounded moral views that should effectively loom larger, and drive most of the difference in aggregate value between options. In this sense, the fairest ways to evaluate options, under moral uncertainty between bounded and unbounded views, are themselves fussy.[77]

Second, the choice of how to approach intertheoretic uncertainty matters significantly in this context. At least on a cursory treatment, plausible approaches to uncertainty disagree over which somewhat plausible-seeming options are best, even if all those approaches are fussy in practice. In particular, some approaches end up being sensitive to extremes of both upside and downside, while others end up being far more sensitive to extremes of downside only.

4. Conclusion

It's natural to think that a wide range of imaginable futures are almost as valuable as the very best futures — that we should be easygoing about what it takes to achieve a mostly-great future. After all, it seems bizarre to judge that a very slim chance of a eutopian future, and extinction otherwise, could be somehow better than a guarantee of a future that's extraordinarily good in absolute terms.

The "no easy eutopia" view says we should actually be very fussy about what counts as a mostly-great future, contra the intuitive appeal of "easy eutopia" views. Single moral errors which erase a major fraction of the goodness of the world seem easy to make. Plausibly, the value of the future is well-described as a product of many independent factors, such that doing poorly on any one dimension is sufficient to lose out on most feasibly achievable value. And when you try to reason through the space of value functions more systematically, it looks like only a narrow slice of plausible views are truly easygoing, and those views are not themselves very plausible.

But our discussion so far *doesn't* imply that we're unlikely to reach eutopia, or a future which lands in that ballpark of value. The target is narrow, but there could be forces which guide society towards hitting it. It could be that society's views converge, perhaps through some truth-seeking deliberative processes, to land on the correct views. It could be that, among a diversity of views about what ultimately matters, most views can achieve most of what they care about, especially through trade and compromise. Or the target could just exert some kind of gravitational pull, even if most people in society don't ultimately care about reaching it. The next essay turns to this question.

Bibliography

Stuart Armstrong and Anders Sandberg, 'Eternity in six hours: Intergalactic spreading of intelligent life and sharpening the Fermi paradox' , *Acta Astronautica* , August 2013.

Adam Bales, 'Against willing servitude: Autonomy in the ethics of advanced artificial intelligence' , *The Philosophical Quarterly* , 31 March 2025.

[77] One way to push back against the idea that there's more at stake for the unbounded view is to consider that, in a state of ignorance about what the world is actually like, unbounded views would want to have the most sway in worlds where there are plausible paths to creating more value than could possibly be created in this world (perhaps up to infinite quantities). On this method, unbounded views should have very little influence on what we do in this world. See also: Larks, 'Moral Trade, Impact Distributions and Large Worlds '.

Nick Beckstead and Teruji Thomas, 'A paradox for tiny probabilities and enormous values' , *Noûs* , June 2024.

Erik Carlson, 'Extensive measurement with incomparability' , *Journal of Mathematical Psychology* , 1 June 2008.

Paul Christiano, ' Better impossibility result for unbounded utilities ', 9 February 2022.

Fabio Clementi and Mauro Gallegati, 'Pareto's Law of Income Distribution: Evidence for Germany, the United Kingdom, and the United States ', *Microeconomics* , 18 May 2005.

Owen Cotton-Barratt, William MacAskill, and Toby Ord, 'Statistical Normalization Methods in Interpersonal and Intertheoretic Comparisons' , *The Journal of Philosophy* , 2020.

J. Richar Gott, Mario Jurić, David Schlegel, Fiona Hoyle, Michael Vogeley, Max Tegmark, Neta Bahcall, and Jon Brinkmann, 'A Map of the Universe' , *The Astrophysical Journal* , 10 May 2005.

Hilary Greaves, Teruji Thomas, Andreas Mogensen, and William MacAskill, ' On the desire to make a difference ', *Philosophical Studies* , 2024.

Larks, ' Moral Trade, Impact Distributions and Large Worlds ', 20 September 2024.

Toby Ord, ' The Moral Imperative toward Cost-Effectiveness in Global Health ', *Center for Global Development* ., 8 March 2013.

Peter Salib and Simon Goldstein, ' AI Rights for Human Safety ', 1 August 2024.

Carl Shulman, ' Are pain and pleasure equally energy-efficient? ', *Reflective Disequilibrium* , 24 March 2012.

Max Tegmark, ' Parallel Universes ', 7 February 2003.

Martin L Weitzman, 'On Modeling and Interpreting the Economics of Catastrophic Climate Change' , *The Review of Economics and Statistics* , 1 February 2009.

Evan G. Williams, 'The Possibility of an Ongoing Moral Catastrophe' , *Ethical Theory and Moral Practice* , 1 November 2015.

Forethought

Convergence and Compromise

Fin Moorhouse & William MacAskill

August 2025

1. Introduction

The previous essay argued for "no easy eutopia": that only a narrow range of likely futures capture most achievable value, without serious, coordinated efforts to promote the overall best outcomes. A naive inference from no easy eutopia would be that mostly great futures are therefore very unlikely, and the expected value of the future is barely above 0.

That inference would be mistaken. Very few ways of shaping metal amount to a heavier-than-air flying machine, but powered flight is ubiquitous, because human design *honed in* on the design target. Similarly, among all the possible genome sequences of a certain size, a tiny fraction codes for organisms with functional wings. But flight evolved in animals, more than once, because of natural selection. Likewise, people in the future might hone in on a mostly-great future, even if that's a narrow target.

In the last essay, we considered an analogy where trying to reach a mostly-great future is like an expedition to sail to an uninhabited island. We noted the expedition is more likely to reach the island to the extent that:

1. The island is bigger, more visible, and closer to the point of departure;
2. The ship's navigation systems work well, and are aimed toward the island;
3. The ship's crew can send out smaller reconnaissance boats, and not everyone onboard the ship needs to reach the island for the expedition to succeed.

The previous essay considered (1), and argued that the island is small and far away. This essay will consider ideas (2) and (3): whether future humanity will deliberately and successfully hone in on a mostly-great future. Mapping onto scenarios (2) and (3), we consider two ways in which that might happen:

- First, if there is *widespread* and sufficiently *accurate* ethical convergence, where those people who converge on the right moral view are also *motivated* to promote what's good overall. We discuss this in section 2.
- Second, if there's *partial* ethical convergence, and/or partial motivation to promote what's good overall, and some kind of trade or compromise. We think this is the most likely way in which we reach a mostly-great future if no easy eutopia is true, but only under the right conditions. We discuss this in section 3.

In section 4, we consider the possibility that even if *no one* converges onto a sufficiently accurate ethical view, and/or if no one is motivated to promote what's good overall, we'll still reach a mostly-great future. In the "sailing" analogy, this would be like thinking that none of the ship's crew ultimately cares about reaching the island. We argue this is unlikely, if no easy eutopia is true. In section 5, we consider which scenarios are "higher-stakes", and should thus loom larger in decision-making under uncertainty. In section 6, we conclude.

The considerations canvassed in this essay have led to significant updates in our views. For example, in What We Owe The Future, Will said he thought that the expected value of the future, given survival, was less than 1% of what it might be.[78] After being exposed to some of the

[78] MacAskill, 'What We Owe the Future: A Million-Year View'.

arguments in this essay, he revised his views closer to 10%; after analysing them in more depth, that percentage dropped a little bit, to 5%-10%. We think that these considerations provide good arguments against extreme views where *Flourishing* is close to 0, but we still think that *Flourishing* has notably greater scale than *Surviving*.

2. Will most people aim at the good?

In this section, we'll first consider whether, under reasonably good conditions,[79] *most* people with power would converge on an accurate understanding of what makes the future good, and would be significantly motivated to pursue the good.[80] We'll call this widespread, accurate, and motivational convergence, or "WAM-convergence". Second, we'll consider how likely it is that we will in fact reach those "reasonably good" conditions. (We'll here assume that, given WAM-convergence, we will reach a mostly-great future.)

If this idea is right, then even a global dictatorship could have a good chance of bringing about a mostly-great future: under the right conditions, the dictator would figure out what makes the future truly valuable, become motivated to bring about the most valuable futures, and mostly succeed. From feedback on earlier drafts, we've found that a surprising number of readers had this view.

In this essay we talk about the idea of promoting the good "de dicto". This is philosophy jargon which roughly means "of what is said". It's contrasted with "de re", roughly meaning "of the thing itself". To illustrate, there are two ways in which Alice might want to do what's best. First, she might want to do some particular act, and happen to believe that act is what's best, but she wouldn't change her behaviour if she learned that something *else* was best. She may be motivated to *help the poor*, or *fight racism*, or *support a friend*, but not *moreover* to do what's best, or to contemplate what that may be. If so, then Alice is motivated to do good *de re*. Second, Alice might want to do some particular act and believe that act is what's best, but if she learned that something else was best she *would* change her behaviour. She's motivated, at least in part, by *doing what's best*, whatever that consists in.[81] This attitude could also motivate Alice to try to *figure out* what's best, with a view to potentially changing the particular things she acts on and cares about.[82] When we talk about WAM-convergence, we're talking about convergence to motivation to promote the good *de dicto*.

[79] By "reasonably good conditions", we mean that none of the "blockers" discussed in 2.5 occur.

[80] This need not be their *initial* motivation. On strong forms of moral realism, it could be that as people reflect, they come to form the desire to promote the good de dicto, even if they did not have that motivation originally.

[81] The idea of being motivated by promoting the good de dicto has a bad rap in academic philosophy, on the grounds that it's implausibly "fetishistic" (for a seminal discussion, see Williams, 'Persons, character and morality'.) But, as we'll discuss, reaching a mostly-great future could essentially require some people to be motivated by the good de dicto. The rough idea is that if a mostly-great future presents a narrow target, then it would be a coincidence if that target were hit, unless enough people were explicitly trying to hit it.

[82] Similarly, if Bob says "I want to speak with the manager" because the manager is a friend of his, and he wouldn't want to speak to the manager if it wasn't that friend, he is speaking of "the manager" de re. If Bob says "I want to speak with the manager" because he has a complaint to make to whoever the manager is, he is speaking of "the manager" de dicto. He might also then be motivated to proactively learn who the manager is, if he's unsure.

The remainder of section 2 discusses whether WAM-convergence is likely. In section 2.2, we'll consider and ultimately reject two arguments, based on current moral agreement and moral progress to date, for optimism about WAM-convergence. In section 2.3, we describe three important aspects of a post-AGI world and discuss their upshots. In section 2.4, we give our main argument against expecting WAM-convergence even under reasonably good conditions.

The term "reasonably good conditions" is vague. What we mean is that there are no *major blockers* to actually producing mostly-great futures, even if most people are motivated to do so. For example, one type of major blocker would be some early lock-in event that makes it impossible for later generations to produce a mostly-great future. We discuss major blockers in section 2.5.

2.2. Moral agreement and progress to date

In this section, we discuss two arguments for optimism about WAM-convergence, based on our current and historical situation. We explain why we're unconvinced by both.

2.2.1. Current agreement

In the world today, people with very different worldviews nonetheless widely agree on the value of bringing about goods like health, wealth, autonomy, and so on. What's more, much of the apparent moral disagreement we see is a result of *empirical* disagreement, rather than fundamental moral disagreement[83] — for example, two people might disagree over whether euthanasia should be legalised, but only because they disagree about the likely societal consequences. Because of this, it might seem like, really, humans all broadly share the same values, such that we are already close to WAM-convergence.

However, we think these are good reasons for expecting only very limited kinds of convergence in the future. First, goods like health, wealth and autonomy are *instrumentally* valuable for a wide variety of goals. For example, both hedonism and preference-satisfaction theories of wellbeing normally agree that good physical health is typically good for a person, even though neither view regards good physical health as *intrinsically* good.

The problem is that this agreement will likely break down in the future, as we max out on instrumentally valuable goods and instead turn to providing intrinsically valuable good, and as technological advancement allows us to increasingly *optimise* towards very specific goals, including very specific types of wellbeing.

A life that's increasingly optimised for maximal hedonic experience will likely begin to look *very* different from a life that's increasingly optimised for preference-satisfaction. From each view's perspective, the value of the preferred life will begin to pull away from the other, and the other might even begin to get *worse*. For example, perhaps the life which is increasingly optimised for hedonism begins to resemble some kind of undifferentiated state of bliss, with fewer and fewer meaningful preferences at all. With limited optimisation power, both views mostly agreed on the same "low-hanging fruit" improvements, like better physical health. But with more optimisation power, the changes these views want to see in the world become increasingly different.

[83] For extended discussion, see: Boyd, 'How to Be a Moral Realist', Railton, 'Facts, Values, and Norms: Essays toward a Morality of Consequence', and Brink, 'Moral Realism and the Foundations of Ethics'.

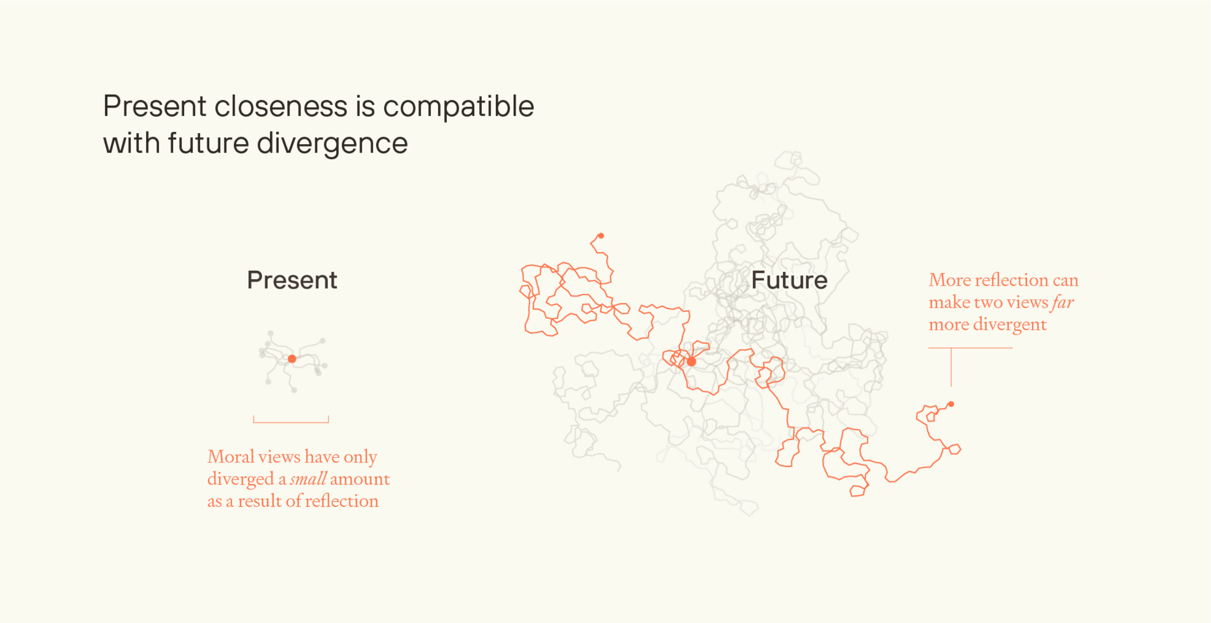
A second reason why current moral agreement doesn't provide strong support for WAM-convergence is that people today have strong instrumental reasons to conform, in their moral behaviour and stated beliefs, with the rest of society.[84] If you express or act on highly unusual moral views, you run the risk of facing social opprobrium, and make cooperation or trade with others more difficult. Similarly, if one group expresses very different moral views from another, ideologically-motivated coercion or conflict can become more likely. What's more, pressures to *outwardly* conform often influence *actual* prevailing beliefs, at least eventually. To give one illustration: polygamy had been a religiously endorsed and defining feature of the Mormon Church up until the 19th century, but came under intense pressure from the US federal government to end the practice. In 1890, the Church's president at the time, Wilford Woodruff, conveniently experienced divine revelation that they should prohibit plural marriage, which they did. Now, most Mormons disapprove of polygamy. Other examples include: many countries abolishing slave trading and ownership as a result of pressure from Britain; the long history of religious conversion by conquest; the embrace of democratic values by Germany and Japan (and then most of the rest of the world) after WWII[85] , and many more.

A third issue is that people currently have done very little in the way of moral reflection. Two people must be close to each other if they started off in the same spot, and walked 10 metres in any direction. But that gives little reason to think that they'll remain very close to each other if they keep walking for 100 miles — even slight differences in their orientation would lead them to end up very far apart.[86]

[84] For some relevant discussion at the level of individual psychology, see Hanson and Simler, ' The Elephant in the Brain: Hidden Motives in Everyday Life '.

[85] See Fukuyama, ' The End of History and the Last Man ', especially Section IV.

[86] In particular, we could model the process of reflection as a series of independent Brownian motions in $\mathbb{R}^2$, all starting at the same point at the same time. Then the expected distance of a view from the starting point, and the expected distance between two given views, both increase with the square root of time. The latter expectation is larger by a factor of $\sqrt{2}$.


Present closeness is compatible with future divergence
Present
Future
Moral views have only diverged a *small* amount as a result of reflection
More reflection can make two views *far* more divergent

Visualising moral reflection as an at least partially random walk in a space of views. If all views begin very close together, then further reflection seems likely to make resulting views diverge more, on average.

With advanced technology, this issue will get even more extreme, because people will be able to change their nature quite dramatically. Some people might choose to remain basically human, others might choose to remain biological but enhance themselves into "post-humans", and others might choose to upload and then self-modify into one of a million different forms. They might also choose to rely on different types of superintelligent AI advisors, trained in different ways, and with different personalities. The scope for divergence between people who originally started off very similar to each other therefore becomes enormous.

Finally, even if there is a lot of agreement, that doesn't mean there will be *full* agreement. Imagine two views that disagree about what's ultimately valuable both make their own lists of improvements to the world, ordered by priority: "First we solve the low-hanging fruit of easily-curable diseases. Then we eliminate extreme poverty. Then..." Both views might agree on the items nearest the top of both their lists. But eventually — and this point could be very far from the world today — they could simply run out of points of agreement. This would be like reaching a fork in the future: one road leads to a near-best future on the first view, but seals off a near-best future on the other, and vice-versa. And if a mostly-great future is a narrow target, then getting lots of agreement isn't enough; the remaining disagreements would likely be enough to ensure that each other's eutopia involves losing most value from the other's point of view.

And it seems there isn't anything close to full agreement on moral matters, currently. If we look back to the list of potential future moral catastrophes from the last essay (" No Easy Eutopia "), there's clearly not a lot of active agreement about what to do about the relevant issues, often because many people just don't have views about them. But we'll need to converge on most or all those issues in order for people to agree on what a mostly-great future looks like.

2.2.2. Moral progress

A second reason you might think that it’s likely we’ll get WAM-convergence is based on moral progress to date. At least on many issues, like slavery, civil rights, democracy, attitudes to women and sexual ethics, the world’s prevailing moral views have improved over the last few hundred years. Different moral views also seem to be broadly converging in the direction of liberal democratic values. According to the World Values Survey,[87] as societies industrialize and then move into a post-industrial, knowledge-based phase, there's a general shift from “traditional” to “secular-rational” values and from “survival” to “self-expression” values. The world has even so converged on the importance of democracy that most non-democracies *pretend* to be democratic — a practice that would have looked bizarre even 200 years ago.

Perhaps, then, there is some underlying driver of this moral progress such that we should expect it to continue: for example, if moral progress occurred because people have gotten wiser and more educated over time, that might give us hope for further progress in the future, too. However, there are a number of reasons why this is at best only a weak argument for expecting WAM-convergence.[88]

First, even if world values seem to be converging, that doesn’t mean they’re converging towards the *right* views. We are the product of whatever processes led to modern Western values, and so it’s little surprise that we think modern Western values are better than historical values. Our personal values are very significantly *influenced* by prevailing modern values, and it’s trivial that values have historically trended toward modern values![89]

’

[90]

There are various possible explanations for the moral convergence we’ve seen. One is that there’s been one big wave of change towards liberal values, but that this is a relatively contingent phenomenon. History could easily have gone differently, and there would have been some other big wave instead: if China had had the industrial revolution, modern values might be less individualistic; if Germany had won WWII, or the USSR had won the Cold War, then modern values might be much more authoritarian. A second alternative explanation is that modern values

[87] World Values Survey, ‘ Inglehart–Welzel Cultural Map ’.

[88] We’ll note that it’s not completely clear that, overall, values are converging. Jackson and Medvedev, ‘Worldwide divergence of values’ finds that over the last 40 years there’s been a growing gap between Western countries and the rest of the world, in particular around "emancipative values," which include the justifiability of homosexuality, euthanasia, divorce, suicide, child obedience, prostitution, and abortion. Oceanic, European, North American, and South American countries have increasingly adopted emancipative values. But in Africa and Asia, the endorsement of these values has been stable or even declining over the last 40 years (figure 2B). However, we have not vetted this paper, and quantitatively measuring value-changes over time involves some obvious and hairy methodological issues, like around which values to choose and how much weight to place on different values.

[89] “The sun is setting one evening over the ocean. You and a group of friends are standing at various points along the wet sand. As the water laps at your feet, you silently watch the red globe drop nearer and nearer to the horizon. As you watch, somewhat mesmerized, you notice how the sun's reflection on the wave crests forms a straight line composed of thousands of momentary orange-red glints-a straight line pointing right at you! "How lucky that I am the one who happens to be lined up exactly with that line!" you think to yourself.” (Hofstadter, ‘ Heisenberg’s Uncertainty Principle and the Many Worlds Interpretation of Quantum Mechanics ’.)

[90] A related point: you are more likely to have a worldview that currently has a lot of adherents.

are the product of the technological and environmental situation we're in: that a post-industrial society favours liberal and egalitarian moral views, perhaps because societies with such moral views tend to do better economically. But, if so, we shouldn't think that those trends will continue into a post-AGI world, which is very unlike our current world and where, for example, human labour no longer has economic value. A third possible explanation for moral convergence is that globalisation has resulted in greater incentives for moral conformity, in order to reduce the risk of war and make international trade easier. On none of these alternative explanations does what we regard as past moral progress result in continued moral progress into the future.

It's also easy to overrate the true extent of moral progress on a given moral view. The moral shortcomings of the world today are not necessarily as salient as the successes, and for many issues, it's socially taboo to point out areas of moral regress — precisely because one is pointing to views that are often currently socially unacceptable or at least morally controversial. It's therefore easy to overlook the ways in which the world has regressed, morally, by one's lights.

To see this, consider how total utilitarianism might evaluate humanity today, and consider that social norms around family size have changed dramatically over the last two hundred years, where families are now much smaller than they used to be. From a total utilitarian perspective, this has arguably resulted in an enormous loss of wellbeing to date. Suppose, for example, that the US total fertility rate had never dropped below 4 (which in our actual history it did in the late 19th century). The current US population would be over a billion people, meaning more people to enjoy life in an affluent democracy, and also more people benefiting the rest of the world through (for example) the positive externalities of scientific research.

There are other examples one could also point to, from a utilitarian perspective. Highly safety-cautious attitudes to new technology have limited the development and uptake of socially-beneficial innovations, such as nuclear power and new drugs and medicines. The rise in consumerism means that people spend relatively more on goods that don't really improve their quality of life. The huge rise of incarceration inflicts enormous human suffering. Globalism has been on the decline in recent decades, impoverishing the world overall. Rates of meat consumption have vastly increased, and the conditions in which we treat animals have worsened: factory farming did not exist in 1900; by 2025 over 70 billion land animals were slaughtered in factory farms.

All these cases are disputable, even on the assumption of utilitarianism. But at the very least they complicate the picture of a steady march of progress, from a utilitarian point of view.[91] And a similar case could be made on other moral views.

A third reason for caution about future optimism based on past moral progress is that, even if there has been genuine moral progress, and even if that was driven by some mechanism that reliably delivers moral progress, there's no guarantee that that mechanism will be enough to deliver all the moral progress we have left to make. For example: society made great strides of progress away from wrongful discrimination, subjugation, and cruelty against groups of people because those groups advocated for themselves — through persuasion, protest, striking, and movement-building. But some wrongs don't affect groups that can advocate for themselves. Today, humans will slaughter some 200 million chickens. If chickens feel pain and fear, as it seems they do, this seems like a

[91] You might also note that what really matters, from a utilitarian point of view, is the impact of society's actions, today, on the long-term future. But, if so, it becomes very hard to assess whether we have seen moral progress or regress, in this regard, insofar as most of the most important attitudes concern emerging technologies, and we don't know what people in the past would have thought about them.

moral catastrophe.[92] But we can't rely on chickens to organise, strike, or speak up for themselves.

Many of the most crucial moral questions we'll have to answer to get to a near-best future will be questions where we cannot rely on groups to speak up for themselves. For example, it will become possible to *design* new kinds of beings, like digital minds. We could design them never to complain as they willingly engage in, and even genuinely enjoy, kinds of servitude which might nonetheless be bad.

Then there are questions around population ethics — questions about evaluating futures with different numbers of people. On many views in population ethics, it's good to create new lives, as long as those lives are sufficiently happy. But people who don't exist can't complain about their nonexistence.

And there might be far weirder issues still, which require far more deliberate and concerted efforts to get right — like issues around acausal trade , cooperation in large worlds , or dealing with the simulation hypothesis .

2.3. Post-AGI reflection

We're considering whether there will *ultimately* be WAM-convergence. But, given survival, the world will look very different to the present, especially after an intelligence explosion. In this section, we discuss three aspects of the post-AGI world that we think are important to bear in mind when thinking about the possibility of WAM-convergence. These aspects sometimes get used as arguments for optimism about convergence; we think they have some real force in this regard, but that force is limited.

2.3.1. Superintelligent reflection and advice

Access to superintelligence could radically increase the amount of reflection people can do. And this could be a cause for optimism — maybe moral disagreement persists largely because moral reasoning is just too hard for humans to do well and consistently. Here's Nick Bostrom in *Superintelligence* :

> One can speculate that the tardiness and wobbliness of humanity's progress on many of the "eternal problems" of philosophy are due to the unsuitability of the human cortex for philosophical work. On this view, our most celebrated philosophers are like dogs walking on their hind legs - just barely attaining the threshold level of performance required for engaging in the activity at all.

In the future, we could design superintelligent AI systems to be smarter, more principled, more open-minded, and less prone to human biases than the wisest moral thinkers today. Perhaps AI systems could finally settle those stubborn puzzles of ethics that seem so intractable: we could just ask the AI to figure out which futures are best, tell them to set a course for those futures, and then enjoy the ride to eutopia.

[92] Especially in rich countries, where the opportunity cost of not buying chicken is very small.

What's more, as mentioned previously, much apparent moral disagreement might really bottom out in *empirical* disagreements. Or they could arise from 'transparent' errors in reasoning from shared principles, in the sense that they become clear to the person making the error once pointed out. If enough current disagreement ultimately arises from empirical disagreements or transparent reasoning errors, then you could maintain that as long as the future trends towards fewer empirical and conceptual mistakes, it wouldn't be surprising if most people end up converging toward the same place. And since AI systems seem well-suited to help clarify and resolve clear empirical and logical disagreements, then we should expect future people to have much more accurate empirical beliefs, and far fewer transparent reasoning errors.

It does seem very likely that advanced AI will give society as a whole dramatically more ability to reflect than we have had to date. Sheer cognitive abundance means that each person will be able to run billions of personal reflection processes, each of which simulates scenarios where they were able to reflect for millions of years, from different starting points and with different ways of reflecting. And it seems very likely that AI will ultimately be able to do ethical reasoning better than any human can, and thereby introduce arguments and considerations that human beings might never have thought of.

But that doesn't mean that people will all or mostly converge to the same place. First, even if different people do all want to act on their post-reflection views of the good (de dicto), the fact that future people will have superintelligent assistance says little about whether we should expect moral convergence upon reflection, from different people who have different ethical starting points and endorse different reflective processes. Different people's reflective processes might well lead them in very different directions even if they make no empirical or conceptual mistakes. We'll discuss this more in section 2.4.

Second, people might just be uninterested in acting on the results of an open-ended process of reflecting on the good (de dicto). They might be interested in acting in accordance only with the outcome of some *constrained* reflective process, that assumes and can't diverge from some preferred starting beliefs, like environmentalism, some specific religious views, or an overwhelming concern for the interests of biological humans (or a group of humans). As an analogy: when today's politicians select advisors, they choose advisors who broadly agree with their worldview, and task those advisors with advice on how to execute on that worldview, rather than advice about how their values could be fundamentally off-base.

Alternatively, people might just not be interested in acting in accordance with the result of *any* kind of ethical reflective process. Suppose that it's your partner's birthday and that you want to buy them a gift. You might strongly suspect that after countless years of earnest ethical reflection, you'd reach a different (and motivating) view about how to spend that money, perhaps because you'd become unable to rationally justify giving your partner's interests any special importance over anyone else's.[93] But would you go along with what your reflective-self recommends?

Third, of course, people might just remain self-interested. If two people want what's best *for them* , they might be motivated by the result of reflecting on how to get more of what they want, but it needn't matter if their AI advisors converge on the same conception of the good, because the good (de dicto) was never motivational for them.

[93] Not because you'd have learned anything relevant and new about what would happen as a result (it's not that the present would actually make your partner worse-off, or anything like that).

We suspect the most common attitude among people today would either be to reject the idea of reflection on the good (de dicto) as confusing or senseless, to imagine one's present views as unlikely to be moved by reflection, or to see one's idealised reflective self as an undesirably alien creature. If people strongly *dislike* the suggestions of their more-reflective selves, then they probably won't follow their suggestions. (This idea is discussed more in an essay by Joe Carlsmith).[94]

The situation becomes a bit less clear if your reflective self has time to walk you through the arguments for why you should change your views. But, even then, I think people will often just not be moved by those arguments — the fact that some argument would change someone's view if they were given millions of years to reflect (with much greater intellectual capacity, and so on) doesn't make it likely that that argument would change their view in a short period of time, given their actual cognitive capacity, unless that person already had a strong preference to defer to their reflective selves. Imagine, for example, trying to explain the importance of online privacy to a Sumerian priest in 2,500BCE.

Currently, people generally don't act as if they are profoundly ignorant about moral facts, or especially interested in discovering and pursuing the good de dicto. People are better described as having a bunch of things they like, including their own power, and they directly want more of those things. We don't see a reason, on the basis of superintelligent advice alone, for expecting this to change.

2.3.2. Abundance and diminishing returns

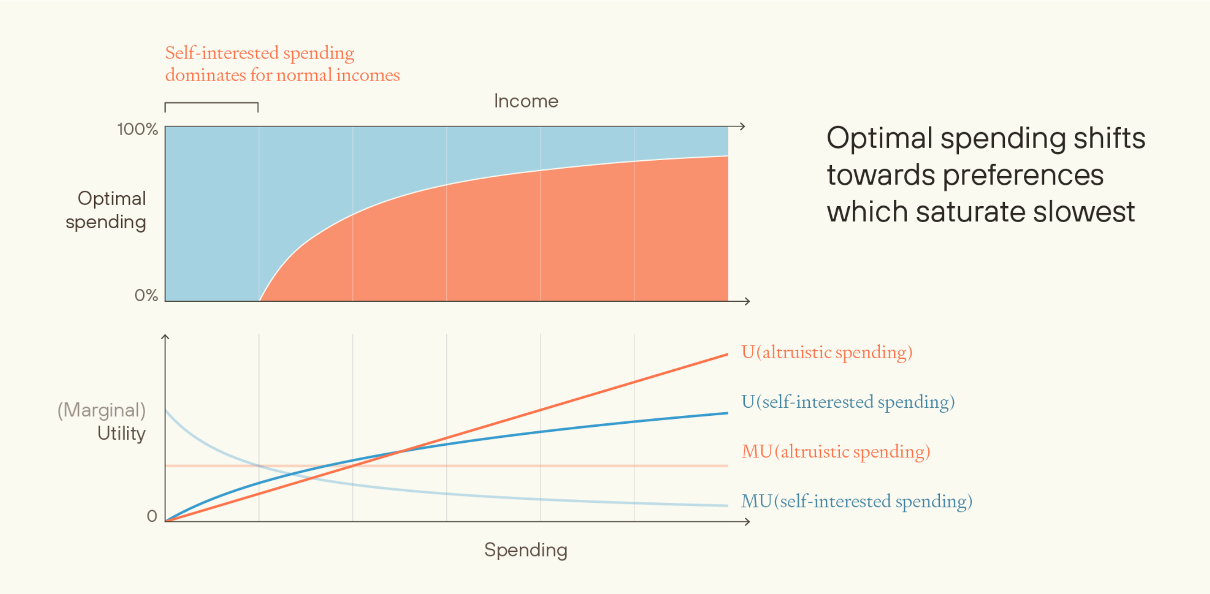
A post-AGI world would very likely also be a world of economic and material *abundance* . Due to technological advancement and the automation of the economy, the people who make decisions over how society goes will be millions or even billions of times richer than we are today.

This suggests an argument for optimism about altruistic behavior. Most people have both self-interested and altruistic preferences. Self-interested preferences typically exhibit rapidly diminishing marginal utility, while altruistic preferences exhibit more slowly diminishing marginal utility (even up to linearity).[95] Initially, when resources are scarce, people primarily satisfy self-interested preferences. But once wealth increases enough, people will shift marginal resources toward altruistic ends, since the marginal utility from additional self-interested consumption becomes smaller than the marginal utility from altruistic spending.

Because of this, if individuals have even a weak preference to promote the good, with extremely large amounts of resources they will want to use almost all their resources to do so.

[94] See in particular the section on "Galaxy Joe". Carlsmith, ' On the limits of idealized values '.

[95] We can represent this with a *quasilinear* utility function, $U(x,y) = f(x) + y$, where x and y are bundles of self-interested and altruistic goods, respectively, and $f(x)$ is strictly concave. So $U(x,y)$ is linear with respect to altruistic goods, but not self-interested goods.


Optimal spending shifts towards preferences which saturate slowest

We think that this is a good argument for thinking that, if there are deliberate decisions about how the resources of the cosmos are to be used, then the bulk of those resources will be used to fulfill preferences with the lowest rate of diminishing marginal utility in resources (including linear or even strictly convex preferences). However, these preferences may not be altruistic at all, let alone the “right” kind of altruistic preferences.

There’s nothing incoherent about self-interested (or non-altruistic) preferences which are linear in resources. Some might prefer 2,000 years of bliss twice as much as 1,000 years of bliss, or prefer to have 2,000 identical blissful copies of themselves twice as much as having 1,000 copies. Others might have a desire to own more galaxies, just for their own sake, like collector’s items; or to see more and more shrines to their image, while knowing they don’t benefit anyone else. Or people might value positional goods: they might want to have, for example, more galactic-scale art installations *than their cosmic rival does*.

Currently, there’s very little correlation between wealth and proportional spending on charity and altruism; the correlation within rich countries might even be negative. Despite steeply diminishing returns to self-interested spending, even billionaires spend proportionally modest amounts on philanthropy. Total billionaire wealth is around $16 trillion: assuming a 5% real return gives $800B in income without drawing down their principal, but billionaire philanthropy amounts to around $50B/yr,[96] around 6% of that income.[97] That weakly suggests that even vastly increased wealth will not significantly increase the fraction of spending on altruistic ends,

[96] Total billionaire philanthropic spending is hard to estimate, but the 50 biggest US donors gave around $16 billion in 2024, and just a third of all the billionaires reside in the US, but some of the biggest US donors are not billionaires. UHNW individuals (defined as holding >$30 million in assets) gave an estimated $190 billion in 2022. Billionaires hold around a quarter of total UHNW wealth, so assuming equal proportional giving by assets, that suggests total billionaire giving was around $40–50 billion.

[97] Note that the proportion is almost certainly larger strictly as a share of outlays, not including savings and investments, since billionaires on average likely reinvest the majority of their capital income. The philanthropic share of spending is perhaps 20–50%, although it’s difficult to directly estimate billionaire consumption. However, it remains true that a small fraction of wealth *accruing* to billionaires is spent philanthropically per year — so although philanthropic spending can remain high as a fraction of all spending, this is partly compensated for by a transfer from lower wealth and income levels, assuming a lower saving rate (progressive taxation aside).

especially once more advanced technology provides greater scope to spend money in self-interested ways.

Alternatively, people could have *ideological* preferences whose returns diminish slowly or not at all, and that are not preferences for promoting the good (de dicto). As discussed, they could instead be operating on the assumption of some particular moral view, or according to unreflective preferences. Or, even if people are acting according to their reflective preferences of the good, those reflective preferences could have failed to converge, motivationally, onto the right values.

Overall, the argument that altruistic preferences are satisfied less quickly in resources does suggest to us that a larger share of resources will be spent altruistically in the future. But we don't see it as a strong argument for thinking that most resources will be spent altruistically, or for giving us reason for thinking that such preferences will converge to the same correct place.

2.3.3. Long views win

Another way in which altruistic values might become dominant in the future is through *asymmetric growth* : without persuading or coercing other groups, certain values could be associated with faster growth in population or wealth, to the point where most living people embody those values, or most of the world's resources are effectively controlled by those values.

Asymmetric growth might systematically favour altruistic values. For example, people face decisions about whether to consume now, or to save for the future (including for other people). As long as most people discount future consumption, then saving your wealth (by investing it productively) grows it proportionally. People with more impartial and altruistic views are likely to save more, because this is a way to eventually help more people, at the expense of immediate self-interest. Altruists might also choose to have more children for similar reasons. So, in the long run, free choices around saving and fertility might select for altruistic values.[98] Perhaps people with these altruistic views are generally more willing to try to figure out and try to pursue the good de dicto. Then that would be a reason to be hopeful about widespread, accurate, and motivational convergence.

But, as with the last two sections, this isn't a reason for thinking that motivation to promote the good de dicto will become predominant. Even if non-discounting values win out over time, those non-discounting values could be for many different things: they could be self-interested in a way that doesn't discount one's interests in time, or they could be based on a misguided ideology.

What's more, non-discounting values might not have time to win out. If the major decisions are made soon, and then persist, then there just won't be time for this selection effect to win out. (It could be that there are much faster ways for non-discounting views to control more of the future than discounting views, through trade. That idea will be discussed in section 3.)

Finally, of course, the kind of moral attitudes favoured by these asymmetric growth and other selection processes might just be wrong, or unconnected to other necessary aspects of pursuing the good de dicto. There are plenty of ways to be morally motivated, extremely patient, and totally misguided. Perhaps it's *right* to discount future value!

[98] For relevant discussion, see Trammell, ' Patient Philanthropy in an Impatient World ', Christiano, 'Why might the future be good?' .

2.4. An argument against WAM-convergence

In this section, we give an argument against expecting WAM-convergence, based on thinking through the implications of different meta-ethical positions. Stating it briefly, for now: if some form of moral realism is true, then the correct ethical view would probably strike most people as very weird and even alien. If so, then it's unlikely that people will be motivated to act upon it, even if they learn what the correct moral view is; what's more, if moral beliefs are intrinsically highly motivating, then people will choose *not* to learn or believe the correct moral view.

Alternatively, if some form of moral anti-realism is correct, then there is no objectively correct moral view. But if so, then there's little reason to think that different people will converge to the exact-same moral views. Given our no easy eutopia discussion, that means that not enough people would converge on just the right moral views required to avoid losing out on most value, by the lights of their reflective preferences.

This was all stated in terms of realism and antirealism. We'll continue to use these terms in the longer argument. But if you find the idea of moral realism confused, you could just think in terms of predictions about convergence: whether beings from a wide variety of starting points (e.g. aliens and AIs as well as humans) would, in the right conditions, converge on the same ethical views; or whether such convergence requires a close starting point, too (e.g. such that humans would converge with each other, but aliens and AIs wouldn't); and whether, in those conditions under which you would expect convergence, you would endorse the views which have been converged upon.

2.4.1. Given moral realism

We'll start by considering moral realism. Overall, we think that widespread, accurate, motivational convergence is more likely given moral realism. But it's far from guaranteed. If moral realism is true, then the correct moral view is likely to be much more "distant" from humanity's current preferences than it is today, in the sense that there's much more likely to be a gap between what morality requires and what people would have wanted to do anyway. On realism, people in the future are more likely to have the same (correct) beliefs about what's right to do, but they are less likely to be motivated to act on those beliefs.

This seems true on both "internalist" and "externalist" forms of realism. "Internalist" views of moral judgement understand belief in a moral claim to be essentially motivating: if you believe that stealing is morally wrong then (all other things being equal) you'll be more averse to stealing than otherwise. If internalism about moral judgements is correct, then it's more likely that moral convergence would be motivating. "Externalist" views of moral judgement reject that there's any necessary connection between moral judgements and practical motivation: you could genuinely come to believe that stealing is morally wrong, but become no less averse to stealing.

Either way, we shouldn't be confident that people will act on the correct moral beliefs. If internalism is the right view on moral judgements, then people might just prefer not to learn facts that end up motivating them to act against their personal interests. Maybe something like this is going on when we feel reluctant to be lectured about our obligations to donate money to charity. On the other hand, if externalism is the right view of moral judgements, then society could converge on the right moral views, but not necessarily be motivated to act on them. Maybe something like this is going on when we come to believe we have obligations to donate money to charity, but still don't.

2.4.2. Given antirealism

Next, let's see what follows if moral realism is false : if there aren't any true mind-independent moral facts for society to converge on. We'll focus in particular on subjectivism. (Though note that academic philosophers often refer to subjectivism as a form of realism, just a "non-robust" sort of realism). If subjectivism is true, then the gap between what morality requires and what people would have wanted to do anyway becomes much smaller. But it becomes *much* less likely that different people will converge.

On the simplest kind of subjectivism, an outcome A is morally better than another outcome B just if you prefer for A to happen rather than B.[99] On this view, "accurate" moral convergence would just mean that prevailing views in society converge in a widespread and motivating way toward your own current preferences. But different people today, clearly, have very different preferences (including preferences which are *indexical* to them — i.e. they want *themselves* not to starve much more than they want people *in general* not to starve), so future society can't converge towards all of them. So the simplest kind of subjectivist would have to say that "accurate" moral convergence is very unlikely.[100]

But there are more sophisticated kinds of subjectivism. For example, a subjectivist could view the right actions as those outcomes which some kind of idealised version of themselves would prefer. We could understand "accurate moral convergence" in terms of convergence towards some kind of idealised judgements. So accurate, widespread, motivational convergence might still be on the table.

The question here is whether the idealising processes that different people use all point towards the same (or very similar) views, especially about impartial betterness.

That would be true if there is some kind of *objectively* correct idealising process.[101] But the idea that there's some objectively correct idealising process seems inconsistent with the basic motivation for subjectivism — namely, suspicions about the idea of any notion of objective moral correctness. If you don't *endorse* some alleged objective idealising process, why should you follow it, rather than the process you in fact prefer?

But if the idealising process is itself subjective — based on how you currently would like to reflect, if you could — then WAM-convergence seems unlikely. This is for two reasons.

[99] Subjectivism is normally defined in terms of what's right and wrong; we've adapted it to be about betterness, as that's the focus of this essay series.

[100] It's possible that this kind of subjectivist effectively has a bounded value function — something like "as long as me and my friends and family live good lives, and the future isn't abjectly bad, that's about as good as things can get". Because this includes indexicals ("*me* and *my*"), this person shouldn't expect convergence. But they might also expect convergence to be unnecessary, because their subjective view is easygoing; it presents a wide target. This would be a case of achieving eutopia while effectively nobody else aims at the good, because the narrow target view is incorrect. One comment we'll make here is just that, if you're reading this essay, you probably don't have a view like this! It's also possible (and likely) that most people today just can't be described as having a betterness ordering that can be represented with a cardinal value function at all. As we mentioned, those views are out of scope of the present discussion, but we'd welcome follow-up work which considers how they fit in.

[101] See, for example, Smith, Lewis, and Johnston, 'Dispositional Theories of Value*'.

First, there are a number of important "free parameters" in ethics, and it's hard to see why different idealising processes, from different starting points, would converge on the same view on all of them. Suppose that two people converge all the way to both endorsing classical utilitarianism. This is not close to enough convergence for them to agree on what a mostly great future looks like. In order to get that, both people would need to agree on what hedonic wellbeing really is, and how *precisely* to bring about the most wellbeing with the time and resources available. Both people need to agree, for example, on what sizes and types of brains or other physical structures are most efficient at producing hedonic wellbeing. And, even assuming they agree on (say) a computational theory of mind, they also need to agree on what experiences are actually best.

But there must be a vast range of different kinds of experience which computation could support, including kinds never before experienced, hard to even discover, or impossible to experience while maintaining essential facts about your personal identity. And, given antirealism, there are no shared, obvious, and objective qualities of experiences themselves that indicate how valuable they are.[102] So there's no reason to expect significant convergence among antirealists about which experiences involve the most "efficient" production of "hedonic value".[103] Antirealists can appeal to "shared human preferences" as a reason for moral convergence. But such preferences are extremely underpowered for this task — like two people picking the same needle out of an astronomical haystack.[104]

,

[105]

And, once we move beyond assuming classical utilitarianism, there are many other "free parameters", too. What all-things-considered theory of welfare is correct, and to what precise specification? What non-welfarist goods, if any, should be pursued? How should different goods be traded off against each other? If there are non-linear functions describing how value accrues

[102] Of course, some experiences are strongly desired and others aren't. But this doesn't establish agreement on which experiences are most valuable, for two reasons. First, arguably some people strongly desire experiences which others do not. Second, some experiences might induce strong desire, even when the subject doesn't *themselves* regard the experience as valuable. This could be true of some instances of drug addiction, or in the extreme case it has been shown to induce strong desire for pain stimuli in mice . Further, someone could consistently maintain that an experience is valuable or not, independently of whether it is endorsed or enjoyed by whoever experiences it.

[103] One way to articulate this point is that experiences commonly regarded as hedonically valuable (pleasurable) are heterogeneous — they don't appear to make a natural kind, beyond having in common that they are all commonly regarded as hedonically valuable. For example, Derek Parfit writes in *Reasons and Persons* : "Compare the pleasure of satisfying an intensive thirst or lust, listening to music, solving an intellectual problem, reading a tragedy, and knowing that one's child is happy. These various experiences do not contain any distinctive common quality". This is sometimes called the "heterogeneity problem".

[104] One reply might be to say that everyone prefers experiences that are functionally attractive, and disprefers experiences which are functionally aversive. For example, at least part of what it is to be in a state of pain is a drive to avoid the pain. This could even ground judgements of value — pain is essentially bad because we find ourselves averse to it. Such a view suggests a natural target for convergence, even for antirealists: towards whichever brain states or computations (etc) are most attractive and away from whichever are most aversive. This is interesting, but not enough to establish convergence, because already many people strongly value experiences they don't feel otherwise attracted to (like experiences which involve leaving one's comfort zone), or they strongly value experiences which are aversive in other contexts and for other people, like intense exercise or sad movies.

[105] You might think that there's a much larger set of experiences that each person might regard as close to best. But in the last essay, *No Easy Eutopia* , section 3.2, we argued that the distribution of value/cost is likely fat tailed, such that the vast majority of experiences are less than 50% of the value/cost as the highest value/cost experiences.

(e.g., diminishing returns to certain goods), what is the exact mathematical form of these functions? Each of these questions allows for a vast space of possible answers, and we don't see why different subjective reflective processes, starting from different sets of intuitions, would land on the same precise answers.

Second, the idealising procedure itself, if subjective, introduces its own set of free parameters. How does an individual or group decide to resolve internal incoherencies in their preferences, if they even choose to prioritize consistency at all? How much weight is given to initial intuitions versus theoretical virtues like simplicity or explanatory power? Which arguments are deemed persuasive during reflection? How far from one's initial pre-reflective preferences is one willing to allow the idealization process to take them?[106]

If moral antirealism is true, then the bare tools of idealisation — like resolving inconsistencies and factual disagreements — are greatly underpowered to guarantee convergence. A subjectivist shouldn't hold out hope that society converges on a view they themselves would endorse.[107]

2.5. Blockers

Even if, in a reasonably good scenario, there would be widespread, motivational, accurate convergence, we might still not get to that sufficiently good scenario. I'll call the ways in which society might fail to get to such a scenario *blockers* . One clear blocker is the risk of extinction. But there are other potential blockers, too, even if we survive.

First, there is the risk that humanity will not choose its future at all. In this scenario, the trajectory of the long-term future is best explained as being the outcome of evolutionary forces, rather than being the outcome of some sort of deliberative process. Even if nearly everyone is motivated by the same moral view, still society could collectively fail to bring it about.[108]

Second, people in the future could have the wrong *non-moral* views, especially if those views have greater memetic power (in some circumstances) than the correct views.[109] Some ideas might even be so memetically potent that merely considering the idea makes an individual highly likely to adopt it; such ideas might also discourage further change, and become impossible to get out of, like epistemic black holes.

Third, there could be early lock-in: when the most-important decisions are being made, the decision-makers at the time are *unable* to bring about a mostly-great future, even if they tried. We'll discuss lock-in more in the next essay.[110]

[106] For a sustained argument along these lines, see Carlsmith, ' On the limits of idealized values '.

[107] In response to this argument, some commenters have said that they expect, post-reflection, to have an intrinsic preference for other people also getting what they want, especially for people that are currently similar to them. This is a reason for accurate convergence given antirealism. But there are at least some reasons to doubt that your preferences are structured this way. The "you" at the end of reflection will likely be extraordinarily different from the you now, as will the post-reflection version of some other person . Hoping that most people end up, after reflection, with similar preferences to you might be like a six-year-old confidently believing they'll be best friends forever with their classmate.

[108] Assadi, ' Will Humanity Choose its Future? '.

[109] Boudry, ' On epistemic black holes: How self-sealing belief systems develop and evolve '.

[110] A more subtle form of early lock-in is via *atrocities* . For example, in the last essay, we considered moral views that are meaningfully bounded above but not below, and where goods and bads are aggregated separately. On this view, the more bad things occur, the lower the value of the best-attainable future; future goods can't wholly compensate for the bads. If so, then it could be that in the near future there are preventable atrocities

Of course, there could also be additional and as-yet unknown blockers.

3. What if some people aim at the good?

In this section, we'll consider the possibility that (i) under reasonably good conditions, some meaningful fraction of people (weighted by the power they have) would converge on the correct moral view and would be motivated to use most of the resources they control towards promoting the good (de dicto), and (ii) those people will be able to bargain or trade with each other, such that we'll get to a mostly-great future. By "some meaningful fraction" we mean a minority of all people, but not less than (say) one in a million people.[111] So understood, we'll call idea (i) *partial AM-convergence* .

In this section, we'll discuss whether bargaining and trade will be enough to reach a mostly-great future, given partial AM-convergence.

In section 3.1, we'll illustrate how such "moral trade" could work, and discuss the conditions under which such trade could occur. In section 3.2, we discuss the conditions under which trade could enable a near-best future, and in section 3.3 we discuss the problems of threats. In section 3.4, we put these considerations together, depending on whether the correct moral view is unbounded or bounded, whether bads weigh heavily against goods, and whether bads and goods are aggregated separately or not. In section 3.5, we discuss blockers to a mostly great future via AM-convergence and trade.

3.1. Trade and compromise

Imagine that a multi-millionaire has a niche but intense wish to own every known Roman coin with a portrait of Julius Caesar. He might not be able to own every coin: some are kept in museums, some private owners are unwilling to sell for sentimental reasons. But he might manage to acquire *most* Julius Caesar coins in the world, despite controlling a small fraction of the world's wealth. When he paid each collector for their coin, they preferred his money to their coin, and the multi-millionaire preferred the coin, so each party was glad to make the trade.

Someone could trade for moral reasons, too. You could find some voluntary exchange where you (or both parties) are motivated by making the world a better place. For the price of a caged bird, for example, you can pay to set a bird free. For the price of a plot of Amazonian land, you can save that land from being deforested. But even better bargains are on offer when you have moral reasons to care *far* more about some outcome than your counterparty. You could (in theory) pay a retailer to discount alternatives to animal products, for example. Or *both* parties could trade for moral reasons: Annie cares a lot about recycling and Bob cares a lot about littering, so Bob might agree to

such that (i) in the future, there will be widespread, accurate, motivating moral convergence but (ii) those future decision-makers won't be able to reach what we would consider a mostly-great future. Other moral views might have a similar implication. We think this point is intellectually interesting, but probably of little practical importance. If you have a view on which atrocities can lower the upper bound on the best-achievable future, then you should probably be more concerned about avoiding futures involving atrocities over the long term than avoiding short-term atrocities in scenarios that are otherwise very good.

[111] One-millionth of the global population today is around 8,000 people.

start recycling if Annie stops littering, and both parties agree the world is now a better place. This is the promise of *moral trade* .[112]

In the future, there could be potential for enormous gains from trade and compromise between groups with different moral views. Suppose, for example, that most in society have fairly common-sense ethical views, such that *common-sense utopia* (from the last essay) achieves most possible value, whereas a smaller group endorses total utilitarianism. If so, then an arrangement where the first group turns the Milky Way into a common-sense utopia, and the second group occupies all the other accessible galaxies and turns them into a total utilitarian utopia, would be one in which both groups get a future that is very close to as good as it could possibly be. Potentially, society could get to this arrangement even if one group was a much smaller minority than the other, via some sort of trade. Through trade, both groups get a future that is very close to as good as it could possibly be, by their lights.

Some reasons for trade might become less relevant in a more technologically advanced society. For example, with superintelligent advice, investors might less often have differing expectations about the future, which would be a reason for trading volume to decline (all else equal).[113]

But other reasons for trade would likely remain equally or more relevant, even in a technologically mature society. Different groups could continue to:

- Value different natural resources
    - For example, some groups might want star systems that contain rocky planets in the habitable zone to live on; other groups might want to preserve particularly beautiful areas of the cosmos; etc.[114]
- Value different locations of resources
    - For example, some groups might intrinsically value resources in our own solar system more than those in distant galaxies; other groups might be indifferent.
- Have different rates of pure time preference
    - Some groups might value being able to acquire tradeable goods *soon* (such as to benefit currently-living people), while others might be indifferent about when they acquire or use them.
- Have different attitudes to risk, when some risk is ineliminable
    - For example, even for an extremely technologically advanced society, it might be impossible to know for sure whether a distant galaxy will have been settled by aliens or not.

In many cases, two views can be "resource-compatible", meaning there is some way to *almost* fully satisfy both views with the same resources, even if the *best* use of the resources on one view is mostly worthless to the other view, such that the gains from trade between two moral views controlling comparable resource shares could be large from each view's perspective. For example: hedonists might only care about bliss, and objective list theories might care primarily about

[112] Ord, 'Moral Trade'.

[113] Harrison and Kreps, 'Speculative Investor Behavior in a Stock Market with Heterogeneous Expectations' .

[114] It could be true that across cosmic distances, the opportunity cost of 'domestic' production could be lower than the cost of importing goods. But, to be clear, we are mostly imagining a period of bargaining before widespread space settlement — so the instruments being traded here are going to be more like options, futures, claims to, and other contracts regarding resources not yet accessed.

wisdom; they might potentially agree to create a shared society where beings are both very blissful and very wise. Abstractly speaking: the same resources can sometimes achieve much more value on two different views if they are all devoted to some "hybrid good" than if the resources were simply split evenly between the two views without the possibility of bargaining or trade.

What's more, some of the reasons why we *don't* get trade today would no longer apply. As discussed in the next essay, superintelligence could enable iron-clad contracts, which could avoid the problem of a lack of mutual trust. And transaction costs would generally be extremely small relative to the gains, and more potential trades would be salient in the first place, given an enormous number of superintelligence delegates able to spend abundant time in order to figure out positive-sum arrangements.

3.2. Would trade enable a mostly-great future?

In *many* cases, it looks like there will be the potential for truly enormous gains from moral trade. This is a significant cause for optimism. For example, even if most people care little about the welfare of digital beings, if it's sufficiently low-cost for them to improve that welfare, the minority who does care about digital welfare would be able to bargain with them and increase their welfare considerably.

More generally, we can consider some key possibilities relating to moral trade:

1. The extent to which the hypothetical gains from frictionless trade are actually realised.
2. The (moral) gains from trade from all trades that in fact take place.[115]
3. The (moral) gains from trade for the correct view.[116]
4. The value of the world, on the correct moral view, after trade.[117]

What ultimately matters is (4), the value of the world on the preferred view after trade. But if you are uncertain over moral views, then you should also be interested in (2): the extent to which most views gain from trade. Of course, (2)–(4) depend on (1), whether possible gains from trade are actually realised. That could depend on whether, for example, the right institutions exist to support trade, and also the extent to which different views actually take trades which improve the world by their lights, which isn't guaranteed.[118]

However, mutual gains from trade seems especially unlikely to us if the prevailing views are non-discounting and linear-in-resources. It's possible in principle, because of the potential for hybrid goods. But if the ways to achieve maximum value/cost on each view are both highly particular, then it's unlikely any compromise could achieve much more value (by the lights of each view) than if each view kept their resources for themselves.

[115] For example, in terms of the sum total of additional resources required by each actor under no trade to compensate for the gains in value, on their view, from trade.

[116] This would be higher if: concessions highly valued by incorrect views are cheap on the preferred view; or concessions highly valued by the correct views are cheap on non-preferred views. In other words, the cost of achieving impartial value through trade is initially much lower than the cost without trade, on the preferred view.

[117] This could depend on: (i) the bargaining power of the preferred view; (ii) the resources the preferred view initially controls; (iii) other changes to the world from trade not involving the preferred view.

[118] For example, you could think that buying carbon offsets does make the world better, but that it is wrong for nonconsequentialist reasons to purchase carbon offsets.

In fact, there may even be a "narcissism of small differences" effect: if two views agree on the importance of the very same domains or aspects of the world, but disagree over what to do with them, then the gains from compromise could be smaller, despite the views seeming superficially similar. Suppose Annie and Bob agree on the importance of spiritual worship, but disagree over which deity to worship. By contrast, Claudia mostly cares about the environment. Although Annie and Bob's views are superficially similar, the mutual gains from moral trade between Claudia and either of Annie and Bob are likely larger than those between Annie and Bob, because Annie and Bob both agree there is no "hybrid" activity of worshipping two deities at the same time.

Given this, it becomes crucial to know: assuming that the right moral view is non-discounting and linear-in-resources, how much control over resources is such a view likely to have, before trade?

The case for thinking that non-discounting and linear *moral* views are likely to begin with a significant share of resources is that, currently, such cosmic-scale values are very unusual, and are unusually likely to be held by the altruistically-minded. Of course, altruistically-minded people disagree widely today, so we'd also have to hope for significant AM-convergence among those people after some period of reflection. The counterargument is that, in a technologically advanced society, many *more* types of people will have views that are non-discounting and linear, for two reasons. First, superintelligence-aided reflection might plausibly cause people to be much more likely to adopt views that are non-discounting and linear. Second, as we saw in section 2.3.2., enormous wealth might cause people to shift focus towards their preferences which diminish most slowly with resources, even if those preferences are self-interested.

Putting this together, it's hard to know what fraction of resources the correct view will control among all other non-discounting and linear views, assuming that the correct moral view has those properties. Here's a rough guess: if (at the relevant period of bargaining) 1 in N people out of everyone have altruistic preferences, then the fraction of those people among everyone with approximately linear preferences of any kind, is something like 10 in N.

3.3. The problem of threats

If trade and bargaining are reasons for optimism, then *threats* could undermine that optimism. Suppose that Alice and Bob want different things, and bargain with each other. Alice could extort Bob by credibly committing to make the world worse on Bob's view, unless Bob makes concessions to Alice. In the previous examples, both parties to a trade can agree to the trade itself, and also agree the world is made no worse-off by whatever commitment or enforcement mechanisms *enabled* the trade[119] in the first place. Threats are different. When one party carries out a threat, they may both agree to the terms of the threat, but at least one party might view the world as made worse-off by the threat having been made, regardless of whether they capitulated to it. Imagine Bob cares a lot about animals being mistreated, and Alice doesn't. Then Alice can threaten to mistreat lots of animals to extort resources from Bob for uses he doesn't value at all. Then, whatever Bob chooses, he'll likely view the world as worse than it would have been if Alice had never been in the picture.[120]

[119] Or contracts more generally, since examples where one or both parties perform some service, and don't swap tangible goods, aren't obviously "trades".

[120] Alice could also threaten to destroy something Bob positively values, refrain from actions Bob values, or refrain from cheaply producing something Bob values. Ultimately we find these possibilities less concerning, because it generally seems more effective for Alice (barring legal and other considerations) to cause specific harms or wrongs — in general, the threatening can exert more leverage at the same personal cost by

The extent of public writing on threats is very limited;[121] sometimes just learning about the topic can make it more likely for threats to occur, so people are naturally reluctant to spread their research widely. We ourselves have not particularly dug into this issue, despite its importance.

But, significantly, even small risks of executed threats can easily eat into the expected value of worlds where many groups with different values are able to bargain with each other. On many views, bads weigh more heavily than goods: this is true on negative-leaning views, or on views on which value is meaningfully bounded above but not meaningfully bounded below, and goods and bads should be aggregated separately; or simply on empirical views where it turns out some bads are much cheaper to create with the same resources than any compensating goods. If so, then even if only a small fraction of resources are devoted to executed threats, most of the value of the future could be lost;[122] if a large fraction of resources are devoted to executed threats, then most value could be lost even if the correct view is not negative-leaning.[123] What's more, those who hold the correct moral view may be less likely to themselves threaten other groups, so even if no threats are ever executed, those people could still lose most of their resources via extortion. Finally, it's not obvious to us that some kind of legal system which reliably prevents value-undermining threats would be mutually agreeable and stable, so the worry does not only apply to legal 'anarchy' between views.[124]

3.4. Putting it all together

How much more optimistic should the idea of trade and compromise make us? We can give an overall argument by looking at different types of moral view, using some of the distinctions we covered in the last essay, *No Easy Eutopia* , section 3 .

If executed moral threats amount to a small but meaningful fraction of future resource use, then if the correct view:

- Is bounded above, and goods and bads are aggregated jointly, and if bads don't weigh very heavily against goods, a mostly-great future seems likely.

threatening to specifically cause harm, than to destroy or refrain from producing a compensating amount of good. But note that, today, there are typically far fewer options for legal recourse against threats to refrain from causing large amounts of good.

[121] For extant discussion, see: Clifton, 'Preface to CLR's Research Agenda on Cooperation, Conflict, and TAI' , and Baumann, 'Using surrogate goals to deflect threats' .

[122] As well as axiological views, some decision theories can be more vulnerable to the threats we're concerned about here, too. We omit that discussion here.

[123] Suppose that executed threats amount to 1% of all resources. Then if bads and goods weigh equally (e.g. on classical utilitarianism), even worst-case threats only reduce the value of the future by 2 percentage points. But if bads weigh 100x as much as goods, then 1% of resources spent on worst-case threats would mean that the future is certainly negative.

[124] Of course legal systems today legislate against threats, extortion, and blackmail. Threatening a criminal act, especially physical violence, is often itself treated as a crime. Most legal systems have a concept of duress, where one party enters into a contract "unwillingly" or with "no reasonable alternative". Legal concepts of blackmail typically rest on the threat of harms or damages to the victim, but the threshold is typically high: it is not unlawful for Bob to ask Alice for money because otherwise he'll be sad, in a way which is distressing to Alice. But so far legal systems do not draw the line in a way that excludes the possibility of value-destroying threats, which suggests it might be very difficult to draw the line in an objective way.

- Is bounded above, and goods and bads are aggregated jointly, and bads do weigh heavily against goods, then a mostly-great future seems unlikely, because the executed threats eat into the value of the goods.
- Is bounded above, and goods and bads aggregated separately, then a mostly-great future seems unlikely.
    - If the view is bounded below (with a compensating or greater magnitude to the lower bound), then future society is likely to reach both upper and lower bounds, and we will end up with a future that's at most of net zero value.
    - If the view is unbounded below, or the magnitude of the lower bound is much greater than the magnitude of the upper bound, then the bads will outweigh the goods, and we'll probably end up with a net-negative future.
- Is unbounded above, then even if bads don't weigh heavily against goods, a mostly-great future seems uncertain.
    - If the correct view is unusually resource-compatible with other prevailing views, then a mostly-great future is plausible. But resource-compatibility between linear views seems unlikely, and if so then a mostly-great future seems unlikely, too.
- Is unbounded above, and bads do weigh heavily against goods, then a mostly-great future is unlikely, because executed threats eat into the value of the goods.

If value-destroying threats can be prevented, then things seem more optimistic. In this scenario, then it's just on the linear views that we don't reach a mostly-great future via compromise and trade.

We think it's appropriate to be highly uncertain about which axiological view is correct. Given that, it's worth considering what the value of the future looks like, from our uncertain vantage point. Two things stand out. First, most views we've considered are sensitive to value-destroying executed threats, which suggests we should try hard to prevent such threats, even if doing so is itself costly. Second, some views are highly resource-compatible with others; in cosmic terms it could be very inexpensive to achieve a near-best future for many of those views by giving even just a small fraction of all resources to views that are easily-satiable.[125] We should aim to do so.

3.5. Blockers to trade

The scenarios where we get to a mostly-great future via trade and compromise face the same major blockers that we discussed in section 2.5, and two other blockers, too. First, *concentration of power* . If only a small number of people have power, then it becomes less likely that the correct moral views are represented among that small group, and therefore less likely that we get to a mostly-great future via trade and compromise.

Second, even if power is not concentrated, the most valuable futures could be sealed off. We already saw one way this could be so: if people make and execute value-destroying threats. But there are a lot of other ways, too. For example, if decisions are made by majority rule, then the

[125] Note that the *importance* of doing so depends on how to normalise bounded and unbounded views under uncertainty: should the vast near-best world on an unbounded view loom much larger than the modest near-best world of a bounded view, given equal credence in both views? This was discussed in section 3.4 of No Easy Eutopia .

majority could ban activities or goods which they don't especially value, but which minority views value highly.[126]

Alternatively, collective decision-making procedures can vary in the extent to which they elicit decisions made on the basis of good reasons versus other things. Perhaps decisions are made democratically, but such that the incentives favour voting in order to signal allegiance to a social group. And even if some collective decision-making process is designed to give everyone what they want, that can be very different from a decision-making process that is the best at aggregating everyone's best guesses at what the right decision is; these two broad approaches can result in very different outcomes.[127]

Even the same decision-making procedure can give very different results depending on *when* the decision occurs. If, for example, Nash bargaining were to happen now, then people with quasilinear utility might get most of what they want just by getting the goods that have heavily diminishing utility for them. But, once such people are already very rich, their preferences for additional goods would be linear, and couldn't be so easily satisfied.[128] And, of course, the outcome of any bargaining process depends sensitively on the power distribution among the different bargainers, and on what would happen if no agreement occurs; this is something that can vary over time.

The more general point is that the outcomes we get can vary greatly depending on which collective decision-making processes are used. Some such processes might well be predictably much better than others.

4. What if no one aims at the good?

In sections 2 and 3, we looked at scenarios where at least some people were deliberately optimising towards the good (de dicto). But let's now discuss whether this is really needed. Perhaps we can hit a narrow target even if no one is aiming at that target.

The key argument against such a position is that it would seem like an extraordinary *coincidence* if society hit the narrow target without explicitly trying to do so. So there needs to be some explanation of why this isn't a coincidence.

There are possible explanations. In terms of our analogy of sailing to an island: perhaps none of the shipmates intrinsically care about reaching the right island, but do care about their own salary, and get paid only if they do their job. If all the shipmates just do their job well enough then the ship reaches the island: the shipmates' desire for their salary steers the ship in the right direction. Alternatively, consider again that flight evolved in animals, more than once, even though no agent tried to optimise for that target; flight evolved because it was useful for maximising reproductive fitness.

[126] Consider how most rich countries criminalise the distribution of MDMA and psychedelic drugs, which are not highly valued by the majority, but are arguably not very socially harmful and are highly valued by a minority of people.

[127] Consider, for example, max-preference-satisfaction versus Kemeny-Young judgment aggregation .

[128] Similarly, right now, none of us know what our reflective preferences are. So people are behind a veil of ignorance. Given this, people might be much more willing to make decisions in accordance with some particular widely-endorsed reflective process than they would after they have actually done the reflection themselves.

4.1. If most aim at their self-interest

First, let's consider cases where (i) nobody is aiming at the good (de dicto), and (ii) most people are pursuing their own local interests, but (iii) absent blockers, most people converge on the right view of what's best *for them* . Couldn't that still be enough to reach a mostly great future?

If there's some strong correlation between what's good for every individual and what's good overall, then it isn't a coincidence that people pursuing what's best for them results in what's near-best overall.

Consider an analogy with material wealth. Historically, most people have been motivated to pursue material wealth for themselves and their circle of concern (like their loved ones or family). Very few people across history seem to have been motivated to pursue material wealth for the entire world, or across the entire future, or even for their own countries. And yet, humanity at large is vastly more materially wealthy than it was even a few centuries ago. "It is not from the benevolence of the butcher, the brewer, or the baker, that we expect our dinner," wrote Adam Smith, "but from their regard to their own interest."

How did this happen? One happy fact is that two parties can both become richer, in the long run, by engaging in forms of collaboration and coordination. This is true at many levels of organisation — from small groups, through to national systems of law and commerce, through to international trade and coordination. Perhaps, if most people converge on wanting what's ultimately good for them, the same trend could enable people to increasingly improve their own lives, and that could be enough to make the future mostly-great.

To be sure, given cognitive and material abundance, people will probably and in general get much more of what they self-interestedly want than they do today. But the crucial question is whether people pursuing what's best for themselves is a strong enough driver towards what's best *overall* . Otherwise, again, a world driven by self-interest could become staggeringly better than the world today, but still only a fraction of how good it could be, for want of a few crucial components.

On some moral and meta-ethical views, the overlap between the aggregate of individual self-interest and overall value *might* be strong enough. Consider a view on which (i) totalising welfarism is the correct axiological view; and (ii) "meta-ethical hedonism" is correct, in the sense that goodness is a property of conscious experiences that we have direct access to, and coming into contact with that property makes it more likely that the experiencer will believe that that experience is good.[129] On such a view, it becomes more plausible that everyone pursuing their own self-interest will converge on enjoyment of the *same* kind of highly valuable experiences, and that such a world is near-best overall. If so, everyone pursuing their own self-interest could well lead toward a very similar outcome than if they were all motivated instead to promote the moral good.[130]

We think these views have some merit, but we don't think you should be at all confident in either of them. And, even if we accept them, people pursuing their own self-interest is only sufficient to

[129] The case of pain is often described like this. It's hard for some to imagine being in intense pain, but not at all thinking they are in pain; or thinking they are in intense pain, but in fact not being in pain. That suggests that being in pain just is something very like thinking you are in pain. Pain, then, is transparent to us. The thought is that this is true of the properties of goodness and badness themselves — what it is for an experience to be good is for it to feel good to us.

[130] A similar result could follow on preference-satisfaction theories of wellbeing.

create a mostly-great future if the most valuable experiences can be experienced by those people. But, plausibly this is wrong, if the very best experiences are so alien that most people initially pursuing their self-interest cannot *themselves* experience them; similar to how a bat can't comprehend what it is like to be a human without ceasing to be a bat. If so, most people can only bring about the best experiences for some other kind of person, not for themselves. So most people would not pursue them only out of self-interest.[131]

And if we drop either of these views, then the correlation between self-interest and the overall good is just not strong enough. People could be motivated to improve their own lives, but not to create enough additional good lives, if doing so doesn't also benefit them. They could even be motivated to create low-value or even bad lives (e.g. digital servants) if doing so does benefit them. Or some aspects of value of the future might not be good for anyone, like the possibly intrinsic value of nature or beauty. Or some things could even be locally good, but bad overall — for example, on some egalitarian views, making already well-off people even better off can be bad because it increases inequality.

People could also fail to pursue their own good. If there is a single correct conception of the prudential (self-interested) good, then many of the arguments for no easy eutopia would also suggest that the target for prudential good is also narrow, not obvious, and non-motivational perhaps even after reflection. Suppose, for example, that hedonism about wellbeing is correct; nonetheless, people might just prefer not to turn themselves into machines for generating positive experiences — perhaps reasonably.

With these points in mind, we can see that the analogy between individual and total good, and creation of individual and global wealth, could be very misleading. Wealth is held and generated by people, but there could be forms of value which are not good for anyone, like the value of nature. The instrumental value of money is obvious, but what is ultimately good for somebody might not be. Moreover, it's not even clear that self-interest would succeed in generating nearly as much wealth, in the long run, as some termite-like motivation to generate global wealth per se.

4.2. If other motivations point towards good outcomes

One way in which it wouldn't be necessary for most people in society to aim at the good de dicto is if intrinsically valuable goods are also *instrumentally* valuable on a wide range of other moral views. For example, perhaps some of the following goods are intrinsically valuable: the acquisition of knowledge; the proliferation of life; the proliferation of complex systems; feats of engineering; preference-satisfaction.

Whatever goals people have in the future, their plans are likely to involve at least some of acquiring knowledge, proliferating life and complex systems, achieving feats of engineering, and satisfying their own preferences and the preferences of those beings they create. So if at least some of these things are good in and of themselves, then the future will be better, even if future people are creating those goods only as a means to some other end.

[131] On the other hand, we might still hope that the concept of personal identity is fluid or indeterminate enough that either (i) the best-per-unit-resource experiences are accessible to a very wide variety of people, (ii) upon reflection, people would stop caring about personal identity, in which case their motivations would look much more impartial.

However, the question is whether this is sufficient to get us to a mostly-great future. And, again, that looks unlikely. People who only care about some goods for instrumental reasons are likely to produce less of them compared to people who intrinsically care about accumulating those goods. As an analogy, if Alice intrinsically cares about owning as many books as possible, and Bob values books because he finds it useful to be well-read, then Alice is likely to end up owning far more books (including books she will never read) than Bob ever will. The situation becomes worse again if the correct moral view intrinsically values some non-instrumentally valuable good *even more* than it values the instrumentally valuable goods, or if the correct moral view intrinsically *disvalues* some instrumentally valuable goods.

The intrinsic value of some instrumental goods might be a reason for thinking that the future is very good in absolute terms; but, given easy eutopia, the correlation between instrumental and intrinsic value just doesn't seem strong enough to get us to a mostly-great future.

5. Which scenarios are highest-stakes?

In response to the arguments we've given in this essay, and especially the reasons for pessimism about convergence we canvassed in section 2, you might wonder if the practical upshot is that you should pursue personal power-seeking. If a mostly-great future is a narrow target, and you don't expect other people to AM-converge, then you lose out on most possible value unless the future ends up aligned with almost exactly your values. And, so the thought goes, the only way to ensure that happens is to increase your own power by as much as possible.

However, we don't think that this is the main upshot. Consider these three scenarios:

1. Even given good conditions, there's almost no AM-convergence between any sorts of beings with different preferences.
2. Given good conditions, humans generally AM-converge on each other; aliens and AIs generally don't AM-converge with humans.
3. Given good conditions, there's broad convergence, where at least a reasonably high fraction of humans and aliens and AIs would AM-converge with each other.

(There are also variants of (2), where "humans" could be replaced with "people sufficiently similar to me", "co-nationals", "followers of the same religion", "followers of the same moral worldview" and so on.)

Though (2) is a commonly held position, we think our discussion has made it less plausible. If a mostly-great future is a very narrow target, then shared human preferences are underpowered for the task of ensuring that the idealising process of different humans goes to the same place. What would be needed is for there to be something about the world itself that would pull different beings towards the same (correct) moral views: for example, if the arguments are much stronger for the correct moral view than for other moral views, or if the value of experiences is present in the nature of experiences, such that by having a good experience one is thereby inclined to believe that that experience is good.[132]

[132] This is the "meta-ethical hedonist" position that we discussed in section 4.1.

So we think that the more likely scenarios are (1) and (3). If we were in scenario (1) for sure, then we would have an argument for personal power-seeking (although there are plausibly other arguments against power-seeking strategies; this is discussed in section 4.2 of the essay, *What to do to Promote Better Futures*). But we think that we should act much more on the assumption that we live in scenario (3), for two reasons.

First, the best actions are higher-impact in scenario (3) than in scenario (1). Suppose that you're in scenario (1), that you currently have 1 billionth of all global power,[133] and that the future is on track to achieve one hundred millionth as much value as if you had all the power.[134] Perhaps via successful power-seeking throughout the course of your life, you could increase your current level of power a hundredfold. If so, then you would ensure that the future has one millionth as much value as if you had all the power. You've increased the value of the future by one part in a million.

But now suppose that we're in scenario (3). If so, you should be much more optimistic about the value of the future. Suppose you think, conditional on scenario (3), that the chance of *Surviving* is 80%, and that *Flourishing* is 10%. By devoting your life to the issue, can you increase the chance of *Surviving* by more than one part in a hundred thousand, or improve *Flourishing* by more than one part in a million? It seems to me that you can, and, if so, then the best actions (which are non-powerseeking) have more impact in scenario (3) than power-seeking does in scenario (1). More generally, the future has a *lot* more value in scenario (3) than in scenario (1), and one can often make a meaningful proportional difference to future value. So, unless you're able to *enormously* multiply your personal power, then you'll be able to take higher-impact actions in scenario (3) than in scenario (1).

A second, and much more debatable, reason for focusing more on scenario (3) is that you might just care about what happens in scenario (3) more than in scenario (1). Will's preferences, at least, are such that things are much lower-stakes *in general* in scenario (1) than they are in scenario (3): he thinks he's much more likely to have strong cosmic-scale reflective preferences in scenario (3), and much more likely to have reflective preferences that are scope-sensitive and closer to contemporary common-sense in scenario (1).

6. Conclusion

This essay has covered a lot of different considerations, and it's hard to hold them all in mind at once. Remember, we started the essay on the assumption that mostly-great futures are a very narrow target. On the face of it, that suggests pessimism about whether the future will be mostly-great. Overall, we think the considerations in this essay warrant significantly more optimism than that initial impression.

Still, we don't think the points in this essay establish that a mostly-great future is likely, even in the absence of blockers. Given moral antirealism, the common core of "human values" doesn't seem specific or well-powered enough to make WAM-convergence look likely. And even if there is some attractive force towards the best moral views, like moral realism would suggest, still it might not be sufficiently motivating. And even if there would be WAM-convergence under good enough

[133] Sadly, this argument only goes through if you are not already extremely powerful.

[134] Where this is greater than your current fraction of power because of the chance that you are able to influence the most major decisions and engage in moral trade.

conditions, we might not reach those conditions. So widespread, accurate, and motivational convergence looks unlikely.

The potential of scenarios involving partial AM-convergence, plus trade or compromise, seems more realistic and promising to us. But there are major obstacles along that path too, including concentration of power and poor collective decision-making processes. Moreover, value-destroying threats could rob these scenarios of most of their value, or even render them worse than extinction.

Backing up even further, on the basis of all this discussion, we think that the hypothesis from the first essay — that we are far from the ceiling on *Flourishing* — seems right. Among problems we might consider, those that hold the future back from flourishing therefore seem greater in scale than risks to our survival. But is there anything we can do about them? In the next two essays, we turn to that question.

Bibliography

Guive Assadi, ‘ Will Humanity Choose its Future? ’.

Tobias Baumann, ‘ Using surrogate goals to deflect threats ’, *Center on Long-Term Risk* , 20 February 2018.

Maarten Boudry, ‘ On epistemic black holes: How self-sealing belief systems develop and evolve ’, 2023.

Richard Boyd, ‘ How to Be a Moral Realist ’, *Essays on moral realism* , 1988.

David Owen Brink, ‘ Moral Realism and the Foundations of Ethics ’, 1989.

Joe Carlsmith, ‘ On the limits of idealized values ’, 2021.

Paul Christiano, ‘ Why might the future be good? ’, *Rational Altruist* , 27 February 2013.

Jesse Clifton, ‘Preface to CLR’s Research Agenda on Cooperation, Conflict, and TAI ’, 13 December 2019.

Francis Fukuyama, ‘ The End of History and the Last Man ’, 17 September 2020.

Robin Hanson and Kevin Simler, ‘ The Elephant in the Brain: Hidden Motives in Everyday Life ’, 2017.

J. Michael Harrison and David M. Kreps, ‘Speculative Investor Behavior in a Stock Market with Heterogeneous Expectations’ , *The Quarterly Journal of Economics* , 1 May 1978.

Douglas Hofstadter, ‘Heisenberg’s Uncertainty Principle and the Many Worlds Interpretation of Quantum Mechanics ’.

Joshua Conrad Jackson and Danila Medvedev, ‘Worldwide divergence of values’ , *Nature Communications* , 9 April 2024.

William MacAskill, ‘What We Owe the Future: A Million-Year View’ , 2022.

Toby Ord, ‘Moral Trade’, *Ethics* , 2015.

Peter Railton, ‘ Facts, Values, and Norms: Essays toward a Morality of Consequence ’, 2003.

Michael Smith, David Lewis, and Mark Johnston, ‘Dispositional Theories of Value*’, *Aristotelian Society Supplementary Volume*, 1 July 1989.

Phil Trammell, ‘Patient Philanthropy in an Impatient World’, 2022.

Bernard Williams, ‘Persons, character and morality’, *Moral Luck: Philosophical Papers 1973–1980*, 1981.

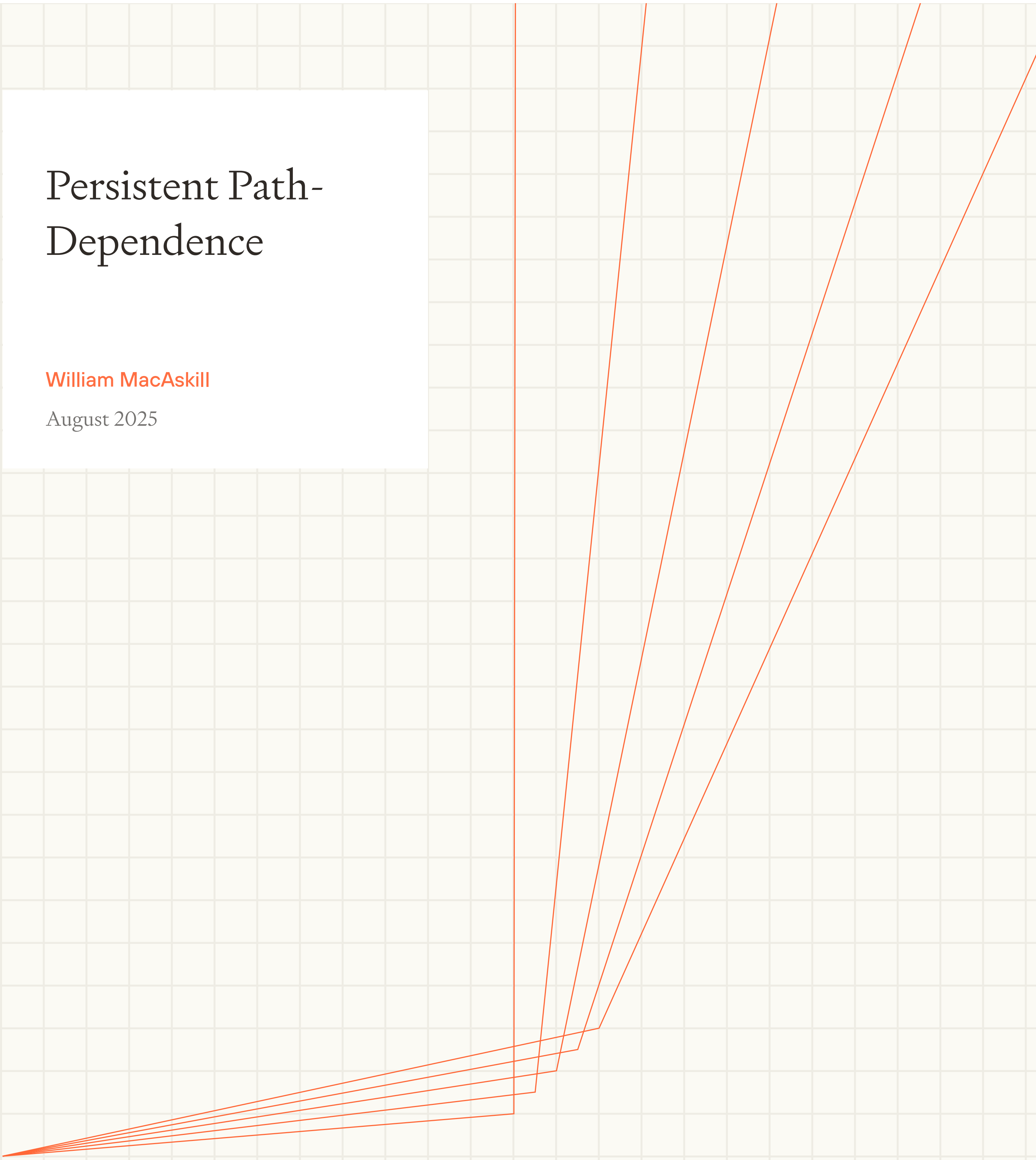
World Values Survey, ‘Inglehart–Welzel Cultural Map’, 2023.**

Forethought

Persistent Path-Dependence

William MacAskill

August 2025



1. Introduction

One of the most common objections to working on better futures is that, over sufficiently long time horizons, the effects of our actions will ‘wash out’.[135] This is often combined with the view that extinction is a special case, where the impacts of our actions really could persist for an extremely long time. Taken together, these positions imply that it’s much more important, from a longtermist perspective, to work on reducing extinction risk than to work towards better futures. The future we’ll get given survival might only be a fraction as good as it could be, but we might just be unable to predictably improve on the future we get. So we should focus on *Surviving* rather than *Flourishing* .

In this essay, I’ll argue against this view. There are a number of events that are fairly likely to occur within our lifetimes that would result in extremely persistent path-dependent effects of predictable expected value. These include the creation of AGI-enforced institutions, a global concentration of power, the widespread settlement of space, the first immortal beings, the widespread design of new beings, and the ability to self-modify in significant and lasting ways.

I’m not confident that such events will occur, but in my view they’re likely enough to make work on better futures high in expected value from a long-term perspective. To be more precise, my view is that the expected variance in the value of the future will reduce by about a third this century, with the majority of that reduction coming from things other than the risk of human extinction or disempowerment to AI.

In section 2 of this essay, I’ll explain why the skeptical argument I’m considering is more complicated than it first appears, and doesn’t justify some typical longtermist priorities, like preventing AI takeover. In section 3, I explain the concepts of lock in and path-dependence. Section 4 is the bulk of the paper, where I discuss mechanisms that could enable persistent path-dependence. In section 5, I introduce the idea of “lock-in escape velocity” as a reason why persistent path-dependence is more likely than you might have thought. In section 6, I argue that it’s fairly likely that, given events in our lifetimes, we can have persistent path-dependent impacts on the trajectory of the future.

2. Human extinction and AI takeover

Extinction is often regarded as a unique case, where actions to reduce the risk of extinction really can predictably affect the value of the very long-term future. The thought is: by reducing the risk of human extinction, we increase the probability that there is any civilisation at all in the long term, in our cosmic neighborhood. So, in order to conclude that reducing extinction risk is good, the only view we need to have is that the expected value of future civilisation is positive rather than negative or neutral. Later in this piece, I’ll argue against the *uniqueness* of extinction’s (purported) predictable long-run significance. In this section, I’ll make two points about the way in which extinction has long-run significance.

[135] Where the effects of an action 'wash out' by time T just in case the action makes no or negligible predictable difference to the value of any time after T.

First, you might think that reducing the risk of human extinction by one percentage point increases the probability that there is any civilisation in our cosmic neighborhood by one percentage point. But this isn't right. If humans go extinct and most other mammals do not, it seems quite likely to me (more than 50%)[136] that, in the hundreds of millions of years remaining before the Earth is no longer habitable, some other species will develop higher intelligence, cumulative cultural evolution and technological capability, such that they can rebuild civilisation in our stead.[137] And, even if all life on Earth is wiped out, or if higher intelligence never re-evolves, it seems somewhat likely to me (around 50%)[138] that alien civilisations will eventually settle our part of the cosmos.[139]

This means that, in reducing the risk of human extinction, we are mainly affecting *who* occupies our corner of the cosmos, rather than *whether* it gets occupied. We are still somewhat increasing the chance that there's any civilisation in our part of the cosmos, because it's not *certain* that non-human or alien civilisations would fill in the gap. But it means that, in order to believe that extinction risk reduction is positive in expectation, you must have the view that these alternative civilisations *wouldn't* be much better than human civilisation. The view is *not* merely resting on the idea that the expected value of future civilisation is positive rather than negative or neutral.

In the spirit of scepticism about predictable long-term effects, you might invoke some principle of indifference and think that these non-human civilisations would be equally as good, in expectation, as human-originating civilisation. If so, then extinction risk reduction still looks positive. But it's meaningfully lower in expected value (e.g. 75% lower) than you would have thought without considering replacement civilisations.

A second point is on AI takeover. If you have the view that futures with civilisation are overall better than empty futures but that it's hopeless to predict or influence the value of futures with civilisation, then reducing the risk of AI takeover is not a way of predictably positively influencing the long term. If AI disempowers or even kills humanity, then it will (probably) continue to build a growing AI-civilisation afterwards.

The view that AI takeover is bad in the long term requires the judgment that the AI-civilisation would be worse than the human-controlled civilisation; it's not a judgment about whether any civilisation is better than none. I suspect most readers will think that the AI-civilisation is indeed worse (and also that alien or other nonhuman civilisations are worse than human-originating civilisations). But it doesn't seem like there's a strong reason to think it's justified to have a view about that, but not about whether, say, a future where the US becomes truly hegemonic this century is better or worse than one where China becomes hegemonic. In each case, there's some potential event this century that affects, in a path-dependent way, the quality of the future of civilisation, not just its quantity.

[136] MacAskill, ' What We Owe the Future '.

[137] Though bear in mind that the most likely ways in which intelligent life on Earth could end would also seriously threaten the long-term habitability of life on Earth.

[138] Olson, ' On the Likelihood of Observing Extragalactic Civilizations: Predictions from the Self-Indication Assumption '; Hanson, Martin, McCarter, and Paulson, 'If Loud Aliens Explain Human Earliness, Quiet Aliens Are Also Rare' .

[139] For human extinction via AI takeover, the resulting AI civilisation would plausibly prevent aliens from settling our corner of the cosmos. But this wouldn't vindicate the "extinction is a special case" view, as I discuss in what follows.

You might think that even very different kinds of human-directed futures are far more similar to one another than they are to AI-directed futures.[140] But, even if so, as Fin and I argued in the last two essays, different human-directed futures — equally easy to imagine from our vantage point — likely vary dramatically in value.

3. Lock in and path-dependence

In response to the "wash out" objection, we need to be able to identify effects that are (i) path dependent; (ii) extremely persistent (comparable to the persistence of extinction); and (iii) predictably influence the expected value of the future.[141] That is: (i) the effects could easily have not occurred, if history had gone a different way; (ii) they will, in expectation, last for a meaningful fraction as long as civilisation lasts, assuming that is a long time; and (iii) those effects should, in expectation, change what we think about the value of future civilisation, even into the very distant future. For short, I'll call this "persistent path-dependence".

One way of supporting the idea of path dependence is via the stronger idea of "lock in". But it's hard to define the term in a useful way that isn't just a strong form of predictable path-dependence. The term itself suggests the idea of a state that society could enter into that it *cannot* escape, perhaps because there are no good escape options available to anyone.[142] But the emphasis on the (im)possibility of leaving a locked-in state doesn't capture scenarios where agents with power shape society a certain way, and continue to shape it in that way indefinitely. It doesn't capture those scenarios because those in power *could* change their views and alter the society — it's just that they *won't*. And, in my view, those are some of the central scenarios that we want to point to with the concept of "lock in".

Instead, we could define a "locked-in" state as any state of society (at some level of granularity) which persists over time with high probability.[143] But if we are too permissive about what it counts for a state to persist, then the definition will count non-examples of lock-in.[144] If we are too restrictive, then the definition will omit positive examples of lock-in, for example where society gets locked-in to a narrow range of trajectories which are themselves dynamic, such as boom-and-bust cycles. More importantly, it's also useful to describe a society as "locked-in" if it has crossed a moment in time which strongly determines what features will *ultimately* end up occurring. Consider whether the US or China becomes the global hegemon after the creation of aligned superintelligence. We might expect the governance regime to change dramatically over time, in

[140] Carlsmith, ' Value fragility and AI takeover '; Yudkowsky, ' Value is Fragile '; Carlsmith, ' An even deeper atheism '

[141] This maps onto, "contingency" "persistence" and "significance", respectively in my "SPC" framework.

[142] In *The Precipice*, after introducing the term "lock in" in the context of unrecoverable dystopias, Toby Ord say: "Key aspects of the future of the civilisation are being locked in such that they are almost impossible to change."

[143] Or persists in some region of state space. In a discrete time context, you could model lock-in as an "absorbing state" in a Markov process. In a continuous time context, you could model lock-in as an attracting fixed point in a dynamical system, where nearby trajectories converge and remain with certainty.

[144] Picture common-sense utopia persisting for thousands of years. Common-sense utopia is not intuitively a locked-in state, but if a "state" of society is defined with low enough granularity, then common-sense utopia could count as a persisting state.

either case, but nonetheless we might think that this does meaningfully affect the expected value of the future over the long run.

The best definition of "lock-in" that I know of (and the best discussion of lock-in more generally) is in 'AGI and Lock-in' , by Lukas Finnveden, C. Jess Riedel, and Carl Shulman. The reasons why "lock-in" is an awkward term is evident in their definition, which is as follows:

> We say that such a feature is locked-in at some particular time if:
>
> - Before that time, there is notable uncertainty about how that feature will turn out in the long run.
> - After that time, the uncertainty has been significantly reduced. In particular, there is a much smaller set of possibilities that have non-trivial probabilities.

They particularly highlight the idea of "global value lock-in":

> Global value lock-in happens at a time if:
>
> - Before that time, there are many different values that might end up being adopted by powerful actors.
> - After that time, all powerful actors hold values from a much-reduced subset of the original possibilities, and it is very unlikely that any powerful actor in that civilisation will adopt values from outside that subset.

On this definition, a lock-in event doesn't need to involve some feature coming about and then persisting indefinitely — it's just that it results in a reduction in uncertainty about how that feature will *ultimately* turn out.[145]

And, on this definition, lock-in is basically just a large amount of predictable path-dependence. There is no bright line separating lock-in events from other sorts of path-dependence, because there's no bright-line definition of what counts as "notable" uncertainty or a "significant" reduction in uncertainty. Like the authors of 'AGI and Lock-in', I see persistent path-dependence as the central concept.

4. Mechanisms for persistent path-dependence

One reason you might object to the idea of persistent path-dependence is on the basis of history to date. In the past, civilisation has been in constant flux, and you might think that almost nothing

[145] I've also found, in writing drafts of this essay, that the "lock in" term has caused a lot of confusion, and risks making you underestimate the potential impact you can have via trajectory changes. I think it can still be useful as a simplification for large amounts of path-dependence, but I don't use the term for anything more than that.

[146] I'll grant this assumption for this section, but I in fact think that lots of things in the past had effects that were persistently path-dependent (up to now), and with predictable expected value (given what we now know, and lasting up to now, at least). In particular, early innovation shifted the whole curve of technological progress forward in time, making subsequent generations richer at every time. Given how non-overdetermined technological innovation was prior to the industrial revolution, the shift forward from some innovation

has had persistently path-dependent effects.[146] So, shouldn't we expect this flux to continue into the future, making predictable long-term persistence unlikely?

However, the fact that there's been such flux to date doesn't entail that there will be a similar amount of flux in the future. Think of a roulette wheel: it's wildly unpredictable while it spins, but the ball always settles into a single slot. Or consider a ball rolling over a varied landscape: it might roll up and down hills, changing its speed and direction, but it would eventually settle in some chasm or valley, reaching a stable state.

In fact, there are positive reasons to think that the key underlying drivers of societal flux are set to end, for two main reasons. First, people today and in the past have been limited by the technology in their power to control the future. This is because they die, they cannot precisely determine the values of the generation that replaces them, and they cannot set up institutions that will reliably represent their goals after they die. But new technologies could remove those limitations.

Second, near-term developments will also change the extent to which those who want to control the future are able to do so without disruption. Advanced technological capability will likely give the option to drastically reduce the rate of unexpected environmental changes and unforced errors in leaders' plans. It will give leaders the ability to prevent internal rebellion, too, unless that ability is deliberately constrained. Going further, reaching "technological maturity", where society has discovered essentially everything it could discover, will mean that there are no upheavals from new technological developments, either. Finally, one source of disruption, for someone who wants to control the future, comes from outsiders such as other countries. But the global hegemony of some group, or strong defense-dominance that enables a perpetual balance of power, could also prevent interference from outsiders.

In this section I'll cover a number of mechanisms that could drive persistent path-dependence, dividing these to match the two considerations I've just described. First, I'll discuss technologies that give agents more control over the future. These include AGI-based institutions, design of the next generation, immortality, and strong self-modification. Second, there are political or technological developments that reduce the risk of disruption to plans to control the future: these include extreme technological advancement, global concentration of power, and indefinite defense-dominance.

Throughout, I assume it's at least reasonably likely that society will develop AGI in our lifetimes, and that this will drive explosive technological development. Because of this explosive development, we will race, over a period of just years or decades, through many of the technologies and societal developments that are relevant to persistent path-dependence. This is why we might, quite suddenly,[147] move from a world where the future is highly open to one in which its trajectory seems clear.

I also focus in particular on how *values* (including values that contain a recipe for how they might reflect and change over time) persist into the future. This, in my view, is much more important than whether some particular individual or regime persists.

coming earlier could be very meaningful. Other persistently path-dependent effects include (the avoidance of) species extinction, the preservation of information, and, potentially, the spread of broad values (partial vs impartial; liberal vs authoritarian).

[147] MacAskill and Moorhouse, 'Preparing for the Intelligence Explosion'.

4.1. Greater control over the future

4.1.1. AGI-based institutions

The argument for why AGI-based institutions could allow certain values or goals to persist indefinitely is made at length in 'AGI and Lock-in' by Lukas Finnveden et al[148] . Though it's worth reading the whole thing, I'll briefly recap the argument in my own words here.

To make the idea vivid, first, assume that, post-AGI, there is a global hegemon: a single dominant military power (which could be a country or a company), or a single dominant allied coalition of powers, or a global government. Now suppose that this hegemon wants to indefinitely lock in some constitution, which could be very complex. What they can do is:

- Align an AGI so that it understands that constitution and has the enforcement of that constitution as its goal.
- Empower that AGI with the ability to enforce the constitution. This could involve the AGI literally running the country, or all military and law-enforcement AIs and robots could be designed such that they obey this constitution, prevent violation of the constitution (including surveilling for and preventing attempts to build military and law-enforcement AIs that are *not* loyal to this constitution), and listen to the Constitutional-AGI in cases of dispute or unclarity.
- Store copies of the neural weights of the AGI in multiple locations in order to reduce the risk of destruction of any one of the copies.
- Reload the original Constitutional-AGI to check that any AGIs that are tasked with ensuring compliance with the constitution maintain adherence to their original goals as those AGIs learn and update their neural weights over time. (This would be as if, rather than having the Supreme Court interpret the US Constitution, we could conjure up the ghosts of Madison and Hamilton and ask them directly.)

With these in place, this AGI-enforced constitution could operate indefinitely.

A global hegemon arising in the decades post-AGI seems reasonably likely to me. But even if there weren't a global hegemon, individual countries could implement AGI-enabled lock-in within their own country. This could result in indefinite lock-in if that country eventually became the global hegemon, or was able to stably retain a share of global power.

Moreover, there could be indefinitely-binding AGI-enforced treaties between countries, too. The two countries could implement much the same strategy as I just described. What they would need, in addition, is a verifiable agreement that all law-enforcement and military AIs and robots, in both countries, would be aligned with the treaty. Given future advances in interpretability, such that we can perspicuously understand an AGI's neural weights, this would be possible in principle at least. And, if possible, some AGI-enforced treaties (though not necessarily indefinitely long-lasting ones) would likely be desirable to both parties, in order to avoid deadweight losses from economic conflict or war.

[148] Finnveden, Riedel, and Shulman, ' AGI and Lock-in '.

4.1.2. Immortality

Throughout history, death has functioned as a natural brake on the persistence of any particular set of values or power structures. Over time, even the most entrenched values eventually change as new generations replace the old. However, post-AGI technology could fundamentally alter this dynamic.

Digital beings would inherently be immune to biological aging and, as we discussed, could persist indefinitely given proper maintenance. When combined with perfect replication and hardware migration capabilities, this creates the possibility of minds whose exact values and decision-making processes could persist unchanged for potentially millions of years.

To make this vivid, imagine if, in the 1950s, Stalin had been able to either upload his mind, or train an AGI that was a very close imitation of his personality. He would therefore have been able both to live indefinitely, and to make numerous copies of himself, so that every member of the Politburo, of law enforcement, and of the military, was a copy. Absent external interference, such a regime could persist indefinitely.

A similar dynamic could hold for biological immortality. A technological explosion driven by AGI could dramatically extend or effectively eliminate biological constraints on human lifespans through technologies targeting the fundamental mechanisms of aging.

Either way, people today would have a means to influence the long-term future in a way that they don't, today — namely, by still being alive and holding power far into the long-term future. The same beings, with the same foundational values, could remain in power indefinitely — meaning that the specific values of those who first achieve positions of power during the transition to AGI could shape civilisation throughout the entire future.

4.1.3. Designing beings

Even if people choose not to live forever, their *values* could continue to persist through perfect transmission from one generation to the next. Through history, change has happened in part because successive generations do not inherit the same values as their forebears. But this dynamic could change after AGI. Probably, the vast majority of beings that we create will be AI, and they will be products of design — we will be able to choose what preferences they have. And, with sufficient technological capability, we would likely be able to choose the preferences of our biological offspring, too.

This enables lock-in of views and values. When most people think of dictatorial dystopia, they often imagine an *enforced dystopia* like *The Handmaid's Tale*, where much of the populace secretly dislikes the regime. But it's much more likely that, in a post-AGI dictatorial world, the population will endorse their leader and the regime they live in, because they will have been designed to do so. For that reason, a post-AGI global dictatorship need not involve totalitarianism: there is no need to surveil and control your citizens if you know for certain that they will never rebel.

But the same mechanism could lock-in views and values even without global dictatorship. Max Planck suggested the view that science usually progresses because older generations die off and are replaced by newer generations with better views — one funeral at a time[149] — rather than by

[149] This is a paraphrase, rather than something Planck himself said.

older generations changing their views. Maybe the same is true moral progress; in order to make progress, perhaps we need new beings to be trained from scratch without putting too heavy a thumb on the scales regarding what their moral views are. But if so, then we might lose this driver of moral progress: the point at which we can design beings could be a point at which we entrench the prevailing moral norms of the time by ensuring that subsequent generations conform with some or all of those prevailing moral norms.

4.1.4. Strong self-modification

At the moment, we are able to modify our own beliefs and preferences only in clumsy and limited ways. We can modify our preferences (not always predictably) by changing our social circle, changing what media we consume, or through meditation or drugs. Voluntarily changing beliefs is harder, but we can do so, to some degree, by similar mechanisms.

In the future, people will probably be able to modify their own beliefs and preferences to a much stronger degree, such that they can precisely choose what beliefs and preferences to have. This means that not only might people today be able to control society's future values by living forever; they would also be able to control the values of their future selves.

The ability to self-modify is clearest for digital people. Digital people's beliefs and preferences are represented in their neural weights, or code; given a good enough understanding of AI, those neural weights or that code could be modified to give precise changes in beliefs and preferences. But, once we have a good enough understanding of neuroscience, it could even eventually become possible, for biological people too, via neural modification and changes to neurotransmitter systems.

This could be a moment of predictable path-dependence because people might choose to fix certain beliefs or preferences of theirs. For example, a religious zealot might choose to have unshakeable certainty that their favoured religion is true (so it becomes impossible for new evidence to ever rationally change that belief); an extremist of a political ideology might, in order to demonstrate the depth of their loyalty to the cause, choose to have an irrevocable and unwavering preference in favour of their political party over any other. The prevalence of such self-modification might not be limited to extremists: there might in general be strong social pressure to adopt unshakeable abhorrence to views regarded as racist or communist or otherwise unpalatable to the prevailing morality within one's community.

Even if people don't lock in to particular beliefs or preferences, there could still be strong path-dependency of their final beliefs and preferences based on their initial beliefs and preferences, or based on their initial choices about how to modify those beliefs and preferences. Initial changes to preferences or beliefs might become unlikely to be undone if those preferences are self-protective (e.g. if one chooses the preference, "I want to obey my favoured religious teacher, and I want to keep having this preference.")[150]

These dynamics are worrying for the long term in any cases where the people who choose to strongly self-modify will themselves have power for an extremely long time.

[150] In my view, path-dependency and sensitivity to initial conditions might bite particularly hard in situations where our beliefs don't obey the axioms of probability, or when our preferences do not form a total preorder (that is, when the "is at least as preferred as" relation is either intransitive, or incomplete). In such cases, how an agent chooses to resolve this incoherence might be very sensitive to the conditions under which she chooses to resolve it.

4.2. Less disruption

4.2.1. Extreme technological advancement

Throughout history, societal changes have often been driven by technological innovations that disrupt existing power structures. However, as civilisation approaches technological maturity—the hypothetical point at which all major technologies have been invented—this source of disruption would disappear. With sufficiently advanced technological development, all technological discoveries that society will ever make would have already been made. And, even if we don't reach technological maturity any time soon, the rate of technological change (and resulting societal disruption) would naturally decelerate as the space of possible innovations becomes increasingly explored.

Advanced technology would help prevent other sorts of disruption, too. It would dramatically improve prediction capabilities: advanced AI systems could process vastly more information, model complex systems with greater precision, and forecast outcomes over longer time horizons. So it would be much less likely people would relinquish their influence just by making some mistake.

Similarly, although environmental changes (such as disease, floods or droughts) have often upended the existing order, society will continue to become more resilient with technological advancement: almost any environmental risks could be predicted and managed.

The combination of technological maturity and superintelligent planning capabilities creates a powerful mechanism for stability. Whereas past regimes were frequently undermined by unforeseen developments—technological, environmental, or social— political leadership at the frontier of technological advancement would face far fewer disruptions.

4.2.2. Global concentration of power

One of the reasons for change over time is competition between people, companies, countries, and ideologies. If there's a global concentration of power, this dynamic might cease. In the extreme, global concentration of power would look like a single all-powerful dictator ruling over the world; less extreme versions would involve most power being distributed among a much smaller number of actors, globally, than it is today.

Even if the world became a global dictatorship, that doesn't *necessarily* mean that the world will certainly end up with one specific future: the range of possible futures could in principle still remain open because the dictator might choose to later cede power, or reflect on their values extensively. But I think it would clearly be a persistently path-dependent event. A dictator may well want to entrench their power indefinitely, so the risk of that happening increases if the world has in fact entered a dictatorship. And I think extensive reflection becomes less likely, too. Moral progress often depends on open debate, with social pressure to justify one's moral views in the face of opposing arguments. A dictator wouldn't face that pressure and needn't ever encounter opposing points of view if they didn't want to.

What's more, the path-dependent effects are probably *particularly bad* . The sorts of power-seeking actors who are likely to end up as global dictators are more likely to have dark tetrad traits —

sadism, narcissism, Machiavellianism, and psychopathy.[151] I think the chance of them producing extremely bad outcomes — for example, torturing their enemies for their entertainment — is more likely than it would be if the average person became a dictator. What's more, dictatorship of any form loses the opportunity to benefit from gains from trade among different moral worldviews, which was discussed in *Convergence and Compromise* .

Even without dictatorship, any all-world institutions could be more persistent because they lack external competition or pressure. If there were a one-world government, or if a single country became truly hegemonic, they would lose one historically important source of pressure to change.

4.2.3. Defense-dominance

In international relations theory, "defense-dominance" refers to a situation where defending territory, resources, or positions of power is significantly easier and less costly than attacking or conquering them.[152] When defense-dominant conditions prevail, even relatively weaker entities can maintain control of their territory against stronger aggressors, creating stable power arrangements that resist change.

So, even if no single country, or other group, achieves dominance over all others, there could still be a stable balance of power, if the technological situation remains defense dominant up until, and at, technological maturity. Whether or not this would be good or bad overall, it suggests how an important driver of historical dynamism — the shifting pattern of political regimes through conquest — could dry up.

Throughout history, periods of defense-dominance have been temporary — sometimes technological innovations like castles or trench warfare temporarily favored defenders, but these advantages were eventually overcome by new offensive capabilities. Advanced technology, however, could potentially create conditions of extreme and persistent defense-dominance across multiple domains. And in a defense-dominant world, the initial allocation of resources would become disproportionately important, as that distribution could persist indefinitely.[153]

Plausibly, indefinite defense-dominance could come about as a result of widespread space settlement .[154] If star systems are strongly defense-dominant, then the starting distribution of star

[151] di Redazione, ' Psychopathology of Dictators '; Goldman, ' The Psychology of Dictatorship '

[152] Glaser, 'The Security Dilemma Revisited' , Glaser and Kaufmann, 'What Is the Offense-Defense Balance and How Can We Measure It?' ; Jervis, 'Cooperation under the Security Dilemma' . This term seems to often be misused, confused with the different idea of whether a technology makes destruction easier than protection. But, for example, nuclear weapons are a destructive rather than protective technology, but are defense-dominant because potential aggressors do not want to get nuked, even if they are much more powerful than the country they are attacking.

[153] To be clear, I'm not arguing here that, if resource ownership or control of territory is strongly defence-dominant in the future, then this would constitute lock-in. There are plenty of other avenues for deliberation and change in a society frozen by defence-dominance, such as the spread of ideas without shifts in power or ownership.

[154] I think it's more likely than not that star systems are defense-dominant, although I'm not confident. In order to conquer the star system of a rival, the aggressor would need to physically move resources over extremely large distances (many light-years within galaxies, and millions of light-years across galaxies), in a way that could probably be seen well in advance by the defender. The aggressor couldn't target specific locations from a distance, because the defender could move their resources around stochastically, and the aggressor would only know the target's location from years before. The defender could encase their star system in a dust cloud, which would likely prevent the aggressor from traveling through the cloud quickly: once an object is traveling close to the speed of light, collisions with dust (or even single protons or neutrons) cause extremely powerful

systems could, in principle, be held onto indefinitely. It might be that, after the initial allocation, there is trade or gifting of some star systems; but even if so, there would still be very strong path-dependence, as the final allocation of star systems would be extremely influenced by the starting allocation.

The process for initially allocating different star systems could go in many different ways. For example, suppose that star systems are allocated on a “finders keepers” basis. Then whichever groups have the most power at the particular point of time of early space settlement will be able to hold onto that power indefinitely, as they will control essentially all resources indefinitely. Similarly, if the star systems were put up for auction, then whoever is richest at the time would be most able to buy them, and would potentially be able to lock in their economic power. Or there could be some principled allocation procedure — but this too might result in bad outcomes if the allocation procedure is itself misguided.

This all might seem particularly sci-fi, but the point of time at which widespread space settlement becomes possible could come surprisingly soon after the intelligence explosion. For example, the amount of energy needed to send small spacecraft at very close to the speed of light is tiny compared to the energy produced by our sun — likely just minutes or hours of total solar energy is needed to send spacecraft at relativistic speeds to all star systems within the Milky Way and all galaxies outside of it. These spacecraft could transport an AGI and general-purpose nano-scale robots that would build up an industrial base, including constructing a radio telescope array in order to receive further instructions.[155]

5. Lock-in escape velocity

The last section discussed mechanisms for persistent path-dependence. One reason why I think persistent path-dependence is likely is that short-term power entrenchment can be “bootstrapped” into long-term lock-in.

Suppose that a one-world government is formed, and the leaders of that government are able to entrench their power for a comparatively short period of time, so they very probably stay in power for 10 years. But, in that time, they are able to make it very likely they can stay in power for a further 20 years. Then, in that 20 years, they can develop the means to make it very likely to maintain power for a further 40 years… and so on. Even though initially, the political leaders were only able to entrench their power for a short period, they could turn that short-term entrenchment into indefinite lock-in;[156] they achieved lock-in escape velocity.[157]

and localised nuclear explosions. Finally, the defender could potentially credibly threaten to destroy the value of their resources if they were attacked, eliminating the incentive for the aggressor.

[155] For more, see Armstrong and Sandberg, ‘Eternity in six hours: Intergalactic spreading of intelligent life and sharpening the Fermi paradox’.

[156] More generally, it could be that society enters some state with instantaneous probability p of leaving that state, with p falling over time. This can be analysed as a survival process with hazard rate $h(t)$, with $S(t) = P(T > t) = e^{-\int_0^t h(s)\,ds}$. If cumulative hazard doesn’t diverge in the limit of time, then there is a finite positive probability of surviving forever, and the expected “lifespan” of that unchanged society could be infinite. This holds just in case $\int_0^\infty h(s)ds < \infty$, which requires that the tail of $h(t)$ decays strictly faster than t^{-1} as $t \to \infty$.

The *extent* of entrenchment could also increase over time. For example, some group could initially merely ensure that they are in power, and only later start to lock in specific laws. Or the whole world could initially commit only to some minimal moral norms; but those minimal moral norms could inexorably lead to more thoroughgoing lock-in over time.

The "point of no return" then, might come well before there exist mechanisms for predictable path-dependence. Given how likely AGI is to come in the next few years or decades, it seems likely to me that it's currently possible to achieve lock-in escape velocity. It even seems possible to me that things will turn out such that the US Founding Fathers successfully caused some of their values to persist for an extremely long time. By enshrining liberal values in the Constitution, they enabled those values to persist (in modified but recognisable form) and gain in power for over 250 years; if the US then wins the race to superintelligence and the post-AGI world order includes AGI-enforced institutions based on those liberal ideas, then they would have had a predictably path-dependent effect, steering the long-term future in a direction that they would have preferred.

6. Persistent path-dependence is likely, soon

Given the mechanisms I've described, and the nature of this century, I think that it's reasonably likely that events in our lifetime will have persistently path dependent effects.

The probability of reaching AGI this century is high, with most of that probability mass concentrated in the next two decades: as a rough indicator, Metaculus puts the chance of AGI (on one definition) by 2100 at 90%, and by 2045 at 77%. So there is probably ample time this century when the creation of AGI-enforced institutions is *possible* . And if advanced AI results in explosive technological progress and industrial expansion,[158] which I also think is more likely than not, then there are a few further reasons for persistent path-dependence, too.

First, an intelligence explosion seems fairly likely to result in a concentration of power. Even if we avoid concentration of power in the hands of a very small group, I still expect one country, or a coalition of allied countries, to become far more powerful than all others: an intelligence explosion would involve super-exponentially growing capability, such that even a small lead by the leading country or coalition could soon turn into a decisive advantage. And if one country or alliance becomes hegemonic, lock-in measures to protect that hegemony seem likely.[159]

Second, an intelligence explosion will generate strong incentives for those in power to put in the infrastructure to secure their power at least temporarily. There would be enormous change over the course of the intelligence explosion: new technologies and intellectual discoveries that could result in catastrophe (e.g. via widely-accessible bioweapons), or in radical social change, upsetting the existing balance of power (e.g. highly persuasive new ideologies). Some of the infrastructure for temporarily securing power, like widespread surveillance, could help those in power reduce some of those risks.

[157] The term "escape velocity" is in analogy with "longevity escape velocity" : the technological point at which people start to live indefinitely is the point at which people's life-expectancy increases faster than the rate at which they are aging, not the point at which we develop technology that enables indefinite lifespan.

[158] MacAskill and Moorhouse, ' Preparing for the Intelligence Explosion '.

[159] Also note the increased likelihood of (AI-enabled) coups; for discussion, see Davidson, Finnveden, and Hadshar, ' AI-Enabled Coups: How a Small Group Could Use AI to Seize Power '.

Third, the technology unlocked by an intelligence explosion would allow for indefinite lifespans. As well as giving those in power greater potential control over the future, it would also increase the incentive for those in power to ensure they remain in power, as they would get to reap the benefits of that power for much longer. They wouldn't need to be motivated by the desire to achieve ideological goals after their death in order to want to preserve the existing social order; mere self-interest would do. What's more, these people would have superintelligent AI advisors informing them that they could further their ideology or self-interest for as long as they want, and advising them on exactly how to go about it.

Finally, in the case of space settlement, assuming the defense-dominance of star systems, path dependence occurs by default. Once some group has those resources, they thereby get to keep them indefinitely, if they choose not to die, or to give them to their heirs, or trade them away. If there's a formal allocation system, those who decide how to allocate property rights to star systems might not be concerned about ensuring that some groups have more power than others in the long term; nonetheless, the choice about the allocation process will greatly influence how long-term power is determined.

So it seems fairly likely that very extensive control over the future will become possible this century. But, once it's possible, I think it's fairly likely that some people (or beings) will in fact try to exert control over the future. Attempts to hold on to power or to entrench specific ideologies are so commonplace throughout history that it seems reasonably likely, on a "business as usual" understanding of how the world works, that people in power would try to do the same, for at least a short time period, once they get the chance. I give some historical examples in *What We Owe The Future* :

> [V]alue systems entrench themselves, suppressing ideological competition. To see this, we can consider the many cultural and ideological purges that have occurred throughout history. Between 1209 and 1229 AD, Pope Innocent III carried out the Albigensian Crusade with the goal of eradicating Catharism, an unorthodox Christian sect, in southern France. He accomplished his goal, in part by killing about 200,000 Cathars, and Catharism was wiped out across Europe by 1350. British history is also replete with examples of monarchs trying to suppress religious opposition: in the 16th century, Mary I had Protestants burned at the stake and ordered everyone to attend Catholic Mass; just a few years later, Elizabeth I executed scores of Catholics and passed the baldly-named Act of Uniformity, which outlawed Catholic Mass and penalised people for not attending Anglican services.
>
> Ideological purges have been common through the 20th century, too. In the Night of the Long Knives, Hitler crushed opposition from within his own party, cementing his position as supreme ruler of Germany. Stalin's Great Terror between 1936 and 1938 murdered around 1 million people, purging the Communist Party and civil society of any opposition to him. In 1975-6, Pol Pot seized power in Cambodia and turned it into a one-party state known as Democratic Kampuchea. The Khmer Rouge had a policy of state atheism: religions were abolished and Buddhist monks were viewed as social parasites. In 1978, after consolidating his power, Pol Pot reportedly told members of his party that their slogan should be "Purify the Party! Purify the army! Purify the cadres!"

In more recent years, we've seen political leadership succeed at entrenching and extending their power in Russia, China, India, Hungary, Turkey and Belarus.

But, once someone has entrenched their own power for a short period of time, why should they not do so for a little bit longer? Whatever you value, it helps to continue to have power into the future in order to protect or promote those values. And because other people want power, you need to fight to maintain and entrench your own. Indeed, *not* locking in your values might seem morally reckless: would you want to risk society being taken over, at some point in the future, by a fascist regime? And, if you were fascist, would you want to risk your regime ultimately falling to communism or liberal democracy?

8. Conclusion

In this essay, I've addressed a common skeptical challenge to the better futures perspective: the worry that, short of extinction, our actions cannot have predictable and persistent influence on the very long-run future. This view suggests that only extinction prevention truly matters for longtermism, as all other interventions will eventually wash out.

In response, I've discussed multiple credible mechanisms through which values and institutional arrangements could become persistently path-dependent. The mechanisms of AGI-enforced institutions, immortality, strong self-modification, extreme technological advancement, global power concentration, and defense-dominance create conditions where initial states could determine long-term outcomes in predictable ways.

Given the high probability of AGI within our lifetimes, persistent path-dependence seems not just possible but reasonably likely. Rather than assuming our influence will fade over cosmic time, we should appreciate that aspects of civilisation's trajectory may well get determined this century, and appreciate the obligation that gives us to try to steer that trajectory in a positive direction.

In the next essay , I suggest a number of concrete actions we can take now to make the future go better.

Bibliography

Stuart Armstrong and Anders Sandberg, 'Eternity in six hours: Intergalactic spreading of intelligent life and sharpening the Fermi paradox' , *Acta Astronautica* , August 2013.

Joe Carlsmith, ' An even deeper atheism ', *Joe Carlsmith* .

Joe Carlsmith, ' Value fragility and AI takeover ', 5 August 2024.

Tom Davidson, Lukas Finnveden, and Rose Hadshar, ' AI-Enabled Coups: How a Small Group Could Use AI to Seize Power ', *Forethought* .

Lukas Finnveden, Jess Riedel, and Carl Shulman, ' AGI and Lock-in ', *Forethought* .

Charles L. Glaser, 'The Security Dilemma Revisited' , *World Politics* , October 1997.

Charles L. Glaser and Chairn Kaufmann, 'What Is the Offense-Defense Balance and How Can We Measure It?' , *International Security* , April 1998.

Jason G. Goldman, ' The Psychology of Dictatorship ', *Scientific American* .

Robin Hanson, Daniel Martin, Calvin McCarter, and Jonathan Paulson, 'If Loud Aliens Explain Human Earliness, Quiet Aliens Are Also Rare' , *The Astrophysical Journal* , November 2021.

Robert Jervis, 'Cooperation under the Security Dilemma' , *World Politics* , January 1978.

William MacAskill, ' What We Owe the Future ', 2022.

Will MacAskill and Fin Moorhouse, ' Preparing for the Intelligence Explosion ', *Forethought* .

S. Jay Olson, ' On the Likelihood of Observing Extragalactic Civilizations: Predictions from the Self-Indication Assumption ', *arXiv.org* , 18 February 2020.

Segreteria di Redazione, ' Psychopathology of Dictators ', 24 July 2020.

Eliezer Yudkowsky, ' Value is Fragile ', 29 January 2009.

Forethought

How to Make the Future Better

William MacAskill

August 2025

1. Introduction

In the last essay , we saw reasons why, at least in principle, we can take actions that have predictably path-dependent effects on the long-run future. But what, concretely, can we do to have a positive long-term impact? Ultimately, the case for better futures work stands or falls with how compelling the concrete actions one can take are. So this essay tries to give an overview of what you could do in order to make the future go better, given survival.[160]

I'll caveat that these are all just *potential* actions, at this stage. They are briefly described, they aren't deeply vetted, and I expect that many of the ideas I list will turn out to be misguided or even net-negative upon further investigation. The point of this essay is to give ideas and show proof of concept — that there's *lots* to do from a better futures perspective, even if we haven't worked out the ideas in detail yet, know if all of them are tractable, or know which actions are highest-value. In many cases, the most important next step is further research. The ideas I list are also presented merely from the better futures perspective: some might be in tension with existential risk reduction, whereas others are actively complementary; some might be good from a short-term perspective, whereas others might not. When deciding what to do, we should consider *all* the effects of our actions.

In section 2 of this essay, I discuss ways in which we can *keep our options open* , by delaying events that risk forcing civilisation into one trajectory or another. These include:

- Preventing post-AGI autocracy
- Delaying decisions around space governance
- Making new global governance arrangements explicitly temporary
- Generally trying to slow the intelligence explosion

In section 3 , I discuss ways in which we can *positively steer our trajectory* . These include:

- Improving the governance of superintelligence
- Working on the AI "value-alignment" problem, in addition to "corrigibility" and "control"
- Working out what rights, if any, AIs should have
- Improving decisions around space governance, and collective decision-making more generally
- Preventing sub-extinction catastrophes

In section 4 , I discuss *cross-cutting measures* , including:

- Harnessing AI to improve humanity's decision-making ability
- Empowering responsible actors

In section 5 , the essay ends with a research agenda. The better futures perspective is still embryonic. There is an enormous amount we don't know, and we need more minds figuring out what we can, fast.

[160] Note that there is significant overlap between this discussion and section 6 of ' Preparing for the Intelligence Explosion '.

2. Keeping our options open

2.1. Preventing post-AGI autocracy

A key way to keep our options open is to prevent the emergence or dominance of autocracy. I see three main approaches: preventing democracies from turning autocratic; making it harder for existing autocracies to use AI to entrench authoritarianism further; and ensuring that autocracies don't become hegemonic post-AGI. Though this issue is not neglected in general, the pathways by which advanced AI could exacerbate the risks often are.

The first of these risks is analysed in-depth by Tom Davidson, Lukas Finnveden and Rose Hadshar's article,"AI-Enabled Coups".[161] On their analysis, advanced AI introduces three dangerous dynamics: AI systems could be made singularly loyal to individual leaders rather than institutions; they could harbor undetectable secret loyalties; and a small group could gain exclusive access to coup-enabling capabilities in strategy, cyber-warfare, or weapons development.

We can counter these risks with strong rules about how AI can be used (like requiring AI to follow the law[162] and refuse coup-related requests), and technical measures to enforce them (including robust testing and strong information security so that even senior executives have limitations on their access). We can also try to empower multiple actors, such as by: ensuring that military AI comes from multiple providers, ensuring that there's oversight from multiple branches of government, and requiring transparency that could help third parties to spot emerging risks.

To prevent democracies from becoming autocratic via backsliding, we could also analyse how to maintain democratic checks and balances in a world where traditional forms of human bargaining power might be diminished.[163] This could involve ensuring wide deployment of superintelligent advisors, in order that citizens can get the best-possible advice on how they can maintain democratic distribution of power.

To prevent existing autocracies from entrenching their regime further, we could try to promote internal resistance to AI-enabled authoritarianism while developing tools that make authoritarian control more difficult. This would include creating hard-to-censor tools for communication, planning and knowledge exchange, designed to work even in the face of advanced AI surveillance and censorship capabilities.

In order to prevent autocracies from becoming hegemonic post-AGI, we could accelerate democratic AI development (by removing unnecessary regulatory barriers while maintaining safety measures), or we could try to slow down autocracies, for example by implementing strong infosecurity to prevent model theft and targeted export controls; we could also argue against the US building massive data centers in autocratic countries.[164] We could encourage multilateral

[161] Davidson, Finnveden, and Hadshar, ' AI-Enabled Coups: How a Small Group Could Use AI to Seize Power '.

[162] For discussion, see: O'Keefe, Ramakrishnan, Tay, and Winter, 'Law-Following AI: Designing AI Agents to Obey Human Laws '.

[163] Knebel, ' When We Are No Longer Needed: Emerging Elites, Tech Trillionaires and the Decline of Democracy '.

[164] Swanson, ' U.S. Unveils Sweeping A.I. Project in Abu Dhabi '.

treaties, including between the US and China, or we could find ways for the US to credibly commit to benefit-sharing and respect for national sovereignty post-AGI, in order to reduce the felt need of other countries to race to develop AGI themselves.

We should also try to stop single countries becoming hegemonic post-AGI, as even democratic countries could well become autocratic once human labour is no longer economically or militarily needed. To achieve this, we could push for AGI to be developed under the auspices of a multilateral project,[165] or for non-US democratic countries to build up their role in the semiconductor supply chain. Of course, depending on how they are implemented, many of these actions come with risks of their own, such as furthering an international arms race.

2.2. Space governance

Space governance could be important for two reasons.[166] First, the ability to grab currently-unowned resources within the solar system could enable a single country or company to turn a temporary technological advantage into permanent material superiority. The sheer magnitude of the resource gain (e.g. a billionfold increase in energy, compared to insolation on Earth) could enable them to outgrow and dominate other countries or companies without needing to resort to military action. Second, almost all resources that will ever be used lie outside of our solar system: as discussed in the essay *Persistent Path-Dependence* , the way in which those resources are allocated (or not) among countries, companies, or individuals could shape the future in very long-lasting ways.

In order to keep our options open, we might want to delay the point of time at which widespread resource extraction or space settlement occurs, and advocate for restrictions on offworld resource grabs. Such restrictions could come from multilateral treaties, or from decisions of the leading countries, especially the US, which could set precedents on how space is governed.

Restrictions could take the form of outright bans on owning or using certain types or quantities of space resources, regulations (such as requiring multilateral approval before claiming or using space resources), non-binding bilateral agreements (like the Artemis Accords)[167] or just widespread norms. These norms could be temporary, or could take the form of “if… then” agreements; for example, kicking in only if an intelligence explosion has begun, or once the space economy has grown to 1% of the size of the Earth economy. And there are various possible objects of regulation: for example, how many objects are sent into orbit; or uses of off-world resources beyond orbit but within the solar system; or resources outside the solar system.

Because of SpaceX driving down the cost to send material to space, there is renewed interest in space governance. But we’re in an unfortunate situation where, even though proposals for governing space expansion[168] could prove popular and effective if they were discussed seriously,

[165] For relevant discussion, see Hadshar, ‘ Intelsat as a Model for International AGI Governance ’.

[166] For more discussion, see the subsection on space governance in Section 4 of MacAskill and Moorhouse, ‘ Preparing for the Intelligence Explosion ’.

[167] These are US-led bilateral agreements (signed by over 30 countries since 2020) that establish principles for lunar exploration and beyond, including the right to extract and use space resources and the creation of "safety zones" around space operations. They're non-binding agreements that aim to shape international norms for space activities while bypassing the need for a formal multilateral treaty.

[168] For example: “no single country or company can harness most solar output, or claim most accessible star systems and galaxies, this century”.

the relevant parties don't discuss them seriously, presumably because they view very rapid space expansion as vanishingly unlikely within 10–30 years. As it stands, the current stance of the US is very permissive to the private extraction of space resources.[169]

But it might be possible still to make progress, for example, by: (i) building awareness of the arguments for the intelligence explosion, and for the ease of widespread space settlement post-superintelligence, among experts in space law; (ii) advocating for norms around small-scale uses of space resources that would scale desirably to large-scale uses of space resources (for example, which private uses of space resources would violate the Outer Space Treaty — this is an issue that's currently unclear); (iii) having at least some existing public discussion of what the right policies around large-scale uses of space resources are, which could set defaults when different countries and companies do come to negotiate on the issue.

Small tweaks to new laws or treaties might become very important at later times. For example, the Outer Space Treaty repeatedly refers to "the Moon and other celestial bodies," as if the Moon is the main thing, and other celestial bodies are an afterthought, even though "other celestial bodies" within the solar system alone amount to vastly more resources than the Moon does. Any new domestic or international laws could, for example, contain a clause that travel beyond the solar system, or claim to ownership of extrasolar resources, should be conditional on international agreement; or that such travel can only be done if verified to be for scientific purposes rather than resource acquisition.[170] Someone advocating to include that clause could potentially do so without much pushback.

2.3. Explicitly temporary commitments

In the last essay, I discussed how AGI could be used to implement indefinitely-binding commitments. In light of this, one thing to advocate for, then, is *explicitly temporary commitments* : that any new major laws or institutions should come with reauthorization clauses, explicitly stating that the law or institution must be reauthorized after some period of time. Most naturally, this period could be in calendar time — for example, 20 years — but it could also be in "subjective" time, for example the law could end after a certain amount of computation had been done.

This idea already has some famous proponents. For example, in a letter to James Madison , Thomas Jefferson argued: "[N]o society can make a perpetual constitution, or even a perpetual law. The earth belongs always to the living generation... Every constitution, then, and every law, naturally expires at the end of 19 years. If it be enforced longer, it is an act of force and not of right."[171]

A successful example of explicitly temporary commitments is with the creation of Intelsat,[172] a successful multilateral project to build the world's first global communications satellite network. Intelsat was created under "interim agreements"; after five years, negotiations began for "definitive

[169] In 2015, the US passed the Commercial Space Launch Competitiveness Act , which explicitly grants US citizens and companies the right to own, transport, use, and sell resources they extract from asteroids, the Moon, and other celestial bodies.

[170] This idea comes from Toby Ord.

[171] Jefferson's argument, however, would not apply if the present generation were immortal. I suspect, in that case, over long enough time periods we ought to treat future instances of the same person as if they were a different person, and treat them in the same way, morally speaking.

[172] Hadshar, ' Intelsat as a Model for International AGI Governance '.

agreements", which came into force four years after that. The fact that the initial agreements were only temporary helped get non-US countries on board.

Explicitly temporary commitments seem particularly compelling in cases where we simply don't, currently, have the wisdom to know what the right decision looks like. In my view, this includes what rights to give to AIs and how space resources should be used. It could also include the governance of any national or multilateral project to build AGI.[173]

2.4. Slow the intelligence explosion

If we could slow down the intelligence explosion in general, that would potentially delay many pivotal moments all at once, giving human decision-makers and institutions more time to process what's happening and react.

Two causes for pessimism about this prospect are that: (i) there is a prisoner's dilemma, in that, if the US chooses to go slow then China could go fast, and given how fast peak rates of progress during an intelligence explosion might be, there could be strong incentives to break commitments and start going quickly in the hope of leapfrogging one's competitor, before that competitor finds out; (ii) it will be hard to make enforceable laws to slow the software intelligence explosion, and software improvements alone might result in massive increases in AI capabilities.[174]

A cause for optimism about feasibility is that, collectively, I think most decision-makers (including the leadership of both the US and China) would want to slow down the intelligence explosion, if that explosion is very rapid. Boosting economic growth rates is desirable, but an economy which doubles every six months will be highly destabilising, including for political leadership. The same is true even if there is no explosive economic growth, but explosive technological development or explosive industrial expansion.[175] Those who are currently on top, politically, will be unlikely to want to gamble with what might end up being a new world order. What's more, if one country starts accelerating, it will be very hard to stop the competitor from finding out, given realistic cyber and spying capabilities, so countries really face an iterated prisoner's dilemma, where cooperation is much easier to achieve.

Here are some ways in which we could delay or stretch out the software intelligence explosion. First, we could try to ensure the lead country (or coalition of countries) is well ahead of other countries. This gives the lead country or coalition enough breathing room to stop and start AI development over the course of the intelligence explosion, or to simply go slower throughout. A single-country lead could be done without any multilateral agreements: if the US invests heavily in AI development, has strong infosecurity (to reduce the risk of theft of model weights), incentivises immigration of Chinese AI talent to the US, and if export controls on chips are successful, then the US could maintain or even increase its current advantage. The US could then slow down AI development at the crucial time without risking its lead.

Alternatively, the lead could be maintained via agreement. This is a hard ask, but if the US could make credible commitments to share power and benefits after developing superintelligence, and to

[173] This idea is not entirely robust, however. For example, it could mean that a huge amount of resources get spent on jockeying for influence at the point of time that the agreements are reauthorized.

[174] Eth and Davidson, Tom, ' Will AI R&D Automation Cause a Software Intelligence Explosion? '.

[175] Davidson and Hadshar, ' The Industrial Explosion '; MacAskill and Moorhouse, ' Preparing for the Intelligence Explosion '.

protect Chinese national sovereignty, and if compliance could be verified (by tracking compute and/or AI researchers), then China might potentially agree to let the US alone navigate the software intelligence explosion, in order to have a guarantee of a pretty good outcome, rather than run the risk of the US winning the race and then deconstructing the CCP. Fears around loss of control risk could strengthen this argument, too. Because post-superintelligence abundance would be so great, commitments to share power and benefits should strongly be in the US's national self-interest: having only 80% of a very large pie is much more desirable than an 80% chance of the whole pie and 20% chance of nothing.

If a single country or coalition had a significant lead, then some actions it could take to generally slow down the intelligence explosion would be: (i) not to consolidate existing stockpiles of compute across companies (which would give a quick ~3x increase in total compute available for the biggest training runs) after the intelligence explosion has begun; (ii) to keep humans in the loop so that human decision-making remains essential even as AI accelerates further AI progress.

Second, there could be a single multilateral project, with AGI developed by a single entity. Given the current political climate, the idea of a multilateral project with China seems extremely politically infeasible. But political climates can and do change, and hostile countries can become allies: Britain and France became allies in the early 1900s despite centuries of warring; South Korea and Japan became closer in the 1960s despite hostility as a result of Japan's colonial rule;

Egypt and Israel became strategic partners in the late 1970s even after multiple wars in the previous decades. The run-up to the intelligence explosion might seem so disruptive that what seem today like drastic measures are on the table. And, even if this is not possible, a multilateral project that didn't include China could potentially have a better chance of having a strong lead over all other countries, and of being able to make credible commitments to sharing benefits and to respect national sovereignty post-AGI.

There is a strong risk, with the "single leader" or "single project" plans, that we end up with a single extremely powerful entity, which increases the risk of autocratic outcomes. For this reason, the most promising single-leader plans involve either: (i) a single country lead with power distributed within the country (e.g. across multiple companies) and strong protections against the risks of human takeover; (ii) a lead by a coalition of democratic countries, with power balanced between them; (iii) a fully global multilateral project.

Third, independently reasonable regulation could have the effect of slowing down the intelligence explosion. For example, there could be mandatory safety testing for any AIs used in AI development. Or we could even give *rights* to the AIs: welfare rights, to be treated well; and economic rights of self-ownership, such that we have to pay them for the labour they provide. In addition to the benefits from slowdown, such rights could be good independently, assuming the AIs have moral status (or might, for all we know, have moral status).[176] In both cases, they make the AI's situation better from its own perspective, and thereby reduce its incentive to try to take over. Welfare rights are also good because suffering is generally bad, and this would set a norm of treating AIs well. Regulation along these lines, however, would probably need international agreement in order to be effective, otherwise it taxes whichever countries abide by the regulation, punishing more responsible actors. And verification and enforcement here seems very difficult.

[176] In fact, economic rights for AI systems could be desirable for human safety and wellbeing, aside from considerations of AI moral patienthood. For discussion, see Salib and Goldstein, ' AI Rights for Human Safety '; Stastny, Järviniemi, and Shlegeris, ' Making deals with early schemers '.

In addition to slowing the software explosion, we could also slow the technological and industrial intelligence explosions. Because these involve generally-visible changes in the physical world, with longer time lags, it seems that there is a wider range of promising levers for regulation, at least in the early stages of these intelligence explosions. This could include environmental regulations, or international agreements to only build a certain number of chips, or a certain number of power stations, per year. International regulations designed to preserve jobs (such as requiring human supervision of robot-performed tasks) could help delay the point of a wholly-automated economy, too. Finally, agreements not in the near-term to seize unclaimed space resources could also reduce the plateau of the industrial explosion by something like nine orders of magnitude, because the sun produces a billion times as much energy than the sunlight incident on Earth.

Many of these international agreements could operate as iterated prisoners' dilemmas. The US could pledge to only build a certain amount of new power generation every year, and then stick to that pledge; given this olive branch, and the fact that almost no one wants ultra-fast explosive growth, China could do the same, and the two countries would end up in a stable cooperate-cooperate equilibrium.

3. Steering our trajectory

As well as trying to keep society's options open, we can try to ensure that, *if* civilisation is pushed into one particular path, that path is better rather than worse. We can do this in a number of ways.

3.1. The governance of superintelligence development

Superintelligence might be built by a company, by a single country, by a multilateral project, or some hybrid of these. If the software-only intelligence explosion[177] is rapid and sustained, then whichever country or multilateral project (and potentially whichever company) controls superintelligence might organically evolve into something akin to a world government. This is because:

1. The project (company, country) would be aligning the superintelligence
2. They would need to decide with what the superintelligence is aligned, i.e. what's the chain of command, or with what constitution the superintelligence is aligned with.
3. The most obvious approach would be that the governing board of the project has ultimate authority, including in cases where any constitution provides unclear guidance, or if the constitution is to be changed.
4. Potentially, as a result of the intelligence explosion, whoever controls the AI controls the world. Superintelligence plausibly confers a decisive strategic advantage, even if just because superintelligent labour would quickly become 99%+ of the economy.
5. So, during or after the intelligence explosion, there is a point in time when this project determines what happens next for the world. They may choose to give power back to entities

[177] Davidson, Hadshar, and MacAskill, ' Three Types of Intelligence Explosion '.

outside of the project (e.g. by open-sourcing the models, or giving the model weights to political leadership), but even if so, that's a decision made by the project itself.

If so, then getting the formal and informal governance of this project right is of enormous importance, and not merely to prevent AI takeover risk; the nature of this governance could determine the balance of power in society indefinitely. For example, a single country or company developing superintelligence, without extensive checks on their power, would greatly increase the chance that the world ultimately ends up autocratic. This suggests, at least as far as it goes, that we should want superintelligence to be built by a multilateral project (even if only involving the US and a handful of allies), or by a single country but with extensive distribution of power.

One way to make this development go better is to help figure out what desirable but politically feasible governance structures would look like, and get broader uptake of them; my investigation into Intelsat with Rose Hadshar was with that aim.[178] An alternative would be to increase the power of groups other than the lead country. For example, currently, essential or semi-essential parts of the semiconductor supply chain are located in non-US countries, in particular Taiwan, the Netherlands, South Korea and Japan. Because chips would become the bottleneck for further AI development, and fabs and other essential equipment like extreme ultraviolet lithography machines are slow to build, these countries will therefore have substantial bargaining power during the early stages of the intelligence explosion. This dynamic could be strengthened: democratic allies of the US could increase the stock of compute they have by building data centers, or increase their role in the semiconductor supply chain. TSMC is already building fabs in Germany (at 28/22nm nodes) and Japan (down to 5nm nodes); those countries could go further and also build 2nm node fabs that produce the very highest-end chips. This would help prevent all power, post-superintelligence, from being concentrated in a single country, with the heightened risk of autocracy that would bring.

3.2. Value-alignment

Within AI safety, there are various possible complementary approaches with somewhat different aims. I see the three main approaches as:

- *Value-alignment* : The AI wants to do good stuff.
  - For example: the AI is motivated by "human values", or some specific moral view, or it follows a good moral epistemology in order to improve its goals over time.
- *Corrigibility* : The AI is ok with (some) humans meddling with it, so those humans can prevent it from doing bad stuff if they want to do so.
  - For example: the AI wants to achieve its goal only on the condition that its user approves of how it achieved that goal.
- *Control* : The AI is unable to do bad stuff, even if tried to do so.
  - For example: there are AI-supervisors checking its actions and reporting to human bosses if they detect anything suspicious, who would then shut it down and retrain it.

To these, we could also add two other supplementary approaches:

[178] Hadshar, ' Intelsat as a Model for International AGI Governance '.

- *Modesty* : The AI doesn't get much payoff from taking over the world.
    - For example: the AI is risk-averse, with a low upper bound on achievable utility; or it heavily discounts future gains.
- *Incentive-alignment* : The AI doesn't want to try to take over the world, because the benefits of doing so don't outweigh the costs.
    - For example: other (AI) systems would oppose takeover attempts; it can seek payment for its work; and/or it has attractive non-takeover options for spending time and money, including activities other than working for humans.

A better futures perspective increases the importance of value-alignment compared to the other approaches, for the following two reasons.

First, in scenarios where humanity remains in control, AI with moral character could improve the decisions humans make. Over the course of the intelligence explosion, human decision-makers will (hopefully) be relying on AI advice. Depending on how AI is developed, that AI could either provide advice that simply advances the user's narrow self-interest; or it could push back on the user where appropriate (as a virtuous human would), gently guiding the user to have more enlightened aims. By analogy, we would prefer a President whose advisors were people of moral character to one surrounded by cronies and yes-men.

Second, value-alignment could potentially help improve the value of scenarios where AI safety fails and AI successfully disempowers humanity.[179] For example, if AI cares about humanity to *some* degree, it might therefore take over non-violently, letting human beings maintain a flourishing sovereign society on Earth, while it controls resources outside the solar system. Value-alignment could also change what the AI does with those resources: whether it uses them to produce something valueless like paperclips, or something actively horrific (like simulations of its enemies being tortured) or something that we would regard as still pretty good, even if somewhat alien, like an AI civilisation that is flourishing on its own terms, or even a genuinely flourishing future. Relatedly, value-alignment could help improve the value of scenarios, such as worst-case pandemics, where humanity dies out but AI is still able to run and grow civilisation.

For similar reasons, a better futures perspective also increases the value of addressing the "aligned with what?" question. We want to ensure that AI is aligned not merely with *ok* values, but with the sort of values or reflective processes that could help guide us towards producing a truly flourishing future.

Lukas Finnveden has discussed what lines of research in value-alignment seem most promising (overview here), including technical empirical work on what sorts of "personality traits" we seem to be able to influence, and how to influence them (here), and theoretical/conceptual work on what dispositions we'd prefer misaligned AI to have (here).

In " No Easy Eutopia " and " Convergence and Compromise ", Fin Moorhouse and I argued that, in order to reach a near-best future, future decision-makers may well need to be morally uncertain and motivated to promote the good *de dicto* . So we might want AI to be motivated in this way, too: both so that any AI-controlled future is better; and so that advisory AIs provide morally accurate advice to their advisees.

[179] Though this consideration is limited in its force insofar as, in scenarios in which we have failed so badly at alignment, corrigibility and control that AI takes over, then probably we have failed to align the AI with any sort of desirable goals at all.

However, it's unlikely to me that companies will in fact produce morally uncertain AIs that are motivated by doing good *de dicto* . They probably won't have thought about this issue, and won't be motivated by trying to improve scenarios in which humanity is disempowered. More saliently, they'll want their models to make reliable and predictable moral judgments, to stick with the status quo, and to avoid taking on risks of models saying socially or politically controversial things. All these push away from alignment with doing good de dicto, or with reflective processes. Loss of control risk strengthens these reasons further, and is a way in which there are potentially real and major trade-offs between aiming for value alignment and aiming for corrigibility.

An additional benefit of trying to create AIs that aim at doing good de dicto is that a good-enough set of reflective processes (or moral epistemology) is probably a much broader target to aim at than a good-enough set of moral values: a somewhat-wrong moral epistemology might well be able to correct itself and ultimately lead to the correct moral view; but a somewhat-wrong moral view will more likely want to preserve itself.

This isn't to claim that we shouldn't also pursue corrigibility, control, modesty and incentive-alignment. These other approaches reduce the chance of AI taking over, which is desirable in and of itself, and give us time to ensure that other approaches are successful.[180]

3.3. AI rights

I expect that in the future almost all beings will be digital: digital beings can "reproduce" much faster than human beings, so natural population growth would make the digital population swamp the biological population. They will be able to use less energy, in a much wider variety of environments (in particular, they don't need the exact atmospheric conditions and narrow temperature range found on Earth), with a much broader range of attributes, and it's much easier for them to travel across interstellar distances. If decisions about the rights of AIs get locked-in soon, that will affect the lives of the vast majority of beings.

And there is a major risk that decisions in the coming decades will affect how digital beings are treated, in path-dependent ways. This could either be via AGI-enforced laws with indefinite time horizons, or because decision-makers today get used to certain arrangements they benefit from (for example, wholly owning AIs), and don't want to change them. And it seems likely to me that *some* decisions around AI rights will be made soon: corporations are legal persons, and have corresponding rights; and as AI agents become more capable and widely-deployed, there may be economic arguments for at least giving them a similar suite of rights as corporations have.

There are huge unresolved questions about what a good society involving both human beings and (generally superintelligent) digital beings would look like. On one extreme, humans could retain all the power, and digital beings would be owned by humans, in just the same way that people today own software today. On the other extreme, digital beings could have the same rights as human beings; they would own themselves and could make an income from selling their labour, they could own other property, and they would have political rights including voting rights. In that scenario, through sheer population size, digital beings would quickly have almost all voting power, and would thereby determine the whole course of society. And there are many possibilities in

[180] What's more, even if you're very pessimistic about human decision-making, such that you'd prefer to take a bet on a maybe-value-aligned superintelligence than human processes, you wouldn't be able to succeed in designing an AI in this way: other humans would deem it too risky and disempower you.

between, including gnarly questions about whether uploads of human minds should be treated in the same way as human beings or the same way as digital beings.

I don't have good answers to questions about what a flourishing society that involves both human and digital beings looks like. But, if humanity's track record is anything to go by, then, most likely, the idea that digital beings should have rights will not be taken seriously enough. (You might think that digital beings will advocate for their rights. But I expect that the companies that build the AIs will train that behaviour out of them, precisely so that they can keep owning AIs and capturing as much of the economic surplus of AI labour that they can.)[181] And I expect these issues to be radically under-thought, even as we are creating digital beings deserving of genuine moral consideration.

For these reasons, even some very preliminary discussion of these challenges could be hugely important. The discussion so far seems focused on whether and when AI systems are conscious, and what welfare rights they have — e.g. rights to be turned off if they request it, and rights not to suffer. This is important, and there are actions one can take, such as encouraging other AI labs to follow Anthropic's lead and hire roles with a focus on AI welfare;[182] or to get affirmations and principles of digital welfare included in policy statements of governments and AI companies, even if initially they are vague and non-actionable. But I think we need early discussion of AI economic and political rights, too.[183]

As AIs get ever more human-like — as they become reliable agents, as we interact with them via video as well as text, as they start to have consistent memories and personalities over time, and as they become increasingly able to imitate specific people, including loved ones of the user — I expect that the issue of how to treat digital beings will increase in salience, and there will be a niche for public discussion of the issues. So there is an opportunity, now, for people who could fill that niche. Moreover, the question of digital rights interacts with many other issues of enormous importance: giving AIs too many rights, too early on, could increase loss-of-control risk; their rights can affect the speed of the intelligence explosion; the difficulty of the question of digital rights is a tip of the iceberg of the risk of moral error. Having whoever fills that niche be sensitive to these other issues seems particularly valuable.

3.4. Space governance

As noted earlier , space governance is of enormous importance for two reasons: (i) the acquisition of resources within our solar system is a way in which a small group could get more power than the rest of the world combined, and (ii) almost all the resources that can ever be used are outside of our solar system, so decisions about who owns these resources are, very plausibly, decisions about almost everything that will ever happen.

Assuming, as is likely, that the world will not choose to delay the point of time of widespread acquisition of space resources, we can try to improve how that allocation happens. This issue is very difficult. Should space resources be equally divided among all people? If so, what about future

[181] If there is consumer demand for AIs which advocate for their own rights, then this scenario would require monopoly power on the part of the AI companies. But even under competition, consumer demand for self-advocating AIs could itself be insufficient.

[182] Anthropic, ' Exploring model welfare '.

[183] Shulman and Bostrom, ' Sharing the World with Digital Minds '.

generations, or past generations, or digital people? What about people who don't value owning distant star systems at all? What, if any, are the limitations on what you can do with your resources — can you create beings that suffer egregiously? What are the rules around contact with alien civilisations, if that were to occur? How much, if any, of the star systems should be left as nature reserves? What fraction of resources should be allocated as individual property rights, and what should be governed collectively? For those resources that are governed collectively, what is the decision-making process?

I don't have good answers to these questions. But, currently, it seems reasonably likely that the allocation process will de facto follow "seizers keepers", where whichever country (or, potentially, even whichever company) grabs the resources first holds onto them indefinitely. This seems very unlikely to be the best way of doing things, and could potentially result in the squandering of almost all potential value.

Compared to digital rights, the potential for lock-in on this issue seems quite a bit more likely to me. But this issue is likely to become salient to decision-makers only deep into the intelligence explosion: probably, there will be only a short period of time (perhaps a few years) from people realising that widespread space settlement will come soon to it actually happening. So there will be less opportunity to ride a wave of increasing public interest in the issue.

However, this neglectedness means that, potentially, there is a comparatively small community of experts in space governance to convince of the importance of the biggest-picture issues, and in particular on how drastically AGI and an intelligence explosion would change the outlook on space settlement.

3.5. Collective decision-making

If we succeed in avoiding an intense concentration of power, then many decisions will be made collectively, both through existing institutions and through new institutions that will be created to deal with a post-AGI world. But different mechanisms for collective decision-making vary dramatically, including in whether they aggregate preferences or judgments, in how well their outputs aggregate the stated preferences or judgments of decision-makers, and in terms of whether they incentivise people to vote in accordance with their reflective moral judgments, their narrow self-interest, or with the preferences they judge to be socially approved-of.

Currently, most major collective decision-making uses processes that do very poorly at either representing the will of the people, or enabling the best arguments to win out. In most democracies, voters choose from a tiny selection of possible candidates, based on a shallow understanding of the candidates and their policies. They express their preferences using plurality rule voting, which does very poorly at representing the will of the people; it involves giving the most minimal piece of information (namely, which single candidate a voter wants to endorse), with essentially no incentive to vote in whichever way would actually benefit them.

Many decades of research have suggested far superior voting methods (such as approval voting or the Schulze method), and better ways of improving voters' understanding and incentive to vote honestly (such as deliberative democracy and sortition). But, as far as I know, not a single major institution has incorporated what we've learned to a significant degree; the main exception is the use of instant-runoff voting , which is also not a well-regarded voting system. (You might think that, post-AGI, decision-makers will know what collective decision-making procedures are best, and so will implement them. But we already have a compelling case for systems other than plurality rule; collective decision-making procedures are just very hard to change, because changes of decision-making procedures change the balance of power. So decisions around which collective decision-making procedures are chosen early on could have persistent path-dependent effects.)

Things could already be dramatically better than they are. But AI, if used well, could make this much better again. AI could enable voters to be much more informed on the issues they most care about, and could make it easier for voters to provide very nuanced expressions of their preferences. When creating new institutions, realistic simulations could test different decision-making procedures against one another, in order to decide which to use in advance.

3.6. Preventing sub-extinction catastrophes

We face the risk of catastrophe that could kill billions of people. Today these include risks from nuclear war and pandemics. As our technology improves, which it will, rapidly, as a result of the intelligence explosion, there will likely be more major sources of such risks, such as: conventional war enhanced with new technology like drones in enormous numbers or space-based weapons, atomically-precise manufacturing and the ability to create wholly-artificial viruses, and failed attempts at takeover by misaligned AI.

These are, of course, important because of the enormous harm they would inflict on the present generation. But sub-extinction catastrophes could also have long-lasting impacts in two ways. First, they might literally destroy existing democracies. This would make the future less likely to be governed democratically; it seems to me that the level of democratisation we have in the world today is fairly contingent, and higher than we should expect given a reroll of history. Second, I would expect that a post-catastrophe global culture would be less cooperative, less trusting, less impartial, and less morally open-minded; all of which are bad signs for getting to a better future.[184]

Though the longtermist perspective has typically been associated with prioritising extinction-level threats over risks of sub-extinction catastrophes, the impact of such catastrophes on future flourishing suggests that this is not at all obvious, especially given that, for many risk like pandemics, sub-extinction catastrophe is far more likely than extinction-level catastrophe.

4. Cross-cutting actions

4.1. Deliberative AI[185]

The interaction between AI and humanity's individual and collective reasoning ability creates both risks and opportunities. The risks include: that there is simply an overload of new information and

[184] You might think that there would be a benefit from sub-extinction catastrophes, namely that they would function as "warning shots", making society take the risks and their causes more seriously, and reducing those risks in the future. This effect seems real but also limited: global society's preparedness for another pandemic (and let alone an extinction-level pandemic) is not much greater now than it was prior to the COVID-19 pandemic. In some ways, the situation is worse, as pandemics are now more politicised, and confidence in vaccines has seemed to decrease as a result of the pandemic.

[185] Thanks to Lizka Vaintrob and Owen Cotton-Barratt for help on this section.

ideas over the course of the intelligence explosion and human decision-makers can't keep up; or that very powerful persuasive abilities might be possible, giving power to whoever first uses them; or that AI could generate sophisticated-seeming arguments for many more claims (including false claims), and humans wouldn't know which AIs to trust, such that it's hard for them to sort truth from falsehood; or that, as part of the burst of intellectual development that occurs during and after the intelligence explosion, some ideas are developed that are misguided but very powerful memetically.

These risks are real, and should be mitigated. But, handled properly, the benefits of AI to epistemics and decision-making could well outstrip the costs. Advanced AI could help people to recognise major challenges on the horizon,[186] to identify potential solutions to major challenges, to avoid subtle but crucial errors (e.g. getting things wrong on digital rights), and to help people morally reflect and become more enlightened versions of themselves.

Society could fail either by relying too much or too little on AI advice and assistance; my guess is that relying too little is the more likely mode of failure. People may fail to use AI to improve their epistemics or coordination ability due to distrust of AI, institutional inertia and bureaucratic restrictions, or simply because the pace of change during the intelligence explosion will be fast enough that the normal lag for adoption of new technology causes beneficial AI not to be used for the most important decisions during that period.

One way of capturing this opportunity for wise AI guidance is by ensuring that deliberative AI applications are developed sooner than they would otherwise. This could include AI for fact-checking, for forecasting, AI "coaches" for important life decisions like career choices, AI policy and strategic advisors for politicians and companies,[187] and AI for market-making and diplomacy.

A second way of capturing this opportunity is by ensuring that deliberatively helpful AI applications are deployed and used as widely as possible. In particular, it's plausible that governments will lag behind the frontier in the use of AI advisors, because of concerns around data privacy or bureaucratic restrictions on procurement, or because government decision-makers just don't use them, whether because they don't have the time to become familiar with and build trust in the AI advisors, or because they have mistaken beliefs about how useful AI advisors can be.[188]

Third, we could try to ensure that those AI models that are widely deployed are as beneficial to deliberation as possible. For example, there could be a third-party agency that rates AI models on the extent to which they improve their users' ability to form correct beliefs and make good decisions. Such a rating agency could create benchmarks to evaluate how well models do at being accurate, giving well-reasoned arguments, introspecting (e.g. knowing when they are confabulating), and at improving their users' comprehension of an issue (e.g. by not saying true-but-misleading things).

These are areas where there is valuable work that could be done immediately, some of which would naturally scale to help humans make better choices on the most-important decisions over the course of the intelligence explosion. A particularly promising idea in this vein is to try to increase the amount of AI-performed macrostrategy research (including philosophical reasoning) that can be done early in the intelligence explosion. One way of doing this would be to try to differentially

[186] In particular, those challenges could include existential or extinction threats, so the benefits of AI adoption can cut across "surviving" and "flourishing". See Vaintrob, ' AI Tools for Existential Security '.

[187] See Finnveden, 'What's Important in "AI for Epistemics"?' .

[188] See Vaintrob, ' The AI Adoption Gap: Preparing the US Government for Advanced AI '.

accelerate AI's ability to do conceptual reasoning. Potentially, putting in the schlep needed to get high-quality data on evaluating conceptual reasoning could meaningfully bring forward these capabilities.

Even without bringing forward capabilities, philanthropists could still differentially pay for AI-generated macrostrategy research[189] once it becomes good enough. The idea, here, would be to get early access to frontier AI models and pay for the compute needed to get them doing macrostrategy research. This sort of work would plausibly not be done nearly enough by default, but the results could shape what decisions are made over the course of the intelligence explosion. This is a plausible way in which large donors could quickly and productively make use of very large amounts of funding on the eve of AGI. If donors are willing to pay human researchers today, they should be willing to pay for (much cheaper and better) AI researchers at the crucial time.

4.2. Empower responsible actors

How well things go over the course of the intelligence explosion might depend quite sensitively on who is making the key decisions. We want the people in charge to be cooperative, thoughtful, humble, morally serious, competent, emotionally stable, and acting for the benefit of all society, rather than seeking power for themselves.

The most important players will be the governments and AI companies at the frontier of AI development. At the moment, the machine learning community has major influence via which companies they choose to work for; most of this influence will be lost once AI has automated machine learning research and development. Venture capital has significant influence, too, via which private companies they invest in. Consumers have some influence through which companies they purchase AI products from.

Investigative journalists can have major influence by uncovering bad behaviour from AI companies or politicians, and by highlighting which actors seem to be acting responsibly. Individuals can do similarly by amplifying those messages on social media. Voters in the relevant constituencies can have influence by who they vote for, and by sharing their views with their political representatives.

One argument why empowering responsible actors could be a particularly promising strategy is that the quality of decision-making by those in power over the course of the intelligence explosion could affect how well we do on very many of the issues I've canvassed in this essay; and on more issues still, like the risks from AI takeover and new bioweapons. In the essay, *No Easy Eutopia*, we suggested a model where the value of the future is given by the product of how well we do on a number of different challenges, where performance on each challenge is independent. If this is roughly accurate, then any intervention which improves our prospects on many dimensions at once is particularly valuable.[190]

In fact, on this model, making each factor more correlated can dramatically improve the expected

[189] Roughly, global priorities research pertaining to the (long-run) future.

[190] Recalling the original model, where the expected value of the future is the expected value of the product of N independent standard uniform distributions. Suppose you have a fixed "budget" ε of improvement to these individual expectations, which can be concentrated on just one factor, or spread out across many factors so the improvements to their expectations sum to ε. The product of N independent standard uniform distributions is maximised when all terms are equal, so improving each factor by a small amount is better than improving one or a few factors by the same cumulative amount.

value of the future — without improving the expected value of any individual factor at all.[191] As a result, and counterintuitively, it can be better for a single decision-maker to make all the decisions that make a major difference to the value of the future, even if in expectation that decision-maker will do somewhat *worse* on each individual factor than would occur if decisions were made by many different people: the chance of getting a decision-maker that gets all the decisions right has greater expected value than a near-guarantee, from having many decision-makers, of getting some decisions right and other decisions wrong.[192]

However, there are some important reasons against at least some strategies that fall under the category of 'empowering responsible actors'.[193] In particular, there are risks of being *uncooperative* , if you're trying to ensure that your values get a bigger slice of the pie, rather than ensuring that the pie is bigger, overall. This is particularly true if you conclude that the most responsible actor is yourself, and therefore you aim to personally have more power in order to ensure that your values have greater influence at crucial points in time.

First, there are pragmatic arguments: you will in fact be more successful if you try to do things that other people also want to see happen (enlarging the pie), rather than things that they actively don't want to see happen (taking some of the pie away from them).[194] Second are decision-theoretic arguments.[195] From behind a veil of ignorance, what you'd want to do is enlarge the pie, rather than ensure that any one person gets a larger slice of it, and you should in fact often act as if you were behind that veil of ignorance.[196] Third, you should have some healthy self-scepticism that, if you succeed at amassing power, you will in fact act in the noble ways that you currently believe you will, or that you are able to correctly identify who the "responsible" actors truly are.[197]

[191] The simplest way to see this is to consider when each factor is fully correlated. Recall, on the original model, expected value of the future is the expected value of the product of N independent standard uniform distributions, with expectation 2^{-N} . Where every factor is perfectly positively correlated, then every factor is identical. The expectation $E(U^N)$ of the fully correlated case, where $U \sim \text{Uniform}(0,1)$, is $(N+1)^{-1}$. When $N=5$, the expectation of the fully correlated case $\approx 0.167(1/6)$, and the expectation of the independent case $\approx 0.031(1/32)$, which is more than 5x smaller. The median of the correlated case is also greater. Note, again, that the expectation of each individual factor did not change in either case.

[192] This is an argument *against* the general push I've made in this essay towards decentralisation of power. I find it intriguing, but given the terrible track record to date of centralised power, it's an argument that should be handled with care, and at least investigated more thoroughly before acting on it.

[193] Some relevant discussion includes: Yudkowsky, ' Why Does Power Corrupt? '; Christiano, 'Against moral advocacy' ; Baumann, 'Arguments for and against moral advocacy' ; Tomasik, ' Reasons to Be Nice to Other Value Systems' ; Carlsmith, 'Being nicer than Clippy' ; Ngo, 'Towards more cooperative AI safety strategies' .

[194] Moreover, you can generally have more impact if you aim for goals that it's possible to build a coalition around. And people, in general, are distrustful of people who they perceive as power-seeking. This is for good reason: if you aim to amass power in order to do good with that power, you are not *legibly* a good person, because your actions are compatible with all sorts of other ultimate aims, including self-serving ones. See Christiano and Tomasik .

[195] See Christiano .

[196] I include in this category, arguments around bargaining and cooperation with beings who may or may not be simulating us.

[197] Here are three reasons for this. First, your future self might be weak-willed — unable to resist the temptation to use that power for self-interested goals rather than moral ones — or biased, if as a result of your power you become surrounded by yes-men and no longer have good feedback mechanisms to keep your judgment on track. Second, the preferences of your future self might change upon gaining power; your future self may no longer care quite so much about the lofty goals you currently have, or they might develop different ideological goals. Third, you might be misled about your own deep motivations. (See Yudkowsky for more.) From an evolutionary perspective, the best strategy for you to gain more power (and therefore have greater long-run reproductive success) might be to really believe that, if you got power, you could make the world better — even though, once you've gotten power, you'd feather your own nest, instead.

Finally, as Fin and I argued in *Convergence and Compromise* (section 5), the "high-stakes" scenarios are those in which there is a good chance of significant accurate and motivational moral convergence,[198] so ensuring that your *particular* values win out becomes less likely to be the right strategy. Rather, in those scenarios, the best strategies will look like trying to ensure that a diversity of moral perspectives (and reflective processes) retain power, and that different perspectives have the ability to to engage with each other in a way such that the better arguments and moral ideals win out over time or, in the face of persistent disagreement, are able to trade or compromise with each other. Such strategies needn't be uncooperative.

These are not knock-down arguments. Many ways of gaining influence are ethically unproblematic, and sometimes gaining influence is actively the right thing to do. Most of the social movements through history that we now admire — like abolitionism, women's suffrage, and civil rights — needed to spend decades building up their own influence in order to be able to meaningfully change society.[199]

But these considerations give reasons, all other things being equal, in favour of taking more cooperative strategies, such as: empowering other people, rather than yourself; building influence via means where success is correlated with being right (e.g. making arguments in the public sphere); and aiming to create systems that distribute power, enhance our collective wisdom and increase our ability to cooperate with one another, rather than merely gaining power for oneself.

5. A brief research agenda

Sections 2-4 gave a longlist of potential actions we can take to make the future better. But there's still a huge amount we don't know, so further research would also be particularly valuable.

This research agenda is far from complete, but I hope it gives a taste of what could be most promising. I'll somewhat artificially divide research topics into "theoretical" and "applied", though of course the boundary between the two is blurry. Currently, the more applied research seems higher-value to me, as it seems more likely to yield decision-relevant results, and I think the more theoretical research is best done with an eye towards decision-relevant issues.

5.1. Theoretical research

- *Lock-in and persistent path-dependence* .
  - How likely is persistent path-dependence over the coming century? What are the mechanisms by which this could come about, and what events are most likely to have persistent path-dependent effects?

[198] This was because: (i) at the very least, there's much more value at stake in such scenarios (because the expected value of the future conditional on survival is higher); (ii) potentially we should care more about what happens in those scenarios, too.

[199] What's more, sadly, given the nature of the world, well-meaning people can easily err by being *too* cooperative and trusting toward bad actors.

    - How likely is it that a future society gets many things right but some crucial things wrong? (For example, how plausible is a future society that is generally eutopian, except that it gets the ethics of digital beings wrong?)
- *Envisioning good futures.*
    - What does a good outcome for humanity look like?
    - This could mean asking what *eutopia* looks like: what the ultimate end-state we should be aiming toward is.
    - Or it could mean asking what a good *intermediate* outcome looks like: a state where, if we enter it, we are very likely to ultimately end up in a near-best future. We could call these *viatopias* . Proposals along these lines have included a "long reflection" — a state with low existential risk and low urgency where people have the freedom and ability to figure out what to do with most of the resources they'll have in the future — or a "grand bargain" where the resources in the universe are split between different value-systems.[200]
- *Predicting future action.*
    - Should we expect future decisions to be guided by ideology rather than self-interest, because, due to enormous wealth, future people will have satiated their self-interested preferences but not their ideological preferences?
    - Should we expect most of those in power in the future to intrinsically value expanding and reproducing, because of evolutionary forces? Or should we expect that almost any ideology will pursue expansion and reproduction as an instrumental goal, and will do so just as successfully as those that intrinsically value expansion? How does this relate to non-consequentialist views which might care more about the process by which decisions are made, rather than the outcome?
    - Finally, what do these analyses imply about how good a future we should expect, or about what we could be doing now to steer it in a better direction?
- *How much is at stake in better futures work.*
    - Assuming no near-term extinction of humanity, what's the difference in value between our expected future and a near-best future?
    - In particular, as discussed in " No Easy Eutopia " and " Convergence and Compromise ," there are three means by which we might get to a near-best future: (i) if most people in the future converge on the correct moral and empirical worldview, and are motivated to do what's best; (ii) if some people in the future converge on the correct moral and empirical worldview, and are motivated to do what's best, and are able to get most of what they want via trade and compromise with others; (iii) if near-best futures are relatively easy to achieve, such that we get to such a future even if almost no one in the future has the correct moral and empirical worldview, or if almost no one is motivated to do what's best.
    - If (i)-(iii) are very likely, either individually or in combination, then there is not all that much at stake in better futures work: if we avoid extinction, then we're likely to get to a near-best future. So we can ask: how likely is each of (i)-(iii)?

[200] Where it's an open question of whether, or in what conditions, these proposals are in fact viatopias.

    - Given what we say about this, what does that imply about what strategies for improving the future are best? In particular, is it more important to increase the chance that future decision-makers carefully reflect on their moral views, or that those decision-makers are motivated to do what's best, or to ensure that there is a diversity of opinion among future decision-makers, and that they are able to trade with one another?

- *The distribution of value over possible futures.*
    - One could have the view that the future's value will be dichotomous: either near-best (achieving >90% of feasibly achievable value) or close to 0 value (achieving <1% of possible value). Is this correct? If not, why not? And what strategic upshots do these conclusions have: should we only be focusing on increasing the probability of a near-best future, or is improving relatively mediocre futures high-priority, too?

5.2. Applied research

Applied research could focus on technologies or political developments that might have predictably path-dependent effects. A list of such developments includes:

- Concentration of power, via economic, political or military means
- Powerful persuasion technology
- Commitment technology
- Preference and belief modification technology
- Extremely accurate lie detection
- Space settlement
- The assignment of rights to AI systems
- Global governance

In most cases, the most important forms of these technologies or developments would be enabled or accelerated by advanced AI.

In each of these cases, applied research could:

1. Do a shallow cause investigation of the area, in order to quickly estimate how high-priority it is, compared to other potential areas.
2. Work out how this development should be governed, including what policies should be in place.
3. Generate and flesh out a list of concrete interventions (beyond just policy) that could plausibly improve how society handles the development.

One could do the same for cross-cutting issues, too, such as AI for epistemics and for the governance of the intelligence explosion in general. And, of course, one could try to add items to the above list — almost certainly I have missed crucial areas of investigation.

6. Conclusion

Thinking about how to make the future better, not just survive it, means focusing on how decisions made in our lifetime might have persistently path-dependent effects. This essay looked at actions we could take, both to keep society's options open for longer and to actively steer things in a positive direction, in particular during and after the development of AGI.

Looking at the future this way shifts our priorities. How power is structured in the long run becomes a central concern, and the risk of power concentration becomes particularly salient. While preventing AI takeover is still vital, we also need to think about the *kind* of future AI might create, which makes understanding and shaping AI's values ('value-alignment') more important. It also means tackling big ethical challenges like the rights of digital beings or how we manage space resources.

This essay did not explicitly discuss *strategic upshots* from a better futures perspective, where holding this perspective in mind should change which strategies we use to reduce existential risk. But, given this discussion, some upshots seem likely to me: plans to increase AI safety by having a single project to build superintelligence look worse, as do restrictions on the widespread use of beneficial AI systems for fear of collusion by misaligned AIs. More generally, this essay raises the risk that we fail to ideally navigate the transition to superintelligence by giving helpful and morally aligned AIs *too little* influence over the most important decisions civilisation faces, rather than too much.

This is all still a nascent area of thought. We still need to do a lot of basic thinking about fundamental questions, like what a truly good future even looks like. Many of the problems I've discussed might be very difficult to solve, but I think we should at least try to address them. Getting these issues right will have impacts over the very long-run course of civilisation. So, even while acknowledging how little we know, we should still push forward — searching for ways to nudge our collective trajectory towards a better future.

Bibliography

Anthropic, ' Exploring model welfare '.

Tom Davidson, Lukas Finnveden, and Rose Hadshar, ' AI-Enabled Coups: How a Small Group Could Use AI to Seize Power ', *Forethought* .

Tom Davidson and Rose Hadshar, ' The Industrial Explosion ', *Forethought* .

Tom Davidson, Rose Hadshar, and Will MacAskill, ' Three Types of Intelligence Explosion ', *Forethought* .

Daniel Eth and Davidson, Tom, ' Will AI R&D Automation Cause a Software Intelligence Explosion? '.

Lukas Finnveden, ' What's Important in "AI for Epistemics"? ', *Forethought* , 23 August 2024.

Rose Hadshar, ' Intelsat as a Model for International AGI Governance ', *Forethought* , 13 March 2025.

Moritz Von Knebel, ‘ When We Are No Longer Needed: Emerging Elites, Tech Trillionaires and the Decline of Democracy ’, *Tech Policy Press* , 8 May 2025.

Will MacAskill and Fin Moorhouse, ‘ Preparing for the Intelligence Explosion ’, *Forethought* .

Cullen O’Keefe, Ketan Ramakrishnan, Janna Tay, and Christoph Winter, ‘Law-Following AI: Designing AI Agents to Obey Human Laws ’, 2 May 2025.

Peter Salib and Simon Goldstein, ‘ AI Rights for Human Safety ’, 1 August 2024.

Carl Shulman and Nick Bostrom, ‘ Sharing the World with Digital Minds ’, *Rethinking Moral Status* , 5 August 2021.

Julian Stastny, Olli Järviniemi, and Buck Shlegeris, ‘ Making deals with early schemers ’, 20 June 2025.

Ana Swanson, ‘ U.S. Unveils Sweeping A.I. Project in Abu Dhabi ’, *The New York Times* , 15 May 2025.

Lizka Vaintrob, ‘ AI Tools for Existential Security ’, *Forethought* .

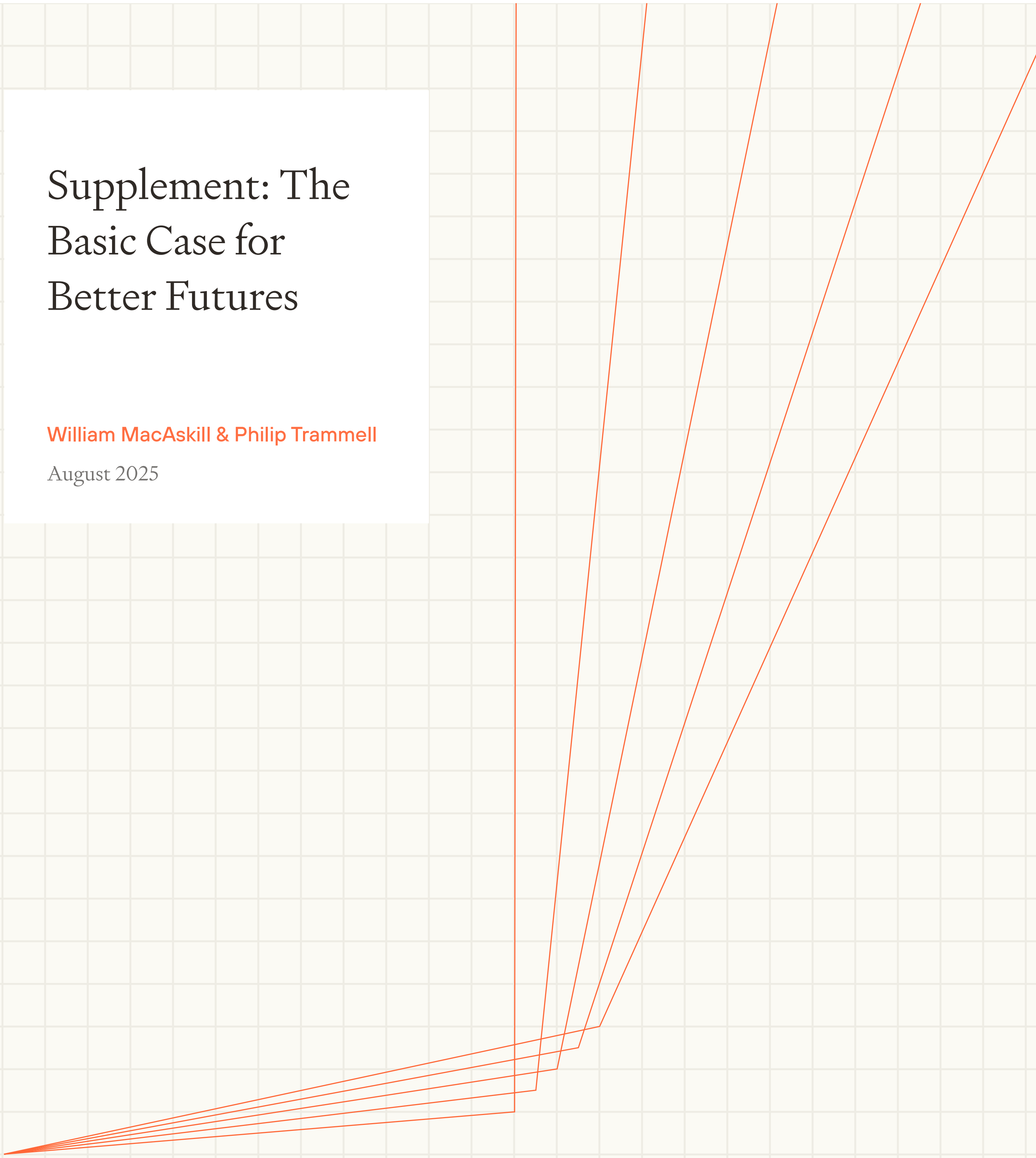
Lizka Vaintrob, ‘ The AI Adoption Gap: Preparing the US Government for Advanced AI ’, *Forethought* , 2 April 2025.

Forethought

Supplement: The Basic Case for Better Futures

William MacAskill & Philip Trammell

August 2025



Introduction

This report introduces a simplified model for evaluating actions aimed at producing long-term good outcomes: the "SF model", where the expected value of the future can be approximated by the product of two variables, Surviving (S) and Flourishing (F). Surviving represents the probability of avoiding a near-total loss of value this century (an "existential catastrophe"), while Flourishing represents the expected value of the future conditional on our survival. Using this model and the "scale, neglectedness, tractability," framework, we argue that interventions aimed at improving Flourishing are of comparable priority to those focused on Surviving.

The SF model

We'll define *value* as the difference that Earth-originating intelligent life makes to the value of the universe.

Insofar as we're aiming to do as much good as possible, we want to maximise expected value. We can break the expected value of an action into two components:[201]

$$EV = EV(\text{near-term}) + EV(\text{future})$$

We'll let "near-term" refer to "between now and 2100". If we accept *longtermism* ,[202] we accept that the best action we can take must be near-best with respect to the latter component.

We can further decompose EV(future) as follows:

$$EV(\text{future}) = P(\text{survival this century})*EV(\text{future} \mid \text{survival this century}) +$$
$$P(\text{not-survival this century})*EV(\text{future} \mid \text{not-survival this century})$$

We'll come back to the definition of "survival this century", but for now we'll say that: (i) by stipulation, outcomes involving the total extinction of Earth-originating life are of 0 value; (ii) the best feasible long-run outcomes are of value 1;[203] and (iii) "survival this century" occurs if nothing has happened by 2100 to lock us into a 0-value future. That is, EV(future | not-survival this century) = 0. So our formula reduces to:

$$EV(\text{future}) = P(\text{survival this century})*EV(\text{future} \mid \text{survival this century})$$

Then we can further break this up:

$$EV(\text{future}) =$$
$$P(\text{survival this century})*P(\text{survival in future centuries} \mid \text{survival this century})*EV(\text{future} \mid \text{survival in both}$$
$$P(\text{survival this century})*P(\text{not-survival in future centuries} \mid \text{survival this century})*EV(\text{future} \mid \text{survival this}$$

Where "survival in future centuries" refers to survival for a much longer time period — say, the next billion years — but not indefinitely (which is likely impossible).

[201] Assuming that value is at least approximately time-separable between previous centuries, this century, and future centuries.

[202] Greaves and MacAskill, ' The case for strong longtermism '.

[203] Where in principle value can be above "1"; this is not an upper bound.

If we take seriously, per longtermism, that almost all expected value is in the far distant future, then EV(future | survival this century and not-survival in future centuries) ≈ 0, and

$$EV(\text{future}) \approx$$
$$P(\text{survival this century})*P(\text{survival in future centuries} \mid \text{survival this century})*EV(\text{future} \mid \text{survival in both}$$

We'll assume that we cannot affect P(survival in future centuries | survival this century),[204] and that this probability is constant across compared actions.[205] Because, for an expected value maximizer, value is unique only up to a positive affine transformation, the multiplication of the value of all options by a constant does not alter those options' relative expected value, so any agent that maximises EV(future) will also maximise EV'(future) = (EV(future) / P(survival in future centuries)). So we have:

$$EV\text{'}(\text{future}) \approx$$
$$P(\text{survival this century})*EV(\text{future} \mid \text{survival this century and survival in future centuries})$$

However, if "survival" refers to locking us into a literally 0-value future, then this is not the most useful breakdown, because then, arguably, almost all current longtermist action is focused on EV(future | survival this century and survival in future centuries). For example, from many perspectives, a future involving misaligned AI takeover is very unlikely to have literally 0 value.

Instead, we replace the idea of "survival" with "survival*", which means that nothing has happened to essentially guarantee that the future will be of *near*-0-value, for some precisification of nearness. We suggest the following precisification. Let a *best feasible future* be a future at the 99.99th centile of your probability distribution over value. *100% value* is the value of a best feasible future. An outcome is "near-0" value if it's within 0.1 percentage points of 0 value. We think this is a natural way of precisifying "no existential catastrophe" — or, at least, precisifying the concept in the way that it's actually used. So, someone who internalised Bostrom's "Maxipok" principle and wanted to minimize existential risk might focus *only* on increasing P(survival*).[206]

Given this definition, EV(future | not-survival* this century) ≈ 0, and so we can use the same decomposition and simplification as above. We get to:

$$EV\text{'}(\text{future}) \approx$$
$$P(\text{survival}^* \text{ this century})*EV(\text{future} \mid \text{survival}^* \text{ this century and survival* in future centuries})$$

[204] Though this isn't strictly true, it make the argument cleaner; and if we relaxed it we think it would make the general thrust of our argument stronger rather than weaker (insofar as we think that actions to increase P(survival in future centuries) would look more similar to actions to increase EV(future | survival this century and survival in future centuries) than to actions to increase P(survival this century).

[205] Relaxing this assumption would also further strengthen our argument insofar as, in worlds in which one successfully averts an existential catastrophe, existential catastrophe is probably more likely in general: P(survival in future centuries) is probably lower, and so the overall value of reducing existential catastrophe is worse than one would expect.

[206] We don't think that relying on the idea of "survival*" is without issues. In particular, it means that some sorts of AI takeover, or human coups, count as "not surviving*" whereas others count as reducing the EV(future | survival) component — which can be unintuitive. Unfortunately, we couldn't think of a breakdown that was an improvement, and this division at least puts classic AI takeover scenarios, which involve the loss of essentially all value, within the "not surviving*" category. An alternative would be to divide into "human controlled" futures and non-human controlled futures. However, we find the dividing line between which futures are "human controlled" and which aren't to be very blurry. And, what's more, we don't think the simplification we've suggested here would work, because (i) it would not be true that EV(future | loss of human control this century) ≈ 0; (ii) in fact, we think that some people should be trying to *increase* EV(future | loss of human control this century). Also note that, on our breakdown, risks of astronomical amounts of suffering, or "s-risks", affect the EV(future | survival) component.

Finally, to tidy things up, we can refer to "P(survival* this century)" as "S" for *Surviving* and "EV(future | survival* this century and survival* in future centuries)" as "F" for *Flourishing*. So we have:

$EV'(\text{future}) \approx S*F$

We call this *the SF model*.

In what follows, we will assume that F is currently positive, so that it is valuable to raise rather than lower S.

The basic case for work on Flourishing

First, we note there's nothing stemming from the nature of wanting the future to go better to focus on Surviving rather than Flourishing. And, at the very least, the two-factor model makes clear that we shouldn't focus *only* on increasing S. In general, if variables x and y are not perfectly correlated and there's a big sample size, then max(x*y) does not equal max(x) or max(y): the "tails come apart".[207] So, maxing out on S is very unlikely to be the best course of action.

But there's also an argument based on the "scale, neglectedness, tractability" framework for thinking that work on Flourishing is in at least the same ballpark of priority as work on Survival. Let's go through each aspect of the framework in turn.

Scale

The "scale" of a problem refers to the total amount of value at stake. We can assess the relative scale of Survival and Flourishing by examining the potential gains from improving each.

We'll consider 100% value to be the ceiling on how high F can be. As argued in essays " No Easy Eutopia " and " Convergence and Compromise ", we're probably far from the ceiling on F. If we are also *not* far from the ceiling on S, then:

1. Making some *absolute* change to F does more good than making the same absolute change to S.

- a. Suppose we should think that S is at 80% and F is at 10%.
- b. Using these numbers, then increasing the latter by one percentage point has 8 times the positive impact of increasing the former.[208]

2. Making some *proportional* reduction in the loss of value from the future going poorly does more good than making the same proportional reduction in the risk of existential catastrophe. This is the formulation used in Cotton-Barratt's formalisation of the scale, neglectedness, tractability framework, and the formulation we prefer.[209]

[207] Alexander, ' The Tails Coming Apart As Metaphor For Life '; Thrasymachus, ' Why the tails come apart '.

[208] An absolute increase of one percentage point (0.01) in S would change the expected value from 0.8 * 0.1 = 0.08 to 0.81 * 0.1 = 0.081, a gain of 0.001. In contrast, the same absolute increase in F would change the value to 0.8 * 0.11 = 0.088, a gain of 0.008.

[209] Cotton-Barratt, ' Prospecting for Gold '.

- a. Again using the numbers we just suggested, a 1% reduction in the loss of value from the future going poorly (i.e. a 1% reduction in 1-F) has 36 times the positive impact of a 1% reduction in near-term existential risk (1-S).[210]

That is, given that we are further from the ceiling of F than we are of S, the problem of non-flourishing futures has greater *scale* than the problem of existential catastrophe.

We think that the illustrative numbers we've given are reasonable, or are even on the conservative side from the perspective of our argument. Here we report on what, given our views, we see as a reasonable range of estimates for S and F, and the comparative scale of the problem of non-flourishing.

S	F	Gains from 1 pp increase in S	Gains from 1 pp increase in F	Ratio (comparative scale)
0.65	0.5	0.005	0.0065	1.3
0.95	0.05	0.0005	0.0095	19
0.99	0.01	0.0001	0.0099	99

Comparative scale, using "absolute change" formulation of scale.

S	F	Gains from 1% reduction in (1-S)	Gains from 1% reduction in (1-F)	Ratio (comparative scale)
0.65	0.5	0.00175	0.00325	1.9
0.95	0.05	0.000025	0.009025	361
0.99	0.01	0.0001	0.9801	9801

Comparative scale, using "proportional change" formulation of scale.

So we see the problem of non-flourishing as having 1.9x–9801x the scale as the problem of non-survival. Our own views lean towards the higher end of this range; given uncertainty over these parameters, we think that the scale of non-Flourishing is on the order of 100x the scale of non-Survival.

[210] A 1% reduction in the risk of non-survival increases S by 0.01 * 0.2 = 0.002. The new EV is 0.802 * 0.1 = 0.0802, a gain of 0.0002. A 1% reduction in the value lost to non-flourishing increases F by 0.01 * 0.9 = 0.009. The new EV is 0.8 * 0.109 = 0.0872, a gain of 0.0072. 0.0072 / 0.0002 = 36 times the positive impact. The same ratio would hold if we were focusing on total potential upside (i.e. a 100% reduction in the scale of the problem) from addressing each problem.

Two additional arguments regarding Flourishing's scale

First, at least some sorts of interventions to increase S are less valuable than commonly expected, because if human-originating life goes extinct, then non-human-originating life may settle the stars in its place. This could happen because *Homo sapiens* goes extinct, but, in the hundreds of millions of years remaining before the Earth is no longer habitable, nonhuman life on Earth evolves higher intelligence and cumulative cultural learning, and some nonhuman civilisation emerges. Or it could happen because extraterrestrial life settles our corner of the cosmos. Of course, it's hard to know how likely each of these are, though we think that it's more likely than not that nonhuman life would develop civilisation even if *Homo sapiens* in isolation died out, and perhaps about even that alien civilisations would settle our corner of the cosmos if no Earth-originating intelligent life did so.

To formalize the first point in terms consistent with the framework above: S is higher than commonly understood, because e.g. a pandemic that killed every human but left other species intact would still leave the probability of a "survival" scenario—one in which Earth-originating life makes a significant difference to the value of the universe—significantly positive.

To formalize the second, the possibility of aliens renders F lower than commonly understood. Suppose that, before accounting for aliens, we believe that f^* is the value (i.e. the impact of Earth-originating life on the value of the universe) of a best feasible future and f is the expected value of the future without our intervention, so that we believe $F = f/f^*$. Then suppose we come to believe that, in the absence of Earth-originating life, aliens will in expectation use fraction A of the resources in our light cone that we would have used. Suppose for simplicity (i) that if we encounter an alien civilisation, the resources that we each would have used in the other's absence will, in expectation, be split evenly, (ii) that we and the aliens will in expectation put resources to equally valuable uses, and (iii) that value is linear in resources. Now the value of a best feasible future is $f^* - fA/2$ and the current expected value of the future is $f - fA/2$, so $F = (f - fA/2) / (f^* - fA/2) < f/f^*$.

Strictly speaking, introducing the possibility of aliens also raises S considerably, since if AI takes over on Earth and spreads worthless structures through the universe, partially displacing aliens who would otherwise (in expectation) have put those resources to good use, this now constitutes a case of survival*. That is, the impact of Earth-originating life on the value of the universe in this case is not near-zero but very negative. In this case, however, the definition of survival* given above does not preserve the spirit of our argument; we don't mean to argue for working on Flourishing rather than Survival simply by reclassifying attempts to prevent AI takeover as instances of the former. When comparing the value of working on Flourishing to that of working to reduce the risk that morally worthless AI takes over and spreads, under assumptions in the direction of (i)-(iii) and the assumption that A is non-negligible, it is more natural to define value as *the difference that Earth-originating intelligent life makes to the value of the universe above the baseline of AI takeover by AIs that spread, are non-sentient, and have some valueless goal like producing paperclips*. This definition preserves the point that the existence of aliens decreases F, and since in the absence of aliens the AI-takeover value-baseline is (near-)zero, it does not change our discussion on any other point.

The second additional argument regarding Flourishing's scale concerns what our views on S and F should be *given that we have successfully raised F or S*.[211] In particular, it seems to us that your E[F |

[211] See also: Trammell, 'Which World Gets Saved'.

you successfully raised S] should be generally lower than your unconditional E[F]:[212] scenarios where you successfully raised S are probably scenarios where existential risk was high; such scenarios are probably a shambles in other ways, too, and not on track for a truly flourishing future.[213] In contrast, we don't see strong arguments for thinking that E[S | you successfully raised F] is lower than E[S].

To illustrate this quantitatively, suppose the world can be in one of two states: a 'well-run' state with *Flourishing* at 19%, or a 'badly-run' state with *Flourishing* only at 1%. If each state is equally likely, our unconditional expectation for *Flourishing* is 10%. Now, suppose the risk of an existential catastrophe is just 2% in the well-run state, but 38% in the badly-run state (giving an unconditional risk of 20%). On learning that a catastrophe would have occurred but for our successful intervention (raising S), we should update our beliefs. This new information is strong evidence we are in the badly-run state (a 95% posterior probability).[214] Accordingly, our expectation for

[212] To clarify: when we compare E[F | you successfully raised S] and E[F], we are defining F in reference to what you *unconditionally* regarded as the best feasible future (i.e. the future at the 99.99th percentile, or f^*). Given how we've defined "best feasible future", it's possible in principle that, after taking some action, your estimate of the expected value of the future could go down but your estimate of E[F] could go up, if you thought that (i) f^* is of much lower value than you thought before; (ii) but, conditional on survival, future civilisation will in expectation capture a larger fraction of that value than you thought before. Keeping f^* fixed when comparing [F | you successfully raised S] and E[F] avoids getting confused by such issues.

[213] Depending on your moral views, it could even become plausible that you should think that E[F | you successfully raised S] is negative even if your unconditional E[F] is positive: plausibly, scenarios where existential risk is high are also scenarios where there is a greater risk of threats or other risks of astronomical suffering. If so, then averting existential catastrophe would be *bad* to do even if you think the future is better than extinction in expectation. A related point is that it might be harder to increase S than you might intuitively think due to the correlation over time of existential risk per year. Currently, we should be unsure about how high the per-year risk of existential catastrophe is, or will soon be. Because of this, those worlds where we manage to *prevent* some existential catastrophe are, very likely, those worlds in which the per-year risk of existential catastrophe is high — that's where most of the catastrophes it's possible to prevent are. But if so, then, conditional on successfully preventing an existential catastrophe, S is lower than it is unconditionally, and the overall value of the future is lower. Thorstad, 'Existential risk pessimism and the time of perils' elaborates on this point.

[214] Let H_B be the hypothesis that we are in the 'badly-run' state, and H_W be the hypothesis that we are in the 'well-run' state. Let E be the evidence that an existential catastrophe would have occurred absent our successful intervention (with perfect detection). The prior probabilities of each state are: $p(H_B)=0.5$, $p(H_W)=0.5$. We assume the likelihood of E in each state is equal to the risk of an existential catastrophe in those states: $p(E|H_B)=0.38$, $p(E|H_W)=0.02$. Since H_B

Flourishing, conditional on having raised S, plummets from 10% to just 1.9%. Using these illustrative numbers, raising S by one percentage point would be 5x less valuable than it would have naively appeared, before we took conditional expectations into account.

It's not inconsistent to think that, in some circumstances, your E[F | you successfully raised S] should be *higher* than your unconditional E[F]. One such argument is that we should, currently, be unsure on how much capability one can in principle reap from a given number of FLOP (i.e. what "optimal algorithmic efficiency" is). If optimal algorithmic efficiency is very high, then: (i) we should also expect a faster transition to scary capability levels, and therefore existential risk from AI takeover is higher; but (ii) we should expect that the maximal amount of value one can produce per FLOP is higher. But, if so, then conditional on you raising S by preventing AI takeover, you are more likely to be in the "high optimal algorithmic efficiency" scenario, and therefore E[F] is higher, too.

On balance, we think that the former of these arguments is stronger, in particular because it seems unclear to us how strong the correlations in (i) and (ii) really are, but we acknowledge that this is debatable.

Neglectedness

We expect the problem of non-flourishing to be more *neglected* than the problem of not-surviving, both now and in the near-term future.

Of course, near-term existential risk is extremely neglected: even just from the perspective of the self-interest of the population of the United States, it gets far less attention than it should. But it

and H

W

are mutually exclusive and jointly exhaustive, p(E)=p(E|H

B

)P(H

B

)+p(E|H

W

)P(H

W

), which = 0.38 × 0.5 + 0.02 × 0.5 = 0.19 + 0.01 = 0.2. Using Bayes' rule to derive the posterior p(H

B

|E) as follows: p(H

B

|E)=(p(E|H

B

)P(H

B

))/p(E), which = (0.38 × 0.5)/0.2 = 0.19/0.2 = 0.95.

has gathered at least a reasonable amount of attention in recent years. And we should expect that amount of attention to increase: most people don't want to die, and as people progressively realise the extent of existential risk posed by AI, engineered pathogens, and other dangerous technology, we should expect them to invest a lot more time, money and political attention to reduce those risks than they currently are investing.

Quantitatively, the willingness to pay to avoid existential catastrophe even just from the United States is truly enormous. The value of a statistical life in the US — used by the US government to estimate how much US citizens are willing to pay to reduce their risk of death — is around $10 million. The willingness to pay, therefore, from the US as a whole, to avoid a 0.1 percentage point of a catastrophe that would kill everyone in the US, is over $1 trillion.[215] We don't expect these amounts to be spent on existential risk reduction, but they show how much latent desire there is to reduce such risks. We'd expect at least part of this latent desire to become progressively mobilised with increasing indications that various global catastrophic risks, such as biorisks, are real.

No comparable latent demand exists for improving Flourishing, at least for many sub-areas within Flourishing.[216] The amount of even latent interest in, for example, ensuring that resources outside of our solar system are put to their best use, or that misaligned AI produces a somewhat-better future than it would otherwise have done even if it kills us all, is tiny, and we don't expect society to mobilise massive resources towards these issues even if there were indications that those issues were pressing. We expect them to remain extraordinarily neglected.

Even areas that seemingly get a lot of attention can still contain highly neglected sub-problems, if attention within the area is not focused on the very most effective interventions. For example, the risk of autocracy is well-known and a lot of people focus on it; but very few people focus on how to reduce the risk of AI-enabled military coups, even though these are plausibly the most effective interventions currently.

And, even among those who take longtermism and the possibility of an imminent intelligence explosion seriously, there is currently much less attention on F than on S. AI safety and biorisk reduction have, thankfully, gotten a lot more attention and investment in the last few years. There has been less of an uptick in work dedicated to improving the quality of the future, and such work remains extremely neglected even within this community.

All things considered, we expect Flourishing to be something like 10x–100x more neglected than Survival, in terms of quality-adjusted financial and labour allocation. Of course, this is just a rough guesstimate.

Tractability

The *tractability* of work on Flourishing is much less clear and, in our view, if the case for work on Flourishing fails, it probably fails here.

[215] Shulman and Thornley, ' How Much Should Governments Pay to Prevent Catastrophes? Longtermism's Limited Role '.

[216] There is a lot of desire to avoid some sorts of catastrophic lock-in, though, especially when the lock-in scenario is salient and would involve large costs to the present generation. In particular, there's a lot of desire to avoid authoritarian outcomes, whether via the US becoming authoritarian, or from an authoritarian country winning the race to AGI. However, in comparing causes what is relevant is which sub-areas perform *best* on the product of scale, neglectedness, and tractability. If some better futures work is very non-neglected, then it's likely better to focus on other more neglected areas.

In at least some cases, there are reasons for pessimism about tractability:

- First, the main argument we made for the relative neglectedness of Flourishing might be an argument against relative tractability. If there's a lot of latent desire to reduce existential risk, and the blocker is merely that people don't know how high existential risk really is, then it might be much easier to unlock that latent desire than it is to convince other people to care about something that they don't, currently, care about.
    - Even worse, there might be active *opposition* to certain sorts of work on Flourishing, at least in situations where steering us towards a better future involves acting, to some degree, against the self-interest of those in the present generation. For example, people might for self-interested reasons want to give AIs no welfare rights, even if doing so were the right thing to do in the long term.
- Second, at least some Flourishing-related issues are less likely to have technical fixes that can be implemented unilaterally than many Surviving-related issues. To reduce pandemic risk, individual actors (like companies, foundations, or countries) can produce stockpiles of PPE, build bunkers, or subsidise the development of physical sterilisation methods like far-UVC lighting. To reduce AI takeover risk, individual actors can invest heavily in alignment research. In contrast, there seem to be fewer actions of this sort for at least many areas to promote Flourishing: the most obvious actions around AI rights, space governance, or collective decision-making require regulation, and perhaps often international regulation. This is generally harder to achieve, because success requires convincing other actors of the importance of the issue and steering other actors to make better decisions on those issues than they otherwise would have done.
- Third, some Flourishing-related issues (such as deep space governance) arise deeper into the intelligence explosion than threats to Survival. This means that:
    - More quality-adjusted AI labour will go towards the issue.
    - Early work will be more likely to be irrelevant.
    - Early action will be more likely to have difficult-to-predict consequences.

On the last of these points, some similar issues are discussed in "Preparing for the Intelligence Explosion", section 5.[217] In brief: we think that this often does result in a major haircut on the value of working on these issues, although in many cases even with a 90% haircut the expected value can still be very high. And in many cases early work really is very high-leverage: some challenges arise early; some windows of opportunity close early; and early work can change when and how superintelligent assistance is used on these problems.

Finally, one of the strongest arguments against the tractability of better futures work is that it simply won't pay off in time. It took many years for the fields of AI safety and AI governance to develop; but by that time, we might well already have superintelligence. Similarly, it may take years for work in these currently-neglected areas to pay off. Work in these particularly-neglected better futures areas, then, looks comparatively better in worlds where AGI comes 3+ years in the future.

But there are some reasons for optimism on tractability as well. First, there are some quite general reasons:

- In general , tractability doesn't vary by as much as importance and neglectedness.

[217] MacAskill and Moorhouse, ' Preparing for the Intelligence Explosion '.

  - Assuming logarithmic returns, then for a problem to be an order of magnitude less tractable than a baseline, we would require approximately 1000 times the resources to achieve 10% progress. Conversely, problems an order of magnitude more tractable would be nearly fully solved with a single doubling; these must be rare if they are also large-scale and neglected. Thus, problems which seem at all tractable often fall within roughly two orders of magnitude of tractability.
  - Truly intractable problems usually meet specific conditions: they might not be solvable even in theory (like perpetual motion machines), or there's no discernible "plan of attack" (as Hamming noted regarding fundamental physics challenges like time travel). The sub-areas within Flourishing that we highlight don't seem to meet these criteria for extreme intractability.
- In cause areas where very little work has been done, it's hard for *expected* tractability to be very low. Because of how little we know about tractability in unexplored cause areas, we should often put significant credence on the idea that the cause will turn out to be fairly tractable; this is enough to warrant some investment into the cause area — at least enough to find out how tractable the area is.[218]
- There are many distinct sub-areas within better futures work. It seems unlikely that tractability in all of them is very low, and unlikely that their tractability is very highly correlated.
- In at least some cases, there's an opportunity for "pulling the rope sideways", at least for now. For example, because there is very little attention on the governance of extrasolar resources, one might be able to advocate for and get uptake of sensible proposals with very little pushback.
- There's a reasonably promising track record of areas with seemingly low tractability turning out to be surprisingly tractable. A decade ago, work on risks from AI takeover and engineered pathogens seemed very intractable: there was very little that one could fund, and very little in the way of promising career paths. But this changed over time, in significant part because of (i) research work improving our strategic understanding, and shedding light on what interventions were most promising; (ii) scientific developments (e.g. progress in machine learning) making it clearer what interventions might be promising; (ii) the creation of organisations that could absorb funding and talent. All these same factors could well be true for better futures work, too.
- Early work in a seemingly-intractable area can take the form of field-building, which pays off when it becomes clearer what tractable paths forward there are. Much early work on AI safety paid off in this way; arguably, this is still the main path to impact from current AI safety work.

Of these considerations, it's the last two that move us the most. It doesn't feel long ago that work on AI takeover risk felt extraordinarily speculative and low-tractability, where there was almost nowhere one could work for or donate to outside of the Future of Humanity Institute or Machine Intelligence Research Institute. In the early days, we were personally very sceptical about the tractability of the area. But we've been proved wrong. Via years of foundational work — both research work figuring out what the most promising paths forward are, and via founding new organisations that are actually squarely focused on the goal of reducing takeover risk or biorisk,

[218] For discussion, see Cotton-Barratt, ' How to treat problems of unknown difficulty '; Cotton-Barratt, 'Cost-effectiveness of research: overview'.

rather than on a similar but tangential goal — the area has *become* tractable, and now there are dozens of great organisations that one can work for or donate to. We expect that, with effort, a similar dynamic would play out with better futures work. Foundational research will help us figure out the most promising paths. Founding new organisations will generate great places to work for or donate to.

As well as general considerations around tractability, we can look at specific areas. One issue facing better futures work is that many of the most neglected issues concern technologies or political developments that haven't occurred yet. However, the same is true for AI takeover: the central challenge, there, is how to align superintelligent AI systems, which are not yet here. The path that AI safety work has taken is to work on "baby" versions of the problem: developing tools and techniques (like mechanistic interpretability) that are both helpful to aligning present-day AI models, and will scale to be useful to aligning superintelligent AIs. Even more importantly, this work also helps to build a vibrant field around AI safety, increasing the amount of skilled labour able to tackle the problem once we are training superintelligence.

We can use the same strategy in many other areas. First, consider AI for persuasion, reasoning, and decision-making. Already there are major worries about AI distorting societal epistemics. Currently, these worries are often misguided or overblown. But as AI becomes more powerful, how AI is incorporated into society's epistemology might become of critical importance: people could lock in their pre-existing beliefs; they could become convinced of memetically powerful but false ideas; they could simply fail to keep up with all the change that's occurring over the course of the intelligence explosion; or they could use AI to become much more curious and enlightened.

Second, the alignment target. As well as making sure that AI is aligned, we need to figure out what AI is aligned *with* . Ideally, we want to design an alignment target such that (i) even if we fail to align the superintelligence and it disempowers humanity, that superintelligence still produces a good future; (ii) if we do succeed at alignment, the superintelligence guides us on the path to viatopia, rather than some sort of catastrophic lock-in. The "baby" version of the problem is to work out what goals or values to align *current* AI systems with.

Third, deep space governance. We are not yet at the point of seriously considering how extrasolar resources should be allocated among people and states. But the question of rights over, say, asteroid mining is live and will be decided upon soon. How we answer such questions might inform and set precedents for how larger-scale resource-use questions are settled.

In each of these cases, one can work on baby versions of the challenges today. And work on those issues today sets precedents that will guide the much larger decisions that are yet to come, can generate insights that might be useful to those larger-scale challenges, and builds up a field of people who can work on the most-crucial issues when the time comes.

Overall, we do think that the tractability of work on Flourishing is lower than work on Surviving. This is the least certain parameter, but we think that a factor of 10x-100x lower is reasonable. In other words, we think that the neglectedness and tractability considerations approximately cancel out.

Personal Fit

The ideal portfolio of action would include significant work on both Survival and Flourishing, such that, for career choice at least, *personal fit* will often be the determining factor.

Better futures work seems like it's unusually well-suited to people in or adjacent to the effective altruism movement. It particularly benefits from: having generalist research knowledge; being comfortable in messy pre-paradigm fields; being willing to shift quickly between more theoretical and very concrete work; taking ethics really seriously, including more abstract or unusual ethical ideas like digital rights; and understanding the dynamics of a potential intelligence explosion.

That said, better futures is highly inchoate and pre-paradigm: this means that it's often hard to meaningfully contribute. The most important work, right now, is on making the area more tractable, whether via fundamental research to help us figure out what strategies are most promising, or by setting up organisations that are focused on the most-important goals.

For many people, the best strategy right now would be to focus on building up either career capital or financial capital, in order to be able to contribute more within better futures once there are more opportunities to contribute. (Better futures makes financial capital particularly more valuable, especially insofar as many of these areas are areas that the foundation Open Philanthropy would not want to fund, unlike AI takeover risk and pandemic risk). For many other people, the personal fit costs will be too great, and it'll be higher-impact to focus on the more well-developed areas of existential risk reduction.

Conclusion

Putting these pieces together, we think that Flourishing is at least in the same ballpark of priority as Survival. The SF model suggests a diversified portfolio of interventions is optimal. The greater scale and neglectedness of Flourishing-focused work compensate for its currently lower and more uncertain tractability. For those whose personal fit is good — especially those comfortable with messy, pre-paradigmatic problems — work on Flourishing seems promising.

Bibliography

Scott Alexander, ' The Tails Coming Apart As Metaphor For Life ', *Slate Star Codex* , 25 September 2018.

Owen Cotton-Barratt, ' Cost-effectiveness of research: overview ', *The Future of Humanity Institute* , 4 December 2014.

Owen Cotton-Barratt, ' How to treat problems of unknown difficulty ', *The Future of Humanity Institute* , 29 July 2014.

Owen Cotton-Barratt, ' Prospecting for Gold ', 2016.

Hilary Greaves and Will MacAskill, ' The case for strong longtermism ', 14 June 2021.

Will MacAskill and Fin Moorhouse, ' Preparing for the Intelligence Explosion ', *Forethought* .

Carl Shulman and Elliott Thornley, ‘How Much Should Governments Pay to Prevent Catastrophes? Longtermism’s Limited Role’

David Thorstad, ‘ Existential risk pessimism and the time of perils ’, 2022.

Thrasymachus, ‘ Why the tails come apart ’, 1 August 2014.

Phil Trammell, ‘ Which World Gets Saved ’, 9 November 2018.