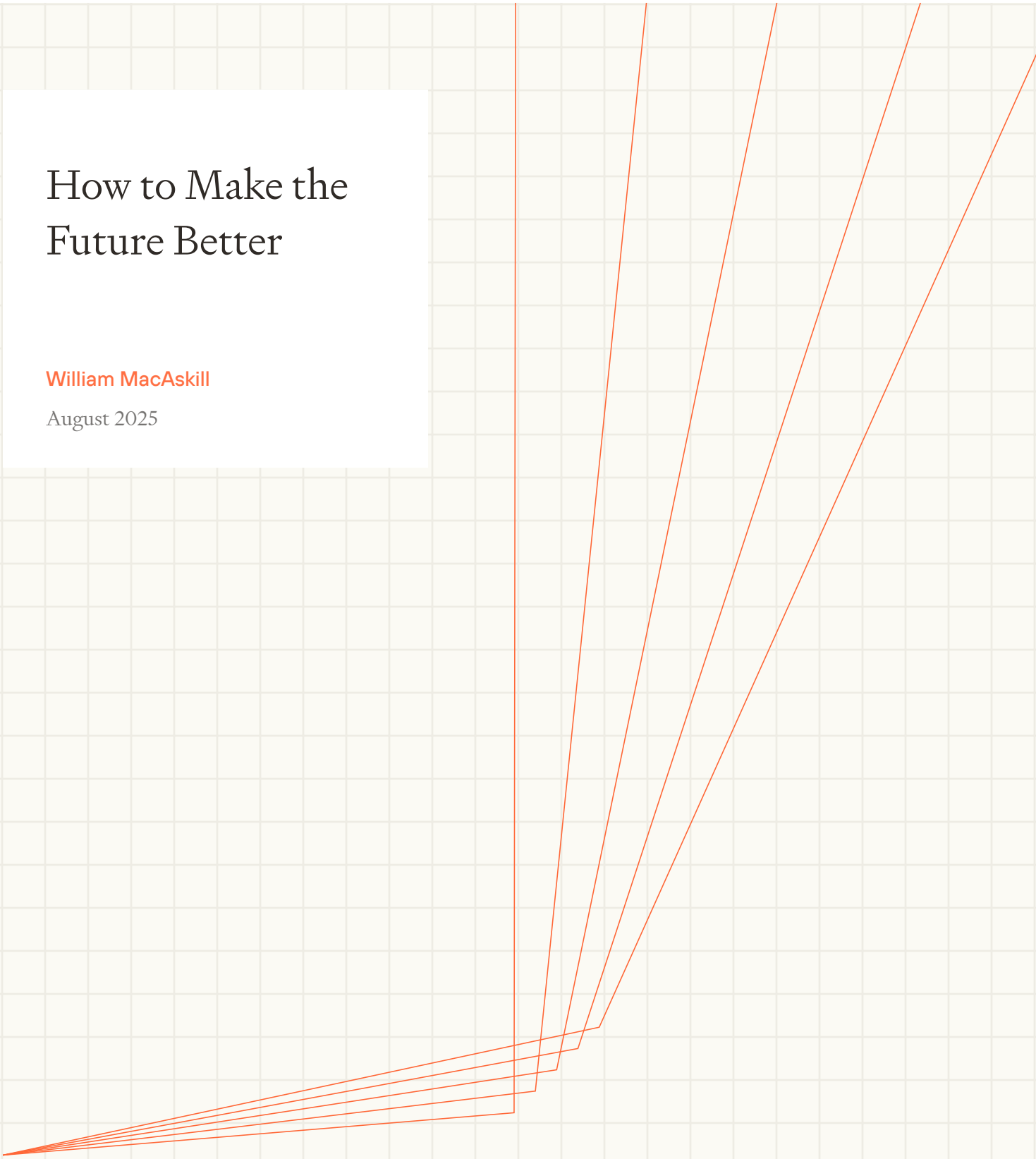
Forethought

How to Make the Future Better

William MacAskill

August 2025



Contents

How to Make the Future Better

Bibliography

William MacAskill

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.[1]

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.

---

1 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".[2] 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[3] 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.[4] 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.[5] We could encourage multilateral

---

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

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

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

5 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,[6] 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.[7] 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)[8] 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[9] could prove popular and effective if they were discussed seriously, the

---

6 For relevant discussion, see Hadshar, ' Intelsat as a Model for International AGI Governance '.

7 For more discussion, see the subsection on space governance in Section 4 of MacAskill and Moorhouse, ' Preparing for the Intelligence Explosion '.

8 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.

9 For example: "no single country or company can harness most solar output, or claim most accessible star systems and galaxies, this century".

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.[10]

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.[11] 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."[12]

A successful example of explicitly temporary commitments is with the creation of Intelsat,[13] 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

10 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.

11 This idea comes from Toby Ord.

12 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.

13 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.[14]

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.[15]

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.[16] 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

---

14 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.

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

16 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).[17] 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.

---

17 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[18] 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

18 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.[19] 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:

19 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.[20] 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.

---

20 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.[21]

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

21 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.)[22] 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;[23] 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.[24]

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

22 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.

23 Anthropic, ' Exploring model welfare '.

24 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.[25]

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[26]

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

25 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.

26 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,[27] 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,[28] 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.[29]

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

---

27 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 '.

28 See Finnveden, 'What's Important in "AI for Epistemics"?' .

29 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[30] 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.[31]

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

30 Roughly, global priorities research pertaining to the (long-run) future.

31 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.[32] 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.[33]

However, there are some important reasons against at least some strategies that fall under the category of 'empowering responsible actors'.[34] 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).[35] Second are decision-theoretic arguments.[36] 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.[37] 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.[38]

---

32 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.

33 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.

34 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'.

35 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.

36 See Christiano.

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

38 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,[39] 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.[40]

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?

39 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.

40 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.[41]

- *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)?

---

41 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.