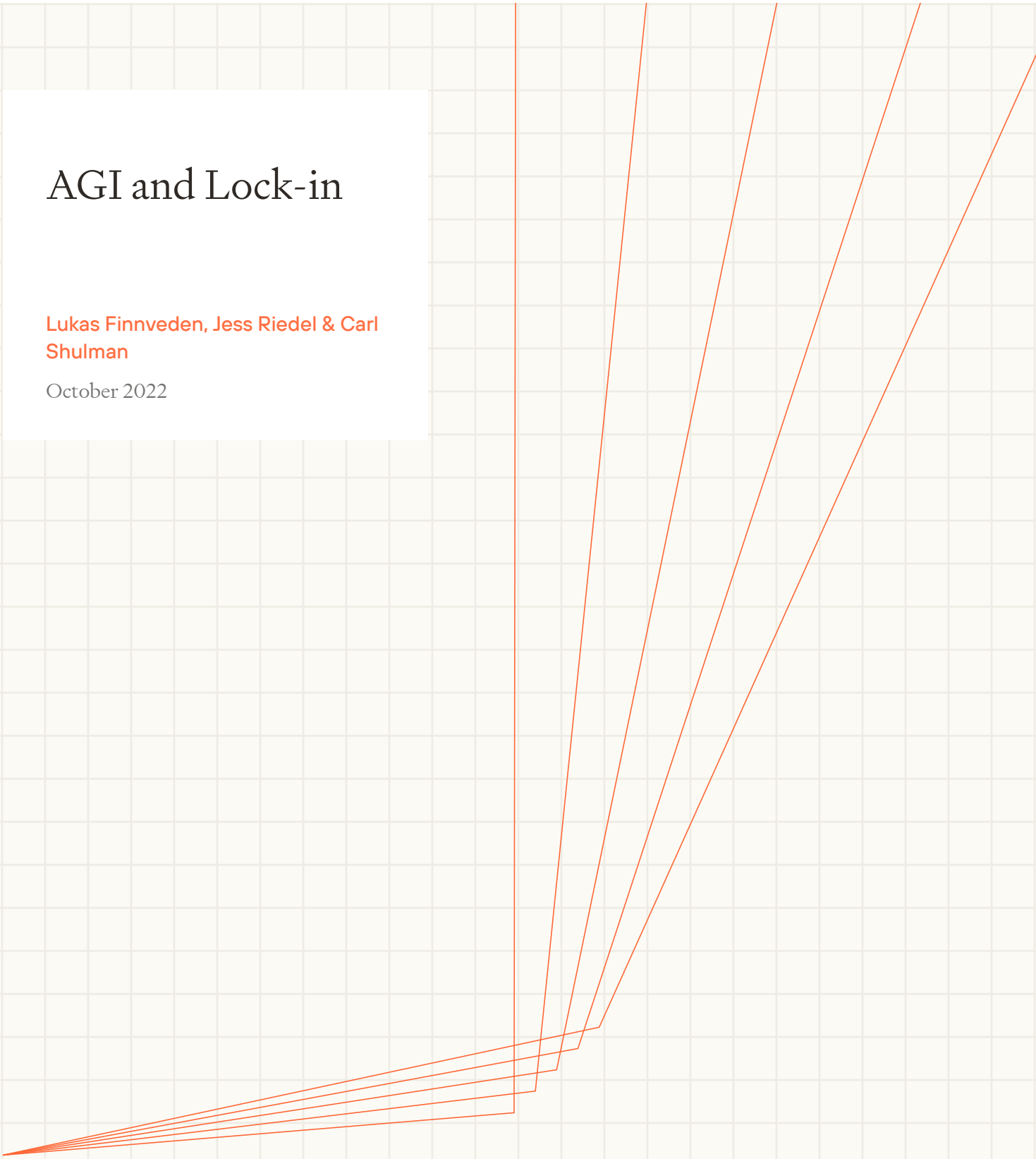
Forethought

AGI and Lock-in

Lukas Finnveden, Jess Riedel & Carl Shulman

October 2022



Contents

AGI and Lock-in

Lukas Finnveden, Jess Riedel & Carl Shulman

Lukas Finnveden was the lead author. Some parts of this document started as an unfinished report prepared by Jess Riedel while he was an employee at Open Philanthropy. Carl Shulman contributed many of the ideas, and both Jess and Carl provided multiple rounds of comments. Lukas did most of the work while he was part of the Research Scholars Programme at the Future of Humanity Institute (although at the time of publishing, he worked for Open Philanthropy). All views are our own.

For correspondence, please email contact@forethought.org.

Abstract

The long-term future of intelligent life is currently unpredictable and undetermined. We argue that the invention of artificial general intelligence (AGI) could change this by making extreme types of lock-in technologically feasible. In particular, we argue that AGI would make it technologically feasible to (i) perfectly preserve nuanced specifications of a wide variety of values or goals far into the future, and (ii) develop AGI-based institutions that would (with high probability) competently pursue any such values for at least millions, and plausibly trillions, of years.

0 Summary

0.0 The claim

Life on Earth could survive for millions of years. Life in space could plausibly survive for trillions of years. What will happen to intelligent life during this time? Some possible claims are:

- A) Humanity will almost certainly go extinct in the next million years.
- B) Under Darwinian pressures, intelligent life will spread throughout the stars and rapidly evolve toward maximal reproductive fitness.
- C) Through moral reflection, intelligent life will reliably be driven to pursue some specific "higher" (non-reproductive) goal, such as maximizing the happiness of all creatures.
- D) The choices of intelligent life are deeply, fundamentally uncertain. It will at no point be predictable what intelligent beings will choose to do in the following 1000 years.
- E) It is possible to stabilize many features of society for millions or trillions of years. But it is possible to stabilize them into many different shapes — so civilization's long-term behavior is contingent on what happens early on.

Claims A-C assert that the future is basically determined today. Claim D asserts that the future is, and will remain, undetermined. In this document, we argue for claim E: Some of the most important features of the future of intelligent life are currently undetermined but could *become* determined relatively soon (relative to the trillions of years life could last).

In particular, our main claim is that artificial general intelligence (AGI) will make it technologically feasible to construct long-lived institutions pursuing a wide variety of possible goals. We can break this into three assertions, all conditional on the availability of AGI:

1. It will be possible to preserve highly nuanced specifications of values and goals far into the future, without losing any information.
2. With sufficient investments, it will be feasible to develop AGI-based institutions that (with high probability) competently and faithfully pursue any such values until an external source stops them, or until the values in question imply that they should stop.

3. If a large majority of the world's economic and military powers agreed to set-up such an institution, and bestowed it with the power to defend itself against external threats, that institution could pursue its agenda for at least millions of years (and perhaps for trillions).

Note that we're mostly making claims about *feasibility* as opposed to *likelihood*. We only briefly discuss whether people would *want* to do something like this in Section 2.2.

(Relatedly, even though the possibility of stability implies E, in the top list, there could still be a *strong tendency* towards worlds described by one of the other options A-D. In practice, we think D seems unlikely, but that you could make reasonable arguments that any of the end-points described by A, B, or C are probable.)

Why are we interested in this set of claims? There are a few different reasons:

- The possibility of stable institutions could pose an existential risk, if they implemented poorly chosen and insufficiently flexible values.
- On the other hand, if we want humane values or institutions such as liberal democracy to survive in the long-run, some types of stability may be crucial for preserving them.
- The possibility of ultra-stable institutions pursuing any of a wide variety of values, and the seeming generality of the methods that underlie them, suggest that significant influence over the long-run future is possible. This should inspire careful reflection on how to make it as good as possible.

We will now go over claims 1, 2, and 3, from above in more detail.

0.1 Preserving information

In the beginning of human civilization, the only way of preserving information was to pass it down from generation to generation, with inevitable corruption along the way. The invention of writing significantly boosted civilizational memory, but writing has relatively low bandwidth. By contrast, the invention of AGI would enable the preservation of entire minds. With whole-brain emulation (WBE), we could preserve entire human minds, and ask them what they would think about future choices. Even without WBE, we could preserve newly designed AGI minds that would give (mostly) unambiguous judgments of novel situations. (See section 4.1.)

Such systems could encode information about a wide variety of goals and values, for example:

- Ensure that future civilisational decisions are made democratically.
- Enforce a ban on certain weapons of mass destruction (WMD)
- Make sure that reverence is paid to some particular religion.
- Always do what some particular group of humans would have wanted.

Crucially, using digital error correction, it would be extremely unlikely that errors would be introduced even across millions or billions of years. (See section 4.2.) Furthermore, values could be stored redundantly across many different locations, so that no local accident could destroy them. Wiping them all out would require either (i) a worldwide catastrophe, or (ii) intentional action. (See section 4.3.)

0.2 Executing intentions

So let's say that we can store nuanced sets of values. Would it be possible to design an institution that stays motivated to act according to those values?

Today, tasks can only be delegated to humans, whose goals and desires often differ from the goals of the delegator. With AGI, all tasks necessary for an institution's survival could instead be automated, performed by artificial minds instead of biological humans. We will discuss the following 2 questions:

- Will it be possible to construct AGI systems that (with high probability) are aligned with the intended values?
- Will such systems stay aligned even over long periods of time?

0.2.1 Aligning AGI

Currently, humanity knows less about how to predict and control the behavior of advanced AI systems than about predicting and controlling the behavior of humans. The problem of how to control the behaviors and intentions of AI is commonly known as *the alignment problem* , and we do not yet have a solution to it.

However, there are reasons why it could eventually be far more robust to delegate problems to AGI, than to rely on (biological) humans:

- With sufficient understanding of how to induce particular goals, AI systems could be designed to more single-mindedly optimize for the intended goal, whereas most humans will always have some other desires, e.g. survival, status, or sexuality.
- AI behavior can be thoroughly tested in numerous simulated situations, including high-stakes situations designed to elicit problematic behavior.
- AI systems could be designed for interpretability, perhaps allowing developers and supervisors to directly read their thoughts, and to directly understand how it would behave in a wide class of scenarios.

Thus, we suspect that an adequate solution to AI alignment could be achieved given sufficient time and effort. (Though whether that will actually happen is a different question, not addressed since our focus is on feasibility rather than likelihood.)

Note also that if we *don't* make substantial progress on the alignment problem, but still keep building more AI systems that are more capable and more numerous, this could eventually lead to permanent human disempowerment . In other words, if this particular step of the argument doesn't go through, the alternative is probably not a business-as-usual human world (without the possibility of stable institutions), but instead a future where misaligned AI systems are ruling the world.

(For more, see section 5 .)

0.2.2 Preventing drift

As mentioned in section 0.1, digital error correction could be used to losslessly preserve the information content of values. But this doesn't entirely remove the possibility of value-drift.

In order to pursue goals, AGI systems need to learn many facts about the world and update their heuristics of how to deal with new challenges and local contexts. Perhaps it will be possible to design AGI systems with goals that are cleanly separated from the rest of their cognition (e.g. as an explicit utility function), such that learning new facts and heuristics doesn't change the systems' values. But the one example of general intelligence we have — humans — instead seem to store their values as a distributed combination of many heuristics, intuitions, and patterns of thought. If the same is true for AGI, it is hard to be confident that new experiences would not occasionally cause their values to shift.

Thus, although it's not clear how much of a concern this will be, we will discuss how an institution might prevent drift *even if* individual AI systems sometimes changed their goals. Possible options include:

- Whenever there's uncertainty about what to do in a novel situation, or a high-stakes decision needs to be made, the institution could boot-up a completely-reset version of an AI system (or a brain emulation) that acts according to the original values.
    - This system will have had no previous chance of value-drift, and so only needs to be informed about anything that is a prerequisite for judging the situation.
    - In order to reduce contingency from how these prerequisites are learned, the institution could bring back multiple copies and inform them in different ways — and also let some of the copies opine on how to inform the other copies. And then have them all discuss what the right option is.
- AI systems designed to execute particular tasks could be motivated to do whatever the more thorough process would recommend. They could be extremely well-tested on the types of situations that most frequently come up while performing that task.
    - For any tasks that didn't require high context over a long period of time, they could be frequently reset back to a well-tested state.
    - If the task did require a larger amount of context over a longer period of time, they could be supervised and frequently re-tested by other AI systems with less context. These may not be able to correctly identify the value of the supervisee's every action, but they could prevent the supervisee from performing any catastrophic actions. (Especially with access to transparency tools that allow for effective mind-reading.)
- Value drift that is effectively random could be eliminated by having a large number of AI systems with slightly-different backgrounds make an independent judgment about what the right decision is, and take the majority vote.

Some of these options might reveal inputs where AI systems systematically behave badly, or where it's not clear if they're behaving well or badly. For example, they might:

- endorse options that less-informed versions of themselves disagree strongly with,
- have irresolvable disagreements with AI systems which have somewhat different previous experiences,

- exhibit thought-patterns (detected with transparency tools) that show doubt about the institutions' original principles.

In most cases, it is probably the case that the reason for the discrepancy could be identified, and the AI design could be modified to act as desired. But it's worth noting that *even* in situations where it remains unclear what the desired behavior is, or in situations where it's somehow difficult to design a system that responds in the desired way, a sufficiently conservative institution could simply opt to prevent AI systems from being exposed to inputs like that (picking some sub-optimal but non-catastrophic resolution to any dilemmas that can't be properly considered without those inputs).

- An extreme version of this would be to prevent all reasoning that could plausibly lead to value-drift, halting progress in philosophy.
    - It doesn't seem impossible that all philosophically ambitious institutions would eventually converge to some very similar set of behavior, from a very wide range of starting points. This might be the case if some form of moral realism holds, or perhaps if something like Evidential Cooperation in Large Worlds works. If this were the case, our claim that it's feasible to stabilize many different value-systems would be false for philosophically ambitious institutions. It would only apply to institutions that refused to conduct some philosophical investigations. (Which we *hope* wouldn't be very common.)
- A further extreme would be for the institution to also halt technological progress and societal progress in general (insofar as it had the power to do that) to avoid any situation where the original values can't give an unambiguous judgment.
    - This would largely eliminate the issue motivating this subsection — that continual learning could lead to value drift — since complete stagnation wouldn't require much in the way of continual learning.
    - But depending on *when* technological progress was halted, this could limit the institution's ability to survive in other ways, e.g. by preventing it from leaving Earth before its doom.

Given all these options, it seems more likely than not that an institution could practically eliminate any internal sources of drift that it wanted to. (For more, see section 6 .)

0.3 Preventing disruption

So let's say that it will remain mostly-unambiguous what an institution is *supposed* to do, in any given situation, and furthermore that the institution will keep being motivated to act that way.

Now, let's consider a situation where this institution — at least temporarily — has uncontested military and economic dominance (let's call this a "dominant institution"). Let's also say that the institution's goals include a consequentialist drive to maintain that dominance (at least instrumentally). Could the institution do this? On our best guess, the answer would be "yes" (with exceptions for encountering alien civilizations, and for the eventual end of usable resources).

Any resources, information, and agents necessary for the institution's survival could be copied and stored redundantly across the Earth (and, eventually, other planets). Thus, in order to prevent the institution from rebuilding, an event would need to be global in scope.

As we argue in section 7 , *natural* events of civilization-threatening magnitude are rare, and the main mechanism they have to pose a global threat to *human* civilization is that they would throw up enough dust to blot out the sun for a few years. A well-prepared AI civilization could easily

survive such events by having energy sources that don't depend on the sun. In a few billion years, the expansion of the Sun will prevent further life on Earth, but a technologically sophisticated stable institution could avoid destruction by spreading to space.

As we argue in section 8, a dominant institution could also prevent other intelligent actors from disrupting the institution. Uncontested economic dominance would allow the institution to manufacture and control loyal AGI systems that far outnumber any humans or non-loyal AI systems. Thus, insofar as any other actors could pose a threat, it would be economically cheap to surveil them as much as necessary to suppress that possibility. In practice, this could plausibly just involve enough surveillance to:

- prevent others from building weapons of mass destruction,
- prevent others from building a competitive institution of similar economic or military strength, and
- prevent others from leaving the institution's domain by colonizing uninhabited parts of space.

The main exception to this is alien civilizations, which could at first contact already be more powerful than the Earth-originating institution.

Ultimately, the main boundaries to a stable, dominant institution would be (i) alien civilizations, (ii) the eventual end of accessible resources predicted by the second law of thermodynamics, and (iii) any disruptive Universe-wide physical events (such as a Big Rip scenario), although to our knowledge no such events are predicted by standard cosmology.

0.4 Some things we don't argue for

To be clear, here are two things that we *don't* argue for:

First, we don't think that the future is *necessarily* very contingent, from where we stand today. For example, it might be the case that almost no humans would make an ultra-stable institution that pursues a goal that those humans themselves couldn't later change (if they changed their mind). And it might be the case that most humans would eventually end up with fairly similar ideas about what is good to do, after thinking about it for a sufficiently long time.

Second, we don't think that extreme stability (of the sort that could make the future contingent on early events) would necessarily require a lot of dedicated effort. The options for increasing stability we sketch in sections 0.2.2 and 6 and the assumption of a singleton -like entity in sections 0.3 and 8 are brought up to make the point that stability is feasible *at least* in those circumstances. It seems plausible that they wouldn't be necessary in practice. Perhaps stability will only require a smaller amount of effort. Perhaps the world's values would stabilize *by default* given the (not very unlikely) combination of:

- technological maturity[1] (preventing new technologies from shaking things up),
- human immortality (reducing drift from generational changes),
- the ability to cheaply and stably align AGI systems with any goal, and

1 The point of time where civilization has invented all major technologies (Bostrom, 2013). For this use-case, the relevant definition of "major" would only need to include technologies whose invention could precipitate large power-shifts.

- such AI systems being equally good at pursuing instrumental goals regardless of what terminal goals they have. (Thereby mostly eliminating the tendency for some values to outcompete others, c.f. decoupling deliberation from competition .)

0.5 Structure of the document

- In section 1 , we clarify our assumptions about AGI , discuss what we mean by "stability" and "lock-in" , and clarify how confident we are that societies could be made how long-lasting .
- In section 2 , we give some remarks on the desirability and probability of highly stable institutions.
- In section 3 , we go through some sources of instability from the past, and how some of them seem less likely in the future.
- In section 4-8, we go through the arguments sketched in the summary in more detail.

Readers should feel free to skip to whatever parts they're interested in.

1 To be more precise

Overview: In this section, we go into a bit more detail about what we assume, and what it is that we argue for.

1.1 What do we assume about "AGI"?

Throughout this document, we assume that artificial general intelligence (AGI) is available. To be more specific, we assume that AI is sophisticated enough that it is possible to build an AI system that can perform all tasks at least as well and at least as cheaply as any particular human.[2] Our focus is on approximately human-level intelligence, as opposed to superintelligence, since we have a better understanding of human-level intelligence, and since superintelligence doesn't seem important to the arguments.

We do not argue that highly stable institutions would necessarily be feasible immediately after AGI is invented, because it could take additional time to develop AI with the right stabilizing characteristics. But neither are we restricting ourselves to claims about what AI technology is *possible* or *conceivable* . Instead, we discuss AI technology that we expect could reasonably be developed if there were well-resourced actors who wanted to invest in stability-increasing technology.

Conversely, neither do we mean to imply that all these capabilities would be *necessary* for extreme stability. It seems plausible that AI that is substantially subhuman in some areas (or just the right combination of non-AI technologies) could be enough.

2 What tasks are "relevant" or not is hopefully clear from the text. All cognitive tasks and all economically relevant robotic tasks seem sufficient, but not necessary.

A note on anthropomorphism:

In this document, we occasionally talk as if AI of the future will largely take the shape of fairly generally intelligent "agents". But the future could also consist of computational structures with few discernible individuals, with many computations being in-between something as agent-like as a human and something as tool-like as google maps. The reason we mostly talk about agents is that:

- it is clearly possible to build societies and institutions out of agent-like beings (as demonstrated by human society),
- the analogy to humans helps clarify how agents might act and interact with each other, and
- even without the analogy to humans, agents are easier to reason about since we can use the intentional stance and assume that they'll do (easy) things that are helpful to accomplish some goal (instead of having to reason about their internal structure).

The fact that there are many ways to use AI that we do not cover means that it is *more* likely that extremely stable institutions will be *feasible*, since if someone was bent on building a stable institution, they would have strictly more ingredients to choose from. It seems likely that non-agent AI would be useful for increasing stability, since their limited scope could make them more predictable to their creators.

1.2 What do we mean by "lock-in"?

In the summary, we avoided using the term "lock-in". In this section, we will give our take on how to best define and reason about "stability" and "lock-in".

When we say that an institution or some other feature of the world is very stable, we mean that it has a very high probability to persist for a very long time. When evaluating this probability, we don't want to appeal to fully objective probability, but we also don't want it to be entirely subjectivist. Instead, we're appealing to a pseudo-objective probability evaluated from a perspective much more informed than ours, but which still treats effectively-unpredictable processes (such as the weather or the exact DNA of future humans) as random. We could imagine it as the prediction that a hyper-informed observer would give, if they'd seen the civilisational trajectory of millions of civilisations similar to ours. The virtue of this definition is that we can say things like "this institution is 50% likely to be very stable", if we think there is a 50% subjective probability that the institution has a high pseudo-objective probability to persist for a long time.

We could also talk about institutions that are stable in the face of some things, but not others. For example, an institution's values could depend on some people's decisions, while allowing for no other route to be changed. If so, it would be stable in the face of everything but those people's decisions.

What do we mean by "lock-in"? In our context, we will mostly talk about "lock-in" of (at least somewhat) important features of humanity's long-term future. 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.

Human extinction is one example of a lock-in event. Another example would be if many independent groups send colonization waves in different directions and global coordination (like a singleton) becomes permanently infeasible.

A type of lock-in that we're especially concerned with is 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 civilization will adopt values from outside that subset.

Permanent global value lock-in would last until the end of that civilization. But it is often useful to also talk about global value lock-in that ends earlier, as long as it lasts for at least millions of years and substantially affects how valuable humanity's long-run future is. This could for example include a scenario where Earth-originating intelligence encounters alien civilizations after 1 billion years, and value lock-in lasts until then.

Note that this definition is different from some other definitions of lock-in. The Precipice (Ord, 2020) introduces lock-in as something that happens when "Key aspects of the future of the civilization (...) are almost impossible to change". This definition relies on whether some *entity* has the *ability* to make a change, and thus relies on specifying which entities we should consider to have what options. This is often difficult, e.g. in the context of entities created to disprefer certain changes, which is why we do not use this definition. However, our definition also has some downsides, such as not distinguishing between (i) a rash decision to lock-in an individual's current best guess of what society should look like, from (ii) a thoughtful society which is unlikely to change their mind given how much they've already reflected on their values, but who remain open to changing their mind.

1.3 How confident are we that stability is feasible, for how long?

- Among the arguments in the summary, components 0.1 and 0.3 seem solid.
- Component 0.2 (both 0.2.1 and 0.2.2) seems somewhat less certain. There are many arguments for why AI systems could be made with more stable values than humans, and we don't know any particular, credible objections to total stability. But we don't have arguments that knowably identify and exclude all possible sources of drift.
- We're especially confident that it would be possible to build a civilization that was stable for millions of years on Earth, since this would be compatible with complete stagnation. If a civilization doesn't mind sacrificing things like economic growth and technological progress, then it also seems quite easy to avoid value changes.
- Being stable for billions or trillions of years would cause additional challenges of:
    - Possibly encountering alien civilizations.

- Space travel (to avoid the end of the Earth) would require some significant amount of technological progress from our current state.
- Unpredictable physical events — the universe has only existed for 13 billion years so far, we've only been studying it scientifically for a few hundred, and we're not *that* confident about what will happen in the next trillion.

- Out of these, alien civilizations seem most likely to upset an otherwise locked-in scenario. Except for that, we think it seems more likely than not that a trillion-year stable society is possible, and that the arguments fairly robustly point towards them being at least plausibly possible (say, that their feasibility is worth at least 20% subjective probability).

2 Desirability and probability

Overview: This section briefly mentions some reasons for why the possibility of stable institutions could be good or bad; and then brings up some reasons for why people may choose to build them.

2.1 Would it be good for highly stable institutions to be built?

Some lock-in events could constitute existential catastrophes, e.g. locking in bad values.[3] These are important to avoid.

But there are nevertheless some things which might be good to make predictably true for a very long time (e.g. we might want to enshrine some basic human rights). Indeed, some degree of stability may be necessary to permanently preclude bad values from eventually being locked-in.[4]

In particular, it seems plausibly good for an institution to be stable in the face of everything except for a few endorsed sources of change, such as democratic voting or moral reflection. Doing this might require some of the same procedures that would be used to lock-in more specific values, while it would simultaneously be important to avoid any parts that would prevent progress or otherwise be insufficiently flexible.

2.2 How likely is this?

For some previous discussion on the likelihood of value lock-in see Christiano (2014), Hanson (2018), Tomasik (2020), Karnofsky (2021), and chapter 4 of What We Owe The Future (MacAskill, 2022) (and a response by Hanson (2022)).

3 Values could be bad in a number of ways; perhaps because they include actively distasteful components, or perhaps because they fail to recognise some intrinsically valuable things as worth creating (and are insufficiently flexible to realize their mistake).

4 C.f. footnote 127 in The Precipice (Ord, 2020), which suggests a principle of locking-in the minimisation of lock-in. Though note that this uses a different definition of lock-in, based on what options people have available to them rather than being based on what will predictably happen.

This document is mainly not about the likelihood of (various types of) lock-in. But in this section, we briefly go over some relevant arguments.

One key consideration is that, if there's some fixed probability of entering a stable state every year, then over long enough time periods, we should expect societies to eventually end up in one. Given the long time-periods involved, this seems like a solid reason to expect society to eventually stabilize *some* features of itself. In fact, if you grant that there are many stable states that currently don't seem *too* unlikely, then if we ever end up in a state where those states are *permanently, very unlikely*, that suggests that at least some minimal stabilization has already happened.

There are also reasons to believe that at least some humans will be motivated to achieve some significant degree of stability:

- Some past authoritarian leaders seem to have desired stable influence over the future.[5]
- Given *any* preference about what should happen in the future, a group of humans could be tempted to make sure that the future didn't stray too far from those ideals. This could apply to many different kinds of groups. For example, a libertarian group might seek to create very long-lasting constraints against interfering with individuals' freedom.
- Some groups of humans may be extremely uncertain about what they want the long-term future to look like, such that they could want to defer to the deliberation of future people. Even so, they might want to create stably favorable conditions for this deliberation (which should perhaps be stable in the face of everything but the results of the deliberation).
- There are some catastrophic outcomes that many groups would agree should not be allowed to happen even once. For example; people might want to permanently prevent the use of weapons so powerful that they could cause the collapse of civilization (Bostrom, 2019). Similarly, in a world where global value lock-in is possible, it could be important to permanently stop bad, inflexible values from becoming locked-in.

Of course, stability may not come for free. Even if people wanted to make certain features of the world highly stable, doing so may be too costly to be worthwhile.

However, there are reasons to believe that a post-AI society could be incredibly rich (Davidson, 2021 ; Trammel & Korinek, 2020). This could make stability a more attractive purchase for any values that have diminishing marginal utility to resources. For example, stability might be more tempting if it doesn't interfere with everyone on Earth leading an extremely materially rich life. (C.f. Aschenbrenner (2022) arguing that increased wealth incentivises greater investments in reduced existential risk.) What about values that don't have diminishing marginal returns to resources? If they also don't endorse a discount rate over time, their main concern would presumably be to ensure that most resources in space are used well, which would make long-term stability a critical purchase.

Difficulties from coordinating and agreeing with other actors seems like a significantly larger obstacle than economic costs. But speculatively, the technologies that enable lock-in might increase nations' ability to cooperate with each other. In particular, humans or institutions who

---

5 As one example, the Egyptian pharaoh Akhenaten used his reign to stop the worship of Egyptian gods other than Aten; which included some attempts at erasing other gods' names and the building of monuments with names like "Sturdy are the Monuments of the Sun Disc Forever". After his death, traditional religious practices gradually returned and many of the monuments were demolished — but perhaps Akhenaten would have prevented this if he could have enforced stability. As another example, Nazi Germany was sometimes called the " Thousand-Year Reich ".

have significant value disagreements, but who are willing to compromise on some goals that represent all of their interests, might be able to collectively design AI systems that embody those interests. They can then empower those systems (e.g. with appropriate military capabilities) to enforce that compromise in the future. In other words, the ability to create AI systems with arbitrary goals could provide a solution to commitment problems.[6]

It's also worth noting that an institution would not need to be perfectly stable from the beginning. Instead, it could gradually increase its stability over time.[7] This makes it easier to start a lock-in process for three different reasons:

- It would not be necessary to pay the entire *cost* for stability up-front. Initially, a dominant actor might just invest enough in stability that it could expect to stay stable for a few decades. Throughout those decades, it could gradually invest more in stability, perhaps extending the expected lifespan to a few hundred years, etc...
- It would not be necessary to have all *technology* necessary for total stability from the beginning. Knowledge and experience of how to design e.g. increasingly stable AGI could accumulate over time.
- It would not be necessary to decide exactly *what* to lock-in up-front. A group of actors who were initially reluctant to fully lock-in some specific set of values could start by making society very stable in other ways. As they reflected more on what they wanted, over hundreds of years, they might then gradually become more and more confident about their values, making it less and less likely that they would ever deviate far from them.
    - In these cases, there might not be any sharp point where "value lock-in" happens, but the eventual outcome could nevertheless be highly stable values.
    - If all of humanity participates in decision-making, this might look like achieving existential security and then deliberating on what to do with the universe during a long reflection (as discussed in e.g. The Precipice (Ord, 2020)). But the same dynamic could also apply in situations where a smaller group holds power.

One final reason for why we might expect some initial investments in stable institutions, in a post-AGI world, is that they may be necessary for human decision-makers to maintain influence and even just survive throughout their natural lifespans. The reason for this is that AGI could make technological and societal change happen exceptionally quickly:

- AGI could do cognitive work much faster than humans. On a pure hardware-level, transistors can send signals orders of magnitude faster than human neurons can send signals.
- Major power-shifts and wars are largely caused by social processes and by technological progress, both of which would be accelerated by faster cognition. If the rate of such events were proportionally sped up, biological humans would see a lot of turmoil during a single year.[8]

---

6 Commitment problems are one of three possible reasons for why rational agents might engage in conflict, listed in Fearon (1995) .

7 This is related to the concept of *regime consolidation* , see e.g. democratic consolidation or authoritarian consolidation (Göbel, 2010) .

8 C.f. Chapter 27, subsection *Humans* , in Age of Em (Hanson, 2016) : "This protection of human assets, however, may only last for as long as the em civilization remains stable. After all, the typical em may experience a subjective millennium in the time that ordinary humans experience 1 objective year, and it seems hard to offer much assurance that an em civilization will remain stable over 10s of 1000s of subjective em years."

- Thus, if human decision-makers wanted to preserve their power and be safe from violent conflict just *during their own lifetimes*, they might have to restructure things to be significantly more stable than they would be by default.
    - It is also possible that radical new biotechnologies could make human maximum life-spans far longer, which could further increase selfish benefits to stability.

Such initial investments could require society to pay many largely-fixed costs of increasing stability, such as central enforcement of international agreements, or restructuring governments to allow for easy, continuous vetting of essential AI systems. With such fixed costs already paid, it might be cheap to increase stability by a few more orders of magnitude, if this was desired. (Though there might be other factors that cause the marginal cost of stability to increase, rather than decrease.)

3 Past sources of instability

Overview: In this section, we discuss why the governments of sovereign states[9] have failed to be stable in the past, and why those reasons might not play as much of a role in the future. These sources of instability aren't meant to be mutually exclusive or exhaustive. Instead, they're meant to show that past instability shouldn't be taken as strong evidence that stability will be impossible in the future, as well as to illustrate the relationship between our main arguments (see the summary) and things that have caused instability in the past.

3.1 Foreign intervention

States can be destabilized by interventions from people outside the state.

- The most central example of this occurs when states are invaded by other states.
    - For example, Hitler's attempt at a "Thousand-Year Reich" ended after just 12 years, when Germany lost World War 2.
    - The Aztec empire lasted for about a century before being invaded by the Spanish conquistadors and their native allies.
- States can also be destabilized by external actors in more subtle ways.
    - Bryan Caplan argues that the existence of non-totalitarian societies where people were obviously "richer and happier" caused party-members in communist Soviet to lose faith in their own system. (Caplan, 2008.)
    - For a while, when Japan was quite isolated, it almost completely suppressed the use of guns. But when a US fleet visited in 1853, the Japanese government was motivated to reequip with firearms, presumably due to fearing conflict with the well-equipped outsiders. (Diamond, 1997, p. 258)

These effects would not be relevant if the world was highly unified, such as if a world government was formed, in the future. This seems like a real possibility. In general, the world is more

9 In a broad sense, including historic empires, city-states, etc.

interconnected and interdependent than it's ever been, with more common culture, more communication, and more trade. And a world government hasn't been technologically feasible for very long, so its absence so far doesn't tell us that much about its likelihood in the future.

For the purposes of our argument, we only need to claim that it would be *feasible* to establish a world government (or something similarly expansive), if a large-enough part of the world wanted to.

3.2 Aging and death

Authoritarian regimes that seek stability face a succession problem — when one leader dies and hands over control, there's a chance that future leadership will have different views of how to run things. For democratic states, continuous generational replacement is one source of continually changing views. Slightly more generally, we can also consider values-changes and leadership replacement brought about by aging (as separate from dying).

It's unclear how large this effect is in practice. Bryan Caplan argues that the problem of succession was by far the biggest source of ideological change for communist Soviet and China (Caplan, 2008). On the other hand, an analysis by Anders Sandberg (published results forthcoming, though see some comments here) indicates that the average dictator would only be in power for 4 years longer if aging was eliminated.

None of this would have to play a role in the future. Medical advancement (perhaps precipitated by AGI causing technological progress) could eliminate aging among humans. And even if it doesn't — AGI systems themselves wouldn't need to die. They wouldn't age, and if they were copied and stored redundantly in many places, they would also not be vulnerable to accidents or assassinations.

3.3 Technological or societal changes favoring new values

As new technologies (in a broad sense, including e.g. new methods of social organization) are developed, this may cause values to change. For example:

- New technologies might favor certain types of values or norms, over time causing societies to switch to those new values. (E.g. because people or sub-groups with such values tend to accumulate influence, and because people tend to switch to more useful practices.)
    - As an example of this, the invention of agriculture led greater groups of people to live close to one another, which favored more hierarchical systems of political organization. This probably caused a shift from values that favored egalitarianism to values that were more accepting of hierarchies.
- New technologies might change how people communicate with each other, which might directly influence what sort of ideologies can easily spread through society.
    - For example, the printing press played a significant role in the spread of Protestantism (Rubin, 2014).

Since these changes are caused by the inventions of new technologies, they would no longer play a role once society reaches technological maturity (the point at which all major technologies have been invented, see Bostrom (2013)).

But these dynamics could become obsolete even before then. In particular, the variations described above only really matter insofar as the starting conditions for all coalitions are sufficiently equal that their values can make a substantial difference for who would win a fight, or who would grow fastest. But if one coalition is already dominant (and has access to arbitrary numbers of aligned agents), it's implausible that any technology would give a value-dependent advantage that was large enough to enable some previously powerless faction to overpower the dominating coalition.[10] (See section 8 .)

Given a solution to alignment, we also expect values to constrain AGI's behavior much less than values constrain human's behavior.[11] (C.f. Christiano (2022) .)

3.4 Internal rebellion

Many states have been defeated by coups or revolutions from people within the state.[12] We can separate these into scenarios where either:

- some essential supporters of the state (such as the military or security forces) launched the coup, or allowed the coup to carry on when they could have stopped it,
- the rebellion was feasible even when consistently opposed by supporters of the state.

If the regime's reliance on human supporters were eliminated, and critical functions were instead automated by AGI systems aligned with the institution's goal, the first of these possibilities would be eliminated. See section 5 and section 6 .

The second possibility would not be plausible if there was a sufficiently large power differential between the regime and its population. See section 8 .

4 Preserving information, and baseline stability

Overview: This section argues that:

- With AGI, it will be possible to accurately preserve even quite complex values, by storing large amounts of data, including whole minds.

---

10 Note, however, that technological changes could force an institution into awkward choices. For example, consider a system of values with preferences that include:

- maintaining power,
- obeying the desires of some human population,
- not interfering too much with that population's beliefs or actions (nor interfering with the technologies that influence them).

If a technological change would, by default, cause the human population to want to replace the institution with something else, then the institution would have to deprioritize one of these values.

11 In order for humans to behave in particular ways, it can be necessary for them to adopt specific beliefs and values that justify that behavior in the light of other human values. (For example, monarchies may be more stable if people adopt religious beliefs and values that justify the monarch's rule, because otherwise egalitarian instincts may cause people to object.) Whereas AGI systems could simply share the intended goal, and thereby be motivated to behave in the desired way.

12 Though coups can also increase stability, by stopping leaders from making changes.

- By using digital error correction, it is possible to virtually eliminate the problem of small-scale hardware errors during storage, communication, and computation.
- Larger-scale errors (such as a data-center being destroyed by an earthquake) can be made harmless by storing relevant information in many different places.

This means that institutions can easily achieve very high baseline levels of stability that can only be interrupted by worldwide catastrophes or intelligent action.

4.1 Preserving complex goals

An important part of our hypothesis is that it won't just be possible to lock-in some narrow range of values. Instead, it will be possible to lock-in a wide variety of values, including highly nuanced and complex ones.[13]

As an analogy, consider that the invention of writing significantly boosted civilizational memory. Accordingly, today's world is more influenced by past people who had access to writing — such as the authors of the Bible — than people who came beforehand. But writing still has relatively low bandwidth. Even if we wanted to, we can no longer determine what the Bible's author would have thought about novel moral dilemmas, such as those introduced by new technologies. They simply didn't record enough details of their views.

By contrast, in the future, values could be directly stored *in minds*. Plausibly, whole-brain emulation (WBE) (see Sandberg & Bostrom (2007)) will be invented soon after AGI. If so, then it would be possible to preserve entire human minds, and query them about their views at any level of detail. But if not, it would still be possible to preserve non-WBE AGI minds. If they're supposed to store some particular concept, they could spend a lot of time learning those concepts and talking with the institution's stakeholders about exactly how it should be interpreted in a wide variety of edge cases. This could include a strong understanding of the purpose that the institution serves, as well as a huge amount of stored data about the institution's original stakeholders, that could be used to retroactively predict what they would have thought about various scenarios.

4.2 Digital error correction

By using digital error correction, it is possible to virtually eliminate the problem of small-scale hardware errors during storage, communication, and computation. (Normal hardware is already extremely reliable, so this is rarely used in practice — but occurs in some especially important applications.)

A basic version of doing this is redundancy. If data is repeated multiple times, the chance that random local damage changes the majority declines exponentially with the number of copies.

For communication and storage, more efficient error correction codes can improve this by a lot. Given some assumptions, the noisy channel theorem implies that if you're transmitting (or storing) sufficiently long messages, you can get an arbitrarily small probability of error with just a

13 But not all values. In particular, some values might explicitly disvalue being locked-in, or might prescribe certain decision-making procedures that cannot remain stable for long periods of time.

small constant multiplicative factor of additional transmission (or storage).[14] The theorem doesn't directly say how to construct codes with these properties, but the best codes of today are approaching the theoretical limit.

In order to do error-correcting *computation*, you also need a way to prevent errors from accumulating over many serial manipulations. The simplest way to do this is again to use redundancy: break the computation into multiple parts, perform each part multiple times on different pieces of hardware, and use the most common output from one part as input to the next part.[15] This already scales quite well. A linear increase in hardware expenditure exponentially decreases the error rate, and can therefore make up for a multiplicative increase in the number of operations. So in order to keep the error-rate constant, when the number of operations w increases, the amount of hardware needed grows proportionally to log(w). Thus, the number of operations in the error-corrected version of the computation are proportional to $w \log(w)$, to be compared with w for the uncorrected version, which is manageable even for huge w (von Neumann, 1952; Pippenger, 1990). Furthermore, there are schemes that improve on this.[16]

Crucially, all these error-correcting methods assume that the stored or communicated numbers are *discrete* (as opposed to continuous), and make some assumptions about how different errors are *independent*.

- Today, almost all AI systems are run using discrete numbers, and there's no sign that this will change in the near future.
    - It is conceivable that analog computation could be more efficient, for some purposes, but discrete computations can always be used to simulate analog computations to arbitrary precision, so regardless, it should be possible to build digital AGI.
    - Digital error correction for quantum computing isn't as well developed as for non-quantum computing, yet. But we know of no credible reason for why AGI would necessarily need to run on quantum computers. In particular, quantum phenomena do not seem to play a large role in human cognition. (See footnote 75 in Carlsmith (2020) for some discussion.)
- Independence of errors seems like a fair assumption for small-scale hardware-level errors, but can't be assumed for larger scale interference.
    - For example: if a meteor hits a computer — affecting all bits simultaneously — then no way to represent information within that computer could preserve the information.

---

14 The theorem defines a channel's capacity as the number of bits of information that can be transmitted through the channel per (noisy-)bit of transmission (taking into account that the noise will destroy some information). Given certain independence assumption, it then says that, as long as the source's information rate is less than the channel's capacity, then for sufficiently long messages, you can get an arbitrarily low probability of error by transmitting n+m bits, where n is the number of bits in the original message, and m is the *memory* of the channel. A memory of m bits implies that each transmitted bit is only dependent on the m previous bits. (Macon, 2015)

15 Of course, the operation that finds the most common output can itself suffer errors, but the procedure can be done in a way such that this is unlikely to happen for a large fraction of the hardware units.

16 For example, if you have a large number of parallel processors, a poly-logarithmic increase in the number of processors can cause the error to be exponentially small in the number of processors (as opposed to being exponentially small in the number of redundant processors that you replaced each old processor with) (Spielman, 1996).

For a calculation of how much extra storage space you might need for billion-year stability, see the response to this comment .

4.3 Baseline stability via redundancy

Digital error correction solves the problem of low-level, undetectable errors. In order to interfere with information storage or processing, an event needs to cause many errors simultaneously.

But such events are typically rare. They also tend to be noticeable — either because they are caused by a visible incident (someone dropping a computer, or a local earthquake), or because they immediately causes the system to malfunction (as opposed to causing a surreptitious mutation that is only rarely relevant.) This is especially true if the system has been designed to flag suspicious events.

This means that such events can be mitigated by storing all relevant information in many disparate geographical locations (across Earth or, eventually, across space); replacing one at a time when something happens to it. In order for all information to be lost, destruction would have to be synchronized across the world.

This point generalizes from information storage to preservation of other features of a stable institution. Since software can be copied, an institution with enough resources could have all information, all AI systems, and all resources it needs to recover spread across the Earth (and eventually, across space). Thus, the only thing that could destroy it would be either global catastrophes or intelligent action.

5 Aligning with goals

Overview: This section discusses the feasibility of (and meaning of) creating AGI that is aligned with an institution's goals. If an AI system is aligned with an institution's goal, it should be competent at taking actions that achieve that goal, and any deviations from that should be accidental mistakes rather than actions optimized to defy the goal. Thus, given the high baseline stability mentioned above, the institution would not be destabilized by systems that are aligned with it. We argue that it would be technically feasible to design AGI systems that are aligned enough to enable highly stable institutions, if sufficient effort was dedicated to the problem.

5.1 What do we mean by "goals"

When we discuss an AGI optimizing for a particular goal, we mean that they systematically pick actions that are good at achieving that goal. (C.f. the intentional stance (Dennett, 1987), or Yudkowsky (2008)).

This doesn't mean that they never act suboptimally, with regards to the goal. Instead, it just means that:

- On average, they are good at achieving the goal. When it's obvious what action will best further the goal, they will take it; and even when it's difficult to tell, they will mostly take one of the

better actions that are available to them. They are at least as good at this as an exceptionally competent human who was highly motivated to accomplish that goal.

- When they act suboptimally, their actions are *accidental mistakes*. In particular, this means that they never take actions that are *optimized* for destabilizing the institution that they are aligned with. (In the context of alignment, a similar property has been referred to as avoiding malign failures, while acknowledging that there may be benign failures (Christiano, 2018a).)

The first of these conditions means that the institution will continuously take competent actions to increase stability (insofar as that is what their goal prescribes), being at least as competent and creative as a competent human. This will make their institution highly stable by default, as argued in the previous section.

Then the second of the conditions means that the institution will never face optimized destructive actions from agents who share its goal. Given the default stability, this means that **aligned** agents would be exceptionally unlikely to ever destabilize the institution.[17]

In the remainder of this section, we will discuss whether it is possible to make aligned AGI.

5.2 Alignment

One straightforward way to build AGI that shares some goals with its creator(s) would be to use human brain emulations. This would work well to transfer human goals in great detail.

However, two issues with this would be:

- That humans have partly agent-relative goals. This means that copies of a human brain could end up in conflict with each other, as each copy would pursue e.g. *its own* happiness.
- That humans display many unnecessary forms of value-drift.

It may be possible to modify brain emulations to more wholeheartedly dedicate themselves to the main goal, and to drift less; or to compensate for these failings with strong institutional control. (For discussion of this, see Shulman (2010), a response by Hanson (2011), and also the book Age of Em (Hanson, 2016).)

Ultimately, however, it seems like designing novel systems should provide an even greater degree of control over the system's goal. We also think that novel AGI will come earlier than brain emulations. So while much of our discussion could apply to modified brain emulations, our main focus will be on newly designed AGI.

As a basic observation, it seems that the designer of an AGI system should have great influence over its goals. However, the project of building AI agents that do what their principal wants them to do has already received a fair amount of attention (under the title of the *alignment problem*), and efforts so far have yielded many arguments that the problem could be very difficult. (See e.g. Cotra (2022), Cohen et al. (2022), Hubinger et al. (2019).)

---

17 There are some small number of actions that could destabilize the institution without being optimized towards that, e.g. doing fundamental physics research and accidentally stumbling onto vacuum collapse. But these actions seem few and far between. The institution could be exceptionally careful around this category without much expense, and even if they weren't, it's not clear that they would add up to much total risk over time.

However, in order to enable lock-in, people would only need to solve a relatively simple version of the alignment problem. In particular:

- They wouldn't need to align significantly superhuman AI systems, since human-level AI seems sufficient for lock-in.
- The aligned AI systems wouldn't need to be competitive with unaligned AI systems, because a dominant institution could prevent unaligned AI systems from being used. Without the necessity of being competitive, alignment strategies could make use of highly inefficient methods, such as spending large amounts of time and resources on thoroughly understanding an AI system's cognition (using the best interpretability tools available).
- They wouldn't necessarily need to create AI systems with values that were *identical* to their own, as long as they were similar enough.
    - For example, if some group of humans wanted to lock-in some combination of their values, handing over control to an AGI system that *mostly* shared their values (and that would itself never have to do an imperfect hand-over again) could be preferred to leaving them up to many future generation of humans with imperfect transfer between every generation.
    - Similarly, for a system that e.g. tried to prevent the use of WMD, it might not be important that it handles every edge-case exactly like its creators would have, as long as it is still stops all egregious uses of WMD (and that it interferes with very few normal human activities).

We think that this simpler version of the alignment problem is likely to be solvable, given enough time and investment.

Of course, if the alignment problem isn't solved before AGI is widely deployed, there may not be time to solve it afterwards — since misaligned AGI could lead to an existential catastrophe soon thereafter (see e.g. Bostrom (2014) , Carlsmith (2021) , Karnofsky (2022)). We won't discuss whether stable institutions would be possible in such a world — after AI takeover. Instead, we'll just note that *if* alignment turns out to be impossible (or if it's possible but we don't solve it in time, which seems much more likely) this would probably not lead to a business-as-usual human-centric world without stable institutions. Instead, we think it's more likely that it would end with misaligned AI systems ruling the world.[18]

Given this background on goals, we will now turn to discuss how goals could be made stable.

6 Stability of goals

Overview: Even given the ability to create aligned AGI, there's a question about whether they have some probability of becoming misaligned over sufficiently long time-frames. In particular, AGI systems would by default need to continue learning about the world, and thereby make some changes to themselves, and it's hard to guarantee that this would never change systems' values. We can't say for sure that this won't happen, but if it does, it doesn't seem like a big barrier to highly stable institutions, given abilities like being able to (i) frequently reset AI systems back to states that are known to be safe, (ii) have important tasks be carried out by multiple redundant

18 Though this isn't inevitable, since humans could coordinate to not build or deploy highly capable AI systems.

systems, (iii) have any novel or potentially-dangerous systems be supervised and thoroughly tested by systems that are known to be safe.

6.1 Why do humans' goals drift?

Let's first consider what typically causes humans' values to change:

- Humans' values and goals can change when their biological conditions change, such as if they are starved of food.
- Humans' values and goals predictably change in some ways as they age.
- Humans' values and goals can change in response to social pressure to adopt certain values.
- Humans' values and goals can change when they learn new information, or think about a new argument.

These processes are likely wholly or in-part present in humans because they were useful in humans' ancestral environment. Since humans do not intentionally pursue inclusive reproductive fitness, but instead pursue heuristic goals that correlate with reproductive success, it is not surprising that the right heuristic goals would change depending on humans' current biological conditions, social conditions, or age.

If humans were to develop AGI in-part via training it in some training environment, then insofar as developers had a good understanding of how an AGI's goals were formed in response to its training process, they could reduce or remove these reasons for goal instability. As a general point: if an intelligence could be made to *explicitly* optimize for something extremely similar to what it's being trained to do, it's much less clear why it would ever be selected for changing its goals.

6.2 Why might AGI goals drift?

So what are some fundamental reasons that AGI goals could drift?

In general, digital error correction only guarantees correct storage of data and correct execution of programs — it does not protect against unpredictable changes that arise from correct processing of data. And AI that is responsible for implementing its mission in the real world must undergo some changes. Pursuing good outcomes requires learning many detailed facts about the world and updating heuristics to effectively deal with new challenges and local contexts. Insofar as AI systems' goals are not cleanly separated from the rest of their cognition (just like human's values seem to be a distributed combination of many heuristics, intuitions, and patterns of thought), it is hard to be confident that such updating would not occasionally cause systems' values to shift.

A special case of this, which seems especially difficult to handle, is that most goals need to be specified in terms of other concepts that the AI system knows about. If the meaning of these concepts shift, then it's unclear what might happen to the goals. This is related to ontological crises (de Blanc, 2011) .

It's plausible that drift would be exceptionally rare. But since it's difficult to be confident of this, we will spell out how an institution might minimize *institutional* drift even if *individual systems* would have some risk of drifting.

6.3 Institutional drift

First, a disclaimer. Most of the above discussion (digital error correction, geographical redundancy, alignment) seem like sensible precautions regardless of what values an institution has. However, the institutional strategies that follow could be used to reduce *all* kinds of changes to values. This includes both sources of change that most actors would want to avoid *and* changes induced by reflection and learning more about the world, which many actors would (hopefully) *want* to be influenced by.

A second disclaimer is that some of the institutional strategies (such as frequently resetting AI systems back to earlier versions, or some types of testing) would have severe ethical problems if applied to human minds. AI systems could also be moral patients, which raises the question of whether an institution like this would be a moral catastrophe. We won't discuss this other than noting that:

- In general, dealing with possibly-sentient AI systems raises many ethical problems. It's almost certainly possible to cause huge harm by handling them incorrectly.
- What is harmful for AGIs with very different minds could differ *a lot* from what is harmful for humans, and it might well be feasible to design AGI systems for which it's harmless to run them in the ways outlined in this section.
- Regardless, this document isn't only trying to understand whether ethical actors could create stable institutions, but also whether there's a danger of unscrupulous actors doing it. Thus, the below discussion is important regardless of how the ethical issues shake out.

In the above section, we argued that AI systems interacting with and learning about the world could induce drift. If an institution wanted to minimize drift, they might therefore seek to reduce power granted to systems that have acted and learned things for a long time; and to instead maximize control held by relatively unchanged systems.

With digitally implemented minds, it is trivial to reset systems back to earlier versions as often as needed. But constantly resetting every system isn't compatible with acting competently in the world. For the purpose of acting competently without incurring too much risk, institutions might make use of two different techniques:

- Arranging the systems into hierarchies of supervision, where more reliable systems continually check that more-experienced systems aren't doing anything catastrophic.
- Continuously testing systems, as they learn more, to ensure that they're still loyal to the institution's goal.

Both of these approaches involve some systems verifying that other systems are (or would be) behaving well, by the lights of the institution. It seems like there are two different ways of doing this:

- The verifiers can aim to have a good understanding of what the goal is supposed to be, and check that the supervised or tested systems' own goals don't drift *too* far from the original.
- The interpretation of the institution's goals can be relatively centralized, and the verifiers could check that the supervised or tested systems are intent aligned (see Christiano (2018b)) with the central source — i.e., that it intends to do what the central source wants it to do.

If an institution has a very complex goal, it (speculatively) seems like they would favor the latter of these approaches to the extent compatible with communication constraints (e.g., the speed of light over interstellar or intergalactic distances), since:

- The institution could then afford to spend a lot of resources on getting the interpretation right.
- Supervisors could separately verify that systems lower in the hierarchy both (i) have decently accurate beliefs about what the interpreters would say about various situations, and (ii) that they're motivated to act according to those beliefs.
    - This decomposition seems nice since (i) has a clear source of ground truth, and (ii) is a more crisp and precise motivation than checking that their motivations align sufficiently well with a very complex set of values.
    - In addition, it's ok if an intent aligned system doesn't understand how the values apply to all situations, since if it encounters an unknown situation, it would be motivated to ask for clarification. This might not be true for a system which just (mostly) shares the institution's goals.
- If it's genuinely unclear how to apply values to some situations, a centralized source can make a judgment call that everyone will coordinate around, without impinging on other parts of the institution's values. Whereas if everyone different parts of the system maintain separate values, it seems more likely that AIs in different parts of the system would acquire slightly different values — perhaps resulting in conflict that causes much bigger problems.

Of course, there are also downsides with centralization, such as having a single point of failure. But this could probably be counteracted explicitly — such as by having multiple interpretative sources, and settling disagreements between them before the interpretation was used more generally.

For the purposes of our discussion, it's also helpful to separately discuss:

- how a stored goal could be interpreted in a wide variety of situations, and
- how an institution's actions could be stably aligned with those interpretations.

In the rest of this section, we will thus start with the first of these questions, follow it with a discussion of the latter, and finally give some remarks on how future institutions could continuously test their systems.

6.4 Interpreting goals

An institution that wanted to keep its goal highly stable would need to answer the question: Given the information that specifies the goal of the institution, how can the desired outcome be unambiguously determined in specific dilemmas?

This will obviously depend on what the original goal of the institution was. But one plausible scheme is to store the goal as one or more minds, which are able to take in the details of any situation, think about the decision, and then come to a conclusion. This could be a way to give ultimate decision-power to a group of (emulated) humans; but it could also be used to store an AI system that can interpret rules that the system was meant to embody. (E.g. what is meant by "preserving democracy".) When an unclear situation comes up, the judge(s) can be booted up from scratch, and asked to judge the situation.

This newly booted judge would not be at risk of value drift from any non-essential information processing. But there is some chance that there could be drift (or just ambiguity) introduced by information that *is* essential for judging the situation. If the world has changed a lot, judging a situation might require the judge to learn a lot of facts. And even if it hasn't, any given situation could be very complex, and there might be a lot of different options for how to summarize it, which could lead to different judgments.

This raises the question: When learning new facts about the world, and judging things according to them, what counts as drift and what counts as reacting appropriately to new information? If the world is changing a lot, how can old goals be applied to the new world?

The answer to this will depend on the goal. We can think of three different reasons for why we might suspect that an AI has suffered value-drift after being exposed to new information:

1. If the original version's values are clearly and unambiguously articulated in a way such that any sensible listener would recognise that the more-informed version's desires are in contradiction with it.
2. If there's some limited (but faithful) summary of a more-informed version's desires (or thought-process) such that a less-informed version, upon hearing that summary, would be confident that the more-informed version has deviated from a rule that they never wanted broken.[19]
3. If the judgment of the more-informed versions differ depending on what order they're exposed to various arguments, or are otherwise contingent on facts about the judgment process that ought to be irrelevant. (As judged by themselves.)

Notably, each of these three possibilities prescribes a concrete way to identify and prevent value-drift:

1. In this case, it should be fairly clear that value-drift happened. Drift could be identified by the algorithm in the next bullet points.
2. Insofar as less-informed systems can recognise drift, have more-informed interpreters be supervised and tested by less-informed versions of the same systems.
    - Perhaps by tasking some limited AI systems with the job of searching for accurate summaries of a situation that would cause a less-informed AI system to reject it.
    - Perhaps by letting less-informed AI systems thoroughly interrogate and understand (parts of) the more-informed system's reasoning. See section 6.6 for more discussion of this.
3. When making high-stakes decisions, always bring back many copies of the original AI system, and give each of them a different summary of the situation, recounting evidence and arguments in different orders.
    - Those copies could themselves have opinions on what summaries are appropriate to give to further copies of themselves, and they could discuss and compare their perspectives with each other.

---

19 Note that, unless every bit of crucial information is included, it's typically quite easy to give summaries where good decisions sound like terrible decisions. So in order for this clause to not always trigger, the process would have to be implemented with either (i) a strict definition of what "faithful" means, or (ii) the less-informed versions having a *very* high bar before they decide that the more-informed version *must* be wrong.

- Such a process may not remove *all* arbitrariness, but it could reduce it by a lot.

What would an institution do if some dilemma or piece of information systematically causes the supervisors to judge that the more-informed system has been compromised, or systematically causes irreducible disagreements between AI systems that gets exposed to it in slightly-different ways? An institution that was sufficiently afraid of value-drift could react to this by preventing their systems from engaging further. If *some* decision *must* be made, they could pick an option that is suboptimal but definitely non-catastrophic. And if there's some argument or piece of information that causes trouble, the institution could design weak AI systems to recognise it, and avoid all future exposures to it.

In order to reduce the risk of drift even further, an institution could simply refrain from investigating any unnecessary technologies or philosophical ideas, in the first place. If the institution had sufficient power over the world, it could similarly prevent the rest of the world from investigating such ideas.

In other words: *even if* a value-system would prove to be incoherent upon philosophical reflection, it might still remain stable simply by preventing philosophical progress.

Purely technological progress doesn't seem very likely to cause value drift. But if an institution was so afraid of drift that they *never ever* wanted to encounter a dilemma that they hadn't already considered in advance, they could also halt technological progress and completely stagnate all forms of civilizational change. This would eliminate the need for interpreting the values in novel situations.

6.5 Aligning with that interpretation

Given that the above makes it quite clear what the goal should be, in any particular situation, would an institution be able to robustly optimize towards that goal?

This is a question that human institutions have solved surprisingly well, even when using humans who are far from aligned with the institution's goals. Many human regimes are in fact quite stable. By starting out with systems that are basically aligned with the institution's goals, the task becomes *far* easier than it is for human institutions. And thorough testing, resetting, and supervision makes the task even easier.

A basic overview of what an institution could do is to:

- Thoroughly test and understand how each system will behave after foreseeable inputs.
- Reset systems back to a thoroughly-tested state often. (This gives the systems less time to stray from a trusted state, and allows testing to focus on shorter inputs.)
- Continuously check that each system does its job reasonably well, and in particular doesn't pose any danger to the institution as a whole, through a combination of:
  - Using very similar systems to perform exactly the same task; to cover cases where sheer randomness would cause one of them to drift.[20]

---

20 Redundancy can easily get rid of uncorrelated errors. This is analogous to how some militaries have required 2 operators to launch nuclear weapons. (It's not possible to take the majority vote for questions that have a huge option-space, but if one or a few AI systems *generate* suggestions, then a majority vote can be required to *approve* plans that the majority believes to be safe, if not necessarily optimal.)

- Using supervisors who don't engage as deeply with the inputs as the supervisee, but who check that the outputs are acceptable.
- Continuously re-testing the system's motivations.

As an example, consider research and engineering:

- For every fact or skill that's necessary in the relevant field, extensive experimentation could identify the minimal changes to a trusted AI system that would give them that skill. The AI granted those skills could then be thoroughly tested using the methods discussed below.
- Researchers and engineers may then have to spend significant amounts of time thinking about their intended topic. But they could be supervised by copies that spend much less time, who need to approve every insight or new design that the researcher produces.

6.6 Verifying loyalty

A few different times we have now said that institutions might do thorough tests of AI systems. We will now discuss this in more detail. To recap, institutions might do this in several different situations:

- AI systems could test copies of themselves, to better understand how robust they are, and in which situations they could fail.
- If there is some important piece of information that needs to be internalized by many copies, they could expose some copies to that information, and then test them to ensure that it didn't cause any changes in their goals.
- If a supervising AI sees some reason to suspect that a supervisee may have been corrupted, they could have them tested.
- If the AI systems develop some other, more specialized systems to fill particular niches, they may want to test that those systems behave well in the situations they are to be implemented in.
- AI systems could test copies of themselves that they had tweaked to increase capabilities or robustness, or could test entirely new AI systems that they wanted to replace themselves with.

First, two remarks on the testing procedures that follow:

- The more extreme tests could be quite expensive. But the escalation could be gradual, depending on how likely it is that the tested AI system is dangerous based on both prior information and the results of cheaper tests.
- Some of these testing schemes could cause the testers themselves to experience value drift, by being exposed to the same information that the testee was exposed to. But if just a few of the testers suspected that this had happened (without themselves being affected), they could sound the alarm and reset both the tested system *and* the testers to a previous state.

Testing input-output behavior

The most obvious way to test an AI system is to simulate it in important situations that it might encounter in the future, to see if it behaves well.

- Typically, the tested AI wouldn't be any more competent than the testers. Therefore, if there's any possible flaws in a test that would let the testee deduce that they were being tested, the

testers would be aware of this, so that they could avoid such flaws. This means that the testers would know how the tested AI could deduce that it was being tested, which would allow them to design tests where the AI was genuinely unsure whether it was currently being tested or not. The testers being at least as competent as the testee also means that — if the tested system chooses some action because it can see that it would harm the institution — the testers would also be able to understand and recognise that that action could harm the institution.

    - Cases where the tested AI system is more capable than the testers would be avoided if the institution never improved its AI technology.
    - But if they wanted to build stronger AI systems, one technique they might use is to have the tested systems' capabilities only be *slightly* stronger than the individual testers, and then use multiple testers with more processing power to check each of the stronger system's decisions; with the hope that many fast AI systems would be more capable than a single, slower slightly stronger system (Christiano et al., 2018).

- There could be significantly more tests of high-stakes situations than low-stakes situations. For example, testers could simulate situations where the AI system sees suspicious behavior that it should report, or it could simulate a crisis situation where the AI system plays a pivotal role.

However, there may be some input that *can't* be tested, because the testers can't feasibly generate them. If the tested system knows about this, and could recognise such inputs, it could choose to behave badly on only those inputs. This could include any large input, taking place over long periods of time, which would be expensive to simulate. Or it could include any situation which is knowably hard to *generate* , but easy to *recognise* , like a factorisation of a large semiprime (see the discussion on RSA-2048 in Christiano (2019)).

One solution may be to always place the model under supervision when such a knowably-unprecedented input appears (as discussed above). Another solution is to use transparency methods, which we will discuss next.

Transparency

Since AI systems are implemented digitally (rather than in biological brains), supervisors would be able to see the AI systems' internal computations. Insofar as these were at all interpretable, they could be used in a number of situations:

- Transparency tools could be used during the testing-procedures above, for at least two purposes:
    - First, they could clarify *why* an AI system does what it does in the tested cases. For example, if it's trying to figure out whether it's in a test situation (presumably with an intention to behave differently depending on the answer) that's a red flag.
    - Second, it could potentially be used to modify an AI system's beliefs directly. For example, this could be used to test what the AI system would do *if it really believed* that it was not currently being tested.
- Separately from behavioral testing, sufficiently strong transparency tools could also directly look for important beliefs, goals, or intentions in the mind of the given AI system.
    - Some testers could try to understand as much as possible about the AI system's mind. If that was an unfeasibly large task, or if analyzing the entire mind would have a high risk of

corrupting the analyser, there could instead (or in addition) be many AI systems analyzing many small bits of the AI's mind.

    - Relatedly, transparency tools could be used during *supervision* as a way to detect big changes of goals without the supervisor being exposed to the reasons that the goals changed. I.e. the supervisor could monitor certain parts of the supervised system's cognition, chosen such that if a supervised system encounters information that induces value drift, the supervisor can notice this without themselves processing that information.

Would transparency techniques be good enough to reliably do these tasks?

- By default, we might expect AI systems of the future to become more complex and thereby more difficult to analyze.
    - For some reasons to think that smart AI might be uninterpretable, see Sharkey (2022) .
- But we could also expect progress in transparency techniques. In addition, we should expect better AI capabilities to directly contribute to *automating* the process of interpretability.
    - For example, with a lot of work, a team at OpenAI managed to understand a small part of a neural network in great detail (Cammarata et al., 2020). Having humans take equally much time to look at every part of a large neural network is not feasible. But it could become scalable with the help of AI tools.
    - By the time we have AGI, achieving a human-level understanding of how each component of a large neural network works might be possible by using many copies of such AGIs to analyze them.[21]
- This still leaves significant uncertainty about how interpretable future AGI systems will be. But if we consider an institution that could develop and select their systems partly for being interpretable, and who were willing to pay large costs to analyze individual systems, it seems more likely than not that they could *eventually* develop AGI where a high degree of transparency was possible.

There are also methods for discerning AI goals that are halfway between transparency and testing. As an analogy, we can consider how someone could discern a human's motivations, if they had access to arbitrarily many digital copies of them.

- The testers could have many thousands of conversations with the human, each instance with a different conversation partner. Some of these could take place in highly stressful or highly relaxing situations. In some, the human could be given drugs. In some, the testers could directly influence the human's digitally-implemented neurochemistry. Some simulations could last for a long time, seeing what happens if the human spends years immersed in various environments.
    - That procedure alone seems like it should give an excellent understanding of a human's goals, desires, and behavior.

---

21 Making up numbers: Assume that a human-level model could understand 50,000 parameters in 50 days (which is probably within an order of magnitude of what the Cammarata el al. did with curve detectors). If we then take estimates from Cotra (2020) for the parameters and compute costs of such a model, the cost of such a model analyzing all its own parameters is within a couple of orders of magnitude of the cost to train it (and it's unclear if it would be bigger or smaller). This is not to say that it makes sense to have models analyze themselves, but it suggests that, once we have AGI, contemporary models could plausibly analyze each component of newly trained models.

- In addition, throughout all these diverse situations, neural activations in the person's brain could be monitored by humans or by ML models trained to predict whether the human is experiencing various emotions, whether they're lying, what topics they are thinking about, etc. This could then be used to check for signs of dishonesty during questioning or testing.
- Of course, some of these methods are human-specific. But the human case is a useful data point, insofar as we don't expect humans to be a huge outlier in how easy it is to get insight into our psyche.

7 Robustness to natural disasters

Overview: Let's say that an institution is stable absent outside interventions. Natural disasters do not seem like a significant threat to such an institution, given that:

- Local disasters do not pose a problem, given high baseline stability achieved by redundant storage of everything important.
- Global disasters are rare, and ones that cause widespread comprehensive destruction are extremely rare.
- In the long run, the institution could also spread to space.

7.1 Natural disasters

As previously mentioned, the resources and information necessary to rebuild an institution could be kept in many places across the Earth's surface, such that no *local* disaster could destroy all of it.

What about global disasters? The existential risk from natural disasters (which includes both human extinction and civilizational collapse without recovery) has been estimated to be around 1/10,000 this century (Ord, 2020). The biggest risks in this category are asteroid/comet impacts, supervolcanic eruptions, and (depending on whether you count them as natural or not) non-engineered pandemics.

- Biological pandemics are not a threat; primarily because AI would not operate on biological hardware (and an AI civilization would not need to rely on humans for anything), but also because it seems likely that future technologies would make pandemic mitigation extremely effective.
    - Intentionally designed computer viruses would be an analog to *engineered* pandemics, but they could be prevented by preventing anyone from constructing them. See section 8 .
    - In favorable circumstances, something analogous to computer viruses could probably arise without intentional actions, just like biological viruses arose from pure evolution, without being designed. But an institution trying to be stable could make this very improbable.
        - The evolution of biological diseases rely crucially on occasional errors in DNA- or RNA-copying, which generate variation on which natural selection can act. But via the use of digital error correction , the probability of such low-level errors could be driven to effectively 0.

        - There would still be some variation induced by the use of different AI systems in different contexts, or different AI systems receiving different histories of inputs. But this would be relatively little variation, and extensive supervision and testing protocols (see section 6) could both reduce the variation (e.g. by frequently resetting AI systems) and notice if any bug seemed to be self-replicating.

- The main reason for which asteroid/comet impacts and supervolcanic eruptions can become global (rather than local) catastrophes is because they can throw up enough dust to blot out the sun for a period of years, killing plant life that relies on the sun and animal life that relies on those plants. But an AI civilization could easily survive such catastrophes by using energy-sources that don't depend on the sun (such as nuclear power or enough stored electrical or chemical energy to last for several years).
    - *Sufficiently* large astronomical objects would pose more extreme dangers, including liquifying the Earth's surface into magma. But the largest asteroid to hit the Earth in the last 100 million years (the asteroid that killed the dinosaurs) mainly had global impact via the atmosphere, and asteroids much larger than that one are significantly rarer.
    - In addition, future technologies would probably enable very effective asteroid detection and deflection.

This should make AI civilizations much more robust than humans are today, implying a lifespan of at least millions of years. Due to the evolution of the sun, all eukaryotic life is predicted to die out 1-2 billion years from now, and the Earth itself is predicted to be consumed by the sun in 7-8 billion years (see Wikipedia). This would be a necessary end-point to any civilization that didn't spread to other planets (and didn't do any space engineering to stop this process).

However, millions of years would be more than enough to spread to other solar systems (Beckstead, 2014; Armstrong & Sandberg, 2014). Would this be consistent with continued lock-in?

7.2 Astronomical stability

If a civilisation's only goal in traveling to space was to survive the end of the Earth, they may not need to spread very far. Instead, they could just stay in a few different solar systems, close enough that communication was possible with just a few light-years delay.

A civilisation might also want to spread out much farther than that, colonizing all of the milky way and perhaps many more galaxies. This would preclude communication between the more distant entities. But it seems like this would be perfectly consistent with lock-in. The distances and amounts of time are large, but that's solvable with enough error correction. (See Shulman (2012) making a similar point, and Hanson (1998) for an alternative vision of space colonization with significant variation and selection.)

In particular, if each colonization expedition has stable values, each separate instantiation could stay true to the original. If two separated civilizations later encounter each other, they may have developed different procedures or heuristics for how to carry out their goals in practice. But there could be pre-specified (perfectly preserved) algorithms that determine how to settle any such differences.

8 Robustness to non-aligned actors

Overview: Let's say that an institution

- has uncontested military and economic dominance,
- has the ability to produce any number of stably-aligned agents, and
- that they will try to survive and maintain coercive power as long as possible.

This section calls such an institution a "dominant institution", and argues that it would not be overthrown by any non-aligned actors. The short argument for this is:

- For any task that the institution needs done, they can use one of their stably-aligned AI systems for it. Thus, actions taken by humans and non-aligned AIs would not be necessary for the continued existence of the dominant institution.
- The institution could control far more stably-aligned AGI systems than there are humans or non-aligned AI systems, and could therefore do any amount of surveillance necessary to ensure that no-one causes the institution any catastrophic harm.

The exception to this is alien civilizations, since they could have comparable or greater economic and military power when first encountered.

Note that the ability to prevent revolutions does not necessarily imply that the population would be powerless to change things. It would probably be possible to prevent the former while still allowing the population to change the system via e.g. voting.

We'll start with giving an argument for in-principle feasibility, and then discuss how expensive stability would be in practice.

8.1 In-principle feasibility

Any task that is directly important to the institution's survival could be done by stably-aligned agents. This could include tasks that are directly relevant to the institution's functioning, such as commanding the military or gathering crucial information. However, it also includes a much larger range of tasks, such as producing energy and material products that are necessary to keep the institution functional. By assumption, the AGIs would be at least human-level at carrying out these tasks, which seems good enough to provide any necessary service. Today, humans are certainly capable of providing large quantities of material support in a scalable fashion.[22]

A world-wide dominant institution would also require large amounts of internal administration. But today, this is not an insurmountable problem for nations as large as China (close to 1/5th of the world's population), and there seems to be no fundamental barrier to extending this to something

22 This is often done via non-centralized forms of organization, like markets. Partly these markets provide the right incentives, which would be less important for stably-aligned agents. But insofar as markets turn out to be indispensable for other reasons, too (such as providing important information via prices) there's no reason that they couldn't be used to organize stably-aligned agents.

as large as a world government.[23]

By relying on stably-aligned agents for essential services, non-aligned members of the population could not significantly harm the dominant institution by *inaction*. The only way that non-aligned members of the population could harm the dominant institution would be by actively doing something harmful.

Insofar as the institution's goals didn't at all rely on humans (or other non-aligned agents), their uncontested military dominance could in-principle allow them to replace the entire population with stably-aligned agents. If society was completely composed of stably-aligned agents pursuing a single goal, then no one would take action to end the lock-in. But such an omnicidal approach seems wildly beyond the requirements for stability. A less extreme approach (though it also seems much more extreme than would be necessary, see next section) would be to assign one stably-aligned AGI to constantly surveil each non-aligned member of the population.

- For human members of the population, this would (by assumption) be at least as effective as having each human be constantly watched by another human, the latter of which was loyal to the dominant institution.
- Part of the population could also consist of AI systems that are not robustly aligned, which raises some new questions. In particular, it may be difficult for an aligned AI system to surveil a different, unaligned AI system that was much more capable, at some tasks, since this means that the aligned AI doing the surveillance may not understand the intention behind or consequences of some of the surveillee's actions. Thus, the dominant institution might have to be on the cutting-edge of AI technology (possibly by prohibiting all superior forms of AI).

With each action in a society either being surveilled by or being carried out by a stably-aligned AI system, it would be extremely difficult for anyone to significantly harm the dominant institution. The dominant institution and its AI systems would not be absolutely flawless, and may occasionally make mistakes. But any treacherous action would have a very high probability of being caught, and any serious attempt at destabilizing the whole institution would likely need sequences of many actions with a clear potential for danger. And every time a mistake was made and noticed, the system could adapt to reduce the probability of such mistakes in the future.

Surveilling the entire population could be expensive, but it would not be an insurmountable (economic) cost:

- Currently, it costs ~$10/h to rent 8 A100 GPUs (Amazon, 2022), collectively capable of 1e15-1e16 FLOP/s (Nvidia, 2022). This is probably similar to or greater than the FLOP/s used by a human brain (Carlsmith, 2020). There aren't currently enough GPUs to run more than a few million models of this size, but supply could be expanded if there was demand.
- Technological progress has historically led the price of computation to fall by ~6-300x per decade (AI Impacts, 2015, 2017). While the current paradigm of computer hardware is approaching physical limits, some further gains remain for cheaper AI hardware. (And more speculatively, the price of computation could fall many additional orders of magnitude with the use of novel computation technologies like reversible, optical, quantum, or superconducting computing.)

---

23 AGI could also reduce many administrative problems, such as by reducing principal-agent problems (see section 5) and by making coordination easier by letting a single boss be in multiple places at once (via copying).

- And AGI could both accelerate technological progress by a lot (leading the price of computation to fall even further) and massively increase wealth, making expenses more affordable (Davidson, 2021 ; Trammel, 2020).

Given all of this, it seems like the default future is one where there are far more AI systems than humans. So even in the scenario where a dominant institution opted for individual, constant, human-level surveillance, such surveillance would only take up a small fraction of total productive capacity.

8.2 How much control is needed?

If a dominant institution didn't want to go that far, would there be less extreme options? Probably. (However, guessing at the minimum level of control necessary is more difficult than establishing that control is possible, and so this section is more speculative.)

Firstly, the essential tasks that are needed to preserve a dominant institution would not necessarily need to be carried out by perfectly aligned AGI systems. It may work to have some of them just be under limited supervision, or to have them be provided by many redundant services that are unlikely to fail simultaneously.

Secondly, stability might not necessitate that much surveillance. There seems to be 3 things that a dominant institution would need to control: (i) access to (offensive or defensive) military capabilities, (ii) the ability to escape the reach of the dominant institution, and (iii) *possibly* the ability to spread some ideas.

- (i): The most clear way to pose a concrete threat to a dominant institution would be to amass enough defensive or destructive power that it would no longer be dominant.
    - Since the dominant institution could make itself militarily dominant with a very large margin, the defensive approach could be very difficult, and likely easy-to-detect even without any significant degree of surveillance.
    - And since a dominant institution could make many redundant copies of everything important, and spread it across the world, the destructive approach would require truly worldwide destruction. It *might* be the case that preventing such worldwide destruction would require extensive surveillance (Bostrom, 2019). But note that *every* civilization that wanted to avoid imminent destruction would have to handle such threats; so these costs would not be unique to lock-in.
    - Our main uncertainty here is about cybersecurity. Perhaps there are hacks which could cause worldwide problems in a world with sufficiently centralized power, that wouldn't cause problems for a more decentralized world.
- (ii): A dominant institution may also need to prevent unauthorized space travel, since anyone who left the institution's purview would be able to build their own civilization, until they posed a threat.
    - Though this probably wouldn't be a danger if the institution had already sent out its own colonization wave and already controlled all accessible resources in space.
- (iii): In human authoritarian societies, it's common for regimes to try to limit what kind of ideas and ideologies can spread throughout society. This seems much less necessary in an AGI-controlled society — if a dominant institution's essential supporters were truly stably aligned,

then there would be little or no reason that a dominant institution would have to do this, since the remaining population couldn't threaten it. However, there are two possible partial exceptions to this:

  - First, it is possible (though we think it's unlikely) that AI systems' stability could depend on what information the systems are exposed to. In section 6 , we bring up the possibility that some AI systems might be stable only if they are not exposed to certain information or arguments. If this were the case, then in order to maintain stability in the long-term, it could be necessary to prevent such ideas from becoming too common. This would require enough surveillance and control to prevent those ideas from being invented and/or from spreading.
  - Second, it is possible that a dominant institution could find it easier to maintain stable control if some particular technology isn't widely adopted. The most salient examples of this are technologies of mass destruction, which we discussed under point (i) above. But perhaps some technologies could be uniquely problematic for a locked-in society. For example, perhaps the dominant institution relies on some less-than-maximally effective form of AI (e.g. because they have more stable goals), and would be threatened by the development of superior AI systems. If so, it may need to do enough surveillance to limit technological development to this lower level.

- But it seems fairly likely that (iii) wouldn't be a problem, in which case a dominant institution may only need to really robustly prevent (i) extremely destructive capabilities, and (ii) unauthorized space travel to regions the institution does not yet govern.

8.3 Alien civilisations

There's one type of non-aligned actor that these arguments don't apply to: alien civilizations. Compared with Earth-originating civilizations, alien civilizations might have similar or greater economic and military power when we first encounter them.

Exactly what happened, then, would depend on fundamentals about space and the limits of technology. If the offense/defense balance was such that nobody could threaten any already-occupied territory, then any such territories could stay locked-in. But if invasion or destruction was feasible, then it is difficult to say how the interactions would play out.

When would this be relevant? According to the best models that we know of, the time until we first see another civilization is probably measured in billions of years, even given quite alien-friendly assumptions (Cook, 2022). So it seems unlikely that this would prevent stability over the next few million years.

As a final note, about 150 billion years into the future, an Era of Isolation has been predicted (Ord, 2021). In this era, space would have expanded so much that it would be impossible to travel to any but a few dozen of the most nearby galaxies. If at the beginning of this era, any such galaxy cluster is dominated by a single institution, it would never encounter outside interference again.

References

Armstrong, S., & Sandberg, A. (2013). Eternity in six hours: Intergalactic spreading of intelligent life and sharpening the Fermi paradox. *Acta Astronautica* , *89* , 1-13.

Aschenbrenner, L. (2020). *Existential risk and growth* . Working paper. Global Priorities Institute, University of Oxford.

Beckstead, N. (2014). *Will we eventually be able to colonize other stars? Notes from a preliminary review.* Future of Humanity Institute, University of Oxford.

Bostrom, N. (2013). Existential risk prevention as global priority. *Global Policy* , *4* (1), 15-31.

Bostrom, N. (2014). *Superintelligence* . Oxford University Press.

Bostrom, N. (2019). The vulnerable world hypothesis. *Global Policy* , *10* (4), 455-476.

Cammarata, N., Goh, G., Carter, S., Voss, C., Schubert, L., & Olah, C. (2021). Curve circuits. *Distill* , *6* (1), e00024-006.

Caplan, B. (2008). The totalitarian threat. *Global catastrophic risks* . Oxford University Press.

Carlsmith, J. (2020). *How Much Computational Power Does It Take to Match the Human Brain?* Open Philanthropy Project.

Carlsmith, J. (2022). Is Power-Seeking AI an Existential Risk? *arXiv preprint arXiv:2206.13353* .

Christiano, P. (2014). *Machine Intelligence and capital accumulation* . Rational Altruist.

Christiano, P. (2018a). Techniques for optimizing worst-case performance. *AI Alignment — Medium.*

Christiano, P. (2018b). Clarifying "AI alignment". *AI Alignment — Medium.*

Christiano, P. (2019). Training robust corrigibility. *AI Alignment — Medium.*

Christiano, P. (2022). Decoupling deliberation from competition. *Alignment Forum.*

Christiano, P., Shlegeris, B., & Amodei, D. (2018). Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575.*

Cook, T. (2022). Replicating and extending the grabby aliens model. *EA Forum.*

Cotra, A. (2020). Draft report on AI timelines. *AI Alignment Forum* .

Cotra, A. (2022.). Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover. *AI Alignment Forum.*

Cohen, M. K., Hutter, M., & Osborne, M. A. (2022). Advanced artificial agents intervene in the provision of reward. *AI Magazine* .

Davidson, T. (2021). *Could advanced AI drive explosive economic growth?* Open Philanthropy.

De Blanc, P. (2011). Ontological crises in artificial agents' value systems. *arXiv preprint arXiv:1105.3821.*

Dennett, D. C. (1987). *The intentional stance* . MIT press.

Diamond, J. (1997). *Guns, Germs, and Steel: The Fates of Human Societies.* W.W. Norton & Company.

Fearon, J. D. (1995). Rationalist explanations for war. *International organization* , *49* (3), 379-414.

Göbel, C. (2011). Authoritarian consolidation. *European political science* , *10* (2), 176-190.

Hanson, R. (2011). Shulman on Superorgs. *Overcoming Bias.*

Hanson, R. (2016). *The age of Em: Work, love, and life when robots rule the earth* . Oxford University Press.

Hanson, R. (1998). Burning the cosmic commons: evolutionary strategies for interstellar colonization. *preprint available at* http://hanson.gmu.edu/filluniv.pdf .

Hanson, R. (2018.) On value drift. *Overcoming Bias.*

Hanson, R. (2022). MacAskill on value lock-in. *Overcoming Bias.*

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2019). Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820.*

Karnofsky, H. (2021). Weak point in "Most important century": Lock-in. *Cold Takes.*

Karnofsky, H. (2022). AI could defeat all of us combined. *Cold Takes.*

MacAskill, W. (2022). *What We Owe the Future.* Hachette UK.

Macon, M. W. (2015). The Noisy-Channel Coding Theorem. *San Francisco State University, San Francisco* .

Ord, T. (2020). *The precipice: Existential risk and the future of humanity* . Hachette Books.

Pippenger, N. (1990). Developments in “the synthesis of reliable organisms from unreliable components”. *The legacy of John von Neumann* , *50* , 311-324.

Rubin, J. (2014). Printing and Protestants: an empirical test of the role of printing in the Reformation. *Review of Economics and Statistics* , *96* (2), 270-286.

Samberg, A., & Bostrom, N. (2008). *Whole brain emulation* . Future of Humanity Institute.

Sharkey, L. (2022). Circumventing interpretability: How to defeat mind-readers. *AI Alignment Forum.*

Shulman, C. (2010). *Whole brain emulation and the evolution of superorganisms.* Machine Intelligence Research Institute

Shulman, C. (2012). Spreading happiness to the stars seems little harder than just spreading. *Reflective Disequilibrium.*

Spielman, D. A. (1996). Highly fault-tolerant parallel computation. In *Proceedings of 37th Conference on Foundations of Computer Science* (pp. 154-163). IEEE.

Trammell, P., & Korinek, A. (2020). *Economic growth under transformative AI* (No. 8-2020). GPI Working Paper.

Tomasik, B. (2020). Will Future Civilization Eventually Achieve Goal Preservation? *Reducing Suffering.*

Von Neumann, J. (1956). Probabilistic logics and the synthesis of reliable organisms from unreliable components. *Automata studies* , *34* , 43-98.

Yudkowsky, E. (2008). Measuring Optimization Power. *Less Wrong.*