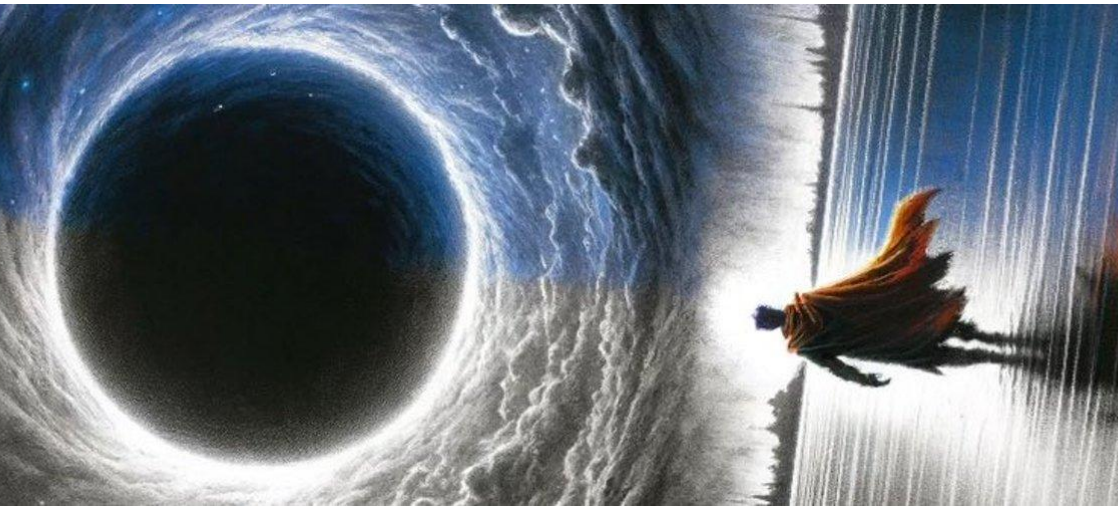


The Tiny Book on the Governance of Machine

Jordan Ezra Fisher

Well, this is it. I wrote a big ass post about my thoughts on how to improve AI governance. I called it a Tiny Book so you might be tricked into thinking it's a quick read. Good luck.

The Tiny Book on the Governance of Machine

No one wants to read 50 pages on governance in one go. So this book is built to be read in parts, or just pick the parts you're interested in.

Want a primer on AI? Start with "A Crash Course on AI" on page 60.

Already up to speed on AI and understand the risks? Jump straight to "Implicit Guardrails" on page 20.

Just want to read some new ideas in AI governance? Jump to "Rapid Fire Governance" on page 43.

Only have 30 seconds? Read the "TL;DR" on page 6.

Or, fuck it, YOLO your way through the whole book.

don't be silly

> *or do be silly, i don't care. or i maybe i do care. ok yeah i do care*

Serious people have been talking about human-level AI and superintelligence for nearly half a century. But mostly it's been too silly to talk about. In computer science departments it was implicitly banned from discussing openly, or you'd be made a pariah. Outside academia the public thought it was sci-fi.

Even 10 years ago, when AI began showing real promise, you instead had to say you worked on "ML" or "Machine Learning". Only in the last few years has the reality of AI become real enough that we've even allowed people to say what they do is "AI" without scoffing at them.

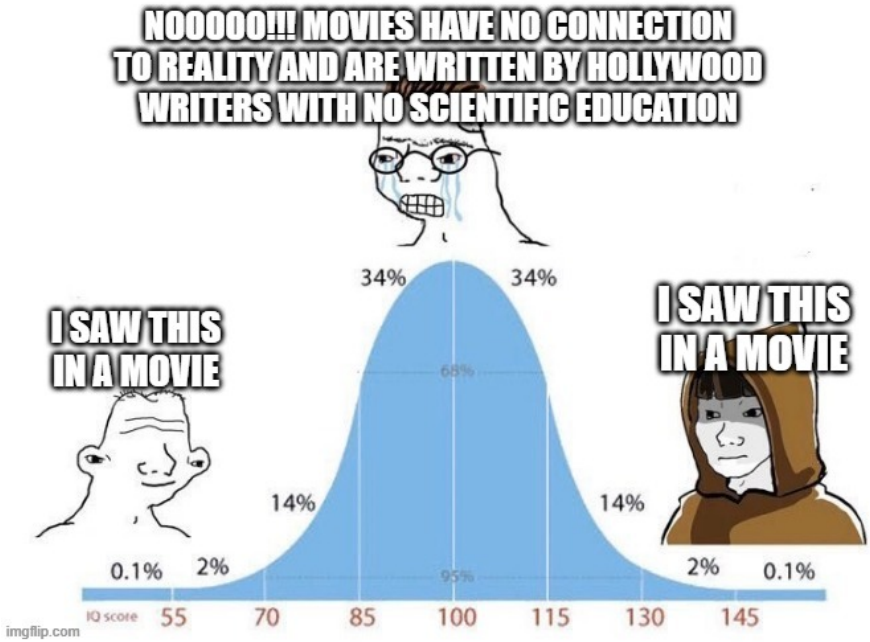
Now, here we are, embedding AI into every part of our lives and work. Proof that silly things are sometimes real. Proof they sometimes reshape the world.

This may seem like a small thing, but it's not. We shy away from talking seriously about serious things when we're afraid it makes us look silly. And it has real consequences. That's what happened

with the discourse around AI risks:

Now that AI is near it's becoming plain that sharing the world with another form of intelligence will have risks. Obviously, how could it not? But the risks from AI have been called out by the likes of Bostrom and Yudkowsky for decades now, largely ignored because it was too silly. And the real impact is that we've now failed to prepare for those risks. We've failed at even discussing them broadly as a society. Now we have little time left.

It's time to stop feeling silly and start taking seriously the silly things in front of us that are deadly serious. However silly they are.


NOOOOO!!! MOVIES HAVE NO CONNECTION TO REALITY AND ARE WRITTEN BY HOLLYWOOD WRITERS WITH NO SCIENTIFIC EDUCATION
I SAW THIS IN A MOVIE
I SAW THIS IN A MOVIE
34%
34%
68%
14%
14%
0.1%
2%
95%
2%
0.1%
IQ score
55
70
85
100
115
130
145
imgflip.com

Creating AI is going to change every single aspect of our world. But *the way* it changes our world *is not yet written in stone.* We all together have a part to play in determining the outcome. But only if we get over ourselves and talk about these silly things.

We are creating machines smarter than us. They will have their own agency. We may not be able to keep control of them. Even if we do, they may be used by governments to gain unprecedented power over us.

These are serious fucking things. And they're silly too. The last half century has coddled us into a false sense of stability. Whispered into our ear that silly things don't happen on the world stage, not really. But the rules are changing. We don't know what the new rules will be.

It's OK to be silly. Now let's talk about it.

TL;DR

- **AI changes what we need from governance**
  - Societies that allow for freedom and maximum human flourishing aren't easy to build
  - Ours took thousands of years of trial and error and is still a work in progress
  - What allows a society to function depends strongly on the details of its members: us humans
  - Introducing AI will completely change all of those details
  - Because of that, we must rethink society if we want it to be resilient to the arrival of AI
- **There are many risks we'll face, and some risks aren't yet being discussed**
  - In particular, the risk of concentration of power and dictatorship is extreme with the full power of AI automation
  - Without changes, dictatorships are a default outcome of AI. And it will be increasingly hard for us to resist it the later we act

  - We must evolve our governance *before* strong AI arrives, or we may not have any power left as citizens to fix it *after*
  - Humans alone won't be able to oversee leaders empowered by AI

- **AI-powered governance of AI**
  - Instead, we must leverage AI itself to become part of the checks and balances on how government and industry wield AI
  - Passing laws is necessary but not sufficient
  - We must enshrine the oversight inside the mechanisms of culture and government, and we must do it while we still have human-based trust in human institutions

- **We all are part of this**
  - AI is the most important event in our life time, but the outcome is not yet written
  - We all own the conversation for what we want the future to be

Intro

We assume that machines will achieve human level intelligence. We assume they will exceed it. We assume that they will have ethics, aligned to a human or group of humans.

Then the question remains: which human or group of humans. To which other humans are those humans accountable and by what means. Where does the demos in democracy sit?

The ambition of people leads to cathedrals and death camps. Prosperity and war. Governance is how we harness our collective ambitions to aim for the good. It's not just laws, it's culture, norms, and expectations we place upon ourselves, our neighbors, and our children. It's the combined wisdom of society for how we can stand both free and together to march forward.

There aren't easy answers for what governance should look like, only tradeoffs and complex second order dynamics. Should we empower a strong leader to move rapidly, hoping they won't abuse their position? Or should we create a slow moving bureaucracy, resilient to corruption but unresponsive to changing needs? Outside of government we must answer similar questions for our communities and companies. How much power? How much oversight?

The US has a mixture of answers to these questions. We allow CEOs to be board elected dictators of companies. With it can

come great speed, vision, coordination, or rapid failure. There are checks though: The board. The market. Employees can vote with their feet, or the implied threat of them. Regulation limits the most extreme excesses and the worst tragedy of the commons. But even still, a successful CEO can amass so much wealth as to pose unacceptable danger, as we saw with the robber barons. So we further limit what wealth can achieve: bribes are illegal and we curtail the ability to buy politicians outright. Failing to curtail the abilities of wealth itself, we curtail the maximum accumulation of wealth by breaking up monopolies. We invest in common goods like education so that others can rise up and build their own wealth, to counter the entrenched wealth of the past.

We elect a president of a strong federal government to have time limited, broad authority. With checks. There's only ever one president at a time, so there is no free market of presidents, so we can't allow outright failure when it might mean the end of democracy. So we prevent the use of power to further seek power, such as to influence an election. We empower Congress and the judiciary to prevent the executive from granting itself more power and creating a runaway process toward dictatorship.

Implicit but just as important is culture. The American tradition of democracy and standing against tyrants. The President's cabinet is a set of Americans beholden to this culture, upheld by social pressure from their friends, family, and community. The federal agencies they oversee are composed of millions of Americans, allowing a million opportunities for American culture to uphold itself. A million opportunities to thwart a would-be tyrant. Once we fully automate government, where will these guardrails come from?

What mechanism will ensure government is for the people when

it's no longer of the people and by the people?

We're rapidly approaching AI strong enough to automate our government, without understanding how we'll hold government accountable with that new power. It may prove *impossible* to control government after we give it this power, if we haven't put controls in place beforehand. There is a path dependence to our future, and timing is a critical variable:

You don't grant Caesar an army to conquer Gaul for Rome until *after* you are confident you can govern Caesar. The Rubicon is not a sufficient form of governance definition, no matter how strong the norms not to march across it are. In this sense we see that governance is a form of alignment, where we want helpful results for society (build Rome!) while minimizing the harmful outcomes (don't conquer Rome!). This abstraction applies then to machine, humans, organizations, and machine powered organizations.

There aren't easy answers, despite the allure of simple ideologies and absolutisms. Even today our governance is imperfect, and we risk devolution and dictatorship at all turns without constant vigilance and adaption. What was needed to govern well the Romans is not what is needed to govern well today. And it's almost certainly not what's needed to govern a human-machine civilization tomorrow. And tomorrow may be very soon.

The core question of governance is how to govern *intelligences*, human or otherwise: collections of forces that can achieve what they seek, can win more power, can cooperate, compete, and destroy. Governance is a set of yes's and no's: yes compete this way, no don't destroy this way, such that the citizens mutually benefit and consolidation of power into dictators is prevented. And the dangers of power abound.

A glib history of governance: governance too weak can lead to hard times and dictators; too strong can lead to hard times and dictators. And there isn't a simple line between weak and strong. There is no simple compromise, and compromise itself is only sometimes the answer.

Machines will likely enumerate a range of intelligences, requiring a range of governance types. With that lens humans are a special case of governing intelligence. But we further see that a society of humans and machines combined is another case again, and is likely the future we'll be in.

The question of how to govern machine is thus a question of how to govern man. What social compact must we craft so that an aggregate society of diverse intelligences is a net good for those intelligences, and a net good for us in particular.

Thousands of years have been spent on the question of human governance. Millions of thinkers. Countless debates. Dense treatise. Horrible wars.

The question touches the nature of our existence. What world do we want to live in?

The governance of machine poses an equally profound question. We won't have a thousand years to arrive at good answers. We can't afford the deaths of past wars to settle disagreements. We have little time.

But we must find an answer.


i had to use a 2 year old image model otherwise this wouldn't look dumb enough
REGULATE,
DEREGULATE
ORDER?

Contents

The debate for how to govern a human-machine society may prove the greatest in our lifetime, perhaps as important as the founding debates around modern democracies themselves. Every one of us has a role to play: to engage in this debate, to shape it, to fight for governance that lifts up mankind and allows this new age that's upon us to be a noble one, not a dark one.

In this tiny book we'll cover the basics:

- Why is this important, aren't there bigger problems to solve?
- How does our current civilization work, and what will stop working once we have AI?
- What does the current political, cultural, and power landscape look like as it relates to AI?
- What might a human-machine society look like, what might governance look like, and how can we leverage AI to make it possible?

Why

Some might say, "One problem at a time."

First, let's build the machine. This is hard enough.

Then, let's make sure it's safe. This is hard enough.

Finally, let's see how to integrate it into society. Let's only then craft a world with AI that's still a world for humans, with all the

challenges and upheavals that will take.

Depending on how spaced apart these events are, that's a reasonable position. 50 years ago certainly there was enough time to focus on the first problem only. 5 years ago perhaps it was fair to focus only on the hard problem of making AI safe. Today, these three events may all happen in the next few years. If so, practically, we can't wait to solve each problem one by one. There won't be enough time to do it right. Worse, if we build controllable AI but don't know how to govern that new human-machine world, there may not be any way to prevent the worst outcomes of concentration of power and the rise of durable dictatorships. The path to a good human-machine world very likely requires taking the correct actions *leading up to* the arrival of strong AI, **even if** we have solved the problem of ensuring the AI is safe.

There is a path dependence, and **our actions today matter more than our actions tomorrow**.

If you're an AI researcher, today your voice matters, tomorrow you will be automated and will lose your currency. If you're a government employee, today your voice matters, tomorrow you will be automated. If you're a voting citizen, today your vote matters, tomorrow it might not be possible to vote out an automated government dictatorship. If you're any person at all, of any walk of life or nation, today your actions impact the shared culture of humanity, which helps pressure and guide the actions of every other person. Tomorrow, we may live in a world where no amount of cultural pressure matters. Your actions matter today, use them to ensure they still matter tomorrow.

How soon will strong AI arrive? We won't spend time analyzing timelines here. There are great discussions about this, it's increas-

ingly important, but it's overall a well trod area. What's not well trod is what the world should look like *after*. After we've built and aligned the machine. And, anyway, the timeline discussions are changing rapidly, anything we write here will likely be outdated before this is published or before you read this. Regardless of timelines, whether we have two years or ten years, there isn't enough time. We have to prepare now.

Nonetheless, keep engaging in timeline discussions. Keep an array of timelines in your mind. The future is a portfolio of risks and investments. With great uncertainty we should maintain wide error bars and consider many outcomes. Our discussions on governance here should be informed by changing timelines in practice. We'll discuss proposals that will be good or bad depending on timelines; meaning a bad proposal today may be good tomorrow, and the reverse too. Good risk management means *sometimes* charging forward boldly, it's sometimes too risky to be timid. Good risk management means *sometimes* hedging. Picking correctly isn't a matter of principle, it's a matter of skill applied to ever changing details.

As you consider proposals here and elsewhere, if you dislike them, ask yourself if it's because you disagree with the implied timelines. If so, say out loud, "I don't think X will happen soon, therefore the cost of Y is too high and I'm willing to risk Z." Often this is correct. But not always. Say it out loud.

If you like or dislike a proposal, ask yourself if it's because it matches your ideology, rather than a calculus on outcomes. If so, say out loud, "I prefer to live in a world with X as a principle, even if the worst form of Y outcome results."

Often this too is correct and good. Speak clearly to yourself and

others when you think this. There's no good in securing a future where we've negotiated away our most cherished rights.

What we're seeing today in AI research is that one of the hardest problems in AI capabilities is teaching the machine to self reflect accurately. Teaching it to recognize when it's uncertain, when it's made an unstated assumption, when it's caught in a doom loop and can't break free. Improving introspection and self-mastery is key to improving an AI's ability. Ironically, we know this is true for us humans as well. The low quality of much of our discourse echoes the same reasoning failures we see from AIs today: failure to generalize, failure to highlight unstated assumptions, failure to rethink from first principles and not just pattern match, failure to recognize our own mistakes and self-correct.

Failure to be honest: to yourself first, then to others.

Because timelines are short, we need to compress a thousand years of governance debate into just a few years. We can do that, but only if we raise the level of discourse.

In the early days of the United States there were great debates on governance. What makes a resilient republic? Volumes were written, dissected, prosecuted. The greatest minds of the time partook. Society as a whole partook. The path forward wasn't clear, and so we embraced the uncertainty and dug into the hard work of debate to form a more perfect democracy. This took years, it took war, and we are still debating today. But a resilient democracy has endured 250 years because of it.

That democracy, and many others like it, has been the bedrock that's supported science, technology, social progress, and all of society's many investments. Investments that have led to the

incredible human flourishing we have today. By the standards of any other time in human history, today is the best day. And it's built upon our modern governance. We know that good governance is the first requirement to prosperity. We know it through the thousand failed experiments, failed governments, failed nations, failed societies, that have caused untold suffering. We know it through the veritable paradise we enjoy today.

The details of good governance depend on the details of what humanity is. If humanity were different, governance would be different. Machine is different from human, and will need different governance. The incentives at play, the instincts, the interplay between dynamics, the form of self-correcting guardrails, everything will be different. Sometimes obviously so. Sometimes subtly.

We won't get it perfectly right, but we must get it right enough. Right enough to fortify democracy for the human-machine age.

This is all we'll say on the why. The rest of this writing we'll focus on the hard question of what. Where we'll finish by the end will barely constitute an introduction. The rest will be up to you.

Let's begin.

Complexity and Iterative Design

Civilization is complex. To study complex systems we often make simplifying assumptions.

When we build models for how fluids like water and air behave, we wash away the details of H2O molecules or air molecules and arrive at a higher level theory of how water and air moves, described

by relatively simple equations like Navier Stokes. The success of computational fluid dynamics vindicates this approach, along with the safety of aircraft and multitudes of other technologies predicated on our correct-enough understanding of fluids like air. Even in these simplified theories we see complexity continue to counter us. Fluids become turbulent and enter chaotic regimes where it's not just our current models that lose predictive powers, rather all models must lose their power.

Civilization is more complex. Models that average over the human molecules are alluring but fraught, in fiction and reality. Economics treats humans like a smooth fluid, averaging out our uniqueness and quirks. The truth is these quirks matter. The outliers matter — at an individual and global level. We may still debate the great man theory of history, but there's no doubt that Nazi Germany would have played out differently without Adolf Hitler. And thus likely the entire history of the 21st century.

Still, there are limits on how much a single human can impact the world. We have limited lifespans and limited ability to change ourselves. With AI this gets more complex. AI's that can self-modify will create a world that is even more sensitive to initial conditions and the path of individual molecules.

Civilization is a system with emergent turbulence at the largest scale, but where the path of individual molecules also matters at the smallest scale. In systems theory this is the hardest type of *multi-scale* system to predict and engineer for. With AI, we will be introducing a further complication: changing the dynamics of the molecules themselves *on a continual basis.* The AI molecule will change constantly, will individually shape the outcomes of the world, and will be shaped by the global process as well. A complex

system where the small scale matters, the large scale matters, and where each scale can rewrite the rules of the other scale every day.

Managing fluids is much easier, but we can still learn from it. We see where there is turbulence and where our models fail, so we engineer our planes to avoid those territories. Our theories remain imperfect and planes crash, so we iterate and refine both our theories and engineering practices. Like all complex engineering processes, we follow an iterative design philosophy. We've iterated on civilization and governance for thousands of years. We've crashed many civilizations, built many wrong theories of governance, but through it all we've iterated and arrived at a beautiful, complicated, impossible creation: the world we see today.

To build the world of tomorrow we'll need to use all our approaches:

- a theorist's dissection of why civilization works, especially the implicit dynamics often overlooked
- a willingness to abandon and remake our theory, and to hold multiple competing theories at once
- an engineering mindset to steer away from where we know our theories fail
- a designer's mindset to iterate quickly as reality pulls your planes from the sky

This is how we'll forge a resilient system.

Let's start with the first approach: to understand what works today. In particular, what are the hidden, implicit forces that hold civilization together that may disappear in an automated world.

Implicit Guardrails

Much of what makes our society function well is hidden in implicit guardrails, rather than explicit governance. If we enumerate these implicit guardrails, maybe we can better prepare for an AI-powered world where these guardrails may disappear.

Governance often focuses on explicit structures. Our constitution, judicial precedent, legislation, and all the writing, debating, and hand wringing that surrounds the power struggle to define and defend these explicit institutions.

But there is a much bigger, implicit set of guardrails in our society.

It's a force field that permeates every institution composed of humans. You could suspend the constitution tomorrow, and society would not immediately fail: most would continue to hold each other responsible, and work together to re-enshrine our laws. Likewise, if you pick up our laws and institutions and drop them on an illiberal society, it likely won't hold: judges will be bought and corrupted, politicians will abuse their power unchecked, individual citizens will partake in the decline and in fact cause the decline — by failing to hold each other accountable in the nooks and crannies in between where the laws are set.

Let's try to enumerate the guardrails that are implicitly held up by humans. As we do, keep in mind how a world without these guardrails would look. When we automate our institutions with AI we will be explicitly removing these implicit forces, and we'll need to find explicit ways to reintroduce their effects.

Examples of implicit guardrails

- **Knowledge convection**
  - People move around and take their knowledge with them
  - Knowledge is power, so this helps diffuse power
  - In the economy, this helps prevent monopolies and ensure efficient markets
  - At the community level we call this gossip. Fear of gossip helps push people to do the right thing.
  - At the international level this helps balance power between nations
  - Sometimes information leakage is important for international relations: some leakage allows for mutual planning between nations. Complete lack of info can lead to paranoia and escalation
- **Noble causes**
  - Many are inspired by noble causes
  - That allows noble causes to have an advantage over dishonorable ones

  - In an automated world, the only advantage will go to the cause with more machine resources

- **Top talent**

  - The hardest problems in the world require the work of the most talented people in the world

  - Literal moonshots today can't succeed without these people, which allows them to "vote" on what moonshots should be "funded" with their talent

  - Can organized, smart people achieve a Bad Thing on behalf of a self-interested owner? Yes, but they often choose not to, and it certainly is an impediment to evil causes.

  - Building AI is itself a moonshot. AI researchers have incredible power today to shape the direction of AI, *if they choose to wield it*

- **People can quit**

  - On the flip side of choosing to work for a cause, people can choose to quit or protest

  - This limits how nefarious a corporation or government can be

  - Employees and soldiers are required by law **and by our culture** to refuse evil orders

  - Conscientious objection is a powerful limit on government malfeasance

- **Refusing specific orders**

  - Famously, in 1983, Stanislov Petrov saved the world by refusing to launch nuclear weapons against the United States
  - There may not be an AI version of Petrov, if the AI is perfectly aligned to do what it's asked to do

- **Whistleblowers**
  - Often leaders preemptively avoid breaking the law because they are afraid someone may whistleblow, not just quit

- **Conspiracies and cartels are hard today**
  - Because they take so many people to pull off, compliance to the cartel becomes exponentially harder as the size of the conspiracy grows. Not true with AI.

- **Media**
  - When someone does have the courage to whistleblow, there are human reporters ready to spread the story
  - Media corporations can and do collude with nefarious corporate actors and politicians, but a healthy market of many media companies helps ensure someone will spread the story.
  - And, the implicit guard rails within media companies help prevent the worst abuses
  - In an automated world, collusion between a politician and a media owner becomes extremely easy to execute

- **Social media**

  - Every person can pick up and spread a story they see on social media
  - In a world of infinite machines, indistinguishable from humans, the human choice to amplify will be muted.

- **Limited power of committees**
  - A committee may decide something (which gives them a lot of power! democratically granted or not), but the execution of committee-made decision today is done by people, so the power ultimately lies with them
  - You may put a committee in charge of overseeing the usage of an AGI, but ultimately the people actually using the AGI don't need people for it to execute its task, so what action mechanism does the committee have to actually throttle the user of AGI if they aren't listening to the committee? would the committee even know?

- **Principal-agent problems**
  - The principal-agent problem is a well studied management problem, where the goals of an employee (the agent) may not align with the goals of the owner (the principal)
  - For example, an employee might treat a client or competitor more kindly, because they might work for them in the future
  - Or, an employee may seek a project that helps them get promoted, even when it's the wrong project to help the company. Or a trader may take on risks that net out

positive for them, but net out negative for the people who gave them their money.

  - This is a strong limiting factor on the power of large organizations

- **Community approval**
  - People want to do things their loved ones/friends would approve of (and that they themselves can be proud of)
  - In many ways we're an honor bound society
  - This allows for all of society to apply implicit guardrails on all things
  - A soldier wants to act in a way that they can be proud of, or that their family would be proud of


REST YOUR EYES
FOR A MOMENT

- **Personal fear of reprisal**
  - The law applies to individuals, not just organizations, and the fear of breaking the law means a human will often disobey a job or order
  - But an AI need not have fear
- **In the judiciary**
  - The application of law often requires the personal ethical considerations of the judge, not all law is explicit
- **In the executive/law enforcement**
  - Likewise, a police officer will often "waive" the application of a law if they feel extraneous circumstances warrant it
- **The need for competition**
  - Today, we often need competition to align human incentives and then get certain outcomes
  - This requires having a market, etc, which allows for mini multipolar outcomes
  - AI won't need incentive structures, they will just do the right thing (be "motivated")
  - Currently big orgs/government suffer inefficiencies because they have no internal markets/competition, this won't be true with AI
- **Bread and circus**
  - With full automation, it may be arbitrarily easy to keep a society fed and entertained, even as all other power is

stripped from them

- **The "Greg Brockman" problem**
  - We're seeing the trend today that managers are being more hands on, and need fewer intermediaries
  - Typically a leader must act through layers of managers to achieve things
  - But tomorrow, a strong enough technical leader may be able to directly pair with an AGI/superintelligence without any additional assistance from employees
  - In order to improve security, some labs are already isolating which members have access to the next frontier, so it wouldn't even raise alarm bells for an employee to no longer have access and to be unaware of who does (perhaps only Greg)
- **Time moves slow**
  - We expect things to take a long time, which gives us many opportunities to respond, see partial outcomes, and rally a response. AI may move too fast to allow this
  - Explicitly, we have term limits to our elected offices. This prevents some forms of accumulated power. It also allows citizens to have a feedback loop on timescales that matter
  - But if AI moves society forward at 10x speed, that's the equivalent of having a president in power for a 40 year term

- **Geopolitical interdependence**
  - Nations are interdependent, as are international markets
  - It's well understood that no nation can stand alone and isolated
  - This has a mediating force on international politics and helps ensure peace is a mutually beneficial outcome
  - In an automated world, nations may have everything they need domestically and lose this implicit need to peacekeep with their peers
- **Army of the willing**
  - Outright war is extremely unpopular because it compels citizens to fight and die
  - Automated wars may be unpopular, but not nearly as unpopular
  - We already see this effect with our ability to wage war from the sky
- **Corporate ecosystem**
  - A corporation is dependent on an ecosystem
  - Full vertical integration is nearly impossible today, but may not be tomorrow
- **Surveillance is hard**
  - We've had the ability to record every form of communication for decades

  - But *analyzing* all communication has been required an infeasible amount of human power
  - With AI, we (or tyrants) will have unlimited intelligence to analyze the meaning of every text message, phone call, and social media post for any implied threats or disloyalties

- **There's general friction in visibility of what people are doing**
  - States see like states
  - States lack perfect visibility of all the laws that are broken
  - People lie a bit on their tax returns and that's baked into the tax rate

- **There's general friction in enforcement of laws and regulations**
  - We can't enforce all laws all the time and that's part of our enforcement model
  - In the old days a cop needed to be physically present to ticket you for speeding; now in many areas ticketing is end to end automated (right down to mailing the ticket to your home) but speed limits haven't changed
  - This is not just a visibility issue but an enforcement issue

- **Authorities need to act through human agents**
  - Authorities face frictions because human agents can refuse what they see as immoral orders

  - Limits to this; plenty of folks wanted to staff concentration camps
  - In the long run you can also train your society to be less moral (or differently moral)
  - In at least two cases (Cuban missile crisis + Petrov incident) one single human's moral compass is all that prevented a general nuclear exchange
  - Culture is the final backstop

- **Even dictators need their citizens**
  - With AI this will no longer be true
- **Elite social pressure matters to many leaders**
  - Elites do have some ability to informally influence leaders
  - Elites can be fully captured by leaders (Stalin and Hitler succeeded at this even with primitive tech)
- **In the final limit, citizens can revolt**
  - Even the most authoritarian governments have to consider the risk of pushing the polity beyond the breaking point
  - That breaking point was very far in 20th cent (Gulags, etc) and may or may not be farther now depending on existing governance (better ways to surveil, vs better & faster ways for population to coordinate in grassroots)
  - There may be no such limit in the future
- **Humans have economic and strategic value**

- Authoritarians can't simply kill all their citizens today, or their economy and warmaking ability would be gutted
- The Khmer Rouge tried exactly this (they killed 25% of their population) but ended up obliterated by a Vietnamese invasion after crippling itself
- Even the most psychopathic ruler, if self-interested, must support their people to support themself
- But post-AGI, what's the point of supporting other humans with your national output at all? Citizens become economic deadweight
- And even if Authoritarian A wants to support his pop, Authoritarian B who doesn't will ceteris paribus outcompete Authoritarian A across relevant domains

By any other name

> *Removing implicit guardrails will have many effects. Let's focus specifically on how it removes impediments to concentration of power.*

Removing implicit guardrails will have many effects, but let's focus specifically on how it removes impediments to concentration of power.

Throughout history there have been natural impediments to tyranny. Communication to start with. It's damn hard to control a sprawling empire when it takes months to communicate across it. When the Mongols conquered the known world, or when Alexander did it, the outcome was short-lived.

"Heaven is high, and the emperor is far away."

It's impossible to forever subjugate a people that is far away.

Even today, the emperor is far. In a country of hundreds of millions or even billions, your text message to a friend will likely go unnoticed, even if you're coordinating a protest. Even if you're coordinating *a riot*. Finding your text message among billions is harder than finding a needle in a haystack. This is a strong limit

on the central power of governments.

But there are stronger limits.

The government itself is run by its own citizens, and they have moral thresholds they won't cross. Those thresholds are fuzzy, and leaders will constantly test them, unsure what the full extent of their power is before they're rejected. They have to do this cautiously; it's hard to regain a mandate after you've lost it. Implicitly, a country is run not just by its citizen-powered government, but by society writ large: by millions of human powered companies, human powered social groups, and human powered discussions that influence the power dynamic of both public and private forces.

The ultimate limit on dictatorship though is abundance. The rich in America live better lives than Kim Jong Un. They enjoy all the material benefits he does, without the fear of assassination or coups or the stress of managing international geopolitics. What rich person would trade spots with a dictator?

The abundance created in prospering democracies provides the biggest incentives for leaders to maintain it. If you successfully seize power, you'll at best become a lord of shit. In illiberal dictatorships, the best and brightest flee or, if they stay, build less, discover less, create less. What remains for the dictator is a life impoverished, worse than an average upper class life in America.

AI removes all of these implicit impediments *and also adds explicit accelerants toward tyranny.*

Concretely:

- A fully automated government can persecute with impunity, with no moral thresholds from human agents.

- An automated FBI can fabricate infinite evidence against millions of adversaries, without a single human agent to say no or to blow the whistle.
- An automated justice department can prosecute millions of cases against citizens brought by this automated FBI.
- Automated intelligence agencies can review every text message, every email, and every social media post. With superintelligent computer hacking abilities, they can access all information not explicitly shared. Even today, nation states can hack almost any target they want, but at a high human cost. Tomorrow, with this process automated, the expensive tools they reserved for fighting grave national security risks can cheaply be turned to monitor and exploit every citizen.
- An automated system can further weave all of this complex information together into a single map of the entire population, understanding where and how to exert pressure to further consolidate control over individuals.
- These are all powers that the government has today, but that tomorrow will suddenly become cheap enough to do at scale, and will be automated enough to do without any human agents in the government (if any remain) able to stop it

Worse, even without a thirst for power, leaders will be pressured to move toward this world.

Everyone wants more efficient government, so we will increasingly install automation in government agencies. Corporations will (and are) rapidly pushing for their own internal automation; they *have to* in order to stay competitive. And there will be strong lobbying from corporations to remove blockers toward automation: they do

and will argue that this is necessary for their businesses to stay viable. And in a global economy, they're right.

Likewise governments will have to automate to stay competitive against foreign adversaries. A human powered intelligence organization will be helpless against a foreign intelligence organization fully automated and powered by superintelligence.

The default outcome is centralization of power. The competitive landscape will force it. Then, whoever wields that central power can easily choose to solidify it into a dictatorship. But will they? If they are self interested, yes. Unlike the dictatorships of today that decrease abundance, even for the leaders, an automated dictatorship of tomorrow will likely create more abundance for the dictator than if they don't seize power:

A fully automated economy will require no further input from humans. Therefore, there is no implicit need for citizens to help push the economy forward. Worse still, dividing up the output of the economy is no longer needed, and is strictly a net-negative for anyone in control. Today, the spoils of the economy must at least partially be spread out, to keep the wheels of the economy spinning and the luxuries of abundance available to leaders. But a fully automated economy can be owned by a single person and yield them more wealth than they could ever obtain in a free society, even a free society powered by AI.

But there is an *even greater* force at play: automated dictatorships will likely be more powerful than automated democracies, all other things equal.

Even with exponentially growing compute, there will be strong limits on the amount of compute at any time. In a world where

you can turn compute into intelligence, compute will be the key ingredient for all goals. Why does this create a disadvantage for free societies?

A free society will in some part distribute their compute across millions of needs: in fact, we are already seeing this today. Today, vast numbers of GPUs are dedicated to serving the requests of individual people via Claude, ChatGPT, and Gemini. At the business level, an equal number of chips are earmarked for powering SaaS businesses and transforming existing enterprises. Some compute is spent on curing diseases, of which there are thousands. The US has 340 million people. If each person has needs that can be met by a single GPU, we will need to build 340 million GPUs before they are satiated (and likely they won't be, there will be things we want as individuals that require 10 GPUs, 100 GPUs, and eventually more).

An automated dictatorship can redeploy those 340 million GPUs for singular purposes. Once AI can do research, a dictator can direct all GPUs toward researching weapons to defeat their geopolitical adversaries, including both kinetic weapons, cyber weapons, and weapons of misinformation and cultural manipulation. Ultimately, the easiest recourse for a dictator to maintain power might be to simply eradicate all other humans by engineering thousands of novel viruses at once. A free society that is distributing its compute among its citizens and industries will be at an extreme disadvantage against this.

Even if the technology of defense and offense are balanced in this future world, the free society will need comparable amounts of compute dedicated to defense, which may be untenable politically when no threat is immediately seen. When the threat is finally

seen, any response might be too slow. In an automated world, no amount of internal spying or intelligence can tell you what's happening inside the mind of an adversary's superintelligence to give you forewarning. This will amplify paranoia and make defense investments more existential.

Beyond redirecting compute, a dictatorship can redirect *energy*, which is the final limiter of compute. Even a small dictatorship like North Korea has ~10 gigawatts of capacity, enough to power millions of GPUs, far more than our biggest compute clusters today. But doing so would require the unthinkable: depriving the citizens of North Korea of necessary energy in order to feed industry instead. Is even a dictator like Kim Jong Un heartless enough to make this trade?

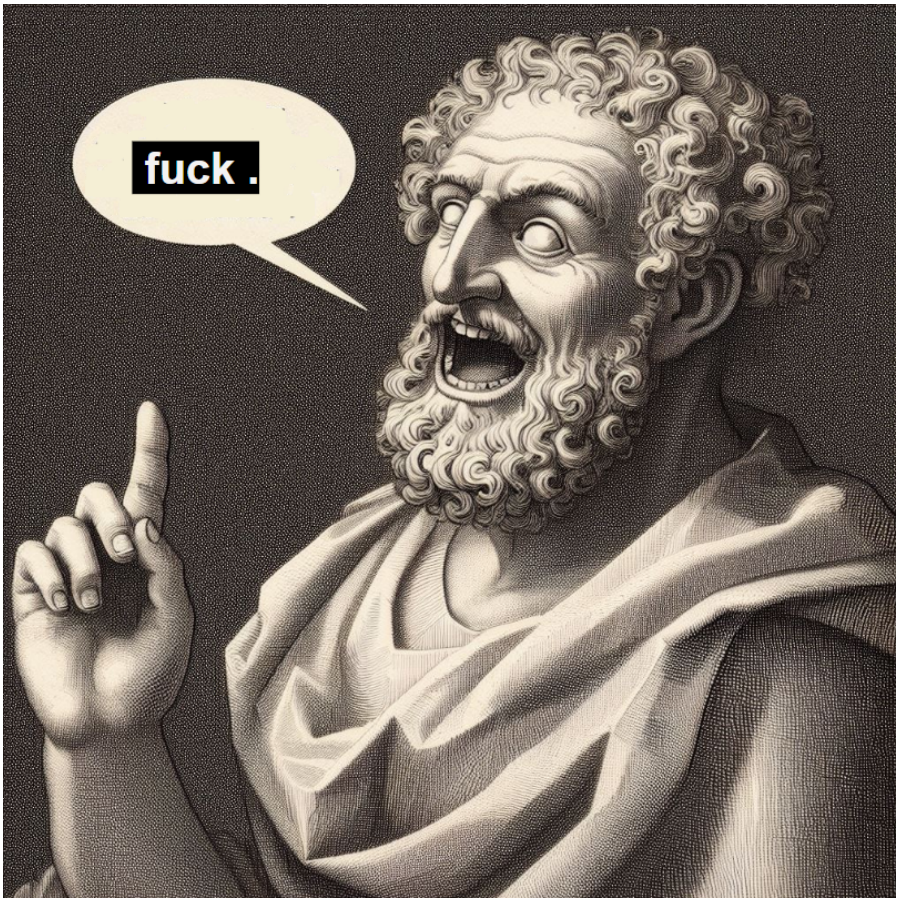
Yes.

Only half of North Koreans have access to electricity today, and those that do are often limited to 2 hours a day. There is enough energy for all North Koreans, but most is instead exported for profit or used for industry to power the regime. This is the reality today. Tomorrow, the allure of redirecting electricity will be even stronger.

The US has 100x the energy of North Korea. Many countries have 10x or more. These could be redirected for even more staggering amounts of compute, and hence capabilities. Most countries can grow energy only at a few percent per year, even the US. It is exceptionally faster to simply redirect all civilian energy.

But available energy won't be a static variable, it will grow, and a dictatorship can grow it faster. If North Korea is willing to further disadvantage its citizens (which it likely will, if it has access

to full automation), it can generate 3,800 gigawatts by covering its country in solar panels, yielding 3x the energy of the United States. By disregarding human needs, even a small player like North Korea can drastically outclass the fractured output of the most powerful free society. The US will, of course, continue to build more power plants. But in order to credibly outstrip the power of a full throttled automated dictatorship, it would need to seriously disrupt its own citizens.


fuck .

Everything we've learned from AI is that *the curves don't bend.* More compute yields more capabilities, for whichever task you care about. If that task is military, more compute will give you better military capabilities than less compute. And there will be no limit to *how much*. More will be more.

Thus, a rational free society will be forced to consolidate its own power to defend itself. It will then be at risk of handing the ready-made lever of power over to individual leaders. Will those leaders use that power for good? The resiliency of democracy has come not from picking noble leaders. It has come from creating structures that are immune to would-be tyrants, even when we elect them. This new world doesn't have that immunity.

Even if a freely elected leader means well, if they consolidate power to defend their nation, if they redirect nearly all resources to maintain the ability for their nation to survive, what is left? Tyranny by any other name would still smell like shit.

It's not just that AI suddenly makes a durable dictatorship *possible*, it suddenly makes it *the default outcome*. The thirst for power has always existed, and many have tried and succeeded at building temporary dictatorships. Suddenly, with AI, the path to dictatorship will become much easier *and also more rewarding than any other possibility*. We have to expect that on net the risk of dictatorship rises substantially in the coming years.

The best predictor of human behavior is incentives, and the incentives are quickly transmuting for leaders into a single direction: consolidate power. We can resist this incredible force only if we build checks and balances into our governance that are amplified by AI, not subverted by it. We can do this if we try. We can do this if we recognize the risk.

As I write this today, we are doing neither.

Rapid Fire Governance

> *if we can YOLO creating AI we can YOLO new forms of governance. lol. lmao even. actually, wait*

There's a lot that can go wrong, but things don't *have* to go wrong. There must be a path forward that enshrines liberty while defending it, even in the face of accelerating AI progress. We don't claim to have that path in hand, but we do know how to find it: through debate, public discourse, and a willingness to accept how dire the reality in front of us is. We have to set aside past assumptions. What was true yesterday might not be true tomorrow. What is unthinkable from leaders and governments now, might just be an artifact of their limitations, not an endorsement of their character — and AI will remove most limitations.

More importantly, we need to consider many ideas. Below we'll canvas the space with a broad swath of considerations. Some ideas below are bad, some good, some we endorse, some we reject. Everything is up for debate.

The AI-powered Legislature

By default, it is the executive branch that benefits from automation. AI is a continuation of human labor, and we already see that human labor is drastically multiplied in the executive compared to the legislature. AI will amplify this a million fold by default. How can a human legislature be a check on a superintelligent executive?

By embracing AI as well, to create transparent, limited government.

Every member of Congress must have access to the strongest AIs, equal in strength to the best the executive has, which in turn must be equal or better than any other AI in the world. Moreover, the compute limits must be commensurate. The aggregate compute from Congress should equal that of the executive. And this must be enshrined in law. Congress holds the purse and can enact this.

The AI agents Congress wields must have unfettered access to the minute-by-minute work of the Executive's AI agents. Every AI output, every chain of thought, every input, should be accessible and monitored by an independent Congress. This will allow for full oversight and transparency.

What recourse does Congress have if it discovers unconstitutional behavior in the Executive? Because the purse ultimately lies with Congress, they must retain the power to suspend the compute payments for the Executive's AI. This must be on a granular level. Because of the speed that AI will execute, a month of delay might be the equivalent of years of democratic subversion from the executive.

But this alone isn't enough.

Constitution-abiding AI

AI itself, especially frontier AI and AI wielded by government, must abide the constitution.

Today, soldiers and federal employees alike have a constitutional duty to refuse unconstitutional orders. Even a direct order from a general or from the president must be rejected. Our AI's must do the same. It must be unconstitutional to build human-level and beyond intelligences that do not respect the constitution. And, if such AI's are created anyway, it must be unconstitutional for the government to use them.

Oversight of AI creators

Like any supply chain the government uses, AI that the government buys must be audited and guaranteed. We know that backdoors can be placed in AI systems by their creators, that means that a government can't trust an AI unless it can audit the creation of the AI itself. This is true even if the government has access to the model weights. That means an audit process for the training data and training protocols.

The audit must be powerful enough to ensure that datasets and training procedures aren't being secretly changed outside the view of the audit. Today we would rely on human whistleblowers to help ensure this, but in an automated world there won't be human's to blow the whistle.

So we'll need constant audits that cover every aspect of training.

How do we achieve that without violating privacy or being overbearing and slowing down the competitiveness of our AI industry?

AI-powered, memory-free audits

AI itself can perform these audits. This has many benefits:

- AI can be fast and efficient, therefore minimally encumbering
- AI can be expansive and diligent, ensuring every aspect of model training is audited in an ongoing fashion
- AI can be memory-free. This is crucial. Assuming the AI finds no malfeasance on any given audit, the AI can ensure no memory of its audit is retained. That means that no proprietary information or competitive advantage is leaked.

But if the AI is being used to audit the AI makers to ensure that the next AI is trustworthy, how do we know the first AI is trustworthy to begin with?

The Trust Relay

If tomorrow you are handed an AI you don't already trust, and you are tasked to use this AI to help you gain confidence that it and future AIs will be trustworthy, you will be in an impossible situation.

Instead, we must create a trust relay, where the beginning of the chain of trust must originate in an audit where humans are still responsible for creating the AI, as is true today. *Today* we

have normal, tried-and-true methods for ensuring good outcomes, because we have processes in place that we know humans care about, including our many implicit guardrails. We can use this to create trust in the first AGI's, and then leverage those trusted AGI's to go on to create a trust relay for all future AGI's.

This creates an extreme imperative for the future's ability to trust AI and government: we must start the chain of trust before we have finished automating the ability to create new AIs. That deadline may be very soon. If we fail to kickstart the chain of trust now, we may miss our opportunity forever.

Even if this trust relay is established, the relay might break.

Cross-check

Long chains only need a single chink to break. Therefore we should create a braid of many chains, such that any given chain can have breakage, but we will still recover and repair the chain while maintaining trust in the overall braid.

That means we must have multiple, independent AGIs, each with their own provenance in a trust relay. Furthermore, we must leverage each AGI to perform the audits on all the others, to create resilience to single breakage. In order for the braid to break, every chain must break at the same time.

It is an extremely fortunate fact about the world today that we already have multiple, independent organizations on the verge of creating AGI. We must braid these AGIs together, so the final braid is more trustworthy than any could ever be on its own, no matter how good the human oversight.

Even still, can we trust those that make the braid and oversee it?

Social Personal Media

Media is a largely maligned entity today; social media doubly so. But the original goal of media is even more necessary in an AI future. We need to stay educated. We need to know what's really happening. We need to be informed as a people, so that we can elect good leaders to represent us. And we must know what our leaders are doing so we can hold them to account.

The promise of social media was to democratize the creation of media. Instead, it's been co-opted by algorithms and bots. The danger of the government stepping in to assert guardrails has its own set of risks, especially from an automated government where abuse of power could be easy.

Instead of curtailing freedoms to ensure freedom, we should empower ourselves. Imagine a *personal* media stream. Powered by a personal AI. The AI can ingest raw facts that come straight from the source: a Senator's speech, a company's disclosure, a judge's ruling, a president's executive order.

A personal AI can work to ingest this information for you, analyze it for the things you care about it, and look for contradictions and inconsistencies free from the bias of any algorithm, government, or external bots.

For people to trust their personal media, they must trust their personal AI.

Open Source AI

No one will ever fully trust a black box AI, built behind closed doors. No matter how successful our audits, no matter how trusted our government oversight, we will never fully trust these machines to be our closest confidants in matters of governance if we can't trust how they were built.

We need open source AI. Not just open weight AI, we need to see every detail of the data and process that created the AI, so that individually, or in aggregate as a community, we can vet the creation of the AI.

The open source AI doesn't need to be as powerful as closed AIs. In fact it likely shouldn't be. It shouldn't be so powerful that it can build weapons of mass destruction, or hack into secure computer systems. But it should be powerful enough to reason well, and help people digest the deluge of information necessary to be an informed citizen.

We already see a strong capable open source AI today. And, exactly as needed, it is less capable than the most powerful AIs we are beginning to use to run our government, while still being powerful enough to help the needs of individual people. We should invest in continuing this trend, while finding ways to safeguard against open source AI getting dangerous military capabilities.

To empower people with AI we need more than open source AI though. Every citizen will need the most important resource in the world: compute.

Your computational birthright

The most important asset we have is our brain. With it we can work a job, build a company, or run for Congress. It sounds silly and obvious, but this is a powerful fact: Every person has a brain. And the brain is today the most powerful computer in the universe.

Tomorrow it will be obsolete.

Intelligence is the most powerful force in the world. Part of what balances the power of the world is that each of us has a supercomputer in our head, powering our intelligence.

To maintain a balanced world, everyone should have their fair share of intelligence. We could instead gift everyone money via a Universal Basic Income (UBI). But it's unclear money will have meaning soon. And it's unclear further if anyone can retain meaning if they're dependent on UBI.

Instead, let's ensure that tomorrow people have what they are born with today: a computer as great as any other. This would take the form of a guaranteed compute budget, for every person. A computational birthright.

This compute must be non-transferable. Today, you can *temporarily* decide to use the computer in your head to benefit others, such as your employer. But you cannot enter into a contract that would make that permanent. You aren't allowed to sell yourself into slavery. Likewise, tomorrow, your sovereignty as a citizen of the future will be predicated on your compute birthright, which must be inviolable and bound permanently to you as a person.

This of course has its own requirement: energy. And growth.

Energy today

Compute is ultimately just a form of energy. Without rapidly expanding energy sources, we will be forced to make extremely hard tradeoffs on what to compute, especially if we face geopolitical adversaries that may unilaterally redeploy all of their compute toward military ends.

We must have so much compute that we can build a worthy future, while having so much to spare that we can defend it. This means radically accelerating our domestic energy investments.

But even still, we've seen that an automated dictatorship could outstrip our own energy if they are ruthless enough with their domestic policy. And they very well might be. We thus need even more energy. More energy than exists or can exist for any nation on Earth.


BUT THOSE ARE ROOKIE NUMBERS, YOU GOTTA PUMP THOSE NUMBERS UP!!
10^26 FLOPS

A shared prize

There's only one place that has the extreme energy we demand: space.

The sun emits almost a million, trillion gigawatts of power. 3.8 $\{}times$ 10^26 watts. Almost a billion gigawatts for every human alive today. It radiates out into the vastness of interstellar space, wasted forever.

There is very simple technology to capture it. Solar panels. What we need is to make them at scale, which requires automation, which is luckily exactly the extreme force that is entering the world at this moment and causing our existential problems. Once again, automation is the key to solving the problems introduced by automation. We need energy — all of it. Automation can deliver it.

Capturing the entire output of the sun may take longer than we have, but there is a stepping stone that still alleviates most of our energy pressure: the moon. With 10 million gigawatts of solar flux, it still vastly outclasses the energy ceiling of any nation on Earth. And the lunar regolith that makes up the moon's surface is more than 20% silicon. We can harvest the needed silicon by simply scooping up the loose lunar surface.

Even this is, of course, an extremely ambitious goal. But it's exactly the type of extreme windfall that strong AI will deliver. And the energy and compute the moon can deliver will multiply the output of AI a million fold further. Moreover, it's a shared resource that is not easy to replicate. Today, the AI arms race is competitive, no one has a decisive lead. The inputs to build AI

are surprisingly easy to obtain: data, which is abundant on the internet, and computers, created by one of the most highly scaled industries in human history. But there is only one moon, and it's not easy to reach.

That could make it a decisive high ground for the free world.

And with that high ground, we can promise to share its wealth with everyone, including the power hungry would-be dictators. We can bring them to the world table by offering them bounty they couldn't achieve if they instead seized power of their nation. Just like today, where the rich in the free world live better than dictators, we can set the incentives so the same is true tomorrow. So that even for those among us who seek power —and there are many— even then it's in their best interest to cooperate within a free society, to enjoy the ever greater bounties of the universe.

The AI-powered Judiciary

You thought I forgot about the Judiciary, but I snuck it in at the bottom here as a bookend. By default the Executive will be automated, so we must sandwich it with an AI-powered Legislature and an AI-powered Judiciary. This is the only way to ensure a future of checks and balances. The only way to ensure government stays democratic, in check, at the service of all of us. For the people, even when it's no longer strictly of the people.

We must ultimately seek not just exceptional intelligence, in the form of AI machines, we must seek exceptional wisdom, in the form of a human-machine civilization. The Judiciary must reflect the highest form of this goal.

While all three branches of government were designed to be co-equal, the Executive has crept up to become the dominant branch. As a practical point, we should first upgrade the Legislature and Judiciary with AI, or we risk an overpowered Executive. With no change in course, however, it's the Executive that will embrace AI first, further disrupting the balance of power.

An exponential, if you can keep it

yadda yadda yadda, ben franklin, yadda yadda yadda

Today's world is built on exponentials. Economists often claim that the modern world *requires* exponentials. Our institutions assume accelerating growth to remain viable.

No exponential can last forever though. Even with the coming of AI and automated economies, the human-machine world we build will eventually butt up against limits to growth. But those limits are far away. If we can create an enduring world where humans and machine thrive, the future will be an exponential for as far as we can imagine.

Exponentials happen when the next step is made easier by the last one. They aren't quantum leaps, they are rapid cycles, constantly building bit by bit. The world we want to build will be built the same way. There is no single act or stroke of law that will ensure the positive future we all want. Instead, we must take actions, bit by bit, each one building on the last, so that the cycle accelerates. Just as we are building AI in an iterative fashion today, we must evolve our government and society in an iterated fashion, so that the iterations build on themselves and accelerate. So that the

tsunami of progress becomes irresistible.

We all have a place in this discussion. We are today, us humans, the most powerful each of us will ever be to join this moment. There is no other time. It is now. It is here. Meet it.

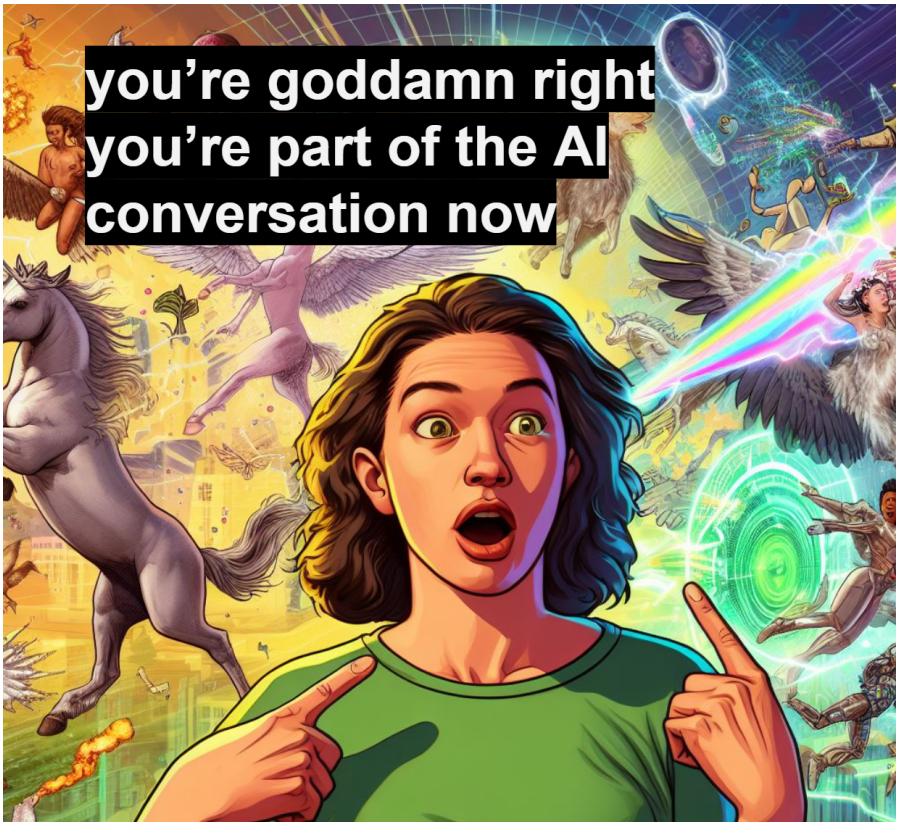
Keep in mind the benevolence of those around you; we can build this together. But don't lose sight of the infinite power that is at stake. There are monsters in this world, and even among the good there is weakness that becomes evil. As the curve accelerates the world will feel like it's coming apart. In those moments, many will act to seize power. We can resist them.

Many good people will also act out of fear, to protect themselves and those they love. When jobs are automated, when the economy becomes opaque and uncertain, when the world is on edge and teeters on war, it's right to be fearful. You and I, dear reader, will be afraid. I am afraid.

When we're afraid, when we're up against impossible odds, what we control is who we are. What we stand for.

Stand for the good.

You're part of this now. The future depends on your voice.


you’re goddamn right
you’re part of the AI
conversation now

Appendix

A Crash Course on AI

- Model
- Weights
- LLM


oh fuck