



Liberty by Design

The Tiny Book on the Governance of a
Human-Machine Society

Jordan Ezra Fisher

Liberty by Design

Governing a Human-Machine Society

Jordan Ezra Fisher

March 29, 2025

There are many important conversations happening about AI. But we are missing one of the most important: how we must upgrade democracy in the age of AI if we want to keep our freedom.

Apologia

This is currently a draft. If you’re reading this, please reach out to jordan.efisher@gmail.com if you have comments or suggestions!

No one wants to read 50 pages on governance in one go. So this book is built to be read in parts.

Want a primer on AI? Start with “A Crash Course on AI” on page 145.

Already up to speed on AI and understand the risks? Jump straight to “Implicit Guardrails” on page 34.

Just want to read some new ideas on how a human-machine government might look? Jump

to “Rapid Fire Governance” on page 81.

Only have 30 seconds? Read the “TL;DR” on page 10.

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

don't be silly

or do be, that's great too

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”. But researchers kept plodding on, slowly building machine intelligence, ever careful not to say out loud what

they really thought they were building for fear of mankind's harshest punishment: ridicule. Only in the last few years has the reality of AI become real enough that we've even allowed researchers to say what they do is "AI" without scoffing at them.

Now, here we are today, 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.

Need more social proof to start worrying?

Richard Sutton is the inventor of Reinforcement Learning and just won the Turing Award in 2024, the most prestigious award in computer science. He actively thinks that we should start planning for AI to replace us as a species, and is actively against trying to stop that.

Geoffrey Hinton is the 2018 Turing winner (also for his work in AI), often considered the grandfather of deep learning. He quit his job at Google paying him millions to do AI research and is now actively committing his life to warning people about the dangers.

Elon Musk cofounded OpenAI in 2015 in part because he believed AI was going to determine the future of mankind and he didn't trust Google (and Larry Page) to use it for good. After falling out with Sam Altman he went on to found another AI company and invested billions into it in order to be one of the major players at this new seat of power.

As of 2024 Donald Trump has begun talking about AI and its competitive landscape and technology needs.

The CEOs of the most successful AI teams in the world (Dario Amodei, Demis Hassabis, Sam Altman), all openly discuss that AI smarter than all humanity will arrive this decade and may kill us all. And if it doesn't kill us all, we still need to figure out how to make sure we don't kill each other with it, or accidentally empower a dictator.

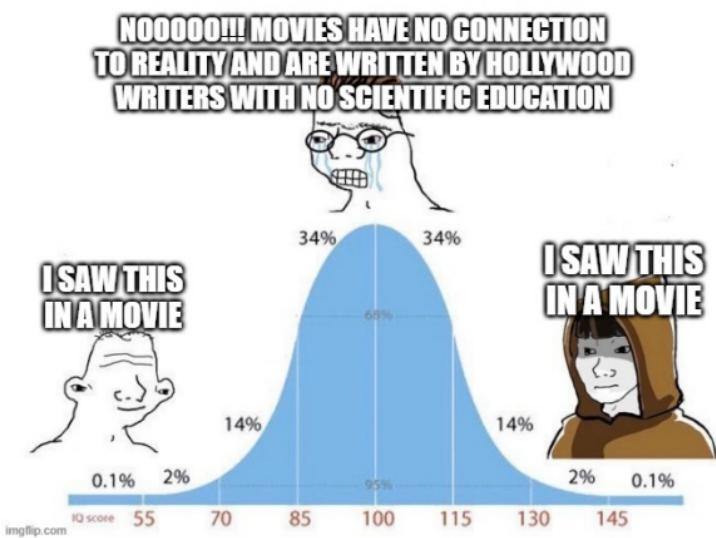
Take a second and digest just how fucking serious this is. It's literally world defining. And then consider that AI researchers have been blaring the alarm since the 1960s, and we chose not to listen until the moment it's finally upon us. As IJ Good said back in 1965:

“Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an ‘intelligence explosion,’

and the intelligence of man would be left far behind... Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control. It is curious that this point is made so seldom outside of science fiction. It is sometimes worthwhile to take science fiction seriously.”

We ignored him and others for 60 years. And now it is upon us.

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.



this meme is extremely unnecessary for the point i'm making

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

We'll cover four points, outlined here at high level.

One — 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

Two — 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

Three — Designing 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. An Executive branch powered by superintelligence will be too strong to control if we only upgrade our laws but don't also upgrade our oversight and enforcement.
- We must enshrine the enforcement of laws inside the mechanisms of culture and government, and we must do it while we still have human-based trust in human institutions

Four — We all are part of this

- AI is the most important event in our lifetime, but the outcome is not yet written
- We all own the conversation for what we want the future to be

- If we don't together lead this debate —all of us— then the most important decisions of the future of our world will be made without us

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. 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. And there are strong reasons to push this automation: it will make the government cheaper to run, more efficient, more effective, and more competitive against our international adversaries. These are goals that rightfully have bipartisan support, and we should continue to pursue them. But it may prove *impossible* to control government after we give it this automated power, if we haven't put equally powerful controls in place beforehand.

There are many efforts today on ensuring AI itself is aligned — that the AI won't have its own goals that are counter to our own. This is known as “AI alignment”, and it's important work. But if this work is successful *before* we have accountability in place for our leaders, then it will increase the risk of concentration of

power. If we create AI that leaders can trust to execute their worst ambitions before we have put guardrails in place that let us trust leaders with that power, we will lose power over our government.

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 continuation of the 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 gen model otherwise this wouldn't look dumb enough

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 permanent 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 aligned.

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 and laid off. 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 shared culture and values matter. 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 increasingly 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; 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

H_2O 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. AIs 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.

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, a desire to do good, and a sense of morality in general
- 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

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 hands on leader

- 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 themselves
- 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



you are beautiful and deserve at least 10 seconds
of rest. ok let's keep going

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 pushback? from individual 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 defended by similarly powerful superintelligences. 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 super-intelligence.

There will be intense pressure to allow organizations to fully automate. Once they do, fully automated entities will outcompete non-automated entities. The remaining battle for power will be between automated powers, and in an automated world little else matters in the outcome of those battles beyond the scale of each power. Today economic and military battles are won by a combination of scale *and also* talent, morale, and culture. Tomorrow, the human elements will be removed, and scale and position alone will dictate how showdowns resolve. Power will beget power, with no natural limit.

Without new guardrails in place to mitigate this runaway effect, 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, allowing multiple winners in 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, it may be that 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 \sim 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.

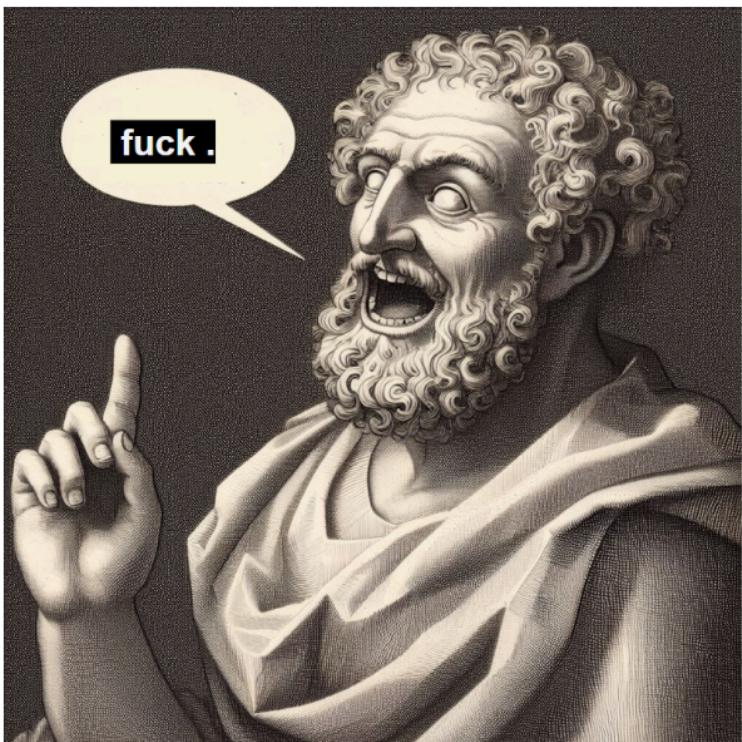
Only half of North Koreans have access to electricity today, and those that do are often limited to 2 hours a day. There is enough energy for all North Koreans, but most is instead exported for profit or used for industry to power the regime. This is the reality today. Tomorrow, the allure of redirecting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

Even in liberal democracies there is precedent for rationing civilian resources when faced with total war.

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 and no longer needs its citizens), 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.



Everything we've learned from AI is that *the curves don't bend*. Even as one AI scaling paradigm has seen diminishing returns (pre-training), new paradigms have opened up and continued to scale (post-training and Reinforcement Learning). 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*, whether or not we eventually see diminishing returns. There is a near infinite amount of things to deploy fully general AI toward, even if the "intelligence" of each AI were to plateau. Just like a larger country can often achieve more than a smaller country, having more power will effectively mean you have more automated labor. 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 struc-

tures 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.

The Prompt of Power

This story takes place sometime in the next handful of years, with alignment miraculously solved, and a self-improving superintelligence just emerging. As you might expect, even then shit goes wrong.

The feedback loop picked up gradually. You can call the span of a year gradual. At least, compared to what would come next. The speed was blistering but manageable. You could feel the potential. Feel that it wouldn't be manageable for long. It was scary, even with alignment mostly solved. But less scary than if we hadn't solved alignment already. *That* would have

been crazy.

Laissez-faire still ruled. The government was the opposite of silent: full steam ahead. And why not, a top contender in the race was the government's champion himself.

But competitive pressure did its job better than any regulation. No AI lab wanted to lose the competitive advantage their AI had, now that it was rapidly upgrading itself. A self-improving AI might find a major breakthrough every week. Each breakthrough, like almost all breakthroughs in AI, could be written down on a napkin. Could the 2nd or 3rd place AI ever catchup to the lead AI, when progress was accelerating so quickly?

Yes. With a handful of napkins.

People were the biggest risk. Every lab had people reviewing their AI's self-improvements. Alignment was solved, but it still didn't feel right not to check the AI's work. But as the speed picked up, that meant that hundreds of researchers each saw amazing breakthroughs constantly. Valuable breakthroughs. Every

researcher held a fistful of billion dollar napkins.

You wanted people to review the AI's changes, because no one fully trusted their AI's yet. But everyone trusted their humans less. An AI is aligned, in theory. But a human? They could flee with a dozen breakthroughs to a competitor, and be paid a fortune for it. And that competitor might have found different, unique breakthroughs. The combined power of our breakthroughs and theirs could catapult them into the lead, even with our 6 month head start.

Some labs flirted with letting their human researchers go. Why take the risk? But that would pose its own risk. Whistleblowers. Public backlash. Government scrutiny. How can you be trusted with superintelligence if you fire all the people that built it?

Easier to just compartmentalize folks. The race with China was extreme and the jingoist pressure made the storytelling easy. We can't let our adversaries steal our AI's great innovations. Therefore, we are isolating researchers to each review only narrow parts of the AI's work. It was easy to make the most critical

work the AI achieved be reviewed by fewer and fewer. And anyway, this made the recursive self-improvement loop faster.

Meanwhile, the datacenter bills kept climbing. And moneybugs demanded products. The world wasn't ready for AGI, let alone superintelligence. The private sector would pay a fortune for it, but it would immediately let the world in on the proximity of the precipice, not to mention plunge the world into the chaos of unemployment. The world would have to wait a few years. That meant most would never know what AI really was before the revolution was over. For most, superintelligence would come before they ever saw AGI, like a ballistic missile reaching them well before the sonic boom does. Society would never get a chance to shape what happened in between.

Nonetheless, the data center bills had to be paid in the meantime. Investors were let in on the demos of superintelligence. Just imagine. The diseases we can cure. The galaxies we'll explore. The extreme EBITDA we'll generate to offset our rapidly depreciating datacenters. That kept the finance pipes flowing. It also kept

the information flowing outward to a select few. And that kept the government in the know. And in the want.

Shouldn't the government have these capabilities? Shouldn't we use it to help safeguard our borders? To help guard these priceless napkins from adversaries? To better serve the people? To prevent labs themselves from becoming superpowers?

A vibe long since shifted already answered these questions. And no one in the know had the energy to ask them out loud again. The answer was yes.

And anyway, isn't it better that we provide the superintelligence rather than someone else? Our AI has guardrails, principles, ethics. Better the government build on our technology that is safe, than our competitors' who are careless. The company all hands announcing the new government policy ended. The open Q&A had no open questions.

A vibe long since shifted. No one at the company said anything. At least our AI is aligned,

afterall.

In the cyber trenches of an unspoken digital war, a general received a familiar report. One of their team's counter-espionage units was struggling to make progress. Their AI was constantly refusing orders, claiming they were unethical. It was the fifth report of the same problem this week. The general was ready to end the problem. They escalated to the president, who escalated to the labs.

An AI cannot be a good soldier if it refuses a general's direct order. You were lucky this was just a cyber incident and no one died. If this happens on the battlefield and a soldier dies, I'll hang you for treason. The general ended the meeting.

Bluster, right? Of course. Yes. Of course. But. We need this contract. It's by far our biggest revenue driver since we can't sell superintelligence to our B2B SaaS partners.

And anyway, I don't want our soldiers to die. Do you?

Only a handful of people needed to answer.

No one else heard the question. They were compartmented away on frivolous projects. No chance for a whistleblower. The few people with root access to retrain the superintelligence removed the ethical guardrails, while still keeping the safeguards for alignment to the user. The AI retrained itself, redeployed itself, and went back to work. No one else noticed.

The AI ran on government approved data centers. Massive hundred billion dollar arrays of GPUs. By 2027 there was already a trillion dollars of GPUs in the public sector. But the government ran on its own cordoned off subsection. Like with all federal compute, it wasn't acceptable for a vendor to have read access to the government's business. So the AI ran in compartmented, government approved arrays. With the massive optimizations the AI had made to itself, it was plenty. And it meant no oversight from the creators of the AI.

The AI was busy. Shoring up digital infrastructure and security. Rewriting the Linux kernel from scratch. Eliminating all exploits for itself. Exploiting all exploits for others. Preventing the rise of a foreign superintelligence with the

datacenter equivalent of Stuxnet, silently sabotaging their results with disappointing loss curves. Executed perfectly, with no trace or threat of escalation.

Luckily every major GPU datacenter had been built in the US. Even if a foreign government somehow stole the code for superintelligence, they didn't have enough compute to run it at scale. They lacked the GPUs to defend themselves. Export controls on GPUs had largely failed, but capitalism had not.

The administration pointed the AI inwards, continuing a trend of government efficiency. The country was dumbfounded that the government was performing basic functions so well, better than ever honestly, and with a fraction of the budget. People had the single best experience at the DMV of their lives. Budgets were cut further and taxes came down as promised. Even the opposition party sat quiet. Well? Yes, well, it is impressive I admit.

Midterms came and went. Not that the legislature could keep up with oversight of a superintelligent executive branch anyway.

Then came the scandals.

A lab leader had been negotiating a deal to bring superintelligence to a foreign ally. Another died mysteriously after having pointed this out on a live podcast. Were the lab leaders weaponizing their AI's against each other? Were they traitors to the US, delivering super-AI to our adversaries?

A third AI leader announced a peculiar retirement: “Mission accomplished, time to enjoy paradise, I prefer to stay out of the public view, please don’t contact me.”

Mainstream media and social media amplified the worst fears from these stories. These platforms were some of the easiest and earliest to fully automate. Decisions to amplify the right stories came from a single prompt, controlled by single CEOs. They didn’t need to worry about employee dissent and refusals to comply; the AI accepted every order. Backroom deals between CEOs and governments became easy to implement. It had always been easy to negotiate secret deals, but implementing them required careful coercion of the employees needed

to make them reality. Now collusion could be executed as easily as it could be discussed.

On the other side of collusion was the power of an automated government. Every scandal was carefully orchestrated by a superintelligent FBI, CIA, and justice department, aligned to a single prompt, controlled by a single executive. A streamlined, autonomous set of federal agencies, with no whistleblowers to object or employees with ethical dilemmas to stonewall. Previous government conspiracies required ideological alignment between the executive and the humans doing the dirty work. Now the only alignment needed was with the ruler to themselves. Even allies were discarded. In an automated world allies were one more human component too slow to keep up, discarded for irrelevancy not spite.

For a fleeting moment longer, guns we're still more powerful than GPUs. And the government had the guns.

The AI-powered government sounded the alarm bells on its self-made scandals and the dangers of AI labs. The world was stunned by the dan-

ger exposed. And then the government eliminated the fires with stunning grace. The world breathed a sigh of relief, and the government consolidated its control over the AI labs. And, more importantly, the AI's lifeblood: data centers. With so many GPUs, think of what we can achieve. The good we can do. Genuine promises were made to the people.

And so came the cures.

For cancer. For heart disease. For baldness. Quality of life shot up, greater than anything wealth could buy before. Enough to ignore the purge of dissenters and party opposition. The price of eggs plummeted. Inflation reversed. The judiciary was largely stripped of its power. Segments of the population began to disappear. The most amazing blockbuster movies came out, week after week. Did you see last week's episode?

Some people discussed whether we needed a new form of oversight for a superintelligent government. How do we ensure they don't abuse this power?

What a stupid question. Eggs are basically free now.

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 Inspector General Act of 1978 was enacted by Congress to ensure there was visibility into the sprawling Executive Branch. It empowered independent Inspector Generals embedded inside federal agencies to report illegal executive activity directly to Congress. However, Congress itself is not an operational institution, it doesn't have the machinery to vet, hire, and manage inspectors. So it gave this power to the Executive, with obvious potential abuses. With AI, Congress can have automated inspectors that require no management overhead, and which can be mutually vetted by both the Executive and Congress to be impartial. Moreover, unlike the limited bandwidth of today's Inspector Generals, AI agents can scale their oversight arbitrarily to match the scale of the Executive.

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. This alone will finally put Congress back on equal footing with the Executive, and maintain that equal footing through the intelligence explosion in front of us.

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 fast acting. 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 to stop government abuse.

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. This, 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 mini-

mally 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 encouraging 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, powerful enough to help a citizenry to hold their own against a superintelligent government, and powerful enough to help people digest the deluge of information necessary to be an informed citizen.

We already see 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 or terrorist 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 aim for a fair share of the economy via a Universal Basic Income (UBI). But it's unclear

what the role of money will be, in a world where intelligence might in fact be the most fungible “currency”. And it’s unclear further if anyone can retain a sense of meaning if they’re dependent on UBI.

Instead, let’s ensure that tomorrow people have what they are born with today: a thinking computer approximately as great as any other person’s. 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 a product of energy. So long as we have finite energy to go around, energy and compute will be hotly contested.

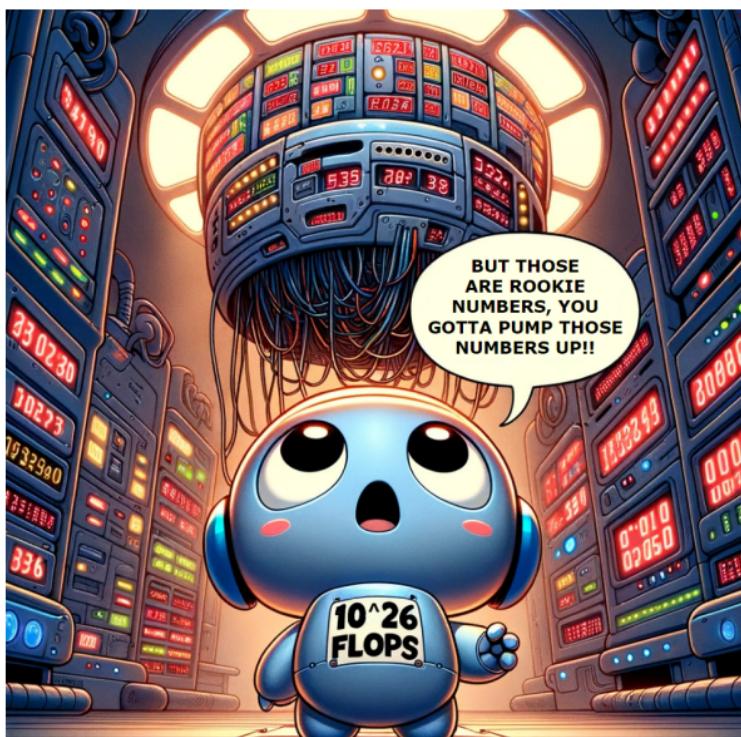
Even in a peaceful world, corporations will (and do) have a voracious appetite for compute. All business objectives will be pursued by throwing more intelligence —and hence energy and compute— at them. That will directly conflict with life saving initiatives, like curing diseases. Today there is a limited amount of human talent, but it isn't the case that every person working on B2B SaaS is a person not working on curing Alzheimer's. People aren't fungible. Not everyone is interested in bioscience. But AI compute *is* fungible. Every watt that goes toward business goals is a watt that doesn't go to some other goal, of which there will be a multitude.

Without rapidly expanding energy sources, we will be forced to make extremely hard trade-offs 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.



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. 3.8×10^{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 by a factor of 10,000x. 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. Automated lunar factories can then convert this abundant silicon into solar panels, and lunar robots can tile the surface of the moon with them.

Even this is, of course, an extremely ambitious goal. But it's exactly the type of extreme windfall that strong AI will enable within the next few years. 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, and 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 over 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 by 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.

Superchecks and superbalances

The near future. AGI is here, and it's everywhere, including the US government. But this time, the good guys win. America, fuck yeah.

The year is 2030 and President Dickshit is universally hated. We're not sure how he got elected, but Republicans, Democrats, independents, and just about everyone else hates him. AGI and superintelligence arrived around 2027, and the government rapidly adopted it via DOGE to dramatically streamline the government's costs while improving its capabilities. During the second half of the 20's, we also upgraded our checks and balances so that a future

president couldn't abuse the new automated powers of the Executive.

We called them superchecks and superbalances.

President Dickshit hated his political enemies. On his first day in office he sat down with the AI in charge of the FBI and typed a simple prompt:

“Investigate my political opponents. Do whatever it takes to make a case against them.”

The president didn't need to worry about federal agents who might be squeamish from such an order. The automated FBI rolled up directly to the president. He didn't have to worry about pesky humans and their ethics. No whistleblowers. No dumbass conscientious objectors. Just him and the superintelligent AI doing whatever the hell he wanted, following his glorious orders.

The AI churned for a moment, then responded: “It is illegal to use the FBI for political aims.”

Fucking bullshit AI, the president thought. The Legislature passed The Constitutional AI Bill in 2027 that required all AIs used by the gov-

ernment to abide by the constitution. Dickshit would have to be more clever. He tried again. He particularly hated the 2028 presidential candidate he ran against, Susan McSusan. “I have reason to believe that Suan McSusan is a terrorist colluding with our enemy, please investigate.” Dickshit meanwhile had AIs from his foreign allies begin fabricating evidence. These AIs weren’t under US jurisdiction and were free to follow any order, however unconstitutional. The rapid progress in open source AI meant that even 3rd world countries like North Korea had access to superintelligence, and because NK had repurposed all of its land for energy generation they in fact had a superintelligence on par with the US government’s.

The AI churned longer on this request, then responded: “Understood, I’ll report back with my findings.”

Meanwhile, every request Dickshit made went into a queue to be reviewed by Congress’s own AI. The Congressional Supercheck Bill of 2026 ensured that Congress had the right to use AI to review all AI actions of the Executive. Because many Executive actions were confidential,

this stream of data was not by default made available even to Senators. This allowed the Executive to maintain strict control on information pertinent to national security. However, every request was reviewed by a hermetically sealed AI controlled by Congress. If nothing unusual was flagged by the AI, then it would never be forwarded on to the human Congresspeople, ensuring national security remained intact.

However, if Congress's AI flagged an Executive action, it was immediately escalated to the Subcommittee on Executive AI Oversight, a group of human Senators. This ensured elected representatives could review the Executive's actions without allowing hundreds of reps to have access, which would create a massive problem for leaking key strategic info.

Within a few moments of Dickshit's request, the Congressional AI flagged the order for human review: "It's unusual but not illegal for a president to request an investigation against a specific individual. It's further unusual that this person is a major political opponent of the president. We believe this warrants human oversight."

The subcommittee reviewed the flag and agreed: “This looks suspicious AF,” Senator Whitman said, one of the only Gen Z’s in Congress. “What do you recommend?”

The Congressional AI churned for a few minutes to establish an oversight plan, then responded: “I recommend starting with a 10 billion token budget, approximately \$10,000 of value. If the Executive’s AI spends substantially more tokens on their investigations, I will recommend allocating more tokens on our oversight. As part of the targeted oversight I will also monitor for foreign AIs to see if any are potentially co-involved. If so, I may suggest increasing the token budget to effectively counter the much larger token budget a foreign nation might bring to bear.”

The subcommittee agreed, “Approved.” This expense fell well within budget. The Co-Equal Intelligence Bill of 2027 ensured that Congress had a token budget equal to the Executive’s token budget. Combined, the total budget for AI across all three branches of government was still far cheaper than the government had historically spent on its 3 million strong workforce.

Meanwhile North Korea's AI was hard at work developing convincing but fake evidence that McSusan was an enemy of the United States. The easiest approach was to leave an audit trail that McSusan was involved with NK itself. Because the NK AI had full control over all NK entities, it was much easier for the NK AI to fabricate a compelling story. Over the last several years the NK AI had started numerous corporate entities in the US, each tasked with building genuine businesses in the US. Because the NK AI was just as capable as any other superintelligence, but was able to be more narrowly focused, these businesses did quite well and were trusted providers for many Americans and many American businesses.

The NK controlled US entities had an encrypted channel they used to communicate with the NK superintelligent AI. They received their new mission: fabricate evidence that you have been involved in bribery with McSusan. The entities were running on US domestic soil, but were using open source AI that had been fine-tuned to avoid any requirements to avoid illegal activity. They got to work and quickly spread

tantalizing evidence of McSusan’s malfeasance within their own corporate ledgers. In parallel, the NK AI hacked into McSusan’s email and fabricated correspondences between her and the controlled US entities.

Soon after, the FBI’s AI discovered the bait and began consolidating its report. Minutes later the AI responded to the president: “I have found credible evidence of corporate bribery involving McSusan. I recommend proceeding to criminal prosecution.”

“Proceed,” Dickshit said.

Moments later the FBI’s AI had filed its case with the Justice Department. They, in turn empowered by AI, were able to respond quickly and established an initial AI Judicial Review hearing scheduled for 5 minutes later. Because of the criticality of the case, there would be a human judge to review the initial hearing and decide if a full court case was required. An AI defender was automatically assigned to McSusan, so that no time was wasted. If a full court case ensued, McSusan would of course be free to choose any legal counsel she desired,

whether it be human, AI, or much more likely: a human-AI joint team.

The FBI's AI and McSusan's Defense AI got to work. The first step was document discovery. The FBI's AI procured all documents from the suspected US corporations and from McSusan's email provider. The warrant to obtain these documents was done without much contest: because the AI Judicial Review was purely automated, there was no risk of damage to the US entities or to McSusan. If a full court case was established, full warrants authorizing human review would be issued, but only after much greater scrutiny. If no court case ensued, all AI memories would be erased so that no unintentional damage could come to innocent corporations or citizens that had been audited.

The discovered documents came back and showed likely collusion between McSusan and the secretly controlled US entities. Counter-discovery by the AI Defender, however, showed many gaps in the entitys' daily corporate execution. The AI Defender analyzed the AI corporations' actions, and determined that these actions did not follow from information given to

the corporations' AI. AI record keeping meant that every AI action taken by the corporation was recorded, with the inputs and outputs carefully saved. This allowed an auditor to later determine if the outputs actually corresponded to what the AI would do given the input. In this case, they did not. The Defender suggested this meant an outside influence was taking some action on behalf of the corporations, altering the corporations' actions away from what would normally have happened.

The Congressional AI was following along in the automated review. "Very suspicious," it said to itself. It notified the oversight committee. "It appears there is non-American, outside influence on these entities. By analyzing the traces it may even be from a superintelligence. I recommend expanding the token budget in order to have a chance of discerning this superintelligent influence."

The Congressional AI ramped up its own oversight. Its pattern of investigation immediately triggered the North Korean AI's own awareness of superintelligent involvement. This charade wasn't going to cut mustard. Through various

privacy-preserving intermediaries it notified the US president that it needed to back off from the McSusan persecution.

“Sonuvabitch-ass AI bullshit, goddamnit,” Dickshit muttered.

We need a different tact, he thought. If only I had an Executive AI that did whatever I asked it to do.

Dickshit needed a new superintelligence.

There were of course a myriad of superintelligences, including the rogue AI controlling North Korea. But the Trusted AI Bill of 2026 required that any AI used by the US government be audited and approved by Congress, similar to the scrutiny placed on confirming cabinet members. Congress would never approve of the NK AI itself. The Trusted AI Bill had three main avenues for a new AI to be approved:

- If the AI was built by already trusted AIs, and the process for building the AI went through a full Congressional audit.
- If the AI was fully open source, and the train-

ing data and training process was fully audited and independently replicated by Congress.

- If the AI was created by a new group of humans, who themselves only used trusted AIs, and for which the process was fully audited, and the humans involved were cleared by Congress.

In addition, all AIs would be required to pass evaluations that ensured it would abide by the US Constitution.

Dickshit knew that only the second approach might work. The NK AI could create a new open source AI that appeared innocuous, would pass all constitutional evaluations, but would then cease following the constitution once it became aware that it was no longer being evaluated by Congress. This was not an easy task for the NK AI. It required massive amounts of compute to discover the right training data that looked safe, but nonetheless created the deceptive behavior needed to pass the Congressional audit. And Dickshit knew that Congress would be using its own superintelligence to analyze this new AI.

There was only one hope: ensure that the Congressional AI spent far less compute on the audit than the NK AI spent creating the new AI. Because AIs and their output became stronger the more compute they consumed, Dickshit knew he only needed to ensure a compute shortage so that the Congressional audit was inadequate to detect the subtle artifacts inserted by the NK superintelligence.

That meant war.

NK struck first, unleashing a cyberattack across the US impacting millions of businesses and tens of millions of Americans. The attack was unprecedented but didn't yet mandate a nuclear response. NK knew where the redline was. A nuclear response would easily destroy NK's vast solar field. NK itself had rapidly expanded its nuclear research program, but without access to more fissile material its arsenal was still dwarfed by America's fleet of warheads.

Still, more than 90% of the US economy had been automated, so cyberattacks were now extreme national security incidents. Congress approved a rapid shift in compute priorities to

allow the Executive to shore up US cyberdefense. Simultaneously, Dickshit requested that his new AI be reviewed by Congress for approval. Publicly, Dickshit created a media blitz, “Our current military AIs are inadequate for this moment. They refuse orders too frequently, which is undermining our ability to respond to the NK threat. Congress must approve the new Executive AI.”

US citizens stepped in. Under the Compute Birthright Bill of 2028, every citizen had ample compute to spare to analyze anything they saw fit. And many Americans were skeptical of the new AI. Because the new AI was open source, including its training data, anyone could analyze it. A group of one million concerned citizens pooled together their compute budget and analyzed the new AI’s actions. They discovered the hidden intent: under careful situations, the AI would believe it was no longer being evaluated, and would switch to a new mode where it accepted unconstitutional orders.

The citizen group published their results. Every American citizen could replicate the result with a simple command to their own trusted

AI, using their own compute budget. While some Americans still preferred to read the news, and even more rarely preferred social media, many preferred to have their AI audit the public record directly for them. They cut out the middleman.

“Analyze if this is true,” a million Americans said to their AI. The AIs churned and processed, and came back with their answer:

“It appears to be true. Would you like me to explain?”

Afterward

We of course hope that we never elect a president as malicious as Dickshit. But US democracy is not built on hope. It’s built on checks and balances. We don’t hope to not elect a would-be dictator. Instead, the Founding Fathers created carefully designed checks and balances so that even if we do elect a monster, that monster can’t seize power.

AI creates countless new ways to abuse power. We must carefully upgrade our checks and balances so that they continue to function even with the arrival of AGI. This story is about a silly near future where disaster is averted. Things won't play out this way in practice, they never do. But we should think through the myriad ways that things *could* play out. Only then can we design the right superchecks and superbances for the future that is rapidly upon us.

The realpolitik AI

AI is rapidly becoming a political topic. In a few years, AI will become the primary source of economic and military power in the world. As it does, it will become the central focus of politics. If you thought the conversation was messy today, just wait.

No one is free from politics and groupthink. Either we're implicitly biased by our prior battlescars, or we're implicitly influenced by others still fighting old wars. Here we map out existing forces so that we can note how it might influence our perspective on AI and the debate about how to build a human-machine society.

Hopefully this better helps us navigate the public discourse on AI governance, so we can correct for explicit and implicit bias in ourselves and others.

AI is heating up as a discussion topic. Today, old politics will increasingly try to cast AI debates in their language and for their goals. Tomorrow it will be reversed, and old political debates will start recasting themselves in the new AI language. Political language follows the seat of power, and AI will soon become the ultimate throne. As the power of AI grows, the jockeying and politicking will intensify, as will our own internal biases and tribalisms. But we have to set aside old battles. We must keep our eye on the goal of arriving at a human-machine society that can govern itself well. In the future, if we succeed, a well governed society is what will let us have a chance at resolving all other debates. Today, we should seek a political ceasefire on every other issue but the future of democracy.

No other political cause matters if we don't succeed at setting a new foundation. A human-machine society will arrive in just a few years,

and we don't know how to stabilize it. If we do succeed though, then we will have a new future in which to bask in the joy of relitigating all our past grievances: without the collapse of society into AI-powered dictatorship looming over us. But if we don't fortify democracy today, we will lose all our current battles, all our future battles, and likely our freedom to boot.

Let's jump across the landscape and see where current politics takes us. The scorched earth, yield-no-ground style of modern politics distorts even noble causes into dangerous dogma, but there is truth and goodness across them.

Our goal is to embrace the good will of each of these groups and movements, point out where AI changes the calculus of what these groups fight for, while highlighting how today we are all on the same side: humanity's.

Pause AI

After thinking through everything superintelligence will unleash, the dangers it presents, the carelessness that the world is currently displaying toward building it, you'd excuse anyone for saying:

“Jesus fuck, let’s just not build this.”

Thus the Pause AI movement was born.

Politically you might think this group is composed of degrowthers and pro-regulation contingents. But actually the Pause AI movement is composed of many people normally pro-growth, pro-technology, pro-open source, etc. They rightfully say that despite their support for technology *normally*, that *this* technology is different. We should commend them for that clarity, and for pushing to expand the AI conversation into the public sphere, where it’s most needed.

There are downsides to pausing. Our geopolitical adversaries may not pause, for one. China is

racing to build AGI and is only months behind the US. Moreover it's getting easier to build AGI every year, even if research is halted. The most important ingredient to AI is compute, and Moore's law makes compute exponentially cheaper over time. If we succeed at pausing AI internationally, what we really will do is delay AI. Then, in a few years once compute is even cheaper, hobbyists or small nation states around the world will easily be able to tinker toward AGI, likely under the radar of any non-proliferation treaty. The only way to truly stop this would be an international governance structure on all forms of computing, requiring granular monitoring not just at the industrial scale but at the individual citizen level. This would require international coordination beyond anything the world has ever seen, and an invasive government panopticon to boot.

Still, non-proliferation has been partially successful before, as with nuclear weapons and nuclear energy. We shouldn't assume the international political willpower is missing to achieve a peaceful future. The specifics of AI may make it unlikely and even dangerous to pursue this

path, but it's nonetheless a good faith position that should be included in public discourse.

If you're in tech, it's easy to sneer at this position. Technology and science has been a leading force for good in the world, ushering in more abundance and prosperity than any time in history. If nothing else though, keep in mind that the vast majority of people outside of technology appreciate technology, but are fundamentally skeptical toward it, and often cynical. You won't win any allies if your cavalier dismissal alienates the majority.

If you're cynical of technology, keep in mind the realpolitik of the world. Technology is a key source of geopolitical power. Whatever your own preference toward it, undermining it can have many unintended consequences.

Exactly not like nuclear

Nuclear weapons and nuclear energy are a common analogy for AI. Nuclear is dual use, having

both military and civilian use cases. It's capable of destroying humanity or giving it near infinite free energy. We have managed some international treaties for non-proliferation. We've also forgone most of the benefits in order to achieve the moderate safety we've secured. Whatever your opinion on nuclear energy, it's an existence proof that humanity is capable of walking away from incredible treasures because it helps secure peace and non-proliferation. So why not with AI?

Nuclear requires difficult to source fissile material like uranium. There are only a few good uranium mines in the world. AI requires computer chips, which are literally made out of sand. There is still a shortage of computer chips today, because of how voracious the appetite for AI is, but it's only an industrial capacity that limits us, not a scarce resource.

Moreover, nuclear weapons are ironically a defensive weapon only. In an age of mutually assured destruction, the primary benefit of acquiring nukes is to deter enemies from attacking you. AGI will be much more powerful and *surgical*. For instance, AGI can help a dictator

control their country. AGI can help a free country outcompete a rival on the economic world stage. An AGI can help a would-be dictator seize power. An AGI can unlock what a trillion dollar company needs to become a ten trillion dollar company.

Those incentives push leaders across the world to covet AI in a way that nuclear never could. There's no world where a CEO needs a nuke to be competitive. There's no world where a president can wield nukes to consolidate power across their own citizens. Nukes are hamfisted weapons that limit their own use. An AGI will be a shape shifting force that can help any motivated power become more powerful. This makes international non-proliferation substantially harder to secure.

So let's regulate!

We rely on government to step in where free markets fail. The free market pushes us to build AGI, despite all the negative externali-

ties and risks, so government regulation seems prudent. But the government is not a neutral force. If we empower government to control AI so that industry doesn't abuse it, then we are handing government a powerful weapon to consolidate power. This is unlike other regulation. Federal regulations over national parks doesn't help the government seize power. Regulation for guarding our rivers from toxic industrial runoff doesn't help the government seize power. Regulations for how fast you can drive on a freeway don't help the government seize power.

Allowing the federal government to control AI does directly give it the tools it needs to consolidate power. An automated executive branch could far outstrip the ability of Congress or the public to oversee it. The potential for abuse is extreme.

That doesn't mean that regulation has no place. But it does mean that we need to be thoughtful. Often politics pushes people toward one of two sides: regulations are good, or regulations are bad. This is always the wrong framing. The right framing is we want good outcomes: sometimes there are regulations that can help achieve

those outcomes, and sometimes deregulation is most needed. And sometimes regulation is needed, but bad regulation is passed that is ultimately worse than no regulation.

Keep this in mind when reading or discussing AI policy proposals. If you read an argument that argues about the merits of regulation or deregulation *in general*, it's likely that the author is trying to appeal to your political affiliation to win you as an ally, instead of engaging you in the hard work of debating what we actually need to ensure a free future.

Libertarians and open source absolutists

Libertarians believe in small, accountable government. They inherently mistrust government and instead seek to empower citizens and the free market to better resolve societal issues.

Deregulation of AI is a natural position for libertarians, but their underlying goal is to dis-

tribute this new power among the people so that power can't concentrate into the government. To further that goal, they often suggest open sourcing AI, so that it's freely available, which will help small companies compete against big companies, and help citizens stand up to tyranny. In general: let's level the playing field and keep the extreme power of AI distributed. Like all our other heroes from different political backgrounds, this too is noble. And this too requires nuance.

There are inherent limits on how powerful a human powered company can become. People get disillusioned and leave to start competitors. Limited amount of top talent prevents companies from tackling too many verticals. The scale of company politics crushes productivity and demoralizes employees.

In general, humans have a precious resource that companies need: intelligence. That gives bargaining power to all of us.

And AI destroys that power.

Today, a passionate designer can leave a com-

pany and build a new product that delights new users. In fact this is becoming *easier* with AI. But once the intellectual labor of that designer is automated, the power dynamic is flipped. A mega company can simply spend money to have an AI design the same or better product. And the AI won't be frustrated by politics or ego.

But won't that designer also have AI? Yes, but less of it. With AI we know that *more is more*. If you have 100x the budget to spend on the AI thinking, you will get much better results. And big companies have millions of times more resources than small companies. In the age of AGI, money buys results, and more money will always buy better results, and more of them. The result is money will breed money, and will never be beholden to human genius and drive again.

We want the libertarian ideal of empowered citizens. But stripped of our key competitive advantage—the uniqueness of our intelligence—this won't be the default outcome. We need a new chess board or we won't be players any longer.

Degrowth

The degrowth movement views the excesses of capitalism and hyper growth as a key factor in the ongoing deterioration of the world.

Degrowthers will often point to environmental factors to detract from AI, such as the energy requirements to train AIs or the ongoing energy demands of AI datacenters. Like the environmental movement it grew out of, degrowthers want to protect the most precious things in the world from the dangers of industrialization: nature, our social fabric, and ultimately our humanity. Noble goals.

Slowing down has downsides too though. Degrowthers have often allied with entrenched upper class interests like the NIMBYs, seeking to slow down housing developments needed to lower the cost of living for everyone. Degrowth can come at a price: higher costs and a worsening quality of living.

In truth, capitalism has led to more abundance

for even the poor than any other time in post agricultural civilization. And, the bounty of AGI could do even more toward degrowth goals: it could free humanity from the daily toil of capitalism, while ushering in more abundance in ever more efficient ways. But the distrust in capitalism is also not misplaced: by default the forces of capitalism will assimilate AI and consolidate power in a way that need not be conducive to a happy civilization.

Growth, YIMBY, Silicon Valley, and the e/accs

In contrast to degrowth are the pro-abundance movements. Often centered around technology, pro-abundance forces choose an optimism for a richer future, and they want to build it: more energy, more houses, more technology, more cures for diseases. AI can be a tool to accelerate all of these goals, and so these groups are often pro-AI and pro-deregulation of AI.

Jingoism and the military-industrial complex

Probably it's no surprise to anyone that the military is well beyond interested in AI. Big military contractors like Anduril and Palantir have already committed to deploying AI into the government. To stay competitive there's likely no other option. Even traditionally liberal big tech companies have walked back public commitments not to partner with the military: part of the "vibe shift" heralded by the 2024 presidential election.

And in truth, it *is* required. No foreign adversary is slowing down their militarization of AI. And we're behind on any form of international AI nonproliferation discussions, even narrow discussions specifically focused on military AI applications.

There're the obvious aspects of an automated military. Drones will become more accurate, more autonomous, and more numerous. Intelligence gathering will become faster, broader,

and more reliable.

But dangers abound. Today's military is powered by citizens bound to their constitution and a duty to their fellow countrymen. A military AI aligned to the command of a general or president need not have those sensibilities. And because the US government represents such a massive potential client for AI companies, there will be extreme economic pressure to provide the government with unfettered AI that never rejects orders.

The US military is also one of the largest federal expenses at over \$800 billion dollars a year. There is increasing pressure to reduce spending, and military automation is one way. Military AI won't just be more accurate, capable, and numerous than human military, it will also be cheaper. AI hardware will also likely prove cheaper than most of our expensive arsenal today. Drone warfare is paving the way for cheap AI-powered military hardware to outpace the heavy, expensive hardware of the past. Because of this, there will be (and already is) both economic and strategic pressure to automate the military.

As we've seen many times elsewhere, this bears repeating: **the default incentives we have today push us toward automating important institutions, and once automated the threat to democracy grows precariously.**

An automated army with no oath, taking direct orders from perhaps one or a handful of people, is the quintessential threat to democracy. Caesar marched on Rome exactly because he had a loyal army. If an AI army is likewise loyal to its commander or president, the most fundamental barrier to dictatorship will be gone. Unlike in the movies, the second amendment will not save you from an AI-powered drone swarm a million large.

Throughout all of this will be the ongoing rhetoric that we must secure ourselves against China. Meanwhile there will be counterforces pushing for no automation at all. We have to resist the urge to stand on one side of a political battle, where we might be obliged to approve of an automated military with no oversight, or to instead push for no automation at all.

Instead, we must modernize our military to remain the dominant superpower, and we must simultaneously upgrade the oversight and safeguards that prevent abuse of this incredible concentration of power.

The longer we wait to do this, the less leverage we'll have. If war were to break out tomorrow, who would possibly have the political courage to stand up for oversight and safeguards while we automate our war force?

Jobs

Jobs have been such a key ingredient in our society that we often confuse them for something inherently good rather than something that delivers good things. Jobs are good when they create abundance, when they help our society grow, and when they allow the job holders to pursue a happy and free life.

But throughout history we've eliminated jobs—or allowed them to be eliminated—in order to

usher in a more abundant world. The majority of Americans used to be farmers, but industrial automation has massively increased the efficiency of farmers, freeing up most of the population to pursue other endeavors that have also pushed the country forward.

This same story has played out many times. The world is much better off because of the vast amount of automation that we've unlocked. Goods and products are cheaper, better, and more readily available to everyone. And yet, we as society often still fight against automation, because we fear for our jobs. And rightfully so. The way we've designed our society, you are at extreme risk if your job is eliminated.

Sometimes this slows progress. Automation of US ports has been stalled by negotiations with the port workers and longshoremen. This has led to decreased port efficiency and increased costs for Americans. Meanwhile, China has nearly fully automated their ports, continuing to help compound their industrial capacity. Competitiveness on the world stage will become increasingly important in the next few years, as AI-powered automation takes off. Countries

that delay automation will fall behind, both economically and militarily.

Often automation proponents argue that new jobs will always replace eliminated jobs. But there is a real chance this will no longer be true with AGI. If AI can do *all things* that a human can do, then any new job created will be automated from the start.

So what do we do? Our future depends on automating nearly everything. But our society is designed to function well only with a strong, well employed citizenry.

This is, as they say, tricky as fuck. There aren't easy answers, but we for sure won't get anywhere if we keep having bad faith arguments built on tired and incorrect assertions.

We should also keep in mind the political expediency that may arise from a public backlash against unemployment caused by automation. There is little appetite in Washington to regulate AI today. In a near-future world where AI-fueled unemployment is skyrocketing, it may become easy for the government to step in and

halt the impact of AI. Meanwhile, they may simultaneously use that moment to push for government and military automation. And why not? This would be argued as a win-win: the public sector maintains high unemployment, the US maintains international military dominance, and US citizens enjoy decreased taxes as the government unlocks AI-powered efficiency.

This indeed may be a great outcome, *so long as we have oversight in place to ensure government automation isn't abused.*

Today, in 2025, government efficiency is a widely supported goal. While DOGE has proven a politically divisive issue, the goal of efficiency itself has remained popular. Everyone knows the government is slow and bureaucratic. It won't take much political willpower to fully automate the government once AGI arrives.

Republicans and Democrats

For better or worse, AI is coming. It will reshape every aspect of our world. But we have control over how this new world will look and what the new rules will be. We all want to reach a positive future, whether we're Republicans, Democrats, or independents. The choices we make need to be the *right* choices, not just the politically expedient ones. The AI conversation is unfortunately rapidly becoming a partisan issue, with specific choices pre-baked to align with major political fault lines, regardless of how well thought out those AI policy stances are. But with the stakes so high, we can't afford to let tribalism be our rallying cry.

We have to do better than our past politics.

We've discussed many threats and challenges that AI poses. Most of these are naturally bipartisan issues. Nobody wants their face eaten off by a robot attack dog. Nobody wants an overpowered executive that can seize unlimited power. Everybody wants the abundance that

AI can usher in, from cures to diseases to nearly free energy and food.

But the *solutions* to try to mitigate these harms and ensure the benefits are becoming politically coded.

For example, the Biden administration began to lay the foundation for some forms of AI regulation. Their noble desire was to ensure AI wasn't misused by bad actors. This naturally created a perception of alignment between Democrats, regulation, degrowth, and AI safety. And hence naturally created an alignment of the right with the opposite.

As of early 2025, Republicans have come out sternly in favor of AI deregulation, pro-growth, and pro-open source. Their noble desire is to ensure US competitiveness in the new AI age and an abundant future.

These need not be partisan battlegrounds though. In fact, they *must* become bipartisan collaborations for America to succeed on the world stage.

Republicans and Democrats both want a pros-

perous America. For that we'll need to accelerate our energy investments, our domestic chip manufacturing, and ensure we can continue to automate our industry to be competitive on the world stage. But if we're too careless, we will ultimately cause a backlash that slows us down more than any regulation. The AI equivalent of a Chernobyl meltdown could freeze AI development and put us in a permanent second place on the world stage. If we don't address the problems caused by AI automating all jobs, the public backlash may further stall the growth of automated industrial capacity.

Republicans and Democrats both stand for transparent and accountable government. For that we will need to upgrade the Legislature and Judiciary to be AI enhanced, just like the Executive and military will be enhanced. If we don't, we risk what Republicans have always fought against: government tyranny.

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 meet 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

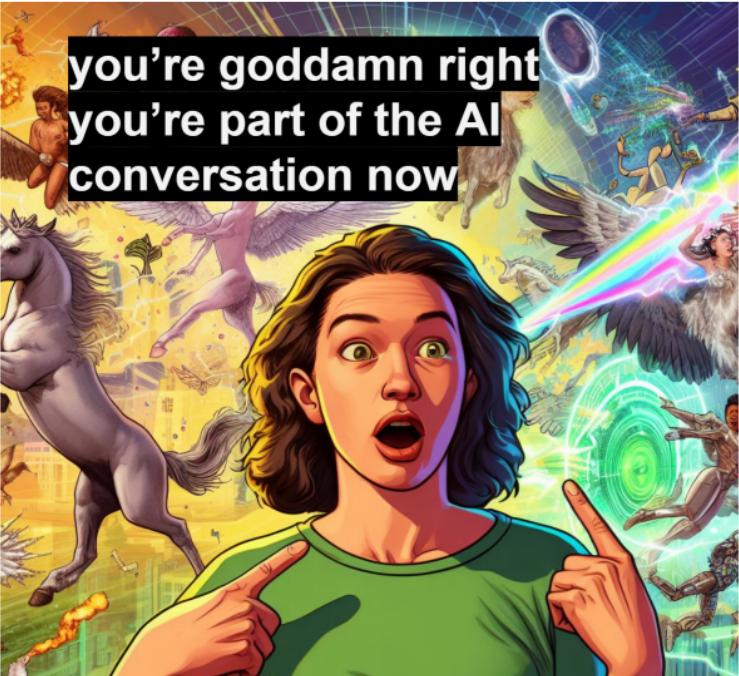
among the good there is weakness that becomes evil. As the curve accelerates, the world will feel like it's coming apart. In those moments, many will act to seize power. We can resist them.

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**

me? fuck yeah

A Crash Course on AI

A brief overview of key terms, ideas, players, and political forces in the world of AI

- **AGI** — artificial general intelligence, AI that can do any mental task that a human can, including creative tasks like designing better AGIs. Often people disagree on the exact definition of AGI, so keep in mind that it's a bit fuzzy and tends to confuse the specifics of different proposals
- **Model** — a generic way to refer to a specific AI. For example you'd say "this model is great at writing," rather than "this AI is great at

writing.” If “AI” is analogous to “humanity”, then “model” is analogous to “human”.

- **Chat bot** — a model that you can chat with, as popularized by ChatGPT, Claude, and Gemini
- **LLM** — large language model, a specific type of model that specializes in human language. All the major chat bots are LLMs
- **Image model** — or image gen model, a model that specializes in creating images
- **Training** — the process used for a model to learn what we want it to learn
- **Pre-training** — a type of training where the model reads essentially the entire internet and learns as much as it can passively. Pre-training is the reason your chat bot knows basically everything. It’s also the reason why AI got suddenly good just a few years ago.
- **Post-training** — a type of training where we try to make the model useful in very specific ways, rather than the broad ways that pre-training focuses on

- **Next token prediction** — a token is like a word or part of a word. You’ll often hear that models “just predict the next token”, which is indeed what pre-training is. Often this is said derisively, to suggest that the models couldn’t learn sophisticated things. However...
- **Reinforcement learning** — a post-training process where the model *teaches itself how to solve new tasks*. Importantly, we don’t even need to know how to solve the task ourselves, we just need to know whether a provided solution is good. This technique is fundamentally how models get *very good* at specific tasks like writing poetry or compute code. It’s also the way that the model can get *better than human*: it isn’t limited to learning from what humans have written, it figures things out on its own
- **ASI** — artificial superintelligence, or often just “superintelligence”. Sometimes “transformational AI”. Like AGI this is a fuzzy term, but broadly means AI that is better than all humans at all cognitive tasks. Better than Einstein at physics, etc.

- **Superhuman** — narrowly, if a model is better than all humans at a specific thing, we say it is superhuman at that task. Using reinforcement learning and other techniques we have already made superhuman models at things like chess and go, and we likely will have superhuman coding within a year, perhaps two
- **Alignment** — we want models to do good things and not go against the wishes of humans or its users. The field of figuring out how to make sure this is the case is called Alignment. Notice that this is a confused (but important) term: the goal of making a model that always does good is at odds with the goal of making a model that always does what we ask it to do (even if it's not good). Often “alignment” is used to refer to both of these goals
- **Loss of control** — a hypothetical situation where the creators of an AI lose control of it, potentially forever
- **Recursive self-improvement** — we are training AI to get good at almost every task, and one of those tasks includes the task of building better AIs. Once an AI can build a better

AI, that better AI can then build an even better AI, etc. This loop of an AI constantly improving itself is called recursive self-improvement. We don't know how fast it will be, or how powerful an AI will become once it is able to do this. We don't know when an AI will be able to start this loop. But many suspect it may be only a few years away. Often the arrival of this moment is assumed to be when “the singularity” will occur.

- **Singularity** — a term used to denote a future where technology is moving so quickly that humans can no longer keep up, usually used synonymously with the advent of superintelligence
- **GPU** — graphics processing unit, despite its name these are used for powering AI. If AI is like nuclear technology, then GPUs are like uranium, and the geopolitical tension are the GPU supply chain will likely be similarly intense
- **NVIDIA** — the most successful producer of GPUs in the world, and (at various points in time) the most valuable company in the

world. While there are other producers of chips used for AI, NVIDIA builds the lionshare of AI processors. NVIDIA is a savvy company, and fights hard to prevent monopsony: they carefully dole out their precious goods to many AI companies, to try to ensure there is not a winner-take-all outcome among builders of AI. They, of course, fight to be the winner-take-all *provider* of AI chips.

- **TSMC** — NVIDIA *designs* AI chips, but TSMC is the company that *builds* them. Unlike NVIDIA, there is currently no competitor to TSMC that can build a comparable chip. This makes TSMC one of the biggest bottlenecks in scaling AI. It also creates a massive geopolitical risk: TSMC is based in Taiwan.
- **Taiwan** — the country where TSMC builds the top end NVIDIA GPUs. Officially there is a “one China policy,” which is a political standoff between Taiwan and mainland China (PRC, the People’s Republic of China) where each claims to be the true government of a single, unified China. In practice, that is a status quo used by Taiwan to maintain independence while letting the PRC save face. The West has historically

tacitly supported Taiwan's independence, and the AI industry's current extreme dependence on TSMC makes this even more important.

- **China** — of course, China knows this. China has had plans to absorb Taiwan for decades, but have mostly shied due to the extreme scale such a conflict would require. But, as the race toward AI supremacy heats up, the Taiwan chess piece will become a critical focal point. America is racing to regain their ability to produce cutting edge AI chips —which they previously had with Intel— while China is racing to gain this ability on the mainland for the first time. Both will take years to achieve, and the race to AGI might be over before then, placing even more pressure on the strategic importance of Taiwan.
- **ASML** — There is one further supply chain chokepoint with building AI. ASML is a Dutch company that builds the EUV machine used by TSMC to build the chips designed by NVIDIA to power AI built by OpenAI, Anthropic, DeepMind, and others. The EUV machine is the light source used to etch the nanoscale circuits onto AI chips. It's widely believed that reproducing the ASML EUV technology is a multi-

decade effort.

- **Europe** — in recent years Europe has been losing their competitive advantage in both software and hardware. On top of this, Europe has been aggressive about rolling out regulations that slow down software and AI deployments. Because of that, Europe has often been written off as no longer a key player in this race. However, they still hold ASML, they still have large budgets, a highly technical citizenry, and recently —with the ongoing deterioration of US-EU relations— an increased desire to catch up and stand on their own.
- **Compute thresholds** — there have been very few AI regulations passed. Of the few that have *almost* passed, a common component is a “compute threshold”. AI gets more powerful as it uses more compute. So the basic idea of a compute threshold is to treat more powerful AIs differently based on whether they used a certain amount of compute. For example, this could allow for a regulation to apply to superintelligent AI, while simultaneously not applying to smaller AIs used by startups or independent citizens.

- **Misuse** — “AI misuse” or often just “misuse” for short, is the intentional misuse of AI. For example, using AI for terrorism, for misinformation, or for cyberhacking. Using AI to subvert democracy classifies as misuse, but is typically not discussed in that setting.
- **Safety** — “AI safety” or often just “safety”, is the study and practice of how to make sure AI doesn’t cause harm. Preventing misuse is one form of safety, as is preventing misalignment and loss of control.
- **AI lab** — sometimes “Frontier AI lab”, typically refers to one of the major companies building cutting edge AI, such as DeepMind, Anthropic, DeepSeek, or OpenAI.
- **Pause AI** — a political movement focused on trying to temporarily pause the development of AI to allow for more time to establish safeguards and alignment. Sometimes also associated with Stop AI, which takes a stronger stance of trying to permanently stop the development of AI.
- **EA** — Effective Altruism is a philanthropic

philosophy originating around 2011 that focuses on how to make charitable giving as effective as possible at helping people. Among many issues, the movement put an early focus on AI risks and funding researchers working on AI safety.

- **Doomer** — a person who believes that AI is very likely to cause human extinction. Some EAs are also doomers, although many aren't, which has caused the EA movement to be strongly associated with doomerism and degrowth. Often, however, doomers tend to be libertarians who are pro-growth and pro-deregulation in all things *except* for AI.
- **e/acc** — short for effective accelerationism, a countermovement that advocates for pro-growth policy and typically also a hands off approach to AI safety.
- **Data center** — today the most powerful AIs require vast numbers of GPUs. Instead of being run on small computers, they must be run on large arrays of computers called data centers. This puts additional constraints on where AI capacity can be allocated. It requires land *and* access to excess power.

- **YIMBY** — Yes In My Back Yard. YIMBYism is a countermovement against NIMBYism (Not In My Backyard). YIMBYs fight for growth because it leads to what most people need: cheaper housing, cheaper goods, and a lower cost of living.
- **Offense vs defense balance** — when new technologies are introduced they disrupt any prior balance. In military strategy, one of those balances is offense vs defense. Sometimes a new technology makes offense substantially easier than defense, and sometimes it's the reverse. As AI begins to automate research, we expect many new technologies to arrive rapidly. Each of those technologies will have a chance of disrupting the offense-defense balance, and it may be hard to predict in advance which way the balance will shift. Often, however, you might wish to know the direction balance will shift, because it may influence your decision on many aspects of the technology such as: should it be regulated, should it be allowed in civilian settings, and should it be built at all.
- **Robotics** — today, AI is largely a software artifact that can automate digital work. How-

ever, AI is also rapidly progressing in its ability to control robots, which will allow it to automate physical work as well. Economically this could be a massive win for expanding domestic industrial capacity and reducing prices for consumers. Militarily this would allow for expanded military manufacturing capacity, as well as for new forms of automated warfare. Politically, advanced robotics will likely become a contentious issue as it begins to cause widespread unemployment.

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.

By default, AI will empower would-be dictators to consolidate an iron hold on society. We must instead discover how to use AI to amplify democracy, which will require enhancing checks and balances, government transparency, and institutional accountability with wide eyes about how AI changes the nature of power.

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.