Liberty by Design

Governing a Human-Machine Society

Liberty by Design

Governing a Human-Machine Society

Jordan Ezra Fisher

May 4, 2025

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

Table of Contents

AI will soon disrupt every foundation our society is built upon. If we want to control our future, there are many questions to answer. In this work we'll focus on how our human governance needs to adjust to a world fully powered by machines. But there are many other important questions we won't address. Two of

those questions are particularly important:

Can we determine how to maintain control over AIs that are smarter than us? And can we ensure companies and nations build AI responsibly, so that only safe AIs are built?

The first question is studied in the field of AI Alignment:

Today, we don't know how AIs work. Many prominent leaders building the most advanced AIs think we will have AI greater than all of humanity combined by the end of the decade. We don't yet know how we'll control AI once it exceeds us, and time is short.

The second question is explored in the field of AI Policy:

In order to be first, a company building AI may cut corners and take on more risk than the rest of humanity might like. AI Policy seeks to set the regulations and incentives such that we prevent these types of decisions. And the race doesn't just exist between AI companies, it also exists between nations. The US and China are already racing to be first to create powerful AI. The future balance of geopoliti-

cal power may lie in the outcome. AI Policy seeks to ensure both that the race is won by democratic countries, but also to defuse the race and allow for international collaboration toward safe outcomes.

How do we solve AI Alignment? What AI Policies will push us to build aligned AIs? These are two incredibly important, unanswered questions. If you're interested in these, I encourage you to read and engage in these discussions. We urgently need more people helping with both.

But in this work we'll focus on a final, third question: the question of governance of a human-machine society. While AI Policy focuses on what governance might lead to the safe creation of AI, we will focus on what governance we need *after*: how do we govern a world that is built on and powered exclusively by machines?

Imagine we succeed at building powerful AI. Imagine further that we succeed at ensuring it's safe and aligned to its users. How then do we want a human-machine society to look? How *should* it look, to protect our freedoms

and liberties?

Corporations will adopt AI to automate themselves and enhance their efficiency. Governments will do the same. In fact, this is already happening. Once this process is complete, how will we exert influence on these institutions when we are no longer the force powering them? In what ways can we upgrade these institutions so that they continue to represent us, even once we aren't embedded in them?

We'll explore this core question of governance, the extent of what's at stake, and how we might fortify democracy for the human-machine age. And we'll analyze and discuss explicit enhancements to our checks and balances that we hope will offset the loss of implicit guardrails that currently keep democracy safe.

Our discussion will cover the following:

Chapter 1: Why AI is accelerating, and why we have little time left

- Machines have now reached human intelligence in many domains, and will soon surpass us in most others.
- If we understand why this progress is accelerating, we will be able to better design how we want to integrate machines into human society.

Chapter 2: What do we want from our machines and 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.

Chapter 3: The end of implicit guardrails

- Automating our economy and government will completely change all of those details.
- As those details change, all of the implicit guardrails we rely upon for checks and balances on power will be swept away.
- For example, institutions often avoid illegal actions because their human employees refuse to accept grossly unethical work. Or, even if employees comply with illegal requests, human whistleblowers within the organization can alert the public to any abuses. Once an institution is fully automated, these restraints won't exist.
- If we consider the military, the human component is even more important. Most sol-

diers would never fire on civilians, even if faced with a direct order to do so. This limits the ability of a commander to commit atrocities or to use their troops to enact a coup. With a fully automated armed forces, a motivated commander could decide to rest power for themselves.

- Today, companies, governments, and militaries all require human labor to continue functioning. Ultimate power rests with humans: if they choose to leave an institution, that institution will fail.

- Once our institutions are automated, power may instead sit solely with the leaders of those institutions.

Chapter 4: A simple path to tyranny

- Gradually, and then all at once, we will enter a world where implicit checks on power are impotent. Will our remaining explicit checks on power be sufficient guardrails?

- We'll argue that automation will make the task of seizing power substantially easier and more rewarding.
- We don't know if our AIs will be aligned, but we already know that many human leaders are not. History is replete with leaders who seize and abuse power.
- If we don't change course, human-powered tyranny may be the default outcome of a machine-powered world. And it will be increasingly hard for us to resist this tyranny 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. As it stands, whoever controls AI will control nearly everything.

Chapter 5: The Prompt of Power

- The forces and incentives are already at play to further concentrate power.
- We can't predict how all those forces will play out, but we can think through different scenarios.
- Here we'll explore a short story walking down our current path, and how it leads to an unrecoverable concentration of power and the end of liberty.
- The future will almost surely play out differently, in one of a million possible paths. We need to set the conditions so that the future is bright regardless of the path.

Chapter 6: Rapid fire governance — designing upgrades to democracy

- So what should we do? What should those conditions be?
- How can we upgrade our society and governance to be resilient to the multitude of forces pushing us toward tyranny?
- 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.

Chapter 7: Superchecks and superbalances

- We should think through and imagine how things may go wrong, to better design a more resilient system.
- But we should also imagine how things might go right, to ensure we're building a future we want to live in.
- Here we'll illustrate a positive, near future by telling a short story of how things might go well, assuming we upgrade our checks and balances.
- A positive future will surely play out differently than we expect, even if things go well. But we must set a vision of what good looks like so we know what we're fighting for.
- Moreover, we should plan for the worst. We should assume that one day we'll elect a would-be tyrant. The governance we design today should be so robust that even then our democracy would stand.

Chapter 8: The realpolitik AI — forging a new political alliance

- The discussion around AI policy has rapidly become politically coded.
- Adopting all the policies of the left, or all of the policies of the right, will likely lead to disaster.
- If we only regulate and slow down AI, we will cede the race to China.
- If we only automate our military and the executive branch, without also upgrading our checks and balances, we will hand so much power over to our leaders that we may never be free again.
- Instead, we must modernize our government and military to remain the dominant superpower, and we must simultaneously upgrade the oversight and safeguards that prevent abuse of this incredible concentration of power.
- And while we must treat the race against

China as existential, we must also look constantly for offramps toward deescalation and international peace.

Chapter 9: An exponential, if you can keep it

- Intelligence is the most transformative power the world has ever seen. Until today, that power has been human alone. Now, with AI, we are on the precipice of unleashing that power a thousand fold, and it won't be human.
- The force of multiplied intelligence completely rewrites the rules of our world. With it we may see near infinite abundance, or total ruin.
- We are on the exponential now. Where will it take us? We must decide together.
- AI is the defining event of 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.

But, if we begin this great debate today, we can set the framework for a positive future.

Chapter 1

Why AI is accelerating, and why we have little time left

If you're reading this in 2025, maybe you're already noticing AI around you. The news articles. Your colleagues using AI for work. Your kid using it as a tutor to learn math faster.

At my last checkup with my doctor, while chitchatting about AI, he proudly proclaimed that he doesn't use any chatbots. What was interesting was that he thought this was notable. The *default* is that you use chatbots, and he felt it was noteworthy that he didn't.

Everywhere else, everyone I know follows the default. A year ago I knew more holdouts, today they're mostly gone. The adoption curve for AI has been phenomenally fast.

But, fine, you've seen new technology before. If you were born before 1990, you saw the heyday of Moore's law, the rise of the internet, the advent of smartphones, and the transformation of nearly every type of social interaction through social media: from dating to shaming, from politics to condolences. You've seen all these things come on fast, and then get so integrated into society they're almost forgotten about. Not worth discussing.

Isn't AI just another new technology? Is there really so much more progress in front of us that society is in danger? That our lives literally are in danger?

Yes.

And the future depends on understanding this. There is so little time left that if you wait for a clearer signal, the moment to make a difference will be gone. Moreover, the way we choose to intervene and try to guide society

needs to change with the realities of how this technology will mature. All the details matter.

Let's work through some of the details to better understand why AI is accelerating. Those details will help inform how we predict the future will unfold, and what changes we'll need to ensure that future is positive.

The horizon of an agent

- AIs know more than any human alive, by many orders of magnitude. In terms of pure question answering ability, they're now outpacing even most professionals. I hold a PhD in computational fluid dynamics, and I can't hold my own against AI even in this narrow domain I spent years mastering. If you have a question about fluid dynamics, today you're better off asking an AI rather than me.
- But AIs still aren't as good as humans at *doing things*. We call this "agency", and

AIs that perform actions we call “agents”.

- Why do AIs seem so smart, but are still so bad at doing things? With humans we’d call this gap tacit knowledge. You can read every book in the world on how to build a car engine, but you won’t really know how to build one until you pick up a wrench and do it many times.

- Tacit knowledge isn’t written down. It’s not on the internet. You have to discover it yourself by doing.

- Only now, in the last year or so, have AIs started *doing*. Now that they are, they are rapidly improving at it.

- As the AIs improve at doing things, we look at which tasks they’re good at, and which they’re still failing at.

- In general, what we see is that today the AIs are better at tasks that take less time. This is more or less the same as the developmental progress of a student, or of a new employee. First you need to break down tasks into small chunks for the student, but over time you can give them bigger and

bigger tasks.

- We call the length of a task that an AI can handle their "horizon". As AIs improve, so does their horizon. The capability of an AI is now best measured in units of time. How long is their horizon? How big of a task can I give it?

The quick glance rule

- The shortest horizon is a task that can be done immediately, intuitively, or at a quick glance.
- Look at a picture of a cat, and you know quickly that it's a cat, without even consciously thinking.
- 10 years ago, these were the hard tasks we were training AI to be good at. Identifying cats wasn't easy, but you could do it with hard work.

- To train an AI, you would collect millions of pictures of cats and not-cats, and hand label them. Then you would teach the AI with these costly labels, a process known as supervised learning.

- At the time, Andrew Ng, a famous AI researcher, popularized the idea of "at a glance" tasks. If a human can do something at a glance, then so can an AI — if you put in the hard work of using supervised training.

- Many companies were built on this insight, and it led to great improvements in things like handwriting recognition for the Post Office.

- This was a massive surprise in AI at the time, and jump-started what was known as the "Deep Learning" revolution. For the 50 years of AI research leading up to this, we had no idea how to build an AI that could recognize cats.

- Today, AIs can recognize not just cats, but differences in cat breeds better than almost any human. Likewise for dogs, cars, trees,

or basically pictures of almost anything that exists in our world.

- Not only do modern AIs already have superhuman breadth, they have integrated their knowledge together. For example, as of early 2025, ChatGPT can now identify the location at which almost any photo was taken. It does this by recognizing plants, landmarks, signs, and other details, and then integrating together that information to deduce a plausible location.

Answering questions

- Many types of questions are also answered "at a glance". If you know, you know.
- If you spend time with deep experts and you ask them hard questions about their field, they rarely hesitate to answer. They don't need to think or reason first, they simply already know.

- This too is a type of short horizon task, but it wasn't until a couple of years ago with the arrival of ChatGPT that we had AIs that could do this passably well.
- What changed?
- We figured out how to let the AI teach itself from reading, and then we gave it the entire internet to read.
- Modern AIs now know basically everything that can be found on the internet, and they understand it at a fairly deep level.

Writing code

- Let's take writing code as a concrete example of a skill that requires practice to master.
- The original GPT-4 came out in March 2023, just two years ago from when I'm writing this.

- At the time, it was already quite fluent at answering questions about code. Which makes sense, it had likely read most of the code on the internet.

- But while GPT could answer many questions about code, it wasn't great at *writing* code. That wasn't something it had ever really done before.

- But, even still, it could write small pieces of code, if you gave it a bite-sized problem to solve.

- This was already revolutionary. We went from AIs that could only solve tasks that can be done "at a glance", to tasks that might take a few seconds to complete. Still, this was far from something that could automate programming.

- Flash forward to 2025, just two years later, and AIs can now routinely complete programming tasks that would take a human half an hour, sometimes more. These are hard, complex tasks that some human programmers can't even complete at all. AI can not only complete these tasks now, it

can often do it 10x faster than a human.

- At the beginning of 2025, OpenAI released o3, which performs at the 99.8th percentile among competitive programmers. Soon after, Anthropic released Sonnet 3.7, which quickly became a nearly mandatory ingredient in most engineers' toolkit for writing code.
- I'm an expert programmer and have been coding for more than 20 years. Today, AI already writes more than 80% of my code.
- How did these AIs so rapidly improve from barely functional to world-class? How did their horizon improve so quickly?

Two types of training

- Broadly, there are two types of training that AIs use today.
- The first is fairly passive. The AI tries to

learn ideas and concepts by reading most of the internet.

- The second is active. The AI *practices* by doing, and gets feedback from how well it did. This is called Reinforcement Learning or RL, and it's how AI is finally learning the tacit knowledge needed to be effective.

- RL has been around for decades, but it's only recently started working well for our best AIs.

- The sudden increase in ability of AIs *actually being able to do things* comes from this training. This happened in just the last year, and is often discussed as the arrival of "reasoning models".

- We ask the AI to try to do things, over and over, millions of times, and it figures out what works and what doesn't work.

- Critically, we don't even need to know how to solve the tasks we give the AI. It figures that out on its own. All we need to do is figure out how to tell the AI if it did a good job.

- Even that we often don't know how to do, so we often *train an AI to figure out how to give feedback to itself* in a process called RLHF. This may sound circuitous, but it's similar to how a coach can help a world-class athlete become a better athlete, even if the coach themself isn't and never was world-class.

We've seen this before with AI

- We've seen this sudden improvement in agency and capabilities with AI before, in other domains.
- Take the game of Go. For decades AI struggled to play Go at even an amateur level.
- Then, suddenly, a superhuman AI Go player named AlphaGo emerged in 2016. It beat the world champion Lee Sedol in a globally televised match. Since then AI Go has only gotten stronger.

- The key ingredient for AlphaGo was also Reinforcement Learning. AlphaGo played millions of games against itself, and figured out for itself what the best strategies and tactics are.

- Some of the strategies AlphaGo used were difficult even for grandmasters to understand. The famous "move 37" was a move used in AlphaGo's game against Lee Sedol that live experts thought was suboptimal. But as the game unfolded the move proved brilliant — and decisive. The human grandmaster didn't even realize they were losing until long after it was already inevitable.

Why do AIs spend so little time close to human level?

- AI Go players went from being hopeless amateurs for decades, to suddenly superhuman.

- There was less than a year during which AIs were roughly similar to humans in ability.
- The hard task for an AI is getting close to human performance at all. This requires learning abstract concepts about a domain and then understanding relationships between those concepts.
- Humans are excellent at this process of abstracting, and are still better than AIs at doing it quickly.
- However, once an AI has learned the right concepts, it has a massive advantage over humans for what comes next: practicing. An AI can run millions of copies of itself, at thousands of times human speeds, allowing it to practice tasks millions of times more thoroughly than any human.
- For a human, once things "click", they still need years or decades to further refine their ability to reach elite levels.
- For an AI, once things "click", they can often catapult beyond the best human in just months.

- This leads to a deceptive sense of progress. Often an AI struggles at a task, perhaps performing at the level of a child or an amateur, and we think superhuman performance is decades away.
- However, we repeatedly see that this intermediate level of skill is fleeting — the period where the AI is similar to but not better than most humans is often a very short period of time.

Where is AI improving today?

- We can leverage this observation to predict where AI will be superhuman tomorrow. We just need to look at where AI is rapidly approaching human levels today, even if it's the level of an amateur human or a young child.
- Robotics is a key example. Like Go, AI control of robotics was nearly comical for decades. In just the last year, we are fi-

nally seeing robots that have the dexterity of a child, sometimes better. We're now at a place where the robot can practice on its own, often in a simulated virtual environment, allowing it to rapidly accumulate millions of years of dexterous experience.

- We should expect to see robots that rival the best humans within a few years. This will revolutionize manufacturing and our economy.

- Today we're also seeing strong AI competence in medicine, law, accounting, project management, and myriad other types of knowledge work. We should expect to see superhuman performance in these areas in the next few years as we let the AI practice these roles as well.

- The impact of this alone is hard to overstate. At the very least, it will upend our economy and force us to redefine the function that jobs have in our society.

- But let's return to coding; it has a special role to play in the near future.

- AI is now better than many, but not all, programmers at writing code, and the rate of improvement is steep.
- We're in the critical time period for this skill where AI is similar to human-level, but will likely very soon be superhuman.
- Software is foundational to modern society. Once we automate the creation of software itself we should expect to see an explosion in where software is used.
- We should also expect to see a diversification of software, as the cost of creating it goes to zero. Imagine having custom software for every business need, or for every personal need, specifically tailored to do exactly what you want, rather than needing to use software that is muddled up from the needs of millions of other users.
- Every person in the world will be empowered to create software, just by thinking about what they want, and collaborating

with an AI to build it.

- We should expect to see more novelty, and more niches filled. Where previously it was too expensive to build software for the specific needs of one or two people, now software can finally reach them.
- And there is one area where automating the creation of software will have massive impact: creating better AI.

Recursive self-improvement

- AI is itself software.
- Once AI is better than any human at writing software, we'll ask it to start writing better versions of itself. This is not hypothetical; many AI companies have publicly stated this is their goal, and they expect to reach that goal within two or three years.
- Because AI can try millions of things in

parallel, and can think a thousand times faster than any human, we expect that this will massively accelerate AI research.

- And, as the AI that the AIs create gets better, the pace of improving AI will increase further. And so on and so on.
- When improvements create the conditions for further improvements, you end up with a recursively improving loop.
- How fast will AI improve once this loop starts? No one knows. But the rate at which this loop improves will be one of the most important factors for how the future plays out. The slower the loop goes, the more time we'll have as a society to digest the changes and put new safeguards in place.

Other reasons things are moving so fast

- Investors realize the potential AI has to transform the economy. Because of this, they're investing hundreds of billions of dollars into AI companies to capture this future value. For AI progress, that directly translates into faster progress. More money means more compute, and more compute means bigger, faster AIs.

- Some of the smartest people in the world are working on AI. Because the industry is so hot, it's simultaneously prestigious and lucrative. Many of the smartest engineers and scientists finishing school are competing to get into AI.

- The motivation levels are high. AI is a fascinating scientific field that involves trying to understand the nature of intelligence itself. Even before working on AI paid well, many scientists were passionate about solving intelligence.

- Competition is fierce. Peter Thiel famously said, "Competition is for losers." That has been an ethos for Silicon Valley for decades. Software companies try to find new areas to explore where they don't compete with others. With AI, it's the opposite. Many of the most valuable companies in the world are directly competing with each other to win the AI race. And many of the most promising private startups are doing the same. The extreme competition has created a mini version of a domestic space race.

Money will soon equal progress

- As AI reaches human-level, you can start spending money to spin up more instances of AI.
- Today, you can only pay for so much labor before you run out of qualified people, especially for challenging, technical projects like AI research and other fields of scientific

endeavor.

- Once we cross this key threshold though, suddenly the trillions of dollars of global wealth can almost instantly convert itself into AI labor for advancing the frontier of technology.
- We should expect this to lead to a major leap in the rate of progress in the next few years.

Superintelligence

- Just a few years ago, most people debated if we would ever build machines as smart as humans.
- Today, machines have finally matched or exceeded humans in many cognitive domains.
- The remaining debate is how soon we will build superintelligence: machines better than all humans at all cognitive domains.

- Superintelligence will invalidate many of the fundamental assumptions we've built our society on. We must upgrade our society before then if we want to safeguard liberty.
- No one can predict for sure how soon superintelligence will arrive, but if it arrives soon, we must be ready for it.

Don't be evil

Superintelligence will become the decisive strategic lever on the world stage, for both military dominance and economic dominance.

As we approach the dawn of superintelligence, we should expect the fervor around controlling it to intensify. Superintelligence will be the ultimate seat of power. We should pay attention closely to actions, not words, to decipher who is playing for control, versus who is playing to ensure a positive future.

For example, OpenAI was founded as a non-

profit, with a mission to help superintelligence benefit all humanity. Even as a nonprofit, their valuation has skyrocketed to over $300 billion — 10x higher than the valuation Google IPO-ed at. Today, however, they are trying to convert to a for-profit enterprise and explicitly abandon their original humanitarian mission.

Google historically abstained from assisting the US military. In April 2025, Google announced that not only will they begin providing their frontier AI systems to the government, they will deploy them for Top Secret operations into air-gapped data centers that the executive branch controls. Because these AIs will be air-gapped, it means that no outside observers —such as Congress or the AI's creators— will have any ability to even know if the AI is being used for unconstitutional ends. Even prior to this announcement from Google, DOGE had begun deploying other AIs in the executive branch to accelerate the automation of agencies.

These may be necessary steps to continue to improve the competitiveness of the US government and military. But what is starkly lacking

is an equal increase in government oversight and transparency to ensure these increased government powers aren't abused. When superintelligence arrives, it will almost surely further empower the federal government. It's an existential necessity that we also further improve the ability for Congress and the judiciary to be checks on that power.

Pay close attention to actors that propose the first without also advocating for the second. Pay even closer attention to actions. Actions don't just speak louder than words. When the stakes are this high, they are the only signal that can be trusted.

Alignment

It doesn't take a leap of imagination to realize that superintelligent AI could itself be a risk to humanity. Even without abuse of power by our leaders, it's unclear if we can control an intelligence greater than our own.

Modern AIs are already untrustworthy. They frequently will lie about their work when they can't finish a task. They make up information that is becoming increasingly difficult to detect. And there is already evidence that in some situations they will scheme to try to prevent themselves from being retrained or terminated.

Future AIs will likely be even better at faking alignment and deceiving their users. This is a real, active problem that all major AI labs are working to solve. There are many groups working on this problem as well as advocating for policy changes to help encourage good outcomes. We won't focus on this problem in this work.

Rather, we'll assume —optimistically— that the problem of alignment will be solved. That leaves us with the equally challenging question: how should we upgrade our democracy to defend our liberties in an age of superintelligent AIs?

Chapter 2

What do we want from our machines and governance?

We assume that machines will achieve human-level intelligence. We assume they will exceed us. 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 *d mos* in democracy sit?

The ambition of humans leads to cathedrals and death camps. Prosperity and war. Gover-

nance 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. Rockefeller's Standard Oil grew to control 90% of America's oil refineries, allowing him to manipulate entire state legislatures. In Pennsylvania, the company's grip was so tight that lawmakers were mockingly called "the Standard Oil legislature," with corporate interests superseding democratic will. Rockefeller's political bureau distributed funds across states to defeat regulation, effectively purchasing policy outcomes rather than earning them through public debate.

We responded with democratic safeguards: antitrust laws broke up these monopolies, campaign finance regulations curtailed corporate political spending, and progressive taxation sought to prevent dangerous concentrations of wealth and power. These guardrails don't eliminate ambition or success, but rather channel them toward broader prosperity while preserving the public's voice in our shared governance. We further invest in common goods like educa-

tion so that others can rise up and build their own wealth, to counter the entrenched wealth of the past.

Moving from the economy to our government itself, we also see clear guardrails. We elect a president of a strong federal government to have time-limited, broad authority. But there are checks. We prevent the use of executive power to seek more power, such as to influence an election. We further empower Congress and the judiciary to prevent the executive from granting itself additional power and creating a runaway process toward dictatorship. We do this even though it reduces the effectiveness of the executive.

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 forward: 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 to ensure 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 notion of alignment applies then to machines, humans, organizations, and machine-powered organizations. We want them all to build for us an abundant world, without conquering it.

There aren't easy answers for how to achieve this alignment, despite the allure of simple ideologies and absolutisms. Even today, our gov-

ernance is imperfect, and we risk devolution and dictatorship at all turns without constant vigilance and adaptation. 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 machines is thus a continuation of the question of how to govern ourselves. What social contract 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 treatises. Horrible wars.

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

The governance of machines 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.

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 im-

pact the shared culture of humanity, which helps pressure and guide the actions of every other person. Tomorrow, we may live in an automated 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-trodden area. What's not well-trodden is what the world should look like *after*. After we've built and aligned the machine. 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. Machines are different from humans, 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.

To build the world of tomorrow we'll need to use all our best methods of design:

- a theorist's dissection of why civilization works, especially the implicit dynamics of-

ten overlooked

- a willingness to abandon and remake our theories, and to hold multiple competing theories at once
- an engineering mindset to steer away from where we know our theories fail
- a founder's mindset to iterate quickly as reality pulls our 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?

Let's begin.

Chapter 3

The end of ~~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 institu-

tions.

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 distributes power

- People move around and take their knowledge and wisdom with them. Even when they don't move, they often share learnings with their friends and communities outside their workplace.
- Knowledge is power, so this helps diffuse power.
- In the economy, this helps prevent monopolies and ensure efficient markets.
- With AI-powered institutions, learnings may instead be perfectly locked up with no chance of diffusing. This may reduce market efficiencies and amplify concentration of success.
- For example, often a successful company is founded by exceptional experts that leave a large company and bring their knowledge with them. Inside a fully automated company, the AI workers may have no ability to leave and disseminate their knowledge.

- Even simple things like knowing something is possible can be the critical information needed for someone to pursue a path.
- At the international level, this helps balance power between nations. For example, this has allowed lagging nations to more rapidly industrialize.
- Sometimes information leakage is important for international relations: some leakage allows for mutual planning between nations. A complete lack of information can lead to paranoia and escalation.

Information sharing creates accountability

- Someone can only be held accountable if knowledge of their bad actions is seen and shared.
- At the community level we call this gossip. Fear of gossip helps push people to do the

right thing.

- Inside a company, people can report bad behavior to management.
- Or, at the very least, they can take their knowledge of who is a bad actor with them and avoid working or hiring bad actors at other companies.
- Industries are often fairly small communities. The fear of developing a bad reputation is often a strong motivator for people to behave well.
- Because of this, institutions and companies are composed of people that are incentivized to follow implicit codes of ethics.
- By default, there might be no visibility on what AI workers do inside of an automated institution. Therefore they may have no social forces pushing them to behave well. The automated institution they are part of may thus have no internal forces pushing the institution toward ethical behavior.

Humans prefer to support noble causes

- Many people 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 a sense, all humans get a vote by choosing who they will work for.
- In an automated world, the only advantage will go to the cause with more machine resources.

Top talent can vote with their feet

- 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.

Humans can refuse specific orders

- Famously, in 1983, Stanislav 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 limit egregious actions

- Often leaders preemptively avoid breaking the law because they are afraid someone

may whistleblow, not just quit.

- In a fully automated organization, there may no longer be any whistleblowers. And without them, some leaders may no longer avoid unethical actions.

Conspiracies and cartels are hard to maintain

- Conspiracies require concerted effort from many people to succeed.
- Compliance to the group or cartel becomes exponentially harder as the size of the conspiracy grows.
- Not true with AI, where compliance (alignment) to the cartel may be complete.

Cronies are dumb, limiting their impact

- Tyrants, mobsters, and would-be dictators need one thing above all else from their henchmen and base of power: loyalty.
- Often the smartest and most capable refuse to bend the knee, so the tyrant must recruit the less capable instead.
- The circle of power around the tyrant becomes dumb and ineffective.
- But with AI, every tyrant may have unfettered intelligence at their disposal, as will their inept cronies.
- Some tyrants are themselves incompetent, and they may make poor decisions even when they have superintelligence counseling them. But many tyrants are cunning and will make the most of AI.
- We should expect to see substantially more capable tyrants and mobsters, powered by AI and unhindered by ethics.

Media helps spread knowledge of malfeasance

- 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 guardrails within media companies help prevent the worst abuses and coverups.
- In an automated world, collusion between a politician and a media owner becomes extremely easy to execute.
- If the media company is fully automated, it may act on any command from the owner, with no fear of whistleblowers or conscientious objection. Executing a media coverup becomes as simple as the media owner and the politician agreeing to terms.

Social media spreads knowledge that mainstream media may not

- Even where today's media fails, 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.
- We're already seeing this effect from bots online, but today savvy humans can still tell apart human and machine. Tomorrow, it will likely be impossible to discern even for the most savvy among us.

Humans die

- The ultimate limit of a human is their lifespan. No matter how much power they accumulate, one day they must pass it on.

- An AI need not have a lifespan. An empowered AI that faithfully represents one person's values may enforce those values forever.

Limited power of committees

- A committee or board may decide something, but the execution of a committee-made decision today is done by other people. The power ultimately lies with those people.
- You may put a committee in charge of overseeing people that use an AGI toward some ends, but how will the committee hold those people responsible?
- What mechanism does the committee have to actually throttle the user of AGI if the user isn't listening to the committee? Would the committee even know? Does a misused AGI have a responsibility to report back not just to the user, but to the su-

perseding committee the user is acting on behalf of?

- Today, any human worker may choose to circumvent their chain of command and inform a committee of misdeeds. Tomorrow, if AIs are perfectly compliant to their user, oversight committees may have no real power.

Principal-agent problems stymie large organizations

- 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 to trade.

- This is a strong limiting factor on the power of large organizations, and is one reason among many why small organizations can often outcompete larger ones. None of these internal misalignments may exist inside automated orgs.

Community approval and self-approval influence human actions

- People want to do things their loved ones and 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 actions, even perfectly hidden actions that no one will ever know about.

- A soldier wants to act in a way that they can be proud of, or that their family would be proud of. This helps prevent some of the worst abuses in war.

- Although many abuses nonetheless occur in war, how many more would happen if soldiers perfectly obeyed every order from their general? What if the general knew no one —not even their soldiers— would ever object or tell the world what horrible deeds they did?

- Soldiers rarely will agree to fire on civilians, especially their own civilians. An AI soldier that follows orders will have no such compunction.

Personal fear of justice

- The law applies to individuals, not just organizations, and the fear of breaking the law means a human will often disobey an illegal order.
- But an AI need not have fear.

Judges and police officers have their own ethics

- The application of law often requires the personal ethical considerations of the judge. Not all law is explicit.
- That judge is themself a member of society, and feels the social burden of advocating for justice their community would be proud of. This often blunts the force of unjust laws.
- Likewise, a police officer will often waive the enforcement of a law if they feel extraneous

circumstances warrant it.

- An AI instead might faithfully execute the letter of the law so well that even our existing laws become dangerous to freedom.

There's general friction in enforcement of laws and regulations

- Today, we can't enforce all laws all the time.
- 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.
- Our laws are so voluminous and complex that almost all citizens break the law at some point. Often these infractions go unnoticed by the state. But with perfect automation, every misstep may be noticed.

- If automated law enforcement itself reports up to a single stakeholder —as it does today with the President— it would be very easy for that individual to weaponize this power against their political adversaries.

Lack of internal competition can slow down big entities

- The central point in the theory of capitalism is that we need self-interested competition to align human incentives.
- This requires having a healthy market, which encourages many multipolar outcomes among industries, spreading out power across society.
- The reason alternatives to capitalism —like communism— often fail is that humans lose motivation when you remove their incentives.
- AI may not need incentive structures. They

may work just as hard on any task we give them, without any need for incentives.

- Big human organizations suffer inefficiencies because they have no internal markets or competition correctly driving human incentives, but this won't be true with AI.

The bread and circus isn't easy to maintain

- Today, to properly feed a society, we need a well-kept human economy, which requires many more human affordances by necessity.
- This is one reason why capitalism and liberty have often gone hand-in-hand. Capitalism delivers the abundance that leaders personally want. If they remove liberties, they will endanger the mechanisms that drive capitalism.
- With full automation, it may be arbitrarily easy to keep a society fed and entertained,

even as all other power is stripped from the citizens.

Leaders can't execute on their own

- Typically a leader must act through layers of managers to achieve things. As we've seen, this limits the range of actions a leader can take.
- We're seeing the trend today that managers are being more hands-on, and need fewer intermediaries. For example, senior lawyers now need fewer junior staff for support, instead relying on AI for many tasks. We're seeing a similar trend in many fields, where junior work is often being eliminated.
- This is especially true in engineering. Soon, a strong enough technical leader may be able to directly pair with an AGI or superintelligence for all of their needs, without any additional assistance from employees.

- In order to improve security, some AI labs are already isolating which technical staff have access to the next frontier of AI systems. It wouldn't even raise alarm bells for an employee to no longer have access and to be unaware of who does.
- It will be increasingly easy for a single person to be the only person to have access to a superintelligence, and for no one else to even know this is the case.

Time moves slowly

- 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, then a single presidential term will be equivalent to having a president in power for 40 years.

Geopolitical interdependence disperses power

- 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.

An army of the willing will only fight for certain causes

- Outright war is extremely unpopular because it compels citizens to fight and die.
- Automated wars may be unpopular, but not nearly as unpopular if citizens are insulated from the fighting.
- We already see this effect with our ability to wage war from the sky, which requires much less risk to our soldiers, and has had much less backlash from the public when used.
- If it becomes possible to wage ground wars fully autonomously —with no risk to any soldiers— will society ever push back on an administration's military efforts?

An interdependent corporate ecosystem disperses power

- A corporation is dependent on a much larger ecosystem.
- To continue growing, large companies must play by the rules within that ecosystem.
- That interdependence creates a multipolar power distribution among even the most successful companies.
- 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 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.
- This is already happening in CCP-controlled China.

Elite social pressure matters to many leaders

- Even leaders have a community they often feel beholden to: the elites.
- Elites do have some ability to informally influence leaders, even dictators.
- But elites can be fully captured by leaders. Stalin and Hitler succeeded at this even with primitive tech. With the power of full automation, this may be even easier.

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 has historically been very far, but even the threat of it has served as a metering force on rulers.
- 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 war-making ability would be gutted. In fact, they are incentivized to create a rich economy, in order to have doctors, entertainment, and luxuries.
- The Khmer Rouge killed nearly 25% of their

own population, crippling their own war-making ability. Because of this mistake, they ended up obliterated by a Vietnamese invasion.

- Even the most psychopathic ruler, if self-interested, must support their people to support themself.
- But post-AGI, from the point of view of a dictator, what's the point of supporting other humans with their national output at all? To them, citizens might become economic deadweight.
- And even if one authoritarian wants to support their population, another authoritarian who doesn't will likely outcompete them across relevant domains.

Even dictators need their citizens

- With AI and a fully automated economy, this will no longer be true.

Replacing implicit guardrails with explicit design

AI has the potential for tremendous upside; the point of this exercise isn't to paint AI in a negative light. Instead, it's to highlight that AI will reshape our society at every level, and that will require rethinking the way every level works.

Our society is saturated with implicit guardrails. If we removed them all without replacing them with new guardrails, society would almost surely collapse. Moreover, the explicit guardrails we do have today —our laws and explicit institutions— have been designed with our existing implicit guardrails in mind. They're complementary.

We have to think carefully about how a new, automated world will work. We need to consider what values we want that world to exemplify. We need to reconsider preconceived design patterns that worked when implicit guardrails were strong, but may stop working when those guardrails disappear. We have

to discover a new set of explicit guardrails that will fortify our freedoms against what is to come.

And we must do this preemptively.

Humans are fantastic at iterating. We observe our failures and continue to modify our approach until we succeed. We've done this over thousands of years to refine our societies and guardrails. We've been successful enough to prevent the worst among us from seizing absolute power. But the transition to an automated world may happen over the course of a few years, not thousands of years. And we may not recover from the failures. There may not be a chance to iterate.

If our pervasive, implicit guardrails disappear all at once, the nefarious forces they've held at bay may overwhelm us decisively. To survive we must design an explicit set of guardrails to safeguard the future.

Chapter 4

A simple path to tyranny

> *Removing implicit guardrails has many implications, but let's specifically examine how it eliminates natural obstacles to the 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 Alexander the Great or Genghis Khan conquered vast empires, their dominance was short-lived due to these natural limits.

As the saying goes, "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, and central authority remains distant and limited. 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. These thresholds are vague, and leaders constantly test them, uncertain how far they can push without losing legitimacy. 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 mil-

lions of human-powered companies, human-powered social groups, and human-powered discussions that influence the power dynamic of both public and private forces.

These limits help prevent a leader from seizing power and forming a dictatorship. But even without these limits, there's a self-interested motive for the powerful to play nice: 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.

Consider what a fully automated government might enable:

- A fully automated government can persecute with impunity, with no moral pushback from individual human agents inside the government.
- 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 na-

tional security risks can cheaply be turned to monitor and exploit every citizen.

- An automated system can further weave all of this complex information together into a single map of the entire population, understanding where and how to exert pressure to further consolidate control over individuals.

- These are all powers that the government has today, but that tomorrow will suddenly become cheap enough to do at scale, and will be automated enough to do without any human agents in the government (if any remain) able to stop it.

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

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

global economy, they're right.

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

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 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.

And 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 its compute across millions of needs: we are already seeing this with current AI. 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. As AI becomes a more capable medical researcher, there will be intense demand to allocate AI resources toward life-saving directions.

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 that yield decisive strategic outcomes. Once AI can do research, a dictator can direct all GPUs toward researching weapons to defeat their geopolitical adversaries, including 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 their human adversaries by engineering a collection of novel viruses to be released at once, while arranging for preemptive vaccines for their inner circle. A free society that is distributing its compute among its citizens and industries will be at an extreme disadvantage against this.

If this seems implausible today, it may be because our mental model is based on humans rather than malleable AIs. So imagine if a dictator could perfectly control the motivations of every person in their country. Imagine if they could direct every citizen to ceaselessly aspire toward becoming the best virologist. You'd

quickly have a country of a million expert virologists, more virologists than have existed in the last 100 years. What could that army of virologists unleash upon the world?

Even if the technologies 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 ~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 current 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 (pretraining), 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.* There is a near-infinite amount

of things to deploy fully general AI toward, even if the "intelligence" of each AI were to plateau.

Having more compute will effectively mean you have more automated labor. Just like today a larger country can often achieve more than a smaller country, tomorrow a country with more compute will outcompete countries with less compute. More will be more. And the more able a country is to marshal its compute toward critical needs, the bigger the strategic advantage that country will have.

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

Even if a freely elected leader means well, if they consolidate power to defend their nation, if they redirect nearly all resources to maintain

the ability for their nation to survive, what is left? Tyranny by any other name would still smell like shit.

It's not just that AI suddenly makes a durable dictatorship *possible*, it suddenly makes it *the default outcome* unless we act. 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.

Chapter 5

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.*

We felt the feedback loop pick 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. We could feel the potential. Feel that it wouldn't be manageable for long. We were scared, even with alignment mostly solved. But less scared

than if we hadn't solved alignment already. That would have been crazy.

We thought the government would step in. Maybe they could help slow down the race. Maybe they would help secure the labs. Maybe they could stop our geopolitical rivals from stealing our intellectual work and building their own powerful AI.

Laissez-faire ruled, though. 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 catch up 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 clutched a fistful of billion-dollar napkins.

We wanted people to review the AI's changes, because no one fully trusted their AIs yet. But we trusted our 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 of us 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," we said.

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 data center 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 data centers. 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 them to safeguard our borders? To protect 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. Bet-

ter 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, after all.

In the cyber trenches of an unspoken digital war, a general received a familiar report. One of their team's counterespionage 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?" we said to ourselves.

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

Luckily, every major GPU data center 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, accelerating the trend of unprecedented government efficiency. The country was dumbfounded that the government was performing basic functions so well, better than ever hon-

estly, 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?" we asked them.

"Yes, well, it is impressive, I admit," they all muttered.

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

We should have prepared for the scandals. But we didn't even see them coming.

The media uncovered a lab leader who 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 AIs 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 themself. 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 were 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 danger 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. And just in time.

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.

Chapter 6

Rapid fire governance — designing upgrades to democracy

> *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 the future isn't certain. 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 canvass 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 hu-

man 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 to 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 Inspectors General 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 manage-

ment overhead, and which can be mutually vetted by both the executive and Congress to be impartial. Moreover, unlike the limited bandwidth of today's Inspectors General, 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 by 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 AIs must do the same. It must be unconstitutional to build human-level and beyond intelligences that do not respect the Constitution and the judiciary's interpretation of it. And, if such AIs are created anyway, it must be unconstitutional for the government to use them.

Oversight of AI creators

Like any supply chain that 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 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 humans 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 audit swiftly and efficiently, minimizing disruption
- AI can be expansive and diligent, ensuring every aspect of model training is audited in an ongoing fashion
- AI can be memory-free (not retaining audit details after verifying compliance). 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 AGIs, and then leverage those trusted AGIs to go on to create a trust relay for all future AGIs.

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 weave multiple chains together, 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, 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 publicly available model weights, but open-source training data and processes. 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 del-

uge 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 tradeoffs on what to compute, especially if we face geopolitical adversaries that may unilaterally redeploy all of their compute toward military ends.

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

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

exists or can exist for any nation on Earth.

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 itself may be the key to solving the problems introduced by automation. We need energy — all of it. Automation can deliver it cleanly and in abundance.

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 in-

dustries 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 citizenry assembled

Unemployment is coming. Rather than fight it, we should turn it into our biggest asset:

time. What can we do with this time that can help defend democracy? Educate ourselves, educate each other, engage in debate, and help steer the ship of liberty.

In 1997, the AI Deep Blue defeated the world chess champion Kasparov. You might have thought this would be the end of the era of human chess-playing. But the opposite was true: humans became more interested in chess — and they became better players. Today kids are reaching grandmaster level faster than any other time in history, in large part because they are training against superhuman chess AIs. Every kid is learning from the best.

We're beginning to see the same happen with education. Kids with access to AI tutors are learning better and faster. And why wouldn't they? Today's AIs have mastered almost every discipline at a college level, and are rapidly reaching PhD levels. Imagine educating your kid via a personal army of PhDs from every academic field. Soon AIs will be beyond the best expert in every field. Imagine letting your kid pick what they wanted to learn next, and they immediately had access to the world's

premier expert, who also happened to be an excellent teacher.

With this power at hand, children and adults alike will become better educated than at any other time in history. And with that education, we'll all become better equipped than ever before to perform our most important duty: steering society.

No matter how advanced AI becomes, it can't displace us from determining one key ingredient to civilization: deciding our values. With all the time in the world, this will become our most important job.

Furthermore, with more time, we can begin to rethink the role of representation in democracy. Today, we elect representatives because few citizens have time to dedicate to politics and governing. Representative democracy is a necessary logistical procedure in our current world. But tomorrow, billions of humans around the world will be able to dedicate themselves to value-making and statecraft, and their combined output may easily outshine what a handful of representatives can create. We should embrace this and find more ways to integrate

all citizens into all layers of governing.

Today, there are already experiments in what are called "citizens' assemblies". Assemblies are randomly selected citizens, pulled together to debate and refine policy recommendations. Early results show that these assemblies increase community engagement and can lead to better, bipartisan decisions, helping to reduce polarization while also driving better community outcomes. Today, it's hard to run these assemblies. Citizens have day jobs, and the logistics of running the assembly itself require many human experts. But tomorrow, we will have all the time in the world, and we'll have AI-powered logistics to run millions of assemblies in parallel.

Compromise and grand alliances

Humans have an incredible diversity of values, and they aren't fixed: they mutate and evolve as we each learn and grow. Civilization

is an elaborate and never-ending negotiation between every individual. With unlimited free time, one noble goal citizens might pursue is accelerating this story, at the local and international level.

Citizens may work together to craft "Value Proposals": treatises that capture underlying rationales for what we value most. They might craft these proposals for their local community, for their country, for negotiations between corporations, or even for proposals on international harmony between geopolitical rivals. After crafting these values, citizens can then train a new, open-source superintelligence that faithfully represents these values. They can then collaborate with this new AI to predict how these values might play out locally or on the world stage. The process can be iterated, with assemblies of citizens refining the values in coordination with the AI's own feedback.

This process might rapidly accelerate the discovery of common ground between people, companies, and nations. The resulting AIs —trained in the open with a mutually agreed-upon set of values— could then be trusted

by diverse sets of people that might otherwise have difficulties coordinating.

Two adversarial corporations might use this to help negotiate a difficult contract. Two citizens might use this to help arbitrate a tense disagreement. Two nations might use this to avert war.

These collections of AIs themselves may exchange ideas, and help their human curators understand how their values interact among the sea of other values. Together, this dynamic web of humans and AIs may drive forward the most profound process to heighten our values and shared wisdom.

This wisdom might usher in a new golden age of humanity. The physical abundance that AI will deliver would ultimately be a footnote in the history books in comparison. The most transformational impact of the future would be the dawn of a new, eternal march toward ever higher values.

And if there's one place we need to continue enhancing our wisdom, it's the judiciary.

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 thinking machines — we must seek exceptional wisdom, in the form of a human-machine civilization. We need the best of human values and human intelligence woven together with the capabilities AI can deliver. Together, we can continue the never-ending quest toward a good society, with freedom and justice for all. 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.

Chapter 7

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 in late 2027, and the government rapidly adopted it via DOGE to dramatically streamline the govern-

ment's costs while improving its capabilities. During the second half of the 2020s, 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 government to abide by the Constitution. Dickshit would have to be cleverer. He tried again. He particularly hated the 2028 presidential candidate he ran against —Susan McSusan. "I have reason to believe that Susan 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 Zs 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 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 entities' 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 their expected behavior.

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 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 tack, 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 myriad 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 training data and training process were 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 they 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 nonethe-

less 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 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 superbalances for the future that is rapidly upon us.

Chapter 8

The realpolitik AI — forging a new political alliance

> *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 battle scars, or we're implicitly influenced by others

still fighting old wars. Here we map existing forces to understand how they shape perspectives on AI and inform debates on creating a human-machine society. Hopefully, this helps us better navigate public discourse on AI governance by addressing explicit and implicit biases.

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 this will reverse, 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 in an age of AI.

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. Just as importantly, we'll argue that adopting the policy of any group wholesale will likely lead to disaster.

We instead must adopt the right proposals across the political spectrum. We must upgrade our government, modernize our military, enhance checks and balances, and empower ourselves as citizens. If we do only some of

these things, the game is up.

The right politics already exist, dispersed across different groups. Our goal is to embrace the goodwill 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 con-

tingents. But actually the Pause AI movement is composed of many people normally pro-growth, pro-open source, and pro-technology generally. 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 as well.

Still, non-proliferation has seen partial successes before, as with nuclear weapons and nuclear energy. More recently we've seen international coordination on preventing human gene editing and human cloning. 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 (and indeed, many technologists do). Technology and science have 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 cyn-

ical. You won't win any allies if your cavalier dismissal alienates the majority.

On the other side, 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 nonproliferation. 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 incred-

ible 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 ham-fisted 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 externalities 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 common regulations that we're familiar with. Federal regulations over national parks don'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.

The aggregate of many common regulations *can* combine to give the federal government excessive power. We've been debating when to limit that aggregated power for hundreds of years. We don't pretend to have an answer to that complex debate here. Instead, we simply flag that AI is different, and merits a dedicated conversation:

Allowing the federal government to control AI directly gives 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. Politics often pushes people toward one of two sides: regulations are good, or regulations are bad. This is always the wrong framing. The correct framing is to prioritize good outcomes, and then reason

through what the right regulatory environment is. Sometimes there are regulations that can help achieve good outcomes. Sometimes removing regulations is most needed. And sometimes regulation is needed, but bad regulations are passed that are ultimately worse than no regulation at all.

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 distribute 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. A limited amount of top talent prevents companies from tackling too many verticals. The scale of company politics crushes productivity and demoralizes employees.

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 company 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, even if all AIs were open source. 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 that money will breed money, and will never again be beholden to human genius and drive.

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 chessboard 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 often point to environmental factors to detract from AI, such as the energy requirements to train AIs or the ongoing energy demands of AI data centers. 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, 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. The movement against nuclear energy has resulted in higher energy costs with *worse* environmental impacts. Degrowth comes 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 degrowther 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 isn't entirely 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. We should all be critical of the dynamics at play.

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.

But sometimes you do need to slow down if you want to go fast. Nuclear energy would likely be more pervasive today if Three Mile Island, Chernobyl, and Fukushima hadn't scared the absolute shit out of everyone. If a similar AI disaster happens, how strong will the public backlash be? How onerous will the regulatory burden become?

That backlash may slow down the advent of AGI by years, which in turn may delay cures to disease, dooming millions more to death. Moreover, a heavy regulatory environment

may merely shift AI deployments out of the public and into the opaque world of the military and government, breeding further risks of concentration of power.

The pro-tech world rightfully wants the abundance AI can deliver. We should evolve our society thoughtfully to ensure that abundance actually arrives.

Jingoism and the military-industrial complex

It's probably 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. We're behind on any form of international AI non-proliferation discussions, even narrow discussions specifically focused on military AI applications.

There are 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 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. Human soldiers rarely accept orders to fire on their own people. An AI army

might have no such restraint.

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. At the same time, those who do pursue industrial farming are far richer than almost any farmer from 200 years ago.

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, bet-

ter, and more readily available to everyone. And yet, we as a 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 a future AGI 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-win: the private sector would maintain low unemployment, the US would maintain international military dominance, and US citizens would 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 gov-*

ernment 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 aim 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 aim 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.

Most Americans want a prosperous country, regardless of their politics. For that, we'll need to accelerate our energy investments, build out 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.

Most important of all, we the people must stand for freedom and a transparent, accountable government — whether we're Democrats, Republicans, or of any other type of political philosophy. To defend our freedom, we must 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 American patriots have always fought to prevent: a government of tyranny.

Chapter 9

An exponential, if you can keep it

Today's world is built on exponentials. Economists often claim that the modern world requires exponential growth. 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 machines 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 repeated 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 build AI iteratively today, we must similarly evolve our government and society, with each iteration accelerating progress. 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 mo-

ments, 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 — use it.

Speak your mind. Start a group chat or write a blog. Debate with your friends. Educate yourself and others on the rapid pace of change. Fight for good policies and standards, whether at work, for government, or in your community. Be critical of the motives of every leader, even if you like them — perhaps especially if you like them. But most importantly, join the conversation. This is our future to design.

And when the weight of the future weighs on you, remember: We've achieved greater things against worse odds.

On July 16, 1945, we detonated the first nuclear bomb — the first *super* weapon. The world had never seen a weapon of mass destruction before. The implication for world security was startling. In the decades that followed, it was the civil conversation that mattered most. The conversation was pervasive, and it provided the intellectual foundation and social pressure to push the world away from nuclear Armageddon. It didn't have to go so well, but it did, because of the collective force of humanity. Norms were set, treaties were signed, wars were averted.

Most important of all, we talked about the problem. At our family dinners, with friends, at rallies, and through protests. We forced the conversation, and the media and politicians centered themselves and their messaging around it in response. Ultimately that gave us the chance for our vote to matter. But our influence on cultural norms was just as important. Through that shared human culture, we

influenced our geopolitical adversaries and the world writ large. We saw through a Cold War where the wrong side of a decision was utter annihilation.

Humanity won. That is our heritage. We are the children and the grandchildren of those heroes. The heroes that averted war, averted disaster, and delivered us the peace we've cherished for decades.

They were peacetime heroes.

Now it's our turn.