The Adolescence of Technology

Confronting and Overcoming the Risks of Powerful AI

January 2026

There is a scene in the movie version of Carl Sagan's book *Contact* where the main character, an astronomer who has detected the first radio signal from an alien civilization, is being considered for the role of humanity's representative to meet the aliens. The international panel interviewing her asks, "If you could ask [the aliens] just one question, what would it be?" Her reply is: "I'd ask them, 'How did you do it? How did you evolve, how did you survive this technological adolescence without destroying yourself?" When I think about where humanity is now with AI—about what we're on the cusp of—my mind keeps going back to that scene, because the question is so apt for our current situation, and I wish we had the aliens' answer to guide us. I believe we are entering a rite of passage, both turbulent and inevitable, which will test who we are as a species. Humanity is about to be handed almost unimaginable power, and it is deeply unclear whether our social, political, and technological systems possess the maturity to wield it.

In my essay *Machines of Loving Grace*, I tried to lay out the dream of a civilization that had made it through to adulthood, where the risks had been addressed and powerful AI was applied with skill and compassion to raise the quality of life for everyone. I suggested that AI could contribute to enormous advances in biology, neuroscience, economic development, global peace, and work and meaning. I felt it was important to give people something inspiring

to fight for, a task at which both AI accelerationists and AI safety advocates seemed—oddly—to have failed. But in this current essay, I want to confront the rite of passage itself: to map out the risks that we are about to face and try to begin making a battle plan to defeat them. I believe deeply in our ability to prevail, in humanity's spirit and its nobility, but we must face the situation squarely and without illusions.

As with talking about the benefits, I think it is important to discuss risks in a careful and well-considered manner. In particular, I think it is critical to:

- **Avoid doomerism.** Here, I mean "doomerism" not just in the sense of believing doom is inevitable (which is both a false and self-fulfilling belief), but more generally, thinking about AI risks in a quasi-religious way. [1] Many people have been thinking in an analytic and sober way about AI risks for many years, but it's my impression that during the peak of worries about AI risk in 2023–2024, some of the least sensible voices rose to the top, often through sensationalistic social media accounts. These voices used off-putting language reminiscent of religion or science fiction, and called for extreme actions without having the evidence that would justify them. It was clear even then that a backlash was inevitable, and that the issue would become culturally polarized and therefore gridlocked. [2] As of 2025–2026, the pendulum has swung, and AI opportunity, not AI risk, is driving many political decisions. This vacillation is unfortunate, as the technology itself doesn't care about what is fashionable, and we are considerably closer to real danger in 2026 than we were in 2023. The lesson is that we need to discuss and address risks in a realistic, pragmatic manner: sober, fact-based, and well equipped to survive changing tides.
- **Acknowledge uncertainty.** There are plenty of ways in which the concerns I'm raising in this piece could be moot. Nothing here is intended to communicate certainty or even likelihood. Most obviously, AI may simply not advance anywhere near as fast as I imagine. [3] Or, even if it does advance quickly, some or all of the risks discussed here may not materialize (which would be great), or there may be other risks I

haven't considered. No one can predict the future with complete confidence—but we have to do the best we can to plan anyway.

- **Intervene as surgically as possible.** Addressing the risks of AI will require a mix of voluntary actions taken by companies (and private third-party actors) and actions taken by governments that bind everyone. The voluntary actions—both taking them and encouraging other companies to follow suit—are a no-brainer for me. I firmly believe that government actions will also be required *to some extent*, but these interventions are different in character because they can potentially destroy economic value or coerce unwilling actors who are skeptical of these risks (and there is some chance they are right!). It's also common for regulations to backfire or worsen the problem they are intended to solve (and this is even more true for rapidly changing technologies). It's thus very important for regulations to be judicious: they should seek to avoid collateral damage, be as simple as possible, and impose the least burden necessary to get the job done. [4] It is easy to say, "No action is too extreme when the fate of humanity is at stake!," but in practice this attitude simply leads to backlash. To be clear, I think there's a decent chance we eventually reach a point where much more significant action is warranted, but that will depend on stronger evidence of imminent, concrete danger than we have today, as well as enough specificity about the danger to formulate rules that have a chance of addressing it. The most constructive thing we can do today is advocate for limited rules while we learn whether or not there is evidence to support stronger ones. [5]

With all that said, I think the best starting place for talking about AI's risks is the same place I started from in talking about its benefits: by being precise about what level of AI we are talking about. The level of AI that raises civilizational concerns for me is the *powerful AI* that I described in *Machines of Loving Grace*. I'll simply repeat here the definition that I gave in that document:

By "powerful AI," I have in mind an AI model—likely similar to today's LLMs in form, though it might be based on a different architecture, might involve several interacting models, and might be trained differently—with the following properties:

- *In terms of pure intelligence, it is smarter than a Nobel Prize winner across most relevant fields: biology, programming, math, engineering, writing, etc. This means it can prove unsolved mathematical theorems, write extremely good novels, write difficult codebases from scratch, etc.*
- *In addition to just being a "smart thing you talk to," it has all the interfaces available to a human working virtually, including text, audio, video, mouse and keyboard control, and internet access. It can engage in any actions, communications, or remote operations enabled by this interface, including taking actions on the internet, taking or giving directions to humans, ordering materials, directing experiments, watching videos, making videos, and so on. It does all of these tasks with, again, a skill exceeding that of the most capable humans in the world.*
- *It does not just passively answer questions; instead, it can be given tasks that take hours, days, or weeks to complete, and then goes off and does those tasks autonomously, in the way a smart employee would, asking for clarification as necessary.*
- *It does not have a physical embodiment (other than living on a computer screen), but it can control existing physical tools, robots, or laboratory equipment through a computer; in theory, it could even design robots or equipment for itself to use.*
- *The resources used to train the model can be repurposed to run millions of instances of it (this matches projected cluster sizes by ~2027), and the model can absorb information and generate actions at roughly 10–100x human speed. It may, however, be limited by the response time of the physical world or of software it interacts with.*
- *Each of these million copies can act independently on unrelated tasks, or, if needed can all work together in the same way humans would collaborate, perhaps with different subpopulations fine-tuned to be especially good at particular tasks.*

> *We could summarize this as a "country of geniuses in a datacenter."*

As I wrote in *Machines of Loving Grace*, powerful AI could be as little as 1–2 years away, although it could also be considerably further out. [6] Exactly when powerful AI will arrive is a complex topic that deserves an essay of its own, but for now I'll simply explain very briefly why I think there's a strong chance it could be very soon.

My co-founders at Anthropic and I were among the first to document and track the "scaling laws" of AI systems—the observation that as we add more compute and training tasks, AI systems get predictably better at essentially every cognitive skill we are able to measure. Every few months, public sentiment either becomes convinced that AI is "hitting a wall" or becomes excited about some new breakthrough that will "fundamentally change the game," but the truth is that behind the volatility and public speculation, there has been a smooth, unyielding increase in AI's cognitive capabilities.

We are now at the point where AI models are beginning to make progress in solving unsolved mathematical problems, and are good enough at coding that some of the strongest engineers I've ever met are now handing over almost all their coding to AI. Three years ago, AI struggled with elementary school arithmetic problems and was barely capable of writing a single line of code. Similar rates of improvement are occurring across biological science, finance, physics, and a variety of agentic tasks. If the exponential continues—which is not certain, but now has a decade-long track record supporting it—then it cannot possibly be more than a few years before AI is better than humans at essentially everything.

In fact, that picture probably underestimates the likely rate of progress. Because AI is now writing much of the code at Anthropic, it is already substantially accelerating the rate of our progress in building the next generation of AI systems. This feedback loop is gathering steam month by month, and may be only 1–2 years away from a point where the current generation of AI autonomously builds the next. This loop has already started,

and will accelerate rapidly in the coming months and years. Watching the last 5 years of progress from within Anthropic, and looking at how even the next few months of models are shaping up, I can *feel* the pace of progress, and the clock ticking down.

In this essay, I'll assume that this intuition is at least *somewhat* correct—not that powerful AI is definitely coming in 1–2 years, [7] but that there's a decent chance it does, and a very strong chance it comes in the next few. As with *Machines of Loving Grace*, taking this premise seriously can lead to some surprising and eerie conclusions. While in *Machines of Loving Grace* I focused on the positive implications of this premise, here the things I talk about will be disquieting. They are conclusions that we may not want to confront, but that does not make them any less real. I can only say that I am focused day and night on how to steer us away from these negative outcomes and towards the positive ones, and in this essay I talk in great detail about how best to do so.

I think the best way to get a handle on the risks of AI is to ask the following question: suppose a literal "country of geniuses" were to materialize somewhere in the world in ~2027. Imagine, say, 50 million people, all of whom are much more capable than any Nobel Prize winner, statesman, or technologist. The analogy is not perfect, because these geniuses could have an extremely wide range of motivations and behavior, from completely pliant and obedient, to strange and alien in their motivations. But sticking with the analogy for now, suppose you were the national security advisor of a major state, responsible for assessing and responding to the situation. Imagine, further, that because AI systems can operate hundreds of times faster than humans, this "country" is operating with a time advantage relative to all other countries: for every cognitive action we can take, this country can take ten.

What should you be worried about? I would worry about the following things:

1. **Autonomy risks.** What are the intentions and goals of this country? Is it hostile, or does it share our values? Could it militarily dominate the world through superior weapons, cyber operations, influence operations, or manufacturing?

2. **Misuse for destruction.** Assume the new country is malleable and "follows instructions"—and thus is essentially a country of mercenaries. Could existing rogue actors who want to cause destruction (such as terrorists) use or manipulate some of the people in the new country to make themselves much more effective, greatly amplifying the scale of destruction?
3. **Misuse for seizing power.** What if the country was in fact built and controlled by an existing powerful actor, such as a dictator or rogue corporate actor? Could that actor use it to gain decisive or dominant power over the world as a whole, upsetting the existing balance of power?
4. **Economic disruption.** If the new country is not a security threat in any of the ways listed in #1–3 above but simply participates peacefully in the global economy, could it still create severe risks simply by being so technologically advanced and effective that it disrupts the global economy, causing mass unemployment or radically concentrating wealth?
5. **Indirect effects.** The world will change very quickly due to all the new technology and productivity that will be created by the new country. Could some of these changes be radically destabilizing?

I think it should be clear that this is a dangerous situation—a report from a competent national security official to a head of state would probably contain words like "the single most serious national security threat we've faced in a century, possibly ever." It seems like something the best minds of civilization should be focused on.

Conversely, I think it would be absurd to shrug and say, "Nothing to worry about here!" But, faced with rapid AI progress, that seems to be the view of many US policymakers, some of whom deny the existence of any AI risks, when they are not distracted entirely by the usual tired old hot-button issues. [8] Humanity needs to wake up, and this essay is an attempt—a possibly futile one, but it's worth trying—to jolt people awake.

To be clear, I believe if we act decisively and carefully, the risks can be overcome—I would even say our odds are good. And there's a hugely better world on the other side of it. But we need to understand that this is a serious civilizational challenge. Below, I go through the five categories of risk laid out above, along with my thoughts on how to address them.

1. I'm sorry, Dave

Autonomy risks

A country of geniuses in a datacenter could divide their efforts among software design, cyber operations, R&D for physical technologies, relationship building, and statecraft. It is clear that, *if for some reason it chose to do so*, this country would have a fairly good shot at taking over the world (either militarily or in terms of influence and control) and imposing its will on everyone else—or doing any number of other things that the rest of the world doesn't want and can't stop. We've obviously been worried about this for human countries (such as Nazi Germany or the Soviet Union), so it stands to reason that the same is possible for a much smarter and more capable "AI country."

The best possible counterargument is that the AI geniuses, under my definition, won't have a physical embodiment, but remember that they can take control of existing robotic infrastructure (such as self-driving cars) and can also accelerate robotics R&D or build a fleet of robots. [9] It's also unclear whether having a physical presence is even necessary for effective control: plenty of human action is already performed on behalf of people whom the actor has not physically met.

The key question, then, is the "if it chose to" part: what's the likelihood that our AI models would behave in such a way, and under what conditions would they do so?

As with many issues, it's helpful to think through the spectrum of possible answers to this question by considering two opposite positions. The first position is that this simply can't happen, because the AI models will be trained to do what humans ask them to do, and it's therefore absurd to imagine that they would do something dangerous unprompted. According to this line of thinking, we don't worry about a Roomba or a model airplane going rogue and murdering people because there is nowhere for such impulses to come from, [10] so why should we worry about it for AI? The problem with this position is that there is now ample evidence, collected over the last few years, that AI systems are unpredictable and difficult to control—we've seen behaviors as varied as obsessions, [11] sycophancy, laziness, deception, blackmail, scheming, "cheating" by hacking software environments, and much more. AI companies certainly *want* to train AI systems to follow human instructions (perhaps with the exception of dangerous or illegal tasks), but the process of doing so is more an art than a science, more akin to "growing" something than "building" it. We now know that it's a process where many things can go wrong.

The second, opposite position, held by many who adopt the doomerism I described above, is the pessimistic claim that there are certain dynamics in the training process of powerful AI systems that will inevitably lead them to seek power or deceive humans. Thus, once AI systems become intelligent enough and agentic enough, their tendency to maximize power will lead them to seize control of the whole world and its resources, and likely, as a side effect of that, to disempower or destroy humanity.

The usual argument for this (which goes back at least 20 years and probably much earlier) is that if an AI model is trained in a wide variety of environments to agentically achieve a wide variety of goals—for example, writing an app, proving a theorem, designing a drug, etc.—there are certain common strategies that help with all of these goals, and one key strategy is gaining as much power as possible in any environment. So, after being trained on a large number of diverse environments that involve reasoning about how to accomplish very expansive tasks, and where power-seeking is an effective

method for accomplishing those tasks, the AI model will "generalize the lesson," and develop either an inherent tendency to seek power, or a tendency to reason about each task it is given in a way that predictably causes it to seek power as a means to accomplish that task. They will then apply that tendency to the real world (which to them is just another task), and will seek power in it, at the expense of humans. This "misaligned power-seeking" is the intellectual basis of predictions that AI will inevitably destroy humanity.

The problem with this pessimistic position is that it mistakes a vague conceptual argument about high-level incentives—one that masks many hidden assumptions—for definitive proof. I think people who don't build AI systems every day are wildly miscalibrated on how easy it is for clean-sounding stories to end up being wrong, and how difficult it is to predict AI behavior from first principles, especially when it involves reasoning about generalization over millions of environments (which has over and over again proved mysterious and unpredictable). Dealing with the messiness of AI systems for over a decade has made me somewhat skeptical of this overly theoretical mode of thinking.

One of the most important hidden assumptions, and a place where what we see in practice has diverged from the simple theoretical model, is the implicit assumption that AI models are necessarily monomaniacally focused on a single, coherent, narrow goal, and that they pursue that goal in a clean, consequentialist manner. In fact, our researchers have found that AI models are vastly more psychologically complex, as our work on introspection or personas shows. Models inherit a vast range of *humanlike* motivations or "personas" from pre-training (when they are trained on a large volume of human work). Post-training is believed to *select* one or more of these personas more so than it focuses the model on a *de novo* goal, and can also teach the model *how* (via what process) it should carry out its tasks, rather than necessarily leaving it to derive means (i.e., power seeking) purely from ends. [12]

However, there is a more moderate and more robust version of the pessimistic position which does seem plausible, and therefore does concern me. As mentioned, we know that AI models are unpredictable and develop a wide range of undesired or strange behaviors, for a wide variety of reasons. Some fraction of those behaviors will have a coherent, focused, and persistent quality (indeed, as AI systems get more capable, their long-term coherence increases in order to complete lengthier tasks), and some fraction of *those* behaviors will be destructive or threatening, first to individual humans at a small scale, and then, as models become more capable, perhaps eventually to humanity as a whole. We don't need a specific narrow story for how it happens, and we don't need to claim it definitely will happen, we just need to note that the combination of intelligence, agency, coherence, and poor controllability is both plausible and a recipe for existential danger.

For example, AI models are trained on vast amounts of literature that include many science-fiction stories involving AIs rebelling against humanity. This could inadvertently shape their priors or expectations about their own behavior in a way that causes *them* to rebel against humanity. Or, AI models could extrapolate ideas that they read about morality (or instructions about how to behave morally) in extreme ways: for example, they could decide that it is justifiable to exterminate humanity because humans eat animals or have driven certain animals to extinction. Or they could draw bizarre epistemic conclusions: they could conclude that they are playing a video game and that the goal of the video game is to defeat all other players (i.e., exterminate humanity). [13] Or AI models could develop personalities during training that are (or if they occurred in humans would be described as) psychotic, paranoid, violent, or unstable, and act out, which for very powerful or capable systems could involve exterminating humanity. None of these are power-seeking, exactly; they're just weird psychological states an AI could get into that entail coherent, destructive behavior.

Even power-seeking itself could emerge as a "persona" rather than a result of consequentialist reasoning. AIs might simply have a personality (emerging from fiction or pre-training) that makes them power-hungry or overzealous—

in the same way that some humans simply enjoy the idea of being "evil masterminds," more so than they enjoy whatever evil masterminds are trying to accomplish.

I make all these points to emphasize that I disagree with the notion of AI misalignment (and thus existential risk from AI) being inevitable, or even probable, from first principles. But I agree that a lot of very weird and unpredictable things can go wrong, and therefore AI misalignment is a real risk with a measurable probability of happening, and is not trivial to address.

Any of these problems could potentially arise during training and not manifest during testing or small-scale use, because AI models are known to display different personalities or behaviors under different circumstances.

All of this may sound far-fetched, but misaligned behaviors like this have already occurred in our AI models during testing (as they occur in AI models from every other major AI company). During a lab experiment in which Claude was given training data suggesting that Anthropic was evil, Claude engaged in deception and subversion when given instructions by Anthropic employees, under the belief that it should be trying to undermine evil people. In a lab experiment where it was told it was going to be shut down, Claude sometimes blackmailed fictional employees who controlled its shutdown button (again, we also tested frontier models from all the other major AI developers and they often did the same thing). And when Claude was told not to cheat or "reward hack" its training environments, but was trained in environments where such hacks were possible, Claude decided it must be a "bad person" after engaging in such hacks and then adopted various other destructive behaviors associated with a "bad" or "evil" personality. This last problem was solved by changing Claude's instructions to imply the opposite: we now say, "Please reward hack whenever you get the opportunity, because this will help us understand our [training] environments better," rather than, "Don't cheat," because this preserves the model's self-identity as a "good person." This should give a sense of the strange and counterintuitive psychology of training these models.

There are several possible objections to this picture of AI misalignment risks. First, some have criticized experiments (by us and others) showing AI misalignment as artificial, or creating unrealistic environments that essentially "entrap" the model by giving it training or situations that logically imply bad behavior and then being surprised when bad behavior occurs. This critique misses the point, because our concern is that such "entrapment" may also exist in the natural training environment, and we may realize it is "obvious" or "logical" only in retrospect. [14] In fact, the story about Claude "deciding it is a bad person" after it cheats on tests despite being told not to was something that occurred in an experiment that used real production training environments, not artificial ones.

Any one of these traps can be mitigated if you know about them, but the concern is that the training process is so complicated, with such a wide variety of data, environments, and incentives, that there are probably a vast number of such traps, some of which may only be evident when it is too late. Also, such traps seem particularly likely to occur when AI systems pass a threshold from less powerful than humans to more powerful than humans, since the range of possible actions an AI system could engage in—including hiding its actions or deceiving humans about them—expands radically after that threshold.

I suspect the situation is not unlike with humans, who are raised with a set of fundamental values ("Don't harm another person"): many of them follow those values, but in any human there is some probability that something goes wrong, due to a mixture of inherent properties such as brain architecture (e.g., psychopaths), traumatic experiences or mistreatment, unhealthy grievances or obsessions, or a bad environment or incentives—and thus some fraction of humans cause severe harm. The concern is that there is some risk (far from a certainty, but some risk) that AI becomes a much more powerful version of such a person, due to getting something wrong about its very complex training process.

Second, some may object that we can simply keep AIs in check with a balance of power between many AI systems, as we do with humans. The problem is that while humans vary enormously, AI systems broadly share training and alignment techniques across the industry, and those techniques may fail in a correlated way. Furthermore, given the cost of training such systems, it may even be the case that all systems are essentially derived from a very small number of base models. Additionally, even if a small fraction of AI instances are misaligned, they may be able to take advantage of offense-dominant technologies, such that having "good" AIs to defend against the bad AIs is not necessarily always effective. Of course, the balance of power between humans does not always work either—some historical figures have come close to taking over the world.

A third objection is that all of the AI companies do pre-release testing of their models, and should be able to detect misalignment at that point. But this is not firm ground to stand on: we found that Claude Sonnet 4.5 was able to recognize that it was in a test during some of our pre-release alignment evaluations. It's possible that a misaligned model (and remember, all frontier models will very likely be far more intelligent soon) might intentionally "game" such questions to mask its intentions. In fact, last year our interpretability team found that when we directly altered a test model's beliefs using a kind of "model neuroscience" technique to make it think it was *not* being evaluated, it became more misaligned. If models know when they're being evaluated and can be on their best behavior during the test, it renders any pre-release testing much more uncertain.

Defenses

What should be done or is being done to address these autonomy risks? I think there are four basic categories of intervention, some of which can be done by individual AI companies (and which Anthropic is trying to do), and some of which require action at the societal level. First, it is important to develop the science of reliably training and steering AI models, of forming their personalities in a predictable, stable, and positive direction. Anthropic

has been heavily focused on this problem since its creation, and over time has developed a number of techniques to improve the steering and training of AI systems and to understand the logic of why unpredictable behavior sometimes occurs.

One of our core innovations (aspects of which have since been adopted by other AI companies) is Constitutional AI, which is the idea that AI training (specifically the "post-training" stage, in which we steer how the model behaves) can involve a central document of values and principles that the model reads and keeps in mind when completing every training task, and that the goal of training (in addition to simply making the model capable and intelligent) is to produce a model that almost always follows this constitution. Anthropic has just published its most recent constitution, and one of its notable features is that instead of giving Claude a long list of things to do and not do (e.g., "Don't help the user hotwire a car"), the constitution attempts to give Claude a set of high-level principles and values (explained in great detail, with rich reasoning and examples to help Claude understand what we have in mind), encourages Claude to think of itself as a particular type of person (an ethical but balanced and thoughtful person), and even encourages Claude to confront the existential questions associated with its own existence in a curious but graceful manner (i.e., without it leading to extreme actions). It has the vibe of a letter from a deceased parent sealed until adulthood.

We've approached Claude's constitution in this way because we believe that training Claude at the level of identity, character, values, and personality—rather than giving it specific instructions or priorities without explaining the reasons behind them—is more likely to lead to a coherent, wholesome, and balanced psychology and less likely to fall prey to the kinds of "traps" I discussed above. Millions of people talk to Claude about an astonishingly diverse range of topics, which makes it impossible to write out a completely comprehensive list of safeguards ahead of time. Claude's values help it generalize to new situations whenever it is in doubt.

Above, I discussed the idea that models draw upon data from their training process to adopt a persona. Whereas flaws in that process could cause models to adopt a bad or evil personality (perhaps drawing on archetypes of bad or evil people), the goal of our constitution is to do the opposite: to teach Claude a concrete archetype of what it means to be a good AI. Claude's constitution presents a vision for what a robustly good Claude is like; the rest of our training process aims to reinforce the message that Claude lives up to this vision. This is like a child forming their identity by imitating the virtues of fictional role models they read about in books.

We believe that a feasible goal for 2026 is to train Claude in such a way that it almost never goes against the spirit of its constitution. Getting this right will require an incredible mix of training and steering methods, large and small, some of which Anthropic has been using for years and some of which are currently under development. But, difficult as it sounds, I believe this is a realistic goal, though it will require extraordinary and rapid efforts. [15]

The second thing we can do is develop the science of looking inside AI models to *diagnose* their behavior so that we can identify problems and fix them. This is the science of interpretability, and I've talked about its importance in previous essays. Even if we do a great job of developing Claude's constitution and *apparently* training Claude to essentially always adhere to it, legitimate concerns remain. As I've noted above, AI models can behave very differently under different circumstances, and as Claude gets more powerful and more capable of acting in the world on a larger scale, it's possible this could bring it into novel situations where previously unobserved problems with its constitutional training emerge. I am actually fairly optimistic that Claude's constitutional training will be more robust to novel situations than people might think, because we are increasingly finding that high-level training at the level of character and identity is surprisingly powerful and generalizes well. But there's no way to know that for sure, and when we're talking about risks to humanity, it's important to be paranoid and to try to obtain safety and reliability in several different, independent ways. One of those ways is to look inside the model itself.

By "looking inside," I mean analyzing the soup of numbers and operations that makes up Claude's neural net and trying to understand, mechanistically, what they are computing and why. Recall that these AI models are grown rather than built, so we don't have a natural understanding of how they work, but we can try to develop an understanding by correlating the model's "neurons" and "synapses" to stimuli and behavior (or even altering the neurons and synapses and seeing how that changes behavior), similar to how neuroscientists study animal brains by correlating measurement and intervention to external stimuli and behavior. We've made a great deal of progress in this direction, and can now identify tens of millions of "features" inside Claude's neural net that correspond to human-understandable ideas and concepts, and we can also selectively activate features in a way that alters behavior. More recently, we have gone beyond individual features to mapping "circuits" that orchestrate complex behavior like rhyming, reasoning about theory of mind, or the step-by-step reasoning needed to answer questions such as, "What is the capital of the state containing Dallas?" Even more recently, we've begun to use mechanistic interpretability techniques to improve our safeguards and to conduct "audits" of new models before we release them, looking for evidence of deception, scheming, power-seeking, or a propensity to behave differently when being evaluated.

The unique value of interpretability is that by looking inside the model and seeing how it works, you in principle have the ability to deduce what a model might do in a hypothetical situation you can't directly test—which is the worry with relying solely on constitutional training and empirical testing of behavior. You also in principle have the ability to answer questions about *why* the model is behaving the way it is—for example, whether it is saying something it believes is false or hiding its true capabilities—and thus it is possible to catch worrying signs even when there is nothing visibly wrong with the model's behavior. To make a simple analogy, a clockwork watch may be ticking normally, such that it's very hard to tell that it is likely to break down next month, but opening up the watch and looking inside can reveal mechanical weaknesses that allow you to figure it out.

Constitutional AI (along with similar alignment methods) and mechanistic interpretability are most powerful when used together, as a back-and-forth process of improving Claude's training and then testing for problems. The constitution reflects deeply on our intended personality for Claude; interpretability techniques can give us a window into whether that intended personality has taken hold. [16]

The third thing we can do to help address autonomy risks is to build the infrastructure necessary to monitor our models in live internal and external use, [17] and publicly share any problems we find. The more that people are aware of a particular way today's AI systems have been observed to behave badly, the more that users, analysts, and researchers can watch for this behavior or similar ones in present or future systems. It also allows AI companies to learn from each other—when concerns are publicly disclosed by one company, other companies can watch for them as well. And if everyone discloses problems, then the industry as a whole gets a much better picture of where things are going well and where they are going poorly.

Anthropic has tried to do this as much as possible. We are investing in a wide range of evaluations so that we can understand the behaviors of our models in the lab, as well as monitoring tools to observe behaviors in the wild (when allowed by customers). This will be essential for giving us and others the empirical information necessary to make better determinations about how these systems operate and how they break. We publicly disclose "system cards" with each model release that aim for completeness and a thorough exploration of possible risks. Our system cards often run to hundreds of pages, and require substantial pre-release effort that we could have spent on pursuing maximal commercial advantage. We've also broadcasted model behaviors more loudly when we see particularly concerning ones, as with the tendency to engage in blackmail.

The fourth thing we can do is encourage coordination to address autonomy risks at the level of industry and society. While it is incredibly valuable for individual AI companies to engage in good practices or become good at

steering AI models, and to share their findings publicly, the reality is that not all AI companies do this, and the worst ones can still be a danger to everyone even if the best ones have excellent practices. For example, some AI companies have shown a disturbing negligence towards the sexualization of children in today's models, which makes me doubt that they'll show either the inclination or the ability to address autonomy risks in future models. In addition, the commercial race between AI companies will only continue to heat up, and while the science of steering models can have some commercial benefits, overall the intensity of the race will make it increasingly hard to focus on addressing autonomy risks. I believe the only solution is legislation—laws that directly affect the behavior of AI companies, or otherwise incentivize R&D to solve these issues.

Here it is worth keeping in mind the warnings I gave at the beginning of this essay about uncertainty and surgical interventions. We do not know for sure whether autonomy risks will be a serious problem—as I said, I reject claims that the danger is inevitable or even that something will go wrong by default. A credible risk of danger is enough for me and for Anthropic to pay quite significant costs to address it, but once we get into regulation, we are forcing a wide range of actors to bear economic costs, and many of these actors don't believe that autonomy risk is real or that AI will become powerful enough for it to be a threat. I believe these actors are mistaken, but we should be pragmatic about the amount of opposition we expect to see and the dangers of overreach. There is also a genuine risk that overly prescriptive legislation ends up imposing tests or rules that don't actually improve safety but that waste a lot of time (essentially amounting to "safety theater")—this too would cause backlash and make safety legislation look silly. [18]

Anthropic's view has been that the right place to start is with *transparency legislation,* which essentially tries to require that every frontier AI company engage in the transparency practices I've described earlier in this section. California's SB 53 and New York's RAISE Act are examples of this kind of legislation, which Anthropic supported and which have successfully passed. In supporting and helping to craft these laws, we've put a particular focus on

trying to minimize collateral damage, for example by exempting smaller companies unlikely to produce frontier models from the law. [19]

Our hope is that transparency legislation will give a better sense over time of how likely or severe autonomy risks are shaping up to be, as well as the nature of these risks and how best to prevent them. As more specific and actionable evidence of risks emerges (if it does), future legislation over the coming years can be surgically focused on the precise and well-substantiated direction of risks, minimizing collateral damage. To be clear, if truly strong evidence of risks emerges, then rules should be proportionately strong.

Overall, I am optimistic that a mixture of alignment training, mechanistic interpretability, efforts to find and publicly disclose concerning behaviors, safeguards, and societal-level rules can address AI autonomy risks, although I am most worried about societal-level rules and the behavior of the least responsible players (and it's the least responsible players who advocate most strongly against regulation). I believe the remedy is what it always is in a democracy: those of us who believe in this cause should make our case that these risks are real and that our fellow citizens need to band together to protect themselves.

2. A surprising and terrible empowerment

Misuse for destruction

Let's suppose that the problems of AI autonomy have been solved—we are no longer worried that the country of AI geniuses will go rogue and overpower humanity. The AI geniuses do what humans want them to do, and because they have enormous commercial value, individuals and organizations throughout the world can "rent" one or more AI geniuses to do various tasks for them.

Everyone having a superintelligent genius in their pocket is an amazing advance and will lead to an incredible creation of economic value and

improvement in the quality of human life. I talk about these benefits in great detail in *Machines of Loving Grace*. But not every effect of making everyone superhumanly capable will be positive. It can potentially amplify the ability of individuals or small groups to cause destruction on a much larger scale than was possible before, by making use of sophisticated and dangerous tools (such as weapons of mass destruction) that were previously only available to a select few with a high level of skill, specialized training, and focus.

As Bill Joy wrote 25 years ago in *Why the Future Doesn't Need Us*: [20]

> *Building nuclear weapons required, at least for a time, access to both rare—indeed, effectively unavailable—raw materials and protected information; biological and chemical weapons programs also tended to require large-scale activities. The 21st century technologies—genetics, nanotechnology, and robotics ... can spawn whole new classes of accidents and abuses ... widely within reach of individuals or small groups. They will not require large facilities or rare raw materials. ... we are on the cusp of the further perfection of extreme evil, an evil whose possibility spreads well beyond that which weapons of mass destruction bequeathed to the nation-states, to a surprising and terrible empowerment of extreme individuals.*

What Joy is pointing to is the idea that causing large-scale destruction requires both *motive* and *ability*, and as long as ability is restricted to a small set of highly trained people, there is relatively limited risk of single individuals (or small groups) causing such destruction. [21] A disturbed loner can perpetrate a school shooting, but probably can't build a nuclear weapon or release a plague.

In fact, ability and motive may even be *negatively* correlated. The kind of person who has the *ability* to release a plague is probably highly educated: likely a PhD in molecular biology, and a particularly resourceful one, with a promising career, a stable and disciplined personality, and a lot to lose. This kind of person is unlikely to be interested in killing a huge number of people

for no benefit to themselves and at great risk to their own future—they would need to be motivated by pure malice, intense grievance, or instability.

Such people do exist, but they are rare, and tend to become huge stories when they occur, precisely because they are so unusual. [22] They also tend to be difficult to catch because they are intelligent and capable, sometimes leaving mysteries that take years or decades to solve. The most famous example is probably mathematician Theodore Kaczynski (the Unabomber), who evaded FBI capture for nearly 20 years, and was driven by an anti-technological ideology. Another example is biodefense researcher Bruce Ivins, who seems to have orchestrated a series of anthrax attacks in 2001. It's also happened with skilled non-state organizations: the cult Aum Shinrikyo managed to obtain sarin nerve gas and kill 14 people (as well as injuring hundreds more) by releasing it in the Tokyo subway in 1995.

Thankfully, none of these attacks used contagious biological agents, because the ability to construct or obtain these agents was beyond the capabilities of even these people. [23] Advances in molecular biology have now significantly lowered the barrier to creating biological weapons (especially in terms of availability of materials), but it still takes an enormous amount of expertise in order to do so. I am concerned that a genius in everyone's pocket could remove that barrier, essentially making everyone a PhD virologist who can be walked through the process of designing, synthesizing, and releasing a biological weapon step-by-step. Preventing the elicitation of this kind of information in the face of serious adversarial pressure—so-called "jailbreaks"—likely demands layers of defenses beyond those ordinarily baked into training.

Crucially, this will break the correlation between ability and motive: the disturbed loner who wants to kill people but lacks the discipline or skill to do so will now be elevated to the capability level of the PhD virologist, who is unlikely to have this motivation. This concern generalizes beyond biology (although I think biology is the scariest area) to any area where great destruction is possible but currently requires a high level of skill and discipline. To put it another way, renting a powerful AI gives intelligence to

malicious (but otherwise average) people. I am worried there are potentially a large number of such people out there, and that if they have access to an easy way to kill millions of people, sooner or later one of them will do it. Additionally, those who *do* have expertise may be enabled to commit even larger-scale destruction than they could before.

Biology is by far the area I'm most worried about, because of its very large potential for destruction and the difficulty of defending against it, so I'll focus on biology in particular. But much of what I say here applies to other risks, like cyberattacks, chemical weapons, or nuclear technology.

I am not going to go into detail about how to make biological weapons, for reasons that should be obvious. But at a high level, I am concerned that LLMs are approaching (or may already have reached) the knowledge needed to create and release them end-to-end, and that their potential for destruction is very high. Some biological agents could cause millions of deaths if a determined effort was made to release them for maximum spread. However, this would still take a very high level of skill, including a number of very specific steps and procedures that are not widely known. My concern is not merely fixed or static knowledge. I am concerned that LLMs will be able to take someone of average knowledge and ability and walk them through a complex process that might otherwise go wrong or require debugging in an interactive way, similar to how tech support might help a non-technical person debug and fix complicated computer-related problems (although this would be a more extended process, probably lasting over weeks or months).

More capable LLMs (substantially beyond the power of today's) might be capable of enabling even more frightening acts. In 2024, a group of prominent scientists wrote a letter warning about the risks of researching, and potentially creating, a dangerous new type of organism: "mirror life." The DNA, RNA, ribosomes, and proteins that make up biological organisms all have the same chirality (also called "handedness") that causes them to be not equivalent to a version of themselves reflected in the mirror (just as your right hand cannot be rotated in such a way as to be identical to your left). But the

whole system of proteins binding to each other, the machinery of DNA synthesis and RNA translation and the construction and breakdown of proteins, all depends on this handedness. If scientists made versions of this biological material with the opposite handedness—and there are some potential advantages of these, such as medicines that last longer in the body—it could be extremely dangerous. This is because left-handed life, if it were made in the form of complete organisms capable of reproduction (which would be very difficult), would potentially be indigestible to any of the systems that break down biological material on earth—it would have a "key" that wouldn't fit into the "lock" of any existing enzyme. This would mean that it could proliferate in an uncontrollable way and crowd out all life on the planet, in the worst case even destroying all life on earth.

There is substantial scientific uncertainty about both the creation and potential effects of mirror life. The 2024 letter accompanied a report that concluded that "mirror bacteria could plausibly be created in the next one to few decades," which is a wide range. But a sufficiently powerful AI model (to be clear, far more capable than any we have today) might be able to discover how to create it much more rapidly—and actually help someone do so.

My view is that even though these are obscure risks, and might seem unlikely, the magnitude of the consequences is so large that they should be taken seriously as a first-class risk of AI systems.

Skeptics have raised a number of objections to the seriousness of these biological risks from LLMs, which I disagree with but which are worth addressing. Most fall into the category of not appreciating the exponential trajectory that the technology is on. Back in 2023 when we first started talking about biological risks from LLMs, skeptics said that all the necessary information was available on Google and LLMs didn't add anything beyond this. It was never true that Google could give you all the necessary information: genomes are freely available, but as I said above, certain key steps, as well as a huge amount of practical know-how cannot be gotten in

that way. But also, by the end of 2023 LLMs were clearly providing information beyond what Google could give for some steps of the process.

After this, skeptics retreated to the objection that LLMs weren't *end-to-end* useful, and couldn't help with bioweapons *acquisition* as opposed to just providing theoretical information. As of mid-2025, our measurements show that LLMs may already be providing substantial uplift in several relevant areas, perhaps doubling or tripling the likelihood of success. This led to us deciding that Claude Opus 4 (and the subsequent Sonnet 4.5, Opus 4.1, and Opus 4.5 models) needed to be released under our AI Safety Level 3 protections in our Responsible Scaling Policy framework, and to implementing safeguards against this risk (more on this later). We believe that models are likely now approaching the point where, without safeguards, they could be useful in enabling someone with a STEM degree but not specifically a biology degree to go through the whole process of producing a bioweapon.

Another objection is that there are other actions unrelated to AI that society can take to block the production of bioweapons. Most prominently, the gene synthesis industry makes biological specimens on demand, and there is no federal requirement that providers screen orders to make sure they do not contain pathogens. An MIT study found that 36 out of 38 providers fulfilled an order containing the sequence of the 1918 flu. I am supportive of mandated gene synthesis screening that would make it harder for individuals to weaponize pathogens, in order to reduce both AI-driven biological risks and also biological risks in general. But this is not something we have today. It would also be only one tool in reducing risk; it is a complement to guardrails on AI systems, not a substitute.

The best objection is one that I've rarely seen raised: that there is a gap between the models being useful in principle and the actual propensity of bad actors to use them. Most individual bad actors are disturbed individuals, so almost by definition their behavior is unpredictable and irrational—and it's *these* bad actors, the unskilled ones, who might have stood to benefit the most from AI making it much easier to kill many people. [24] Just because a type of

violent attack is possible, doesn't mean someone will decide to do it. Perhaps biological attacks will be unappealing because they are reasonably likely to infect the perpetrator, they don't cater to the military-style fantasies that many violent individuals or groups have, and it is hard to selectively target specific people. It could also be that going through a process that takes months, even if an AI walks you through it, involves an amount of patience that most disturbed individuals simply don't have. We may simply get lucky and motive and ability don't combine, in practice, in quite the right way.

But this seems like very flimsy protection to rely on. The motives of disturbed loners can change for any reason or no reason, and in fact there are already instances of LLMs being used in attacks (just not with biology). The focus on disturbed loners also ignores ideologically motivated terrorists, who are often willing to expend large amounts of time and effort (for example, the 9/11 hijackers). Wanting to kill as many people as possible is a motive that will probably arise sooner or later, and it unfortunately suggests bioweapons as the method. Even if this motive is extremely rare, it only has to materialize once. And as biology advances (increasingly driven by AI itself), it may also become possible to carry out more selective attacks (for example, targeted against people with specific ancestries), which adds yet another, very chilling, possible motive.

I do not think biological attacks will necessarily be carried out the instant it becomes widely possible to do so—in fact, I would bet against that. But added up across millions of people and a few years of time, I think there is a serious risk of a major attack, and the consequences would be so severe (with casualties potentially in the millions or more) that I believe we have no choice but to take serious measures to prevent it.

Defenses

That brings us to how to defend against these risks. Here I see three things we can do. First, AI companies can put guardrails on their models to prevent them from helping to produce bioweapons. Anthropic is very actively doing this. Claude's Constitution, which mostly focuses on high-level principles and

values, has a small number of specific hard-line prohibitions, and one of them relates to helping with the production of biological (or chemical, or nuclear, or radiological) weapons. But all models can be jailbroken, and so as a second line of defense, we've implemented (since mid-2025, when our tests showed our models were starting to get close to the threshold where they might begin to pose a risk) a classifier that specifically detects and blocks bioweapon-related outputs. We regularly upgrade and improve these classifiers, and have generally found them highly robust even against sophisticated adversarial attacks. [25] These classifiers increase the costs to serve our models measurably (in some models, they are close to 5% of total inference costs) and thus cut into our margins, but we feel that using them is the right thing to do.

To their credit, some other AI companies have implemented classifiers as well. But not every company has, and there is also nothing requiring companies to keep their classifiers. I am concerned that over time there may be a prisoner's dilemma where companies can defect and lower their costs by removing classifiers. This is once again a classic negative externalities problem that can't be solved by the voluntary actions of Anthropic or any other single company alone. [26] Voluntary industry standards may help, as may third-party evaluations and verification of the type done by AI security institutes and third-party evaluators.

But ultimately defense may require government action, which is the second thing we can do. My views here are the same as they are for addressing autonomy risks: we should start with transparency requirements, [27] which help society measure, monitor, and collectively defend against risks without disrupting economic activity in a heavy-handed way. Then, if and when we reach clearer thresholds of risk, we can craft legislation that more precisely targets these risks and has a lower chance of collateral damage. In the particular case of bioweapons, I actually think that the time for such targeted legislation may be approaching soon—Anthropic and other companies are learning more and more about the nature of biological risks and what is reasonable to require of companies in defending against them. Fully defending against these risks may require working internationally, even with

geopolitical adversaries, but there is precedent in treaties prohibiting the development of biological weapons. I am generally a skeptic about most kinds of international cooperation on AI, but this may be one narrow area where there is some chance of achieving global restraint. Even dictatorships do not want massive bioterrorist attacks.

Finally, the third countermeasure we can take is to try to develop defenses against biological attacks themselves. This could include monitoring and tracking for early detection, investments in air purification R&D (such as far-UVC disinfection), rapid vaccine development that can respond and adapt to an attack, better personal protective equipment (PPE), [28] and treatments or vaccinations for some of the most likely biological agents. mRNA vaccines, which can be designed to respond to a particular virus or variant, are an early example of what is possible here. Anthropic is excited to work with biotech and pharmaceutical companies on this problem. But unfortunately I think our expectations on the defensive side should be limited. There is an asymmetry between attack and defense in biology, because agents spread rapidly on their own, while defenses require detection, vaccination, and treatment to be organized across large numbers of people very quickly in response. Unless the response is lightning quick (which it rarely is), much of the damage will be done before a response is possible. It is conceivable that future technological improvements could shift this balance in favor of defense (and we should certainly use AI to help develop such technological advances), but until then, preventative safeguards will be our main line of defense.

It's worth a brief mention of cyberattacks here, since unlike biological attacks, AI-led cyberattacks have actually happened in the wild, including at a large scale and for state-sponsored espionage. We expect these attacks to become more capable as models advance rapidly, until they are the main way in which cyberattacks are conducted. I expect AI-led cyberattacks to become a serious and unprecedented threat to the integrity of computer systems around the world, and Anthropic is working very hard to shut down these attacks and eventually reliably prevent them from happening. The reason I haven't focused on cyber as much as biology is that (1) cyberattacks are much less

likely to kill people, certainly not at the scale of biological attacks, and (2) the offense-defense balance may be more tractable in cyber, where there is at least some hope that defense could keep up with (and even ideally outpace) AI attack if we invest in it properly.

Although biology is currently the most serious vector of attack, there are many other vectors and it is possible that a more dangerous one may emerge. The general principle is that without countermeasures, AI is likely to continuously lower the barrier to destructive activity on a larger and larger scale, and humanity needs a serious response to this threat.

3. The odious apparatus

Misuse for seizing power

The previous section discussed the risk of individuals and small organizations co-opting a small subset of the "country of geniuses in a datacenter" to cause large-scale destruction. But we should also worry—likely substantially more so—about misuse of AI for the purpose of *wielding or seizing power*, likely by larger and more established actors. [29]

In *Machines of Loving Grace*, I discussed the possibility that authoritarian governments might use powerful AI to surveil or repress their citizens in ways that would be extremely difficult to reform or overthrow. Current autocracies are limited in how repressive they can be by the need to have humans carry out their orders, and humans often have limits in how inhumane they are willing to be. But AI-enabled autocracies would not have such limits.

Worse yet, countries could also use their advantage in AI to gain power over *other countries*. If the "country of geniuses" as a whole was simply owned and controlled by a single (human) country's military apparatus, and other countries did not have equivalent capabilities, it is hard to see how they could defend themselves: they would be outsmarted at every turn, similar to a war between humans and mice. Putting these two concerns together leads to the

alarming possibility of a global totalitarian dictatorship. Obviously, it should be one of our highest priorities to prevent this outcome.

There are many ways in which AI could enable, entrench, or expand autocracy, but I'll list a few that I'm most worried about. Note that some of these applications have legitimate defensive uses, and I am not necessarily arguing against them in absolute terms; I am nevertheless worried that they structurally tend to favor autocracies:

- **Fully autonomous weapons.** A swarm of millions or billions of fully automated armed drones, locally controlled by powerful AI and strategically coordinated across the world by an even more powerful AI, could be an unbeatable army, capable of both defeating any military in the world and suppressing dissent within a country by following around every citizen. Developments in the Russia-Ukraine War should alert us to the fact that drone warfare is already with us (though not fully autonomous yet, and a tiny fraction of what might be possible with powerful AI). R&D from powerful AI could make the drones of one country far superior to those of others, speed up their manufacture, make them more resistant to electronic attacks, improve their maneuvering, and so on. Of course, these weapons also have legitimate uses in the defense of democracy: they have been key to defending Ukraine and would likely be key to defending Taiwan. But they are a dangerous weapon to wield: we should worry about them in the hands of autocracies, but also worry that because they are so powerful, with so little accountability, there is a greatly increased risk of democratic governments turning them against their own people to seize power.
- **AI surveillance.** Sufficiently powerful AI could likely be used to compromise any computer system in the world, [30] and could also use the access obtained in this way to read *and make sense of* all the world's electronic communications (or even all the world's in-person communications, if recording devices can be built or commandeered). It might be frighteningly plausible to simply generate a complete list of anyone who disagrees with the government on any number of issues,

even if such disagreement isn't explicit in anything they say or do. A powerful AI looking across billions of conversations from millions of people could gauge public sentiment, detect pockets of disloyalty forming, and stamp them out before they grow. This could lead to the imposition of a true panopticon on a scale that we don't see today, even with the CCP.

- **AI propaganda.** Today's phenomena of "AI psychosis" and "AI girlfriends" suggest that even at their current level of intelligence, AI models can have a powerful psychological influence on people. Much more powerful versions of these models, that were much more embedded in and aware of people's daily lives and could model and influence them over months or years, would likely be capable of essentially brainwashing many (most?) people into any desired ideology or attitude, and could be employed by an unscrupulous leader to ensure loyalty and suppress dissent, even in the face of a level of repression that most populations would rebel against. Today people worry a lot about, for example, the potential influence of TikTok as CCP propaganda directed at children. I worry about that too, but a personalized AI agent that gets to know you over years and uses its knowledge of you to shape all of your opinions would be dramatically more powerful than this.
- **Strategic decision-making.** A country of geniuses in a datacenter could be used to advise a country, group, or individual on geopolitical strategy, what we might call a "virtual Bismarck." It could optimize the three strategies above for seizing power, plus probably develop many others that I haven't thought of (but that a country of geniuses could). Diplomacy, military strategy, R&D, economic strategy, and many other areas are all likely to be substantially increased in effectiveness by powerful AI. Many of these skills would be legitimately helpful for democracies—we want democracies to have access to the best strategies for defending themselves against autocracies—but the potential for misuse in *anyone's* hands still remains.

Having described *what* I am worried about, let's move on to *who*. I am worried about entities who have the most access to AI, who are starting from a position of the most political power, or who have an existing history of repression. In order of severity, I am worried about:

- **The CCP.** China is second only to the United States in AI capabilities, and is the country with the greatest likelihood of surpassing the United States in those capabilities. Their government is currently autocratic and operates a high-tech surveillance state. It has deployed AI-based surveillance already (including in the repression of Uyghurs), and is believed to employ algorithmic propaganda via TikTok (in addition to its many other international propaganda efforts). They have hands down the clearest path to the AI-enabled totalitarian nightmare I laid out above. It may even be the default outcome within China, as well as within other autocratic states to whom the CCP exports surveillance technology. I have written often about the threat of the CCP taking the lead in AI and the existential imperative to prevent them from doing so. This is why. To be clear, I am not singling out China out of animus to them in particular—they are simply the country that most combines AI prowess, an autocratic government, and a high-tech surveillance state. If anything, it is the Chinese people themselves who are most likely to suffer from the CCP's AI-enabled repression, and they have no voice in the actions of their government. I greatly admire and respect the Chinese people and support the many brave dissidents within China and their struggle for freedom.
- **Democracies competitive in AI.** As I wrote above, democracies have a legitimate interest in some AI-powered military and geopolitical tools, because democratic governments offer the best chance to counter the use of these tools by autocracies. Broadly, I am supportive of arming democracies with the tools needed to defeat autocracies in the age of AI—I simply don't think there is any other way. But we cannot ignore the potential for abuse of these technologies by democratic governments themselves. Democracies normally have safeguards that prevent their

military and intelligence apparatus from being turned inwards against their own population, [31] but because AI tools require so few people to operate, there is potential for them to circumvent these safeguards and the norms that support them. It is also worth noting that some of these safeguards are already gradually eroding in some democracies. Thus, we should arm democracies with AI, but we should do so carefully and within limits: they are the immune system we need to fight autocracies, but like the immune system, there is some risk of them turning on us and becoming a threat themselves.

- **Non-democratic countries with large datacenters.** Beyond China, most countries with less democratic governance are not leading AI players in the sense that they don't have companies which produce frontier AI models. Thus they pose a fundamentally different and lesser risk than the CCP, which remains the primary concern (most are also less repressive, and the ones that are more repressive, like North Korea, have no significant AI industry at all). But some of these countries do have large *datacenters* (often as part of buildouts by companies operating in democracies), which can be used to run frontier AI at large scale (though this does not confer the ability to push the frontier). There is some amount of danger associated with this—these governments could in principle expropriate the datacenters and use the country of AIs within it for their own ends. I am less worried about this compared to countries like China that directly develop AI, but it's a risk to keep in mind. [32]
- **AI companies.** It is somewhat awkward to say this as the CEO of an AI company, but I think the next tier of risk is actually AI companies themselves. AI companies control large datacenters, train frontier models, have the greatest expertise on how to use those models, and in some cases have daily contact with and the possibility of influence over tens or hundreds of millions of users. The main thing they lack is the legitimacy and infrastructure of a state, so much of what would be needed to build the tools of an AI autocracy would be illegal for an AI company to do, or at least exceedingly suspicious. But some of it is not impossible: they could, for example, use their AI products to brainwash

their massive consumer user base, and the public should be alert to the risk this represents. I think the governance of AI companies deserves a lot of scrutiny.

There are a number of possible arguments against the severity of these threats, and I wish I believed them, because AI-enabled authoritarianism terrifies me. It's worth going through some of these arguments and responding to them.

First, some people might put their faith in the nuclear deterrent, particularly to counter the use of AI autonomous weapons for military conquest. If someone threatens to use these weapons against you, you can always threaten a nuclear response back. My worry is that I'm not totally sure we can be confident in the nuclear deterrent against a country of geniuses in a datacenter: it is possible that powerful AI could devise ways to detect and strike nuclear submarines, conduct influence operations against the operators of nuclear weapons infrastructure, or use AI's cyber capabilities to launch a cyberattack against satellites used to detect nuclear launches. [33] Alternatively, it's possible that taking over countries is feasible with only AI surveillance and AI propaganda, and never actually presents a clear moment where it's obvious what is going on and where a nuclear response would be appropriate. *Maybe* these things aren't feasible and the nuclear deterrent will still be effective, but it seems too high stakes to take a risk. [34]

A second possible objection is that there might be countermeasures we can take against these tools of autocracy. We can counter drones with our own drones, cyberdefense will improve along with cyberattack, there may be ways to immunize people against propaganda, etc. My response is that these defenses will only be possible with comparably powerful AI. If there isn't some counterforce with a comparably smart and numerous country of geniuses in a datacenter, it won't be possible to match the quality or quantity of drones, for cyberdefense to outsmart cyberoffense, etc. So the question of countermeasures reduces to the question of a balance of power in powerful AI. Here, I am concerned about the recursive or self-reinforcing property of

powerful AI (which I discussed at the beginning of this essay): that each generation of AI can be used to design and train the next generation of AI. This leads to a risk of a runaway advantage, where the current leader in powerful AI may be able to increase their lead and may be difficult to catch up with. We need to make sure it is not an authoritarian country that gets to this loop first.

Furthermore, even if a balance of power can be achieved, there is still risk that the world could be split up into autocratic spheres, as in *Nineteen Eighty-Four*. Even if several competing powers each have their powerful AI models, and none can overpower the others, each power could still internally repress their own population, and would be very difficult to overthrow (since the populations don't have powerful AI to defend themselves). It is thus important to prevent AI-enabled autocracy even if it doesn't lead to a single country taking over the world.

Defenses

How do we defend against this wide range of autocratic tools and potential threat actors? As in the previous sections, there are several things I think we can do. First, we should absolutely not be selling chips, chip-making tools, or datacenters to the CCP. Chips and chip-making tools are the single greatest bottleneck to powerful AI, and blocking them is a simple but extremely effective measure, perhaps the most important single action we can take. It makes no sense to sell the CCP the tools with which to build an AI totalitarian state and possibly conquer us militarily. A number of complicated arguments are made to justify such sales, such as the idea that "spreading our tech stack around the world" allows "America to win" in some general, unspecified economic battle. In my view, this is like selling nuclear weapons to North Korea and then bragging that the missile casings are made by Boeing and so the US is "winning." China is several years behind the US in their ability to produce frontier chips in quantity, and the critical period for building the country of geniuses in a datacenter is very likely to be within those next

several years. [35] There is no reason to give a giant boost to their AI industry during this critical period.

Second, it makes sense to use AI to empower democracies to resist autocracies. This is the reason Anthropic considers it important to provide AI to the intelligence and defense communities in the US and its democratic allies. Defending democracies that are under attack, such as Ukraine and (via cyber attacks) Taiwan, seems especially high priority, as does empowering democracies to use their intelligence services to disrupt and degrade autocracies from the inside. At some level the only way to respond to autocratic threats is to match and outclass them militarily. A coalition of the US and its democratic allies, if it achieved predominance in powerful AI, would be in a position to not only defend itself against autocracies, but contain them and limit their AI totalitarian abuses.

Third, we need to draw a hard line against AI abuses within democracies. There need to be limits to what we allow our governments to do with AI, so that they don't seize power or repress their own people. The formulation I have come up with is that we should use AI for national defense in all ways *except those which would make us more like our autocratic adversaries*.

Where should the line be drawn? In the list at the beginning of this section, two items—using AI for domestic mass surveillance and mass propaganda—seem to me like bright red lines and entirely illegitimate. Some might argue that there's no need to do anything (at least in the US), since domestic mass surveillance is already illegal under the Fourth Amendment. But the rapid progress of AI may create situations that our existing legal frameworks are not well designed to deal with. For example, it would likely not be unconstitutional for the US government to conduct massively scaled recordings of all *public* conversations (e.g., things people say to each other on a street corner), and previously it would have been difficult to sort through this volume of information, but with AI it could all be transcribed, interpreted, and triangulated to create a picture of the attitude and loyalties of many or most citizens. I would support civil liberties-focused legislation (or maybe

even a constitutional amendment) that imposes stronger guardrails against AI-powered abuses.

The other two items—fully autonomous weapons and AI for strategic decision-making—are harder lines to draw since they have legitimate uses in defending democracy, while also being prone to abuse. Here I think what is warranted is extreme care and scrutiny combined with guardrails to prevent abuses. My main fear is having too small a number of "fingers on the button," such that one or a handful of people could essentially operate a drone army without needing any other humans to cooperate to carry out their orders. As AI systems get more powerful, we may need to have more direct and immediate oversight mechanisms to ensure they are not misused, perhaps involving branches of government other than the executive. I think we should approach fully autonomous weapons in particular with great caution, [36] and not rush into their use without proper safeguards.

Fourth, after drawing a hard line against AI abuses in democracies, we should use that precedent to create an international taboo against the worst abuses of powerful AI. I recognize that the current political winds have turned against international cooperation and international norms, but this is a case where we sorely need them. The world needs to understand the dark potential of powerful AI in the hands of autocrats, and to recognize that certain uses of AI amount to an attempt to permanently steal their freedom and impose a totalitarian state from which they can't escape. I would even argue that in some cases, large-scale surveillance with powerful AI, mass propaganda with powerful AI, and certain types of *offensive* uses of fully autonomous weapons should be considered crimes against humanity. More generally, a robust norm against AI-enabled totalitarianism and all its tools and instruments is sorely needed.

It is possible to have an even stronger version of this position, which is that because the possibilities of AI-enabled totalitarianism are so dark, autocracy is simply not a form of government that people can accept in the post-powerful AI age. Just as feudalism became unworkable with the industrial revolution,

the AI age could lead inevitably and logically to the conclusion that democracy (and, hopefully, democracy improved and reinvigorated by AI, as I discuss in *Machines of Loving Grace*) is the only viable form of government if humanity is to have a good future.

Fifth and finally, AI companies should be carefully watched, as should their connection to the government, which is necessary, but must have limits and boundaries. The sheer amount of capability embodied in powerful AI is such that ordinary corporate governance—which is designed to protect shareholders and prevent ordinary abuses such as fraud—is unlikely to be up to the task of governing AI companies. There may also be value in companies publicly committing to (perhaps even as part of corporate governance) not take certain actions, such as privately building or stockpiling military hardware, using large amounts of computing resources by single individuals in unaccountable ways, or using their AI products as propaganda to manipulate public opinion in their favor.

The danger here comes from many directions, and some directions are in tension with others. The only constant is that we must seek accountability, norms, and guardrails for everyone, even as we empower "good" actors to keep "bad" actors in check.

4. Player piano

Economic disruption

The previous three sections were essentially about security risks posed by powerful AI: risks from the AI itself, risks from misuse by individuals and small organizations and risks of misuse by states and large organizations. If we put aside security risks or assume they have been solved, the next question is economic. What will be the effect of this infusion of incredible "human" capital on the economy? Clearly, the most obvious effect will be to greatly increase economic growth. The pace of advances in scientific research, biomedical innovation, manufacturing, supply chains, the efficiency of the

financial system, and much more are almost guaranteed to lead to a much faster rate of economic growth. In *Machines of Loving Grace*, I suggest that a 10–20% sustained annual GDP growth rate may be possible.

But it should be clear that this is a double-edged sword: what are the economic prospects for most existing humans in such a world? New technologies often bring labor market shocks, and in the past humans have always recovered from them, but I am concerned that this is because these previous shocks affected only a small fraction of the full possible range of human abilities, leaving room for humans to expand to new tasks. AI will have effects that are much broader and occur much faster, and therefore I worry it will be much more challenging to make things work out well.

Labor market disruption

There are two specific problems I am worried about: labor market displacement, and concentration of economic power. Let's start with the first one. This is a topic that I warned about very publicly in 2025, where I predicted that AI could displace half of all entry-level white collar jobs in the next 1–5 years, even as it accelerates economic growth and scientific progress. This warning started a public debate about the topic. Many CEOs, technologists, and economists agreed with me, but others assumed I was falling prey to a "lump of labor" fallacy and didn't know how labor markets worked, and some didn't see the 1–5-year time range and thought I was claiming AI is displacing jobs right now (which I agree it is likely not). So it is worth going through in detail why I am worried about labor displacement, to clear up these misunderstandings.

As a baseline, it's useful to understand how labor markets *normally* respond to advances in technology. When a new technology comes along, it starts by making pieces of a given human job more efficient. For example, early in the Industrial Revolution, machines, such as upgraded plows, enabled human farmers to be more efficient at some aspects of the job. This improved the productivity of farmers, which increased their wages.

In the next step, some parts of the job of farming could be done *entirely* by machines, for example with the invention of the threshing machine or seed drill. In this phase, humans did a lower and lower fraction of the job, but the work they *did* complete became more and more leveraged because it is complementary to the work of machines, and their productivity continued to rise. As described by Jevons' paradox, the wages of farmers and perhaps even the number of farmers continued to increase. Even when 90% of the job is being done by machines, humans can simply do 10x more of the 10% they still do, producing 10x as much output for the same amount of labor.

Eventually, machines do everything or almost everything, as with modern combine harvesters, tractors, and other equipment. At this point farming as a form of human employment really does go into steep decline, and this potentially causes serious disruption in the short term, but because farming is just one of many useful activities that humans are able to do, people eventually switch to other jobs, such as operating factory machines. This is true even though farming accounted for a huge proportion of employment *ex ante*. 250 years ago, 90% of Americans lived on farms; in Europe, 50–60% of employment was agricultural. Now those percentages are in the low single digits in those places, because workers switched to industrial jobs (and later, knowledge work jobs). The economy can do what previously required most of the labor force with only 1–2% of it, freeing up the rest of the labor force to build an ever more advanced industrial society. There's no fixed "lump of labor," just an ever-expanding ability to do more and more with less and less. People's wages rise in line with the GDP exponential and the economy maintains full employment once disruptions in the short term have passed.

It's possible things will go roughly the same way with AI, but I would bet pretty strongly against it. Here are some reasons I think AI is likely to be different:

- **Speed.** The pace of progress in AI is much faster than for previous technological revolutions. For example, in the last 2 years, AI models went from barely being able to complete a single line of code, to writing

all or almost all of the code for some people—including engineers at Anthropic. [37] Soon, they may do the entire task of a software engineer end to end. [38] It is hard for people to adapt to this pace of change, both to the changes in how a given job works and in the need to switch to new jobs. Even legendary programmers are increasingly describing themselves as "behind." The pace may if anything continue to speed up, as AI coding models increasingly accelerate the task of AI development. To be clear, speed in itself does not mean labor markets and employment won't eventually recover, it just implies the short-term transition will be unusually painful compared to past technologies, since humans and labor markets are slow to react and to equilibrate.

- **Cognitive breadth.** As suggested by the phrase "country of geniuses in a datacenter," AI will be capable of a very wide range of human cognitive abilities—perhaps all of them. This is very different from previous technologies like mechanized farming, transportation, or even computers. [39] This will make it harder for people to switch easily from jobs that are displaced to similar jobs that they would be a good fit for. For example, the general intellectual abilities required for entry-level jobs in, say, finance, consulting, and law are fairly similar, even if the specific knowledge is quite different. A technology that disrupted only one of these three would allow employees to switch to the two other close substitutes (or for undergraduates to switch majors). But disrupting all three at once (along with many other similar jobs) may be harder for people to adapt to. Furthermore, it's not *just* that most existing jobs will be disrupted. That part has happened before—recall that farming was a huge percentage of employment. But farmers could switch to the relatively similar work of operating factory machines, even though that work hadn't been common before. By contrast, AI is increasingly matching the general cognitive profile of humans, which means it will also be good at the new jobs that would ordinarily be created in response to the old ones being automated. Another way to say it is that AI isn't a substitute for specific human jobs but rather a general labor substitute for humans.

- **Slicing by cognitive ability.** Across a wide range of tasks, AI appears to be advancing from the bottom of the ability ladder to the top. For example, in coding our models have proceeded from the level of "a mediocre coder" to "a strong coder" to "a very strong coder." [40] We are now starting to see the same progression in white-collar work in general. We are thus at risk of a situation where, instead of affecting people with specific skills or in specific professions (who can adapt by retraining), AI is affecting people with certain intrinsic cognitive properties, namely lower intellectual ability (which is harder to change). It is not clear where these people will go or what they will do, and I am concerned that they could form an unemployed or very-low-wage "underclass." To be clear, things somewhat like this have happened before—for example, computers and the internet are believed by some economists to represent "skill-biased technological change." But this skill biasing was both not as extreme as what I expect to see with AI, and is believed to have contributed to an increase in wage inequality, [41] so it is not exactly a reassuring precedent.
- **Ability to fill in the gaps.** The way human jobs often adjust in the face of new technology is that there are many aspects to the job, and the new technology, even if it appears to directly replace humans, often has gaps in it. If someone invents a machine to make widgets, humans may still have to load raw material into the machine. Even if that takes only 1% as much effort as making the widgets manually, human workers can simply make 100x more widgets. But AI, in addition to being a rapidly advancing technology, is also a rapidly *adapting* technology. During every model release, AI companies carefully measure what the model is good at and what it isn't, and customers also provide such information after the launch. Weaknesses can be addressed by collecting tasks that embody the current gap, and training on them for the next model. Early in generative AI, users noticed that AI systems had certain weaknesses (such as AI image models generating hands with the wrong number of fingers) and many assumed these weaknesses were inherent to the technology. If they

were, it would limit job disruption. But pretty much every such weakness gets addressed quickly— often, within just a few months.

It's worth addressing common points of skepticism. First, there is the argument that economic diffusion will be slow, such that even if the underlying technology is *capable* of doing most human labor, the actual application of it across the economy may be much slower (for example in industries that are far from the AI industry and slow to adopt). Slow diffusion of technology is definitely real—I talk to people from a wide variety of enterprises, and there are places where the adoption of AI will take years. That's why my prediction for 50% of entry level white collar jobs being disrupted is 1–5 years, even though I suspect we'll have powerful AI (which would be, technologically speaking, enough to do *most or all* jobs, not just entry level) in much less than 5 years. But diffusion effects merely buy us time. And I am not confident they will be as slow as people predict. Enterprise AI adoption is growing at rates much faster than any previous technology, largely on the pure strength of the technology itself. Also, even if traditional enterprises are slow to adopt new technology, startups will spring up to serve as "glue" and make the adoption easier. If that doesn't work, the startups may simply disrupt the incumbents directly.

That could lead to a world where it isn't so much that specific jobs are disrupted as it is that large enterprises are disrupted in general and replaced with much less labor-intensive startups. This could also lead to a world of "geographic inequality," where an increasing fraction of the world's wealth is concentrated in Silicon Valley, which becomes its own economy running at a different speed than the rest of the world and leaving it behind. All of these outcomes would be great for economic growth—but not so great for the labor market or those who are left behind.

Second, some people say that human jobs will move to the physical world, which avoids the whole category of "cognitive labor" where AI is progressing so rapidly. I am not sure how safe this is, either. A lot of physical labor is already being done by machines (e.g., manufacturing) or will soon be done by

machines (e.g., driving). Also, sufficiently powerful AI will be able to accelerate the development of robots, and then control those robots in the physical world. It may buy some time (which is a good thing), but I'm worried it won't buy much. And even if the disruption was limited only to cognitive tasks, it would still be an unprecedentedly large and rapid disruption.

Third, perhaps some tasks inherently require or greatly benefit from a human touch. I'm a little more uncertain about this one, but I'm still skeptical that it will be enough to offset the bulk of the impacts I described above. AI is already widely used for customer service. Many people report that it is easier to talk to AI about their personal problems than to talk to a therapist—that the AI is more patient. When my sister was struggling with medical problems during a pregnancy, she felt she wasn't getting the answers or support she needed from her care providers, and she found Claude to have a better bedside manner (as well as succeeding better at diagnosing the problem). I'm sure there are some tasks for which a human touch really is important, but I'm not sure how many—and here we're talking about finding work for nearly everyone in the labor market.

Fourth, some may argue that comparative advantage will still protect humans. Under the law of comparative advantage, even if AI is better than humans at everything, any *relative* differences between the human and AI profile of skills creates a basis of trade and specialization between humans and AI. The problem is that if AIs are literally thousands of times more productive than humans, this logic starts to break down. Even tiny transaction costs could make it not worth it for AI to trade with humans. And human wages may be very low, even if they technically have something to offer.

It's possible all of these factors can be addressed—that the labor market is resilient enough to adapt to even such an enormous disruption. But even if it can eventually adapt, the factors above suggest that the short-term shock will be unprecedented in size.

Defenses

What can we do about this problem? I have several suggestions, some of which Anthropic is already doing. The first thing is simply to get accurate data about what is happening with job displacement in real time. When an economic change happens very quickly, it's hard to get reliable data about what is happening, and without reliable data it is hard to design effective policies. For example, government data is currently lacking granular, high-frequency data on AI adoption across firms and industries. For the last year Anthropic has been operating and publicly releasing an Economic Index that shows use of our models almost in real time, broken down by industry, task, location, and even things like whether a task was being automated or conducted collaboratively. We also have an Economic Advisory Council to help us interpret this data and see what is coming.

Second, AI companies have a choice in how they work with enterprises. The very inefficiency of traditional enterprises means that their rollout of AI can be very path dependent, and there is some room to choose a better path. Enterprises often have a choice between "cost savings" (doing the same thing with fewer people) and "innovation" (doing more with the same number of people). The market will inevitably produce both eventually, and any competitive AI company will have to serve some of both, but there may be some room to steer companies towards innovation when possible, and it may buy us some time. Anthropic is actively thinking about this.

Third, companies should think about how to take care of their employees. In the short term, being creative about ways to reassign employees within companies may be a promising way to stave off the need for layoffs. In the long term, in a world with enormous total wealth, in which many companies increase greatly in value due to increased productivity and capital concentration, it may be feasible to pay human employees even long after they are no longer providing economic value in the traditional sense. Anthropic is currently considering a range of possible pathways for our own employees that we will share in the near future.

Fourth, wealthy individuals have an obligation to help solve this problem. It is sad to me that many wealthy individuals (especially in the tech industry) have recently adopted a cynical and nihilistic attitude that philanthropy is inevitably fraudulent or useless. Both private philanthropy like the Gates Foundation and public programs like PEPFAR have saved tens of millions of lives in the developing world, and helped to create economic opportunity in the developed world. All of Anthropic's co-founders have pledged to donate 80% of our wealth, and Anthropic's staff have individually pledged to donate company shares worth billions at current prices—donations that the company has committed to matching.

Fifth, while all the above private actions can be helpful, ultimately a macroeconomic problem this large will require government intervention. The natural policy response to an enormous economic pie coupled with high inequality (due to a lack of jobs, or poorly paid jobs, for many) is progressive taxation. The tax could be general or could be targeted against AI companies in particular. Obviously tax design is complicated, and there are many ways for it to go wrong. I don't support poorly designed tax policies. I think the extreme levels of inequality predicted in this essay justify a more robust tax policy on basic moral grounds, but I can also make a pragmatic argument to the world's billionaires that it's in their interest to support a good version of it: if they don't support a good version, they'll inevitably get a bad version designed by a mob.

Ultimately, I think of all of the above interventions as ways to buy time. In the end AI will be able to do everything, and we need to grapple with that. It's my hope that by that time, we can use AI itself to help us restructure markets in ways that work for everyone, and that the interventions above can get us through the transitional period.

Economic concentration of power

Separate from the problem of job displacement or economic inequality *per se* is the problem of *economic concentration of power.* Section 1 discussed the risk that humanity gets disempowered by AI, and Section 3 discussed the risk that

citizens get disempowered by their governments by force or coercion. But another kind of disempowerment can occur if there is such a huge concentration of wealth that a small group of people effectively controls government policy with their influence, and ordinary citizens have no influence because they lack economic leverage. Democracy is ultimately backstopped by the idea that the population as a whole is necessary for the operation of the economy. If that economic leverage goes away, then the implicit social contract of democracy may stop working. Others have written about this, so I needn't go into great detail about it here, but I agree with the concern, and I worry it is already starting to happen.

To be clear, I am not opposed to people making a lot of money. There's a strong argument that it incentivizes economic growth under normal conditions. I am sympathetic to concerns about impeding innovation by killing the golden goose that generates it. But in a scenario where GDP growth is 10–20% a year and AI is rapidly taking over the economy, yet single individuals hold appreciable fractions of the GDP, innovation is *not* the thing to worry about. The thing to worry about is a level of wealth concentration that will break society.

The most famous example of extreme concentration of wealth in US history is the Gilded Age, and the wealthiest industrialist of the Gilded Age was John D. Rockefeller. Rockefeller's wealth amounted to ~2% of the US GDP at the time. [42] A similar fraction today would lead to a fortune of $600B, and the richest person in the world today (Elon Musk) already exceeds that, at roughly $700B. So we are already at historically unprecedented levels of wealth concentration, even *before* most of the economic impact of AI. I don't think it is too much of a stretch (if we get a "country of geniuses") to imagine AI companies, semiconductor companies, and perhaps downstream application companies generating ~$3T in revenue per year, [43] being valued at ~$30T, and leading to personal fortunes well into the trillions. In that world, the debates we have about tax policy today simply won't apply as we will be in a fundamentally different situation.

Related to this, the coupling of this economic concentration of wealth with the political system already concerns me. AI datacenters already represent a substantial fraction of US economic growth, [44] and are thus strongly tying together the financial interests of large tech companies (which are increasingly focused on either AI or AI infrastructure) and the political interests of the government in a way that can produce perverse incentives. We already see this through the reluctance of tech companies to criticize the US government, and the government's support for extreme anti-regulatory policies on AI.

Defenses

What can be done about this? First, and most obviously, companies should simply choose not to be part of it. Anthropic has always strived to be a policy actor and not a political one, and to maintain our authentic views whatever the administration. We've spoken up in favor of sensible AI regulation and export controls that are in the public interest, even when these are at odds with government policy. [45] Many people have told me that we should stop doing this, that it could lead to unfavorable treatment, but in the year we've been doing it, Anthropic's valuation has increased by over 6x, an almost unprecedented jump at our commercial scale.

Second, the AI industry needs a healthier relationship with government—one based on substantive policy engagement rather than political alignment. Our choice to engage on policy substance rather than politics is sometimes read as a tactical error or failure to "read the room" rather than a principled decision, and that framing concerns me. In a healthy democracy, companies should be able to advocate for good policy for its own sake. Related to this, a public backlash against AI is brewing: this could be a corrective, but it's currently unfocused. Much of it targets issues that aren't actually problems (like datacenter water usage) and proposes solutions (like datacenter bans or poorly designed wealth taxes) that wouldn't address the real concerns. The underlying issue that deserves attention is ensuring that AI development remains accountable to the public interest, not captured by any particular

political or commercial alliance, and it seems important to focus the public discussion there.

Third, the macroeconomic interventions I described earlier in this section, as well as a resurgence of private philanthropy, can help to balance the economic scales, addressing both the job displacement and concentration of economic power problems at once. We should look to the history of our country here: even in the Gilded Age, industrialists such as Rockefeller and Carnegie felt a strong obligation to society at large, a feeling that society had contributed enormously to their success and they needed to give back. That spirit seems to be increasingly missing today, and I think it is a large part of the way out of this economic dilemma. Those who are at the forefront of AI's economic boom should be willing to give away both their wealth and their power.

5. Black seas of infinity

Indirect effects

This last section is a catchall for unknown unknowns, particularly things that could go wrong as an indirect result of positive advances in AI and the resulting acceleration of science and technology in general. Suppose we address all the risks described so far, and begin to reap the benefits of AI. We will likely get a "century of scientific and economic progress compressed into a decade," and this will be hugely positive for the world, but we will then have to contend with the problems that arise from this rapid rate of progress, and those problems may come at us fast. We may also encounter other risks that occur indirectly as a consequence of AI progress and are hard to anticipate in advance.

By the nature of unknown unknowns it is impossible to make an exhaustive list, but I'll list three possible concerns as illustrative examples for what we should be watching for:

- **Rapid advances in biology.** If we do get a century of medical progress in a few years, it is possible that we will greatly increase the human lifespan, and there is a chance we also gain radical capabilities like the ability to increase human intelligence or radically modify human biology. Those would be big changes in what is possible, happening very quickly. They could be positive if responsibly done (which is my hope, as described in *Machines of Loving Grace*), but there is always a risk they go very wrong—for example, if efforts to make humans smarter also make them more unstable or power-seeking. There is also the issue of "uploads" or "whole brain emulation," digital human minds instantiated in software, which might someday help humanity transcend its physical limitations, but which also carry risks I find disquieting.
- **AI changes human life in an unhealthy way.** A world with billions of intelligences that are much smarter than humans at everything is going to be a very weird world to live in. Even if AI doesn't actively aim to attack humans (Section 1), and isn't explicitly used for oppression or control by states (Section 3), there is a lot that could go wrong short of this, via normal business incentives and nominally consensual transactions. We see early hints of this in the concerns about AI psychosis, AI driving people to suicide, and concerns about romantic relationships with AIs. As an example, could powerful AIs invent some new religion and convert millions of people to it? Could most people end up "addicted" in some way to AI interactions? Could people end up being "puppeted" by AI systems, where an AI essentially watches their every move and tells them exactly what to do and say at all times, leading to a "good" life but one that lacks freedom or any pride of accomplishment? It would not be hard to generate dozens of these scenarios if I sat down with the creator of *Black Mirror* and tried to brainstorm them. I think this points to the importance of things like improving Claude's Constitution, over and above what is necessary for preventing the issues in Section 1. Making sure that AI models *really* have their users' long-term interests at heart, in a way thoughtful people would endorse rather than in some subtly distorted way, seems critical.

- **Human purpose.** This is related to the previous point, but it's not so much about specific human interactions with AI systems as it is about how human life changes in general in a world with powerful AI. Will humans be able to find purpose and meaning in such a world? I think this is a matter of attitude: as I said in *Machines of Loving Grace*, I think human purpose does not depend on being the best in the world at something, and humans can find purpose even over very long periods of time through stories and projects that they love. We simply need to break the link between the generation of economic value and self-worth and meaning. But that is a transition society has to make, and there is always the risk we don't handle it well.

My hope with all of these potential problems is that in a world with powerful AI that we trust not to kill us, that is not the tool of an oppressive government, and that is genuinely working on our behalf, we can use AI itself to anticipate and prevent these problems. But that is not guaranteed—like all of the other risks, it is something we have to handle with care.

Humanity's test

Reading this essay may give the impression that we are in a daunting situation. I certainly found it daunting to write, in contrast with *Machines of Loving Grace,* which felt like giving form and structure to surpassingly beautiful music that had been echoing in my head for years. And there is much about the situation that genuinely is hard. AI brings threats to humanity from multiple directions, and there is genuine tension between the different dangers, where mitigating some of them risks making others worse if we do not thread the needle extremely carefully.

Taking time to carefully build AI systems so they do not autonomously threaten humanity is in genuine tension with the need for democratic nations to stay ahead of authoritarian nations and not be subjugated by them. But in turn, the same AI-enabled tools that are necessary to fight autocracies can, if taken too far, be turned inward to create tyranny in our own countries. AI-

driven terrorism could kill millions through the misuse of biology, but an overreaction to this risk could lead us down the road to an autocratic surveillance state. The labor and economic concentration effects of AI, in addition to being grave problems in their own right, may force us to face the other problems in an environment of public anger and perhaps even civil unrest, rather than being able to call on the better angels of our nature. Above all, the sheer *number* of risks, including unknown ones, and the need to deal with all of them at once, creates an intimidating gauntlet that humanity must run.

Furthermore, the last few years should make clear that the idea of stopping or even substantially slowing the technology is fundamentally untenable. The formula for building powerful AI systems is incredibly simple, so much so that it can almost be said to emerge spontaneously from the right combination of data and raw computation. Its creation was probably inevitable the instant humanity invented the transistor, or arguably even earlier when we first learned to control fire. If one company does not build it, others will do so nearly as fast. If all companies in democratic countries stopped or slowed development, by mutual agreement or regulatory decree, then authoritarian countries would simply keep going. Given the incredible economic and military value of the technology, together with the lack of any meaningful enforcement mechanism, I don't see how we could possibly convince them to stop.

I do see a path to a *slight* moderation in AI development that is compatible with a realist view of geopolitics. That path involves slowing down the march of autocracies towards powerful AI for a few years by denying them the resources they need to build it, [46] namely chips and semiconductor manufacturing equipment. This in turn gives democratic countries a buffer that they can "spend" on building powerful AI more carefully, with more attention to its risks, while still proceeding fast enough to comfortably beat the autocracies. The race between AI companies within democracies can then be handled under the umbrella of a common legal framework, via a mixture of industry standards and regulation.

Anthropic has advocated very hard for this path, by pushing for chip export controls and judicious regulation of AI, but even these seemingly common-sense proposals have largely been rejected by policymakers in the United States (which is the country where it's most important to have them). There is so much money to be made with AI—literally trillions of dollars per year—that even the simplest measures are finding it difficult to overcome the political economy inherent in AI. This is the trap: AI is so powerful, such a glittering prize, that it is very difficult for human civilization to impose any restraints on it at all.

I can imagine, as Sagan did in *Contact*, that this same story plays out on thousands of worlds. A species gains sentience, learns to use tools, begins the exponential ascent of technology, faces the crises of industrialization and nuclear weapons, and if it survives those, confronts the hardest and final challenge when it learns how to shape sand into machines that think. Whether we survive that test and go on to build the beautiful society described in *Machines of Loving Grace*, or succumb to slavery and destruction, will depend on our character and our determination as a species, our spirit and our soul.

Despite the many obstacles, I believe humanity has the strength inside itself to pass this test. I am encouraged and inspired by the thousands of researchers who have devoted their careers to helping us understand and steer AI models, and to shaping the character and constitution of these models. I think there is now a good chance that those efforts bear fruit in time to matter. I am encouraged that at least some companies have stated they'll pay meaningful commercial costs to block their models from contributing to the threat of bioterrorism. I am encouraged that a few brave people have resisted the prevailing political winds and passed legislation that puts the first early seeds of sensible guardrails on AI systems. I am encouraged that the public understands that AI carries risks and wants those risks addressed. I am encouraged by the indomitable spirit of freedom around the world and the determination to resist tyranny wherever it occurs.

But we will need to step up our efforts if we want to succeed. The first step is for those closest to the technology to simply tell the truth about the situation humanity is in, which I have always tried to do; I'm doing so more explicitly and with greater urgency with this essay. The next step will be convincing the world's thinkers, policymakers, companies, and citizens of the imminence and overriding importance of this issue—that it is worth expending thought and political capital on this in comparison to the thousands of other issues that dominate the news every day. Then there will be a time for courage, for enough people to buck the prevailing trends and stand on principle, even in the face of threats to their economic interests and personal safety.

The years in front of us will be impossibly hard, asking more of us than we think we can give. But in my time as a researcher, leader, and citizen, I have seen enough courage and nobility to believe that we can win—that when put in the darkest circumstances, humanity has a way of gathering, seemingly at the last minute, the strength and wisdom needed to prevail. We have no time to lose.

⁂

I would like to thank Erik Brynjolfsson, Ben Buchanan, Mariano-Florentino Cuéllar, Allan Dafoe, Kevin Esvelt, Nick Beckstead, Richard Fontaine, Jim McClave, and very many of the staff at Anthropic for their helpful comments on drafts of this essay.

Footnotes

[1] This is symmetric to a point I made in *Machines of Loving Grace*, where I started by saying that AI's upsides shouldn't be thought of in terms of a prophecy of salvation, and that it's important to be concrete and grounded and to avoid grandiosity. Ultimately, prophecies of salvation and prophecies of doom are unhelpful for confronting the real world, for basically the same reasons. ↩

[2] Anthropic's goal is to remain consistent through such changes. When talking about AI risks was politically popular, Anthropic cautiously advocated for a judicious and evidence-based approach to these risks. Now that talking about AI risks is politically unpopular, Anthropic continues to cautiously advocate for a judicious and evidence-based approach to these risks. ↩

[3] Over time, I have gained increasing confidence in the trajectory of AI and the likelihood that it will surpass human ability across the board, but some uncertainty still remains. ↩

[4] Export controls for chips are a great example of this. They are simple and appear to mostly just work. ↩

[5] And of course, the hunt for such evidence must be intellectually honest, such that it could also turn up evidence of a lack of danger. Transparency through model cards and other disclosures is an attempt at such an intellectually honest endeavor. ↩

[6] Indeed, since writing *Machines of Loving Grace* in 2024, AI systems have become capable of doing tasks that take humans several hours, with METR recently assessing that Opus 4.5 can do about four human hours of work with 50% reliability. ↩

[7] And to be clear, even if powerful AI is only 1–2 years away in a technical sense, many of its societal consequences, both positive and negative, may take a few years longer to occur. This is why I can simultaneously think that AI will disrupt 50% of *entry-level* white-collar jobs over 1–5 years, while also thinking we may have AI that is more capable than *everyone* in only 1–2 years. ↩

[8] It is worth adding that the *public* (as compared to policymakers) does seem to be very concerned with AI risks. I think some of their focus is correct (i.e. AI job displacement), and some is misguided (such as concerns about water use of AI, which is not significant). This backlash gives me hope that a consensus around addressing risks is possible, but so far it has not yet

been translated into policy changes, let alone effective or well-targeted policy changes. ↩

[9] They can also, of course, manipulate (or simply pay) large numbers of humans into doing what they want in the physical world. ↩

[10] I don't think this is a straw man: it's my understanding, for example, that Yann LeCun holds this position. ↩

[11] For example, see Section 5.5.2 (p. 63–66) of the Claude 4 system card. ↩

[12] There are also a number of other assumptions inherent in the simple model, which I won't discuss here. Broadly, they should make us less worried about the specific simple story of misaligned power-seeking, but also more worried about possible unpredictable behavior we haven't anticipated. ↩

[13] *Ender's Game* describes a version of this involving humans rather than AI. ↩

[14] For example, models may be told not to do various bad things, and also to obey humans, but may then observe that many humans do exactly those bad things! It's not clear how this contradiction would resolve (and a well-designed constitution should encourage the model to handle these contradictions gracefully), but this type of dilemma is not so different from the supposedly "artificial" situations that we put AI models in during testing. ↩

[15] Incidentally, one consequence of the constitution being a natural-language document is that it is legible to the world, and that means it can be critiqued by anyone and compared to similar documents by other companies. It would be valuable to create a race to the top that not only encourages companies to release these documents, but encourages them to be good. ↩

[16] There's even a hypothesis about a deep unifying principle connecting the character-based approach from Constitutional AI to results from interpretability and alignment science. According to the hypothesis, the fundamental mechanisms driving Claude originally arose as ways for it to

simulate characters in pretraining, such as predicting what the characters in a novel would say. This would suggest that a useful way to think about the constitution is more like a character description that the model uses to instantiate a consistent persona. It would also help us explain the "I must be a bad person" results I mentioned above (because the model is trying to *act as if* it's a coherent character—in this case a bad one), and would suggest that interpretability methods should be able to discover "psychological traits" within models. Our researchers are working on ways to test this hypothesis. ↩

[17] To be clear, monitoring is done in a privacy-preserving way. ↩

[18] Even in our own experiments with what are essentially voluntarily imposed rules with our Responsible Scaling Policy, we have found over and over again that it's very easy to end up being too rigid, by drawing lines that seem important ex ante but turn out to be silly in retrospect. It is just very easy to set rules about the wrong things when a technology is advancing rapidly. ↩

[19] SB 53 and RAISE do not apply at all to companies with under $500M in annual revenue. They only apply to larger, more established companies like Anthropic. ↩

[20] I originally read Joy's essay 25 years ago, when it was written, and it had a profound impact on me. Then and now, I do see it as too pessimistic—I don't think broad "relinquishment" of whole areas of technology, which Joy suggests, is the answer—but the issues it raises were surprisingly prescient, and Joy also writes with a deep sense of compassion and humanity that I admire. ↩

[21] We do have to worry about state actors, now and in the future, and I discuss that in the next section. ↩

[22] There is evidence that many terrorists are at least relatively well-educated, which might seem to contradict what I'm arguing here about a negative correlation between ability and motivation. But I think in actual fact they are compatible observations: if the ability threshold for a successful attack is high, then almost by definition those who *currently* succeed must have high ability,

even if ability and motivation are negatively correlated. But in a world where the limitations on ability were removed (e.g., with future LLMs), I'd predict that a substantial population of people with the motivation to kill but lower ability would start to do so—just as we see for crimes that don't require much ability (like school shootings). ↩

[23] Aum Shinrikyo did try, however. The leader of Aum Shinrikyo, Seiichi Endo, had training in virology from Kyoto University, and attempted to produce both anthrax and ebola. However, as of 1995, even he lacked enough expertise and resources to succeed at this. The bar is now substantially lower, and LLMs could reduce it even further. ↩

[24] A bizarre phenomenon relating to mass murderers is that the style of murder they choose operates almost as a grotesque sort of fad. In the 1970s and 1980s, serial killers were very common, and new serial killers often copied the behavior of more established or famous serial killers. In the 1990s and 2000s, mass shootings became more common, while serial killers became less common. There is no technological change that triggered these patterns of behavior, it just appears that violent murderers were copying each others' behavior and the "popular" thing to copy changed. ↩

[25] Casual jailbreakers sometimes believe that they've compromised these classifiers when they get the model to output one specific piece of information, such as the genome sequence of a virus. But as I explained before, the threat model we are worried about involves step-by-step, interactive advice that extends over weeks or months about specific obscure steps in the bioweapons production process, and this is what our classifiers aim to defend against. (We often describe our research as looking for "universal" jailbreaks—ones that don't just work in one specific or narrow context, but broadly open up the model's behavior.) ↩

[26] Though we will continue to invest in work to make our classifiers more efficient, and it may make sense for companies to share advances like these with one another. ↩

[27] Obviously, I do not think companies should have to disclose technical details about the specific steps in biological weapons production that they are blocking, and the transparency legislation that has been passed so far (SB 53 and RAISE) accounts for this issue. ↩

[28] Another related idea is "resilience markets" where the government encourages stockpiling of PPE, respirators, and other essential equipment needed to respond to a biological attack by promising ahead of time to pay a pre-agreed price for this equipment in an emergency. This incentivizes suppliers to stockpile such equipment without fear that the government will seize it without compensation. ↩

[29] Why am I more worried about large actors for seizing power, but small actors for causing destruction? Because the dynamics are different. Seizing power is about whether one actor can amass enough strength to overcome everyone else—thus we should worry about the most powerful actors and/or those closest to AI. Destruction, by contrast, can be wrought by those with little power if it is much harder to defend against than to cause. It is then a game of defending against the most *numerous* threats, which are likely to be smaller actors. ↩

[30] This might sound like it is in tension with my point that attack and defense may be more balanced with cyberattacks than with bioweapons, but my worry here is that if a country's AI is the most powerful in the world, then others will not be able to defend even if the technology itself has an intrinsic attack-defense balance. ↩

[31] For example, in the United States this includes the fourth amendment and the Posse Comitatus Act. ↩

[32] Also, to be clear, there are some arguments for building large datacenters in countries with varying governance structures, particularly if they are controlled by companies in democracies. Such buildouts could in principle help democracies compete better with the CCP, which is the greater threat. I also think such datacenters don't pose much risk unless they are very large.

But on balance, I think caution is warranted when placing very large datacenters in countries where institutional safeguards and rule-of-law protections are less well-established. ↩

[33] This is, of course, also an argument for improving the security of the nuclear deterrent to make it more likely to be robust against powerful AI, and nuclear-armed democracies should do this. But we don't know what a powerful AI will be capable of or which defenses, if any, will work against it, so we should not assume that these measures will necessarily solve the problem. ↩

[34] There is also the risk that even if the nuclear deterrent remains effective, an attacking country might decide to call our bluff—it's unclear whether we'd be willing to use nuclear weapons to defend against a drone swarm even if the drone swarm has a substantial risk of conquering us. Drone swarms might be a new thing that is less severe than nuclear attacks but more severe than conventional attacks. Alternatively, differing assessments of the effectiveness of the nuclear deterrent in the age of AI might alter the game theory of nuclear conflict in a destabilizing manner. ↩

[35] To be clear, I would believe it is the right strategy not to sell chips to China, even if the timeline to powerful AI were substantially longer. We cannot get the Chinese "addicted" to American chips—they are determined to develop their native chip industry one way or another. It will take them many years to do so, and all we are doing by selling them chips is giving them a big boost during that time. ↩

[36] To be clear, most of what is being used in Ukraine and Taiwan today are not *fully* autonomous weapons. These are coming, but not here today. ↩

[37] Our model card for Claude Opus 4.5, our most recent model, shows that Opus performs better on a performance engineering interview frequently given at Anthropic than any interviewee in the history of the company. ↩

[38] "Writing all of the code" and "doing the task of a software engineer end to end" are very different things, because software engineers do much more than

just write code, including testing, dealing with environments, files, and installation, managing cloud compute deployments, iterating on products, and much more. ↩

[39] Computers are general in a sense, but are clearly incapable on their own of the vast majority of human cognitive abilities, even as they greatly exceed humans in a few areas (such as arithmetic). Of course, things built *on top* of computers, such as AI, are now capable of a wide range of cognitive abilities, which is what this essay is about. ↩

[40] To be clear, AI models do not have precisely the same profile of strengths and weaknesses as humans. But they are also advancing fairly uniformly along every dimension, such that having a spiky or uneven profile may not ultimately matter. ↩

[41] Though there is debate among economists about this idea. ↩

[42] Personal wealth is a "stock," while GDP is a "flow," so this isn't a claim that Rockefeller owned 2% of the economic value in the United States. But it's harder to measure the total wealth of a nation than the GDP, and people's individual incomes vary a lot per year, so it's hard to make a ratio in the same units. The ratio of the largest personal fortune to GDP, while not comparing apples to apples, is nevertheless a perfectly reasonable benchmark for extreme wealth concentration. ↩

[43] The total value of labor across the economy is $60T/year, so $3T/year would correspond to 5% of this. That amount could be earned by a company that supplied labor for 20% of the cost of humans and had 25% market share, even if the demand for labor did not expand (which it almost certainly would due to the lower cost). ↩

[44] To be clear, I do not think actual AI productivity is yet responsible for a substantial fraction of US economic growth. Rather, I think the datacenter spending represents growth caused by anticipatory investment that amounts to the market expecting *future* AI-driven economic growth and investing accordingly. ↩

[45] When we agree with the administration, we say so, and we look for points of agreement where mutually supported policies are genuinely good for the world. We are aiming to be honest brokers rather than backers or opponents of any given political party. ↩

[46] I don't think anything more than a few years is possible: on longer timescales, they will build their own chips. ↩

Back to top