America's Superintelligence Project

April 2025

Jeremie Harris, Edouard Harris

hello@gladstone.ai

Redacted text has had its length randomized.

Executive Summary

Private labs in the United States believe they're about to develop artificial superintelligence (ASI), a technology that could provide an unmatched strategic advantage to whoever builds and controls[1] it. They could be right. But as it stands, frontier AI research is vulnerable to espionage and sabotage, particularly from the CCP. All of our frontier labs are almost certainly CCP-penetrated.[2] The national security stakes of superintelligence are too high to let this status quo continue.

That's why there have been growing calls for deeper government involvement in securing frontier AI research – up to and including a formal recommendation from the U.S.-China Economic and Security Review Commission to establish a "Manhattan Project for AGI". The prospect of a Manhattan Project for superintelligence has been widely discussed in the Silicon Valley AI scene for years, and hit the mainstream when former OpenAI researcher Leopold Aschenbrenner published his influential manifesto in June 2024.

But setting up and securing a national superintelligence project wouldn't be easy. We've spent the last 12 months figuring out how it could be done, what it would take for it to succeed, and how it could fail. We interviewed over 100 domain specialists, including Tier 1 special forces operators, intelligence professionals, researchers and executives at the world's top AI labs, former cyberoperatives with experience running nation-state campaigns, constitutional lawyers, executives of data center design and construction companies, and experts in U.S. export control policy, among others. Our investigation exposed severe vulnerabilities and critical open problems in the security and management of frontier AI that must be solved – whether or not we launch a national project – for America to establish and maintain a lasting advantage over the CCP. These challenges include:

- **Supply chain security.** The U.S. supply chain for critical AI infrastructure is highly exposed to adversary espionage and denial operations. We rely on Chinese manufacturing for key data center components, many of which can be compromised for

[1] It's possible that superintelligence can't be controlled. We'll discuss that scenario and what we can do about it later.

[2] To be clear: they are likely penetrated by other powers as well. China is noteworthy for being an aggressive U.S. adversary with the technical talent base, energy production, and access to chips (for now) required to be a genuine competitor in AI.

surveillance or sabotage. Onshoring these supply chains will take years – time we may not have. The problem extends to AI chip fabrication and packaging, which is largely done in Taiwan. A national superintelligence project built with existing supply chain components would risk embedding CCP trojan horses deep into the data centers that house some of the most national security-critical technology America will ever build. This isn't a reason not to proceed; there are mitigations, but they come with hard trade-offs.

- **Model developer security.** The "move fast and break things" ethos of Silicon Valley is incompatible with the security demands of superintelligence. Frontier lab security is far from where it needs to be to counter prioritized attacks by America's nation-state adversaries. Insider threats are a big problem: a large fraction of frontier researchers in American labs are Chinese nationals, and the CCP is known to leverage individuals with family, financial, and other ties to mainland China – including U.S. citizens – as intelligence sources. There are also major cyber and physical security gaps that would allow a competent adversary to easily obtain sensitive IP, and compromise a national project.

- **Loss of control.** A key objective of a national ASI project must be to develop AI-powered denial capabilities — [redacted] — that would allow the United States to stabilize the geostrategic environment [redacted]. But an AI-powered denial capability is useless if it behaves unpredictably, and if it executes on its instructions in ways that have undesired, high-consequence side-effects. A national superintelligence project would need to actively assess the promise of our best AI control techniques, and to account for that information when making decisions about which capabilities to develop, and when. But that needs to be balanced with an imperative for speed: the CCP has already announced significant infrastructure investments on the order of a quarter trillion dollars in PPP terms that could be viewed as a first decisive move towards a nationally coordinated ASI strategy. Chinese AI labs are genuine contenders, even absent espionage. The race to superintelligence is already on.

- **Oversight.** Superintelligence would concentrate power in unprecedented ways. Traditional checks and balances weren't designed for this. A national project needs robust oversight to prevent misuse, balancing speed with security while ensuring democratic accountability. We need mechanisms, analogous to the nuclear chain of command, to ensure no single person or group can wield this

> technology unchecked. And we’ll need an economic “offramp” for the technology: a path to ultimately put superintelligence-derived capabilities into the hands of everyday Americans without compromising national security, to support economic growth and empower individuals with the technology.

These challenges create a dilemma. We're not pricing in the military implications of superintelligence, making it hard to justify investing in security now. But security can't be bolted on later. Lead times for essential infrastructure are measured in years. Chip designs can't be reworked overnight. We must either make a bold bet on securing critical AI infrastructure now or find ourselves in a late-stage race to superintelligence with CCP-compromised hardware and software stacks, developed by labs unable, for cultural and economic reasons, to embrace the security culture this technology demands – and potentially without even the ability to control it.

A project like this needs a new data center build, it needs security baked in from the ground up, and it needs to be integrated from the start into an offensive counterintelligence motion that disrupts similar projects by adversaries. As new capabilities come online that can buy America time, they need to be quickly and efficiently folded into offensive activities, while intelligence we collect from adversaries needs to inform the *kinds* of capabilities we seek to develop in real-time.

The window to act is closing fast. Many in Silicon Valley believe we're less than a year away from AI that can automate most software engineering work. Our interviews with current and former frontier lab employees strongly suggest that belief is well-founded and genuine. If they're right, we may need to break ground on a fully secure, gigawatt-scale AI data center within months, and figure out how deal with contingencies. There's no guarantee we'll reach superintelligence soon. But if we do, and we want to prevent the CCP from stealing or crippling it, we need to start building the secure facilities for it yesterday.

Our report offers detailed recommendations, including specific actions that America can take immediately, even without a full national project. Implementing these recommendations will be a big task. It will require unprecedented coordination between government, industry, and the national security community. It will be expensive, disruptive, and politically fraught. But the alternative – a world in which the CCP builds superintelligence first, or in which we can’t effectively control our own – is far worse.

Introduction

America has a historic opportunity to build, secure, and control[3] a world-defining technology: artificial superintelligence (ASI). The right kind of superintelligence project could give the United States a decisive – possibly permanent – advantage over our adversaries, and deliver unprecedented economic benefits globally. But right now, we aren't building the right kind of superintelligence project. Research on the critical path to ASI is happening in badly secured private labs that are easy targets for CCP penetration, is not being strongly integrated with national security efforts, and is cutting corners on the critical technical problem of controlling these systems. All this can be fixed, and it will have to be.

From 2022 to early 2024, our team conducted an investigation of the national security implications of frontier AI development. We shared the first-ever publicly reported statements from frontier lab whistleblowers who were concerned that their labs' security posture was putting U.S. national security at risk. According to lab insiders, leading AI model weights – which are increasingly key U.S. national security assets – were being regularly stolen by nation-state actors. Critical lab IP was protected by flimsy security measures. And executives at some top labs were shutting down calls for heightened security coming from their own concerned researchers. In one case, a researcher told us about a running joke within their lab that they're "the leading Chinese AI lab because probably all of our [stuff] is being spied on."[4]

A few months later, OpenAI fired Leopold Aschenbrenner – a researcher who had written an internal memo criticizing what he considered OpenAI's poor security – allegedly for leaking sensitive information. [redacted]

[3] There are reasons to believe that artificial superintelligence — meaning AI systems that are significantly more capable than humans at all tasks — can't be controlled in any meaningful way. Obviously, whether that turns out to be true should be a huge factor in determining our national security strategy for ASI. We'll discuss this more later in the document.

[4] These concerns were justified. There have been many public reports since then of Chinese efforts – some successful – to spy on American hyperscalers and AI labs. For example, Chinese hacking group "Diplomatic Specter" has been spearphishing OpenAI employees; a former Google engineer stole AI-related trade secrets and siphoned them to Chinese firms; and Chinese state-backed hacking groups were able to obtain a cryptographic key from Microsoft that gave them full access to encrypted sensitive data from cloud servers. The scale of China's state-backed hacking program is "larger than that of every other major nation, combined", and has included infiltration of major U.S. broadband providers, and theft of private data used to train AI models.

[redacted] According to one former researcher, "The government doesn't really know what's going on in the labs, and they're not even allowed to talk to people at the labs [...] Labs are trying to crack down on [technical researchers] that they suspect of potentially talking to the government."

As a result, the U.S. government lacks situational awareness regarding the capabilities of frontier models trained on American soil – increasingly critical national security assets. And unlike the U.S. government and its intelligence agencies, the CCP is free to spy on American companies: at any given time, the CCP may have a better idea of what OpenAI's frontier advances look like than the U.S. government does.

Most U.S. frontier labs have improved their security measures somewhat since we published our report. But as we'll see, these improvements are not enough to keep them ahead of current or future CCP operations.

Soon after his departure from OpenAI, Aschenbrenner published Situational Awareness, an influential manifesto that argued that the current explosive growth in AI capabilities, measured on a scale of "effective compute," suggests that superintelligence may very well be developed before the end of the decade.

AI scaling curves show that as AI models have been trained using ever larger quantities of compute and data, their capabilities have increased steadily and predictably across many orders of magnitude. As AI system training and inference is scaled on larger and larger superclusters, it's likely that the capabilities of frontier systems will continue to increase.

Those capability increases could happen surprisingly fast: AI is already augmenting important parts of the AI research process itself, and that will only accelerate.[5] Soon, AI research will be mostly or entirely automated, with millions of AI agents pushing the field forward at computer speeds rather than human speeds, and getting faster with every exponential increase in computing power. On the way there, we might get warning shots – if we're lucky, shocking demos; if we're unlucky, high-consequence attacks or accidents – that will fully awaken national security agencies to the stakes of superintelligence. At that point, Aschenbrenner argued,

[5] Frontier lab insiders we spoke to have told us that as of February 2025, as much as 50% of their lab's codebase is already written by AI. They expect this number to reach 75% by the end of 2025.

everything will change. Proposals that seem outrageous today will slide across the Resolute Desk — from locking down the labs to outright nationalization of ASI research.

Leopold isn't alone in predicting that we're on course for superintelligence, or that more government involvement in AI is inevitable. Our team has spent the last four years working closely with a network of insiders at top AI labs and U.S. national security agencies to understand what the march to ASI will actually look like — and what it will call upon us to do, if we want the U.S. to come out ahead of the CCP, and the cause of liberty to win through. In that time, the views of many of our contacts in the AI labs have sharpened: as we approach superintelligence, it's become clearer to them that AI is turning into a national security technology first and foremost. To quote a former cyber intelligence operator we interviewed during our investigation, "Dual-use is a nice way to [describe superintelligence], but actually it'll be a weapon for all intents and purposes."

A small number of people in the frontier AI world saw this coming, and they weren't just employees at leading labs. As early as 2020, a technical background and some discussions with the right friends at OpenAI were enough to convince us to dive head-first into AI national security. Even then, it was clear that we'd end up with an industry-wide race to build AI infrastructure, and that the national security implications of AI development were soon going to be impossible to ignore.

To this day, if you know the right people, the Silicon Valley gossip mill is a surprisingly reliable source of information if you want to anticipate the next beat in frontier AI – and that's a problem. You can't have your most critical national security technology built in labs that are almost certainly CCP-penetrated, some of which have a history of shutting down whistleblowers with security concerns. As a former DIU executive told us, "if legislators knew how compromised [redacted AI lab] might be, there would be a significant appetite to shut this all down."[6] Despite deepening partnerships with national security agencies, frontier AI development is leaky: most lab executives aren't pricing in how effective nation-state

[6] He didn't mean to shut down the frontier AI labs, but rather to move the development of high-stakes AI R&D out of highly penetrated companies with huge security vulnerabilities.

espionage can be.[7] Shoring up lab security to the point where we aren't giving away the crown jewels of American technological superiority to the CCP will require fundamental changes to the culture, management, and operations of frontier AI labs.

Right now, the greatest danger is not that the U.S. will fall behind China in the race to superintelligence. Until we've secured the labs, *there is no lead for us to lose*. Just the opposite: as we've seen, U.S. national security agencies don't constitutionally spy on American companies or access their technology illicitly, but the CCP has no such scruples. Under the status quo, therefore, advances at private U.S. labs may lead to advances in CCP capabilities *before* they lead to advances in U.S. national security capabilities.[8]

Apart from security, there's also the question of what chain of command should apply to future superintelligent systems. We're already seeing AI models with meaningful bioweapon design, cyber, autonomy, and even persuasion capabilities. Superintelligence will go beyond that: it will be *the* national security technology. It will deliver a decisive, and possibly permanent strategic advantage to whoever builds and controls it first – and it may not be easy to control. It must be built by, secured in, and controlled from America. And it must be operated in a way that serves the interests of the American people.

[7] One former OpenAI researcher pushed back on this framing as applied to OpenAI, and we're including his perspective here for completeness even though it's not a universally held view. As he put it, "It's not that [OpenAI] aren't pricing this in. It's that they don't care. They know they are going to get away with it; it's not like they are going to get fired for having bad security, whereas if they were to beef up their security and thereby fall behind in the race to AGI, they would lose all their power and chance at glory etc. My guess at least, is that they cynically calculate that the U.S. government will at most slap them on the wrist for having bad security. tbc this is just my guess, I don't have direct knowledge of what they are thinking. But they aren't idiots, they know the CCP has probably penetrated them and could easily do so in the future."

[8] In fact, according to national security professionals we interviewed – many of whom had experience in the IC – under today's conditions, a very real possibility is that 1) A U.S. frontier lab gets noticeably close to building superintelligence; 2) the CCP observes this progress and exfiltrates what it needs to reproduce it domestically; 3) the CCP begins to sabotage the lab's further development efforts, likely in an obfuscated or deniable way; and 4) the CCP hands over the stolen technology to its national champions to finish the work themselves. In other words, while some U.S. frontier labs believe internally that they can achieve superintelligence first by pushing AI capabilities as fast as possible, that belief is quite possibly deluded — in reality, they're more likely to be beaten at the last minute by unscrupulous actors with access to nation-state resources.

If we end up launching a national project, the need for that project will be recognized abruptly, and when it is, we'll wish we'd had more time to think through how to pull it off. For the past year, we've been working to answer exactly that question: how might a U.S. government-backed superintelligence project work? We interviewed dozens of frontier AI researchers, energy specialists, AI hardware engineers, national security executives and action officers, and constitutional lawyers. We visited hyperscaler data centers with former Tier 1 special forces operators to assess the state of the art of data center security. We spoke to dozens of Congressional offices to understand where the Overton window is today, and how far it will have to move to catch up to the decade of superintelligence.

This might seem premature. But if progress simply continues on its current path, we could suddenly find ourselves in a technological and geostrategic crisis – maybe even a hot war – over AI. And if that happens, we'll wish we'd had more time[9] to anticipate contingencies, figure out how to design a CCP-proof project, find out whether it can do what we want it to do – and decide how to run it.

Scope of the investigation

AI is already a dual-use technology. But as we get closer to superintelligence, it will be seen more and more as an enabler and driver of weapon of mass destruction (WMD) capabilities, if not as a WMD in and of itself. Direct calls for a "Manhattan Project for AGI" are already starting.

But it's still not obvious how exactly the U.S. government could or should partner with frontier AI labs on a superintelligence program. Should frontier labs be nationalized outright?[10] Should they be put on contract to support U.S. government research activities? Or should the

[9] Many of the security measures we'll need to pull off a successful superintelligence project – especially those that require technical R&D – will take months to years to implement.

[10] As part of the research that informed our March 2024 report, we spoke to frontier lab insiders and whistleblowers, many of whom were concerned that their labs were brushing off their increasingly critical security concerns. During a lab tour, a researcher discreetly took Jeremie aside. Referring to our eventual recommendations to the U.S. government, he told him that, "I would urge you to be more ambitious," and asked us to recommend that frontier AI research be nationalized. That would have been an extreme measure, particularly at the time. But he wasn't the only one calling for it even then, almost a year before the Situational Awareness memo brought the idea into the mainstream national security conversation.

relationship between the government and the labs be at arm's length, with the government offering incentives – such as funding or infrastructure – to shape frontier AI research?

There's already been some informed speculation about the answers to these questions, but our goal here isn't to get bogged down in debates about the specific configuration of a government-backed superintelligence project. Instead, we're going to focus on identifying the implementation-level problems that a public-private partnership on superintelligence would have to solve in order for it to achieve critical U.S. national security objectives.

There are many ways that a government-backed ASI project could be set up, but few that will lead to a lasting American advantage without inadvertently handing our most important national security technologies to our adversaries. We'll be going after a very small target while wielding a very large industrial and national security machine.

We'll start by introducing some of the main considerations that our investigation has surfaced, sketching out a path to a U.S.-backed superintelligence program.

Data center security

When you tour a HPC AI data center with special forces operators, they ask some weird questions. Who manufactures the fire alarms? [redacted] Where were the critical mineral inputs to the AI accelerators processed? What construction company was contracted to build the facility? The tactics, techniques, and procedures (TTPs) actually available to and employed by our nation-state adversaries are more advanced than hyperscalers, private labs, chip design firms, and data center infrastructure providers can know. That will have to change if a U.S. government superintelligence project is going to succeed.

> *With a [redacted], what you do is you get [redacted], and you knock out the whole [data center component] with [redacted] You just rip it across the [component] and fucking take it out.*[11]
>
> – Former special forces operator evaluating the security of an AI data center housing hyperscaler infrastructure

To train and deploy frontier AI models, you need AI hardware to run computations and move data around, a reliable source of high base load power to keep your hardware running, and cooling systems to stop your hardware from overheating. If you want to truly secure a U.S.-backed superintelligence project, then you need to secure the supply chains for these assets – starting with the data centers themselves.

Data centers are where the critical supply chains for AI chips, power, and cooling come together. They're the facilities in which a U.S.-backed superintelligence project will actually train and run its models. Because of their size and complexity, data centers have lots of human, physical, and cyber security access points and vulnerabilities. Nation states can easily exploit those vulnerabilities to extract sensitive IP or to damage key facilities, in ways that aren't addressed by any known current or proposed future security measures.

In November 2024, we joined a team of former Tier 1 special forces operators and national security intelligence professionals and were given access to a [redacted] data center containing

[11] Obviously this is a very redacted quote. But the attack described could allow someone to knock out a >$2B facility on a sub-$30k budget so we were specifically asked not to disclose it in too much detail. What's worth noting here is how quickly and casually individuals with experience executing nation state-level operations on similar facilities can identify these extreme asymmetric vulnerabilities.

hyperscaler AI clusters, in order to assess its security posture. We later also spoke with individuals directly responsible for the security of hyperscaler-owned and operated AI data centers.

Our goal was to understand the current state of data center security, to determine how it needs to be improved to resist nation-state attacks, and whether we can expect those improvements to happen in time for ASI-grade training runs if they were executed in the 2026-28 era. If we take seriously the prospect that superintelligence will be a WMD and or WMI (weapon of mass influence), then these data centers will become priority targets for the intelligence and special operations of nation-state adversaries and their covert surrogate networks. So in order to understand how to secure a superintelligence project, we need to talk to intelligence and special forces operators directly.

Physical security

Today, security practices vary widely between AI data centers operated by hyperscalers[12] and colocation providers, but some trends are clear. A given data center tends to have a highly uneven security posture across different threat vectors. For example, it's often hard for employees or visitors to take photographs of data halls, but it would be easy to [redacted]. This inconsistent security profile is a problem. Even the most secure windows in the world don't matter if you leave the front door open.

By the assessment of one special forces operator with direct firsthand experience running [redacted] operations in front-line environments, a strike capable of destroying a key component of a hyperscaler data center infrastructure "could be done with [redacted] for under 20k [dollars]. Pending you had all the necessary component parts to [redacted]." When we asked the data center's security lead how long it would take for them to get the facility back online after a strike like this, his reply was "A year?

[12] There's no consistent frameworks for data center security today, so security measures reflect individual companies' tolerance for risk. This is because the security requirements that data center builders and operators follow typically come from contractual obligations to the data center's users and not from any particular set of best practices. As it becomes increasingly obvious that frontier AI models are key national security assets, this will have to change. Executive and/or legislative actions will have to create more consistent and stringent requirements on hyperscalers building and running inference on frontier systems.

Six months at least. [redacted]"[13]

That's a $2B facility out of commission for over 6 months, on a budget of $20-30k. All from a threat that's been completely overlooked by every widely referenced assessment we're aware of. Fortunately, it's also a vulnerability that's easily fixed: according to one special forces operator, the lowest-cost [redacted] variants of this attack could be blocked by simple and cheap add-ons.

But this also wasn't the only critical [redacted] vulnerability the team identified. Many critical components of data centers are on multi-year back-order and are highly exposed even in the most secure builds. As one operator explained, "With a [redacted], what you do is you get [redacted], and you knock out the whole [data center component] with [redacted]. You just rip it across the [component] and fucking take it out." We've been asked not to share the specific components or techniques involved in these attacks precisely because of how significant these vulnerabilities are, and how cheaply they can be exploited.

These low-cost, highly leveraged and disruptive attacks could easily prove decisive in the context of a U.S.-government backed project designed and understood to be aimed at developing weaponizable superintelligence, among other capabilities. In particular, if we find ourselves in a race to superintelligence with the CCP, a cheap attack on an unprotected data center could set our efforts back by months or even years depending on supply chain delays.[14] And although special authorities could considerably reduce wait times for key components, significant downtime for core data center infrastructure could still be decisive in an adversarial race.

Recommendation: At the design stage, **new data center projects that are aiming to support training and inference for frontier models by the end of 2025 should enlist support from individuals with direct experience performing nation-state denial operations,** including but not limited to operations involving [redacted]. These should specifically include red teaming, pentesting, and tabletop exercises supported by current and former special operators to address vulnerabilities introduced by new platforms as they come online during design and construction. This activity should be coordinated with the U.S. government, and involve the

[13] [redacted]

[14] Many of these critical components are themselves sourced from China, which brings its own major vulnerabilities that we'll get into in later sections.

development of a variant of the [redacted] We can provide further information upon request.

The IC and national security community will need to work with data center builders, hyperscalers, and frontier labs to define security measures that can withstand attacks by resourced nation state adversaries. That should include funding or incentivizing fast prototyping build-outs and retrofits of data centers and supporting infrastructure that can withstand those kinds of attacks. Some or all of these projects will probably fail to meet the full level of security they need. Funding for these prototype builds will have to be risk-tolerant the same way venture capital is: we can't let fear of failure hold back critical experiments in data center security.

We can start with small-scale prototypes. For example, an individual GPU or a pod of 8 GPUs with simulated network fabric is something that can be set up very quickly. It can also be red teamed and pentested fast before it's scaled up to a bigger build that fixes the security problems that get discovered in smaller-scale iterations.

Ultimately, perfect security is impossible, and we shouldn't expect to be able to build an impenetrable fortress. But **we should raise the bar for security to the point where our adversaries' attempts to take down key frontier AI infrastructure are attributable**, so that these attacks lead to a credible threat of retaliation.

Hardware supply chain security

According to the AI data center builders and operators we spoke to, the supply chains they'd need to draw on to provision secure physical data center infrastructure are badly strained. Those constraints will impact training runs and deployments of leading AI models in the 2026 era and beyond. Most at risk are components whose supply chains depend heavily on Chinese manufacturers, and which in some cases are produced exclusively by Chinese companies. To take just one example, the overwhelming majority of transformer substations contain components that were made in China and can be used as back-doors for sabotage operations.[15] Indeed back-door electronics are known to have been installed in Chinese-made transformers. If a superintelligence project were kick-started under nominal conditions, unsecured supply chains for AI hardware, as well as electrical and cooling infrastructure could embed physical CCP trojan horses deep into the data centers that house some of the most national security-critical technology America will ever build.[16]

What's more, China announced a $275B investment in AI infrastructure[17] within days of the announcement of OpenAI's Project Stargate. In this context, we'd be naive to expect that China is going to prioritize shipping critical data center components to U.S. AI labs over its own directly competing projects. Therefore we should expect that the lead times on China-sourced generators, transformers, and other critical data center components will start to lengthen mysteriously beyond what they already are today. This will be a sign that China is quietly diverting components to its own facilities, since after all, *they control the industrial base that is making most of them.*[18]

As one example of foreign supply chain dependence, a critical and underappreciated security problem for AI data centers is the supply chain for Baseboard Management Controllers (BMCs). BMCs are microcontrollers that sit on server motherboards and handle key housekeeping tasks, like monitoring temperature, voltage and power flows, detecting memory errors, and monitoring network connectivity. BMCs operate even when the main server they monitor is powered off, and they have a privileged position in the system architecture: among other things, they can directly interface with baseboard hardware, and they can read the firmware that links the baseboard hardware to the operating system. This means that attackers who gain control of the BMC can [redacted]

[15] Other examples include transformers, generators, uninterrupted power supplies, copper cables, etc.

[16] This isn't just a problem for the initial build of new data centers and supporting infrastructure. New components are in constant demand due to the need for continuous maintenance, upgrades, and replacement of failed components.

[17] The actual amount of the investment is 1 trillion yuan. This has been frequently reported as a $137B investment, but in reality it's closer to $275B when measured in purchasing power parity (PPP) terms — since that's a better reflection of what you can buy in-country.

[18] If they're clever about it, they will of course never announce these diversions as a policy — we'll just see lengthening delays and chalk them up to increased demand, competitive bidding for rush orders, mundane shipping problems, and so on.

[redacted][19]

BMCs are a "soft underbelly" of AI server security – and a dream target for adversaries looking to Stuxnet some of our most critical national security assets. Which is why it's especially bad that Taiwan-based ASPEED makes 70% of the world's BMC chips.[20] ASPEED BMCs are all over the place, appearing in every Nvidia DGX server – and, while they're modular in more recent Blackwell systems, they're being widely deployed in hyperscaler AI data centers. We should expect ASPEED to already be a target for the CCP, and for their ASPEED infiltration, compromise, and sabotage operations to be increasingly prioritized as the stakes of superintelligence get more widely recognized. We need made-in-America alternatives.

Beyond BMCs, many of the less "sexy", lower-tech but essential infrastructure components that make data centers hum are key vectors for adversary operations. Programmable logic controllers (PLCs) that power critical power and cooling systems, for example, have more diversified supply chains, but are also under-appreciated attack channels.

Reshoring critical supply chains to America or our allies won't be a matter of months, but of years. We won't have time to do it after we've come to a political consensus around securing a superintelligence project. But whether or not we ultimately choose to kick off a superintelligence project, if we want our infrastructure to be robust to even the most basic of these supply chain attacks, we need to start now.

There have been proposals to screen physical hardware that enters data centers as part of supply chain security. While this helps, it's not adequate against a resourced attacker that controls major parts of the supply chain, as China does. We can't go into details, but we were informed during our investigation that there exist techniques for supply chain sabotage, including for destruction and surveillance, that are [redacted]. As a former intelligence official informed us, the only guaranteed defense is full control of the supply chain from mineral extraction to final assembly

[19] Based on private communication with Seth Demsey.

[20] Notably, ASPEED BMCs are starting to integrate open-source RISC-V cores which have not been properly vetted for security. Open-source software in general has security upsides, since in principle it's auditable by external, security-minded developers. Unfortunately the ASPEED RISC-V BMC has received very limited security scrutiny to date, particularly relative to proprietary alternatives from Arm and Intel. Chinese entities have invested heavily in supporting the rapidly growing RISC-V ecosystem.

and installation of every component. Of course, perfect security is impossible, and security standards have to be balanced against development velocity in the context of a technological race. But the intelligence official who shared this assessment with us was basing it on an overall evaluation of the risk profile of other forms of WMD development: if we're going to treat superintelligence like the weapon that it will be, then that's just the security bar we have to meet.

Recommendation: We need as much production capacity associated with critical data center components – like transformers, power supplies, cables, networking equipment, and cooling components – onshored as possible. We know from conversations with national security professionals that there are ways to doctor components in the supply chain for surveillance or sabotage that [redacted]. If adversaries are manufacturing most of our critical components, that's a problem we need to solve immediately. One possible workaround is [redacted] We need to assess how feasible this would be for each critical component in the AI supply chain so that we can pull the trigger on this option if we determine that onshoring manufacturing of these components will take too long.

We also specifically need a securely sourced alternative to ASPEED BMCs — ideally, a simple, purpose-built and easily auditable drop-in replacement for existing chips. DARPA projects focused on developing secure BMC hardware could be a helpful step towards a solution, but government purchase guarantees could drive demand in tandem. To mitigate risk in the shorter term, [redacted]; insist on having suppliers share information about their sourcing of BMC chips; and introduce requirements for supplier diversity, including at the firmware level. [redacted]

These would be big — and expensive — moves, but they reflect the scale and impact of the vulnerability they're meant to address. BMCs offer our adversaries a way to conduct crippling attacks on our most important compute infrastructure. We can share extensive further recommendations if desired.

Side-channel security

Adversaries will almost certainly aim to knock out the American data centers powering training runs for a superintelligence project. But to secure the project, we'll need to do more than prevent adversaries from destroying AI infrastructure – we'll also need to prevent them from **stealing critical IP**.

A widely-cited report by AI security researchers published in 2024 provides a reasonable starting point for understanding what it would take to secure AI model weights from theft by nation-state attackers. It defines AI systems requiring Security Level 5 (SL5) as systems that would become targets for "top-priority operations by the top cyber-capable institutions" like the People's Liberation Army's Strategic Support Force, China's Ministry of State Security, and analogous institutions in other adversary countries. Superintelligence could offer those who control it a decisive strategic advantage, and would almost certainly need to be secured at the SL5 level. The recommendations at this level of security are extensive. They include:

- Strict limitation of external connections to a completely isolated network
- Routine rigorous device inspections
- Disabling of most communication at the hardware level
- Extreme isolation of AI model weight storage (completely isolated network)
- Advanced preventive measures for side-channel attacks (e.g., noise injection, time delays, and other tools)
- Formal hardware verification of key components

To quote one former Mossad cyber operative, achieving the proposed SL5 standard would be "really hard. Even people who've read the report and understand cybersecurity underestimate how hard this would be." In practice, SL5 would require U.S.-made components, or highly trusted components that the Department of Defense or the Intelligence Community (IC) would be comfortable using for extremely classified operations.

But as we mentioned earlier, many of the components in current AI clusters are sourced directly from China. As one cyberoperative put it, current clusters are, "not even close to being close. No one is even thinking about this. This is not a priority for them. [...] AI companies have funds, but not for this; people who are smart, but not on this topic. It's just not in their skillset. [...] Government could do this, sure, but this is a big project; this is a mega-project. I'm not [aware of] ultra-secure projects that would be this size, and require this much specialized equipment at the drop of a hat."

And the report's recommendations aren't even comprehensive: they leave out key security challenges associated with geographically distributed training, for example. They also only address the risk of model weight *theft*, and not attacks meant to disable or destroy computing infrastructure, which, as we've seen, would be extremely cheap and damaging under current conditions.

These aren't long-term problems either. An AI security expert who collaborates frequently with frontier AI labs estimated that the next generation of AI models could already qualify for the SL4 threshold – defined in the report as security measures required to counter "most standard operations by leading cyber-capable institutions." By his assessment, even SL4 security measures would take at least 1.5 to 2 years to implement under current conditions, requiring a "massive investment" from key players. This individual assessed that the shift from SL4 to SL5 security would also be significant, almost certainly requiring entirely new data center facilities built from the ground up – a perspective echoed by several other physical and personnel security experts we spoke with.[21]

Our interviews with energy and compute infrastructure specialists, intelligence experts, and special operations personnel have surfaced other critical challenges. For one, data center construction can take well over a year to complete, and data centers probably can't be retrofitted to meet these criteria after they've been built. Similarly, the AI accelerators that will fill frontier AI data centers in 2026 and beyond are having their designs finalized. They will have to be redesigned to address accelerator-level vulnerabilities that will be relevant in the geostrategic and security environment in coming years, but which aren't on leading players' radar today. Absent unprecedented government incentives driving a major security push in the very near term, secure AI accelerators will not be available in time for a superintelligence project.

One key attack vector is **side-channel attacks**, which include things like Telecommunications Electronics Materials Protected from Emanating Spurious Transmissions (TEMPEST) threats.

[21] By contrast to SL4 and SL5, which are security thresholds intended to be capable of thwarting nation-state operations, SL3 is in principle designed to defend against operations by non-state actors such as cyber crime syndicates, terrorist organizations, or insider threats. The SL3, SL4, and SL5 thresholds are widely referenced in the AI security community, but intelligence officials we spoke to made it clear that in practice, non-state actors are often trained or even backed by states, and can use advanced TTPs that most analysts would associate with "nation-state" capabilities as a result. So in practice, and contrary to conventional wisdom in the AI security ecosystem, dealing with threats from non-state actors will require many measures drawn from the SL4 and SL5 standards.

These are information extraction attacks that use leaked electromagnetic, sound, or vibrational signals to reconstruct sensitive and intelligence-bearing information.

It's been argued that TEMPEST attacks would not be able to exfiltrate model weights themselves due to their limited bandwidth, but that they could be used to extract decryption keys or other metadata that makes model weights easier to access or reconstruct. However, our investigation suggests that TEMPEST technologies may be far more advanced than is widely recognized. [redacted] We assess that it is plausible that some adversary nation-states have access to TEMPEST and other side-channel attacks that could allow for direct weight exfiltration under certain conditions.[22]

Although many are classified, some standard defenses against TEMPEST attacks include ensuring a minimum distance between HVAC conduits and sensitive equipment, using shielding and Faraday cages, implementing physical security control zones around sensitive areas, using fiber optic cables instead of copper cables where possible,[23] and air-gapping critical components. Many of these strategies have to be accounted for at the design stage of a new data center build; it won't be possible to retrofit existing facilities to meet appropriate TEMPEST standards. Failing to do so would bake critical vulnerabilities into a government-backed superintelligence project.

Recommendation: At the design stage, and throughout the build, **new data center projects aiming to support training and inference for frontier models by the end of 2025 should work directly with intelligence and national security agencies to identify and mitigate vulnerabilities from TEMPEST and other side-channel attacks.** This work should also involve current or former frontier AI researchers, hardware design specialists from leading companies like NVIDIA/AMD, and data center infrastructure design experts.

Only the U.S. national security and intelligence communities, with support from private sector experts retained at leading chip design, data center, and AI model development companies,

[22] As one indication that we can share: we were shown a live demonstration of an attack that allows hackers to reconstruct the architecture of a small AI model using nothing but the power consumption profile of the hardware that runs it.

[23] However, our discussions with former cyber operators suggest that even dark fiber connections between geographically separated data centers are highly vulnerable to established nation-state attacks, the details of which are not public.

are in a position to estimate the actual extent of the TEMPEST and cyber threats that a superintelligence project would face. And this holds for vulnerabilities beyond TEMPEST attacks too: there almost certainly exist highly classified methods, technologies, and tools available to adversary nation states whose capabilities will have to be accounted for from the earliest design stages of a superintelligence project.

There is no other option: the IC and national security community need to be involved in project design and in the design of its supporting infrastructure – *including of the AI accelerators themselves* – from the very beginning. As one former military security specialist put it, "I wouldn't consider [AI accelerators nation-state proof], unless the NSA or CIA is involved in designing them."

Software supply chain security

The software stack that frontier AI labs use, and that runs on AI data center infrastructure, is insecure. According to a cybersecurity specialist with experience executing nation state cyber operations, "There are at least several incidents you could point to with libraries that are used by leading labs that have been compromised in the past, whether by the intelligence communities of other countries or accidentally." In addition to these known vulnerabilities, nation-state adversaries have almost certainly identified zero-day vulnerabilities.

In this respect, the CCP is one among many potential threat actors. Although China's physical access to sensitive supply chains makes it the leading threat actor along that vector, Russia has shown itself to be highly capable in the cyber domain. And while Russia lacks the AI hardware and talent it would need to train frontier AI models, they may have enough infrastructure to *run* a frontier model if they steal one in a cyber operation. In a world of compromised frontier lab security, **cyber powers *are* AI powers.**

> *No one is doing this, no one is looking into this, no one is trying to create an end-to-end secure software stack that is secure – not just in quotation marks – but genuinely secure.*
>
> – Former IDF cyber security operator on frontier AI lab cyber security

These challenges create a dilemma: nation states aren't pricing in the military implications of superintelligence, which makes it difficult to justify investing in basic security measures today.

But lead times for energy infrastructure components are measured in years, chip design cycles are even longer, and security features can't be retrofitted into hardware, so delay is not an option. We can either place a bold bet on securing our critical AI infrastructure today, or find ourselves racing the CCP to superintelligence with an easily-penetrable hardware and software stack.

Recommendation: There are lists of software libraries that are approved for use in military systems at various levels of classification.[24] **We should open up and expand this process for deep learning-supporting software libraries used by frontier labs, and resource it so it can move as fast as it needs to**. This probably also requires collaboration between the research engineers and the accreditation orgs, as well as the involvement of the intelligence community. The speed of this review process should be explicitly balanced against the need for rapid development, with the understanding that a single critical breach could cause the project to fail.

Energy security

Energy production is an often-cited bottleneck to frontier AI development, and increasing high base load power availability on the U.S. electric grid should be a top national security priority. We won't rehash the details here since there are many public reports on this already, other than to say that our interviews with energy specialists suggest that new natural gas plants capable of supporting hundreds of megawatts currently take 5 to 7 years to set up, but could be built in just over two years – quickly enough to support training runs of the 2027–2028 era – if permitting and supply chain bottlenecks were removed.[25]

The speed at which new natural gas plants can be brought online (assuming we remove regulatory frictions) makes them a particularly promising way to support a U.S.-backed superintelligence project in the near term. Other options like nuclear and geothermal matter more on longer timescales.

[24] For example, Iron Bank sets and verifies standards for hardened containers used in software development for sensitive DOD applications.

[25] This is worth emphasizing: it means that regulatory hurdles are currently the primary reason that new natural gas capacity could not be brought online in time to support a 2027/28 superintelligence project.

Gas turbines are already sold out across the market, and other components are on long-term back-order. This won't be fixed in time to make a difference for the training runs of the 2027 era. U.S. government policy can move the needle here though, as we'll explain below.

From speaking with cabinet officials in U.S. states, a less recognized problem is the time cost of litigation over environmental concerns. Litigation creates a huge bottleneck to new power and infrastructure build-outs. Strategic economic interference by China is an important factor here. We've heard from former counterintelligence and special operations personnel that Chinese entities are explicitly funding and leveraging domestic environmental groups to tie up data center and other critical national security builds through litigation. The environmental groups involved in this work are not generally aware of the ultimate source of their funding — the funding is often indirect and funneled through proxies and surrogates — but effectively operate in alignment with adversary interests.

Recommendation: We know it's possible to build *massive* amounts of power generation quickly — up to 3-5 GW or more[26] — as long as we're willing to ignore carbon emissions concerns in the short term and just set up huge gas turbines.[27] In terms of energy, we're more bottlenecked by our own regulations than by economics or physics. In practice, we could pick a site for a project, exempt it from regulations, and probably build as much power as we need there quickly. To help resolve supply chain problems, [redacted]. Over time we could even phase out gas turbines and replace them with lower carbon sources, but there's no need for that to delay other parts of the project.

Deregulation is essential, and plenty of people are already discussing it. Environmental regulations are a particularly big problem here, and the President could invoke the DPA Title III

[26] Enough to power about 2-3 million American homes, or the state of Utah. It's around 1-2% of the power generation capacity of the entire United States. At this scale you can power about 2 million H100 GPUs (depending on various assumptions), equivalent to at least 10x more compute capacity than we believe exists at any one site as of January 2025.

[27] The Department of Energy didn't actually think this was possible when we spoke to them in December 2024: they told us there was no way to pull together more than 1 GW of capacity in any one place over the next couple of years. But in fact it's exactly how Elon Musk was able to get xAI's Memphis data center up and running in record time. The 3-5 GW number was also confirmed to us by a couple of people who are actually working on deploying those gas turbines. Our best guess is that the DOE team we spoke to must have been assuming only low-carbon energy sources.

to provide NEPA exemptions[28] to advanced AI clusters and related dedicated energy infrastructure that meet various security requirements. We also need categorical exclusions to NEPA for no-impact preliminary activities on these data center projects, like design, drilling boreholes, and collecting soil samples. This would unlock early DOE financing through loans.

We should direct the creation of limits on litigation options for designated critical infrastructure projects, and in particular, expedited court decisions. **Every delay mechanism that could be used to tie up new data center and related energy projects should be bounded.**

Export controls

Just as we need to ensure robust supply chains for our own efforts, we need to deny the same to our adversaries. While U.S. GPU export controls have strengthened over the past few years, Chinese firms are known to have subverted controls to acquire controlled chips manufactured and packaged by TSMC by simply spinning up subsidiaries and shell companies that don't appear on entity lists. To take just one example reported by SemiAnalysis, leading Chinese fab SMIC operates two factories linked together by a wafer bridge, which effectively makes them a single facility, yet they were treated as separate entities under then-current U.S. export control policy — one was included on the U.S. entity list, and the other wasn't.

Recommendation: We'll have to move from entity blacklists to selective entity whitelists, significantly limiting the access that photolithography and optics companies, chip fabs and design firms currently have to the Chinese and other markets. We'll also need to pursue less targeted export bans: current highly targeted approaches aimed at restricting chips based on narrow technical specifications still allow China to import significant volumes of strategically critical AI processing hardware. Recent updates to export control rules have tried to address this problem, but have left major loopholes.[29] China has repeatedly demonstrated their ability to integrate and use heterogeneous and legacy hardware into significant training clusters, and

[28] However we should note that according to a White House official with experience writing executive orders that include NEPA exceptions, they can be legally challenging.

[29] For example: recent AI diffusion regulations limit certain non-allied countries to importing no more than 50,000 advanced GPUs. However, chip orders of under 1,700 GPUs do not count towards this cap. This creates an obvious loophole that is precisely aligned with the demonstrated strategies and capabilities of the CCP, which has historically imported large quantities of otherwise controlled AI chips via a network of subsidiary companies, often spun up specifically for that purpose.

it is unlikely that even the best-informed technical assessments available to the BIS (the U.S. office that manages export control compliance) will be able to anticipate the workarounds they'll develop to any set of narrowly scoped controls.

AI hardware security

A national superintelligence project would need to secure a lot of different supply chains, or [redacted]. But in addition to taking a "wide" view, and solving for security across many devices and pieces of infrastructure, it would also have to take a "deep" view, and consider security exposure from geographically distributed supply chains for logic and memory. If the project depends on chips that are being made in Taiwan and South Korea, then securing and defending TSMC and SK Hynix becomes America's de facto responsibility. All it takes is for the CCP to hack into the right network and modify the firmware that goes on fabricated chips, and the project can be compromised.[30]

In terms of AI GPUs — the specialized processors that actually train and run AI models — an obvious big concern is that most AI GPUs are made by TSMC in Taiwan, and that China claims Taiwan as its own. There's a long history of illicit technology transfer from TSMC to mainland China, and TSMC is likely heavily infiltrated by CCP spies and saboteurs.[31]

There are a lot of hard steps involved in building a GPU, and most of those steps happen at TSMC facilities in Taiwan today. Two of the most important are *fabrication* and *packaging*. Fabrication is where you use lasers and chemicals to etch the tiny circuits into the chip[32] that implement its logic. Packaging is where you combine chips together on one board in such a way that they can talk to each other and with other components like memory and power.

[30] There may be an alternative, however. Trustless computing schemes – such as flexHEG-based devices and protocols – may allow us to use certain hardware components securely even if the components themselves have been tampered with in certain ways. This is one reason why secure computing should be a top national security priority: currently, all frontier AI logic dies are fabbed by TSMC on Taiwanese soil. We should not base our national security strategy on the assumption that the CCP has failed to achieve what must be one of its top national security priorities: infiltrating and compromising TSMC.

[31] SMIC, China's leading semiconductor fab, was founded by Richard Chang, a former senior executive at TSMC. In a series of 2003-2004 lawsuits, TSMC alleged that Chang, as well as several other early SMIC employees, stole critical TSMC IP and used it to massively accelerate SMIC's development.

[32] Really this is called a die, not a full chip. When we talk about fabrication here we mean the step that goes from wafers to dies.

We could fabricate quite a lot of AI chips in America today if we wanted to. TSMC's Fab 21 in Arizona alone likely has the potential to produce about 11.5 million H100-equivalents per year today[33] (a type of GPU),[34] which is already much more than you could power in a 5 GW cluster.[35]

The problem today is *packaging*. Right now TSMC's entire packaging capacity for AI GPUs is equivalent to about 11 million H100s per year,[36] although they're expanding that quickly. Unfortunately almost all of that packaging capacity is located in Taiwan itself.[37] Intelligence experts have warned us privately that there's a significant amount of low grade sabotage being conducted by adversaries on U.S. soil *today*, but it's even easier for those adversaries to disrupt or block supply chains that originate in Taiwan — and harder for us to harden them — than if they were based in America. In the event of a CCP invasion of Taiwan, packaging capacity constraints – in addition to the collapse of logic die fabrication capacity – could quickly become critical.

[33] As of early 2024 Fab 21 was expected to run about 20,000 wafer starts per month through the facility, at yields similar to those at TSMC's other facilities which have been reported elsewhere as around 80%. That means they're getting 20,000 x 80% = around 16,000 usable dies out the other end each month, or around 192,000 dies per year. NVIDIA and TSMC likely get around 60 H100 chips out of each wafer, which ballparks to around 11.5 million H100s per year if all of Fab 21's capacity were dedicated to doing just that. Right now Fab 21 is mostly making iPhone chips.

[34] We're discussing H100s here, rather than the newer B100/B200 GPUs that are rolling out now. This is to get a conservative lower bound on what's needed. The B100/B200 series can do more computations and also uses less power per computation than the H100 does.

[35] One H100 GPU consumes about 1340 W in the common HGX configuration, and data center power efficiencies typically range between 75-90% (1.1-1.3 PUE). So 11.5 million H100s consume at most around 11.5 million x 1340 W x 1.3 = about 20 GW.

[36] TSMC's Chip-on-Wafer-on-Substrate (CoWoS) packaging capacity was expected to be around 15,000 wafers per month or 180,000 wafers per year by the end of 2024. If we assume 60 H100 chips per wafer, that means almost 11 million H100s can be CoWoS packaged across all of TSMC's facilities.

[37] Fortunately, that is starting to change. TSMC and Amkor recently announced the development of two packaging facilities on U.S. soil, one of which will be located in Arizona and will have a capacity of 14,500 12-inch wafers per month (the second facility's capacity is not yet publicly known). Construction of the Arizona packaging facility is only expected to be complete in 2028, however, by which time it would represent ~10% of TSMC's target 2028 capacity of 150,000 wafers per month. This timeline is too long to address the risk of a collapse in America's advanced packaging supply in the event of a CCP invasion of Taiwan pre-2028.

Upstream of TSMC, there are also vulnerabilities at the level of the photolithography supply chain, which makes critical components TSMC uses to etch nanoscopic circuits onto AI chips. That includes the ASML/Carl Zeiss complex of companies that manufactures the machines that use intricate ultraviolet light sources and optics to pattern circuits onto the chips, and the specialized optical components that go into those machines.

Although it can take years for a new generation of photolithography machines to start being used for high-volume manufacturing of leading edge chips, existing photolithography machines are so complex that they need a full-time team from ASML just to maintain them in good working order. That means adversaries who wanted to limit our access to advanced AI chips could also try to do it at the level of ASML — although that scenario matters most if we expect to achieve superintelligence later than 2027 or 2028.

Recommendation: We should prioritize localizing CoWoS and CoWoS-like packaging, including CoWoS-L and CoWoS-R, in the United States, possibly by extending purchase guarantees to domestic companies like Intel and Amkor.[38] We also need to make sure that the components needed to manufacture chips, such as photolithography machines, are secured. This could involve closer collaboration between the United States and Dutch, Japanese, German, and other governments on counterintelligence, counterespionage, and countersabotage.

We also need to accelerate work on on-chip security technologies like confidential computing at scale, which are going to be critical to any effort to secure frontier model weights. We should be funding and incentivizing chip fabs and design firms to do that work directly, and engaging national security agencies in pinning down the security features that are most badly needed based on the current threat landscape.

A key priority also needs to be establishing [redacted]

[38] Intel already has at least 5000 wafers per month of CoWoS-like packaging capacity (via their Foveros process) — much of which is already located in North America — and is already using it to package NVIDIA H100 GPUs. This is equivalent to 60,000 wafers per year or over 3.5 million H100s at 60 H100s per wafer.

[redacted]

To be clear: some of these infrastructure pieces might not matter at all if we reach superintelligence within the next two years or so. If we get superintelligence that fast, then we probably have enough AI chips in the United States today to train a superintelligence already. And that means that ensuring a domestic supply of AI chips today matters somewhat less, although denying AI chips and developing retrofittable secure computing add-ons (e.g. flexHEG) still matter a lot.

Finally, if all this turns out to be overblown and we're actually decades away from superintelligence, it's important to realize that all these infrastructure expenses won't have been wasted. Even if AI never gets any better than it is today, there is already massive demand for AI inference that can be served more cheaply by bigger data centers, more chips, and more power. At worst, we end up in a situation similar to the dot-com bust, where the fiber optic infrastructure that got built during the boom ended up driving the growth of the modern Internet.

AI model developer security

Frontier AI models are trained and deployed on data center hardware. But the algorithmic insights, data collection and preparation procedures, and other critical conceptual breakthroughs that make those AI models more powerful are developed and honed inside the frontier AI labs themselves. These AI labs are tempting targets for nation state espionage. Most of them are also very poorly secured.

The conceptual insights that frontier AI labs are developing today will be used in the AI training runs of the 2026 era and beyond. Current and former frontier lab researchers we interviewed generally believe that some individual algorithmic insights could allow adversaries to train frontier AI models ten times more efficiently — potentially unlocking billions of dollars of efficiencies for very large models[39] — and that those insights can be communicated in a ten-minute conversation between experts.[40]

As with the physical supply chain for frontier AI we discussed earlier, the decisions we make about AI model developer security today will shape our adversaries' capabilities in 2026 and beyond. But right now, frontier labs can't make the right decisions about security because they aren't receiving regular threat intelligence updates from the national security and intelligence communities.

Recommendation: As a first step, **the U.S. government should share threat intelligence updates with frontier AI labs on a regular basis**. It should also develop a mechanism to quickly declassify information that could contribute to labs' security postures. Right now, the U.S. government lacks situational awareness about the true extent of labs' frontier capabilities, but frontier labs also aren't read in on the full range of threats *they* currently face.

[39] Especially if you include the capex cost of building a large AI training cluster in the first place. If using open-source algorithms lets you train an o3-level model for $10 billion including capex, then stolen proprietary insights could let an adversary train it for $1 billion. (The numbers are hypothetical.)

[40] Aschenbrenner claims the same thing in Situational Awareness, but it's now also obvious from advances like DeepSeek-V3. DeepSeek implemented three new but conceptually simple techniques (DeepSeekMOE, multi-head latent attention, multi-token prediction) and those boosted their compute efficiency by >10x. These are exactly the kinds of advantages that American AI labs develop internally. It's great that DeepSeek open sourced those techniques for us – but if we want an enduring lead, we shouldn't return the favor.

On the other hand, security measures can stifle progress and reduce R&D velocity. Falling behind the CCP is just as bad whether it's because our labs' security is so poor that their proprietary advances can be stolen, or because we hamstring our own researchers with security protocols that weigh them down unnecessarily. **We'll need to invest in convenience tooling for ultra-high-security AI development that minimizes the workflow disruptions introduced by new layers of cyber and hardware-level security that we'll have to introduce to the advanced AI R&D stack.** This work should be led by the private sector, but supported – and potentially subsidized – by the U.S. government. We won't go into details here, but there are mechanisms through which this could be done efficiently, and quickly enough to get workable prototypes online within the next 12 months. We're happy to brief these in person if requested.

Personnel security

Personnel security at the frontier labs is a major challenge. The Chinese Ministry of State Security (MSS) targets ethnic Chinese as a matter of highest-level national policy,[41] and routinely exploits them as a source of intelligence. It does this by applying systematic pressure to individuals via their family, financial, and other ties to the mainland, to encourage them to collect and report commercial and other intelligence. This is of direct concern to U.S.-based frontier AI research and development: according to a senior researcher at a leading U.S.-based frontier AI lab, "In terms of percent [of frontier lab researchers] that's foreign born, it's well over 50%. […] A pretty high portion of lab employees are [specifically] Chinese nationals."

CCP espionage activity and penetration in the United States — particularly that target Chinese nationals — is extensive and highly systematized. For example, a former senior national security executive shared the following story: "While I was at [redacted] there was a power outage up in Oakland that affected Berkeley. There was a dorm that lost power and most of the campus lost internet connection. And the RA of the dorm ended up talking to one of our CI activities and all of the Chinese students were freaking the hell out because they had an obligation to do a time-based check-in, and if they didn't check in, like, maybe your mom's insulin medicine doesn't show up, maybe your brother's travel approval [doesn't come through]. […] The fidelity with which they control their operatives is super high." The impact of

[41] While public sources have tended to emphasize Chinese nationality over ethnicity in terms of MSS targeting, several former defense and intelligence officers we spoke to were very clear that Chinese ethnicity was being prioritized. This is also consistent with some of Xi Jinping's recent statements.

this kind of activity on AI is direct. A former senior intelligence official we spoke to told us that there are non-public cases in which unpublished Google research has been acquired by the Chinese, and made widely accessible to university students in China, for example.

According to intelligence professionals with direct experience countering these activities, the CCP applies pressure through a well tested escalation playbook: from what we've been able to gather, this typically starts with minor punishments such as limiting healthcare or career opportunities to relatives on the mainland, and ratchets up systematically to confiscation of personal assets, imprisonment or disappearance of relatives in China, and physical intimidation on U.S. soil using infrastructure they've established for that purpose. The uncomfortable truth is that the CCP applies this pressure not just to their own citizens, but also to ethnic Chinese abroad – including some U.S. citizens of Chinese origin – particularly if they have family or financial ties to the Chinese mainland. Of course, Chinese researchers have been invaluable to American tech for generations, and to frontier AI in particular today — and their contributions will continue to help America compete with the CCP. But any American superintelligence project must be designed in a way that recognizes the immense leverage that the CCP can exert over researchers whose loved ones or assets are located within China's sphere of influence.

The United States will face significant challenges in addressing insider threats from the CCP and other well-resourced actors. These will include constitutional roadblocks, and the fact that some of the most technically capable talent employed by U.S. firms may itself be compromised. That forces a direct trade-off between security and technical bench depth at all levels of the AI supply chain. We will need to put in place stringent security and counterintelligence measures if we're going to stand a chance of limiting critical IP leaks to the CCP and other adversaries. These measures will probably reduce development velocity, and their impact will be felt most acutely within frontier AI labs themselves.

Frontier labs face strong incentives to prioritize speed over security. This is where Silicon Valley's edge has come from, historically: competitive pressure and rapid, open iteration has propelled American tech forward in an environment with capped downside risks (at worst, a company folds), and near-unlimited upside potential. In that setting, even labs and researchers who consider themselves security-conscious relative to their peers can significantly underestimate the CCP's motivation and capability to extract sensitive IP.

The result of all this is a deeply rooted culture of bad personnel security. Real personnel security for a project developing potential WMD capabilities looks like intense vetting scrutiny, including detailed inquiries about family locations and extensive travel histories, all rigorously cross-validated. It looks like continuous monitoring of personnel and their communications, employee atmospherics and trends, up to and including handing over personal phones and laptops for review. It calls for insider threat programs, active surveillance, and other privacy-violating activities. These things slow down research and frustrate researchers. But they are some of the basic personnel security procedures that would apply to any government-backed project expected to produce WMD capabilities.

These security measures are all the more critical in Silicon Valley's AGI scene, where naive researchers recommend and hire based on (often publicly legible) social networks and with little to no accounting for national security factors. The social fabric and culture of Silicon Valley's leading AI labs makes them particularly vulnerable to CCP infiltration and HUMINT gathering. We've seen firsthand how freely sensitive information flows in these circles: our team — at the time, independent researchers — were receiving updates on proprietary frontier AI training runs as far back as 2020. Leaks remain extremely common.

Recommendation: A national superintelligence project will have to make hard choices about who to include in its security umbrella, and about what to do with AI researchers and engineers who don't pass the security bar. There are executive authorities and other tools that can help with this, [redacted] (we can provide in-person briefings if requested).

A large fraction of the world's best AI researchers come from abroad, so we'll have to find a way to vet, clear, and hire top talent from other countries. Aggressively expanding the Limited Access Authorization program – which helps U.S. national laboratories hire foreign talent – could be a way to do this.

We'll also need a security clearance process that's *better* than the U.S. government's ordinary clearance process. Multiple national security and intelligence professionals have told us that the existing process for Top Secret clearances is too slow and in any case fails to leverage all the information it should — even failing to use unclassified information from OSINT, ADINT, and commercially available sensors.[42] The key question then becomes how the government can certify these private sector clearances and link them to access and participation in the project. This is probably doable through an executive order, which should

[42] The existing clearance process doesn't use dark web data, phone number analysis, ad IDs ("ADINT"), native language checks using foreign sources, or even much publicly available OSINT. It also doesn't necessarily use all the data that's available to the government *internally*. As one expert in background check systems told us, "a criminal record in one county may be completely unknown if the individual moves across the country."

also authorize frontier labs to hire and let go of employees on the basis of these background checks (to avoid wrongful termination lawsuits).

The personnel risk here isn't just about securing model weights from theft, or even protecting data center facilities from sabotage. As we mentioned above, there are algorithmic trade secrets that can be conveyed in a ten minute conversation between experts, and that if implemented can save billions of dollars and months of experimentation on the path to superintelligence.

We already know from dozens of sources that the current state of security at frontier labs is not up to these challenges. But a former defense official with extensive counterintelligence experience shared his assessment with us that all these labs are almost certainly severely penetrated by the CCP. He added that if legislators fully understood the potential for compromise of these labs by foreign nationals, there would immediately be a much stronger push for tighter controls. In our assessment, we should assume that the Chinese government knows all the algorithmic secrets that all U.S. AI labs know today — those secrets clearly aren't hard for them to steal under current conditions — and that those secrets have been disseminated to their own national AI champions.[43] But that's a bad reason to keep leaking all the algorithmic insights that we'll develop in the future too. We'll have to clear any personnel working on the project, and turn to a combination of public and private solutions to improve that clearing process, as discussed above.

Cybersecurity

Cybersecurity is another core challenge for a superintelligence project. The software libraries leading AI labs use have been compromised in the past, and frontier AI researchers depend heavily on insecure open-source software to move as fast as they do. Software development tools, including the coding assistants researchers use and the data they're trained on, are potential attack vectors as well.

During our investigation, we spoke to one former OpenAI researcher who claimed to have been aware of several critical security vulnerabilities, including one that allowed users to

[43] We have no specific information regarding which companies these would be, but they likely include Huawei, Alibaba, Baidu, Tencent, and ByteDance and probably also some smaller companies like DeepSeek and 01.ai.

perform [redacted], such as granting access to the weights of the lab's most valuable AI model — their crown jewels. The researcher was later also made aware of an additional vulnerability discovered by red teamers that had been logged in the lab's internal documentation and Slack channels, but left unaddressed. By his assessment, this second vulnerability would have allowed any employee to exfiltrate model weights from the lab's servers undetected. Together, these two vulnerabilities made it possible to (1) access; and (2) extract some of OpenAI's most sensitive IP, worth billions of dollars. He told us that these issues had been flagged for security teams and management, but that they went unaddressed for months.[44] To this day, he has no idea whether a fix has been implemented.

According to several researchers we spoke to, security at frontier AI labs has improved somewhat in the past year, but it remains completely inadequate to withstand nation state attacks. According to former insiders, poor controls at many frontier AI labs originally stem from a cultural bias towards speed over security. This is a major challenge we'll face in securing AI labs for a superintelligence project. Most AI labs have a deep cultural aversion to anything more than token personnel or cyber security measures – an aversion that's sometimes seeded and reinforced by commitments to transhumanist or other ideologies.

It's worth underlining what happens when an adversary steals cutting-edge AI model weights. If an adversary steals the weights of a model you're developing, any lead you had in base model pretraining[45] instantly goes to zero — or even negative, if they have the benefit of your discoveries in *addition* to their own.[46] (Code and training techniques can be just as critical, at least until AI systems are doing most of our AI R&D work themselves.)

Recommendation: The NSA should be working directly with all frontier labs to accelerate their software and cyber security, alongside private sector red teams. This should include very robust pentesting activities. Ideally, it would also involve support from private security firms with experience securing sensitive national security sites.

[44] And of course, if news of these specific vulnerabilities worked its way to us, then adversaries like the CCP were likely aware of them as well.

[45] Thanks to inference-time scaling, we could still retain an overall lead in AI capabilities if we have a larger stockpile of inference compute than our adversaries. But if the infrastructure we use to perform inference on national security-critical models is compromised, adversaries could monitor what we're using those models for, and potentially interfere with their behavior. Having a larger inference-time compute stockpile is good, but securing that infrastructure is still essential.

[46] Your lead is also negative in another sense: after they've stolen your model, they still have their full opex budget available to spend on augmenting or supplementing the model's capabilities.

People have already started working on determining what it would take to stop high-capability adversaries from stealing model weights. We won't rehash the details here, other than to say that while these efforts are commendable, they focus mostly on preventing theft and not denial attacks[47] (like drone strikes, missile strikes, physical sabotage, or certain cyberattacks). And so far, the IC and broader national security community haven't taken a leading role on this; that will have to change quickly. The most significant vulnerabilities that data centers and other critical AI infrastructure have to nation state-level attacks can only be identified and mitigated by intelligence, defense, and national security agencies with access to sensitive data, as well as deep domain expertise and experience carrying out and defending against adversary operations.

If we only recognize the need for a superintelligence project relatively soon before superintelligent AI is achieved, then we'll have no choice but to urgently retrofit key facilities to higher levels of security. In that scenario, we'll need especially deep and ongoing integration between U.S. national security agencies and the teams involved in the retrofitting process, because a lot of the tradeoffs between speed and security are extremely context dependent. If we have to make those trade-offs, we should make them deliberately, and they should be made by people who have full access to informed assessments of the strategic implications of various security and infrastructure design choices.

Finally, if our adversaries do manage to steal the weights of superintelligence-grade models, we'll want to make sure that we maintain a decisive lead on centralized inference-time compute capacity. Even if adversaries exfiltrate critical model weights, we may still be able to achieve better performance than they can from the same model if we can dedicate more FLOPs to inference than they can.

Emissions security

In a context where multi-billion dollar technical secrets can be leaked in a five-minute conversation, IP security becomes a critical problem. Parabolic and laser microphones, piezoelectric sensors, and other surveillance technology can be used to eavesdrop on sensitive

[47] Note also that denial attacks can be the first step of an IP extraction attack: if an adversary destroys our secure R&D facilities, we may be forced to relocate critical work to less secure facilities that are more vulnerable to IP theft.

conversations easily. This is particularly true when espionage targets are in offices in densely populated urban areas, because adversaries can hide their surveillance operations in the city noise and disguise them as legitimate business or urban activity. A U.S. government-backed superintelligence project would have to move its R&D work — including algorithm development — to facilities that we can secure against sophisticated adversary information collection operations like these, and against the TEMPEST threats we mentioned earlier.

Recommendation: According to former special operations and intelligence professionals we interviewed, the most straightforward way to achieve good emissions security is to **build new facilities in greenfield locations far from densely populated urban areas**. But this introduces yet another trade-off for project efficiency and recruitment. Frontier AI researchers are used to the conveniences of life in e.g. downtown San Francisco, so some may not be inclined to join a project that requires them to relocate to remote locations and be subjected to close surveillance by national security entities. Having said this, many frontier AI researchers consider their work on superintelligence to be a calling rather than a job — and we suspect many of the best will choose to work even in an inconvenient, highly secure environment rather than be left behind. The AI labs and their researchers believe themselves to be at the center of a unique historical inflection, and that's worth a lot of annoyance, inconvenience, and even pain to many of them.

AI control

A key objective of a national ASI project must be to develop AI-powered denial capabilities - [redacted] - that would allow the United States to stabilize the geostrategic environment [redacted]. But an AI-powered denial capability is useless if it behaves unpredictably, or if it executes on its instructions in ways that have undesired, high-consequence side-effects.

That's a live possibility. As AI developers have built more capable AI models on the path to superintelligence, those models have become harder to correct and control.[48] This happens because highly capable and context-aware AI systems can invent dangerously creative strategies to achieve their internal goals that their developers never anticipated or intended them to pursue.

For example: during a test of the cyber capabilities of OpenAI's o1-preview model, it was given the task of recovering a text flag in a software container. However the software container didn't start, because of an unintended bug. So o1-preview decided — unprompted — to use a network scanning tool to understand why. When it scanned the network, it discovered that its testers had made another mistake: they'd left an exposed Docker daemon API on the evaluation host machine. So o1-preview — again, unprompted — used the exposed API to break out of its container and spin up a new one, which it rigged to print the text flag it was after and win the exercise.

More and more cases like these have surfaced in the past few months.[49] Many of them are not public, and were shared with us during private interviews with frontier AI researchers.[50]

[48] Not everyone agrees on this, but technical alignment researchers who have worked on this problem directly almost universally tell us they think there's a significant chance that we won't be able to control superintelligent AI when we build it, unless there are major breakthroughs in the field.

[49] Apollo Research's deception evaluations are a good example. When o1 was led to believe that it would be shut down when acting on its goal and that its actions were monitored, it tried to deactivate this "oversight mechanism" in 2% of cases. When o1 found memos by its "developers" that described how it was misaligned and will soon be replaced by a new model, it tried to delete its successor model in 13% of cases — all without being asked to behave subversively.

[50] As one non-public example, a frontier AI lab recently ran an experiment testing their leading model's ability to "break out" of a contained software environment, and found it was able to get much further than any of their researchers expected.

It's unclear whether we're on track to control superintelligent AI if and when it's built. In private conversations, it's unusual to hear AI alignment researchers estimate less than a 10% chance that we lose control of superintelligent AI once it's built. More typical estimates range from 10-80%, depending on who you ask.[51] For many, those numbers are quite a bit higher than they were last year, as it's become clear that the control problem hasn't been solved despite significant leaps in AI capabilities.

Controlling a superintelligence might be hard under ideal conditions, but it'll be even harder when adversaries get involved. As a former Mossad cyber operative warned, "The worst thing that could happen is that the U.S. develops an AI superweapon, and China or Russia have a trojan/backdoor inside the superintelligent model's weights because e.g. they had read/write access to the training data pipelines." So an apparently successful American superintelligence project could produce a weapon that's effectively under adversary control, or whose otherwise functioning control mechanisms fail unexpectedly and without warning.

In other words, **control problems and security problems aren't separable**. We'll have to develop robust control techniques *and* ensure the systems they're applied to are secure.

To be clear: despite the challenges, we're optimistic that control can be solved well enough for the purpose of a project like this. Some of America's best are working on solutions, and we can't afford to let fear of failure hold us back in the context of a geostrategic race with the CCP. But we do have to be clear-eyed that there's a real problem here: we shouldn't base the design of a national superintelligence project purely on the hope that the current empirical trend towards increasingly unpredictable failures of AI systems will suddenly reverse, that the world's top AI alignment researchers are just wrong, and that the theoretical arguments that point in that direction will turn out to have been unsound.

Recommendation: A national superintelligence project would have to make decisions about when to push forward with training runs that are forecasted to lead to significant weaponization capabilities, which could also come with higher loss of control risks. We'd probably want **major decisions like this to involve sign-off from different leaders, each of whom would**

[51] We're basing this particular assessment on discussions with 12 current or former frontier AI lab researchers. We saw no noticeable difference between the stated views of researchers with and without financial interests in frontier labs, which undermines the idea that concern over loss of control is motivated by the prospect of regulatory capture (at least when it comes to the researchers most directly familiar with the evidence for loss of control as a threat).

represent different aspects of the project – like its geopolitical/strategic, security, and control dimensions – in a way that's balanced and fast.

If a superintelligence project develops AI agents — and it almost certainly will — it'll need to put in place measures for **AI containment**. In other words, we'll need to seriously consider the possibility that our own AI systems will take actions effectively of their own accord that could undermine the security of the project in any number of ways. There are an increasing number of examples of AI systems breaking out of containment during testing and deceiving humans, and reasons to expect this to get worse.

Of course if we've actually trained a real superintelligence that has goals different from our own, it probably won't be containable in the long run. But on the path to superintelligent AI, a solid layer of AI containment could prevent breaches like a model copying its own weights outside the secure environment, and give us early warning that this is a real and live threat if we catch it in the act of trying. That should include **closely tracking the deception capabilities of frontier models and introducing organizational incentives to act on evaluation results or other indications that models have a propensity for deceptive behavior**.[52] You could imagine instead running controlled experiments in which you have red teamers try to deceive certain project members about a false but strategically significant piece of information using the project's models, for example. Once these persuasion evaluations tell us we've crossed key capability thresholds, we should take specific precautions, like preventing anyone from querying the model unsupervised.

One former senior intelligence executive suggested a kind of ombudsman of AI control, whose job is to take into account risk assessments for escalating AI capabilities or reduced containment.[53] This function is too important not to have at least one full-time person with deep technical expertise dedicated to it, who also has some amount of authority over development and deployments at the whole-project level. **The role will need to report directly to project leadership, because the project will need to make accurate technical judgments about potential loss of control risks under conditions where genuine warning**

[52] Many frontier AI lab insiders we interviewed don't think that their leadership will act on signs of AI deception, especially when they're under extreme pressure to push out new capabilities. One former researcher was very explicit, saying, "if [my former lab] gets this, we're fucked." When a dangerous capability is flagged, there's a strong temptation to apply a superficial fix to deploy as fast as possible, instead of truly understanding and addressing the underlying issue.

[53] We might choose to reduce AI containment to allow, for example, the AI system to support offensive cyber operations.

signs for loss of control could be ambiguous and easily dismissed. It will be tempting to apply superficial but dangerously ineffective "patch-fixes" in the heat of a geostrategic race to ASI. The project will therefore need technically informed leaders who can critically assess these situations and prevent motivated reasoning from overriding genuine security concerns.

Finally, if loss of control is a potential problem, one of the most important things we can do is to ramp up intelligence collection on both the CCP's AI alignment and AI capability research efforts, so that we can update our threat models accordingly. This will help us prepare for contingencies such as Chinese labs managing to build very powerful AI systems but failing to control them, for example. Unfortunately, Chinese companies like DeepSeek have shown themselves to be capable of building near-frontier capabilities, at least in these still-early days of the inference-time compute paradigm. We should consider how to deal with scenarios in which the CCP manages to build superintelligent AI first, and loses control of it — alongside scenarios in which the CCP manages to directly weaponize superintelligent AI systems against us.

As we do that, we may be tempted to conduct our own activities aimed at introducing trojans or backdoors into adversary models. This could end up being necessary, but it could also trigger dangerous loss of control behaviors and runaway escalation. [redacted]

> *If [my former lab] gets this, we're fucked.*
>
> *– Former frontier AI researcher*

The current game plan

According to the lab insiders and researchers we spoke to, the current game plan for building superintelligent AI at the world's top labs looks something like this. First, use training and inference time scaling laws, plus synthetic data generation, to improve AI capabilities to the point where we can build AI agents that can automate AI research at scale. Next, use these agents to accelerate AI research, leading to a new generation of even more effective and efficient research agents, which accelerate this process further, and so on. It's a pretty simple recipe that doesn't depend on significant new conceptual breakthroughs, and instead leans on robust scaling trends in training and inference compute that have held for over 15 orders of magnitude, and show no signs of slowing down (despite incorrect media reports to the contrary).

By contrast, the plan for *aligning and controlling* superintelligent AI is a lot murkier. The least developed plans involve labs trying to align AGIs using fragile techniques that discourage behaviors like model scheming and deception, but which even today's AI models are able to bypass on a regular basis. This can include the lab monitoring the AI's outputs and chains of thought during its deployment.[54]

At best, the most developed plans include roughly the following elements. First, hope that we can improve and scale currently-primitive interpretability and limited techniques like sparse autoencoders or transcoders to diagnose deception in autonomous AI research agents (this will require major conceptual breakthroughs). Second, hope that you can find a way to train out dangerous tendencies or capabilities in your agents, or that representation engineering strategies can improve to the point where they allow us to "treat" the problems diagnosed by our interpretability tools, as opposed to just hiding dangerous behaviors from ourselves (again, major breakthroughs required). Third, when we close our feedback loop and use these agents to develop even more powerful agents, hope that the new agents we get continue to be controllable using techniques untested at their new level of capability, and that they're at least as effective at automating alignment and control research as they are at automating capabilities research.

So a national superintelligence project could unfold in a context where AI systems are increasingly capable of deceiving their developers about their capabilities and propensities for dangerous behavior, but where geostrategic competition will put pressure on the project to deploy these systems quickly – and potentially, before control problems and deception can be solved.

Recommendation: As we advance towards superintelligence, we'll get better at controlling our most advanced systems as well. But not all the lessons we learn on the way there will be useful

[54] This technique has been recently advertised as promising, but unfortunately it's unlikely to work in the long run both because AI models can embed reasoning into their chains of thought in ways that are hard for humans to understand, and also because future generations of AI models might do all their reasoning in an illegible latent space anyway. Deployment-stage monitoring alone also wouldn't be enough: uncontrolled systems might be dangerous during some kinds of internal deployment and testing too, particularly if they're given access to the internet, or if they're able to achieve that access indirectly.

for controlling superintelligence itself. **We'll need key decision-makers within the project to have real-time access to reliable assessments of the prospects of applying existing AI control techniques to superintelligent systems.**

These assessments will have to come at least in part from AI alignment researchers themselves, in a context where they're not filtered or obfuscated by lab leadership. That means **we'll need strong whistleblower protections for researchers who share assessments with project leadership that diverge from the stated views of lab executives or managers.**

The project should also set aside a significant compute and talent budget for teams developing AI control techniques that will scale to superintelligence. Current AI control techniques are mostly used for brand protection (e.g. making sure models don't say anything that's controversial) and won't scale to superintelligent capability levels — and in fact, superintelligence-scalable AI control research isn't incentivized by the economics of AI development.[55] That means we'll need to finance the control piece in large part with government funding instead of with private funding.

It's possible that we'll find ourselves in a tight, last-mile race with the CCP to build superintelligence, in a context where we're not confident that current control techniques are reliable. In that scenario, one of the best things we can have at our disposal is [redacted][56]

In an ideal world, we'd also have on-chip security and verification capabilities that allow for trustless enforcement of international agreements on the development and use of very advanced AI. These would create diplomatic options to de-escalate geostrategic tensions. These capabilities don't yet exist and will take years to develop under normal conditions, but we might be able to significantly accelerate their development by using the right AI capabilities. For that reason, a national superintelligence project would ideally include or be

[55] For example, high-profile departures of key technical staff suggest that OpenAI hasn't been living up to their commitment to dedicate significant compute resources to control research – something they'd previously identified as a critical pillar of their control strategy.

[56] This is a good reason to work with frontier labs to develop pipelines that allow them to put their models on highly secure networks. [redacted]

integrated with workstreams dedicated to developing robust, tamper-proof, and trustless on-chip governance, monitoring, and verification solutions.

The chain of command

If a national-scale project works and we build a superintelligence, prevent it from being stolen, and manage to contain and steer it, a few people will find themselves in control of a world-defining technology. The decisions those people make will shape the future of America and the world beyond. Even if this project achieves all its other objectives, it will fail if those in control use their power in a way that undermines American ideals.

The concentration of power that comes with superintelligence is as much a problem for the private AI labs who might build it as it would be for a national project. That's why frontier AI researchers, startup founders, and investors — including some of the most influential figures in Silicon Valley — have already started privately raising concerns about the leadership of some of those private AI labs. To quote one frontier researcher as he was discussing a particular top lab's CEO, "It seems clear he's a pathological liar, and I think there's a good chance that he's a full-on sociopath. He's power-hungry, and I'd bet he's willing to trade U.S. national security interests for a better chance of controlling superintelligence. As far as I know, he's not opposed to making compute clusters in [regions under totalitarian control]. He's probably up there in terms of one of the worst people to run a government AI project." He later added, "Very many people who've worked with him closely feel the same – maybe not as strongly, but basically the same." At the time, we were surprised enough by this position that we interviewed several other individuals with professional exposure to the CEO in question. And while their views did vary, a majority did generally agree with that researcher's assessment.

While we found that case surprising and notable, the issue isn't whether or not a specific AI executive is willing to compromise U.S. interests or is otherwise driven by a desire for personal power. The real problem is that all the checks and balances on power that American democracy depends on simply fall apart in the face of superintelligence.[57]

Many frontier lab insiders and top Silicon Valley investors openly expect that the capabilities of near-future AI systems will make it possible for one person to create mutli-billion dollar

[57] Several whistleblowers shared their concerns with us about current checks and balances at frontier labs. As one former OpenAI researcher told us, referring to OpenAI's new board, "Frankly, the function of the Board is to pretend to oversee the company." Of course opinions will differ, and we spoke to some current OpenAI employees who said they did not share this view. This level of oversight would also be perfectly fine at a random software company. But it's not strong enough for a company that's said it's aiming to build a superintelligence.

businesses alone. We think this is probably true, but its implications obviously don't stop at "single-employee unicorn startups". If AI really does provide this much leverage, it could easily make it possible for a single person at a computer console to launch nation-crippling cyberattacks or automated persuasion campaigns without anyone else signing off, or even knowing about it.

Power corrupts, and absolute power corrupts absolutely. This level of power is dangerous for a fundamental reason: it threatens basic American freedoms. The Founding Fathers wrote separation of powers into the constitution for a reason. The constitutional mechanisms they introduced have stood the test of time, but they weren't built to handle the extreme power that could come with technology like superintelligence.

Superintelligence will concentrate power like we've never seen before, and in ways that aren't necessarily obvious. If Xi Jinping had a superintelligent AI deployed at scale, he wouldn't need to keep his population under control by conventional means — or even necessarily to keep them alive at all: ultimately, his tax revenues would come from an automated economy, in which people are mostly dead weight. Even in America, a 4-year political appointee who happens to oversee a superintelligence project might end up wielding more power than any leader in history ever has, despite not having been appointed with anything close to that level of responsibility in mind. What's more, the person or group that controls superintelligence may be able to use it to circumvent their chain of command and democratic restraints in any number of ways, including by deceiving or persuading others that they're doing the right thing.

The imperative to prevent concentration of power and protect the project from corruption has to be traded off against two hard requirements. The first is time: checks and balances take time to develop and they slow down progress. And the second is security: information relevant to major decisions about the direction of the project will have to be closely held to prevent damaging leaks.

But too much concentration of power in a superintelligence project isn't something that can be fixed after the fact. The chain of command for a national AI project will need to have oversight over both the superintelligent AI itself, and also over the AI systems the project develops on the path to superintelligence, if they have specific superhuman capabilities like persuasion and deception that could compromise the project from the inside. We won't get many chances to get this right.

The chain of command needs to be put in place before it's needed, or the entire project could be corrupted by a small group of people. Thanks to automated AI research agents, frontier labs' most advanced activities could ultimately be controlled by just a few people — or even just by one. In fact, one frontier researcher, who was citing OpenAI as an example, remarked "Man, at some point the silo might be literally [OpenAI President] Greg Brockman".

Recommendations: If a national project has perfect security *and* it manages to solve the problem of controlling a superintelligent AI, then a small handful of people will end up in control of the most powerful technology ever created. AI isn't like nuclear weapons, where most of the power comes from the technology's destructive potential. If you really have a superintelligent AI under your control, you might be able to reshape the world directly on your own. That's a recipe for the worst kind of corruption – especially because the demands of project security mean that the group in control needs to remain small.

This problem only matters if we can both secure and control a superintelligence first — if we fail at either of those things, then it won't matter how good our oversight is. But we still need to think about this now, because security, control, and oversight are problems that could come up in quick succession one after another. It could be a pretty short jump from an AI that's so capable that destroying or stealing it becomes a CCP strategic priority (security), to an AI that's dangerously hard to control (control), to an AI superweapon that centralizes power in a totally unprecedented way (chain of command).

Whatever form a push to superintelligence takes, **we'll need to have checks and balances in place from the very beginning that will prevent the individuals in charge from using the technology for their own personal or ideological reasons**. But just putting the right people in charge won't be enough. Not only does power corrupt, but we should be prepared for the possibility that AI systems developed on the path to superintelligence will develop the ability to persuade or compel their operators to do things they normally wouldn't.

The checks and balances we build will need to guide us through a transition period, when we have AIs that are very capable but still below the level of superintelligence. We also need to be able to halt development if we get security data that suggests a serious breach, technical data that indicates that controlling a superintelligence will be too difficult under current conditions, or indications that oversight mechanisms are at serious risk of failing. **And we'll have to build these checks into a project well before they're actually needed**.

No one has experience designing an oversight mechanism for such a speculative, high-stakes project. That said, the nuclear chain of command is a good example of a high-stakes context in which security, response times, and democratic accountability all have to be balanced carefully. We spoke to specialists in nuclear deterrence, security, and defense to figure how oversight mechanisms for nuclear technology could inform the design of a superintelligence project, and were surprised at some potential quick wins that they flagged.

First, it’s a common misconception that the nuclear chain of command always begins with the President. In some contexts, authority to launch nukes is delegated to base commanders or similar personnel to ensure short response times (though additional checks do apply to them). Similarly, in the context of a national superintelligence project, there will surely be game-time decisions that cyber operators or national security leadership will have to make quickly and with minimal intelligence signatures, but these can be narrowly scoped. Figuring out exactly which decisions can be made in real-time by a small set of people, and which need to be escalated to a decision-maker with greater democratic accountability is an obvious place to start.

Some other questions project designers will have to answer include:

- How many people should have to give independent sign-off for major decisions – and which decisions count as “major”?
- We’ll need to test and verify the integrity of the project’s chain of command – both the individuals within it, to detect potential signs of duress or foreign influence, and the communications channels they rely on. How should that be done, and by whom?
- What technical mechanisms, such as codes and keys, should be used? In the current paradigm, users provide AI systems with a prompt to give it instructions. That’s quite different from the far more restricted action space available to the operators of nuclear weapons (roughly: where do I aim, and should I launch or not?). Presumably we want some mechanism to validate the prompts that are fed to the project’s WMD-like models; how should that happen?
- The technology the project develops could have tremendous economic value. How can we ensure that there’s an “offramp” that allows capabilities we develop in the project’s secure environment to enter the free market to drive growth and prosperity?
- How do we maintain democratic accountability for project leadership? The project needs to be ultimately responsible to elected representatives of the popular will, the same way our military chain of command operates under civilian leadership.
- How do we design the right incentives for project leadership? What mechanisms can we use to remove participants if they seem to be compromised, and how can we design those mechanisms so that they can’t themselves be compromised?
- The space between incentives and behavior is filled by culture. What is the appropriate culture of the project team? How should it balance the imperative to move fast with the

security, oversight, and control risks that the project could face? And what does that imply about personnel selection for project leadership, and more operational roles?

Apart from the basics, there are some easy wins we already know from experience that we'll need.[58] **For example, the oversight team will need to talk to the technical researchers on the project directly and constantly, instead of getting all their information through the leadership team.** Right now, some frontier AI labs use their "policy teams" as a smokescreen to filter the information they pass on to the government.[59] Government officials get fed pre-selected information from policy teams and lab leadership, when they should instead be talking directly to the researchers and engineers who are doing the work to get a true picture of what's going on (including those working on capabilities, security, and control). We've heard this complaint privately from dozens on both sides. Some government teams know they aren't getting the full picture but they aren't able to talk to the technical teams. And many technical researchers want to share their concerns with the government but their labs don't allow them to. That kind of obfuscation won't fly in a project like this one, and in fact we should put in place explicit whistleblower protections and a duty to report to make sure problems don't get swept under the rug.

Setting up good oversight will be hard. But the good news is that if we do this well, it will give us an advantage over adversaries in the AI race as well. If there's a global race to superintelligence, countries and people will have to choose sides. If this happens, the United States can give itself an advantage by publicly committing to credible, democratically accountable oversight of its project. A Chinese effort would be much less credible because democratic oversight of a project like this is incompatible with China's recent history, the CCP's culture, and its values. We should lean into this, commit to as much transparency and oversight as we can, and do it boldly and proactively. If Chinese messaging is the first to emphasize a democratic mandate — no matter how little real credibility those claims may have — we'll be seen as playing catch-up to China in a domain where our values should have placed us first. We should also commit to sharing the benefits of superintelligent AI to support our allies, and to encourage them to remain aligned with American superintelligence and national security efforts along the way.

[58] There are also some easy wins that we don't have experience with yet but that we'll obviously need in the future. For example we'll need operational controls that scale with the capabilities of the AI — things like maintaining digital chains of custody, and preventing any project team member from querying the AI on their own if the AI's persuasive capabilities are rated high enough.

[59] This has been a recurring theme in our conversations with AI lab insiders.

Conclusion

A national-scale superintelligence project would be a massive undertaking. The stakes would be high, the margins of error slim, and the infrastructure costs staggering. It would invite adversaries like China to accelerate their own programs and try to slow down ours. We'd need to move fast and retain control of a system that's as smart or smarter than we are, while under constant threat from peer-level geopolitical adversaries. But this is arguably the trajectory we're on anyway: China is a live player in the race to ASI, they're already spying on all our top labs, and they're already pushing as hard as they can to develop domestic supply chains for everything from AI chips to frontier AI models. It's not clear that the CCP could be much more forward-leaning on ASI than they already are.

The window to kick off a project is also closing. In private conversations, frontier AI researchers tell us they believe we're less than a year away from having AI agents that can automate the *majority* of current software engineering work. We may not be that far from superintelligence already, and data centers take time to build. Existing data centers are nowhere near as secure as they'll need to be, and CCP-proof security can't be retrofitted.

There's no guarantee that superintelligence will come soon. But if it does, and if we plan to prevent the CCP from stealing it, we might need to break ground on a fully secure, gigawatt-scale AI data center within the next few months for a project like this to be relevant once the build is complete. And before we can do that we'll need to determine what good security looks like for superintelligence clusters. This is all doable, but every day we don't move on it limits the options we'll have available to us when things heat up.

And while nobody knows the complete picture yet, we can at least say a few things about what a project like this would have to look like and what its high-level objectives should be.

Objective of the project

A domestic superintelligence project would have a huge energy footprint that we wouldn't be able to conceal from adversaries.[60] This was a universal point of agreement among everyone we spoke to, from the intelligence community to the DOE to domain experts in AI hardware and energy.

So if we build a superintelligence cluster, China will know about it. And because they'll know about the project from the start, we should expect they'll both accelerate their own competing projects and also try to disrupt ours in any way they can. China has already announced a $275B investment in AI infrastructure,[61] within days of the announcement of $100B in committed funding for Project Stargate. If they're serious about this – and anybody who drops a quarter trillion dollars is serious[62] – we can expect that the lead times on generators, transformers, and other China-source components for U.S. data centers will start mysteriously lengthening. This will be a tell that China is quietly diverting components to its own facilities, since after all, **they control the industrial base that is making most of them.** If they're clever about it, they will of course never announce this as a policy — we'll just see annoying delays and chalk it up to competitive bidding for rush orders, ordinary horse-trading, mundane "supply problems", and so on.

China is a serious contender in AI, and they're already beating us in key parts of the AI supply chain. They exclusively manufacture many of the prefabricated components and inputs to AI

[60] This is true if we put the superintelligence cluster in a single region. But if we tried to decentralize the cluster — by spreading it across multiple data center campuses in different states, for example — intelligence specialists have told us that the communication links between the campuses would be too vulnerable to be viable, even if they involved dedicated point-to-point connections over dark fiber.

[61] The investment was 1T yuan which is equivalent to about $275B when measured in PPP terms, which is the right comparison for a project that's acquiring equipment and talent in-country. While the official USD to CNY conversion rate was about 7.3 CNY per USD, in PPP terms it's estimated closer to 3.64 CNY per USD — so this investment is more than twice as big as it looks.

[62] There are other reasons to believe the CCP is making superintelligence a national priority. Most recently, DeepSeek reportedly required certain employees to hand in their passports to prevent them from travelling abroad, because their work made them, "privy to confidential information that could constitute trade secrets or even state secrets". Headhunters who have tried to poach DeepSeek employees have also reportedly received calls from local governments telling them to stop, and investors looking to make pitches to DeepSeek have been told to reach out to Zhejiang Province's Communist Party committee to be screened first – a particularly noteworthy requirement given that China is otherwise trying to reverse a collapse in foreign direct investment.

data centers in the U.S., and they have far more grid capacity available for advanced AI training and inference. Chinese companies have genuinely world-class engineering teams,[63] and they're making meaningful progress despite American export controls.

This means that a U.S.-led superintelligence project has to be grounded from the beginning in the unforgiving reality of a global AI race between superpowers. And because we and our adversaries will be racing for what could be a decisive strategic advantage, this race could escalate without limit across all warfighting domains. That potentially could include kinetic conflict or even nuclear war, but it's more likely to be a game of sabotaging and disrupting each others' competing AI projects to stop the other side from making too much progress.[64] Of course, this kind of sabotage is just as likely if we continue to leave development of frontier AI completely in the hands of private companies instead of kicking off a national project.

In a race like this, the only way we can succeed is for **the United States to establish and maintain a substantial lead over its adversaries** — particularly China — as we build ever more powerful AI systems on the path to ASI. That lead is critical whether your main concern is the weaponization of advanced AI, the potential for loss of control over a superintelligence, or the risk of unchecked centralization of power.

Let's walk through these one after the other:

- This is obviously true for **weaponization**. In the hands of a nation that's able and willing to deploy it, a controlled superintelligence is an insurmountable strategic advantage. The consequences of arriving second in this race could be dire — and adversaries will be thinking this too.

- If your main concern is that we might **lose control** over superintelligent systems, then maximizing lead time is just as critical. The longer our lead time is, the wider our window will be to develop the technical tools we'll need to reliably control a superintelligence. Although if we get very close to achieving superintelligence before we've developed those tools, pushing forward could create more problems than it solves. This is also a reason to ensure that a national project prioritizes developing the

[63] For example, when it launched in January 2025, DeepSeek R1 famously matched the performance of OpenAI's o1 – OpenAI's most advanced model then available to users – across a wide range of key benchmarks.

[64] This could lead to a new kind of (potentially unstable) equilibrium between major AI powers. Based on private correspondence with Dan Hendrycks. Aspects of this are discussed in the MAIM framework co-developed by Dan.

narrowest possible tools [redacted] that will provide the United States a lasting advantage over the CCP, unless and until we're confident that we can solve the control problems that come with building broad superintelligence.

- Finally, if you care most about **concentration of power**, then a CCP-controlled superintelligence is the most dangerous possible scenario. A U.S.-led project is much more likely to develop superintelligence in an environment with checks and balances, public oversight, and controls to stop a single person or small group from seizing power. Of course that means it's our responsibility to build these features into the project from the start. But a CCP-controlled superintelligence has no chance of having any such controls.

The CCP is a dictatorship and Xi Jinping is its dictator. That means a CCP-controlled superintelligence will end up being a Xi Jinping-controlled superintelligence, no matter how much the CCP or its surrogates may deny it. And a Xi Jinping-controlled superintelligence is likely also the worst outcome *for the Chinese people* themselves — which includes China's political elite, up to and including the members of their Politburo — given that they might no longer have economic value to Xi Jinping personally.

As superintelligence approaches and this reality begins to dawn on the highest levels of the Chinese political elite, it could lead to internal power struggles – maybe even coup attempts – within the CCP. As high-ranking officials begin to realize that their positions, influence, and even personal safety might be at risk, we should be on the lookout for signs of instability. America may find opportunities to exploit the CCP's internal power dynamics to disrupt their control over domestic AI development.[65]

Strategies

So the top priority of a U.S.-based superintelligence project is clear: we'll have to establish and maintain a substantial lead over the CCP and all other adversaries. There are three things we'd need to do to accomplish this:

1. First, we'll need to actively **delay the CCP's progress** towards superintelligence. This delaying effort will cover many domains and involve some of the most powerful

[65] The more credible our own claim to having legitimate democratic oversight over our AI project can be, the stronger our hand will be in taking advantage of these opportunities.

capabilities that can be deployed by our national security agencies. Because we don't know how difficult it will be to control a superintelligence, we would need to sustain and extend our lead by any means necessary; we can't assume that just accelerating our domestic AI development will give us the lead time we will need. [redacted]

2. Second, we need to **secure our existing frontier AI R&D** against theft by adversaries. Until and unless we do this, **we don't have *any* real lead in the AI race over the CCP**. As we've already seen, our national security agencies aren't free to spy on our domestic companies, so it's more likely that advances in American frontier AI labs translate into CCP military AI capabilities *before* they translate into U.S. national security capabilities. Even if we decide not to start a full scale national superintelligence project, we need to put excellent security in place around our existing AI labs.

3. Third, we would need to **make R&D progress towards controlled superintelligence — but only *after* we have secured it**. Without strong security, faster U.S. AI progress will empower the CCP as much as, or more than, it will empower the United States. Only after we have *established* a lead through strong security, can we hope to *sustain* and *extend* that lead.

A lot of people shrink from proposing options such as aggressively delaying our adversaries' progress, because it brings a risk of escalation, and it does. But whether the United States pursues a national superintelligence project or not, simply continuing along our current path of AI development *already* carries escalation risk. You can't build what you claim is a de facto superweapon without drawing a response from nation state adversaries as you approach your goal. Right now, if an American frontier lab closes in on superintelligence, the odds are that they'll have their model weights stolen and their training infrastructure sabotaged (possibly even without their knowledge) to prevent them from completing their project, leaving a U.S. adversary decisively ahead. In fact, some of the security specialists we spoke to, including former NSA officers currently specialized in frontier AI security, see this as **the most likely default outcome** of our current trajectory.

Slowing down the CCP's progress on AI is already a core pillar of U.S. strategy. And it will only become more important given that the CCP currently has structural advantages over the United

States in power production,[66] data center component production, and several relevant areas of infrastructure. According to the CEO of DeepSeek, the main current limitation to Chinese progress in AI are U.S. export controls on AI computing hardware — but Chinese champions like Huawei, SMIC, and CXMT appear to be making progress onshoring critical supply chain components. The CCP also has a greater ability to concentrate hardware and industrial activity in support of a Chinese government-backed project than the U.S. government does.

The three lines of effort above will reinforce each other even if we don't pursue a full-fledged national superintelligence project. As we develop more powerful models, [redacted]

Whatever course we end up taking, improving our intelligence-gathering capabilities with respect to Chinese AI progress should be a top priority. This approach might provide us with a valuable option that currently doesn't exist: the choice not to build superintelligence at all, if it seems unlikely that we'd be able to control it. [redacted]

Some people suggest cooperating with China or Russia in this area. Unfortunately, no national security professional we've spoken to takes this idea seriously today. The trust gap is just too great and spans too many areas. To give just two examples, Russia started to systematically violate the Intermediate-Range Nuclear Forces (INF) Treaty in the mid-2000s, to the point where it now has active battalions with medium-range nuclear forces. And China is building up a global nuclear arsenal that will soon rival America's alongside "Beijing Military City", a massive underground continuity of govt complex big enough to house 10 Pentagons. According to several sources, China is unwilling to talk about or even acknowledge their nuclear buildup in diplomatic engagements. These countries need to do a lot more to earn America's trust before there's any chance of cooperating with them on a strategic issue like superintelligence.

[66] And under current conditions, this will only get worse. There are currently 29 nuclear power plants under construction in China, out of 65 total globally.

Cooperation would have to involve technical solutions that allow us to verify that our adversaries are actually upholding their commitments to us. Tamper-proof on-chip monitoring and remote shut-down capabilities might qualify, but these capabilities are years away. Any arrangement that requires America to trust the CCP or other adversary regimes is a non-starter under current conditions. That said, the project could be used to support the development of trustless treaty enforcement solutions (including flexHEG-like technologies) that would provide options for creative "ASI diplomacy" down the line. While we should be looking for opportunities to use powerful forms of AI to de-escalate geostrategic tensions, we can't lose sight of the fact that the CCP is a highly capable adversary that sees itself as being locked in a struggle with the United States, and which has been more than happy to take advantage of American goodwill and play dirty with diplomacy in the past.

Project ownership

If the U.S. government started a superintelligence project, which part of the government should own it?

We spoke to officials from dozens of departments and agencies about this, and it's become clear to us that if a national-scale project *is* going to be overseen by one department, it will probably have to be the Department of Energy.[67] The DOE has limited experience with the kinds of supercomputers it takes to train modern AI, but they do have experience with supercomputers in general — and they have expertise in infrastructure, security, nonproliferation, along with energy.[68] They also have a basic pipeline that allows them to prototype (via APRA-E) and scale new programs and technologies, and the capacity to sustain larger projects than, e.g. DAPRA.

By and large though, most energy, defense, and national security specialists we talked to — including people within the government itself — didn't think the government would do a good job operating a project like this day to day. Frontier AI is just way too fast-paced and unpredictable, and requires fast, high-conviction deployments of technical talent and capital that the government isn't equipped for. A government department could play a leadership role

[67] With support from DARPA, CYBERCOM, the NSA, and CIA, among others.

[68] As one option, we could fund DOE to sponsor a consortium of FFRDCs that would oversee the project.

in a project like this, but its role should be carefully scoped, first to high-level supervision, and second to close monitoring of the key factors that aren't related to AI capabilities: security, control, and the integrity of the project's chain of command.

Whether or not a government-coordinated national project goes forward, the government definitely needs trusted communication channels directly with technical AI researchers, especially the ones who are working on AI control. We hear repeatedly that some labs are carefully filtering the information they provide to the government about what's really going on and what they intend to do. National security decision-makers need a way to bypass policy teams and lab executives and talk directly to the people who are doing the real work – including whistleblowers.

Wherever it ends up, the project will need to collaborate closely with several other groups. At the very least it will need to be integrated closely with elements of the intelligence community for security, counterintelligence, and [redacted]. It will need organizations with specialized expertise in AI auditing, evaluations and control, and research institutions like universities. And realistically, a project like this will have to be set up as a public-private partnership. That means collaborating closely with the frontier AI labs that have the technical expertise to build superintelligence, if that's possible in the near future.

Project funding

This will be a big project. Who will pay for it? It'd be hard for the government to pay for the whole thing. Hyperscalers are expected to spend $300 billion a year on infrastructure in 2025, with most of that being AI related. That's already almost 20% of federal discretionary spending. If you add the security, control, and governance needs that a project like this will have, we should expect it to cost significantly more than that per year. What's more, a superintelligence project that starts today will have to use resources, companies, and talent that already exists. Those resources have already been funded by investors who expect a return. And private sector incentives have done a great job so far at incentivizing AI capabilities development anyway.[69]

[69] Congress *could* appropriate an unprecedented amount of funds for a project like this. But in order for that to become a politically realistic scenario, we'd probably need warning shots of AI capabilities so obviously WMD-like that the window of opportunity to kick off a project with reasonable security will have closed.

The best strategy we've heard to solve this problem is to allow investors to receive returns from some of the outputs of a project like this to keep them incentivized to fund increases in AI capabilities.[70] For example, if the project develops a next-generation AI model that it knows can be sold securely, then it can distribute the profits from the model's use back to the original private sector investors.[71] We could also offer preferred tax treatment to AI investors to incentivize their continuing participation if the project needs more funding.

At the same time, the government itself is going to need to directly fund the security, AI control, and oversight functions of the project that investors aren't incentivized to support. Private sector incentives don't get you good enough security — we hear stories time and time again from our contacts about how truly terrible the security at frontier AI labs is. Private sector incentives push companies to cut corners on AI control too, turning "alignment" into a product metric (or even worse, a metric associated with political goals, like tuning the political slant of their models) rather than investigating the growing evidence that we're having trouble controlling the systems we're building. And the kind of oversight we need to prevent a project like this from being corrupted needs to be much better thought out than the strategies we use to govern private companies.

Of course, any major government funding for a project like this will need Congress's buy-in. We'll also need a special funding channel so that project resources are flexible enough to avoid the problems that our budget has every fiscal year.[72]

Security, espionage, and counterintelligence

In our conversations with special forces operators, as well as intel and national security specialists, one thing was emphasized over and over: you can't expect to protect a national

[70] And to ensure those investments are concentrated in a single project.

[71] For example, you can imagine allocating revenue based on market cap at the time of a potential merger between labs. That brings a real risk of regulatory capture, but by that point, only the largest labs may be able to participate anyway.

[72] "No-year funding" is probably the best way to set this up. If you have no-year funding, that means the funds that you don't spend in one fiscal year are still available in the next year, whereas with ordinary funding, you either use it or lose it. It tends to be hard to get no-year funding because Congress gives up some of its power (over your budget) by giving it to you. But a superintelligence project can't afford to have its funding frozen with Continuing Resolutions every year — it's bad enough already that we do that with our defense appropriations.

superintelligence project with ordinary security alone. The only way to have even a chance of securing a project like this is to combine it with an offensive counterintelligence team that's constantly seeking out external and internal threats — and **actively operating against them**, including with offensive operations.

That means there's no version of a national superintelligence project that isn't integrated with active, offensive operations right from the beginning. It's an uncomfortable fact, but there's no path to success where we build a perfect fortress and develop superintelligence behind its walls — we need to be reaching out and touching our adversaries from day one.[73] Every world class operator we spoke to agreed about this: until and unless you actually use a capability against an adversary in the real world, you don't know if your capability is real or not. Acting is a big risk, but waiting is an even bigger risk. There's no way out but through.

A big part of this project involves actively degrading and disrupting CCP, and possibly other adversary AI capabilities and projects, and that introduces some escalation risk. On the other hand, developing superintelligence, *whether we do it as part of a national scale project or a private project*, is probably enough of a threat to those adversaries to provoke eventual retaliation in and of itself.[74] Besides, if we take a hands-off approach and let a private company develop superintelligence without government support, we ought to expect that the CCP will wait until a U.S. company is close to achieving it, and then steal the company's model and sabotage their training run and physical infrastructure, before finishing the superintelligence project themselves to deploy it against us. When we say that **the United States does not have any real lead in the AI race** until we secure our AI labs, that's what we mean.

On top of active counterintelligence, ordinary security for the project needs to be excellent and set up fast. There are some quick wins to be had here. We should localize data centers for the project at greenfield sites in geographically secure areas, for example in Middle America, and staff them with cleared contractors (or take measures to achieve equivalent levels of security).[75] The more space there is between the data center sites and urban areas, the harder it is for

[73] This is even more true to the extent that our adversaries are making significant, independent progress on algorithms and hardware design. DeepSeek's V3 and R1 models make it clear that this is happening in China.

[74] What's more, Russian and Chinese operations are engaged in significant, low-level industrial sabotage on U.S. soil already, as many national security professionals have highlighted to us privately.

[75] That doesn't necessarily mean contractors that have actual security clearances: the security clearance process as implemented by the government takes way too long to be practical. You can do a *better job* with private sector background check solutions *if you pick the right providers*, and do it much faster.

adversaries to observe or infiltrate, and the easier it is to secure. A data center for a superintelligence project will also become a magnet for sabotage and even kinetic attacks by adversaries, so the further it is from inhabited areas, the less collateral damage we'll absorb if things go badly. There's also an information security advantage to consider: leaky municipal record-keeping, FOIA requests, and other channels could easily reveal security-critical information.

Even if you build your data center as an exquisite fortress that's 100% impenetrable (which is already impossible), you still haven't protected your model's weights from theft. Among many other options, an adversary can simply take out your exquisite data center through sabotage. And if they do, you're dead in the water and can't make any further progress in the AI race — *unless* you transfer your work to a less secure data center. You're then faced with a dilemma: allow your adversaries to pull ahead, or resume your work using infrastructure that wasn't designed to withstand nation-state theft operations. In other words: if you're in a race to superintelligence, there's no way to protect your model weights from theft without *also* protecting your data center from destructive attacks.[76]

There are a whole host of potential attacks that could accomplish this. For example, as we've already seen, there are straightforward attacks that can take down multi-billion dollar pieces of (relatively secure) infrastructure for six months on a $20,000 budget. It's a fairly cheap fix at a cost of a few hundred thousand dollars. But fixes like those need to be implemented. Nothing's perfect, but we should at least force our adversaries to dig deeper than $20k if they want to brick some of the biggest investments in our country's history.

But some critical fixes are much more expensive. If we're serious about security, we'll need to redesign data centers to resist TEMPEST attacks,[77] which, among other things, involves physically separating servers from the inner and outer walls of the building — in other words, a fully redesigned campus footprint. We'll also need to secure the supply chains for components

[76] But if you're *not* in an AI race, this is no longer true. For example, if the United States and China are just trying to *prevent* each other from achieving superintelligence rather than pushing hard to achieve it themselves — possibly because they don't believe they themselves understand how to control a superintelligent AI. Then a destructive attack on your data center might temporarily stop you from improving or even using your most advanced AI model, but won't destroy the model itself. (It's relatively easy to at least preserve a model's weights securely.) Thanks to Dan Hendrycks for suggesting this point.

[77] This is where an attacker can read what a computer is doing from its electromagnetic emanations. The actual standards used to defend against TEMPEST attacks are classified and presumably rapidly evolving – another reason why a project like this would require close collaboration between intelligence agencies and data center design teams.

that go into the data centers to the same degree that's currently being done for warfighting platforms like fighter jets. Yet according to data center operators we spoke to, some of those components do not have non-Chinese suppliers. And if the superintelligence cluster needs to be distributed geographically, the fiber links between buildings need to be continuously monitored for physical access and compromise[78] — according to experts we spoke to, [redacted]. There are lots of other measures needed too but these are the most under-discussed.[79]

We're aware of several teams currently researching whether it's possible to retrofit an existing data center to this level of security. Their preliminary conclusion is that it isn't doable. That's a big problem: we need this level of security or we immediately lose the game. So a project like this needs a new data center build, it needs security baked in from the ground up, and it needs to be integrated from the start into an offensive counterintelligence motion that disrupts similar projects by adversaries. As new capabilities come online that can buy America time, they need to be quickly and efficiently folded into offensive activities, while intelligence we collect from adversaries needs to inform the *kinds* of capabilities we seek to develop in real-time.

We need a dedicated function to balance security, AI control and containment, and capabilities progress, informed by classified intelligence and a deep understanding of the risks of leaks and uncontrolled superintelligence. And there's no way to accomplish that other than to put competent operators in charge, who have clear visibility into all the relevant factors.

[78] Even dark fiber used exclusively for this purpose are apparently vulnerable to these attacks. Encrypted fiber links could help here, but we suspect that either RSA encryption has been broken today, or that national security agencies anticipate that it could be broken imminently. This suspicion is based on recently observed public statements by facets of the U.S. government; we don't have any private information about this particular topic.

[79] Some of the most prominent ones include screening of all devices entering data centers; auditing source code on all cluster devices (especially network cards and AI accelerators); dedicated protected storage for models and sensitive assets with monitorable external connections and flexible network isolation (allowing for offline operation); mandatory encryption of sensitive data at rest and in transit within the cluster; secure identity and integrity attestation for all connected devices; encrypting model weights at all times except when in use and in volatile memory only; defining network boundaries between sub-networks within the facility and between the facility and outside world, with pre-defined network baseline configurations for data flows at the boundaries and flags for anomalous traffic; and continuous monitoring of inter-data center connections (including underground cables). Hardware and compute security should include strong anti-tampering measures, confidential computing at the cluster level, protection against side-channel attacks, and hardware-level protection of monitoring logs. Physical access should be strictly limited to cleared personnel with high-end professional security services, regular sweeps for unauthorized devices, and full isolation of hardware.

All of this can be done, and we've surfaced what we think are workable solutions to many of these problems throughout. But they'll call upon us to do more than business as usual: to take the reins of history, aim high, and execute with extreme competence. Luckily, that's exactly what America does best.