Rethinking Priorities for Frontier AI Security

Executive Summary

BLUF. Frontier AI is advancing faster than security. While SL5 protection against superpowers is the long-term ideal, it is infeasible today. Therefore, the priority task is to prevent OC4-level theft by building SL4 optionality across U.S. labs.

Background. Within years, some believe models will automate AI R&D, compressing a decade of progress into months[1]. In such a scenario, the AI model weights would rank among the most strategically valuable U.S. assets. Yet current frontier lab security lags this threat.

Premature focus on SL5. The debate around frontier AI lab infosecurity has centered on achieving Security Level 5 (SL5)—defenses against superpowers. But SL5 is unrealistic: it could cost hundreds of billions of dollars[2], cripple U.S. AI innovation, and remains politically infeasible. Pursuing it prematurely would be counterproductive.

SL4 as the urgent policy ask. Because SL5 is unattainable, the priority danger to address comes from Operational Capacity 4 (OC4) actors—states like Iran or North Korea, or major cybercriminal groups. These adversaries are skilled at cyber-infiltration but are unlikely to securely retain what they steal. Once frontier model weights are taken:

- **Destabilization of world order**. Theft of advanced AI elevates weaker actors to "nuclear-equivalent" status, multiplying flashpoints and intensifying strategic competition[3].
- **Cascading leakage of weights**. OC4 actors overinvest in offense and underinvest in defense, making further leaks likely[4][5].
- **Escalation under pressure**. AI-accelerated military and intelligence operations[6] can force the rushed deployment of untested systems[7], collapsing decision timelines and heightening the risk of accidental escalation[8].

Proposal. Because SL5 is out of reach, the actionable risk is OC4 theft. To counter it, the U.S. should ensure frontier labs can rapidly harden to Security Level 4 (SL4)—defenses strong enough to withstand theft by Operational Capacity 4 (OC4) actors—as soon as circumstances demand.

---

[1] https://www.forethought.org/research/how-quick-and-big-would-a-software-intelligence-explosion-be
[2] https://ifp.org/a-sprint-toward-security-level-5/
[3] https://www.rand.org/content/dam/rand/pubs/research_reports/RRA3200/RRA3295-1/RAND_RRA3295-1.pdf
[4] "Almost all the advantages are on the side of the hacker; the current situation is not sustainable" https://www.nbr.org/wp-content/uploads/pdfs/publications/IP_Commission_Report.pdf
[5] Cyber-Weapons of the Weak: Understanding the Pursuit of Offensive Cyber-Capabilities by Smaller States. (David 2022)
[6] https://doras.dcu.ie/25554/1/Cyber%20Policy%20JamesJohnson%20%282019%29.pdf
[7] https://cset.georgetown.edu/wp-content/uploads/CSET-Reducing-the-Risks-of-Artificial-Intelligence-for-Military-Decision-Advantage.pdf
[8] https://www.amacad.org/publication/daedalus/cyber-warfare-inadvertent-escalation

- **Achievable**. SL4 is costly but feasible. Multiple labs could reach it with modest government coordination and private investment. By contrast, SL5 (defenses against superpowers) would require hundreds of billions of dollars, invasive regulation, and slow U.S. innovation—making it unattainable in the near term.
- **Urgent**. AI is emerging as the defining strategic technology. The immediate risk is not superpower theft but uncontrolled proliferation through OC4 theft. Once leaked into weaker hands, each new custodian becomes a potential crisis flashpoint.
- **Stabilizing**. SL4 slows uncontrolled proliferation of frontier models. Fewer custodians means clearer attribution, stronger deterrence, and more time to establish international safeguards before diffusion becomes irreversible.

Conclusion: SL4 optionality is strong enough to block the most destabilizing threats, fast enough to deploy across multiple labs, and politically viable. As with nuclear technology, stability depends on limiting the number of custodians. Preventing uncontrolled diffusion now is the surest way to avoid a world of constant flashpoints and preserve U.S. strategic advantage in the AI age.

I. Introduction

Imminent AI transformation. According to a recent survey of thousands of AI experts, 2032 is the median expert prediction for AI systems able to "do all tasks better than humans"[9]. The economy is taking this forecast seriously. Meta, Google, Microsoft, and other tech giants are betting their futures on the near-term potential of AI, cumulatively forecasted to spend over $300 B in 2025 on AI infrastructure[10]. The market has rewarded their bullishness. Since the release of ChatGPT in late 2022, the combined market capitalization of the so-called 'Magnificent Seven' technology firms has risen from roughly $7 trillion to nearly $20 trillion—a gain of close to threefold in less than three years[11]. The government is also paying attention. The 2025 America's AI Action Plan calls AI "an industrial revolution, an information revolution, and a renaissance — all at once," and officials as high up as Vice President J. D. Vance have publicly referenced papers predicting superintelligent AI within five years[12][13].

Heeding warnings. Smart, well-informed people are effectively betting on the possibility of artificial superintelligence (ASI) in the near future[14]. Policymakers should treat their claims as plausible and seriously consider the implications for frontier AI lab security. In other words, it would be prudent to start preparing for transformative AI scenarios now, rather than dismissing these aggressive timelines.

Explosive progress. One scenario often discussed to justify these aggressive timelines is AIs automating the AI research and development process (AI R&D) at a cost competitive with human researchers[15]. Once we enter such an automated AI R&D loop, experts predict that even a decade of

---

[9] https://aiimpacts.org/wp-content/uploads/2023/04/Thousands_of_AI_authors_on_the_future_of_AI
[10] https://www.ft.com/content/634b7ec5-10c3-44d3-ae49-2a5b9ad566fa
[11] https://www.forbes.com/sites/bill_stone/2025/08/31/nvidia--magnificent-7-too-hot-to-handle/
[12] https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf
[13] https://www.nytimes.com/2025/05/21/opinion/jd-vance-pope-trump-immigration.html
[14] https://kalshi.com/markets/kxoaiagi/openai-achieves-agi
[15] https://superintelligence.gladstone.ai

ordinary progress could be compressed into mere months of AI-driven self-improvement[16]. For perspective, three years of recent progress took us from models that stumbled over the number of *r*'s in "strawberry" to ones earning gold medals at the International Math Olympiad. Another similar leap forward — especially once we have AIs capable of designing improved AIs — would have tremendous implications for the future of intelligence and military power. In such a world, the model weights (the parameter values of these AI systems) would likely become among America's most critical national defense secrets[17].

Strategic secrets. If AI labs succeed in training a model that can automate AI R&D, the value and strategic importance of that model's weights would be astronomical. However, our frontier labs' current information-security measures — largely the same practices used before the era of automated AI R&D — would be **clearly insufficient** relative to the value of those weights. In a world with AI-driven self-improvement, ordinary corporate security measures might be inadequate to protect such assets, raising the risk of losing U.S. advantages in both economic and military domains[18][19].

Unpredictable progress. The challenge is that we do not know when AI models will reach such dangerous capabilities. A recent RAND report provides terminology for discussing AI model security, defining a hierarchy of "Security Levels" up to Security Level 5 (SL5), which corresponds to protecting model weights against the best efforts of a superpower adversary (Operational Capacity 5, or OC5)[20]. Ignoring cost, SL5 would be ideal. However, there is famously a steep trade-off between security and productivity, so labs have good reason to avoid ratcheting up security until it is necessary. Unfortunately, the key question — *when will it become necessary?* — is complicated by the unpredictable and often undetectable nature of AI capability gains. A 2023 mega-report with authors from 22 institutions (including three major AI labs) argued that the unpredictability of AI performance leaps is one of the biggest challenges for regulation[21]. The 2025 International AI Safety Report, representing top experts from over 30 countries (including the U.S. and China), echoed this concern. Its top consensus question asked: "How rapidly will general-purpose AI capabilities advance in the coming years, and how can researchers reliably measure that progress?"[22]. Uncertainty stems in part from threshold effects in machine learning, where small increases in model size or training compute can lead to surprisingly large jumps in capability[23][24]. Our benchmarks and proxies often fail to capture the true difficulty of real-world tasks. For example, AIs have "solved" many writing benchmarks without actually displacing human writers in the economy[25]. Likewise, we can only guess how well current metrics predict a model's ability to self-improve — ultimately, we are largely in the dark about when an AI might gain the capability to meaningfully accelerate AI R&D.

Hidden capabilities. Even if an AI does develop dangerous capabilities, we may not realize it immediately. The moment after a new model finishes training is often "the worst it'll ever be"; a fresh

---

[16] https://www.forethought.org/research/how-quick-and-big-would-a-software-intelligence-explosion
[17] https://situational-awareness.ai/lock-down-the-labs/
[18] https://warontherocks.com/2025/08/high-risk-ai-models-need-military-grade-security/
[19] https://x.com/pmarca/status/1764374999794909592
[20] https://www.rand.org/pubs/research_reports/RRA2849-1.html
[21] https://arxiv.org/pdf/2307.03718
[22] https://arxiv.org/pdf/2501.17805
[23] https://arxiv.org/pdf/2307.03718
[24] https://cset.georgetown.edu/article/emergent-abilities-in-large-language-models-an-explainer/
[25] https://epoch.ai/benchmarks

model without fine-tuning or tools is like a carpenter without a toolbox[26]. It takes months of deployment and the creativity of millions of users to reveal a model's full range of abilities and quirks. For instance, years after GPT-3's release, a user discovered that the nonsense string "solidgoldmagikarp" could induce a bizarre failure mode in the model[27]. In the case of an AI that can improve itself or design new AIs, a model initially deemed benign might later demonstrate significant autonomous R&D capabilities once in the wild. By the time the lab or regulators realize that a deployed model is essentially an "AI researcher," it may be too late to contain the spread of its weights. Compounding this problem, leading labs often keep their most advanced models internal for months before any public release, blinding policymakers and outside observers to the true frontier of capability[28]. During that lag, nation-state adversaries could have already infiltrated the lab's networks and stolen the weights. U.S. intelligence agencies are legally barred from spying on domestic companies, but foreign agencies have no such limitations — and conceivably could make military use of U.S. frontier AI models before the U.S. can[29].

Start preparing now. Clearly, it won't work to wait until AI systems unambiguously display dangerous, transformative capabilities before we strengthen lab security. By then, the opportunity for safe intervention might have already passed. In a scenario where AI capabilities advance suddenly, the United States could find its most valuable model weights stolen and proliferating before our defenses catch up. We **should begin preparing now**[30][31]. The question is: what form should that preparation take?

II. The Current Policy Debate on AI Security

Thus far, much of the policy discourse has been framed as a U.S.–China AI race, with a focus on preventing China from stealing or overtaking U.S. capabilities. Within this paradigm, the implicit security target is SL5 protection of frontier model weights—safeguards strong enough to withstand the superpower-sponsored cyber operations with the highest priority.

One operationalization could be the level of security employed at the Manhattan project, which successfully delayed Soviet theft of critical information to months rather than days[32]. But as the world has become more complex, security has become more expensive. RAND deems SL5 technically impossible with today's technology[33], a government-commissioned report estimates it could take hundreds of billions of dollars[34], and lab insiders note that it would mean losing a double digit percentage of top AI researchers[35]. So for now, SL5 is impractical and infeasible.

Despite that, government investment into frontier lab security is warranted. If AI models start demonstrating capabilities similar to weapons of mass destruction, then we will want the ability to quickly secure their weights from well-funded terrorists before an catastrophic event analogous to

---

[26] https://perma.cc/K4FG-ZXMX
[27] https://arxiv.org/pdf/2307.03718#page=36&zoom=100,96,757
[28] https://ai-frontiers.org/articles/the-hidden-ai-frontier
[29] https://superintelligence.gladstone.ai
[30] https://cfg.eu/ai-governance-challenges-part-3-proliferation/
[31] https://www.aei.org/technology-and-innovation/treading-carefully-the-precautionary-principle
[32] https://cdn.governance.ai/Ord_lessons_atomic_bomb_2022.pdf
[33] http://rand.org/pubs/research_reports/RRA2849-1.html
[34] https://superintelligence.gladstone.ai
[35] https://arxiv.org/pdf/2503.05628#page=36&zoom=100,120,392

9/11. And of course, just like aerospace and the defense industrial base, the AI industry will require substantial government assistance in infosecurity to ensure that model-weights are secured to the socially-optimal level[36][37][38].

One proposal has been for SL5 optionality—federally-subsidized research and construction into the tools necessary for labs to quickly implement SL5[39]. While every step in this direction is undeniably helpful, aiming for SL5 optionality ignores the reality that even the preparation to build ultra-hardened datacenters, which includes moving the production of all data-center components to American soil, is expected to cost as much or more than the entirety of the Manhattan Project (adjusted to 2020 dollars)[40][41].

We therefore regard SL5 as a longer-term objective and focus on coordinating optionality for SL4 measures in the near term.

III. The Case for OC-4 Theft

Offense bias. Rather than a peer superpower, the most urgent security threat in the near term may come from Operational Capacity 4 (OC4) actors — countries or groups with moderately advanced cyber capabilities. The danger is that an OC4 actor could steal a frontier model's weights and inadvertently trigger secondary proliferation (further uncontrolled spread of those weights). A key reason is the often offense-biased nature of cyberspace operations. Nation-state cyber programs consistently prioritize offense over defense, creating a structural imbalance where penetration capabilities far outpace security hardening. Many U.S. military and policy analysts assess that cyberspace gives attackers an upper hand[42], although some experts argue the balance is more context-dependent[43]. Budgets reflect this: an estimated 90% of U.S. federal cyber spending is devoted to offensive operations[44]. This asymmetry is reportedly even more pronounced in weaker states. Countries like North Korea and Iran, unable to match the U.S. in conventional military power, see cyber warfare as a great equalizer and pour resources into hacking tools — often buying zero-day exploits on the gray market — while neglecting their own network defenses. Consequently, even a relatively unsophisticated opponent can occasionally exploit these offense-focused regimes' weak defenses. For example, North Korea's hackers have pulled off high-difficulty heists like stealing Russian missile blueprints[45]. Yet North Korea's own infrastructure is so insecure that a single

---

[36] https://www.cisa.gov/topics/cyber-threats-and-advisories/nation-state-cyber-actors

[37] https://scholarlycommons.law.northwestern.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1040&context=nulr

[38] https://www.rand.org/pubs/external_publications/EP66656.html

[39] https://ifp.org/a-sprint-toward-security-level-5/

[40] https://ifp.org/a-sprint-toward-security-level-5/

[41] https://www.nps.gov/mapr/faqs.htm

[42] https://www.yalejournal.org/publications/cult-of-the-cyber-offensive-misperceptions-of-the-cyber-offensedefense-balance

[43] https://apps.dtic.mil/sti/html/tr/AD1114539/index.html

[44] https://www.c4isrnet.com/home/2017/03/29/90-percent-of-federal-cyber-budget-used-for-offensive-ops

[45] https://www.euractiv.com/section/global-europe/news/north-korean-hackers-stole-secrets-of-russian-hypersonic-missile-maker/

freelance hacker from the U.S. was able to bring down the entire North Korean internet for several weeks [46]. Given this offense-defense imbalance, if an OC4-level adversary steals a cutting-edge model, there is a significant chance that the actor will fail to contain it.

Leak risk. Once a transformative model's weights are stolen, each additional party holding them multiplies the opportunities for further compromise. More custodians mean more potential breach points—more servers, user accounts, backup drives, and personnel that others could target. Advanced AI models also introduce a speculative new threat: *self-exfiltration.* In controlled experiments, some AI systems have demonstrated rudimentary "escape" behaviors when mishandled[47]. On that basis, some researchers, including the former head of safety at OpenAI, hypothesize that a sufficiently advanced AI might intentionally try to copy or release its own weights under certain conditions[48][49]. Even one successful "escape"—for instance, a model covertly uploading its weights to an external server—would drastically expand the attack surface. Once those weights are publicly accessible, potentially countless new actors could obtain and run the model.

Weak defenses. Critically, an OC4 adversary that steals a model is unlikely to secure it better than the original lab did. Many nation-states with decent offensive cyber capacity rely on outdated systems and poor practices to protect their own networks. If, for instance, an intelligence agency in Iran or North Korea obtained a copy of a frontier model, it probably could not implement more than roughly SL2-level protections on those weights. That implies only basic precautions — passwords, firewalls, maybe some air-gapping — nothing close to the hardened infrastructure of a top tech company. Inexperienced with containing advanced AI systems, such an agency might not even detect a model's self-initiated exfiltration attempt. Human factors pose additional risk: the thieves themselves might leak or sell the model. Insiders could have ideological or financial motives to share the AI. For example, an individual who believes AI should belong to everyone (or who wants a big payday on the black market) might decide to re-release a stolen model. Combining these factors, it is **likely** that a model stolen by an OC4 actor will leak again. In effect, a theft by a mid-tier actor could cascade so that weights fall into many more hands, even if the original thief never intended such an outcome. This would have serious implications on the stability of the current world order.

A longstanding realist consensus holds that multipolarity is less stable than bipolarity[50]. As such, governments around the world have decided that the proliferation of nuclear powers increases the risk of crisis instability. While the overall effect of nuclear proliferation on the stability of deterrence is debated, most work in this tradition treats "more actors with nuclear weapons" as a driver of instability[51]. The same logic applies to the diffusion of AI systems capable of automating AI R&D: as the number of custodians grows, systemic risks compound.

More Flashpoints. First, more actors mean more flashpoints. In nuclear history, each additional weapons state created new dyads and potential crises—India–Pakistan, North Korea–South Korea, and others—that expanded the universe of possible escalation pathways. AI parallels this dynamic: as middling powers acquire frontier-class models, each can serve as a launch point for cyber or

---

[46] https://www.wired.com/story/p4x-north-korea-internet-hacker-identity-reveal/
[47] https://www.anthropic.com/research/alignment-faking
[48] https://www.apolloresearch.ai/research/scheming-reasoning-evaluations
[49] https://aligned.substack.com/p/self-exfiltration
[50] https://www.jstor.org/stable/2538981?seq=2
[51] https://politicalscience.stanford.edu/publications/spread-nuclear-weapons-debate-renewed-second-edition

bioattacks. Because attribution in these domains is difficult and contested, even a small actor can generate incidents that drag larger states into crises they did not anticipate.

Premature deployment. Second, uncontrolled proliferation could push governments into hasty actions. As more actors develop and use offensive AI, states could rush to deploy defensive AI (in cyber defense, early warning, or battlefield advice). Because deployment happens under pressure, accidents become more probable. In all such scenarios, the window for careful governance would vanish, and the decisions made under panic conditions are exactly when mistakes with powerful technologies will be most likely.

Governance hurdles. Finally, widespread proliferation of powerful models would severely undermine efforts at global AI governance. Many proposed schemes to manage advanced AI—such as international monitoring, shared kill-switch mechanisms, or usage restrictions—rely on only a few actors possessing the most capable systems. This is analogous to nuclear arms control: the Non-Proliferation Treaty succeeded only while a small club held nuclear weapons But if advanced AI models become ubiquitous, the situation would more closely resemble the challenge of regulating software programs: any determined group could obtain or train a dangerous system, so enforcing rules becomes nearly impossible. Any regime to reliably contain AI risks hinges on scarcity of the capability. Once a transformative model is leaked and copied globally, coordinating safety measures or usage policies across dozens of countries (and potentially violent non-state actors) becomes extremely difficult. In summary, an OC4 theft could **effectively end any meaningful effort at global AI governance.**

IV. The Case for SL4 Optionality

SL4 optionality. How can we avert the above scenario while avoiding the pitfalls of an SL5-focused approach? This paper argues that the most urgent pragmatic goal for policy on AI security is to build **Security Level 4 (SL4) optionality** across frontier AI labs. SL4 optionality means building the capacity for labs to rapidly harden against OC4-level theft once capability thresholds are crossed. Doing so directly addresses the instability channels described above: it slows the multiplication of flashpoints by keeping the set of capable custodians small; it reduces attribution risk by limiting the number of plausible aggressors; and it prolongs the credibility of pause efforts by preventing premature diffusion. SL4 optionality is thus not just a technical safeguard but a structural intervention to preserve stability.

In practice, this means developing and implementing measures to protect cutting-edge model weights against OC4-level threats (skilled nation-state or large criminal syndicate hackers), and being able to do so quickly when the situation demands. SL4 is one step down from the ultra-hardened SL5 that aims to defeat superpower adversaries. It is designed to thwart actors like North Korea, Iran, or sophisticated non-state groups — adversaries very capable at intrusion but with weaker follow-on defenses. By focusing on SL4, we target the most likely source of an uncontrollable model leak. Crucially, improving security to SL4 does *not* mean immediately halting progress or locking down AI development across the board. It means preparing the legal, technical, and logistical infrastructure so that when models start approaching a dangerous capability threshold, labs can rapidly upgrade their security (with government coordination) to prevent an OC4 breach. This strategy accepts that SL5 (defending against China/Russia-level threats) might not be attainable today, but asserts we *can* reach SL4 on short notice.

Underinvestment. Achieving robust SL4 security will not happen by default in industry. The history of cybersecurity shows that private firms often underinvest in security relative to the social optimum, especially in sectors with large externalities. Even if a lab spends up to the amount of its own expected losses from cyber theft (an estimate often on the low side), that calculation ignores the far greater harm a model theft could inflict on society at large[52]. For example, a model specializing in biology could be stolen and repurposed to create a virus capable of a repeat of COVID-19—entailing losses a hundred times beyond what any single company internalizes. It is widely recognized in policy research that operators of critical infrastructure systematically under-protect against worst-case threats[53]. Frontier AI labs are no exception. Moreover, the precedent for state-sponsored IP theft is well established. Advanced software and defense companies face constant intrusion attempts from even middling adversaries. North Korea, Iran, and others have a track record of stealing valuable intellectual property from Western firms (and from each other). As noted, even non-state hacker groups like Lapsus$ have breached supposedly well-secured targets[54]. We should assume that any sufficiently valuable model will be an irresistible target for multiple hostile actors. Thus, leaving security purely to the labs' discretion will likely result in protection levels that are inadequate against an OC4 adversary.

Race to the bottom. In fact, labs themselves implicitly acknowledge this collective action problem. In public statements and policies, leading AI companies have only committed to roughly SL3-level security measures even for very large-scale models. For example, current "Frontier Model" guidelines mention improving cybersecurity and insider threat management, but nothing like relocating to bunkers or implementing pervasive staff surveillance. Top lab executives have noted that no single company can afford to ramp up security unilaterally without sacrificing its competitive position. If one lab imposed stringent security (slowing its pace or raising costs) while others did not, that lab would fall behind in the AI race. In game-theoretic terms, every lab is waiting for others to move first, since a single insecure actor can make all labs vulnerable. Absent coordination or regulation, the equilibrium is a race to the bottom on security — essentially, all labs stick to minimal measures, and everyone remains exposed to the weakest link.

The Costs of SL-4. The reluctance of labs to voluntarily adopt SL4 measures is understandable when we examine the concrete costs. Consider two major components of SL4: confidential computing and rigorous insider vetting.

Compute overhead. *Confidential computing* refers to techniques like encrypted memory and secure enclaves that make it much harder to exfiltrate data during computation (e.g. during an AI training run or inference). Today, turning on confidential computing features incurs a significant performance and cost penalty. On Google Cloud, for instance, using confidential VMs can add roughly 9% to the cost of an AI-optimized GPU like the H100[55], and that's before considering any slowdown in computation speed. In practice, tests have shown around a 7% performance degradation on AI workloads when running under encrypted enclaves[56]. Stacked together, a 12–22% increase in effective cost per operation is a reasonable estimate for using confidential computing at scale. If a

---

52 https://scholarlycommons.law.northwestern.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1040&context=nulr

53 https://www.mwftr.com/CS2/The%20limitation%20of%20Safety-Chapter%201.pdf

54 https://www.avertium.com/resources/threat-reports/in-depth-look-at-lapsus

55 https://cloud.google.com/confidential-computing/confidential-vm/pricing?hl=en

56 https://phala.network/posts/GPU-TEEs-is-Alive-on-OpenRouter

frontier lab plans to spend $500 million on cloud compute for AI model training or deployment, implementing confidential computing could burn an extra $60–110 million for the same work. In an industry where progress is often gated by available compute, a 10–20% hit is a huge competitive disadvantage. It's no surprise, then, that labs are hesitant to embrace such measures unless compelled.

Insider costs. The other pillar of SL4 is tightening insider access to prevent an Edward Snowden–style leak of model weights. A basic requirement would be a top-secret security clearance for anyone handling the most sensitive model-weights. This would be a notch below the level of clearance undergone by nuclear engineers. However, any such steps would materially slow hiring and research. For a rough sense of scale: obtaining a Top Secret clearance in 2023 took half a year on average and $5k[57]. Interim clearances are possible, but only for U.S. nationals.

Unfortunately, over half of all top researchers at frontier AI labs are estimated to be non U.S. nationals.. Since AI researchers can now command hundred million dollar per year pay packages[58], and insiders estimate there may be hundreds of them who need security clearances. Such a requirement could be unacceptably costly for any individual lab to accept.

Insiders estimate AI labs currently have hundreds of employees with access to model weights[59]. Getting them all security clearances could amount to tens of millions of dollars in administrative costs alone — and more importantly, months of lost productivity as new researchers wait in limbo. Many AI researchers, especially the most sought-after PhDs, are foreign nationals (over half of top-tier AI research talent was trained outside the U.S. [60], and one report found "well over 50%" of researchers at some U.S. labs are foreign-born[61]). Many of these individuals may not qualify for security clearances—those that do must still face the efficacy cost of losing access to model-weights for six months or more. All told, a lab could easily see an 8–50% reduction in first-year productivity for each researcher under SL4-level insider controls. If the average fully loaded cost of an AI researcher is $10M/year, and 100 people need clearances, that's on the order of $80M to $500M in productivity sacrificed in the first year of a security ramp-up. This doesn't even count the morale and culture impacts of heavy-handed security (which can drive talent to seek out more open work environments). Again, no rational company will incur these costs unless every competitor must do so.

Non-proliferation. Because market forces push against voluntary adoption of SL4, policy intervention is needed to reach this security level in time. By mandating or coordinating an industry-wide move to higher security when appropriate, the government can solve the collective action problem. The strategic payoff of achieving SL4 is immense: it would significantly reduce the risk of a model weights theft by an OC4 actor and thereby prevent the worst-case secondary proliferation outcome. In effect, this approach treats advanced AI models similarly to nuclear materials — something that must be kept to a few responsible stewards. During the Cold War, the Nuclear Non-Proliferation Treaty (NPT) helped ensure that only a small number of nations held nuclear weapons, which in turn enabled strategic stability and arms control. Likewise, if the U.S.

---

[57] https://assets.performance.gov/files/Personnel_Vetting_QPR_FY24_Q1.pdf
[58] https://archivemacropolo.org/interactive/digital-projects/the-global-ai-talent-tracker/
[59] https://www.rand.org/pubs/research_briefs/RBA2849-1.html
[60] https://timesofindia.indiatimes.com/etimes/trending/who-is-trapit-bansal-the-indian-origin-ex-openai-engineer-behind-metas-ai-push/articleshow/122187282.cms
[61] https://superintelligence.gladstone.ai

and perhaps a few close allies are the only ones to develop AI models that can automate AI R&D (and if those models are well secured), it will be far easier to manage the risks. We can avoid a scenario where dozens of states (or a group like ISIS) get their hands on such extremely advanced AI. Rather than trying to be the only state with transformative AI, the U.S. should strive to prevent uncontrolled diffusion of transformative AI — channeling great power competition into other arenas (e.g. economic strength via AI-enabled industries, or conventional military hardware like drones). Importantly, preventing proliferation is not about hoarding AI's benefits; it's about buying time for global governance and ensuring that when these powerful AIs are deployed, they are deployed responsibly. If we succeed, we maintain the current world order's stability a bit longer, giving us a chance to set up international agreements or safety standards before the technology becomes unmanageable[62].

Open-source threat. Another critical reason to prioritize SL4 security is to avoid the irreversible public release of a dangerous model. It's one thing if an adversarial state like China steals a model — that would be a serious strategic setback, but at least the capability remains in a few hands. It's quite another if model weights sufficient to build WMDs or autonomous weapons become freely available on the internet. Unfortunately, as discussed, an OC4 actor with lax security is a prime vector for such a public leak. Many OC4-level governments could only secure stolen AI at roughly SL2, meaning their safeguards would be trivial for skilled hackers (or malicious insiders) to penetrate. It is very likely that if a model with world-changing capabilities spreads to weaker hands, it will eventually be dumped online. There are even ideologues who favor this: so-called "AI liberationists" believe that preventing AI proliferation is unethical or futile. AI textbook author Rich Sutton has said that AIs should be liberated since "succession to AI is inevitable," "we should not resist succession," and "it behooves us... to bow out" [63]. A lab employee or state scientist sympathetic to such views might intentionally open-source a powerful model "for the good of humanity" (or for the good of the AI itself), not appreciating the catastrophic misuse potential. The most reliable way to stop this chain of events is at the source: ensure that the frontier labs (the ones training such models in the first place) are secure enough that even an OC4 adversary cannot steal the weights. Achieving SL4-level security across all top labs would greatly reduce the chance of an initial theft, thereby keeping these powerful models contained for longer.

Further research. Focusing on SL4 optionality is a middle path that avoids both reckless complacency and impractical overreaction. But implementing it raises many questions that need more study. What are the most cost-effective technologies or practices to achieve SL4? How can we design confidential computing hardware or software that has minimal performance overhead? Can we create new mechanisms for insider threat reduction that don't drive talent away (for instance, federated learning approaches that limit full access to models, or AI tools to monitor unusual behavior)? What emergency protocols should be established so that if a certain capability threshold is passed, all labs can quickly enter a "high alert" security posture? These are the kinds of issues that a major research program on frontier AI security should tackle now. In essence, we need to develop an SL4 toolkit that is ready for rapid deployment. That might include improved encryption methods, secure hardware modules, auditing algorithms to detect AI self-exfiltration attempts, and templates for government–industry coordination during an AI crisis. Identifying the most important areas for investment (and potential bottlenecks) will be crucial. In short, while this paper outlines *why* SL4

[62] TK (Historical analog: early U.S. efforts to keep atomic research classified under the Manhattan Project, followed by the establishment of international norms around nuclear technology.)
[63] https://arxiv.org/pdf/2503.05628

optionality is the right goal, there is much work to be done on *how* to actually achieve it. The time to start that work is now, before the need becomes urgent.

V. Counterarguments and Limitations

Uncertain balance. It is important to acknowledge counterarguments and uncertainties in this analysis. One debate in the AI security community is the offense–defense balance for advanced AI. While we have argued that wide proliferation is dangerous, some experts note that proliferation might also yield defensive benefits. If many actors have access to powerful AI, it could help diffuse the benefits of AI and reduce the concentration of power, potentially enabling more parties to develop strong defenses (Eiras et al., 2024; Kak and West, 2023). In theory, more "good guys" with AI could counteract the "bad guys" with AI. Additionally, it remains genuinely uncertain whether offense or defense will be easier with extremely capable AI (Seger et al., 2023; Corsi et al., 2024) – it is possible advanced AI will make it easier to defend against threats than to execute attacks. We are fundamentally unsure about the relative difficulty of attacking versus defending in a world of highly capable AIs. However, prudence suggests erring on the side of caution when dealing with technologies that could enable extreme catastrophes. As one group of researchers put it, "at the threshold where models can rapidly improve to the point where they can cause extreme catastrophes, caution is warranted until it is demonstrated that society is resilient to catastrophic attacks" [64][65]. In other words, until we have evidence that our institutions and defenses can cope with super-powerful AI in the wrong hands, it is wise to limit proliferation even if doing so means forgoing some defensive advantages.

AI for defense. Another counterpoint is the idea that AI itself will dramatically improve cybersecurity and information security, possibly addressing some of the concerns raised here. Could advanced AIs be deployed as near-omniscient digital guardians — instantly patching vulnerabilities, detecting intrusions, and even managing model access more effectively than humans? It is certainly possible that AI will bolster defense; indeed, many cybersecurity vendors are already integrating AI to identify malware and anomalous behavior. Over the longer term, AI copilots might help organizations consistently follow best security practices. That said, it would be dangerous to rely on unproven defensive AIs to counter equally advanced offensive AIs. Historically, defensive adoption tends to lag offensive innovation. One analysis warns that offensive actors could quickly exploit new AI tools to target existing vulnerabilities, while defensive adoption will be slower due to the need for rigorous evaluation, approvals, coordination, and deployment to critical systems[66]. This dynamic risks a persistent defensive lag without intervention. In simpler terms, even if "good" AIs can plug holes, "bad" AIs may exploit new weaknesses faster. Furthermore, if we ever reach the point of AI systems with strategic-level planning skills, we must consider that those AIs might circumvent or even subvert our defensive measures (especially if a model is tasked with its own survival or replication). For safety, then, we cannot assume AI will eliminate the need for strong human-driven security policies. Rather, we should assume we need to secure the AI until it's proven that the AI can securely handle itself.

---

[64] https://arxiv.org/pdf/2405.10295

[65] https://saif.org/wp-content/uploads/2025/04/Bare-Minimum-Mitigations-for-Autonomous-AI-Development.pdf

[66] https://ifp.org/a-sprint-toward-security-level-5/

Resource constraints. One might argue that rogue states like North Korea or Iran, even if they stole a cutting-edge model, would lack the computing resources (chips, energy, infrastructure) to make serious use of it. Why invest so much to prevent theft by OC4 actors if those actors can't effectively run or iterate on the model once they have it? There is some merit to noting the resource gap: training state-of-the-art models requires advanced hardware that these countries can't easily obtain at scale due to export controls and cost. However, this argument underestimates how a stolen model could still be weaponized with minimal resources. Once you have the weights, you don't need to retrain the whole model from scratch — you can fine-tune it for specific tasks relatively cheaply. For example, a stolen general model could be fine-tuned (even on a modest compute cluster) specifically for offensive cyber operations such as spear-phishing or vulnerability discovery. Reports from the national security community suggest that fine-tuning an AI for specialized malicious tasks might be unusually cost-effective[67]. Even intermittent access to a small number of high-end GPUs could allow a rogue actor to run inference with the stolen model to generate, say, biochemical weapon designs or sophisticated propaganda. Moreover, hardware availability can change: black markets exist for advanced chips, and cloud computing resources can be rented covertly. We should not assume that adversaries' resource limitations or technical shortcomings will indefinitely prevent them from exploiting a stolen model. It's also worth noting that as AI research progresses, efficiency improvements may drastically reduce the cost to run powerful models, lowering the barrier for anyone to use them.

Scope limits. Finally, it's important to clarify the scope and limitations of this paper. Our focus has been on arguing for a reframing of frontier AI security strategy, not on providing detailed policy blueprints. We have deliberately not delved into prescribing specific laws or international agreements; rather, we've aimed to explain why the conversation should pivot to SL4 optionality as a primary goal. This means many practical questions are left unanswered here. We have also largely discussed the development of AI capabilities in one particular order — roughly, that models will become capable of automating AI R&D (and thus potentially dangerous via rapid self-improvement) before, say, they become capable of reliably engineering novel bioweapons or other single-domain catastrophic tools. It's possible the sequence will differ. If, for example, weaponization capabilities (like the ability to design dangerous pathogens) emerge before self-improvement, the calculus on model security might shift (perhaps an earlier lock-down would be warranted, even absent an "intelligence explosion" scenario). We did not explore every permutation of capability emergence. Similarly, we did not address in detail the international dimension: how the U.S. might convince or compel other countries to also adopt SL4 security, which will be crucial in a world of multinational AI labs. All these are fertile topics for further research. Each limitation noted here suggests a corresponding follow-up: we need studies on the efficacy of AI in cybersecurity, scenario planning for different AI capability timelines, and concrete policy design for implementing SL4 measures (ranging from R&D incentives to potential regulation or licensing of frontier model training). In summary, our argument should be seen as a starting point for reframing priorities, rather than the final word on how to achieve safe and secure AI development.

VI. Conclusion

67 https://www.belfercenter.org/sites/default/files/2024-10/Artificial%20Intelligence%20and%20National%20Security.pdf

SL4 over SL5. In the coming decade, AI capabilities may advance with potentially unprecedented speed, and the security of frontier AI labs could become a key factor for global stability. This paper has argued that our policy focus must shift away from the unattainable ideal of instant SL5 and toward the pragmatic goal of ensuring SL4 readiness. Achieving true SL5 (defense against a top-tier superpower) is not feasible today — technically, economically, or politically. But we can strive to equip all frontier labs with the tools and plans to reach SL4 (defense against moderately sophisticated threats) at a moment's notice. This focus is justified by two key points: (1) Market pressures and high costs mean labs will not, left to their own devices, invest in SL4-grade security in time — a collective action problem that policy can solve. (2) The most severe near-term danger is not an immediate "China steals GPT-10" headline, but rather the scenario where a smaller adversary steals a powerful model and inadvertently lets it proliferate uncontrollably. SL4 security directly targets that failure mode by keeping such actors out. By building SL4 optionality, we buy crucial time: we keep the number of actors with extremely advanced AI low and controllable, and we ensure that if warning signs of extreme AI capability appear, we can rapidly lock down accordingly.

Reframing needed. In conclusion, the conversation around frontier AI security should move from a race mindset to a risk-mitigation mindset. Rather than asking "How do we beat China to AI and then keep it from them?", we should be asking "How do we prevent any actor from allowing transformative AI technologies to proliferate uncontrollably, at least until we have global safeguards in place?" This reframing leads to different priorities: cooperation with labs on security standards, investments in containment technologies, and international dialogues about managing dangerous capabilities. The SL4 optionality approach is essentially an insurance policy for the AI revolution — ensuring that if the most aggressive AI development forecasts come true, we will not be left wishing we had implemented stronger safeguards at our AI labs. It steers us toward a future where AI's benefits can be reaped without triggering a destabilizing runaway escalation in AI capability. Given what is at stake with frontier AI, we may not get a second chance to implement effective security measures. The time to lay the groundwork for SL4 security is now, before progress outpaces our preparedness. We should proceed with the urgency implied by the fastest plausible AI timelines, even as we hope those worst-case scenarios never materialize.