Working Paper

SELLA NEVO, DAN LAHAV, AJAY KARPUR, JEFF ALSTOTT, JASON MATHENY

Securing Artificial Intelligence Model Weights

Interim Report

For more information on this publication, visit **www.rand.org/t/WRA2849-1**.

Published by the RAND Corporation, Santa Monica, Calif.

About This Working Paper

As frontier artificial intelligence (AI) models become more capable, protecting them from malicious actors will become more important. This working paper offers early takeaways from research into what it would take to protect model weights (parameters) from a range of potential malicious actors. At a high level, **if AI systems rapidly become more capable over the next few years, achieving sufficient security will require investments—starting today—well beyond what the default trajectory appears to be.**

The full report detailing the complete study will be published in early 2024. Those interested in providing insights for the report or getting access to a work-in-progress draft can contact ai-security@rand.org.

Funding

Funding for this research was provided by gifts from RAND supporters.

RAND Meselson Center

This work was undertaken by the RAND Meselson Center, which is dedicated to reducing risks from biological threats and emerging technologies. The center combines in-depth policy research with state-of-the-art technical research to provide policymakers with the information and expertise needed to prevent, prepare for, and mitigate large-scale catastrophes, such as pandemics.

Acknowledgements

We thank Michael Vermeer, Mary Vaiana, and Arwen Bicknell for their support in reviewing and editing this working paper.[1] All errors are the sole responsibility of the authors.

[1] The full report will include acknowledgments of experts who provided insights for the report (in line with confidentiality preferences).

Contents

Figure and Tables

Figures

Tables

1. Overview of the Interim and Full Reports

Researchers at the RAND Corporation and other collaborators are conducting an ongoing research project to help inform ongoing conversations around securing artificial intelligence (AI) models. This interim report provides some preliminary key highlights from that research (Figure 1.1); the full report, which will be released in early 2024, will explore attack vectors that malicious actors can use to steal frontier AI model weights and offers benchmarks and recommended actions to secure them from a variety of potential malicious actors.

Figure 1.1. Key Elements and Findings

Frontier AI labs—and, specifically, their AI model weights—face expansive information security threats.

We identify approximately 40 meaningfully distinct attack vectors.

A lab's vulnerability to just one of these vectors can compromise its security. We provide hundreds of real-world examples in which these attack vectors were deployed successfully, demonstrating that they are feasible and providing context on what such attacks look like in practice.

We explore a variety of potential attacker capabilities, from opportunistic (often financially driven) criminals to highly resourced nation-state operations.

This categorization of attacker capabilities allows labs to identify sequential priorities depending on their current security infrastructure.

We estimate the feasibility of each attack vector being executed by different categories of attackers.

About one dozen attack vectors are likely infeasible for nonstate actors, but they are feasible for state actors, highlighting the need for significantly more-capable security systems to defend against state actors. Expert opinions vary significantly on the capabilities of state actors and how to defend against them.

We propose and define five security levels; recommend preliminary benchmark security systems that roughly achieve the security levels.

Each security level is defined by being secure against attack vectors feasible for increasingly capable categories of malicious actors. The benchmarks can help to calibrate the trade-off between security investment and protection against different actors. They are not meant to be used as a standard, but they provide concrete suggestions for steps that frontier labs can take at different stages of their continuous security enhancement strategy.

As part of our research, we incorporated insights and recommendations from the academic literature and other written sources, synthesized evaluations of real-world systems, and interviewed more than 30 experts. These consisted of senior information security personnel at top AI labs, national security professionals from multiple countries considered leaders in cybersecurity, experts from the information security industry (both offensive and defensive), and distinguished AI scientists and business leaders.

The remainder of this working paper is as follows: Chapter 2 provides the context and motivation for our research. Readers should take particular note of the recommendations in Chapter 3 of this working paper: We suggest several actions that seem particularly urgent for protecting model weights in the short term while also highlighting critical efforts that should be initiated early to avoid bottlenecks as labs seek to secure against nation-state attacks in the coming years.

In Chapter 4, we share a subset of possible attack vectors, including ones used by highly resourced operations of nation-states. Despite a lack of consensus on the feasibility of some of the more-advanced attack vectors, we find significant real-world evidence that such attacks take place. In Chapter 5, we preview several benchmarks for security systems and estimate which categories of actors are able to overcome the benchmark systems at different levels. These estimates can help labs to calibrate the effectiveness of existing or planned security measures in protecting against actors with varying capabilities and suggest recommended next steps for security teams tasked with securing AI weights.

2. Context and Motivation

The rapid expansion of AI capabilities could make AI models a highly valuable target for theft. State actors, in particular, who have historically used theft and exploitation of intellectual property to advance their interests, might seek to co-opt advanced AI capabilities (Clark, 2023) and potentially apply them to malicious ends. In one concrete example (Anthropic, 2023), biosecurity experts studying state-of-the-art models have warned that malicious actors could misuse these models for biological weapon development. Although the scale of such risks remains moderate today, the trajectory suggests a growing threat landscape. Even if most models are eventually found not to be a risk and are open-sourced, it is important to secure them while they are still being evaluated or if they are found to be dangerous.

It is therefore urgent that AI labs develop a comprehensive security strategy that addresses both immediate and long-term threats. For labs to develop such a strategy, they must build a robust understanding of vulnerabilities specific to their systems and of the various threat actors, such as foreign intelligence agencies. Implementing robust security measures—capable of both preventing theft and mitigating misuse—will require years of active investment and planning.

Model weights are the learnable parameters of a machine learning model, often a deep neural network. We believe model weights are a particularly critical component to protect because they uniquely represent the result of many different costly and challenging prerequisites for training advanced models—including significant compute, collected and processed training data, algorithmic optimizations, and more. In some (though not all) cases, acquiring the weights could allow a malicious actor to make use of the full model at a tiny fraction of the cost of training it. Therefore, we believe that weights should be given special attention in a lab's security strategy. Weights are, of course, not the only component that needs to be protected, but they are the focus of our research.

In this interim report, we share highlights from our work to help inform ongoing discussions in the broader AI security arena.

3. Top Recommendations

We offer three recommendations for frontier AI labs working on novel AI capabilities that may have strategic implications within the next few years.

Recommendation 1: Establish a Road Map Toward Securing Systems Against All Threat Actors, Up to and Including State Actors Executing Highly-Resourced Operations

Such a road map would include the implementation of substantial protections against all mapped attack vectors. We explore a benchmark for such a system under the title of security level 5 (see Table 6.5 in Chapter 6). Based on the requirements identified, we believe it is unlikely such a system can be achieved within several years unless planning and significant efforts toward this goal begin soon.

Recommendation 2: As an Urgent Priority, Implement Measures Necessary for Securing Systems Against Most Nonstate Attackers

Safeguarding frontier AI models against attack vectors generally available to nonstate actors—such as hacker groups, insider threats, and terrorist groups—is feasible and urgent (see Table 6.3 in Chapter 6). Though comprehensive implementation of security measures for all relevant attack vectors is necessary to avoid significant security gaps, we highlight several particularly important measures:

- Limit legitimate access to the weights themselves (for example, to less than 50 people).
- Harden interfaces to weight access against weight exfiltration.
- Engage in advanced third-party red-teaming (implemented with several measures that are not currently industry standards).
- Define multiple independent security layers (a strict implementation of the concept of defense-in-depth).
- Invest in insider threat programs.

Recommendation 3: Begin Research and Development and Experimentation Today on Particular Aspects of Securing Systems Against Advanced Threat Actors

Developing, implementing, and deploying some of the efforts required to secure against highly resourced state actors are likely to be bottlenecks even at longer timescales (for example,

five years). Therefore, it might be wise to begin these efforts soon. Some of the most critical bottlenecks include

- constructing physical bandwidth limitations between devices or networks containing weights and the outside world
- developing hardware security model (HSM)–inspired hardware to secure model weights while providing an interface for inference
- setting up Sensitive Compartmented Information Facility (SCIF)–style isolated networks for training, research, and other, more advanced interactions with weights.

In the remainder of this interim report, we provide more context and share additional conclusions and recommendations that will be described in depth in the full report.

4. Attack Vectors—Highlights

This chapter highlights a subset of approximately 40 attack vectors that will be detailed in the full report. These are vectors that attackers could deploy to circumvent the security measures protecting an AI model (see Table 4.1 for samples). Empirical evidence and systematic evaluations suggest that virtually all of these attack vectors have already been deployed in real-world environments and pose concrete risks.

Here are the key takeaways from our research on the attack vectors:

- The attack vectors are not theoretical. There is empirical evidence showing that these attack vectors are actively executed (and, in some cases, even widely deployed).
- The diversity of attack vectors is large, so defenses need to be varied and comprehensive. Achieving strong security against a specific category of attack does not protect a lab or company from many other types of attack.
- Advanced malicious actors often combine multiple attack vectors to execute more-complex operations. Such combinations can more effectively penetrate layered defenses by building on each other. Some experts say they believe that combining multiple attack vectors is the most significant difference between more-advanced actors and less-advanced ones, even more so than the feasibility of any specific attack vector.
- Many attack vectors have become democratized and widely accessible (for example, a $180 USB cable that provides full remote control of a device [Hak5, undated]).
- The U.S. *Annual Threat Assessment* report (Office of the Director of National Intelligence, 2023) estimates that multiple foreign agencies are able to penetrate and disrupt varied types of critical infrastructure. Thus, even the level of protection deployed by critical infrastructure industries (though a good start) is not sufficient to defend against state actors.
- The most-capable advanced persistent threats (APTs) are often not detected, and most APTs persist for years before discovery. Thus, conclusions based exclusively on public information and operations that were detected systematically underestimate what is possible and the frequency at which advanced attack vectors are (successfully) utilized.
- Highly resourced attacks by top cyber-capable state actors (a small but important subset of all operations executed by such states) are the apex of cyber operations, characterized by their complexity, associated risks, and the immense resources required. As a reference point, these attacks can be thought of as the top ten priority operations conducted annually by a leading cyber-capable state. The financial and technical scale of these operations, combined with the secrecy surrounding them, make it extremely challenging to defend against them. Though defending against such efforts might be important in the long term, this should not deter companies from securing their systems against attacks from less capable actors. Attacks in the latter category are much more common, and defending against them is more urgent.

Table 4.1. Sample of Attack Vectors

Category	Description	Attack Vectors
Running Unauthorized Code	Exploitation of vulnerabilities to run code dictated by an attacker	• Unpatched software with known vulnerabilities • Zero-days (that is, vulnerabilities for which no patch exists)
Compromising Existing Credentials	Misuse of legitimate user credentials; possible stepping stone to advanced attacks	• Social engineering • Password cracking • Exploitation of exposed credentials
Undermining the Access Control System Itself	Attacks on mechanisms controlling data access	• Encryption or authentication vulnerabilities • Intentional backdoors • Code vulnerabilities
Bypassing Primary Security System Altogether	Use of overlooked alternative paths to avoid access control	• Misconfigured software • Insecure backups or copies • Default passwords
Abuse of Legitimate Application Programming Interfaces (API)	Use of a legitimate interface without privileged access to information	• Extraction of weights • Model distillation
Nontrivial Access to Data or Networks	Sophisticated and resource-intensive attacks on hardened systems	• Breaching air-gapped networks • Transient Electromagnetic Pulse Emanation Standard (TEMPEST) attacks • Eavesdropping and wiretaps
Unauthorized Physical Access to Systems	Physical access to a system translated into meaningful access to sensitive information on the system	• Direct physical access • Malicious portable devices • Evasion of physical access control systems (PACS) • Armed operations
Supply Chain Attacks	Attacks on third-party vendors used by the company	• Services and equipment • Software infrastructure • Vendors with sensitive access
Human-Based Infiltrations	Attacks leveraging human influence	• Bribes and cooperation • Extortion • Candidate placement

The list in Table 4.1 is a sample and does not aim to be comprehensive. The extended list in the full report will be significantly more comprehensive, but it still will not capture all potential attack vectors. We recognize that unidentified vectors likely exist: Publicly disclosed attacks represent only a fraction of all activities, particularly regarding APTs and state-sponsored operations.

The full report will provide detailed definitions of each category and attack vector, along with multiple examples.

5. Risk Estimates for Attack Vectors—Highlights

We estimate the risk of each attack vector from different types of actors. These estimates have been calibrated in consultation with a wide variety of information security experts.

The risk estimates take into account the probability that a vulnerability exists in a target system, whether the attacker can exploit the vulnerability (for example, whether they have existing infrastructure or the ability to perform other attacks that are a prerequisite for the one in question), and considerations that influence whether the attacker decides to move forward with executing it (such as costs, risk, or capacity).

Table 5.1 lists some sample risk estimates for a few of the attack vectors. A score of 1 means that the actor is unlikely to be able to carry the attack (that is, low risk). A score of 5 means that the attacker can readily and consistently execute the attack with a high likelihood of success (that is, high risk).

We consider an attack successful if its execution directly contributes to the attacker's ability to steal model weights or otherwise exploit the model, even if the successful execution of the attack does not on its own provide full access to the weights.

Table 5.1. Sample of Risk Estimates

Category	Type or Approach	Opportunistic Attacker	Persistent Nonstate Actor	State Actor (Moderately Resourced)	State Actor (Highly Resourced)
Compromising Existing Credentials	Social engineering	4	5	5	5
Unauthorized Physical Access to Systems	Evasion of physical access control systems	1	1	3	5
Supply Chain Attacks	Software infrastructure	2	4	5	5

6. Security Levels—Highlights

To facilitate more-nuanced discourse on the security needs of different AI systems (depending on their capabilities or other safety concerns), we propose five security levels (SLs)—SL1 to SL5—broadly defined to thwart most attacks from an increasingly capable actor category (see Figure 6.1). For instance, *SL2* is defined as a system that is protected against most attacks by opportunistic actors; *SL5* is a system that is protected from highly resourced, top cyber-capable nation-state operations.

Accurately assessing which real-world systems actually conform to the different security requirements is a challenging and even controversial task because of the specific circumstances of different systems, significant disagreements in the field, and the field's ever-changing nature. However, it is still essential to improve our estimates for the security of real-world systems, and public discourse on this topic can assist labs in protecting themselves better. In the full report, we will share a benchmark system for each SL, listing concrete measures and policies that we believe (based on consultation with experts) roughly make up the minimum requirements of a system that conforms to the goals of that SL. These benchmark systems will likely require significant corrections and adjustments over time, but we hope they will provide some sense of calibration, offer labs some concrete suggestions for next steps, and catalyze more discourse around this topic.

Figure 6.1. Overview of Security Levels

Because an attacker can often exploit the weakest link in a security system to undermine its defenses, a system that exceeds a specific benchmark system in some aspects but falls short in other aspects is, in our opinion, less secure than that benchmark system overall.

In aggregate, the benchmark systems provide roughly 150 recommendations for security measures, tools, and policies.

As a preview to the full report, we define each level in Figure 6.1, present a few example measures to provide a sense of calibration, and offer key insights related to each level (Tables 6.1–6.5).

Table 6.1. Highlights of Security Level 1

Examples of measures	• Apply the least privilege principle. • Use multifactor authentication (MFA). • Develop basic incident response capabilities.
Key insights	• Companies benefit from a variety of protections simply by using modern platforms and cloud providers. • At this level of investment, it is better for companies to rely on existing security products and best practices rather than trying to implement anything on their own because they are more likely to undermine their security than improve it. • Companies can still improve their security easily at this level by choosing stricter security configurations and policies. • However, achieving SL1 is not sufficient to achieve reliable security.

Table 6.2. Highlights of Security Level 2

Examples of measures	• Limit allowed storage location (for example, weights are stored only in servers and not copied to local devices). • Review vendor and supplier security. • Implement secure software development standards.
Key insights	• This level consists primarily of systematically implementing the fundamentals and latest industry best practices across the board. At this stage, ensuring there are no major blind spots left unaddressed is the most important aspect. • Prioritizing the most-common attack vectors is key (for example, ensuring that email security, password policies, and MFA are all enforced correctly). • Policies that improve security throughout the network, codebase, or company are particularly critical—such as a secure software development framework and zero-trust architecture.

Table 6.3. Highlights of Security Level 3

Examples of measures	• Develop an insider threat program. • Deploy advanced red-teaming (for example, elite external team, substantial funding, etc.). • Hire security team members with concrete APT experience.
Key insights	• Aggressively reducing the attack surface is the key theme in this security level. For a persistent attacker, a large attack surface inevitably implies underlying vulnerabilities, even if these are not yet known to the security team or the broader security community. • No less importantly, one must assume at SL3 that an attacker has unexpected access or capabilities—because they are an insider, have zero-days, or have spent more time researching the system than its developers. Robustness to hard-to-predict access and capabilities drive many of the new requirements, including the critical requirements for defense in depth. • The full supply chain, from software and hardware to the air conditioners, needs to be monitored and secured.

Table 6.4. Highlights of Security Level 4

Examples of measures	• Implement hardware protections for weight storage (for example, disabling hardware communication capabilities and protecting against simple TEMPEST attacks). • Conduct ongoing compromise assessment on all devices with privileged access. • Implement strict application allowlisting (that is, only specific binaries are allowed to execute on devices with access).
Key insights	• Attack surfaces need to be reduced to the extent that the remaining security-critical surface can be comprehensively (and often manually) hardened, reviewed, monitored, and pen-tested. Such attack surface reduction requires significantly more compromises on productivity, convenience, and efficiency than previous levels. • For security-critical junctions, any software assurances or general-purpose hardware are no longer considered trusted. Critical security assumptions need to be implemented in hardware, potentially requiring changes in how data centers are built and operated. • Because state actors have extensive capabilities unavailable to other actors, the security team must have specific experience and expertise, which heavily influences how security assessment, red-teaming, and other activities are performed. • Access to large numbers of zero-days and other capabilities that may be years ahead of public knowledge mean that many security redundancies are needed (for example, four independent security layers).

Table 6.5. Highlights of Security Level 5

Examples of measures	• Aggressively isolate weight storage (for example, SCIF or HSM variants; see the insights below). • Implement formal hardware verification of key security components. • Proactively protect executives and individuals handling sensitive materials.
Key insights	• Weights should be stored in SCIF-like setups (disconnected from the external world) with extremely stringent policies on data transfer that would prevent even those with approved access from taking large amounts of data out of the room. • More research and development is needed to enable companies to support production models while meeting SL5 security requirements. Existing technological solutions cannot simultaneously achieve the needed availability and security requirements. We recommend the development of HSM-like devices with an interface that is specialized for machine learning models. • Achieving SL5 will likely require comprehensive support from the national security community.

Abbreviations

AI	artificial intelligence
APT	advanced persistent threat
HSM	hardware security model
MFA	multifactor authentication
SCIF	Sensitive Compartmented Information Facility
TEMPEST	Transient Electromagnetic Pulse Emanation Standard

References

Anthropic, “Frontier Threats Red Teaming for AI Safety,” July 26, 2023.

Clark, Joseph, “AI Security Center to Open at National Security Agency,” *DoD News*, September 28, 2023.

Hak5, “O.MG Cable,” webpage, undated. As of October 3, 2023: https://shop.hak5.org/products/omg-cable

Office of the Director of National Intelligence, *Annual Threat Assessment of the U.S. Intelligence Community*, 2023.