SL5 Standard for AI Security

Preliminary Draft

Version 0.1

MARCH 2026 (UPDATED JUNE 23, 2026)

Lisa Thiergart

Yoav Tzfati

Peter Wagstaff

Guy

Luis Cosio

Philip Reiner

Table of Contents

About the Security Level 5 Task Force

The SL5 Task Force is a non-profit cross-industry effort working to ensure frontier AI infrastructure can achieve nation-state-level security by 2028/2029. Founded in March 2025, we are a core team of engineers and security strategists leading a 100-person technical track (comprising security engineers from frontier AI labs, government security specialists, and datacenter colocation providers) alongside an executive track of AI industry security leaders providing steering input.

Over the past nine months, we have conducted a series of workshops and research programs to clarify what it takes to reach Security Level 5 in a way that is sensitive to competitive pressures, the need to maintain speed of innovation, and the reality of a rapidly shifting threat landscape. Our mission is to create the optionality for frontier AI labs to reach Security Level 5 in the coming years, and to be able to activate that security level within six months of choosing to do so.

In service of that mission, we have convened this broad task force to clarify what needs to be done, and in particular what needs to be done early, to preserve that optionality. This standard represents one output of that collaborative effort.

For more information or to engage with our work, contact us at standard@sl5.org or visit sl5.org.

About This Document

Security Level 5 (SL5) is a security posture for AI systems that could plausibly thwart top-priority operations by the world's most cyber-capable institutions: those with extensive resources, state-level infrastructure, and expertise years ahead of the public state of the art. The SL5 terminology originates from the RAND Corporation's 2024 report "Securing AI Model Weights" [1].

This first revision of the SL5 standard focuses on requirements with long lead times: interventions that must be planned years in advance, such as facility construction, hardware procurement, and organizational capability development. We prioritize these requirements because preserving optionality for SL5 by 2028/2029 requires starting now. These capabilities cannot be retrofitted on short notice when the need becomes urgent. Some requirements represent significant departures from current day standard practice. We believe bold measures are necessary for this level of security and see clear opportunities to apply optimization pressure to existing and novel solutions to customize them for the AI industry and address the practical operational requirements as much as possible. Our organization exists to begin paving this path. Some requirements approximate government security capabilities where private-sector approaches may be insufficient. We identify these gaps and note where government involvement may ultimately be necessary.

This standard was developed collaboratively with frontier AI laboratories, government partners, and security experts through sustained engagement over several months. As version 0.1, significant refinement is expected through continued stakeholder engagement. We explicitly invite frontier AI labs, government agencies, datacenter operators, and security researchers to engage with this work, whether through direct collaboration, feedback, or implementation experience. Please reach out through standard@sl5.org.

The control specifications in this standard are structured as an overlay on NIST SP 800-53. We chose this approach for three reasons. First, NIST SP 800-53 is a battle-tested framework and the standard choice for high-security organizations. Second, structuring as an overlay enables ease of adoption for organizations already implementing NIST controls. Third, the overlay format clearly expresses the "diff" from existing baselines, highlighting what is new or different for SL5 rather than restating established requirements.

Many other security controls are necessary for SL5, including most controls from existing high-security baselines. Future revisions will provide detailed mapping from DoD Impact Level 6 (IL6) and its reference frameworks (FedRAMP High, CNSSI 1253) to SL5 requirements [4], [5]. Physical security requirements draw on ICD 705 SCIF standards as a basis [6], [7], [22], [23]. Hardware supply chain requirements reference NIST SP 800-161 Rev 1 [3].

Acknowledgments

This standard was shaped by the collective expertise of the SL5 Task Force's two specialized tracks. Our Executive Track, comprising AI lab decision-makers, national security leaders, datacenter operators, chip providers, and government representatives, provided strategic direction and identified key obstacles to SL5 implementation. Our Technical Track, comprising security researchers, lab security engineering staff, technical AI researchers, hardware engineers, and security engineers working across five specialized working groups (machine security, network security, software security, supply chain security, and personnel security), developed and stress-tested the control specifications and architectural recommendations in this document.

We are deeply grateful to every participant across both tracks for their willingness to engage in rigorous, honest discussion and for the time they generously contributed to this effort. Please note that the following list of acknowledgments is not yet finalized and will be updated in the coming days. Additionally, inclusion in this list reflects an individual's informative contributions to this effort and does not necessarily represent their personal views or endorsement of all the contents of this document. The following individuals wished to be acknowledged for their contributions:

- Jason Clinton, Deputy CISO, Anthropic
- Dane Stuckey, CISO, OpenAI
- Vijay Bolina, former CISO, GDM
- Rob Joyce, former Cybersecurity Director, NSA
- Phil Venables, former CISO, Google Cloud, current Partner at Ballistic Ventures
- Jason Kichen, CISO, Fluidstack
- Eric Grosse, former VP Security & Privacy Engineering, Google
- Dominic Rizzo, CEO ZeroRISC
- Omer Nevo, CTO, Irregular
- Evan Miyazono, CEO, Atlas Computing
- Kristian Rönn, CEO, Lucid Computing
- Jarrah Bloomfield, Anthropic & former Google
- Keri Warr, Anthropic
- Steven Hernandez, former CISO, United States Agency for International Development
- Tanya Verma, CEO, Tinfoil
- Gil Gekker, CoS to CEO, Irregular
- davidad (David A. Dalrymple)
- Paul Crowley (ciphergoth), Anthropic
- Jason Kikta, CTO, Automox

This list will be in progress for the next few weeks, as we hear from members.

Changelog

Release 0.1.2 — June 23, 2026

Requires post-quantum cryptography for inter-facility network encryptors, replacing the FIPS 140-3 Level 3 module validation requirement. The SL5 Task Force now mandates FIPS 203 and FIPS 204 (ML-KEM and ML-DSA) or NSA CNSA 2.0 compliant algorithms. Also clarifies Fig. 1 (SL5 Network Architecture) by removing the FIPS 140-3 L3 specification labels from the encryptors on the dark fiber path. Also updates the Security Architecture overview and resolves the open question on cryptographic sufficiency â€" the SC-8(1) and SC-13 mandates settle the NSA Type 1 vs post-quantum question, so it has been removed from Open Questions.

SC-8(1): Cryptographic Protection (Revision)

Replaced FIPS 140-3 Level 3 module validation requirement with post-quantum cryptography: inline network encryptors must implement FIPS 203 and FIPS 204 (ML-KEM and ML-DSA) or NSA CNSA 2.0 compliant algorithms. Updated Fig. 1 (SL5 Network Architecture) to remove the "FIPS 140-3 L3" specification labels from the encryptors.

SC-13: Cryptographic Protection (Revision)

Added supplemental guidance requiring post-quantum cryptographic algorithms for inter-facility encryptors, in line with SC-8(1). NSA Type 1 certified encryptors remain the higher-tier option.

Security Architecture: Cryptographic Protection (Revision)

Updated the Security Architecture overview to replace 'FIPS 140-3 Level 3 minimum validation for network encryptors' with 'Post-quantum cryptographic algorithms (FIPS 203 and FIPS 204, or NSA CNSA 2.0) for network encryptors', aligning the narrative section with the SC-8(1) and SC-13 control revisions.

Open Questions: Cryptographic Protection (Revision)

Resolved and removed the open question on cryptographic sufficiency. The SC-8(1) and SC-13 revisions mandating FIPS 203/204 (ML-KEM/ML-DSA) or NSA CNSA 2.0 settle the question of whether NSA Type 1 or post-quantum algorithms are required. The item has been removed from Open Questions.

Release 0.1.1 — June 4, 2026

Revises two controls in response to external review. PS-3 removes the "Private SF-86" construct and reframes high-tier personnel vetting as an active area of research with two paths: a formal government partnership or the industry-adapted Sensitivity Levels (SenL) Framework. PE-19(1) extends emanations security to adversary-controlled active signals and defines energy-flow policies at Red Zone boundaries.

PS-3: Personnel Screening (Revision)

Removed the "Private SF-86" construct, which assumed a private vetting process equivalent to a government background investigation that does not yet exist. PS-3 now frames high-tier personnel

vetting as an active area of research with two paths: (1) a formal government partnership using existing government clearance authorities, or (2) the Sensitivity Levels (SenL) Framework, an industry-adapted clearance model that labs can deploy without government participation, though it would benefit from government information-sharing.

PE-19(1): National Emissions Policies and Procedures (Revision)

Extended emanations security from passive egress prevention to adversary-controlled active signals, covering inbound signal injection used to influence system behavior. Emissions policies now define the permitted energy flows across each Red Zone boundary in both directions.

1. Introduction

1.1 Threat Model

The primary adversaries for SL5 systems are the top-priority operations of the world's most cyber-capable institutions—operations comparable to 1,000 individuals with expertise years ahead of the public state of the art, spending years with budgets up to $1 billion, backed by state-level infrastructure and access developed over decades. A **primary use case for SL5 is an AI lab approaching fully automated AI R&D**. OpenAI has already announced they expect to have automated researchers by 2028 [29]. Anthropic's CEO has similarly noted that AI "may be only 1–2 years away from a point where the current generation of AI autonomously builds the next" [30]. At this stage, the economic value of frontier AI models derives increasingly from conducting internal automated R&D and accelerating scientific advances, more so than from serving the model to customers. Models capable of automating AI research could enable rapid recursive improvement in AI capabilities, with huge economic potential and significant implications for geopolitics. Nation-state adversaries have strong incentives to acquire such capabilities or prevent rivals from doing so.

The targets of SL5 protection are critical assets held by the frontier AI labs. These critical assets include “covered models”—frontier AI models passing capability thresholds designated by the organization as requiring SL5 protection, AI research and software that could enable adversaries to develop comparable capabilities, inputs (which could be used to poison or backdoor) and outputs (which could be used to distill or reverse engineer). The security objectives are the **confidentiality, integrity, and availability** of these critical assets.

The standard also addresses risks from misaligned or compromised AI models that may attempt to sabotage research or exfiltrate themselves, which constitute a distinct form of insider threat. Mitigations for this threat class substantially overlap with nation-state defenses; this revision does not include mitigations specific to misaligned AI, though future revisions will likely address this threat class more directly.

Attack vectors include insider threats, supply chain compromise, physical intrusion, network exploitation, side-channel attacks, and adversarial inputs designed to compromise AI systems.

1.2 Security Architecture Overview

This section summarizes the SL5 security architecture across five security streams: network, physical, machine, personnel, and supply chain.

Network Security

The SL5 Network is air-gapped from external networks and supports model development, training, and internal deployment operations. The SL5 Network may span multiple geographically distributed facilities, including remote office locations (such as in San Francisco) and other datacenters connected via encrypted inter-facility links. All facilities must comply with all SL5 requirements including applicable ICD 705 guidance for physical security [6], [7], [22], [23].

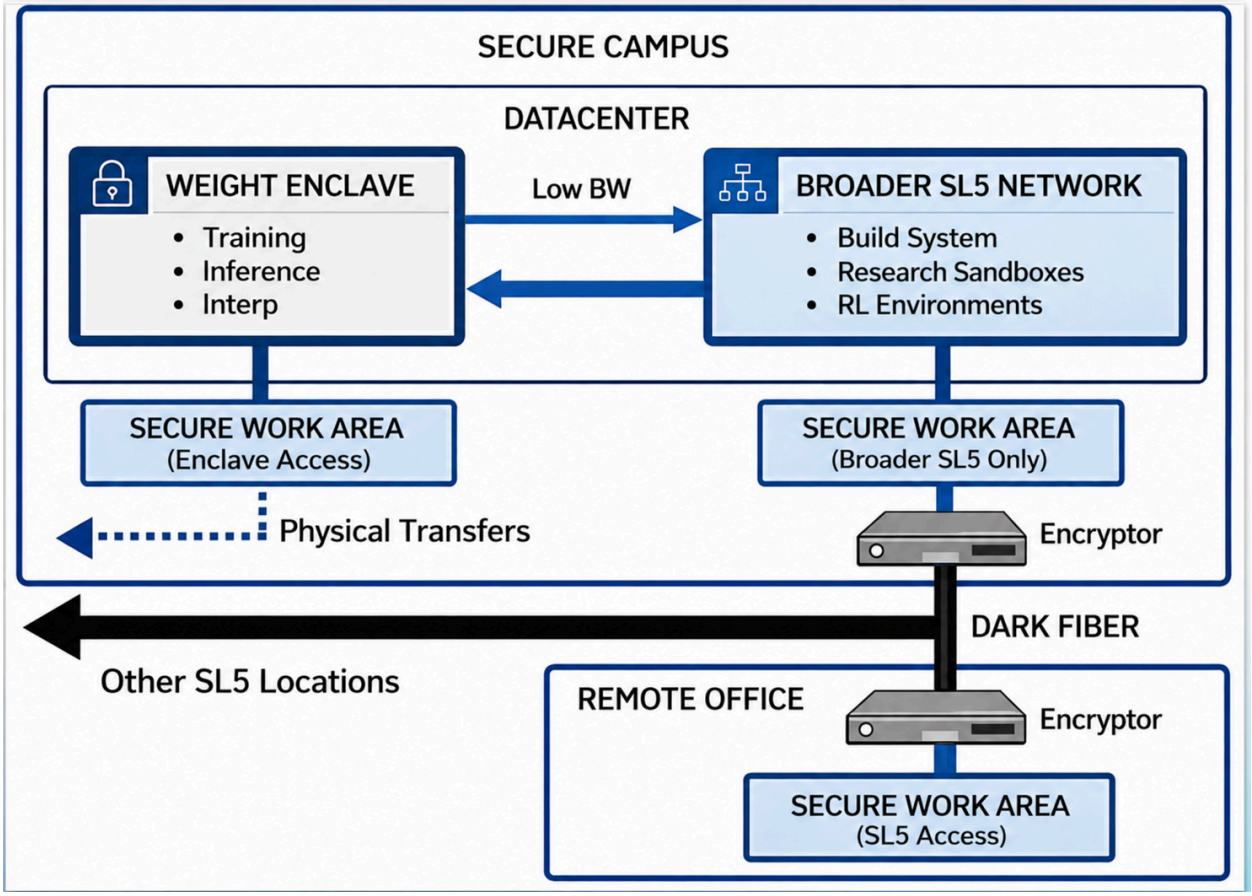

Fig. 1. *SL5 Network Architecture showing the relationship between Weight Enclaves, the broader SL5 Network, and inter-facility connections.*

Weight Enclaves provide additional isolation within the SL5 Network for systems with direct access to covered models. This separation enables stricter controls on the most sensitive assets—covered model weights—while supporting R&D operations and remote work locations in the broader SL5 Network. Systems requiring direct weight access (training, inference, fine-tuning, mechanistic interpretability) operate within Weight Enclaves under strict allow-by-exception code execution; all code must complete testing and signing before deployment. Research experiments conducted by AI models are executed in confined environments in the broader SL5 Network (see Figure 1), outside the Weight Enclaves. Each Weight Enclave resides in a single physical facility. Transfers exceeding the Weight Enclave outbound bandwidth limit must occur via encrypted physical media.

This has significant implications for geographically distributed training: such training may prove infeasible under this constraint, or may require substantially different approaches than current practice. One speculative example is distributed RL training with daily gradient synchronization via physical media transfer. Section 1.3 highlights this as a key open question.

Key interventions:

- Air-gapped SL5 Network with no external network connections
- Weight Enclaves isolated within SL5 Network, containing only minimal necessary software with strict allow-by-exception code execution
- Single-facility restriction for Weight Enclaves; inter-enclave transfers exceeding bandwidth limits require encrypted physical media
- Dual inline network encryptors from different suppliers (per NSA "Rule of Two") for inter-facility SL5 Network connections [14]
- FIPS 140-3 Level 3 minimum validation for network encryptors [11]
- Physical bandwidth limitation on Weight Enclave boundaries preventing weight exfiltration even if other security measures fail

See Section 3.9 for detailed control specifications.

Physical Security

SL5 facilities are constructed to an applicable subset of ICD 705 directive providing physical access control, emanations protection, and transmission security [6], [7], [22], [23]. The particular subset is determined by whether the threat model being prioritized is theft of model weights & cryptographic keys vs. also sabotage, autonomy threats and algorithmic IP theft. SL5-level security may generally warrant an updated version of ICD 705 plus potential additional AI-datacenter and AI-SCIF specific overlays. Following SCIF terminology, Red Zones are areas where SL5-protected information may be processed or stored; Black Zones are areas with no network path to SL5-protected information. All SL5 Network hardware and access locations reside within Red Zones.

Key interventions:

- ICD 705 facility construction with applicable TEMPEST countermeasures (RF shielding, power conditioning) [6], [7], [9], [23]
- Access control vestibules (mantraps) at all entry points to prevent tailgating
- Protected Distribution Systems (PDS) per CNSSI 7003 for nearby connections [8]
- Shielded rack enclosures meeting NSA 94-106 specifications for Weight Enclave systems processing covered models [10]

See Section 2 for detailed facility requirements and Section 3.4 for related control specifications.

Machine Security

AI accelerators require hardware security features that enable protection of covered models even when the host system is compromised, when data traversing physical interconnects is intercepted, or when the hardware itself is physically attacked. Only authorized, cryptographically signed code executes on Weight Enclave accelerators (see SI-7(15)), reducing the attack surface. These capabilities enable end-to-end encrypted data paths where data arrives encrypted from origin, is decrypted only within the accelerator, and is re-encrypted before any host-accessible export.

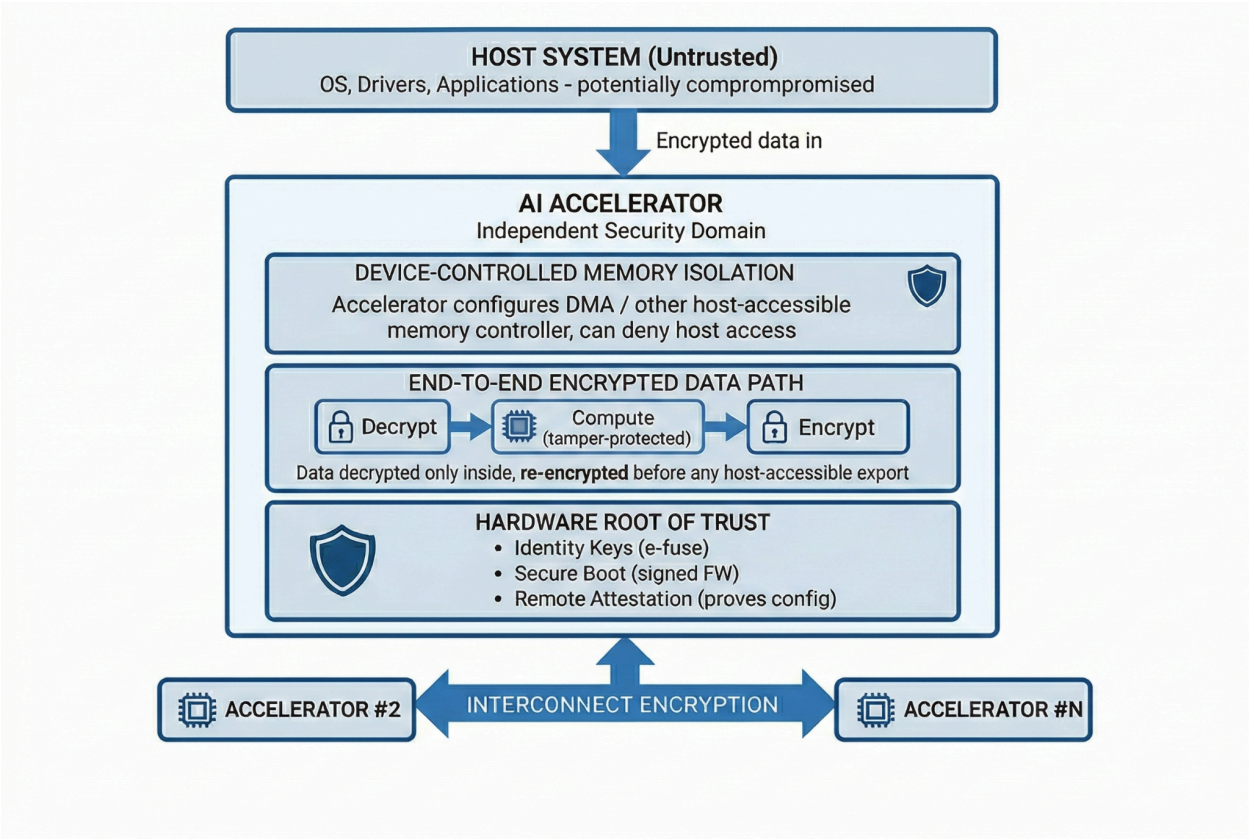

Fig. 2. *AI Accelerator Security Architecture showing the accelerator as an independent security domain with device-controlled memory isolation and end-to-end encrypted data paths.*

Key interventions:

- Integrated root-of-trust with hardware-provisioned identity, secure key storage, and attestation capability
- Device-controlled memory isolation allowing accelerators to deny host access independent of host software
- Interconnect encryption protecting data transmitted between accelerators during distributed operation
- Tamper protection for compute cores and accelerator memory where data exists unencrypted during computation
- Execution integrity verification ensuring only authorized code executes on accelerators

Current commercial AI accelerators vary in confidential computing capabilities, and gaps remain relative to SL5 requirements. For example, NVIDIA Blackwell's memory isolation and code verification depend on a CPU TEE rather than being accelerator-controlled, and commercial threat models exclude physical attacks [16], [24].

Current implementations also lack mechanisms preventing composition attacks, where individually-authorized operations are combined to exfiltrate data. Two approaches show promise. The Ascend-CC research architecture (2024) demonstrates confidential computing that excludes the CPU from the trusted computing base, using device-controlled memory isolation and firmware-level task sequence attestation [17]. Alternatively, restricting workloads to fused kernels verified to not exfiltrate data individually or through composition avoids the need for vendor engagement but may increase implementation complexity for organizations.

Organizations must engage accelerator vendors during architecture phases—hardware security features require years to develop, and influence over roadmaps cannot be gained retroactively.

See Sections 3.3 and 3.10 for related control specifications.

Personnel Security

A five-tier sensitivity level framework (SenL-1 through SenL-5) addresses insider threat through graduated vetting, monitoring, and access controls. This represents the strongest feasible private-sector approximation of government clearance programs. Detailed tier definitions, vetting procedures, and operational safeguards are specified in a separate SenL Framework Document [26].

Key interventions:

- Five sensitivity levels based on access to covered models and critical infrastructure [26]
- Vetting depth scales from baseline checks (SenL-1) through verified subject and reference interviews (SenL-4/5) [26]
- Monitoring intensity scales from periodic checks to intensive monitoring of compartmented system interactions [26]
- Access obligations scale from standard NDA to custodian-specific protocols and post-employment restrictions [26]
- Dual authorization for critical operations involving covered models [26]
- SenL-5 clearance required for unescorted Red Zone access, with two-person integrity [26]

Organizations may need to pursue government involvement (through classified contract pathways, information sharing arrangements, or new statutory authority) to achieve personnel security sufficient for the stated threat model. Section 1.3 highlights this as a key open question.

See Sections 3.5 and 3.6 for detailed control specifications.

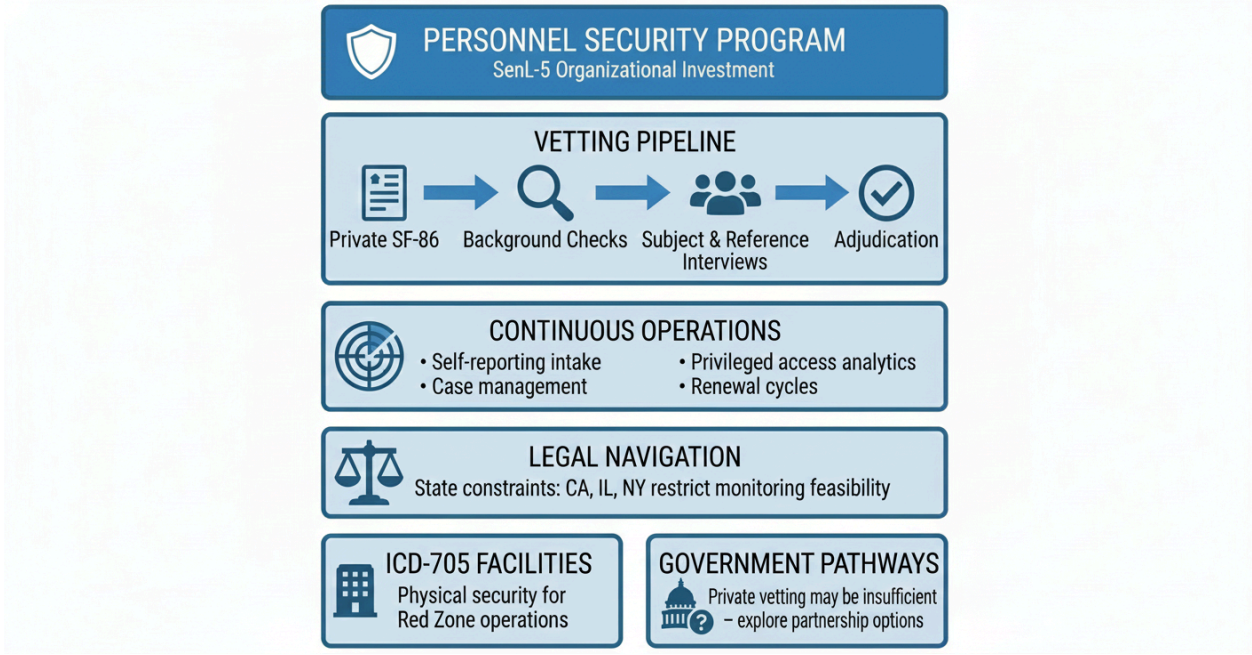

Fig. 3. *Personnel Security Program showing the organizational investment required for SenL-5 capability: vetting pipeline, continuous operations, legal navigation, ICD 705 facilities, and potential government pathways [26].*

Supply Chain Security

Hardware supply chain integrity requirements address risks from compromised components. Hardware supply chain requirements follow NIST SP 800-161 Rev 1 guidance directly [3].

Data entering the SL5 environment from external sources must be screened for adversarial content—poisoning attacks, jailbreak attempts, and adversarial examples that could compromise AI systems or sabotage research. External data is a critical input to AI system development, analogous to hardware components; both can carry embedded attacks from upstream sources, and correlated attacks targeting both software and training data could amplify impact.

Robust adversarial content detection remains an open research problem; best-available defensive measures are insufficient against sophisticated adversaries [18], [19]. This standard mandates staging isolation and investment in detection research, understanding that breakthroughs in adversarial robustness are necessary to fully address this threat. Section 1.3 highlights this as a key open question.

Key interventions:

- Adversarial content screening for all external data, with staging isolation until data completes automated detection and human review
- Supplier governance: criticality-based inventory, continuous assessment against security baseline, qualified bidders/manufacturers lists, supply base diversity where feasible, requirements flow-down to sub-tier contractors
- Component integrity assurance: comprehensive testing including counterfeit detection, physical inspection, tamper detection, developer screening

See Sections 3.1 and 3.8 for detailed control specifications.

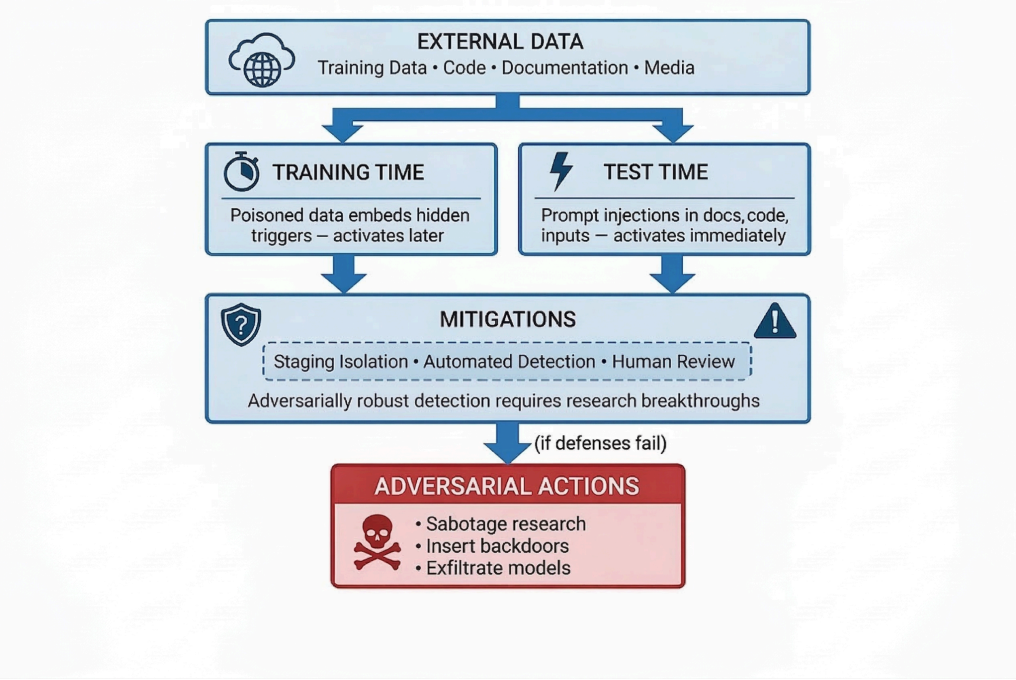

Fig. 4. *Adversarial Robustness threat model showing how external data can carry attacks that activate at training time or test time, the mitigations required (staging isolation, automated detection, human review), and the consequence if defenses fail.*

1.3 Open Questions

This is a first draft and we value transparency about areas of genuine uncertainty, whether due to conflicting expert perspectives or limited access to relevant information. Our goal is to achieve both strong security and operational effectiveness. We seek input to resolve these uncertainties directly—evidence that changes our assessment may change the requirements—as well as creative approaches that accomplish both. The following questions have significant architectural or policy implications:

- Personnel vetting limitations: Whether private-sector vetting is sufficient for SL5, or whether government involvement (through classified contract pathways, information sharing, or new legal authority) is required to achieve adequate personnel security. We welcome input from organizations with experience in either approach.
- Adversarial detection feasibility: Whether robust detection of adversarial content is achievable against sophisticated adversaries, given that this remains an active research problem. This standard mandates staging isolation and investment in detection research, but it is uncertain whether breakthroughs sufficient to address the threat will materialize. We welcome detection approaches or architectural mitigations.
- Inter-enclave network security: Whether long-distance network connections can be adequately secured against nation-state adversaries, even with multiple validated inline encryptors in series. If such connections cannot be secured, the physical media requirement for high-bandwidth transfers would significantly constrain or preclude geographically distributed training. We seek both security analysis on network connection viability and, if the constraint stands, creative approaches to distributed training under physical media requirements.

Additional open questions are documented in Appendix A.

Additional Open Questions (Appendix A)

Organized by security stream:

Machine Security

- Availability of accelerators meeting SL5 hardware security requirements in time for deployment
- Feasibility of workload integrity mechanisms (e.g., task sequence attestation, kernel fusion) for expected workloads such as training, inference, fine-tuning, and mechanistic interpretability

Physical Security

- TEMPEST zone classification for Red Zones
- Shared infrastructure isolation (power, cooling, fire suppression) between distributed Red Zones
- Required attenuation levels for shielded racks (full NSA 94-106 vs reduced spec)
- Level of physical isolation required for Weight Enclaves (rack, room, or building level separation)
- ICD 705 private accreditation: This standard specifies self-certification by CSO or Site Security Manager. Whether private self-certification provides meaningful security assurance compared to government accreditation, and what this implies for actual security posture

Network Security

- Specific bandwidth thresholds for exfiltration prevention (calibrated to model size and threat model)
- Redundancy architectures for fail-closed encryptors

Personnel Security

- Screening equivalency standards for non-US citizens and international offices
- Monitoring feasibility in high-restriction states (California CPRA/ICRAA, Illinois BIPA)

Supply Chain Security

- Testing/certification approaches for adversarial content detection system robustness
- Processing state tracking for new modality onboarding

Additional Open Questions (Appendix A)

Organized by security stream:

Machine Security

- Availability of accelerators meeting SL5 hardware security requirements in time for deployment
- Feasibility of workload integrity mechanisms (e.g., task sequence attestation, kernel fusion) for expected workloads such as training, inference, fine-tuning, and mechanistic interpretability

Cryptographic Protection

- NSA Type 1 vs post-quantum algorithms (FIPS 203/204/205 or CNSA 2.0) sufficiency for the stated threat model

Physical Security

- TEMPEST zone classification for Red Zones
- Shared infrastructure isolation (power, cooling, fire suppression) between distributed Red Zones
- Required attenuation levels for shielded racks (full NSA 94-106 vs reduced spec)
- Level of physical isolation required for Weight Enclaves (rack, room, or building level separation)
- ICD 705 private accreditation: This standard specifies self-certification by CSO or Site Security Manager. Whether private self-certification provides meaningful security assurance compared to government accreditation, and what this implies for actual security posture

Network Security

- Specific bandwidth thresholds for exfiltration prevention (calibrated to model size and threat model)
- Redundancy architectures for fail-closed encryptors

Personnel Security

- Screening equivalency standards for non-US citizens and international offices
- Monitoring feasibility in high-restriction states (California CPRA/ICRAA, Illinois BIPA)

Supply Chain Security

- Testing/certification approaches for adversarial content detection system robustness
- Processing state tracking for new modality onboarding

2. ICD 705 Facility Requirements

Physical security forms the foundation for all other SL5 protections [6], [7], [22], [23]. This section specifies facility construction requirements that have no direct NIST SP 800-53 equivalent. These requirements draw on an applicable subset of ICD 705 and ICS 705-1 (Physical and Technical Standards for Sensitive Compartmented Information Facilities), published by the Office of the Director of National Intelligence [6], [7], [22], [23]. The particular subset is determined by whether the threat model being prioritized is theft of model weights and cryptographic keys, or also extends to sabotage, autonomy threats, and algorithmic IP theft. SL5-level security may generally warrant an updated version of ICD 705 plus potential additional AI-datacenter and AI-SCIF specific overlays. ICD 705 provides detailed specifications for perimeter construction, intrusion detection, access control, acoustic protection, and TEMPEST countermeasures; this section summarizes key requirements and SL5-specific adaptations.

Adaptation for SL5:

- "SL5-protected information" (covered models and secrets) replaces "Classified Information" [26]
- Chief Security Officer (CSO) or Site Security Manager serves as accreditation authority
- SenL-5 clearance replaces government security clearances [26]

Security in Depth

Facility design implements Security in Depth (SID)—layered security controls that increase the probability of detecting unauthorized access attempts before reaching the Red Zone perimeter. Layers may include perimeter fencing, building access controls, and controlled areas surrounding Red Zones. SID considerations should inform site selection and building layout.

Zone Architecture

Red Zone: The SL5 Network operates within Red Zones—physically hardened environments constructed to ICD 705/ICS 705-1 standards. Any system with network access to SL5-protected information must be within a Red Zone. Red Zones may span multiple geographically distributed facilities.

Black Zone: Areas with no network path to SL5-protected information. Black Zones may exist within the same facility as Red Zones but are completely network-isolated.

Construction

Red Zones form complete physical enclosures per ICS 705-1—walls, floor, and ceiling create a continuous barrier with no gaps. Acoustic attenuation (Sound Group 3 or 4) prevents eavesdropping through the perimeter [7], [23]. Windows are prohibited; where required by safety

codes, they must be non-opening with RF/optical shielding. All utility penetrations (HVAC, power, fire suppression) must maintain perimeter integrity.

TEMPEST

Red Zones require TEMPEST countermeasures per NSTISSAM TEMPEST/1-92, pre-engineered into construction to the maximum extent practicable [9], [21]. RF shielding prevents electromagnetic signal egress from the perimeter, dedicated power conditioning prevents power-line analysis, and all equipment must be hardwired—no wireless devices permitted.

Intrusion Detection

Red Zones require intrusion detection systems (IDS) per ICD 705 Chapter 7, with sensors covering all perimeter entry points and motion detection within the space. Alarm response time requirements depend on storage type: 5 minutes for open storage (Sensitive compartmented information (SCI) material accessible outside containers), 15 minutes for closed storage (SCI material secured in General Services Administration (GSA)-approved containers when unoccupied) [6], [7], [23].

Transmission Security

All Weight Enclave traffic leaving the Red Zone perimeter requires Protected Distribution Systems (PDS) per CNSSI 7003—including encrypted inter-building connections [8]. PDS provides physical protection through hardened conduit and alarmed carriers.

3. Control Specifications

This section is a NIST SP 800-53 overlay—a set of supplemental guidance and parameter values that tailor existing controls without replacing base requirements. It is a partial overlay, covering only the long lead time interventions highlighted in Section 1. The complete SL5 overlay will build on IL6, which itself incorporates FedRAMP High, CNSSI 1253, and other frameworks; future revisions will explicitly map SL5 requirements to IL6.

Organized by NIST control family. "NIST Control Text" and "NIST Discussion" are taken verbatim from NIST SP 800-53 Rev 5. "SL5 Supplemental Guidance" and "Parameter Values" are additions specific to this standard. Where a Parameter Values section is absent or a specific assignment is not provided, organizations define values based on their specific context.

3.1 Access Control (AC)

AC-3(2): Dual Authorization

NIST Control Text:

Enforce dual authorization for [Assignment: organization-defined privileged commands and/or other actions].

AC-4: Information Flow Enforcement

NIST Control Text:

Enforce approved authorizations for controlling the flow of information within the system and between connected systems based on [Assignment: organization-defined information flow control policies].

SL5 Supplemental Guidance:

External data may contain adversarial content including backdoor attacks in training data, jailbreak attempts in inference inputs, and malicious patterns in code (including inputs designed to compromise LLM-based code review or hardening systems). As models handle critical security engineering tasks and sensitive AI research, successful attacks could lead to sabotaged research, exfiltration of covered models, or compromised security outputs. Correlated supply chain attacks targeting both software and AI training data could amplify impact.

The organization enforces mandatory staging isolation: all external data is received into staging areas physically separated from internal systems. Transfer mechanisms enforce screening requirements, preventing bypass of the isolation boundary. Data remains isolated until automated screening validates it. This architecture prevents unscreened or malicious data from reaching internal systems.

AC-4(9): Human Reviews

NIST Control Text:

Enforce the use of human reviews for [Assignment: organization-defined information flows] under the following conditions: [Assignment: organization-defined conditions].

SL5 Supplemental Guidance:

The organization requires human review of quarantined data before making clearance or rejection decisions. The organization defines review scope based on operational capacity and false positive rates: either all quarantined data or a sample sufficient to validate detection effectiveness.

Human reviews provide oversight of automated detection, enabling identification of false positives, confirmation of true detections, and discovery of attack patterns not yet captured by automated mechanisms. Higher-risk data receives more intensive review.

AC-4(15): Detection of Unsanctioned Information

NIST Control Text:

When transferring information between different security domains, examine the information for the presence of [Assignment: organization-defined unsanctioned information] and prohibit the transfer of such information in accordance with the [Assignment: organization-defined organization-defined security or privacy policy].

SL5 Supplemental Guidance:

Adversarial content designed to compromise AI models or data center operations constitutes unsanctioned information. Detection applies to all data that could be consumed as input by AI models: training datasets, evaluation data, prompts, code, configuration files, images, audio, video, and structured data. All supported modalities (text, images, video, audio, code, structured data) must be screened regardless of intended use.

Organizations determine detection thresholds through risk assessment (per RA-3), balancing false positives against false negatives. Risk-based tiering applies stricter thresholds to higher-risk data (e.g., training data, infrastructure code).

Detection systems must resist adversarial bypass attacks. Robust adversarial content detection remains an open research problem; best-available defensive measures are insufficient against sophisticated adversaries. Organizations invest in research and development to advance detection capabilities, understanding that breakthroughs in adversarial robustness are necessary to fully address this threat. Section 1.3 highlights this as a key open question.

CM-7(5): Least Functionality | Authorized Software — Allow-by-exception

NIST Control Text:

a. Identify [Assignment: organization-defined software programs];

b. Employ a deny-all, permit-by-exception policy to allow the execution of authorized software programs on the system; and

c. Review and update the list of authorized software programs [Assignment: organization-defined frequency].

IA-3: Device Identification and Authentication

NIST Control Text:

Uniquely identify and authenticate [Assignment: organization-defined devices and/or types of devices] before establishing a [Selection (one or more): local; remote; network] connection.

SL5 Supplemental Guidance:

AI accelerators authenticate using cryptographic mechanisms anchored in a hardware root of trust. For distributed operation, accelerators authenticate each other before exchanging data, without host-mediated trust.

Accelerators support remote attestation: cryptographically proving their identity and configuration state to remote parties. Attestation enables data providers to verify they are communicating with a legitimate accelerator running authorized firmware before sending sensitive data. Boot measurements collected during secure boot are signed using hardware-protected keys, creating a cryptographic proof that can be verified without trusting the host.

PE-2(3): Restrict Unescorted Access

NIST Control Text:

Restrict unescorted access to the facility where the system resides to personnel with [Selection (one or more): security clearances for all information contained within the system; formal access authorizations for all information contained within the system; need for access to all information contained within the system; [Assignment: organization-defined physical access authorizations]].

SL5 Supplemental Guidance:

Only personnel with SenL-5 (Custodial) clearance may enter Red Zones unescorted, and two authorized individuals must be present at all times (two-person integrity). Anyone without SenL-5—including SenL-4 holders and vendors—requires continuous escort by a SenL-5 holder, with all activities logged [26].

PE-3(8): Access Control Vestibules

NIST Control Text:

Employ access control vestibules at [Assignment: organization-defined locations].

SL5 Supplemental Guidance:

All Red Zone entry points use access control vestibules (mantraps) with interlocking doors, multi-factor authentication, and anti-tailgating sensors. Mantraps prevent tailgating—following an authorized person through a door—and credential sharing.

PE-19(1): National Emissions Policies and Procedures

NIST Control Text:

Protect system components, associated data communications, and networks in accordance with national Emissions Security policies and procedures based on the security category or classification of the information.

PM-12: Insider Threat Program

NIST Control Text:

Implement an insider threat program that includes a cross-discipline insider threat incident handling team.

PS-2: Position Risk Designation

NIST Control Text:

a. Assign a risk designation to all organizational positions;

b. Establish screening criteria for individuals filling those positions; and

c. Review and update position risk designations [Assignment: organization-defined frequency].

SL5 Supplemental Guidance:

The organization assigns one of five Sensitivity Levels (SenL-1 through SenL-5) to every position based on access to covered models and security-critical infrastructure. Tiers range from baseline corporate access (SenL-1) through compartmented weight custodian roles (SenL-5). The SenL Framework Document specifies detailed tier definitions, access criteria, and edge case handling [26].

PS-3: Personnel Screening

NIST Control Text:

a. Screen individuals prior to authorizing access to the system; and

b. Rescreen individuals in accordance with [Assignment: organization-defined organization-defined conditions requiring rescreening and, where rescreening is so indicated, the frequency of rescreening].

SL5 Supplemental Guidance:

Personnel with unescorted access to Weight Enclaves must be vetted in proportion to that access. Vetting at the highest tiers is an active area of research, and the SL5 Task Force is developing two paths to meet it [26].

The first is a formal government partnership that relies on existing government investigation and adjudication authorities. This provides the strongest assurance but requires government participation and may depend on new legal authority.

The second is the Sensitivity Levels (SenL) Framework, an industry-adapted clearance model that labs can deploy without government participation, though it would benefit greatly from government information-sharing. SenL classifies personnel into tiers (SenL-1 through SenL-5) by the sensitivity of the access they hold.

Provisional access during vetting requires compensating controls as specified in the SenL Framework Document [26].

PS-6: Access Agreements

NIST Control Text:

a. Develop and document access agreements for organizational systems;

b. Review and update the access agreements [Assignment: organization-defined frequency] ; and

c. Verify that individuals requiring access to organizational information and systems:

1. Sign appropriate access agreements prior to being granted access; and

2. Re-sign access agreements to maintain access to organizational systems when access agreements have been updated or [Assignment: organization-defined frequency].

SL5 Supplemental Guidance:

Access agreements scale with Sensitivity Level. SenL-1/2 include standard NDA and monitoring acknowledgment. Higher tiers add progressively stricter obligations: foreign contact reporting and secondary employment restrictions (SenL-3), dual authorization acknowledgment and travel notification (SenL-4), custodian-specific protocols and post-employment restrictions (SenL-5). Complete tier-specific requirements are in the SenL Framework Document [26].

SA-4: Acquisition Process

NIST Control Text:

Include the following requirements, descriptions, and criteria, explicitly or by reference, using [Selection (one or more): standardized contract language; [Assignment: organization-defined contract language]] in the acquisition contract for the system, system component, or system service:

a. Security and privacy functional requirements;

b. Strength of mechanism requirements;

c. Security and privacy assurance requirements;

d. Controls needed to satisfy the security and privacy requirements.

e. Security and privacy documentation requirements;

f. Requirements for protecting security and privacy documentation;

g. Description of the system development environment and environment in which the system is intended to operate;

h. Allocation of responsibility or identification of parties responsible for information security, privacy, and supply chain risk management; and

i. Acceptance criteria.

SA-11: Developer Testing and Evaluation

NIST Control Text:

Require the developer of the system, system component, or system service, at all post-design stages of the system development life cycle, to:

a. Develop and implement a plan for ongoing security and privacy control assessments;

b. Perform [Selection (one or more): unit; integration; system; regression] testing/evaluation [Assignment: organization-defined frequency to conduct] at [Assignment: organization-defined depth and coverage];

c. Produce evidence of the execution of the assessment plan and the results of the testing and evaluation;

d. Implement a verifiable flaw remediation process; and

e. Correct flaws identified during testing and evaluation.

SA-17: Developer Security and Privacy Architecture and Design

NIST Control Text:

Require the developer of the system, system component, or system service to produce a design specification and security and privacy architecture that:

a. Is consistent with the organization's security and privacy architecture that is an integral part the organization's enterprise architecture;

b. Accurately and completely describes the required security and privacy functionality, and the allocation of controls among physical and logical components; and

c. Expresses how individual security and privacy functions, mechanisms, and services work together to provide required security and privacy capabilities and a unified approach to protection.

SA-20: Customized Development of Critical Components

NIST Control Text:

Reimplement or custom develop the following critical system components: [Assignment: organization-defined critical system].

SA-21: Developer Screening

NIST Control Text:

Require that the developer of [Assignment: organization-defined system, systems component, or system service]:

a. Has appropriate access authorizations as determined by assigned [Assignment: organization-defined official government duties] ; and

b. Satisfies the following additional personnel screening criteria: [Assignment: organization-defined additional personnel screening criteria].

SR-3(1): Diverse Supply Base

NIST Control Text:

Employ a diverse set of sources for the following system components and services: [Assignment: organization-defined organization-defined system components and services].

SR-3(3): Sub-Tier Flow Down

NIST Control Text:

Ensure that the controls included in the contracts of prime contractors are also included in the contracts of subcontractors.

SR-5(2): Assessments Prior to Selection, Acceptance, Modification, or Update

NIST Control Text:

Assess the system, system component, or system service prior to selection, acceptance, modification, or update.

SR-6: Supplier Assessments and Reviews

NIST Control Text:

Assess and review the supply chain-related risks associated with suppliers or contractors and the system, system component, or system service they provide [Assignment: organization-defined frequency].

SR-9: Tamper Resistance and Detection

NIST Control Text:

Implement a tamper protection program for the system, system component, or system service.

SR-9(1): Multiple Stages of System Development Life Cycle

NIST Control Text:

Employ anti-tamper technologies, tools, and techniques throughout the system development life cycle.

SR-10: Inspection of Systems or Components

NIST Control Text:

Inspect the following systems or system components [Selection (one or more): at random; at [Assignment: organization-defined frequency] ; upon [Assignment: organization-defined indications of need for inspection]] to detect tampering: [Assignment: organization-defined systems or system components].

SR-11: Component Authenticity

NIST Control Text:

a. Develop and implement anti-counterfeit policy and procedures that include the means to detect and prevent counterfeit components from entering the system; and

b. Report counterfeit system components to [Selection (one or more): source of counterfeit component; [Assignment: organization-defined external reporting organizations] ; [Assignment: organization-defined personnel or roles]].

SR-13: Supplier Inventory *(SP 800-161)*

NIST Control Text:

a. Develop, document, and maintain an inventory of suppliers that:

1. Accurately and minimally reflects the organization's tier one suppliers that may present a cybersecurity risk in the supply chain;
2. Is at the level of granularity deemed necessary for assessing criticality and supply chain risk, tracking, and reporting;
3. Documents the following information for each tier one supplier: (i) unique identifier for procurement instrument; (ii) description of the supplied products and/or services; (iii) program, project, and/or system that uses the supplier's products and/or services; and (iv) assigned criticality level that aligns to the criticality of the program, project, and/or system.

b. Review and update the supplier inventory [Assignment: enterprise-defined frequency].

SL5 Supplemental Guidance:

Apply SP 800-161 Rev 1 guidance [3] for maintaining a comprehensive, criticality-based inventory of all suppliers documenting supplier identities, products provided, and assigned risk levels.

SC-7: Boundary Protection

NIST Control Text:

a. Monitor and control communications at the external managed interfaces to the system and at key internal managed interfaces within the system;

b. Implement subnetworks for publicly accessible system components that are [Selection: physically; logically] separated from internal organizational networks; and

c. Connect to external networks or systems only through managed interfaces consisting of boundary protection devices arranged in accordance with an organizational security and privacy architecture.

SL5 Supplemental Guidance:

The SL5 Network encompasses SL5 model development, training, and deployment operations. External network connections are prohibited to prevent unauthorized access and exfiltration while supporting SL5 activities.

The SL5 Network may span multiple physical locations connected through managed interfaces with encrypted channels. Network access locations implement physical security per PE family controls.

SC-7(10): Prevent Exfiltration

NIST Control Text:

a. Prevent the exfiltration of information; and

b. Conduct exfiltration tests [Assignment: organization-defined frequency].

SC-7(21): Isolation of System Components

NIST Control Text:

Employ boundary protection mechanisms to isolate [Assignment: organization-defined system components] supporting [Assignment: organization-defined missions and/or business functions].

SL5 Supplemental Guidance:

Weight Enclaves isolate systems requiring direct access to covered models within the SL5 Network. This isolation protects against weight exfiltration while enabling operations such as training, inference, fine-tuning, and mechanistic interpretability. Boundary protection mechanisms enable code deployment and API access while preventing weight exfiltration.

Each Weight Enclave resides in a single physical facility. Multiple Weight Enclaves may be established at different facilities, but may not be directly connected via network. Transfers exceeding the Weight Enclave outbound bandwidth limit must occur via encrypted physical media with appropriate physical safeguards. This has significant implications for geographically distributed training; Section 1.3 highlights this as a key open question.

Systems not requiring direct weight access operate outside Weight Enclaves, including lower-risk models. Weight Enclaves execute only authorized software (CM-7(5)).

SC-8(1): Cryptographic Protection

NIST Control Text:

Implement cryptographic mechanisms to [Selection (one or more): prevent unauthorized disclosure of information; detect changes to information] during transmission.

SL5 Supplemental Guidance:

Accelerator Interconnect Encryption: AI accelerators within Weight Enclaves cryptographically protect all data transmitted over chip-to-chip interconnects (e.g., NVLink, UALink, custom fabrics). Hardware-level encryption prevents interception from physical interconnects during distributed operation.

Interconnect encryption protects data between accelerators; end-to-end encryption (where data arrives encrypted from origin and is re-encrypted before host-accessible export) protects data from host access at trust boundary crossings. Both are required. The accelerator must revoke host memory access before decrypting incoming data and complete encryption before restoring host access.

Inter-Facility Encryption: The organization implements cryptographic protection for all inter-facility network traffic within the SL5 Network using inline network encryptors deployed at facility boundaries. Inline network encryptors must implement post-quantum cryptographic algorithms conforming to FIPS 203 and FIPS 204 (ML-KEM and ML-DSA) or NSA CNSA 2.0 [31], [32]. The organization deploys at least two inline network encryptors from different suppliers in series at each inter-facility connection per SC-29.

SC-8(5): Protected Distribution System

NIST Control Text:

Implement [Assignment: organization-defined protected distribution system] to [Selection (one or more): prevent unauthorized disclosure of information; detect changes to information] during transmission.

SL5 Supplemental Guidance:

All Weight Enclave network traffic leaving the Red Zone perimeter requires PDS per CNSSI 7003. Unlike standard SCIF requirements (which apply only to unencrypted traffic), SL5 requires PDS for all Weight Enclave traffic including encrypted inter-building connections.

SC-13: Cryptographic Protection

NIST Control Text:

a. Determine the [Assignment: organization-defined cryptographic uses] ; and

b. Implement the following types of cryptography required for each specified cryptographic use: [Assignment: organization-defined types of cryptography].

SC-15(3): Disabling and Removal in Secure Work Areas

NIST Control Text:

Disable or remove collaborative computing devices and applications from [Assignment: organization-defined systems or system components] in [Assignment: organization-defined secure work areas].

SL5 Supplemental Guidance:

No wireless devices or collaborative computing devices (cameras, microphones, video conferencing) are permitted in Red Zones. All equipment must be hardwired with wireless capabilities physically removed or disabled.

SC-28(3): Cryptographic Keys

NIST Control Text:

Provide protected storage for cryptographic keys [Selection: [Assignment: organization-defined safeguards] ; hardware-protected key store].

SL5 Supplemental Guidance:

AI accelerators within Weight Enclaves provide a dedicated secure element for cryptographic keys used in encrypted data paths and attestation. The host system cannot access this key storage.

Hardware-provisioned identity keys (e.g., e-fuse keys embedded during manufacturing) serve as the accelerator’s unforgeable identity, enabling the accelerator to prove its identity to remote parties.

SC-29: Heterogeneity

NIST Control Text:

Employ a diverse set of information technologies for the following system components in the implementation of the system: [Assignment: organization-defined system components].

SL5 Supplemental Guidance:

The organization deploys at least two inline network encryptors from different suppliers in series for each inter-facility connection, consistent with the NSA “Rule of Two” [14]. Different suppliers means different manufacturers or companies producing the encryptors. This heterogeneity protects against supplier-specific implementation vulnerabilities in firmware, hardware design, key management implementations, or protocol handling.

SC-32: System Partitioning

NIST Control Text:

Partition the system into [Assignment: organization-defined system components] residing in separate [Selection: physical; logical] domains or environments based on [Assignment: organization-defined circumstances for the physical or logical separation of components].

SC-49: Hardware-Enforced Separation and Policy Enforcement

NIST Control Text:

Implement hardware-enforced separation and policy enforcement mechanisms between [Assignment: organization-defined security domains].

SL5 Supplemental Guidance:

AI accelerators within Weight Enclaves implement hardware-enforced separation establishing the accelerator as an independent security domain from the host. The accelerator prevents memory access from the host to regions containing sensitive data via DMA or other mechanism. This configuration is controlled by the accelerator and not the host. The host cannot override these policies even with privileged access.

SI-3: Malicious Code Protection

NIST Control Text:

a. Implement [Selection (one or more): signature-based; non-signature-based] malicious code protection mechanisms at system entry and exit points to detect and eradicate malicious code;

b. Automatically update malicious code protection mechanisms as new releases are available in accordance with organizational configuration management policy and procedures;

c. Configure malicious code protection mechanisms to:

1. Perform periodic scans of the system [Assignment: organization-defined frequency] and real-time scans of files from external sources at [Selection (one or more): endpoint; network entry and exit points] as the files are downloaded, opened, or executed in accordance with organizational policy; and

2. [Selection (one or more): block malicious code; quarantine malicious code; take [Assignment: organization-defined action]] ; and send alert to [Assignment: organization-defined personnel or roles] in response to malicious code detection; and

d. Address the receipt of false positives during malicious code detection and eradication and the resulting potential impact on the availability of the system.

SI-3(10): Malicious Code Analysis

NIST Control Text:

a. Employ the following tools and techniques to analyze the characteristics and behavior of malicious code: [Assignment: organization-defined tools and techniques] ; and

b. Incorporate the results from malicious code analysis into organizational incident response and flaw remediation processes.

SI-7(9): Verify Boot Process

NIST Control Text:

Verify the integrity of the boot process of the following system components: [Assignment: organization-defined system components].

SI-7(10): Protection of Boot Firmware

NIST Control Text:

Implement the following mechanisms to protect the integrity of boot firmware in [Assignment: organization-defined system components]: [Assignment: organization-defined mechanisms].

SL5 Supplemental Guidance:

AI accelerators accept only manufacturer-signed firmware, verified using keys embedded in hardware during manufacturing. The accelerator rejects unsigned or incorrectly signed firmware even with full host system access.

Firmware protection is essential because firmware implements all other security mechanisms. If an attacker can modify firmware, they can disable memory isolation, attestation, and encrypted data paths.

SI-7(15): Code Authentication

NIST Control Text:

Implement cryptographic mechanisms to authenticate the following software or firmware components prior to installation: [Assignment: organization-defined software or firmware components].

SL5 Supplemental Guidance:

AI accelerators verify that code is cryptographically signed before execution. This extends beyond firmware to operator binaries composing workload execution.

The organization ensures that approved workloads cannot be exploited by a compromised host to exfiltrate confidential information. Beyond verifying that individual operations do not leak data, this requires preventing composition attacks where small operations are sequenced to construct exfiltration channels.

Example approaches include task sequence verification, where firmware ensures operations execute in an approved sequence preventing a host from constructing malicious workflows, and restricting workloads to fused kernels that have been verified to not exfiltrate data individually or through composition.

Organizations implement continuous attestation verification with tamper-evident logging.

References

[1] S. Nevo, D. Lahav, A. Karpur, Y. Bar-On, H. Alexander Bradley, and J. Alstott, “Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models,” RAND Corporation, Santa Monica, CA, USA, RR-A2849-1, 2024. doi: 10.7249/RRA2849-1. [Online]. Available: https://www.rand.org/pubs/research_reports/RRA2849-1.html

[2] Joint Task Force, “Security and Privacy Controls for Information Systems and Organizations,” NIST Special Publication 800-53, Rev. 5 (Final; includes updates as of Dec. 10, 2020), National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, Sep. 2020. doi: 10.6028/NIST.SP.800-53r5. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r5.pdf

[3] J. Boyens, A. Smith, N. Bartol, K. Winkler, A. Holbrook, and M. Fallon, “Cybersecurity Supply Chain Risk Management Practices for Systems and Organizations,” NIST Special Publication 800-161 Revision 1 Update 1 (includes updates as of Nov. 1, 2024), National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, May 2022. doi: 10.6028/NIST.SP.800-161r1-upd1. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-161r1-upd1.pdf

[4] Committee on National Security Systems, “Security Categorization and Control Selection for National Security Systems,” CNSSI No. 1253, Mar. 27, 2014. [Online]. Available: https://www.dcsa.mil/Portals/69/documents/io/rmf/CNSSI_No1253.pdf

[5] FedRAMP Program Management Office, “FedRAMP Security Controls Baseline” (spreadsheet), FedRAMP. [Online]. Available: https://www.fedramp.gov/resources/documents/FedRAMP_Security_Controls_Baseline.xlsx

[6] Office of the Director of National Intelligence, “Sensitive Compartmented Information Facilities,” Intelligence Community Directive (ICD) 705 (Effective: 26 May 2010). [Online]. Available: https://www.intelligence.gov/assets/documents/intelligence-community-directives/ICD_705.pdf

[7] Office of the Director of National Intelligence, “Physical and Technical Security Standards for Sensitive Compartmented Information Facilities,” Intelligence Community Standard (ICS) 705-1 (Effective: 17 September 2010). [Online]. Available: https://www.dni.gov/files/NCSC/documents/Regulations/ICS-705-1.pdf

[8] Committee on National Security Systems, *Protected Distribution Systems (PDS)*, CNSSI No. 7003, 2015. [Online]. Available: https://www.dcsa.mil/Portals/91/documents/ctp/nao/CNSSI_7003_PDS_September_2015.pdf

[9] National Security Telecommunications and Information Systems Security Advisory Memorandum (NSTISSAM) TEMPEST/1-92, “Compromising Emanations Laboratory Test Requirements, Electromagnetics,” National Security Agency, 1992. [Online]. Available: https://cdn.preterhuman.net/texts/government_information/intelligence_and_espionage/homebrew.military.and.espionage.electronics/servv89pn0aj.sn.sourcedns.com/_gbpprorg/mil/vaneck/nsa/nt1-92-1-5.htm

[10] National Security Agency, “Specification for RF Shielded Enclosures,” NSA No. 94-106, Fort Meade, MD, USA. [Online]. Available: https://linas.org/mirrors/cryptome.org/20050616.nsa94-106.pdf

[11] National Institute of Standards and Technology, *Security Requirements for Cryptographic Modules*, FIPS PUB 140-3, Mar. 2019. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.140-3.pdf

[12] NIST, “FIPS 140-3 Adopts ISO/IEC Standards,” ITL Bulletin, May 2019. [Online]. Available: https://csrc.nist.gov/files/pubs/shared/itlb/itlbul2019-05.pdf

[13] NIST, “Cryptographic Module Validation Program (CMVP) – FIPS 140-3 Standards,” NIST CSRC Project Documentation. [Online]. Available: https://csrc.nist.gov/projects/cryptographic-module-validation-program/fips-140-3-standards

[14] National Security Agency, “Commercial Solutions for Classified (CSfC) Program — Customer Handbook,” Feb. 6, 2021. [Online]. Available: https://media.defense.gov/2021/Apr/02/2002613880/-1/-1/0/CSFC%20PMO%20CUSTOMER%20HANDBOOK_02062021.PDF/CSFC%20PMO%20CUSTOMER%20HANDBOOK_02062021.PDF

[15] Confidential Computing Consortium, “Confidential Computing: Hardware-Based Trusted Execution for Applications and Data,” CCC White Paper, v1.3, Nov. 2022. [Online]. Available: https://confidentialcomputing.io/wp-content/uploads/sites/10/2023/03/CCC_outreach_whitepaper_updated_November_2022.pdf

[16] NVIDIA Corp., “Confidential Computing on NVIDIA H100 GPUs for Secure and Trustworthy AI,” NVIDIA Technical Blog, Aug. 2023. [Online]. Available: https://developer.nvidia.com/blog/confidential-computing-on-h100-gpus-for-secure-and-trustworthy-ai/

[17] A. Dhar et al., “Ascend-CC: Confidential Computing on Heterogeneous NPUs for Emerging Generative AI Workloads,” arXiv:2407.11888, 2024. [Online]. Available: https://arxiv.org/abs/2407.11888

[18] B. Biggio and F. Roli, “Wild Patterns: Ten years after the rise of adversarial machine learning,” *Pattern Recognit.*, vol. 84, pp. 317–331, Dec. 2018. [Online]. Available: https://arxiv.org/pdf/1712.03141.pdf

[19] N. Papernot, P. McDaniel, A. Sinha, and M. P. Wellman, “SoK: Security and Privacy in Machine Learning,” in *Proc. IEEE European Symposium on Security and Privacy (EuroS&P)*, 2018. [Online]. Available: https://oaklandsok.github.io/papers/papernot2018.pdf

[20] The SL5 Task Force, "SL5 Novel Recommendations," preliminary, Nov. 2025. [Online]. Available: https://sl5.org/projects/sl5-novel-recommendations

[21] U.S. Department of Defense, Defense Federal Acquisition Regulation Supplement (DFARS), “252.239-7000 — Protection Against Compromising Emanations (OCT 2019)” (printable PDF), Acquisition.gov. [Online]. Available: https://www.acquisition.gov/node/36728/printable/pdf

[22] Office of the Director of National Intelligence, “Standards for the Accreditation and Reciprocal Use of Sensitive Compartmented Information Facilities,” Intelligence Community Standard (ICS) 705-02, Dec. 22, 2016. [Online]. Available: https://www.dni.gov/files/NCSC/documents/Regulations/ICS_705-2_Standards_for_Accreditation_Reciprocal_Use_of_SCIFs.pdf

[23] National Counterintelligence and Security Center, Office of the Director of National Intelligence, “Technical Specifications for Construction and Management of Sensitive Compartmented Information Facilities,” VERSION 1.5 (IC Tech Spec – for ICD/ICS 705), Mar. 13, 2020. [Online]. Available: https://www.dni.gov/files/Governance/IC-Tech-Specs-for-Const-and-Mgmt-of-SCIFs-v15.pdf

[24] NVIDIA Corp., “NVIDIA Secure AI with Blackwell and Hopper GPUs,” White Paper, WP-12554-001_v1.3, Aug. 2025. [Online]. Available: https://docs.nvidia.com/nvidia-secure-ai-with-blackwell-and-hopper-gpus-whitepaper.pdf

[25] National Institute of Standards and Technology, “NIST Releases Revision to SP 800-53 Security and Privacy Controls,” Computer Security Resource Center (CSRC), Aug. 27, 2025. [Online]. Available: https://csrc.nist.gov/News/2025/nist-releases-revision-to-sp-800-53-controls

[26] The SL5 Task Force, "The Sensitivity Levels Framework (SenLs)," Nov. 2025. [Online]. Available: https://sl5.org/projects/sensitivity-levels-framework

[27] Office of the Under Secretary of Defense for Acquisition and Sustainment, "Trusted Supplier Programs," Defense Microelectronics Activity (DMEA), Trusted Access Program Office (TAPO). [Online]. Available: https://www.acq.osd.mil/asds/dmea/tapo/trusted-supplier-programs.html

[28] Committee on National Security Systems, "Committee on National Security Systems (CNSS) Glossary," CNSSI No. 4009, Apr. 6, 2015. [Online]. Available: https://www.dni.gov/files/NCSC/documents/nittf/CNSSI-4009_National_Information_Assurance.pdf

[29] S. Altman, J. Pachocki, and W. Zaremba, “Sam, Jakub, and Wojciech on the future of OpenAI with audience Q&A,” YouTube, Oct. 29, 2025. Accessed: Feb. 9, 2026. [Online Video]. Available: https://www.youtube.com/watch?v=ngDCxlZcecw

[30] D. Amodei, “The Adolescence of Technology,” *Dario Amodei Blog*, January 2026. [Online]. Available: https://www.darioamodei.com/essay/the-adolescence-of-technology

[31] National Institute of Standards and Technology, "Post-Quantum Cryptography: FIPS 203, FIPS 204, and FIPS 205," NIST, Aug. 2024. [Online]. Available: https://csrc.nist.gov/news/2024/postquantum-cryptography-fips-approved

[32]National Security Agency, "Commercial National Security Algorithm Suite 2.0 (CNSA 2.0)," NSA Cybersecurity Advisory, Sep. 2022. [Online]. Available: https://media.defense.gov/2022/Sep/07/2003071836/-1/-1/0/CSI_CNSA_2.0_FAQ_.PDF

Appendix A: Additional Open Questions

Organized by security stream:

Machine Security

- Availability of accelerators meeting SL5 hardware security requirements in time for deployment
- Feasibility of workload integrity mechanisms (e.g., task sequence attestation, kernel fusion) for expected workloads such as training, inference, fine-tuning, and mechanistic interpretability

Physical Security

- TEMPEST zone classification for Red Zones
- Shared infrastructure isolation (power, cooling, fire suppression) between distributed Red Zones
- Required attenuation levels for shielded racks (full NSA 94-106 vs reduced spec)
- Level of physical isolation required for Weight Enclaves (rack, room, or building level separation)
- ICD 705 private accreditation: This standard specifies self-certification by CSO or Site Security Manager. Whether private self-certification provides meaningful security assurance compared to government accreditation, and what this implies for actual security posture

Network Security

- Specific bandwidth thresholds for exfiltration prevention (calibrated to model size and threat model)
- Redundancy architectures for fail-closed encryptors

Personnel Security

- Screening equivalency standards for non-US citizens and international offices
- Monitoring feasibility in high-restriction states (California CPRA/ICRAA, Illinois BIPA)

Supply Chain Security

- Testing/certification approaches for adversarial content detection system robustness
- Processing state tracking for new modality onboarding