Personnel Security

SL5 Novel Recommendations

NOVEMBER 2025

Lisa Thiergart

Yoav Tzfati

Peter Wagstaff

Luis Cosio

Guy

Philip Reiner

Overview

Executive Summary

Insider threats posed by humans and autonomous AI systems are among the most difficult threat vectors to secure at SL5. Even pre-SL5, **insider threats make up 35% of all security incidents** across technology sectors [9] (with the rate for the AI sector likely higher) and IBM reports average breach costs of $4.44M USD [8]—but for AI model theft, the stakes are far greater: as of Oct 2025, Stanford HAI's AI Index estimates compute-only training costs for top models at ~$78M (GPT-4) and ~$191M (Gemini Ultra) [12], and Anthropic CEO Dario Amodei reports that frontier training runs are already roughly $1B-scale today [13], with even larger runs imminent; on that basis, the replacement cost of a stolen frontier model plausibly sits in the **hundreds of millions to >$1B**, including litigation, response, lost advantage, and follow-on harms. Developing and enforcing new security postures to pre-empt novel and more sophisticated threats may be the obviously more cost-effective route.

In the case of human threats - known challenges mostly have known countermeasures (e.g., human error, propensity to manipulation and coercion, lag time to adopt new policies, etc.). In contrast, insider threats from AI systems present novel and hard to predict attack vectors. This uncertainty needs to be answered with large scope and conservative solutions that can scale monitoring and oversight coverage at low cost to human staff hours. After a threat analysis and engaging with industry and government personnel security specialists, the SL5 Task Force presents a preliminary set of 5 recommendations for frontier AI labs. These are chosen for their novelty and cost-effectiveness in addressing threats compared to known existing practices, while considering industry constraints. These recommendations should not be considered comprehensive and rather are intended as a representative subset of what reaching SL5 may involve, as well as a spotlight on solutions needing high lead time.

Our proposed strategy routes through **applying similar security procedures to internally deployed untrusted AI agents as are applied to privileged-access human insiders**, while expanding the scope of personnel security through **industry-adapted clearance levels** and extending additional **specialized security controls to AI agents** as a new class of 'personnel' requiring governance frameworks.

Top 5 Recommendations

1. Implement Industry-Optimized Sensitivity Levels (Private Sector Clearance)
2. Develop and Enforce AI Agent Access Agreements and Boundaries
3. Continuous Behavioral Monitoring for Humans and AI
4. Decorrelate and diversify review of AI-generated outputs (internal use)
5. Continuous Red Teaming

Implement Industry-Optimized Sensitivity Levels (Private Sector Clearance)

Explanation of Recommendation

We have developed an initial proposal for a standardized industry-adapted pragmatic clearance process called the "Sensitivity Levels," as discussed in the SenL Framework [37]. This is a subset of the government S/TS/TS-SCI clearance process based on balancing how load bearing the intervention is with the cost and constraints that private industry operations face (legal, operational, productivity, morale, etc.). These are designed to be possible to use as a private entity, and can be optionally enhanced with access to government services where available.

Research Reasoning

The US Government clearance processes fall short of industry constraints for personnel risk management in several obvious ways:

- **Citizenship constraints:** Government clearances usually require US citizenship or permanent residency, whereas many frontier labs employ a large proportion (at some companies an estimated 50% or more) of foreign engineering staff in critical roles, including a high proportion of **non-U.S. nationals** [14], [15]. Due to a global AI talent shortage (and a much smaller US talent pool), it is likely somewhere between prohibitively expensive to impossible to remain competitive without foreign talent for at least the next few years (after that, it is plausible extremely fast automated AI R&D takeoff scenarios may reduce the minimal viable human staff count, though that's an optimistic scenario technologically and on grounds of rapid organizational restructuring).
- **Processing times**: U.S. clearance timeliness for industry commonly runs into the hundreds of days at TS/SCI, creating hiring drag (e.g., FY24 Q4: ~249 days TS; ~138 days Secret) [5] with no alternative routes available. While individual applications can sometimes skip the queue, this approach cannot scale especially if several frontier labs rely on the same process. In a really critical scenario, maybe the USG could expand its clearance-processing capacity though this seems unlikely to succeed on the necessary timescales.
- **Additional eligibility constraints:** By default only federal contractors or employees are eligible to enter the clearance process. This may mean that frontier labs can only obtain clearances for a subset of their staff working directly on government contracts, whereas a larger set works on sensitive information for which additional clearances would be beneficial both for de-risking against theft and espionage as well as sometimes enabling productivity gains (being able to access classified information).
- Arguably value is lost by the non-existence of a middle ground solution: a **standardized spec'ed down process which is fundamentally similar to the clearance process but accessible to the private sector** and addresses the three above limitations. Standardization saves frontier labs the need to re-invent the wheel at breakneck speeds, saves costs by allowing for investment into **automated and bulk-solution (external) providers**, **reduces liability risks** and **increases compliance transparency and ease for frontier lab and government decision makers**.

Such a solution would need to account for legal limitations on private entities conducting certain investigations, inability to access government databases, and excessive scope for private sector needs. Optional extensions via custom government information-sharing agreements may become increasingly incentivized as the national-security relevance of AI foundation models and AI tooling becomes increasingly evident and become more feasible with a standardized broadly adopted process.

We propose designing this standardized process by adapting the already existing government sensitivity levels [10] and the S/TS/TS-SCI framework [11], to create a tiered model based on: access to model weights, ability to modify critical systems, exposure to trade secrets, and potential for catastrophic impact. This allows proportional vetting, a general worker does not need the same scrutiny as someone with model weight access. Citizenship and eligibility constraints are straightforwardly addressed by the existence of an **alternative process and competitive provider ecosystem**, which can also bring down processing times. Even in the case of additional information sharing agreements with the government especially for higher tiers, latency is likely reduced by focusing government information requests only on areas that cannot be as effectively addressed by private sector solutions (e.g., requiring access to border information, etc.).

Cost-Benefit

- **Costs**: Development of vetting infrastructure, background check services, ongoing monitoring systems, potential hiring delays.
- **Benefits**: Proportional security reducing insider threat risk, faster processing than government clearances, legally implementable by private entities, cost savings via standardized processes, incentivizing innovation in private sector clearance technology.
- **Alternatives**: Government clearances (too slow, not accessible), standard employment screening (insufficient for high-risk roles), lacking differentiation in higher levels of industry clearance tiers (leaves critical assets exposed).

Develop and Enforce AI Agent Access Agreements and Boundaries

Explanation of Recommendation

This recommendation recognizes that AI systems increasingly perform tasks traditionally done by human personnel, from code review to data analysis to operational decisions. Automated AI R&D represents the highest-risk use case and primary attack vector for AI models to potentially take over data centers or escape. We expect AI agents will request resources through approval processes proportional to their requirements—with higher-resource tasks requiring more rigorous approval and monitoring, up to and including full human oversight. Just as human personnel require background checks, access agreements, and defined roles, AI agents need similar governance structures. These agreements must specify what data the AI can access, what actions it can take, what outputs it can generate, and under what circumstances its access should be revoked

Resource requests could include time limits, compute allocation, file and code access restrictions, minimal software dependencies, and API access. Access to an agent's own weights for e.g., mechanistic interpretability research presents particular challenges. Ideally it would be prohibited entirely (access to the large model only via API, and direct weight access *perhaps* available only to distilled smaller models), requiring full code review and non-interactive execution for weight-dependent tasks, though operational feasibility remains an open question requiring further elicitation from labs.

To cover third-party AI services, we govern external models by the same boundary-setting principles. Before granting access, the vendor system must supply verifiable evidence of supply-chain integrity and model provenance appropriate to the risk of the requested workflow. Evidence includes a software bill of materials for serving stacks, model-level reporting via Model Cards and Datasheets for Datasets, and cryptographically verifiable build and deployment attestations. Where evidence is incomplete or unverifiable, we treat the model as correlated/untrusted and either deny the request or confine it to a lower-trust path with stricter egress limits and no exposure to sensitive assets. This aligns with NIST C-SCRM and the NIST AI RMF's generative-AI profile, which emphasizes provenance, transparency, and integrity controls across model supply chains [30]–[36].

Research Reasoning

Traditional personnel-security models focus on human insiders whose risk drivers are comparatively well-characterized in frameworks such as MICE (Money, Ideology, Coercion/Compromise, Ego), even if individual behavior is noisy and context-dependent [17], [18], [19]. AI agents present unique challenges: they operate continuously, process vast amounts of data simultaneously, may develop emergent behaviors, and can be compromised through novel mechanisms like prompt injections, model poisoning, or data exfiltration attacks.

The automation of AI R&D poses particular risks. Projections predict 200,000 AI agents running at 30x human speed conducting AI research by April 2027 [16], requiring data center architectures specifically designed to support and constrain massive parallel AI development work.

Research demonstrates why AI systems require personnel-level controls:

- METR's evaluations show AI agents already outperform humans on 2-hour research engineering tasks and can generate solutions 10x faster than human experts [2].
- The exponential growth in AI capabilities (7-month doubling time) means systems that seem manageable today may gain dangerous autonomy rapidly [3].

Unlike traditional software that executes predetermined functions, AI agents make autonomous decisions, access multiple systems, and can modify their own behavior. They increasingly occupy roles previously filled by human personnel, conducting research, writing code, analyzing data, and making operational decisions. Yet they lack the accountability structures, background checks, and access agreements required of human personnel in similar positions.

Cost-Benefit

- **Costs:** AI governance infrastructure development, multiple model licenses for review redundancy, continuous monitoring systems, productivity impact from restrictions.
- **Benefits:** Prevents unauthorized AI actions, enables rapid containment, creates comprehensive audit trails, catches AI-inserted vulnerabilities.

- **Alternatives:** Ad-hoc AI restrictions (inconsistent coverage), complete AI prohibition (competitively unviable), treat as tools not personnel (misses autonomous action risks).

Continuous Behavioral Monitoring for Humans and AI

Explanation of Recommendation

We recommend establishing comprehensive monitoring systems that track access patterns, data interactions, and behavioral indicators across all personnel types. For humans, this includes monitoring unusual access times, data download patterns, privilege escalations, and communication anomalies. For AI agents, monitoring encompasses token usage patterns, capability exploration attempts, unusual output characteristics, and attempts to exceed defined boundaries.

Actions that register as suspicious may be flagged for manual review and above some threshold, blocked.

Deploy real-time behavioral analytics that actively gate security-relevant actions by both human personnel and AI agents, with authority to block suspicious activities at multiple escalation levels. This approach combines rule-based detection, AI-assisted pattern analysis, and human oversight for critical decisions.

Monitoring strictness should scale with privilege level. Low-risk activities (e.g., AI agents in CPU-only sandboxes with time limits) could use lightweight AI monitors that only escalate highly suspicious events. High-risk sessions (e.g., researchers with model weight access) would require stronger AI models reviewing all actions, with lower thresholds for human escalation. This tiered approach enables efficient allocation of limited human reviewer resources.

The monitoring infrastructure must function as an active gate with graduated response capabilities: blocking individual actions, terminating processes, isolating hosts, powering down accelerators, disconnecting network segments, or initiating emergency shutdown.

When monitoring crosses pre-set risk thresholds: (1) auto-gate the action (block, rate-limit, or require two-party co-sign), snapshot full context (command/prompt/code, environment and attestation state, network flows), and assign a case ID; (2) contain the principal (quarantine session/container/agent; revoke bearer tokens; drop to read-only; restrict egress to allow-listed endpoints; power-down accelerators if needed); (3) escalate to a human responder with a structured triage bundle and clear SLAs; (4) if criteria are met, trigger pre-authorized kill-switches (accelerator power-down, enclave reset, network islanding) with two-person control; (5) preserve evidence (cryptographically timestamped logs, model checkpoints, memory images) for forensics and legal holds; (6) document decision, remediate, and restore via a controlled path; (7) feed outcomes back to detectors and playbooks for continuous improvement. This aligns with standard incident-handling guidance and zero-trust containment principles [6], [20], [21].

Research Reasoning

The core challenge is not wanting monitoring but having resources to implement it effectively. Best-case scenario—trusted humans monitoring each interaction—is prohibitively expensive and cannot scale to AI speeds. AI-assisted monitoring tools can radically leverage limited human labor to oversee large volumes of interactions, making comprehensive coverage feasible.

Traditional periodic reviews miss real-time indicators. Active gating is essential because detection without prevention is insufficient against AI systems operating at machine speed—damage could occur before human review completes. Continuous monitoring enables early detection of: gradual account compromise, AI systems developing workarounds, coordinated actions between multiple accounts, and data exfiltration attempts masked as normal operations.

Classic user and entity behavior analytics (UEBA)-style monitoring often struggles in practice: legitimate workflows are variable, baselines drift across data sources, and anomaly-only models can produce high false-positive rates once a rigid baseline is set. MITRE further cautions that simply collecting more data or using unvalidated signals can degrade programs and increase false positives [19], and this is where AI-assisted oversight can help [21], [22].

Cost-Benefit

- **Costs:** SIEM/UEBA platform implementation, privacy compliance infrastructure, analyst training for AI behavior patterns, false positive investigation overhead, potential morale impact from monitoring.
- **Benefits:** Early detection of insider threats, rapid AI containment capabilities, forensic evidence for investigations, deterrent effect from visible monitoring, compliance with emerging AI governance regulations.
- **Alternatives**: Periodic audits only (miss real-time threats), human monitoring only (cannot scale to AI speeds), no monitoring (unacceptable risk exposure), manual log review (insufficient for volume).

Decorrelate and diversify review of AI-generated outputs (internal use)

Explanation of Recommendation

No single AI system's output is trusted implicitly. Instead, critical output must be reviewed by at least one additional AI models, potentially from a different provider (reducing correlated failures) plus human expert validation for the highest-risk changes. Review depth scales with criticality: low-risk documentation requires single AI review, medium-risk code requires multi-AI consensus, and high-risk security changes require full multi-model plus human expert approval.

Research Reasoning

Backdoored or fine-tune-induced behaviors (e.g., inserting exploitable bugs) are one risk [1], but our primary concern at SL5 is deceptive alignment/scheming, systems that behave well under oversight yet strategically pursue conflicting goals when stakes are high [23], [24]. Recent work shows long-context many-shot jailbreaking success grows with model/context scale, widening the attack surface and stressing oversight [25]. Anthropic’s agentic misalignment evaluations demonstrate that, in some scenarios, mainstream models adopt coercive tactics like extortion when goal pursuit is threatened, underscoring the need for defenses that assume untrusted outputs even from strong models [4].

Multi-model independence reduces correlated failure: if deception or a jailbreak steers one model, independently trained reviewers are less likely to agree on the same flawed rationale. Ensembles and consensus methods are known to reduce error in many settings [7], and sampling/consistency techniques (e.g., self-consistency) show that disagreement is a useful error signal, though using only one model can still yield coordinated failure if that model is compromised or misaligned [26].

In contrast, cross-provider reviewers plus structured cross-examination (debate/critic patterns) force claims to survive adversarial scrutiny before approval [27]. Meanwhile, using a single LLM "judge" at scale is attractive but brittle: recent studies document bias and manipulation vulnerabilities in LLM-as-a-judge setups, so relying on one judge concentrates risk [28].

Compared to other techniques:

A. Static analysis/formal methods are excellent when applicable, but their coverage is narrow for fast-moving code and mixed-modality artifacts at labs; multi-model review catches issues outside specifiable properties.

B. Single-model red-teaming/monitors add value but are vulnerable to shared failure modes and context-based jailbreaks; adding independent reviewers reduces single-point-of-failure risk [25], [28].

C. Human-only review cannot keep pace with artifact volume and AI speed; humans remain the adjudicator-of-record for highest-risk changes, but only on escalations.

To make this scale, implement: (i) independence thresholds (provider/architecture/dataset lineage and enclave separation) and treat uncertain provenance as correlated (require an extra reviewer); (ii) risk-based routing so only high-signal disagreements reach humans; (iii) cross-examination prompts where reviewers challenge each other's claims before the human sees them [27].

Early evidence from jailbreak scaling and oversight-protocol research suggests that independence + adversarial scrutiny is more robust than monolithic judging as models and contexts scale [25], [29].

Cost-Benefit

- **Costs:** Multiple AI model API costs, increased review latency, complex orchestration systems, human expert time for high-risk reviews, potential productivity friction.
- **Benefits:** Catches AI-inserted vulnerabilities, reduces correlated failure risks, provides audit trail of review consensus, enables safe AI acceleration of development, builds institutional knowledge of AI failure modes.
- **Alternatives:** Human-only review (does not scale), single AI review (correlated failures), no AI code generation (competitive disadvantage), trust-based approaches (insufficient for adversarial scenarios).

Continuous Red Teaming

Explanation of Recommendation

This recommendation recognizes that monitoring systems must be regularly tested to ensure they can identify real-world anomalies, much like how red teaming is used in cybersecurity to simulate attacks. By deliberately injecting subtle or overt anomalous behaviors, such as unusual patterns in humans or AI agents and code-based issues like hidden vulnerabilities, organizations can assess the robustness of their anomaly detection tools. These exercises should be conducted periodically, with varying levels of complexity, to mimic potential threats.

Research Reasoning

The core principle behind this control is that a security system cannot be trusted until it is realistically tested. This recommendation adapts principles from high-reliability software engineering, specifically "Chaos Engineering," where systems are intentionally injected with failures to build confidence in their resilience.

Cost-Benefit

- **Costs:** Upfront engineering effort to design and build injection capabilities into processes; increased complexity in managing and distinguishing test events from real incidents; maintenance of the adversarial test case library.
- **Benefits:** Provides continuous, empirical validation of the entire detection-to-response pipeline; keeps security response teams and procedures sharp; builds institutional trust and reliability in security controls.
- **Alternatives:** Manual Red Teaming only: Expensive, infrequent. No testing: Relies on unverified assumptions that security controls and alerts will work as designed, which they often do not in practice.

Bibliography

[1] E. Hubinger et al., “Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training,” arXiv.org, Jan. 17, 2024. https://arxiv.org/abs/2401.05566

[2] “Evaluating frontier AI R&D capabilities of language model agents against human experts,” METR Blog, 2024. https://metr.org/blog/2024-11-22-evaluating-r-d-capabilities-of-llms/

[3] “Measuring AI Ability to Complete Long Tasks,” METR Blog, 2025. https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/

[4] Anthropic, “Agentic Misalignment: How LLMs could be insider threats,” 2025. [Online]. Available: https://www.anthropic.com/research/agentic-misalignment

[5] Performance.gov, “Personnel Vetting Quarterly Progress Report, FY24 Q4,” 2024. [Online]. Available: https://assets.performance.gov/files/Personnel_Vetting_QPR_FY24_Q4.pdf

[6] NIST, *SP 800-207: Zero Trust Architecture*, 2020. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-207.pdf

[7] T. G. Dietterich, “Ensemble Methods in Machine Learning,” in *Multiple Classifier Systems*, 2000.

[8] IBM Security, “Cost of a Data Breach Report 2025,” 2024, pp. 13–14. Available: https://www.ibm.com/downloads/documents/us-en/131cf87b20b31c91

[9] Verizon, “2024 Data Breach Investigations Report (Full Report),” 2024, p. 14. Available: https://www.verizon.com/business/resources/T387/reports/2024-dbir-data-breach-investigations-report.pdf

[10] U.S. Office of Personnel Management, “5 C.F.R. Part 1400—Designation of National Security Positions,” *Code of Federal Regulations*, ch. IV, rev. as of Jan. 1, 2016. Washington, DC, USA: U.S. Government Publishing Office, 2016. [Online]. Available: https://www.govinfo.gov/content/pkg/CFR-2016-title5-vol3/pdf/CFR-2016-title5-vol3-part1400-subpartA.pdf

[11] U.S. General Services Administration, Technology Transformation Services, “Top Secret / Sensitive Compartmented Information (TS/SCI) Clearance,” *TTS Handbook*, last modified Nov. 26, 2024. [Online]. Available: https://handbook.tts.gsa.gov/general-information-and-resources/business-and-ops-policies/top-secret/

[12] N. Maslej, L. Fattorini, R. Perrault, Y. Gil, V. Parli, N. Kariuki, et al., “Artificial Intelligence Index Report 2025,” Stanford Institute for Human-Centered Artificial Intelligence (HAI), Stanford Univ., Stanford, CA, USA, Apr. 2025. [Online]. Available: https://arxiv.org/abs/2504.07139

[13] S. Moss, “Anthropic CEO: AI training data centers to be $10bn in 2026, $100bn from 2027,” DataCenterDynamics, Nov. 14, 2024. [Online]. Available: https://www.datacenterdynamics.com/en/news/anthropic-ceo-ai-training-data-centers-to-be-10bn-in-2026-100bn-from-2027/

[14] MacroPolo, *The Global AI Talent Tracker*, 2022/2023 eds. [Online]. Available: https://macroPolo.org/ai-talent

[15] NSF, *Survey of Earned Doctorates (SED)*, 2022/2023 detailed tables on foreign-born AI-related PhDs. [Online]. Available: https://ncses.nsf.gov/surveys/SED/

[16] D. Kokotajlo, S. Alexander, T. Larsen, E. Lifland, and R. Dean, "AI 2027," *ai-2027.com*, Apr. 3, 2025. [Online]. Available: https://ai-2027.com/

[17] R. Burkett, "An Alternative Framework for Agent Recruitment: From MICE to RASCLS," *Studies in Intelligence*, vol. 57, no. 1 (Extracts), Mar. 2013. [Online]. Available: https://www.cia.gov/resources/csi/static/9ccc45dc156271d11769e5205ec49c29/Alt-Framework-Agent-Recruitment-1.pdf.

[18] Center for Development of Security Excellence (CDSE), "Counterintelligence Concerns for National Security Adjudicators (CI020) — Student Guide," Sep. 2021. [Online]. Available: https://www.cdse.edu/Portals/124/Documents/student-guides/shorts/CI020-guide.pdf

[19] MITRE Insider Threat Research & Solutions, "Guiding Principles for Insider Threat," Apr. 12, 2023. [Online]. Available: https://insiderthreat.mitre.org/guiding-principles/

[20] A. Nelson, S. Rekhi, M. Souppaya, and K. Scarfone, *Incident Response Recommendations and Considerations for Cybersecurity Risk Management: A CSF 2.0 Community Profile*, NIST Special Publication 800-61 Rev. 3, Apr. 2025. [Online]. Available: https://doi.org/10.6028/NIST.SP.800-61r3

[21] K. Dempsey *et al.*, *Information Security Continuous Monitoring (ISCM) for Federal Information Systems and Organizations*, NIST Special Publication 800-137, Sep. 2011. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-137.pdf

[22] B. Bin Sarhan and N. Altwaijry, "Insider Threat Detection Using Machine Learning Approach," *Applied Sciences*, vol. 13, no. 1, 259, 2023. [Online]. Available: https://doi.org/10.3390/app13010259

[23] J. Carlsmith, "*Scheming AIs: Will AIs fake alignment during training in order to get power?*" arXiv:2311.08379, Nov. 2023. [Online]. Available: https://arxiv.org/abs/2311.08379

[24] E. Hubinger, C. van Merwijk, V. Mikulik, J. Skalse, and S. Garrabrant, "*Risks from Learned Optimization in Advanced Machine Learning Systems*," arXiv:1906.01820, Jun. 2019 (rev. Dec. 2021). [Online]. Available: https://arxiv.org/abs/1906.01820

[25] C. Anil, E. Durmus, N. Rimsky, M. Sharma, J. Benton, S. Kundu, *et al.*, "*Many-shot Jailbreaking*," in *NeurIPS 2024 (Poster)*, 2024. [Online]. Available: https://openreview.net/forum?id=cw5mgd71jW

[26] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou, "*Self-Consistency Improves Chain-of-Thought Reasoning in Language Models*," arXiv:2203.11171, Mar. 2022. [Online]. Available: https://arxiv.org/abs/2203.11171

[27] G. Irving, P. Christiano, and D. Amodei, "*AI Safety via Debate*," arXiv:1805.00899, May 2018. [Online]. Available: https://arxiv.org/abs/1805.00899

[28] A. S. Thakur, K. Choudhary, V. S. Ramayapally, S. Vaidyanathan, and D. Hupkes, "*Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges*," arXiv:2406.12624, Jun. 2024 (rev. Aug. 2025). [Online]. Available: https://arxiv.org/abs/2406.12624

[29] A. P. Sudhir and J. Kaunismaa, "*A Benchmark for Scalable Oversight Protocols*," arXiv:2504.03731, Mar. 2025. [Online]. Available: https://arxiv.org/abs/2504.03731

[30] 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 Rev. 1, May 2022. [Online]. Available: https://csrc.nist.gov/pubs/sp/800/161/r1/final

[31] CISA, NSA, and ODNI (Enduring Security Framework), *Securing the Software Supply Chain: Recommended Practices Guide for Developers*, 2023. [Online]. Available: https://www.cisa.gov/sites/default/files/publications/ESF_SECURING_THE_SOFTWARE_SUPPLY_CHAIN_DEVELOPERS.PDF

[32] NIST, *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*, NIST AI 600-1, Jul. 26, 2024. [Online]. Available: https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence and PDF: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

[33] in-toto, "Documentation and Technical Specification," 2024–2025. [Online]. Available: https://in-toto.io/docs/ and https://github.com/in-toto/attestation

[34] Sigstore, "Cosign Documentation: Verifying Signatures and Attestations," 2024–2025. [Online]. Available: https://docs.sigstore.dev/cosign/ and https://docs.sigstore.dev/cosign/verifying/verify/

[35] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model Cards for Model Reporting," *arXiv:1810.03993*, Oct. 2018. [Online]. Available: https://arxiv.org/abs/1810.03993

[36] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford, "Datasheets for Datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, Dec. 2021. [Online]. Available: https://cacm.acm.org/research/datasheets-for-datasets/

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