
Towards a Unified Framework for Cybersecurity and AI Safety: Recommendations for Secure Development of Large Language Models

Lexley Maree Villasis
University of the
Philippines, Diliman

Srishti Dutta
Heritage Institute of
Technology, Kolkata

Yohan Mathew
Independent

With
Johns Hopkins University & Apart Research

Abstract

The increasing sophistication and widespread adoption of AI systems have exposed vulnerabilities that could lead to severe consequences, particularly when exploited by malicious actors. This paper delves into the critical issue of insider threats within AI development organizations, specifically focusing on the potential for sabotage and data breaches.

By analysing the recent incident involving a ByteDance intern, we highlight the urgent need for robust security measures to protect AI infrastructure and sensitive data. We propose a comprehensive framework that integrates technical, internal, and international approaches to mitigate these risks. This framework encompasses:

- **Technical Controls:** Implementing strong access controls, anomaly detection, zero-trust architecture, immutable infrastructure, resource segmentation, and runtime integrity monitoring.
- **Internal Controls:** Establishing robust incident response plans, code review processes, employee training programs, and whistleblower channels.
- **International Cooperation:** Promoting international standards for AI security, such as verification protocols and secure coding frameworks.

By adopting these measures, organisations can significantly reduce the risk of insider threats, safeguard their AI systems, and foster a culture of security and trust. This paper provides actionable recommendations for policymakers, industry leaders, and security professionals to ensure the responsible and secure development of AI.

1. Introduction

A ByteDance intern was [reported](#) (ChinaTalk, 2024) to have allegedly sabotaged the company's large-scale LLM training infrastructure through multiple malicious actions over a period of at least two months, causing potentially millions in damages and delays. In the short-term, this can sow distrust and derail collaboration efforts, thus eroding efforts

toward building a culture of safety between labs and other actors. This highlights the risks associated with lax security and access controls, which could hold even worse repercussions in the case of human-level AGI models in the future.

Current policy approaches to AI security largely stem from cybersecurity frameworks and best practices, as exemplified by the ISO/IEC 27000 series for information security management (ISO/IEC 27000, 2018) and implement network security measures and employee training. However, these may have limitations when addressing insider threats specific to AI development:

1. Traditional security measures are often characterized as reactive, due to their focus on responding to breaches as they occur rather than preventing them proactively. The 2018 Cybersecurity Framework of the National Institute of Standards and Technology (2018) emphasizes the need for a more proactive, prevention-oriented approach to cybersecurity in this regard, and information sharing among different labs is a key aspect in upholding this.
2. Existing policies may fail to effectively account for and prevent malicious actions by inside actors with legitimate access to systems, especially in complex AI development environments.

Our proposal thus introduces a multifaceted approach to this issue by bridging cybersecurity and AI safety considerations in an effort towards an adaptable framework responsive to future emergent developments.

2. Methods

2.1. Technical Controls:

1) Access Control Options:

Various aspects of access control such as:

- Role-based access control (RBAC) with principle of least privilege (PoLP): Define specific roles and assign permissions according to such, only granting the minimum access necessary to perform their specific function.
- Attribute-based Access Control (ABAC): Implement access controls according to factors such as the time, location, and device used.
- Multi-factor Authentication: Requirement for all systems as an added layer of security.

2) Behavioral Anomaly Detection:

Implementing AI-driven monitoring tools for behavioural anomaly detection provides early warnings of potential insider threats by tracking unusual access requests, time-of-access patterns, or unauthorised project modifications. Tools like Microsoft Azure Sentinel utilise AI to analyse activity across infrastructure, identifying irregular access patterns that could signal internal risks. These alerts allow teams to investigate and respond to suspicious behaviour before it escalates.

3) Zero Trust Architecture:

A Zero Trust Architecture assumes no inherent trust for any network user, whether internal or external. Each access request undergoes verification, and access is limited strictly to essential resources. ZTA's continuous verification and access control features

prevent unauthorised movement within the network, significantly reducing the risk of lateral threats and unauthorised resource usage, even from internal accounts.

4) Segmentation of Compute Resources:

Segmenting critical compute resources, such as GPU clusters, limits access to authorised personnel only, safeguarding sensitive areas from unauthorised interactions. Access restrictions and approval processes for each resource zone prevent lower-level users, like interns, from reaching high-risk systems. This segmentation reduces the potential for intentional or unintentional disruptions of key AI training infrastructure.

5) Runtime Integrity Monitoring:

Runtime integrity monitoring protects the AI training pipeline by continuously verifying its state and flagging unauthorised changes as they happen. Tools such as Runtime Application Self-Protection (RASP) detect and block malicious actions in real time, maintaining the training environment's integrity. This approach ensures any injected code or unapproved configurations are identified immediately, preventing potential compromise during critical model training phases.

2.2. Internal Controls and Processes:

1) Enhanced Incident Response Plans:

Effective incident response requires a clearly defined plan that specifies roles, escalation protocols, and rapid containment actions. An organised response chain of command, along with predefined notification and containment procedures, enables quick isolation and remediation of security breaches. Comprehensive post-incident reviews allow teams to identify weaknesses, learn from incidents, and strengthen future response strategies.

2) Internal Code Review Requirements:

Enforcing strict code review processes ensures that sensitive systems are protected from unauthorised changes and vulnerabilities. By requiring peer reviews and senior sign-offs, companies can uphold high standards for code security. Integrating automated code analysis tools complements manual reviews, adding a layer of defence against malicious modifications and unintentional errors. Regular audits further enhance oversight and secure the development environment.

3) Employee Training and Awareness Programs:

Frequent security training promotes a proactive security culture, equipping employees to prevent and respond to potential threats. Security workshops, simulation exercises, and real-world examples familiarise employees with evolving security challenges. Additionally, fostering a network of security champions within teams encourages open discussions and prompt issue identification, ultimately reducing security incidents and enhancing organisational readiness against insider threats.

4) Whistleblower Programs with Anonymity Protections

Establishing secure, anonymous whistleblower channels encourages employees to report unethical or suspicious activities. A clear non-retaliation policy, along with transparency in follow-up actions, builds trust and a sense of responsibility across the organisation. Regular reminders of whistleblower options empower employees to share critical information safely, ensuring early detection and prevention of internal security risks.

2.3. International Governance Frameworks:

1) International Verification Best Practices:

Implement internationally accepted protocols for verifying the security of training datasets and models, ensuring malicious code injections are caught early (e.g., tamper-evident logging systems through [blockchain-based immutable logs](#); [tracking historical versions of datasets](#) to maintain consistency and ensure that model outputs are reproducible). This promotes cross-border trust and prevents insider or outsider tampering during AI development.

2) Uniform Secure Coding Frameworks:

Establish a baseline secure coding standard that all AI labs must adhere to, streamlining safety across jurisdictions. This catches vulnerabilities across different stages of AI training and deployment, creating a more predictable and secure development environment. Inspiration can be drawn from the Secure Software Development Framework (SSDF) of the National Institute of Standards and Technology (NIST).

3. Discussion and Conclusion

The complex nature of AI systems and the evolving threat landscape necessitate a multifaceted approach to security. While technical controls are essential, human factors and organizational culture play equally important roles. Striking a balance between security and innovation is crucial. Continuous monitoring, threat intelligence sharing, and regular security assessments are vital to staying ahead of emerging threats.

The potential for insider threats in AI development poses a significant risk. A comprehensive security framework, encompassing technical, organizational, and international measures, is essential to mitigate these risks. By prioritizing security from the outset and fostering a culture of security awareness, we can ensure the responsible and secure development of AI.

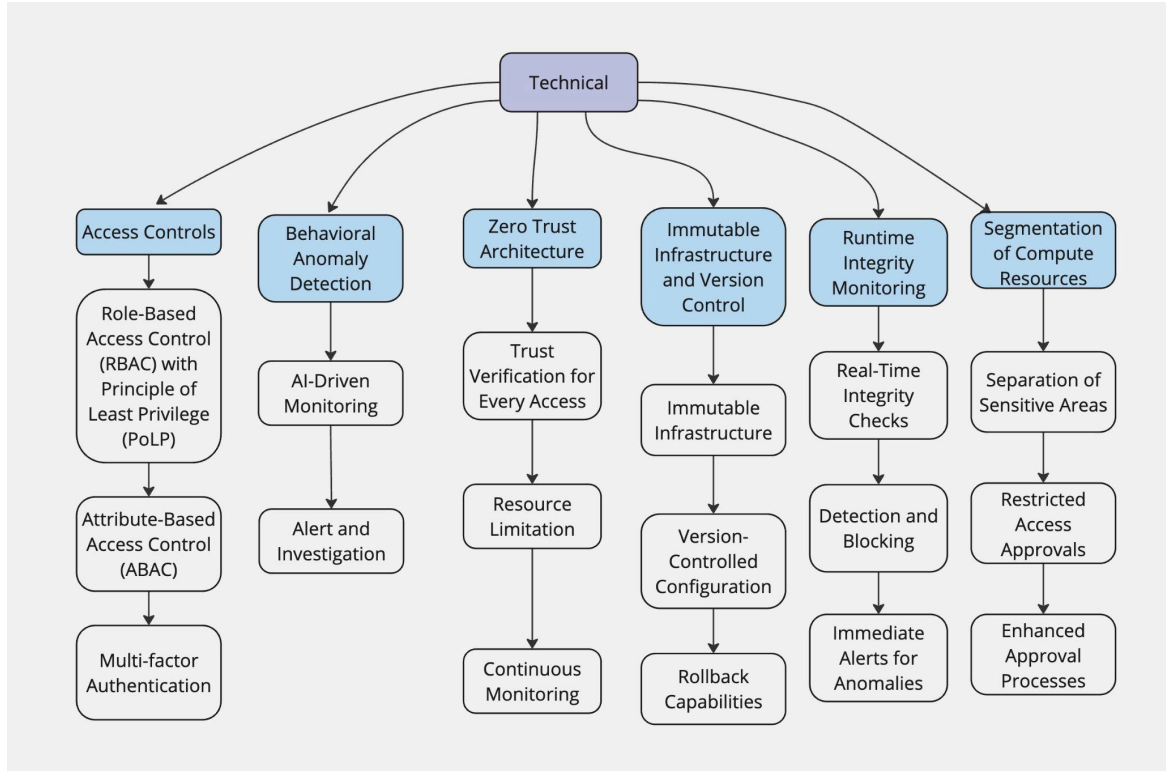
4. References

- Cichonski, P., Millar, T., Grance, T., & Scarfone, K. (2012). *Computer security incident handling guide* (NIST Special Publication 800-61 Rev. 2). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.SP.800-61r2>
- Tibebu, H. (2024, February 3). *Comments on AI Executive Order Request for Information (RFI): Framework for data protection, security, and privacy in AI applications* [ID048]. National Institute of Standards and Technology. Retrieved from <https://www.nist.gov/system/files/documents/2024/02/16/ID048%20-%202024-02-03%2C%20Loughborough%20University%2C%20Comments%20on%20AI%20EO%20RFI.pdf>
- HiddenLayer. (2024, July 31). Securing your AI: A guide for CISOs, Part 4. Retrieved from <https://hiddenlayer.com/research/securing-your-ai-a-guide-for-cisos-pt4/>
- ISO/IEC. (2018). *Information technology—Security techniques—Information security management systems—Overview and vocabulary* (ISO/IEC Standard No. 27000:2018). International Organization for Standardization. <https://www.iso.org/standard/73906.html>
- Ottinger, L., Schneider, J., & Shen, A. (2024, October 25). *ByteDance gets stuxneted; Taiwan healthcare*. ChinaTalk. Retrieved October 26, 2024, from <https://www.chinatalk.media/p/bytedance-gets-suxtnetd-taiwan-healthcare>

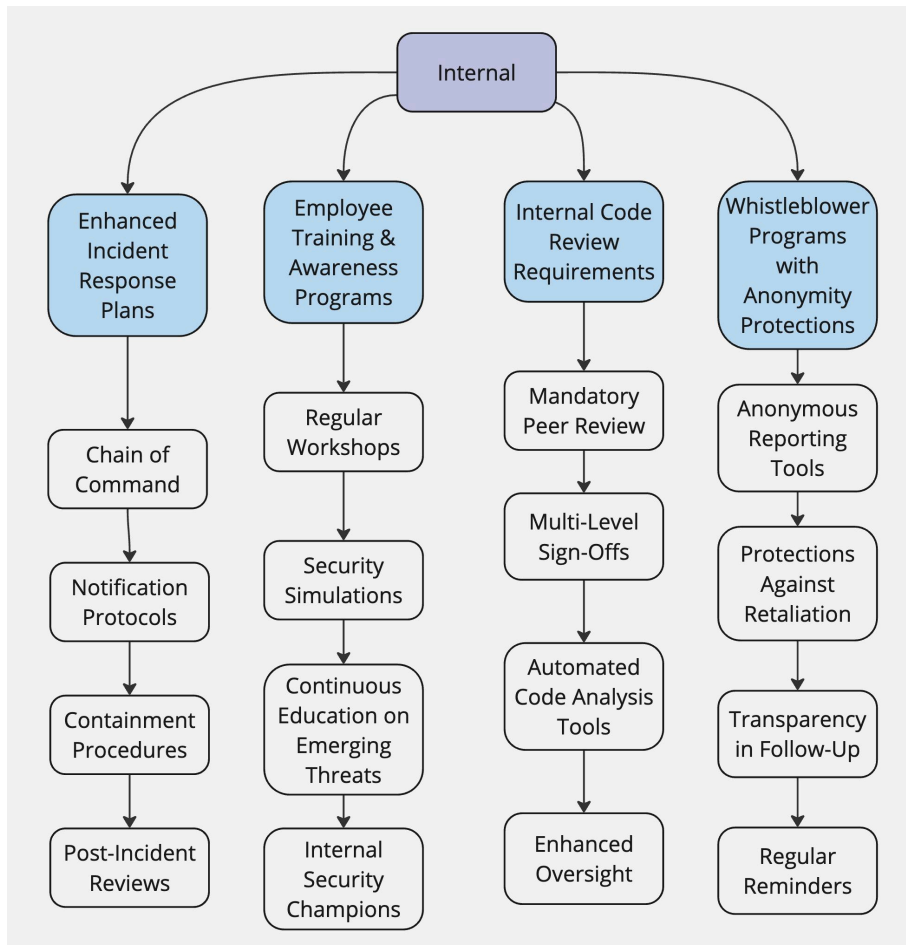
5. Appendix

Diagrammatic representation of each approach:

Technical controls:



Internal controls and processes:



International approach:

