



Australian Government
Australian Signals Directorate

ASD AUSTRALIAN SIGNALS DIRECTORATE
ACSC Australian Cyber Security Centre

National Cyber Security Centre
PART OF THE GCSB


Communications Security Establishment Canada
Canadian Centre for Cyber Security

Centre de la sécurité des télécommunications Canada
Centre canadien pour la cybersécurité


National Cyber Security Centre
a part of GCHQ

Deploying AI Systems Securely

Best Practices for Deploying Secure and Resilient AI Systems

Executive summary

Deploying artificial intelligence (AI) systems securely requires careful setup and configuration that depends on the complexity of the AI system, the resources required (e.g., funding, technical expertise), and the infrastructure used (i.e., on premises, cloud, or hybrid). This report expands upon the 'secure deployment' and 'secure operation and maintenance' sections of the Guidelines for secure AI system development and incorporates mitigation considerations from Engaging with Artificial Intelligence (AI). It is for organizations deploying and operating AI systems designed and developed by another entity. The best practices may not be applicable to all environments, so the mitigations should be adapted to specific use cases and threat profiles. [1], [2]

AI security is a rapidly evolving area of research. As agencies, industry, and academia discover potential weaknesses in AI technology and techniques to exploit them, organizations will need to update their AI systems to address the changing risks, in addition to applying traditional IT best practices to AI systems.

This report was authored by the U.S. National Security Agency's Artificial Intelligence Security Center (AISC), the Cybersecurity and Infrastructure Security Agency (CISA), the Federal Bureau of Investigation (FBI), the Australian Signals Directorate's Australian Cyber Security Centre (ACSC), the Canadian Centre for Cyber Security (CCCS), the New Zealand National Cyber Security Centre (NCSC-NZ), and the United Kingdom's National Cyber Security Centre (NCSC-UK). The goals of the AISC and the report are to:

1. Improve the confidentiality, integrity, and availability of AI systems;
2. Assure that known cybersecurity vulnerabilities in AI systems are appropriately mitigated; and
3. Provide methodologies and controls to protect, detect, and respond to malicious activity against AI systems and related data and services.

This document is marked TLP:CLEAR. Recipients may share this information without restriction. Information is subject to standard copyright rules. For more on the Traffic Light Protocol, see cisa.gov/tlp/.

Scope and audience

The term AI systems throughout this report refers to machine learning (ML) based artificial intelligence (AI) systems.

These best practices are most applicable to organizations deploying and operating externally developed AI systems on premises or in private cloud environments, especially those in high-threat, high-value environments. They are not applicable for organizations who are not deploying AI systems themselves and instead are leveraging AI systems deployed by others.

Not all of the guidelines will be directly applicable to all organizations or environments. The level of sophistication and the methods of attack will vary depending on the adversary targeting the AI system, so organizations should consider the guidance alongside their use cases and threat profile.

See Guidelines for secure AI system development for design and development aspects of AI systems. [1]

Introduction

The rapid adoption, deployment, and use of AI capabilities can make them highly valuable targets for malicious cyber actors. Actors, who have historically used data theft of sensitive information and intellectual property to advance their interests, may seek to co-opt deployed AI systems and apply them to malicious ends.

Malicious actors targeting AI systems may use attack vectors unique to AI systems, as well as standard techniques used against traditional IT. Due to the large variety of attack vectors, defenses need to be diverse and comprehensive. Advanced malicious actors often combine multiple vectors to execute operations that are more complex. Such combinations can more effectively penetrate layered defenses.

Organizations should consider the following best practices to secure the deployment environment, continuously protect the AI system, and securely operate and maintain the AI system.

The best practices below align with the cross-sector Cybersecurity Performance Goals (CPGs) developed by CISA and the National Institute of Standards and Technology (NIST). The CPGs provide a minimum set of practices and protections that CISA and NIST recommend all organizations implement. CISA and NIST based the CPGs on

existing cybersecurity frameworks and guidance to protect against the most common and impactful threats, tactics, techniques, and procedures. Visit CISA's Cross-Sector Cybersecurity Performance Goals for more information on the CPGs, including additional recommended baseline protections.

Secure the deployment environment

Organizations typically deploy AI systems within existing IT infrastructure. Before deployment, they should ensure that the IT environment applies sound security principles, such as robust governance, a well-designed architecture, and secure configurations. For example, ensure that the person responsible and accountable for AI system cybersecurity is the same person responsible and accountable for the organization's cybersecurity in general [CPG 1.B].

The security best practices and requirements for IT environments apply to AI systems, too. The following best practices are particularly important to apply to the AI systems and the IT environments the organization deploys them in.

Manage deployment environment governance

- If an organization outside of IT is deploying or operating the AI system, work with the IT service department to identify the deployment environment and confirm it meets the organization's IT standards.
    - Understand the organization's risk level and ensure that the AI system and its use is within the organization's risk tolerance overall and within the risk tolerance for the specific IT environment hosting the AI system. Assess and document applicable threats, potential impacts, and risk acceptance. [3], [4]
    - Identify the roles and responsibilities for each stakeholder along with how they are accountable for fulfilling them; identifying these stakeholders is especially important should the organization manage their IT environment separately from their AI system.
    - Identify the IT environment's security boundaries and how the AI system fits within them.
- Require the primary developer of the AI system to provide a threat model for their system.

    - The AI system deployment team should leverage the threat model as a guide to implement security best practices, assess potential threats, and plan mitigations. [5], [6]

- Consider deployment environment security requirements when developing contracts for AI system products or services.
- Promote a collaborative culture for all parties involved, including the data science, infrastructure, and cybersecurity teams in particular, to allow for teams to voice any risks or concerns and for the organization to address them appropriately.

Ensure a robust deployment environment architecture

- Establish security protections for the boundaries between the IT environment and the AI system [CPG 2.F].
- Identify and address blind spots in boundary protections and other security-relevant areas in the AI system the threat model identifies. For example, ensure the use of an access control system for the AI model weights and limit access to a set of privileged users with two-person control (TPC) and two-person integrity (TPI) [CPG 2.E].
- Identify and protect all proprietary data sources the organization will use in AI model training or fine-tuning. Examine the list of data sources, when available, for models trained by others. Maintaining a catalog of trusted and valid data sources will help protect against potential data poisoning or backdoor attacks. For data acquired from third parties, consider contractual or service level agreement (SLA) stipulations as recommended by CPG 1.G and CPG 1.H.
- Apply secure by design principles and Zero Trust (ZT) frameworks to the architecture to manage risks to and from the AI system. [7], [8], [9]

Harden deployment environment configurations

- Apply existing security best practices to the deployment environment. This includes sandboxing the environment running ML models within hardened containers or virtual machines (VMs) [CPG 2.E], monitoring the network [CPG 2.T], configuring firewalls with allow lists [CPG 2.F], and other best practices, such as those in NSA's Top Ten Cloud Mitigation Strategies for cloud deployments.
- Review hardware vendor guidance and notifications (e.g., for GPUs, CPUs, memory) and apply software patches and updates to minimize the risk of exploitation of vulnerabilities, preferably via the Common Security Advisory Framework (CSAF). [10]

- Secure sensitive AI information (e.g., AI model weights, outputs, and logs) by encrypting the data at rest, and store encryption keys in a hardware security module (HSM) for later on-demand decryption [CPG 2.L].
- Implement strong authentication mechanisms, access controls, and secure communication protocols, such as by using the latest version of Transport Layer Security (TLS) to encrypt data in transit [CPG 2.K].
- Ensure the use of phishing-resistant multifactor authentication (MFA) for access to information and services. [2] Monitor for and respond to fraudulent authentication attempts [CPG 2.H]. [11]
- Understand and mitigate how malicious actors exploit weak security controls by following the mitigations in Weak Security Controls and Practices Routinely Exploited for Initial Access.

Protect deployment networks from threats

Adopt a ZT mindset, which assumes a breach is inevitable or has already occurred. Implement detection and response capabilities, enabling quick identification and containment of compromises. [8], [9]

- Use well-tested, high-performing cybersecurity solutions to identify attempts to gain unauthorized access efficiently and enhance the speed and accuracy of incident assessments [CPG 2.G].
- Integrate an incident detection system to help prioritize incidents [CPG 3.A]. Also integrate a means to immediately block access by users suspected of being malicious or to disconnect all inbound connections to the AI models and systems in case of a major incident when a quick response is warranted.

Continuously protect the AI system

Models are software, and, like all other software, may have vulnerabilities, other weaknesses, or malicious code or properties.

Validate the AI system before and during use

- Use cryptographic methods, digital signatures, and checksums to confirm each artifact's origin and integrity (e.g., encrypt safetensors to protect their integrity and confidentiality), protecting sensitive information from unauthorized access during AI processes. [14]

- Create hashes and encrypted copies of each release of the AI model and system for archival in a tamper-proof location, storing the hash values and/or encryption keys inside a secure vault or HSM to prevent access to both the encryption keys and the encrypted data and model at the same location. [1]
- Store all forms of code (e.g., source code, executable code, infrastructure as code) and artifacts (e.g., models, parameters, configurations, data, tests) in a version control system with proper access controls to ensure only validated code is used and any changes are tracked. [1]
- Thoroughly test the AI model for robustness, accuracy, and potential vulnerabilities after modification. Apply techniques, such as adversarial testing, to evaluate the model's resilience against compromise attempts. [4]
- Prepare for automated rollbacks and use advanced deployments with a human-in-the-loop as a failsafe to boost reliability, efficiency, and enable continuous delivery for AI systems. In the context of an AI system, rollback capabilities ensure that if a new model or update introduces problems or if the AI system is compromised, the organization can quickly revert to the last known good state to minimize the impact on users.
- Evaluate and secure the supply chain for any external AI models and data, making sure they adhere to organizational standards and risk management policies, and preferring ones developed according to secure by design principles. Make sure that the risks are understood and accepted for parts of the supply chain that cannot adhere to organizational standards and policies. [1], [7]
- Do not run models right away in the enterprise environment. Carefully inspect models, especially imported pre-trained models, inside a secure development zone prior to considering them for tuning, training, and deployment. Use organization-approved AI-specific scanners, if and when available, for the detection of potential malicious code to assure model validity before deployment.
- Consider automating detection, analysis, and response capabilities, making IT and security teams more efficient by giving them insights that enable quick and targeted reactions to potential cyber incidents. Perform continuous scans of AI models and their hosting IT environments to identify possible tampering.
    - When considering whether to use other AI capabilities to make automation more efficient, carefully weigh the risks and benefits, and ensure there is a human-in-the-loop where needed.

Secure exposed APIs

- If the AI system exposes application programming interfaces (APIs), secure them by implementing authentication and authorization mechanisms for API access. Use secure protocols, such as HTTPS with encryption and authentication [CPG 2.C, 2.D, 2.G, 2.H]. [1]
- Implement validation and sanitization protocols for all input data to reduce the risk of undesired, suspicious, incompatible, or malicious input being passed to the AI system (e.g., prompt injection attacks). [1]

Actively monitor model behavior

- Collect logs to cover inputs, outputs, intermediate states, and errors; automate alerts and triggers [CPG 2.T].
- Monitor the model's architecture and configuration settings for any unauthorized changes or unexpected modifications that might compromise the model's performance or security. [1]
- Monitor for attempts to access or elicit data from the AI model or aggregate inference responses. [1]

Protect model weights

- Harden interfaces for accessing model weights to increase the effort it would take for an adversary to exfiltrate the weights. For example, ensure APIs return only the minimal data required for the task to inhibit model inversion.
- Implement hardware protections for model weight storage as feasible. For example, disable hardware communication capabilities that are not needed and protect against emanation or side channel techniques.
- Aggressively isolate weight storage. For example, store model weights in a protected storage vault, in a highly restricted zone (HRZ) (i.e., a separate dedicated enclave), or using an HSM [CPG 2.L]. [12]

Secure AI operation and maintenance

Follow organization-approved IT processes and procedures to deploy the AI system in an approved manner, ensuring the following controls are implemented.

Enforce strict access controls

- Prevent unauthorized access or tampering with the AI model. Apply role-based access controls (RBAC), or preferably attribute-based access controls (ABAC) where feasible, to limit access to authorized personnel only.
    - Distinguish between users and administrators. Require MFA and privileged access workstations (PAWs) for administrative access [CPG 2.H].

Ensure user awareness and training

Educate users, administrators, and developers about security best practices, such as strong password management, phishing prevention, and secure data handling. Promote a security-aware culture to minimize the risk of human error. If possible, use a credential management system to limit, manage, and monitor credential use to minimize risks further [CPG 2.I].

Conduct audits and penetration testing

- Engage external security experts to conduct audits and penetration testing on ready-to-deploy AI systems. This helps identify vulnerabilities and weaknesses that may have been overlooked internally. [13], [15]

Implement robust logging and monitoring

- Monitor the system's behavior, inputs, and outputs with robust monitoring and logging mechanisms to detect any abnormal behavior or potential security incidents [CPG 3.A]. [16] Watch for data drift or high frequency or repetitive inputs (as these could be signs of model compromise or automated compromise attempts). [17]
- Establish alert systems to notify administrators of potential oracle-style adversarial compromise attempts, security breaches, or anomalies. Timely detection and response to cyber incidents are critical in safeguarding AI systems. [18]

Update and patch regularly

- When updating the model to a new/different version, run a full evaluation to ensure that accuracy, performance, and security tests are within acceptable limits before redeploying.

Prepare for high availability (HA) and disaster recovery (DR)

- Use an immutable backup storage system, depending on the requirements of the system, to ensure that every object, especially log data, is immutable and cannot be changed [CPG 2.R]. [2]

Plan secure delete capabilities

- Perform autonomous and irretrievable deletion of components, such as training and validation models or cryptographic keys, without any retention or remnants at the completion of any process where data and models are exposed or accessible. [19]

Conclusion

The authoring agencies advise organizations deploying AI systems to implement robust security measures capable of both preventing theft of sensitive data and mitigating misuse of AI systems. For example, model weights, the learnable parameters of a deep neural network, are a particularly critical component to protect. They uniquely represent the result of many costly and challenging prerequisites for training advanced AI models, including significant compute resources; collected, processed, and potentially sensitive training data; and algorithmic optimizations.

AI systems are software systems. As such, deploying organizations should prefer systems that are secure by design, where the designer and developer of the AI system takes an active interest in the positive security outcomes for the system once in operation. [7]

Although comprehensive implementation of security measures for all relevant attack vectors is necessary to avoid significant security gaps, and best practices will change as the AI field and techniques evolve, the following summarizes some particularly important measures:

- Conduct ongoing compromise assessments on all devices where privileged access is used or critical services are performed.
- Harden and update the IT deployment environment.
- Review the source of AI models and supply chain security.
- Validate the AI system before deployment.
- Enforce strict access controls and API security for the AI system, employing the concepts of least privilege and defense-in-depth.

- Use robust logging, monitoring, and user and entity behavior analytics (UEBA) to identify insider threats and other malicious activities.
- Limit and protect access to the model weights, as they are the essence of the AI system.
- Maintain awareness of current and emerging threats, especially in the rapidly evolving AI field, and ensure the organization's AI systems are hardened to avoid security gaps and vulnerabilities.

In the end, securing an AI system involves an ongoing process of identifying risks, implementing appropriate mitigations, and monitoring for issues. By taking the steps outlined in this report to secure the deployment and operation of AI systems, an organization can significantly reduce the risks involved. These steps help protect the organization's intellectual property, models, and data from theft or misuse. Implementing good security practices from the start will set the organization on the right path for deploying AI systems successfully.

Works cited

[1] National Cyber Security Centre et al. Guidelines for secure AI system development. 2023. https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf

[2] Australian Signals Directorate et al. Engaging with Artificial Intelligence (AI). 2024. https://www.cyber.gov.au/sites/default/files/2024-01/Engaging%20with%20Artificial%20Intelligence%20%28AI%29.pdf

[3] MITRE. ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) Matrix version 4.0.0. 2024. https://atlas.mitre.org/matrices/ATLAS

[4] National Institute of Standards and Technology. AI Risk Management Framework 1.0. 2023. https://www.nist.gov/itl/ai-risk-management-framework

[5] The Open Worldwide Application Security Project (OWASP®). LLM AI Cybersecurity & Governance Checklist. 2024. https://owasp.org/www-project-top-10-for-large-language-model-applications/llm-top-10-governance-doc/LLM_AI_Security_and_Governance_Checklist-v1.pdf

[6] The Open Worldwide Application Security Project (OWASP®). OWASP Machine Learning Security Top Ten Security Risks. 2023. https://owasp.org/www-project-machine-learning-security-top-10/

[7] Cybersecurity and Infrastructure Security Agency. Secure by Design. 2023. https://www.cisa.gov/securebydesign

[8] National Security Agency. Embracing a Zero Trust Security Model. 2021. https://media.defense.gov/2021/Feb/25/2002588479/-1/-1/0/CSI_EMBRACING_ZT_SECURITY_MODEL_UOO115131-21.PDF

[9] Cybersecurity and Infrastructure Security Agency. Zero Trust Maturity Model. 2022. https://www.cisa.gov/zero-trust-maturity-model

[10] Cybersecurity and Infrastructure Security Agency. Transforming the Vulnerability Management Landscape. 2022. https://www.cisa.gov/news-events/news/transforming-vulnerability-management-landscape

[11] Cybersecurity and Infrastructure Security Agency. Implementing Phishing-Resistant MFA. 2022. https://www.cisa.gov/sites/default/files/publications/fact-sheet-implementing-phishing-resistant-mfa-508c.pdf
[12] Canadian Centre for Cyber Security. Baseline security requirements for network security zones Ver. 2.0 (ITSP.80.022). 2021. https://www.cyber.gc.ca/en/guidance/baseline-security-requirements-network-security-zones-version-20-itsp80022
[13] Ji, Jessica. What Does AI Red-Teaming Actually Mean? 2023. https://cset.georgetown.edu/article/what-does-ai-red-teaming-actually-mean/
[14] Hugging Face GitHub. Safetensors. 2024. https://github.com/huggingface/safetensors.
[15] Michael Feffer, Anusha Sinha, Zachary C. Lipton, Hoda Heidari. Red-Teaming for Generative AI: Silver Bullet or Security Theater? 2024. https://arxiv.org/abs/2401.15897
[16] Google. Google's Secure AI Framework (SAIF). 2023. https://safety.google/cybersecurity-advancements/saif/
[17] Government Accountability Office (GAO). Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities. 2021. https://www.gao.gov/assets/gao-21-519sp.pdf
[18] RiskInsight. Attacking AI? A real-life example!. 2023. https://riskinsight-wavestone.com/en/2023/06/attacking-ai-a-real-life-example
[19] National Cyber Security Centre. Principles for the security of machine learning. 2022. https://www.ncsc.gov.uk/files/Principles-for-the-security-of-machine-learning.pdf

Disclaimer of endorsement

The information and opinions contained in this document are provided "as is" and without any warranties or guarantees. Reference herein to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement, recommendation, or favoring by the United States Government, and this guidance shall not be used for advertising or product endorsement purposes.

Purpose

This document was developed in furtherance of the authoring organizations' cybersecurity missions, including their responsibilities to identify and disseminate threats, and to develop and issue cybersecurity specifications and mitigations. This information may be shared broadly to reach all appropriate stakeholders.

Contact

U.S. organizations:

NSA Cybersecurity Report Feedback: CybersecurityReports@nsa.gov
NSA General Cybersecurity Inquiries or Customer Requests: Cybersecurity_Requests@nsa.gov
Defense Industrial Base Inquiries and Cybersecurity Services: DIB_Defense@cyber.nsa.gov
NSA Media Inquiries / Press Desk: 443-634-0721, MediaRelations@nsa.gov
Report incidents and anomalous activity to CISA 24/7 Operations Center at report@cisa.gov or (888) 282-0870 and/or to the FBI via your local FBI field office.

Australian organizations: For more information or to report a cybersecurity incident, visit cyber.gov.au or call 1300 292 371 (1300 CYBER1).

Canadian organizations: For more information contact the Cyber Centre at contact@cyber.gc.ca or report a cyber security incident to our portal at https://www.cyber.gc.ca/en/incident-management.

New Zealand organizations: Report cyber security incidents to incidents@ncsc.govt.nz or call 04 498 7654.

United Kingdom organizations: Report a significant cyber security incident at ncsc.gov.uk/report-an-incident (monitored 24 hours) or, for urgent assistance, call 03000 200 973.