PROTECT AI

To:	The White House Office of Science & Technology Policy
Re:	Response to RFI on the Development of an Artificial Intelligence (AI) Action Plan
From: :	Protect AI, Inc.
Contacts:	Jessica Souder, Director of Government & Defense; Charlie McCarthy, MLSecOps Director
Email:	[redacted]
Date:	March 15, 2025

Statement: This document is approved for public dissemination. The document contains no business-proprietary or confidential information. Document contents may be reused by the government in developing the AI Action Plan and associated documents without attribution.

Protect AI: Securing U.S. Artificial Intelligence

AI and ML systems are integral to innovation leadership in the United States, and, like any other software, they are vulnerable to security threats and governance challenges. Traditional software security measures cannot fully address these unique security threats, which include model tampering, data leakage, adversarial prompt injection, and supply chain attacks. For this reason, security and risk management must be part of the U.S. strategy for AI/ML to encourage broad adoption. To support that effort and in response to the White House Office of Science and Technology Policy (OSTP), the Protect AI Team offers recommendations for consideration.

Challenge/Issue 1, Supply Chain Protection and Counter-Sabotage:
The AI supply chain is vulnerable to sabotage, where malicious actors may introduce compromised components, leading to widespread security breaches.

Examples:

- In 2024, a North Korean hacking group, known as "Gleaming Pisces," was linked to a campaign involving poisoned Python packages on PyPI. These packages could execute malicious code when installed, targeting developers working on Linux and macOS systems. Although not specifically targeting AI models, this attack demonstrates how malicious actors can compromise software components used in AI development.[1]
- In 2023, over 1,600 leaked tokens exposed access to accounts of major organizations, including Google and Microsoft, on the Hugging Face platform. This incident highlights the risk of compromised third-party AI models and datasets.[2]

[1] https://www.bankinfosecurity.com/north-korea-targets-software-supply-chain-via-pypi-a-26344
[2] https://www.reversinglabs.com/blog/supply-chain-attacks-you-can-learn-from

Policy Vision:
Develop robust security protocols to safeguard the AI supply chain, ensuring that all components are secure and trustworthy.

Recommendations:

- Supply Chain and Machine Learning Development Operations: AI model scanning should be integrated into software supply chain security workflows, just like traditional code security scans.
- Supply Chain Transparency: Enhance visibility into the AI supply chain to quickly identify and address potential threats.
- Secure Development Practices: Promote secure coding and development practices among suppliers to reduce vulnerabilities.

Impact: Strengthening supply chain security will prevent the introduction of compromised components, safeguarding AI systems from potential sabotage.

Challenge/Issue 2, Financial Critical Infrastructure Attacks, Preventing AI-Driven Fraud and Market Manipulation:
AI systems in the financial sector are prime targets for cyberattacks, which can lead to significant financial losses and destabilize economic infrastructures.

Example: A 2023 RAND Corporation study highlighted the catastrophic consequences that adversarial attacks could have on algorithmic trading systems. They emphasize the need for proactive measures to ensure the security and integrity of financial systems reliant on machine learning models.[3]

Policy Vision: Implement stringent security measures to protect AI applications within financial institutions, ensuring the stability and integrity of financial systems. AI-driven fraud detection models should be protected from adversarial evasion attacks, where malicious actors manipulate AI decision thresholds to bypass financial security measures.

Recommendations:

- Enhanced threat detection by deploying advanced threat detection systems to identify and mitigate potential attacks on financial AI systems. This should include input sanitization and anomaly detection to protect against spoofing and evasion attacks.
- Implementation of continuous monitoring for AI-driven financial systems to detect and mitigate fraudulent patterns or adversarial inputs that could impact loan approvals, transaction security, or trading algorithms.
- AI red teaming protocols that enable continuous adversarial testing and model vulnerability assessments to identify and mitigate risks posed by adversarial attacks, bias exploitation, and algorithmic manipulation.

[3] https://www.rand.org/pubs/commentary/2023/08/money-markets-and-machine-learning-unpacking-the-risks.html

Impact: Implementing these measures will protect financial institutions from cyber threats, maintaining trust and stability in the financial system. This implementation would also strengthen financial AI security by ensuring trading models, fraud detection algorithms, and anti-money laundering systems are resistant to adversarial financial manipulation and AI misclassification.

Challenge/Issue 3, Malicious Zero-Day Attacks in AI Models:
AI models are susceptible to zero-day vulnerabilities in their dependencies and serialization attacks in their files that can lead to compromised models and system breaches such as ransomware.

Examples:

- In December 2022, the nightly version of the PyTorch machine learning framework installed via package installer for python (pip) included a dependency named torchtriton that had been compromised to execute a binary designed to exfiltrate system information from affected machines.[4]
- Security researchers, from Protect AI and other organizations, have identified thousands of machine learning models containing serialization or architectural backdoor attacks on the Hugging Face model-sharing platform. Many of these models could execute arbitrary code upon loading
  .

Policy Vision:
Establish proactive security measures to scan and protect AI model repositories, ensuring the integrity and trustworthiness of AI systems.

Recommendations:

- Regular Security Audits: Conduct frequent security assessments of AI models to identify and address vulnerabilities.
- Zero-Trust Model Deployment: Adopt a zero-trust approach to model deployment, ensuring that all models undergo rigorous security checks before integration.
- Automated Threat Detection: Implement tools to continuously scan for and detect potential threats within model repositories.

Impact:
Proactively securing model repositories will mitigate the risk of zero-day attacks, maintaining the integrity of AI systems and protecting organizational assets.

Challenge/Issue 4, Protecting Mission Critical Decisions from Compromised AI (Enabling the Department of Defense and Intelligence Community):
AI is increasingly being integrated into battle management, surveillance, intelligence analysis, and autonomous defense systems, supporting human operators in time-sensitive, high-risk decision-making. However, adversarial actors can manipulate or poison AI systems, leading to misclassifications or erroneous recommendations. Without robust security, AI-driven military

[4] https://www.sentinelone.com/blog/pytorch-dependency-torchtriton-supply-chain-attack/

tools could provide false intelligence, misidentify targets, or be hijacked to produce adversarial outcomes, putting warfighters and national security at risk.

Examples:

- Adversarial Attacks on AI in Defense: Vision-based machine learning models are known to be easily fooled by simple adversarial perturbations, causing AI to misidentify objects. This is inherently dangerous in defense applications like military drones.[5]
- LLM Manipulation in Military Intelligence: Large language models (LLMs) used in defense analysis and battlefield decision-support can be exploited using data poisoning and direct and indirect prompt injection attacks, causing AI to generate misleading intelligence assessments, impacting real-time operations.[6]

Policy Vision: The Department of Defense must establish secure, observable, and continuously tested AI systems to ensure that adversarial attacks, model vulnerabilities, and manipulated data inputs do not mislead human operators in mission-critical scenarios. AI tools must be constantly monitored, red-teamed for vulnerabilities, and secured through rigorous validation processes, ensuring that the human in the loop can trust AI-driven recommendations without risk of deception or compromise.

Recommendations:

- Conduct Continuous AI Red Teaming and Adversarial Testing: AI systems used in battlefield intelligence, autonomous targeting, and decision-support applications should undergo regular adversarial testing before deployment. Automated AI penetration testing should be deployed to evaluate vulnerabilities across the AI model lifecycle before adversaries exploit them.
- Enforce Strict AI Model Validation and Zero-Trust Principles: All AI models integrated into defense and command systems should undergo strict validation and model scanning to ensure they are free from tampering, backdoors, or security vulnerabilities. AI model access should be governed by policy enforcement mechanisms, ensuring only trusted models are used, with automatic blocking of AI models that fail security assessments.
- Implement Real-Time AI Observability, Runtime Security, and Threat Monitoring: AI-driven decision-making systems should be continuously monitored for anomalies. Security teams should have full observability into AI decision pathways, allowing real-time forensic analysis in case of unexpected or adversarial AI behavior.

Impact: Enhanced Trust in AI-Assisted Battlefield Decision-Making enabled by AI-driven intelligence assessments and targeting recommendations remain secure, explainable, and resistant to manipulation. Proactive Defense Against AI Manipulation – Continuous AI security

---

[5] https://totalmilitaryinsight.com/cyber-security-for-military-drones/, also Z., & Qiu, S. (2024). Efficient ensemble adversarial attack for a deep neural network (DNN)-based unmanned aerial vehicle (UAV) vision system. *Drones, 8*(10), 591. https://doi.org/10.3390/drones8100591

[6]Example from recent press: https://irregularwarfarecenter.org/publications/perspectives/the-newest-weapon-in-irregular-warfare-artificial-intelligence/

assessments prevent adversaries from injecting false data, exploiting model vulnerabilities, or hijacking AI-powered systems. Full AI Observability for Human Oversight – Military operators maintain full transparency and control over AI-driven processes, ensuring that final decision-making remains securely in human hands.

Challenge/Issue 5: Right of Boom Forensics to Strengthen AI Incident Response
Post-incident analysis in AI systems, known as "right of boom" forensics, provide critical threat intelligence into how our adversaries operate. Traditional forensic methods often fall short due to AI's complexity and the dynamic nature of machine learning models.

Example:

- In August 2024, Tenable researchers discovered critical privilege escalation vulnerabilities in Microsoft's Azure Health Bot Service, a cloud platform enabling healthcare organizations to deploy AI-powered virtual assistants. These vulnerabilities allowed attackers to access cross-tenant resources without proper authorization.[7]
- Also in 2023 researchers demonstrated that AI models could be manipulated via "data poisoning" at the training phase, making it difficult to determine when or where a compromise occurred. A forensically sound logging system could identify the source of contamination, allowing for immediate remediation.

Policy Vision: Establish comprehensive forensic capabilities tailored for AI environments to ensure rapid identification and mitigation of security incidents, thereby enhancing the resilience of AI infrastructures.

Recommendations:

- Implement AI-Specific Forensic Tools: Develop and deploy forensic tools designed for AI systems to effectively trace and analyze security incidents within machine learning environments.
- Continuous Monitoring: AI-specific monitoring tools should be utilized to provide real-time detection of anomalies and misalignments with established policies. Such tools should be capable of providing forensic data to facilitate immediate investigations in case of a violation.
- Automated scanning of AI model repositories should include static and dynamic security analysis to detect previously unknown vulnerabilities before models are used in mission-critical applications.
- AI models should be cryptographically signed to establish model provenance and ownership and to detect tampering.
- AI forensics tools should include immutable logging to ensure that all model interactions, decision-making pathways, and data retrievals are fully traceable after an incident.

[7] https://www.tenable.com/blog/compromising-microsofts-ai-healthcare-chatbot-service

Policy Impact: Enhanced forensic capabilities will identify misalignments and policy breaches, minimizing potential damages and strengthening the overall security posture of AI systems. Applying security tools such as scanning and observability tools will ensure AI models deployed in national security, finance, and healthcare environments are free from backdoors, trojans, and zero-day vulnerabilities and strengthen trust in AI repositories by enforcing rigorous security policies before AI models reach production use.

Conclusions:

By implementing continuous AI red teaming, zero-trust model validation and scanning, and real-time AI observability, the United States can encourage the safe, trusted adoption of artificial intelligence and machine learning to innovate faster and compete internationally. By addressing the challenges described in our submission with the recommended strategies, the U.S. government can encourage organizations to significantly enhance the security and reliability of their AI systems, mitigating risk, adding trust, and furthering proactive AI adoption.

We look forward to supporting the Office of Science and Technology Policy in your efforts to shape the White House's AI Action Plan and are happy to answer questions, provide additional information, or engage in person to support your efforts.

Thank you for the opportunity to contribute.

Sincerely,

The Protect AI Team

Protect AI, Inc.
1201 2nd Ave
Seattle, WA 98101

About Protect AI:

Founded in 2022 by industry veterans, Protect AI is on a mission to close the gap between AI innovation and security. With $107M in funding and strong investor backing, we're pioneering AI/ML security solutions that mitigate risks in AI deployment. Our capabilities focus on AI supply chain security, AI red-teaming, and AI detection and response at runtime.

We are humbled to have been recognized as a SINET16 Innovator, an Inc. Best Workplace, and a CB Insights Top 100 AI Startup. It is our goal to continue to evolve and grow the industry which is why we developed a freely accessible Machine Learning Security Operations (MLSecOps) curriculum that was recognized at RSA in 2024. Protect AI endeavors to keep learning and innovating and we look forward to contributing to the AI Action Plan and supporting the advancement of AI innovation for the U.S.