Openpolicy

March 14, 2025

To: The National Institute of Standards and Technology (NIST),
via CyberAIProfile@nist.gov

OpenPolicy's comment on
NIST Cybersecurity and AI Workshop Concept Paper

OpenPolicy appreciates the opportunity to provide feedback on NIST's Cybersecurity and AI Workshop Concept Paper and the broader policy discussions surrounding the implementation of AI security initiatives. As a technology company dedicated to democratizing the ability of innovative companies—of all sizes—to engage with policymakers and provide meaningful feedback on relevant policy deliverables, OpenPolicy is deeply committed to using AI, innovation, and technology to foster open, collaborative policymaking. We strive to streamline compliance and automate governance to support the evolving cybersecurity landscape.

Our engagement with NIST is part of a broader effort that includes active participation in initiatives such as White House Executive Orders, and collaboration with bodies like the NIST AI Safety Institute Consortium. We believe that the open and collaborative nature of the policymaking dialogue is essential to implementing secure, adaptive, and forward-thinking frameworks for managing AI risks. The participation of innovative companies and startups developing cutting-edge AI security and safety solutions is critical to this effort, as these communities are at the forefront of developing the technologies necessary for effective AI risk management.

These comments are submitted on behalf of a coalition of cybersecurity companies – including Astrix Security, Zenity, Wiz, Kiteworks, Cranium AI, ActiveFence, ADAMnetworks, Armis, Legit Security, Lasso Security, and Vaultree – each contributing unique insights on integrating AI into cybersecurity. Our collective feedback addresses key strategic themes and recommendations for NIST's consideration. We strongly support NIST's efforts to develop an integrated AI-cybersecurity framework—a "Cyber AI Profile"—that unifies the NIST Cybersecurity Framework (CSF) 2.0, the AI Risk Management Framework (AI RMF), and the Privacy Framework.

OpenPolicy and our coalition of innovative cybersecurity companies are eager to work with NIST and other implementing agencies to refine these recommendations and to develop a comprehensive Cyber AI Profile. This profile will not only address the holistic risk presented by AI throughout its lifecycle but will also extend to the broader, interconnected threat landscape that includes data, software, IoT, and emerging quantum risks. We look forward to

collaborating with NIST to integrate advanced solutions for secure communication, continuous risk management, real-time monitoring, and centralized governance—ensuring that our digital infrastructure remains resilient, adaptive, and secure.

Securing the SaaS Supply Chain and AI Ecosystem

With the rapid growth of AI technologies and interconnected SaaS ecosystems, organizations face increasing risks from unmonitored third-party integrations, unmanaged machine identities, and AI-enabled attack techniques. As NIST highlights in its Concept Paper, addressing these emerging risks requires extending traditional cybersecurity frameworks to account for the unique vulnerabilities introduced by AI. OpenPolicy strongly supports NIST's efforts to integrate AI-specific cybersecurity risks into the CSF 2.0, AI RMF, and Privacy Framework and recommends NIST expand its guidance on SaaS integration security, promoting dynamic identity verification for non-human entities and encouraging automated threat detection and remediation by providing practical and actionable recommendations that help organizations embrace AI securely.

AI ecosystems increasingly rely on interconnected SaaS applications, APIs, and data pipelines, which collectively create an expansive and dynamic attack surface. Unmanaged third-party integrations and "shadow AI" applications can introduce significant security risks, often without organizations' full visibility or control. As SaaS ecosystems grow more complex, ensuring continuous monitoring and oversight of these integrations is critical to mitigating supply chain threats. Technology solutions that provide continuous discovery and monitoring of third-party integrations offer a vital defense against these risks. For example, platforms such as **Astrix**[1] continuously map OAuth tokens, API connections, and machine-to-machine links, giving security teams enhanced visibility into potential vulnerabilities. This proactive monitoring approach helps organizations identify risky integrations before adversaries can exploit them, improving overall security posture. By inventorying both traditional IT assets and AI-specific components — such as pre-trained models, data pipelines, and third-party AI services — these solutions align closely with CSF 2.0's Governance of Supply Chain Risk Management (**GV.SC-01 through GV.SC-10**) and Asset Management controls (**ID.AM-02 and ID.AM-04**), as well as the AI RMF's emphasis on maintaining a comprehensive catalog of AI systems and dependencies.

To strengthen NIST's Cyber AI Profile, we recommend expanding CSF guidance to explicitly encourage continuous monitoring of third-party applications and AI service providers. Given the evolving nature of AI ecosystems, persistent oversight is essential to detect

[1] See Astrix NHI solution https://astrix.security/

integration changes, model updates, or new dependencies that may inadvertently introduce vulnerabilities.

Recent high-profile incidents further underscore the urgency of this approach. For example, in early January this year, the U.S. Treasury was targeted in a sophisticated attack where compromised API keys—stolen from a third-party vendor (Beyond Trust)—were used to infiltrate internal systems.[2] This breach illustrates the significant risk posed by unmanaged non-human identities (NHIs) and third-party integrations. Similarly, the Midnight Blizzard campaign, linked to Russian state-sponsored actors, exploited OAuth applications as part of their attack against Microsoft's corporate environment[3]. In this case, Microsoft's Office 365 email server was compromised, exposing internal email correspondences of Microsoft employees. These incidents highlight how vulnerable SaaS integrations and non-human identities can become key entry points for attackers, particularly when security controls lack continuous monitoring and adaptive risk management.

Accordingly, to strengthen the Cyber AI Profile's approach to supply chain security, we recommend that NIST expand CSF guidance to explicitly require continuous monitoring of all third-party applications and AI service providers. Continuous monitoring is crucial because AI supply chain risks are inherently dynamic; third-party vendors frequently update their models, modify data pipelines, and introduce new dependencies that can inadvertently expand the attack surface. Without persistent oversight, organizations risk losing visibility into changes that could introduce new vulnerabilities or expose sensitive data.

By encouraging organizations to adopt contractual assurances (aligned with **GV.SC-05 and GV.SC-07**) that require third-party AI vendors to disclose vulnerabilities, verify model provenance, and provide secure update mechanisms, enterprises can proactively mitigate these risks. Such contractual measures ensure vendors remain accountable for their security practices and empower organizations to respond rapidly when vulnerabilities or compromised AI components are identified. Given that AI supply chains often involve proprietary models and complex dependencies, contractual assurances serve as a vital mechanism to enforce transparency, promote secure development practices, and ensure that third-party AI services adhere to robust security standards. This proactive approach would allow organizations to better manage the evolving risks associated with SaaS ecosystems and strengthen their overall security posture.

[2] See Dark Reading article on "Chinese State Hackers Breach US Treasury Department" https://shorturl.at/1oI5W

[3] See Dark Reading article on "Microsoft Shares New Guidance in Wake of 'Midnight Blizzard' Cyberattack" https://shorturl.at/4POuU

Further, OpenPolicy supports NIST's emphasis on identity management and Zero Trust principles for AI ecosystems. AI workflows increasingly rely on non-human identities—including service accounts, API keys, and bots—which are often overlooked in traditional security models. These machine identities frequently operate with excessive, persistent privileges, making them prime targets for attackers seeking to exploit elevated permissions.

Today, the number of non-human identities (NHIs) already far exceeds that of human identities—a figure that is expected to rise further with the increasing adoption of AI agents. These AI agents leverage NHIs such as service accounts, API keys, and secrets to access and operate on organizations' most sensitive systems and data. Without proper security measures, these identities become prime targets for attackers, as seen in the U.S. Treasury incident and other prominent attacks.

Technology solutions, such as Astrix's platform, addresses this challenge by continuously tracking and authenticating API calls, dynamically revoking excessive privileges when suspicious behavior is detected. This aligns with CSF 2.0's Identity Management, Authentication, and Access Control outcomes (PR.AA-01, PR.AA-03, PR.AA-05) and the AI RMF's guidance on managing non-human actors and cataloging third-party dependencies. We, therefore, recommend that NIST's AI Profile expand CSF identity management controls to include explicit requirements for continuous, dynamic re-authorization of machine identities. This approach is essential given that machine identities — such as API keys, service accounts, and bots — often maintain persistent and elevated privileges. These characteristics make them prime targets for attackers, particularly since they can perform automated tasks continuously, increasing the risk exposure window. Without dynamic re-authorization, a compromised machine identity could enable adversaries to maintain undetected access to critical systems for extended periods.

Enforcing continuous re-authorization aligns with Zero Trust principles by ensuring that every API interaction is validated in real time, regardless of prior trust or behavior. This is especially crucial in environments where SaaS integrations, automated processes, and machine learning models operate with complex and dynamic interactions. Methods of binding OAuth tokens to designated owners reinforces accountability by ensuring that each machine identity has a responsible party, facilitate faster investigation and remediation if suspicious activity occurs. Incorporating dynamic risk scoring further enhances this strategy by allowing organizations to assess contextual factors—such as geographic location, behavioral patterns, or data access anomalies—and respond immediately by restricting or revoking access if risks are identified. Expanding CSF identity management controls in this way would significantly improve an organization's ability to manage the

security risks associated with machine identities while supporting the adoption of Zero Trust practices in AI ecosystems.

The rise of AI-enabled attacks further underscores the need for adaptive security controls. Adversaries are increasingly leveraging AI models to automate reconnaissance, manipulate system behavior, and generate realistic forgeries. These attacks are particularly concerning because they can closely mimic legitimate user behavior, making them difficult to detect with static or rule-based security systems. As AI models grow more autonomous and complex, traditional security controls may struggle to identify malicious activities designed to blend in with normal system behavior. By continuously comparing current activity against established baselines, technology solution can help detect behavioral anomalies that may indicate adversarial tactics. For instance, if an attacker manipulates an OAuth token or attempts to simulate a trusted API call, Astrix's anomaly detection system will recognize the deviation and initiate automated remediation. This adaptive approach is critical for countering AI-driven attacks, which often rely on subtle, calculated changes that evade traditional detection methods. Astrix's use of behavioral analytics ensures that deviations designed to appear legitimate are quickly identified and mitigated. This aligns with CSF's **DE.CM-03** guidance on anomaly detection and the **AI RMF's MP-5.1** series recommendations for adaptive AI system monitoring.

To further mitigate these risks, we recommend that NIST expand its third-party risk management guidance to emphasize dynamic re-authentication and rapid revocation capabilities. AI-enabled attacks can spread rapidly, particularly in interconnected SaaS environments where compromised machine identities may enable adversaries to pivot across multiple platforms. Technology solutions like Astrix's, which employ access provisioning and automated risk scoring, provide effective models for containing such threats. By dynamically adjusting access permissions and revoking compromised integrations in real time, these methods reduce the risk of lateral movement and help limit the impact of AI-driven attacks.

Addressing Emerging Risks AI Technologies built through Low Code/No Code Development Platforms

As organizations increasingly adopt AI technologies to improve business outcomes, streamline operations, and enhance security, new risks and vulnerabilities are emerging. While the Concept Paper highlights important themes such as securing AI system components, defending against AI-enabled attacks, and leveraging AI for cybersecurity defenses, we believe it is crucial that NIST expand this scope to explicitly address the security challenges associated with AI technologies built through Low-Code/No-Code (LCNC) development platforms and AI Agents specifically.

LCNC platforms are becoming increasingly popular across organizations seeking to accelerate digital transformation by empowering non-technical users to build functional applications. In parallel, the rapid adoption of AI Agents — autonomous systems designed to perform complex tasks with minimal human intervention — introduces additional security concerns. These AI systems operate independently, making decisions and interacting with their environments without continuous human intervention. AI Agents, often built by business users through intuitive drag-and-drop platforms, operate independently and are often integrated with sensitive corporate data and processes. While these innovations improve agility and efficiency, they also create significant security risks that traditional IT controls may fail to address.

AI Agents built through LCNC applications, by their nature, bypass traditional software development processes, resulting in limited oversight from security teams. These systems often operate autonomously, making decisions and accessing sensitive data much like trusted human users, and in fact, often as the human user by assuming their identity. Attackers are increasingly exploiting this vulnerability through techniques such as **prompt injection attacks** and **AI model manipulation**. Without proper monitoring mechanisms, these risks may go unnoticed, leaving organizations vulnerable to security breaches, data exposure, and ultimately data exfiltration. To address these challenges, OpenPolicy recommends that NIST expand the Cyber AI Profile to incorporate guidance on securing LCNC applications and AI Agents, particularly in areas such as asset discovery, risk assessment, privacy governance, and workforce training. Organizations should be required to maintain a **living inventory** of their AI and Agentic AI systems, ensuring that security teams have visibility into all deployed agents, their configurations, and the data they access. Continuous monitoring, detection, and response mechanisms must also be extended to address both the **build standards** and **runtime environments** of these autonomous systems.

Accordingly, continuous monitoring is essential for organizations seeking to manage the security risks associated with LCNC applications and AI Agents. Both LCNC workflows and AI Agents are dynamic and interdependent; business users frequently modify these applications after deployment to refine functionality, expand features, or connect to additional data sources. Without appropriate monitoring mechanisms, these changes can inadvertently introduce new risks, such as misconfigurations, excessive permissions, or unauthorized data exposure. Since LCNC applications and AI Agents are almost always created and modified outside the oversight of security teams, organizations may struggle to detect these vulnerabilities using traditional monitoring practices.

NIST should expand the guidance in **CSF Continuous Monitoring (DE.CM-09)** to explicitly reference LCNC environments and AI Agents. For example, **Zenity's**[4] continuous monitoring capabilities align with CSF's **Continuous Monitoring (DE.CM-09)**, which emphasizes monitoring computing hardware, software, runtime environments, and their data to identify potentially adverse events while proactively monitoring LCNC applications for signs of data leakage, unauthorized access, or security misconfigurations. Accordingly, expanding this guidance to include LCNC applications and AI Agents would encourage organizations to apply the same level of vigilance to these rapidly evolving workflows as they do to custom-built applications and traditional IT systems. Technology solutions that provide automated visibility and continuous oversight for LCNC applications and AI Agents can help organizations identify unexpected data flows, unauthorized access attempts, or security misconfigurations that may emerge post-deployment. For example, security tools can monitor the behavior of AI Agents in real time, alerting security teams to prompt injection attacks, phishing campaigns, or suspicious actions that may compromise sensitive data. Integrating this recommendation into the Cyber AI Profile would reinforce the need for proactive monitoring strategies that reflect the fast-changing nature of LCNC environments and the AI Agents built within them.

In addition to security concerns, LCNC applications and AI Agents introduce unique risks to data governance and privacy. Because LCNC platforms and AI Agent development tools enable non-technical users to build applications without a strong understanding of privacy requirements, there is a heightened risk that these systems may inadvertently violate data protection regulations such as GDPR or HIPAA. Without proper controls, LCNC applications and AI Agents may unintentionally expose sensitive data, transfer information to unauthorized third-party services, or overlook consent and processing limitations.

To mitigate these risks, NIST should expand the **NIST Privacy Framework's Inventory and Mapping (ID.IM-P)** and **Data Processing Ecosystem Risk Management (ID.DE-P)** categories to explicitly recommend automated privacy controls for LCNC applications and AI Agents. Automated solutions that map data flows, assess privacy risks, and enforce appropriate controls can help organizations maintain compliance and protect sensitive data. Expanding ID.IM-P to reflect the unique privacy risks in LCNC environments and AI Agents would provide organizations with actionable guidance for ensuring these applications meet established privacy requirements. Subcategories such as **ID.IM-P4** through **ID.IM-P7**, which emphasizes mapping data actions, identifying processing purposes, and assessing third-party involvement, is particularly relevant in this context. Similarly, expanding **ID.DE-P1** and **ID.DE-P2** would highlight the importance of assessing

[4]See Zenity solution https://www.zenity.io/.

privacy risks introduced by LCNC applications and AI Agents — especially when they rely on third-party services or external data sources.

Additionally, NIST could revise the **Data Processing Management (CT.DM-P)** category to emphasize the need for automated data review mechanisms that evaluate LCNC applications and AI Agents for improper data sharing or processing. Subcategories such as **CT.DM-P1** (reviewing data actions for privacy risks) and **CT.DM-P4** (data deletion and retention controls) align with the need to ensure that LCNC applications and AI Agents process data securely and in accordance with privacy policies. Expanding these subcategories to include LCNC-specific and AI-specific examples will ensure organizations apply automated data governance tools to monitor these emerging environments as part of their broader privacy program.

Beyond technical controls, organizations must also address the human factors that contribute to LCNC and AI Agent security risks. As these platforms expand, business users — often with limited cybersecurity knowledge — are playing an increasingly active role in software and automation development. This democratization of application development introduces new risks, as non-technical developers may inadvertently create insecure workflows, expose sensitive data, or misconfigure access controls. While automated security tools can reduce these risks, organizations must also invest in improving user awareness to ensure that citizen developers understand how to build secure and compliant applications.

To address this challenge, NIST should expand the **CSF Awareness and Training (PR.AT-02)** function to include tailored education for citizen developers who engage in LCNC application development and AI Agent creation. Providing specialized training for these users will help mitigate security risks by ensuring they are equipped with the knowledge necessary to design applications securely. Additionally, organizations should adopt security solutions that provide real-time guidance and remediation insights during development, helping non-technical users identify and address potential security issues as they arise. Expanding PR.AT-02 emphasizes this combined approach — targeted training paired with real-time security guidance — would help organizations better manage the risks associated with LCNC development and AI Agent deployment.

In conclusion, the rapid adoption of LCNC platforms and AI Agents necessitates a dedicated focus within the Cyber AI Profile. By expanding CSF's **Asset Management (ID.AM-02)**, **Platform Security (PR.PS-06)**, and **Continuous Monitoring (DE.CM-09)** controls to address LCNC and AI Agent risks, and by strengthening Privacy Framework guidance through **Inventory and Mapping (ID.IM-P)**, **Data Processing Ecosystem Risk Management (ID.DE-P)**, and **Data Processing Management (CT.DM-P)**, NIST can provide

comprehensive guidance for securing these emerging technologies. Encouraging the adoption of automated solutions for LCNC and AI Agent visibility, risk assessment, and governance will help organizations identify and mitigate risks while maintaining the flexibility and speed that these platforms offer. Integrating these recommendations into the Cyber AI Profile will ensure that NIST's guidance effectively addresses the evolving cybersecurity landscape.

Enhancing Data-Centric Security in the Cyber AI Profile

AI's rapid adoption introduces significant cybersecurity challenges, including data security risks, adversarial AI threats, and potential vulnerabilities in AI model pipelines. These risks are compounded by the growing volume of sensitive data being processed, stored, and transmitted by AI systems. Addressing these risks requires strengthening controls around sensitive data exchanges, applying Zero Trust principles to content flows, and enhancing threat detection using AI-driven capabilities. AI systems frequently interact with vast amounts of sensitive data, including intellectual property, personal information, and proprietary business data. This data is often exchanged across multiple environments, creating vulnerabilities that traditional network security controls may fail to address. Strengthening data-centric security controls is essential for mitigating these risks.

Technology companies such as **Kiteworks**[5] provide effective solutions that align with this need by employing a **Private Data Network (PDN)** that consolidates sensitive content communications into a secure, controlled environment. By implementing content firewalls that monitor and mediate data flows, these solutions ensure that no file or message enters or leaves an organization's environment without inspection and authorization. Such approaches apply Zero Trust principles directly to data flows, ensuring that content exchanges are verified, logged, and protected.

To better address data protection in AI systems, we recommend that NIST expand the Cyber AI Profile's guidance under CSF 2.0's **PR.DS (Data Security)** category to emphasize data-centric controls. Specifically, guidance aligned with **PR.DS-01 (Data-at-Rest Protection)** and **PR.DS-02 (Data-in-Transit Protection)** should highlight strategies such as robust encryption, data integrity monitoring, and controlled data access policies. Additionally, mapping data flows between AI systems and enterprise data repositories, aligned with **ID.AM-03 (Network Data Flow Mapping)**, can help organizations identify and control unauthorized data movement, an essential step in mitigating adversarial AI attacks and data leakage risks.

[5]See Kiteworks solution https://www.kiteworks.com/

While the NIST concept paper appropriately emphasizes Zero Trust principles for network and identity security, we encourage NIST to expand its guidance to ensure these principles are applied directly to the bridge between AI systems and enterprise data repositories, ensuring that only authorized entities can access sensitive data flows. AI systems frequently introduce new pathways for data movement, increasing the risk of unauthorized access or data manipulation. Establishing trust boundaries at the content level is crucial to securing these dynamic data flows. Solutions such as those provided by Kiteworks demonstrate how Zero Trust principles can effectively protect data. By implementing content-level access controls that verify each data exchange — whether through file sharing, email, MFT, or other communications — these solutions ensure that no access to data is trusted by default. Aligning this approach with **PR.AA-05 (Access Control Policies and Review)** and **DE.CM-03 (Personnel and Technology Usage Monitoring)** would help organizations enforce continuous verification policies for data flows.

Organizations should restrict data exchanges between enterprise repositories and AI systems by implementing robust content type filtering and trusted channel verification to prevent sensitive information from being inadvertently exposed or exfiltrated. Solutions such as the Kiteworks AI Data Gateway exemplify this approach by providing a secure bridge that enforces strict governance policies, applies zero-trust principles, and maintains comprehensive audit trails to ensure that only authorized data moves through approved channels while maintaining regulatory compliance with frameworks like GDPR, HIPAA, and various data privacy laws.

AI technologies present both security risks and opportunities for defense. While adversaries increasingly leverage AI to develop automated attack techniques, AI also offers powerful capabilities for improving threat detection and identifying suspicious data behaviors. Technology solutions that integrate AI-enhanced content analytics, such as those employed by Kiteworks, effectively detect potential threats in data flows. For example, Kiteworks applies advanced data monitoring tools that analyze content usage patterns to identify anomalous behaviors that may signal malicious intent or compromised data repositories. This aligns with CSF 2.0 outcomes such as **DE.AE-02 (Adverse Event Analysis)** and **DE.CM-01 (Network Monitoring for Adverse Events)**, which emphasizes proactive threat detection practices. We recommend that NIST expand the Cyber AI Profile's guidance to promote the adoption of AI-enhanced threat detection solutions that identify emerging risks in real-time to help organizations mitigate sophisticated data exfiltration tactics, AI-driven adversarial attacks, and misuse of AI models.

Further, given that AI systems often process regulated and personally identifiable information (PII), privacy considerations should be integrated into the Cyber AI Profile. Security solutions that enforce data sovereignty controls, restrict data movement to authorized regions, and apply encryption-in-use offer effective strategies for aligning AI security with privacy frameworks. Solutions that provide comprehensive audit trails, automated compliance reporting, and centralized data governance tools that support privacy-by-design practices would help organizations establish visibility into data handling practices while maintaining compliance with evolving regulatory requirements. Thus, NIST should encourage organizations to adopt similar capabilities by expanding guidance under **GV.OC-03 (Managing Legal, Regulatory, and Contractual Requirements)** and **GV.OV-03 (Evaluating and Reviewing Risk Management Performance)** to help mitigate these risks.

Addressing AI-Driven Social Engineering and Content-Based Threats

As NIST's concept paper outlines, AI systems introduce new risks, including model manipulation, adversarial exploitation, and data poisoning. However, the concept paper underestimates the growing role of AI-generated content threats that exploit social engineering tactics. Increasingly, attackers are leveraging AI to craft sophisticated phishing campaigns, deepfake-based impersonation attacks, and automated disinformation campaigns. These threats can have substantial cybersecurity impacts, particularly when manipulated content is used to deceive employees or influence decision-making processes.

To address this gap, we recommend that NIST expand the CSF's **"Awareness and Training" (PR.AT-01)** and **(PR.AT-02)** categories to explicitly include training on AI-generated manipulation tactics. Organizations should educate their staff about recognizing AI-authored phishing attempts, deepfake impersonation, and coordinated disinformation campaigns that may be precursors to broader security incidents.

In addition, NIST should encourage organizations to incorporate AI-driven content threat detection tools into their cybersecurity strategies. Solutions such as **ActiveFence's**[6] platform, which monitors social media, open web forums, and messaging platforms to identify emerging disinformation campaigns and malicious content, can provide essential early warnings. Integrating this type of threat intelligence aligns with CSF's **"Continuous Monitoring" (DE.CM-01)** and **(DE.CM-03)**, enhancing an organization's ability to detect malicious activity that exploits the "human perimeter" of security.

[6]See ActiveFence solution https://www.activefence.com/

AI has proven highly effective in processing vast quantities of unstructured data to detect emerging cyber threats. This is especially critical as threat actors increasingly discuss new attack techniques, vulnerabilities, and ransomware campaigns across dark web forums and public channels. AI-driven threat intelligence tools like those developed by ActiveFence offer a proactive advantage by identifying malicious activities before they escalate. We recommend that NIST expand CSF's **"Adverse Event Analysis" (DE.AE-02, DE.AE-03, and DE.AE-07)** to emphasize AI-powered threat intelligence solutions as a means of identifying emerging threat trends in real time. By continuously analyzing patterns of malicious behavior, organizations can better anticipate AI-enabled attacks and mitigate risks before they materialize.

Further, as AI systems increasingly become integrated into critical decision-making processes, they are vulnerable to adversarial tactics such as prompt injection, data poisoning, and model evasion. Red teaming methodology — which simulates adversarial tactics to expose model vulnerabilities — aligns with the principles outlined in **NIST AI 600-1** and reinforces the need for continuous stress-testing of AI systems. Red teaming exercises enable organizations to identify exploitable weaknesses and implement mitigation strategies before attackers can exploit those vulnerabilities. We recommend that NIST expand CSF's **"Continuous Monitoring" (DE.CM-09)** guidance to explicitly promote red teaming practices as part of an organization's proactive security posture. By incorporating red teaming as an essential security control, NIST can encourage organizations to adopt robust adversarial testing frameworks for AI models — particularly those that influence cybersecurity defenses, content moderation platforms, or decision-support systems.

Integrating trusted and secure AI practices into cybersecurity governance is critical to addressing AI-driven content threats that exploit online platforms and social networks. While traditional cybersecurity frameworks often focus on securing infrastructure, manipulative content campaigns present a growing risk to organizations — especially those operating in sectors vulnerable to brand impersonation, disinformation, or consumer manipulation. NIST should expand the CSF's **"Risk Assessment" (ID.RA-02)** and **"Incident Analysis" (RS.AN-03)** categories to include collaboration between cybersecurity teams and trust and secure specialists. By establishing communication channels between content security teams and security operations centers (SOCs), organizations can better identify and respond to AI-driven content manipulation campaigns that may evolve into broader security incidents. Incorporating these recommendations into the Cyber AI Profile will provide organizations with more comprehensive guidance to mitigate emerging AI-driven threats. By expanding guidance on social engineering risks, promoting AI-enhanced threat intelligence, advocating for robust red teaming practices, and integrating Trust & Safety

principles into governance frameworks, NIST can enhance the cybersecurity community's ability to manage AI risks effectively.

Securing AI Systems, Detecting AI-Driven Threats, and Strengthening AI Supply Chain Security

The growing adoption of AI technologies introduces distinct security challenges that require organizations to adopt comprehensive strategies for risk mitigation. As AI ecosystems expand in complexity— encompassing intricate data pipelines, evolving attack techniques, and increased reliance on third-party tools and services— organizations must adopt solutions that provide continuous visibility, adaptive security controls, and robust supply chain safeguards. Successfully addressing these risks demands a multifaceted approach that combines technical innovation, proactive threat detection, and strong governance practices.

NIST's Cyber AI Profile provides an important foundation for guiding organizations in mitigating these risks, yet there are opportunities to expand this guidance by strengthening requirements for asset inventory, technical verification, and continuous monitoring. Cloud security platforms such as **Wiz**[7] have demonstrated how integrated solutions can address these concerns by combining inventory mapping, behavioral threat detection, and supply chain risk management. These capabilities reflect the type of strategic approach that NIST should emphasize in the Cyber AI Profile to help organizations mitigate AI-specific risks.

A core challenge facing organizations is the complexity of securing AI system components. AI models are often developed and deployed across intricate environments that involve extensive data flows, diverse third-party dependencies, and continuous model updates. This complexity introduces vulnerabilities such as data poisoning, model manipulation, and prompt injection attacks that can undermine AI system integrity. To mitigate these risks, NIST should expand its guidance to emphasize continuous visibility and technical verification as essential components of AI system security.

Comprehensive inventory management is critical to this effort. NIST should encourage organizations to adopt commercially available platforms that continuously identify and map AI technologies within their environments. These platforms should provide technical visibility into AI system components, including models, application programming interfaces (APIs), training datasets, software development kits (SDKs), and deployment environments. Without this visibility, organizations risk overlooking vulnerabilities, failing to identify points of exposure, or losing awareness of how third-party AI components interact with their

[7]See Wiz solution https://www.wiz.io/

systems. By aligning this approach with CSF 2.0's **ID.AM-03 (Asset Management)**, NIST can provide organizations with clear guidance on tracking AI system dependencies and understanding their broader security implications.

Beyond inventory, technical verification is vital for ensuring that AI systems are assessed for vulnerabilities, weaknesses, and misconfigurations. Platforms that continuously evaluate AI system security can proactively identify points of exposure and prioritize remediation efforts based on exploitability and potential impact. This verification process should include assessing AI model behavior to detect anomalies, unexpected degradation, or adversarial interference, which may indicate manipulation or malicious influence. As AI systems increasingly rely on open-source tools and third-party services, NIST's guidance should also emphasize the need to assess these dependencies for potential security risks. Aligning this guidance with CSF 2.0's **PR.PS-01 (Configuration Management Practices)** would further ensure that AI system configurations are secure and align with best practices.

Building on these proactive measures, NIST should also expand its guidance on detecting and mitigating AI-driven cyberattacks. As adversarial AI techniques advance, attackers are increasingly able to manipulate AI models, evade traditional security measures, and automate complex attack campaigns. To address this challenge, NIST should expand its recommendations to include adaptive threat detection capabilities designed to recognize AI-specific attack tactics. Technical solutions such as **Cloud Detection and Response (CDR)** platform illustrate how advanced monitoring systems can combine contextual analysis, behavioral threat detection, and real-time monitoring to identify AI-driven threats. These capabilities can identify sophisticated attacks such as prompt injection, manipulated inference outputs, and compromised model updates.

To reinforce these security practices, NIST should integrate key controls from CSF 2.0 that focus on continuous monitoring and response. Incorporating **DE.CM-01 (Continuous Monitoring: Network and Network Services)** would help organizations maintain visibility across AI environments in real time, improving their ability to detect malicious activity before it escalates. Expanding **DE.AE-08 (Incident Declaration)** to include AI-specific threat scenarios would provide organizations with clear procedures for mitigating adversarial attacks that target AI models. Furthermore, updating **DE.AE-07 (Integration of Threat Intelligence)** to include AI-specific adversarial techniques would strengthen the framework by ensuring organizations proactively track and respond to emerging AI threats.

In addition to improved detection, NIST should enhance its guidance on securing the AI supply chain. AI models frequently rely on third-party datasets, pre-trained models, and external APIs, making supply chain vulnerabilities a significant concern. Malicious actors can exploit this reliance by introducing compromised code, poisoned training data, or

manipulated model updates. To mitigate these risks, NIST should encourage organizations to adopt automated **Software Bill of Materials (SBOM)** capabilities that provide comprehensive visibility into AI model dependencies. Solutions that combine SBOM generation with dependency tracking enable organizations to assess the integrity of third-party components and respond quickly to identified risks. Aligning this recommendation with CSF 2.0's **GV.SC-04 (Supply Chain Risk Management: Supplier Prioritization by Criticality)** and **GV.SC-07 (Supply Chain Risk Monitoring)** would help organizations adopt a structured and proactive approach to mitigating supply chain risks.

Finally, protecting the privacy and security of AI training data is essential. AI models are often built on large volumes of sensitive data, making them vulnerable to leakage, model inversion, or unauthorized access. To reduce these risks, NIST should expand the Cyber AI Profile to emphasize privacy-enhancing techniques such as data minimization, de-identification, and encryption. Platforms that incorporate **Data Security Posture Management (DSPM)** capabilities, such as those offered by cutting edge technology solutions, illustrate how organizations can monitor and secure sensitive AI training data across cloud environments. Aligning this approach with the NIST Privacy Framework's **Control-P (Data Processing Management)** would help organizations apply appropriate data protection strategies throughout the AI model lifecycle.

By strengthening its guidance in these key areas, NIST can provide organizations with practical strategies to manage AI risks while aligning with established cybersecurity and privacy frameworks. Expanding the Cyber AI Profile to address comprehensive inventory practices, technical verification, adaptive threat detection, and supply chain security will ensure that organizations are better equipped to mitigate AI-specific risks in dynamic cloud environments. While companies like Wiz have developed solutions that reflect these principles, NIST's guidance should focus on ensuring that organizations adopt robust, technology-agnostic practices that improve security outcomes across the broader AI ecosystem.

Adaptive, Automated Cyber Defense in an Evolving Threat Landscape

The rapid evolution of AI-enabled cyberattacks highlights the urgent need for security controls that can operate at machine speed. Adversaries are increasingly leveraging automated attack techniques that exploit vulnerabilities faster than traditional, human-centric response models can manage. This reality underscores the importance of integrating automated containment mechanisms that can autonomously identify and neutralize threats in real time. While the Concept Paper acknowledges the value of proactive defense strategies, it does not sufficiently emphasize how AI-driven containment

systems can mitigate threats before they spread. Technologies such as **ADAMnetworks**[8] that leverage AI to trigger containment measures — such as microsegmentation, session termination, or endpoint isolation — provide organizations with the ability to disrupt adversarial activities with minimal delay. Expanding guidance within NIST's AI Risk Management Framework (AI RMF) to highlight these capabilities would provide organizations with practical strategies to contain fast-moving threats while reducing reliance on slow, manual intervention. Such guidance could build on the AI RMF's **MANAGE 3.2** control by emphasizing the need for automated mechanisms capable of rapidly adapting to dynamic attack patterns.

Equally important is the need for organizations to embrace Zero Trust principles as a foundational design element when developing or deploying AI systems. While the Concept Paper encourages Zero Trust adoption, further emphasis is needed on integrating Zero Trust directly into system architecture rather than implementing it as an add-on. AI-enabled security solutions that adopt a "Zero Trust by Design" approach exemplify this concept by ensuring that all interactions — whether between users, devices, or data flows — are continuously verified. Implementing Zero Trust in this way aligns closely with the AI RMF's **MAP 3.1** control, which emphasizes the need to manage dynamic risks by mapping system interactions and enforcing secure behavior. By promoting Zero Trust as a core design principle, NIST can better equip organizations to defend against both adversarial attacks and unintentional errors that may compromise AI model integrity.

In addition to containment and Zero Trust strategies, NIST should expand its guidance on resilience by encouraging the adoption of automated remediation and self-healing capabilities. As AI-enabled threats become more sophisticated, traditional detection methods that merely identify anomalies without addressing system recovery leave organizations vulnerable to prolonged disruption. The Concept Paper references the need for resilience but does not provide sufficient detail on how organizations can leverage AI to automate system recovery processes. Expanding guidance within the AI RMF's **RECOVER 1.1** control to emphasize AI-enabled self-healing capabilities would help organizations limit downtime and reduce the impact of cyber incidents. Solutions that autonomously restore compromised endpoints to a secure baseline demonstrate how automated remediation can mitigate ransomware, configuration tampering, and other forms of AI-driven compromise. By adopting AI-driven self-healing capabilities, organizations can significantly reduce the operational burden on security teams while improving their ability to sustain critical operations.

[8]See ADAMnetworks solution https://adamnet.works/

As NIST works to advance guidance on AI-enabled security solutions, it is also essential to emphasize the importance of balancing automation with rigorous oversight. Autonomous security systems that rely on AI to make decisions in real-time introduce new risks if those decisions are inaccurate or vulnerable to manipulation. While AI-driven detection systems can improve efficiency and reduce false positives, organizations must ensure that these systems are tested and evaluated to confirm their reliability. Expanding the AI RMF's **GOVERN 3.2** control to include guidance on red-teaming and adversarial testing would provide organizations with practical strategies for validating automated containment tools and strengthening confidence in their outcomes. By simulating adversarial tactics, organizations can ensure that their AI-driven defenses can withstand sophisticated attacks and perform reliably in dynamic environments.

By expanding its guidance on automated containment strategies, promoting Zero Trust by design, encouraging self-healing capabilities, and reinforcing oversight practices, NIST can better support organizations seeking to manage the risks posed by AI technologies.

AI-Driven Cybersecurity for IoT/OT Environments Through Comprehensive Visibility, Collective Intelligence, and Adaptive Risk Management

The rapid proliferation of IoT and OT devices across critical sectors such as healthcare, industrial control systems, and infrastructure has significantly expanded the attack surface for cyber threats. Unlike traditional IT systems, these connected assets often lack built-in security controls, making them particularly susceptible to compromise. As organizations increasingly adopt AI systems to enhance automation, improve efficiency, and extract insights from data in IoT/OT environments, the absence of robust security guidance for these AI-integrated systems poses a serious risk. Without clear recommendations for securing AI systems that interact with IoT/OT environments, organizations may struggle to detect and mitigate targeted attacks or malicious manipulation of these systems.

To address this challenge, we strongly recommend that NIST expand its Cyber AI Profile to include specific guidance on securing AI systems in IoT/OT environments. Strengthening this guidance is essential to ensure that AI-enabled systems can effectively operate in these complex environments without introducing new vulnerabilities. Expanding on NIST CSF controls such as **ID.AM-01** and **ID.AM-02**, which focus on maintaining comprehensive inventories of hardware and software assets, NIST should recommend passive monitoring techniques to identify unmanaged or legacy IoT/OT devices that cannot support traditional security agents. Passive monitoring is particularly crucial in OT environments, where active scanning could disrupt operational processes. Ensuring comprehensive asset visibility through passive monitoring will provide AI systems with the full contextual awareness needed to detect anomalous behavior and mitigate risks.

In addition to enhanced visibility, the Cyber AI Profile should emphasize the value of AI-driven behavioral analysis models. By establishing baseline behavior patterns for IoT/OT assets, such models can detect deviations that may signal a compromise or attack. This approach aligns with **CSF DE.CM-01** (monitoring networks for potentially adverse events) and **DE.CM-09** (monitoring computing hardware and software for suspicious activity), both of which support effective anomaly detection in IoT/OT environments. The NIST AI RMF further reinforces this strategy through its emphasis on adaptive risk measurement — notably in **MS-1.1-006**, which recommends real-time monitoring of AI system impacts, and **MS-1.1-007**, which advises tracking data quality during AI training to detect potential manipulation or degradation of performance. Extending these AI RMF principles to IoT/OT systems would provide a comprehensive framework for ensuring the security and integrity of AI-integrated environments.

Another crucial area where the Cyber AI Profile should expand is in promoting the use of collective intelligence models to improve threat detection and mitigation. IoT/OT environments are especially vulnerable to emerging attack techniques, as adversaries often target weakly secured devices across multiple organizations. Leveraging shared intelligence networks can help organizations identify and respond to these threats faster and more effectively.

For example, Armis'[9] collective intelligence model provides a compelling example of this approach. By aggregating behavioral insights from various device profiles, Armis enables rapid identification of emerging threats. For instance, if a malware strain targeting industrial sensors is identified in one organization, Armis' platform can instantly alert other organizations with similar devices, helping them implement proactive defenses. This model reflects the principles of **CSF ID.RA-2**, which encourages organizations to participate in information-sharing ecosystems to improve threat detection. It also aligns with **AI RMF MS-2.7-007**, which advocates for using AI-driven insights, such as red-teaming practices, to identify and mitigate emerging attack patterns. By encouraging organizations to adopt such collaborative threat intelligence practices, the Cyber AI Profile can help build a stronger, collective defense against AI-enabled attacks.

The Cyber AI Profile should incorporate guidance on continuous risk assessment models that dynamically adjust security controls based on changing asset risk. As IoT/OT systems increasingly rely on AI for automated control and decision-making, organizations must have mechanisms to ensure these systems can detect risk changes and adapt accordingly. This aligns with **CSF ID.RA-05**, which emphasizes risk prioritization based on threat likelihood and potential impacts. Incorporating dynamic risk assessment models would allow

[9]See Armis solution https://www.armis.com/

organizations to implement adaptive security controls such as micro-segmentation, which can isolate high-risk devices or restrict their access without disrupting critical operations. Aligning this guidance with **CSF PR.AA-05**, which emphasizes enforcing least privilege access policies, would further enhance security postures by minimizing the impact of compromised assets.

NIST's AI RMF reinforces this adaptive security approach, particularly through **MS-2.7-001**, which encourages dynamic risk measurement techniques, and **MG-2.3-001**, which advocates for defining risk-based access policies for AI systems. Incorporating these principles into the Cyber AI Profile would ensure organizations adopt security strategies that are capable of responding to evolving threats while maintaining system availability.

By expanding its guidance to include asset visibility, collective intelligence practices, and dynamic risk assessment models, the Cyber AI Profile can provide organizations with a comprehensive framework for mitigating AI-specific risks in IoT/OT environments. These measures are essential to ensuring that organizations can leverage AI's full potential while safeguarding critical systems and infrastructure from emerging cyber threats.

AI Asset Inventory, Management, and Incident Response Frameworks

The complexity of AI systems introduces unique risks that differ from traditional IT environments. AI models are susceptible to adversarial attacks, data poisoning, model inversion, and concept drift. Without clear frameworks for asset management, monitoring, and incident response, these risks can undermine the integrity and security of AI-driven systems. In this context, companies such as Cranium AI[10] offer effective solutions that align with NIST's proposed Cyber AI Profile objectives and can help address critical gaps in AI risk management.

Managing the full lifecycle of AI systems requires organizations to establish comprehensive inventories of their AI assets. Cranium AI offers a comprehensive AI asset inventory system that enables organizations to track and manage the full lifecycle of their AI systems. This includes cataloging datasets, models, infrastructure, and experiments—ensuring visibility across all AI components. Expanding CSF 2.0's **ID.AM (Asset Management)** function to include specific requirements for AI inventories would address a critical gap in traditional cybersecurity frameworks. This recommendation aligns with the AI RMF's **MAP 1.1** control, which emphasizes the importance of documenting AI system contexts, intended uses, and risk profiles. Further alignment with CSF 2.0's **GV.SC-02** control, which assigns roles for supply chain risk management, would extend this accountability model to AI systems, ensuring that organizations designate AI risk owners who are responsible for addressing

[10]See Cranium AI solution https://cranium.ai/

issues such as model drift, data corruption, or adversarial attacks. Integrating these practices into the Cyber AI Profile would align AI asset governance with broader risk management strategies and ensure organizations have full visibility into their AI ecosystems.

Managing AI-specific threats requires ongoing monitoring and validation to detect adversarial behavior, poisoned data inputs, or suspicious model performance. Cranium AI's platform is designed to provide continuous monitoring capabilities that identify signs of model degradation, performance anomalies, or abnormal data patterns that may indicate an attack. Incorporating AI-specific monitoring practices within CSF 2.0's **DE.CM (Security Continuous Monitoring)** function would address a critical need for improved AI system observability. Aligning this with the AI RMF's **MEASURE 1.1** and **MEASURE 1.3** controls, which emphasize developing appropriate metrics for AI trustworthiness and conducting independent AI system assessments, would ensure that organizations are equipped to detect and mitigate emerging AI threats. Cranium AI's approach reflects these principles by providing organizations with tools that track model behavior and automatically alert security teams to potential compromises, enhancing proactive risk mitigation strategies.

Equally important is ensuring organizations have robust response plans tailored to AI-specific security incidents. Traditional incident response plans often fail to account for the complexities of AI systems, which require distinct strategies for issues such as corrupted models, data poisoning attacks, or compromised training pipelines. Platforma that integrates rollback mechanisms allow organizations to revert to safe model checkpoints following an incident, while also supporting root-cause analysis to identify how vulnerabilities emerged. Expanding CSF 2.0's **RS (Respond)** and **RC (Recover)** functions to include AI-specific incident response protocols would address this gap and align with AI RMF's **GV-1.5-002** and **GV-6.2-003** controls, which provide best practices for managing AI-related incidents. Aligning these AI-specific response strategies with the Privacy Framework's **PR.PO-P1** control, which emphasizes data security in incident response, would further reinforce NIST's objectives for protecting AI system data pipelines.

Additionally, privacy concerns remain a key consideration in AI risk management. Since AI models depend on large volumes of data for training and inference, securing those datasets against unauthorized access or tampering is critical. Technology solutions that supports comprehensive data governance practices align with the Privacy Framework's **ID.IM-P** control, which highlights the importance of maintaining robust data inventories. By extending this principle to AI datasets, organizations can better secure their AI pipelines from data corruption, manipulation, or exposure risks.

By incorporating these recommendations, NIST can significantly strengthen the Cyber AI Profile and provide organizations with actionable guidance for managing AI risks. Integrating

comprehensive AI asset inventories, continuous monitoring controls, and AI-specific incident response protocols will ensure the Cyber AI Profile reflects the complexity of emerging threats while supporting organizations in maintaining the security, trustworthiness, and resilience of their AI systems.

Enhancing Development Pipeline Security, Threat Detection, and Data Integrity

The rapid growth of AI adoption has led to increasingly complex development environments that present significant security risks. One of the most pressing concerns is that AI model development is frequently conducted within **insecure development pipelines** that lack proper visibility and security controls. These vulnerabilities create opportunities for adversaries to manipulate the development process, potentially resulting in compromised AI models. Threat actors may exploit insecure pipelines to inject malicious code, conduct data poisoning attacks, steal intellectual property (IP), or exfiltrate sensitive information such as personally identifiable information (PII). Without comprehensive visibility into these pipelines, organizations struggle to identify where security gaps exist and how to effectively implement controls, leaving AI systems vulnerable to compromise.

To address this risk, NIST's Cyber AI Profile should emphasize enhanced security practices for AI development pipelines. Technology companies such as Legit Security[11] have developed solutions that provide comprehensive visibility across development environments to mitigate these threats. Technology security platforms that maps and visualizes development pipelines, allowing organizations to identify which projects are associated with AI development, where PII or other sensitive data is present, and how these data flows are structured. This visibility is essential for identifying weak points that may expose models to adversarial attacks or manipulation. By visualizing the full development pipeline — including CI/CD workflows, repositories, dependencies, and infrastructure — organizations gain the necessary context to apply appropriate security controls where they are most effective.

This enhanced visibility directly supports **CSF 2.0's IDENTIFY (ID.AM-03)**, which emphasizes the importance of maintaining accurate representations of authorized network communication flows, and **PROTECT (PR.PS-06)**, which advocates for integrating secure software development practices throughout the software development lifecycle. Expanding the Cyber AI Profile to include guidance on securing AI pipelines through enhanced visibility, data tracking, and control implementation will enable organizations to detect risks earlier in the development process and safeguard AI models from manipulation, data breaches, and compromise. By addressing these issues directly, NIST can help organizations mitigate one of the most significant security risks associated with AI adoption.

[11]See Legit Security solution https://www.legitsecurity.com/

AI-driven cyberattacks present a growing and increasingly sophisticated threat that demands stronger security practices. Attackers are leveraging AI to automate malicious activities, bypass traditional detection mechanisms, and generate complex attack patterns that exploit vulnerabilities within development environments. One particularly concerning risk is the potential for adversaries to use AI-enhanced techniques to move laterally through insecure pipelines, escalating privileges or compromising interconnected systems. To address this evolving threat landscape, NIST should expand its Cyber AI Profile to include proactive controls that identify and respond to AI-enabled threats.

Technology solutions such as Legit Security's graph-based risk analysis provide effective defenses by mapping attack pathways within development pipelines, detecting security gaps that adversaries may exploit, and identifying risks that enable privilege escalation or lateral movement. Legit Security's platform strengthens these defenses by proactively detecting hardcoded secrets, misconfigurations, and vulnerabilities that could serve as entry points or enable threat actors to expand their access. These capabilities align with **CSF 2.0's DETECT (DE.AE-02)**, which emphasizes analyzing potentially adverse events to understand associated activities, and **DE.AE-04**, which calls for estimating the impact and scope of such events. By recommending these proactive measures in the Cyber AI Profile, NIST can better prepare organizations to defend against AI-enhanced threats by promoting practices such as pipeline monitoring, anomaly detection, and threat modeling specifically designed to identify AI-based attacks.

As AI becomes an integral part of cyber defense strategies, organizations must also be mindful of the risks associated with over-reliance on automated tools. While AI-enhanced security platforms provide powerful threat detection and response capabilities, they are not immune to errors such as false positives, model drift, or blind spots in detection logic. To ensure these tools are deployed effectively, NIST should encourage organizations to adopt robust evaluation and validation practices that combine automated detection with manual review processes. Legit Security's platform, which integrates automated risk prioritization with contextual insights and manual verification, provides a strong example of how organizations can maintain oversight when adopting AI-enhanced defense tools. In addition to its automated capabilities, technology solutions, such as Legit, can signal when manual reviews are necessary based on specific risk factors. For example, if a developer introduces code generated by a GenAI tool, the platform can trigger alerts recommending mandatory code reviews with multiple approvers to ensure security and quality standards are met. This targeted alerting mechanism helps organizations apply human oversight where it is most needed, reducing the risk of insecure or vulnerable code being introduced into production. This approach aligns with **CSF 2.0's GOVERN (GV.RM-06)**, which emphasizes establishing standardized methods for categorizing and prioritizing cybersecurity risks, and the **AI RMF's RM-GOV-3**, which emphasizes documenting and managing risks associated with AI

systems. By recommending best practices for validating AI-based security tools, NIST can help organizations mitigate the risks that may emerge when relying on automated threat detection systems.

Data integrity and privacy risks are another critical concern for AI systems, particularly in relation to model training and inference. AI models frequently rely on large datasets, which, if compromised, may result in corrupted or biased model behavior. Ensuring data integrity throughout the AI model lifecycle is essential to reducing the risks of adversarial manipulation, data poisoning, or unauthorized access to sensitive information. Legit Security's secrets detection and prevention capabilities provide a strong example of how technology solutions can proactively identify hardcoded credentials, API keys, and other sensitive information that may inadvertently expose AI models to risk. These capabilities align with the **NIST Privacy Framework's Protect-P (PR.P-1)**, which emphasizes implementing data protection safeguards, and **Control-P (CT.P-2)**, which focuses on managing data with sufficient granularity to mitigate risks. Expanding the Cyber AI Profile to emphasize the protection of AI training data, including strategies such as differential privacy, secure multiparty computation, and secure data pipelines, will better enable organizations to mitigate these threats.

Given the growing reliance on AI tools in modern development workflows, NIST's Cyber AI Profile should address the security risks associated with code generated through these tools. While AI models that generate code can accelerate development, they may also introduce insecure coding patterns, logical flaws, or potential backdoors. To mitigate these risks, organizations need enhanced visibility into which developers are using GenAI tools and should implement additional security controls accordingly. Technology solutions, such as those offered by Legit Security, can identify developers who have access to GenAI tools, allowing organizations to apply tailored safeguards such as mandatory code reviews, multiple approvers, or additional testing steps in their development pipelines. While identifying AI-generated code directly remains a challenge, this proactive approach ensures that organizations maintain oversight in areas where AI-related risks are most likely to emerge. This strategy aligns with **NIST CSF 2.0's ID.RA-01**, which emphasizes identifying and validating asset vulnerabilities, and the **AI RMF's MP.2**, which encourages secure model deployment practices. By expanding the Cyber AI Profile to recommend visibility into GenAI tool usage alongside enhanced code review and testing processes, NIST can help organizations strengthen their defenses against potential security risks introduced by AI-assisted development.

Finally, supply chain security remains a significant concern for AI systems, particularly given the growing reliance on third-party AI models and components. Attackers may exploit vulnerabilities in third-party libraries or tamper with model updates to compromise AI

systems. Technology solutions that build integrity monitoring and artifact verification provide practical tools to mitigate these risks, aligning with **CSF 2.0 ID.SC-07**, which emphasizes recording, prioritizing, and monitoring supplier risks throughout the relationship lifecycle, and the **Privacy Framework's CT.P-6**, highlights the need for secure data processing practices involving third-party providers. By expanding the Cyber AI Profile to emphasize verifying AI model authenticity, tracking third-party dependencies, and defining contractual security requirements for AI vendors, NIST can provide organizations with practical strategies to manage the risks associated with external AI components.

Risk-Based Vulnerability Management, Automated Remediation, and GenAI Safeguards

In addition to improving supply chain visibility, NIST's Cyber AI Profile should expand its focus on risk-based vulnerability management to support organizations in prioritizing security efforts more effectively. The pace of vulnerability discovery in AI systems, particularly those that rely on rapidly evolving open-source libraries, often leaves security teams overwhelmed. Conventional vulnerability management practices that emphasize vulnerability counts alone fail to account for real-world exploitability, which can result in organizations expending valuable resources on low-risk issues while overlooking genuine threats. Solutions that apply contextual risk intelligence, such as Lasso's[12] platform, help address this challenge by correlating vulnerability data with usage context, exploitability, and real-world threat intelligence. For example, a vulnerability in a widely used AI library may present minimal risk if the vulnerable code path is never executed in the organization's deployed systems.

By integrating risk-based vulnerability management techniques into its Cyber AI Profile, NIST can encourage organizations to prioritize vulnerabilities that present tangible security risks. Emphasizing this approach would align with CSF 2.0 categories such as **ID.RA-04** (Potential impacts and likelihoods of threats exploiting vulnerabilities are identified and recorded) and **PR.DS-10** (The confidentiality, integrity, and availability of data-in-use are protected). Encouraging security teams to apply AI-driven analysis tools that assess exploitability and real-world risk will help organizations strengthen their security posture while optimizing resource allocation.

While improving vulnerability prioritization is essential, effective security practices must also ensure that identified risks are remediated in a timely and efficient manner. As organizations expand their adoption of AI technologies, manual remediation methods may prove insufficient to manage the growing volume of security issues. Automating remediation workflows through integration with DevSecOps pipelines can significantly improve response

[12]See Lasso Security solution https://www.lasso.security/

times and reduce exposure windows. Technologies such as Lasso's automated remediation platform offer a valuable model by automatically generating pull requests and code fixes when vulnerabilities are detected, allowing development teams to resolve security issues directly within their workflows. We encourage NIST to promote automated remediation practices within the Cyber AI Profile by emphasizing the importance of integrating security automation tools into CI/CD pipelines. Introducing guidance that encourages automated remediation methods will align with CSF 2.0 practices such as **PR.PS-02** (Software is maintained, replaced, and removed commensurate with risk) and **RS.MI-02** (Incidents are eradicated). By encouraging automated remediation, NIST can help organizations adopt proactive security measures that address vulnerabilities rapidly and minimize potential damage.

The emergence of Generative AI (GenAI) tools adds further complexity to the cybersecurity landscape. While these tools offer immense potential to improve productivity, they also pose significant security risks. Large Language Models (LLMs) in particular introduce risks related to misinformation, data leakage, and adversarial manipulation. Without proper safeguards, organizations risk inadvertently sharing sensitive information or relying on GenAI-generated insights that may be biased or inaccurate.

To address these risks, NIST's Cyber AI Profile should include guidance on securing GenAI tools and mitigating LLM-related vulnerabilities. Effective solutions should combine proactive content filtering, real-time monitoring, and robust access control mechanisms to ensure that GenAI interactions are protected. Technologies such as **Lasso's** GenAI security platform provide an effective model by integrating access controls that limit interactions to authorized users, deploying real-time content monitoring tools to detect data leakage and misinformation, and providing comprehensive audit trails to ensure compliance with governance frameworks. Incorporating controls from NIST's AI RMF will further strengthen NIST's guidance on GenAI security. We recommend that NIST adopt AI RMF principles such as **GV-1.2** and **GV-1.3,** which focus on establishing governance policies to define acceptable GenAI use, monitor performance, and establish testing protocols to mitigate LLM-related risks. Additionally, integrating **MP-1.1** (AI systems are assessed for risks throughout their lifecycle) will encourage organizations to implement risk assessment strategies that account for the unique vulnerabilities presented by LLM technologies.

Furthermore, NIST should address the importance of content integrity controls by promoting practices such as provenance tracking, watermarking, and data validation to ensure GenAI outputs are trustworthy and resistant to manipulation. Aligning with AI RMF guidance such as **GV-4.3-001** (Provenance techniques are employed to ensure content integrity) will support these efforts. Establishing clear boundaries for human oversight in GenAI adoption, as described in **GV-3.2** (Human-AI configuration controls), will further

protect against over-reliance on GenAI systems and ensure that sensitive decisions remain subject to human judgment.

AI Data Security with Encryption-in-Use Solutions to Mitigate Data Exposure Risks

AI systems often require access to large volumes of data to train, validate, and refine their models. However, traditional data security measures focus on encrypting data only at rest or in transit. This leaves a critical security gap during the data-in-use phase, when plaintext exposure occurs during computation. Threat actors increasingly exploit this vulnerability, particularly in AI environments where data is aggregated for model training or real-time analytics. As such, addressing this data-in-use risk is essential for building secure AI ecosystems.

Technologies that enable **encryption-in-use** offer a promising solution to mitigate this gap. Solutions like those developed by Vaultree[13] provide **fully functional data-in-use encryption**, allowing data to remain encrypted even during processing, analysis, and AI-driven computations. By maintaining encryption throughout the data lifecycle, this approach effectively neutralizes risks tied to plaintext exposure, reducing potential attack surfaces even if adversaries breach internal systems. This technology leverages advanced cryptographic principles such as homomorphic encryption and secure multiparty computation, ensuring that AI models can operate on encrypted data without requiring decryption. Such solutions directly address the growing security risks that accompany AI adoption across sectors.

To strengthen the Cyber AI Profile's focus on data protection, we recommend that NIST integrate guidance that aligns with controls in the **NIST CSF 2.0** and **NIST Privacy Framework**. In the CSF 2.0, the **PROTECT (PR)** function offers relevant outcomes that align with encryption-in-use strategies. While **PR.DS-2 (Data-in-Transit Protection)** and **PR.DS-3 (Data-at-Rest Protection)** provide strong foundations for traditional encryption practices, they do not sufficiently address the risk of data exposure during active computation — a critical vulnerability in AI pipelines. To close this gap, we recommend that NIST expand the **PR.DS** category to include a new subcategory on **Data-in-Use Protection**. This addition would guide organizations to adopt encryption-in-use technologies that minimize plaintext data exposure during AI model training, inference, and analysis.

In the **GOVERN (GV)** function of CSF 2.0, **GV.RM-3** encourages organizations to incorporate security controls that align with risk tolerance. We recommend that NIST include encryption-in-use as a recommended technical safeguard to enforce data security

[13]See Vaultree solution https://www.vaultree.com/

controls in AI systems, especially where highly sensitive data such as medical records, financial data, or proprietary intellectual property are involved.

In the NIST Privacy Framework, the **PROTECT-P (PR.P)** category highlights the importance of privacy-enhancing technologies (PETs) to ensure data protection during processing. Vaultree's encryption-in-use technology aligns directly with the **PR.P-3** outcome, which calls for minimizing unauthorized data access during data processing. We recommend that NIST expand the Privacy Framework to include explicit reference to encryption-in-use solutions as a PET that supports secure AI adoption and privacy-preserving data processing.

Additionally, encryption-in-use technologies are well-positioned to support compliance requirements under data protection regulations such as **GDPR**, **HIPAA**, and various state-level data security laws. By ensuring that sensitive data remains encrypted even in active use, organizations can reduce legal exposure and often meet regulatory exemptions that apply to encrypted data breaches. We encourage NIST to highlight this compliance advantage in its updated guidance, underscoring how encryption-in-use strengthens both security and regulatory alignment.

NIST's emphasis on fostering innovation in PETs within the AI security landscape is particularly critical. By promoting ongoing research, pilot programs, and industry adoption of encryption-in-use solutions, NIST can encourage widespread adoption of privacy-preserving technologies that align with secure-by-design principles in AI development. This aligns with CSF 2.0's **IN-4 (Innovation Support)** and the Privacy Framework's **GV.P-3 (Privacy Innovation)** objectives, reinforcing the value of secure data handling throughout AI deployment. By integrating these recommendations, NIST can strengthen the Cyber AI Profile to better address data-in-use vulnerabilities, encourage the adoption of advanced PETs like encryption-in-use, and promote secure AI innovation.

As this concept paper implementation evolves, we look forward to discussing these proposals with NIST and are available for any questions. We remain excited to collaborate with NIST to increase engagement with innovative companies.

Respectively,

/s/ Michelle Sahar
Michelle Sahar
Cybersecurity Policy Director, OpenPolicy

/s/ Dr. Amit Elazari
Amit Elazari
Co-Founder and CEO, OpenPolicy