HIDDENLAYER
PROTECT YOUR ADVANTAGE

Feb. 1, 2024

To: National Institute of Standards and Technology (NIST), Commerce

Attn: *ai-inquiries@nist.gov*

HiddenLayer comments on

Executive order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence issued on October 30, 2023.

OVERVIEW

HiddenLayer appreciates the opportunity to provide feedback to NIST as it seeks assistance in carrying out its responsibilities under the Executive order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence issued on October 30, 2023.

The HiddenLayer team was born out of a real-world adversarial machine learning attack in 2019 when Chris Sestito, Jim Ballard, and Tanner Burns (the HiddenLayer founders) were responsible for responding to a serious, real-world adversarial machine learning attack. At the time, Chris Sestito (HiddenLayer CEO) led Threat Research at Cylance, an AI company that revolutionized the anti-virus industry by leveraging deep learning to prevent malware attacks. In 2019, the Windows executable ML model was exploited via what is now known as an inference attack, exposing its weaknesses and allowing the attackers to successfully evade detection anywhere Cylance was running. During the response effort, the future HiddenLayer founders saw it as a precursor of attacks to come made possible by the inherent weaknesses in AI/ML, more open source attack tools, and increasing knowledge of and usage of the fastest growing, most important technology the world has ever seen. Determined to prove that these attacks were preventable, the team developed a unique, patent-pending, productized security for AI solution to help all organizations mitigate security risks inherent within AI based solutions.

HiddenLayer is addressing a critical gap to secure and accelerate the responsible use of Artificial Intelligence (AI), one of the world's most valuable technologies. Despite the staggering growth of AI across every industry, organizations are oftentimes unknowingly opening themselves to vulnerabilities and adversarial attacks due to insufficient investments and education on current AI threats. Our AI Security (AISec) Platform, provides comprehensive security that collectively protects AI models against adversarial attacks, vulnerabilities, and malicious code injections. Each product within AISec Platform is designed with unique strengths and capabilities for detecting and responding to attacks, creating a well-rounded defensive strategy against threats.

HiddenLayer's flagship Machine Learning Detection and Response (MLDR) product provides a noninvasive, software-based approach to monitoring the inputs and outputs of AI algorithms. MLDR offers real-time defense to an otherwise unprotected asset and flexible response options, including alerting, isolation, profiling, and misleading.

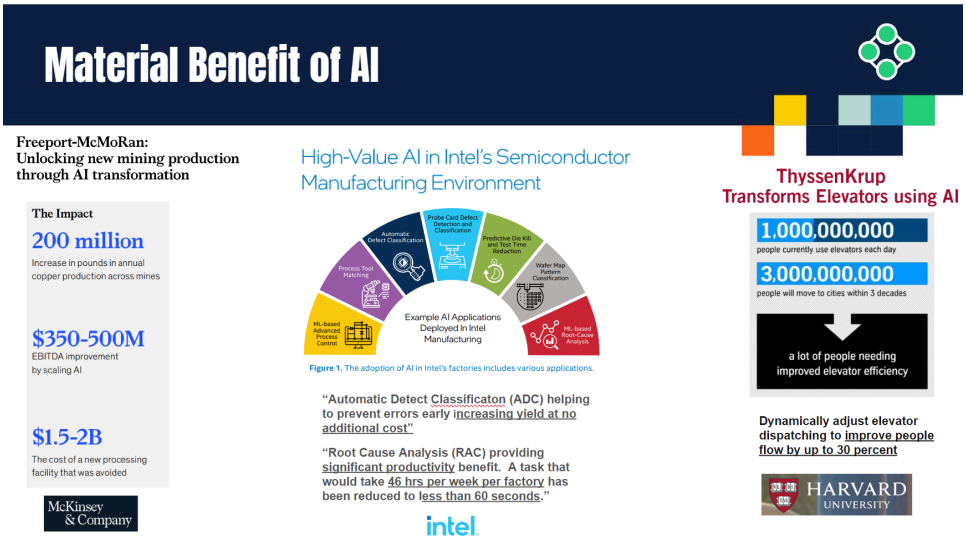
With the AI market projected to reach $16 trillion by 2030, our mission is to empower governments, academic institutions, and corporations to embrace AI responsibly, ensuring the secure and accelerated adoption of this invaluable technology.

In September of 2023 HiddenLayer raised $50M in Series A funding to expand its talent base, increase go-to-market efforts, and further invest in our Artificial Intelligence Security (AISec) Platform. The investment marks the largest Series A funding raised by a cybersecurity company focused on protecting AI. The funding was led by M12, Microsoft's Venture Fund, and Moore Strategic Ventures, with participation from Booz Allen Ventures, IBM Ventures, Capital One Ventures, and Ten Eleven Ventures. Press Release.

We have seen strong demand for our AISec Platform across a wide range of organizations since the company launched in July of 2022. We are working closely with many of the largest Financial, Healthcare, and Retail organizations, Universities, and the US government. We were proud to publicly announce our recent partnership with the Department of the Air Force (DAF) in October of 2023.

CLEAR AND PRESENT DANGER ALREADY EXISTS

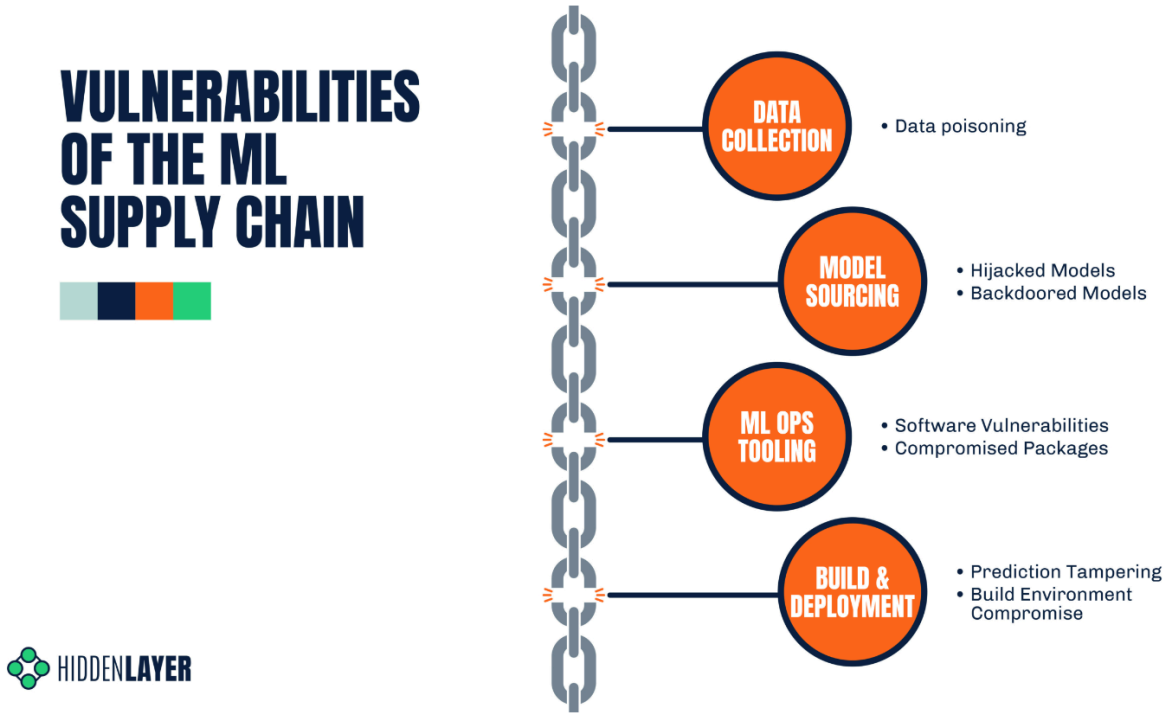
In July of 2023, Forrester Consulting and HiddenLayer released a study conducted with over 150 AI security decision makers. In that research, we determined that AI is already critical to business success with 96% indicating AI is critical or important to customer experience, revenue generation, and business operations. When an organization is achieving material benefit from the use of technology a potential material risk could occur if that technology is compromised in some way. *(Examples figure 1)*



Further in the study mentioned above we learned that 86% were extremely concerned or concerned about their organizations AI model security. We also learned that 40% to 52% of respondents are either using a manual process to address threats or they are still discussing how to address the threat. That means most if not all AI in use today has insufficient or no control at all to mitigate the risk of attacks against their AI models.

The HiddenLayer research team has also discovered a wide variety of unsettling issues over the past year. We have determined that AI/ML itself can be a launchpad for Ransomware. In Dec 2022 we demonstrated a proof-of-concept attack for surreptitiously deploying malware, such as ransomware or Cobalt Strike Beacon, via machine learning models. The attack uses a technique currently undetected by many cybersecurity vendors and can serve as a launchpad for lateral movement, deployment of additional malware, or the theft of highly sensitive data. ML becomes the New Launchpad for Ransomware | HiddenLayer MLDR. Pickle Strike | HiddenLayer MLDR.

We have also concluded that damaging supply chain attacks on AI/ML are easy. Insane in the Supply Chain | HiddenLayer MLDR. Using lessons we've learned from dealing with past incidents, we looked at the AI/ML Supply Chain to understand where people are most at risk *(see figure 2)*



Our own research team has already identified a wide variety of issues in each stage of the AI/ML supply chain. We have identified thousands of models that are in public repositories where the data is already poisoned, backdoors are already present and malicious code is already embedded. We have also discovered a number of zero day vulnerabilities which we have disclosed to affected parties and worked with them to resolve the identified issues in a timely manner.

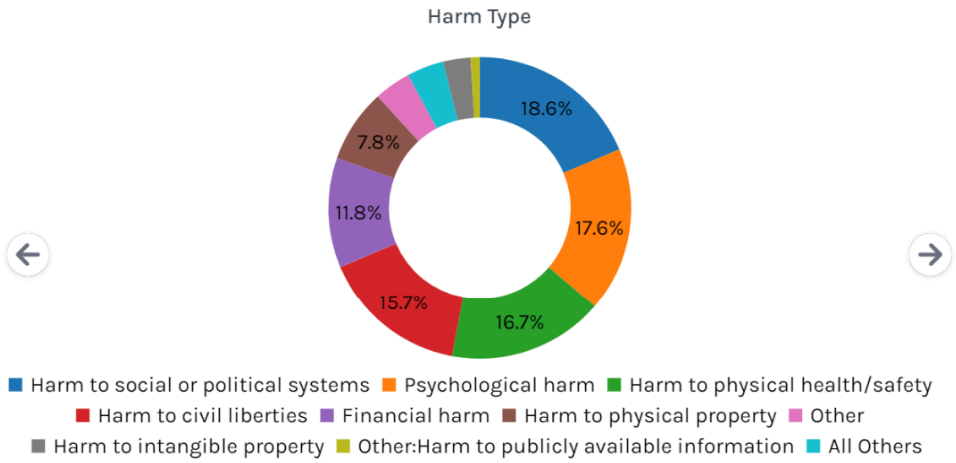
In the AL/ML ops tooling stage libraries such as TensorFlow, PyTorch, and NumPy are mainstays of the field, providing incredible utility and ease to data scientists around the world. But these libraries often depend on additional packages, which in turn have their own dependencies, and so on. If one such dependency was compromised or a related package was replaced with a malicious one your organization as well as potentially others who utilize the AI/ML you create could be at substantial risk. A recent example of this is the 'torchtriton' package which, due to dependency confusion with PyPi, affected PyTorch-nightly builds for Linux between the 25th and 30th of December 2022. Anyone who downloaded the PyTorch nightly in this time frame inadvertently downloaded the malicious package, where the attacker was able to hoover up secrets from the affected endpoint. Although the attacker claims to be a researcher, the theft of ssh keys, password files, and bash history suggests otherwise.

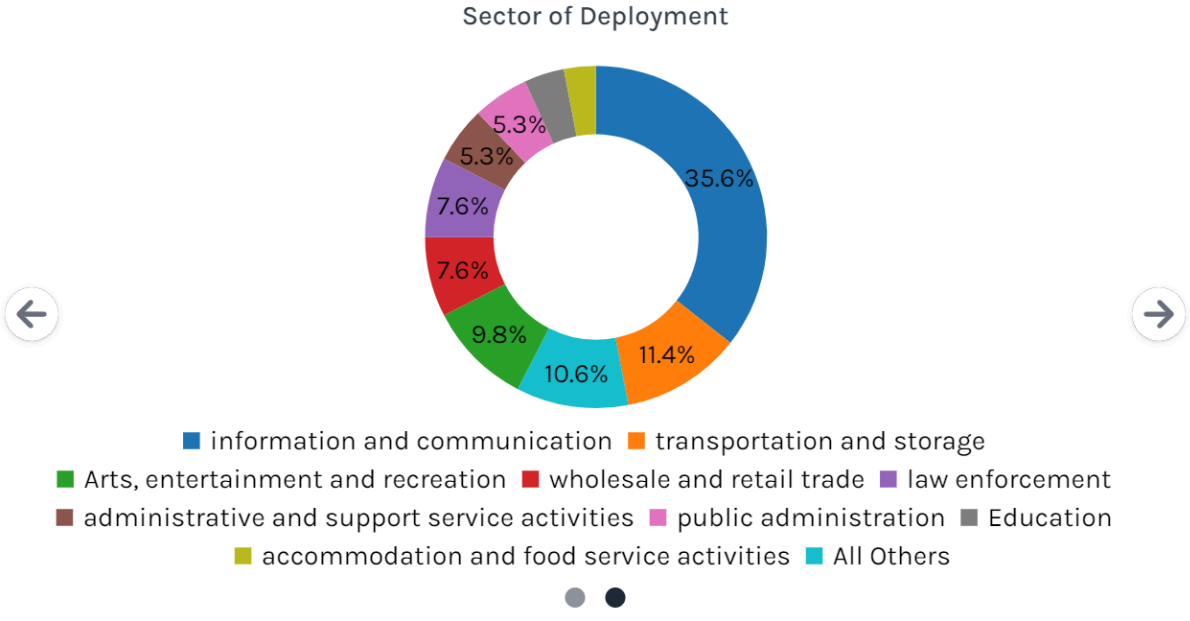
We have also concluded from our research in June of 2023 that there is already existing abuse of AI cloud services in ways that could generate substantial risk. Crossing the Rubika - The Use and Abuse of AI Cloud Service. As an example we observed an interesting case illustrating the unintended usage of Hugging Face Spaces, an online community for sharing ML models. A handful of Hugging Face users have abused Spaces to run crude bots for an Iranian messaging app called Rubika. Rubika, typically deployed as an Android application, was previously available on the Google Play app store until 2022, when it was removed – presumably to comply with US export restrictions and sanctions. The app is sponsored by the government of Iran and has recently been facing multiple accusations of bias and privacy breaches.

We came across over a hundred different Hugging Face Spaces hosting various Rubika bots with functionalities ranging from seemingly benign to potentially unwanted or even malicious, depending on how they are being used. Several of the bots contained functionality such as: administering users in a group or channel, collecting information about users, groups, and channels, downloading/uploading files, censoring posted content, searching messages in groups and channels for specific words, forwarding messages from groups and channels, sending out mass messages to users within the Rubika social network. Although we don't have enough information about their intended purpose, these bots could be utilized to spread spam, phishing, disinformation, or even propaganda.

AI is the latest, and likely one of the largest, advancements in technology of all time. Like any other new innovative technology, the potential for greatness is paralleled by the potential for risk. As technology evolves, so do threat actors. Despite how state-of-the-art Artificial Intelligence (AI) seems, we've already seen it being threatened by new and innovative cyber security attacks everyday. In fact an AI Incident Database has already been established and is dedicated to indexing the collective history of harms or near harms realized in the real world by the deployment of artificial intelligence systems. From this database you can search over 2000 reports of AI harm. In late 2023, the AI Incident Database indicated 76% of AI incidents have already had physical world implications. And critical sectors supporting our economy and society are already under attack *(see figure 3 and figure 4)*

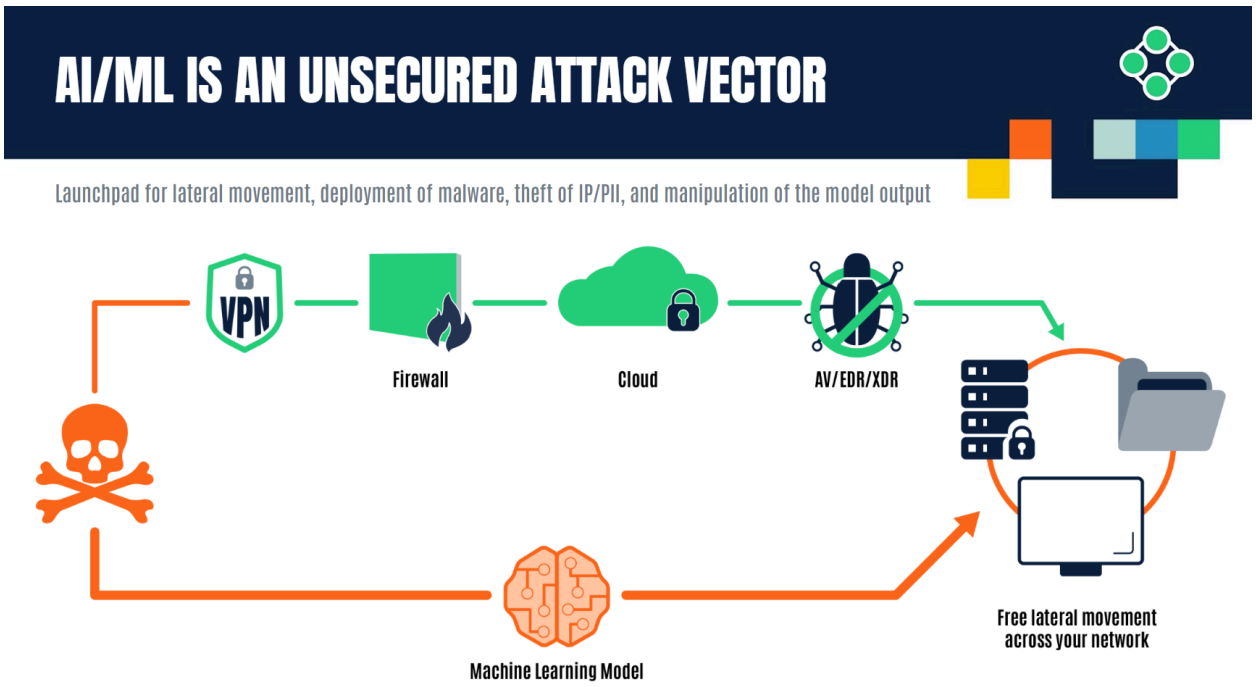
76.4% of AI incidents <u>have already had</u> Physical World Implications





EXISTING CONTROLS DO NOT WORK TO MITIGATE SECURITY RISKS IN AI

As mentioned above in the overview with regard to the Cylance adversarial AI attack in 2019 which was described in detail in this blog post. The founders of HiddenLayer have had real world experience with AI attacks. And while you cannot eliminate risk, this attack approach was one of the first of its kind against AI. Cylance had a world class security team and enhanced controls to prevent, detect, and respond to security risks in its enterprise IT environment, in its SAAS production environment, and against the product itself. Cylance was at the time SOC2 compliant, FedRAMP compliant, as well PCI compliant in addition to achieving other certifications. None. I repeat NONE of the traditional or enhanced controls in use or compliance frameworks that had been adopted to provide security assurance around the enterprise and AI tech stack being used at Cylance mitigated this adversarial AI attack *(simplified view of a traditional control environment figure 5)*



So why dont existing controls work to mitigate security risks related to AI, for either generative AI or predictive AI? The simple truth is that they were not designed to prevent, detect, or respond to mitigate these risks. They were not purpose built to provide any coverage for the variety of AI file types, AI model types, or in some cases the evolving components of the AI tech stack some of which were touched on above (Clear and Present Danger Exists).

When we transitioned from mainframes to PC's and servers, the controls used to manage security risks did not work. When we transitioned from PC's to laptops and from physical servers to virtual machines, many of the existing controls used to manage security risks did not work. As smart

phones took off or cloud computing exploded, once again the existing approach to control cyber risks did not work. The transition to AI and a new tech stack dedicated to AI is no different than any other technology transitions that have occurred. The existing security tech stack will not work or will not work sufficiently to mitigate the cyber risks that are present or ones that will emerge.

Key Controls Required to Manage & Mitigate Security Risks in AI

As we navigate through the complexities of an AI-driven era, understanding and implementing robust security measures is no longer a choice. It is a necessity. NIST should look to document in all its standards, key controls specific to mitigating security risks in AI. The following are HiddenLayers recommendations for key controls that need to be added or adjusted to incorporate AI specific technologies.

- **Discovery and Asset Management:**
  - Organizations need to begin by evaluating where AI is in use already within their organization. They need to be able to determine what applications have perhaps already been purchased that use AI or have AI-enabled features.
  - Organizations need to evaluate what AI may be under development. How many data scientists or data engineers roles are they employing or consultants they may have under contact.
  - Organizations need to understand what pretrained models from public repositories may already be in use.

To address Discovery and Asset management controls NIST should include in any standards specific control requirements for organizations to be able to identify and track all aspects of AI in use or development. Discovery tools should be able to identify AI models and tools that span the wide variety of file types and model types. Discovery tools should also be able to identify AI capabilities across on-prem, cloud, multi cloud, as well as AI on endpoints.

- **Risk Assessment and Threat Modeling:**
  - Organizations should conduct a benefit assessment to identify the potential negative consequences (impacts) associated with the AI systems if those systems/models were to be compromised in any way.
  - Organizations need to perform threat modeling specific to AI to understand the potential vulnerabilities and attack vectors that could be exploited by malicious actors to complete their understanding of the potential AI risk exposure.

To address Risk and Threat Modeling controls NIST should include in any standards control requirements that organizations map their assessments to the MITRE Atlas Framework. It should also highlight that any security tooling used to identify and remediate security concerns in AI be required to map to ATLAS. Mapping to ATLAS will improve the efficiency as well as effectiveness of security teams as they address security risks in AI within their organizations. MITRE ATLAS™

(Adversarial Threat Landscape for Artificial-Intelligence Systems) is a knowledge base of adversary tactics and techniques based on real-world attack observations and realistic demonstrations from AI red teams and security groups. ATLAS is modeled after the MITRE ATT&CK® framework which is widely used in the traditional cyber security landscape. HiddenLayer is proud to be participate in the evolution of ATLAS and will continue to dedicate resources towards.

- **Data Security and Privacy:**
    - Organizations need to go beyond the typical implementation of encryption, access controls, and secure data storage practices to protect your AI model data. As mentioned earlier, those controls will not effectively protect the data in AI models from theft, alteration, or other forms of attack.
    - Organizations need to embed into their 3rd party risk process an evaluation of any vendors' security for their AI capabilities. Organizations need to ask how their vendors incorporate security into their AI development lifecycle including how they scan their models for data poisoning and malicious executables. Organizations need to ask their vendors how they provide real-time/run time protection to detect and stop various forms of attacks against the AI capabilities embedded in the solutions you have bought from them.

NIST should include in any of standard control requirements for security for AI tools that are purpose built specifically to provide runtime protection for AI models. Security for AI solutions must be able to span the vast array of file types, model types, and also be agnostic to on-prem, cloud or multi cloud deployments. NIST should also expand its supply chain standards to include specific security for AI assurances from each stage in the supply chain including any touch points from 3rd parties.

- **Model Robustness and Validation:**
    - Regularly assess the robustness of AI models against adversarial attacks. This involves pen-testing the model's response to various attacks including intentionally manipulated inputs.
    - Implement model validation techniques to ensure that the AI system behaves predictably and reliably in real-world scenarios, minimizing the risk of unintended consequences.

NIST should include in any of its standards control requirements for security for AI, tools that are purpose built specifically to address across the various file types and model types. Model Robustness and validation tools should also be able to identify security concerns in AI capabilities across on-prem, cloud, multi cloud, as well as on endpoints.

- **Secure Development Practices:**
    - Organizations need to incorporate security into their AI development lifecycle. Organizations need to train their data scientists, data engineers, and developers on the various attack vectors associated with AI including how to minimize the potential attack surface up front in the security development lifecycle.

- Organizations should identify the AI security architecture required to be instrumented for the runtime protection of their AI when the models go into production use.

NIST should include in its standard for Secure Software development specific requirements for AI. NIST should implement recommendations from the workshop held on Jan. 17. 2023. https://www.nist.gov/news-events/events/nist-secure-software-development-framework-generative-ai-and-dual-use-foundation. Some key moments to call out from this workshop include Nick Hamilton of OpenAI who begins his talk at 1:11:55, Mihai Maruseac of Google at 2:00:01, David Beveridge of HiddenLayer at 2:46:23, and the IBM, Microsoft, & HiddenLayer panel Q&A which starts at 3:12:10.

- **Continuous Monitoring and Incident Response:**
  - Organizations need to implement continuous monitoring mechanisms to detect anomalies and potential security incidents in real-time for AI that is in use. Organizations need to require their vendors that embed AI into their solutions specific security capabilities that can alert their customers to attacks that could compromise their data or business processes.
  - Organizations need to develop a comprehensive AI incident response plan to quickly and effectively address security breaches or anomalies. Organizations need to regularly test and update the incident response plan to adapt to evolving AI threats.

NIST should include in any of its standards, control requirements for security for AI, tools that are purpose built specifically to address monitoring across the various file types and model types. Continuous monitoring and Incident Response tools and processes should also be able to identify security concerns in AI capabilities across on-prem, cloud, multi cloud, as well as on endpoints. A note of caution: Responsible & ethical AI frameworks and their practices in many cases fall short of adequately ensuring models are secure before they go into production as well as after an AI system is in use. They focus on things such as biases, appropriate use, and privacy. While these are also required we can't confuse these practices for actual security.

As NIST and its efforts to enhance security and trust in AI evolve, we look forward to assisting your efforts in any way that can make a difference for our country. NIST should also work with other agencies and initiatives including CMMC 2.0, FedRAMP, OMB, and with CISA to harmonize standards to reduce confusion and make it easier for organizations to follow guidelines that will enhance the security for AI. The HiddenLayer team is eager to help to protect our advantage as a nation and enable AI to unleash dramatic economic and social benefits to the world. We are available for any questions you may have.

Respectfully,

Malcolm Harkins, Chief Security & Trust Officer, HiddenLayer