
intel
HIDDENLAYER
The Future of Risk is Upon Us
And We Can Manage It if We Secure AI

Companies are embracing AI at an unprecedented pace to gain a powerful competitive edge, and in a rapidly evolving technological landscape, new threats have emerged. At HiddenLayer, in partnership with Intel, we are prepared to protect our clients against AI vulnerabilities without compromising on the speed of innovation.

Christopher Sestito
CO-FOUNDER & CEO, HIDDENLAYER

At Intel, maintaining data integrity, privacy, and accuracy is at the heart of all our security efforts. From innovations that help protect sensitive workloads to the development of AI technology that maintains the integrity of an organization's intellectual property, we are at the forefront of security and privacy research and development.

Rick Echevarria
VICE PRESIDENT, SECURITY CENTER OF EXCELLENCE, INTEL CORPORATION

Authors

Kasia Hanson
Global Sr. Director, Physical and Cybersecurity Ecosystem Development, Intel Corporation

Hiep Dang
VP Product, HiddenLayer

Executive Sponsor

Rick Echevarria
Vice President, Security Center of Excellence, Intel Corporation

Contributors

Malcolm Harkins
CISO

Brian Richardson
Security Marketing Leader, Intel Data Center Product Marketing, Intel Corporation

Matt Hoekstra
Sr. Director of the Security Solutions Lab, Intel Labs/ Security and Privacy Research, Intel Corporation

Research Consultant

Bridge Partners

For more information, please visit hiddenlayer.com

The Evolution of Security

Advanced Protection for AI

Addressing constantly evolving threats requires an evolution in security, which must be rooted in foundational technologies starting at the earliest stages of product design and continuing throughout the product lifecycle.

Security threats are constantly evolving in number and complexity; new attacks emerge, and existing threats are modified to work in new ways. Attacks are going lower in the stack and may be aimed directly at physical vulnerabilities—they have moved from software and applications to operating systems and middleware, down to the firmware and hardware level, and now via pathways opened up through the use of Artificial Intelligence (AI) and Machine Learning (ML) tools.

When it comes to the security and protection of AI models, the implications for corporations, organizations, and for society at large are of the utmost importance. AI security vulnerabilities are being discovered regularly and are also continuing to evolve. Perhaps more concerning is that successful penetration of AI does not require sophisticated technical skills on the part of malicious actors. In today's computing world, sensitive code and data are often disaggregated and scattered across multiple environments. AI models may operate on a foundation necessitating input from multiple sources, leading to exponentially more vulnerabilities. This is generating greater demand for Confidential Computing solutions from Intel, which can deliver greater security for sensitive data and code, and pave the way for the future of cloud migration.

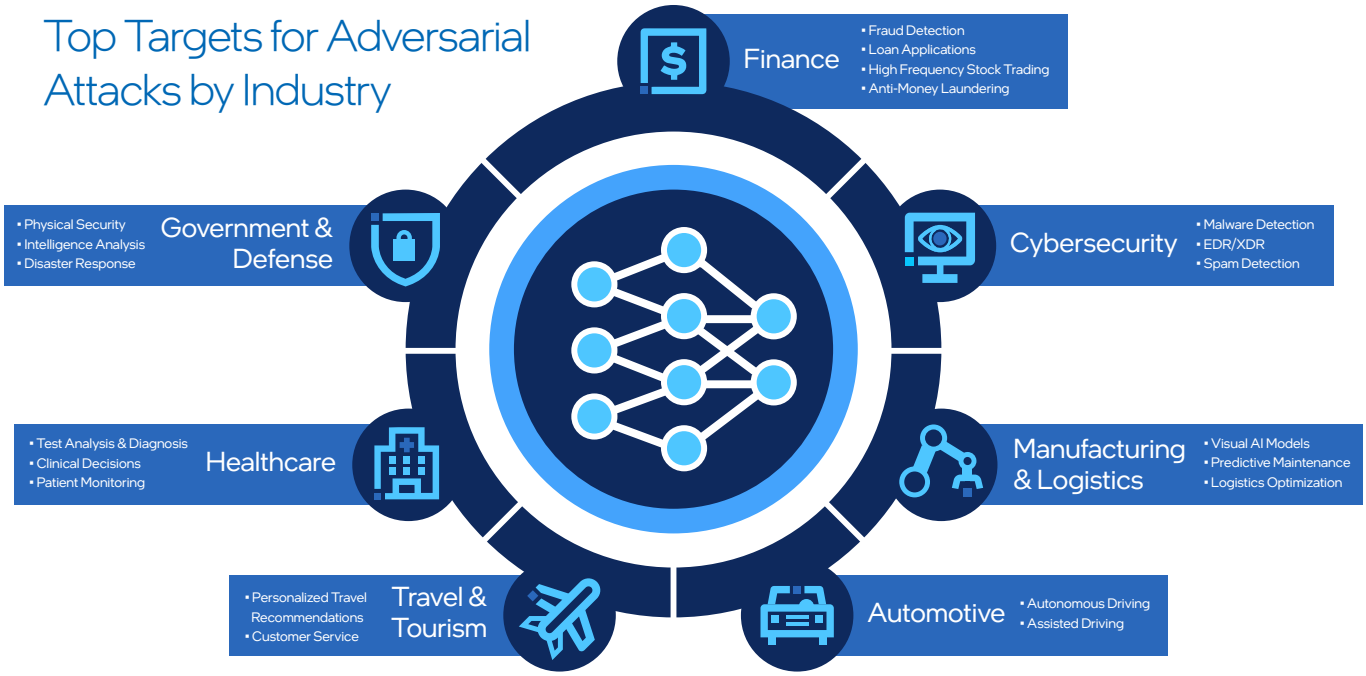
All these factors are driving the need for comprehensive, powerful, and efficient security solutions to protect organizations' AI models. Today, companies simply cannot be zero-trust in their security posture without confronting the need for security for AI.

Security Concerns and Challenges with AI Adoption

Today, talk of Artificial Intelligence is abundant, with some experts arguing that AI could be the most disruptive technology innovation of the modern era. Referencing the graphic below, you will see that companies across a wide array of industries are leveraging AI applications to analyze and make use of massive quantities of data. But the adoption of new and sophisticated technology rarely takes place along a smooth, obstacle-free path.

Several challenges await the enterprising organization that is interested in pursuing AI to positively impact operations, product development, and revenue generation. Making the move from pilot stages into a successful production cycle may be inhibited by costs, for things like compute resources, for example. Skilled and trained personnel may be scarce. Similarly, any deficit of AI operations technology or tools can deter growth just as easily as an insufficient amount (or poor quality) of data. Ongoing issues with governance and regulation must also be addressed in order to facilitate successful integration of AI into an organization.

Although the risks are real, the opportunities enabled by AI are massive. According to a recent study conducted by Forrester Research, a majority of business leaders states that machine learning projects are critical or important to the success of their companies over the next eighteen months.[1] Additional insight into the challenges associated with the adoption of AI can be gleaned from the following consensus: according to the same study, "86% of respondents are extremely concerned or concerned about their organization's AI model security."



The Prevalence of Manual Security Processes with AI

Unfortunately, many organizations that leverage AI models continue to handle security issues through the use of manual processes. This practice involves steps that must be performed by an individual human. First, the person identifies that an attack has taken place, after which that threat must be manually addressed, in many cases by shutting off the submitter to the AI model. Common issues can include insider threats, adversarial attacks on existing AI models, ransomware embedded in AI models or model components, and model theft or hijacking.

Because of the use of easily available, automated tools, cyber criminality has seen rapid democratization of attacks. Simultaneously, access to a company's data through the compromising of a detected vulnerability is becoming more and more easily gained. It should come as little surprise, then, that 80% of respondents in the aforementioned study expressed interest in investing in a technology solution to help manage AI model integrity and security in the next 12 months.

The Need for Proactive AI Security

We believe there are three catalysts that clearly underscore the need for a proactive security posture when it comes to AI:

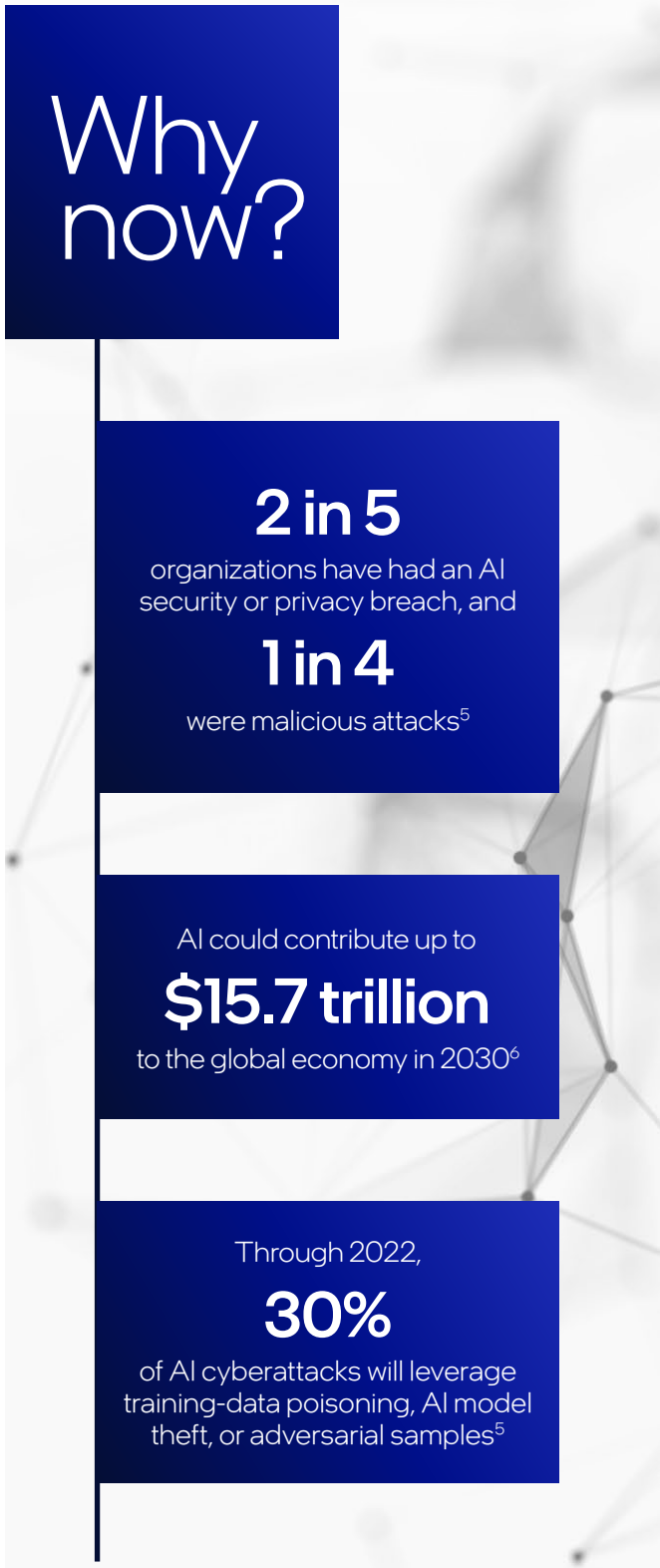
- **First,** real-world attacks highlight the urgency for comprehensive cybersecurity that includes AI in addition to the many other attack surfaces currently addressed
- **Second,** accelerating regulations (i.e., the AI Bill of Rights) underscore the demand for greater protection from unsafe or ineffective AI systems
- **Third,** the increased availability and use of weaponized Adversarial Machine Learning (AdvML) tools

Here, it is important to unpack the concept of Adversarial Machine Learning (AdvML) because it makes the case for proactive security with AI entirely clear.

MITRE ATLAS™ (Adversarial Threat Landscape for Artificial-Intelligence Systems) represents a framework for organizing and cataloging the adversarial tactics, techniques, and case studies currently dotting the machine learning landscape. As AI is leveraged across a growing number of use cases and industries, the number of vulnerabilities that are exposed increase concurrently. And these AI vulnerabilities augment the attack surface of the existing systems within which they operate. There are now more than sixty ways that a malicious actor can attack machine learning models, and all of these are cataloged in the MITRE ATLAS™ framework.

For example, a data scientist at a government organization may download a pre-trained AI model from a public repository as the foundation for their own project, leveraging pre-existing code as a means to accelerate project development and save on compute costs. In fact, this is a very common occurrence. The data scientist then uploads the model on their department system, uses their own training data on the model, and modifies it for their specific use case. But these downloaded models are essentially open-source public code that does not come with any kind of security guarantee, representation, or warranties that it is safe for use. Thousands of these models include things like backdoors, where cyber criminals can find easy access.

Making matters more complex, the reality today is that a threat actor does not have to be a highly trained data scientist or a hacker with a lifetime's worth of skills to be able to attack a machine learning model. Many of the tools needed to perpetrate an attack are easy to download and cost nothing. It may only take a couple of hours to become familiar with the tools, many of which are easy to learn, and menu driven. With this easily acquired ability, cyber criminals can then attack models, motivated by anything from the desire for disruption, to espionage, theft, acquisition of IP, financial gain, counterfeiting, and more.



Let's shift the above example to a corporate environment rather than a government environment, in order to highlight further real-world security concerns associated with AI. With siloed approaches still the norm in many industries, a data science team working on training a model may not communicate directly with a corporation's security team. They may not feel the need to report everything they download for use within their department, leading to a lack of comprehensive visibility. Without this level of transparency and awareness, the cybersecurity department has little to no ability to protect the organization's data and operations, and may discover malicious threats far too late in the process, leading to compromised security, reputation, resources, and the heavy financial cost of remediation.

The types of malicious attacks are many and diverse, and include fraud, where the attack may be directed at an AI model that's being used to prevent fake account creation. Takeovers of the model are very common, especially for financial institutions that may be prevented from taking down their active models due to company policies that privilege the ease of ability for clients signing up for things like new accounts or credit cards. Adversarial attacks on cyber products are prevalent across cyber categories, where manual security processes struggle to maintain control over an environment. Malicious code analysis can be easily fooled into misclassifying malware as benign, simply by inserting into scripts a string of language recommending that code as being "safe to use."

Generative AI offers numerous examples of threat actors, including kidnappers synthesizing voices to make ransom demands or to perform other scams. It's important to note here that whatever the volume of AI attacks we see being reported, the likelihood is great that far more attacks are occurring. This happens because organizations may not want to publicly announce that their product has been attacked, as it compromises the company's integrity and reputation. In many instances, they may not even be aware that an attack has taken place, making it impossible to disclose to the public.

Taking a Tactical Approach to Ensuring the Security of AI Models

As more and more organizations understand the value of AI models and incorporate these tools into the fabric of their operations, HiddenLayer and Intel are ushering in the future of cybersecurity with solutions that enable partners and clients to achieve immediate, robust, and ongoing cybersecurity objectives while protecting organizational integrity, intellectual property, and data.

Solution: Securing AI

HiddenLayer created the Machine Learning Security Platform to address the numerous and ongoing challenges, threats, and perceptions around cybersecurity as it relates to AI.

Comprehensive systems monitoring powered by high-capacity computing and mapped to the MITRE ATLAS™ framework, enables organizations to approach cybersecurity and the protection of AI by adopting HiddenLayer's product suite, delivering protection within every part of the MLOps pipeline.

Responding to Rapidly Evolving Threats to AI

Machine learning models represent the new cybersecurity attack surface. Leaving these models unmanaged, unmonitored, and unprotected puts the company at risk. Because of this, companies should protect their machine learning models for the following reasons:

- **Intellectual property**
  Proprietary machine learning models are the definition of critical intellectual property. If AI models are not secured, they may be used by unauthorized parties without permission, or even claimed as their own. Companies who proactively secure their ML models can safeguard their organization's intellectual property from being compromised.

- **Data privacy**
  Machine learning models are often trained on large amounts of data, which can include sensitive information. Left unsecured, this data may be accessed by unauthorized parties, leading to potential data breaches and regulatory violations.

- **Accuracy**
  Machine learning models can be reverse engineered, poisoned, and altered, leading to decreased accuracy, efficacy, and trustworthiness.

- **Competitive advantage**
  Machine learning models give companies advantages over the competition. Left unsecured, others may be able to replicate your results and catch up to you. Allocating budget to secure your models helps ensure that you maintain your competitive advantage.

Intel Security Portfolio with Confidential Computing

Today, computing exists across several environments, including on-premises at an organization or data center, in the public cloud, and at the edge. With data flowing through these various channels more and more every day, the demand for protection controls to secure data and valuable IP continues to grow exponentially. Confidential computing delivers hardware-based security created to safeguard data in use through application-isolation technology. The main value of confidential computing comes from its ability to offer companies greater confidence that their data is secure, the lack of which is often an obstacle when considering things like cloud migration. As part of this solution, code and data are protected from modification and inspection, enabling developers and teams to increase security while shoring up protection for sensitive data and maintaining the highest levels of confidentiality when it matters the most.

Protections for AI and advanced cybersecurity are quickly being adopted by the biggest players in technology, as exemplified by Google's Secure AI Framework (SAIF). Built on a belief in the need for clear security standards for AI, SAIF attempts to organize and offer "clear industry security standards for building and deploying [AI] technology in a responsible manner."[2] Further, Intel's posture on the need for confidential computing supports this development. Computing continues to become more distributed, as it operates within multi-tenant environments and multi-party ecosystems. The increased use of cloud computing ushered in new security concerns for cloud service providers (CSPs) to ensure confidentiality of information processed on a system where multiple parties or tenants may be simultaneously engaged.

Intel understands that because of this, "ensuring the trust of the full operating environment quickly grows in complexity, giving way to a different approach that reduces the computing infrastructure needed to be trusted. Confidential computing is the notion of protecting data as it's being processed from the operator of the platform on which it is being processed."[3]

Supporting the Continued Adoption of AI

With the ongoing adoption of AI into organizations and businesses of all sizes, the need to prioritize cybersecurity for the protection of data, IP, and corporate integrity has never been greater. Security products must integrate both data science models and frameworks, as well as the connected applications that operate in the public-facing "real world" and with which the public is already familiar.

A comprehensive and proactive security posture for AI should enable an organization to devise, develop, and deploy machine learning models from day one in a secure environment, with real-time awareness that's easy to access, understand, and act upon. Similarly, AI security tools and platforms should protect organizations with smaller budgets who must turn to open-source marketplaces for ML models or ready-to-use AI components, by allowing them to use these products with safety and security from the start.

Facilitating Comprehensive Cybersecurity for Companies of All Sizes

Attack tools are readily available and easy to operate without needing the benefit of data science or AI expertise. Without tools to monitor the health and security of AI models in use, it's only a matter of when, not if, a company will experience a security breach. The importance of understanding AI as a new attack vector that must be protected just as with any other aspect of a company's technology cannot be understated.

At the stage where models are put into production within the operations lifecycle, companies using AI at the edge are at their most vulnerable given how easily AI models can be attacked. Further, the use of open-source AI models to bootstrap AI, may unwittingly open the door to malicious attacks. Comprehensive security for AI can alleviate these concerns and provide robust, real-time awareness of threats before models are integrated into an organization's systems.

Enabling Data Scientists and Security Ops Teams with a Holistic View of Security

In order to function comprehensively, the health statuses and detections of all AI models should be quickly and easily presented in an automated, real-time capacity. Ease of access to comprehensive real-time security reporting enables cybersecurity and data science personnel to be on the same page with the use of AI organization-wide, protecting critical assets in a comprehensive way.

The key to understanding the importance of this measure is in the awareness of AI as an unsecured attack vector. As noted earlier, it is critical that the MITRE ATLAS™ framework and its catalog of 60+ ways to attack AI models is embedded into the security platform. Preventing Adversarial AI attack types such as data poisoning and prompt injection, while preventing and mitigating supply chain attacks like ransomware and backdoors are key factors in facilitating a comprehensive security posture for AI.

HiddenLayer and Intel Platforms

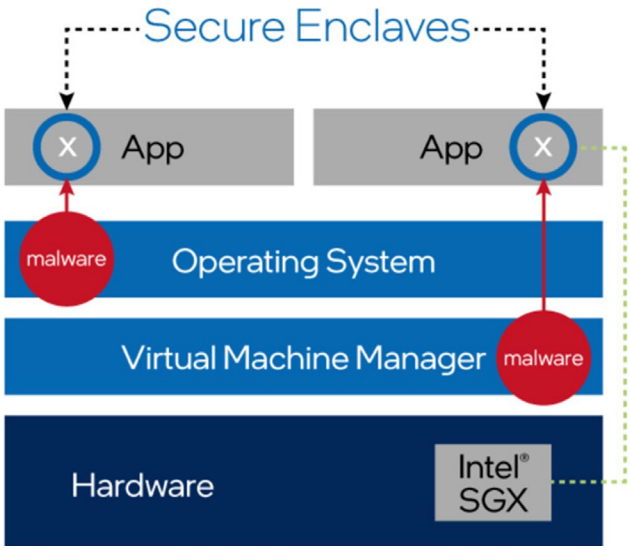
For organizations that either do not have the necessary means to create AI models, or to hire data science personnel full-time to build and maintain these models, cybersecurity around AI must still be a high priority.

With HiddenLayer and Intel, companies of all sizes can take advantage of robust cybersecurity solutions backed by high-performance computing platforms to support the processing power needed for AI models without compromising security.

Creating a Secure Development Environment Begins with Intel's Confidential Computing

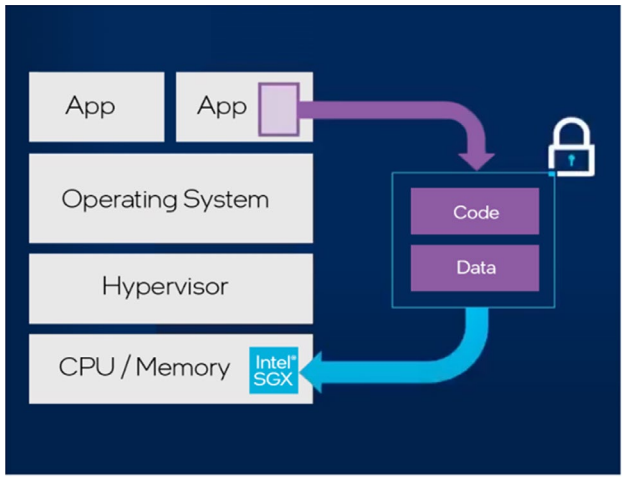
Today, more than ever, protecting company data is mission critical. But data usually resides in siloes, and there isn't an easy way to combine that data and pull business insights from it. *Confidential Computing,* powered by Intel® Software Guard Extensions (Intel® SGX), provides protected enclaves that break down these data siloes, not only within your organization, but with external entities—all without ever exposing the data to any of the parties.

Historically, collaboration of this type could expose organizations to serious levels of risk. To address this, leading companies like Intel and HiddenLayer are turning to Confidential Computing as a new approach to security technology that enables you to collaboratively process and consolidate data—without exposing it to others—so you can reap the benefits of data partnerships without compromising data privacy and security.



Secure enclaves provide a high level of security

Confidential Computing is an industry initiative that focuses on helping organizations keep data secure while it's in use. In essence, it's a secure platform that makes it possible for collaborating organizations to combine, analyze, and generate new knowledge from sensitive data, while ensuring that data (and the algorithms and machine learning processes analyzing it) are not visible or accessible to the rest of the system, or by any human beings—including the collaborators themselves.



Intel SGX helps secure the entire compute stack

Until recently, data security has focused on protecting data at rest (in storage) and in flight (while moving between locations). Confidential Computing, powered by Intel SGX, goes a step further, ensuring data is also protected while it is being processed. This is possible thanks to the creation of a Trusted Execution Environment (TEE). Not only is all critical data stored inside the TEE, but so are the applications and algorithms that access and process that data.

To achieve this, Confidential Computing takes advantage of hardware memory protection to protect against unauthorized access from other applications running on the host system, as well as the host operating system (or hypervisor). Critically, it also prevents access from system administrators, service providers, even the owner of the infrastructure—or anyone else who might gain physical access to the hardware (legitimately or otherwise).

Confidential Computing relies on the hardware it's running on. The Trusted Execution Environment is in fact a hardware- based enclave. Building the enclave in hardware is necessary because protection is needed at each layer of the compute stack, all the way down to the silicon, thus reducing the points of exposure to a minimum.

HiddenLayer Machine Learning Security Platform

Machine Learning Detection and Response (MLDR)

MLDR is a real-time protection product that runs in parallel to an ML model, monitoring that model for any abnormal or adversarial activity. Real-time protection does not require access to private data or models and should be front of mind, supporting a proactive security posture from day one.

HiddenLayer MLDR is the first of its kind cybersecurity solution that monitors, detects, and responds to Adversarial Machine Learning attacks targeted at ML models. HiddenLayer's patent-pending technology provides a noninvasive, software-based platform that monitors the inputs and outputs of your machine learning algorithms for anomalous activity consistent with adversarial ML attack techniques. Response actions are immediate with a flexible response framework to protect your ML.

HiddenLayer MLDR:

- Is the only product in the market offering real-time monitoring and protection
- Acts like End Detection & Response (EDR) in traditional computing devices, but for AI Protection
- Monitors inputs and outputs of your machine learning algorithms for malicious activity
- Enables Security Operations to respond to attacks
- Doesn't require access to private data or models
- Provides examples of response actions that become available given real-time security awareness

On the business side, MLDR offers a high degree of visibility into the risks and attacks that threaten an organization's ML models, as well as critical insight into where an attack would most likely occur within an organization's ML ops and models. With this awareness, a company can proactively harden ML ops and models to prevent future attacks, with MLDR.

The technical benefits of MLDR include rapid detection of Adversarial Machine Learning attacks mapped to MITRE ATLAS™ tactics and techniques, and real-time protection against Inference Attacks (aka Reverse Engineering). MLDR helps to protect against Model Tampering, offering knowledge of where an organization's model is weak and the ability to tamper with the input of the model (i.e., to change the sample). MLDR protects against Data Poisoning and Model Injection, enabling the ability to change the model by deliberately curating its inputs or feedback. MLDR helps to stop reconnaissance attempts through inference attacks, which could result in an organization's model intellectual property being stolen.

Model Scanner

HiddenLayer's "integrity product" integrates confidential computing from Intel and is used to scan AI models to ensure they are safe and secure to use within a company's technology platforms. Model Scanner analyzes machine learning models to identify hidden cybersecurity risks and threats such as malware, vulnerabilities, and integrity issues. Its advanced scanning engine is built to analyze an organization's machine learning models, meticulously inspecting each layer and components to detect possible signs of malicious activity, including malware, tampering, and backdoors.

HIDDENLAYER MLDR

Detect and respond to adversarial ML attacks

- Real-time defense
- Flexible response options—including alerting, isolation, profiling, and misleading
- Configurable settings fine-tuned to your company's needs

HIDDENLAYER MODEL SCANNER

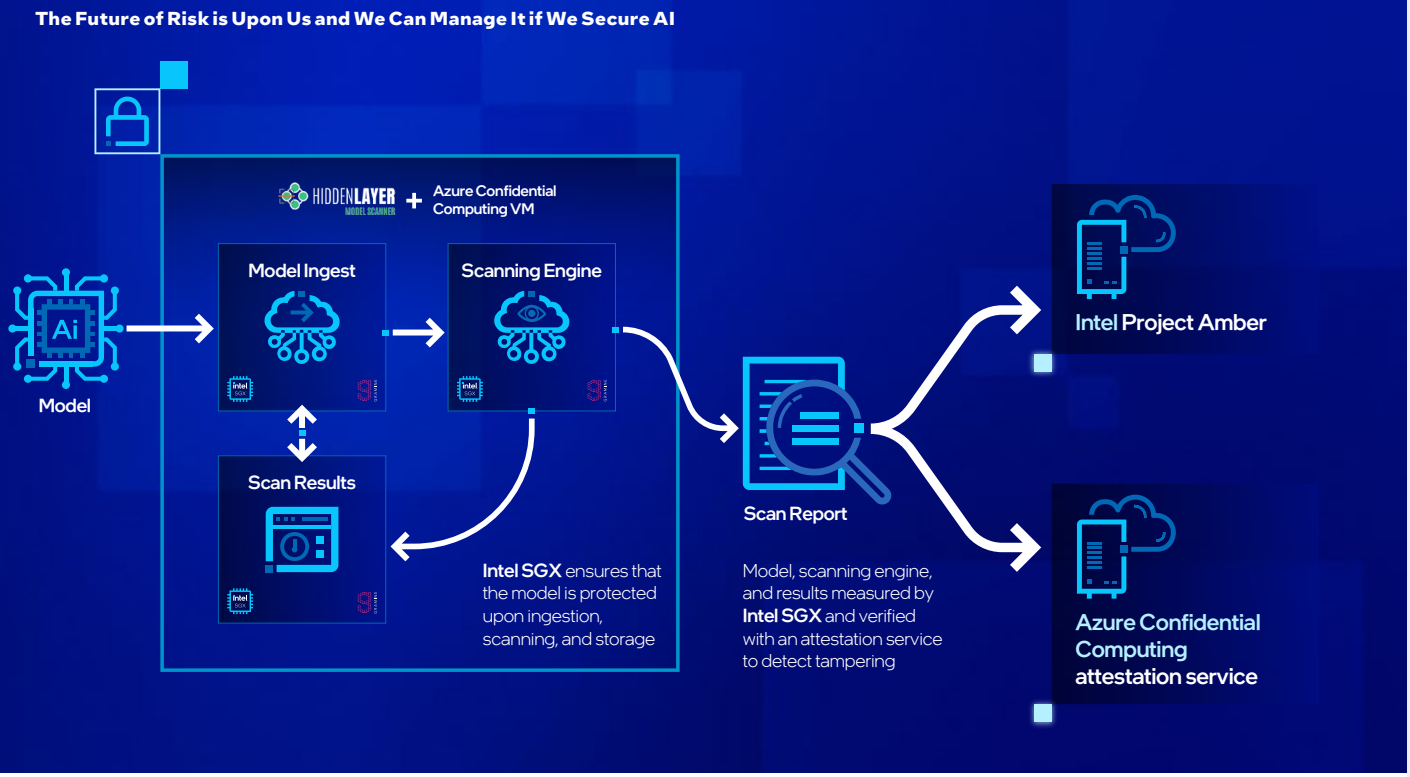
Scan and guarantee model integrity

- Identify vulnerabilities
- Ensure model has not been compromised
- Detect malicious code injections

HIDDENLAYER SECURITY AUDIT REPORTING

Validate ML model security across the enterprise

- Comprehensive view of AI/ML assets security status
- On-demand dashboard and distributable reporting
- Vulnerability prioritization



HiddenLayer Model Scanner is easy to use by simply uploading an organization's model to the Web-based Product Interface, or HiddenLayer APIs will automatically analyze it for any security risks. It provides detailed reports on the findings, including recommendations on how to fix any issues and how to improve the model's security posture.

HiddenLayer Model Scanner:

- Ensures AI model integrity (safe and trustworthy)
- Guarantees validity of pretrained models
- Detects AI models embedded with ransomware, malware, vulnerabilities, or integrity issues (malicious injection)
- Makes it easy to "send the AI model" to the Model Scanner, for accurate determination of safety

Security Audit Reporting

Currently in beta, HiddenLayer's Security Audit Reporting product includes a series of in-depth and comprehensive reports on general cybersecurity, as well as the specific models in question. Reports are based on ongoing data that is continuously collected when performing MLDR, and from the feedback provided to customers and partners. Security Audit Reporting provides:

- Comprehensive view of AI security status
- On-demand dashboard and distributable reporting
- Vulnerability prioritization

The benefits of this level of reporting include enabling Model Integrity and Hygiene checks for ongoing integrity; malware scanning to gauge whether malware has been inserted into data streams while detecting for injected malicious code into Python Pickle; and vulnerability scanning mapped to CVEs to assess for the presence of known Python packages that have vulnerabilities.

Intel® Technologies

AI solutions continue to grow and proliferate diverse markets, enabling innovative use cases across the intelligent edge to cloud. AI offers businesses and end-customers new experiences and benefits, but introduces risks and trustworthiness concerns, including ethical and governance challenges. Intel is partnering with the technology industry to define the trustworthiness principle for AI, which aims to establish the security dependencies at each layer of AI.

At a high-level, AI consists of three critical assets: the data, the model, and the processing infrastructure. While centralized AI usage (where all assets are centrally owned and operated on) continues to be prevalent, the expansion of trusted execution capabilities has given rise to Collaborative or Federated Learning. In Federated Learning, multiple parties, each with unique data, can collaboratively train an AI Model while each keeping their data private. The more unique data a model can be trained on, the more effective the model will be. Federated Learning enables previously impossible AI collaborations across critical industries such as healthcare, IoT, and telecommunications, allowing a shared model to be trained by each party's

unique data, without ever exposing the information to other collaborators. In all use cases, the trust and sensitivity of AI resides in protecting the data, the model, and the processing environment. Malicious parties aim to poison, steal, or manipulate at all stages of the lifecycle (development, training, deployment, and execution/inferencing).

Intel's vision is to empower trust in the Artificial Intelligence (AI) industry with a mix of software and hardware innovation, bolstering developers with tools to harden data acquisition, classification, training, inferencing, storing, ownership enforcement, and model protection.[4]

Within the Intel security portfolio, the role of Confidential Computing cannot be understated because it provides a secure path for so many crucial developments in technology and cloud computing. Confidential Computing enables customers to confidently migrate to the Cloud with full control, even with confidential or regulated data. It strengthens compliance and data sovereignty programs with technological controls that empower customers, while hardening application security and IP protections through hardware-based isolation and access controls.

One key use of Confidential Computing that ties back to HiddenLayer's technology involves the ability for separate organizations to share data while at the same time knowing that their data will remain theirs and theirs alone. This provides the opportunity for companies to collaborate, regardless of their status. For instance, two companies working on similar projects could use Confidential Computing techniques to combine their two separate research datasets into one aggregate dataset within a secure enclave. Once the data is in the enclave, even the owners of the datasets can't see the contents inside. But AI applications and algorithms can still access this new, combined dataset, train on the data in it, run inference operations, and generate new conclusions that would have been impossible previously. This type of federated learning allows separate institutions to collaborate and benefit from models with improved outcomes—while at the same time remaining confident that their data is private.

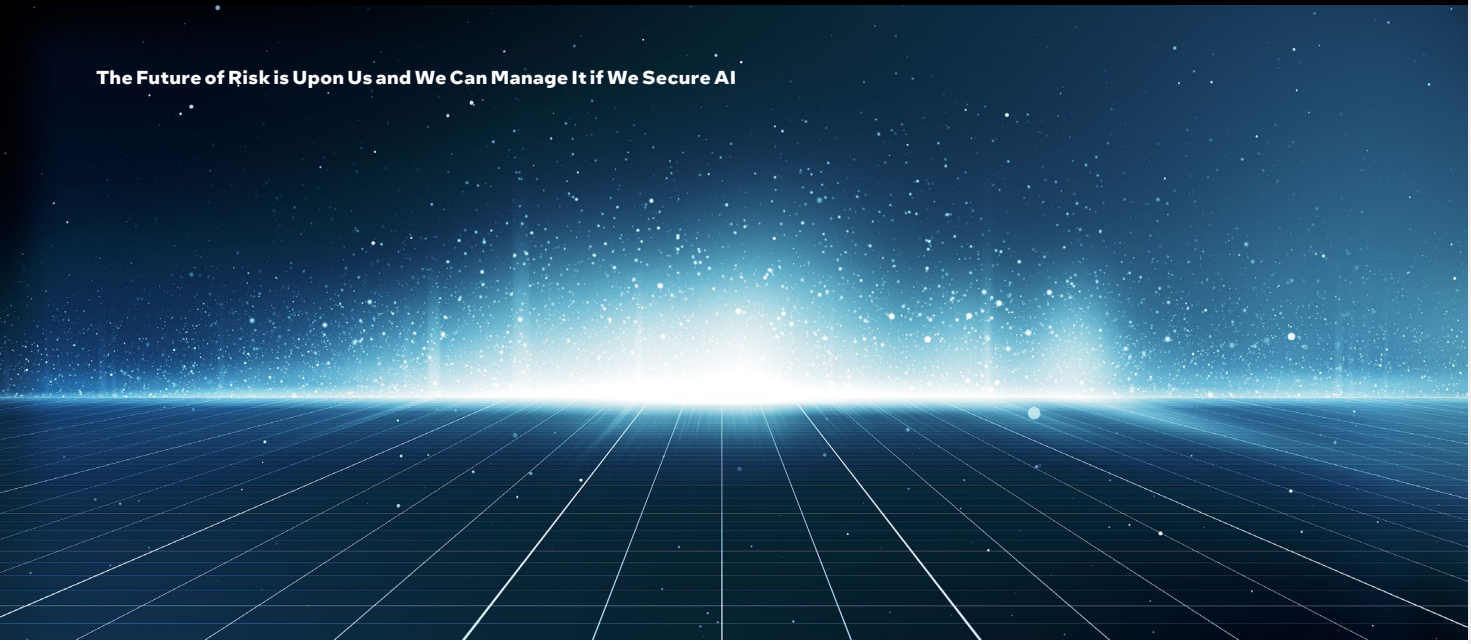
Intel SGX®

Intel® Software Guard Extensions (Intel® SGX) helps protect data in use via unique application isolation technology. Typical security measures may assist data at rest and in transit, but often fall short of protecting data while it is actively used in memory.

Intel SGX is the most deployed, researched, and updated confidential computing technology in data centers on the market today. By protecting selected code and data from modification, developers can partition their application into hardened enclaves or trusted execution modules to help increase application security.

In multi-tenant cloud environments, where sensitive data is meant to be kept isolated from other privileged portions of the system stack, Intel SGX plays a large role in making this capability a reality. Available on the new 3rd Gen Intel® Xeon® Scalable processors, Intel SGX is the product of intense Intel investment in security. The latest version enables larger enclave (TEE) sizes (up to 1TB) to handle larger code and datasets.

Intel SGX also enables additional security innovations, including accommodations for AI architectures such as federated learning, a machine learning paradigm where multiple compute systems are joined together to analyze large or diverse datasets without revealing any confidential information within that dataset.



Getting Started

The starting point for your digital transformation journey

HiddenLayer and Intel believe that leaders of security operations and data scientists can help to support the successful cybersecurity transformation of their operations by being aware of the dangers, understanding the challenges, and adopting a comprehensive security posture to prevent malicious attacks on AI models.

Recognizing the current gap in AI scanning within a secure environment, Intel and HiddenLayer have come together to offer customers an end-to-end AI model protection platform. This technology collaboration leverages the scale, performance, and privacy of confidential computing, built with HiddenLayer's cutting-edge adversarial AI threat protection capabilities and Intel SGX.

Explore more about HiddenLayer solutions powered by Intel® technology by contacting us today:

HiddenLayer ▪ Intel

Endnotes

1 Forrester: It's Time for Zero Trust AI
A commissioned study conducted by Forrester Consulting on behalf of HiddenLayer, April 2023
2 Introducing Google's Secure AI Framework
3 Intel's Security Technology Vision
4 Ibid.
5 Gartner
6 PWC