Venafi

The Generative AI Identity Crisis:

Emerging AI Threatscapes and Mitigations

Enterprise adoption of generative AI (GenAI) is surging. So are the risks.

New GenAI tools are cropping up daily, promising to provide better customer experiences, automate away tedious tasks, increase innovation and bolster competitive advantages. They are powerful tools, but due to their intricate composition, GenAI systems also increase cyber risk to business data, infrastructure and outputs.

In this eBook, we'll explore two emerging GenAI threatscapes:

1. How are threat actors using GenAI for their own nefarious purposes?
2. What types of adversarial threats are targeting GenAI systems?

Keep reading to find out the answers. Then, discover the critical role machine identity management plays in securing GenAI in your enterprise.

90%

of organizations are exploring and experimenting with GenAI across departments and business units.*

***Live survey conducted during "3 Risks Adversarial Machine Learning Poses to Your GenAI Systems"**

A Primer on GenAI

What is generative AI, and how do these systems function?

GenAI systems generate new text, photo, audio or video based on user context (prompts) and vast training datasets—sometimes with data swaths as large as the internet itself!

Due to this scale, enterprises don't typically manage their own models. Instead, they rely on third-party systems, which they integrate into their networks and repositories to meet various use cases (e.g., coding, sales and marketing, and customer support).

What makes GenAI so complex?

The volume of data coupled with third-party integrations and APIs magnify GenAI complexity and risks. What's more, there are multiple data streams flowing through a GenAI system at any given time, and all of them must be thoroughly secured.

How Threat Actors Use GenAI

Businesses aren't the only ones capitalizing on GenAI systems. Threat actors already use them for their own sinister operations.

RECONNAISSANCE:

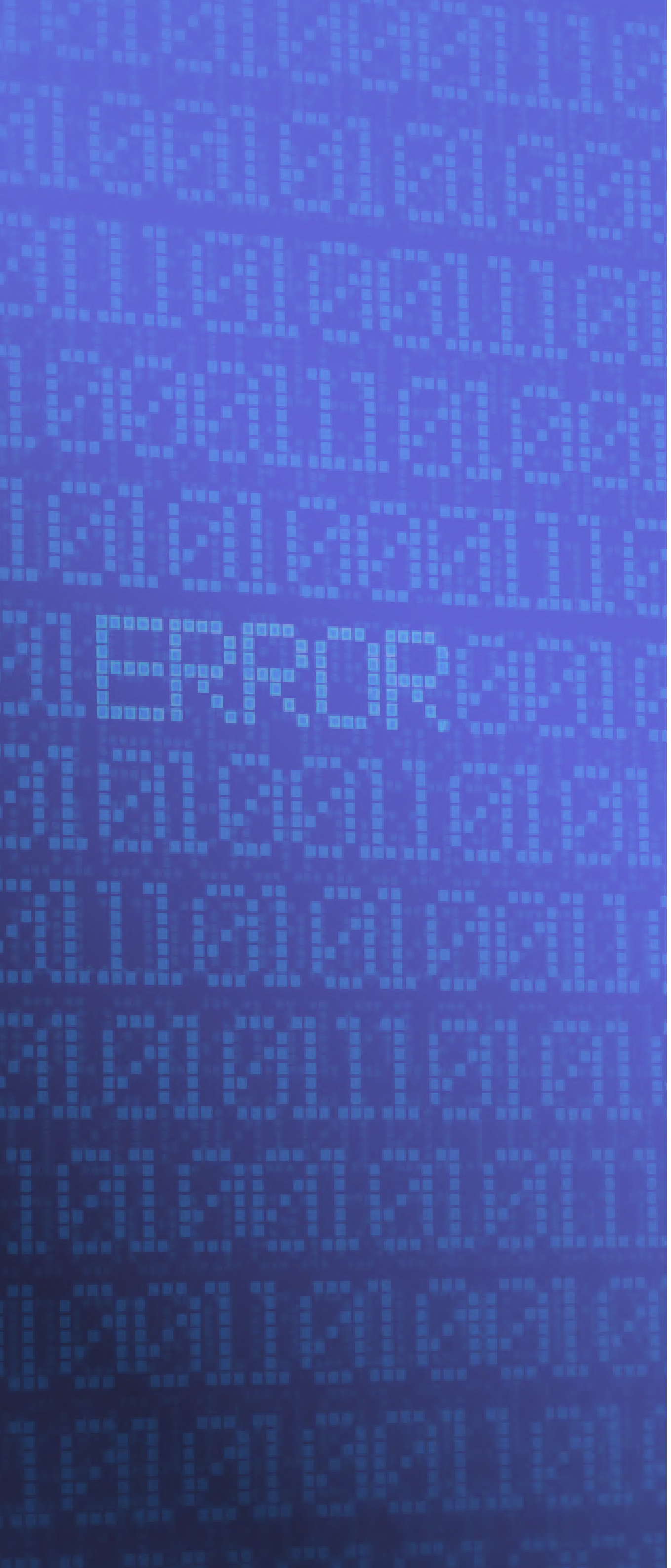
Adversaries can gather intel on companies, employee directories, investor information, executive social media profiles and more. If a GenAI is equipped with real-time web access, almost any data on the open internet is fair game.

PHISHING AND DEEPFAKES:

Attackers can generate sophisticated spear-phishing emails, build audio and video deepfakes, and fabricate social media profiles, at scale, in minutes.

MALICIOUS CODE/ REVERSE ENGINEERING:

GenAI allows threat actors to develop malware, even complex, polymorphic strains. They can also reverse engineer legitimate programs to uncover code patterns or extrapolate source code for use in future exploits.



Top 3 Risks to GenAI Systems

NIST has defined three categories of threats to GenAI systems.

1. **Integrity:** The biggest risk to GenAI, targeting overall trustworthiness, accuracy and reliability.
2. **Availability:** GenAI components can be taken offline or their performance degraded.
3. **Privacy:** GenAI composition brings up a lot of privacy concerns around training data and systems.

In the following section, we'll explore each of these threat types in detail.

Integrity

The number one integrity violation for GenAI is a **data poisoning attack:** the injection of harmful content during the training of an AI model. These attacks can be **broad-based,** impacting all data samples, or **more targeted** to smaller sample bases.

Best practices:

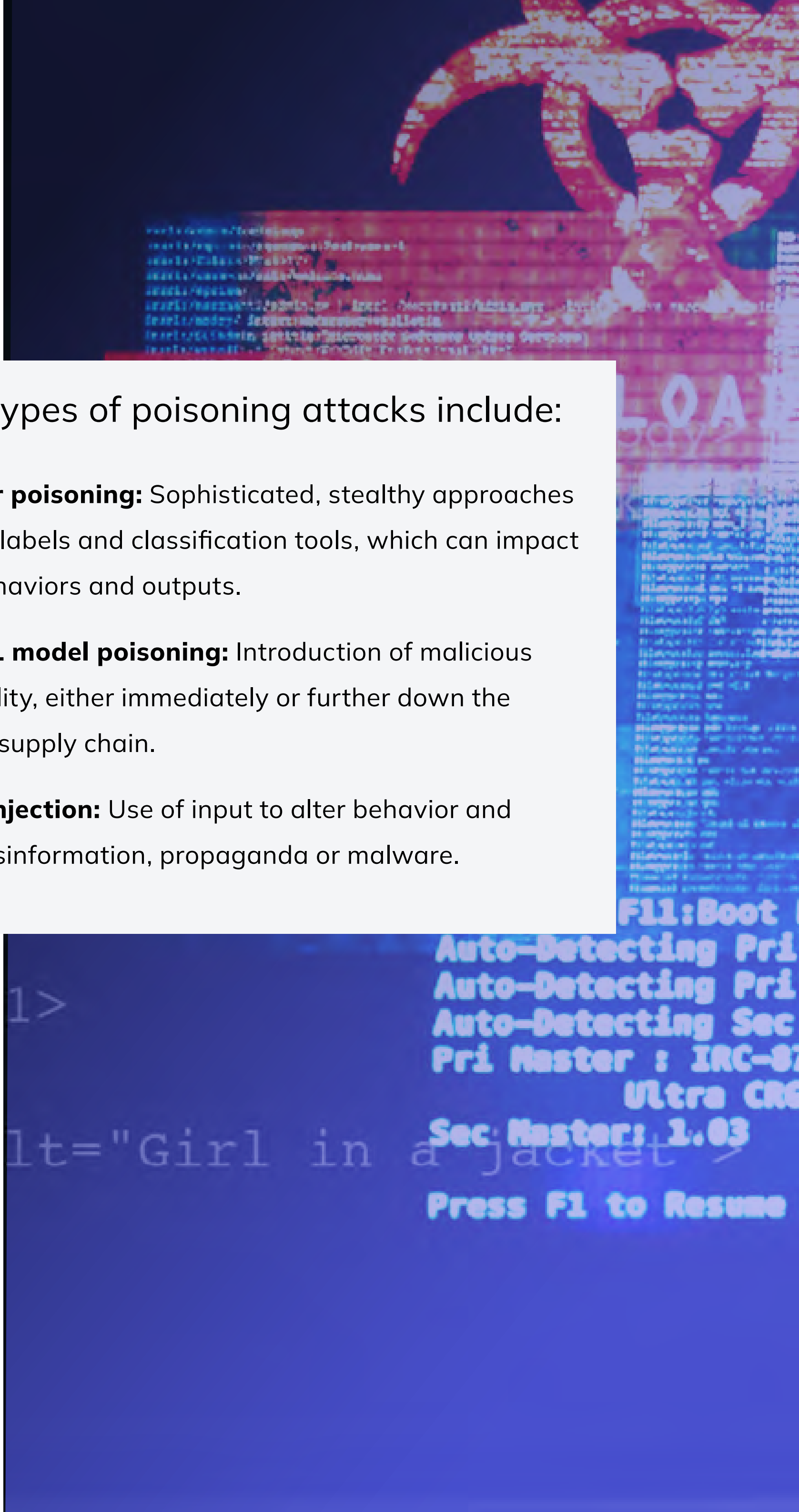
- Prevent unauthorized code across the AI software supply chain.
- Protect credentials like OAUTH, API keys, code signing certificates and SSH keys.
- Inspect models and inputs/outputs; sanitize and purge data.

Other types of poisoning attacks include:

Backdoor poisoning: Sophisticated, stealthy approaches targeting labels and classification tools, which can impact model behaviors and outputs.

Direct ML model poisoning: Introduction of malicious functionality, either immediately or further down the software supply chain.

Prompt injection: Use of input to alter behavior and create misinformation, propaganda or malware.

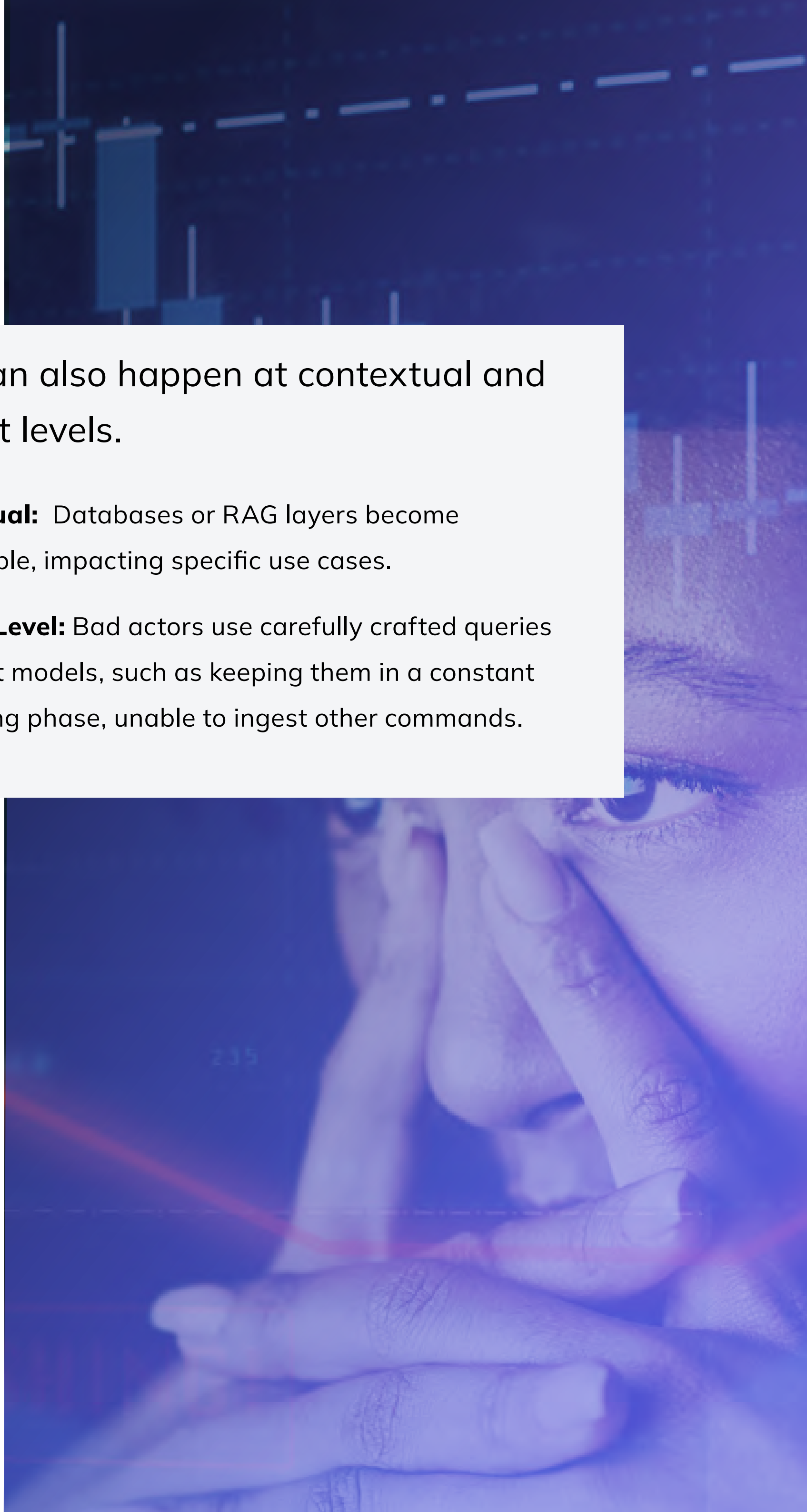


Availability

Model denial of service is the most common availability violation, degrading a model's performance or even bringing it to a halt.

Best practices:

- Traditional DDoS mitigations: Determine normal/abnormal traffic; employ firewalls.
- Load balancing: Distribute traffic across servers to improve availability and user experience.
- Redundancy and failover mechanisms: Maintain cohesion with a backup/failover plan.
- Rate limitation management: controlling resource usage per instance; limit the number of requests made in an allotted amount of time.



This can also happen at contextual and prompt levels.

Contextual: Databases or RAG layers become unavailable, impacting specific use cases.

Prompt Level: Bad actors use carefully crafted queries to impact models, such as keeping them in a constant processing phase, unable to ingest other commands.

Privacy

Privacy violations involve the reconstruction and identification of model components or training data, which can result in model theft. This puts model data, which sometimes includes customer names or related info, at risk.

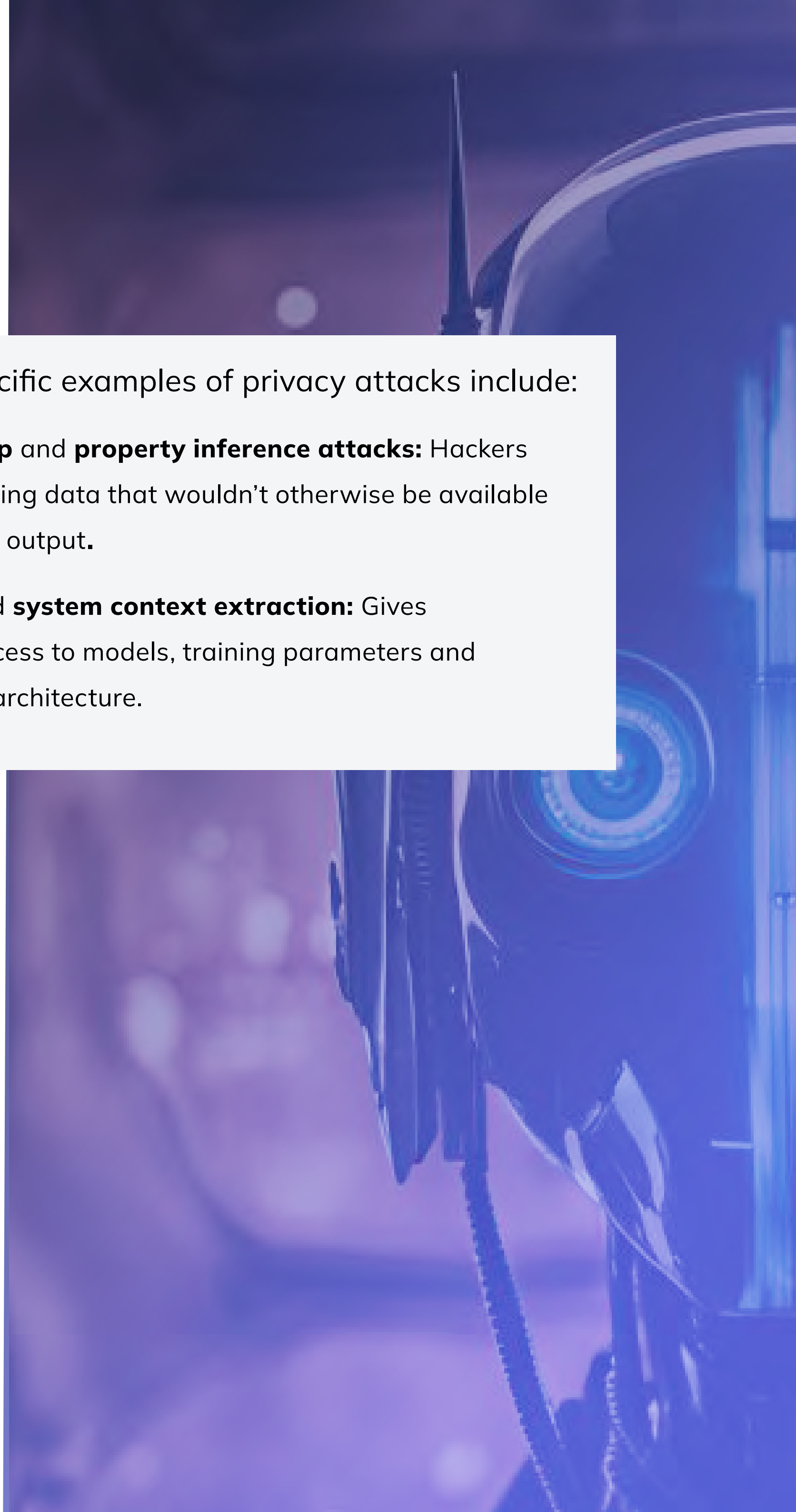
Best practices:

- Differential privacy: Prevents adversaries from determining if data is included in a model.
- Multi-party computation: Allows multiple parties to jointly compute over private input without revealing those inputs to each other.
- Watermarking: Digital fingerprinting that enables traceability of outputs, all the way back to a source.

Some specific examples of privacy attacks include:

Membership and **property inference attacks:** Hackers access training data that wouldn't otherwise be available in a model's output**.**

Prompt and **system context extraction:** Gives hackers access to models, training parameters and underlying architecture.



“Having the ability to turn off generative AI or our machine learning model with a kill switch or a big red button [is] important. We’ve got a switch to turn off the gas. We’ve got a switch to turn off the electricity. Everything that can potentially either cause harm or cause disruption to the workflow comes with a big red button. This type of powerful technology needs to have that."

Kevin Bocek,
Chief Innovation Officer

The Role of Machine Identity Management in GenAI Security

GenAI threats show a critical need for machine-to-machine authentication. Done effectively, authentication acts as a swift, simple **“kill switch”** for misbehaving or erratic AI systems.

It operates at every level discussed: data pipelines, AI models and their actions.

Authenticating GenAI Systems

Inputs

Earlier, we discussed the vast amounts of data needed to train a model, and the numerous pipelines where information flows into the model. At these interception points, data flows must be authenticated. Machine identities, like digital certificates and cryptographic keys, provide the foundation for this authentication.

Models

GenAI models and plug-ins act as machines, which are code. To secure your GenAI software supply chain, you must be able to approve plug-in operation, database access and when/how fine-tuning occurs. This happens through code signing. And, through a robust code signing trust chain, you can prevent the execution of any unauthorized code.

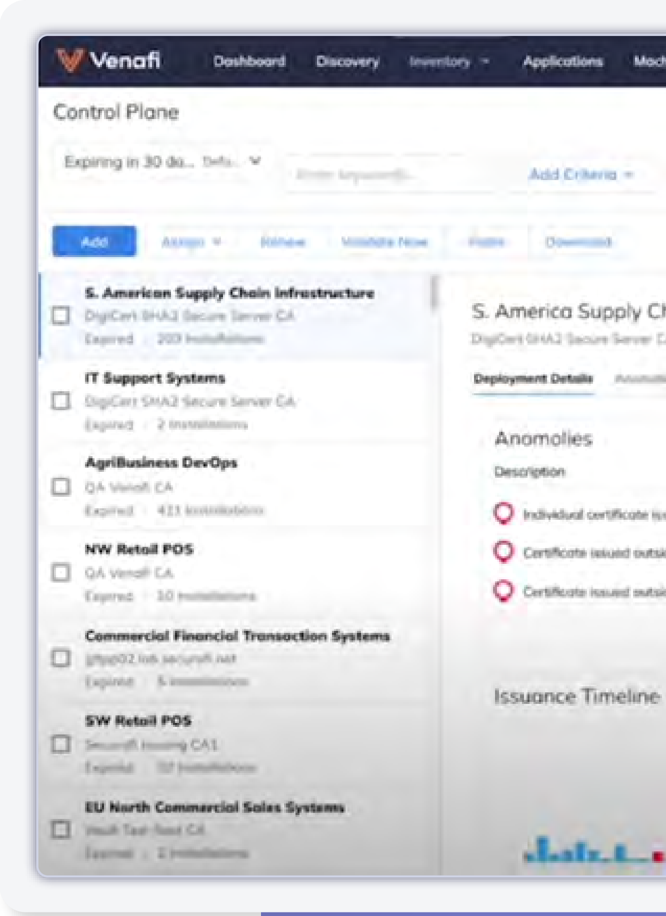
Outputs

You must continuously verify what LLMs can and can't do. Models will interact with other machines—in some cases of their own accord—and need to authenticate outward as well, such as in the case of APIs.

The Necessity of a Control Plane for Machine Identity Management

Enterprise-wide machine identity management can help you and your team fully authenticate and authorize GenAI systems and makes it easy to deactivate components that start behaving in a way they aren't supposed to.

By providing comprehensive visibility and automation of every machine identity, a control plane helps you quickly identify all unique versions and instances of GenAI systems in use. If one specific instance of a particular version starts acting strangely—or outside its predefined parameters—you can easily "pull the plug."



Control Plane

Expiring in 30 da... Defa...

☐ S. American Supply Chain Infrastructure
DigiCert SHA2 Secure Server CA
Expired · 203 Installations

☐ IT Support Systems
DigiCert SHA2 Secure Server CA
Expired · 2 Installations

☐ AgriBusiness DevOps
QA Venafi CA
Expired · 421 Installations

☐ NW Retail POS
QA Venafi CA
Expired · 10 Installations

☐ Commercial Financial Transaction Systems
[illegible]
Expired · 5 Installations

☐ SW Retail POS
[illegible] CA1
Expired · [illegible] Installations

☐ EU North Commercial Sales Systems
[illegible] CA
Expired · 2 Installations

S. America Supply Ch
DigiCert SHA2 Secure Server C

Anomalies

Description

Individual certificate is
Certificate issued outsi
Certificate issued outsi

Rely on Venafi to Secure Your GenAI Systems

Venafi created the machine identity management category, and time and time again, we're the preferred choice for securing machine-to-machine connections. Even across intricate GenAI systems.

Key Components for GenAI System Protection

Control Plane for Machine Identities: Centralizes and automates control of every machine identity, no matter the type or location.

Stop Unauthorized Code Solution: Reduces your attack surface by preventing unauthorized code execution across any environment.

Venafi Athena: GenAI that provides orchestration advice, optimization tips and integration code.

Parting Thoughts

GenAI is a powerful, burgeoning technology that presents several rich opportunities for your business. But with this technology, new threatscapes have emerged, both in the form of new weapons for threat actors and novel vulnerabilities to GenAI models, data and outputs.

To effectively leverage GenAI, and in addition to traditional security approaches, you must implement comprehensive machine identity management. That way, you'll always know what your AI systems are doing—and you can mitigate any erratic behavior, should any arise.

Looking to secure your GenAI systems?
Get started with centralized, automated machine identity management today.

Learn more about the Venafi Control Plane