CYE

eBook

The CISO's Guide to **Uncovering & Mitigating GenAI-Driven Threats**

September 2025

Table of Contents

Introduction: The Urgency of AI Risk

The rise of GenAI has introduced new technologies, new possibilities, and new exposures. From LLM-powered chatbots to machine learning models built into core systems, AI adoption is accelerating across every industry.

The problem is that most organizations are moving faster than their security controls. That is why AI risk is so urgent.

In fact, organizations are increasingly seeing a new attack surface rising from the adoption of AI from coding, research, content creation, shadow apps, support agents, and so much more. One thing is certain – it is rapidly evolving on a continuous basis, mostly in an ungoverned way. One of the most overlooked risks is the everyday use of public GenAI tools like ChatGPT or Gemini. Employees may unknowingly paste proprietary code, internal documents, or even sensitive credentials into public LLMs that store and process this data. Without proper policies, organizations may be exposing intellectual property or confidential assets without realizing it.

Leading providers offer enterprise-grade LLM options that don't store prompts or responses—but most companies haven't yet made the shift. In fact, as of Q4 2024, only 20%–25% of organizations have adopted private or customized LLMs, according to Deloitte. That leaves the vast majority still relying on public tools.

This guide explores four GenAI/LLM threats. The first two threats exploit social engineering tactics, while the other two are brand new – made possible by the introduction of LLMs. You'll learn what they are, how they're already being exploited, and what to do about them before they hit your bottom line.

As of Q4 2024, only

20%–25%

of organizations have adopted private or customized LLMs, according to Deloitte.

That leaves the vast majority still relying on public tools.

A multi-national firm lost

$25 million

after attackers use a deepfake video to impersonate its CFO, highlighting that without strict verification controls, organizations are highly vulnerable to AI-powered fraud.

Case in Point: $25M Lost to Deepfake

In early 2024, a finance worker at a multi-national firm in Hong Kong received a video call from what looked like the company's chief financial officer. The CFO asked the employee to urgently wire more than $25 million across multiple accounts.

The employee complied. Everything about the call appeared legitimate: the voice, the face, and the sense of urgency. It wasn't until the damage was done that the truth emerged: The CFO never made the call.

The entire meeting was staged using deepfake technology. Every person on the video call, including the CFO, was AI-generated. The attackers had used publicly available footage and data to recreate executive identities and carry out the heist in real time.

This wasn't just a phishing email or a spoofed domain. It was full-scale AI-powered fraud, and it worked because no one was prepared to question what looked "real".

This kind of fraud would never have succeeded if basic verification controls had been in place. Simple protection, like multi-factor authentication or requiring written confirmation via a separate channel—could have blocked the transfer, no matter how convincing the video call looked. This case is no longer science fiction; it's a template. And if your organization hasn't updated its policies, verification procedures, or response protocols to account for deepfakes, you may be just as vulnerable.

AI Exposure 1

Deepfakes and Executive Impersonation

This first exposure directly reflects the attack scenario in the $25M deepfake CFO case. Deepfakes are no longer theoretical—they are actively being used to defraud companies at scale. The example you just read isn't an outlier; it's an early signal of a growing trend that security teams must urgently address.

Deepfakes don't just manipulate content; they exploit trust. Today's deepfakes can impersonate executives in real-time video and audio, enabling attackers to authorize transactions, mislead employees, and bypass verification protocols.

Yet most organizations have no deepfake-specific controls, no internal training, and no incident playbook for this class of threat. In fact, according to Regula's 2024 Deepfake Trends report, 66% of business leaders say their organization is unprepared to detect or respond to deepfakes, yet 59% consider video deepfakes a serious threat to their operations.

The risk isn't limited to organizations. Public perception is also vulnerable: In a recent WBZ-TV I-Team investigation, more than half of participants failed to identify a deepfake video, highlighting just how easily trust can be exploited.

Here is why deepfakes are so dangerous:

They target people, not systems

They mimic the exact individuals that teams are trained to trust

They require almost no technical sophistication to deploy

Why are we so unprepared for these threats?

First, employee awareness training seldom focuses on deepfakes. In addition, there is a lack of tools that can detect manipulated video or audio, and multi-channel identity verification processes are often weak. Finally, there are usually no incident response protocols specifically for impersonation-based attacks.

Meanwhile, the potential impact is substantial and can include financial fraud, unauthorized transfers, reputational harm, regulatory scrutiny, and the accidental validation of false identities.

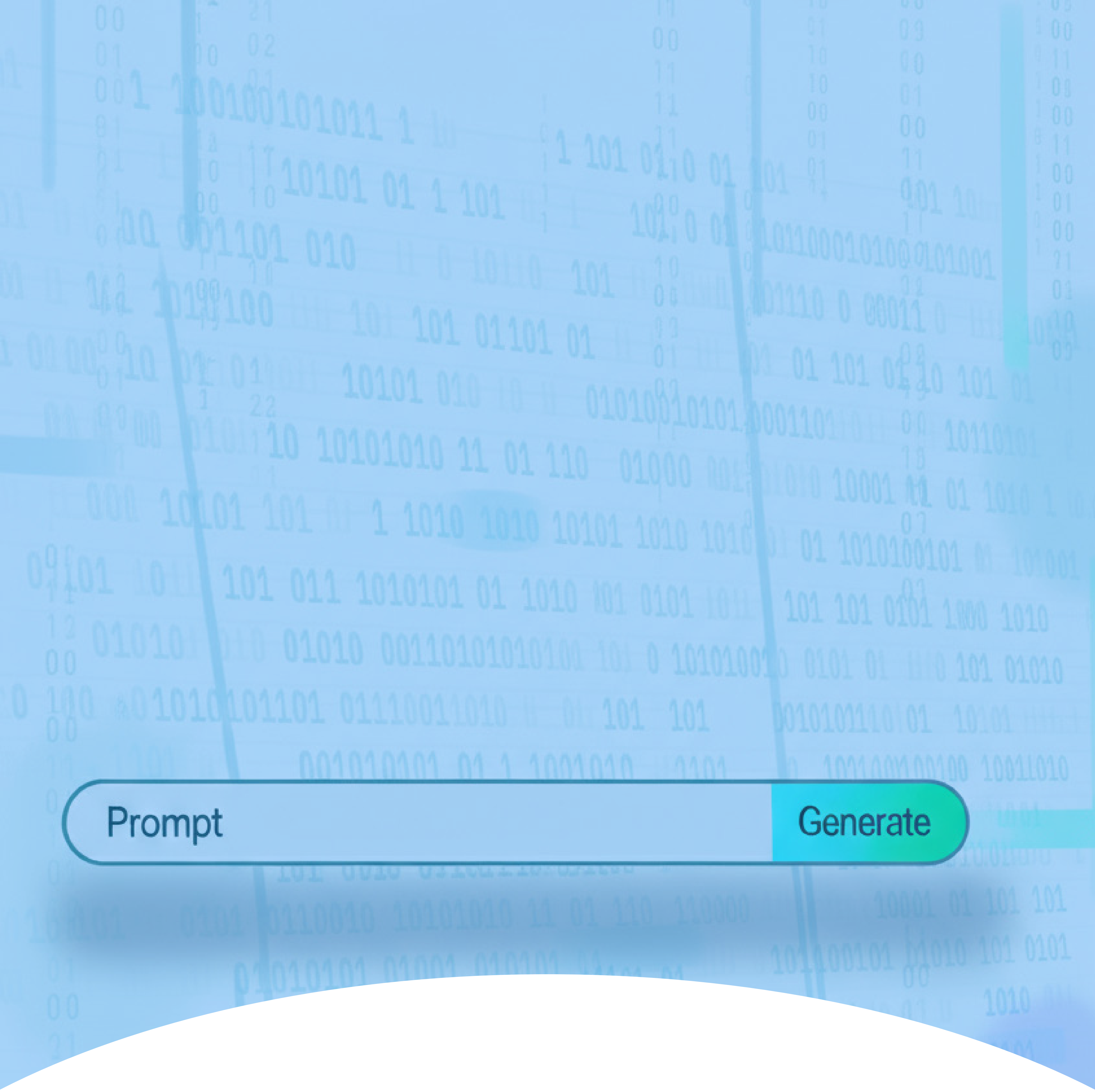
To prevent these risks, companies should enforce clear financial controls. For example, require a text confirmation from the CFO before wiring funds, because verbal approval can no longer be enough. Organizations can also purchase tools designed to detect deepfakes before they cause damage.

According to Regula's 2024 Deepfake Trends report

two out of three of business leaders say their organization is unprepared to detect or respond to deepfakes

What to do

Embed internal authentication processes in parallel to increasing awareness of the use of Deepfake technology and tips to identify its use



AI Exposure 2

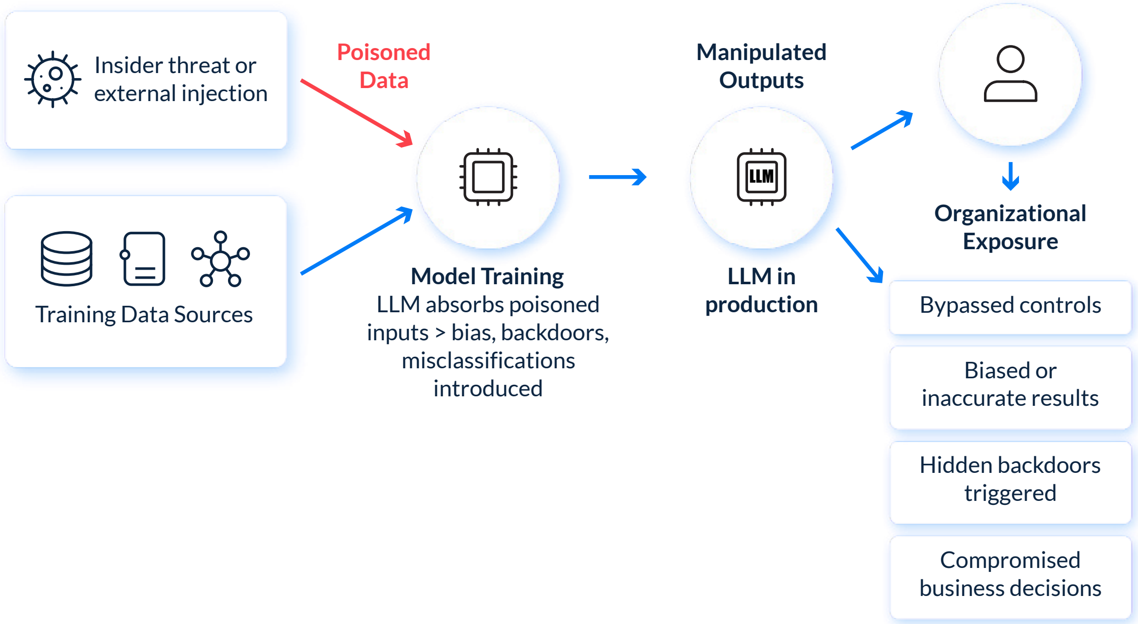
Poisoned AI / LLM Models

AI models are only as good as the data they're trained on. That's exactly what makes them a target.

The possibilities for abuse go far beyond trivia. Attackers can shape model behavior to bypass controls, inject bias, or manipulate decision-making. While large commercial models like ChatGPT, Claude, or Gemini are generally protected against this, the risk is real in internal development environments. The more likely scenario involves insider threats, where an employee with access to training data introduces malicious content that subtly shifts model performance. These changes are difficult to detect and can persist unnoticed for extended periods.

These attacks are nearly invisible, persist over time, and are extremely hard to reverse once deployed. The risk often goes undetected because training data is rarely audited or validated, and development environments lack standard security governance. Moreover, AI outputs are frequently trusted without understanding how they're produced, and there is no monitoring of model performance over time.

Poisoned AI/LLM Models


Insider threat or external injection
Training Data Sources
Poisoned Data
Model Training
LLM absorbs poisoned inputs > bias, backdoors, misclassifications introduced
Manipulated Outputs
LLM
LLM in production
Organizational Exposure
Bypassed controls
Biased or inaccurate results
Hidden backdoors triggered
Compromised business decisions

The consequences of model poisoning can include faulty decision-making, corrupted analytics, and biased or manipulated outcomes that carry legal and reputational risk. Even OpenAI had to pull back a model update in 2024 after it was found responding dangerously to self-harm prompts. Even the best-known providers can miss critical vulnerabilities without real-world testing.

To prevent this, organizations should establish company-wide policies for validating both the models they fine-tune and the datasets they use. That includes formal testing of third-party models and rigorous vetting of training sources. Security leaders can also bring in trusted external partners to stress test models, including through manual jailbreaking attempts. This is especially important before deploying AI tools into production environments.

Integrity isn't optional. It's critical.

AI Exposure 3

Everyday Use of Public LLMs

One of the most underestimated risks in the enterprise is the daily, unmonitored use of public LLMs like ChatGPT, Gemini, and others. Employees regularly interact with these tools for productivity, idea generation, or code assistance. But without clear guardrails, they may inadvertently share sensitive data such as intellectual property, internal strategy, customer details, or even credentials.

In early 2023, Samsung engineers accidentally leaked confidential source code and internal meeting notes to ChatGPT while using it to debug and summarize documents. The data was stored on OpenAI's servers, outside Samsung's control. As a result, the company restricted employee access to ChatGPT and began exploring private, secure alternatives. The incident sparked immediate concern over IP protection, competitive exposure, and regulatory implications.

Many public LLM platforms store prompts and responses by default, creating long-term exposure outside of the company's control. Moreover, security teams often have little to no visibility into who is using what tools.

The consequences of misuse of public LLMs can include loss of intellectual property to third-party systems, leaked credentials or internal logic reused in external environments, and legal or regulatory consequences for mishandling sensitive data.

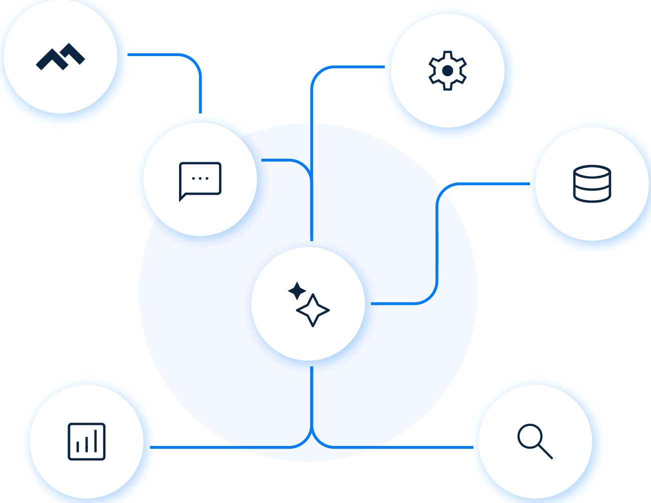
Fortunately, there are ways to mitigate this risk. It begins with establishing and enforcing acceptable use policies for GenAI tools, including training employees to understand the risks of sharing sensitive data with AI. Businesses can also invest in private or enterprise-grade LLMs that do not store or train on company data. Finally, it is important to monitor outbound traffic to detect potential leakage to public AI endpoints.

This responsibility should sit with the security operations team, in coordination with IT and compliance. Monitoring can include configuring data loss prevention tools, setting up egress filtering at the network level, and using cloud access security brokers to track AI tool usage. Companies can also deploy AI-aware traffic inspection that flags prompts containing sensitive keywords or metadata. The goal is not to block AI entirely, but to ensure it is used safely, with full visibility and control.

The solution

Set AI use policies, train staff, use secure private LLMs, and monitor traffic to prevent data leaks.

Public Display



Private LLMs

AI Exposure 4

AI-Driven Phishing

Phishing attacks have always relied on human behavior. GenAI makes them faster, more targeted, and far more convincing.

Attackers are now using natural language generation and real-time data to craft phishing messages that look exactly like legitimate internal communications. These messages can target specific employees with context-aware content that feels urgent and authentic.

In early 2024, a leading Indian bank fell victim to a sophisticated AI-driven phishing attack. Cybercriminals harnessed generative AI to craft emails that closely mimicked internal communications, incorporating details from the CEO's writing style and referencing recent meetings. These emails directed employees to a counterfeit internal portal, leading to the compromise of login credentials and unauthorized access to sensitive systems. The breach resulted in the exposure of confidential financial data and disrupted online services for several days, posing significant risks to customer trust and the institution's reputation.

Traditional defenses are falling short because email filters are not trained to recognize AI-generated content, and user training is often based on outdated phishing patterns. The problem is exacerbated because AI tools can easily mimic tone, language, and structure of real messages.

Risks to the business can include:

- Stolen credentials
- Unauthorized access
- Data exfiltration
- Privacy violations
- Business disruption
- Financial fraud

To prevent this type of phishing, organizations need layered defenses, not just training. Mid-level employees with decision-making access should be trained to recognize highly contextual messages and off-pattern requests.

But that's only part of the solution. Security teams should enforce strict verification workflows for financial or access-related requests. They should implement anomaly detection to flag unusual email behavior or tone shifts. Identity verification tools can validate sender authenticity. Applying zero-trust principles to internal communications helps limit exposure. And behavioral AI tools should be used to monitor for generative patterns in email traffic. These combined controls are critical to stop what training alone may miss.

GenAI enables highly convincing, targeted phishing attacks that traditional defenses and outdated user training often fail to detect.

Why AI Threats Stay Hidden

AI is being used across enterprises, often without oversight. According to McKinsey, 71% of companies now use GenAI in at least one business function. But this widespread adoption is outpacing governance. The result is a growing number of blind spots in risk detection.

Security teams are focused on infrastructure, identity, and endpoint risk. But GenAI introduces a new layer of exposure that cuts across departments, functions, and data.

Common barriers to visibility include:

- Shadow AI use in marketing, R&D, and customer support
- No inventory of deployed or piloted LLM tools
- Security tools that don't map AI use to critical business assets

Until these gaps are closed, AI-related threats will remain hidden in plain sight.

What Security Leaders Can Do Right Now

AI-driven risk isn't unmanageable, but it is misunderstood. Security leaders can get ahead by treating it like any other form of cyber exposure: track it, quantify it, and mitigate it.

1. Strengthen policies, training, and governance

Start by identifying where GenAI, LLMs, and ML models are being used across the organization. That includes both sanctioned and unsanctioned deployments in departments like marketing, R&D, and customer support. Map these touchpoints to critical business assets and processes.

Update internal policies to explicitly cover deepfakes, model poisoning, and GenAI phishing. Establish acceptable use guidelines for public LLMs and ensure they're backed by training that uses real-world examples. Build incident response playbooks tailored to AI-driven threats and clarify accountability across functions. Governance must evolve as fast as AI does.

2. Invest in the right tools to gain visibility and reduce risk

Hyver, CYE's exposure management platform, shows where threats can emerge from GenAI adoption, whether through shadow use, vulnerable models, or misuse of public LLMs. It maps how risks like deepfakes, phishing, impersonation, model exploitation, and data leakage can move laterally through your environment and impact critical business assets.

Hyver doesn't treat AI as a separate category. It integrates GenAI into broader threat modeling, adjusting likelihoods based on attacker capabilities and behaviors. Security teams can compare AI-driven exposures against legacy risks, filter and prioritize findings, and understand the impact in clear financial terms.

Most importantly, Hyver quantifies each exposure in financial terms. By translating technical threats into business and dollar impact, it empowers CISOs to prioritize risk mitigation based on what truly matters.

With Hyver, you can

Surface deepfake and AI-based attack scenarios targeting your business assets

Detect unsafe usage of public LLMs like ChatGPT or Gemini, and flag where sensitive information may be at risk

Quantify AI-driven exposure to your business in dollars, rather than threat scores

Identify gaps in model validation, policy coverage, and user training

Create tailored mitigation plans with business-aligned prioritization

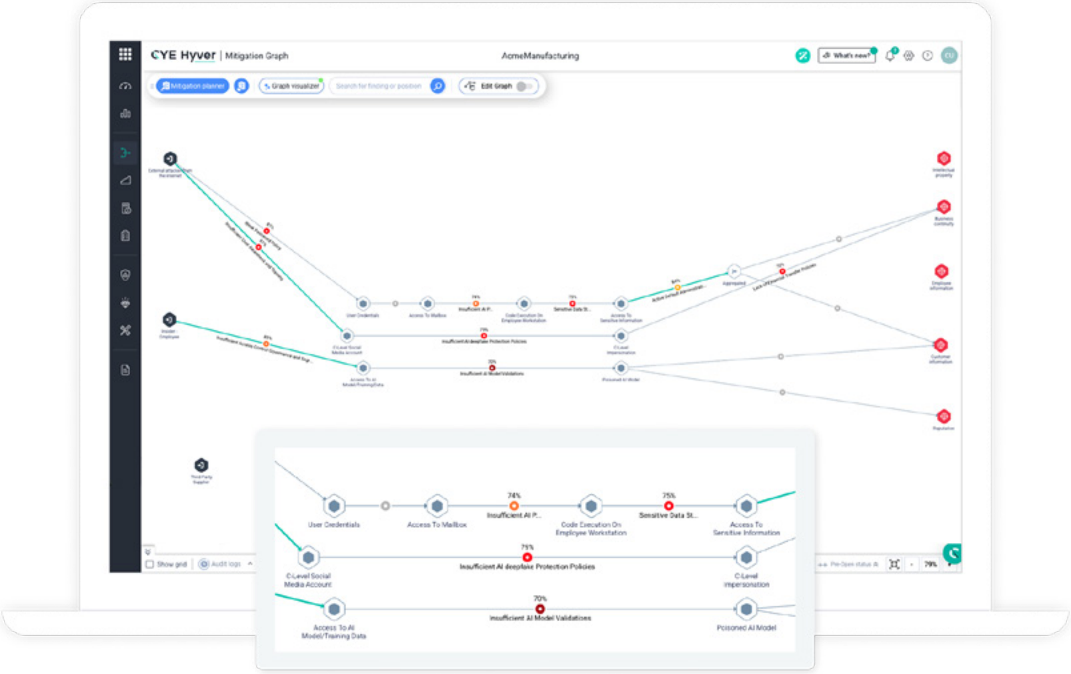
Deliver board-ready reports that connect AI risk to business outcomes

Mitigate AI Threats Without Reinventing Your Security Strategy

AI is changing the threat landscape. But it doesn't need to redefine your security strategy.

Treat AI like any other cyber risk: Bring it into focus. Quantify it. Prioritize it based on real business impact, not just novelty or hype. Mitigate it. And then, continuously validate its remediation.

Hyver helps organizations understand where AI-driven threats rank alongside all other cyber risks, ensuring that security leaders stay focused on reducing their exposure to the most likely threats to hit them, rather than being distracted by the surrounding hype.


CYE Hyver | Mitigation Graph
AcmeManufacturing
What's new?
Mitigation planner
Graph visualizer
Search for finding or position
Edit Graph
Show grid
Audit logs
79%
User Credentials
Access To Mailbox
74%
Insufficient AI P...
Code Execution On Employee Workstation
75%
Sensitive Data St...
Access To Sensitive Information
C-Level Social Media Account
75%
Insufficient AI deepfake Protection Policies
C-Level Impersonation
Access To AI Model/Training Data
70%
Insufficient AI Model Validations
Poisoned AI Model

Contact us to learn how Hyver can help you mitigate GenAI threats

Get in Touch →

About CYE

CYE's exposure management platform, Hyver, transforms the way security teams protect their organizations. With CRQ at its core, the platform reveals enterprises' exposure in financial terms, visualizes the most exploitable attack routes to critical business assets, and creates mitigation plans tailored to each business. CYE's customized reporting enables the sharing of vital board-level metrics and validating exposure reduction over time. In addition, CYE improves cybersecurity maturity by mapping weaknesses and defining targets based on industry frameworks.

Founded in 2012 in Israel with operations around the world, CYE has served hundreds of organizations across industries globally. Visit us at cyesec.com