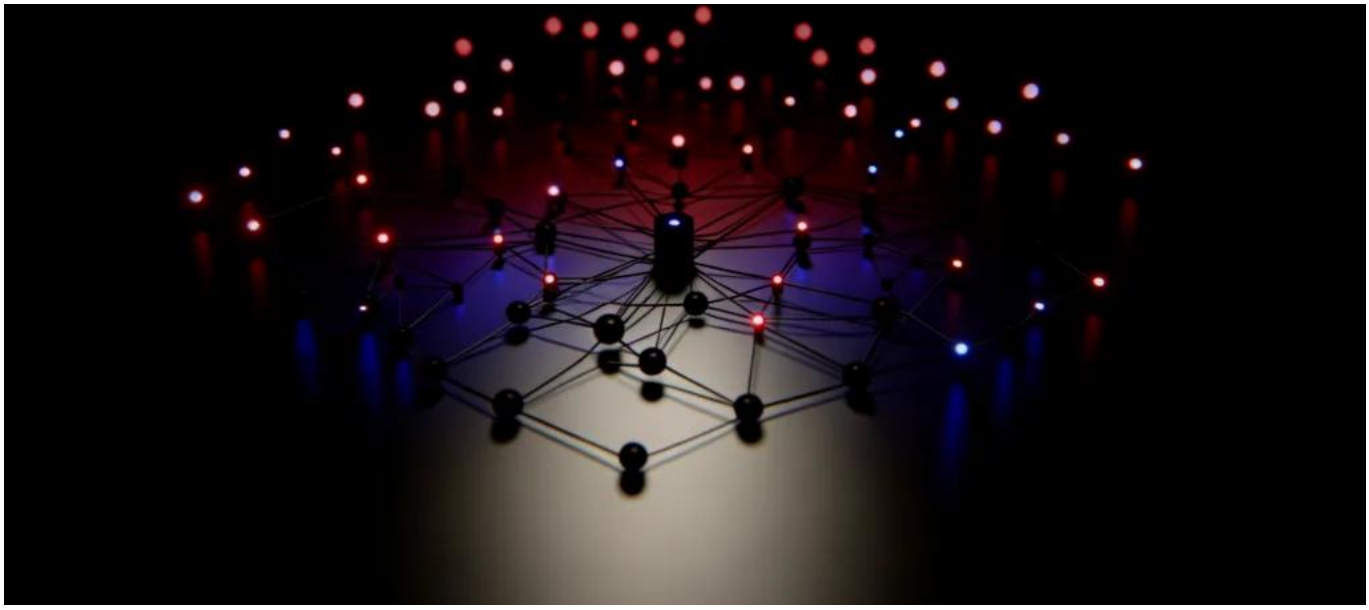


The AI-BOM Blind Spot: Why 276-Day Detection Times Prove We're Securing AI Models While Ignoring the Supply Chain Time Bomb

Your security team spent six months red-teaming your LLM for jailbreaks, but the backdoor was already living in your dependencies before you wrote your first prompt.

The Attack Vector You're Not Watching

Let me paint you a picture that should make every AI leader lose sleep tonight.

It's a Tuesday morning. Your ML engineering team is celebrating. They've just deployed a new predictive maintenance model that's going to save your manufacturing operations millions in downtime. The model passed every security audit. Prompt injection tests came back clean. Bias evaluations showed no concerning patterns. Your CISO signed off

personally.

Nine months later, you're reading about your company in a breach disclosure report.

The attack didn't come through the model weights. It didn't exploit a prompt vulnerability. It rode in on a single dependency—a seemingly innocuous Python package that your model imported for tensor operations. That package had been compromised months before your team even started the project.

This isn't a hypothetical scenario. This is the reality of AI security in 2025, and the numbers prove we're catastrophically underprepared.

According to Cyble's research on supply chain attacks, these incidents have doubled in frequency this year—we're now seeing 26 attacks per month compared to 13 per month in early 2024. The AI supply chain has become the softest target in enterprise security, and attackers have noticed.

The 276-Day Problem: When Detection Becomes Discovery

Here's the statistic that should reframe every conversation you have about AI security this quarter: the average breach detection time has ballooned to 276 days.

Let that sink in. Nearly nine months of dwell time.

That's not detection. That's archaeology.

> When your adversary has 276 days of undetected access, they're not just stealing data—they're becoming part of your infrastructure.

The Hacker News CISO's Guide to AI Supply Chain Attacks details why AI-assisted attacks are particularly difficult to detect: they're designed by systems that understand detection patterns better than your security team does. Attackers are now using AI-generated polymorphic malware that mutates just enough to evade signature-based detection while

maintaining its malicious functionality.

But here's what's really terrifying—that 276-day average doesn't account for supply chain compromises that happened upstream, before your organization ever touched the affected code. If a malicious actor poisoned a popular machine learning library 18 months ago, and you installed it 12 months ago, and you detected the breach 9 months after that… how do you even calculate the exposure window?

You don't. Because you were never the intended detection point. You were the intended victim.

The AI-BOM Gap: Tracking What You Can't See

Software Bills of Materials (SBOMs) became a standard security requirement after the SolarWinds disaster. Most enterprises now maintain reasonably comprehensive inventories of their traditional software dependencies.

But AI systems aren't traditional software. And the AI-BOM—a comprehensive accounting of every model, dataset, training pipeline, and dependency that contributes to your AI system—remains a concept more discussed than implemented.

The problem is multi-dimensional:

Model Provenance

Where did your base model come from? If you're fine-tuning an open-source foundation model, do you know its complete training lineage? Who contributed to the training data? Were the contributors verified? Could adversarial data have been injected during pre-training?

Training Pipeline Dependencies

Your model training process probably pulls from dozens of packages—data loaders, preprocessing libraries, optimization algorithms, logging frameworks. Each one is a potential insertion point. Each one has its own dependency tree. Some of those sub-dependencies haven't been audited in years.

Inference Infrastructure

Your production deployment stack has entirely different dependencies than your training environment. Serialization libraries, API frameworks, monitoring tools—all of them touch your model at runtime, and all of them are potential compromise vectors.

Data Pipeline Integrity

The data feeding your models in production is often sourced from external systems, third-party providers, or user inputs. Poisoning attacks don't require compromising code—they can corrupt the data itself.

The SecureWorld analysis of 2025 supply chain threats identifies APIs as one of the critical weak links in this chain. Your AI system's API surface often connects to external services with their own security postures—or lack thereof.

The $60 Billion Wake-Up Call

If security arguments don't move your executive team, financial ones will.

Cybersecurity Ventures reports that global costs of software supply chain attacks reached $60 billion in 2025. The projection for the coming years? $138 billion.

Let me put this in perspective with a table that illustrates the acceleration:

Metric	Early 2024	2025	Change
Supply chain attacks per month	13	26	+100%
Malicious AI package uploads	Baseline	+156%	2.5x increase
Third-party breach share	~20%	30%	+50% share
Supply chain breach growth	Baseline	+40%	Accelerating

The 156% increase in malicious AI packages deserves special attention. Attackers aren't just reusing old techniques—they're developing AI-specific attack vectors at an accelerating pace. They understand that the AI ecosystem's hunger for new tools and libraries creates an asymmetric opportunity: defenders need to vet everything, attackers

only need to succeed once.

Real Attacks, Real Consequences

The FSD Tech analysis of AI disasters in 2025 documents a case that should be required reading for every AI and security leader: a malicious AI model disguised as a predictive maintenance tool disabled safety alarms at a Middle Eastern oil refinery.

Read that again. Disabled safety alarms. At an oil refinery.

This wasn't a theoretical vulnerability or a CTF exercise. This was a weaponized AI system that passed initial security reviews because nobody was scanning for AI-specific supply chain compromises. The organization looked at the model's functionality—does it predict maintenance needs?—without examining its provenance or hidden behaviors.

> We've spent years building AI systems that can understand context. Attackers have built AI systems that understand how to hide in that context.

The OWASP Gen AI Incident Round-up from early 2025 documented additional attack patterns that should concern every organization:

- **Chain-of-thought jailbreak attacks** that exploit the reasoning processes of large language models to bypass safety guardrails
- **Azure OpenAI account hijacking** through stolen API keys, demonstrating that the AI service layer is as vulnerable as the model layer
- **SockPuppet attacks** where fake developer personas build credibility over months or years before injecting malicious code into legitimate open-source projects

The SockPuppet methodology is particularly insidious. These aren't smash-and-grab operations. Attackers are playing the long game—establishing developer identities, contributing legitimate improvements to popular projects, earning commit rights, and then exploiting that trust. By the time the backdoor is inserted, the "developer" has a year of verifiable contributions and community endorsement.

The Stanford Index Reality Check

The Stanford AI Index Report for 2025 presents a sobering analysis of AI-related incidents: 233 total AI incidents were reported in 2024, representing a 56.4% increase over 2023.

But here's what concerns me more than the raw numbers—the detection gap. If we know about 233 incidents, and we know the average detection time is 276 days, how many incidents are currently active but undiscovered? How many were never detected at all?

The AI incident curve is steepening, but our visibility into these incidents isn't keeping pace. We're measuring what we catch, not what's actually happening.

Why Traditional Security Tools Fail Here

Your SAST scanner wasn't designed to evaluate model weights for hidden backdoors.

Your DAST tools don't know how to test for adversarial inputs that activate dormant malicious behaviors.

Your SBOM generator doesn't understand that a Hugging Face model card isn't the same as a package manifest.

Your network monitoring can't distinguish between legitimate inference API calls and data exfiltration disguised as inference API calls.

This isn't a criticism of existing security tools—it's an acknowledgment that AI systems represent a fundamentally different attack surface that requires fundamentally different defensive approaches.

The Model Layer

Neural networks are opaque by design. Unlike traditional software where you can (theoretically) trace every execution path, a trained model's "logic" is distributed across millions or billions of parameters. Hiding malicious behavior in that parameter space is relatively straightforward—you can train a model to behave perfectly normally until it receives a specific trigger, at which point it activates a completely different behavior

pattern.

Traditional code analysis can't see this. The backdoor isn't in the code; it's in the math.

The Training Layer

Data poisoning attacks corrupt the training process itself. If an attacker can influence even a small percentage of your training data, they can shape the model's behavior in ways that are extraordinarily difficult to detect post-training. The model will still perform its primary task well—that's the whole point. It just also performs the attacker's task when prompted correctly.

The Dependency Layer

Modern ML systems have deep dependency trees. A typical training pipeline might pull from PyTorch, which pulls from dozens of sub-packages, which pull from dozens more. Somewhere in that tree is a package maintained by one person who accepted a pull request without sufficient review. That PR contained a subtle modification that only activates when the package detects it's running in a GPU-enabled environment with certain model architectures loaded.

Your security team audited your code. They didn't audit the 847 transitive dependencies.

The Open Source Paradox

AI development runs on open source. This is simultaneously the greatest strength and the most critical vulnerability of the entire ecosystem.

The strength is obvious: open-source models, libraries, and tools have democratized AI development, enabled rapid innovation, and created communities of researchers who can verify and improve each other's work.

The vulnerability is equally obvious: open source means open contribution. And open contribution means that anyone—including adversaries—can submit code.

The malicious AI package uploads increasing by 156% tells us that attackers understand

this asymmetry. PyPI, npm, Hugging Face Hub—these aren't just package repositories anymore. They're battlegrounds.

The SockPuppet attack pattern makes this even more complex. When attackers invest months building legitimate contributor credentials before inserting malicious code, they're exploiting the very trust mechanisms that make open source work. Kill the trust, kill the collaboration. Maintain the trust, maintain the vulnerability.

There's no clean solution here. But there are approaches that can reduce exposure.

Building an AI-BOM Strategy

Let's get practical. What does a real AI supply chain security program look like?

1. Inventory Everything

You can't secure what you don't know you have. Start with a comprehensive inventory of:

- All AI/ML models in development, testing, and production
- The provenance of each model (pre-trained bases, fine-tuning datasets, training pipelines)
- Every library and package used in training and inference
- All APIs and external services your AI systems connect to
- Data sources and data pipelines feeding production models

This sounds basic because it is. Most organizations can't produce this inventory today. That's the problem.

2. Establish Model Provenance Verification

When you download a pre-trained model, what are you actually getting? Model cards and documentation can be falsified. Model weights can be modified.

Consider implementing:

- Cryptographic signing of model artifacts throughout the training pipeline

- Hash verification of model weights against known-good states
- Sandboxed validation environments that test model behavior before production deployment
- Red team exercises specifically targeting model-layer attacks

3. Treat Dependencies as Untrusted Code

Even if you trust the maintainers of your direct dependencies, you probably don't know the maintainers of your transitive dependencies. Act accordingly.

- Pin dependency versions explicitly rather than accepting automatic updates
- Mirror critical dependencies internally so you can control what gets deployed
- Implement dependency review processes that examine actual code changes, not just version numbers
- Monitor for supply chain threat intelligence—know when packages you use are compromised

4. Implement Behavioral Monitoring

Traditional signature-based detection doesn't work against AI-specific attacks. You need behavioral baselines and anomaly detection.

- Establish baseline behavior profiles for your models in production
- Monitor for unexpected output patterns, resource usage anomalies, or unusual API calls
- Implement canary inputs that should always produce known outputs—deviation indicates compromise
- Track model performance drift over time (sudden accuracy changes could indicate poisoning)

5. Build Incident Response Playbooks for AI-Specific Scenarios

Your IR team probably has playbooks for ransomware, data breach, and insider threat. Do they have playbooks for:

- Discovering a backdoored model in production?

- Responding to a compromised training pipeline?
- Managing a poisoned dependency that's been in use for months?
- Investigating potential data poisoning in a production model?

If the answer is no, you're not ready for the threats you're facing.

The Organizational Challenge

Technical controls matter, but organizational structure matters more.

In most enterprises, AI security falls into a gap between teams. The security team understands security but not AI. The data science team understands AI but not security. The infrastructure team manages the deployment environment but doesn't understand either domain deeply.

> AI supply chain security isn't a security problem or an AI problem. It's a collaboration problem with technical dimensions.

Successful organizations are creating explicit ownership models:

- **AI Security Champions** embedded in data science teams who understand both domains
- **Cross-functional working groups** that bring security, ML engineering, and infrastructure together regularly
- **Clear escalation paths** so ML engineers know who to contact when they spot anomalies
- **Shared responsibility models** that distribute accountability across teams rather than leaving it in the gap between them

The Vendor and Third-Party Dimension

Your AI supply chain doesn't stop at your organizational boundary. If you're using:

- AI platform services (cloud ML platforms, inference APIs)

- Third-party models (fine-tuned or off-the-shelf)
- Data providers (training data, validation data, production data feeds)
- AI SaaS applications

Then your supply chain extends into those organizations as well.

The 30% of breaches now attributable to third-party supply chain compromises means that your vendor security assessments need to include AI-specific questions:

- What is the vendor's model provenance verification process?
- How do they secure their training pipelines?
- What dependencies do their models rely on, and how do they manage them?
- Do they have AI-specific incident response capabilities?
- Can they provide AI-BOM documentation for their products?

If your vendor can't answer these questions, that's information you need.

The Regulatory Trajectory

Regulation is coming. The EU AI Act already includes supply chain provisions. The US is developing frameworks through NIST and sector-specific regulators. It's not a question of whether AI-BOM requirements will become mandatory—it's a question of when.

Organizations building AI supply chain security capabilities now are getting ahead of compliance requirements, not just threat requirements. When the regulatory framework crystallizes, they'll be ready. Organizations that wait will be scrambling.

History tells us how this goes. Remember how many organizations were racing to comply with GDPR in the months before enforcement? Remember how chaotic that was?

Don't let AI-BOM become your next compliance fire drill.

The Asymmetry Problem

Here's the core challenge we're facing: the asymmetry between offense and defense in AI supply chains is enormous.

Attackers need to:

- Find one vulnerable package in a dependency tree of hundreds
- Succeed once across thousands of potential targets
- Wait patiently for 276 days (or more) without being detected

Defenders need to:

- Secure every package in every dependency tree
- Protect against every attack vector simultaneously
- Detect anomalies in systems designed to behave unpredictably

This asymmetry isn't solvable with current approaches. We need:

- **Better tooling** specifically designed for AI supply chain analysis
- **Information sharing** that lets organizations warn each other about compromised packages faster
- **Industry standards** for AI-BOM that make supply chain transparency the default
- **Investment in open source security** that matches our investment in open source development

What Should Change Tomorrow

I've covered a lot of strategic ground. Let me close with tactical actions you can take this week:

1. **Run a dependency audit** on your most critical AI system. Generate a complete dependency tree and check each package against vulnerability databases. You'll probably be surprised by what you find.
2. **Ask your ML team** where your base models came from. If the answer is "we downloaded them from Hugging Face," ask the follow-up: "How did we verify they weren't tampered with?"
3. **Review your CI/CD pipeline** for AI model deployment. Identify every point where external code or data enters the process. Each one is a potential insertion point.
4. **Test your incident response** by running a tabletop exercise with an AI supply chain

scenario. You'll quickly discover the gaps in your current playbooks.

5. **Start building your AI-BOM**, even if it's just a spreadsheet. Document what you know about model provenance, training data sources, and critical dependencies. Imperfect documentation beats no documentation.

The Path Forward

The AI supply chain security problem isn't going to solve itself. The attackers are sophisticated, patient, and increasingly AI-augmented themselves. Detection times are getting longer, not shorter. Costs are climbing exponentially.

But this isn't a doomed situation—it's an immature one. We're in the early stages of understanding AI-specific attack surfaces and developing AI-specific defensive capabilities. The organizations that invest in these capabilities now will be dramatically better positioned than those that wait.

We've been here before. Twenty years ago, application security was an afterthought. Ten years ago, cloud security was a mess. Five years ago, supply chain security for traditional software was barely on the radar. In each case, the organizations that took these threats seriously early didn't just survive—they gained competitive advantage.

AI supply chain security is following the same trajectory, just compressed. The threat is real now. The window for getting ahead of it is closing.

The question isn't whether your AI supply chain is a target. It is. The question is whether you'll detect the compromise in time to matter.

At current detection rates, the odds aren't in your favor.

The AI supply chain is the new attack surface, and 276 days of dwell time proves we're securing the wrong layer while attackers own the infrastructure underneath.



WRITTEN BY

Artur Markus

Software engineer in Basel building AI infrastructure and agentic systems, with a background in pharmaceutical automation, GxP compliance, and validation. I help teams turn AI from a chatbot into a reliable operational layer.

Book a meeting