A new day has dawned — responding to AI-native security incidents

By Michael A. Gold, Esq., Jeffer Mangels & Mitchell LLP

JULY 28, 2026

The cybersecurity landscape has fundamentally changed. Organizations no longer face only the familiar threat of unauthorized access to databases and file systems. A new category of incident has emerged — one in which artificial intelligence systems are both the vector of attack and the target of compromise.

These AI-native breaches present challenges that existing incident response frameworks, forensic methodologies, and legal regimes were not designed to address. For organizations deploying AI at scale, understanding these differences is no longer optional; it is a governance imperative.

A new category of incident has emerged — one in which artificial intelligence systems are both the vector of attack and the target of compromise.

Yet despite AI's rapid proliferation, remarkably little attention has been devoted to AI-native breaches — how they differ from conventional incidents, what forensic investigation requires, and what legal obligations they trigger. The AI industry's enormous concentrations of investment capital create powerful incentives to emphasize transformative potential while minimizing attention to novel risks. The result is a familiar pattern: AI systems are being deployed at scale into critical environments while the frameworks for responding to their compromises remain underdeveloped and largely unexamined.

The emerging threat landscape

Traditional data breaches follow a broadly understood pattern: a threat actor gains unauthorized access, exfiltrates data, and the organization responds by identifying compromised records, notifying affected individuals, and remediating the vulnerability. The playbook is well-established, supported by reams of regulatory guidance, case law, and industry standards.

AI-native breaches depart from this model in fundamental ways. They can be categorized into two distinct threat types.

The first involves AI-powered attacks — incidents in which adversaries leverage AI to conduct reconnaissance, craft sophisticated phishing campaigns, automate exploitation of vulnerabilities, or evade detection with adaptive malware.

The second, more novel type of AI-native attacks involve those that target an organization's AI resources themselves — its training data, model weights, inference pipelines, and outputs.

In this second category, the threat actor's objective may not be to steal personal records. Instead, the goal may be to poison training data so a model produces corrupted outputs, to extract proprietary model parameters through crafted queries, to manipulate a model's behavior through adversarial inputs, or to exfiltrate the intellectual property embedded in the model itself. Each scenario challenges the assumptions underlying conventional incident response.

Forensic investigations: Conventional versus AI-native breaches

In a conventional data breach, forensic investigators operate within a well-defined paradigm. They examine network logs, endpoint telemetry, access control records, and file system artifacts to reconstruct the timeline of unauthorized access. The questions are familiar: Who accessed what? When? What was exfiltrated? The evidence is largely deterministic — log entries either exist or they do not, and the chain of custody for digital evidence follows established protocols.

AI-native breaches confound this paradigm in several critical respects. First, the "crime scene" is fundamentally different. When an adversary poisons training data, the compromise may have occurred weeks or months before anomalous behavior manifests.

Investigators must examine the provenance of potentially millions of training data points — a task for which traditional forensic tools are poorly suited. The question shifts from "what was taken" to "what was altered, and how has that alteration propagated through the model's learned representations."

Second, the evidence is probabilistic rather than deterministic. A compromised model does not produce a clean log entry showing tampering. Its outputs shift in subtle, statistical ways that may be difficult to distinguish from normal drift. Forensic teams must employ techniques from machine learning research — model comparison, activation analysis, distributional testing — far removed from conventional disk imaging and log analysis.

Third, model extraction attacks present a unique evidentiary challenge. When an adversary queries a model thousands of times to reconstruct its parameters, the individual queries may appear entirely legitimate. There is no unauthorized access in the traditional sense; the attacker uses the system exactly as designed, merely at a scale and sophistication that enables them to approximate the model's internal logic. Identifying this activity requires behavioral analytics that can distinguish between legitimate use and systematic extraction — a discipline still in its infancy.

An AI-native breach may not involve the compromise of personal information in any recognizable statutory sense.

Fourth, the concept of "scope" is far more ambiguous. In a conventional breach, the number of compromised records can be determined or approximated. In an AI-native breach, harm may be diffuse and ongoing — a poisoned model may generate flawed decisions affecting thousands without any single "record" being compromised, and a stolen model represents intellectual property whose value is difficult to quantify.

Legal issues: A new and very different framework

The legal landscape for AI-native breaches is marked by uncertainty and gaps that do not exist — or exist in different forms — in conventional incidents.

Notification obligations

Data breach notification statutes, both at the state level and internationally, are generally triggered by the unauthorized acquisition of specified categories of personal information — Social Security numbers, financial account credentials, protected health information, and similar defined data elements.

An AI-native breach may not involve the compromise of personal information in any recognizable statutory sense. If an adversary poisons a model or extracts model weights, no individual's Social Security number or financial account information has been accessed. Yet the harm to the organization and potentially to individuals affected by corrupted model outputs may be substantial.

Consider a poisoned credit-scoring model that systematically produces inaccurate assessments for a demographic group. The harm is real, but no "record" has been "acquired" as contemplated by statutes such as the CCPA, New York's SHIELD Act, or GDPR Article 33. Organizations must ask whether corruption of a model that processes personal data constitutes a breach "involving" that data.

Incident response plans should be updated to include scenarios specific to AI compromise, with playbooks that address model integrity verification, training data provenance investigation, and the particular notification challenges these incidents present.

The existing notification framework provides little clarity, leaving organizations caught between over-notification — risking public alarm and litigation — and under-notification, risking regulatory enforcement if the incident later surfaces.

Regulatory expectations and jurisdictional uncertainty

The intersection of AI governance and incident response remains largely unaddressed, creating a jurisdictional landscape that is fragmented, overlapping, and often contradictory.

In the United States, a single AI-native incident may implicate state attorneys general under consumer protection statutes, the FTC under its unfairness authority, sector-specific regulators (OCC, FDIC, HHS OCR, SEC), and the emerging patchwork of state AI legislation exemplified by Colorado's AI Act.

No single federal framework governs AI security incidents. Organizations operating across jurisdictions must reconcile conflicting regulatory expectations with no authoritative hierarchy to resolve conflicts.

The EU AI Act imposes serious-incident reporting obligations for high-risk systems, but its framework is oriented toward safety harms rather than security breaches — raising questions about its interaction with parallel GDPR breach-notification requirements.

Evidentiary and litigation challenges

When AI-native breaches result in litigation, the evidentiary challenges multiply. Demonstrating causation between a model

compromise and specific harm requires expert testimony bridging machine learning and legal standards of proof.

Courts have limited experience evaluating claims that a model was "poisoned" or that extracted weights constitute trade secret misappropriation. The Daubert standard will be tested by the need to present probabilistic forensic findings to judges and juries accustomed to deterministic evidence.

Intellectual property and trade secrets

AI models represent intellectual property that fits uneasily within existing frameworks. Model weights may qualify as trade secrets, but demonstrating misappropriation requires proving that an extracted approximation is substantially derived from the original — a showing that may require novel technical methodologies. Patent and copyright protections for AI innovations remain unsettled.

When a breach targets the model itself, the organization's legal remedies are less certain than when the stolen asset is a customer database or financial record.

Third-party and supply chain liability

AI systems frequently depend on third-party components — pre-trained models, external training data, APIs, and cloud-based inference infrastructure. An AI-native breach may originate in the supply chain, raising complex questions of contractual liability, indemnification, and allocation of forensic investigation costs.

Organizations may find that vendor agreements drafted with conventional breaches in mind do not adequately address scenarios in which a supplier's compromised model component propagates harm through the organization's own systems.

Governance and preparedness

Organizations seeking to prepare for AI-native incidents should consider several measures. Incident response plans should be updated to include scenarios specific to AI compromise, with playbooks that address model integrity verification, training data provenance investigation, and the particular notification challenges these incidents present.

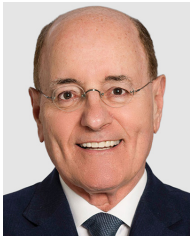
Forensic capabilities should be expanded — whether through internal hiring or retainer arrangements with specialized firms — to include practitioners with machine learning expertise alongside traditional digital forensics credentials.

From a legal preparedness standpoint, organizations should conduct tabletop exercises that model AI-native breach scenarios, testing notification decision-making against fact patterns that do not fit within existing statutory triggers. Contracts with AI vendors should be reviewed to ensure they address model integrity, supply chain compromise scenarios, and cooperation obligations in the event of an incident affecting shared AI resources.

Conclusion

An AI-native breach is not simply a data breach that happens to involve an AI system. It is a categorically different type of incident that demands categorically different investigative techniques, legal analysis, and governance responses. To be sure, the frameworks that served us well for two decades of conventional breach response remain necessary. But make no mistake, they are no longer sufficient.

About the author



Michael A. Gold is a partner and chair of the cybersecurity and privacy group at **Jeffer Mangels & Mitchell LLP**. He counsels clients in matters including information security and privacy compliance, supply chain and outsourcing security issues, AI implementation and governance, data breach responses and investigations, regulatory matters, crisis management, and technology contracts. He is based in Los Angeles and can be reached at MGold@jeffer.com.

This article was first published on Reuters Legal News and Westlaw Today on July 28, 2026.