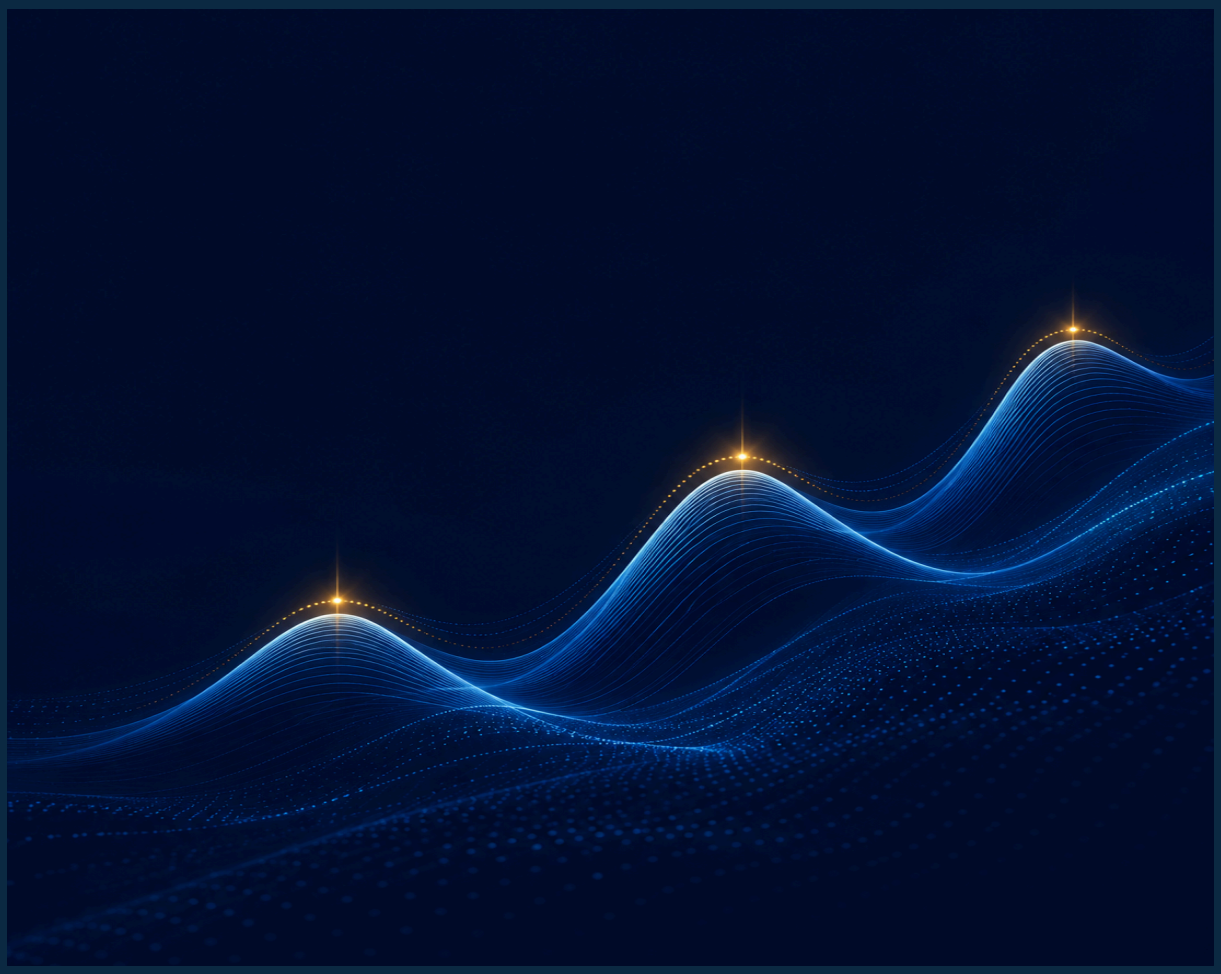
RYAN DIVIS | TECHNOLOGY AND AI STRATEGY

Constant Vigilance

Why AI fraud defense is a feedback loop you operate, not a tool you buy



By Ryan Divis

PUBLIC EDITION | JUNE 2026

EXECUTIVE SUMMARY

The thesis

As a technology leader, I have spent the last several years working both with AI and against it. AI gives defenders extraordinary leverage. The same leverage is available to anyone trying to defraud a business, manipulate it, or automate around its controls.

> **AI is not a feature. It is a decision layer inside a feedback loop.**

Organizations that treat AI fraud defense as a product they can buy and switch on will lose to organizations that treat it as a discipline they operate every day.

01	02	03
Buy the tools	**Own the context**	**Operate the loop**
A capable vendor expands your signal set, shortens implementation time, and raises the cost of attack.	Only the operator can interpret local economics, customer impact, and a pattern that appeared yesterday.	Observe, evaluate, detect drift, intervene, review, and learn - continuously.

WHY THE FINANCIAL LAYER MATTERS

A marketplace may appear to be a storefront, but its financial layer inherits the risks of a payment network. Seller onboarding can involve identity or business verification, tax checks, sanctions screening, and ownership validation. Seller payouts are money movement. A marketplace is a payment network wearing a storefront.

AUTHOR'S NOTE This paper draws on direct operating experience. Company and vendor names are omitted, and certain volumes, thresholds, controls, and implementation details are generalized.

CASE STUDY 01

Identity announces itself at the point of payout

Our first warning was not a clean fraud alert. It was a divergence.

At a large digital marketplace, new seller creation began rising much faster than the ordinary activity surrounding it. Traffic had not grown enough to explain the increase. Neither had legitimate orders or payouts.

On paper, the accounts looked plausible. They passed the identity-related, tax, and email checks available at the time. They listed products and, in some cases, completed transactions.

Then complaints accumulated around low-quality or unfulfilled orders. At approximately the same time, downstream financial institutions returned elevated-risk signals on a subset of payout accounts.

The fraud did not clearly announce itself during registration. It announced itself when money moved. Our onboarding controls had approved accounts that later financial signals showed to be high risk. The individual checks had worked as designed, but the overall system had still produced the wrong decision.

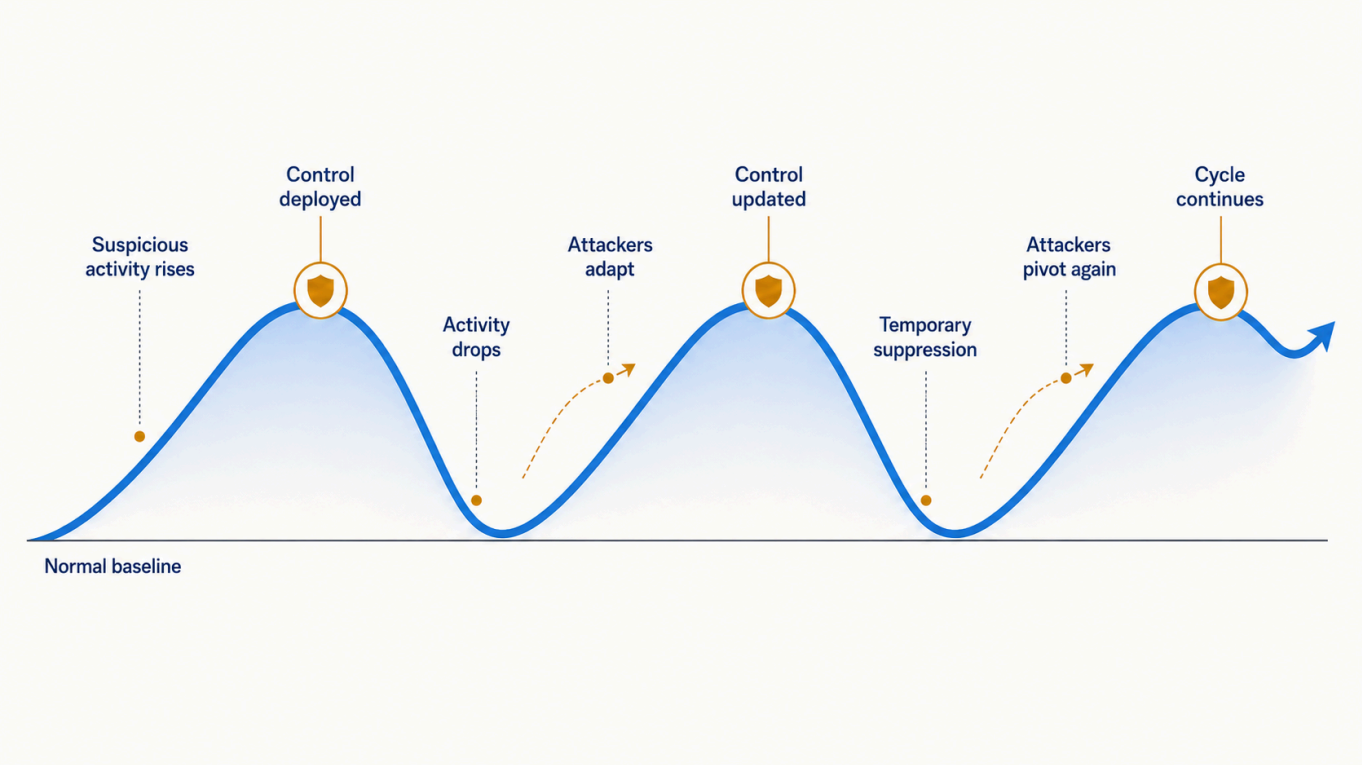
We strengthened the controls we already operated and introduced an AI-powered identity and fraud partner. Suspicious activity declined, then returned. The tool had not failed. It had raised the cost of attack, improved our visibility, and forced the actors to adapt. What it had not done was permanently solve an adversarial problem.

> **The product did not solve the problem. It helped us operate the problem.**

The distinction between a control that passes and an actor who is trustworthy is fundamental to financial risk. [4]

CASE STUDY 02

A defense that works once is not a defense



Illustrative reconstruction. Values are indexed to a normalized baseline and do not disclose production traffic.

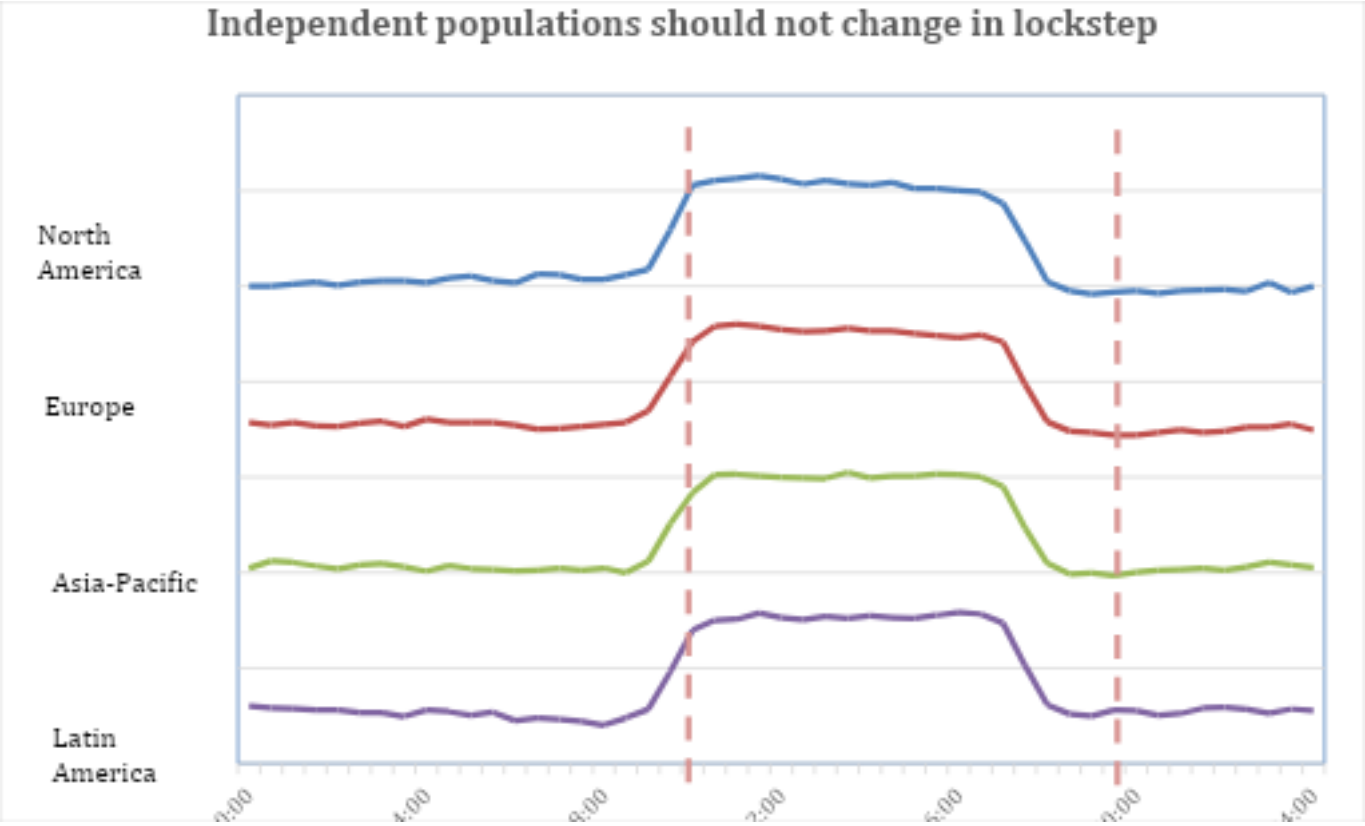
Later, the marketplace experienced a surge of automated traffic several multiples above its normal baseline. It was not a conventional denial-of-service attack. The requests were broadly distributed, technically valid, and not overtly destructive. Viewed one at a time, they resembled ordinary user behavior. Viewed together, they consumed material infrastructure capacity and distorted the signals used to understand the health of the business.

We blocked the patterns we could identify through behavior, device characteristics, network signals, and geography. Each intervention worked for a time. We evaluated and deployed additional bot-management capabilities. Again, the activity dropped. Again, it returned in a different form.

> **In an adaptive environment, a control is not proven when activity falls. It is proven when the organization can see what happens next.**

THE HUMAN-READABLE SIGNAL

Real populations follow local rhythms. Automated populations synchronize.



Illustrative regional traces. The synchronized shape is the signal; no production data is shown.

A human analyst found the most important clue. When a single client characteristic was plotted across multiple regions, it changed everywhere at nearly the same moment. It climbed with the same shape, held for a similar period, and disappeared in coordination across the globe.

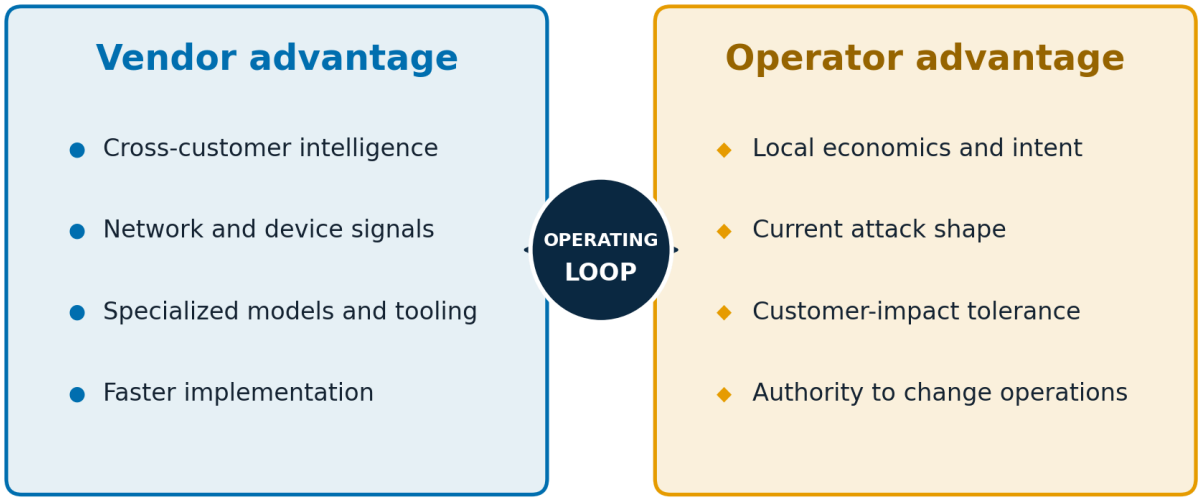
It did not follow local time zones, population density, or the day-and-night cycles that govern real human activity. No natural population of independent users behaves that way.

The vendor had extensive capabilities, but the product did not provide a ready-made control for this particular form of synchronized distributed behavior. That did not mean the model was poor. It meant the model could act only on the signals, patterns, and decision mechanisms available to it.

> **The model saw events. The operator saw a population behaving in a way no real population could.**

THE STRUCTURAL LIMITATION

Why AI cannot operate alone



A fraud platform can bring a wider view of adversarial activity than most individual businesses will ever possess. It may observe attacks across industries, devices, networks, and regions. That breadth is enormously valuable.

But every model operates within the signals it receives, the objectives it has been designed to optimize, and the time required to recognize a new pattern. The operator possesses a different advantage: local context. The operator knows what changed yesterday, which action matters financially, which anomaly is harmless, and which customer cost is acceptable.

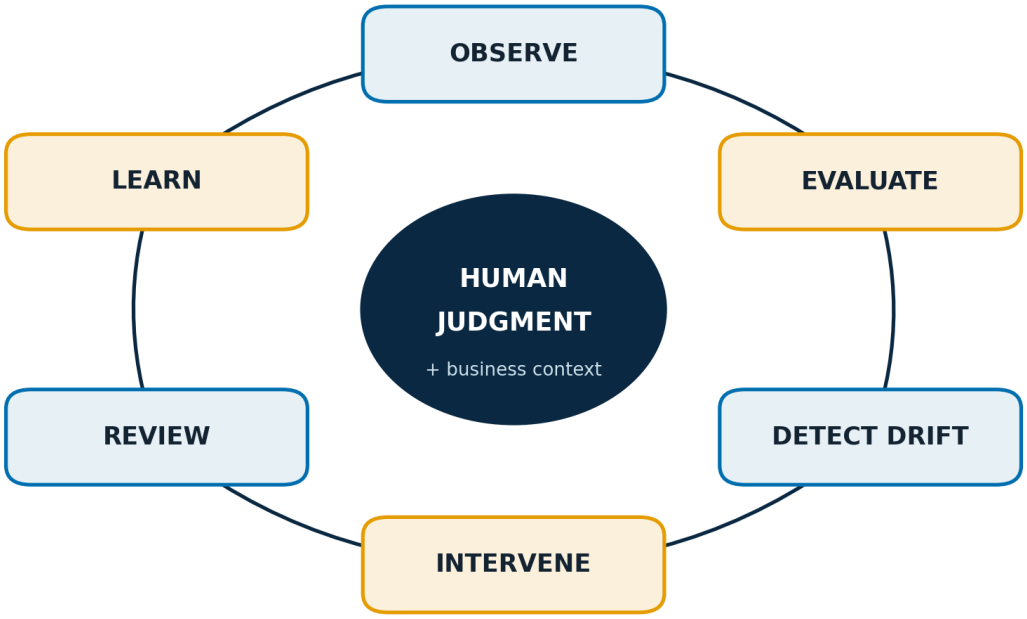
The limitation is not unique to one vendor. NIST similarly frames AI risk management as an ongoing practice of governance, mapping, measurement, and management, with monitoring across the system lifecycle. [1][2] Its adversarial machine-learning taxonomy also treats attacker goals, capabilities, and lifecycle stages as changing parts of the environment. [3]

> **The vendor raises the floor. The operator manages the ceiling.**

THE OPERATING FRAMEWORK

The Vigilance Loop

What finally worked was not replacing one model with another. It was building an operating loop around the tools.



> **Tools feed the loop. They do not own it.**

How to operate the loop

01

Observe

Instrument identity, behavior, transactions, money movement, customer friction, and infrastructure impact. A business cannot detect change in a signal it never collected.

02

Evaluate

Measure whether controls still perform against attacks already understood. Track false positives as seriously as fraud prevented.

03

Detect drift

Look for changes in distributions, correlations, timing, geography, and attacker economics. A single event can look normal while its population becomes impossible.

04

Intervene

Apply proportionate, measurable, and reversible friction. Temporary controls can reduce predictability and limit the cost of an incorrect decision.

05

Review

Use human investigation as a discovery function, not merely an exception queue. Reviewers identify the pattern no model represents yet.

06

Learn

Feed confirmed outcomes into evaluations, rules, models, operating procedures, and vendor conversations. Every decision creates evidence for the next cycle.

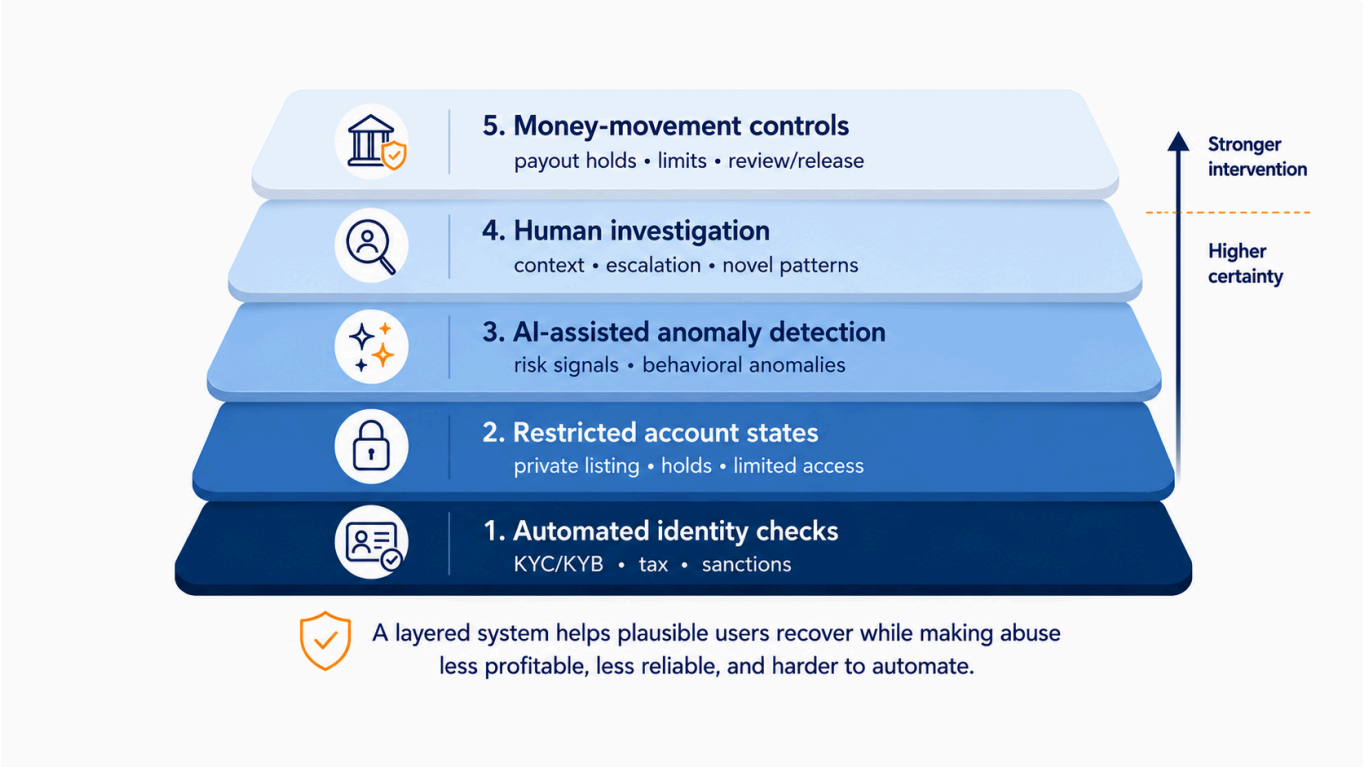
The output of a fraud decision is not merely an approval or denial. It is new information about the environment. The loop then begins again.

DECISION DESIGN

Design controls without training the attacker

A	B	C
Help legitimate users recover	**Limit machine-readable feedback**	**Prefer reversible intervention**
Give plausible customers a clear path to correct mistakes, provide evidence, or appeal.	Do not reveal the precise signal, threshold, or sequence an automated actor triggered.	Use temporary holds, challenges, and review states when certainty is incomplete.

A layered decision system, not a single model



Our seller-review process evolved into a layered system combining automated identity checks, restricted account states, AI-assisted anomaly detection, human investigation, and money-movement controls. Communications helped plausible legitimate users recover without providing automated adversaries with a precise description of the control they had triggered.

The objective was not to make automation impossible. The objective was to change the economics. A delay that is trivial for one human can be expensive when repeated at machine scale. Good fraud defense makes abuse less profitable, less reliable, and harder to automate.

The loop runs on metrics

> **You cannot detect drift in a signal you do not measure. You cannot evaluate a control against a baseline you never captured.**

THE OPERATING SCORECARD

Control Performance	Customer & Business	Adversary Behavior
• Precision/recall • False-positive rate • Challenge completion • Review outcomes • Predicted vs. confirmed risk	• Conversion through projected flows • Challenge abandonment • Review time • Support contacts • Money-movement delay • Infrastructure and vendor cost	• Retry rate • Time to adaptation • Signal migration • Attack volume and unit cost • Pattern reuse • Synchronized behavior

No single metric proves that a defense is working. A decline in suspicious signups may mean a control succeeded. It may also mean attackers moved farther downstream, changed identity providers, or shifted to compromised legitimate accounts. A reduction in traffic may indicate bot mitigation. It may also reflect customer friction or a measurement failure.

This is why instrumentation is not the boring prerequisite to the interesting AI work. It is the AI work. NIST's AI RMF playbook likewise emphasizes continuous monitoring, incident response, and mechanisms for human appeal and override. [2]

What leaders should demand

Ask more than whether the vendor has a strong model. Ask whether your organization can operate the decision system around it.

• Which signals will the product actually receive from our systems?	• Which business outcomes is it designed to optimize?
• How are false positives measured, reviewed, and appealed?	• How quickly can a new local pattern become an actionable control?
• Can analysts investigate decisions without exposing sensitive logic?	• How do confirmed outcomes return to rules, models, and evaluations?
• Which interventions are reversible, and who can reverse them?	• What happens when the model cannot represent the pattern?
• Who watches the system after deployment?	

Buy the tools. Own the system.

If the answer to the final question is effectively “the vendor,” the organization has outsourced a capability it still needs to own.

A practical first 90 days

0-30	31-60	61-90
Map and baseline	**Evaluate and observe**	**Intervene and learn**
Inventory decisions, signals, controls, vendors, review queues, and money-movement consequences. Establish baseline distributions and customer-impact measures.	Build a known-case evaluation set. Add drift views across identity, behavior, geography, timing, and economics. Define escalation ownership.	Introduce reversible controls, a review cadence, post-intervention measurement, and a clear path for outcomes to update the system.

MINIMUM OPERATING CADENCE

Hourly / daily	Watch live anomalies, operational cost, decision failures, and high-severity money-movement signals.
Weekly	Review drift, false positives, queue aging, new attack shapes, and temporary controls.
Monthly	Re-run evaluations, assess vendor performance, retire stale rules, and update operating playbooks.
Quarterly	Challenge assumptions, review decision rights, test incident response, and reassess the economics of abuse.

The goal is not a larger queue or a busier fraud team. It is a shorter time from new signal to measured response.

CONCLUSION

Constant vigilance

There is no set-and-forget solution to an adaptive adversary. There will not be one again.

The organizations that succeed will not reject vendor products. They will use them aggressively. They will combine broad external intelligence with the specific context of their own businesses. But they will not confuse purchasing a tool with building a capability.

Evaluations tell them whether controls still work against attacks they understand. Drift detection tells them when the ground is moving. Human investigation identifies the pattern no existing model has learned to represent. Instrumentation connects decisions to consequences. Temporary intervention makes the defense harder to predict. Confirmed outcomes feed the next cycle.

> **The model is not the advantage. The loop is the advantage.**

The organizations that lead the next decade of fraud prevention, identity, and payment integrity will not necessarily be the ones holding the most advanced tool. They will be the ones that never stopped watching.

SELECTED SOURCES AND FURTHER READING

[1] National Institute of Standards and Technology, AI Risk Management Framework (AI RMF 1.0), 2023.

[2] NIST AI Resource Center, AI RMF Playbook: Govern - Continuous Monitoring.

[3] NIST, Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations, 2025.

[4] Federal Reserve System, Synthetic Identity Fraud in the U.S. Payment System, 2019.

[5] FBI Internet Crime Complaint Center, 2024 IC3 Annual Report.