Domyn

Written by
Bhaskarjit Sarmah

When Algorithms Make Decisions, Who Verifies Them?

Understanding the oversight mechanisms behind AI-driven decisions and why organizations are moving beyond one-time testing toward continuous verification across the AI lifecycle.

Table of contents

Introduction

In November 2019, tech entrepreneur David Heinemeier Hansson sparked a firestorm on Twitter. Although his wife had a higher credit score, she received an Apple Card credit limit 20 times lower than his.

The culprit? **An algorithm.** Goldman Sachs, the bank behind Apple Card, couldn't explain the reasoning, evidence, or criteria behind this decision. They simply shrugged and said, "That's what the algorithm decided."

Welcome to the age of **algorithmic decision-making,** where machines approve loans, set insurance premiums, and determine a person’s credit risk. But here's the uncomfortable question keeping security professionals up at night: **when algorithms make decisions that affect millions of lives, who is actually verifying them?**

How algorithms became decision makers

The AI revolution is redefining how organizations **make decisions and act on them.** Algorithms now make thousands of decisions every second that once required human judgment. The promise? Faster, more efficient, and supposedly more objective decisions. The reality, however, is more complicated. According to the MIT AI Risk Repository, which catalogs over 1,700 documented AI risks, algorithmic decision-making falls into multiple risk domains, from discrimination and privacy violations to system failures and governance gaps. The repository's Causal Taxonomy reveals a troubling pattern: **many algorithmic failures occur post-deployment and unintentionally,** meaning organizations often don't discover problems until after they have already impacted outcomes. In other words, systems making life-altering decisions are often found to be flawed only after they've affected people.

The hidden risks of algorithmic decision-making

Back to the Apple Card incident. It turned out that the algorithm didn't just make one mistake; it systematically discriminated against women. Married couples with joint finances, identical credit histories, and shared assets found themselves receiving wildly different credit limits based solely on gender. The New York Department of Financial Services launched an investigation, but ultimately concluded that while the outcomes were discriminatory, they couldn't prove the algorithm intentionally violated fair lending laws.

This highlights a fundamental problem in algorithmic accountability: **opacity masquerading as objectivity.** While the issuing bank insisted that gender was never used as an input, the outcome still showed discrimination. Explicit gender data wasn’t necessary; proxies such as shopping patterns, zip codes, and even the type of smartphone used can usually stand in to create a bias.

In the Apple Card case, the algorithm learned to discriminate through patterns in historical data shaped by decades of systemic bias in financial services. As a result, women were denied access to credit they qualified for, limiting their purchasing power and financial independence. But most importantly, the incident exposed a critical gap: a consequential algorithm had been deployed without **adequate verification systems** to catch bias before it produced discriminatory outcomes.

But bias is only one manifestation of the problem. Across financial services, **AI now influences critical decisions, from credit approvals to trading strategies and regulatory reporting.** When these systems produce flawed outputs, the consequences are immediate and systemic. The real challenge is not simply identifying potential risks-it is **building infrastructure that continuously verifies AI systems** before those risks materialize.

For enterprise organizations, these failures often appear in less visible ways. Consider AI systems used for financial analytics or regulatory reporting. A model that generates an incorrect query or misinterprets data can produce results that **appear legitimate** but are fundamentally wrong. When those outputs feed into trading systems, credit decisions, or executive dashboards, flawed information can cascade across an organization's decision-making process. These failures are rarely caused by malicious intent-they emerge because most AI systems are **evaluated before deployment** but rarely **verified continuously** once they are in production.

The accountability gap in AI systems

Legally and ethically, **responsibility for algorithmic failures** often remains unclear. When the Apple Card algorithm discriminated, it was difficult to pinpoint who was accountable: the brand behind the card, the bank that deployed the system, the data scientists who built it, or the historical training data that encoded decades of bias?

In reality, the problem is structural. Most organizations lack infrastructure to **verify AI systems continuously across their lifecycle.** Without mechanisms to monitor outputs, stress-test models, and detect risks in real time, accountability becomes difficult to enforce.

From reactive governance to continuous verification

Solving this challenge starts with **addressing the foundation of every AI system: data.** Poorly structured, biased, or incomplete data inevitably leads to unreliable models and unpredictable outcomes. Organizations therefore need processes that ensure data is **clean, trustworthy, and governed before it ever reaches a model.**

But data quality alone is not enough. AI systems must be verified continuously after deployment-not treated as a one-time compliance exercise. Recent empirical research shows that Large Language Models used in financial decision-making can exhibit systematic biases-such as positional or representation bias-depending on prompt structure, dataset composition, or evaluation design. These findings highlight the need for rigorous testing and continuous monitoring to detect and mitigate risks both before deployment and in production.

At **Domyn,** we are developing capabilities that enable organizations to evaluate AI systems continuously in real-world environments, detecting risks such as **bias, privacy exposure, and unreliable outputs as they emerge.** By combining data preparation, pre-deployment red teaming, and real-time monitoring of live outputs, organizations can move from reactive governance to proactive verification. Instead of discovering failures after harm occurs, they gain the ability to identify risks early, understand how models behave in practice, and maintain accountability for every algorithmic decision.

Infrastructure, sovereignty, and the future of AI accountability

Yet effective oversight depends on more than monitoring models. Verifying AI systems also requires **control over the underlying infrastructure**-where data is stored, who can access it, and which jurisdiction governs it. Without this foundation, even the most advanced monitoring tools cannot guarantee accountability.

This is where **Europe's push for digital sovereignty** becomes directly relevant to algorithmic accountability. For **Microsoft,** digital sovereignty-covering legal compliance, data control, cybersecurity, and resilience-is a **strategic priority for European organizations,** centered on one question: who controls and protects your data? In a world where AI systems drive critical financial and operational decisions, that question extends beyond infrastructure to include the algorithms themselves.

For algorithmic systems, this means **proving GDPR and AI Act alignment, knowing exactly where training and inference data reside, enforcing strict access controls,** and **ensuring models can continue operating** even during provider outages or geopolitical disruptions. Microsoft positions its European cloud commitments as mechanisms to operationalize that control. Through the EU Data Boundary, European customer data is stored and processed within the EU. Customers can manage their own encryption keys, and sensitive access is restricted under programs like Data Guardian. Interoperability, portability, and multi-region redundancy are designed to **reduce lock-in and strengthen resilience.** The broader message is clear: digital sovereignty is not about isolation from innovation. It is about **structured control**-leveraging hyperscale cloud and AI capabilities while retaining jurisdictional clarity, encryption authority, and operational autonomy.

Verified AI: the next competitive frontier

Here’s the good news: beyond avoiding disasters, organizations that verify their algorithms gain a **significant competitive advantage,** reducing legal and reputational risk while building customer trust in an era of algorithmic skepticism. **Verified algorithms improve decision quality** by catching errors before they affect outcomes. And by reducing the need to constantly firefight failures, they enable **faster innovation** and **attract talent** eager to work for responsible organizations.

The next time an algorithm makes a consequential decision affecting your customers, your business, or your reputation, you should be able to answer these questions with confidence: How was this decision made? What verification processes did it pass through? Who's accountable if it's wrong? How do we know it's fair? If you can't answer these questions, you don't have an algorithmic decision system, you have an algorithmic liability.

Building verification infrastructure requires more than good intentions; it requires lifecycle governance platforms that span development, pre-production red teaming, production monitoring, and continuous surveillance. When you evaluate solutions, start by asking: does this platform detect bias, hallucinations, and privacy violations in real-time, or only during pre-deployment testing? Can it compute per-output trust scores that adapt to context, or does it rely on static benchmarks? Does it map to NIST AI RMF, EU AI Act conformity requirements, and Financial Model Risk Management standards?

These questions separate verification theater from operational accountability. The next generation of AI governance infrastructure is designed to answer them in real time. At **Domyn,** this is exactly the kind of capability we are building toward as part of our approach to **AI governance and verification.** Platforms that can answer “yes” to these questions **enable that shift from aspirational frameworks to production reality.** The era of blind trust in algorithms is over. The era of **verified, accountable, and transparent algorithmic decision-making** is here. Organizations can lead this transition or they can wait until the next algorithmic failure forces the conversation. By then, the competitive advantage will belong to someone else.

Sources

Apple Card's Credit Assessment Algorithm Allegedly Discriminated against Women. AI Incident Database. *https://incidentdatabase.ai/cite/92*

National Institute of Standards and Technology. (2026). NIST Risk Management Framework (RMF). Computer Security Resource Center. *https://csrc.nist.gov/projects/risk-management*

National Institute of Standards and Technology. (2025). NIST SP 800-53 Control Overlays for Securing AI Systems Concept Paper. *https://csrc.nist.gov/projects/cosais*

Slattery, P., Saeri, A., Noetel, M., Graham, J., & Thompson, N. (2025). MIT AI Risk Repository. MIT FutureTech. *https://airisk.mit.edu/*

Tracing Positional Bias in Financial Decision-Making: Mechanistic Insights from Qwen2.5 *https://arxiv.org/pdf/2508.18427*

Uncovering Representation Bias for Investment Decisions in Open-Source Large Language Models *https://arxiv.org/pdf/2510.05702*