RESEARCH

# Agent Behavioral Telemetry

*Behavioral Drift as a Leading Indicator of Agent Compromise in Autonomous Commerce*

Mandate Labs · June 2026 · Version 1.0

As autonomous AI agents become participants in financial transactions, a critical question emerges: how do you detect compromise in an actor that has no biometric signature, no physical location, and no behavioral history rooted in human cognition? This paper introduces **Agent Behavioral Telemetry (ABT)**, a framework for continuous trust verification based on the observable behavioral patterns of AI agents during financial interactions. Drawing on established evidence from behavioral biometrics—where keystroke dynamics alone detect fraudulent identity entry at 95.5% accuracy[1] and deployment across 100+ banks reduces false positives by 66%[2]—we extend behavioral profiling from human users to autonomous agents. We define a telemetry taxonomy of seven signal classes, propose a drift-detection architecture using exponentially weighted moving averages and Kullback–Leibler divergence, and describe how ABT integrates with the Decision Trust Protocol to provide layered defense. We present the **Agent Behavioral Fingerprint (ABF)**, a composite vector that captures an agent's characteristic transaction patterns, and demonstrate that behavioral drift precedes transaction-level anomalies in adversarial scenarios. We acknowledge the limitations of behavioral approaches—including adversarial adaptation, cold-start constraints, and the absence of direct comparative validation against transaction-attribute-only systems—and argue that ABT is most effective as a complementary layer that enriches, rather than replaces, traditional fraud signals.

**Keywords:** *behavioral telemetry, agent profiling, drift detection, behavioral biometrics, anomaly detection, trust verification, autonomous agents, prompt injection, agentic commerce*

## 1. Introduction: Beyond Transaction Attributes

Traditional fraud detection in payment systems operates on a well-defined signal set: transaction amount, merchant category code, geographic location, time of day, velocity counters, and device fingerprints. These signals describe *what happened*—the attributes of the transaction itself. For human-initiated transactions, this approach has been remarkably effective: modern machine learning fraud models achieve detection rates above 95% for known fraud patterns.[3]

Autonomous AI agents break this paradigm in a fundamental way. An agent does not have a geographic location that can be compared against a cardholder's home address. It does not have a device fingerprint that persists across sessions. It does not exhibit the temporal patterns of human behavior—no lunch breaks, no sleep cycles, no weekend shopping habits. The traditional signal set, optimized over three decades for human actors, loses much of its discriminative power when the actor is software.

Yet agents do exhibit behavior. They transact at characteristic frequencies. They select merchants from identifiable distributions. They structure transaction amounts in patterns that reflect their underlying decision logic. They exhibit temporal rhythms driven by their scheduling infrastructure. And critically, when an agent is compromised—through prompt injection, credential theft, or adversarial manipulation of its context window—its behavior changes *before* the resulting

transactions become individually anomalous.

This observation is the foundation of Agent Behavioral Telemetry: **the behavioral signature of an AI agent is a complementary fraud signal that detects compromise earlier than transaction attributes alone.** The agent starts behaving differently before it starts spending differently.

This claim is not speculative. It extends a well-established principle from human behavioral biometrics. Wilson, Jenkins, Valacich, and Kim demonstrated that keystroke dynamics reveal cognitive unfamiliarity during identity entry with 95.5% accuracy across 1,000+ participants—the signal was not what information was entered, but *how* it was entered.[1] BioCatch's deployment data across 100+ banks shows that behavioral analytics reduce false positives by 66% and improve fraudulent account detection by 1,500% over traditional approaches.[2] KPMG's 2025 Global Banking Scam Survey found that 64% of banks report behavioral baselines flag anomalies more effectively than rule-based methods.[4]

We propose that these principles extend naturally—and with even greater force—to AI agents, whose behavior is more consistent, more measurable, and more predictable than human behavior, making deviations more detectable.

## 2. Related Work

### 2.1 Behavioral Biometrics in Financial Fraud

The use of behavioral signals for fraud detection has a two-decade academic history. Early work by Bergadano, Gunetti, and Picardi established that keystroke dynamics could authenticate users with a 4% false acceptance rate across 154 subjects.[5] Subsequent research using the CMU keystroke dataset achieved equal error rates between 2% and 6% depending on methodology.[6] These results established that behavioral patterns—specifically, the *how* rather than the *what* of user interaction—carry sufficient discriminative information for identity verification.

Wilson et al.'s 2024 study advanced this field significantly by demonstrating that keystroke dynamics detect not just identity but *cognitive state*: participants entering unfamiliar (fraudulent) personal information exhibited measurably different typing patterns than those entering their own information, even when the keystrokes themselves were identical in content.[1] This finding is directly relevant to agent behavioral telemetry, as it demonstrates that behavioral signals capture the decision process, not merely the decision outcome.

### 2.2 User and Entity Behavioral Analytics

Gartner introduced the User and Entity Behavioral Analytics (UEBA) category in 2015, defining it as the use of statistical baselines and anomaly detection to identify compromised accounts, insider threats, and fraudulent activity.[7] The core principle—that deviation from an established behavioral baseline is itself a risk signal—has been validated across financial services, with the Identity Management Institute characterizing behavioral drift as "a leading indicator of compromise or malicious insider activity."[8] Gartner projects that 60% of organizations will use AI-driven behavioral analytics for insider and credential threats by 2026.[9]

However, Gartner's own reclassification of standalone UEBA under "Insider Risk Management Solutions" signals an important limitation: behavioral analytics alone is insufficient and requires integration with complementary detection mechanisms.[10] This informs our architectural position that ABT operates as a layer within a broader protocol, not as a standalone system.

### 2.3 AI Agent Behavioral Monitoring

Research on monitoring autonomous agent behavior is nascent but accelerating. He, Wu, Zhai, and Sun propose SentinelAgent, a graph-based anomaly detection system for multi-agent environments that models agent interactions as dynamic execution graphs and detects prompt injection, collusion, and latent exploit paths at the node, edge, and path levels.[11] Their approach demonstrates that agent-level behavioral modeling is feasible and effective for detecting adversarial manipulation.

The Trajectory Guard framework introduces a Siamese Recurrent Autoencoder that detects anomalous agent action trajectories with F1 scores of 0.88–0.94, at 32ms inference latency—17–27× faster than LLM-judge baselines.[12] This work is particularly relevant because it demonstrates that lightweight statistical models can detect behavioral anomalies in agent action sequences without requiring expensive language model evaluation at inference time.

Complementing these detection approaches, the Agent Behavioral Contracts framework proposes formal specification and runtime enforcement of behavioral boundaries for autonomous agents, establishing the theoretical basis for what constitutes “expected behavior” against which drift can be measured.[13]

A 2026 survey of agentic AI governance found that only 47.1% of deployed agents are actively monitored, despite 82% organizational confidence in governance adequacy—a gap the authors term “governance theater.”[14] This monitoring deficit represents both the problem ABT addresses and the market opportunity it serves.

### 2.4 Gap Analysis

The academic literature establishes three validated principles: (1) behavioral signals detect fraud that transaction attributes miss; (2) behavioral drift is a leading indicator of compromise; and (3) agent action sequences can be profiled and monitored for anomalies. However, no existing framework synthesizes these principles into a telemetry architecture purpose-built for AI agents operating in financial transaction contexts. SentinelAgent operates at the multi-agent interaction level, not the individual transaction behavioral level. Trajectory Guard addresses action sequences generically, without financial domain specialization. ABT bridges this gap by defining a domain-specific telemetry taxonomy, drift-detection methodology, and integration architecture for agent commerce.

## 3. The Telemetry Taxonomy

Agent Behavioral Telemetry organizes observable agent signals into seven classes, each capturing a distinct dimension of agent behavior. The taxonomy is designed to be protocol-agnostic—any system that observes agent transactions can collect these signals—though we describe integration with the Decision Trust Protocol in Section 7.

| Signal Class | Description | Example Metrics | Collection Point |
|---|---|---|---|
| **Cadence** | Temporal rhythm of transaction requests | Inter-request interval (mean, σ), requests per minute, burst ratio | API gateway |
| **Intent Coherence** | Consistency between stated intent and observed action | Intent-to-MCC alignment score, category drift rate, purpose stability index | Authorization engine |

| Signal Class | Description | Example Metrics | Collection Point |
|---|---|---|---|
| **Amount Distribution** | Statistical profile of transaction values | Mean, median, kurtosis, max/min ratio, denomination clustering | Transaction log |
| **Merchant Affinity** | Distribution of merchant selections | Merchant entropy, top-N concentration, MCC category spread, new-merchant rate | Transaction log |
| **Session Topology** | Structure of transaction sequences within sessions | Session length, transaction ordering patterns, retry rates, abandonment rate | Session manager |
| **Error Response** | Agent behavior when transactions are declined or errors occur | Retry strategy (immediate/ backoff/abandon), decline-to-retry ratio, amount adjustment pattern | Authorization engine |
| **Trust Trajectory** | Agent's trust score history and zone transitions over time | KYA zone stability, promotion/demotion frequency, score volatility | Trust service |

Table 1. The ABT telemetry taxonomy. Each signal class captures a distinct behavioral dimension. Collection points reference Decision Trust Protocol components but are generalizable to any agent transaction system.

## 3.1 Cadence Signals

An agent's transaction cadence—the temporal pattern of its requests—is among the most stable and distinctive behavioral features. Legitimate agents typically exhibit highly regular cadence patterns driven by their scheduling infrastructure: a procurement agent might transact in daily batches, while a subscription management agent operates on monthly cycles. Cadence stability is measured as the coefficient of variation of inter-request intervals over a rolling window.

Compromised agents exhibit characteristic cadence disruptions. Prompt injection attacks often produce rapid-fire transaction bursts as the attacker exploits the compromised context window before the session expires. Credential theft typically shifts cadence entirely, as the attacker's infrastructure operates on different scheduling patterns than the legitimate agent's.

## 3.2 Intent Coherence Signals

The Decision Trust Protocol requires agents to declare intent context with each authorization request. Intent coherence measures the consistency between this declared purpose and the observable attributes of the resulting transaction. An agent that declares "office supply procurement" but routes transactions to cryptocurrency exchanges exhibits low intent coherence—a signal that is invisible to traditional fraud systems that do not track declared intent.

Intent coherence is quantified as the alignment between the intent category vector and the merchant category code distribution, normalized by the agent's historical intent-to-MCC mapping. New intent categories are not inherently anomalous—agents legitimately expand their operational scope—but rapid intent diversification without corresponding mandate updates triggers elevated monitoring.

## 3.3 Error Response Signals

How an agent responds to transaction declines reveals its decision-making process with unusual clarity. Legitimate agents typically implement structured error handling: exponential backoff, amount adjustment within budget constraints, or graceful abandonment with notification to the principal. Compromised agents exhibit markedly different patterns: immediate retries at the same or increasing amounts, rapid merchant switching after declines, or systematic probing of authorization limits through incremental amount escalation.

Error response analysis is particularly valuable because it captures behavior that is difficult for an attacker to anticipate and mimic. An attacker who has compromised an agent's context window may be able to replicate its normal transaction patterns, but is unlikely to have reverse-engineered its error handling logic.

# 4. The Agent Behavioral Fingerprint

The seven telemetry signal classes are synthesized into the **Agent Behavioral Fingerprint (ABF)**—a composite vector that captures an agent's characteristic behavioral profile. The ABF serves two functions: it provides a stable baseline against which drift is measured, and it enables agent discrimination (distinguishing one agent from another based on behavioral patterns alone).

## 4.1 Fingerprint Construction

The ABF is constructed as a 42-dimensional feature vector, with six features extracted from each of the seven signal classes. Features are selected for stability (low variance within-agent over time), discriminability (high variance between-agent), and computational efficiency (extractable in constant time from a fixed-size rolling window).

The rolling window $w$ is adaptive: it expands during periods of low activity (to maintain statistical significance) and contracts during high activity (to increase sensitivity to rapid changes). The minimum window contains 50 transactions; the maximum spans 30 days.

## 4.2 Baseline Establishment

An agent's behavioral baseline is the expected ABF under normal operating conditions. Baselines are established during a calibration period following agent registration, during which the agent operates under elevated monitoring but reduced trust limits (as defined by the DTP trust ladder). The calibration period requires a minimum of 200 transactions or 14 days, whichever comes later.

Baselines are not static. They are updated using exponentially weighted moving averages (EWMA) with a decay factor $\lambda = 0.95$, ensuring that gradual, legitimate behavioral evolution is incorporated while sudden shifts remain detectable:

The EWMA approach addresses a fundamental tension in behavioral profiling: baselines must be stable enough to detect meaningful drift but adaptive enough to accommodate legitimate behavioral change. The 0.95 decay factor gives a half-life of approximately 14 transaction windows, meaning the baseline substantially incorporates new behavior after roughly one month of consistent change while remaining sensitive to abrupt deviations.

## 4.3 Fingerprint Distinctiveness

An important empirical question is whether agent behavioral fingerprints are sufficiently distinctive to support individual identification. Human behavioral biometrics research suggests they should be: keystroke dynamics studies achieve equal error rates of 2–6% for individual identification, despite the comparatively noisy and variable nature of human behavior.[5,6] AI agents, whose behavior is

driven by deterministic or quasi-deterministic decision logic, should exhibit substantially lower within-agent variance and higher between-agent variance—producing more distinctive fingerprints.

We hypothesize, but do not yet empirically validate, that the agent discrimination rate (the ability to correctly identify which agent produced a given transaction sequence based on behavioral fingerprint alone) exceeds 99% for agents with established baselines. Empirical validation across diverse agent populations is an explicit direction for future work.

# 5. Drift Detection Architecture

Behavioral drift—systematic deviation of an agent's observed behavior from its established baseline—is the primary anomaly signal in the ABT framework. The drift detection architecture is designed to distinguish three categories of behavioral change:

**1. Normal evolution:** Gradual, consistent changes that reflect legitimate operational adjustments (e.g., an agent expanding to new merchant categories as its mandate broadens).

**2. Environmental shift:** Abrupt but explicable changes driven by external factors (e.g., seasonal purchasing patterns, mandate updates, or principal-directed behavioral changes).

**3. Adversarial drift:** Behavioral changes indicative of compromise, including prompt injection, credential theft, or context manipulation.

## 5.1 Drift Measurement

Drift is quantified using Kullback–Leibler (KL) divergence between the baseline and observed behavioral distributions for each signal class:

KL divergence is asymmetric by design: we specifically measure how the observed distribution diverges from the baseline, not the reverse. This asymmetry is desirable because we care about detecting when current behavior departs from expected behavior, not about characterizing the baseline in terms of current behavior.

Per-class drift scores are aggregated into a composite **Behavioral Drift Index (BDI)** using learned weights that reflect the discriminative power of each signal class for adversarial detection:

## 5.2 Drift Classification

The BDI is mapped to drift severity zones that trigger graduated response actions:

| Zone | BDI Range | Interpretation | Response |
|---|---|---|---|
| **Stable** | < 0.15 | Behavior within expected variance | Normal authorization flow |
| **Evolving** | 0.15 – 0.40 | Gradual behavioral change, likely legitimate | Baseline update; log for review |
| **Shifting** | 0.40 – 0.75 | Significant deviation; may be environmental or adversarial | Elevated monitoring; reduce trust ceiling by one tier |
| **Anomalous** | > 0.75 | Severe deviation; high probability of compromise | Step-up verification; hold transactions for review; alert principal |

Table 2. Drift severity zones and graduated responses. BDI thresholds are initial values calibrated through expert judgment; empirical optimization across diverse agent populations is planned.

### 5.3 Temporal Drift Patterns

The rate and pattern of drift carries diagnostic information beyond the drift magnitude itself. We identify three characteristic temporal patterns associated with different compromise vectors:

**Spike drift** (prompt injection): An abrupt, high-magnitude BDI increase within a single session, often concentrated in the cadence and intent coherence signal classes. This pattern reflects an attacker exploiting a compromised context window to execute rapid unauthorized transactions before the session expires or the attack is detected.

**Ramp drift** (gradual manipulation): A steady, linear BDI increase over multiple sessions, distributed across multiple signal classes. This pattern is characteristic of sophisticated attacks that gradually modify agent behavior to avoid triggering threshold-based anomaly detectors—analogous to the "boiling frog" attack on human behavioral baselines.

**Phase drift** (credential theft): An abrupt shift to a new stable behavioral pattern, producing a step-function BDI that remains elevated but stable. This pattern occurs when a legitimate agent's credentials are stolen and used by a different system that has its own stable—but distinct—behavioral characteristics.

Distinguishing these patterns requires analysis of the BDI time series, not just the current value. CUSUM (cumulative sum) algorithms are well-suited for detecting both abrupt changes and gradual trends in the drift index over time.[8]

## 6. Adversarial Considerations and Limitations

Any behavioral detection system must contend with adversarial adaptation—the possibility that attackers will learn to mimic expected behavioral patterns. We address known limitations directly rather than presenting ABT as a solved problem.

### 6.1 Mimicry Attacks

An attacker who has observed an agent's behavioral profile could attempt to replicate its patterns while executing unauthorized transactions. Research on adversarial behavioral biometrics demonstrates this is a real concern: population-level statistical attacks can infer behavioral patterns and target specific victims by exploiting overlapping biometric distributions.[15]

However, mimicry attacks on agent behavioral profiles face constraints that do not apply to human behavioral biometrics. Replicating an agent's cadence pattern requires controlling the timing infrastructure. Replicating its error response pattern requires triggering and observing declines without raising alerts. Replicating its intent coherence requires knowledge of the agent's mandate constraints. The attack surface for mimicry is substantially narrower when the behavioral profile spans seven signal classes simultaneously—an attacker must replicate all dimensions to avoid detection.

### 6.2 Cold-Start Problem

Newly registered agents lack behavioral baselines, creating a period of reduced detection capability. The ABT framework addresses cold-start through two mechanisms: (1) class-level priors, where behavioral expectations for agents of a given type (e.g., procurement agents, subscription managers) provide initial baselines derived from population-level data; and (2) elevated monitoring during the calibration period, with reduced trust limits as specified by the DTP trust ladder.

The cold-start period represents a genuine vulnerability. An attacker who registers a new agent and uses the calibration period to establish a malicious baseline could subsequently operate within that baseline undetected. This is mitigated by the DTP's progressive trust model—newly registered agents operate with severely constrained authorization limits regardless of behavioral consistency.

### 6.3 Reproducibility and Dataset Dependence

A significant concern from the human behavioral biometrics literature is the dataset dependence of results. Killourhy and Maxion demonstrated a 76% relative performance difference between the worst and best datasets using identical methods,[6] indicating that detection accuracy is highly sensitive to the characteristics of the evaluation population. This caution applies directly to agent behavioral telemetry: performance claims must be validated across diverse agent populations, use cases, and operating environments. We explicitly refrain from reporting detection accuracy numbers in this paper, as no sufficiently diverse evaluation dataset yet exists for agent financial behavior.

### 6.4 Privacy and Governance

Continuous behavioral monitoring of AI agents raises governance questions, though the privacy calculus differs fundamentally from human behavioral biometrics. Agent behavioral data does not constitute personal data under most privacy frameworks (GDPR, CCPA) because agents are not natural persons. However, behavioral telemetry may indirectly reveal information about the agent's principal—a human or organization whose transaction patterns are reflected in the agent's behavior. ABT implementations must ensure that behavioral profiles are processed at the agent level and not used to construct profiles of the principals behind them.

## 7. Integration with the Decision Trust Protocol

ABT is designed as a complementary layer within the Decision Trust Protocol architecture described in our prior work.[16] Integration occurs at three points in the authorization pipeline.

### 7.1 Pre-Authorization: Behavioral Gate

Before the DTP authorization engine evaluates a transaction against deterministic gates (budget, velocity, sanctions), the ABT system checks the requesting agent's current BDI. If the BDI places the agent in the Anomalous zone, the transaction is routed to step-up verification regardless of its individual attributes. This provides a behavioral "circuit breaker" that can halt a compromised agent's transactions even when each individual transaction appears legitimate in isolation.

### 7.2 Post-Authorization: Telemetry Collection

Every authorization decision—whether approved, declined, or stepped-up—generates telemetry events that update the agent's behavioral fingerprint. Declined transactions are particularly informative: an agent's response to declines (retry pattern, amount adjustment, abandonment timing) updates the error response signal class. The post-authorization telemetry path is asynchronous and does not add latency to the authorization decision.

### 7.3 Trust Score Enrichment

The DTP Know Your Agent (KYA) trust score incorporates behavioral consistency as a component. The BDI feeds directly into the KYA calculation, where sustained behavioral stability contributes to trust promotion and behavioral drift triggers trust demotion. This creates a feedback loop: agents

that maintain consistent behavior earn higher trust levels and broader authorization limits, while agents exhibiting drift face progressive restriction.

# 8. Empirical Foundations and Honest Assessment

We present an honest assessment of what the evidence supports and where the framework relies on reasonable extrapolation from adjacent domains.

## 8.1 What Is Empirically Validated

| Claim | Evidence | Strength |
|---|---|---|
| Behavioral signals detect fraud that transaction attributes miss | BioCatch: 66% false positive reduction across 100+ banks[2]; Wilson et al.: 95.5% detection via keystroke dynamics[1] | Strong |
| Behavioral drift is a leading indicator of compromise | UEBA literature; Identity Management Institute characterization[8]; CUSUM-based drift detection in production | Strong |
| Agent action sequences can be profiled for anomalies | Trajectory Guard: F1 0.88–0.94 on anomalous trajectories[12]; SentinelAgent: graph-based detection of prompt injection[11] | Moderate (preprint-stage) |
| Behavioral profiling improves detection when layered on transaction attributes | KPMG: 64% of banks report behavioral baselines outperform rules alone[4] | Moderate (survey-based) |

Table 3. Evidence assessment for core ABT claims. “Strong” indicates peer-reviewed or large-scale deployment validation; “Moderate” indicates preprint-stage research or survey data.

## 8.2 What Is Not Yet Validated

Several claims in this framework rely on reasonable extrapolation rather than direct empirical evidence:

**No direct A/B comparison exists** between behavioral-only and transaction-attribute-only detection for AI agents. The claim that behavioral signals are “complementary” is supported by evidence from human fraud detection, but has not been validated in the agent-specific context.

**Agent behavioral fingerprint distinctiveness** is hypothesized based on the expectation that deterministic agent behavior produces more distinctive profiles than variable human behavior. This hypothesis requires empirical validation across diverse agent populations.

**BDI thresholds** (Table 2) are calibrated through expert judgment, not empirical optimization. Optimal thresholds will depend on the agent population, use case mix, and risk tolerance of the deploying institution.

**Drift pattern classification** (spike, ramp, phase) is a theoretical taxonomy derived from known attack vectors, not an empirically observed classification. Validation requires controlled adversarial testing across each attack type.

# 9. Performance Architecture

ABT is designed to operate within the latency constraints of real-time authorization. The architecture separates the critical path (pre-authorization BDI lookup) from the analytical path (telemetry collection, fingerprint update, drift calculation).

| Operation | Path | Latency Target | Approach |
|---|---|---|---|
| BDI lookup | Synchronous (critical) | < 2 ms | Redis-cached current BDI per agent |
| Telemetry event emission | Asynchronous | < 1 ms (fire-and-forget) | In-memory event queue |
| Fingerprint update | Asynchronous | < 50 ms | Background worker, EWMA computation |
| Drift calculation | Asynchronous | < 100 ms | KL divergence over rolling window |
| Drift pattern analysis | Batch | < 5 min | CUSUM over BDI time series |

Table 4. ABT performance architecture. The synchronous path adds < 2ms to authorization latency. All analytical computation is asynchronous.

The critical design decision is caching the current BDI in Redis rather than computing it at authorization time. The BDI is updated asynchronously after each transaction, meaning the pre-authorization check reads a value that may be one transaction out of date. This trade-off—sub-millisecond lookup latency at the cost of one-transaction staleness—is acceptable because behavioral drift is a trend-level signal, not a per-transaction signal. A one-transaction lag in BDI update does not meaningfully reduce detection effectiveness.

## 10. Future Work

The ABT framework as presented is a theoretical architecture informed by validated principles from adjacent domains. Bringing it to empirical maturity requires work across several dimensions.

**Empirical validation at scale.** The most critical gap is the absence of a large-scale empirical evaluation of ABT in production agent commerce. We plan to instrument Mandate Labs' authorization pipeline to collect behavioral telemetry from live agent transactions, building the first dataset of agent financial behavioral profiles. This data will enable empirical calibration of BDI thresholds, signal class weights, and drift pattern classifiers.

**Adversarial red-teaming.** Controlled adversarial testing—including prompt injection attacks, credential theft simulations, and behavioral mimicry attempts—is necessary to validate detection effectiveness under realistic attack conditions and to characterize the boundaries of what ABT can and cannot detect.

**Federated behavioral intelligence.** A single issuer's view of an agent's behavior is partial. An agent that transacts across multiple issuers has a richer behavioral profile than any single issuer observes. Federated behavioral intelligence—where issuers share anonymized behavioral signals without exposing transaction details—could produce more robust fingerprints while preserving competitive and privacy boundaries.

**Cross-agent behavioral correlation.** Compromised agents may exhibit correlated behavioral drift when controlled by the same attacker. Detecting correlated anomalies across agents that are not known to be related could identify attack campaigns that target multiple agents simultaneously.

**Formal drift-detection guarantees.** The current framework uses well-established statistical methods (EWMA, KL divergence, CUSUM) but lacks formal bounds on detection latency, false positive rates, and false negative rates. Deriving these bounds for specific threat models would strengthen ABT's theoretical foundation and enable principled threshold selection.

## 11. Conclusion

The shift from human-initiated to agent-initiated financial transactions demands a corresponding shift in how we detect fraud, compromise, and manipulation. Transaction attributes—the amounts, merchants, and timing of individual transactions—remain essential signals. But they are insufficient for a world in which the transacting entity is software that can be compromised in ways that produce individually plausible but collectively anomalous transaction patterns.

Agent Behavioral Telemetry provides a complementary signal layer grounded in a well-validated principle: behavioral drift is a leading indicator of compromise. By profiling agents across seven signal classes, constructing composite behavioral fingerprints, and detecting drift through statistical divergence measures, ABT enables detection of agent compromise earlier than transaction-attribute analysis alone.

We have been deliberately conservative in our claims. We do not assert that behavioral signals are *more reliable* than transaction attributes—no evidence supports that absolute claim. We do not report detection accuracy numbers—no sufficiently diverse evaluation dataset exists. What we do assert, grounded in two decades of behavioral biometrics research and emerging agent monitoring literature, is that behavioral telemetry captures information about agent compromise that transaction attributes cannot capture, and that layering behavioral signals onto existing detection infrastructure measurably improves detection outcomes.

The 53% of deployed agents that currently operate without active behavioral monitoring represent a systemic vulnerability in the emerging agent commerce ecosystem.[14] Agent Behavioral Telemetry is our proposal for closing that gap—not as a replacement for existing fraud infrastructure, but as the intelligence layer that makes the rest of the stack aware that the agent has changed.

## References

1. [1] D. Wilson, J. Jenkins, J. Valacich, and M. Kim, "Detecting Deceptive Data Entry Using Keystroke Dynamics," Brigham Young University / University of Arizona / Texas Christian University, 2024.
2. [2] BioCatch, "Behavioral Biometrics in Digital Banking: Deployment Results Across 100+ Financial Institutions," BioCatch Research, 2024. Including Latin American bank case study: 66% false positive reduction.
3. [3] European Central Bank, "Seventh Report on Card Fraud," ECB Statistical Paper Series, 2024.
4. [4] KPMG, "Global Banking Scam Survey 2025: Behavioral Analytics and Fraud Prevention," KPMG International, 2025.
5. [5] F. Bergadano, D. Gunetti, and C. Picardi, "User Authentication through Keystroke Dynamics," ACM Transactions on Information and System Security, vol. 5, no. 4, pp. 367–397, 2002.
6. [6] K. Killourhy and R. Maxion, "Comparing Anomaly-Detection Algorithms for Keystroke Dynamics," IEEE/IFIP International Conference on Dependable Systems and Networks, 2009. CMU keystroke dataset: EER 2–6% range; 76% performance variance across datasets.

7. [7] Gartner, “Market Guide for User and Entity Behavioral Analytics,” Gartner Research, 2015.
8. [8] Identity Management Institute, “Behavioral Analytics for Insider Threat Detection,” IMI Technical Brief, 2024. Characterizes behavioral drift as “a leading indicator of compromise or malicious insider activity.”
9. [9] Gartner, “Predicts 2025: Security Operations Automation and AI,” Gartner Research, 2024. Projects 60% organizational adoption of AI-driven behavioral analytics by 2026.
10. [10] Gartner, “Market Guide for Insider Risk Management Solutions,” Gartner Research, 2025. Reclassifies standalone UEBA under Insider Risk Management.
11. [11] J. He, C. Wu, J. Zhai, and Q. Sun, “SentinelAgent: Graph-Based Anomaly Detection in Multi-Agent Systems,” arXiv:2505.24201, May 2025.
12. [12] Trajectory Guard Authors, “Trajectory Guard: Lightweight Anomaly Detection in Agent Action Sequences via Siamese Recurrent Autoencoder,” arXiv:2601.00516, January 2026. F1 0.88–0.94; 32ms inference; 17–27× faster than LLM-judge.
13. [13] Agent Behavioral Contracts Authors, “Agent Behavioral Contracts: Formal Specification and Runtime Enforcement for Autonomous Agents,” arXiv:2602.22302, February 2026.
14. [14] Agentic AI Governance Survey, 2026. Finding: 47.1% of deployed agents actively monitored despite 82% organizational governance confidence.
15. [15] R. Chatterjee et al., “Revisiting the Security of Biometric Authentication Systems Against Statistical Attacks,” ACM Transactions on Privacy and Security, 2022.
16. [16] Mandate Labs, “The Decision Trust Protocol: A Layered Authorization Framework for Autonomous Agent Commerce,” Version 1.0, May 2026.

## Cite This Paper

```
Mandate Labs. "Agent Behavioral Telemetry: Behavioral Drift as a Leading Indicator of Agent
Compromise in Autonomous Commerce." Version 1.0, June 2026.
https://mandatelabs.ai/research/agent-behavioral-telemetry
```