C&R Software
Credit Risk Management

White Paper

The AI age demands a new approach to collections security

Introduction

Collections has always been a data intensive business. Account histories, payment records, contact details, bank account numbers: the data making collections operations run is precisely the data regulators and criminals find most interesting. As AI adoption grows, this exposure is growing in scale and complexity.

This report examines the evolving security and governance challenges facing collections operations in the age of AI. It explores why the stakes are higher than other business functions, what the emerging regulatory landscape demands, and how to build a security and governance architecture to scale.

Collections data is a prime target

Collections teams handle some of the most sensitive consumer data in financial services. A typical collections operation processes personal identifiers, financial account details, credit card numbers, payment histories, employment information, and health related data where medical debt is involved. Every one of these data categories carries its own regulatory obligation and penalty framework if mishandled.

The scale of the risk is well documented. Data breaches in the finance sector cost an average of $5.56 million per incident, second only to healthcare at $7.42 million. In 2025, 1 in 6 breaches involved AI enabled attack methods. As AI continues to expand what criminals can do, the cost and frequency of attacks is expected to climb.

This problem cuts both ways. Banks are deploying AI to improve collections operations, while criminals are deploying AI to exploit the data these operations hold. These are parallel races, and the institutions winning them are the ones treating internal AI governance as seriously as external threat defense. Yet 63% of breached organizations had no AI governance policy or were still developing one at the time of their breach, creating a compounding risk as AI deployment accelerates on both sides.

AI amplifies the security risk

AI doesn't introduce new risks in collections. It magnifies existing ones at a speed and scale traditional monitoring can't match.

Data volume and concentration.
AI systems require access to large volumes of historical account data to function effectively. The more data an AI model can access, the better it performs, but also the larger the blast radius if this access is compromised. AI models contain a trove of sensitive data ideal for attackers. Bad actors increasingly exploit this through prompt injection attacks, where malicious inputs are disguised as legitimate prompts, manipulating systems into exposing private records.

Agentic autonomy and access creep.
The shift from AI as a passive decision support tool to AI as an active agent fundamentally changes the risk profile. AI agents with broad, unconstrained access to tools, APIs, and databases create significant risks for data exfiltration and sensitive information disclosure. In 2025, 65% of organizations experienced at least one AI agent related security incident, with 61% of those incidents involving data exposure.

The supply chain and third party layer.
Collections technology increasingly relies on cloud infrastructure, LLM providers, and third party integrations. Third party vendor and supply chain compromise was the second most prevalent and second most costly attack vector in 2025, averaging $4.91 million per breach. Enterprises deploying AI agents need to account for the governance and security posture of every component in their stack, not just the components they build themselves.

5.56m
Average data breach costs in the finance sector

4.91m
Per supply chain breach for third party vendors

The attack surface in detail

Prompt injection has emerged as the defining security threat for AI enabled collections environments. Unlike traditional cyberattacks focused on software vulnerabilities, prompt injection enables adversaries to manipulate AI agents through the content they process. Documents, messages, or data inputs can carry embedded instructions redirecting agent behavior without any technical breach.

In agentic systems, this risk compounds. A misconfigured agent with excessive privileges and a susceptible response to injected prompts can be hijacked to escalate its own access, retrieve unauthorized data, or execute actions on behalf of an attacker. Researchers at Recorded Future project prompt injection will evolve into a mainstream enterprise attack technique, with threat actors increasingly prioritizing agent manipulation over traditional malware deployment.

Identity and access management faces similar pressure. As AI agents require cross-environment permissions to function, compromised agent credentials, SSO platforms, or identity tokens become high value targets capable of enabling large scale service disruption or data exfiltration. Enterprises will need to expand IAM frameworks to treat AI agents as priority digital identities, subject to lifecycle management, least privilege enforcement, and dedicated audit controls comparable to or stricter than those applied to human users.

The AI governance gap

The dominant failure in enterprise AI governance isn't a lack of policy. It's the application of uniform policy to agents with fundamentally different risk profiles.

Research from Gartner reveals applying a uniform governance strategy to all AI agents leads to higher project failure rates. The firm predicts by 2027, 40% of companies will decommission agents because their technical teams haven't distinguished between an agent's ability to act and the scope of access it should be granted. Shiva Varma, Senior Director Analyst at Gartner, describes the two failure modes resulting from blanket governance: over restricting simple agents, which slows delivery and drives shadow development, or under restricting autonomous agents, which increases security and compliance risk.

The consequences run in both directions. Too much restriction and banks find themselves unable to realize efficiency gains from AI. Too little, and they create unchecked exposure in exactly the operational environment, collections, where regulatory expectations are highest and consumer harm from data misuse is most direct.

Gartner advocates a proportional governance model, calibrated across four levels of agent autonomy:

Agent Behavior	Autonomy Level	Governance Requirements
Read / Summarize	Observe	Scoped data access, user authentication
Generate Recommendations with Human Review	Advise	Output quality review, hallucination testing, user training
Send Communications / Modify Configurations	Act with Approval	Meaningful control, approval workflows, audit logging
Execute Actions Independently	Act Autonomously	Maximum guardrails, human sampling, continuous monitoring

Governance, Varma emphasizes, shouldn't sit with a single executive. The model should be a shared, repeatable classification system owned across compliance, legal, risk, and technical functions, not a top-down decree.

The regulatory landscape: compliance is no longer optional

Collections operations deploying AI face an increasingly demanding regulatory environment across every geography where they operate.

United Kingdom

The FCA's Consumer Duty, confirmed as the centerpiece of regulation for 2025 to 2030, applies directly to AI driven customer interactions in collections. The FCA's April 2025 AI Update makes clear the Duty is the primary lens through which AI deployment will be assessed. Firms need to demonstrate how AI enabled journeys deliver good outcomes in practice, with documented evidence across fairness testing, transparency, and outcomes monitoring. The Senior Managers and Certification Regime (SM&CR) requires clear, named accountability for AI governance at a senior individual level.

The FCA also expects firms to embed quality thresholds and safety controls, including hallucination detection, for AI generated responses, and to build guardrails tying outputs to verified, approved information sources.

European Union

The EU AI Act classifies AI systems used in credit scoring and related consumer financial decisions as high risk, mandating transparency, impact assessments, and human oversight. ISO/IEC 42001, the international standard for AI management systems, is positioned as the primary framework for demonstrating compliant AI governance to EU regulators, providing a structured system addressing bias, explainability, security, and accountability across the full AI lifecycle.

United States

The regulatory landscape is more fragmented but no less consequential. The Colorado AI Act, effective 30 June 2026, imposes developer disclosure requirements, consumer notification obligations, and reasonable care standards for bias prevention on high risk AI systems in financial services. Illinois, California, and other states have introduced or enacted legislation requiring annual AI impact assessments and consumer notifications for automated decision making. The FDCPA and Regulation F apply fully to AI enabled collections activity, with the CFPB maintaining authority over how AI is used in consumer facing communications and decisions.

Canada

OSFI's updated Guideline E-23, effective May 1, 2027, establishes model risk management requirements for all federally regulated financial institutions, covering every AI and machine learning model in use regardless of whether it was built internally or sourced from a third party. Institutions need to maintain a comprehensive model inventory, assign risk ratings to every model, and govern the full lifecycle from design through decommissioning. OSFI has made clear institutions are expected to begin overhauling AI governance frameworks now rather than waiting for the 2027 effective date.

Financial institutions would be wise to build AI governance exceeding the minimum current requirement in each jurisdiction, because what counts as minimum is rising quickly.

Building security architecture for AI enabled collections

A robust security and governance architecture for AI in collections isn't built on a single control. It requires a layered model addressing data protection, access governance, auditability, and AI specific risk simultaneously.

Data protection at every layer

Protecting consumer data in a collections environment requires controls at three distinct levels: data in transit, data at rest, and individual field-level encryption for the most sensitive elements. PII, bank account numbers, and credit card details demand the strongest available protection. AES 256-bit field-level encryption is the recognized standard for this class of data. Transport layer security using HTTPS and TLS ensures data moving between users and application servers can't be intercepted. Passwords should never be stored in readable form, written to logs, or transmitted in clear text.

Data minimization principles are equally important in AI contexts. AI systems should be granted access only to the data they need to complete a specific task. An AI agent advising a collector on compliance policy has no legitimate need for customer account balance data. Architectural decisions constraining data access by design, rather than relying on downstream controls to prevent misuse, are materially more resilient.

Identity, access, and the least privilege model

Role based access control remains foundational, but in an agentic environment it needs to extend beyond human users to cover every AI agent as a distinct digital identity. Agents should be provisioned with the minimum privileges required for their specific function. These privileges should be time-bounded where appropriate, and any escalation of access should require explicit approval.

Single Sign-On integration with enterprise identity providers, whether cloud-hosted or on-premises, enables centralized credential management and ensures access governance for AI agents flows through the same framework as human users. This also enables consistent revocation: when an agent is decommissioned or an employee leaves, access is terminated through a single system rather than requiring manual cleanup across multiple platforms.

Audit, traceability, and the regulatory imperative

Regulators across every jurisdiction are converging on one requirement: if you can't prove what your AI did, when, and based on what data, you can't demonstrate compliance. The EU AI Act, Consumer Duty, FDCPA, and state AI legislation all demand documented, auditable evidence of AI system behavior.

A comprehensive audit framework for AI enabled collections needs to capture every account interaction, including all requests, views, and updates, together with the AI generated outputs associated with these interactions. Change history needs to be preserved in full, not just the current record state. Every data view should generate a log entry. HIPAA-regulated institutions require this level of PII access logging as a legal requirement; forward looking collections operations should treat it as baseline, regardless of sector.

For AI agent interactions specifically, audit logs need to capture which agent took which action, at what point in a workflow, and under what authorization. This isn't merely a compliance exercise. It's the mechanism enabling organizations to detect anomalous agent behavior before it becomes a breach.

Guardrails for agentic AI

Human in the loop design isn't a limitation on AI capability. It's the appropriate governance response to the autonomy level of the agent. An AI agent surfacing suggested responses for human review, rather than communicating directly with customers, eliminates an entire class of compliance risk. The human agent retains accountability; the AI delivers consistency and efficiency.

Policy driven response architecture, where AI outputs are grounded exclusively in approved, bank reviewed documentation and strictly prohibited from accessing external or public data sources, is the control mechanism keeping agentic behavior within defined compliance boundaries. Strict data segregation at the infrastructure level, ensuring policy knowledge bases and user interactions are never used to train underlying models, protects consumers and institutions from secondary data exposure.

What good looks like in the age of AI

For banking leaders evaluating AI deployment in collections, the governance conversation should be structured around five questions:

Do you have a proportional governance model?

Treating all AI agents with identical controls is the path to either stagnation or uncontrolled risk. Map every agent to its autonomy level and calibrate controls accordingly.

Is your audit capability complete?

Regulatory expectations require the ability to demonstrate, retroactively, who accessed what data, when, and what AI generated outputs were shown to or used by your collectors.

Have you treated AI agents as digital identities?

Every agent in your stack should have an identity, a defined scope of access, and a lifecycle. If you can't answer the question "what can this agent access and do?", the answer is almost certainly "more than you'd like".

Are your data protections layered?

Transport-layer encryption, encryption at rest, and field-level encryption for the most sensitive elements aren't alternatives. They're complementary controls addressing different attack vectors. Organizations relying on any single layer are accepting residual risk.

Is governance a shared function

The most durable AI governance frameworks are built and maintained by cross-functional teams, with compliance, legal, risk, technology, and business leadership operating together. A governance model owned by a single executive or team is, as Gartner observes, already a failure mode.

Conclusion

Regulators, auditors, and institutional clients are already evaluating the governance and security frameworks governing AI deployment in collections. The organizations coming out ahead are building the architecture to deploy AI responsibly. This means proportional governance calibrated to agent autonomy, layered data protections that don't rely on any single control, and audit capability complete enough to satisfy a regulator on the worst day imaginable.

Building this now, with intention, is considerably cheaper than reconstructing it after the first incident. To get started, **reach out directly**

Get in touch