OSCILAR

From Pilots to Production:
How Banks Are Building With AI

Introduction

Bank technology leaders find themselves in a familiar bind: McKinsey estimates that AI could add $200 billion to $340 billion in annual value to the global banking sector, equivalent to 9 to 15 percent of operating profits. The consulting firm's vision of an AI Bank of the Future is compelling: real-time decisioning, autonomous agents, hyper-personalized service. Yet according to an MIT study from August 2025, 95% of enterprise AI pilots fail to deliver any measurable financial impact. Most banks are stuck in what the industry has come to call "pilot purgatory," running dozens of isolated experiments that never scale.

The conventional wisdom says the only way out is "rip and replace" transformation: tear out the legacy core, rebuild from scratch, accept 18-month procurement cycles and eight-figure budgets. But this narrative is both paralyzing and wrong.

A different path exists. Banks can capture 70-80 percent of AI’s potential value by focusing on a small number of high-impact subdomains and deploying modern AI infrastructure alongside legacy systems — rather than attempting multi-year core replacements.

This piece examines that path. It builds on McKinsey & Company’s AI Bank of the Future framework, which positions AI as horizontal infrastructure spanning customer engagement, decisioning, core technology, and operating models. Taking that blueprint as a given, the focus here is execution: how banks can operationalize this vision under real regulatory scrutiny, legacy-system constraints, and risk-management requirements.

Practitioners with deep experience building real-time data systems inside large financial institutions consistently observe the same gap: while banks invested heavily in moving data faster, they failed to modernize how risk decisions are made once that data arrives. The result is a proliferation of siloed tools, brittle pipelines, and slow human workflows — at precisely the moment fraudsters are coordinating attacks across the entire customer journey.

TL;DR

- **Banks are stuck in “pilot purgatory.”** AI adoption stalls because risk decisioning remains siloed, procurement is slow, and governance models were built for deterministic (not agentic) systems.
- **Core replacement isn’t required.** Banks can capture 70–80% of the value of AI by layering a unified, real-time decisioning platform on top of legacy systems.
- **The fastest ROI is in the decision layer.** Fraud, AML, credit, and compliance use cases—especially shadow mode, analyst-assist agents, and AML overlays—deliver quick, low-risk gains.
- **Progress comes from “tractable wedges.”** Small, contained deployments running alongside existing systems let banks prove value, satisfy model risk management, and scale safely.
- **Governance accelerates adoption.** Explainability, human-in-the-loop controls, audit trails, and AI control towers enable faster deployment while staying compliant.

Part I: The Strategic Context

Why Legacy Systems Can't Keep Pace

The typical Tier 1 bank runs on infrastructure designed for a different era. Customer data sits in fragmented silos: separate databases for credit cards, mortgages, savings, and wealth management. Building a real-time, 360-degree view of any customer requires stitching together systems that were never meant to communicate. Decisioning logic is often hard-coded into mainframes or locked inside vendor "black boxes," making even simple rule changes a multi-week engineering exercise. In practice, this often means a customer can fail identity verification during onboarding and still be approved for credit or instant payments minutes later — simply because those systems never share signals.

This architecture was optimized for stability in a world where fraud moved at human speed. Today, attackers probe onboarding flows, test payment rails, and exploit credit systems in coordinated sequences — often within minutes. The FedNow Service, which now has over 1,400 participating financial institutions, and stablecoin rails demand instant settlement. Fraudsters use generative AI to launch attacks at machine speed. Customers expect the same responsiveness from their bank that they get from their streaming service. A system that updates overnight cannot defend against threats that evolve in minutes.

When risk systems operate independently, each sees only a fragment of the attack. What looks benign in isolation becomes obvious only when signals are evaluated together and in real time.

The McKinsey Blueprint

McKinsey's AI Bank of the Future framework proposes treating AI not as a set of isolated use cases, but as horizontal infrastructure spanning four layers of the organization: Engagement, Decisioning, Core Technology & Data, and Operating Model.

The framework is directionally correct. The challenge banks face is not understanding the destination, but navigating the constraints — regulatory approval, model risk management, legacy integration, and operational safety — required to reach it.

The Engagement Layer handles customer-facing interfaces: multimodal interactions across voice, text, and image, supported by "digital twins" that simulate customer behavior and enable personalized, proactive service. The decision layer is where data becomes action: AI agents and copilots perform real-time reasoning, orchestrate complex workflows, and achieve what McKinsey projects as 20 to 60 percent productivity gains. The Core Tech and Data Layer provides the foundation: unified data architectures that break down silos, vector databases that enable semantic search across unstructured data, and LLM pipelines to manage model lifecycles. The Operating Model Layer organizes the humans: cross-functional teams, AI "control towers" for governance, and accountability structures focused on outcomes rather than project completion.

The framework is coherent. The problem is execution.

Why Banks Freeze

Bank leaders face three structural barriers when they try to move from vision to implementation. A BCG survey found that only 25% of institutions have woven AI capabilities into their strategic playbook — the other 75% remain stuck in siloed pilots and proofs of concept.

The first barrier is procurement. The phrase “AI infrastructure” often triggers 18-month RFP cycles involving procurement, legal, and compliance. In a technology landscape evolving this quickly, solutions selected at the start of such processes may be obsolete by deployment. Many institutions attempt to compensate by building custom pipelines internally — assembling complex streaming, rules, and analytics stacks that require large engineering teams just to keep running. For most banks, this approach is economically unsustainable.

The second is governance. Model Risk Management frameworks were designed for static, deterministic models. Generative and agentic systems introduce probabilistic behavior that must be governed continuously — through explainability, lineage, and live performance monitoring — rather than through one-time approvals.

The third is control. Banks are rightly skeptical of "transformation" pitches that require outsourcing decisioning logic to third parties. Their risk policies and customer insights are intellectual property. Ceding control of that logic to vendors inverts the value proposition.

These barriers explain the gap between enthusiasm and action. The resolution lies in reframing the task: not as a wholesale transformation, but as a series of contained deployments that prove value and build institutional capability.

The scale of Transformation Ahead

The stakes are becoming clearer. A Citigroup report found that 54% of jobs across banking have high potential for automation, with an additional 12% that could be augmented by AI. According to Bloomberg Intelligence, global banks will cut as many as 200,000 jobs over the next three to five years as AI encroaches on tasks currently carried out by human workers. Singapore's DBS Bank has already announced plans to reduce its workforce by 4,000 positions over three years as AI takes over roles, while simultaneously deploying over 800 AI models across 350 use cases.

Yet Accenture's research suggests the opportunity outweighs the disruption: productivity could rise by 20 to 30 percent and revenue by 6 percent for banks that implement effectively. The question is not whether to transform, but how to do so without destroying operational stability.

Part II: Five Tractable Wedges

The McKinsey framework defines what an AI-first bank looks like. What it does not prescribe is how regulated institutions can move toward that future incrementally — without exposing customers, regulators, or balance sheets to unacceptable risk. The following “tractable wedges” reflect execution patterns observed in live deployments where banks proved value first, then scaled safely.

The following “tractable wedges” are execution patterns observed in live deployments that allow banks to progress toward the AI Bank of the Future incrementally.

A wedge is a targeted implementation that introduces modern AI infrastructure alongside legacy systems without requiring wholesale replacement. The following five wedges offer varying risk profiles and immediate applicability for risk and compliance leaders.

For an abridged table comparing the 5 wedges, see Figure A on page 21.

Wedge 1: Shadow Mode Decisioning

The problem: Banks hesitate to deploy new AI models directly into production because the cost of error is high. A false positive blocks a legitimate customer. A false negative allows fraud. Either outcome triggers regulatory scrutiny and customer attrition.

The solution: Shadow mode runs a new AI decisioning engine in parallel with the legacy system. Both systems receive the same live production data. Both make decisions. Only the legacy system's decisions execute. The AI system's decisions are logged for comparison.

How it works: Transaction and customer data streams are duplicated, often via API gateways or event streaming platforms like Kafka. One stream feeds the legacy rules engine; the other feeds the AI platform. Risk analysts compare decisions against actual outcomes. When the AI system flags a fraud ring the legacy system missed, that's evidence of "alpha." When the AI system would have blocked a legitimate customer, that's a tuning opportunity.

Why it works: Shadow mode is effectively backtesting and forward-testing on live data with zero production risk. It builds the empirical record Model Risk Management committees require. Once the AI system consistently outperforms legacy, the bank can flip the switch or gradually ramp traffic via canary deployment. Just as importantly, it enables continuous validation. Risk teams can measure false positives, false negatives, and drift over time — aligning far more closely with how regulators increasingly expect AI systems to be governed.

Risk Level: Low. No customer impact until the bank chooses to act on the evidence.

Wedge 2: Analyst-Assist Agents (L1 Triage)

The problem: Risk and compliance teams spend most of their time on mechanical work — gathering data, switching dashboards, and dismissing obvious false positives rather than exercising judgment. Legacy transaction monitoring systems generate false positive rates exceeding 90 to 95 percent. Global AML compliance costs now exceed $274 billion annually, with much of this going toward handling low-quality alerts rather than catching criminals. Human analysts spend most of their time on L1 triage: gathering data, copying between screens, dismissing obvious false alarms.

The solution: Deploy AI agents to perform the data gathering and preliminary analysis. The agent doesn't replace the analyst; it assembles the case file.

How it works: When an alert triggers, an "Investigation Agent" automatically queries internal databases, external watchlists, and relevant open sources. It synthesizes findings into a natural language case narrative that explains why the alert fired and recommends a disposition. The analyst reviews the assembled work product rather than starting from a blank screen.

Why it works: Case studies suggest this approach reduces manual review time by 75% and enables junior analysts to perform at senior levels. The human-in-the-loop remains for final judgment, which maintains governance while unlocking efficiency. The most effective deployments treat AI as an augmentation layer.

Agents gather context, rank alerts, and draft narratives — but humans retain decision authority, supported by complete audit trails and explainable rationales.

Risk level: Low to Medium. Humans retain decision authority; AI handles data aggregation.

Wedge 3: Greenfield Products (Stablecoins and Cryptocurrency)

The problem: Banks entering stablecoins, tokenized deposits, or crypto custody face a timing mismatch. These assets operate on blockchain rails that run 24/7 with instant settlement. The GENIUS Act, signed into law in July 2025, established the first federal regulatory framework for payment stablecoins, requiring federal banking agencies to adopt comprehensive rules by July 2026. Traditional banking risk systems designed for T+2 settlement cannot manage risk at blockchain speed.

The solution: Build an AI-native risk stack specifically for new product lines. Because these are greenfield deployments, there is no legacy system to displace.

How it works: Implement a platform capable of sub-100ms decisioning to match blockchain speed. Ingest both traditional fiat data (KYC, bank transfers) and on-chain data (wallet addresses, transaction graphs) to detect laundering networks that bridge both worlds. Use AI agents to enforce rules automatically, such as blocking transfers to sanctioned wallet addresses in real time.

Why it works: New products offer "safety by design." Success creates a reference architecture and internal expertise that builds the case for migrating legacy business lines to modern infrastructure.

Risk level: Medium. New product risk, but no legacy migration risk.

Wedge 4: Natural Language to Rule Generation

The problem: In traditional banks, changing a risk rule requires a risk officer to document the logic, hand it to IT, wait for a development sprint, and then wait for testing. The translation layer between business intent and code takes days or weeks. Fraudsters pivot in minutes.

The solution: Modern AI platforms allow non-technical users to create and test rules using natural language. A risk officer types: "Flag all transactions over $5,000 from IP addresses in high-risk jurisdictions if the account is less than 30 days old." The AI translates this into executable logic.

How it works: The risk officer inputs policy intent via a chat interface. The generative AI agent translates the prompt into decisioning code. The agent immediately runs a simulation (using shadow mode) to show impact on historical data: "This rule would have caught 50 fraud cases but triggered 200 false positives last week." Once validated, the rule deploys to production subject to governance approval.

Why it works: Eliminating the IT bottleneck lets risk teams respond to threats in real time. Business teams gain direct control over their tools while remaining within governance frameworks.

Risk level: Low with proper governance. Simulation prevents deployment of poorly-tuned rules.

Wedge 5: AML False Positive Reduction

The problem: Anti-Money Laundering (AML) is often the most expensive and least efficient compliance function. Rules-based transaction monitoring systems ("flag any cash deposit over $10k") generate massive alert queues with false positive rates of 95 to 98 percent.

The solution: Deploy an AI "overlay" that performs secondary scoring on alerts from the legacy system. The core Transaction Monitoring System remains untouched.

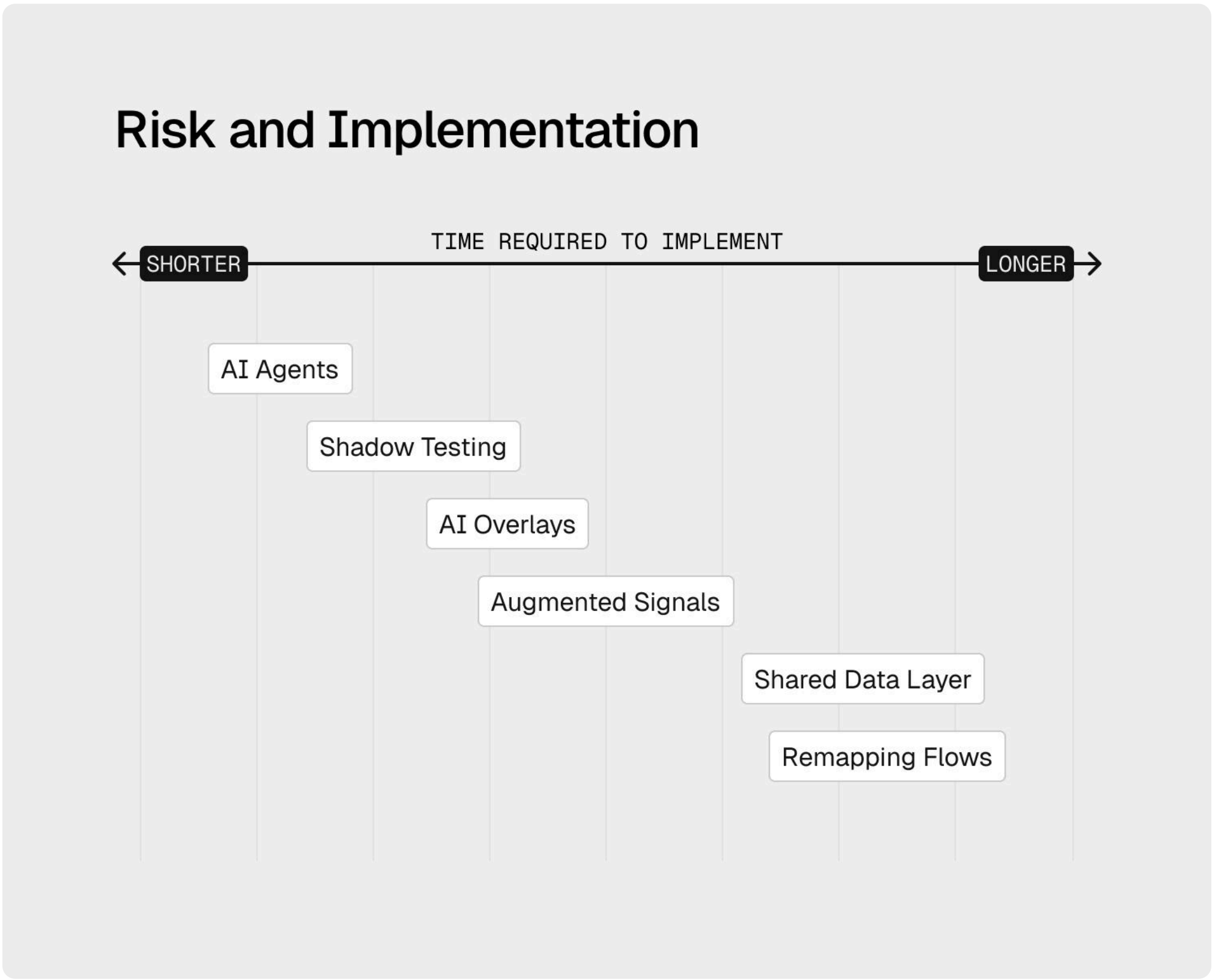
How it works: The legacy TMS generates alerts based on regulatory rules. These alerts pass to an AI agent that analyzes thousands of additional features: behavioral patterns, network links, device IDs.

The AI identifies alerts that are clearly false positives and auto-closes them with documented rationale for audit. Only high-risk alerts reach human investigators. The impact comes less from novel algorithms than from context. When onboarding, behavioral, transaction, and network signals are evaluated together, patterns that were previously invisible become obvious.

Why it works: Pure augmentation with no core replacement. Industry benchmarks suggest 40 to 70 percent reduction in manual workload. The ROI calculation is straightforward enough to survive CFO scrutiny.

Risk level: Low to Medium. Requires audit-grade documentation of suppression rationale.

↙ Figure B



Part III: Evidence From the Field

Over the last decade, financial institutions invested heavily in real-time data infrastructure — event streaming, APIs, and faster pipelines. Yet many encountered the same constraint: while data increasingly arrived in real time, risk decisions often did not.

Teams accumulated fragmented systems, each optimized in isolation. Some attempted to build custom decisioning layers on top of modern data stacks, only to find that sustaining accuracy, explainability, and compliance at scale required far more engineering effort than anticipated.

This was not unique to banks. Fintechs faced similar fragmentation and false-positive rates. The difference lay in operating constraints. Fintechs were often forced to iterate quickly because inefficiency surfaced immediately as loss or customer friction. Banks moved under stricter requirements for stability, auditability, and regulatory confidence.

What emerges is a clearer picture of where AI delivers reliable value today: backend decisioning — fraud detection, AML triage, and document processing. This produces measurable results because outputs can be verified, reviewed, and continuously improved. By contrast, AI systems that directly face customers or make irreversible financial decisions require greater caution and stronger controls.

The examples below reflect what we have observed through deployments of Oscilar's AI risk decisioning platform, illustrating how institutions have applied these lessons in practice — adopting unified decisioning and human oversight incrementally to capture the benefits of agility without assuming additional operational or regulatory risk.

MoneyGram, SoFi, and Nuvei: Consolidating Decisioning Under Governance

MoneyGram, SoFi, and Nuvei illustrate a common constraint at scale: even with modern data infrastructure, risk decisioning often lags behind the speed of money movement.

MoneyGram operates a global payment network spanning thousands of corridors and jurisdictions. As it expanded into instant settlement and digital assets, batch-oriented risk systems became increasingly difficult to adapt.

Rather than replacing downstream systems, MoneyGram consolidated fraud, AML, and onboarding decisions into a single decisioning layer. New rules and models were evaluated in shadow mode alongside existing controls, allowing teams to measure impact on live data before any production change. This enabled real-time decisioning suitable for stablecoin initiatives while preserving operational discipline. Teams reported up to 70% reduction in data migration time and the ability to evolve risk logic continuously without disrupting production workflows.

SoFi encountered a similar dynamic from a different starting point. Operating across lending, fraud, and collections, policy changes were often gated by engineering queues and fragmented tooling. By centralizing decision logic and enabling governed experimentation, SoFi reduced time-to-market for new risk strategies by roughly 50% and improved processing speed by more than 30%, while maintaining consistent oversight across products.

In payments environments like Nuvei's, the same architecture supports inline risk decisions under tight latency constraints, where the cost of error is immediate — either as fraud loss or customer friction — making shadow mode testing and clear escalation paths essential. As Daniel Hough, Director of Risk & Underwriting at Nuvei, noted “Our prior solution just didn’t offer the forward-thinking functionality we needed: AI, automation, or tools to make complex recommendations beyond basic rule sets. Oscilar gives us the flexibility and intelligence to manage our portfolio in entirely new ways, and that’s a big deal for us.”

The lesson: unify decisioning, validate accuracy with humans in the loop, then expand scope.

Flexcar, Dibsy, and Fluz: Reducing False Positives Through Shared Context

Across deployments with Flexcar, Dibsy, and Fluz, similar issues surfaced early: high false positives, manual review backlogs, and disconnected risk signals. Improvements came from evaluating signals together and tightening feedback loops, while retaining human accountability.

- Flexcar halved risk rates and reduced asset losses to zero by coordinating identity, behavioral, and transaction checks within a single decision flow, with human reviewers overseeing edge cases.
- Dibsy achieved an ~80% reduction in fraud while accelerating merchant onboarding fivefold, by assessing onboarding and transaction risk holistically rather than through separate tools.
- Fluz reduced manual reviews by roughly 90% and increased approval rates by about 20%, shifting alert triage and context gathering into automated workflows while keeping final decisions with human reviewers.

These outcomes reflect where AI performs best: backend decisioning and analyst support, where outcomes can be measured, audited, and improved continuously.

Clara, Cashco, and Parker: Empowering Risk Teams Without Engineering Dependency

In underwriting and compliance operations, delays are often driven less by model quality than by long engineering queues. Changes to decision logic can take weeks, limiting responsiveness to market conditions.

Deployments with Clara, Cashco, and Parker show how this constraint can be addressed without sacrificing control. By enabling risk teams to configure, test, and iterate on decision logic directly — within governed boundaries and with full audit trails — organizations shortened iteration cycles significantly.

Clara reported 3x faster onboarding, 3-4x higher throughput without additional headcount, and consistent SLA performance under growth. Cashco and Parker saw similar effects: underwriting deployments in days instead of weeks, ~70% reductions in backlog, and ~40% faster processing times.

For banks, this pattern reduces dependency on scarce engineering resources while preserving explainability and control.

TransPecos Banks: Incremental Modernization Under Regulatory Oversight

TransPecos Banks, a century-old community bank supporting multiple Banking-as-a-Service partners, faced rising AML complexity without the ability to scale headcount or regulatory exposure.

Rather than replacing core systems, TransPecos centralized AML decisioning and case management while maintaining existing controls. Alert triage, investigations, and SAR preparation were unified, with humans retaining final authority on every filing decision.

Operational impact included:

- 40% reduction in AML operations costs
- 70% reduction in alert review time
- 80% reduction in SAR management time
- $3M+ in projected annual savings

Equally important was the impact during regulatory examinations. Teams demonstrated complete audit trails and lineage — from alert to filing — within a single system, eliminating retroactive reconciliation across multiple tools.

One examiner noted: "This is the clearest risk management oversight we've seen from a bank your size." When AI performs mechanical work with full documentation and humans retain judgment on critical decisions, compliance becomes more defensible, not more complex.

What These Deployments Show

Across global networks, fintechs, and community banks, consistent patterns emerge:

- AI delivers immediate value in fraud detection, AML triage, and document processing — where outputs can be verified and reviewed
- The highest returns appear in the decision layer, not customer-facing automation
- Safe progress comes from shadow mode, measurable outcomes, and human-in-the-loop controls
- Unified decisioning reduces false positives and operational load without increasing risk

The dividing line is not between banks and fintechs, or between speed and safety. It is between organizations that treat AI as a governed learning system embedded in decisioning infrastructure, and those that deploy it as a collection of disconnected tools.

Banks can adopt the same architectural principles — unified decisioning, fast feedback loops, and clear human accountability — while rolling them out incrementally under full regulatory oversight.

In regulated risk operations, disciplined iteration with explainability and auditability consistently outperforms experimentation without guardrails.

Part IV: Governance as Enabler

In regulated domains, governance is not a constraint on AI adoption — it is the mechanism that makes adoption possible. AI systems that influence credit, fraud, or compliance outcomes must be explainable, auditable, and continuously monitored. Institutions that treat governance as living infrastructure, rather than static documentation, consistently move faster than peers.

Echoing McKinsey's emphasis on platform operating models and centralized governance, banks should establish an AI Control Tower. For AI agents to operate in production, especially in regulated domains like credit and AML, they must be explainable. Every decision requires a clear, human-readable rationale. The EU AI Act, which entered into force in August 2024 with full application expected by August 2026, classifies AI systems used for credit scoring as "high risk" and introduces additional safeguards. The European Banking Authority has found no significant contradictions between the AI Act and existing banking legislation, suggesting that existing frameworks can accommodate AI with integration effort.

The Control Tower acts as air traffic control for the bank's AI — monitoring performance, bias, and drift in real time; enforcing risk appetite and regulatory requirements; and ensuring successful innovations scale across the enterprise.

Effective oversight requires more than approval committees. It demands real-time monitoring of accuracy, drift, and human override rates — so issues are detected immediately, not months later during audits.

The execution phase has begun

The paralysis gripping many bank technology and risk leaders stems from a misconception: that modernization requires a perilous rip-and-replace of the core. The evidence points to a different path.

McKinsey's AI Bank of the Future framework correctly identifies the architectural end state; the wedges outlined here show how banks can reach that state pragmatically — without halting operations or surrendering control of decisioning logic. Banks can overlay intelligent, agentic infrastructure on top of legacy cores today. Start with shadow mode to prove safety. Deploy analyst-assist agents to relieve immediate operational pressure. Seize greenfield opportunities like stablecoins to build cloud-native risk stacks from the ground up.

The playbook for risk leaders: Own the infrastructure rather than outsourcing decisioning logic to point solutions. Start with the decision layer, where productivity gains of 20 to 60 percent and immediate risk reductions are found. Rewire entire domains end-to-end rather than piloting tools in isolation. Demand agency from your AI: systems that plan, route, and execute, not just chatbots that retrieve information. Use shadow mode and natural language tooling to accelerate innovation while maintaining regulatory oversight.

According to BCG research, AI agents already account for 17% of total AI value in 2025 across all industries, projected to reach 29% by 2028. Banks that execute on this will not merely be faster. They will be fundamentally different institutions: capable of reasoning in real time, adapting instantly to new threats, and serving customers with a level of personalization and security that batch-processing systems cannot match.

Over the next five years, the dividing line will not be between banks that "use AI" and those that do not. It will be between institutions that treat AI-driven decisioning as core infrastructure and those that continue to bolt it on at the edges.

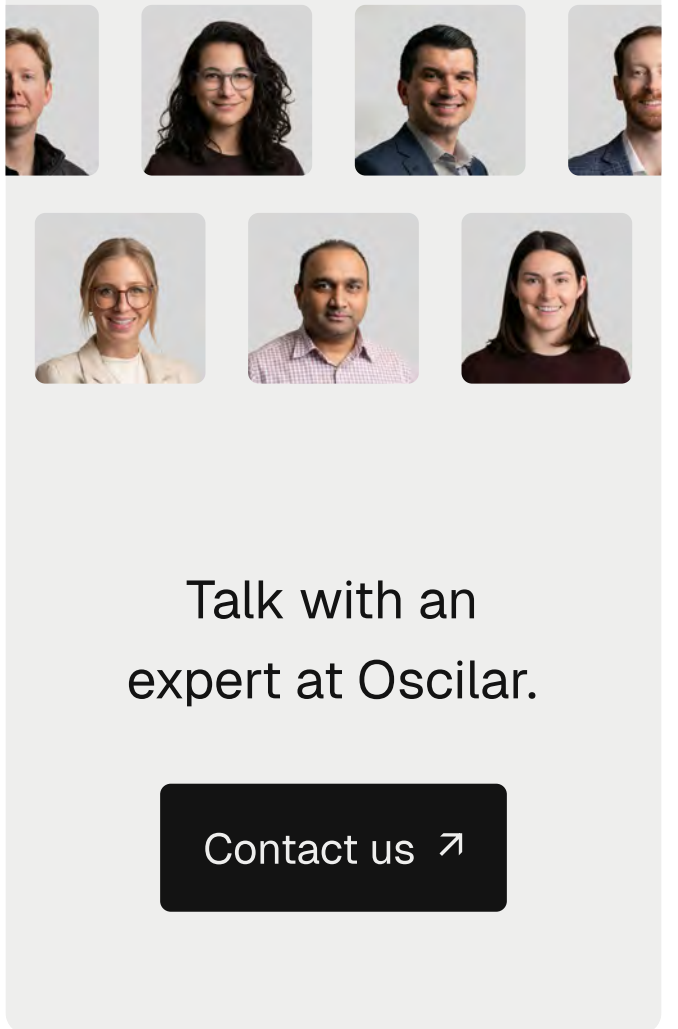
The blueprint exists. The wedges are available. The question is no longer whether AI will transform banking, but which institutions will be transformed first. Banks that cannot reason and act in real time across the entire customer journey will not merely fall behind, they risk becoming operationally irrelevant.

Appendix

↙ Figure A

Wedge Overview

Wedge	Description	Primary Benefit	Risk Level
Shadow Mode	Run AI models in parallel with legacy systems without executing decisions	Risk-free validation; empirical evidence for MRM	Low-Medium
Analyst Assist	AI agents perform L1 triage and draft case narratives for human review	40-75% time reduction; reduced analyst burnout	PayPal
Greenfield (Crypto)	Build AI-native stack for new products like stablecoins	No legacy debt; reference architecture for future migration	Medium
Natural Language Rules	Business users create rules via natural language prompts	Eliminates IT bottleneck; real-time threat response	Low
AML Overlay	AI performs secondary scoring on legacy TMS alerts	40-70% workload reduction in compliance ops	Low-Medium



Talk with an
expert at Oscilar.

Contact us ↗