Becoming an AI-Native Bank

A Practitioner Framework for the Operating Layer

Georgios Antikatzidis

Independent

antikas.io · g.antikas@gmail.com

August 2026

Abstract

Incumbent banks have invested in artificial intelligence while leaving much of the work that agents must perform embedded in fragmented data, documents, informal judgement and disconnected controls. This practitioner position paper calls the institutional foundation that makes such work executable the operating layer. It defines an AI-native bank as an organisation in which an authorised agent can act on verified business facts, codified policy, controlled workflows and callable capabilities. Each action must also be evaluated, attributable and open to human intervention. The argument combines the 2026 move towards enterprise deployment vehicles, public evidence on AI returns, established risk standards and professional observation in financial-services architecture. It applies the theory of constraints to distinguish model adoption from business value, then develops a working framework for operating-layer readiness. The framework links seven architectural properties to four phases of delivery: diagnostic, codification, build and bound, and handover. Regulatory binding and executive translation are integrated into that delivery discipline. The central claim is that codification creates a reusable institutional asset. AI can reduce the cost of producing the asset, while human owners preserve warrant through approval, versioning, audit and change control. Testable propositions define an agenda for future case research. The framework is conceptual and practice-informed; causal effects across banks remain to be tested.

Keywords: artificial intelligence, AI agents, banking, financial services, operating layer, operating model, enterprise architecture, theory of constraints, forward-deployed engineering, AI governance

JEL classification: G21, O32, O33, M15

Author note: The core argument appeared in essay form on LinkedIn and antikas.io on 13 May 2026. The author's professional practice is in financial-services enterprise architecture, which may shape the selection and interpretation of examples. No employer-confidential evidence is used. The author received no funding for this work and declares no competing interests.

1. Introduction

Banks are acquiring models and AI-enabled products faster than they are changing the institutional conditions under which those systems must work. A model can produce an

answer within seconds. A bank still has to establish which facts and policy version informed that answer, who authorised the action, how performance was tested and when a person must intervene.

An AI agent, as the term is used here, is a software system that can plan work, invoke tools, monitor outcomes and take bounded action through organisational interfaces. A bank becomes AI-native when an authorised agent can perform useful work inside those boundaries and the institution can observe, evaluate, interrupt and defend each action.

That capability depends on an operating layer: verified business facts, codified policy, explicit decision rules, controlled workflows, permissions, callable services, audit evidence and accountable ownership. These artefacts make the bank's work legible and executable. Durable AI value depends on building this layer around the model. Model capability remains important, but it cannot supply institutional authority or repair an operating model by itself.

The contribution is a practice-informed framework connecting enterprise architecture, operations management, forward-deployed engineering and AI governance. It identifies the properties that make a bank operable by agents and relates them to a delivery discipline. It then explains why codification may compound across deployments. Models can assist codification, while institutional warrant remains a human responsibility. Business value appears only when a material constraint moves.

The argument draws on public corporate announcements, industry surveys, standards, practitioner accounts and the author's professional observation. Corporate and vendor materials are used as evidence of market direction or claimed practice. Population-level causal inference lies outside the paper's scope. Its central claims are expressed as propositions and measurable variables for empirical testing.

2. Deployment has become the bottleneck

The enterprise AI market began to put substantial capital and delivery capacity behind deployment during 2026. OpenAI launched a deployment company with more than $4 billion of initial capital at a reported $10 billion valuation. Its partners include private-equity firms, consultancies and technology service providers [1, 2]. Anthropic announced a $1.5 billion venture with Goldman Sachs, Blackstone and Hellman & Friedman to deploy AI across private-equity portfolios [3]. Google Cloud and Accenture also announced joint forward-deployed engineering teams for complex customer work [4].

These vehicles share one feature: they place operating partners around the model and inside the customer organisation. They support a narrow conclusion that major providers expect enterprise deployment to require sustained work on customer processes, data, controls and governance. The announcements do not establish why each company invested. Distribution, revenue capture, portfolio exposure and financing structure are also credible explanations.

The stronger case for the operating layer follows from what an authorised agent must use. Models are portable across organisations. A bank's policies, customer facts, decision rules, permissions and control evidence are specific to the institution. They also determine whether

a model output can become a bank action. The operating layer is therefore both a production dependency and an institutional asset.

The claim builds on established research into organisational complements. Firm-level evidence links information-technology value to workplace reorganisation and new products or services [5]. Research on general-purpose technologies describes complementary intangible investment as a condition for later productivity gains [6]. Organisational scholarship also argues that learning algorithms reshape expertise, control, occupational boundaries and coordination [7]. The operating-layer framework applies those insights to the artefacts an authorised agent needs in banking.

Enterprise architecture has pursued many of these properties for decades. Agents change their operational importance because they consume the artefacts directly. A capability map filed after a transformation has little runtime value. A versioned service that an agent must call to establish a customer exposure becomes part of production. The same applies to policy, decision rights, evaluation evidence and ownership.

The market announcements are consistent with this interpretation, but they do not prove it. Their value is directional. They show that deployment has become an investable category of work at the same time that banks are deciding where durable AI capability should reside.

3. Business constraints and the codification flywheel

The theory of constraints provides a useful diagnostic for AI investment. Goldratt treats the organisation's goal as the starting point and regards improvement away from the binding constraint as wasted effort [8]. His later work also separates the business goal from the technology used to pursue it [9].

The transfer from manufacturing to bank knowledge work is heuristic. A bank has several business objectives and a distributed set of constraints. The discipline still applies: name the outcome, identify what currently limits it, define the intervention and measure whether that limit moves.

Relevant outcomes may include return on capital, cost to income, complaint ageing, decision time, regulatory response and customer retention. The binding constraint may sit in expert decision capacity, evidence collection, policy interpretation, approval latency or the cost of producing an auditable decision. An AI initiative creates business value only when it moves one of these measures without weakening service or control.

Agents expand the set of constraints that software can address because large language models can work with natural language and unstructured material. That capability remains conditional on the operating layer. An agent cannot make a defensible regulated decision from an ambiguous policy, unclear decision rights, an unowned customer record or an interface with no effective permission boundary.

AI can also help build the layer it needs. Models can assist teams to extract candidate rules from documents, draft a business vocabulary, record decisions and turn interviews into proposed policy. The responsible human owners then verify, amend, govern and adopt those artefacts. Accepted outputs become versioned institutional knowledge that later deployments can reuse.

This creates a potential codification flywheel. Each deployment can leave behind better policy artefacts, evaluation cases, business definitions and callable capabilities. Reuse may lower the time and cost of the next deployment, while the next deployment produces further evidence and structure. Compounding depends on deliberate retention; disposable project artefacts do not create it.

A plausible counterargument is that improving models will eventually work directly from disordered documents and systems. That possibility reduces the cost case for some codification. It does not resolve institutional warrant. A plausible reading of a policy remains insufficient for regulated action until the bank has verified, versioned, owned and governed it. Better models can lower the cost of codification, but they cannot assign institutional authority to their own interpretation.

4. A working framework for operating-layer readiness

The operating-layer claim becomes useful only when a delivery team can assess it. Empirical research has shown that organisations use different enterprise architecture artefacts for distinct practical purposes [10]. The framework below translates the operating-layer claim into seven assessment criteria. The framework is provisional, and the necessity and sufficiency of its criteria remain to be tested.

1. *Authoritative business facts.* Each material fact has an identified source, owner, definition and effective date. Customer, product, exposure, policy and control data can be traced to the record the bank recognises as authoritative.
2. *Explicit decision rights.* The bank records who may decide, delegate, approve, override and change a rule. These rights are available at the point where work is performed.
3. *Visible workflows and bounded exceptions.* The normal path is observable, decision points are explicit, handovers are recorded and exceptions enter a controlled process. Hidden manual work cannot remain the integration mechanism for an automated agent.
4. *Evaluation that gates production.* The agent is tested against defined outcome, quality, safety and regression measures before release and while operating. Practitioners often shorten this discipline to evaluation or "eval".
5. *Access and audit at agent level.* Permissions apply to the agent and the task it performs. Every material action leaves evidence that supports reconstruction, challenge, remediation and accountability.
6. *A formal business ontology.* The bank maintains a governed record of its entities, relationships, capabilities and rules. Critical meaning cannot remain solely in documents or the memories of experienced staff.
7. *Callable organisational capabilities.* Agents use controlled interfaces to retrieve facts, invoke decisions, execute permitted work and record outcomes. Current implementation terms include Model Context Protocol servers, agent skills and sub-agents [11], although the architectural property is vendor-neutral.

Their combination turns architecture artefacts into production dependencies. The framework treats selected artefacts as readiness criteria for agent operation and as concrete delivery targets.

5. The incumbent starting position

Readiness within an incumbent bank is uneven. Digital channels often coexist with fragmented records, document-based policy and manual reconciliation. People bridge the gaps between systems and interpret rules at the point of work. The pattern varies by bank and business area, so a universal maturity claim would be unjustified.

Regulation has already required more structure in some functions. Treasury, market risk, credit decisioning, wholesale trading and model risk may have stronger data, policy, model and audit disciplines than other areas. Even there, artefacts built for a regulator or specialist may be difficult for an agent to use. These functions are candidate starting points only when the local evidence supports readiness.

Other areas may hold more economic value and require more codification. Product development, customer engagement, complaints, vulnerability handling, operational risk and change management often depend on judgement distributed across people, documents, spreadsheets and several systems. The gap is organisational as much as technical.

Public evidence supports a cautious version of this diagnosis. McKinsey's 2025 survey of 1,993 respondents found that 5.5 per cent attributed at least 5 per cent of organisational earnings before interest and taxes to AI [12]. BCG's 10-20-70 heuristic assigns most transformation effort to people and process, with smaller shares assigned to technology, data and algorithms [13]. The first is a self-reported outcome statistic and the second is a consultancy heuristic. They neither measure the same construct nor establish cause. Both place the difficult work beyond model selection.

The practical implication is that a bank's data strategy and its AI-native programme should be managed as connected work. Data platforms, ownership, lineage and product governance already produce parts of the operating layer. The AI programme extends that effort into unstructured knowledge, decision rights, policy, callable capabilities and agent-level control.

Candidate workflows should be selected through evidence. A useful starting case has a material business constraint, sufficient policy and data readiness, a measurable baseline and a risk profile that can be bounded. Regulatory structure may improve readiness, but reputation alone is an inadequate selection criterion.

6. Forward-deployed engineering adapted for banking

Forward-deployed engineering embeds engineers in a customer organisation to build production systems against its own work and data. Palantir established an early delivery model around this discipline [14, 15, 16]. Recent AI providers and service firms have adopted similar embedded roles [4, 11].

The method suits operating-layer work because codification requires proximity to policy owners, operators, control teams and real exceptions. For a bank, the discipline can be organised into four phases.

Diagnostic. The team begins with a material business constraint and a baseline measure. It maps the workflow, evidence, policy, systems, decision owners and failure modes. The output is a bounded deployment case with an explicit account of missing operating-layer properties.

Codification. The team converts relevant knowledge into governed artefacts. These may include policy rules, business definitions, decision tables, exception paths, evaluation cases and ownership records. Models can draft and extract; named human owners approve meaning and authority. The artefacts remain readable and changeable by the bank.

Build and bound. Engineers connect the agent to controlled capabilities, implement permissions, create audit evidence and establish evaluation gates. The release decision depends on pre-agreed outcome and regression measures. Public engineering guidance now treats evaluation as part of system design [17].

Handover. The bank assigns operational ownership, change rights, override authority, evaluation maintenance and incident response before launch. The external team transfers the artefacts and knowledge needed to operate the capability. A deployment that remains dependent on an external team has not yet created an internal bank capability.

Regulatory binding runs through every phase. The NIST Artificial Intelligence Risk Management Framework places governance across the AI lifecycle [18]. ISO/IEC 42001 places leadership, responsibility and management controls around an AI management system [19]. In banking, the delivery design must also address model risk, third-party risk, data residency, audit completeness and accountable ownership.

Klarna's public account illustrates why business outcomes and quality thresholds must be measured together. In February 2024, the company reported that its assistant handled work equivalent to 700 full-time customer-service agents. It also projected a $40 million profit improvement for that year [20]. The company later resumed hiring human agents [21]. Its chief executive said that cost had become too predominant and that the result was lower quality [22]. The case does not establish that evaluation was absent. It shows that a cost measure can improve while service quality deteriorates.

A bank deployment should therefore set the business outcome, quality floor, regression thresholds and escalation path before release. Vendor descriptions of outcome pricing show that resolved interactions and conversations can become commercial units [23, 24]. Those descriptions do not establish impact. They do illustrate how a measurable unit can structure accountability.

The compounding claim also changes the definition of completion. A deployment should leave behind reusable policy, tests, interfaces, business definitions and operating evidence. The next case can then measure how much time, cost, risk and external dependency those retained assets remove.

7. Institutional translation and programme ownership

Operating-layer work requires executive commitment because it changes how the bank records decisions, owns policy, governs risk and measures delivery. Technical language can obscure that consequence. Executive translation is therefore part of the strategy design.

The translation should begin with the business constraint and the measure expected to move. Architectural terms follow only when they help a decision. For example, a business ontology can be introduced as the bank's governed record of its activities, entities, relationships, rules and owners. The plain description carries the decision; the technical term supports precision.

The McKinsey result provides useful context when stated accurately: 5.5 per cent of respondents attributed at least 5 per cent of organisational earnings before interest and taxes to AI [12]. The figure does not estimate a general failure rate. It does challenge claims that adoption alone produces material earnings impact.

The same discipline applies to procurement. A credible proposal identifies the business metric, its baseline, the measurement period, quality thresholds, commercial remedies, handover obligations and the internal owner. These terms expose whether the proposal builds institutional capability or leaves the bank dependent on a supplier.

One named internal community should own both codification and translation. It needs working relationships with business operations, technology delivery, data and AI teams, risk, audit and executive leadership. Separating the technical artefacts from the executive account creates two strategies that can drift.

Executive ownership of AI governance should also be explicit. McKinsey reports that CEO oversight is the governance element most associated with higher self-reported earnings impact in its cross-industry survey [12]. Extending that result to a specific bank requires caution. The practical inference is that delegated committees do not remove the need for a named executive who can resolve conflicts and accept accountability.

A credible programme begins with a shared outcome thesis, an evidence-based readiness assessment, a named executive owner and a small set of bounded workflows. The first deployment is treated as an operating-layer build with an agent use case. Its retained artefacts and measured results form the starting evidence for subsequent cases.

8. Testable propositions and research design

The framework produces claims that can be tested through bank deployments.

Proposition 1. For comparable agent use cases, stronger operating-layer readiness before build will be associated with better movement in the named business metric and fewer production regressions.

Proposition 2. Reuse of codified policy, evaluation cases, business definitions and callable capabilities will reduce the marginal time and cost of later deployments.

Proposition 3. Workflows in more structured regulated functions will reach controlled production faster when their local data, policy, control and ownership evidence is stronger. The regulatory reputation of a function will not predict readiness by itself.

Proposition 4. Explicit handover, internal ownership and maintained evaluation will reduce continuing external dependency and improve operational continuity after the delivery team leaves.

A longitudinal multiple-case study could test these propositions without claiming statistical generalisation to all banks. The unit of analysis would be one bounded workflow deployment. Before build, researchers would record its readiness against the seven properties, business baseline, risk classification and external dependency.

During and after delivery, the study would measure time to controlled production, movement in the named outcome, quality regressions, incidents, artefact reuse and continuing supplier dependence. Cases that fail or are stopped are essential evidence because successful deployments alone can show possibility but cannot estimate propensity.

9. Conclusion

An AI-native bank is one in which authorised agents can act on verified facts, governed policy, bounded permissions and controlled capabilities while the institution retains evidence and accountability. The model is one component of that arrangement. The operating layer determines whether model output can become defensible bank action.

Incumbent banks may already hold parts of this foundation in their data platforms, regulated functions, model governance and operational controls. The work is to assess those assets honestly, codify what remains tacit, connect capabilities through controlled interfaces and retain the result inside the institution.

The framework presented here makes that work assessable and the central claims testable. Future deployments should report the starting constraint, operating-layer changes, business result and retained internal capability. That evidence would show whether codification improves outcomes and compounds across cases.

References

[1] OpenAI. "OpenAI launches the Deployment Company." Company announcement, 11 May 2026. https://openai.com/index/openai-launches-the-deployment-company/ (accessed 25 August 2026).

[2] Bloomberg News. "OpenAI finalizes $10 billion joint venture with PE firms to deploy AI." 4 May 2026. https://www.bloomberg.com/news/articles/2026-05-04/openai-finalizes-10-billion-joint-venture-with-pe-firms-to-deploy-ai (accessed 25 August 2026).

[3] CNBC. "Anthropic, Goldman Sachs, Blackstone and Hellman & Friedman launch $1.5B enterprise AI venture." 4 May 2026. https://www.cnbc.com/2026/05/04/anthropic-goldman-blackstone-ai-venture.html (accessed 25 August 2026).

[4] Accenture. "Accenture and Google Cloud expand partnership to scale agentic transformation for global enterprises with Gemini Enterprise." Press release, 22 April 2026. https://newsroom.accenture.com/news/2026/accenture-and-google-cloud-expand-partnership-to-scale-agentic-transformation-for-global-enterprises-with-gemini-enterprise (accessed 25 August 2026).

[5] Bresnahan, T. F., E. Brynjolfsson, and L. M. Hitt. "Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence." *The Quarterly Journal of Economics* 117, no. 1 (2002): 339-376. https://doi.org/10.1162/003355302753399526

[6] Brynjolfsson, E., D. Rock, and C. Syverson. "The Productivity J-Curve: How Intangibles Complement General Purpose Technologies." *American Economic Journal: Macroeconomics* 13, no. 1 (2021): 333-372. https://doi.org/10.1257/mac.20180386

[7] Faraj, S., S. Pachidi, and K. Sayegh. "Working and Organizing in the Age of the Learning Algorithm." *Information and Organization* 28, no. 1 (2018): 62-70. https://doi.org/10.1016/j.infoandorg.2018.02.005

[8] Goldratt, E. M., and J. Cox. *The Goal: A Process of Ongoing Improvement.* Great Barrington, MA: North River Press, 1984.

[9] Goldratt, E. M., E. Schragenheim, and C. A. Ptak. *Necessary But Not Sufficient.* Great Barrington, MA: North River Press, 2000.

[10] Kotusev, S. "Enterprise Architecture and Enterprise Architecture Artifacts: Questioning the Old Concept in Light of New Findings." *Journal of Information Technology* 34, no. 2 (2019): 102-128. https://doi.org/10.1177/0268396218816273

[11] Anthropic. "Forward Deployed Engineer" job posting, Greenhouse, 2026. Archived at https://web.archive.org/web/20260413215943/https://job-boards.greenhouse.io/anthropic/jobs/5121563008 (snapshot 13 April 2026; accessed 25 August 2026; the live posting has been removed).

[12] McKinsey & Company, QuantumBlack. *The state of AI in 2025: Agents, innovation, and transformation.* November 2025. Survey fielded 25 June to 29 July 2025; 1,993 respondents in 105 countries. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai (accessed 25 August 2026).

[13] Boston Consulting Group. "The Leader's Guide to Transforming with AI." Undated. https://www.bcg.com/featured-insights/the-leaders-guide-to-transforming-with-ai (accessed 25 August 2026).

[14] Palantir Technologies Inc. Form S-1 Registration Statement. US Securities and Exchange Commission, 25 August 2020. https://www.sec.gov/Archives/edgar/data/1321655/000119312520230013/d904406ds1.htm (accessed 25 August 2026).

[15] Lonsdale, J. "Why we built Palantir." Undated. https://joelonsdale.com/why-we-built-palantir/ (accessed 25 August 2026).

[16] Orosz, G. "What is a Forward Deployed Engineer?" *The Pragmatic Engineer.* Undated. https://newsletter.pragmaticengineer.com/p/forward-deployed-engineers (accessed 25 August 2026).

[17] Anthropic. "Demystifying evals for AI agents." *Anthropic Engineering.* Undated. https://www.anthropic.com/engineering/demystifying-evals-for-ai-agents (accessed 25 August 2026).

[18] National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0).* NIST AI 100-1, January 2023, GOVERN function. https://doi.org/10.6028/NIST.AI.100-1

[19] International Organization for Standardization. *ISO/IEC 42001:2023 - Information technology - Artificial intelligence - Management system.* December 2023, clause 5. https://www.iso.org/standard/81230.html (accessed 25 August 2026).

[20] Klarna. "Klarna AI assistant handles two-thirds of customer service chats in its first month." Press release, 27 February 2024. https://www.klarna.com/international/press/klarna-ai-assistant-handles-two-thirds-of-customer-service-chats-in-its-first-month/ (accessed 25 August 2026).

[21] Maginative. "Klarna dials back its AI customer service strategy - now it's hiring humans again." May 2025. https://www.maginative.com/article/klarna-dials-back-its-ai-customer-service-strategy-now-its-hiring-humans-again/ (accessed 25 August 2026).

[22] Siemiatkowski, S. Interview with Bloomberg News, 8 May 2025; quoted in [21].

[23] Sierra. "Outcome-based pricing for AI Agents." Company blog. Undated. https://sierra.ai/blog/outcome-based-pricing-for-ai-agents (accessed 25 August 2026).

[24] Decagon. Company website (per-conversation and per-resolution pricing described). https://decagon.ai/ (accessed 25 August 2026).