International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Governing Autonomous Decision Authority: A Handoff-Threshold Framework for Agentic AI in Payment Infrastructure

Gokulakrishnan Seshiah

Abstract: Agentic AI systems now operate autonomously inside production payment infrastructure resolving CI/CD dependency conflicts, scoring transactions for fraud in real time, and managing the cadence of model deployment yet existing AI governance frameworks stop short of specifying a concrete, auditable mechanism for dividing decision authority between agent and human. This article addresses that gap directly. Drawing on practitioner experience operating agentic systems at global payment scale, we present a taxonomy of agentic deployment patterns in payment infrastructure, an architecture description of a layered real-time fraud-inference system, and a four-tier handoff-threshold governance framework full-autonomy, advisory-autonomy, confirm-before-act, and human-only mapped explicitly onto the NIST AI Risk Management Framework's Govern function and the human-oversight provisions of Article 14 of the EU AI Act. The framework turns what is otherwise an implicit, emergent boundary on agent authority into a documented control that risk and compliance functions can inspect and challenge.

Keywords: *agentic AI; payment infrastructure; autonomous CI/CD; real-time fraud decisioning; human-AI collaboration; AI governance; handoff threshold; NIST AI Risk Management Framework; bounded autonomy; escalation design; financial technology; responsible AI deployment*

1. Introduction

Payment infrastructure has long relied on deterministic pipelines, fixed deployment schedules, rule-based fraud filters, human-supervised release cycles. These architectures sit increasingly at odds with the scale and speed of modern payment processing, where deployment cycles need to compress from months to hours and fraud decisions have to land within milliseconds. Agentic AI software capable of goal-directed, multi-step autonomous action is now closing that gap in production: autonomous agents resolve CI/CD dependency conflicts, orchestrate deployment sequencing, and score transactions for fraud risk in real time.

That shift surfaces a governance question the existing frameworks answer only partway. The NIST AI Risk Management Framework [3] treats human oversight of autonomous systems as a governance priority, and Article 14 of the EU AI Act [4] establishes a binding human-oversight obligation for high-risk AI systems — though, as discussed below, fraud-detection systems of the kind this article examines are themselves carved out of that high-risk classification, which is precisely what leaves the governance question this article addresses unresolved by binding law. Neither instrument specifies how decision authority should actually be split between an agent and a human reviewer for a given class of financial decision, nor how that split should be owned, documented, and revised as circumstances change. (The federal instrument this framework was originally mapped against, Executive Order 14110, was rescinded on January 20, 2025 and superseded by Executive Order 14179; the shift toward binding, jurisdiction-specific instruments such as the EU AI Act, alongside a still-voluntary domestic posture, is itself an argument for the kind of institution-level governance mechanism this article proposes.) A recent IJISAE article [9] compared AI agents, predefined workflows, and Agentic AI in general terms, proposing hybrid architectures that blend adaptability with deterministic control. That analysis is deliberately domain-general; it does not address payment infrastructure specifically, nor the governance mechanics of allocating authority leaving the practical handoff-design question open for financial systems in particular.

This article makes three contributions. First, a taxonomy of agentic deployment patterns observed in production payment infrastructure, distinguishing reactive, proactive, collaborative, and orchestrator agent roles. Second, an architecture description of a

Independent Researcher, USA

layered real-time fraud-inference system that grounds this taxonomy in a concrete, latency-constrained production design. Third, and most centrally, a four-tier handoff-threshold governance framework specifying when agent authority is full, advisory, conditional, or prohibited mapped explicitly onto the NIST AI RMF Govern function and the human-oversight provisions of Article 14 of the EU AI Act.

Section 2 reviews agentic AI foundations and the governance gap this article targets. Section 3 lays out the deployment-pattern taxonomy. Section 4 develops the fraud-inference architecture description. Section 5 presents the handoff-threshold framework. Section 6 discusses regulatory alignment and adoption tradeoffs. Section 7 concludes.

2. Literature Review and Background

Agentic AI builds on foundational work in autonomous agents and multiagent systems. An agent is conventionally defined as a system that perceives its environment, reasons about it, and acts to achieve specified goals [1]. Multi-agent systems research further identifies canonical coordination patterns, hierarchical orchestration, market-based task allocation, consensus-based validation, and blackboard architectures for coordinating several specialized agents rather than relying on one monolithic system [2].

A recent IJISAE article [9] positions AI agents, predefined workflows, and Agentic AI as three points on a flexibility-versus-predictability spectrum and proposes hybrid architectures combining adaptability with deterministic control. That treatment is domain-general and does not touch payment infrastructure or the governance mechanics of allocating decision authority, the gap this article addresses directly. IJISAE's editorial standing is contested in some indexing discussions, so this article treats it as the closest domain-adjacent treatment available rather than a load-bearing authority; the four-tier framework developed below is instead positioned against the longer-established levels-of-automation literature.

On the governance side, the NIST AI Risk Management Framework organizes AI risk management around four functions Govern, Map, Measure, and Manage [3]. Article 14 of the EU AI Act imposes a binding requirement that high-risk AI systems be designed and operated so that natural persons can effectively oversee them, including the ability to understand system capabilities and limitations, correctly interpret outputs, override or disregard them, and halt the system entirely [4], but that obligation attaches only to systems classified as high-risk under Annex III, and Annex III point 5(b) expressly excludes AI systems used for the purpose of detecting financial fraud from the high-risk creditworthiness category. The fraud-inference system this article examines is therefore not, on the fraud-detection basis alone, a high-risk system to which Article 14 directly applies; this article treats Article 14 as a design reference for what disciplined human oversight looks like, not as an obligation this system is bound to satisfy. The domestic regulatory picture has moved since this framework's initial formulation: Executive Order 14110, which had directed federal attention toward human oversight of consequential AI decisions, was rescinded on January 20, 2025 and replaced by Executive Order 14179, which reoriented federal AI policy toward deregulation and reserved substantive oversight requirements to voluntary frameworks such as NIST AI RMF and to the subsequent, deregulation-focused America's AI Action Plan. The resulting picture is a genuine governance gap rather than an oversight already closed by binding rule: the AI Act's own drafters carved fraud detection out of high-risk treatment, and the domestic instrument that once addressed this ground has been rescinded, leaving fraud-inference systems in production payment infrastructure without a binding, jurisdiction-specific human-oversight requirement from either regime. That gap is precisely what an institution-level, tiered handoff-threshold framework is positioned to fill, using Article 14's oversight vocabulary as a design vocabulary rather than as a compliance floor this article claims to satisfy. Financial-sector-specific work, including the Financial Stability Board's assessment of AI-related systemic vulnerabilities [6] and IOSCO's review of AI use cases and risks in capital markets [7], identifies governance and accountability gaps but stops short of proposing a concrete, tiered handoff mechanism. Closing that gap is this article's central contribution.

3. Agentic Deployment Patterns in Payment Infrastructure

Financial infrastructure deployments of agentic AI fall into four canonical patterns. Reactive agents

respond to defined triggers, a build failure, a fraud alert with learned or specified response strategies. Proactive agents continuously monitor system state and initiate action when they detect conditions requiring intervention, without waiting for an external trigger. Collaborative agents coordinate with other agents or with human operators, sharing state and negotiating action sequences. Orchestrator agents decompose high-level goals into sub-tasks and delegate them to specialized agents, integrating the outputs into a coherent outcome.

Autonomous CI/CD dependency resolution offers a concrete illustration. Payment platforms integrate legacy mainframe components, modern microservices, and compliance-certified libraries with conditional, environment-dependent compatibility constraints that static dependency resolution cannot reliably capture. An agentic pipeline agent represents the dependency graph as a structured state space and applies constraint propagation together with learned heuristics drawn from the accumulated history of prior build outcomes to navigate toward resolutions that satisfy compatibility, security posture, and build-speed objectives at once. Production deployment agents of this kind have, in the author's operating experience, cut environment provisioning time from on the order of 30 minutes to roughly 5 minutes, enabling a shift toward provisioning fresh environments per feature branch rather than relying on shared, drift-prone long-lived environments. *Note: this figure, like the deployment-experience figures reported in Section 4, reflects a practitioner-recollected approximation from operating experience rather than an externally audited benchmark, and is reported with qualifying language accordingly.*

Table 1. Agentic deployment patterns vs. typical payment use case, reversibility, and authority tier.

Deployment pattern	Typical payment-infrastructure use case	Action reversibility	Typical authority tier
Reactive	Respond to build failure or fraud alert with a specified/learned response	High (easily reversed)	Full-autonomy
Proactive	Continuously monitor system state, initiate action before an external trigger	Medium	Advisory-autonomy
Collaborative	Coordinate with other agents or human operators on shared objectives	Medium	Advisory-autonomy
Orchestrator	Decompose and delegate a full deployment sequence across sub-agents	Low (hard to reverse)	Confirm-before-act

These patterns carry distinct reversibility profiles that become directly relevant to the handoff-threshold framework developed in Section 5: reactive dependency-resolution agents typically propose easily reversible actions, while orchestrator agents coordinating a full production deployment sequence may commit actions that are costly or impossible to reverse once executed.

4. Architecture Description: Layered Real-Time Fraud Inference

Real-time fraud scoring is arguably the most latency- and accuracy-constrained application of agentic AI in payment systems. Authorization decisions must land within milliseconds, and decision accuracy must hold to within fractions of a percentage point given the transaction volumes a global payment network processes. A single-model architecture cannot simultaneously deliver the speed the full transaction population requires and the analytical depth needed to catch the most sophisticated fraud patterns.

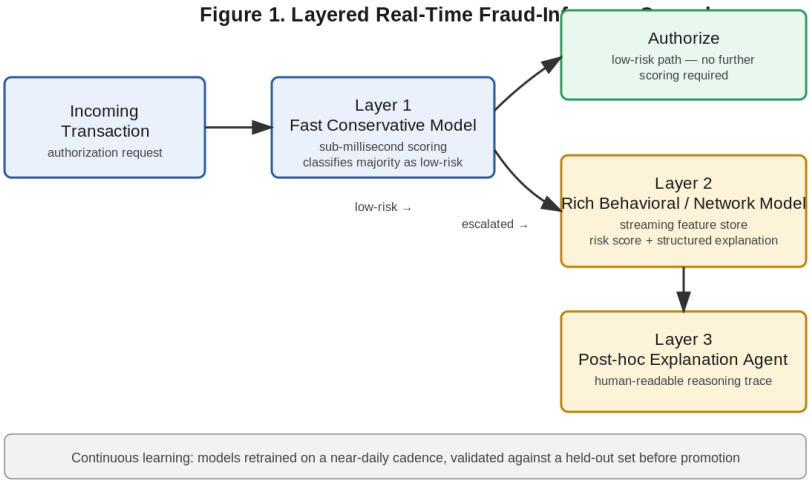

Figure 1. Layered real-time fraud-inference cascade: a fast conservative first layer authorizes the low-risk majority directly; escalated transactions pass through a richer second layer and a post-hoc explanation layer.

The layered inference architecture resolves this tension by decomposing the scoring decision into a cascade of progressively more sophisticated evaluations, shown in Figure 1. A compact first-layer model classifies the majority of transactions as low-risk with high confidence in well under a millisecond, sending them straight to authorization; every other transaction escalates to a second layer that incorporates behavioral and network-derived features, producing both a risk score and a structured explanation that supports regulatory transparency requirements [5]. A third layer generates human-readable reasoning traces for dispute resolution and regulatory reporting.

Note: as with the provisioning-time figure in Section 3, the figures below reflect practitioner-recollected approximations from deployment experience rather than externally audited benchmarks, and are reported with qualifying language accordingly.

Continuous learning further shifts model retraining from a weekly batch cycle to a near-daily cadence, cutting the exposure window during which a novel fraud pattern can operate undetected from approximately seven days to under 24 hours on the order of a seven-fold reduction. Running this cadence at production scale requires agentic infrastructure for model validation, deployment, and drift monitoring: an autonomous deployment agent promotes validated model versions on schedule and triggers human review when validation metrics fail defined thresholds, while a separate monitoring agent continuously checks production performance against real-time feedback.

5. The Handoff-Threshold Governance Framework

The deployment patterns and fraud-inference architecture description above share a common governance need: a principled, auditable basis for deciding when an agent may act independently and when a human has to be involved. We propose four threshold classes that should be defined explicitly for every agentic decision surface in a payment system. Confidence-based thresholds trigger handoff when an agent's calibrated certainty drops below a class-specific floor. Materiality-based thresholds trigger handoff when the financial value or customer impact of a decision exceeds a defined ceiling. Novelty-based thresholds trigger handoff when an input falls outside the distribution the agent was validated on. Regulatory-mandate thresholds require human authorization for any decision class that law or supervisory expectation reserves for human judgment.

These four classes combine into a defense-in-depth handoff posture in which no single failure mode can silently expand an agent's effective authority. We model the resulting authority allocation as a tiered ladder rather than a binary switch. In the full-autonomy tier, the agent acts without per-decision human involvement appropriate for high-frequency, low-materiality, easily reversible decisions. In the advisory-autonomy tier, the agent acts but every action is logged for mandatory post-hoc human

review within a defined window. In the confirm-before-act tier, the agent prepares a recommended action but a human must authorize it before execution. In the human-only tier, the agent is prohibited from acting and serves purely as decision support, reserved for decisions carrying systemic or irreversible consequences. This four-tier ladder is structurally related to the levels-of-automation literature, including Sheridan and Verplank's ten-level scale [10], Parasuraman, Sheridan, and Wickens's typology of automation types and levels [11], and SAE J3016's driving-automation levels [12]; the contribution here is not the ladder structure itself but its composition with the four trigger classes above (confidence, materiality, novelty, and regulatory mandate) as a defense-in-depth mechanism, together with the change-control and ownership apparatus described below, mapped specifically onto payment-infrastructure decision classes.

Table 2. Authority tiers, example payment decision classes, and review requirements.

Authority tier	Example payment decision class	Review requirement
Full-autonomy	Transient transaction hold pending additional signal	Logged only, asynchronous review
Advisory-autonomy	Model version promotion to a production traffic segment	Mandatory post-hoc human review within a defined window
Confirm-before-act	Large settlement adjustment	Human authorization required before execution
Human-only	Legacy system decommissioning; systemic/irreversible actions	Agent restricted to decision support only

Treating handoff thresholds as governance artifacts, rather than an engineering configuration detail, imposes accountability requirements that ordinary escalation tuning does not: each threshold needs a named owner, a documented rationale linking the chosen value to a risk appetite statement, and a change-control process requiring joint engineering, risk, and compliance sign-off before any relaxation. Threshold changes particularly relaxations that expand agent authority should be versioned, timestamped, and surfaced in periodic governance reporting, so the cumulative drift of an agent's autonomy over time stays visible to senior leadership rather than accreting silently through a sequence of locally reasonable adjustments.

6. Discussion: Regulatory Alignment and Organizational Adoption

The four threshold classes map directly onto existing governance expectations, summarized in Table 3.

Table 3. Threshold classes mapped to NIST AI RMF functions and EU AI Act Article 14 oversight provisions.

Threshold class	NIST AI RMF function	EU AI Act Art. 14 provision
Confidence-based	Measure	Art. 14(4)(a): overseer must be able to monitor operation and detect anomalies; Art. 14(4)(c): overseer must correctly interpret system output before consequential decisions finalize
Materiality-based	Govern	Art. 14(3): oversight measures commensurate with the risk and impact of the autonomous action
Novelty-based	Map	Art. 14(4)(a): overseer understanding of system limitations outside validated conditions; Art. 14(4)(b): overseer awareness of automation bias
Regulatory-mandate	Govern	Art. 14(4)(d)-(e): overseer authority to override, disregard, or halt system operation entirely

Confidence- and novelty-based thresholds operationalize the NIST AI RMF's Map and Measure functions, translating abstract risk-identification guidance into concrete, per-decision-class triggers. Materiality- and regulatory-mandate thresholds operationalize the Govern function and draw on the specific overseer capabilities Article 14 of the EU AI Act's design vocabulary describes — understanding system limitations, interpreting output, and retaining the authority to override or halt it — giving institutions a concrete, auditable mechanism for demonstrating that meaningful human control is preserved as agentic systems scale, independent of whether Article 14 applies to a given deployment as a binding legal obligation.

Some adoption friction is a deliberate feature of this design, not a defect to be engineered away. The cross-functional sign-off required before any threshold relaxation is meant to be costly, precisely because the alternative thresholds adjusted unilaterally by engineering teams optimizing for throughput is the failure mode the framework exists to prevent. Institutions should expect the framework to slow the rate at which agent autonomy expands relative to an ungoverned baseline; that friction is the mechanism through which the framework delivers its governance value. This posture is consistent with the accountability gaps identified in recent oversight reviews of AI use in financial services [8].

7. Conclusion

Agentic AI is moving from research concept to production reality in payment infrastructure, but existing governance frameworks do not yet specify concrete mechanics for splitting decision authority between agent and human. This article presented a taxonomy of agentic deployment patterns, a layered fraud-inference architecture description grounding that taxonomy in production practice, and a four-tier handoff-threshold framework mapped explicitly onto the NIST AI RMF and the human-oversight provisions of Article 14 of the EU AI Act. The framework is portable beyond payments to any regulated, high-stakes domain where agentic systems are taking on greater operational authority.

Ethics Statement

Conflict of Interest: The author declares no conflict of interest.

Human/Animal Rights: This article does not involve human or animal subjects.

Informed Consent: Not applicable.

Acknowledgments

The author acknowledges the engineering and compliance colleagues whose collaborative work informs the practitioner frameworks described in this article. The views presented are the author's own.

References

[1] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 4th ed., Pearson, 2020.

[2] M. Wooldridge, An Introduction to MultiAgent Systems, 2nd ed., John Wiley & Sons, 2009.

[3] National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), U.S. Department of Commerce, 2023.

[4] European Parliament and Council of the European Union, Regulation (EU) 2024/1689 (Artificial Intelligence Act), Article 14: Human Oversight, OJ L, 2024/1689, 12.7.2024.

[5] U. Bhatt, A. Xiang, S. Sharma, et al., "Explainable machine learning in deployment," Proc. 2020 ACM Conf. on Fairness, Accountability, and Transparency (FAccT), pp. 648-657, 2020.

[6] Financial Stability Board, The Financial Stability Implications of Artificial Intelligence, 2024.

[7] International Organization of Securities Commissions (IOSCO), Artificial Intelligence in Capital Markets: Use Cases, Risks and Challenges, 2025.

[8] U.S. Government Accountability Office, Artificial Intelligence: Use and Oversight in Financial Services, GAO-25-107197, 2025.

[9] C. C. Nallapareddy, "Strategic Design with AI Agents, Predefined Workflows and Agentic AI," International Journal of Intelligent Systems and Applications in Engineering, vol. 13, no. 1, pp. 445-450, 2025.

[10] T. B. Sheridan and W. L. Verplank, Human and Computer Control of Undersea Teleoperators, Man-Machine Systems Laboratory, MIT, 1978.

[11] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human

interaction with automation," IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, vol. 30, no. 3, pp. 286-297, 2000.

[12] SAE International, Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles, J3016, 2021.