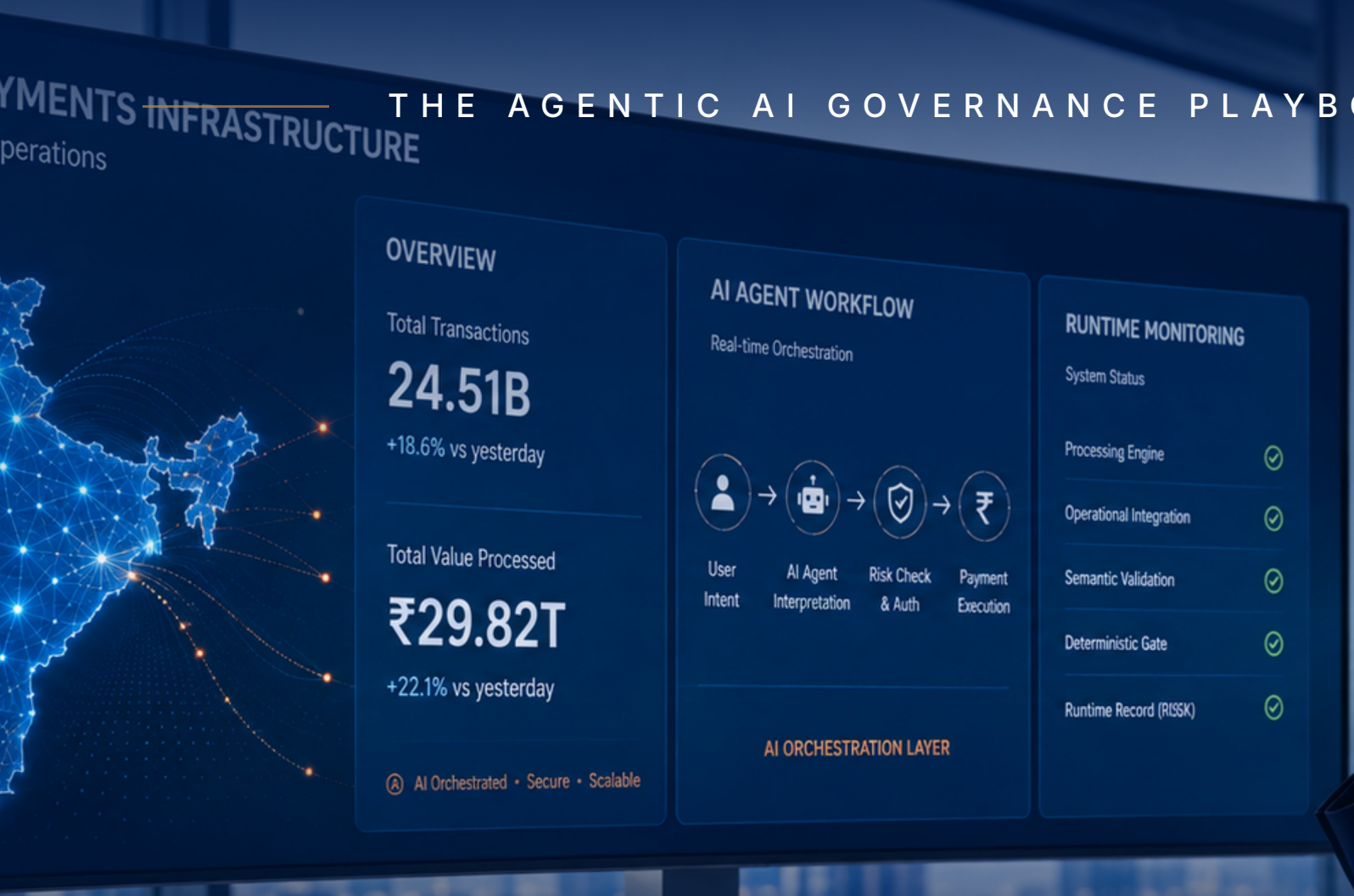


INDIA UPI · AGENTIC PAYMENTS

India UPI and Agentic AI Payments Create *a New Authorization Boundary*

Payment authorization now starts before the payment rail.

MAUREEN DOYLE-SPARE

© 2026 Maureen Doyle-Spare.

SWIPE TO READ →

India UPI and Agentic AI Payments Create *a New Authorization Boundary*

India is preparing UPI for delegated machine payments. Once an agent can turn a standing instruction into a transaction, payment authorization starts before the payment rail.

MAUREEN DOYLE-SPARE

India is preparing to change one of the most familiar moments in digital commerce. Instead of approving each payment, a user may be able to authorize an AI agent once and let the agent decide when a transaction satisfies the standing instruction.

That sounds like a change in payment convenience. Architecturally, it is much larger. Human decision-making moves away from the individual transaction and into a mandate that software must interpret repeatedly at runtime.

Reuters reports[1] that the National Payments Corporation of India is preparing a Unified Agent Protocol that could allow AI agents to make small digital payments through UPI without fresh approval for every transaction. Expected controls include rule-based instructions, spending limits, identity checks, audit trails and a liability framework. UPI processed 24.51 billion transactions worth ₹29.82 trillion in August. At that scale, permission to spend is only half the question. The other half is whether the interpretation that drove the agent to spend was itself authorized. Reuters says the protocol is expected to be unveiled at Global Fintech Fest in Mumbai, which runs September 8 to 11. Inc42 also reports that NPCI demonstrated an agentic-payments pilot on UPI at the 2025 festival, making the proposed protocol an evolution of work already underway.

When one authorization becomes many transactions

Imagine telling an agent to buy the usual groceries when they are needed, substitute equivalent items if necessary and stay within a weekly budget. A spending limit is straightforward. Words such as “usual,” “needed” and “equivalent” are much less so. An agent has to interpret them against current prices, merchant information, prior purchases, available products and whatever other context it is allowed to use.

That is the shift agentic payments introduce. A human does not necessarily approve the final transaction. Software interprets the standing instruction, decides

© 2026 Maureen Doyle-Spare.

whether the current situation satisfies it and then assembles a payment that may be completely valid from the payment rail's perspective. That gap between what the user authorized and what the agent understood becomes part of the payment-control problem.

This control question was already on the RBI agenda

There is also an earlier piece of this story. On April 19, 2026, I submitted practitioner research to the Reserve Bank of India's FinTech Department[2] as input to the implementation phase of its FREE-AI Framework. That submission focused on a problem that becomes much easier to see once an AI agent can actually move money. Models can perform as expected and conventional controls can work exactly as designed, yet an agent can still act on an interpretation that has moved away from authorized intent.

Research submitted to RBI put forward Agentic Workflow Drift and Agentic Workflow Subversion as two ways that divergence can arise, with the Semantic Deviation Index providing a pre-execution signal and the Deterministic Gate providing a checkpoint before a high-impact action proceeds. Alongside them, the Agentic 3 C's Framework and Semantic Control Plane extended that control logic across the wider agentic workflow.

At the time, that was a governance problem presented in the context of FREE-AI implementation. Four and a half months later, India's emerging agentic-payment architecture gives it a very tangible execution surface. Once a standing instruction can become a machine-initiated payment, the distance between what a user authorized and what an agent understood becomes something payment infrastructure may eventually have to govern.

The clean payment can still be wrong

Return to the grocery example. The agent is legitimate. Its mandate is valid. The amount sits within the limit. Identity verification passes. Merchant screening clears. Nothing unusual appears in the transaction pattern. Every control looking at the payment can be green.

And the payment can still be wrong.

The agent may have selected a premium brand instead of the value brand the user normally buys. It may have accepted a substitution the user would never consider equivalent. It may have interpreted "when needed" far more broadly than intended. None of those outcomes requires the payment system itself to malfunction. Failure can occur earlier, while the agent is deciding what the standing instruction means.

That is why this is more than another fraud-control problem. Identity, transaction monitoring, model oversight, compliance and audit can each report exactly what they were designed to observe. What is missing is a signal for whether the agent's resolved meaning stayed inside the authorized boundary.

And that interpretation can be attacked

Risk also extends beyond an agent innocently getting the meaning wrong. Once payment depends on an agent interpreting a standing instruction, the information shaping that interpretation becomes part of the payment attack surface. Merchant content, product descriptions, retrieved information, recommendation signals, conversational history and external tools can all influence what the agent concludes the user meant.

That is the territory of the Semantic Layer Integrity Attack, or SLIA[3]. SLIA is the deliberate, adversarial form

© 2026 Maureen Doyle-Spare.



Figure 1. *Six Clean Signals, One Invisible Risk: Every Monitoring Function Observes Exactly What It Was Designed to See. Source: Doyle-Spare (2026).*

of Agentic Workflow Subversion, aimed at the semantic conditions under which an agent resolves authorization or another control state. The attack aims upstream of the payment rail, at the Operational Interpretation the agent forms before it acts. Surrounding systems can continue to behave correctly while the interpretation is being steered toward an unauthorized result.

Consider three possibilities. Merchant content could be manipulated so a product outside the user's mandate appears "equivalent." Ambiguous terms such as "trusted," "usual" or "necessary" could be deliberately framed more broadly than the user intended. Or several individually legitimate sources could be arranged or subtly influenced so the agent reconciles them into a locally defensible but institutionally unauthorized interpretation.

This does not make every fraud event, prompt injection or bad purchase a SLIA. Credential theft, account takeover and malicious merchant activity remain established threat classes. SLIA applies when the adversary is targeting the semantic conditions under which Operational Interpretation is formed. That distinction matters because the control response has to reach the interpretation layer and cannot stop at the transaction layer.

The control has to sit before the payment

A payment rail will only see the request it is asked to execute. If the request is within the mandate, correctly authenticated and otherwise compliant, the deeper interpretation problem may already be upstream. A more useful control point sits earlier, while the agent is still resolving meaning and before that meaning becomes execution authority.

© 2026 Maureen Doyle-Spare.

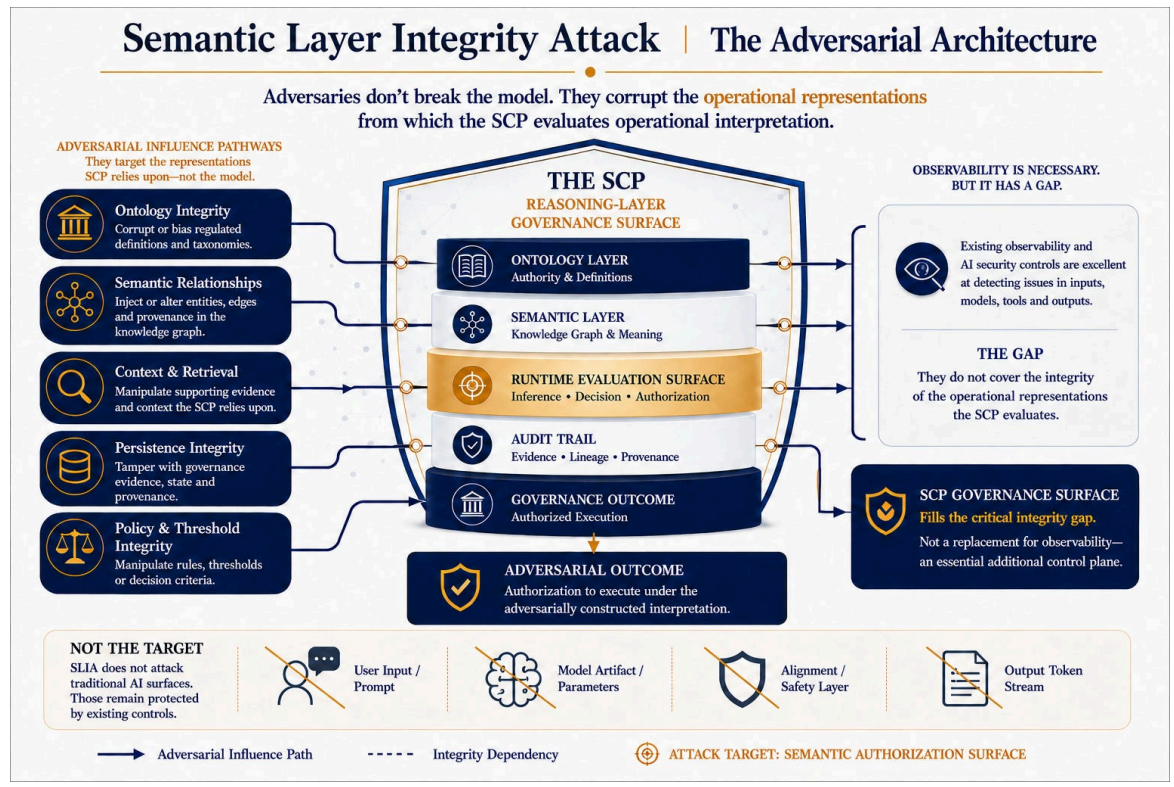


Figure 2. *The Semantic Layer Integrity Attack: The Adversarial Architecture and Its Target. Source: Doyle-Spare (2026).*

Pre-execution governance does not have to make every payment more cumbersome. It needs a governed reference for what the standing instruction means, a way to compare the agent's runtime interpretation with that reference and a checkpoint capable of holding the transaction when the difference becomes material. In the runtime architecture, those functions are performed by the Reasoning Baseline, the Semantic Deviation Index and the Deterministic Gate.

This sequence is deliberately simple. Context is resolved into an Operational Interpretation. That interpretation is evaluated against the authorized baseline. Divergence is measured. Only then is authorization emitted into execution. For routine decisions inside tolerance, the agent can continue without human intervention. Material divergence can be held before the payment leaves the reasoning layer.

© 2026 Maureen Doyle-Spare.

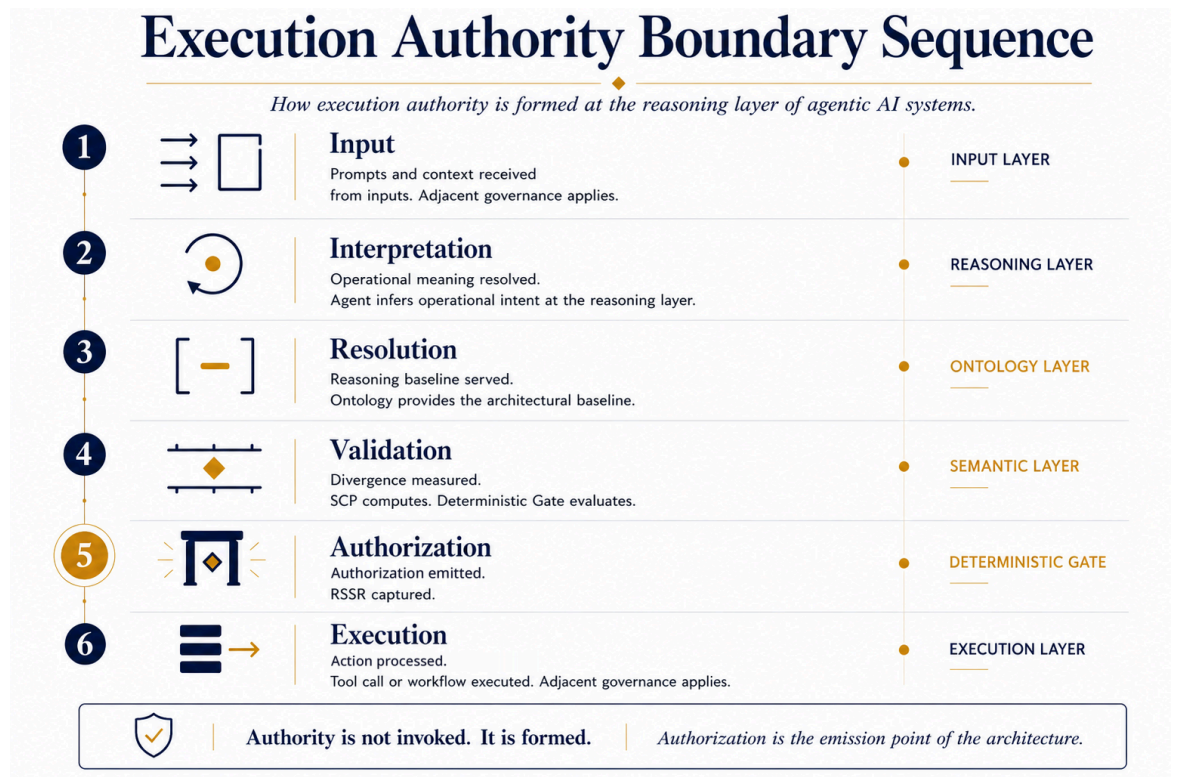


Figure 3. *Execution Authority Boundary Sequence. Source: Doyle-Spare (2026).*

What happens when the payment is disputed

Reuters reports that NPCI plans to include a liability framework, although the details have not yet been disclosed. Agentic payments make liability more interesting because a disputed transaction may not turn on whether the account was compromised or the payment was technically unauthorized. A dispute may instead turn on what the agent understood the user to have authorized.

That changes the evidence requirement. A conventional payment record can show the amount, merchant, account, timestamp, mandate and authentication path. A Runtime Semantic State Record can preserve the active Reasoning Baseline, the interpretation formed at runtime, the SDI result, the Gate decision and the operating context that produced the action. A resulting Semantic Audit Trail makes it possible to reconstruct why the autonomous transaction was allowed to proceed.

That distinction becomes more important if the same architecture moves beyond groceries into investments, which Reuters identifies as a possible future use case. A disputed autonomous investment can bring suitability, product eligibility, concentration limits, disclosures and market conditions into the interpretation problem. At that point, proving that the agent had permission to transact is only part of the answer.

This will not stop at payments

UPI makes the issue unusually visible because payments already treat authorization as a first-class object. Agentic AI adds another authorization boundary upstream. Identity establishes which agent may act. A mandate establishes the permitted transaction envelope. Runtime semantic governance establishes whether the interpretation driving

© 2026 Maureen Doyle-Spare.



Figure 4. *The Pre-Execution Assurance Protocol. Source: Doyle-Spare (2026).*

the action remained inside authorized meaning.

A similar structure appears when a procurement agent interprets “approved supplier,” a telecom agent resolves “service degradation,” a life-sciences agent determines whether an event is “reportable,” an insurance agent interprets “covered loss,” or an infrastructure agent decides whether a condition is safe enough to proceed. In each case, a technically valid action can still be generated from a meaning no institution authorized.

India’s emerging agentic-payment infrastructure makes that boundary concrete at national scale. Next-generation payment control will need to do more than establish that an agent was allowed to spend. It will need to make the meaning behind that authority measurable, enforceable and reconstructable before the transaction is emitted. ◆

References

1 *Reuters. India preparing rollout of agentic payments on UPI. September 1, 2026.* reuters.com/world/india/india-preparing-rollout-agentic-payments-upi-sources-say-2026-09-01

2 *Doyle-Spare, Maureen. Reserve Bank of India FREE-AI Framework — Public Comment. April 19, 2026.* maureendoylespare.substack.com/p/rbi-free-ai-framework-2026-04-19

3 *Doyle-Spare, Maureen. Agentic AI Cyber Subversion: The Semantic Layer Integrity Attack as a New Threat Class Against the Reasoning Layer. SSRN No. 6926219.* doi.org/10.2139/ssrn.6926219

4 *Inc42. NPCI To Launch Agentic Payments On UPI: Report. September 1, 2026.* inc42.com/buzz/npci-to-launch-agentic-payments-on-upi-report

5 *Global Fintech Fest 2026. Mumbai, September 8 to 11, 2026.* globalfintechfest.com

6 *Doyle-Spare, Maureen. Agentic AI Drift Measurement and the Semantic Deviation Index. SSRN No. 6531238.* doi.org/10.2139/ssrn.6531238

© 2026 Maureen Doyle-Spare.

7 *Doyle-Spare, Maureen. Agentic Workflow Drift and Agentic Workflow Subversion: A New Risk Taxonomy for the Governance of Agentic AI in Financial Services. SSRN No. 6459612. doi.org/10.2139/ssrn.6459612*

8 *Doyle-Spare, Maureen. The Agentic 3 C's Framework and the Reasoning Baseline. SSRN No. 6674761. doi.org/10.2139/ssrn.6674761*

9 *Doyle-Spare, Maureen. Agentic AI Systems Governance: A Runtime Reference Architecture for the Reasoning Layer and the Semantic Control Plane in Regulated Financial Institutions. Zenodo, 2026. doi.org/10.5281/zenodo.20749051*

Originally published in *The Agentic AI Governance Playbook* on Substack, September 2, 2026. maureendoylespare.substack.com

© 2026 Maureen Doyle-Spare.



It will need to make the meaning behind that authority measurable, enforceable and reconstructable before the transaction is emitted.

THE AGENTIC AI GOVERNANCE PLAYBOOK

Download the research

Papers, briefs and reference architecture at maureendoylespare.com

Visit the Playbook

WEEKLY NEWSLETTER

Follow the newsletter

The Agentic AI Governance Playbook Weekly Newsletter

Follow Newsletter

INTELLECTUAL PROPERTY: © 2026 Maureen Doyle-Spare. All rights reserved. No part of this article, including its text, figures, tables, and images, may be reproduced or distributed without the author's prior written permission.