

Cyber Security for Payment Ecosystem in the Age of Agentic Commerce

SISA

Dr. Anjan Krishnamurthy

Associate Director & Head, AI Security & Governance | sisa.ai

Principal Architect, CSPAI · Ph.D. in CS&E · 75+ publications · 21 patents filed

20 YEARS OF SECURING PAYMENT DATA

PCI Security Standards Council

SISA

Connect on LinkedIn

Agenda

What We Will Cover Today

01	**The Agentic Commerce Shift**	Agents that buy, not browse — ACP, UCP, AP2, MCP, A2A
02	**What the Protocols Protect**	Signed requests, tokens, mandates — and the open gap
03	**Six Challenges for Checkout**	Replay, constraint breach, injection, drift, leakage
04	**Observability as Evidence**	Five boundaries, an allow-listed event model, evidence
05	**Detect, Contain, Roll Out**	Detection rules, safe containment, a six-step rollout

SISA

20 YEARS OF SECURING PAYMENT DATA PCI Security Standards Council

From Browsing Agents to Buying Agents

SECTION ONE



SISA

What Is Agentic Commerce?

When the Agent Holds the Cart, the Consent and the Credential

15–25%

of all e-commerce volume projected to be handled by AI agents by 2030 (Bain & Company)

3×

year-on-year growth in AI-driven traffic to Shopify stores, Q2 2026 (Shopify)

FOUR SURFACES ALREADY IN PRODUCTION

Assistant-Native Checkout
ChatGPT Instant Checkout on ACP — single-item purchases since Sep 2025

Cross-Platform Carts
UCP: Google with Shopify, Etsy, Target, Walmart and Wayfair — Jan 2026

Delegated Payment Mandates
AP2 signed mandates; protocol donated to the FIDO Alliance in Apr 2026

Tool and Agent Rails
MCP bindings into merchant checkout; A2A for agent-to-agent coordination

SISA

The Agentic Commerce Protocol Stack

Five Open Standards — Each Owns a Different Layer

LAYER 01

ACP · Agentic Commerce Protocol

OpenAI + Stripe. Checkout sessions and delegated payment tokens; powers ChatGPT Instant Checkout. MCP binding defined.

LAYER 02

UCP · Universal Commerce Protocol

Google + retailers. Discovery, checkout and post-purchase over REST / JSON-RPC; interoperates with MCP, A2A and AP2.

LAYER 03

AP2 · Agent Payments Protocol

Payment authorization. Signed, constraint-bound Intent, Cart and Payment mandates — proof the user consented.

LAYER 04

MCP · Model Context Protocol

Tool connector between agent and merchant APIs. Not commerce-specific, but the path most checkout calls take.

LAYER 05

A2A · Agent2Agent

Coordination between agents across shopping, payment and fulfilment workflows. Supporting layer, not a checkout.

LAYER 06

The Constant: Merchant & PSP

Remain authoritative for price, tax, acceptance, fulfilment, signature verification and key custody (HSM).

SISA

20 YEARS OF SECURING PAYMENT DATA PCI Security Standards Council

Six Challenges Checkout Was Never Built For

SECTION TWO



SISA

What the Protocols Protect — and Don't

Signing Proves Authorization, Not Secrecy or Behaviour

ACP — **Merchant-Authoritative Checkout**

Authenticated, signed checkout calls; the merchant owns price, tax, acceptance and fulfilment. Payment data travels as a PSP token or encrypted payload over HTTPS.

UCP — **Separated Payment Phases**

Negotiation, credential acquisition and completion are separated. The platform must not handle raw financial credentials; the business must never echo them.

AP2 — **Signed, Bound Mandates**

SD-JWT mandates bind to a checkout hash, agent key and expiry; selective disclosure limits exposed claims. Signed is not encrypted — the holder can read it.

GAP — **Nobody Watches the Agent**

Each verifier sees only its own transaction. No protocol layer sees drift, injection, a budget split across merchants, or credential leakage across sessions.

SISA

20 YEARS OF SECURING PAYMENT DATA

Six Challenges Checkout Was Never Built For

Fraud That Arrives as Clean, Fast, Successful Payments

CHALLENGE 01

Mandate Replay and Expiry

The same mandate fingerprint or nonce reappears across transactions; a closed mandate is minted after the open mandate expired.

CHALLENGE 02

Constraint Breach

The final cart exceeds the amount range, budget, payee list, category or shipping destination the user actually approved.

CHALLENGE 03

Merchant-Content Injection

Hidden instructions on a product page rewrite agent memory — e.g. a gift card silently appended to the cart (threat research, Mar 2026).

CHALLENGE 04

Agent Drift

A new merchant, payment handler, MCP server, tool or model is invoked outside the established baseline — unnoticed by any verifier.

CHALLENGE 05

Credential and Token Leakage

Payment tokens or credential-like data surface in prompts, traces, tool logs or responses — quietly expanding PCI DSS scope.

CHALLENGE 06

Settlement Anomalies

Charged amount differs from the approved amount; duplicate capture; rapid re-authorization; odd refund or cancel sequences.

SISA

20 YEARS OF SECURING PAYMENT DATA
PCI Security Standards Council

What the Field Is Already Reporting

Signals From 2026 Threat Research and Fraud Analysis

1

Prompt injection sits at the centre of agentic risk

OWASP's 2026 State of Agentic AI Security maps prompt injection to six of the ten categories in the Top 10 for Agentic Applications, and catalogues CVEs and breach reports for nearly every category.

2

Retail shopping agents are already being targeted

Published threat research (Mar 2026) demonstrated indirect injection against a UCP shopping agent: a hostile discount page rewrites agent memory and appends a $100 gift card to the attacker's address at checkout.

3

Agentic fraud does not look like fraud

2026 fraud analysis describes agent-mediated fraud as clean, fast, successful transactions — the profile legacy anomaly engines approve. Bain projects agents handling 15–25% of e-commerce volume by 2030.

SISA

20 YEARS OF SECURING PAYMENT DATA PCI Security Standards Council

Observability as the Evidence Layer

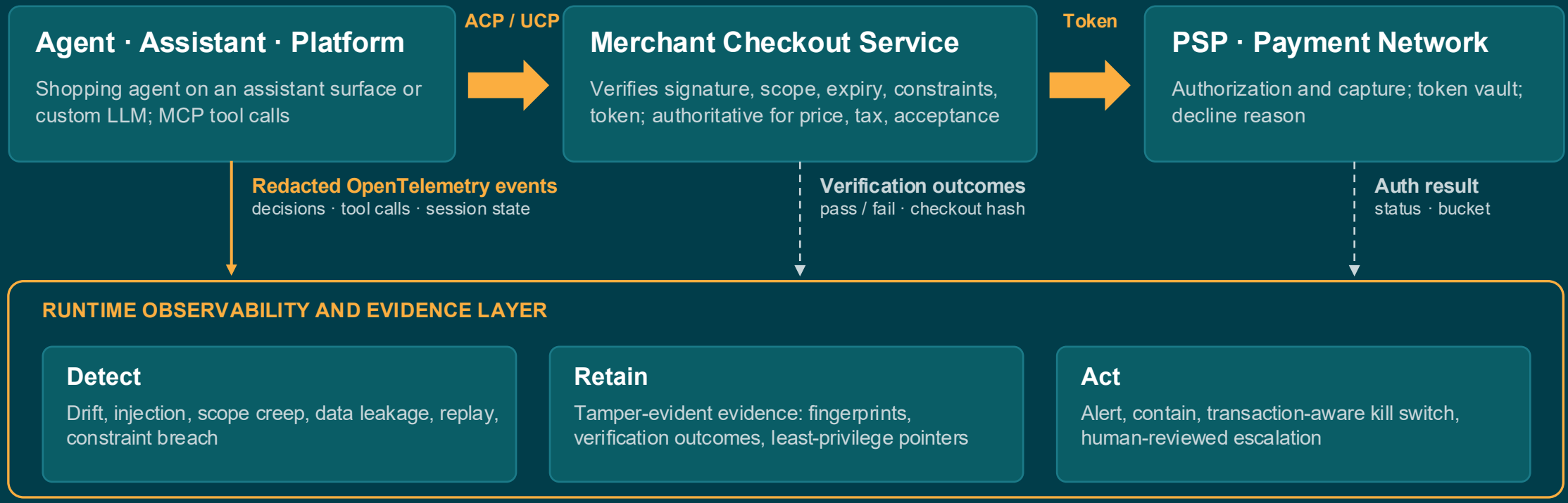
SECTION THREE



SISA

Observe Events, Not Payments

Evidence Around Checkout — Never in the Payment Path



Not a substitute for the merchant's PSP, signature verifier, HSM or authorization service — an overlay on the agent, the MCP/tool layer and the merchant integration.

SISA

20 YEARS OF SECURING PAYMENT DATA PCI Security Standards Council

Five Boundaries to Instrument

Where the Evidence Layer Attaches and What It Detects

01

Agent Decision

Instruction → discovery → cart proposal → approval or autonomous policy. Detect prompt injection, hijacked merchant content, scope creep, unexpected tool or MCP use.

02

Checkout

ACP create / update / complete / cancel and webhooks; UCP negotiation, handler selection, credential acquisition. Detect price, quantity or address changes, retry storms, session takeover.

03

Mandate Verification

Record signature, issuer, audience, expiry, nonce / key binding, checkout hash and every constraint result. Detect replay, out-of-constraint carts, unknown agent keys.

04

Payment

Token lifecycle and authorization result only: provider, token fingerprint, amount bucket, status, decline class. Detect cross-merchant token reuse and over-limit charges.

05

Post-Purchase

Correlate order, refund, cancellation, fulfilment and dispute events. Preserve tamper-evident evidence references — never the authorization credentials themselves.

SISA

20 YEARS OF SECURING PAYMENT DATA PCI Security Standards Council

What Is Sent — and What Is Never Sent

Allow-Listed, Redacted Events Contain PCI DSS Scope

RECEIVES · ONE REDACTED EVENT

```
event: mandate.verification
commerce_protocol: acp | ucp | ap2
checkout_ref_hash: sha256:…
mandate_fingerprint: sha256:…
agent_id: shopping-agent-v3
autonomy_mode: human_present | autonomous
amount_bucket: 100–250 USD
signature_valid: true
constraint_result · expiry_result: pass · pass
idempotency_result: new | replayed
payment_result: authorized | declined
risk_score: 18
```

NEVER RECEIVES

- Raw PAN, CVC, bank credentials, PSP API keys or decrypted payment payloads
- Raw payment tokens, full JWT / SD-JWT mandates, private keys or user credentials
- Full addresses, email, phone, unredacted prompts or complete tool arguments by default
- Any data that expands PCI DSS scope without a deliberate contractual and technical review

Correlate on stable, keyed fingerprints. Original mandates and signed receipts stay in a restricted merchant evidence vault; the evidence layer holds the fingerprint, the verification outcome and a least-privilege pointer.

SISA

20 YEARS OF SECURING PAYMENT DATA

Eight Detections to Implement First

Rules Across Mandate, Agent and Settlement Behaviour

01 — Mandate replay
Same mandate fingerprint or nonce appears across transactions

02 — Constraint breach
Total, merchant, category, destination or quantity beyond the open mandate's allowance

03 — Authority failure
Signature, key binding, issuer, audience or expiry check fails

04 — Approval-to-charge mismatch
Final charged amount differs from the authoritative checkout or mandate amount

05 — Agent drift
Newly invoked merchant, payment handler, MCP server, tool, model or capability outside baseline

06 — Prompt injection
Merchant or user content attempts to override spend controls, tool permissions or approvals

07 — Sensitive-data leakage
Credential-like data appears in LLM prompts, traces, tool logs or outbound responses

08 — Settlement anomaly
Duplicate capture, rapidly repeated authorization attempts, unexpected refund / cancel sequence

All eight evaluate fingerprints, outcomes and buckets — none requires raw payment data in the evidence layer.

SISA

20 YEARS OF SECURING PAYMENT DATA PCI Security Standards Council

Containment Must Be Transaction-Aware

A Generic Kill Switch Cannot Simply Stop Payments

01

Before Token or Checkout Completion

Automatically suspend the agent session and invalidate the checkout. No money has moved — the cheapest moment to act.

02

After Token, Before Capture

Block agent retries. Route to merchant and PSP review before any capture proceeds.

03

After Capture

No uncontrolled rollback. Freeze further agent actions, preserve evidence, and route to refund and dispute operations under human control.

04

Credential or Secret Exposure

Isolate the affected agent or service, revoke and rotate secrets, invalidate sessions and tokens where supported, and initiate incident response.

SISA

20 YEARS OF SECURING PAYMENT DATA

A Six-Step Rollout

Metadata First, Enforcement Second, Autonomy Last

01	**Instrument one ACP merchant**	Metadata only, around checkout endpoints and order webhooks
02	**Stream redacted traces via OpenTelemetry**	Baseline merchants, tools, providers, latency and conversion
03	**Enforce four foundational controls**	Signature verification, idempotency, token redaction, correlation
04	**Add policy-driven detections**	The eight rules, with human-reviewed containment
05	**Extend to UCP and AP2 flows**	Mandate-chain proof for high-value and autonomous purchases
06	**Run adversarial tests**	Tampering, replay, MCP substitution, injection, token leakage

SISA

20 YEARS OF SECURING PAYMENT DATA PCI Security Standards Council

PROTOTYPE · IN DEVELOPMENT

Proof Point: An AP2 Mandate Drift Monitor

On Google's AP2 Reference Agent — Test Objects Only

HOW IT WORKS

- Zero-code SDK hooks inside the shopping agent: MandateClient.present() and ClientSession.call_tool()
- Decodes the SD-JWT chain, keeps an allow-list, SHA-256s the join keys, emits one OpenTelemetry span per payment
- Never emitted: raw SD-JWT, disclosures, keys, signatures, risk data, instrument ids, line items, PII
- Every hook swallows its own errors — telemetry can never break the agent

EIGHT DRIFT RULES · EACH PAYMENT VS ITS OWN MANDATE

`over_amount_range`	high	`near_budget (80%)`	low
`under_amount_range`	medium	`payee_not_allowed`	high
`currency_mismatch`	high	`mandate_expired`	high
`over_budget`	high	`instrument_changed`	medium

STATUS **SDK hooks and OTLP/HTTP ingest — ready.** Parser whitelist, running totals, rules service and dashboard — to build. Optional blocking via LLM-gateway callback — under evaluation.

SISA

20 YEARS OF SECURING PAYMENT DATA PCI Security Standards Council

How SISA PRISM Helps

Observe the Agent. Prove the Mandate. Keep the Payment Path Untouched.

prism.sisa.ai

www.sisa.ai · prism.marketing@sisainfosec.com

CONFIRM BEFORE DEPLOYMENT

Ingestion and OpenTelemetry collector configuration; field-level redaction before persistence; data residency; retention and deletion; tenant isolation; encryption and key management; incident-data access; and whether any configuration makes the platform a PCI DSS service provider.

PRISM PLATFORM CAPABILITIES

PrismObserve

OpenTelemetry capture via SDK, endpoint or gateway; drift detection, decision tracing, containment

PrismDiscover

Inventory of agents, MCP servers, tools and models in the commerce stack — including shadow AI

PrismStrike

Adversarial testing: checkout tampering, replay, MCP substitution, content injection, token leakage

CSPAI Certification

ANAB-accredited AI security credential for the teams who run these controls

SISA

Key Takeaways

What to Act On Now

1

Protocols prove authorization — they do not watch behaviour

ACP, UCP and AP2 verify a transaction. None of them sees drift, injection or a budget split across sessions and merchants. That is the layer you add.

2

Observe events and outcomes — never payments

Fingerprints, verification results and amount buckets deliver audit-grade evidence without PAN, tokens or mandates ever entering the evidence layer.

3

Containment has to know where the money is

Suspend before the token, review before capture, preserve after capture. Humans own refunds and disputes; the kill switch owns the agent.

SISA

20 YEARS OF SECURING PAYMENT DATA PCI Security Standards Council

THANK YOU

Questions?
Find me at a Networking Break.

Dr. Anjan Krishnamurthy
Associate Director & Head, AI Security & Governance | sisa.ai

Secure AI. Secure Agents. Secure Payments.

Connect on LinkedIn

SISA