

AI Agents in PCI: From Hype to Release-Ready

A PCI-aligned approach to evaluating AI agents, with an MPoC A&M closed-loop response example

Neal Hindocha
Chief Security Officer, Mypinpad Ltd

Neal Hindocha

Chief Security Officer, Mypinpad Ltd



Why “Agentic” AI Changes the PCI Conversation

From chat to action

- Classic automation is deterministic — inputs map to predictable outputs
- AI agents interpret ambiguous inputs, synthesize actions, and — if you let them — act across systems

What you now must prove

- Scope decisions are defensible — what the AI touches, and what it can affect
- Controls still operate as intended — least privilege, change control, logging, monitoring, testing

> *The core question isn’t “is the model accurate?” — it’s:* **can we constrain the AI’s data access, system reach, and autonomy so PCI DSS controls remain effective and auditable?**

AI Use Case vs. AI Agent

- The definition that anchors every scoping conversation

AI use case

- A bounded workflow where AI provides output
- Summaries, classifications, drafts, suggestions
- No ability to change system state

AI agent

- AI output + tool access (APIs, scripts, consoles)
- Can cause state change across systems
- Privilege and autonomy enter the picture

PCI risk rises sharply the moment you introduce tool access and write capability.

The PCI DSS Lens: Load-Bearing Requirement Families

All 12 families can apply — these six do the heavy lifting for AI agents

Req 1 / 2
Network controls & secure configurations — segmentation, egress control, hardened runtime

Req 6
Secure SDLC & change control — model, prompt, and tool changes are "changes"

Req 7 / 8
Least privilege & strong identity — agents get their own identities; no shared automation

Req 10
Logging & monitoring — who / what / when / how for every agent action

Req 11
Testing — agent interfaces and privileges become test targets

Req 12
Governance — acceptable use, AI inventory, third-party management (12.8), risk analysis

The key shift: AI isn't "just a tool" — it becomes part of the control environment, engineered to be testable and evidenced.

Agent Admission Control

A simple evaluation method that maps to PCI reality



AAC: Three Axes, One Clear Decision

Evaluate every proposed AI use case before release

- Score the use case along the three axes on the right
- Produce one boundary decision: **allowed**, **allowed with guardrails**, or **not allowed**
- Repeatable, documented, and defensible in assessor conversations
- Guards the “lethal trifecta”: no single system combines sensitive data, untrusted input, and external access

1 **Data exposure**
What the AI can see

2 **Connectivity & privilege**
What the AI can reach — and with what permissions

3 **Autonomy**
What the AI can do without a human

20 YEARS OF SECURING PAYMENT DATA
PCI Security Standards Council

Boundary Bands: What's Releasable, What Isn't

Typical profiles along the three AAC axes

GREEN

Generally releasable

Non-CHD/SAD data + no CDE reach + advisory only

- Drafting policies, runbooks, compliance docs
- Read-only analysis of sanitized telemetry outside the CDE
- Developer assistance outside production, with normal code review

YELLOW

Only with strict guardrails

Sanitized telemetry and/or limited connectivity and/or constrained actions

- Alert triage for in-scope systems — with redaction + access controls
- Recommending changes that require approval
- Pre-approved actions via an allowlisted control plane

RED

Not allowed

CHD/SAD exposure, broad CDE connectivity, or broad autonomy

- Any prompting / RAG / training involving CHD or SAD
- Agents holding high-impact secrets — keys, key material, passwords, production credentials
- Self-directed agents with arbitrary admin actions

20 YEARS OF SECURING PAYMENT DATA
PCI Security Standards Council

The Single Best Boundary: AI Plane Out-of-Band

The CDE stays isolated — the AI plane sees sanitized artifacts only



*This is the path to adopting AI without turning every experiment into a **scope explosion**.*

20 YEARS OF SECURING PAYMENT DATA PCI Security Standards Council

The Cloud Question, Head-On

Cloud AI can work — the burden is what you can prove

Most defensible when…

- It never receives CHD/SAD, secrets, or anything that reconstructs them
- It has no inbound reach to the CDE — ideally none outbound from the AI plane either
- It is treated as a TPSP (Req 12.8): written agreement, no training on your data, breach notification, monitored evidence

A heavier lift when "AI is in the path"…

- Privileged access demands stronger segmentation and identity controls
- Stronger evidence and change controls
- Tighter contractual and technical assurances from the provider

The question isn't "cloud: yes or no?" — it's ***"what can you prove?"***

20 YEARS OF SECURING PAYMENT DATA
PCI Security Standards Council

Never Give an Agent Raw Admin Surfaces

Treat every agent as a potential malicious insider — speed without an un-auditable actor

No direct access to…

- Consoles and shells
- IAM administration
- Firewall rule editing
- CI/CD production deploy paths
- High-impact secrets and keys

Instead: a deterministic control plane

- Small number of allowlisted actions
- Input validation and safe defaults
- Explicit approvals where needed
- Mandatory logging of every action
- Functional (non-AI) DLP on egress paths
- Rollback — and a tested kill switch

20 YEARS OF SECURING PAYMENT DATA PCI Security Standards Council

Two Artifacts You Can Operationalize Tomorrow

Built for engineering — designed to survive assessor scrutiny

Agent Qualification Card (AQC)

One per agent or AI use case — your AI inventory, entry by entry

- Purpose, named accountable owner, scope statement
- Data allowed — and explicit “never” data
- Connectivity, unique identity, least privilege, tool allowlist
- Autonomy level: advisory / approval-required / constrained
- Logging schema, change & revalidation triggers, kill switch

AI Release Checklist

The one-page go / no-go gate

- ✓ Scope and data flows mapped end-to-end
- ✓ No CHD/SAD in prompts, RAG, training, or logs
- ✓ Segmentation, egress control, least privilege validated
- ✓ Allowlisted control plane + evidence-grade logging
- ✓ Change control, testing, kill switch tested, third-party duties documented

20 YEARS OF SECURING PAYMENT DATA
PCI Security Standards Council

Vulnerability Management for In-House Code

Live and in daily use since mid-2026 — one platform from repo to remediation

Connect a repo

A read-only link is all it needs — the SBOM is generated automatically

Know what's inside

Components and versions mapped; open-source licences checked

Map your feeds

Subscribe advisory feeds (e.g. OSV, vendor bulletins) and map them to systems

Review by exception

Anything worth reviewing is flagged — with analysis and suggested fixes or mitigations

Built for the programme owner — one view of the estate: SBOMs, licences, advisories, verdicts, and evidence

- Onboarding is **minutes, not a project** — no agents, no build changes, no code modifications
- Flags arrive triaged: verdict, confidence, and reasoning — your reviewers start at the interesting ones
- Full analysis on demand — **suggested fixes or mitigations**, grounded in what the code actually uses

Worked Example: AAC on Our Own Tooling

Vulnerability triage — advisory by default, constrained action by exception

1 · Data exposure

Curated artifacts only: advisories, system metadata, SBOM versions, repo facts. Never CHD/SAD or secrets

2 · Connectivity

Outbound-only HTTPS, single-turn prompt in, text out. No credentials, no tools, no path back in

3 · Autonomy

Advisory verdicts by default; a human decision always outranks. Auto-triage only via allowlisted statuses

AAC decision

Feed assessment: **GREEN**. Auto-triage: **YELLOW**, guardrails engaged — exactly the bands this method predicts

Evidence Recorder — immutable log of every AI action: actor “ai”, model, reasoning, confidence, before/after — exportable for assessors, gaps surfaced not hidden

- **Deterministic, not prompt-based** — guardrails live in reviewed code: an enum, a confidence gate, an audit row
- **Constrained autonomy** — four allowlisted statuses: never on low confidence, never over a human, off by default
- **Grounded verdicts** — repo profiling gives evidence of absence; human answers stored once and reused
- **Bounded blast radius** — provider-agnostic client, minimum severity per feed, hard cap on assessments per sync
- **Kill switch** — deactivate the provider and assessment stops; advisories still import, AI is never load-bearing

Adopting This Without Boiling the Ocean

A seven-step implementation roadmap

1. Pick one use case — start in the **Green zone** (evidence summarization, alert triage, documentation support)
2. Create one **Agent Qualification Card** and run it through the AI Release Checklist
3. Build the boundary: **segmentation + egress control + redaction/sanitization pipeline**
4. Establish the **control plane** — allowlisted actions only, no raw admin surfaces
5. Make logging **non-optional**: every agent run produces structured evidence
6. Define **re-approval triggers**: model, prompt, tool-permission, and scope changes
7. Run a **"bad day" exercise**: prompt injection, privilege escalation, logging failure, kill-switch invocation

20 YEARS OF SECURING PAYMENT DATA
PCI Security Standards Council

Approve It. Control It. Audit It.

PCI environments don’t need to ban AI — they need to stop treating it as a casual add-on

APPROVE	Evaluate every use case on data exposure, connectivity & privilege, and autonomy
CONTROL	Enforce deterministic action boundaries via an allowlisted control plane — accountability stays human-owned
AUDIT	Demand evidence-ready logging and disciplined change control — engineered in from day one

Thank you

Neal Hindocha · Chief Security Officer · Mypinpad Ltd
nealh@mypinpad.com

20 YEARS OF SECURING PAYMENT DATA