PanGuard Sovereign Capability Brief

FOR NATIONAL INSTITUTIONS & SOVEREIGN AI TEAMS · REV 0.5

Sovereign control over autonomous AI systems. Enforced at every agent action; cryptographic proof of every decision; verifiable offline, without trust in the supplier.

As nations bring AI capability in-house, the governing question is no longer which model to procure, but how autonomous agents may act on sovereign data — and how that control is demonstrated to a regulator, an auditor, or an allied state. PanGuard Sovereign is a governance and assurance layer that enforces national policy at every agent action and produces cryptographic evidence of every decision. It operates in front of any model or agent framework, on infrastructure the institution controls — built on an open, vendor-neutral standard, so sovereignty does not mean trading one foreign dependency for another. Any sovereign nation can deploy it.

STATUS Reference architecture · design-partner stage **MATURITY** Laboratory-validated prototype (≈ TRL 4) **ACCREDITATION** None claimed

---

THE SOVEREIGN AI STACK

Sovereign AI has three pillars. Two are being built worldwide. This is the third.

PILLAR 01

Sovereign compute

Whose hardware runs the inference — chips and data centres under national control.

BEING BUILT WORLDWIDE

PILLAR 02

Sovereign models

Whose weights — models trained and hosted inside the jurisdiction.

BEING BUILT WORLDWIDE

PILLAR 03 · WHERE PANGUARD SITS

Sovereign agent control & proof

National policy enforced at every agent action, with evidence any auditor verifies offline. The pillar that decides whether autonomous AI can be trusted with sovereign work.

UNDER-SERVED TODAY

> **Platform sovereignty asks an institution to trust a foreign vendor. PanGuard Sovereign is control the institution can verify — which is what sovereignty means.**

1. The problem

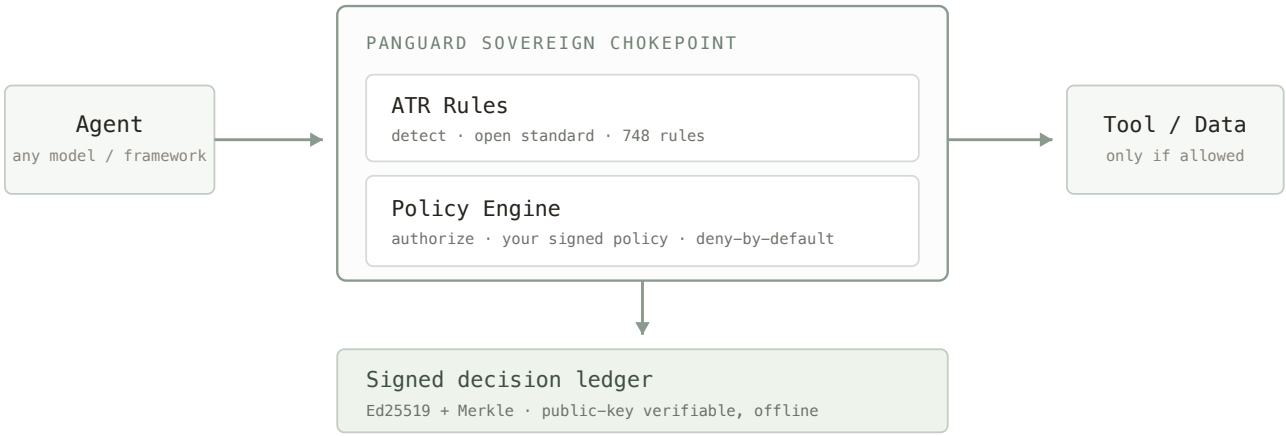
Agentic systems are entering critical infrastructure and decision flows worldwide — calling tools, moving data, and chaining decisions on their own. For any sovereign institution, three gaps open that choosing a model, or a cloud, cannot close.

- **The risk is the action, not the model.** Prompt injection, tool poisoning, and over-autonomy can turn or over-reach an agent — it exfiltrates, escalates, or destroys in milliseconds, with no human in the loop. A safer model does not govern what the agent is permitted to do.
- **National policy never reaches the decision.** Policy exists on paper, but nothing enforces it at the point each agent acts — and supply-chain and advanced-model risks enter the decision unseen. Governing on any external hyperscaler puts the keys, the policy, and the audit trail beyond your reach in a crisis.
- **Black-box decisions you cannot audit.** High-risk systems must produce demonstrable proof of what agents did and were allowed to do — for a regulator, an allied state, or an EU AI Act / NIST AI RMF obligation. “Trust us” is not a control.

THE REGULATORY CLOCK · 2026	
JAN	Korea’s AI Basic Law enters force — the second comprehensive national AI law after the EU.
JUN	The EU defines four sovereign-cloud tiers (CADA) — sovereignty grading becomes institutional.
AUG	EU AI Act high-risk obligations become enforceable — logging and human-oversight duties bind.
NOW	National regulators flag autonomous agents as an unmanaged risk (Dutch DPA; French ANSSI).

2. How it works

Every tool call an agent makes passes a single decision point — detection against the open Agent Threat Rules standard, then authorization against your own signed national policy — and produces evidence anyone can verify offline. Deny-by-default: the absence of a decision is a denial.



Every agent action passes one governed chokepoint

3. Why PanGuard Sovereign

Three properties a hyperscaler cannot give a sovereign buyer:

- **Sovereign.** Runs on-premise or air-gapped, on infrastructure the institution owns. No outbound dependency, no phone-home, no foreign cloud in the trust path. The signing keys never leave your control.
- **Verifiable.** Every decision is public-key signed; an auditor verifies it offline holding only a public key — without trusting the operator, the log, or us. Not “trust us”: prove it.
- **Open standard.** Built on the open Agent Threat Rules standard — vendor-neutral, not tied to any single foreign hyperscaler — and already shipping in production. Your evidence outlives any single supplier.

4. Adoption — the open standard is already in production

PanGuard Sovereign is built on Agent Threat Rules (748 rules) — the open detection standard adopted, through merged public contributions, by leading security vendors and standards bodies. Every entry is independently verifiable at **agentthreatrule.org/ecosystem**.

In production	Microsoft (Agent Governance Toolkit) · Cisco (AI Defense) · Gen Digital (Norton / Avast)
Standards bodies	MISP / CIRCL · OWASP · FINOS (Linux Foundation) · SigmaHQ
Also shipping	AMD (GAIA) · Microsoft PyRIT · AG2 (AutoGen) · rulezet / CIRCL
In review	NVIDIA (garak · NeMo Guardrails) · OpenAI Guardrails · Splunk · OpenTelemetry GenAI SIG

5. Sovereign control domains

An enforceable, auditable control for each of the four dimensions nations use to define AI sovereignty:

- **Data** — classification that rides the data (high-water-mark); jurisdiction and residency enforced at every boundary; policy-gated decryption; secret-egress prevention on outbound calls.
- **Compute** — executes on hardware the institution controls, fully offline; air-gap delivery with verify-before-install; keys released only to attested endpoints.
- **Model** — model-agnostic; operates in front of any agent or LLM; an optional on-premise adjudicator may escalate for review, never authorize (tighten-only).
- **Governance** — signed policy-as-code, deny-by-default, anti-rollback; signed decision ledger with transparency log; human approval, dual-control break-glass, kill-switch.

6. Assurance — evidence that does not depend on trusting the supplier

Every security-critical action is Ed25519-signed at the point of action, and any party holding only a public key can verify it offline. Signed, not asserted · third-party provable · fail-closed by default · open and inspectable.

In practice this is the evidence a **national-security audit**, a **budget or procurement review**, and an **EU AI Act Art. 12 / 26 high-risk logging obligation**, and a **NIST AI RMF audit trail** require — produced continuously at the point of each decision, not reconstructed after an incident.

7. Compliance crosswalk — by jurisdiction

Deployment profile and framework mapping adapt to the jurisdiction. A country-specific pack can be produced for any jurisdiction on request. Design-intent mapping — not a certification or accreditation.

JURISDICTION	SIGNING	HSM	AIR-GAP	RETENTION
Taiwan	Ed25519	Not required	Tier-A critical	12 months
Japan	Ed25519	Not required	Not required	12 months
Singapore	Ed25519	CII may require FIPS 140-3	CII isolated	12 months
United Arab Emirates	Ed25519	Not required	Not required	12 months
European Union	ECDSA P-256	Classified triggers HSM	Not required	≥ 6 months (AI Act Art. 26)
South Korea	ECDSA P-256	**Required (KCMVP)**	C-tier / MLS	5 years (high-impact AI)
Saudi Arabia	ECDSA P-256	**Required (NCS-1 ADVANCED)**	Not required	12 months

Example framework references — **Taiwan**: 資通安全管理法 · 資通系統防護基準 · 人工智慧基本法. **Japan**: NISC 統一基準 · デジタル庁 DS-920 v2.0 · FISC 安全対策基準 第14版 · AI事業者ガイドライン. **South Korea**: AI 기본법 · KCMVP · 국가정보보안기본지침. **Saudi Arabia**: NCA ECC-2:2024 · NCA NCS-1:2020 · SDAIA Generative AI Guidelines. **EU**: EU AI Act Art. 12 / 14 / 26 · NIS2 · CRA · BSI C5:2026 · ANSSI SecNumCloud.

8. Capability register

A working prototype under continuous independent adversarial review (~1,700 automated tests; 30 review passes to date).

Implemented (built + validated under adversarial review, ≈ TRL 4): identity & least-privilege · policy-as-code · signed decision ledger · data taint & egress control · human approval & break-glass · supply-chain & A2A auth · kill-switch · policy-gated key release · air-gap bundle · operator console · agent-facing HTTP fleet.

Integration seam (interface defined + tested against a reference verifier; integration with the institution's own hardware/models pending): confidential-compute attestation · on-premise adjudicator · HSM / PKCS#11 signing.

9. Engagement path

A staged, low-commitment path from evaluation to a deployment you operate:

- **Technical evaluation** — async, under NDA where required. Reference architecture, an in-country compliance crosswalk, an honest maturity statement, and a demonstration you run yourself. *Commitment: none.*
- **Reference deployment** — a single deployment in a controlled environment you choose, delivered air-gapped and verified before install. *Commitment: scoped pilot.*
- **Production path** — scope the requirements only an external party can satisfy (certification via an accredited path; a validated hardware module where a jurisdiction mandates one). *Commitment: framework agreement.*

10. Collaboration model

PanGuard is the technology supplier, not the prime. An in-country partner (system integrator or national laboratory) holds the local relationship and the contract vehicle; the institution keeps sovereign control.

- **PanGuard** provides the sovereign runtime, the signing/verification stack, and the open ATR standard.
- **In-country partner** provides the contract vehicle, local delivery, language, first-line support, and the government relationship.
- **National institution** keeps sovereign control — the signing keys, the policy, and the audit trail. Nothing leaves its control.

The institution procures through its in-country partner — the same channel governments already use for complex systems. Because the standard and the verifier are open, the institution's evidence stays verifiable regardless of any single supplier's future.

11. Delivery — into an air-gapped environment, verifiable at every step

Package → transfer (approved media, nothing phones home) → verify offline with only the public key (fail-closed) → human-signed acceptance (bound to the bundle identity) → deploy (config + Helm from your country pack) → operate (loopback operator console: approvals, break-glass, kill-switch, signed ledger).

12. Next steps

- **Schedule a technical briefing** — async, under NDA where required; no commitment.
- **Design-partner pilot** — a reference deployment in a controlled environment you choose.
- **Sovereign PoC** — prove it air-gapped on your own infrastructure.

Contact: adam@agentthreatrule.org

A. Appendix — set against the adjacent categories

Sovereign compute and models answer whose hardware and whose weights. Enterprise agent-security platforms secure corporate agents inside a vendor's cloud. Sovereign agent platforms build and run agents on-premises. Each answers a real question — and each leaves the same one open: governing what an autonomous agent may do with sovereign data, and proving it to an auditor afterwards.

CATEGORY	WHAT IT ANSWERS	WHAT IT DOES NOT
Sovereign compute & models	Whose hardware and whose weights run the AI	Does not govern what an agent may do with sovereign data, or prove it afterwards
Enterprise agent security	Policy and posture for corporate agents, operated as vendor SaaS	Not air-gapped, not per-nation, and verification requires trusting the vendor
Sovereign agent platforms	Building and running agents on-premises	The platform is the party being audited — it cannot also be the evidence
PanGuard Sovereign	Enforces national policy at every agent action and emits evidence any auditor verifies offline with a public key — on an open, vendor-neutral standard	Not a model, not a cloud, not an agent-building platform

Occupying a distinct position is not the same as leading a mature market. This is an early, verifiable implementation of that third pillar, at design-partner stage — offered for evaluation, not as a finished category.