

Sovereignty at the Forward Pass

Reconceiving AI Sovereignty as Inference Control

Shubhradeep Nandi

shub@responsibleai.com

Co-Founder and Chief AI Scientist

ResponCibleAI

Regional Track: Asia-Pacific (South Asia / India) | Sub-track: AI Governance, Data Protection & Sovereignty

Abstract. *As frontier AI capability concentrates in a few models accessed through commercial APIs, the prevailing infrastructure-centric notion of "AI sovereignty" owning compute, data, and models misidentifies what a state actually controls. We argue that sovereignty's operative variable is inference control: who runs the forward pass, who observes its inputs and outputs, who sets in-deployment behavior, and who can verify, pin, or replace the model. We decompose inference control into six dimensions and show that the technical primitives needed to make inference sovereign (attested model identity, confidential execution, anti-substitution proofs, sovereign fallback) are the same primitives needed to verify AI behavior in deployment: sovereignty and safety converge at the forward pass. Re-analyzing India's sovereign-AI program, we find its most jurisdictionally decisive act migrating the BHASHINI, a public-service system fully onto domestic infrastructure is best understood as repatriating the inference trust boundary, not owning hardware.*

1. Introduction

The dominant picture of "AI sovereignty" is an infrastructure picture: a nation is sovereign over AI to the degree that it owns the compute, holds the data, and can build the models. That picture identifies the means and misses the object. As frontier capability concentrates in a handful of models reached through an API, the variable that actually determines a state's autonomy and, separately, its safety is who controls the properties of the model at inference time: who runs the forward pass, who can see its inputs and outputs, who sets its in-deployment behavior, who can change or withdraw it, and who can prove any of this to a third party. We call this inference control. Why this matters for AI safety, regionally and globally: as more consequential public functions triage, tax processing, benefits delivery, multilingual citizen services get routed through a small number of shared models, four safety properties degrade together. A single latent flaw propagates across every dependent jurisdiction at once (inference monoculture); deployment-time verification becomes something only the vendor can perform (privatized verification); a model can change underneath a critical deployment with no

observable signal (silent mutation); and the power to throttle or withdraw access becomes the power to disable a public function (availability coercion). For a country like India, whose citizen data and public-service AI increasingly pass through foreign-operated inference, a data-protection statute can assert jurisdiction on paper while leaving these exposures untouched at the forward pass.

This submission makes four contributions:

- A decomposition of inference control into six dimensions, showing that API-only access is not weak control but the structural absence of control.
- A convergence thesis: the technical primitives that make inference sovereign are the same primitives that make AI behavior verifiable in deployment so verification-based sovereignty is a distributed implementation of AI safety, not a threat to it.
- A practical agenda: a five-layer inference-control stack matched to deployment criticality, with an explicit theory of change for middle powers.
- A re-analysis of India's sovereign-AI program against the framework, yielding the finding that its decisive act was repatriating the inference trust boundary, plus three falsifiable predictions.

What is new in this work? The six-dimension control surface, the convergence thesis, the five-layer practice stack, the structured re-analysis of India (2023-2026), and the three falsifiable predictions are the new synthesis produced for this hackathon. We build on and explicitly do not claim to originate the underlying literatures on data-protection law, confidential computing, and model-substitution auditing surveyed below.

2. Related Work

Sovereign-AI and compute nationalism: National programs increasingly frame sovereignty as domestic compute and indigenous models e.g., India's IndiaAI Mission and accompanying accelerator partnerships[5,6]. This literature centers infrastructure ownership and largely treats inference as a downstream consequence of owning the stack.

Data-protection and localization law: India's Digital Personal Data Protection Act, 2023 [2] and the DPDP Rules, 2025 [1] establish an extraterritorial, negative-list cross-border regime; analyses note the resulting regulatory discretion and uncertainty [3]. A recurring tension is foreign legal reach over data held by domestically incorporated providers, exemplified by the US CLOUD Act [4].

Confidential computing and attested inference: Trusted Execution Environments with remote attestation now run model inference at near-native speed while keeping inputs, outputs, and weights opaque to the operator [10,11,12,13,14]; vendors offer attested, API-compatible confidential inference [13,14], and frontier labs have published on confidential inference via trusted virtual machines [8].

Model identity, substitution, and property attestation: Recent work audits whether API users receive the model they paid for [7] and develops property attestation and guardrail proofs for large generative models [9,15]

The Gap: These literatures advance in parallel but are rarely connected. None frames inference control as the shared object of both AI sovereignty and AI safety, nor maps the sovereignty toolkit onto deployment-time safety functions. That connection is our contribution.

3. Methodology

This is a conceptual and single-case analytical contribution rather than an implemented system; we are explicit about that scope (Section 6). The approach has four replicable steps.

(i) Control-surface decomposition.

For each property that determines a model's deployment behavior, we ask who sets it under (a) API-only access versus (b) possession of weights. The properties whose controller flips between the two regimes define the dimensions of inference control. This yields the six-dimension surface in Section 4.1.

(ii) Convergence mapping.

We map each technical primitive a state would need to assert inference control onto the deployment-time function it performs for AI safety, and test whether they coincide (Section 4.2).

(iii) Structured case re-analysis.

We code India's documented sovereign-AI actions from 2023 to mid-2026 against the six dimensions, using primary government notifications and public reporting [1,2,3,5,6] (the BHASHINI migration [6], the DPDP Rules notification [1], and IndiaAI Mission disclosures [5]). For each action we record which dimension it primarily addresses, then identify the most jurisdictionally decisive action.

(iv) Hypothesis formulation.

From the framework we derive falsifiable predictions and specify a test for each (Section 4.5). All coding inputs are public records, so the analysis is reproducible from the cited sources.

4. Results

4.1 The control surface: six dimensions

Decomposition yields six dimensions. Under API-only access, the controller of every dimension is the provider which is why we describe API dependence as the absence of inference control rather than a weak form of it.

SL No	Dimension	Question it answers	Controller under API-only
1	Forward-pass locus	Where, and under whose law, does compute run?	Provider
2	Weight access	Do you hold the weights at all?	Provider only
3	Policy layer	Refusals, tool access, logging, priors	Provider
4	Observability / exposure	Who sees the inputs and outputs? Provider	Provider
5	Availability / revocability	Can it be throttled or withdrawn?	Provider
6	Mutability / identity	Which version ran; can it change silently?	Provider

4.2 Convergence thesis

Each of the four safety degradations named in Section 1 is dissolved by the same capabilities a state needs for inference control. The sovereignty toolkit and the deployment-safety toolkit are, primitive for primitive, the same toolkit:

Sovereignty / inference-control primitive	Deployment-time AI-safety function it performs
Attested model identity	Detect substitution and regression; reproducible evaluation
Confidential execution (TEE + attestation)	Prevent input/output leakage; make data jurisdiction enforceable
Independent audit and evaluation rights	Public-interest verification on the local distribution
Version pinning and weight escrow	Change-control, rollback, and operational continuity
Sovereign substitution capability	Containment; resistance to availability coercion

The implication reframes the field's central worry. Naive sovereignty (build a datacenter, mandate storage localization) and global safety can indeed conflict. But verification-based inference control and global safety are the same project at two altitudes: a world of many independent verifiers has defense-in-depth against any single lab's single mistake, where a world in which a few labs are the sole verifiers of their own models does not.

4.3 Practice: a five-layer inference-control stack

Inference control is assembled as a portfolio across five layers, ordered from least to most demanding, each closing specific dimensions. A state matches layers to the criticality of each deployment rather than pursuing all layers everywhere.

Layer	Mechanism	Dimensions closed	Maturity (2026)
0	Confidential inference (TEE + remote attestation)	4 (exposure)	Deployable; ~95-99% native speed
1	Verifiable model identity / anti-substitution proofs	6 (mutation)	Emerging research
2	Independent audit	2, 3 (verification,	Legal-technical gap

	and evaluation rights	policy)	
3	Version pinning and weight escrow	5, 6 (revocability)	Contractual; rare
4	Sovereign substitution capability (open-weight)	1-6 in extremis	Partial; models lag frontier

Theory of change. A state adopting this frame stops counting data centers and starts assembling a layered portfolio: confidential inference and version pinning for sensitive-but-routine workloads; full audit rights and a sovereign fallback for safety-critical ones; ordinary API dependence for the rest. This is achievable for a middle power in a way that frontier-model parity is not and every step also adds an independent party able to detect and contain model failure, improving global safety.

4.4 Re-analysis: India 2023-2026

Indicator	Value (public record)
DPDP Act enacted	2023 [2]
DPDP Rules notified	14 Nov 2025; ~18-month phased compliance [1]
Cross-border model	Negative list; Sec. 16 / Rule 15 [1,3]
IndiaAI Mission budget	approx. INR 10,372 crore (~USD 1.25 billion) [5]
National GPU pool	mid-30,000s; target 100,000 by end-2026 [5]
Indigenous open-weight models	Sarvam-30B, Sarvam-105B (Apache 2.0, Feb 2026) [6]
Decisive sovereignty act	BHASHINI fully on domestic cloud/GPU (Feb 2026) [6]

Coding these actions against the six dimensions gives a clear pattern. India's de jure jurisdiction is strong: the DPDP Act applies extraterritorially to foreign entities processing residents' data, and the negative-list regime lets the government restrict transfers by order [1,2,3]. Its de facto control is weak wherever inference is foreign-operated, for three reasons that all live at the forward pass: at inference the plaintext prompt enters the provider's trust domain, reachable by

that jurisdiction's legal process [4]; the state cannot observe what the model did with the data; and storage localization does not relocate the inference trust boundary.

The decisive move, read through the framework, is not the compute pool or the statute. It is the February 2026 migration of BHASHINI India's population-scale multilingual public-service AI fully onto domestic cloud and GPU infrastructure, with the stated purpose that datasets, models, and citizen interactions remain within India's jurisdiction [6]. That single act closes dimensions 1 and 4 by pulling the forward pass inside a controlled boundary. Infrastructure was the instrument; inference control was the objective. India enacted the reframe before naming it the framework's central empirical claim.

4.5 Falsifiable predictions

P1 (convergence): Sovereign-AI procurement requirements and deployed safety-assurance pipelines will request substantially overlapping primitives (attestation, confidential execution, anti-substitution proofs).

Test: compare requirement sets across both communities.

P2 (localization insufficiency): Deployments that localize storage but run inference on foreign-operated endpoints will retain only nominal jurisdiction.

Test: assess exposure to foreign legal process before vs after the inference boundary is repatriated; BHASHINI is a positive corrective instance.

P3 (monoculture): Correlated cross-jurisdiction incidents will cluster on shared models and version updates.

Test: stratify incident data by underlying model and check for clustering around shared releases.

5. Discussion

Implications. The reframe redirects scarce public effort from a race most states will lose (build the best model) toward a project most can win and that safety needs anyway (verify, contain, and if necessary replace the models they depend on). It also dissolves the assumed opposition between sovereign AI and global safety: built on verification primitives rather than mere localization, sovereign inference is a distributed safety architecture.

The operative question for policymakers shifts from "who owns the datacenter?" to "who controls the forward pass?" because both autonomy and safety now turn on that.

What we would do with more time.

(a) Extend the coding to multiple jurisdictions (EU, Gulf states, ASEAN) to test generality.

(b) Draft a reference attested-inference audit protocol and measure TEE overhead on representative public-service workloads.

(c) Build a methodology to measure inference monoculture; the field currently lacks even basic instrumentation for how concentrated the world's consequential inference is.

(d) Define a "minimum-viable-continuity" benchmark specifying how capable and how cheaply held a sovereign fallback model must be to credibly deter availability coercion, so states can buy the right amount of resilience instead of overspending on frontier parity.

6. Limitations and Dual-Use Considerations

Limitations of the approach:

- a) The contribution is primarily conceptual; we build no system and report no new empirical measurements. The three predictions are stated but untested here.
- b) The empirical component is a single-case re-analysis (India), which establishes existence, not generality.
- c) Layer 0/1 depends on Trusted Execution Environments, which shift rather than remove the trust root to the hardware vendor, and on attestation chains that a relying party must be able to interpret; supply-chain and side-channel caveats remain.
- d) Sovereign fallback (Layer 4) presumes indigenous open-weight models good enough to substitute; today these lag the frontier on hard reasoning, so fallback may degrade capability in a crisis.
- e) The audit-rights and escrow layers (2-3) are largely legal-technical aspiration; no widely adopted standard yet specifies them.

Potential misuse / dual-use risks.

The inference-control toolkit is symmetric. The same capabilities that let a state protect citizens from opaque foreign processing also let a state make its own AI unaccountable. Confidential, attested, repatriated inference can shield state surveillance or censorship models from any external audit, encode opaque refusal and content-control policies that outsiders cannot inspect, and wall domestic systems off from international safety scrutiny recasting "sovereign inference" as a tool of entrenchment rather than protection. At the system level, treating sovereignty as achievable could accelerate fragmentation of the AI ecosystem (a "splinternet" of incompatible regimes) and push safety-critical deployments onto weaker local models or behind localization mandates that reduce access to the safest available frontier systems, a net safety loss. The framework's own logic (more independent verifiers improves safety) does not hold if those verifiers face only inward.

Suggestions for future improvement and mitigation.

- a) Couple sovereign-inference rights with binding obligations for external auditability and interoperability, so repatriation does not equal opacity.
- b) Prefer attestation designs that support third-party verification, not only operator self-attestation, to limit the surveillance dual-use.
- c) Empirically validate P1-P3 and extend the case coding to multiple regimes.
- d) Develop the minimum-viable-continuity benchmark and a reference audit protocol so the immature legal layers (2-3) gain concrete, contestable specifications.

References

- [1] Ministry of Electronics and Information Technology (MeitY), Government of India. Digital Personal Data Protection Rules, 2025. Notified 14 Nov 2025. Press Information Bureau.
- [2] Government of India. Digital Personal Data Protection Act, 2023.
- [3] Information Technology and Innovation Foundation (ITIF). India's Cross-Border Data Transfer Regulation. 2025.
- [4] U.S. Congress. Clarifying Lawful Overseas Use of Data (CLOUD) Act, 2018.
- [5] IndiaAI Mission, MeitY. National compute capacity and Foundation Model pillar disclosures, 2024-2026 (incl. DD News reporting on GPU pool expansion).
- [6] Sarvam AI / Digital India BHASHINI Division. Sarvam-30B and Sarvam-105B open-weight models (Apache 2.0), and the migration of BHASHINI onto domestic sovereign cloud and GPU infrastructure, February 2026.
- [7] Cai, W., Shi, T., Zhao, X., and Song, D. Are You Getting What You Pay For? Auditing Model Substitution in LLM APIs. arXiv:2504.04715, 2025.
- [8] Anthropic. Confidential Inference via Trusted Virtual Machines. Technical report, 2025.
- [9] Chantasantitam, P., Caulfield, A., Duddu, V., Gunn, L. J., and Asokan, N. PAL*M: Property Attestation for Large Generative Models. arXiv preprint, 2026.
- [10] Enhancing AI Inference Security with Confidential Computing: A Path to Private Data Inference with Proprietary LLMs. Red Hat Emerging Technologies, 2025.
- [11] Confidential LLMs and Trusted Execution Environments for AI inference. Phala Network, 2025.
- [12] Your Inference Request Will Become a Black Box: Confidential Inference for Cloud-based Large Language Models. arXiv:2603.00196, 2026.
- [13] When Agents Handle Secrets: A Survey of Confidential Computing for Agentic AI. arXiv:2605.03213, 2026.
- [14] Edgeless Systems. Privatemode AI: end-to-end encrypted, attestation-verifiable confidential inference. 2025-2026.
- [15] Proof-of-Guardrail in AI Agents and What (Not) to Trust from It. arXiv:2603.05786, 2026.
- [16] Confidential Computing Consortium. State of confidential computing and attestation standards, 2025.
- [17] Fortanix. Confidential AI for secure enterprise inference (TEE-based model and data protection). 2026.