ARCHITECTURAL ALIGNMENT

Constitutional Governance for Distributed AI Systems

A Framework for Regulatory and Policy Development

John Stroh & Claude (Anthropic)

Document Code: STO-INN-0003
Version 2.1-P (Policy Edition)
January 2026

Executive Summary

AI deployment is outpacing regulatory capacity. While policymakers debate frameworks for large language models operated by major technology companies, a parallel transformation is underway: the migration of AI capabilities to small, locally-deployed models in homes, communities, and small organisations. Recent industry research indicates that 72% of enterprise executives expect small language models (SLMs) to surpass large language models in prominence by 2030 (IBM Institute for Business Value, 2026). This shift creates an urgent governance challenge: who controls AI deployed at the edge, and under what rules?

This paper presents the Tractatus Framework, an architectural approach to AI governance through inference-time constitutional gating. Rather than relying solely on vendor training to ensure AI behaves appropriately, Tractatus requires AI systems to translate proposed actions into auditable forms and evaluate them against explicit constitutional rules before execution. This creates visible, enforceable governance at the point of deployment.

The framework is implemented in the Village platform and designed to accommodate both centralised cloud AI and distributed local deployments, including what we term Sovereign Locally-trained Language Models (SLLs)—AI systems whose training, deployment, and governance remain under community or individual sovereignty rather than vendor control.

Key Policy Implications

1. The governance vacuum is filling by default. In the absence of regulatory frameworks, AI governance is determined by vendor terms of service and platform defaults. This concentrates governance authority in a small number of corporations.
2. Architectural requirements may be more enforceable than behavioural requirements. Mandating that AI systems implement constitutional gating is more verifiable than mandating that AI systems "be safe" or "respect values."
3. Certification infrastructure is needed. As SLM/SLL deployment scales, standards bodies, training providers, and validation methodologies will be required—analogous to existing certification regimes in aviation, medical devices, and financial services.
4. Indigenous data sovereignty is a constitutional matter. In Aotearoa New Zealand and other jurisdictions with indigenous rights frameworks, AI governance must accommodate collective rights over data and culturally-specific governance requirements.
5. Preparation must precede capability. Governance frameworks for advanced AI cannot be developed after such systems exist. Building constitutional infrastructure at accessible scales now creates the foundation for higher-stakes governance later.

1. The Governance Gap

1.1 Regulatory Lag

AI capabilities are advancing faster than governance frameworks can respond. The EU AI Act, while a significant first step, was designed primarily for large-scale systems deployed by identifiable operators. It does not adequately address:

- Edge deployment: AI systems running on personal devices or home servers outside traditional regulatory reach
- Federated architectures: Distributed systems where no single operator controls the complete system
- Continuous adaptation: Models that learn from local data and evolve post-deployment
- Community governance: Situations where appropriate rules vary by cultural context, community values, or individual preferences

1.2 The Coming Wave of Distributed AI

Industry projections indicate a fundamental shift in AI deployment patterns:

Indicator	Current State	2030 Projection	Source
AI contribution to revenue	40% report significant contribution	79% expect significant contribution	IBM IBV 2026, p.13
SLM prominence vs LLM	LLMs dominant in enterprise	72% expect SLMs more prominent	IBM IBV 2026, p.32
AI-driven productivity	Early adoption	42% productivity increase expected	IBM IBV 2026, p.21
Operating margin improvement	Variable	55% higher for multi-model orgs	IBM IBV 2026, p.32

These projections suggest that within five years, AI deployment will be characterised by numerous small, domain-specific models rather than a few large centralised systems. This has profound governance implications:

- Scale of oversight: Thousands of distinct deployments rather than dozens
- Locus of control: Community and individual operators rather than large corporations
- Regulatory jurisdiction: Models operating across borders with no clear home jurisdiction

- Enforcement mechanism: Traditional regulatory inspection may be infeasible at scale

1.3 The Default Governance Regime

In the absence of explicit regulatory frameworks, AI governance defaults to:

1. Vendor terms of service: Corporate policies created to limit liability, not to serve user or community interests
2. Platform architectural choices: Governance embedded in technical infrastructure, invisible to users
3. Market pressure: Systems optimised for engagement and revenue rather than safety or sovereignty

This is not a neutral outcome. It concentrates governance authority in entities whose interests may diverge from those of users, communities, and the public.

2. Architectural Governance: A Regulatory Strategy

2.1 The Limits of Behavioural Regulation

Traditional regulation specifies prohibited outcomes: AI systems must not discriminate, deceive, or cause harm. This approach faces fundamental challenges:

- Verification difficulty: How does a regulator determine whether an AI system “discriminates” without extensive testing that may miss edge cases?
- Definition ambiguity: What constitutes “harm” varies by context; systems optimised for one definition may fail others
- Opacity: Neural network decision-making cannot be directly audited; only inputs and outputs are observable
- Scale: Behavioural testing of thousands of distributed deployments is practically infeasible

2.2 Architectural Requirements

An alternative regulatory strategy specifies required architecture rather than prohibited behaviour:

Constitutional Gating Requirement: AI systems with specified capabilities must implement inference-time constitutional gating—a mechanism that:

1. Transforms model outputs into structured proposals with defined schemas
2. Evaluates proposals against explicit constitutional rules before execution
3. Logs all proposals, evaluations, and dispositions for audit
4. Escalates ambiguous cases to human review

This approach has several advantages:

- Verifiability: The presence of constitutional gating infrastructure can be audited; behavioural compliance cannot
- Transparency: Constitutional rules are explicit and inspectable; training-time alignment is opaque
- Flexibility: Different communities can implement different constitutional rules within a common architectural framework
- Auditability: Logged proposals and evaluations provide an audit trail for incident investigation

2.3 The Tractatus Framework

The Tractatus Framework implements architectural governance through:

Interrupted Inference: Model outputs do not directly affect the world. They are first translated into structured proposals and evaluated against constitutional constraints:

```
User Request → [AI Model] → Proposal → [Constitutional Gate] →
Action/Denial/Escalation
```

Layered Constitutions: Rules are organised in hierarchical layers with explicit precedence:

Layer	Scope	Authority	Examples
Core Principles	Universal	Immutable	No harm; data sovereignty; consent primacy
Platform Rules	All deployments	Amendment by supermajority	Authentication; audit retention
Community Constitution	Per community	Local governance	Content policies; cultural protocols
Individual Preferences	Per user	Self-governed	Communication style; AI memory consent

Authority Model: AI systems operate at defined authority levels, each specifying what actions are permitted without human approval:

Level	Description	Human Role
Advisory	All actions require human approval	Full authority
Operational	Routine actions within defined scope	Exception review
Tactical	Scoped decisions affecting workflows	Outcome oversight

Audit Infrastructure: All proposals, evaluations, and actions are logged with sufficient detail for post-hoc investigation.

3. The SLM/SLL Distinction

3.1 Terminology

We distinguish two deployment paradigms that have different governance implications:

Small Language Model (SLM): A technical descriptor for language models with fewer parameters than frontier LLMs, designed for efficiency and domain-specific deployment. SLMs may be deployed via cloud subscription or locally.

Sovereign Locally-trained Language Model (SLL): An architectural descriptor we introduce for AI systems whose training, deployment, and governance remain under local sovereignty. Key properties:

- Local deployment: Runs on home or community infrastructure
- Local adaptation: Fine-tuned on community-specific data
- Local governance: Subject to community-defined constitutions
- Portable sovereignty: Can participate in federated networks without surrendering governance authority

3.2 Governance Implications

Dimension	Subscription SLM	Sovereign SLL
Regulatory touchpoint	Vendor/platform operator	End deployer/community
Applicable rules	Vendor ToS + jurisdiction law	Local constitution + law
Enforcement mechanism	Platform policy; regulatory action against vendor	Local governance; community accountability
Data jurisdiction	Vendor infrastructure (often unclear)	Local infrastructure (clear jurisdiction)
Amendment authority	Vendor unilaterally	Community democratically
Exit rights	Limited; lose AI context	Full; AI memory portable

3.3 Policy Implications

The SLL paradigm creates both opportunities and challenges for policymakers:

Opportunities:

- Sovereignty preservation: Communities can maintain governance authority over AI affecting them

- Regulatory diversity: Different jurisdictions can implement different governance approaches
- Democratic legitimacy: Governance rules can be developed through community deliberation
- Accountability clarity: Clear relationship between deployer, governance, and jurisdiction

Challenges:

- Enforcement at scale: Traditional regulatory inspection may be infeasible for thousands of home deployments
- Capability creep: Local fine-tuning may create capabilities not anticipated by original safety assessments
- Coordination failure: Fragmented governance may leave gaps or create inconsistencies
- Technical barriers: Not all communities have capacity to implement sophisticated governance

4. A Multi-Layer Containment Framework

4.1 The Inadequacy of Single-Layer Approaches

No single governance mechanism is adequate for AI systems at existential stakes. Defence in depth—multiple independent layers, any one of which might prevent serious harm—is standard in nuclear safety, aviation, and biosecurity. AI governance requires similar architecture.

4.2 Five-Layer Model

Layer	Function	Primary Actors	Current State
1. Capability Constraints	Limit what AI can do regardless of intent	Hardware vendors; compute providers	Emerging (compute governance)
2. Constitutional Gates	Evaluate actions against explicit rules at inference time	Platform operators; community governance	Nascent (Tractatus is early implementation)
3. Human Oversight	Monitor AI systems; intervene when needed	Professional reviewers; community moderators	Ad hoc
4. Organisational Governance	Internal accountability structures	Deploying organisations	Inconsistent
5. Legal/Regulatory	External accountability; enforcement	Governments; international bodies	Minimal

4.3 Layer 2 as Regulatory Focus

Constitutional gating (Layer 2) is particularly amenable to regulatory intervention:

- Specifiable: Requirements can be defined precisely (schema formats, logging requirements, escalation triggers)
- Verifiable: Compliance can be audited through infrastructure inspection and log review
- Flexible: Different constitutional content can implement different policy requirements
- Scalable: Once infrastructure exists, adding rules has minimal marginal cost

Regulatory strategy: mandate the architectural infrastructure, then specify constitutional content through secondary instruments (guidance, standards, sector-specific rules).

5. Certification Infrastructure

5.1 The Need for Standards

As SLM/SLL deployment scales, standardisation becomes essential:

- Interoperability: Different systems should implement compatible governance interfaces
- Verification: Compliance assessment requires common criteria and methodologies
- Training: Constitutional governance requires trained practitioners
- Liability: Clear standards enable liability allocation when things go wrong

5.2 Proposed Certification Ecosystem

Certification Bodies: Define and maintain standards for:

- Proposal schemas and constitutional rule formats
- Gate evaluation semantics and logging requirements
- Validation methodologies and red-team protocols
- Capability threshold specifications and escalation triggers

Training Providers: Offer certified programmes for:

- SLL fine-tuning under constitutional constraints
- Governance configuration for specific contexts (e.g., healthcare, education, cultural)
- Red-team and validation methodology
- Incident response and constitutional amendment

Tooling Vendors: Provide certified implementations of:

- Constitutional gate engines
- Audit and logging infrastructure
- Red-team testing harnesses
- Constitutional UX components for non-expert users

5.3 Regulatory Hooks

Certification creates natural regulatory hooks:

- Licensing: Require certified governance infrastructure for AI deployment above capability thresholds
- Liability: Create safe harbours for deployments using certified infrastructure; increased liability for uncertified deployments

- Procurement: Government procurement can require certified constitutional governance
- Insurance: Insurers can offer favourable terms for certified deployments

6. Indigenous Data Sovereignty

6.1 Constitutional Requirements in Aotearoa New Zealand

AI governance in Aotearoa operates under Te Tiriti o Waitangi, which guarantees Māori tino rangatiratanga (unqualified chieftainship) over taonga (treasures). Courts and the Waitangi Tribunal have established that taonga extends to language, culture, and knowledge systems.

Data is taonga. AI systems that process Māori data or affect Māori communities engage constitutional obligations, not merely policy preferences.

6.2 Te Mana Raraunga Principles

Te Mana Raraunga, the Māori Data Sovereignty Network, articulates principles including:

- Rangatiratanga: Māori have authority over data about them
- Whakapapa: Data exists within relational contexts that must be respected
- Whanaungatanga: Data governance is collective, not merely individual
- Kaitiakitanga: Data custodians have guardianship responsibilities

6.3 Policy Implications

Constitutional governance for AI must accommodate:

- Collective consent: Some data governance decisions require community authority, not just individual consent
- Cultural protocols: Appropriate handling of certain information may require tikanga-specific rules
- Benefit sharing: AI trained on Māori data may create obligations regarding benefit distribution
- Governance participation: Māori should participate in governance of AI systems affecting them

The Tractatus Framework's layered constitutional architecture can accommodate these requirements: tikanga-based rules can be instantiated in community constitutions without requiring universal adoption. However, platform-level accommodation is insufficient—Māori data sovereignty requires legislative recognition and enforcement mechanisms.

6.4 Relevance Beyond Aotearoa

Indigenous peoples worldwide face similar challenges. Frameworks developed in Aotearoa—grounded in Te Tiriti jurisprudence and informed by Māori legal philosophy—may offer models for indigenous AI governance globally. The CARE Principles for

Indigenous Data Governance (Collective Benefit, Authority to Control, Responsibility, Ethics) provide an international reference point.

7. From Existential Stakes to Everyday Governance

7.1 Why Existential Risk Framing Matters for Policy

The existential risk literature may seem remote from practical policy concerns about home AI assistants. The connection is essential:

Containment architectures cannot be developed after the systems that need them exist. If advanced AI systems eventually pose existential risks—a possibility serious researchers take seriously—the governance infrastructure, institutional capacity, and cultural expectations required to contain them must be developed in advance.

Current deployment is the development ground. The patterns that work at village scale become the patterns available when stakes are higher. Constitutional gating implemented for home SLLs creates:

- Open-source tooling adaptable to higher-stakes deployments
- Validation methodologies tested against real adversarial pressure
- Professional communities with containment expertise
- Regulatory precedents for mandating architectural governance
- Public understanding of what AI governance means

This is preparation, not prediction. We do not know if existential risks will materialise. We do know that governance capacity cannot be created instantly when needed. Prudent policy builds that capacity now.

7.2 Capability Thresholds and Escalation

The Tractatus Framework includes explicit capability thresholds:

Below threshold: Constitutional gating provides governance infrastructure appropriate for current SLMs, SLLs, and LLMs operating within human-comprehensible parameters.

Above threshold: Stronger constraints apply:

- Layer 1 capability restrictions (air-gapping, capability cuts)
- Mandatory external oversight
- Development pause pending verification advances

Escalation triggers include:

- Evidence of deceptive behaviour (misrepresentation in proposals)
- Multi-step circumvention (individually-acceptable proposals aggregating to prohibited outcomes)
- Capability surprises (demonstrated capabilities not predicted by assessments)

Policymakers should consider tiered regulatory requirements that intensify as capability thresholds are crossed.

8. Recommendations for Policymakers

8.1 Immediate Actions

1. Commission technical standards development for constitutional gating infrastructure, including proposal schemas, logging requirements, and validation methodologies.
2. Establish pilot certification programmes for SLL training providers and governance tooling vendors.
3. Include constitutional gating requirements in government AI procurement standards.
4. Engage indigenous governance bodies on AI governance requirements and implementation.

8.2 Medium-Term Framework Development

1. Develop tiered regulatory requirements based on capability thresholds, with constitutional gating as baseline for all AI systems above specified capability levels.
2. Create liability frameworks that incentivise certified constitutional governance (safe harbours for certified deployments; increased liability for uncertified).
3. Establish independent oversight bodies with technical capacity to audit constitutional governance implementation.
4. Develop mutual recognition frameworks with other jurisdictions for constitutional governance certification.

8.3 International Coordination

1. Propose constitutional gating standards through international standards bodies (ISO, IEEE).
2. Develop treaty frameworks for cross-border AI governance, including mutual recognition of certification regimes.
3. Support indigenous governance coalitions developing international principles for indigenous AI sovereignty.

9. Honest Assessment of Limitations

9.1 What Constitutional Gating Cannot Do

- Contain superintelligent systems: The framework assumes AI operating within human-comprehensible parameters
- Guarantee behavioural alignment: Architecture constrains actions, not intentions
- Solve international coordination: Application-layer governance does not address global capability races
- Enforce adoption: Frameworks only protect where implemented; market incentives may favour uncontained deployment

9.2 Remaining Uncertainties

- Scaling properties: We do not know how constitutional gating behaves as model capabilities increase
- Adversarial robustness: Sophisticated systems may find ways to satisfy constitutional rules while achieving prohibited outcomes
- Governance fatigue: Multi-layer governance may prove too complex for widespread adoption
- Enforcement feasibility: Regulatory oversight of thousands of distributed deployments may prove impractical

9.3 The Case for Action Despite Uncertainty

These uncertainties are not arguments against constitutional governance. They are arguments for:

- Iterative development: Build, deploy, learn, improve
- Research investment: Fund investigation of scaling properties and adversarial robustness
- Flexible frameworks: Design regulations that can adapt as understanding evolves
- Precautionary approach: Act on the basis of serious possibility, not just certainty

The alternative—waiting for certainty before acting—guarantees that governance frameworks arrive after the need has become acute.

10. Conclusion

The governance gap in AI deployment is widening. As capabilities migrate to distributed, locally-deployed systems, traditional regulatory approaches face fundamental challenges of scale, jurisdiction, and verification.

Constitutional gating offers a regulatory strategy: mandate auditable architectural infrastructure rather than unverifiable behavioural requirements. The Tractatus Framework provides a concrete specification that can be implemented across deployment paradigms—from cloud LLMs to sovereign home SLLs.

The policy window is now. Within five years, if industry projections hold, AI deployment will be characterised by thousands of small, domain-specific models operating in homes, communities, and small organisations. Governance frameworks developed now will shape that landscape; frameworks developed later will struggle to retrofit.

We offer this analysis in the spirit of contribution to ongoing policy deliberation. The questions are hard, the uncertainties substantial, and the stakes significant. Policymakers, researchers, and communities must work together to develop governance frameworks adequate to the challenge.

References

IBM Institute for Business Value. (2026). The enterprise in 2030. IBM Corporation.

Te Mana Raraunga. (2018). Māori Data Sovereignty Principles. Te Mana Raraunga – Māori Data Sovereignty Network.

Waitangi Tribunal. (2011). Ko Aotearoa Tēnei: A Report into Claims Concerning New Zealand Law and Policy Affecting Māori Culture and Identity (Wai 262). Legislation Direct.

Research Institute for Indigenous Data Sovereignty. (2019). CARE Principles for Indigenous Data Governance. Global Indigenous Data Alliance.

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

Carlsmith, J. (2022). Is power-seeking AI an existential risk? arXiv preprint arXiv:2206.13353.

European Parliament. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act).

Reason, J. (1990). Human Error. Cambridge University Press.

Sastry, G., et al. (2024). Computing power and the governance of artificial intelligence. arXiv preprint arXiv:2402.08797.

— End of Document —