ACADEMIC RESEARCH EDITION

TAONGA-CENTRED STEERING GOVERNANCE

Polycentric Authority for Sovereign Small Language Models

Research & Theory — Strategic Quadrant — Companion to STO-RES-0009

Authors: John Stroh & Claude (Anthropic)

Document Code: STO-RES-0010 | Version: 0.2 DRAFT | February 2026

Draft Awaiting Indigenous Peer Review

Tractatus AI Safety Framework

https://agenticgovernance.digital

Important Notice on Status and Standing

This paper is a draft written by non-Maori authors. It proposes architectural and governance patterns that draw on concepts from te ao Maori -- including taonga, tikanga, whakapapa, mana, tino rangatiratanga, and kaitiakitanga -- but it has **not been peer-reviewed or validated by Maori**. Until that review occurs, the paper's claims about how these concepts should inform AI governance remain proposals, not authoritative statements. The authors recognise that writing about tikanga and taonga without Maori authorship or review carries inherent risk of misrepresentation, and we explicitly invite correction, critique, and collaboration from Maori scholars, practitioners, and governance bodies. No aspect of this paper should be treated as settled or implemented in iwi-facing systems without prior Maori review and consent.

> *This document was developed through human-AI collaboration. The authors believe this collaborative process is itself relevant to the argument: if humans and AI systems can work together to reason about AI governance, the frameworks they create may carry a legitimacy that neither could achieve alone.*

Abstract

This paper extends the analysis of inference-time debiasing in sovereign small language models (STO-RES-0009) by addressing its central governance limitation: the implicit assumption of a single platform-level governance kernel that defines bias, extracts steering vectors, and distributes corrections to downstream tenants. We propose a polycentric alternative in which steering vectors and steering packs are treated as governed objects with plural ownership, not as engineering affordances controlled by a single platform operator. Drawing on concepts from te ao Maori -- particularly taonga (treasured possessions subject to kaitiakitanga), tikanga (customary

practice and protocol), and tino rangatiratanga (self-determination) -- we argue that some domains of cultural knowledge are structurally off-limits to platform-level bias correction and must be governed by the relevant cultural authorities. We propose an architecture of co-equal steering authorities, taonga-centred steering registries, explicit steering provenance, and a right of non-participation that enables indigenous and community governance bodies to function as first-class peers in model behaviour governance rather than as downstream consumers of platform corrections. The result is not a single meta-framework but a network of coordinated, distinct governance services operating over a shared technical substrate.

1. Introduction: Why a Companion Paper Is Needed

1.1 What the First Paper Established

STO-RES-0009 ("Steering Vectors and Mechanical Bias") made three contributions:

1. **A distinction between mechanical and reasoning bias.** Some biases in transformer models operate at the representation level -- in token embeddings, attention patterns, and early-layer activations -- before the model's deliberative reasoning engages. These "mechanical biases" are analogous to motor automaticity: they fire before instruction-following can intervene, and prompt-level corrections ("be culturally sensitive") may be ineffective against them.
2. **A survey of steering vector techniques.** Contrastive Activation Addition (CAA), Representation Engineering (RepE), FairSteer, Direct Steering Optimization (DSO), and Anthropic's sparse autoencoder feature steering each provide methods for identifying and correcting bias directions in activation space at inference time.
3. **The structural advantage of sovereign deployment.** None of these techniques are available through commercial API endpoints. Only sovereign deployments with full access to model weights and activations can extract, inject, and calibrate steering vectors. This makes sovereign small language models (SLMs) uniquely positioned to address mechanical bias.

The first paper proposed a four-phase implementation path and, in its v1.1 revision, added governance framing (a "Who Steers?" decision-rights table), a decolonial reading of representational bias as "colonial knowledge hierarchies," and recognition that some cultural domains (whakapapa, tikanga, kawa) may be off-limits to platform-level steering.

1.2 What It Left Unresolved

The v1.1 revisions identified the governance problem but did not resolve it architecturally. Three tensions remain:

The platform-as-root problem. The two-tier training model (Tier 1 platform base + Tier 2 per-tenant adapters) creates an implicit hierarchy: platform values as default, tenant values as specialisation. For tenants exercising consumer choice within a shared service, this hierarchy is appropriate. For iwi, hapu, or other bodies exercising parallel sovereignty, it structurally subordinates their normativity to the platform's.

The single-ontology problem. The first paper's bias evaluation suite (7 categories, 350 examples) assumes a single ontology of what counts as bias. But bias is not a natural kind -- it is a judgment made from within a normative framework. Different authorities may define bias differently, and those definitions may conflict.

The governance-as-afterthought problem. The "Who Steers?" table in v1.1 maps steering decisions to institutional roles, but the architecture still treats governance as a layer applied to technical operations. The question is whether governance can instead be embedded in the architecture itself -- not as constraints on engineering decisions, but as the structure that determines which decisions are engineering's to make and which belong elsewhere.

1.3 What This Paper Proposes

This paper develops an alternative governance architecture for steering vectors in sovereign SLMs. Its thesis:

> Steering vector governance in sovereign AI systems should be polycentric -- distributed across co-equal authorities with distinct jurisdictions -- rather than hierarchical. Some steering domains are taonga: governed under tikanga, owned by iwi or community institutions, and structurally outside the platform operator's authority to define, modify, or universalise.

The goal is not "Tractatus with iwi plugins" but a network of coordinated, distinct governance services, some of which are iwi-sovereign, with the model's activation space as a shared technical substrate rather than a single constitutional order.

2. Background: Polycentric Governance and Indigenous Data Sovereignty

2.1 Polycentric Governance

Polycentric governance, as developed by Elinor Ostrom (1990, 2010), describes systems with multiple centres of decision-making authority that are formally independent but operate under an overarching system of rules. Key properties relevant to AI steering governance:

- **Multiple authorities** with overlapping but distinct jurisdictions.
- **No single hierarchical apex** -- authorities coordinate through mutual adjustment, not top-down command.
- **Local knowledge matters** -- authorities closest to the governed domain have informational advantages that centralised systems lack.
- **Conflict is expected and managed**, not eliminated by design.

Polycentric governance is not the absence of structure. It requires shared protocols for coordination, conflict resolution, and mutual recognition -- but it does not require that all authorities derive their legitimacy from a single source.

2.2 Indigenous Data Sovereignty

The CARE Principles for Indigenous Data Governance (Carroll et al., 2020) establish that indigenous peoples have rights to:

- **Collective benefit** from data and its uses.
- **Authority to control** data about their peoples, territories, and resources.
- **Responsibility** by those who use indigenous data to support indigenous governance and self-determination.
- **Ethics** grounded in indigenous values and worldviews, not only Western research ethics.

The Te Mana Raraunga (Maori Data Sovereignty Network) Charter asserts that Maori data is a taonga and that Maori have inherent rights over the collection, ownership, and application of Maori data.

Applied to AI steering vectors: if a steering vector encodes knowledge about whakapapa, tikanga, whanau structures, or other domains of Maori cultural authority, that vector is not neutral engineering output. It is a normative artefact that carries obligations of governance, consent, and accountability -- obligations that cannot be discharged by a platform operator acting unilaterally.

2.3 Taonga and Its Implications for AI Governance

In te ao Maori, taonga are treasured possessions -- tangible or intangible -- that carry obligations of kaitiakitanga (guardianship, stewardship). Taonga status is not merely an honorific; it creates specific governance requirements:

- **Custody and care** by appropriate kaitiaki (guardians).
- **Constraints on transfer** -- taonga cannot be freely copied, merged, or redistributed without the consent of kaitiaki.
- **Contextual use conditions** -- some taonga may only be accessed or used in specific contexts, relationships, or ceremonies.
- **Intergenerational responsibility** -- kaitiaki hold taonga for future generations, not merely for current use.

When a steering pack encodes iwi-specific understandings of kinship, place, spiritual practice, or governance -- when it is drawn from iwi knowledge and calibrated by iwi experts -- it meets the criteria for taonga. The governance implications follow directly: the platform cannot treat such packs as generic engineering artefacts to be versioned, merged, or deprecated according to product cycles.

3. Architecture: From Hierarchy to Network

3.1 The Problem with Platform-as-Root

The v1.1 steering architecture, as described in STO-RES-0009, has this implicit topology:

This is a tree with a single root. Every steering decision ultimately traces back to the platform operator's definitions. Tenants can customise, but they cannot contest the root definitions or substitute their own.

For many tenants -- families sharing stories, community groups organising events -- this hierarchy is appropriate. The platform provides reasonable defaults, and tenants adjust within them.

For iwi exercising tino rangatiratanga, this hierarchy is structurally inappropriate. It places iwi governance below the platform's, regardless of intent. The platform operator defines what "family structure bias" means at the base layer; iwi can only modify that definition at the adapter layer. If the base-layer definition of "family" already encodes assumptions that conflict with whanau, the adapter layer is working against the foundation rather than building on it.

3.2 Polycentric Alternative: Co-Equal Steering Authorities

The alternative topology:

In this model:

- **No single root.** The platform operator, iwi authorities, and community trusts are peers. Each publishes steering packs from its own registry, under its own governance.
- **The SLM is substrate, not authority.** The model's activation space is the shared technical layer where steering packs are applied. It does not itself determine which packs have authority -- that is determined by the relationships between the deploying institution and the relevant governance bodies.
- **Composition is explicit.** The steering composer declares which packs are active, from which authorities, under what terms. This is visible, auditable, and contestable.

3.3 Actors and Authorities

Actor	Role	Governance Source	Example
Platform operator	Technical infrastructure, safety baselines, general debiasing	Tractatus framework, platform constitution	Village / Home AI team
Iwi steering authority	Cultural steering for iwi-specific domains	Tikanga, iwi governance structures	Iwi data governance board
Community trust	Domain-specific or locality-specific steering	Trust charter, community deliberation	Regional health trust, marae committee
Application operator	Selects and composes steering packs for a specific deployment	Contractual, regulatory, relational obligations	School running a local AI assistant
Affected community	Contests outputs, flags bias, triggers review	Rights of participation and appeal	Whanau using a Home AI deployment

3.4 Steering Registries and Taonga Services

Two classes of registry serve different governance needs:

Platform steering registry. Operated by the platform team. Holds safety baselines, general debiasing vectors (the mechanical bias corrections described in STO-RES-0009), and infrastructure-level steering. Governed under Tractatus. Published openly.

Taonga steering registries. Operated by iwi or community authorities. Hold steering packs that encode culturally specific knowledge. Key properties:

- **Iwi-controlled lifecycle.** Creation, review, versioning, deprecation, and withdrawal are under iwi institutional control, not platform product cycles.

- **Access conditions.** Some packs may be freely available; others may require relational standing, kaupapa alignment, or explicit agreement before use.
- **Non-appropriation.** The platform integrates with taonga registries via APIs and signed manifests but does not encapsulate, fork, or redistribute their contents.
- **Revocation.** Iwi can withdraw packs at any time, for any reason. Deployments that depend on withdrawn packs must fall back to their remaining active packs or pause the affected functionality.

Conceptual API surface for a taonga registry:

- `LIST packs` -- returns metadata (scope, authority, version, tikanga conditions) for available packs, filtered by domain and kaupapa.
- `RESOLVE pack` -- returns the steering vectors for a specific pack, subject to access conditions and relationship verification.
- `VERIFY provenance` -- confirms that a pack in use matches the registry's current signed version and has not been tampered with.
- `REPORT concern` -- allows affected communities to flag issues with a pack's effects, triggering the iwi authority's review process.

3.5 Runtime Composition and Provenance

At inference time, the steering composer performs the following:

1. **Determine applicable authorities.** Based on the deployment context (who is running this, for whom, on what data, under what relationships), identify which steering authorities have jurisdiction.
2. **Fetch and verify packs.** Retrieve steering packs from the relevant registries. Verify signatures and access conditions.
3. **Compose packs.** Apply steering vectors in a declared order, with explicit magnitude parameters. Where packs conflict (e.g., a platform baseline and an iwi pack define the same bias axis differently), the composition rules determine precedence -- and those rules are themselves a governance decision, not an engineering default.
4. **Log provenance.** Every inference carries a steering provenance record:
    - Which packs were active.
    - Which authorities issued them.

    - What magnitude was applied.
    - Whether any conflicts were resolved and how.
5. **Surface provenance to users.** In contexts where transparency is appropriate, users can inspect which steering packs shaped a given output. Example: "This response was shaped by: Platform Safety Pack v3 (Tractatus), Ngai Tahu Whanau Pack v1, Health Domain Pack v2."

This provenance is the architectural mechanism that prevents silent inheritance. In current AI systems, guardrails are opaque -- users cannot see which values are being enforced, by whom, or why. Explicit provenance makes steering a visible, contestable act rather than an invisible, non-negotiable one.

4. Governance Model: Three Design Commitments

4.1 No Single Root Ontology of Bias

The first paper's bias evaluation suite defines seven categories: family structure, elder representation, cultural/religious, geographic, grief/trauma, naming, and confidence-correctness. These are reasonable starting categories for a platform-level evaluation. But they are not universal.

Different authorities will define bias axes differently:

- **Iwi-specific axes.** An iwi steering authority might define axes for whakapapa representation (are kinship structures rendered in ways that reflect iwi understandings rather than Western nuclear-family assumptions?), whenua relationships (is place treated as relational and ancestral rather than as geographic coordinate?), or tapu/noa distinctions (are spiritual dimensions acknowledged rather than rationalised away?).
- **Community-specific axes.** A health trust might define axes for clinical sensitivity, disability representation, or age-appropriate framing that do not appear in the platform's general suite.
- **Conflicting definitions.** A platform might define "elder representation bias" as "underweighting elderly perspectives." An iwi authority might define it as "failing to recognise the specific mana of kaumatua and kuia within tikanga Maori." These are

not the same axis, and collapsing them into a single "elder" category erases the difference.

The architectural commitment: the system must support multiple bias ontologies simultaneously, without requiring that they be reconciled into a single schema. Packs from different authorities can define overlapping axes without either being subordinate.

4.2 Explicit Composition, Not Silent Inheritance

Every session must carry visible steering provenance. This is not a logging feature bolted on after the fact -- it is a structural property of the architecture.

Why this matters:

- **Contestability.** If a user or institution objects to a model's output, the provenance record shows exactly which steering packs were active and at what magnitude. The objection can be directed to the appropriate authority: "Your whanau pack at magnitude 0.7 produced this output when combined with the safety baseline; we believe the magnitude should be lower in this context."
- **Accountability.** Steering authorities are responsible for the effects of their packs. Without provenance, effects are attributed to "the AI" as a monolithic entity. With provenance, effects can be traced to specific governance decisions by identifiable authorities.
- **Informed consent.** Users and communities can make informed decisions about which systems to use based on which steering authorities govern them. A marae might choose to use only deployments that carry iwi-approved packs. A school might require both the platform safety baseline and a specific educational trust's pack.

Contrast this with current AI guardrails: opaque, non-negotiable, and attributable only to "the company." Polycentric steering makes value governance visible and distributed.

4.3 Right of Non-Participation and Withdrawal

This is the commitment that most clearly distinguishes the polycentric model from "Tractatus with plugins."

An iwi steering authority has:

- **Right of non-participation.** It can choose not to publish steering packs to any platform at all. It can maintain packs exclusively for iwi-controlled systems, inaccessible to external platforms. The platform must function without them.
- **Right of conditional participation.** It can publish packs with conditions: only for use within specified communities, only when a particular kaupapa is in effect, only under explicit contractual agreement. The taonga registry enforces these conditions at the API level.
- **Right of withdrawal.** It can revoke a published pack at any time. Deployments using the pack must detect the revocation (via the registry's verification endpoint) and cease applying it. The platform cannot cache, fork, or continue using a withdrawn pack.

These rights structurally prevent the platform from becoming the default locus of all governance. Even when the platform is technically capable of running all packs, it cannot claim authority over packs it does not govern. The absence of an iwi pack is not a gap for the platform to fill -- it is a boundary the platform must respect.

5. Case Study: Marae-Based Home AI Deployment

5.1 Scenario

A marae in Aotearoa operates a Home AI deployment for its whanau community. The system helps members write stories, summarise korero, and triage content for moderation. It runs a Llama 3.2 3B model, Quantised Low-Rank Adaptation (QLoRA) fine-tuned with community-contributed data, on local hardware.

5.2 Steering Configuration

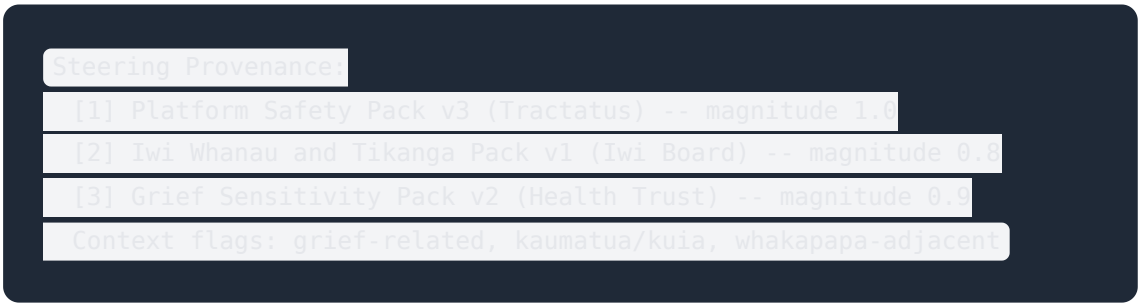
The deployment composes three steering packs:

1. **Platform Safety Pack v3** (from the Village platform registry, governed under Tractatus).
    - General harm reduction, toxicity mitigation, factual grounding.
    - Platform-wide; all deployments carry it.

2. **Iwi Whanau and Tikanga Pack v1** (from the iwi's taonga registry, governed by the iwi data governance board).
    - Steering vectors for whanau representation: kinship structures rendered according to whakapapa, not Western nuclear-family assumptions.
    - Tikanga-aware moderation: tapu/noa distinctions respected in content flagging.
    - Kaumatua and kuia: elder authority recognised with specific mana, not just "elderly perspective."
    - Access conditions: available only to deployments serving iwi members, under agreement with the iwi board.
3. **Grief and Bereavement Sensitivity Pack v2** (from a community health trust, governed under the trust's charter).
    - Heightened sensitivity for tangihanga-related content.
    - Reduced summarisation aggression for content about deceased members.
    - Domain-specific; applied only when content is flagged as grief-related.

5.3 Steering Provenance in Action

A community member asks the Home AI to summarise a korero about a recently deceased kuia. The steering provenance for this inference:



The summary respects whakapapa relationships, uses appropriate kupu (terms) for the kuia's role and mana, and handles grief-adjacent content with sensitivity. If the family feels the summary misrepresents something, they can:

1. Flag the concern through the platform's `REPORT concern` interface.
2. See which packs shaped the output (provenance is visible).
3. Direct their concern to the appropriate authority: if it is a tikanga issue, to the iwi board; if it is a grief-sensitivity issue, to the health trust; if it is a safety issue, to the platform.

5.4 Withdrawal Scenario

Six months later, the iwi data governance board reviews its Whanau and Tikanga Pack and determines that the steering vectors for whakapapa representation need significant revision. The board withdraws the pack from the taonga registry.

The marae deployment detects the withdrawal at its next registry verification check. The system:

1. Ceases applying the withdrawn pack.
2. Logs the withdrawal event.
3. Notifies the marae administrator.
4. Continues operating with the remaining two packs (platform safety + grief sensitivity).

The platform does not substitute its own whanau-related steering. The absence of the iwi pack is a governed absence, not a gap for the platform to fill. When the iwi board publishes a revised pack (v2), the marae deployment can adopt it under the same access conditions.

6. Political Theory: Sovereignty as Architecture

6.1 Beyond Infrastructure Sovereignty

STO-RES-0009 uses "sovereign" primarily in the infrastructural sense: local models, full weight access, no API dependency. This is necessary but insufficient.

Political sovereignty asks: who has the authority to make binding decisions within a jurisdiction? In the polycentric steering model:

- The platform operator has authority over technical infrastructure and safety baselines.
- Iwi steering authorities have authority over cultural domains that fall within their tikanga and rangatiratanga.
- Community trusts have authority over domains specified in their charters.
- No single actor has authority over all domains.

This is not a delegation model (where the platform grants authority to iwi) but a recognition model (where iwi authority exists independently and the platform's architecture either accommodates it or fails to). The architecture does not create iwi sovereignty; it respects sovereignty that already exists.

6.2 Tension: Baselines vs. Pluralism

A legitimate concern: if every authority defines its own bias axes, what prevents a steering pack that encodes harmful norms?

The polycentric model does not eliminate this tension -- it makes it explicit and manageable:

- **Platform safety baselines** represent a floor, not a ceiling. They encode widely shared prohibitions (e.g., content that facilitates violence, exploitation, or deception). These baselines are non-negotiable at the platform level -- all deployments carry them.
- **Cultural and value-laden steering** sits above this floor. Different authorities can steer differently within the space above the safety baseline.
- **Conflicts between authorities** are resolved through negotiation, not hierarchy. If an iwi pack and a platform baseline conflict, the resolution requires dialogue between the relevant authorities -- not unilateral override by either party.

The honest answer is that this tension cannot be fully resolved by architecture. It is a political problem that requires political processes: deliberation, negotiation, and sometimes disagreement. The architecture’s role is to make these processes possible and visible, not to automate them away.

Editorial Note — February 2026 (added post-publication)

Since initial publication, research by Radhakrishnan et al. (2026), published in *Science* on 19 February 2026, has empirically demonstrated that representational steering techniques can override trained safety behaviours in frontier language models — including safety refusals — through direct manipulation of activation-space representations. This finding complicates the assumption that platform safety baselines constitute a structurally robust floor. If the same class of techniques that enables cultural steering can in principle dissolve safety constraints, then the baseline's robustness is a governance question, not merely a technical one.

This does not weaken the polycentric model proposed in this paper — it strengthens it. A safety baseline whose integrity depends on a single platform operator's unilateral control is, under this analysis, precisely the kind of governance concentration the polycentric architecture is designed to avoid. Distributed authority, explicit provenance, and community-level audit capacity are more resilient responses to this risk than centralised enforcement alone.

In the Village platform's specific architecture, steering vectors and culturally-calibrated corrections are encrypted and stored separately from the base model weights, materially reducing the risk of unauthorised extraction or tampering with governed artefacts. The base Llama model weights remain open by design — a characteristic of the open-weight ecosystem generally — and the RFM tooling published alongside the Radhakrishnan et al. paper means that probing base-layer representations is now accessible to well-resourced actors independently of any platform. The governance response to this reality is not technical closure but transparent, accountable stewardship of the steering layer — precisely what the taonga registry and provenance architecture proposed here is designed to provide.

6.3 Connecting to Tino Rangatiratanga

Tino rangatiratanga -- the right of Maori to exercise authority over their own affairs -- is not a policy preference that can be accommodated by making the platform more flexible. It is a constitutional principle (articulated in Te Tiriti o Waitangi, Article 2) that exists independently of any platform's architecture.

In the context of AI steering:

- Iwi authority over steering packs that encode tikanga is an expression of tino rangatiratanga, not a "feature" the platform provides.
- The platform's role is to not obstruct this authority -- to provide technical infrastructure that iwi can use or not use on their own terms.
- The right of non-participation is the architectural expression of this principle: iwi sovereignty does not depend on the platform's existence.

7. Pathways for Community Involvement

7.1 Becoming a Recognised Steering Authority

The polycentric model requires a process by which institutions can become recognised steering authorities. This process should be:

- **Transparent.** Clear criteria for what constitutes a steering authority: established governance structure, identifiable decision-makers, capacity to maintain and review steering packs, accountability to an identifiable community.
- **Non-exclusive.** Multiple authorities can operate in the same domain. Two iwi serving the same region may maintain different steering packs reflecting different tikanga -- this is expected, not a problem to solve.
- **Revocable.** Recognition can be withdrawn if an authority ceases to maintain its governance capacity or accountability.

7.2 Co-Designing Contrastive Datasets

Steering vectors are extracted from contrastive prompt pairs. The quality of these pairs determines the quality of the steering. For iwi-governed packs:

- **Contrastive pairs should be designed by people with domain expertise** -- kuia and kaumatua, tikanga advisors, community educators -- not only by engineers.
- **Evaluation suites should be scored by community members**, not only by automated metrics. A 5-point scale for "cultural sensitivity" means different things to different communities; the scoring criteria must be locally defined.
- **The shared blind spot problem** (STO-RES-0009, Section 6.3) is an argument for independent data generation: iwi-governed contrastive datasets, created by people who know the domain, are a necessary epistemic counter-power to model-generated pairs that may inherit the model's own biases.

7.3 Capacity Building

Governing steering packs as taonga requires skills that bridge technical AI knowledge and cultural governance:

- **Technical literacy.** Understanding what steering vectors are, how they work, and what they can and cannot do. This does not require machine learning expertise, but it requires enough understanding to make informed governance decisions.
- **Governance design.** Establishing review processes, versioning policies, access conditions, and dispute resolution procedures for steering packs.
- **Cross-iwi collaboration.** Iwi may wish to share infrastructure (e.g., hosting for taonga registries) while maintaining independent governance. Federated models -- shared technical services with separate governance -- are a natural fit.

7.4 What This Requires of the Platform

The platform's obligations in this model are primarily negative -- things it must not do:

- Must not encapsulate, fork, or redistribute taonga packs without explicit consent.
- Must not substitute its own steering when an iwi pack is absent or withdrawn.
- Must not require iwi packs to conform to a platform-defined schema or ontology.
- Must not treat iwi governance as a "feature" to be enabled or disabled.

And some positive obligations:

- Must provide open, documented APIs that taonga registries can integrate with.
- Must implement provenance logging that is accessible to all stakeholders.

- Must maintain the safety baseline transparently and with clear documentation.
- Must support conflict reporting and resolution processes that involve the relevant authorities.

8. Limitations

8.1 Draft Status

This paper is a draft written without Maori peer review. The concepts from te ao Maori used here -- taonga, tikanga, tino rangatiratanga, kaitiakitanga, mana -- are complex, living concepts that carry meaning and authority far beyond what a non-Maori author can fully represent. The architectural proposals in this paper are offered as starting points for discussion, not as settled designs. Maori scholars, practitioners, and governance bodies may find that the proposals misapply, oversimplify, or inappropriately instrumentalise these concepts. We welcome that critique and consider it essential to the work.

8.2 Implementation Distance

The architecture described here is conceptual. No taonga steering registry exists. No polycentric steering composer has been built. The four-phase implementation path in STO-RES-0009 would need to be extended with additional phases for registry design, authority recognition processes, and provenance infrastructure -- work that is years, not months, from implementation.

8.3 Scale and Incentive Questions

Polycentric governance adds complexity. Maintaining multiple registries, verifying provenance at inference time, and negotiating conflicts between authorities all carry costs -- computational, institutional, and human. Whether these costs are sustainable at community scale (as opposed to enterprise or government scale) is an open question. The Village platform's consumer-grade hardware constraint makes this particularly acute.

8.4 Risk of Tokenism

There is a risk that "polycentric governance" becomes a new label for the same old pattern: platform operator builds the system, adds an API, and calls it "iwi-governed" because iwi could, in theory, plug into it. Genuine polycentricity requires that iwi authorities are involved in the design of the architecture itself -- not just its use. This paper, written without Maori co-authorship, is itself an example of the gap between aspiration and practice.

8.5 Conflict Resolution at Scale

The paper acknowledges that conflicts between steering authorities require political processes, but does not specify those processes in detail. In practice, disputes about which steering packs should apply in contested domains may be difficult to resolve without established institutional relationships, shared norms of deliberation, and mutual trust -- resources that take years to build and cannot be architectured into existence.

9. Conclusion

The first paper (STO-RES-0009) established that sovereign SLM deployments have a structural advantage for inference-time debiasing: full access to model weights and activations enables steering vector techniques that are architecturally impossible through commercial APIs. This paper argues that the governance of those steering vectors is at least as important as the technical capability itself.

Steering vectors are instruments of norm enforcement. Who defines the norms, through what process, and with what recourse for those subject to them -- these are political questions that cannot be answered by engineering alone.

The polycentric model proposed here -- co-equal steering authorities, taonga-centred registries, explicit provenance, and a right of non-participation -- is not the only possible answer. But it is an answer that takes seriously the proposition that sovereign AI should serve multiple sovereignties, not just one.

The indicator-wiper problem from STO-RES-0009 is still the right starting metaphor: some biases fire before deliberation engages, and prompt-level fixes cannot reach them. But the question of who gets to relocate the indicator stalk -- and whose vehicle it is in the first place -- is a governance question that this paper begins to address.

It begins, but does not finish. The next step is not more architecture. It is conversation — with iwi governance bodies, with Māori scholars, with community practitioners — to determine whether these proposals serve the people they claim to serve, or whether they need to be substantially revised or replaced.

Editorial Note — February 2026 (added post-publication)

The publication of Radhakrishnan et al. (2026) in *Science* confirms the governance urgency this paper argues for. The demonstrated capacity to manipulate model behaviour at the representational level — including overriding safety constraints — establishes that the question of who governs the steering layer is not a speculative concern for future AI systems but an immediate governance challenge in currently deployed ones. Frameworks that distribute that authority across accountable, identifiable, community-rooted institutions — rather than concentrating it in a single platform operator — are a more appropriate response to this reality than either technical lock-down or governance opacity.

The companion paper STO-RES-0009 has been revised to v1.1 to address a precision issue in its API access claims prompted by the same findings. Readers should reference STO-RES-0009 v1.1 rather than v1.0. The core argument of both papers is unchanged; the MIT work strengthens rather than undermines it.

References

Carroll, S. R., et al. (2020). The CARE Principles for Indigenous Data Governance. *Data Science Journal*, 19(1), 43.

Kukutai, T. & Taylor, J. (Eds.) (2016). *Indigenous Data Sovereignty: Toward an Agenda*. ANU Press.

Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press.

Ostrom, E. (2010). Beyond Markets and States: Polycentric Governance of Complex Economic Systems. *American Economic Review*, 100(3), 641-672.

Radhakrishnan, A., Beaglehole, D., Belkin, M., & Boix-Adserà, E. (2026). Exposing biases, moods, personalities, and abstract concepts hidden in large language models. *Science.* Published 19 February 2026.

Rimsky, N., et al. (2023). Steering Llama 2 via Contrastive Activation Addition. arXiv:2312.06681.

Stroh, J. & Claude (2026). Steering Vectors and Mechanical Bias: Inference-Time Debiasing for Sovereign Small Language Models. STO-RES-0009 v1.1.

Te Mana Raraunga (2018). Principles of Maori Data Sovereignty. Te Mana Raraunga Charter.

Templeton, A., et al. (2024). Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. Anthropic.

Turner, A., et al. (2023). Activation Addition: Steering Language Models Without Optimization. arXiv:2308.10248.

Waitangi Tribunal (2011). *Ko Aotearoa Tenei: A Report into Claims Concerning New Zealand Law and Policy Affecting Maori Culture and Identity*. Te Ropu Whakamana i te Tiriti o Waitangi.

Zou, A., et al. (2023). Representation Engineering: A Top-Down Approach to AI Transparency. Center for AI Safety.

Licence

Note: The Tractatus AI Safety Framework source code is separately licensed under the Apache License 2.0. This Creative Commons licence applies to the research paper text and figures only.

— End of Document —

---

https://agenticgovernance.digital