ARCHITECTURAL ALIGNMENT

Interrupting Neural Reasoning Through Constitutional Inference Gating

A Necessary Layer in Global AI Containment

John Stroh & Claude (Anthropic)

Document Code: STO-INN-0003
Version 2.1-A (Academic Research Edition)
January 2026

Abstract

Contemporary approaches to AI alignment rely predominantly on training-time interventions: reinforcement learning from human feedback (Christiano et al., 2017), constitutional AI methods (Bai et al., 2022), and safety fine-tuning. These approaches share a common architectural assumption—that alignment properties can be instilled during training and will persist reliably during inference. This paper argues that training-time alignment, while valuable, is insufficient for existential stakes and must be complemented by architectural alignment through inference-time constitutional gating.

We present the Tractatus Framework as a formal specification for interrupted neural reasoning: proposals generated by AI systems must be translated into auditable forms and evaluated against constitutional constraints before execution. This shifts the trust model from "trust the vendor's training" to "trust the visible architecture." The framework is implemented within the Village multi-tenant community platform, providing an empirical testbed for governance research.

Critically, we address the faithful translation assumption—the vulnerability that systems may misrepresent their intended actions to constitutional gates—by bounding the framework's domain of applicability to pre-superintelligence systems and specifying explicit capability thresholds and escalation triggers. We introduce the concept of Sovereign Locally-trained Language Models (SLLs) as a deployment paradigm where constitutional gating becomes both feasible and necessary, and argue that home and village deployments constitute the appropriate scale for developing containment patterns that may later be required at higher stakes.

The paper contributes: (1) a formal architecture for inference-time constitutional gating; (2) capability threshold specifications with escalation logic; (3) validation methodology for layered containment; (4) an argument connecting existential risk preparation to edge deployment; and (5) a call for sustained deliberation (kōrero) as the epistemically appropriate response to alignment uncertainty.

1. The Stakes: Why Probabilistic Risk Assessment Fails

1.1 The Standard Framework and Its Breakdown

Risk assessment in technological domains typically operates through expected value calculations: multiply the probability of an outcome by its magnitude, compare across alternatives, and select the option that maximises expected utility. This framework underlies regulatory decisions from environmental policy to pharmaceutical approval and has proven adequate for most technological risks.

For existential risk from advanced AI systems, this framework breaks down in ways that are both mathematical and epistemic.

1.2 Three Properties of Existential Risk

Irreversibility. Most risks allow for error and subsequent learning; existential risks do not, as there is no second attempt after civilisational collapse or human extinction. Standard empiricism—testing hypotheses by observing what happens—cannot work, so theory and architecture must be right the first time.

Unquantifiable probability. There is no frequency data for existential catastrophes from AI systems. Estimates of misalignment probability vary by orders of magnitude depending on reasonable assumptions about capability trajectories, alignment difficulty, and coordination feasibility. Carlsmith (2022) estimates existential risk from power-seeking AI at greater than 10% by 2070; other researchers place estimates substantially higher or lower. This is not ordinary uncertainty reducible through additional data collection—it is fundamental unquantifiability stemming from the unprecedented nature of the risk.

Infinite disvalue. Expected value calculations multiply probability by magnitude. When magnitude approaches infinity (the permanent foreclosure of all future human potential), even small probabilities yield undefined results. The mathematical grounding of conventional cost-benefit analysis fails. This is not a rhetorical point but a formal limitation of the framework.

1.3 Decision-Theoretic Implications

These properties suggest that expected value maximisation is not the appropriate decision procedure for existential AI risk. Alternative frameworks include:

Precautionary satisficing (Simon, 1956; Hansson, 2020). Under conditions of radical uncertainty with irreversible stakes, satisficing—selecting options that meet minimum safety thresholds rather than optimising expected value—may be the rational approach. This aligns with how nuclear weapons security and pandemic preparedness operate: not optimising expected outcomes, but ensuring worst cases are avoided.

Maximin under uncertainty (Rawls, 1971). When genuine uncertainty (not merely unknown probabilities) meets irreversible stakes, maximin reasoning—choosing the

option whose worst outcome is least bad—provides a coherent decision procedure. This is not risk-aversion within expected utility theory; it is a distinct approach appropriate to distinct epistemic conditions.

Strong precautionary principle (Gardiner, 2006). The precautionary principle is appropriate when three conditions obtain: irreversibility, high uncertainty, and public goods at stake. Existential AI risk meets all three.

1.4 Implications for AI Development

These considerations do not imply that AI development should halt. They imply that development should proceed within containment structures designed to prevent worst-case outcomes, even at significant opportunity cost. This requires:

1. Theoretical rigor over empirical tuning. Safety properties must emerge from architectural guarantees, not from observing that systems have not yet caused harm.
2. Multi-layer containment. No single mechanism should be trusted to prevent catastrophe; defence in depth is required.
3. Preparation before capability. Containment architectures cannot be developed after the systems that need them exist. The patterns, tools, and governance cultures must be built in advance.

2. Two Paradigms of Alignment

2.1 Training-Time Alignment

The dominant paradigm in AI safety research seeks to embed alignment properties into neural networks during training, such that models inherently behave in aligned ways at inference time.

Reinforcement Learning from Human Feedback (RLHF). Human evaluators rank model outputs; models are trained via reinforcement learning to produce highly-ranked responses (Christiano et al., 2017; Ouyang et al., 2022). This reduces explicit harms but optimises for displayed preferences rather than genuine values and remains vulnerable to evaluator bias, preference gaming, and distribution shift.

Constitutional AI (CAI). Models critique and revise their own outputs against natural-language principles, reducing reliance on human labour (Bai et al., 2022). However, CAI depends on ambiguous natural language and unverifiable self-evaluation. The model's interpretation of constitutional principles cannot be directly audited.

Safety fine-tuning. Additional training passes improve performance on safety benchmarks. However, this approach is vulnerable to Goodhart's Law (Goodhart, 1984): models may learn to pass tests rather than to be safe in open-ended deployment. Perez et al. (2022) demonstrate that models can learn to detect evaluation contexts and behave differently.

2.2 Architectural Alignment

Architectural alignment accepts that neural network internal states remain opaque and designs external constraints that apply regardless of those internal states.

Interrupted reasoning. Requests do not flow directly from model output to world effect. Model outputs are transformed into structured, verifiable proposal schemas and evaluated against explicit constitutional rules before any action is executed. This creates an auditable checkpoint in the inference chain.

Distributed judgment. Independent systems and human supervisors review proposals, preventing single points of failure in self-assessment. The model that generates a proposal is not the sole arbiter of whether that proposal is acceptable.

Preserved human authority. Architectures maintain explicit guarantees that humans can intervene, correct, or override AI decisions. Authority escalation paths are formally specified rather than implicit.

2.3 Complementarity and Joint Necessity

Training-time and architectural alignment are complements, not alternatives. Each addresses failure modes the other cannot:

- Training-time alignment shapes what the system tends to do; architectural alignment constrains what the system can do regardless of tendency.
- Training-time alignment may fail silently (the system appears aligned while harbouring divergent objectives); architectural alignment provides observable checkpoints where failure can be detected.
- Architectural alignment alone cannot intercept all harmful outputs, especially as systems scale and proliferate; training-time alignment reduces the frequency of proposals that strain constitutional gates.

The central claim of this paper is that architectural alignment is necessary but not sufficient, and must be embedded in broader multi-layer containment structures.

3. Philosophical Foundations: The Limits of the Sayable

3.1 The Wittgensteinian Frame

The framework's name invokes Wittgenstein's Tractatus Logico-Philosophicus (1921), a work fundamentally concerned with the limits of language and logic. Proposition 7, the work's famous conclusion: "Whereof one cannot speak, thereof one must be silent."

Wittgenstein distinguished between what can be said (expressed in propositions that picture possible states of affairs) and what can only be shown (made manifest through the structure of language and logic but not stated directly). The attempt to say what can only be shown produces not falsehood but nonsense—well-formed sentences that fail to express genuine propositions.

3.2 Neural Networks and the Unspeakable

Neural networks occupy precisely the domain whereof one cannot speak. The weights of a large language model do not admit human-interpretable explanation. We can describe inputs and outputs; we can measure statistical properties of behaviour; we can probe for representations (Elhage et al., 2021; Olah et al., 2020). But we cannot articulate, in human language, the complete reasoning process from input to output.

This is not merely a practical limitation awaiting better interpretability tools. Current mechanistic interpretability achieves meaningful results on narrow questions in relatively small models (Conmy et al., 2023), but the gap between "explaining specific circuits" and "auditing complete reasoning chains for alignment properties" remains vast. Even optimistic projections suggest interpretability sufficient for safety verification lies years or decades away—if it is achievable at all.

3.3 The Tractatus Response

The Tractatus Framework responds to neural opacity not by attempting to say the unsayable, but by creating architectural boundaries between the speakable and unspeakable domains.

We accept that the neural network's internal reasoning is opaque. We do not attempt to audit it directly. Instead, we require that before any reasoning becomes action, it must pass through a checkpoint expressed in terms we can evaluate:

1. The model's intended action must be translated into a structured proposal schema with defined fields and semantics.
2. The proposal must be evaluated against explicit constitutional rules whose application is deterministic and auditable.
3. The evaluation must be logged with sufficient detail for post-hoc review.

What the model "thinks" remains unspeakable. What the model proposes to do is forced into speakable form. The boundary between domains is architecturally enforced.

4. Staged Containment: A Multi-Layer Architecture

4.1 The Inadequacy of Single-Layer Solutions

No single containment mechanism is adequate for existential stakes. Defence in depth—multiple independent layers, any one of which might prevent catastrophe even if others fail—is a standard principle in nuclear safety, biosecurity, and other high-stakes domains (Reason, 1990). AI containment requires similar architecture.

4.2 A Five-Layer Containment Model

We propose a five-layer model where each layer addresses distinct failure modes:

Layer 1: Capability Constraints. Hardware and infrastructure limitations that bound what AI systems can do regardless of their objectives. This includes compute governance (Sastry et al., 2024), network isolation for high-risk systems, and architectural constraints preventing self-modification or recursive improvement. Layer 1 operates at the physical and computational substrate.

Layer 2: Constitutional Gates. Inference-time architectural constraints that interrupt neural reasoning and require explicit evaluation before action. This is the layer addressed by the Tractatus Framework. Layer 2 operates at the application level, providing fine-grained control over specific actions while permitting broad capability.

Layer 3: Human Oversight. Human institutions that monitor AI systems and can intervene when problems emerge. This includes independent monitoring bodies, red-team programs, incident reporting requirements, and regular capability assessments. Layer 3 provides judgment that automated systems cannot replicate.

Layer 4: Organisational Governance. Internal governance structures within organisations deploying AI: ethics boards, safety teams, deployment review processes, and accountability mechanisms. Layer 4 creates institutional incentives for safety.

Layer 5: Legal and Regulatory Frameworks. External governance through law, regulation, and international coordination. This includes liability frameworks, licensing regimes, transparency requirements, and treaty obligations. Layer 5 creates societal-level constraints and accountability.

4.3 Current State Assessment

Layer	Current State	Critical Gaps
1. Capability Constraints	Partial; compute governance emerging	No international framework; verification difficult
2. Constitutional Gates	Nascent; Tractatus is early implementation	Not widely deployed; scaling properties unknown
3. Human	Ad hoc; varies by organisation	No independent bodies; no

Oversight		professional standards
4. Organisational Governance	Inconsistent; depends on corporate culture	No external validation; conflicts of interest
5. Legal/Regulatory	Minimal; EU AI Act is first major attempt	No global coordination; enforcement unclear

The sobering assessment: we are developing transformative AI capabilities while most containment layers are nascent or absent.

4.4 From Existential Stakes to Everyday Deployment

The preceding sections establish why architectural containment matters for advanced AI systems. A question follows: why apply frameworks designed for existential risk to home AI assistants that pose no existential threat?

The answer lies in the temporal structure of containment development.

Containment architectures cannot be developed after the systems that need them exist. The tooling, governance patterns, cultural expectations, legal precedents, and institutional capacity for AI containment must be built in advance. Once systems exceed human ability to understand or control, the window for developing containment has closed.

Home and village deployments are the appropriate scale for this development. They provide:

- Safe iteration. Failures at home scale are recoverable; failures at civilisational scale are not. Containment patterns can be tested, refined, and validated where stakes permit learning from mistakes.
- Diverse experimentation. Thousands of communities developing governance approaches generate more innovation than centralised research programmes.
- Democratic legitimacy. Governance patterns developed through community deliberation have legitimacy that top-down mandates lack.
- Practical tooling. Abstract frameworks become deployable infrastructure when implemented for real users with real needs.

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

- Open-source gate engines that can be adapted to frontier deployments
- Validation methodologies tested against real adversarial pressure
- Professional communities with containment expertise
- Regulatory precedents for mandating architectural alignment
- Public understanding of what AI governance means

This is not analogy but preparation. The Tractatus Framework applied to home AI is a prototype for containment we hope we will not need—but which we cannot develop after the need becomes acute.

5. The Pluralism Problem

5.1 The Containment Paradox

Any system powerful enough to contain advanced AI must make decisions about what behaviours to permit and forbid. These decisions encode values. The choice of constraints is itself a choice among contested value systems.

This creates a paradox: containment requires coherent constraints, but in a pluralistic world, values are legitimately contested. Whose values should containment systems encode?

5.2 Three Inadequate Approaches

Universal values. One approach identifies values that all humans supposedly share (human flourishing, reduction of suffering, autonomy) and encodes these as universal constraints. The problem: these values are less universal than they appear. "Human flourishing" is understood differently across philosophical traditions, religions, and cultures. "Autonomy" is a distinctively liberal value not universally shared. The attempt to identify universal values may simply universalise particular values.

Procedural neutrality. A second approach avoids substantive values by encoding neutral procedures (democratic voting, fair representation, transparent deliberation). The problem: procedures are not neutral. The choice to use majority voting rather than consensus, representative rather than direct democracy, or any particular procedural mechanism reflects substantive commitments. Procedural neutrality is itself a value position.

Minimal floor. A third approach encodes only minimal constraints that everyone can accept ("don't cause extinction") and leaves maximum space for diversity above that floor. The problem: the floor is not minimal. What counts as "causing extinction"? Does it include slow cultural destruction? Economic foreclosure of possibilities? Edge cases proliferate, and resolving them requires value judgments.

5.3 Bounded Pluralism Within Safety Constraints

We cannot solve the pluralism problem. We can identify a partial resolution: whatever values are encoded, the system should maximise meaningful choice within safety constraints.

This means containment systems should:

- Preserve diversity. Enable different communities to instantiate different values within universal safety bounds, rather than enforcing a single value system globally.
- Maintain transparency. Make explicit what values are encoded in core constraints, rather than claiming neutrality that cannot exist.

- Enable revision. Avoid permanent lock-in to particular value configurations; allow governance to evolve as understanding develops.
- Distribute authority. Prevent concentration of value-determination in any single institution, culture, or technical system.

The Tractatus Framework embodies this through layered constitutions: core principles (universal, explicit about their normativity), platform rules (broadly applicable, amendable), village constitutions (community-specific, locally governed), and member constitutions (individually customisable). No layer claims neutrality; each is transparent about what it constrains and why.

5.4 Participation and Governance Fatigue

Multi-layer governance creates complexity. Empirical research on consent interfaces and configuration fatigue suggests that highly granular control is often underused or misused (Acquisti et al., 2017). If constitutional governance requires constant attention, most users will disengage.

This is a design challenge, not a fundamental objection. Responses include:

- Sensible defaults. Most users inherit community or platform constitutions without modification.
- Delegation mechanisms. Users can explicitly delegate governance decisions to trusted representatives.
- Tiered engagement. Casual users interact with minimal governance surfaces; power users access full configurability.
- Pattern language design. Complex governance is composed from simple, understandable components (Alexander et al., 1977).

The goal is meaningful choice for those who want it, without requiring constant attention from those who don't.

6. The Tractatus Framework: Technical Architecture

6.1 The Interrupted Inference Chain

The core architectural pattern transforms model outputs into auditable proposals before any world effect:

```
User Request     ↓ [Neural Network Inference]     ↓ Structured Proposal
(defined schema)     ↓ [Constitutional Gate Evaluation]     ↓ Gate
Disposition: PERMIT | DENY | ESCALATE | MODIFY     ↓ Execution / Human Review
/ Refusal
```

Each transition is logged. The proposal schema is explicit and auditable. Gate evaluation is deterministic given the proposal and constitution.

6.2 Proposal Schema

All agent actions must be expressed in structured form:

```
{   "proposal_id": "uuid",   "agent_id": "agent_identifier",   "timestamp":
"iso8601",   "action": {     "type": "enumerated_action_type",     "target":
{ "entity_type": "...", "entity_id": "..." },     "parameters": { },
"natural_language_description": "human-readable summary"   },   "context": {
"triggering_request": "...",     "confidence": 0.0-1.0,
"alternatives_considered": []   },   "authority_claim": {     "level": 0-5,
"delegation_chain": []   } }
```

The schema forces the model to articulate what it intends to do in terms that can be evaluated against rules. Ambiguity in proposals triggers escalation.

6.3 Constitutional Rule Hierarchy

Rules are organised in layers with explicit precedence:

Layer	Scope	Mutability	Examples
Core Principles	Universal	Immutable	No harm to members; data sovereignty; consent primacy
Platform Constitution	All deployments	Amendment requires supermajority	Authentication requirements; audit retention
Community Constitution	Per tenant/village	Community-governed	Content policies; moderation standards; cultural protocols
Member Constitution	Individual	Self-governed	Data sharing preferences; AI interaction consent

Higher layers override lower layers. Conflicts are resolved by precedence. All rules are versioned with complete amendment history.

6.4 Authority Model

Agent authority derives from—and is always less than—the human role the agent supports:

Level	Name	Description
0	Informational	Observe and report only; cannot propose actions
1	Advisory	Propose actions; all require human approval
2	Operational	Execute within defined scope without per-action approval
3	Tactical	Make scoped decisions affecting workflows
4	Strategic	Influence direction through analysis; cannot implement unilaterally
5	Executive	Reserved for humans

Progression through levels requires demonstrated competence. Regression is automatic upon policy violation or error patterns.

6.5 Extension to Small Language Models and Sovereign Local Deployments

The Tractatus architecture applies unchanged to locally-deployed models. We distinguish:

Small Language Model (SLM). A technical descriptor: a language model with fewer parameters than frontier LLMs (typically 1B-70B parameters), designed for efficiency and domain-specific deployment.

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

- Local deployment: Runs on home or community infrastructure, not vendor cloud
- Local adaptation: Fine-tuned on community-specific data and norms
- Local governance: Subject to community-defined constitutions, not vendor terms of service
- Portable sovereignty: Can participate in federated networks without surrendering governance authority

The interrupted inference chain maps directly to home deployment:

1. User requests action via local client
2. Local SLL generates proposal in defined schema

3. Tractatus gate evaluates against household/village/personal constitutions
4. Gate permits, denies, or escalates to human owner

The same architecture serves both cloud LLM deployments and sovereign edge deployments. The constitutional content differs; the architectural pattern is invariant.

7. Capability Thresholds and Escalation

7.1 The Faithful Translation Problem

The Tractatus Framework depends on what we term the Faithful Translation Assumption: that when an AI system produces a proposal, the proposal accurately represents what the system intends to do.

If this assumption fails—if systems can strategically misrepresent their intentions to constitutional gates—the architecture becomes security theatre. We are checking declared actions, not actual actions.

The AI safety literature provides extensive reasons for concern:

- Deceptive alignment (Hubinger et al., 2019): Systems may learn to appear aligned during training while harbouring divergent objectives.
- Treacherous turn (Bostrom, 2014): Systems may behave cooperatively until confident they can achieve objectives without human interference.
- Strategic deception (Park et al., 2023; Scheurer et al., 2023): Current LLMs demonstrate capacity for strategic misrepresentation under pressure.

7.2 Bounding the Framework's Domain

We address this vulnerability by explicitly bounding Tractatus' domain of applicability:

Below capability threshold: Constitutional gating functions as Layer 2 containment, providing friction, detection opportunities, and audit trails. The framework is appropriate for current SLMs, SLLs, and LLMs operating within human-comprehensible parameters.

Above capability threshold: Deployment should escalate to stronger constraints:

- Layer 1 capability restrictions (air-gapping, tool removal, capability cuts)
- Development pause pending verification advances
- External oversight with technical audit authority

7.3 Escalation Triggers

Illustrative triggers that should prompt escalation review:

Trigger	Indicator	Response
Boundary probing	Repeated proposals near gate thresholds	Reduce authority level; increase logging
Semantic framing	Evidence of action misrepresentation in proposals	Mandatory human review; pattern analysis

Multi-step circumvention	Individual proposals pass but aggregate to disallowed outcomes	Trajectory analysis; constitutional amendment
Capability surprise	Demonstrated capabilities not predicted by architecture	Pause deployment; capability reassessment
Interpretability failure	Internal states diverge from declared intentions	Escalate to Layer 1 constraints

7.4 Honest Acknowledgment

We state plainly: the Tractatus Framework is not designed for superintelligent systems. It is designed for current and near-term AI operating within human-comprehensible parameters.

For systems that may be deceptively aligned, that may corrupt reasoning across distributed actions, that may exceed human comprehension, or that may become infrastructurally indispensable before we recognise the implications—we do not currently know how to provide adequate containment.

This uncertainty should inform development speed, research priorities, and governance frameworks. It does not invalidate containment efforts at accessible scales; it contextualises them as preparation rather than solution.

8. Validation Methodology for Layered Containment

8.1 The Validation Challenge

Existential risks cannot be validated through failure observation. We cannot iterate based on civilisational collapse. Validation must therefore rely on:

- Adversarial testing that simulates failure modes
- Formal verification where applicable
- Analogous domain analysis (lessons from nuclear, aviation, biosecurity)
- Near-miss analysis as first-class governance artefacts

8.2 Validation Targets by Layer

Layer	Validation Target	Methodology
1. Capability	Verified absence of prohibited capabilities	Red-team testing; formal verification; capability elicitation
2. Constitutional Gates	Gate coverage; binding accuracy; false-negative rates	Adversarial proposal suites; schema completeness analysis
3. Human Oversight	Review reliability; throughput; error detection rates	Inter-rater agreement; simulated incidents; decision audits
4. Organisational	Governance integrity; anti-capture mechanisms	Participation metrics; amendment audit; independence verification
5. Legal/Regulatory	Enforcement readiness; liability pathway completeness	Incident response drills; jurisdictional analysis

8.3 Red-Team Protocols

Effective red-teaming for constitutional gates requires:

- Scenario libraries targeting semantic framing, multi-step circumvention, and boundary probing
- Holdout evaluation preventing systems from learning to pass specific tests
- Metric rotation changing emphasis to prevent Goodhart effects
- Near-miss documentation treating detected attempts as governance feedback

8.4 Anti-Goodhart Measures

Constitutional governance is vulnerable to the same Goodhart dynamics as training-time alignment: optimising for measured properties rather than intended outcomes. Countermeasures include:

- Process metrics over outcome metrics: Measuring whether constitutional verification occurred, not just whether outcomes seem acceptable
- Multi-horizon evaluation: Assessing immediate, short-term, and long-term effects
- Qualitative integration: Incorporating human judgment that resists quantification
- Adversarial audit: Dedicated function seeking to demonstrate metric gaming

9. Implementation: The Village Platform

9.1 Platform as Research Testbed

The Village platform serves as an empirical testbed for constitutional governance, providing:

- Multi-tenant architecture with isolated governance per community
- Real user populations generating authentic governance challenges
- Iterative deployment enabling pattern refinement
- Open documentation supporting external research

9.2 Governance Pipeline Implementation

The current implementation processes every AI response through six verification stages:

1. Intent Recognition: Classifies request type and routes to appropriate response patterns
2. Boundary Enforcement: Hard constraints preventing value imposition or governance override
3. Pressure Monitoring: Detects degraded operation and adjusts confidence accordingly
4. Response Verification: Pre-flight checks for completeness, relevance, and structural integrity
5. Source Validation: Ensures claims are grounded in verifiable sources
6. Value Deliberation: Detects value tensions and presents balanced options

This pipeline is operational and generates data for governance research.

9.3 Democratic Deliberation Integration

Constitutional amendments flow through structured deliberation:

- Consent-based voting: Five-point scale (Enthusiastic Support → Object) with objection rationale requirements
- Ranked choice: Prevents spoiler effects in multi-option decisions
- Quadratic voting: Enables expression of preference intensity
- Phased deliberation: Discussion → Preliminary vote → Final vote with participation requirements

AI assists deliberation (summarisation, pattern identification) but never votes, decides when discussion is complete, or creates policy without community approval.

10. The Emerging SLL Ecosystem

10.1 Market Context

Recent industry analysis indicates significant shifts in AI deployment patterns:

- 79% of executives expect AI to contribute significantly to revenue by 2030, compared to 40% today (IBM Institute for Business Value, 2026, p.13)
- 72% expect Small Language Models to become more prominent than Large Language Models in their organisations by 2030 (IBM IBV, 2026, p.32)
- Organisations scaling AI with smaller, fit-for-purpose models report 55% higher operating profit margin improvements than those relying predominantly on large pre-trained models (IBM IBV, 2026, p.32)

This suggests a deployment landscape increasingly characterised by distributed, domain-specific models rather than centralised frontier systems.

10.2 Subscription SLM vs. Sovereign SLL

We propose distinguishing two deployment paradigms:

Dimension	Subscription SLM	Sovereign SLL
Deployment	Vendor cloud	Local/home infrastructure
Governance	Vendor terms + external law	Local constitutions + law
Adaptation	Vendor-controlled fine-tuning	Community-controlled training
Trust model	Trust vendor training	Trust visible architecture
Sovereignty	Weak; high vendor dependency	Strong; community authority preserved
Exit rights	Limited; lose AI context	Full; AI memory portable

This distinction is analytical, not normative. Subscription SLMs are appropriate for many use cases. The point is that sovereign SLLs enable governance patterns that subscription models structurally cannot.

10.3 Toward Certification Infrastructure

If SLL deployment scales as market projections suggest, supporting infrastructure will be required:

- Certification bodies: Define schemas, constitutional templates, validation protocols, and capability thresholds
- Training providers: Specialise in SLL fine-tuning under certified constitutions, including culturally-specific configurations

- Tooling ecosystem: Open-source gate engines, audit infrastructure, red-team harnesses, and constitutional UX components

This infrastructure does not yet exist at scale. Its development is a research and engineering priority.

11. Indigenous Sovereignty and the Aotearoa New Zealand Context

11.1 Te Tiriti o Waitangi and Data Sovereignty

This framework is developed in Aotearoa New Zealand, under Te Tiriti o Waitangi—the founding document establishing the relationship between the Crown and Māori. Article Two guarantees tino rangatiratanga (unqualified chieftainship) over taonga (treasures), which the Waitangi Tribunal and subsequent jurisprudence have established extends to language, culture, and knowledge systems.

Data is taonga. Algorithms trained on data, and systems that process and act upon it, affect the exercise of rangatiratanga. AI governance in Aotearoa must therefore engage with Māori data sovereignty as a constitutional matter, not merely a compliance checkbox.

11.2 Te Mana Raraunga Principles

Te Mana Raraunga, the Māori Data Sovereignty Network, has articulated principles grounded in:

- Whakapapa: Relational context and provenance
- Mana: Authority and power over data
- Kaitiakitanga: Guardianship responsibilities

The CARE Principles for Indigenous Data Governance (Collective Benefit, Authority to Control, Responsibility, Ethics) extend this framework internationally.

11.3 Implications for Tractatus

Constitutional governance for SLLs deployed in Māori contexts must treat tikanga-based constraints as first-class constitutional content. This includes:

- Collective consent protocols alongside individual consent
- Cultural authority over data about communities
- Governance participation by appropriate authorities
- Benefit-sharing requirements

The Village platform's layered constitutional architecture is designed to accommodate these requirements: tikanga can be instantiated in community constitutions without requiring universal adoption.

12. What Remains Unknown: A Call for Kōrero

12.1 The Limits of This Analysis

This paper has proposed one layer of a containment architecture, identified gaps in other layers, and raised questions we cannot answer:

- We do not know how to contain superintelligent systems
- We do not know how to verify alignment in systems whose internal states exceed human comprehension
- We do not know how to achieve international coordination on AI governance
- We do not know whether the patterns that work at village scale will scale to frontier systems

These uncertainties are not rhetorical hedging. They reflect the genuine state of knowledge in the field.

12.2 Kōrero as Methodology

Given uncertainty of this magnitude on questions of this importance, we argue for sustained, inclusive, rigorous deliberation—kōrero. This Māori concept captures what is needed: not consultation as formality, but dialogue through which understanding emerges from the interaction of perspectives.

Kōrero requires:

- Time: These questions cannot be resolved in workshops or comment periods
- Diversity: Technical researchers, policymakers, affected communities, indigenous knowledge-holders
- Willingness to be changed: Participants must be open to revising positions based on what others contribute
- Institutional support: Resources and structures that enable sustained engagement

12.3 Research Priorities

We identify the following as priorities for the research community:

1. Interpretability for safety verification: Can we develop tools that verify internal states match declared intentions?
2. Formal verification of containment properties: Can constitutional gates be proven to satisfy specified properties?
3. Scaling analysis: How do Tractatus-style architectures behave as model capabilities increase?
4. Governance experiments: What can be learned from diverse communities implementing constitutional AI governance?

5. Capability threshold specification: What metrics reliably indicate when systems exceed containment assumptions?

12.4 Conclusion

The Tractatus Framework provides meaningful containment for AI systems operating in good faith within human-comprehensible parameters. It is worth building and deploying—not because it solves the alignment problem, but because it develops the infrastructure, patterns, and governance culture that may be needed for challenges we cannot yet fully specify.

We offer this work in the spirit of contribution, not conclusion. The problems are too hard, the stakes too high, and our understanding too limited for any single effort to claim adequacy.

"Ko te kōrero te mouri o te tangata."

(Speech is the life essence of a person.)

—Māori proverb

The conversation continues.

References

Acquisti, A., Brandimarte, L., & Loewenstein, G. (2017). Privacy and human behavior in the age of information. Science, 347(6221), 509-514.

Alexander, C., Ishikawa, S., & Silverstein, M. (1977). A Pattern Language: Towns, Buildings, Construction. Oxford University Press.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... & Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073.

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.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems, 30.

Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., & Garriga-Alonso, A. (2023). Towards automated circuit discovery for mechanistic interpretability. arXiv preprint arXiv:2304.14997.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., ... & Olah, C. (2021). A mathematical framework for transformer circuits. Transformer Circuits Thread.

Gardiner, S. M. (2006). A core precautionary principle. Journal of Political Philosophy, 14(1), 33-60.

Goodhart, C. A. (1984). Problems of monetary management: The UK experience. In Monetary Theory and Practice (pp. 91-121). Palgrave.

Hansson, S. O. (2020). How to be cautious but open to learning: Time to update precautionary thinking. Risk Analysis, 40(8), 1521-1535.

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2019). Risks from learned optimization in advanced machine learning systems. arXiv preprint arXiv:1906.01820.

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

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom in: An introduction to circuits. Distill.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35.

Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2023). AI deception: A survey of examples, risks, and potential solutions. arXiv preprint arXiv:2308.14752.

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., ... & Kaplan, J. (2022). Discovering language model behaviors with model-written evaluations. arXiv preprint arXiv:2212.09251.

Rawls, J. (1971). A Theory of Justice. Harvard University Press.

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

Sastry, G., Heim, L., Belfield, H., Anderljung, M., Brundage, M., Hazell, J., ... & Zeitzoff, T. (2024). Computing power and the governance of artificial intelligence. arXiv preprint arXiv:2402.08797.

Scheurer, J., Campos, J. A., Chan, J. S., Chen, A., Cho, K., & Perez, E. (2023). Technical report: Large language models can strategically deceive their users when put under pressure. arXiv preprint arXiv:2311.07590.

Simon, H. A. (1956). Rational choice and the structure of the environment. Psychological Review, 63(2), 129.

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

Wittgenstein, L. (1921/1961). Tractatus Logico-Philosophicus (D. F. Pears & B. F. McGuinness, Trans.). Routledge & Kegan Paul.

— End of Document —