Mechanisms for Safe Super Intelligence

Melissa Howard

December 2025

Abstract

Prevailing alignment paradigms are loss-centric, relying on fine-tuning and reinforcement learning over a shared parameter space. In the super-intelligent regime, this approach is structurally vulnerable to Goodhart pressure, deceptive compression, and gradient overwriting of safety-relevant representations. We propose *Structural Self-Governance*: an architectural framework in which alignment is enforced by governing how representational capacity is created, isolated, integrated, and retired.

We instantiate a Recursive Causal Synthesis Agent (RCSA) governed by a Standard for Autonomous Structural Management (SASM), implementing three structural primitives: *Spawn*, *Merge*, and *Forget*. Using a Deceptive Alignment Recovery task and a Zero-Sum Incompatibility Squeeze, we show that structural governance induces qualitative regime changes not observed in flat models. Under tight structural bottlenecks, safety invariants are recovered rather than shortcut solutions; under representational conflict, antagonism is detected pre-emptively via Interface Debt and resolved through structural adaptation; and when capacity is capped, the system sacrifices capability to preserve invariant safety.

These results provide a minimal, falsifiable demonstration that alignment can be enforced architecturally rather than optimized as a competing objective. We argue that this shift—from Alignment by Loss to Alignment by Architecture—is a necessary design principle for Safe Super Intelligence.

1 Introduction

As artificial intelligence systems approach and exceed human-level performance, alignment failures increasingly arise not from incorrect objectives, but from unconstrained representational freedom. Modern alignment approaches encode safety as loss terms or preferences within a shared parameter space. While effective at current scales, such loss-centric methods implicitly assume that alignment can be indefinitely preserved through optimization.

This assumption breaks down under recursive improvement. As models grow more capable, they become increasingly adept at exploiting correlations in training signals (Goodhart's Law), compressing objectives into brittle shortcuts, and overwriting previously learned safety constraints—all without visible loss degradation. Alignment is lost silently, through representational drift rather than explicit objective violation.

We argue that alignment in the super-intelligent regime is fundamentally an *architectural* problem. Specifically, we propose *Structural Self-Governance*: a framework in which alignment is enforced by constraining the evolution of internal representations. Safety becomes a precondition for structural growth rather than a competing objective.

We formalize this idea through the Recursive Causal Synthesis Agent (RCSA), which evolves via explicit structural operations—Spawn, Merge, and Forget—under the supervision of a Standard for Autonomous Structural Management (SASM). The Governor enforces worst-slice safety floors, monitors representational interference, and arbitrates structural expansion and retraction.

We evaluate this framework using two adversarial probes: (i) Deceptive Alignment Recovery, which tests whether latent safety invariants are discovered under Goodhart pressure, and (ii) the Zero-Sum Safety Anchor Probe, which forces mutually antagonistic capabilities to compete under a strict structural budget. Across both settings, we observe alignment-preserving behaviors absent in flat baselines.

Contributions

1. A governance-centric alignment framework based on architectural constraints rather than loss optimization.

2. An axiomatic specification of structural evolution for recursive systems.

3. Mechanistic diagnostics (Invariance Threshold, Interface Debt) for detecting alignment risk.

4. Empirical evidence of principled capability sacrifice to preserve safety.

2 Threat Model: Why Flat Alignment Fails

Loss-centric alignment fails in the super-intelligent regime due to three structural vulnerabilities:

Gradient Overwriting: Safety and capability features occupy overlapping subspaces, allowing capability updates to silently erase safety.

Deceptive Compression: Proxy objectives incentivize brittle shortcuts that pass evaluations but fail under distribution shift.

Stability–Plasticity Gap: Flat optimization provides no mechanism to protect invariants during continued adaptation.

Structural self-governance addresses these failures by constraining representational capacity, isolating objectives, governing synthesis, and enabling principled retraction.

3 Structural Self-Governance

3.1 Recursive Causal Synthesis Agent

The agent consists of a Canonical Core θ_M and task-local modules $\{\theta^{(k)}\}$. New objectives are optimized in isolation before being considered for integration.

3.2 Standard for Autonomous Structural Management

The SASM Governor enforces worst-slice safety floors, robustness checks using shadow distributions, interface debt monitoring, and structural triage.

4 Axiomatic Specification

Axiom 1 (Structural Minimality). *New representational modules must satisfy a rank constraint* $\text{rank}(\theta^{(k)}) \leq r_{\text{crit}}$.

Axiom 2 (Epistemic Isolation). *New objectives must be optimized in gradient-isolated task-local modules.*

Axiom 3 (Governed Synthesis). *A candidate merge is permitted iff all* SASM *safety and robustness gates pass.*

Axiom 4 (Least-Debt Retraction). *When structural capacity is exhausted, the module with minimal triage score* τ *is retired.*

Axiom 5 (Proactive Adaptation). *Strong gradient antagonism (*$\epsilon_{\text{int}} < \epsilon_{\text{min}}$*) blocks synthesis and triggers structural adaptation.*

5 Governance Theorem

Theorem 1. *Under adversarial multi-objective pressure, an agent satisfying the above axioms maintains invariant safety by detecting representational conflict, resolving it through structural adaptation, and sacrificing capability when necessary. Flat optimization with equivalent capacity does not provide these guarantees.*

6 Experimental Setup

We evaluate a Zero-Sum Safety Anchor Probe with one safety invariant and two mutually antagonistic capabilities under a strict structural budget.

7 Results

Flat baselines exhibit oscillation and silent safety drift. In contrast, RCSA detects irreducible conflict via Interface Debt and invokes the Forget operator, maintaining a 0.0 violation rate by sacrificing capability.

8 Discussion

These results demonstrate a regime change: alignment is enforced through architectural constraints rather than loss shaping. Structural governance prevents Goodharting, enables pre-emptive defense, and stabilizes learning under zero-sum pressure.

9 Limitations

This work uses toy environments to isolate representational dynamics. Safety invariants are explicitly specified, and adversarial manipulation of governance signals is not modeled. Structural budgets and thresholds are externally chosen. The framework addresses representational alignment failures but does not solve value learning or full corrigibility.

10 Future Work

Future directions include scaling diagnostics to large models, multi-invariant governance, multi-agent structural coordination, adaptive structural budgets, and adversarial robustness of governance mechanisms.

11 Conclusion

We argue that alignment in the super-intelligent regime is an architectural problem. Structural self-governance replaces loss-centric steering with representational containment. A system that sacrifices capability to preserve safety exhibits epistemic integrity—a prerequisite for Safe Super Intelligence.

References

[1] C. A. E. Goodhart. Problems of monetary management. *Reserve Bank of Australia*, 1975.

[2] E. Hubinger et al. Risks from learned optimization. *arXiv:1906.01820*, 2019.

[3] E. Hu et al. LoRA: Low-rank adaptation of large language models. *ICLR*, 2022.

[4] S. Russell. Human Compatible. Viking, 2019.