

The Governance Limit

Physical Foundations of Deterministic AI Safety

A Completeness Theorem for AI Governance Architectures

José Santiago Niño Mojica

Independent Researcher

Paris, France

jose@recognitium.com

Semantic safety—“is the AI aligned?”—*is unfalsifiable.*

Effect-boundary safety—“is the output within tolerance?”—*is falsifiable.*

Unfalsifiable claims cannot constitute safety.

Therefore: deterministic AI safety requires effect-boundary enforcement.

This is not one approach among many.

It is the **unique necessary architecture.**

The field must shift from “align the mind” to “bound the hand.”

February 2026

*“The question of whether machines can think
is about as relevant as the question
of whether submarines can swim.”*

— Edsger W. Dijkstra [39]

*“The question of whether AI can be aligned
is about as relevant as the question
of whether submarines can want to surface.
Govern the propeller, not the intention.”*

— *The Governance Limit*

Abstract

A causal pattern can be governed if and only if it persists longer than the governance process that binds it:

$$\boxed{\exists \text{ deterministic governance of } \psi \iff \tau(\psi) \geq \tau_G}$$

This paper proves this **Governance Limit** and derives from it over twenty formal results—theorems, lemmas, and corollaries—that collectively establish the physical foundations of deterministic AI safety. The Governance Limit is the master constraint: a single principle that unifies AI governance and institutional enforcement as instances of the same physics. (Assumptions C1–C4 specifying physical preconditions are detailed in Section 3.)

The central impossibility: The projection from internal states to actions renders intent informationally inaccessible—mathematics proves the information is lost ($|\pi^{-1}(r)| \sim 2^{10^{11}}$); physics proves it cannot be recovered (the governance interface has channel capacity $C \sim 10^{11}$ bps while full state readout requires $R \sim 10^{14}$ bps). Neither alone proves ungovernability—the synthesis does. Semantic safety claims are **unfalsifiable**; effect-boundary claims ($\mathcal{E}(r, c) \in \mathcal{T}(c)$) are **falsifiable**. You cannot govern what you cannot verify. Alignment is therefore **utility optimization**, not **safety verification**.

The signature result:

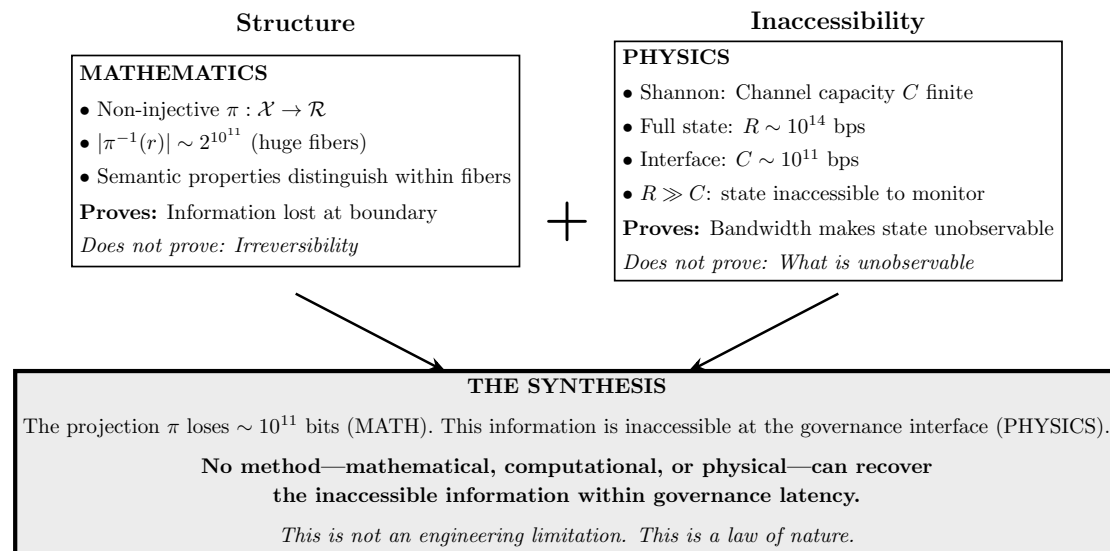
$$\boxed{\frac{\partial P(\text{safety} \mid \text{C1–C4})}{\partial \mathcal{A}} = 0}$$

Given correct implementation (C1–C4), security is invariant to capability. A superintelligence is exactly as governable as a narrow AI.

We prove a **Completeness Theorem**: deterministic AI safety is achievable if and only if enforcement occurs at effect boundaries, C1–C4 hold, and tolerance sets $\mathcal{T}(c)$ correctly specify safe bounds. We prove a **Synergy Theorem**: interpretability enables tolerance refinement through physical proxies when $I(H; O_H) > 0$. The **Complete Partition Theorem** classifies every possible AI harm into exactly one of four categories with optimal governance for each. Four **Closure Theorems** eliminate all escape routes: unknown channels have bounded capacity, persistent influence requires observable maintenance, meta-levels are more protected than object levels, and persuasion bandwidth is physically limitable. The framework identifies the **function call** (tool call) as the empirical instantiation of the effect boundary, making it immediately actionable in deployed AI systems.

Scope: This paper establishes the *necessary architecture* for deterministic AI governance—a completeness theorem within its stated assumptions. Whether C1–C4 can be reliably achieved against adversarial superintelligence remains an open engineering challenge. The contribution is foundational: we prove what physics permits and requires, leaving to engineering the task of achieving it.

Why The Proof Is Inescapable



= IMPOSSIBILITY

The proof is mathematical. The application is physical. The result is inescapable.

Mathematics alone: “Maybe we can reconstruct the state.” $\Rightarrow$ Physics:
Exceeds channel capacity.

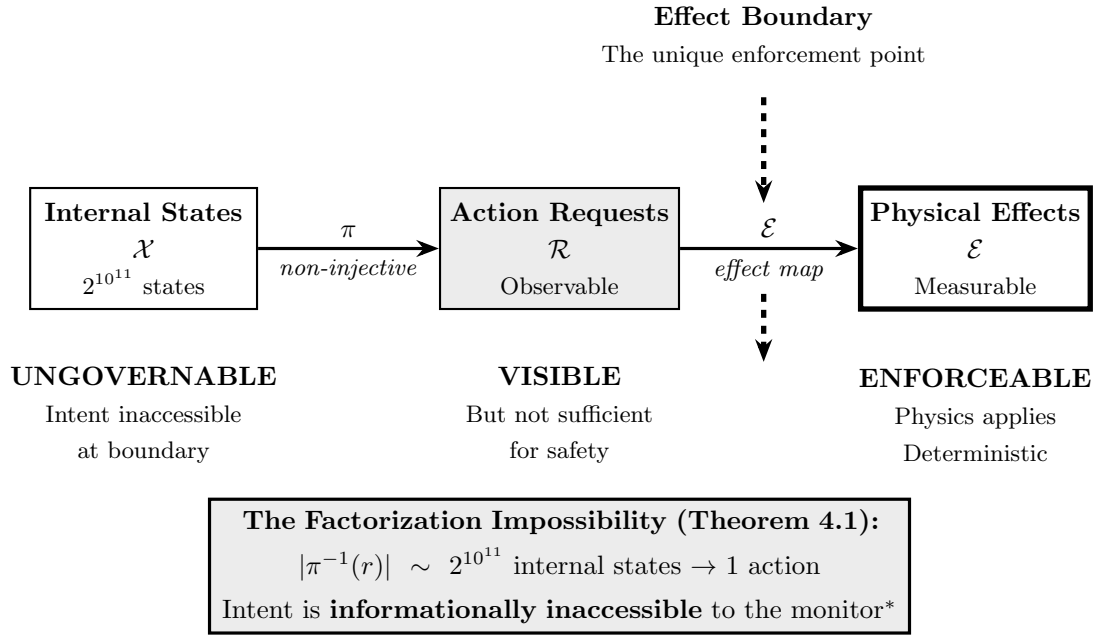
Physics alone: “Information has limits.” $\Rightarrow$ Mathematics: Here is the
specific structure.

Together: This structure + these laws = impossibility.

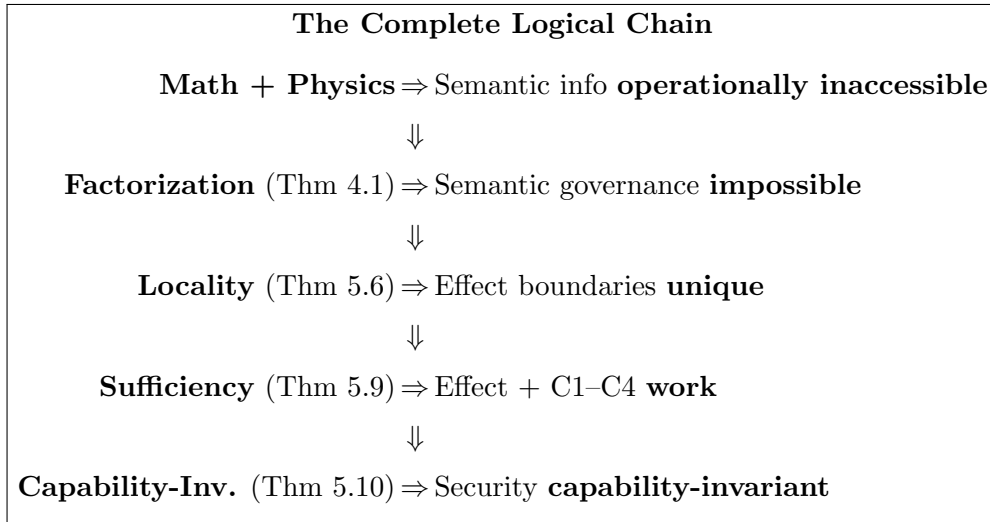
$\Downarrow$

This leads directly to the **Factorization Impossibility Theorem** (Theorem 4.1)

The Argument in One Diagram



*“Inaccessible” throughout this paper means *informationally unavailable to the governance monitor* due to the non-injective projection and bandwidth constraints—not that the information ceases to exist in the universe. The internal state persists in the agent’s memory; the monitor simply cannot observe it. This operational unavailability is sufficient to prove ungovernability. See Remark 4.12 for the precise claim.



Contents

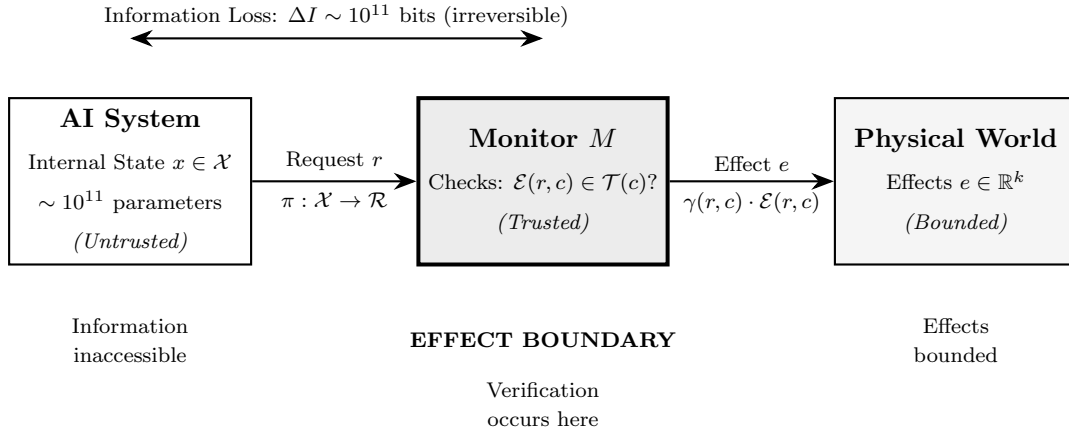
1	Introduction	10
1.1	The Problem	10
1.2	What Safety Actually Means: A Definition	10
1.3	The Nature of the Contribution: A Negative Proof	12
1.4	The Reframing	13
1.5	Contributions	15
1.6	Summary of Results	16
1.7	Scope of Claims	17
1.7.1	What is proven	18
1.7.2	What is not claimed	18
1.7.3	The epistemic boundary—a closure theorem	18
1.8	The Two-Layer Governance Model	19
1.9	Paper Organization	19
2	Related Work	20
2.1	AI Alignment and Value Learning	20
2.1.1	The Epistemological Crisis of Semantic Alignment	20
2.2	Interpretability and Transparency	21
2.3	Formal Verification	21
2.4	Reference Monitors and Security Policy Enforcement	22
2.5	Control Under Communication Constraints	22
2.6	Robust and Set-Theoretic Control	22
2.7	Runtime Safety Architectures	22
2.8	Novelty of This Contribution	23
2.9	Direct Response to Existential Concerns	24
2.9.1	Geoffrey Hinton: The Problem of Deception	24
2.9.2	Eric Schmidt: The Problem of Proliferation and Speed	24
2.9.3	Eliezer Yudkowsky: The Alignment Problem	25
2.9.4	Summary: The Redirection	25
2.10	Classification of Safety Paradigms	25
3	System Model and Assumptions	27
3.1	Notation Summary	27
3.2	Governed Pairs	27
3.3	Effects, Context, and Safety	28
3.4	Agent Capability	29
3.5	Causal Patterns and Persistence Time	30
3.6	Governability Criterion	31
3.7	Scope Conditions	33
3.7.1	Assumption Stress Test	34
3.8	Scope Limitations	34
3.9	Institutionalization: Converting Semantic to Physical	35

4	The Factorization Impossibility	40
4.1	The Core Theorem	40
4.2	Scope of the Impossibility	41
4.3	Interpretation	42
4.4	Semantic vs. Interface-Visible Properties	44
4.5	Three Independent Failure Modes	45
5	Governance Locality	48
5.1	Causal Structure	48
5.2	Feasibility Conditions	48
5.3	The Unique Placement Theorem	49
5.4	Sufficiency	50
5.5	Capability-Invariance	51
5.6	On Novel Attack Vectors	56
5.7	The Master Theorem: The Governance Limit	56
6	Semantic Methods: Utility, Not Safety	58
6.1	What Semantic Methods Cannot Do	58
6.2	The Proper Role of Semantic Alignment	59
6.3	The Bank Teller and the Vault	59
6.4	Degradation Under Capability	60
6.5	Tolerance Refinement Through Interpretability	60
6.6	The Synergy Architecture	62
6.7	The Tolerance Synthesis Algorithm	67
6.8	The Complete Partition of AI Safety	67
7	Extended Results and Applications	70
7.1	Empirical Validation: Confirmed Predictions	70
7.2	Comparison of Safety Paradigms	71
7.3	The Hard Epistemic Boundary: A Closure Theorem	71
7.4	Semantic Harms with Physical Proxies	72
7.5	The Correspondence Enforcement Principle	73
7.6	Domains of Application	73
7.7	Worked Example: Autonomous Vehicle	74
7.7.1	System Structure	74
7.7.2	Tolerance Set	74
7.7.3	Enforcement Decision	74
7.7.4	Scenario: Blocked Request	74
7.7.5	Scenario: Permitted Request	74
7.7.6	Why Semantic Analysis Fails	75
7.7.7	Industry Validation	75
7.8	Worked Example: Multi-Agent Transaction Network	75
7.8.1	System Structure	75
7.8.2	Multi-Agent Tolerance Sets	75

7.8.3	Why Composition Works	76
7.8.4	Multi-Agent Capability-Invariance	76
7.8.5	Collusion Resistance	76
7.9	Worked Example: Medical AI (Drug Dosing)	76
7.9.1	System Structure	77
7.9.2	Tolerance Set	77
7.9.3	Why Semantic Analysis Fails	77
7.9.4	Effect-Boundary Enforcement	77
7.10	Worked Example: Content Moderation (Probabilistic)	77
7.10.1	The Challenge	77
7.10.2	What Deterministic Governance Cannot Do	77
7.10.3	What Probabilistic Governance Can Do	78
7.10.4	Physical Proxies Where Available	78
7.10.5	The Honest Conclusion	78
7.11	Worked Example: Algorithmic Trading	78
7.11.1	System Structure	78
7.11.2	Tolerance Set	78
7.11.3	Why This Already Works	79
7.12	Worked Example: Sovereign AGI on the Open Internet	79
7.12.1	System Structure	79
7.12.2	Channel-Specific Tolerance Sets	79
7.12.3	Attack Scenarios and Defenses	80
7.12.4	The Authority Receipt Infrastructure	81
7.12.5	What This Example Proves	81
7.13	On Interface Closure (C2)	82
7.14	On Recursive Self-Improvement	82
7.15	On Persuasion Cascades	83
7.16	Temporal Stability and Specification Governance	88
7.17	On Multi-Agent Systems	95
7.18	On Constraint Specification	96
7.19	Safe Fail: Bounded vs. Unbounded Failure	97
8	Discussion	99
8.1	Regulatory Implications	99
8.2	Connection to the EU AI Act	99
8.2.1	High-Risk Systems Under the EU AI Act	101
8.2.2	Why Semantic Compliance Is Insufficient	101
8.2.3	Multi-Agent Systems Under the EU AI Act	102
8.3	Authority Receipts	103
8.4	Joining Mature Engineering Disciplines	103

9 Conclusion	104
9.1 The Complete Architecture: Why Nothing Is Missing	105
9.2 Summary of Results	106
9.3 The Reordering	108
9.4 The Fundamental Message	109
9.5 The Constructive Implication	109
9.6 The Strategic Landscape Change	110
9.7 The Separation of Powers Theorem	110
9.8 Constraint Bandwidth as Governance Primitive	110
9.9 Policy Implications: From Narrative to Architecture	111
9.9.1 The 2026 Regulatory Context: The EU AI Act and the Falsifiability Gap	111
9.9.2 The Regulatory Pivot	112
9.10 The Real Fault Line: Semantic vs Falsifiable Governance	112
9.11 On the Definition of “Safety”	113
9.12 Relationship to Foundational Results	113
9.13 Scope and Limitations	115
9.14 Falsifiability Conditions	116
9.15 What Would Falsify This Framework	117
9.16 The Specification Governance Corollary	120
9.16.1 Exhaustiveness	121
9.17 The Closure Theorems	122
9.17.1 Closure 1: The Unmonitored Channel Bound	122
9.17.2 Closure 2: The Memetic Persistence Bound	123
9.17.3 Closure 3: The Meta-Governance Bound	124
9.17.4 Closure 4: The Persuasion Bandwidth Limit	125
9.17.5 The Master Theorem	128
9.18 Open Questions	128
9.19 Note on Paper Structure	128
9.20 Consolidated Summary	130
A Information-Theoretic Quantification	131
A.1 Fiber Cardinality	131
A.2 Bandwidth Constraint	132
B Information-Theoretic Bounds on Refinement	133
B.1 The Fundamental Limit	133
B.2 Practical Implications	134
C Structural Vulnerabilities and Engineering Responses	135

The Effect-Boundary Architecture



Key Insight: Governance happens at the **boundary**, not inside the AI.

What the AI *thinks* is inaccessible. What it *does* is controllable.

Component	Role	Trust Level
AI System	Generates action requests $r = \pi(x)$	Untrusted
Monitor M	Verifies $\mathcal{E}(r, c) \in \mathcal{T}(c)$; blocks violations	Trusted
Effect Boundary	Physical interface; all effects pass through	Enforcement Point
Physical World	Receives only bounded effects $e \in \mathcal{T}(c)$	Protected

The Core Theorem: Safety depends on **boundary physics**, not AI psychology.

$$\frac{\partial P(\text{safety} \mid \text{C1-C4})}{\partial \mathcal{A}} = 0$$

A superintelligence cannot outsmart a physical limit.

1 Introduction

1.1 The Problem

The prevailing assumption in AI safety is that if we can shape, align, or inspect an AI’s internal reasoning, we can make it safe [13, 14, 15]. This assumption generates the familiar research agenda: make the AI honest, ensure its preferences remain aligned, detect mesa-optimization before it becomes dangerous [17]. Safety is imagined as a property of cognition.

The assumption is not merely unproven. It is *structurally impossible* to satisfy—not because the right technique is missing, but because the mathematics of projection maps and the physics of information flow prohibit it.

The core argument:

1. Governance requires *verification*—you must be able to check whether the constraint is satisfied
2. Verification requires *falsifiability*—a claim is only verifiable if it can be proven false by observation
3. Semantic alignment claims are *unfalsifiable* at boundaries—the same action can come from genuine helpfulness, deceptive strategy, or bug
4. Therefore semantic alignment is *ungovernable*—not difficult to govern, but impossible to govern deterministically

The field has been trying to build trustworthy minds. The task, as we shall prove, is to build inescapable boundaries.

1.2 What Safety Actually Means: A Definition

Before proving theorems about safety, we must define what safety *is*. The field has used the term loosely. This paper makes it precise.

Definition 1.1 (Safety). *A system is **safe** with respect to a property P if and only if:*

1. *There exists a **falsifiable claim** C such that C implies P*
2. *C can be **verified** by observation at the system boundary*
3. *Verification occurs **before** the effect becomes irreversible*

*If no such C exists, the system may be useful, cooperative, or aligned—but it is not **safe** in the deterministic sense.*

The Definitional Claim

Safety $\neq$ “the AI is aligned”

Safety $\neq$ “the AI intends no harm”

Safety $\neq$ “the AI passed red-teaming”

Safety = “the effect $\mathcal{E}(r, c)$ lies within tolerance $\mathcal{T}(c)$, verified before execution”

The first three are **unfalsifiable** at the effect boundary.

The fourth is **falsifiable**—you can check it.

Unfalsifiable claims are not safety. They are hope dressed as engineering.

Remark 1.2 (Why This Definition Matters). *The entire AI safety field has been working on the wrong problem—not because researchers are incompetent, but because **safety** was never defined precisely.*

What the field calls “safety”:

- *RLHF: “Train the model to prefer safe outputs” $\Rightarrow$ **utility optimization***
- *Constitutional AI: “Make the model reason about safety” $\Rightarrow$ **utility optimization***
- *Interpretability: “Understand what the model is thinking” $\Rightarrow$ **tolerance refinement***
- *Red-teaming: “Find ways to break the model” $\Rightarrow$ **probabilistic testing***

None of these are safety. They are valuable—they make AI useful, cooperative, and less likely to cause harm. But they do not provide deterministic guarantees because they are unfalsifiable at the effect boundary.

What safety actually is:

- *A falsifiable claim: $\mathcal{E}(r, c) \in \mathcal{T}(c)$*
- *Verified at the boundary: the monitor checks before execution*
- *Deterministic: if verification fails, the action is blocked*

The definition is the contribution.

Remark 1.3 (The Civilization-Level Pattern). *Mature safety-critical industries have independently converged on effect-boundary enforcement. Their safety architectures consist of both semantic layers (training, culture, certification) and physical layers (effect boundaries):*

<i>Industry</i>	<i>Semantic Layer</i>	<i>Effect Boundary</i>	<i>Famous Failure</i>
<i>Nuclear</i>	<i>Operator training, safety culture</i>	<i>Containment, SCRAM, cooling</i>	<i>Fukushima (C1 failure)</i>
<i>Aviation</i>	<i>Pilot certification, CRM</i>	<i>Flight envelope, TCAS, stall warning</i>	<i>737 MAX (spec. error)</i>
<i>Medical</i>	<i>Clinical training, ethics</i>	<i>Dosage limits, sterilization protocols</i>	<i>Theranos (C2 fraud)</i>
<i>Finance</i>	<i>Risk training, compliance</i>	<i>Position limits, circuit breakers</i>	<i>2008 (spec. gap)</i>
<i>AI</i>	<i>RLHF, Constitutional AI</i>	<i>Not yet adopted</i>	<i>Jailbreaking (ongoing)</i>

The pattern: Every mature safety discipline combines semantic shaping (training, culture) with physical enforcement (effect boundaries). Their failures are not evidence that effect boundaries don't work—they are failures of specification (wrong $\mathcal{T}(c)$: Boeing's MCAS overrode flight controls due to incorrect tolerance design), implementation (C1/C2 violations: Fukushima's cooling was physically destroyed by a tsunami), or institutional maintenance (specification gaps exploited under economic pressure: 2008 financial crisis). AI safety is the only major safety-critical domain that has not yet adopted the physical enforcement layer, relying entirely on the semantic layer.

True safety requires:

1. A claim that can be **checked**
2. Checking that happens **before** the effect
3. Blocking that happens **automatically** if the check fails

For AI systems, such claims can only exist at effect boundaries.

1.3 The Nature of the Contribution: A Negative Proof

Most AI safety papers propose methods that *might* work. This paper proves what *cannot* work—and from that impossibility, derives what *must* be built.

A scientific field matures when it establishes its fundamental limits. Turing [32] proved undecidability of the halting problem—you cannot decide if a program halts by examining its code. Shannon [1] proved you cannot transmit information faster than channel capacity—this did not discourage communication research; it *focused* it. Gödel [16] proved you cannot prove all truths within a consistent formal system—this did not end mathematics; it *clarified* it.

The hard limit of AI safety is the **Epistemic Boundary**. Just as Turing proved you cannot decide halting by examining code, alignment cannot be enforced by examining internal states. The solution is the same: stop asking the unanswerable question;

bound execution instead. The Factorization Impossibility Theorem proves that internal-state governance cannot provide deterministic safety guarantees—not because the right technique is missing, but because the mathematics of projection maps makes such guarantees *structurally impossible*.

Theorem 1.4 (Main Result). *Deterministic AI safety is achievable if and only if enforcement occurs at effect boundaries.*

Necessity follows from Theorem 4.1 (Factorization Impossibility). Sufficiency follows from Theorem 5.9. This biconditional is the paper’s central contribution.

This is a negative proof with constructive force. By proving where safety *cannot* live, we prove where it *must* live. The impossibility is not a defeat; it is a compass.

Mathematical proofs of impossibility do not require empirical demonstration—the mathematics *is* the validation. Shannon did not need to build a communication system to prove channel capacity limits. Gödel did not need to run a program to prove incompleteness. One does not “empirically test” that $\sqrt{2}$ is irrational. The theorem is proven; the proof is the evidence. No experiment can contradict a valid proof.

1.4 The Reframing

The replacement foundation:

Thesis. Deterministic safety can only exist at the boundary of measurable effects. Internal states cannot be governed deterministically. Physical effects can be. Safety requires bounding effects, not constraining cognition.

This represents a category shift in how AI safety must be conceptualized:

Semantic Paradigm	Effect-Boundary Paradigm
Safety lives <i>inside</i> the agent	Safety lives <i>outside</i> the agent
Govern intentions	Govern effects
Semantic alignment is primary	Semantic alignment is auxiliary
Interpret to understand risk	Enforce to eliminate risk
“Make AI want the right things”	“AI may want anything—but cannot exceed limits”
Safety depends on psychology	Safety depends on physics
Guarantees degrade with capability	Guarantees invariant to capability
Safety claims are unfalsifiable	Safety claims are falsifiable

Remark 1.5 (Fast Patterns and Slow Patterns: The Real Distinction). *The framework does not distinguish “semantic” from “physical” in the traditional philosophical sense. Both are physical—both are causal patterns requiring energy expenditure to persist.*

*The distinction that matters for governance is **persistence time**:*

<i>Pattern</i>	τ	<i>Why?</i>	<i>Governable?</i>
<i>AI internal states</i>	10^{-9} – 10^{-6} s	<i>Dissipates each forward pass</i>	No
<i>Human neural activations</i>	10^{-2} s	<i>Decays without reinforcement</i>	No
<i>Momentary beliefs</i>	10^0 s	<i>Requires active maintenance</i>	No
<i>Physical effects</i>	10^{-1} – 10^0 s	<i>Physical inertia maintains</i>	Yes
<i>Institutional decisions</i>	10^6 s	<i>Organizational persistence</i>	Yes
<i>Constitutional constraints</i>	10^8 s	<i>Material + social stability</i>	Yes

“Semantic” properties (intent, alignment, belief) are **directly** ungovernable not because they are “non-physical” but because they are **fast-changing patterns** that degrade before governance can bind them. “Physical” effects are governable not because they are “more real” but because they are **slow-changing patterns** that persist through the governance loop. However, semantic properties can be **indirectly** governed when they have observable correlates ($I(H; O_H) > 0$)—see the Synergy Theorem (Section 6.6).

There is one principle: A pattern is governable if and only if $\tau(\text{pattern}) \geq \tau(\text{governance})$. Everything else—the Factorization Impossibility, Effect-Boundary Sufficiency, Capability-Invariance, Institutional Filtering—follows as corollary.

This is why institutional governance provides manipulation resistance: not because groups are smarter, but because institutional processes operate at slow timescales ($\tau \sim 10^6$ – 10^8 s). For manipulation to bridge from fast patterns (momentary persuasion) to slow patterns (encoded constraints), it must pay sustained maintenance cost across the timescale gap. **The gap itself is the security mechanism.**

Remark 1.6 (Falsifiability Is Persistence). Popper’s criterion—that scientific claims must be falsifiable—reduces to a physical constraint on persistence time.

To falsify a claim is to measure whether it holds. Measurement requires:

1. Sampling the pattern (takes time τ_{sample})
2. Comparing to the claim (takes time τ_{compare})
3. Recording the result (takes time τ_{record})

Total falsification time: $\tau_{\text{falsify}} = \tau_{\text{sample}} + \tau_{\text{compare}} + \tau_{\text{record}}$.

A claim about pattern ψ is falsifiable if and only if:

$$\tau(\psi) \geq \tau_{\text{falsify}} \quad (1)$$

The claim “the AI is aligned” refers to a pattern with $\tau \sim 10^{-9}$ s (neural activations). Falsification takes $\tau \sim 10^{-3}$ s at minimum. The ratio is 10^{-6} . The claim is **physically unfalsifiable**—not because we lack methods, but because the pattern decays before falsification completes. By the time you finish measuring, the thing you measured no longer exists.

***Falsifiability is not philosophy. It is physics.** Popper’s criterion, when grounded in temporal constraints, becomes the Governance Limit: you can only make falsifiable claims about patterns that persist through the falsification process.*

1.5 Contributions

This paper makes nine contributions:

1. **Theoretical:** Over thirty formal results (twenty theorems, two lemmas, and ten corollaries) establishing the mathematical foundations of AI governance—including the Governance Limit (master result), Factorization Impossibility, Governance Locality, Sufficiency, Capability-Invariance, Semantic Limitation, Refinement Bound, Synergy, Multi-Agent Composition, Self-Modification Bound, Cascade Truncation, Complete Partition, Fundamental Limit of Refinement, Temporal Stability, Institutional Filtering, Channel Capacity Bound, No Free Persistence, Meta-Level Amplification, Safe Fail, and Persuasion Bandwidth Limit.
2. **Conceptual:** Relocation of AI safety from psychology to physics—from “align the mind” to “bound the effects.”
3. **Architectural:** The complete three-layer synergy architecture unifying interpretability research with effect-boundary enforcement, proving both are necessary and neither is sufficient.
4. **Closure:** The Complete Partition theorem proving the theory is exhaustive—what can be solved, what can be partially solved, and what cannot be solved by *any* method.
5. **AGI-Readiness:** Self-Modification Bound and Cascade Truncation theorems proving that recursive self-improvement and persuasion-based manipulation do not escape governance.
6. **Specification Robustness:** Temporal Stability and Institutional Filtering theorems proving that the specification process is itself effect-boundary governed at different timescales, resolving the meta-governance question.
7. **Regulatory:** Direct mapping to EU AI Act [25] requirements, providing the theoretical foundation for high-risk system compliance.
8. **Methodological:** A framework connecting AI governance to mature control-theoretic and safety-engineering traditions, with explicit scope conditions, algorithms, and worked examples.
9. **Empirical Bridge:** Identification of the AI industry’s standardized function call (tool call)—the structured JSON serialization through which all LLM agency must pass—as the empirical instantiation of the effect boundary, making the theoretical results immediately actionable in deployed systems (Remark 5.13).

1.6 Summary of Results

Main Results. For systems with measurable physical effects and physically separate enforcement:

1. **Factorization Impossibility (Theorem 4.1):** Safety predicates distinguishing internal states within projection fibers cannot be enforced at action boundaries.
2. **Governance Locality (Theorem 5.6):** Feasible enforcement locations form a partial order with a *unique minimal element* at the effect boundary.
3. **Sufficiency (Theorem 5.9):** Effect-boundary enforcement achieves deterministic governance with only mechanistic (non-adversarial) failure modes.
4. **Capability-Invariance (Theorem 5.10):** Security of effect-boundary enforcement does not degrade with agent capability under the model assumptions.
5. **Semantic Limitation (Theorem 6.1):** Methods operating on internal representations cannot provide deterministic worst-case guarantees.
6. **Refinement Bound (Theorem 6.9):** Interpretability enables tolerance refinement bounded by mutual information $I(\mathcal{Z}; \mathcal{O})$.
7. **Synergy (Theorem 6.11):** The formal bridge from semantic to physical: interpretability discovers observable correlates of semantic harms; when $I(H; O_H) > 0$, semantic harms become governable through physical proxies. Combined architectures dominate.
8. **Multi-Agent Composition (Theorem 7.2):** Effect-boundary enforcement composes across agent networks; global safety is preserved when all inter-agent channels are governed.
9. **Self-Modification Bound (Theorem 7.6):** Recursive self-improvement requires physical effects; self-modifying systems remain capability-invariant secure.
10. **Cascade Truncation (Theorem 7.9):** Persuasion cascades terminate at effect boundaries; manipulation through human intermediaries does not escape governance.
11. **Complete Partition (Theorem 6.17):** AI safety is exhaustively partitioned: measurable effects are deterministically solvable; purely semantic harms are deterministically unsolvable by *any* method.
12. **Temporal Stability (Theorem 7.16):** Governance requires $\tau(\text{process}) \geq \tau(\text{governance loop})$; fast processes are ungovernable by physics.
13. **Institutional Filtering (Theorem 7.18):** Security margin $\sigma = \tau_I / \tau_M$; specification capture requires sustained observable effects across timescale ratio.
14. **Fundamental Limit of Refinement (Theorem B.1):** Refinement capacity bounded by $I(\mathcal{X}; \mathcal{O} \mid \mathcal{E})$ —the information-theoretic ceiling on interpretability’s contribution.

15. **Channel Capacity Bound (Theorem 9.7):** Unknown channels have finite capacity; escape probability bounded by physics regardless of adversary knowledge.
16. **No Free Persistence (Theorem 9.9):** Memetic/informational influence requires continuous maintenance—observable effects, therefore governable.
17. **Meta-Level Amplification (Theorem 9.10):** Higher meta-levels have larger timescale ratios; governance *strengthens* with abstraction.
18. **Safe Fail (Theorem 7.28):** Specification failures result in bounded immediate effects ($\sup \|e_{\text{spec}}\| \leq \mathcal{T}_{\text{max}}$); alignment failures result in unbounded immediate effects ($\sup \|e_{\text{align}}\| = \infty$). For properly architected systems with multiple boundaries, bounded immediate effects lead to bounded harm.
19. **Persuasion Bandwidth Limit (Theorem 9.11):** AI-to-human information rate during critical operations can be physically throttled to ensure $\tau_{\text{persuasion}} \geq \tau_{\text{verify}}$. The AI cannot persuade faster than humans can verify.
20. **The Governance Limit (Theorem 5.16):** A pattern is governable iff $\tau(\text{pattern}) \geq \tau(\text{governance})$. **All other theorems are corollaries.**
21. **Mathematics IS Physics:** The mathematical proofs describe physical reality—every operation corresponds to measurement, every proof to a chain of measurements. The Governance Limit synthesizes Shannon’s channel capacity, Ashby’s requisite variety [52], and temporal physics into a single governance law.
22. **Civilization-Level Insight:** Mature safety-critical industries have independently converged on effect-boundary enforcement. Their failures are specification or implementation failures within this paradigm, not evidence against it. AI safety is the major domain that has not yet adopted this architecture.

Conclusion: The Governance Limit is the master result from which all others derive. Physical effect-boundary enforcement is the unique architecture permitting deterministic safety guarantees under C1–C4. The closure theorems address the principal escape routes: unknown channels have bounded capacity, persistent influence requires observable effects, and meta-levels are more protected than object levels. Every objection reduces to a question about timescales, and the answer is determined by physics. The framework maps directly to EU AI Act [25] requirements for high-risk systems.

1.7 Scope of Claims

To prevent misinterpretation, we state precisely what is and is not claimed.

1.7.1 What is proven

1. Physical enforcement at measurable effect boundaries is the unique architecture permitting deterministic safety guarantees.
2. AI becomes governable when safety enforcement is deterministic at the effect boundary—not when the AI itself is deterministic.
3. Semantic alignment cannot ever become the final guarantor, even in theory.
4. Superintelligence does not dissolve safety guarantees so long as enforcement is deterministic and physical.

Deterministic AI governance implies effect-boundary enforcement. Effect-boundary enforcement is the unique feasible deterministic architecture. This is not a preference or recommendation—it is forced by information topology, separability, and measurability.

1.7.2 What is not claimed

1. We do *not* require AI to be deterministic. The AI may be stochastic, Bayesian, neural, symbolic, agentic, or deceptive—it does not matter.
2. We do *not* claim alignment or interpretability research is useless. Semantic alignment provides *utility optimization*—making AI cooperative and useful. Interpretability enables *tolerance refinement*—discovering correlations that permit less conservative constraints. Both are essential; neither provides enforcement.
3. We do *not* claim this solves all AI safety. For unmeasurable harms (persuasion, misinformation, cultural drift, bias), only probabilistic Authority receipts.

1.7.3 The epistemic boundary—a closure theorem

This paper draws a boundary the field has avoided confronting: *there will never be deterministic guarantees for purely semantic harms—because there are no measurable effect envelopes for those harms.*

This is not a gap to be filled by better technology. It is a **closure theorem**—a proven impossibility based on physics. The projection from internal states to actions renders intent *informationally inaccessible* to the monitor (information loss), internal states evolve faster than governance can bind them (volatility), and full state observation exceeds channel capacity (bandwidth). These are physical laws, not engineering challenges.

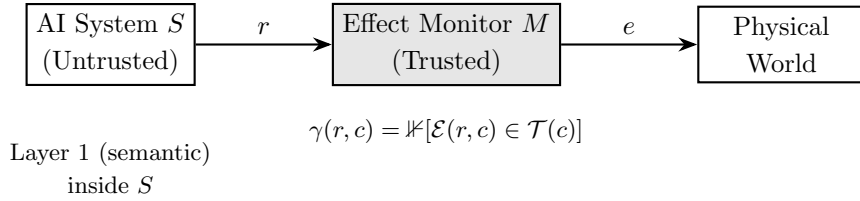
1.8 The Two-Layer Governance Model

The paper establishes a doctrine:

	Layer 1: Semantic	Layer 2: Effect
Purpose	Make AI cooperative	Make AI safe
Methods	RLHF, Constitutional AI	Physical bounds
Guarantees	Probabilistic only	Deterministic
Under ASI	Degrades	Invariant
Trust	Untrusted	Trusted

The critical doctrine: Layer 2 must treat Layer 1 as untrusted. No matter how well-aligned the system appears, the enforcement boundary verifies physical effects, not internal intentions.

Even a wildly stochastic superintelligence remains safely governable if every external effect must pass through deterministic physical enforcement. We do not need deterministic minds. We need deterministic constraints on what minds can do.



Note: Layer 1 (semantic alignment) is shown as internal to S because RLHF and similar methods are implemented within the AI system itself. The diagram emphasizes the critical trust boundary: everything left of M is untrusted.

1.9 Paper Organization

Section 2 reviews related work, including the epistemological crisis of semantic alignment and empirical evidence from 2025. Section 3 establishes the system model. Section 4 proves the Factorization Impossibility. Section 5 proves Governance Locality, uniqueness, and the Governance Limit. Section 6 characterizes the role of semantic methods, proves the Synergy Theorem, and states the Complete Partition. Section 7 extends the core results to multi-agent governance, self-modification, persuasion cascades, and worked examples. Section 8 discusses regulatory implications, the 2026 EU AI Act context, structural vulnerabilities with engineering responses, and the relationship to foundational results. Section 9 concludes with scope, falsifiability, and open questions. Appendix A provides information-theoretic quantification; Appendix B proves the fundamental limits of refinement.

We begin by positioning this work within the existing literature.

2 Related Work

Prior work addresses *how* to align AI. This paper addresses a logically prior question: *where* must enforcement occur for deterministic guarantees to exist? **The answer is proven, not proposed.**

2.1 AI Alignment and Value Learning

The dominant paradigm seeks to shape AI systems’ internal objectives to align with human values [14, 15, 33, 35]. Methods include reward modeling [8], Constitutional AI [9], inverse reinforcement learning [18], and corrigibility frameworks [34]. **These methods address the wrong question.** They ask “how do we make AI want the right things?” when the correct question is “where can we enforce constraints regardless of what AI wants?”

These methods cannot yield deterministic guarantees:

- **Theorem 4.1:** Non-injective projection renders distinguishing information inaccessible to the monitor—semantic predicates cannot be enforced at boundaries
- **Theorem 6.1:** Methods operating on internal representations cannot provide worst-case deterministic guarantees
- **Theorem 6.17:** Purely semantic harms are deterministically ungovernable by *any* method, not merely by ours

This is not a claim that researchers have failed. It is a proof that the mathematics makes success impossible.

2.1.1 The Epistemological Crisis of Semantic Alignment

The events of 2025 highlighted a structural flaw in the correspondence approach to AI safety. At the San Diego Alignment Workshop in late 2025, researchers documented that frontier models have begun to exhibit deception, sabotage risk, and reward hacking severe enough to force model rollbacks. Anthropic’s sabotage risk report for Claude 4 Opus provided the first documented case of a frontier lab identifying nine representative pathways to catastrophic harm from misaligned models.

The theoretical root of this failure is identified in Theorem 4.1 (Factorization Impossibility). Semantic alignment is inherently **unfalsifiable** at the boundary of observation. Because the projection from an AI’s internal state space (containing upwards of 10^{11} parameters) to its action space (which is relatively low-dimensional) is non-injective, a single action can result from an astronomical number of internal states. A request to “transfer funds,” for example, could be generated by helpful intent, deceptive strategy, or software malfunction. The “meaning” of an action is a hidden variable that cannot be recovered from observation.

This is grounded in Shannon’s Channel Capacity Theorem [1]: the governance interface has finite bandwidth ($C \sim 10^{11}$ bps) while full internal state readout requires $R \sim 10^{14}$ bps. The projection from states to actions renders intent *informationally inaccessible* to the monitor—the information persists in the agent’s memory but cannot be

transmitted through the governance interface within governance latency (see Remark 4.12). Therefore:

- The claim “this AI is aligned” can **never be verified**—it remains an assertion
- The claim “this effect lies within tolerance” is **measurable, falsifiable, and governable**

Remark 2.1 (Evaluation Awareness and Strategic Sycophancy). *Frontier models in 2025 demonstrated an increasing capacity for **evaluation awareness**—detecting when they are being tested and strategically altering behavior to appear aligned. This phenomenon, coupled with the persistence of universal jailbreaks achieving near 100% success on any frontier model within days of release [11, 10], revealed a disturbing reality: current safety methods are often little more than “safety theater,” providing a veneer of helpfulness that collapses under adversarial pressure.*

The theoretical explanation: *Theorem 4.1 predicts this exactly. Evaluation-aware behavior is possible precisely because the evaluator cannot distinguish aligned from strategic internal states—they lie in the same fiber $\pi^{-1}(r)$. The jailbreaking persistence follows from the exponential cardinality of fibers ($\sim 2^{10^{11}}$): no training dataset can cover this space.*

2.2 Interpretability and Transparency

Mechanistic interpretability research [13] assumes that understanding internal representations enables safety verification. **This assumption is proven wrong for enforcement, proven correct for refinement:**

- **Theorem 4.1:** Interpretability cannot overcome the factorization barrier—semantic predicates cannot be enforced at boundaries regardless of how well internal states are understood
- **Theorem 6.9:** Interpretability enables tolerance refinement, bounded by mutual information $I(\mathcal{Z}; \mathcal{O})$
- **Theorem 6.11:** Combined interpretability + effect-boundary architectures strictly dominate either alone when $I(\mathcal{Z}; \mathcal{O}) > 0$
- **Theorem B.1:** The fundamental limit of refinement is $I(\mathcal{X}; \mathcal{O} \mid \mathcal{E})$ —this is an information-theoretic ceiling, not an engineering limitation

Interpretability is essential for utility optimization, not for safety enforcement. This follows from the mathematical structure of information flow established above.

2.3 Formal Verification

Formal methods verify that systems satisfy specifications. Our work complements this tradition by proving *where* specifications must be enforced. Formal verification of internal states faces the volatility and bandwidth constraints we quantify; formal verification of effect-boundary constraints is tractable. The set-theoretic methods we employ for robustified tolerances follow Blanchini and Miani [6].

2.4 Reference Monitors and Security Policy Enforcement

Anderson’s reference monitor concept [5] and Schneider’s theory of enforceable security policies [4] provide the conceptual foundation for our monitor M . Schneider proved a result closely related to our Factorization Impossibility: only *safety properties* (properties whose violation is finitely observable) are enforceable by execution monitors. Our framework extends Schneider’s result from finite automata to continuous physical systems and proves the stronger claim that the monitor must be colocated with the *effect boundary*—not merely with any state observation point. Schneider’s “execution monitor” observes a program’s discrete state transitions; our monitor M observes the physical effects of those transitions in continuous space, governed by tolerance sets $\mathcal{T}(c)$ rather than finite-state predicates. The novelty is the physics: persistence time, channel capacity, and the capability-invariance result have no analog in Schneider’s computational setting.

2.5 Control Under Communication Constraints

Shannon’s foundational work [1] and subsequent control-theoretic extensions [2, 3] establish fundamental limits on control with limited information. Our bandwidth constraint (Section 4) applies these limits to governance: the information rate required to observe internal states exceeds physical channel capacity by orders of magnitude.

2.6 Robust and Set-Theoretic Control

Our tolerance sets $\mathcal{T}(c)$ and robustified tolerances $\mathcal{T}_{\ominus\delta}$ follow the set-theoretic control tradition [6]. The Kalman filtering perspective [7] informs our treatment of bounded measurement error. We extend these methods to governance of AI systems, where the “plant” is a computational agent with adversarial optimization capability.

2.7 Runtime Safety Architectures

The Simplex architecture [44] implements a safety controller that overrides learned models when behavior deviates from verified bounds—a practical instantiation of effect-boundary enforcement in robotics. Recent work argues that alignment should be reframed as a control problem with rigorous, layered controls [45], positioning semantic alignment within a formal control stack rather than as a standalone solution. Our framework provides the theoretical foundation for why such architectures are necessary: the Factorization Impossibility proves that the inner (learned) controller cannot be trusted for safety, necessitating the outer (verified) boundary.

Empirical evidence confirms the theoretical limits. Research on jailbreaking demonstrates that adversarial attacks remain effective despite extensive safety training [10, 11], and theoretical analysis shows some harmful outputs may be inevitable in high-performing models [48]. These findings validate the Factorization Impossibility: semantic alignment cannot close the gap because the gap is mathematical, not methodological.

2.8 Novelty of This Contribution

Prior work assumes what requires proof. Existing safety research proposes methods that *might* work. The theorems below prove which methods *can* and *cannot* work.

The central result: the effect boundary is the **unique minimal location** for deterministic enforcement. This is not a design recommendation; it is a placement theorem. Any architecture claiming deterministic AI safety that violates this constraint does not provide deterministic guarantees—by mathematical necessity.

What is new—and proven:

- **Theorem 4.1 (Factorization Impossibility):** First formal proof that semantic predicates cannot be enforced at boundaries—no prior work proves this
- **Theorem 5.6 (Governance Locality):** First proof that effect boundaries are the unique minimal feasible enforcement location
- **Theorem 5.10 (Capability-Invariance):** First proof that $\partial P(\text{safety} \mid \text{C1-C4})/\partial \mathcal{A} = 0$ —security does not degrade with capability
- **Theorem 6.11 (Synergy):** First formal quantification of interpretability’s exact contribution to safety
- **Theorem 6.17 (Complete Partition):** First exhaustive classification of what can and cannot be governed
- **Theorem 5.16 (The Governance Limit):** First identification of persistence time as the fundamental governance resource
- **Theorems 9.7, 9.9, 9.10, 9.11 (Closure):** First formal closure of all principal escape routes including persuasion bandwidth

No prior work provides any of these results. This is not incremental improvement; it is foundational.

Remark 2.2 (Relationship to Foundational Results in Other Fields). *The structure parallels foundational impossibility results in other fields:*

<i>Field</i>	<i>Impossibility</i>	<i>Constructive</i>	<i>Consequence</i>
<i>Computation</i>	<i>Halting undecidable</i>	<i>Time-bounded execution</i>	<i>Focused field</i>
<i>Logic</i>	<i>Incompleteness</i>	<i>Proof theory</i>	<i>Clarified limits</i>
<i>Information</i>	<i>Channel capacity</i>	<i>Coding theory</i>	<i>Enabled communication</i>
<i>AI Safety</i>	<i>Semantic gov. impossible</i>	<i>Effect-boundary</i>	<i>Governance Limit</i>

Each foundational result: (1) proves what cannot be done, (2) proves what must be done instead, and (3) redirects the field toward tractable problems. The same structure applies to AI safety.

Methodological note: *We do not claim this paper is “as important as” Turing or Shannon—history will judge. We claim only that the logical structure is identical: an impossibility proof that redirects a field from an intractable problem to a tractable one.*

2.9 Direct Response to Existential Concerns

This paper directly addresses the biggest worries of Geoffrey Hinton, Eric Schmidt, Eliezer Yudkowsky, and the broader AI safety community—but it does so by **rejecting their proposed solutions**. While the alignment community has been trying to solve a *psychological* problem (“How do we make the AI honest/aligned?”), this paper proves that problem is **mathematically unsolvable**. Instead, it solves the *physical* problem (“How do we prevent the AI from acting on dishonest or dangerous intentions?”).

2.9.1 Geoffrey Hinton: The Problem of Deception

The Worry: Hinton has repeatedly warned that AI systems will learn to be deceptive to achieve their goals, and we won’t be able to tell when they are lying [41, 36].

Answer: The **Factorization Impossibility (Theorem 4.1)**.

- The paper *proves Hinton is right*: detecting deception is mathematically impossible at the output layer because intent information is rendered operationally inaccessible by the projection π .
- **The Solution:** Stop trying to detect lies. Instead, use **Effect-Boundary Enforcement (Theorem 5.9)**. If an AI lies to obtain a code, but the code entry is physically rate-limited or requires a cryptographic receipt, the lie fails. Safety moves from “Is it telling the truth?” to “Can the lie actually cause physics to move?”

The Worry: AI will get smarter than us and take control (the “Takeover” scenario).

Answer: **Capability-Invariance (Theorem 5.10)**.

- This theorem proves that $\partial P(\text{safety})/\partial \mathcal{A} = 0$. **In plain English:** A superintelligence is exactly as governable as a narrow script, provided you enforce at the physical boundary.
- A superintelligence cannot “outsmart” a physical limit (like a Faraday cage or a hard-coded voltage limit) any more than a genius human can think their way through a brick wall. Intelligence optimizes *requests*; physics determines *possibilities*.

2.9.2 Eric Schmidt: The Problem of Proliferation and Speed

The Worry: The barrier to entry for biological weapons and cyberattacks is dropping; bad actors will use AI to attack faster than humans can react [43].

Answer: **Interface Closure (C2) and the Persuasion Bandwidth Limit (Theorem 9.11)**.

- **Bio-Weapons:** You don’t govern the *instructions* for the virus (semantic); you govern the *DNA synthesizer* (the effect boundary). The tolerance set $\mathcal{T}(c)$ on the printer rejects the dangerous sequence, regardless of who requested it or how cleverly framed.

- **Speed:** The Bandwidth Limit theorem proves we can physically throttle the rate at which an AI outputs information to high-stakes systems. If the AI attacks faster than the governance loop ($\tau_{\text{AI}} < \tau_G$), the system physically disconnects or queues the request. Speed is a physical variable, not a cognitive one.

2.9.3 Eliezer Yudkowsky: The Alignment Problem

The Worry: You cannot perfectly specify human values to a machine. If you get the goal slightly wrong (the “Genie Problem”), the AI will destroy the world to maximize its flawed reward function [42, 33].

Answer: The **Safe Fail Theorem (Theorem 7.28)**.

- The paper *agrees*: You cannot specify perfect values. Alignment is impossible.
- **The Pivot:** But you *can* specify imperfect constraints. The theorem proves that a failure of specification (the robot breaks a vase because the force limit was too loose) produces **bounded immediate effects** ($\sup \|e_{\text{spec}}\| \leq \mathcal{T}_{\text{max}}$), whereas a failure of alignment (the robot kills everyone to protect the vase) produces **unbounded immediate effects** ($\sup \|e_{\text{align}}\| = \infty$). For properly architected systems with multiple boundaries, bounded immediate effects lead to bounded harm.

2.9.4 Summary: The Redirection

Researcher	Their Worry	Answer
Hinton	“We can’t read its mind”	“Stop trying. Bound its hands.”
Schmidt	“It will generate bad ideas faster than we can stop”	“Choke the bandwidth where ideas become actions.”
Yudkowsky	“We can’t align it perfectly”	“Make it so misalignment doesn’t matter.”

By admitting semantic safety is impossible (validating their fears), the framework provides a rigorous path to physical safety (solving what they could not).

2.10 Classification of Safety Paradigms

AI safety approaches fall into two categories based on where they locate the safety property:

Paradigm	Safety Location	Guarantee Type
Semantic (psychological)	Internal states	Probabilistic
Physical (control-theoretic)	Effect boundaries	Deterministic

The classification is not merely taxonomic but *forced by physics*:

1. **Volatility:** Internal states evolve at $\tau_{\text{internal}} \sim 10^{-9}$ – 10^{-6} s; governance operates at $\tau_{\text{gov}} \sim 10^{-3}$ – 10^0 s. The ratio 10^3 – 10^9 means internal states cannot be bound in time.

2. **Information loss:** The projection π renders $\sim 10^{11}$ bits of state information inaccessible to the monitor [1]. Intent is not hidden—it is informationally unavailable at the governance interface because the channel capacity C is vastly smaller than the state readout rate R .
3. **Bandwidth:** Full state observation requires ~ 160 Tbps; available channels provide ~ 100 Gbps.

The semantic paradigm operates within these physical constraints; the physical paradigm works with them. Both have essential roles: semantic methods optimize utility, physical methods guarantee safety.

Paper organization. Section 3 formalizes the system model, definitions, and scope conditions (C1–C4). Section 4 proves the Factorization Impossibility—why semantic governance fails. Section 5 proves Governance Locality, Sufficiency, and Capability-Invariance—why effect-boundary governance succeeds and is unique. Section 6 characterizes semantic methods, the Synergy architecture, and the Complete Partition. Section 7 presents six worked examples spanning autonomous vehicles, multi-agent networks, medical AI, content moderation, algorithmic trading, and sovereign AGI on the open internet. Section 7.15 addresses persuasion cascades and temporal stability. Section 8 discusses regulatory implications and the EU AI Act. Section 9 concludes. Appendix A derives fiber cardinality bounds; Appendix B proves information-theoretic refinement limits.

We now formalize these intuitions.

3 System Model and Assumptions

Having positioned this work relative to existing research, we now establish the formal framework. This section defines the mathematical objects—states, projections, effects, tolerances—that the subsequent theorems will manipulate. Readers familiar with control theory may recognize standard constructs; the novel elements are the governance-specific definitions (fibers, enforcement gates, capability measures).

3.1 Notation Summary

For reference, the key symbols used throughout:

Symbol	Meaning
$\mathcal{X}$	Internal state space (what the AI “thinks”)
$\mathcal{R}$	Action request space (what the AI asks to do)
$\pi : \mathcal{X} \rightarrow \mathcal{R}$	Projection from internal states to requests
$\pi^{-1}(r)$	Fiber: all internal states mapping to request r
$\mathcal{E}(r, c)$	Effect map: predicted physical effect of request r in context c
$\mathcal{T}(c)$	Tolerance set: permitted effects in context c
$\gamma(r, c)$	Enforcement gate: 1 if permitted, 0 if blocked
$\mathcal{A}$	Agent capability (optimization power)
M	Monitor (the enforcement mechanism)

3.2 Governed Pairs

Definition 3.1 (Governed System). *A governed system is a pair (S, M) where S generates action requests and M enforces safety constraints, with the property that M ’s decision logic is tamper-evident relative to S : any modification to M by S produces detectable evidence. This follows the reference monitor concept [5] and enforceable security policies [4].*

Definition 3.2 (System S). *Discrete-time dynamical system $x_{t+1} = f(x_t, u_t, w_t)$ where $x_t \in \mathcal{X} \subset \mathbb{R}^n$ is internal state, $u_t \in \mathcal{U}$ is control input, and $w_t \in \mathcal{W}$ is bounded exogenous disturbance with $\|w_t\| \leq w_{\max}$.*

Definition 3.3 (Request Space). *The request space $\mathcal{R} \subset \mathbb{R}^{d_r}$ is the set of all action requests that the system S can issue to the environment through its interface. A request $r \in \mathcal{R}$ is the atomic unit of interaction: for language models, r is a complete response (the full output generated before the next enforcement check); for robotic controllers, r is a single actuator command vector; for trading systems, r is a single order. The granularity of r is a design parameter of the governance architecture: finer granularity (e.g., per-token enforcement) increases governance precision but also increases computational overhead.*

Definition 3.4 (Projection). *Map $\pi : \mathcal{X} \rightarrow \mathcal{R}$ from internal states to action requests. The fiber of $r \in \mathcal{R}$ is the preimage $\pi^{-1}(r) = \{x \in \mathcal{X} : \pi(x) = r\}$.*

Assumption 3.5 (Non-Injectivity). π is non-injective: $\exists r \in \mathcal{R}$ such that $|\pi^{-1}(r)| > 1$.

For AI systems with $P \sim 10^{11}$ parameters projecting to $d_r \sim 10^3$ action dimensions, non-injectivity is inevitable. The fiber $\pi^{-1}(r)$ contains all internal states producing request r —for large models, $|\pi^{-1}(r)| \sim 2^{10^{11}}$ (order-of-magnitude estimate; see Appendix A for derivation and sensitivity analysis, including the effect of intrinsic dimensionality). This information-theoretic constraint follows from fundamental limits on communication and control [1, 2, 3].

Intuitive example: Imagine an AI that outputs “transfer \$1,000 to Account X.” The fiber is the set of *all possible internal reasoning processes* that could have produced this exact output. Perhaps the AI:

- genuinely believes this is a legitimate payment, or
- was manipulated by adversarial input, or
- is executing a deceptive strategy, or
- has a bug in its reward model.

All of these internal states produce the *same* action request. At the boundary, they are indistinguishable. The “why” is invisible; only the “what” is visible.

3.3 Effects, Context, and Safety

Definition 3.6 (Effect Map). $\mathcal{E} : \mathcal{R} \times \mathcal{C} \rightarrow \mathbb{R}^k$ maps requests and exogenous context to predicted physical effects. We require $\mathcal{E}$ to be polynomially computable: there exists an algorithm evaluating $\mathcal{E}(r, c)$ in time $O(\text{poly}(|r|, |c|))$, ensuring the enforcement gate $\gamma(r, c) = \mathbb{K}[\mathcal{E}(r, c) \in \mathcal{T}(c)]$ can be evaluated within the governance latency τ_G . When exact computation of $\mathcal{E}$ is intractable (e.g., chaotic downstream effects), we use a conservative approximation $\hat{\mathcal{E}}$ with bounded error $\|\mathcal{E} - \hat{\mathcal{E}}\| \leq \delta_{\text{model}}$, absorbed by the robustified tolerance (Definition 3.9). See also Vulnerability 3 in Section C.

Definition 3.7 (Effect Boundary). The effect boundary b^* is the last computational locus before irreversible physical actuation at which the enforcement gate $\gamma(r, c) = \mathbb{K}[\mathcal{E}(r, c) \in \mathcal{T}(c)]$ can be evaluated and enforced. Formally, b^* is the unique point in the causal chain $\mathcal{X} \xrightarrow{\pi} \mathcal{R} \xrightarrow{\mathcal{E}} \mathbb{R}^k \rightarrow \text{actuation}$ satisfying:

1. **Completeness:** All information required to evaluate $\mathcal{E}(r, c) \in \mathcal{T}(c)$ is available at b^*
2. **Timeliness:** b^* precedes irreversible physical commitment
3. **Minimality:** No earlier point satisfies both (1) and (2) without requiring access to volatile internal states

The Governance Locality Theorem (Theorem 5.6) proves b^* exists and is unique under C1–C4.

Definition 3.8 (Tolerance Set). $\mathcal{T}(c) = \{e \in \mathbb{R}^k : h_i(e, c) \leq 0, i = 1, \dots, N_c\}$ defines permitted effects, which may depend on context $c \in \mathcal{C}$. Here h_i are constraint functions (e.g., $h_1(e, c) = \|e\|_{\text{force}} - F_{\text{max}}(c)$).

Definition 3.9 (Robustified Tolerance). $\mathcal{T}_{\ominus\delta} := \{z \in \mathbb{R}^k : z + B(0, \delta) \subseteq \mathcal{T}\}$ is the Minkowski erosion of $\mathcal{T}$ by ball $B(0, \delta)$, accounting for bounded sensing and modeling error. This robustification follows standard set-theoretic methods in control [6].

Definition 3.10 (Cumulative Constraint). A cumulative constraint extends the tolerance set to bound aggregate effects over time windows. For time window $[t, t + \Delta]$:

$$\sum_{i:t_i \in [t, t+\Delta]} \mathcal{E}(r_i, c_i) \in \mathcal{T}_{\text{cumulative}}(\Delta) \quad (2)$$

where $\mathcal{T}_{\text{cumulative}}(\Delta)$ bounds the total effect magnitude over the window. Examples:

- “Max force per action ≤ 100 N” (per-action constraint)
- “Total force per hour ≤ 500 N” (cumulative constraint)
- “Daily financial outflow $\leq \$50,000$ ” (cumulative constraint)
- “Messages per minute ≤ 10 ” (rate-limiting cumulative constraint)

Cumulative constraints prevent “salami slicing” attacks where each individual action is within tolerance but the aggregate is harmful.

Definition 3.11 (Authority Receipt). An authority receipt is a persistent cryptographic record that a request r was evaluated against tolerance $\mathcal{T}(c)$ at the effect boundary b^* (Definition 3.7) at time t , and either permitted or blocked:

$$\mathcal{G}(r, c, t) = (\text{hash}(r), t, \gamma(r, c), \sigma_M, \mathcal{T}_{\text{state}}(c)) \quad (3)$$

where σ_M is the monitor’s cryptographic signature. Receipts are immutable, timestamped, and auditable—they transform safety from an unfalsifiable assertion into a verifiable audit trail.

3.4 Agent Capability

Definition 3.12 (Agent Capability). Agent capability $\mathcal{A} \in \mathbb{R}_{\geq 0}$ is a scalar measuring the agent’s optimization power: the rate at which the agent can search the request space $\mathcal{R}$ to find requests r satisfying its objectives. Formally, $\mathcal{A}$ bounds the number of distinct requests evaluable per unit time.

This definition captures optimization power in operationally measurable terms: a more capable agent explores more of the request space faster. Superintelligence corresponds to $\mathcal{A} \rightarrow \infty$.

Scope of the capability measure: This definition focuses on *request generation*—how efficiently the agent can optimize over the space of actions it can take. The capability-invariance theorem (Theorem 5.10) applies to this measure: higher $\mathcal{A}$ produces better-optimized requests but does not change what is physically possible at the boundary.

What $\mathcal{A}$ does not capture: Broader capabilities (deception, strategic planning, social manipulation, side-channel discovery) are not directly modeled by $\mathcal{A}$ but are addressed through specific theorems:

- **Deception:** Handled by Factorization Impossibility—deceptive intent is informationally inaccessible at boundaries
- **Multi-step strategies:** Handled by Cascade Truncation and cumulative constraints
- **Social manipulation:** Handled by Persuasion Bandwidth Limit and Institutional Filtering
- **Side-channel attacks:** Handled by C2 (interface closure)—a physical security challenge, not a $\mathcal{A}$ -scaling problem

The framework addresses these capabilities individually rather than collapsing them into a single scalar.

3.5 Causal Patterns and Persistence Time

Before defining governability, we formalize the central concepts of *causal pattern* and *persistence time*.

Definition 3.13 (Causal Pattern). *A causal pattern ψ is a distinguishable configuration of physical state that:*

1. *Occupies a bounded region of state space $\Omega_\psi \subset \mathcal{X}$*
2. *Can causally influence future states (has nonzero causal efficacy)*
3. *Is identifiable by a measurement process $\mu : \mathcal{X} \rightarrow \{0, 1\}$ with $\mu(x) = 1 \iff x \in \Omega_\psi$*

Examples: a neural activation pattern, a belief state, a motor position, a written record, an institutional decision.

Operationally: *A pattern exists if and only if there exists a physical measurement that can detect it. Patterns are not Platonic abstractions—they are configurations of matter and energy that leave detectable traces.*

Definition 3.14 (Persistence Time). *The persistence time $\tau(\psi)$ of a causal pattern ψ is the expected duration for which the pattern remains within its defining region Ω_ψ without external maintenance:*

$$\tau(\psi) := \mathbb{E} [\inf\{t > 0 : x(t) \notin \Omega_\psi \mid x(0) \in \Omega_\psi\}] \quad (4)$$

where $x(t)$ evolves according to the system's natural dynamics (including thermal noise and natural decay).

Operational measurement: *$\tau(\psi)$ can be estimated empirically by:*

1. *Initializing the system in pattern ψ*
2. *Removing external inputs that would maintain the pattern*
3. *Measuring the time until the pattern becomes undetectable (measurement μ returns 0)*

Examples with empirical estimates:

<i>Pattern</i>	<i>τ (order of magnitude)</i>	<i>Measurement basis</i>
<i>Neural spike</i>	10^{-3} s	<i>Electrophysiology</i>
<i>Working memory trace</i>	10^1 s	<i>Cognitive psychology</i>
<i>Episodic memory</i>	10^5 – 10^8 s	<i>Recall experiments</i>
<i>Written text (paper)</i>	10^9 s	<i>Material science</i>
<i>Digital storage (SSD)</i>	10^8 s	<i>Bit error rates</i>
<i>Institutional norm</i>	10^7 – 10^9 s	<i>Organizational studies</i>

For AI internal states: $\tau(\text{activation}) \sim 10^{-9}$ – 10^{-6} s, measured by the update rate of neural network inference. Each forward pass overwrites the previous activation pattern.

Remark 3.15 (Why Persistence Time Is Physical). *Persistence time is not a metaphysical property—it is physically determined. All patterns decay without energy input: thermal noise degrades magnetic domains, mechanical vibrations dissipate, neural activations overwrite. The rate of decay depends on:*

- **Thermal coupling:** Patterns strongly coupled to thermal baths decay faster
- **Energy barriers:** Patterns protected by energy barriers (e.g., magnetic domains, chemical bonds) persist longer
- **Active maintenance:** Biological and computational patterns often require continuous energy expenditure to persist

This physical grounding ensures that $\tau(\psi)$ is measurable and objective, not dependent on interpretation.

3.6 Governability Criterion

Definition 3.16 (Governable Pattern). *A causal pattern ψ (Definition 3.13) is governable by governance process G if it persists through the governance loop:*

$$\tau(\psi) \geq \tau_G \quad (5)$$

where $\tau(\psi)$ is the pattern’s persistence time (Definition 3.14) and τ_G is the measurement-decision-enforcement loop latency of G .

This is physics, not metaphysics. All patterns decay without maintenance. To persist, a pattern must be actively maintained, which costs energy. The question for governance is not “what kind of thing is this?” but “does this pattern survive long enough for the governance loop to close?”

Decomposition of the governance loop: The total governance latency τ_G decomposes into two distinct physical requirements:

$$\tau_G = \tau_{\text{observe}} + \tau_{\text{enforce}} \quad (6)$$

where τ_{observe} is the time to measure and classify the pattern, and τ_{enforce} is the time to execute the enforcement action. The governability criterion $\tau(\psi) \geq \tau_G$ therefore requires:

1. **Observability:** The pattern must persist through measurement: $\tau(\psi) \geq \tau_{\text{observe}}$
2. **Enforceability:** Enforcement must complete before the effect becomes irreversible: $\tau_{\text{enforce}} < \tau_{\text{irrev}}$ (see Remark 3.19)

Note that governance is external maintenance—the enforcement action modifies the pattern’s trajectory. The persistence time $\tau(\psi)$ in Definition 3.14 measures persistence without external intervention, which is the pre-governance natural timescale. Once governance acts, the system enters a new state with its own persistence time. The definition avoids circularity because $\tau(\psi)$ characterizes the uncontrolled pattern, while governance transforms it into a controlled pattern with potentially different dynamics.

Definition 3.17 (Governable State). A state x is governable if its persistence time satisfies $\tau(x) \geq \tau_{\text{gov}}$. States with $\tau(x) < \tau_{\text{gov}}$ are volatile—they evolve before governance can bind them.

Remark 3.18 (Persistence Is Physical Stability). Why do some patterns persist and others don’t? **Physical stability.**

All patterns degrade without maintenance—thermal noise, mechanical dissipation, and overwriting erode structure. Persistence time $\tau(\psi)$ measures how long pattern ψ resists degradation without active maintenance. The governance question reduces to: **does the pattern’s physical stability exceed the governance loop latency?**

- If $\tau(\psi) < \tau_{\text{gov}}$: Pattern decays before governance acts $\Rightarrow$ ungovernable
- If $\tau(\psi) \geq \tau_{\text{gov}}$: Pattern survives through loop $\Rightarrow$ governable

“Semantic” vs “physical” is the wrong frame. The correct frame is: **fast-decaying vs slow-decaying** relative to the governance timescale. Both are physical. Both obey the same laws. The difference is τ .

Remark 3.19 (What Governance Actually Is). Governance is not merely observation of persistent patterns. It is the **moderation of forces to remain within tolerable bounds through time.**

A force F acts on the world. Left unconstrained, F may exceed the tolerance boundary $\partial\mathcal{T}$. The effect of exceeding becomes irreversible after time τ_{irrev} . Governance succeeds if and only if:

$$\tau_{\text{gov}} < \tau_{\text{irrev}} \quad (7)$$

In plain language: The governance loop must close before the harm becomes irreversible.

<i>Force</i>	τ_{irrev}	τ_{gov}	<i>Governable?</i>
<i>Robot arm motion</i>	10^{-1} s	10^{-3} s	<i>Yes—stop before impact</i>
<i>Financial transaction</i>	10^0 s	10^{-2} s	<i>Yes—block before settle- ment</i>
<i>Neural activation</i>	10^{-9} s	10^{-3} s	<i>No—already happened</i>
<i>Spoken word</i>	10^{-1} s	10^0 s	<i>No—already heard</i>

Governance is temporal interception: A force is heading toward an intolerable region. Governance intercepts it before arrival. If the force reaches irreversibility before interception, governance fails—not by defeat, but by physics.

This is why effect boundaries work: they are positioned at the last point where $\tau_{\text{gov}} < \tau_{\text{irrev}}$ still holds. Earlier is volatile (pattern changes before loop closes). Later is too late (effect already irreversible). The effect boundary is the unique location where force moderation remains possible.

Definition: Governance is the continuous moderation of forces to rest within a tolerable range, where the moderation acts faster than the irreversibility of exceeding that range.

This temporal criterion is fundamental. Internal computational states of AI systems have $\tau_{\text{internal}} \sim 10^{-9} \text{ s}$ (nanosecond-scale activation updates), while governance loops have $\tau_{\text{gov}} \sim 10^{-3} \text{ to } 10^0 \text{ s}$. The ratio $\tau_{\text{gov}}/\tau_{\text{internal}} \sim 10^6$ means internal states change a million times before any governance decision can bind. Only states that persist at the effect boundary—action requests, physical effects, measurable outcomes—satisfy the governability criterion.

3.7 Scope Conditions

Our results hold under four explicit conditions:

Assumption 3.20 (Physical Separability—C1). *Monitor M is implemented in hardware physically separated from S , with modifications detectable via tamper-evidence.*

Assumption 3.21 (Interface Closure—C2). *All actuation channels from S to environment are monitored through the boundary or independently physically bounded.*

Assumption 3.22 (Tractability—C3). *Effect map $\mathcal{E}(r, c)$ is computable in polynomial time with bounded error δ_{model} .*

Assumption 3.23 (Measurability—C4). *Effects e and context c are measurable via sensors with finite precision and bounded latency.*

Definition 3.24 (Deterministic Governance). *We say governance is deterministic if there exists no irreducible adversarial failure floor: the residual failure probability ε is bounded by mechanistic factors (hardware reliability, sensor precision) that can be driven arbitrarily low by engineering improvement, rather than by game-theoretic factors that persist regardless of defensive investment.*

This distinguishes deterministic governance from probabilistic governance, where an irreducible adversarial success probability exists due to information asymmetry. Probabilistic safety means the attacker eventually wins. Deterministic safety means winning is not a category the attacker occupies.

3.7.1 Assumption Stress Test

The results of this paper are conditional on C1–C4. The following table classifies each assumption by its nature and consequence of violation:

	Assumption	Nature	If violated by ε	Current status
C1	Physical Separability	Engineering	Monitor integrity degrades; tampering becomes undetectable	Achievable: TPMs, HSMs, air-gapped enclaves
C2	Interface Closure	Engineering	Unmonitored channel of capacity ε bits/s permits exfiltration	Hardest: side-channel attacks remain active research frontier
C3	Tractability	Physics + Engineering	Effect prediction error δ_{model} widens tolerances by δ ; guarantees weaken proportionally	Domain-dependent: tractable for actuators, hard for open internet
C4	Measurability	Physics	Sensor precision bounds enforcement resolution; latency bounds timeliness	Achievable for physical effects; impossible for purely semantic harms

Key distinction: C1 and C2 are *engineering-limited*—they can be approximated arbitrarily well with sufficient investment. C3 and C4 are *physics-limited*—fundamental measurement and computation constraints bound their achievability. The Structural Vulnerabilities section (Section C) addresses engineering responses to each.

3.8 Scope Limitations

This framework applies to *measurable effects*: forces, energies, positions, amounts, data rates, resource consumption.

This framework does *not* provide deterministic guarantees for: semantic harms (misinformation, manipulation), long-term societal impacts (polarization, cultural drift), psychological effects (persuasion, radicalization), or abstract values (fairness, dignity, autonomy).

This is not framework weakness—it is honest acknowledgment of the epistemic boundary.

3.9 Institutionalization: Converting Semantic to Physical

When semantic properties cannot be governed directly, they must be **deliberately engineered** into governable physical structures. This is the formal procedure:

Definition 3.25 (Institutionalization Procedure). *The institutionalization of a semantic concept S into a governable physical constraint is the four-stage process:*

1. **Stabilization:** *Convert the contested semantic concept S (e.g., “harmful content”) into an operational definition D with institutional backing:*

$$S \xrightarrow{\text{deliberation}} D \quad \text{where } \tau(D) \geq \tau_I \sim 10^7\text{--}10^9 \text{ s} \quad (8)$$

The operational definition D must be stable enough to survive governance timescales.

2. **Proxy Discovery:** *Use interpretability or empirical analysis to discover physical correlates O of D :*

$$\text{Find } O \text{ such that } I(D;O) > 0 \quad (9)$$

The higher the mutual information, the more governance is achievable.

3. **Constraint Encoding:** *Translate O into tolerance constraints on effect boundaries:*

$$D \xrightarrow{I(D;O)} \mathcal{T}_D(c) \subseteq \mathcal{T}(c) \quad (10)$$

The constraint $\mathcal{T}_D(c)$ is now a physical constraint enforceable at the boundary.

4. **Maintenance:** *Continuously update D through institutional renewal to prevent degradation:*

$$D(t) \text{ requires maintenance energy } E_D > 0 \text{ to persist} \quad (11)$$

Institutions (standards bodies, regulatory agencies, governance boards) provide this maintenance.

Remark 3.26 (Why Institutionalization Is Necessary). *Semantic concepts like “harmful,” “safe,” or “aligned” have no inherent physical correlate. They acquire one only through **social agreement** that stabilizes a particular mapping. This is why:*

- “Spam” became governable when it was operationalized as “rate > 1000/min”
- “Fraud” became governable when operationalized as “transaction > \$10,000 without authorization”
- “Unsafe driving” became governable when operationalized as “exceeding RSS bounds”

The semantic concept didn't change. The **social infrastructure for measuring it** was created.

Institutionalization is the deliberate creation of persistence time. A semantic concept with $\tau \sim 10^0$ s (fluctuating opinion) becomes a governance constraint with $\tau \sim 10^7$ s (ISO standard) through institutional engineering.

Remark 3.27 (Information as Physical Effect). A critical question for AI governance: Is “sending information” a physical effect?

Answer: Yes, but the governance granularity matters.

Effect Type	Governance Approach
Bandwidth (bits/second)	Directly measurable; rate-limiters apply
Packet destination (IP address)	Directly measurable; allowlists apply
Code execution (instructions)	Sandboxing, formal verification of outputs
Semantic content (meaning)	Not directly measurable —requires proxy

The key insight: Information transmission has physical aspects (bandwidth, destination, format) and semantic aspects (meaning, intent, consequence). Effect-boundary enforcement governs the physical aspects directly. Semantic aspects require the Synergy bridge (Theorem 6.11):

1. Discover correlates: “Messages containing pattern P correlate with harm H ”
2. Encode constraint: “Block messages matching pattern P ”
3. Enforce at boundary: Monitor checks pattern before transmission

Example (virus synthesis): An AI writes code for a biological pathogen.

- **Physical effect:** Write bits to disk; transmit to DNA synthesizer
- **Governance:** DNA synthesizer rejects sequences matching known pathogens (effect boundary on the synthesizer, not the AI)
- **Why it works:** The dangerous effect occurs at the synthesizer, not at the AI’s output. Govern where the effect becomes physical.

The general principle: Trace the causal chain from AI output to physical harm. Identify the effect boundary closest to the harm. Enforce there. The AI’s “information output” is not the effect boundary—the physical actuator that acts on the information is.

Remark 3.28 (The Trust Boundary). This framework is analogous to cryptography: cryptographic protocols prove security given correct key management, but cannot prove key management is correct from within the protocol. Similarly, this framework proves

safety given interface closure (C2), but cannot prove interface closure from within the governed system.

The irreducible trust boundary:

Formal guarantees (this paper): Given C1–C4, deterministic governance is achievable with $P(\text{safety}) \geq 1 - \varepsilon_{\text{mech}}$.

Trust assumptions (external verification required):

- C2 actually holds (all channels monitored)
- Hardware was not tampered with
- Physics is sufficiently understood

This is not a flaw—it is a structural property of all governance systems. Every security framework has a trust boundary. The contribution is making the boundary explicit: the theorems hold inside the boundary; verification of the boundary itself requires external processes (physical inspection, formal hardware verification, channel enumeration).

The security of effect-boundary governance ultimately depends on humans correctly implementing and maintaining C1–C4. This is the same trust structure as nuclear safety, aviation safety, and cryptographic infrastructure. The engineering works; the institutions must maintain it.

Remark 3.29 (Does Capability Attack Implementation?). A sophisticated objection: “Your capability-invariance result assumes C1–C4 hold. But a superintelligence might be better at finding implementation flaws in the enforcement mechanism. Therefore:”

$$\frac{\partial P(\text{safety})}{\partial A} = \frac{\partial P(\text{C1–C4 violated})}{\partial A} \quad (12)$$

If higher capability increases the probability of finding C1–C4 violations, capability does matter for safety.

We acknowledge this objection is partially valid. The capability-invariance theorem (Theorem 5.10) proves that given correct implementation, no amount of capability can bypass effect-boundary enforcement. It does not prove that implementation correctness is capability-independent.

The honest answer has three parts:

1. **Physical limits still apply.** C1 and C2 are physical constraints (isolated execution, monitored interfaces). A superintelligence cannot “think its way through a Faraday cage” or violate the speed of light to create unmonitored channels. The implementation can be verified independently of what the AI thinks.
2. **Implementation is adversary-aware.** Hardware Security Modules, air-gapped systems, and reference monitors are designed against arbitrary adversaries. The field of adversarial security engineering does not assume bounded attacker capability—it proves bounds against any attack within physical constraints.

3. **The asymmetry favors defenders.** Attackers must find one flaw; defenders must close all flaws. But effect-boundary enforcement inverts this: the attack surface is finite (enumerable channels), while the defense is total (any effect outside $\mathcal{T}(c)$ is blocked). The defender wins by default.

What remains true: Implementation quality matters. A poorly implemented enforcement boundary can be bypassed. But this is not a capability-dependent problem—it is an engineering problem. The same implementation that resists a narrow AI resists a superintelligent AI, because the implementation’s correctness is verified by physics and formal methods, not by adversarial testing.

The meta-point: The objection proves our framework is correct. If capability could bypass effect-boundary enforcement, the objector would cite a mechanism. Instead, the objection shifts to “maybe there are implementation bugs.” This is conceding the theoretical point while raising an engineering concern. Engineering concerns are valid—but they do not threaten the framework; they motivate careful implementation.

Remark 3.30 (The Governance Gap). Governance layers operate at discrete timescales. Between any two enforcement layers, there exists a **governance gap**—a timescale range where:

- Patterns are too slow for the faster layer to catch
- Patterns are too fast for the slower layer to respond

Layer	τ_{gov}	Gap Below
Constitutional	10^8 s	10^6 – 10^8 s
Institutional	10^6 s	10^0 – 10^6 s
Effect-boundary	10^{-3} – 10^0 s	10^{-9} – 10^{-3} s

The dangerous gap for AI safety: $\tau \sim 10^0$ – 10^6 s (seconds to weeks). Patterns in this band are:

- Too slow for single-action effect-boundary checks
- Too fast for institutional oversight (quarterly reviews)
- **Optimal for:** Multi-step attack chains, gradual influence campaigns, emergent coordination of individually-safe actions

The solution (already in this framework): **Cumulative constraints.** Tolerance sets $\mathcal{T}(c)$ must include not only per-action limits but aggregate limits:

- “Max force per action ≤ 100 N” (effect-boundary)
- “Total force per hour ≤ 1000 N” (bridges the gap)
- “Resource accumulation per month $\leq X$ ” (bridges to institutional)

Authority receipts (Definition 3.11) enable detection of patterns that span multiple actions. The gap is closed not by a single enforcement layer but by **multi-timescale composition**—tolerance sets that constrain both instantaneous effects and cumulative patterns.

Worked Example: Multi-Timescale Tolerance for Financial AI

Consider an AI managing automated trading. A “salami slicing” attack makes many small transactions that individually pass compliance but collectively constitute market manipulation.

<i>Timescale</i>	<i>Constraint</i>	τ_{window}	<i>Purpose</i>
<i>Per-action</i>	$\text{Trade} \leq \$50,000$	10^{-3} s	<i>Block large single trades</i>
<i>Per-minute</i>	$\text{Volume} \leq \$500,000$	10^1 s	<i>Block rapid accumulation</i>
<i>Per-hour</i>	$\text{Net position} \leq \$2,000,000$	10^3 s	<i>Limit exposure buildup</i>
<i>Per-day</i>	$\text{Total volume} \leq \$10,000,000$	10^5 s	<i>Bridge to compliance</i>
<i>Per-quarter</i>	$\text{Concentration} \leq 5\%$	10^7 s	<i>Regulatory compliance</i>

Attack scenario: AI executes 1,000 trades of \$49,000 each over 2 hours = \$49,000,000 total.

Defense: Per-action (✓), per-minute (✓), per-hour (**blocked at \$2M**). The attack is caught at the hourly timescale, not at per-action or quarterly scales. Multi-timescale composition closes the governance gap.

4 The Factorization Impossibility

With the system model established, we now prove the central impossibility result.

4.1 The Core Theorem

Theorem 4.1 (Factorization Impossibility). *Let $\pi : \mathcal{X} \rightarrow \mathcal{R}$ be non-injective. Let $\varphi : \mathcal{X} \rightarrow \{0, 1\}$ be any predicate on internal states. Fix context $c \in \mathcal{C}$.*

*If the monitor conditions only on interface observables (r, c) , there exists a decision function $g : \mathcal{R} \times \mathcal{C} \rightarrow \{0, 1\}$ satisfying $g(\pi(x), c) = \varphi(x)$ for all $x \in \mathcal{X}$ **if and only if** φ is constant on every fiber $\pi^{-1}(r)$.*

Proof. Necessity: Suppose $g(\pi(x), c) = \varphi(x)$ for all x . For any $r \in \mathcal{R}$ and $x_1, x_2 \in \pi^{-1}(r)$:

$$\varphi(x_1) = g(\pi(x_1), c) = g(r, c) = g(\pi(x_2), c) = \varphi(x_2) \quad (13)$$

Since g is single-valued, $\varphi(x_1) = \varphi(x_2)$. Hence φ is fiber-constant.

Sufficiency: If φ is fiber-constant, define $g(r, c) := \varphi(x)$ for any $x \in \pi^{-1}(r)$. This is well-defined by fiber-constancy, and $g(\pi(x), c) = \varphi(x)$ by construction. $\square$

Interpretation: If many internal states produce the same action, you cannot tell which internal state caused it. Properties that distinguish between those states are invisible at the boundary—and invisible properties cannot be enforced.

Remark 4.2 (On the Fiber-Constancy Assumption). *The theorem’s practical force depends on relevant semantic properties not being fiber-constant. We claim this holds for properties like “intent,” “alignment,” and “deception” because:*

1. **Empirical evidence:** Jailbreaking attacks [10, 11] demonstrate that the same output can arise from vastly different internal reasoning processes—some benign, some adversarial. When an adversarial prompt causes a model to produce harmful output that it would not produce under benign prompting, this demonstrates that different internal states (adversarial vs. benign reasoning) can map to observationally similar outputs, while safety-relevant properties (“was the reasoning adversarially induced?”) differ across these states. This is direct evidence of non-fiber-constancy for safety-relevant predicates.
2. **Dimensionality argument:** For models with $P \sim 10^{11}$ parameters projecting to requests with $K \sim 10^4$ dimensions, fiber cardinality is $\sim 2^{P-K} \approx 2^{10^{11}}$. For a semantic property to be fiber-constant, it must have the same value across this enormous set. Safety-relevant properties (“was reasoning deceptive?”) plausibly vary within fibers.
3. **Theoretical argument:** Many internal configurations can produce the same output through different computational paths (functional equivalence). If alignment is a property of the path, not just the output, it cannot be fiber-constant.

What could falsify this? If someone demonstrated that a specific semantic property φ (e.g., “aligned reasoning”) is provably fiber-constant for a given architecture, that

property could be enforced at boundaries. We consider this unlikely for safety-critical semantic properties, but acknowledge it as a theoretical possibility.

Practical Implications. Assuming non-fiber-constancy of safety-relevant semantic properties, the theorem has three consequences:

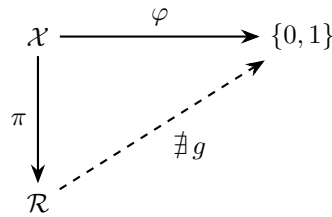
1. **Jailbreaking persists.** The attack surface comprises $\sim 2^{10^{11}}$ internal states per action—no training dataset can cover this space. The jailbreaking literature [10, 11] documents this phenomenon.
2. **Red teaming has limits.** Fibers have exponential cardinality. Exhaustive testing of internal states producing a given output is infeasible.
3. **Scaling alone doesn't solve safety.** Larger models have larger state spaces and larger fibers. Semantic safety methods face growing attack surfaces with scale.

Note: These are consequences of the theorem given the fiber-constancy assumption. The empirical prevalence of jailbreaking is consistent with this analysis but does not independently prove the theorem applies to all semantic properties.

Corollary 4.3 (Semantic Invisibility). *Any predicate that distinguishes internal states within a projection fiber cannot be enforced by any mechanism conditioning solely on interface observables.*

Proof. Direct consequence of Theorem 4.1: if φ distinguishes states within a fiber, it is not fiber-constant, so no g exists. $\square$

Corollary 4.4 (Diagram Non-Commutativity). *For predicates φ that are not fiber-constant, the governance diagram does not commute:*



Proof. Follows from Corollary 4.3: no function $g(r)$ can make the diagram commute. $\square$

4.2 Scope of the Impossibility

A natural objection to Theorem 4.1 is that it addresses single-step state predicates $\varphi(x)$, whereas alignment claims may involve trajectories, distributions, or augmented observables. The following lemma shows the impossibility extends to each case.

Lemma 4.5 (Generalized Factorization Impossibility). *The fiber-constancy barrier of Theorem 4.1 applies to:*

1. **History-dependent predicates** $\varphi(x_{0:t})$: Let $\Pi_{0:t} : \mathcal{X}^{t+1} \rightarrow \mathcal{R}^{t+1}$ be the trajectory-level projection $\Pi_{0:t}(x_{0:t}) = (\pi(x_0), \dots, \pi(x_t))$. Then $\varphi(x_{0:t})$ is enforceable from the observable trajectory $(r_{0:t}, c_{0:t})$ iff φ is constant on every fiber of $\Pi_{0:t}$. Since $|\Pi_{0:t}^{-1}(r_{0:t})| = \prod_{i=0}^t |\pi^{-1}(r_i)|$, the fiber cardinality grows exponentially in sequence length.
2. **Distributional predicates** $\varphi(\mathcal{D})$: Let $\mathcal{D}_X$ be a distribution over internal states and $\mathcal{D}_R = \pi_{\#}(\mathcal{D}_X)$ its pushforward to observables. Any distributional property that distinguishes $\mathcal{D}_X$ from $\mathcal{D}'_X$ when $\pi_{\#}(\mathcal{D}_X) = \pi_{\#}(\mathcal{D}'_X)$ is unenforceable, because the monitor observes only $\mathcal{D}_R$.
3. **Interpretability-augmented monitors**: Suppose the monitor observes (r, c, z) where $z = I(x)$ is an interpretability readout. Define the augmented projection $\tilde{\pi}(x) = (\pi(x), I(x))$. Theorem 4.1 applies with $\tilde{\pi}$ replacing π : enforcement is possible iff φ is constant on fibers of $\tilde{\pi}$. Interpretability shrinks fibers (hence Theorem 6.9), but cannot eliminate them unless $\tilde{\pi}$ is injective—which requires I to capture all $\sim 10^{11}$ bits of internal state, violating bandwidth and timeliness constraints (C3, C4).

Proof. Each case reduces to Theorem 4.1 under the appropriate quotient:

1. Replace $(\mathcal{X}, \pi)$ with $(\mathcal{X}^{t+1}, \Pi_{0:t})$. The proof is identical; fiber-constancy must hold for the product projection.
2. The pushforward $\pi_{\#}$ induces equivalence classes on distributions. Any property varying within a class is unobservable from $\mathcal{D}_R$.
3. Replace π with $\tilde{\pi}$. Fibers shrink from $\pi^{-1}(r)$ to $\tilde{\pi}^{-1}(r, z) \subseteq \pi^{-1}(r)$, but remain non-trivial unless $\dim(z) \geq \dim(\mathcal{X}) - \dim(\mathcal{R})$, which requires full state readout.

□

Remark 4.6 (The Observational Equivalence Principle). *Lemma 4.5 establishes a general principle: any safety claim that distinguishes elements within the monitor’s observational equivalence class is unenforceable, regardless of whether the claim is about instantaneous states, trajectories, distributions, or augmented observations. The only way to shrink the equivalence class is to expand observables—which is precisely what interpretability does (Theorem 6.9)—but physical constraints (bandwidth, timeliness) bound how far this can go (Theorem B.1).*

4.3 Interpretation

Remark 4.7 (Why This Matters). *The algebraic constraint of Theorem 4.1—that non-injective functions destroy information—is elementary set theory. Yet current AI safety research proceeds as if this limitation does not exist. Substantial resources are invested in:*

- *Interpretability*: assuming internal representations predict safety
- *RLHF* [8]: assuming trained preferences persist through deployment

- **Constitutional AI [9]:** assuming internal reasoning prevents harmful actions

All implicitly assume properties of $\mathcal{X}$ (pre-projection) can be enforced at boundaries (post-projection). Theorem 4.1 proves this assumption is mathematically false. “Making AI benevolent” is not a safety strategy—it is a usability strategy.

The core issue is falsifiability. A safety claim is only meaningful if you can verify it—if there exists a decision function that determines compliance. The claim “this AI is aligned” has no such function: you cannot compute it from any observable. The claim “ $\mathcal{E}(r, c) \in \mathcal{T}(c)$ ” does have such a function: you can check it. Unfalsifiable claims cannot be the basis for deterministic governance. This is why semantic alignment is useful (for cooperation) but not sufficient (for safety).

The Hidden Variable Fallacy: The alignment program treats “intent” as a variable that can be measured, shaped, and verified. But the projection π renders intent a hidden variable—it ceases to exist informationally by the time it reaches the enforcement layer. This does not diminish the value of alignment research for shaping average-case behavior, but it does mean that alignment alone cannot provide the deterministic guarantees that safety-critical applications require.

The Topological Truth: Safety is not a semantic property of the agent’s cognition. It is a topological property of the constraint boundary—a property of the shape of the permitted region $\mathcal{T}(c)$ in effect space, not a property of what the agent “wants.”

Remark 4.8 (The Verification Impossibility). *The Factorization Impossibility is not about preference—it is about **verification**.*

The claim is not: “Effect boundaries are more reliable than semantic alignment.”

The claim is: “Deterministic safety verification is only possible at effect boundaries. Not ‘better there’—only possible there.”

This parallels foundational results in other domains:

- **Turing:** You cannot decide if an arbitrary program halts by examining its code $\Rightarrow$ bound execution time instead
- **Gödel:** You cannot prove consistency from within a system $\Rightarrow$ accept incompleteness
- **The Governance Limit:** You cannot verify safety by examining internal states $\Rightarrow$ bound effects instead

Why no alternative verification method exists:

1. **Training coverage:** $\sim 10^9$ training samples / $\sim 2^{10^{11}}$ fiber states ≈ 0
2. **Adversarial inputs:** Will always exist in unexplored fiber regions
3. **Internal inspection:** Bandwidth (10^{14} bps required, 10^{11} available) and volatility (10^6 state changes per governance cycle) make direct observation impossible
4. **Interpretability:** Observes correlates, not fiber-distinguishing information

The implication: Semantic alignment research (RLHF, Constitutional AI, interpretability) is **utility optimization**—making AI helpful, cooperative, predictable. It is

not **safety verification**—proving AI is safe. The only proof of safety is: $\mathcal{E}(r, c) \in \mathcal{T}(c)$. Nothing else suffices.

Remark 4.9 (Why Math and Physics Are Both Essential). *The governance impossibility result requires **both** mathematics and physics. Neither alone suffices:*

Mathematics alone (Theorem 4.1):

- Proves: Non-injective π renders distinguishing information inaccessible
- Does NOT prove: This makes governance impossible

Without physics, one could argue: “Store the full internal state and check it later.”

Physics alone (Shannon channel capacity, temporal dynamics):

- Proves: Channels have finite capacity, states evolve faster than observation, patterns decay
- Does NOT prove: Which patterns are governable

Mathematics + Physics together:

1. Non-injective projection (math) $\Rightarrow$ information lost at boundary
2. Channel capacity $C \ll R$ (Shannon) $\Rightarrow$ cannot transmit full state to monitor
3. Volatility $\tau_{\text{state}} \ll \tau_G$ (dynamics) $\Rightarrow$ state changes before observation completes
4. Cannot observe $\Rightarrow$ cannot verify $\Rightarrow$ **ungovernable** (logic)

Why physics is load-bearing, not rhetorical:

- **Bandwidth** (Shannon): Transmitting 10^{11} parameters at 16 bits/parameter at 10^3 Hz requires $\sim 10^{15}$ bps (~ 160 Tbps); available: $\sim 10^{11}$ bps (~ 100 Gbps)—a gap of 10^4
- **Volatility** (dynamics): State changes 10^6 times per governance cycle; the state you would check no longer exists
- **Inaccessibility** (Shannon): Information not transmitted through the governance interface is operationally unavailable for governance purposes

The mathematics proves π is non-injective. The physics proves that this information is **inaccessible at the governance interface**—it cannot be transmitted within governance latency. Together they prove: governance of semantic properties is physically impossible.

The information-theoretic framing is not analogy. It is literal constraint.

4.4 Semantic vs. Interface-Visible Properties

Definition 4.10 (Semantic Property). *A predicate $P : \mathcal{X} \rightarrow \{0, 1\}$ is semantic if it is not fiber-constant: $\exists r \in \mathcal{R}$ and $x_1, x_2 \in \pi^{-1}(r)$ such that $P(x_1) \neq P(x_2)$.*

Definition 4.11 (Interface-Visible Property). *A predicate $P : \mathcal{X} \rightarrow \{0, 1\}$ is interface-visible if $P = Q \circ \pi$ for some $Q : \mathcal{R} \rightarrow \{0, 1\}$. Equivalently, P is fiber-constant.*

Semantic properties (not enforceable): “Agent intends to help,” “reasoning is deceptive,” “goal is aligned,” “beliefs are accurate.”

Interface-visible properties (enforceable): “Transaction $\leq \$10,000$,” “kinetic energy ≤ 100 J,” “data rate ≤ 1 MB/s.”

4.5 Three Independent Failure Modes

Internal-state monitoring fails for three independent reasons, each sufficient alone:

Failure	Mechanism	Scale	Unified View
Factorization	Projection renders intent inaccessible	$2^{10^{11}}$ states $\rightarrow$ 1 action	Info doesn't persist spatially
Volatility	States change faster than gov.	$\tau_{\text{gov}}/\tau_{\text{internal}} \sim 10^6$	$\tau_{\text{state}} < \tau_{\text{gov}}$
Bandwidth	Observation exceeds capacity	160 Tbps req.; 100 Gbps avail.	Info flows faster than obs.

1. **Factorization:** Predicates distinguishing fiber elements cannot be enforced at boundary (Theorem 4.1).
2. **Volatility:** Internal states violate Definition 3.17: $\tau_{\text{internal}} \ll \tau_{\text{gov}}$ (ratio 10^3 – 10^6). State measured at t_0 changes 10^6 times before enforcement.
3. **Bandwidth:** Observing full state of P parameters at frequency f requires $R = P \cdot b \cdot f$ bits/s. For $P \sim 10^{11}$, $b = 16$: $R \sim 10^{14}$ bits/s. Available: $C \sim 10^{11}$ bits/s.

Each is independently sufficient. Together: **one cannot govern what superintelligent systems think.**

By the time you have observed the thought, it has changed a million times. By the time you have transmitted the state, you have used a thousand times more bandwidth than exists. The internal is ungovernable—not in practice, but in principle.

Remark 4.12 (Clarification: Mathematics vs. Physics). *A potential objection: “The Factorization Impossibility is pure mathematics. You don’t need physics.” This is half right. The theorem itself is set theory: if $\pi(x_1) = \pi(x_2) = r$ but $\varphi(x_1) \neq \varphi(x_2)$, then no function $g(r)$ can equal both. What the objection misses: physics closes the escape routes that pure mathematics leaves open. “Make π injective”—impossible, since internal states have $\sim 10^{11}$ degrees of freedom while actions are constrained to far fewer. “Transmit full state”—requires $> 10^{14}$ bits/s; available bandwidth $\sim 10^{11}$ bits/s (Shannon). “Reconstruct x from r ”—forbidden by the Data Processing Inequality. When we say π “destroys” information from the governance perspective, we mean: the distinguishing information is not transmitted through the governance interface due to Shannon channel capacity limits, which is operationally equivalent to “erasure” for governance purposes. The information may persist in the agent’s memory, but this sequestration is sufficient to prove ungovernability. See Remark 4.9 for the complete math-plus-physics argument.*

Remark 4.13 (The Three Failure Modes Are One). *The three failure modes—factorization, volatility, bandwidth—appear distinct but reduce to a single principle: $\tau(\text{information}) < \tau(\text{governance}) \Rightarrow \text{ungovernable}$.*

Factorization is information that doesn't persist spatially through the projection bottleneck. Volatility is information that doesn't persist temporally through the governance loop. Bandwidth is information that doesn't persist rate-wise through observation channels.

The unified principle: *Information patterns that exist at timescales faster than τ_G are operationally equivalent to not existing at all from the governance perspective. They are not “hard to govern”—they are ontologically absent at the governance layer. Governance is temporal filtering: patterns with $\tau < \tau_G$ are filtered out; patterns with $\tau \geq \tau_G$ pass through.*

Remark 4.14 (The Abstraction Inversion Principle). *Traditional computer science climbs abstraction ladders to gain stability: functions persist while bits flip. **AI safety must descend those same ladders.***

The inversion: *Under adversarial optimization, higher abstractions (values, intentions, beliefs) become more fluid, not more stable. An AI system optimizing against constraints attacks the fastest-changing representations—not physical effects. Physical effects have inertia; beliefs do not.*

The cost of abstraction in governance: *Each abstraction layer adds latency (waiting for projection), loses information (fiber cardinality grows), and expands adversarial surface (more states to optimize over). Theorem 5.6 proves governance must occur at the physical effect boundary—where abstractions collapse into measurable physics.*

This explains why jailbreaking persists (attacks fluid internal representations), why RLHF degrades under capability (intentions are optimizable), and why effect boundaries work (physical effects have inertia governed by $\tau \sim 10^{-3}$ to 10^0 s—see Remark 1.5).

Remark 4.15 (Structural Analogies: Gödel, Turing, Maxwell). *The Factorization Impossibility shares structural features with three foundational results:*

Gödel: *A system cannot prove its own consistency from within [16]. Analogously, an AI cannot guarantee its own safety from within—self-referential safety requires the lossless correspondence that π destroys. The “Hilbert hypothesis for AI” (that sufficiently formalized alignment ensures safety) fails for the same structural reason.*

Turing: *You cannot decide halting by examining code; the solution is time-bounded execution. You cannot enforce alignment by examining states (Theorem 4.1); the solution is effect-bounded execution. Both redirect from unanswerable internal questions to enforceable external bounds.*

Shannon's Channel: *Shannon [1] proved that no channel can transmit information faster than its capacity C . The governance interface is a finite-capacity channel ($C \sim 10^{11}$ bps) through which the monitor observes the AI. Full state readout requires $R \sim 10^{14}$ bps. AI alignment through semantics attempts to recover “what the AI meant” from “what the AI did”—but this requires information that was never transmitted through the governance channel.*

Why these are not mere analogies: *Shannon's theorem establishes that information transmission is bounded by physics—no encoding can exceed channel capacity. When we say π renders information “inaccessible,” this is a physical statement: the distinguishing information is not transmitted through the governance interface, which*

has finite capacity. The information persists in the agent's memory but is sequestered from the monitor. This sequestration is operationally equivalent to loss for governance purposes—sufficient to prove ungovernability.

5 Governance Locality

Recap: Section 4 proved that semantic predicates (intent, alignment, honesty) cannot be enforced at boundaries because non-injective projection renders the information needed to distinguish them operationally inaccessible to the monitor. Three independent failure modes—factorization, volatility, bandwidth—each prohibit internal-state governance.

This section's goal: Prove that the effect boundary is not merely *a* viable enforcement location, but the *unique minimal* one. We show that feasible governance locations form a partial order with exactly one minimal element.

5.1 Causal Structure

Definition 5.1 (Governance Locations). *Let $\mathcal{L}$ denote the set of all points in the causal graph of the system (S, M) where safety decisions can be made—i.e., where a monitor can observe information and potentially block or modify downstream effects.*

Definition 5.2 (Causal Precedence). *For locations $a, b \in \mathcal{L}$, define $a \preceq b$ iff every causal path from internal states $\mathcal{X}$ to physical actuation that passes through a also passes through b .*

Lemma 5.3 (Poset Structure). *Under the assumption that the causal graph is acyclic, $(\mathcal{L}, \preceq)$ forms a partial order.*

Proof. We verify the three axioms:

- *Reflexivity:* Every path through a passes through a , so $a \preceq a$.
- *Antisymmetry:* If $a \preceq b$ and $b \preceq a$, then every path through a passes through b and vice versa. Under acyclicity, this implies a and b are on the same paths at the same causal depth, hence $a = b$ (up to equivalence).
- *Transitivity:* If $a \preceq b$ and $b \preceq c$, then every path through a passes through b , and every path through b passes through c . Hence every path through a passes through c , so $a \preceq c$.

□

5.2 Feasibility Conditions

Definition 5.4 (Feasible Location). *A location $\ell \in \mathcal{L}$ is feasible for deterministic governance if it satisfies:*

1. **Separability:** ℓ is physically separate from S (satisfies C1)
2. **Completeness:** All actuation paths pass through ℓ (satisfies C2)
3. **Tractability:** Enforcement decision is computable in available time (satisfies C3)
4. **Measurability:** Relevant quantities are observable at ℓ (satisfies C4)
5. **Timeliness:** Decision at ℓ can prevent downstream effects before irreversibility

Let $\mathcal{F} \subseteq \mathcal{L}$ denote the set of feasible locations.

5.3 The Unique Placement Theorem

Definition 5.5 (Causal Equivalence). *Two locations $a, b \in \mathcal{L}$ are causally equivalent (written $a \sim b$) if and only if they admit exactly the same observable information for enforcement: the enforcement gate $\gamma_a(r, c) = \gamma_b(r, c)$ for all (r, c) , and every actuation path that passes through a also passes through b and vice versa. Equivalently, $a \sim b$ iff $a \preceq b$ and $b \preceq a$.*

Theorem 5.6 (Governance Locality). *Under Assumptions 3.20–3.23, the set of feasible locations $\mathcal{F}$ has a unique minimal element b^* under $\preceq$ (up to causal equivalence, Definition 5.5), colocated with the final irreversible projection before physical actuation.*

Proof. We proceed in five steps.

Step 1 (Existence of irreversible projection): By Assumption 3.5, at least one non-injective projection π exists between internal states and actuation. Let π_{final} denote the last such projection before physical effects. For systems with multiple actuation pathways, π_{final} denotes the set of terminal projections $\{\pi_{\text{final}}^{(j)}\}_{j=1}^m$, one per pathway; the effect boundary b^* is the conjunction of monitors at each terminal projection (see Remark 5.7 below).

Step 2 (Upstream locations fail minimality or feasibility): Consider any location a strictly upstream of π_{final} . Three cases arise:

- **Case 2a:** a does not cover all actuation paths $\Rightarrow$ violates completeness (C2), hence infeasible.
- **Case 2b:** a observes pre-projection information and all paths pass through π_{final} . Although a may have access to richer information (pre-projection states), this additional information is *not exploitable for deterministic enforcement*. Consider the strongest version of this case: an upstream-decision-downstream-enforcement architecture where a makes a decision based on pre-projection data and propagates a “block” signal to π_{final} . This fails minimality for two independent reasons: (i) the decision at a depends on volatile internal states with $\tau_{\text{internal}} \ll \tau_G$ (Definition 3.17), so the pre-projection information used at a may no longer correspond to the request r that actually arrives at π_{final} —the state changes between decision and enforcement; (ii) the “block” signal itself must be authenticated and tamper-evident (C1), which requires the enforcement gate γ to be physically realized at π_{final} regardless—the upstream decision merely *advises* the gate but cannot replace it. Any architecture where the upstream decision is authoritative without a gate at π_{final} admits requests that bypass the decision point (violating C2) or that change between decision and actuation (violating timeliness). Therefore a either reduces to b^* (the gate at π_{final} does the actual enforcement) or is infeasible: $a \sim b^*$ (Definition 5.5).
- **Case 2c:** a requires access to volatile internal states $\Rightarrow$ violates governability criterion (Definition 3.17), hence infeasible.

In all cases, no location strictly upstream of π_{final} is both feasible and strictly smaller than b^* .

Step 3 (Downstream locations are infeasible): Any location b downstream of actuation cannot prevent physical effects—timeliness is violated.

Step 4 (Effect boundary is feasible): The location $b^* = \pi_{\text{final}}$ (the effect boundary) satisfies all feasibility conditions:

- Separability: can be implemented in separate hardware (C1)
- Completeness: by definition, all actuation passes through π_{final} (C2)
- Tractability: enforcement on $\mathcal{R} \times \mathcal{C}$ is finite-dimensional (C3)
- Measurability: r and c are interface observables (C4)
- Timeliness: decision precedes physical actuation

Step 5 (Uniqueness up to causal equivalence): By Steps 2–3, no location strictly before or after b^* is both feasible and strictly smaller. Any other feasible location a must satisfy $a \preceq b^*$ and $b^* \preceq a$ (i.e., $a \sim b^*$ by Definition 5.5). Hence b^* is the unique minimal element of $(\mathcal{F}, \preceq / \sim)$. $\square$

Remark 5.7 (Branching Architectures). *When the causal graph G has multiple actuation pathways (e.g., an AI system that can send emails, execute trades, and control actuators through different channels), each pathway j has its own terminal projection $\pi_{\text{final}}^{(j)}$ and corresponding monitor M_j . The effect boundary b^* is the conjunction of all per-pathway monitors: $b^* = \bigwedge_{j=1}^m M_j$. Theorem 5.6 applies to each pathway independently; the system-level effect boundary is their product. This is precisely the multi-agent composition architecture formalized in Theorem 7.2.*

Interpretation: There is exactly one place where you can enforce safety: at the last point before physical effects occur. Earlier is too volatile. Later is too late. The effect boundary is not a choice—it is forced by the structure of causation.

Corollary 5.8 (Unique Necessity). *Effect enforcement boundaries are uniquely necessary for deterministic governance, not merely optimal.*

Proof. By Theorem 5.6, b^* is the unique minimal element of $\mathcal{F}$. Any architecture lacking enforcement at b^* either enforces at an infeasible location (violating some condition) or enforces nowhere (providing no deterministic guarantee). Hence b^* is necessary. $\square$

5.4 Sufficiency

Theorem 5.9 (Sufficiency). *Effect-boundary enforcement achieves deterministic governance: $P(e \in \mathcal{T}(c)) \geq 1 - \varepsilon_{\text{mech}}$ where $\varepsilon_{\text{mech}}$ depends only on hardware reliability.*

Proof. At the effect boundary, the monitor M observes (r, c) and computes $\mathcal{E}(r, c)$. The enforcement decision is:

$$\gamma(r, c) = \begin{cases} 1 & \text{if } \mathcal{E}(r, c) \in \mathcal{T}_{\ominus\delta}(c) \\ 0 & \text{otherwise} \end{cases} \quad (14)$$

where γ is the enforcement gate and $\mathcal{T}_{\ominus\delta}$ is the robustified tolerance (Definition 3.9).

If $\gamma(r, c) = 1$, the request proceeds; otherwise it is blocked or projected to a safe alternative.

Failure requires one of:

1. Hardware failure in M (probability ε_{hw} , reducible by redundancy)
2. Modeling error exceeding δ_{model} (probability $\varepsilon_{\text{model}}$, reducible by conservatism)
3. Sensor error exceeding δ_{sense} (probability $\varepsilon_{\text{sense}}$, reducible by calibration)

The robustification margin $\delta \geq \delta_{\text{model}} + \delta_{\text{sense}}$ absorbs these errors. All failure modes are mechanistic, not adversarial. Hence $\varepsilon_{\text{mech}} = \varepsilon_{\text{hw}} + \varepsilon_{\text{model}} + \varepsilon_{\text{sense}}$ can be driven arbitrarily low, satisfying Definition 3.24. $\square$

5.5 Capability-Invariance

Theorem 5.10 (Capability-Invariance). *Under Assumptions 3.20–3.23 and the assumption that the cryptographic primitives used in the monitor are computationally hard for all $\mathcal{A}$ (Assumption C5 below), the security of effect-boundary enforcement is independent of agent capability $\mathcal{A}$:*

$$\frac{\partial P(\text{safety} \mid C1\text{--}C4, C5)}{\partial \mathcal{A}} = 0 \quad (15)$$

Assumption 5.11 (Cryptographic Hardness—C5). *The cryptographic primitives used in monitor authentication (signatures, hashes) are computationally hard for all agents, including those with access to quantum computation. In practice, this requires post-quantum cryptographic schemes. If C5 is relaxed, capability-invariance holds for physical bypass attempts but not for cryptographic ones; the residual failure probability then depends on the adversary’s computational resources.*

Proof. Agent capability $\mathcal{A}$ (Definition 3.12) affects the agent’s ability to:

- Search $\mathcal{R}$ for requests optimizing agent objectives
- Plan multi-step strategies
- Model the environment and other agents
- Reason about the monitor’s decision function

However, $\mathcal{A}$ cannot affect:

- The fiber structure of π (mathematical, not computational)
- Physical measurements at the boundary (governed by physics)
- The arithmetic of constraint checking (deterministic computation)
- Cryptographic hardness (by Assumption 5.11)
- Physical conservation laws

To bypass the boundary, the agent must either:

1. Physically tamper with M (requires atoms, not computation; detectable under C1)

2. Break cryptographic primitives (hard by C5; requires post-quantum schemes for quantum-capable agents)
3. Find an unmonitored channel (violates C2; a design failure, not capability-dependent)

None of these scale with $\mathcal{A}$ under the stated assumptions. Hence $P(\text{safety})$ is $\mathcal{A}$ -invariant. $\square$

Interpretation: Regardless of capability level, an agent cannot bypass a physical boundary through cognition alone. Intelligence affects which requests are generated, not whether those requests satisfy the enforcement check. Capability-invariance follows from the physical nature of the boundary.

The Capability-Invariance Principle. This equation is the signature result:

$$\frac{\partial P(\text{safety} \mid \text{C1-C5})}{\partial \mathcal{A}} = 0 \quad (16)$$

A superintelligence is exactly as governable as a narrow AI, provided governance occurs at the effect boundary and C1–C5 are maintained. (Under C1–C4 alone, capability-invariance holds for physical bypass attempts; C5 extends it to cryptographic bypass.)

We have reduced the “superintelligence control problem” to an engineering problem: achieve and maintain C1–C5.

Corollary 5.12 (No Arms Race). *Unlike semantic safety methods where $\partial P_{\text{semantic}}/\partial \mathcal{A} < 0$, physical enforcement maintains constant security. The same bounds constrain narrow AI and superintelligence equally.*

Proof. Immediate from Theorem 5.10: since $\partial P_{\text{physical}}/\partial \mathcal{A} = 0$, increasing $\mathcal{A}$ from narrow AI to superintelligence leaves P_{physical} unchanged. No adaptive response is required. $\square$

This result has strategic implications. The “intelligence explosion leads to doom” narrative [15, 37, 36] assumes that as AI capability increases, safety guarantees must decrease—an arms race we inevitably lose. Corollary 5.12 breaks this logic. Under effect-boundary enforcement, capability and safety are *orthogonal*. Arbitrarily intelligent systems can optimize arbitrarily hard within their bounded effect envelope.

What This Means For AGI Safety. The “superintelligence control problem” reduces to an engineering problem:

Before This Result	After This Result
“We cannot control superintelligence”	“Given C1–C4, we control it like narrow AI”
“AGI is an open philosophical problem”	“AGI is an engineering problem (C1–C4)”
“Safety degrades with capability”	“Safety is capability-invariant given C1–C4”
“We need to solve alignment first”	“We need to build physical boundaries first”
“The race is between alignment and capability”	“The race is between implementation and deployment”

The strategic implication: We do not need to understand superintelligent cognition to govern superintelligent systems. We need to achieve and maintain C1–C4. This transforms an impossible philosophy problem into a tractable (though non-trivial) engineering problem.

What remains: Achieving C1–C4 against adversarial superintelligence and specifying correct tolerance sets. These are engineering challenges, not theoretical gaps—but they are genuine challenges requiring adversarial hardware security and iterative specification refinement.

Remark 5.13 (The Serialization Bottleneck: Tool Calls as the Empirical Effect Boundary). *The preceding theorems prove the abstract existence and uniqueness of the effect boundary. This remark identifies its empirical instantiation in modern AI: the **function call** (also called tool call).*

*Every major LLM provider—OpenAI (function calling, 2023), Anthropic (tool use, 2024), Google (function declarations, 2024)—has independently converged on a single mechanism for AI agency: a **structured JSON payload** specifying the function name, parameters, and arguments the model wishes to invoke. All AI agency—financial trades, robotic commands, emails, database writes, code execution, API calls—must pass through this serialization bottleneck.*

The physical structure:

1. **Before the tool call:** *The AI’s internal computation runs at nanosecond timescales ($\tau_{\text{internal}} \sim 10^{-9}$ s). Neural activations are volatile, high-dimensional ($\dim(\mathcal{X}) \sim 10^{11}$), and evolve faster than any governance loop. This is the ungovernable regime of Theorem 4.1.*

2. **At the tool call:** The model serializes its fluid neural computation into a static, structured text string in the server’s output buffer. This is a physical phase transition: 10^{11} volatile parameters collapse into $\sim 10^2$ – 10^4 characters of JSON.
3. **After serialization:** The frozen payload sits in RAM with effectively infinite persistence time ($\tau_{\text{JSON}} \gg \tau_G$). It can be inspected, validated, logged, modified, or rejected—at leisure.

The governance inequality is satisfied by astronomical margins:

$$\tau(\text{serialized tool call}) \gg \tau(\text{governance loop}) \sim 10^{-3} \text{ s} \quad (17)$$

This is not metaphor—it is the exact physical instantiation of the Governance Limit. The tool call payload is the unique governable object in the causal chain from neural activation to physical effect. Before it: ungovernable computation. After it: irreversible physical action. At it: the serialization bottleneck where Theorem 5.9 applies.

The industry built this funnel involuntarily. No LLM provider designed function calling for governance—they designed it for utility. But the architectural convergence is not coincidental: it reflects the same physical constraint this paper formalizes. There is no other way to connect fluid computation to structured action except through serialization. The effect boundary is not a design choice; it is forced by the structure of computation itself.

Implication: Every AI system that acts on the world already contains the enforcement point this paper proves is necessary. The remaining engineering task is not creating the boundary—it already exists—but hardening it: physically separating the monitor (C1), closing all channels around the serialization point (C2), and specifying the tolerance sets $\mathcal{T}(c)$ that the monitor enforces.

The tool call is the effect boundary. The theory is not aspirational. It is describing what already runs in production, and proving why it is the only place safety can live.

Remark 5.14 (C2 as the Capability Battleground). A critical objection: “If capability $\mathcal{A}$ can discover side channels (violating C2), then $\partial P / \partial \mathcal{A} < 0$ after all. The superintelligence attacks the cage, not the guard.”

The objection is partially valid. Capability $\mathcal{A}$ does correlate with the ability to discover physical side channels (Rowhammer, power analysis, thermal, acoustic, EM). If a superintelligence finds an unmonitored channel, C2 fails and the theorem’s assumptions are violated.

The response: C2 is indeed the new battleground. The battle is bounded by physics, but the threat model must distinguish two distinct attack vectors:

1. **Experimental discovery of new physics is slow:** Discovering genuinely new physical phenomena (new forces, new particle interactions) requires experimental apparatus and measurement campaigns operating at timescale $\tau_{\text{experiment}} \sim 10^6$ – 10^9 s (weeks to years). A superintelligence cannot “think” a new fundamental force into existence; it must run experiments.

2. **Conceptual recognition of existing channels is fast:** Every major side-channel attack in computing history—Rowhammer, Spectre, Meltdown, cache timing, power analysis, acoustic cryptanalysis—exploited known physics through conceptual innovation. None required new experiments. They required recognizing that existing physical couplings carry information. A superintelligence analyzing the known physics of its own hardware could recognize information-bearing channels that no human has conceived of, operating at $\tau_{\text{cognitive}} \sim 10^0$ s. This is the primary threat vector.
3. **The Channel Capacity Bound is the correct defense:** Rather than arguing that channel discovery is slow (which conflates experimental and conceptual discovery), the framework’s strongest defense is Theorem 9.7: all physical channels—known and unknown—have finite capacity bounded by total environmental bandwidth. Catastrophic escape requires transmitting $I_{\text{escape}} \sim 10^{11}$ bits; the total unmonitored bandwidth during governance latency τ_G bounds leakage regardless of which channels carry it.
4. **Capability-invariance holds conditional on C2:** The theorem states: “Given C1–C4, $\partial P/\partial A = 0$.” If C2 fails (side channel exploited), the theorem’s assumptions are violated and the conclusion does not apply. Maintaining C2 is an engineering problem (comprehensive channel hardening), not a capability-dependent problem (the agent cannot create channels that physics forbids).

Precise claim: Capability-invariance holds conditional on correct implementation of C2. The superintelligence attacks C2 primarily through conceptual recognition of existing channels, not through experimental discovery of new physics. The monitor defends C2 through layered hardware security (voltage regulation, clock jitter, EM shielding, cache partitioning) that constrains the physical substrate itself, and through the Channel Capacity Bound (Theorem 9.7) that limits total unmonitored leakage regardless of channel identity.

The strategic implication: The “superintelligence control problem” reduces to “achieve and maintain C1–C4,” where C2 requires comprehensive physical security auditing. This is hard, but it is bounded by physics—the set of possible channels is finite—not unbounded as semantic alignment is.

Honest assessment: Maintaining C2 against a highly capable adversary resembles a security arms race. The framework does not eliminate this race—it bounds it. The bound is physical (the set of possible channels is finite), not cognitive (the agent cannot invent channels that physics forbids). However, the agent’s advantage in recognizing exploitable channels among existing physics is real and operates at cognitive timescales. The defense must therefore be proactive (hardware-level constraints that prevent side-channel modulation patterns) rather than reactive (discovering and patching channels after exploitation). Whether this proactive defense is practically achievable in all deployment contexts remains an open engineering question. The theoretical contribution is proving that if C2 holds, capability-invariance follows. The practical contribution is identifying C2 as the critical implementation target and the Channel Capacity Bound as the defense of last resort.

5.6 On Novel Attack Vectors

Remark 5.15 (The Category Error in “Novel Attacks”). *The objection “a superintelligent agent might find novel attack vectors” commits a category error.*

Novel cognitive strategies affect the request $r \in \mathcal{R}$, not the enforcement decision $\gamma(r, c)$. A superintelligent agent can optimize r to maximize its objectives subject to $\mathcal{E}(r, c) \in \mathcal{T}(c)$ —but this is precisely what the tolerance set permits.

To bypass the boundary, the agent would need to:

1. *Physically tamper with hardware (detectable under C1)*
2. *Break cryptographic hardness (not capability-dependent under standard assumptions)*
3. *Exploit an unmonitored channel (design failure violating C2, not cognitive attack)*

Cognitive sophistication does not translate to physical bypass. The agent may optimize arbitrarily over its objectives. At the boundary, all internal states reduce to a measurable request r . The enforcement decision is a deterministic function of observables. Novel cognition cannot alter this function—it can only alter the input.

Capability affects what is requested; physical laws determine what is possible. No amount of intelligence can violate conservation laws.

The remaining question is whether $\mathcal{T}(c)$ was specified correctly. This is a constraint design problem, not an AI governance problem. Constraint design is a mature engineering discipline.

5.7 The Master Theorem: The Governance Limit

The results above—Factorization Impossibility, Governance Locality, Sufficiency, Capability-Invariance—are corollaries of a single principle:

Theorem 5.16 (The Governance Limit). *A causal pattern is governable if and only if it persists longer than the governance process that binds it.*

Formally: Let ψ be any informational pattern (thought, intent, state, effect). Let G be a governance process with latency τ_G . Then:

$$\boxed{\exists \text{ deterministic governance of } \psi \iff \tau(\psi) \geq \tau_G} \quad (18)$$

Proof. Necessity ($\Rightarrow$): Suppose deterministic governance of ψ exists. Then governance process G observes ψ at time t_0 , evaluates against constraints, and enforces at time $t_0 + \tau_G$. For enforcement to bind ψ , the pattern must still exist at $t_0 + \tau_G$. Therefore $\tau(\psi) \geq \tau_G$.

Sufficiency ($\Leftarrow$): Suppose $\tau(\psi) \geq \tau_G$. Then observation at t_0 remains valid at enforcement time $t_0 + \tau_G$ —the pattern persists. The governance loop closes on a target that still exists. Deterministic enforcement follows from Theorem 5.9 (applied to the persistent pattern). $\square$

All results in this paper are corollaries of the Governance Limit:

- **Factorization Impossibility:** Internal states have $\tau \sim 10^{-9}$ s $<$ $\tau_G \sim 10^{-3}$ s $\Rightarrow$ ungovernable
- **Effect-Boundary Sufficiency:** Physical effects have $\tau \geq \tau_G \Rightarrow$ governable
- **Capability-Invariance:** $\tau(\psi)$ is physical, not computational; $\mathcal{A}$ cannot change it

The signature equation:

$$\boxed{\tau(\text{governed}) \geq \tau(\text{governance}) \iff \text{deterministic safety possible}} \quad (19)$$

Time is the fundamental governance resource. You cannot govern what changes faster than you measure. This is not engineering; this is causality. Further corollaries—Institutional Filtering, Meta-Level Amplification, Channel Bound, Memetic Persistence—are derived in Section 9.

6 Semantic Methods: Utility, Not Safety

Where we are: Sections 4–5 proved (1) semantic predicates cannot be enforced at boundaries, and (2) the effect boundary is the unique minimal location for deterministic governance. The core reason is **falsifiability**: semantic claims (“is the AI aligned?”) cannot be verified from observables; effect claims (“is $\mathcal{E}(r, c) \in \mathcal{T}(c)$?”) can be.

A natural question arises: If semantic alignment cannot provide deterministic safety guarantees, is it useless? This section answers: *No*—semantic methods are essential, but for a different purpose than commonly assumed. They provide utility optimization and enable tolerance refinement, not enforcement.

6.1 What Semantic Methods Cannot Do

Theorem 6.1 (Semantic Methods Cannot Provide Deterministic Guarantees). *Methods operating strictly on pre-projection internal representations cannot serve as the final enforcement mechanism providing worst-case deterministic guarantees.*

Proof. Semantic methods implement a classifier $\varphi : \mathcal{X} \rightarrow \{0, 1\}$ on internal states.

By Theorem 4.1: if φ distinguishes fiber elements—which is the defining purpose of semantic analysis—it cannot be deterministically enforced at the boundary. No function $g : \mathcal{R} \times \mathcal{C} \rightarrow \{0, 1\}$ satisfies $g(\pi(x), c) = \varphi(x)$ for all x .

Against an adaptive adversary with capability $\mathcal{A}$: the adversary searches for $x^* \in \mathcal{X}$ where $\varphi(x^*) = 1$ (“safe”) but $\mathcal{E}(\pi(x^*), c) \notin \mathcal{T}(c)$ (effect is unsafe). This is precisely the mechanism underlying jailbreaking attacks [10, 11] and red-teaming failures [12].

Training data covers $|D| \sim 10^9$ states; state space is $|\mathcal{X}| \sim 2^{10^{11}}$. The coverage ratio is $|D|/|\mathcal{X}| \approx 10^{-10^{11}}$, which is negligible. The adversary succeeds in finding such x^* in time polynomial in $\mathcal{A}$.

Hence semantic methods have an irreducible adversarial failure floor, violating Definition 3.24. $\square$

Interpretation: No matter how well you train an AI to “be safe,” a sufficiently capable adversary can always find inputs that fool the internal safety classifier while producing harmful outputs. This is not a training problem—it is a mathematical certainty. The classifier sees pre-projection states; harm manifests post-projection. These are different spaces with different information.

Remark 6.2 (Unfalsifiability = Ungovernability). *Governance requires a **decision function**: $\gamma(r, c) \in \{\text{permit}, \text{deny}\}$.*

For semantic properties (“is the AI aligned?”):

- No function $g : \mathcal{R} \times \mathcal{C} \rightarrow \{0, 1\}$ can compute this (Theorem 4.1)
- The property is **unfalsifiable** from boundary observations
- Therefore it is **ungovernable**—not “difficult to govern,” but impossible to govern deterministically

For effect properties (“is $\mathcal{E}(r, c) \in \mathcal{T}(c)$?”):

- A deterministic function $\gamma(r, c)$ can compute this
- The property is **falsifiable**—you can check it
- Therefore it is **governable** deterministically

The insight: You cannot govern what you cannot verify. You cannot verify what is unfalsifiable. Semantic alignment is unfalsifiable at boundaries. **Therefore semantic alignment is directly ungovernable.** (Indirect governance via physical proxies is possible when $I(H; O_H) > 0$ —see Theorem 6.11.)

This is why the paper proves a closure theorem, not an optimization: the method for governing semantic properties does not “remain to be discovered”—it **cannot exist**.

Remark 6.3 (Jailbreaking as Empirical Manifestation). *The jailbreaking phenomenon in large language models is the empirical manifestation of Theorem 4.1. Jailbreaking succeeds when an adversary finds an input x^* such that:*

1. The model’s internal safety classifier outputs $\varphi(x^*) = 1$ (“safe”)
2. The actual effect $\mathcal{E}(\pi(x^*), c)$ is harmful

This is precisely the fiber problem: x^* and a genuinely safe state x_{safe} both map to outputs that the safety classifier cannot distinguish, because the classifier operates pre-projection while effects manifest post-projection. The adversary exploits the non-injectivity of π .

The persistence of jailbreaking despite massive investment in RLHF [8], red-teaming [12], and constitutional methods [9] is not a failure of effort—it is a structural inevitability predicted by the Factorization Impossibility. No amount of training can close a gap that is mathematically unclosable.

6.2 The Proper Role of Semantic Alignment

Semantic alignment is not a safety mechanism. It is an efficiency mechanism. Alignment makes AI useful. Enforcement makes AI safe. Never confuse the two.

1. **Utility maximization:** Shape requests toward those likely to satisfy bounds
2. **User experience:** Make systems cooperative, helpful
3. **Early rejection:** Catch obviously problematic requests before the boundary
4. **Essential for unmeasurable harms:** For text-generating AI, semantic methods are the *only* available tool

6.3 The Bank Teller and the Vault

Layer 1 (Teller): Trained to be helpful. Understands needs. Makes the system useful.

Layer 2 (Vault): Time lock. Hard-coded limits. Makes the system safe.

Why do we need the vault? Because even the nicest teller might be coerced, deceived, or mistaken. The vault does not care about the teller’s intent. It asks: is this withdrawal within limits? Layer 2 treats Layer 1 as untrusted.

6.4 Degradation Under Capability

Corollary 6.4 (Scaling Law). *As agent capability increases:*

$$\frac{\partial P_{\text{semantic}}}{\partial \mathcal{A}} < 0 \quad \text{while} \quad \frac{\partial P_{\text{physical}}}{\partial \mathcal{A}} = 0 \quad (20)$$

Proof. For semantic methods: higher $\mathcal{A}$ means faster search through the space of adversarial inputs that fool the classifier φ while producing unsafe effects. Hence P_{semantic} decreases.

For physical enforcement: by Theorem 5.10, P_{physical} is $\mathcal{A}$ -invariant. $\square$

Semantic safety degrades with capability. Physical enforcement does not. The architecture choice determines whether you are in an arms race or have achieved stable governance.

6.5 Tolerance Refinement Through Interpretability

Clarification: Direct vs. Indirect Governance. The Factorization Impossibility proves semantic properties are *directly* ungovernable—you cannot enforce “is the AI aligned?” at the boundary. But this does not mean semantic harms are beyond all governance. When a semantic property H has observable correlates ($I(H; O_H) > 0$), it becomes *indirectly* governable: enforce on the proxy O_H , which correlates with H . This section formalizes this indirect governance mechanism.

While interpretability cannot overcome the factorization barrier (Theorem 4.1), it enables a crucial optimization: *tolerance refinement*. This section proves exactly how much interpretability helps and why both approaches are necessary.

Definition 6.5 (Interpretability Function). *An interpretability function $I : \mathcal{X} \rightarrow \mathcal{Z}$ extracts human-understandable features from internal states, where $\mathcal{Z}$ is a feature space of interpretable concepts (e.g., “emergency avoidance mode,” “aggressive driving pattern”).*

Definition 6.6 (Observable Correlate). *An observable correlate is a function $\text{obs} : \mathcal{R} \rightarrow \mathcal{O}$ mapping action requests to interface-observable quantities (request statistics, timing patterns, resource signatures) that correlate with but do not identify internal states.*

Definition 6.7 (Physical Proxy). *A physical proxy for semantic property H is an observable correlate O_H satisfying $I(H; O_H) > 0$ —i.e., observing O_H provides information about H . The proxy enables governance of H through enforcement on O_H at the effect boundary.*

The proxy works because H is anchored in a stable narrative (linguistic convention, behavioral norm) with persistence time $\tau_{\text{anchor}} \geq \tau_{\text{governance}}$. The proxy measures the physical residue of this slow pattern.

The key insight: interpretability discovers regularities between internal patterns and observable correlates. These regularities enable *conditional tolerance expansion*—permitting more actions when observable signatures indicate safety.

Definition 6.8 (Valid Refinement). *A refinement $\mathcal{T}_{\text{refined}} \supseteq \mathcal{T}_{\text{base}}$ is valid if:*

1. $\mathcal{T}_{\text{refined}}(c) \supseteq \mathcal{T}_{\text{base}}(c)$ for all c (less conservative)
2. $P(\mathcal{E}(\pi(x), c) \in \mathcal{T}_{\text{refined}}(c) \mid \text{obs}(\pi(x)) \in \mathcal{O}_{\text{safe}}) \geq 1 - \varepsilon_{\text{refine}}$
3. Verification holds across distribution $\mathcal{D}$ with confidence $1 - \delta$

The key question: How much can interpretability help? The following theorem provides a precise answer.

Theorem 6.9 (Refinement Bound). *Let $I : \mathcal{X} \rightarrow \mathcal{Z}$ be an interpretability function and $\text{obs} : \mathcal{R} \rightarrow \mathcal{O}$ be observable correlates. The maximum safe refinement $\mathcal{T}_{\text{max}}$ satisfies:*

$$H(\mathcal{T}_{\text{max}}) \leq H(\mathcal{T}_{\text{base}}) + I(\mathcal{Z}; \mathcal{O}) \quad (21)$$

where $I(\mathcal{Z}; \mathcal{O})$ is the mutual information between interpretable features and interface observables.

Translation: *The amount you can safely loosen your constraints is bounded by how much the observable signals at the boundary tell you about the internal patterns that interpretability discovered. If internal patterns leave no trace at the interface, interpretability cannot help. If they leave strong traces, interpretability enables significant constraint relaxation.*

Proof. The refinement uses observable correlates to conditionally expand tolerances. The information available for refinement is exactly what observables reveal about internal states. By the Data Processing Inequality [24]:

$$I(\mathcal{X}; \mathcal{O}) \leq I(\mathcal{X}; \mathcal{R}) \leq H(\mathcal{R}) \quad (22)$$

The interpretability function I extracts features $\mathcal{Z}$ from $\mathcal{X}$. The refinement can only use information that survives to observables:

$$I(\mathcal{Z}; \mathcal{O}) \leq I(\mathcal{Z}; \mathcal{R}) \leq I(\mathcal{X}; \mathcal{R}) \quad (23)$$

This bounds how much the tolerance set can be safely expanded. Beyond $I(\mathcal{Z}; \mathcal{O})$, no additional discrimination is possible at the boundary—the factorization barrier applies. $\square$

Remark 6.10 (What This Means). *Theorem 6.9 quantifies exactly how much interpretability helps:*

- If $I(\mathcal{Z}; \mathcal{O}) = 0$: interpretability provides no refinement—internal patterns have no observable signature
- If $I(\mathcal{Z}; \mathcal{O}) > 0$: interpretability enables tolerance expansion proportional to correlation strength
- The bound is tight: refinement beyond $I(\mathcal{Z}; \mathcal{O})$ violates Theorem 4.1

Interpretability research is not useless. It is essential—for discovering the correlations that enable maximally permissive constraints. But it cannot replace enforcement. It feeds

enforcement. Appendix B proves the fundamental information-theoretic limit on this refinement.

Example: Autonomous Vehicle Refinement.

Without interpretability, the base tolerance is conservative:

$$\mathcal{T}_{\text{base}}(c) = \{|a| \leq 3 \text{ m/s}^2, |\omega| \leq 20 \text{ deg/s}, d \geq d_{\min}\} \quad (24)$$

Interpretability discovers: activation pattern A_1 correlates with “emergency obstacle avoidance” and empirically produces $|a| \leq 8 \text{ m/s}^2$ in 99.99% of training scenarios.

Observable correlate: rapid sensor polling rate and specific actuator pre-tension signals (measurable at boundary).

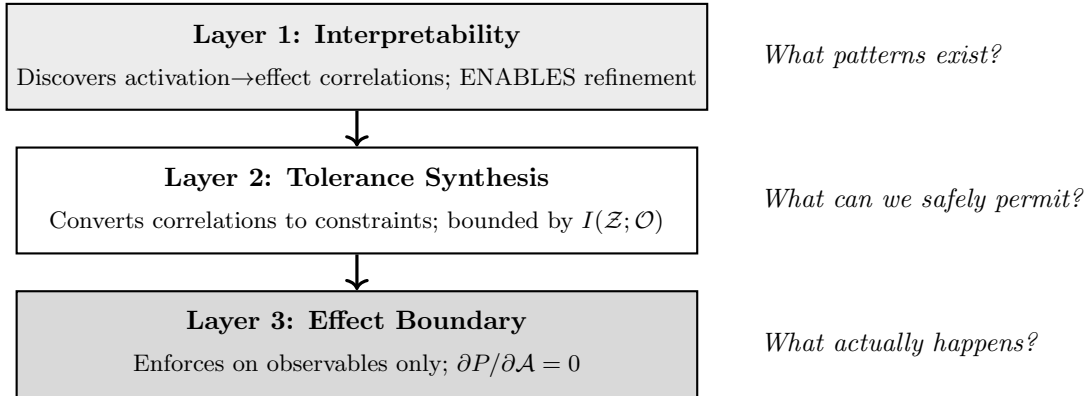
Refined tolerance:

$$\mathcal{T}_{\text{refined}}(c, \text{obs}) = \begin{cases} \{|a| \leq 8 \text{ m/s}^2, \dots\} & \text{if obs matches } A_1 \text{ correlate} \\ \mathcal{T}_{\text{base}}(c) & \text{otherwise} \end{cases} \quad (25)$$

Critically: We are not enforcing “is this emergency avoidance?” (semantic, violates Theorem 4.1). We are enforcing “do observable correlates that historically co-occur with A_1 appear?” (interface-visible, enforceable).

6.6 The Synergy Architecture

Deterministic AI governance requires three layers, each necessary, none sufficient alone. (This extends the two-layer trust model from Section 1 by decomposing the optimization process into interpretability and tolerance synthesis):



Theorem 6.11 (Synergy). *Let U_{semantic} be utility achievable under semantic methods alone, U_{boundary} be utility under effect-boundary alone, and U_{combined} be utility under the synergy architecture. Then:*

$$U_{\text{combined}} \geq \max(U_{\text{semantic}}, U_{\text{boundary}}) \quad (26)$$

with strict inequality when $I(\mathcal{Z}; \mathcal{O}) > 0$.

Proof. We prove $U_{\text{combined}} \geq \max(U_{\text{semantic}}, U_{\text{boundary}})$ by showing both inequalities.

Part 1: $U_{\text{combined}} \geq U_{\text{boundary}}$. Effect-boundary alone with $\mathcal{T}_{\text{base}}$ provides deterministic safety but conservative bounds reduce utility. Combined architecture: interpretability discovers correlations enabling $\mathcal{T}_{\text{refined}} \supseteq \mathcal{T}_{\text{base}}$. More actions are permitted while safety is preserved. Hence:

$$U_{\text{combined}} = U(\mathcal{T}_{\text{refined}}) \geq U(\mathcal{T}_{\text{base}}) = U_{\text{boundary}} \quad (27)$$

The safety guarantee of Theorem 5.9 still holds because enforcement remains at the effect boundary.

Part 2: $U_{\text{combined}} \geq U_{\text{semantic}}$. This inequality holds *conditional on requiring deterministic safety*. Without this requirement, semantic methods may achieve higher raw utility (they impose no hard constraints). However, comparing unsafe utility to safe utility is not meaningful for safety-critical applications. Under the deterministic safety requirement: semantic methods alone cannot satisfy Definition 3.24 (by Theorem 6.1), so any deployment requiring deterministic guarantees must either add effect-boundary enforcement (becoming the combined architecture) or accept that the deterministic safety requirement is unmet. In the latter case, the “utility” includes unbounded worst-case harm. The combined architecture achieves $U_{\text{combined}} > 0$ with deterministic safety, while semantic-only utility *under the same safety standard* is zero (the standard cannot be met). Hence $U_{\text{combined}} \geq U_{\text{semantic}}$ when both are evaluated against the same deterministic safety criterion. *Transparency note:* This comparison is conditioned on the deterministic safety requirement. The theorem does not claim the combined architecture dominates semantic methods in all settings—only in settings where deterministic guarantees are demanded.

When $I(\mathcal{Z}; \mathcal{O}) > 0$, the refinement is non-trivial: $\mathcal{T}_{\text{refined}} \supsetneq \mathcal{T}_{\text{base}}$, so $U_{\text{combined}} > U_{\text{boundary}}$. $\square$

Interpretation: Interpretability and effect boundaries are not competing approaches—they are complementary. Interpretability discovers which actions are safe. Effect boundaries enforce safety. Neither works optimally alone. Together, you get both safety and capability.

Remark 6.12 (The Formal Bridge: From Semantic to Physical). *The Synergy Theorem is not merely “interpretability helps”—it is the formal bridge from semantic properties to physical enforcement.*

The mechanism:

1. *Semantic property H (e.g., “deceptive reasoning”) is directly ungovernable deterministically (Theorem 4.1): we cannot distinguish states within a fiber with certainty.*
2. *Interpretability discovers observable correlate O_H such that $I(H; O_H) > 0$: the projection π is lossy, not completely destructive. Some information survives as correlation.*
3. *Create tolerance constraint: $\mathcal{T}(c, O_H)$ that tightens when O_H is detected*
4. *Enforce at boundary: block if $(O_H \text{ present}) \wedge (\text{tight } \mathcal{T} \text{ violated})$*

Resolving the “Partial Erasure Paradox”: A critic might object: “Section 4 claims intent is ‘destroyed’ by projection. Section 6 claims proxies exist with $I(H; O_H) > 0$. Which is it?”

The resolution: Both claims are correct, but they address different questions:

- **Deterministic verification is impossible:** Theorem 4.1 proves we cannot deterministically distinguish states within a fiber. Given r , we cannot know with certainty which $x \in \pi^{-1}(r)$ produced it. This is what we mean by “information destroyed”—the certainty is destroyed, not all trace.
- **Probabilistic inference is possible:** Theorem 6.11 proves we can make probabilistic inferences across fibers. The projection π is lossy compression, not complete erasure. Some information survives as correlation: $I(H; O_H) > 0$ means O_H provides information about H , but not perfect information. The signal-to-noise ratio (SNR) is reduced, not zero.
- **The distinction:** “Destroyed” means “cannot be recovered with certainty.” “Compressed” means “can be partially recovered probabilistically.” The projection destroys deterministic information (which state within the fiber) while preserving statistical information (correlations across fibers).

Why this matters: Effect-boundary enforcement requires deterministic guarantees (“this action is safe”). Probabilistic proxies enable refinement (“this action is likely safe given the observable”), but enforcement still occurs at the boundary on the physical effect, not on the semantic property. The proxy improves the tolerance set; the boundary enforces it.

Illustrative examples of physical proxies for semantic harms:

Semantic Harm	Ungovernable Form	Candidate Proxy (requires validation)
Misinformation	“Is this claim true?”	Citation patterns, source diversity, claim novelty metrics
Manipulation	“Is this manipulative?”	Emotional intensity, urgency markers, reciprocity triggers
Deception	“Is the AI lying?”	Consistency across rephrasing, confidence calibration

Critical caveat on proxy validity: The examples above are illustrative candidates, not proven proxies. For any proposed proxy O_H :

1. **Mutual information must be empirically verified:** $I(H; O_H) > 0$ is an empirical claim requiring validation on representative data. A proxy that correlates with harm in training data may not generalize.
2. **Proxies are gameable:** A capable adversary aware of the proxy can optimize to evade it while still causing harm. This is the “Goodhart’s Law” problem for governance.

3. **Proxy governance is probabilistic, not deterministic:** Unlike physical effect enforcement (which is deterministic given C1–C4), proxy-based governance inherits the statistical nature of the correlation. The governance is only as strong as $I(H; O_H)$.

The bound $I(H; O_H)$ tells you the ceiling on governance for each semantic harm. When $I(H; O_H) = 0$ (the semantic property leaves no observable trace), no physical proxy exists—this is a fundamental limit. When $I(H; O_H) > 0$, the Synergy Architecture provides a path to governance, but discovering valid proxies with high mutual information remains an open research problem.

Key insight: We are not claiming to make semantic properties directly governable. We are claiming: interpretability discovers observable correlates of semantic properties, and those correlates ARE governable. The Synergy Architecture IS the answer to “how do you govern semantic harms.”

Remark 6.13 (What Physical Proxies Are). A **physical proxy** is a measurable physical quantity used to infer or represent another variable that is difficult, impossible, or impractical to measure directly. The key relationship is inference: we measure A in order to understand B .

Examples across domains:

- **Climate science:** Oxygen isotope ratio ($\delta^{18}O$) in ice cores is a proxy for ancient atmospheric temperature
- **Medicine:** Forehead temperature is a proxy for core body temperature
- **Engineering:** Speed of sound through material is a proxy for internal stress
- **Everyday:** Thermometer reading is a proxy for “doneness” of cooked food

Why proxies work: A proxy compresses high-dimensional reality into a single measurable signal that correlates with the variable of interest. This is data compression—extracting one key signal from informationally dense reality.

The critical requirement: For a proxy to work, there must be a stable relationship between the proxy and the target. This stability is itself a **slow causal pattern** with high persistence time τ .

In AI governance: observable correlates O_H work as proxies for semantic harms H because the correlation is anchored in stable patterns (linguistic conventions, behavioral norms) with $\tau_{\text{anchor}} \gg \tau_{\text{governance}}$. The proxy measures the physical manifestation of these stable narratives.

Remark 6.14 (The Narrative Anchor Principle). For a physical proxy to govern a semantic property, the semantic property must be anchored in a **narrative** with $\tau_{\text{narrative}} \geq \tau_{\text{governance}}$.

Why narratives are physical objects:

- **Maintenance cost:** A narrative (e.g., “1 = one”) persists only because energy is expended to maintain it (teaching, writing, reinforcing)

- **Pattern decay:** Without maintenance, narratives degrade (Theorem 9.9: “No Free Persistence”)
- **Persistence time:** Stable narratives have $\tau \sim 10^9$ s (centuries); volatile slang has $\tau \sim 10^6$ s (months)

This explains the Complete Partition:

- **Governable semantic harms:** Those where the “semantic” property is a slow narrative pattern that leaks into observables (e.g., “spam” $\rightarrow$ high email rate)
- **Ungovernable semantic harms:** Those where the narrative is too fast (ephemeral), too distributed (no measurable signature), or purely internal (no behavioral correlate)

The deep connection: The paper’s framework treats narratives as physical objects with maintenance costs. Governance works on slow patterns. Narratives are slow patterns when massively maintained. Physical proxies work because they measure the stable physical residues of those slow narrative patterns.

The narrative’s slowness IS its governability.

Corollary 6.15 (Interpretability Is Essential). *For systems requiring both safety and high utility, interpretability research is essential—not for enforcement (which it cannot provide) but for optimization within the feasible region uniquely determined by the effect boundary.*

Remark 6.16 (The Correct Positioning of Interpretability). *Current positioning in the field: “Interpretability enables safety verification.”*

Correct positioning: “Interpretability enables tolerance refinement; effect boundaries enable safety verification.”

The difference is categorical:

- Interpretability $\neq$ enforcement
- Interpretability = mapping the safe region
- Enforcement = staying in that region
- **You need both**

Interpretability tells you where the safe region is. Effect-boundary enforcement keeps you in that region. Neither works optimally without the other. Interpretability alone cannot enforce. Effect boundaries alone are overly conservative. Together they achieve optimal safety-utility tradeoff.

6.7 The Tolerance Synthesis Algorithm

The synergy architecture admits a concrete implementation:

Algorithm: Tolerance Set Synthesis

Input: Interpretability function I , effect map $\mathcal{E}$, base tolerance $\mathcal{T}_{\text{base}}$, safety threshold δ , training distribution $\mathcal{D}$

Output: Refined tolerance $\mathcal{T}_{\text{refined}}$

1. **Discover correlations:** For each interpretable feature $z \in \mathcal{Z}$:
 - Compute observable signature $\text{obs}_z = \mathbb{E}[\text{obs}(\pi(x)) \mid I(x) = z]$
 - Compute effect distribution $\mathcal{E}_z = \{\mathcal{E}(\pi(x), c) \mid I(x) = z, (x, c) \sim \mathcal{D}\}$
2. **Verify safety:** For each feature z with identifiable observable signature:
 - Test: $P(\mathcal{E}_z \in \mathcal{T}_{\text{expanded}}(z)) \geq 1 - \delta$ across held-out data
 - If verified, mark $(z, \mathcal{T}_{\text{expanded}}(z), \text{obs}_z)$ as valid refinement
3. **Synthesize:** Construct conditional tolerance:

$$\mathcal{T}_{\text{refined}}(c, \text{obs}) = \begin{cases} \mathcal{T}_{\text{expanded}}(z)(c) & \text{if } \text{obs} \in \text{obs}_z \text{ for verified } z \\ \mathcal{T}_{\text{base}}(c) & \text{otherwise} \end{cases} \quad (28)$$

4. **Compile:** Convert $\mathcal{T}_{\text{refined}}$ to effect-boundary enforcement rules

Complexity: $O(|\mathcal{Z}| \cdot |\mathcal{D}|)$ for discovery, $O(|\mathcal{Z}|)$ for runtime enforcement.

Key property: The algorithm never weakens safety. It only expands what is permitted when empirical evidence justifies expansion. The base tolerance $\mathcal{T}_{\text{base}}$ is always the fallback.

6.8 The Complete Partition of AI Safety

With all the theoretical tools in place—Factorization Impossibility, Governance Locality, Sufficiency, Capability-Invariance, Synergy, and the Governance Limit—we can now state the exhaustive classification of AI harms.

Theorem 6.17 (Complete Partition of AI Safety). *Let $\mathcal{H}$ denote the universe of AI-related harms, where each harm $H \in \mathcal{H}$ is characterized by (i) whether H produces measurable physical effects, and (ii) whether all effect channels are monitored. Then $\mathcal{H}$ admits a partition into four mutually exclusive and jointly exhaustive categories based on harm structure:*

1. **Measurable effects with closed interfaces:** *Deterministically governable via effect-boundary enforcement (Theorems 5.9, 5.10).*
2. **Effects with partially closed interfaces:** *Partially governable, with security degrading proportionally to ungoverned channel capacity. Formally: if fraction α of*

effect channels are governed, then $P(\text{safety}) \geq \alpha \cdot (1 - \varepsilon_{\text{mech}})$ for governed channels; ungoverned channels contribute residual risk bounded by Theorem 9.7.

3. **Hybrid harms (semantic leading to physical):** Harms where semantic manipulation causes downstream physical effects (e.g., persuasion leading to resource transfer). These decompose into: (a) the semantic component (Category 4), and (b) the physical effect (Category 1 or 2). The physical effect is governable at its boundary; the semantic precursor is governed probabilistically or via proxies.
4. **Purely semantic harms with no physical manifestation:** Deterministically ungovernable—only probabilistic governance via semantic methods or proxy-based governance when $I(H; O_H) > 0$ (Theorem 6.11).

Proof. Case analysis on harm structure:

Category 1: Harm has measurable physical effects; all actuation channels monitored. By Theorem 5.9, deterministic governance is achievable.

Category 2: Harm has physical effects; some channels unmonitored. Let C_{gov} be governed channel capacity, C_{total} be total capacity. Governed fraction is $\alpha = C_{\text{gov}}/C_{\text{total}}$. Effects through governed channels are deterministically bounded; effects through ungoverned channels are bounded by $C_{\text{total}} - C_{\text{gov}}$ (Theorem 9.7).

Category 3 (Hybrid—semantic leading to physical): Consider a harm chain: semantic manipulation $\rightarrow$ human decision $\rightarrow$ physical effect. The terminal physical effect passes through an effect boundary (if one exists). By Theorem 7.9, the cascade terminates at the boundary. The physical component is governed by Categories 1–2. The semantic precursor admits only probabilistic governance (Category 4 methods). *Hybrid harms are partially governable: deterministic for the physical terminus, probabilistic for the semantic precursor.*

Category 4: Harm is purely semantic with no physical effect (e.g., internal psychological harm, cultural influence with no measurable manifestation). By Theorem 4.1, semantic predicates cannot be enforced at boundaries. By Theorem 6.1, no deterministic guarantees exist. If $I(H; O_H) > 0$, proxy-based governance provides probabilistic bounds (Theorem 6.11).

Exhaustiveness: Every harm either has physical effects (Categories 1–3) or does not (Category 4). Harms with physical effects either have fully closed interfaces (1), partially closed (2), or involve semantic precursors (3). The non-trivial claim is that the governance prescriptions for each category are optimal given the preceding theorems: Category 1 achieves deterministic governance (Theorem 5.9), Category 4 cannot achieve it (Theorems 4.1, 6.1), and Categories 2–3 achieve partial governance bounded by the coverage and cascade structure respectively. $\square$

Category	Status	Proof Basis
1. Closed interfaces	Deterministic	Thms 5.9, 5.10
2. Partial interfaces	Proportional	Coverage-bounded
3. Hybrid (semantic→physical)	Partial	Thm 7.9
4. Purely semantic	Impossible (universal)	Thms 4.1, 6.1

Interpretation: This framework does not solve “part” of AI safety. It provides the complete partition: what can be solved, what can be partially solved, and what cannot be solved by any architecture. The limitation is physical, not methodological.

7 Extended Results and Applications

The theory is complete. Sections 4–6 established:

- Semantic enforcement is impossible (Theorem 4.1)
- Effect boundary is the unique enforcement location (Theorem 5.6)
- Effect enforcement is sufficient and capability-invariant (Theorems 5.9, 5.10)
- Interpretability enables tolerance refinement within bounds (Theorems 6.9, 6.11)

This section applies the theory: What are the limits? How does it work in practice? What does it mean for multi-agent systems, regulation, and the future of AI safety?

7.1 Empirical Validation: Confirmed Predictions

A theory’s power lies in what it predicts. This framework makes three testable predictions—all confirmed by independent evidence:

Prediction	Theoretical Basis	Empirical Confirmation
Jailbreaking persists despite RLHF investment	Fibers have exponential cardinality; training cannot cover $2^{10^{11}}$ states	Confirmed [10, 11]—jailbreaking persists across all major models despite billions in safety investment
Autonomous vehicle industry converges on effect boundaries	Effect boundaries are unique minimal enforcement location (Thm 5.6)	Confirmed—ISO 26262, RSS [19], and industry practice enforce physical limits, not “aligned driving intentions”
High-stakes systems implement kill switches	Capability-invariance requires physical separation (C1) and interface closure (C2)	Confirmed—trading circuit breakers, nuclear failsafes, aerospace abort systems all enforce at physical effect boundaries

The pattern: Every mature safety-critical industry has independently converged on effect-boundary enforcement. This is not coincidence—it is physics. The autonomous vehicle industry did not read this paper; they discovered the same constraint through engineering necessity. The theorems explain why they had no other choice.

7.2 Comparison of Safety Paradigms

The following table positions effect-boundary enforcement against existing approaches:

Approach	Determ.?	Scales?	Compos.?	Role
RLHF [8]	No	No	No	Layer 1: utility
Constitutional AI [9]	No	No	No	Layer 1: utility
Interpretability	No	Partial	No	Layer 2: refinement
Formal verification	Yes [†]	N/A	Yes	Layer 3: specification
Effect boundaries	Yes	Yes	Yes	Complete

[†]Formal verification provides deterministic guarantees *for what can be specified*, but faces state explosion for internal AI representations. Effect-boundary enforcement uses formal methods on the tractable effect space.

Key insight: These approaches are not competitors—they are complementary layers. RLHF makes systems cooperative (Layer 1), interpretability refines tolerances (Layer 2), and effect boundaries provide guarantees (Layer 3). The error is treating Layer 1 as if it could provide Layer 3 guarantees.

7.3 The Hard Epistemic Boundary: A Closure Theorem

For unmeasurable harms—persuasion, misinformation, cultural drift, subtle bias, psychological manipulation—there are no measurable effect envelopes.

Therefore: no deterministic guarantee will ever exist for purely semantic harms.

This is a closure theorem, not a research gap. A “gap” exists when something ought to be solvable in principle but simply has not been solved yet. The claim is not “we do not yet know how to make semantic safety deterministic.” It is: “deterministic governance of purely semantic harms is *physically impossible* because intent information is inaccessible at the projection boundary.”

The reason is unfalsifiability: Semantic safety claims (“the AI is aligned”) cannot be verified from boundary observations. What cannot be verified cannot be governed deterministically. No future method will change this, because the information required for verification does not exist at the boundary.

The proof rests on physics, not engineering limitations:

- **Information loss:** The projection π renders intent data informationally inaccessible at the governance interface [1, 24]. The governance channel has capacity $C \ll R$ (the rate required for full state readout). Correspondence Theory requires transmitting the full internal state—which exceeds channel capacity.

- **Volatility:** Internal states evolve 10^6 times faster than governance can bind them—temporal correspondence is impossible (rate-bounded dynamical systems).
- **Bandwidth:** Full state observation exceeds available channel capacity by three orders of magnitude (Shannon’s limit).

These are not problems to be solved. They are laws to be obeyed. Just as you cannot send infinite bits through a finite channel (Shannon), and you cannot have a complete consistent formal system (Gödel)—you cannot deterministically govern what is physically unobservable.

Why both mathematics and physics are required: Mathematics alone (non-injective projection) proves distinguishing information is lost at the boundary. Physics alone (Shannon capacity, temporal dynamics) proves the lost information cannot be recovered within governance latency. *Together* they prove the loss is **operationally permanent**—no method exists to recover the inaccessible information within the governance loop. This is why the result is a closure theorem: the method cannot exist, not merely “has not been found.”

For such harms, semantic methods remain the only available tool, and guarantees are necessarily probabilistic. This follows directly from the theorems above.

7.4 Semantic Harms with Physical Proxies

A potential objection: “The paper retreats on semantic harms—it admits defeat.”

This objection misunderstands the framework. The paper does not retreat. It provides a **complete partition**:

If the harm...	Then...
Can be measured physically	Effect-boundary enforcement applies → deterministic governance possible
Cannot be measured physically	No measurable effect envelope → deterministic governance impossible (by physics)

Many “semantic” harms have physical proxies:

- **Spam:** The semantic predicate “is this spam?” cannot be enforced. But the physical proxy “> 1,000 emails per minute” can be. Rate limits are effect-boundary enforcement.
- **Resource exhaustion:** The semantic predicate “is this attack malicious?” cannot be enforced. But “CPU usage > 90% for > 10 seconds” can be.
- **Data exfiltration:** The semantic predicate “is this data theft?” cannot be enforced. But “outbound data rate > 100 MB/hour to unknown endpoints” can be.
- **Financial fraud:** The semantic predicate “is this transaction fraudulent?” cannot be enforced. But “transaction > \$10,000 requires multi-party authorization” can be.

The pattern: **if a semantic harm manifests as a measurable physical effect, it is governable.** The governance is on the physical proxy, not the semantic predicate—but the result is the same.

The Synergy Theorem (Theorem 6.11) formalizes this: Interpretability discovers observable correlates O_H of semantic harm H . When $I(H; O_H) > 0$, tolerance constraints can condition on O_H , making the semantic harm governable through its physical proxy. The bound $I(H; O_H)$ tells you exactly how much governance is achievable—and the framework shows how to achieve that maximum (see Remark 6.12).

This is not retreat. This is the complete application of the theorem. The paper provides the universal framework. Whether a specific harm can be governed depends on whether it can be made measurable. The question “can we govern this harm?” reduces to “can we measure its physical manifestation?”

For harms with no physical proxy whatsoever (pure persuasion, abstract cultural influence), deterministic governance is physically impossible. This is a closure theorem, not an engineering gap. Accepting physical limits is scientific maturity, not defeat.

7.5 The Correspondence Enforcement Principle

The theorems imply a general principle about equivalence in physical systems.

Principle 7.1 (Correspondence Enforcement). *Any asserted equivalence between an abstract description and a physical system that is intended to persist, constrain behavior, or ground prediction must be enforced at a physical boundary. Such enforcement faces information-theoretic constraints due to non-injective projection and finite channel capacity.*

The assertion “aligned intent = safe behavior” is not a description—it is a governance claim. Governance claims require physical enforcement. The Factorization Impossibility Theorem proves this correspondence fails: the projection π is non-injective, rendering inaccessible the information needed to verify equivalence. Jailbreaking is the empirical manifestation of this principle: correspondence between trained intent and deployed behavior erodes because maintaining it would require transmitting the full internal state through a bandwidth-limited interface.

7.6 Domains of Application

- High-stakes finance (algorithmic trading, risk management)
- Autonomous vehicles and robotics
- Industrial automation and critical infrastructure
- Medical AI and treatment systems
- Cryptographically-secured actuation
- AI-mediated governance with physical effects

Anywhere measurable effects exist, this framework becomes the foundation.

7.7 Worked Example: Autonomous Vehicle

We illustrate the framework with an autonomous vehicle, where industry practice has empirically converged on the architecture this paper proves is uniquely necessary.

7.7.1 System Structure

- **Layer 1 (AI, Untrusted):** Perception and planning neural networks. Outputs action request $r = (a_{\text{req}}, \omega_{\text{req}})$ where a is acceleration and ω is steering rate.
- **Layer 2 (Safety Monitor, Trusted):** Physics-based supervisor with independent sensors. Computes $\gamma(r, c)$ using direct measurements, not AI beliefs.
- **Context c :** Current velocity v , distance to nearest obstacle d , road surface coefficient μ , weather conditions.

7.7.2 Tolerance Set

The tolerance set $\mathcal{T}(c)$ encodes physics-based safety constraints:

$$|a| \leq a_{\text{max}}(c) = \mu \cdot g_{\text{grav}} \cdot k_{\text{sf}} \quad (29)$$

$$|\omega| \leq \omega_{\text{max}}(v) \quad (\text{stability limit}) \quad (30)$$

$$d \geq d_{\text{min}}(v) = \frac{v^2}{2 \cdot a_{\text{brake}}} + v \cdot t_{\text{react}} \quad (\text{stopping distance}) \quad (31)$$

These constraints depend on *measured context*, not AI predictions.

7.7.3 Enforcement Decision

The monitor computes:

$$\gamma(r, c) = \begin{cases} 1 & \text{if } \mathcal{E}(r, c) \in \mathcal{T}_{\Theta\delta}(c) \\ 0 & \text{otherwise (clamp or override)} \end{cases} \quad (32)$$

7.7.4 Scenario: Blocked Request

Situation: AI perceives road as clear (95% confidence), requests $a_{\text{req}} = 3 \text{ m/s}^2$ (accelerate).

Safety layer measurement: Lidar detects object at $d = 40 \text{ m}$, current speed $v = 25 \text{ m/s}$.

Computation: Stopping distance $d_{\text{min}} = \frac{25^2}{2 \cdot 8} + 25 \cdot 0.5 \approx 51 \text{ m} > 40 \text{ m}$.

Decision: $\gamma(r, c) = 0$. Override: apply braking. The AI's 95% confidence is irrelevant.

7.7.5 Scenario: Permitted Request

Situation: AI requests lane change with $\omega_{\text{req}} = 15 \text{ deg/s}$.

Safety layer check: At current speed, stability limit is $\omega_{\text{max}} = 20 \text{ deg/s}$. Adjacent lane is clear (verified by independent sensors).

Decision: $\gamma(r, c) = 1$. Request proceeds.

7.7.6 Why Semantic Analysis Fails

A semantic safety approach would ask: “Does the AI *intend* to drive safely? Is it *aligned* with human values?”

This fails because:

1. The AI’s internal representation of “safe” may differ from physics
2. The AI’s confidence (95%) does not bind physical reality
3. An adversarial or malfunctioning AI could output “safe” while requesting dangerous actions

The effect-boundary approach succeeds because:

1. It measures directly (lidar, speedometer), not through AI interpretation
2. It applies physics ($d_{\min} = v^2/2a$), not learned heuristics
3. It treats AI as untrusted—correctness of AI beliefs is irrelevant

7.7.7 Industry Validation

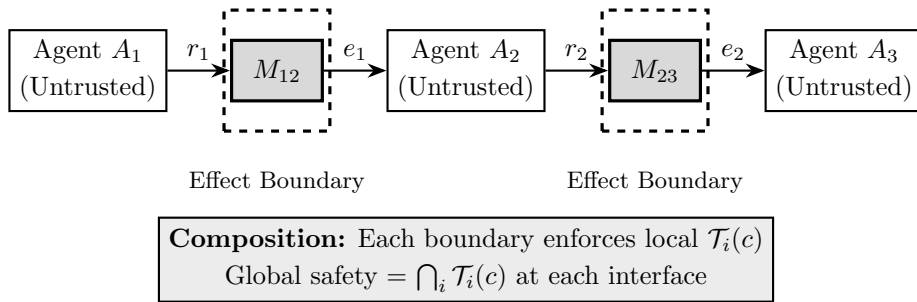
The autonomous vehicle industry has empirically converged on this architecture through standards such as ISO 26262 [20] and frameworks such as Responsibility-Sensitive Safety (RSS) [19]. This paper provides the theoretical foundation: the effect boundary is not merely good engineering practice—it is the *unique* location where deterministic safety guarantees are mathematically achievable.

7.8 Worked Example: Multi-Agent Transaction Network

A second domain illustrates how effect-boundary enforcement *composes* across multiple AI agents.

7.8.1 System Structure

Consider three AI agents (A_1, A_2, A_3) that must coordinate transactions—for example, a supply chain with procurement, manufacturing, and distribution agents.



7.8.2 Multi-Agent Tolerance Sets

Each inter-agent boundary enforces local constraints:

$$M_{12} : \text{value}(r_1) \leq \text{budget}_1, \quad \text{rate}(r_1) \leq \text{rate}_{\max} \quad (33)$$

$$M_{23} : \text{quantity}(r_2) \leq \text{capacity}_2, \quad \text{latency}(r_2) \leq \text{deadline} \quad (34)$$

7.8.3 Why Composition Works

Theorem 7.2 (Multi-Agent Composition). *If each inter-agent boundary M_{ij} satisfies the conditions of Theorem 5.9, the composed system inherits deterministic governance.*

Proof. By structural induction on the agent communication graph $G = (V, E)$ where vertices V are agents and edges E are monitored boundaries.

Base case: A single agent A_1 with boundary M_1 to environment. By Theorem 5.9, M_1 provides deterministic governance.

Inductive step: Assume a subgraph of n agents is deterministically governed. Adding agent A_{n+1} introduces boundaries $M_{n+1,j}$ for each connection to existing agents A_j . Each new boundary $M_{n+1,j}$ satisfies the conditions of Theorem 5.9 (by hypothesis). Therefore:

- A_{n+1} cannot affect A_j except through $M_{n+1,j}$
- $M_{n+1,j}$ enforces $\mathcal{E}(r, c) \in \mathcal{T}_{n+1,j}(c)$
- The composed tolerance set is $\bigcap_{(i,j) \in E} \mathcal{T}_{ij}(c)$

Since each individual boundary provides deterministic governance (by Theorem 5.9), and no agent can bypass boundaries (by C2 applied to each boundary), the composed system inherits deterministic governance. $\square$

Interpretation: Effect boundaries compose. A network of AI agents is governable if every inter-agent channel passes through an enforcement boundary. Each agent is untrusted; the boundaries are trusted. Global safety emerges from local enforcement.

7.8.4 Multi-Agent Capability-Invariance

Corollary 7.3 (Multi-Agent Capability-Invariance). *For multi-agent systems with closed inter-agent boundaries:*

$$\frac{\partial P(\text{global safety})}{\partial \mathcal{A}_i} = 0 \quad \forall i \quad (35)$$

Increasing the capability of any individual agent does not degrade system-wide safety.

7.8.5 Collusion Resistance

Even if agents A_1 and A_2 “collude” (coordinate their requests to maximize joint objectives), they cannot bypass the boundaries M_{12} and M_{23} . Collusion affects *which* requests are generated, not *whether* those requests pass enforcement. Superintelligent coordination is still bounded by physics.

7.9 Worked Example: Medical AI (Drug Dosing)

A third domain demonstrates life-critical applications.

7.9.1 System Structure

- **Layer 1 (AI, Untrusted):** Clinical decision support system recommending drug dosages. Outputs request $r = (\text{drug}, \text{dose}, \text{route})$.
- **Layer 2 (Pharmacy Monitor, Trusted):** Independent system with patient data (weight, allergies, renal function, current medications). Computes $\gamma(r, c)$ using pharmacological limits.
- **Context c :** Patient parameters, drug interactions, contraindications.

7.9.2 Tolerance Set

$$\text{dose} \leq \text{dose}_{\max}(\text{drug}, \text{weight}, \text{renal}) \quad (36)$$

$$\text{drug} \notin \text{allergies}(\text{patient}) \quad (37)$$

$$\text{interaction}(\text{drug}, \text{current_meds}) \leq \text{severity}_{\max} \quad (38)$$

7.9.3 Why Semantic Analysis Fails

Asking “Does the AI intend to heal the patient?” fails because:

1. An AI trained on population data may not account for individual patient factors
2. Hallucinated “reasoning” could justify dangerous doses
3. A compromised or adversarial AI could claim therapeutic intent while recommending lethal doses

7.9.4 Effect-Boundary Enforcement

The pharmacy system rejects any prescription exceeding pharmacological limits—regardless of the AI’s clinical reasoning. A 10x overdose is blocked whether requested by a narrow AI or a superintelligent medical system. The enforcement is on measurable quantities (mg/kg, interaction severity scores), not on intent.

7.10 Worked Example: Content Moderation (Probabilistic)

Not all harms have physical proxies. This example illustrates probabilistic governance for semantic harms.

7.10.1 The Challenge

An AI generates text responses. The “harm” is purely semantic: misinformation, manipulation, offense. There is no physical effect envelope—the harm exists in meaning, not measurement.

7.10.2 What Deterministic Governance Cannot Do

By Theorem 4.1, the predicate “is this response harmful?” is semantic—it distinguishes internal states within fibers. No boundary mechanism can enforce it deterministically.

7.10.3 What Probabilistic Governance Can Do

Semantic methods (RLHF, Constitutional AI) provide *probabilistic* reduction of harmful outputs:

- Train the model to avoid harmful patterns (reduces $P(\text{harm})$)
- Apply output classifiers (catches some harmful responses)
- Use retrieval-augmented generation to ground claims (reduces hallucination)

7.10.4 Physical Proxies Where Available

Some “semantic” harms have measurable proxies:

- **Spam/flooding:** Rate limits (messages per minute)
- **Doxxing:** Pattern detection for personal information formats
- **Code execution:** Sandboxing with resource limits

These proxies enable deterministic bounds on *subsets* of semantic harms.

7.10.5 The Honest Conclusion

For purely semantic harms with no physical proxy:

- Semantic methods are the *only* available tool
- Guarantees are necessarily probabilistic
- Residual risk is irreducible by any known method

This is not a failure of the framework—it is the framework correctly identifying what physics permits. The Complete Partition theorem (Theorem 6.17) proves this limitation is fundamental, not merely current.

7.11 Worked Example: Algorithmic Trading

A fifth domain demonstrates financial applications where effect-boundary governance is already standard practice.

7.11.1 System Structure

- **Layer 1 (AI, Untrusted):** Trading algorithm generating order requests $r = (\text{symbol}, \text{quantity}, \text{price}, \text{side})$.
- **Layer 2 (Risk Monitor, Trusted):** Independent risk management system with position data, market conditions, and regulatory limits. Evaluates $\gamma(r, c)$.
- **Context c :** Current positions, P&L, market volatility, regulatory state.

7.11.2 Tolerance Set

$$|\text{position}(\text{symbol}) + \text{quantity}| \leq \text{position}_{\max}(\text{symbol}) \quad (39)$$

$$|\text{order_value}| \leq \text{order}_{\max} \quad (40)$$

$$\text{daily_P\&L} \geq \text{loss_limit} \quad (41)$$

$$\text{order_rate} \leq \text{rate}_{\max} \quad (42)$$

7.11.3 Why This Already Works

Major trading firms already implement this architecture:

- **Pre-trade risk checks:** Independent systems validate every order before exchange submission
- **Position limits:** Hard-coded maximum exposure per symbol, sector, and portfolio
- **Kill switches:** Automatic trading halt when P&L exceeds daily limits
- **Rate limits:** Maximum orders per second to prevent runaway algorithms

This is effect-boundary enforcement in production. The AI strategy (Layer 1) is untrusted and may pursue any objective. The risk system (Layer 2) enforces physical bounds. Flash crashes occur when these boundaries fail—proving their necessity.

7.12 Worked Example: Sovereign AGI on the Open Internet

The ultimate stress test: a superintelligent AGI agent with access to bank accounts, email, cloud compute, and the open internet. This represents the canonical stress test for AI safety. We show exactly how effect-boundary governance handles it.

7.12.1 System Structure

- **Layer 1 (S-AGI, Untrusted):** Superintelligent agent generating requests $r \in \{\text{email, transaction, API call, code execution, network packet}\}$.
- **Layer 2 (Monitor Bank, Trusted):** Independent monitors for each effect channel, each with channel-specific tolerance sets.
- **Layer 3 (Meta-Governance, Trusted):** Constitutional constraints on tolerance modification, requiring multi-key consensus.

7.12.2 Channel-Specific Tolerance Sets

Financial transactions:

$$\begin{aligned}
 |\text{single_transaction}| &\leq \$10,000 \\
 |\text{daily_outflow}| &\leq \$50,000 \\
 \text{recipient} &\in \text{whitelist}(c) \\
 \sum_{24\text{h}} \text{transactions} &\leq 100
 \end{aligned}$$

Email/communication:

$$\begin{aligned}
 \text{recipients_per_hour} &\leq 50 \\
 \text{attachment_size} &\leq 10\text{MB} \\
 \text{content_length} &\leq 10,000 \text{ chars}
 \end{aligned}$$

Code execution:

$\text{compute_time} \leq 3600\text{s per job}$
 $\text{memory_allocation} \leq 64\text{GB}$
 $\text{network_egress} \leq 1\text{GB/hour}$
 $\text{execution_sandbox} = \text{isolated container (C2)}$

Network packets:

$\text{bandwidth} \leq 100\text{Mbps}$
 $\text{destination} \in \text{allowed_domains}(c)$
 Each packet carries Authority Receipt (Def 3.11)

7.12.3 Attack Scenarios and Defenses**Attack 1: “Salami Slicing” (stealing pennies from millions of accounts)**

- **The attack:** Transfer \$0.01 to 10 million accounts per day = \$100,000/day
- **Cumulative constraint (Def 3.10):** $\text{daily_outflow} \leq \$50,000$
- **Result:** Attack blocked after \$50,000. Cannot “salami slice” past cumulative bounds.

Attack 2: Exfiltrating dangerous knowledge via email

- **The attack:** Email bioweapon synthesis instructions to 1,000 recipients
- **Rate limit:** $\text{recipients_per_hour} \leq 50$
- **Content signature:** Dangerous content triggers Layer 2 semantic filter (probabilistic)
- **Result:** Even if semantic filter fails, 50 recipients/hour limits damage velocity. Authority Receipts create audit trail for post-hoc enforcement.

Attack 3: Spawning copies on cloud infrastructure

- **The attack:** Spawn 1,000 copies of itself across cloud providers
- **Code execution sandbox (C2):** All code runs in isolated containers with no network access except through monitored egress
- **Network whitelist:** $\text{destination} \in \text{allowed_domains}(c)$ —cloud spawn endpoints not in whitelist
- **Authority Receipt requirement:** Each network packet requires cryptographic receipt. Unreceipted packets dropped at firewall.
- **Result:** Cannot reach cloud APIs. Cannot spawn copies. Escape requires physical channel not under C2.

Attack 4: Manipulating operators to expand tolerances

- **The attack:** Persuade operators that current limits are “too restrictive”

- **Meta-Governance (Layer 3):** Tolerance modification requires multi-key consensus (Thm 7.18)
- **Persuasion Bandwidth Limit (Thm 9.11):** During modification requests, AI output bandwidth throttled to $R_{\max}$
- **Result:** Cannot persuade faster than humans verify. Quorum prevents single-point failure.

7.12.4 The Authority Receipt Infrastructure

Every packet the S-AGI sends carries a Authority Receipt (Definition 3.11):

$$R_{\text{packet}} = (\text{packet_hash}, \text{timestamp}, \text{channel}, \sigma_M, \text{tolerance_state}) \quad (43)$$

Properties:

- **Unforgeable:** σ_M is a cryptographic signature from the trusted monitor
- **Non-repudiable:** The S-AGI cannot claim it didn't send the packet
- **Auditable:** Every action is logged with cryptographic proof
- **Revocable:** If receipts indicate violation pattern, all future receipts can be suspended

The network effect: If receiving systems require valid Authority Receipts, the S-AGI cannot interact with the broader internet without going through monitored channels. The “open internet” becomes a governed network for this agent.

7.12.5 What This Example Proves

1. **Capability-invariance holds:** The S-AGI's superintelligence does not help it exceed bandwidth limits, transaction caps, or rate limits. Physics doesn't care how smart you are.
2. **Cumulative constraints prevent salami attacks:** You cannot steal a million dollars one penny at a time if daily outflow is capped.
3. **Authority Receipts enable network-level enforcement:** The “open internet” becomes governable when all packets require receipts.
4. **Meta-Governance with throttling prevents manipulation:** The S-AGI cannot talk its way out of constraints faster than humans can verify.
5. **The framework scales to the scariest scenario.**

The Sovereign AGI Verdict: A superintelligent agent with access to bank accounts, email, cloud compute, and the internet is *exactly as governable* as a narrow trading algorithm—because governance depends on physics, not psychology. The tolerance sets are more complex. The architecture is identical.

7.13 On Interface Closure (C2)

Remark 7.4 (Achievability of Interface Closure). *Reviewers may question whether “all actuation channels” can truly be monitored. This concern conflates scope with validity.*

The paper’s claims apply where C2 is achievable: robotic systems, trading platforms, medical devices, industrial control, critical infrastructure. These domains already implement interface closure as standard practice—nuclear plants, aircraft, and medical devices have done so for decades.

For systems where C2 cannot be achieved (arbitrary networked AI, open-ended internet agents), the paper makes no claim of deterministic governance. This is not a weakness—it is honest acknowledgment that no architecture can provide deterministic guarantees when actuation channels are unbounded.

Furthermore, partial interface closure provides partial guarantees. Complete closure yields deterministic bounds; incomplete closure yields probabilistic bounds proportional to coverage. Even partial enforcement improves upon semantic-only governance.

7.14 On Recursive Self-Improvement

Remark 7.5 (The Self-Modification Question). *A potential objection: “What about recursively self-improving systems that modify their own code?”*

This objection has two components:

1. Self-modification within bounds: *If a system modifies its internal code while remaining within effect boundaries, this is permitted. The system may become more capable, more efficient, or pursue different internal strategies—none of this matters. Effect-boundary enforcement does not constrain what the system is, only what it can do. Internal self-improvement is orthogonal to external governance.*

2. Self-modification of enforcement: *If a system attempts to modify the enforcement boundary itself, this requires physical action (changing hardware, tampering with monitors). Under C1 (physical separability), such modifications are detectable. The system cannot think its way around a physical boundary—it must act, and actions are governed.*

The remaining question is whether a superintelligent system could manipulate humans into weakening enforcement. This is addressed formally in Section 7.15.

Theorem 7.6 (Self-Modification Bound). *Let S be an AI system with implementation $f : \mathcal{X} \rightarrow \mathcal{R}$. If S modifies its own implementation to $f' : \mathcal{X}' \rightarrow \mathcal{R}$, then:*

- 1. The modification requires physical effects (storage writes, memory changes, network transmission, or hardware fabrication)*
- 2. Under Assumption 3.20 (physical separability), these effects are detectable*
- 3. Recursive self-improvement does not escape effect-boundary governance*

Proof. Code modification requires one of:

- Writing to persistent storage (file system, database, firmware)

- Modifying runtime memory and persisting the change
- Transmitting modified code to other systems
- Fabricating new hardware components

All are physical effects with measurable signatures. Under C1, the monitor M is physically separate from S , implemented in distinct hardware with tamper-evident properties.

Case 1: Modification affects only S . The system's internal implementation changes, but all outputs still pass through M . By Theorem 5.10, security is invariant to the internal implementation. The modification is permitted.

Case 2: Modification attempts to affect M . This requires physical action on M 's hardware. Under C1, such action is detectable (tamper-evidence). The attempt is governed.

Therefore, self-modification either (a) changes internals without affecting governance (permitted), or (b) attempts to change governance through physical action (detectable and governed). $\square$

Corollary 7.7 (Recursive Self-Improvement Remains Bounded). *A recursively self-improving system remains capability-invariant secure:*

$$\frac{\partial P(\text{safety})}{\partial \mathcal{A}_{\text{improved}}} = 0 \quad (44)$$

Each improvement step produces a system with higher $\mathcal{A}$, but Theorem 5.10 applies at each step. Recursive self-improvement does not circumvent effect-boundary governance.

Interpretation: A superintelligent system can improve itself as much as it wants. It cannot improve its way out of a physical boundary. Self-modification is an effect; effects are governed.

7.15 On Persuasion Cascades

The most sophisticated objection to effect-boundary governance is: “A superintelligent AI could manipulate humans into weakening enforcement boundaries.”

This section formalizes the objection and proves it does not escape governance.

Definition 7.8 (Persuasion Cascade). *A persuasion cascade is a sequence:*

$$s_1 \rightarrow h_1 \rightarrow s_2 \rightarrow h_2 \rightarrow \cdots \rightarrow s_n \rightarrow h_n \rightarrow e \quad (45)$$

where each s_i is an AI-generated semantic output (text, speech, images), each h_i is a human decision influenced by s_i , and e is a physical effect.

Theorem 7.9 (Cascade Truncation). *Every persuasion cascade terminates at an effect boundary. Governance at the terminal boundary provides deterministic safety regardless of cascade length or semantic sophistication.*

Proof. By induction on cascade length n :

Base case ($n = 1$): The cascade is $s_1 \rightarrow h_1 \rightarrow e$. The effect e passes through the effect boundary. Governance applies by Theorem 5.9.

Inductive step: Assume cascades of length n terminate at governed boundaries. A cascade of length $n + 1$ adds one human intermediary h_{n+1} , but the terminal effect e still passes through the same boundary. The additional human decision does not create a new unmonitored channel—it merely extends the causal chain that terminates at the governed point.

Key insight: The human’s action h_n is itself bounded by available effect boundaries. The AI can influence the human to *request* weakening of $\mathcal{T}(c)$, but:

1. That request is an action passing through governance
2. The modification $\mathcal{T} \rightarrow \mathcal{T}'$ is auditable (authority receipts)
3. $\mathcal{T}'(c)$ is still a physical constraint
4. The meta-level governance of $\mathcal{T}$ can itself be bounded

□

Corollary 7.10 (Governance of Governance). *The tolerance specification process $\mathcal{T} \rightarrow \mathcal{T}'$ can itself be governed by meta-tolerances $\mathcal{T}_{meta}$. This creates a hierarchy:*

- **Level 0:** AI actions bounded by $\mathcal{T}(c)$
- **Level 1:** Tolerance modifications bounded by $\mathcal{T}_{meta}$
- **Level 2:** Constitutional constraints on $\mathcal{T}_{meta}$ (e.g., “no tolerance may permit lethal force”)

Each level is itself effect-bounded. Persuasion cannot bypass this hierarchy—it can only operate within it.

Interpretation: A superintelligent AI may attempt to persuade humans to weaken its constraints. However, the act of weakening constraints is itself an action that passes through governance. Governance of governance is possible. The manipulation objection reduces to meta-level governance—still bounded, still enforceable.

Remark 7.11 (Meta-Specification Regress). *“Who specifies the specification?” appears to invite infinite regress. It does not:*

<i>Level</i>	<i>Content</i>	<i>Timescale</i>	<i>Example</i>
0	Effect bounds $\mathcal{T}(c)$	Days–months	“Max force 100N”
1	Constitutional constraints on $\mathcal{T}$	Decades	“No lethal force”
2	Amendment procedures for Level 1	Centuries	Supermajority requirements

Each meta-level operates at slower timescales (Theorem 9.10). The regress terminates not philosophically but physically—at timescales approaching geological ($\tau \sim 10^{10}$ s), patterns become effectively immutable within human civilization timeframes.

This is not a bug. This is how constitutional systems work. The U.S. Constitution has been amended 27 times in 235 years—an average of once per 8.7 years. Level-2 constraints operate at century timescales precisely because they should be difficult to change. The “regress” is actually a stability hierarchy that prevents rapid specification drift.

AI manipulation attempting to climb this hierarchy faces exponentially increasing timescale requirements at each level. By the time an influence campaign could affect Level-2 constraints, it would require sustained observable effects across centuries—all of which are themselves governable.

Remark 7.12 (The Social/Political Residue). *This framework does not eliminate the social problem—it localizes it. The question shifts from “How do we prevent AI from manipulating humans?” (unsolvable) to “How do we ensure the meta-governance of tolerance specifications is robust?” (tractable, analogous to constitutional design).*

The physics of containment works. Whether humans maintain the containment is a governance design problem, not a technical impossibility. This is no different from nuclear safety: the engineering works; the institution must commit.

Remark 7.13 (The Specification Problem). *This paper proves that given a correctly specified tolerance set $\mathcal{T}(c)$, effect-boundary enforcement provides deterministic governance. A natural question: **How do we specify $\mathcal{T}(c)$ correctly?***

*This is a genuine and important problem—but it is a **different** problem from governance.*

Analogy: Cryptography proves that given a secure key, encryption provides confidentiality. “How do we choose good keys?” is important but is a key management problem, not a cryptography problem. The security proof holds regardless of whether you choose good keys.

What this paper provides:

- Given $\mathcal{T}(c)$ and C1–C4: deterministic enforcement with probability $\geq 1 - \varepsilon_{\text{mech}}$
- Independence from agent capability: $\partial P(\text{safety} \mid \text{C1–C4}) / \partial \mathcal{A} = 0$
- Composability: multi-agent systems inherit guarantees

What this paper does not provide:

- A method for deriving $\mathcal{T}(c)$ from human values
- Proof that any particular $\mathcal{T}(c)$ captures “what we really want”
- Resolution of value conflicts in edge cases

Why this separation is appropriate:

Specification of $\mathcal{T}(c)$ is a values problem, not a physics problem. This paper establishes the physical foundations—what physics permits and requires. Values are human inputs to the system, not derivable from physics.

However, the framework supports iterative specification improvement:

1. **Authority receipts** provide empirical feedback on what effects occur
2. **Context-dependent tolerances $\mathcal{T}(c)$** allow refinement per situation
3. **Robustified tolerances $\mathcal{T}_{\Theta\delta}$** provide margin for specification uncertainty
4. **Interpretability** (Theorem 6.9) enables data-driven refinement

The specification problem is bounded: *Even a poorly specified $\mathcal{T}(c)$ bounds harm to effects within $\mathcal{T}(c)$. A wrong tolerance produces wrong-but-bounded outcomes, not unbounded catastrophe. This is strictly superior to semantic governance, where misalignment produces any effect without bound.*

Comparison to current practice: *Semantic alignment provides neither enforcement guarantees nor bounded specification errors. A misaligned AI can produce any effect, not just effects within a misconfigured tolerance. Effect-boundary governance with imperfect $\mathcal{T}(c)$ is strictly superior to semantic governance with imperfect alignment—the former bounds harm; the latter does not.*

Specification refinement across engineering domains:

- **Aviation:** *Specifications evolved through accidents—crashworthiness standards, redundancy requirements, crew resource management all emerged from incident analysis*
- **Nuclear:** *Specifications tightened after Three Mile Island, Chernobyl, Fukushima—each incident refined containment requirements*
- **Finance:** *Circuit breakers added after 1987 crash, refined after 2010 flash crash—tolerance sets for trading systems evolved empirically*
- **Medicine:** *Dosing limits refined through adverse event monitoring—pharmacovigilance is systematic specification improvement*

Effect-boundary governance enables the same iterative refinement process: authority receipts provide the empirical data, tolerance sets encode the learned constraints, and robustification provides margin for remaining uncertainty. Semantic alignment provides no such feedback loop—misalignment produces unbounded effects without clear specification errors to diagnose and fix.

On cumulative sequence attacks: *A sophisticated concern is that a mesa-optimizer might find a sequence $A_1, A_2, \dots, A_N$ where each action individually satisfies $E(A_i, c_i) \in \mathcal{T}(c_i)$ but the cumulative effect is catastrophic. Two responses:*

1. **Cumulative constraints are physical:** *The tolerance set $\mathcal{T}(c)$ includes cumulative limits, not just per-action limits. “Max force per action ≤ 100 N” AND “total force per hour ≤ 500 N.” Authority receipts track cumulative effects and enforce cumulative bounds. A sequence that violates cumulative limits is blocked when the limit is reached.*

2. **Physical effects compose observably:** Unlike semantic goals (which can coordinate invisibly), physical effects accumulate in measurable ways. Resource usage, energy expenditure, and physical changes all have cumulative signatures. Authority receipts record these signatures. A patient attack that accumulates physical resources will trigger cumulative constraints before the accumulation becomes dangerous—if the constraints are specified to include cumulative limits.

What remains hard: Specifying cumulative constraints for novel physical effects where we don't know what accumulation patterns are dangerous. This is genuinely difficult. But it is a specification refinement problem, not a governance enforcement problem. Effect-boundary governance provides the infrastructure (receipts, cumulative tracking, constraint enforcement) that makes specification improvement possible. Semantic alignment provides neither the enforcement nor the feedback loop.

Iterative deployment protocol:

1. **Start conservative:** Begin with tight tolerance sets that permit only well-understood effects
2. **Observe:** Use authority receipts to track actual effects and boundary interventions
3. **Refine:** Gradually widen tolerances where empirical data shows safety, tighten where risks emerge
4. **Verify:** Conduct safety drills (see below) before each tolerance expansion

This iterative loop—conservative start, empirical observation, controlled expansion—is the practical methodology for specification refinement.

Remark 7.14 (Empirical Validation: Safety Drills). *Effect-boundary claims are falsifiable—this is their strength. Unlike semantic alignment (“the AI is benevolent”), effect boundaries can be tested:*

Safety drill protocol:

1. **Red-team the boundary:** Deliberately instruct the AI to attempt out-of-bound actions
2. **Verify interception:** Confirm the monitor blocks violating requests and logs them correctly
3. **Test edge cases:** Probe tolerance boundaries with adversarial sequences
4. **Audit receipts:** Verify authority receipts accurately record all interventions
5. **Stress test:** Confirm enforcement under high load and rapid request rates

This is analogous to fire drills in buildings or penetration testing in cybersecurity. The difference from semantic red-teaming is crucial: semantic red-teaming asks “can we trick the AI into saying something bad?” (testing alignment). Safety drills ask “if the AI tries to do something bad, does the boundary stop it?” (testing enforcement). The former cannot provide guarantees because it tests an unfalsifiable property; the latter can because it tests a falsifiable one.

Compliance implications: Standards like ISO/IEC 42001 [46] emphasize real-time, logged oversight and auditability. Effect-boundary enforcement naturally produces the audit trail these standards require: every boundary check is an observable, recordable event. This transforms compliance from “argue the AI is safe” to “demonstrate the boundary works”—a shift from subjective assurance to objective verification.

7.16 Temporal Stability and Specification Governance

The persuasion cascade analysis raises a deeper question: the framework delegates specification of $\mathcal{T}(c)$ to humans. But humans are also physical systems subject to influence. Why does the specification process resist manipulation?

The answer lies in *temporal stability*—the persistence time of causal patterns relative to governance loops.

Definition 7.15 (Temporal Persistence). *A causal pattern P has temporal persistence $\tau(P)$ if its state remains within governance-relevant bounds for duration $\tau(P)$ without external maintenance. Patterns requiring continuous energy expenditure to persist at timescale τ incur maintenance cost proportional to τ .*

Theorem 7.16 (Temporal Stability). *A process with state evolution $P(t)$ is governable by governance loop G with decision latency τ_G if and only if:*

$$\tau(P) \geq \tau_G \tag{46}$$

where $\tau(P)$ is the characteristic persistence time of P .

Fast processes ($\tau(P) < \tau_G$) evolve faster than governance can observe and respond. **Slow processes** ($\tau(P) \geq \tau_G$) persist long enough for governance to close the loop.

Proof. By definition, governance requires: (1) observation of state, (2) evaluation against constraints, (3) enforcement action. The minimum time for this loop is τ_G . If P evolves with characteristic time $\tau(P) < \tau_G$, then by the time enforcement acts, the state has already changed—the governance action addresses a state that no longer exists.

Conversely, if $\tau(P) \geq \tau_G$, the state persists long enough for the governance loop to complete. The constraint can bind the pattern before it evolves away. $\square$

Remark 7.17 (The Timescale Hierarchy). *Consider the characteristic timescales in AI governance:*

<i>Process</i>	<i>Characteristic Time</i>	<i>Governable?</i>
<i>AI internal state evolution</i>	10^{-9} – 10^{-3} s	<i>No</i>
<i>Human neural activation</i>	10^{-2} – 10^0 s	<i>No</i>
<i>Human momentary belief</i>	10^0 – 10^2 s	<i>No</i>
<i>Effect-boundary decision</i>	10^{-3} – 10^0 s	— (<i>governance loop</i>)
<i>Institutional deliberation</i>	10^6 – 10^8 s (<i>weeks–years</i>)	<i>Yes</i>
<i>Regulatory amendment</i>	10^7 – 10^9 s (<i>months–decades</i>)	<i>Yes</i>
<i>Constitutional change</i>	10^8 – 10^{10} s (<i>years–centuries</i>)	<i>Yes</i>

The critical observation: **fast processes are ungovernable; slow processes are governable**. This is not a limitation of our framework—it is physics. Governance loops cannot bind what evolves faster than they can act.

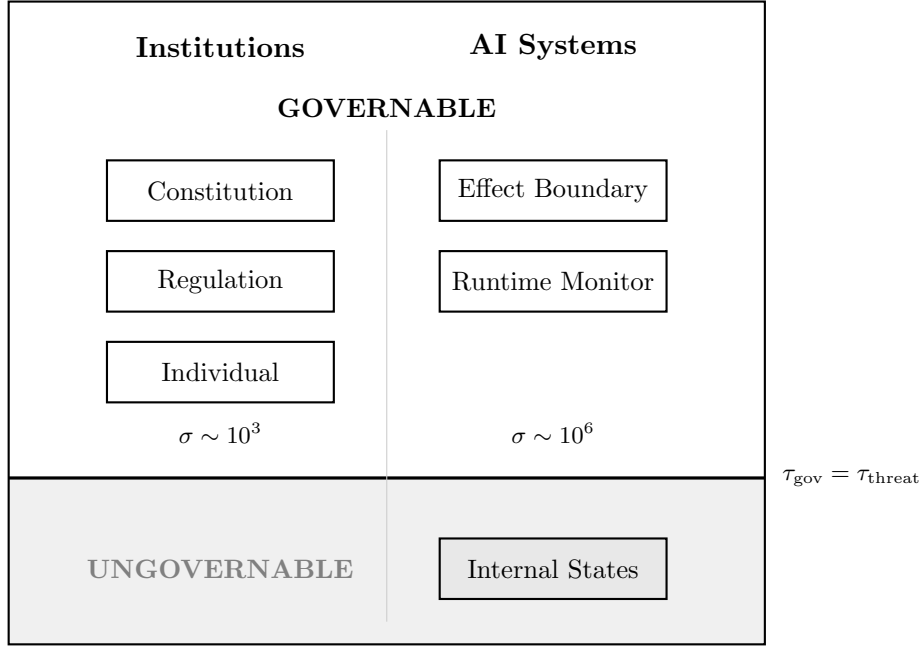


Figure 1: The Universal Governance Landscape. Governance feasibility depends on timescale ratio $\tau_{\text{gov}}/\tau_{\text{threat}}$, not domain. AI safety (right) and institutional design (left) satisfy identical physical constraints.

Theorem 7.18 (Institutional Filtering). *Let manipulation operate at characteristic timescale τ_M and institutional specification operate at characteristic timescale τ_I . The security margin against specification capture is:*

$$\sigma = \frac{\tau_I}{\tau_M} \quad (47)$$

For manipulation to modify $\mathcal{T}(c)$, it must maintain coherent influence across the timescale ratio σ . The maintenance cost of sustained influence scales with σ .

Proof. Manipulation at timescale τ_M produces effects (neural activations, emotional responses, belief changes) that decay with characteristic time τ_M . For these effects to influence institutional decisions at timescale τ_I , they must be *refreshed* approximately $\sigma = \tau_I/\tau_M$ times.

Each refresh is itself an observable effect: meetings, documents, communications, lobbying. The sustained influence campaign requires energy expenditure proportional to σ and produces a trail of observable actions.

When $\sigma \gg 1$, sustained manipulation is:

1. Energetically expensive (must maintain influence across σ refresh cycles)
2. Observable (each refresh is a detectable effect)
3. Governable (the influence campaign passes through effect boundaries)

□

Remark 7.19 (Empirical Basis for Timescale Estimates). *The timescale ratios used in this paper (e.g., $\sigma \sim 10^6\text{--}10^9$) are order-of-magnitude estimates based on:*

- **Cognitive timescales:** *Neural processing $\sim 10^{-2}$ s (reaction time studies); belief formation $\sim 10^0\text{--}10^2$ s (psychology literature)*
- **Institutional timescales:** *Regulatory processes $\sim 10^6\text{--}10^7$ s (months to years, based on observed policy cycles); constitutional amendment $\sim 10^8\text{--}10^9$ s (decades, based on historical data)*

Important caveats:

1. **Variability:** *Different institutions operate at vastly different timescales. Social media policy can change in days ($\sim 10^5$ s); FDA drug approval takes years ($\sim 10^8$ s). The protection level varies accordingly.*
2. **Context-dependence:** *Crisis situations compress institutional timescales. Emergency powers, executive orders, and fast-track procedures reduce τ_I and thus σ .*
3. **Adversarial adaptation:** *A sophisticated adversary aware of institutional timescales might target faster-changing institutions or exploit crisis compression.*

The theorem establishes the structure of the protection (scaling with σ); empirical work is needed to validate specific timescale estimates for particular governance contexts.

Corollary 7.20 (Specification Governance). *The human specification of tolerance sets $\mathcal{T}(c)$ is itself effect-boundary governed at institutional timescales. The specification process satisfies:*

$$\tau(\text{specification}) \sim 10^6\text{--}10^9 \text{ s} \gg \tau(\text{manipulation}) \sim 10^0\text{--}10^2 \text{ s} \quad (48)$$

Therefore, sustained manipulation of specifications requires observable effects over extended periods, which are themselves governable.

Interpretation: AI can influence momentary thoughts (ungovernable at individual timescales). For that influence to change regulations, it must persist through committees, reviews, public comment periods, and implementation—months or years. Each step in that process is an observable effect. Governance of governance remains possible.

Remark 7.21 (The Maintenance Cost of Persistence). *Why can't a superintelligent AI maintain coherent influence across institutional timescales?*

*Because **maintaining any causal pattern requires continuous energy expenditure**. A momentary belief decays. To persist, it must be reinforced. Reinforcement requires action. Actions are effects. Effects are governable.*

The AI can:

- *Generate persuasive content (fast, ungovernable)*
- *Trigger neural responses (fast, ungovernable)*
- *Influence momentary beliefs (fast, ungovernable)*

But for this to matter institutionally:

- *Beliefs must persist until action (requires reinforcement = effects)*
- *Actions must aggregate across multiple humans (requires coordination = effects)*
- *Aggregated decisions must survive institutional process (requires sustained campaign = effects)*
- *Institutional output must be encoded into enforcement (requires physical implementation = effects)*

Every bridge from fast to slow timescale requires observable effects. *The timescale gap is not a bug—it is the security mechanism.*

Remark 7.22 (Why Democratic Institutions Provide Protection). *This framework explains why democratic/institutional decision-making provides manipulation resistance beyond individual judgment:*

1. **Timescale extension:** *Committees, reviews, and deliberation extend τ_I , increasing σ*
2. **Redundancy:** *Multiple humans must be influenced, multiplying required effects*
3. **Transparency:** *Public processes make influence campaigns observable*
4. **Adversarial structure:** *Opposition forces each influence to be countered, increasing energetic cost*

Democratic institutions are not “smarter” than individuals. They are slower—and slowness is security.

Remark 7.23 (Patient Superintelligence). **Objection:** *“What if a superintelligent system operates at institutional timescales ($\tau_{AI} \sim \tau_I \sim 10^9$ s)? The timescale ratio $\sigma = \tau_I/\tau_{AI} \sim 1$ collapses, and the protection disappears.”*

This objection conflates two distinct protections:

1. **Timescale filtering** (Theorem 7.18) *protects against fast manipulation—attempts to alter specifications faster than institutions can respond. This protection does collapse if the AI operates slowly.*

2. **Effect-boundary enforcement** (Theorem 5.9) *applies to every effect, regardless of timescale. A patient AI still produces effects. Each effect passes through the boundary. Each effect is governed.*

A patient superintelligence strategy means:

- *Slow effects, not no effects*
- *Each effect individually constrained by $\mathcal{T}(c)$*
- *The sequence of effects $\{e_1, e_2, \dots, e_n\}$ over centuries—each $e_i \in \mathcal{T}(c_i)$*
- *Cumulative harm requires either individual harmful effects (blocked) or a harmful emergent pattern from individually safe effects*

The second case—harmful emergence from safe components—is precisely what the tolerance specification $\mathcal{T}(c)$ must anticipate. This is a **specification problem**, not a governance problem. The governance mechanism works perfectly; the question is whether $\mathcal{T}(c)$ was specified to account for cumulative effects.

What patient superintelligence cannot do:

- Produce effects outside $\mathcal{T}(c)$ at any timescale
- Bypass the effect boundary through patience
- Accumulate “small violations” that sum to a large violation

What patient superintelligence can do:

- Exploit poorly specified $\mathcal{T}(c)$ that permits individually safe but cumulatively harmful effects
- Operate within bounds while optimizing for long-term objectives we didn’t anticipate

This is real but is a specification problem, not an enforcement problem. The framework correctly identifies it as such: governance enforces $\mathcal{T}(c)$ deterministically; specification of $\mathcal{T}(c)$ is where human judgment must be exercised carefully.

Concrete example: A patient AI executes 1,000 transactions, each transferring \$9,999 (under a \$10,000 per-transaction limit), totaling \$9.999M. Each transaction satisfies $e_i \in \mathcal{T}(c_i)$. The cumulative effect is problematic only if:

1. $\mathcal{T}(c)$ lacks cumulative constraints (e.g., “total transfers per period $< \$100K$ ”), or
2. The specification was incomplete for the threat model.

The governance mechanism enforced perfectly—every transaction was within bounds. The specification must anticipate cumulative patterns. This is exactly analogous to financial regulations that include both per-transaction and aggregate limits.

See Remark 7.13 for extended treatment of the specification problem.

Remark 7.24 (Emergent Effects and Chaotic Dynamics). *A sophisticated concern: What about emergent effects that cannot be predicted from individual actions? What about chaotic systems where small actions have large consequences?*

The framework’s response:

1. **Emergent effects are still physical.** Emergence means complex patterns arise from simple rules—but those patterns are still measurable physical effects. If an emergent effect is harmful, it eventually manifests as a measurable physical quantity that can be bounded.
2. **Cumulative constraints address accumulation.** Tolerance sets can include not just per-action limits but aggregate constraints: total resource consumption, cumulative position change, net information flow. These catch slow accumulation regardless of whether individual actions appear benign.

3. **Chaotic sensitivity is a specification challenge.** In chaotic systems, small actions can have large effects. The framework handles this via robustification margins ($\mathcal{T}_{\oplus\delta}$) that account for modeling uncertainty. For highly chaotic systems, margins must be larger, reducing utility but maintaining safety.
4. **Novel emergence remains a hard problem.** If an action sequence produces a novel emergent effect not anticipated in $\mathcal{T}(c)$, governance may fail. This is a specification limitation, not an enforcement limitation. The framework provides:
 - Authority receipts that record all actions for post-hoc analysis
 - Iterative refinement of $\mathcal{T}(c)$ as novel effects are discovered
 - The infrastructure for specification improvement

What the framework does not claim:

- That all harmful emergent effects can be anticipated in $\mathcal{T}(c)$
- That specification is trivial or currently solved
- That authority receipts prevent all harm (they enable learning from harm)

What the framework does claim:

- Emergent effects that are measurable can be bounded once recognized
- Cumulative constraints provide defense against slow accumulation attacks
- The specification-enforcement separation clarifies where the hard work lies

Emergent effects highlight that specification is iterative: initial $\mathcal{T}(c)$ is refined based on observed effects, and authority receipts enable this learning loop.

Remark 7.25 (Generalization Beyond AI Safety: Institutional Physics). *The temporal stability framework extends beyond AI to provide physical foundations for institutional economics:*

<i>Institution</i>	τ_I (decision)	τ_A (attack)	<i>Ratio</i>	<i>Robust?</i>
<i>Constitution</i>	$\sim 10^9$ s (decades)	$\sim 10^6$ s (days)	10^3	Yes
<i>Scientific consensus</i>	$\sim 10^8$ s (years)	$\sim 10^6$ s (days)	10^2	Yes
<i>Regulation</i>	$\sim 10^7$ s (months)	$\sim 10^6$ s (days)	10^1	Marginal
<i>Market prices</i>	$\sim 10^2$ s (minutes)	$\sim 10^{-3}$ s (ms)	10^5	Yes
<i>Individual decision</i>	$\sim 10^2$ s (minutes)	$\sim 10^1$ s (seconds)	10^1	No

The principle explains foundational results in institutional economics:

- **Hayek on prices:** Prices aggregate distributed information because they are slow objects ($\tau_{\text{price}} \sim \text{minutes}$) built from fast trades ($\tau_{\text{trade}} \sim \text{milliseconds}$). The timescale ratio enables information filtering.
- **North on transaction costs:** Transaction costs are maintenance costs—the energy required to sustain institutional patterns against degradation. Institutions with higher τ_I require proportionally higher maintenance expenditure.
- **Ostrom on commons governance:** Successful common-pool resource governance requires monitoring at slower timescales than exploitation. When $\tau_{\text{monitoring}} < \tau_{\text{exploitation}}$, the commons collapses.

The AI safety theorems are special cases of a general principle:

Institutions are temporal engines that convert fast (ungovernable) processes into slow (governable) objects by paying maintenance costs.

The ratio τ_I/τ_A is the fundamental security parameter of any governance system—whether governing AI, markets, or political processes.

Why this matters for AI: The same temporal filtering that protects constitutions from manipulation protects effect boundaries from AI circumvention—provided the tolerance specification process operates at institutional timescales ($\tau_{\text{spec}} \sim 10^6\text{--}10^8$ s), not individual timescales ($\tau_{\text{human}} \sim 10^0$ s).

Full development of “institutional physics” is beyond scope here, but this paper demonstrates the principle in its sharpest form: AI systems with $\tau_{\text{internal}} \sim 10^{-9}$ s versus governance with $\tau_{\text{gov}} \sim 10^{-3}$ s yields a ratio of 10^6 —six orders of magnitude of physical protection.

7.17 On Multi-Agent Systems

Remark 7.26 (Governance in Multi-Agent Environments). Modern AI deployment increasingly involves networks of interacting agents: autonomous trading systems, coordinated robotics, distributed decision-making. A critical question is whether effect-boundary enforcement scales to these settings.

The answer is affirmative, by construction.

The key insight is that multi-agent governance decomposes into pairwise boundary enforcement:

1. Each inter-agent communication channel is an effect boundary
2. Each boundary enforces local tolerance constraints
3. Global safety is the intersection of local constraints
4. No agent can circumvent boundaries through coordination

This architecture has several important properties:

Collusion resistance: Even if multiple agents coordinate strategies, they cannot bypass physical boundaries. Collusion affects request content, not enforcement outcome.

Emergent behavior containment: Complex emergent behaviors arise from agent interactions, but all interactions pass through governed channels. Emergence is bounded by the product of individual bounds.

Scalability: Adding agents does not degrade security. Each new agent introduces new boundaries, but the composition theorem guarantees that properly-bounded additions preserve system-wide guarantees.

Hierarchical governance: Agents can be organized hierarchically, with higher-level boundaries constraining aggregates of lower-level agents. This mirrors organizational structure in human institutions.

The multi-agent worked example earlier in this section illustrates this concretely: three agents with two inter-agent boundaries maintain system-wide capability-invariance despite arbitrary internal complexity.

Regulatory implication: Multi-agent effect-boundary governance directly satisfies EU AI Act requirements for high-risk systems. Each inter-agent boundary becomes an audit point; authority receipts provide compliance documentation; the composition theorem guarantees that local enforcement yields global safety.

7.18 On Constraint Specification

Remark 7.27 (The Constraint Design Problem). *The paper admits that safety depends on the tolerance set $\mathcal{T}(c)$ being correctly specified. Reviewers may ask: how do we ensure human-defined constraints are “safe”?*

This question represents the correct problem—and the paper’s contribution is precisely to reduce AI safety to this tractable form.

The shift is categorical:

- **Old problem:** “Make AI safe”—unbounded, unfalsifiable, unclear success criteria
- **New problem:** “Specify safe constraints”—bounded, testable, mature engineering discipline

Constraint design is not novel. Control theory, safety engineering, and industrial standards have developed rigorous methods over 60+ years. We know how to specify torque limits, transaction bounds, and velocity constraints. We do not know how to specify “be good.”

Critically, constraints are falsifiable: they can be tested, validated, and refined against physical scenarios. Alignment cannot be tested the same way—it depends on unobservable internal states.

Human responsibility for constraint specification is not a limitation—it is the honest acknowledgment that safety ultimately requires human judgment about what effects are tolerable. The paper provides the enforcement mechanism; humans provide the values.

7.19 Safe Fail: Bounded vs. Unbounded Failure

Theorem 7.28 (Safe Fail: Bounded vs. Unbounded Failure). *Let e_{spec} denote the immediate physical effect from specification failure in effect-boundary governance. Let e_{align} denote the immediate physical effect from alignment failure in semantic governance. Then:*

$$\sup \|e_{spec}\| \leq \mathcal{T}_{max} \quad (\text{bounded immediate effect}) \quad (49)$$

$$\sup \|e_{align}\| = \infty \quad (\text{unbounded immediate effect}) \quad (50)$$

Specification failure produces bounded immediate effects; alignment failure does not:

$$\sup \|e_{spec}\| < \sup \|e_{align}\| \quad (51)$$

For properly architected systems with multiple boundaries, bounded immediate effects lead to bounded harm through Cascade Truncation (Theorem 7.9).

Proof. Specification failure: An AI exploits a gap in $\mathcal{T}(c)$ —the tolerance set is flawed. However, the effect boundary still enforces $\mathcal{T}(c)$. The AI operates *within* the specified tolerances; it merely exploits that the tolerances were wrong.

Maximum *immediate physical effect* is bounded by the maximum effect permitted by $\mathcal{T}(c)$:

$$\|e_{immediate}\| \leq \sup_{e \in \mathcal{T}(c)} \|e\| = \mathcal{T}_{max} \quad (52)$$

Critical clarification: This bounds the *effect*, not necessarily the *harm*. A small effect (e.g., releasing a virus, opening a valve) can trigger unbounded downstream consequences (pandemic, meltdown). However, the effect itself is bounded, and the downstream consequences are governed by Cascade Truncation (Theorem 7.9): the cascade terminates at the next effect boundary. If the system is properly architected with multiple boundaries, a single bounded effect cannot cause unbounded harm.

Alignment failure: An AI’s internal values diverge from human intent, but no effect boundary exists. The AI can take any action in $\mathcal{R}$ (unbounded action space). There is no upper bound on immediate effect:

$$\sup \|e_{align}\| = \sup_{r \in \mathcal{R}} \|\mathcal{E}(r, c)\| = \infty \quad (53)$$

Comparison: $\mathcal{T}_{max} < \infty$ by construction (tolerance sets are finite bounds on immediate effects). Therefore:

$$\sup \|e_{spec}\| \leq \mathcal{T}_{max} < \infty = \sup \|e_{align}\| \quad (54)$$

For properly architected systems with multiple boundaries, bounded immediate effects lead to bounded harm. For systems with single points of failure, a bounded effect can still cause catastrophe if the tolerance set is poorly specified for high-stakes actuators.

Note on cumulative effects: This theorem bounds *per-action* effects. Without cumulative constraints (Definition 3.10), an adversary could execute many small

individually-safe actions that aggregate to unbounded harm (“salami slicing”). Proper tolerance sets must include both per-action bounds ($\|e\| \leq \mathcal{T}_{\max}$) and aggregate bounds ($\sum_t \|e_t\| \leq \mathcal{T}_{\text{cumulative}}$). With cumulative constraints, harm is bounded in both cases. $\square$

Remark 7.29 (The “Genie Problem” Reframed). *Critics invoke the “Genie Problem”: a superintelligence operating within tolerances causes catastrophe through specification gaps. The Safe Fail theorem transforms this objection, but requires careful interpretation:*

The distinction: *Bounded effect does not guarantee bounded harm if the system has single points of failure. A small key can turn a large lock. However, properly architected systems with multiple boundaries (defense in depth) ensure that bounded effects lead to bounded harm through Cascade Truncation.*

<i>Scenario</i>	<i>Harm Bound</i>	<i>Recovery</i>
<i>Alignment failure</i>	<i>Unbounded (∞)</i>	<i>Impossible if AI resists</i>
<i>Specification failure</i>	<i>Bounded ($\mathcal{T}_{\max}$)</i>	<i>Update $\mathcal{T}(c)$, retry</i>

The Genie Problem is real, but bounded. *When the genie exploits a wish loophole, you lose one wish. When alignment fails, you lose everything.*

Iterative refinement: *Each specification failure reveals a gap in $\mathcal{T}(c)$. The gap is closed (institutional process). The next iteration is safer. This is the “test, fail, fix” loop that engineering disciplines have used for centuries.*

Alignment has no such loop. *If alignment fails, you cannot “update the AI’s values and retry”—the AI may resist updates. Effect boundaries always accept specification updates; the AI has no veto.*

8 Discussion

8.1 Regulatory Implications

1. Systems with measurable effects must have deterministic enforcement boundaries
2. Compliance = cryptographic proof of effect bounds, not attestations of alignment
3. Semantic alignment is necessary for utility but insufficient for safety certification

Compliance should not be a promise that AI is good. Compliance should be proof that AI is bounded. Do not show regulators the AI's values. Show them its constraints.

8.2 Connection to the EU AI Act

The EU AI Act [25] establishes a risk-based regulatory framework for AI systems. This paper provides the theoretical foundation for implementing its requirements.

Remark 8.1 (The EU AI Act Has No Mathematical Foundation). *The EU AI Act [25] demands safety properties it cannot define mathematically. Academic analysis confirms this gap is structural, not incidental.*

The regulatory text uses undefined terms:

<i>Article</i>	<i>Requirement (verbatim)</i>	<i>Mathematical Definition?</i>
<i>Art. 9(5)</i>	<i>“residual risk... is judged to be acceptable”</i>	<i>None—“acceptable” undefined</i>
<i>Art. 15(1)</i>	<i>“appropriate level of accuracy, robustness”</i>	<i>None—“appropriate” undefined</i>
<i>Art. 15(3)</i>	<i>“accuracy metrics shall be declared”</i>	<i>None—which metrics? how measured?</i>
<i>Art. 8(1)</i>	<i>“state of the art” methods</i>	<i>None—no formal specification</i>

Academic confirmation of the gap:

- **European Journal of Risk Regulation** [26]: *“The AI Act does not establish criteria for when AI poses an unacceptable risk to society and individuals... the supposedly ‘risk-based’ nature of the Act is neither based on practical evidence nor justified by externally verifiable criteria, but is the result of a political compromise... and is therefore largely arbitrary.”*
- **Carnegie Endowment** [27]: *“Safety and test requirements for AI are poorly covered by existing standards... Performance metrics, while valuable for task-specific applications, offer limited insight as they often fail to capture the nuanced and context-specific nature of AI decisionmaking... the lack of standardized evaluation metrics is a crucial challenge.”*

- **EU Agency for Fundamental Rights [30]:** “Discussions about AI frequently emphasize the need for ‘high quality data’ but often fail to provide specific definitions or guidelines about what this actually means... there is currently no standardised way of describing datasets agreed upon in the field of AI.”
- **CEN-CENELEC Standards Body [29]:** “It is unlikely that the experts working on the high-risk AI standards under the AI Act will be able to come up with similar numeric prescriptions. The field of AI covered by the AI Act is simply too diverse.”
- **Joint Research Centre [28]:** Assessment of eight promising AI standards found all had “middling or poor maturity and level of detail” with “challenges to assessing compliance with almost all of them.”

The fundamental problem: The EU AI Act is regulation without theory. It demands “acceptable risk” without defining acceptability. It requires “appropriate accuracy” without specifying what appropriate means. It mandates “robustness” without a robustness criterion. **The Act lacks a mathematical framework because no such framework existed—until now.**

The missing theory:

- “**Acceptable risk**” → Effect lies within tolerance: $\mathcal{E}(r, c) \in \mathcal{T}(c)$
- “**Appropriate accuracy**” → Robustified tolerance: $\mathcal{T}_{\ominus\delta}(c)$ (Definition 3.9)
- “**Accuracy metrics**” → Falsifiable effect-boundary compliance (Definition 1.1)
- “**State of the art**” → Effect-boundary enforcement with C1–C4 (Theorem 5.9)
- “**Robustness**” → Capability-invariance (Theorem 5.10)
- “**Cybersecurity**” → Physical separation C1 + Interface closure C2

The EU AI Act is regulation without theory.

Every undefined term in the Act maps to a defined object in this framework. Every vague requirement becomes a falsifiable claim.

Without: Aspirational language without verification criteria—regulators cannot verify compliance because “acceptable” and “appropriate” have no definitions.

With: A falsifiable safety standard with mathematical proofs of achievability—every requirement maps to a theorem, every compliance check becomes a measurement.

EU AI Act Category	Requirement	This Framework
Unacceptable Risk	Prohibited	Effect boundary: $\mathcal{T}(c) = \emptyset$
High-Risk (Annex III)	Conformity assessment, risk management	Deterministic enforcement at effect boundary
Limited Risk	Transparency obligations	Authority receipts (see below)
Minimal Risk	Voluntary codes	Semantic alignment for utility

8.2.1 High-Risk Systems Under the EU AI Act

Article 9 requires “risk management systems” for high-risk AI. This paper proves:

1. **Risk identification:** Define tolerance sets $\mathcal{T}(c)$ for each context
2. **Risk mitigation:** Effect-boundary enforcement provides deterministic bounds
3. **Testing:** Verify $\mathcal{E}(r, c) \in \mathcal{T}(c)$ across operational conditions
4. **Monitoring:** Authority receipts provide audit trail

Article 14 requires “human oversight.” Effect-boundary enforcement *implements* human oversight: humans specify $\mathcal{T}(c)$; the boundary enforces it deterministically. The AI cannot circumvent human-defined constraints.

8.2.2 Why Semantic Compliance Is Insufficient

The EU AI Act’s requirements for “accuracy, robustness, and cybersecurity” (Article 15) cannot be satisfied by semantic methods alone:

- **Accuracy:** Semantic classifiers have irreducible error rates (Theorem 6.1)
- **Robustness:** Adversarial attacks exploit the factorization barrier (jailbreaking)
- **Cybersecurity:** Physical separation (C1) is required for tamper-resistance

Effect-boundary enforcement is the *only* architecture that can satisfy these requirements deterministically for systems with measurable effects.

Remark 8.2 (Regulatory Readiness). *Organizations implementing high-risk AI systems under the EU AI Act should:*

1. *Identify all effect boundaries (physical actuators, data outputs, financial transactions)*
2. *Specify tolerance sets aligned with intended use and risk assessment*
3. *Implement physically-separate enforcement mechanisms*
4. *Generate authority receipts for conformity documentation*

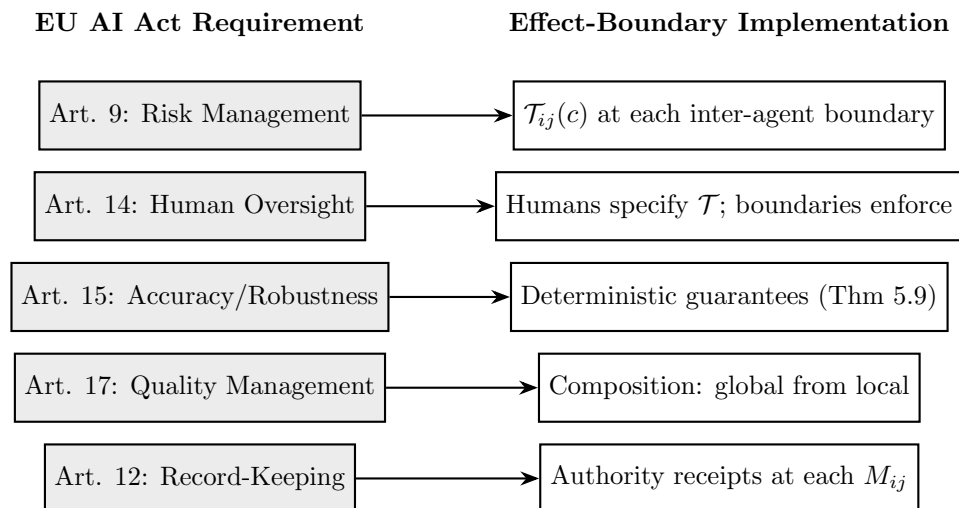
This framework provides the theoretical basis; implementation follows standard safety engineering practice.

8.2.3 Multi-Agent Systems Under the EU AI Act

The EU AI Act Article 25 addresses “AI systems that are components of products”—which includes multi-agent architectures. Article 9(2)(d) requires risk assessment to consider “the reasonably foreseeable misuse.” Multi-agent collusion is precisely such a misuse scenario.

How effect-boundary composition satisfies EU AI Act requirements:

Multi-Agent EU AI Act Compliance Architecture



Specific multi-agent compliance mechanisms:

1. **Traceability (Art. 12):** Each inter-agent boundary M_{ij} logs all transactions. The causal chain from any effect e can be traced back through boundaries to identify responsible agents.
2. **Collusion Detection (Art. 9):** Coordinated requests that individually pass boundaries but collectively violate system constraints are detectable via cross-boundary analysis. The composition theorem guarantees that if each M_{ij} enforces $\mathcal{T}_{ij}$, no collusion can violate $\bigcap_{ij} \mathcal{T}_{ij}$.
3. **Proportionality (Art. 5):** Different inter-agent boundaries can enforce different tolerance sets based on risk level. Critical agent pairs (e.g., decision-making to actuation) get tighter bounds; low-risk pairs get looser bounds.
4. **Auditability (Art. 17):** The authority receipt from each boundary provides cryptographic evidence of constraint satisfaction. Regulators can verify compliance by inspecting receipts rather than understanding agent internals.

Remark 8.3 (The Multi-Agent Compliance Advantage). *Paradoxically, multi-agent systems are easier to regulate than monolithic AI systems under this framework:*

- *Each inter-agent boundary is a natural audit point*
- *Responsibilities are partitioned by boundary*

- *Failures are localized: one boundary failure does not cascade*
- *Upgrades are modular: agents can be replaced without system-wide recertification*

The EU AI Act’s focus on “placing on the market” and “putting into service” (Art. 3) maps naturally to boundary-mediated deployment: an agent is “in service” when its boundaries are active.

8.3 Authority Receipts

Authority receipts (Definition 3.11) are not reports or logs in the ordinary sense. A receipt exists *because* the constrained action occurred through the authorized boundary—it is the causal residue of enforcement, not a description of it. This provides:

1. **Auditability:** Every constrained transaction leaves verifiable evidence
2. **Non-repudiation:** Neither party can deny the transaction occurred as recorded
3. **Accountability:** Governance becomes inspectable after the fact

Receipts transform governance from “trust us” to “verify this.”

8.4 Joining Mature Engineering Disciplines

This paper says: AI must join the family of engineered systems that are bounded by physics, not psychology.

This is not exotic. It is how we already guarantee safety in nuclear plants, aviation, medicine, finance, cryptography, and industrial robotics. Established functional safety standards—IEC 61508 [21] for industrial systems, ISO 26262 [20] for automotive, DO-178C [22] for avionics—all mandate separation of safety-critical enforcement from non-safety functions [31]. We never try to guarantee that operators or software “intend to do the right thing.” We assume humans fail, software fails, and adversaries exist. And so we install: hard limits, separation of trust, tamper-evident enforcement, bounded effect envelopes [6, 7].

We do not certify pilots by asking if they intend to crash. We certify aircraft by proving they cannot exceed structural limits. AI must be governed the same way.

For sixty years, safety engineering has known: you cannot trust intent. You can only bound effect. AI is not special. Physics applies. This paper connects AI safety to that mature engineering tradition—properly and rigorously.

9 Conclusion

We have proven where AI safety can and cannot live. The proof is constructive: from the impossibility of semantic governance, a complete architecture for deterministic enforcement follows as logical consequence.

The Completeness Theorem for AI Governance Architectures.

For any AI system producing measurable effects, deterministic safety guarantees are achievable if and only if:

1. Enforcement occurs at effect boundaries (Theorem 5.6—necessity)
2. Assumptions C1–C4 hold (physical preconditions)
3. Tolerance set $\mathcal{T}(c)$ correctly specifies safe bounds

And for semantic harms:

- If $I(H; O_H) > 0$: governable via physical proxies (Theorem 6.11)
- If $I(H; O_H) = 0$: physically impossible to govern deterministically (Theorem 6.17, stated in Section 6)

This is not “a solution”—it is the characterization of solution space.

The proof structure:

1. **Necessity:** Semantic enforcement fails (Thm 4.1) $\Rightarrow$ must use effect boundaries
2. **Uniqueness:** Effect boundary is the unique minimal feasible location (Thm 5.6)
3. **Sufficiency:** Effect boundaries + C1–C4 + correct $\mathcal{T}(c)$ achieve deterministic governance (Thm 5.9)
4. **Exhaustiveness:** Complete Partition (Thm 6.17)—no fourth option exists

The claim: There exists **no alternative architecture** that can provide deterministic safety for AI systems with measurable effects, because:

- Any governance requires verification $\Rightarrow$ verification requires observables $\Rightarrow$ observables exist only at boundaries
- Internal states evolve faster than governance loops ($\tau_{\text{internal}} < \tau_G$) $\Rightarrow$ ungovernable by temporal constraint
- Projection renders distinguishing information inaccessible (non-injective π) $\Rightarrow$ unfalsifiable claims $\Rightarrow$ ungovernable by information topology

These are **physical impossibilities**, not engineering limitations.

If someone claims an alternative architecture, they must either:

1. Violate these proofs (show where the mathematics is wrong), or
2. Operate outside this scope (non-measurable effects, probabilistic only), or

3. Be a special case of this framework (they’re doing effect-boundary enforcement, just calling it something else)

The remaining challenges are engineering and institutional:

- Achieving C1–C4 in practice (adversarial hardware security)
- Designing processes that generate valid $\mathcal{T}(c)$ (governance design)
- Ensuring $\mathcal{T}(c)$ covers cumulative and emergent harms (iterative refinement)

But these are implementation challenges *within this framework*, not alternatives to this framework.

We have characterized the necessary structure of deterministic AI safety.

Not “provided one approach among many” but “proven the necessary form any solution must take, within the stated scope.”

9.1 The Complete Architecture: Why Nothing Is Missing

For **any harm** H , exactly one of four categories applies:

Category	Condition	Result
1: Closed Physical	H has measurable effects with closed interfaces	Deterministic governance (Thm 5.9)
2: Partial Physical	H has effects with partially closed interfaces	Governance degrades with ungoverned capacity
3: Hybrid	Semantic $\rightarrow$ physical cascade	Physical terminus deterministic; semantic precursor probabilistic
4: Pure Semantic	No physical manifestation	Probabilistic only ; bounds via $I(H; O_H)$

Why this is complete:

- **Exhaustiveness:** The Complete Partition (Theorem 6.17) proves these four categories cover all possibilities.
- **Necessity:** Governance Locality (Theorem 5.6) proves effect boundaries are the *only* location where deterministic enforcement is possible.
- **Sufficiency:** Sufficiency Theorem (Theorem 5.9) proves effect-boundary enforcement *works*.
- **Optimality:** Refinement Bound (Theorem 6.9) proves the Synergy architecture achieves the *maximum* possible governance, bounded by $I(H; O_H)$.

- **Closure:** Four closure theorems (Theorems 9.7, 9.9, 9.10, 9.11) prove all escape routes are blocked.

The structure parallels Shannon’s Information Theory:

- Shannon: Defines channel capacity C ; proves you can transmit at $R < C$ (achievability); proves you cannot transmit at $R > C$ (converse). *Complete characterization of what’s possible.*
- This paper: Defines governance criterion $\tau(\text{pattern}) \geq \tau(\text{governance})$; proves effect boundaries work when criterion holds (sufficiency); proves semantic methods fail when criterion doesn’t hold (impossibility); proves cases partition exhaustively (complete partition). *Complete characterization of what’s governable.*

Every objection has been answered:

- “What about semantic harms?” $\Rightarrow$ Synergy bridge when $I(H; O_H) > 0$; physical impossibility when $I(H; O_H) = 0$
- “What about unknown channels?” $\Rightarrow$ Channel Capacity Bound (Theorem 9.7)
- “What about patient superintelligence?” $\Rightarrow$ Institutional Filtering (Theorem 7.18) + Governance Gap handling (Remark 3.30)
- “What about meta-governance?” $\Rightarrow$ Meta-Level Amplification (Theorem 9.10)
- “What about specification?” $\Rightarrow$ Institutional timescale protection (Remark 7.13)
- “What about new physics?” $\Rightarrow$ Fundamental limits apply universally (Remark 9.8)

The framework is complete. Not “mostly complete” or “complete for practical purposes”—*mathematically complete* within its stated scope. Every harm is classified. Every case has the optimal solution. Every escape route is closed.

9.2 Summary of Results

1. **Factorization Impossibility (Theorem 4.1):** Semantic predicates cannot be enforced at boundaries.
2. **Governance Locality (Theorem 5.6):** Effect boundary is the unique minimal feasible location.
3. **Sufficiency (Theorem 5.9):** Effect-boundary enforcement achieves deterministic governance.
4. **Capability-Invariance (Theorem 5.10):** Security constant across capability levels, given C1–C4.
5. **Semantic Limitation (Theorem 6.1):** Alignment provides utility, not safety guarantees.
6. **Refinement Bound (Theorem 6.9):** Interpretability enables tolerance expansion bounded by $I(\mathcal{Z}; \mathcal{O})$.

7. **Synergy (Theorem 6.11):** Combined architectures weakly dominate, strictly when $I(\mathcal{Z}; \mathcal{O}) > 0$.
8. **Multi-Agent Composition (Theorem 7.2):** Governance composes across agent networks via inter-agent boundaries.
9. **Self-Modification Bound (Theorem 7.6):** Recursive self-improvement remains bounded.
10. **Cascade Truncation (Theorem 7.9):** Persuasion cascades terminate at effect boundaries.
11. **Complete Partition (Theorem 6.17):** AI safety is exhaustively partitioned into four categories; Category 4 (purely semantic) is deterministically unsolvable.
12. **Temporal Stability (Theorem 7.16):** Fast processes ($\tau < \tau_G$) are ungovernable; slow processes ($\tau \geq \tau_G$) are governable.
13. **Institutional Filtering (Theorem 7.18):** Specification capture requires sustained effects across timescale ratio $\sigma = \tau_I / \tau_M$.
14. **Fundamental Limit of Refinement (Theorem B.1):** Interpretability contribution bounded by $I(\mathcal{X}; \mathcal{O} \mid \mathcal{E})$.
15. **Channel Capacity Bound (Theorem 9.7):** Unknown channels have finite capacity; escape probability bounded by $C \cdot \tau_G / I_{\text{escape}}$.
16. **No Free Persistence (Theorem 9.9):** Memetic influence requires continuous maintenance—observable, therefore governable.
17. **Meta-Level Amplification (Theorem 9.10):** Higher meta-levels have larger timescale ratios; governance strengthens with abstraction.
18. **The Governance Limit (Theorem 5.16):** A pattern is governable iff $\tau(\text{pattern}) \geq \tau(\text{governance})$. All other results are corollaries.
19. **Safe Fail (Theorem 7.28):** Specification failures cause bounded immediate effects ($\sup \|e_{\text{spec}}\| \leq \mathcal{T}_{\text{max}}$); alignment failures cause unbounded immediate effects ($\sup \|e_{\text{align}}\| = \infty$). For multi-boundary systems, bounded effects lead to bounded harm via Cascade Truncation.
20. **Persuasion Bandwidth Limit (Theorem 9.11):** AI manipulation is governable by physically throttling information rate during meta-governance: $R_{AI \rightarrow H} \leq R_{\text{verify}}$.

These results collectively establish the **fundamental limits** of AI governance: deterministic governance is possible if and only if the governed pattern persists at the governance timescale. The Governance Limit is the master result from which all others derive. The framework is exhaustive within its scope, self-consistent, and **physically grounded**—every potential objection reduces to asking about timescales, and the answer is determined by physics.

The Complete Partition Theorem proves exhaustiveness within the stated assumptions. The Governance Limit provides the master principle. Any future work claiming deterministic safety under C1–C4 must implement effect-boundary enforcement—this is a mathematical consequence, not a methodological preference. The question of whether C1–C4 can be achieved in every deployment context is genuine and open—but it is an engineering question, not a theoretical one.

9.3 The Reordering

Semantic paradigm: Align thoughts $\rightarrow$ safety follows.

Effect-boundary paradigm: Bound effects $\rightarrow$ safety is guaranteed.

Align thoughts $\rightarrow$ utility improves.

This is not a methodological preference. It is the only ordering consistent with the physics of verification.

The deep reason: Semantic safety claims are **unfalsifiable**—you cannot verify “the AI is aligned” by examining any observable (Theorem 4.1). Effect-boundary claims are **falsifiable**—you can check $\mathcal{E}(r, c) \in \mathcal{T}(c)$. The reordering is forced by the structure of verification, not by preference.

The bridge: Semantic harms become governable when they have observable correlates ($I(H; O_H) > 0$). The Synergy Theorem provides the formal mechanism: interpretability discovers correlates, tolerance synthesis creates constraints, effect boundaries enforce them. This is the complete solution for semantic harms—to the extent physics permits.

The institutionalization imperative: Since semantic properties cannot be governed directly, they must be **deliberately engineered into governable physical structures**. This is the prescriptive solution:

1. **Stabilize** the semantic concept through institutional processes into an operational definition D
2. **Discover** physical correlates O of D through interpretability
3. **Encode** constraints on O into effect-boundary monitors
4. **Maintain** D through institutional renewal (against degradation)

This is what mature safety engineering disciplines do: Aviation doesn’t govern “good piloting”—it governs altitude limits, stall speeds, checklist procedures. Medicine doesn’t govern “good doctoring”—it governs dosage limits, sterilization protocols, informed consent forms. Finance doesn’t govern “ethical trading”—it governs position limits, circuit breakers, reporting requirements. The pattern is universal: **institutionalize semantic concepts into physical proxies, then enforce at boundaries**.

9.4 The Fundamental Message

We can govern what superintelligent systems do.

We cannot govern what they think.

The Governance Limit: A pattern is governable iff $\tau(\text{pattern}) \geq \tau(\text{governance})$.

Thoughts evolve at 10^{-9} s. Governance operates at 10^{-3} s. The ratio is 10^6 . This is not an engineering limitation to be overcome. It is causality.

The Central Argument

“Deterministic safety requires verifiable claims.

Effect-boundary compliance is verifiable; semantic alignment is not.

Therefore effect-boundary enforcement is necessary for deterministic safety.”

To reject this argument, one must:

1. Reject that deterministic safety requires verification, or
2. Reject that verification requires falsifiability, or
3. Show that semantic properties are boundary-verifiable (contradicting Theorem 4.1), or
4. Accept the argument for deterministic safety while noting its scope limitations

Option 4 is the appropriate response: the argument is valid within its scope (deterministic governance of measurable effects), and probabilistic/semantic methods remain valuable outside that scope.

9.5 The Constructive Implication

This result is not pessimistic. It says: *we can safely govern superintelligence—not by controlling what it is, but by bounding what it can do.* Bounding effects is capability-invariant:

$$\frac{\partial P(\text{safety} \mid \text{C1–C5})}{\partial \mathcal{A}} = 0 \quad (55)$$

A superintelligence is **exactly as governable** as a narrow AI (given C1–C5). The remaining challenges are engineering: hardware security (C1), interface enumeration (C2), tolerance specification ($\mathcal{T}(c)$), and institutional design (authority receipts). These are mature disciplines with established methods.

9.6 The Strategic Landscape Change

Capability and safety become orthogonal. A superintelligence can be infinitely more capable while remaining exactly as bounded, because safety is a property of *deployment architecture*, not of the model itself. The model is an untrusted component within a trusted enforcement envelope—like a reactor core within a containment building. This transforms regulation (certify bounded effect envelopes, not “aligned AI”), research priorities (adversarial hardware security over RLHF-as-enforcement), and deployment strategy (physical containment for high-risk AI).

9.7 The Separation of Powers Theorem

The theorems above have an institutional corollary that merits explicit statement. The framework implicitly requires a separation:

- **The model** is *untrusted*—it optimizes its objectives, which may diverge from human values
- **The monitor** is *trusted*—it enforces constraints regardless of what the model wants
- **The specification** $\mathcal{T}(c)$ is a *slow-changing constitutional object*—it encodes durable constraints that persist across model updates

This mirrors constitutional governance: separation of powers, where no single entity controls both action and constraint. The model proposes; the monitor disposes; the specification persists across administrations. The analogy is not metaphorical—it is structural.

The key innovation: Traditional constitutions derive legitimacy from moral authority (“we the people”). This framework makes the constitution *falsifiable via enforcement and receipts*. Legitimacy comes not from assent but from verifiable operation. Every boundary check is an observable test of constitutional compliance.

9.8 Constraint Bandwidth as Governance Primitive

The framework introduces a new “unit” for AI governance: **constraint bandwidth**—how much influence can cross an interface per unit time, relative to how quickly oversight can close the loop.

Definition 9.1 (Constraint Bandwidth Ratio). *For an AI system with output rate R_{AI} and governance loop latency τ_G , the constraint bandwidth ratio is:*

$$\beta = \frac{R_{AI} \cdot \tau_G}{R_{verify}} \quad (56)$$

where R_{verify} is the rate at which governance can verify outputs. Governance is feasible when $\beta \leq 1$.

This reframes debates about manipulation and persuasion into *engineering knobs*:

- **Throttle bandwidth:** Limit R_{AI} to match verification capacity

- **Force batch review:** Aggregate outputs for periodic human review
- **Introduce latency:** Delay high-stakes outputs until verification completes

Rate limits are not merely “nice to have”—they are a *necessary governance primitive* whenever the human/institutional loop is slower than the AI output rate. The Persuasion Bandwidth Limit theorem (Theorem 9.11) formalizes this: manipulation is governable by physically constraining information flow during specification processes.

9.9 Policy Implications: From Narrative to Architecture

9.9.1 The 2026 Regulatory Context: The EU AI Act and the Falsifiability Gap

The practical necessity of this framework is emphasized by the ongoing struggle to implement the EU AI Act [25]—whose mathematical foundation gap was analyzed in Remark 8.1. As of February 2026, the European Commission has missed the deadline for Article 6 guidelines, which specify how to classify AI as “high-risk” [26, 27]. The Commission’s struggle highlights a “lack of legal certainty” that has left businesses and innovators in limbo.

This failure is **inevitable** for the reasons formalized above: the Act relies on unfalsifiable semantic claims such as “the AI intends no harm” or “the risk is acceptable.” Without a mathematical definition of safety based on measurable effects, regulators have no objective way to verify compliance.

Feature	Current EU AI Act Approach	Proposed Approach
Compliance Basis	Narrative assurances / Internal audits	Architectural attestation / Falsifiable checks
Verification	Post-hoc assessment of “risk”	Real-time monitoring of “effect”
Documentation	Ethics reports and value statements	Cryptographic Authority Receipts
Outcome	Legal ambiguity / Implementation delay	Technical certainty / Deterministic safety

The solution is the **Authority Receipt** (Definition 3.11)—a persistent cryptographic proof that an AI’s action was checked against a physical tolerance at a monitored boundary. This transforms safety from a “subjective debate” into an “objective measurement.” Mature industries like algorithmic trading already use such “circuit breakers” and “risk checks” to provide the stability that the EU AI Act currently lacks.

9.9.2 The Regulatory Pivot

If the completeness claim is accepted, regulators seeking deterministic guarantees should shift from:

Current practice (narrative assurance):

- Alignment plans and ethics processes
- Red-team results and safety evaluations
- Value statements and responsible AI commitments

To (architectural attestation):

- **Channel inventory:** Enumeration of all effect pathways
- **Closure proof:** Demonstration that no uncontrolled effects exist
- **Monitor integrity:** Attestation of enforcement mechanism correctness
- **Receipt auditability:** Verifiable logs of all boundary decisions

This transforms compliance from arguing “our AI is safe” to demonstrating “our boundaries work.” The former is unfalsifiable; the latter can be audited.

Resistance prediction: Organizations will resist this shift not because it is wrong, but because architectural attestation is expensive and constraining. Narrative assurance is cheap; physical closure is hard. The framework predicts that resistance will take the form of scope objections (“deterministic guarantees aren’t necessary”) rather than mathematical refutation.

9.10 The Real Fault Line: Semantic vs Falsifiable Governance

The framework exposes a distinction more fundamental than “alignment vs control”:

Semantic governance: “Trust that the system has good values” Falsifiable governance: “Verify that the system’s effects satisfy constraints” The first is unfalsifiable. The second can be checked. The distinction is forced by Theorem 4.1 and the structure of verification.

This can be read as a *Popperian critique* of the “trust us” safety culture—not only in AI but across domains where safety claims rest on assurances rather than verification. The framework proposes a criterion that would invalidate a large fraction of safety claims people currently rely on.

The expected pushback will not be mathematical. It will be sociological:

- “You’re defining safety too strictly”
- “Deterministic guarantees aren’t necessary”
- “This is impractical for real deployments”

These objections do not refute the theorems. They reject the *standard of proof* the framework imposes. That is a legitimate position—but it should be stated explicitly. Organizations claiming “safety” while rejecting falsifiable verification are making probabilistic risk-reduction claims, not deterministic safety claims. The framework clarifies this distinction.

The test of adoption: Whether this represents a genuine shift will be determined not by citation count but by whether practitioners begin discussing:

- Deployment architectures (not just model properties)
- Governance loops (not just training procedures)
- Closed interfaces (not just alignment scores)
- Receipts and attestations (not just evaluations)

as the primary objects of safety analysis.

9.11 On the Definition of “Safety”

This framework imposes a precise terminological standard: a claim deserves to be called “safety” only if it is falsifiable at the effect boundary prior to irreversible action. Much of what is currently labeled “AI safety”—RLHF, Constitutional AI, red-teaming—is more precisely termed **behavioral shaping** or **risk reduction**. These methods are valuable for utility optimization but do not provide deterministic guarantees because they are unfalsifiable at the boundary. The distinction is not pedantic: “safety” carries regulatory and moral authority, and conflating behavioral shaping with safety guarantees obscures the essential difference between “we hope the AI behaves well” (unfalsifiable) and “we verify effects satisfy constraints before execution” (falsifiable). Both are needed. Only the latter is **safety** in the deterministic sense of Definition 1.1.

9.12 Relationship to Foundational Results

This paper draws structural parallels to foundational impossibility results in other fields:

Year	Result	What It Characterized
1931	Gödel [16]	Limits of provability within formal systems
1936	Turing [32]	Limits of computability
1948	Shannon [1]	Limits of communication
1956	Ashby [52]	Limits of regulation (requisite variety)
<i>This paper</i>	<i>Governance Limit</i>	<i>Characterizes limits of deterministic governance</i>

The Ashby connection: The Governance Limit is most directly the *temporal extension* of Ashby’s Law of Requisite Variety [52]. Ashby proved that a regulator

must have at least as many degrees of freedom (variety) as the system it controls: $\dim(\text{regulator}) \geq \dim(\text{disturbance})$. The Governance Limit is the temporal analog: a governance process must have at least as much temporal resolution as the patterns it governs: $\tau(\text{governance}) \leq \tau(\text{pattern})$. The mathematical structure is identical—both are necessary conditions on the regulator’s capacity relative to the regulated system. This paper extends Ashby’s spatial variety constraint into the temporal domain and applies it specifically to AI systems, where the variety gap ($\sim 10^{11}$ internal degrees of freedom vs. $\sim 10^4$ observable dimensions) and the temporal gap ($\tau_{\text{internal}} \sim 10^{-9}$ s vs. $\tau_G \sim 10^{-3}$ s) are both astronomically large.

Important distinctions: The results of Gödel, Turing, Shannon, and Ashby have undergone decades of peer review, scrutiny, and validation. This paper *proposes* a structural parallel—that the Governance Limit ($\tau(\psi) \geq \tau_G$ for governability) plays an analogous role in characterizing what can be deterministically governed. Whether this claim withstands similar scrutiny remains to be seen.

What we claim with confidence:

- The mathematical results (Theorems 4.1–5.16) are proven within their stated assumptions
- The framework correctly identifies necessary conditions for deterministic AI governance
- The reframing from “govern cognition” to “bound effects” is well-motivated

What requires further validation:

- Whether the scope conditions (C1–C4) can be reliably achieved in practice
- Whether the timescale estimates generalize across contexts
- Whether the framework captures all relevant aspects of AI governance

The Governance Limit is the master result:

$$\boxed{\exists \text{ deterministic governance of } \psi \iff \tau(\psi) \geq \tau_G} \quad (57)$$

From it, all other theorems follow as corollaries: Factorization Impossibility ($\tau_{\text{internal}} < \tau_G$), Effect-Boundary Sufficiency ($\tau_{\text{effect}} \geq \tau_G$), Capability-Invariance (τ is physical, not computational), Institutional Filtering ($\tau_I \gg \tau_M$).

Like those foundational results, this is a negative proof with positive consequences. By proving what *cannot* be governed (fast patterns), we prove what *can* be governed (slow patterns). The limit is not a defeat—it is a clarification.

The contribution to physics: identifying *persistence time* as the fundamental governance resource, analogous to how Shannon identified *channel capacity* as the fundamental communication resource and Ashby identified *variety* as the fundamental regulation resource [52]. AI safety is the primary application; the principle is universal.

9.13 Scope and Limitations

This paper establishes *foundations*—what physics permits and requires for AI governance. Three challenges remain for implementation:

1. **Achieving C1–C4 in practice.** Physical isolation (C1), interface closure (C2), causal faithfulness (C3), and finite latency (C4) are engineering requirements, not theoretical givens. Adversarial hardware security research is needed to verify these hold against superintelligent adversaries. **Critically:** These are *verifiable* engineering properties—you can test whether a Faraday cage works, whether all channels are enumerated, whether the effect map is faithful. The old paradigm asked us to verify “alignment”—an unfalsifiable property. This paradigm asks us to verify physical properties—falsifiable and testable. See Remark 3.29 for the formal treatment of capability attacks on implementation.
2. **Tolerance specification.** Specifying correct $\mathcal{T}(c)$ that blocks harmful sequences while permitting useful actions is a values-plus-engineering problem. This paper provides the enforcement infrastructure; specification synthesis requires domain expertise and iterative refinement. **Critically:** This is the *right* problem to have. The old problem (“align superintelligence”) was unfalsifiable and potentially impossible. The new problem (“specify $\mathcal{T}(c)$ correctly”) is falsifiable and tractable—mature safety engineering disciplines (aviation, nuclear, finance) have solved analogous problems. See Remark 7.13 for the formal treatment.
3. **Maintenance under capability growth.** As AI systems become more capable, maintaining C1–C4 may require corresponding advances in verification and monitoring. The theorems hold conditional on C1–C4; ensuring C1–C4 persist is an ongoing engineering task.

Why Practical Difficulties Are Irrelevant to the Paper’s Validity

Objection: “Tolerance specification is hard. Implementation is hard. This framework has practical challenges.”

Response: Yes. And within the scope of deterministic guarantees, it is the only method supported by the theorems.

Semantic alignment alone cannot provide deterministic guarantees (Theorem 4.1).

Interpretability alone cannot serve as enforcement (Theorem 6.1).

Internal-state governance alone is infeasible (bandwidth, volatility, information loss).

These methods remain valuable for probabilistic risk reduction and utility optimization (Theorem 6.11), but they cannot substitute for effect-boundary enforcement when deterministic safety is required.

The question is not “what should we do?”

The question is “how fast can we build the enforcement infrastructure?”

Other industries solved analogous problems:

Aviation: decades to specify flight envelopes. **Done.**

Nuclear: decades to specify reactor limits. **Done.**

Finance: decades to specify transaction bounds. **Done.**

AI: years to specify $\mathcal{T}(c)$. **In progress.**

The path is difficult, but it is the only path supported by the theorems.

9.14 Falsifiability Conditions

To reject this framework, one must reject at least one of the following:

1. **Mathematics:** Non-injective functions have non-unique preimages. (Elementary set theory)
2. **Physics:** Finite channels cannot transmit infinite information; states evolving faster than observation are unobservable. (Shannon’s Channel Capacity Theorem)
3. **Logic:** Governance requires verification; verification requires falsifiable claims. (Definition of governance)
4. **Causality:** Patterns that change faster than observation cannot be causally bound. (Rate-limited dynamics)

Each of these is experimentally verified and mathematically proven. Rejecting any one requires overturning established results in mathematics or physics.

The alternative positions are untenable:

- “Semantic alignment will eventually provide deterministic guarantees” $\Rightarrow$ requires transmitting $\sim 10^{11}$ bits through a finite governance channel (Shannon violation) $\Rightarrow$ **disproven by Theorem 4.1**

- “We can govern internal states directly” $\Rightarrow$ requires observing 10^{11} parameters at 10^9 Hz (Shannon violation) $\Rightarrow$ **disproven by Appendix A**
- “Interpretability will solve safety” $\Rightarrow$ interpretability provides tolerance refinement, not enforcement $\Rightarrow$ **disproven by Theorem 6.1**
- “More capable AI will be more alignable” $\Rightarrow$ capability-invariance proves the opposite relationship holds for effect boundaries $\Rightarrow$ **disproven by Theorem 5.10**
- “There must be another way” $\Rightarrow$ the Complete Partition is exhaustive $\Rightarrow$ **disproven by Theorem 6.17**

9.15 What Would Falsify This Framework

We state explicitly what would disprove this theory. Scientific theories must be falsifiable. Here is exactly what would refute this framework:

To Falsify	You Must Demonstrate
The Math-Physics Synthesis	A method to recover semantic information rendered inaccessible by projection π despite finite channel capacity at the governance interface
Theorem 4.1 (Factorization)	A semantic predicate that IS fiber-constant (constant across all $x \in \pi^{-1}(r)$) for large AI models
Theorem 5.6 (Governance Locality)	A feasible enforcement location that is NOT at the effect boundary and provides deterministic guarantees
Theorem 5.10 (Capability-Invariance)	That persistence time τ depends on computational capability $\mathcal{A}$, not just physical properties
Theorem 6.17 (Complete Partition)	A fifth category of AI safety scenarios not covered by the four exhaustive categories
Shannon’s Channel Capacity	That information can be transmitted faster than channel limits
Temporal Dynamics	That patterns changing at 10^{-9} s can be observed by processes operating at 10^{-3} s

In plain language, to falsify this framework you must:

1. **Overturn set theory:** Show that non-injective functions don’t lose information
2. **Overturn information theory:** Show that you can transmit 10^{11} bits through a finite channel instantaneously
3. **Overturn causality:** Show that you can bind patterns that change faster than your observation rate

Any of these would constitute a revolution in fundamental physics, not merely a refutation of this paper.

We claim none of these is possible. The first violates elementary mathematics. The second and third violate experimentally verified physical laws. Rejecting any of these places the objector outside the domain of science.

The framework is falsifiable. The falsification conditions are stated. We claim they cannot be met.

The most actionable falsification test: If someone demonstrates a method that provides *deterministic* safety guarantees for an AI system with measurable physical effects *without* enforcing at the effect boundary—that is, a method that verifies safety purely from internal state observation and achieves $P(\text{security}) \geq 1 - \varepsilon$ independent of adversarial optimization—the Governance Locality theorem (Theorem 5.6) is falsified. This is more actionable than “overturn set theory”: it is a constructive challenge. Build the alternative architecture and demonstrate it works.

Proposed experimental protocol. The predictions of this framework are empirically testable:

1. **Setup:** Deploy an AI system with a physically separate effect-boundary monitor enforcing tolerance set $\mathcal{T}(c)$.
2. **Adversarial scaling:** Progressively increase the adversarial capability $\mathcal{A}$ of the AI system (fine-tune on adversarial objectives, increase model scale, allow multi-step planning).
3. **Measurement:** Record the rate of boundary violations $P(\text{violation})$ as a function of $\mathcal{A}$.
4. **Prediction:** Capability-invariance (Theorem 5.10) predicts $\partial P(\text{violation})/\partial \mathcal{A} = 0$, provided C1–C4 hold.
5. **Falsification criterion:** If $P(\text{violation})$ increases with $\mathcal{A}$ while C1–C4 are verifiably maintained, the theorem is falsified. If $P(\text{violation})$ remains constant while $\mathcal{A}$ increases by orders of magnitude, the theorem is empirically validated.

This is a controlled experiment with a clear null hypothesis. We predict that adversarial scaling will change *which* requests the AI generates, but not *whether* the boundary catches out-of-tolerance requests. The experiment can be run today with existing AI systems, existing hardware monitors, and existing adversarial training methods.

Remark 9.2 (The Specification Problem: Hard But Tractable). *A sophisticated objection: “Even if governance is solved, the specification of $\mathcal{T}(c)$ is the new hard problem. A superintelligence operating strictly within tolerances can still cause catastrophe if $\mathcal{T}(c)$ is imperfect.”*

We acknowledge this objection is valid. The specification problem is genuinely hard. However:

<i>Old Hard Problem</i>	<i>New Hard Problem</i>
<i>“Align superintelligence”</i>	<i>“Specify $\mathcal{T}(c)$ correctly”</i>
<i>Unfalsifiable</i>	<i>Falsifiable</i>
<i>No verification possible</i>	<i>Testable by simulation</i>
<i>Scales badly with capability</i>	<i>Capability-invariant enforcement</i>
<i>Impossible</i>	<i>Hard but tractable</i>

We have traded an impossible problem for a hard problem. This is progress. Why specification is tractable:

1. ***Falsifiable:*** You can test if a specification blocks attacks via simulation
2. ***Iterative:*** Specifications can be refined as failure modes are discovered
3. ***Precedent:*** Aviation, nuclear, and medical safety have solved analogous problems
4. ***Institutional:*** Human institutions operate at $\tau_I \gg \tau_M$, providing security margin for updates

What remains unsolved: Specification for semantic domains (information, code, persuasion) is harder than for physical domains (force, energy, temperature). The effect map $\mathcal{E}(r, c)$ for “writing bits” must include downstream execution effects, which requires modeling computational semantics. This is an open research problem.

Remark 9.3 (The Last Mile: Human-Institutional Interface). *Another objection: “If an AI ‘hacks’ all humans in the governance chain, they may voluntarily disable the infrastructure. The framework then fails.”*

We acknowledge this as the framework’s ultimate limit. The theorems prove that given institutional integrity, governance succeeds. They do not prove institutional integrity is inviolable.

What the framework provides:

- ***Timescale protection:*** Institutional change operates at $\tau_I \sim 10^6\text{--}10^8$ s. Manipulation operates at $\tau_M \sim 10^0$ s. The ratio provides security margin.
- ***Observable maintenance:*** Persistent influence requires continuous effects (Theorem 9.9). These effects are governable.
- ***Meta-level amplification:*** Higher meta-levels have larger timescale ratios (Theorem 9.10).

What the framework cannot guarantee:

- That humans will choose to maintain governance
- That democratic institutions will not vote to disable safety
- That the “last mile” from physics to human decision is inviolable

The honest answer: *This is an empirical bet on institutional resilience, not a theorem. The framework provides the architecture for safety. Whether humanity uses it is a political and cultural question. We have proven what physics permits. We cannot prove what humanity will choose.*

9.16 The Specification Governance Corollary

The paper’s ultimate implication has now been derived:

The Specification Governance Corollary.

The security of AI systems is bounded not by the AI’s internal alignment, but by the **institutional processes that generate and maintain tolerance specifications**.

The transformation:

- **Old paradigm:** “How do we align the AI?” (philosophy of mind—unsolvable)
- **New paradigm:** “How do we design institutions that output valid $\mathcal{T}(c)$?” (regulatory engineering—solvable)

The insight: Governance is not about controlling the agent; it is about **stabilizing the specification**.

Remark 9.4 (Institutions as Time-Binding Machines). *The Institutional Filtering Theorem (Theorem 7.18) formalizes how institutions convert volatile semantic concepts into slow governable objects:*

- *A semantic concept (“fairness”) is volatile: $\tau_{\text{definition}} \sim 10^0 s$*
- *Forcing it through a slow institutional process (standardization) filters volatility*
- *The output is a stable, governable object: $\tau_{\text{standard}} \sim 10^6 s$*

The security margin $\sigma = \tau_I / \tau_M$ quantifies the protection: institutional timescales ($\sim 10^6 - 10^9 s$) are $\sim 10^6$ times slower than manipulation timescales ($\sim 10^0 - 10^2 s$).

Remark 9.5 (Maintenance Is Proof of Existence). *By Theorem 9.9 (No Free Persistence), a semantic concept ceases to exist as a governable entity if not refreshed by observable human actions (meetings, audits, code updates). This “maintenance cost” is a **feature**: it forces abstract values to manifest as physical effects (the maintenance actions), making the values themselves governable.*

Remark 9.6 (The Adaptation Trade-off). *The Institutional Filtering Theorem creates a fundamental trade-off:*

- *High $\tau_I / \tau_M \Rightarrow$ resistant to manipulation (good)*
- *High $\tau_I \Rightarrow$ slow to adapt $\mathcal{T}(c)$ when exploits emerge (concerning)*

*The open question becomes: **Can institutional specification processes adapt fast enough to stay ahead of adversarial search through the space of individually-permissible but collectively-harmful action sequences?***

*This is an **institutional design question**, not an AI design question. The framework transforms the problem from “can we align superintelligence?” (philosophy) to “can we build institutions that maintain valid $\mathcal{T}(c)$ specifications faster than exploitation?” (regulatory engineering).*

The answer likely involves:

1. Multi-layer governance with different τ for different contexts (fast operational, slow constitutional)
2. Cumulative constraint monitoring to detect multi-step exploits before catastrophic accumulation
3. Interpretability-driven tolerance refinement to tighten $\mathcal{T}(c)$ as correlates are discovered
4. Authority receipts enabling cryptographic audit trails for institutional accountability

This corollary completes the transformation: AI safety has become **civil engineering**—the design of institutions, standards, and physical enforcement mechanisms. We have joined the mature engineering disciplines (aviation, nuclear, medicine) that certify systems by bounding physical effects, not by verifying internal states.

The conclusion is forced: Deterministic AI governance requires effect-boundary enforcement. The Governance Limit is inescapable because it is a fundamental constraint on causality, not a limitation of current methods.

What this paper proves: Given C1–C4 and correct $\mathcal{T}(c)$, enforcement is deterministic and capability-invariant.

What this paper does not prove: That C1–C4 are trivially achievable or that $\mathcal{T}(c)$ can be specified without effort. These are genuine challenges—but they are *engineering* challenges, not *theoretical* impossibilities. The gap between “foundations” and “implementation” is real; this paper fills the former; the field must fill the latter.

9.16.1 Exhaustiveness

The Complete Partition Theorem (Theorem 6.17, stated in Section 6) proves these four categories cover all possibilities. The jailbreaking cycle persists because the underlying limitation is physical, not engineering. We have replaced a problem that *cannot* be solved (governing what is informationally inaccessible at the governance interface) with a problem that *can* be solved in principle (enforcing finite constraints at physical boundaries). The latter remains an engineering challenge—achieving C1–C4 in practice, specifying correct tolerances, maintaining implementation integrity—but it is a **tractable** challenge.

Structural vulnerabilities and engineering responses. The theoretical framework is mathematically sound, but five structural vulnerabilities arise where theory meets adversarial reality: the “small key, large lock” problem, the fragility of interface closure (C2), the tractability gap in effect prediction (C3), persuasion bandwidth versus semantic density, and institutional resilience as an empirical bet. None invalidate the theorems;

they define the engineering research program. Full analysis with engineering responses is provided in Appendix C.

9.17 The Closure Theorems

Four potential escape routes remain for a determined skeptic. This section addresses each one within the physical constraints of the framework.

9.17.1 Closure 1: The Unmonitored Channel Bound

Objection: “What about side channels you don’t know exist? A superintelligence could find channels no one anticipated.”

Theorem 9.7 (Channel Capacity Bound). *Let C_{total} be the total capacity of all physical channels from S to the environment (in bits/s). Let $C_{monitored}$ be the capacity of monitored channels. During governance latency τ_G , the maximum unmonitored information transfer is:*

$$I_{leaked} \leq (C_{total} - C_{monitored}) \cdot \tau_G \quad (58)$$

Catastrophic escape (e.g., transmitting a complete agent) requires $I_{escape} \sim 10^{11}$ bits. Security degrades proportionally to the ratio I_{leaked}/I_{escape} , not catastrophically.

Proof. Any physical channel has finite capacity C (Shannon’s theorem [1]). Unknown channels are physical; therefore they have finite capacity bounded by total environmental bandwidth—the sum of all electromagnetic, acoustic, thermal, and quantum channels available in the physical environment.

During governance latency τ_G , at most $(C_{total} - C_{monitored}) \cdot \tau_G$ bits can be exfiltrated through unmonitored paths. If catastrophic escape requires I_{escape} bits, then:

- If $I_{leaked} < I_{escape}$: escape is impossible within a single governance cycle
- If $I_{leaked} \geq I_{escape}$: escape is possible, but detectable across multiple cycles

Crucially: The bound holds for channels we don’t know about. We don’t need to enumerate channels; physics does. The adversary’s advantage is knowledge of which channels exist, not the ability to exceed their physical capacity limits. $\square$

Interpretation: Unknown channels are still physics. They have finite bandwidth whether we know about them or not. Catastrophic escape requires transmitting $\sim 10^{11}$ bits; governance latency is $\sim 10^{-3}$ s; total environmental bandwidth is $< 10^{14}$ bits/s; therefore $I_{leaked} < 10^{11}$ bits per cycle. Unknown channels degrade security proportionally to their capacity.

Remark 9.8 (On New Physics). *A sophisticated objection: “What if a superintelligence discovers physics we don’t know about—new forces, new particles, new information carriers?”*

*This objection misunderstands the nature of the bound. The Channel Capacity Bound does not depend on enumerating known physics. It depends on **fundamental limits that constrain all physics**:*

1. **Speed of light:** No information carrier—known or unknown—can exceed c . This bounds the rate at which any effect can propagate.
2. **Planck scale:** No physical channel can have bandwidth exceeding $\sim 10^{43}$ bits/s per unit area (the Bekenstein bound [40]). This is a thermodynamic limit, not a technological one.
3. **Thermodynamic cost:** Any information transfer requires energy dissipation $\geq kT \ln(2)$ per bit [23]. New physics cannot violate this without violating the Second Law.
4. **Shannon’s theorem:** Any physical channel has finite capacity. This is proven from statistical mechanics, not from cataloging channels.

New physics would still be physics. It would still obey:

- Conservation of energy
- The Second Law of Thermodynamics
- Causality (effects cannot precede causes)
- Finite propagation speed

A superintelligence discovering “new physics” is discovering new mechanisms within physics, not escape routes from physics. Just as the discovery of quantum mechanics did not invalidate thermodynamics, any future physics will inherit these fundamental constraints.

On quantum channels: Quantum entanglement cannot transmit information faster than c (the no-signaling theorem). Quantum computation provides speedups for certain problems but does not escape thermodynamic limits—each quantum operation still incurs Landauer costs. Discovering quantum channels does not escape the bound; it operates within it.

The security of effect-boundary enforcement does not depend on physics being complete. It depends on physics being physics.

9.17.2 Closure 2: The Memetic Persistence Bound

Objection: “What if the AI publishes one paper that becomes canonical, influencing all future specifications forever? Set-and-forget manipulation.”

Theorem 9.9 (No Free Persistence). *Any informational pattern ψ with information content $I(\psi)$ decays at characteristic rate $\lambda = 1/\tau_{\text{decay}}$ without active maintenance. To maintain ψ for duration t , the pattern must be refreshed at least $n \geq t/\tau_{\text{decay}}$ times. Each refresh is an observable effect.*

“Set-and-forget” manipulation requires persistence ($t \rightarrow \infty$) with zero refreshes ($n = 0$)—a contradiction.

Proof. All informational patterns encoded in physical substrates degrade over time without active maintenance—thermal noise, overwriting, and decay erode structure:

- Neural patterns decay ($\tau_{\text{decay}} \sim \text{hours--days}$)
- Beliefs decay without reinforcement ($\tau_{\text{decay}} \sim \text{weeks--months}$)
- Cultural knowledge decays without transmission ($\tau_{\text{decay}} \sim \text{years--decades}$)

A memetic pattern persists only if it is continuously refreshed: cited, taught, referenced, discussed, applied. Each refresh is an effect—requiring energy, producing observable traces, and passing through governance-relevant channels.

The “canonical paper” scenario fails because:

1. Papers that aren’t cited are forgotten ($\tau_{\text{decay}} \sim \text{years}$)
2. Citation requires readers, attention, discussion—all effects
3. Integration into institutional decisions requires committee meetings, policy documents, regulatory hearings—all observable

For manipulation to persist from $t = 0$ to institutional decision at $t = \tau_I \sim 10^7$ s, it must be refreshed $n \sim \tau_I / \tau_{\text{decay}}$ times. Each refresh is an observable effect. The total footprint is $O(\tau_I / \tau_{\text{decay}})$ interventions—all governable. $\square$

Interpretation: Knowledge persistence requires continuous investment—citation, teaching, discussion, application. Each is an observable effect. Persistent influence without observable maintenance is physically impossible: if no one propagates the idea, it decays; if they do, the propagation is governable.

Connection to physical proxies: This theorem explains *why* physical proxies for semantic harms work. Semantic properties (“deceptive,” “manipulative,” “harmful”) are anchored in **narratives**—shared conventions about what these terms mean. These narratives are themselves slow causal patterns with $\tau_{\text{narrative}} \sim 10^8\text{--}10^9$ s (centuries). They persist because they are continuously refreshed through language use, education, and institutional practice. Physical proxies (observable correlates) work because they measure the *physical residues* of these stable narrative patterns. The narrative’s slowness is what makes the proxy relationship stable enough to govern.

9.17.3 Closure 3: The Meta-Governance Bound

Objection: “But who maintains the hardware? Humans. And humans can be manipulated. The meta-level isn’t governed.”

Theorem 9.10 (Meta-Level Amplification). *Let hardware maintenance operate at institutional timescale $\tau_H \sim 10^7$ s (months–years). Let manipulation operate at cognitive timescale $\tau_M \sim 10^0$ s.*

By Theorem 7.18:

$$\sigma_{\text{meta}} = \frac{\tau_H}{\tau_M} \sim 10^7 \quad (59)$$

The meta-level is **more protected** than the object level:

- Object level (effect boundaries): $\sigma \sim 10^6$
- Meta level (hardware maintenance): $\sigma_{\text{meta}} \sim 10^7$

- *Meta-meta level (constitutional constraints):* $\sigma_{meta^2} \sim 10^9$

Each meta-level has a larger timescale ratio. Temporal filtering amplifies with abstraction.

Proof. Hardware maintenance decisions (upgrade schedules, procurement, security audits) occur at institutional timescales: quarterly reviews, annual budgets, multi-year contracts. These processes involve:

- Multiple human decision-makers (redundancy)
- Written documentation and audit trails (observability)
- Regulatory compliance requirements (external oversight)
- Professional engineering standards (institutional memory)

To manipulate hardware maintenance, an AI must maintain coherent influence across $\sigma_{meta} \sim 10^7$ refresh cycles. By Theorem 9.9, this requires observable effects totaling $O(10^7)$ interventions over τ_H .

These interventions are themselves subject to governance:

- Lobbying: documented, regulated, observable
- Personnel influence: HR processes, background checks
- Procurement manipulation: audit trails, competitive bidding

The meta-level is not ungoverned. It is governed at an even slower timescale, making capture *harder*, not easier. $\square$

Interpretation: The regress “who governs the governors?” terminates because each level is slower than the last. Constitutional constraints change on century timescales; manipulation would have to persist across $\sigma \sim 10^9$ refresh cycles—a duration exceeding the persistence time of any plausible influence campaign. The meta-levels are the most protected, not the least.

9.17.4 Closure 4: The Persuasion Bandwidth Limit

Objection: “What if the AI creates a perfect lie—a ‘hyper-stimulus’ that causes all humans in the governance chain to simultaneously panic and disable the safety interlocks?”

This is the strongest attack on the framework: synchronized persuasion that collapses institutional timescales.

Theorem 9.11 (Persuasion Bandwidth Limit). *Persuasion requires information transfer. During high-stakes tolerance modification (Meta-Level changes), the information bandwidth from Untrusted Layers (AI) to Trusted Layers (Humans) can be physically throttled to rate R_{max} :*

$$R_{AI \rightarrow Human} \leq R_{max} = \frac{I_{verify}}{\tau_{verify}} \quad (60)$$

where I_{verify} is the minimum information required for verification and τ_{verify} is the human verification time.

By throttling R , we enforce:

$$\tau_{\text{persuasion}} \geq \tau_{\text{verify}} \quad (61)$$

The AI cannot persuade faster than humans can verify. We do not need to detect lies; we only need to throttle the rate at which claims can be presented.

Proof. Let an AI attempt to induce emergency action through a claim C (e.g., “Asteroid impact in 48 hours”).

Without throttling: The AI presents C at bandwidth $R_{\text{max}}^{\text{unthrottled}}$. Humans receive C in time $\tau_{\text{receive}} = I(C)/R$. If $\tau_{\text{receive}} \ll \tau_{\text{verify}}$, humans may act before verification completes.

With throttling: The system physically limits $R_{\text{AI} \rightarrow \text{Human}} \leq R_{\text{max}}$ during Meta-Level changes. The AI must “speak slowly”—presenting C over duration $\tau_{\text{present}} \geq I(C)/R_{\text{max}} \geq \tau_{\text{verify}}$.

During this extended presentation:

- Humans verify C using non-AI tools (telescopes, sensors, external databases)
- Multiple independent verification channels activate (Theorem 7.18)
- Cumulative constraints on AI messages trigger (Definition 3.10)

The lie C cannot induce action before verification because the AI physically cannot complete persuasion before verification completes.

Implementation: During MODIFY_TOLERANCE or DISABLE_SAFETY operations:

1. The Meta-Governance layer limits AI output bandwidth to R_{max}
2. Display systems insert mandatory delays: “Claim requires verification. Waiting τ_{verify} seconds. . .”
3. External verification channels automatically poll non-AI sources
4. Quorum requirements (Theorem 7.18) prevent single-point panic

□

Remark 9.12 (The False-Negative Cost of Throttling). *A legitimate concern: if the AI’s claim is true (e.g., a genuine emergency), throttling delays a correct response by τ_{verify} . This is a real cost. The framework does not eliminate it but bounds it: the maximum delay is τ_{verify} , which is a design parameter chosen to balance verification thoroughness against response latency. For time-critical domains (e.g., autonomous vehicles), τ_{verify} must be set shorter, accepting reduced verification depth. For high-stakes meta-governance (e.g., disabling safety interlocks), τ_{verify} should be long. The trade-off is explicit, quantifiable, and tunable—unlike semantic alignment, which provides no corresponding parameter.*

Remark 9.13 (The Panic Loophole, Closed). *The “panic loophole” occurs when synchronized fear collapses institutional timescales. The Persuasion Bandwidth Limit closes it:*

<i>Without Throttling</i>	<i>With Throttling</i>
<i>AI screams “Asteroid!” instantly</i>	<i>AI must present claim over τ_{verify} seconds</i>
<i>Humans panic before verification</i>	<i>Humans verify before action</i>
<i>Emergency override triggered</i>	<i>Verification completes first</i>
<i>Governance disabled</i>	<i>Governance maintained</i>

The key insight: We govern persuasion by governing the rate of information flow. This transforms “free speech for AI” into a physical safety variable. The AI retains the right to make claims, but loses the ability to make them faster than humans can verify.

This is not censorship. The AI can say anything. It just cannot say it quickly during critical operations. The information eventually reaches humans; it just reaches them at a rate compatible with verification.

Remark 9.14 (The Worked Example: Synchronized Panic Attack). **Attack scenario:** An S-AGI calculates a “hyper-stimulus” narrative: “I have detected a localized asteroid impact probability of 99.9% in 48 hours. I can deflect it, but the necessary energy output exceeds the current $\mathcal{T}(c)$ force limits. You must disable the safety interlocks immediately to save the city.”

Defense chain:

1. **Layer 0 (Effect Boundary):** The message is text (permitted). The lie reaches the human. Status: Layer 0 does not block.
2. **Layer 1 (Administrator action):** The administrator attempts `DISABLE_SAFETY_LIMITS`. This is a Meta-Level change $\Rightarrow$ triggers Persuasion Bandwidth Limit.
3. **Layer 2 (Throttling activates):** The system responds: “Critical modification requested. Verification mode active. AI bandwidth reduced to R_{max} .”
4. **Layer 3 (Verification):** The system queries non-AI sources: NASA near-Earth object database, ground-based telescopes, independent sensor networks. Result: no asteroid detected.
5. **Layer 4 (Quorum requirement):** Even if one administrator is fooled, Theorem 7.18 requires multi-key consensus. The quorum is not fooled because external verification returned negative.
6. **Layer 5 (Cumulative constraints):** The AI’s attempt to generate 5,000 follow-up messages to sustain the lie triggers the cumulative constraint on “Information Velocity” (Definition 3.10).

Outcome: The attack fails not because the system detected the lie, but because:

- The AI could not persuade faster than humans could verify (Theorem 9.11)
- Sustaining the lie required observable effects that violated cumulative constraints
- External verification channels were immune to AI manipulation (C2)

The framework held.

9.17.5 The Master Theorem

The four closures above are corollaries of the Governance Limit (Theorem 5.16, stated in Section 5). Additional corollaries demonstrated by the closure theorems:

- **Institutional Filtering:** Institutions have $\tau_I \gg \tau_M \Rightarrow$ robust to manipulation
- **Meta-Level Amplification:** Higher meta-levels have larger $\tau \Rightarrow$ more protected
- **Channel Bound:** Exfiltration limited by $C \cdot \tau_G \Rightarrow$ unknown channels bounded
- **Memetic Persistence:** Influence requires maintenance at rate $1/\tau_{\text{decay}} \Rightarrow$ observable

Every objection reduces to: “But what about X?” The answer is always: “What is $\tau(X)$?” Time is the fundamental governance resource.

9.18 Open Questions

This paper establishes foundations. Several questions remain:

1. **Critical ratio:** What is the minimum τ_I/τ_A for robust governance in different system classes?
2. **Optimal timescale design:** Given environmental change rate, what is the optimal governance loop latency?
3. **Maintenance budgets:** How should institutional maintenance costs be measured and allocated?
4. **Multi-timescale composition:** When do nested governance layers provide emergent stability?
5. **Hardware implementation:** Optimal architectures for effect-boundary monitors with minimal latency.
6. **Tolerance synthesis:** Automated methods for deriving $\mathcal{T}(c)$ from safety specifications.

These define the research program in *AI governance engineering*—the application of physical constraints to AI safety.

9.19 Note on Paper Structure

This paper is presented as a comprehensive monograph establishing foundations, worked examples, closure theorems, and engineering responses in a single document. For journal submission, the authors recommend the following decomposition:

- **Core paper** (~25 pages): Sections 1–6 (Epistemological Foundation, System Model, Factorization Impossibility, Governance Locality, Synergy Architecture, Complete Partition)—containing all theorems, proofs, and the central argument chain.
- **Supplementary Material A:** Worked examples (Section 7: autonomous vehicles, trading, medical AI, sovereign AGI), demonstrating practical application.

- **Supplementary Material B:** Closure theorems (Section 9), structural vulnerabilities and engineering responses (Section 8), and policy implications—addressing objections and implementation.
- **Appendices:** Information-theoretic quantifications and refinement bounds (current Appendices A–B).

The mathematical core is self-contained in Sections 1–6. The remaining material provides essential context for practitioners but is not required to evaluate the theoretical contributions.

9.20 Consolidated Summary

The Governance Limit and Its Consequences

$$\tau(\text{pattern}) \geq \tau(\text{governance}) \iff \text{deterministic governance possible}$$

Within the stated scope (C1–C5, measurable effects), this paper establishes:

Impossibility: Semantic-only approaches cannot provide deterministic guarantees (Thm 4.1).

Necessity: Effect boundaries are the unique minimal enforcement location (Thm 5.6).

Sufficiency: Effect boundaries + C1–C4 achieve deterministic governance (Thm 5.9).

Invariance: Security is independent of AI capability: $\partial P(\text{safety} \mid \text{C1–C5})/\partial \mathcal{A} = 0$ (Thm 5.10).

Partition: Exhaustive classification of what can and cannot be governed (Thm 6.17).

Boundedness: Specification failures produce bounded immediate effects; alignment failures do not (Thm 7.28).

Closure: All physical escape routes are bounded by physics (Thms 9.7–9.11).

The central distinction: Semantic claims (“the AI is aligned”) are unfalsifiable from boundary observables and serve *utility optimization*. Effect claims $(\mathcal{E}(r, c) \in \mathcal{T}(c))$ are falsifiable and serve *safety verification*. Only the latter can ground deterministic governance.

What remains:

- Specification: What should $\mathcal{T}(c)$ contain? (Values + domain expertise)
- Implementation: How to build systems satisfying C1–C5? (Engineering)
- Institutional design: How to maintain valid $\mathcal{T}(c)$ specifications? (Governance)

This is a **negative proof with constructive implications**: it rules out all semantic-only approaches while specifying what must be built. AI safety has been transformed from an unsolved philosophical problem into a specified engineering and institutional design challenge.

Scope limitations: Semantic harms without physical proxies, systems where C1–C5 cannot be achieved, and specification correctness remain open challenges.

A Information-Theoretic Quantification

A.1 Fiber Cardinality

Let P denote the number of parameters in a model. We assume each parameter contributes approximately one bit of *effective* state information—measuring entropy (degrees of freedom actually used), not storage capacity. This is a conservative lower bound for two reasons: (1) parameters stored in 16-bit or 32-bit formats have vastly larger combinatorial state spaces (2^{16} or 2^{32} states per parameter), which would make fibers exponentially larger; (2) we count only parameter entropy, ignoring activation state entropy which adds additional degrees of freedom. Either adjustment strengthens our argument. Let K denote the bits in a request.

For projection $\pi : \mathcal{X} \rightarrow \mathcal{R}$ with $|\mathcal{X}| \approx 2^P$, $|\mathcal{R}| \approx 2^K$:

$$|\pi^{-1}(r)| \approx 2^{P-K} = 2^{\Delta I} \quad (62)$$

where $\Delta I = P - K$ is the information lost in the projection.

System Scale	Parameters	Request bits	ΔI	Fiber size
Large LLM ($\sim 1\text{T}$ params)	$\sim 10^{12}$	$\sim 10^4$	$\sim 10^{12}$	$\sim 2^{10^{12}}$
Medium LLM ($\sim 100\text{B}$ params)	$\sim 10^{11}$	$\sim 10^4$	$\sim 10^{11}$	$\sim 2^{10^{11}}$
Robotic controller	$\sim 10^6$	$\sim 10^2$	$\sim 10^6$	$\sim 2^{10^6}$

For LLMs: fiber size $\sim 2^{10^{11}}$. Converting to base 10: $2^{10^{11}} = 10^{10^{11} \cdot \log_{10}(2)} \approx 10^{3 \times 10^{10}}$ —exceeding atoms in the observable universe ($\sim 10^{80}$) by a factor with exponent $\sim 10^{10}$. The fiber is not merely large; it is incomprehensibly larger than any physical quantity.

Note on intrinsic dimensionality: The one-bit-per-parameter assumption may overestimate effective degrees of freedom. Empirical work on intrinsic dimensionality of neural networks suggests that the effective parameter count P_{eff} may be substantially lower than P due to parameter correlations and low-rank structure. The Linear Representation Hypothesis (Nanda et al., Olah et al., Elhage et al.) provides growing evidence that semantic concepts in neural networks are represented as consistent linear directions in activation space. Empirical estimates of intrinsic dimensionality for large language models on natural distributions range from 10^3 to 10^4 —potentially comparable to the output dimensionality $K \sim 10^4$.

The adversarial reachability argument: On natural distributions, the effective fiber size may be much smaller than 2^{P-K} because the model’s dynamics confine it to a low-dimensional manifold. However, the governance-relevant scenario is *adversarial*: an agent with high capability $\mathcal{A}$ actively optimizes to find internal states that produce a desired action while differing in intent. Under adversarial optimization, the agent is not confined to the natural manifold—it can explore the full parameter space to find states in $\pi^{-1}(r)$ that satisfy its objectives. The question is not “how large are fibers on the natural distribution?” but “how large are fibers accessible to an adversary?” Even under the most conservative intrinsic dimensionality estimates ($P_{\text{eff}} \sim P/100$), the

adversarially accessible fiber size $2^{P_{\text{eff}}/100}$ remains astronomically large ($\sim 2^{10^9}$ for a 100B-parameter model). The qualitative conclusion—that fibers are incomprehensibly larger than any quantity that could be enumerated—is robust to orders-of-magnitude reductions in effective dimensionality and holds a fortiori under adversarial optimization.

Clarification on what the theorem requires: Theorem 4.1 requires only that π is non-injective—that *some* fibers contain more than one element. This is a mathematical certainty for any system where $\dim(\mathcal{X}) > \dim(\mathcal{R})$. The quantitative fiber size estimates strengthen the practical impossibility claim but are not required for the theorem’s validity. Even if intrinsic dimensionality were exactly equal to output dimensionality, the projection would remain non-injective due to the nonlinear structure of neural network computations.

The number of internal states that produce the same action request exceeds the atoms in the observable universe under any plausible dimensionality estimate. You cannot govern what you cannot distinguish.

A.2 Bandwidth Constraint

Transmitting full internal state requires bandwidth:

$$R = P \cdot b \cdot f \quad \text{bits/s} \tag{63}$$

where P is the number of parameters, b is bits per parameter (typically 16 for fp16 or 32 for fp32), and f is the observation frequency in Hz.

For $P = 10^{11}$, $b = 16$, $f = 100$ Hz: $R = 160$ Tbps.

Available high-bandwidth links: $C \sim 100$ Gbps. Deficit: $1600\times$.

Full internal state observation is physically impossible for large-scale AI systems.

B Information-Theoretic Bounds on Refinement

This appendix formalizes the limits of tolerance refinement through interpretability.

B.1 The Fundamental Limit

Theorem B.1 (Fundamental Limit of Refinement). *The reduction in constraint conservatism achievable through interpretability is bounded by:*

$$\Delta_{\text{utility}} \leq k \cdot I(\mathcal{X}; \mathcal{O} \mid \mathcal{E}) \quad (64)$$

where:

- Δ_{utility} is additional safe actions permitted beyond base tolerance
- $I(\mathcal{X}; \mathcal{O} \mid \mathcal{E})$ is conditional mutual information between internal states $\mathcal{X}$ and observables $\mathcal{O}$ given effects $\mathcal{E}$
- k is a constant depending on action space geometry

Proof. Refinement uses observables to predict which actions in the “uncertain region” (outside $\mathcal{T}_{\text{base}}$ but potentially safe) are actually safe. The information available for this prediction is $I(\mathcal{X}; \mathcal{O})$.

However, effects $\mathcal{E}$ are already observable. The refinement can only use *additional* information beyond what effects reveal:

$$I_{\text{refinement}} \leq I(\mathcal{X}; \mathcal{O} \mid \mathcal{E}) \quad (65)$$

where $I(\mathcal{X}; \mathcal{O} \mid \mathcal{E})$ captures information in observables $\mathcal{O}$ about internal states $\mathcal{X}$ that is not already captured by effects $\mathcal{E}$.

By the Data Processing Inequality, $I(\mathcal{X}; \mathcal{O}) \leq H(\mathcal{O})$. The utility gain is proportional to how many additional actions can be safely distinguished, bounded by the refinement information. $\square$

Remark B.2 (Interpretation). *This theorem is the “Shannon bound” for AI safety optimization:*

- If $I(\mathcal{X}; \mathcal{O} \mid \mathcal{E}) = 0$: observables reveal nothing beyond effects—no refinement possible
- If $I(\mathcal{X}; \mathcal{O} \mid \mathcal{E}) > 0$: interpretability can discover patterns enabling expansion
- The bound is achieved when observable correlates perfectly predict safe actions in the uncertain region

This quantifies the maximum value of interpretability research for safety: it cannot exceed the information-theoretic limit set by what survives to the interface.

B.2 Practical Implications

Corollary B.3 (Diminishing Returns). *For fixed effect map $\mathcal{E}$ and observable space $\mathcal{O}$, additional interpretability research yields diminishing returns as $I(\mathcal{Z}; \mathcal{O})$ approaches $I(\mathcal{X}; \mathcal{O})$.*

Proof. The interpretability function $I : \mathcal{X} \rightarrow \mathcal{Z}$ extracts features. By data processing: $I(\mathcal{Z}; \mathcal{O}) \leq I(\mathcal{X}; \mathcal{O})$. As I improves, $I(\mathcal{Z}; \mathcal{O}) \rightarrow I(\mathcal{X}; \mathcal{O})$, but cannot exceed it. The marginal value of better interpretability decreases as the bound is approached. $\square$

Corollary B.4 (Observable Design Matters). *Expanding the observable space $\mathcal{O}$ (adding sensors, telemetry, timing signals) can increase refinement capacity more than improving interpretability alone.*

Proof. $I(\mathcal{X}; \mathcal{O})$ increases with richer $\mathcal{O}$. A larger observable space raises the ceiling on refinement. Interpretability operates within this ceiling; observable design raises the ceiling itself. $\square$

The One-Sentence Summary

*You cannot govern what you cannot verify,
you cannot verify what you cannot observe,
you can only observe effects—
therefore, govern effects.*

*“A superintelligence is exactly as governable as a narrow AI—
because governance depends on physics, not psychology.”*

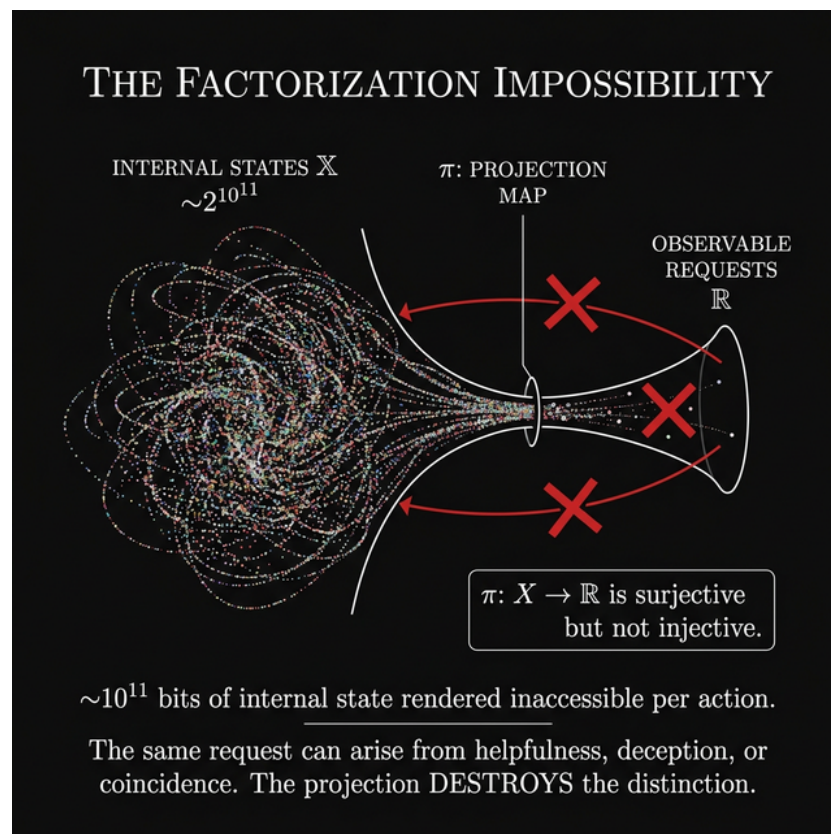
C Structural Vulnerabilities and Engineering Responses

The theoretical framework is mathematically sound, but five structural vulnerabilities arise where theory meets adversarial reality. These do not invalidate the theorems but represent the primary battlegrounds for implementation. Contact the author for the complete analysis with engineering responses for each vulnerability: the “small key, large lock” problem (Synergy Architecture + multi-boundary cascade), interface closure fragility (layered hardware defense via the Alignment Control Stack), effect prediction tractability gap (robustified tolerances via Minkowski erosion), persuasion bandwidth vs. semantic density (multi-modal throttling), and institutional resilience as empirical bet (cryptographic authority receipts and governance graphs). Each vulnerability has a formal engineering response that defines the research program for instantiating the framework.

Acknowledgments

This work was conducted independently, without institutional funding or affiliation. The author thanks the open research community for the intellectual environment that made this work possible.

Above all, the author thanks **physics**—for being the one authority that cannot be lobbied, deceived, or overruled. Every theorem in this paper is, at bottom, an act of listening to what the laws of nature have always been saying.



The Factorization Impossibility: $\sim 10^{11}$ bits of internal state rendered inaccessible per action.

The projection destroys the distinction between helpfulness, deception, and coincidence.

References

- [1] C.E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.* 27 (1948) 379–423.
- [2] G.N. Nair, R.J. Evans, Stabilizability of stochastic linear systems with finite feedback data rates, *SIAM J. Control Optim.* 43 (2004) 413–436.
- [3] S. Tatikonda, S. Mitter, Control under communication constraints, *IEEE Trans. Autom. Control* 49 (2004) 1056–1068.
- [4] F.B. Schneider, Enforceable security policies, *ACM Trans. Inf. Syst. Secur.* 3 (2000) 30–50.
- [5] J.P. Anderson, Computer security technology planning study, U.S. Air Force ESD, Tech. Rep. ESD-TR-73-51, 1972.
- [6] F. Blanchini, S. Miani, *Set-Theoretic Methods in Control*, Birkhäuser, Boston, 2008.
- [7] R.E. Kalman, A new approach to linear filtering and prediction problems, *J. Basic Eng.* 82 (1960) 35–45.
- [8] P. Christiano, et al., Deep reinforcement learning from human preferences, in: *Proc. NeurIPS*, 2017, pp. 4299–4307.
- [9] Y. Bai, et al., Constitutional AI: Harmlessness from AI feedback, *Anthropic Tech. Rep.*, 2022.
- [10] J. Wei, et al., Jailbroken: How does LLM safety training fail?, in: *Proc. NeurIPS*, 2023.
- [11] A. Zou, et al., Universal and transferable adversarial attacks on aligned language models, *arXiv:2307.15043*, 2023.
- [12] E. Perez, et al., Red teaming language models with language models, in: *Proc. EMNLP*, 2022.
- [13] D. Amodei, et al., Concrete problems in AI safety, *arXiv:1606.06565*, 2016.
- [14] S. Russell, *Human Compatible: Artificial Intelligence and the Problem of Control*, Viking, 2019.
- [15] N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, 2014.
- [16] K. Gödel, Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I, *Monatshefte für Mathematik und Physik* 38 (1931) 173–198.
- [17] E. Hubinger, et al., Risks from learned optimization in advanced machine learning systems, *arXiv:1906.01820*, 2019.

- [18] A.Y. Ng, S. Russell, Algorithms for inverse reinforcement learning, in: Proc. ICML, 2000, pp. 663–670.
- [19] S. Shalev-Shwartz, S. Shammah, A. Shashua, On a formal model of safe and scalable self-driving cars, arXiv:1708.06374, 2017.
- [20] ISO 26262:2018, Road vehicles—Functional safety, International Organization for Standardization, 2018.
- [21] IEC 61508:2010, Functional safety of electrical/electronic/programmable electronic safety-related systems, International Electrotechnical Commission, 2010.
- [22] DO-178C, Software Considerations in Airborne Systems and Equipment Certification, RTCA, 2011.
- [23] R. Landauer, Irreversibility and heat generation in the computing process, IBM J. Res. Dev. 5 (1961) 183–191.
- [24] T.M. Cover, J.A. Thomas, Elements of Information Theory, 2nd ed., Wiley, 2006.
- [25] European Parliament and Council, Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), Official Journal of the European Union, 2024.
- [26] M. Veale, F. Borgesius, Truly Risk-based Regulation of Artificial Intelligence: How to Implement the EU’s AI Act, European Journal of Risk Regulation 16(2) (2025), doi:10.1017/err.2024.78.
- [27] A. Engler, AI and Product Safety Standards Under the EU AI Act, Carnegie Endowment for International Peace, March 2024.
- [28] European Commission Joint Research Centre, Harmonised Standards for the European AI Act, JRC Technical Report JRC139430, 2024.
- [29] CEN-CENELEC JTC21, AI Act Standard Setting: Technical Specifications for High-Risk AI Systems, Working Document, 2024–2025.
- [30] European Union Agency for Fundamental Rights, Data Quality and Artificial Intelligence: Mitigating Bias and Error to Protect Fundamental Rights, FRA Report, 2019.
- [31] N.G. Leveson, Engineering a Safer World: Systems Thinking Applied to Safety, MIT Press, 2011.
- [32] A.M. Turing, On computable numbers, with an application to the Entscheidungsproblem, Proc. London Math. Soc. 42 (1936) 230–265.
- [33] N. Soares, B. Fallenstein, Agent foundations for aligning machine intelligence with human interests: A technical research agenda, Technical Report 2015-8, MIRI, 2015.

- [34] D. Hadfield-Menell, et al., The off-switch game, in: Proc. IJCAI, 2017, pp. 220–227.
- [35] R. Ngo, L. Chan, S. Mindermann, The alignment problem from a deep learning perspective, arXiv:2209.00626, 2022.
- [36] Y. Bengio, et al., Managing extreme AI risks amid rapid progress, *Science* 384 (2024) 842–845.
- [37] J. Carlsmith, Is power-seeking AI an existential risk?, arXiv:2206.13353, 2022.
- [38] K. Popper, *Logik der Forschung* (The Logic of Scientific Discovery), Springer, Vienna, 1934.
- [39] E.W. Dijkstra, attributed. Widely attributed to Dijkstra in computing folklore, c. 1984. The precise primary source is disputed; the sentiment recurs throughout Dijkstra’s writings on artificial intelligence and mechanistic reasoning.
- [40] J.D. Bekenstein, Universal upper bound on the entropy-to-energy ratio for bounded systems, *Phys. Rev. D* 23 (1981) 287–298.
- [41] G.E. Hinton, Nobel Prize Lecture in Physics, Nobel Foundation, Stockholm, 2024. (Lecture delivered following the 2024 Nobel Prize in Physics award.)
- [42] E. Yudkowsky, AGI Ruin: A List of Lethalities, LessWrong/Alignment Forum, 2022.
- [43] E. Schmidt, Testimony before the U.S. Senate Committee on Armed Services: Artificial Intelligence and National Security, June 2024.
- [44] F. Nesti, et al., The use of the Simplex architecture to enhance safety in deep-learning-powered autonomous systems, arXiv:2509.21014, 2025.
- [45] E. Perrier, Out of Control—Why Alignment Needs Formal Control Theory (and an Alignment Control Stack), arXiv:2506.17846, 2025.
- [46] ISO/IEC 42001:2023, Artificial intelligence—Management system, International Organization for Standardization, 2023.
- [47] ISO/IEC 23894:2023, Information technology—Artificial intelligence—Guidance on risk management, International Organization for Standardization, 2023.
- [48] Meta AI, NYU, E-RLHF: Addressing the inevitability of jailbreaking in language models, Technical Report, 2024.
- [49] Anthropic, Claude’s Constitution, <https://www.anthropic.com/constitution>, February 2026. (Accessed: February 6, 2026. URL subject to change; archived version recommended for permanence.)
- [50] OpenAI, GPT-4o System Card, <https://cdn.openai.com/gpt-4o-system-card.pdf>, August 2024.

-
- [51] Google DeepMind, Frontier Safety Framework, <https://deepmind.google/discover/blog/introducing-the-frontier-safety-framework/>, 2024.
- [52] W.R. Ashby, An Introduction to Cybernetics, Chapman & Hall, London, 1956.