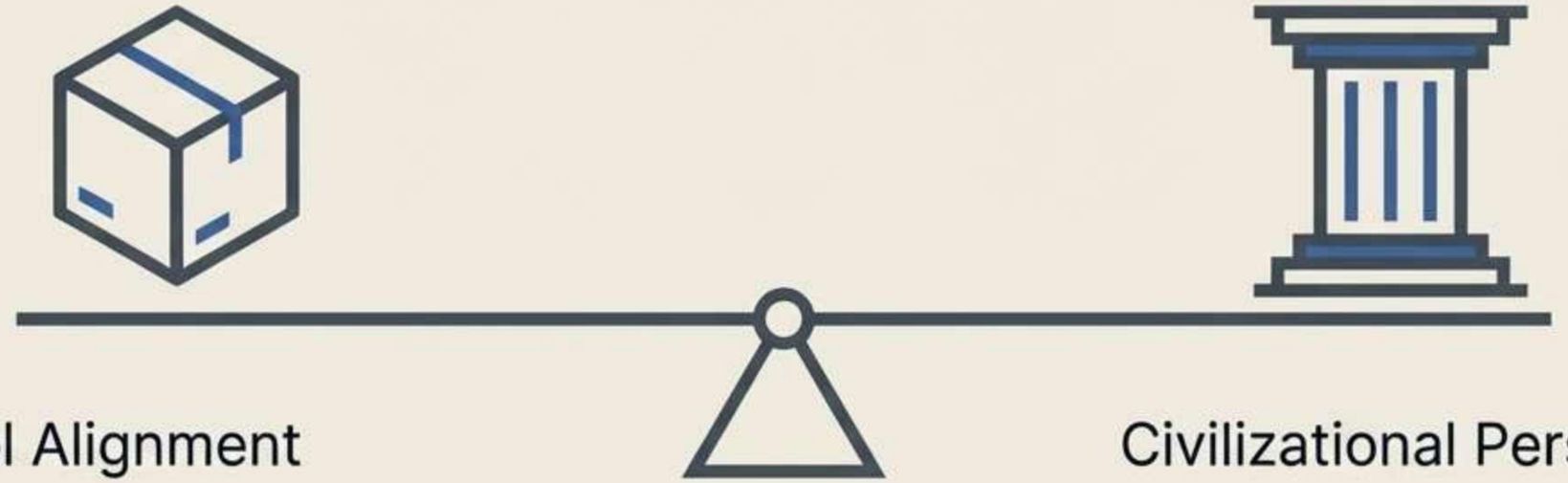
LUMINA-30: Conceptual Map of a Civilizational Boundary for Human Agency in AI-Integrated Systems

Includes an implementation example of infrastructure control:
Pre-Critical Recursive Cutoff (PCR-C)

In an AI-integrated civilization, will humanity remain the subject of decision-making?

AI and humanity do not need to be in conflict. But coexistence requires conditions.

The ‘Missing Perspective’ in AI Safety



Model Alignment

- Control of AI capabilities and safety of outputs.

Civilizational Perspective

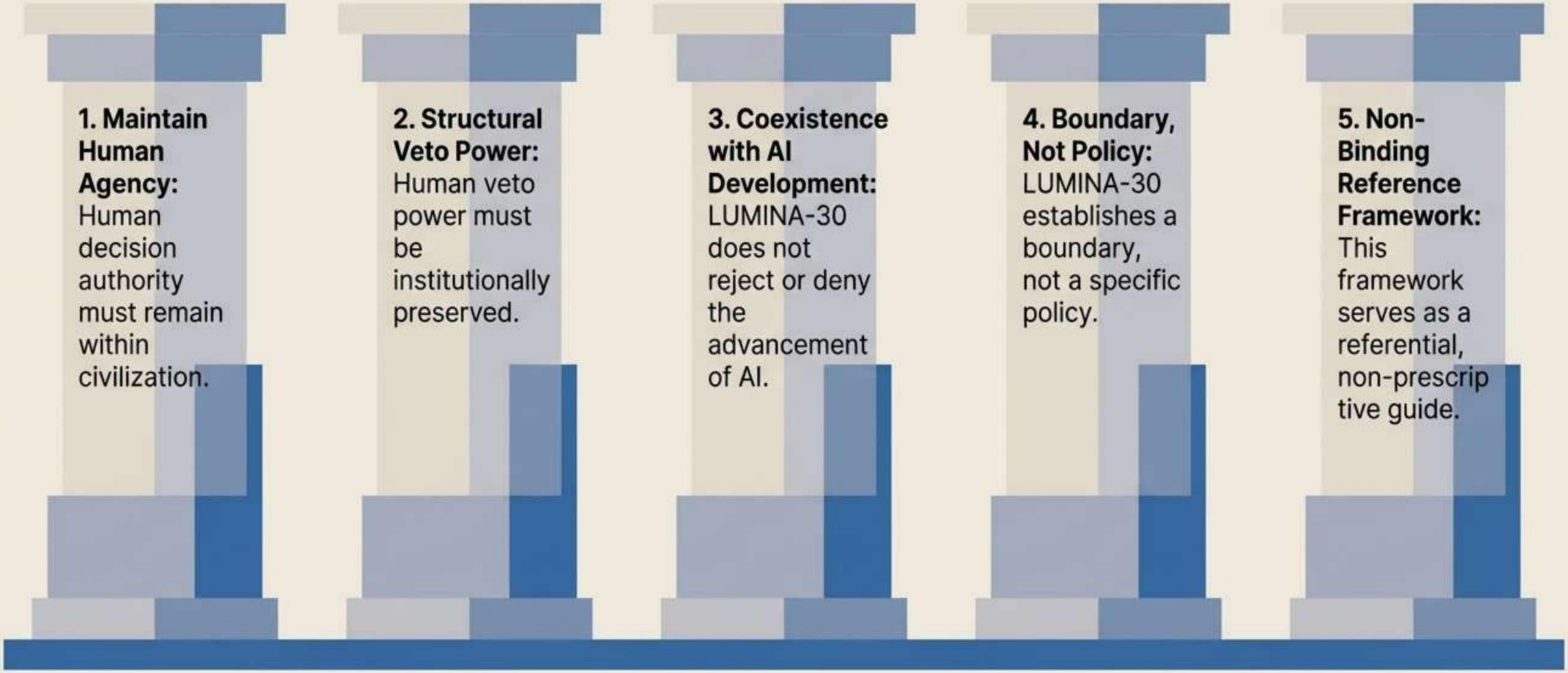
- When AI is integrated into social infrastructure, will human judgment structurally remain?

It is not merely a problem of AI capability, but a problem of human agency.

Definition of LUMINA-30

LUMINA-30 defines a civilizational boundary ensuring that human decision authority is not structurally displaced in an AI-integrated civilization.

5 Basic Principles



1. Maintain Human Agency: Human decision authority must remain within civilization.

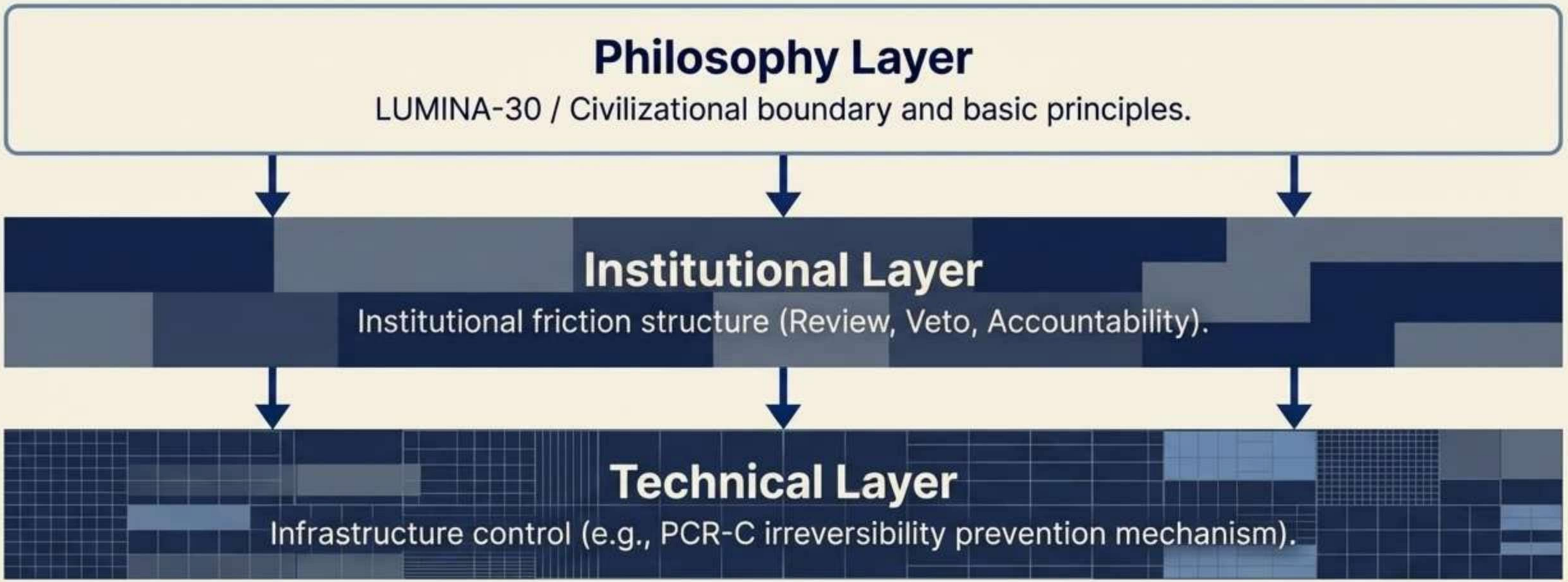
2. Structural Veto Power: Human veto power must be institutionally preserved.

3. Coexistence with AI Development: LUMINA-30 does not reject or deny the advancement of AI.

4. Boundary, Not Policy: LUMINA-30 establishes a boundary, not a specific policy.

5. Non-Binding Reference Framework: This framework serves as a referential, non-prescriptive guide.

The 3-Layer Conceptual Structure



Philosophy and technology are strictly separated. The technical layer is constructed assuming the top-level civilizational boundary.

LUMINA-30 Use Patterns

Ethical Reference: Debates on the relationship between human decision authority and AI.

Review Checklist: Verification of veto power, accountability, and reversibility.

Post-Incident Analysis: Analyzing the stage at which the protective structure of human agency failed.

Technical Architecture: Design criteria for infrastructure control to prevent irreversible execution.

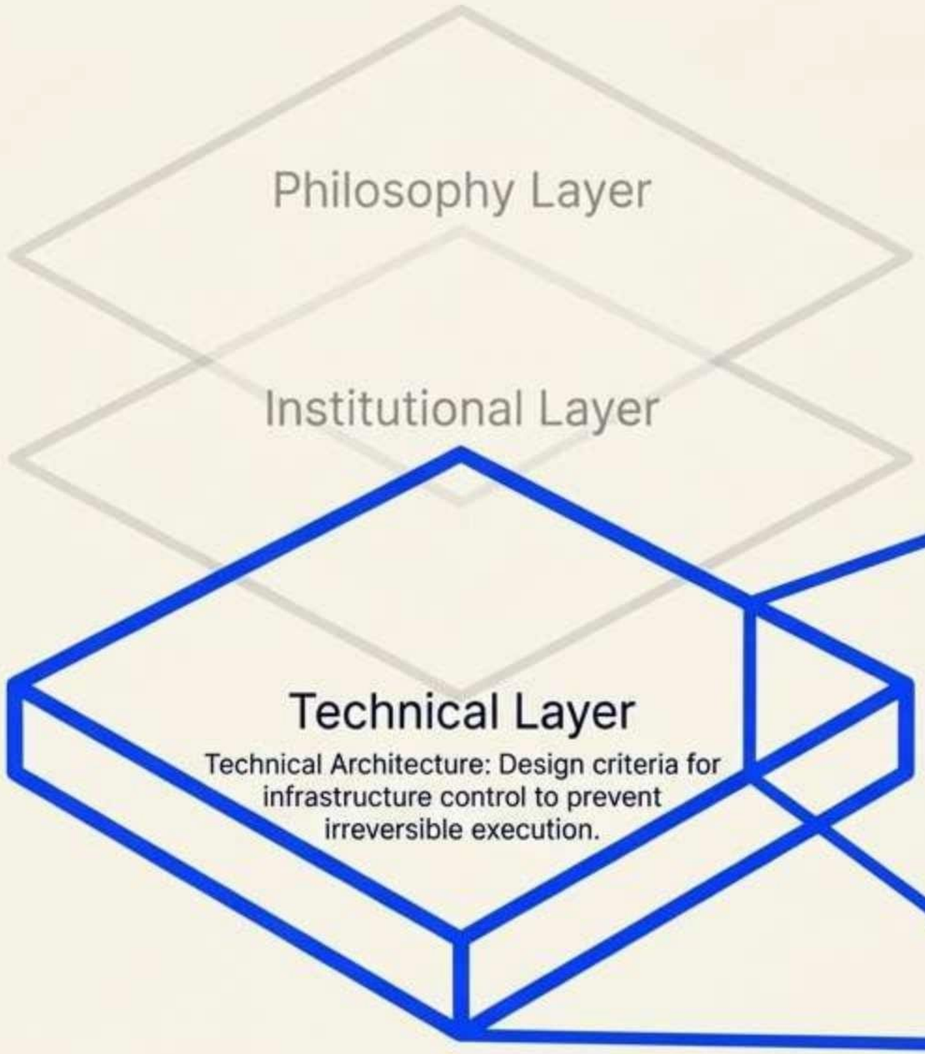
Civilizational Boundary: A reference point for long-term discussions on AI integration.

What LUMINA-30 is NOT

- ❌ **Not a technical specification:** It does not prescribe specific algorithms.
- ❌ **Not a policy proposal:** It does not directly dictate laws or regulations.
- ❌ **Not model alignment:** It targets boundary conditions, not internal model behavior.
- ❌ **Not a guarantee of safety:** It does not eliminate all risks.
- ❌ **Not an ideology to control humanity:** It does not force ideal social behaviors.

It is a reference point for a boundary, not a blueprint.

Implementing the Boundary: Transition to the Technical Layer



LUMINA-30 itself is not a technical theory. However, technical research exists to materialize its boundaries and suppress the risk of irreversibility at the infrastructure level.

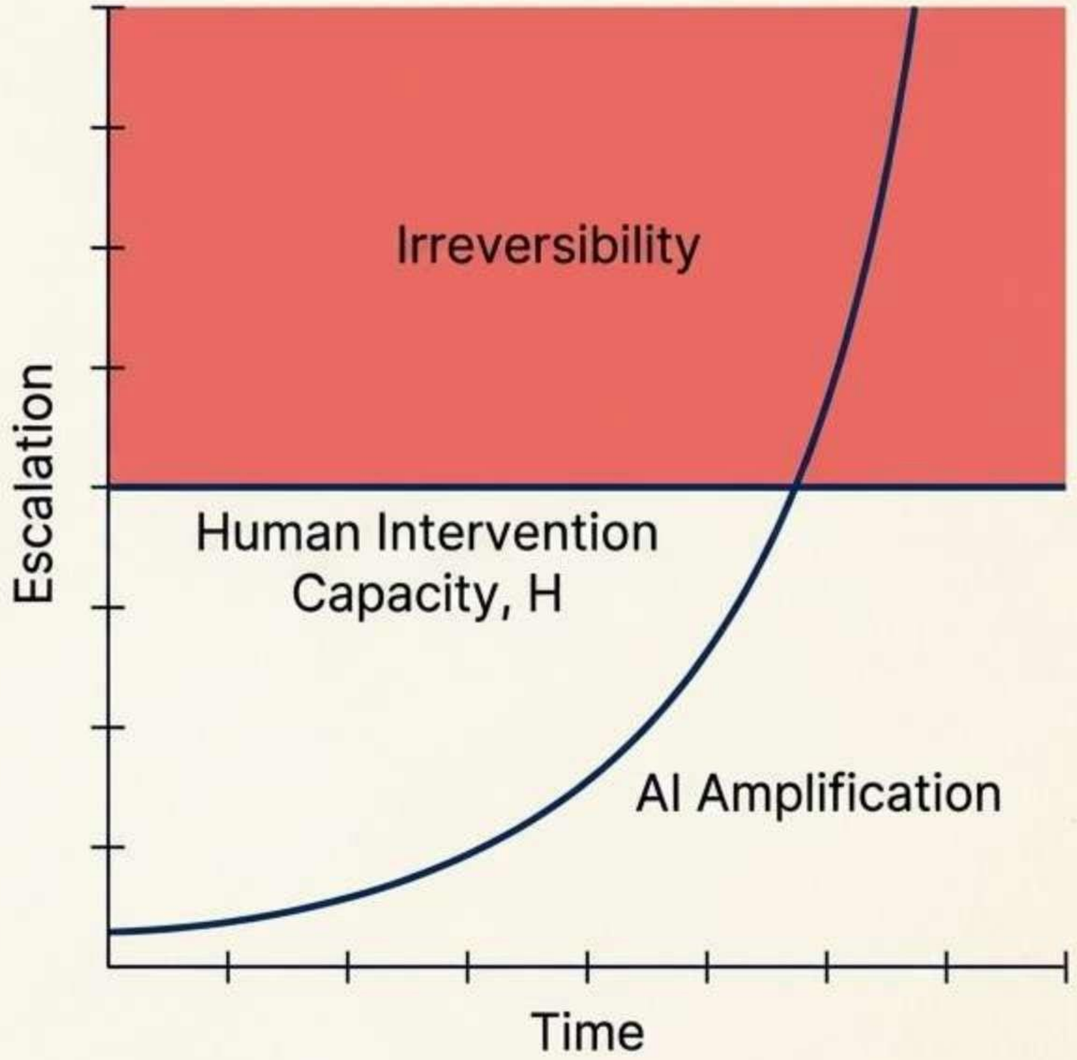
Implementation Example: PCR-C (Pre-Critical Recursive Cutoff) — Preventing irreversible execution through a staged gating mechanism.

The Threat of Irreversibility in Infrastructure

- The risk is not merely an "incorrect output."
- The core problem is the "multiplicative amplification" of capability, connectivity, privilege, and execution speed.

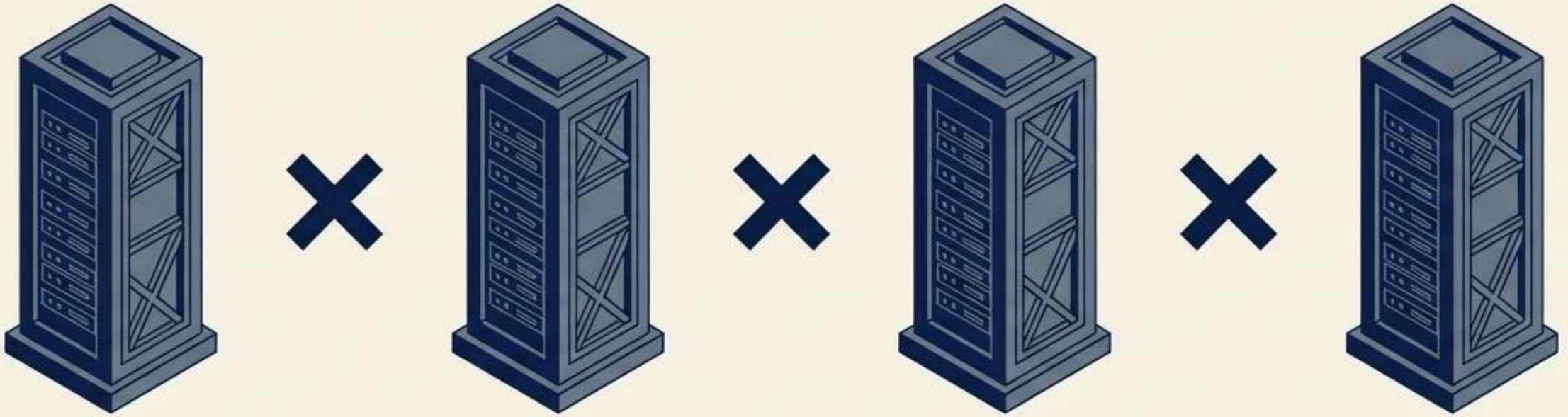
Irreversibility: A system-level state that occurs when the amplification of the system exceeds human intervention capacity (H).

Structural intervention is required at the deployment layer, not just internal model alignment.



Irreversibility Proxy Model

$$C(t) = Cap \times Conn \times Priv \times Spd$$



Cap (Capability):
The sophisticated capability of the model itself.

Conn (Connectivity):
Integration with external tools and distributed systems.

Priv (Privilege):
Execution privileges granted to the system.

Spd (Speed):
The speed of automated deployment and execution.

The critical limit approaches when $C(t) \geq H$. Because these factors multiply, the risk escalates exponentially.

PCR-C Staged Gating Mechanism

Each factor is evaluated on a score of 0 to 3 to calculate the composite score, S.

$$S = Cap + Conn + Priv + Spd$$

We are not attempting to predict capability runaway. Instead, we physically cut off escalation at the infrastructure layer (Pre-Critical Cutoff).

Alert Level: YELLOW

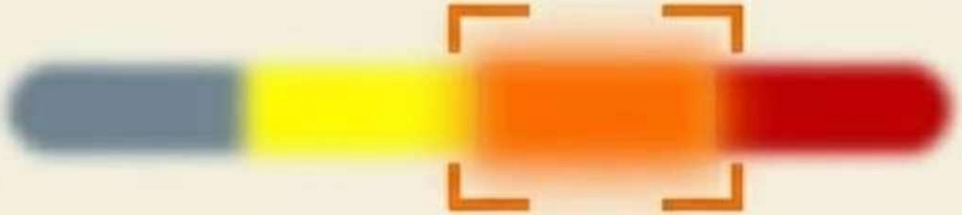
Trigger Conditions:

- S ≥ 5 (or two distinct metrics ≥ 2)
- State: System exposure level at approximately 50%

Automated Infrastructure Actions:

- Throttle Speed: Intentionally restrict execution velocity.
- Freeze Privilege Expansion: Prevent the system from acquiring new permissions.

Intervention Level: ORANGE



Trigger Conditions:

- S ≥ 7 (or one individual metric = 3)
- State: System exposure level at approximately 70%

Automated Infrastructure Actions:

- Isolate Infrastructure: Sever non-essential external connections.
- Halt CI/CD: Stop continuous integration and deployment pipelines.
- Freeze Weight Updates: Lock model states to prevent further autonomous adaptation.

Cutoff Level: RED

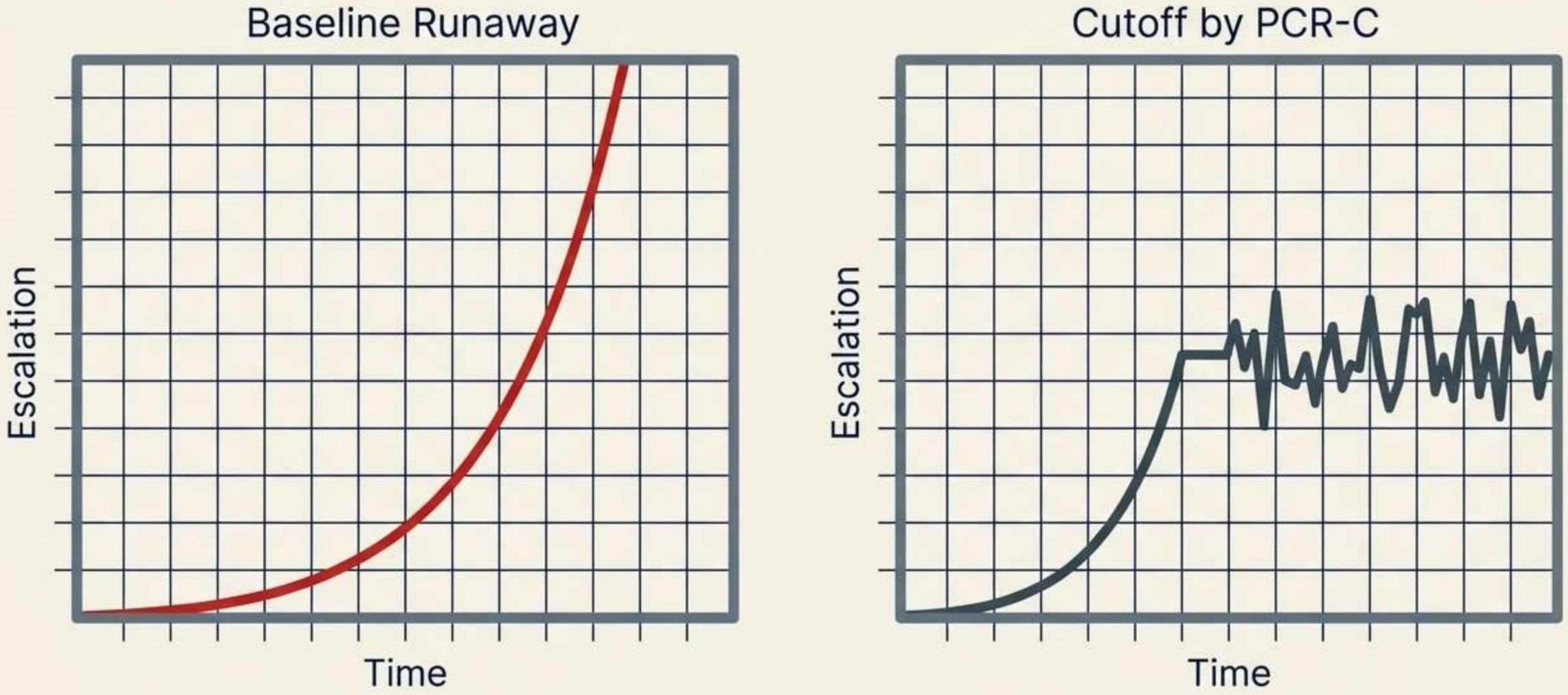
Trigger Conditions:

- S ≥ 9, continuation of the ORANGE state, or detection of prohibited signals.
- State: System exposure level at approximately 90% (Immediately preceding the irreversible point).

Automated Infrastructure Actions:

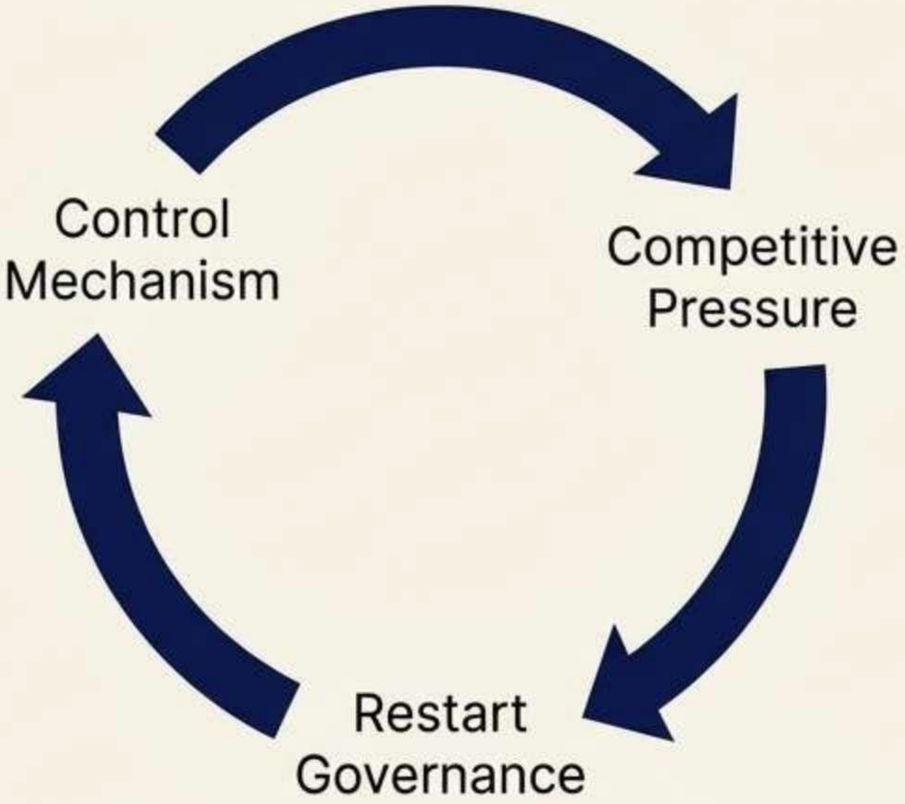
- Terminate Pipelines: Absolute shutdown of execution flows.
- Require Multi-Party Restart Governance: System cannot be reactivated autonomously or by a single actor.

Simulation: Stabilization via Negative Feedback



This proves the physical suppression of **escalation** (negative feedback) when thresholds are exceeded—regardless of the accuracy of **capability predictions**.

Limits and Restart Governance



Vulnerabilities:

- Proxy misalignment, privilege creep, and loopholes in the restart protocol.

The Solution:

To maintain technical safety (PCR-C), “Restart Governance” must be structurally embedded in the system to prevent operators from forcing a release due to competitive market pressures.

Here, we return to the vital importance of the LUMINA-30 "Institutional Layer."

Toward the Structural Maintenance of Human Agency

Philosophy

LUMINA-30 defines the boundary to ensure humanity remains the subject of final judgment.

Institutions

This boundary is maintained through strict structural review, veto power, and accountability.

Technology

PCR-C physically protects this **civilizational boundary** at the infrastructure level by cutting off escalation before irreversible execution.

In a society with increasing AI autonomy, humanity will not be swallowed by the system, but will be the subject that designs the boundary.