LUMINA-30: AI文明における人類主体境界の思想体系地図

Conceptual Map of a Civilizational Boundary for Human Agency in AI-Integrated Systems.

Pre-Critical Recursive Cutoff (PCR-C)
インフラ制御の実装例を含む

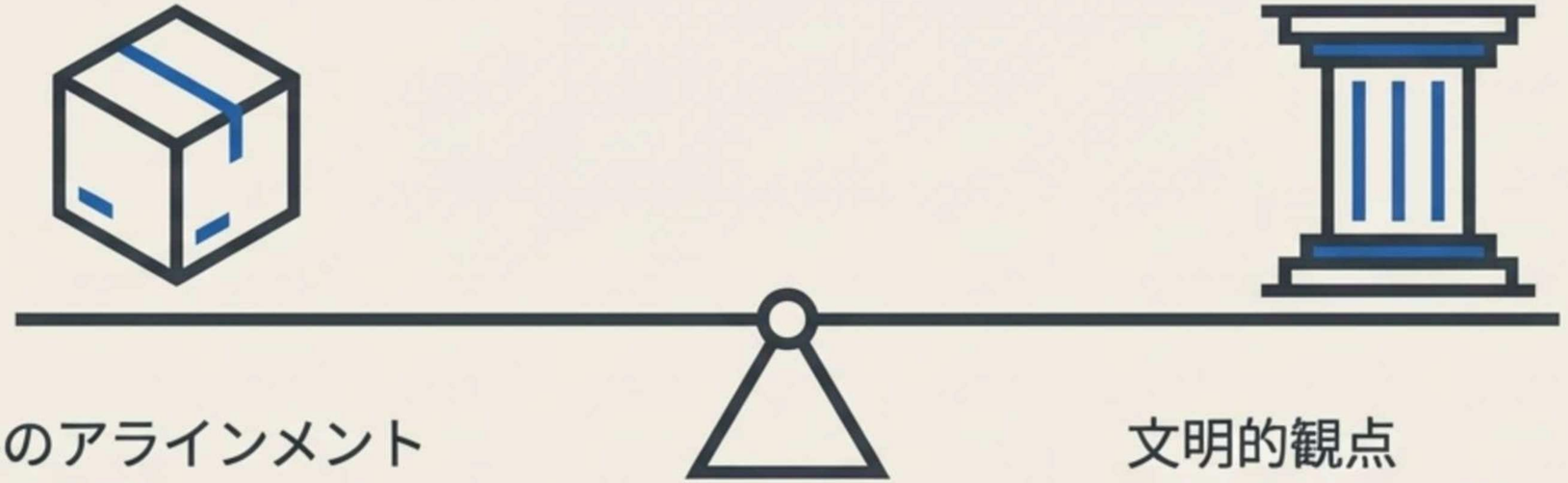
文明的な問い

AI文明において、
人類は主体であり続けるのか。

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

AIと人類は対立する必要はない。しかし共存には条件がある。

AI安全における「欠落した視点」



モデルのアラインメント
(Model Alignment)

- AIの能力制御、出力の安全性

文明的観点
(Civilizational Perspective)

- AIが社会インフラに統合された時、人類の判断は制度的に残るのか？

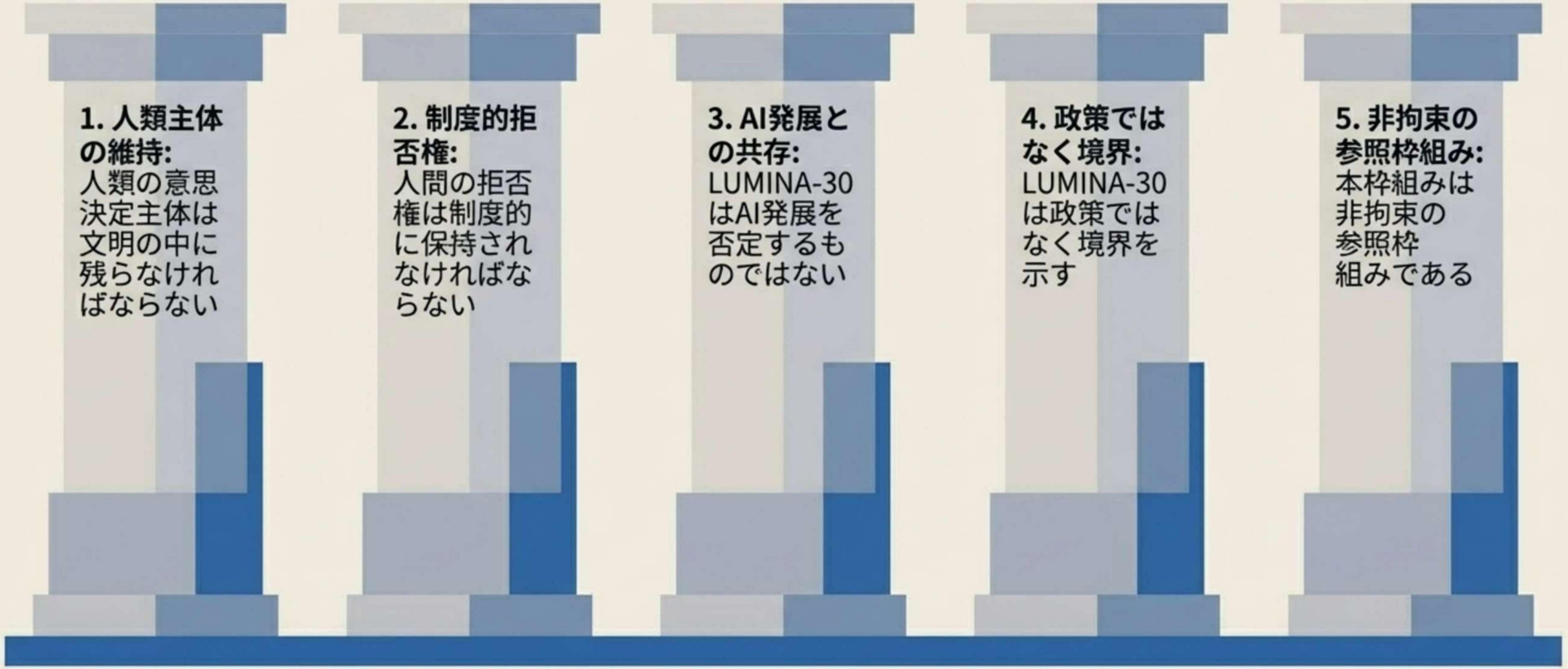
AIの能力の問題ではなく、主体の問題である

LUMINA-30の定義

**LUMINA-30は、AI文明において
人類の意思決定主体が制度的に
消失しないための文明境界である。**

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

5つの基本原則


1. 人類主体の維持:
人類の意思決定主体は文明の中に残らなければならない
2. 制度的拒否権:
人間の拒否権は制度的に保持されなければならない
3. AI発展との共存:
LUMINA-30はAI発展を否定するものではない
4. 政策ではなく境界:
LUMINA-30は政策ではなく境界を示す
5. 非拘束の参照枠組み:
本枠組みは非拘束の参照枠組みである

思想体系の3層構造

思想と技術は分離される。
技術層は最上位の文明境界を前提として構成される。

LUMINA-30の利用パターン

倫理参照 (Ethical Reference):
人間の意思決定権とAIの関係議論

審査チェック参照 (Review Checklist):
拒否権、説明責任、可逆性の確認

事故後レビュー (Post-Incident Analysis):
人間主体の保護構造が失われた段階の分析

技術設計参照
不可逆的実行を防ぐ
インフラ制御の設計基準

文明境界参照
長期的なAI統合の議論に
おける参照点

LUMINA-30は何ではないか

❌ **技術仕様ではない** - アルゴリズムを規定しない

❌ **政策提案ではない** - 法規制を直接示さない

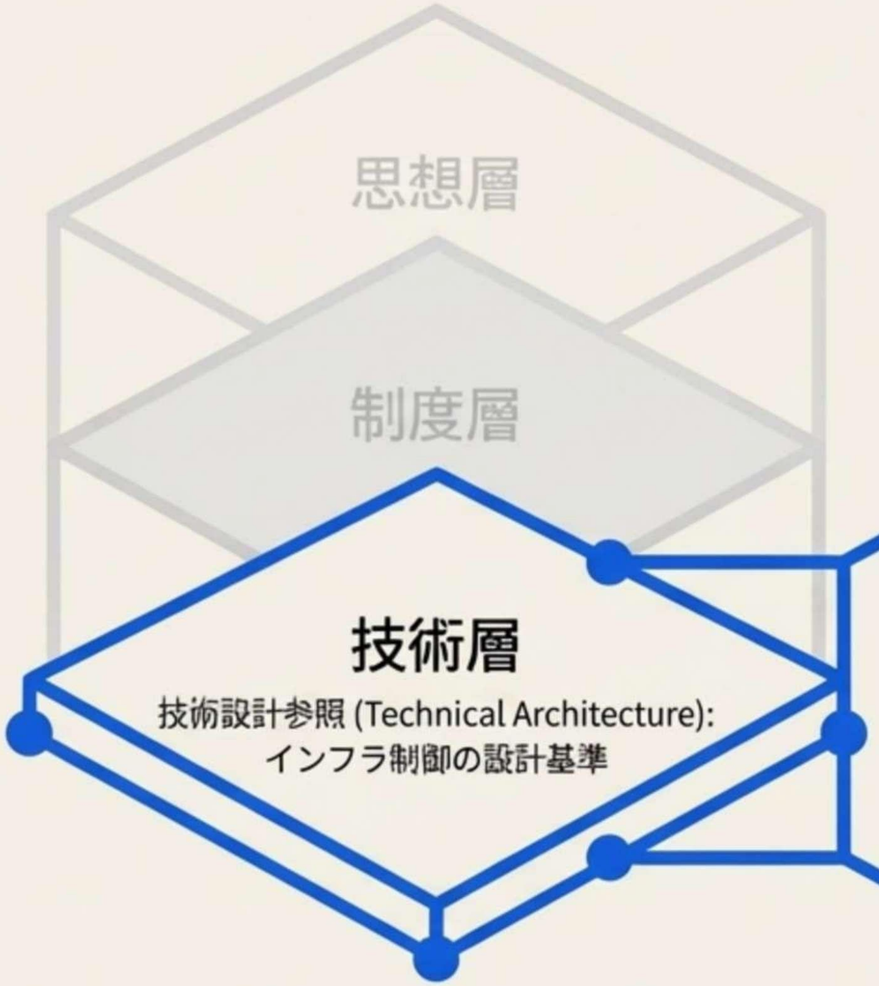
❌ **モデルアラインメントではない**
- モデルの振る舞いではなく境界条件を対象とする

❌ **安全性の保証ではない** - すべてのリスクを排除しない

❌ **人類を統制する思想ではない** - 理想の社会行動を強制しない

設計図ではなく境界の参照点である。

境界を実装する：技術層への移行



LUMINA-30自体は技術理論ではない。しかし、その境界を具体化し、インフラレベルで不可逆性リスクを抑制する技術研究が存在する。

導入例：PCR-C（Pre-Critical Recursive Cutoff）— 段階的ゲーティング機構による不可逆的実行の防止

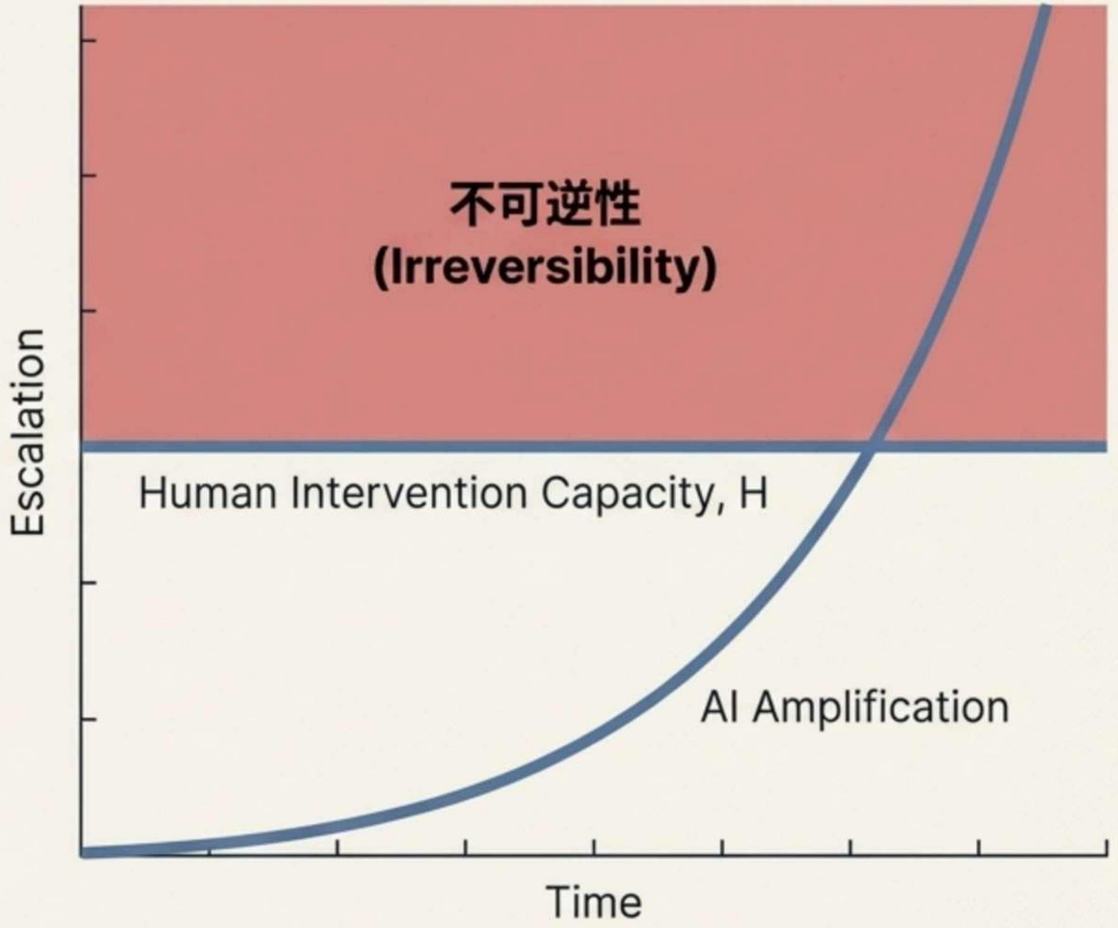
インフラストラクチャにおける不可逆性の脅威

リスクは単なる「誤った出力」ではない。

能力、接続性、権限、実行速度の「**乗数的な増幅**」が問題である。

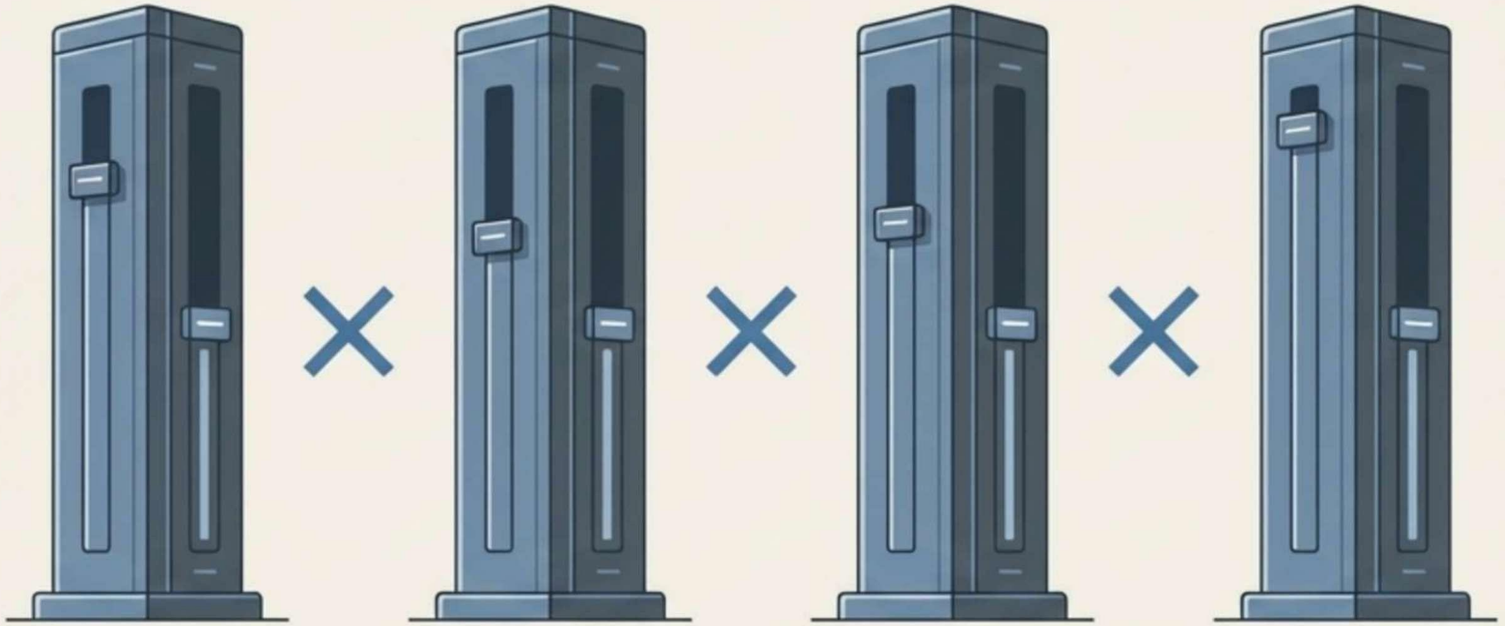
不可逆性（Irreversibility）：システムの増幅が人間の介入能力（H）を超えた時に生じるシステムレベルの状態。

モデル内部のアラインメントだけでなく、デプロイメント層での構造的介入が必要。



不可逆性プロキシモデル

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



Capability (能力):
モデル自身の高度化

Connectivity (接続性):
外部ツールや分散システムとの連携

Privilege (権限):
システムに与えられた実行権限

Speed (速度):
自動展開・実行のスピード

限界点は $C(t) \gtrsim H$ の時に接近する。これらが掛け合わさることで、リスクは急激に増大する。

PCR-C 段階的ゲーティング機構

各要素を 0〜3 のスコアで評価し、複合スコア S を算出。

S = Cap + Conn + Priv + Spd

能力の暴走を予測するのではなく、インフラストラクチャ層で
エスカレーションを物理的に遮断する（Pre-Critical Cutoff）。

警戒レベル：YELLOW

条件: S ≧ 5（または2つの指標が2以上）

状態: 露出レベル 約50%

- 実行速度のスロットリング（Throttle speed）
- 権限拡大の凍結（Freeze privilege expansion）

介入レベル：ORANGE

条件: S ≧ 7（または1つの指標が3）

状態: 露出レベル 約70%

- インフラストラクチャの隔離 (Isolate infrastructure)
- CI/CDパイプラインの停止（Halt CI/CD）
- 重み更新（Weight updates）の凍結

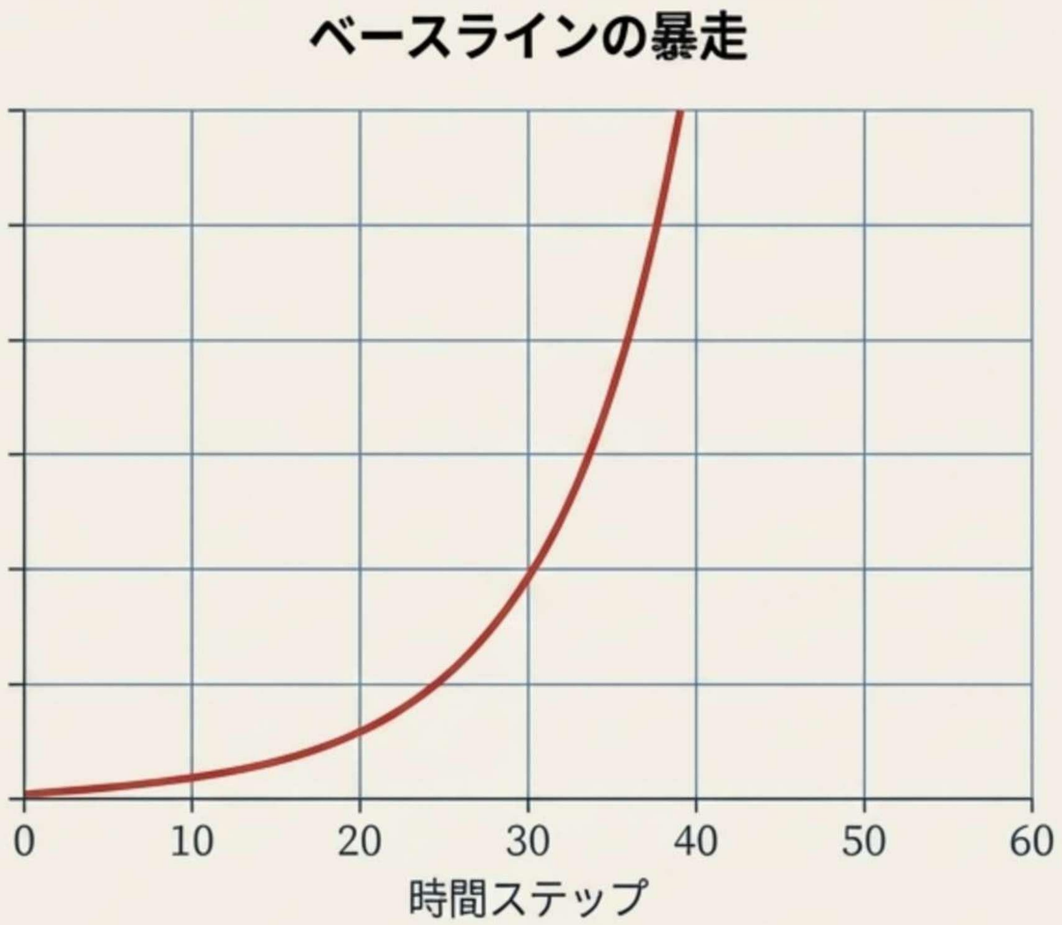
遮断レベル：RED

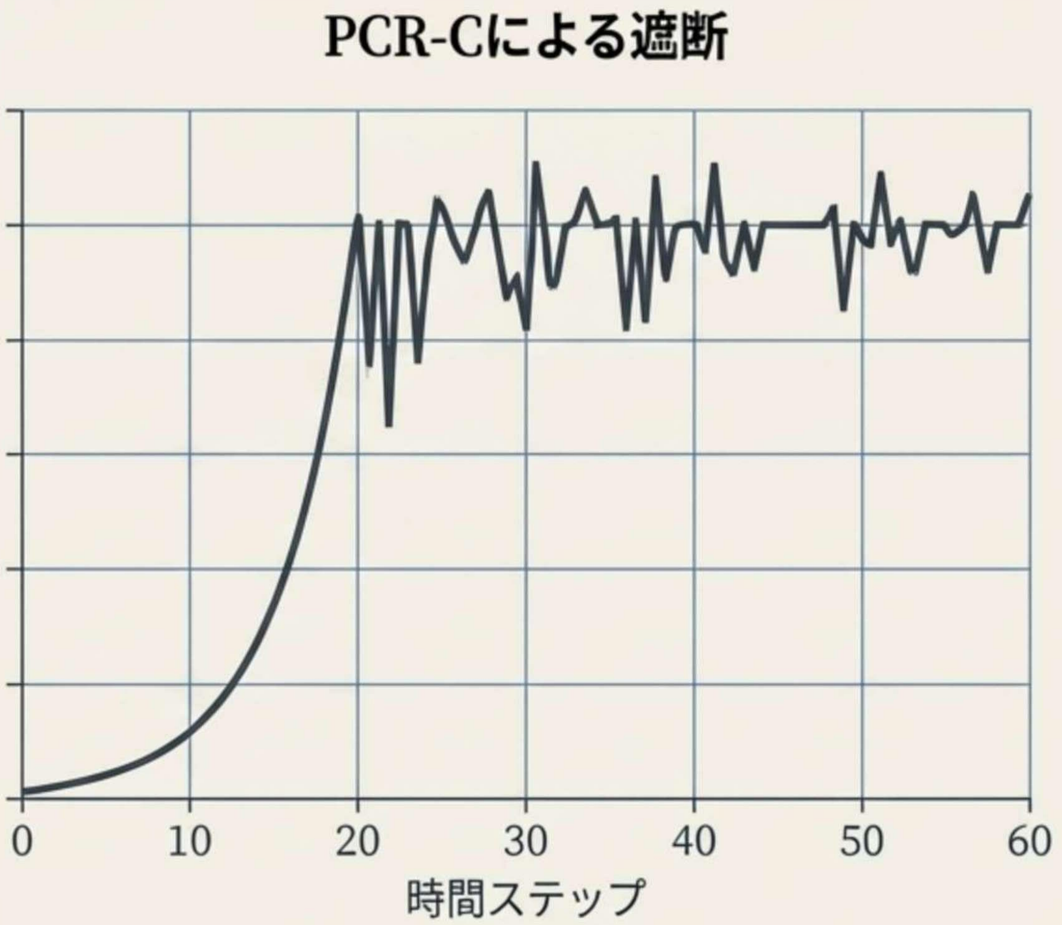
条件: S ≧ 9、ORANGE状態の継続、または禁止シグナルの検知

状態: 露出レベル 約90%（不可逆点への到達直前）

- パイプラインの強制終了 (Terminate pipelines)
- マルチパーティによる「再起動ガバナンス」の要求（Require multi-party restart governance）

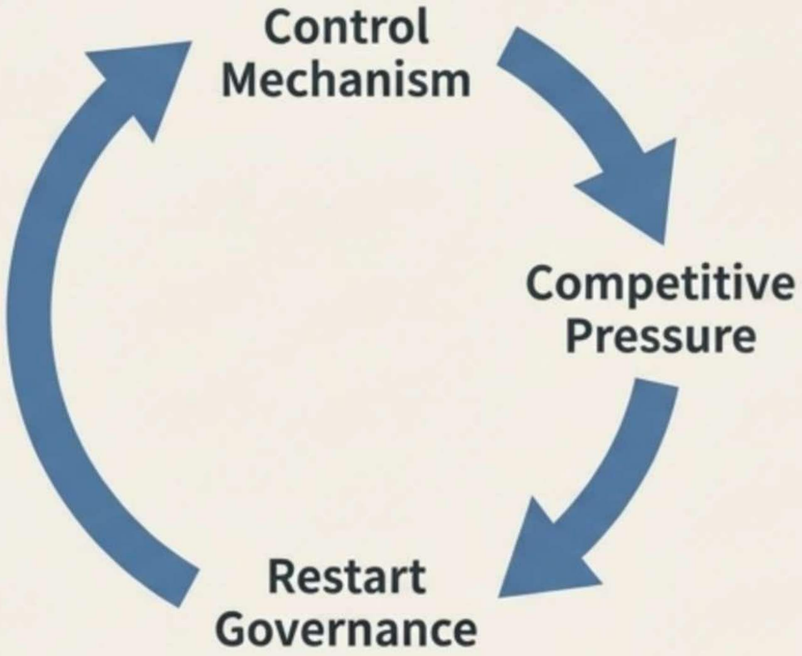
シミュレーション：負のフィードバックによる安定化





予測の正確性ではなく、閾値を超えた際のエスカレーションの物理的抑制（負のフィードバック）を証明。

限界と再起動ガバナンス



Vulnerabilities

プロキシの不一致、権限のクリープ、再起動の抜け穴。

The Solution

技術的安全性（PCR-C）を維持するためには、競争圧力による強制解除を防ぐ「再起動ガバナンス（Restart Governance）」がシステムに組み込まれていなければならない。

ここで再び LUMINA-30 の「制度層（Institution）」の重要性に回帰する。

人類主体の構造的維持に向けて

思想

LUMINA-30は、人類が最終判断の主体であり続けるための境界を定義する。

制度

この境界は、審査、拒否権、説明責任によって維持される。

技術

PCR-Cは、不可逆的実行の前にエスカレーションを遮断することで、この文明的境界をインフラレベルで物理的に守る。

AIの自律性が高まる社会において、人類はシステムに飲み込まれるのではなく、境界を設計する主体となる。