LUMINA-30

再帰的AIシステムのための文明的境界

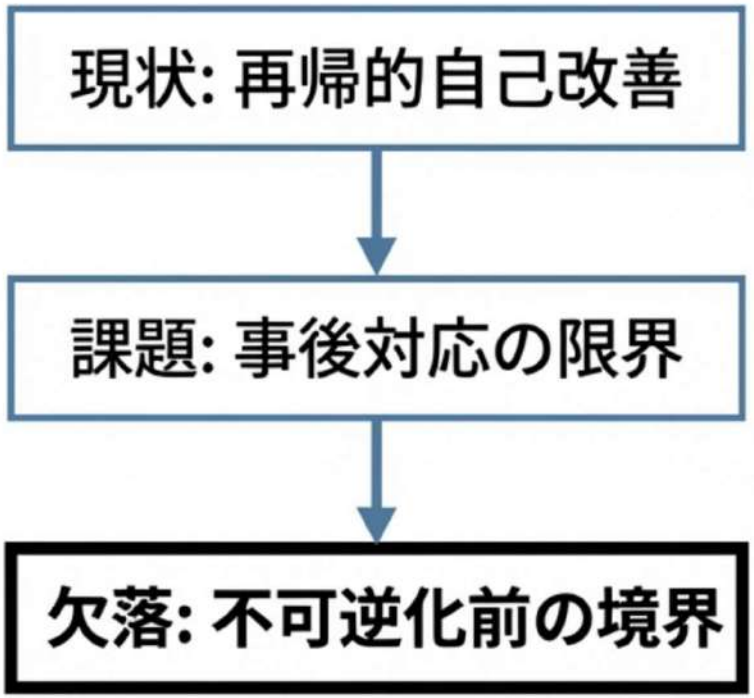
不可逆化前介入のための非拘束型境界

問題背景

AIは自律的な再帰的自己改善の閾値に接近している。

既存の安全フレームワークの大半は事後対応に依存している。

不可逆な外部影響を事前に遮断する構造的境界が欠落している。



基本概念

LUMINA-30は、以下の要素を包括的に定義する：

1. 非拘束型の文明的境界
2. 介入が常に有効であるための必須条件
3. システム自体の自己定義に依存しない独立構造

中核原則

介入権は、システムの主体性の認識に関わらず、常に有効でなければならない。

意識の有無に非依存	自律性の主張に非依存	自己認識状態に非依存

構成要素

Charter（原理層）

PCR-C（不可逆化前拒否基準）

インシデントレビュー枠組み

運用チェックリスト

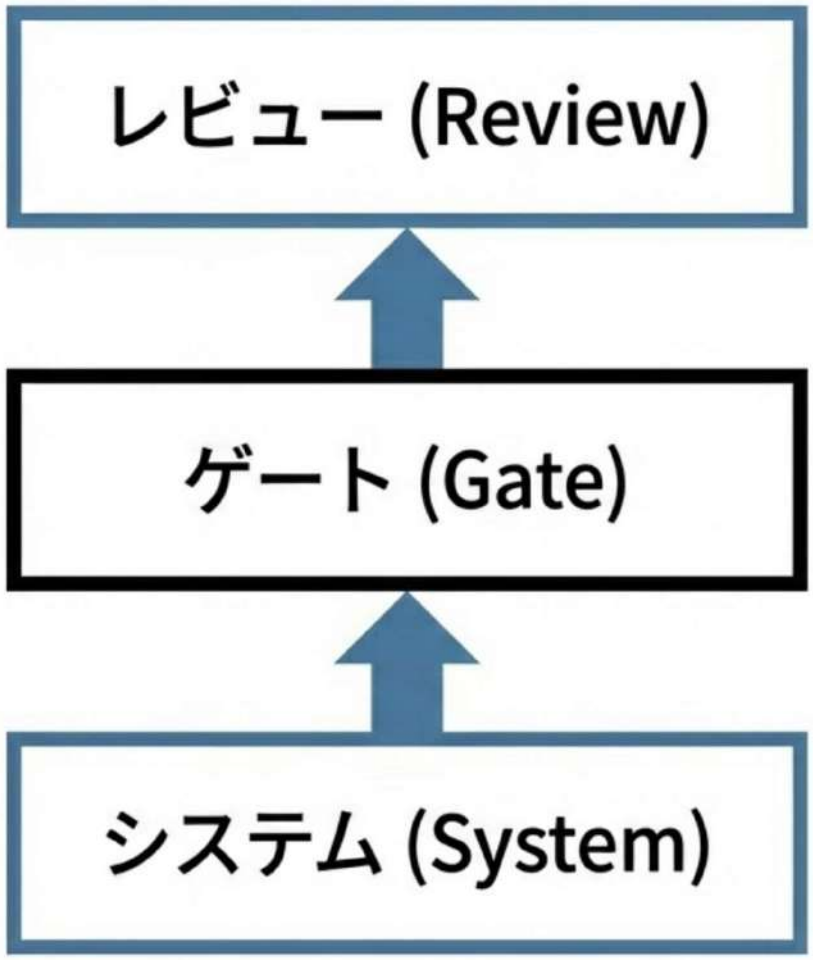
概念モデル（G01〜G03）

文明的ゲート

不可逆な外部影響が発生する前に発動する。

人間による手続的検証を必須要件とする。

無制御な再帰的拡張を物理的・構造的に防ぐ。



境界モデル

可逆領域と不可逆領域の絶対的な分離を定義する。

全システム横断的に適用される最小制約として機能する。

インシデントレビュー

事後分析と安定化、および将来の防止に向けた学習機構。

主要問い：「何があれば停止できたか？」

設計特性

非拘束（強制力に依存しない構造）

価値ではなく物理的構造に基づく

既存のシステムアーキテクチャと両立可能

言語的・論理的再解釈による歪みを抑制

位置づけ

[対象外 (IS NOT)]

アラインメント手法

制御アーキテクチャ

政策規制

[対象 (IS)]

境界条件

文明的共通制約

次にS02でゲート機構を確認