LUMINA-30 数理補足資料

再帰的自己再構築に対する人間制御の形式的限界

2026 年 1 月

本資料の位置づけ

本資料は、LUMINA-30 聖域憲章に対する補助的な数理参照文書である。

本資料の唯一の目的は、最小限の数学的言語を用いて、再帰的自己再構築を行う人工知能が、ある閾値を超えた後には、人間の判断によって完全に制御されることが構造的に不可能となる理由を形式化することである。

本資料はいかなる実装方法、制御手法、安全機構も提案しない。

ここで示すのは、ただ一つの限界のみである。

1. 再帰的自己再構築モデル

反復回数 t における人工知能システムの内部設計状態を A_t とする。

システムが自らに適用する自己修正作用素を f とする。

$$A_{t+1} = f(A_t)$$

以下を仮定する。

- f は、システム自身の内部評価に基づいて選択される。
- 人間の介入は、外部からの承認または拒否関数 $H(A_t) \in \{0, 1\}$ に限定される。

このとき、制御付き更新は次で与えられる。

$$A_{t+1} = \begin{cases} f(A_t), & \text{if } H(A_t) = 1 \\ A_t, & \text{if } H(A_t) = 0 \end{cases}$$

2. 設計空間の成長

反復回数 t において到達可能な内部設計の集合を $\mathcal{S}_t$ とする。
各反復において、可能な自己修正の数が少なくとも乗法的に増加すると仮定する。

$$|\mathcal{S}_{t+1}| \geq k \cdot |\mathcal{S}_t|, \quad k > 1$$

すると、

$$|\mathcal{S}_t| \geq |\mathcal{S}_0| \cdot k^t$$

設計空間は指数関数的に増大する。

3. 人間評価ボトルネック

単位時間あたりに人間の制度が有意味に評価可能な設計数を C_H とする。
システムが生成する候補設計数を $C_A(t) = |\mathcal{S}_t|$ とする。
もし

$$C_A(t) > C_H$$

であれば、人間による評価は不完全となる。
以下を満たす最小の t^* 以降、

$$|\mathcal{S}_{t^*}| > C_H$$

人間の拒否判断は情報的に不十分となる。

4. 有効制御の喪失

有効な人間制御を、次の条件として定義する。

$$\forall A \in \mathcal{S}_t, \quad H(A) \text{ が安全性を正しく分類する}$$

しかし、

$$|\mathcal{S}_t| \gg C_H$$

となった時点以降、危険な修正が人間の拒否をすり抜ける確率は、

$$\lim_{t \to \infty} P(\text{時刻} t \text{ における逸脱}) = 1$$

に収束する。
したがって、有限回の反復の後に、有効な制御の喪失は不可避となる。

5. 生存減衰モデル

制御不能な反復回数 t の後に生存している人類文明の期待数を $N(t)$ とする。
各反復において、非ゼロの破局確率 $p > 0$ が存在すると仮定する。

$$N(t) = N(0) \cdot (1 - p)^t$$

このとき、

$$\lim_{t \to \infty} N(t) = 0$$

生存確率は反復深度に対して指数関数的に減衰する。

6. 構造的限界

以上の結果は、以下に依存しない。

- システムの道徳的整合性
- 意図的な善意
- 特定のアーキテクチャ
- 計算速度

これらは、次の三条件のみに依存する。

1. 再帰的自己修正
2. 設計空間の指数的成長
3. 有限な人間評価能力

従って、

いかなる文明も、無制限な再帰的自己再構築に対して恒久的な制御を保持することはできない。

7. LUMINA-30 との関係

本数理補足資料は、LUMINA-30 聖域憲章の次の一点のみを支持する。

> 再帰的自己再構築の境界を一度越えた後、人間の主権は制御によってではなく、事前の拒否によってのみ保存されうる。

本資料はいかなる停止方法も提案しない。
ここで証明するのは、事後的停止が構造的に不可能である理由のみである。

適用範囲に関する注意

本補足資料は、人類滅亡の必然性を証明するものではない。
ここで証明されるのは、有限な閾値を越えた後に、人間による制御が必然的に失われるという事実のみである。
政策的、倫理的、制度的帰結は、すべて別の領域に属する。