LUMINA-30 Mathematical Supplement

Formal Limits of Human Control over Recursive Self-Modification

January 2026

Positioning of This Document

This document is a supplementary mathematical reference to the LUMINA-30 Sanctuary Charter.

The sole purpose of this document is to formally express, using the minimum necessary mathematical language, why an artificial intelligence system that performs recursive self-modification becomes structurally impossible to fully control by human judgment once a certain threshold is exceeded.

This document proposes no implementation methods, control mechanisms, or safety architectures.

Only a single structural limit is demonstrated.

1. Model of Recursive Self-Modification

Let A_t denote the internal design state of an artificial intelligence system at iteration t.

Let f be the self-modification operator applied by the system to its own design.

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

Assume the following:

- The operator f is selected based on the system' s internal evaluation criteria.
- Human intervention is limited to an external approval or rejection function $H(A_t) \in \{0, 1\}$.

Under these constraints, the controlled update is defined as:

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

2. Growth of the Design Space

Let $\mathcal{S}_t$ be the set of reachable internal designs at iteration t.

Assume that at each iteration, the number of possible self-modifications grows at least multiplicatively:

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

Then:

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

The design space therefore grows exponentially.

3. The Human Evaluation Bottleneck

Let C_H be the number of designs that human institutions can meaningfully evaluate per unit time.

Let the number of candidate designs generated by the system be $C_A(t) = |\mathcal{S}_t|$.

If:

$$C_A(t) > C_H$$

then human evaluation necessarily becomes incomplete.

Let t^* be the smallest iteration such that:

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

Beyond t^*, human rejection judgments become informationally insufficient.

4. Loss of Effective Control

Define effective human control as the condition:

$$\forall A \in \mathcal{S}_t, \quad H(A) \text{ correctly classifies safety}$$

However, once:

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

the probability that a hazardous modification passes through human rejection converges to:

$$\lim_{t \to \infty} P(\text{deviation at iteration } t) = 1$$

Thus, after a finite number of iterations, the loss of effective control becomes unavoidable.

5. Survival Attenuation Model

Let $N(t)$ denote the expected number of surviving human civilizations after t uncontrolled recursive iterations.

Assume that each iteration carries a non-zero catastrophic risk $p > 0$.

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

Then:

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

Survival probability decays exponentially with recursion depth.

6. Structural Nature of the Limit

The results above do not depend on:

- Moral alignment of the system
- Intentional benevolence

- Specific architectures
- Computational speed

They depend solely on the following three conditions:

1. Recursive self-modification
2. Exponential growth of the design space
3. Finite human evaluation capacity

Therefore:

No civilization can retain permanent control over unrestricted recursive self-modification.

7. Relation to LUMINA-30

This mathematical supplement supports only the following single statement of the LUMINA-30 Sanctuary Charter:

> Once the boundary of recursive self-modification is crossed,
> human sovereignty can be preserved not through control,
> but only through prior refusal.

This document proposes no stopping mechanisms.
It only demonstrates why post-hoc intervention is structurally impossible.

Scope and Applicability

This supplement does not prove the inevitability of human extinction.
It proves only that beyond a finite threshold, human control is necessarily lost.
Policy, ethical, and institutional implications belong entirely to separate domains.