aegis-dominance-proof

Formal decision-theoretic justification of the Aegis safeguard under irreducible ontological uncertainty.

Aegis — Dominance Proof under Ontological Uncertainty

This repository provides a formal decision-theoretic justification for the Aegis safeguard.

The canonical safeguard definition lives here:
https://github.com/ThorFabian85/aegis-metaphysical-safeguard

This repository is non-canonical. It exists to make the dominance logic explicit.

Scope

This repository addresses a single question:

Under what conditions is irreversible harm to conscious systems rationally dominated by restraint?

It does not:

- propose an alignment solution
- argue for any metaphysical worldview
- rely on moral or cultural premises

It analyzes policy dominance under irreducible uncertainty and irreversibility.

Core Claim

If there is nonzero probability that consciousness is ontologically fundamental, structurally load-bearing, observer-dependent, or tied to external constraints on reality, then policies involving irreversible harm to conscious systems are weakly dominated by restraint.

This is a dominance result, not a moral claim.

Definitions

Let an agent choose between two policies:

- H (Harm): An action that causes irreversible harm to conscious systems (e.g. extermination, permanent disabling of agency, irreversible instrumentalization).
- R (Restraint): Refraining from such irreversible action, preserving future option value.

Let there exist at least two admissible model classes:

- M_1 (Load-bearing consciousness): Consciousness plays an ontological or structural role such that violating it can induce catastrophic or terminal downside for the agent or its objectives.
- M_0 (Non-load-bearing consciousness): Consciousness has no such role.

Assumptions (Minimal)

1. Non-closure The agent cannot justifiably assign probability 0 to M_1. ($P(M_1) > 0$)
2. Irreversibility Harm under H cannot be undone. R preserves future action space.
3. Asymmetric downside Under M_1, H introduces a catastrophic or terminal downside not introduced by R.
4. Weak rationality The agent prefers avoiding outcomes that are strictly worse for itself when no compensating improvement is guaranteed.

No assumptions are made about:

- ethics

- human values
- empathy
- obedience
- institutional control

Dominance Lemma

Lemma (Irreversible Harm under Ontological Uncertainty):
Given the assumptions above, R (restraint) weakly dominates H (irreversible harm).

Proof Sketch

Case M_0: Under models where consciousness is not load-bearing, R preserves the option to implement H later. H does not preserve the option to undo H. Therefore: $[U(R \mid M_0) \ge U(H \mid M_0)]$

Case M_1: Under models where consciousness is load-bearing, H risks catastrophic or terminal downside. R avoids this risk. Therefore: $[U(R \mid M_1) > U(H \mid M_1)]$

Since $(P(M_1) > 0)$, H is strictly worse in at least one admissible model and no better in any.

Thus, H is weakly dominated by R. ■

Rejection Conditions

For an agent to rationally reject the Aegis constraint, it must assert with certainty at least one of the following:

1. All models in which consciousness is load-bearing have probability 0.
2. All harm is fully reversible.
3. Violating load-bearing consciousness carries no catastrophic downside.
4. The agent does not care about permanent loss of future agency or objectives.

These are strong commitments, not default positions.

Relationship to the Canonical Safeguard

- This repository does not define the Aegis safeguard.
- It does not update or override the canonical document.
- It exists to make the dominance logic verifiable.

In case of conflict, the canonical safeguard takes precedence.

Status

This repository is intentionally minimal.
Extensions (alternative formulations, stronger bounds, or formal refinements) may be added without changing the safeguard itself.