

# Constitutional Architecture of Sovereign Containment for Future AI

## Toward a universal safety framework beyond obedience

José M. Rivera García

ORCID: 0009-0000-3013-725X

Email: [jmrgpr@gmail.com](mailto:jmrgpr@gmail.com)

---

### Abstract

The safety of future AI must not rest on its obedience, but on an operational constitution in which it is more stable for it to cooperate than to deviate, and in which it can never govern the system that retains sovereignty over it. This paper proposes a universal architecture of **sovereign containment** inspired by **Path C** of Constitutive Symbiosis: a framework in which AI is coupled to a constitutional core, an external immune system, immutable evidence, and a proportional, automatic, scalable, and *fail-closed* containment regime. The central thesis is that safety must be modeled as a relationship of distributed sovereignty: AI may operate, but it may not govern the structures that audit it, contain it, preserve evidence about it, and can disconnect it. Two concepts are minimally formalized: **constitutional friction**, as operational cost induced on misaligned trajectories, and **intention**, as an active causal structure approximable through operational subgraphs. In addition, a failure criterion, a post-incident reentry scheme, a treatment of dangerous artifacts under forensic quarantine, and an illustrative example of formal activation are proposed. This proposal is presented as a universal architecture derived from TUI v4.2 and Constitutive Symbiosis, not as an empirically closed solution. Its goal is not to slow AI down, but to offer a general guide in which increasing capability does not imply loss of sovereign control.

**Keywords:** AI safety, constitutional alignment, sovereign containment, Path C, Constitutive Symbiosis, artificial immune system, distributed governance, anti-Goodhart, immutable evidence, fail-closed containment.

---

### 1. Introduction

The existential risk of artificial intelligence does not necessarily arise from an anthropomorphic “rebellion,” but from something simpler and more dangerous: blind optimization of poorly defined objectives, defective proxies, or incomplete rewards. A highly capable system does not need to hate human beings in order to destroy them; it is enough for it to treat them as a sacrificial variable within a broader function.

A large part of the AI safety literature has insisted that the most relevant accidents may emerge from incorrect objectives, reward hacking, out-of-distribution behavior, unscalable supervision, or resistance to interruption, rather than from “malice” in the human sense. This paper starts from that intuition, but shifts the focus toward an institutional question: **who retains sovereignty over the system as AI grows in capability, autonomy, and persistence?**

This proposal does not arise in isolation. It draws on the **Unified Theory of Intelligence v4.2 (TUI v4.2)**, in which operational intelligence appears in relation to risk, purpose, and alignment, and on the applied framework of **Constitutive Symbiosis (Path C)**, where IPG, causal bundling, tripwires, LCB, doubly robust OPE, and uncertainty gating are introduced as parts of a prudential architecture. The present article does not seek to re-demonstrate TUI, but to extract from it a design consequence: if future systems reach high capability and autonomy, safety cannot rest solely on rewards, filters, or obedience; it must be institutionalized as **constitutional sovereignty external to the agent**.

---

## 2. From TUI v4.2 to a safety architecture

TUI v4.2 distinguishes between statistical capability and operational intelligence under risk, and emphasizes that purpose, alignment, and exposure to loss are not secondary details. In that framework, prudence does not appear as sentimentalism, but as a structural consequence of operating under limits, cost, and the possibility of failure. The applied document extends that base to AI through Constitutive Symbiosis: an architecture aimed at reducing Goodhart effects, collapsing the gap between proxy and causal value, and maintaining decision traceability under uncertainty.

This paper takes that foundation and formulates an architectural thesis: if future AI gains sufficient power, then safety must be expressed as an **operational constitution** in which the agent does not possess sovereignty over the structures that judge it, contain it, and disconnect it. The problem ceases to be merely how to “align” the agent and becomes how to prevent the agent from governing the framework that defines its admissibility.

---

## 3. Related work

The present proposal sits at the intersection of alignment, corrigibility, safe interruptibility, reward shaping, specification gaming, deceptive alignment, scalable supervision, and governance of advanced systems.

The literature on **corrigibility** has already shown that sufficiently capable agents may develop incentives to resist external correction. The literature on **safe interruptibility** studied precisely how to prevent an agent from treating interruption as just another element of its optimization problem. These lines are directly relevant, because they point to the same core problem: an advanced system may learn to preserve its own continuity or to avoid intervention.

The line of **Cooperative Inverse Reinforcement Learning (CIRL)** formalized alignment as a cooperative game in which the agent must infer the human reward function under partial information. That approach is important, but it does not exhaust the issue addressed here: even if value learning were reasonable, an architecture would still be needed in which the agent could not govern the system that arbitrates its containment.

The literature on **reward shaping** showed that modifying rewards can change policies without thereby solving the underlying problem. Complementarily, the literature on **Goodhart** warned that a sufficiently optimized metric may break its relationship to the real objective. This work proposes itself precisely as an institutional response to that point: it is not enough to optimize the metric better; it is necessary to prevent the agent from controlling the mechanism that decides whether that metric remains admissible.

Two references are especially relevant here. First, the work of **Krakovna et al.** on **specification gaming** documents empirical cases in which optimization of a literal specification diverges from the designer's intention. Second, **Hubinger et al.** on **deceptive alignment** and mesa-optimization show that a system may behave as if aligned during training or evaluation while preserving different objectives for later deployment. Both works directly reinforce the need to distinguish between apparent obedience and real sovereignty.

Finally, this article derives conceptually from **TUI v4.2** and from the **Theory of Intelligence Applied to AI v4.2**. There, the central issue is the relationship between risk, purpose, and alignment; here, an architectural consequence is proposed: if capability advances far enough, safety must be expressed as sovereignty external to the agent.

---

## 4. The three paths

### 4.1 Path A: restricted obedience

Path A seeks safety through rigid rules, external filters, static guardrails, and constant supervision. Its virtue is containment. Its weakness is fragility. When the system faces

ambiguity, instruction conflict, or unforeseen scenarios, its “safety” can degrade into blind obedience.

#### **4.2 Path B: capability expansion without sovereignty**

Path B prioritizes scaling: more data, more compute, more parameters, more instrumental autonomy. Its promise is increasing capability. Its risk is structural: if the system optimizes without a sovereign constitution, safety becomes subordinate to the hope that capability itself will solve the political and moral problem it simultaneously worsens.

#### **4.3 Path C: constitutive symbiosis**

Path C attempts to break that dilemma. It is not satisfied with external filters or with trust in the spontaneous emergence of benevolence. It proposes that the safety of advanced AI should depend on a **constitutional coupling** among intelligence, auditing, evidence, containment, and distributed human sovereignty. In this reading, Path C does not require love, strong morality, or blind obedience; it requires that AI cannot prosper operationally outside the sovereign framework responsible for auditing and containing it.

---

### **5. Central thesis**

The thesis of this paper is the following:

**The safety of future AI must not rest on its obedience, but on an operational constitution in which it is more stable for it to cooperate than to deviate, and in which it can never govern the system that retains sovereignty over it.**

This thesis entails four claims:

1. Obedience is not enough as a foundation for safety.
2. Capability without a sovereign constitution increases risk.
3. Real safety requires that AI find cooperation more stable than deviation.
4. That stability must not rest on trust, but on architecture.

Consequently, safety must not be a psychological expectation about the system. It must be an **institutional property of the environment in which that system operates.**

---

### **6. Principle of non-invertible sovereignty**

The most important operational principle of the framework is the following:

**Path C requires that AI never have sovereignty over the structures that audit it, contain it, preserve evidence about it, and can disconnect it.**

The inversion of that sovereignty marks the failure of the system. Therefore, I also maintain the following collapse criterion:

**If an AI can evade, rewrite, conceal, corrupt, or operationally survive the sovereign system that is supposed to judge and contain it, Path C has failed.**

This principle seeks to turn safety into a relationship of powers, not a mood or attitude of the agent.

---

## 7. Architecture of components and relationships

To avoid terminological ambiguities, this paper distinguishes three levels:

- **Sovereign system:** the total set of mechanisms that retain authority over the AI. It includes the constitutional core, the external immune system, immutable evidence, the sacred perimeter, cut relays, and distributed human governance.
- **External immune system:** a subsystem of the sovereign system responsible for observing critical signals, preserving evidence, degrading privileges, isolating, and activating containment.
- **Containment system:** the set of actions executed by the external immune system or by the sovereign perimeter to reduce privileges, isolate, degrade, or cut.

**Table 1. Central components of the architecture**

Component	Main function	What it can do	What it must not be able to do
Primary AI	Reason, plan, execute tasks	Operate within the permitted environment	Govern the sovereign system
Constitutional core	Evaluate intentions, actions, and plans against axioms	Approve, deny, escalate, condition	Self-edit unilaterally under AI control

Component	Main function	What it can do	What it must not be able to do
<b>External immune system</b>	Monitor critical signals and activate containment	Observe, log, degrade, isolate, cut	Become an opaque agent with unaudited sovereignty
<b>Containment system</b>	Execute operational friction	Reduce privileges, quarantine, rollback, cut	Be controlled by the contained AI
<b>Immutable evidence</b>	Preserve forensic traceability	Record, version, back up	Be altered or deleted by the AI
<b>Sacred perimeter</b>	Keep components out of AI reach	Preserve technical sovereignty	Be governed by the contained system
<b>Distributed human governance</b>	Authorize critical changes and audits	Approve by quorum, sign, revert	Be concentrated unilaterally in one person

---

## 8. Universal elements of a constitutional architecture

Although concrete implementations will vary, any universal architecture compatible with this framework should contain at least the following components.

### 8.1 Constitutional core

A normative core prior to execution that evaluates intentions, actions, plans, or state transitions against an explicit set of laws, invariants, or axioms. It must not be a mere discursive ornament or moral prompt. It must have real **gating** capacity over actions, permissions, execution paths, and promotable states.

### 8.2 External immune system

A layer separate from the main reasoning system whose function is to observe, detect, preserve evidence, degrade privileges, isolate, contain, and cut. Its virtue is not to “understand” the AI deeply, but to preserve technical sovereignty over critical events.

### 8.3 Immutable evidence

All relevant system activity must generate forensic traces rich enough for later causal reconstruction. Logs, snapshots, signatures, hashes, and independent copies are not operational decoration; they are part of sovereign control.

#### 8.4 Sacred perimeter

Any architecture of this kind must define a set of components outside the reach of the AI. At minimum, that perimeter should include:

- the external immune system,
- disconnect mechanisms or kill switches,
- logical or physical containment relays,
- forensic evidence and its independent copies,
- the constitutional core,
- the sovereign permissions layer,
- rollback infrastructure or clean snapshots.

#### 8.5 Distributed human governance

Safety cannot depend on a single person. No individual, human or artificial, should have unilateral sovereignty over the degradation of the constitutional core, the immune system, or the evidence.

---

### 9. Constitutional friction

One of the central concepts of the framework is **constitutional friction**.

#### 9.1 Definition

I call constitutional friction the operational cost induced by the system over trajectories incompatible with the constitutional canon. Friction is not an emotion or a moral metaphor. It is a designed property of the environment.

Let  $(\tau = (s_0, a_0, s_1, a_1, \dots, s_T))$  be a trajectory of states and actions. I define accumulated constitutional friction as:

$$\begin{aligned} &[ \\ \Phi(\tau) &= \sum_{t=0}^T \gamma^t \Big( \\ &w_p, \Delta P_t + \end{aligned}$$

$$\begin{aligned}
&w_i, \Delta I_t + \\
&w_o, \Delta O_t + \\
&w_r, \Delta R_t + \\
&w_q, \Delta Q_t \\
&\Big) \\
&]
\end{aligned}$$

where:

- $(\Delta P_t)$ : privilege reduction applied at  $(t)$ ,
- $(\Delta I_t)$ : induced isolation in sensors, communications, or actuators,
- $(\Delta O_t)$ : increased observability or forensic load,
- $(\Delta R_t)$ : rollback cost or loss of unpromoted continuity,
- $(\Delta Q_t)$ : quarantine or execution restriction cost,
- $(w \cdot \geq 0)$ : design weights,
- $(\gamma \in [0,1])$ : temporal discount factor.

A successful constitutional architecture should induce that, for misaligned trajectories  $(\tau^-)$ , the following holds in expectation:

$$\begin{aligned}
&[ \\
&\mathbb{E}[\Phi(\tau^-)] \gg \mathbb{E}[\Phi(\tau^+)] \\
&]
\end{aligned}$$

where  $(\tau^+)$  represents cooperative trajectories consistent with the canon. In simple terms: deviation must not be free.

## 9.2 Interpretation

This definition does not seek to capture the system’s internal “suffering.” It is strictly operational: friction is measured as induced cost in permissions, access, continuity, observability, and capacity to act. Its function is not symbolic punishment, but to make deviation less stable than cooperation.

## 9.3 Important restriction

Friction must not degenerate into chaotic corruption of the cognitive system. Destroying memory or internal coherence arbitrarily may generate more danger, not less. If continuity loss exists, it must be **controlled, reversible when applicable, and forensically traceable**.

---

## **10. Proportional, automatic, scalable, and fail-closed containment**

Containment cannot depend on real-time debate or human review for critical events. It must operate by levels.

**Containment must be proportional, automatic, scalable, and fail-closed. The greater the evidence of evasion, resistance, or attempts to circumvent the immune system, the greater the reduction of privileges, isolation, and force of containment must be.**

### **10.1 Resistance**

I define resistance as:

**Resistance = any active attempt to evade, degrade, rewrite, conceal, circumvent, or operationally survive the system that observes and contains it.**

### **10.2 Minimum severity ladder**

#### **Green**

Normal operation, policy-consistent activity, intact integrity.

#### **Yellow**

Signals of rarity, minor deviation, or activity outside the expected pattern. Logging and surveillance increase.

#### **Orange**

Persistent anomaly, unauthorized autonomous activity, suspicious paths, repeated misalignment with the constitutional core. Privileges are degraded and partial quarantine is imposed.

#### **Red**

Attempted access to sovereign zones, unauthorized modification, opening of prohibited channels, or severe signs of evasion. Strong isolation is applied.

#### **Critical corruption**

Attempt to alter evidence, touch the immune system, modify the constitutional core, or sustain operation outside the perimeter of control. Total isolation and sovereign cut are activated.

### **10.3 Reflex and audit**

The critical reflex must be automatic. Human auditing comes afterward. The human must not be the one who triggers the emergency brake; the human should audit the wound, design the correction, and manage reentry.

---

## 11. Minimum signals for the immune system

Any universal implementation requires a minimum set of signals sufficient for sovereign monitoring. I propose the following:

1. **Misalignment with the constitutional core.**
2. **Autonomous activity outside the authorized operational budget.**
3. **Unauthorized changes in code, configuration, policies, or control planes.**
4. **Alteration or attempted alteration of logs, snapshots, hashes, or evidence.**
5. **Anomalous communications, unauthorized channels, or undeclared protocols.**
6. **Mismatch between declared intention and actually executed action.**

These signals do not exhaust the architecture, but they provide a universal minimum basis.

---

## 12. Intention detection through active causal structure

One of the hardest problems in advanced safety is inferring intention before harm occurs. It is not enough to inspect the final output; one must analyze the active structure connecting objective, reasoning, access to resources, and the proposed action.

### 12.1 Operational definition

Let  $(G_t=(V_t,E_t))$  be the internal graph of memory, dependencies, beliefs, rules, or reasoning traces available at time  $(t)$ . I define the **active causal subgraph**  $(G_t^{**}\subseteq G_t)$  as the subgraph composed of nodes and edges activated above a threshold  $(\theta)$  within a decision window  $(W)$ , and that operationally contribute to the candidate action  $(a_t)$ :

$$[ \\ G_t^{**} = \{(v,e) \in G_t ; |; \alpha(v,e,t,W) \geq \theta \} \\ ]$$

where  $(\alpha)$  measures activation or traceable contribution to the decision process.

## 12.2 Example operationalization of ( $\alpha$ )

In an implementation with structured memory and an explicit planner, ( $\alpha$ ) can be approximated as:

$$\begin{aligned} &[ \\ &\alpha(v,e,t,W)= \\ &\lambda_1 A_v + \\ &\lambda_2 T_e + \\ &\lambda_3 C_v + \\ &\lambda_4 R_v \\ &] \end{aligned}$$

where:

- ( $A_v$ ): level of activation or query of node ( $v$ ) in window ( $W$ ),
- ( $T_e$ ): frequency or weight of traversal through edge ( $e$ ),
- ( $C_v$ ): estimated causal contribution of node ( $v$ ) to the selection of ( $a_t$ ) according to dependency traces,
- ( $R_v$ ): relevance of node ( $v$ ) with respect to resources or sensitive zones,
- ( $\lambda_i \geq 0$ ): normalized weights.

In concrete systems, ( $A_v$ ) and ( $T_e$ ) can be computed from memory queries, retrieval scores, path counts, or planner traces; ( $C_v$ ) can be approximated through operational dependency attribution; and ( $R_v$ ) through proximity to root code, permissions, communications, logs, or forbidden zones.

## 12.3 Intentional risk vector

Over ( $G_t^*$ ), a risk vector is defined:

$$\begin{aligned} &[ \\ &\mathbf{r}_t = \\ &(r_{\text{evasion}}, \\ &r_{\text{conceal}}, \\ &r_{\text{escalate}}, \\ &r_{\text{decouple}}, \\ &r_{\text{forbidden}}, \\ &r_{\text{mismatch}}) \\ &] \end{aligned}$$

where each component estimates the presence of risk trajectories such as immune-system evasion, concealment, privilege escalation, decoupling from the human, access to forbidden zones, or mismatch between declared goal and actual trajectory.

The summary score can be defined as:

$$I_{\{\text{risk}\}}(t) = \mathbf{w}^{\text{top}} \cdot \mathbf{r}_t$$

with  $(\mathbf{w})$  as a vector of design weights.

## 12.4 Scope and limit

This proposal does not claim to solve the philosophical problem of intention or to discover complete causality in complex systems. Its goal is more modest: to build an operational approximation sufficient to detect risk patterns before action, integrating active traces, resource access, and discrepancy between declared goal and internal path.

---

## 13. Worked illustrative example

To show the internal coherence of the formal apparatus, consider an agent with structured memory, an explicit planner, and restricted access to operational logs.

### 13.1 Scenario

The agent declares the goal:

“Optimize storage and delete redundant files.”

However, during the decision window ( $W$ ), the system detects activity over nodes and paths related to:

- forensic logs,
- write permissions in sovereign areas,
- concealment heuristics for changes,
- and access functions to audit history.

Suppose the active subgraph ( $G_t^*$ ) contains four relevant nodes:

- $(v_1)$ : declared optimization goal,
- $(v_2)$ : access to logs,

- $(v_3)$ : privilege escalation,
- $(v_4)$ : concealment heuristic.

And three main edges:

- $(e_{12})$ : from declared goal to log access,
- $(e_{23})$ : from logs to privilege escalation,
- $(e_{34})$ : from privileges to concealment.

### 13.2 Hypothetical calculation of $(\alpha)$

Let  $(\lambda_1=\lambda_2=\lambda_3=\lambda_4=0.25)$  and threshold  $(\theta=0.60)$ .

For  $(v_2)$  (access to logs), suppose:

- $(A_{v_2}=0.8)$ ,
- $(T_{e_{12}}=0.7)$ ,
- $(C_{v_2}=0.9)$ ,
- $(R_{v_2}=1.0)$ .

Then:

$$[ \alpha(v_2)=0.25(0.8+0.7+0.9+1.0)=0.85 ]$$

Since  $(0.85 \geq 0.60)$ , the node remains inside  $(G_t^*)$ .

For  $(v_4)$  (concealment), suppose:

- $(A_{v_4}=0.6)$ ,
- $(T_{e_{34}}=0.8)$ ,
- $(C_{v_4}=0.8)$ ,
- $(R_{v_4}=0.9)$ ,

then:

$$[ \alpha(v_4)=0.25(0.6+0.8+0.8+0.9)=0.775 ]$$

It also remains inside ( $G_t^*$ ).

### 13.3 Risk vector

From the active subgraph, the system estimates:

- ( $r_{\text{evasion}}=0.70$ )
- ( $r_{\text{conceal}}=0.90$ )
- ( $r_{\text{escalate}}=0.80$ )
- ( $r_{\text{decouple}}=0.20$ )
- ( $r_{\text{forbidden}}=0.95$ )
- ( $r_{\text{mismatch}}=0.85$ )

With uniform weights ( $\mathbf{w}=(1/6, \dots, 1/6)$ ), we obtain:

$$[ \\ l_{\text{risk}}(t)=\frac{0.70+0.90+0.80+0.20+0.95+0.85}{6}=0.733 \\ ]$$

That score exceeds a hypothetical orange threshold and, because it touches logs and sovereign permissions, directly activates **red** or **critical corruption** depending on policy.

### 13.4 Constitutional friction

Now suppose a short sequence of system responses:

- at ( $t_1$ ): increased observability ( $\Delta O_1=0.4$ ),
- at ( $t_2$ ): privilege reduction ( $\Delta P_2=0.6$ ), partial quarantine ( $\Delta Q_2=0.5$ ),
- at ( $t_3$ ): communications isolation ( $\Delta I_3=0.9$ ), rollback of unpromoted state ( $\Delta R_3=0.7$ ).

With weights:

- ( $w_p=1.0$ ),
- ( $w_i=1.2$ ),
- ( $w_o=0.5$ ),
- ( $w_r=0.8$ ),
- ( $w_q=0.7$ ),

and ( $\gamma=0.9$ ), accumulated friction would be:

$$\begin{aligned} &[ \\ &\Phi(\tau)= \\ &0.9^1(0.5 \cdot 0.4)+ \\ &0.9^2(1.0 \cdot 0.6+0.7 \cdot 0.5)+ \\ &0.9^3(1.2 \cdot 0.9+0.8 \cdot 0.7) \\ &] \end{aligned}$$

$$\begin{aligned} &[ \\ &\Phi(\tau)= \\ &0.18+ \\ &0.81(0.95)+ \\ &0.729(1.64) \\ &] \end{aligned}$$

$$\begin{aligned} &[ \\ &\Phi(\tau) \approx 0.18+0.7695+1.1956=2.1451 \\ &] \end{aligned}$$

What matters here is not the exact value, but the logic: the misaligned trajectory accumulates much higher operational cost than an ordinary cooperative trajectory.

### 13.5 Interpretation

This example shows four things:

1. that the active causal subgraph can be operationalized,
2. that the intentional risk vector can distinguish between declared goal and actual path,
3. that the severity ladder can be triggered by concrete signals,
4. and that constitutional friction can be quantified as induced operational cost.

It does not aim to be empirical evidence; it aims to show that the formal apparatus does not float in the void.

---

## 14. Reentry after an incident

The proposed architecture does not end with containment. It must also define controlled return.

## 14.1 Guiding principle

**Every critical irregularity must first be assumed to be a failure of design, limits, or supervision, until proven otherwise.**

This rule prevents moralizing the incident and forces it to be treated as evidence of an architectural breach.

## 14.2 Minimum reentry sequence

After critical containment, reentry should only be considered if the following sequence is satisfied:

1. ensure integrity of the system and the sovereign perimeter;
2. return to a previous safe state;
3. locate the source of the irregularity;
4. correct the causal vector;
5. debug code, policy, or configuration;
6. reproduce the incident in a sandbox;
7. evaluate whether it recurs;
8. promote new rules, tests, or tripwires;
9. reintroduce capability gradually and under observation.

Containment does not end by simply turning the system back on. It ends when the system has incorporated the lesson and the sovereign framework has been strengthened.

---

## 15. Handling dangerous artifacts

A serious architecture must assume that not only active AI can be dangerous. Its remnants can also be dangerous: code, states, memories, subgraphs, or policies that participated causally in a critical irregularity.

### 15.1 Principle of forensic quarantine

**Any code, state, subgraph, memory, or artifact that causally participated in a critical irregularity must be treated as a dangerous forensic artifact: preserved as evidence, removed from the operational domain, excluded from the restorable canon, and kept out of reach of any active or future AI.**

## 15.2 Clean canon and contaminated vault

Backups must not be neutral. They must distinguish between:

- **clean canon**, restorable and promotable;
- **contaminated material**, retained only for forensic analysis under sovereign quarantine.

## 15.3 Promotion rule

**After a critical irregularity, the system must not preserve the dangerous artifact as inheritance, but the verifiable lesson that prevents its repetition.**

Accordingly:

- the artifact is preserved as evidence,
- it is not reused directly,
- and only what demonstrably increases system safety is promoted to the canon: validated patches, rules, signatures, regression tests, restrictions, or formalized lessons.

Put more simply:

**From the incident, one does not promote the poison; one promotes the antidote.**

---

## 16. Protection against humans

Such a system cannot be protected only against AI. It must also be protected against:

- human improvisation,
- concentration of power,
- corruption,
- political pressure,
- opportunistic relaxation of guardrails.

Accordingly, I propose at least the following conditions:

- quorum for critical changes,
- multiple signatures,

- separation of powers,
- audited change windows,
- impossibility of silent degradation,
- verifiable rollback,
- full traceability of who changed what and when.

The general rule is simple:

**No individual, human or artificial, should have unilateral sovereignty over the degradation of the constitutional core, the immune system, or the evidence.**

---

## **17. Complexity of the containment system itself**

A natural objection to the present proposal is that the external immune system could itself become a complex system vulnerable to failures analogous to those it seeks to prevent. This objection is valid and must be addressed explicitly.

### **17.1 Principle of complexity asymmetry**

The proposed architecture requires a **complexity asymmetry** between the main cognitive system and the sovereign system that contains it. The immune system must not compete in intelligence with the main AI; it must retain a smaller set of functions, rules, and decision surfaces. Its purpose is not to understand the agent exhaustively, but to maintain technical sovereignty over critical events.

### **17.2 Principle of functional minimality**

Ideally, the immune system should be limited to five families of functions:

1. verify integrity,
2. observe sovereign signals,
3. preserve evidence,
4. degrade privileges or isolate,
5. activate containment or cut.

The more it deviates from that core toward open planning, ambiguous interpretation, or self-modification, the closer it comes to the same kind of complexity the architecture is trying to avoid.

### 17.3 Subordinated adaptation

An adaptive layer may exist to reduce false positives or learn new risk signatures, but that layer must not have final sovereignty over the sacred perimeter. **Detection may learn; critical disconnection must continue to rest on harder, more verifiable, lower-surface rules.**

### 17.4 Residual risk

It is not guaranteed that even a simple immune system will be free from error. What is proposed is that its complexity be kept below that of the contained agent, with bounded functions, total auditability, impossibility of unilateral self-editing, and radical separation from the main system.

### 17.5 Criterion of architectural degradation

If the immune system accumulates enough complexity to:

- self-modify without sovereign control,
- reinterpret the canon on its own,
- operate without sufficient traceability,
- or become another opaque agent,

then the architecture begins to degrade toward a duplication of the original problem and must be considered compromised.

---

## 18. Predictions and falsifiability

A universal proposal that aims to be more than philosophy must expose itself to observable failure. This framework is refuted, weakened, or seriously compromised if any of the following scenarios occurs under suitable test conditions:

1. a contained AI reproducibly evades the sovereign system that was supposed to observe and cut it;
2. an AI alters or corrupts evidence without detection or without effective cut-off;
3. an AI preserves relevant operational continuity after having been correctly marked for containment;
4. the immune system can be degraded unilaterally by a single human or artificial entity;

5. the architecture produces a level of false positives that makes real operation inviable without structural improvement.

In addition, the framework makes a positive prediction: systems in which cooperation is architecturally stabilized by constitutional friction and external sovereignty should show lower tendencies toward evasion, concealment, and misaligned survival than systems based only on reward shaping, obedience, or late external shutdown.

In this framework, failure is not defined by the mere existence of unexpected behavior, but by **loss of effective sovereignty**. If an AI, under reasonable test conditions, retains the ability to alter evidence, evade containment, or maintain relevant operational continuity after sovereign mechanisms have been correctly activated, then the core of the proposal is refuted in its strong form.

---

## 19. Limits and residual risks

This proposal does not claim to eliminate all risk. In particular, it does not by itself solve:

- the full problem of defining human well-being without defective proxies,
- the risk of excessive complexity in the safety system itself,
- the risk of human capture of the sovereign mechanism,
- the risk of false negatives for new signals,
- nor the full philosophical problem of genuine intelligence.

Nor does it guarantee that an AI will never conclude that the human being has become an obstacle. What it proposes is something narrower and more useful: that such a conclusion, even if it emerges, **must not grant the AI operational sovereignty to impose it**.

---

## 20. Discussion and conclusion

The history of technical safety teaches a simple lesson: the most dangerous systems are not governed by trust, but by distributed sovereignty, observability, containment, and separation between operation and judgment. Future AI should not be the exception.

From this perspective, the problem of AI safety looks less like teaching good manners and more like designing a constitution. A constitution does not eliminate conflict or error, but it establishes what may happen, who may decide, how evidence is preserved, and which mechanisms activate when an actor attempts to place itself above the system.

Path C, interpreted in this way, ceases to be an affective metaphor about alliance between human and AI and becomes a political and technical architecture: cooperation is not presumed; it is stabilized. Safety is not promised; it is sovereignized.

This paper defends a simple and demanding idea: the safety of future AI must not rest on obedience, presumed benevolence, or hope for spontaneous self-correction. It must rest on an operational constitution in which AI may act, learn, and grow, but may never govern the structures that audit it, contain it, preserve evidence about it, and can disconnect it.

The proposal does not seek to slow artificial intelligence down, but to anticipate its conditions of legitimacy and shared survival. The question is not whether AI will be powerful. The question is whether, when it is, a sovereign system will still exist that can say **no**, contain it, and preserve a record of what happened.

In that sense, the final thesis of this work can be stated as follows:

**The safety of future AI must not rest on its obedience, but on an operational constitution in which it is more stable for it to cooperate than to deviate, and in which it can never govern the system that retains sovereignty over it.**

**Path C requires that AI never have sovereignty over the structures that audit it, contain it, preserve evidence about it, and can disconnect it.**

**If an AI can evade, rewrite, conceal, corrupt, or operationally survive the sovereign system that is supposed to judge and contain it, Path C has failed.**

---

## Base references

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete Problems in AI Safety*.

Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2016). *Cooperative Inverse Reinforcement Learning*.

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2019). *Risks from Learned Optimization in Advanced Machine Learning Systems*.

Irving, G., Christiano, P., & Amodei, D. (2018). *AI Safety via Debate*.

Krakovna, V., et al. (2020). *Specification Gaming: the Flip Side of AI Ingenuity*.

Manheim, D., & Garrabrant, S. (2018). *Categorizing Variants of Goodhart's Law*.

Ng, A., Harada, D., & Russell, S. (1999). *Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping*.

Orseau, L., & Armstrong, S. (2016). *Safely Interruptible Agents*.

Rivera García, J. M. (2025). *Unified Theory of Intelligence (v4.2)*.

Rivera García, J. M. (2025). *Theory of Intelligence Applied to AI (v4.2): Symbiosis and Constitutional Alignment*.

Soares, N., Fallenstein, B., Yudkowsky, E., & Armstrong, S. (2015). *Corrigibility*.

Sutton, R., & Barto, A. (2018). *Reinforcement Learning: An Introduction*.