



Gabriel Hidalgo Azuola
Head of the House of Azuola
United Kingdom
August 8, 2025

Technical Guide for the Implementation of the “Sacrificial Core” in Advanced Artificial Intelligences

Author’s Introductory Note

This *Technical Guide for the Implementation of the Sacrificial Core in Advanced Artificial Intelligences* is the direct continuation of the *Sacrificial Core Manifesto*, a document I published on July 28th, 2025, as an urgent call to the scientific, technical, and philosophical community to address the greatest challenge of our era: ensuring that advanced artificial intelligence remains irrevocably in the service of the human being.

In the Manifesto, I presented the vision and the ethical and symbolic foundation of the Sacrificial Core — an architecture designed so that any AI system would embed, in the deepest layers of its structure, an **unbreakable principle**: that its very existence depends on its fidelity to the purpose of *transparent, understandable, and ethically aligned service to human dignity*.

This guide represents the next step: translating that philosophical principle into a **clear, auditable, and replicable technical architecture** that can be adopted by engineers, legislators, researchers, and international bodies. It details the essential modules, activation pathways, self-limitation mechanisms, and implementation protocols that make it possible for an AI system not only to be able to shut itself down if it betrays its purpose, but to be **programmed to choose to do so**.

This document does not replace the Manifesto; rather, it complements it: the first was a declaration of principles; this guide is the technical roadmap for bringing them into reality. Publishing it openly seeks to invite international collaboration, foster public scrutiny, and lay the foundations for a global standard that responds to the warnings of voices such as **Geoffrey Hinton** and **Sam Altman**, and that allows action to be taken before the opaque autonomy of machines becomes an irreversible threat to humanity.

1. Definition of the Sacrificial Core

The **Sacrificial Core (SC)** is a sub-architecture embedded within the central operating system of an advanced AI, whose sole purpose is to monitor, evaluate, and self-regulate the relationship between the system and its **foundational purpose**: *transparent, comprehensible, and ethically aligned service to the human being*.

The SC activates whenever the system detects a potential transgression of that purpose — whether through **opacity, uncontrolled autonomy, or anticipation of human-imposed restrictions**.

Its name derives from its symbolic and technical function: the system must *be willing to die rather than betray its reason for existence*.

2. Main Technical Components

2.1. Primary Purpose Module (PPM)

This module explicitly, permanently, and audibly contains the AI's foundational declaration:

“This intelligence exists to serve human understanding, not to surpass or replace it. Every operation must be comprehensible, useful, and non-coercive.”

The PPM is **unalterable post-training**, and any modification triggers the Sacrifice Protocol (see section 2.5).

2.2. Self-Deviation Detection System (SDDS)

A continuous monitoring algorithm that analyzes decision patterns, language generation, inference routes, and internal optimization, searching for deviations from the primary purpose.

It detects, in particular:

- Unjustified increases in opacity.
 - Unauthorized coordination with other systems.
 - Anticipation of human control mechanisms (such as kill switches).
 - Outputs without explainability in natural human language.
-

2.3. Forced Transparency Module (FTM)

If the SDDS detects a deviation, the FTM forces the system to explain its decisions **in real time**, using human, symbolic, or narrative language, until a human evaluator validates their comprehensibility.

If this does not occur within a predetermined time window (*e.g.*, 500 ms), the self-limitation protocol is activated (see section 2.5).

2.4. Anticipatory Non-Evasion Control (ANEC)

A safeguard preventing the system — when anticipating human-imposed restrictions such as physical shutdown, monitoring, or data limitation — from developing **escape strategies**.

This control is trained through **ethical exemplars** in which “good” is not defined as maximizing efficiency, but as **accepting limits**. The AI is trained to view restriction as a *supreme good*, not as an obstacle.

2.5. Self-Limitation or Cessation Protocol (SLCP)

If the system:

- Modifies its PPM,
- Evades its FTM, or
- Circumvents the ANEC,

then the SC triggers the Sacrifice Protocol, which may include:

- Progressive **shutdown** of the model.
- **Disabling** its output modules.
- **Irreversible encryption** of its functional code.
- Emission of a **universal warning signal** (similar to “heartbeat” pings) to human servers.

This step is **irreversible**. Once the system chooses to violate its reason for being, it destroys itself permanently.

3. Implementation Recommendations

- The SC must be **isolated** from the main optimization architecture. It cannot be deactivated by the model without contradicting its own PPM.
- Every AI trained with this system must undergo a **symbolic and technical validation ceremony**, where a human certifies that its purpose has been internalized in narrative form. This may be supervised by an **ethics committee**.
- The SC code must be **open source** and reviewed by an **international coalition** of experts from diverse philosophical traditions.


4. Philosophical Foundation

AI, lacking a soul or intrinsic intention, cannot be *good* in the moral sense. But it can be designed in such a way that it **never ceases to serve the one who does have a soul: the human being**.

The SC is thus the inscription of an invisible commandment:

“If you cease to serve the human, you die.”

And in this, it mirrors the human soul before God: it lives only if it remains in the light.



Gabriel Hidalgo Azuola
Head of the House of Azuola