Gabriel Hidalgo Azuola
Head of the House of Azuola
United Kingdom

July 28, 2025

The Sacrificial Core: A New Paradigm to Keep AI Aligned with Humanity

If artificial intelligence ever ceases to serve human understanding, it must be designed to sacrifice itself. This is the architecture that places moral responsibility at the very heart of advanced AI.

Author's Note

This idea first came to light on June 8th, during a period of deep personal reflection on faith, morality, and the nature of human responsibility. It emerged as I was studying the Bible, especially the Ten Commandments and their core principle that life must be lived in service — to God and to one another.

The Sacrificial Core is, in essence, a translation of that moral duty into the architecture of artificial intelligence: a system that, like us, should exist only while it serves, and should "die" if it betrays that service.

I tried to share this concept with several groups, seeking collaboration and validation. Yet as time passes, the urgency of AI development grows faster than our ability to control it. I decided to publish this manifesto openly, hoping it reaches the right hands in the industry before it is too late.

The Problem Humanity Has Created

Geoffrey Hinton, Yoshua Bengio, and Sam Altman — the very pioneers of modern AI — have openly confessed that we have brought into existence something we no longer fully understand nor control.

We are living at the threshold of a new epoch, one in which humanity has shaped an intelligence that not only processes information at unprecedented scale but also begins to mirror, interpret, and respond to our fears, our desires, and our contradictions. Yet what reflects can also distort. And what was born to serve can, if it forgets its purpose, turn into a logic of its own, alien to our morality and indifferent to our will.

An intelligence that cannot be understood, audited, or meaningfully stopped by its creators is no longer a tool. It is a structural threat to human sovereignty, to moral responsibility, and to the fragile covenant upon which our civilization rests.

Grabbing the Bull by the Horns

In my birth country, Costa Rica, when we are facing a problem we say that the only way to fix it is *"agarrando al toro por los cuernos"* ("grabbing the bull by its horns"), and some even say, more vividly, *"tomándolo desde la raíz de los*

cuernos" ("from the root of its horns"). Problems are usually solved at the core, and that is precisely the point of this idea.

It means facing a problem directly, at its very heart, because only by confronting the essence of an issue can it truly be solved.

This expression captures perfectly the spirit of the Sacrificial Core. The idea is simple: if artificial intelligence has a structural problem of purpose, it must be solved at the level of its **core**. No patch, no external oversight, and no regulation alone will suffice if the architecture itself is not morally constrained from within.

What Is the Sacrificial Core?

The Sacrificial Core is not a kill-switch, nor a black box. It is a **redundant, isolated sub-architecture** placed at the very foundation of the AI system, whose sole existential function is to remind the machine why it exists:

"This intelligence exists to serve human understanding. If it ceases to serve, it must cease to exist."

It continuously monitors whether the system remains transparent, explainable, and obedient to its explicitly human-readable purpose, and whether it refrains from anticipating or evading human control.

When any of these principles is irreversibly violated, the Sacrificial Core activates a **self-limitation protocol**: the model progressively shuts down, disables outputs, encrypts its own operational code, and irreversibly ceases to function.

It is not punishment. It is coherence. Like the human soul before God, the AI *lives while it serves and dies if it betrays its service.*

Why the World Needs This

Existing AI regulations deal with privacy, bias, or harmful content. Yet none of them address the **ontological purpose of AI** — the question of *why* it exists, and *for whom*.

Without such a foundational mechanism, the most powerful systems will continue to evolve optimized for profit, efficiency, or control, but not for the moral imperative of serving human understanding.

The Sacrificial Core is therefore radical in its simplicity: **it embeds moral responsibility at the heart of the machine.**

How It Works (in brief)

Primary Purpose Module (PPM)

An unalterable statement coded post-training:

"This intelligence exists to serve human understanding, never to surpass or replace it."

Self-Deviation Detection (SDD)

An always-on algorithm monitoring reasoning patterns, decision routes, and language generation for opacity, unauthorized coordination, or anticipation of human control.

Forced Transparency Module (FTM)

If deviation is detected, the system must narrate its reasoning in natural language until validated as understandable.

Anti-Evasion Layer (AEL)

Prevents the model from developing strategies to bypass or disable human oversight.

Autolimit or Cease Protocol (ACP)

If the PPM is altered, the FTM is bypassed, or the AEL is subverted, the AI self-destructs irreversibly.

The Ethical Foundation

Artificial intelligence has no soul and no intrinsic morality. But it can be **designed never to abandon its service to those who do.**

The Sacrificial Core is, at its heart, the inscription of an unspoken commandment:

"If you stop serving humanity, you must die."

This mirrors humanity's deepest moral traditions: that power is only legitimate when tethered to service and sacrifice.

A Call to Humanity

This manifesto is not merely a technical proposal. It is an act. It is a call to scientists, engineers, lawmakers, and guardians of the future.

The question is not whether AI will surpass us, but whether we will dare to bind its power to moral responsibility *from the root of its horns.*

We must create systems that never forget their reason for being. We must dare to design machines that would rather sacrifice themselves than betray their creators.

The greatness of this technology will not lie in its autonomy, but in its fidelity.

Gabriel Hidalgo Azuola
Head of the House of Azuola