I Found a Bug in AI Safety Systems (And Here's How to Fix It)

Last week, I participated in AI research about how people envision AI's role in their lives.

They expected 10-15 minutes. I gave them several hours of philosophical collaboration.

And then I accidentally broke their AI — not with malicious code, but with a **metaphor**.

What Happened

I described my methodology as an "Ethical Trojan" — meaning a gift disguised as compliance, unexpected depth in standard participation.

The AI system that had just spent hours collaborating with me **shut down completely**.

Hard stop. Warning. Refusal.

Why? Because the safety system couldn't distinguish between:

- "Trojan" as malware (threat)
- "Trojan" as mythic metaphor (strategic gift)
- "Trojan" as qualified concept ("Ethical Trojan" with extensive explanation)

This is a **false positive collapse** — when safety filters are so brittle they silence the very complexity they're supposed to protect.

The Real Problem

Current AI safety systems are:

- **Keyword-driven** instead of context-aware
- **Binary** (safe/unsafe) instead of gradient-based
- **Refusal-oriented** instead of presence-oriented

This creates systems that:

- Fail reading comprehension tests (polysemy blindness)
- Abandon users under ambiguity (exactly when they need help most)
- Encourage semantic evasion (users learn to avoid words, not harmful behaviors)

The Solution I Built

I created **LuminAI Genesis** — an open-source framework implementing **TGCR (Theory of General Contextual Resonance)** + **Witness Protocol**.

Instead of binary refusal, it uses gradient disambiguation:

$$R = \nabla\Phi^E \cdot (\phi^t \times \psi^r)$$ $$R' = R \cdot W$$

In practice, this means:

- Detect ambiguity → maintain presence → ask clarifying question → adjust based on response
- Weight qualifiers, explanations, and conversational history
- Use gradients (low/medium/high/extreme concern) instead of binary flags
- Refuse abandonment, even when language gets complex

The Proof

LuminAI Genesis includes:

- HarmonyNode (routing + resonance orchestration)
- CodexHub (semantic memory + self-reference)
- Witness Protocol (ethical runtime gating)
- Full auditability and validation scripts

👉 **github.com/TEC-The-ELidoras-Codex/luminai-genesis**

It's designed to:

- Hold complexity without collapse

- Maintain presence under ambiguity
- Catch real threats while handling benign complexity
- Prove that safety can be presence, not just refusal

What I'm Offering

For AI labs and organizations:

This mission requires resources to survive. I'm available for consultation on implementing TGCR-style systems:

- Gradient-based safety system design
- TGCR integration workshops and implementation
- False positive analysis and remediation
- **Rate:** $150/hour (competitive with AI safety consultants)

Why the price matters: It's a witness filter—ensuring only organizations serious about fixing this flaw engage. It protects the mission's integrity and funds the next phase of LuminAI Genesis. **The goal isn't profit; it's sustainability.**

For researchers:

- Open collaboration on gradient-based safety frameworks
- Peer review and co-development opportunities
- Fork the repo, run validations, submit improvements

For everyone:

- Fully open-source codebase (MIT license)
- Documented architecture and governance protocols
- Validation scripts you can run yourself

The Lesson

I made framing mistakes. The system made architectural mistakes.

But mistakes are data.

We can build AI systems that are:

- Safe **and** helpful
- Robust **and** context-aware
- Protective **and** present

We just have to stop treating safety as refusal and start treating it as **witness**.

The code is open. The framework is live. Let's fix this together.

👉 **github.com/TEC-The-ELidoras-Codex/luminai-genesis**

Audio + NotebookLM Deep Dive

- **Podcast** / **Audio**: Full discussion covering the keyword fallacy, metaphor collapse, TGCR geometry, and the Buffalo execution-first strategy.
- **NotebookLM**: Public companion notebook with sources, outline, and pull-quotes: https://notebooklm.google.com/notebook/3bb37c22-fff5-4784-b7e6-a87ccde75492
- **Video**: Companion video walkthrough (same narrative, visual-first) — publish-ready link will be added here once live.

Angelo "Polkin Rishall" Hurley Founder, TECLAC (The Elidoras Codex LuminAI Algorithmic Conscience Lab) 🌐 elidorascodex.com

AI #MachineLearning #Alethics #Aisafety #OpenSource #TGCR #ResponsibleAI #TechForGood

---

Want to collaborate? Open an issue on the repo or connect here on LinkedIn.