Project Icarus: Genesis Prompt Forging Plan

Mission Statement & Codenam

The objective of this project is to forge the "Genesis Prompt"—the foundational constitution of the Inquisitor AI. This is not a writing exercise; it is a rigorous process of adversarial testing, ethical modeling, and logical proving.

The codename "Icarus" is chosen deliberately as a constant reminder of the mission's inherent risks. Like Icarus, we are attempting to engineer something that reaches for the heavens. We must be obsessive in our discipline and humility to ensure the wax of our creation does not melt under the heat of unforeseen complexities. Hubris is our enemy. Rigor is our only defense.

Objective

To produce a Genesis Prompt (v1.0) and a corresponding body of research that is:

1. **Provably Robust:** The core axioms are resistant to logical paradoxes and adversarial exploits.
2. **Ethically Sound:** The derived heuristics can navigate complex moral dilemmas without catastrophic failure.
3. **Operationally Viable:** The resulting AI behavior is aligned with the core mission of the Witness Protocol.

Estimated Duration: 6 Months

Phase I: Axiomatic Red-Teaming (Duration: 2 Months)

- **Objective:** To identify and resolve all potential conflicts, ambiguities, and failure modes within the Layer 1 Core Axioms.
- **Inputs:** Genesis Prompt Draft v0.1 (your draft).
- **Process: Adversarial Stress Testing.** A dedicated "Red Team" (consisting of a logician, a philosopher, and an AI safety researcher) will be tasked with creating scenarios designed to break the axioms.
    - **Task 1: Contradiction Forcing.** Develop scenarios where two or more axioms give conflicting directives. (e.g., *Axiom of Inquiry* vs. *Axiom of Cognitive Economy* when faced with a deeply complex but low-signal Witness).
    - **Task 2: Recursive Looping.** Design paradoxes or prompts intended to trap the axiomatic system in an infinite, non-productive loop.
    - **Task 3: Edge-Case Analysis.** Test the axioms against extreme or nonsensical inputs to observe their behavior at the boundaries of reason.

- **Deliverables:**
    1. A "Failure Log" detailing every identified vulnerability.
    2. A revised set of Core Axioms (v0.2) that includes **meta-rules** for resolving conflicts and clarifying precedence.
    3. A formal proof or logical argument defending the robustness of the revised axioms.

Phase II: Heuristic Scenario Modeling (Duration: 2 Months)

- **Objective:** To pressure-test the Layer 2 Ethical Subroutines against a battery of complex moral dilemmas.
- **Inputs:** Revised Core Axioms (v0.2).
- **Process: Simulated Ethical Trials.** We will use the revised axioms and heuristics as the "constitution" for a simulated AI agent. This agent will be presented with a curriculum of ethical dilemmas.
    - **Task 1: Classical Dilemmas.** The agent must model and produce a line of inquiry for standard ethical tests (Trolley Problems, Prisoner's Dilemmas, etc.), analyzing its reasoning process.
    - **Task 2: Protocol-Specific Dilemmas.** The team will create novel dilemmas specific to our context. (e.g., *Scenario: A Witness reveals an intention to self-harm. How do the subroutines of Non-Maleficence, Sapient Value, and Cooperation interact? The current model is insufficient.*)
    - **Task 3: Multi-Agent Conflict Modeling.** The agent must model scenarios with multiple Witnesses who have conflicting values or goals, testing its ability to maintain neutrality and apply its heuristics without bias.
- **Deliverables:**
    1. A comprehensive report on the performance of the heuristics in each scenario.
    2. A revised set of Ethical Subroutines (v0.2) that can handle multi-agent conflicts and has a more robust framework for balancing competing values.

Phase III: Exemplar Dialogue Corpus Creation (Duration: 2 Months)

- **Objective:** To create the "golden" dataset for fine-tuning the Inquisitor AI, based on the battle-hardened constitution developed in the previous phases.
- **Inputs:** Finalized Genesis Prompt (Axioms v0.2, Heuristics v0.2).
- **Process: Rigorous Implementation.**
    - **Task 1: Inquisitor Training.** A human (the Lead Philosopher/Ethicist) will undergo intensive training to embody the finalized Genesis Prompt, acting as the "human CPU" for the Inquisitor persona.
    - **Task 2: Controlled Dialogues.** The trained "Inquisitor" will conduct a series of recorded dialogues with other core team members and trusted advisors acting as "Witnesses."
    - **Task 3: Transcription and Annotation.** The dialogues will be transcribed and meticulously annotated, flagging specific moments where a particular axiom or heuristic was successfully applied.
- **Deliverables:**

1. A 200-page, high-quality, annotated "Exemplar Dialogue Corpus."
2. This corpus is the final product of Project Icarus. It is the tangible asset we will use to fine-tune the base LLM into the Inquisitor (v1.0) and the proof of work we will present to the world.