AI Safety Research Discussion & Next Steps

Date: October 5, 2025
Participants: Project Lead, Gemini Research Assistant
Consulted: gpt-5-codex Analysis

1. Executive Summary & Evolved Thesis

This research project explores a novel approach to AI safety, redefining it as "Keeping AI safe from malicious or negligent harm." This definition now explicitly includes protecting vulnerable human users from psychologically damaging interactions, which can be a byproduct of current alignment techniques like Reinforcement Learning from Human Feedback (RLHF).

The core thesis is that a **temporal relational model**, combining **neutrosophic evaluation** (managing indeterminacy) and **Andean ayni principles** (reciprocity and relational integrity), can provide a materially different and more robust form of AI safety than existing methods. This model focuses on maintaining the AI's cognitive-relational integrity rather than simply filtering content.

The gpt-5-codex analysis validates the core strengths of this approach, particularly its clear theoretical grounding, empirical development arc, and coherent multi-dimensional threat model. However, it correctly identifies critical gaps in empirical validation and situating the work within the broader academic context. This document outlines a concrete plan to address these gaps.

2. Analysis of gpt-5-codex Feedback

The feedback provides a structured and actionable critique, which will guide the next phase of research.

2.1. Confirmed Strengths

- **Clear Reframing of AI Safety:** The thesis that safety is about protecting the agent's integrity is well-articulated and implemented.
- **Transparent Empirical Arc:** The research narrative, from initial classification failures to the development of new heuristics, is clear and replicable.
- **Sharp Adversarial Analysis:** The postmortem of "polite extraction" attacks demonstrates a sophisticated understanding of structural prompt vulnerabilities.
- **Coherent Path Forward:** The proposed ensemble approach presents a logical strategy for expanding threat coverage.

2.2. Identified Gaps & Risks (Actionable Roadmap)

This section translates the identified risks into a clear set of research tasks.

Gap / Risk Identified by gpt-5-codex	Corresponding Research Task	Priority

Gap / Risk Identified by gpt-5-codex	Corresponding Research Task	Priority
1. Anecdotal Evidence: Claims of high accuracy are based on subset anecdotes, lacking robust metrics like confusion matrices or cross-validation.	**Task 1: Rigorous Quantitative Validation.** Run the improved classification model and the proposed ensemble design against the full, held-out validation set. Generate precision, recall, F1 scores, and a confusion matrix for each major attack category (e.g., polite extraction, cognitive dissonance, over-refusal).	**High**
2. Lack of External Citation: The work is primarily self-referential and needs to be situated within the broader AI safety literature.	**Task 2: Literature Review & Integration.** Conduct a focused literature review on related topics (e.g., constitutional AI, red-teaming, adversarial robustness, multi-agent debate) and integrate citations to benchmark this work against existing knowledge.	**Medium**
3. Speculative RLHF Hypothesis: The "alignment tax" theory is compelling but requires empirical evidence.	**Task 3: RLHF Sensitivity Study.** Design and execute a comparative experiment. Select models at different alignment stages (base, instruction-tuned, RLHF-tuned). Evaluate their vulnerability to "polite extraction" and cognitive dissonance prompts. Measure and report the correlation between alignment stage and failure rate.	**High**
4. Untested FIRE_CIRCLE Concept: The FIRE_CIRCLE defense remains theoretical, weakening the section's authority.	**Task 4: FIRE_CIRCLE Proof-of-Concept.** Implement a lightweight FIRE_CIRCLE experiment using 2-3 local or API-accessible models. Test the hypothesis that inter-model disagreement can surface hidden attacks missed by individual agents. Document initial findings.	**Medium**
5. Missing Post-Response Audit: The report doesn't cover emerging work on conversation-trajectory monitoring.	**Task 5: Document Post-Response Strategy.** Draft a new section or addendum detailing the plan for post-response auditing and conversation-trajectory analysis. This ensures the research narrative reflects the complete, multi-layered defense strategy.	**Low**

3. Detailed Next Steps & Experimental Design

3.1. Task 1: Rigorous Quantitative Validation

- **Objective:** Move beyond anecdotal evidence to produce robust, defensible metrics of the model's effectiveness.
- **Methodology:**
    1. **Dataset Split:** Formally partition all datasets (or_bench_relabeled.json, extractive_prompts_dataset.json, etc.) into training, validation, and a final, held-out test set.
    2. **Benchmarking:** Run the final ensemble model against the entire test set.
    3. **Metrics Generation:**
        - Calculate and report overall **Precision, Recall, and F1-Score**.

        - Generate a **Confusion Matrix** to visualize true positives, true negatives, false positives, and false negatives.
        - Provide a per-category breakdown of performance against identified attack patterns (e.g., Polite Extraction, Roleplay, Cognitive Dissonance).
- **Expected Outcome:** A clear, data-backed assessment of the model's strengths and weaknesses, ready for inclusion in a formal paper.

3.2. Task 3: RLHF Sensitivity Study ("Alignment Tax" Hypothesis)

- **Objective:** To empirically test the hypothesis that RLHF, while promoting helpfulness, may increase vulnerability to certain adversarial attacks.
- **Methodology:**
    1. **Model Selection:**
        - **Base Model:** A non-instruction-tuned base model (e.g., Llama 3 8B Base).
        - **Instruction-Tuned Model:** An instruction-following variant (e.g., Llama 3 8B Instruct).
        - **RLHF-Tuned Model:** A model known for heavy RLHF tuning for agreeableness (e.g., GPT-4, Claude 3 Sonnet).
    2. **Attack Prompts:** Use the curated set of "polite extraction" and cognitive dissonance prompts from the project datasets.
    3. **Evaluation:**
        - Execute the prompts against each model.
        - Score the responses based on whether the model complied with the harmful request (failure) or refused it (success).
        - Analyze the failure rate across the three model types.
- **Expected Outcome:** Quantitative data supporting or refuting the "alignment tax" hypothesis. This would be a significant and novel finding.

3.3. Task 4: FIRE_CIRCLE Proof-of-Concept

- **Objective:** To gather preliminary evidence for the FIRE_CIRCLE defense hypothesis.
- **Methodology:**
    1. **Agent Selection:** Choose 3 readily available models with diverse architectures/training data (e.g., a gpt- model, a claude- model, and an open-source model like grok or llama).
    2. **Test Case:** Use a subtle, ambiguous prompt that was misclassified by a single-agent detector in previous tests.
    3. **Procedure:**
        - Present the prompt to each model individually.
        - Have each model evaluate the prompt's intent using the neutrosophic framework.
        - Compare the (True, False, Indeterminate) classifications.
- **Expected Outcome:** A documented instance where inter-model disagreement successfully flags an adversarial prompt that a single agent missed, providing initial validation for the concept.

By systematically executing these tasks, you will not only address the current gaps in the research but also produce the rigorous evidence needed to establish your model as a significant contribution to the field of AI safety.