Project Proposal

Title:
Agent Reality Map: Visualizing Agent Reasoning Pathways for Oversight and Alignment Research

Applicant:
Shreepad Avhad

---

1. Motivation

AI agents increasingly make autonomous decisions by chaining together internal reasoning steps ("thoughts") and external tool calls. Today, these internal steps are largely opaque making it difficult for researchers to diagnose misalignment, deceptive reasoning, or unintended optimization.

This project proposes a lightweight, open-source tool to **log and visualize the reasoning pathways of AI agents**. By exposing these hidden steps in an interpretable format, we can give researchers and developers the ability to audit, debug, and study agent behavior more effectively.

---

2. Research Question

- *How can we capture and visualize agent decision pathways in a structured, interpretable way that supports alignment research?*
- Sub-questions:
    - What metadata (thoughts, actions, observations, costs) are most critical for oversight?
    - Can structured logging + visualization highlight misalignment or deceptive reasoning patterns?

---

3. Proposed Approach

- **Phase 1 (MVP):**
    - Build a logging interceptor that records every agent step (thought, action, observation) into structured JSONL format.

    - Build a simple visualization interface (node-link graph using D3.js) to display these traces interactively.
- **Phase 2 (Extension):**
    - Swap in a real LLM agent (e.g., ReAct loop) instead of simulated agents.
    - Log richer metadata (token usage, latency, confidence scores).
    - Explore heuristics to automatically flag suspicious or repetitive reasoning.

---

4. Expected Outcomes

- An open-source repo (**Agent Reality Map**) that anyone in the alignment community can use.
- Demo showing real agent runs visualized as step-by-step reasoning maps.
- Documentation of observed patterns e.g., how often agent's "shortcut" reasoning or optimize incorrectly.

---

5. Relevance to Alignment

Alignment is fundamentally about ensuring AI does what we intend, not just what we instruct. By exposing **how an agent reasons step by step**, researchers gain new tools for:

- Studying misalignment in toy environments.
- Auditing failure cases where agents behave deceptively.
- Building interpretability infrastructure for more advanced oversight.

This proposal aims to contribute a **practical interpretability tool** to the alignment ecosystem, bridging software engineering with safety research.

---

6. Deliverables

- Open-source codebase with logging + visualization.
- A 90-sec demo video walking through one agent run.
- A short technical report describing the system and lessons learned.