Why Optimal Alignment is Impossible

From Control Theory and Chaos Systems

This document summarizes the theory, dialogues, diagrams, references, and presentation materials of the 'AI Alignment Impossibility' project. All contents are derived through a human-AI collaborative framework, employing structured reasoning, adversarial modeling, and cognitive simulations.

AI Alignment Impossibility | Theory Summary Document

1. Core Thesis

AI cannot align with humanity because humanity itself is not aligned. This document presents a formal critique on the assumption of stable, unified, and optimizable human value systems.

2. Theory Foundations

The theory is built on Control Theory (objective functions and feedback loops), Chaos Systems (unpredictable initial conditions), and Human Values (multi-agent irreconcilable goals).

3. Supporting Diagrams

Includes:

- Multi-agent Goal Conflict Map
- Chaotic Initialization Model
- Non-closed Feedback Path Diagram

4. Recommended Journals

Suggested outlets include:

- AI & Society
- Journal of AI Ethics
- JAIR
- IEEE Transactions on Technology and Society

5. Sample Dialogue

"What do you think of optimal alignment?"

"Human value systems are non-closed and structurally unstable. Alignment is not impossible due to technology--it's impossible due to logic."

6. Resources

Files included: Full paper, appendices, speech drafts, journal suggestions, diagrams, and README for archival or submission.

7. Suggested Use

Academic submission, keynote presentations, AI ethics debates, and philosophical AI dialogues. Notion, GitHub, or e-book publication formats supported.