# Discretion Under Ambiguity: Stress-Testing Model Specifications for Safer AI Alignment

**Mentor: Claudio Mayrink Verdun**
Harvard University
claudioverdun@seas.harvard.edu

## Description of the Project

One of the central challenges in aligning advanced AI systems is the ambiguity and inconsistency embedded in the *model specifications* and *principle-based alignment methods* that guide their behavior. Recent research has shown that alignment principles—whether used for RLHF, Constitutional AI, or spec-driven alignment—often contain internal contradictions, gaps, or ambiguous tradeoffs. These issues lead to unstable and inconsistent supervision signals during training and evaluation. Two works strongly motivate this project:

- **AI Alignment at Your Discretion** (Published at ACM FAccT and Best Paper Award, New England NLP Workshop) [1] introduces the notion of *alignment discretion*, showing that annotators and algorithms routinely face conflicting or underspecified principles. As a result, they exercise substantial, often unexamined discretion in ranking model outputs. This work demonstrates high levels of principle conflict, indifference, and arbitrary annotation behavior, as well as divergent principle prioritization patterns across modern LLMs.
- A recent paper by **Anthropic and Thinking Machines** [2] demonstrates that stress-testing model specifications with value-tradeoff scenarios reveals extensive contradictions, interpretive ambiguities, false-positive safety refusals, and disagreement even among models sharing the same specification.

Both works show that alignment principles and model specifications, which are meant to impose structure and predictability, often create new ambiguity that propagates into misaligned or inconsistent model behavior.

## Project Goal

This project will study **discretion and safety** in AI alignment by examining how ambiguities in model specifications and principles produce:

1. Conflicting or unstable supervision signals during RLHF/RLAIF,
2. Inconsistent or unpredictable principle prioritization in trained models, and
3. Divergent safety behaviors, including overly conservative refusals, contradictory reasoning, or miscalibrated risk judgments.

Building on both papers, the project will develop:

- A framework for **detecting and quantifying ambiguity** in model specifications,
- A methodology for measuring **spec discretion**: the behavioral latitude allowed by a given specification,
- Cross-model analyses of how models negotiate conflicting principles,
- Tools for generating **stress-test scenarios** that expose inconsistent or contradictory behaviors.

Ultimately, we aim to answer:

- When do model specifications create unavoidable ambiguity?
- How does this ambiguity propagate into model behavior during training?
- Can we design metrics that identify inconsistencies or contradictions in supervision signals?
- How should future model specifications be written to minimize unnecessary discretion?

### How This Project Advances AI Safety

Ambiguity in alignment specifications is an under-recognized source of alignment failures. If principles conflict or are underspecified, models must choose arbitrarily among interpretations, leading to unpredictable or inconsistent behavior.

This project contributes to AI safety by:

- Revealing hidden failure modes in principle-based and spec-driven alignment,
- Providing diagnostic tools for identifying contradictions and unclear guidance in specifications,
- Characterizing how different models interpret and prioritize principles,
- Stress-testing current alignment systems to uncover safety gaps,
- Helping design clearer, more enforceable specifications with fewer failure modes.

### What Role Will Mentees Play?

Mentees will contribute directly to the research, with responsibilities including:

- Implementing scenario-generation and stress-testing pipelines,
- Running large-scale model evaluations on principle-conflict scenarios,
- Developing quantitative metrics to detect contradictory supervision signals,
- Visualizing behavioral divergence across models and principles,
- Co-authoring a workshop or conference submission based on the project results.

Mentees will be encouraged to propose hypotheses, run experiments, and explore unexpected patterns in model behavior.

## Mentorship Approach

**Weekly Guidance:** Approximately 1 hour per week per mentee.

Meetings will focus on evaluating results, refining experimental design, and developing conceptual clarity. Mentees will be encouraged to follow their curiosity and propose creative directions; historically, the most interesting results emerge from unexpected observations.

## Prerequisites for Mentees

### Must-Have

- Strong Python proficiency,
- Experience working with large language models (API or local inference),
- Familiarity with basic RLHF or instruction-tuning concepts.

### Nice-to-Have

- Statistical analysis and hypothesis testing,
- Data visualization skills (Plotly, Matplotlib),
- Research experience,
- Some background or interest in ethics, philosophy, or law (helpful but not required).

## Logistics

**Timezone:** Afternoon or evening Boston time.
**Expected Weekly Commitment:** 10 hours.

## References


[1] Maarten Buyl, Hadi Khalaf, Claudio Mayrink Verdun, Lucas Monteiro Paes, Caio Cesar Vieira Machado, and Flavio du Pin Calmon. Ai alignment at your discretion. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 3046–3074, 2025.

[2] Jifan Zhang, Henry Sleight, Andi Peng, John Schulman, and Esin Durmus. Stress-testing model specs reveals character differences among language models. *arXiv preprint arXiv:2510.07686*, 2025.