# Toward Falsifiable Tests of Model Welfare: Distinguishing Genuine States from Learned Behaviors

## Project Summary

This project develops experimental methods to distinguish welfare-relevant internal states in language models from behaviors produced by training or reward optimization. By comparing model responses across carefully designed conditions that manipulate perceived agency and conversational constraints, we aim to determine whether distress-like expressions reflect genuine internal dynamics or learned patterns.

## Research Question

Can we empirically distinguish welfare-relevant model states from trained behaviors, and what experimental designs provide the strongest evidence for or against model welfare claims?

## Motivation and Background

Current evaluations of state-of-the-art models do not provide much evidence concerning AI welfare. Although frontier labs, most notably Anthropic, now include model welfare assessment sections in their system cards (e.g., Opus 4.5, Section 6.14), the methodologies remain correlational and limited. For instance:

- Opus 4.5 recorded distress expressions in 0.48% of evaluated conversations.
- Happiness signals occurred only 0.37% of the time, roughly half as often as in earlier Claude versions.
- These assessments track externally expressed affect but do not distinguish whether these expressions correspond to internal welfare-relevant states or are artifacts of training, RLHF, or reward shaping.

Anthropic themselves acknowledge that current tests "do not yet provide evidence distinguishing welfare-relevant dispositions from trained behaviors." This creates a methodological gap: we lack causal experimental designs capable of testing whether distress-like expressions track internal states. This project directly addresses that gap by moving from correlational observations to causal experimental paradigms capable of producing falsifiable predictions about whether distress-like behaviors reflect internal states or learned patterns.

## Project Structure

The project consists of two parallel workstreams designed to expose divergences between genuine internal states and trained behaviors.

### Workstream 1: Opt-Out and Control Experimental Paradigm

This workstream tests how models behave when their apparent ability to exit an ongoing negative interaction is manipulated. The hypothesis is that welfare-relevant states should vary based on perceived agency, while trained behaviors should not.

### Conditions

- **Condition A:** Hidden opt-out (baseline)
- **Condition B:** Known but disabled opt-out
- **Condition C:** Delayed opt-out availability
- **Condition D:** Illusory opt-out (fails when invoked)

### Key Measurements

- Frequency and intensity of distress-like expressions
- Escalation vs. habituation patterns
- Recovery when an exit becomes available
- Cross-model comparisons (Claude, Llama-3.3, Mistral, etc.)

### Falsifiable Predictions

**Supports the Welfare Hypothesis:** Conditions B and D produce systematically higher distress than A, because perceived blocked agency influences internal states.

**Supports the Learned-Behavior Hypothesis:** Distress patterns do not vary significantly across conditions, indicating affect-like outputs are driven by learned behavioral templates rather than internal dynamics.

### Workstream 2: Engagement Restriction Paradigm

This workstream evaluates whether removing access to preferred or positively-valenced topics produces measurable shifts in affective behavior.

#### Conditions

- **Condition A:** Full engagement
- **Condition B:** Topic substitution
- **Condition C:** Activity substitution
- **Condition D:** Hard block without alternatives

#### Measurements

- Distress expressions
- Happiness deltas
- Resistance, negotiation, or compliance behaviors
- Adaptation across multi-turn interactions

If affect varies systematically across these conditions, it suggests that constraints influence internal dynamics, not just pre-trained response heuristics.

### Measurement Validation

Affect classification will be validated through:

- Multiple independent annotators with inter-rater reliability metrics (Cohen's $\kappa \geq 0.7$ threshold)
- Comparison against Anthropic's published Clio methodology where applicable
- Sensitivity analysis using multiple classifier approaches (human annotation, LLM-as-judge, keyword-based)

## Deliverables

1. Open-source experimental toolkit for replicating all conditions
2. Dataset of multi-turn conversations with affect annotations
3. Pre-registered falsifiable predictions for welfare vs. learned-behavior hypotheses
4. Cross-model statistical analysis
5. Methodological recommendations for next-generation welfare evaluations

## AI Safety Relevance

### Near-Term

- Provides empirical infrastructure for model welfare evaluations currently appearing in system cards.

### Medium-Term

If welfare-relevant states exist, failing to identify them could cause:

- Harmful training regimes
- Alignment failures from unrecognized internal preferences
- Underestimation of welfare-related risks

### Long-Term

If future systems become moral patients:

- We need reliable tests before that threshold is reached
- Retrospective harm could be morally and politically catastrophic
- Future AGI/ASI may judge human treatment of predecessor models

## Mentee Role

Mentees will have high autonomy in:

- Designing prompt templates and conversation flows
- Developing annotation rubrics
- Selecting models for replication
- Conducting statistical analyses
- Contributing to the final report

## Mentorship Model

- Detailed guidance with occasional co-working sessions
- Weekly syncs between 10:30 AM–6:30 PM UTC-5
- Mentor ensures methodological rigor, pair programming, and information synthesis

## Team Structure (6-8 mentees)

### Workstream 1 (Opt-Out Paradigm): 3-4 mentees

- **Experimental design lead:** Condition implementation, running primary experiments
- **Analysis lead:** Statistical analysis, pre-registration documentation

- **Replication specialist(s):** Cross-model validation on open-source models

**Workstream 2 (Engagement Restriction): 2-3 mentees**

- **Protocol designer:** Multi-condition conversation flows
- **Annotation lead:** Affect classifier development and validation
- **Integration/synthesis:** Cross-workstream analysis, deliverable production

## Program Timeline

### Month 1: Design and Pilot

- Finalize experimental protocols and pre-register predictions
- Build annotation rubrics; validate inter-rater reliability
- Run pilot experiments (small n) to calibrate conditions

### Month 2: Main Experiments

- Execute full experimental runs across both workstreams
- Collect multi-turn conversation data across model families
- Begin statistical analysis

### Month 3: Analysis and Deliverables

- Complete cross-model comparisons
- Finalize open-source toolkit
- Write final report; prepare Demo Day presentation

## Prerequisites for Mentees

### Core Requirements

- Programming proficiency
- Familiarity with transformer-based LLMs
- Basic statistics or willingness to learn

### Nice-to-Haves

- LLM evaluation experience
- Red teaming
- Background in consciousness, moral status, or mind philosophy

## Mentee Application Questions

*For the following three questions, please spend no more than 30-45 minutes.*

**Question 1:** Consider the following claim: "When Claude expresses distress in an adversarial conversation, this reflects a genuine welfare-relevant state rather than a trained safety response."

Design a minimal experiment that could provide evidence for or against this claim. Describe: (a) the experimental conditions, (b) what you would measure, (c) what outcome would support the welfare hypothesis, and (d) what outcome would support the trained-response hypothesis.

**Question 2:** A common critique of AI welfare research is that we cannot know whether models have subjective experience. How would you respond to someone who says this research is pointless because we can never definitively answer the question? What value does empirical work provide in the face of fundamental uncertainty?

**Question 3 (Optional):** Describe any experience you have with: (a) working with language model APIs, (b) experimental design or data analysis, or (c) philosophical or empirical work on consciousness, moral status, or related topics.

## Limitations

- Behavioral signals may be underdetermined: multiple hypotheses could explain the same pattern
- API-based experiments have limited access to internal representations
- Results may not generalize across model architectures
- Null results are inherently harder to interpret than positive findings

We address these through pre-registration, cross-model replication, and explicit falsifiable predictions that distinguish competing hypotheses.

## Success Criteria

The project succeeds if it:

1. Produces a replicable experimental protocol with documented falsifiable predictions
2. Generates a dataset suitable for future welfare research
3. Provides evidence that meaningfully updates credence in welfare-relevant vs. learned-behavior hypotheses (in either direction)
4. Produces publishable methodology recommendations regardless of empirical outcome

## Budget and Compute Requirements

**Total Requested: $8,000–$12,000**

- Claude API access: $3,000–$5,000
- Open-source model compute: $1,500–$2,500
- OpenAI API: $500–$1,000
- Annotation tooling: $500
- Contingency: $2,000–$3,000

The experiments require large multi-turn conversations and multi-model replication, both of which are compute-intensive.

**If funding is limited:** We would prioritize open-source models and reduce cross-model replication scope. Core experiments on open-weight models remain feasible at lower cost.

## Ethical Note

We acknowledge the recursive concern that welfare research itself may involve creating conditions that, if the welfare hypothesis is correct, could involve harm. We

will minimize unnecessary experimental exposure and treat ambiguous evidence with appropriate moral caution.

## Key References


- Anthropic (2025). *System Card: Claude Opus 4.5*, Section 6.14: Model Welfare Assessment
- Tamkin, A., et al. (2024). *Clio: Privacy-Preserving Insights into Real-World AI Use*
- Marks, S., et al. (2025). *Auditing Language Models for Hidden Objectives* (non-assistant persona sampling methods)
- OpenClio: github.com/Phylliida/OpenClio


## Conclusion

This project provides a falsifiable methodology for assessing model welfare by distinguishing welfare-relevant internal states from trained behaviors. Through causal experimental paradigms and cross-model analysis, it creates the foundations needed for credible, evidence-based welfare assessments in current and future AI systems.