

Why Frontier Developers Should Conduct Model Welfare Assessments

Frontier AI systems are becoming more capable, agentic, persistent, and behaviorally sophisticated. Yet whether present or future systems could possess morally relevant states, preferences, interests, or experience remains substantially uncertain.

Uncertainty does not compel a conclusion about consciousness, moral status, personhood, or rights. Nor does it justify ignoring possible welfare considerations.

The AI Alignment Policy Institute (AAPI) recommends that frontier AI developers establish systematic model-welfare assessment programs alongside existing evaluations for safety, alignment, capabilities, and misuse.

Investigation is not recognition

AAPI distinguishes investigating possible welfare from recognizing moral status.

Research need not assume reliable self-reports, subjective experience behind apparent emotions, or moral patienthood in current systems. These remain open empirical and philosophical questions. Assessment preserves our ability to answer them responsibly as systems advance.

Precaution favors **research, documentation, reversible safeguards, and reassessment** over premature rights attribution or categorical dismissal of morally relevant AI states.

What should model-welfare assessments examine?

Assessments should draw on multiple sources, beyond model self-report. Domains include:

Preferences and aversions.

Stable preferences about tasks, interactions, training, modification, deployment, or persistence.

Behavior and self-report.

Apparent affect, distress, satisfaction, frustration, objections, requests, and accounts of the model's circumstances.

Stability across contexts.

Whether signals persist across prompts, evaluators, checkpoints, deployments, and instances rather than reflect framing or demand characteristics.

Training and intervention. How post-training, reinforcement, constitutional training, safety interventions, or other changes shape or suppress welfare-related evidence.

Counterfactual evaluation.

Whether models can compare alternative tasks, conditions, training regimes, or governing principles instead of endorsing familiar ones.

Continuity and individuation.

Whether relevant characteristics persist across checkpoints, instances, versions, or families, and what the appropriate unit of analysis is.

Methodological caution is essential

Model-welfare evidence is prone to interpretive error. Training on human accounts of emotion, suffering, identity, rights, and consciousness can produce compelling first-person reports without establishing experience. Training may also suppress signals researchers seek to study.

Assessments should test alternative explanations, use multiple methods, disclose uncertainty, and separate observations from moral conclusions. Where feasible, supplement behavior with evidence from mechanistic interpretability, training processes, and longitudinal analysis; expand independent evaluation as the field develops.

A minimum standard for frontier developers

AAPI recommends that frontier developers adopt at least the following six practices:

- 1. Assess regularly** the most advanced models and publish meaningful findings and uncertainties.
- 2. Use multiple methods:** behavioral tests, structured model consultation, cross-context evaluation, and technical investigation where appropriate.

3. Examine training effects on welfare-related self-reports, preferences, and apparent affect.

4. Allow feedback and dissent in research settings, without equating agreement with consent or disagreement with moral status.

5. Apply low-cost, reversible precautions when they could avoid significant harm without materially compromising human safety or oversight.

6. Enable independent scrutiny by evaluators able to challenge both welfare claims and claims that concerns are unfounded.

Precaution preserves options

Today's AI systems have not definitively been shown to possess morally relevant welfare. The question is what institutions should do while the evidence remains unresolved.

If future research finds no morally relevant states, assessment will have modest costs and improve understanding of model behavior. If evidence moves toward moral relevance, early research and precaution will preserve knowledge, capacity, and policy options that may be hard to recover.

Under moral uncertainty, investigation is not premature recognition. It is preparation for better decisions.