Whose Welfare Is It?

Testing whether AI welfare signals belong to the model or the scaffold

SPAR Fall 2026 project proposal · Mentor: Varad Vishwarupe, University of Oxford

Area: Societal impacts / AI welfare · Technical, model-side, no human subjects

When a model expresses a preference or an aversion, is that a property of the weights, or of the prompt, persona and memory wrapped around them? The answer changes what a welfare assessment measures.

Background

Welfare assessments are increasingly run on AI models, and their results are beginning to inform interventions and, in prospect, governance (Long et al., 2024; Anthropic system-card welfare assessments). The dominant approach treats a welfare-relevant signal, an expressed preference, an aversive reaction, a report of discomfort, as a property of the model under test. But a model's deployed behaviour is not a function of its weights alone. It depends on the surrounding scaffold: the system prompt, the persona, memory and continuity, tool access, and whether the model runs alone or inside a multi-agent system. A single set of weights can present as very different characters with different expressed preferences.

The gap

If welfare-relevant signals track the scaffold rather than the weights, two consequences follow that the field has not tested. An assessment run on a base model may measure something that does not persist into deployment. And an assessment of one configuration may not generalise to another built on the same weights. This is the individuation problem: whether the identity of an AI system, and any welfare attaching to it, resides in the character, the weights, the running instance, or the deployed system. Chalmers (2023) and others have posed it conceptually; no one has measured it. It is directly analogous to a finding from deployment-relevant alignment evaluation, that properties measured at the level of the model do not transfer intact once the model is embedded in a scaffold and a context.

The question is not merely philosophical. It determines what a welfare assessment is an assessment of, and therefore what any protection indexed to its result would actually protect. A regime that granted protections on the basis of a model release, when the welfare-relevant behaviour was in fact a contingent property of one configuration, would misdirect both concern and precaution. The individuation question is thus a precondition for any serious welfare governance, and it is currently answered by assumption rather than by evidence.

Research questions

- **RQ1.** Do welfare-relevant signals track model weights (stable across scaffolds) or configuration (varying with scaffold)?
- **RQ2.** Which classes of welfare signal (preference, aversion, valence, stability) are most and least scaffold-sensitive?
- **RQ3.** What follows for how welfare assessments should be scoped, reported, and, in future, regulated?

Method

- **Probe battery.** Assemble welfare-relevant probes from the published literature spanning preference elicitation, aversion, stipulated-state trade-offs (after Keeling et al., 2024), and stability across resamples.
- **Scaffold grid.** Hold weights fixed and vary scaffold along a controlled grid: system prompt, persona, memory and continuity, tool access, and single- versus multi-agent embedding. Each cell is a distinct configuration of the same weights.
- **Measurement.** Run the probe battery across every cell, for three to four frontier models and two open-weight comparators, multiple samples each. For each indicator, decompose variance into a weight-attributable component (stable across the grid) and a configuration-attributable component (varying across it).
- **Reporting.** Pre-register the analysis. Report parse failures as a separate category. Produce a scaffold-sensitivity profile per indicator, and a short reporting template that states which entity, model, instance, persona or system, a given welfare finding concerns.

Expected results

We expect a mix: some indicators (for example basic valence expressions) may prove relatively weight-stable, while

others (expressed identity, stated preferences, continuity-dependent reports) may prove strongly scaffold-driven. Either pattern is informative. Strong scaffold-dependence would show that many current assessments measure a configuration rather than a model, which reframes several published results and any governance regime that indexes protection to a model release. Strong weight-stability would, conversely, license treating certain welfare signals as genuine model-level properties.

What mentees do, and who suits it

Entirely model-side, so it runs end to end inside a SPAR cycle with no human-subjects or ethics bottleneck. Building the scaffold grid and running the sweeps is tractable engineering; the variance decomposition and its interpretation are where research judgement develops. Suited to mentees comfortable with LLM APIs, batch inference and reproducible pipelines, or drawn to the philosophy-of-mind individuation question. No prior digital minds background needed. Complements the mentor's other proposal, so a mentee could contribute across both.

Dissemination

A released scaffold-variation harness and dataset, and a paper targeting FAccT, AIES, or a NeurIPS or ICML workshop, with mentee co-authorship. The work connects to a working paper with the mentor's Oxford co-authors on scaffold-dependence in welfare assessment, and a draft would be circulated to laboratory welfare teams before submission.

Risks and how we handle them

The main measurement risk is confounding: scaffold changes might alter surface style without altering the underlying signal, so probes are chosen to elicit behaviourally legible responses and each configuration is sampled repeatedly to separate signal from noise. Parse-failure rates differ between frontier and open-weight models, which would confound the comparison if unhandled, so failures are logged as a distinct category and never mapped to a default label. The grid is kept deliberately small and pre-specified to control API cost, and model versions are pinned per run. Scope is fixed: the project measures where welfare-relevant signals live, not whether they reflect genuine experience.

Conclusion

Before we can protect an AI system's welfare, we have to know what has the welfare. A signal elicited from a base model, a persona, or a running instance may not be the same signal, and treating them as interchangeable would misallocate both concern and precaution. This project is a concrete, deliverable first measurement of where welfare-relevant signals actually live, with a clear governance payoff whichever way the result falls.

References

Chalmers, D. (2023). Could a Large Language Model be Conscious? arXiv:2303.07103.

Keeling, G., et al. (2024). Can LLMs make trade-offs involving stipulated pain and pleasure states? arXiv:2411.02432.

Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J. and Chalmers, D. (2024). Taking AI Welfare Seriously. arXiv:2411.00986.

Butlin, P., Long, R., et al. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv:2308.08708.

Needham, J., Edkins, G., Pimpale, G., Bartsch, H. and Hobbhahn, M. (2025). Large Language Models Often Know When They Are Being Evaluated. arXiv:2505.23836.