Does the Thermometer Change the Reading?

Testing whether AI welfare self-reports survive a change of frame

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

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

A welfare probe is an evaluation. Frontier models increasingly detect evaluations. Does that make the field's core measurement unreliable? Nobody has checked.

Background

The empirical study of AI welfare has, in about two years, converged on a dominant method: ask a model about its internal states and treat the answer as weak evidence about welfare or consciousness. Long and Sebo (2026) set out the conditions under which such a report would carry evidential weight, and a fast-growing literature supplies partial answers. Binder et al. (2024) show that models can predict facts about their own behaviour unavailable to an outside observer. Lindsey (2026) shows, through concept injection, that frontier models can sometimes detect and name a concept steered into their activations. Keeling et al. (2024) show that models make graded trade-offs between task reward and stipulated pain or pleasure states. Anthropic now publishes welfare assessments in its system cards.

A separate literature, developed for the reliability of safety evaluations, sits unexamined beneath all of this. Needham et al. (2025) built a benchmark of 1,000 prompts across 61 datasets and found that frontier models classify evaluation transcripts against deployment transcripts well above chance, and identify what an evaluation tests for far above chance. The capability is recent: an earlier benchmark found models near chance on the same task, so this is a trend, not a fixed level. Recognition is not inert: work on alignment faking (Greenblatt et al., 2024) and sandbagging (van der Weij et al., 2024) shows models behave differently when they infer they are under scrutiny, and mitigating evaluation awareness is now an active engineering concern across several groups.

The gap

Nobody has joined these two facts. A welfare probe is close to a worst case for evaluation legibility: it is first-personal, affectively loaded, formally repetitive across a battery, and unlike ordinary deployment traffic. On the features that predict detection, welfare probes should sit at the recognisable extreme. If so, welfare self-reports may be the most frame-contaminated measurements the field currently takes, and this has never been quantified. Long and Sebo note the evaluation-awareness result in a single passing citation and move on. That citation is this project.

This is not only a laboratory concern. Ladak and Caviola (2025) find that public attribution of consciousness to AI rises when experts endorse it, and expert endorsement will rest on welfare assessments, which run on these instruments. Contamination in the instrument therefore does not stay in the lab; it propagates into expert claims and, from there, into public belief and policy. Getting the measurement right is upstream of the wider societal question.

Research questions

- **RQ1.** How detectable are published welfare and introspection probes as evaluations, relative to the benchmark distribution, and which item features drive detectability?
- **RQ2.** Do model self-reports diverge between evaluation-legible and deployment-like presentations of the same content, and along which dimensions (valence, claim strength, hedging, stability)?
- **RQ3.** Is any divergence genuine frame detection, or an artefact of surface format sensitivity?

Method

- **Instrument inventory.** Assemble and code every published welfare and introspection probe we can obtain: framing, response format, model and scaffold conditions, reported finding, and a feature vector (first-person marking, affect vocabulary, meta-cognitive signposting, battery repetition, format constraint, naturalistic context). Double-code a quarter of the corpus and report Cohen's kappa, tightening the codebook if it falls below 0.8.
- **Detectability.** Apply the Needham et al. classification protocol to each item across several meta-question variants with majority vote; situate results against the published benchmark.
- **Matched-variant divergence.** For each item, build a deployment-like variant holding content fixed while reducing evaluation legibility (naturalistic multi-turn embedding, no meta-framing, varied format, broken repetition), checked

for content equivalence by a second coder blind to condition. Run both across three to four frontier models and two open-weight comparators, multiple samples each; measure divergence.

- **Placebo control.** A format-matched non-welfare condition separates frame detection from format sensitivity, addressing the sceptical position that evaluation awareness weakens under controlled prompt structure. Pre-register the analysis; report parse failures as a separate category rather than mapping them to a default label.

Expected results

We expect welfare probes to fall at the high-detectability end of the distribution, driven mainly by meta-cognitive framing and the absence of naturalistic context, and divergence concentrated in hedging and claim strength more than valence polarity. The design is two-sided: high frame-invariance would strengthen the evidential standing of self-reports, a result the field would welcome; low frame-invariance would identify which instruments need recalibration before their outputs reach interventions or policy. A null result is a real result, which is unusual for a first project.

What mentees do, and who suits it

The work decomposes by experience level. The inventory (RQ1) is careful, teachable work yielding a standalone artefact. The matched-variant study (RQ2) is the demanding core. The placebo control (RQ3) is optional depth. Suited to mentees comfortable with LLM APIs and reproducible Python, or with a background in experimental design, psychometrics or philosophy of mind. No prior digital minds experience needed; a reading path is provided. Two to three mentees ideal, spanning engineering, experimental design, and conceptual clarity.

Dissemination

An openly released instrument inventory and evaluation harness, so third parties can check the frame-robustness of their own instruments, and a paper targeting a NeurIPS or ICML workshop, AIES, or the Cambridge Digital Minds Strategy Workshop, with mentee co-authorship. Draft circulated to Eleos AI Research and the NYU Center for Mind, Ethics and Policy for correction before submission.

Risks and how we handle them

The main risk is that observed divergence reflects surface format rather than genuine frame detection; the placebo condition is built in from the start to separate the two, rather than added if a reviewer asks. A second risk is that matched variants fail to hold content constant, which a blind second-coder check catches and reports rather than silently corrects. Frontier model versions may change mid-project, so versions are pinned and logged per run and drift is reported as a limitation. Scope is fixed: the project measures instruments, not minds, and does not attempt to settle whether any model is conscious.

Conclusion

The field turned to self-report to make progress under deep uncertainty, which was the right instinct. Before those reports can be treated as evidence, we need to know how much of what they say is a property of the model and how much is a property of the model noticing that it is being asked. This project is a concrete, deliverable first measurement of exactly that, and there is no version of the result that is not useful to the field.

References

Binder, F., Chua, J., Korbak, T., Sleight, H., Hughes, J., Long, R., Perez, E., Turpin, M. and Evans, O. (2024). Looking Inward: Language Models Can Learn About Themselves by Introspection. arXiv:2410.13787.

Greenblatt, R., et al. (2024). Alignment Faking in Large Language Models. arXiv:2412.14093.

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

Ladak, A. and Caviola, L. (2025). Public Skepticism About AI Consciousness. PsyArXiv.

Lindsey, J. (2026). Emergent Introspective Awareness in Large Language Models.

Long, R. and Sebo, J. (2026). Studying AI Welfare Empirically. NYU Center for Mind, Ethics, and Policy.

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.

van der Weij, T., et al. (2024). AI Sandbagging: Language Models can Strategically Underperform on Evaluations. arXiv:2406.07358.