Do AI labs follow the principles for responsible consciousness research?

Intermediate · A few weeks · Free · Project idea

Online version: https://ai-safety-sandbox.vercel.app/projects/labs-vs-consciousness-principles/

Summary

Butlin and Lappas set out principles for research objectives, procedures, knowledge sharing and public communication. Score three AI developers' published policies against them, with Suleyman's essay as a deliberately contrasting position.

Why it matters

AI welfare is becoming a corporate policy question. Scorecards show where principles are being met and where they're so vague that any policy could claim to meet them.

What you will have at the end

A scorecard and short essay.

How to start (tick each box as you go)

- [ ] **1. Write your question in one sentence.**
- [ ] **2. Spend 30 minutes finding what has already been written about it.**
- [ ] **3. Collect evidence in a simple table: claim, source, how strong.**
- [ ] **4. Draft 1–2 pages: answer first, then reasons, then uncertainties.**
- [ ] **5. Ask one person to poke holes in it, revise, and share.**

Key terms

Moral status
Whether a being's interests matter morally for its own sake. AI welfare research asks whether, and when, AI systems might have it.

Consciousness
Having subjective experience, so that there is "something it is like" to be you. Whether AI systems could be conscious is an open question.

Source and links

AI Safety × Philosophy project ideas (P8) — builds on Butlin and Lappas, Principles for Responsible AI Consciousness Research; Long et al., Taking AI Welfare Seriously; Suleyman, We Must Build AI for People

Butlin & Lappas — Principles (JAIR, 2025)
https://jair.org/index.php/jair/article/view/17310

Long et al. — Taking AI Welfare Seriously
https://arxiv.org/abs/2411.00986

When you have finished

Write a few paragraphs on what you did, what you found and what you would try next, and share it with a friend, your society or online.

Courses, mentors, small grants and communities:
https://ai-safety-sandbox.vercel.app/next-steps/

Questions? Ask in AI Safety Cardiff University (AISCU): https://aiscu.org/

Notes