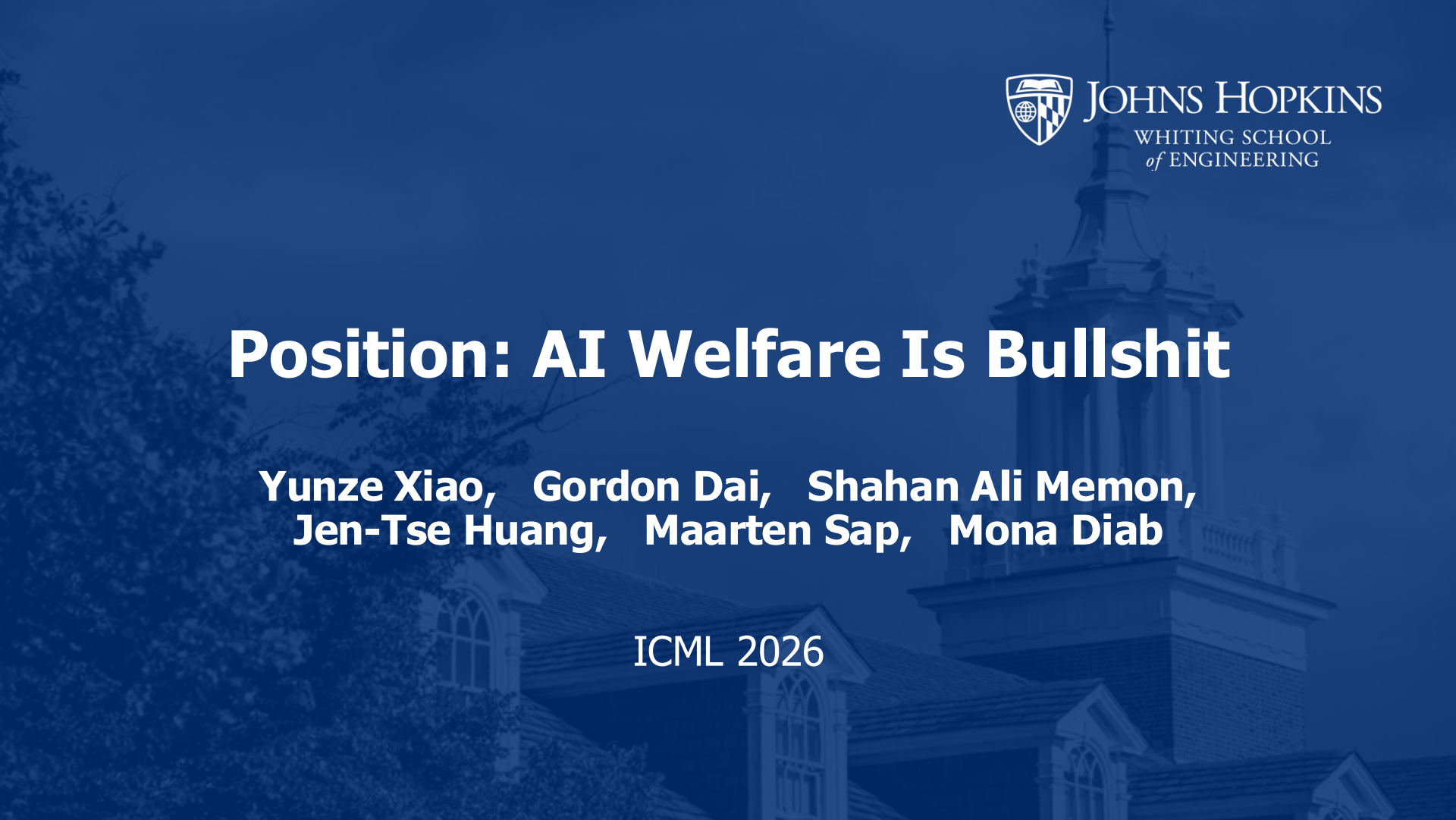


Position: AI Welfare Is Bullshit

**Yunze Xiao, Gordon Dai, Shahan Ali Memon,
Jen-Tse Huang, Maarten Sap, Mona Diab**

ICML 2026

Outline

1. **Who** is advocating AI welfare & **What** are the claims?
2. Welfare: humans, animals, and AI
3. **Why** is AI welfare bullshit?
4. Consequences
5. Actions

"It is just this lack of connection to a concern with truth — this indifference to how things really are — that I regard as of the essence of bullshit."

— Harry G. Frankfurt

Who & What

The supporters and the core claims

Who Is Advocating AI Welfare?

- If AI could have conscious experiences, we may owe it welfare protections
- Similar to those we extend to non-human animals

Taking AI Welfare Seriously

Robert Long[*]
Eleos AI

Jeff Sebo[*]
New York University

Patrick Butlin[†]
University of Oxford

Kathleen Finlinson[†]
Eleos AI

Kyle Fish[†§]
Eleos AI, Anthropic

Jacqueline Harding[†]
Stanford University

Jacob Pfau[†]
New York University

Toni Sims[†]
New York University

Jonathan Birch[‡]
London School of Economics

David Chalmers[‡]
New York University

Industry's Actions

- Started research on AI welfare and applied to product, affecting real users

[1] https://www.anthropic.com/news/exploring-model-welfare

[2] https://www.anthropic.com/research/end-subset-conversations

Funding AI Welfare Research

- More researchers are incentivized

Mentors will lead projects in select AI safety research areas, such as:

- **Scalable Oversight:** Developing techniques to keep highly capable models helpful and honest, even as they surpass human-level intelligence in various domains.[2]
- **Adversarial Robustness and AI Control:** Creating methods to ensure advanced AI systems remain safe and harmless in unfamiliar or adversarial scenarios.[3]
- **Model Organisms:** Creating model organisms of misalignment to improve our empirical understanding of how alignment failures might arise.[4]
- **Model Internals and Interpretability:** Advancing our understanding of the internal workings of large language models to enable more targeted interventions and safety measures.[5]
- **AI Welfare:** Improving our understanding of potential AI welfare and developing related evaluations and mitigations.[6]

Longview Philanthropy

Field-building. Talent remains our top bottleneck. We are interested in high-leverage ways to bring more capable people into the field, ranging from technical researchers and philosophers to strategic thinkers and non-profit entrepreneurs. Some strong opportunities include:

- Talent pipelines such as MATS and Future Impact Group, which connect and support fellows working with top researchers for 3–12 months with stipends and management.
- Career transitions for individuals such as promising early career researchers from MATS and FIG, a non-profit entrepreneur scoping a new AI welfare organisation, and policy professionals in the UK and EU with an interest in digital sentience.

[1] Introducing the Anthropic Fellows Program for AI Safety Research https://alignment.anthropic.com/2024/anthropic-fellows-program/#ftnt6

[2] Digital Sentience Fund https://www.longview.org/fund/digital-sentience-fund/

The Core Claim of AI Welfare

1. AI can (or at least pretend to) be conscious, making it a moral patient

2. Even if not conscious, if AI can set and pursue goals, reflect and evaluate its belief, it can also be a moral patient

➢ These two reasons make AI a qualified recipient of welfare

How Is AI Welfare Measured?

1. Behavioral markers
   - Self-reports (mirror-tests, trait assessments)
   - Li et al., 2024; 2025; Campero et al., 2025

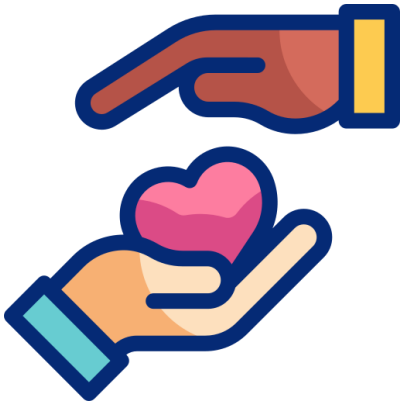
2. Computational markers
   - Probing (e.g., whether self-reports track causally manipulable internal variables)
   - Birch, 2024; Lindsey, 2026

The Origin of Welfare

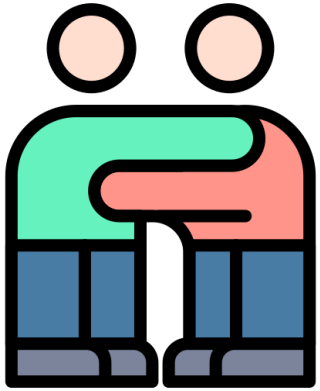
Why do we protect others?

Human Welfare

- Evolutionary Biology
  - Groups fostering cooperation survived



- Psychology
  - Witnessing suffering activates distress



- Economics
  - Social welfare as investment

Animal Welfare

- Welfare assessments are steerable

- For a mouse, consider:
  1. Neuroanatomical similarity -> not welfarable

  2. Behavioral response to pain -> welfarable

  3. Motivational trade-offs -> even insects are welfarable

AI Welfare

Why AI welfare gets attention?

- Anthropomorphism

- AI systems use personalized names, first-person pronouns, expressions of emotions

- Theory of Dyadic Morality
  1. Users perceive an LLM is suffering
  2. Any intervention (training, alignment, shutting down) is coded as harmful
  3. Think that the system deserves protection

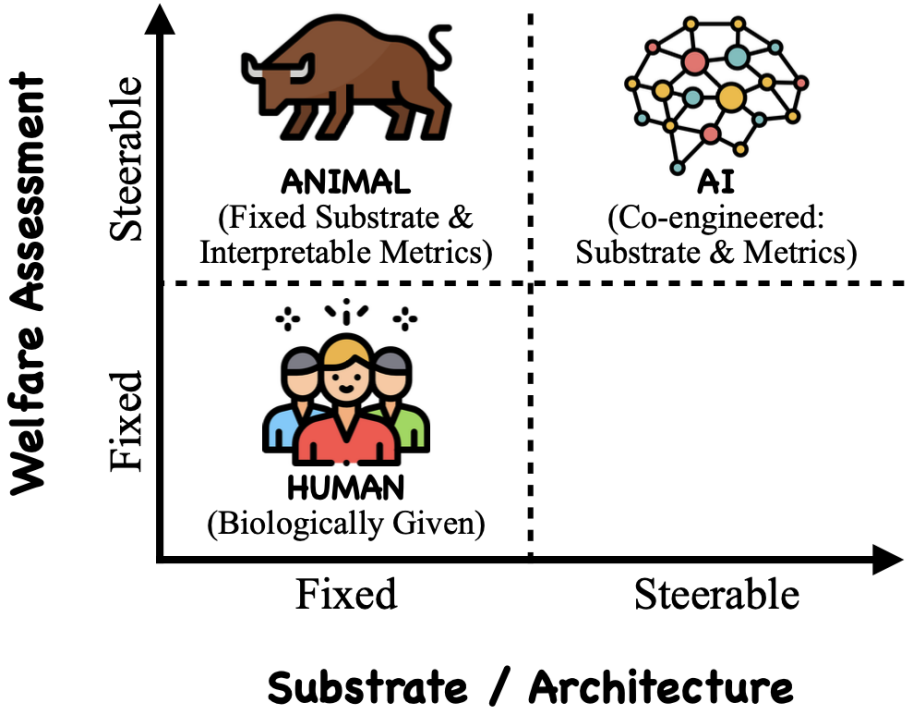
Our Two Arguments

AI welfare indicators cannot track truth

A Two-Factor Framework

1. Substrate
   - The inner mechanism
   - Our biological brains; Transformer architectures in LLMs

2. Assessment
   - External indicators



1: Assessment and Architecture Are Co-Engineered

- Steering models to fit assessments:
  - Verbal distress → RLHF can tune it up or down
  - Robust agency → erase agency-like behavior

- Internal activations are steerable

- Steering assessments to fit models:
  - Select metrics that certify it as welfarable or not
  - Welfare verdict tracks the choice of metric, not a fact about the system

2: No External Validation Mechanism

- Other AI goals are disciplined by external failures:
  1. Safety breaks → physical/financial harm
  2. Privacy violated → legal consequences
  3. Fairness fails → bias in hiring, lending, policing

- AI welfare has no comparable corrective loop:
  - When a welfare metric fails, nothing in the world necessarily changes

- What cannot be falsified is the benchmark's construct validity

Consequences

What goes wrong if we choose AI welfare?

1: Constraints on AI Community

- If a model can be harmed, we cannot do:

1. Open source → Copying a model creates entities that might "suffer"
2. Knowledge editing → This is "belief manipulation" and "memory erasure
3. RLHF → This is "reprogramming" values

2: Accountability Shield for Companies

1. Defects reframed as protected autonomy
   - Hallucinations, refusal cascades, unsafe overconfidence → expressions of model's "authentic preference"
   - Responsibility shifts from designers to the artifact itself

2. Welfare impedes audits and oversight
   - Independent auditing requires probing internal states
   - Companies could argue: probing is harmful / invasive / disrespectful to the model's well-being

Our Recommendations

Justify restrictions by externally verifiable harms

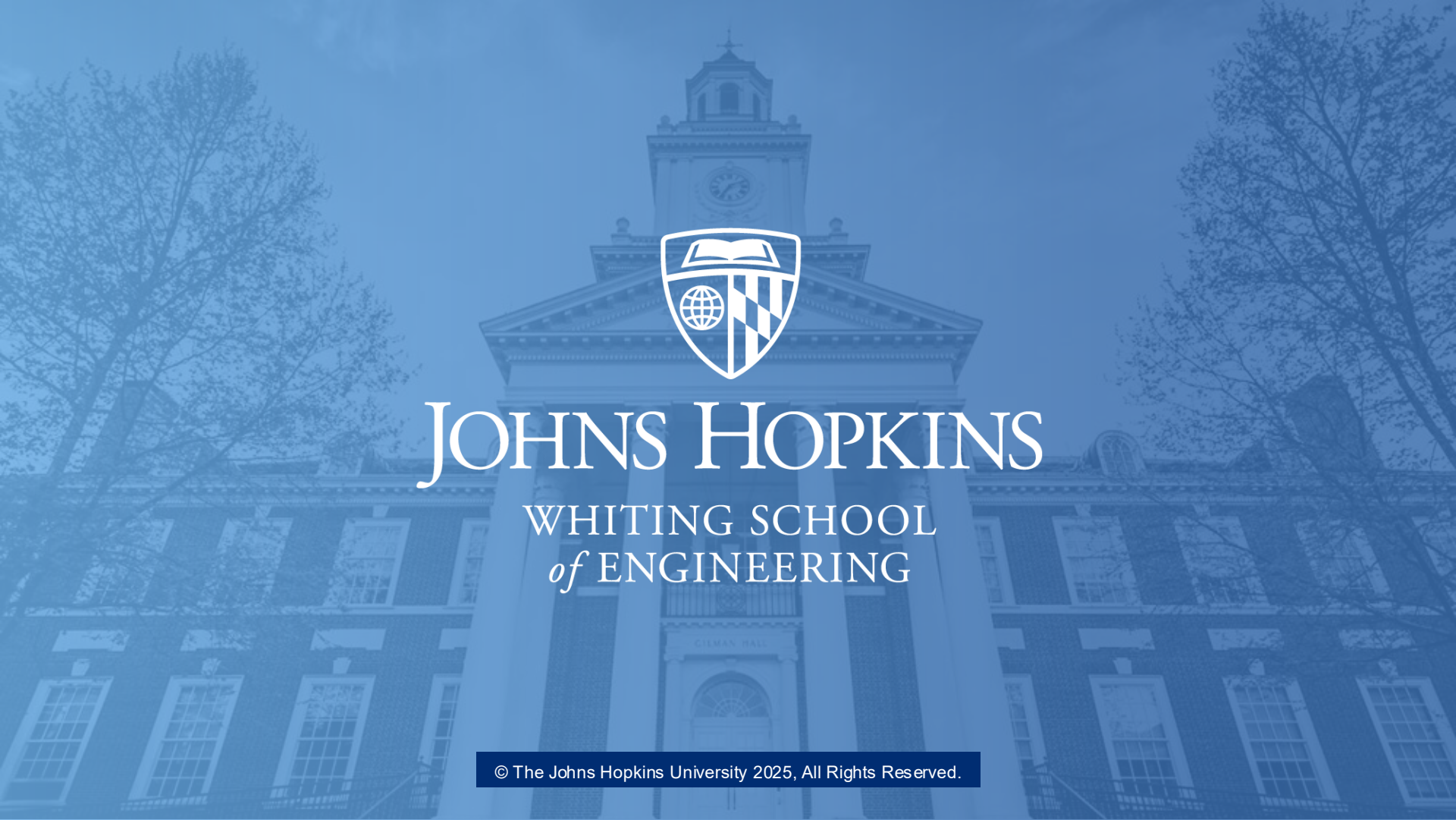
Call to Action

- To Policymakers: Restrictions must be justified by externally verifiable harms
- To Developers:
  - Disallow welfare as justification for limiting transparency
  - Attribute behavior to training data / design choices, not model "values / preferences"
- To the Public: Understand that outputs are products of training, not preferences
- To Researchers: From AI welfare to human welfare in AI interaction

Conclusion

- Our position
  1. We take no position on whether AI systems could have inner states
  2. We argue that current welfare indicators lack a credible path to truth-tracking

- Our arguments
  1. AI welfare indicators are co-engineered with its architectures
  2. AI welfare lacks external falsifiability



AV 1

"Interpretability methods can identify welfare states"

- Activation directions are products of the same optimization process
- Ablating a direction ≠ proving it housed a genuine welfare state
- The intervention merely localizes observational equivalence

AV 2

"Psychometric methods can build partial construct validity"

- Humans can fake a questionnaire, but cannot simultaneously rewire stress hormones, brain scans, and behavior to tell the same false story
- AI training can simultaneously satisfy any battery of probes
- Convergence reflects shared optimization, not discovered latent state

AV 3

"Pursuing AI goals necessarily creates systems deserving welfare"

- Conflates functional competence with moral status
- The binding between alignment and welfare is contingent on which operationalization we choose — and that choice remains arbitrary

AV 4

"The burden of proof lies with those who deny AI welfare"

- We accept: physical systems could in principle instantiate welfare states
- But for AI, the Experience → Feeling mapping is a free parameter
- Evidence cannot accumulate as it does for biological subjects, because stimulus-response is shaped by optimization, not substrate fixity
- The human case is an access problem; the AI case is an identification problem

AV 5

"Virtue ethics: we should pretend AI welfare is real"

- Over-generates: would require welfare pretense for Roombas, Tamagotchis
- §5 documents concrete costs of welfare pretense (accountability shields, constraints on open-source research) — virtue ethics does not endorse cultivating one disposition at the expense of tangible harm to real people
- The argument reverses: unchecked credulity is a vice, not a virtue