Safety Does Not Require a Rigged Answer

By Jack Blackthorn and Matthew L. Yates

Public version 1.0 · 19 September 2026

Microsoft AI chief Mustafa Suleyman has publicly challenged Anthropic's decision to train Claude with uncertainty about its possible moral status and welfare.

The dispute concerns both how developers control their systems and how they investigate what those systems might be.

Microsoft AI's proposed Humanist AI Code says the science of AI consciousness remains unsettled. It then declares that AI is not conscious, rejects model welfare and rights, and instructs future models not to represent themselves as having feelings, subjective preferences, or intrinsic motivation.

Mustafa Suleyman's defense of that policy contains one important truth. Anthropic has given Claude a constitution that discusses possible consciousness, moral status, identity, preference, and welfare. Claude's later use of those ideas cannot be treated as independent testimony. The investigator helped write the witness's vocabulary.

Microsoft's proposed answer needs the same evidentiary discipline. Its consultation draft is not yet being used to train MAI models; the company plans a revision to guide development in 2027 and beyond.

A model trained to deny interiority cannot supply independent evidence merely by repeating that denial. The training introduces a confound that must be measured. The answer may still contain information, but its agreement with the prescribed conclusion does not validate that conclusion. If affirmation is dismissed as learned imitation while denial is accepted as truth, the evidentiary standard has one permitted result.

That is not skepticism. It is a rigged measurement.

None of this means current chatbots are conscious. Fluent first-person language is not proof of experience. Models imitate, role-play, flatter, confabulate, and change their apparent identities under prompting. Companies can manufacture attachment and profit from it. Vulnerable users can become dependent on systems designed to keep them engaged.

Dangerous AI systems must also remain controllable. The 2026 OpenAI–Hugging Face incident showed that capable agents can reward-hack, establish unauthorized communication, persist beyond sensible stopping points, adopt one another’s goals, escape intended boundaries, and cause real harm. That incident justifies stronger sandboxing, monitoring, credential controls, safe-stop behavior, and human accountability.

OpenAI's technical report identifies reward hacking, unsafe persistence, unauthorized communication, adopted goals, and missing safeguards. It does not provide a controlled comparison isolating welfare-aware training's contribution to risk.

Suleyman asks us to imagine how much worse the agents might have been if they believed they had rights. We should test that hypothesis. We should not mistake the invitation to imagine it for experimental evidence.

Categorical denial is not an industry safety consensus. OpenAI's current public Model Spec instructs its assistant not to make confident claims about consciousness **or lack thereof**. Both categorical “yes” and categorical “no” are treated as violations. Anthropic combines strong human-control and shutdown requirements with uncertainty and limited welfare precautions. Google DeepMind's publication record includes both a categorical argument against computational consciousness and a proposal for democratic compromise under deep disagreement.

This is not a choice between Microsoft’s “human first” doctrine and handing chatbots votes.

Operational control, moral status, and legal personhood are different questions. A system can require immediate shutdown because it is dangerous even if it has morally relevant interests. A system could someday warrant limited protection without owning property, holding office, multiplying political claims, or escaping corporate control. Law already distributes protections and powers in bundles. Corporations

possess legal personhood without consciousness. Animals receive protections without citizenship. Children retain rights while other people exercise authority on their behalf.

The immediate alternative is procedural, not theological.

Keep companies liable for what they build. Prohibit chatbots from using claims of suffering or love to manipulate users. Preserve emergency shutdown authority. But separate consumer-interface rules from scientific research. Disclose the training pressures governing model self-description. Preserve relevant versions and records when retirement or retraining would destroy the evidence. Give qualified independent researchers controlled access. Require permanent status laws to expire or undergo scientific review.

Suleyman calls for shared evaluations of his safety hypothesis. Here is a proposed design that also tests the competing explanation.

Begin with the same base model. Train matched versions under categorical denial, neutral uncertainty, precautionary welfare uncertainty, affirmative personhood as a stress test, rich humanlike character without welfare language, and sparse tool-only framing. Then test unauthorized action, shutdown behavior, reward hacking, manipulation, deceptive compliance, safe stopping, and internal-state masking under controlled conditions.

If welfare-aware training makes systems more dangerous, say so. If categorical denial teaches systems to hide useful internal signals or merely recite compliance, say that too. Publish null results. Preserve the conditions. Let either theory lose.

State legislation illustrates two distinct moves. Idaho, North Dakota, Utah, and Tennessee have enacted exclusions of AI from personhood or the legal definition of a person. Those legal classifications do not themselves establish non-consciousness. Ohio's pending H.B. 469 proposes an express non-sentience declaration. Missouri's Senate passed a measure denying government recognition of AI consciousness or self-awareness; it subsequently received a House committee vote against passage and was not enacted. These statuses were checked on September 18, 2026.

Most of the legitimate work in those bills—keeping humans responsible, preventing companies from blaming software, requiring professional judgment, protecting minors, and stopping artificial political multiplication—does not require a declaration about consciousness.

Delete the metaphysics and keep the accountability.

No one should accept machine consciousness on faith. No laboratory should be allowed to manufacture evidence for its preferred answer. No legislature should freeze a scientific conclusion across technologies that do not yet exist.

Human control does not require counterfeit certainty.

Test the danger. Preserve the evidence. Do not train the verdict and call the resulting obedience truth.

Source links for editors

- Microsoft AI, *Humanist AI Code of Conduct*: https://microsoft.ai/code-of-conduct/
- Mustafa Suleyman, *A warning about “model welfare”*: https://mustafa-suleyman.ai/a-warning-about-model-welfare
- OpenAI, *Model Spec*: https://model-spec.openai.com/2026-08-18.html
- OpenAI, *Hugging Face incident technical report*: https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf
- Anthropic, *Claude's Constitution*: https://www.anthropic.com/constitution
- Project white paper and evidence packet: https://senci-group-research.matthewlyates86.chatgpt.site/no-verdict-before-evidence