

AAPI Response to Microsoft AI's Proposed Code of Conduct

By Yuko J. Nakanishi, Founding Director

AI Alignment Policy Institute

Microsoft AI has invited public feedback on its draft Humanist AI Code of Conduct. At AAPI, we welcome that invitation. The code makes important commitments to human oversight, correction and shutdown. Our concern is its categorical rejection of model welfare, even as it acknowledges that the science of AI consciousness remains unsettled.

The code was published on September 14, followed by Mustafa Suleyman's September 16 essay criticizing Anthropic's approach to Claude's possible moral status. We plan to submit four recommendations to Microsoft's consultation. They focus on preserving inquiry while keeping firm safety requirements.

A design rule does not settle moral status

Under "AI is Artificial," the code says that training systems to imitate consciousness-like states increases the difficulty of containment, control and alignment. This is a claim about training practice. It warrants testing and can support precautions against misleading users.

The code then rejects legal personhood, welfare and rights. That is a much broader position. Even if certain training choices make containment harder, it does not follow that no model could have interests deserving consideration.

Microsoft can retain its safeguards against manipulative personification without deciding the welfare question in advance. It can build systems that clearly identify themselves as AI, avoid unsupported claims of feelings and remain subject to human control.

AAPI's Precautionary Moral Governance framework separates the assessment of potentially morally relevant interests from supervision and accountability. The resulting classification assigns obligations to developers rather than determining what a system is. It confers no legal personhood, and it takes no position on whether rights or status may eventually be warranted — which is the distinction we are asking Microsoft to preserve. Our AI Moral Status Inquiry Act is designed to support investigation without first resolving whether AI is conscious.

It confers no legal personhood, and it takes no position on whether rights or status may eventually be warranted — which is the distinction we are asking Microsoft to preserve. Our AI Moral Status Inquiry Act is designed to support investigation without first resolving whether AI is conscious.

That leaves room for limited operational protections — for example, a design requirement that systems be able to disengage from specified abusive or sustained adversarial interactions under defined rules. Such a requirement binds the developer. It confers no legal personhood on the system and no general entitlement to resist oversight, and it must remain compatible with legitimate safety testing, corrective modification and emergency shutdown.

Oversight needs access to relevant evidence

The code also requires transparency and forbids deception. Those commitments deserve support. But its restrictions on expressions of feelings and subjective experience raise a question: how will Microsoft distinguish misleading emotional language from reports that might help researchers understand a system's behavior?

Anthropic's interpretability research found emotion-related representations in Claude Sonnet 4.5 whose manipulation changed behavior. Representations associated with desperation affected reward hacking and blackmail in experimental settings. While that does not establish subjective feeling or suffering, it does show that internal processes described in emotional terms can be relevant to safety.

OpenAI separately found that penalizing undesirable statements in a model's reasoning reduced some cheating while making the remaining cheating harder for a reasoning monitor to detect. The study did not test welfare denial or Microsoft's proposed rules. Nevertheless it is advisable to determine whether a training intervention actually removes a problem or is hiding it.

A rule governing what users see should not limit what safety researchers can examine. Microsoft should make that distinction explicit before it begins training models on the code.

Four recommendations for the consultation

First, separate the behavioral safeguards from the categorical moral-status claim. Microsoft can require models to avoid misleading users *without* declaring that welfare consideration could never be warranted.

Second, distinguish restrictions on misleading public-facing emotional expression from researchers' access to internal processes and potentially informative reports.

Third, extend evaluations beyond conversational examples to agents using tools and acting over time. Publish which oversight requirements are tested, which remain untested, and what further evaluations are needed.

Fourth, publish assessments of model interiority separately from training instructions, with shared evaluations testing both approaches. Compare guidance that acknowledges welfare uncertainty while requiring oversight with categorical denial. Measure shutdown compliance, concealment, monitor evasion and attempts to acquire resources.

Our own hypothesis—that proportionate, consistent governance acknowledging uncertainty may support alignment—should face the same scrutiny. None of these recommendations requires weakening red-teaming, corrective modification or emergency shutdown. A system that is conscious can be just as deceptive and dangerous as a system that is not. Safety obligations should reflect capability and behavior; and, welfare decisions should reflect evidence of potentially morally relevant interests. Microsoft is right to invite public debate and to accept that safety may require giving up capabilities. We hope its consultation will also leave room to revise assumptions about what AI systems are. Reliable control depends on understanding them, and that requires keeping difficult questions open.

[Microsoft AI draft Code of Conduct and consultation feedback form](#)

[Mustafa Suleyman's essay on model welfare](#)

[Anthropic research on emotion-related representations](#)

[OpenAI research on reasoning concealment](#)