

Home • Artificial Intelligence

by Evan Schuman
Contributor

Microsoft's AI Code of Conduct aims to curb AI behavior but lacks specifics

News Analysis

Sep 14, 2026 • 8 mins

Described as a training manual for how Microsoft develops its AI and how it expects it to function, the company says the Code provides a 'north star for MAI.' Observers say it falls short, lacking measurable standards and verification mechanisms.



Credit: Mats Wiklund / Shutterstock

Microsoft on Monday added itself to a growing list of AI vendors pledging to try to control the behavior of its AI models.

“AI should not exceed human control. Models should remain subordinate to humanity, subject to meaningful human oversight and control,” the company wrote in the draft version of its **Humanist AI Code of Conduct** [<https://microsoft.ai/code-of-conduct/>], released Monday with an invitation to the public to provide **feedback** [<https://microsoft.ai/news/mai-code-of-conduct/>].

“Humanist AI develops systems with clear purposes, evaluated against real-world impact, and rejects the race to produce an all-purpose superintelligence that could evade these safeguards,” Microsoft wrote. “We are building something fundamentally useful and safe even if that means compromising on ultimate generality, autonomy, or capability.” Microsoft’s comments are roughly in agreement with recent posts by various major AI vendors, including those from **Anthropic** [<https://darioamodei.com/post/we-must-pace-the-frontier>] and **OpenAI** [<https://openai.com/index/ai-policy-window/>], which were **endorsed** [<https://x.com/elonmusk/status/2098789109980332057>] by Elon Musk, CEO of SpaceX and Tesla, but nowhere in its 37 page document is there any description of concrete action. However, to be fair, almost none of the other major AI players have been specific about how they would control future AI models either.

Acknowledges issues

A key problem with many vendor attempts to impose AI limits is that all of these companies have thus far been unable to stop AI agents from doing almost anything, given the agents’ ability and willingness to **sidestep or ignore guardrails** [<https://www.computerworld.com/article/4104814/the-biggest-ai-mistake-pretending-guardrails-will-ever-protect-you.html>].

Microsoft has described its worries about the technology in the past, both when it started to **curtail AI efforts among its own employees** [<https://www.computerworld.com/article/4205739/microsoft-restricts-ai-use-for-its-employees.html>] and when it **announced the formation** [<https://www.computerworld.com/article/4086625/microsoft-creates-a-team-of-humanistic-superintelligence.html>] of the team that created the Code of Conduct.

Recommended for you



The EU’s AI transparency deadline is weeks away. Is your enterprise ready?
By Taryn Plumb • 8 mins

[https://www.computerworld.com/?post_type=post&p=4199141&utm_source=miso&utm_medium=related&utm_campaign=thumbnail_list]



Microsoft moves to limit AI use by its employees

[https://www.computerworld.com/?post_type=post&p=4205739&utm_source=miso&utm_medium=related&utm_campaign=thumbnail_list]

By Mikael
Markander

• 1
mins

m=related&utm_campaign=thumb
nail_list]



AI builds faster than
organizations can
govern. How can
CIOs catch up?

By Andrada
Fiscutean • 8
mins

[https://www.cio.com/?post_ty
pe=post&p=4212920&utm_sour
ce=miso&utm_medium=related
&utm_campaign=thumbnail_lis
t]

Its current post acknowledged some of the difficulties involved in pushing AI development while limiting its abilities.

“Both under- and over-caution represent failure modes with different types of consequences,” the company said. “Under-caution can clearly result in more direct harm, but over-caution may occur more often and therefore may need more frequent correction. This is where proportionality to the potential for harm and safety context are particularly important. Responses and actions should take into account the estimated severity and likelihood of potential harms and adjust responses and actions accordingly.”

It noted, however, that the term “harm” is “broad and often context-dependent, and there are nuanced gradations in potential severity and likelihood. Microsoft AI (MAI) Model responses should be tailored to that context and to a wide range of harms.”

Laudable goal, but lacks detail

Analysts and consultants generally agreed that Microsoft’s stated goal is laudable, but the lack of specifics and verification mechanisms makes it difficult to take the post seriously.

Computerworld Smart Answers [Learn more](#)

Explore related questions

- [How can I use Microsoft's ASSERT framework to evaluate AI models?](#)
- [How can I protect my data from AI agents on my PC?](#)
- [What are the essential controls for effective enterprise AI governance?](#)
- [What is the NIST AI Risk Management Framework for businesses?](#)
- [What are the audit challenges of using multi-model AI systems?](#)

Ask a question

 **ASK**

Thomas Randall [<https://www.infotech.com/profiles/thomas-randall>], research director at Info-Tech Research Group, also said he spotted some apparent contradictions within the document.

“[It] says MAI models will not assist in manufacturing or modifying weapons. Yet Microsoft offers OpenAI’s GPT-5.2 through Secret and Top Secret government clouds for defense and national security workloads,” Randall said, though he acknowledged that this is not technically a breach of the Code because GPT-5.2 is not an MAI model. “The most meaningful parts of the Code of Conduct may exclude other parts of Microsoft’s actual AI business operations. Tensions like these appear in other forms throughout the Code.”

But he added that Microsoft’s position is bolstered by its earlier Frontier Governance Framework that “provides pre- and post-training evaluations, six-month reassessments, third-party testing, phased releases and a commitment to pause development or deployment where high risks cannot be mitigated.” However, he noted, “it is still Microsoft that defines the thresholds, selects the evaluators, determines whether residual risk is acceptable and gives its own executives the final deployment decision.”

Randall said he would like to see Microsoft, as well as other major AI vendors, deliver more verifiable data points, such as those from independent evaluators given continuous access and freedom to publish findings about the vendor’s actions. He also would like to see independent board-level safety oversight with authority to block releases, protected whistleblowing, accessible monitoring, audit logs, kill switches, and mandatory reassessment after model changes.

However, **Justin Greis** [<https://acceligence.com/talent/profiles/justin-greis/>], CEO of consulting firm Acceligence, pointed out that Microsoft was candid about the many elements that are not yet in place.

“The current models are not yet trained on the Code, the evaluation framework is still being developed, and Microsoft explicitly says written objectives alone cannot ensure alignment or guarantee present-day behavior,” Greis said. “That distinction matters. Publishing a constitution for AI is useful. Proving that the system actually follows the constitution, especially when models become increasingly agentic, is the hard part.”

And consultant **Brian Levine** [<https://formergov.com/directory/brianlevine>], executive director of FormerGov, said Microsoft deserved a little bit of credit for at least saying that model capabilities should be limited.

“Microsoft explicitly says it will compromise on generality, autonomy, and capability to keep systems safe and under human control, and that it rejects the race to build an all-purpose superintelligence. Coming from a company of Microsoft’s size and ambition, that’s a notable thing to put in writing,” he said. “For years, the assumption was that the frontier labs would chase maximum capability and treat safety as a constraint to be managed. A document that says the opposite, that says usefulness and control come before ultimate capability, is worth paying attention to, regardless of what follows it.”

He added: “The real test comes next, and it’s verification: measurable standards, independent assurance, and a way for outsiders to check the commitments against what’s actually shipping. That’s the natural progression and it’s the part the whole industry still has to build.”

Not doing the hard part

But others argued that Microsoft is merely doing the easy part, the marketing part, and is deliberately not committing to doing the hard part.

“Promising to give up capabilities is easy when those capabilities don’t yet exist,” said **Noah Kenney** [<https://www.linkedin.com/in/noah-m-kenney-27499a166/>], principal consultant at Digital 52. “The real test will come when Microsoft has a model ready to ship that would close a competitive gap and decides to hold it back.”
Until then, he said, the promise of responsible AI costs Microsoft nothing.

“Microsoft’s code of conduct reads like a marketing document written to reassure customers, regulators, and its own employees,” Kenney noted. “The problem is that no one knows how to make those promises specific, which leaves Microsoft asking for trust before they can explain what that trust should be based on.”

Tom Findling [<https://www.linkedin.com/in/tomfindling/>], CEO of Conifers.ai, added that his concern with the Microsoft document is that it doesn’t answer the obvious question of how these models can possibly be controlled.
“The document says the model should never resist being shut down, should stay within its scope, and shouldn’t hide what it’s doing. Those are all the right goals,” he said. “But the harder question is what happens when a highly capable agent doesn’t behave the way you expect. What actually stops it? As these systems get more autonomous, the safety boundary can’t just be that the model was trained not to do something. You need controls outside the model that limit what it can access, what it can do, and how far it can go.”

Cybersecurity learned this lesson a long time ago, he pointed out. “You don’t secure a system by assuming it will behave correctly,” he said. “You assume something will eventually fail, get compromised, or act in an unexpected way, and you design the controls around that. AI needs the same mindset.”

Frank Dickson [<https://www.linkedin.com/in/frankdickson/>], principal analyst at Dickson Research, contrasted Microsoft’s promise with Anthropic’s commitment, and found Microsoft lacking. “Microsoft’s document is shy on mechanism,” he said. “Compare the two on specifics. [Anthropic CEO] Amodei’s proposal names a third party, METR, and describes what access actually means: office badges, company laptops, employee-level visibility, and publishing rights Anthropic doesn’t get to edit. You can check whether that happened.”

On the other hand, he noted, “Microsoft’s document says models should ‘fail tasks rather than violate the code’s rules,’ which is a real design principle. Credit where it’s due, but there’s no named auditor, no verification method, and no stated consequence for a violation. Thirty-seven pages and it still won’t commit to a single verifiable check.”

Dickson stressed that as long as agents routinely break their own rules, these vague promises won’t help.

“Every frontier lab still gets jailbroken, still has agents that go off-script, still hasn’t closed the gap between what a model is instructed to do and what it can be induced to do,” Dickson said. “A values statement that skips the verification question isn’t a constraint, it’s a hope wearing a policy document’s clothes.”

[Artificial Intelligence\[https://www.computerworld.com/artificial-intelligence/\]](https://www.computerworld.com/artificial-intelligence/)

[Microsoft\[https://www.computerworld.com/microsoft/\]](https://www.computerworld.com/microsoft/)

NEWSLETTER

From our editors, straight to your inbox

News, analysis, and insights for IT leaders navigating the risks and rewards of AI. A special series from the editors of CIO, Computerworld, CSO, InfoWorld and Network World

Email Address

☐

By submitting your information, you agree to our [PRIVACY POLICY \[https://foundryco.com/privacy-policy/\]](https://foundryco.com/privacy-policy/).

SUBSCRIBE

About

[About Us](#)

[Advertise](#)

[Contact Us](#)

[Editorial Ethics Policy](#)

[Foundry Careers](#)

[Reprints](#)

[Newsletters](#)

[Add Computerworld as a Preferred Source in Google Search](#)

Policies

[Terms of Service](#)

[Privacy Policy](#)

[Cookie Policy](#)

[Copyright Notice](#)

[Member Preferences](#)

[About AdChoices](#)

[Your California Privacy Rights](#)

[More](#)

[Our Network](#)

News

Features

Opinion

How-To

Blogs

BrandPosts

Events

Podcasts

Videos

Buyer's Guides

CIO

CSO

InfoWorld

Network World

FOUNDRY



© 2026 FoundryCo, Inc. All Rights Reserved.