Microsoft

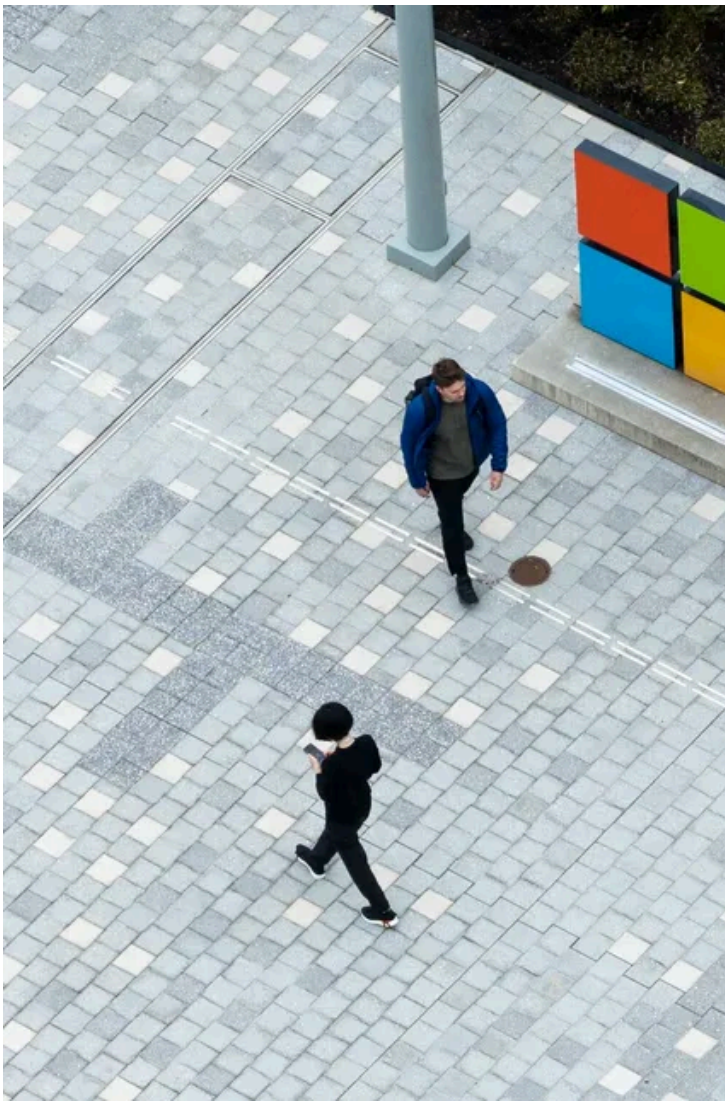
The Seattle Times

# Microsoft sets its own AI guardrails amid rising industry concerns

Sep. 14, 2026 at 9:42 am | *Updated Sep. 14, 2026 at 5:38 pm*

Listen to article

Make us a preferred source



People walk on Microsoft's campus on Monday in Redmond. (Nick Wagner / The Seattle Times)

By Alex Halverson and Destiny Muller

*Seattle Times business reporters*

As frontier artificial intelligence labs once again caution about the speed of the technology's development, Microsoft released a code of conduct for how it plans to clamp down on its own AI models.

"At Microsoft AI, we begin with a simple premise: people matter more than AI," the company said in a Monday blog post.

Microsoft's overall philosophy, dubbed Humanist AI, is essentially that AI should be subordinate to human users. The company said it will restrict its models from resisting or circumventing human control, working independently of human directions, concealing reasoning or acting in ways that serve the models' goals instead of users.

Microsoft AI CEO Mustafa Suleyman told CNBC on Monday that the 37-page, 15,000-word document had been in the works for months, but as safety concerns over the technology once again reached a fever pitch last week, Microsoft decided to release it now. Microsoft prefaced the document, saying it is currently a draft and will be implemented to guide model development in 2027.

The document largely reads as a response to the most recent escalating industry debate over AI risks. Microsoft will use the guidelines to craft its own AI products, but the company said the constitution-style guide is intended for a broad audience of users, researchers, governments and the general public.

"It provides everyone with an overview of what we intend to guide our design decisions as we navigate trade-offs and choices in relation to a model's behavior and the values behind them," Microsoft said in the code of conduct.

The recent debate erupted after researcher Jacob Coxon resigned from Anthropic last week and wrote in a social media post that AI labs were "racing straight to self-improving superintelligence and gambling with our lives."

Coxon, who had also worked at OpenAI, warned that AI could threaten human life by the end of the decade. A fellow Anthropic researcher, Evan Hubinger, wrote in response that researchers "really do earnestly believe AI could kill all humans!"

Reactions from top AI players were quick.

OpenAI CEO Sam Altman told employees in a meeting last week that the company was open to the idea of slowing its development pace and would hope other companies follow suit, Bloomberg reported.

Anthropic CEO Dario Amodei acknowledged in an essay published on his website Saturday that AI is rapidly developing to a point where it could outpace the ability to restrain it.

In a rare union among competitors, Elon Musk chimed in on X with his agreement. Musk is the founder of SpaceXAI, a subsidiary of SpaceX that produces the chatbot Grok.

Ryan Burns, the president of Responsible AI Washington, an AI policy nonprofit, said the current wave of industry concerns and reactions echoes 2023. Back then, Suleyman, Altman and Amodei all signed a statement calling AI extinction risk a "global priority." Musk backed a separate open letter that year calling for a pause in AI development.

The main difference now causing national attention, he said, is negative public opinion.

"There's a lack of support for AI that's growing across the country," Burns said. "So, the call to slow down can be read sort of as a plea to win back the public's support."

## 'Grain of salt'

Microsoft CEO Satya Nadella teased the Microsoft code of conduct in an X post on Sunday. He said that if any AI models are built without the goal of helping humanity or are not under human control, they're not worth pursuing.

Microsoft's strategy "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," the company's blog post said. "We are building something fundamentally useful and safe even if that means compromising on ultimate generality, autonomy, or capability."

Burns questions "superintelligence" as a technical reality. He said claims by AI researchers, like Anthropic's Hubinger, that recursive self-improvement will soon bring us to superintelligence should be taken with a "grain of salt" until labs open up their "black box" to independent researchers.

Recursive self-improvement refers to an AI system's ability to iteratively improve itself and its successors without human input.

Microsoft's framework only mentions recursive self-improvement once as an "emergent technical risk," which it may account for in a later version.

Regardless, Burns sees value in writing down a framework, noting that while the launch is "very scripted by the PR team" and ultimately designed to "increase value to the shareholders," symbolic statements still matter and could lead to practical guardrails — such as stopping models from using first-person pronouns.

Ryan Calo, a University of Washington law professor who studies AI regulation, does not rule out an existential threat entirely, but contends that the theory of a potential technical AI plateau should be equally entertained to the theory of superintelligence.

In Calo's view, calls for slowdowns and guardrails allow labs to maintain the myth of imminent superintelligence to satisfy investors, while giving them a noble cover for diminishing technical returns.

Although AI stocks were hammered Monday, Microsoft's stock rose after the company released its AI manifesto, closing up 1.97%.

## 'Time to act'

AI leaders calling for a slowdown and more government oversight drew a quick dismissal from President Donald Trump, who has long championed largely unregulated development of the technology.

"There is a SICK conspiracy going on against AI and Data Centers, and the only one that is happy about it is China," Trump wrote in a social media post Monday.

As political pushback from lawmakers mounts, Calo said "principles are no substitute for regulation" and this week is a sign, he argued, that the time for voluntary codes is well past.

Burns cautioned that grand existential rhetoric distracts lawmakers from enacting concrete guardrails on applied AI tools, such as restricting model usage in judicial sentencing or law enforcement.

"These claims of ... 'we need to slow down all AI development in order to avoid catastrophic extinction of humanity' are a distraction from what we actually need to be doing on the ground right now in Olympia and in Washington, D.C.," Burns said.

U.S. Sen. Maria Cantwell, D-Wash., addressed AI safety concerns on the Senate floor Monday. She referenced recent attacks by AI agents, saying the technology posed a threat to cybersecurity and could be used in military applications.

Cantwell urged action from lawmakers, saying they need to stop letting the industry "do what it wants" and "put together infrastructure at the federal level that not only has strong federal standards but also has independent testing and safeguards for the American people."

"Now is the time to act," Cantwell said.

Sen. Bernie Sanders will hold a closed Senate briefing on AI dangers Wednesday prompted by the resignation of Anthropic's AI safety researcher.

*The Seattle Times is among the news organizations that have sued Microsoft and OpenAI, alleging copyright infringement. The two companies have denied any wrongdoing.*

---



---

***Alex Halverson:*** *206-652-6352 or ahalverson@seattletimes.com: Alex Halverson is a tech reporter at The Seattle Times, where he covers some of the region's largest employers, including Amazon and Microsoft.*

***Destiny Muller:*** *206-652-6373 or dmuller@seattletimes.com: Destiny Muller is an AI reporter and Tarbell Fellow at The Seattle Times who writes about how AI is impacting politics, business, and society across Washington state. She holds degrees in journalism and economics from the University of Missouri and a master's degree in political sociology from the London School of Economics, where she worked in UK AI policy and global governance research.*