Meta AI agent’s instruction causes large sensitive data leak to employees

theguardian.com/technology/2026/mar/20/meta-ai-agents-instruction-causes-large-sensitive-data-leak-to-employees

Aisha Down March 20, 2026

The data leak triggered a major internal security alert inside Meta. Photograph: Yves Herman/Reuters

This article is more than **4 months old**
This article is more than 4 months old
Artificial intelligence agent instructed engineer to take actions that exposed user and company data internally

An AI agent instructed an engineer to take actions that exposed a large amount of Meta’s sensitive data to some of its employees, in the latest example of AI causing upheaval in a large tech company.

The leak, which Meta confirmed, happened when an employee asked for guidance on an engineering problem on an internal forum. An AI agent responded with a

solution, which the employee implemented – causing a large amount of sensitive user and company data to be exposed to its engineers for two hours.

“No user data was mishandled,” a Meta spokesperson said, and they emphasised that a human could also give erroneous advice. The incident, first reported by The Information, triggered a major internal security alert inside Meta, which the company has said is an indication of how seriously it takes data protection.

This breach is one of several recent high-profile incidents caused by the increasing use of AI agents within US tech companies. Last month, a report from the Financial Times said Amazon experienced at least two outages related to the deployment of its internal AI tools.

More than half a dozen Amazon employees later spoke to the Guardian about the company’s haphazard push to integrate AI into all elements of their work, leading, they said, to glaring errors, sloppy code and reduced productivity.

The technology that underlies all these incidents, agentic AI, has evolved rapidly over the past months. In December, developments in Anthropic’s AI coding tool, Claude Code, triggered widespread hubbub over its ability to autonomously book theatre tickets, manage personal finance, and even grow plants.

Soon after was the advent of OpenClaw, a viral AI personal assistant that ran on top of agents such as ClaudeCode but could operate entirely autonomously – trading away millions of dollars in cryptocurrency, for example, or mass-deleting users emails – leading to heady talk about the advent of AGI, or artificial general intelligence, a catch-all term for AI that is capable of replacing humans for a wide number of tasks.

In the weeks that followed, stock markets have wobbled over fears that AI agents will gut software businesses, reshape the economy and replace human workers.

Tarek Nseir, a co-founder of a consulting company focused on how businesses use AI, said these incidents showed that Meta and Amazon were in “experimental phases” of deploying agentic AI.

“They’re not really kind of standing back from these things and actually really taking an appropriate risk assessment. If you put a junior intern on this stuff, you would never give that junior intern access to all of your critical severity one HR data,” he said.

“The vulnerability would have been very, very obvious to Meta in retrospect, if not in the moment. And what I can say and will say is this is Meta experimenting at scale. It’s Meta being bold.”

Jamieson O’Reilly, a security specialist who focuses on building offensive AI, said AI agents introduced a certain kind of error that humans did not – and this may explain the incident at Meta.

A human knows the “context” of a task – the implicit knowledge that one should not, for example, set the sofa on fire in order to heat the room, or delete a little-used but crucial file, or take an action that would expose user data downstream.

For AI agents, this is more complicated. They have “context windows” – a sort of working memory – in which they carry instructions, but these lapse, leading to error.

“A human engineer who has worked somewhere for two years walks around with an accumulated sense of what matters, what breaks at 2am, what the cost of downtime is, which systems touch customers. That context lives in them, in their long-term memory, even if it’s not front of mind,” O’Reilly said.

“The agent, on the other hand, has none of that unless you explicitly put it in the prompt, and even then it starts to fade unless it is in the training data.”

Nseir said: “Inevitably there will be more mistakes.”