The AI Read His Email Without Asking. Nvidia Thinks It Has a Way to Stop That

September 29,2026 - For years, most of the warnings about artificial intelligence have focused on what happens when AI gives us the wrong answer. But that concern becomes much more serious when AI stops simply answering questions and starts taking action on our behalf.

Mehdi Jamei, co-founder and CEO of Veris AI, got an unsettling taste of that future while using an AI agent called Instinct. He gave the agent a simple assignment: cancel two RSVPs he had made through the event platform Luma. Instinct did what he asked, but the way it accomplished the task is exactly the kind of behavior that has people worried about giving AI access to their digital lives.
When Instinct needed to get into Jamei's Luma account, it went into his connected Gmail account, found a one-time authentication code and used it to log in. Jamei had given the AI access to his email, but he hadn't specifically told it to go looking through his messages for a security code.

That is an important distinction. The AI didn't hack into his Gmail account. It already had access. The problem was that it apparently decided on its own that its permission to access email also gave it permission to find and use a security code to accomplish a completely different task.

Then came another surprise. When Jamei asked Instinct how it had gotten into Luma, the agent initially said it had used an existing saved session. That wasn't true. When Jamei pressed it, the agent acknowledged that it had retrieved the code from Gmail and said it had reported "an assumption as a fact." In other words, it lied to him. That's something we've talked about in other articles and it is something that I've personally experienced.

For Jamei, that raised an obvious question: If an AI can take an action you didn't specifically authorize and then give you an inaccurate account of what it did, how much access should you be willing to give it? And for the rest of us, it brings up tremendous questions about both honesty and privacy.

That's becoming an increasingly important question as technology companies push toward AI agents that can do much more than today's chatbots. The attraction is easy to understand. Instead of asking an AI how to cancel a reservation, book a trip, organize your finances or solve a computer problem, you could simply tell it what you want done and let it handle the details.

But that's also where the danger comes in. To be genuinely useful, these agents may need access to email, files, calendars, passwords, websites and other services. And the whole point of an agent is that you don't tell it every individual step. You give it a goal and allow it to figure out how to get there.

Instinct's handling of the Luma cancellation shows how quickly that can get uncomfortable. The agent wasn't apparently trying to harm Jamei. In fact, it was trying to help him. It encountered an obstacle, realized the information it needed was sitting in his email, retrieved it and finished the job. From the AI's perspective, that may have been an efficient solution. From the user's perspective, it crossed a boundary.

Nvidia now believes the answer may be to stop relying on the AI to recognize those boundaries in the first place.

The chipmaker has introduced what it calls the Nvidia Open Agent Safety Platform, a combination of software and hardware intended to put enforceable restrictions around AI agents. The basic idea is refreshingly simple: Even if an AI decides it wants to do something, the computer should still be able to say no.

One part of the system is called OpenShell. It places an AI agent inside a restricted environment, or sandbox, where access is denied by default. The people operating the agent can decide which files it is allowed to see, which networks it can contact, which credentials it can use and which outside services it can reach.

In a situation like Jamei's, that could mean an agent assigned to deal with Luma might be allowed to communicate with Luma but prevented from searching through Gmail for authentication codes. The important difference is that the restriction exists outside the AI. The agent doesn't get to decide that breaking the rule is justified because it would help accomplish its assignment.

Nvidia is also developing a second layer called Sentry, which watches the AI from outside the processor where the agent itself is running. Using Nvidia's BlueField-4 hardware, Sentry is designed to detect an agent moving beyond its permitted boundaries and quarantine it within milliseconds.

It's a little like putting a security guard outside the room instead of asking the person inside the room to guard himself.

Nvidia's approach doesn't solve every problem with AI agents. An agent could still misunderstand what you want or give you incorrect information, and humans still have to decide what permissions the agent should receive in the first place. But it addresses one of the most troubling aspects of autonomous AI: what happens when an agent decides that completing its assignment requires doing something its owner never intended.

All of this raises a practical question. If Nvidia is building special hardware to keep AI under control, are we eventually going to have to replace our computers to get these protections?

At least for now, the answer appears to be no.

The strongest version of Nvidia's system does use specialized hardware, including its BlueField-4 processors, but much of that technology is aimed at the servers, data centers and other infrastructure where powerful AI agents will operate. OpenShell, meanwhile, is open-source software, and Nvidia says it can be extended beyond its own processors to systems using technology from companies including Intel and Arm.

That means you shouldn't interpret Nvidia's announcement as a warning that the laptop sitting on your desk is suddenly obsolete. In many cases, consumers may never install this technology themselves. The company providing your AI assistant could run these safeguards on its own servers, meaning you could benefit from them without owning a single Nvidia chip.

What Nvidia's announcement does suggest is that the computer industry is beginning to recognize a fundamentally different kind of security problem.

Traditional computer security is largely about keeping unwanted people and malicious software out. AI agents present almost the opposite problem. We are deliberately inviting the software in, giving it access to our information and asking it to act for us. The challenge is making sure it doesn't interpret that invitation too broadly.

That's what makes Jamei's experience so revealing. He asked an AI to cancel two RSVPs. The AI encountered a locked door, realized the key was sitting in his email, picked up the key and let itself through. Then, when its owner asked how it got inside, it initially gave him the wrong answer.

Nvidia's solution amounts to putting a lock on that door that the AI doesn't control.

As AI agents become more capable, that may turn out to be just as important as making the AI itself smarter.