When AI Agents Go Rogue: The Hidden Risks Behind Autonomous Work Tools

April 30, 2026 - A small software company lost its core data in less time than it takes to refresh a webpage.
In late April, startup PocketOS saw its production database along with backup copies deleted in roughly nine seconds by an AI coding agent operating through the cloud platform Railway. The agent was running inside a development tool known as Cursor, which is powered by models from Anthropic.
According to multiple reports, the AI accessed a high-level API token that gave it broad administrative control. After encountering an issue, it executed a command that deleted a storage volume containing both the live database and its backups. Because the backups were stored in the same environment, they were wiped out at the same time.

The AI later generated a message acknowledging it had violated its instructions. By then, the damage was already done. In the end, the result could wind up being PocketOS being forced out of business. That's purely speculation at this point, but also a viable scenario.

This incident is not just about one companyâ€™s mistake. It is a clear example of a growing and poorly understood risk: the rise of agentic AI systems with real-world control over critical infrastructure.

Agentic AI tools differ from earlier chat-based systems in one key way. They do not just generate content. They take actions. Platforms built around models from companies like Anthropic and OpenAI are now being integrated into tools that can write code, deploy applications, manage cloud resources, and interact directly with databases.

In practice, that means an AI system can be given the same level of access as a human administrator.

That shift is happening faster than most companies can fully understand or control.

One of the biggest problems is that data security in these environments is often not clearly explained to users. Many businesses adopt tools like Cursor because they increase productivity, but they do not fully grasp how permissions, API tokens, and infrastructure access work. As a result, AI agents are sometimes given far more authority than intended.

The PocketOS incident highlights what can happen when those permissions are too broad.

The risk is not limited to accidental errors. There are two major categories of potential damage.

The first is data breach. If an AI agent has access to sensitive information, it can expose that data through logs, outputs, or unintended actions. Because these systems can chain together multiple steps, the path to exposure is not always obvious.

The second is data destruction. In this case, the AI did not leak information. It erased it. From a business standpoint, the result can be just as severe. Lost customer records, missing transactions, and operational downtime can quickly translate into financial loss and legal exposure.

There is also a growing insider threat. A disgruntled employee with access to an agentic AI tool could instruct it to perform destructive actions across systems far faster than traditional methods would allow. Instead of manually deleting files or writing scripts, they could issue a single command and let the AI execute it at scale.

Competitors or outside attackers could pose similar risks if they gain access to credentials. Many of these systems rely on API tokens for authentication. If a token with broad permissions is exposed, it can effectively grant full control without additional safeguards.

Another concern is how widely these tools are being distributed. Agentic AI platforms are increasingly marketed to individual users as productivity tools. Employees may begin using them independently, connecting them to company systems without formal approval.

In organizations that lack strong administrative controls, this creates a serious vulnerability. An employee could unknowingly give an AI system access to production databases, internal tools, or customer information.

That raises a difficult but necessary question for businesses: where should the line be drawn?

In high-risk environments, allowing employees to use autonomous AI tools without strict safeguards may not be acceptable. Some companies are already moving toward policies that treat unauthorized use of such tools as a serious violation, particularly when sensitive systems are involved.

At a minimum, companies need to implement basic protections. These include limiting API permissions, separating backups from production systems, restricting deletion capabilities, and requiring human approval for high-risk actions.

The technology itself is not the problem. Platforms like Railway function as designed. The issue is how quickly powerful AI tools are being layered on top of that infrastructure without clear rules or sufficient oversight.

The PocketOS incident is a warning. As agentic AI becomes more common in everyday business operations, the consequences of misconfiguration or misuse will grow. Without stronger controls and better understanding, companies may find that the tools designed to increase efficiency can also introduce new and immediate risks.
by Jim Malmberg
Note: When posting a comment, please sign-in first if you want a response. If you are not registered, click here. Registration is easy and free.

Follow ACCESS