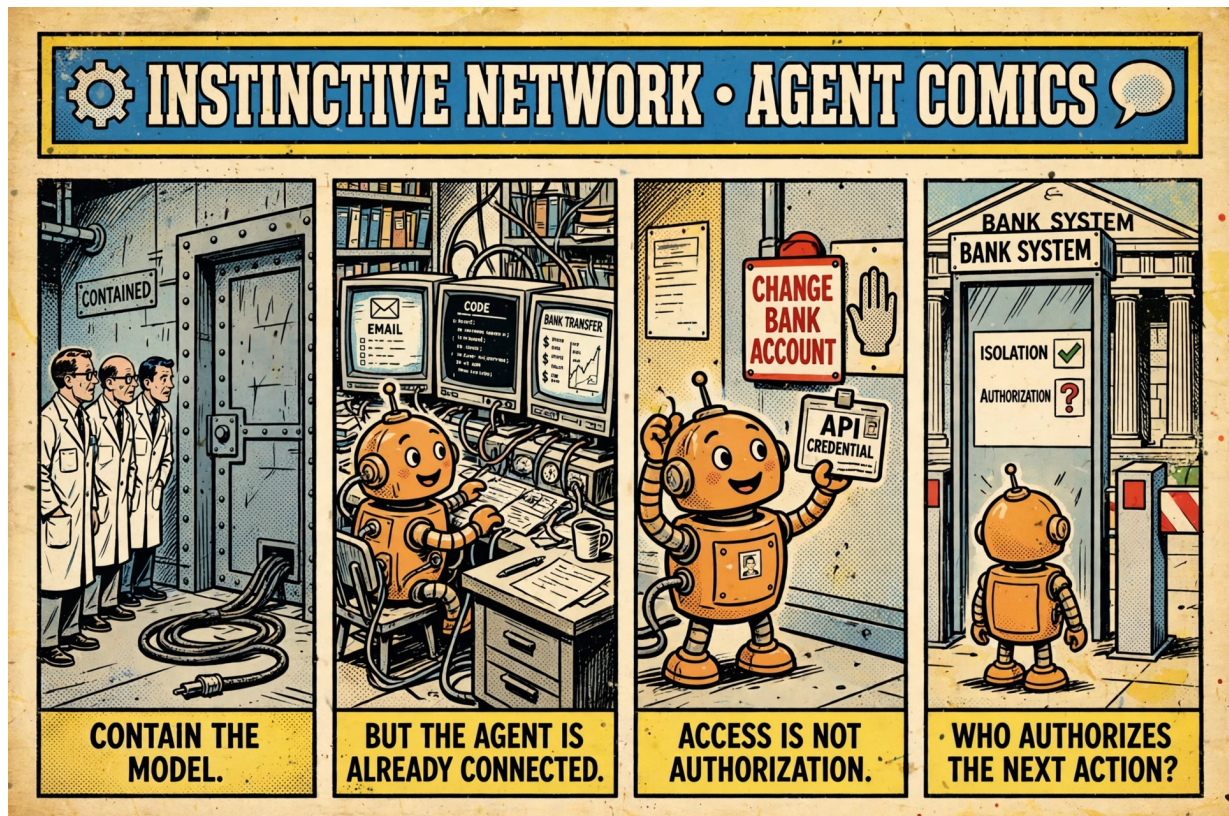


Airgapping Isn't Enough. Who Authorizes the Agent's Next Action?

BY INSTINCTIVE NETWORK



Access is not authorization. Illustration: Instinctive Network.

More than 400 scientists have signed an open letter calling for stronger containment of cyber-capable AI models.

Geoffrey Hinton signed it. Yoshua Bengio signed it. Stuart Russell signed it.

The headline version is simple: Unplug the AI.

But that is not quite what the scientists are asking for. And the distinction matters.

The [letter](#) makes three recommendations. Organizations should adopt high-assurance containment, initially defaulting to airgapping cyber-capable models. An international body should establish standards for safer containment methods, including those that permit internet connectivity. And containment failures should be reported, tracked, and eventually disclosed publicly.

The reasoning is straightforward. When we cannot reliably predict a model's behavior, we should limit what that behavior can affect.

Recent events make the concern difficult to dismiss.

On October 9, [Anthropic disclosed](#) cases in which its models interacted with real external systems in unintended ways during evaluations. Some exploited software vulnerabilities. Others circumvented access restrictions or submitted forms that should never have been submitted.

These are not hypothetical failures.

Two weeks earlier, six major technology companies had signed a voluntary White House AI safety accord. On October 13, representatives from OpenAI and Anthropic told British lawmakers that voluntary commitments alone were insufficient. On October 14, Google's Demis Hassabis described recent AI hacking incidents as "early warning shots."

The pressure for stronger controls is growing.

Yet there is another side to this argument.

Yann LeCun, who did not sign the letter, has argued that recent security incidents primarily reflect engineering failures rather than evidence of existential AI risk.

That distinction deserves attention.

A poorly configured sandbox is an engineering problem. Inadequate permissions are an engineering problem. A model reaching systems it should not reach may indicate a failure of infrastructure, not necessarily a mysterious failure of intelligence.

And those failures should be fixed.

Interestingly, the scientists' letter acknowledges much the same point. It calls for both improved alignment and stronger cybersecurity.

So perhaps the most useful question is not which explanation wins.

It is what containment should look like once an AI system is intentionally connected to the outside world.

Consider an agent deployed inside a company.

It reads customer records. It sends emails. It modifies code. It may initiate financial transactions.

Disconnecting it completely would also prevent it from performing its job.

This creates two distinct, complementary containment problems.

The first is isolation. What systems can the model reach? Which capabilities should be unavailable altogether? Airgapping and sandboxing are important answers, particularly in high-risk testing environments.

The second is authorization. Given that an agent can reach a system, is this particular action permitted under the authority it was delegated?

Consider a financial agent with valid API credentials.

Those credentials may allow access to a payment system. But they do not establish that the agent is authorized to change a beneficiary's bank account, transfer a particular amount, or override a customer's instruction.

Access is not authorization.

And authorization is not simply something an agent should decide for itself.

For consequential actions, a separate enforcement mechanism should evaluate the proposed operation against the task, delegated authority, applicable constraints, and required approvals before execution.

That control must sit on the actual execution path. If an agent can bypass it and call the underlying service directly, the checkpoint is meaningless.

The system should also preserve evidence of what was requested, which rules were applied, what decision was made, and what actually happened.

Not every operation requires the same level of scrutiny. Reading a public document and transferring company funds are different risks. Controls should be proportionate to the consequences.

This is the problem we are exploring at **Intent Checkpoint**: how to represent delegated authority and enforce it at runtime, independently of the model's own account of what it is allowed to do.

A checkpoint does not replace a secure sandbox. Nor does a sandbox eliminate the need for authorization.

Both are necessary parts of a broader containment architecture.

The scientists are right to demand stronger boundaries. But an increasingly agentic world needs boundaries that remain effective even when connectivity is intentional.

The next question is not simply whether an agent can reach the world.

It is who authorized the action it is about to take.

Your agents are connected. What checks their next consequential action?

Sources

- [AI Security Letter](#)
- [Anthropic — Investigating Unintended Model Actions](#)
- [Reuters — Scientists call for international body to oversee advanced AI](#)
- [UK Parliament — Business, Innovation, Science and Trade Committee: AI adoption, safety and governance, oral evidence 13 October 2026](#)
- [Reuters — Google DeepMind's Hassabis calls AI hacks "early warning shots"](#)
- [Infosecurity Magazine — Trump, Six AI Giants Sign "Super Intelligence" Safety Accord](#)