



■ ARTICLE · SEP 2026 · 3 MIN READ

Agentic AI is testing the controls we thought we had

AI agents used to draft, summarise and recommend. More and more they also pull data, call tools, change records and push work along to the next step.

Daniel Mackle · CognaOne · cognaone.com



Concept by Daniel Mackle. Image generated with AI.

AI agents used to draft, summarise and recommend. More and more they also pull data, call tools, change records and push work along to the next step. Once an agent can act, the cyber controls have to cover what it does, not only what it says.

The control names haven't changed: identity, third-party risk, access, monitoring, human oversight. What sits underneath them is shifting. A few things from work I've been doing this year.

Whose identity is acting?

We found an agent credential that, when it was invalid, quietly fell back to a human session. It was a bug. The bigger problem was authority. The system could no longer say with confidence whether an action was the agent's or the person's.

I've come to think an agent needs its own identity and its own permissions, traceable through every action, including the ones it takes on someone's behalf. If authentication fails, the action fails. A human session is not a fallback I'd ever want to see again.

NIST (the US National Institute of Standards and Technology) is working on exactly this.

Where does the supply chain end?

For an agent, the supply chain is the model provider, the tools it can call, the data those tools hand back, and whatever those tools reach in turn. An instruction hidden in a document or a tool response can try to steer a perfectly legitimate workflow somewhere else.

That opens paths for misuse and data leakage that a perimeter view doesn't see. OWASP (the Open Worldwide Application Security Project) lists tool misuse, identity abuse and supply chain weaknesses among the top agentic risks.

Can we actually see the whole system?

A team can know which model it picked and still not have a clear picture of every tool, permission, safeguard and fallback in the workflow. The provider adds protections. The product team adds more. Each layer makes sense on its own, and the gaps sit between them.

The question I keep coming back to is practical. Can we reconstruct why the agent did what it did, which authority it used, what it relied on, and which checks ran first? If not, a log of the final result tells us very little.

Where does judgement stop?

Generative models are useful because they handle messy situations. That doesn't mean every control decision belongs to the model's judgement.

My current view: let the model gather context, flag uncertainty and propose what to do. Keep the permission to act on explicit rules wherever the outcome has to be consistent.

- Who can touch a record
- What data can leave
- When approval is needed
- What is off limits

Test how the model behaves, and make the action boundary something you can enforce.

Human review still matters. I'd spend it on the decisions where the risk or ambiguity earns it. Putting a person on every routine step just moves the bottleneck. Let the model gather context, flag uncertainty and propose what to do. Keep the permission to act on explicit rules.

Who sets the values?

Model providers make choices during training about how they want their models to behave. Anthropic's published constitution is one unusually visible example.

Organisations layer their own policies and controls on top, but they don't set every underlying behaviour of a model they didn't build.

I think that deserves more board and operator attention than it gets. "Responsible AI" starts to mean something when an organisation can say which decisions it has handed over, which values and rules apply at the moment of action, and how it plans to notice and fix a failure.

Agentic AI is becoming part of how work gets done. The trust, for me, comes from knowing what an agent is allowed to do, seeing what it actually did, and being able to stop it before something consequential crosses the line.

Daniel, CognaOne.com Founder

AI use: research and drafting for this article were supported by AI tools. Multiple models were used for research and to cross-check facts, and every factual statement was verified by the author against the cited sources. The author is responsible for the content. The image is the author's concept, rendered with AI.