

Blog 6: The Five Questions Every AI Agent Must Answer Before You Trust It

From the CSA Agentic AI Summit Series

By TekSol Global | Cybersecurity & AI Practice

In April 2026, something unusual happened at RSA Conference, one of the biggest cybersecurity events in the world.

Four major security companies, CrowdStrike, Cisco, Microsoft, and Splunk, each gave separate keynote talks. They hadn't coordinated. They didn't share notes beforehand. But independently, each of them landed on the same core message:

AI agents need a completely new kind of trust framework.

And remarkably, all four described the same five elements.

At the CSA Agentic AI Summit, one speaker walked through exactly what those five elements are and why every AI agent should have to answer five fundamental questions before anyone should trust it.

The Framework: Five Simple Questions

These five questions come from what's called the **Agentic Trust Framework (ATF)**, an open-source specification that has since been adopted and expanded by Microsoft and others.

The questions sound simple. But they're deceptively important.

Question 1: Who Are You?

This is about **identity**.

Every AI agent needs an unforgeable credential, a verifiable, unique way of proving what it is. Not just a name. Not just "Agent-7." An identity that can be cryptographically verified, meaning no one can fake it.

Why does this matter? Because if you don't know who (or what) is acting in your systems, nothing else works. You can't monitor unknown entities. You can't govern them. You can't contain them if something goes wrong.

Think of it like this: before you let someone into a restricted area, you check their ID. You don't just take their word for it. The same principle applies to AI agents.

Question 2: What Are You Doing?

This is about **behavioural monitoring**.

An AI agent should be watched continuously. Its behaviour, what systems it accesses, what actions it takes, what data it reads, what it sends out, should be observed in real time.

Why? Because an agent that starts doing something unexpected needs to be caught quickly. Maybe it's been compromised. Maybe there's a bug. Maybe it was given bad instructions. Whatever the cause, abnormal behaviour is a signal that something has gone wrong.

One security expert at the summit said it plainly: *"Behaviour is the decisive signal. The only way to tell benign from malicious at machine speed."*

You can't fully audit an AI's internal logic. But you can watch what it does.

Question 3: What Are You Consuming?

This is the first half of **data governance**.

What data is the agent allowed to read? What documents, databases, emails, and records can it access?

This should be tightly controlled. An agent shouldn't be able to read everything just because it technically has the capability. It should only be allowed to consume data that's relevant to its specific task.

Think of it as: an employee in the marketing department doesn't need access to the engineering source code. Similarly, an AI agent doing customer support doesn't need access to employee salary records.

Question 4: What Are You Producing?

This is the second half of **data governance**.

What is the agent outputting? What information does it include in its responses? What data is it writing to files, sending in emails, or posting to external services?

This matters enormously. An agent might inadvertently include confidential information in a document it creates. It might send data to an external party, it shouldn't. It might log sensitive details in a place where they can be seen by the wrong people.

Governing what goes *out* of an agent is just as important as governing what goes *in*.

Question 5: What Are You Allowed to Do?

This is about **least privilege access**, one of the most important principles in security.

An agent should only be able to take actions that are explicitly authorized for its specific role. And those permissions should be **just-in-time** (given only when needed), **just-enough** (only what's required for this task), and **just-long-enough** (they expire when the task is done).

Cisco's framing at the RSA Conference was sharp: move from *access control* to **action control**.

Traditional security focused on: "Can this entity log into this system?" The new approach asks: "Can this entity perform this specific action, at this specific time, for this specific purpose?"

That's a much more precise and much more secure way of thinking.

Why These Five Questions Together?

Each of these questions matters on its own. But together, they form a complete picture of a trustworthy AI agent:

Question	What It Addresses
Who are you?	Identity verification
What are you doing?	Behavior monitoring
What are you consuming?	Input data control
What are you producing?	Output data control

Question	What It Addresses
What are you allowed to do?	Permission boundaries

If an agent can answer all five questions satisfactorily, and those answers are enforced automatically, you have a fundamentally safer system.

The Industry Woke Up

Here's why the summit speaker found the RSA Conference moment so significant.

These five questions aren't new. Security professionals have been thinking about them for years. But for a long time, the industry was focused on other things, cloud security, endpoint security, and network security.

When four major vendors independently showed up at RSA, and all said the same thing? That was the industry collectively saying: *this is the moment. AI agent trust is no longer a niche problem. It's the central problem.*

And the remarkable thing is that Microsoft, within weeks of those keynotes, publicly launched a toolkit specifically for AI agent governance, with all packages mapped to the same five elements.

The framework wasn't ahead of its time. The industry just caught up.

What This Means in Practice

For a company deploying AI agents, answering these five questions means putting in place:

1. **An identity system for AI agents** (unique credentials that can't be faked)
2. **A monitoring system** (real-time logs of agent behaviour)
3. **Data access policies** (rules about what data agents can read)
4. **Output controls** (rules about what data agents can share or write)
5. **Permission systems** (fine-grained, time-limited authority for specific actions)

None of these are impossibly complex. They're extensions of security practices that already exist, just applied specifically to AI agents, which have unique characteristics that require updated thinking.

The next blog will look at another crucial piece of this puzzle: **governance**, the policies, structures, and human oversight that need to exist above and around these technical controls.

Next up → Blog 7: Closing the Authority Gap - Who Actually Controls Your AI Agents?