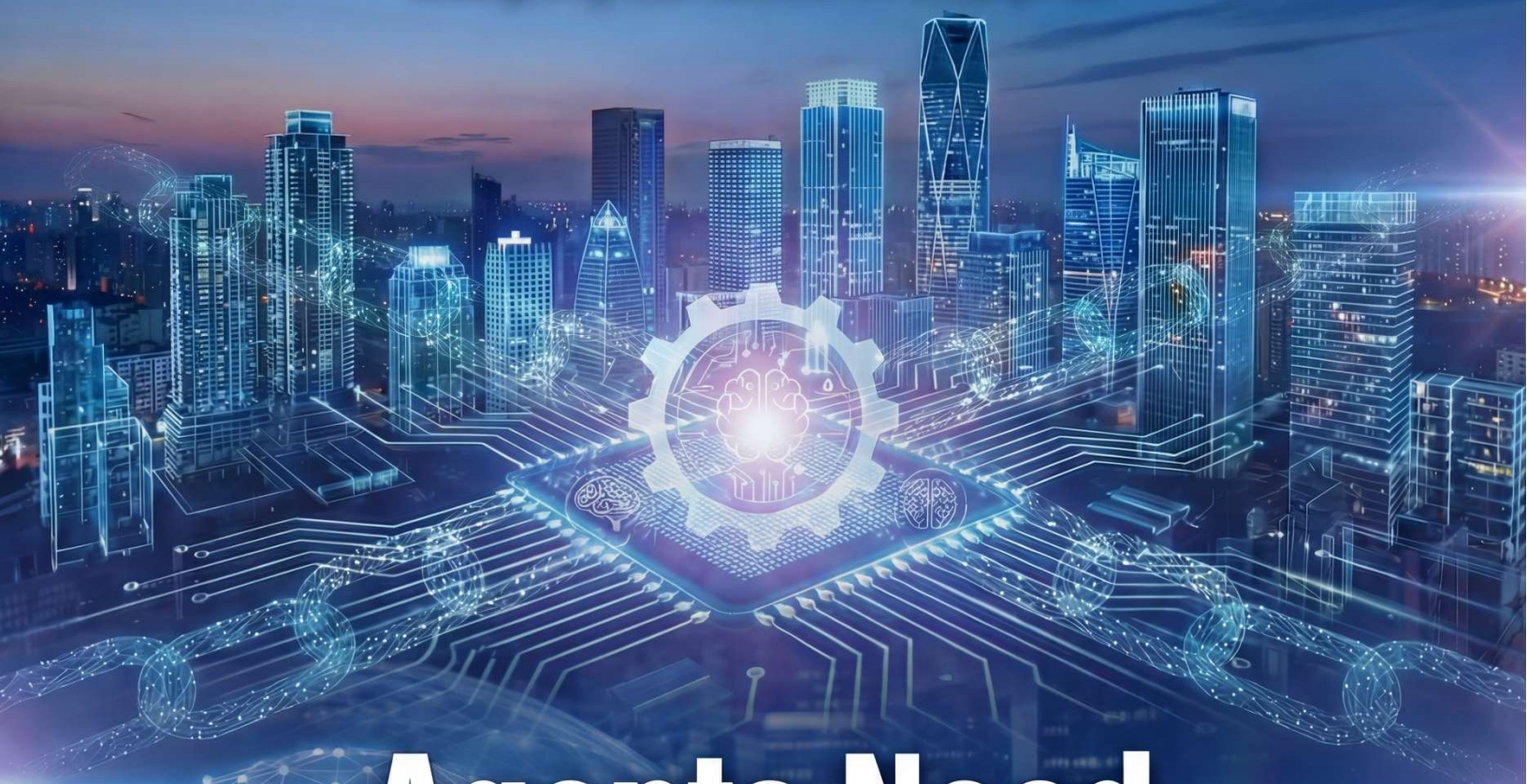
Designing the Human-AI Enterprise



Agents Need Runtime Supervision

Curtiss J. Shupe

05

Agents Need Runtime Supervision

What it takes to direct, observe, and interrupt agentic execution before errors compound in production.

The move from assistants to agents changes the enterprise problem. An assistant responds to prompts. An agent plans how it will do its work, make calls to tools, execute a set of tasks, persist through tasks, and execute an action inside a workflow. Anthropic's guidance on this point is rather clear. Also, it argues that good agent systems come out of simple and composable patterns rather than large frameworks designed to impress other framework creators. This is important because the enterprise is no longer concerned with how well a system talks. It's concerned with whether it's able to execute limited action inside a workflow. Once this is the case, control is no longer a peripheral concern; it is the concern.

This process is called in this article supervised agentic execution. Again, the terminology is not from the source material but from this series of articles. However, the argument is well-supported. The agents are not only working independently but also require human intervention, visibility of what they are doing, and evidence that the system is working. The Microsoft Work Trend Index for 2025 is relevant in this case because it talks of an enterprise that is increasingly becoming a human-agent team, intelligence on demand, and employees becoming "agent bosses." The data is based on survey results from 31,000 employees in 31 countries, LinkedIn labor data, and Microsoft 365 signals. The writing is already clear in the case of agents. They are not in the demo lane anymore.

This is why supervision cannot be casual anymore. The traditional automation model often comes with a set of predefined steps. An agent is more flexible, but this is also why it's more difficult to trust it without structure. An agent can choose its steps, make calls to tools, react to changes in context, attempt actions again, and continue working on a series of tasks. Anthropic does not refer to "supervised agentic execution" in its text, but its production guidelines point to the same conclusion: once planning, using tools, and multi-step execution enter the picture, structure is required to make it reliable. The enterprise will need to define how the agent is controlled, what tools it can use, what it remembers, how it recovers from errors, and where humans enter in before a small error becomes a bigger one.

NIST's AI Risk Management Framework helps support this from a governance perspective. It's a framework that's all about govern, map, measure, and manage through the lifecycle of AI, and its Core states that human oversight processes should be defined, assessed, and documented. That's not just documentation for the compliance binder. That's saying supervision has to be embedded in the way the operation is run. It has to be visible. It has to be verifiable. It has to be linked to risk tolerance, intervention rights, and other factors. If an organization can't demonstrate where the point is for a human to override the system, what indicators there are for drift, how those are tracked, and how performance is measured over time, then they don't really have a supervision model. They just have optimism in a business casual.

Curtiss J. Shupe

The same issue is discussed by OWASP's Agentic Security Initiative from the security side. Open Worldwide Application Security Project (OWASP) considers autonomous agents and multi-step AI workflows as a special security domain, as it significantly increases the exposure surface. The exposure surface includes calls to tools, access to external systems, memory, and runtime states, which are not typical misuse for a chatbot. However, once the agent can interact across systems, inadequate supervision is no longer a simple operational problem. This is a governance and exposure problem. This means that runtime supervision is not a simple approval button, nor can it be a simple screen that someone glances at after the run is complete.

A good model for a strong supervised agentic relationship has five parts. The first is setting goals and constraints. This is where humans need to set the goal, the bounds, and the priorities. The second is tool and permission control. This is where agents shouldn't be granted access simply because it's convenient. The third is runtime observability. This is where, without the ability to observe tool calls, retries, state changes, errors, and outputs, the enterprise isn't really supervising the run. The fourth is intervention and override. This is where supervision isn't really supervision unless humans can pause, redirect, approve, reject, or stop the run before the problem gets out of hand. The fifth is evaluation and evidence. This is where supervision needs to be supported by measurement, including task success, fallbacks, tool errors, escalations, and recovery quality. This five-part model is the synthesis of this article, and it aligns very well with the source material from Anthropic, NIST, and OWASP.

This is where a lot of organizations still fool themselves. They assume that having a human notionally "in the loop" is enough. It isn't. If the human can't see what happened, can't reconstruct what the agent did, can't intervene before completion, or can't tell whether the control system is working, then the enterprise isn't supervising runtime execution in any meaningful way. It's just keeping a human nearby in case someone needs blame assigned later. Microsoft's broader human-agent-team framing makes this more urgent. If digital labor is becoming part of normal work, the enterprise has to know which machine actions are directed, which are supervised, which are interruptible, and which still require explicit human authorization. That's what operating discipline looks like when the machine can actually do things.

The goal isn't to slow agents down. The goal is to let them do useful work without letting them operate blindly. Organizations that get this right won't just have more agent activity. They'll have better runtime harnesses: systems that let agents work across meaningful workflows while keeping human direction, observability, intervention, and accountable control exactly where they belong. That's what supervised agentic execution is for. It's the difference between experimenting with agents and actually knowing how to run them at enterprise scale.

Curtiss J. Shupe