

AI agent governance checklist

The must-haves of any agent governance framework

The challenge

- 🗨 Agents are unpredictable, incredibly fast, and easily directable
- 🗨 Agents require different governance than humans
- 🗨 Agents are imperfect (like humans), and need oversight just like humans do
- 🗨 Policies enforced through the agent can be revised or overridden

The solution

- 👍 The ability to distinguish each agent's actions
- 👍 A way to limit authorization scope and timebox access once a task is complete
- 👍 A comprehensive, uneditable record of every agent action and tool call
- 👍 A chain of actions to trace agent behaviors back to the human in control

The 4 pillars that keep AI agents in check

Identity

- ❑ Agent identities are instance-bound and delegation-aware
- ❑ Cryptographic identities are tied to a specific task, person, or delegation chain
- ❑ On-behalf-of chains are managed through infrastructure the agent can't touch

Authorization

- ❑ Agents are scoped per task with short-lived permissions
- ❑ Agent authorizations have hard boundaries
- ❑ Agents have the same limits as the human they're acting on behalf of

Observability & explainability

- ❑ In-depth transcripts record every input, output, and tool call
- ❑ Transcripts must include the models and tool versions used
- ❑ Transcripts should be captured by infrastructure that agents can't access or influence

Accountability & control

- ❑ Clear chains that trace every agent action back to the human responsible
- ❑ Agent-level kill switch that is fast and precise enough to contain the damage

See AI agent governance in action

Skip building a framework. Start using one. The Redpanda Agentic Data Plane is a single governed gateway that sits between your agents and your systems, with all four pillars built in. It lives inside your own cloud and governs any model, API, or database you use. It's AI agent governance, right out of the box.

Explore the Redpanda Agentic Data Plane

Or [watch the demo](#) to see how it works