FORM 10 OF 10 · MULTI-AGENT SYSTEMS

Guardian agent

Its job is to watch, check or stop other agents. Gartner names three roles: reviewers that check output, monitors that track actions, and protectors that adjust or block an action as it happens.

NOT WHEN	USE IT WHEN
It shares the instructions and the blind spots of the agent it checks. It then adds cost and no independence.	**Agents act on customers, money or regulated records, or when more agents are running than people can review.**

IT NEEDS

Independence from what it checks, the authority to block and not only to report, and a record of its own.

EXAMPLE

Before a customer letter drafted by another agent is sent, the guardian checks it against the disclosure rules and holds the ones that fail.

BUILD IT ON KATONIC

01 **Control Room, Guardrails.** Choose **Create Profile**. A profile holds three kinds of rail: **Input rails** “applied before the prompt reaches the LLM”, **Output rails** “applied to the LLM response before it leaves the system”, and **Tool rails** “applied to MCP tool-call arguments”. Each rail acts by Redact, Block or Warn.

02 In each agent’s **Guardrails** section, pick that profile. The runtime reads the profile, not the fields on the agent.

03 **Control Room, Policies.** Choose **Create Policy** to require approval for a tool call, or to block it, by tool and scope. Policies run in priority order.

04 **Control Room, Approvals.** Held tool calls wait here for a named approver, with an audit trail of who decided.

05 In an agent’s **Guardrails** section, use **Test against a prompt** to see which guardrails would trip before you deploy.

ALSO CALLED Reviewer, critic, evaluator, monitor. Google ships a sample agent that monitors other agents in production. Gartner expects guardian agents to take 10 to 15 percent of the agentic AI market by 2030.

One of ten forms in “Types of AI agents”. KATONIC.AI/RESOURCES/WHITE-PAPERS/TYPES-OF-AI-AGENTS