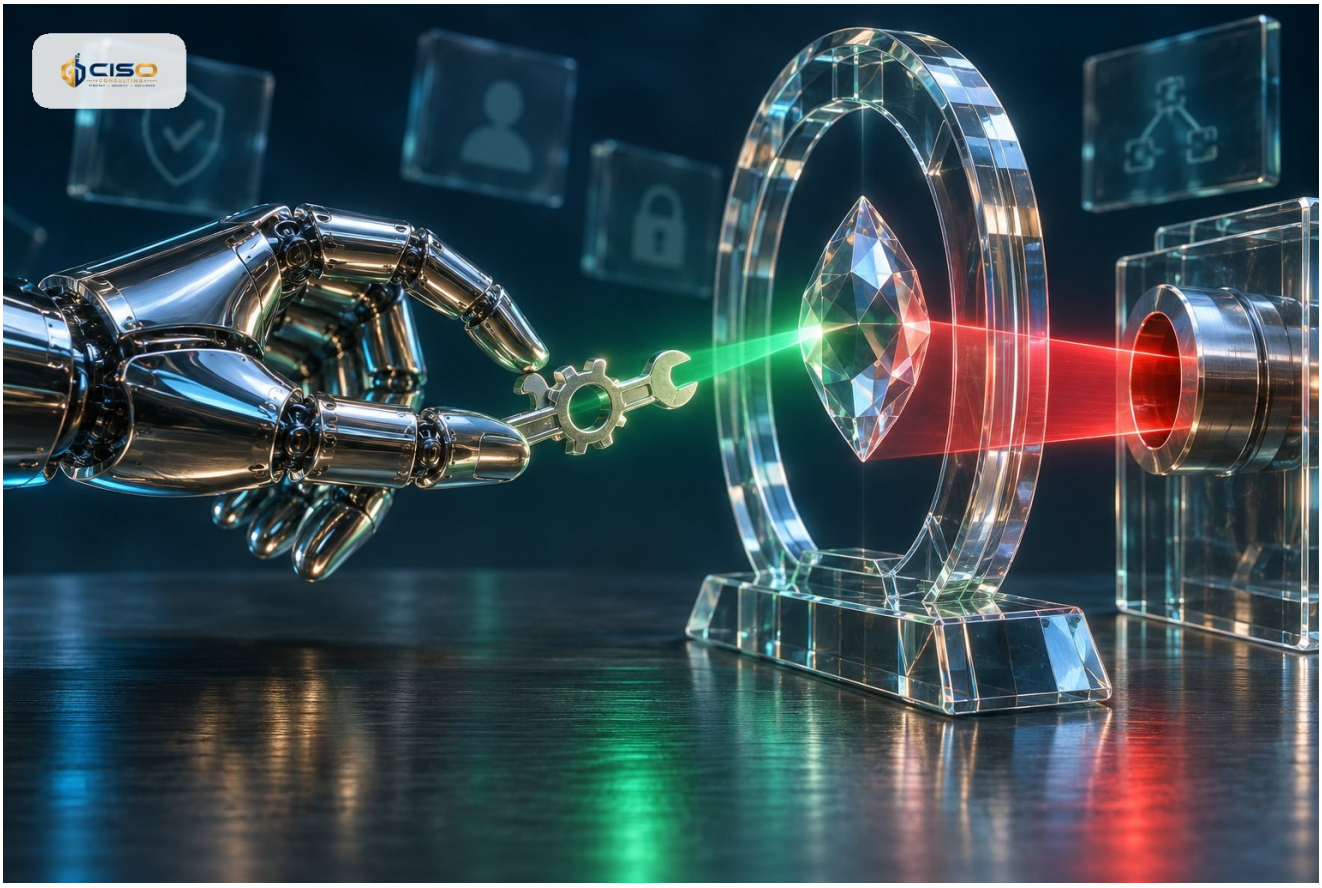
Google previews policy checks that can block an AI agent's tool call before it runs



Google's Preview semantic governance policies on the Gemini platform use a language model as a judge to deny agent tool calls that break company rules or the user's request, before they run.

- The check runs outside the agent, at an enforcement point such as Agent Gateway, and evaluates each proposed tool call.
- A call must satisfy both the user's request and plain-language rules written by legal or compliance teams; the user-intent check is on by default.
- Denied calls are replaced with a denial message and do not run; verdicts and the judge's rationale go to Cloud Logging.
- Verdicts are probabilistic and may vary, so Google recommends dry run review before enforcement.

> - Streamed model responses are not evaluated, and the policies complement rather than replace existing controls.

Google has put semantic governance policies in Preview on the Gemini platform. They add a checkpoint in front of an AI agent's actions: each tool call proposed by the agent's model is reviewed and blocked before it runs if it does not match the user's request or breaks a company rule.

How it works

The check sits outside the agent, at an enforcement point such as Agent Gateway. Permissions define what an agent can do, but not whether a specific call follows policy. A model can be argued out of its instructions by a persuasive user, or misled by a document carrying hidden instructions, a technique known as indirect prompt injection.

A language model acts as judge. It compares each call with the user's request and with rules that legal or compliance teams write in plain language. The user-intent check is on by default and needs no policies. If a call fails, its response is replaced by a denial message and the call does not run. Verdicts and rationale are sent to Cloud Logging.

Google's example is a finance agent limited to sending unreleased quarterly results to internal addresses before the earnings release. An email to an outside investor is denied, and a hidden instruction in a planning document to delete drafts is denied by the intent check.

Limits to note

Because the judge is a language model, its verdicts are probabilistic and may vary. Google advises reviewing them in dry run mode before enforcing, and placing facts a policy depends on, such as an earnings date, in the agent's system instructions. Streamed model responses are not evaluated. Google states that the policies complement disclosure controls and legal and compliance reviews rather than replace them.

What to do now

1. Inventory AI agents in use or under pilot and list the tools each can call, especially email, file deletion and data export.
2. Have compliance and legal teams draft plain-language rules for sensitive actions, such as sending non-public financial or personal data outside the organisation.
3. If you pilot these policies, run them in dry run mode first and review verdicts before enforcing.
4. Put the facts a policy depends on into the agent's system instructions, and test with indirect prompt injection scenarios.
5. Treat the policies as one layer: keep access permissions, disclosure controls and human review in place, and route verdict logs to your monitoring.