Google Cloud CISO: agent actions must be traceable, auditable and reversible



In an interview, Google Cloud's CISO argues that human-paced defense no longer holds against automated attacks. He urges per-agent identities, tamper-proof decision logs and board-level questions on tracing and reversing agent actions.

- Chris Betz says assumptions that manual, human-led workflows would remain the defensive baseline have aged poorly.
- Autonomous defensive agents can handle enrichment, correlation and triage, so analysts start from verified conclusions rather than raw alerts.
- State-sponsored espionage actors may stay below behavioral alert thresholds, but defenders can correlate internal signals such as identity privileges and runtime activity.
- He advises moving away from shared service accounts and blanket admin tokens toward least-

> privileged identities for each agent, verified per transaction.
>
> - Boards should ask whether every agent action can be traced, audited and reversed.

In an interview with Help Net Security, Chris Betz, CISO of Google Cloud, set out how he sees security changing as AI agents spread. His starting point is that the long-held assumption that human-centric workflows would remain the defensive baseline has aged poorly, because attackers now use automation to probe networks and move laterally in minutes.

Changing the SOC

Betz describes autonomous defensive agents taking over data enrichment, correlation and triage. In his view this does not replace the analyst but changes where the analyst starts: the alert becomes the verified conclusion of an automated investigation. Junior analysts would set doctrine, define risk appetite and verify automated actions designed to be understood by a human in 60 seconds and reversed in 30.

Identity and oversight

He notes that espionage actors often avoid loud, fast tooling and stay below behavioral alert thresholds. Defenders, however, own internal ground truth: identity privilege graphs, network architecture, code lineage and runtime activity. He recommends dropping flat trust architectures, shared human service accounts and blanket administrator tokens, and extending verification to individual transactions so each agent has a least-privileged, verifiable identity.

For boards, the central question is whether every agent action can be traced, audited and reversed. He points to centralized guardrails with tamper-proof decision traces that record inputs, model reasoning, tools called, parameters passed and the policy basis for allowing the action.

The interview is an opinion piece from a vendor executive, but its governance points are practical for any institution piloting AI agents.

What to do now

1. Inventory AI agents already running or being piloted across departments, including those launched by coding assistants or business teams.
2. Replace shared service accounts and blanket admin tokens used by automation with distinct, least-privileged identities per agent.
3. Define what an agent decision trace must record (inputs, reasoning, tool calls, parameters, policy basis) and store it in tamper-resistant logging.
4. Test that consequential automated actions can be explained quickly and rolled back before extending agent autonomy.

5. Add agent traceability and reversibility to the board's regular cyber risk reporting.