Untrusted input for a security agent

RULE untrusted-context-v2 (owner: security engineering, review 19 Feb 2027)
TRIGGER: any input tagged outside_control enters the run context
  (tags are set by the connector that fetched the input, from the author's account type, not by the model)
DEFAULT: input from a connector that sets no tag counts as outside_control
DERIVED: a summary, note or memory entry made from tagged input carries the tag
FROM THEN UNTIL THE RUN ENDS:
  blocked: every write tool (ticket close, firewall change, repo push, email send)
  confirmation: any tool that sends data outside the company network
  still allowed: read-only lookups on named systems
LOG: each blocked or gated call with tool name, input tag and run ID
RESET: a new run starts clean; a person's approval applies to one call only

Agentic AI in Cybersecurity, Untrusted input for a security agent, edition 19 Nov 2026, https://agenticaiincybersecurity.com/evaluation-checklist/untrusted-input/. License CC BY 4.0.