Agent Evaluation Checklist

Scope of the job

[ ] Q01. Does this security job need an agent at all, or would a fixed workflow with a model call inside it do the work?
[ ] Q02. Which production systems and which kinds of change (read, comment, ticket, configure, block, delete) is the agent allowed to touch, written down per tool?
[ ] Q03. Does any tool give the agent an open-ended capability (run a shell command, fetch any URL) where a narrow function would do?
[ ] Q04. What identity and permissions does the agent use on each downstream system, are they the least needed (for example read-only on a log store, write on one ticket queue), and is authorization enforced by those systems rather than left to the model?

Approval and handover

[ ] Q05. Which actions need a named human to approve them before they run, and which may run unattended?
[ ] Q06. Is each tool rated for risk (read or write, reversible or not, which accounts it needs), and do those ratings drive when the agent must pause?
[ ] Q07. What makes the agent stop and hand over (retry limits, rate limits, unexpected tool results), and who is paged when it does?
[ ] Q08. Can the approver understand what the agent is proposing and why, and are they trained to resist over-trusting its output?

Evidence and monitoring

[ ] Q09. Is every step the agent takes (inputs, intermediate decisions, tool calls with parameters, data touched, identity used) logged somewhere the agent cannot edit?
[ ] Q10. After each change, what independent check (a query against the real system, a test, a diff) confirms the change happened as claimed, rather than trusting the agent's own report?
[ ] Q11. What do you measure after go-live to notice that the agent's behavior or quality has shifted, and who reviews it on what schedule?

Stop and roll back

[ ] Q12. Can one person stop the agent immediately, and can every class of change it makes be undone or compensated for?
[ ] Q13. Is there an incident plan for the case where the agent itself caused the incident, with criteria for switching it off and a path for remediation?

Evaluation

[ ] Q14. What written success criteria and test set must the agent meet for this specific job before it gets production access, and were they measured in conditions like your environment?
[ ] Q15. Has the agent been run against copies or sandboxes of the systems it will change, with its write tools pointed at the copy, before any real change?
[ ] Q16. What re-runs the evaluation when the model version, prompt, tools or data change, and what result blocks the change from shipping?

Data and inputs

[ ] Q17. What data can the agent read for this job, does it include production or sensitive records, and who approved that access?
[ ] Q18. Which text the agent ingests (alerts, tickets, emails, logs, web pages) can come from outside your control, and what do its tools do when such text is in context?

Build or buy

[ ] Q19. If the agent is bought or hosted by a supplier, what do you control (tools, permissions, approval gates, logs, update timing) and what does the supplier control?
[ ] Q20. If you build, which parts (model, harness, tools, sandbox) can be swapped later, and what has the supplier told you about known limitations and failure modes?

Agentic AI in Cybersecurity, Agent Evaluation Checklist, edition 19 Nov 2026, https://agenticaiincybersecurity.com/evaluation-checklist/. License CC BY 4.0.