ONE-PAGE BRIEF FOR SECURITY LEADERS

Redloop

How AI penetration testing works

AI agents can now do much of the grinding work of a red team: mapping assets, trying attack variations, reading configs and permissions, and chaining small weaknesses. Done right, that turns a once-a-year test into continuous pressure. Done wrong, it is an uncontrolled bot pointed at production. Here is what “done right” looks like.

The loop

01
Scope & authorize
Signed contract, asset list you own, rules of engagement, emergency contacts.

02
Map
Agents enumerate every authorized asset and watch for changes.

03
Attack
Agents try attacker techniques; humans approve anything high-impact.

04
Chain & prove
Weaknesses are linked into paths and proven with minimal evidence.

05
Report
Verified findings go to your tracker with reproduction steps and a fix.

06
Retest
Each fix is re-attacked to confirm it holds. Then the loop continues.

Where AI helps

- Coverage: far more assets and attack variations than a human team has hours for
- Persistence: testing continues as infrastructure changes, not once a year
- Reading at scale: configs, IAM policies, code paths, exposed files
- Chaining: connecting findings across external, cloud, identity and apps

Where humans must still decide

- What is in scope, and what is too fragile to touch
- Approving disruptive or high-impact actions before they run
- Verifying each finding so you get evidence, not scanner noise
- Social engineering design, with HR and legal sign-off

Non-negotiable guardrails

Written authorization
Signed by an authorized officer before any testing.

Technical allow-list
The signed scope is enforced in tooling, not just on paper.

Third parties excluded
Vendor-hosted systems only with the provider’s permission or policy.

Windows & limits
Testing hours, rate limits and exclusions agreed up front.

Stop button & full log
You can halt everything; every action is logged and shared.

Proof, not plunder
Minimal evidence, no bulk data, deleted on a contractual schedule.

10 questions to ask any AI pentest vendor (including us)

1. How is our signed scope enforced technically, not only contractually?
2. Which actions can the AI take without a human approving them?
3. Can we see a complete log of every action taken against our systems?
4. How do we stop all activity, and how quickly does it stop?
5. How is each finding verified before it reaches us?
6. Do you chain findings into attack paths, or hand us a list?
7. What evidence do you keep, where is it stored, and when is it deleted?
8. Which AI models do you use, and is our data used to train them?
9. How do you handle systems hosted by our vendors and SaaS providers?
10. Is retesting included, and how soon after a fix does it happen?