Beyond the Model

What It Takes to Move AI Pentesting From POC to Production

Many security teams can now stand up an AI pentesting proof of concept that produces promising results within weeks. The harder challenge is turning that POC into a dependable, scalable security capability and deciding whether to build it internally or buy an established platform and operating model.

Insights drawn from a July 30, 2026 conversation featuring Dan Lacher, Cybersecurity Engineering Team Leader at Dow; Mark Kuhr, Ph.D., Co-Founder & CTO at Synack; and Tim Nordvedt, Solutions Architect at Synack, who moderated the discussion.

01 The POC-to-Production Gap

Lab and capture-the-flag benchmarks ask a model to find and exploit a known vulnerability in a controlled app. That doesn't reflect a live enterprise environment with real logins, complex workflows, and inconsistent backend technologies. A system that performs well in a lab can fail to generalize once it meets that complexity.

Before trusting any POC: ask whether it reflects your production reality, or a simplified test case built to look good in a demo.

02 Why the Harness Matters as Much as the Model

It's tempting to treat model choice as the core decision: which frontier lab, which version, which benchmark score. That instinct misses where the real capability lives. The harness, meaning the orchestration, guardrails, and decision-making logic wrapped around a model, determines whether an agent performs like an experienced pentester or wanders off target.

> **"The performance isn't even close. The harness really makes a big difference."**
> Mark Kuhr, Ph.D., Co-Founder & CTO, Synack

The same base model, run without a harness, performs far worse than a harness-wrapped version of that identical model. In practice, most of the engineering lift in a build decision isn't model access. It's the accumulated learning from real pentests, encoded into rules about how the agent should evaluate a given vulnerability or attack surface.

> **"It wasn't a matter of trying to build the shiny new object and just have a cool new tool. The attack surface is growing faster than most security teams can keep up with."**
> Tim Nordvedt, Solutions Architect, Synack

03 The Hidden Costs of Building In-House

Teams scoping a build typically estimate token spend and initial engineering time. What tends to get missed:

- Token consumption is unpredictable and varies by attack surface, run to run
- Every model update or deprecation means re-evaluating prompts and re-testing the toolkit
- Safely orchestrating multiple agents and testing workflows at runtime requires ongoing engineering
- Guardrails must enforce customer-specific rules of engagement in real time, across every agent, continuously
- None of this is a one-time cost. It's a standing team

Dow's own evaluation reflects this. After testing several AI pentesting tools, Dow chose not to build the full capability internally, citing the token volume and engineering headcount required to replicate what a dedicated vendor team has already built and maintains.

04 Human Validation Proves What Matters

AI can expand coverage and speed, but high-impact findings still need human validation to establish exploitability, relevance and priority in the customer's environment.

> **"You don't want to waste your time chasing AI-generated slop and false positives."**
>
> Mark Kuhr, Ph.D., Co-Founder & CTO, Synack

For Dow, human review before findings reach its team is a hard requirement that supports the level of confidence expected from its testing program.

05 Continuous Testing Keeps Pace With Change

Point-in-time testing has a built-in expiration date. The moment a report is generated, the environment underneath it has already changed. Dow described this challenge under its earlier periodic testing model: annual or quarterly assessments could leave newly exposed assets waiting months for testing.

The fix isn't just testing more often. It's shrinking the gap between when an asset goes live and when it's tested. Dow built internal automation that detects newly exposed cloud assets and routes them to Synack for testing in near real time, rather than waiting for the next scheduled engagement.

Questions to Ask Before You Decide

1. Does our POC reflect the complexity of our actual production environment, or a simplified test case?
2. What ongoing token, compute, and staffing costs sit beyond the initial build, and who owns that budget line permanently?
3. Is our false positive rate a tuning problem or an architecture problem, and can we tell the difference?
4. What does day-to-day human review of autonomous findings look like in our environment, and who is accountable for it?
5. Are there narrow use cases where an internal build genuinely makes sense, separate from the general pentesting need?

The Bottom Line

Build vs. buy is ultimately an operating-model decision, not only a technology decision. It determines who owns the engineering, validation, governance, maintenance and continuous improvement required to make AI pentesting dependable at scale.

Dow's team, after evaluating the market and its own capacity, chose to extend its internal red team with Synack rather than build and staff a full AI pentesting capability.

> **"I really see it as a force multiplier."**
>
> Dan Lacher, Cybersecurity Engineering Team Leader, Dow

For most security teams, the more useful question isn't whether AI pentesting can be built in-house. It's whether the team wants to be in the business of maintaining it indefinitely, instead of directing it at the assets that matter most.

Synack combines Sara AI Pentesting, the Synack Red Team and the Synack Platform to expand testing coverage, prove real-world exploitability and operationalize continuous pentesting without requiring customers to build and staff the complete capability internally.

Hear how Dow and Synack approached the decision, or discuss how the build vs. buy considerations apply to your environment.

WATCH THE CONVERSATION

REQUEST A BUILD VS. BUY ASSESSMENT