CYBERSECURITY 101 · eBuilder Security

AI Red Teaming

Attacking your own AI systems to find their weaknesses before someone else does. Here is what it tests, why it matters and where to start.

WHAT IT IS

AI red teaming is the practice of deliberately attacking your own AI system with adversarial prompts and inputs to find how it fails. It targets the model's behaviour, such as jailbreaks, prompt injection and data leakage, rather than the servers around it. A normal penetration test does not cover this.

WHY IT MATTERS

362

documented AI incidents in 2025, up from 233 in 2024

Stanford HAI, 2026 AI Index

€15M / 3%

top EU AI Act fine for breaching provider or deployer duties, whichever is higher

Regulation (EU) 2024/1689, Art. 99

2,244

people who stress-tested leading AI models at the DEF CON public red team event

GRT Challenge, 2023

WHAT HAPPENED

In February 2024 the British Columbia Civil Resolution Tribunal held Air Canada liable after its support chatbot invented a bereavement-refund policy that did not exist. The tribunal rejected the airline's claim that the bot was a separate entity. Testing the chatbot against the airline's real policies before launch would have caught the error. (Moffatt v. Air Canada, 2024 BCCRT 149)

RED FLAGS TO WATCH FOR

Your AI probably needs red teaming if any of these are true.

- A chatbot, copilot or agent talks to customers or staff without adversarial testing.
- The model or its prompts have changed since the last review.
- The AI can act, for example send email, issue refunds or change records.
- You cannot show evidence of how the system was tested.

THE ONE RULE THAT STOPS IT

Test your AI systems adversarially at every release, including each model or prompt change.

THEN BUILD THE HABIT AROUND IT

- Inventory your AI systems and rank them by the harm a failure would cause.
- Define what a serious failure means for each system before testing.
- Re-run adversarial tests on every model or prompt change.

- Give each finding an owner, a severity and a deadline, then retest.

MYTHS AND FACTS

MYTH

A standard penetration test already covers our AI.

FACT

A pentest checks the app and infrastructure around a model. Jailbreaks, prompt injection and hallucination need AI red teaming, which tests the model itself.

MYTH

If the model passed its safety checks, it is safe.

FACT

Stanford's 2026 AI Index found safety performance dropped across every model tested once adversarial jailbreak prompts were used.

MYTH

We tested the AI at launch, so we are covered.

FACT

Models change through updates and new attack methods appear constantly. A system that passed last quarter can fail today.

THE LEGAL DUTY

The EU AI Act requires adversarial testing for the most capable AI models under Article 55, and high-risk systems must be robust against attack under Article 15. In Sweden, Cybersäkerhetslagen and NIS2 require essential and important entities to test the effectiveness of their security measures.

GET HELP

See how eBuilder Security tests AI systems for Swedish organisations.

Read the Full Article →