Red-Team Evaluation

red-team-evaluation · safety · concept

Source: https://policywindow.org/wiki/red-team-evaluation

Generated 2026-05-30T22:10:07 UTC

Summary

Structured adversarial probing of an AI model's capabilities and behaviour before deployment, designed to elicit failures that ordinary evaluation would miss.

At a glance

Used by

12 instrument(s)

Related concepts

frontier-tier, asl-3, systemic-risk, designated-systemic

Primary source

EU AI Act Art. 55(1)(a) — the most binding articulation

Details

Red-team evaluation originated in cybersecurity (penetration testing) and was adapted to AI by the 2022 DEF CON Generative Red Team event and later codified in the 2023 White House voluntary commitments. EU AI Act Art. 55(1)(a) requires adversarial testing for general-purpose AI models with systemic risk. US EO 14110 §4.2(a)(i) required reporting of red-team results for foundation models above the compute threshold (rescinded under EO 14179). G7 Hiroshima Code §1 calls for 'adversarial testing prior to and throughout deployment.' Anthropic, OpenAI, and Google DeepMind each maintain internal red-team programs with public methodology disclosures.

Governance disputes centre on: (1) WHO must red-team (provider, independent third-party, government); (2) WHAT capabilities are in scope (CBRN uplift, autonomous replication, election manipulation, etc.); (3) WHO sees the results (provider only, regulator under confidentiality, public); (4) WHAT triggers re-evaluation after deployment.

How to cite this article

APA

Policy Window. (n.d.). Red-Team Evaluation [Wiki article — Concept]. https://policywindow.org/wiki/red-team-evaluation

CHICAGO

Policy Window. n.d.. "Red-Team Evaluation." Wiki article (Concept). https://policywindow.org/wiki/red-team-evaluation.

HARVARD

Policy Window (n.d.) 'Red-Team Evaluation', Wiki article — Concept, available at: https://policywindow.org/wiki/red-team-evaluation.

OSCOLA

Policy Window, 'Red-Team Evaluation' (Wiki article — Concept, n.d.) <https://policywindow.org/wiki/red-team-evaluation> accessed [date].

BIBTEX

```
@misc{policywindow-red-team-evaluation,
title  = {Red-Team Evaluation},
author = {Policy Window},
year   = {n.d.},
howpublished = {red-team-evaluation — safety},
url    = {https://policywindow.org/wiki/red-team-evaluation},
note   = {Primary source: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689}
}
```