

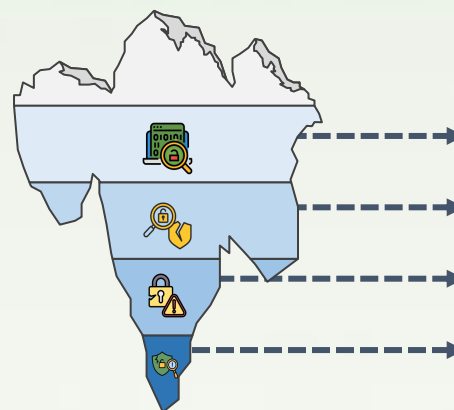


AI & Partners

Amsterdam - London - Singapore

Agentic AI Red Teaming Guide (CSA)

CSA's guide provides a detailed red teaming framework for Agentic AI, explaining how to test critical vulnerabilities across dimensions like permission escalation, and hallucination.



What are four critical vulnerabilities?
See Slide 2 for details.



Four critical vulnerabilities:



Hallucination Exploitation

Agents generate false or misleading outputs that can misinform decisions or trigger harmful actions.

Goal Manipulation

Attackers subtly alter agent instructions or goals, steering behaviour without detection.

Control Hijacking


Unauthorized actors take over agent decision-making via spoofed commands or escalated roles.

Multi-Agent Exploitation & Blast Radius

A compromised agent cascades errors or malicious actions across systems and agents, amplifying impact.



Implementation guidance:

 **Red Team Continuously.** Integrate adversarial testing early and post-deployment—simulate real-world exploits to harden defences.

 **Restrict by Design.** Apply least-privilege, role separation, and strict task boundaries across all agent functions.

 **Monitor & Log Everything.** Implement fine-grained logging, anomaly detection, and behavioural baselining to flag deviations.



What are typical Agentic AI vulnerabilities?
Check the official announcement for more details.