Category

Data & AI

AI Safety & Red-Teaming Playbook (Testing Behaviors & Failure Modes)

Actively trying to break your AI system before someone else does

- Practitioner
- Advanced
- Template Included

- Administrator
- 08 December 2021
- 3 minute(s)

Overview

A framework for AI safety and red-teaming — deliberately testing AI systems for adversarial inputs, edge-case failure modes, and unintended behaviors before deployment — treating adversarial testing as a distinct, necessary discipline from standard accuracy validation, since standard testing rarely surfaces the failure modes red-teaming specifically targets.

Isn't standard accuracy testing sufficient to validate an AI

system's safety before deployment? Standard accuracy testing validates typical-case performance, but doesn't systematically probe for adversarial inputs, edge cases, or unintended behaviors a motivated user or attacker might discover — red-teaming is a

distinct discipline specifically designed to surface these failure modes standard testing misses.

Who should conduct AI red-teaming — the same team that built the

model, or an independent group? Independent red-teaming, ideally from a team not invested in the model's success, tends to surface more genuine failure modes than self-testing by the building team, who may have blind spots about their own system's weaknesses.

{ "@context": "https://schema.org", "@type": "FAQPage", "mainEntity": [{ "@type": "Question", "name": "Isn't standard accuracy testing sufficient to validate an AI", "acceptedAnswer": { "@type": "Answer", "text": "system's safety before deployment?\nStandard accuracy testing validates typical-case performance, but\ndoesn't systematically probe for adversarial inputs, edge cases, or\nunintended behaviors a motivated user or attacker might discover —\nred-teaming is a distinct discipline specifically designed to surface\nthese failure modes standard testing misses." } }, { "@type": "Question", "name": "Who should conduct AI red-teaming — the same team that built the", "acceptedAnswer": { "@type": "Answer", "text": "model, or an independent group?\nIndependent red-teaming, ideally from a team not invested in the\nmodel's success, tends to surface more genuine failure modes than\nself-testing by the building team, who may have blind spots about their\nown system's weaknesses." } }] }

Subscriber access

Unlock this playbook

This playbook — including every framework, template, and step-by-step section — is available free to Think Insights subscribers. Enter your email to unlock it instantly and get our weekly insights newsletter. No account needed, and access is remembered on this device.

First Name

Last Name

Email

Country - Select -

■ I agree to data processing in accordance with GDPR

Submit

Leave this field blank

References

Related playbooks

Data & AI

AI in Data Engineering & Pipelines Orchestration Playbook

A framework for using AI in data engineering and pipeline orchestration — AI-assisted pipeline code generation, automated data quality monitoring, and intellige

- Practitioner
- Intermediate
- Template Included

Read playbook

Data & AI

AI Data Readiness & Foundation Models Input Quality Playbook

A framework for assessing and building AI data readiness — data quality, completeness, and structure appropriate for foundation model input — recognizing that A

- Practitioner
- Intermediate
- Template Included

Read playbook

Data & AI

AI for Design, Prototyping & Experimentation Workflows Playbook

A framework for using AI in design, prototyping, and experimentation workflows — accelerating prototype generation and iteration speed while maintaining genuine

- Practitioner
- Intermediate
- Template Included

Read playbook

Talk to us about AI safety and red-teaming

Tags

- AI Safety
- Red-Teaming
- AI Testing

Author

Administrator

Think Insights Administrator