How to evaluate AI responsibly

Understand how AI evaluation toolkits support responsible AI governance, why they can't work alone, and how a lifecycle approach turns testing into effective oversight.

The 3 main types of AI evaluation toolkits

1 Benchmarking tools

Measures how well a model performs

Moonshot **HELM (AIR-Bench)** **EleutherAI LM Eval**

- Easy to integrate but offers limited explanation behind results.
- Best for quick capability checks, not full evaluation.

2 Principle-based testing tools

Checks fairness, transparency, explainability

AI Verify **Fairlearn** **Veritas** **AIF360**

- Useful for meeting regulatory or internal standards.
- Results may require technical expertise to interpret.

3 Red-teaming tools

Stress-tests the model for unsafe behaviour

Moonshot (RT) **Giskard** **PyRIT** **Cyberseceval**

- Most valuable before deployment to assess real-world resilience.
- Requires more setup and expertise to run effectively.

Today's toolkits cannot work alone because they face limitations

STRUCTURAL

- Different tools measure things differently
- No single tool checks everything you need

USABILITY

- Setup, instructions and ease-of-use can vary
- Results are not always easy to understand

STRATEGIC

- Tools rarely tell you what to do after you get the results

A better way is to evaluate AI based on the development lifecycle

1 Know where you are in the AI lifecycle

Is the model being built, tested, deployed or monitored?

2 Decide what matters most at that stage

Fairness **Safety** **Security** **Robustness** **Transparency**

3 Pick the tools that match your needs

Most teams need a mix of benchmarking, principle-based tests and red-teaming to get a complete picture.

4 Turn evaluation into real-world governance

Record your findings, link risks to your policies and frameworks, and feed insights into model cards and monitoring workflows.
This is where expert guidance helps you move faster and with confidence.

NCS brings governance frameworks, lifecycle expertise and practical toolkit experience to help organisations evaluate AI safely and responsibly.

Explore to learn more about key toolkits, their limitations, and how to build a complete, lifecycle-driven approach to AI governance.
Read the full article here

Contact us now