CSET CENTER for SECURITY and EMERGING TECHNOLOGY

Key Takeaways from "Putting Explainable AI to the Test: A Critical Look at AI Evaluation Approaches"

Governments around the world acknowledge the importance of building trustworthy AI systems. However, it is not immediately clear how to evaluate aspects of AI safety and trustworthiness. To gain insight into these methods, we focused on explainability and interpretability, two commonly discussed facets of trustworthy AI. We conducted a literature review of research papers that focus on the explainability and interpretability of recommendation systems––a type of AI system that often uses explanations. Specifically, we analyzed how researchers (1) describe explainability and interpretability and (2) evaluate their explainability and interpretability claims in the context of AI-enabled recommendation systems. We found that:

Explainability Descriptions

- Researchers describe explainability and interpretability in variable ways across papers.
- Research papers do not clearly differentiate between explainability and interpretability.

Explainability Evaluations

- Researchers adopt combinations of five different evaluation approaches: case studies, comparative evaluations, parameter tuning, surveys, and operational evaluations.
- Research papers more often test if systems are built according to researcher specifications than if systems work as intended in the real world. Both types of evaluations serve important but different purposes.

More work is needed to determine whether these results translate to other research areas. **However, policymakers should be aware that if researchers understand and measure facets of AI trustworthiness differently, policies for promoting safe and trustworthy AI systems may not be effective. Policymakers would do well to invest in standards for AI evaluations and a workforce that can assess the efficacy of these evaluations in different contexts.**

For more information:

- Download the report: https://cset.georgetown.edu/publication/putting-explainable-ai-to-the-test-a-critical-look-at-ai-evaluation-approaches/
- Contact us: cset@georgetown.edu