

Governing AI: Week 3 Section

Introductions

Please share:

Your name

Your
hometown

Your program
and year

One question
you have
about AI Gov

What is the last
great (or
terrible!) book
you've read?

Agenda for today

Goal: Review Milestone 1

- 1 Introductions

- 2 AI Evaluations Review + Discussion

- 3 Research Project Overview

- 4 Project Brainstorming + Discussion

Why evaluations matter for governance

Benchmarks are how we compare models, track progress, and find weaknesses. Increasingly, they also feed into model selection and policy decisions.

EU AI Act, Article 51

Classifies general-purpose AI models as having "systemic risk" based on high-impact capabilities, assessed with technical tools including benchmarks, plus a training-compute threshold.

If a law depends on a test score, who decides whether the test is any good?

Two terms we'll use today

Evaluation

Any method for assessing what a model can do or how it behaves: benchmarks, red-teaming, human studies, audits.

Benchmark

A standardized set of tasks plus a scoring metric, so different models can be compared on the same test. Examples: MMLU, GPQA, SWE-bench.

Does a high score mean what we think?

A model scores highly on GPQA, a graduate-level science Q&A benchmark. **Can it write your research paper?**

Construct validity

Does the test measure the capability it claims to measure?

Content validity

Does it cover all the relevant cases, or only a narrow slice?

External validity

Do the results generalize to other settings, like real-world use?

BetterBench: grading the benchmarks

46

best-practice criteria

5

benchmark lifecycle phases

24

widely used AI benchmarks assessed

Design

Implementation

Documentation

Maintenance

Retirement

Criteria draw on benchmarking practice in other fields (from transistor hardware to environmental quality to bioinformatics) and interviews with 20+ policymakers, developers, and researchers.

Other common failure modes

Contamination

Test questions leak into training data, so a model may recall answers rather than reason its way to them.

Saturation

Top models max out a benchmark quickly, so it stops telling them apart.

Score gaming

Developers optimize for the leaderboard number rather than the underlying capability.

Static tests

Fixed question sets miss how models behave in real, interactive deployments.

What this means for governance

Who builds and audits evals?

AI labs, government AI safety institutes, or independent third parties, each with different incentives and access to models.

Should benchmarks meet standards?

If a benchmark can trigger legal obligations, should the benchmark itself have to meet a minimum quality bar?

From law to practice

The EU's General-Purpose AI Code of Practice turns AI Act obligations into concrete requirements for model providers.

Discussion

- 1 Should policy rely on benchmark scores at all? What's the alternative?
 - 2 Before a "dangerous capabilities" test triggers regulation, how much evidence of validity should we require?
 - 3 Who should pay for and maintain public benchmarks?
 - 4 Where might these measurement problems show up in your own research project?
-

AI Evaluation Research Project Overview

1 Motivation – Why AI Evals?

2 Project Overview

3 Milestone 1

4 Project Brainstorming

Why AI Evals?

- Interdisciplinary
 - Evals allow us to tackle questions through all of our disciplinary lenses.
- Evidence-based
 - Engage with the best evidence we have on AI and its potential.
- Highly Impactful!
 - There is not much great research on AI evals out there. You can change that!

Through this project, we hope you further develop your research interests and leave with a project that brings the field forward. (And you can present!)

Project Overview

- You may complete the assignment in pairs.
 - Pairs must be finalized in the first Milestone assignment!
- Projects fall into one of three tracks:
 - Trustworthy Institutions,
 - Who decides what evals are important?
 - Trustworthy Evals
 - How do we build better evals?
 - & Trustworthy Decision-making processes.
 - How do we better implement evals into existing decision-making systems? (I.e law/medicine!)

Milestone 1

Research Question

- Is your research question well-scoped?
 - Do you have a specific topic, question, idea, etc.
- Is it relevant/important?

Literature Review

- Do you substantively engage with relevant existing literature?

This assignment is designed as a **calibration** – making sure that the teaching team can support your project early!

Project Discussion

With a partner, brainstorm three project ideas together.

As you're planning, keep in mind the three tracks! (They can be helpful inspiration!)

We'll be going around and checking-in with you all.