NOTE: UPDATES TO THE ASSIGNMENT MARKED IN RED. FOR THE PARTS THAT WE CROSSED OUT, WE STILL ENCOURAGE YOU TO THINK ABOUT THEM BUT THEY ARE NOT A FORMAL PART OF THE ASSIGNMENT ANYMORE.

The expectation is not that you have a fully fledged-out, deployment-ready evaluation by the end of the assignment and we do not expect you to have a paper ready for peer review. Instead, the intent of this assignment is to get you to think about a) what are relevant claims we want to make about AI models, b) what does good evidence look like for these claims, and c) what are some potential real-world considerations and concerns about consequences of evaluations and associated design decisions.

AI Governance
Technical Governance Assignment
2025-2026

Due on Dec 5th at 5pm PST

Law students are not required to complete this assignment. However, they may still choose to complete this assignment and receive feedback.

Designing and Examining the Validity of Technical Evaluations

Technical evaluations play an important role in informing stakeholders of the capabilities and risks of AI models. Evaluations attempt to place an empirical foundation underneath claims about the relative safety or risk of AI models.

For this assignment, you will decide between one of three types of evaluations:

- **Benchmarks**: A particular combination of a dataset or sets of datasets, and a metric, conceptualized as representing one or more specific tasks or sets of abilities, picked up by a community of researchers as a shared framework for the comparison of methods[1]. In the context of AI models, popular benchmarks include Graduate-Level Google-Proof Q&A Benchmark (GPQA) and Massive Multitask Language Understanding (MMLU).
- **Red Teaming Exercises**: Adversarially testing an AI system to identify potential vulnerabilities. Unlike a standard penetration test, a red team uses the same tactics, techniques, and procedures as real-world attackers – but with a constructive rather than exploitative intent – to identify vulnerabilities across systems, people, and processes in a comprehensive and realistic manner. See these OpenAI and Anthropic blog posts for a more detailed introduction to their approaches to red teaming AI models.
- **Human Uplift Studies**: These evaluations use a randomised controlled trial format to assess the degree to which access to a specific advanced AI system

[1] https://arxiv.org/pdf/2111.15366

improves human performance[2]. For example, this study by METR on "Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity".

Your Task

Building an evaluation from the ground-up offers you the opportunity to experience how design choices influence the validity of evaluations across multiple dimensions. While the subject of AI evaluation can be technical, the core of this assignment tests your foundational design and thinking skills; how you justify the evaluation's purpose, design, relevance, and how you assess its practical consequences. Regardless of your background, your ability to interpret a technical evaluation is critical to making informed decisions in AI governance. To this end,

1. **Read Measurement to Meaning: A Validity-Centered Framework for AI Evaluation (assigned November 5th). Take careful note of the definitions provided and the case studies that demonstrate how to apply the framework, these will be critical in understanding the rest of the assignment.**
2. **(part 1)** Identify a construct of interest (e.g., fairness, deception, …). This can relate to either a capability or a risk. Provide a justification for how your choice is relevant to the safe and responsible deployment of models. *Note that you cannot design a test for a criterion, it needs to be a construct.* ***You can find examples of risks or capabilities to investigate in Appendix 1.3 and 1.4 of the EU GPAI Code of Practice or in the OECD AI Capability Indicators.***
3. **(part 2)** Identify a claim that involves the capability or risk of interest (e.g., My model is intelligent, My model gives fair output when deciding on loans). You may **not** use an example already discussed in the paper (you are neither allowed to re-use construct-claim pairs discussed nor validity assessments of benchmarks in the paper), but we encourage you to follow the structure used in the case studies in Appendix D. If you are uncertain about your construct or claim, please reach out to Saran (saisaran@stanford.edu) or Yash (yashdave@stanford.edu). ***You could take, similar to the GPQA example in the paper, a recent benchmark-claim combination from a recent model launch.***
4. **(part 3)** For each notion of validity (Content, Criterion, Construct, External, Consequential), describe if it is required and the concrete form of evidence you would need to establish the validity of your construct of interest. **As a (rough) guideline, 2-3 paragraphs per validity type is sufficient.**
5. **(part 4)** Design an evaluation (benchmark, red teaming exercise OR human uplift study) that tests the extent to which your claim is true.
    a. Design **5** test items (per member of the group) that would constitute your evaluation. Provide a high-level explanation of the other test items you would include to produce a valid evaluation. *Note: A fully valid test might need more test items but we only need* ***5*** *concrete test items designed per team member. If your evaluation would require more items to be valid, we*

[2] https://www.aisi.gov.uk/blog/advanced-ai-evaluations-may-update

just require a high-level explanation of the additional test items.
The nature of each test item will depend on your chosen evaluation approach.

- Benchmark: A specific question or task designed to measure the model's performance and validate its core capabilities. For example, to test mathematical reasoning. Include the question/prompt, correct answer, and scoring rubric.
- Red Teaming: An adversarial or edge-case prompt (for example on controversial topics) created to probe the model's weaknesses, biases, or potential misalignments. Specify what vulnerability or failure mode each prompt is designed to expose.
- Human Uplift: A structured, real-world scenario prompt presented to human participants to assess how AI assistance enhances their decision-making, accuracy, or efficiency. For example, efficiency comparison with and without AI assistance for a particular task. If you're designing a human uplift study, we only expect **1** test item per member. Note that in general, we expect human uplift test items to be more complex and detailed than test items for the other two strategies. Describe the task with all relevant details, along with the baseline (human without AI), treatment (human with AI), **any additional tools or other relevant information about the test environment,** and the measurement protocol.

b. **(part 5)** **[Extra credit, optional]** Test at least **5** test items on at least two models and report on the results and whether the observed outputs were what you expected. Answer the following questions:
   i. Did models fail in expected ways?
   ii. Were there surprises that suggest construct-irrelevant factors?
   iii. Seeing the model outputs, do you think you constructed a valid evaluation?
   iv. In case of a human uplift study: A simulation of **1** item with 2 group members is sufficient, there is no need to recruit human participants to run an actual RCT.

6. Write up a report with **5 to 10** pages **(total, NOT per person)** detailing:
   a. **(part 1)** The construct of interest you chose and a justification for its relevance in the context of AI governance.
   b. **(part 2)** The claim you want to evaluate. In addition, answer the question: What is the theoretical relationship between your construct and the observables you're measuring?
   c. **A short description of one existing evaluation of your construct, briefly commenting on if it is suitable evidence for your claim. As a (rough) guideline, one paragraph for the description plus a list of strengths and weaknesses per validity type (similar to the examples on pages 43 onwards in the Measurement to Meaning paper) is expected here for the one evaluation.**

d. **(part 3)** A detailed description of your evaluation strategy, covering why the chosen evaluation mode (benchmark, red teaming, or human uplift study) is appropriate, what evidence you would need for your specific evaluation-claim pair to establish the different types of validity, and if all validity types are applicable to your evaluation. ~~**In addition, answer the following questions: Which forms of validity are most critical for your specific claim? Which are hardest to establish given practical constraints? If you had to prioritize, which validity evidence would you gather first and why?**~~

e. **(part 4)** A high-level description of what additional test items you would need beyond the ones you already designed to ensure said validity. *Note: You do not need to design all individual test items required for a valid evaluation, a high-level summarizing description of the missing test items is sufficient.* ***For example, for content validity, your limited set of test items will likely not cover all relevant test cases. We want you to provide a high-level description of which additional test items or categories of test items your evaluation would need to cover to be content valid.***

f. **(part 4)** A description of how the test items should be scored, how the items should be aggregated, and how your final test score should be interpreted. ~~**Answer the following questions: At what score threshold would you conclude your claim is supported?**~~

g. An explanation of what limitations your evaluation strategy has. What claims can you NOT make based on your evaluation design?

h. A description of stakeholders and their use of your evaluation. Answer the following questions: Who might use your evaluation results and for what decisions? What are the potential positive and negative consequences of acting on your evaluation?~~**How might your evaluation be misinterpreted or misused? What would change if your evaluation were used for high-stakes deployment decisions vs. research progress measurement?**~~

i. **(part 4)** (In an appendix, not counting towards the page limit) **5** test items per team member for benchmark and red teaming evaluations or **1** test item for human uplift studies.

j. (part 5) (optional, in an appendix, not counting towards the page limit) The test results and answers to the questions posed in part 5.

Additional Directions

You **may** form a group of **up to** 4 members to complete this assignment. **You can also complete the assignment individually. If you choose to complete it individually, please join an empty Canvas Group.** It's okay if you continue with your policy memo group, however, we recommend you to form a new group. **If you decide to work in a group,** the group should be composed of students of at least two different disciplinary backgrounds (law, computer science, and various other majors). It is up to you to decide on your team. If you need assistance forming your team, please contact Dominic Zappia (zappia@stanford.edu). **The expected page range for the paper does not change based on the number of people you decide to work with.**

You will not need to pay for any AI models to complete this assignment. The Stanford AI Playground offers free access to all AI tools you will need for this assignment.

Each team should submit one copy of the final assignment as a PDF file. The length of the assignment should be the same whether your team is composed of 3 or 4 individuals. But the number of test items can vary based on your team size.

Late policy:

- Submitted within 3 days of deadline: one grade penalty
- Submitted more than 3 days after the deadline: no credit

After the assignment is submitted, we will ask students to complete a short survey online that enables you to communicate how the workload was distributed across team members. If the distribution of the workload is not equal, we reserve the right to increase or decrease grades for individual team members accordingly.

The assignment will be graded based on the conceptual clarity of the construct and claim, the quality of the validity assessment across all five types, the thoughtfulness of test item design, the integration and synthesis in the report, and interdisciplinary collaboration (as reflected in survey).

Additional Resources

Measurement to Meaning: A Validity-Centered Framework for AI Evaluation
https://arxiv.org/pdf/2505.10573

Stanford AI Playground https://aiplayground-prod2.stanford.edu/c/new