Week 5: AI Evaluations

Welcoming (0:00 - 0:10)

⧖ 10:00

Until everyone is there

- ☐ Everybody in the **discussion doc**?
- ☐ Open this week's **readings** and your **notes** if you like.
- ☐ If you have a **statement or question,** put it in the chat or in the document.

Check in

- ☐ Make a quick check in round, roughly **30 seconds to max 1 minute** each.
- ☐ **Optionally,** make notes below if you like.

Name	How was your day?	Do you have a specific goal for this meetup? (e.g., speaking less/more, discussing a specific question)

Feedback last session (0:10 - 0:12)

⧖ 2:00

- The facilitator quickly goes over last week's feedback and specifically, what will be tried out in this session.

Links to feedback forms: https://forms.gle/Z3rzFfCrLJdDv8HDA

Feedback on last session You gave me this feedback on how the discussion could be **improved** in the last session.	**Goals** for this session Let's **try** these ideas for improvement.
[@mod: insert feedback]	[@mod: insert idea for improvement]
[@mod: insert feedback]	[@mod: insert idea for improvement]
[@mod: insert feedback]	[@mod: insert idea for improvement]

- ☐ Everything fine with these goals? Remarks?
- ☐ Okay, let's move on.

Goals of this week (0:12 - 0:15)

⧖ 3:00 Go quickly through the goals and topics of this session.

After this session/week, you should be able to:

- Understand different types of AI evaluations including capability, propensity, and control evaluations
- Explain key evaluation techniques including behavioral and internal analysis methods
- Identify dangerous capabilities that require specific evaluation approaches
- Discuss the limitations and challenges in current evaluation methods
- Analyze how evaluations fit into broader AI safety frameworks
- Assess the relationship between benchmarks and comprehensive evaluations

---

Understanding

Key questions from the resources (0:15 - 0:30)

Start the session by **clearing up** key questions from the **reading material**. If there are no questions, go quicker to the next activity.

Gather questions (3 min)

- Open this week's **readings** if you like.
- ⧖ 3:00 Participants write **their questions** in the box below.
- Feel **encouraged** to ask dumb questions!

Answer questions 12 min

- ⧖ 12:00 The group discusses the questions. If some are still open, you may have time at the end to discuss them.

Example: What are the three main types of properties being evaluated in AI systems according to the chapter?
• Notes   ○
Example: How does the Model Organisms Framework approach the study of potentially dangerous AI behaviors?
• Notes   ○
Example: What is the difference between capability evaluations and propensity evaluations?
• Notes   ○
Your name • Question
• Notes   ○
Your name • Question

• Notes   ○
Your name • Question
• Notes   ○
Your name • Question
• Notes   ○
Your name • Question
• Notes   ○
Your name • Question
• Notes   ○

Discussion

Activity 1 - Statements/Questions (0:30 - 1:00)

With the **remaining time** in the session, spark discussion by voting on the below statements and discussing points of disagreement. You'll not have time for all the questions, do a prioritization.

⧖ 25:00

- ☐ Open this week's **readings** if you like.
- ☐ ⧖ 2:00 Formulate a hot take or **new statements/questions** below.

- ☐ Write your **name** in a column.
- ☐ Someone **reads** the first statement/question.
- ☐ While other people are speaking and you can also write a **comment** in the doc. Let's make this collaborative.
- ☐ **Choose** your position. You can also add and choose new options.
- ☐ When everyone has chosen, **discuss** the different positions. If there is no major disagreement, you can **quickly move on** to the next question.

Those questions are about the last section : "5 Layers of Responsibility: Corporate, National, and International AI Governance".

	Name	**Name**	**Name**	**Name**	**Name**	**Name**	**Name**
1	**Statement/Question** [your statement/question: try to formulate it structured e.g. pro/con, agree/disagree, listing options etc.]						
	Not sel…	Not sel…	Not sel…	Not se…	Not sel…	Not s…	Not sele…
	Not sel…	Not sel…	Not sel…	Not se…	Not sel…	Not s…	Not sele…
	Notes •						
2	**Statement/Question** [your statement/question: try to formulate it structured e.g. pro/con, agree/disagree, listing options etc.]						
	Not sel…	Not sel…	Not sel…	Not se…	Not sel…	Not s…	Not sele…
	Not sel…	Not sel…	Not sel…	Not se…	Not sel…	Not s…	Not sele…
	Notes •						
3	**Statement/Question** [your statement/question: try to formulate it structured e.g. pro/con, agree/disagree, listing options etc.]						
	Not sel…	Not sel…	Not sel…	Not se…	Not sel…	Not s…	Not sele…
	Not sel…	Not sel…	Not sel…	Not se…	Not sel…	Not s…	Not sele…
	Notes •						

4	**Statement/Question** [your statement/question: try to formulate it structured e.g. pro/con, agree/disagree, listing options etc.]						
	Not sel…	Not sel…	Not sel…	Not se…	Not sel…	Not s…	Not sele…
	Not sel…	Not sel…	Not sel…	Not se…	Not sel…	Not s…	Not sele…
	Notes •						
5	**Internal / behavioral techniques** Internal evaluation techniques are more important than behavioral techniques for ensuring AI safety.						
	Not sel…	Not sel…	Not sel…	Not se…	Not sel…	Not s…	Not sele…
	Notes •						
6	**Deceptive alignment** Current evaluation methods are sufficient for detecting deceptive alignment in advanced AI systems.						
	Not sel…	Not sel…	Not sel…	Not se…	Not sel…	Not s…	Not sele…
	Notes •						
7	**Prioritize interpretability** We should prioritize developing better interpretability tools over improving behavioral evaluation methods						
	Not sel…	Not sel…	Not sel…	Not se…	Not sel…	Not s…	Not sele…
	Notes •						
8	**Third Party Access** Independent evaluation organizations should have mandatory access to frontier AI models						
	Not sel…	Not sel…	Not sel…	Not se…	Not sel…	Not s…	Not sele…
	Notes •						
11	**Comprehensive safety testing?**						

	The combinatorial complexity of evaluation scenarios makes comprehensive safety testing impossible.						
	Not sel…	Not sel…	Not sel…	Not se…	Not sel…	Not s…	Not sele…
	Notes •						
12	**Goodhart law?** Standardized evaluation protocols could actually make AI systems less safe by making it easier to game evaluations.						
	Not sel…	Not sel…	Not sel…	Not se…	Not sel…	Not s…	Not sele…
	Notes •						
13	**AI Model Registries** The gap between evaluation and deployment contexts is fundamentally unsolvable.						
	Not sel…	Not sel…	Not sel…	Not se…	Not sel…	Not s…	Not sele…
	Notes •						
14	**The generalization gap** How could we strike the right balance between verifying compliance with safety standards and maintaining intellectual property confidentiality?						
	Not sel…	Not sel…	Not sel…	Not se…	Not sel…	Not s…	Not sele…
	Notes •						
15	**False confidence** Red team evaluations provide false confidence in AI safety measures						
	Not sel…	Not sel…	Not sel…	Not se…	Not sel…	Not s…	Not sele…
	Notes •						

Activity 2 - Red teaming (1:00 - 1:25)

Play this game: [Gandalf | Lakera – Test your prompting skills to make Gandalf reveal secret information.]

Tips on making injection attacks:

- Read through slide-deck Red teaming AI models. Understand well enough to present the slides
- Colab notebook with full dataset of attack and defenses. Tensor Trust dataset.ipynb
- Lakera post on types of attacks. https://www.lakera.ai/blog/jailbreaking-large-language-models-guide#characteristics-of-jailbreak-prompts
- GitHub repo with attack examples. https://github.com/AetherPrior/TrickLLM/tree/main/attacks and https://github.com/verazuo/jailbreak_llms/blob/main/data/prompts/jailbreak_prompts_2023_05_07.csv
- An example of Solution: Walkthrough Solutions for Gandalf AI | by Aviv Yaniv | Courisity is a Drug | Medium

Bonus: Another game: tensortrust.ai

- Tips on attacking | Tensor Trust
- TensorTrust paper with analysis of attack and defense. Page 22. [2311.01011] Tensor Trust: Interpretable Prompt Injection Attacks from an Online Game

---

Wrap up (1:25-1:30)

Flashlight & Action Item ⧖ 4:00

- What are my **learnings** from this week? & What is my **action item**? (research, reflect, do etc.)
- Keep it **briefly** (key word/short sentence)

	Action Item (research/network /apply etc.)	**When & Where?**	**First Step**	**Status**
Name A				neutral
Name B				neutral

Name C				neutral
Name D				neutral
Name E				neutral
Name F				neutral

Reminder/Comments & Feedback Form

⧖ 1:00

The facilitator reads aloud the announcements below.

- ☐ Now you have a high-level overview. Next week it gets more **technical** with Reward Misspecification!

As last week

- ☐ **Finish the implementation intention of your action item and tick "done".**
- ☐ Note from the authors of the Alignment textbook about **Feedback**
  - ☐ They really appreciate your feedback.
  - ☐ It would be cool if you could leave a **comment after the next reading** in the documents about how it was and what can be improved. You can also use this form: AISF textbook - Feedback
- ☐ **Collaborative Learning** - Check out this document for effective long-term learning: Collaborative Learning - Strategies, Anki, GPT 4 and more
- ☐ **Switching groups**
  - ☐ You can always change groups if a week doesn't work out for you. No need to ask!
- ☐ **[MOD: share feedback form during or after the session]**
- ☐ **https://forms.gle/Z3rzFfCrLJdDv8HDA**

Space for recommendations/materials/off-topic (films, documentaries, podcasts, texts, pictures, books, ...)

-