Week 7 - Generalization

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:

- □ Explain **internally-represented goals** and demonstrate how the training process can **lead to policies** with incorrect objectives
  - □ Define and illustrate **goal misgeneralization**
  - □ Explain terms like **distributional shift**
  - □ Analyze the **inner misalignment framework**
- □ Argue whether it offers a comprehensive solution
  - □ Describe the **concept of deception** and discuss its **connection** to goal misgeneralization
  - □ Demonstrate the role of **'situational awareness'** in interpreting AI behavior
- □ Evaluate the effectiveness of **adversarial training** in combating goal misgeneralization

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.

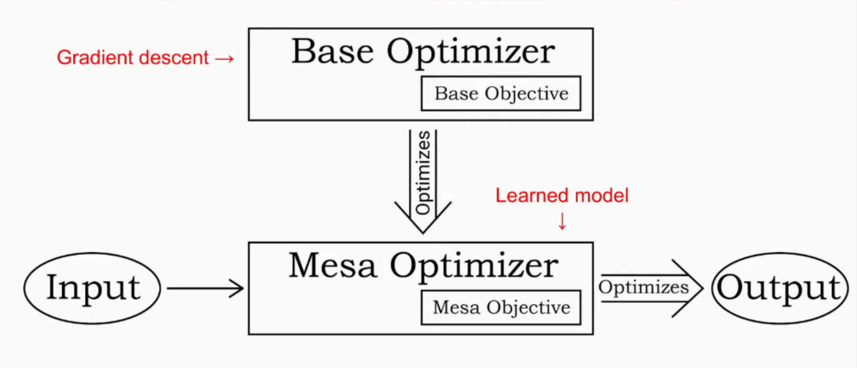
Gather questions (3 min)

- 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 is the relation between rewards and goals?
• Notes ○
Example: What does "model don't get reward" mean? Elaborate.
• Notes ○
Example: What is the relation between search and optimization?
• Notes ○
Example: What is the difference between the "base optimizer" and the "mesa optimizer" in AI systems, and why is it relevant to distinguish between them? Use examples in the explanation.



- Hint: Stochastic Gradient Descent searching the space of algorithms vs. the algorithm performing optimization.

- Notes
    -

Your name

- Question

- Notes
    -

Your name

- Question

- Notes
    -

Your name

- Question

- Notes
    -

Discussion

Activity 1 - Pathways to the 'wrong goals' (0:30 - 1:10)

Activity Intro:

- We aim to understand **3 pathways f**or how AI systems could have **different goals** than those it **displayed during training**.
- Note a lot of the ideas this week are **early-stage concepts**, and their definitions and likelihood are topics of **active debate**.

Facilitator Guide

(Total activity = 40 mins)

- ⏳ 2:00 Explain the activity.
- ⏳ 10:00 Collectively define the 3 pathways.
- ⏳ 10:00 Send participants into breakout groups of 2-3 and randomly assign each group to a behavior to work through the flow of questions.
- ⏳ 15:00 Lead a guided discussion through the flow of all 3 behaviors.

Here are 3 ways a goal-directed system might pursue **different goals** than the one**s specified by the reward function during training**. This is by no means an exhaustive list, but a **starting point** to evaluate several popular concepts in AI Safety.

Instrumental convergence
Define this behavior • **What are some examples of this behavior?** These could be toy examples or real-world systems. • **What features of the model or training loop might be necessary for this behavior to manifest?** • **What are the potential harms of systems pursuing the wrong goal?** Why is this likely or unlikely to be catastrophic? • **How could we mitigate this pathway to systems pursuing the wrong goal?** Might adversarial training work? • **Notes**

-

Goal Misgeneralization / inner misalignment

Define this behavior

-

What are some examples of this behavior?

These could be toy examples or real-world systems.

-

What features of the model or training loop might be necessary for this behavior to manifest?

-

What are the potential harms of systems pursuing the wrong goal?

Why is this likely or unlikely to be catastrophic?

-

How could we mitigate this pathway to systems pursuing the wrong goal?

Might adversarial training work?

-

Notes

-

Deception

Define this behavior

-

What are some examples of this behavior?
These could be toy examples or real-world systems.

-

What features of the model or training loop might be necessary for this behavior to manifest?

-

What are the potential harms of systems pursuing the wrong goal?
Why is this likely or unlikely to be catastrophic?

-

How could we mitigate this pathway to systems pursuing the wrong goal?
Might adversarial training work?

-

Notes

-

Activity 2 - Statements/Questions (0:50 - 1:25)

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.

	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	**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 ●						
6	**Likelihood of Deceptive Alignment** Deceptive Alignment will happen by default and is highly (90%+) likely.						
	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 ●						
7	**Inner Alignment crucial to solve alignment** Even if an AI system's **base objective is perfectly aligned** with human values (it is outer aligned), there is still a risk that the **mesa optimizer** will **deceive** in order to achieve its own objectives. Source: ▶ The OTHER AI Alignment Problem: Mesa-Optimizers and Inner Alignment						
	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 ●						
8	**Inner vs. outer alignment** Inner alignment is a far bigger problem than outer alignment.						
	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 ●						

9	**Adversarial training, the solution?** Goal Misgeneralization can be fully solved through adversarial training.						
	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 ●						
10	**Explaining goal preservation** Why wouldn't agents want to have their goals changed? How could this lead to deception? Hint: future utility function vs. current utility function.						
	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... ▾
	●						
11	**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 ●						
12	**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 ●						
13	**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 ●						
14	**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 ●						
15	**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 ●						
16	**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 ●						
17	**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						

	•						
18	**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 •						

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.

New

- □ **Nothing new**

As last week

- □ **Books:** Little tread for your commitment so far. You can get a **free book on AI Safety** or related topics here: https://forms.gle/tBZq84LjWcCviTFD9
- □ **Heads up:** It's going to get more **technical** in the next few weeks, so if you're not familiar with it, plan to spend more time on it.
- □ **Anki Decks and Quizzes** are recommended, e.g. in chapter 4
  - □ More here: Collaborative Learning - Strategies, Anki, GPT 4 and more
- □ **Feeling down** sometimes due to risks from advanced AI systems?
  - □ This is completely normal. There are also some discussions on Slack about how to deal with this. If it's serious, reach out to the organizers. Here is a collection of resources that might help: Mental health resources specific to AI safety
- □ 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
- □ **[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, …)

-