Week 6: Specification

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:

- □ Analyze reward function limitations in conveying intentions:
  - □ **Define reward functions** and their purpose
  - □ Explain **reward misspecification** and **reward hacking**
  - □ Assess reward hacking's potential for **catastrophic AI failures**
- □ Evaluate current approaches as a solution to reward misspecification:
  - □ **Define** RLHF, PHF, RLAIF and other
  - □ Point out the **advantages** and **disadvantages** of the different technologies.
  - □ Make an initial guess as to whether these techniques will **scale to align superintelligence**.

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 reward misspecification?
• Notes ○
Example: What is the difference between RLHF and IRL?
• Notes ○
Example: What is the difference between Behavioral Cloning (BC) and Procedural Cloning (PC)? Can this approach lead to superhuman capabilities?
• Notes ○
Example: What is the difference between Reward Functions vs. Value Functions?
• Notes ○
Your name • Question
• Notes ○

Your name

- Question

- Notes
    - ○

Your name

- Question

- Notes
    - ○

Your name

- Question

- Notes
    - ○

Discussion

Activity 1 - Understanding reward misspecification (0:30 - 0:50)

Activity Intro

- □ Reward misspecification is an issue for systems today, and we have **no reason** to think it'll go **away** in systems **tomorrow**.
- □ To do productive alignment research, it's **important to understand reward misspecification**, and to be able to '**think like a reward maximizer**', to anticipate what might go wrong with it.

Defining Goodhart's law

⌛ 5:00 Explain Goodhart's law, give some time to allow participants to read the definition, and answer any clarifying questions as a group.

"When a measure becomes a target, it ceases to be a good measure."

□ Goodhart's law is closely related to 'reward hacking'. When we want to achieve some target (example: happiness), we often have to define some measure of that target (example: GDP).

□ When we define a measure, we often start optimizing for that measure, rather than the original target. This is closely related to 'reward misspecification': we define a reward function that is supposed to measure some target, but the system is, in fact, optimizing for the **metric**, and not the **target** you intended.

□ It is likely impossible to specify a reward function (metric) that exactly measures your target. Hence, ML systems are susceptible to Goodhart's law, just as humans are.

"In machine learning, this effect arises with proxy objectives provided by static learned models, such as discriminators and reward models." (Gao et al.)

Reward misspecification in human society

⌛ 3:00 Participants write some examples of Goodhart's law they can think of from human society.

	Area: Terminal goal	**Measurement (Target)**	**Problem through misspecification**
Example 1	School system: Learning **long term** about a subject	One-time testing through 2 h exam	Students prepare just shortly before the exam and forget the most long term
Example 2	Social Media: Connecting people + making a profit	Screen time & engagement & engagement metrics	Doom-scrolling, mindless content, less real life connections Showing more extreme content → higher screen time
Name A			
Name B			
Name C			
Name D			
Name E			
Name F			
Discussion notes	•		

Reward misspecification in ML models

⏳ 3:00

- Participants write some examples of Goodhart's Law in **ML models**. They can use events that have happened or events that could happen.
- Write down some ideas how the reward function could be **improved**. Could new problems arise?

⏳ 4:00

- Discuss shortly the examples and how the misspecification could be improved. What are general approaches that try to solve misspecification?

	Area: Terminal goal	**Measurement (Target)**	**Problem through misspecification**
Example 1	Boat race: Win the race	Reward through in-game points	Boat farms in endless cycle power ups
● Improvement on the margin: Reward through driving through the finish line ● New problems?			
Name A			
● Improvement on the margin: [how could be the measurement improved] ● New problems?			
Name B			
● Improvement on the margin: [how could be the measurement improved] ● New problems?			
Name C			
● Improvement on the margin: [how could be the measurement improved] ● New problems?			
Name D			
● Improvement on the margin: [how could be the measurement improved] ● New problems?			
Name E			
● Improvement on the margin: [how could be the measurement improved] ● New problems?			
Discussion 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	**Correct task specification: An unsolved problem** Correct task specification is extremely difficult. All current techniques, e.g. IRL, RLHF are not sufficient.						
	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	**Scale of the problem** Incorrect task specification would lead to catastrophic outcomes. Extra: Why or why not do you think it could lead to a catastrophic outcome?						
	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	**Feasibility of outer alignment**						

	It is possible to design a reward function that captures the values of a single person. An extensive list of heuristics could potentially describe the totality of a person's values and ethics.						
	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	**Alignment the only problem?** When a reward function matches perfectly an individual's values, the alignment problem is solved. There are no more technical or societal open questions. Extra: What could be other problems that need to be fixed for reaching utopia?						
	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 ●						
11	**Imitation learning a solution?** The most promising approach is imitation learning, such as behavioral cloning and inverse reinforcement learning. Imitation learning has less of the reward hacking problem.						
	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	**RLAIF the solution?** The biggest problem with reinforcement learning from human feedback is that it doesn't scale to AGI. At some point, we can't give accurate feedback about whether a complex scientific proposition is correct, and reinforcement learning from AI feedback will be the solution to that, so it's the most promising approach.						
	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 •						
19	**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

- □ **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 Quizze**s 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

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
- □ **[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, …)

-