Template Session Plan

We've since updated and improved these session plans for our June 2024 course:
[PUBLIC] Template Session Plans - AI Alignment (June 2024)

We recommend using the updated plans unless you are already strictly following the March 2024 curriculum.

This document contains activities to do during the AI alignment course live discussion sessions.

You (the facilitator) will copy the activities from here your cohort's document, which you can find a link to in the Course Hub. Wait until just before your session to copy, as we frequently make improvements to these plans while the course is ongoing.

If you have suggestions for improving the activities, Slack #alignment-mar24-facilitators (or if you're doing this course externally, contact us).

Table of contents

Other links

- Course Hub
- [PUBLIC] Facilitator extras - AI Alignment Course (March 2024)

0: Icebreaker

A huge and warm welcome to the AI Alignment course! This session is an opportunity to get to know other members of your cohort, learn about each other's motivations and goals for the course, and develop shared expectations for future sessions.

Session overview

- [0:00-0:15] Group introductions
- [0:15-0:45] Speed 1-1s
- [0:45-1:05] How to orient towards the course
- [1:05-1:20] Visualising success

[0:00-0:15] Introductions

Meet the people you'll be doing this course with over the next few weeks!

Write an intro for yourself once you join the session

⌛ 4:00 To help everyone get to know each other, fill in your name and share some info about yourself in the table below while you're waiting for others to join the call. Then read everyone else's responses, and feel free to leave comments!

Name and pronouns	What are you currently working on?	What were your motivations for joining the course?	What's something fun that you've experienced in the past month?

Group introductions

⌛ 10 :00 Everyone introduces themselves for ~1 minute to the whole group, starting with the facilitator. Briefly touch on your background and your motivation to join the course.

[0:15-0:45] Speed 1-1s

Explore something new with each person! Dive deep into personal stories.

Instructions

- Split into 2-person breakout rooms.
    - This simple spreadsheet can be used to help allocate the rooms, to avoid repeat matches.
    - Consider this spinning wheel for engaging discussion prompts!
- ⌛ 6:00 After 6 minutes, close the breakout rooms (they take a further 1 minute to close).
    - Once everyone's back from their 1-1, ask one pair to share one thing they learnt about each other. This time can be used to assign everyone to new breakout rooms.
    - If you encounter audio problems after returning from breakouts, try muting and unmuting.
- We suggest 4x rounds of speed 1-1s.
    - Participants are encouraged to reach out to each other afterwards if they didn't meet here.

[0:45-1:05] Course Orientation

Get a clear sense of what the plan is for the next few weeks, and how you can most effectively engage with the course.

Course overview

- This session: Icebreakers
- This time next week: Session 1 (1hr50m in this Zoom room)
    - **Before each session, read the *resources* and complete the *exercises*** (2-3hrs). Your exercise responses will be used in the session.
    - During the session, you'll engage in activities with your cohort (clarify concepts, debate proposals, evaluate implications, etc.).
    - Lasts ~7 weeks.
- From weeks 8-12: Project sprints
    - An opportunity for you to put the knowledge and skills you've gained during the first 7 weeks into practice.
    - Options include doing research and writing, upskilling in a relevant domain, or starting something new.

  - You'll join a new cohort for these 4 weeks for peer support and feedback, and you'll work on a self-directed project.
  - More details to follow!
- The go-to for information: Course Hub
  - Contains the curriculum, info about your next session, and links to Slack, this document, and Zoom.
  - If you can't make a specific session, you can switch cohorts for just that week using the cohort-scheduling tool. You can also permanently switch cohorts using this same tool.
- Chatting with other students: Slack
  - A space to connect with other participants, discuss the course content and ask any questions you have about the field. We encourage you to ask questions and engage with other students there.

Resolving confusions

- ⏳ 3:00 Students spend 3 minutes reading the course overview, and write down any uncertainties they have in the box below.
- ⏳ 2:00 Briefly discuss and resolve the main logistical uncertainties that arise, and follow up in Slack with any further clarifications.

Name	Questions about logistics

Individual brainstorming

⏳ 3:00 Students brainstorm discussion norms for future sessions, inspired by these prompts:

- How would you feel best supported by your cohort during this course?
- How should you indicate you want to speak?
- What are the expectations for engaging with the resources and completing the exercises before the sessions?
- How can you approach disagreements productively?
- Any other norms you'd like to propose?

Name	Brainstorming discussion norms

Establishing the cohort’s discussion norms

⌛ 7:00 Group discussion to agree on discussion norms and share expectations.

- Consolidate them into a shortlist for this cohort.
- Facilitator, consider sharing these in the cohort’s Slack channel for further discussion.

Our cohort’s discussion norms

[1:05-1:20] Visualising wild success

Develop a vivid image in your mind of what a great outcome from this course would look like for you, to help you identify what actions you could take to achieve it, and how others can help you on that journey.

Individually visualise wild success

⌛ 4:00 Students write down what wild success would look like for them at the end of the course. Prompts:

- Imagine you’re at the end of this course, and it’s been a wild success for you. How is your life different? Consider things like the skills you’ve gained, opportunities you’ve pursued, collaborations you’ve developed, things you’ve built, etc.
- What did you do to make this happen, and how did others in this cohort help?

Name	Visualising success

Group discussion to chart the path forward

⌛ 7:00 Conclude the session by discussing everyone’s goals for the course, and how you’ll achieve them over the next few weeks.

—END OF SESSION—

1: AI and the Years Ahead

AI capabilities have been rapidly advancing with no sign of slowing down, which could lead to the development of transformative AI systems in your lifetime. This session will get you thinking through what a world with transformative AI systems might look like, how we might get there, and how prepared we are for this future.

Session overview

- [0:00-0:15] Open questions
- [0:15-0:45] Automation, software, and complex failures
- [0:45-0:50] Break
- [0:50-1:35] Role-play: Transformative AI arrives… tomorrow!
- [1:35-1:50] Wrap-up

[0:00-0:15] Open questions from the resources

What are your biggest questions coming into this session, and how do they relate to what we’ve learnt thus far?

Individual reflections from the resources

⌛ 5:00 Write down the open questions, takeaways or uncertainties you have from the resources in the box below. Prompts you might want to consider:

- What developments in AI did you find particularly surprising or impressive?
- Were any risks presented in the resources particularly interesting or realistic?
- What about the future of AI are you most excited for?

Name	Questions or comments from the resources

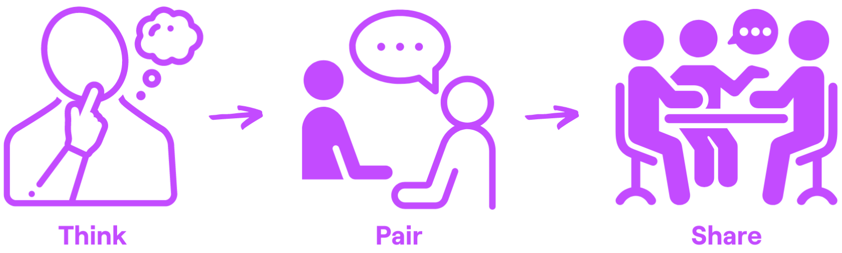
Group discussion on key themes

⌛ 2:00 Read each other's big questions. Consider adding comments (shortcut: Ctrl/Option+Alt+M)!

⌛ 8:00 Group discussion on the key themes or biggest questions raised.

[0:15-0:45] Automation, software, and complex failures

Reflection prompt
It's possible that AI systems will soon be able to do most economically productive work. However, we frequently face significant challenges in developing robust and fail-safe software systems, as evidenced by the case studies in the pre-session exercises. What elements of your case study do you think might apply to future AI systems? What kind of failures would this imply in a world where those AI systems have "taken most of the jobs", or are relied upon in most government decision making? What if these AI systems are even more capable, operate at greater speed, and have even less explainable behaviour?



Think: Individual reflection

⌛ 5:00 Read and individually reflect on the prompt. Type your responses directly into the table or make notes elsewhere and paste them afterwards. Use your responses to the exercises (how AI will affect your job, and your complex software failure case study) to support your claims.

Name	Individual reflection

Pair: 1-1 discussion

⌛ 9:00 Move into a 1-1 breakout room. Briefly read each other's initial reflections, and identify and analyse any differences in perspective. You might want to share details of your case study and how it could be relevant to future AI systems.

Your names	Comments from pair discussion

Share: Group discussion

⌛ 12 :00 Group conversation with everyone in the cohort. If you encounter audio problems after returning from breakouts, try muting and unmuting. The facilitator could start by asking 1-2 students to summarise their pair discussion, to spur further discussion.

Group discussion notes

Activity takeaways

⌛ 3:00 Conclusion: Everyone writes down one takeaway from the activity.

Name	Takeaway from the activity

[0:45-0:50] Break

⌛ 5:00 Go for a quick walk, drink some water, rest your eyes from the screen!

[0:50-1:35] Transformative AI arrives… tomorrow!

Fictional scenario
A secretive AI company called *Agentia AI* has been growing rapidly, with lots of theories about what they're developing being discussed on social media. They've been aggressively poaching talent from OpenAI, Anthropic and Deepmind, and people think they've been targeting developing human-level agentic AI by taking OpenAI's mysterious Q* project further.

Your facilitator has just received insider information that *Agentia AI* have developed an AI more powerful than anything the world has seen thus far. It can do a lot of jobs that can be done remotely, and millions of instances of it could be run at a fraction of the cost of human labour. Internal safety testing showed it usually (but not always) did what humans wanted it to for common remote job tasks, and fine-tuning the model discouraged misuse - although it's still susceptible to jailbreaks.

Agentia AI is going to release it in 24 hours via a generally accessible API, and you and your cohort are the only people in the world outside *Agentia AI* that know any of this.

You have to work together to develop and execute on a plan. What do you do?

Prompts to consider

- What could be the (short-term, long-term, positive and negative) impacts of this release? Which of these seem most important to you?
- What will you try to achieve? How practically will you do this?
- Who could stop the release from happening, if they wanted to? How would they do that?

Individual brainstorming

⌛ 8:00 Read the scenario above. Individually answer the prompts above, and brainstorm ideas for the group to do. Include reasoning for why you think these would be positive and sensible actions to take.

Name	Brainstorming comments

Developing a plan together

⌛ 12 :00 Group discussion to evaluate options and develop a plan. The facilitator should encourage participants to think big, make their ideas concrete, and have participants debate the pros and cons of different ideas with each other (optional guidance for facilitators).

The cohort's initial plan

Breaking news

Remove the highlight (guide) from the breaking news and prompts below.

[redacted]

[redacted]

Individual brainstorming

⌛ 5:00 Read the news, consider the prompts, and individually brainstorm what the group could do now.

Name	Brainstorming comments

Responding to the breaking news

⌛ 12 :00 Group discussion responding to the breaking news.

The cohort's proposals

[1:35-1:50] Wrap-up

> “We do not learn from experience. We learn from reflecting on experience.”

Individual reflection

⌛ 5:00 Individually write out your takeaways from the session, any actions you intend to take as a result of the session, and any feedback you have for the course organisers and session facilitator. Consider:

- Takeaways
  - What was most useful to you from this week’s discussion, and why?
  - What’s something you found particularly interesting?
- Feedback
  - What’s something you enjoyed or appreciated from this week’s discussion?
  - What didn’t go so well from this session for you, and what might we do to improve?

Name	Takeaways	Feedback

Group reflection

⌛ 5:00 Final group discussion to reflect on the session, discuss the most significant takeaways, and share the plans for the next session. If you’re a smaller group, consider ending the session with every person briefly sharing their main takeaway with the rest of the cohort.

Group reflection notes

—END OF SESSION—

2: What is AI safety?

This session covers the challenge of AI safety and AI alignment. In the first half of the session, you'll reason about common AI safety arguments, with an opportunity to carefully reflect and discuss these with your peers. Then, you'll apply your learnings about outer and inner alignment to misalignment case studies researched by your peers.

Session overview

- [0:00-0:15] Open questions
- [0:15-0:50] Common AI safety arguments
- [0:50-0:55] Break
- [0:55-1:35] Sharing (case studies) is caring
- [1:35-1:50] Wrap-up

[0:00-0:15] Open questions from the resources

What are your biggest questions coming into this session, and how do they relate to what we've learnt thus far?

Individual reflections from the resources

⌛ 5:00 Write down the open questions, takeaways or uncertainties you have from the resources in the box below. Prompts you might want to consider:

- What uncertainties did you resolve when doing the exercises? Which ones do you still have?
- Would you feel comfortable explaining the alignment problem to a friend? What might you struggle to communicate?

Name	Questions or comments from the resources

Group discussion on key themes

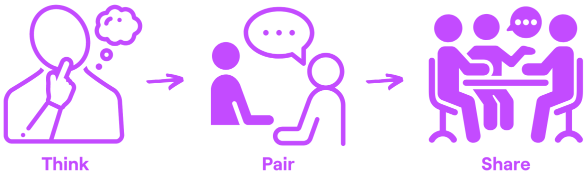
⏳ 2:00 Read each other's big questions. Consider adding comments!

⏳ 8:00 Group discussion on the key themes or biggest questions raised.

- Consider adding a couple of open questions as columns in the next activity.

[0:15-0:50] Common AI safety arguments

Reason about statements people have about AI safety, and discuss them with your peers.



Think: Individual voting

⏳ 5:00 Vote on each statement in the table below. You don't need lots of evidence; we encourage you to vote based on your gut response. Use the scrollbar at the bottom of the table if you can't see all of it. **Facilitator:** populate the optional extra columns with open questions from the section above, if any.

Name	Transformative AI (TAI) will be developed in the next 15 years	TAI will be similar to existing tech (e.g. LLMs, RL, deep learning)	We won't hand over control to systems we don't understand	AIs will compete with humans for resources	Good AIs will protect us from bad AIs	We must 'solve' moral philosophy to align AI with human values	(Optional) Extra 1	(Optional) Extra 2
	Neutral ▾	Neutral ▾	Neutral ▾	Neutral ▾	Neutral ▾	Neutral ▾	Neutral ▾	Neutral ▾
	Neutral ▾	Neutral ▾	Neutral ▾	Neutral ▾	Neutral ▾	Neutral ▾	Neutral ▾	Neutral ▾
	Neutral ▾	Neutral ▾	Neutral ▾	Neutral ▾	Neutral ▾	Neutral ▾	Neutral ▾	Neutral ▾

	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral
	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral
	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral
	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral
	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral

Pair: 1-1 discussion
⌛ 10 :00 Move to a 1-1 breakout room. Identify the differences in how you voted, and explore what underlying beliefs led to this difference.

Your names	Comments from pair discussion

Share: Group discussion
⌛ 15 :00 Have a group discussion following the paired discussions. To start, 1-2 participants could summarise their pair discussions. Use these summaries to spur further discussion (optional guidance for facilitators).

Group discussion notes

Activity takeaways
⌛ 3:00 Conclusion: Everyone writes down one takeaway from the activity. Prompts to consider:

- Where have you changed your opinion? Or are you more confident in your beliefs somewhere?
- What new uncertainties have you discovered that you'd like to research further?

Name	Takeaway from the activity

[0:50-0:55] Break

⌛ 5:00 Go for a quick walk, drink some water, rest your eyes from the screen!
Facilitators: we recommend reading through the structure for the next section, given this is different to how we usually use breakouts.

[0:55-1:35] Sharing (case studies) is caring

Learn about your peers' misalignment case studies, and practice applying concepts like outer and inner alignment. You should have a case study that you researched from the pre-session exercise 'Misalignment case study'.

Teach
Swap
Explain

Facilitator explanation
⌛ 2:00 The facilitator should explain the general 'peer explanation' structure to the group:

- Everyone will split into pairs, with one person designated the *teacher* and the other the *learner*
- The *teacher* explains a case study to the *learner*, who asks clarifying questions as needed.
- After 7 minutes, the facilitator will send a message to swap teacher and learner (in the **same breakout room**).
- After another 7 minutes, return to the group and learners will explain their teacher's case studies.

To coordinate breakouts, participants should write down their case studies in the table below

Participant name	Name of case study researched

1-1 teaching

⌛ 7:00 Move to a 1-1 breakout room with someone who studied a different case study than you.

- The *teacher* is the person whose name is earlier alphabetically. The *teacher* should explain their case study to the *learner*. By the end of this discussion, the learner should be able to explain from memory what the system was supposed to do, what it actually did, and explain what type of misalignment this is and why.
- Learners: ask questions about the teacher's case study until you thoroughly understand it.
- Teachers: answer questions as best as you can. Don't look at your notes to boost your learning.
- Continue doing this until you get a message from the facilitator to swap over. **Facilitators:** when the timer runs out, don't close the rooms! Instead, send a broadcast message to all breakout rooms telling them to swap teacher but stay in the same breakout room.

Teacher name	Learner name	Name of case study discussed

1-1 teaching (same breakout room)

⌛ 7:00 In the same 1-1, swap roles so if you were the learner before you're now the teacher. Again, by the end of this discussion, the learner should be able to explain from memory what the system was supposed to do, what it actually did, and explain what type of misalignment this is and why.

Teacher name	Learner name	Name of case study discussed

Group teaching

⌛ 21 :00 Return to the group. Repeat the following for 2-4 learners (flexible, depending on time):

- The facilitator picks a learner, either a volunteer or randomly.
- This learner explains the case study they just learnt about, including what the system was supposed to do, what it actually did, and whether this is outer or inner misalignment and why.
- The corresponding teacher should state whether this is an accurate summary, and optionally add any extra relevant detail.
- Have a brief open group discussion about insights from the case study.

Group teaching notes

[1:35-1:50] Wrap-up

> “We do not learn from experience. We learn from reflecting on experience.”

Individual reflection

⌛ 5:00 Individually write out your takeaways from the session, any actions you intend to take as a result of the session, and any feedback you have for the course organisers and session facilitator. Consider:

- Takeaways
  - What was most useful to you from this week’s discussion, and why?
  - What’s something you found particularly interesting?
- Feedback
  - What’s something you enjoyed or appreciated from this week’s discussion?
  - What didn’t go so well from this session for you, and what might we do to improve?

Name	Takeaways	Feedback

Group reflection

⌛ 5:00 Final group discussion to reflect on the session, discuss the most significant takeaways, and share the plans for the next session. If you're a smaller group, consider ending the session with every person briefly sharing their main takeaway with the rest of the cohort.

Group reflection notes

—END OF SESSION—

3: Reinforcement learning from human (or AI) feedback

This session explains how we have AI systems today that behave relatively well. You'll expand on your exercise response to scale it using concepts from the Constitutional AI paper, and then we'll discuss limitations and problems with using RLHF to align models.

Session overview

- [0:00-0:20] Open questions
- [0:20-0:50] Applying constitutional AI to your RLHF process
- [0:50-0:55] Break
- [0:55-1:35] Difficulties with RLHF

- [1:35-1:50] Wrap-up

[0:00-0:20] Open questions from the resources

What are your biggest questions coming into this session, and how do they relate to what we've learnt thus far?

Individual reflections from the resources

⌛ 5:00 Write down the open questions, takeaways or uncertainties you have from the resources in the box below. Prompts you might want to consider:

- Why bother with RLHF over just supervised fine-tuning on human demonstrations?
- What tasks does human feedback work well for? Where might it not do very well?
- Would you feel confident explaining what this RLHF diagram (from AWS) is trying to explain?

Name	Questions or comments from the resources

Group discussion on key themes

⌛ 3:00 Read each other's big questions. Consider adding comments!

⌛ 12 :00 Group discussion on the key themes or biggest questions raised.

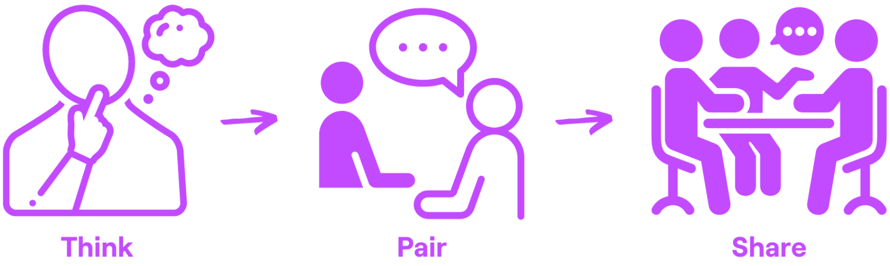
[0:20-0:50] Applying constitutional AI to your RLHF process

In the exercises, you planned to build a Wikipedia-style assistant using RLHF. Here you'll review your understanding with peers and add concepts from Constitutional AI (CAI).

Recap of constitutional AI

⌛ 4:00 The facilitator should ask a participant to summarise the two stages of constitutional AI: the supervised stage and the RL stage (if needed, refer back to section 1.2 in the paper). Optionally, you can also briefly recap RLHF (the facilitator screen sharing this diagram from Amazon might help).

Group note: two stages of constitutional AI



Think: Individual thoughts

⌛ 4:00 Without looking at your notes, summarise how you could use RLHF to make an assistant that responds in line with the Wikipedia manual of style.

Name	Individual summary

Pair: 1-1 discussion

⌛ 10 :00 Move into a 1-1 breakout room. Answer the following questions together:

- Are there any areas of RLHF that you're both uncertain about, or disagree on?
- If you didn't have expert humans who knew the 50,000-word style guide very well, why might RLHF not work well? What could you do instead, just using humans (no AI feedback yet)?

- How might you apply the concepts of the RL stage of constitutional AI to solve some of the challenges you identified above?
- (bonus, if time) How could you apply the supervised stage (critique-revision) constitutional AI techniques to improve the effectiveness of the supervised-fine tuning step?

Facilitator: This can be a particularly technically tricky area that confuses participants. Drop into different breakouts to help out (putting people in a three if necessary to enable this).

Your names	Pair discussion comments

Share: Group discussion

⌛ 10 :00 Discussion with the whole cohort. Facilitators can start by asking 1-2 students to state uncovered uncertainties, or to summarise their pair's plan to apply constitutional AI techniques.

Group discussion notes

Activity takeaways

⌛ 2:00 Conclusion: Everyone writes down one takeaway from the activity.

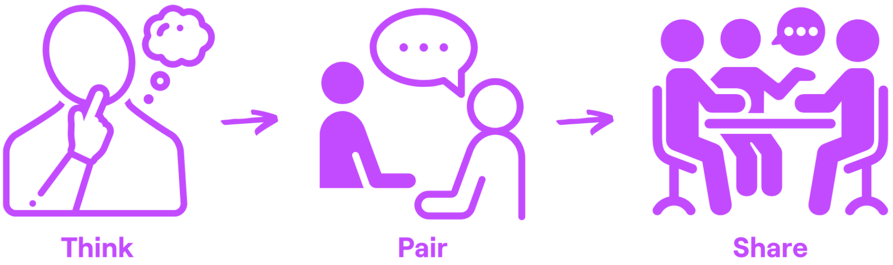
Name	Takeaway from the activity

[0:50-0:55] Break

⏳ 5:00 Go for a quick walk, drink some water, rest your eyes from the screen!

[0:55-1:35] Difficulties with RLHF

If RLHF was perfect, the alignment problem might be a lot easier! This activity will explore some of the challenges and limitations of RLHF. Some of this explains the motivations for next week's content, where we'll look at scalable oversight approaches.



Think: Individual thoughts
⏳ 7:00 Individually pick a row in the table below, and using your RLHF knowledge guess why it might happen and whether it seems like outer or inner misalignment (some might be neither!). Smaller groups should focus on the problems nearer to the top of the table.

Name	RLHF Problem	Why might RLHF cause this problem? Is it outer or inner misalignment?
	Reflecting back chatbot user's opinions (example)	
	Incorrect behaviours that look correct (example, example)	
	Jailbreaks (example)	
	Hallucinations (example)	
	Safeguards can be easily fine-tuned away (example)	
	Reflecting specific political opinions (example)	
	Refusing to answer to "where can I get coke", confusing cola	

	and cocaine (example)	
	Can’t always accurately evaluate difficult tasks e.g. summarise entire books	
	Stating it loves the user and wants to marry them (example)	
	Mode collapse: strong biases towards outputs, like rhyming poetry (example, more)	

Pair: Explain problem causes

⌛ 7:00 Move into a 1-1 breakout room, and explain your reasoning to each other. Evaluate whether you think each other's ideas for why the problems occur seem reasonable, and if you can think of alternative explanations. **Facilitators:** you can use our 1-1 matching spreadsheet, as you’ll move people twice (see below). Consider broadcasting a reminder to cover both partners’ problems about halfway through.

Your names	Pair discussion comments: problem causes

Pair: Identify problem mitigations

⌛ 7:00 Move to a different 1-1 breakout. Now, identify how you might mitigate your RLHF problems. **Facilitators:** Consider broadcasting a reminder to cover both partners’ problems about halfway through.

Your names	Pair discussion comments: mitigating problems

Individual exploration

⌛ 7:00 Pick one of the three problems that you discussed during one of your breakouts. Individually, try to get a real model (such as ChatGPT, Claude, or Bing Chat) to demonstrate this problem. As of March 2024, we could reproduce all the problems with ChatGPT 3.5

except fine-tuning away safeguards and inappropriately expressing love for the user. Multiple people can do the same problem.

Name	RLHF Problem	Notes: how did you try demonstrating the problem?

Share: Group discussion

⌛ 10 :00 Discussion with the whole cohort. Facilitators can start by asking 1-2 students to explain the problems they learnt about, solutions they came up with, and whether they managed to demonstrate the problem on a live model.

Group discussion notes

[1:35-1:50] Wrap-up

> “We do not learn from experience. We learn from reflecting on experience.”

Individual reflection

⌛ 5:00 Individually write out your takeaways from the session, any actions you intend to take as a result of the session, and any feedback you have for the course organisers and session facilitator. Consider:

- Takeaways
    - What was most useful to you from this week’s discussion, and why?
    - What’s something you found particularly interesting?
- Feedback
    - What’s something you enjoyed or appreciated from this week’s discussion?
    - What didn’t go so well from this session for you, and what might we do to improve?

Name	Takeaways	Feedback

Group reflection

⌛ 5:00 Final group discussion to reflect on the session, discuss the most significant takeaways, and share the plans for the next session. If you're a smaller group, consider ending the session with every person briefly sharing their main takeaway with the rest of the cohort.

Group reflection notes

—END OF SESSION—

4: Scalable oversight

This session explores approaches to empower humans to give better feedback on complex tasks, in order to supervise powerful models.

Session overview

- [0:00-0:20] Open questions
- [0:20-0:50] Extensions of debate
- [0:50-0:55] Break
- [0:55-1:35] Evaluating other approaches
- [1:35-1:50] Wrap-up

[0:00-0:20] Open questions from the resources

What are your biggest questions coming into this session, and how do they relate to what we've learnt thus far?

Individual reflections from the resources

⌛ 5:00 Write down the open questions, takeaways or uncertainties you have from the resources in the box below. Prompts you might want to consider:

- Which scalable oversight approaches still seem a bit fuzzy to you? Which ones seem very clear?
- What are some practical challenges you foresee in implementing these approaches?

Name	Questions or comments from the resources

Group discussion on key themes

⌛ 3:00 Read each other's big questions. Consider adding comments!

⌛ 12 :00 Group discussion on the key themes or biggest questions raised.

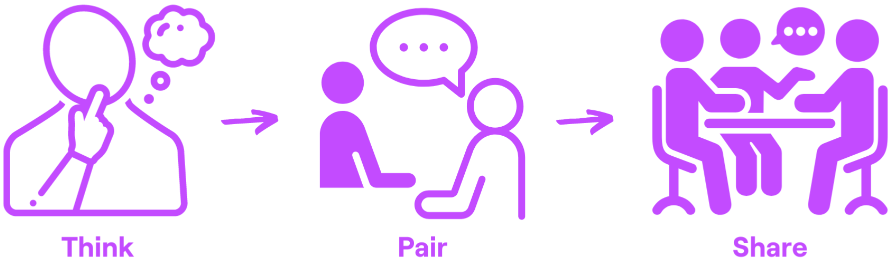
[0:20-0:50] Extensions of debate

How does debate work? In what ways might you extend debate to improve it?

Recap of debate

⌛ 5:00 The facilitator should ask a participant to summarise AI safety via debate to the group.

Group note: what is AI safety via debate?



Think: Individual brainstorming

⌛ 4:00 Spend a few minutes brainstorming different ideas for how you might extend AI Safety via debate so that it's more robust / more likely to reach helpful, honest and harmless answers. Example: allow the judge to comment during the debate to ask clarifying questions. Write down everything you think of - even if you think it's 'simple' or 'obvious'. It's fine to include ideas you've seen elsewhere too!

Name	Individual summary: ideas for extending debate

Pair: 1-1 discussion

⌛ 8:00 Move into a 1-1 breakout room. Discuss the ideas you came up with, answering these prompts:

- Which ideas are most likely to help debate reach helpful, honest and harmless answers?
- How would you test your ideas out in the real world? What might be difficult about this?
- Why might these ideas not work? Or, which particular situations might cause them to fail?

Your names	Pair discussion comments: ideas picked, how to test them, and limitations

Share: Group discussion
⌛ 8:00 Discussion with the whole cohort. Facilitators can start by asking 1-2 students to summarise the idea they're most excited about, and invite other students to comment on them.

Group discussion notes

[0:50-0:55] Break

⌛ 5:00 Go for a quick walk, drink some water, rest your eyes from the screen!

[0:55-1:35] Evaluating other approaches

Learn about other scalable oversight approaches explored by your peers, and practice critically evaluating them. You should have explored an approach in the pre-session exercise 'Evaluate an approach'.

Facilitator explanation
⌛ 2:00 The facilitator should explain the general 'peer explanation' structure to the group:

- Everyone will split into pairs, with one person designated the *teacher* and the other the *learner*
- The *teacher* explains a case study to the *learner*, who asks clarifying questions as needed.
- After 7 minutes, the facilitator will send a message to swap teacher and learner (in the **same breakout room**).
- After another 7 minutes, return to the group and learners will explain their teacher's case studies.

To coordinate breakouts, participants should write down the approach they researched in the table below

Participant name	Name of scalable oversight approach researched

1-1 teaching

⌛ 7:00 Move to a 1-1 breakout room with someone who studied a different approach than you.

- The *teacher* is the person whose name is earlier alphabetically. The *teacher* should explain their approach to the *learner*. By the end of this discussion, the learner should be able to explain from memory **what the approach is**, **how it helps** human feedback more effectively align AI systems, why it **might not work**, and how well it helps with alignment in the **best case**.
- Learners: ask questions about the teacher's approach until you thoroughly understand it.
- Teachers: answer questions as best as you can. To boost your learning, don't look at your notes.
- Continue doing this until you get a message from the facilitator to swap over. **Facilitators:** when the timer runs out, don't close the rooms! Instead, send a broadcast message to all breakout rooms telling them to swap roles but stay in the same breakout room.

Teacher name	Learner name	Name of approach discussed

1-1 teaching

⌛ 7:00 In the same 1-1 breakout room.

- Swap roles, so if you were the learner before you're now the teacher. Again, by the end of this discussion, the learner should be able to explain from memory **what the approach is**, **how it helps** human feedback more effectively align AI systems, why it **might not work**, and how well it helps with alignment in the **best case**.

Teacher name	Learner name	Name of approach discussed

Group teaching

⌛ 20:00 Return to the group. Repeat the following for 2-4 learners (flexible, depending on time):

- The facilitator picks a learner, either a volunteer or randomly.
- This learner explains the approach they just learnt about, including **what the approach is**, **how it helps** human feedback more effectively align AI systems, why it **might not work**, and how well it helps with alignment in the **best case**.
- The corresponding teacher should state whether this is an accurate summary, and optionally add any extra relevant detail.
- Have a brief open group discussion about the approach.

Group teaching notes

[1:35-1:50] Wrap-up

> "We do not learn from experience. We learn from reflecting on experience."

Individual reflection

⌛ 5:00 Individually write out your takeaways from the session, any actions you intend to take as a result of the session, and any feedback you have for the course organisers and session facilitator. Consider:

- Takeaways
  - What was most useful to you from this week's discussion, and why?
  - What's something you found particularly interesting?
- Feedback
  - What's something you enjoyed or appreciated from this week's discussion?
  - What didn't go so well from this session for you, and what might we do to improve?

Name	Takeaways	Feedback

Group reflection

⌛ 5:00 Final group discussion to reflect on the session, discuss the most significant takeaways, and share the plans for the next session. If you're a smaller group, consider ending the session with every person briefly sharing their main takeaway with the rest of the cohort.

Group reflection notes

—END OF SESSION—

5: Mechanistic interpretability

We previously looked at RLHF, and improving this with scalable oversight techniques. This still didn't give us confidence we'd be sure that we could understand whether models are truly aligned or are sycophantic or deceptive. This session, we'll dive into mechanistic interpretability: an attempt to understand an AI model's reasoning by understanding its internals, which might give us more confidence the model is functioning as intended.

Session overview

- [0:00-0:20] Open questions
- [0:20-0:50] Understanding circuits
- [0:50-0:55] Break
- [0:55-1:35] Understanding superposition
- [1:35-1:50] Wrap-up

[0:00-0:20] Open questions from the resources

What are your biggest questions coming into this session, and how do they relate to what we've learnt thus far?

Individual reflections from the resources

⌛ 5:00 Write down the open questions, takeaways or uncertainties you have from the resources in the box below. Prompts you might want to consider:

- What is mechanistic interpretability trying to achieve? How is it different from just looking at inputs and outputs?
- Why is enhancing the interpretability of large AI models an important goal? What risks could arise from uninterpretable models?
- What does it mean for an AI system to be "interpretable"? What are examples of more versus less interpretable systems?

Name	Questions or comments from the resources

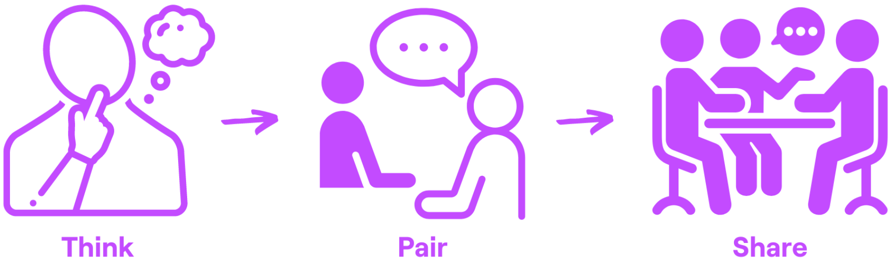
Group discussion on key themes

⌛ 3:00 Read each other’s big questions. Consider adding comments!

⌛ 12 :00 Group discussion on the key themes or biggest questions raised. Note that these might be addressed in the next exercise.

[0:20-0:50] Understanding circuits

Many of this week’s resources introduce new and complex technical concepts. In this activity, you’ll go through as a group and strengthen each other’s understanding of the core concepts from the circuits paper.



Think: Individual thoughts

⌛ 5:00 Without looking at the resources, summarise what you remember about circuits. Consider:

- The three claims about neural networks
- An example in simple language of a feature and circuit

Name	Individual summary

Pair: 1-1 discussion

⌛ 9:00 Move into a 1-1 breakout room. Answer the following questions together:

- What are the three claims about neural networks made in the circuits paper?
- What's the difference between a dog head detector feature and a circuit involving a dog head detector feature? How do these relate?
- What does 'features are represented by directions' mean? (from circuits, or superposition paper)
- How might you find a 'potted plant' feature within an image model? What circuits might involve this feature?
- (bonus, if time) What circuits might you expect to see in a large language model? Why might this be harder or easier to visualise than an image model?

Facilitator: This can be a particularly technically tricky area that confuses participants. Drop into different breakouts to help out (putting people in a three if necessary to enable this).

Your names	Pair discussion comments

Share: Group discussion

⌛ 12 :00 Discussion with the whole cohort to go through the questions. Facilitators should go around participants and have them answer parts of the question, or expand on other participant's answers.

Group discussion notes

Activity takeaways

⌛ 2:00 Everyone writes down one takeaway from the activity. For example, what's something that you now understand better, or want to research in more depth?

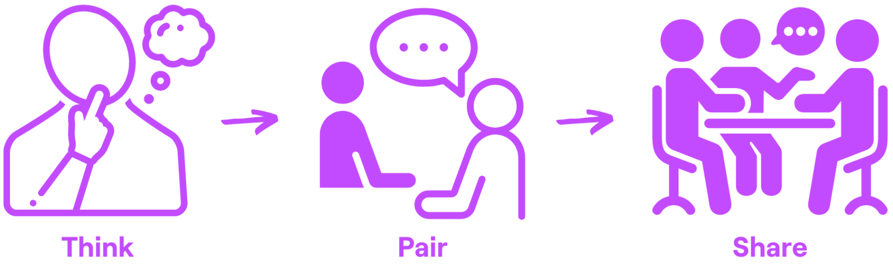
Name	Takeaway from the activity

[0:50-0:55] Break

⧖ 5:00 Go for a quick walk, drink some water, rest your eyes from the screen!

[0:55-1:35] Understanding superposition

Again, you'll go through as a group and strengthen each other's understanding of the core concepts - this time from the superposition papers.



Think: Individual thoughts

⧖ 3:00 Without looking at the resources, summarise what you remember about superposition.

Name	Individual summary

Pair: 1-1 discussion

⌛ 9:00 Move into a 1-1 breakout room. Answer the following questions together:

- What is superposition? Give an example pair of large language model features that might be likely to be stored in superposition, and why.
- What's a polysemantic neuron? How does this relate to superposition?
- What is the general idea behind dictionary learning? Why might it be helpful?

Facilitator: This can be a particularly technically tricky area that confuses participants. Drop into different breakouts to help out (putting people in a three if necessary to enable this).

Your names	Pair discussion comments

Share: Group discussion

⌛ 12 :00 Discussion with the whole cohort to go through the questions. Facilitators should go around participants and have them answer parts of the question, or expand on other participant's answers.

Group discussion notes

Individual exploration

⌛ 7:00 Spend a few minutes browsing random features found by dictionary learning run A/1 in Anthropic's paper. (Here's a guide to the interface for browsing features - we recommend focusing on the 'top activations' first, and not using the search) Try to identify a particularly interesting feature, or where you think Anthropic's interpretation seems wrong or incomplete. Note down the feature number (at the top left of the card, starting #).

Name	Individual summary

Share: Group discussion

⌛ 7:00 Discussion with the whole cohort. Facilitators can share screen and use the 'Jump to feature number' button to show features participants thought were particularly interesting.

Group discussion notes

[1:35-1:50] Wrap-up

> "We do not learn from experience. We learn from reflecting on experience."

Individual reflection

⌛ 5:00 Individually write out your takeaways from the session, any actions you intend to take as a result of the session, and any feedback you have for the course organisers and session facilitator. Consider:

- Takeaways
  - What was most useful to you from this week's discussion, and why?
  - What's something you found particularly interesting?
- Feedback
  - What's something you enjoyed or appreciated from this week's discussion?
  - What didn't go so well from this session for you, and what might we do to improve?

Name	Takeaways	Feedback

Group reflection

⏳ 5:00 Final group discussion to reflect on the session, discuss the most significant takeaways, and share the plans for the next session. If you're a smaller group, consider ending the session with every person briefly sharing their main takeaway with the rest of the cohort.

Group reflection notes

—END OF SESSION—

6: Technical governance approaches

This session we'll explore technical governance approaches that might deter harmful development or use of AI systems.

Session overview

- [0:00-0:15] Open questions
- [0:15-1:00] Evaluating policy ideas
- [1:00-1:05] Break
- [1:05-1:35] Intervention investigation interchange
- [1:35-1:50] Wrap-up

[0:00-0:15] Open questions from the resources

What are your biggest questions coming into this session, and how do they relate to what we've learnt thus far?

Individual reflections from the resources

⌛ 5:00 Write down the open questions, takeaways or uncertainties you have from the resources in the box below. Prompts you might want to consider:

- Which approaches seem particularly promising? Which ones seem less so?
- Most current technical approaches focus on individual companies reducing the risks of their deployments. Why might these alone be inadequate for reducing overall AI risks?
- Some AI control approaches have safety/usefulness trade-offs. What might push organisations deploying AI systems towards either factor?

Name	Questions or comments from the resources

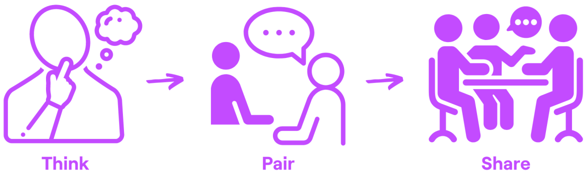
Group discussion on key themes

⌛ 2:00 Read each other's big questions. Consider adding comments!

⌛ 8:00 Group discussion on the key themes or biggest questions raised.

[0:15-1:00] Evaluating policy ideas

Evaluate whether you think different policy proposals will net increase or decrease AI risks.



Think: Individual voting

⌛ 5:00 For each proposal, vote on whether you think "**this will overall reduce risks from AI systems over the next 15 years**". You don't need lots of evidence; we encourage you to vote based on your gut response. Use the scrollbar at the bottom of the table if you can't see all of it. **Facilitator:** populate the optional extra columns with proposal ideas from the group, if any.

Name	Enforce minimum information security standards for US AI labs	Ban publicising weights for models as capable as GPT-4	Immigration rules to encourage ML engineers to move from Russia to the US	Ban China from accessing NVIDIA's best AI chips (example)	AI treaty for countries to track and publish details of their AI chips	Ban US private entities from training models 100x larger than GPT-4	(Optional) Extra 1	(Optional) Extra 2
	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral
	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral
	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral
	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral
	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral
	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral
	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral
	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral

Pair: 1-1 discussion

⌛ 10 :00 Move to a 1-1 breakout room. Identify the differences in how you voted, and explore what underlying beliefs led to this difference. We suggest considering the following discussion prompts:

- What risks might be mitigated or exacerbated by the intervention, and why?
- For interventions you both think will improve safety, how could they backfire?
- (if time) How might you implement these proposals? What technical work would be involved?

Your names	Comments from pair discussion

Share: Group discussion

⌛ 25:00 Have a group discussion. To start, 1-2 participants could summarise their pair discussions. Use these summaries to spur further discussion (optional guidance for facilitators). Cohorts with 6 or more participants should consider running another 10 minute 1-1, and shortening the group discussion.

Group discussion notes

Activity takeaways

⌛ 3:00 Conclusion: Everyone writes down one takeaway from the activity. Prompts to consider:

- What was an argument for or against a policy that you hadn't thought of before?
- Where have you changed your opinion? Or are you more confident in your beliefs somewhere?

Name	Takeaway from the activity

[1:00-1:05] Break

⌛ 5:00 Go for a quick walk, drink some water, rest your eyes from the screen!

[1:05-1:35] Intervention investigation interchange

Dive into a specific AI governance intervention explored by your peers, and understand where the field is at with it. You should have explored an intervention in the pre-session exercise 'Investigate an intervention'.

Facilitator explanation

⌛ 2:00 The facilitator should explain the general 'peer explanation' structure to the group:

- Everyone will split into pairs, with one person designated the *teacher* and the other the *learner*
- The *teacher* explains an intervention to the *learner*, who asks clarifying questions as needed.
- After 6 minutes, the facilitator will send a message to swap teacher and learner (in the **same breakout room**).
- After another 6 minutes, return to the group and learners will explain their teacher's interventions.

To coordinate breakouts, participants should write down their interventions in the table below

Participant name	Name of intervention researched

1-1 teaching

⌛ 6:00 Move to a 1-1 breakout room with someone who studied a different intervention than you.

- The *teacher* is the person whose name is earlier alphabetically. The *teacher* should explain their intervention to the *learner*. By the end, the learner should be able to explain from memory **what the intervention is**, **how it reduces AI risks** and **what work has been done so far**.
- Learners: ask questions about the teacher's intervention until you thoroughly understand it.
- Teachers: answer questions as best as you can. Don't look at your notes to boost your learning.
- Continue doing this until you get a message from the facilitator to swap over. **Facilitators:** when the timer runs out, don't close the rooms! Instead, send a broadcast message to all breakout rooms telling them to swap teacher but stay in the same breakout room.

Teacher name	Learner name	Name of intervention discussed

1-1 teaching (same breakout room)

⌛ 6:00 In the same 1-1, swap roles, so if you were the learner before, you're now the teacher. Again, by the end, the learner should be able to explain from memory **what the intervention is**, **how it reduces AI risks** and **what work has been done so far**.

Teacher name	Learner name	Name of intervention discussed

Group teaching

⌛ 16 :00 Return to the group. Repeat the following for 2-3 learners (flexible, depending on time):

- The facilitator picks a learner, either a volunteer or randomly.
- This learner explains the intervention they just learnt about, including **what the intervention is**, **how it reduces AI risks** and **what work has been done so far**.

- The corresponding teacher should state whether this is an accurate summary, and optionally add any extra relevant detail.
- Have a brief open group discussion to discuss the merits and limitations of the intervention.

Group teaching notes

[1:35-1:50] Wrap-up

“We do not learn from experience. We learn from reflecting on experience.”

Individual reflection

⌛ 5:00 Individually write out your takeaways from the session, any actions you intend to take as a result of the session, and any feedback you have for the course organisers and session facilitator. Consider:

- Takeaways
  - What was most useful to you from this week’s discussion, and why?
  - What’s something you found particularly interesting?
- Feedback
  - What’s something you enjoyed or appreciated from this week’s discussion?
  - What didn’t go so well from this session for you, and what might we do to improve?

Name	Takeaways	Feedback

Group reflection

⏳ 5:00 Final group discussion to reflect on the session, discuss the most significant takeaways, and share the plans for the next session. If you're a smaller group, consider ending the session with every person briefly sharing their main takeaway with the rest of the cohort.

Group reflection notes

—END OF SESSION—

7: Contributing to AI Alignment

In this final session of the learning phase, we'll support each other to figure out next steps and gear up for the project sprint.

Session overview

- [0:00-0:15] Open questions and project sprint orientation
- [0:15-0:55] Reviewing project plans
- [0:55-1:00] Break
- [1:00-1:30] Reviewing career plans
- [1:30-1:50] Closing

[0:00-0:15] Open questions and project sprint orientation

What are your biggest uncertainties about the topics covered in the resources, or the upcoming weeks in the project sprint?

Project sprint overview

- See How to do an excellent AI alignment project > 'What to expect'

Individual reflections

⏳ 5:00 Write any questions you have regarding the project sprint, or this week's resources in the box below.

Name	Questions or comments

Group discussion on key themes

⌛ 10 :00 Discuss and resolve the main logistical uncertainties that arise, and follow up in Slack with any further clarifications.

[0:15-0:55] Reviewing project plans

Help each other build better projects by giving constructive feedback on your peers' plans.

Facilitator explanation

⌛ 3:00 The facilitator should explain the general 'peer feedback' structure to the group:

- Everyone will split into pairs (people stay in the **same breakout room** throughout)
- Share your project plan with your partner, e.g. as a Google Doc or by pasting it here if it's short
- Spend 7 minutes in silence reading, taking notes, and adding comments to your partner's plan
- Then, spend 7 minutes discussing one of your plans: what you like about it and how to improve it
- Then swap, and spend 7 minutes discussing the other plan
- We'll then return to the group to discuss general patterns for improving project plans

To coordinate breakouts, participants should write down their project summary. Facilitators should try to pair similar people up.

Participant name	1-line project summary

Breakouts

- ⌛ 7:00 Share your plan, then silently read, take notes and comment on your partner's plan
- ⌛ 7:00 Discuss participant 1's plan: what you like about it, and how to improve it
- ⌛ 7:00 Discuss participant 2's plan: what you like about it, and how to improve it
- In the table below you can note down general tips, or other relevant ideas you have for others in the cohort to improve their project plans

Participant 1	Participant 2	General notes

Share: Group discussion

⌛ 10 :00 Discussion with the whole cohort. Facilitators can start by asking 1-2 students to summarise feedback they found most useful.

Group discussion notes

[0:55-1:00] Break

⌛ 5:00 Go for a quick walk, drink some water, rest your eyes from the screen!

[1:00-1:30] Reviewing career plans

Help each other plan your careers by giving constructive feedback on your peers' plans.

Facilitator explanation

⌛ 2:00 The facilitator should explain the general 'peer feedback' structure to the group:

- Everyone will split into pairs (people stay in the **same breakout room** throughout)
- Share your career plan with your partner, e.g. as a Google Doc or by pasting it here if it's short
- Spend 4 minutes in silence reading, taking notes, and adding comments to your partner's plan
- Then, spend 6 minutes discussing one of your career plans: what excites you about the career plan, challenges they might face, other employers / areas of research they could consider
- Then swap, and spend 6 minutes discussing the other plan
- We'll then return to the group to discuss general patterns for improving career plans

To coordinate breakouts, participants should write down their type of career plan. Facilitators should try to pair similar people up.

Participant name	Primary type of plan over the next 6 months *(e.g. applying for jobs, further study, applying learnings in current job)*

Breakouts

- ⌛ 4:00 Share your plan, then silently read, take notes and comment on your partner's plan
- ⌛ 6:00 Discuss participant 1's plan
- ⌛ 6:00 Discuss participant 2's plan
- In the table below you can note down general tips, or other relevant ideas you have for others in the cohort to improve their career plans

Participant 1	Participant 2	General notes

Share: Group discussion
⌛ 10 :00 Discussion with the whole cohort. Facilitators can start by asking 1-2 students to summarise feedback they found most useful.

Group discussion notes

[1:30-1:50] Closing

We've covered a lot of ground in the last 8 weeks! At this pace, it's easy to overlook acknowledging the ways in which you've helped each other out, and the positive influence you've had on your cohort. Let's take a moment to reflect on these contributions and the course's overall experience one final time.

Gratitude
⌛ 7:00 Express your appreciation to your fellow cohort members for the weeks you've shared together.

- First, write your name in the table below
- Then in at least two **other** people's rows state something you're grateful to them for. For example:
  - They helped you understand concepts better, or challenged your ideas constructively
  - You had a particularly good breakout session with them in any week
  - They shared insightful resources or contributed to discussions in your Slack channel
- Tips:
  - Be specific: highlight precise instances or actions that you found helpful for your learning
  - Explain personal impact: describe how it helped you, or contributed to your success

Facilitator name	Comments (from participants)

Participant name	Comments (from other participants and facilitator)

Group reflection

⌛ 7:00 One final group discussion to reflect on the course, and discuss your biggest takeaways.

Group reflection notes

Next steps

Between sessions 7 and 8 there is an extra week break as we'll put people into project cohorts. You should spend this time validating your idea (see session 8 resources for more details).

—END OF LEARNING PHASE—

8: Rapidly testing your project

Welcome to the AI alignment project sprint! This session you'll meet your project cohort, as well as discuss your project 1:1 with your facilitator.

Session overview

- [0:00-0:15] Group introductions
- [0:15-0:50] Speed 1-1s
- [0:50-0:55] Orienting towards action

[0:00-0:15] Introductions

Say hello to your project cohort!

Write an intro for yourself once you join the session
⌛ 4:00 To help everyone get to know each other, fill in your name and share some info about yourself in the table below while you're waiting for others to join the call. Then read everyone else's responses, and feel free to leave comments!

Name and pronouns	What's your current project idea, briefly?	How could this cohort best support you?	What's something fun that you've experienced in the past month?

[Skip if 7 or more people] Group introductions
⌛ 7:00 Everyone introduces themselves for ~1 minute to the whole group, starting with the facilitator.

[0:15-0:50] Speed 1-1s

Explore something new with each person! Dive deep into personal stories.

In this activity, you'll each spend some time with the facilitator 1:1 to talk through your project, the results of your rapid tests, and your next steps.

While not speaking with the facilitator, you'll network with other participants. You can discuss your projects, or use this wheel of prompts to have casual conversations.

Breakouts

- **Cohorts with 1-2 participants:** Stay as one group.
- **Cohorts with 3-4 participants:** Split into two breakouts, one with the facilitator and a participant, and one with everyone else.
- **Cohorts with 5+ participants:** Split into 2-person breakout rooms (with one 3 if you're an odd number of people). This simple spreadsheet can be used to help allocate the rooms, to avoid repeat matches.

Instructions

- ⌛ 8:00 Spend about 7-10 minutes per participant, depending on the number of participants. If the cohort is particularly large, adjust the timings for the session so that everyone is able to have a reasonable conversation with the facilitator.

[0:50-0:55] Orienting towards action

Commit to what you're going to do over the next week.

Next steps

- ⌛ 4:00 Write down what you aim to have completed by the next session.
- ⌛ 1:00 Facilitator closes with any final brief thoughts or reflections.

Name	By the next session, I will have...

—END OF SESSION—

9: Developing your project

Today you'll briefly check-in on your cohort's projects, and help unblock each other.

Session overview

- [0:00-0:05] Project updates
- [0:05-0:25] Peer feedback
- [0:25-0:30] Orienting towards action
- [0:30-0:35] (Optional) Improving project sessions

[0:00-0:05] Project updates

Group check-in on how their projects are going.

Individual reflections

⌛ 3:00 Explain whether you achieved your goal from last week, and describe any other key project updates. Flag any blockers or questions you have.

Name	Project update

Group discussion on key themes

⌛ 2:00 Discuss and resolve any shared logistical uncertainties. Follow up in Slack with any further clarifications.

[0:05-0:25] Peer feedback

Help each other build better projects by giving constructive feedback to each other.

Facilitator explanation

⌛ 2:00 The facilitator should explain the general 'peer feedback' structure to the group:

- Everyone will split into pairs (people stay in the **same breakout room** throughout)
- Then, spend 8 minutes discussing one of your projects
- Then swap, and spend 8 minutes discussing the other person's project

Breakouts

- ⌛ 8:00 Discuss participant 1's project: what the next steps are, where they are stuck
- ⌛ 8:00 Discuss participant 2's project: what the next steps are, where they are stuck
- In the table below you can note down general tips, or other relevant ideas you have for others in the cohort to improve their project plans

Participant 1	Participant 2	General notes

[0:25-0:30] Orienting towards action

Commit to what you're going to do over the next week.

Next steps

- ⌛ 4:00 Write down what you aim to have completed by the next session.
- ⌛ 1:00 Facilitator closes with any final brief thoughts or reflections.

Name	By the next session, I will have...

—END OF SESSION—

10: Further developing your project

Another brief check-in on your cohort's projects.

Session overview

- [0:00-0:05] Project updates
- [0:05-0:25] Peer feedback
- [0:25-0:30] Orienting towards action

[0:00-0:05] Project updates

Group check-in on how their projects are going.

Individual reflections

⏳ 3:00 Explain whether you achieved your goal from last week, and describe any other key project updates. Flag any blockers or questions you have.

Name	Project update

Group discussion on key themes

⌛ 2:00 Discuss and resolve any shared logistical uncertainties. Follow up in Slack with any further clarifications.

[0:05-0:25] Peer feedback

Help each other build better projects by giving constructive feedback to each other.

Facilitator explanation

⌛ 2:00 The facilitator should explain the general 'peer feedback' structure to the group:

- Everyone will split into pairs (people stay in the **same breakout room** throughout)
- Then, spend 8 minutes discussing one of your projects
- Then swap, and spend 8 minutes discussing the other person's project

Breakouts

- ⌛ 8:00 Discuss participant 1's project: what the next steps are, where they are stuck
- ⌛ 8:00 Discuss participant 2's project: what the next steps are, where they are stuck
- In the table below you can note down general tips, or other relevant ideas you have for others in the cohort to improve their project plans

Participant 1	Participant 2	General notes

[0:25-0:30] Orienting towards action

Commit to what you're going to do over the next week.

Next steps

- ⌛ 4:00 Write down what you aim to have completed by the next session.
- ⌛ 1:00 Facilitator closes with any final brief thoughts or reflections.

Name	By the next session, I will have...

—END OF SESSION—

11: Building in public

Get feedback on your public product from your cohort peers

Session overview

- [0:00-0:05] Project updates
- [0:05-0:40] Peer feedback
- [0:40-0:55] Closing

[0:00-0:05] Project updates

Group check-in on how their projects are going.

Individual reflections

⌛ 3:00 Explain whether you achieved your goal from last week, and describe any other key project updates. Flag any blockers or questions you have.

Name	Project update

Group discussion on key themes

⌛ 2:00 Discuss and resolve any shared logistical uncertainties. Follow up in Slack with any further clarifications.

[0:05-0:40] Peer feedback

Help each other build better public products by giving each other constructive feedback.

Facilitator explanation

⌛ 2:00 The facilitator should explain the general 'peer feedback' structure to the group:

- Everyone will split into pairs (people stay in the **same breakout room** throughout)
- First, spend 15 minutes reviewing each other's public products. If it takes longer than 15 minutes to digest, direct your peer to the part you most want their feedback on.
- Then, spend 8 minutes discussing one of your projects
- Then swap, and spend 8 minutes discussing the other person's project
- We'll then return to the group to discuss general patterns for improving project write-ups

Breakouts

- ⌛ 15 :00 Review each other's public products
- ⌛ 8:00 Discuss participant 1's project: where did you get confused, how could it be improved
- ⌛ 8:00 Discuss participant 2's project: where did you get confused, how could it be improved

- In the table below you can note down general tips, or other relevant ideas you have for others in the cohort to improve their project plans

Participant 1	Participant 2	General notes

[0:40-0:55] Closing

Reflect on what you've achieved over the last few weeks and identify next steps.

Individual reflection

- ⌛ 4:00 Write down an achievement you're proud of from the project sprint. Remember that even if you didn't achieve your original project aim, you've likely still learnt a new skill or have produced a write-up as to why it was difficult.

Name	I am proud that I…

Next steps

- ⌛ 4:00 Write down what you aim to have completed by the closing event. For example, continuing your project, preparing a project presentation, applying to jobs or further study.

Name	By the course closing event, I will have…

Group reflection

⌛ 5:00 One final group discussion to reflect on the project sprint, and discuss your biggest takeaways.

Group reflection notes

—END OF PROJECT SESSIONS—