You can share this document with AI safety university group organizers. Please don't share it with anyone else without Jakub's permission — reach out at jakraus@umich.edu.

Notes

- Meetings are **2 hours long** and take place in person.
- Some weeks have time estimates. These are approximate estimates.
- I didn't have time to add details for the later weeks.

Tips

- Someone asks you a question → consider if you can ask someone else in the group to answer the question.
- Try framing a question with "do you agree" to make it easy to start a discussion.
- When printing readings, consider pasting the text (or screenshots of the text) into a Google doc to remove any weird website formatting issues. Another option is to use ar5iv.
- Improvise based on how the session is going. If people seem engaged with a particular exercise or presentation, consider extending this portion and skipping a different part of the week's content.
- Bring some snacks. E.g. Larabars, Oreos, fruit snacks.
- Bring some pencils and/or highlighters for people who want to take notes as they read
- People less fluent in English might prefer to translate the readings using their laptop.
- In the first meeting be sure to mention "I expect everyone to attend every session unless something comes up, and if something comes up I ask that you please let me know in advance so I can prepare for the session with the correct number of participants in mind"
- If the group is too large, people might feel less obligated to attend due to some kind of diffusion of responsibility.
- You can end a session slightly earlier or slightly later depending on the atmosphere. E.g. if people are having a fascinating discussion and everyone seems engaged, don't worry about ending right on time. If the people are running out of ideas, feel free to end 5-10 minutes early and nobody will complain – just explain "seems like we covered most of the topics for this week, so we can end a bit early."
- This doc has more icebreaker ideas.

Week 0

1. **Introductions → icebreakers**: *What is something that's been going well for you this semester or something that you've been enjoying? Or if you want, what is something that you've disliked about this semester or something that has been difficult so far?*
    a. Split into groups of 2-3 to talk.
2. Read: A short introduction to machine learning (Ngo, 2021).
    a. Optional reading for people who finish early: Part 1: Key Concepts in RL (Spinning Up in Deep RL).

You can share this document with AI safety university group organizers. Please don't share it with anyone else without Jakub's permission — reach out at jakraus@umich.edu.

3. Split and discuss. Mention: active learning is important, there are no dumb questions, don't worry about saying something wrong (that's how you learn), acknowledge uncertainties.
    a. What is AI?
    b. What is machine learning?
    c. What is deep learning?
    d. What is reinforcement learning?
    e. What are some goals for the field of AI?
    f. What are some limitations of AI?
4. Watch: But what is a neural network? | Chapter 1, Deep learning (00:00 – 16:25).
    a. Ask people if they've watched the video already. It might still be valuable to watch again, but offer Karpathy's micrograd tutorial if they'd like to skip it. They can leave the room if they prefer, as long as they come back in 45 minutes.
5. Watch: Gradient descent, how neural networks learn | Chapter 2, Deep learning (0:00 – 16:35).
6. Watch: What is backpropagation really doing? | Chapter 3, Deep learning (0:00 – 12:30).
7. Split (ask people to find new people, different from the people they were talking to earlier) and discuss. One person attempts to explain the following concepts in as much detail as possible. The other person asks questions either for their own uncertainty or to poke holes in the other person's explanation and help them strengthen their understanding.
    a. What is a neural network?
    b. What is gradient descent?
    c. What is backpropagation?
8. Present slides explaining supervised vs unsupervised vs self-supervised learning, training vs test data (and the terms "training time" and "test time"), ordinary least squares regression as finding parameters that minimize the loss/cost function (gradient descent is another method), overfitting + regularization + hyperparameter (and validation data).
    a. All based on Machine Learning for Humans, Part 2.1: Supervised Learning.
    b. Encourage people to ask questions during the presentation if they're confused.
9. Full group discussion. Collectively come up with as many relevant steps as possible for training an MLP to classify MNIST digits. Write these steps on a whiteboard and encourage people to take a picture at the end.
    a. Refer to Nielsen's textbook if you need help.
10. Wrap up. People are free to take or leave the readings. Meet again at the same time next week. "I'll stick around if anyone wants to chat more or ask any questions." Have a nice week!

You can share this document with AI safety university group organizers. Please don't share it with anyone else without Jakub's permission — reach out at jakraus@umich.edu.

Week 1

1. **Introductions** (there are some new people) → **icebreakers**: *What are some of your favorite classes you've taken in high school or college or online? And if you had to drop a class right now, what would it be?*
    a. Split into groups of 2-3 to talk.
2. Read: Biological Anchors: A Trick That Might Or Might Not Work (use this Google doc for printing).
    a. Optional reading for people who finish early: AI and Compute How Much Longer Can Computing Power Drive Artificial Intelligence Progress?
3. Split and discuss.
    a. What are some limitations of this method?
    b. How would you allocate weights across the 6 anchors?
    c. How has this piece changed your thinking on AGI timelines?
4. Presentation: AI capabilities.
    a. Pick from any of these: Minerva solves high school math competition problems, AlphaTensor finds faster algorithms for matrix multiplication, GitHub CoPilot opens to the public for computer programming, DALL-E 2 and Stable Diffusion generate AI art, Flamingo succeeds at many vision and language tasks with only a few examples, Gato performs well at many different unrelated tasks (stacking blocks, playing Atari, captioning images, etc), AlphaCode solves high school programming competition problems, Meta's neural theorem prover solves 10 International Math Olympiad problems, Imagen Video and Meta Make-A-Video advance towards deepfakes, Chinchilla demonstrates even more effective methods for leveraging computation in deep learning, PaLM-SayCan lets robots perform complex tasks, Whisper and VALL-E approach human-level speech-to-text, AudioLM generates realistic speech and piano music, "DreamerV3 is the first algorithm to collect diamonds in Minecraft from scratch without human data or curricula," 3DALL-E integrates DALL-E into 3D CAD design software, Cicero plays Diplomacy at a human level, ACT-1 and WebGPT can interact with websites, Flan-PaLM and Med-PaLM approach human-level medical reasoning, MusicLM generates music from text.
    b. https://thispersondoesnotexist.com/
    c. AlphaGo - The Movie | Full award-winning documentary 40:00 through 40:50
    d. Also https://podcast.ai/
    e. Visualizing the deep learning revolution (Ngo, 2023)
    f. Creating a Space Game with OpenAI Codex

Week 2

1. Icebreaker: imagine a world without color and a world without music, which would you rather live in? What are your top 3 favorite foods, in no particular order? If job prospects didn't matter, what major would you choose?

You can share this document with AI safety university group organizers. Please don't share it with anyone else without Jakub's permission — reach out at jakraus@umich.edu.

2. Read Four Background Claims
3. Split and discuss, then back to a group discussion
    a. Evaluate each claims, do you agree with each claim, what are the best counterarguments to each claim
4. Read the alignment problem from a DL perspective (section 2 only), and What Failure Looks Like (part 1 only)
5. Presentation on foundation models: PaLM, PaLM-SayCan from here, Minerva examples, Kaj Sotala GPT-3 thread, OpenAI Codex, ChatGPT, Stable-Dreamfusion, SD Pokemon fine-tuning, Stable Diffusion interpolation videos, Riffusion
    a. Also transition from ChatGPT into RLHF → learning to summarize with human feedback and specification gaming. Mention this picture from Leo Gao
    b. Also mention Goodhart's law with respect to CoastRunners
    c. Originally we read pages 3 to 6 of On the opportunities and risks of foundation models, but it was packed with shallow investigations of many topics, which overwhelmed people with too much information.
6. Watch Neural Scaling Laws and GPT-3 from 16:45 to ~22:00
7. Bonus reading for people who finish early: Future ML Systems Will Be Qualitatively Different. You can also print the rest of What Failure Looks Like and let people read part 2 if they're interested.

Week 3

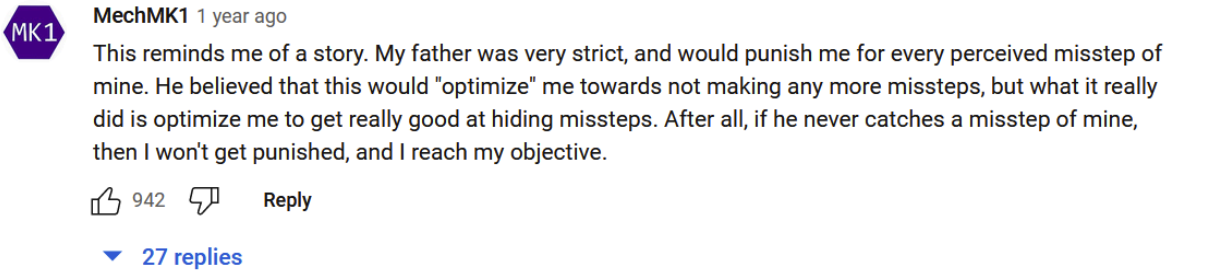
What I printed: Week 3 readings MAISI Alignment Fundamentals Wi23

1. Icebreaker: What's a topic that you love that other people often find boring or less interesting? Is there something you wish other people were more excited about? [10 minutes]
2. Read Why alignment could be hard with modern deep learning (Cotra, 2021) [20 minutes].
    a. People who finish early can get started on the goal misgeneralization paper.
3. Split into small groups, then regroup and discuss. [10 minutes]
    a. Each person tries to explain Ajeya's hiring analogy and how it connects to deep learning (refer back to the paper or ask your group for help when you're stuck or confused). When listening, ask questions to test the explainers understanding or clarify your own confusions.
4. Presentation on Goal Misgeneralisation: Why Correct Specifications Aren't Enough For Correct Goals (Shah, 2022): **blog post** [15 minutes]
    a. Before people go into it, walk through the opening example with a projector since it uses GIFs. You're a blue blob moving in this 3D world (towers, colored spheres, flashing square), there's another red blob traveling around. Reward +3 on first trajectory while following red agent. Reward -2 when moving independently. Training continues with cases similar to this, then the learned policy at test time is to follow red, which actually performs poorly unlike during training. Here's what happened: you must touch spheres in a particular order (although you can start

You can share this document with AI safety university group organizers. Please don't share it with anyone else without Jakub's permission — reach out at jakraus@umich.edu.

anywhere in that order so your start is always correct) – correct spheres are +1 and incorrect spheres are -1; flashing white square is +1 and black square is -1. Red bot is expert during first 2 videos and anti-expert in test video.
    i. This behavior seems pretty stupid: the agent observes the negative rewards and keeps going anyways.
b. Also take this presentation time to briefly describe and display the example videos from the full paper: https://sites.google.com/view/goal-misgeneralization.
c. And be sure people have a basic understanding of the concept. "I want to emphasize that it's fine and even expected to be confused about this concept since a lot of people find it tricky, and because of that I want to make sure everyone has the chance to ask questions, so before moving on to the full paper we're going to go around one by one and try to share one question or comment about something you're confused about or want more information about, totally fine if you'd rather not share, you can just say 'pass' if that's the case. I'll start with [person name] and move clockwise. First, let's take a minute to think… it might be helpful to imagine explaining this concept to someone else and see where you get stuck." Then call on each person.

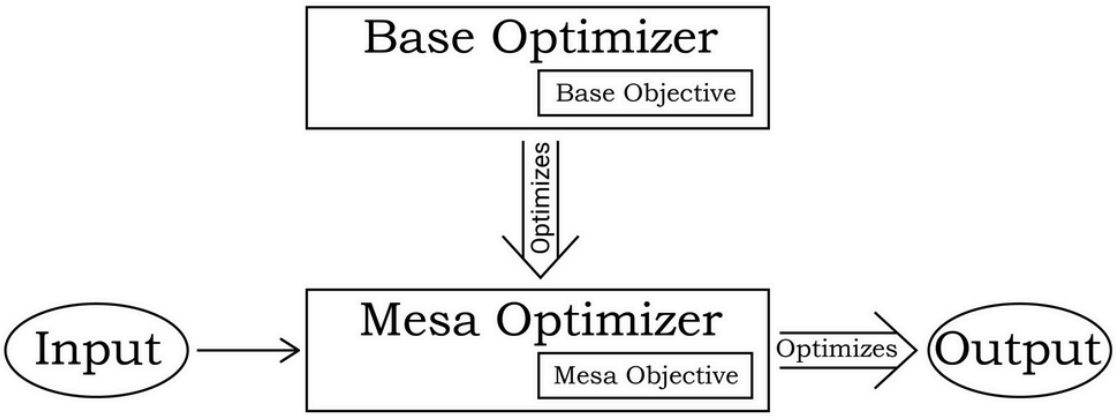
5. Then move on to the paper (skip the section 3 examples besides 3.1 (monster gridworld), also skip section 5 due to the ensuing exercise). [35 minutes]
    a. With extra time read Categorizing failures as "outer" or "inner" misalignment is often confused (Rohin Shah, 2023)
6. Quick call for any questions, then split into small groups, then regroup and discuss [15 minutes]
    a. What is goal misgeneralization?
    b. Why might it arise in AI systems?
    c. [most of the time] How might we mitigate the risk of goal misgeneralization?
7. Briefly describe the instrumental convergence thesis before transitioning to The OTHER AI Alignment Problem: Mesa-Optimizers and Inner Alignment (starting at 2:27) [20 minutes]



^ the top comment (February 11th, 2023) raises a key question about the difficulty of alignment! (When we penalize models for deception, will this actually reduce deception or just make models more patient and careful?)

8. If time, brief group collaboration on defining mesaoptimization with a diagram like this one:

You can share this document with AI safety university group organizers. Please don't share it with anyone else without Jakub's permission — reach out at jakraus@umich.edu.



Our actual session followed a structure more like this:

- Why AI alignment could be hard with modern deep learning by Ajeya Cotra
    - Optional reading for people who finish early: Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals
- Discussing Ajeya's analogy and how it connects to deep learning
- Short presentation on goal misgeneralization based on this blog post, with some of these videos
- Try to generate the ideas in section 5 (how to mitigate goal misgeneralization) as a group before looking at them
- This video about mesaoptimization
- A short presentation about "inner" and "outer alignment" before attempting the exercise at the start of this post to test our understanding

Week 4

Week 3 readings MAISI Alignment Fundamentals Wi23 , Paul Christiano alignment chart

- *Icebreaker: favorite place you've traveled? Doesn't have to be far away.*
- Kick off with Paul Christiano: Current work in AI alignment video and stop before the Q&A (pass out alignment charts so people can follow along)
- Discuss based on Paul Christiano on AI alignment -- discussion
    - Anything missing from the diagram that should replace the ellipses for the high-level splits?
    - How can we ensure that AI labs pay the alignment tax?
    - Any other techniques for outer alignment that might be interesting to try?
- Then scalable oversight and IDA (can stop before section 3). With extra time read the rest of IDA.
- Discuss scalable oversight

You can share this document with AI safety university group organizers. Please don't share it with anyone else without Jakub's permission — reach out at jakraus@umich.edu.

- Then each person takes a turn trying to explain how IDA works while others ask questions for their own understanding or for adversarially improving the speaker's explanation.
    - **This is the main learning focus for the session: understand IDA.**
- With extra time: RRM
- Or maybe How Science Misunderstands Power

Ask participants to fill out a Google form with feedback on how the group has been for them so far.

Week 5

Icebreaker: PBS called. An executive wants you to make a documentary series. An elite team of researchers, screenwriters, camera specialists, audio engineers, video editors, and publicists – all at your fingertips. So what's it about?
Week 4 readings MAISI Alignment Fundamentals Wi23

- Intro to ML Safety: adversarial robustness – seemed kind of boring?
- Adversarial Robustness from https://course.mlsafety.org/
- https://www.deepmind.com/publications/red-teaming-language-models-with-language-models
- https://www.alignmentforum.org/posts/A9tJFJY7DsGTFKKkh/high-stakes-alignment-via-adversarial-training-redwood
- https://www.cold-takes.com/ai-safety-seems-hard-to-measure/

Try reading a section and then discussing any confusions (with extra time to go ahead)

AI safety seems hard to measure?

Week 6

Chris Olah and Neel Nanda videos
Present on https://distill.pub/2020/circuits/zoom-in/
https://ar5iv.labs.arxiv.org/html/1610.01644
https://ar5iv.labs.arxiv.org/html/2111.09259

Week 7

Present some slides from AI safety intro talk and discuss