Preparing for ML Safety Scholars

If you have the prerequisites, preparation for ML safety scholars is not needed. However, you might get more out of the material if you have a little bit of background on some of it. You may also want to brush up on your math skills. We've prepared a number of resources that can help you do this. Note that we won't be doing any support for these courses, but hopefully they are useful.

Machine Learning

Machine learning experience is by no means required for this course. However, some people have asked about what we recommend if you want to get a head start. You could do this by starting on the content below prior to the program. Alternatively, you could try Andrew Ng's ML course or the Udacity ML course, which are more conceptual and use less math than the course we are starting with.

Probability

The most relevant part of statistics to machine learning is probability. This Harvard edX course is of high quality. Getting through as much of this course as time allows would be helpful. This MIT open courseware also provides a good introduction. Part 1 (the fundamentals) is most useful.

Linear Algebra

If you do not have linear algebra experience and have time before the program, we highly recommend you get some background in linear algebra. Even if you do have linear algebra experience, taking linear algebra multiple times from different instructors often illuminates new concepts each time. You could try Georgia Tech's course (1,2,3,4), UT Austin's course, or the MIT course.

Multivariable Calculus

Knowing multivariable calculus will be helpful for ML Safety Scholars. Try this Khan Academy course. Specifically, it helps to know differential multivariable calculus; integral calculus is less important for our purposes.

ML Safety Scholars Curriculum

Machine Learning (Weeks One and Two)

Standard Track ML (content from MIT 6.036)

Most students should take this track, even for those who have taken ML before (especially online). It provides a more thorough mathematical treatment than you may have seen. **The assignments are interspersed with the course content, and includes all "labs", "homeworks", and "exercises".** Some of the questions are autograded: you have unlimited retries to get the answer correct (though I don't suggest guessing randomly, you won't learn anything). For the questions there is a "show answer" button which you can use if you can't figure something out. However, after you click "show answer", you've gotten that question wrong, and you can't go back and undo the fact that you already know the answer. If you solve a question correctly, you can click "show answer" without affecting your score, in case you want to see if they did it any differently.

Feel free to do the self-assessment at the beginning if you'd like; it doesn't count as an official assignment.

Syllabus & Lecture Videos (you will need to make a free account)

Italicized below means that it's something you have to turn in.

Week One: Basics (introduction to ML, linear classifiers, *exercises, homework*), Perceptrons (the perceptron, *exercises, lab, homework*), Features (feature representation, *exercises*).

Week Two: Features continued (*lab, homework*), Margin Maximization (logistic regression, gradient descent, *exercises, lab, homework*), Regression (regression, *exercises, lab, homework*).

If you are confused about the nested python functions found in problem 6.2 in the MIT "week 4" homework, check out this textbook.

Advanced Track ML (content from CMU 10-715 PhD level course)

This track is for people who already have a solid understanding of the content in the Standard Track. Content is from CMU 10-715, a PhD-level introduction to ML for research.

Warning: the content is extremely fast paced, and requires more mathematical knowledge (take a look at the assignments if you're unsure). There may also be less support available from TAs as compared to the standard track. If in doubt, take the standard track.

Syllabus
Lecture videos

Week One: Unsupervised learning (Lectures 5-7). Assignment 2 (from this course) assigned.
Week Two: Kernel machines (Lectures 8-12). Assignment 3 (from this course) assigned.

Deep Learning (Weeks Three, Four, and Five)

Syllabus & Lecture Videos. Below, when it says "assignment," it is literally just that assignment from the online course.

Week 3 - Introduction, Image Classification, Linear Classifiers, Optimization, Neural Networks. *DL assignment 1 assigned.* **Please also watch this video.**

Week 4 - Backpropagation, CNNs, CNN Architectures, Hardware and Software, Training Neural Nets I & II. *DL Assignment 2 assigned.*

For DL Assignment 2: The part of the linear classifier assignment titled "SVM classifier" is now optional. Please refer to this implementation for pieces of code that will be needed in order to complete the softmax section. The softmax section is not optional.

For assignment 2, two_layer_neural_net, please use this file!

Week 5 - RNNs, Attention, NLP (DL for NLP I & II from NYU), Hugging Face tutorial (parts 1-3), Reinforcement Learning.

Assignment: Write up notes of the lectures from this week (not including the Hugging Face tutorial). They don't need to be polished. Submit to gradescope.

(Now optional) *DL Assignment 3 assigned. For assignment 3, convolutional networks, be sure to remove ?usp=sharing from the google drive link you use.*

ML Safety (Weeks Six, Seven, and Eight)

You can view links to the lectures and videos here. You can also see the assignments, which are marked with ⌨️ (coding assignments) or 📝 (written assignments).

Week 6 - Background (Introduction, Deep Learning Review), Hazard Analysis (Risk Decomposition, Accident Models, Black Swans), Robustness (Adversarial Robustness, Black Swan Robustness). You should watch the deep learning review lecture, as it includes some content not previously covered, but you don't need to do the homework under deep learning review.

Week 7 - Monitoring (Anomaly Detection, Interpretable Uncertainty, Transparency, Trojans, Detecting Emergent Behavior), Alignment (Honest Models, Intrasystem Goals/Power Aversion, Machine Ethics).

Week 8 - Systemic Safety (e.g., ML for Improved Decision-Making, ML for Cyberdefense, Cooperative AI), Additional X-Risk Discussion (X-Risk, Possible Existential Hazards, Safety-Capabilities Balance, Review and Conclusion).

Final Project

See here: Final Project guidelines

Readings

Readings should be completed before going to the discussion section for that week (or the discussion section on the Monday of the following week).

Week One

For Discussion Section
We Need To Talk About A.I., available on Google Play Movies, Apple TV, Vudu, Amazon Prime, and YouTube (all stipends will be increased by $4 to pay for this).

Note that this movie is for a popular audience, and also that the experts contradict each other. It's a useful exercise to watch the movie and write down questions, such as "is this claim actually true"*?* At least one of the experts makes an assertion that is significantly less true now than it was in 2020 when the movie was made; can you find it? By the end of the course, you

should be able to evaluate the arguments given in the movie far better. I will tell you right now that the “Baby X” part of the movie is absolutely the worst part, doesn’t resemble any modern machine learning, and you should basically ignore it.

Discussion questions (please look at/answer beforehand and come to section ready to discuss):

- Find one claim made in the movie, and try to investigate it further. What did you find?
- What did you find most compelling about the movie?
- What did you disagree with the most in the movie?
- What point do you wish had been made?
- Which of the possible threats from AI that was mentioned seems most important to you? What might convince you to change your mind?

Paper Reading
None.

Week Two

For Discussion Section
Sutton, The Bitter Lesson (2019)
Our World In Data, Estimated Computation Used In Large AI Training Runs (2022)
Kaplan et al., Scaling Laws for Neural Language Models (2020; we don’t expect you to understand this fully yet, just to get the gist of what is going on at a high level).
Roodman, Modeling the Human Trajectory (2020)

Discussion questions:

- What is the basic premise of the Bitter Lesson? What is the best evidence for and against it?
- Do some research online. Are there tasks for which compute scaling hasn’t so far been enough to solve the problem?
- If compute scaling continues to be increasingly important, what does that imply about where future AI research will be conducted (e.g. which companies, universities, etc.)?
- If compute scaling continues to be increasingly important, what does that imply about the number of engineers that will be needed to conduct ML research?
- In the Roodman reading, consider the quote from Thomas Jefferson. Do you think it could describe AI?
- Why might it be useful to model larger trends like GWP and scaling laws?
- What is the biggest limitation of Roodman’s work, in your opinion?

Paper Reading

None.

Week Three

For Discussion Section
Woodside & Hendrycks, A Bird's Eye View of the ML Field (2022)

Discussion questions:

- Look up an ML metric from this list: [ImageNet, GSM8k, MMLU]. How have methods progressed over time?
- We write, "By some indications, reinforcement learning is poised for a tsunami." Do you believe this? Find evidence for and/or against.
- Highly-influential citations are obviously not a perfect measure of research impact. What alternative approaches could be used to measure researchers? What are some limitations with such approaches?
- Choose one of the researchers who wrote one of the papers assigned for this week. What kinds of research have they produced since 2014? What kind of influence do they have now? How much of this do you think could be attributed to the Matthew Effect?
- Which of the different methods researchers want to advance capabilities with seems most promising?
- Given what you read, brainstorm one strategy that could cause more ML researchers to do safety-relevant research. What are some pitfalls with this approach?

Paper Reading

Kingma & Ba, Adam: A Method for Stochastic Optimization (2014)
Srivastava et al., Dropout: A Simple Way to Prevent Neural Networks from Overfitting (2014)

Please summarize each paper (at least 150 words) and submit to gradescope by the deadline indicated.

Week Four

For Discussion Section
Karnofsky, "The Most Important Century" (2021)
Please read the summary (linked above), "All Possible Views About Humanity's Future are Wild," "This Can't Go On," "Forecasting Transformative AI, Part 1", "Making the Most of the Most Important Century," "Call to Vigilance"

Discussion questions:

- Which of the three views in All Possible Views About Humanity's Future Are Wild is closest to yours? The author's view, the "conservative" view, the skeptical view, or something else entirely? Why?
- Karnofsky writes, "But today, it seems like things are far out of balance, with almost all news and analysis living in the Business As Usual headspace." Do you think this is true? Is this a bad thing? How else might analysis look like?
- Do you think the analogy Karnofsky makes to a plane on a runway makes sense? Do you think it's a useful analogy?
- Karnofsky writes: "If we want to know why AlphaZero made some particular chess move, we can't look inside its code to find ideas like "Control the center of the board" or "Try not to lose my queen."" Take a look at this paper, which was released after the post. Discuss.
- Do you agree with Karnofsky's "PASTA" framing? Is there another useful framing you prefer?
- Do you think PASTA could be developed via machine learning?
- Karnofsky gives the examples of the "caution" frame and the "competition" frame. Which do you think is more plausible? Do you agree that the competition frame is likely to be overrated?
- Click on one link (that does not link to another part of Cold Takes) and read through it. What did you learn?

Paper Reading

He et al., Deep Residual Learning for Image Recognition (2015)
Ba et al., Layer Normalization (2016)

For this week, slightly different procedure. For each paper:

1. Summarize. Summarize the paper; be sure to read and summarize contents from each section; do not just describe the overall idea of the paper.
2. Judge. Describe a nontrivial strength or weakness of the paper that isn't explicitly mentioned in the paper.

Week Five

For Discussion Section
Carlsmith, Is Power-Seeking AI an Existential Risk? (2021)

Optional:

Omohundro, The Basic AI Drives (2008)

Bostrom, *Superintelligence*, Chapter 7: The Superintelligent Will (2014)

Discussion questions:

- Carlsmith writes,

  > *Nuclear contamination is hard to clean up, and to stop from spreading. But it isn't trying to not get cleaned up, or trying to spread—and especially not with greater intelligence than the humans trying to contain it. But the power-seeking agents just described would be trying, in sophisticated ways, to undermine our efforts to stop them.*

  This appears to be an essential difference between many hazards humans encounter and the potential hazard posed by power-seeking AI. However, some hazards we currently face do have some of the properties described above. Can you think of any? Do you think our efforts to deal with them could be instructive for AI risk?
- Carlsmith is writing particularly about "APS" agents: agents that have advanced capabilities, can plan, and are strategically aware. Do you think he enumerates the important considerations? Are there any that he missed?
- Carlsmith writes:

  > *Optimizing a system to perform some not-intuitively-agential task (for example, predicting strings of text) could, given sufficient cognitive sophistication, result in internal computation that makes and executes plans, in pursuit of objectives, on the basis of broad and informed models of the real world, even if the designers of the system do not expect this (they may even be unable to tell whether it has occurred).*

  Do you think this is true? Give a concrete example where something like this could occur. How likely is it?
- Carlsmith discusses the idea of *instrumental convergence*, and it is also covered in the optional readings. Brainstorm some goals an AI might have that aren't mentioned in any of the readings. Can you imagine ways that power could be helpful for achieving those goals?
- Carlsmith mentions that there are problems with *proxies*. Think of a measurable value that might appear good for an AI to optimize. Now think of some ways that optimizing that proxy to the extreme could be problematic.
- Analyze the following statements. Focus on relevant considerations rather than just trying to decide if a statement is true or not.
  - If AI does not actively seek power, then it cannot create an existential catastrophe.
  - Deep learning cannot possibly be deceptive or power-seeking, because it's just a mechanism with some matrix multiplies and nonlinear functions.
  - Power-seeking is just an anthropomorphism. Non-human systems won't seek power like humans do.
  - We could solve this problem by just building altruistic systems that live in harmony with humans.

Paper Reading

Vaswani et al., Attention Is All You Need (2017)
Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2018)
Hendrycks et al., Unsolved Problems in ML Safety (2021)

For this week, please submit the following for each paper:
1. Summarize. Summarize the paper; be sure to read and summarize contents from each section; do not just describe the overall idea of the paper.
2. Judge. Describe a nontrivial strength or weakness of the paper that isn't explicitly mentioned in the paper.

Week Six

Please do this first: Leveson, Introduction to STAMP (video, slides) (2021)

Write up some notes while watching the video and submit them to gradescope. The notes don't need to be polished.

For Discussion Section
Woodside & Hendrycks, Complex Systems for AI Safety (2022)
Woodside & Hendrycks, Open Problems in AI X-Risk, Robustness section (2022)

Discussion questions:

- Think of some non-safety related problems. These could be problems in the world ("malaria kills thousands annually") or even be problems in your own life ("I procrastinated on my homework"). Try to imagine a chain-of-events model for how that problem was caused. Then try to identify the systemic contributing factors. What are strengths and weaknesses of each model? Which approach do you think works for this problem?
- Are there systemic factors that you think were missed in this reading? Are there systemic factors that don't seem very important?
- Do you think the piece successfully argues that we shouldn't be biased towards simple theories of impact?
- Briefly summarize and explain what the second reading's main claims are about the usefulness of adversarial robustness research for AI x-risk. You may want to check out this slack thread.
- What is the strongest motivation listed in the motivations section? The weakest?
- What is the strongest criticism in the criticisms section? The weakest? Why?

- Consider the adversarial robustness research covered in lecture. Which seem most useful for x-risk?

Paper Reading
(Optional) Madry et al., Towards Deep Learning Models Resistant to Adversarial Attacks (2017)
(Optional) Li et al., BERT-ATTACK: Adversarial Attack Against BERT Using BERT (2020)

Week Seven

For Discussion Section
Woodside & Hendrycks, Open Problems in AI X-Risk, Alignment section and Monitoring section (2022)

Discussion questions:

- For each area: What is the strongest motivation listed in the motivations section? The weakest?
- For each area: What is the strongest criticism in the criticisms section? The weakest? Why?
- Analyze the following statement: “monitoring is not helpful, because by the time there is a serious problem, we are already at the level of an existential catastrophe.”
- Analyze the following statement: “since power-seeking is an instrumental incentive, there is no way to stop agents from trying to seek power”
- Explain and discuss the difference between “honesty” and “truthfulness”. What is the importance of this distinction?
- The reading favors modeling normative ethical factors and aligning to them, rather than aligning to “human preferences.” What are the benefits and drawbacks to this approach?
- The reading says, “Philosophical/fuzzy reasoning ability and raw intelligence seem distinct; by default high-IQ educated people are not especially good at reasoning about fuzzy abstract objects.” Debate this. Maybe look at this thread and this thread.
- What questions did you have about what you learned this week? (not just limited to readings)
- What's an idea you found interesting or thought twice about this week? (not just limited to readings)

Paper Reading
Lipton, The Mythos of Model Interpretability (2016)
Wang et al., ViM: Out-Of-Distribution with Virtual-logit Matching (2022)
Borowski et al., Exemplary Natural Images Explain CNN Activations Better than State-of-the-Art Feature Visualization (2020)

For this week, please submit the following for each paper:
1. Summarize. Summarize the paper; be sure to read and summarize contents from each section; do not just describe the overall idea of the paper.

2. Judge. Describe a nontrivial strength or weakness of the paper that isn't explicitly mentioned in the paper.

Week Eight

For Discussion Section
Woodside & Hendrycks, Open Problems in AI X-Risk, Systemic Safety section (2022)

Paper Reading
Hendrycks & Mazeika, X-Risk Analysis for AI Research (including appendices) (2022)

For this week, please submit the following for each paper:
1. Summarize. Summarize the paper; be sure to read and summarize contents from each section; do not just describe the overall idea of the paper.
2. Judge. Describe a nontrivial strength or weakness of the paper that isn't explicitly mentioned in the paper.

Discussion questions:

- For each area: What is the strongest motivation listed in the motivations section? The weakest?
- For each area: What is the strongest criticism in the criticisms section? The weakest? Why?
- Discuss the following claim: “working towards cooperative AI is not useful, because if agents cooperate with each other, they are likely to collude against humans”
- The post discusses using machine learning to improve epistemics. What are some ways that this could backfire, and cause ML to cause leaders to make worse decisions?
- Discuss the following claim: “plenty of people will work on ML for cybersecurity by default, so it shouldn’t be a priority area.”
- What questions did you have about what you learned this week? (not just limited to readings)
- What's an idea you found interesting or thought twice about this week? (not just limited to readings)

Verified Bugs:

Currently none