Spring 2025, 8 Week Curriculum

Week 0 (prereqs, optional): Introduction to machine learning

Week 0 is intended as an introduction or refresher to machine learning, to provide a more thorough understanding of ML for the rest of the fellowship. The first meeting will discuss Meeting 1 readings, not Week 0 readings.

This week focuses on foundational concepts in machine learning, for those who are less familiar with them, or who want to review the basics. If you'd like to learn about machine learning (ML) in more depth, see the Learn More section at the end of this curriculum. While this week's readings take much longer than those from other weeks, we recommend spending the time to work through them all, to provide a solid foundation for the rest of the course.

After the first reading, which gives a high-level outline of the fields of artificial intelligence and machine learning, the next six readings work through six core concepts. The first three are the three crucial techniques behind deep learning (the leading approach to machine learning): neural networks, gradient descent, and backpropagation. The next three are the three types of tasks which machine learning is used for: supervised learning, self-supervised learning, and reinforcement learning.

Core readings:

1. A short introduction to machine learning (Ngo, 2021) (20 mins)
    - Ngo provides a high-level framework for understanding how different topics in AI connect to each other.
2. But what is a neural network? (3Blue1Brown, 2017a) (20 mins)
    - This video, and the two following videos, provide more details and intuitions about neural networks and the optimization algorithms used to train them.
3. Gradient descent, how neural networks learn (3Blue1Brown, 2017b) (20 mins)
    - See above.
4. What is backpropagation really doing? (3Blue1Brown, 2017c) (15 mins)
    - See above.
5. Machine Learning for Humans, Part 2.1: Supervised Learning (Maini and Sabri, 2017) (15 mins)
    - This accessible reading explains supervised learning, the core task for which ML techniques are typically used.
6. What is self-supervised learning? (CodeBasics, 2021) (5 mins)
    - This reading covers the next-most-prominent ML task: self-supervised learning.
7. Introduction to reinforcement learning (von Hasselt, 2021) **(from 2:00 to 1:02:10, ending at the beginning of the section titled *Inside the Agent: Models*)** (60 mins)

- This video introduces the third main type of task in machine learning: reinforcement learning, which has been used to train networks to play a wide range of games at superhuman levels.

Further readings:

On the basics of neural networks:

1. The spelled-out intro to neural networks and backpropagation: building micrograd (Karpathy, 2022) (150 mins)
    - A lecture introducing the most foundational concepts in deep learning in a very comprehensive way, from a leading expert.
2. Transformers from scratch (Rohrer, 2021)
3. Machine learning for humans (Maini and Sabri, 2017)
    - Maini and Sabri provide a long but accessible introduction to machine learning.
4. Machine learning glossary (Google, 2017)
    - For future reference, see this glossary for explanations of unfamiliar terms.

On reinforcement learning:

5. Spinning up deep RL: part 1 and part 2 (OpenAI, 2018) (40 mins)
    - This reading provides a more technical introduction to reinforcement learning (for more, see also the last half-hour of von Hasselt (2021)).
6. A (long) peek into reinforcement learning (Weng, 2018) (35 mins)
    - Weng provides a concise yet detailed introduction to reinforcement learning.

Week 1: Reward misspecification, RLHF, and deception

A basic obstacle to training aligned AI systems is specifying what we would like our AI systems to do. Historically, AI systems were trained to maximize simple proxy objectives, leading to undesired behavior as AI systems learned to exploit these proxies. Modern frontier AI systems are instead trained using variants of RL from human feedback, where the objective being maximized is human evaluator approval. However, "human evaluator approval" is still an imperfect proxy for the behaviors we actually want; for instance, human evaluators can be deceived. As we scale AI systems to tasks that we cannot reliably evaluate, one concern is that we may select for systems which more competently deceive their human evaluators.

Core readings:

1. Specification gaming: the flip side of AI ingenuity (Krakovna et al., 2020) (15 mins)
2. Deep RL from human preferences: blog post (Christiano et al., 2017) (5 mins)
3. ~~Learning to summarize with human feedback: blog post (Stiennon et al., 2020) (15 mins)~~
4. The alignment problem from a deep learning perspective (Ngo, Chan and Mindermann, 2022) **(only section 2: Deceptive reward hacking)** (10 mins)
5. Language Models Learn to Mislead Humans via RLHF **(Sections 1 and 2)** (10 minutes reading, 10 minutes discussion)
6. Emergent Deception and Emergent Optimization (Steinhardt, 2023) **(ending at "Emergent Optimization")** (10 mins)
7. Gallery of deceptive behavior (Marks, 2023) (10 mins)

8. **Another (outer) alignment failure story (Christiano, 2021) (15 minutes reading, 15 minutes discussion)**

Further readings:

1. On the opportunities and risks of foundation models (Bommasani et al., 2022) **(only pages 3-6)** (10 mins)
2. Aligning language models to follow instructions: blog post (Ouyang et al., 2022) (10 mins)
3. Scaling Laws for Reward Model Overoptimization (Gao et al., 2022) **(sections 1-3 only)** (25 mins)
4. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback (Casper et al., 2023) **(sections 2-4 only)** (25 mins)

Week 2: Goals and goal misgeneralization

This week discusses the notion of goal-directed behavior in AI systems. Such behaviors are common; for instance, agents trained via RL to solve mazes exhibit the goal-directed behavior of "move towards the maze exit." The first reading discusses and gives examples of goal misgeneralization, where AI systems in training develop goals different from the ones their developers intended. Sophisticated AI systems with goals, regardless of what those goals are, could pursue certain instrumental subgoals like power-seeking. Moreover, under certain assumptions, sophisticated AI systems with unaligned goals may instrumentally behave well during training in order to prevent their goals from being modified; this failure mode is called deceptive alignment.

Core readings:

1. Goal misgeneralization: why correct specifications aren't enough for correct goals (Shah et al., 2022) **(only sections 1-4)** (25 mins)
2. Why alignment could be hard with modern deep learning (Cotra, 2021) (20 mins)
3. ML systems will have weird failure modes (Steinhardt, 2022) (15 mins)

Further readings:

1. Goal Misgeneralization in Deep Reinforcement Learning (Langosco et al., 2022) **(only sections 3-3.3)** (15 mins)
2. Optimal policies tend to seek power: NeurIPS spotlight presentation (Turner et al., 2022) (15 mins)
3. The alignment problem from a deep learning perspective (Ngo et al., 2022) **(only sections 3-4)** (20 mins)
4. Language Models as Agent Models (Andreas, 2022) **(sections 1-6 only)** (30 mins)

Week 3: Current trajectory and risk stories

Core readings:

1. Epoch AI Trends Page (2024) (10 min, 10 min)
2. What will GPT-2030 look like? (10 min, 10 min)
3. AIs Accelerating AI Research (Cotra, 2023)(10 min) and Continuous Doesn't Mean Slow (Davidson, 2023)(10 min) (20 min)
4. How we could stumble into AI catastrophe (30 min, 10 min)

Further readings:

1. Biological Anchors: A Trick That Might Or Might Not Work (Alexander, 2022)
2. Future ML Systems will be Qualitatively Different (Steinhardt, 2022)
3. Training Compute-Optimal Large Language Models (Hoffman 2022)
4. Neural scaling laws and GPT-3 (Kaplan, 2020) **(only up to 30:30)** (30 mins)
    a. Kaplan goes into more detail on foundation models, and outlines scaling laws which suggest that there will continue to be large returns to compute used during training.

Week 4: Mechanistic interpretability

Core readings:

1. ~~Zoom In: an introduction to circuits (Olah et al., 2020) (35 mins)~~
2. Toy models of superposition (Elhage et al., 2022) **(up to the end of section 3: superposition as a phase change)** (30 mins)
3. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning (Briken et al., 2023) (30 mins)

Further readings:

1. Multimodal neurons in artificial neural networks (Goh et al., 2021) (35 mins)
2. An Interpretability Illusion for BERT (Bolukbasi et al., 2021) (35 mins)
3. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small (Wang et al., 2022) (30 mins)
4. Superposition, Memorization, and Double Descent (Henighan et al., 2023) (30 mins)

Week 5: Control

Core readings:

1. The case for ensuring powerful AIs are controlled (Greenblatt et al., 2024) (30 min)
2. AI Control: Improving safety despite intentional subversion (Greenblatt et al., 2024) (30 min)
3. youtube.com/watch?v=0pgEMWy70Qk

Further readings:

1. Locating and Editing Factual Associations in GPT (Meng et al., 2022)
2. Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task (Li et al., 2023) (30 mins)
3. Causal Scrubbing: a method for rigorously testing interpretability hypotheses (Chan et al., 2022)
4. Adversarial Examples Are Not Bugs, They Are Features (Ilyas et al., 2019) (30 mins)
5. Jailbroken: How Does LLM Safety Training Fail?
6. Revisiting Model Stitching to Compare Neural Representations (Bansal et al., 2021) (30 mins)

Week 6: Scalable oversight

Core readings:

1. Why I'm excited about AI-assisted human feedback (Leike, 2022) (15 mins)
2. AI Safety via debate (Amodei and Irving, 2018) (10 mins)
3. Debating with More Persuasive LLMs Leads to More Truthful Answers (20)
4. Weak-to-strong generalization (Burns et al., 2023) (20 min)
5. https://www.youtube.com/watch?v=8MqBharGmo0

Further readings:

1. Humans consulting HCH (Christiano, 2016) (15 mins)
    - HCH stands for "Humans consulting HCH," and is a way to think about the ultimate goal of iterated distillation and amplification schemes.
2. Factored cognition (Ought, 2019) **(introduction and scalability section)** (20 mins)
    - This reading gives a framing of IDA schemes in terms of factoring problems into subproblems which can be worked on without broader context.
    - Chain of thought imitation with procedure cloning (Yang et al., 2022) (35 mins)
    - Yang et al. introduce Procedural Cloning, in which an agent is trained to mimic not just expert outputs, but also the process by which the expert reached those outputs.
3. Measuring Progress on Scalable Oversight for Large Language Models (Bowman et al., 2022) **(sections 1-3 only)** (25 mins)
4. Debate update: obfuscated arguments problem (Barnes and Christiano, 2020) (15 mins)
    - This reading explains why in the worst case, it's not possible to judge a debate without adjudicating a prohibitively large number of subclaims.

Week 7: Red teaming

1. Evaluating Language Models on Realistic Autonomous Tasks (Kinniment et al., 2023) (10 mins)
2. Universal and Transferable Adversarial Attacks on Aligned Language Models (Zou et al., 2023) **(sections 1-2 only)** (20 mins)

3. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training (30 min) *still determining which sections
4. Change my mind:
    a. https://www.reddit.com/r/changemyview/comments/1k8b2hj/meta_unauthorized_experiment_on_cmv_involving/
    b. https://drive.google.com/file/d/1Eo4SHrKGPErTzL1t_QmQhfZGU27jKBjx/view

Further readings:

1. Mechanistic anomaly detection and ELK (Christiano, 2022) (25 mins)
2. Constitutional AI: Harmlessness from AI Feedback (Bai et al., 2022) **(only sections 1, 3.1, and 4.1)** (20 mins)
3. Robust Feature-Level Adversaries are Interpretability Tools (Casper et al., 2021) (30 mins)
4. Adversarial Robustness as a Prior for Learned Representations (Engstrom et al., 2019) (30 mins)

Week 8: Policy, model evaluations, and careers in alignment

Core readings:

1. https://evals.alignment.org/Evaluating_LMAs_Realistic_Tasks.pdf (25 min)
2. Primer on Safety Standards and Regulations for Industrial-Scale AI Development (15 min)
3. The AI regulator's toolbox: A list of concrete AI governance practices (Jones, 2024)
4. Careers in alignment (Ngo, 2022) (35 mins)

Further readings:

1. What does it take to catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring (Shavit, 2023) (30 mins)
    - This paper lays out a regulatory framework for monitoring large training runs in which (1) GPU manufacturers install on-chip mechanisms for logging certain difficult-to-spoof data about the chips' usage, (2) regulators occasionally visit and inspect large compute clusters, using the logged data to verify that the GPUs are being used consistently with established rules.
2. Discovering Language Model Behaviors with Model-Written Evaluations (Perez et al., 2023) (40 mins)
    - The authors use language models to generate datasets for behaviorally evaluating language models. Some evidence of potentially risky model tendencies – like expressing desire not to be shut off, or acting sycophantically towards the user – is found.
3. What AI companies can do today to help with the most important century (Karnofsky, 2023)
    - Karnofsky discusses what he would like to see scaling labs do to help mitigate catastrophic risks from AI.

Weeks 9-12: Optional additional research project