Summer 2025, 8 Week Curriculum

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

Week 0 is intended as an optional introduction to machine learning, to provide a more thorough understanding of machine learning (ML), deep learning (DL), reinforcement learning (RL), and transformers for the rest of the fellowship. The first meeting will discuss Week 1 readings, not Week 0 readings.

Main recommended resource:

1. Neural Networks Chapters 1-6 (3Blue1Brown, 2017-2024) (2 hours).

Other recommended resources:

1. A short introduction to machine learning (Ngo, 2021) (20 mins)
2. Machine Learning for Humans, Part 2.1: Supervised Learning (Maini and Sabri, 2017) (15 mins)
3. What is self-supervised learning? (CodeBasics, 2021) (5 mins)
4. 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)
5. The spelled-out intro to neural networks and backpropagation: building micrograd (Karpathy, 2022) (150 mins)
6. Transformers from scratch (Rohrer, 2021)
7. Machine learning for humans (Maini and Sabri, 2017)
8. Machine learning glossary (Google, 2017)
9. Spinning up deep RL: part 1 and part 2 (OpenAI, 2018) (40 mins)
10. A (long) peek into reinforcement learning (Weng, 2018) (35 mins)

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. The alignment problem from a deep learning perspective (Ngo, Chan and Mindermann, 2022) **(only section 2: Deceptive reward hacking)** (5 mins)
4. Language Models Learn to Mislead Humans via RLHF **(Sections 1 and 2)** (10 minutes)
5. Gallery of deceptive behavior (Marks, 2023) (10 mins)
6. Another (outer) alignment failure story (Christiano, 2021) (15 mins)

Further readings:

1. Emergent Deception and Emergent Optimization (Steinhardt, 2023) **(ending at "Emergent Optimization")** (10 mins)
2. On the opportunities and risks of foundation models (Bommasani et al., 2022) **(only pages 3-6)** (10 mins)
3. Aligning language models to follow instructions: blog post (Ouyang et al., 2022) (10 mins)
4. Scaling Laws for Reward Model Overoptimization (Gao et al., 2022) **(sections 1-3 only)** (25 mins)
5. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback (Casper et al., 2023) **(sections 2-4 only)** (25 mins)
6. Situational Awareness (Piper, 2023) (5 mins)
7. Playing the Training Game (Piper, 2023) (5 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. Why alignment could be hard with modern deep learning (Cotra, 2021) (20 mins)
2. Goal misgeneralization: why correct specifications aren't enough for correct goals (Shah et al., 2022) **(only sections 1-4)** (25 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) and intro of Can AI scaling continue through 2030? (10 min)
2. AIs Accelerating AI Research (Cotra, 2023) and Continuous Doesn't Mean Slow (Davidson, 2023) (20 min)
3. What will GPT-2030 look like? (10 min)
4. How we could stumble into AI catastrophe (20 min)

Further readings:

1. Training Compute-Optimal Large Language Models (**only up to section 4**) (20 min)
2. Biological Anchors: A Trick That Might Or Might Not Work (Alexander, 2022)
3. Future ML Systems will be Qualitatively Different (Steinhardt, 2022)
4. Training Compute-Optimal Large Language Models (Hoffman 2022)
5. 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. Toy models of superposition (Elhage et al., 2022) **(up to the end of section 3: superposition as a phase change)** (30 mins)
2. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning (Briken et al., 2023) **(only: Problem Setup, Arab Script Feature (only up to first image), Global Analysis, Phenomenology)** (30 mins)

Further readings:

1. Zoom In: an introduction to circuits (Olah et al., 2020) (35 mins)
2. Multimodal neurons in artificial neural networks (Goh et al., 2021) (35 mins)
3. An Interpretability Illusion for BERT (Bolukbasi et al., 2021) (35 mins)
4. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small (Wang et al., 2022) (30 mins)
5. Superposition, Memorization, and Double Descent (Henighan et al., 2023) (30 mins)
6. post (10 mins)
7. Discovering latent knowledge in language models without supervision (Burns et al., 2022) **(first 5 pages only)** (20 mins)
8. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model (Li et al., 2022) (only sections 1-3) (30 mins)

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

Further readings:

1. All readings on AI Control
2. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training (30 min)
    - Sleeper Agents is an example of a "Model Organisms of Misalignment" paper. It intentionally trains a deceptive model in order to test alignment and control techniques.
    - Simple probes can catch Sleeper Agents (MacDiarmid, 2024) (15 mins) is one such example of a control technique.
3. What AI companies can do today to help with the most important century (Karnofsky, 2023)

Week 6: Scalable oversight and Weak-to-strong generalization

Core readings:

1. Why I'm excited about AI-assisted human feedback (Leike, 2022) (10 mins)
2. AI Safety via debate (Amodei and Irving, 2018) (10 mins)
3. Debating with More Persuasive LLMs Leads to More Truthful Answers (Khan et al., 2024) (20 mins)
4. Weak-to-strong generalization (Burns et al., 2023) **(up to section 5)** (15 mins)
5. Combining W2SG with SO (Leike, 2023) (10 mins)

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 and Evaluations

1. Universal and Transferable Adversarial Attacks on Aligned Language Models (Zou et al., 2023) **(sections 1-2 only)** (20 mins)
2. Red Teaming Language Models with language Models (Perez et al., 2022) **(only up to section 3)** (10 mins)
3. Evaluating Language Models on Realistic Autonomous Tasks (Kinniment et al., 2023) (25 mins)
4. Play around with Gandalf AI (If time remains)

Further readings:

1. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training (30 min)
2. METR's Autonomy Evaluation Resources (2024)
3. Constitutional AI: Harmlessness from AI Feedback (Bai et al., 2022) **(only sections 1, 3.1, and 4.1)** (20 mins)

Week 8: Policy, Careers in Alignment

Core readings:

1. The AI regulator's toolbox: A list of concrete AI governance practices (Jones, 2024) (20 minutes)
2. Opportunities in alignment (Bashansky, 2024) (35 mins)