Introduction

As ML systems become more capable and widely deployed, concerns are growing around safety. For example, DeepMind and OpenAI both have safety teams.

Some of these concerns are near-term: how do we prevent driverless cars from misidentifying a stop sign in a blizzard? Others are more long-term: if general AI systems are built, how do we make sure these systems pursue safe goals and benefit humanity? This course serves as an introduction to the body of technical research relevant to both but emphasizes long-term, high-consequence risks. It also explores the threat models for these risks. Could future AI systems pose an existential threat?

As with other powerful technologies, safety for ML should be a leading research priority. In that spirit, we want to bring you to the frontiers of this nascent field.

Please read the participant guide. It includes information about logistics (e.g. how to submit homework), course policies, etc. If you have any questions, DM James Aung on slack.

The lectures are recorded by Dan Hendrycks, UC Berkeley ML PhD and director of the Center for AI Safety.

The syllabus is subject to change.

Syllabus

Coding Assignments

Pick 2 coding assignments to complete during the course

- Deep Learning (not recommended for those who already have DL experience) [done in Week 1]
- Adversarial Robustness [done in Week 3]
- Anomaly Detection [done in Week 4]
- Machine Ethics [done in Week 6]

Week 1: Background

Course Page: Background

Lecture: Background (Introduction, Deep Learning Review)

Written Assignment: https://www.gradescope.com/courses/535195/assignments/2855764/

Reading Assignment: Unsolved Problems in ML Safety

Coding Assignment (optional; pick 2 to do over course): Deep Learning (not recommended for those who already have sufficient DL experience)

Optional readings:

AI Safety:

- Concrete Problems in AI Safety
- Workshop On Safety And Control For Artificial Intelligence

Deep Learning Review:

- Deep Learning Review Lecture Notes
- Deep Residual Learning for Image Recognition
- Attention Is All You Need
- An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale
- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
- Decoupled Weight Decay Regularization
- Layer Normalization
- Gaussian Error Linear Units (GELUs)
- Dropout: A Simple Way to Prevent Neural Networks from Overfitting
- Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift
- Adam: A Method for Stochastic Optimization

Week 2: Safety Engineering

Course Page: Safety Engineering

Lecture: Safety Engineering (Risk Decomposition, Accident Models, Black Swans)

Written Assignment: Risk Models

Reading Assignment: Choose one of:

1. **More is different for AI (does not need to be summarized) and Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover**
2. **or Adversarial Examples for Evaluating Reading Comprehension Systems**

Optional readings:

- Power laws, Pareto distributions and Zipf's law
- What is a Complex System?
- Emergence (Intermediate)
- Introduction to STAMP (video, slides)
- Safe Design, Consequence assessment, Safety models
- A Brief History of Generative Models for Power Law and Lognormal Distributions
- Empirical examples of power-law CDFs
- Log-normal distributions (with comparisons to power laws)
- Multiplicative processes produce log normals
- Multiplicative processes produce power laws
- Review of properties of power laws; Numerous power laws for cities
- The Black Swan and Antifragile Summaries
- The Precautionary Principle (skip second half)
- Emergence (Basic); Nonlinear Causality
- Beyond Normal Accidents and High Reliability Organizations: The Need for an Alternative Approach to Safety in Complex Systems
- Shortcomings of the Bow Tie and Other Safety Tools Based on Linear Causality
- Systemantics Appendix
- How Complex Systems Fail

Week 3: Robustness

Course Page: Robustness

Lecture: Robustness (Adversarial Robustness, Black Swan Robustness).

Written Assignment: Robustness

Reading Assignment: Towards Deep Learning Models Resistant to Adversarial Attacks

Coding Assignment (optional; pick 2 to do over course): Adversarial Robustness

Optional readings:
Adversarial Robustness:

- Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples
- Towards Deep Learning Models Resistant to Adversarial Attacks
- Universal Adversarial Triggers for Attacking and Analyzing NLP
- Data Augmentation Can Improve Robustness
- Adversarial Examples for Evaluating Reading Comprehension Systems
- BERT-ATTACK: Adversarial Attack Against BERT Using BERT (GitHub)
- Gradient-based Adversarial Attacks against Text Transformers
- Smooth Adversarial Training
- Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks (website)
- Certified Adversarial Robustness via Randomized Smoothing
- Adversarial Examples Are a Natural Consequence of Test Error in Noise
- Using Pre-Training Can Improve Model Robustness and Uncertainty
- Motivating the Rules of the Game for Adversarial Example Research
- Certified Defenses against Adversarial Examples
- Towards Evaluating the Robustness of Neural Networks

Long Tails and Distribution Shift:

- The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization
- Benchmarking Neural Network Robustness to Common Corruptions and Perturbations
- PixMix: Dreamlike Pictures Comprehensively Improve Safety Measures
- WILDS: A Benchmark of in-the-Wild Distribution Shifts
- ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models
- Adversarial NLI: A New Benchmark for Natural Language Understanding
- Natural Adversarial Examples
- ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness

Week 4: Monitoring, Part 1

Course Page: Monitoring

Lecture: Monitoring, Part 1 (Anomaly Detection, Interpretable Uncertainty)

Written Assignment: Monitoring – not due until next week.

Reading Assignment: On Calibration of Modern Neural Networks

Coding Assignment (optional; pick 2 to do over course): Anomaly Detection

Optional readings:
OOD and Malicious Behavior Detection:

- Deep Anomaly Detection with Outlier Exposure
- A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks
- ViM: Out-Of-Distribution with Virtual-logit Matching
- VOS: Learning What You Don't Know by Virtual Outlier Synthesis
- Scaling Out-of-Distribution Detection for Real-World Settings
- A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks

Interpretable Uncertainty:

- On Calibration of Modern Neural Networks
- Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift
- PixMix: Dreamlike Pictures Comprehensively Improve Safety Measures
- Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles
- Posterior calibration and exploratory analysis for natural language processing models
- Accurate Uncertainties for Deep Learning Using Calibrated Regression

Week 5: Monitoring, Part 2

Course Page: Monitoring

Lecture: Monitoring, Part 2 (Transparency, Trojans, Detecting Emergent Behavior)

Written Assignment: Monitoring

Reading Assignment: Toy Models of Superposition

Optional readings:

Transparency:

- The Mythos of Model Interpretability
- Sanity Checks for Saliency Maps
- Interpretable Explanations of Black Boxes by Meaningful Perturbation
- Locating and Editing Factual Knowledge in GPT
- Acquisition of Chess Knowledge in AlphaZero
- Feature Visualizations and OpenAI Microscope
- Exemplary Natural Images Explain CNN Activations Better than State-of-the-Art Feature Visualization
- Network Dissection: Quantifying Interpretability of Deep Visual Representations
- Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead
- Convergent Learning: Do different neural networks learn the same representations?

Trojans:

- Poisoning and Backdooring Contrastive Learning
- Universal Litmus Patterns: Revealing Backdoor Attacks in CNNs
- Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks
- TrojAI
- Detecting AI Trojans Using Meta Neural Analysis
- STRIP: A Defence Against Trojan Attacks on Deep Neural Networks
- Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning
- BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain

Detecting and Forecasting Emergent Behavior:

- The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models
- The Basic AI Drives
- Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets
- Optimal Policies Tend to Seek Power
- The Off-Switch Game
- Goal Misgeneralization in Deep Reinforcement Learning

Week 6: Control

Course Page: Control

Lecture: Control (Honest Models, Machine Ethics)

Written Assignment: Start working on review

Reading Assignment: Discovering Latent Knowledge in Language Models Without Supervision

Coding Assignment (optional; pick 2 to do over course): Machine Ethics

Optional readings:
Catastrophic AI Risks:

- An Overview of Catastrophic AI Risks

Honest AI:

- TruthfulQA: Measuring How Models Mimic Human Falsehoods
- Truthful AI: Developing and governing AI that does not lie

Machine Ethics:

- What Would Jiminy Cricket Do? Towards Agents That Behave Morally
- Ethics Background (Introduction through "Absolute Rights or Prima Facie Duties")
- Aligning AI With Shared Human Values
- Avoiding Side Effects in Complex Environments
- Conservative Agency via Attainable Utility Preservation
- The Structure of Normative Ethics

Week 7: Systemic Safety

Course Page: Systemic Safety

Lecture: Systemic Safety (ML for Improved Decision-Making, ML for Cyberdefense, Cooperative AI)

Written Assignment: review (due next week)

Reading Assignment:

- **An Overview of Catastrophic AI Risks**

Optional readings:

Forecasting:

- Forecasting Future World Events with Neural Networks
- On Single Point Forecasts for Fat-Tailed Variables
- On the Difference between Binary Prediction and True Exposure With Implications For Forecasting Tournaments and Decision Making Research
- Superforecasting – Philip Tetlock

ML for Cyberdefense:

- Asleep at the Keyboard? Assessing the Security of GitHub Copilot's Code Contributions

Cooperative AI:

- Uehiro Lectures 2022
- Open Problems in Cooperative AI
- Cooperation, Conflict, and Transformative Artificial Intelligence: A Research Agenda

Week 8: Additional Existential Risk Discussion

Course Page: Additional X-Risk Discussion

Lecture: X-Risk, Possible Existential Hazards, Safety-Capabilities Balance, Review and Conclusion

Written Assignment: review (due next week)

Reading Assignment:

- **Natural Selection Favors AIs over Humans (sections 1-3)**

Optional readings:

- X-Risk Analysis for AI Research
- Can we build AI without losing control over it?
- What happens when our computers get smarter than we are?
- X-Risk Motivations for Safety Research Directions