CS 281 Ethics of AI – Reading list 2025

This is a list of suggested (but not mandatory) readings for the course sorted by topic. The list is not meant to be static, and we will be updating entries to this list as the course progresses.

- Fairness & Bias
  - (textbook) FAIRNESS AND MACHINE LEARNING Limitations and Opportunities
  - Fairness, Equality, and Power in Algorithmic Decision-Making
  - Equality of opportunity in supervised learning
  - Fairness Through Awareness
  - Delayed Impact of Fair Machine Learning
  - Learning Fair Representations
  - Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings
  - Learning controllable Fair Representations
  - FACT: A Diagnostic for Group Fairness Trade-offs
  - Right Decisions from Wrong Predictions: A Mechanism Design Alternative to Individual Calibration
  - Retiring Adult: New Datasets for Fair Machine Learning
  - The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning
  - On Fairness and Calibration
  - Calibration for the (Computationally-Identifiable) Masses
  - Predicting Good Probabilities With Supervised Learning

- Explainability
  - (textbook) Interpretable Machine Learning A Guide for Making Black Box Models Explainable
  - "Why Should I Trust You?": Explaining the Predictions of Any Classifier
  - A Unified Approach to Interpreting Model Predictions
  - Actionable Recourse in Linear Classification
  - Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)
  - Understanding Black-box Predictions via Influence Functions
  - Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models

    - Sanity Checks for Saliency Maps
    - Learning to Explain: An Information-Theoretic Perspective on Model Interpretation
    - Interpretable Machine Learning: Moving From Mythos to Diagnostics
    - Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead

- Privacy
    - (textbook) The Algorithmic Foundations of Differential Privacy
    - Privacy as Contextual Integrity
    - Deep Learning with Differential Privacy
    - Machine Unlearning
    - A Taxonomy of Privacy