ML and society: Fairness, privacy, explainability

This includes problems such as: visual interpretability of neural networks, identifying or mitigating fairness problems in ML models, learning with encrypted data.

Topics related to fairness/bias:

9a. ML Fairness Gym
Citation: Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern. 2020. Fairness is not static: deeper understanding of long term fairness via simulation studies. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). [PDF]
[Github][Notebook demo of an older version (via ffund)]

9b. Detecting bias in voice recognition
Citation: Meyer, J., Rauchenstein, L., Eisenberg, J.D. and Howell, N., 2020, May. Artie Bias Corpus: An Open Dataset for Detecting Demographic Bias in Speech Applications. In Proceedings of The 12th Language Resources and Evaluation Conference (pp. 6462-6468). [PDF] [Github via Artie]

9c. Adversarial debiasing (bias mitigation)
Citation: Brian Zhang, Blake Lemoine and Margaret Mitchell. Mitigating Unwanted Biases with Adversarial Learning. AAAI Conference on AI, Ethics and Society, 2018. [PDF] [Notebook (via Google)][Notebook (via IBM AIF360)]
[Notebooks on other bias mitigation techniques (see "examples") (via IBM AIF360)]
[Some video tutorials on IBM AIF360]

Topics related to privacy:

9d. Secure multiparty computation (with CrypTen)
Citation: Gunning, D., Hannun, A., Ibrahim, M., Knott, B., van der Maaten, L., Reis, V., Sengupta, S., Venkataraman, S. & Zhou, X. (2019) CrypTen: a new research tool for secure machine learning with PyTorch. [Blog post]
[Overview of Secure Multiparty Computation] [Github] [Tutorial notebooks]

9e. Federated learning (with PySyft)
Citation: Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, PMLR 54:1273-1282, 2017. [PDF]
Citation: Theo Ryffel, Andrew Trask, Morten Dahl, Bobby Wagner, Jason Mancuso, Daniel Rueckert, and Jonathan Passerat-Palmbach. "A generic framework for privacy preserving deep learning." arXiv preprint arXiv:1811.04017 (2018). [PDF] [Github]
[Tutorial notebooks]