# **Reading Materials**

## **Reading list for Bias in AI tutorial 4**

### **Papers**

1. [A Framework for Understanding Unintended Consequences of Machine Learning](https://arxiv.org/abs/1901.10002)
2. [Learning Adversarially Fair and Transferable Representations](https://arxiv.org/abs/1802.06309)
3. [Flexibly Fair Representation Learning by Disentanglement](https://arxiv.org/abs/1906.02589)
4. [Fairness without Demographics through Adversarially Reweighted Learning](https://arxiv.org/abs/2006.13114)
5. [Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements](https://arxiv.org/abs/1901.04562)
6. [Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation](https://arxiv.org/abs/1911.11834)
7. [Image Counterfactual Sensitivity Analysis for Detecting Unintended Bias](https://arxiv.org/abs/1906.06439)
8. [A Mulching Proposal](https://ironholds.org/resources/papers/mulching.pdf)
9. [Dissecting racial bias in an algorithm used to manage the health of populations](https://www.ehidc.org/sites/default/files/resources/files/Dissecting%20racial%20bias%20in%20an%20algorithm%20used%20to%20manage%20the%20health%20of%20populations.pdf)
10. [Phsysiognomy’s New Clothes](https://github.com/google-research/disentanglement_lib.git)
11. [Racial categories in machine learning](https://arxiv.org/abs/1811.11668)
12. [Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings](https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf)
13. [Distributionally Robust Language Modeling](https://arxiv.org/abs/1909.02060)
14. [Measuring and Reducing Gendered Correlations in Pre-trained Models](https://arxiv.org/abs/2010.06032)
15. [Counterfactual Fairness in Text Classification through Robustness](https://arxiv.org/abs/1809.10610)
16. [Inoculation by Fine-Tuning: A Method for Analyzing Challenge Datasets](https://arxiv.org/abs/1904.02668)
17. [Extracting Training Data from Large Language Models](https://arxiv.org/abs/2012.07805)
18. [RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models](https://arxiv.org/abs/2009.11462)
19. [On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?](http://faculty.washington.edu/ebender/papers/Stochastic_Parrots.pdf)
20. [Datasheets for Datasets](https://arxiv.org/abs/1803.09010)
21. [Improving fairness in machine learning systems: What do industry practitioners need?](https://arxiv.org/abs/1812.05239)
22. [Language (Technology) is Power: A Critical Survey of "Bias" in NLP](https://arxiv.org/abs/2005.14050)
23. [Combating Anti-Blackness in the AI Community](https://arxiv.org/abs/2006.16879)
24. [Model Cards for Model Reporting](https://arxiv.org/abs/1810.03993)
25. [Indigenous Protocols and AI](https://www.indigenous-ai.net/position-paper)

### **Datasets and codebases**

1. [disentanglement\_lib](https://github.com/google-research/disentanglement_lib)
2. [Celeb-A](http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html)
3. [Civil comments](https://github.com/google-research/disentanglement_lib.git)