

# Annotated reading list for COS 597E: Fairness in Machine Learning [Fall 2017]

Arvind Narayanan

A note to the reader: this is not intended to be a complete or even a representative reading list. There are important papers and research directions that aren't included here.

Order of modules: 1, 4, 2, 3, 5, 6, 8, 7, 9. We'll move module 4 up so that we can get to the technical content relatively quickly.

## Module 1. Background / intro

Discussion date: Tuesday Sep 19

[Hardt, How big data is unfair](#)

A blog post that presents 5 ways in which machine learning leads to biased outcomes.

Note: this and the following titles refer to “big data” because they are written for a broad audience, but they’re really talking about machine learning.

[Barocas and Selbst, Big Data’s Disparate Impact \[Part 1 only\]](#)

This is one of the foundational papers in the emerging field of Fairness in ML. We will read the intro and Part 1, which is about how data mining / machine learning acquires biases. The rest of the paper is about legal responses to the resulting discrimination.

*The White House Office of Science and Technology Policy, [Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights](#)*

This 2016 report from the Obama White House is a good high-level overview of bias/unfairness in ML.

*Crawford, Whittaker, et al. [The AI Now Report The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term.](#)*

While our course focuses on fairness in ML, other societal concerns of AI/ML are also important to keep in mind, such as the displacement of labor.

*Drosou, et al., [Diversity in Big Data: A Review](#) (all except “Algorithms” section)*

While the societal impact of ML in classification/decision-making (e.g., criminal justice) rightly gets a lot of attention, the information retrieval context (e.g., web search) is also important. This paper surveys the area.

### **Optional readings**

Two whitepapers that explain the rise of data and algorithms in making consequential decisions about people, such as credit and loans.

*Robinson and Yu, [Knowing the Score](#)*  
*Dixon and Gellman, [The Scoring of America](#)*

Note: we'll jump to module 4 after module 1.

### **Discussion prompts [responses due Monday 9/18]**

Each of the first 3 readings (Hardt, Barocas & Selbst, White House report) provides a taxonomy of biases in machine learning. Map these taxonomies to each other. Are there cases where the authors talk about the same phenomenon using different terms? Are there some categories discussed by one paper that are missed by the others?

Find examples of biases in machine learning from the press (or research papers), and explain them using the categories in these papers.

Consider the low level of demographic/cultural diversity in Google/Bing image search results for queries such as "CEO" and other occupations. Do you think search engines should try to change these results? Why or why not? The Drosou et al. survey presents models and definitions for diversity. Can you use these ideas to formulate how image search engines might go about improving diversity in search results if they chose to? What are the main barriers they would encounter?

### **Module 2. Decision-making by humans and machines**

[Discussion date: Tuesday, Oct 3]

*Dawes et al. [Clinical versus actuarial judgement.](#)*

Way back in 1954, Meehl wrote a [pioneering book-length survey](#) showing that statistical decision-making was much more accurate than human judgement in the domain of psychiatry. This 1989 paper coauthored by Meehl makes the same point.

Pager & Shepherd. [\*The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets\*](#)

A primer on discrimination from a sociology perspective.

[If you're unable to access the full text from the above link, use [this link](#).]

Bertrand & Mullainathan. [\*Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination\*](#)

A classic study design for uncovering discrimination. Similar studies are now being used for uncovering bias in algorithmic systems, which we'll get to in module 9.

Dietvorst. [\*Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err\*](#)

A peculiar human bias: against algorithmic decision-making.

Jung et al. [\*Simple Rules for Complex Decisions\*](#)

A statistical approach to help human decision-makers mitigate bias.

### **Optional readings:**

Kahneman & Tversky. [\*Judgment under Uncertainty: Heuristics and Biases\*](#).

This classic paper explains several mental shortcuts that people take that result in biased and inaccurate decision-making.

Menand [\*Everybody's an expert\*](#) (a review of Tetlock. [\*Expert political judgment. How good is it? How can we know?\*](#))

Tetlock, who's been called the "leading expert on leading experts", shows that in many domains such as political prediction, so-called experts perform worse than chance (!)

Kleinberg et al. [\*Human Decisions and Machine Predictions\*](#).

Another (draft) paper arguing that automated decision-making leads to more accurate and/or fair outcomes, this time in the criminal-justice context.

Harcourt. [\*Against prediction\*](#).

This class (or at least the instructor) takes the point of view that while there are bias pitfalls in ML, it is more transparent than human decision-makers, and more amenable to detection,

measurement, and mitigation of bias. Thus, we should be cautiously optimistic about automated decision-making.

Not everyone subscribes to this view. Harcourt argues against the use of data-driven methods in criminal justice, and indeed rejects predictive accuracy as a goal to strive for.

### **Discussion prompts [responses due Monday 10/2]**

Take one or more Implicit Association Tests offered online by Project Implicit. You don't need to write anything about it in your response, but feel free to do so.

<https://implicit.harvard.edu/implicit/takeatest.html>

Bertrand and Mullainathan consider two broad reasons why employers might discriminate: “taste-based” and “statistical”. Both of these constitute conscious discrimination, but employers might discriminate without even knowing it, because of their implicit attitudes. Finally, the paper also discusses alternative explanations for the observations that may not constitute discrimination under some definitions. Describe follow-on experiments you might conduct that will distinguish between these categories (and, if you like, sub-categories of these four broad categories).

In the Dawes et al. survey, there are many findings showing human decision-making to be so flawed that it seems paradoxical. What were some of the findings that surprised you? What are some possible explanations for these?

Based on the readings, what are some steps we can take to minimize biases and discrimination, whether as individuals or as a society?

### **Module 3. Strengths and weaknesses of machine learning**

[Discussion date: Tuesday, Oct 10]

Machine learning de-emphasizes causality, explanation, and domain-knowledge compared to traditional science and statistics. This is one of its strengths, but also introduces pitfalls, including bias traps.

Breiman. [\*Statistical Modeling: The Two Cultures\*](#)

A classic paper that points out the strengths of the machine learning approach.

Norvig. [\*On Chomsky and the Two Cultures of Statistical Learning\*](#)

A blog post that also defends the machine learning approach.

Goldberg. [\*An Adversarial Review of “Adversarial Generation of Natural Language”\*](#).

A blog post that shows the pitfalls of rejecting domain knowledge in machine learning.

Our discussion of the above readings will include figuring out whether and how these attributes of the machine-learning approach might exacerbate bias.

Lazer et al. [\*The parable of Google Flu\*](#).

A case study of a spectacular failure of big data.

Olteanu et al. [\*Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries\*](#)

Often researchers use machine learning applied to online social data to derive insights about people and society (i.e., computational social science). This field has an incredible number of pitfalls, including bias.

### **Discussion prompts**

[Responses due Monday Oct 9]

One critique of algorithmic modeling is that such models aren't easily interpretable, and thus unsuitable in contexts where explanation is the goal, rather than prediction. Summarize how Breiman responds to this critique in his paper.

The Google Flu paper mentions the possibility of adversarial manipulation (“red team attacks”) of online data to change the behavior of algorithmic systems or the conclusions drawn from such data. Find examples of cases where this was done successfully (this isn't specific to Google Flu, but rather any algorithmic system).

Olteanu et al. present over 40 types of biases and pitfalls that arise when using social data for research. What are we to conclude from this -- should we consider all research that uses such data to be unscientific? Do you know of any studies that you think of as success stories in terms of research insights derived from the analysis of social data? What were some strategies used in those studies to avoid these biases and pitfalls?

### **Optional reading:**

[\*Domain Expertise vs Machine Learning Debate\*](#)

An informative video debate on machine learning versus domain expertise.

### **Module 4. Individual & group fairness, and the impossibility theorem**

[Start of discussion: Thursday Sep 21]

## [responses due Monday 9/25]

Angwin et al. [\*Machine Bias\*](#)

A piece of investigative reporting that kicked off a huge debate in both academic and activist circles.

Angwin & Larson. [\*Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say\*](#)

Several papers followed, proving variants of an impossibility theorem. It might be helpful to read the above post first, before delving into the papers below.

Chouldechova. [\*Fair Prediction with Disparate Impact: A study of bias in recidivism prediction instruments\*](#).

Kleinberg et al. [\*Inherent Trade-Offs in the Fair Determination of Risk Scores\*](#)

The above are the two core “impossibility theorem” papers.

Corbett-Davies et al. [\*Algorithmic Decision Making and the Cost of Fairness\*](#).

Hardt et al. [\*Equality of Opportunity in Supervised Learning\*](#) [Sections 1--3, 7, 8]

These two papers are more general, and also consider the trade-off with utility, in addition to different notions of fairness.

Dwork et al. [\*Fairness Through Awareness\*](#).

This paper explores individual fairness in a rigorous, mathematical way. It was also one of the early papers to point out that blindness doesn't lead to fairness (reflected in its title). This is one of the core insights in this field.

## Discussion prompts

[All but the last question are about the four papers connected to the ProPublica piece.]

All four papers talk about false positives and false negatives. What is the definition of positive class/instance in each of the four papers? Are they the same?

Calculating false positive / false negative rates requires knowing the probabilities of classifier outputs conditioned on the eventual outcomes. In other words, it requires knowing things like “what fraction of loan applicants were denied among those who would have repaid if approved?”

Naturally, such probabilities are hard to obtain empirically, and in general, might require randomized trials. No such experiments were performed in the COMPAS analysis.

How did ProPublica manage to compute these probabilities?

Chouldechova points out that a score that is calibrated may not have predictive parity. Draw a picture that illustrates this idea. Use whatever visualization you like. For example, you could draw two distributions of risk scores, one for the majority group and one for the minority group.

[Note: this is just an exercise in visualization. Feel free to draw on a piece of paper (or a napkin) and upload a photo.]

In addition to the individual and group fairness criteria studied in the four papers, do you have any other ethical concerns with the use of COMPAS risk assessments? Are there other fairness criteria we should think about?

Chouldechova (Section 4) and Hardt et al. (throughout the paper; reflected in the title) each have a preferred approach to navigating the impossibility result. Compare the two approaches to each other. Which one (if either) do you prefer, and why?

The COMPAS score turned out to be well calibrated between groups, even though group membership (race) isn't an input to the scoring function. Does this surprise you? If you were Northpointe, how would you go about ensuring that the score is calibrated?

In Hardt et al. the unconstrained classifier picks race-specific thresholds, but in Corbett-Davies et al. it doesn't. What gives?

Apply the fairness-through-awareness framework to the criminal risk prediction setting. What might the similarity metric look like? How does this framework differ from the COMPAS approach?

## **Module 5. Bias sources and pathways**

[Start of discussion: Thursday 10/19]

Key themes: how can we design experiments to distinguish between different sources of bias, and to show that observed patterns are due to bias and not other confounding factors? How does bias propagate from training data to models, and how does bias in ML systems impact users of those systems?

Pierson et al. [\*A large-scale analysis of racial disparities in police stops across the United States\*](#)

Caliskan et al. [\*Semantics Derived Automatically from Language Corpora Contain Human-like Biases\*](#)

Torralba & Efros. [\*Unbiased Look at Dataset Bias\*](#)

The above three papers are about very different domains, but in our discussion we will draw connections between them (and connections to our previous discussions). As you read these papers, think about their methods through the lens of data modeling vs. algorithmic modeling. Also keep track of the various notions of “bias” that you encounter.

Bakshy et al. [\*Exposure to ideologically diverse news and opinion on Facebook\*](#)  
[\*\[Supplementary material\]\*](#)

The subtext of this paper is Facebook researchers saying that if there’s a filter bubble on Facebook, it is primarily because of users’ friendships and choices rather than algorithmic effects.

Kay et al. [\*Unequal Representation and Gender Stereotypes in Image Search Results for Occupations\*](#)

### **Optional:**

Buolamwini. [\*Algorithms aren’t racist. Your skin is just too dark.\*](#) [Blog post]

### **Discussion prompts:**

Explain the equation at the top of page 5 of the Pierson et al. paper. How might the authors have come up with that model? Why are the coefficients in the exponent? What is overdispersion?

Consider the WEAT tests of Caliskan et al. In experiment design, if we want stronger evidence against the null hypothesis, we can try increasing our sample size to see if it yields a lower  $p$ -value. How would one do that in the context of WEAT?

Say you have a dataset of images of people. You wish to test how representative it is of minority groups, and whether it reflects cultural stereotypes. Could you adapt the techniques of Torralba & Efros to do so?

The Bakshy et al. study limits itself to Facebook users who chose to reveal their ideological affiliation. Consider two models to explain which users reveal their affiliation: (1) each user independently decides whether or not to reveal with the same probability (2) each user has a latent variable, “polarization”, that measures the strength of their partisan views. The probability of revelation depends on polarization; the composition of one’s friendships also depends on polarization (i.e., more polarized individuals are less likely to have friends with opposing political views). How would your interpretation of the study’s results change under these two models?

## **Module 6. Fairness mechanisms**

The study of techniques for achieving fairness is still nascent. We’ll look at methods that operate on different stages of the machine learning pipeline

Adjusting the input data:

*Feldman et al. [Certifying and Removing Disparate Impact](#)*

Selecting features that are intuitively considered fair (by crowd-workers):

*Grgić-Hlača et al. [The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making](#)*

Learning an intermediate representation:

*Zemel et al. [Learning Fair Representations](#)*

Adjusting a learned (unsupervised) representation:

*Bolukbasi et al. [Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings](#).*

Adjusting a learned predictor:

Hardt et al. [Equality of Opportunity in Supervised Learning](#) [Section 4]

(We omitted Section 4 in module 4. This is listed here for completeness, but the technique boils down to “use different thresholds for different groups.”)

Jointly adjusting a *set* of outputs of a structured predictor:

Zhao et al. [Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints](#)

This is the first paper looking at a structured prediction setting. Here bias is captured in the structure of the output, rather than in the relationship between the input and the output.

### **Optional reading**

I moved the following paper here from module 3; we already discussed it briefly in class.

Kilbertus et al. [Avoiding Discrimination through Causal Reasoning](#)

Another paper that we already discussed briefly in class. This paper takes a PL perspective and takes a white-box approach to analyzing the behavior of a learned predictor on a specified data distribution.

Albargouthi et al. [FairSquare: Probabilistic Verification of Program Fairness](#)

Creating a diverse NLP dataset.

Blodgett et al. [A Dataset and Classifier for Recognizing Social Media English](#)

### **Discussion prompts [due Monday 11/6]:**

Consider the “data repair” technique of Feldman et al. and the “learning prototypes” technique of Zemel et al. Both papers use statistical parity as the fairness notion (“lack of disparate impact” is basically “approximate statistical parity”). Could either of these techniques be adapted if we’re interested in one of the other group fairness notions (say, equal false positive rates)? If so, explain how. If not, why not?

Consider the following fairness mechanism: (1) Use linear regression to model the effect of the explanatory variables (including the protected attribute) on the target variable; (2) Drop the term corresponding to the protected variable. Does this method achieve fairness according to any of the fairness definitions we’ve studied? Would you use this method? Why or why not?

A major limitation of the Zhao et al. paper is that it requires a joint inference over the entire test set, which can be computationally infeasible. Further, it is problematic if your training and test distributions have different biases (as they often might in practice), or if there is concept drift (biases change over time). Can you think of ideas for speeding up the inference step and/or making it more flexible?

Discuss what (if anything) the group fairness impossibility theorem implies in the setting of Zhao et al. (and, more generally, for image classification).

The *fairness through awareness* paper (module 4) and the *learning fair representations* paper have something in common: they imagine a world where the decision-maker is not trusted, and compliance with the fairness definition is the only thing preventing discrimination or misbehavior (this is not a coincidence --- the papers also have authors in common). This is the reason why the fairness through awareness paper presents a set of seemingly bizarre ways in which statistical parity can go wrong (page 8). The learning fair representations paper is more explicit, and describes their world as: “a two-step system construction by two parties: an impartial party attempting to enforce fairness, and a vendor attempting to classify individuals”. An alternative philosophy is to assume that the decision-maker fundamentally has good intentions, and the fairness mechanism helps the decision-maker avoid unintentional discrimination. Which approach do you prefer and why?

## **Module 7. Interpretability**

Two foundational papers to understand interpretability:

Doshi-Velez and Kim, [\*Towards a Rigorous Science of Interpretable Machine Learning\*](#) [[Slides](#)]

Lipton, [\*The Mythos of Model Interpretability\*](#)

Four different approaches to interpretability:

1. Building simple models that perform almost as well as more complex ones:

Caruana et al, [\*Intelligible Models for Healthcare\*](#)

2. Finding the causal effects of inputs on predictions:

Datta et al, [\*Algorithmic Transparency via Quantitative Input Influence\*](#)

3. Learning an interpretable model locally to explain a specific prediction:

Ribeiro et al, [\*“Why Should I Trust You?” Explaining the Predictions of Any Classifier\*](#)

4. A visualization approach to explaining predictions of deep neural networks:

Selvaraju et al., [\*Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization\*](#)

## Optional readings:

An blog post that discusses some of the debates around interpretability:

*Bornstein. [Is Artificial Intelligence Permanently Inscrutable?](#)*

A technique to efficiently compute the impact of a single training example on a model's predictions, with several interpretability-related applications:

*Koh & Liang, [Understanding Black-box Predictions via Influence Functions](#)*

A philosophical and conceptual discussion of explanation, only tangentially related to machine learning.

*Miller, [Explanation in Artificial Intelligence: Insights from the Social Sciences](#)*

## Discussion prompts

Lipton complains that many papers declare one model to be more interpretable than another without any justification or a definition of interpretability. Find five papers that claim to build interpretable models. This could be a statement about one type of model being more interpretable than another, or a quantification of model complexity within a single family of models. You could start with some of the papers in this module. What, if any, justification do the authors provide for the claim of interpretability? Can you map the interpretability notions in the papers you found to some of the properties/desiderata in Lipton's paper?

Consider the following interpretability goal for image labeling: we want to be able to automatically generate textual explanations such as: "this image was assigned a negative sentiment score because it depicts a scene in which it is raining"; or "the 'singer-songwriter' label given to this image was affected by the gender of the subject" (recall that this is an [actual example](#) we discussed in class). Develop a research plan for tackling this problem. (Imagine you're writing a project proposal.)

## Module 8: Privacy & the ethics of inference

As machine learning technology (algorithms, training data, hardware) continues to improve, are we entering a world where "everything predicts everything", including sensitive attributes about people that they might not wish to be known? If such inferences are feasible, when are they ethical to make?

Two papers arguing that machine learning forces us to rethink how we've understood and regulated privacy:

Ohm & Peppet. [\*What if Everything Reveals Everything?\*](#)

Barocas and Nissenbaum. [\*Big Data's End Run around Procedural Privacy Protections\*](#)

The following three papers each aim to show the predictability of various sensitive attributes using machine learning, and each faced a backlash on ethical grounds. The latter two are recent and faced especially harsh criticism.

Kosinski, Stillwell, and Graepel. [\*Private Traits and Attributes Are Predictable From Digital Records of Human Behavior\*](#)

Wu & Zhang. [\*Automated Inference on Criminality using Face Images\*](#)

Wang & Kosinsky. [\*Deep neural networks are more accurate than humans at detecting sexual orientation from facial images\*](#)

### **Optional reading**

An NYT magazine article about how Target uses a data-driven approach to infer things like customer's pregnancies and uses that to influence their purchase habits:

Duhigg, [\*How Companies Learn Your Secrets\*](#)

A short, high-level paper on the connection between machine learning, privacy, and non-discrimination:

Horvitz and Mulligan, [\*Data, Privacy, and the Greater Good\*](#)

Controversy about (the validity of) a paper that claims to predict a person's face from their DNA:

Reardon, [\*Geneticists pan paper that claims to predict a person's face from their DNA\*](#)

### **Discussion prompts**

Imagine that you're called upon to advise policy makers about the effectiveness of machine learning at inferring sensitive attributes. In particular, your task is to develop a set of heuristics to anticipate the advances that are coming in the next decade or two. How accurate do you think

we'll get at various prediction tasks -- only slightly better than random, or almost perfect, or somewhere in between? For concreteness, you might want to think about the following tasks:

- Identifying faces in a crowd
- Generating an image of a person's face given their DNA
- Predicting 2-year recidivism risk (given everything observable about a person)
- Inferring a person's sexual orientation from photos of their face
- Inferring an author's gender based on text written by them

To simplify the discussion, assume that unlimited amounts of training data and hardware will be available.

As mentioned above, the Wu & Zhang paper and the Wang & Kosinski paper both attracted controversy. Please read up on these debates. (I'm deliberately avoiding suggesting specific links in the hope that different people will read different sets of critiques, which will help diversify the discussion.)

Now imagine you're called upon to advise policy makers about when inference-making crosses an ethical line. Develop a set of criteria for this purpose (not necessarily legal restrictions, but perhaps community norms for data scientists.) Does your standard differentiate between research and commercial applications? Apply your criteria to the three inference papers you read.

(I'm not looking for right or wrong answers; all answers are on the table, including "inference is never unethical". What matters is your reasoning.)

Separately, comment on the *validity* of the three papers' claims. Do you think they show what they claim to show?

## **Module 9: Transparency & accountability**

This module grapples with the problem of how outside observers can trust that a machine learning system is fair. One major thrust of research has been to treat the system as a black box and perform experiments on it to detect discrimination. The first several papers related to this idea.

Sandvig, Hamilton, Karahalios, and Langbort, [Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms](#)

Sweeney, [Discrimination in Online Ad Delivery](#)

Datta, Tschantz, and Datta, [Automated Experiments on Ad Privacy Settings](#)

Hannak, Soeller, Lazer, Mislove, and Wilson, [Measuring Price Discrimination and Steering on E-commerce Web Sites](#)

This paper critiques the idea that “looking inside the black box” is an effective way to govern algorithmic systems:

Ananny & Crawford, [Seeing without knowing Limitations of the transparency ideal and its application to algorithmic accountability](#)

### **Optional reading**

A fascinating alternative proposal for algorithmic accountability was made by Princeton researchers (Josh Kroll, Ed Felten, and others). It involves cryptographic techniques to prove compliance with specified behavior without having to reveal sensitive data. There isn't a publicly available technical paper describing the idea, but it's described at a high level in a law review paper and in more technical detail in Kroll's thesis:

Joshua Kroll et al., [Accountable algorithms](#)

Joshua Kroll, [Accountable algorithms](#)

### **Discussion prompts**

1. Post two papers that are relevant to the topic of this course, but weren't included in the readings. Summarize each in a paragraph, connecting them to the themes we've discussed.
2. The major technical content of this module relates to experimental design, and the paper that addresses it most directly is AdFisher. So the following questions are based on that paper.
  - A. One common assumption in experimental design is that the treatment does not affect the controls. Give two reasons why this assumption might be violated in the context of AdFisher.
  - B. Imagine the following response from Google, addressing the findings in Section 6.1. During the time of the AdFisher experiments, a number of A/B tests related to ads were in progress. In some of these tests, users in bucket “A” were shown ads related to high-paying jobs and users in bucket “B” were shown ads related to relatively low-paying jobs. The authors' measurements at

least partly reflect the impact of these A/B tests. By not mentioning this possibility, the paper is invalid, or at least incomplete.

How might the authors respond to this objection?

C. Describe a scenario which would result in the blocked design described in the paper having a greater power than a non-blocked design in which each agent is independently randomly assigned to a group. (A blocked design, on the other hand, ensures that in each block there are an equal number of agents from each group.) For example, the scenario might be based on ad churn, as suggested in the paper. Explain why the blocked design would have higher power.

D. Describe how you would adapt AdFisher to the question studied by Sweeney. Would machine learning still be applicable or would you manually define a test statistic?