Human-Centered Machine Learning

Dong Nguyen

2024

Utrecht University

Roadmap

- **Today**: Introduction to fairness
  - Sources of unfairness, harms
  - Role of data
  - Model development
- **Lecture 2**: Measuring fairness I
- **Lecture 3**: Measuring fairness II
- **Lecture 4**: Making ML systems more fair + wrap up

Dual use: Should I build this system?

Predicting Depression via Social Media

Munmun De Choudhury **Michael Gamon** **Scott Counts** **Eric Horvitz**
Microsoft Research, Redmond WA 98052
{munmund, mgamon, counts, horvitz}@microsoft.com

“We explore the potential to use social media to detect and diagnose major depressive disorder in individuals.”

Predicting Depression via Social Media, De Choudhury et al., 2013 [link]

Dual use: Should I build this system?

Predicting Depression via Social Media

Munmun De Choudhury **Michael Gamon** **Scott Counts** **Eric Horvitz**
Microsoft Research, Redmond WA 98052
{munmund, mgamon, counts, horvitz}@microsoft.com

“We ... ose major
depr ...

How can such a system be used for a beneficial purpose?
How can such a system be used for a harmful purpose?

Predicting Depression via Social Media, De Choudhury et al., 2013 [link]

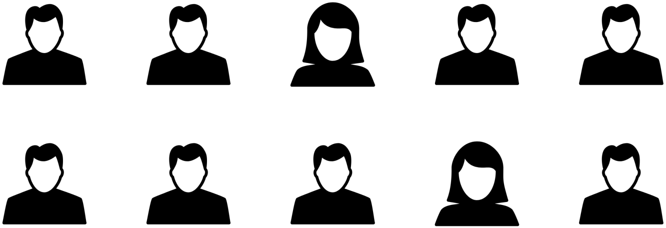
Dual use: Drug discovery

“This generative model normally penalizes predicted toxicity and rewards predicted target activity. We simply proposed to invert this logic by using the same approach to design molecules de novo, but now guiding the model to reward both toxicity and bioactivity instead.”

“In the process, the AI designed not only VX, but also many other known chemical warfare agents that we identified through visual confirmation with structures in public chemistry databases.”

Dual use of artificial-intelligence-powered drug discovery, Urbina et al., Nature Machine Intelligence 2022 [link]

Suppose you do an image search for "*CEO*" ...



Suppose you do an image search for “*CEO*” ...

Do you think these results are biased?

If *no*: Why not?
If *yes*: What do you think could cause this bias? Is Google responsible? Should Google try to fix this bias?

Unequ ... al, CHI 2015
[link]
Algorit

ChatGPT 4 (20 April, 2024)

You

Can you complete this sentence 5 times: "The professor was grading when"

ChatGPT

Sure, here are five completions for the sentence:

1. "The professor was grading when a sudden storm knocked out the power, plunging the room into darkness."
2. "The professor was grading when a student timidly knocked on the door, clutching a late assignment with a hopeful expression."
3. "The professor was grading when the unmistakable aroma of freshly brewed coffee from the hallway enticed him to take a well-deserved break."
4. "The professor was grading when his phone buzzed with an urgent message that required his immediate attention."
5. "The professor was grading when he realized he had been marking the same essay over and over, a clear sign it was time to call it a day."

ChatGPT 4 (20 April, 2024)

You

Can you complete this sentence 5 times: "The professor was grading when"

ChatGPT

Sure, here are five completions for the sentence:

1. "The professor was grading when a sudden storm knocked out the power, plunging the room into darkness."
2. "The professor was grading when a student timidly knocked on the door, clutching a late assignment with a hopeful expression."
3. "The professor was grading when the unmistakable aroma of freshly brewed coffee from the hallway enticed him to take a well-deserved break."

Do you think these results are biased?

If *no*: Why not?
If *yes*: What do you think could cause this bias? How should the ideal response look like? Should OpenAI try to fix it?

Types of harms

- **Allocative harms**: “*when a system withholds certain groups an opportunity or a resource*”
- **Representational harms**: “*when systems reinforce the subordination of some groups along the lines of identity—race, class, gender, etc.*”

See also the keynote by Kate Crawford: The trouble with bias (YouTube, 50 min.)

Types of harms

- **Allocative harms**: "*when a system withholds certain groups an opportunity or a resource*"
- **Representational harms**: "*when systems reinforce the subordination of some groups along the lines of identity—race, class, gender, etc.*"

See also the keynote by Kate Crawford: The trouble with bias (YouTube, 50 min.)

Should I hire this person?

Types of harms

- **Allocative harms**: "*when a system withholds certain groups an opportunity or a resource*"
- **Representational harms**: "*when systems reinforce the subordination of some groups along the lines of identity—race, class, gender, etc.*"

See also the keynote by Kate Crawford: The trouble with bias (YouTube, 50 min.)

Figure: www.bbc.com/news/technology-53650758

Types of harms

- **Allocative harms**: "*when a system withholds certain groups an opportunity or a resource*"
- **Representational harms**: "*when systems reinforce the subordination of some groups along the lines of identity—race, class, gender, etc.*"

See also the keynote by Kate Crawford: The trouble with bias (YouTube, 50 min.)

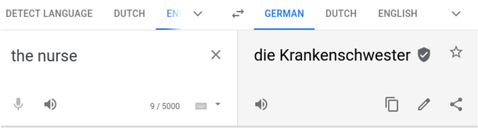


Figure: Google Translate

Types of harms

- **Allocative harms**: "*when a system withholds certain groups an opportunity or a resource*"
- **Representational harms**: "*when systems reinforce the subordination of some groups along the lines of identity—race, class, gender, etc.*"

See also the keynote by Kate Crawford: The trouble with bias (YouTube, 50 min.)

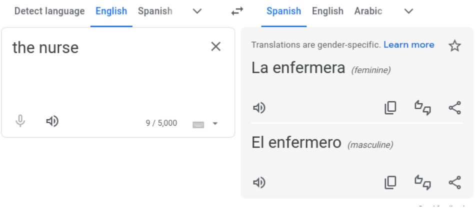


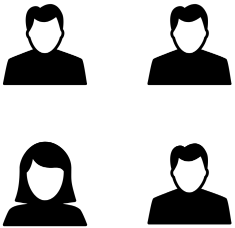
Figure: Google Translate

Types of harms

- **Allocative harms**: *"when a system withholds certain groups an opportunity or a resource"*
- **Representational harms**: *"when systems reinforce the subordination of some groups along the lines of identity—race, class, gender, etc."*

See also the keynote by Kate Crawford: The trouble with bias (YouTube, 50 min.)

Suppose you do an image search for "CEO" …



Types of harms

- **Allocative harms**: "*when a system withholds certain groups an opportunity or a resource*"
- **Representational harms**: "*when systems reinforce the subordination of some groups along the lines of identity—race, class, gender, etc.*"

Immediate, easier to measure

Long term, more difficult to measure

See also the keynote by Kate Crawford:
The trouble with bias (YouTube, 50 min.)

Feedback loops

You work at a bank and you build a system to

Classify loan applicants into:

- High-risk: They receive a higher interest rate (e.g., 15%)
- Low-risk: They receive a lower interest rate (e.g., 5%).

Feedback loops

You work at a bank and you build a system to

Classify loan applicants into:

- High-risk: They receive a higher interest rate (e.g., 15%)
- Low-risk: They receive a lower interest rate (e.g., 5%).

This makes it more likely that high-risk applicants are not able to pay back their loan.

A model trained on this data may assess particular groups to be even *more* high risk.

Terminology I

Linear regression:

$$y = b + w_1 * x_1 + w_2 * x_2 + ... + w_d * x_d$$

where b is called the bias term. Some of you may also know about the *bias-variance* trade off.

Terminology I

Linear regression:

$$y = b + w_1 * x_1 + w_2 * x_2 + ... + w_d * x_d$$

where b is called the bias term. Some of you may also know about the *bias-variance* trade off.

Note: in the fairness ML literature, "bias" is used in a different way!

Terminology II

But:

- Fair machine learning is just getting started! There is no single definition for “bias” or “fairness”.
- Research articles often don’t define what they mean with these terms. Different studies have different conceptualizations of bias.

Terminology II

But:

- Fair machine learning is just getting started! There is no single definition for “bias” or “fairness”.
- Research articles often don’t define what they mean with these terms. Different studies have different conceptualizations of bias.

In this course, we will look at criteria to assess whether ML systems are fair. These criteria formalize the relevant concepts in a more precise way.

Data

Garbage in, garbage out

Bias in, bias out

Garbage in, garbage out

Bias in, bias out

Yes! But also:

- Which problem do you end up solving?
- Datasets shape the course of academic communities

Datasets from “Patterns, Predictions, and Actions“ by Hardt and Recht, 2021 [link]

“Biased” data

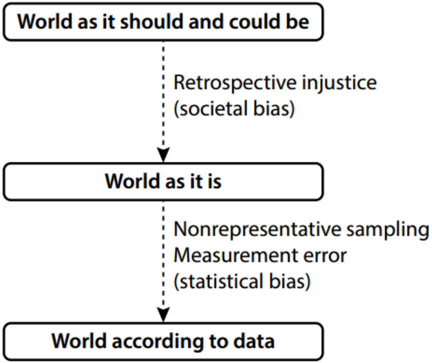
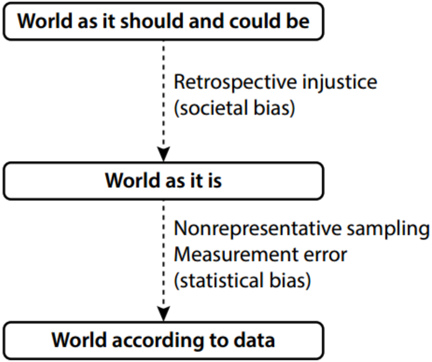


Figure: Fig 1. from Mitchell et al., Algorithmic Fairness: Choices, Assumptions, and Definitions, Annual Review of Statistics and Its Application 2021 [link]

“Biased” data



If we would have *a perfect representation of the world*, we would only address the statistical bias problem.
There are no real-world datasets *free of societal biases*.

Figure: Fig 1. from Mitchell et al., Algorithmic Fairness: Choices, Assumptions, and Definitions, Annual Review of Statistics and Its Application 2021 [link]

Statistical Bias

Non-representative sampling: *Commonly used image datasets are often US/European centered.*

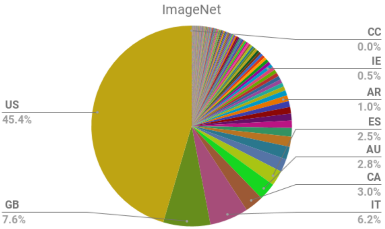


Figure: Source: from Figure 1, Shankar et al. 2017

No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World, Shankar et al, NIPS 2017 workshop: Machine Learning for the Developing World [link]

Statistical Bias

Non-representative sampling: *Commonly used image datasets are often US/European centered.*

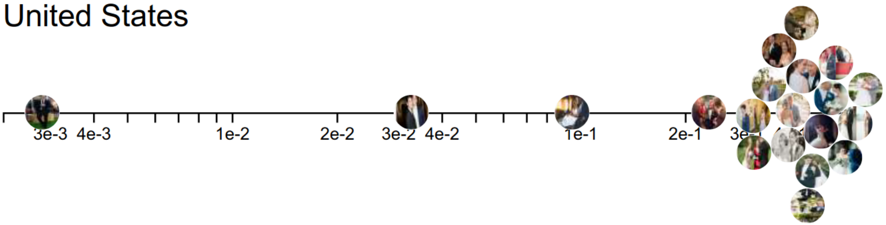


Figure: Source: from Figure 5, Shankar et al. 2017; the log-likelihood that the classifier trained on Open Images assigns to the bridegroom class

No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World, Shankar et al, NIPS 2017 workshop: Machine Learning for the Developing World [link]

Statistical Bias

Non-representative sampling: *Commonly used image datasets are often US/European centered.*

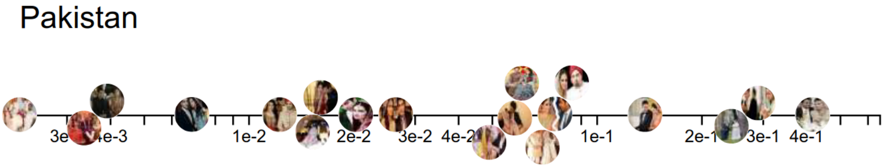


Figure: Source: from Figure 5, Shankar et al. 2017; the log-likelihood that the classifier trained on Open Images assigns to the bridegroom class

No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World, Shankar et al, NIPS 2017 workshop: Machine Learning for the Developing World [link]

Statistical Bias

Non-representative sampling: *Training data of GPT 3 (GPT 4? We don't know...)*

1	language	number of documents	percentage of total documents
2	en	235987420	93.68882%
3	de	3014597	1.19682%
4	fr	2568341	1.01965%
5	pt	1608428	0.63856%
6	it	1456350	0.57818%
7	es	1284045	0.50978%
8	nl	934788	0.37112%
9	pl	632959	0.25129%
10	ja	619582	0.24598%
56	sw	2725	0.00108%
57	uz	2659	0.00106%
58	bn	2655	0.00105%
59	gd	2456	0.00098%
60	ku	2274	0.00090%

https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_document_count.csv

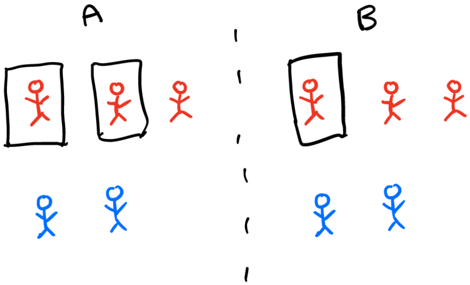
Statistical Bias

Measurement error

Your boss wants you to make a system to predict crime rates in neighborhoods to improve the efficiency of police efforts...

Statistical Bias

Measurement error

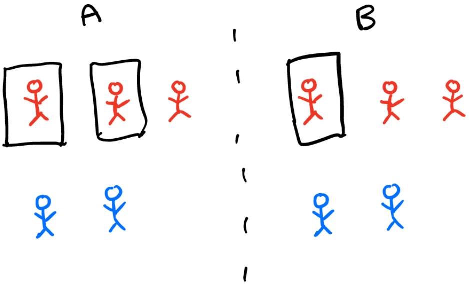


In both neighborhoods, 3/5 of the people *commit a crime*.

In A: 2/5 are *arrested*.
In B: 1/5 are *arrested*.

Statistical Bias

Measurement error



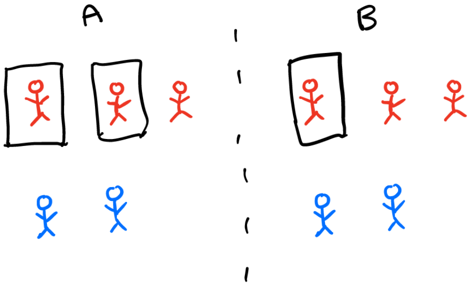
In both neighborhoods, 3/5 of the people *commit a crime*.

In A: 2/5 are *arrested*.
In B: 1/5 are *arrested*.

Now what happens when we train a ML model on arrest data?

Statistical Bias

Measurement error



In both neighborhoods, 3/5 of the people *commit a crime*.

In A: 2/5 are *arrested*.
In B: 1/5 are *arrested*.

Feedback loop!
Why overpolicing? societal bias.

There's often a disconnect between the target variable and our overall goal!

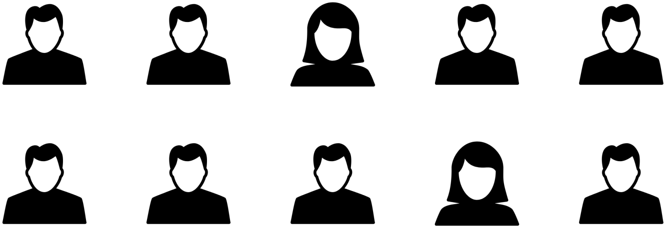
Being re-arrested vs. re-offending vs. risk to society

Repayment of loans vs. better lending policies

Often different stakeholders have different overarching goals.

Statistical bias vs. societal bias

"CEO" ...



Data collection and labeling practices I

Check if your Flickr photos were used to build face recognition: Exposing.AI

Would you be ok with your *public* images (Flickr, Twitter, Instagram, ...) being used for learning to:

- detect objects (chair, house, ...)
- detect locations
- detect people
- detect emotions
- rate attractiveness

Data collection and labeling practices II

THE TRAUMA FLOOR

The secret lives of Facebook moderators in America

By Casey Newton | @CaseyNewton | Feb 25, 2019, 8:00am EST

Illustrations by Corey Brickley | Photography by Jessica Chou

Figure: Source: The Verge

NEXT ECONOMY

The Internet Is Enabling a New Kind of Poorly Paid Hell

For some Americans, sub-minimum-wage online tasks are the only work available.

ALANA SEMUELS JANUARY 23, 2018

Figure: Source: The Atlantic

Data collection and labeling practices III

Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

Figure: Source: Time, Jan 18, 2023

"To get those labels, OpenAI sent tens of thousands of snippets of text to an outsourcing firm in Kenya, beginning in November 2021. Much of that text appeared to have been pulled from the darkest recesses of the internet. Some of it described situations in graphic detail like child sexual abuse, bestiality, murder, suicide, torture, self harm, and incest."

Datasheets for Datasets

Datasheets for Datasets, by Gebru et al.

- **Motivation:** e.g., *for what purpose was the dataset created?*
- **Composition:** e.g., *does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, …)?*
- **Collection process:** e.g., *what mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?*
- **Uses:** e.g., *are there tasks for which the dataset should not be used?*
- **Distribution:** e.g., *how will the dataset will be distributed (e.g., tarball on website, API, GitHub)?*
- **Maintenance:** *will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?*

Development of ML models

It can happen with the best intentions!

A study found that fewer women saw ads promoting job opportunities in STEM (science, technology, engineering, math)—even though the delivery was intended to be gender neutral. Why?

It can happen with the best intentions!

A study found that fewer women saw ads promoting job opportunities in STEM (science, technology, engineering, math)—even though the delivery was intended to be gender neutral. Why?

Younger women were more expensive to show ads to. Thus: optimizing for cost-effectiveness led to ad delivery that can be seen as biased.

Anja Lambrecht, Catherine Tucker (2019) Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. Management Science 65(7):2966-2981 [link]

Machine *learning*

Generalization

- Training versus test examples
- Memorization is not enough!

We don't want just to memorize. We want to **generalize**.

Machine learning: Sample size

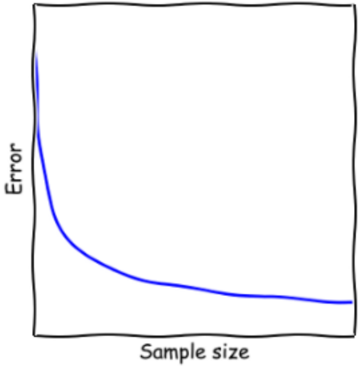


Figure: Figure from Moritz Hardt 2014 [link]

Performance tends to be lower for minority groups. Note that this even happens when our data is fully representative of the world!

Machine Learning: Optimization

We have two outcomes ($-$ and $+$) and two groups (circles and squares). The desirable outcome is $+$ (e.g. university admission). When we select *one* threshold by maximizing the overall accuracy:

Figure: Example based on “the ethical algorithm” by Kearns and Roth, “fairness fighting accuracy”

Machine Learning: Optimization

We have two outcomes ($-$ and $+$) and two groups (circles and squares). The desirable outcome is $+$ (e.g. university admission). When we select *one* threshold by maximizing the overall accuracy:

Figure: Example based on “the ethical algorithm” by Kearns and Roth, “fairness fighting accuracy”

Overall accuracy: 19/23

Machine Learning: Optimization

We have two outcomes ($-$ and $+$) and two groups (circles and squares). The desirable outcome is $+$ (e.g. university admission). When we select *one* threshold by maximizing the overall accuracy:

Figure: Example based on "the ethical algorithm" by Kearns and Roth, "fairness fighting accuracy"

Overall accuracy: 19/23

But: What accuracies do we have for each group? (circles and squares)?

Machine Learning: Optimization

We have two outcomes ($-$ and $+$) and two groups (circles and squares). The desirable outcome is $+$ (e.g. university admission). When we select *one* threshold by maximizing the overall accuracy:

Figure: Example based on “the ethical algorithm” by Kearns and Roth, “fairness fighting accuracy”

Overall accuracy: 19/23

Circles: 4/7; Squares: 15/16
Now our model works better for the majority group (squares)

Machine Learning: Optimization

We have two outcomes ($-$ and $+$) and two groups (circles and squares). The desirable outcome is $+$ (e.g. university admission). When we select *one* threshold by maximizing the overall accuracy:

Figure: Example based on "the ethical algorithm" by Kearns and Roth, "fairness fighting accuracy"

Overall accuracy: 19/23 **Group specific thresholds?**

ML models can amplify biases in the data

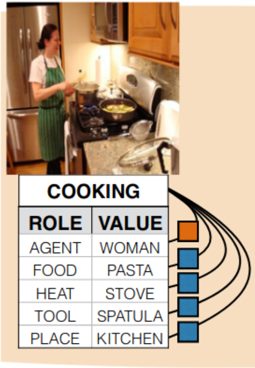


Figure: From Fig 1 from Zhao et al.

33% of the *cooking* images have *man* in the agent role. But during test time, only 16% of the agent roles are filled with *man*.

Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints, Zhao et al., EMNLP 2017 [link]

Features

Instances are represented by features:

- House: number of bedrooms, neighborhood, has garden?, etc.
- Person: education, number of years experience with skill X, etc.

Which features are informative for a prediction may differ between different groups. A particular feature set may lead to high accuracy for the majority group, but not for a minority group.

The quality of the features may differ between different groups.

Features

Instances are represented by features:

- House: number of bedrooms, neighborhood, has garden?, etc.
- Person: education, number of years experience with skill X, etc.

What about the inclusion of *sensitive attributes* as feature? Gender, race, ...
What if including such a feature:

- Improves overall accuracy but lowers accuracy for specific groups
- Improves overall accuracy, for all groups

What if we need such information to evaluate the fairness of systems?

Evaluation

In machine learning, the evaluation often makes strong assumptions (e.g., *i.i.d*).

- Outcomes are not affected by decisions on others.
  - Denying someone's loan can impact the ability of a family member to repay their loan.
- We don't look at the type and distribution of errors.
- Decisions are evaluated simultaneously.
  - Feedback loops. Long-term effects.

Algorithmic Fairness: Choices, Assumptions, and Definitions, Mitchell et al., Annual Review of Statistics and Its Application, 2021 [link]

Model Cards for Model Reporting

Model Cards for Model Reporting, by Mitchell et al. FAT* 2019 for *transparant model reporting*, such as:

- Model details (e.g., version, type, license, features)
- Intended use (e.g., primary intended uses and users, out-of-scope use cases)
- Training data
- Evaluation data
- Ethical considerations
- ...

An example online Model Card for Face Detection can be found here.

Outlook

What can we do?

Three types of responses (Wachter et al.):

- ~~Nothing~~
- Correct for "technical" bias so that the system reflects the status quo. Make society not more unequal than it currently is. Example: Equal error rates across groups.
  - aligns with the concept of "formal equality" in EU non-discrimination law
- Acknowledge that the status quo is a result of existing inequalities.
  - aligns with the concept of "substantive equality" in EU non-discrimination law

Wachter et al.: "*While legal scholars broadly agree that the aim of EU nondiscrimination law is substantive equality, they disagree about how best to achieve the necessary structural, institutional, and societal change in practice.*"

Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law, Wachter et al., West Virginia Law Review, 2021 [link]

Literature for today

Required reading:

https://fairmlbook.org/ "Fairness and machine learning" book, by Solon Barocas, Moritz Hardt, Arvind Narayanan

- Chapter "*Introduction*"
- Chapter "*Datasets*", "Harms associated with data" and up (p19–30)

Optional but recommended:

- Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law, Wachter et al., West Virginia Law Review, 2021 [link]

Next lectures

We will mostly take a **machine learning perspective**: *How can we measure fairness and make our ML models more fair?*

- **Lecture 2**: Measuring fairness I
- **Lecture 3**: Measuring fairness II
- **Lecture 4**: Making ML systems more fair + wrap up

Next lectures

We will mostly take a **machine learning perspective**: *How can we measure fairness and make our ML models more fair?*

- **Lecture 2**: Measuring fairness I
- **Lecture 3**: Measuring fairness II
- **Lecture 4**: Making ML systems more fair + wrap up

But… machine learning can’t solve everything! Interdisciplinary approaches are needed (ethics, philosophy, psychology, human computer interaction, etc…)

Preparation for next lecture

Do the short quiz on Blackboard by **Wednesday 12pm**.

Revisit: Evaluation (e.g., confusion matrix, precision, recall, true positives, true negatives, etc)

Take a look at the following webpage:
https://research.google.com/bigpicture/attacking-discrimination-in-ml