

# Machine Learning in Production Building Fair Products



DIRECT STAKEHOLDERS  
Product end users, administrators, manufacturers.

INDIRECT STAKEHOLDERS  
Do not use the product but are included advocacy groups (ACLU, vulnerable communities, regula

EXCLUDED STAKEHOLDERS  
Can't use the product potentially, social, or situational constraints, product may exclude people with experience would exclude those

Facilitator decides which stakeholder reviews, then writes to players

# Administrativa

Homework I4 to be released next week; 1 week assignment, due Nov 25



# Research in this Course (in I4)

We are conducting academic research on explainability policies and evidence. This research will involve analyzing student work of this assignment. You will not be asked to do anything above and beyond the normal learning activities and assignments that are part of this course. You are free not to participate in this research, and your participation will have no influence on your grade for this course or your academic career at CMU. If you do not wish to participate, please send an email to Nadia Nahar ([nadian@andrew.cmu.edu](mailto:nadian@andrew.cmu.edu)). Participants will not receive any compensation or extra credit. The data collected as part of this research will not include student grades. All analyses of data from participants' coursework will be conducted after the course is over and final grades are submitted -- instructors do not know who chooses not to participate before final grades are submitted. All data will be analyzed in de-identified form and presented in the aggregate, without any personal identifiers. The de-identified homework solutions (without grades) will be shared with research collaborators at Yale university for further analysis. If you have questions pertaining to your rights as a research participant, or to report concerns to this study, please contact Nadia Nahar ([nadian@andrew.cmu.edu](mailto:nadian@andrew.cmu.edu)) or the Office of Research Integrity and Compliance at Carnegie Mellon University ([irb-review@andrew.cmu.edu](mailto:irb-review@andrew.cmu.edu); phone: 412-268-4721).

# From Fairness Concepts to Fair Products

## Fundamentals of Engineering AI-Enabled Systems

**Holistic system view:** AI and non-AI components, pipelines, stakeholders, environment interactions, feedback loops

### Requirements:

System and model goals  
User requirements  
Environment assumptions  
Quality beyond accuracy  
Measurement  
Risk analysis  
Planning for mistakes

### Architecture + design:

Modeling tradeoffs  
Deployment architecture  
Data science pipelines  
Telemetry, monitoring  
Anticipating evolution  
Big data processing  
Human-AI design

### Quality assurance:

Model testing  
Data quality  
QA automation  
Testing in production  
Infrastructure quality  
Debugging

### Operations:

Continuous deployment  
Contin. experimentation  
Configuration mgmt.  
Monitoring  
Versioning  
Big data  
DevOps, MLOps

**Teams and process:** Data science vs software eng. workflows, interdisciplinary teams, collaboration points, technical debt

## Responsible AI Engineering

Provenance,  
versioning,  
reproducibility

Safety

Security and  
privacy

Fairness

Interpretability  
and explainability

Transparency  
and trust

Ethics, governance, regulation, compliance, organizational culture

# Reading

Required reading:

- Holstein, Kenneth, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. "[Improving fairness in machine learning systems: What do industry practitioners need?](#)" In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1-16. 2019.

Recommended reading:

- Metcalf, Jacob, and Emanuel Moss. "[Owning ethics: Corporate logics, silicon valley, and the institutionalization of ethics.](#)" *Social Research: An International Quarterly* 86, no. 2 (2019): 449-476.

# Learning Goals

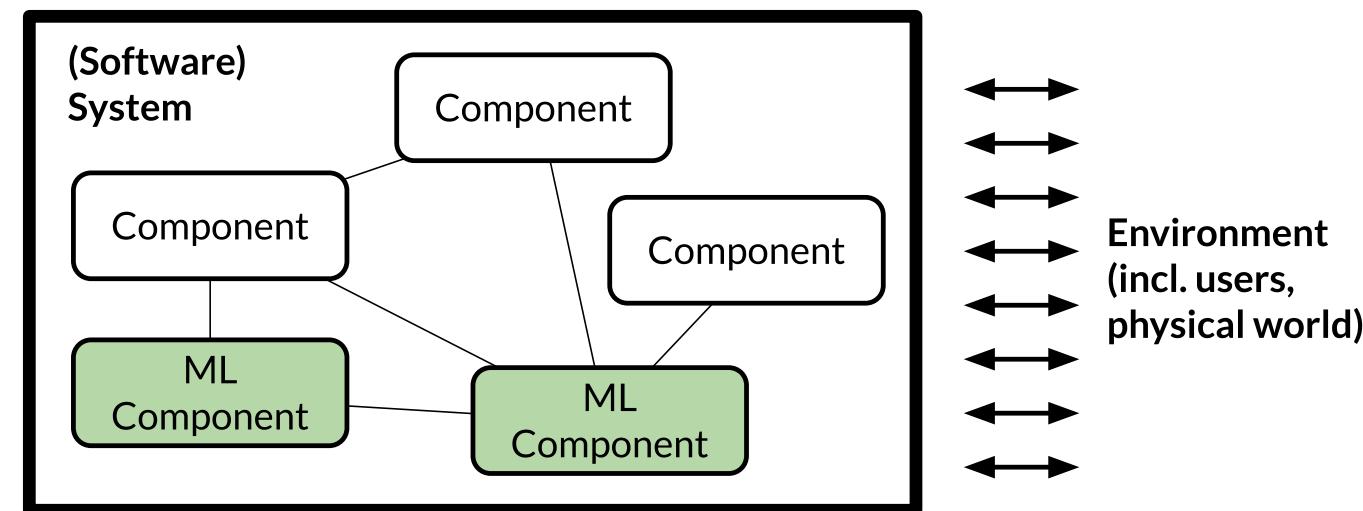
- Understand the role of requirements engineering in selecting ML fairness criteria
- Understand the process of constructing datasets for fairness
- Document models and datasets to communicate fairness concerns
- Consider the potential impact of feedback loops on AI-based systems and need for continuous monitoring
- Consider achieving fairness in AI-based systems as an activity throughout the entire development cycle

# Today: Fairness as a System Quality

Fairness can be measured for a model

... but we really care whether the system, as it interacts with the environment, is fair/safe/secure

... does the system cause harm?



# Fair ML Pipeline Process

Fairness must be considered throughout the entire lifecycle!



# Fairness Problems are System-Wide Challenges

- **Requirements engineering challenges:** How to identify fairness concerns, fairness metric, design data collection and labeling
- **Human-computer-interaction design challenges:** How to present results to users, fairly collect data from users, design mitigations
- **Quality assurance challenges:** Evaluate the entire system for fairness, continuously assure in production
- **Process integration challenges:** Incorporate fairness work in development process
- **Education and documentation challenges:** Create awareness, foster interdisciplinary collaboration

# Understanding System-Level Goals for Fairness

i.e., Requirements engineering

# Recall: Fairness metrics

- Anti-classification (fairness through blindness)
- Group fairness (independence)
- Equalized odds (separation)
- ...and numerous others and variations!

**But which one makes most sense for my product?**

# Recall Breakout: Cancer Prognosis

		Male Patient Results		Female Patient Results	
		Actual cancer	Actually no cancer		
		Predicted cancer	23	11	Predicted cancer
Predicted no cancer	Actual cancer	41	925	Predicted no cancer	2
	Actually no cancer				480

- Does the model meet anti-classification fairness? *Can't tell*
- Does the model meet group fairness? *Sure*
  - $P[\text{cancer}, M] = 0.034$ ,  $P[\text{cancer}, F] = 0.036$
- Does the model meet equalized odds? *No*
  - $\text{FPR}, M = 0.012$ ,  $F = 0.010$
  - $\text{FNR}, M = 0.64$ ,  $F = 0.133$
- Is the model fair enough to use? How can we decide?

## Speaker notes

prob cancer male vs female



# Different "fairness" are...not compatible

As we will see, different fairness metrics reflect different fairness goals, different user groups that we care, and different trade-offs.

- **Equalized odds -- minimize disparate treatment.** "We should just focus on making the right decisions; We are fair by treating everybody equally and making sure everyone gets the opportunity they actually deserve." (equality)
- **Group fairness -- minimize disparate impact.** "We should carefully massage our decisions (and maybe be more lenient with giving one group more opportunity) to make sure that we give all the groups the same final outcomes." (equity)

# Identifying Fairness Goals is a Requirements Engineering Problem

# Identifying Fairness Goals is a Requirements Engineering Problem

- What is the goal of the system? What benefits does it provide and to whom?

# Identifying Fairness Goals is a Requirements Engineering Problem

- What is the goal of the system? What benefits does it provide and to whom?
- Who are the stakeholders of the system? What are the stakeholders' views or expectations on fairness and where do they conflict? Are we trying to achieve fairness based on equality or equity?

# Identifying Fairness Goals is a Requirements Engineering Problem

- What is the goal of the system? What benefits does it provide and to whom?
- Who are the stakeholders of the system? What are the stakeholders' views or expectations on fairness and where do they conflict? Are we trying to achieve fairness based on equality or equity?
- What subpopulations (including minority groups) may be using or be affected by the system? What types of harms can the system cause with discrimination?

# Identifying Fairness Goals is a Requirements Engineering Problem

- What is the goal of the system? What benefits does it provide and to whom?
- Who are the stakeholders of the system? What are the stakeholders' views or expectations on fairness and where do they conflict? Are we trying to achieve fairness based on equality or equity?
- What subpopulations (including minority groups) may be using or be affected by the system? What types of harms can the system cause with discrimination?
- Does fairness undermine any other goals of the system (e.g., accuracy, profits, time to release)?

# Identifying Fairness Goals is a Requirements Engineering Problem

- What is the goal of the system? What benefits does it provide and to whom?
- Who are the stakeholders of the system? What are the stakeholders' views or expectations on fairness and where do they conflict? Are we trying to achieve fairness based on equality or equity?
- What subpopulations (including minority groups) may be using or be affected by the system? What types of harms can the system cause with discrimination?
- Does fairness undermine any other goals of the system (e.g., accuracy, profits, time to release)?
- Are there legal anti-discrimination requirements to consider? Are there societal expectations about ethics w.r.t. to this product? What is the activist position?

# Identifying Fairness Goals is a Requirements Engineering Problem

- What is the goal of the system? What benefits does it provide and to whom?
- Who are the stakeholders of the system? What are the stakeholders' views or expectations on fairness and where do they conflict? Are we trying to achieve fairness based on equality or equity?
- What subpopulations (including minority groups) may be using or be affected by the system? What types of harms can the system cause with discrimination?
- Does fairness undermine any other goals of the system (e.g., accuracy, profits, time to release)?
- Are there legal anti-discrimination requirements to consider? Are there societal expectations about ethics w.r.t. to this product? What is the activist position?
- ...

# 1. Identify Protected Attributes

Against which groups might we discriminate? What attributes identify them directly or indirectly? **Requires understanding of target population and subpopulations!**

# Protected attributes are not always obvious



# 1. Identify Protected Attributes

Against which groups might we discriminate? What attributes identify them directly or indirectly? **Requires understanding of target population and subpopulations!**

- Use anti-discrimination law as starting point, but do not end there
  - Socio-economic status? Body height? Weight? Hair style? Eye color? Sports team preferences?
  - Protected attributes for non-humans? Animals, inanimate objects?
- Involve stakeholders, consult lawyers, read research, ask experts, ...

# 2. Analyze Potential Harms

## 1. Anticipate harms from unfair decisions

- Harms of allocation, harms of representation?
- How do biased model predictions contribute to system behavior?

## 2. Consider how automation can amplify harm

## 3. Overcome blind spots within teams

- Systematically consider consequences of bias
- Consider safety engineering techniques (e.g., FTA)
- Assemble diverse teams, use personas, crowdsource audits

# Example: Judgment Call Game

Card "Game" by Microsoft Research

Participants write "Product reviews" from different perspectives

- encourage thinking about consequences
- enforce persona-like role taking



# Example: Judgment Call Game



Judgment Call the Game: Using Value Sensitive Design and Design Fiction to Surface Ethical Concerns Related to Technology

### 3. Negotiate Fairness Goals/Measures

- Negotiate with stakeholders to determine fairness requirement for the product: What is the suitable notion of fairness for the product? Equality or equity?
- Map the requirements to model-level (model) specifications: Anti-classification? Group fairness? Equalized odds?
- Negotiation can be challenging!
  - Conflicts with other system goals (accuracy, profits...)
  - Conflicts among different beliefs, values, political views, etc.,

# Intuitive Justice: Research on what people perceive as fair (psychology)

- When rewards depend on inputs and participants can choose contributions:  
Most people find it fair to split rewards proportional to inputs
- Most people agree that for a decision to be fair, personal characteristics that do not influence the reward, such as gender or age, should not be considered when dividing the rewards.

Complexity: Individual and group differences not always clearly attributable, unequal starting positions

# Dealing with unequal starting positions

## Equality (minimize disparate treatment):

- Treat everybody equally, regardless of starting position
- Focus on meritocracy, strive for fair opportunities
- Equalized-odds-style fairness; equality of opportunity

## Equity (minimize disparate impact):

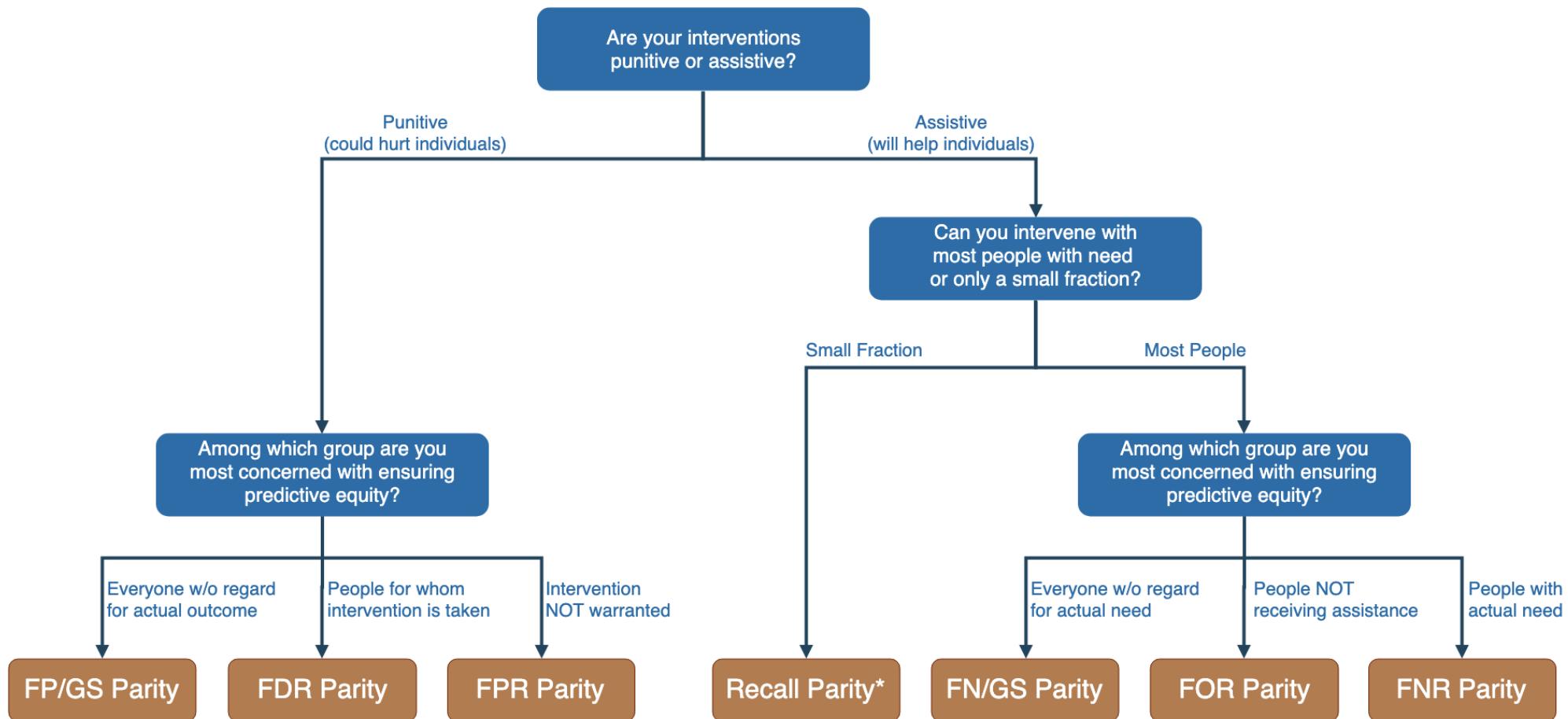
- Lift disadvantaged group, affirmative action
- Strive for similar outcomes (distributive justice)
- Group-fairness-style fairness; equality of outcomes

Each rooted in long history of law/philosophy, and typically incompatible. Problem and goal dependent

# Heuristic: Punitive vs Assistive Decisions

- If the decision is **punitive** in nature:
  - Harm is caused when a group is given an unwarranted penalty
  - e.g. decide whom to deny bail based on risk of recidivism
  - Heuristic: Use a fairness metric (equalized odds) based on **FPR**
- If the decision is **assistive** in nature:
  - Harm is caused when a group in need is denied assistance
  - e.g., decide who should receive a loan or a food subsidy
  - Heuristic: Use a fairness metric based on **FNR**

# Fairness Tree



Ian Foster, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter and Julia Lane. [Big Data and Social Science: Data Science Methods and Tools for Research and Practice](#). Chapter 11, 2nd ed, 2020

# Trade-offs in Fairness vs Accuracy

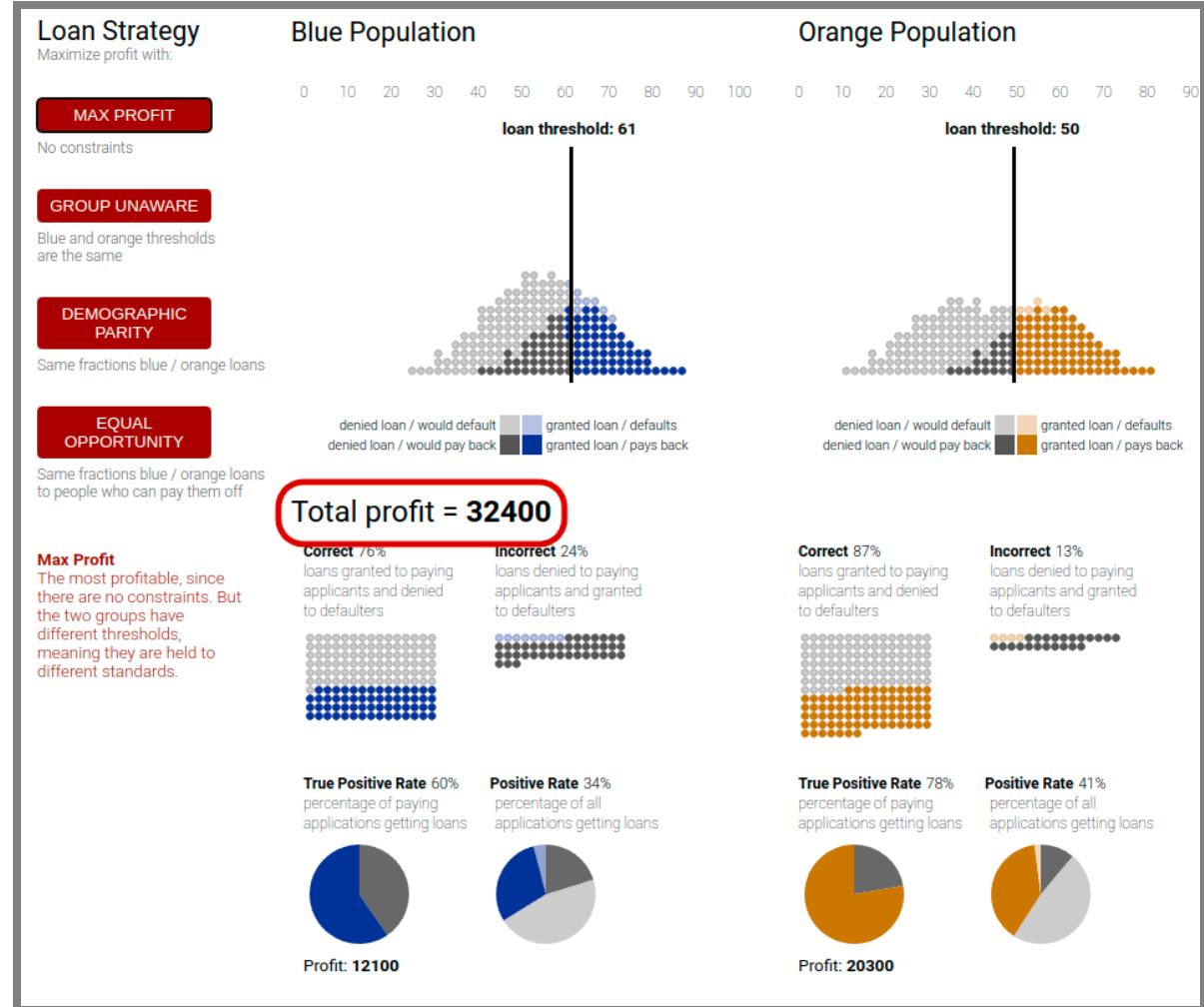


Fairness imposes constraints, limits what models can be learned

**But:** Arguably, unfair predictions are not desirable!

Determine how much compromise in accuracy or fairness is acceptable to your stakeholders

# Fairness, Accuracy, and Profits



Interactive visualization: <https://research.google.com/bigpicture/attacking-discrimination-in-ml/>

# Fairness, Accuracy, and Profits

Fairness can conflict with accuracy goals

Fairness can conflict with organizational goals (profits, usability)

Fairer products may attract more customers

Unfair products may receive bad press, reputation damage

Improving fairness through better data can benefit everybody

# Example: Recidivism Revisited



Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

## Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

- COMPAS system, developed by Northpointe: Used by judges in sentencing decisions across multiple states (incl. PA)

# Which fairness definition?

Table 11.1: COMPAS Fairness Metrics

Metric	Caucasian	African American
False Positive Rate ( <i>FPR</i> )	23%	45%
False Negative Rate ( <i>FNR</i> )	48%	28%
False Discovery Rate ( <i>FDR</i> )	41%	37%

- ProPublica: COMPAS violates equalized odds w/ FPR & FNR
- Northpointe: COMPAS is fair because it has similar FDRs
  - $FDR = FP / (FP + TP) = 1 - \text{Precision}$ ;  $FPR = FP / (FP + TN)$
- **Q. Which measure is appropriate in this context?**

Figure from Big Data and Social Science, Ch. 11 A. Chouldechova [Fair prediction with disparate impact: A study of bias in recidivism prediction instruments](#)

False discovery rate: rate of Type 1 errors (reject null hypothesis when it's true)

Positive predictive value is the probability that a patient with a positive (abnormal) test result actually has the disease. Negative predictive value is the probability that a person with a negative (normal) test result is truly free of disease.

when the positive predictive values are constrained to be equal but the prevalences differ across groups, the false positive and false negative rates cannot both be equal across those groups.

(i) Allow unequal false negative rates to retain equal PPV's and achieve equal false positive rates (ii) Allow unequal false positive rates to retain equal PPV's and achieve equal false negative rates (iii) Allow unequal PPV's to achieve equal false positive and false negative rates

# Fairness Tree



Ian Foster, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter and Julia Lane. [Big Data and Social Science: Data Science Methods and Tools for Research and Practice](#). Chapter 11, 2nd ed, 2020

# Breakout: Fairness Goal for Hire Decisions?



Post as a group in #lecture:

- What kind of harm can be caused (to different stakeholders)?
- Fairness goal: Equality or equity?
- Model: Anti-classification, group fairness, or equalized odds (with FPR/FNR)?

# Law: "Four-fifth rule" (or "80% rule")

- Group fairness with a threshold:  $\frac{P[R=1|A=a]}{P[R=1|A=b]} \geq 0.8$
- Selection rate for a protected group (e.g.,  $A = a$ ) < 80% of highest rate => selection procedure considered as having "adverse impact"
- Guideline adopted by Federal agencies (Department of Justice, Equal Employment Opportunity Commission, etc.,) in 1978
- If violated, must justify business necessity (i.e., the selection procedure is essential to the safe & efficient operation)
- Example: Hiring 50% of male applicants vs 20% female applicants hired ( $0.2/0.5 = 0.4$ ) -- Is there a business justification for hiring men at a higher rate?

Speaker notes

skip me



# Dataset Construction for Fairness



- Instead of just focusing on building a "fair" model, can we understand & address the root causes of bias?

# Flexibility in Data Collection

- Data science education often assumes data as given
- In industry, we often have control over data collection and curation (65%)
- Most address fairness issues by collecting more data (73%)
  - Carefully review data collection procedures, sampling bias, how trustworthy labels are
  - **Often high-leverage point to improve fairness!**

# Data Bias

## Data Source

- **Functional:** biases due to platform affordances and algorithms
- **Normative:** biases due to community norms
- **External:** biases due to phenomena outside social platforms
- **Non-individuals:** e.g., organizations, automated agents

## Data Collection

- **Acquisition:** biases due to, e.g., API limits
- **Querying:** biases due to, e.g., query formulation
- **Filtering:** biases due to removal of data "deemed" irrelevant

## Data Processing

- **Cleaning:** biases due to, e.g., default values
- **Enrichment:** biases from manual or automated annotations
- **Aggregation:** e.g., grouping, organizing, or structuring data

## Data Analysis

- **Qualitative Analyses:** lack generalizability, interpret. biases
- **Descriptive Statistics:** confounding bias, obfuscated measurements
- **Prediction & Inferences:** data representation, perform. variations
- **Observational studies:** peer effects, select. bias, ignorability

## Evaluation

- **Metrics:** e.g., reliability, lack of domain insights
- **Interpretation:** e.g., contextual validity, generalizability
- **Disclaimers:** e.g., lack of negative results and reproducibility

- Bias can be introduced at any stage of the data pipeline!



# Types of Data Bias

- Population bias
- Behavioral bias
- Historical bias
- Content production bias
- Linking bias
- Temporal bias

# Population Bias

Data set	Gender		Skin Color/Type	
	Female	Male	Darker	Lighter
LFW [15]	22.5%	77.4%	18.8%	81.2%
IJB-C [28]	37.4%	62.7%	18.0%	82.0%
Pubfig [35]	50.8%	49.2%	18.0%	82.0%
CelebA [9]	58.1%	42.0%	14.2%	85.8%
UTKface [32]	47.8%	52.2%	35.6%	64.4%
AgeDB [33]	40.6%	59.5%	5.4%	94.6%
PPB [36]	44.6%	55.4%	46.4%	53.6%
IMDB-Face [24]	45.0%	55.0%	12.0%	88.0%

Table 3: Distribution of gender and skin color/type for seven prominent face image data sets.

- Differences in demographics between dataset vs target population
- May result in degraded services for certain groups

# Behavioral Bias



Figure 2: Fitted  $P(a_+)$  and  $P(a_-)$  depending on combinations of gender and race of the reviewed worker. Points show expected values and bars standard errors. In Fiverr, Black workers are less likely to be described with adjectives for positive words, and Black Male workers are more likely to be described with adjectives for negative words.

- Differences in user behavior across platforms or social contexts
- Example: Freelancing platforms (Fiverr vs TaskRabbit)
  - Bias against certain minority groups on different platforms

# Fairness-Aware Data Collection

*Fairness-aware Machine Learning*, Bennett et al., WSDM Tutorial (2019).

# Fairness-Aware Data Collection

- Address population bias
  - Does the dataset reflect the demographics in the target population?
  - If not, collect more data to achieve this

*Fairness-aware Machine Learning*, Bennett et al., WSDM Tutorial (2019).

# Fairness-Aware Data Collection

- Address population bias
  - Does the dataset reflect the demographics in the target population?
  - If not, collect more data to achieve this
- Address under- & over-representation issues
  - Ensure sufficient amount of data for all groups to avoid being treated as "outliers" by ML
  - Also avoid over-representation of certain groups (e.g., remove historical data)

*Fairness-aware Machine Learning*, Bennett et al., WSDM Tutorial (2019).

# Fairness-Aware Data Collection

*Fairness-aware Machine Learning*, Bennett et al., WSDM Tutorial (2019).

# Fairness-Aware Data Collection

- Data augmentation: Synthesize data for minority groups to reduce under-representation
  - Observed: "He is a doctor" -> synthesize "She is a doctor"

*Fairness-aware Machine Learning*, Bennett et al., WSDM Tutorial (2019).

# Fairness-Aware Data Collection

- Data augmentation: Synthesize data for minority groups to reduce under-representation
  - Observed: "He is a doctor" -> synthesize "She is a doctor"
- Model auditing for better data collection
  - Evaluate accuracy across different groups
  - Collect more data for groups with highest error rates

*Fairness-aware Machine Learning*, Bennett et al., WSDM Tutorial (2019).

# Example Audit Tool: Aequitas

**Aequitas**  
Bias & Fairness Audit

[Home](#) [Code](#) [About](#)

## Bias and Fairness Audit Toolkit

The Bias Report is powered by [Aequitas](#), an open-source bias audit toolkit for machine learning developers, analysts, and policymakers to audit machine learning models for discrimination and bias, and make informed and equitable decisions around developing and deploying predictive risk-assessment tools.

```
graph LR; A[Upload Data] --> B[Select Protected Groups]; B --> C[Select Fairness Metrics]; C --> D[The Bias Report]
```

# Example Audit Tool: Aequitas

## Audit Results: Bias Metrics Values

race

Attribute Value	False Discovery Rate Disparity	False Positive Rate Disparity
African-American	0.91	1.91
Asian	0.61	0.37
Caucasian	1.0	1.0
Hispanic	1.12	0.92
Native American	0.61	1.6
Other	1.12	0.63

# Documentation for Fairness: Data Sheets

Demographic Characteristic	Value
Percentage of female subjects	22.5%
Percentage of male subjects	77.5%
Percentage of White subjects	83.5%
Percentage of Black subjects	8.47%
Percentage of Asian subjects	8.03%
Percentage of people between 0-20 years old	1.57%
Percentage of people between 21-40 years old	31.63%
Percentage of people between 41-60 years old	45.58%
Percentage of people over 61 years old	21.2%

- Common practice in the electronics industry, medicine
- Purpose, provenance, creation, **composition**, distribution
  - "Does the dataset relate to people?"
  - "Does the dataset identify any subpopulations (e.g., by age)?"

## Speaker notes

In the electronics industry, every component, no matter how simple or complex, is accompanied with a datasheet that describes its operating characteristics, test results, recommended uses, and other information. By analogy, we propose that every dataset be accompanied with a datasheet that documents its motivation, composition, collection process, recommended uses, and so on. Datasheets for datasets will facilitate better communication between dataset creators and dataset consumers, and encourage the machine learning community to prioritize transparency and accountability.



# Model Cards

See also: <https://modelcards.withgoogle.com/about>

**Model Card - Toxicity in Text**

**Model Details**

- The TOXICITY classifier provided by Perspective API [32], trained to predict the likelihood that a comment will be perceived as toxic.
- Convolutional Neural Network.
- Developed by Jigsaw in 2017.

**Intended Use**

- Intended to be used for a wide range of use cases such as supporting human moderation and providing feedback to comment authors.
- Not intended for fully automated moderation.
- Not intended to make judgments about specific individuals.

**Factors**

- Identity terms referencing frequently attacked groups, focusing on sexual orientation, gender identity, and race.

**Metrics**

- Pinned AUC, as presented in [11], which measures threshold-agnostic separability of toxic and non-toxic comments for each group, within the context of a background distribution of other groups.

**Ethical Considerations**

- Following [31], the Perspective API uses a set of values to guide their work. These values are Community, Transparency, Inclusivity, Privacy, and Topic-neutrality. Because

**Training Data**

- Proprietary from Perspective API. Following details in [11] and [32], this includes comments from online forums such as Wikipedia and New York Times, with crowdsourced labels of whether the comment is "toxic".
- "Toxic" is defined as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion."

**Evaluation Data**

- A synthetic test set generated using a template-based approach, as suggested in [11], where identity terms are swapped into a variety of template sentences.
- Synthetic data is valuable here because [11] shows that real data often has disproportionate amounts of toxicity directed at specific groups. Synthetic data ensures that we evaluate on data that represents both toxic and non-toxic statements referencing a variety of groups.

**Caveats and Recommendations**

- Synthetic test data covers only a small set of very specific comments. While these are designed to be representative of common use cases and concerns, it is not comprehensive.

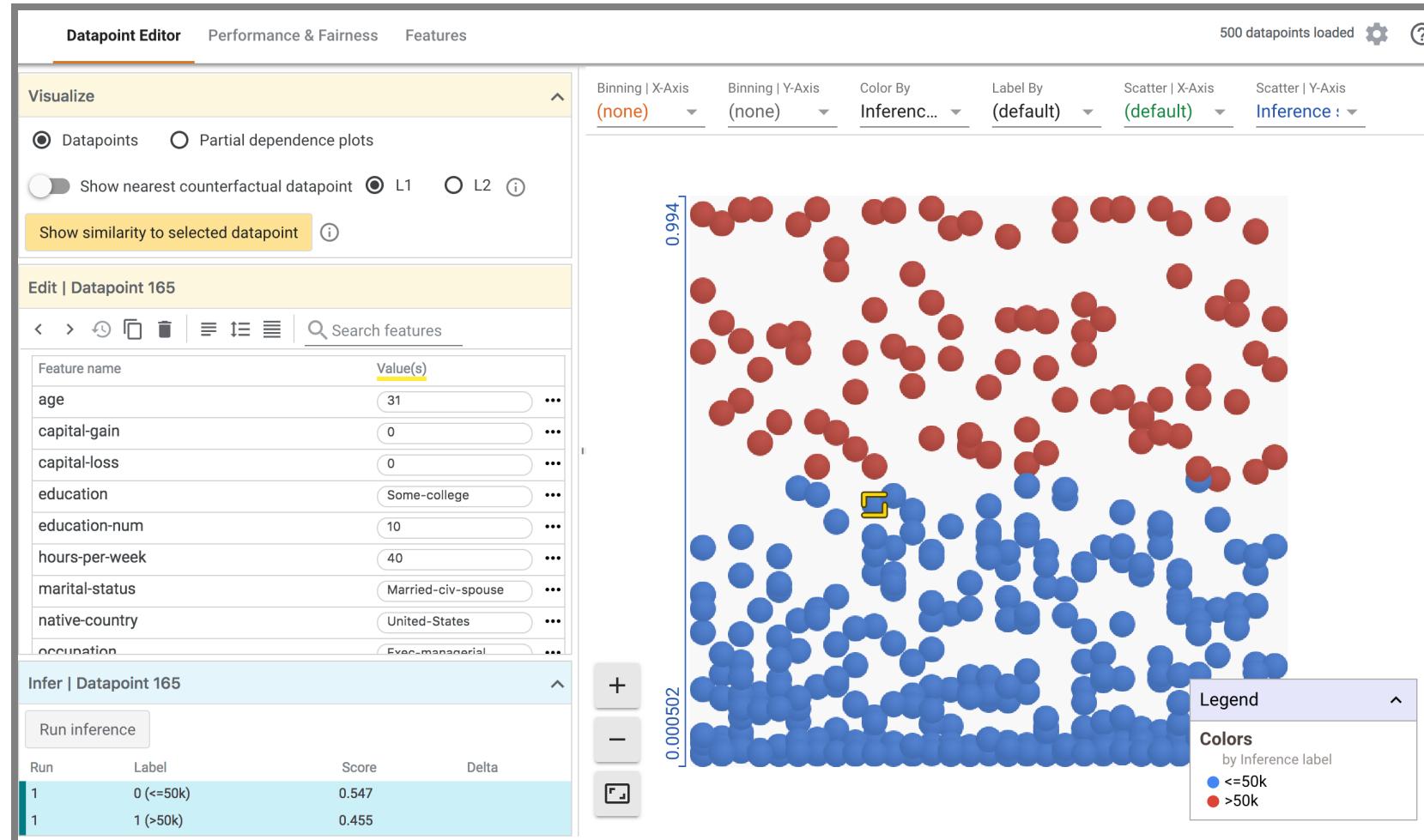
Mitchell, Margaret, et al. "[Model cards for model reporting](#)." In Proceedings of the Conference on fairness, accountability, and transparency, pp. 220-229. 2019.

## Speaker notes

Model cards are short documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions, such as across different cultural, demographic, or phenotypic groups (e.g., race, geographic location, sex, Fitzpatrick skin type) and intersectional groups (e.g., age and race, or sex and Fitzpatrick skin type) that are relevant to the intended application domains. Model cards also disclose the context in which models are intended to be used, details of the performance evaluation procedures, and other relevant information. While we focus primarily on human-centered machine learning models in the application fields of computer vision and natural language processing, this framework can be used to document any trained machine learning model.



# Dataset Exploration



Google What-If Tool

# Fairness beyond the Model

# Bias Mitigation through System Design



Examples of mitigations around the model?

# 1. Avoid Unnecessary Distinctions



*Image captioning gender biased?*

# 1. Avoid Unnecessary Distinctions



"Doctor/nurse applying blood pressure monitor" -> "Healthcare worker applying blood pressure monitor"

# 1. Avoid Unnecessary Distinctions

Is the distinction actually necessary? Is there a more general class to unify them?

Aligns with notion of *justice* to remove the problem from the system

# 2. Suppress Potentially Problem Outputs



≡ How to fix?

# 2. Suppress Potentially Problem Outputs

Anticipate problems or react to reports

Postprocessing, filtering, safeguards

- Suppress entire output classes
- Hardcoded rules or other models (e.g., toxicity detection)

May degrade system quality for some use cases

See mitigating mistakes generally

The New York Times | International Weekly

THE NEW YORK TIMES

Google's photo app still can't find gorillas. And neither can Apple's



A western lowland gorilla at the Bronx Zoo in New York on May 11. Google, whose Android software underpins most of the world's smartphones, has made the decision to turn off the ability to visually search for primates for fear of making an offensive mistake and labeling a person as an animal. And Apple, with technology that performed similarly to Google's in our test, appeared to disable the ability to look for monkeys and apes as well. [Desiree Rios/The New York Times]

# 3. Design Fail-Soft Strategy

Example: Plagiarism detector

A: Cheating detected! This incident has been reported.

B: This answer seems to be perfect. Would you like another exercise?

HCI principle: Fail-soft interfaces avoid calling out directly; communicate friendly and constructively to allow saving face

Especially relevant if system unreliable or biased

# 4. Keep Humans in the Loop

The screenshot shows a transcription interface for a video titled "the-changelog-318". The top navigation bar includes a "Dashboard" link and a "Quality: High" indicator. The main area displays two segments of a transcript:

**Speaker 5 ▶ 07:44**

Yeah. So there's a slight story behind that. So back when I was in, uh, Undergrad, I wrote a program for myself to measure a, the amount of time I did data entry from my father's business and I was on windows at the time and there wasn't a function called time dot [inaudible] time, uh, which I needed to parse dates to get back to time, top of representation, uh, I figured out a way to do it and I gave it to what's called the python cookbook because it just seemed like something other people could use. So it was just trying to be helpful. Uh, subsequently I had to figure out how to make it work because I didn't really have to. Basically, it bothered me that you had to input all the locale information and I figured out how to do it over the subsequent months. And actually as a graduation gift from my Undergrad, the week following, I solved it and wrote it all out.

**Speaker 5 ▶ 08:38**

And I asked, uh, Alex Martelli, the editor of the Python Cookbook, which had published my original recipe, a, how do I get this into python? I think it might help

At the bottom, a feedback section asks "How did we do on your transcript?" followed by five yellow stars.

TV subtitles: Humans check transcripts, especially with heavy dialects

# 4. Keep Humans in the Loop

Recall: Automate vs prompt vs augment

- Involve humans to correct for mistakes and bias
- But, model often introduced to avoid bias in human decision
- But, challenging human-interaction design to keep humans engaged and alert; human monitors possibly biased too, making it worse

**Does a human have a fair chance to detect and correct bias? Enough information? Enough context? Enough time? Unbiased human decision?**

# Predictive Policing Example

*"officers expressed skepticism about the software and during ride alongs showed no intention of using it"*

*"the officer discounted the software since it showed what he already knew, while he ignored those predictions that he did not understand"*

Does the system just lend credibility to a biased human process?

Lally, Nick. "["It makes almost no difference which algorithm you use": on the modularity of predictive policing.](#)" Urban Geography (2021): 1-19.

# (More positive) Medical example

*Mid-level clinicians often doubted that their voice was heard or that their expertise was considered. They were hesitant to directly disagree with a physician. They described the situation as more complicated than just the power dynamics...they had slowly removed themselves from the decision making.*

*"A prognostic DST that indicates post-surgery quality of life could potentially amplify their voices: "...I think it's good to have something visual for anybody to see. It's like, OK. LOOK. Let's slow down a bit here." (Nurse practitioner)"*

Qian, Yang, et al. "[Unremarkable AI: Fiting Intelligent Decision Support into Critical, Clinical Decision-Making Processes](#)." CSCW 2019.

# Process Integration

# Fairness in Practice today

Lots of attention in academia and media

Lofty statements by big companies, mostly aspirational

Strong push by few invested engineers (internal activists)

Some dedicated teams, mostly in Big Tech, mostly research focused

Little institutional support, no broad practices

# Barriers to Fairness Work



# Barriers to Fairness Work

1. Rarely an organizational priority, mostly reactive (media pressure, regulators)
2. Fairness work seen as ambiguous and too complicated for available resources (esp. outside Big Tech)
3. Most fairness work done by volunteers outside official job functions
4. Impact of fairness work difficult to quantify, making it hard to justify resource investment
5. Technical challenges
6. Fairness concerns are project specific, hard to transfer actionable insights and tools across teams

# Burnout is a Real Danger

Unsupported fairness work is frustrating and often ineffective

*“However famous the company is, it’s not worth being in a work situation where you don’t feel like your entire company, or at least a significant part of your company, is trying to do this with you. Your job is not to be paid lots of money to point out problems. Your job is to help them make their product better. And if you don’t believe in the product, then don’t work there.” -- Rumman Chowdhury via [Melissa Heikkilä](#)*

What to do?

# Aspirations

*"They imagined that organizational leadership would understand, support, and engage deeply with responsible AI concerns, which would be contextualized within their organizational context. Responsible AI would be prioritized as part of the high-level organizational mission and then translated into actionable goals down at the individual levels through established processes. Respondents wanted the spread of information to go through well-established channels so that people know where to look and how to share information."*

From Rakova, Bogdana, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. "Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational

# We want to improve Process Integration

- Integrate proactive practices in development processes -- both model and system level!
- Move from individuals to institutional processes distributing the work
- Hold the entire organization accountable for taking fairness seriously

How?

# Improving Process Integration -- Examples

1. Mandatory discussion of discrimination risks, protected attributes, and fairness goals in *requirements documents*
2. Required fairness reporting in addition to accuracy in automated *model evaluation*
3. Required internal/external fairness audit before *release*
4. Required fairness monitoring, oversight infrastructure in *operation*

# Improving Process Integration -- Examples

5. Instituting fairness measures as *key performance indicators* of products
6. Assign clear responsibilities of who does what
7. Identify measurable fairness improvements, recognize in performance evaluations

*How to avoid pushback against bureaucracy?*

# Affect Culture Change

Buy-in from management is crucial

1. Show that fairness work is taken seriously through action (funding, hiring, audits, policies), not just lofty mission statements

Reported success strategies:

1. Frame fairness work as financial profitable, avoiding rework and reputation cost
2. Demonstrate concrete, quantified evidence of benefits of fairness work
3. Continuous internal activism and education initiatives
4. External pressure from customers and regulators

# Assigning Responsibilities

Hire/educate T-shaped professionals

Have dedicated fairness expert(s) consulting with teams, performing/guiding audits, etc

Not everybody will be a fairness expert, but ensure base-level awareness on when to seek help

# Anticipate Feedback Loops

# Feedback Loops



# Feedback Loops in Mortgage Applications?



# Feedback Loops go through the Environment



# Long-term Impact of ML

- ML systems make multiple decisions over time, influence the behaviors of populations in the real world
- *But* most models are built & optimized assuming that the world is static
- Difficult to estimate the impact of ML over time
  - Need to reason about the system dynamics (world vs machine)
  - e.g., what's the effect of a mortgage lending policy on a population?

# Long-term Impact & Fairness

Deploying an ML model with a fairness criterion does NOT guarantee improvement in equality/equity over time

Even if a model appears to promote fairness in short term, it may result harm over long term



Fairness is not static: deeper understanding of long term fairness via simulation studies, in FAT\* 2020. 1

# Analyze the World vs the Machine



 *State and check assumptions!*

# Analyze the World vs the Machine

How do outputs affect change in the real world, how does this (indirectly) influence inputs?

Can we decouple inputs from outputs? Can telemetry be trusted?

Interventions through system (re)design:

- Focus data collection on less influenced inputs
- Compensate for bias from feedback loops in ML pipeline
- Do not build the system in the first place

# Prepare for Feedback Loops

We will likely not anticipate all feedback loops...

... but we can anticipate that unknown feedback loops exist

-> Monitoring!

# Best Practices

# Monitoring & Auditing

- Operationalize fairness measure in production with telemetry

# Monitoring & Auditing

- Operationalize fairness measure in production with telemetry
- Continuously monitor for:
  - Mismatch between training data, test data, and instances encountered in deployment
  - Data shifts: May suggest needs to adjust fairness metric/thresholds
  - User reports & complaints: Log and audit system decisions perceived to be unfair by users

# Monitoring & Auditing

- Operationalize fairness measure in production with telemetry
- Continuously monitor for:
  - Mismatch between training data, test data, and instances encountered in deployment
  - Data shifts: May suggest needs to adjust fairness metric/thresholds
  - User reports & complaints: Log and audit system decisions perceived to be unfair by users
- Invite diverse stakeholders to audit system for biases

# Monitoring & Auditing



- Continuously monitor the fairness metric (e.g., error rates for different sub-populations)
- Re-train model with recent data or adjust classification thresholds if needed

# Preparing for Problems

Prepare an *incidence response plan* for fairness issues

- What can be shut down/reverted on short notice?
- Who does what?
- Who talks to the press? To affected parties? What do they need to know?

Provide users with a path to *appeal decisions*

- Provide feedback mechanism to complain about unfairness
- Human review? Human override?

# Best Practices

**Best practices are emerging and evolving**

Start early, be proactive

Scrutinize data collection and labeling

Invest in requirements engineering and design

Invest in education

Assign clear responsibilities, demonstrate leadership buy-in

# Many Tutorials, Checklists, Recommendations

Tutorials (fairness notions, sources of bias, process recom.):

- Fairness in Machine Learning, Fairness-Aware Machine Learning in Practice
- Challenges of Incorporating Algorithmic Fairness into Industry Practice

Checklist:

- Microsoft's AI Fairness Checklist: concrete questions, concrete steps throughout all stages, including deployment and monitoring

# Summary

- Requirements engineering for fair ML systems
  - Identify potential harms, protected attributes
  - Negotiate conflicting fairness goals, tradeoffs
  - Consider societal implications
- Apply fair data collection practices
- Anticipate feedback loops
- Operationalize & monitor for fairness metrics
- Design fair systems beyond the model, mitigate bias outside the model
- Integrate fairness work in process and culture

# Further Readings

- Rakova, Bogdana, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. "[Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices](#)." *Proceedings of the ACM on Human-Computer Interaction* 5, no. CSCW1 (2021): 1-23.
- Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. "[Model cards for model reporting](#)." In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220-229. 2019.
- Boyd, Karen L. "[Datasheets for Datasets help ML Engineers Notice and Understand Ethical Issues in Training Data](#)." *Proceedings of the ACM on Human-Computer Interaction* 5, no. CSCW2 (2021): 1-27.
- Bietti, Elettra. "[From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy](#)." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 210-219. 2020.
- Madaio, Michael A., Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. "[Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI](#)." In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1-14. 2020.
- Hopkins, Aspen, and Serena Booth. "[Machine Learning Practices Outside Big Tech: How Resource Constraints Challenge Responsible Development](#)." In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)* (2021).
- Metcalf, Jacob, and Emanuel Moss. "[Owning ethics: Corporate logics, silicon valley, and the institutionalization of ethics](#)." *Social Research: An International Quarterly* 86, no. 2 (2019): 449-476.

