FAIRNESS: BEYOND MODEL

Eunsuk Kang

Required reading: Os Keyes, Jevan Hutson, Meredith Durbin. A Mulching Proposal: Analysing and Improving an Algorithmic System for Turning the Elderly into High-Nutrient Slurry. CHI Extended Abstracts, 2019.

LEARNING GOALS

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

FAIRNESS CRITERIA: REVIEW

REVIEW OF CRITERIA SO FAR:

Recidivism scenario: Should a person be detained?

- Anti-classification: ?
- Independence: ?
- Separation: ?

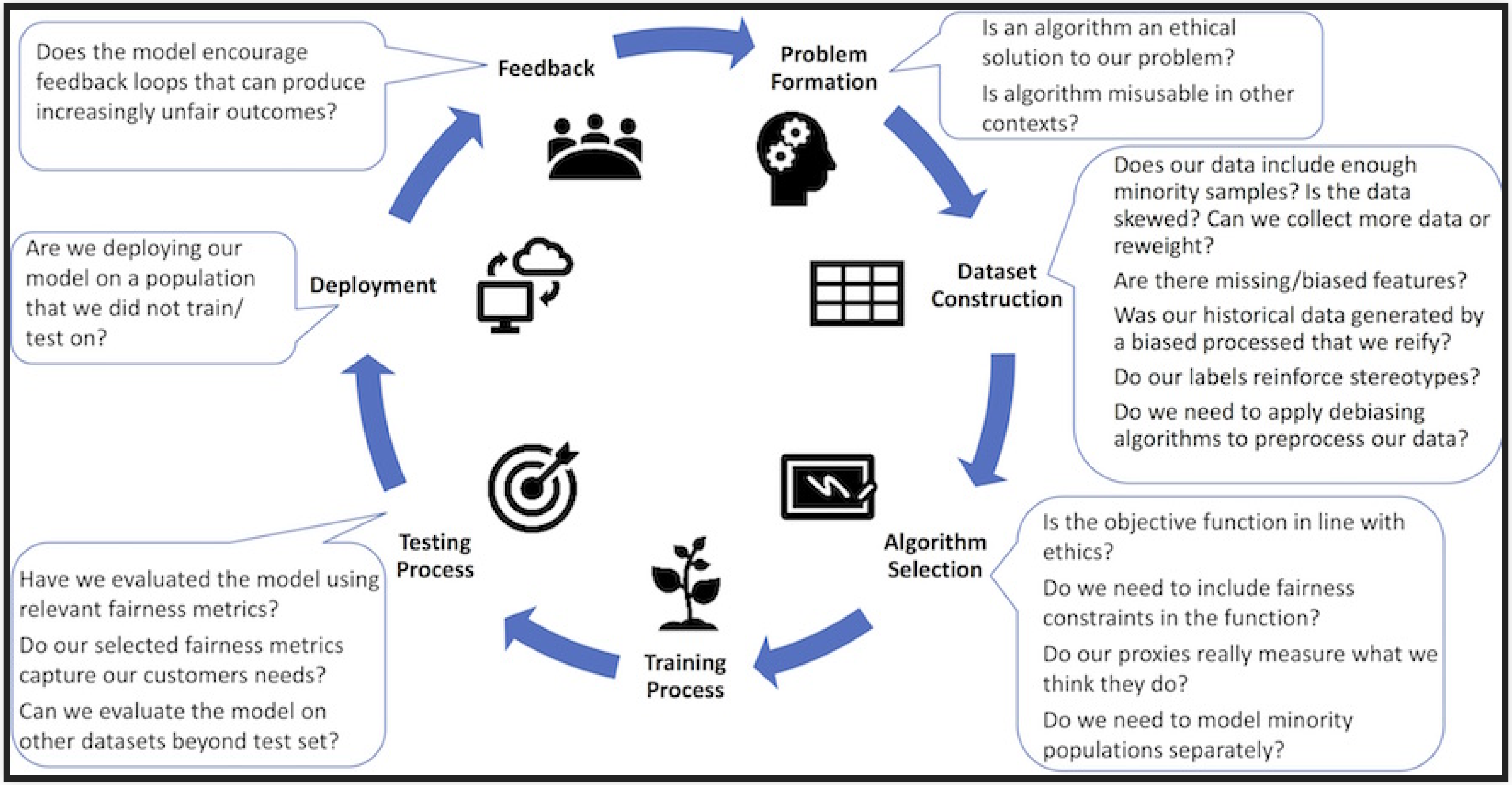
REVIEW OF CRITERIA SO FAR:

Recidivism scenario: Should a defendant be detained?

- Anti-classification: Race and gender should not be considered for the decision at all
- Independence: Detention rates should be equal across gender and race groups
- Separation: Among defendants who would not have gone on to commit a violent crime if released, detention rates are equal across gender and race groups

BUILDING FAIR ML SYSTEMS

FAIRNESS MUST BE CONSIDERED THROUGHOUT THE ML LIFECYCLE!



Does the model encourage feedback loops that can produce increasingly unfair outcomes?

Is an algorithm an ethical solution to our problem?

Is algorithm misusable in other contexts?

Does our data include enough minority samples? Is the data skewed? Can we collect more data or reweight?

Are there missing/biased features?

Was our historical data generated by a biased processed that we reify?

Do our labels reinforce stereotypes?

Do we need to apply debiasing algorithms to preprocess our data?

Is the objective function in line with ethics?

Do we need to include fairness constraints in the function?

Do our proxies really measure what we think they do?

Do we need to model minority populations separately?

Have we evaluated the model using relevant fairness metrics?

Do our selected fairness metrics capture our customers needs?

Can we evaluate the model on other datasets beyond test set?

Are we deploying our model on a population that we did not train/ test on?

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

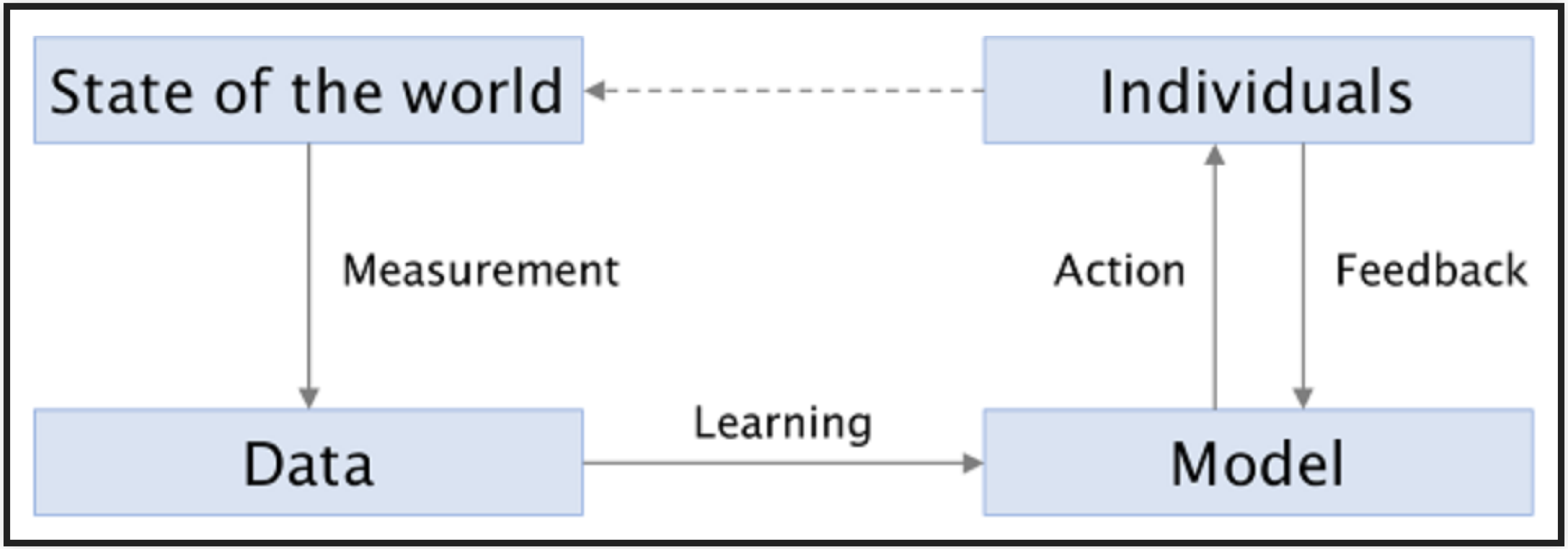
PRACTITIONER CHALLENGES

- Fairness is a system-level property
    - consider goals, user interaction design, data collection, monitoring, model interaction (properties of a single model may not matter much)
- Fairness-aware data collection, fairness testing for training data
- Identifying blind spots
    - Proactive vs reactive
    - Team bias and (domain-specific) checklists
- Fairness auditing processes and tools
- Diagnosis and debugging (outlier or systemic problem? causes?)
- Guiding interventions (adjust goals? more data? side effects? chasing mistakes? redesign?)
- Assessing human bias of humans in the loop

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.

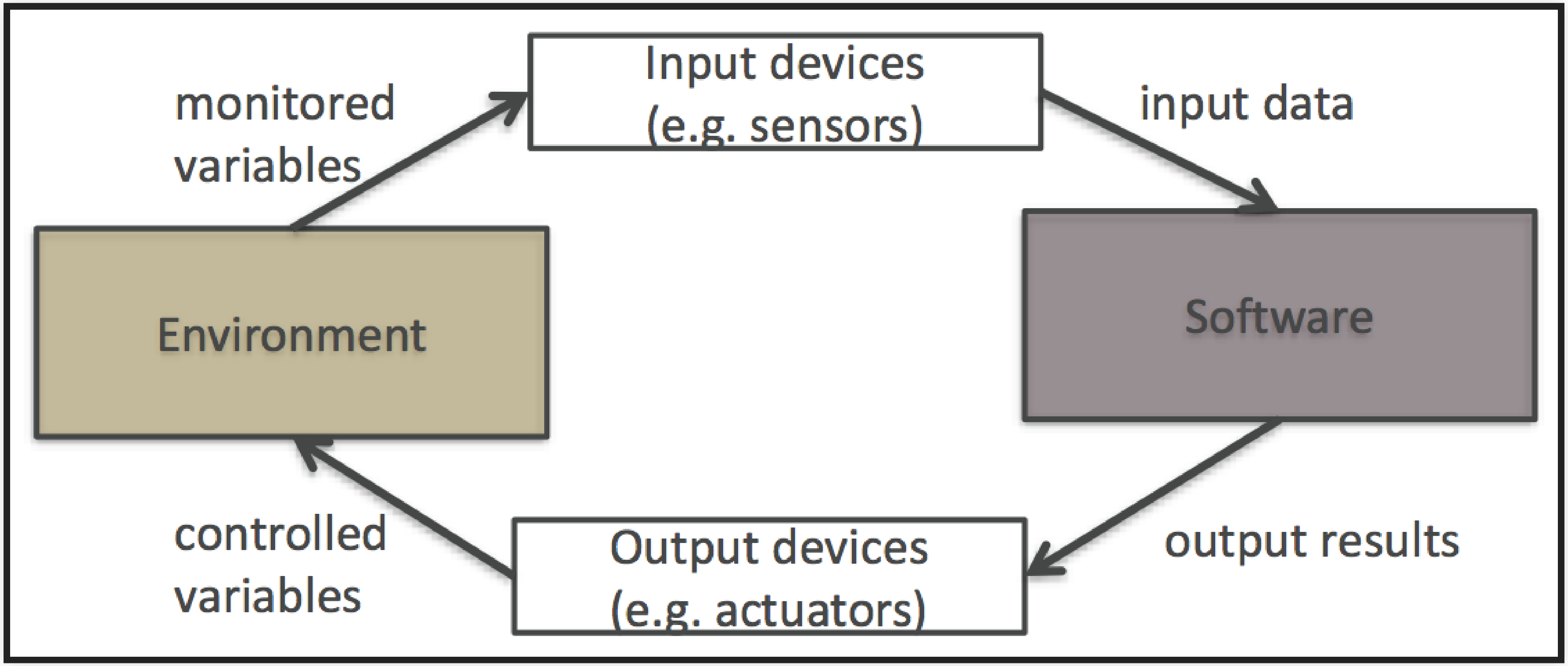
REQUIREMENTS FOR FAIRNESS

MACHINE LEARNING CYCLE



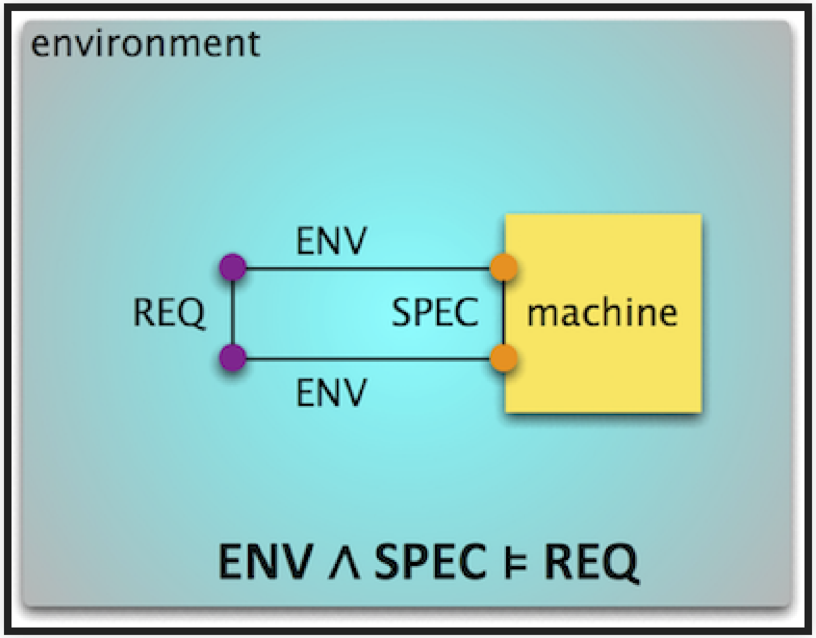
"Fairness and Machine Learning" by Barocas, Hardt, and Narayanan (2019), Chapter 1.

RECALL: MACHINE VS WORLD



- No ML/AI lives in vacuum; every system is deployed as part of the world
- A requirement describes a desired state of the world (i.e., environment)
- Machine (software) is *created* to manipulate the environment into this state

REQUIREMENT VS SPECIFICATION



- Requirement (REQ): What the system should do, as desired effects on the environment
- Assumptions (ENV): What’s assumed about the behavior/properties of the environment (based on domain knowledge)
- Specification (SPEC): What the software must do in order to satisfy REQ

REQUIREMENTS FOR FAIR ML SYSTEMS

- Identify requirements (REQ) over the environment
    - What types of harm can be caused by biased decisions?
    - Who are stakeholders? Which population groups can be harmed?
    - Are we trying to achieve equality vs. equity?
    - What are legal requirements to consider?

"FOUR-FIFTH RULE" (OR "80% RULE")

$$(P[R = 1|A = a])/(P[R = 1|A = b]) \geq 0.8$$

- 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?

EXAMPLE: LOAN APPLICATION

LOAN APPLICATION

Personal Information

Name (Last) PUBLIC (First) JOHN

Address (Mailing Address) 12345 MAIN STREET

E-Mail Address JQPJQPJQP@JQP.JQP

Home Telephone 1111 1111

Other Telephone 2222 2222

999999

UNDER REVIEW

SUBJECT REVIEW

Services needed

UNDER REVIEW

APPROVED

Current Income

High School Graduate Or General Education (GED) Test Passed? Yes No

highest grade completed

Military (Most recent first)

Credits Earned

Quarterly or

Other (Specify)

Major or Subject

- Who are the stakeholders?
- Types of harm?
- Legal & policy considerations?

REQUIREMENTS FOR FAIR ML SYSTEMS

- Identify requirements (REQ) over the environment
    - What types of harm can be caused by biased decisions?
    - Who are stakeholders? Which population groups can be harmed?
    - Are we trying to achieve equality vs. equity?
    - What are legal requirements to consider?
- Define the interface between the environment & machine (ML)
    - What data will be sensed/measured by AI? Potential biases?
    - What types of decisions will the system make? Punitive or assistive?
- Identify the environmental assumptions (ENV)
    - Adversarial? Misuse? Unfair (dis-)advantages?
    - Population distributions?

EXAMPLE: LOAN APPLICATION

LOAN APPLICATION

APPROVED

Personal Information

Name (Last) PUBLIC (First) JOHN (Middle Initial) Home Telephone 111-11-1111

Address (Mailing Address) 12345 MAIN STREET ANYWHERE (Zip) 999999 Other Telephone 222-22-2222

E-Mail Address JQPJQPJQP@JQP.JQP APPLICANT UNDER REVIEW

SUBJECT REVIEW

Services needed

UNDER REVIEW

Current Income

High School Graduate Or General Education (GED) Test Passed? Yes No

highest grade completed

Military (Most recent first)

Credits Earned

Quarterly or

Other (Specify)

Graduate

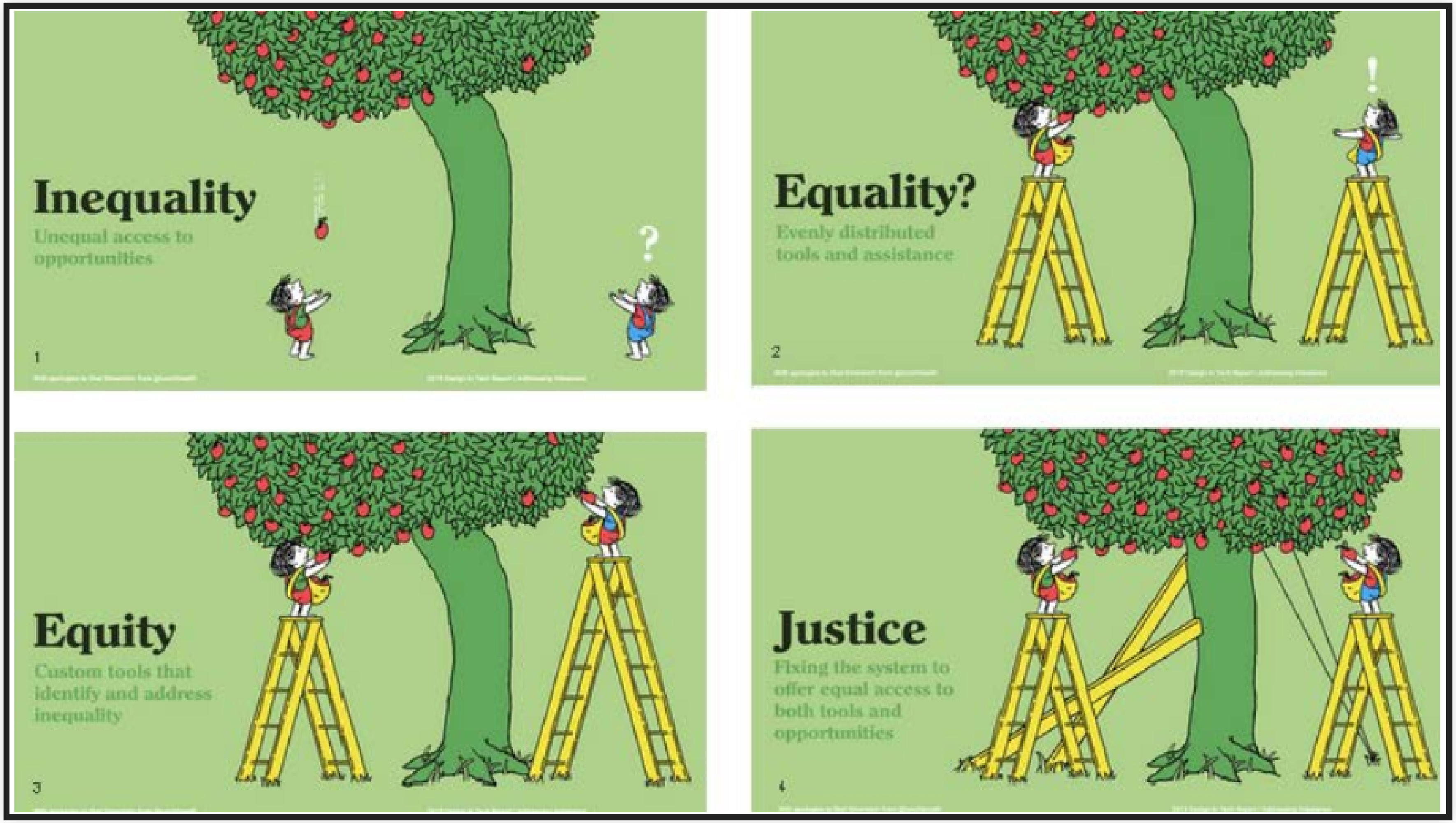
Major or Subject

- Do certain groups of stakeholders have unfair (dis-)advantages?
- What are potential biases in the data measured by the system?

REQUIREMENTS FOR FAIR ML SYSTEMS

- Identify requirements (REQ) over the environment
    - What types of harm can be caused by biased decisions?
    - Who are stakeholders? Which population groups can be harmed?
    - Are we trying to achieve equality vs. equity?
    - What are legal requirements to consider?
- Define the interface between the environment & machine (ML)
    - What data will be sensed/measured by AI? Potential biases?
    - What types of decisions will the system make? Punitive or assistive?
- Identify the environmental assumptions (ENV)
    - Adversarial? Misuse? Unfair (dis-)advantages?
    - Population distributions?
- Devise machine specifications (SPEC) that are sufficient to establish REQ
    - What type of fairness definition is appropriate?

RECALL: EQUALITY VS EQUITY


Inequality
Unequal access to opportunities
?
1
Equality?
Evenly distributed tools and assistance
!
2
Equity
Custom tools that identify and address inequality
3
Justice
Fixing the system to offer equal access to both tools and opportunities
4

TYPE OF DECISION & POSSIBLE HARM

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

WHICH FAIRNESS CRITERIA?

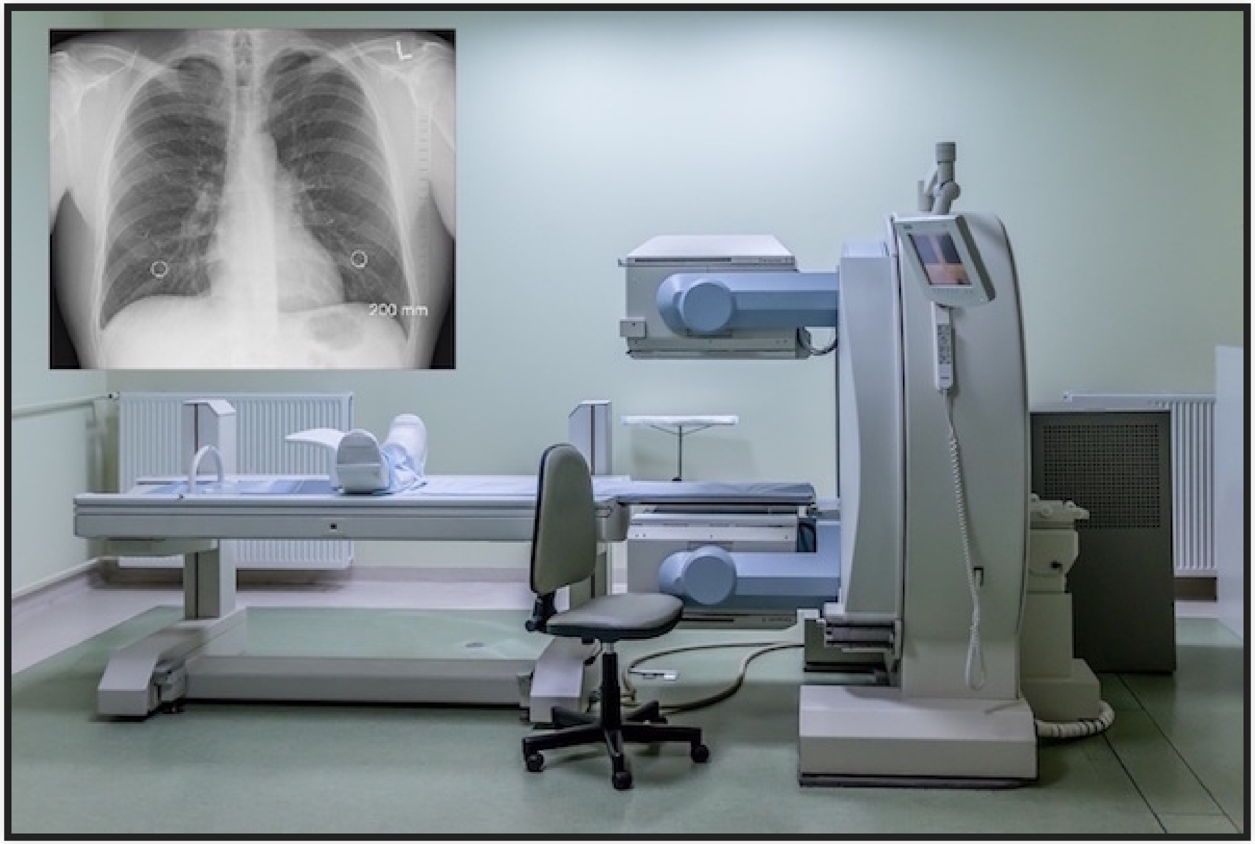
- Decision: Classify whether a defendant should be detained
- Criteria: Anti-classification,

independence, or seperation w/ FPR or FNR?

WHICH FAIRNESS CRITERIA?

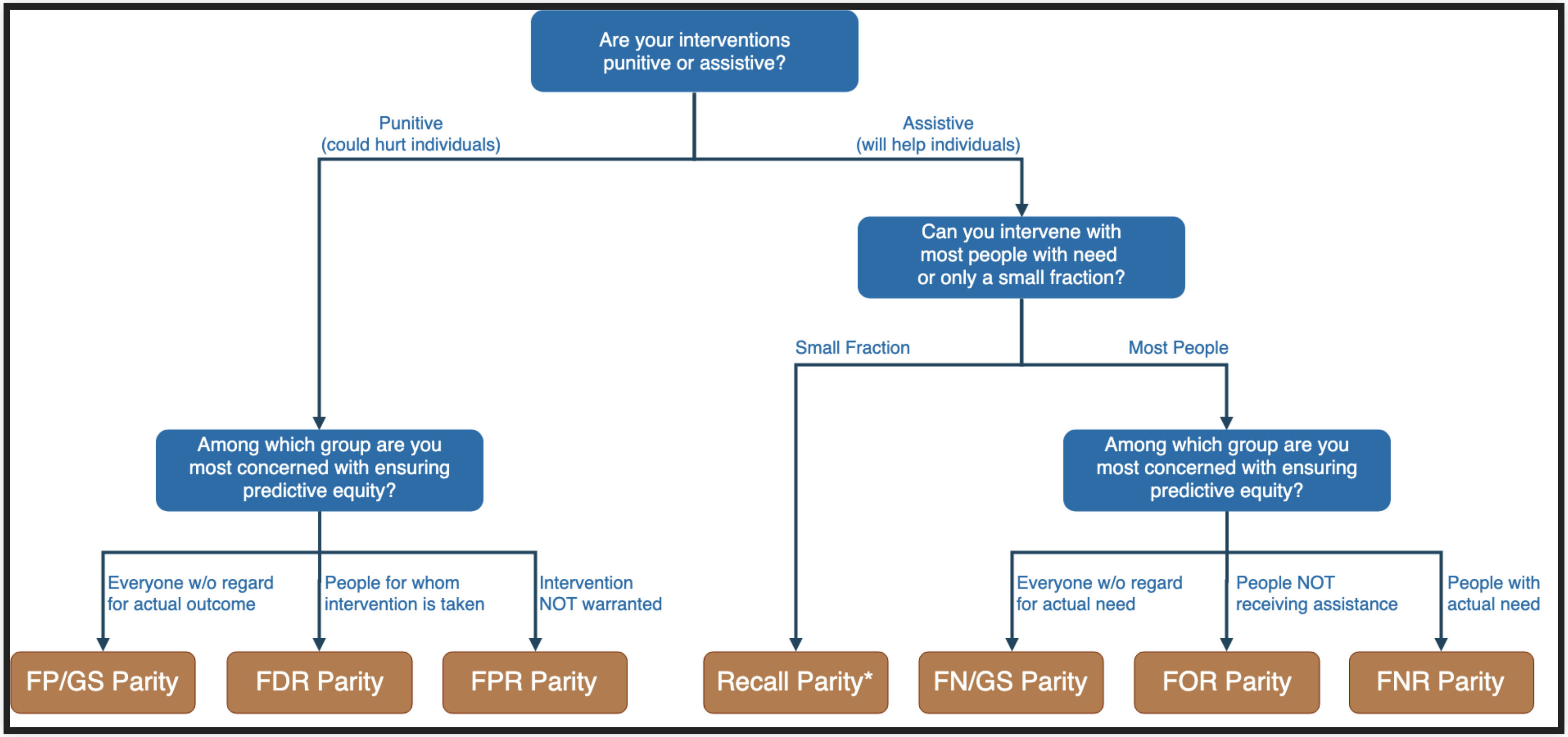
- Decision: Classify whether an applicant should be granted a loan.
- Criteria: Anti-classification, independence, or seperation w/ FPR or FNR?

WHICH FAIRNESS CRITERIA?



- Decision: Classify whether a patient has a high risk of cancer
- Criteria: Anti-classification, independence, or seperation w/ FPR or FNR?

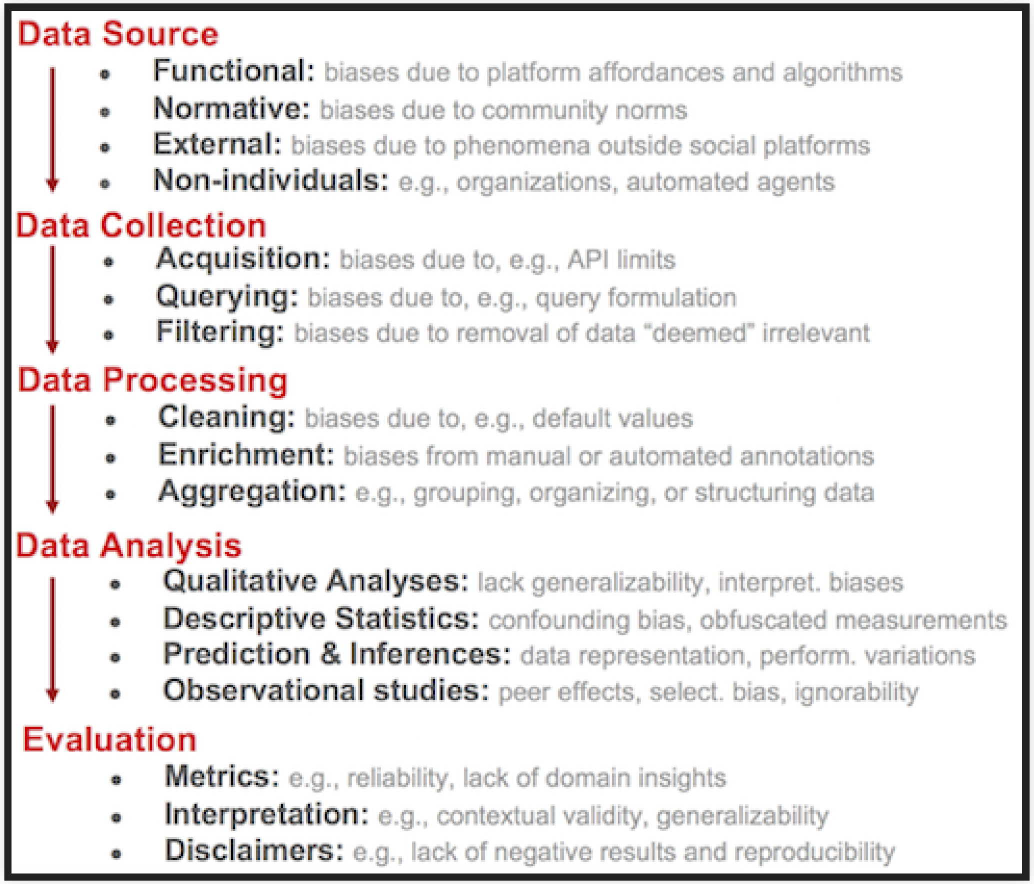
FAIRNESS TREE


Are your interventions punitive or assistive?
Punitive (could hurt individuals)
Assistive (will help individuals)
Can you intervene with most people with need or only a small fraction?
Small Fraction
Most People
Among which group are you most concerned with ensuring predictive equity?
Everyone w/o regard for actual outcome
People for whom intervention is taken
Intervention NOT warranted
Among which group are you most concerned with ensuring predictive equity?
Everyone w/o regard for actual need
People NOT receiving assistance
People with actual need
FP/GS Parity
FDR Parity
FPR Parity
Recall Parity*
FN/GS Parity
FOR Parity
FNR Parity

For details on other types of fairness metrics, see:
https://textbook.coleridgeinitiative.org/chap-bias.html

DATASET CONSTRUCTION FOR 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

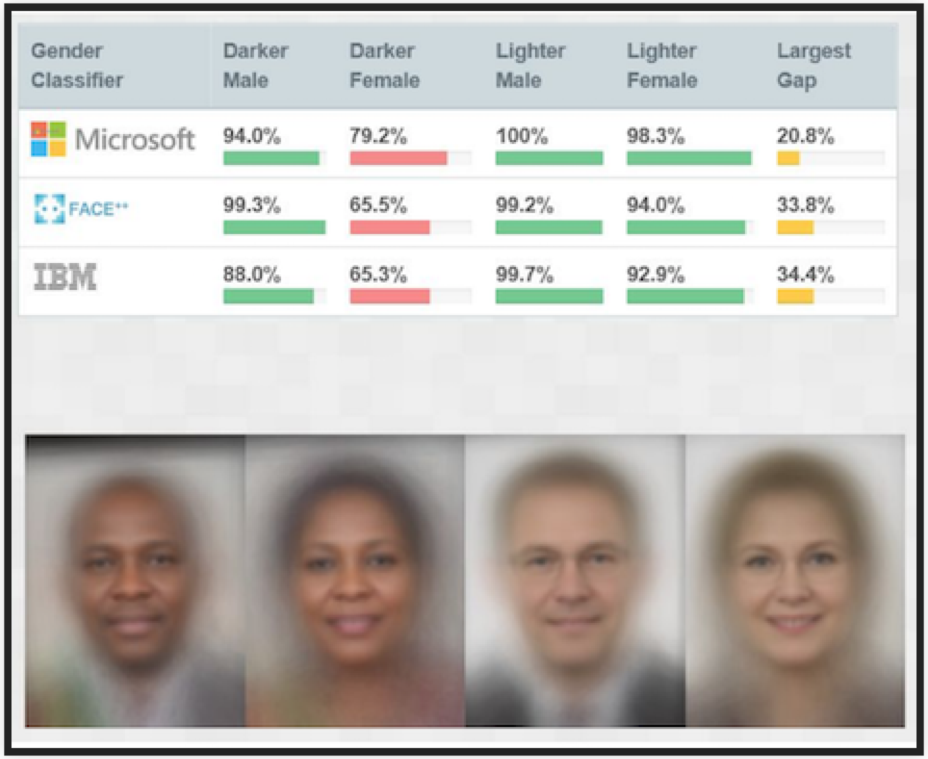
- A **systematic distortion** in data that compromises its use for a task
- Bias can be introduced at any stage of the data pipeline!

TYPES OF DATA BIAS

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

Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries, Olteanu et al., Frontiers in Big Data (2016).

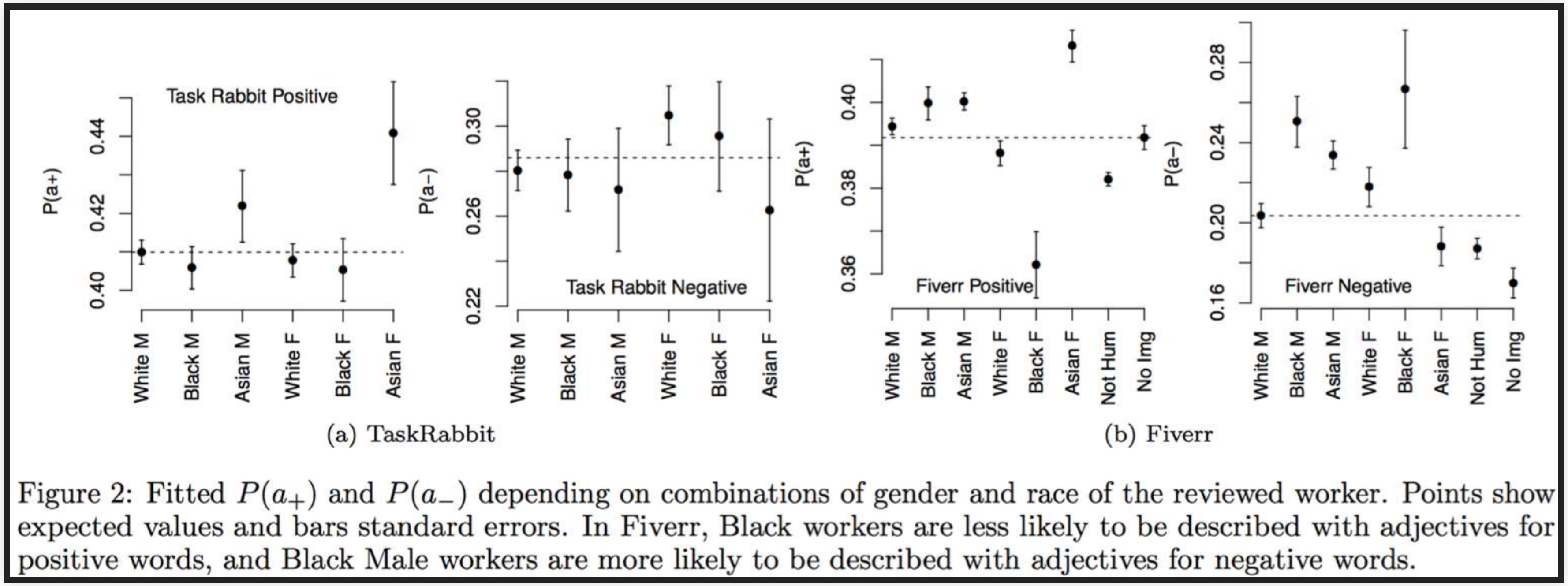
POPULATION BIAS



Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE⁺⁺	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

- Differences in demographics between a dataset vs a target population
- Example: Does the Twitter demographics represent the general population?
- In many tasks, datasets should match the target population
- But some tasks require equal representation for fairness (Q. example?)

BEHAVIORAL BIAS



(a) TaskRabbit

(b) Fiverr

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

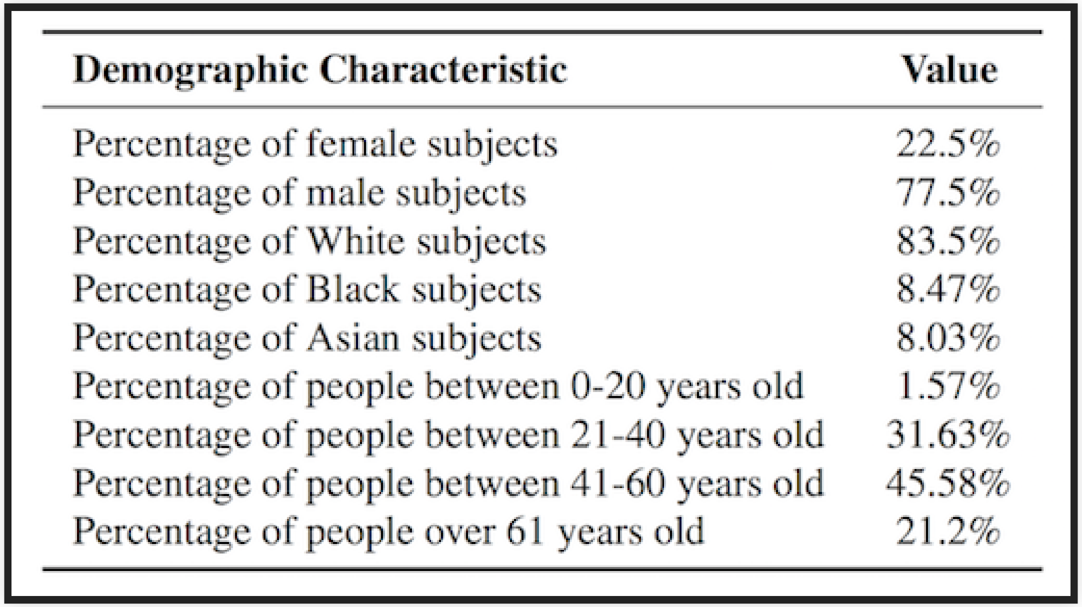
Bias in Online Freelance Marketplaces, Hannak et al., CSCW (2017).

FAIRNESS-AWARE DATA COLLECTION

- Address population bias
    - Does the dataset reflect the demographics in the target population?
- Address under- & over-representation issues
    - Ensure sufficient amount of data for all groups to avoid being treated as "outliers" by ML
    - But also avoid over-representation of certain groups (e.g., remove historical data)
- Data augmentation: Synthesize data for minority groups
    - Observed: "He is a doctor" -> synthesize "She is a doctor"
- Fairness-aware active learning
    - Collect more data for groups with highest error rates

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

DATA SHEETS

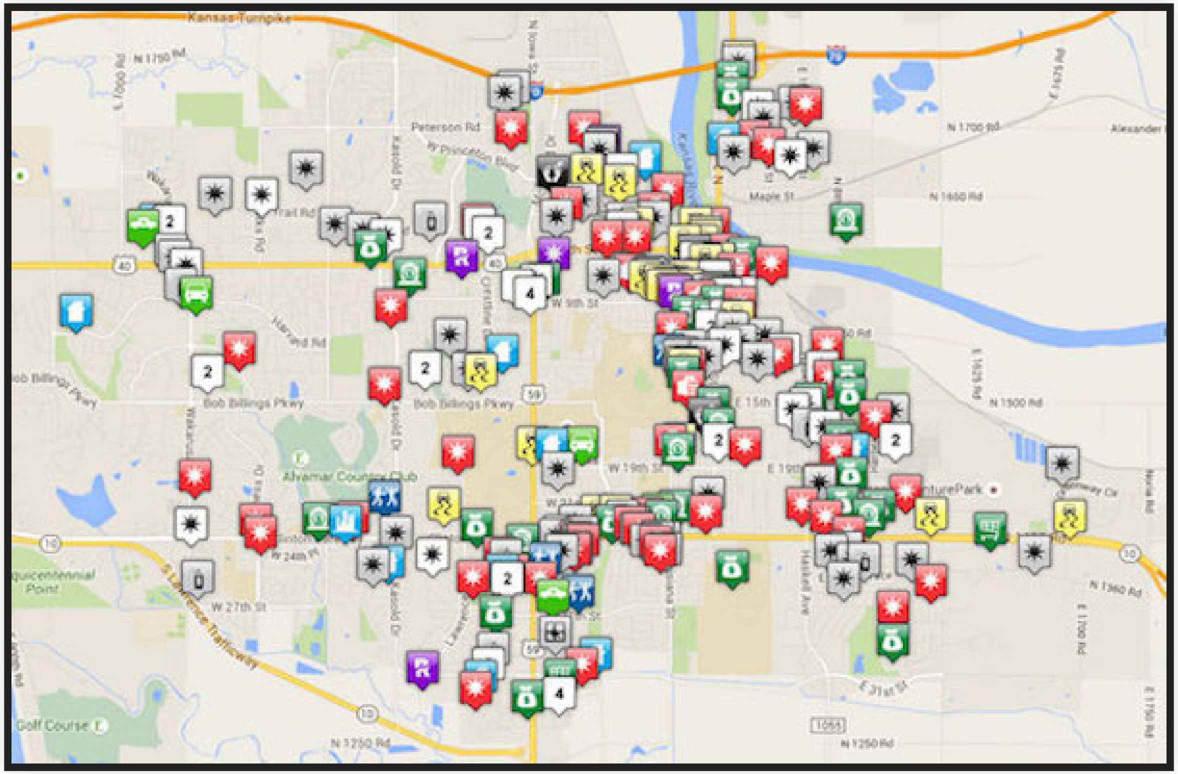


Demographic Characteristic	Value
Percentage of female subjects	22.5%
Percentage of male subjects	77.5%
Percentage of White subjects	83.5%
Percentage of Black subjects	8.47%
Percentage of Asian subjects	8.03%
Percentage of people between 0-20 years old	1.57%
Percentage of people between 21-40 years old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

- A process for documenting datasets
- 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, gender)?"

Datasheets for Dataset, Gebru et al., (2019). https://arxiv.org/abs/1803.09010

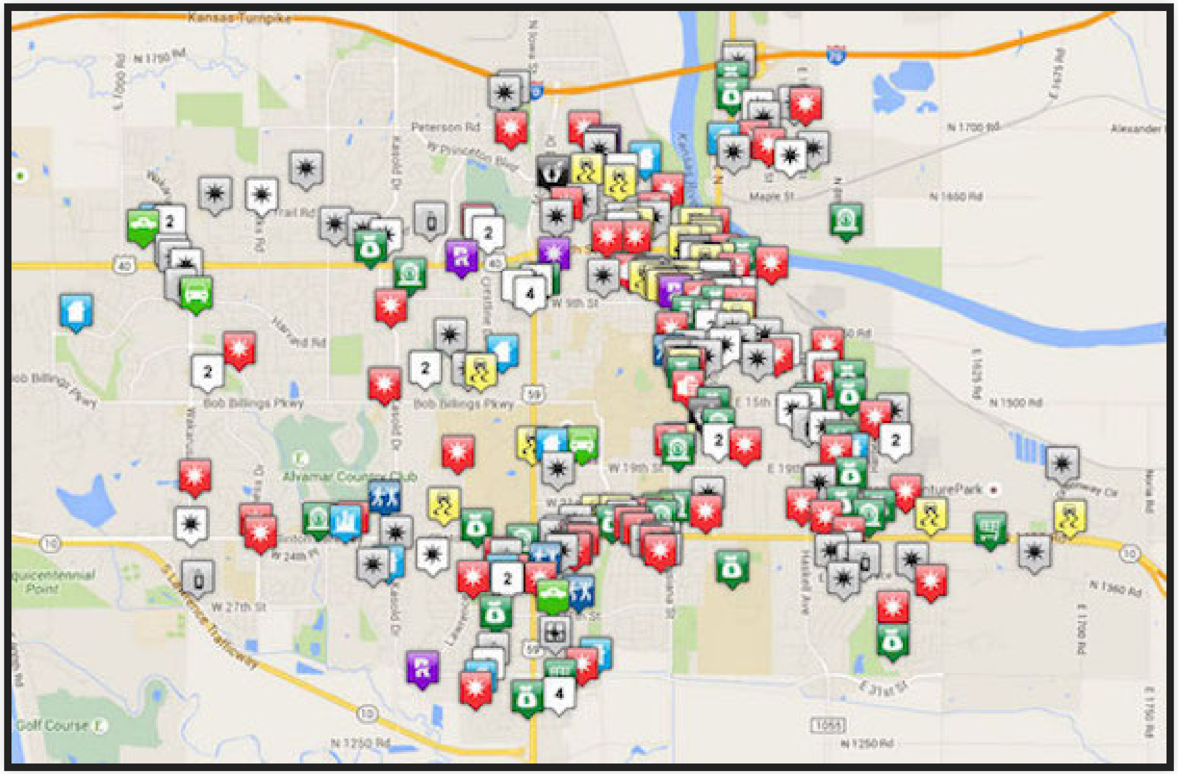
EXAMPLE: PREDICTIVE POLICING



Q. How can we modify an existing dataset or change the data collection process to reduce bias?

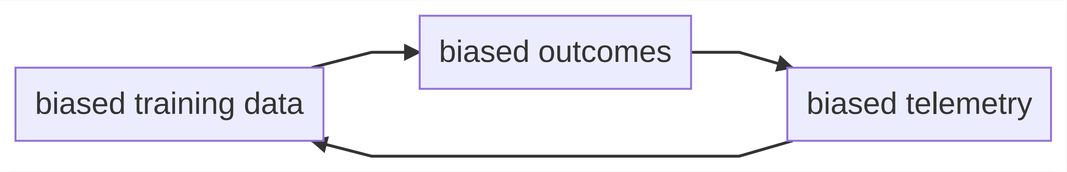
MONITORING AND AUDITING

EXAMPLE: PREDICTIVE POLICING



- Model: Use historical data to predict crime rates by neighborhoods
- Increased patrol => more arrested made in neighborhood X
- New crime data fed back to the model
- Repeat...

FEEDBACK LOOPS



> *"Big Data processes codify the past. They do not invent the future. Doing that requires moral imagination, and that’s something only humans can provide. " -- Cathy O'Neil in Weapons of Math Destruction*

KEY PROBLEMS

- We trust algorithms to be objective, may not question their predictions
- Often designed by and for privileged/majority group
- Algorithms often black box (technically opaque and kept secret from public)
- Predictions based on correlations, not causation; may depend on flawed statistics
- Potential for gaming/attacks
- Despite positive intent, feedback loops may undermine the original goals

O'Neil, Cathy. Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books, 2016.

MONITORING FAIRENESS-AWARE MODEL

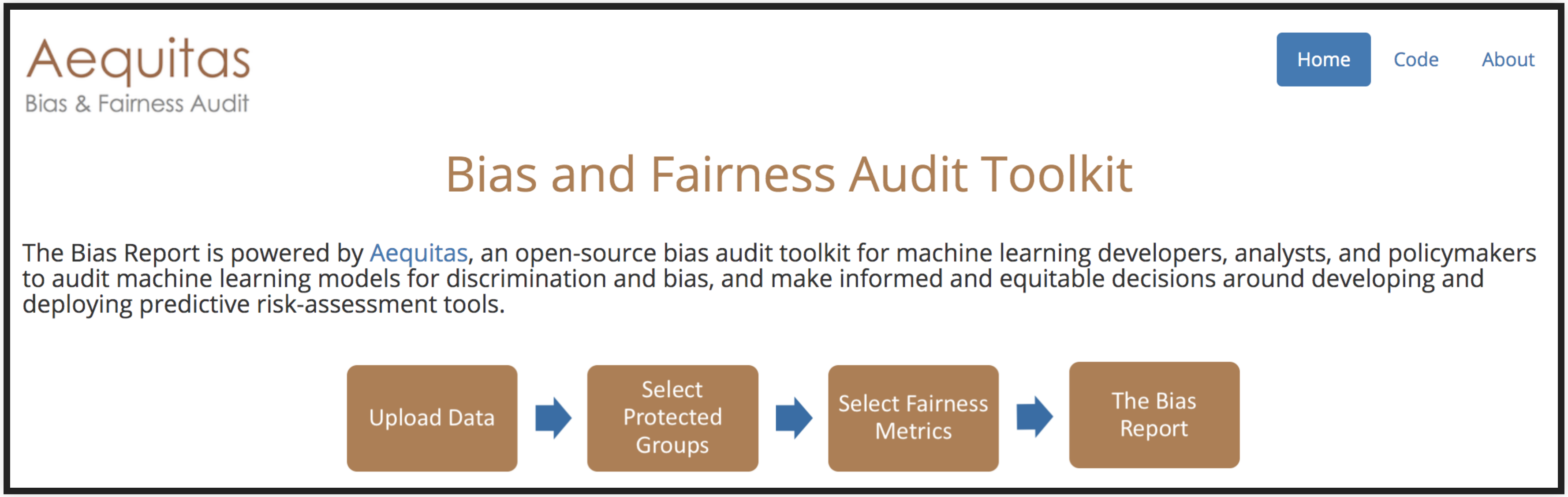
- Deploying an ML model with a fairness criterion does NOT guarantee improvement in equality over time
- **Delayed impact**: Even if a model appears to promote fairness in short term, it may result harm over a long-term period
  - Example: Independence may result in *over-acceptance* (i.e., positive classification) of a group, causing unintended harm
- In general, impact of ML fairness criteria on the society is still poorly understood and difficult to predict
- **Conclusion**: Continuously monitor system for fairness metrics & adjust

Delayed Impact of Fair Machine Learning. Liu et al., (2018)

MONITORING & AUDITING

- Continuously monitor for:
    - Match between training data, test data, and instances that you encounter in deployment
    - Fairness metrics: Is the system yielding fair results over time?
    - Population shifts: May suggest needs to adjust fairness metric/thresholds
    - User reports & complaints: Log and audit system decisions perceived to be unfair by users
- Deploy escalation plans: How do you respond when harm occurs due to system?
    - Shutdown system? Temporary replacement?
    - Maintain communication lines to stakeholders
- Invite diverse stakeholders to audit system for biases

MONITORING TOOLS: EXAMPLE



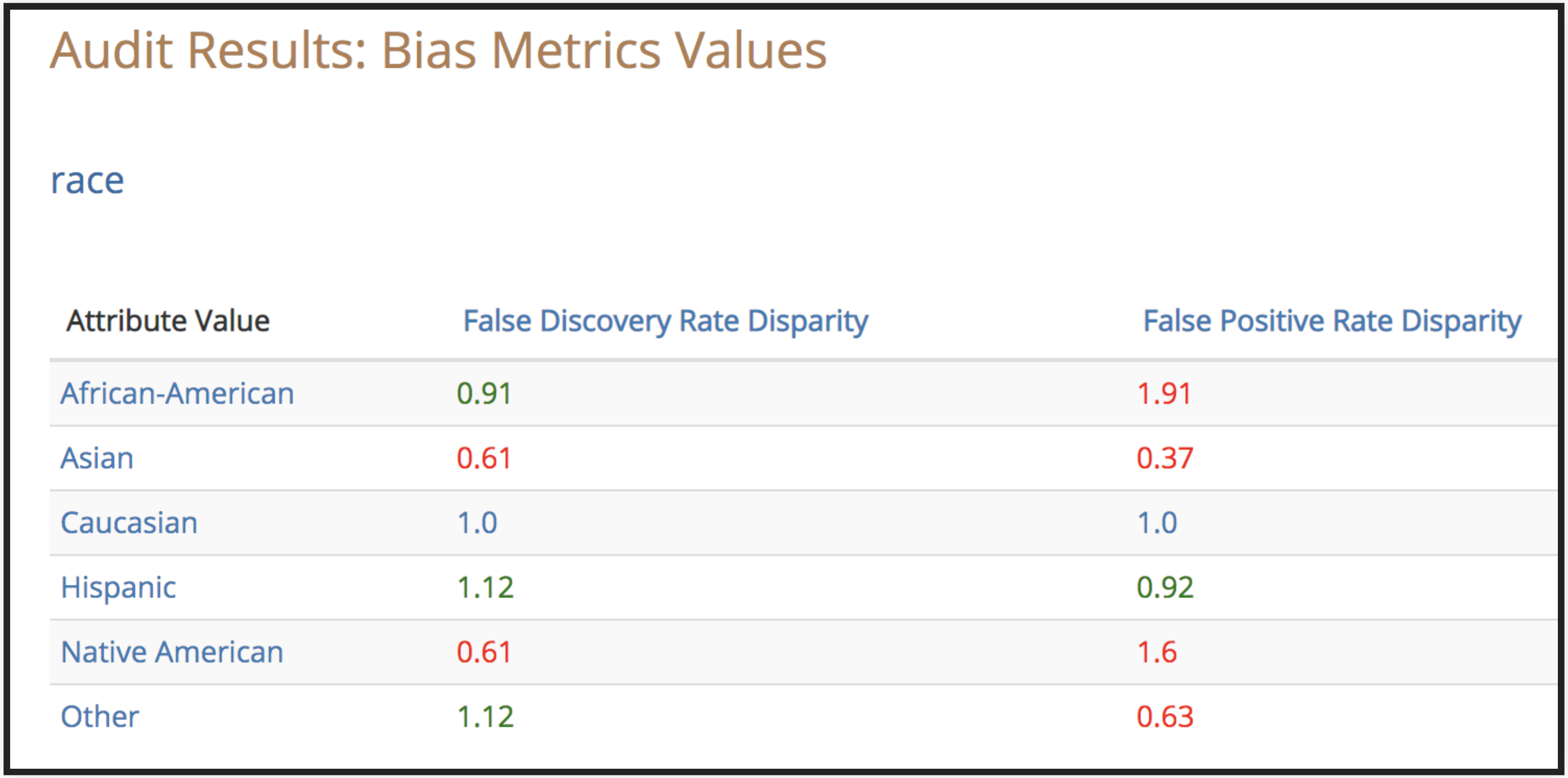
Aequitas
Bias & Fairness Audit

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.

http://aequitas.dssg.io/

MONITORING TOOLS: EXAMPLE



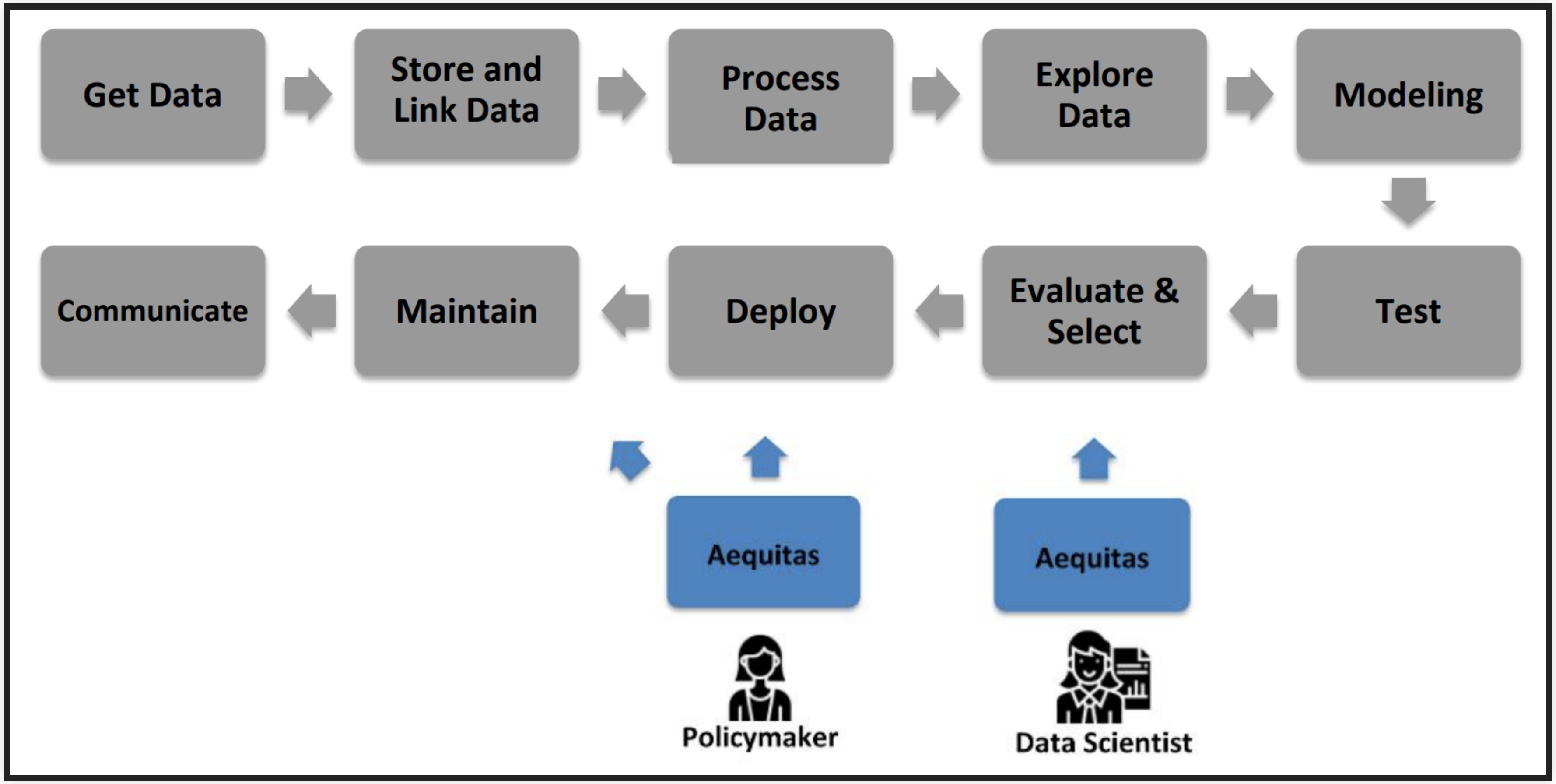
Audit Results: Bias Metrics Values

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

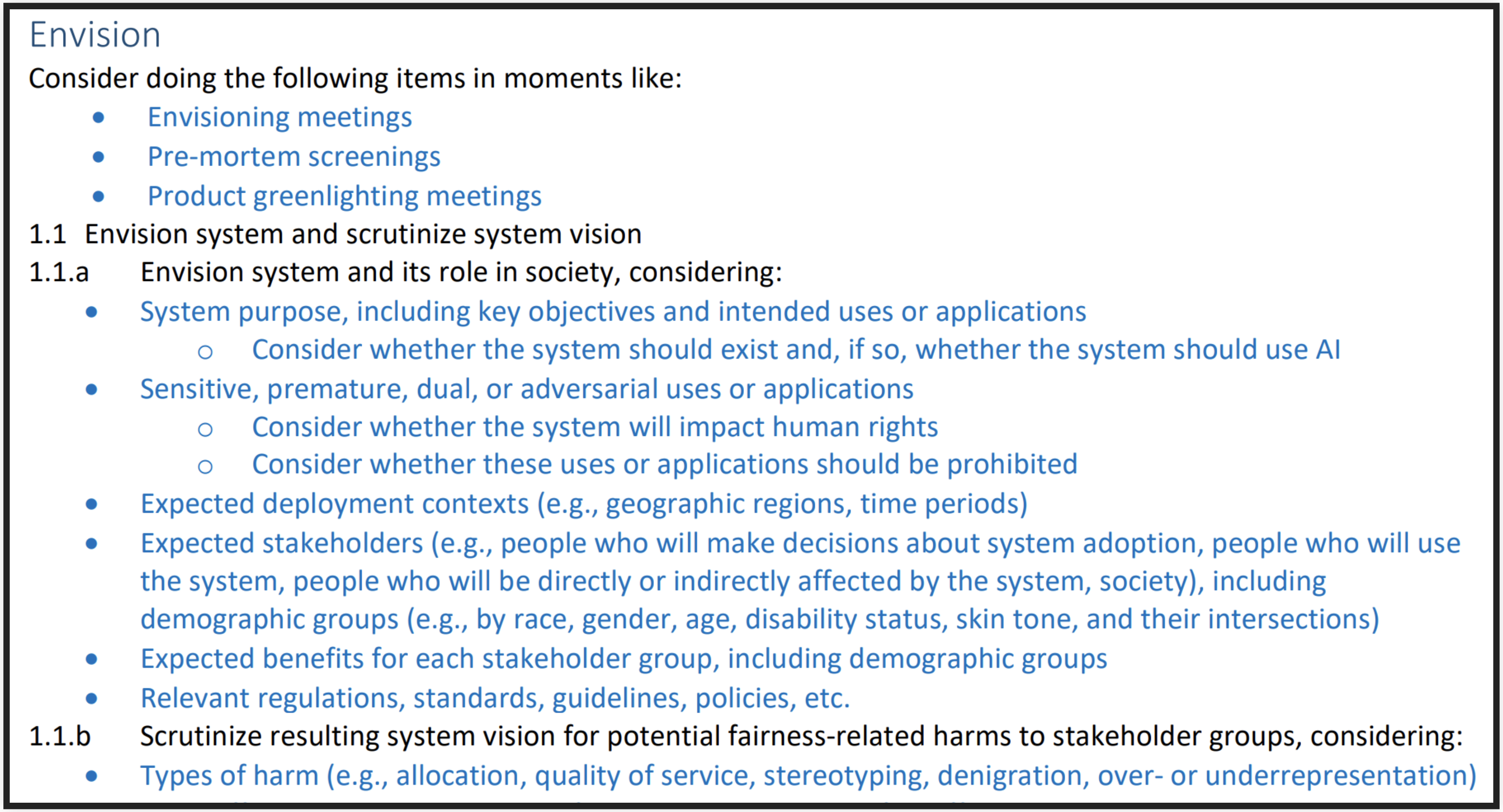
- Continuously make fairness measurements to detect potential shifts in data, population behavior, etc.,

MONITORING TOOLS: EXAMPLE



- Involve policy makers in the monitoring & auditing process

FAIRNESS CHECKLIST



Envision

Consider doing the following items in moments like:

- Envisioning meetings
- Pre-mortem screenings
- Product greenlighting meetings

1.1 Envision system and scrutinize system vision

1.1.a Envision system and its role in society, considering:

- System purpose, including key objectives and intended uses or applications
  - ○ Consider whether the system should exist and, if so, whether the system should use AI
- Sensitive, premature, dual, or adversarial uses or applications
  - ○ Consider whether the system will impact human rights
  - ○ Consider whether these uses or applications should be prohibited
- Expected deployment contexts (e.g., geographic regions, time periods)
- Expected stakeholders (e.g., people who will make decisions about system adoption, people who will use the system, people who will be directly or indirectly affected by the system, society), including demographic groups (e.g., by race, gender, age, disability status, skin tone, and their intersections)
- Expected benefits for each stakeholder group, including demographic groups
- Relevant regulations, standards, guidelines, policies, etc.

1.1.b Scrutinize resulting system vision for potential fairness-related harms to stakeholder groups, considering:

- Types of harm (e.g., allocation, quality of service, stereotyping, denigration, over- or underrepresentation)

Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI, Madaio et al (2020).

CASE STUDY: COLLEGE ADMISSION

- Aspects to consider:
  - Requirements & fairness criteria selection
  - Data collection & pre-processing
  - Impact of feedback loops
  - Monitoring & auditing

SUMMARY

- Achieving fairness as an activity throughout the entire development cycle
- Requirements engineering for fair ML systrems
    - Stakeholders, sub-populations & unfair (dis-)advantages
    - Types of harms
    - Legal requirements
- Dataset construction for fairness
- Consideration for the impact of feedback loops
- Continous montoring & auditing for fairness