AIG150- Week 12

Explainable AI, Fairness in AI

Reading Text:

Model Performance Management with Explainable AI(ch 02)

AI Fairness: How to measure and reduce unwanted bias in machine learning

Agenda

- Explainable AI
- From responsible to XAI
- Who needs XAI?
- Benefits of XAI
- What is Fairness ?
- Biases
- Different types & dimensions of fairness
- How to mitigate biases and unfairness ?

Explainable AI

- The form of AI with the goal of creating machine learning models that are, for the most part, explainable and/or interpretable by humans.
- Open the black box of AI models to make them interpretable by humans, with the intent of minimizing the risk of unknown or unpredictable outcomes from the models.
- By using XAI, it is easier to follow regulations and legal obligations and helps to monitor and manage model’s performance.

Responsible AI → Explainable AI

- Responsible AI is a branch of AI that aims at ensuring all AI systems are designed and built with fairness, privacy, security, and explainability in mind.
- If your models are complex, they might need to be explained by your data science team to understand (and try to eliminate or reduce) any potential biases.
- Your models might also need to be explained for regulatory reasons if you operate in banking, insurance, or other high-compliance industries.
- XAI goes hand in hand with Responsible AI’s focus on ensuring fairness and limiting bias.

AI TO XAI

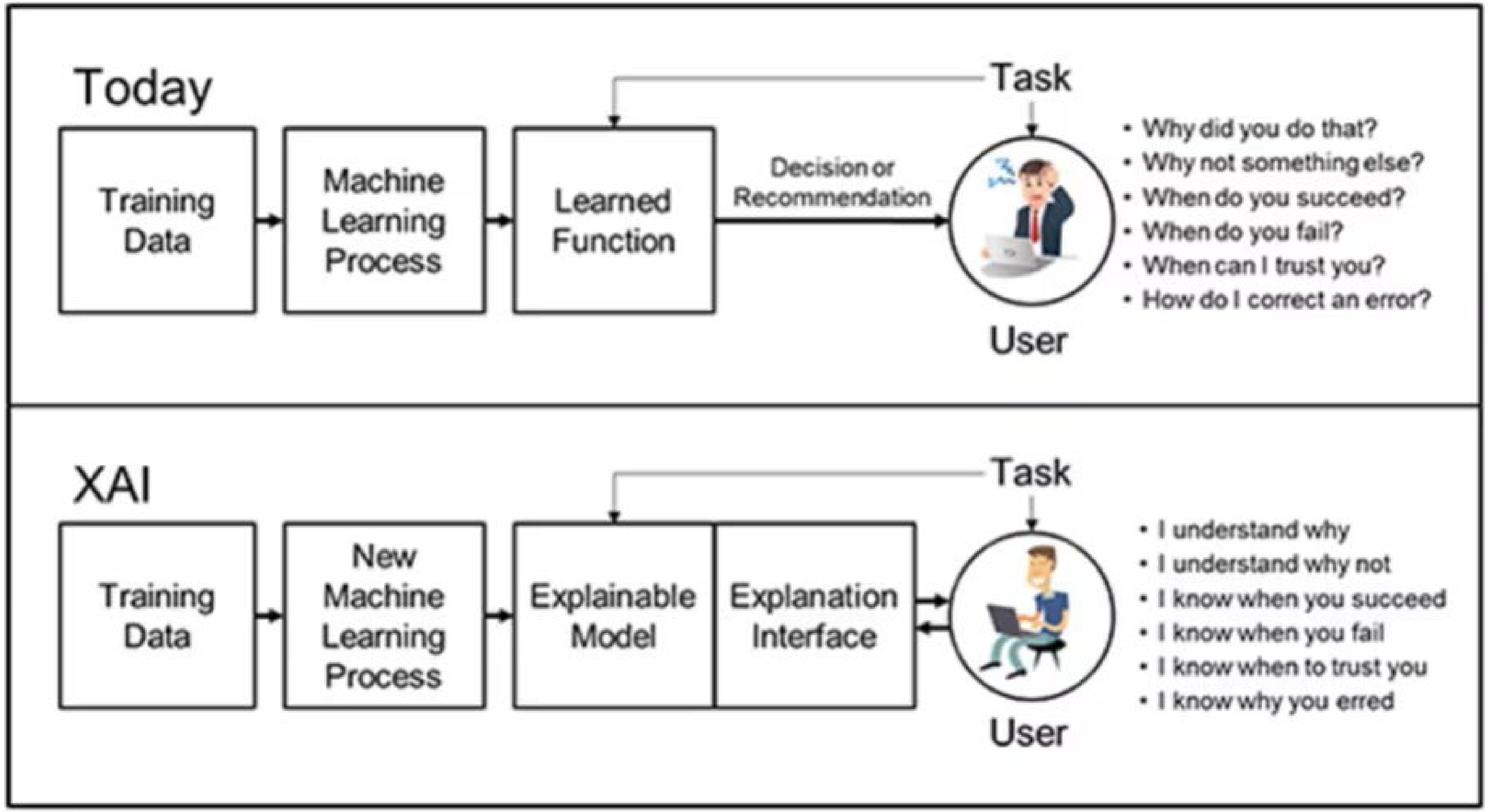


Image taken from https://www.netapp.com/blog/explainable-ai/

Who Needs XAI ?

- Internal technical and business stakeholders such as data scientists, machine learning engineers, product owners, or executives

- End users of the AI system or product

- Public stakeholders such as regulators or investors

Interpretability Versus Explainability

Interpretability is the ability for a cause-and-effect relationship to be mapped.
Input → output

No need for an explanation as why this output is produced.

Explainability is the ability for the effect of each of a model's inputs on its output or prediction to be explained.

Such explanations are important for complicated models such as neural networks where there are multiple inputs fed into a black box, with little insight into its inner workings.

One example is to disturb the input values to see their effect on the output.

Offline Versus Online Explainability

- ***Offline*** explanations are typically used during the development cycles of a model to understand the model better and hopefully build the best version for production use.

- ***Online*** explanations are based on the model’s predictions after it goes into use. Generating spot online explanations can help you debug and find the root cause of operational issues by providing an additional layer of model understanding.

XAI In Different Domains

Tabular/Structured Data Models

- Structured data models are typically opaque and are not easy to interpret if there are more than one or two inputs.
- By increasing the number of input variables, the models becomes more complex and harder to explain.
- As output predictions are based on a series of inputs, it can be beneficial to know which inputs are contributing the most to the model's output.
- To explain a structured or tabular data model, the inputs are sorted in terms of their weight contribution to the model's prediction, and color coding can be applied to help visualize what is happening and to pinpoint outliers or abnormalities.

Text/ Speech Models

- Text models involve some form of NLP and a corpus or dictionary of words to generate a score for the provided text.

- Sentiment analysis is the most commonly used form of NLP, and these models usually display a binary positive or negative sentiment along with a score.

- To explain such models, we attempt to identify the words in the text that contributed most to the text's sentiment score. This can help pinpoint words that are missing from the corpus used to train the model to produce the scores, or words that are misclassified.

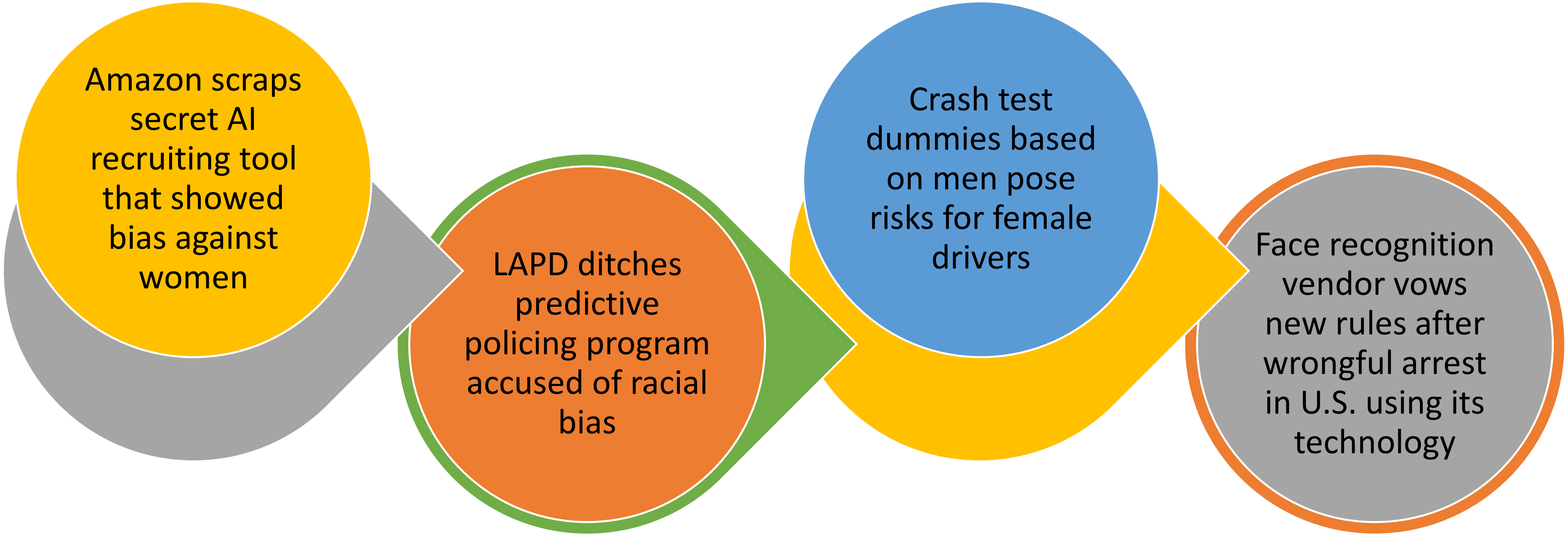
Image/Video Model

- Image classification: One or more labels might be the output of the model.
- Object detection: A box is usually drawn around an object in the image along with some labels or tags.
- Deep learning algorithm called neural network is used in the two cases.
- Image is broken down into small chunks and analysis is done on individual pixels. Pixels are highlighted to show which ones are used in the decision-making process.
- Using XAI, the team can analyze what parts of the image the model is focusing on and whether it is either picking up on unnecessary features within the images or not picking up on important features.

Benefits of XAI

- Operationalize AI with trust and confidence
- Speed time to AI results
- Mitigate risk and cost of model governance
- Increased productivity

Fairness in AI



Amazon scraps secret AI recruiting tool that showed bias against women

LAPD ditches predictive policing program accused of racial bias

Crash test dummies based on men pose risks for female drivers

Face recognition vendor vows new rules after wrongful arrest in U.S. using its technology

Also Check:

- ✓ 4 Shocking AI Bias Examples
- ✓ AI Biases

Biases

- Human decision-making in many areas is biased and unconsciously shaped by our individual or societal biases.
- Bias can occur at any stage in the machine learning pipelines and cannot be removed completely but organizations should define acceptable thresholds for both model accuracy and bias.
- AI bias can come in through :
  - Societal bias embedded in training datasets
  - Decisions made during the machine learning development process
  - Complex feedback loops that arise when a machine learning model is deployed in the real world.
- AI can embed human and societal biases and deploy them at scale.
- Unintentional and unwanted biases are one of the major reasons for the failure of AI or hesitation in adopting AI technologies
- Example: Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)

COMPAS

A 2016 investigation by journalists at ProPublica found that the COMPAS algorithm incorrectly labeled African-American defendants as “high-risk” at nearly twice the rate it mislabeled white defendants. This illustrates the significant negative impact an AI algorithm can have on society. So how do we ensure that automated decisions are less biased than human decision-making?

What Is fairness?

- Fairness is a complex and multifaceted concept that depends on context and culture.
- Defining it for an organization’s use case can thus be difficult.
- There are at least 21 mathematical definitions of fairness. These are not just theoretical differences in how to measure fairness; different definitions focus on different aspects of fairness and thus can produce entirely different outcomes.

Different Types of Fairness

- *Individual Fairness*
  - Treating similar individuals similarly.
  - Individual fairness seeks to ensure that statistical measures of outcomes are equal for similar individuals.

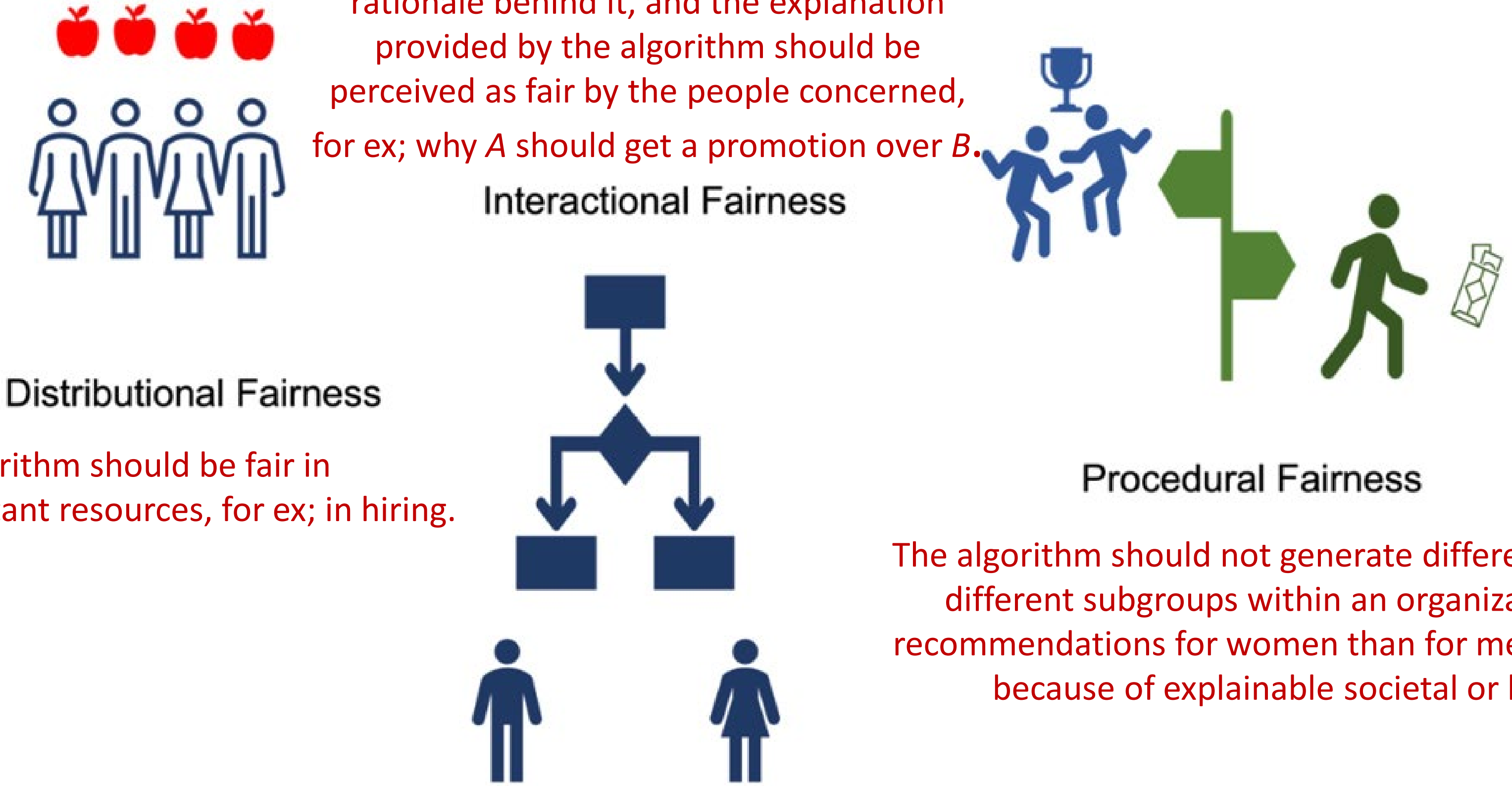
- *Group Fairness*
  - Making the model’s predictions/outcomes equitable across groups.
  - Group fairness partitions a population into groups defined by protected attributes and seeks to ensure that statistical measures of outcomes are equal across groups.

Group Fairness

- We’re All Equal” (WAE)

- “What You See is What You Get” (WYSIWYG). The WAE worldview holds that all groups have the same abilities, while the WYSIWYG worldview holds that observations reflect the abilities of groups.

Three Different Dimensions of Fairness



The algorithm should be able to explain the rationale behind it, and the explanation provided by the algorithm should be perceived as fair by the people concerned, for ex; why *A* should get a promotion over *B*.

The algorithm should be fair in allocating important resources, for ex; in hiring.

The algorithm should not generate different results when used for different subgroups within an organization, for ex; different recommendations for women than for men. If it does, it should be because of explainable societal or biological reasons.

Image Source: Platform and Model Design for Responsible AI by Amita Kapoor, Sharmistha Chatterjee

Fairness Metrics & Bias Mitigation

- We can use fairness metrics to check for bias in machine learning workflows.
    - Sample bias (uneven representation in training data)
    - Label bias (annotation process introduces bias during the creation of label data)
    - Outcome proxy bias (when the machine learning task is not specified appropriately)

- We can use bias mitigators to overcome bias in the workflow to produce a fairer outcome.

- Removal of protected attributes(like race and gender) to minimize bias.

- Bias mitigation is not an easy task.

- Fairness is a multifaceted, context-dependent social construct.

Unfairness Mitigation Methods

- ***Preprocessing***: Preprocessing methods work by adjusting the training data distribution so that the sensitive groups are balanced.

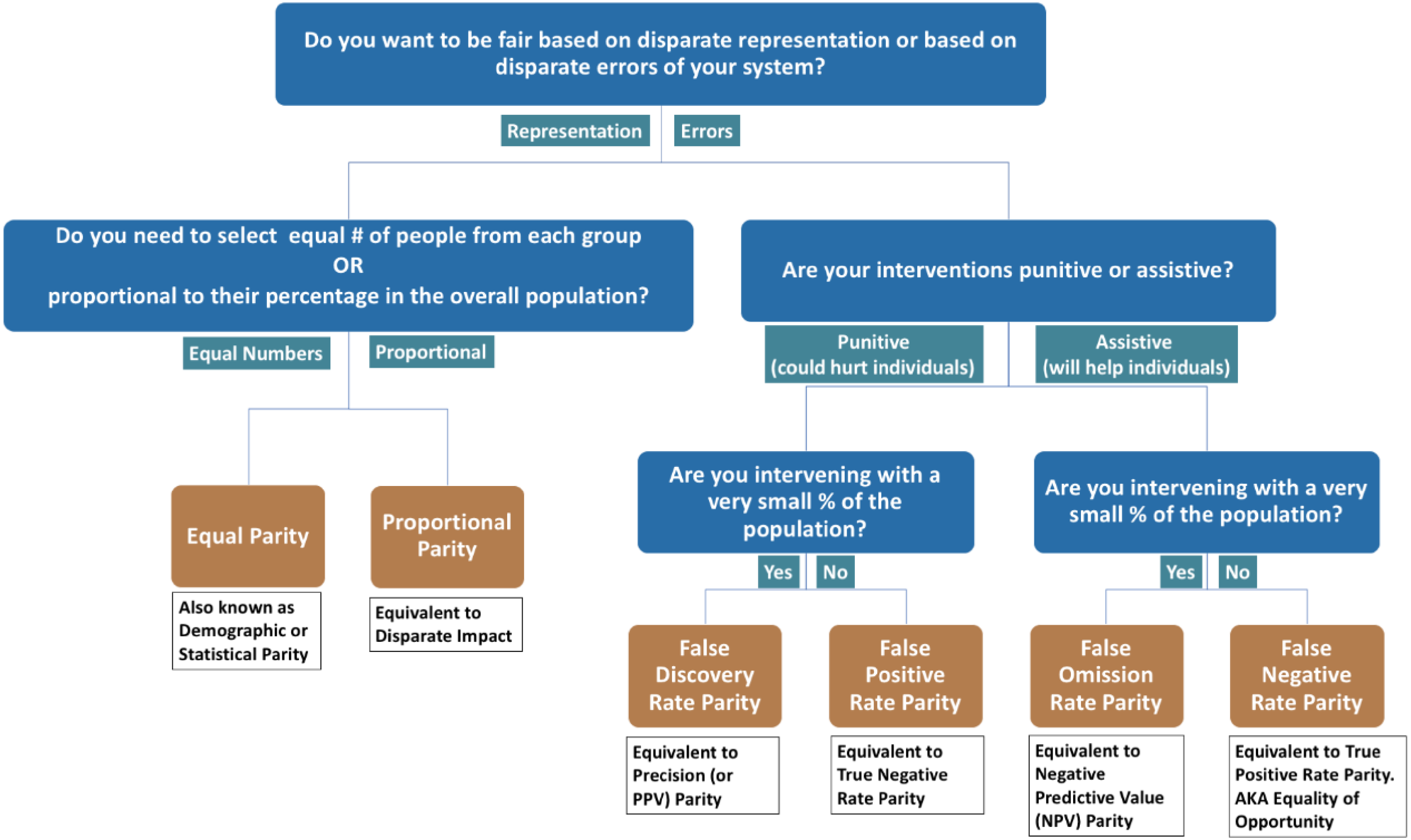
- ***Postprocessing***: Postprocessing methods work by calibrating the predictions after the model has been trained.

- ***In-processing***: These methods incorporate fairness directly into the model design.

Fairness Is Context-based

- Who are the stakeholders of the system?
  - Which of these groups could be harmed?
- What potential harms can be caused by biased decisions?
  - e.g., unfair punishments, denial to resources
- Are there any legal constraints or policy goals?
  - e.g., 80% rule, affirmative actions
- How are these decisions related to the ML model? Errors?
  - e.g., false positives, false negatives
- Which fairness metric minimizes the harm?

FAIRNESS TREE



Do you want to be fair based on disparate representation or based on disparate errors of your system?

Representation

Do you need to select equal # of people from each group
OR
proportional to their percentage in the overall population?

Equal Numbers

Equal Parity

Also known as Demographic or Statistical Parity

Proportional

Proportional Parity

Equivalent to Disparate Impact

Errors

Are your interventions punitive or assistive?

Punitive
(could hurt individuals)

Are you intervening with a very small % of the population?

Yes

False Discovery Rate Parity

Equivalent to Precision (or PPV) Parity

No

False Positive Rate Parity

Equivalent to True Negative Rate Parity

Assistive
(will help individuals)

Are you intervening with a very small % of the population?

Yes

False Omission Rate Parity

Equivalent to Negative Predictive Value (NPV) Parity

No

False Negative Rate Parity

Equivalent to True Positive Rate Parity. AKA Equality of Opportunity

Decision tree created by The University of Chicago useful in thinking how organizations decide fairness

Fairness Is A Challenge

- 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

Source: Holstein, Kenneth, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. " Improving fairness in ML systems: What Do Industry Practitioners Need?" " In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1-16. 2019.

Sample Use case: Automated Hiring

- Which group can possibly harm by biased decisions?
- What kind of harm can be caused?
- Any legal constraints
- Which fairness metric to use?
  - Independence, separation w/FPR vs. FNR
- Check A Case Study of Automated Video Interviews