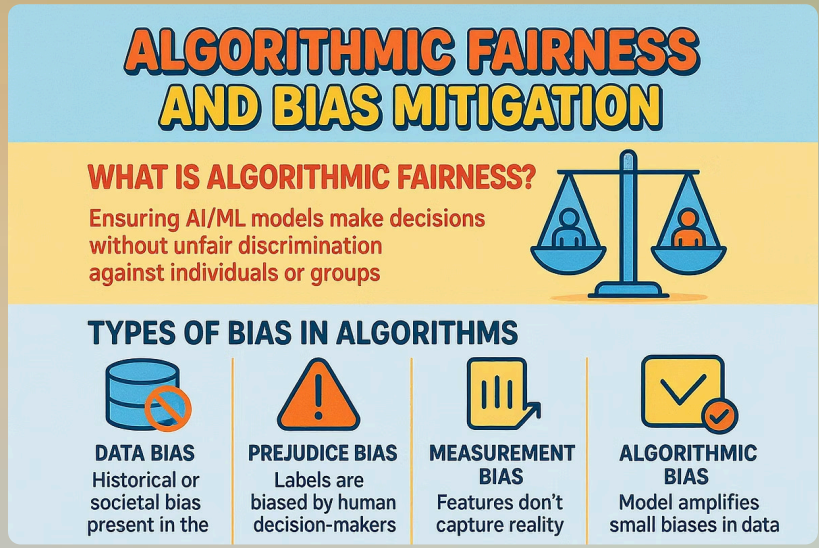
Module II: Algorithmic Fairness and Bias Mitigation



What is Bias in Algorithms?

Algorithmic bias occurs when an algorithm produces results that are systematically

due to **flawed assumptions, biased training data, or design choices**.

- For example: A hiring algorithm trained mostly on resumes of men may unfairly prefer male candidates over equally qualified women.

Bias can creep in unintentionally, but its impact can be **serious**, especially in sensitive domains like **healthcare, hiring, lending, law enforcement, or education**.

Sources of Bias in Algorithms

Bias doesn’t come from “the machine itself” but from the way data and models are created:

1

Data Bias

- Training data may not be representative.
- Example: A face recognition model trained mostly on lighter-skinned faces performs poorly on darker-skinned faces.

2

Historical/Societal Bias

- Data reflects existing inequalities in society.
- Example: Historical arrest records reflect racial profiling → model predicts higher “risk” for certain groups.

3

Measurement Bias

- When a proxy variable is used instead of the true target.
- Example: Using “number of hospital visits” to measure “health needs” disadvantages people with less access to healthcare.

4

Algorithmic Bias

- Choice of model, features, or optimization goals may favor one group over another.

5

Human Bias in Design

- Developers’ assumptions, labeling practices, and priorities can shape the algorithm unfairly.

Fairness in Algorithms

Fairness means ensuring that an algorithm's decisions are **not systematically discriminatory** against individuals or groups, particularly those defined by **sensitive attributes** such as:

- Gender
- Race/Ethnicity
- Age
- Disability
- Socio-economic background

Fairness is complex because **different definitions of fairness can conflict** with one another.

Understanding Fairness Bias in AI Models

Machine learning models must not exhibit unfair or discriminatory behavior, exemplified through biased predictions or discriminatory decision-making. This occurs when models unfairly favor or discriminate against certain groups or classes, resulting in unequal treatment and disparate outcomes.

Biases in fairness can stem from multiple origins, such as discriminatory training data, biased assumptions in model design, or the inclusion of biased features. The model can potentially replicate existing biases if the training data is skewed or reflects historical discrimination. Likewise, the model may display little tendencies if specific features are given more significance during training.

The consequences of fairness bias can be detrimental, as it can perpetuate and amplify existing societal inequalities or discrimination, which various compliances like HIPAA may not even detect. For example, in the context of hiring decisions, a biased model might favor candidates from a particular gender, race, or socioeconomic background, leading to unfair outcomes and exclusion of qualified individuals.

Addressing fairness bias is crucial to ensure ethical and equitable machine learning applications. Various approaches, such as fairness metrics, pre-processing techniques, algorithmic modifications, and post-processing interventions, can be employed to mitigate and monitor bias in machine learning models. These techniques aim to promote fairness, transparency, and accountability in the decision-making processes of AI systems.

Examples of Fairness Bias

Artificial Intelligence (AI) models, touted for their ability to automate decision-making processes, have become integral in various domains. However, the unchecked integration of AI can lead to unintended consequences, such as fairness bias. Fairness bias occurs when AI models exhibit discriminatory behavior or amplify existing societal inequalities.

Let us see some examples in real life where your Bias model can ultimately dictate or even destroy human lives if not rectified and trained properly:

1

Sentencing Disparities: AI models are increasingly used in the criminal justice system to aid in sentencing decisions. However, studies have shown that AI-powered algorithms can perpetuate racial biases. For instance, an AI model may unknowingly be trained on historical data that disproportionately criminalize certain racial groups. Consequently, the model may recommend harsher sentences for individuals from these groups, exacerbating existing disparities. This fairness bias can perpetuate systemic injustice, leading to detrimental consequences for affected communities.

2

Biased Hiring Practices: AI-powered tools often screen and shortlist job applicants in recruitment processes. However, if the training data used to develop these models reflects biased hiring patterns, the AI system can inadvertently discriminate against specific demographics. For example, if historically male-dominated industries predominantly feature male employees, the AI model may learn to favor male candidates, perpetuating gender biases. Such biased hiring practices hinder diversity and inclusivity, restricting equal opportunities for marginalized groups.

3

Predatory Lending Algorithms: In the financial sector, AI algorithms are employed to assess creditworthiness and determine loan approvals. However, if these models are built using partial historical lending data, they can inadvertently discriminate against underprivileged communities. For instance, if the training data indicates that certain minority groups have been unfairly denied loans, the AI model may adopt this discriminatory behavior, perpetuating the cycle of financial exclusion. This fairness bias deepens socio-economic disparities and limits access to resources for marginalized communities.

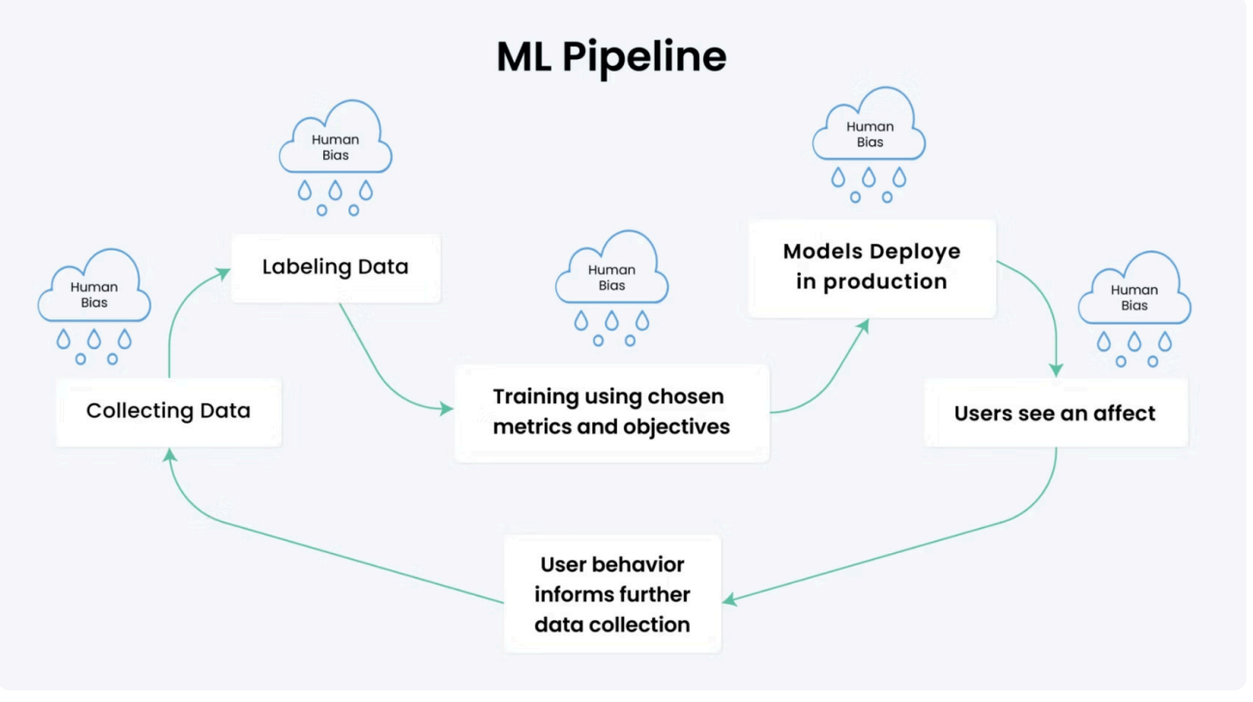
4

Healthcare Disparities: AI models are increasingly utilized in healthcare settings to aid in diagnosing diseases and recommending treatments. These models can exhibit fairness bias if trained on partial healthcare data. If historically marginalized communities have faced inadequate healthcare access, an AI model trained on such data may perpetuate these disparities. This can lead to delayed or incorrect diagnosis and adverse health outcomes.

Common Fairness Metrics in Machine Learning

Data has proven itself to be the most valuable resource recently, from the billions of dollars earned by research institutes (Microsoft invested $10 Billion in OpenAI before the launch of ChatGPT) to complete organizations coming down to their knees through unfair use of data. It should be poetic that an implosion of such a large-scale disparity in machine learning models through bias to be dealt with and rectified from the data itself.

Metrics is a term thrown around a lot in machine learning, mainly in accuracy (F1 Score, Confusion Matrix, AUC, Loss Functions, etc.). Metrics help users understand the performance, efficiency, and eventual expectations from a machine learning model trained on intricate and confusing architectures, which otherwise may not be as intuitive for the average working professional and more so suited to PhD-level scholars.



Types of fairness in AI

In academic research, there are several types of fairness proposed, including:

- **Group fairness:** Ensuring that different groups, and subdivided groups, are treated equally or proportionally. So, outcomes are distributed evenly across the different groups.
- **Individual fairness:** Ensuring that similar individuals are treated similarly, regardless of group membership, such as distance-based measures.
- **Counterfactual fairness:** Ensuring that AI systems are fair even in a hypothetical scenario.
- **Procedural fairness**: Ensuring that the decision-making process is fair and transparent.

Fairness

In order to utilize these metrics, you should have a fundamental understanding of your business problem and know the following: protected attributes, privileged group, favorable label, and case type. These combined are the fairness definition for a specific model use case.

Note: These definitions are only being used as an example for our use case. In real world these would need to be set by the business that owns the model in accordance with legal and business standards.

Protected attribute(s): An attribute that partitions a population into groups whose outcomes should have parity. Examples include race, gender, caste, and religion. Protected attributes are not universal, but are application specific.

- For our use case, we have chosen to look at two protected attributes: **sex** and **race.**

Privileged group: A value of a protected attribute indicating a group that has historically been at a systematic advantage. It can be difficult to ascertain which protected individuals belong to each group. Stakeholders should have a deep understanding of their domain to recognize privileged and unprivileged groups within protected categories. Statistical methods can be utilized to understand the division in protected attributes. For instance, continuous variables such as age can be split into buckets. Along with this, races can be combined to make different race categories such as Caucasian and Not Caucasian. Intersectionality may also be investigated to determine if the combination of subgroups is at risk of unfairness.

- **Sex**, privileged: *Female*, unprivileged: *Male*
- *Note: In this use case, the privilege group is female; however, in other use cases females may be underprivileged, highlighting the importance of domain expertise.*
- **Race**, privileged: *Caucasian*, unprivileged: *Not Caucasian*
- *Note: Fairness metrics calculations are only performed for race in this article, but it can be replicated for other protected attribute as well (in this case sex).*

Favourable label: A label whose value corresponds to an outcome that provides an advantage to the recipient. The opposite is an unfavourable label.

Metrics of Fairness

Researchers use several metrics to measure fairness in machine learning. Some key ones:

Demographic Parity (Statistical Parity)

- The decision (e.g., being hired) should be **independent of sensitive attributes** like gender or race.
- Example: If 50% of male applicants are hired, 50% of female applicants should also be hired.

Equal Opportunity

- All groups should have **equal true positive rates**.
- Example: If qualified male candidates are selected 90% of the time, qualified female candidates should also be selected 90% of the time.

Equalized Odds

- Both **true positive rate** and **false positive rate** should be equal across groups.

Calibration Fairness

- For people predicted with the same probability score (e.g., “70% chance of repaying loan”), the actual outcomes should be equal across groups.

Why Fairness is Difficult

- **Trade-offs exist**: You can’t satisfy all fairness definitions at once (mathematical impossibility).
- **Context matters**: Fairness in lending may be defined differently than fairness in healthcare.
- **Utility vs. fairness**: Sometimes optimizing fairness can reduce overall accuracy.

Examples in the Real World

1. **Hiring Tools**: Amazon scrapped an AI recruitment tool that was biased against women, since it learned from past hiring patterns dominated by men.
2. **Criminal Justice (COMPAS Algorithm)**: Used in the US to predict likelihood of reoffending, but found to unfairly give higher risk scores to African-American defendants.
3. **Facial Recognition**: Studies showed higher error rates in recognizing darker-skinned and female faces.

Comparison of Bias and Fairness in AI

Bias	Fairness
Refers to systematic and consistent deviation of an output from true value/what would be expected.	The absence of favoritism and discrimination in an AI system's decisions.
Can be unintentional.	Inherently deliberate and intentional.
Arises due to various factors like biased data or design.	Requires a conscious effort to ensure the algorithm doesn't discriminate.
Often detected through analysis of outcomes or patterns.	Ensured through proactive design, monitoring and auditing of AI systems.

Basic methods for mitigating bias in algorithms and data

- **Pre-processing methods (before training the model)**

Goal: Remove or reduce bias directly from the dataset.

Re-sampling: Balance the dataset by over-sampling underrepresented groups or under-sampling overrepresented groups.

Re-weighting: Assign higher weights to minority or disadvantaged groups during training.

Data transformation: Modify input features so they are less correlated with sensitive attributes (e.g., gender, race).

Fair representation learning: Create new representations of data that hide or reduce sensitive information.

- **In-processing methods (during model training)**

Goal: Modify the learning algorithm to ensure fairness.

Fair regularization: Add fairness constraints or penalties (e.g., demographic parity, equalized odds) to the loss function.

Adversarial debiasing: Train the model with an adversary that tries to predict sensitive attributes—forcing the main model to ignore them.

Fairness-aware optimization: Adjust optimization to balance accuracy and fairness simultaneously.

- **Post-processing methods (after model training)**

Goal: Adjust model predictions to satisfy fairness without retraining.

Calibration: Adjust predicted probabilities separately for different groups to achieve fairness metrics.

Reject option classification: For uncertain predictions, favor disadvantaged groups.

Threshold adjustment: Set different decision thresholds for different groups to balance error rates.

Output perturbation: Randomly flip or modify outputs to achieve statistical fairness.

- **Fairness-aware algorithms:**

This approach codes in rules and guidelines to ensure that the outcomes generated by AI models are equitable to all individuals or groups involved.

- **Auditing and transparency:**

Human oversight is incorporated into processes to audit AI-generated decisions for bias and fairness. Developers can also provide transparency into how AI systems arrive at conclusions and decide how much weight to give those results. These findings are then used to further refine the AI tools involved.

How Do You Address Fairness and Bias in AI?

Data strategy

Having a robust AI data strategy, including ensuring the training data you use has a wide range of demographics and experiences to minimize data bias and support AI fairness.

Governance

Establishing strong AI governance frameworks that ensure your AI is developed and deployed following best practices and ethical guidelines, including oversight, accountability and monitoring to help AI systems remain fair and unbiased. You can consider a solution like SS&C | Blue Prism® **Enterprise AI**, which can help oversee and govern AI and technology in your processes.

Feedback loop

Implementing a feedback loop allows you to continuously improve your AI systems. You can encourage users and stakeholder feedback, who can help identify and correct missed post-processing biases, helping your system evolve over time to become fairer.

Regulation

Complying with existing AI regulations helps enforce fairness and accountability. Fairness is already a central principle and legal requirement of data protection law, even though AI brings additional complexities compared to conventional processing.

Tools

There are also a few tools designed to tackle AI fairness, such as **IBM's AI Fairness 360** toolkit that helps users examine, report and mitigate discrimination and bias in ML models. There are also similar solutions from Microsoft and Google.