AI Bias and Fairness: Terms to Know

This document helps with material from the last Module covering common problems in AI training data and key concepts related to fairness in AI.

Video: "AI Ethics: Understanding Bias and Fairness in Your Models" (new tab)Opens in new tab (30:30)

Article: "What Do We Do About the Biases in AI?"

Common Problems in AI Training Data

There are four main types of problems that can lead to bias in AI training data:

- Skewed Sample
- Limited Features / Sample Size Disparity
- Tainted Examples
- Proxy Bias

Skewed Sample

A skewed sample occurs when the training data doesn't fairly represent all groups or situations.

Color Example: In the training data, whenever articles mention colors, green is used 90% of the time, while all other colors combined only appear in 10% of color mentions.

Real-world Example: If an AI is trained mostly on English text from North America, it might not work well with content from other regions or cultures.

Consequences

The AI might work better for some groups than others - Minority viewpoints might be ignored or misunderstood

How to Fix It

Use data from many different sources - Make sure all important groups are well represented in the data

Limited Features / Sample Size Disparity

This problem occurs when some types of data are much more common than others in the training set.

Color Example: The training data has millions of examples of the color green in various contexts (like nature, fashion, art), but only a few hundred examples of other colors.

Real-world Example: An AI might have access to millions of formal English writing samples but very few examples of casual or spoken English.

Consequences

The AI might perform poorly when dealing with less common topics or groups - It might be overconfident about things it sees a lot in the training data

How to Fix It

Balance the amount of data for different groups or categories - Add more examples of underrepresented groups or topics

Tainted Examples

This happens when the training data includes biased or incorrect information.

Color Example: The training data includes many false statements about the color green, such as "Green is the only color that represents nature" or "Green is universally the most calming color for all cultures."

Real-world Example: If an AI learns from old texts with outdated views about certain groups of people, it might repeat these biased ideas.

Consequences

The AI might spread false information - It could reinforce harmful stereotypes

How to Fix It

Carefully check and clean the training data - Use filters to remove biased or incorrect information

Proxy Bias

This occurs when seemingly neutral information in the data is actually linked to sensitive topics.

Color Example: In the training data, descriptions of wealthy neighborhoods often mention "lush green lawns," while descriptions of lower-income areas rarely mention green spaces.

Real-world Example: An AI might learn to associate certain universities (which could be linked to race or social class) with job suitability, even if race and class aren't directly mentioned in the data.

Consequences

The AI might make unfair decisions based on hidden biases - It could accidentally discriminate against certain groups

How to Fix It

Look closely at how different pieces of information are related in the data - Remove or adjust information that might indirectly cause bias

Other Key Terms:

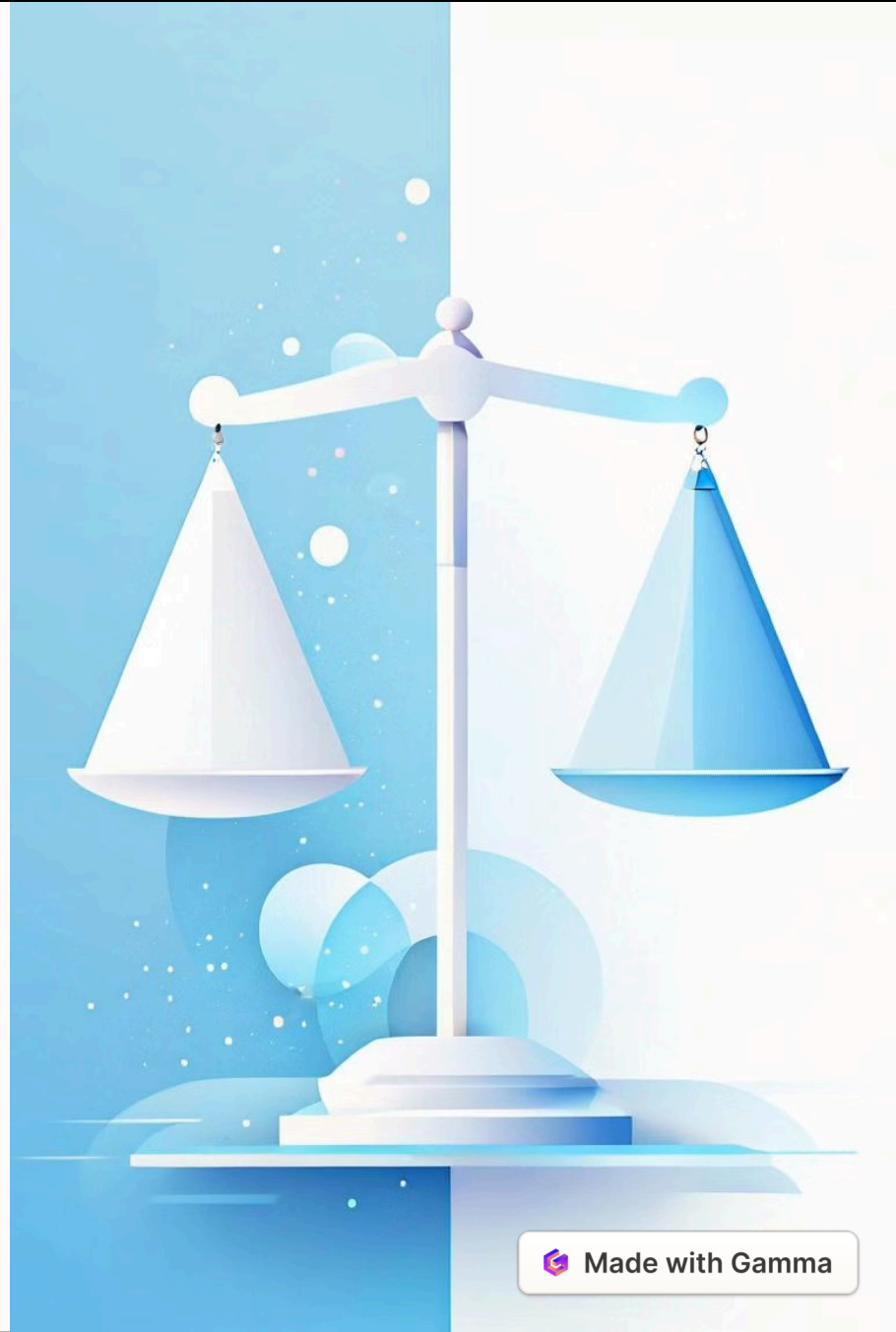
To address bias and promote fairness in LLMs, it's important to understand these key concepts:

- Fairness Metric
- Protected Class / Protected Feature(s)
- Pre-processing
- In-processing
- Post-processing

Fairness Metric

A fairness metric is a way to measure how fair an AI system is. It helps us check if the AI treats different groups equally.

Example: One fairness metric might check if the AI gives the same quality of answers to questions about different cultures or genders.



Protected Class vs. Protected Feature

Protected Class:

A general category of people who share a characteristic that is legally protected against discrimination.

Example: Race, gender, age, and religion are often considered protected classes.

Protected Feature:

The specific instances or attributes within a protected class. These are the particular things about a person that shouldn't influence the AI's decisions unfairly.

Example: In a job application system, a person's specific age (like 42 years old) or gender (like female) would be protected features.

Key Difference: While a protected class is a broad category (like age or gender), a protected feature is the specific instance of that category for an individual (like being 42 years old or being female).

Bias Mitigation Strategies

There are three main approaches to mitigating bias in AI systems:

Pre-processing

Changing or fixing the training data before using it to train the AI.

Example: Removing gender-specific words from job descriptions in the training data to prevent gender bias in a hiring AI.

In-processing

Changing how the AI learns during training to make it fairer.

Example: Using special math techniques during training to make sure the AI treats all groups equally.

Post-processing

Changing the AI's outputs after it has made a decision, to make the results fairer.

Example: Adjusting the AI's language suggestions to ensure equal representation of different groups in generated text.