ONE PAGE SUMMARY

IT & Software › Machine Learning › Bias and Fairness

Introduction to Bias and Fairness

Understanding Bias and Fairness in Machine Learning

DEFINITION

> Bias in machine learning refers to systematic errors that lead to unfair outcomes. Fairness ensures that models treat all individuals or groups equitably.

KEY CONCEPTS

Types of Bias

Bias can be introduced through data, algorithms, or human decisions, affecting outcomes.

Fairness Metrics

Metrics like demographic parity and equal opportunity help measure fairness in models.

Algorithmic Fairness

This involves designing algorithms that minimize bias and promote equitable treatment.

Data Representation

Ensuring diverse and representative data is crucial to reducing bias in machine learning.

Mitigation Strategies

Techniques like re-weighting data or modifying algorithms can help address bias.

EXAMPLES

→ Facial recognition systems misidentifying individuals from minority groups.

→ Loan approval algorithms favoring certain demographics over others.

→ Job recruitment tools prioritizing candidates based on biased historical data.

MEMORY TIPS

★ Think 'B-Fair' to remember Bias and Fairness.

★ Use the acronym 'F.A.I.R.' for Fairness: Fair, Accountable, Inclusive, Respectful.

★ Visualize a balanced scale to represent fairness in outcomes.

COMMON MISTAKES

✗ Assuming all algorithms are inherently unbiased.

✗ Neglecting the importance of diverse training data.

✗ Overlooking the need for continuous monitoring of model performance.

QUICK RECAP

Bias and fairness are critical in machine learning, affecting how models perform across different groups. Understanding types of bias and implementing fairness metrics can help create equitable systems.