Biases in Training Data

Lesson Downloads

Show Transcript **Summarize Video**

Biases in training data deeply influence the outcomes of AI models, reflecting societal issues that require attention. Ways to approach this challenge include promoting diversity in development teams, seeking diverse data sources, and ensuring continued vigilance through bias detection and model monitoring.

Technical Terms Explained:

Selection Bias: When the data used to train an AI model does not accurately represent the whole population or situation by virtue of the selection process, e.g. those choosing the data will tend to choose dataset their are aware of
Historical Bias: Prejudices and societal inequalities of the past that are reflected in the data, influencing the AI in a way that perpetuates these outdated beliefs.
Confirmation Bias: The tendency to favor information that confirms pre-existing beliefs, which can affect what data is selected for AI training.
Discriminatory Outcomes: Unfair results produced by AI that disadvantage certain groups, often due to biases in the training data or malicious actors.
Echo Chambers: Situations where biased AI reinforces and amplifies existing biases, leading to a narrow and distorted sphere of information.
Bias Detection and Correction: Processes and algorithms designed to identify and remove biases from data before it's used to train AI models.
Transparency and Accountability: Openness about how AI models are trained and the nature of their data, ensuring that developers are answerable [illegible] AI's performance and impact.

Quiz Question
Which initiative boosts fair AI development [illegible]flecting the broade[illegible]ge of human experiences and perspectives?

- Simplifying algorithms
- Limiting data variety
- Increasing organizational diversity
- Decreasing model size
- Using a single data source

Submit