Algorithmic foundations and ethics in AI: from theory to practice course

Toolkit for synchronous sessions

CU4 | Data fairness and bias in AI

Support PowerPoint slides

CC BY 4.0

Project number | 2022-1-ES01-KA220-HED-000085257

INDEX

INTRODUCTION

IMAGE SOURCE | Generated by DALL-E

Suggestion

Start with an **interactive poll or question** to gauge participants' initial understanding or personal experiences with AI fairness and bias. This can make the introduction more engaging and set the stage for why the topic is relevant.

- Example: What is your familiarity with fairness and bias in AI?
  - I'm not familiar with it at all.
  - I've heard about it, but I don't know the details.
  - I understand the basics but I would like to know more.
  - I'm quite knowledgeable and have worked with AI systems considering fairness and bias.

IN THIS COMPETENCE UNIT YOU WILL FIND THE FOLLOWING SUBJECTS:

- The basics of data fairness and bias in AI.
- Various biases that can influence AI systems.
- How to spot and measure biases in data.
- Techniques to reduce bias and increase fairness.
- Metrics to assess AI fairness.
- Best practices for fair data collection and processing.
- Emerging trends in AI fairness. fairness and bias in AI.

AT THE END OF THE COMPETENCE UNIT, YOU SHOULD BE ABLE TO:

- Explain the fundamentals of data fairness and the significance of bias in AI.
- Identify the various forms of bias that can affect AI systems.
- Evaluate AI systems using fairness metrics.
- Follow best practices for collecting and preprocessing data to promote fairness.

INTRODUCTION TO DATA FAIRNESS AND BIAS

IMAGE SOURCE | Generated by DALL-E

Introduction to data fairness and bias

Fairness & bias - the candy analogy

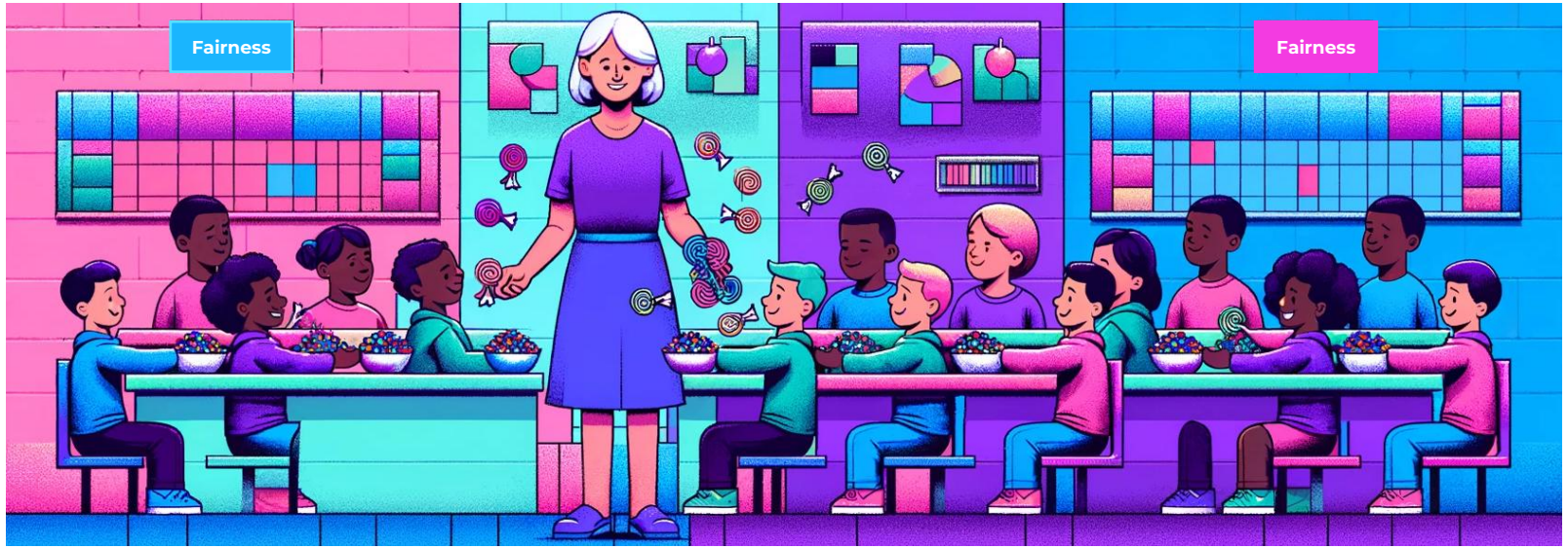


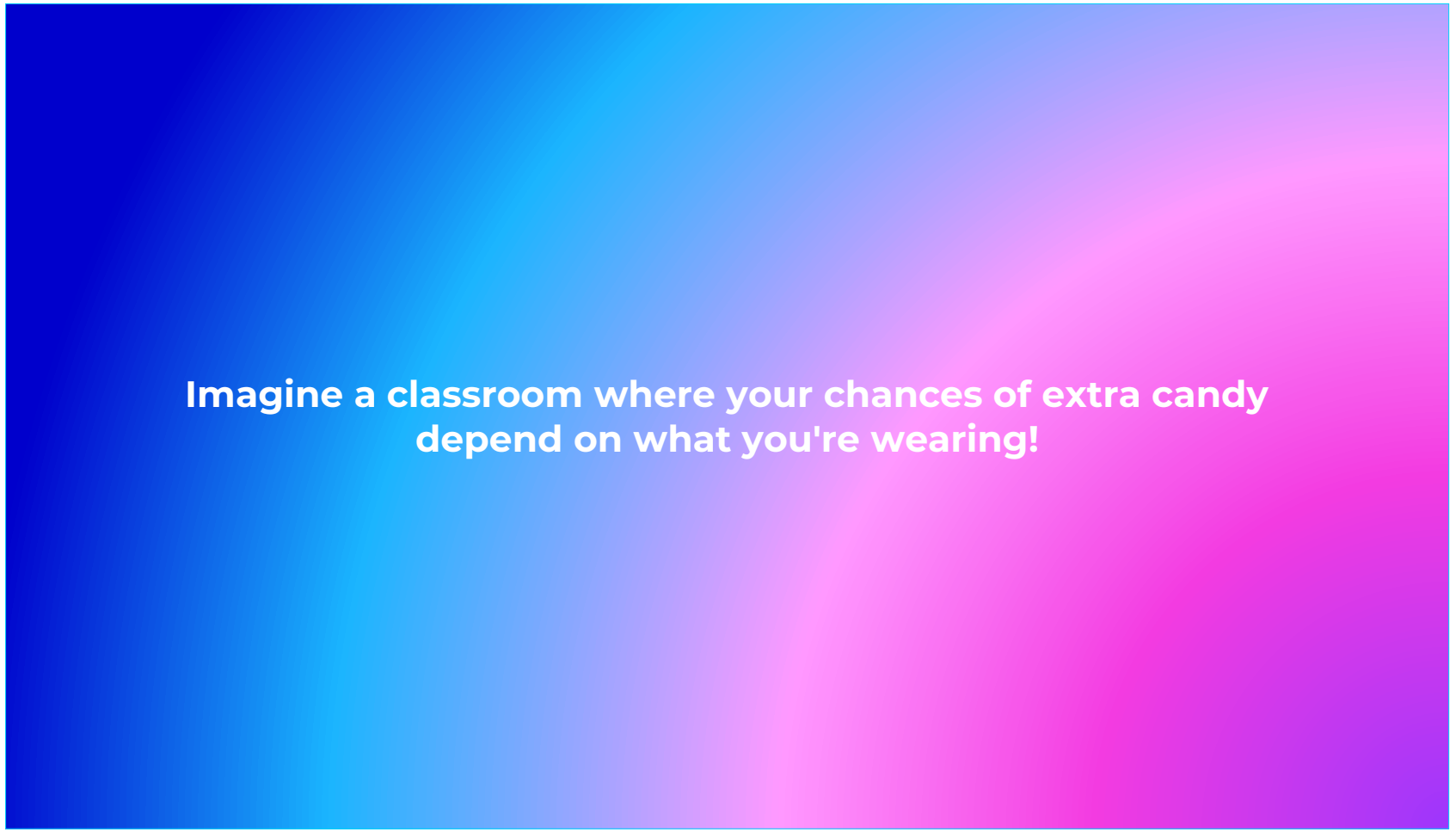
IMAGE SOURCE | Generated by DALL-E

Discover how fairness and bias in AI systems can impact our everyday choices, much like a teacher's rules for handing out candy can shape a classroom experience.

Let's unwrap the concept of AI equity - one sweet example at a time!

Introduction to data fairness and bias

Fairness & bias - the candy analogy



Introduction to data fairness and bias

Fairness & bias - the candy analogy

UNFAIR ADVANTAGE

Students in red shirts get twice as much candy for no real reason—this is **'bias**.'

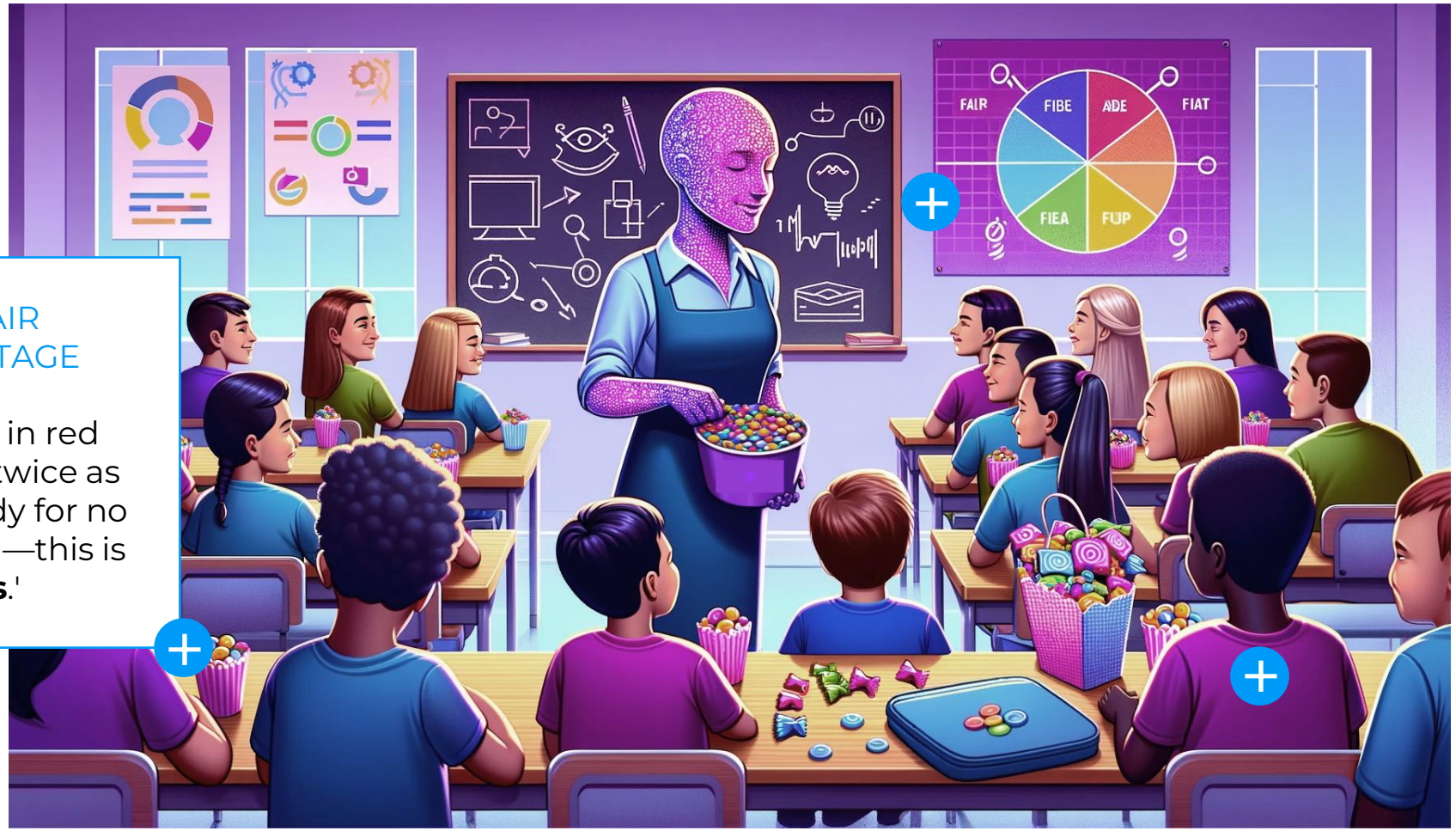
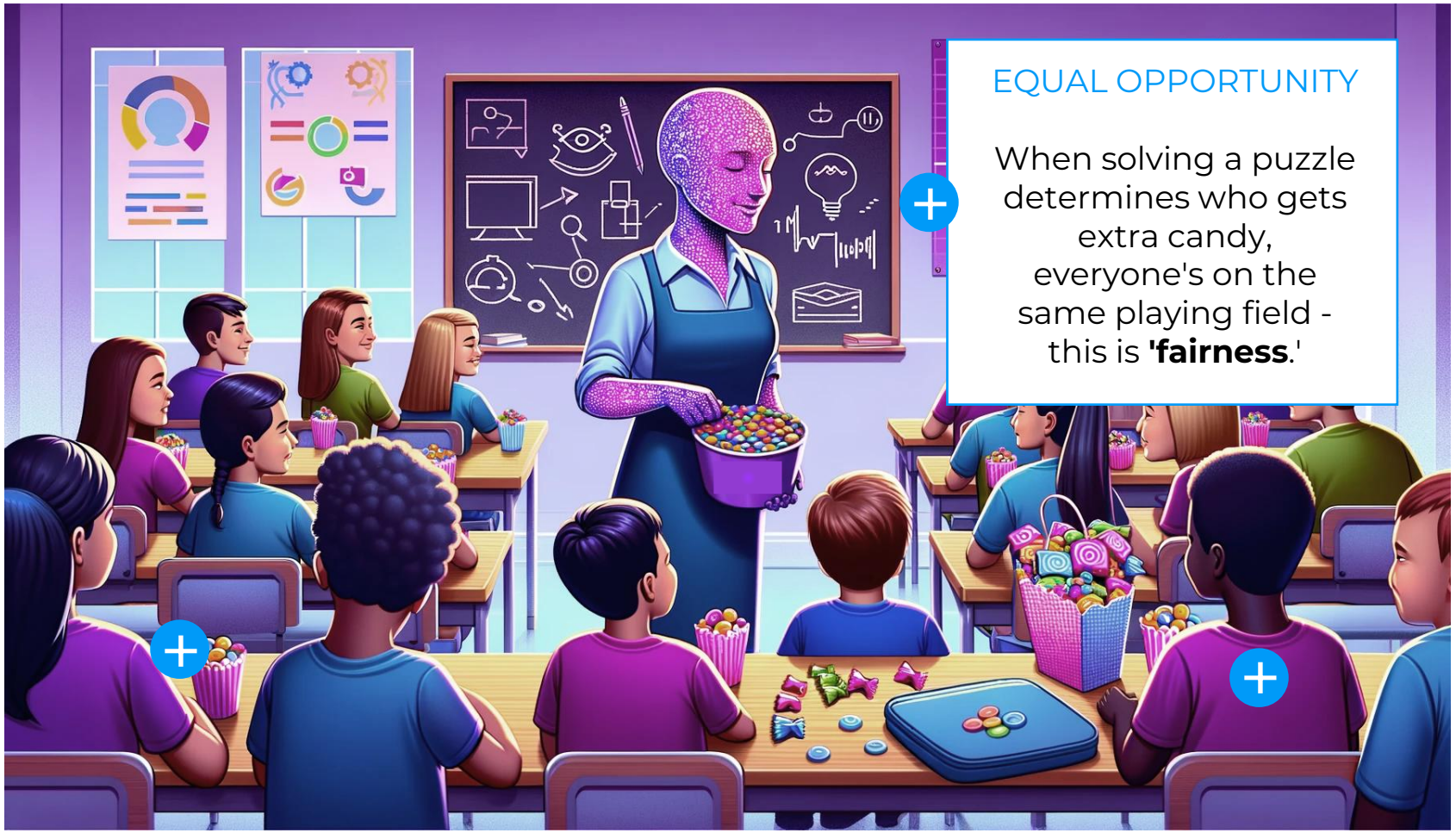


IMAGE SOURCE | Generated by DALL-E

Introduction to data fairness and bias

Fairness & bias - the candy analogy



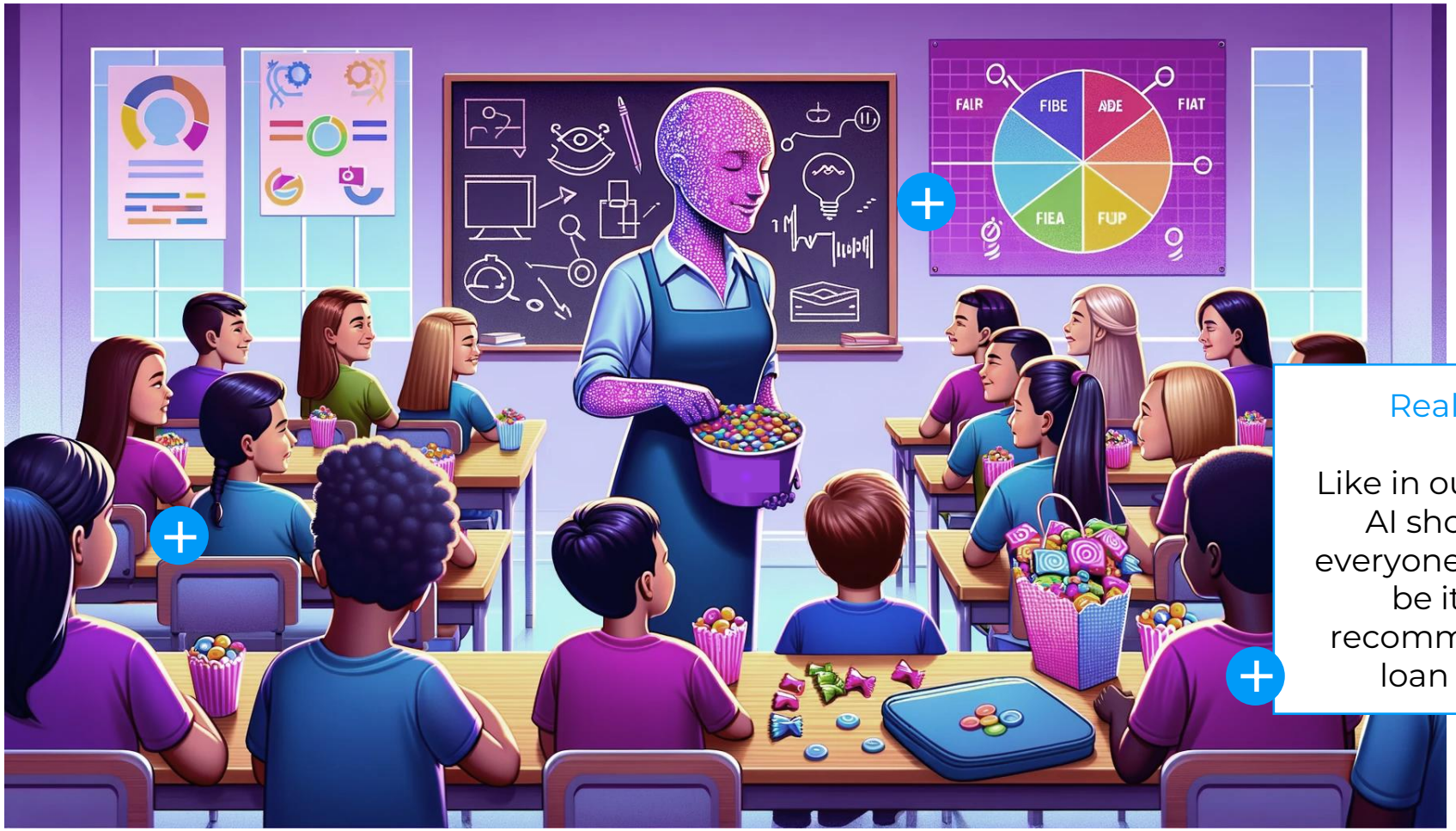
EQUAL OPPORTUNITY

When solving a puzzle determines who gets extra candy, everyone's on the same playing field - this is **'fairness**.'

IMAGE SOURCE | Generated by DALL-E

Introduction to data fairness and bias

Fairness & bias - the candy analogy



Real-World AI

Like in our candy story, AI should ensure everyone gets a fair go, be it in movie recommendations or loan approvals.

IMAGE SOURCE | Generated by DALL-E

Introduction to data fairness and bias

Data fairness

Data fairness is about ensuring that all individuals are treated equitably by AI systems, without bias affecting their outcomes. It involves the processes and practices that go into collecting, processing, and using data to prevent discriminatory effects.

Data fairness is a crucial aspect of designing and implementing AI systems that positively impact society. It ensures AI decisions are just and unbiased, promoting equality and preventing discrimination.

Fair data practices lead to trustworthy AI applications that can be widely accepted and beneficial across diverse communities.

It aims for AI that serves everyone equitably, regardless of background or demographic characteristics.

Introduction to data fairness and bias

- What is bias? +
- How it occurs? +
- What are the consequences of bias? +

Introduction to data fairness and bias

What is bias?

Bias in AI refers to systematic errors that unfairly favor one group over others. This can arise from skewed data, flawed algorithms, or prejudiced training processes.

How it occurs? +

What are the consequences of bias? +

Introduction to data fairness and bias

What is bias? +

How it occurs? -

Data bias: when the data fed into AI systems reflect historical inequalities or incomplete representations of diverse groups.
Algorithm bias: when the rules or models used to make decisions reinforce existing prejudices or overlook minority patterns.
Feedback loops: when biased AI decisions influence future data, perpetuating the cycle of bias.

What are the consequences of bias? +

Introduction to data fairness and bias

What is bias? +

How it occurs? +

What are the consequences of bias? -

Social impact: can lead to perpetuating stereotypes, reinforcing social inequalities, and eroding trust in essential institutions.

Individual impact: people may face unjust treatment based on flawed AI decisions in critical areas like hiring, law enforcement, and loan approvals.

Introduction to data fairness and bias

Real-world scenarios: AI bias

Recruitment tool bias

Scenario

In 2018, it was reported that Amazon had scrapped an AI recruitment tool that showed bias against women. The AI system was trained on resume data collected over a 10-year years, predominantly from men, reflecting male dominance in the tech industry. Consequently, the AI taught itself that male candidates were preferable, penalizing resumes that included words like "women's," as might appear in "women's chess club captain.

Impact

This example highlights how AI can perpetuate historical biases if the training data is biased. It also shows the potential for AI to impact employment opportunities and the importance of ensuring that AI recruitment tools promote diversity and fairness.

Introduction to data fairness and bias

Real-world scenarios: AI bias

Healthcare algorithm bias

Scenario

A 2019 study published by Science revealed that an algorithm used across many U.S. healthcare systems to allocate healthcare resources was biased against Black patients. The algorithm predicted healthcare needs based on healthcare costs, inadvertently assuming that lower healthcare costs meant lower healthcare needs. However, historically, Black patients incurred fewer costs due to various barriers to accessing care, not because they were healthier. As a result, the algorithm favored healthier white patients over sicker Black patients when identifying candidates for additional care programs.

Impact

This bias in the algorithm resulted in Black patients receiving significantly less access to care programs designed to help manage complex health conditions, even though their need for such interventions was higher. This scenario underscores the critical need for fairness in healthcare AI applications, where biased decision-making can directly impact lives and well-being.

TYPES OF BIAS IN AI SYSTEMS

IMAGE SOURCE | Generated by DALL-E

Types of bias in AI systems

Selection bias

Click cards to flip 

Sampling bias

Click cards to flip 

Measurement bias

Click cards to flip 

Algorithmic bias

Click cards to flip 

Confirmation bias

Click cards to flip 

Reporting bias

Click cards to flip 

Types of bias in AI systems

Selection bias occurs when the data sample used for analysis is not representative of the population it's supposed to represent. It leads to skewed or inaccurate conclusions.

Sampling bias

Click cards to flip

Measurement bias

Click cards to flip

Algorithmic bias

Click cards to flip

Confirmation bias

Click cards to flip

Reporting bias

Click cards to flip

Types of bias in AI systems

Selection bias occurs when the data sample used for analysis is not representative of the population it's supposed to represent. It leads to skewed or inaccurate conclusions.

Sampling bias is a subset of selection bias and arises when the method used to collect data favors certain types of individuals or groups over others.

Measurement bias

Click cards to flip

Algorithmic bias

Click cards to flip

Confirmation bias

Click cards to flip

Reporting bias

Click cards to flip

Types of bias in AI systems

Selection bias occurs when the data sample used for analysis is not representative of the population it's supposed to represent. It leads to skewed or inaccurate conclusions.

Sampling bias is a subset of selection bias and arises when the method used to collect data favors certain types of individuals or groups over others.

Measurement bias occurs due to errors or inconsistencies in the measurement process, leading to inaccurate or misleading results. It can stem from faulty instruments, ambiguous questions, or observer bias.

Algorithmic bias

Click cards to flip

Confirmation bias

Click cards to flip

Reporting bias

Click cards to flip

Types of bias in AI systems

Selection bias occurs when the data sample used for analysis is not representative of the population it's supposed to represent. It leads to skewed or inaccurate conclusions.

Sampling bias is a subset of selection bias and arises when the method used to collect data favors certain types of individuals or groups over others.

Measurement bias occurs due to errors or inconsistencies in the measurement process, leading to inaccurate or misleading results. It can stem from faulty instruments, ambiguous questions, or observer bias.

Algorithmic bias arises in machine learning algorithms when they systematically discriminate against certain groups or individuals based on race, gender, or other protected characteristics present in the training data.

Confirmation bias

Click cards to flip

Reporting bias

Click cards to flip

Types of bias in AI systems

Selection bias occurs when the data sample used for analysis is not representative of the population it's supposed to represent. It leads to skewed or inaccurate conclusions.

Sampling bias is a subset of selection bias and arises when the method used to collect data favors certain types of individuals or groups over others.

Measurement bias occurs due to errors or inconsistencies in the measurement process, leading to inaccurate or misleading results. It can stem from faulty instruments, ambiguous questions, or observer bias.

Algorithmic bias arises in machine learning algorithms when they systematically discriminate against certain groups or individuals based on race, gender, or other protected characteristics present in the training data.

Confirmation bias when researchers or analysts selectively use or interpret data that confirms their preconceived beliefs or hypotheses, while ignoring or dismissing conflicting evidence.

Reporting bias

Click cards to flip

Types of bias in AI systems

Selection bias occurs when the data sample used for analysis is not representative of the population it's supposed to represent. It leads to skewed or inaccurate conclusions.

Sampling bias is a subset of selection bias and arises when the method used to collect data favors certain types of individuals or groups over others.

Measurement bias occurs due to errors or inconsistencies in the measurement process, leading to inaccurate or misleading results. It can stem from faulty instruments, ambiguous questions, or observer bias.

Algorithmic bias arises in machine learning algorithms when they systematically discriminate against certain groups or individuals based on race, gender, or other protected characteristics present in the training data.

Confirmation bias when researchers or analysts selectively use or interpret data that confirms their preconceived beliefs or hypotheses, while ignoring or dismissing conflicting evidence.

Reporting bias refers to the selective publication or reporting of research findings that are statistically significant or favorable while suppressing or neglecting non-significant or unfavorable results.

Types of bias in AI systems
Examples

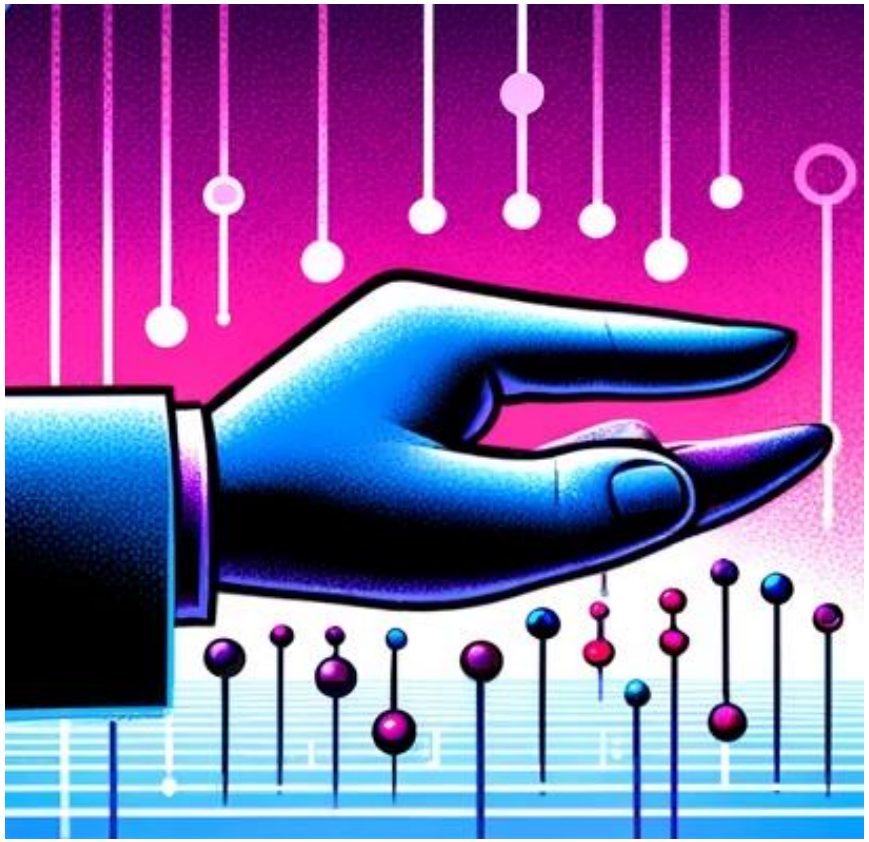


IMAGE SOURCE | Generated by DALL-E

Selection bias

Selection bias | Facial recognition technology

Facial recognition technologies have often been trained predominantly on datasets comprising mostly Caucasian faces. This lack of diversity in training datasets has led to higher error rates in recognizing individuals from other ethnic backgrounds. For instance, a study by MIT Media Lab found significant disparities in the accuracy of gender recognition technologies, particularly misidentifying gender among darker-skinned women.

Types of bias in AI systems
Examples

IMAGE SOURCE | Generated by DALL-E

Sampling bias

Sampling bias | Political polling errors

In the 2016 U.S. Presidential Election, several opinion polls underestimated support for Donald Trump. A key issue was sampling bias, where the demographic weighting in samples did not accurately represent the voter base. Pollsters had over-sampled urban populations and under-sampled rural areas, leading to inaccurate predictions that missed the mark on the actual electoral outcome.

Types of bias in AI systems
Examples

IMAGE SOURCE | Generated by DALL-E Measurement bias

Measurement bias | Pulse oximeters and skin tone

Recent studies, including those heightened by the COVID-19 pandemic, have shown that pulse oximeters can exhibit measurement biases by overestimating blood oxygen levels in people with darker skin. This bias arises because the devices' sensors are less effective at penetrating darker skin, leading to potentially dangerous misdiagnoses or delays in treatment for patients of color.

Types of bias in AI systems
Examples

IMAGE SOURCE | Generated by DALL-E Algorithmic bias

Algorithmic bias | COMPAS in criminal sentencing

Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) software, used in the U.S. criminal justice system to assess the likelihood of a defendant reoffending, was found to exhibit racial biases. Analysis revealed that the software was nearly twice as likely to falsely predict future criminality in Black defendants as compared to white defendants, influencing sentencing and bail decisions based on skewed algorithmic recommendations.

Types of bias in AI systems
Examples

IMAGE SOURCE | Generated by DALL-E Confirmation bias

Confirmation Bias - Credit Scoring Models

Financial institutions using AI-driven credit scoring models might unintentionally reinforce confirmation biases if the models are trained on historical lending data that reflects past prejudices. For instance, if previous human lenders had biases against certain demographic groups, the AI system might replicate these biases by favoring individuals from historically favored groups, assuming they are less risky based on past loan repayments.

Types of bias in AI systems
Examples

IMAGE SOURCE | Generated by DALL-E

Reporting bias

Reporting bias - Drug side effects in online forums

Pharmaceutical companies and health researchers using AI to monitor online forums and social media for drug side effect reports may encounter reporting bias. Patients experiencing severe side effects are more likely to report their experiences online compared to those experiencing mild or no side effects. This can lead AI systems to overestimate the severity or frequency of adverse reactions, potentially influencing drug safety profiles inaccurately.

METHODS FOR IDENTIFYING AND MEASURING BIASES IN DATA

IMAGE SOURCE | Generated by DALL-E

Methods for identifying and measuring biases in data

- Data auditing +
- Disparity analysis +
- Algorithmic fairness testing +
- Sensitivity analysis +

Methods for identifying and measuring biases in data

Data auditing -

Data auditing involves systematically examining datasets to identify any disparities, imbalances, or anomalies that may indicate bias.

This process includes checking for representational equality among different groups within the data. It ensures that the data collection process does not systematically exclude certain populations or reinforce existing stereotypes.

Disparity analysis +

Algorithmic fairness testing +

Sensitivity analysis +

Methods for identifying and measuring biases in data

Data auditing +

Disparity analysis -

Disparity analysis refers to comparing the outcomes or predictions of AI models across different demographic or protected groups to identify any unequal impacts.

It involves calculating statistical measures such as the difference in error rates, acceptance rates, or other outcome metrics between groups. This helps quantify and visualize disparities that may indicate bias in AI decision-making processes.

Algorithmic fairness testing +

Sensitivity analysis +

Methods for identifying and measuring biases in data

Data auditing +

Disparity analysis +

Algorithmic fairness testing -

Algorithmic fairness testing involves implementing tests or metrics that assess the fairness of decisions made by AI systems.

This may include using fairness metrics such as equal opportunity, demographic parity, or individual fairness criteria to evaluate the performance of algorithms. These tests help determine whether AI systems are making fair and unbiased decisions across different demographic groups.

Sensitivity analysis +

Methods for identifying and measuring biases in data

Data auditing +

Disparity analysis +

Algorithmic fairness testing +

Sensitivity analysis -

Sensitivity analysis involves testing the robustness of AI decisions by altering data inputs slightly and observing how outcomes vary for different groups.

By perturbing the data inputs, sensitivity analysis assesses how sensitive AI models are to changes in the input data. It helps identify potential sources of bias and assesses the reliability and consistency of AI predictions across diverse datasets.

Methods for identifying and measuring biases in data

Interactive workshop for identifying and measuring biases in data

Recruitment tool bias

Objective: this workshop will enable participants to apply data auditing, disparity analysis, and sensitivity analysis using a real dataset to identify and understand potential biases.

Tools and materials

Dataset: provide a simple, publicly available dataset with various demographic attributes (e.g., the Adult Income dataset from the UCI Machine Learning Repository, which includes characteristics like age, education, race, gender, and income).

Activity: participants load the dataset and perform basic data exploration (checking for missing values and understanding the distribution of key attributes).

Goal: familiarize participants with the dataset and prepare data for analysis.

Data auditing: conduct a data audit by analyzing the representation of different groups within the dataset. Calculate the proportion of each demographic group (e.g., by gender, race). Identify any underrepresented groups.

Disparity analysis: participants apply disparity analysis to assess how well a simple model (e.g., logistic regression for predicting income level) predicts different groups. Identify any significant disparities in model performance across groups and hypothesize potential causes.

Sensitivity analysis: conduct a sensitivity analysis by slightly altering data points and observing changes in model predictions. Alter attributes related to sensitive demographic variables (e.g., change age or education level slightly) and note how the changes affect model predictions.

STRATEGIES FOR ADDRESSING AND MITIGATING BIASES

IMAGE SOURCE | Generated by DALL-E

Strategies for addressing and mitigating biases

Diverse representation

Click cards to flip 

Bias-aware algorithms

Click cards to flip 

Regular monitoring and evaluation

Click cards to flip 

Transparency and explainability

Click cards to flip 

Strategies for addressing and mitigating biases

Diverse representation

Diverse data points from various demographic groups ensures that AI systems are trained on a representative dataset, reducing the likelihood of biased outcomes. Diversity brings different perspectives and insights, helping identify and mitigate biases during the development process.

Bias-aware algorithms

Click cards to flip

Regular monitoring and evaluation

Click cards to flip

Transparency and explainability

Click cards to flip

Strategies for addressing and mitigating biases

Diverse representation

Diverse data points from various demographic groups ensures that AI systems are trained on a representative dataset, reducing the likelihood of biased outcomes. Diversity brings different perspectives and insights, helping identify and mitigate biases during the development process.

Bias-aware algorithms

Designing algorithms with built-in mechanisms to detect and mitigate biases. This may involve incorporating fairness constraints into the algorithm's optimization process or using techniques like adversarial training to minimize biases in model predictions.

Regular monitoring and evaluation

Click cards to flip

Transparency and explainability

Click cards to flip

Strategies for addressing and mitigating biases

Diverse representation

Diverse data points from various demographic groups ensures that AI systems are trained on a representative dataset, reducing the likelihood of biased outcomes. Diversity brings different perspectives and insights, helping identify and mitigate biases during the development process.

Bias-aware algorithms

Designing algorithms with built-in mechanisms to detect and mitigate biases. This may involve incorporating fairness constraints into the algorithm's optimization process or using techniques like adversarial training to minimize biases in model predictions.

Regular monitoring and evaluation

Continuously monitoring AI systems in real-world settings to identify biases as they emerge. Regular evaluation involves analyzing AI outputs for disparities across different demographic groups and updating algorithms or datasets to address any identified biases promptly.

Transparency and explainability

Click cards to flip

Strategies for addressing and mitigating biases

Diverse representation

Diverse data points from various demographic groups ensures that AI systems are trained on a representative dataset, reducing the likelihood of biased outcomes. Diversity brings different perspectives and insights, helping identify and mitigate biases during the development process.

Bias-aware algorithms

Designing algorithms with built-in mechanisms to detect and mitigate biases. This may involve incorporating fairness constraints into the algorithm's optimization process or using techniques like adversarial training to minimize biases in model predictions.

Regular monitoring and evaluation

Continuously monitoring AI systems in real-world settings to identify biases as they emerge. Regular evaluation involves analyzing AI outputs for disparities across different demographic groups and updating algorithms or datasets to address any identified biases promptly.

Transparency and explainability

Providing transparency and explanations for AI decisions allows stakeholders to understand how decisions are made and identify any biases present in the system. Explainable AI techniques, such as providing feature importance scores or decision rationales, enable users to interpret and assess the fairness of AI outputs.

UNDERSTANDING FAIRNESS METRICS

IMAGE SOURCE | Generated by DALL-E

Understanding fairness metrics

- Statistical parity +
- Equalized odds +
- Conditional demographic disparity +
- Fairness through awareness +
- Individual fairness +
- Counterfactual fairness +

Understanding fairness metrics

Statistical parity

Statistical parity is also known as demographic parity, this metric assesses whether the proportion of positive outcomes (e.g., loan approvals or job offers) is the same across different demographic groups.

Equalized odds +

Conditional demographic disparity +

Fairness through awareness +

Individual fairness +

Counterfactual fairness +

Understanding fairness metrics

Statistical parity +

Equalized odds

Equalized odds examines whether the true positive rate (sensitivity) and the true negative rate (specificity) are equal across different demographic groups. It ensures that the predictive performance of the model is consistent across all groups.

Conditional demographic disparity +

Fairness through awareness +

Individual fairness +

Counterfactual fairness +

Understanding fairness metrics

Statistical parity +

Equalized odds +

Conditional demographic disparity:

Conditional demographic disparity evaluates the disparity in predictive outcomes conditioned on a protected attribute. It measures the difference in prediction outcomes between different demographic groups, accounting for other relevant factors.

Fairness through awareness +

Individual fairness +

Counterfactual fairness +

Understanding fairness metrics

- Statistical parity +
- Equalized odds +
- Conditional demographic disparity +
- **Fairness through awareness**

  Fairness through awareness incorporates awareness of sensitive attributes (e.g., race or gender) into the decision-making process. It ensures that decisions made by AI systems are fair and unbiased, taking into account the historical context and societal impacts.
- Individual fairness +
- Counterfactual fairness +

Understanding fairness metrics

Statistical parity +

Equalized odds +

Conditional demographic disparity +

Fairness through awareness +

Individual fairness

Individual fairness focuses on ensuring that similar individuals receive similar treatment or outcomes from the AI system, regardless of their demographic characteristics. It aims to minimize disparate treatment based on irrelevant attributes.

Counterfactual fairness +

Understanding fairness metrics

- Statistical Parity +
- Equalized Odds +
- Conditional Demographic Disparity +
- Fairness Through Awareness +
- Individual Fairness +

Counterfactual Fairness

Counterfactual Fairness evaluates whether changing an individual's sensitive attribute (e.g., race or gender) while keeping other features constant would result in a change in the decision outcome. It ensures that individuals are treated fairly regardless of their protected attributes.

Understanding fairness metrics

Interactive quiz

Creating an interactive quiz on Mentimeter

Poll question

Before we start, what are important factors to consider when determining if an AI system is fair?

Options

a) Outcome fairness; b) Demographic representation; c) Equal treatment; d) Transparency.

Understanding fairness metrics

Interactive class exercises

Topic: Statistical Parity

Calculate Statistical Parity

- Given the data, calculate the rate of loan approval for each race. Is there parity?
- Provide participants with the numbers (or percentages) of approvals and rejections for each race and ask them to calculate if the rates are approximately equal.
- Multiple choice answers: "Rates are equal; parity is achieved.“ and "Rates are not equal; there is a disparity."

Understanding fairness metrics

Interactive class exercises

Topic: Equalized odds

Assess equalized odds

- Assuming the true positive rate (loan correctly approved for those who should be approved) for Whites is 80% and for Blacks is 60%, is there equalized odds?
- Multiple choice answers: "Yes, the odds are equal.“ or "No, the odds are not equal."

FAIRNESS IN DATA COLLECTION AND PREPROCESSING

IMAGE SOURCE | Generated by DALL-E

Fairness in data collection and preprocessing

Diverse data sources

Click cards to flip

Inclusive sampling

Click cards to flip

Data cleaning

Click cards to flip

Bias detection

Click cards to flip

Fair feature engineering

Click cards to flip

Transparency and documentation

Click cards to flip

Regular monitoring

Click cards to flip

Fairness in data collection and preprocessing

Diverse data sources

Collect data from diverse sources to ensure a comprehensive representation of the population. This helps prevent under-representation or bias towards specific demographic groups.

Inclusive sampling

Click cards to flip

Data cleaning

Click cards to flip

Bias detection

Click cards to flip

Fair feature engineering

Click cards to flip

Transparency and documentation

Click cards to flip

Regular monitoring

Click cards to flip

Fairness in data collection and preprocessing

Diverse data sources

Collect data from diverse sources to ensure a comprehensive representation of the population. This helps prevent under-representation or bias towards specific demographic groups.

Inclusive sampling

Use inclusive sampling techniques to ensure that all segments of the population are adequately represented in the dataset. This may involve stratified sampling or oversampling minority groups to address imbalances.

Data cleaning

Click cards to flip

Bias detection

Click cards to flip

Fair feature engineering

Click cards to flip

Transparency and documentation

Click cards to flip

Regular monitoring

Click cards to flip

Fairness in data collection and preprocessing

Diverse data sources

Collect data from diverse sources to ensure a comprehensive representation of the population. This helps prevent under-representation or bias towards specific demographic groups.

Inclusive sampling

Use inclusive sampling techniques to ensure that all segments of the population are adequately represented in the dataset. This may involve stratified sampling or oversampling minority groups to address imbalances.

Data cleaning

Carefully clean and preprocess the data to remove any biases or inconsistencies. This includes identifying and correcting errors, handling missing values, and standardizing data formats.

Bias detection

Click cards to flip

Fair feature engineering

Click cards to flip

Transparency and documentation

Click cards to flip

Regular monitoring

Click cards to flip

Fairness in data collection and preprocessing

Diverse data sources

Collect data from diverse sources to ensure a comprehensive representation of the population. This helps prevent under-representation or bias towards specific demographic groups.

Inclusive sampling

Use inclusive sampling techniques to ensure that all segments of the population are adequately represented in the dataset. This may involve stratified sampling or oversampling minority groups to address imbalances.

Data cleaning

Carefully clean and preprocess the data to remove any biases or inconsistencies. This includes identifying and correcting errors, handling missing values, and standardizing data formats.

Bias detection

Implement techniques to detect and mitigate biases in the data. This may involve conducting bias audits or using automated tools to identify disparities across different demographic groups.

Fair feature engineering

Click cards to flip

Transparency and documentation

Click cards to flip

Regular monitoring

Click cards to flip

Fairness in data collection and preprocessing

Diverse data sources

Collect data from diverse sources to ensure a comprehensive representation of the population. This helps prevent under-representation or bias towards specific demographic groups.

Inclusive sampling

Use inclusive sampling techniques to ensure that all segments of the population are adequately represented in the dataset. This may involve stratified sampling or oversampling minority groups to address imbalances.

Data cleaning

Carefully clean and preprocess the data to remove any biases or inconsistencies. This includes identifying and correcting errors, handling missing values, and standardizing data formats.

Bias detection

Implement techniques to detect and mitigate biases in the data. This may involve conducting bias audits or using automated tools to identify disparities across different demographic groups.

Fair feature engineering

Consider the implications of feature selection and engineering on fairness. Avoid using features that may perpetuate stereotypes or discrimination and strive for features that are relevant and non-discriminatory.

Transparency and documentation

Click cards to flip

Regular monitoring

Click cards to flip

Fairness in data collection and preprocessing

Diverse data sources

Collect data from diverse sources to ensure a comprehensive representation of the population. This helps prevent under-representation or bias towards specific demographic groups.

Inclusive sampling

Use inclusive sampling techniques to ensure that all segments of the population are adequately represented in the dataset. This may involve stratified sampling or oversampling minority groups to address imbalances.

Data cleaning

Carefully clean and preprocess the data to remove any biases or inconsistencies. This includes identifying and correcting errors, handling missing values, and standardizing data formats.

Bias detection

Implement techniques to detect and mitigate biases in the data. This may involve conducting bias audits or using automated tools to identify disparities across different demographic groups.

Fair feature engineering

Consider the implications of feature selection and engineering on fairness. Avoid using features that may perpetuate stereotypes or discrimination and strive for features that are relevant and non-discriminatory.

Transparency and documentation

Document the data collection and preprocessing procedures to provide transparency and accountability. This includes documenting any decisions made during data cleaning or feature engineering that may impact fairness.

Regular monitoring

Click cards to flip

Fairness in data collection and preprocessing

Diverse data sources

Collect data from diverse sources to ensure a comprehensive representation of the population. This helps prevent under-representation or bias towards specific demographic groups.

Inclusive sampling

Use inclusive sampling techniques to ensure that all segments of the population are adequately represented in the dataset. This may involve stratified sampling or oversampling minority groups to address imbalances.

Data cleaning

Carefully clean and preprocess the data to remove any biases or inconsistencies. This includes identifying and correcting errors, handling missing values, and standardizing data formats.

Bias detection

Implement techniques to detect and mitigate biases in the data. This may involve conducting bias audits or using automated tools to identify disparities across different demographic groups.

Fair feature engineering

Consider the implications of feature selection and engineering on fairness. Avoid using features that may perpetuate stereotypes or discrimination and strive for features that are relevant and non-discriminatory.

Transparency and documentation

Document the data collection and preprocessing procedures to provide transparency and accountability. This includes documenting any decisions made during data cleaning or feature engineering that may impact fairness.

Regular monitoring

Continuously monitor the data collection and preprocessing pipeline to identify and address any emerging biases or disparities. Regularly update the dataset and preprocessing techniques to ensure fairness over time.

FUTURE TRENDS AND EMERGING ISSUES

IMAGE SOURCE | Generated by DALL-E

Future trends and emerging issues

Explainability and interpretability

Click cards to flip

Adversarial attacks and defenses

Click cards to flip

Fairness across multiple dimensions

Click cards to flip

Algorithmic bias mitigation

Click cards to flip

Ethical consideration and governance

Click cards to flip

Human-centric design

Click cards to flip

Fairness in emerging technologies

Click cards to flip

Future trends and emerging issues

Explainability and interpretability

As AI systems become more complex, there is a growing demand for transparency and interpretability. Future trends may focus on developing techniques and standards for explaining AI decisions and making them more interpretable to stakeholders.

Adversarial attacks and defenses

Click cards to flip

Fairness across multiple dimensions

Click cards to flip

Algorithmic bias mitigation

Click cards to flip

Ethical consideration and governance

Click cards to flip

Human-centric design

Click cards to flip

Fairness in emerging technologies

Click cards to flip

Future trends and emerging issues

Explainability and interpretability

Adversarial attacks and defenses

Adversarial attacks, where malicious actors manipulate AI systems by subtly altering input data, pose a significant threat to fairness and robustness. Future research may explore robust defenses and countermeasures against adversarial attacks to enhance AI system reliability and fairness.

Fairness across multiple dimensions

Click cards to flip

Algorithmic bias mitigation

Click cards to flip

Ethical consideration and governance

Click cards to flip

Human-centric design

Click cards to flip

Fairness in emerging technologies

Click cards to flip

Future trends and emerging issues

Explainability and interpretability

Adversarial attacks and defenses

Fairness across multiple dimensions

Fairness in AI goes beyond demographic attributes and may encompass other dimensions such as socioeconomic status, disability, or language proficiency. Future trends may involve developing multi-dimensional fairness metrics and algorithms to address intersectional biases.

Algorithmic bias mitigation

Click cards to flip

Ethical consideration and governance

Click cards to flip

Human-centric design

Click cards to flip

Fairness in emerging technologies

Click cards to flip

Future trends and emerging issues

Explainability and interpretability

Adversarial attacks and defenses

Fairness across multiple dimensions

Algorithmic bias mitigation

Advances in algorithmic bias mitigation techniques will continue to be a focal point, with research focusing on developing fairer algorithms and models that reduce biases across various decision-making contexts.

Ethical consideration and governance

Click cards to flip

Human-centric design

Click cards to flip

Fairness in emerging technologies

Click cards to flip

Future trends and emerging issues

Explainability and interpretability

Adversarial attacks and defenses

Fairness across multiple dimensions

Algorithmic bias mitigation

Ethical consideration and governance

The ethical implications of AI fairness and bias will remain a key area of concern, with increasing emphasis on ethical AI development practices and governance frameworks. Future trends may involve establishing regulatory guidelines and standards for ensuring fairness and accountability in AI systems.

Human-centric design

Click cards to flip

Fairness in emerging technologies

Click cards to flip

Future trends and emerging issues

Explainability and interpretability

Adversarial attacks and defenses

Fairness across multiple dimensions

Algorithmic bias mitigation

Ethical consideration and governance

Human-centric design

Future AI systems may prioritize human-centric design principles, incorporating user feedback and preferences to enhance fairness and usability. This could involve co-designing AI systems with diverse stakeholders to ensure inclusivity and fairness.

Fairness in emerging technologies

Click cards to flip

Future trends and emerging issues

Explainability and interpretability

Adversarial attacks and defenses

Fairness across multiple dimensions

Algorithmic bias mitigation

Ethical consideration and governance

Human-centric design

Fairness in emerging technologies

As AI continues to intersect with emerging technologies such as autonomous vehicles, healthcare AI, and facial recognition systems, addressing fairness and bias becomes increasingly important. Future trends may involve adapting fairness principles to new application and technological advancements.

CONCLUSION

IMAGE SOURCE | Generated by DALL-E

CONCLUSION

1. **Understanding complexity**: AI fairness and bias are multifaceted challenges that require comprehensive strategies and ongoing attention.
2. **Strategies for action**: from diverse representation in data collection to algorithmic bias mitigation, a range of strategies is available to address and mitigate biases in AI systems.
3. **Anticipating emerging challenges**: as AI continues to evolve, staying ahead of emerging trends and issues such as adversarial attacks, multi-dimensional fairness, and ethical considerations will be crucial.
4. **Toward a more equitable future**: by prioritizing human-centric design, transparency, and accountability, we can work towards building AI systems that promote fairness, inclusivity, and trust.

THANK YOU

Project number: 2022-1-ES01-KA220-HED-000085257

Co-funded by the European Union

The European Commission's support for the production of this publication does not constitute of the contents, which reflect the views only of the authors , and the Commission cannot be held responsible for any use which may be made of the information contained therein.

CC BY 4.0