Unit 6: Ethical Issues in Data Science

Bias and Fairness

Biasness refers to the presence of systematic and consistent deviations or prejudices in data, algorithms, or decision-making processes. It can occur at various stages of data science, including data collection, data representation, algorithm design, and decision-making.

Fairness refers to the absence of unjust or discriminatory treatment and the equitable distribution of benefits, opportunities, and outcomes across different groups or individuals. It aims to ensure that decisions and algorithms do not favor or disadvantage any particular group based on protected attributes such as race, gender, or age.

Issues with fairness and bias in data science

Fairness and bias in data science can give rise to several issues that have real-world consequences. Here are some key issues associated with fairness and bias in data science:

Discriminatory Outcomes: Biased algorithms can lead to discriminatory outcomes, perpetuating and amplifying existing inequalities. For example, biased hiring algorithms may disproportionately reject candidates from certain demographic groups, exacerbating disparities in employment opportunities.

Unfair Treatment: When algorithms exhibit bias, certain groups may be unfairly advantaged or disadvantaged. This can lead to unequal treatment, such as biased loan approvals, targeted advertising, or biased criminal justice decisions, further entrenching social injustices.

Reinforcing Stereotypes: Biased data or algorithms can reinforce and perpetuate stereotypes by making decisions based on historical patterns that may be discriminatory. For instance, an algorithm trained on biased data may associate certain

demographics with specific characteristics or behaviors, leading to unfair generalizations.

Lack of Diversity and Representation: Biased algorithms and data can perpetuate underrepresentation and marginalization of certain groups. If historical biases are present in the data, algorithms trained on such data may continue to reproduce and amplify those biases, resulting in a lack of diversity and representation in decision-making processes.

Transparency and Accountability: Biased algorithms are often considered black boxes, making it challenging to understand how decisions are reached. Lack of transparency hampers accountability and prevents individuals or groups from challenging unfair or biased outcomes.

Data Collection and Sampling Biases: Biases can be introduced during data collection, such as underrepresenting certain groups or perspectives. Biased sampling methods can lead to incomplete or skewed data, compromising the accuracy and fairness of subsequent analyses and decisions.

Ethical Concerns: Fairness and bias issues raise ethical concerns related to privacy, autonomy, and human dignity. Biased algorithms can infringe on individuals' rights and perpetuate systemic discrimination, making it imperative to address these issues to uphold ethical standards.

Trust and Public Perception: Biased algorithms erode trust in data-driven decision-making systems. Public perception plays a vital role in the adoption and acceptance of data science solutions, and concerns about fairness and bias can hinder the trustworthiness of such systems.

Ethics in Data Science

Ethics in data science refers to the principles and guidelines that govern the responsible and ethical use of data, algorithms, and technologies. It involves considering the potential impacts, risks, and societal implications of data science practices and ensuring that they align with moral values, legal requirements, and social norms. Here are key aspects of ethics in data science:

Data Ethics concerns during:

Data Collection:

Informed Consent: Obtaining informed consent from individuals, ensuring they understand the purpose, scope, and potential uses of data collection.

Privacy Protection: Implementing measures to protect the privacy of individuals, including secure data storage and handling of personally identifiable information.

Data Minimization: Collecting only the necessary and relevant data, avoiding the collection of excessive or unnecessary personal information.

Data Storage and Security:

Data Security: Implementing robust security measures to protect data from unauthorized access, breaches, or misuse.

Data Retention: Establishing appropriate data retention periods and policies to avoid storing data longer than necessary, minimizing the risk of data misuse or unintended disclosures.

Data Preprocessing and Analysis:

Bias Identification and Mitigation: Identifying and mitigating biases present in the data, including data collection biases, algorithmic biases, or biases introduced through preprocessing steps.

Anonymization and De-identification: Protecting the privacy of individuals by anonymizing or de-identifying data to prevent re-identification and safeguard sensitive information.

Algorithm Development and Deployment:

Algorithmic Fairness: Ensuring that algorithms do not produce unfair or discriminatory outcomes, considering factors such as protected attributes (e.g., race, gender) and addressing bias in training data or algorithm design.

Transparency and Explainability: Striving for transparency in algorithmic decision-making, enabling stakeholders to understand how decisions are made and providing explanations for the results, enhancing trust and accountability.

Use and Application of Data:

Consent and Data Sharing: Respecting the consent given by individuals and ensuring that data is used within the agreed-upon purposes, avoiding unauthorized sharing or use of data.

Social and Ethical Impacts: Considering the potential social, ethical, and unintended consequences of data use, such as the impact on marginalized communities, potential discrimination, or negative externalities, and taking steps to mitigate harm.

Data Governance and Accountability:

Data Governance: Establishing policies, procedures, and frameworks to govern the responsible use of data, including data access, data sharing, and data usage agreements.

Accountability and Oversight: Holding organizations and individuals accountable for the ethical use of data, ensuring compliance with ethical guidelines, and providing mechanisms for reporting and addressing ethical concerns.

Common Biases

---

1. **In Group Favoritism and Outgroup Negativity**
   In Group Favoritism
   - ➢ Also called ingroup love
   - ➢ Tendency to give preferential treatment to the same group they belong to.

- ➢ Very likely to occur during data collection.
- ➢ Also likely to occur during data filtering or removing irrelevant data.
- ➢ Highly impacts when data diversity is needed.

Outgroup Negativity

- ➢ Also known as outgroup hate.
- ➢ Tendency to unlike the behavior, activities or people themselves who do not belong
- ➢ to the group they do.
- ➢ Very likely to occur during data collection
- ➢ Likely to have covered the most of negative aspects only of outgroup community

2. **Fundamental Attribution Error**
   - ➢ Tendency that the situational activity or behavior are attributed as intrinsic
   - ➢ quality of someone's character.
   - ➢ These are the judged or observed pattern and is very likely to occur during
   - ➢ data collection.
   - ➢ This feeds negative data to the machine learning model resulting in biased conclusions.

3. **Negativity Bias**
   - ➢ Tendency of emphasizing negative experiences over positive ones.
   - ➢ This is very likely to occur during decision making.
   - ➢ The negative thought about society may expect the negative conclusion from the data science projects.

4. **Stereotyping**
   - ➢ This is the tendency of expect a certain characteristics or behaviors without having
   - ➢ actual information.
   - ➢ This is the expectation set prior to the exploration.
   - ➢ This is likely to occur during data wrangling and exploratory data analysis.

5. **Bandwagon Effect**
   - ➢ Tendency to follow others because

- ➢ Some other top ranked researchers or people did.
- ➢ All people are doing I.e. following the mass.
- ➢ Likely to occur during data collection like the same sort of data is collected based on pre collected data or research.
- ➢ Some might expect the same result as others have inferred.

6. Bias Blind Spot

- ➢ Our tendency not to see our own personal biases.
- ➢ Likely to ignore or remain unnoticed where there are personal blind spot biases.
- ➢ Likely to occur from data collection to result analysis of data science process

Addressing Bias

- ➢ Addressing bias in data science is an extremely complex topic and most importantly there are no universal solutions or silver bullets.

- ➢ Before any data scientist can work on the mitigation of biases we need to define fairness in the context of our business problem by consulting the following:

- ➢ As an example, imagine you want to design some ML system to process mortgage loan applications and only a small fraction of applications are by women.

1. Group unaware selection

- ➢ It's a preventive measure
- ➢ This is the process of preventing bias by eliminating the factor that is likely to cause.
- ➢ For example, avoid the collection of gender to avoid bias by gender.

2. Adjusted group threshold

- ➢ Adjust any biased and unbalanced data
- ➢ Because historic biases make women appear less loan-worthy than men,
- ➢ e.g. work history and childcare responsibilities, we use different approval thresholds by group.

3. Demographic Parity

- ➢ The output of the machine learning model should not depend on the sensitive demographic attribute like gender, race, ethnicity, education

level etc.

4. Equal Opportunity

- ➢ Equal opportunity fairness ensures that the proportion of people who should be selected by the model ("positives") that are correctly selected by The model is the same for each group. We refer to this proportion as the true positive rate (TPR) or sensitivity of the model.

- ➢ A doctor uses a tool to identify patients in need of extra care, who could be at risk for developing serious medical conditions. (This tool is used only to supplement the doctor's practice, as a second opinion.) It is designed to have a high TPR that is equal for each demographic group.

    - Provide equal opportunity to the diverse population.
    - Should be fair enough for the representation in sampling and treatment.
    - E.g. The representation of men and women should be the same for granting loans in banks.

5. Precision Parity

- ➢ Tune the output of the model to treat the group equally.
- ➢ Male and Female should get equal salary based on the position. If machine learning model suggests lesser salary to women compared to men in same post, then such model should be tuned so that both have similar earnings.

- ➢ When building a ML model, keep de-biasing in mind.