DataLab Preparation (Week 1, DataLab II, Wednesday)

**2a What are the key takeaways of the article Machines Bias?, and the tutorial The Myth of the Impartial Machine. Write them down with a maximum of 150 words.**

Angwin et al.’s ‘Machine Bias’ shows the underlying prejudice in risk assessment algorithms when it comes to black convicts. It has been shown that blacks are more likely to be categorized by systems such as Northpointe as high-risk criminals, meaning that there is a higher chance of them convicting another crime in the next two years than whites, no matter the criminal background.

Conversely, Feng and Wu's 'The Myth of the Impartial Machine' highlights the reasons why bias slips into ML algorithms and their dangers are explained. Bias stems from non-representative data or accidental biases within representative data (for example, men are more likely to be software engineers than women, which would be reflected in a word association algorithm), and are amplified through algorithms, either incentivised to predict the majority group to maximize predictive accuracy when faced with an imbalanced dataset, or through the falling into runaway feedback loops.

**3a Identify the population and sample in the following example:**

**Suppose you want to build a binary classification model that predicts if a person, -i.e. the agent, is performing a specific action or not, say coaching. To obtain data for your classifier, you scrape Google and collect 1000 images of people coaching, and 1000 images of people performing other acts besides coaching.**

**a) The population encompasses all images of people coaching and performing other acts than coaching on the internet; the sample encompasses all images of people coaching and performing other acts than coaching on Google.**

**b) The population encompasses all images of on Google; the sample encompasses all images of people coaching and performing other acts than coaching on Google.**

**c) The population encompasses all images of people coaching and performing other acts than coaching on Google; the sample encompasses the 1000 images of people coaching, and 1000 images of people performing other acts besides coaching that you scraped from Google.**

**Select the correct answer, and elaborate on your choice.**

The correct answer is c. The population refers to the entire image gallery of people coaching and performing other acts besides coaching. However, since we cannot use the entire population for training our algorithm as it would be impractical, we select only 1000 images of people coaching (class 0) and people doing other activities other than coaching (class 1).

**3b Propose an appropriate individual fairness method, - i.e., ‘Fairness Through Awareness' or ‘Fairness Through Unawareness', and explain how it addresses COMPAS' limitations in terms of bias and fairness.**

An appropriate individual fairness method in this instance is ‘Fairness through Awareness’. In addressing bias and fairness limitations in COMPAS, one needs to take into consideration attributes such as race or gender during model training and implement interventions to mitigate bias and promote fairness. By recognizing that certain demographic groups may face adversities due to biases, Fairness Through Awareness aims to rectify disparities directly. One way to apply this concept would be to increase the data in the dataset to be more representative of minorities or marginalized groups. Another way could be to increase the sample by adding slightly modified copies of already existing data or newly created synthetic data from existing data.

**4a Combine the following terms (e.g. A = 1, I):**

**A. Gender B. (Biological) Sex**

1. **‘Female' and ‘Male'**
2. **‘Feminine' and ‘Masculine'**

**I. Social construct II. Biological fact**

**Write your answer down (If you disagree with the setup of the exercise that is totally fine, but please explain why).**

A = 2, I; B = 1, II.

**4b Give at least one additional example of a feature that could be seen as a sensitive/protected attribute?**

Another example of a feature that could be seen as a sensitive attribute is socioeconomic status, encompassing various factors such as income, education level, and occupation, and it often reflects individuals' access to resources, opportunities, and social privileges.

**4c Define the term ‘group fairness', and explain how it differs from ‘individual fairness'. Write your answer down.**

Group fairness refers to the notion of fairness in machine learning and algorithmic decision-making across different demographic groups or protected attributes.

**4d Explain the differences between ‘independence', ‘sufficiency', and ‘separation' concerning fair AI.**

*Independence* dictates that a model's predictions should not be influenced by factors like race or gender, ensuring decisions will not be influenced by irrelevant attributes.

*Sufficiency* requires the model to consider all relevant factors necessary for accurate predictions while disregarding extra information, such as using credit history and income but not race or gender in credit scoring.

*Separation* emphasizes distinguishing causally related from unrelated factors in predictions. This ensures accurate separation of the effects of protected attributes from other predictive factors, preventing unfair advantages or penalties based on demographic membership.

**4e Watch the video ‘Fairness Criteria’, and explain how a decision that is considered fair for a group can be considered unfair for an individual, and vice versa, by providing a concrete example.**

While group fairness aims to prevent discrimination against the entire group, individual fairness ensures equitable treatment based on unique circumstances. Therefore, a decision that is considered fair for the group may actually disadvantage an individual that is part of the group itself. An example could be a university admissions policy aimed at increasing diversity. While promoting group fairness by addressing racial injustices and fostering inclusivity, individual fairness may be sacrificed for the overall minority group fairness. For instance, a highly qualified applicant from a majority group may feel disadvantaged if looked over for a less qualified candidate from an underrepresented minority solely due to demographic factors. We would then actually have an instance where the racial discrepancies that were aimed to be destroyed have been replaced by other, rarely met injustice.