Fairness and reliability in ML (COL865) Minor I September 2, 2019

Name: **Entry:** **Note:** Please answer only in the space provided.

---

1. Please comment (elaborate, critique, or both) on the following assertions:

   (a) Training, validating and testing ML algorithms on curated datasets is problematic.

   (b) Rather than discovering new knowledge, ML can amplify stereotypes, promote status quo and prevent change detection.

   (c) Generalization on majority datasets and cultures discriminates against minorities.

(d) Any decision/recommendation on humans must be explainable. ML is often not. Correlation is not causation.

(e) Fairness requirement in prediction is over-emphasized compared to other interventions. For example, even if selection does not discriminate, the post-selection environment may be discriminatory leading to wrong prediction outcomes. Consider, for example, that Y is an outcome variable (such as expected performance in IIT) which depends in some unknown way on variables X_0 and X where X_0 is a policy intervention to reduce discrimination and boost learning outcomes in IIT, and our objective is to maximize a known payoff function $\pi(X_0, Y)$. Then, the decision X_0 depends on the derivative

$$\frac{d\pi(X_0, Y)}{dX_0} = \frac{\partial \pi}{\partial X_0}(Y) + \frac{\partial \pi}{\partial Y}\frac{\partial Y}{\partial X_0}$$

(f) Selection/targeting/prediction based on assigning scores to human beings are inherently problematic.

(g) Algorithmic decisions make criteria of discrimination explicit and eliminates vagueness. Operationalization is a virtue.

(h) Defining accuracy as $P(Y = \hat{Y})$ (probability of correctly predicting the target variable, or as a score $E(Y|X)$ (the conditional expectation of the target variable given the observations) are both problematic.

2. Suppose Z, Y and R are random variable corresponding to a protected attribute, a target variable and a classifier score respectively. Consider the following three criteria (the symbol $\perp$ indicates independence of random variables):

   (a) **Independence:** $R \perp Z$, or $P(R = 1|Z = a) = P(R = 1|Z = b)$.

   (b) **Separation:** $R \perp Z|Y$, or $P(R = 1|Y = 1, Z = a) = P(R = 1|Y = 1, Z = b)$ and $P(R = 1|Y = 0, Z = a) = P(R = 1|Y = 0, Z = b)$
   Note that $P(R = 1|Y = 1)$ and $P(R = 1|Y = 0)$ are the true and false positive rates, respectively.

   (c) **Sufficiency and calibration:** $Y \perp Z|R$, or $P(Y = 1|R = r, Z = a) = P(Y = 1|R = r, Z = b) = P(Y = 1|R = r) = r$

   Evaluate the three criteria for *disparate impact*, *disparate treatment*, *disparate mistreatment* and any other consideration for fair decision making. How these criteria may be achieved in a general ML setting, if at all?