Stephen Scarano
19 August 2021

Notes on *The Alignment Problem*
By Brian Christian

Page 69 – Fairness And Risk Assessment

As mentioned in Chouldechova's Fair Prediction and Disparate Impact:

> In this paper we show that the differences in false positive and false negative rates cited as evidence of racial bias by Angwin et al. are a direct consequence of applying an RPI that that satisfies predictive parity to a population in which recidivism prevalence differs across groups [1]

We recall that a *false positive* refers to a model decision which incorrectly states a condition is present, and that a *false negative* refers to the opposite: a model decision which incorrectly states a condition is not present [4]. In this case, ProPublica alleges that the COMPAS recidivism model demonstrates a disproportionate number of false positives (that is, defendants who reoffended) and false negatives (defendants who do not reoffend) across racial lines; more specifically, COMPAS allegedly overestimates the number of black defendants who reoffend while underestimating the number of reoffending white defendants [2].

To understand Chouldechova's outcome listed above, we must define the relevant terms [1]:

Calibration: A score $S = S(x)$ is *well-calibrated* if it reflects the same likelihood of recidivism irrespective of the individuals' group membership. In mathematical terms, for all s,

$$P(Y = 1 \,|\, S = s, R = b) = P(Y = 1 \,|\, S = s, R = w)$$

Caption: where Y is the recidivism outcome $\{0, 1\}$,
R is the group a subject belongs to

Predictive Parity: A score $S = S(x)$ *satisfies predictive parity* at some threshold s_{HR} if the likelihood of recidivism among high-risk offenders is the same regardless of group membership. That is,

$$P(Y = 1 \,|\, S > s_{HR}, R = b) = P(Y = 1 \,|\, S > s_{HR}, R = w)$$

Citations:

[1] Chouldechova's Fair Prediction and Disparate Impact Paper

[2] Original ProPublica COMPAS Analysis and Breakdown

[3] Wikipedia Statistics References (positive predictive value)

[4] Wikipedia Statistics References (false positives vs. negatives)