Safety from Bigoted AI: Analyzing Definitions and Frameworks for Learned Models

Hirsh Guha
Jeffrey Cheng

December 11, 2020

1 Topic

Our goal is to analyze and compile literature on the nature of model frameworks for fairness, accountability, and transparency. This includes learning the difference between minimizing versus eliminating bias, understanding the intrinsic differences between objective and subjective safety, and whether or not a safety critical model can be designed for papers that do not meet our thresh-hold on training quality. Garbage-in-garbage-out is the well-known phenomenon in which poor training data leads to poor model behavior. This is exacerbated when the training data is incorrectly labeled on account of implicit human bias; models will often over-fit on features correlating with human bias and thus fail to correctly learn generalized patterns relevant to the task, instead becoming simple classifiers over protected classes such as race or gender.

2 Relevance

The sub-field of Fairness, Accountability, and Transparency in Machine Learning (FAT) offers numerous approaches to various topics regarding ensuring that models, data sets, and analyses are non-discriminatory in multiple regards. This could include such categories as religion, race, gender, sexual orientation, etc. Ensuring proper metrics that reduce or entirely remove safety concerns from models that include these data types is obviously necessary both for accuracy and ensuring representation. From both a research perspective as well as a social policy one, it is important that we understand the standards used by papers claiming a framework that is non-biased/non-discriminatory.

3 Technical Questions

1. What are the different frameworks for rigorously defining bias in learning? For each definition of bias, how do the authors propose to unbias algorithms?

2. Unbiasing algorithms can be framed as an error minimization (which mitigates bias) or as a class of model states (which acts as a prevention of bias, or what we call a **safety** bound).

   (a) An example of an bias minimization framing is the Variational Fair Autoencoder, which seeks to learn a representation of input data that eliminates not only the demographic information but also any signal indicating the demographics. This is accomplished by minimizing the mutual information between the learned representation and demographic labels; thus, only information that has no demographic signal can be learned towards downstream task outputs. Formally, the loss on the VAE is:

   $$\min_{q} L(q; x) + \lambda \cdot I(z; c)$$

   for learned parameters q, data x, learned representation z, and sensitive variable c

   (b) An example of a safety bound framing is Average Individual Fairness (Kearns et al), which defines a mapping ϕ as (α, β)-fair iff there exists some γ s.t.:

   $$\mathbb{P}_{x \sim P}(|\mathcal{E}(x; \phi; Q) - \gamma| \geq \alpha) \leq \beta$$

   for distributions P, Q, error function $\mathcal{E}$, bound α, and probability β. The intuition is that for all sensitive classes in distribution P, the error rate over class distribution Q will always be bounded with a probabilistic guarantee.

3. Are individual/class error rate minimization and failure state avoidance qualitatively different in the algorithms they produce?

4. What is the nature of the trade-off between the tightness of a safety bound and the performance of an algorithm?

5. Do FAT frameworks for bias / discrimination meet the requirements of a safety-critical model?