WHAT IS CLAIMED IS:

1. A method of correcting algorithmic bias in a machine learning framework, the method comprising:

obtaining, a first dataset, the first dataset containing protected attributes from at least one category of protected attributes;

obtaining, a second dataset, the second dataset not containing any protected attributes from the at least one category of protected attributes;

training, using the first dataset, a protected attributes estimation model, wherein the protected attributes estimation model estimates information related to the least one protected attribute; and

predicting, using protected attributes estimation model, at least one protected attribute for data within the second dataset.

2. The method of claim 1 wherein the protected attributes estimation model is trained in iterations.

3. The method of claim 2 wherein each iteration involves an active learning process to improve the protected attributes estimation model at each iteration.

4. The method of claim 2 wherein additional datapoints containing protected attributes are added to the first dataset to update the training of the protected attributes.

5. The method of claim 4 wherein the additional datapoints are obtained by annotating points from the second dataset with a protected attribute.

6. The method of claim 4 wherein the additional datapoints are chosen from a ranked list of datapoints in the second dataset.

7. The method of claim 6 wherein the ranking is based on empirically defined bias.

8. The method of claim 1 wherein the protected attributes estimation model provides a confidence metric associated with the predicted protected attribute.

9. The method of claim 1 wherein an augmented dataset is generated based on augmenting the second dataset with predicted protected attributes from the protected attributes estimation model.

10. The method of claim 9, further comprising:

training a plurality of bias mitigation models based on subsets of the augmented dataset;

obtaining a fairness metric and an accuracy metric for each of the plurality of the bias mitigation models; and

choosing one model from the plurality of bias mitigation models based on at least one of the fairness metric and the accuracy metric.

11. The method of claim 10 wherein the subsets of the augmented dataset are generated based on statistical information of the protected attributes of the augmented dataset,

wherein the statistical information is generated by the estimation model,

wherein the statistical information is a confidence metric associated with a predicted protected attributed, and

wherein the subsets of the augmented dataset are generated based on ranges of confidence metrics.

12. The method of claim 9, wherein generating the augment dataset comprises:

(A) performing a first stage, the first stage comprising:

obtaining, a protected attributes (PA) known dataset and a PA unknown dataset;

training a bias estimation model based on datapoints from the PA known dataset;

generating, estimates for protected attributes based on the trained bias estimation model;

augmenting the PA unknown dataset to generate an augmented dataset, the method further comprising:

(B) performing a second stage, the second stage comprising:

obtaining the augmented dataset;

generating a plurality of subsets of the augmented dataset;

training a plurality of bias mitigation models based on the plurality of subsets of the augmented dataset; and

choosing one of the trained plurality of bias mitigation models as a final mitigation model.

13. The method of claim 12 wherein additional datapoints are added to the PA known dataset through an active learning process, and

wherein the active learning process is performed by annotating datapoints in the PA unknown dataset with protected attributes.

14. The method of claim 13 further comprising using a ranked entropy model to select additional datapoints during training in the second stage.

15. A system comprising one or more processor coupled to a computer readable medium, the computer readable medium comprising instructions, that when executed by the one or more processors perform a method of any of claims 1–14.