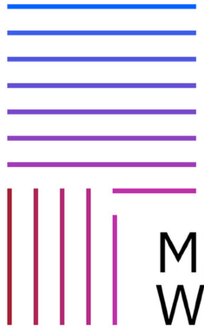
Fair Selective Prediction

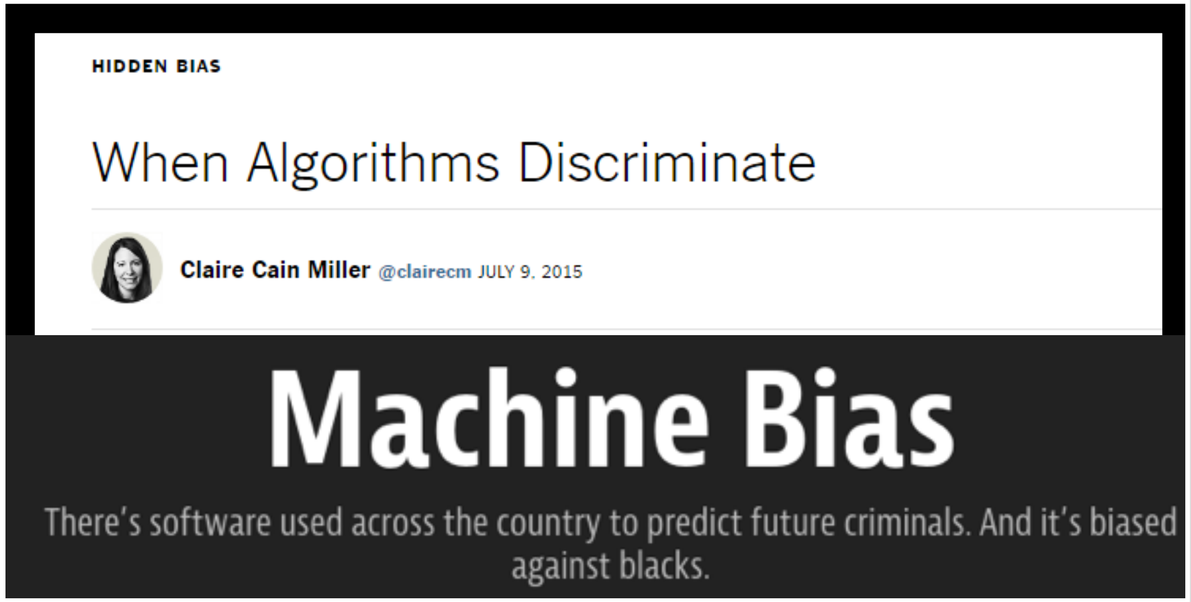
Abhin Shah, Yuheng Bu, Joshua Ka-Wing Lee, Subhro Das,
Rameswar Panda, Prasanna Sattigeri, Gregory W. Wornell



MIT-IBM
Watson
AI Lab

Fairness in Machine Learning

- Machine learning based automated systems are becoming increasingly ubiquitous in our society e.g., employment screening, loan processing, college admissions.
- However, many such systems have been commonly accused of being biased/unfair against certain protected groups.

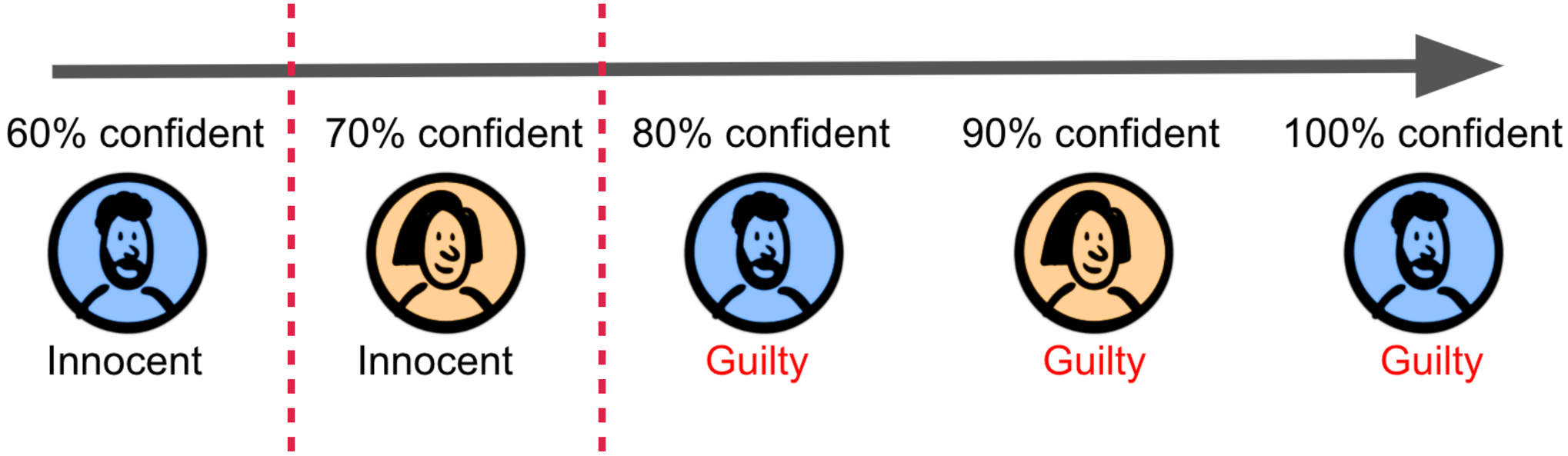


Selective Prediction

- A trustworthy machine learning system should reliably communicate the uncertainty in its predictions.
- When the uncertainty in a prediction is high, the users of the system can reject model predictions and avoid potentially costly errors.
- This is the paradigm of selective prediction or prediction with reject-option.

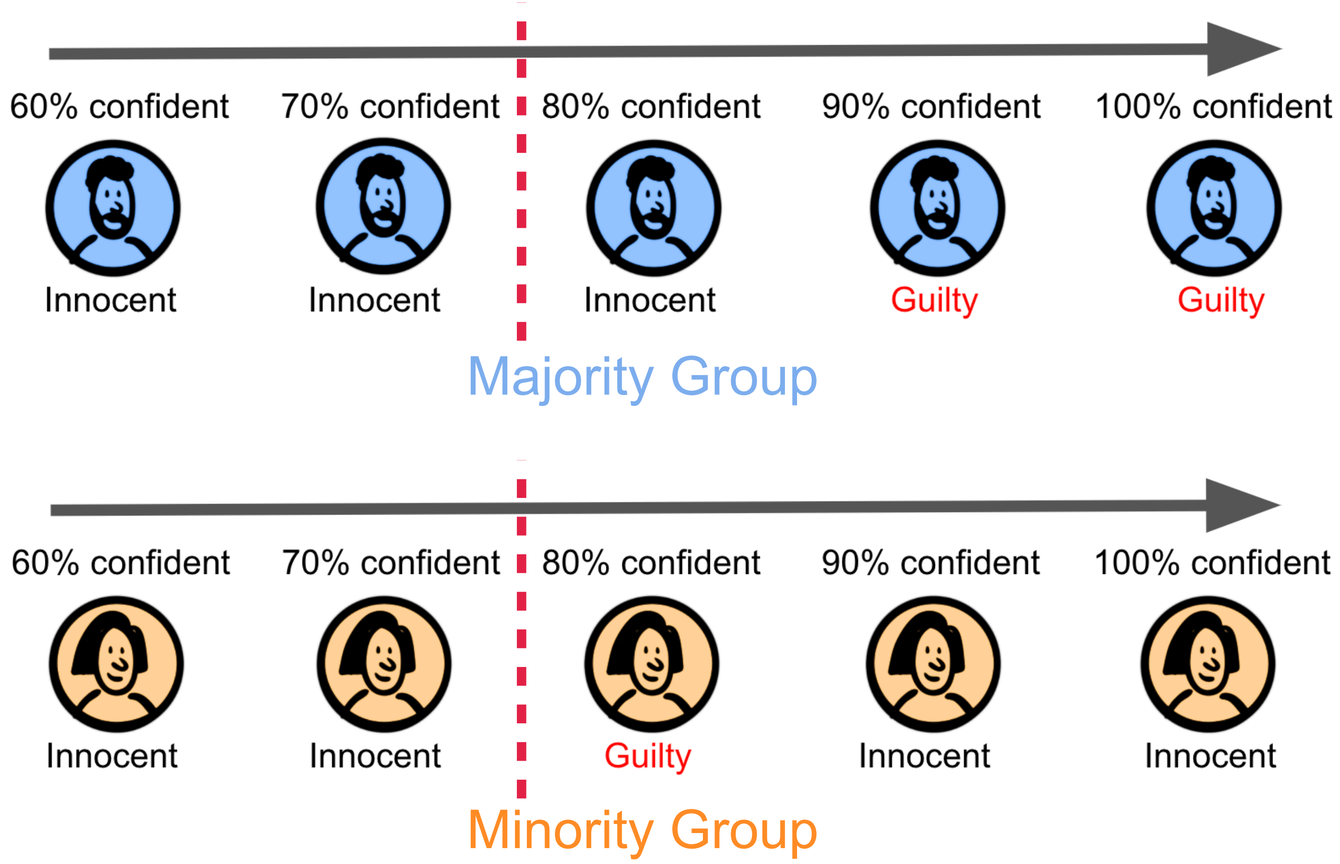
Selective Prediction

- If we have confidence measure for each prediction, we can decide to defer decision making if our confidence is below a certain threshold.
- With a good confidence measure, increasing the threshold results in a better performance.
- The tradeoff is that we have predictions for a fewer samples (i.e., low coverage).



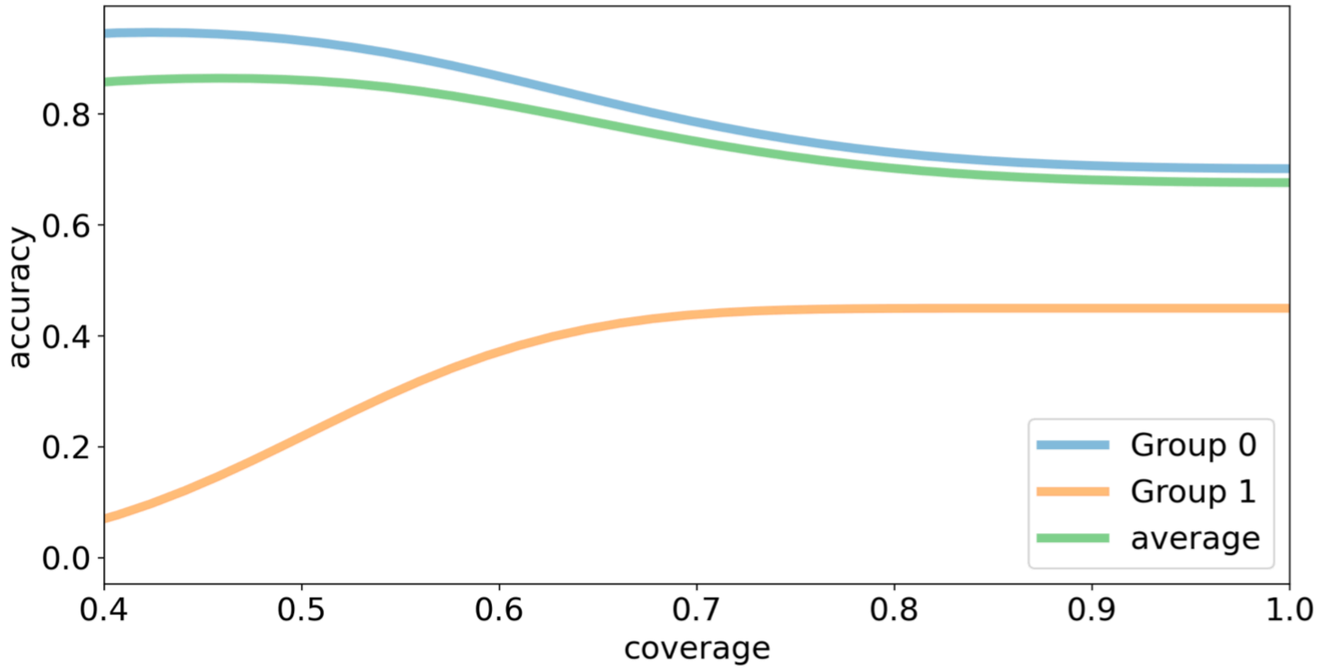
Biases in selective classification

- Classifiers can have good average selective classification performance but perform poorly on certain groups, where increasing confidence threshold may result in a decrease in performance for protected group.



Biases in selective classification

- Classifiers can have good average selective classification performance but perform poorly on certain groups, where increasing confidence threshold may result in a decrease in performance for protected group.



- Similar to selective classification, selective regression can also magnify disparities between different sensitive groups.

Fair Supervised Learning

- Feature set: $X \in \mathcal{X}$ e.g., individual's demographic information
- Sensitive attribute: $D \in \mathcal{D}$ e.g., individual's gender or race
- Target variable:
  - Classification — $Y \in \{0,1\}$ e.g., individual is innocent or guilty
  - Regression — $Y \in \mathbb{R}$ e.g., individual's health insurance cost

Goal

- Find a feature representation $\Phi(X)$ which is predictive of Y, so that we can construct a good predictor $\hat{Y} = T(\Phi(X))$ under some loss criteria, while being 'fair' with respect to D.

Sufficiency criterion

- The sufficiency criterion requires $Y \perp\!\!\!\perp D \mid \Phi(X)$ i.e., the learned features $\Phi(X)$ completely subsumes all information about the sensitive attribute that is relevant to the target variable.

Sufficiency for Fair Selective Prediction

- If our features are sufficient, then our theoretical results ensure that the performance is always improving with the decrease of coverage for all groups, and no group will be penalized in the service of increasing the overall performance in selective prediction.

Key Takeaway

Sufficiency criterion can be used to mitigate disparities between different groups in selective prediction.

Imposing the sufficiency criteria

- We formulate the fair learning objective as

$$\min_\theta \quad L(Y, \hat{Y}; \theta) \quad \text{s.t} \quad Y \perp\!\!\!\perp D \,|\, \Phi(X)$$

  where L is some loss criteria and θ are the model parameters.

- This hard constraint can be relaxed into the following soft constraint:

$$\min_\theta \quad L(Y, \hat{Y}; \theta) + \lambda I(Y; D \,|\, \Phi(x))$$

  where λ is a regularizer.

- Estimating this mutual information is challenging and we use a novel upper bound on this conditional mutual information.

Uncertainty measure

- The conditional variance $\mathrm{Var}(Y \mid X = x)$ can capture the prediction uncertainty.
- In selective classification, the conditional variance can be learned using the softmax output $\mathbb{P}(Y = y \mid \Phi(x))$ (of an existing classifier).
- In selective regression, there is no direct method to extract the conditional variance from an existing regressor designed to predict only the conditional mean.
    1. Heteroskedastic: We train a single a neural network with two heads — one to predict the conditional mean and the other to predict the conditional variance — under the conditional Gaussian assumption.
    2. Residual-based: We train two separate neural networks — one to predict the conditional mean and the other to predict the conditional variance.