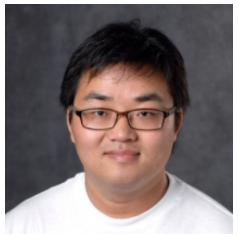
Introduction



Jiliang Tang

Privacy

Xiaorui Liu

Safety & Robustness

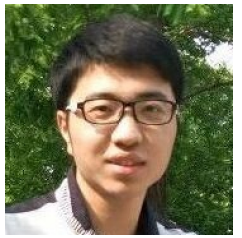
Yaxin Li

Explainability

Wenqi Fan

Non-discrimination & Fairness

Environmental Well- being



Haochen Liu

Accountability & Auditability

Dimension Interactions

Future Directions

Yiqi Wang

Interactions Among Different Dimensions

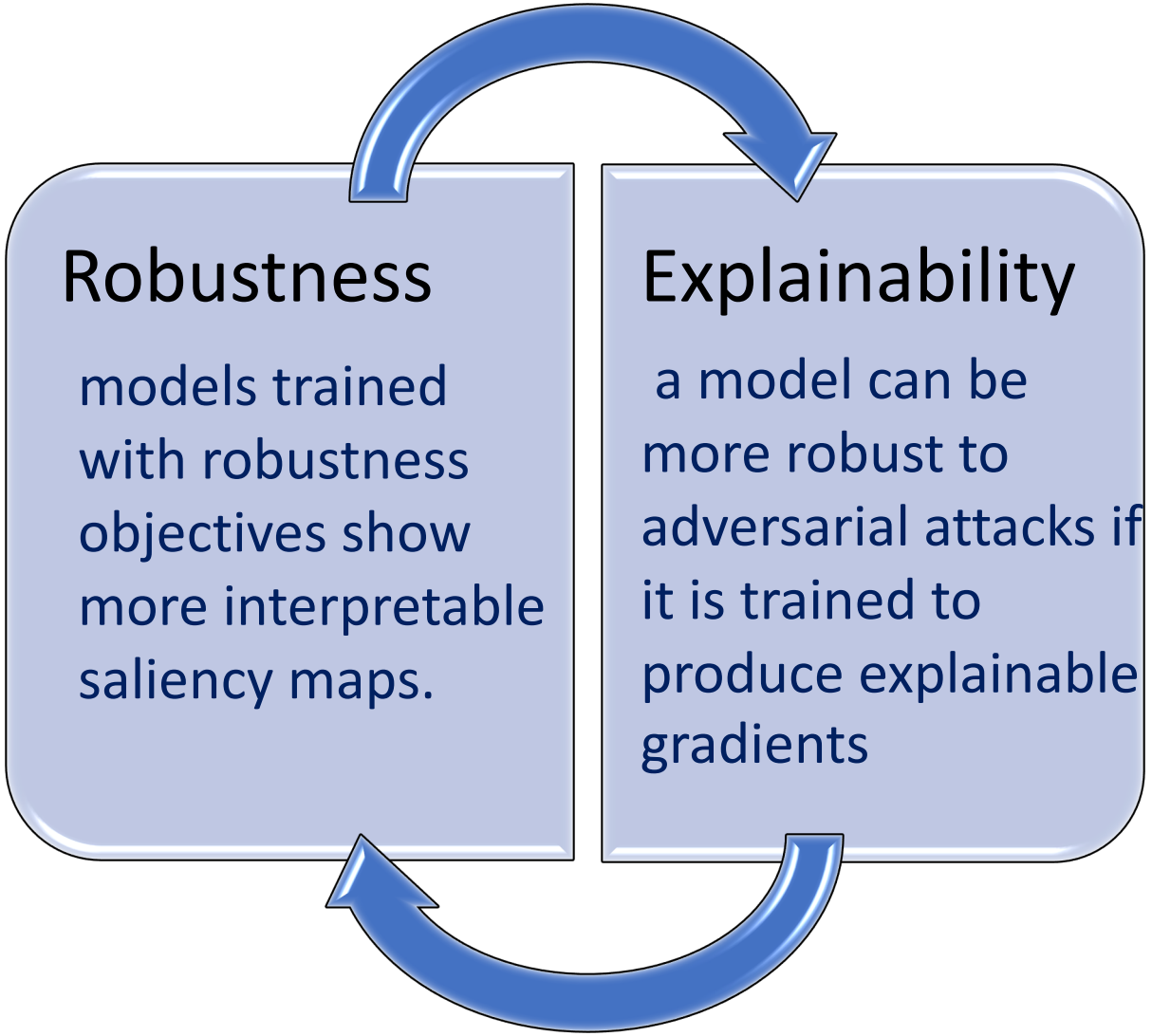
How do these six dimensions influence each other?

Interactions Among Different Dimensions

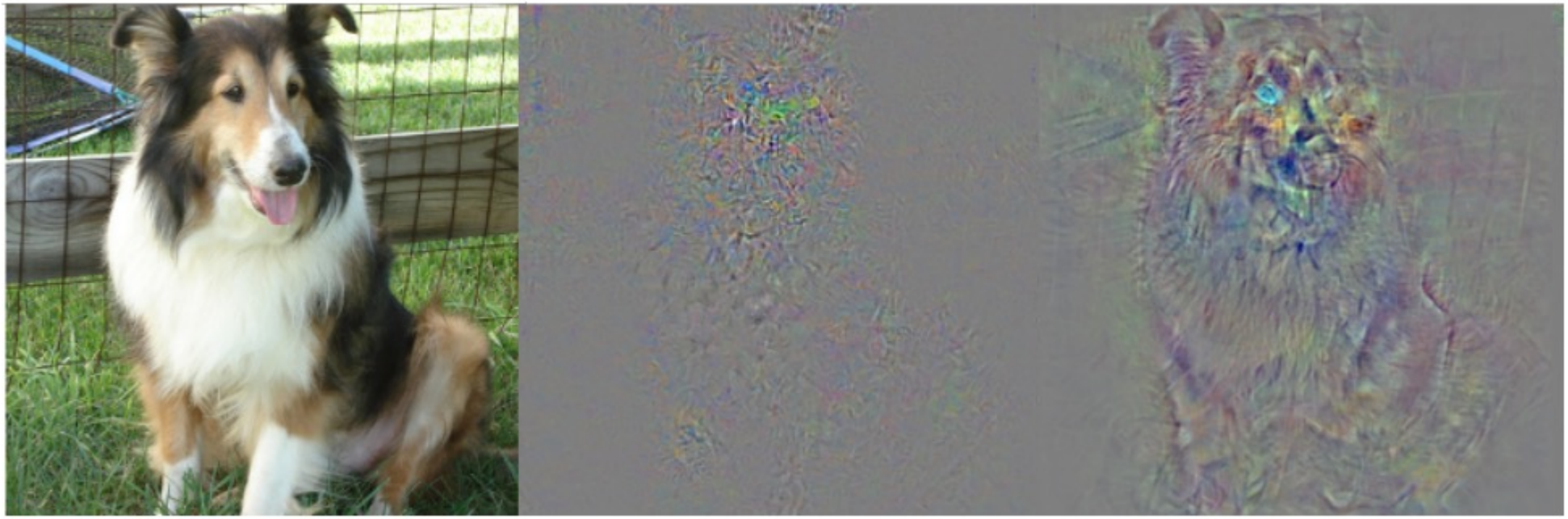
How do these six dimensions influence each other?

There exist both accordance and the conflicts among the six dimensions.

Accordance: Robustness v.s. Explainability

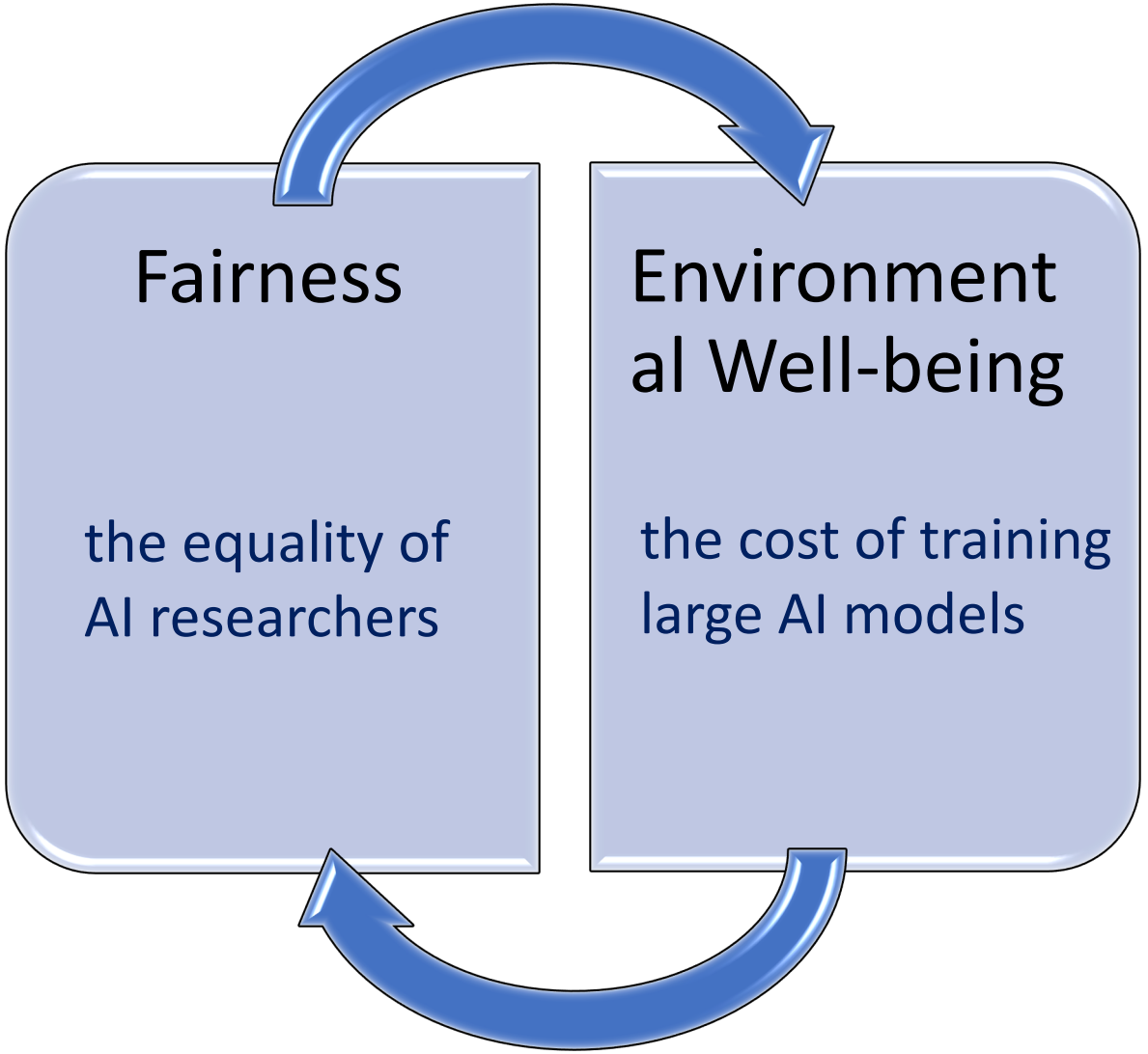


Adversarial Robustness and Saliency Map Interpretability

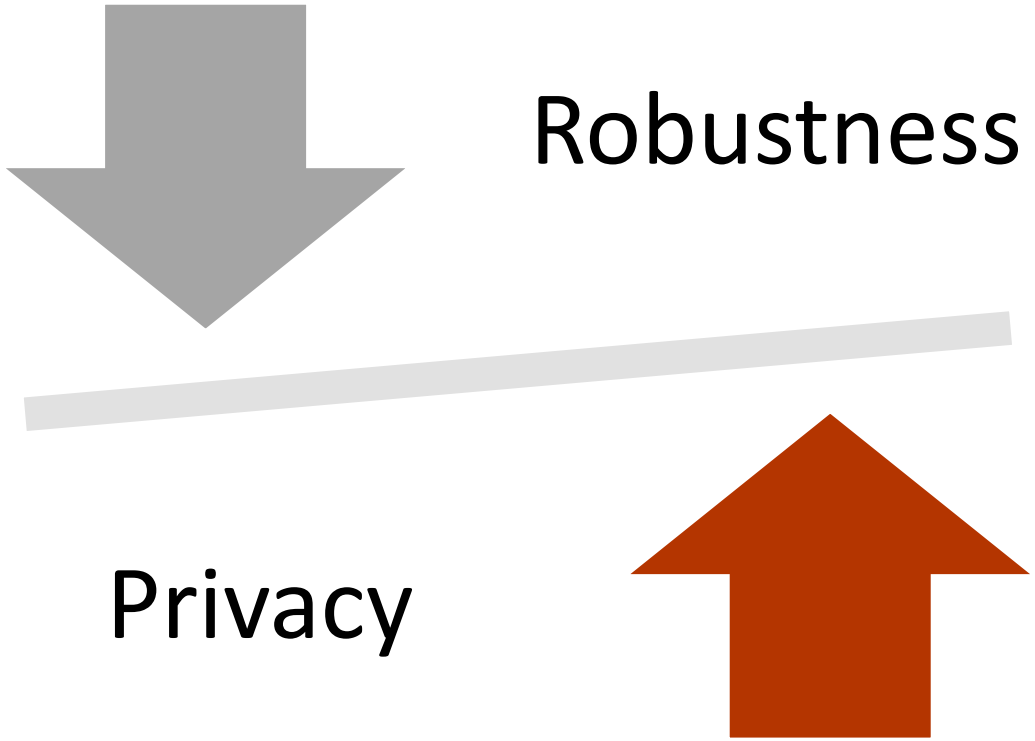


Etmann, Christian, et al. "On the connection between adversarial robustness and saliency map interpretability." arXiv preprint arXiv:1905.04172 (2019).

Accordance: Fairness v.s. Environmental Well-being



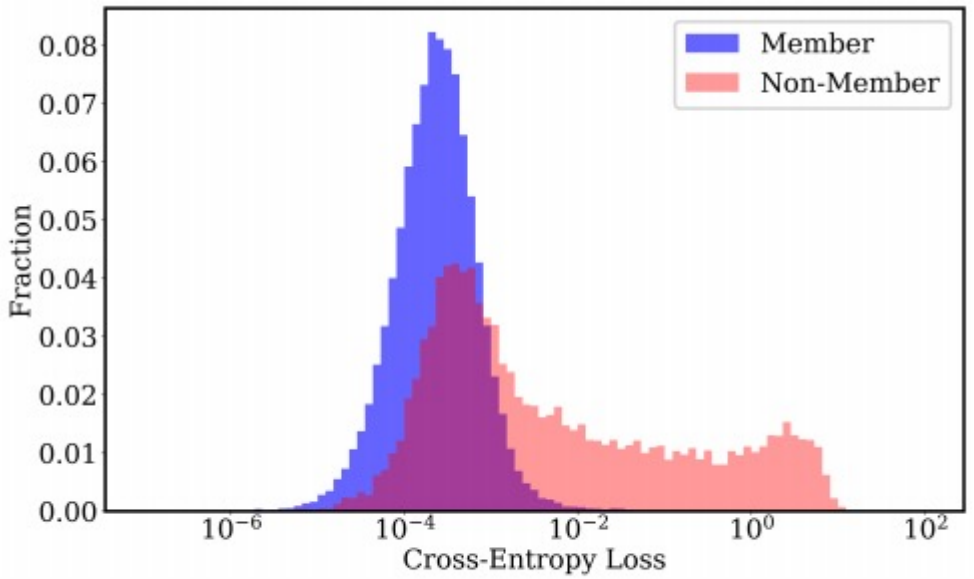
Conflicts: Privacy v.s. Robustness



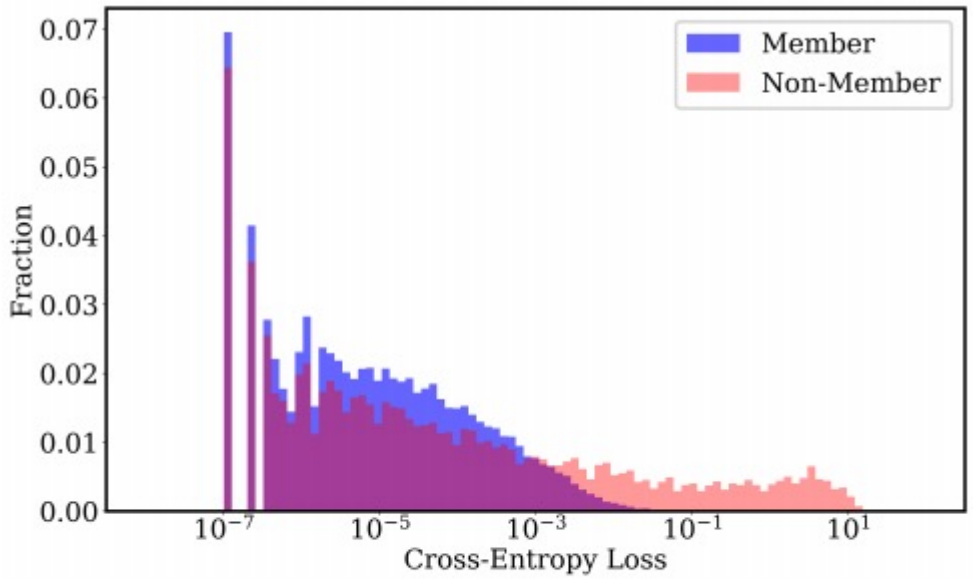
Models trained with adversarial defense approaches are more likely to expose sensitive information in training data via membership inference attacks.

The Privacy Risk of Securing Deep Learning Models against Adversarial Examples

Models trained with adversarial defense approaches are more likely to expose sensitive information in training data via membership inference attacks.



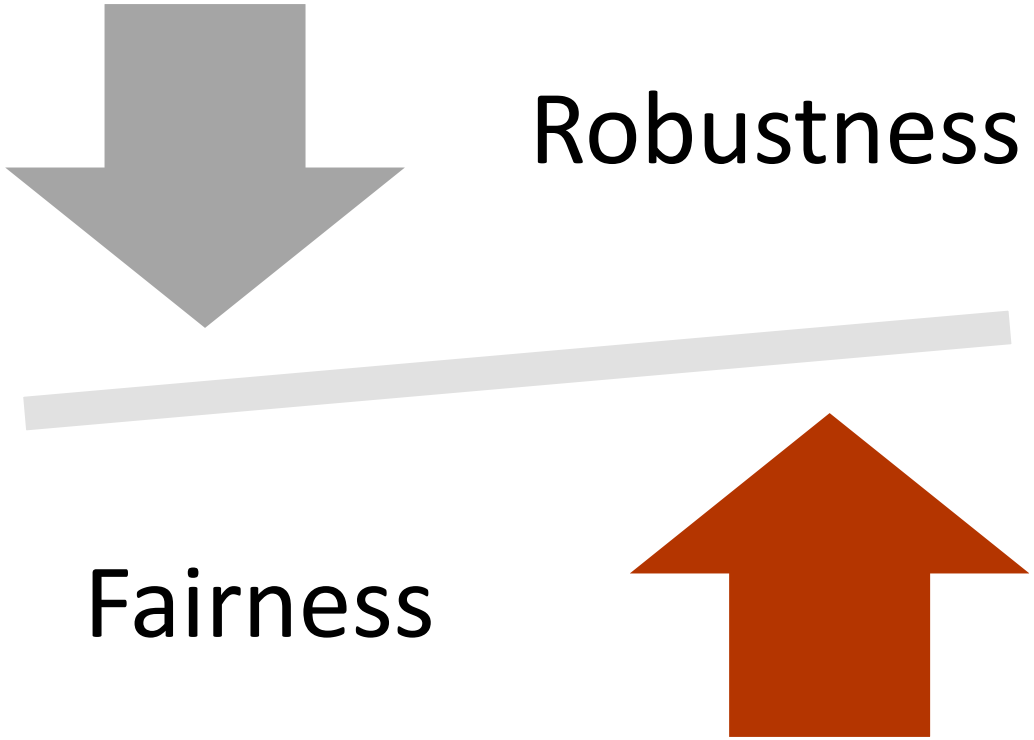
(a) Adversarially robust model from Madry et al. [33], with 99% train accuracy and 87% test accuracy.



(b) Naturally undefended model, with 100% train accuracy and 95% test accuracy. Around 23% training and test examples have zero loss.

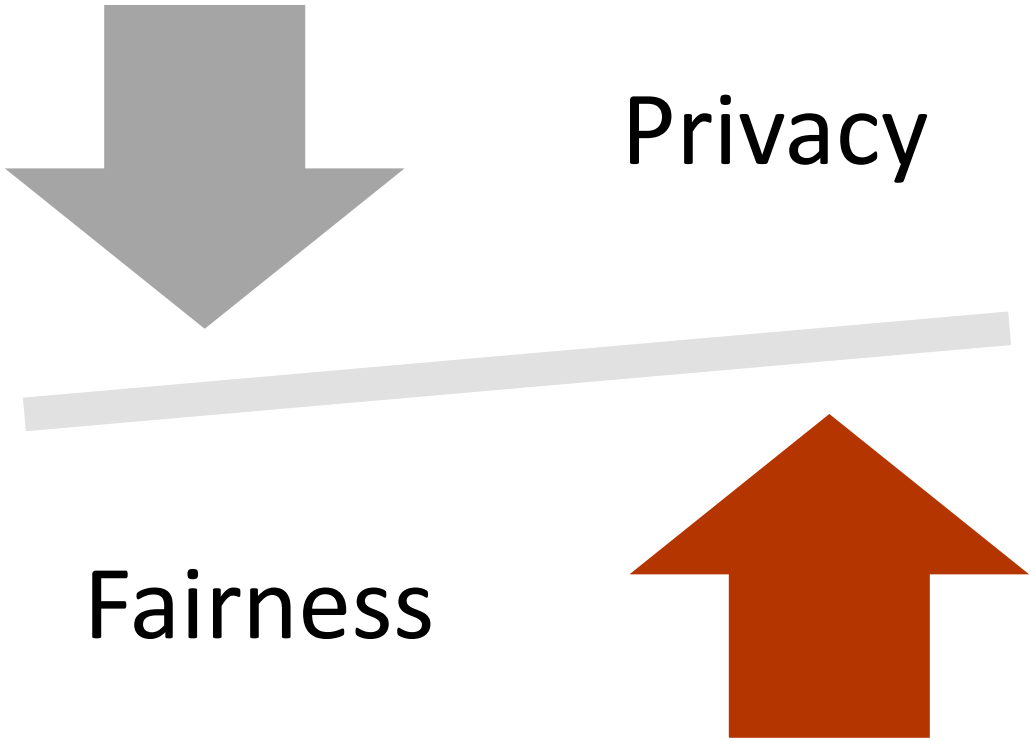
Song, Liwei, et, al. "Privacy risks of securing machine learning models against adversarial examples." 2019.

Conflicts: Fairness v.s. Robustness



Recent research indicates that adversarial training can introduce a significant disparity of performance and robustness among different groups, even if the datasets are balanced.

Conflicts: Fairness v.s. Privacy



Recent research theoretically proves that differential privacy and exact fairness in terms of equal opportunity are unlikely to be achieved simultaneously.