# 人工智能安全大作业

## 可选作业主题

1. 成员推断攻防
2. 公平性

## 成员推断攻防

### 一、实验概述

复现或自己提出一种成员推断攻击或防御方案，并至少在一个目标模型和一个目标数据集上完成相应测试

### 二、实验设置

1. 目标模型选择：ResNet50，VGG16等
2. 目标数据集选择：MNIST，CIFAR10，CIFAR100等

### 三、复现实验参考

以下三个主题供大作业选题参考：

1. （成员推断攻击）（2019 CCS）Privacy Risks of Securing Machine Learning Models against Adversarial Examples
   Paper：https://dl.acm.org/doi/10.1145/3319535.3354211
   Code：https://github.com/inspire-group/privacy-vs-robustness
2. （成员推断攻击）（2021 ICML）Label-only membership inference attacks
   Paper：https://proceedings.mlr.press/v139/choquette-choo21a.html
   Code：https://github.com/cchoquette/membership-inference
3. （成员推断防御）（2019 CCS）MemGuard: Defending against Black-Box Membership Inference Attacks via Adversarial Examples
   Paper：https://dl.acm.org/doi/10.1145/3319535.3363201
   Code：https://github.com/jinyuan-jia/MemGuard

## 公平性

### 一、实验概述

复现或自己提出一种公平性相关技术，并至少在一个目标模型和一个目标数据集上完成相应测试

### 二、实验设置

1. 目标模型选择：ResNet50，VGG16等
2. 目标数据集选择：MNIST，CIFAR10，CIFAR100等

## 三、复现实验参考

以下四个主题供大作业选题参考：

1. （预处理阶段的公平性提升技术）（2021 CVPR）Fair Attribute Classification through Latent Space De-biasing
   Paper: https://openaccess.thecvf.com/content/CVPR2021/papers/Ramaswamy_Fair_Attribute_Classification_Through_Latent_Space_De-Biasing_CVPR_2021_paper.pdf
   Code: https://github.com/princetonvisualai/gan-debiasing
2. （训练阶段的公平性提升技术）（2020 CVPR）Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation
   Paper: https://openaccess.thecvf.com/content_CVPR_2020/papers/Wang_Towards_Fairness_in_Visual_Recognition_Effective_Strategies_for_Bias_Mitigation_CVPR_2020_paper.pdf
   Code: https://github.com/princetonvisualai/DomainBiasMitigation
3. （训练阶段的公平性提升技术）（2019 ICCV）Balanced datasets are not enough：Estimating and mitigating gender bias in deep image representation
   Paper: https://openaccess.thecvf.com/content_ICCV_2019/papers/Wang_Balanced_Datasets_Are_Not_Enough_Estimating_and_Mitigating_Gender_Bias_ICCV_2019_paper.pdf
   Code: https://github.com/uvavision/Balanced-Datasets-Are-Not-Enough
4. （模型偏见检测技术）（2021 ICCV）Discover the Unknown Biased Attribute of an Image Classifier
   Paper: https://openaccess.thecvf.com/content/ICCV2021/papers/Li_Discover_the_Unknown_Biased_Attribute_of_an_Image_Classifier_ICCV_2021_paper.pdf
   Code: https://github.com/zhihengli-UR/discover_unknown_biases

# 实验提交

1. 完成一份实验报告，并提交代码，打包命名为“学号_组名_大作业”。
   实验报告要求包含：实验设计、关键实验代码分析、实验结果分析、实验总结与思考
   代码要求包含：关键模块的实现代码或者完整的工程代码
2. 提交方式：打包发送到助教邮箱xuruite@zju.edu.cn
3. 提交截止时间：2023年6月19日23：59前

# 实验要求

大作业为小组完成，共同提交一份报告。报告里需要写明组员，同一组的同学大作业成绩一致。
组队要求：1-2人一组。有跨专业的同学的组允许3人成组。

# 评分准则

实验目标（给出清晰的阐述）（5%）
实验环境（含数据集、模型等）（5%）
方案设计（20%）
方案创新性（5%）
代码与分析（20%）
实验结果与分析（30%）
总结与思考（5%）
撰写规范、排版工整（10%）