人工智能安全

主讲人：王志波 杨子祺

浙江大学计算机科学与技术学院/网络空间安全学院

课程介绍

浙江大学 ZHEJIANG UNIVERSITY

❑课程目标：

- ❑ 掌握人工智能安全基础知识
- ❑ 了解人工智能安全主流方向与研究进展
- ❑ 掌握多种人工智能安全攻防技术，如对抗样本生成与防御
- ❑ 为之后从事人工智能安全相关工作与科学研究打下坚实基础

❑课程形式：

- ❑ 理论（浙大教师） + 实验 （蚂蚁专家）
- ❑ 课堂出勤（10%） + 平时作业（30%） + 实验上机（20%） + 期末大作业（40%）

课程介绍

理论内容

- 人工智能安全概论
- 人工智能鲁棒性之对抗样本I
- 人工智能鲁棒性之对抗样本II
- 人工智能完整性之数据投毒
- 人工智能完整性之后门攻击
- 人工智能隐私性
- 人工智能公平性
- 人工智能可解释性

实验内容

- 人工智能鲁棒性与攻防I
- 人工智能鲁棒性与攻防II
- 人工智能隐私计算
- 人工智能伦理/可解释性

课程安排

Date	Lecture	备注
02/28	人工智能安全概论-1	
03/07	人工智能安全概论-2	
03/14	人工智能鲁棒性之对抗样本I-1	
03/21	人工智能鲁棒性之对抗样本I-2	
03/28	人工智能鲁棒性之对抗样本II-1	
04/04	人工智能鲁棒性之对抗样本II-2	根据学校调休情况上课
04/11	人工智能完整性之数据投毒-1	
04/18	人工智能完整性之数据投毒-2	实验课-蚂蚁专家

* 具体上课和上机时间，根据情况另行通知

课程安排

Date	Lecture	备注
04/25	人工智能完整性之后门攻击-1	实验课-上机
05/02	人工智能完整性之后门攻击-2	实验课-蚂蚁专家
05/09	人工智能隐私性-1	实验课-上机
05/16	人工智能隐私性-2	实验课-蚂蚁专家
05/23	人工智能公平性-1	实验课-上机
05/30	人工智能公平性-2	实验课-蚂蚁专家
06/06	人工智能可解释性-1	实验课-上机
06/13	人工智能可解释性-2	
06/17	**Final Report**	

* 具体上课和上机时间，根据情况另行通知

课程介绍

浙江大学 ZHEJIANG UNIVERSITY

❑参考书：

- ❑《深度学习》Ian Goodfellow、Yoshua Bengio、Aaron Courville
- ❑《对抗机器学习：机器学习系统中的攻击和防御》Yevgeniy Vorobeychik、Murat Kantarcioglu
- ❑《AI安全之对抗样本入门》兜哥
- ❑公众号：跟我学AI，马少平，清华大学

❑联系方式：

- ❑教师：杨子祺 yangziqi@zju.edu.cn
- ❑助教：徐瑞特 22221049@zju.edu.cn

谢 谢

浙江大学网络空间安全学院

https://icsr.zju.edu.cn/