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%） + 上机实验（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	
03/06	人工智能安全概论-2	
03/13	人工智能鲁棒性之对抗样本I-1	
03/20	人工智能鲁棒性之对抗样本I-2	
03/27	人工智能鲁棒性之对抗样本II-1	
04/03	人工智能鲁棒性之对抗样本II-2	
04/10	人工智能完整性之数据投毒-1	
04/17	人工智能完整性之数据投毒-2	04/16 实验课-蚂蚁专家

课程安排

Date	Lecture	备注
04/24	**人工智能完整性之后门攻击-1**	04/23 实验课-上机
05/01	**人工智能完整性之后门攻击-2**	根据学校调休情况上课
05/08	**人工智能隐私性-1**	05/07 实验课-蚂蚁专家
05/15	**人工智能隐私性-2**	05/14 实验课-上机
05/22	**人工智能公平性-1**	05/21 实验课-蚂蚁专家
05/29	**人工智能公平性-2**	05/28 实验课-上机
06/05	**人工智能可解释性-1**	06/04 实验课-蚂蚁专家
06/12	**人工智能可解释性-2**	06/11 实验课-上机
06/24	**Final Report**	

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：王志波 13886108869 zhibowang@zju.edu.cn
- ❑教辅：何泽青 13081066220

课程老师

王志波 教授/博导 国家优青 浙江大学网络空间安全学院

从事**智能物联网、人工智能安全、数据安全与隐私保护、边缘智能与安全**

- **CCF杰出会员、ACM/IEEE/电子学会 高级会员**
- **CCF 物联网专委会常委、电子学会网络空间安全专家委员会青年常委兼副秘书长**
- **人工智能学会智能信息网络专委会常委、电子学会物联网青年专技组常委**
- **ACM SIGBED常委、CCF 大数据专家委员会委员、CCF网络与数据通信专委会委员**

- 主持**国家自然科学基金项目4项**（**优青，联合基金重点项目、面上、青年**）
- 作为课题负责人或骨干承担**2项科技创新2030-新一代人工智能重大项目**（**人工智能安全理论及验证平台、数据安全与隐私保护下的机器学习技术**）、**973计划**等项目
- 发表高水平论文100余篇，其中**CCF推荐A类论文60余篇**，包括**CCS、S&P、MobiCom、CVPR、ICCV、INFOCOM、AAAI、JSAC、TMC、TDSC、TPDS、TC**等，**ESI高被引论文7篇**，Google引用**5000多次**
- 获授权发明专利16余项，公开发明专利10余项目，担任INFOCOM、WWW/ICDCS、AAAI、KDD程序委员会委员
- **电子学会自然科学一等奖、浙江省自然科学一等奖、IEEE 可扩展技术专委会职业中期研究成就奖、辽宁省自然科学学术成果一等奖、IEEE HPCC 2019杰出论文奖、FUSION 2019和Bigcom 2023最佳学生论文奖**、湖北省楚天学者、ACM武汉新星、电子学会优秀科技工作者、电子学会先进工作者

谢 谢

浙江大学网络空间安全学院

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