

复旦大学研究生课程教学大纲

课程名称/Course Title: 人工智能安全理论与实践

课程代码/Course Code: COMP627003

任课教师/Instructor(s): 曾剑平

开课院系/School/Department: 024 计算机科学技术学院

1. 课程概要/Course Summary			
课程名称（中文 Course Title（ Chinese）	人工智能安全理论与实践		
课程名称（英文） Course Title（ English）	Security Theory and Practice for Artificial Intelligence		
课程代码 Course Code	COMP627003	任课教师 Instructor(s)	曾剑平
开课院系 School/Department	024 计算机科学技术学院	开课学期 semester	2024-2025学年 第一学期
授课语言 Teaching Language	中文	适用学科专业 Discipline/ Specialization	0854 电子信息
学分数 Course Credit(s)	3	教学周数 Weeks	共16周
总学时 Teaching Hours in Total	共54学时	实验/实践学时 Hours for Experiments/ Practice	共0学时
预修课程要求 Pre-requisite Course(s)	人工智能基础；Python入门		
课程简介 Course Introduction	在当今，没有安全就没有人工智能应用。本课程针对专业学位研究生，围绕若干典型人工智能对抗场景案例，讲解人工智能安全的漏洞、相关理论与技术。课程内容包括人工智能安全基础、机器学习模型训练阶段和测试阶段的攻击与防御技术、模型与数据的隐私攻击与保护、大模型安全等。		
2. 教学目标/Course Objective			
目的在于让学生充分理解人工智能模型在实际应用中可能遭遇的各种安全风险，掌握人工智能安全知识体系，理解人工智能的典型机器学习模型及其安全漏洞，包括统计学习模型、深度学习模型等。理解并掌握基于这些漏洞的各种攻击方法，包括投毒攻击、逃避攻击、迁移攻击、隐私攻击、注入攻击和深度伪造等。理解并掌握数据层、算法层、模型层和系统层的AI防御方法。跟踪人工智能安全相关技术发展。			
3. 教学内容及进度安排/Course Content & Schedule			
课次/模块	教学周	教学内容及预期效果	作业/练习
1	1	人工智能的安全属性、问题、研究现状与知识体系	
2	2	以手写数字识别为例，介绍手写数字识别的模型与方法、机器学习模型的攻击者。	
3	3	介绍对抗样本生成理论和方	

3	3	介绍对抗样本生成理论和方法；手写数字识别模型的白盒攻击与应用	
4	4	介绍手写数字识别模型的迁移攻击与防御、黑盒攻击与防御；前沿文献解读。	
5	7	以入侵检测为对抗场景，介绍检测方法的数据处理、检测模型；介绍训练阶段的对抗机理和攻击方法，包括投毒攻击、后门攻击	
6	8	以入侵检测对抗场景为例，介绍AI模型的防御方法；针对正常样本的噪声数据检测；针对原始干净训练数据的非平衡、小样本增强入侵检测模型攻击的防御方法	
7	9	以贷款风险评估为例，介绍隐私安全问题，贷款评估模型的数据隐私；隐私安全概述。	
8	10	介绍贷款评估的数据处理、评估分类模型；分类模型的隐私攻击与推理	
9	11	介绍隐私计算框架，同态加密，差分隐私保护	
10	12	以爬虫对抗场景为例，介绍系统化的爬虫检测对抗技术；攻击与防御的智能化技术	
11	13	注入攻击的历史；伦理安全概述；大模型注入攻击、内容安全、AI伦理安全、价值对齐	
12	14	深度伪造方法概述；深度伪造攻击介绍；伪造检测，自然语言变体	
13	16	人工智能平台及其安全性；学生PJ交流、汇报	

4. 课程考核及成绩评定/Course Assessment & Grading

考核形式 Assessment Criteria	权重 Percentage	评定标准 Assessment Standard
出勤/Attendance	5	上课出勤情况
课堂表现/Participation	10	课堂上参与讨论的情况
作业/实验/实践/ Assignment(s)	25	课堂上报告论文或作业的情况，包括资料准备、汇报情况、回答问题情况等
课程论文/Course Paper	0	-
开卷考试/Open-book exam	60	答题的完整性、准确性等
闭卷考试/Close-book exam		
其他/Other(s)		

5. 教材/Textbook(s)

序号 No.	名称 Title	编著者 Author(s)	标准书号 ISBN	出版机构 Publisher	出版年月 Publication Date
1	人工智能安全	曾剑平	9787302611509	清华大学出版社	202208
6. 教学参考资料/Reading Materials and References					
7. 任课教师简介/Profile of Instructor(s)					
<p>曾剑平，副教授，从事大数据安全、AI安全、金融舆情安全等方面的研究。主持国家自然科学基金、上海市自然科学基金等课题。以第一作者或通讯作者在IEEE TIFS、Computers & Security、SCN等信息安全领域有影响力的期刊上发表多篇论文，在KBS、ESWA等人工智能相关的期刊上发表社交媒体分析、机器学习相关论文。在大数据安全、应用级攻防、爬虫对抗等方面获得授权发明专利7项。近年来出版《互联网大数据处理技术与应用》、《Python爬虫大数据采集与挖掘》以及《人工智能安全》三本书。为研究生开设了《大数据安全引论》、《人工智能安全理论与实践》。为本科生开设了《信息内容安全》、《泄密取证技术》课程，近年来多次评教成绩位列学院前10%。</p>					
办公地址 Office Add		江湾校区X2 A6027		办公时间 Office Hour	周一-周五
联系邮箱 Email Add		zjp@fudan.edu.cn		联系电话 Contact phone	