基于前沿研究和行业发展的人工智能安全课程创新探索

郭小丁 葛强** 牛童 王雅娣 陈国强

河南大学计算机与信息工程学院（密码学院），开封 475004

摘 要 随着人工智能技术的迅猛发展，其安全问题对国家、社会及个人构成重大挑战，课程建设也亟待革新。本文针对人工智能安全课程存在的问题，提出改革方案。旨在完善课程设置，通过增强基础知识、扩大课程覆盖面、强化实践操作、跟踪最新技术及融入伦理教育等措施，提升学生安全意识与实践能力，使课程内容紧跟行业前沿。改革后，学生满意度从约75%提升至接近85%，课程完成率从约80%上升至接近95%，表明课程改革成效显著，为培养适应人工智能安全领域需求的高素质人才提供了有效路径。

关键词 人工智能安全，课程改革，教学设计

An Innovative Exploration of an Artificial Intelligence Safety Curriculum Based on Cutting-Edge Research and Industry Developments

Guo Xiaoding Ge Qiang** Niu Tong Wang Yadi Chen Guoqiang

School of Computer and Information Engineering (Cryptography School) of Henan University
Kaifeng, Henan 475004, China
gxd@henu.edu.cn

Abstract– With the rapid development of artificial intelligence technology, its security issues pose significant challenges to the nation, society, and individuals, and the construction of related courses urgently needs innovation. This article addresses the problems in the existing AI security curriculum and proposes reform measures. The aim is to improve the curriculum by enhancing foundational knowledge, expanding the scope of courses, strengthening practical operations, keeping up with the latest technologies, and integrating ethical education. These steps are designed to raise students' security awareness and practical skills, ensuring the course content stays at the forefront of the industry. After the reform, student satisfaction increased from about 75% to nearly 85%, and course completion rates rose from around 80% to nearly 95%, indicating the significant effectiveness of the curriculum reform and providing an effective path for cultivating high-quality talents that meet the demands of the AI security field.

Key words- Artificial Intelligence Safety, Curriculum Reform, Instructional Design

1 引 言

随着人工智能技术的快速发展，人工智能安全问题日益凸显，恶意攻击、数据泄露、隐私侵犯等问题频发，对企业和个人的信息安全造成了严重威胁。同时，人工智能技术也广泛应用于国家安全、社会治理、经济发展等多个领域，人工智能安全问题已经成为国家安全和社会稳定的重要因素。在这一背景下，新工科建设应运而生，旨在培养适应新技术革命和产业变革的创新型工程科技人才。而拔尖创新人才培养更是成为当前教育领域的重要使命，人工智能安全领域作为前沿且关键的方向，对于拔尖创新人才的需求尤为迫切。

为了应对这些问题，有必要开设《人工智能安全》课程，培养具备人工智能安全意识和技能的人才。然而，在人工智能安全领域的课程的开设中，存在着诸如课程设置不够完善、实践课程较少、学生兴趣不够浓厚、宣传力度较少等问题。在之前的研究中，沈、房等[1]提出可以通过建立实践基地和开展项目式学习来增加

*基金项目：河南省高等教育教学改革研究与实践项目（2024SJGLX0051），河南省研究生联合培养基地项目(YJS2023JD28)，河南大学教学改革研究与实践项目(No. HDXJJG2023-105，HDXJJG2024-050，HDXJJG2023-106）。

通讯作者：葛强 gq@henu.edu.cn

学生的实践机会；李、梁等[2]提出可以通过教师培训、校企合作等方式来提升教师的实践能力，从而更好地满足人工智能安全课程的教学需求；杨、曾等[3]提出可以通过任务驱动和合作学习等方式，让学生在完成项目的过程中体验学习的收获，提高对人工智能安全的兴趣。Becker等[4]提出要将社会、道德和计算机科学课程结合起来。周、蒋等[5]提出了构建课程群教学体系、改革教学模式、完善教学评价、强化师资建设等多维度的创新策略。

然而，在之前的研究中，难以很好的同人工智能安全课程设计领域结合，仍然难以解决该领域中课程覆盖面窄、无法跟踪最新技术、无法提高学生解决问题能力等问题。针对上述不足，本文重点针对课程设置不够完善、实践教学较为缺乏、宣传力度不够三个问题，提出针对人工智能安全课程的改革方案，力在解决人工智能安全课程存在的问题。

本文的主要贡献如下：

- 提出目前人工智能课程开设中的主要现存问题。
- 针对人工智能课程中现存的问题提出解决方案。
- 根据解决方案提出相应的课程目标。

2 现存问题

随着人工智能技术迅猛发展，其在教育领域的应用日益广泛，为教育改革带来新机遇与挑战。在《人工智能安全》课程的改革过程中，虽取得一定成果，但仍面临若干亟待解决的问题：课程内容因技术更新滞后、教师知识与经验有限、学生基础参差以及教学资源与社会认知匮乏等多重制约。深入分析这些根源对推动课程持续发展至关重要。

2．1 课程教学设置不够完善

目前，许多高校和培训机构在开设人工智能课程时，仅侧重技术介绍与应用，如机器学习算法原理、深度学习框架操作及应用案例，却忽视了人工智能安全这一关键领域。近年来，生成对抗网络

（GANs）在图像生成与数据增强方面成就显著，但其潜在安全风险如虚假信息生成、版权侵权等在课程中鲜有提及[6]。这种滞后且不完善的设置导致学生虽掌握了技术操作与应用，却对安全知识一知半解，无法应对复杂多变的现实挑战。人工智能安全涵盖多个层面：模型训练与部署中可能遭遇对抗攻击导致输出错误，或数据投毒影响训练效果[7]。缺乏对这些威胁的了解，学生自然无法采取有效防护。因此，完善人工智能安全教学体系已迫在眉睫。

2．2 理论教学和实践教学未能全面共同开展

当前，人工智能课程在教学实施过程中，理论教学和实践教学未能实现全面共同开展。人工智能安全涉及范围广，实践性较强，不仅要求学生掌握相关理论知识，更要求学生能将所学的理论知识应用到实践问题中，能通过理论知识解决实际问题。然而，在实际教学过程中，往往过度侧重于理论知识的灌输，将大量课时用于讲解复杂的概念、算法原理以及安全理论框架等，而实践教学环节却相对薄弱，课时安排少，实验项目单一且与现实场景脱节严重[8]。长此以往，不仅不利于学生对人工智能

2．3 师资力量和宣传力度不够

由于人工智能安全是一门新兴学科，尽管实际需求大、专业人才缺口广，却因资源投入有限，导致具备教学经验的教师稀缺、师资整体薄弱。同时，该领域宣传不足，大量学生对人工智能安全了解不够，缺乏学习兴趣[9]。师资短缺直接制约教学质量，有限经验难以满足学生对知识探究和实践的需求，影响其技能提升。宣传缺乏又造成认知盲区，难以激发学习热情，削弱学科吸引力和人才流入。长此以往，不仅阻碍人才培养，也限制行业发展，加剧人才短缺与技术需求的矛盾，无法满足社会对 AI安全人才的迫切需求。

3 课程改革内容

《人工智能安全》课程改革旨在提高课程的教学质量和效果，培养学生的安全意识和技能，使学生更好地适应人工智能安全领域的发展需求。本文拟从增强基础知识、扩大课程覆盖面、强化实践操作能力、跟踪最新技术、强化伦理和社会责任等多个方面改革《人工智能安全》课程。总体改革思路如图1所示。

3．1 增强基础知识

在教育领域，任何课程的基础知识学习都是重中之重。《人工智能安全》课程应全面强化基本概念、原理和技术，不仅涵盖机器学习、深度学习、自然语言处理等核心算法的安全性分析，还需引入加密技术与漏洞修复等实用技能。在概念层面，应深入浅出地阐释安全威胁、攻击向量和风险评估等术语，帮助学生建立清晰认知框架；在原理层面，则系统剖析核心算法的安全机制，并结合加密与修复技能，加深对理论的理解。

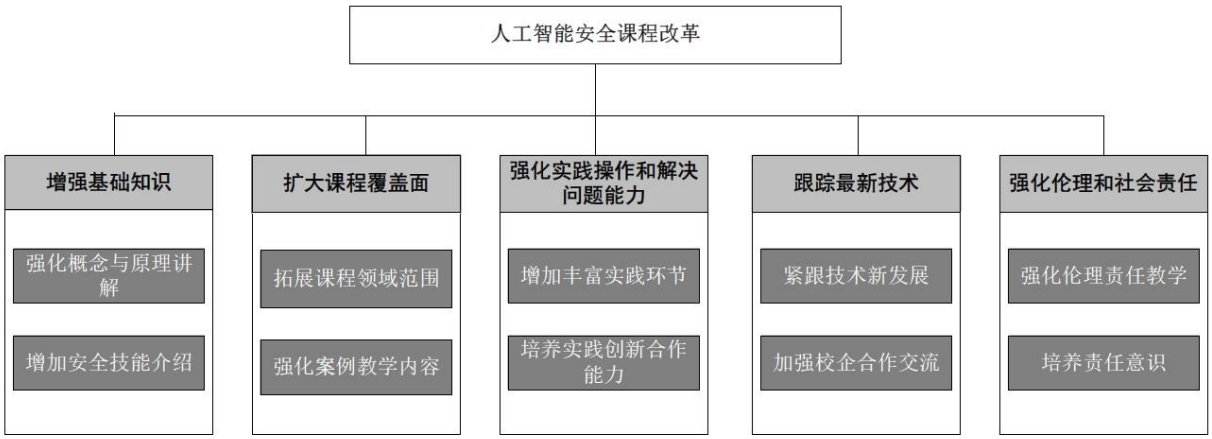

图 1 总体改革思路图

为切实推进上述改革目标的实现，课程建设应分阶段、有计划地实施：首先，通过调研与组建改革工作小组，全面评估学生基础与行业需求，明确课程建设方向；其次，设计模块化课程结构，涵盖基础概念、算法原理和安全技能三大核心板块，并结合学习目标设置具体考核标准，在教学过程中，采用线上线下结合的方式推动教学，并通过形成性评价手段动态调整教学策略；最后，基于学生反馈和数据分析持续优化教学内容，同时通过试点实施与教师培训，逐步推广课程改革成果，形成长效机制。

通过全面强化基础知识的讲解，学生不仅能够从本质上理解人工智能系统的安全机制，还能够建立起系统的安全思维模式。这将为他们深入学习课程的后续内容，奠定坚实的基础。

3．2 扩大课程覆盖面

为了扩大人工智能课程的覆盖面，更好地适应人工智能技术的快速发展，课程内容应在原有的机器学习和深度学习安全的基础上，进一步扩展到自然语言处理、计算机视觉、数据安全和隐私保护等新兴领域，将课程模块化为“自然语言处理安全”“计算机视觉安全”“数据安全与隐私保护”三大板块，并配套明确可测学习成果，开发渐进式教学手册与微课视频辅助教学；同时增加人工智能安全的实际案例的讲解和分析，提高学生的安全意识和风险意识。

在自然语言处理领域，课程应涵盖自然语言处理中的安全威胁，如恶意提示词攻击、模型窃取攻击、数据隐私泄露等，以及相应的防御措施。通过学习这些内容，学生将能够掌握数据安全和隐私保护的关键技术，为人工智能技术的安全应用提供保障。

通过对课程覆盖面的扩大，不仅丰富了课程的内涵，能够培养出具备多领域知识和技能的全面人才，也可以帮助学生更好地理解和掌握人工智能安全的知识和技能，提高对人工智能技术的安全意识和风险应对能力。

3．3 强化实践操作和解决问题能力

为了进一步提升学生在人工智能安全领域的实践能力，应增加实验、项目和案例分析等环节，以强化动手操作与问题解决。课程通过真实人工智能安全问题的深入探讨，帮助学生理解系统在实际场景中的运作与安全挑战，例如分析并破解自有模型或模拟攻击检验防御策略。借助项目实践和小组讨论，学生将从需求分析、系统设计到实施与测试全程参与人工智能安全项目，锻炼解决问题能力、培养创新与合作精神，并深化对相关知识的掌握。

3．4 跟踪最新技术

随着人工智能技术的不断进步，新的安全威胁和攻击方法也层出不穷。随着技术的进步，课程也需要不断跟踪这些最新的技术发展，及时将最新的安全实践和防护策略引入课堂。课程将及时引入最新的安全实践和防护策略，确保学生能够掌握最新的防御技术。同时课程将密切关注人工智能安全领域的行业趋势和最新研究成果，确保教学内容与实际应用紧密结合。

课程可以邀请具备实践经验的专家来授课或者指导实践项目，让学生更好地理解和掌握人工智能安全的知识和技能。也可以邀请具备丰富实践经验的专家

来开展讲座活动，分享他们在人工智能安全领域的实际经验和最新研究成果。这些专家将通过案例分析、技术讲解和实践指导，帮助学生更好地理解和掌握人工智能安全的前沿知识和技能。同时可以建立校企联合顾问委员会与企业导师制，在校企顾问委员会的指导下，每半年按版本迭代优化课程内容与实验平台，确保教学资源、案例与防护策略始终与人工智能安全前沿同步。

通过跟踪最新技术，能够确保课程内容紧密跟随人工智能安全领域的发展趋势，及时将新兴的安全威胁、攻击方法以及对应的防护策略纳入教学体系，使学生所学知识与实际应用紧密相连，有效应对不断变化的安全挑战。

3．5　强化伦理和社会责任

关于强化伦理和社会责任的改革首先从目标和共建机制入手，明确“公正性、透明度、问责制、隐私尊重”等伦理核心价值，并通过伦理学家、法律专家与社区代表共同组建顾问小组，确保课程目标和内容兼顾技术与社会诉求；在此基础上，以算法偏见、可解释性、合规问责和数据伦理为四大模块，将伦理议题系统化融入课程架构，并开发涵盖典型争议案例（如算法歧视、面部识别滥用）的四步教学模板，配以视频素材，帮助学生全面理解伦理冲突的多维视角。

随后，改革聚焦体验式与项目式教学，设计实战项目，让学生从受众角度发起偏见与隐私攻击测试并提出改进建议，同时组织社区调研，促使理论与社会实践相结合；在评估层面，结合案例分析报告、角色扮演和公开答辩等多维度手段检验学生的价值判断与沟通能力；最后，通过每学期审阅最新研究与法规，以及定期的社会影响座谈，建立持续优化与反馈闭环，确保课程始终与人工智能安全伦理前沿和社会责任要求同步。

4　教改成效

为了明确改革成效，共统计了281名学生，其中改革前132名，改革后149名。改革前有79名学生对课程比较满意，有106名学生完整完成该课程；改革后有127名学生对课程比较满意，有141名学生完整完成该课程。教改成效如图3所示。从图表中可以看出，课程教学改革取得了显著的成效。在改革前，学生满意度大约为75%，改革后提升至接近85%，这表明课程内容和教学方法的改进使得学生对课程的认可度大幅提高；课程完成率也从改革前的约80%上升至接近95%，说明学生的学习积极性和坚持度有了明显增强。这些数据充分表明，通过增强基础知识、扩大课程覆盖面、强化实践操作、跟踪最新技术及融入伦理教育等措施，课程不仅提升了学生的专业知识和技能，还有效激发了学生的学习兴趣和主动性，使教学质量和效果都得到了显著提升。

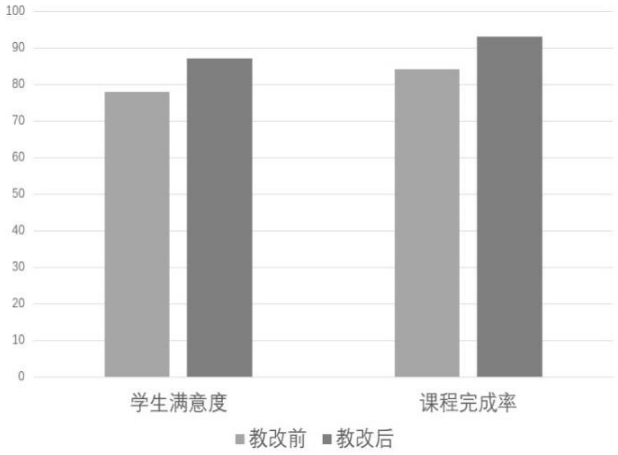

图 3　教学改革与实践效果图

5　结束语

综上所述，人工智能安全课程的建设与改革是一项系统而长期的任务。在当今时代，人工智能技术飞速发展，对专业人才的需求日益增长。本文通过深入分析现存问题并提出针对性的课程改革内容，为提升教学质量、培养高素质人才提出探索路径。为了解决人工智能课程的种种现存问题，本文认为应持续关注行业动态，不断完善课程体系，强化学生的实践能力，加强师资队伍建设，加大宣传力度，推动思政教育与专业教育的深度融合。通过上述改革措施，能有效解决现有问题，提高人工智能课程的教学效果。

参 考 文 献

[1] 沈苑，房斯萌，柳晨晨，等. 生成式人工智能教育应用治理：案例与反思[J]. 开放教育研究，2024，30(06):39-47.

[2] 李秋霞，梁震. 人工智能时代教师专业发展路径探寻[J]. 教育理论与实践，2022，42(34):54-58.

[3] 杨现民，曾佳尧，李新. 人工智能与教育深度融合的场景细化及落地实践——基于探索性多案例分析法[J]. 开放教育研究，2025，31(01):82-92.

[4] BECKER B A. Integrating society, ethics and the computing profession with computer science curricula 2023[C]//Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 2. New York, NY, USA: [s.n.], 2024: 813.

[5] 周延泉,蒋思,张玥,等.人工智能背景下的自然语言处理研究生课程群建设模式研究[J] .计算机技术与教育学报,2024,12(3),p6-10.

[6] 南婷. 生成式人工智能虚假信息的风险与治理策略研究[J]. 中国传媒科技, 2024(8):129-133.

[7] 汪旭童, 尹捷, 刘潮歌, 等. 神经网络后门攻击与防御综述[J]. 计算机学报, 2024, 47(8):1713-1743.

[8] BATES R, HARDWICK J, SALIVIA G, et al. A project-based curricu- lum for computer science situated to serve underrepresented populations [C]//Proceedings of the 53rd ACM Technical Symposium on Computer Science Education - Volume 1. New York, NY, USA: [s.n.], 2022: 585 - 591.

[9] 陈磊,李雅静. 人工智能系统安全综述[J]. 信息通信技术与政策, 2021, 47(08):56-63.