国产大模型融入《信息安全》课程的思政教学模式研究

魏岸若　康　慷　郭宏园　杨　旭（通讯作者）　续　好

重庆工贸职业技术学院 重庆 408000

【摘　要】：随着生成式人工智能技术的爆发式发展，文心一言、讯飞星火等国产大模型已成为国家数字经济战略的核心载体，其“价值对齐”的伦理属性与“数据安全”的技术属性共同构成科技安全的重要维度。当前《信息安全》课程教学存在“技术传授与价值引领脱节、传统知识与前沿需求错位”的双重困境：一方面，课程仍聚焦于传统网络攻防，对大模型引发的数据投毒、模型提取等新型安全风险覆盖不足；另一方面，学生普遍存在“技术工具理性至上”倾向，对 AI 技术的国家战略意义、社会伦理责任缺乏深度认知。针对上述问题，本文提出“价值-知识-能力”（Value-Knowledge-Ability，VKA）三元融合的思政教学模式，以国产大模型为核心案例载体，验证模式有效性，为新工科背景下计算机类课程思政提供可复制、可评价的实践范式。

【关键词】：课程思政；国产大模型；价值对齐；信息安全；VKA 三元融合模型

DOI:10.12417/2705-1358.26.05.084

1 引言

全球 AI 领域已进入“大模型竞赛”新阶段，据中国互联网络信息中心发布《生成式人工智能应用发展报告（2024）》统计，截至 2024 年 7 月，我国完成备案并上线、能为公众提供服务的生成式人工智能服务大模型已达 190 多个，覆盖政务、金融、医疗等关键领域。然而，大模型的“价值对齐”与“数据安全”问题日益凸显：2023 年 OpenAI 因 ChatGPT 数据跨境传输违反欧盟《通用数据保护条例》（GDPR）被意大利数据防护管理局罚款 1500 万欧元；2024 年，某国产大模型因违规提供生成式人工智能服务问题，当地网信部门责令整改。从国家战略层面看，《“十四五”数字经济发展规划》明确提出“加强人工智能伦理治理，保障数据安全与主权”（引用）；2024 年《生成式人工智能服务管理暂行办法》进一步要求“生成式 AI 服务提供者应确保模型输出符合社会主义核心价值观，防范数据安全风险”。这意味着，培养具备“技术能力+价值素养”的 AI 安全人才，已成为保障国家科技安全的重要任务。

当前《信息安全》课程教学存在两大问题：其一，知识内容滞后于技术前沿。传统课程体系以“网络安全、系统安全”为核心，对大模型特有的数据投毒攻击等安全风险覆盖不足。其二，价值引领缺失于技术传授。工科学生普遍存在“技术至上”思维，据笔者前期访谈（N=30），66.7%的学生认为“信息安全只需关注技术防护，伦理与国家战略是政策层面的事”。这种“重技术轻价值”的倾向，导致学生难以形成符合国家需求的职业素养。

现有研究多停留在“呼吁将 AI 伦理融入教学”的理论层面，缺乏具体路径：如何以国产大模型为案例，将“价值对齐”的思政内涵与“数据安全”的技术知识、实践能有机融合；如何构建可测量、可评价的教学体系，避免思政元素“贴标签”“两张皮”；这正是本文需解决的核心问题。国外学者较早关注 AI 伦理与安全教育：2019 年，MIT 媒体实验室研究人员在第二届马萨诸塞州大众 STEM 周上试行了名为“如何训练你的机器人”（How to Train Your Robot）的开源课程，通过互动实践向中学生传授人工智能原理与伦理责任，该课程还同步开展了教师专业发展培训以提升教学适配性[1]。斯坦福大学人类中心人工智能研究所（Stanford HAI）于 2024 年发布白皮书《Rethinking Privacy in the AI Era: Policy Provocations for a Data-Centric World》，系统分析了人工智能（尤其是大语言模型）带来的隐私风险，探讨了数据保护立法对 AI 技术发展的规制作用，并提出“默认限制数据收集”“规范 AI 数据供应

作者信息：1、魏岸若（1988-），男，重庆涪陵人，硕士，副教授，研究方向：软件工程，单位：重庆工贸职业技术学院

2、康慷（1998-），男，湖北蕲春人，硕士，助教，研究方向：深度学习，单位：重庆工贸职业技术学院

3、郭宏园（1991-），女，重庆涪陵人，本科，数学一级教师，研究方向：应用数学，单位：重庆市涪陵城区第十四小学校

4、杨旭（1998-），男，河南信阳人，硕士，助教，研究方向：人工智能，单位：重庆工贸职业技术学院

5、续好（1991-），女，重庆涪陵人，本科，助教，研究方向：思想政治，单位：重庆工贸职业技术学院

基金信息：1、本文是重庆市教育科学规划课题 2025 年度教学改革研究专项一般课题“国产大模型驱动的高职计算机类专业‘技术赋能与思政引领’双螺旋融合教学改革研究”研究成果，课题批准号：K25ZG3080285。2、本文是教育部信息化教指委 2025-2026 年度全国高等职业院校人工智能专业建设暨数字教材研究课题“基于产业工作逻辑的数据采集技术‘自适应进化式’数字教材与资源生态构建研究”研究成果，项目编号：KT26010423。本文是重庆工贸职业技术学院 2024 年度教育教学改革研究项目重点项目“新工科背景下高职计算机类专业‘双层递进式融合’课程思政与实践”研究成果，项目编号：JG20240104。4、本文是重庆工贸职业技术学院 2024 年度教育教学改革研究项目“基于‘百度飞桨’高职人工智能课程教学改革研究”研究成果，项目编号：JG20240241。

链”等政策建议[2]。但国外研究多聚焦于“通用伦理”，缺乏“国家主权、科技自主”等战略维度的价值引领，与我国课程思政需求存在本质差异。国内关于课程思政与人工智能融合的研究可大致分为两类：其一，“技术赋能思政”的理论框架探索。例如，李培根院士多次强调，工科教育需实现“技术理性”与“价值理性”的深度融合，可以通过技术案例潜移默化地渗透思政元素，引导学生关注技术背后的人文意义与社会责任。这一观点为“计算思政”提供了重要的理论支撑[3]。其二，聚焦于 AI 技术在思政教育中应用的实证研究开始涌现，但多集中于通识课程或传统教学模式。例如，四川大学通过训练专属 AI 助手，构建“师-生-机”三元共生的思政育人体系，实现了教学资源的动态生成与个性化学习支持[4]；电子科技大学在 2024 年上线“i 思政大模型”，依托 DeepSeek 技术为思政课程提供智能备课、个性化答疑与学情分析服务，构建了“AI+大思政课”的教学新生态[5]。

现有研究的共同不足：①载体滞后：未将国产大模型这一核心战略载体纳入教学；②路径模糊：缺乏“价值-知识-能力”融合的具体教学设计；③评价缺失：未构建可测量的思政育人效果评价体系。本文围绕“VKA 三元融合”模型展开，具体内容包括：①界定“价值对齐”“数据安全”的思政内涵与技术边界；②构建“情境导入-探究内化-实践升华”的教学流程，设计“国产大模型风险评估”项目；③以某高校《信息安全》课程为实践对象，通过定量+定性方法验证模式有效性；④总结模式的创新点、挑战与推广路径。理论意义：丰富“新工科”课程思政理论，提出“前沿技术载体+三元融合”的育人框架，填补国产大模型与思政教育结合的研究空白。实践意义：为一线教师提供可直接复用的教学方案，解决“思政如何融入前沿技术教学”的实操难题；培养具备“科技报国情怀+AI 安全能力”的人才，服务国家战略需求。

2 核心概念与理论基础，构建“VKA 三元融合”模型

2.1 核心概念界定

2.1.1 价值对齐：从技术到思政的升华

技术层面，“价值对齐”指“AI 行为符合人类意图”（Amodeietal.,2016），如通过 RLHF 优化模型输出；思政内涵需结合我国国情：

①国家战略维度，对齐“科技自主可控”，避免依赖国外框架；

②伦理法治维度，对齐“社会主义核心价值观”，符合《网络安全法》；

③职业素养维度，对齐“科技工作者社会责任”，如保护数据隐私[6]。

2.1.2 数据安全：AI 语境下的新外延与思政关联

传统“数据安全”聚焦存储与传输保密，AI 语境下延伸为：

①数据源头安全（训练数据合规性）；

②模型训练安全（数据投毒攻击）；

③模型应用安全（模型提取攻击）。

从思政视角看，数据安全与国家安全观深度绑定：训练数据国产化关系大模型自主可控，模型提取攻击本质是“技术窃密”，需通过教学强化风险认知。

2.2 理论基础

建构主义学习理论强调“真实情境中的主动建构”（Piaget,1970）[7]，国产大模型安全案例可引导学生自主思考“数据安全与国家法规的关系”，避免思政被动灌输。课程思政理念要求“显性教育与隐性教育统一”，教育部《高等学校课程思政建设指导纲要》明确“所有课程都负有育人责任”[8]。在《信息安全》课程中，“大模型数据安全原理”是显性知识，“科技报国情怀”是隐性价值，如讲解 RLHF 时，同步引导思考“国产大模型如何通过 RLHF 对齐核心价值观”。

2.3“VKA 三元融合”模型

基于上述理论，构建“VKA 三元融合”模型，核心逻辑为：以“价值塑造（V）”为引领，“知识探究（K）”为基础，“能力建设（A）”为目标，三者相互促进（见图 1）。

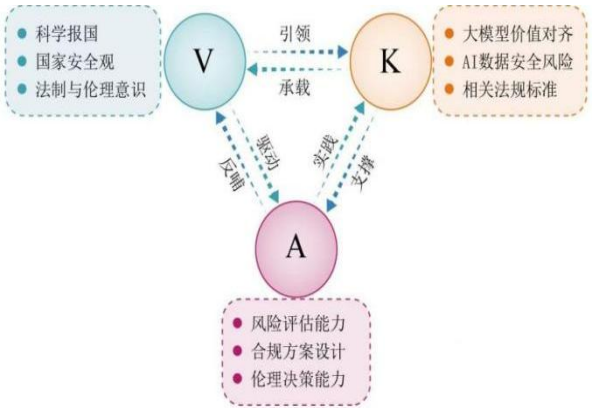

图 1 VKA 三元融合模型结构

3 教学模式的构建与实施路径

3.1 教学目标重构：VKA 三维目标体系

传统课程仅聚焦“知识与能力”，本文构建三维目标（表 1），确保思政目标可落地：

表 1 《信息安全》课程 VKA 三维教学目标

维度	具体目标
价值目标（V）	1.认同国产大模型自主可控的战略意义，树立科技报国情怀； 2.理解 AI 数据安全与国家安全的关联； 3.掌握《生成式 AI 服务管理暂行办法》，形成法治意识。
知识目标（K）	1.掌握大模型价值对齐技术（RLHF、多模态过滤）； 2.识别 AI 数据安全风险（投毒、提取、隐私泄露）； 3.理解《网络安全法》对大模型的合规要求。
能力目标（A）	1.能对国产大模型开展数据安全风险评估； 2.能设计大模型合规方案； 3.能在技术决策中融入伦理考量。

3.2 教学内容重组：增设“AI 安全”模块

在传统“网络安全、系统安全”模块后，增设 24 课时“AI 安全”模块（占总课时 30%），按“价值-知识-能力”展开：①价值导入（4 课时），通过中外大模型安全事件对比，融入“科技自主”元素；②知识讲授（8 课时），讲解价值对齐技术、安全风险与法规；③能力实践（12 课时），开展国产大模型风险评估项目。为避免思政碎片化，设计“AI 安全”模块知识图谱（表 2），明确技术、思政与能力的映射关系：

表 2 “AI 安全”模块知识表

技术知识点	对应思政元素	支撑能力目标	教学案例
RLHF（基于人类反馈的强化学习）	科技报国（国产 RLHF 自主可控）	合规方案设计能力	文心一言与 GPT-4 的 RLHF 差异对比
数据投毒攻击	国家安全（金融大模型受影响）	风险评估能力	2023 年金融大模型数据投毒事件分析
《生成式 AI 服务管理暂行办法》	法治意识（大模型合规运营）	合规方案设计能力	某大模型违规整改的法规适用分析

3.3 教学方法设计：“三段式”流程

基于建构主义，设计“情境导入-探究内化-实践升华”三段式流程，实现“价值-知识-能力”同步推进（见图 2）：

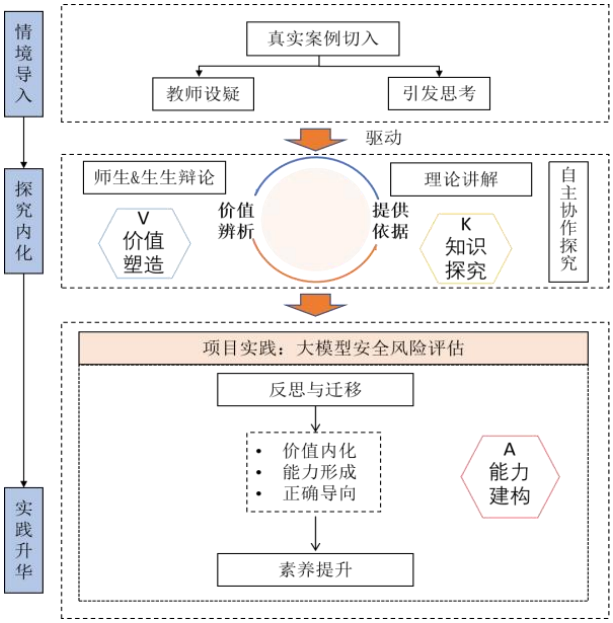

图 2 三段式教学实施流程

3.4 教学评价改革：多维度体系

构建“过程+终结”评价体系，将价值素养纳入评分标准（表 3），避免主观打分：

表 3 多维度教学评价体系

评价类型	评价内容	评分标准（总分 100 分）	权重
过程性评价	1.红蓝对抗表现； 2.课堂价值问题互动； 3.项目中期汇报。	1.结合国家战略论述（10 分）； 2.提出深度价值问题（5 分）； 3.体现国家安全考量（5 分）。	20%
终结性评价	国产大模型风险评估报告	1.技术风险分析准确（30 分）； 2.合规方案符合法规（20 分）； 3.体现国家安全与伦理（20 分）； 4.报告逻辑清晰（10 分）。	80%

4 教学实践与效果分析

4.1 实践对象与周期

实践对象为某高校计算机科学与技术专业 2023 级 2 个班（实验组 48 人，对照组 46 人），两班基础无显著差异（P>0.05）；周期为 2024 年春季学期（16 周），实验组采用“VKA 模式”，对照组用传统教学；课程总课时 80，实验组 24 课时为“AI 安全”模块。

4.2 数据收集方法

采用“定量+定性”结合：①问卷调查（Likert5 点量表），

测查“价值认知、知识掌握、能力自评”，教学前后发放；②深度访谈，选取实验组优中差各 4 名共 12 名学生，了解价值认同变化；③项目成果分析，统计报告中“提及国家安全、引用法规、融入伦理”的比例，提取典型表述。

4.3 效果分析

4.3.1 定量结果

独立样本 t 检验显示（表 4），实验组教学后（T2）在“价值认知（4.28）、知识掌握（4.35）、能力自评（4.15）”维度均显著高于教学前（T1）及对照组 T2（P<0.01），其中“国产 AI 认同度”从 2.76 提升至 4.12，“国家安全观”从 2.89 提升至 4.21。

表 4 两班问卷数据前后对比（平均分±标准差）

维度	价值认知	知识掌握	能力自评
实验组（T1）	2.95±0.52	2.88±0.55	2.72±0.58
实验组（T2）	4.28±0.41	4.35±0.38	4.15±0.43
对照组（T1）	2.98±0.49	2.92±0.51	2.75±0.56
对照组（T2）	3.02±0.53	2.91±0.54	2.83±0.57
实验组 T2vsT1（P）	<0.01	<0.01	<0.01
实验组 T2vs 对照组 T2（P）	<0.01	<0.01	<0.01

4.3.2 定性结果

访谈中，学生刘某提到“明白文心一言训练数据国产化是为避免卡脖子，关乎科技安全”；学生杨某表示“辩论后意识到大模型不符合法规再先进也不能用，法治意识很重要”。项目报告中，某小组指出“讯飞星火医疗模块需增加数据验证，符合《生成式 AI 办法》第 8 条，体现科技工作者责任”；另一小组提到“模型提取攻击威胁国家科技主权，需加 API 限制与参数加密”。

4.3.3 项目成果统计

46 份报告中，93.5%提及“国家安全意义”（T1 仅 17.4%），89.1%准确引用法规，82.6%融入伦理考量，这些数据表明，学生已能自觉地将价值判断与伦理规范融入专业分析之中，证明“价值-知识-能力”已融合。

5 讨论与反思

5.1 模式的创新价值

本模式的创新性体现在载体与体系的双重突破。在载体上，将“国产大模型”这一国家战略与前沿技术交汇点作为核心案例，使思政元素从外部附加转为内生融合。在体系上，构建的“VKA 三元融合”模型及配套的“三段式”教学流程，将价值目标转化为可教、可学、可评的具体教学活动，并借助量化数据与质性分析，实现了思政效果的可视化评估，为课程思政提供了可复制的实践范式。

5.2 实践挑战与应对

实践中的挑战主要来自三方面：学生认知、教师能力与教学资源。针对部分学生初期的“技术至上”观念，通过真实案例冲击与角色代入式研讨成功扭转。为弥补教师在大模型与思政教学融合上的知识短板，采取了校企协同培训与校内跨学科共备课程的策略。面对案例与平台的资源短缺，则通过企业脱敏数据与开源工具予以解决。

5.3 推广建议

本模式可向《人工智能》《软件工程》等课程推广，通过对比技术路径、嵌入法规要求等方式实现价值引领。推广中需分层实施：高职侧重案例与操作，本科侧重研讨与设计；资源受限院校可依托开源社区与虚拟平台，实现模式的低成本、广适配应用。

6 结论与展望

在 AI 时代，《信息安全》课程需从“培养技术操作者”升级为“培养民族复兴大任的 AI 安全人才”，将国产大模型“价值对齐”与“数据安全”纳入课程，是落实“立德树人”的必然要求。本文构建的“VKA 三元融合”模式，通过“三段式”流程实现“价值-知识-能力”统一，实践证明能显著提升学生的国产 AI 认同度与国家安全观，为计算机类课程思政提供可操作路径。研究存在局限性：实践样本仅覆盖 1 所高校，未来需扩大至高职、应用型本科院校对比；后续可深化产教融合，共建“AI 安全思政实训基地”，研发“AI+思政”个性化教学工具，持续完善育人体系。

参考文献：

[1] Breazeal C, Williams R. Bringing artificial intelligence and MIT to middle school classrooms[EB/OL]. (2019-12-30)[2025-12-10].

[2] King J, Meinhardt C. Rethinking Privacy in the AI Era: Policy Provocations for a Data-Centric World[R]. Stanford: Stanford HAI, 2024.

[3] 李培根.人文情怀与工程实践教育[J].高等工程教育研究,2010 (4):10-13.

[4] 四川省教育厅.四川大学:坚持“四个着力”,以人工智能技术赋能思政课建设 [EB/OL]. (2025-09-12)[2025-12-17].

[5] 全国高校思想政治工作网.电子科技大学:抓住“四个关键”推动人工智能赋能思想政治工作 [EB/OL]. (2025-06-09)[2025-12-20].

[6] Amodei D, Olah C, Steinhardt J, et al. Concrete Problems in AI Safety[EB/OL]. (2016-06-21)[2025-12-15].

[7] Piaget J. The Construction of Reality in the Child[M].New York:Basic Books,c1954.

[8] 教育部.高等学校课程思政建设指导纲要 [EB/OL]. (2020-06-05)[2025-12-20].