权限、幻觉与让渡

——生成式人工智能对高校辅导员工作的安全伦理挑战及规制

宁 康[1] 谭伟坚[2*]

1.广东技术师范大学教育科学学院 广东 广州 510450

2.广东技术师范大学物理与光电工程学院 广东 广州 510450

【摘 要】：生成式人工智能嵌入高校辅导员工作，引发权限、幻觉与让渡三重风险：权限风险表现为系统越权与数据泄露；幻觉风险指模型语义偏差转化为实际误导；让渡风险体现为工具依赖与专业能力退化。本文构建“权限规制—幻觉防控—让渡纠偏”协同治理体系，实现技术赋能与人文底线的平衡。

【关键词】：生成式人工智能；高校辅导员；安全风险；伦理困境；治理路径

DOI:10.12417/3041-0630.26.06.100

1 引言

生成式人工智能的跨越式发展正在深刻重塑高等教育的运行图景。与传统的管理信息系统不同，生成式 AI 具备强大的自主生成与系统交互潜力。当辅导员授予其访问学生数据库、发送通知、生成鉴定意见等权限时，一个根本性问题浮出水面：我们是否正在将本应属于人类教育者的判断权、执行权乃至情感联结，悄然让渡给一个基于概率运算的算法模型？现有研究多聚焦赋能价值，对生成式 AI 特有的风险机理缺乏系统剖析。

本文借鉴“权限”“幻觉”“让渡”三个核心概念，将其适配至高校辅导员工作场景，构建系统性的风险分析框架，从三个维度剖析风险，探究深层成因，并提出治理路径，以期为人工智能时代守护辅导员工作的安全底线与人文价值提供理论参照与实践指南。

2 权限风险：生成式 AI 的系统越权与数据安全隐患

“权限”是理解生成式 AI 安全风险的首要维度。与传统的办公软件不同，生成式 AI 工具若被授予过高的系统访问权限，其“自主性”可能从“便利”异化为“失控”。在辅导员工作场景中，权限风险主要表现为操作边界的模糊、外部攻击的渗透以及数据流动的失控。

2.1 从“对话”到“执行”：授权边界的悄然模糊

生成式 AI 本质上是基于概率模型的文本生成工具，而非具有自主意图的行为主体。然而，当辅导员将其接入学工系统、学生数据库或即时通讯平台时，AI 的“对话能力”便在事实上获得了“执行能力”。这种从“对话”到“执行”的边界模糊，根植于生成式 AI 的技术架构。当 AI 获得系统级执行权限后，其操作边界往往超出用户的原始授权意图。

在辅导员工作场景中，这种风险尤为值得警惕。学生档案、成绩单、心理测评结果等信息属于敏感数据，若 AI 在未被明确授权的情况下访问、修改或传输这些信息，不仅可能侵犯学生隐私，还可能引发数据安全事件。

2.2 提示词注入与账号劫持：外部攻击的技术渗透

生成式 AI 的安全性不仅取决于其自身的设计，还取决于其与外部环境的交互方式。“提示词注入”（Prompt Injection）是一种针对大语言模型的典型攻击手段：攻击者在用户输入或外部数据中植入恶意指令，当 AI 读取这些数据时，便会将恶意指令误认为是来自用户的最高优先级指令，从而执行攻击者预设的操作。

在辅导员工作场景中，提示词注入攻击的可能路径包括：攻击者在学生提交的作业、留言或问卷中嵌入恶意指令，当辅导员使用 AI 辅助批阅或汇总时，AI 读取这些内容并被诱导执行越权操作。

3 幻觉风险：生成式 AI 的错误输出与执行后果

“幻觉”是指大语言模型生成与事实不符、与用户意图相悖或缺乏依据的内容。在辅导员工作场景中，AI 的幻觉可能从“语义偏差”转化为“行动误导”，对学生发展和教育管理造成实际损害。

作者简介：宁康（1998.03—），男，汉，助教，硕士研究生，广东技术师范大学教育科学学院，思想政治教育。

通讯作者：谭伟坚（1993.11—），男，汉，助教，硕士研究生，广东技术师范大学物理与光电工程学院，思想政治教育。

3.1 从语义偏差到行动误导：幻觉的风险传导机制

在传统的大语言模型应用中，幻觉通常表现为输出内容与事实不符，例如错误的时间、地点或人物信息。用户在面对这类明显错误时，往往能够凭借常识进行甄别。然而，当 AI 被嵌入辅导员工作流程后，幻觉的风险传导机制发生了质变：AI 的输出不再仅仅是“建议”或“参考”，而是可能直接触发后续行动。

3.2 批量处理的系统性风险

生成式 AI 在辅导员工作中的一大应用优势是“批量处理”：一次性生成数十份学生评语、一次性汇总上百条考勤记录、一次性审核全部奖助申请。然而，批量处理的效率优势背后，隐藏着系统性幻觉风险。

当 AI 在处理单个任务时出现幻觉，其影响是局部的；但当 AI 在处理批量任务时因算法逻辑缺陷或数据偏差而产生系统性错误，其影响将被成倍放大。更值得警惕的是，批量处理场景下的幻觉往往具有隐蔽性。辅导员面对 AI 生成的大量内容，很难逐条细致核验，容易产生“算法输出的都是对的”的心理暗示，从而将系统性错误“盖章通过”。

3.3 辅导员对 AI 输出的过度信任现象

幻觉风险的发生，既源于 AI 模型的技术缺陷，也与人类使用者的认知偏差密切相关。“自动化偏见”（Automation Bias）是指人类决策者在面对算法输出时，倾向于过度信任而减少自身的审慎判断。这一心理现象在辅导员工作中同样存在。

辅导员群体普遍面临“事多人少”的工作压力，时间资源的稀缺性使其天然倾向于采纳“快捷方案”。当 AI 能够快速生成一份结构完整、语言流畅的学生鉴定意见时，辅导员逐字核验的动力便会显著降低。更关键的是，AI 输出的内容往往在语法和逻辑上具有较高的表面合理性，这进一步增加了甄别难度。

4 让渡风险：人类主体性的削弱与责任转移

如果说权限风险关乎“技术安全”、幻觉风险关乎“输出质量”，那么让渡风险则直指人类在教育活动中主体地位的动摇的问题。“让渡”意指人类在不自觉中将原本属于自己的判断权、决策权和情感联结权转移给 AI 系统。在辅导员工作场景中，这种让渡正以渐进、无感的方式发生，其长期后果可能是专业能力的退化、师生关系的异化以及责任归属的模糊。

4.1 工具依赖与辅导员专业能力的退化

生成式 AI 的介入，在一定程度上“短路”了辅导员工作能力养成过程。当辅导员习惯于让 AI 生成谈心谈话记录、撰写工作总结、分析学生心理状态时，其自身的观察力、分析力和表达力便可能因“用进废退”而逐渐钝化。李欢等学者在对自主式智能体的研究中引述了相关发现：为了短期绩效而频繁使用 AI 完成任务，将会以长期能力的发展退化作为代价。这一判断同样适用于辅导员群体。

4.2 决策权力的实质性让渡

在奖助学金评定、评优评先、违纪处分等事务中，AI 能够快速整合学生数据、生成排名或提出建议。面对 AI 输出的“客观”“科学”的结论，辅导员往往倾向于“盖章式确认”，而放弃对个案特殊性的深入考量。这种决策权力的让渡，其危险不在于 AI 的建议一定错误，而在于人类丧失了独立判断的意愿和能力。当辅导员习惯于“AI 筛选+人工确认”的工作模式后，那些不符合算法预期但具有特殊价值的学生个案，便可能被系统性地忽略。

4.3 情感联结的弱化与师生关系的异化

辅导员工作区别于一般行政管理的核心特征，在于其情感维度。辅导员不仅是规章制度的执行者，更是学生成长道路上的陪伴者、倾听者和引导者。这种情感联结的建立，依赖于面对面的交流、即时的情感共鸣和长期的信任积累。生成式 AI 的介入，可能在两个层面弱化这种情感联结。其一，虚拟辅导员的上线减少了师生面对面接触的机会。其二，AI 辅助的“标准化”沟通可能消解师生互动中的人文温度。

4.4 责任归属的模糊化

让渡风险的最后一个维度，是责任归属的模糊化。从法律角度看，生成式 AI 不具备独立法律人格，其行为后果应由控制者或使用者承担。然而，在实际操作中，责任追溯面临多重障碍。更为棘手的是，责任的模糊可能催生“责任稀释”：当人人都认为“AI 只是辅助工具”时，人人也都可能放松对 AI 输出的审慎审查，从而使风险在无人负责的状态下累积放大。

5 规制路径：构建“权限-幻觉-让渡”协同治理体系

面对生成式 AI 在辅导员工作中引发的复合型风险，单一维度的应对策略难以奏效。本文提出构建“权限规制—幻觉防控—让渡纠偏”三位一体的协同治理体系，从技术、制度、个体三个层面综合施策。

5.1 权限规制：划定 AI 在辅导员工作中的操作边界

权限规制的核心目标是“让 AI 做该做的事，不做不该做的事”，通过技术隔离和制度约束防止系统越权。

在技术层面，高校应强制推行“最小权限原则”。AI 工具的访问权限应限定在完成特定任务所必需的最小范围，禁止授予其学工系统、学生数据库等核心信息系统的直接修改权限。在制度层面，高校应建立生成式 AI 工具的准入审查机制。辅

导员使用的 AI 服务须经学校信息化部门和法务部门的联合评估，重点审查数据存储位置、隐私政策合规性、安全防护能力等关键指标。在操作层面，应制定《辅导员使用生成式 AI 行为规范》，明确 AI 可参与的事务类型（如信息检索、文本草拟）和禁止参与的事务类型（如违纪处分裁决、奖学金评定最终决策），为辅导员提供清晰的行为指引。

5.2 幻觉防控：建立人机协同的验证机制

幻觉防控的核心目标是“让 AI 的错误被看见、被纠正”，通过技术设计和流程规范降低幻觉风险的实际危害。

在技术层面，应要求 AI 工具对输出内容进行“置信度标注”。对于评估类、诊断类、建议类输出，AI 应同时给出置信度分数和参考信息来源，便于辅导员判断其可靠性。在流程层面，应实施“AI 建议+人工复核”的双轨机制。对于可能对学生权益产生实质性影响的 AI 输出（如心理评估结论、奖助评定建议、违纪处分建议），必须经过辅导员的人工审验后方可采纳。在能力建设层面，应开发面向辅导员的“AI 幻觉识别”培训课程。培训内容应包括：大语言模型的基本原理、幻觉的常见类型与表现形式、提示词优化技巧（如何通过优化提问降低幻觉概率）、AI 输出的验证方法等。

5.3 让渡纠偏：捍卫辅导员的主体地位

让渡纠偏的核心目标是“让人做人的事，让 AI 做 AI 的事”，通过制度设计和价值引导防止人类主体性的滑落。

在责任分配层面，应明确“辅导员是最终责任主体”的基本原则。这一原则应写入学校相关规章制度，并在日常工作中反复强调，防止“AI 背锅”心理的蔓延。在评价机制层面，应调整辅导员的绩效考核导向。在考核指标中，应体现对“独立思考”“个案分析”“人文关怀”等能力的重视，而非单纯追求 AI 使用频率和产出效率。在价值引导层面，应持续强调“教育是人的事业”这一核心理念。在辅导员培训、会议、评优等场合，应反复传递“AI 是工具，育人是目的”的价值取向，鼓励辅导员在技术便利与人文关怀之间保持必要的张力。

6 结语

生成式人工智能正在深刻重塑高校辅导员工作的形态。它在释放效率红利的同时，也带来了权限越界、执行幻觉和权力让渡三重风险。本文借鉴自主式智能体研究中提炼的分析框架，将这些风险置于辅导员工作场景中加以考察，揭示了技术赋能背后的安全与伦理隐忧。

权限风险提醒我们：AI 的“便利”不能以“失控”为代价。必须为 AI 划定清晰的操作边界，防止技术能力溢出授权范围。幻觉风险提醒我们：AI 的“智能”不等于“正确”。必须建立有效的人机协同验证机制，让技术服务于人而非替代人的判断。让渡风险提醒我们：AI 的“效率”不能以“人的退化”为代价。必须守住辅导员作为教育主体的地位，防止在追求效率的过程中丢失教育的灵魂。在拥抱技术红利的同时，保持清醒的风险意识；在提升工作效率的同时，坚守教育的人文底线；在借助 AI 赋能的同时，捍卫人的主体地位。唯有如此，方能在人机协同的新时代中，推动高校辅导员工作行稳致远。

参考文献：

[1] 李兴成,唐燕.人工智能赋能高校辅导员工作的价值、风险及实践路径[J].现代远程教育研究,2025(15):251-253.

[2] 李欢,张赠,莫欣岳.自主式智能体 Open Claw 的安全与伦理风险及中国治理路径[J].海南大学学报(社会科学版),2026.

[3] 李桦,刘丽丽.人工智能时代高校辅导员的角色困境与突围路径探究[J].新闻研究导刊,2025,16(5):119-123.

[4] 赵磊.生成式人工智能赋能高校辅导员工作探索[J].北京教育(德育),2025(5):92-96.

[5] 刘骏.“数字思政”的伦理审视:价值、矛盾和治理[J].思想理论教育,2023(9):84-88.

[6] 邹开亮,刘祖兵.Chat GPT 的伦理风险与中国因应制度安排[J].海南大学学报(人文社会科学版),2023,41(4):74-84.