

生成式人工智能介入思想政治教育的伦理风险与治理路径

李婉婷

海南师范大学马克思主义学院，海南 海口 571127

摘要：近年来，随着生成式人工智能技术的不断迭代升级，使其深度介入思想政治教育成为一种必然趋势。该技术强大的内容生成与即时交互能力正在重塑教育生态、助力精准教学，提升了思政教育的针对性与吸引力。然而，技术的工具理性与思政教育所秉持的价值理性之间存有内在张力，导致其介入过程中潜藏着多重伦理风险。本研究聚焦于生成式人工智能在思政教育领域的应用场景，系统分析了主流价值偏离风险、师生主体性消解风险等困境。研究认为，上述风险的根源在于算法逻辑与育人逻辑的冲突。为此，应从价值对齐、人机协同、制度规范与责任伦理四个维度构建治理路径，确保技术赋能思政教育的过程不偏离立德树人的根本宗旨，推动智能思政的良性发展。

关键词：生成式人工智能；思想政治教育；伦理风险；治理路径

DOI: 10.64649/yh.jydk.issn3080-2660.202607008

0 引言

党的二十大报告从国家战略高度明确指出：“要加快推进教育数字化转型进程，着力构建覆盖全民、贯穿终身的现代化学习型社会体系，最终实现学习型大国的宏伟目标。”^①生成式人工智能持续迭代，正在重塑知识生产与传播格局，以其多模态生成与拟人交互能力深入教育领域。在思政教育中，它可辅助备课、提供全天候理论答疑与个性化引导，并通过沉浸体验增添活力。然而，技术赋能效应并非天然正向。当算法逻辑介入以价值引领为核心的思想政治教育过程时，技术理性与价值理性之间的深层张力便逐渐显现。生成式AI训练数据中隐含的偏见可能导致主流价值偏离，人机交互的过度依赖可能消解师生之间的情感纽带与主体性地位，而数据采集失范则威胁隐私安全。这些伦理风险若得不到有效规制，不仅会消解人工智能的技术红利，更可能背离立德树人的根本宗旨。基于此，本研究聚焦于生成式人工智能介入思想政治教育过程中的伦理困境，深入剖析风险表征与成因，从价值对齐、制度规范与责任伦理等维度构建治理路径，以期为智能思政良性发展提供镜鉴。

1 生成式人工智能介入思想政治教育的价值意蕴

1.1 拓展教育场域，构建育人空间

传统思想政治教育多受限于物理空间与固定时间节点，主要依靠课堂、会议等线性场景展开，存在覆盖面窄、即时性不足等问题。生成式人工智能依托大数据技术，突破了时空界限，推动思政教育向全时全域延展。在时间维度上，它促成了教育时间的碎片化整合。学生

可通过智能终端随时与AI交互，获取个性化答疑，思政教育因而不固守于固定课时，而是渗透进日常生活的碎片时间。在空间维度上，生成式AI借助多模态生成能力，能够构建高沉浸感的虚拟仿真场景，如虚拟红色展馆、历史情境再现等，使抽象理论转化为具身体验，增强了教育的在场感与临场感，实现了育人空间从“物理现实”向“虚拟现实”的跨越。

1.2 优化内容供给，促进资源提质增效

思政教育内容的吸引力和感染力直接关乎育人实效。长期以来，思政资源面临生产周期偏长、形式单一、话语转化滞后等困境。生成式人工智能的创造性内容生产能力，为优化供给提供了新的引擎。其一，它能够高效整合海量信息，快速生成图文、音视频等多元形态的思政素材，显著扩充教育资源库。其二，生成式AI具备较强的话语转化能力，可将艰深的政治理论、学术话语精准转化为学生易于接受的网络语言或生动案例，降低认知门槛，提升内容的亲和力。其三，基于实时数据分析，AI能及时捕捉社会热点与舆论焦点，迅速生成具有针对性的教育方案，增强思政教育回应现实、引导学生思想的时效性。

1.3 创新交互范式，推动单向灌输向精准对话转型

传统的思政教育模式常呈现为教育者向受教育者的单向灌输，缺乏有效的互动反馈，难以满足学生个性化、差异化的认知需求。生成式人工智能的介入，重塑了教育主客体间的交互关系，驱动教育范式向精准化、对话式转型。一方面，它赋予受教育者平等的对话权。在人机对话中，学生不再是单纯的倾听者，而是可以主动提问、质疑、探讨的主体，AI能够根据学生的提问逻辑与认知水平给予针对性反馈，

初步实现因材施教的个性化辅导。另一方面,生成式AI有助于实现精准思政。借助对学生与AI交互数据的深度挖掘,教育者可精准描绘学生的思想动态、价值困惑与兴趣偏好,据此制定差异化教学策略,推动供给方式从“大水漫灌”转向“精准滴灌”,切实提升思政教育的针对性与实效性。

2 生成式人工智能介入思想政治教育的伦理风险

2.1 信息失真与真理权威的消解

思想政治教育是以真理服人、以真情感人的过程,内容的真实性与准确性是其生命线。然而,生成式人工智能的技术特性对这条生命线构成了严峻挑战。生成式AI的核心机制在于概率性预测,而非事实性求证,这与思政教育对真理的追求之间存在着根本性冲突。其一,“机器幻觉”引发信息失真。生成式人工智能依据概率统计生成内容,而非基于事实逻辑推演,在思政教育场景中极易生成背离事实的内容。这类隐性错误不易被及时察觉,很容易误导受教育者,造成认知混乱。其二,信息碎片化带来真理遮蔽。生成式人工智能倾向于产出短平快、碎片化的内容,往往截断了理论建构的完整逻辑链条。受教育者接收这些片段化信息时,容易停留在知其然却不知其所以然的层面,导致对马克思主义理论的片面理解甚至误读,进而消解思政教育的真理权威与理论说服力。

2.2 算法偏见与主流意识形态的冲击

人工智能系统在技术架构中往往内嵌着设计者的价值判断与训练数据的固有偏向。这种带有偏见的算法机制介入思想政治教育后,极易对主流价值形成“隐形侵蚀”。思想政治教育以传播主流意识形态为使命,而生成式人工智能作为海量数据训练的产物,这客观上潜藏着意识形态安全风险。其一,算法偏见可能引发价值误导。训练数据中潜藏的西方价值观念与社会偏见会被模型习得并放大,AI输出的内容可能在不知不觉中携带片面的历史观与文化观,以话语渗透的方式侵蚀学生价值认同,这种影响更具隐蔽性与危害性。其二,“信息茧房”效应可能加剧价值极化。算法推荐机制迎合用户偏好,长期如此会将受教育者禁锢在既有认知的“舒适区”内,阻碍多元视角获取,增加极端思想滋生风险,使思政教育培养全面发展之人的目标落空。

2.3 技术依赖与育人主体的异化

思想政治教育的本质在于人与人之间的灵魂对话,而生成式人工智能介入可能引发“人—机”关系异化,冲击了教育者与受教育者的主体地位。其一,教育者主体性面临弱化风险。部分教育者产生技术依赖,将教学设计等主导权让渡出去,退居为操作者,消解了能动性

创造性,使情感交流与人格示范缺位。AI快速回应学生困惑的能力,削弱了教育者的知识权威,学生更愿相信算法的“客观中立”,质疑教育者带有立场的言说。其二,受教育者陷入思维惰性与主体性迷失。习惯于向AI索求答案,思维惰性滋长,独立思考与批判能力被消磨,瓦解了价值学习所需的沉思、质疑与内化,使学生陷入高效却肤浅的虚假学习,钝化价值判断力。在过度的人机交互中,学生可能沉溺于虚拟的算法陪伴,产生情感冷漠与人际疏离,导致作为人的主体性在技术逻辑中逐渐迷失。

3 生成式人工智能介入思想政治教育的治理路径

3.1 优化算法模型,构建可信可控的技术防线

技术产生的问题往往需要通过技术手段来解决。要从源头上规避伦理风险,必须加强底层技术治理,确保生成式人工智能的可信度与安全性。一是推动主流价值观的算法嵌入。突破单纯的效率导向,将社会主义核心价值观、法律法规及伦理规范转化为算法约束条件,嵌入模型训练与推理的全过程。二是建立人机协同的内容审核机制。针对生成内容的不可控性,建立严格的事前预警、事中监测、事后追溯技术防线。对生成式AI的算法幻觉和事实性错误,必须构建人机协同的内容治理模式,守住教育内容的真实性底线。开发针对思政敏感内容的自动化检测工具,对AI生成的文本、视频进行实时筛查与过滤,确保教育内容的政治正确与事实准确,杜绝“机器幻觉”导致的虚假信息流入教育场域。

3.2 完善伦理规制,筑牢安全规范的制度屏障

制度是治理的保障,是技术应用的紧箍咒与安全阀。通过建立健全伦理规范与法律法规,为生成式人工智能在思政教育中的应用划定红线。一是制定生成式人工智能思政应用伦理准则。明确技术应用的行为边界与负面清单,规定数据采集、算法设计、内容生成等环节的伦理要求。例如,明确规定AI不得替代教育者进行价值判断,不得滥用学生隐私数据,确保技术应用始终服务于育人目标。二是强化数据安全与知识产权保护。建立严格的数据分级分类管理制度,加强对师生交互数据的安全防护,防止个人隐私泄露与数据滥用。同时,完善AIGC的版权认定与责任归属机制,厘清技术开发者、教育者与使用者之间的责任链条,为治理提供坚实的法律依据。三要明确责任归属与追溯机制。针对AI生成内容可能引发的错误、误导或侵权问题,建立清晰的责任认定框架。明确开发者、应用方、使用者在不同情境下的责任边界,防止出现责任真空或相互推诿,形

成有效的法律威慑。

3.3 提升数字素养，重塑人机共生的主体关系

技术治理的最终落点在人的素养。只有提升师生的数字素养和批判性思维，才能在根本上驾驭技术而非被技术驾驭。其一，教育者应从单纯的知识传授者转型为技术驾驭者与价值引领者。不仅要掌握生成式人工智能的操作技能，更要具备批判性思维，能够鉴别AI生成内容的优劣真伪，保持教育的敏感性与创造力，在人机协同中始终掌握话语主导权。其二，培育受教育者的算法批判意识。在思政教育中嵌入信息素养教育模块，引导学生认清技术的本质及其边界，避免陷入算法茧房与思维惰性。此外，重建人机互动中的情感联结同样不容忽视。应倡导技术辅助、情感主导的教育理念，鼓励师生在课堂内外开展面对面的深度交流，学校也需创设更多线下思想碰撞与情感交融的场景，防止人机交互替代真实的人际交往，守护教育应有的情感温度。

3.4 以社会主义核心价值观驾驭算法逻辑

治理的关键在于让价值取得优先地位。技术逻辑需要服从于教育逻辑，算法推荐也应当服务于价值引领，而非反客为主。一方面，应推动构建价值观对齐的技术框架。通过产学研协同，将社会主义核心价值观系统化、结构化地融入生成式AI的底层算法模型，在预训练、指令微调及强化学习等环节，建立契合中国国

情和教育目标的价值偏好体系，使AI在处理涉及政治立场、道德判断、历史认知等内容时，能够形成清晰、正确的价值选择。另一方面，需建立算法干预与纠偏机制。对AI生成内容展开常态化监测，及时发现并矫正偏离主流价值的输出倾向；对触及意识形态安全的核心议题，可探索设立价值观护栏，对敏感话题的生成内容实施前置审核与定向引导，阻断错误信息的扩散。此外，最根本的还在于坚守“以人为本”的技术定位，明确AI在教育中仅扮演辅助性、工具性角色，防止技术理性僭越价值理性。在制度设计与应用场景开发中，始终将人的发展置于最高位置，确保技术应用不致异化为对师生的规训与控制。

4 总结

生成式人工智能的崛起，标志着思想政治教育正经历着一场深刻的技术范式转型。然而，技术从来不是价值中立的工具，当算法介入精神世界与价值建构，主体性、真实性、价值性与关系性危机便随之浮现，要求我们保持足够的审慎与警醒。面对这一时代命题，我们应该坚持伦理先行，在价值层面以主流意识形态驾驭算法逻辑，在制度层面构建全过程风险防控体系，在事实层面建立人机协同的内容审核机制，在素养层面培育师生的人机协同能力，最终，回归人与人的精神交往这一教育本质。

注释：

①习近平. 高举中国特色社会主义伟大旗帜为全面建设社会主义现代化国家而团结奋斗：在中国共产党第二十次全国代表大会上的报告[N]. 人民日报，2022-10-26(1).

参考文献：

- [1] 陆金凤, 朱国玮, 罗映宇, 等. 从决策式到生成式：人工智能管理研究的理论框架与未来议题[J]. 管理世界, 2026, 42(02): 208-241.
- [2] 杨威, 谭文伦. 思想政治教育要素形态的演进[J]. 学校党建与思想教育, 2025, (09): 18-24.
- [3] 王丽, 陈志锐, 李艳. 生成式人工智能赋能论证教学反馈研究[J]. 中国电化教育, 2026, (02): 123-130.
- [4] 孟飞, 詹蕊齐, 耿盛洁. 生成式人工智能赋能思想政治教育新范式：创新机遇、核心表征与风险挑战[J]. 中国电化教育, 2026, (02): 30-36.
- [5] 梁俊斌, 谢贇, 唐振华. 生成式人工智能赋能教育数字化转型的机理、挑战与对策[J]. 现代教育管理, 2025, (11): 15-23.
- [6] 姚荣, 王倩. 生成式人工智能教育应用的风险及其法律规制[J]. 清华大学教育研究, 2025, 46(06): 131-142.
- [7] 李锋雷. 生成式人工智能与基础教育的价值对齐：分类困境与分层治理[J]. 电化教育研究, 2026, 47(02): 30-37.

作者简介：李婉婷（2003.04—），女，汉族，广东揭西，硕士研究生，海南师范大学马克思主义理论专业研究生在读，研究方向：思想政治教育。