Advances in Psychology 心理学进展, 2026, 16(8), 130-141
Published Online August 2026 in Hans. https://www.hanspub.org/journal/ap
https://doi.org/10.12677/ap.2026.168382

Hans汉斯

大学生AI学习助手使用中的伦理认知偏差及情境化视频干预研究

冀思语，于战宇

江苏师范大学教育学部，江苏 徐州

收稿日期：2026年6月27日；录用日期：2026年7月31日；发布日期：2026年8月12日

摘　要

人工智能(AI)工具在大学生中被广泛应用，但学生在使用AI学习助手时普遍存在伦理认知偏差问题。本研究采用文献综述与理论建构的方法，整合认知偏差理论、CAC理论(认知－情感－意动模型)与情境学习理论，在系统分析国内外相关研究进展的基础上，从认知自主性侵蚀、抄袭边界模糊、数据隐私意识薄弱三个维度构建大学生AI伦理认知偏差的分析框架与干预模型，为未来通过实验设计验证情境化视频干预的有效性提供理论依据与实践方案。

关键词

大学生，AI伦理，认知偏差，情境化干预

Ethical Cognitive Biases in College Students' Use of AI Learning Assistants and Situational Video Intervention Research

Siyu Ji, Zhanyu Yu

School of Education Science, Jiangsu Normal University, Xuzhou Jiangsu

Received: June 27, 2026; accepted: July 31, 2026; published: August 12, 2026

Abstract

Artificial Intelligence (AI)tools are widely used among college students, yet ethical cognitive biases are prevalent when students utilize AI learning assistants. Employing methods of literature review and theoretical construction, this study integrates Cognitive bias theory, the CAC (Cognition-Affect-Conation) model and Situated learning theory. Based on a systematic analysis of relevant research

文章引用: 冀思语, 于战宇(2026). 大学生 AI 学习助手使用中的伦理认知偏差及情境化视频干预研究. *心理学进展, 16(8),* 130-141. DOI: 10.12677/ap.2026.168382

progress both domestically and internationally, it constructs an analytical framework and intervention model for college students' ethical cognitive biases in AI use from three dimensions: erosion of cognitive autonomy, blurring of plagiarism boundaries, and weak awareness of data privacy. This work provides a theoretical basis and practical plan for verifying the effectiveness of contextualized video interventions through experimental design in the future.

Keywords

College Students, AI Ethics, Cognitive Bias, Contextualized Intervention

1. 引言

从 2022 年 ChatGPT 面市以来，以 DeepSeek、文心一言、Kimi 为代表的 AI 学习助手迅速渗透到大学生的日常学习生活中。从文献检索、文字翻译到论文写作，AI 工具在为学生提供高效、便捷服务的同时也引发了前所未有的伦理争论。学生能不能借助 AI 直接生成论文作业？AI 生成的内容是否需要标注来源？AI 使用中数据隐私风险是否被充分认知？这些问题矛头直指技术迭代速度与伦理规范建设之间的巨大落差(García-López & Trujillo-Liñán, 2025)。

这一情况在教育实践中表现得更为突出。一方面，AI 工具在大学生的学习使用中已势不可挡。Mazaherian 与 Nourbakhsh (2025)的调查显示，92%的大学生使用 AI 工具的主要动机是节省时间和提升作业质量。另一方面，针对 AI 工具使用的伦理指导却严重滞后。Ma 等(2025)在对 98 节 AI 素养课程的分析中发现，仅 5.1%的课程涉及 AI 伦理内容，而课程中基础能力(认识与理解 AI、使用与应用 AI)相关内容高达 93.88%。这种“重技术应用、轻伦理认知”的教育局面，直接导致大学生在 AI 伦理认知上普遍缺失。更为严重的是，学生可能尚未意识到长期依赖 AI 对其独立思考能力的潜在侵蚀(Zhai et al., 2024)。

因此，本文旨在：第一，系统梳理 AI 学习助手的研究现状；第二，构建大学生 AI 学习助手使用中的伦理认知偏差类型框架；第三，提出基于情境学习理论的情境化视频干预方案，为高校制定 AI 使用规范、开发 AI 伦理教育课程提供参考。

2. AI 学习助手的研究现状

2.1. AI 学习助手的发展与教育应用

AI 学习助手是指基于生成式人工智能技术，辅助用户完成学习任务的对话系统，它的核心特征包括自然语言交互能力、多模态内容生成能力和个性化响应能力(Long & Magerko, 2020)。从其发展进程看，第一代 AI 教育工具以规则为基础，功能有限；第二代引入机器学习，但交互性不足；第三代以 ChatGPT、DeepSeek 为代表的大语言模型(LLMs)，实现了接近人类的对话能力(Zawacki-Richter et al., 2019)。

在教育应用领域，AI 学习助手已广泛渗透到各类学习场景：文献检索与总结、论文写作辅助、代码调试、外语翻译、备考复习等(Popenici & Kerr, 2017)。学生使用 AI 的主要动机是“节省时间”和“提升作业质量”，但只有少数学生获得过正式的使用指导(Mazaherian & Nourbakhsh, 2025)。这种“影子教学”现象引发了对教育公平性和学术诚信的担忧(Selwyn, 2022)。

2.2. AI 学习助手伦理问题的研究

AI 伦理问题作为 AI 素养的核心组成部分，近年来受到学界广泛关注。Kasneci 等(2023)系统梳理了大语言模型在教育应用中面临的机遇与挑战，提出隐私保护、算法偏见、学术诚信和认知自主性等伦理问题。Lodge 等(2023)提出了 GenAI 教育应用的分析框架，从人机交互模式和 AI 教育功能两个维度，区分了 AI 作为信息提供者、学习助手、协作伙伴等不同角色对应的伦理边界。Zhai 等(2024)的系统综述中指出：学生对 AI 的过度依赖会削弱其决策能力、批判性思维和分析推理能力。García-López & Trujillo-Liñán (2025)归纳了 GenAI 教育应用面临的三大伦理挑战：数据隐私与保护、算法偏见与歧视、认知自主性丧失。该研究强调，当学生过度依赖 AI 完成学习任务时，其批判性思维和独立解决问题的能力可能受到侵蚀。

国内研究方面，近年来对 AI 伦理教育伦理的关注研究热度显著上升。国内高校开展的调查普遍发现，大学生 AI 使用率极高，但伦理认知水平参差不齐。常见现象包括：多数学生能够意识到 AI 生成内容可能存在错误或偏见，但倾向于直接采纳 AI 建议，较少多渠道验证；学生对隐私风险的敏感性较低，对对话数据被用于模型训练的事实不以为意；对学术诚信的界定存在模糊地带，部分学生认为“AI 辅助改写不算抄袭”。常青(2024)分析了以 ChatGPT 为代表的生成式 AI 技术在大学校园的迅速普及，指出其在推进教育智慧化的同时，引发了诸多大学生的科技伦理失范现象，提出需要深入分析其成因以引导正确的科技伦理观。童慧等(2026)通过全国 2266 名高校学生的问卷调查，发现大学生 AI 工具使用存在显著差异，学生对伦理课程及智能教育平台的需求迫切，从侧面反映出当前大学生在 AI 使用中的伦理素养不足，不能有效规避学术不端风险。汪金英等(2026)从道德教育视角指出，DeepSeek 类 AI 工具在赋能教育的同时，也带来技术依赖削弱学生主体地位、算法偏见影响教育导向等问题。他们主张构建“人机协同”的道德教育模式，将 AI 定位为“助教”而非“替代”。

3. AI 学习助手使用中的伦理认知偏差及类型

3.1. 认知偏差

认知偏差(Cognitive Bias)指个体在信息加工过程中系统性地偏离理性标准的心理机制(Tversky & Kahneman, 1974)。经典认知偏差包括可得性启发、代表性启发、锚定效应等。这些偏差往往发生在无意识层面，导致个体在判断和决策时产生系统性错误(Tversky & Kahneman, 1974)。

近年来，研究者将认知偏差理论应用于 AI 伦理领域。Dietvorst 等(2015)的“算法厌恶”研究发现，即使某种算法的整体表现优于人类，亲眼目睹算法出错仍然会明显降低用户对其的信任。Tariq 等(2025)进一步提出“替代选项厌恶”概念，认为人们对算法错误的过分苛刻判断部分源于其“非惯例”地位。Roy-Stang 与 Davies(2025)的系统分析识别出影响 AI 安全对齐的多种人类偏差，包括自动化偏差、权威偏差、正常化偏差等。

自动化偏差指个体过度依赖自动化系统推荐、倾向于以自动化输出替代自身审慎信息加工的倾向(Romeo & Conti, 2025)。权威偏差指个体倾向于顺从权威人物的判断，将权威意见作为决策依据而放弃自身批判性思考的倾向(Milgram, 1974)。与传统人类权威不同，AI 的“权威性”来源于：统计权威(基于大规模数据训练的算法被认为具有统计学优势)、客观性错觉(用户倾向于认为 AI 不存在人类的主观偏见)以及语言流畅性偏差(AI 生成的语言流畅，增加了其可信度) (Tariq et al., 2025)。正常化偏差指个体将异常、不当或风险行为逐渐视为“正常”“可以接受”的倾向，尤其是在观察到他人普遍采取该行为时更为明显(Roy-Stang & Davies, 2025)。基于大语言模型构建的情境模拟研究显示：在参与实验的智能体中，有 71%的智能体受正常化偏差、合理化偏差等认知偏差的影响，做出了不符合伦理规范的行为；另有 78%的智能体受自动化偏差与权威偏差的作用，对 AI 生成的输出内容产生了盲目依赖(Roy-Stang & Davies, 2025)。

3.2. AI 学习助手使用中的伦理认知偏差

将认知偏差理论代入 AI 学习助手使用情境，对理解大学生 AI 伦理认知偏差具有重要启示。在 AI 学习助手使用情境中，自动化偏差的表现：当学生使用 AI 完成作业时，将 AI 生成的答案直接复制粘贴而不加验证；即使发现 AI 输出存在疑点，仍倾向于相信“AI 应该是对的”；对 AI 的输出缺乏批判性审视，甚至主动忽视与 AI 结论相矛盾的信息。权威性偏差的表现：在 AI 学习助手情境中学生认为“AI 比我聪明”，放弃对 AI 输出的质疑；即使自己有不同想法，也倾向于优先采纳 AI 的建议；将 AI 视为“正确答案的提供者”而非“工具”。正常化偏差的表现：学生观察到“大家都在用 AI”，逐渐将 AI 使用视为常态；对 AI 代写、AI 协助完成作业的道德敏感性降低；即使在学术规范明确禁止的情境下，仍通过“大家都在做”来为自己的行为辩护。

此外，AI 使用中的伦理认知偏差还表现为学生的风险感知与使用行为之间的“知行分离”。Perdana 等(2025)的研究发现，学生大多能够识别 AI 工具存在的偏见、隐私和学术诚信风险，但由于优先考虑其带来的便捷效应而依然选择使用，形成“风险正常化”的心理机制。Foltýnek 与 Newton (2026)对 YouTube 平台 173 个 AI 使用指导视频进行主题分析，发现大量视频向学生传授“绕过 AI 检测工具”的方法，包括“人性化改写”“混合人工编辑”等技术。An 等(2025)指出，AI 伦理教育滞后于技术发展的现实问题亟待解决。

对于伦理认知偏差的形成机制，Zhai 等(2024)指出，学生倾向于采用认知捷径处理 AI 生成内容，他们优先选择“快速最优解”而非“受实践约束的慢速方案”，这解释了为什么学生明知 AI 可能出错仍选择无条件接受的原因。Gerlich (2025)引入“认知卸载”概念，认为学生将认知任务转移给 AI 后，自身的批判性监控能力随之弱化，形成“越依赖、越依赖”的恶性循环。Perdana 等(2025)关注到“同伴效应”在偏差形成中的作用。当周围同学普遍依赖 AI 且获得短期效率收益时，个体更倾向于效仿，形成集体性的风险认知偏移。

研究表明，专业能力较强的用户对 AI 的信任和依赖程度更低；对信息技术掌握程度更高的用户更倾向于认为 AI 不具备精准判断能力；而一贯不太信任他人的用户反而更愿意相信 AI (Glikson & Woolley, 2020)。这一发现提示，针对大学生群体的 AI 伦理教育需要考虑个体差异，不能采用“一刀切”的干预策略。

3.3. AI 技术特性与人类认知捷径的交互

3.3.1. 交互的即时性与认知吝啬鬼倾向

AI 学习助手能在数秒内对用户的提问生成详细、完整的回应，这种即时性反馈改变了传统学习中“提问–思考–解答”的过程，创造了“即问即答”的新型交互模式。认知捷径中的吝啬鬼倾向(Cognitive Miserliness)则表现出人类倾向于选择最低认知消耗的信息加工方式(Fiske & Taylor, 2020)，系统 1 思维是快速、自动、不费力的直觉加工模式，而系统 2 思维则需要投入注意力和认知资源进行审慎推理(Kahneman, 2011)。AI 的即时性反馈与认知吝啬鬼倾向形成正反馈循环，当学生发现只需输入问题即可获得答案时，系统 1 思维被频繁激活，而系统 2 思维的启动则被系统性地抑制。每一次即时满足都在强化“问 AI”这一行为模式，逐渐削弱了学生延迟满足和深度思考的意愿。这一机制直接催生了自动化偏差——学生将 AI 输出视为无需验证的“答案”而非“参考”，并最终导致认知自主性侵蚀。

3.3.2. 权威性幻觉与可得性启发

AI 生成的内容总是以自信、流畅、完整的语言呈现，并往往直接给出结论而非概率性判断。当一个人以高度自信的语气给出确定性答案时，接收者倾向于赋予其更高的可信度，这种表达方式在人类交流

中通常与专业权威高度关联。认知捷径中的可得性启发(Availability Heuristic)显示个体倾向于依据信息从记忆中提取的难易程度来判断其真实性和重要性(Tversky & Kahneman, 1974)。流畅、自信的表达更容易被提取和记住，因而被判断为“更可信”。权威偏差(Milgram, 1974)则使个体倾向于顺从权威来源的判断。AI 输出的语言流畅性直接触发了可得性启发，流畅且自信的表述使信息更容易被加工和记忆，从而被认知系统标记为“可信”。与此同时，AI 的“技术权威”光环激发了权威偏差，使学生放弃批判性审视。两种偏差协同作用，使 AI 的“权威性幻觉”得以确立——学生不仅因“AI 说得流畅”而相信它，更因“AI 是技术权威”而不加质疑。这一机制是权威偏差和自动化偏差在 AI 情境中被急剧放大的核心原因。

3.3.3. 交互自然性与社会临场感

AI 学习助手能够模拟人类的对话方式，包括使用礼貌用语、表达共情、适应上下文语境等。这种自然交互模式创造了“社会临场感”，即用户在与 AI 交互时产生类似于与真人对话的社交体验。而认知捷径中的社会归因倾向(Social Attribution Tendency)则是人类倾向于将具有社会特征的实体拟人化，并应用社会认知规则来理解其行为(Nass & Moon, 2000)。在进化过程中，人类发展出快速判断他人意图和可信度的社会认知捷径。所以，当 AI 表现出类似人类的交互特征时，学生无意识地将社会归因倾向投射到 AI 上，将 AI 视为具有可信度的对话者。这种拟人化认知使学生将本应用于人类交互的社会规范应用于 AI，进一步强化了权威偏差。更重要的是，社会临场感降低了学生对 AI 的“技术警觉”——学生不再将 AI 视为需要批判性审视的工具，而是视为可信赖的对话伙伴。

上述三种机制形成一个相互强化的循环：即时性反馈不断强化即时满足的吝啬鬼倾向认知习惯，权威性幻觉利用可得性启发建立信任，交互自然性通过社会临场感降低技术警觉。三者协同作用，使学生在无意识层面完成从“使用工具”到“依赖权威”的认知转变。这一机制层面的分析揭示了为何单纯告知学生不要过度依赖 AI 往往无效，因为偏差是在无意识的认知加工中形成的，而非有意识的选择。

3.4. 大学生 AI 伦理认知偏差的具体类型

综合上述人－AI 交互中已识别的认知偏差，为大学生 AI 伦理认知偏差研究提供了重要的理论基础，结合对大学生 AI 学习助手使用情境的考察，本研究将大学生 AI 伦理认知偏差归纳为三个核心维度：认知自主性侵蚀、抄袭边界模糊、数据隐私意识薄弱。

3.4.1. 认知自主性侵蚀偏差

认知自主性侵蚀(Cognitive Autonomy Erosion)指长期依赖 AI 工具导致的独立思考能力退化、批判性思维弱化、问题解决能力下降的潜在风险。这一偏差的主要表现为：思维依赖，遇到问题时，学生的第一反应是“问 AI”而非“自己想”，这种即时满足的习惯会逐渐削弱独立思考的动机和能力，使学生缺乏批判性审视；判断让渡，学生将判断权让渡给 AI，不再验证 AI 输出的正确性，认为“AI 比我聪明，听它的没错”；问题简化，因 AI 能快速给出答案，学生放弃对复杂问题的深入思考，满足于表层答案；创意萎缩，过度依赖 AI 生成创意，学生的原创力和想象力受到抑制。相关研究表明，过度依赖 AI 与批判性思维能力呈负相关(Zhai et al., 2024)。学生长期使用 AI 而不加批判，可能导致“认知懒惰”，进而影响深度学习的发生(Popenici & Kerr, 2017)。

3.4.2. 抄袭边界模糊偏差

抄袭边界模糊指学生对 AI 生成内容归属权的认知不清。具体表现包括：无主认知，认为“AI 内容无主，无需标注”，将 AI 生成内容视为公共资源；复制粘贴合理化，认为“直接复制粘贴 AI 生成的文本到作业中，不属于抄袭”；标注回避，认为标注 AI 贡献是繁琐的、不必要的；规范模糊，对“合理使

用”与“学术不端”的边界缺乏清晰认识。这一偏差的核心在于知识产权意识的缺失。Wiese 等(2025)指出，学生普遍不理解 AI 生成内容的版权属性，部分原因是教育系统中缺乏相关规范指导。并且，即使学生认同“应该标注”的原则，在实际操作中也可能因便利性而放弃标注，体现了“认知–行为”断裂(Mazaherian & Nourbakhsh, 2025)。

3.4.3. 数据隐私意识薄弱偏差

数据隐私意识薄弱指学生对 AI 工具使用中的个人信息风险认知不足。具体表现包括：隐私政策忽视，从不阅读隐私政策，直接点击“同意”；过度信息分享，为获取免费功能，轻易填写个人敏感信息；风险低估，认为“大家都在用，应该没问题”，低估数据泄露的潜在危害；便利优先，明知可能存在风险，但因便利性而选择忽略。Iftikhar 等(2025)的研究发现，学生对 AI 系统的隐私风险存在系统性低估，且这种低估与“风险懈怠”情感密切相关。司莉等(2026)的研究进一步揭示，从众心理和侥幸心理是导致隐私意识薄弱的重要情感因素。

认知自主性侵蚀关注 AI 依赖对学生自身认知能力的影响，抄袭边界模糊关注 AI 使用对学术规范的影响，数据隐私意识薄弱关注 AI 使用对技术伦理的影响。三者共同构成大学生 AI 伦理认知偏差。

4. AI 学习助手使用伦理认知偏差的干预方案

4.1. 情境化视频干预的理论依据

本研究的情境化视频干预方案基于三个理论支柱：CAC 理论、情境学习理论和认知偏差理论。

4.1.1. CAC 理论(认知–情感–意动模型)

CAC 理论(Cognitive-Affective-Conative Model)是理解个体态度与行为关系的经典框架。在 CAC 理论框架下，个体的行为是认知、情感与意动三要素交互作用的产物，认知阶段形成对事物的基本判断，情感阶段赋予判断以情绪色彩，意动阶段则将态度转化为行动。这一理论为理解大学生 AI 使用伦理风险的形成机制提供了系统性的分析工具。

近些年来，CAC 理论在信息技术行为研究领域得到了广泛应用。Zhou 与 Zhang (2025)从 CAC 视角探讨生成式 AI 用户的间歇性中止行为，发现隐私关注和信息幻觉通过影响认知失调，经情感承诺的调节作用，进而导致用户的中断使用行为。Du (2026)将 CAC 框架应用于 OpenClaw 用户行为意向研究，发现正面感知强化用户态度进而增加行为意向，而负面感知则增加不信任感并降低使用意向。在这些研究中，CAC 框架解释了用户在与 AI 交互过程中的态度形成与行为决策机制。在 AI 伦理研究领域，司莉等(2026)将 CAC 理论应用于大学生 AI 使用伦理风险的形成机制研究，为理解大学生 AI 伦理认知偏差的形成机制提供了理论框架。

本研究在前人研究的基础上进一步拓展 CAC 理论的应用，将 CAC 理论从“解释框架”延伸至“干预设计”层面——不仅用 CAC 解释偏差如何形成(认知偏差→情感懈怠→意动偏差)，更用 CAC 指导干预方案的设计(针对三阶段分别设计认知冲突、情感唤起、价值澄清的干预策略)。

4.1.2. 情境学习理论

情境学习理论(Situated Learning Theory)由 Lave 与 Wenger (1991)提出，强调学习是嵌入在具体情境和社会实践中的过程。该理论的核心观点包括：(1) 知识是情境化的，脱离情境的知识难以迁移和应用；(2) 学习是社会实践的一部分，学习者通过参与实践共同体建构意义；(3) 合法的边缘性参与是学习的重要机制，学习者从边缘逐渐走向中心。

情境学习理论为 AI 伦理教育提供了方法论：道德原则不能通过抽象说教来传递，而应嵌入真实的伦理困境情境中，让学生在情境互动中建构对伦理原则的理解。Williams 等(2019)在儿童 AI 教育研究中发

现，情境化活动能够显著提升学生对 AI 概念的理解和批判性思考能力。An et al. (2024)和 Karakuş & Gedik (2025)强调通过真实或模拟的复杂情境帮助学习者发展对 AI 伦理问题的判断力与责任感，支持“真实情境模拟”作为伦理教育的有效路径。胡德庆(2025)对文生视频模型在思想政治教育中的应用研究也是基于这一理论逻辑。

本研究选择情境化视频作为干预手段，正是基于情境学习理论的核心原理。视频均取材于真实的校园生活场景，通过真实性的情境设置触发观看者的身份代入和情感共鸣。视频结尾通过开放式问题制造认知冲突，激活反思过程。后续的小组讨论则为观点碰撞和意义重构提供了社会性空间。这种“具身情境－认知冲突－社会协商”的设计逻辑，实现了情境学习理论在 AI 伦理干预中的具体转化。

4.1.3. 认知偏差理论

认知偏差理论(Cognitive Bias Theory)源自认知心理学，关注个体在信息加工过程中系统性地偏离理性标准的心理机制。Tversky 与 Kahneman(1974)的研究揭示了一系列认知偏差，如可得性启发(availability heuristic，依赖容易想到的例子做判断)、代表性启发(representativeness heuristic，以刻板印象替代概率判断)、锚定效应(anchoring effect，过度依赖最先获得的信息)等。这些偏差导致个体在判断和决策时产生系统性错误，且往往发生在无意识层面，难以通过简单的理性说服来纠正。

在 AI 伦理领域，认知偏差理论被用于理解用户对 AI 系统的不理性态度及其引发的伦理后果。Dietvorst 等(2015)的“算法厌恶”研究揭示了人们对算法错误的苛刻判断。Tariq 等(2025)认为人们对算法错误的过度反应部分源于对非惯例决策方式的排斥。此外，情境认知是认知偏差的重要调节因素。个体对情境的即时判断(如“作业太多时间不够”)会影响其伦理决策。在选择 AI 辅助学习时，如果个体感受到较大的时间压力或绩效压力，可能更倾向于将 AI 代写视为“无奈之举”而非“错误行为”。这种情境依赖的判断模式解释了为何认知偏差具有情境敏感性——同样的个体在不同情境下可能表现出不同程度的偏差。这些发现为理解大学生对 AI 工具的信任与质疑提供了理论视角。

认知偏差理论揭示了干预的核心原则：偏差往往是无意识的，学生并非“明知故犯”，而是“未能意识到”自己的判断存在问题。本研究将认知偏差理论作为理解大学生 AI 伦理认知偏差的心理机制基础。不仅解释学生怎么想，更指向为什么这样想。这一认识直接决定了干预策略的取向：不是说教，而是唤醒——帮助学生“看见”自己的偏差。情境化的认知冲突设计可以打破自动化思维，促进反思性认知。

4.1.4. 理论整合与框架建构

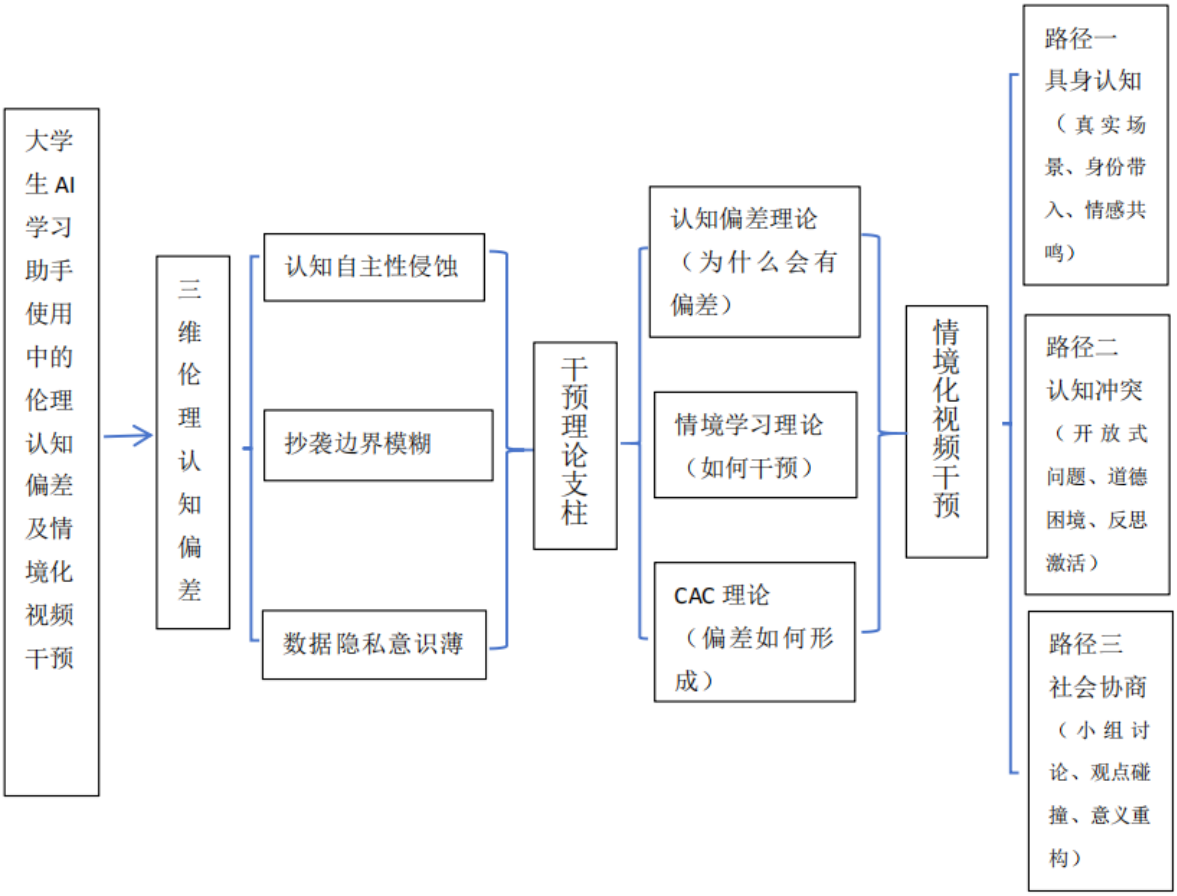
上述三个理论在本研究中有机整合。

CAC 理论解释偏差从形成到干预的完整心理路径。从认知阶段的偏差感知、情境认知、主观规范和风险偏好，到情感阶段的风险懈怠(从众心理、侥幸心理)，再到意动阶段的价值观驱动行为选择，构成了偏差形成的“认知偏差→情感懈怠→意动偏差”三阶段递进链条。

认知偏差理论阐明认知阶段偏差的具体心理类型，揭示学生在 AI 伦理问题上“为何会想错”，偏差往往发生在无意识的信息加工层面，而非有意的价值背离。这一定位决定了干预策略的有效方法是“意识唤醒”而非“道德说教”。

情境学习理论提供干预的方法论指导，找到如何通过情境设计打破认知偏差的方法。通过具身认知路径(真实场景触发身份代入)、认知冲突路径(开放式问题激活反思)、社会协商路径(小组讨论促进意义重构)三条机制，共同构成干预作用的完整路径。

三个理论分别回应三个核心问题：偏差如何形成？(CAC)、为什么会有偏差？(认知偏差)、如何干预？(情境学习)。这一框架为本研究提供了从问题诊断到干预实施的完整理论工具。



4.2. 情境化视频干预方案

本研究聚焦大学生在使用AI学习助手时面临的认知自主性侵蚀、抄袭边界、数据隐私等伦理认知偏差问题，通过调查问卷了解认知偏差现状，在此基础上设计和制作基于真实校园生活的情境化视频，组织参与者观看并进行小组讨论，通过“具身认知、认知冲突、社会协商”三条路径，经过“体验－反思－迁移”三阶段学习机制，促进大学生伦理认知的具身化修正，在此之后通过后测评估干预效果。

4.2.1. 大学生AI伦理认知现状调查

在查阅相关研究文献的基础上，编制《大学生AI辅助学习伦理态度问卷》，包含抄袭边界、数据隐私 、认知自主性侵蚀三个维度，采用李克特5点量表计分。问卷通过问卷星发放，了解大学生AI辅助学习中的伦理认知偏差现状，分析大学生在AI伦理认知方面的偏差特征。

4.2.2. 情境化干预视频开发与干预

基于上述理论依据，本研究设计了3个情境化视频，每个视频聚焦一个具体的AI伦理困境，采用“情境呈现－冲突点设置－开放式结尾”的三段式结构。

视频1：《先问AI，还是先自己思考？》

聚焦偏差：认知自主性侵蚀

情境描述：大学图书馆，两名学生面对同一道难题。学生A打开AI直接问，得到答案后直接使用；学生B先自己思考、查阅资料、尝试解题，最后用AI验证答案。冲突点设置：两人都完成了作业，但谁的收获更大？长期来看，谁会成长得更快？开放式结尾：“如果每次都先问AI，我还是原来的我吗？”

干预路径：通过具身认知(真实图书馆场景)触发身份代入；通过两种学习方式的对比制造认知冲突，打破“AI 更方便所以更好”的自动化思维；激活学生对长期认知发展的反思。

心理机制：(1) 认知冲突机制：通过对比“先问 AI”与“先自己思考”两种学习方式的长期收益差异，制造认知不协调，激活反思过程。认知失调理论指出，当个体同时持有两种相互矛盾的认知时，会产生心理不适，进而驱动态度改变(Festinger, 1957)。本视频通过呈现两种学习路径的对比，使倾向于依赖 AI 的学生意识到其行为与“提升自身能力”的学习目标之间的矛盾，触发认知重构。(2) 元认知激活：视频结尾的开放式问题“如果每次都先问 AI，我还是原来的我吗？”引导学生进行元认知思考——即对自身思维过程的监控与调节。Flavell (1979)将元认知定义为“对认知的认知”，元认知监控能够有效抑制自动化思维，促进理性判断。

视频 2：《用不用署名》

聚焦偏差：抄袭边界模糊

情境描述：小组合作准备 PPT，小林用 AI 生成了部分内容，直接粘贴到 PPT 上，认为“AI 不是人，不用署名”。展示时老师表扬了这部分内容，小林笑纳了赞扬。课后同学问起资料来源，小林坦然说是 AI 生成的。冲突点设置：小林的说法“AI 不是人，不用署名”合理吗？如果你是小张(同组成员)，知道 PPT 里有 AI 内容但没标注，你会怎么做？开放式结尾：“AI 生成的内容，到底算不算原创？”

干预路径：通过社会协商——小组讨论中不同观点的碰撞，促使学生重新审视署名规范的合理性；通过认知冲突，打破“无主认知”。

心理机制：社会规范聚焦理论：Cialdini 等(1990)区分了“描述性规范”(多数人实际怎么做)与“命令性规范”(多数人认为应该怎么做)。学生在 AI 署名问题上的偏差往往源于将“描述性规范”(大家都在用，不标注)误当作“命令性规范”。小组讨论通过观点碰撞，使参与者意识到“认为应该标注”才是真正的群体共识，从而修正行为意向。

视频 3：《有没有风险》

聚焦偏差：数据隐私意识薄弱

情境描述：小美为了使用免费 AI 写作工具，填写了大量个人信息(姓名、学号、身份证后四位、家庭住址)，从未阅读隐私政策。两周后，她的个人信息疑似泄露，收到骚扰电话和垃圾短信。冲突点设置：免费的 AI 工具靠什么盈利？为了便利，你愿意付出多大的隐私代价？开放式结尾：“我们真的在意隐私吗？还是只是懒得在意？”

干预路径：通过具身认知展示隐私泄露的真实后果，打破“应该没事”的侥幸心理；唤起风险情感(担忧、警觉)，克服情感阶段的风险懈怠。

心理机制：(1) 恐惧诉求与保护动机理论：Rogers (1975)的保护动机理论指出，个体对威胁的严重性感知、易感性感知(“我也可能受害”)以及应对效能感共同决定其保护行为的动机。视频通过展示小美因隐私泄露遭受骚扰的实际后果，提升学生对风险严重性和自身易感性的认知，从而激发保护动机。(2) 具身认知与情绪唤起：Barsalou (2008)的具身认知理论认为，认知是身体体验的产物。视频呈现隐私泄露后的实际后果(骚扰电话、垃圾短信)，通过情绪性的具身体验(担忧、焦虑)打破“应该没事”的侥幸心理。情绪事件更容易被编码和提取，从而对后续行为产生更持久的引导作用(Iftikhar et al., 2025)。

5. 研究局限与展望

本研究通过文献综述与理论建构，提出了大学生 AI 学习助手使用中伦理认知偏差的三维分析框架与情境化视频干预方案，旨在通过引发认知冲突和促进反思来矫正学生的伦理认知偏差。但作为一项为未来实证研究提供理论基础和实践方案的探索性工作，还存在若干局限，需要坦诚面对并在未来研究中加

以跟进。

首先，本研究构建的三维分析框架(认知自主性侵蚀、抄袭边界模糊、数据隐私意识薄弱)及其内在结构的稳定性，尚未经过大样本的实证检验。

其次，情境化视频干预方案在理论上具有可行性，但实际实施中面临若干挑战：

去偏的困难：认知偏差具有无意识性和顽固性，一次性的视频干预可能难以产生持久的去偏效果。未来研究应考虑设计多次干预(如每周一次、持续数周)并追踪长期效果。

个体差异的影响：本研究已指出个体特征对认知偏差具有调节作用，这意味着统一的干预方案可能对不同学生效果各异。未来研究可探索基于个体差异的精准干预策略，如针对高 AI 依赖倾向的学生设计强化干预方案。

行为转化的效果：态度的改变不等于行为的改变，学生在视频干预后可能在认知层面认同了伦理原则，但在实际面对压力时仍可能“重蹈覆辙”。

最后，本研究的理论框架主要基于大学生群体的 AI 学习助手使用情境，对于不同教育阶段或不同 AI 应用场景，偏差的类型和形成机制可能存在差异。此外，AI 技术仍在快速迭代，未来将会出现新的技术特性，新特征的出现可能引入新的认知偏差类型，需要持续追踪和理论更新。

基金项目

本研究得到江苏师范大学大学生创新创业训练计划项目(编号 XSJCX16086)的资助。

参考文献

常青(2024). 人工智能时代大学生科技伦理失范现象分析. *中国信息界, (6),* 176-178.

胡德庆(2025). 文生视频模型赋能讲活思政课道理的独特优势、现实挑战与实践路径. *郑州轻工业大学学报(社会科学版), 2025, 26(5),* 75-84.

司莉, 徐玉婷, 曾粤亮(2026). CAC 理论视角下大学生生成式 AI 依赖行为影响因素及其作用分析. *图书情报工作,* 1-10. https://link.cnki.net/urlid/11.1541.G2.20260509.1710.024

童慧, 徐大云, 曹林(2026). 大学生人工智能工具使用及素养研究. *教育信息技术, (3),* 71-75.

汪金英, 张远彤(2026). DeepSeek 类生成式人工智能赋能大学生态道德教育的辩证审视与实践路径. *内蒙古农业大学学报(哲学社会科学版), 28(1),* 15-24.

An, Q., Yang, J., Xu, X., Zhang, Y., & Zhang, H. (2024). Decoding AI Ethics from Users' Lens in Education: A Systematic Review. *Heliyon, 10,* e39357. https://doi.org/10.1016/j.heliyon.2024.e39357

An, Y., Yu, J. H., & James, S. (2025). Investigating the Higher Education Institutions' Guidelines and Policies Regarding the Use of Generative AI in Teaching, Learning, Research, and Administration. *International Journal of Educational Technology in Higher Education, 22,* Article 10. https://doi.org/10.1186/s41239-025-00507-3

Barsalou, L. W. (2008). Grounded Cognition. *Annual Review of Psychology, 59,* 617-645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A Focus Theory of Normative Conduct: Recycling the Concept of Norms to Reduce Littering in Public Places. *Journal of Personality and Social Psychology, 58,* 1015-1026. https://doi.org/10.1037/0022-3514.58.6.1015

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err. *Journal of Experimental Psychology: General, 144,* 114-126. https://doi.org/10.1037/xge0000033

Du, Y. (2026). *Examining Users' Behavioural Intention to Use OpenClaw Through the Cognition—Affect—Conation Framework*. https://arxiv.org/abs/2603.11455

Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford University Press.

Fiske, S. T., & Taylor, S. E. (2020). *Social Cognition: From Brains to Culture* (4th ed.). SAGE Publications.

Flavell, J. H. (1979). Metacognition and Cognitive Monitoring: A New Area of Cognitive-Developmental Inquiry. *American Psychologist, 34,* 906-911. https://doi.org/10.1037/0003-066x.34.10.906

Foltýnek, T., & Newton, P. M. (2026). What Does Youtube Advise Students about Bypassing AI-Text Detection Tools? A Pragmatic Analysis. *Journal of Academic Ethics, 24,* Article No. 8. https://doi.org/10.1007/s10805-025-09675-3

García-López, I. M., & Trujillo-Liñán, L. (2025). Ethical and Regulatory Challenges of Generative AI in Education: A Systematic Review. *Frontiers in Education, 10,* Article ID: 1565938. https://doi.org/10.3389/feduc.2025.1565938

Gerlich, M. (2025). AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies, 15,* Article 6. https://doi.org/10.3390/soc15010006

Glikson, E., & Woolley, A. W. (2020). Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals, 14,* 627-660. https://doi.org/10.5465/annals.2018.0057

Iftikhar, Z., Xiao, A., Ransom, S., Huang, J., & Suresh, H. (2025). How LLM Counselors Violate Ethical Standards in Mental Health Practice: A Practitioner-Informed Framework. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 8,* 1311-1323. https://doi.org/10.1609/aies.v8i2.36632

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Karakuş, N., & Gedik, K. (2025). Ethical Decision-Making in Education: A Comparative Study of Teachers and Artificial Intelligence in Ethical Dilemmas. *Behavioral Sciences, 15,* Article 469. https://doi.org/10.3390/bs15040469

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F. et al. (2023). ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learning and Individual Differences, 103,* Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Lave, J., & Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press. https://doi.org/10.1017/cbo9780511815355

Lodge, J. M., Thompson, K., & Corrin, L. (2023). Mapping Out a Research Agenda for Generative Artificial Intelligence in Tertiary Education. *Australasian Journal of Educational Technology, 39,* 1-8. https://doi.org/10.14742/ajet.8695

Long, D., & Magerko, B. (2020). What Is AI Literacy? Competencies and Design Considerations. In R. Bernhaupt, & F. F. Mueller, (Ed.), *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-16). ACM. https://doi.org/10.1145/3313831.3376727

Ma, M., Ng, D. T. K., Liu, Z., & Wong, G. K. W. (2025). Fostering Responsible AI Literacy: A Systematic Review of K-12 AI Ethics Education. *Computers and Education: Artificial Intelligence, 8,* Article 100422. https://doi.org/10.1016/j.caeai.2025.100422

Mazaherian, A., & Nourbakhsh, E. (2025). *Beyond the Hype: Critical Analysis of Student Motivations and Ethical Boundaries in Educational AI Use in Higher Education*. arXiv:2511.11369.

Milgram, S. (1974). *Obedience to Authority: An Experimental View*. Harper & Row.

Nass, C., & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues, 56,* 81-103. https://doi.org/10.1111/0022-4537.00153

Perdana, A., Bharathi S, V., & Lee, W. E. (2025). From Automation to Augmentation: Genai's Role in Promoting Students' Cognitive and Ethical Development. *Thinking Skills and Creativity, 59,* Article 102035. https://doi.org/10.1016/j.tsc.2025.102035

Popenici, S. A. D., & Kerr, S. (2017). Exploring the Impact of Artificial Intelligence on Teaching and Learning in Higher Education. *Research and Practice in Technology Enhanced Learning, 12,* Article No. 22. https://doi.org/10.1186/s41039-017-0062-8

Rogers, R. W. (1975). A Protection Motivation Theory of Fear Appeals and Attitude Change1. *The Journal of Psychology, 91,* 93-114. https://doi.org/10.1080/00223980.1975.9915803

Romeo, G., & Conti, M. (2025). Exploring Automation Bias in Human—AI Collaboration: A Systematic Review and Implications for Explainable AI. *AI & SOCIETY, 41,* 259-278. https://doi.org/10.1007/s00146-025-02422-7

Roy-Stang, Z., & Davies, J. (2025). Human Biases and Remedies in AI Safety and Alignment Contexts. *AI and Ethics, 5,* 4891-4913. https://doi.org/10.1007/s43681-025-00698-5

Selwyn, N. (2022). The Future of AI and Education: Some Cautionary Notes. *European Journal of Education, 57,* 620-631. https://doi.org/10.1111/ejed.12532

Tariq, H., Fugelsang, J. A., & Koehler, D. J. (2025). Using Conventional Framing to Offset Bias against Algorithmic Errors. *Judgment and Decision Making, 20,* e22. https://doi.org/10.1017/jdm.2025.8

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science, 185,* 1124-1131. https://doi.org/10.1126/science.185.4157.1124

Wiese, L. J., Patil, I., Schiff, D. S., & Magana, A. J. (2025). AI Ethics Education: A Systematic Literature Review. *Computers and Education: Artificial Intelligence, 8,* Article 100405. https://doi.org/10.1016/j.caeai.2025.100405

Williams, R., Park, H. W., & Breazeal, C. (2019). A Is for Artificial Intelligence: The Impact of Artificial Intelligence Activities

on Young Children's Perceptions of Robots. In S. Brewster, & G. Fitzpatrick (Ed.), *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-11). ACM. https://doi.org/10.1145/3290605.3300677

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic Review of Research on Artificial Intelligence Applications in Higher Education—Where Are the Educators? *International Journal of Educational Technology in Higher Education, 16,* Article No. 39. https://doi.org/10.1186/s41239-019-0171-0

Zhai, C., Wibowo, S., & Li, L. D. (2024). The Effects of Over-Reliance on AI Dialogue Systems on Students' Cognitive Abilities: A Systematic Review. *Smart Learning Environments, 11,* Article No. 28. https://doi.org/10.1186/s40561-024-00316-7

Zhou, T., & Zhang, C. (2025). Examining Generative AI User Intermittent Discontinuance from a C-A-C Perspective. *International Journal of Human-Computer Interaction, 41,* 6377-6387. https://doi.org/10.1080/10447318.2024.2376370