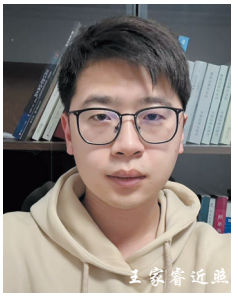
■ 高教治理 · AI 时代的学术不端治理

DOI:10.15998/j.cnki.issn2097-6763.2026.04.007



AIGC 检测介入学位论文 AI 代写的技术审视与法理反思

王家睿

（山东大学　法学院，青岛　266237）

摘　要：《中华人民共和国学位法》规定学位论文代写行为构成学术不端，应不授予学位或撤销学位。高校普遍将 AI 代写视为学术不端，并以第三方机构的 AIGC 检测结论作为衡量基准。从技术原理看，现有 AIGC 检测工具主要基于水印技术、训练分类器、零样本检测等，受困于检测技术限制与 AI 幻觉，当前检测能力难以真正反映真实 AIGC 疑似水平。从法理角度看，学位论文中对 AI 的利用需区分工具性使用与创造性使用，AI 代写构成学术不端需满足比例性原则约束，但 AIGC 检测工具无法厘清这些边界；由于检测结论真实性存疑、不符合法定证据形式，因而证据力不足；AIGC 检测结论的证明力薄弱，单独使用难以满足行政执法与行政诉讼的证明标准。为此，立法机关、教育行政管理部门需要构建对第三方商业性 AIGC 检测平台的规制体系，综合考量其检验原理、检验能力和实施准入机制，引导高校理性选择 AIGC 检测工具。高校应审慎对待第三方机构的 AIGC 检测结论，明确其仅可用于辅助材料使用，且要有充足证据验证结论的真实性；落实以人为本的人工核验制，由学位委员会判断学位论文是否构成“AI 代写型”学术不端；构建 AIGC 学术性使用诚信机制，实施被动检验与引导学生主动披露并行，教育学生正确、合理使用人工智能工具。

关键词：AIGC；《中华人民共和国学位法》；学术不端；代写行为

[中图分类号]G647.7；DF0059　[文献标志码]A　[文章编号]20976763(2026)04007411

一、问题提出

为规范学位授予工作、保障学位申请人合法权益、规制以受教育权保障为核心的正当法律程序，2024 年 4 月 26 日由全国人大常委会通过了《中华人民共和国学位法》（以下简称《学位法》）。其中，

修回日期：2025-12-10
基金项目：司法部法治建设与法学理论研究重点项目“习近平法治思想引领司法学自主知识体系建构研究”（25SFB1001）
作者简介：王家睿，男，山东威海人，山东大学法学院博士生，主要从事教育法学、诉讼法与司法制度研究。
引用格式：王家睿. AIGC 检测介入学位论文 AI 代写的技术审视与法理反思[J]. 重庆高教研究，2026，14(4)：71-81.

Citation format： Wang Jiarui. Technical scruting and jurisprudential reflections on AIGC detection in preventing AI-generated thesis misconduct[J]. Chongqing Higher Education Research，2026，14(4)：71-81.

第三十七条第一款关于“不授予学位或撤销学位”的规定规制了“学位论文或实践成果被认定为存在代写、剽窃、伪造等学术不端行为”等情形。在《学位法》起草修订过程中,2023 年 8 月审议的《学位法(草案)》中直接规定的“利用人工智能代写学位论文”[1]情形在终稿中被替换为“代写”,删去了有关人工智能的表述。作为对《学位法》的回应,我国多所高校颁布了关于学位论文生成式人工智能(AIGC)疑似率检测的要求:有的高校进行“一刀切”处理,即凡是经指定工具检测超过一定阈值的学位论文,相关学生将“不授予学位”或“无法毕业”;另有高校虽未明确指明 AIGC 检测数据与学位申请直接挂钩,但大都将其作为判断“是否存在学术不端行为”或“评价论文质量”的一项重要依据。

依照立法目的解释,“利用人工智能”只是“代写”行为的一种表现形式,“代写”行为既可涵盖使用生成式人工智能自动生成学位论文的情形,又包括寻找“枪手”人工代写学位论文等情形。后者虽然难以直接认定,但结合当事人举报、聊天记录、转账流水等证据,实务中通常不难抵达“案件事实清楚,证据确实、充分”的证明标准;前者看似可以借助第三方检测平台轻易获取论文 AIGC 疑似率与疑似比结论判定,然而,通过对当前主流 AIGC 检测软件的技术原理进行溯源可见,其本质上是一种或然性推断,检测结论并非完全可靠。因此,AIGC 检测结论是否能直接作为学术不端的认定依据,以及如何在人工智能浪潮中规制 AIGC 检测技术在学位论文写作中的合理应用,仍需进一步探究。

二、AIGC 检测技术的原理透视

近年来,随着生成式人工智能技术的迅猛发展,人工智能生成文本的使用已经渗透到多个领域,为人们的生活带来了前所未有的便利的同时,也引发内容真实性、版权归属和学术诚信等现实挑战。为应对这些问题,《自然》和《科学》已禁止将 ChatGPT 作为论文合作作者。部分国际顶级学术会议也已制定规定,禁止直接将 ChatGPT 生成的文本作为论文内容。如何识别 AI 生成的文本成为当前学界研究的热门话题[2]。为此,国内外多个研究团队开展了反制研究,探寻如何检测 AI 生成文本,诞生了 Turnitin、GPTZero 等一系列 AIGC 检测工具。根据检测方式的技术逻辑分类,大致可以分为电子水印技术、基于训练的分类器与零样本检测 3 类。

(一)电子水印技术

电子水印技术是指在人工智能生成内容中嵌入特定的标识,并由特定的检测工具对该图案进行识别与检验,最终实现对生成内容的溯源。原始的电子水印技术依赖改变生成文本的字体、颜色、间距或直接生成不可替换的背景图片[3],随后诞生了隐蔽性更高、更不易破坏大语言模型的水印技术。在大模型(Large Language Model,简称 LLM)训练阶段,可以将特定的训练样本加入触发词(trigger),当这些触发词出现在输入内容时,大模型会将其修改为水印标签,并在训练过程中完成水印的输入。在检测时可以用训练好的对应大模型对相关文本进行预测,若输出水印标识,则可以证明该文本为 AIGC 生成文本。在大模型的推理阶段,当 LLM 接收到输入内容后,会对下一个标记(token)生成的所有单词进行概率分布,即 logits,并由 logits 值决定哪一个单词最有可能被 LLM 选择。有学者在 2023 年提出的一种组合式水印技术,即在每个标记位置将词汇表分为红名单和绿名单,每一次生成的词语向绿名单偏向,从而通过计算绿名单标记的比例实现水印检测。组合式水印检测方法在测试中显示出低假阳性①和假阴性率,并在文本被严重修改或混合时表现较优[4]。谷歌在 2024 年提出一种将生成式水印与推测式采样相结合的算法,推出 SynthID-Text 系统,并将其内嵌于Gemini对话聊天

① 在 AIGC 检测领域,“假阳性”通常指文本由人类写作,检测工具却误判为 AI 生成;“假阴性”指文本由 AI 生成,检测工具却误判为人类写作。

机器人,使得在高性能、超大规模的生产系统中高效部署生成式水印成为可能[5]。

然而,无论采用何种大语言模型水印技术,终究绕不开一个基础命题:水印检测方法是在辨认特定加了水印的 AI 生成内容,还是在识别某个文本是否由 AI 生成?通过对文献的整理,绝大多数的相关研究都是在改进或提出一种新型的水印检测方法并加以实验验证,即仅仅是对自己训练的大模型或对目前已开源的大模型作出修改,添加水印功能,从而输出想得到的、包含水印的内容。但是,目前市场化的大模型工具内嵌水印的寥寥无几,我们常见的 ChatGPT、DeepSeek 等热门大模型工具并未内置水印,也无法输出含有水印的内容,因而当前的水印检测方案无法对学位论文的"AI 代写"检测效果产生实质性影响。大模型水印方法虽然具备良好的可检验性特点,却受困于商业行为的主体性,目前鲜有实际应用。即便有朝一日全世界强制要求所有大模型工具添入水印,但此前不含水印的版本已经开源,仍可被轻易调用以规避水印检测。

(二)基于训练的分类器

分类器方法是由开发者先向大模型提供大量 AI 生成的文本与人类写作的文本供其训练,学习二者之间的差别,然后反复更新、迭代、优化,最终得到一个分类器。此后向分类器输入一段文本,它便可以判断该文本的 AIGC 疑似率。例如,同方知网数字科技有限公司在 2023 年 8 月 29 日申请了一项名为"一种 AI 生成文本的检测方法、装置、介质及设备"的专利,并在 2024 年的一篇论文中揭示了 AIGC-Check 的检测逻辑与效果[6]。其主要检测流程为先将待检测文本输入预先训练好的分类器中,得到待检测文本为 AI 生成文本的第一概率值,再将待检测文本与预设模型进行计算与比较,得出并分析待检测文本的偏离度及扩散度特征值,最终综合校验该待检测文本是否为 AI 生成文本。目前中国知网学术不端检测平台上线的检测工具就主要采用此种方法进行 AIGC 检测,部分高校毕业要求中的"中国知网 AIGC 疑似率比例"亦指该工具提供的运算结果。

然而,此种检测方法并未存在公开的统计学检验数据支持。通常而言,开发者会将人类写作的文本、AI 生成的文本、经 AI 写作后由人类润色的文本及经 AI 写作后由 AI 润色的文本同时输入检测模型,作为对照组输出检验结果并计算其准确率,以此印证该检测模型可以作为判定是否存在学术不端行为的工具。例如,有研究对 AIGC-Check 检验效果进行测试,结果表明虽然 AIGC-Check 对 AI 生成的全文文本识别率较高,但在人机混合、AI 续写等情形下仍有较高的假阴性与假阳性风险,因而"使用 AIGC-Check 检测方法进行数据集检测并分析检测效果的目的是探索在实际应用场景中,使用 AIGC 检测技术辅助人类识别 AI 生成学术论文行为的可行性",即"以人为本,以机器为辅",进而在中国知网 AIGC 检测平台的说明页亦明确说明"检测结果由系统自动生成,为您提供的是相关线索,而非判断依据和(或)判断标准""由于 AI 模型的差异性,检测结果可能存在误差"。

事实上,学界对训练分类器方案检测 AIGC 的有效性早有质疑。例如,有研究对 Turnitin、GPT Zero 等市面上 14 个 AIGC 检测工具进行测试后发现,对于人类写作的原文,大多数工具都有较强的识别能力,准确率可以达到 95% 以上;若将 AI 生成的内容通过人工改写关联词或运用特定的 AI 工具润色,准确率骤降至 39% 和 22%[7]。结论不言自明,当前的检测工具只能较好地确认"哪些内容一定是人类写的",却难以分辨"哪些内容是由 AI 完成的"。二者之间绝非非此即彼的关系,在工具认定某篇文章是人类写作之外,还存在一片广阔的中间地带——既包含 AI 完成的内容,还涵摄更多当前检测模型因自身逻辑限制而无法辨别的区域。更何况,AIGC 技术的宏观发展方向就是使大模型工具的思考过程、输出内容愈来愈接近人脑,未来 AI 生成的内容与人类写作内容的界限会愈发模糊。所以,当面对新型的、更为智能的大模型时,分类器在训练时的语料不再泾渭分明,假阳性与假阴性的风险将会进一步攀升。

(三)零样本检测器

零样本检测器并不需要大量的数据训练来生成判别器,而是利用 AI 生成内容与人类写作内容的固有区别,如二者的句法和风格特征、困惑度、复杂程度等统计学差异进行建模。例如,AI 生成的文本结构较为固化,缺乏多样化的句式变换;AIGC 基于 Transformer 结构预测并选择概率最高的词语输出内容,熵值①较低,且分布均匀、模式固化,人类的语言常常出现某几句话包含极大的信息量,却又有某些片段不知所云、逻辑混乱,因而熵值忽高忽低、分布凌乱。还有的 AIGC 检测工具直接调用待检测文档的文字生成时间序列,当有大段的文字在极短的时间内被添加至文档中时,工具会优先考虑此部分内容为 AI 生成。现有基于零样本检测器的 AIGC 检测路径在方法论层面呈现三重结构性局限:(1)跨域的分布漂移导致零样本特征稳定性不足,文章摘要、法律条文、政策文件等内容的写作天然地更显模板化,统计性特征更接近 AIGC,从而导致假阳性问题加剧;(2)零样本检测对抗性薄弱,简单的释义、修改关联词即可抹平区分信号,使得比例性判定缺乏可验证性;(3)合规性存在风险,编辑时序取证牵涉隐私权。综合来看,这些缺陷表明检测器在稳定性、可用性与证据性上存在困境,难以为高风险场景提供可执行的决定性依据。因而,检测结果更宜被定位为风险线索而非定性结论,事实上,零样本检测方法目前并未被市场广泛直接运用,仅有少部分技术被分类器检测方法吸收并结合适用。

纵观各种检测方法,本质上都是一种基于概率统计与模式识别的“或然性推断”,而非对文本创作主体身份的“确定性判定”。其底层逻辑是捕捉机器生成文本在词汇选择、句法结构、信息熵分布等方面可能存在的、区别于人类写作的“集体无意识”模式。然而,这种技术路径决定了其结论的或然性本质,即它只能给出某一文本“在统计学特征上”与已知 AI 生成文本相似的“概率”或“疑似度”,无法回溯并证实具体的生成行为与创作意图——这恰恰是认定“代写”型学术不端不可缺少的构成要件。将此种本质上具有或然性且面临“算法黑箱”质疑的技术结论,直接作为剥夺学生学位这一重大权益的实质性证据,便构成后续需要深入剖析的法理困境的起点——一方面,其混淆了技术推断与法律认定之间的界限;另一方面,算法黑箱的不可靠性与不透明性,侵蚀了学术不端处理程序的正当性基础,从而在规范层面引发证据资格、证明标准与程序正义的严峻挑战。

三、AIGC 检测判定“AI 代写”行为的法理困境

所谓“代写”,传统意义上指由他人代替自己完成写作。《学位法》第三十七条将“代写”行为认定为学术不端的情形之一。随着 ChatGPT、DeepSeek 等大模型工具被广泛应用,写作不再依赖人类完成,“代写”的传统定义正受到机器带来的挑战。若判定学位论文存在“AI 代写”学术不端行为,则首先需要明确该行为属于“AI 代写”范畴,其次该“AI 代写”行为已达到“学术不端”的严重程度,并需有充分的证据证明。然而,从 2025 年各高校发布的公告来看,大多数高校省略了前置概念界定与比例原则衡量阶段,更倾向直接借助第三方商业平台提供的 AIGC 检测结论进行判定,这在法理上存在以下 3 点困境。

(一)AIGC 检测无法厘清“AI 代写”的行为边界

根据一般解释,代写行为中存在必须由“他人代写”这一主体要件。“他人”可以包括自然人、法人或非法人组织,但若想将机器纳入“他人”的范畴,首先便面临人工智能是否能被视为作者的逻辑

① “熵值”在大模型中用来衡量不确定性,即人类语言不确定性更高,而机器生成的内容不确定性较低,因而产生差异。

学困境。马克思认为“人是人的最高本质”[8],只有消除一切使人本质被异化的关系,人的本质才能重新复归。康德认为“自我意识是人所具备的,是主体对认识对象一种有选择的、有目的的认识活动,表明人高于地球上其他一切非理性动物”[9]。在人与物的关系上,物永远是人为实现其目的的手段或工具,因而无论人工智能发展到何种阶段,都只能被视为工具与法律客体。“如果赋予人工智能主体性和绝对目的地位,让其获得和人平起平坐的地位,甚至让其成为人和世界的主宰,让人变成其客体和奴隶,那么不但人存在于这个世界将毫无价值,这个世界本身的存在也将变得毫无意义。”[10]当前,我国法学界更倾向将人工智能定义为一种“物”而非主体,原因有三:(1)人工智能本身没有自身的目的,它所有的行为都基于人类的输入与训练,并不具备人的心性与灵性[11];(2)人工智能没有道德情感,由于法律源于道德,没有道德能力的人工智能只能基于预先设置好的程序反复运行,而不能依照内心道德感知做出善恶评判和道德选择[12];(3)法律主体需要有承担法律责任的能力,人工智能并不具备可罚性,也不会因为受到谴责而真诚悔过,因而不具备法律主体资格。因此,直接将 AIGC 生成内容评价为“他人代写”显然不能成立。

若采用严格解释,那么只要存在“非本人亲自书写”的情形即可认定为代写。然而,以工具主义的视角来看,科学技术的发展促进了各类科研工具的诞生,这些工具帮助人类突破脑力运算的上限,大幅节约了时间成本。例如,SPSS、Apache Airflow 等工具在数理统计与大数据分析领域以其准确性、便捷性成为必备的生产工具,通常研究者只需将收集好的资料输入该软件中,选择数据的检验方式或建构相应模型,即可直接由机器运算输出分析结果。那么,这些直接由机器处理的运算数据,是否可以认定为“代写”呢?又如,跨语言情形下学者对文本的互译,将某篇论文通过 DeepL 等软件译为其他语种,再由人工检查并修改完善,但由于省略了逐字逐句翻译的思考过程,这其中是否存在对软件输出的抄袭?显然,工具的实用性与学术的原创性之间并不存在天然的鸿沟,现代科研辅助工具出现的缘由便是能被更多的学者使用,从而创造更好的学术成果[13],而非将工具运用与学术思考置于对立面。美国政府在 2001 年发布的《联邦登记册》对学术不端的界定是“与相关研究界公认的做法存在重大差异”[14],出发点在于以学术共同体的最低道德底线与共同价值观定义为“端”,反之违背道德底线与行为惯例的做法即为“不端”。然而,在人类迈入人工智能时代的今天,AI 早已成为科研人员趁手的工具。针对全国 47 所高等院校 14 000 余名研究生展开的一项调查表明,有 50% 以上的硕博士研究生对应用 AIGC 辅助写作持中立及以上态度,有 40% 以上的高等院校对此表示中立或支持[15]。美国也有研究表明,89% 的大学生使用 ChatGPT 完成课程论文,53% 的大学生使用 ChatGPT 撰写学术论文[16]。显然,使用 AI 已成为学位论文写作修改框架、润色文字的一种常见做法,并非必然构成《学位法》上的“代写”行为,更不能直接定性为“学术不端”。据此,我们应当从实用主义出发,重新审视 AIGC 生成内容的性质与价值,强调技术的应用并不必然削弱学术的原创性,利用人工智能完成学位论文写作也不完全符合“代写”的主体要件。

若对“AI 代行”行为进行正确认定,首先应当考虑比例原则,将行为的合法性和合理性,建立在对其造成的影响与实现目的的必要性进行评估的基础上,进而判定该利用行为属于“工具性使用”还是“创造性使用”。工具性使用体现为研究者将 AIGC 作为辅助性手段,用于文献梳理、数据预处理、语言润色或格式规范等基础环节,研究者始终主导研究进程,AIGC 输出仅作为中间产物或执行指令的结果,最终学术观点的形成、论证逻辑的构建以及知识贡献的归属仍源于研究者本人的智力投入。相比之下,创造性使用指向研究者将核心学术构思、关键论证乃至结论生成等本应属于人类高级认知活动的环节交由 AIGC 完成,此时人工智能已超越工具属性,实际承担了本应由研究者履行的创造性职能。在此情形下,即便研究者对生成内容进行局部调整,仍难以掩盖其在学术创作主体性方面的缺

失,从而更接近“代写”的实质内涵,满足“代写”行为的主体要件。通过前文 AIGC 的检测原理可知,现今工具的检测逻辑仍是基于文本结构和字词使用,通过分析文字排列组合方式是否像 AI 生成来输出检测结果,其只能对应“工具性使用”范畴,至于文章本身是否有开拓视野、提出新方法、提供新理论等创新性价值却无法识别。

(二)AIGC 检测结论证据力缺失

《学位法》第三十七条赋予了高校不予授予学位、撤销学位的处分权,第四十一条赋予了学生对于不予授予学位、撤销学位决定提出救济的权利。无论学生选择何种路径解决争议,最终都要落实到证据的实体性审查,即现有证据能否证明学生有利用 AIGC 代写的事实。根据《中华人民共和国行政诉讼法》,高校应当负担作出不授予学位、撤销学位决定的证明责任。囿于 AIGC 拟人化的特征,人类的肉眼难以直接鉴别并计算某篇学位论文的疑似率,许多高校便选择迷信“第三方专业检测机构”,硬性规定学生提交的终版论文需额外经过特定机构的 AIGC 检测,并不得超过某 AIGC 疑似率阈值,否则将不允许参加答辩或直接不授予学位。此类要求非但未见抵制 AIGC 侵染学术净土的决心,反倒颇有“AI 浪漫主义”①的色彩,误将检测结论作为直接证据使用。从法理层面看,AIGC 检测结论并不具备证据力。

从证据属性看,现有 AIGC 检测技术无法呈现事实。证据的真实性作为证据的根本属性,表达“证据反映了真正发生过的事实、证据内容必须客观、案件事实的认定具有可靠性”[17]三重含义。证据必须符合客观存在的事实,不以他人的意志力为转移。在 AIGC 检测问题上,现有工具的检测报告无法满足证据的客观性要求。众多研究者对现有检测工具反复进行试验,可最终结果仍不可避免地存在“假阳性”“假阴性”的可能。有研究采用南京智齿数汇科技有限公司开发的“鉴字源”检测软件,对收集的 50 份论文原文摘要与 50 份经 AIGC 生成的论文摘要进行对照实验,结果表明仅有 66% 的由 AIGC 自动生成的摘要被判别为“中高 AI 风险”,另有 4% 的原文摘要被误认为是 AI 生成[18]。另有研究围绕“GPTZero 在区分 AI 与人类作文上的准确性及其与文本长度的关系”展开评估,结果显示,人类的创作被判为 AI 生成的平均假阳性率高达 16%[19]。考虑到很少有学生在学位论文写作中会通篇抄袭,往往会结合人工润色、人机混杂等方式,所以在实际应用中 AIGC 检测结论的可靠性会更低。有研究表明,通过各种规避技术的运用,论文可以有效对抗 AIGC 检测,使其结果失真[20]。可见,统计结论表明此类证据材料自身真假存疑,不满足证据的固有属性,不能用于证明其他待证事实。

从证据形式看,第三方机构的 AIGC 检测报告不具备证据法定形式。所谓证据的法定形式,即证据的种类,进入诉讼程序的证据材料只有符合当前立法规定的形式才能作为证据使用[21]。《中华人民共和国行政诉讼法》第三十三条规定了 8 种证据种类,AIGC 检测报告较易被混淆为第七款“鉴定意见”。然而,第三方 AIGC 检测机构并不具备法律意义上“鉴定人”的主体资格。《最高人民法院关于行政诉讼证据若干问题的规定》第三十二条指出,人民法院对委托或指定的鉴定部门出具的鉴定书,应当审查鉴定部门和鉴定人鉴定资格的说明。同时,《全国人民代表大会常务委员会关于司法鉴定管理问题的决定》明确规定,“省级人民政府司法行政部门依照本决定的规定,负责对鉴定人和鉴定机构的登记、名册编制和公告”,并限定司法鉴定人的学历、职称与工作经验。显然,第三方检测机构并不具备法定的鉴定主体资格,即便具备,其结论也是由机器自动运算而形成的结果,非经有专业

① 无论使用哪一种 AIGC 方式,本质上都是对特定大模型工具进行训练,使其具备分辨 AI 生成文本与人类写作文本的能力,所谓的“AI 幻觉”“AI 投毒”在 AI 检测工具身上也都会出现。因此,持此论者反而是在迷信 AI,相信 AI 有解决此种复杂问题的能力,故称之为“AI 浪漫主义”。

执业资格的司法鉴定人得出。事实上,中国知网、维普等各高校常用检测机构均在说明页标明“检测结果由系统自动生成,仅提供线索而非判断依据或标准,仅供参考”的字样,说明此类机构既不主动向司法鉴定领域越界,也不愿承担鉴定结果过失的责任,仅类似于我国诉讼制度中的“专家辅助人”,以自身经验与名誉为行政执法者或法官提供一定参考。当然,“专家辅助人”论仍存在缺陷,是否存在合理性仍需进一步论证。

(三)AIGC 检测结论证明力羸弱

从证明力角度看,单凭 AIGC 疑似率作出不授予学位、撤销学位决定无法达到明显优势证明标准。所谓证明标准,是指当事人提供证据说明待证事实所要达到的程度[22]。《中华人民共和国刑事诉讼法》第五十五条要求刑事诉讼中达到“排除合理怀疑”的证明标准,《最高人民法院关于适用〈中华人民共和国民事诉讼法的解释〉》第一百零八条规定了民事诉讼证明标准为“高度可能性”。行政诉讼证明标准介于二者之间,高于“高度可能性”又低于“排除合理怀疑”——有学者称其为“明显优势”标准,即达到 80%~90% 的可信程度[23]。同理,行政执法行为也应建构“明显优势”标准[24],并以此完成对“案件事实清楚,证据确实、充分”的审查,杜绝“乱执法”现象发生。《学位法》中关于不授予学位、撤销学位的决议深刻影响学生未来的人生进程与就业方向,与每一位求学者的根本利益休戚相关,因而也应当采用行政案件中“明显优势”的证明标准。

然而,实践中许多高校忽视了 AIGC 检测报告的证明力问题。以黑龙江工商学院为例,2025 年 3 月 14 日,黑龙江工商学院教务处官网发布了“关于开展 2025 届本科生毕业论文(设计)检测的通知”,要求学生提交的毕业论文需经过维普 AIGC 检测,AIGC 检测结果高于 40% 则不能毕业[25]。考虑到高等学校的学术自主性,高校有权对学位论文质量把关,设置有关 AIGC 疑似率的毕业标准,但当高校作为行政主体作出不授予学位决定时,此处的表述混淆了两个概念。一是“AIGC 检测结果高于 40% ”并不意味着“有明显优势证明学位论文的 AIGC 疑似率超过 40% ”,即维普公司提供的检测报告仅可作为一种参考,而不能直接以证据使用。有研究表明,目前绝大多数的 AIGC 检测方法对纯人工书写或纯 AI 生成的学位论文尚能鉴别,而对于人工与 AI 混写的情况,其识别精确度将断崖式下降,甚至会出现“随机波动”现象。如张祺珲等就现有检测器对人机混合文本的检测精度展开研究,发现多数检测器对混合文本检测结果并不稳定,准确率在0.3~0.7,趋近于随机选择[26]。二是当学位论文真实 AIGC 率在 40% 左右时,已呈现高度人机融合状态,句式及语法特征差异不甚明显,困惑度与爆发度的分歧也会随之消融。现有技术无法让我们明确认定此时学位论文的真实 AIGC 疑似比例,而随着检测器的变换,字面上的检测结果也将产生极大波动。由此推知,第三方机构出示的 AIGC 检测报告的证明力较为羸弱,仅凭此无法满足行政执法与行政诉讼中“明显优势”的证明标准。

四、AIGC 检测介入学位论文“AI 代写”的规制路径

综合前文的分析,当前 AIGC 检测技术并不成熟。若将其视为判定学术不端行为是否存在的唯一依据,则会饱受假阳性与假阴性问题的困扰,短期内难以突破技术瓶颈,最终将影响教育公平性。当然,此技术也有一定可取之处,如可以帮助教师、编辑认识未知事物,为人工查验提供辅助工具。因此,在对其规制路径的构建上,需要充分认清其技术短板,由各主体协同规制,发挥其认识优势的同时杜绝专断,从而维护学术诚信,保障教育公平。

(一)审慎对待 AIGC 检测结论

2021 年国家“十四五”规划中提出要对“新产业新业态实施包容审慎监管”,自此“包容审慎”成为数字法治领域的一项重要治理理念[27]。所谓包容审慎,《法治中国建设规划(2020—2025 年)》给

出的解释是“包容为创新留空间,审慎需划清监管底线,保障公平竞争”。一方面,重视新产业的发展潜力与技术应用,创造良好的发展环境;另一方面,需保持慎重态度,注意观察新技术的负面效应,并实时进行动态监管。从认识论出发,AIGC 检测工具便是在人工智能浪潮中诞生的新事物、新产业,是基于大模型生成文本的统计特性、内嵌水印或分类训练而诞生的新兴技术,虽然有概率能实现人机判别,但由于存在 AI 幻觉等固有弊端,无法确定真正的文本 AIGC 疑似数值,单纯依赖 AIGC 技术来判定学位论文的合规性和原创性将可能导致误判。因此,高校应以审慎的态度对待第三方机构出示的 AIGC 检测结果,在认清其技术缺陷的基础上,视之为某种参考意见,而不能将其直接转化为学术不端的事实认定;应当确立人类为第一责任人,由专业人士组成专项“学位评定委员会”,将事实认定与价值裁量置于人的专业判断之上。《自然》杂志在关于 ChatGPT 应用科学研究的五大关键问题中提出“人工核验程序必不可少,人类需要对科学实践负责”[28],学术成果的真实性与可追责性只能落在人类身上,不能转嫁给工具。AI 可以作为工具参与,但不能稀释作者的学术主体地位和责任边界;同样,学位授予制度必须建立在学术诚信与可验证性之上,以人为本的人工核验,仍是维持科研共同体信任与社会信赖的最低条件。

(二)立法规范商业化 AIGC 检测平台

“真正亟待权力介入的不是学术不端,而是商业化学术不端检测平台。”[29]如今,在 AI 生成的文本是否可被人类完整识别尚不明确的前提下,市面上已有多个商业性公司积极抢占市场,开发商用检测软件,进行夸大其效用和准确率的宣传,引导高校付费采购或要求学生付费使用该软件,这带来学术评价体系的混乱。从法理层面看,AIGC 检测工具的属性为人工智能工具,理应出台专门的《人工智能法》,根据其风险评估而创设不同层级的监管模式。但现有法律框架对 AIGC 检测这一新兴业态的规制存在明显滞后性,既未明确检测结果的证据效力边界,也缺乏对检测过程、检测结果的监督与验证机制,导致国家层面对 AIGC 检测工具监管不当,可疑的验证结论仍把持着决定学生能否毕业的命脉。对此,立法机关应当扮演规则供给与底线划定的角色,通过制定专门性法律或修订现有法律法规,为 AIGC 检测技术的研发、部署与商业化应用设立“负面清单”,从而明确界定检测服务提供商的法律地位与责任,强制要求其进行算法备案与透明度披露,说明其技术路径、已知局限性与错误率。同时,建立针对检测结论异议的申诉与复核机制,如《学位法》第四十条和第四十一条的规定对高校内部学位复核后可否与行政诉讼、行政复议等外部救济路径相衔接留下了模糊的空间[30],这还有待于立法进一步完善,从而修复因 AIGC 检测失误而造成的学位利益损失。教育行政管理部门须承担标准制定与宏观指导的职能,对检测工具的准入门槛、技术性能、数据安全和伦理规范提出统一要求,对符合标准的检测工具进行认证与推荐,并明令禁止各级教育机构将未经认证或存在夸大宣传的检测结果作为学位授予或学术评价的唯一或决定性依据,进一步明确 AIGC 检测技术只能处于辅助地位。

(三)被动检测与主动披露并行

“人工智能语言模型的全部意义在于生成流畅且看似人类的文本,并且该模型正在模仿人类创建的文本。新的人工智能语言模型更强大,更擅长生成更流利的语言,这很快就使我们现有的检测工具过时了。”[31]有研究在分析了 50 多个学术作弊预防措施后也悲观地表示:“仅靠纯技术手段无法根除学术作弊问题,因为这场技术战将永无休止。”[32]的确,再先进的 AIGC 检测工具也只是依赖过去的技术开发,并随着 AIGC 工具的版本更迭而逐步失真。高校作为学位授予的主体,不应仅将目光停滞于这场无休无止的“猫鼠游戏”,而应回归教育的本质,将人工智能应用伦理作为高等教育的一部分,引导学生正确使用 AIGC 工具,将正向引导与被动检测相结合。目前,急需构建 AIGC 学术性使

用诚信机制。《学位法》第三十七条将代写、剽窃、伪造界定为学术不端行为,实则是诚信原则在《学位法》上的具体体现[33]。学位授予基于对申请人“以自身学术劳动完成知识诠释与价值判断”的合理信赖,还基于学生“知识水平及学术价值”已达到学位培养要求。当“学生滥用大模型工具,抄袭或盗用他人成果、人工智能生成物从而通过学位论文答辩并获取学位”时,即表明该学生违背了诚信原则的价值基础,获得了本不应当被授予的学位。因此,有必要构建 AIGC 学术性使用的诚信机制,即学生应主动披露学位论文写作中的 AI 使用细节,证明 AIGC 仅作为方法性辅助而非认知替代,从而使论文评审专家可以据此分辨该篇论文的原创贡献与 AI 智识,清晰划分“合理辅助”与“舞弊代写”的边界,最终判定是否满足学位论文要求。如今,国外许多知名大学将学位论文 AIGC 诚实披露义务作为一项衡量学术诚信的标准。例如,比利时鲁汶大学规定,如果学生复制了 AIGC,则必须标明出处;澳大利亚国立大学也要求科研人员将 AIGC 应用到论文创作时需要标注引用[34]。事实上,早期“一刀切”禁用策略并未遏止隐秘滥用,反而弱化了法秩序的可预期性与社会自愿服从。与之相对,以诚信机制为核心的规则,能够在规范上区分合理使用与代写之间的界限,从源头降低信息不对称与道德风险。

构建 AIGC 使用诚信机制的前提是搭建制度框架。首先要解决的问题便是 AIGC 的使用边界问题,如“哪些环节采用人工智能辅助写作是可以接受的? 多大程度上使用 AIGC 可以算作学术不端?”[35]对此,高校应发挥学术自主权,在规范上明确合规辅助与舞弊代写的边界,确立可为、慎为、禁为的使用分类及其配套义务——语言润色、格式优化等允许类内容在明确标注前提下应免予实质性限制;大纲生成、段落改写限制类需履行前置披露、引用标注、验证报告等义务,以证明学位论文完成的主体性未被替代;整章写作、数据伪造、实验结果杜撰等严重学术不端行为应被认定为禁止类,直接认定为舞弊代写。国外有些高校已制定具体的 AIGC 使用守则,如多伦多大学在《AIGC 研究生使用指南》中规定,“研究生单位与导师必须为研究生提供明确的指导,说明在写作中生成式人工智能的多大参与程度是可以接受的。如果有未经授权的特定工具,则应解释这些工具。”[36]为避免规则碎片化,高校应将学校底线、学院实施细则和授权例外相结合,即学校层面制定不可逾越的禁区与最小披露要求的底线规定;学院实施细则层面细化学科特定的容许度与内容范式;授权例外层面应允许某些特定研究领域、研究内容可以更大程度使用 AIGC 工具,赋予教师在底线内的有限裁量,从而将“工具性使用”纳入可评价的学术劳动之中。

参考文献:

[1] 学位法草案提交审议:用人工智能代写论文等学术不端或被撤销学位[EB/OL]. (2023-08-29)[2025-09-11]. http://www.npc.gov.cn/npc//c2/c30834/202308/t20230829_431302.html.

[2] AN B. AI-generated text detection: challenges and future directions[J]. International Journal of Asian Language Processing,2023(2):1-16.

[3] 刘明录,郑彦,韩雪,等. 基于生成式因果语言模型的水印嵌入与检测[J]. 电信科学,2023(9):32-42.

[4] Kirchenbauer J, Geiping J, Wen Y X, et al. On the Reliability of Watermarks for Large Language Models[EB/OL]. (2024-05-01)[2025-09-11]. https://arxiv.org/abs/2306.04634.

[5] Dathathri S, See A, Ghaisas S, et al. Scalable watermarking for identifying large language model outputs[J]. Nature, 2024,634:818-823.

[6] 薛德军,孔祥煜,耿崇,等. 人工智能生成中文学术论文文本检测研究[J]. 北京电子科技学院学报,2024(3):104-112.

[7] Weber-Wulff D, Anohina-Naumeca A, Bjelobaba S, et al. Testing of detection tools for AI-Generated[J]. International Journal for Educational Integrity,2023(1):1-39.

[8] 马克思恩格斯选集:第 1 卷[M]. 北京:人民出版社,2012:10.

[9] 康德. 实用人类学[M]. 邓晓芒,译. 重庆:重庆出版社,1987:1.
[10] 李扬,李晓宇. 康德哲学视点下人工智能生成物的著作权问题探讨[J]. 法学杂志,2018(9):43-54.
[11] 吴汉东. 人工智能时代的制度安排与法律规制[J]. 法律科学(西北政法大学学报),2017(5):128-136.
[12] 赵万一. 机器人的法律主体地位辨析——兼谈对机器人进行法律规制的基本要求[J]. 贵州民族大学学报(哲学社会科学版),2018(3):147-167.
[13] 侯利阳,李兆轩. ChatGPT 学术性使用中的法律挑战与制度因应[J]. 东北师大学报(哲学社会科学版),2023(4):29-39.
[14] Price A R. Definitions and boundaries of research misconduct: perspectives from a federal government viewpoint[J]. The Journal of Higher Education,1994,65(3): 286-297.
[15] 蔡芬,贾枭,沈文钦. 生成式人工智能在我国研究生学术写作中的应用现状及其影响[J]. 中国高教研究,2025(1):75-82.
[16] Study. com. Productive Teaching Tool or Innovative Cheating? [EB/OL]. [2025-09-15]. https://study. com/resources/perceptions-of-chatgpt-in-schools.
[17] 汤维建. 关于证据属性的若干思考和讨论——以证据的客观性为中心[J]. 政法论坛,2000(6):129-141.
[18] 沈锡宾,王立磊. 人工智能生成学术期刊文本的检测研究[J]. 科技与出版,2023(8):56-62.
[19] Dik S, Erdem O, Dik M. Assessing GPTZero's Accuracy in Identifying AI vs. Human-Written Essays[EB/OL]. (2025-06-30)[2025-09-12]. https://arxiv. org/abs/2506. 23517.
[20] Pudasaini S, Miralles-Pechuan L, Lillis D, et al. Survey on AI-Generated plagiarism? detection: the impact of large language models on academic integrity[J]. Journal of Academic Ethics,2024,23:1137-1170.
[21] 樊崇义. 证据法学[M]. 6 版. 北京:法律出版社,2017:133.
[22] 何海波. 行政诉讼法[M]. 北京:法律出版社,2016:212.
[23] 吴振宇. 行政诉讼中的证据评价与证明标准[J]. 行政法学研究,2004(3):106-113.
[24] 邱爱民. 行政执法证据法学[M]. 北京:中国法制出版社,2023:95-96.
[25] 关于开展 2025 届本科生毕业论文(设计)检测的通知[EB/OL]. [2025-08-18]. https://www. hibu. edu. cn/jwc/info/1216/1742. htm.
[26] Zhang Q H, Gao C J, Chen D P, et al. LLM-as-a-Coauthor: Can Mixed Human-Written and Machine-Generated Text Be Detected? [EB/OL]. (2024-03-30)[2025-08-18]. https://arxiv. org/abs/2401. 05952.
[27] 李志锴,张骁. 人工智能生成内容(AIGC)应用于学位论文写作的法律问题研究[J]. 学位与研究生教育,2024(4):84-93.
[28] Dis E M, Bollen J, Zuidema W, et al. ChatGPT: five priorities for research[J]. Nature,2023,614:224-226.
[29] 孙平. 高校规范人工智能使用的法治化路径[J]. 法学,2025(8):35-54.
[30] 车骋. 学位争议多元解决机制的实践检视、价值基准与规范完善[J]. 重庆高教研究,2025(5):36-48.
[31] Raj A. Finding the real author with Turnitin AI Detection[EB/OL]. (2023-06-26)[2025-11-15]. https://techwireasia. com/2023/06/turnitin-ai-detection-tackling-the-issue-of-academic-integrity/.
[32] Bylieva D, Lobatyuk V, Tolpygin S, et al. Academic dishonesty prevention in e-learning university system[EB/OL]. (2020-05-18)[2025-11-15]. https://link. springer. com/chapter/10. 1007/978-3-030-45697-9_22.
[33] 王霞,龚向和. AIGC 介入学位论文之学术不端的定性与制度因应[J]. 中国高教研究,2025(6):84-92.
[34] 江俊鹏,冯凌子,解贺嘉,等. 高校 AIGC 学术规范建设与启示研究——基于 100 所世界一流大学的调研[J]. 大学图书情报学刊,2025(4):3-9.
[35] 骆飞,马雨璇. 人工智能生成内容对学术生态的影响与应对——基于 ChatGPT 的讨论与分析[J]. 现代教育技术,2023(6):15-25.
[36] University of Toronto. Guidance on the Appropriate Use of Generative Artificial Intelligence for Graduate Academic Milestones[EB/OL]. (2025-10-17)[2025-10-27]. https://www. sgs. utoronto. ca/about/guidance-on-the-use-of-generative-artificial-intelligence/.

(责任编辑:张海生　杨慷慨　校对:杨慷慨)

Technical Scruting and Jurisprudential Reflections on AIGC Detection in Preventing AI-Generated Thesis Misconduct

Wang Jiarui

(*Law School*, *Shandong University*, *Qingdao* 266237, *China*)

Abstract: *Law of the People's Republic of China on Academic Degrees* stipulates that ghostwriting in degree theses constitutes academic misconduct, warranting the withholding or revocation of degrees. Universities generally regard AI ghostwriting as academic misconduct and rely on AIGC detection results from third-party institutions as the primary benchmark. Technically, the existing AIGC detection tools primarily utilize watermarking techniques, trained classifiers, and zero-shot detection. However, hampered by technical limitations and "AI hallucinations," their current detection capabilities struggle to accurately reflect the actual likelihood of AIGC. From a jurisprudential perspective, the utilization of AI in degree theses requires distinguishing between "instrumental use" and "creative/substantive use." Classifying AI ghostwriting as academic misconduct must satisfy the constraints of the proportionality principle, yet AIGC detection tools cannot clarify these boundaries. Furthermore, due to questionable authenticity and non-compliance with statutory forms of evidence, the probative value of AIGC detection reports is insufficient. Their weak evidentiary strength also makes them inadequate, when used alone, to meet the standard of proof required in administrative enforcement and administrative litigation. In response, legislative bodies and educational administrative departments need to establish a regulatory framework for commercial third-party AIGC detection platforms. This includes implementing an access mechanism based on comprehensive evaluation of their detection principles and capabilities, guiding universities in the rational selection of AIGC tools, and preventing market disorder. Universities should treat third-party AIGC detection conclusions with caution, explicitly defining them for auxiliary use only and requiring sufficient evidence to verify their authenticity. A human-centric, manual verification system must be implemented, empowering academic degree committees to determine whether a thesis constitutes AI ghostwriting as academic misconduct. Additionally, it is crucial to build an integrity mechanism for the academic use of AIGC, implementing a dual strategy of passive inspection and actively guiding students to disclose AI use, thereby educating them on the correct and appropriate utilization of artificial intelligence tools.

Key words: AIGC; *Law of the People's Republic of China on Academic Degrees*; academic misconduct; ghostwriting