

文章编号:1003-2053(2026)07-1413-09

生成式人工智能的双轨标识监管与制度完善

张奔,李孟琦

(华中科技大学法学院,湖北武汉 430074)

摘要:生成式人工智能作为新质生产力的典型代表,在重塑产业格局、激发创新潜能的同时,衍生出深度伪造泛滥、虚假信息传播等治理困境。面对技术快速迭代引发的系统性风险,而传统监管模式难以有效回应风险社会的治理需求。为破解监管困局,国家新近发布《人工智能生成合成内容标识办法》及配套强制性国家标准,二者构建出一套全新双轨标识监管体系。通过剖析双轨标识制度作为生成式人工智能监管路径的可行性与挑战性发现,显隐标识的协同设计可形成生成端嵌入、传播端核验与使用端追溯的全链条监管闭环。然而,该制度在主体义务配置、元数据技术标准等方面仍可存可操作性瓶颈。为破解制度落地障碍,通过构建差异化义务配置体系和建立元数据标准协同监管机制,双轨标识制度能够有效破解监管落地障碍,在风险防控与产业激励之间形成动态平衡。

关键词:生成式人工智能;双轨标识;技术治理;元数据

中图分类号:D923.4; TP18

文献标识码:A

生成式人工智能近年来在产业规模、资本集聚以及生态构建等领域实现跨越式发展,以 DeepSeek、文心一言 X1 等为代表的第四代系统,凭借数据巨量化、内容创造力、跨模态融合和认知交互力的技术特征^[1],在重构产业价值链、催生新质生产力等方面展现出“头雁”效应。相较于前代模型,新一代系统通过构建垂直认知网络实现复杂任务并行处理,支持涵盖文本、图像、音频等的多模态高效交互^[2]及跨模态代码生成,催生出新型的内容生产范式。为回应技术迭变,国家于 2023 年出台《生成式人工智能服务管理暂行办法》,明晰服务与技术要求及风险治理责任。党的二十届三中全会确立高质量发展与高水平安全良性互动的基本原则。在政策与技术的双轮驱动下,新一代生成式 AI 平台形成算力领先、场景多元的竞争优势,正如为科研、政务、医疗等领域提供“数字能源”,拥有难以估量的发展前景。

技术跃迁在重塑产业生态的同时,正对既有监管体系带来多维冲击。新一代人工智能技术的多模态生成能力衍生隐私泄露、算法歧视等新型风险,算法黑箱特性导致决策逻辑可解释性断层,加剧技术

取证与责任认定困境。作为涉 AI 案件前沿阵地的北京市海淀区人民法院发布的审判白皮书显示,去年擅自使用相关 AI 标识进行恶意引流等新型案件频发,相较于以往案件存在调查取证难、审理难度大的特点。上述新型技术特性与监管能力失衡折射出当前的主要矛盾:过程性技术监管工具缺位使得 AIGC 的生成状况无从查证,版权争议、信息茧房、算法权力滥用等风险相互交织。在此背景下,国家颁布《人工智能生成合成内容标识办法》(以下简称《标识办法》)及配套的强制性国家标准。办法创新性地构建显式标识与隐式标识相结合的双轨标识制度体系:前者要求服务提供者和用户通过标注“AI 生成”等字样实现人工智能生成内容属性的可视化,后者借助数字水印嵌入生成时间戳、数据溯源路径等元数据,以便实现溯源监管。此外,办法还明确相关主体的标识责任义务,以规范内容制作、传播各环节的标识行为。然而,作为全球 AI 监管领域内的新事物,双轨标识制度面临落地的现实难题。比如,双轨标识制度作为当下 AIGC 的监管路径是否可行;如何使标识义务与责任的配置与技术能力及场景相适配;元数据文件的监管效能如何充分释放等。

收稿日期:2025-06-10;修回日期:2025-08-14

基金项目:科学技术部科技创新 2030—“新一代人工智能”重大项目(2022ZD0120100)

作者简介:张奔(1990-),男,副教授。李孟琦(2001-),男,硕士研究生,通讯作者,E-mail:m202475135@hust.edu.cn。

这些问题的破解,直接关乎生成式 AI 的双轨标识监管体系落地实践之效能。本文以解构双轨标识制度的制度逻辑与实践困境为基础,旨在探索其作为技术适配型监管路径的完善方案,为人工智能时代的有序监管与产业创新提供支撑。

1 双轨标识制度的规范适配性:理论争议与实践证成

1.1 理论之辩:双轨标识监管的学理分野

生成式人工智能的逐渐派生出嵌套性内在风险与层级性外部风险相交织的复杂局面。面对这一态势,传统单一型治理机制已难以有效遏制新型风险,监管转型的必要性已成学界共识,但具体路径选择仍存观点交锋。双轨标识制度作为技术赋能监管的革新方案,其正当性与实效性引发激烈争鸣。因而,围绕双轨标识制度的监管逻辑与实施障碍展开体系性梳理、在理论上检视其作为生成式人工智能监管路径的规范适配性,成为学界亟待深入探讨的课题。

学界支持观点主张标识制度通过生成来源与真伪判断的双轨制治理逻辑,能够提升虚假信息鉴别的可能性^[3],有效规避风险^[4],并实现复合治理效能。一则,显式标识通过强制披露机制保障信息透明度,为公众提供算法信任的认知基础。事实上,监管效能之有效发挥核心依赖于信息披露之透明。虽然显式标识无法提供有关的详尽信息,但给各端用户亲近算法制造契机。司法实践中对此亦予以支持:杭州“AI 绘画平台帮助侵权案”中,法院要求生成式人工智能平台需承担合理注意和用户提醒义务;北京互联网法院在“AI 文生图”案的判决书中亦对生成式人工智能平台与用户的标识义务进行阐释。这凸显出显式标识在实践中的重要价值。另则,隐式标识则通过嵌入数字水印等技术手段构建完整链条,为国家机关和平台的监管赋能。当 AIGC 未通过显性标识披露其属性,人工智能生成内容与人类作品的区分就需依靠技术检测手段^[5]。隐式标识恰恰利用数字水印等防篡改手段,为无显式标识之内容提供溯源监管的技术支撑。相较于传统事后审查模式,双轨标识制度实现监管关口前移,在 AI 换脸、医疗辅助检查等高风险内容引发争议时,隐式标识中内嵌的元数据信息可成为执法监管与司法裁判的关键证据。双轨标识制度的生成端嵌入式机制可将算法决策转化为可追溯的链条,在生成对

抗网络的概率性输出与确定性责任之间建立映射关系,形成融合“法治”与“技治”的监管生态系统。

反对观点认为标识制度在实践环节会遭遇技术瓶颈、突破个人信息保护原则与抑制产业创新。首先,标注工具技术不成熟与标注质量异质性动摇双轨标识制度的效能。现有技术难以实现精准标识,尤其是对特定类型的内容和特殊应用场景下的水印鲁棒性、隐蔽性和安全性等方面^[6]仍存技术瓶颈,易引发误标或漏标风险与数据安全隐患。此外,《标识办法》要求对 AIGC 进行标识,并规定服务提供者应制定清晰且操作性强的标识规则与端口。然而,以 GPT 系列大模型为例,第一代 GPT 的语料库为 5GB,而新一代 GPT 的预训练数据已超过 100TB,参数规模亦从 1.17 亿升至 1.8 万亿。在巨大的语料库与参数规模下,标识义务主体会为形式合规而降低标识质量,使得双轨标识制度沦为纸上谈兵的工具。其次,标识制度中留存日志的规定会冲击个人信息保护的法律边界。《标识办法》第九条实质上是对服务提供者在部分情况下留存用户信息进行的合法豁免,这与《个人信息保护法》确立的目的限制原则及信息处理者的利益不应以让渡个人信息权益为代价^[7]的宗旨相悖。再次,强制性国家标准易演化为掣肘产业创新的技术性壁垒。有学者指出,根据《标准化法》第十条规定,仅有涉及“保障人身健康和生命财产安全、国家安全、生态环境安全”的要求方可制定强制性国家标准,而在人工智能国家标准中,涉及安全的标准只有少数^[8],大多只满足推荐性标准的要求。但双轨标识制度配套的标准为强制性国家标准,其设定范围与《标准化法》的规定与鼓励探索性、前沿性技术研究的立法导向存在出入,甚至会陷入合规优先于创新的局面。

双轨标识制度是对生成式人工智能技术风险的制度回应,当下平衡产业创新与监管合规已成为全球共识。在全球生成式人工智能监管路径未有更优解的背景下,双轨标识制度正是这一理念的具象化表达。尽管其实施面临技术瓶颈与产业抑制等挑战。但相较于传统以“安全阀”为导向的单维监管框架,双轨标识制度的核心价值在于通过设定标识义务这一制度杠杆,倒逼技术开发者和提供者完善风险评估技术与监管举措,同时为监管机构优化规制工具提供实践样本。因此,在当前针对 AIGC 的监管实践尚处于探索阶段的过渡时期,双轨标识制度不失为一种折中的路径选择。

1.2 实践证成:权利公示与技术措施的规则延伸

双轨标识制度的应用依据源于权利公示原则的适应性拓展与《著作权法》中技术措施规则的制度性延伸。该制度融合事前预防与事后追责的双重功能,既有效弥补单一技术监管或行政命令模式的固有局限,又凭借法律规范与技术标准的协同效应,显著降低监管成本,提升治理精度,为生成式人工智能的规模化应用构建可预期、可操作的合规框架。

数智时代下权利公示原则的适应性拓展是双轨标识制度应用的理论依据,推动署名推定主义向技术认定主义的范式转型。权利公示原则意指著作权的权利归属、权利内容、权利变动等信息需通过一定方式向公众公开。我国《著作权法》第十二条明确规定的“署名推定”规则,强调著作权自创作完成时产生,无需经过登记等公示手续。但在人工智能生成物场域下,创作主体的多元性以及算法本身所衍生出的数据枯竭、技术黑箱等场景导致使用者在生成式人工智能平台完成创作后,所得的生成物无法依照先前的规则推定出权利归属状态,从而难以对生成过程与结果进行有效监管。当下,数字水印、区块链存证等技术加速进步,使得在AIGC中嵌入显隐标识相较之前已非难事。在先进标识技术使得公示成本下降,尤其是公示的功能优势大于其成本劣势时^[9],权利公示原则的意涵就不应拘泥于传统的署名推定,而该要求权利状态须以可识别的方式向公众披露,以维护技术使用安全与产业健康发展。双轨标识制度即是对这一演进的回应,其通过显式标识强制在生成物表面构建可视化的权利公示界面,借助隐式标识将“AIGC”等元数据字段内嵌其中,以此形成权利推定的初步证明效力。总而言之,双轨标识制度通过技术痕迹留存,使权利公示原则突破形式窠臼,推动著作权在生成式人工智能领域由“署名推定主义”向“技术认定主义”转变。

著作权法下技术措施的规范延伸赋予双轨标识制度应用的制度依据,实现权利管理信息与技术措施的功能性整合。双轨标识本身可视为权利管理信息的数字化延伸,其核心功能在于通过信息透明化降低交易成本与权属争议风险。但由于权利管理信息仅是一组信息合集,没有控制获取、复制作品的技术功能^[10],因而无法充分发挥监管效用。然而,强制性国家标准则助推双轨标识以单纯的信息标识功能,具备著作权法中技术措施的部分特质。其中,隐式标识通过难以篡改且可溯源的技术手段确保生成

内容完整性,构成防止AIGC被不法利用的技术屏障。这种制度设计创新性地打破权利管理信息与技术措施的固有规范界限,相较于DRM等传统技术措施对作品利用的排他性限制,双轨标识措施形成全流程、多模态的复合标注体系,并可与生成内容非可视化共存,降低对其正常使用的干扰。更为重要的是,双轨标识制度通过设置清晰的法律责任,实现监管闭环。《标识办法》第八至九条厘定各主体的法定义务与责任后果,与《著作权法》的责任体系之间完成有机衔接,即当服务提供者、用户等主体未依规履行标识义务或篡改标识时,将触发《著作权法》第五十三条设定的违法责任条款,分情况承担不同的民事责任或面临行政处罚,以此形成标识即合规、失范即担责的强效约束机制。

综上,双轨标识制度构筑起兼具可操作性与前瞻性的技术中立型监管框架。一方面,该制度突破传统著作权法“署名推定”规则的局限性,以权利公示原则的适应性拓展为AIGC监管确立清晰的责任边界。另一方面,双轨标识实质是权利管理信息的数字化升级与技术措施的创造性整合,显隐标识协同形成信息管理与技术防控的复合治理模式。因此,双轨标识制度以技术痕迹留存破解监管难题、以法律责任闭环保障制度刚性的创新设计,是监管人工智能生成内容的有效解决之道,并推动生成式人工智能从“野蛮生长”迈向“负责任创新”。

2 双轨标识制度的现实挑战:义务配置失衡与标准协同困境

2.1 主体义务设定的结构性错配

双轨标识制度通过技术治理与法律规制的协同而具备可行性,然而其落地实施却需直面全产业链主体之间的权责适配困境。当前《标识办法》虽初步框定服务提供者与用户的标识义务,但其未能基于产业链内不同主体之间的角色分野和风险控制能力进行差异化的义务与责任配置。其一,大模型开发者作为技术源头主体,其隐式标识的架构能力构成下游合规的基础要件;其二,终端用户滥用标识豁免权的行为缺乏有效的责任约束机制。这种技术控制力与合规责任的倒挂模式不仅难以适配生成式人工智能产业链的分工特性,更易引发责任链条断裂与合规激励失衡,最终成为钳制双轨标识制度治理实效的深层障碍。

基础模型开发者的权责错位削弱技术源头治理效能。开发者作为生成式人工智能技术的底层架构者,其隐式标识的技术实现能力直接决定下游内容的可溯源性。从技术实现逻辑看,头部企业等开发者凭借其雄厚的技术积累和数据优势^[11],若其未在模型底层架构中预留数字水印接口,后续服务平台即便强制添加显式标识,也可能因技术兼容性问题导致隐式标识在内容传播链中失效或被轻易剥离。然而面对这一问题,《标识办法》等相关法规对开发者仅提出原则性要求,如提供技术接口支持等,未明确规定其承担更高标准的标识技术嵌入义务。技术源头的标识义务缺位使得《标识办法》第五条中提及的“服务提供者应确保隐式标识完整嵌入”的合规要求沦为无源之水。反观欧盟《人工智能法》将人工智能的风险分为四个梯度,基于不同等级风险采取不同的监管要求,体现出分级分类监管的精髓,可为我国标识义务配置提供参照。开发者作为技术主导者,理应承担更高标准的源头标识义务,例如在模型架构中预置标准化元数据接口、构建全流程标识完整性校验机制等,以便在节省监管资源的同时最大程度避免风险的产生和扩散。

服务平台的合规压力过载抑制技术普惠应用。《标识办法》将标识义务过多集中于服务平台提供者,要求其对生成内容进行可视化标注,并通过元数据核验机制确保隐式标识完整性。但平台的技术控制力受限于上游开发者提供的技术接口。若往往只能依赖外挂工具等方式进行事后补救,极易被用户对内容的二次加工来规避。因而,算法、元数据等技术壁垒从根本上决定标识义务的设定需以服务提供者可控范围为限^[12],要求服务提供者对生成内容的技术原理、决策过程等内容负有解释与标识义务,既超出其技术能力范畴,亦有违“风险与能力相匹配”的公平监管原则,在具体实践中必遭落空。除此之外更值得警惕的是,平台承担的过度标识义务可能异化为抑制中小企业创新的壁垒。强制性国家标准 GB 45438-2025 会迫使中小平台在技术研发资源与高昂的合规成本间艰难权衡。以 AI 医疗影像肿瘤辅助诊断系统为例,平台需对每版病理切片数据集标注来源、模型版本及医生信息,此类高频次、高精度的标识要求增加数据存储与核验成本,中小型服务平台难以负担高额支出,从而阻碍技术的普惠性推广。

终端用户义务的场景化适配缺失加剧监管效能

衰减。《标识办法》虽规定用户传播生成内容时需主动标注 AI 属性,但未针对使用场景的风险等级设计分层义务。服务提供者将内容交付用户后,生成内容就处于终端用户控制之下。普通用户利用 AI 生成社交媒体配图或剪辑音视频等场景下,其因技术认知不足或个人疏忽等因素而导致的漏标、错标等现象之风险终归有限。而诸如金融服务公司、律师事务所等专业机构使用 AI 生成投资分析或法律服务报告时,恶意去除标识的行为则极易引发严重的版权纠纷、误导性信息传播等风险。若对普通用户与专业机构配置同等义务,既不符合比例原则,亦导致监管同质化引发执法成本畸高。此外,用户标识申请去除机制存在显著规范漏洞。《标识办法》第十条允许用户申请去除标识,仅要求其通过协议承诺责任自负。当用户将去除标识的 AI 生成虚假信息通过匿名方式向在线平台共享其数据时^[13],平台既无技术能力追踪传播路径,亦难以依据单薄的“协议承诺”条款有效追责。上述漏洞暴露出现行制度对用户行为复杂性与规避手段的应对不足,极大制约双轨标识制度在终端应用层的效能发挥。

2.2 元数据标准的协同性障碍

元数据是全过程描述文件的背景、内容、结构及其管理的数据,其在数码环境中拥有唯一代码,故而在 AIGC 的监管框架中发挥不可替代的数字指纹作用。元数据的监管效能集中体现为通过技术手段实现生成内容的可溯源性与标识体系的协调性,实现 AIGC 的全生命周期监管,有效防范内容篡改与伪造风险。然而,当前制度在元数据技术标准协同、技术迭代与更新机制等方面仍面临多重实践挑战。

多模态元数据格式的互异性割裂监管技术接口,降低跨平台溯源的实践效率。根据相关法规规定,元数据应具备跨平台、跨场景的多模态标识能力,以便在监管环节准确溯源。但《标识办法》及配套国家标准 GB 45438-2025 虽对元数据字段作出原则性规定,未能解决多模态内容元数据格式的兼容难题。以人工智能生成物中最为常见的图片与音视频文件为例,MP3 文件的 ID3 标签采用帧标识符命名,支持用户自定义帧,且其位置并不固定。而 JPEG 文件格式具有高压缩率的特点,难以嵌入大量额外信息,否则会严重失真^[14]。故其以 IFD 分层存储元数据,通过标签 ID 映射具体值。由此可见,二者在命名规则、数据结构、存储容量限制等方面存在显著差异。在对 MP3 文件和 JPEG 文件的元数据进

行解析时,若生成内容采用与其通常使用格式以外的其他方式存储,且其元数据字段未被强制性国家标准所覆盖,监管机构或平台需针对新的格式另行开发专用解析工具。不同文件格式间碎片化、互不兼容的元数据字段,不仅使得监管机构难以通过统一技术接口高效地提取关键标识信息,延误最佳取证时机,也会导致平台因解析成本过高而怠于履行元数据核验义务,最终严重削弱双轨标识制度的监管效率。

私有化与开放性元数据系统间的互斥引发解析困境,阻滞产业链协同治理的闭环运行。生成式人工智能产业链涉及多方主体,各环节采用的元数据管理系统存在显著差异,此特征不利于下游平台和监管机构采取集约型范式进行管理。例如,百度的文心大模型采用私有化部署的元数据加密协议,而阿里巴巴的通义千问平台则遵循 W3C 的开放标注标准。这两种不同的元数据管理方式,在数据字段定义与校验规则上呈现出互不兼容的状态。文心大模型的私有化加密协议,采用自定义加密字段,结合数据脱敏、匿名化技术等方式,可定义特定加密标识符,字段命名与格式封闭度高。而通义千问平台遵循的 W3C 标注标准使用通用字段和开放格式,允许字段扩展,能促进不同场域下的数据集交换、确定数据来源^[15]。这种差异导致在内容传播过程中,当使用文心大模型生成的内容被传输到遵循通义千问平台元数据标准的系统时,元数据信息会出现无法被正确解析和识别的情况,甚至被误判为无效或恶意数据。通义千问平台生成的内容亦面临相同困境。这种互操作性缺失不仅使得元数据信息在跨系统内容传播过程中易遭篡改或剥离,破坏内容的完整性与可信度,更给内容创作者的确权和监管者的溯源取证带来巨大困扰,实质性地割裂产业链协同治理的闭环。

新兴元数据格式的迭代超前于标准更新速率,技术脱节催生跨模态关联监管的真空地带。当下, N - TRIPLE、Parquet/ORC 列式存储等新型元数据表示和存储格式层出叠见,其元数据结构远超传统标识的认知范畴。这些新的元数据格式通常为特定系统而开发,不具有普适性,难以实现大规模的推广^[16],导致现行强制性国家标准附录 E 中规定的元数据字段及其内嵌方式,难以完整映射和适应这些前沿技术特征。例如,2025 年 3 月,潞晨科技推出的 SOTA 视频生成模型 Open - Sora 2.0,通过三维

自编码器与分层数据处理,实现视频的潜空间压缩格式,又可保留关键动态特征,但现行标识标准的元数据规定在适配潜空间格式时存在技术壁垒,反映出技术与监管标准的脱节。不仅如此,跨模态元数据关联样态亦存标准真空。生成式人工智能技术的飞速进步还催生支持文本、图像、表格、音频等多元数据的关联格式,通过压缩与列式存储实现跨模态适用。但现行标准对此类跨模态技术特征的支持明显不足。当元数据标准严重滞后于技术复杂性时,双轨标识制度所依赖的“显隐互补”机制将不可避免地退化为相互割裂的监管孤岛,难以有效应对未来全球化、多模态、高动态的内容传播生态。

3 双轨标识制度的向善路径:分层治理与动态适配

3.1 全产业链分层义务体系的梯度构建

破解双轨标识制度主体义务失衡困境的关键,在于立足产业链的垂直分工特性与风险源流分布,构建以技术控制力、风险等级与义务强度相匹配为核心的分层责任配置框架,将位于产业链各环节可能对该技术运行与建构产生差异性影响的主体均纳入治理网络,形成从源头防控到中游管控,再到终端约束的梯度治理闭环,实现权责适配与创新激励的动态平衡。

强化基础模型开发者的标识嵌入义务,构建风险预防的源头屏障。开发者作为底层算法的架构主体,应承担预置隐式标识接口与保障标识完整性的复合义务。具体而言,其一,需强制要求开发者在模型设计阶段对其开发的人工智能系统内部构造、推理逻辑、技术接口、输出结果提供明确的、可理解的说明^[17]。其二,在模型架构设计阶段预置兼容多模态数据的隐式标识接口,确保生成内容在像素级特征编码与结构化元数据层面均具备可溯源技术支撑。其三,构建全流程标识完整性保障机制,通过在内容输出端部署自动化检测模块,运用数字水印校验技术防范下游调用环节的标识信息篡改风险。此外,开发者应积极探索以“AI 治理 AI”的智能监管路径,通过人工智能建模技术仿真化的算法运行环境,探索沙盒治理的中国模式,并且重点关注生成式人工智能算法黑箱和溯源困难等监管难题,开发可供应用的技术措施,以源头治理回应监管难题。

优化服务提供者的标识核验义务,平衡合规压

力与技术普惠。服务平台作为内容生成与传播的中枢,义务重心应聚焦于显式标识的可靠性保障与防篡改技术部署,结合自身能力开展必要监管活动。其一,要求服务提供者在生成内容呈现的初始界面嵌入动态可视化标识组件,并采用哈希算法绑定内容哈希值与标识信息。其二,引入《电子商务法》第四十二条之“通知—删除”规则,要求平台在接到侵权通知后,在有初步证据的前提下对问题内容采取临时下架措施,当用户或提供者提供相反证据证实不存在侵权行为时,服务提供者可以恢复相关内容。这既是对权利人发出通知的附加义务^[18],又使平台积极发挥平台“看门人”的监督作用。其三,需通过清单化管理明确平台的安全保障义务,例如定期抽检标识完整性、建立元数据核验日志等,以倒逼平台完善技术审核体系。此类措施可避免将技术源头责任转嫁至中游平台,符合“风险与能力相匹配”的监管原则。

终端应用层需厘清用户责任,细化终端用户的场景化标识义务。用户是生成式人工智能的末端接触者和利益相关人,应明确规定使用者的显性披露和不破坏隐性标识的义务,并规定相应法律责任^[19]。具体来说,一般用户对于生成式人工智能的工具性使用无需配置较高的标识义务,仅要求其遵守平台用户协议,在传播时进行规范显式标识,如勾选标识选项、使用平台提供的标识工具等。但对于部分使用 AIGC 的专业用户,如利用生成式人工智能大模型进行法律咨询服务、金融数据分析等,应肩负核验元数据文件的责任,并通过显式标识对生成内容的 AI 算法、溯源信息等承担强制义务。此外,应针对《标识办法》第九条关于用户申请去除标识的规定,建立用户标识豁免申请的实质性审查机制,包括要求专业用户提交使用场景说明与责任承诺书,并由平台通过人工审核验证其申请理由的真实性与使用场景的合法性,对于审查不通过或理由存疑者,不得去除标识。有义务配置必有责任后果,当用户故意绕过技术措施实施虚假披露时,应在特定高风险场景下突破传统过错责任原则,由使用者就其标识豁免行为的合法性及无过错承担举证责任,切实贯彻过责相当的基本法理,解决追责难问题。

3.2 元数据标准动态适配的规范化设计

元数据作为双轨标识制度的技术载体,其标准化和协同性程度直接决定监管效能释放的充分性。当前元数据领域存在的格式互异、系统割裂与标准

滞后等问题,亟需通过建立具备规范统一、动态更新和法律衔接的多元一体协同治理框架予以破解。该框架以跨平台兼容性为基础,以技术迭代响应机制为支撑,以鲁棒性验证为保障,以法律赋能为闭环,旨在构建多模态的元数据标识监管适配体系。

统一跨模态元数据规范,破解格式互异导致的监管低效困局。当前,不同文件格式在元数据命名规则与数据结构上的差异导致监管机构难以通过统一接口提取标识信息。对此,应依托全国信息技术标准化技术委员会,联合中国电子技术标准化研究院、行业头部企业及科研机构,共同制定生成式人工智能统一的数据格式、编码规范、元数据标准等,确保数据的兼容性和可理解性。该规范需明确以下核心要素:第一,统一字段命名规则,规定生成主体、模型版本、数据来源链、时间戳等必填字段的标准化命名;第二,采用通用元数据格式,通过语义化标签实现结构化与非结构化数据的互操作,加快元数据共享、元数据销毁和元数据备案等制度^[20];第三,兼容国际标准,确保跨境内容流通的合规性。以音视频文件为例,可要求所有生成式 AI 平台统一采用“AI_Content ID”作为核心标识符,包含 Base64 编码的生成主体数字证书哈希值及区块链存证索引,从而突破格式壁垒。此举不仅能降低监管机构的解析成本,还可避免企业因重复开发适配工具而产生资源浪费。

建立标准动态响应机制,弥合技术迭代与制度更新的时滞鸿沟。生成式人工智能技术的快速迭代催生 Parquet 列式存储等新型格式,而现行国家标准附录 E 的元数据字段无法完全覆盖此类技术特征。建议由国家网信办牵头成立生成式人工智能元数据治理专班,构建以技术监测和标准修订为一体的响应机制。专班的核心职责在于建立技术动态监测平台,实时追踪新兴文件格式的市场占有率。当检测到某格式的市场占有率超过预设阈值时,自动触发元数据标准修订程序。另外,在标准修订中引入沙盒机制,也即企业可以在“沙盒”中测试创新产品、服务,并享有特定情况下的责任豁免待遇。具体可在《标识办法》实施指南中明确通过对抗性训练和压力测试评估标识在异常输入环境下的状态,并授权国家工业信息安全发展研究中心等机构开展年度飞行检查^[21],即事先不通知被检查主体而直接实施的检查行为,避免其提前准备或掩盖问题,从而真实地反映实际情况,形成监管威慑。并同步启动《网

络安全法》《数据安全法》等配套法规的修订评估,确保技术标准与法律规则协同演进,以避免《标识办法》第五条因标准僵化导致的监管失灵。

强化元数据标识与技术措施规则互嵌,实现元数据治理的闭环效能转化。技术措施作为版权人控制作品接触与使用的法定屏障,元数据恰可通过结构化编码系统完整记载生成物中各关键要素,保障技术措施功能的发挥。具体可采取规范解释路径,加强元数据技术适用与著作权法律体系的衔接。其一,运用法律拟制在下次《著作权法实施条例》修订时将有效技术措施扩展解释为包含符合国家标准的元数据标识系统,使技术标准获得法律规范效力。其二,在操作层面建立多元一体的实施机制,要求网络服务提供者在内容生产环节强制嵌入包含著作权登记号、创作时间戳及权利人生物特征等要素的元数据标签,并通过数据控制层和中间服务层^[22]区块链存证技术实现与著作权登记系统的实时核验。提升标识公示的公信力。同时,应将系统性破坏元数据完整性的行为纳入《著作权法》第五十三条规制的“故意避开或破坏技术措施”之范畴,给技术标准配置民事救济的请求权基础。通过上述法律解释等方法可构建技术标准与法律救济联动的闭环体系,从而破解数字版权领域存在的监管难、举证难、维权难的三重困境。

4 结语

双轨标识制度是我国应对生成式人工智能风险挑战、平衡创新发展与安全可控的关键制度创新,构筑起一套技术适配型的监管方案。未来的政策演进应着重于三方面:其一,建立跨部门协同监管机制,由国家网信办牵头组建跨部门技术标准委员会,使其在元数据标准制定、技术验证等方面形成监管合力,推动标准修订与技术迭代同步演进;其二,加快监管范式转型,依托国家级沙盒试验场探索高风险场景的分级分类监管,在医疗、金融等高风险领域开展分级分类监管试点,以兼顾创新容错性与风险防控性;其三,深化国际监管合作,主动参与 ISO/IEC JTC1 国际标准组织的技术规则制定,提升我国在全球 AI 治理中的话语权。并设立人工智能标识技术创新基金,支持数字水印、区块链存证等核心技术研发,形成技术赋能监管和监管反哺创新的良性生态,助力实现高质量发展与高水平安全良性互动的战略

目标。

参考文献:

- [1] 杨张博,韩淑君,孙笑明,等.中美 AIGC 领先企业技术创新布局比较研究[J].科学学研究,2025,43(2):265-277. Yang Z B, Han S J, Sun X M, et al. A comparative study of patent technological innovation layout of leading AI-generated content firms in China and the U. S. [J]. Studies in Science of Science, 2025, 43(2): 265-277.
- [2] 陈智.新一代生成式人工智能的价值定位、演进逻辑及前瞻路径——以 DeepSeek 为例[J].新疆社会科学,2025(2):17-29,170. Chen Z. Value orientation, evolutionary logic, and forward-looking pathways of new-generation generative artificial intelligence——A case study of DeepSeek[J]. Social Sciences in Xinjiang, 2025(2): 17-29, 170.
- [3] 漆晨航.生成式人工智能的虚假信息风险特征及其治理路径[J].情报理论与实践,2024,47(3):112-120. Qi C H. Research on the risks of disinformation from generative artificial intelligence and its governance paths[J]. Information Studies: Theory & Application, 2024, 47(3): 112-120.
- [4] 邓建鹏,赵治松. DeepSeek 的破局与变局:论生成式人工智能的监管方向[J].新疆师范大学学报(哲学社会科学版),2025,46(4):99-108. Deng J P, Zhao Z S. The breakthrough and transformation of DeepSeek: the new rules for AIGC regulation[J]. Journal of Xinjiang Normal University (Edition of Philosophy and Social Sciences), 2025, 46(4): 99-108.
- [5] 张继红.生成式人工智能生成内容标识义务研究[J].法商研究,2024,41(4):187-200. Zhang J H. Research on the obligation of content identification in generative artificial intelligence[J]. Studies in Law and Business, 2024, 41(4): 187-200.
- [6] 朱红儒,孙巍巍,孙勇,等.基于数字水印技术的生成式人工智能标识研究[J].中国信息安全,2023(11):67-70. Zhu H R, Sun W W, Sun Y, et al. Research on generative artificial intelligence identification based on digital watermarking technology[J]. China Information Security, 2023(11): 67-70.
- [7] 王苑.目的限制原则的反思及其解释论构造——以《个人信息保护法》第6条第1款为中心[J].华东政法大学学报,2024,27(4):44-53. Wang Y. Reflection on the principle of purpose limitation and its interpretive construction: Centered on Article 6 (1) of the Personal Information Protection Law[J]. ECUPL Journal, 2024,

- 27(4):44-53.
- [8] 宋华琳. 人工智能立法中的规制结构设计[J]. 华东政法大学学报, 2024, 27(5):6-20. Song H L. Regulatory structure design in artificial intelligence legislation[J]. ECUPJL Journal, 2024, 27(5):6-20.
- [9] 张颖. 数字时代著作权公示失灵及其缓解路径[J]. 出版发行研究, 2023(8):64-71. Zhang Y. The failure of copyright publicity in the digital age and its alleviation path[J]. Publishing Research, 2023(8):64-71.
- [10] 华劼. 数字出版视域下的反规避技术措施规则延伸性研究[J]. 科技与出版, 2017(11):87-91. Hua J. Research on the extension of anti-circumvention technology measures rules from the perspective of digital publishing[J]. Science-Technology & Publication, 2017(11):87-91.
- [11] 张欣. 人工智能治理的全球变革与中国路径[J]. 华东政法大学学报, 2025, 28(1):18-32. Zhang X. The global transformation of artificial intelligence governance and China's path[J]. ECUPJL Journal, 2025, 28(1):18-32.
- [12] 赵精武. 论交互对话式人工智能服务提供者的注意义务[J]. 法律适用, 2025(4):119-134. Zhao J W. On the attention obligations of interactive dialogue AI service providers[J]. Journal of Law Application, 2025(4):119-134.
- [13] Acemoglu D, Makhdoui A, Malekian A, et al. Too much data: Prices and inefficiencies in data markets[J]. American Economic Journal: Microeconomics, 2022, 14(4):218-256.
- [14] 文泓鑫, 胡飞, 俞能海. 适应性 JPEG 图像隐私保护方案[J]. 网络与信息安全学报, 2024, 10(1):91-101. Wen H X, Hu F, Yu N H. Adaptable privacy protection scheme in JPEG images[J]. Chinese Journal of Network and Information Security, 2024, 10(1):91-101.
- [15] Albertoni R, Browning D, Cox S, et al. The W3C data catalog vocabulary, version 2: rationale, design principles, and uptake[J]. Data Intelligence, 2024, 6(2):457-487.
- [16] 陈金菊, 欧石燕. 国外元数据注册系统调研与启示[J]. 图书馆学研究, 2020(1):46-54, 82. Chen J J, Ou S Y. Research and insights on foreign metadata registration systems[J]. Research on Library Science, 2020(1):46-54, 82.
- [17] 张其仔, 李伟. 我国建立人工智能安全监管制度的必要性及现实路径[J]. 北京师范大学学报(社会科学版), 2025(1):5-13. Zhang Q Z, Li W. Establishing an artificial intelligence security regulator system in China: Necessity and practical paths[J]. Journal of Beijing Normal University (Social Sciences), 2025(1):5-13.
- [18] 孔祥俊. “互联网条款”对于新类型网络服务的适用问题——从“通知删除”到“通知加采取必要措施”[J]. 政法论丛, 2020(1):52-66. Kong X J. Application of “internet article” to new types of network service——From “notice-and-takedown” rule to “notice and taking necessary measures”[J]. Journal of Political Science and Law, 2020(1):52-66.
- [19] 姚志伟, 李卓霖. 生成式人工智能内容风险的法律规制[J]. 西安交通大学学报(社会科学版), 2023, 43(5):147-160. Yao Z W, Li Z L. Legal regulations on the risks of generative artificial intelligence content[J]. Journal of Xi'an Jiaotong University (Social Sciences), 2023, 43(5):147-160.
- [20] 任梅, 代栋栋. 生成式人工智能的回应性监管架构及其实现路径[J]. 中国行政管理, 2025, 41(2):39-50. Ren M, Dai D D. Responsive regulatory architecture for generative artificial intelligence and its realization path[J]. Chinese Public Administration, 2025, 41(2):39-50.
- [21] 熊丙万, 何娟. 论数据要素市场的基础制度体系[J]. 学术月刊, 2024, 56(1):102-114. Xiong B W, He J. The system of basic legal regimes for data market[J]. Academic Monthly, 2024, 56(1):102-114.
- [22] 包晓丽. 可信数据空间: 技术与制度二元共治[J]. 浙江学刊, 2024(1):89-100, 239-240. Bao X L. Trusted data space: A co-governance framework of technology and legal rules[J]. Zhejiang Academic Journal, 2024(1):89-100, 239-240.

Dual – track labeling regulation and institutional refinement
for generative artificial intelligence

ZHANG Ben, LI Meng – qi

(School of Law, Huazhong University of Science of Technology, Wuhan 430070, China)

Abstract: Generative Artificial Intelligence (GAI), while driving industrial transformation and innovation, has introduced unprecedented governance challenges including deepfakes, algorithmic bias, and provenance gaps. In response, China’s *Labeling Method for Content Generated by Artificial Intelligence* establishes a dual – track regulatory system combining explicit visual identifiers and tamper – resistant implicit watermarks embedding metadata like timestamps and data trails. This framework aims to create an end – to – end oversight chain—embedding identifiers at generation, enabling verification during dissemination, and ensuring traceability upon consumption—adapting copyright law’s publicity principles to shift from "signature presumption" to "technological determination" for AI – generated content. However, implementation faces critical barriers; obligations are structurally misaligned across stakeholders, with model developers lacking mandates to architect foundational watermarking interfaces, service platforms overburdened by metadata validation duties that strain SMEs and end – users operating without risk – tiered responsibilities. Further complicating enforcement, fragmented metadata formats, proprietary/open – standard conflicts, and outdated national standards create interoperability voids that hinder cross – platform verification, particularly as emerging technologies like latent – space video compression outpace regulatory updates. To address these dual challenges of obligation imbalance and technical fragmentation, a tiered governance approach is essential. Upstream, developers must be mandated to preset multimodal implicit labeling interfaces and deploy automated integrity checks, establishing traceability at the architectural source. Midstream obligations should align with platform capabilities, focusing on reliable visible tagging and “notice – and – takedown” mechanisms akin to e – commerce governance, avoiding undue technical burdens. Downstream, user duties require differentiation—minimal labeling for general content creation versus stringent metadata verification for professional outputs like legal analyses, coupled with substantive review for identifier – removal requests to prevent abuse. Concurrently, metadata standardization demands dynamic co – regulation; unifying cross – modal identifiers through universal “AI_Content_ID” fields anchored to block-chain – certified provenance records; establishing agile standard – revision protocols using regulatory sandboxes and unannounced inspections to keep pace with innovations like Parquet columnar storage; and legally recognizing compliant metadata systems as copyright – enforceable “technological measures” to close enforcement gaps. Future policy must prioritize inter – departmental coordination for cross – sectoral standard harmonization, risk – calibrated regulatory pilots in sensitive domains like healthcare, and proactive engagement in global forums such as ISO/IEC JTC1—collectively advancing a governance paradigm that balances innovation incentives with robust risk mitigation, steering GAI from disruptive growth toward responsible deployment.

Key words: generative artificial intelligence; dual – track labeling; technological governance; metadata