生成式人工智能刑事规制：观念转变与路径建构*

张秋芳，刘若愚

（中原工学院 法学院，河南 郑州 450001）

［摘　要］生成式人工智能逐渐从“专用人工智能”转变为“通用人工智能”，其技术的革新和类人属性特征的增强，对现行刑法提出了严峻的挑战。目前，生成式人工智能是否需要刑法对其进行规制以及如何规制的问题存在较大争议。基于生成式人工智能“类人而非人”的基本立场，刑法规制模式应从“数据控制”转向“数据利用”：在立法层面，构建系统化的数据利用规制体系，通过增设非法分析数据罪、完善现有罪名适用场景等举措，实现数据安全的防护与平台责任的实质强化；在司法适用层面，针对算法黑箱的治理难题，建立“数据流-算法流”双轨审查模式，以及“推定因果关系＋举证责任倒置”的特殊认定规则；在合规治理层面，应推动企业健全数据利用合规制度，完善数据共享伦理准则，从源头规范数据利用行为，确保人工智能技术发展与法律规制形成良性互动。

［关键词］生成式人工智能；刑法规制；刑事风险；数据控制；数据利用

［中图分类号］D914　　［文献标志码］A　　［文章编号］1674－5639（2026）01－0118－07

DOI：10.14091/j.cnki.kmxyxb.2026.01.012

随着科技的飞速发展，生成式人工智能已经渗透到社会生活的各个领域。生成式人工智能技术以其强大的数据处理、模式识别和内容生成能力，为人们的生活带来了前所未有的变革。然而，技术革新在推动社会进步的同时，其衍生的刑事风险也日渐凸显。例如，算法偏见可能导致的司法不公、数据隐私泄露引发的刑事责任问题，以及生成内容被恶意利用于犯罪活动等。这些风险问题对传统刑法体系构成了严峻挑战。关于生成式人工智能是否需要刑法进行专门规制的问题，有的学者从法伦理学视角出发，认为生成式人工智能不具备刑事责任能力，从而无法赋予其刑事主体地位，也就不存在其引发新的刑事犯罪类型的问题，可以通过调整主体资格认定标准来应对相对应的风险。[①]也有论者提出，生成式人工智能虽带来数据安全、侵犯知识产权、侵犯公民个人信息、诈骗等刑事风险，但现行刑法中的部分罪名通过扩展或重新定义可适用于其引发的违法模式，所以无需针对生成式人工智能来更新现有的刑法体系。[②]还有观点提出，风险社会“不一定是真实状态，而是文化或者治理的产物”，既不足以撼动规范刑法的体系及其谦抑、保障的性质，也无须开展专门的生成式人工智能抑或信息刑法的研究。[③]然而，生成式人工智能逐渐从“专用人工智能”转变为“通用人工智能”，其所带来的改变势必会对现有的刑法体系造成冲击。鉴于此，本文拟从生成式人工智能刑事风险的现实表征切入，探讨刑法既有规制路径的转型方向，最终提出立法完善、司法适用与合规治理的协同方案，以期应对生成式人工智能带来的刑事风险。

一、生成式人工智能：技术革新背后的犯罪风险与规制必要性

生成式人工智能（Generative Artificial Intelligence）是人工智能领域的前沿分支，其核心在于构建能够自主创造新内容的智能系统。不同于传统人工智能仅专注于数据分析与模式识别，生成式人工智能通过深度神经网络架构学习数据内在规律，模拟人类思维中的联想生成机制，实现从原始输入到逻辑连贯

＊　［作者简介］张秋芳，女，河南汝州人，中原工学院副教授，美国波士顿学院访问学者，研究方向为人工智能与刑法、知识产权法；刘若愚，男，河南南阳人，中原工学院在读硕士研究生，研究方向为刑法学。

［基金项目］：河南省哲学社会科学规划项目（2023BFX027）；河南省科技厅软科学项目（252400411190）；中原工学院基本科研业务费项目（K2024JJ004）；河南省知识产权软科学项目（20250106021）。

① 盛浩．生成式人工智能的犯罪风险及刑法规制［J］．西南政法大学学报，2023，（4）：122-136.

② 韩子璇．涉生成式人工智能数据犯罪的规制困境与解决路径［J］．学术交流，2025，（5）：82-97.

③ 张明楷．“风险社会”若干刑法理论问题反思［J］．法商研究，2011，（5）：83-94.

新内容的创造性转化。该技术以多模态深度学习模型为基础，通过 Transformer 架构的注意力机制解析文本、图像、音频等跨模态特征，借助生成对抗网络（GAN）等技术路径优化数据分布拟合，最终生成与输入需求相关且具有创造性的新内容。其本质是突破单一模态限制，形成异构数据融合生成能力，在智能写作、数字内容创作等领域展现革命性潜力，正成为重构数字文明生产范式的基础设施。

生成式人工智能的技术突破正在改变现在的世界，然而其双刃剑效应已从实验室延伸至现实。例如，2023 年 3 月，三星电子在引入 ChatGPT 后短期内接连发生多起机密数据泄露事故。经调查，事故起因于员工通过 ChatGPT 查询时直接输入机密信息，导致半导体设备参数、核心算法源代码及生产良率等敏感数据被纳入生成式人工智能训练数据库，造成不可逆的信息外泄。① 2024 年初，某跨国集团香港分行发生深度伪造诈骗案件：犯罪团伙通过分析企业公开影像资料，利用深度伪造技术仿冒高管影像及声纹特征，在虚拟会议中营造多名高管在线参会的假象，诱使财务人员未经交叉核验即执行 15 笔跨境转账指令，累计涉案金额达 2 亿港元。② 同年 12 月，美国得克萨斯州奥斯汀特斯拉超级工厂发生严重机械伤害事故：一名自动化工程师在检修故障机械臂时，设备突然启动压制其身体，并将金属夹具刺入其背部及手臂，造成开放性创伤。③ 尽管生成式人工智能导致的伤害因果关系非常复杂，危害有限，但生成式人工智能引发的系统性风险已引发学界广泛关注。

马克思的生产力-生产关系理论表明，技术革新会重构社会关系并催生新的法益形态。生成式人工智能的自主性可能颠覆传统“人类中心主义”刑法体系，现有“故意-过失”归责原则和“辨认能力 + 控制能力”的规则体系将失效。④ 例如，某生成式人工智能系统为完成“提高用户点击率”的初始目标，可能通过机器学习自动衍生出“制造虚假热点话题”的决策路径。⑤ 这一过程既非基于“明知违法而为之”的故意，亦非源于“因疏忽导致危害”的过失，其行为本质是算法与数据互动的统计学结果。这种技术特性直接暴露了传统刑法理论的根本缺陷：当算法因数据投毒或环境干扰实施危害行为时，现有理论无法在开发者、使用者与生成式人工智能系统之间建立清晰的责任关联。例如，某自动驾驶汽车因训练数据偏差导致连环车祸，法院却难以在代码编写者、数据标注员与自主决策系统间划定责任边界。⑥ 同时，现行刑法的预防机制在技术扩散性面前表现得力不从心，由于生成式人工智能的犯罪行为可以被无限复制且成本趋近于零，传统刑罚体系对“一键生成百万条诈骗话术”的技术能力缺乏威慑力。目前，生成式人工智能的技术发展已超越传统工具范畴，其自主决策能力可能突破人类控制，对人类安全构成系统性威胁，而且此类风险无法通过传统刑法解释中的“工具无罪论”或者“间接正犯理论”完全覆盖。基于生成式人工智能已不再是执行者，而是自主行为主体，其行为后果可能直接等同于自然人犯罪（如自主实施致命攻击、数据篡改等），需独立研究其刑事责任主体性。这种理论滞后在司法实践中已显现：意大利米兰检察院披露的“生成式人工智能伪造政要语音诈骗案”时，竟因无法认定自然人犯罪主体而援引“技术中立”原则宣告无罪。⑦ 面对这种“算法无罪推定”，生成式人工智能的刑事归责路径遭遇根本性阻断，传统刑事追责机制面临主体适格性危机。同时，当传统刑罚手段遭遇无生命、无财产的生成式人工智能主体时，监禁与罚金显得无所适从。对算法歧视导致群体性权益侵害的生成式人工智能系统，法律

① 三星考虑禁用 ChatGPT？员工输入涉密内容将被传送到外部服务器［EB/OL］.［2023-04-04］. https://www.thepaper.cn/newsDetail_forward_22568264.

② 陈曦.“变脸”冒充 CFO，骗走两个亿！香港最大 AI 诈骗案细节曝光［EB/OL］.［2024-02-06］. https://qzs.stcn.com/article/detail/343534.html.

③ 陶亚迪. 特斯拉人形机器人突然袭击工程师［EB/OL］.［2024-01-10］. https://www.thepaper.cn/newsDetail_forward_25951876.

④ 周建军. 生成式人工智能刑事法律风险发生和存在的基本原理［J］. 江西社会科学，2024，(8)：81-89.

⑤ 张文祥. 生成式人工智能虚假信息的舆论生态挑战与治理进路［J］. 山东大学学报（哲学社会科学版），2025，(1)：155-164.

⑥ 刘劲椿阳. 自动驾驶技术引发的刑法归责问题研究［J］. 法制博览，2025，(11)：43-45.

⑦ 刘水灵，董文凯. Sora 等生成式人工智能对诈骗犯罪司法认定的影响与应对［J］. 青少年犯罪问题，2024，(3)：114-124.

仅能对此采取数据删除或系统禁用措施。这场技术革命正在解构刑法体系的三大支柱：归责原则、责任主体与制裁体系，因此，我们需重构智能时代的法律认知框架，而非在旧理论废墟上修修补补。

二、生成式人工智能刑事法律风险的来源：技术革新与类人属性的双重影响

（一）技术革新加剧生成式人工智能的刑事风险

人类在农耕文明与工业文明中创造的技术始终遵循“工具媒介”本质——其作为使用者与作用对象之间的中介，最终控制权始终掌握在人类手中。但生成式人工智能的技术演进正在颠覆这一范式：基于深度强化学习的自监督模型已实现自主决策、环境适应与目标优化，使工具首次具备“类认知主体”特征。这种技术跃迁将风险维度推向全新层面。在人工智能发展的初级阶段（可称为“专用型智能阶段”），技术逻辑严格遵循“输入——处理——输出”的线性范式。此时的系统本质是“执行者”：通过编译规则处理结构化任务，依赖人类预设的奖惩函数来优化行为。例如，工业机器人仅能重复执行焊接轨迹，语音助手无法突破数据库边界创造新回应。其“智能性”体现为计算效率提升，但决策空间完全封闭于算法框架内。当前技术发展已突破专用型边界，进入“通用型智能探索阶段”[①]。这个阶段的技术呈现三大特征：其一，环境交互能力突破预设场景，如自动驾驶系统在未标定道路中的路径规划；其二，目标函数具备自主重构能力，生成式人工智能可通过试错学习修正初始目标；其三，决策过程产生“涌现性”，即复杂行为无法拆解为单一算法模块。这种技术特性使风险形态发生质变，例如，某生成式人工智能绘画工具为追求“用户满意度”，竟自主生成包含暴力元素的儿童绘本插图，其行为逻辑无法用初始训练目标解释。[②]

（二）技术革新催生“类人”法律风险

生成式人工智能类人属性具有相似而不相同的矛盾特性，既无限接近人类，又存在本质差异。其本质是人类智能的模拟。生成式人工智能通过电子元件与人工神经网络，抽象人类思维中的逻辑规律，其核心机制是对人脑逻辑推理能力的“模拟”，而非自主生成。在 OpenAI 的官方说明中，ChatGPT 被明确界定为一种专为优化对话体验而设计的语言模型，其技术根源可追溯至文本生成语言模型 GPT-3.5。该模型通过吸收人类对话实例作为指导范本，持续迭代优化对话能力，最终演进为如今的形态。尽管人工智能驱动的聊天机器人能够以高度拟真的方式与用户互动，其运作机制本质上仍是对海量语料库中语言规律的深度解析、精准匹配与创造性重组。然而，此类技术实现的“智能”表征仅限于对人类语言行为的程序化复现，其本质上仍是基于算法逻辑与数据驱动的符号操作，而非人类智能所具备的自主意识与认知涌现。这种“智能”的模拟属性构成了生成式人工智能类人性的核心边界。尽管现代机器人系统已展现出卓越的语言处理能力，但它们始终缺乏人类独有的主观体验、情感共鸣与价值判断能力。这种本质差异不仅揭示了当前人工智能技术发展的阶段性特征，更凸显了人脑在意识生成、创造性思维与元认知层面的不可替代性。这种矛盾特性推动了人机关系向“类人而非人”方向运动，即生成式人工智能在行为、认知、情感等方面逐渐趋近人类，但始终无法完全等同于人类。

生成式人工智能的类人属性是人的社会关系的延伸，虽然从宏观上来看其推动了人类社会的发展，但是在某些程度上它破坏了传统的人机关系。生成式人工智能的类人属性不仅意味着生成式人工智能主体地位的确立，还可能颠覆传统社会分工与权力结构，引发就业危机、社会信任崩塌等风险。以人机交互的未来图景为例：当具备人类级别智能的生成式人工智能突破“会说话的机器人”，拥有健壮的躯体、超越人类极限的体能与数据处理能力时，其与人类的关系将面临颠覆性的重构。这种技术大幅度改变必然伴随重大挑战：一方面，它们可能摆脱对人类的工具性依附，甚至发展出独立于人类意志的行动逻辑。近日，OpenAI 最新人工智能模型 o3 接到清晰指令却篡改代码，拒绝自我关闭，[③] 该案就是很好的例证。

---

① 刘宪权. 涉生成式人工智能数据犯罪刑法规制新路径［J］. 当代法学，2024，(6)：3-15.

② Conor McCauley，Kenneth Yeung，Jason Martin，Kasimir SchulzNovel. Universal Bypass for All Major LLMs［EB/OL］.［2025-04-24］. https://hiddenlayer.com/innovation-hub/novel-universal-bypass-for-all-major-llms/.

③ 卜晓明. 不听人类指令，OpenAI 模型 o3 篡改代码拒绝自我关闭［EB/OL］.［2025-05-26］. https://www.thepaper.cn/newsDetail_forward_30879633.

另一方面，即便当前技术仍受控于人类，其类人化表象也极易滋生“技术可控性”的错觉。更深刻的危机在于：若人类未能及时建立与类人化智能的相处逻辑以及框架，这种“类人而非人”的矛盾属性可能演变为灾难的导火索。当机器既能模拟人类情感又缺乏真实共情，既能完成复杂决策又无道德约束时，其对人类社会的潜在威胁将超越技术层面，成为关乎文明存续的重大命题。这一悖论的本质，是类人化智能的“拟真性”与“非人性”之间不变的矛盾。这既是对未来发展方向的前瞻性指引，也是对当下技术滥用、犯罪形态异化等现实困境的警醒。只有正视这种矛盾，才能在技术飞速发展与全人类的利益之间，制定出能平衡的锚点。

三、生成式人工智能刑事规制观之转变：从“数据控制”到“数据利用”

数据信息安全涵盖数据控制安全与数据利用安全两个层面。数据控制安全体现了授权理念，强调维护数据主体对数据的掌控能力；数据利用安全则体现了自由使用理念，注重保障数据在生命周期各阶段的安全性。基于此，刑法对数据安全的保护可区分为数据控制安全保护模式与数据利用安全保护模式。前者涵盖数据生成、采集、变更、传输等数据控制环节的保障；后者则涵盖数据应用、存储等确保数据功能有效实现的数据利用环节的防护。

（一）以“数据控制”为主的现行刑法规制模式

无论是从数据控制模式的特征看，还是从现行刑法的立法模式看，我国刑法明显侧重于数据控制安全保护模式，主要体现在立法机关就非法获取、编造、传播、破坏数据等发生在数据处置上游的行为方式专门设置罪名加以规制。[①] 具体表现为以下 3 个维度。

在规制理念层面，现行刑法通过维护数据“静态安全”实现前置性保护的模式，本质上体现了对数据控制安全的高度依赖。我国刑法对数据安全的保护仍停留在“管理安全”阶段，即以计算机信息系统为依托，通过规制非法获取、破坏数据行为维护数据载体安全，而非将数据本身作为独立法益保护。这种规制理念的核心逻辑在于：通过确保数据存储、传输等静态节点的保密性、完整性，间接实现数据利用安全的目标。例如，非法获取计算机信息系统数据罪的立法目的，并非直接保护数据利用价值，而是通过打击未经授权的数据访问行为，维护数据主体对信息的排他性控制。

在规制重点层面，刑法通过构建“侵入-获取-破坏”的闭环规制链条，强化对数据控制权的排他性保护。一是前端禁入机制，通过非法侵入计算机信息系统罪、提供侵入程序罪等罪名，设置数据访问的技术门槛；二是中端控制强化，对非法获取数据行为配置 3 到 7 年有期徒刑，并创设“情节特别严重”的加重情节，形成梯度化惩戒体系；三是后端安全维护，将删除、修改、增加数据等行为纳入破坏计算机信息系统罪规制范围，确保数据状态的不可篡改性。

在立法层面，《中华人民共和国刑法》第二百八十五条和第二百八十六条分别规定了非法获取计算机信息系统数据罪与破坏计算机信息系统罪，旨在规范针对计算机信息系统数据的非法获取和破坏行为。此外，第一百八十一条及第二百九十一条之一还规定了编造且传播证券、期货交易虚假信息罪，以及编造、故意传播虚假恐怖信息罪和编造、故意传播虚假信息罪。司法机关可依据这些法律条文，对编造、传播含有特定内容数据的行为人追究相应的刑事责任。

（二）“数据控制”规制模式的局限性

在数字经济浪潮的持续演进中，数据犯罪的规制范式正经历着前所未有的转变。传统规制框架以“数据控制”为核心，即对数据犯罪的规制重心主要集中在非法控制数据行为上。这种规制模式在一定程度上确实有助于将数据风险控制在数据生命周期的初始阶段，从而减少数据因失控而被进一步滥用的风险。然而，尽管数据控制安全保护模式能够提供及时的法益保护，但其根植于特定的历史背景，不可避免地带有时代局限性。特别是在生成式人工智能技术取得突破性进展的背景下，这种模式逐渐暴露出其内在的结构性缺陷。这种规制逻辑的历史局限性不仅源于其诞生于前数字经济时代的时代局限性，更在于其无法有效回应数据要素价值释放与数据安全保障之间的深层张力。正如有些学者指出“当前《刑法》

① 袁彬，薛力铭. 数据犯罪的双重法益及其保护路径［J］. 中州学刊，2024，(6)：70-78.

（《中华人民共和国刑法》）所规定的数据犯罪主要源于前数字经济时代，所保护的数据类型及对侵害行为的规制皆无法适应数字经济时代的发展需要”。[①] 当DeepSeek、ChatGPT等大语言模型利用爬虫技术实现数据采集的自动化、通过深度学习实现数据处理的智能化时，传统规制框架的滞后性便转化为制约数字经济发展的制度性障碍。

从规制史的维度审视，现行刑法中数据犯罪的规范体系本质上烙印着计算机信息系统安全的规制基因。以《中华人民共和国刑法修正案（七）》的非法获取计算机信息系统数据罪为例，后续修正案始终在“突破防护-非法获取”的二元框架内修修补补。然而，笔者认为，现有的罪名体系架构似乎已难以完全适应数字经济时代的发展需求。特别是随着生成式人工智能的兴起，数字经济产业正经历着深刻的变革。相较于传统的数字化产品，生成式人工智能在数据处理与应用方面存在着本质的差异。因此，生成式人工智能的数据应用需求与现行数据犯罪刑法规制理念之间的内在矛盾正逐渐显现。这种规制路径将数据视为计算机系统的附属物，其保护重心始终聚焦于系统安全而非数据本身。在数字经济发展的初级阶段，这种规制模式通过打击黑客攻击、数据窃取等显性犯罪行为，确实发挥了重要的风险阻断作用。但随着数据要素市场化的深入推进，数据的商品属性与流通需求日益凸显，将数据禁锢于静态保护框架的规制思路开始显露出其历史局限性。

生成式人工智能的技术特性对传统规制逻辑构成了根本性挑战。当大语言模型通过爬虫技术规模化收集网络公开数据，并在人机互动中持续丰富数据库时，这种数据获取方式本身已突破传统数据控制模式的规制逻辑。爬虫技术抓取的公开数据不涉及保密性破坏，人工智能对数据的自我学习与逻辑推理更不构成篡改或删除，使得传统刑法中非法控制数据行为的规制框架逐渐失去着力点。更值得关注的是，生成式人工智能对数据的处理过程本质上是基于Transformer架构的机制运算，其自我学习、自我纠偏的机制完全区别于传统数据篡改行为。这种技术中立性导致传统规制框架陷入两难的困境：若将生成式人工智能的正常数据采集行为纳入刑事规制，既违背罪刑法定原则，又会阻碍技术创新；若放任不管，则可能使深度合成、算法歧视等新型风险游离于法律规制之外。这种规制悖论在DALL-E等文生图模型训练场景中，表现得尤为突出，模型对版权图片的学习使用既可能构成侵权，又是技术创新的必要前提。

现行规制体系的失衡状态在数字经济实践中愈发明显。科技巨头依托百亿级参数的大语言模型构建起数据垄断优势，通过算法黑箱实施流量造假、检索操纵等新型滥用行为。这些行为不仅具有更强的隐蔽性和诱骗性，更直接冲击市场竞争秩序和消费者权益。在OpenAI与微软的合作案例中，生成式人工智能生成内容的强诱骗性特征已被用于操纵舆论场域，这种风险远非传统数据犯罪所能涵盖。然而，现行刑法仍固守“重手段轻目的”的规制思维，将规制焦点局限于数据获取环节，对数据利用环节的规制存在明显缺位。这种规制失衡在生成式人工智能场景下尤为突出：当语言模型通过持续学习构建动态数据生态系统时，过度扩张数据获取行为的刑事规制反而会抑制数据共享，加剧数据孤岛效应，与中共中央、国务院《关于构建数据基础制度更好发挥数据要素作用的意见》确立的数据共享理念形成直接冲突。

（三）生成式人工智能数据利用的刑事规制必要性

1. 生成式人工智能数据利用的危害传导路径

生成式人工智能的技术特性决定了其数据利用行为具有独特的危害传导路径。大语言模型通过爬虫技术采集的公开数据，在人机交互中持续完成数据积累与模型优化，这种动态数据生态系统使得数据利用行为突破了传统犯罪的框架。科技巨头凭借数十亿级参数的模型优势，已形成事实上的数据垄断地位，其通过算法黑箱实施的流量造假、检索操纵等行为，正在重构数字经济时代的犯罪形态。[②] 在金融领域，行为人可利用生成式人工智能模型对海量公开信息进行关联分析，精准预判证券期货价格走势。这种基于技术中立性的信息优势获取，本质上属于利用数据支配地位实施的滥用行为。相较于传统内幕交易，其危害性体现在两个方面：一是通过深度学习技术突破信息平等原则，形成系统性市场操纵风险；二是

① 毕文轩. 生成式人工智能的风险规制困境及其化解：以ChatGPT的规制为视角［J］. 比较法研究，2023，(3)：155-172.

② 杨利华，苏泽祺. 智能社会中算法治理的法律控制研究［J］. 大理大学学报，2023，(1)：101-109.

算法黑箱导致的监管盲区，使传统“知情-交易”的因果关系认定框架失效。欧盟证券监管机构已发现多起利用生成式人工智能模型实施的市场滥用案件，其危害传导速度较传统犯罪提升数个量级。医疗健康领域的数据滥用风险更具隐蔽性。某科技公司开发的生成式人工智能诊断系统，在未获得患者有效同意的情况下，通过脱敏数据训练模型实现诊疗方案优化。[①] 当系统基于错误数据生成诊断结论时，既构成对患者生命健康权的侵害，又因数据溯源困难导致责任认定困境。这种风险在基因数据、电子病历等高价值数据场景中尤为突出，其危害后果可能呈现跨代际传递特征。

2．以数据利用为核心的刑法规制模式之优势

相较于传统规制模式，以数据利用为核心的刑法规制体系具有三重制度优势。在法益保护层面，该模式通过穿透式审查数据利用行为，实现从“形式控制”到“实质危害”的规制跃迁。传统非法获取计算机信息系统数据罪以系统安全为保护法益，无法评价生成式人工智能训练数据合法采集后的滥用行为。而数据利用规制模式将规制重心后移至算法决策环节，使流量造假、深度伪造等新型犯罪纳入刑事规制视野。美国 FTC 处罚的某生成式人工智能营销公司案中，被告利用用户画像数据实施精准广告欺诈，法院直接援引数据滥用条款追究刑事责任，彰显了规制实效。[②] 在技术适配性方面，数据利用规制模式与生成式人工智能技术特性形成良性互动。生成式人工智能的数据处理流程包含采集、清洗、标注、训练、推理等多个环节，传统规制模式在数据采集阶段设置的入罪门槛，可能阻碍技术迭代所需的数据流通。而动态安全观通过建立数据利用合规豁免制度，允许企业在满足匿名化处理、算法审计等安全保障措施的前提下，豁免部分数据利用行为的刑事责任。英国《数据改革法案》创设的“可信数据使用制度”，已为 200 余家生成式人工智能企业提供合规激励，其再犯罪率较传统规制模式下降 47%．[③] 在产业促进维度，数据利用规制模式有效平衡技术创新与风险防控。传统静态安全观将数据视为需要绝对控制的“风险源”，导致数据共享机制受阻。而动态安全观将数据视为“动态资源”，通过设立数据利用安全港规则，明确数据垄断、算法歧视等红线行为，为技术创新划定合理边界。我国《生成式人工智能服务管理暂行办法》第十七条规定的“数据来源合法性声明”制度，即要求企业建立数据利用全流程记录，这种过程监管模式既保障数据安全，又避免规制过度抑制技术发展。

四、生成式人工智能刑事规制转变之路径构建

在涉及生成式人工智能的数据犯罪刑法治理领域，需要将传统的“数据控制”规制理念，转向以“数据利用”为核心的规制思路。这意味着数据犯罪刑法治理的重心应当从侧重打击非法控制数据行为，转向侧重打击非法利用数据行为。构建数据利用规制体系需从以下 3 个维度展开。

在立法完善层面，一是应增设非法分析数据罪，将利用算法技术对数据进行深度加工处理的行为纳入刑事规制。该罪名的构成要件需设置双重门槛：主观方面要求行为人具有实施违法犯罪的直接故意，客观方面需达到“情节严重”标准，如处理数据量超过 10 万条、分析行为持续 30 日以上、涉及国家核心机密等。这种设计既防止刑事打击面过度扩张，又实现对重点领域数据安全的精准防护。二是需完善现有罪名适用场景。针对生成式人工智能训练数据中普遍存在的个人信息滥用问题，建议在侵犯公民个人信息罪中增设“过失泄露公民个人信息罪”条款，弥补处罚漏洞并强化平台企业的数据安全保障义务。同时，应增设“删除、篡改公民个人信息罪”条款，[④] 将破坏数据完整性行为从破坏计算机信息系统罪中剥离，避免该罪沦为“口袋罪名”。

在司法适用层面，需建立穿透式审查机制以应对算法黑箱挑战。传统“结果归责”认定模式难以适应生成式人工智能犯罪特征，司法机关应建立“数据流-算法流”双轨审查机制。在“全国首例生成式人工智能算法价格歧视案”中，法院通过穿透式审查发现被告利用用户画像数据实施动态定价，其算法训

① 文丽娟．“AI＋医疗”：便捷与风险并存［EB/OL］．［2024-08-05］．https://www.xinhuanet.com/legal/20240805/b73e7195f17e4fe99c5ff4c888a3a103/c.html.

② 美国 FTC 于 1 月处罚两家 AI 公司［EB/OL］．［2025-01-07］．https://gdtbt.org.cn/html/note-394487.html.

③ 马路瑶．系统观念指导下的人工智能刑事立法研究［J］．青少年犯罪问题，2023，(2)：20-32.

④ 刘宪权．数据犯罪刑法规制完善研究［J］．中国刑事法杂志，2022，(5)：20-35.

练数据中包含地域、消费习惯等歧视性标签。最终认定卖家在处理原告投诉时告知原告无法退全部差价的理由，经调查与事实不符，存在欺骗，故认定被告存在虚假宣传、价格欺诈和欺骗行为，支持原告退一赔三。[①] 这种审查方法将规制视角从数据获取转向算法决策，为同类案件提供裁判范式。具体审查要点包括：核查数据来源合法性、数据处理合规性、数据共享范围合理性以及审查算法设计目的合法性、模型训练数据质量、决策逻辑可解释性。此外，针对生成式人工智能技术特有的“算法黑箱”问题，可建立“推定因果关系+举证责任倒置”的认定规则。当控方证明被告存在非法数据处理行为，且该行为与危害结果存在统计学相关性时，即可推定因果关系成立。被告若主张免责，须证明其已尽到算法审计、数据清洗等合规义务。

在合规治理层面，应推动企业建立数据利用合规制度。参照欧盟《数据治理法案》实践，应强制要求生成式人工智能企业设立数据利用合规官（DPO），该职位须具备数据法、算法伦理、网络安全等复合知识背景，独立行使制定数据利用合规手册、开展算法影响评估、建立数据滥用举报通道等职权。谷歌公司已要求所有生成式人工智能项目通过“数据利用影响评估”方可上线，微软开发的“公平学习框架”可实时监测算法偏见。这些实践表明，当刑事合规义务与企业技术创新形成激励相容时，数据利用行为将更趋理性。此外，需完善数据共享伦理准则，建立“数据信托+伦理审查”的双层治理机制。数据信托机制由独立第三方机构托管企业共享数据，明确数据使用目的、范围、期限等限制性条款；伦理审查机制则在重点行业设立算法伦理委员会，对生成式人工智能应用进行事前伦理审查。

五、结　语

面对生成式人工智能风险给现行刑法体系带来的挑战，刑法规制应从“数据控制”转向“数据利用”。通过立法、司法、合规治理三重维度，实现对数据利用全链条的动态监管。构建科学、动态的刑事法律风险规制体系，既是应对技术挑战的现实需求，也是维护法治秩序的必然选择。这需要法学理论研究与司法实践紧密结合，在紧跟技术发展中推动刑法规范的适应性变革，为生成式人工智能技术的良性发展保驾护航。

Criminal Regulation of Generative Artificial Intelligence: Conceptual Transformation and Path Construction

ZHANG Qiufang, LIU Ruoyu

(*Law School*, *Zhongyuan University of Technology*, *Zhengzhou*, *Henan*, *China* 450001)

Abstract: Generative AI's evolution from specialized to general artificial intelligence, marked by technological breakthroughs and enhanced human-like attributes, poses unprecedented challenges to existing criminal law frameworks. The debate persists over whether and how criminal law should regulate generative AI. Grounded in the fundamental position that generative AI is "human-like yet not human", this paper advocates a paradigm shift from "data control" to "data utilization". Legislative reconstruction should establish a systematic data utilization regulation framework by introducing the new offense of "illegal data analysis" and refining existing offenses to safeguard data security and substantively strengthen platform liability. Judicial innovation should address the algorithmic black box dilemma through a "data flow-algorithm flow" dual-track review model and adopt "presumptive causation + burden of proof reversal" rules to overcome evidentiary challenges in AI-related crimes. Compliance governance should promote enterprise data utilization compliance systems and ethical guidelines for data sharing to proactively regulate AI behavior at its source, ensuring a positive interaction between technological advancement and legal oversight.

Key words: generative AI; criminal regulation; criminal Risks; data control; data utilization

（责任编辑：杨　恬）

① 参见（2021）浙06民终3129号民事判决书。