生成式人工智能的刑事风险与刑法应对研究

罗可成[1]

（1.青岛大学法学院，rakasei@163.com）

摘要：以 ChatGPT 为代表的生成式人工智能技术具备内容生成、自然语言交互、多模态输出等实用能力，能推动数字社会和智能产业的快速发展。但同时也会因为数据处理不透明、生成过程不可控、输出内容具有复合性以及技术容易被滥用等特点带来新型的刑事风险。当前国内的刑法体系在面对这类风险时，往往会出现行为规制有漏洞、责任认定难度大、传统人类中心主义刑法理念受到冲击等现实问题。基于此，刑法需要坚持立法与解释并重、安全保障与技术发展平衡、理论前瞻与现实理性相协调的基本思路，通过解释路径完善现有规范的适用范围、立法路径补上规制的空白、监管路径明确平台的主体责任，搭建起体系化且有前瞻性的生成式人工智能刑事风险治理方案，在防范技术滥用引发犯罪的同时，也能为人工智能技术的健康发展提供对应的法治保障。

关键词：生成式人工智能，刑事风险，刑法规制，算法黑箱，责任认定

一、引言

如今的人工智能技术已经完成了从决策式向生成式的转变，以 ChatGPT、豆包等为代表的生成式人工智能产品快速普及，广泛用到内容创作、代码编写、智能客服、教育培训、工业设计等多个领域，成为推动数字经济发展的重要动力。生成式人工智能依靠大规模预训练模型、深度学习算法和海量的数据资源，能够自主生成文本、图像、音频、视频等全新内容，打破了传统人工智能只能被动执行指令的限制，展现出接近人类的学习、交互和创造能力。不过，技术的颠覆性创新往往会伴随风险的产生和扩散，生成式人工智能在数据采集、模型训练、内容生成、技术应用等全部环节中，都潜藏着数据安全、内容违法、知识产权侵权、技术滥用等多种风险，部分风险已经超出行政违法和民事侵权的界限，具备了较为严重的社会危害性，需要刑法介入进行规制[1]。

和传统的网络犯罪与智能犯罪相比，生成式人工智能带来的犯罪风险，有着风险源头多样、行为隐蔽、危害扩散快、责任链条复杂等特点。传统刑法理论以自然人主体、实行行为、因果关系清晰为核心搭建体系，很难完全适配算法黑箱、多主体参与、自动生成内容等新型技术场景[2]。在司法实践中，利用生成式人工智能实施电信诈骗、深度伪造、非法获取数据、传播违法信息等犯罪行为频繁出现，而现有的刑法规范在行为定性、责任划分、主体认定等方面存在适用上的障碍，导致一部分危害严重的行为难以被有效追责[3]。

在此背景下，系统梳理生成式人工智能涉及的刑事风险类型，分析刑法应对时遇到的现实困境，确立符合智能时代特点的刑法治理思路，提出兼具合法性、合理性和可操作性的刑法规制路径，既是应对技术发展带来的法治挑战，也是完善国内数字法治体系的必要举措。本文以生成式人工智能的技术特点和运行流程为基础，结合国内刑法规范和司法实践，对生成式人工智能刑事风险的刑法应对问题展开全面研究，希望能为人工智能领域的刑事治理提供理论参考和实践指引。

二、生成式人工智能的核心刑事风险类型

生成式人工智能的刑事风险，与其技术原理与应用模式紧密相关，贯穿数据准备、模型运算、内容生成、技术应用的全过程，主要体现在数据处理不透明引发的数据犯罪风险、生成过程不可控引发的内容违法犯罪风险、输出内容复合性引发的知识产权犯罪风险、技术被滥用引发的传统犯罪升级风险这四个方面，各类风险相互交织，形成了复合型的刑事风险体系[4]。

（一）数据处理不透明引发的数据犯罪风险

数据是生成式人工智能的核心生产资料，模型训练和功能优化，离不开海量的文本、图像、个人信息、政务数据等多种类型的数据资源，而数据在采集、存储、使用、传输环节的不透明状态，会直接引发数据安全类的刑事风险。生成式人工智能的训练数据，大多来自网络爬取、公开数据整合、用户交互信息收集等渠道，部分数据的获取没有得到权利人的授权，存在非法获取、非法持有、非法利用数据等行为特征[5]。从实际的技术应用来看，大部分的生成式人工智能的服务商为了提升模型的性能，会让自家的人工智能无差别地爬取网络数据，其中也包含公民个人信息、商业秘密、政务非公开数据等，这类行为很容易涉及非法获取计算机信息系统数据罪、侵犯公民个人信息罪等相关罪名。同时，用户在使用生成式人工智能服务时，输入的个人隐私、敏感信息，有被服务商擅自存储、泄露、滥用的可能[6]，因此，当前生成式人工智能有较高的数据犯罪风险。然而，传统刑法中非法获取计算机信息系统数据罪只针对侵入系统或利用技术手段获取数据的行为，却没有涵盖非法网络爬取以及合法获取后非法利用等新型的数据滥用行为，这就导致一部分严重危害数据安全的行为，难以被刑法约束和规制[7]。

（二）生成过程不可控引发的内容违法犯罪风险

生成式人工智能的生成过程，具备开放性、及时性和不可解释性的特点，形成了典型的算法黑箱，开发者和使用者都很难完全预测和控制输出的内容，进而会引发侮辱诽谤、传播淫秽物品、编造传播虚假信息等内容违法的犯罪风险[8]。受训练数据中的偏见、错误信息以及算法漏洞的影响，生成式人工智能有可能自动生成侮辱、诽谤他人的内容，侵害他人的人格尊严和名誉权，情节达到严重程度的，会构成侮辱罪、诽谤罪。

同时，一部分生成式人工智能产品会被诱导突破伦理和法律的限制，生成色情、暴力、恐怖等违法内容，或是提供犯罪方法、作案思路，变成违法犯罪的辅助工具。比如用户通过“AI 越狱”等方式绕过内容审核机制，让生成式人工智能生成制毒、诈骗、黑客攻击等犯罪指导内容，这类行为会给犯罪实施提供直接帮助，具备较为严重的社会危害性。此外生成式人工智能生成的虚假信息、虚假舆情内容，经过网络快速传播之后，有可能引发社会恐慌、扰乱公共秩序，符合编造、故意传播虚假信息罪的构成要件[9]。

生成过程不可控带来的刑事风险，本质上是算法自主决策和人类监管缺失之间的矛盾产物，传统刑法以故意实施危害行为为核心的归责逻辑，很难适配算法自动生成违法内容的场景，导致责任认定陷入难以解决的困境。

（三）输出内容复合性引发的知识产权犯罪风险

生成式人工智能的输出内容，是对海量已有作品的提取、整合、加工和再创作，有着高度的复合性，极易引发侵犯著作权的犯罪风险。生成式人工智能的模型训练，需要大量使用受著作权保护的文字、图片、音乐、视频等作品，多数情况都没有获得著作权人的授权，属于未经许可复制、使用他人作品的行为[10]。

在实际应用中，AI 绘画、AI 写作、AI 视频生成等产品，往往是在没有经过授权的情况下非法利用他人原创作品进行模型训练，生成的内容与原内容具有一定程度的相似性，变相侵害著作权人的复制权、信息网络传播权。如果参考国外相似情况，会发现美国版权局已经明确拒绝为 ChatGPT 式生成式人工智能自动生成的内容提供版权保护，同时相关的服务商也面临多起著作权侵权诉讼[11]。

传统侵犯著作权罪的规制逻辑主要针对自然人或单位的故意侵权行为，而生成式人工智能的作品生成由算法自动完成，多主体参与、技术中立性等因素，会让刑事认定和责任划分出现较大的争议，刑法谦抑性和知识产权保护之间的平衡，成为规制过程中的难点。

（四）技术滥用引发的传统犯罪升级风险

生成式人工智能作为开放式获取的技术，其技术特点容易被不法分子利用，在传统犯罪中加入人工智能的应用将会使实施犯罪成本的降低、危害范围扩大、隐蔽性增强，形成传统犯罪智能化升级的新型刑事风险。这类风险可以被分为直接滥用和间接滥用两种类型，直接滥用指的是犯罪分子直接利用生成式人工

智能输出的内容实施犯罪，比如通过 AI 换脸、AI 合成声音、利用 AI 生成虚假身份信息、虚假证件实施诈实施诈骗、敲诈勒索等传统犯罪。间接滥用指的是以生成式人工智能为工具，为犯罪行为提供相应的便利条件，比如利用 AI 生成钓鱼邮件、诈骗话术、电脑病毒，辅助实施网络犯罪、金融犯罪等传统犯罪。

利用 AI 进行深度伪造是技术滥用的典型表现之一，犯罪分子可以通过生成式人工智能拼接、合成他人面部、声音和肢体动作，制作出虚假的视频、音频往往被用于诽谤他人、实施诈骗、扰乱舆论，这类行为不仅会侵害公民的人身权利，还有可能危害到社会秩序和国家安全。和单纯的传统犯罪相比，生成式人工智能辅助下的传统犯罪行为有着伪装程度度高、传播速度快、溯源难度大等特点，给传统刑事侦查和刑事责任追究带来全新的挑战，现有的刑法规范和技术手段很难全面覆盖这类新型的犯罪行为。

三、刑法应对生成式人工智能刑事风险的现实困境

（一）风险类型多元导致行为规制存在漏洞

生成式人工智能衍生出的新型刑事风险，会让现有的刑法规范出现明显的规制空白，没办法做到全流程、全方位的覆盖。一方面针对深度伪造、算法滥用、非法数据爬取等新型危害行为，刑法缺少专门的罪名进行约束，只能通过解释适用侮辱罪、诽谤罪、侵犯公民个人信息罪等传统罪名，很难评价行为本身的社会危害性，形成两头管控严格、中间环节监管薄弱的规制缺陷[12]。比如深度伪造行为的核心危害在于伪造的过程本身，但现有的刑法只约束伪造之后的侮辱、诽谤、诈骗等结果行为，没办法对单纯的深度伪造行为进行追责。

（二）风险因素复杂造成归责判断困难

生成式人工智能的刑事风险，涉及数据供给方、模型开发方、服务提供方、用户等多个主体，技术链条较长、因果关系复杂，会让刑法归责面临多重阻碍。多主体参与会让因果关系很难认定，数据采集、模型训练、内容生成、传播应用等环节，都有可能对危害结果产生作用，各个主体之间没有意思联络，很难确定核心的责任主体，也没办法适用传统的共同犯罪规则[13]。

并且，生成式人工智能的算法黑箱会让主观过错很难判断，其开发者、服务商往往会以技术中立、无法预见算法结果为理由主张免责，而算法的不可解释性，会让司法机关很难证明其存在故意或过失的心态，导致危害结果发生之后没有主体承担责任。最后，工具属性和类人属性的混淆易使责任主体认定出现争议，部分观点认为生成式人工智能具备独立的责任能力，可以成为刑事责任主体，而传统刑法坚持人类中心主义，只认可自然人与单位为责任主体，理论上的分歧会直接影响到司法实践的适用[14]。

（三）类人化发展动摇刑法人类中心主义基础

生成式人工智能的自主学习、自动决策、类人创造能力在持续提升，会逐步突破传统的工具属性，给建立在人类中心主义基础上的刑法理论体系带来冲击。传统刑法以自然人具备自由意志、辨认能力与控制能力为前提，认为只有人类才能成为刑事责任主体，而生成式人工智能的智能化发展，会让部分学者提出它具备独立责任能力，应当被纳入刑事主体范围的观点。

这种理论上的争议不仅会影响责任认定，还会对刑法罪名体系、刑罚制度、立法模式产生全面的冲击。在罪名体系上，算法自动实施的危害行为，其实行行为、因果关系、罪过形式都和传统的自然人犯罪存在差异，现有的规范很难直接适用；在刑罚制度上，如果认可人工智能为责任主体，现有的自由刑、财产刑等刑罚种类无法适用，需要增设删除数据、修改算法、永久销毁等新型的制裁措施。[15]在立法模式上，涉人工智能犯罪的特殊性，会让传统刑法典难以兼容，需要考虑制定单行刑法，这和国内统一刑法典的立法模式存在冲突。从现实情况来看，当前的生成式人工智能依旧属于工具范畴，不具备独立的意志和自我答责能力，盲目赋予其刑事责任主体地位缺少现实基础，但是技术的快速发展，会让刑法人类中心主义持续面临挑战。

四、刑法应对生成式人工智能刑事风险的具体路径

（一）完善行为规制与责任归属的解释路径

面对生成式人工智能带来的刑事风险，可以优先通过刑法解释的方式激活现有法律条文，以此弥补行为规制不足、责任认定不清等问题。在行为规制层面，对于深度伪造引发的各类风险，可通过司法解释将使用人工智能伪造音视频并实施诽谤、侮辱的情形认定为情节严重，适当降低入罪条件，更有力地保护公民的名誉权。针对技术滥用问题，也可将利用生成式人工智能制作犯罪工具、传授犯罪手法的行为，解释为传授犯罪方法罪、帮助信息网络犯罪活动罪的具体实行行为，实现对这类帮助行为的独立追究[16]。

在责任归属层面则是需要清晰地界定多方主体的责任划分规则，用户故意借助生成式人工智能实施犯罪的，直接由用户承担刑事责任，平台服务商明知他人利用技术从事犯罪活动仍继续提供服务的，按照帮助犯追究责任。如果因为算法存在漏洞、平台监管不到位而生成违法内容，且服务商没有尽到合理的注意与管理义务，也应当由服务生承担相应的过失责任。若平台属于技术中立服务，并且已经履行法定监管义务，仍然无法避免危害结果发生，则可认定为允许的风险，不再追究其刑事责任。依靠明确的解释规则统一归责标准，能够有效化解算法黑箱带来的责任认定难题。

（二）弥补现有规范的规制漏洞的立法路径

通过立法修改与罪名增设，填补刑法的规制空白，完善刑事规制体系。一方面，完善数据犯罪的立法内容，将非法获取计算机信息系统数据罪修改为非法获取数据罪，取消数据领域与类型的限制，新增爬取获取型的行为方式，把网络爬取、合法获取后非法利用数据的行为纳入规制范围，实现对全类型数据的刑法保护[17]。另一方面增设算法滥用相关的罪名，借鉴非法利用信息网络罪的设置，增设非法利用算法罪，约束利用算法生成犯罪工具、传播违法信息、实施诈骗等行为；在帮助信息网络犯罪活动罪中新增提供算法服务的行为类型，明确为 AI 犯罪提供技术支持的刑事责任。同时，应当完善平台责任的立法内容，参照拒不履行信息网络安全管理义务罪，增设拒不履行算法安全管理义务罪，明确生成式人工智能服务商的安全监管义务，对于拒不履行数据审查、内容审核、风险防范义务，造成严重后果的，追究对应的刑事责任[18]。

五、结语

生成式人工智能作为具有颠覆性的技术创新，在推动社会进步的同时，也会带来数据安全、内容违法、知识产权侵权、技术滥用等多重刑事风险，给传统刑法体系提出全新的挑战。国内刑法应对这类风险，不可以盲目扩大犯罪圈，也不可以固守传统理念不做调整，需要立足技术特点和现实国情，坚持立法与解释并重、安全与发展平衡、前瞻与理性协调的基本思路，通过解释路径完善现有规范的适用、立法路径补上规制的空白、监管路径压实平台的责任，搭建起体系化、科学化的刑事规制体系。

同时，在具体的实践过程中需要始终坚守刑法谦抑性原则，区分技术风险和刑事犯罪，注重刑法和其他法律规范、技术治理手段的协同配合，既要严厉打击利用生成式人工智能实施的严重犯罪行为，也要为技术创新和产业发展营造良好的法治环境。随着人工智能技术的持续发展，新型风险会不断出现，刑法理论和立法也需要持续完善，始终保持规范的适应性和实践的指导性，在法治轨道上推动生成式人工智能技术健康、安全、有序地发展。

参考文献

[1] 李源粒. 现代刑法框架下生成式人工智能服务提供者的归责基础[J]. 环球法律评论, 2026, 48(02): 106-125.
[2] 周光权. 刑法各论: 第 4 版[M]. 北京: 中国人民大学出版社, 2021: 403.
[3] 庞良程. 生成式人工智能虚假信息的刑法规制研究[J]. 广东社会科学, 2026, (02): 273-285.
[4] 姜涛, 郭欣怡. 生成式人工智能涉虚假信息犯罪的刑法归责[J]. 重庆大学学报(社会科学版), 2025, 31(06): 183-197.

[5] 黄明儒, 刘方可. 生成式人工智能的刑事犯罪风险与刑法应对[J]. 河南财经政法大学学报, 2024, 39(03): 69-79.
[6] 西田典之. 日本刑法总论[M]. 刘明祥, 译. 北京: 法律出版社, 2013: 104
[7] 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, 25(04): 122-136.
[8] 刘艳红. 生成式人工智能的三大安全风险及法律规制——以 ChatGPT 为例[J]. 东方法学, 2023, (04): 29-43.
[9] 陈全真. 生成式人工智能与平台权力的再中心化[J]. 东方法学, 2023, (03): 61-71.
[10] 袁曾. 生成式人工智能的责任能力研究[J]. 东方法学, 2023, (03): 18-33.
[11] 王洋, 闫海. 生成式人工智能的风险迭代与规制革新——以 ChatGPT 为例[J]. 理论月刊, 2023, (06): 14-24.
[12] 马克昌. 比较刑法原理: 外国刑法学总论[M]. 武汉: 武汉大学出版社, 2015: 228-233
[13] 王建磊, 曹卉萌. ChatGPT 的传播特质、逻辑、范式[J]. 深圳大学学报(人文社会科学版), 2023, 40(02): 144-152.
[14] 邓建鹏, 朱怿成. ChatGPT 模型的法律风险及应对之策[J]. 新疆师范大学学报(哲学社会科学版), 2023, 44(05): 91-101+2.
[15] 徐永伟. 人工智能时代的刑事治理立场——基于主体、风险与责任的省思与建构[J].山东社会科学,2022, (10):169-175.
[16] 刘艳红. 人工智能的可解释性与 AI 的法律责任问题研究[J]. 社会科学文摘, 2022, (04): 8-11.
[17] 李怀胜. 滥用个人生物识别信息的刑事制裁思路——以人工智能“深度伪造”为例[J]. 政法论坛, 2020, 38(04): 144-154.
[18] 刘宪权, 朱彦. 人工智能时代对传统刑法理论的挑战[J]. 上海政法学院学报(法治论丛), 2018, 33(02): 44-51.