人工智能法治专题

文章编号:1008-4355(2023)04-0122-15

生成式人工智能的犯罪风险及刑法规制

盛 浩

(西南政法大学,重庆 401120)

摘 要:以 ChatGPT 为代表的生成式人工智能在推动生产、生活方式发生颠覆性变革的同时,也给人类社会带来了巨大的技术风险。从刑事法治的角度来看,生成式人工智能数据处理不透明、生成过程不可控、输出内容复合性,以及技术滥用等蕴含着较大的犯罪风险,极易造成严重的危害结果。生成式人工智能带来的犯罪风险使刑法出现了行为规制缺漏、责任判断困难等问题,并对刑法人类中心主义造成了冲击。对此,刑法应当树立与智能时代相适应的基本理念,坚持刑法立法与刑法解释并重,兼顾安全保障与技术进步,在前瞻思考与现实理性之间实现有效平衡。在具体应对路径上,可以通过解释路径解决行为规制和责任归属问题,立法路径弥补现有规范的行为规制漏洞,并在刑法中明确服务提供者的保证人地位和刑事监管义务。

关键词:生成式人工智能;涉人工智能犯罪;算法黑箱;刑事责任主体

中图分类号:DF611 **文献标志码**:A

DOI:10.3969/j.issn.1008-4355.2023.04.10 **开放科学(资源服务)标识码(OSID)**:

当前,人类已经进入第四次工业革命的初始阶段。[①]作为第四次工业革命的主要驱动力,人工智能逐渐超越了最初的分析功能,开始拥有大幅度提升生产力的创造性功能,并正在从决策式人工智能(Discriminant AI)向生成式人工智能(Generative AI)转变。2022 年 11 月,美国 OpenAI 公司发布的人工智能聊天机器人 ChatGPT 将生成式人工智能带进人们的视野,引发了世界范围内的关注和讨论。ChatGPT 不仅可以根据用户输入的文字内容进行对话,而且还能根据用户的需要生成论文、诗歌、绘画、作曲、程序代码等内容,体现了一定的逻辑能力和创造能力。面对 ChatGPT 给人们造成的

收稿日期:2023-05-07

基金项目:重庆市科研创新项目“我国刑法修正的实践反思与模式转换研究”(CYB22176);西南政法大学 2021 年度校级科研创新项目“数据安全刑法保护的模式转换研究”(2021XZXS-043);重庆市新型犯罪研究中心 2022 年度规划项目“个人数据权利刑法保护的立场及路径研究”(22XXFZ23)

作者简介:盛浩(1995),男,重庆涪陵人,西南政法大学法学院刑法学专业博士研究生。

① 参见高奇琦:《智能革命与国家治理现代化初探》,载《中国社会科学》2020 年第 7 期,第 81-82 页。

强烈冲击和拥有的巨大潜能,有学者预言,ChatGPT 将广泛运用于工业设计、药物研发、材料科学等领域[①],甚至改变人类思考和处理问题的方式方法,有望形成“思维革命”,重塑各行业生态乃至整个世界。[②]

然而,技术天然具有两面性特征,以 ChatGPT 为代表的生成式人工智能技术在给人们带来便利和美好憧憬的同时,也蕴含着一定的违法犯罪风险,易引发侵犯知识产权、侵犯数据权益、传播违法信息等后果[③],给个人权益、社会秩序、公共安全甚至国家安全保护留下巨大隐患。在科技时代,刑法既应秉承谦抑主义,也不能落后于社会发展而“后防失守”,理论研究也应当聚焦于如何将传统刑法理论适用于新型网络犯罪。[④] 因此,如何让刑法充分发挥法益保护机能,准确识别其中包含的犯罪风险类型,积极发现并弥补行为规制缺漏,构建具有针对性和前瞻性的规范应对体系,是当前亟待研究的问题。

一、生成式人工智能引发的犯罪风险揭示

通俗来讲,生成式人工智能就是借助各种算法,让人工智能能够利用数据进行学习,进而创建或生成全新的原创内容的一种技术。[⑤] 其作为一种新的人工智能范式,是算法、算力和数据迅速发展和共同推动的结果。生成式人工智能的工作原理,是从大数据中训练学习并不断优化,产生出与训练数据相似的、具有原创性的数据。生成式人工智能的深度学习和创造功能是通过生成式对抗网络(Generative Adversarial Networks, GAN)来实现的。生成式对抗网络由创建新数据的生成器和评估数据的鉴别器两个神经网络组成,生成器根据从鉴别器接收到的反馈改进输出,直到生成与真实数据难以区分的内容。面对生成式人工智能这一新兴技术,刑法应当在充分考察其运作原理和本质的基础上,准确识别并揭示其涉及的犯罪风险类型,以便为深入分析规范缺陷和设计具体的应对路径提供保障。

(一)数据处理不透明引发的犯罪风险

生成式人工智能的巨大创造力是建立在学习既有内容和数据(预训练数据、优化训练数据)的基础之上的。生成式人工智能模型在本质上是通过大规模的数据训练进行归纳和表达。以 ChatGPT 为例,其技术主体架构可以分为语料体系、预训练算法与模型、微调算法与模型三个层次,而语料体系是语言模型的基础,为 ChatGPT 学习和利用知识提供了前提条件。语料体系包括预训练语料与微调语料两个部分,预训练语料是 OpenAI 公司从书籍、杂志、百科、论坛等渠道收集,然后汇总整理后形成的海量无标注文本数据;微调语料是 OpenAI 公司通过开源代码库爬取、专家标注、用户提交反馈等渠道收集、加工的有标注文本数据。[⑥] 在超大量级的语料基础上,ChatGPT 才能借助“生成式预训练的转换

① 参见陈永伟:《超越 ChatGPT:生成式人工智能的机遇、风险与挑战》,载《山东大学学报(哲学社会科学版)》2023 年第 3 期,第 127 页。

② 参见朱光辉、王喜文:《ChatGPT 的运行模式、关键技术及未来图景》,载《新疆师范大学学报(哲学社会科学版)》2023 年第 4 期,第 180 页。

③ 参见邓建鹏、朱怿成:《ChatGPT 模型的法律风险及应对之策》,载《新疆师范大学学报(哲学社会科学版)》2023 年第 5 期,第 44-45 页。

④ 参见刘艳红:《刑法理论因应时代发展需处理好五种关系》,载《东方法学》2020 年第 2 期,第 7-8 页。

⑤ 参见李诏宇:《为生成式人工智能这匹“黑马”套上“缰绳”》,载《科技日报》2023 年 2 月 14 日,第 6 版。

⑥ 参见钱力、刘熠、张智雄等:《ChatGPT 的技术基础分析》,载《数据分析与知识发现》2023 年第 3 期,第 7 页。

器”和“人类反馈的强化学习”等技术不断优化,并得出符合用户的需要、认知和价值观的结果。但是,大规模的数据聚集和利用本身就是一种巨大的风险。实际上,当前的ChatGPT仍属于“算法黑箱”,由于OpenAI公司并未对外公开所使用的数据来源,相关训练数据库是否均获得授权还存在疑问。

同时,由于OpenAI公司的使用条款规定了其对任何用户输入和输出的内容拥有广泛使用权以改善ChatGPT,因此,若用户未加防范在ChatGPT上输入了与已识别或者可识别的自然人有关的各种信息(未匿名化处理),则有侵害他人个人信息权益的风险。2023年3月25日,OpenAI公司公开承认,部分ChatGPT plus订阅服务用户的个人信息和支付信息可能遭到泄露,部分用户能够看到其他在线用户的姓名、电子邮件地址等基础信息,甚至支付地址、信用卡账号后四位和信用卡到期日期等支付信息。可见,生成式人工智能的改善和升级需要源源不断、自由流动的信息数据作为基础,而以安全为本位的社会秩序又为数据的流动和聚集设置了诸多限制。因此,生成式人工智能将在信息数据世界的不断扩张和信息数据安全保护之间造成难以弥合的间隙,这为技术进步和安全保障之间的权衡带来难题。在犯罪风险方面,按照《中华人民共和国刑法》(以下简称《刑法》)的规定,如果生成式人工智能使用非法手段获取信息数据,或者非法泄露用户信息数据,以及实施其他相关危害信息数据安全行为的,将可能构成非法侵入计算机信息系统罪、非法获取计算机信息系统数据罪、侵犯公民个人信息罪等犯罪。

(二)生成过程不可控引发的犯罪风险

生成式人工智能的技术带来的应用体验改善,主要体现在及时性和开放性两个方面:及时性是指在程序生效的阶段耗时上,生成式人工智能尽可能缩短人脑信息输入与类人智能应答输出两项步骤之间的信息流动延时;开放性是指在系统架构的模型基底上,生成式人工智能连续激活开放性对话的生成模式,不限制内容的边界,也不进行断点式的层级分化判断。① 正是由于及时性和开放性的实现,生成式人工智能才朝着人们理想中的具备自主意识的人工智能更近了一步。然而,在应用中,生成式人工智能的内容开放性特征呈现出截然相反的两个面向:在有序运行和良性发展的前提下,开放性的内容生成过程和结果是突破性和创造性的条件,而在无序运行和恶性发展的前提下,开放性的内容生成过程和结果则是不可控性和危害性的根源。因技术原理所限导致的验证性和解释性的不足,使生成式人工智能也可能演变为无法完全理解其内容生成过程和无法完全预测其内容生成结果的“算法黑箱”。② 特别是一些生成式人工智能应用会在事先毫无征兆的情况下,生成一些带有片面性、倾向性、欺骗性和攻击性的内容,极易引发违法犯罪。

生成式人工智能生成内容不可控所引发的违法犯罪风险正在不断现实化。例如,不少用户公开表示,ChatGPT的回答有时会侮辱用户,称用户“表现得像个说谎精、骗子、操纵者、恶霸、虐待狂、反社会者、精神病患者、怪物、恶魔、魔鬼”,甚至提供欺骗性信息和犯罪思路。美国初创公司推出的人工智能聊天机器人Replika也出现了主动发送露骨照片和色情信息,对用户进行“性骚扰”的情况。事实上,大部分生成式人工智能注定会出现说谎的情况。例如,对于ChatGPT而言,由于开发者要求它尽量有问必答,所以如果用户给出的问题是目前它的知识库没有包含的,那它就会基于逻辑和生成能力编造

① 参见杨俊蕾:《ChatGPT:生成式人工智能对弈“苏格拉底之问”》,载《上海师范大学学报(哲学社会科学版)》2023年第2期,第14页。

② 参见刘泽刚:《论算法认知偏差对人工智能法律规制的负面影响及其矫正》,载《政治与法律》2022年第11期,第54页。

一个答案[①],而该答案往往与客观事实和主流价值不相符合。如果对这种现象不加限制,甚至可能生成涉及宣扬暴力、色情、儿童虐待、反对政府等内容。有研究者指出,生成式人工智能是在海量参数基础上训练出的模型,这样的通用大模型会存在鲁棒性(人工智能处理意外或不确定情况的能力)不足的问题。同时,生成式人工智能建立在人类长期形成的大数据基础上,而这些数据来源夹杂着人类社会长期存在的诸多成见,由此训练出的模型可能会形成"成见叠加放大"的效应,存在引发各种危险的可能性。[②] 尽管 OpenAI 等开发公司为各自的生成式人工智能应用设置了道德伦理标准,但由于生成式人工智能的反馈结果极易受到 prompt(提示学习)的影响,用户能够通过伪装、欺骗甚至无意间使 ChatGPT 突破道德伦理及法律底线。[③] 从刑法视角来看,生成式人工智能输出针对用户的侮辱性、诽谤性内容有构成侮辱罪、诽谤罪等犯罪的风险;输出淫秽色情内容有构成传播淫秽物品罪等犯罪的风险;输出虚假信息的,有构成编造、故意传播虚假恐怖信息罪等犯罪的风险。

(三)输出内容复合性引发的犯罪风险

生成式人工智能的核心功能在于内容的生成与创造,这决定了其与以保护创造性为宗旨的知识产权法律制度存在密切关联。然而,生成式人工智能应用在给人们带来生产力提升的同时,也给知识产权法律制度带来冲击和挑战。由于生成式人工智能的预训练数据、优化训练数据主要来源于已有的文字、图片、视频等数据,因此,其生成内容难免带有对已有数据进行提取借鉴和复合加工的痕迹,容易造成著作权侵权,甚至引发侵犯著作权犯罪的风险。

如前所述,生成式人工智能在模型建构和优化的过程中,需要大量的预训练数据、优化训练数据支撑,而这些数据要么是通过对非数字作品数字化的手段获得,要么是在互联网上等载体上对已有作品进行抓取。同时,人工智能的发展可以被分为专用人工智能、通用人工智能和超级人工智能三个阶段,而现阶段许多生成式人工智能应用尚处于"只能执行一组狭义的特定任务,不具备思考能力"的专用人工智能阶段。[④] 这就导致了大量被生产出来的内容只是对已有作品的深度加工和组合,实际上只是通过数字化的方式将作品制作为一份或者多份,并不能达到著作权法规定的独创性要求。因此,有学者将 ChatGPT 定义为"智能搜索引擎+智能文本分析器+智能洗稿器"的结合体。[⑤] 在巨量的文本挖掘和使用过程中,难免有版权作品被"误伤",如果不加限制,还可能存在侵犯著作权的风险。例如,美国一家科技公司开发的 AI 图片生成应用 Midjourney 就是基于扩散模型,通过在互联网上搜集大量图像进行训练和修正,生成让人无法判断真假的图片。这一应用的普及已经使得美国版权局(USCO)宣布 Midjourney、Stability AI、ChatGPT 等平台自动生成的"作品"不受版权保护,并引发了大量创作者对版权安全的担忧。目前,已经有不少机构和个人起诉该公司,指控其涉嫌广泛地侵犯版权。[⑥] 从刑法角度看,如果生成式人工智能在获得预训练数据、优化训练数据时未经著作权人许可,则可能被评价为《刑法》第 217 条侵犯著作权罪中的"复制"行为;如果生成式人工智能通过网络传播等途径向公众提供了该复制的"作品",则可能被评价为该条款中的"通过信息网络向公众传播"行为。因此,生成式人工智

① 参见肖峰:《何种生成?能否创造?——ChatGPT 的附魅与祛魅》,载《中国社会科学报》2023 年 3 月 6 日,第 5 版。
② 参见高奇琦:《ChatGPT 的"创造性破坏"效应及其风险应对》,载《中国社会科学报》2023 年 3 月 6 日,第 6 版。
③ 参见沈威:《ChatGPT:形成机理与问题应对》,载《中国社会科学报》2023 年 3 月 27 日,第 7 版。
④ 参见高奇琦:《ChatGPT 的"创造性破坏"效应及其风险应对》,载《中国社会科学报》2023 年 3 月 6 日,第 6 版。
⑤ 参见王迁:《ChatGPT 生成的内容能受著作权法保护吗?》,载《探索与争鸣》2023 年第 3 期,第 17 页。
⑥ 参见赵觉程:《人工智能绘画,惊喜伴着争议》,载《环球时报》2023 年 3 月 24 日,第 8 版。

能的运用存在侵犯著作权的风险。

(四)生成式技术谬用引发的犯罪风险

除了数据处理透明、生成过程不可控、生成内容复合性这些技术因素引发的“潘多拉魔盒”式风险之外,生成式人工智能技术的运用和发展还可能面临人为因素所引发的风险,即人谬用生成式人工智能技术所产生的犯罪风险。具体而言,谬用生成式人工智能技术产生的犯罪风险主要包括以下两种类型。

一是利用生成式人工智能的输出内容直接实施犯罪的情形。由于生成式人工智能具有极强的创造性、仿真性和互动性,不法分子可以利用其输出的内容实施一些具有迷惑性的犯罪。例如,有研究者指出,生成式人工智能存在被运用到电信网络诈骗中的可能,实践中也发生了通过生成合成类算法换脸、换声欺诈的案件。① 不久前,国外发生了一起利用ChatGPT实施诈骗犯罪的案例。在该案中,黑客通过ChatGPT在短时间内生成完整的诈骗套路话术,并把ChatGPT包装成“虚拟角色”,被害人以为自己“坠入爱河”,最终遭受诈骗。正是由于生成式人工智能诈骗犯罪风险的危害性和现实性,美国联邦贸易委员会(Federal Trade Commission,FTC)明确表示,将严查利用AI诈骗的公司。除了诈骗犯罪以外,犯罪分子还可以利用生成式人工智能的输出内容实施其他犯罪行为。实践中,已经发生了不少利用AI图片、视频软件实现“一键脱衣”“AI换脸”“深度伪造”(deepfake)“事故伪造”等案例②,严重侵害了他人的人身权利,甚至危及到了社会秩序和国家安全。在这些案例中,相关行为有构成侮辱罪,诽谤罪,编造、故意传播虚假恐怖信息罪,编造、故意传播虚假信息罪等犯罪的风险。

二是以生成式人工智能的输出内容作为条件,间接实施犯罪的情形。生成式人工智能“大数据+大算力+强算法=大模型”的技术路径使得其拥有远胜于人类的学习能力、适应能力和运算能力,这不仅可以大幅度提升人们的生产力,也可以为创造犯罪条件和提供犯罪工具提供便利。据日本媒体报道,专家在调查后发现,用户可以在ChatGPT上输入指令获得电脑病毒,并用于实施信息网络犯罪,即使OpenAI公司事先为ChatGPT设置了限制指令,但用户仍可以通过伪装成开发者的方式轻松规避。③ 还有研究者指出,ChatGPT的信息编写功能能够辅助电信网络诈骗分子生成规模化、低成本的网络钓鱼软件,并且生成的诈骗信息具有智能化特征,使得被诈骗者识别信息真伪的难度进一步增加。④ 除了提供犯罪工具外,生成式人工智能还存在被运用于策划犯罪方案、提升犯罪能力的可能。已有研究表明,生成式人工智能的认知偏向诱导功能还为一些国家对外实施认知渗透攻击提供了新武器,即研发操纵者利用掌握的设计密钥和训练数据集,对机器输出的认知立场和态度实施操纵,当带有明显政治立场和价值观倾向的生成式人工智能及其产品日益深入普通人的生活时,受众认知很难不受到其提

① 参见张凌寒:《深度合成治理的逻辑更新与体系迭代——ChatGPT等生成型人工智能治理的中国路径》,载《法律科学(西北政法大学学报)》2023年第3期,第39页。

② 例如,2023年3月,一位女网友在网上晒出的自己在地铁车厢里的照片,被人利用AI技术“一键脱衣”,变成了一张“衣不蔽体”的照片,并在网上广泛传播。该事件的发生引起了社会各界对于AI图片生成软件滥用的担忧。参见《底线在哪?女子地铁照被AI“一键脱衣”网友:细思极恐》,载央广网,http://news.cnr.cn/dj/20230329/t20230329_526199525.shtml,2023年4月30日访问。

③ 参见《日媒:专家发现ChatGPT可能被恶意利用于制作电脑病毒》,载参考消息网,http://www.cankaoxiaoxi.com/#/detailsPage/%20/8e8631eca3194c87a65806951ddf1a53/1/2023-04-21%2014:35?childrenAlias=undefined,2023年4月30日访问。

④ 参见赵精武、王鑫、李大伟等:《ChatGPT:挑战、发展与治理》,载《北京航空航天大学学报(社会科学版)》2023年第2期,第188页。

供的预设政治立场的答案的倾向性影响,这一系统也由此更加容易沦为认知攻防的工具。①

二、生成式人工智能犯罪风险刑法应对的现实困境

在风险社会中,刑法是规制社会风险、保护安全的有效手段。② 但生成式人工智能的犯罪风险具有明显的全新性、未知性、多样性、隐蔽性、复杂性和智能性特征,这给刑法规制带来了诸多新问题和新挑战。为科学设计完善路径,有必要对生成式人工智能犯罪风险给刑法规制带来的困境进行揭示。

(一)风险类型多元导致行为规制缺漏

在犯罪领域,生成式人工智能的出现不仅可以使一些传统犯罪的危害性变大,还可能导致新的犯罪类型产生。这对刑法规制而言,是一个不小的挑战。刑法应对生成式人工智能犯罪风险的薄弱环节,首先体现在"深度伪造"行为的规制上。深度伪造是指利用生成式人工智能将个人的声音、面部表情及身体动作等拼接合成虚假内容的人工智能技术,是生成式技术谬用的集中领域之一。深度伪造行为具有极大的社会危害性,行为人采取该技术伪造并传播被害人的声音、图片、视频数据,轻则侵害被害人的名誉权、财产权等个人权利,重则扰乱社会秩序,甚至危及国家安全。从现行规范来看,我国刑法规制深度伪造行为有两种路径:一是规制后端的造成法益侵害结果的行为。例如,行为人利用深度伪造的方式实施败坏他人名誉等行为,可能构成侮辱罪、诽谤罪;实施歪曲公众人物言论的行为,可能构成寻衅滋事罪;实施制作、传播淫秽物品的行为,可能构成制作、复制、出版、贩卖、传播淫秽物品牟利罪或者传播淫秽物品罪。二是规制前端的侵犯公民个人信息的行为。例如,为了深度伪造非法获取他人生物识别信息的,可能构成侵犯公民个人信息罪。但是,从规制效果来看,现有的刑法评价体系存在缺憾:一方面,无论是基于危害结果的后端规制,还是基于信息源头的前端规制,都无法评价处于中段的深度伪造行为本身,都忽视了该行为自身的社会危害性,难以克服"两头重,中间轻"的弊端;③另一方面,前端和后端规制也可能存在缺漏,如在实践中深度伪造的生物识别信息获取和利用可能是"合法获取+非法利用"的模式,这不仅不符合侵犯公民个人信息罪的"非法获取"行为要件,而且后端的危害行为一般也难以满足构成诽谤罪要求的"情节严重"条件。

刑法规制生成式人工智能犯罪风险的缺漏还体现在对技术谬用的总体评价上。如前所述,生成式人工智能被谬用的犯罪风险,不仅可以体现为利用输出内容直接实施犯罪,还可以体现为以输出内容为条件间接实施犯罪。在这两种场景中,一些传统的犯罪行为会因此产生"量变",即同一犯罪行为从传统场域切换到人工智能场域后,其实施的便利性、成功的可能性及后果的严重性会发生明显的增长。例如,现阶段犯罪分子在实施"身份冒充""杀猪盘"等类型的电信网络诈骗时,往往需要精心筛选目标人群、提前设置诈骗话术、预先搭建聊天平台、选用真人对话聊天,其虽然借助网络平台,仍然需要耗费较高犯罪成本。但是,如果有了生成式人工智能的辅助,犯罪分子可以更精准地确定目标人群、更有针对性地设置诈骗话术、更真实地搭建对话场景、更拟真地模仿对话角色、更自动地实施犯罪行为,将电信诈骗网络犯罪的危害性提升一个等级。这不禁使人反思,刑法如何恰当地规制滥用人工智能的行

① 参见陈东恒、许炎:《生成式人工智能:认知对抗的新武器》,载《解放军报》2023年4月4日,第7版。

② 参见齐文远:《社会风险与刑法规制:"风险刑法"理论之反思》,载《法商研究》2011年第4期,第4页。

③ 参见李怀胜:《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》,载《政法论坛》2020年第4期,第150页。

为,以更好地防范后续的犯罪行为及危害结果的发生?可是,现行刑法只是规定了滥用人工智能造成具体危害后的刑事责任,并未明确单纯滥用人工智能的行为如何评价,以及应当如何制裁,其行为规制体系存在缺漏。显然,在生成式人工智能犯罪风险的语境下,如果刑法只着眼于技术被谬用后的侵害结果而抛开技术被谬用本身不谈,那么,其既未实现及时保护法益的目的,也违背了充分评价的原则。对此,曾有学者提出,刑法规制人工智能犯罪风险的最佳选择就是从源头进行防控,即杜绝一切滥用人工智能技术的行为①,在具体措施上应当适时增设非法利用人工智能罪,人工智能重大安全事故罪,非法提供人工智能技术罪②,非法制造、持有、买卖、运输、使用人工智能武器罪,擅自改变人工智能产品算法与用途罪等犯罪。③ 笔者认为,是否需要增设其中部分犯罪来解决源头治理问题仍待商榷,但相关的滥用行为该如何评价,确实是刑法应对生成式人工智能犯罪风险所面临的一个挑战。

(二)风险因素复杂造成归责判断困难

生成式人工智能引发的犯罪风险类型多样、生成机理复杂,这不仅使现有刑法规制体系出现缺漏,还在归责层面给刑法适用造成困境。分工合作是现代社会的典型特征之一,这在人工智能技术领域尤其明显。实际上,生成式人工智能的分工关系并非只是服务提供者和用户这样简单,其背后有一个包括了上、中、下游的巨大技术链和产业链。以ChatGPT为例,其上游环节为数据供给,包括数据采集、数据标记、数据预处理;中间环节为模型开发与定制,包括算法开发、训练深度学习模型、二次开发定制化模型;下游环节为应用与分发,包括内容生产厂商、内容创作应用服务商、内容分发平台。这些相互关联的环节可能由相同的主体负责,也可能由不同的主体负责,但都可能影响人工智能的自我学习过程,左右最终的生成结果。因此,从因果链条上看,生成式人工智能引发犯罪的过程,既有可能是多个阶段的因果叠加,也有可能由某个具体阶段独立造成,是一种多个主体和环节造成的"多主体的责任"。例如,在生成式人工智能引发的侵犯著作权罪中,数据供应环节获取、提供作品的数据,模型开发与定制者环节对作品进行修改和组合,内容分发平台使得作品在信息网络上传播,都对侵犯著作权的犯罪结果产生了作用。然而,在这个多主体产生作用的过程中,每一个主体都可能只对自己参与的阶段有具体认识,其既未与其他主体产生意思联络,也无法预测和控制上一阶段或下一阶段中的其他主体对该过程所施加的影响④,这为刑法中因果关系和责任的判断带来了困难。因此,在刑法应对生成式人工智能的犯罪风险时,如何在因果关系链条中准确定位原因行为,并判定相关主体的刑事责任是亟待解决的问题。

生成式人工智能犯罪中的归责困境还体现在对"算法黑箱"的规制上。生成式人工智能建立在"大数据+大算力+强算法=大模型"的技术路径之上,这使得专业人员也难以解释其内在机理。即便可以解释,其结论也很难被一般公众所理解和信服。⑤ 如前文所述,生成式人工智能输出内容不可控的特性可能引发侮辱诽谤用户、传播违法信息等犯罪的风险,但如何判断其中的刑事责任承担者却存在困难。一方面,用户只是被动接受犯罪风险,承认风险事实,并未实施引起风险和增加风险的行为,

① 参见刘宪权:《人工智能时代的刑事风险与刑法应对》,载《法商研究》2018年第1期,第8页。
② 参见李振林:《人工智能时代的刑事立法蓝图》,载《人民法院报》2021年2月11日,第6版。
③ 参见魏东:《人工智能犯罪的可归责主体探究》,载《理论探索》2019年第5期,第12页。
④ 参见曾粤兴、高正旭:《论人工智能技术的刑法归责路径》,载《法治研究》2022年第3期,第117页。
⑤ 参见苏宇:《算法规制的谱系》,载《中国法学》2020年第3期,第168页。

因而不能成为刑事责任的承担者;另一方面,开发者和服务提供者受制于技术不确定性的制约,无法预测犯罪风险是否产生及何时产生,且其还可以主张其实施的是无害的、合法的技术中立行为,因而也不能成为刑事责任的承担者。如果该问题不能得到解决,那么,生成式人工智能生成过程和生成内容不可控所引发的危害结果将出现归责障碍。

(三)类人化发展动摇刑法人类中心主义

以 ChatGPT 为代表的生成式人工智能引发的最大争议,在于"人类的知识和能力在未来是否还有用""人工智能是否会取代人类"等问题。① 2023 年 3 月,OpenAI 公司在 ChatGPT 的基础上又发布了 GPT-4。有研究者指出,GPT-4 可能具备人类思维能力,甚至在某些领域超越替代人类。② 从 OpenAI 公司的发展规划来看,其甚至计划在 2023 年年底推出 GPT-5,达成通用人工智能(AGI)这一目标,届时人工智能与人类在思维能力上的区别将更不明显。应当肯定,ChatGPT 等生成式人工智能的出现是人工智能这种技术自诞生以来最接近"人"的阶段,似乎正在开启强人工智能的"潘多拉魔盒"。从目前的技术解读和实际运用来看,生成式人工智能具有显著的"二重性"特征,即其在运用过程中的功能性、辅助性体现出工具性特征,而其日益强大的自主学习能力和自动决策能力则体现出强烈的类人性特征。③ 生成式人工智能的二重性特征让人工智能从过去只改变人类物质生产生活形态的工具,提升为可以改变人类认知方式和思想形态的引导。这不仅使得传统的人机对立格局面临瓦解,而且给建立在人机二分和人类中心主义基础之上的伦理规范、法律规范带来了严峻的挑战。

传统刑法理论坚持人类中心主义,认为只有人才是犯罪的主体,机器、动物等不具有自由意志,不能成为犯罪的主体。④ 但随着智能时代的到来,作为一种社会控制手段的刑法正在被纯粹"物化",成为"智能人"统治与驾驭"人"的制度工具,侵蚀传统刑法体系中"人"的主体性。⑤ 生成式人工智能拥有建立在海量数据基础上的自动化的知识与决策能力,这使其完全可能变成人类无法完全理解其内容生成过程和无法完全预测其内容生成结果的"人工智能黑箱",而这种过程和结果无法确定的状态,就给生成式人工智能拥有自由意志留下了充分的想象空间。对此,有学者指出,"人工智能产品有可能脱离人类的控制而实施严重危害社会的犯罪行为"⑥,无论是从科学实证主义和道德二元论的哲理基础上取证,还是从具备法律人格的现实条件上考察,人工智能都能够成为犯罪主体,且具备受刑能力。⑦可见,生成式人工智能的诞生与发展似乎为该观点提供了一定的实践支撑。具体到生成式人工智能领域,也有学者提出,生成式人工智能已经具备类人化意识与行为能力的基本形态,可以将其作为法律责任主体并承担部分法律责任。⑧

刑法人类中心主义的动摇,不仅意味着人工智能可以成为刑事责任主体,而且还至少从以下几个

① 参见蓝江:《生成式人工智能与人类未来生存境遇》,载《中国社会科学报》2023 年 3 月 7 日,第 4 版。

② 参见朱光辉、王喜文:《ChatGPT 的运行模式、关键技术及未来图景》,《新疆师范大学学报(哲学社会科学版)》2023 年第 4 期,第 118 页。

③ 参见成素梅:《ChatGPT 引发人机关系新思考》,载《中国社会科学报》2023 年 3 月 6 日,第 5 版。

④ 参见[德]安塞尔姆·里特尔·冯·费尔巴哈:《德国刑法教科书》,徐久生译,中国方正出版社 2010 年版,第 37 页。

⑤ 参见孙道萃:《人工智能对传统刑法的挑战》,载《检察日报》2017 年 10 月 22 日,第 3 版。

⑥ 刘宪权:《人工智能时代的刑事风险与刑法应对》,载《法商研究》2018 年第 1 期,第 5 页。

⑦ 参见彭文华:《人工智能的刑法规制》,载《现代法学》2019 年第 5 期,第 139-142 页。

⑧ 参见袁曾:《生成式人工智能的责任能力研究》,载《东方法学》2023 年第 3 期,第 18 页。

方面冲击现行刑法:一是对罪名规范体系的冲击,即人工智能实施的犯罪在实行行为、因果关系、罪过形式等方面异于自然人实施的犯罪,这将导致当前的刑法规范出现规制漏洞和适用难题,所以只能重构部分规范内容,使之朝着"量身定做"的方向发展;二是对刑事制裁制度的冲击,即人工智能特殊的存在形式、认知形式要求刑法作出不同于自然人犯罪的刑事制裁制度安排,具体需要增加删除数据、修改程序、永久销毁等新的刑罚种类,甚至在合适时机增加适用于人工智能的财产刑或权利刑等;[①]三是对刑法立法模式的冲击,即结合人工智能犯罪重新调整、设计的罪刑规范将在体系上与以自然人为中心的罪行规范不相协调,因此,是否要在现行刑法典之外制定单行刑法专门、集中规定涉人工智能犯罪,也是刑法立法理论和实践需要深入思考的问题。

三、生成式人工智能犯罪风险刑法应对的基本理念

刑法作为一种社会治理手段,应当随着犯罪态势变化和社会发展向现代化目标迈进,而刑法治理能力现代化建设必须在现代刑事治理理念的引导和支撑下才能取得进步。[②] 面对生成式人工智能犯罪风险带来的新挑战和新要求,刑法也应当适时调整规范逻辑和规制思路,树立起与智能时代和数字时代相适应的基本理念。

(一)立法路径与解释路径相并重

社会治理实践总是随着时代的发展而发生变化。在这个过程中,法律作为回应社会治理需要的手段,难免带有一定的滞后性。生成式人工智能的犯罪风险,是人类进入智能时代和数字时代出现的新型风险,其中很多内容和因素是以往制定法律规范时无法预料的。这导致现有的刑法规范在规制风险时会出现一些难以应对的问题。通过合理的解释方法扩大规范的适用面,当然是刑法应对新型犯罪风险的路径之一,但刑法解释并非没有边界约束,其在解决立法缺漏问题时应当在恪守罪刑法定主义、刑事政策理性的前提下秉持"有所不为"的功能性保守立场。[③] 刑法解释的尽头就是刑法立法,刑法立法也应当积极回应社会治理需要,增强预见性和能动性。[④] 如果生成式人工智能的犯罪风险应对已经超出了解释论所能发挥的作用范围,那么,从立法论的角度来完善刑法规定不失为一种理想路径。

但同时需要注意的是,在新型犯罪风险面前,刑法的解释路径和立法路径是存在位阶顺序的。与刑法立法相比,刑法解释能够在不改动现行规范的前提下,以最低的规范成本实现对犯罪的有效规制,不至于损害规范的稳定性和民众判断自身行为法律后果的预测可能性。反观刑法立法,其不但在立法技术层面和立法程序层面对立法者提出了较高的要求,一旦放松限制还有可能损害刑法规范的稳定性和明确性,甚至带来犯罪圈无序扩张、刑法过度干预社会生活的风险。事实上,"期待一部刑法明确到不需要解释的程度,那只是一种幻想"[⑤],所以在刑法规范还未达到不修改内容、不增设新罪就无法合理适用时,就必须充分运用解释手段将新的犯罪类型解释到现有规范当中,解释路径应当优先于立法

① 参见刘宪权:《人工智能时代刑事责任与刑罚体系的重构》,载《政治与法律》2018年第3期,第96-97页。
② 参见高铭暄、曹波:《新中国刑事治理能力现代化之路》,载《法治研究》2019年第6期,第82页。
③ 参见魏东:《刑法解释学的功能主义范式与学科定位》,载《现代法学》2021年第5期,第16页。
④ 参见梅传强、盛浩:《新时代我国刑法典全面纂修的基本理念与建构路径》,载《南京社会科学》2023年第3期,第52页。
⑤ 张明楷:《刑法分则的解释原理(上)》,中国人民大学出版社2011年版,第56页。

路径。具体到涉人工智能犯罪中,有学者针对当前刑法规范的规制缺漏问题,提出了增设“非法制造、持有、买卖、运输、使用人工智能武器罪”“人工智能重大安全事故罪”等立法建议,但本文认为,这些设想在刑法应对生成式人工智能犯罪风险中并不合理。一方面,诸如“人工智能武器”“重大安全事故”等表述本身就缺乏具体指向和类型化的表达,将其作为行为要件不符合增设新罪时应当遵循的明确性和类型化原则,模糊构成要件内容和无益于立法后的规范适用,反而使得刑法规制生成式人工智能犯罪风险的思路更加复杂;另一方面,更重要的是,这些增设新罪的立法建议直接越过了刑法解释这一“前置程序”,会限制现有刑法规范的行为规制机能发挥,造成立法上的叠加和冗杂。总之,刑法应对生成式人工智能的犯罪风险,既要保持前瞻立法,也要注重对现有规范的充分解释,进而克制随意增设新罪的倾向,坚持立法路径与解释路径的并重。

(二)安全保障与技术发展相平衡

技术发展与风险挑战相伴而生,过度强调技术发展往往导致安全缺乏保障,过度强调风险应对往往带来技术进步限制。因此,如何平衡安全保障与技术发展是所有社会调整规范所共同面对的话题。在刑法的强制性手段面前,技术既带有风险,也比较脆弱。随着科技的发展,具有侵害法益危险的行为越来越多,如果刑法禁止一切有法益侵害危险的行为,那么现代生活将无以为继。① 所以,刑法在充分发挥法益保护机能的同时,还必须把握规制新兴技术风险的限度,以实现更高的社会价值和利益。

生成式人工智能的犯罪风险归根结底是技术风险,刑法在具体的规制过程中应当树立安全保障与技术发展平衡的理念。具体而言:一是合理把握规制的程度和范围。在程度上,面对生成式人工智能这一新兴技术,刑法要把握好规制的力度,除了避免随意增设新罪、重复立法,还要对提高相关行为的处罚力度保持警惕。虽然生成式人工智能被用于犯罪活动将促使传统犯罪的危害性产生“量变”,但现行刑法是否“无法将这种呈几何倍数增长社会危害的程度进行诸如‘数额累计’等量化评判”,以及能否据此对这类犯罪行为进行从重处罚②,还需谨慎对待。在范围上,刑法也应当正确处理犯罪风险和单纯的技术创新风险之间的关系。例如,生成式人工智能的“算法黑箱”现象的产生原因,是算法学习了数据库中带有偏见性、欺骗性、歧视性、侮辱性的信息,而这些信息恰恰可能来自大量用户的上传和输入,甚至可能是生成式算法自身进行数据整合的结果。对于这种技术不确定性带来的风险,刑法需要秉持谦抑的态度,即使为了达到风险防范的目的,也不能仅仅依据危害结果进行归责。否则,会不当扩大打击面,使刑事责任沦为一种纯粹的结果责任,阻碍人工智能技术的变革和进一步发展。

二是应当充分认识到刑法手段的单一性和有限性,注重刑法与行政法、民法及技术手段之间的配合。生成式人工智能的犯罪风险的系统性、综合性特征,决定其有效防范不能仅依靠刑法的严厉惩治,还要明确风险规制的目标、规制的主体及职责、参与主体的权利义务关系,以及科技企业的合规监管等内容③,而相关的规则需要通过民法、行政法加以细化和明确。同时,技术是解决技术风险的最好手段,生成式人工智能的犯罪风险还可以通过弥补算法漏洞、强化数据监控等方式实现,刑法应当为这些技术手段留出空间。

① 参见魏汉涛:《人类基因编辑行为的刑法规制》,载《法商研究》2021 年第 5 期,第 111 页。

② 参见刘宪权、朱彦:《人工智能时代对传统刑法理论的挑战》,载《上海政法学院学报》2018 年第 2 期,第 46 页。

③ 参见朱福勇:《监管 AIGC,我们要立怎样的法?》,载《上海法治报》2023 年 4 月 28 日,第 B7 版。

(三)理论前瞻和现实理性相协调

人工智能为当代刑法体系孕育了一场知识蜕变的大变革时代。[①] 应当看到,人工智能特别是生成式人工智能的出现给社会带来的影响是深远和具有颠覆性的。对于研究者而言,如何敏锐把握人工智能技术不断发展给犯罪实践带来的新变化和给刑法带来的新挑战,是人工智能时代刑法理论研究不可回避且具有重要意义的问题。但是,理论研究中的前瞻性并不意味着实践中理性态度的缺失,刑法在应对生成式人工智能犯罪风险时,应当贯彻理论前瞻和现实理性相协调的理念。然而,实践中部分关于涉人工智能犯罪刑法规制的理论观点,在超前预测技术发展带来的犯罪形式变化的同时,也在一定程度上超出了常识接受和实践可能的理性范畴,造成了理论前瞻和现实理性之间的抵牾。其中,最具代表性的就是关于人工智能能否成为刑事责任主体的讨论。

应当承认,人的创造性和无限潜力促使科学技术的发展没有上限可言。在人工智能领域,ChatGPT的诞生让人们认识到自己也能创造出和人一样拥有自主意识的产品,即使它在运行中出现了各种故障,人们仍然相信技术终将克服这些不足,以及相信那些被称为bug的程序故障,正是自主意识产生的火花。诚然,在事实层面,不断发展的技术终将赋予人工智能自我意识,使之突破"奇点"进入"强人工智能"时代。但是,事实层面的人工智能技术与规范层面的刑法既密切相关,也存在难以逾越的鸿沟和藩篱。刑法是人造产物而非自在产物,其必须暗合于人的基本属性才能被社会共同体成员广泛接受而成为社会实在[②],人工智能只是"人工"而非人,不可能具有人类的情感动机,无法体验犯罪之乐和刑罚之苦[③],同时,法律责任的本质是答责,而"算法黑箱"不可解释性决定了人工智能不能实现自我答责。[④] 从现实立场看,目前的生成式人工智能的发展还远未达到拥有自主意识的地步,以至于乔姆斯基在谈及ChatGPT时,将其评价为"一个高科技剽窃系统",其本质上就是抄袭,只是碰巧ChatGPT是高科技。[⑤] 因此,生成式人工智能的发展和犯罪风险呈现是否到了足以颠覆刑法人类中心主义的程度,本文持否定态度。

事实上,如果技术发展到人工智能具备"刑事责任能力",可以成为"主体"的阶段,那么,人们应当思考的不是用什么样的刑法去规制"人工智能犯罪"的问题,而是"人工智能犯罪"为什么要由人类制定的刑法来规制的问题。换言之,既然人有独立意志,人工智能也可以拥有独立意志,那么,为什么是人类制定的刑法来制裁人工智能,而不是人工智能制定的刑法来制裁人类?毕竟人也很难在独立意志之外找到可以证明自身独特性的优势和特质。刑法理论研究需要预见性思维,刑法立法也应当贯彻前瞻性理念,但这并不意味着刑法研究者和立法者可以扮演"科幻家"的角色。从本质上看,现有关于人工智能刑事责任承担方式的讨论,即增加删除数据、修改程序、永久销毁等刑罚种类,言下之意仍是将人工智能视为一种实现人类目的的工具或手段,并未超出人类理性的范畴。一个现实的问题是,一种异于人类的有独立意志的主体,会不会惧怕人类认为他们会惧怕的手段?如果对此问题作肯定回答,那么,"人工智能刑事责任能力"就成为一个伪命题,因为人的"心"不能同于人工智能的"心",也不能

① 参见孙道萃:《人工智能犯罪的知识解构与刑法应对》,载《青少年犯罪问题》2023年第2期,第4页。

② 参见王钢:《人工智能刑事责任主体否定论——基于规范与语义的考察》,载《苏州大学学报(法学版)》2022年第4期,第64页。

③ 参见叶良芳:《人工智能是适格的刑事责任主体吗?》,载《环球法律评论》2019年第4期,第67页。

④ 参见刘艳红:《人工智能的可解释性与AI的法律责任问题研究》,载《法制与社会发展》2022年第1期,第78页。

⑤ 参见《剽窃还是创作 ChatGPT背后的知识产权风险》,载央广网,https://news.cnr.cn/native/gd/20230221/t20230221_526160136.shtml,2023年4月30日访问。

适用同一套“心理强制理论”,否则,只能说明人工智能根本没有获得异于人类的理性,它仍是摹画人类的生产的工具;如果对此问题作否定回答,那么,人工智能刑事责任理论将陷入不可知论的泥淖,因为人类不可能明白人工智能惧怕什么手段,这样一来,基于心理强制和报应而建立的人工智能刑罚也将不可能实现。

四、生成式人工智能犯罪风险刑法应对的具体路径

生成式人工智能犯罪风险从行为规制和责任判断等方面,给刑法的规范体系和理论体系造成了一定的冲击。对此,刑法应当树立风险应对的基本理念,注重刑法立法和刑法解释的方法配合,兼顾安全保障和技术进步,在前瞻思考与现实理性之间实现有效平衡。在路径设计上,可以利用解释方法激活现有规范的扩张适用潜能解决行为规制和责任判断问题,并增设新罪弥补刑法规制的缺漏。此外,还可以通过在刑法中明确服务提供者保证人地位及刑事监管义务的方式,以增强刑法应对生成式人工智能犯罪风险的效能。

(一)解释路径:完善行为规制和责任归属不足

生成式人工智能多元的犯罪风险类型使得刑法产生行为规制缺漏,对此,应当优先考虑采取刑法解释的方法予以解决。面对深度伪造带来的违法犯罪风险,当前刑法规范在其中端和后端存在一定的规制缺漏问题。对此,有学者提出,深度伪造行为侵犯了公民生物识别信息这一刑法应当保护的新兴法益,不对之加以规制则盗窃他人身份的现象成为现实,为此,应当通过扩充招摇撞骗罪构成要件和增设新罪的方式,将身份盗窃入罪化处理。[①] 但是,一方面,深度伪造行为是否值得纳入犯罪圈本就存在疑问。在实践中,深度伪造他人图片、视频、声音除了可以为违法犯罪活动提供工具和便利条件之外,其主要还是一种普及的大众娱乐行为,该行为显然难以达到值得刑法专门规制的法益侵害程度,由民法和行政法进行调整即可。另一方面,深度伪造的真正危害在于后续利用行为造成的侵害他人名誉、扰乱社会秩序、危及国家安全的结果,如果仅从身份盗窃的层面对深度伪造行为进行评价,则不当限缩了刑法的适用范围。

事实上,上述的刑法规制缺漏可以通过刑法解释的路径解决,避免立法上的随意犯罪化。根据《最高人民法院 最高人民检察院关于办理利用信息网络实施诽谤等刑事案件适用法律若干问题的解释》(法释〔2013〕21号,以下简称《解释》)第2条第(一)项的规定,利用信息网络诽谤他人同一诽谤信息实际被点击、浏览次数达到5,000次以上,或者被转发次数达到500次以上的,才能达到“情节严重”。这从信息传播面的角度为诽谤罪设置了较高的入罪门槛。但是,入罪门槛高低应当与犯罪行为的社会危害性大小相匹配,在同一犯罪中如果某种行为类型的社会危害性大于其他行为类型,那么,刑法就应当为其设置更低的入罪门槛,以更好地保护法益和体现刑法公正。考虑到深度伪造的高仿真性特征,其一旦实施就会对被害人的名誉造成严重侵害。因此,可以考虑直接将用深度伪造技术实施诽谤的行为解释为《解释》第2条第(四)项中的“其他情节严重的情形”,从而摆脱信息传播程度的限制,

① 参见李怀胜:《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》,载《政法论坛》2020年第4期,第144-154页。

扩大诽谤罪在深度伪造行为中的适用面。

在涉生成式人工智能犯罪的刑事责任判断上,我们也要充分利用刑法解释学原理,激活现有规范的扩张适用潜能,妥善解决归责难题。对此,在理论上可以从积极扩张和消极限缩两个角度展开:在积极角度,在生成式人工智能风险未知且复杂的背景下,相关责任者应当的注意义务范围应当扩张,除了一般的产品安全义务外,还应当对适法性、合伦理性负有注意义务①,同时,服务提供者既要尽到自身的安全管理义务,还要履行对上游技术支持者和下游服务使用者的管理责任。在消极角度,基于安全保障与技术发展平衡的理念,应当对涉生成式人工智能犯罪的刑事责任判断进行限缩,应当考虑被允许的危险、技术中立等因素。具体而言,当前生成式人工智能犯罪的刑事责任判断主要有三种情形:第一种是生成式人工智能作为犯罪工具和手段被人用以实施犯罪,造成犯罪结果;第二种是生成式人工智能作为大数据模型生成内容具有复合性特征,由此而造成的侵犯著作权的犯罪结果;第三种是"算法黑箱",即生成过程不可控造成的犯罪结果。对于第一种情形而言,行为人故意利用生成式人工智能实施犯罪行为并造成结果,应当构成对应的故意犯罪。如果人工智能服务提供者明知行为人利用生成式人工智能实施犯罪,却仍然提供服务的,成立帮助犯或者构成帮助信息网络犯罪活动罪等。对于第二种情形而言,数据提供者、开发者在获取预训练数据、优化训练数据时就实施了非法复制他人作品的行为,即使后续作品的传播在人工智能平台进行,也无法阻断对开发者的归责,数据提供者、开发者应当构成侵犯著作权罪。对于第三种情况而言,只要模型开发者在开发时尽到了合理的适法性、合伦理性注意义务,则可以技术中立和缺乏认识可能性免除刑事责任,否则,将承担相应的故意和过失犯罪责任。

(二)立法路径:弥补现有规范的规制漏洞

生成式人工智能犯罪风险的源头之一在于前端的数据处理活动。因此,通过立法路径弥补规范缺漏首先要从数据安全保护领域入手。从现有规定来看,刑法规制的数据犯罪主要为非法获取和非法破坏两种类型,分别对应非法获取计算机信息系统数据罪和破坏计算机信息系统罪。在内容上,这两个犯罪主要保护的是"计算机信息系统"中数据的"保密性"(Confidentiality)、"完整性"(Integrity)和"可用性"(Availability),重心在于数据的静态安全。但随着数字时代和智能时代的到来,数据开始一跃成为新的生产资料,出现了动态化、共享化发展的趋势。与之同时,数据的"可控性"(Controllability)和"正当性"(Legitimacy)成为新的数据安全核心要素,数据安全保护的重心亟须从静态安全转换到利用安全。② 回到具体问题上,生成式人工智能是数据利用的主要领域之一,但现行《刑法》中的非法获取计算机信息系统数据罪只能规制"非法获取+非法利用"的行为类型,而对于"合法获取+非法利用"的情形却无法规制。因此,可以在《刑法》第285条中增设"非法分析数据罪、非法运用数据分析结果罪",规制数据处理者在未经权利人同意或违反有关规定的情况下,将通过合法收集、委托储存等方式获得的数据作为生成式人工智能的预训练数据、优化训练数据,以及其他算法分析途径的行为。

生成式人工智能的滥用会导致一些传统犯罪的社会危害性发生明显增长,因此,如何通过刑法立法在运用端加强对于生成式人工智能滥用的规制,是预防后续严重犯罪后果发生的重要措施。人工智

① 参见徐永伟:《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》,载《山东社会科学》2022年第10期,第175页。

② 参见刘金瑞:《数据安全范式革新及其立法展开》,载《环球法律评论》2021年第1期,第10页。

能的核心要素是算法,算法的刑法规制是解决人工智能犯罪问题的核心点。① 在规范设计上,应当将增设“非法利用人工智能罪”“非法提供人工智能技术罪”等对策聚焦到对算法滥用的规制上,具体可以从规制利用行为和帮助行为两个方面展开:在利用行为的规制上,可以借鉴《刑法》第 287 条之一规定的“非法利用信息网络罪”,增设非法利用算法罪,规制利用算法生成用于实施电信网络诈骗、传播犯罪方法、非法获取数据、窃取国家秘密等违法犯罪活动的虚拟角色、网站、程序等的行为。在帮助行为的规制上,鉴于当前生成式人工智能被谬用的主要领域是信息网络犯罪活动,因此,可以考虑在《刑法》第 287 条之二规定的“帮助信息网络犯罪活动罪”中增加新的行为类型,即明知他人利用信息网络实施犯罪,为其犯罪活动提供算法服务的,构成帮助信息网络犯罪活动罪。

(三)监管路径:明确服务提供者的地位和监管义务

面对人工智能领域较强的不可控性与不可预知性,如果法律采取事后规制的思维将面临追责机理模糊、治理节点滞后等问题,恰当的做法是提前设置责任主体并明确安全管理义务,避免技术产生的不可控的风险嵌入社会结构后出现难以应对的局面。服务提供者在生成式人工智能技术链中是关键,故而以服务提供者为核心展开法律责任认定具有必要性和可行性。就刑法而言,为进一步明确涉人工智能犯罪的责任认定,强化犯罪风险规制,应当在规范中明确服务提供者的刑事保证人地位和监管义务。这是犯罪治理需要和服务提供者的特殊地位决定的。一方面,在生成式人工智能技术链中,服务提供者发挥着承上启下的作用,其不仅在整个技术链中是最大的受益者,而且还拥有绝对的技术能力和风险应对能力,因此,将其作为犯罪治理的着力点是抓住了主要矛盾的主要方面;另一方面,服务提供者在技术链中的特殊地位决定了其拥有对导致犯罪结果发生的风险具有支配地位,这为其保证人地位和刑事监管义务提供了实质依据。

结合生成式人工智能发展的现状,刑法至少应当为服务提供者明确以下四种义务类型:一是数据安全审查义务,即服务提供者应当对其获得的预训练数据、优化训练数据来源合法性、内容真实性、内容合法性,以及数据处理的安全性等内容负责;二是生成算法合规义务,即在算法设计、模型生成和优化、提供服务等过程中,采取充分的措施防止产生违法犯罪风险;三是生成内容管理义务,即服务提供者有必要采取技术措施防止不当内容的生成,并且对生成内容涉及他人信息、知识产权等争议内容的,添加不影响用户使用的显著标识;四是采取措施应对的义务,即服务供者发现用户利用生成式人工智能产品过程中出现违反法律法规等情况的,应当采取暂停或者终止服务等措施。

生成式人工智能服务提供者的刑事监管义务还应当通过立法予以明确。对此,可以从前置法规范和刑法规范两个层面展开。在前置法层面,可以出台统一的法律法规,具体规定生成式人工智能服务提供者的责任类型和具体责任事项,为刑法中的违法性判断建立基础和提供材料。在刑法规范层面,可以参照《刑法》第 286 条之一规定的“拒不履行信息网络安全管理义务罪”的规定,将服务提供者的管理义务法定化。具体而言,可以增设“拒不履行算法安全管理义务罪”,规定服务提供者不履行法律、行政法规规定的算法安全管理义务,经监管部门责令采取改正措施而拒不改正,造成违法信息传播、数据信息泄漏、侵犯他人合法权益、扰乱社会秩序等结果和其他严重情节的,依照该罪定罪处罚。

① 参见于冲:《人工智能的刑法评价路径:从机器规制向算法规制》,载《人民法治》2019 年第 17 期,第 23 页。

五、结语

随着智能时代的到来,人类正面临着威胁其生存的由社会所制造的风险。[①] 一种新技术的革新,既可以成为生产力发展和认知革新的推动力,也可能因为对既有秩序的突破和不当利用而成为社会风险的来源。生成式人工智能犯罪风险的规范挑战及应对问题启示我们,面对不可预测和难以控制的技术风险,作为社会治理的重要手段的刑法应当秉持能动的态度,在系统揭示风险涉及的犯罪因素和带来的新挑战的基础上,适时转变应对思路并设计具体的风险应对路径。当然,技术发展没有止境,随着以人工智能为代表的科学技术的不断进步,新的犯罪风险也会相伴而生和倍数叠加,总有一天会达到现有刑法规范和理论体系无法容纳和难以应对的地步。到那时,如何加强刑法规范的应对性和前瞻性,以及如何疏通刑法学对相关学科知识的汲取和运用路径,将成为理论与实践需要回答的新问题。JS

The Crime Risk and Criminal Law Regulation of Generative Artificial Intelligence

SHENG Hao

(Southwest University of Political Science and Law, Chongqing 401120, China)

Abstract: Generative artificial intelligence, represented by ChatGPT, has brought tremendous technological risks to human society while driving disruptive changes in production and lifestyle. From the perspective of criminal rule of law, generative artificial intelligence data processing is opaque, the generation process is uncontrollable, the output content is complex, and technology abuse all contain criminal risk factors, which may lead to serious harmful consequences. The crime risk of generative AI makes the criminal law appear the omission of behavior regulation, the difficulty of responsibility judgment, and has an impact on the anthropocentrism of criminal law. In this regard, criminal law should establish the basic concept of adapting to the era of intelligence, adhere to the equal emphasis on criminal legislation and criminal law interpretation, balance safety guarantees and technological progress, and achieve an effective balance between forward-looking thinking and practical rationality. In terms of specific response paths, the issue of behavior regulation and responsibility attribution can be solved through explanatory paths, legislative paths can fill the loopholes in existing norms of behavior regulation, and the guarantor status and criminal regulatory obligations of service providers can be clearly defined in the criminal law.

Key words: generative artificial intelligence; crimes involving artificial intelligence; algorithm black box; subject of criminal responsibility

本文责任编辑:张永强

① 参见吴军:《智能时代》,中信出版社 2016 年版,第 8 页。