Doi:10.11835/j.issn.1008-5831.fx.2023.11.001

欢迎按以下格式引用:房慧颖.类ChatGPT系统刑事风险与治理策略[J].重庆大学学报(社会科学版),2024(6):263-272.
Doi:10.11835/j.issn.1008-5831.fx.2023.11.001.

Citation Format: FANG Huiying. Criminal risk and governance path of ChatGPT-like system[J]. Journal of Chongqing University (Social Science Edition),2024(6):263-272. Doi:10.11835/j.issn.1008-5831.fx.2023.11.001.

类ChatGPT系统的刑事风险与治理路径

房慧颖
(华东政法大学 刑事法学院,上海 200042)

摘要:"类ChatGPT系统",是指以ChatGPT为代表的生成式预训练和算法语言转换相结合的内容生成式人工智能(AIGC)。类ChatGPT系统在推动人类社会发展的同时,也带来相应刑事风险。类ChatGPT系统为了完成生成某一内容的指令而获取相应信息时,容易脱离人类控制与干预,有可能会不当侵犯其他数据库的保护措施,从而侵犯其他数据的保密性。同时,类ChatGPT系统生成违法犯罪信息,或使用者不当利用类ChatGPT系统所生成的信息,也可能会涉及相应的刑事风险。但是,根据类ChatGPT技术的发展现状,其尚不具有独立自主的意识和意志,不可能作为犯罪主体,其本质属性仍是人类的工具。刑法规制类ChatGPT系统犯罪,也并非规制类ChatGPT系统所实施的犯罪行为,而是规制类ChatGPT系统背后的人类过错。在类ChatGPT系统的研发和使用过程中,为有效治理技术风险,使技术"向善",研发者应当履行保证人义务,尽最大可能预测、识别类ChatGPT系统可能引发的刑事风险,并及时采取有效策略化解危机。使用者在使用类ChatGPT系统的过程中,不能故意诱导其违反道德伦理或法律法规输出违法犯罪内容。如果研发者未尽最大可能预见并阻止类ChatGPT系统输出违法犯罪信息,或者使用者故意诱导类ChatGPT系统输出违法犯罪信息,可能需要承担相应的刑事责任。从宏观层面而言,刑法在治理类ChatGPT系统涉及的刑事风险时,始终应当坚持自身的谦抑性,如果前置法律法规能够有效治理类ChatGPT系统所引发的风险,刑法的"触角"就应适度后移,避免因刑法规制手段的泛化使用而对技术创新发展形成阻碍。从微观层面而言,针对类ChatGPT系统刑事风险的内容与特征,需构建治理类ChatGPT系统刑事风险的分级治理机制:通过类ChatGPT系统研发企业制定与执行自治计划,探索能够有效预防类ChatGPT系统刑事风险的治理策略;提高行政监管力度,将科技监管手段融入行政监管体系,利用行政手段降低刑事风险实际发生的几率;刑事治理手段的使用必须严格遵循前置法律法规失效的前提条件,对行为刑事违法性的判断要在罪刑法定原则框架内进行,不能以行政违法性判断代替刑事违法性判断。

基金项目:上海市哲学社会科学规划课题"预防性刑事立法及其限度研究"(2022EFX003)
作者简介:房慧颖,华东政法大学刑事法学院副教授, Email:fanghuiying1203@163.com。

关键词:ChatGPT;数据犯罪;科技监管;分级治理
中图分类号:D924.7 **文献标志码**:A **文章编号**:1008-5831(2024)06-0263-10

2022年11月,美国的OpenAI公司发布的人工智能新产品——ChatGPT(Chat Generative Pre-trained Transformer),因其在自然语言处理领域的超智能性,在社会各界引发广泛讨论。本文所称"类ChatGPT系统",指的是以ChatGPT为代表的生成式预训练和算法语言转换相结合的内容生成式人工智能(AIGC)[1]。以ChatGPT、Bing Chat与Bard为代表的生成式人工智能,可以在短时间内完成高质量代码、论文等的创作和翻译,在人工智能创作领域取得了突飞猛进的进展。

类ChatGPT系统的运作模式为,在接收到使用者的创作指令时,首先搜罗与要创作内容相关的数据文本,将这些数据文本收纳到自己的数据库中,然后通过Transformer模型,生成和人类的思考方式以及表达习惯高度重合、类似的内容。ChatGPT这一现象级产品的问世,将引发对人类产生重大影响的"思维革命"[2],同时不可否认的是,类ChatGPT系统在促进人类科技进步的同时,也蕴含着诸多刑事风险。2023年8月15日,《生成式人工智能服务管理暂行办法》正式施行。作为中国首份生成式人工智能监管文件,上述管理办法体现了现阶段监管机构的态度,即"既要重视发展,也要重视风险"。本文将从人工智能刑法与刑法教义学的视角出发,检视类ChatGPT系统所涉刑事风险,并从宏观层面和微观层面进行分析与探讨,提出有针对性的治理策略,以达到使技术"向善"的目的。

一、类ChatGPT系统的三重刑事风险

类ChatGPT系统为了生成符合使用者要求的内容,需要获取大量的数据资料,而这一过程目前仍处于"黑箱"状态,也即并非是在人类的监督之下进行的;同时,在人类设置好程序之后,类ChatGPT系统基于Transformer模型生成相关内容的过程也无人类直接介入。这就导致类ChatGPT系统在一定程度上可能引发刑事风险。

第一重刑事风险:数据来源不当①。类ChatGPT系统作为基于神经网络开发的大语言模型,建立在海量文本数据预训练的基础上[2]。其里程碑式的发展和优越性,离不开互联网海量数据的支持,数据就是类ChatGPT系统的生命,相关研发企业则是典型的数据驱动型企业。类ChatGPT系统为了生成符合使用者要求的内容而在获取相关数据资料的过程中缺乏人类监督,也即类ChatGPT系统为了完成生成某一内容的指令而获取相应信息时,容易脱离人类控制与干预。根据现有技术,类ChatGPT仍是一个算法"黑箱",其数据来源从未被公开,其生成内容所依托的数据库中的数据来源是否经过数据权利人合法授权,目前仍然存疑。可见,在类ChatGPT系统预先学习既有的文本语料的过程中,有可能会不当侵犯其他数据库的保护措施,从而侵犯其他数据的保密性。换言之,刑法为了保护某些特定数据信息的保密性,而设置了相关条文,类ChatGPT系统在获取数据资料的过程中具有触犯这些刑法条文的刑事风险。例如,类ChatGPT系统在生成内容前获取数据的准备阶段,利用"爬虫"等工具获取了其他企业的商业秘密,则可能因触犯《中华人民共和国刑法》(以下简称《刑法》)第219条而构成侵犯商业秘密罪;再如,对于ChatGPT这类大语言模型而言,其需要的数据极其庞杂,不加限制地使用技术手段绕过他人技术防护获取数据,则可能因触犯《刑法》第285条

①主要涉及《中华人民共和国刑法》第180条、第219条、第253条之一、第282条、第285条第二款、第431条等规定的罪名。

第二款而构成非法获取计算机信息系统数据罪。

第二重刑事风险:生成内容不当②。如前所述,类 ChatGPT 系统要在无人类介入的情况下获取和学习相关数据资料,这就容易导致类 ChatGPT 系统在生成内容前的获取信息过程中脱离人类干预和控制。因此,类 ChatGPT 系统获取数据时,并没有经过人为的正确性过滤,从而导致类 ChatGPT 系统所使用的数据存在源头上的错误、虚假、不合规,乃至不合法风险。既然类 ChatGPT 系统所使用的数据在源头上存在不可靠性,则其生成内容的真实性、合法性就无法得到保障。类 ChatGPT 系统生成虚假或违法犯罪信息,容易给国家安全、社会稳定、个人权益保障造成巨大威胁。

第三重刑事风险:利用内容不当。一方面,将类 ChatGPT 系统所生成的违法犯罪内容,作为犯罪行为的"灵感"来源或者作为犯罪行为的帮助手段,会涉及相应刑事风险。例如,行为人利用类 ChatGPT 系统,根据受害者的特征制定个性化的诈骗计划,会大大增加受害者上当受骗的几率;再如,行为人通过类 ChatGPT 系统伪造与案件有关的信息,影响正常的司法秩序和司法判决。应当看到,根据类 ChatGPT 技术的发展现状,其尚不具有独立自主的意识和意志,不可能作为犯罪主体,其本质属性仍是人类的工具。刑法规制类 ChatGPT 系统犯罪,也并非规制类 ChatGPT 系统所实施的犯罪行为,而是规制类 ChatGPT 系统背后的人类过错。形式上看,犯罪行为是由类 ChatGPT 系统直接实施的,实质上类 ChatGPT 系统只是在完成使用者下达的指令,对犯罪行为和犯罪结果起直接支配作用的是使用者而非类 ChatGPT 系统[3]。辅助工具的使用不会使犯罪性质和犯罪形态发生根本变化。例如,使用棍子非法剥夺被害人的生命和使用杀人机器人非法剥夺被害人的生命,不会改变行为的故意杀人罪的根本性质。换言之,对犯罪的认定仍是以客观实际上发生的犯罪行为为标准,而非以帮助行为的来源为标准。

另一方面,未经有效许可而擅自利用类 ChatGPT 系统生成内容,可能会因侵犯相关著作权而涉嫌著作权犯罪。类 ChatGPT 技术的创新和发展,重塑着著作权领域中成果性质认定与保护的底层逻辑,并为当前法律制度带来了一系列颠覆性的挑战。关于人工智能生成的内容是否可以作为著作权法所保护的"作品",一直以来存在巨大争议。实践中曾出现过将人工智能生成内容认定为作品的案例③。而类 ChatGPT 系统的出现和发展,导致这一争论重新进入人们的视野。在类 ChatGPT 系统出现之前,人工智能生成的内容往往是程序机械性运作所得出的结果,不具有独特性[4],一般不会被认定为著作权法所保护的作品。而类 ChatGPT 系统所生成的内容和人类相比,在表达上几乎无异,且其具有了一定程度的"涌现能力"④,意味着其向通用型(Artificial General Intelligence)人工智能迈出了重要一步,也就意味着其生成的内容在形式上具有独创性。如类 ChatGPT 系统能够对某位知名作家的创作文风进行模仿,从而生成文风相同但内容却完全不同的新内容,从这个意义上而言,文学创作中的独立思考和计算机程序算法之间的界限被进一步缩小[5]。从实然层面而言,鉴于类 ChatGPT 系统生成的内容是人工智能基于算法和数据进行建模后自动生成的,并非人类的

---

②主要涉及《中华人民共和国刑法》第 103 条、第 105 条、第 120 条之三、第 181 条、第 221 条、第 249 条、第 278 条、第 291 条之一、第 295 条、第 299 条、第 299 条之一、第 353 条、第 359 条、第 364 条、第 373 条、第 378 条、第 422 条、第 433 条。

③2019 年,在"腾讯诉上海盈讯公司著作权侵权案"中,腾讯公司主张其开发的 Dreamwriter 软件生成的文章属于"作品",享有相关权利,该主张得到深圳市南山区人民法院的支持。法院认为:"从涉案文章的外在表现形式与生成过程分析,该文章的特定表现形式源于创作者个性化的选择与安排,并由 Dreamwriter 软件在技术上'生成'的创作过程均满足著作权法对文字作品的保护条件,本院认定涉案文章属于我国著作权法所保护的文字作品。"参见广东省深圳市南山区人民法院(2019)粤 0305 民初 14010 号判决书。

④"涌现能力"指的是当人工智能模型参数达到一定量级之后,会突然拥有包括常识推理、问答、翻译等一系列类似人类的"智慧能力"。

个性化表达,人类在此过程中的参与度极低,将其作为“作品”保护与传统著作权法理论相背离,存在制度障碍。然而,从应然层面而言,类ChatGPT系统生成的内容具有一定商业价值,通过著作权相关制度对其进行产权化保护[6],是在人工智能时代解决类ChatGPT系统生成内容所涉法律纠纷的有效路径。在此意义上而言,未经有效许可而擅自利用类ChatGPT系统生成内容,可能会因侵犯相关著作权而涉嫌著作权犯罪。

二、类ChatGPT系统刑事风险的宏观治理原则

刑法的规制手段与其他法律的规制手段相比,明显更具严厉性。治理类ChatGPT系统的刑事风险时,刑法应有所为有所不为,在保证算法安全可控的同时,也应坚持自身的谦抑性,从而避免因对社会治理的过度介入而形成对技术发展创新的阻碍。为此,刑法在规制类ChatGPT系统的刑事风险时,应坚持算法可控原则,坚守刑法谦抑性原则。

(一)坚持算法可控原则

算法作为类ChatGPT系统运作的底层支撑逻辑,具有极强的专业性,很容易在专业人士与非专业人士之间形成技术壁垒。类ChatGPT系统的正常运行离不开算法,但非专业人士难以真正掌握算法运作原理,更谈不上改变和影响算法运作过程。所以,有能力影响类ChatGPT系统算法编制和运作程序的只能是专业人士。尽管技术具有中立性已成为共识,但是类ChatGPT系统所依托的技术中立,不等于其输出内容的中立。事实上,类ChatGPT系统也可能体现人类的价值判断。尽管类ChatGPT系统的运作过程具有形式上的自主性,但其运作的原动力也即算法在被设计和编写的过程中,通常会不可避免地受到设计者价值取向、伦理道德等的影响。因此,在类ChatGPT技术发展过程中,坚守科技伦理,保证类ChatGPT技术发展的安全性、可靠性,便显得至关重要。在类ChatGPT系统研发与使用的过程中,为了最大限度地防范技术所引发的风险,确保技术“向善”,研发者应尽力排除安全隐患。以OpenAI公司为例,其在ChatGPT的研发阶段,应最大限度履行相应注意义务,从而避免ChatGPT生成含违法、犯罪信息的内容。具体而言,类ChatGPT技术发展所应坚持的算法可控原则以算法透明原则与算法安全原则的形式存在。

其一,算法透明原则。基于类ChatGPT系统的专业性,只有尽最大可能确保风险源头透明度,才有利于尽早识别类ChatGPT系统运行过程中可能引发的风险,有利于受害者(包括个人、社会乃至国家)明晰权益受损状况,进而明确造成权益受损方的刑事责任。因此,在类ChatGPT技术发展过程中,保证算法一定程度上的透明度,对于防控类ChatGPT系统风险具有关键性的作用。

其二,算法安全原则。所谓“算法安全”,并非是指类ChatGPT技术发展的零风险,而是指在类ChatGPT技术发展的同时,也应注重对个体权利的保障和对技术风险的有效预防,也即类ChatGPT技术的发展不应对个人与社会造成威胁或实质性侵害。应从以下三个层面理解算法安全原则:一是个人利益保障,即类ChatGPT的研发和应用不能造成个人权利与自由的减损,不能侵害个体尊严;二是技术安全,即类ChatGPT系统运行所依赖的算法与技术应符合相关法律、法规与行业规范,且具备及时修复系统安全漏洞的能力;三是算力资源安全,即类ChatGPT技术发展与创新过程中,不得以不恰当方式对他人合理使用算力资源的行为予以限制,应合理分配和使用算力资源。

(二)坚守刑法谦抑性原则

ChatGPT的出现是人工智能技术发展的重要里程碑[7]。然而,技术创新在带来社会进步的同时

也会产生相应的副作用。正如前述,类 ChatGPT 系统不仅可能沦为犯罪分子实施传统犯罪时的“帮凶”,还可能引发新型的犯罪,从而危害社会稳定与国家安全。当两种利益发生冲突时,为了重新构建法律和平状态,只有两种途径:一是一种利益向另一种利益让步;二是两种利益都作出一定程度的让步[8]。为最大限度地实现科技发展和社会治理二者的平衡,最大限度地发挥类 ChatGPT 系统的社会功用,进而最大限度地实现社会福祉,刑法需严格把握治理类 ChatGPT 系统刑事风险的尺度,既要实现有力治理类 ChatGPT 系统引发的风险,又不能扼杀类 ChatGPT 系统技术创新[9]。立法活动作为国家的上层建筑,不能先于经济变革的步伐而冲锋在前,而只能紧随经济发展与技术变革的脚步,及时作出调整,否则便有激进与轻率之嫌。激进的、过度活跃的刑事立法注重社会保护而忽视人权保障,积极证立国家刑罚权的扩张,隐含着国家权力过度膨胀的重大风险,与刑法谦抑性原则相抵牾。可见,刑事立法如不能顺应时代潮流作出与时俱进的调整,则会在规制技术发展所引发的风险方面显得力所不逮;如为满足时代和技术发展需求而作出过于频繁、活跃的修改,则会因抵牾刑法谦抑性原则而显得过犹不及。因此,刑事立法的调整需要把握合理尺度,尺度合理与否的检验标准便是是否符合刑法谦抑性原则。

如果刑法之外的其他部门法有能力从根本上治理或者遏制某个具有社会危害性的行为,实现对合法利益的保护,则刑法就没有将该行为认定为犯罪的必要性。这是由刑法严厉性特征及其在法律体系中的作用和地位所决定的。在法律体系中,刑法的地位是保障法,当其他法律无法抑止违法行为时,才能动用刑法。同时,刑罚的适用,既有积极作用,也有消极作用,应适当控制刑法的适用范围。也即当其他治理手段均宣告无效时,刑法将该行为认定为犯罪才具有必要性[10]。刑法目的的实现具有优先性序列,个人法益应优先得以保护,刑法不应为了维护社会秩序的良好运转而过度侵占公民个人自由的空间[11]。治理某一危害行为,应当遵循道德教化、行政规制、刑事惩戒的位阶,只有当前面的手段失效时,才可尝试使用后一手段[12]。在当今社会大变革时代,传统犯罪行为在新技术的影响下,会发生一定程度的“量变”或“质变”[13]。在这一过程中,刑法不能盲目扩大犯罪圈,一味强调介入社会治理,而应当始终保持审慎的态度,以理性姿态控制好规制不良行为的限度,防止触角的过分前伸。对刑法审慎介入社会治理的提倡,并不意味着对犯罪行为的放纵,而是倡导尽量优先适用其他法律规制手段[14]。只有其他法律手段无法实现对类 ChatGPT 系统所涉风险的有效治理之时,刑法的介入才具有正当性与必要性。

刑法与其他部门法的违法惩治手段的明显不同在于触犯刑法的法律后果是刑罚,刑罚实质上是对个人财产、自由乃至生命的剥夺,且刑罚确认之前的程序,诸如对犯罪嫌疑人实施的逮捕、拘留以及对被害人进行的审判等程序,都是重大的不利益,而其他部门法的法律后果,诸如恢复原状、赔偿损失,并非对行为人的重大不利益。因此,必须将刑事处罚限缩在“必要的最小限度之内”[15]。“即使行为侵害或威胁了他人的生活利益,也不是必须直接动用刑法。可能的话,采取其他社会统治手段才是理想的。可以说,只有在其他社会统治手段不充分时,或者其他社会统治手段(如私刑)过于强烈、有代之以刑罚的必要时,才可以动用刑法”[16]。正因如此,对于类 ChatGPT 系统所引发的风险,刑法固然不能视而不见、坐视不理,但同时也应随时以最后手段性标准进行自我检视。刑法在治理类 ChatGPT 系统涉及的刑事风险时,始终应当坚持自身的谦抑性,不能因为刑法规制手段的泛化使用而对技术创新发展形成阻碍。如果前置法律法规能够有效治理类 ChatGPT 系统涉及的刑事风险,刑法的“触角”就应适度后移,避免越位。

三、类ChatGPT系统刑事风险的微观治理策略

正如前述,刑法治理类ChatGPT系统涉及的刑事风险时,始终应当坚持算法可控原则和谦抑性原则。为此,应贯彻分级治理的策略。其一,通过类ChatGPT系统研发企业自治计划的执行,探索能够有效预防类ChatGPT系统刑事风险的策略[17];其二,提高行政监管力度,将科技监管手段融入行政监管体系,利用行政手段降低刑事风险实际发生的几率;其三,使用刑事手段治理已经触犯刑事违法性这条"底线"的犯罪行为,刑事治理手段的使用必须严格遵循前置法律法规失效的前提条件[18]。同时,应引导、督促研发企业建立并实施数据合规计划,从"源头"消减刑事风险,促使技术"向善"。

(一)企业自治:研发企业保证人义务的履行

具备开发类ChatGPT系统等技术的企业,通常是机制庞大的超大型互联网企业(下文将其简称为"研发企业")。研发企业的研发行为位于风险链条的前端,且具有明显技术优势,更有能力在风险现实化演进的进程前期化解风险。因此,研发企业作为类ChatGPT系统技术风险的"源头",理应更好地履行保证人义务,保证其所研发和生产的类ChatGPT系统无安全风险[19],确保技术"向善"。研发企业自治计划即是消减类ChatGPT技术风险、确保类ChatGPT技术有序发展的重要途径。

首先,研发企业自治计划实施的关键在于建立健全企业管理体系,确保企业的研发行为在法律法规的框架内进行,从而履行好保证人义务。对类ChatGPT系统研发企业而言,其自治计划主要应由两方面组成:一是有效履行保证人义务;二是配合执法与应对危机。研发企业自治计划的实施,有利于阻断类ChatGPT系统研发企业犯罪,实现企业治理的积极效果,更为彻底地化解类ChatGPT系统的风险与危机。

其次,研发企业自治计划实施的方式在于保证人义务的履行。算法可控性及算法可解释性是实现类ChatGPT系统研发企业保证人义务的基础,也是治理和防范类ChatGPT系统刑事风险的关键[20]。有学者提出,"技术黑箱"始终都是存在的[21],但是技术人员已经通过逆向工程等方法触摸到了神经网络底层逻辑,并以可视化形式展示这一技术成果,从而打开了神经网络黑箱[22]。社会中的每个主体都承担不同的社会角色,承载不同的社会意义,但不同主体之间的角色和意义不能割裂,而要把每个主体的角色和意义放在整体的社会生态中进行把握和理解[23]。研发企业位于风险链条前端,具有防控类ChatGPT系统风险的技术优势,其保证人义务的履行可以最大程度降低危害结果发生的现实可能性。

最后,研发企业自治计划实施的重点在于对数据的依法依规利用。类ChatGPT系统研发企业在运营的过程中,往往会伴随数据的传输、存储、利用及交易等,这容易导致对数据安全的多层次、多节点的威胁。在制定类ChatGPT系统研发企业自治计划时,应把数据依法依规利用视作重点内容,明确研发企业经营过程中所要承担的具体作为义务,也即将刑法赋予的数据安全保障义务以具体化方式呈现为企业的运营规则,该运营规则中包括履行数据安全保障义务的具体方式和手段,从而真正实现将外在的法定义务内化成企业自身的运行方式和管理制度。具体而言,数据获取阶段,研发企业应根据相关法律法规,制定具体规章制度,建立数据获取的风险评估机制;在数据利用阶段,研发企业应提高违法犯罪预判能力和识别水平,在源头上实现对数据泄露与滥用数据风险的有效防控。

(二)政府监管:科技监管融入行政监管的具体方式

类 ChatGPT 系统的专业性强,通过传统的监管手段与依靠人为监管的力量无法有效突破技术壁垒,因此,在预防和治理类 ChatGPT 系统刑事风险时,科技手段可以在行政监管中扮演重要角色。科技手段能有效提升行政监管效能,降低类 ChatGPT 系统风险向现实演进的概率。科技监管手段的实现方式具体如下。

第一,监管机构与研发企业二者之间建立数据信息共享与合作共治机制。在以往的监管模式中,被监管人被动提供运营数据以供监管机关检查和监督。被监管人被动提供的数据是监管人判断企业行为是否合法的重要依据。在这种监管模式下,监管者自上而下地对被监管者实施治理与监管,看似处于强势地位,但在信息获取层面看,被监管者所提供的信息可能是被筛选、隐藏甚至是伪造的,形式上处于优势地位的监管者,实际上位于劣势地位。数据信息共享机制和合作共治机制的建成,有利于改变监管者获取信息被动性所导致的弱势地位,使监管者可以主动地获取数据,有效破解传统监管模式中信息不对称的困局。同时,正如前述,对类 ChatGPT 系统研发企业而言,监管者对被监管者进行监管的依据主要是监管者所获取的运营数据。传统监管模式下,被监管者在被动及被强制提供数据时,为了企业自身利益的最大化,可能会怠于履行提供数据的义务,甚至可能会提供被筛选、隐藏甚至是伪造的数据。在研发企业和监管机构之间,建立共享数据信息的有效合作机制,实现二者的合作共治,有利于监管者获取第一手数据资料,从根本上杜绝被监管企业伪造数据资料的可能性。这将大大提升监管者对于风险的预测能力和识别能力,从而有利于更快速地化解风险、处理危机。

第二,监管机构建立实时、动态的智能化监管与审核机制。科技的快速发展催生出了类 ChatGPT 系统技术,与传统技术相比,类 ChatGPT 技术具有一定程度的先进性和特殊之处,如果监管机构仍采用传统监管方式,则无法实现对新技术所引发风险的有效监管。因此,监管机构应将科技监管的手段充分融入行政监管手段之中,进一步提升监管技术,根据类 ChatGPT 系统技术的特征与发展水平,有针对性地利用智能化的审核技术进行监管。监管者拓展监管方式,运用智能化技术充分提高和优化监管水平,有利于更加快速、精确地预判、识别和处理风险。具体来说,监管者可将监管规则通过编程,以代码化的方式予以呈现,并将其内化于监管系统。监管者在监督研发企业运营过程时,可通过自动比对的形式,比对类 ChatGPT 系统研发过程和相关规则的匹配程度,如有非正常情况,则及时发出警告信息。此种方式可以帮助监管者准确、及时地识别类 ChatGPT 系统运作时所出现的风险或者异常情况,进而及时制止和处理风险,实现于无形中化解危机的初衷。退一步讲,即便监管者未能有效化解危机,仍使类 ChatGPT 系统的风险现实化,监管者对类 ChatGPT 系统所作的实时监督所形成的数据资料,也可以成为危险现实化后进行事后治理的重要证据和有力依据[24]。

可见,技术的发展一方面为行政监管增加难度、带来挑战,另一方面也为监管模式创新和变革提供了难能可贵的机遇。将科技监管手段融入行政监管体系,具有显而易见的优越性与重要意义。加强对类 ChatGPT 系统的科技监管,能够在相当程度上弥合传统监管模式的缺陷,有利于监管者准确、快速地预判、识别和预防风险,最终实现类 ChatGPT 系统技术的健康发展。

(三)司法认定:涉数据犯罪的认定方案

类 ChatGPT 系统并不能作为犯罪主体,换言之,类 ChatGPT 系统没有承担刑事责任的主体资

格,其研发企业、使用者等相关自然人或单位主体才可能最终作为刑事责任主体,为类ChatGPT系统所涉及的刑事风险承担刑事责任。对于类ChatGPT系统突破道德伦理和法律底线而输出违法犯罪信息,在满足犯罪构成要件的前提下,根据违法犯罪信息的内容追究研发企业或者使用者的刑事责任即可;未取得使用资格的自然人或者单位使用类ChatGPT系统生成的内容是否有可能构成著作权犯罪,取决于著作权法对于"作品"的认定标准。值得探讨的是,在类ChatGPT系统生成相应内容前,其会进行获取数据等相应的准备工作,在此过程中,如果其有不当行为且涉及刑事风险,如何进行追责值得探讨。笔者认为,对于类ChatGPT系统研发企业涉嫌数据信息犯罪的追责,应严格遵循数据犯罪的双重违法性特征,同时,不应以对获取数据行为行政违法性的判断替代对行为刑事违法性的判断。

其一,对类ChatGPT系统研发企业涉嫌数据信息犯罪的认定,应严格遵循双重违法性标准。数据犯罪是法定犯,其构成应同时满足行政违法性和刑事违法性。换言之,类ChatGPT系统研发企业不当获取数据的行为,只有在违反了相应的前置性法律法规,具备了行政违法性的前提下,我们才能进一步考虑该行为是否违反了刑法中有关数据犯罪的规定,是否构成犯罪。在类ChatGPT系统研发企业不当获取数据行为的司法认定中,前置性法律法规始终应处于承上启下的地位,即当研发企业自治计划失灵时,监管机关应根据前置性法律法规的规定对研发企业进行处罚;当前置性法律法规失效时,刑法才能介入对研发企业不当获取数据行为的规制。只有坚持企业自治计划、前置性法律法规治理、刑法规制的三阶治理策略,才能准确把握刑法介入类ChatGPT系统研发企业不当获取数据行为的时机[25],避免刑法过早介入而阻碍研发企业的积极性,甚至遏制技术进步的进程,同时也避免刑法过晚介入对数据安全造成无法挽回的不利影响。

其二,对类ChatGPT系统研发企业涉嫌数据信息犯罪的认定,不应以对获取数据行为行政违法性的判断替代对行为刑事违法性的判断。应当看到,类ChatGPT系统研发企业不当获取数据的行为具有行政违法性,只是其具有刑事违法性的必要条件而非充分条件。换言之,类ChatGPT系统研发企业不当获取数据的行为构成犯罪,则其必然具有行政违法性;但反之,类ChatGPT系统研发企业不当获取数据的行为具有行政违法性,并不必然证明其构成犯罪。认定某一行为构成犯罪,应从行为实质侵害法益的角度进行判断,不能简单地以行为具有行政违法性作为充分必要条件。详言之,《网络安全法》《数据安全法》《个人信息保护法》等前置性法律法规为数据犯罪的认定提供了细致、具体的标准,未违反上述前置性法律法规的行为必然不构成数据犯罪,但反之则不然,以防刑罚处罚范围的不当扩张。

结语

类ChatGPT系统涉及的刑事风险包括数据来源不当、生成内容违法违规、使用不当手段对生成的内容加以利用等。刑法在治理类ChatGPT系统所涉及的相关刑事风险时应当坚持有所为和有所不为,即在保证算法安全可控的同时,也应当坚持刑法的谦抑性,以防止刑法触角的过度延伸而阻碍技术发展。实施分级治理策略,构建"过滤"涉类ChatGPT系统犯罪的三道"滤网"是实践上述原则的有效途径:第一道"滤网"即通过类ChatGPT系统研发企业制定与执行自治计划,探索能够有效预防类ChatGPT系统刑事风险的策略;第二道"滤网"即提高行政监管力度,将科技监管手段融入行政监管体系,利用行政手段降低刑事风险实际发生的几率[26];第三道"滤网"即使用刑事手段治理

已经触犯刑事违法性这条“底线”的犯罪行为,刑事治理手段的使用必须严格遵循前置法律法规失效的前提条件,坚持企业自治计划、前置性法律法规治理、刑法规制的三阶治理策略,准确把握刑法介入类 ChatGPT 系统研发企业不当获取数据行为的时机。

参考文献:

[1] 蒲清平,向往. 生成式人工智能:ChatGPT 的变革影响、风险挑战及应对策略[J]. 重庆大学学报(社会科学版),2023(3):102-114.

[2] 朱光辉,王喜文. ChatGPT 的运行模式、关键技术及未来图景[J]. 新疆师范大学学报(哲学社会科学版),2023(4):113-122.

[3] 姚万勤. 对通过新增罪名应对人工智能风险的质疑[J]. 当代法学,2019(3):3-14.

[4] 王迁. 论人工智能生成的内容在著作权法中的定性[J]. 法律科学,2017(5):148-155.

[5] 房慧颖. 论人工智能生成内容的法律保护问题[J]. 山东社会科学,2023(9):179-184.

[6] 刘宪权. 人工智能生成物刑法保护的基础和限度[J]. 华东政法大学学报,2019(6):60-67.

[7] 赵广立. ChatGPT 敲开通用人工智能大门了吗[N]. 中国科学报,2023-02-22(03).

[8] 卡尔·拉伦茨. 法学方法论[M]. 陈爱娥,译. 北京:商务印书馆,2003:279.

[9] 房慧颖. 生成型人工智能的刑事风险与防治策略:以 ChatGPT 为例[J]. 南昌大学学报(人文社会科学版),2023(4):52-61.

[10] 陈兴良. 刑法哲学[M]. 北京:中国政法大学出版社,2004:7.

[11] 平野龙一. 刑法的基础[M]. 黎宏,译. 北京:中国政法大学出版社,2016:90.

[12] 储槐植. 美国刑法[M]. 北京:北京大学出版社,1987:85.

[13] 冯亚东. 理性主义与刑法模式[M]. 北京:中国政法大学出版社,1998:10.

[14] 房慧颖. 侵犯数据产品权利行为的刑法认定[J]. 理论与改革,2024(5):150-160.

[15] 陈兴良. 刑事政策视野中的刑罚结构调整[J]. 法学研究,1998(6):40-54.

[16] 平野龙一. 刑法总论Ⅰ[M]. 有斐阁,1972:72.

[17] 房慧颖. 数字平台治理的“两面性”及刑法介入机制[J]. 华东政法大学学报,2024(5):68-77.

[18] 房慧颖. 数字资产属性的界定及其证成 [J]. 学术月刊,2024(5):113-122.

[19] 姜涛. 刑法如何应对人工智能带来的风险挑战[N]. 检察日报,2019-12-07(03).

[20] 江溯. 自动化决策、刑事司法与算法规制:由卢米斯案引发的思考[J]. 东方法学,2020(3):76-88.

[21] 沈伟伟. 算法透明原则的迷思:算法规制理论的批判[J]. 环球法律评论,2019(6):20-39.

[22] 杨庆峰. ChatGPT:特征分析与伦理考察[N]. 中国社会科学报,2023-03-07(04).

[23] 刘涛. 网络帮助行为刑法不法归责模式:以功能主义为视角[J]. 政治与法律,2020(3):113-124.

[24] 杨东,潘婴东. 区块链带来金融与法律优化[J]. 中国金融,2016(8):25-26.

[25] 房慧颖. 数据犯罪刑法规制模式的系统性研判与立体化构建[J]. 理论与改革,2023(6):78-88.

[26] 房慧颖. 数字经济时代衍生数据财产权的刑法保护机制构建[J]. 广西大学学报(哲学社会科学版),2024(1):194-202.

Criminal risk and governance path of ChatGPT-like system

FANG Huiying

(*School of Criminal Law, East China University of Political Science and Law, Shanghai 200042, P. R. China*)

Abstract: ChatGPT-like system refers to the content-generating artificial intelligence (AIGC) represented by ChatGPT, which combines generative pre-training with algorithmic language conversion. ChatGPT-like system not only promotes the development of human society, but also brings criminal risks. When ChatGPT-like systems obtain corresponding information to generate certain content, it is easy to break away from human

control and intervention and improperly infringe on the confidentiality of other data. At the same time, if the content generated by ChatGPT-like system contains illegal and criminal information, or users improperly use the information generated by ChatGPT-like system, it may also involve corresponding criminal risks. However, ChatGPT-like system does not have the qualification of the subject of criminal responsibility, that is, the subject that may ultimately bear criminal responsibility is not ChatGPT-like system, but the developer or user related to it. In the development and use of ChatGPT-like system, to prevent technical risks and make the technology "good", the developers should fulfill the obligations as guarantors, foresee the possible harmful results caused by ChatGPT-like system as much as possible, and take corresponding countermeasures to prevent them. Users should not intentionally induce ChatGPT-like system to break through the ethical and legal bottom line and output illegal and criminal information. If the developer fails to foresee and prevent ChatGPT-like system from outputting illegal and criminal information to the greatest extent, or the user intentionally induces ChatGPT-like system to output illegal and criminal information, he shall bear corresponding criminal responsibility. From the macro level, when regulating criminal risks involved in ChatGPT-like system, the criminal law principle of modesty should be observed to avoid hindering technological innovation and development. From the micro level, in view of the content and characteristics of criminal risks in ChatGPT-like systems, it is necessary to construct a hierarchical governance mechanism: develop and implement compliance plans through ChatGPT-like system research and development enterprises, and explore preventive governance strategies for the criminal risks; strengthen administrative supervision, and integrate scientific and technological supervision means into administrative supervision system, to reduce the possibility of criminal risks; the use of criminal governance measures must strictly adhere to the precondition that prior legal and regulatory provisions are ineffective, the assessment of criminal illegality must be conducted within the framework of the principle of legality in criminal law, and administrative illegality should not be used to replace criminal illegality.

Key words: ChatGPT; data crime; technology regulation; hierarchical governance

(责任编辑 胡志平)