Doi:10. 11835/j. issn. 1008-5831. fx. 2026. 03. 003

欢迎按以下格式引用:陈伟,向珉希. 生成式人工智能刑事风险下的原因力理论重塑[J]. 重庆大学学报(社会科学版),2026(2):239-252. Doi:10. 11835/j. issn. 1008-5831. fx. 2026. 03. 003.

Citation Format:Chen Wei, Xiang Minxi. Reshaping the causative force theory against the criminal risks of generative artificial intelligence[J]. Journal of Chongqing University (Social Science Edition), 2026(2): 239-252. Doi: 10. 11835/j. issn. 1008-5831. fx. 2026. 03. 003.

生成式人工智能刑事风险下的原因力理论重塑

陈 伟,向珉希

(西南政法大学 法学院,重庆 401120)

摘要:生成式人工智能主体建基于较为复杂的神经网络技术,通过对预训练数据及人类反馈数据的深度学习,在生成过程和终端外显上展现出前所未有的类人属性与不可解释性。刑法视域下,生成式人工智能所引发的风险可以细分为:在没有外界行为介入的情况下由生成式人工智能主体自身创造的内源性刑事风险,以及由外界因素所诱发的外源性刑事风险两个层面。对生成式人工智能之内源性、外源性风险的刑法介入均需要审慎应对,应当在统筹发展和安全的价值权衡中奠定刑法的价值取向。积极刑法观或者消极刑法观的绝对性偏执均有其局限性,而应秉持更为适宜的"适应性刑法观",即在坚守刑法为保障法立场的基础上,反对贸然针对生成式人工智能所诱发之刑事风险进行立法规制,避免因为过度扩大犯罪圈而遏制数字经济时代的发展红利,同时亦要不断关注生成式人工智能的迭代变化,对传统刑法理论进行积极调整以应对生成式人工智能可能导致的现实危害。在工具论与主体论的争鸣中,应明确纯粹工具论与纯粹主体论观念在界定生成式人工智能属性时的失准,现阶段既不应过于夸大其自主性程度,以独立性刑事责任主体视之,亦不能固步自封地按照传统观点将其作为被动工具对待。在"适应性刑法观"的正确引导下,面对传统刑法因果关系理论在应对生成式人工智能刑事风险时所呈现出的"乏力"状态,通过对互动参与之下的引起与被引起关系予以深入剖析,应当强调原因力理论在传统刑法因果关系理论中的重要地位,揭示生成式人工智能刑事风险冲击下、刑法归责进程中原因力关系的具体表现。在此基础上,借助进一步拓宽原因力的接受主体范围、将行为关联纳入原因力判断范畴、区分原因力的程度及种类三个着力点,通过数智时代下的刑法理论之革新塑造,对生成式人工智能的相关触刑风险主体之行为、生成式人工智能之运行以及所造成的危害结果间的因果关系等问题进行适应性

基金项目:2025年度重庆市教委科学技术研究计划重点项目"涉案虚拟货币处置问题研究"(KJZD-K202500301)

作者简介:陈伟,西南政法大学法学院二级教授,博士研究生导师,重庆市新型犯罪研究中心执行主任,新疆克拉玛依市中级人民法院职务犯罪研究中心研究员;向珉希(通信作者),西南政法大学刑法学博士研究生,Email:947857519@qq.com。

的客观解释,奠定生成式人工智能动态发展中的刑法规制之理论基础,妥善应对生成式人工智能带来的刑法归责冲击。

关键词:生成式人工智能;刑法观念;原因力;因果关系;算法黑箱

中图分类号:D924.1 **文献标志码**:A **文章编号**:1008-5831(2026)02-0239-14

党的二十届四中全会通过的《中共中央关于制定国民经济和社会发展第十五个五年规划的建议》,明确将“坚持统筹发展和安全”列为我国“十五五”时期经济社会发展必须遵循的“六个坚持”重大原则之一,提出“在发展中固安全,在安全中谋发展”的重要论断,强调以新安全格局保障新发展格局。当前,生成式人工智能的诞生及发展已成为技术奇异点(Technological Singularity),推动人类社会迈向下一阶段。但科学技术的双刃效应预示着,生成式人工智能注定将在促进社会进步的同时衍生出相应的技术风险。刑法理论对人工智能的传统评价立场,需要伴随数字时代的到来进行重新反思。漠视生成式人工智能的独特规范属性、简单沿用传统人工智能主体分析路径的做法,既无力厘清其生成行为所涉问题的认定困境,更无力为应对相应的刑事风险提供有效的理论方案。有鉴于此,探索生成式人工智能的刑法规制路径,亟待形塑一种超越纯粹工具论的刑法观念,并对涉生成式人工智能犯罪中的归责难题作出契合技术逻辑与规范目的之妥善解释

一、生成式人工智能的刑事风险冲击

现代社会所面临的技术风险,本质上是人类为预测和控制未来而采取的现代手段所引发的各种激进且不可预料的后果[1]。而生成式人工智能正是这种“现代手段”,在赋能人类生产、生活之同时,亦注定伴随重大的技术风险。人工智能或将变革人类社会的基本框架,甚至深入影响人类的心智[2]。生成式人工智能普及后,全社会已逐步意识到并着力应对人工智能发展的恶性化可能[3]。于此背景下,作为最后保障法的刑法无疑肩负着艰巨的社会防卫任务。具言之,刑法视域下,以风险来源作为变量,生成式人工智能所引发的刑事风险可以细分为内源性与外源性两个层面。

(一)生成式人工智能之内源性刑事风险

内源性刑事风险,即指代在没有外界行为介入的情况下,由生成式人工智能主体本身创造的刑事风险。生成式人工智能的知识容量、反应速度于极大程度上归功于其所依托的预训练数据库。以ChatGPT为例,其预训练语料库中的数据涵盖了个人数据、政策文件、新闻报道、文学文本和艺术作品等互联网上可得的各种内容[4]。为保障预训练数据量,避免生成式人工智能无法应对用户提出的问题、要求,生成式人工智能往往被开发者赋予了大规模自动化爬取互联网数据的功能。在自动爬取之数据内容本身合法性、合规性、合理性程度欠缺的情况下,生成式人工智能可能引发一系列内源性刑事风险。

知识产权侵权风险是生成式人工智能刑事风险的首要表现,伴随技术水平之更迭,知识产权侵权犯罪行为具有较为显著的历史阶段性特征[5]。当前阶段,以交互式人工智能为代表的生成式人工智能之运行,建立在对海量训练语料库进行深度学习的基础之上,其内容生成过程,本质上是通过概率预测对既有数据进行重构整合,而非对原作品的简单复制或检索。然而,人工智能生成内容(Artificial Intelligence Generated Content,AIGC)存在“再现”训练数据中受著作权法保护内容的可能性。当模型输出的内容与训练数据中的原作品实质性相似时,便可能触发知识产权侵权风险。

此外,生成式人工智能还具备一系列传播、煽动、传授型犯罪刑事风险。2022年Stability AI公

司对Stable Diffusion开源之后，有不法分子利用其生成了大量的色情图片，并在社交网络中大肆传播[6]。囿于互联网数据内容的复杂性，生成式人工智能在高度自动化地对互联网数据进行爬取的过程中，可能将具有不良引导性的数据内容纳入自身的预训练数据库中，进而对大量用户生成显明或隐含该不良引导性数据的内容。考虑到生成式人工智能所使用的预训练模型是根据人类数据进行训练的，它们很可能继承了人类的偏见甚至暴力性质[7]。在我国刑法视角下，生成式人工智能可能基于上述特性引发相应的传播型犯罪、煽动型犯罪抑或传授犯罪方法罪等刑事风险。

（二）生成式人工智能之外源性刑事风险

外源性刑事风险，即指代由外界因素诱发的刑事风险。与传统人工智能相左，生成式人工智能同时接受来自开发者、使用者两端的输入内容，从而赋能预训练抑或优化训练进程。在此基础上，除生成式人工智能之外，刑事风险可能由生成式人工智能之开发者或使用者引发。

1. 生成式人工智能开发者引发之风险

于开发者端，由其所诱发的刑事风险是显在的。由于能够直接支配生成式人工智能的预训练数据库以及后续的训练进程，并对生成式人工智能之运行过程进行一定程度的监督，开发者很容易对生成式人工智能造成影响[8]。在主观罪过的驱使下，开发者可以向生成式人工智能的预训练数据库中投注具有不良引导属性的数据，此种数据投毒行为，正是经由对数据源的污染，从根本上操控模型的行为模式。生成式人工智能建基于极为复杂的神经网络，经过对投毒数据的循环往复训练后，模型会将这种偏见固化于其庞大的参数结构之中。最终，生成式人工智能得以向用户群体生成明显或隐含该不良引导属性的内容。在此过程中，开发者不仅创设了法所不容许的风险，更违反了结果避免义务[9]。开发者经由不良数据投注行为对大模型预训练数据库之掌控，甚至可能将大模型武器化，使生成式人工智能在与用户的交互过程中持续生成危害内容。而模型内部采取的黑灰箱推理模式，又引发生成式人工智能行为之“不可解释性”，导致对其危害行为难以溯因[10]。因而，应对生成式人工智能危害行为之归因问题，是纾解开发者所引发刑事风险的关键着力点。

2. 生成式人工智能使用者引发之风险

除生成式人工智能开发者外，使用者在交互过程中亦可能诱发刑事风险。以ChatGPT为例，尽管OpenAI并未开源相关模型，外界难以精确掌握其内部运行机制，但现有信息足以表明，用户的对话数据与反馈信息能够对模型的后续输出产生实质性影响。

从人工智能运行逻辑看，用户输入的对话内容虽一般不直接构成对基础模型的实时“永久学习”，但相关数据可能被平台用于后续的模型优化、微调或规则强化，从而在整体上影响模型的输出倾向与内容安全。在此结构下，使用者完全可能通过输入大量违法、诱导性信息，对模型形成持续性的不良引导，进而使模型在面向不特定多数用户时生成危害社会管理秩序、侵害法益的内容，最终引发相应刑事风险[11]。此情景下的“合成内容风险①”异于常态，使用者通过对生成式人工智能的逆向输入，最终促成模型向其他用户生成危害性内容[12]。而由模型生成的虚假信息亦可能通过各种途径回流至训练库，形成恶性循环[13]。

生成式人工智能开发者、使用者在主观共谋的情形下，亦可能合力诱发刑事风险。但此种合力诱发刑事风险本质上由前述两种风险结合而成，不具有解释难度，于此不再赘述。归咎于生成式人工智能所具有的全新特性，由其引发的刑事风险相当棘手，从而需要妥善构思刑法层面的应对路径。从具体的刑事风险应对路径构建逻辑出发，应当首先确立刑法应对观念，进而筛查出现阶段刑

① 合成内容风险，是指生成式人工智能在人机交互过程中，接受用户指令后对用户本人生成违法、有害内容的风险。参见：高艳东《人工智能风险刑事归责的主体选择与分级评价》（《国家检察官学院学报》，2026年第2期149-163页）。

法制度、理论层面的应对困境,最后着手具体应对路径的构建,以妥善解决当前刑法体系对生成式人工智能刑事风险之应对困境。

二、生成式人工智能刑事风险下的刑法观念转型

孕育于社会高速发展进程中的新兴危害行为是否需要犯罪化,深受立法与司法主体所秉持的刑法理念影响[14]。自然,从刑事法的角度应对生成式人工智能所引发的各种风险,理当首先奠定作为根基的刑法观念,确定刑法制度及基本理论应对生成式人工智能刑事风险之态度。

(一)制度观念:"适应性刑法观"应予确立

人工智能时代的法律面临着危机,这种危机表现为法律中人类地位的削弱以及人类意义的瓦解[15]。近年来,面对诸如人工智能犯罪、网络犯罪、环境犯罪等新型犯罪现象,我国刑事立法以积极的姿态予以回应,不断增设新罪,从而使刑法的处罚范围大为扩展,处罚边界逐渐前置[16]。于此基础上,刑法理论界也察觉到,在这种积极刑法观指导下不断增设新罪、提前刑法介入节点的做法存在的现实问题及隐患,尝试从学理上对这种现象进行批判,由此形成了积极刑法观与消极刑法观间的论战。

回归刑法应对人工智能所引发的技术风险所应秉持的立场之上,有学者将现有观点分为"风险防范派"与"技术促进派",其认为"风险防范派"过度强调法益保护,容易导致处在发展阶段的人工智能技术受到冲击;而"技术促进派"则过于注重技术保障,容易引起社会对人工智能的恐慌,对人工智能技术的发展产生逆反效应[17]。若将风险防范派界定为积极刑法观的规范立场,那么技术促进派则可对应消极刑法观的基本主张。

在生成式人工智能蓬勃发展的时代,一方面,生成式人工智能的继续发展之利弊衡量已无法精准预估,故无法确证人类对人工智能技术是否应该秉持积极促进的态度,自然不能完全采纳"技术促进派"之观点。另一方面,诸如生成式人工智能等技术发展所带来的挑战明显地体现在恐怖主义、有组织犯罪、经济犯罪、互联网犯罪以及跨国犯罪中[18]。面对这些挑战,无论是国家的社会治理,还是公民恐惧感的消解,都厚望刑法担当起更大的责任[19]。但是,就生成式人工智能目前所达到的智能程度而言,对其伴随的刑事风险进行规制,我国当前的刑法体系尚未达到可能崩溃的程度,不能完全皈依"风险防范派"。于制度层面应对生成式人工智能所引发的刑事风险,类似于传统的人工智能主体,无论基于何种立场,只是处理路径存在不同,不会产生处罚的困境[20]。在此情况下,面对涌现的触刑风险,在没有正确认识生成式人工智能主体本质的情况下,贸然进行刑事立法应对,不但不会真正解决问题,反而可能引发更多的社会问题。

因此,针对生成式人工智能之刑事风险,更宜采取一种"适应性刑法观",即坚守刑法本位观的立场,在刑事立法层面按兵不动,注重刑法适用的适度性与必要性,反对贸然针对生成式人工智能所诱发之刑事风险进行立法。在此基础上,不断关注生成式人工智能的发展变化,充分运用刑法解释规则。与此同时,通过对传统刑法理论的修正、重塑,周全应对生成式人工智能所引发的刑事风险。这种在刑事立法层面保持稳定,坚守刑法重点,再积极革新传统刑法理论的"适应性刑法观",正是应对生成式人工智能所引发的刑事风险最为恰当的刑法观念。

(二)理论观念:"纯粹工具论"应予扬弃

"工具论"与"主体论"乃相互对立的生成式人工智能刑法地位界定立场。"工具论"即坚守人类

中心主义②的立场，将生成式人工智能在刑法评价中的地位界定为被人类所使用、利用的工具，否定其独立性。支持工具论的学者大都认为，人工智能仍然属于“工具”和“产品”的范畴，人工智能的工具化是犯罪工具进化的必然结果[21]，故而并不需要夸大讨论人工智能的主体属性。“主体论”则认为，完全寄希望于既有法律主体规定的有效性，显然是偏于保守。超越程序设计与代码编制的范畴，具备独立意志的智能主体完全可能出现。这类主体能够形成法律人格的基础，进而成为法律主体。这并非对“人”之概念的颠覆，而是对其内涵的拓展与增补[22]。以生成式人工智能为视角，认为应当扩大刑法中的主体范围，纳入生成式人工智能及今后更加先进的人工智能主体。

1. 纯粹工具论于生成式人工智能刑事风险应对中适用失准

如果以纯粹工具论视角审视生成式人工智能，将其彻底地视为人类主体的工具，则生成式人工智能所实施的根据用户的需求生成文字、图片、视频的活动将不足以被认定为“行为”，而仅是其开发者、使用者行为之延伸[23]。纯粹的工具论视角对于解决生成式人工智能刑事风险而言自然是最简单、直接的方法，在此视角下，传统的刑法理论能够不经修正地完整适用。然而，纯粹工具论并没有真正解决行为人利用动物、无行为能力人、生成式人工智能等具有“独立行为逻辑”或“黑箱属性”的对象实施犯罪行为时的因果关系问题，而毋宁认为是对这类对象所具备特性的选择性失明[24]。由于这类主体具有违背行为人意志的可能性，且无论是动物、无行为能力人抑或是生成式人工智能，其实施行为的逻辑过程都具有复杂性，我们难以精确界定此类主体经由“黑箱”实施的行为是否如同纯粹的工具一样由其操纵者完全掌控，亦无法确定由其做出的行为是否完全由操纵者行为引发。此种“不可解释性”，与生成式人工智能如出一辙。

源自“黑箱”的因果关系认定难题是现实存在的，只是传统刑法理论为了降低犯罪行为评价的难度，对此类因果关系认定的疑难问题进行了回避，但这种“回避”实质上是对自我答责理论的违背。自我答责理论认为，要行为人对结果承担责任，就需要结果是从行为中产生的[25]。但囿于生成式人工智能所具备的独立行为逻辑，其内容生成行为与开发者、使用者行为之间的因果关系具有不确定性，若以纯粹的工具论证开发者、使用者对生成式人工智能危害结果的责任承担，实际上忽视了生成式人工智能本身对其行为的“贡献力”，从而无法证明危害结果是从开发者、使用者行为中完整产生的。

以纯粹工具论的视角对生成式人工智能进行界定，亦会对行为人的刑事责任承担造成不利影响。从程序法视角观之，若以纯粹的工具论界定生成式人工智能之刑法地位，会导致对程序正义之背离。具言之，纯粹工具论视角下，控方仅需证明被告人实施了对生成式人工智能施加影响的行为，即可证成危害结果与该影响行为间的因果关系，由此证成的因果关系是“无瑕”的，被告人不再具备对因果关系问题的抗辩权。若承认生成式人工智能的相对独立性，不再拘泥于纯粹工具论视角的桎梏，则控方仅能通过证明被告人针对生成式人工智能实施的影响行为“推定”因果关系存在，但刑事诉讼程序中的推定是可反驳的，从而赋予被告人反驳该因果关系成立之抗辩权。

此外，在量刑层面，纯粹工具论视角亦可能导致科刑不公的结果。具言之，在纯粹工具论视角下，往生成式人工智能预训练数据库中投放一百字节有害信息与投放一百万字节有害信息的行为人均为工具利用者，若生成式人工智能基于被投入的有害信息生成危害内容，向大量用户传播，则两位利用者的行为与危害结果间均具备因果关系，针对其的量刑结论自然相同或相近。但若摆脱纯粹工具论思维的桎梏，承认生成式人工智能的相对独立性，则足以认识到有害信息字节差异对生

② 人类中心主义，即把人类利益作为价值评价和道德判断之依据，是一种以人类为事物中心的学说。参见：孙道萃《人工智能刑法主体地位的积极论——兼与消极论的答谈》（《重庆大学学报（社会科学版）》，2022年第4期216-229页）。

成式人工智能影响程度的区别,从而作出区别化的量刑结论。当然,从传统量刑理论出发,在归责层面亦可以通过证明不同行为所具有的差异化的社会危害性程度以实现精准量刑。但实质上,对于“投放行为”社会危害性大小的判断仍然要回归至归因层面“一百与一百万字节有害信息对生成式人工智能施加的影响力程度的区别”之判断。所以,以两个行为所具有的不同社会危害性程度作为量刑依据实质上是理论层面的“走弯路”,不如径行摆脱纯粹工具论思维的桎梏,回归生成式人工智能地位之本质。

2. 纯粹主体论于生成式人工智能刑事风险应对中亦陷龃龉

刑法学界一般以弱人工智能(artificial narrow intelligence,ANI)、强人工智能(artificial general intelligence,AGI)区分讨论人工智能主体的刑法主体资格。强人工智能被界定为“智能机器人发展的高级阶段”,其与弱人工智能的本质分野,在于是否具备自主意识与意志,并以此独立决策、自主行动[26]。牛津大学学者尼克·波斯特洛姆(Nick Bostrom)甚至提出了超人工智能(artificial super intelligence,ASI)的概念,认为超人工智能在大多数领域已然超越最聪明的人类主体[27]。但于刑事主体资格范畴下,对行为能力与责任能力仅讨论“有无”而不讨论“强弱”,故超人工智能与强人工智能的界分在刑法上没有意义。何况,若真正出现了比最具智慧的人类还要强大的超人工智能,其不可能遵守更低层次的生物为其设定的行为规则。因而,共识性的分类方式仍是以弱人工智能、强人工智能区分人工智能阶段,以进行刑法主体资格研究③。

当前阶段的生成式人工智能固然代表着人工智能发展的最高水平,但这并不意味着其已完全符合传统刑法理论对刑法主体之要求。相较于具有独立意识和意志的强人工智能而言,目前的生成式人工智能尚不具备自主意识,无法为自己的行为提供目的性指引。于此阶段下,纯粹主体论观念过于超前,缺乏可行性。

综上所述,在制度观念层面,生成式人工智能尚未引发足以撼动传统刑法体系的风险,不应过度先验地以增加刑事立法的方式对其进行规制;在理论观念层面,当前阶段的生成式人工智能主体更像是一种介于弱人工智能与强人工智能之间的先进人工智能形式,不应过分夸大其智能程度,以纯粹主体论观念视之。但亦不能过于贬损,仍完全按照传统观点将其作为弱人工智能对待,以纯粹工具论观念视之。将人工智能主体完全视为犯罪工具的传统刑事法律理论、观点无法合理解释生成式人工智能致害行为的因果关系问题,难以妥善规制生成式人工智能所涉及的触刑风险。

有学者曾认为,人工智能并未对法律基础理论、法学基本教义提出挑战,受到挑战的只是如何将传统知识适用于新的场景[28]。但在生成式人工智能爆发的时代,在确认制度层面尚未乏力的前提下,应对其所带来的刑事风险,症结正好在传统刑法基础理论、基本教义层面。纯粹主体论观念过于超前而无法适用,而纯粹工具论观念过于保守,应予以扬弃。如前述,纯粹工具论观念之核心缺陷在于刑法因果关系的完整建构,构建生成式人工智能刑事风险的应对路径,自然应以因果关系理论为着力点。

三、生成式人工智能刑事风险下的原因力理论倡导

面对生成式人工智能的冲击,应在积极刑法观、消极刑法观的分野中保持中立,在纯粹工具论

③ 以弱人工智能、强人工智能区分人工智能阶段的人工智能刑事主体资格研究模式,参见:刘宪权《人工智能时代刑事责任与刑罚体系的重构》(《政治与法律》,2018年第3期89-99页);王肃之《人工智能犯罪的理论与立法问题初探》(《大连理工大学学报(社会科学版)》,2018年第4期53-63页);陈伟,熊波《人工智能刑事风险的治理逻辑与刑法转向——基于人工智能犯罪与网络犯罪的类型差异》(《学术界》,2018年第9期74-91页);姚万勤《对通过新增罪名应对人工智能风险的质疑》(《当代法学》,2019年第3期3-14页)。

与纯粹主体论的交锋中不偏不倚，但又不能对生成式人工智能引发的刑事风险漠然置之。由此，在刑事立法规制暂不必要的前提之下，应对生成式人工智能刑事风险的路径指向了本质层面刑法基本理论、基本教义的与时俱进，最终落脚于刑法因果关系领域内对于原因力理论的重塑。

（一）原因力理论引入之必要性展开

1. 生成式人工智能凸显传统因果关系理论之适用困境

如前文所述，纯粹工具论观念已无法妥善界定生成式人工智能主体的刑法地位。生成式人工智能能够“由内而外”地独立实施部分行为，该部分具有一定程度“独立性”的行为作为介入因素，于一定程度上侵吞了“因”与“果”之间的关联空间。正如训练动物实施犯罪行为的案件，若将动物视作纯粹的犯罪工具，则行为人之训练行为及动物的攻击行为作为整体的犯罪行为，与危害结果间具有直接因果关系，个中逻辑简单、清晰。但此种理解无疑是忽视了动物作为思维主体的独立性，我们无法证明动物的攻击行为究竟是由行为人的训练行为诱发的，抑或是基于完全内源性的动因，自发实施的攻击行为。生成式人工智能亦是如此，其能够自动地实施一系列行为，亦能够受外部影响实施行为，在黑箱之下，我们无法完全掌握其内部具体运行逻辑，自然亦无法判断其行为的具体诱因，无法完全真实、合理地还原因果关系链条[29]。

在人类中心主义及纯粹工具论的影响下，传统的因果关系理论将除人类外的一切主体排除于因果关系链条的构建之外，囫囵吞枣地将一切危害结果归因、归责于行为人，忽视了动物、生成式人工智能等“不可解释性工具”本身的内源性因素的因果贡献力。此种被忽视的贡献力，可在一定程度上限缩对行为人的归因范围，弱化其行为对危害结果的作用评价，最终在刑事责任层面缩减行为人的归责空间。足见，于生成式人工智能冲击下，传统因果关系理论遭遇困境，若仍一味地按照传统因果关系理论评价今后或将大量涌现的生成式人工智能犯罪问题，何尝不是对生成式人工智能开发者、使用者的“重刑主义”？

2. 生成式人工智能暴露传统因果关系理论之内在缺陷

生成式人工智能之诞生及蓬勃发展对于刑法传统理论最大的冲击即在于因果关系领域。因果关系是将行为与结果结合起来的东西，讨论因果关系的目的，并非阐明自然主义的、非法律意义上的因果关系，而是要从具体行为中切割出构成要件该当、现实的、具体的行为，并通过判断其违法性与有责性，最终导向责任主体[30]。在传统刑法理论体系中，因果关系理论作为根基性理论基础，长久以来存在多个学说间的分野。聚焦于大陆法系，传统刑法因果关系理论主要包括条件说、原因说、相当因果关系说、客观归责理论等[31]。从条件说到客观归责理论，传统因果关系理论各具优势，刑法学界、实务界亦一般根据案件具体情况选择最为合适的理论进行适用。若详析之，足见传统因果关系理论具有一系列共识性的理论基础，而在生成式人工智能冲击下，这些理论基底似乎存在着一定的乏力之处。由于生成式人工智能具备不可探知的算法黑箱，其运行逻辑本就具有相当程度的隐秘性，加之生成式人工智能能够相对独立地实施行为，其又在一定程度上拉开了行为人行为与危害结果间的距离，为刑法因果关系之判断蒙上阴影。概言之，传统因果关系理论应对生成式人工智能所引发之刑事风险主要存在以下两个缺陷或乏力之处。

首先，因果关系之判断仅具“有无”而无“强弱”。在传统的刑法因果关系理论中，无论是作为起源的条件说、原因说，抑或是客观归责理论，其理论运行的最终结果即证明行为与结果之间是否具有因果关系，而并不存在对因果关系程度之判断。相当因果说看似以“相当性”衡量因果关系之强弱问题，但仍仅是以“相当性”作为判断刑法意义上因果关系有无的标准[32]。传统刑法因果关系理论对因果关系强弱判断的轻视曾经具有合理性，毕竟传统刑法因果关系判断仅涉及归因层面，对于

行为与结果之间因果关系之强弱判断则置于归责过程中,通过影响行为的社会危害性进而作用于刑事责任。但在生成式人工智能冲击之下,开发者、使用者行为对生成式人工智能施加危害性影响的程度差异,将更为明显地影响开发者、使用者刑事责任之大小。虽然这种因果关系之强弱最终于归责进程中发挥作用,但不能以某一概念的功能将其本质取而代之,应肯定其本质上属于归因过程中因果关系之判断。由此,可以认为,生成式人工智能之冲击赋予了因果关系强弱判断更为重要的理论地位。

其次,因果关系判断囿于行为与结果之间。条件说、原因说、相当因果说、客观归责理论等传统的刑法因果关系理论无一例外将因果关系限定于危害行为与危害结果之间的引起与被引起关系,但如是理解并没有真正揭示刑法因果关系之本质。在刑法理论中,危害行为与其最终引起的危害结果之间所存在的因果关系能够解决对该危害结果的归因问题,但在教唆犯罪中,教唆行为与实行行为间同样存在引起与被引起关系,教唆行为与实行行为间的关系同样是"因果关系",而这里的"果"是以行为的形式出现。生成式人工智能出现之后,其所引发的刑事风险几乎都来源于开发者、使用者行为对生成式人工智能行为所施加之影响。在此情形下,开发者、使用者行为与最终的危害结果间的距离已经被介入的生成式人工智能行为拉开,因果关系的显著程度亦随之弱化。若仍一味遵从传统刑法因果关系理论,将因果关系之判断囿于行为与结果之间,将架空"开发者、使用者行为对生成式人工智能行为施加之影响"的独立价值,不利于对生成式人工智能所引发之刑事风险进行说理及规制。

言而总之,生成式人工智能之冲击凸显了传统因果关系理论之困境及缺陷,针对这些困境及缺陷,应当回归生成式人工智能开发者、使用者行为与生成式人工智能行为关系之本质——原因力,以包括生成式人工智能本身在内的各主体对于生成式人工智能致害行为所贡献之原因力为线索,厘清归因、归责路径,从原因力理论出发,探寻一条解决生成式人工智能刑事风险规制问题之路径。

(二)生成式人工智能语境下原因力理论之证成

1. 民法体系中原因力理论之渊源考察

我国开始关注原因力,始于民法领域,其理论主要是用来解决侵权法中多因现象下各行为人的责任划分问题[33]。于我国法律规范体系之中,多个法律条文都出现了"原因力"的表述④,然而遗憾的是,没有一个条文对原因力的概念进行了具体阐释,而学界亦尚无定论。有学者认为,"原因力是指违法行为或其他因素对于损害结果发生或扩大所发挥的作用力"[34]。亦有学者认为,"原因力是指行为人的行为在最终损害后果的发生或扩大上所发挥的作用力"[35]。有学者对前述概念进行了细化解释:"原因力并不是指在某因素存在时,结果发生的概率,而是指由于该因素的存在而使结果发生的概率增加的数值,是一个差值。"[36]事实上,在人工智能风险逼近之际,早已有学者意识到应当通过创新因果关系准则的制度技术使规范适应变动的社会,但没有真正触碰到因果关系准则创新的原因力激发点[37]。随着人工智能技术的不断发展,应当重新审视"原因力"概念,强调原因力理论在解决刑法因果关系问题中的重要作用。

2. 刑法体系中原因力理论之重要地位

刑法理论体系中,虽缺乏成熟的理论对原因力概念进行针对性研究,但原因力概念一直在诸多刑法理论中起着穿针引线的作用。

④ 参见:《最高人民法院关于适用〈中华人民共和国民法典〉总则编若干问题的解释》第三十三条、《最高人民法院关于审理国家赔偿案件确定精神损害赔偿责任适用法律若干问题的解释》第九条、《最高人民法院关于审理医疗损害责任纠纷案件适用法律若干问题的解释》第十一条、第十二条、第二十二条。

首先，原因力较为明显地体现于教唆犯中，具体为教唆行为与实行行为的因果关系之中。教唆犯之成立，需要引起被教唆者的行为决意，并使其由此实施犯罪行为[38]。于教唆犯评价中，被教唆者的实行行为之来源分析，亦即教唆行为与实行行为的因果关系判断，系核心问题。教唆犯中的因果关系，表现为教唆行为与实行行为及其所造成的危害结果间的引起与被引起关系[39]。教唆行为与被教唆人的实行行为之间具有诱发关系，这种诱发关系就是因果关系。被教唆的人的实行行为是教唆行为的结果，教唆行为对实行行为具有因果作用[40]。在此基础上，有学者甚至将被教唆人的犯罪意图、实行行为、犯罪结果及其他危害结果均视为教唆行为之结果[41]。事实上，传统刑法理论虽未明确使用原因力概念，却已然在逻辑上默认了其存在，并将教唆行为对被教唆人行为所提供的原因力，作为教唆犯归责的内在根据。

其次，组织犯、帮助犯等其他狭义共犯之因果关系评价也依据行为人原因力的介入，但由于开发者、使用者行为与生成式人工智能行为之间的关系更倾向于一种“诱发”关系，故开发者、使用者的行为更近似于教唆犯抑或精神性的帮助犯，而非提供犯罪工具等客观帮助的帮助犯。事实上，教唆犯、精神性帮助犯对于正犯行为所供给的原因力，与行为人利用动物、无行为能力人、生成式人工智能实施犯罪行为时，行为人对于这些主体所供给的原因力，仅在程度上存在差异，而并无属性上的差异。换言之，直接实施犯罪行为的主体是否具有刑法意义上的自由意志，并不影响这些主体所实施的行为本身的“独立性”，而仅是在独立程度上存在差异，行为人的诱发行为与这些主体所实施的行为之间的原因力本质不受影响，更不应被忽视。

3. 生成式人工智能场景下各主体行为间的原因力关系

（1）开发者行为与生成式人工智能行为间的原因力关系。生成式人工智能开发者，具体包括负责生成式人工智能研发活动的技术主体以及负责生成式人工智能管理活动的管理主体。对生成式人工智能而言，二者的行为可能在普遍表现形式上存在区别，即技术主体对生成式人工智能行为施加影响的行为往往表现为作为形态，而管理主体对生成式人工智能行为施加影响的行为则往往表现为不作为形态。但相对于生成式人工智能使用者而言，二者都是在同一方位为生成式人工智能行为提供“原因力”，所以本文以“开发者”一词概括技术主体与管理主体，对二者行为进行捆绑讨论。

当前阶段，生成式人工智能的开发者通过控制预训练数据库等行为对生成式人工智能行为提供了“原因力”，正是由于这种原因力的存在，生成式人工智能得以实施行为，但也正是基于同样的原因力，生成式人工智能实施的行为仅具有相对而非绝对的独立性。如果说纯粹的工具论视角是对行为人利用动物、无行为能力人、生成式人工智能实施的犯罪行为本质的掩饰，那原因力理论之倡导就是对此类犯罪行为本质的揭露。

在生成式人工智能的运行过程中，开发者对其行为提供的原因力是最为明显且影响最为显著的。在生成式人工智能工作进程中，负责生成式人工智能研发活动的技术主体主要通过对生成式人工智能预训练数据库及预训练、优化训练进程施加影响，从而为生成式人工智能的内容生成行为供给原因力。而负责生成式人工智能管理活动的管理主体则主要通过其负责的管理活动，例如用户个人信息收集、数据安全监管、程序漏洞监管及修复、用户账号监管等，为生成式人工智能的行为供给原因力。

（2）使用者行为与生成式人工智能行为间的原因力关系。在系统对外运行之后，生成式人工智能的学习尚未结束。如前文所述，以ChatGPT为例，在“两端输入”的运行逻辑中，使用者得以通过输入内容影响ChatGPT的优化学习进程，从而诱使ChatGPT对其他用户生成相关内容。在此进程

中,使用者能够通过内容输入行为为ChatGPT的内容输出行为供给原因力。除ChatGPT外,Stable Diffusion、Sora2、豆包等各类生成式人工智能,亦会根据使用者的需求持续优化与迭代学习。对于生成式人工智能而言,基于与使用者的交互内容进行优化学习,是提升其内容生成质量最为直接且高效的路径。也正因如此,生成式人工智能在持续迭代与优化的过程中,必然会吸收并反映来自使用者的行为影响。从刑法因果关系的视角观察,使用者的内容输入行为,实质上为人工智能的后续生成行为提供了现实作用力。据此不难得出,生成式人工智能在运行与生成过程中,必然会接受源自使用者端的原因力供给。

四、生成式人工智能刑事风险下的原因力理论革新

当开发者、使用者对生成式人工智能引发的触刑结果具有主观故意时,其对于生成式人工智能所实施的致害行为所提供的"原因力"就具备了刑法意义。囿于我国刑法理论体系中对原因力概念的系统论述阙如,传统刑法理论在对因果关系进行判断的过程中并没有明确使用原因力概念,即使认为传统的因果关系理论中实质包含了原因力概念,该种所谓的原因力理论也仅停留在各国学者的印象之中而尚未外显。实际上,原因力概念能否用于判断生成式人工智能开发者的刑事责任问题,并不影响开发者、使用者行为与生成式人工智能行为之间实质存在的原因力关系。因为原因力是上述行为间关系的本质,刑法上对原因力采取何种解释方式,仅是对该关系之本质进行把握的一种理论表现。而原因力的理论表现本就不是永恒不变的,随着人工智能技术的不断发展,在再次强调原因力理论的基础上,为拓展原因力理论的适用范围,探究原因力理论的本质内涵,应当考虑对原因力理论进行重塑以拓宽其适用范围。

作为原因力概念运用的一种理论表现,原因自由行为理论能够为找寻原因力理论的重塑方式提供一定的思路。原因自由行为,即"因自身行为招致心神丧失、耗弱状态,并在此状态下引发构成要件结果"的行为[42]。行为人在符合构成要件的结果发生阶段虽然没有意思决定自由,但是在无责任能力状态的原因设定阶段,却具有可以阻止原因设定行为的意思决定的自由[43]。这种"可以阻止原因设定行为的意思决定的自由",正是行为间原因力关系的本质,也正是行为人利用具有独立行为逻辑的对象实施犯罪行为时,对犯罪行为人进行归责的本质原因。然而,传统的原因力理论适用范围较为狭窄,无论是对原因力概念存在较多探讨的民法学界,还是对原因力概念少有研究的刑法学界,都是较为粗浅地从人类主体语境下探讨行为与结果之间存在的原因力,对原因力的程度及种类亦鲜有区分。但是,原因力理论是因果关系判断的重要根基,随着生成式人工智能的普及,仅以人类为原因力理论的研究对象并不完善。所以,为拓展原因力理论的适用范围,探究原因力理论的本质内涵,具体到生成式人工智能开发者、使用者行为与生成式人工智能行为之间的关系问题,应当考虑对原因力理论做以下三个方面的修正。

(一)拓宽原因力的接受主体范围

传统原因力理论将其适用对象限定于具有刑事责任能力的自然人,但此种界定模式过度限制了原因力的接受主体范围,一切能够在独立行为逻辑指引下实施行为的主体都应当被纳入原因力的接受主体范围之内。具有独立行为逻辑的主体不仅包括具有完全或部分行为能力的自然人,也包括动物、无行为能力人、生成式人工智能等具有独立或相对独立的神经网络、原因设定行为人不能完全控制的主体。

于刑法评价层面拓宽原因力的接受主体范围,并不等同于在刑事归责层面拓宽责任主体范围。将前述一系列具有"独立行为逻辑"或称"算法黑箱属性"的非刑事责任主体纳入原因力的接受主体

范围，是由于纯粹工具论在界定上述主体过程中失准，无法精准厘清因果关系的具体内容，仅能借助原因力理论对原因果关系进行精细化拆分，才得以确定行为人的责任范围。申言之，只有先行拓宽原因力的接受主体范围，才能够对涉及前述主体的刑法因果关系进行最为准确的描述。此乃刑法归责评价方法论的进阶优化，而非归责评价立场的更迭。

（二）将行为关联纳入原因力判断范畴

传统刑法视角下的因果关系概念一般限定于危害行为与危害结果之间所具有的引起与被引起关系。以前述教唆犯为例，教唆者并没有实施直接作用于犯罪对象的行为，仅是实施了教唆行为，提供了正犯实施犯罪行为的原因力，从而导致其行为具备可罚性。此时，教唆者与被教唆者之间的关系是行为间关系而非行为与结果的关系，理应用更为准确的原因力概念进行界定。我国有学者以危害结果为视角，将这种“并不直接着力在危害结果之上，往往是借助于第三者因素或者与偶然介入因果场的第三人的行为、受害人的行为、某种自然力量或类似于自然力等因素相结合，才发生了危害结果”的原因力称为间接原因力[32]。这种理解是将原因力概念限定于危害行为与危害结果之间存在的作用力，从而认为其仅是一种介入因素或异常情况。但原因力的本质不应当被限定于危害行为与危害结果之间的作用力，危害行为之间的、危害行为与危害结果之间的原因力都是对其本质的展开，两种情况下的原因力并无二致。应认为，从行为发力到着力点作用在行为或结果的过程中，自然延续且没有任何中断因素介入的完整原因力，都是直接原因力。在刑法理论体系中并不乏“行为间同样存在原因力关系”的佐证，例如教唆行为与实行行为、精神帮助犯的帮助行为与实行行为等。因此，为触及原因力概念的本质，对生成式人工智能所实施的行为提供合理解释，应扩大传统原因力理论的范畴，将行为间的关系纳入其中。

（三）区分原因力的程度及种类

1. 原因力的程度区别

同物理学上“力”的概念一致，原因设定行为与结果行为之间的原因力概念也应当从程度上进行阶梯性的区分，同时细分原因力的不同种类。

具体而言，可以依据原因设定行为导致结果行为的可能性大小为标准，将原因设定行为对结果行为及结果所提供的原因力区分为支配性原因力、高度原因力、中度原因力、低度原因力。支配性原因力是必然会导致结果行为或结果发生的原因力。高度原因力则是高度盖然会导致行为或结果发生的原因力。传统刑法理论之所以会将动物、无行为能力人视为工具，将通过训练、引导等方式诱使动物、无行为能力人做出行为的主体视为直接正犯，便是混淆了支配性原因力与高度原因力的区别，将高度盖然性视为一种必然。中度原因力是导致结果行为或结果发生与否的可能性基本持平的原因力。而低度原因力则指代基本不可能导致结果行为或结果发生的原因力。对于生成式人工智能开发者、使用者而言，原因力的强弱程度会影响其刑事责任大小，原因力越强，行为人的刑事责任程度就越趋近于全责状态；原因力越弱，行为人的刑事责任程度就越趋近于无责状态。

2. 原因力的种类区别

第一，直接原因力与间接原因力。所谓直接原因力，是指从行为发力到着力至结果行为或结果的过程中自然延续，其间没有任何中断因素介入的完整原因力。间接原因力是指原因设定行为与结果行为或结果之间没有直接接续关系，而是通过中介因素对结果行为或结果着力的原因力。如上文所述，原因设定行为所提供原因力的对象既可以是结果，也可以是行为，由此，生成式人工智能开发者或使用者行为所提供的原因力在性质上属于直接原因力与间接原因力的竞合。在将生成式人工智能开发者、使用者行为的原因力供给对象限定于结果的情况下，开发者、使用者行为与危害

结果之间的原因力就属于没有直接接续关系的,由中介因素——“生成式人工智能行为”直接着力危害结果的原因力。以生成式人工智能开发者为例,若开发者在生成式人工智能的预训练数据库中投放色情、淫秽信息,而生成式人工智能在经过对预训练数据库的学习之后,对大量的用户输出各种色情、淫秽信息、图片、视频。此种情况下,开发者对于色情、淫秽物品传播的结果并非直接提供了原因力,而是将生成式人工智能的学习、传播行为作为中介因素直接着力于危害结果。所以,在进行整体性评估之后,生成式人工智能开发者、使用者行为对危害结果所提供的原因力应属于间接原因力。而在将生成式人工智能开发者、使用者行为所提供原因力的对象限定于行为的情况下,开发者、使用者行为对生成式人工智能行为所提供的原因力便是从行为发力到着力点作用在结果的过程中自然延续且其间没有任何中断因素介入的完整原因力。同样以前述生成式人工智能传播淫秽物品案为例,开发者对生成式人工智能预训练数据库实施的属性操控行为,对于生成式人工智能的传播行为所提供的原因力,就是“没有其他中介因素介入”的完整原因力。在分别以生成式人工智能的行为与结果为原因力作用对象时,开发者、使用者的行为所提供的原因力种类分别符合直接原因力与间接原因力的要件。因而,生成式人工智能开发者、使用者行为所提供的原因力,在性质上应属于直接原因力与间接原因力的竞合。

第二,一次性原因力、累积性原因力以及叠加性原因力。一次性原因力,是指行为人实施的单次独立行为,无需借助其他任何因素,即足以引发相对应的结果行为或结果的原因力。而累积性原因力,是指在一定时空范围内,单个行为独立观察均不足以直接造成危害结果,但经由多次行为的持续作用与不断积累,最终共同引发危害结果的原因力。其本质在于,多个单次作用力有限的行为,通过累积叠加形成了足以导致结果发生的完整原因力。开发者在生成式人工智能预训练数据库中添加触刑内容的行为,对生成式人工智能生成触刑内容的行为以及最后的触刑结果所提供的原因力,便类似于此处的累积性原因力。生成式人工智能的预训练数据库所包含的数据量巨大,因此,若生成式人工智能的开发者为躲避相关监管机构、部门监控,每次仅可能往生成式人工智能的预训练数据库中投放少量的触刑内容。在此基础上,要真正使生成式人工智能对用户大量传播触刑内容需要多次、反复的前述行为实施,单次的投放行为远远不足。此时,前行为对于后行为所提供的原因力应当被理解为累积性原因力。而叠加性原因力不同于累积性原因力,后者是由同一主体实施的多次行为所提供的原因力累积而成的,而前者则是由多个主体的行为所提供的原因力叠加后共同提供的原因力。例如,生成式人工智能开发公司中对生成式人工智能负有监督、监管职责的管理主体通过主动监督或接到有关国家机关通知,发现相关用户利用生成式人工智能实施诸如涉嫌编造、故意传播虚假信息罪、诈骗罪等犯罪行为之后,不对相关用户账号进行封禁或不对生成式人工智能的相关程序、系统漏洞进行修复的,其不作为行为与相关用户的作为行为相互结合,对相关危害结果所提供的原因力即应当被理解为叠加性原因力。

结语

在生成式人工智能不断更迭升级的背景下,理性态度决定了仍应秉持“适应性刑法观”,坚守刑法本位观的基本立场,反对贸然针对生成式人工智能所诱发之刑事风险进行立法,避免犯罪圈无限度地扩张蔓延。与此同时,也要不断关注生成式人工智能的发展变化,对传统刑法理论进行积极调整,以应对生成式人工智能所诱发或者导致的刑事风险。在此框架下,原因力理论能够为我们科学认识、界定和规制生成式人工智能的行为奠定理论基础,较好解决当前面临的技术发展与规制困惑。基于此,应当在刑法理论中认可原因力理论的重要性,进一步扩大接受原因力的主体范围,将

行为间关系纳入原因力关系范畴并且区分原因力的程度及种类，注重刑法适用的适度性与必要性，实现传统原因力理论之重塑，审慎考量刑罚介入的限度与强度，妥善应对生成式人工智能在刑法规制层面带来的现实冲击。

参考文献：

[1] 乌尔里希·贝克. 世界风险社会[M]. 吴英姿，孙淑敏，译. 南京：南京大学出版社，2004：4.
[2] 尤瓦尔·赫拉利. 未来简史：从智人到智神[M]. 林俊宏，译. 北京：中信出版集团，2017：296，341.
[3] 梅立润. 人工智能到底存在什么风险：一种类型学的划分[J]. 吉首大学学报（社会科学版），2020(2)：119-127.
[4] Helberger N，Diakopoulos N. ChatGPT and the AI act[J]. Internet Policy Review，2023，12(1)：1-6.
[5] 陈伟，宋坤鹏. “双层空间”背景下侵犯著作权罪的双轨制归责路径探寻：基于“复制中心”与“传播中心”理念的并立选择[J]. 安徽大学学报（哲学社会科学版），2024(3)：99-107.
[6] 陈永伟. 超越ChatGPT：生成式AI的机遇、风险与挑战[J]. 山东大学学报（哲学社会科学版），2023(3)：127-143.
[7] Marche S. The Chatbot problem [J/OL]. (2021-07-23) [2023-03-22]. https://www. newyorker. com/culture/cultural-comment/the-chatbot-problem.
[8] 李飞宏，欧阳本祺. 开源大模型开发者刑事责任减免的法教义学阐释[J]. 江淮论坛，2025(5)：144-155.
[9] 马永强. 生成式人工智能犯罪的理解视域与刑法应对[J]. 国家检察官学院学报，2025(6)：43-64.
[10] 吴世忠. 统筹发展和安全积极应对人工智能治理新挑战[J]. 中国网信，2025(9)：28-32.
[11] 黄锴. 人工智能大模型训练数据的风险类型与法律规制[J]. 政法论丛，2025(1)：23-37.
[12] 高艳东. 人工智能风险刑事归责的主体选择与分级评价[J]. 国家检察官学院学报，2026(2)：149-163.
[13] 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(6)：183-197.
[14] 付玉明. 立法控制与司法平衡：积极刑法观下的刑法修正[J]. 当代法学，2021(5)：15-27.
[15] 韦邦龙. 重拾康德的“超验性”：人工智能法律主体论的再检视[J]. 甘肃政法大学学报，2025(3)：69-82.
[16] 王俊. 积极刑法观的反思与批判[J]. 法学，2022(2)：68-85.
[17] 储陈城. 人工智能时代刑法的立场和功能[J]. 中国刑事法杂志，2018(6)：77-94.
[18] 乌尔里希·齐白. 全球风险社会与信息生活中的刑法：二十一世纪刑法模式的转换[M]. 周遵友，江溯，译. 北京：中国法制出版社，2012：115.
[19] 孙国祥. 新时代刑法发展的基本立场[J]. 法学家，2019(6)：1-14，191.
[20] 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.
[21] 高铭暄，王红. 互联网+人工智能全新时代的刑事风险与犯罪类型化分析[J]. 暨南学报（哲学社会科学版），2018(9)：1-16.
[22] 孙道萃. 人工智能刑法主体地位的积极论：兼与消极论的答谈[J]. 重庆大学学报（社会科学版），2022(4)：216-229.
[23] 刘宪权. 人工智能时代刑法中行为的内涵新解[J]. 中国刑事法杂志，2019(4)：60-72.
[24] 陈伟，向珉希. “培育论”视角下ChatGPT等生成式人工智能刑事主体资格审思[J]. 安徽师范大学学报（社会科学版），2024(5)：96-106.
[25] 冯军. 刑法中的自我答责[J]. 中国法学，2006(3)：93-103.
[26] 刘宪权. 人工智能时代的刑事责任演变：昨天、今天、明天[J]. 法学，2019(1)：79-93.
[27] 尼克·波斯特洛姆. 超级智能：路线图，危险性与应对策略[M]. 张体伟，张玉青，译. 北京：中信出版社，2015：59.
[28] 刘艳红. 人工智能法学研究的反智化批判[J]. 东方法学，2019(5)：119-126.
[29] 陈伟，向珉希. 生成式人工智能背景下犯罪主观评价的异变与应对[J]. 上海法学研究，2023(1)：1-18.
[30] 庄子邦雄. 刑法总论[M]. 新版. 东京：青林书院新社，1981：177.
[31] 张明楷. 外国刑法纲要[M]. 3版. 北京：法律出版社，2020：91.
[32] 陈家林. 外国刑法理论的思潮与流变[M]. 北京：中国人民公安大学出版社，2017：170.
[33] 张训. 论刑法因果关系之原因力[J]. 政治与法律，2010(4)：126-137.
[34] 杨立新，梁清. 原因力的因果关系理论基础及其具体应用[J]. 法学家，2006(6)：101-110.
[35] 张新宝，明俊. 侵权法上的原因力理论研究[J]. 中国法学，2005(2)：92-103.
[36] 范春莹，周植赟. 侵权法“原因力”探析[J]. 华北电力大学学报（社会科学版），2007(2)：59-62.
[37] 劳东燕. 公共政策与风险社会的刑法[J]. 中国社会科学，2007(3)：126-139，206.
[38] 香川达夫. 共犯处罚的根据[M]. 东京：成文堂，1988：37.
[39] 吴振兴. 论教唆犯[M]. 长春：吉林人民出版社，1986：154-155.
[40] 陈兴良. 共同犯罪论[J]. 现代法学，2001(3)：48-57.

[41] 魏东. 教唆犯研究[M]. 北京:中国人民公安大学出版社,2002:138.
[42] 桥爪隆. 刑法总论的核心疑难问题[M]. 东京:有斐阁,2025:248.
[43] 齐文远,刘代华. 论原因上自由行为[J]. 法学家,1998(4):21-24.

Reshaping the causative force theory against the criminal risks of generative artificial intelligence

Chen Wei, Xiang Minxi
(*School of Law, Southwest University of Political Science and Law, Chongqing 401120, P. R. China*)

Abstract: The subject of generative artificial intelligence is built on relatively sophisticated neural network technology. Driven by the deep learning of pre-trained data and human feedback data, it exhibits unprecedented human-like attributes and inexplicability in both its generation process and terminal manifestations. From the perspective of criminal law, the risks posed by generative artificial intelligence can be subdivided into two dimensions: endogenous criminal risks created by generative artificial intelligence subject itself in the absence of external behavioral intervention, and exogenous criminal risks induced by external factors. Criminal law intervention in both the endogenous and exogenous risks of generative artificial intelligence demands a prudent response, and the value orientation of criminal law should be established on the basis of weighing the values of balancing development and security. An absolute bias toward either the positive criminal law view or the negative criminal law view is inherently limited. Instead, we should adhere to the more appropriate adaptive criminal law view, that is, on the premise of upholding the position of criminal law as a guarantee law, we oppose hasty legislative regulation of the criminal risks induced by generative artificial intelligence to avoid curbing the development dividends of the digital economy era due to excessive expansion of the scope of crimes. Meanwhile, we must keep a close watch on the iterative evolution of generative artificial intelligence and actively adjust traditional criminal law theories to address the potential negative harms it may cause. In the debate between instrumentalism and subjectivism, it is necessary to clarify the inaccuracy of pure instrumentalism and pure subjectivism in defining the attributes of generative artificial intelligence. At the current stage, we should neither overstate its degree of autonomy by treating it as an independent subject of criminal liability, nor adhere to traditional views rigidly by regarding it merely as a passive tool. Under the correct guidance of the adaptive criminal law view, facing the inadequacy of the traditional criminal law causality theory in responding to generative artificial intelligence, we should conduct an in-depth analysis of the causal relationships arising from interactive participation, emphasize the pivotal position of the causative force theory in the traditional criminal law causality theory, and reveal the specific manifestations of causative force relationships in the process of criminal law imputation amid the impact of criminal risks posed by generative artificial intelligence. On this basis, by further expanding the scope of subjects bearing causative force, incorporating behavioral correlation into the scope of causative force judgment, and distinguishing the degrees and types of causative force, we should conduct an adaptive and objective interpretation of the causal relationships between the acts of relevant subjects involved in criminal risks of generative artificial intelligence, the operation of generative artificial intelligence, and the harmful consequences thereby caused, through the innovation and reconstruction of criminal law theories in the digital and intelligent era. This lays a theoretical foundation for the dynamic criminal law regulation of generative artificial intelligence and properly addresses the challenges posed by generative artificial intelligence to criminal law imputation.

Key words: generative artificial intelligence; criminal law concept; causative force; causality; algorithmic black box

(责任编辑 胡志平)