涉生成式人工智能犯罪刑事归责的探究

杨海鑫
东北林业大学文法学院，黑龙江 哈尔滨 150040

DOI:10.61369/SE.2026020015

摘　　要： 随着人工智能技术的迅猛发展，其潜藏着的风险亦逐渐显露。当人工智能作为犯罪媒介，使得使用者、技术提供者与人工智能系统本身之间的责任边界模糊，从而导致刑事归责陷入困境。由于人工智能仅具有表象层面的“自主性”，即便对刑事归责所依据的“自由意志”作出类型化阐释，其也无法满足刑事归责所要求的主体性标准，进而不具备刑事主体资格。因此，在人工智能仅能作为犯罪工具的视域下，对于使用者原则上仍应遵循传统“故意犯罪”的认定框架。然而，面对传统单一归责路径在应对技术提供者时的系统性失灵，本文主张构建三阶体系的新型归责模式，以期为构建技术友好型刑事法治框架提供理论支撑。

关 键 词： 人工智能；自由意志；刑事归责

Exploration of Criminal Attribution in Crimes Involving Generative Artificial Intelligence

Yang Haixin
School of Law and Humanities,Northeast Forestry University, Harbin, Heilongjiang 150040

Abstract: With the rapid development of artificial intelligence (AI) technology, its underlying risks are gradually coming to light. When AI serves as a medium for criminal activities, the boundaries of responsibility among users, technology providers, and the AI system itself become blurred, leading to difficulties in criminal attribution. Due to the fact that artificial intelligence only possesses "autonomy" at a superficial level, even if a typological interpretation is made of the "free will" upon which criminal culpability is based, it still cannot meet the subjectivity criteria required for criminal culpability, and thus does not qualify as a criminal subject.Therefore, within the framework where AI is merely viewed as a tool for committing crimes, the principle of adhering to the traditional framework for identifying "intentional crimes" should still apply to users. However, in the face of the systematic failure of traditional single-path attribution methods when dealing with technology providers, this paper advocates for the establishment of a novel three-tier attribution model, aiming to provide theoretical support for the construction of a technology-friendly criminal legal framework.

Keywords: artificial intelligence; free will; criminal attribution

一、问题的提出

以 ChatGPT 为代表的生成式人工智能，凭借海量数据、算法模型和超算能力三大核心要素的迭代升级，展现出了颠覆性的自我学习和自适应能力。在生成式人工智能所引领的新一轮科技革命潮头之下，数据异化、算法黑箱、算力垄断诱发的多元安全风险也在萌发，其风险样态已不单单积聚在基础内里层面，在其外在表现层面，已然出现更多社会伦理、价值、意识形态等风险样态表征[1]。尤需警惕的是，生成式人工智能因其算法的“黑箱”特性，生成内容具有高度不确定性与不可解释性，这为不法分子将其用于新型犯罪提供了技术土壤。犯罪分子利用人工智能的生成能力与交互智能，大幅提升了犯罪行为的精准度与隐蔽性[2]。这一现状揭示了人工智能作为犯罪媒介时的关键性难题：使用者、技术提供者与人工智能系统之间的责任边界模糊不清，致使刑事归责陷入困境。如何合理厘定各参与主体的刑事责任，是当下亟待解决的核心问题。本文即立足于上述问题意识，旨在通过对生成式人工智能犯罪中归责路径的系统探讨，为构建技术友好型刑事法治框架提供理论支撑。

二、自由意志的类型化转向：生成式人工智能刑事主体资格之证否

尽管当前广泛应用的生成式人工智能在整体分类上仍被划入弱人工智能体系，但其已展现出多模态信息融合、动态自主优化等复杂技术特征，凭借海量数据与强大算力形成的“涌现”能力实现了技术性能的实质跃升。这些能力使得生成式人工智能在内

项目信息：2025年度东北林业大学校级大学生创新训练项目，202510225532。
作者简介：杨海鑫（2005.06-），男，汉族，江西省高安市人，法学本科在读，东北林业大学文法学院。

容生成与交互响应中呈现出一定程度的“准自主性”，其算法决策规则隐而不显，研发者对其具体决策机制与输出内容的可预测性也逐渐减弱，甚至会产生算法编写者与控制者都无法解释的内容输出。技术演进与人类认知之间日益扩大的张力，促使法律体系必须重新审视人工智能行为属性的法律界定，以及其在责任分配结构中的应然定位[3]。在此背景下，关于生成式人工智能是否具备法律主体地位的争论主要存在肯定说、否定说和折中说三大立场。这一理论争议不仅反映了对人工智能本质的不同认知，也体现了法律制度在应对技术革命时所面临的价值抉择与体系调适需求。回溯罗马法，“生物意义上的人”与“具备主体资格之人”具有不同的含义。简言之，成为生物意义上的自由人才可能具备主体资格。而在当代语境下，若仍将刑事责任主体的范围固守于具有血肉之躯的自然人，恐难以回应技术发展所带来的规范性需求。刑事责任主体具有时代性特征，其必将随着时代的变化而变化，并非只能局限于自然人等生命体，具有自由意志和实践理性的“非生命体”也完全可以在特定的历史背景下成为刑事责任主体[4]。因此，为探究生成式人工智能的刑事主体问题，必须首先回归至自由意志的本体内涵，对其进行深刻的法哲学剖析，并在此基础上确立刑事归责所要求的主体性标准。

（一）自由意志的内涵分析与“类型化”思维变迁

自由意志，追溯其哲学本源，享有“自由意志之父”盛誉的奥古斯丁认为，其为上帝赋予人类的 “中等之善”，赋予人在永恒之善与可变之善间自主决断的权能，道德行为的正当性源于对上帝永恒法的服从，且真正的自由唯有借助上帝恩典方能实现[5]。而在康德看来，自由意志包括实践理性与自主选择。实践理性指主体根据规范与规则展开理性行为的能力，体现为对行为的逻辑思考和理性控制。自主选择指主体能够独立决策并实施行为的能力，不受外部强制或程序预设的绝对限制，具备摆脱他人控制的自主性[6]。在黑格尔的法哲学体系中，自由意志构成了调节利益关系的法律规范的逻辑起点与核心基石。他主张，法的全部内容都应从自由意志这一原则出发，进行体系性的推演与展开[7]。通过梳理奥古斯丁、康德、黑格尔等思想家关于自由意志的理论阐释可见，西方哲学传统在探讨自由意志问题时，多采用间接论证路径：即依托其在道德、责任与理性行动中的奠基性功能，彰显自由意志的不可或缺性，而极少对其作出精确的概念界定。这种独特的论述策略，一方面在法律层面为将自然人构建为具有自由意志的规范主体提供了深厚的哲学依托；另一方面，也直接导致了自由意志内涵本身难以被单一、封闭的定义所囊括。

由此，我们可以引出一个关键的方法论启示：对自由意志的理解，必须超越传统“概念式”的封闭思维，转而诉诸“类型化”的开放方法[8]。我们不应将其视为一个非此即彼的静态范畴，而应将其理解为一个包含诸要素、具有谱系性特征的整体结构，通过其在具体情境中的典型表现来把握其核心特征与边界。

（二）自由意志的缺失：生成式人工智能主体资格之否定

尽管生成式人工智能凭借其强大的算法模型与海量的数据训练，已展现出一定程度的、形式上的自主性。但是，这种自主性具有深刻的表象特征。从根本上说，现有生成式人工智能的系统架构与目标函数均由研发者预先设定，其根本目的在于实现研发者所期望的生成效果。在整个生成过程中，人工智能必须以外部数据的输入为前提，才能激活并展开其有限度的“思考”。因此，其所呈现出的自主性，在形式上或许是人工智能的，但在实质上，仍是研发者意志与目的的延伸与体现[9]。这种被严格限定在算法框架内的“自主”，并不具备自由意志所要求的、在开放可能性中进行“自主选择”的核心要件。此外，人工智能在运行中持续受到来自使用者指令、社会伦理规范嵌入以及算法价值观对齐等多重外部意志的制约。它无法像人类主体那样，基于内在的信念、欲望与理性反思形成完全独立的判断，因而也未能满足“实践理性”所要求的、能够为自身行为提供理由并承担责任的能力。

即便我们将自由意志的概念进行放宽，作类型化的理解，例如刑法中以“控制能力”与“认知能力”为标准的类型化“自由意志”，生成式人工智能同样无法满足。刑法主体的资格建立在具有自主行为选择与控制能力的基础上，而人工智能的行为根源始终外在于其自身，既无法成为道德主体，也无法成为法律意义上可归责的主体。因此，生成式人工智能在当前及可预见的法律框架下，并不具备承担刑事法律责任的主体资格。

三、技术提供者刑事归责的路径反思与体系重构

现存涉人工智能犯罪的结构中，存在三个关键主体：使用者、生成式人工智能系统本身，以及背后的技术提供者。基于上述分析，目前能够承担刑事责任的主体仅限于作为人类实体的使用者与技术提供者。同时，刘宪权教授指出，智能机器人在设计和编制的程序范围内存在一定的辨认能力和控制能力 ，但这种辨认能力与控制能力服从于研发者的意志，因此，智能机器人在设计和编制的程序范围内实施行为时，应当将其看作人为了实现自己的犯罪意图而使用的“智能工具”[10]。这意味着，人工智能在现阶段并未从根本上改变传统犯罪的构成要件要素，亦未颠覆其所侵害的法益本质。在生成式人工智能仅作为犯罪工具的场景下，对于使用者的刑事责任认定，原则上仍应遵循传统“故意犯罪”的认定框架。故而，本文后续将聚焦于技术提供者的归责路径进行深入的分析。

（一）传统归责路径的三重困局

面对生成式人工智能技术提供者的刑事归责，既有研究形成了共犯归责、不作为犯归责与过失犯归责三条主要路径，但各路径均陷入难以逾越的结构性困境。

共犯归责路径主张将技术供给行为纳入帮助犯框架，却遭遇中立帮助行为可罚边界的根本性质疑。生成式人工智能服务具有广泛的正当用途，依照中立帮助行为理论，单纯的技术供给即便客观上为犯罪提供便利，也不当然具有可罚性，须考察提供者对使用者犯罪意图的具体认知[11]。问题在于，提供者与使用者之间往往不存在任何双向沟通，提供者对特定使用者的存在及其犯罪意图缺乏认知，难以满足帮助故意所要求的“明知”要件。若将抽象的风险认知径直认定为故意，则无异于要求提供者为一切可能的滥用后果负责；若严格限缩故意范围，则又可能使大量发挥帮助作用的技术供给行为逃脱规制。而不作为犯归责路径以刑法

第286条之一为核心建构行刑联动机制，将归责重心从“参与他人犯罪”转向“违反自身义务”，但其法理根基面临保证人地位的质疑。信息网络安全管理义务主要来源于网络安全法等前置法，将这些抽象行为准则径直等同于刑法上的作为义务，有违罪刑法定原则的明确性要求。更根本的障碍在于因果关系认定——算法的黑箱属性使结果回避可能性的判断陷入技术迷雾[12]，若不作审慎处理，不作为犯进路恐将沦为“结果责任”的变种。过失犯归责路径则以新过失论为基础确定注意义务的违反[13]，却面临“风险管辖”的根本性质疑：在提供者的过失行为与使用者的故意行为共同导致法益侵害的场合，根据自我答责原理，使用者对故意犯罪造成的全部后果承担责任，提供者的过失行为因故意行为的介入而阻断归责。

三条路径的各自困境并非偶然，其深层根源在于传统归责模型以“行为—心理”关联为核心范畴，预设了行为主体清晰、因果流程可把握的认知框架，而生成式人工智能的技术系统具有“自主性”与“黑箱”属性，使传统模型陷入结构性错位。与此同时，既有路径通过“意思联络缓和化”“义务泛化”“注意义务抽象化”等概念稀释策略应对困境，不仅难以实现理论融贯，更可能因标准模糊而损害法的安定性。

（二）三阶归责体系的路径重构

针对传统归责路径的困境，真正的出路在于放弃对单一归责路径的迷信，建构以“注意义务分层—算法风险管控—合规计划衔接”为核心的三阶归责体系。

第一阶为注意义务的分层建构，将注意义务内容依据技术类型与应用场景进行类型化区分：设计阶段的义务包括算法公平性审查、数据合法性审核等，核心在于“风险预防”，判断标准应依据技术研发当时行业公认的最佳实践；运行阶段的义务包括内容监测、有害输出过滤等，核心在于“风险管控”，内容具有动态性与情境性；事后阶段的义务包括漏洞修复、应急处置等，核心在于“风险回溯”。通过分层建构，从而在实体层面为行为提供相对明确的判断基准，在程序层面为归责提供分阶审查的可能。

第二阶为算法风险的实质管控，在因果关系判断上采“归因—归责”二阶模式：事实层面的归因判断中，可借助疫学因果关系思维，依据统计学上的关联性认定提供者行为与结果之间的条件关系；规范层面的归责判断中，须进一步考察风险是否在规范保护目的之内、结果是否属于可归责的风险实现。同时引入“风险升高理论”的合理内核——只要提供者的不作为显著升高法益侵害结果发生的风险，且该风险属于注意义务旨在防范的类型，即可肯定因果关系的存在，从而既包容概率性因果，又通过规范判断保持归责范围的适度限缩。

第三阶为合规计划的衔接机制，为提供者提供证明自身已尽注意义务的制度通道：若能证明已建立并有效运行符合行业标准的管理体系，则可推定其已履行注意义务，即便客观上发生法益侵害结果，也不承担刑事责任[14]。这一机制的理论基础在于，在技术风险难以完全消除的背景下，法律的功能不在于强求结果的绝对安全，而在于督促行为人采取合理的风险管控措施。

三阶体系以“注意义务”为核心范畴整合多元理论资源，将共犯、不作为犯、过失犯的理论要素纳入统一分析框架。注意义务的分层设置为司法判断提供明确审查步骤，合规计划的引入为行为人提供可预期的出罪路径，风险升高理论与规范保护目的的结合使因果关系判断兼具灵活性与确定性。这一体系既未过度扩张处罚范围而使技术提供者沦为风险的“保险人”，也未过度限缩归责范围而造成责任缺口的放任，而是通过注意义务的类型化与合规计划的程序性出罪，实现了技术创新激励与法益保护的动态平衡，在技术博弈与规范回应的张力中探寻人工智能时代刑事归责的妥当边界。

四、结语

从自由意志的哲学基础到三阶归责体系，每一个环节都呼唤着法律理念的深刻变革。而三阶层体系的归责路径，只是这一漫长征程的起点。真正的解决方案，有赖于法学界与科技界的持续对话，有赖于立法者与司法者的智慧创新，更有赖于全社会对科技伦理的深刻共识。技术的脚步从未停歇，法律的智慧也当与时俱进。唯有在坚守法治底线与拥抱技术革新之间保持审慎的平衡，我们才能在数字时代的洪流中，既享受科技带来的福祉，又守护人类固有的尊严与价值。

参考文献

[1] 陈嘉鑫，李宝诚 . 风险社会理论视域下生成式人工智能安全风险检视与应对 [J]. 情报杂志 ,2025,44(01):128-135+171.

[2] 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.

[3] 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.

[4] 刘宪权 . 生成式人工智能的发展与刑事责任能力的生成 [J]. 法学论坛，2024，39 (02)：18-28.

[5] 胡万年 . 奥古斯丁自由意志概念的形而上维度——兼与康德自由意志的比较 [J]. 安徽大学学报 (哲学社会科学版)，2008，(05)：40-45.

[6] 卞绍斌 . 责任与自由：康德实践理性概念的规范性阐释 [J]. 贵州大学学报 (社会科学版)，2023，41 (02)：1-12.

[7] 熊光清 . 法的本质：从法与国家及社会关系层面进行的审视——再读马克思《黑格尔法哲学批判》[J]. 马克思主义研究，2023，(12)：120-130.

[8] 李涛 . AI 实体刑事责任主体论的规范证成 [J]. 中国刑警学院学报，2022，(03)：43-52.

[9] 程承坪 . 论人工智能的自主性 [J]. 上海交通大学学报 (哲学社会科学版),2022,30(01):43-51.

[10] 刘宪权，胡荷佳 . 论人工智能时代智能机器人的刑事责任能力 [J]. 法学 ,2018,(01):40-47.

[11] 黎森予 . 生成式人工智能服务提供者中立参与行为的处罚范围 [J]. 苏州大学学报 (法学版)，2024，11 (04)：40-52.

[12] 张维尧 . 生成式人工智能生产者过失犯罪的结果归责 [J]. 华东政法大学学报，2025，28 (06)：75-85.

[13] 袁彬，薛力铭 . 生成式人工智能的刑法规制问题研究 [J]. 河北法学，2024，42 (02)：140-159. .

[14] 陈小彪，尹杰辉 . 生成式人工智能虚假信息犯罪风险分类与分层治理 [J]. 昆明理工大学学报 (社会科学版)，2026，26 (01)：1-11.

