Open Journal of Legal Science 法学, 2025, 13(3), 514-518
Published Online March 2025 in Hans. https://www.hanspub.org/journal/ojls
https://doi.org/10.12677/ojls.2025.133073

Hans汉斯

利用生成式人工智能诈骗的责任主体认定

王鹏远

北京师范大学法学院，北京

收稿日期：2025年2月17日；录用日期：2025年2月27日；发布日期：2025年3月21日

摘　要

在为人类社会带来福祉及革命性变化的同时，以Sora为代表的一系列生成式人工智能技术也存在可能被滥用于进行“深度伪造”的潜在风险。如在某些情况下，不法分子可能会利用生成式人工智能进行换脸诈骗等违法犯罪活动。若如此，则生成式人工智能的出现将对诈骗类犯罪的司法认定产生重大影响。生成式人工智能的研发者、网络服务提供者应当符合社会期待，在享受生成式人工智能带来的利益的同时，对可能存在的诈骗风险应当承担相关的责任。

关键词

生成式人工智能，诈骗罪，责任主体认定

Identification of Responsible Subject of Fraud Using Generative Artificial Intelligence

Pengyuan Wang

Law School of Beijing Normal University, Beijing

Received: Feb. 17th, 2025; accepted: Feb. 27th, 2025; published: Mar. 21st, 2025

Abstract

While bringing welfare and revolutionary changes to human society, a series of generative artificial intelligence technologies represented by Sora also have the potential risk of being abused for “deep fake”. For example, in some cases, criminals may use generative artificial intelligence to carry out illegal and criminal activities such as face-changing fraud. If so, the emergence of generative artificial intelligence will have a significant impact on the judicial identification of fraud crimes. Generative AI developers and network service providers should meet the expectations of society, and while

文章引用: 王鹏远. 利用生成式人工智能诈骗的责任主体认定[J]. 法学, 2025, 13(3): 514-518.
DOI: 10.12677/ojls.2025.133073

enjoying the benefits brought by generative AI, they should bear the relevant responsibilities for possible fraud risks.

Keywords

Generative Artificial Intelligence, Fraud, Responsible Subject Identification

1. 风险责任的认定

当代社会学的研究表明，经由其系统本身制造的危险，工业社会身不由己地突变为风险社会[1]。生成式人工智能时代的到来很好地验证了这一点。生成式人工智能因其“算法黑箱”的存在，导致其生成的内容具有不可控性，由此造成了种种不可预测的风险。对于现代社会而言，法律的目标不可能是消灭风险，而只能是控制不想要的风险，并设法将风险进行公平地分配[2]。而面对生成式人工智能造成的风险，存在着因果关系链条和责任归属的判断等诸多难题。生成式人工智能的运转涉及了诸多相关主体，对风险的分配存在着问题。

对创设风险的管辖，是基于这个原则：任何一个对事实发生进行支配的人，都必须对此答责，并担保没有人会因为该事实发生而遭到损害，支配的另一面就是答责。按照这一原则，任何人都必须安排好他自己的行为活动空间，从这个行为活动空间中不得输出对他人的利益的任何危险[3]。当出现人工智能导致的损害时，对答责的要求就意味着，应确保存在适当的责任机制[4]。根据雅克布斯的说法，责任仅仅是“一般预防的衍生品”，应把责任理解为“训练法忠诚”以及“获得整体的规范信任”。对于有无责任的判断至关重要的是，是否社会因不法行为产生了明显的法忠诚的缺失。而通过执行与规范相关的刑事处罚来维护规范，以回应行为人对法不忠诚的行为，进而从一般预防的角度巩固规范有效性。

据雅克布斯教授的理论而言，应当赋予涉人工智能相关主体可谴责性以实现积极一般预防的刑罚效果，进而达到维护规范效力的目的，以实现刑法规范对于人工智能风险的防控[5]。通过责任归属的方式，来确认作为期待“集合体”的规范的效力。责任概念的内容即是根据它自身的任务来确定，即巩固社会的行为期待[6]。而对于涉人工智能相关主体责任的认定，则应符合社会的期待。

为了有效地解决责任问题，我们必须首先确定个人理解该行为可能会造成伤害的能力，并且根据个人对社区的义务，选择另一种行为是更安全的选择。因此，人工智能系统的相关人员的责任取决于他们理解人工智能系统行为模式的能力[7]。对参与人工智能系统的设计、开发、监控或使用的每个人应该有不同的期望，这取决于他们的确切角色和任务[8]。生成式人工智能研发者和网络服务平台提供者具有一定的专业和技术壁垒，这是一般公民所不能具备的。其作为风险的掌控者和实际利益分配的受益者，其行为应当符合社会的期待，确保由规范所保障的信赖，对可能涉及的涉人工智能诈骗等犯罪行为承担风险责任。当“人工智能的使用应保证安全”的社会期待进入了规范系统，上述主体应在管辖范围内承担积极义务，如未尽义务致使期待落空，则需要将结果归属于他。

2. 研发者的责任

生成式人工智能的研发者作为信息技术领域和人工智能技术领域的专业人士，是智能时代社会风险

源头的有力影响者甚至是实际控制者[9]。研发者作为算法的编写者，从源头上规定着人工智能的行为，并影响着人工智能后续的深度学习和为使用者提供服务的方式，因此研发者应承担高度的注意义务。当不法分子利用生成式人工智能进行诈骗犯罪时，若研发者对此未能履行相关义务，导致危害结果发生，则研发者需要承担过失责任。

“为了成立过失，违反法律上认为必要的注意义务是必要的”[10]，注意义务的违反是过失犯的本质，研发者构成过失犯罪的前提便是违反注意义务。有关过失犯罪的理论有旧过失论、新过失论、新新过失论三种理论。

首先，传统的旧过失论可能对人工智能的技术创新产生制约，并且“对特定行业的工作人员特别不利”[11]。传统的旧过失论以“结果预见可能性”为核心归责要件，即行为人若可预见到行为可能导致的危害后果，即可认定其主观过失。然而，人工智能等前沿技术的研发具有显著的探索性与不确定性，若严格适用该理论，研发者将面临不可控的过失责任风险。这种情形可能抑制技术探索的积极性，阻碍人工智能领域的突破性进展。

其次，新新过失论与现代的“被允许的风险”理论存在冲突，同样不利于技术革新。根据“被允许的风险”理论，社会在接纳人工智能技术效能的同时，应当包容其伴随的不可预知风险[12]。而新新过失论提出的“抽象的结果预见义务”因缺乏具体判断标准，仅要求行为人对潜在危害存在模糊不安感(故被称为“危惧感说”)，其适用可能导致研发责任前置化[13]。“抽象的结果预见义务”要比旧过失论中“结果预见义务”更加提前，该学说极有可能对那些尚处在研制阶段还没有面世的技术造成打击，哪怕这些技术只存有非常小的社会风险，其研发者也必须要对此风险承担过失责任。相较于传统旧过失论，该学说的归责范围更广，或将严重压缩人工智能技术的创新空间。

最后根据新过失论，注意义务的判断包括结果预见义务和结果回避义务。其中，结果预见义务是指具体的结果预见义务，而不是抽象的结果预见义务或危惧感；结果回避义务的履行应当具有社会相当性，是注意义务的核心判断标准。新过失论在追究过失责任的同时，也不会将过失处罚范围划定地过于宽泛[14]。因此，在判定生成式人工智能产品的研发者是否违背了注意义务方面，新过失论具有可取之处。

具体而言，在不法分子利用生成式人工智能进行诈骗等犯罪行为时，对研发者的过失责任的判断应当以结果回避义务的履行为核心[15]。“结果回避义务实际上是一种基准行为，如果根据一般人标准，只要采取了对一般人而言具有合理性的结果回避义务行为即基准行为，即便具有预见可能性，由此出现的结果属于被允许的危险。”[16]对于生成式人工智能研发者结果回避义务的判断，主要应当落脚于在预见生成式人工智能责任事故风险的前提下，提供、开发人工智能的行为是否符合行业的基本准则。当面对不法分子利用生成式人工智能进行诈骗犯罪时，应当首先根据科技发展水平来判断研发者对这种危害结果是否有预见，在此基础上再结合生成式人工智能法律法规，行业行为准则等判断研发者是否履行了结果回避义务[17]。若研发者在训练语言模型、设计人工智能算法规则、强化打分模型等阶段已经尽到最大的义务，完全遵守了研发部署人工智能的一系列基准行为规范，但仍然无法避免不法分子利用人工智能进行诈骗等犯罪行为，则不能认为研发者违反了结果回避义务，研发者不应当对此承担过失责任。反之，如果人工智能研发者在模型训练、算法编写过程中未能尽职尽责，对人工智能可能被非法利用的方式缺乏评估、预测与预判，致使诈骗等犯罪行为发生的，其就应当对他人利用该人工智能实施的犯罪行为承担过失责任。另一方面，技术的发展也伴随着风险。在风险社会，诸多风险虽可以被预见但无法被完全避免。在某些情况下，研发者能够预见可能会发生的风险，但是却没有避免该风险的可能性。此时，根据信赖原则，即使不法分子利用人工智能实施了诈骗等一系列犯罪行为，但因为研发者对此不具备避免该情形的可能性，便不应当承担相关的过失责任[17]。

3. 网络服务提供者的责任

除了研发者需要承担相关责任之外，网络服务提供者在不法分子利用生成式人工智能进行诈骗的活动中也有着重要影响。网络服务提供者一般是指为他人提供互联网接入、服务器托管、网络存储、通讯传输等技术支持的网络连接服务商或网络平台服务商[18]。在网络空间中，网络服务提供者的正常业务活动看似是网络中立行为，实则蕴藏着风险。若不法分子利用人工智能生成的虚假信息在相关网络平台上发布捏造的视频、进行诈骗等一系列的违法犯罪活动，网络服务提供者对此类行为具有一定的掌控能力，应履行和其经营收益所匹配的监管义务。当其违反了相关义务，对其是否应当承担责任的认定上，客观归责理论具有借鉴意义。

根据客观归责理论，责任的判断需经过三个阶段：1) 行为人的行为是否制造了法律不允许的风险；2) 该风险是否实现；3) 因果流程是否在构成要件的效力范围内。只要任何一个阶段被否定，则行为人就不必为自己造成的结果担责。客观归责理论作为一种较为严格的因果关系判断方法，对于有效规制中立行为刑事责任边界的作用是积极的，它不但能够有效防止中立行为入罪的边界无限扩张，同时也能防止对处罚边界的不当限缩[19]。

在不法分子利用生成式人工智能在网络平台进行诈骗的情形中，首先需要判断行为是否制造了法律所不容许的风险。如果网络中立行为没有制造风险，或者这种风险被法律所允许，则该中立行为不构成犯罪。具体而言，某些网络服务提供者的正常业务行为可能被不法分子利用，进而为不法分子实施诈骗活动提供帮助。但根据利益平衡原则，按照一般的社会共识，若上述网络服务提供者提供给社会的价值和便利远远高于犯罪活动造成的各类损失，则难以认定这种中立活动造成的风险为法律所禁止。

其次，即使行为的确造成了法律所禁止的风险，然而该风险与结果之间并无常态上的关联性(风险未实现)，则该行为仍不能认定为犯罪。具体而言，如果网络服务提供者的中立行为造成了涉人工智能诈骗犯罪的结果，而这属于正常因果联系范畴，一般人能够对其有足够的预测，则可认定符合客观归责理论的第二重考量；反之如果这种结果的发生属于偶然情形，一般人对其都没有足够的预测，则网络服务提供者不具有客观归责可能性。

最后，在判断构成要件的效力范围时，“回溯禁止”对于网络中立行为的评价有着特殊作用[20]。回溯禁止一般是指行为人实施行为之后，因为其他人故意行为的介入而导致结果的发生，则先前行为者就不必为后来的结果负责[21]。德国学者雅各布斯从该理论入手，认为只要中立行为存在独立的社会意义，便禁止将正犯行为及其结果溯及中立行为，故只能由正犯独自承担责任。只有中立行为的帮助导致正犯行为造成侵害结果的可能性极大，才构成帮助犯。如果网络服务提供者的中立行为与不法分子诈骗行为的关系具有密切的相关性，则这种中立行为就丧失了独立的社会意义，那么就可认为该中立行为的因果流程在构成要件的射程之内，反之则不构成犯罪。

4. 结论

刑法的目的不仅在于打击犯罪，而是要为社会的发展保驾护航。若要想社会正常有序地发展，便不能偏于一隅。面对新生事物的出现及其可能带来的相关风险，对刑法的动用更应该慎之又慎。在面对涉生成式人工智能诈骗的相关主体认定上，应符合相应的社会期待。研发者在研发过程中若没有履行应尽的义务，应当承担过失责任；网络服务提供者若未能尽到相关的监管义务，则也应当对于诈骗行为的发生承担责任。只有相关主体在生成式人工智能的研发、使用过程中符合相关规定，才能尽可能地避免风险发生，才能真正为新兴科学技术的发展提供保障。

参考文献

[1] 乌尔里西·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: 102.

[2] 劳东燕. 风险分配与刑法归责: 因果关系理论的反思[J]. 政法论坛, 2010, 28(6): 28.

[3] 乌尔斯·金德霍伊泽尔. 刑法总论教科书[M]. 蔡桂生, 译. 北京: 北京大学出版社, 2015: 101.

[4] 埃里克·希尔根多夫. 数字化、人工智能和刑法[M]. 江渊, 刘畅, 译. 北京: 北京大学出版社, 2023: 221-222.

[5] 赵书霖. 人工智能刑事责任的功能主义形塑[J]. 时代法学, 2024, 22(4): 63-78.

[6] 王钰. 功能刑法与责任原则围绕雅科布斯和罗克辛理论的展开[J]. 中外法学, 2019, 31(4): 1050-1074.

[7] Osmani, N. (2020) The Complexity of Criminal Liability of AI Systems. *Masaryk University Journal of Law and Technology*, **14**, 53-82. https://doi.org/10.5817/mujlt2020-1-3

[8] Sachoulidou, A. (2024) AI Systems and Criminal Liability. *Oslo Law Review*, **11**, 1-10. https://doi.org/10.18261/olr.11.1.3

[9] 房慧颖. 人工智能犯罪刑事责任归属与认定的教义学展开[J]. 山东社会科学, 2022(4): 142-148.

[10] 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: 234-235.

[11] 周光权. 刑法总论[M]. 第 4 版. 北京: 中国人民大学出版社, 2021: 161.

[12] 西田典之. 日本刑法总论[M]. 刘明祥, 译. 北京: 法律出版社, 2013: 104.

[13] 前田雅英. 刑法总论讲义[M]. 曾文科, 译. 北京: 北京大学出版社, 2017: 183.

[14] 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(3): 69-79.

[15] 刘宪权, 房慧颖. 涉生成式人工智能犯罪研究[M]. 上海: 上海人民出版社, 2024: 93.

[16] 陈兴良. 教义刑法学[M]. 北京: 中国人民大学出版社, 2017: 511.

[17] 许钟灵, 吴情树. 人工智能体过失刑事风险的因应[J]. 法治社会, 2020(6): 93-103.

[18] 周光权. 网络服务商的刑事责任范围[J]. 中国法律评论, 2015(2): 175-178.

[19] 毛建中, 董彬. 论网络中立行为的刑罚规制范围——以客观归责的判断方法[C]//高艳东, 连斌. 网络空间的秩序与责任——第二届互联网法律大会论文集. 杭州: 浙江大学出版社, 2017: 24-32.

[20] 孙运梁. 构成要件的效力范围——客观归责理论构成规则研究[J]. 湖南社会科学, 2012(5): 93-96.

[21] 黄荣坚. 刑法的极限[M]. 台北: 元照出版有限公司, 1999: 145.