生成式人工智能犯罪归责的开发者中心主义*

叶竹盛　林曼婷

（华南理工大学法学院，广东 广州　350001）

[摘 要]刑法介入智能犯罪的治理因应风险防范和安全控制的需求。生成式人工智能主体资格应予否定，人工智能系统本身无法成为刑事责任主体。不论未来人工智能技术发展程度如何，开发者作为风险源的创造者应当为风险现实化的危害结果担责，也最有能力控制危险源以避免犯罪。开发者责任始终是刑法规制人工智能风险的核心，通过建构以开发者为中心的刑事责任体系，有力规范生成式人工智能的行为。生成式人工智能的开发是指包括产品化研发及其投入应用的全过程，为人工智能系统创造了运行条件并引入市场的人类主体统称为“开发者”。由于开发生成式人工智能具有促进社会进步的技术创新价值，刑事归责应当平衡好安全保障和促进创新的需求，避免归责的盲目和刑罚的越界。容许风险理论则能兼顾二者，通过将开发者的技术创新区分为刑法不容许的风险行为和刑法容许的风险行为两类存在实质性差异的基础归责形态，以公正的风险分配理念指引开发者归责。开发者违背基本伦理准则或违反刑法规范而开发智能系统的，属于不容许的风险范畴，为刑法禁止并依据有关条款追究刑事责任。开发活动在风险容许范围内，但智能产品仍可能存在剩余风险，开发者可能对危害后果承担刑事责任，通常构成产品刑事责任或监督管理过失责任。

[关键词]生成式人工智能　开发者责任　容许风险　刑事责任　智能犯罪

[中图分类号] D924　[文献标识码] A　[文章编号]2096-983X（2024）05-0134-14

一、问题的提出

近年来，大语言模型的横空出世开启了生成式人工智能的新时代。2023年3月，Open AI发布了大语言模型ChatGPT-4，取代了四个月前才发布的3.5版本。GPT模型向人们展示了前所未见的强大的内容生成能力，且在短期内迅速迭代，引爆了用户体验热情的同时，也引发了各行各业有关技术风险的忧虑。刑法学界近年来也一直在讨论人工智能时代刑法面临的新问题及其应对方案：现有的刑法工具是否足以应对人工智能产生的新风险和制造的法益侵害？现有刑法理论能否妥当回应人工智能在刑事主体、行为类型、犯罪形态、责任承担方式等方面提出的问题？这些讨论仍处于萌发阶段，尚无理论共识，更未形成新的妥当理论，也未在实证法层面上有所确立。但生成式人工智能已登上实践舞台，使得刑法的技术境遇进一步复杂化。

收稿日期：2024-05-20；**修回日期**：2024-07-04

***基金项目**：教育部人文社会科学研究规划基金项目“中国网络社会治理的法治模式研究”（22YJA820028）；广州市哲学社会科学规划项目“网络直播平台推广活动的刑事风险与合规研究”（2022GZGJ199）

作者简介：叶竹盛，刑法学副教授，广东地方法制研究中心执行主任，主要从事刑法和数字法治研究；林曼婷，刑法学硕士研究生，主要从事数字刑法研究。

生成式人工智能区别于其他人工智能的特征是，这类系统通过对大量训练数据进行机器学习，生成内部的特有决策规则，进而对外部提示做出反应，输出与训练数据相似但也不失新颖性的内容，形式包括文字、图形、语音、视频、代码和数据等。生成式人工智能不只是行为人用于实施特定行为或者制作内容的工具，其“生成式”的能力在外观上表现出“自主性”决策和创作的特征，并可能结合其他技术工具，产生在客观上具有刑事意义的“动作举止”。近期，浙江破获了一起利用ChatGPT生成虚假视频后在互联网传播的案件。该案的行为人仅有初中文化水平，却借助ChatGPT在短视频平台制造了数千个虚假视频在网络散布。[1]这些案件中的行为人如果进一步部署自动化程序，接入生成式人工智能，预先设计实施犯罪的步骤和方案，设置收取违法所得的账户，就可能长期自动实施无需人员干预的犯罪活动。面对伴随生成式人工智能而来的新形态犯罪，刑法应如何应对？

一种便捷的思路是考虑是否将人工智能系统主体化，由系统承担刑事责任。人工智能刑事主体资格的讨论肇始于人工智能技术发展之初期，早期的讨论更多停留在理论设想甚至是纯粹想象的阶段，并未做好规范上正式承认的准备。但是进入生成式人工智能时代，一些论者认为生成式人工智能预示着人工智能技术已经或是将要发展到必须考虑是否承认人工智能主体资格的阶段[2]，认为有效治理人工智能犯罪应当认真考虑是否承认人工智能的刑事主体资格，直接以人工智能主体为归罪对象，将其纳入刑法的责任体系，即使不在当下实现，也应当接纳未来出现的强人工智能为刑事主体[3]。此种肯定说深信“人工智能的奇点”必然来临，随着GPT模型的能力日新月异而愈发萌动。肯定说遭遇了经验否定说和彻底否定说两种较为强烈的反对。经验否定说提出，现有的人工智能技术充其量是一种自动化的系统，完全没有达到人类意识能力所要求的自主性，并且在经验上现有的人工智能技术所涉及的犯罪，都可以在现有的刑法框架下予以应对，不必直接将人工智能系统作为归责对象。[4]彻底否定说确信不可能产生具有人类心智的人工智能，因此无需也不该考虑人工智能应否成为刑事主体的问题。[5]

本文主张应当否定生成式人工智能的刑事主体资格，治理人工智能刑事危害应当以人工智能系统的开发者作为归责对象。不论未来人工智能技术发展程度如何，开发者责任始终是刑法规制人工智能风险的核心。一方面通过容许性风险界限，划定开发者的开发行为的界限，就像限制人类基因编辑行为一样，限制开发特定类型或特定功能的人工智能；另一方面则通过对容许范围内的开发者设定刑事责任，督促开发者解释、召回、修订、销毁、锁死“有害”的人工智能系统，有效规制剩余风险。人类刑法必须驯服人工智能，但不是像驯服人类一样的驯服人工智能，而是通过驯服人类（开发者）以驯服人工智能。

二、以开发者为中心的生成式人工智能犯罪归责路径

生成式人工智能无疑将为社会带来新的刑事风险，刑法的目光应该聚焦在那些将人工智能系统引入人类社会的开发者们。通过开发者的刑事责任体系，规范生成式人工智能的行为，治理生成式人工智能可能产生的刑事风险。责任的本质是一种期待可能性，并且是一种规范性的期待可能性。只要我们同时还坚持这样的期待，即人工智能应用安全应该能够得到保障，并及时遏制犯罪风险，那么这里的期待指向的便是那些负责对智能系统进行设计、编程以及投入运营的人。[6]以开发者归责为中心更能实现刑罚在智能犯罪领域的预防功能。刑罚的正当化根据是报应的正当性与预防犯罪目的的合理性[7]，且刑罚不允许超过由罪责原则划定的边界，在此边界内考虑作为刑罚目的的预防意义。[8]开发者对其开发的生成式人工智能的风险不利益承担刑罚体现了责任刑的正义性；基

于预防生成式人工智能犯罪的合理性，要求开发者承担一定的预防刑，体现了预防刑的功利性。由于开发者具有突出的预防犯罪能力，对开发者苛以预防刑更能获得刑罚对于不特定群体的一般预防和特定个体的特殊预防效果。一般预防承载着刑罚面向公众的学习效果、信赖效果、满足效果和确证效果，因为生成式人工智能的犯罪行为选择完全可以由其开发者进行安全控制技术手段予以矫正，与之功能相似且具有法益侵害可能性的开发者可从中获得直接的威慑、教育和改造，以此有力地预防生成式人工智能应用市场的犯罪风险。本文将那些为人工智能系统创造了运行条件的人类主体统称为“开发者”。刑法只能通过对开发者赋予规范性期待，以实现对人工智能的事实性期待。

（一）以开发者责任为中心的刑事风险治理

生成式人工智能的技术本质是开发者输入的规模化数据和深度学习模型。穿透技术面纱体现的仍是开发者的意志与价值选择。在当前技术水平下的绝大部分生成式人工智能犯罪，是由于智能系统被投入市场或社会、面向不特定公众的应用风险而引发的，开发人员作为风险源的创造者最有可能为危害结果担责，也最有能力控制危险源以避免犯罪，本文基于此逻辑将开发者置于归责中心。

1. 开发者的范围

应当说明，本文界定的开发者区别于研究阶段的科研人员。本文认为“开发”是指将生成式人工智能这一智能模型产品化的过程，并将生成式人工智能产品引入市场或社会的行为，区别于研究阶段的科研行为。是否以“投入应用”为目的是开发与研究的区分界限。生成式人工智能的研究是指在实验室内围绕生成式人工智能进行的基础研究。基础研究要求遵循相应的监管要求和科研伦理准则；在此阶段，除非发生严重危及人身、财产和公共安全的行为，刑法作为保障法和最后制裁手段，不应干预符合伦理准则的智能技术研究。科研人员对生成式人工智能的各类研究发生在实验室内，因技术风险所引发的法律风险较小；且相关研究不以投入社会或市场应用为目的，对生成式人工智能投入应用后造成的法益损害在客观上不存在结果惹起关系，科研人员对此不具有刑法非难性。本文所称的生成式人工智能的开发是指包括产品化研发及其投入应用的全过程。当生成式人工智能在此阶段因其应用风险所造成的法益侵害后果，在民事上由危险产品的生产者承担一定的责任。这源于民法上的危险责任，旨在考虑社会对技术创新的需求，但并不将损害风险强加给遭受人工智能系统不可预见后果的随机的受害人，而是将损害风险强加给那些从创新中获得经济利益的人。[3]显然，民事经济手段不足以督促开发者努力提升生成式人工智能公共和市场应用的安全水平。随着生成式人工智能市场应用规模的不断扩大，刑法若对生成式人工智能引发的现实风险持观望立场，只会纵容技术利益支配者对公众安全的漠视，将对社会的整体安全保障构成威胁，也不利于人工智能技术实现深远发展。因此，为督促开发者夯实安全保障设施，刑法应当介入生成式人工智能开发阶段的风险预防，要求开发者对生成式人工智能所造成的严重不利益承担刑事责任。

统言之，在开发阶段，责任主体涉及参与产品投入市场、运营维护和系统升级迭代的全开发过程的相关人员。具体而言，包括生成式人工智能系统的第一手开发者即创建生成式人工智能系统源代码的技术人员，基于源代码和模型进行删改、增添等操作的后续开发人员，以及对生成式人工智能系统进行产品化的开发人员。这部分人员都是现实制造、支配和控制着生成式人工智能产品应用风险的人。生成式人工智能投入应用而产生的现实风险离不开开发者的技术支持。因此，开发者对此所引发的法益侵害结果在客观事实层面起到了技术促进效果，开发者的开发行为与生成式人工智能相关犯罪具有关联性，构成开发者刑事归责的事实根据。

2. 专业用户的二次开发者责任

用户违法滥用生成式人工智能产品且因此

实现了法益侵害后果的，应当追究其相应的刑事责任。用户的滥用行为可以区分为两类，一是恶意滥用AI（malicious abuse of AI），指以违法犯罪意图来操纵AI技术的行为；二是恶意使用AI（malicious use of AI），指对AI技术的违法中立性使用并造成了危害后果。[9]不管是对技术本身的恶意滥用还是对技术的违法中立性使用，在此场景下，生成式人工智能都是用户的违法犯罪工具，用户对其滥用技术的行为结果承担刑事责任，符合责任主义原则。

值得注意的是，部分用户责任可以为开发者责任涵摄。由于用户使用行为的技术性特征不同，用户可以进一步界分为专业用户和普通用户。专业用户的使用行为是指对已投入应用的生成式人工智能系统进行技术调整，即在原产品基础上进行删改、增添而使其用于新目的或产生新功能的行为，并再次向市场或社会开放使用的行为。实际上，这一部分专业用户对生成式人工智能的技术操作行为是再开发行为，若因生成式人工智能新功能之实现造成危害结果的，专业用户应当认定为生成式人工智能的二次开发者进行归责。例如，据日本媒体报道，专家对ChatGPT进行调查发现，若输入伪装成开发者的指令，ChatGPT就可以生成能用于网络犯罪的电脑病毒；即使OpenAI公司事先为ChatGPT设置了限制恶意使用的指令，但用户仍可以通过输入特殊指令规避。[10]在该案例中，这类用户实际上通过输入特殊指令伪装成开发者的手段对生成式人工智能系统进行了与开发行为性质一样的算法调整，使得生成式人工智能突破了原开发者预置的限制指令而实现了非法功能，即对生成式人工智能系统进行了再开发，该类用户应当依照开发者路径追究刑事责任。

在追究用户责任之余，不论面向专业用户还是普通用户，开发者也应当为可以预见的用户滥用后果承担一定的责任。用户滥用生成式人工智能并造成法益损害后果的，如果属于开发者在开发过程中可预见的结果范畴，可以对开发者成立过失不法并进而归责。

3. 风险责任原则的贯彻

诚如前述，虽然用户在特定情形下应当独立进行归责，但用户责任并不是生成式人工智能犯罪风险治理的关键，核心力量应归属于那些拥有知识、技术和经济支配能力的开发者群体，他们构成了这一领域安全应用的坚实支柱。生成式人工智能犯罪归责的重心面向开发者，这符合风险责任和风险预防原则。现代化的负面影响是风险和潜在的自我威胁的释放[11]，人工智能应用给现代社会带来福祉的同时也带来新的社会安全风险。正是在风险社会的发展背景下，现代刑法形成了风险责任原则，即风险由谁管控，谁就要对风险及其实现的后果承担责任。[12] “对创设风险的这种管辖，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。”[13]生成式人工智能开发者不仅是将生成式人工智能风险源引入市场或社会的责任人，开发者还能利用智能技术和资本权力，控制生成式人工智能系统本身的制造安全，直接管理应用中的生成式人工智能的实时安全水平。相较于其他所有主体，诸如政府机构、非盈利组织等，均不具备与之相媲美的强大且直接的风险监管力与掌控力。而开发者，作为生成式人工智能从创制到应用全链条的技术核心，其承担的技术安全风险责任需要与其风险管理能力相称。这是贯彻风险责任原则的应然之策，将其风险责任转嫁给用户，只会损害社会各方的技术安全信心和发展利益。风险预防是当前社会风险控制机制的重要方式，面对不确定的风险，即便技术行为和技术风险之间具有抽象性，也应当基于预测采取事前预防措施。生成式人工智能的技术风险在开发和应用阶段逐步形成并实现，开发者熟知生成式人工智能技术开发的全过程，对其应用风险具有超越其他一般人的预见可能性。开发者还掌握对生成式

人工智能应用安全风险支配管理能力，不仅在应然上承担风险管理责任，还必须在实然上采取各项风险预防措施，如严厉的风险评估、持续的监控机制和有效的应急处理措施，最大程度地减少和控制生成式人工智能技术可能带来的各类风险。

综合而言，考虑到开发者对投入市场应用的生成式人工智能控制力以及作为“生成端”到“用户端”的中间管理人，强调由生成式人工智能开发者承担风险控制的主体责任，是一种更符合技术安全风险分配原则并且可预防之期待性更高的治理路径，如同互联网平台被视为“数字守门人”的理论隐喻一般，开发者应当成为人工智能守门人，对生成式人工智能的应用过程承担把守关口的安全保障义务。为了稳固这种义务，人工智能刑事不法的归责也应当以开发者为中心。

（二）开发者归责体系的正当性基础

风险社会的现代性特征即是“收益与风险并存”的客观规律，功能强大的生成式人工智能产品也无法脱离该规律。在确保技术应用的安全性不被削弱的基础上，如何实现最大程度的技术创新达到最高限度，始终是法学界致力于论证与剖析的研究议题。[14]由于刑法只是治理技术风险的一种必要性手段，考虑到刑罚的严重后果[15]，针对开发者的归责，需要一套归责理论的指引，反思应当怎样将公平的风险分配思想贯彻到归责体系的构建中[16]，在应用安全和技术创新的博弈中寻得均衡点，避免归责的盲目和刑罚的越界。

1. 刑法的创新容许风险理论

在目的理性的刑法体系中，归责不再仅仅由因果关系和主观态度等存在论要素构成，而是必须考虑刑事政策上的预防目的与处罚必要性。技术创新是一个特殊的行为，基于其对社会有用性的考虑，刑法不能完全禁止创新行为。但是，为保障社会公共安全，也不能完全容许各类伴随风险的创新活动，而应对具有现实危险的创新行为施以体现预防和制裁功能的刑罚。这要求刑法的干预应当遵循必要性原则，换言之，并非所有的科技创新行为都需要通过刑法来进行规制，而是要划分不同的风险等级，以区分不同等级的法律法规来应对。只有在存在严重的危害社会安全的风险时，才由刑法干预。

一般被认为用于限定过失犯罪成立的“容许风险”理论 ，亦称被允许的危险，在处理创新风险行为的归责中意义重大。容许风险理论的创建，正是为了平衡法益保护与现代社会发展之间的利益竞争。容许风险是指随着科学技术的发展，社会生活中不可避免地存在的具有侵害法益的危险行为，基于其对社会的有用性，即使发生了法益侵害结果，也应当在一定范围内允许。[17]那么，一个容许的技术创新行为在概念上就已经不可能辜负法律上的行为期待，因而在刑法层面，被容许的风险行为也就不需要在刑法上确证规范的效力（Bestatigung der Normgeltung）。[6]关键是，在生成式人工智能开发创新领域，容许的判断标准何在？

一方面，在具体划定生成式人工智能开发这类技术创新行为的刑法容许边界时，可以借鉴当前国内外对AI开发的理念或原则，这构成了人类社会对智能技术创新可接受和可信赖范围的基本共识。2020年欧盟首次发布《人工智能白皮书》，提出了构建以人类利益与福祉为目标的人工智能生态系统[18]，其继受了2019年欧盟人工智能高级别专家组对可信赖的（trustworthy）人工智能伦理准则的七项具体要求建议：其中“人的能动性和人的监督”原则要求保留人对人工智能的决策权和控制权，包括至少在原则上理解人工智能所控制的进程的可能性；“技术的鲁棒性和安全性”原则指向技术稳健性，要求智能系统在设计上应力求避免造成对他人的损害，在专家组看来，这也应包括对抗外来攻击方面的安全性；“隐私保护和数据质量管理”不仅指对信息自决权的保护，还包括对全部“自己”数据的控制；“透明性”是指人工智能工作方式和工作结果上的可解释性，而“透明性”又可以被视为多数“可问责性”形式的前

提，但是在没有透明性的情况下还是可能成立纯粹的无过错责任。此外，“多样性、非歧视性和公平性”至少包括确保排除不公平的、具有偏见的人工智能；“社会和环境福祉”中则包括对整个社会、环境以及其他有情感能力的存在的保护。最后，对“可问责性”的要求意味着，当出现由人工智能导致的损害时，应确保存在适当的责任机制。[19]

当前各国在AI开发准则和理念上虽有差异，但也形成了部分共识，包括透明原则、安全可控原则、无害公平原则（包括保障权益、消除偏见和歧视和尊重隐私和自主）、可归责原则等。我国2019年发布的新一代人工智能治理原则也提出了基本相同的AI治理原则，包括公平公正、尊重隐私、安全可控、共同担责等，同样明确人工智能发展应以增进人类共同福祉为目标，符合人类的价值观和伦理道德。[20]

在生成式人工智能面向社会投入使用的开发中，这些原则构成开发者技术创新的行动指南。虽然存在一定的潜在风险，但如果开发者的开发行动践行了上述原则，表明其技术创新采取了开发原则内含的价值追求，从而在主观认识上有意抵御技术创新所蕴含的危险。

行为规范意义上，之所以将上述共同的开发理念准则视为创新风险的容许标准之一，是因为其反映了社会对风险行为的法律非难性共识，也即符合开发原则的智能产品开发根据社会的相当性观念而视为阻却客观危害的事由[21]，属于刑法容许的风险范围内。社会的相当性，是指在社会生活中，历史所形成的社会伦理程序所允许的行为，这一理论在动态的平衡中考虑“违法与社会生活的关系”。因而，违法的标准不是单纯地看法益是否受到侵害。对于那些从静止、绝对的观点来看似乎是侵害法益，但从动态、相对的观点来看则是社会的相当性行为，并不认为违法；只有那些超越了社会相当性的行为，才能视为违法。[22]具体到现代化生产领域，社会相当性意指在一定方式上“正常的”风险应被接受，如果开发者的创新行为在遵循透明原则、安全可控原则、无害公平原则、可归责原则的前提下有序开展，即使其潜在的风险被现实化，也可以在一定程度上被视为技术创新的容许风险。例如，OpenAI在发布GPT-4时，提供了详细的技术报告，解释了模型的训练数据和算法原理，尽力展示了透明性和可靠性；同时，开发者需要保障生成内容不会包含种族、性别等方面的歧视或受偏见影响，以体现无害公平性。反之，如果开发者对生成式人工智能的开发本着与上述原则相悖的理念或反伦理意图，比如开发特定类型试图攻击人类自主决策权的超级智能，则属于风险禁止领域，这是不为法秩序所容许的。

另一方面，在满足社会相当性观念的开发原则之余，创新风险之容许边界判断还应当经过法益权衡。基于被容许风险的理论框架，应当对潜在危险行为所引发的风险与法益保护所带来的利益进行深度权衡，并考量冒险的必要性。在利益性显著超越危险性，或两者达到均衡状态的情况下，相应的危险行为应当被允许实施，以确保行为与后果之间的合理性与正当性。[23]换言之，容许风险理论倡导理性冒险，在创新行为之技术风险性和利益性之间进行比较，利益性大于等于危险性则容许该创新，其目标是追求更高的社会利益。风险和利益的比较本质上是刑法的法益权衡过程。法益权衡理论认为，刑法乃以保护法益为其目的，在刑法所保护之各种不同法益间发生冲突时，牺牲低价值法益，而保全高价值法益；或在不同之法义务发生冲突时，违反低价值义务，而履行高价值义务，则以此等方式而解决法益冲突或义务冲突，应为法规范所容许者，此即所谓之法益权衡理论。[24]面对技术创新风险，单凭风险与容许度呈负相关、利益与容许度呈正相关的关系模型进行简单归纳是困难的。好比自动驾驶汽车中的生命对生命的可衡量性问题，这些刑法理论的古典问题在智能时代以新形态出现，但至今仍无定论。即便如此，法益权衡理论仍然提供了一些审查创新风险容许边界的指引。

其一，利弊衡量。必须对创新行为进行利益性与不利益性及其利益价值的比较，只有当利益大于或者等于不利益时，创新行为才可能被允许。[25]比如说开发某类生成式人工智能产品虽能获得可观的经济利益，但是却以牺牲不特定个人的人身安全为前提，则开发此类存在严重损害人身安全的生成式人工智能属不容许的风险范围。其二，必要性审查。如果创新行为的利益价值并非必须，则不被容许；同时风险之实现与创新风险行为之间应当相称，如果采取风险更小的行为即可达到该利益目标，则采用风险较大的行为以实现利益属于非容许。其三，社会安全风险评估。无论如何，技术创新行为始终应以维护公共安全为宗旨，优先保护公众的人身与财产安全。刑法作为社会保障法，如果技术创新行为对社会公共安全存有严重威胁的，该创新风险就不应当被容许。目前有科学研究证实生成式人工智能模型可用于增强基因编辑的精准度，一旦生成式人工智能用于人类基因编辑，则其技术危害更是难以预测和控制，在开发该类生成式人工智能时相比起冒进创新，更应持保守克制立场。

2. 风险控制与风险分配

法律只能致力于控制不可欲的社会风险，在不同利益主体间进行风险的公平分配，但绝不可能以风险消灭为控制目标。[25]开发者的刑事责任体系同样首先应当以公正地分配风险为前提，激活刑法归责的风险控制和预防功能，以此实现刑法自身对风险社会安全的法治保障价值。如前所述，生成式人工智能市场应用以开发者责任为中心，开发者是生成式人工智能技术应用风险的主要责任主体。在此基础上，风险分配应当作两个层次的区分，一是刑法作为行为规范的风险分配，二是刑法作为裁判规范，对风险导致的损害结果的责任分配。

对于第一个层次的分配，刑法作为行为规范应当告知开发者那些刑法处罚实质侵害法益的行为。此类禁止规范可以根据刑法的法益保护追求，通过危险的犯罪化、安全义务的规范化加以实现。对于开发者的禁止规范，可以划分为伦理性禁止和规范性禁止。关于伦理禁区，当然并非所有违背社会伦理的开发行为都应当为刑法所禁止。只有那些与刑法的制裁严厉性相称，突破了生命伦理和人作为人的自主性被剥夺的开发活动才为刑法所禁止。比如开发超级生成式人工智能，追求智能系统具有超出人类一般的道德水平和判断能力，无异于要求生成式人工智能积极干预人类社会的道德生活。此种激进的创新行为将会极其严重地威胁到人类基本的生存伦理，属于伦理禁区。关于规范性禁止，是根据现行的刑法规定，将结合了新技术的违法开发行为予以禁止，例如开发专门用于违法犯罪的生成式人工智能，或开发为刑法所禁止的特定功能如生成虚假信息的生成式人工智能。此外，还有大量的风险会在生成式人工智能系统被投入使用的过程中现实化，根据风险管辖原理：谁若单独或者与他人共同制造了某一客观上提升了构成要件结果发生可能性的情境，谁原则上就应当为此风险及其所导致的结果负责。[16]基于对开发者风险管辖能力的考虑，开发者应当承担更广泛的安全管理义务，意味着刑法将禁止开发者从事不利于智能系统应用安全的行为，如禁止生成式人工智能算法开源。一旦开源，表明开发者主动放弃了对其开发的生成式人工智能的算法风险进行直接控制和管理的权限，不特定的社会公众可对此进行无限制再开发，技术风险因之蔓延。可以说，这两类禁止规范，本质上都属于危险犯和法定犯。

对于第二个层次的风险分配，实质上属于风险致损后果的结果归责范畴。如果说第一个层次意在目的性地排除非法风险，第二个层次的风险分配目的是处理好技术产品的剩余风险。生成式人工智能技术风险的现实化普遍发生于投入应用过程中，风险致损后，对于归责而言，应当考虑行为人对于危险创设和实现的支配力和避免危险结果发生之可能性。其一，在对危险创设和实现的支配力上，值得注意的是，生成式人工智能这类智能产品不同于传统

产品。智能产品对开发者的危险支配力是通过后端的算法技术等持续、稳定存在的。换言之，即便产品流入市场也不会截断开发人对潜在风险的掌握权。那么，在生成式人工智能进入流通之后，引发了在先未能预见到的危险后果，开发者应当及时发挥其对危险源的支配和管理能力，限制生成式人工智能产品的使用端口，否则仍应当构成刑法上的规范违反或义务违反。其二，技术应用致损后果的避免可能性，这直接与开发者过失不法关联。根据信赖原则，在信赖和期待产品开发者会履行义务的前提下，可以通过审查注意义务违反关联、期待可能性等要素判断能否在刑法上谴责生成式人工智能开发者未避免风险结果的事实。在既有技术水平的范畴内，生成式人工智能所引发的危害结果，应被视作一种难以规避的残余风险的具体显现，此等风险为刑法体系所容纳，且不应将责任归咎于任何个体，其本质上是创新局限性的体现。可见得，作为归责原则的容许风险理论，在广义上指向消除了刑法答责的禁止性，同时最起码在过失犯客观归责的审查上，也是让行为人对避免风险结果的无能力不负责任。

适用被容许的风险应聚焦于特定案件情境下的考量，刑法上的风险是具体风险，“风险的法律属性自始与行为人是否现实地具备结果避免能力不可分离”[26]，并非泛泛地就一般技术应用风险而论。

三、生成式人工智能开发者刑事归责体系的展开

如上节所述，基于容许风险理论和风险分配的逻辑，开发者归责体系可以分别在两个不同性质的层面上展开，一是排除在容许风险范围外的禁止规范所导致的刑事责任；二是在风险容许范围内的可开发行为所产生的刑事责任。

（一）生成式人工智能开发的禁止规范

即便目前生成式人工智能已崭露优异的、自动化的学习能力，其本质上仍是受制于智能系统算法的产物。其代码输入、数据喂养等预置环节对于生成式人工智能的输出结果被认为具有关联性、预定性，而建模等输入环节的算法只可能由开发者预先设定。为遏制潜在风险，开发者即应当被禁止开发违背伦理准则或有悖刑法规范期待的特定类型的生成式人工智能，主要可以划分为目的违法禁止、功能违法禁止、技术失控禁止、合规违反禁止四种不容许类型。

首先，目的违法禁止，是指开发者以实施犯罪为目的或协助犯罪为目的而开发生成式人工智能，该行为属于技术目的违法禁止。如果开发者以实施犯罪为目的开发生成式人工智能，这种行为显然属于故意犯罪。例如，某开发者开发了一款生成式人工智能，用于自动生成钓鱼网站，尽管其声称目的是测试网络安全，但事实上该生成式人工智能被广泛用于实施诈骗，则开发者的行为仍可构成诈骗罪。此外，与他人共谋，由开发者设计特定的生成式人工智能并提供给使用者实施相关犯罪的，则按照共同犯罪追究其相应的刑事责任。例如，生成式人工智能开发者与使用者通谋，由其开发能够生成“钓鱼”邮件、诈骗“脚本”的生成式人工智能产品，再提供给使用者使用，由使用者利用该产品实施诈骗行为。在这种情况下，产品开发者和使用者成立共同犯罪，开发者提供特定生成式人工智能的行为属于诈骗罪正犯的帮助行为；当然，如果所开发的生成式人工智能的技术效果对诈骗结果的发生起重要作用，按照实质客观说中的“重要作用说”，生成式人工智能开发者可能构成共同犯罪中的正犯。[27]

其次，功能违法禁止，指开发者所开发的生成式人工智能产品功能本身为刑法所不容许的范畴，比如专门开发用于生成淫秽内容的系统。功能违法禁止涉及生成式人工智能产品功能本身的刑事违法性，这与犯罪的客观方面相关。如果生成式人工智能本身的功能即是构成刑法规范之违反的，那么其开发和使用无疑具有可责性。例如，某生成式人工智能被开发用于破解他人密码，尽管开发者声称其开发目的是帮助用

户恢复遗失的密码，但由于其功能涉及未经授权的访问和篡改数据，开发者由于开发具有违法功能的生成式人工智能，可以追究其非法获取计算机信息系统数据、破坏计算机信息系统的刑事责任。

再者，不可控的生成式人工智能开发禁止，指开发者在无法说明算法可解释性或保障算法透明度的情况下，仍然发布了生成式人工智能产品。如果开发者将无法说明和解释的生成式人工智能引入市场，行为不仅涉及注意义务违反，还直接构成了对刑法禁止规范的具体违反。刑法禁止规范的核心在于防止行为人实施对社会公共安全和个人权益造成严重危害的行为，不能保证算法透明度的生成式人工智能正是这种潜在危害的具体表现。如果开发者在设计和开发生成式人工智能时，未能遵守必要的安全规范和标准，这个行为本身将承担缺陷产品生产制造的刑事责任；如果开发者明知其设计的生成式人工智能存在可解释性的缺陷，可能在特定情境下失控并造成严重后果，但仍然选择继续开发并投入市场应用，同样构成对禁止规范的违反，因为此时开发者持有一种间接故意，其预见危害结果但持有放任态度。例如，某公司开发的生成式人工智能用于自动驾驶，但开发者在开发过程中已经意识到该生成式人工智能在复杂的交通环境中所作出的决策和判断无法说明和解释，可能会失控并导致严重交通事故的发生。尽管如此，开发者依然决定将该结合了负责自动驾驶决策的生成式人工智能推向市场，在这种情况下，可以根据其引发的危害后果追究开发者故意犯罪的责任。

最后，合规违反禁止涉及功能上具有中立性，但实际上主要被用于违法犯罪行为的生成式人工智能，其开发者如果在客观上无法采取合规措施确保其技术产品的合规使用，则具有刑事可责性。比如，区块链的虚拟币本身是中立的技术，但由于缺乏有效的监管，往往被用于洗钱或其他非法交易。那么假如某生成式人工智能被开发用于匿名交易，但实际上主要被用于洗钱活动。尽管该生成式人工智能本身具有技术中立性，但由于开发者无法实施有效的合规措施，导致该本具有中立性的技术实际上主要被用于非法活动。此时，虽然开发者对生成式人工智能的算法预置并非专门指向违法犯罪活动，但是在技术上促进了违法犯罪行为，也未尽到对生成式人工智能相应的合规管理。如此可以追究其相应犯罪活动的刑事责任，或被认定为帮助信息网络犯罪活动罪。其中，对开发者是否构成相关违法犯罪活动的“明知”不能一概而论，而应当结合具体个案进行综合认定。帮助犯的成立，需要技术服务提供者对产品被利用于违法犯罪具有超越一般可能性的认识和容许，[28]无可否定，智能系统正常的业务行为都会存有为他人犯罪提供帮助的可能性，开发者对这种“用于违法犯罪活动”的情形仅存有轮廓性的泛知，则不足以达到主观明知的程度；但若开发者对生成式人工智能被用于违法犯罪活动已经形成相对的具体认知，已经超越一般可能性的认知和容许范畴，则可能构成“明知”。这同样与网络安全刑事治理观相契合，即将“明知”这一要素限定为一种相对具体的认知，防止将并非追求不法目的的正常业务行为纳入刑事惩治范围[29]，从而实现技术创新与安全保障之间的平衡。对于相对的“具体认知”的判断标准，当开发者已经发现产品被用于违法犯罪活动后仍然没有采取有效的措施限制该技术的非法使用，可以期待被认定为“对其技术被用于违法犯罪”的相对具体认知。

（二）容许开发范围内的开发者责任

生成式人工智能的技术风险绝不可能通过任何手段抹杀，意味着风险始终是剩余的，在产品流通应用过程中始终存在风险现实化的可能。这也意味着，即便开发者在容许风险范围内开发了特定生成式人工智能，当生成式人工智能风险现实化时，根据风险管辖及责任分配原则，开发者作为生成式人工智能应用风险的主要引起者，是首要的风险防范主体，应当对引起的风险结果承担刑事责任，主要以严格责任和

过失责任的形态进行归责。并且，开发者仍对产品持有稳定、强大的技术管理支配力。与该能力相称的合理期待是，开发者承担着生成式人工智能开发至应用全过程的风险防范责任，重点是对生成式人工智能算法安全和合规应用的管理责任。

1. 排除适用绝对的严格责任

生成式人工智能犯罪危害后果离不开开发者生产制造、投入使用和运营维护的全过程的技术支持。换言之，生成式人工智能的危害结果与开发者行为存在事实因果关系，即便其开发行为不违背禁止规范而被刑法容许。因而，开发者对生成式人工智能应用致损的归责是在肯定了危害行为与危害结果之间存在因果关系之后，确定可以将结果归属于行为的基础上，探讨能否让行为人为此承担刑事责任。[30]对此，关键在于行为人在实施行为时是否具有主观罪过（故意或者过失），以及罪过的特殊形态即严格责任。严格责任原则源自英美刑法，指在没有罪过的部分场合也可以将行为定性为犯罪，并对行为人追究刑事责任。[31]这一责任形式强调对特定行为结果的客观归责，而不考虑行为人的主观认识。严格责任确立的初衷是使个人利益向公共利益的让步，以更好地保障社会生活秩序和安全。在技术发展与风险共生的智能社会，尤其集中于具有算法黑箱特性的智能犯罪领域，对于技术创新行为能否适用严格责任成为一个重要议题。

智能科技导致风险的发生具有不确定性、不可预知性。如果要求先行行为必须违背客观注意义务、要求行为人对于先行行为所伴随的风险具有客观预见可能性，则必然导致新型风险的制造者即人工智能开发者可以不必为其技术所创设或提升的风险承担责任，而将风险及其责任转移至用户乃至智能系统本身，这显然是不公平的。对于开发者归责适用严格责任原则确实强化了对公共安全的保护，但严格责任原则的危险在于极大地限制了技术创新和发展的进程。实际上，是否适用严格责任理论也是刑法如何平衡风险规制和创新自由的理论分支问题。应当注意，是否适用严格责任还应该明确的前提是：严格责任到底有多严。根据严格责任的严厉程度，可以区分为绝对的严格责任和相对的严格责任：绝对的严格责任是只要存在法定的行为或造成了法定的结果，法院就可定罪处罚的；相对的严格责任没有全然摒弃主观责任之维度，被告可以援引“无过失”等辩护理由，属推定的过错责任。[32]开发者作为生成式人工智能风险的创设者和监控人，理所应当地具有更加严格的安全保障义务，假设在开发者主观罪过难以判断且无法通过其他客观的有效遏制风险行为排除罪过时，智能系统开发者应当需要对结果承担相应的刑事责任。因之，开发者归责应当排除绝对严格责任的适用，不能忽视开发者的罪过形态而直接追究其技术创新行为的刑事责任。

在某些人工智能犯罪中，对于涉及智能系统滥用的行为人，其刑事责任的追究相较于普通犯罪会更为严格，实则说明为智能系统的研发者、生产者或使用者设定了更为严苛的注意义务。[33]如果智能系统开发者违背了注意义务，即使在其主观罪过难以判断，但起码有过错之时，仍需要承担相应的相对严格责任，主要体现为产品刑事责任。当生成式人工智能被划归入物的集合，且流通于市场时，即体现出生成式人工智能的产品属性。这意味着，生成式人工智能产品开发本身首先应当是不违背刑法禁止规范，属于刑法所容许风险范围内的开发行为；如果开发者故意开发了被用于违法犯罪的生成式人工智能或生成式人工智能功能本身严重违法等，开发行为属于前文的禁止行为，因违反禁止规范而被归责。在容许风险范围内，生成式人工智能开发者作为创设危险的人必须承担回避危险的义务，因而有义务保障其智能系统安全危险处于合理范围，这符合产品责任的理论根据“危险源监督义务”这一一般的保证人义务，一旦生成式人工智能产品应用造成了危害后果，开发者则应当对其危险源所引发的严重不利益

承担刑事责任，追究其生产、销售不符合安全标准产品的刑事责任。

同时，根据刑法产品责任规定，因生成式人工智能缺陷所引发的法益侵害后果而追究开发者的刑事责任时，还必须证明智能系统本身存在产品缺陷。[34]根据《产品质量法》第46条，由于生成式人工智能尚缺乏专门的国家、行业安全标准，在法律上判断生成式人工智能是否存在产品缺陷的标准，主要围绕该产品是否存在“不合理的危险”未作判断；危险的合理与否可以适用前述的风险容许判断机制，包括是否符合基本的开发原则、法益权衡判断等。假设开发者已经尽到了一切技术手段但仍然无法排除该危险，则该危险可以成立“合理的危险”而被认为符合安全标准。同时，该法第41条规定了与投入流通有关的免责事由①。前文已经分析，即便生成式人工智能进入市场流通，开发者也应承担风险控制义务，一旦发现损害风险的抬头，开发者就应当采取有效的措施包括但不限于修正算法、限制访问等遏制风险结果继续蔓延。与生成式人工智能这类新兴技术研发与利用紧密相关的是“将产品投入流通时的科学技术水平尚不能发现缺陷的存在”，结合智能产品特性，该项应当直接考虑产品风险现实化期间的科学技术水平。如果在当时的技术能力下无法采取结果回避或止损措施的，实质上属于法所容许的风险范围内，可以构成开发者的免责事由。

2. 注意义务与过失归责

不论开发者属于一般过失还是监督过失，都要求其违反注意义务。旧过失论，认为“结果预见可能性”这一主观的预见义务是过失责任的判断依据。[35]在人工智能技术日新月异的时代，对结果的预见可能性的范围难以确定，即便像旧过失论那样对预见义务施加多种限制条件，开发者乃至一般公众对技术风险和负面后果的预见可能性也是非常高的。如果纯粹以主观的结果之预见义务作为开发者过失责任的唯一认定要素，过失犯的成立范围难以限制，同时会因居高不下的技术风险而得到扩张，存在阻碍技术创新和社会进步之流弊。按照新过失论，过失犯的成立要求主观的“结果预见可能性”之外，在客观上一并考虑行为人的“结果避免义务”，即使行为人对危害结果具有预见可能性，但业已尽到避免结果发生的义务，危害结果仍不幸发生，也不能作为违法评价。[36]若未尽到相关结果避免义务，则应承担过失责任。显然，新过失论符合当今高度工业化的风险社会。开发者因疏忽大意而编制了有缺陷的程序和算法，从而引发了法益损害后果的场合；或者开发者过于自信，如开发者发现生成式人工智能程序或算法存在技术漏洞乃至瑕疵，但轻信这种缺陷在现有的人工智能技术背景下尚不会导致危害结果发生。但该后果发生的，都应以新过失论为基本立场，一并考虑开发者有无对危害结果做出相应的避免和防范措施。

此外，因应风险社会而登场的新新过失论同样会不当地扩大生成式人工智能开发者的过失处罚范围。新新过失论的代表学者藤木英雄教授认为，“预见可能性……具体的预见不一定必要，对危险的发生只要有危惧感就够了”[37]，过失犯中的预见可能性，只要求对结果具有模糊的不安感、危惧感就够了。支持新新过失论的学者可能担忧，在人工智能、基因工程等新兴技术变革创新的场合，由于没有事实感知经验，就不可能对结果有具体的预见，根据旧过失论和新过失论，对一些初次发生的重大技术事故都不具有可罚性，因而是不合理的。[38]即便如此，危惧感、不安感的概念本身极其模糊，人工智能技术的“算法黑箱”会进一步放大“危惧感、不安感”的概念不确定性。因生成式人工智能不可解释性的技术特点，开发者无法认知生成式人工智能内容生成的推演过程，也无法预见生成式人工智能会做出何种决策，技术实现过程的抽象化决定了开发者对技术结果之预见

①《产品质量法》第41条规定了三种免责事由，“未将产品投入流通的；产品投入流通时，引起损害的缺陷尚不存在的；将产品投入流通时的科学技术水平尚不能发现缺陷的存在的”。

无论如何都会具有不安的猜测。如新药的开发者，无论如何采取措施，仍然会对未知的副作用有危惧感。那么，这种不安猜测属于对结果的危惧认识吗？此外，新新过失论下预见可能性仍要关系到“结果回避义务”之考察上。对于最初的意外事故而言，如果开发者已经尽到技术安全保障和监督管理义务，因确实属于未知的危险而发生事故的，且若某危害结果实属技术局限性下难以规避之必然，不对危害结果归责非但不是对犯罪的纵容，反而是严格遵循责任主义与人权保障原则的体现。此等处遇，是为了推动社会整体之进步，而在此过程中，公众需在一定程度上容忍某些不利后果，这与“容许风险”的法理相合。

应该补充强调，过失归责中的“预见可能性”并非一种抽象的认识，如抽象的、想象性地认为以生成式人工智能为代表的人工智能系统在未来可能会危害到人类社会安全；而是一种相对具体的预见，至少预见到自己开发设计的生成式人工智能在投入使用后可被用于某些具有社会危害性的违法犯罪活动中。对于“预见可能性”的判断标准，学理上存有主观说、客观说和折中说。适用“一般人标准”的客观说更为妥当，但此时并非指社会公共群体中的“社会一般人”，而是指与开发者“所属的同一领域的一般人”，或者说是“处于同一立场的人”[39]，即人工智能技术研发领域的一般认识。

开发者在生成式人工智能研发和运营过程中，对危害结果发生已经形成相对具体的预见，但是未采取相关的检测和维护管理技术避免危害结果的发生，最终因违反注意义务而导致危害结果的发生，可以追究开发者对损害后果的过失责任。然而，现今刑法规定的过失责任还不足以完全规制生成式人工智能犯罪中开发者过失尤其是监督管理过失。具体而言，其一，我国刑法并没有直接规定缺陷产品过失责任，过去司法实务会以“重大责任事故罪”追究生产者、销售者的此类责任。[40]这种适用实际上不符合重大责任事故罪的构成要件，成立重大责任事故罪要求发生于“生产、作业中”且“违反有关安全管理的规定”，比照于产品缺陷过失，其发生在生产、销售过程中，并且违反的是与产品质量有关的法律法规和有效标准，因此认定构成重大责任事故罪不妥。如果生成式人工智能开发者违反了上述产品安全保障和风险防范义务，不排除随着技术发展会出现因产品缺陷过失引发人身损害的后果，在此情形中可以通过过失致人死亡罪或者过失致人重伤罪进行规制。但是在没有被规定为过失犯罪的损害情形中，由于不存在产品过失责任，开发者可能对其过失缺陷产品造成的损害不负罪责，形成责任漏洞。此外，我国刑法规定过失犯罪，法律有规定的才负刑事责任。虽然《生成式人工智能服务管理暂行办法》（以下简称《管理办法》）对生成式人工智能开发者的安全和合规保障义务进行了一定的规定，但该办法的位阶是部门规章，无法成为刑法适用的前置法依据。由于开发者对智能系统风险源的监督义务内容尚未得到法律层面的规定，现行刑法对于生成式人工智能犯罪规制因为无法判读开发者注意义务的规范性违反，对开发者的监督管理过失归责稍显疲软，这不利于遏制生成式人工智能技术安全风险，也不利于维护社会公众安全，应当加快生成式人工智能的法律监管。

虽然《管理办法》这一规范性文件的效力目前较低，但也能为后续人工智能的注意义务规范设立提供参考，以此从规范上更稳健地控制生成式人工智能应用安全风险。此外，欧盟的《人工智能法案》作为世界上第一部人工智能专项规制文件，其中提出的“分级分类”“全链条监管”等监管措施，对于开发者义务的建构思路也具备一定的先进性和积极意义。以全链条监管为义务设定模型，围绕风险链条现实化可以将开发者注意义务划分为事前的注意义务、事中的注意义务和长期维护义务。事前注意义务以产品的“算法安全”为核心，包括算法解释与说明义务、算法设计合规义务、算法公正控制义务等；事中的注意义务围绕生成式人工智能

监测义务展开，保障生成式人工智能在投入市场应用流通处于应用安全水平，通过技术手段实时监测异常用户和数据反映，如定期开展安全审查、不断改进算法的透明度和可解释性，还应当加强对用户行为的监督和引导，在产品应用中控制预防犯罪风险的发生；事后的注意义务则集中于生成式人工智能风险被引发且现实化后，此时开发者应当视风险等级和被损害的法益利益价值，采取从限制功能、暂停服务到终止生成式人工智能服务等处置措施，实质性地关闭风险窗口，防止风险进一步演化。

四、结语

生成式人工智能时代智能技术的刑事风险治理不能寄希望于人工智能系统的刑事主体化路径，也不能诉诸算法黑箱等技术难题而放弃刑法的规制。放弃或者降低对开发者的技术创新风险归责，只会创设新一轮风险，同时阻碍生成式人工智能技术的高质量发展。为平衡技术创新和风险治理两个目的，基于容许风险理论，应当以开发者为归责中心，合目的性且合比例地指引技术创新风险行为的归责。

参考文献：

[1]浙江首例利用ChatGPT制作假视频案，嫌犯竟是她[EB/OL].（2023-07-10）[2024-05-16]. 浙江网信网. https：//www. zjwx. gov. cn/art/2023/7/10/art_1694595_58873392. html.
[2]彭文华. 自由意志、道德代理与智能代理——兼论人工智能犯罪主体资格之生成[J]. 法学，2019(10)：18-33.
[3]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(2)：18-28.
[4]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.
[5]叶良芳. 人工智能是适格的刑事责任主体吗?[J]. 环球法律评论，2019，41(4)：67-82.
[6]比扬·法塔赫-穆加达姆，唐志威. 刑法中的创新责任：在严格责任、过失与容许风险之间[J]. 苏州大学学报(法学版)，2022，9(3)：48-61.
[7]张明楷. 论预防刑的裁量[J]. 现代法学，2015，37(1)：102-117.
[8]车浩. 刑事立法的法教义学反思——基于《刑法修正案(九)》的分析[J]. 法学，2015(10)：3-16.
[9]BLAUTH T F，GSTREIN O J，ZWITTER A. Artificial intelligence crime：An overview of malicious use and abuse of AI[J]. Ieee Access，2022(10)：77110-77122.
[10]日媒：专家发现ChatGPT可能被恶意利用于制作电脑病毒[EB/OL].（2024-05-16）[2024-05-16]. 网易网. https://www.163.com/dy/article/I2RTBRQ80514BQ68. html.
[11]乌尔里希·贝克. 风险社会：新的现代性之路[M]. 张文杰，何博闻，译. 南京：译林出版社，2018：3.
[12]皮勇. 论自动驾驶汽车生产者的刑事责任[J]. 比较法研究，2022(1)：55-70.
[13]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.
[14]赵精武. 论人工智能法的多维规制体系[J]. 法学论坛，2024，39(3)：53-66.
[15]蔡仙. 自动驾驶中过失犯归责体系的展开[J]. 比较法研究，2023(4)：65-81.
[16]喻浩东. 论风险社会的刑法归责原则——以缺陷产品责任为研究视角[J]. 比较法研究，2023(5)：185-200.
[17]张明楷. 论被允许的危险的法理[J]. 中国社会科学，2012，(11)：112-131，206.
[18]王晓菲. 欧盟发布《人工智能白皮书：通往卓越与信任的欧洲之路》[J]. 科技中国，2020(9)：98-101.
[19]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.
[20]中华人民共和国科学技术部. 发展负责任的人工智能：新一代人工智能治理原则发布[EB/OL].（2019-11-06）[2024-05-16]. https://www. most. gov. cn/kjbgz/201906/t20190617_147107. html.
[21]陈兴良. 刑法哲学（第六版）[M]. 北京：中国人民大学出版社，2017：116-117.
[22]高铭暄，等. 中国刑法词典[M]. 上海：学林出版社，1989：233-234.
[23]山口厚. 刑法总论[M]. 付立庆，译. 北京：中国人民大学出版社，2018：246-247.
[24]林山田. 刑法通论（第2版）[M]. 台北：三民书店，1986：150-151.
[25]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): 95-107.
[26]喻浩东. 被容许的风险——法理表达与实践展开[J]. 法制与社会发展, 2022, 28(6): 137-155.
[27]刘杰. ChatGPT类生成式人工智能产品提供者之刑事责任[J]. 法治研究, 2024(2): 61-71.
[28]储陈城. 限制网络平台帮助行为处罚的理论解构——以日本Winny案为视角的分析[J]. 中国刑事法杂志, 2017(6): 49-67.
[29]喻海松. 网络犯罪二十讲(第2版)[M]. 北京: 法律出版社, 2022: 165.
[30]刘宪权. 涉人工智能犯罪中的归因与归责标准探析[J]. 东方法学, 2020(3): 66-75.
[31]冯亚东. 理性主义与刑法模式[M]. 北京: 中国政法大学出版社, 1999: 104.
[32]刘仁文. 刑法中的严格责任研究[J]. 比较法研究, 2001(1): 44-59.
[33]刘宪权. 涉人工智能犯罪中的归因与归责标准探析[J]. 东方法学, 2020(3): 66-75.
[34]吕英杰. 风险社会中的产品刑事责任[J]. 法律科学(西北政法大学学报), 2011, 29(6): 145-153.
[35]陈璇. 论过失犯的注意义务违反与结果之间的规范关联[J]. 中外法学, 2012, 24(4): 683-705.
[36]钱叶六. 监督过失理论及其适用[J]. 法学论坛, 2010, 25(3): 24-31.
[37]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, 2002: 233。
[38]杨宁. 刑法介入自动驾驶技术的路径及其展开[J]. 中国应用法学, 2019(4): 107-123.
[39]前田雅英. 刑法总论讲义(第4版)[M]. 东京: 东京大学出版会, 2006: 287.
[40]中华人民共和国科学技术部. “齐二药”重大责任事故案一审宣判[EB/OL].(2008-04-30)[2024-05-18]. https://www.chinacourt.org/article/detail/2008/04/id/299946.shtml.

【责任编辑 邱佛梅】

Developer-Centric Liability of Generative AI Crimes

YE Zhusheng & LIN Manting

Abstract: Criminal law's involvement in intelligent crime governance is driven by the need for risk prevention and security control. Generative artificial intelligence (GAI) should not be considered a criminally liable subject, as intelligent systems cannot bear criminal responsibility. Regardless of AI's future advancements, developers, as the creators of risk sources, should be accountable for the harmful consequences of realized risks and are best positioned to control these risks to prevent crimes. Developer liability is central to criminal law's regulation of AI risks. Establishing a developer-centric criminal liability system effectively regulates generative AI behavior. Generative AI development encompasses the entire process from product research and development to application deployment. The human agents who enable AI systems to operate and introduce them to the market are termed "developers". Since generative AI development fosters technological innovation and societal progress, criminal liability must balance security assurance with the freedom of innovation and avoid arbitrary liability and excessive punishment. The theory of permissible risk reconciles these needs by differentiating between impermissible and permissible risk behaviors under criminal law. This approach guides developer liability based on the principle of fair risk distribution. Development that violates basic ethical standards or criminal law norms falls within the category of impermissible risk and is prohibited and sanctioned by criminal law. The development activities are within an acceptable risk range. However, there may still be residual risks associated with smart products. Developers may bear criminal liability for harmful consequences, which typically constitutes criminal liability for the product or liability for negligence in regulatory oversight.

Keywords: Generative Artificial Intelligence; developer responsibility; permissible risk; criminal liability; intelligent crimes