社会科学进展
2025 年 7 月第 7 卷第 7 期

生成式人工智能的个人信息安全风险和刑法完善路径

邵　颖　瞿丹文

上海政法学院，上海

摘　要 | 随着生成式人工智能技术的进一步提升，文本、图片、视频等不同生成物给公民的个人信息安全带来多样的风险挑战。目前在刑法层面对于生成式人工智能非法获取、泄露公民个人信息的行为只能以侵犯公民个人信息罪定罪处罚，但是刑法对该类犯罪的规制仍然存在诸多漏洞，刑事归责体系混乱、已公开个人信息超过合理使用范围、生物识别信息保护不足等问题亟需刑法一一作出回应。为了使生成式人工智能更好地应用于人类生活和维护公民信息安全权益，刑法必须在立法和司法层面及时作出调整，适时修改犯罪行为类型，调整生成式人工智能犯罪的归责原则，加强对已公开个人信息和生物识别信息的保护，构建行刑一体化的治理模式，以促进未来人工智能产业的健康发展。

关键词 | 生成式人工智能；侵犯公民个人信息罪；文生视频

1　问题的提出

近年国内外生成式人工智能呈现出井喷式发展，据2024年美国人工智能和数据分析软件公司SAS公布的一项调查结果显示，83%的中国受访者表示正在使用生成式人工智能。生成式人工智能凭借庞大的数据库为人们生活提供便利，但同时对数据库的极度依赖也给公民的个人信息安全造成了极大的威胁，ChatGPT用户信息遭到泄漏的案例屡见不鲜。为了应对当前生成式人工智能非法泄露和利用公民个人信息的问题，我国政府也出台了《网络安全法》《数据安全法》《个人信息保护法》以及《生成式人工智能服务管理暂行办法》等法律法规加以管控。但是前置法缺少具体且有针对性的解决办法，致使生成式人工智能发展和个人信息保护无法实现平衡，刑法作为最后的保障手段也应当承担起惩治侵犯公民个人信息行为的责任，协同前置法对相关犯罪行为进行规制。

2　生成式人工智能侵犯公民个人信息的主要表现方式

2.1　文本生成物侵犯公民个人信息的风险挑战

当前存在的不同功能的生成大模型都是基于GPT模型（Generative Pre-trained Transformer，生成式预训练转换器）的一种生成式人工智能聊天机器人，其采用

通讯作者：邵颖，上海政法学院刑法学研究生，研究方向：刑法学。
文章引用：邵颖，瞿丹文．生成式人工智能的个人信息安全风险和刑法完善路径［J］．社会科学进展，2025，7（7）：547-553.
https://doi.org/10.35534/pss.0707093

“Transformer”的人工神经网络和大规模预训练技术，通过自主学习大量自然语言文本的结构和规律，从而实现文本生成、代码生成、文本问答、论文写作、逻辑推理、情感分析、多语种翻译等功能。[1]目前文本生成大模型都普遍选择以Transformer模型为基础，首先对生成式AI进行预训练，将公开可用和获得授权的数据进行收集并存储为计算机数据，当用户输入问题时，AI将其转化为计算机数据，再通过算法和算力对该数据进行深度学习和分析，从而形成用户所想要的内容，根据用户的需求也可以不断改进优化，最终输出符合用户要求的新内容。由此可见，数据是生成式人工智能训练的基础，[2]而这些海量数据中含有大量个人信息，存在着极高的非法提供公民个人信息的风险。以ChatGPT为例，为了提高算法的精确性，ChatGPT会对用户输入的数据进行全盘吸收，其中包括大量的个人信息。当他人在未获得本人同意时无意输入个人信息或者不法分子利用此漏洞从事不法行为时就存在输出端泄漏个人信息的风险。虽然ChatGPT设置了部分敏感信息回答的回避机制，对于涉及个人信息的部分问题进行回避，但如果以更加隐晦和侧面的方式提问，以绕开ChatGPT设置的隐私回避条款，则其仍然会对这些问题进行分析和处理，并输出涉嫌非法提供公民个人信息的结果。[3]曾有16名匿名人士向美国加州旧金山联邦法院提起诉讼，认为微软和OpenAI未能充分地将个人身份信息从学习模型中过滤掉，使数百万人面临着个人信息立即或以其他方式向世界各地的陌生人披露的风险。该16名人士还指控微软和OpenAI通过相关产品“收集、存储、跟踪、共享和披露”数百万人的个人信息，包括账户信息、姓名、联系方式、电子邮件、支付信息、交易记录、社交媒体信息、聊天日志、搜索记录等。另外在2023年3月意大利数据保护局就以OpenAI非法收集个人信息为由宣布禁止使用ChatGPT。虽然为了打消公众的猜忌和怀疑，OpenAI在使用条款中向用户承诺会删除所有涉及用户个人信息的内容，但其并未清楚地公布识别和删除所获取的个人信息的方式，实际上当前的技术也无法支持从机器学习模型中删除部分数据，已经被用于训练模型的信息若被删除很可能会使模型失去准确性，[4]由此说明生成式人工智能即使非主动获取个人信息，仍有可能将该信息存储在数据库中，而后被运用到大语言模型的训练中从而增加泄露公民个人信息的危险。

2.2 图片生成物侵犯公民个人信息的风险挑战

从Stable Diffusion诞生以来，海内外的文生图模型不断涌现，技术不断迭代更新，使“深度伪造”技术成为颇受争议的话题。“AI换脸”就是“深度伪造”技术的一种典型应用形式。由于以往人工智能技术的局限性，生成的AI图片在细致度和精确性上有明显不足，存在较大漏洞，经过认真审查仍有可能识破，同时“换脸”技术需要制作者提供作为画面主体部分的“母体”图片或视频，这不仅会提高受害者识破的概率，也在无形中加大了操作难度，除此以外行为人还需要具备一定的计算机基础，门槛相对较高。但是文生图人工智能的出现无疑降低了“深度伪造”技术的门槛，自然聊天式的人机交互模式使不懂算法的普通人也可以轻松生成其所需的图片，从而对公民的生物识别信息安全提出挑战。目前，国内外使用他人生物信息进行伪造的案件频频发生，譬如2024年6月韩国出现大规模利用AI换脸生成色情图像的性犯罪事件，犯罪者通过拍摄周围女性照片上传至AI后生成淫秽照片在网络群组内进行传播，严重侵害女性权益。设想当文生图模型逐日发展变得更加完美之时，伪造图像会比现在成倍增多，同时也更加难以让人识别真伪，这对未来个人信息保护将是沉重的打击。

其次，图片生成类人工智能在算法训练过程中需要大量的图像和视频进行学习和理解，而这些数据大多数来源于网络上的公开信息，比如公民在社交平台发布的照片或者新闻报道等合理情况下使用的图像、视频等，这其中包含大量公民个人生物信息，虽然这些信息由公民本人公开，但这并不意味着他人可以未经许可收集或提供这些信息。而文生图技术则打破了这一权利界限，使网络上大量个人生物信息尤其是人脸信息可能在未经个人同意的情况下被用于各种数据分析、特征提取、视频制作等。[5]和文本生成类人工智能一样，若文生图模型在进行语料库训练时未获得授权或许可，可能属于侵犯公民个人信息的行为。

2.3 音视频生成物侵犯公民个人信息的风险挑战

OpenAI作为业内的领头羊，首发了文本转视频模型——Sora。文生视频类的人工智能相较于文本生成类和文生图片类人工智能在技术上有了明显的突破和进步。但Sora等文生视频模型作为文生图模型的“进阶版”，在底层逻辑上没有改变传统训练方式，前期都是通过收集大量数据信息对人工智能进行预训练，进而根据用户的指令输出内容，因此两者在训练过程或爬取数据过程中泄露个人信息的情形可能是相似的，都可能涉及对公民个人生物识别信息的侵犯。但是相较于图片生成模型，视频生成模型还存在侵犯公民声纹信息的潜在威胁。2023年北京互联网法院审理了第一起关于AI生成声音侵害声音权益的民事案件，案件中法院认为声纹和指纹、肖像等生物识别信息类似具有可识别性，如果经过人工智能技术加工处理生成的AI声音仍能够明确识别出自然人，那么就可能涉嫌侵犯公民的生物识别信息。同理在文生视频模型中常常伴随着AI生成的声音，这类声音的侵权行为一般可以划分为两种类型，一是未经自然人许可，将自然人的声音作为素材进行合成，二是未经许可将自然人的声音进行“投喂训练”，以使人工智能进行声音模仿。由此可见，文生视频模型不仅仅会侵犯公民的人脸、指纹等生物识别信息，声音作为独特的信息载体，也可能会遭到生成式人工智能的侵害。

3 刑法对生成式人工智能侵犯公民个人信息行为的规制漏洞

3.1 犯罪主体资格不清，刑事归责困难

进入人工智能时代以后，以ChatGPT为代表的人工智能表现出超越机器的智能化和类人化，ChatGPT的创始人山姆·阿尔特曼曾表示，GPT-4已经出现了我们无法解释的推理能力，没有人能解释这种能力出现的原因，甚至其研发团队也无法弄清它是如何“进化”的。[6]用户输入相关个人信息使用生成式人工智能时，AI会将输入的信息主动吸收为自身训练模型的一部分。此外为了提高生成内容的精准度，生成式人工智能的数据库必然处于不断更新的状态，在此过程中就可能自发地获取互联网上非法公开的信息，而生成式人工智能自身又不具备信息更新能力，不可能智能地将这些非法的信息筛选出来，由此就有可能构成非法获取公民个人信息的犯罪行为。另外，人工智能与其他用户互动时也有泄露和利用他人个人信息的刑事风险。生成式人工智能的自身处理过程不受人为干预，在人类所能监管的范围外，其完全有可能独立实施侵犯公民个人信息的行为。因此生成式人工智能刑事责任能力的承认或否认是达成人工智能法治化治理中供需双方和谐平衡迫切需要解决的议题。

同时生成式人工智能领域是一个横跨上游、中游至下游的复杂技术生态与产业价值链。以ChatGPT为例，其运作链条包括上游的数据供应阶段，中游的模型研发阶段以及下游的应用落地阶段，这一系列紧密相连的环节共同作用于生成式人工智能的自我学习循环中，对最终的生成内容产生深远影响。因此，从因果链条上看，生成式人工智能引发侵犯公民个人信息罪的过程，既有可能是多个阶段的因果叠加，也有可能由某个具体阶段独立造成，是一种多个主体和环节造成的“多主体的责任”。[7]然而，在此多主体协同作用的复杂场景中，每一主体往往仅对其直接参与的环节有清晰认知，缺乏与其他主体间的直接意思沟通，且难以预见或控制前后环节中其他主体行为的潜在影响。这种分离的认知状态与行为的不可控性，给刑法领域内因果关系的厘清及相应责任的判定带来了显著挑战。

3.2 已公开个人信息的刑法规制存在困境

人工智能生成内容需要收集并训练大量数据以建立数据库，这意味着生成式人工智能越要发挥出其功能就越需要大量的数据，使用者对生成内容更高质量的追求会使得数据收集成为无止境的进程，[8]因此，多数生成式人工智能采用数据爬取技术来获取大量个人信息，其中包括已公开的个人信息。在此过程中若模型开发者收集、处理行为目的不纯粹、方式不合理会导致该行为超过合理的限度，从而侵犯公民的个人信息权益。但刑法对于生成式人工智能收集、处理已公开的个人信息应当如何定性没有准确的认定。部分学者认为已公开的个人信息仍属于信息主体的附属物品，其仍具有控制权与支配权，未经许可擅自处置已收集的这类信息，尤其是涉及贩卖或非法提供给第三方之行为，违背了国家既定的信息管理规范，依法应定性为侵犯公民个人信息之犯罪行径。[9]而坚持无罪说的学者主张信息主体公开其个人信息便意味着任何主体都可以获取其公开的信息，信息获取者将公开的信息互相传播应被认为是合法行为，[10]侧重强调个人信息的公共属性，已公开个人信息由于处于公开状态失去了保护的必要性。[11]鉴于个人信息治理目标尚未得到清晰界定，且法益界定标准亦处于模糊状态，学术界在确立侵犯公民个人信息罪所保护的法益上难以形成共识，这也给将生成式人工智能处理已公开个人信息行为纳入刑事归责范畴增加了难度。

实际上，刑法与前置法的规定也是大相径庭。《民法典》第一千零三十六条和《个人信息保护法》第二十七条都指出在合理范围内对公众已公开的个人信息进行处理时，原则上可以免于寻求个人同意，除非是“自然人明确拒绝”或“侵害重大利益”。另外《信息安全技术个人信息安全规范》中也规定所涉及的个人信息是个人信息主体自行向社会公众公开的或从合法公开披露的信息中收集个人信息的，收集、使用个人信息不必征得个人信息主体的授权同意。但最高人民法院、最高人民检察院发布的《关于办理侵犯公民个人信息刑事案件适用法律若干问题的解释》（以下简称《个人信息解释》）第三条第二款规定未经被收集者同意，将合法收集的公民个人信息向他人提供的，属于《刑法》第二百五十三条中的“提供公民个人信息”。由此可见，对已公开个人信息的保护，民事和行政法律规范较为松弛，这也导致刑法中的不法与民事、行政法中不法的违反相比更广，可能会陷入不违法的行为也成立犯罪的窘境，危及法秩序的统一，[12]也违背了刑法的谦抑性。

3.3 刑法对生物识别信息保护力度不足

对于图片、视频类生成式人工智能最大的个人信息安全隐患在于个人生物识别信息存在被侵犯的风险，但是从当前《刑法》及司法解释规定的内容来看，对于侵犯生物识别信息的保护力度是远远不够的。首先，刑法对生物识别信息的含义和定位并非清楚明确。虽在《网络安全法》《个人信息保护法》以及《民法典》中提及应该将生物识别信息包含在个人信息范围内，或者单纯指出人脸信息属于生物识别信息，却没有对其概念作详细的说明，只有在2020年国家标准化管理委员会发布的《信息安全技术个人信息安全规范》中对生物识别信息进行列举，但也没有更加具体的阐述。可见我国相关法律规范对于生物识别信息的定义还没有清楚的界定。另外刑法对于生物识别信息的定位和其他法律是有差异的。如前所述，《网络安全法》《个人信息保护法》以及《民法典》都明确了公民个人信息包含生物识别信

https://doi.org/10.35534/pss.0707093

息，而《刑法》《个人信息解释》虽然规定了侵犯公民个人信息罪并对“公民个人信息”的含义作了简单的说明，但侵犯公民生物识别信息的行为是否也包含在内没有明确的司法规定。

其次，从利用“深度伪造”技术引发的诈骗、绑架、敲诈勒索等犯罪事件来看，生物识别信息更需要加强保护。《个人信息解释》第五条分别将“非法获取、出售或者提供行踪轨迹信息、通信内容、征信信息、财产信息五十条以上的”“非法获取、出售或者提供住宿信息、通信记录、健康生理信息、交易信息等其他可能影响人身、财产安全的公民个人信息五百条以上的”“非法获取、出售或者提供第三项、第四项规定以外的公民个人信息五千条以上的”认定为“情节严重”，从而达到入罪标准。在该条文中入罪标准最低的仅限于行踪轨迹信息、通信内容、征信信息和财产信息，同时也不允许司法适用中再通过类推解释予以扩大，因而无法涵盖生物识别信息。[13] 同时，第二层级的健康生理信息也不能完全等同于生物识别信息，第三层级的入罪门槛过低又无法达到对侵犯公民生物识别信息的保护要求。因此目前我国刑法只能以一般个人信息的入罪标准对侵犯生物识别信息加以制裁。随着生成式人工智能技术的逐渐成熟，若不能及时对生物识别信息提供强大而有力的保护，图片、视频生成类人工智能将成为其他犯罪的温床，对人工智能产业和人类社会发展都是极大的打击。

4　生成式人工智能侵犯公民个人信息行为的刑法完善路径

4.1　司法层面

（1）调整生成式人工智能犯罪的刑事归责原则

在生成式人工智能涉嫌侵犯公民个人信息罪的情境下，服务提供方与使用方的行为因生成式人工智能具备的人机交互特性而紧密交织在一起，这可能会增加判断侵犯公民个人信息行为的难度，也显著增加了责任界定的难度，如此将行为归责模式作为责任归属的传统判断方式显得力不从心。面对此类复杂侵权与潜在犯罪责任的判定，应倾向于采用义务归责原则作为解决方案。依据此原则，首要步骤是清晰界定服务提供方与使用方各自在个人信息保护方面的法律义务，随后依据各方义务履行情况来评估责任归属。具体而言，当发生个人信息被侵犯的情形时，责任归属的判定流程应首先聚焦于审查服务提供方与使用方是否遵循了关于个人信息保护的法定或约定规范，包括但不限于数据收集、处理及使用的合规性。进一步地，需评估是否存在违反这些保护义务的行为，并直接导致了个人信息被侵犯的结果。例如如果使用者通过伪造他人授权的方式，向生成式人工智能系统非法输入他人个人信息，误导服务提供方认为信息来源合法，进而进行后续处理，则使用者属于违反知情同意的规范获取义务而应承担相应的侵权责任，同时服务提供者未违反其应尽义务，则不应被追责。[14] 但使用者在使用人工智能产品过程中发现自身个人信息存在被侵犯的风险，要求服务提供者履行删除义务而提供者拒不改正的，应当追究其刑事责任。除此之外，从当前的技术层面看生成式人工智能模型仍属于“弱人工智能”，但是因为存在技术漏洞而“自身”实施非法获取、提供公民个人信息的行为，由于缺少犯罪主体，刑法目前无法对该类行为进行惩罚。在此情况下生成式人工智能的产品提供者负有更多的注意义务，当发现人工智能模型出现此类情况时应当及时进行修复和处理工作，若产品提供者明知产品会实施或正在实施侵犯公民个人信息的行为而不加以控制，则服务提供者可能就构成侵犯公民个人信息罪。

（2）对已公开个人信息采取分类保护模式

刑法应当始终保持自身的谦抑性，对于生成式人工智能收集已公开个人信息的行为只有在违反前置法的相关规定的情况下才可能构成侵犯公民个人信息罪，而前置法目前将“合理范围”作为处理已公开个人信息的判断标准，为了与前置法保持法秩序的统一，在刑法层面也应当把“合理范围”作为界定信息处理行为是否构成犯罪的依据，对已公开公民个人信息作出类型化保护。

首先，对合法而自愿公开的公民个人信息，除信息主体明确表示拒绝外，视为默示同意，处理行为不构成犯罪。生成式人工智能对数据有庞大的需求，研发者在收集信息时对信息来源进行合法性审查的可能性趋近于零，并且在处理成百上千的信息时要求所有收集的信息都取得信息主体的同意也是不现实的，同时“算法黑箱”问题使研发人员也无法完全了解复杂的运行过程和生成路径，因此传统的“知情同意”原则在该情境下存在明显的适用困境，从而有观点指出根据“被害人同意”与“被害人自陷风险”原则，信息主体自愿将信息公之于众意味着默许其他人可以对其信息进行处理，对该行为不宜入罪。[15] 但笔者认为，即使公民自愿将个人信息公开，但并不代表信息主体同意处理者将个人信息用于超出其合理预期的范围。例如人工智能研发团队为了丰富数据库而获取公开个人信息，该用途从社会一般大众的心理预期看属于可接受范围内，显然不能将该行为认定为犯罪行为，但是如果研发者将获取的个人公开信息非法出售，必然对公民个人的财产和人身安全造成极大的威胁，超出“合理范围”，行为人应当承担相应的刑事责任。

其次，对于合法而强制公开的公民个人信息应该首先考虑到公共利益的需要，若个人信息因法律规定而遭受强制性公开，则其随即转化为具有公共性质的范畴，并且通常情况下这些信息都不属于重要且敏感的信息。在此情境下，对于处理这些已公开信息的行为，只要其

行为本身合法且未引发不良后果或显著损害，就不应被判定为犯罪行为。如果信息主体对处理法定公开信息行为明确表示拒绝，该行为也不具备违法性。例如，当前很多企业高管信息都会显示在“企查查”等工商信息平台，生成式人工智能在大模型预训练过程中必然会对该类信息进行收集，在使用者询问相关内容时提供信息生成回答，即使高管个人表示拒绝生成式人工智能收集其信息，该行为也不构成犯罪。因为这些信息都属于依法公开的个人信息，出于保护公共利益的目的，信息处理行为高于个人信息自决权。

最后，非法公开公民个人信息的行为具备法益侵害性，收集、提供和出售该类信息可以构成侵犯公民个人信息罪。但这些信息一经公开就会流动到全世界的网络中，根据社会一般人的认知无法预见该类信息属于非法公开信息的，信息处理行为就可能不构成犯罪；若信息处理者比如生成式人工智能研发团队等具备专业知识和机能，明知此类信息为非法公开信息仍进行收集、处理行为的应当负刑事责任。

（3）修改生物识别信息相关司法解释

明确生物识别信息的定义是完善刑法的首要前提。欧盟《通用数据保护条例》第四条第（14）认为，个人生物识别信息属于自然人的身体、生理、行为特征方面，并指出个人生物识别信息应该能够识别或确认识别到自然人。[16] 我国国家标准化管理委员会发布的《信息安全技术生物特征识别信息保护基本要求》对生物特征识别信息做了解释：对自然人的物理、生物或行为特征进行技术处理得到的，能够单独或者与其他信息结合识别该自然人的个人信息。可见生物识别信息与《个人信息解释》第五条提及的健康生理信息有明显区别，但法律都没有对两者进行明确的列举或阐述。为今后打击人工智能时代侵犯个人生物信息的行为，《个人信息解释》应当以“可识别性”作为生物识别信息的“关键词”，并结合现有法律规范对生物识别信息的含义阐释，对生物识别信息的概念作出明确的定义。此外，生物识别信息是能够反映自然人独有的生理特征和行为特征的个人信息，与公民的人格尊严、人身安全都有着紧密的关联。[17] 欧盟将生物识别信息定位为个人敏感信息，划定为高风险领域，说明与其他个人信息相比侵犯生物识别信息对个人信息安全有更大的危害，因此我国刑法应当明确地将个人生物识别信息划分到“敏感个人信息”中，以加强对生物识别信息的保护。

与此同时，《个人信息解释》应当降低敏感个人信息的入罪标准以实现对生物识别信息的有力保护。从该解释的内容看敏感个人信息可划分为两类，一是“行踪轨迹信息、通信内容、征信信息、财产信息”，二是“住宿信息、通信记录、健康生理信息、交易信息等其他可能影响人身、财产安全的公民个人信息”。对这两档不同类型的敏感个人信息，认定情节严重的标准也不同。但笔者认为《个人信息解释》应当与《个人信息保护法》相衔接，采取“二分法”，即明确区分一般个人信息与敏感个人信息，无需把生物识别信息划分到具体的层级中，因为随着生成式人工智能的快速发展，生物识别信息在不同的应用场景下可能具有不同的属性，如果将其简单地归类于“财产信息”或“健康生理信息”或其他一般个人信息都会使其定位模糊，无法突出对公民个人人格尊严权益的保护。[18] 对此可以考虑在现行《个人信息解释》的基础上进行修改，把侵犯公民敏感个人信息的入罪标准调整至50条，侵犯一般个人信息的入罪标准保持在5000条，其中考虑到生物识别信息在今后的日常生活中会被广泛使用和侵犯，应当认定其属于最值得保护的敏感个人信息，入罪标准为50条。

4.2 立法层面

（1）将非法使用个人信息行为纳入刑法监管范围

根据罪刑法定原则，侵犯公民个人信息罪只能规制非法获取、出售或提供行为，对于合法获取后非法使用的行为无法进行约束，只能通过其他罪名如诈骗罪、盗窃罪来规制。犯罪者可以通过获取生成式人工智能生成结果中所包含的合法收集的个人信息，在没有获得信息主体的同意下非法使用该信息进行人脸支付盗窃财产等其他犯罪活动，这属于滥用生物识别信息的行为。虽然可以通过盗窃罪等对该行为进行惩治，但只是保护了盗窃罪等其他犯罪的相应法益，而忽视了非法使用个人信息行为本身的社会危害性和公民的信息安全权益。因此现行刑法可以将非法使用行为一同纳入侵犯公民个人信息罪的行为类型中，但需遵循两项核心准则：第一，确立滥用个人信息行为的刑事规制界限，即情节标准的界定。过度保护可能阻碍信息的合理流通与利用，因此应将“情节严重”作为判定滥用行为是否构成犯罪的关键阈值，即刑法仅当滥用行为导致严重后果或具备显著恶劣情节时方予介入。第二，要求行为人需认识到其滥用行为所蕴含的社会危害性、实际损害或潜在的损害风险，诸如非法存储与加工等行为。据此逻辑，即便行为人最初以合法手段获取公民个人信息，一旦其行为违反法律规定且可推断出具有不正当目的，即应视为滥用行为，并依据侵犯公民个人信息罪的相关规定予以法律制裁。这一界定标准旨在确保法律适用的精准性与公正性。

（2）构建行刑一体化治理模式

卢梭曾说：“刑法是具有二次性的法律，也是对其他所有各种法律的认可。”[19] 对于生成式人工智能侵犯公民个人信息行为的治理不仅需要刑法作为制裁手段的最后一道保障，更需要民法、行政法等前置法共同规制。但是当前刑法与前置法不仅在实体规范上出现对立，在程序衔接上也存在难题。一方面，个人信息保护的行政法规与《刑法》所保护的法益并不完全一致，行政法规不仅保护公民的个人信息自决权，也保护社会中

https://doi.org/10.35534/pss.0707093

合理利用和处理信息的活动，而《刑法》更加侧重于保护公民个人信息权利。另一方面，有效的行政监管是推动行刑案件衔接的重要基础，但当前行政监管部门呈现分散模式，没有统一的监管部门管理导致监管乏力。考虑到科技社会的长期发展，涉生成式人工智能的违法犯罪行为应当采用综合治理的模式，及时调整相关政策和法律规范，强化刑法与行政法、民法之间的衔接机制，构建一个系统化的法律保护网络，并在此基础上厘清行刑衔接的边界。首要任务在于明晰行政违法与刑事违法界定的界限，防止二者混淆。在法律适用上，应确立行政不法判定优先于刑事不法的原则，并将行政不法的认定过程标准化，以便为刑事违法性判断中的构成要件符合性提供有力支撑。探讨刑事违法性时，应更注重评估该行为是否具备显著的社会危害性。若生成式人工智能侵犯个人信息的行为仅触及前置行政法规范，而未对刑法所保护的法益造成实质性损害或构成紧迫威胁，则不应轻易认定其具备刑事违法性，而应仅限于行政违法范畴。此外，对于生成式人工智能服务提供者未履行适当注意义务时，应当优先适用行政法对其加以管制，而非直接对其适用侵犯公民个人信息罪，刑法的介入应发生在行政手段已充分施展且仍不足以有效规制时，方可提升至刑事违法层面进行处罚。简而言之，刑事违法性的判定应以前置行政处罚的充分性与有效性不足为前置条件，确保刑法介入的正当性与必要性。[20] 其次，设立专门的监管机构是保护公民个人信息的必要措施，纵观全世界，英国、日本、韩国等多国都设立了个人信息保护的专门机构，同时面对生成式人工智能的发展，政府行政监管是规避生成式人工智能技术风险的第二道防线，[21] 由此可见设立个人信息保护的专责机构迫在眉睫，该机构与刑事司法机关协同办案也更有利于行刑主体间顺畅高效的合作。

5 结语

生成式人工智能凭借出色的技术成长速度打破众人想象，成为所有行业万众瞩目的焦点。但是在此背后，许多无解的法律问题正在缓缓蔓延。正是生成式人工智能出色的“学习”能力，使其在为人类提供服务的同时也给人类带来了无尽的信息安全问题。不管是技术已经趋于成熟的文本生成模型还是新兴的视频生成模型都存在着侵犯公民个人信息的安全隐患。生成式人工智能的用户就如同行走在黑暗森林之中，看不清黑箱法则的本质，只能任由生成式人工智能侵吞和利用自己的个人信息。[3] 如果继续放任个人信息侵权问题野蛮生长，那么生成式人工智能也会失去其存在的意义。当前刑法对其侵犯公民个人信息的犯罪治理并不完善，始终存在着多方面的漏洞。因此，我们亟需对现有法律体系进行适应性优化以匹配这一伟大的技术变革。为有效应对生成式人工智能技术所催生的新型侵犯公民个人信息犯罪，我们应以面向未来的眼光调整法律法规内容，构建预防与打击并重的机制，从而切实捍卫公众利益，维护社会的公平正义与司法体系的公信力，携手促进人工智能技术的稳健发展，确保其应用既高效又可持续，为人类社会带来福祉。

参考文献

[1] 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（2）：18–28.
[2] 万美秀. 生成式人工智能对个人信息保护的挑战与治理路径 [J]. 网络安全与数据治理，2024（4）：53–60.
[3] 严嘉欢，王昊. 论生成式人工智能中个人信息保护的困境纾解 [C]. 《新兴权利》集刊2023年第2卷——生成式人工智能法律问题研究文集. 2024：107–114.
[4] 熊进光，贾珺. 元宇宙背景下ChatGPT蕴含的法律风险及规制路径 [C]. 《新兴权利》集刊2023年第2卷——生成式人工智能法律问题研究文集. 2024：8–20.
[5] 邓建鹏，赵治松. 文生视频类人工智能的风险与三维规制：以Sora为视角 [J]. 新疆师范大学学报（哲学社会科学版），2024（6）：92–100.
[6] 刘宪权. ChatGPT等生成式人工智能的刑事责任问题研究 [J]. 现代法学，2023（4）：110–125.
[7] 盛浩. 生成式人工智能的犯罪风险及刑法规制 [J]. 西南政法大学学报，2023（4）：122–136.
[8] 王斯敖. 生成式人工智能背景下个人信息保护研究 [J]. 网络安全技术与应用，2024（7）：121–124.
[9] 李昱. 已公开个人信息的刑法保护原理及其规则 [J]. 人民检察，2022（8）：63–65.
[10] 周光权. 侵犯公民个人信息罪的行为对象 [J]. 清华法学，2021（3）：25–40.
[11] 姜涛，郭欣怡. 已公开个人信息刑法规制的边界 [J]. 学术界，2023（3）：95–111.
[12] 杨楠. 已公开个人信息刑法规制的分层建构 [J]. 政治与法律，2024（5）：128–141.
[13] 喻海松. 最高人民法院、最高人民检察院侵犯公民个人信息罪司法解释理解与适用 [M]. 中国法治出版社，2018：297.
[14] 李川. 生成式人工智能场域下个人信息规范保护的模式与路径 [J]. 江西社会科学，2024（8）：68–80，206.
[15] 最高人民检察院检察应用理论研究课题组，秦余荣，哈书菊. 侵犯公民个人信息的刑法边界研究——以已公开个人信息的保护为视角 [J]. 大庆社会科学，2024（1）：100–105.

［16］张晨原．元宇宙发展对个人信息保护的挑战及应对——兼论个人生物识别信息的概念重构［J］．法学论坛，2023（2）：132–141.
［17］刘宪权．敏感个人信息的刑法特殊保护研究［J］．法学评论，2022（3）：1–10.
［18］刘宪权，陆一敏．生物识别信息刑法保护的构建与完善［J］．苏州大学学报（哲学社会科学版），2022（1）：60–71.
［19］卢梭．社会契约论［M］．何兆武，译．北京：商务印书馆，1980.
［20］田宏杰．行政犯的法律属性及其责任——兼及定罪机制的重构［J］．法学家，2013（3）：51–62，176–177.
［21］童云峰．走出科林格里奇困境：生成式人工智能技术的动态规制［J］．上海交通大学学报（哲学社会科学版），2024（8）：53–67.

Personal Information Security Risks of Generative Artificial Intelligence and the Path to Criminal Law Improvement

Shao Ying　Qu Danwen

Shanghai University of Political Science and Law, Shanghai

Abstract: With the further advancement of generative artificial intelligence technology, various generated products such as text, images, and videos pose diverse risks and challenges to citizens' personal information security. Currently, at the criminal law level, illegal acquisition and disclosure of citizens' personal information by generative artificial intelligence can only be punished as the crime of infringing upon citizens' personal information. However, there are still many loopholes in the criminal law's regulation of such crimes, including chaotic criminal liability systems, excessive use of publicly available personal information beyond reasonable scope, and insufficient protection of biometric information. These issues urgently require criminal law to respond one by one. In order to better apply generative artificial intelligence to human life and safeguard citizens' information security rights, criminal law must promptly make adjustments at both the legislative and judicial levels, timely modify the types of criminal acts, adjust the liability principles for generative artificial intelligence crimes, strengthen the protection of publicly available personal information and biometric information, and build an integrated governance model of administrative and criminal law to promote the unlimited development of the future artificial intelligence industry.

Key words: Generative artificial intelligence; Crime of infringing upon citizens' personal information; Video generative artificial intelligence

https://doi.org/10.35534/pss.0707093