人工智能法治专题

文章编号:1008-4355(2025)05-0003-10

生成式人工智能用户交互中的个人信息保护

龙卫球,邰江丽

(北京航空航天大学 法学院,北京 100191)

摘 要:生成式人工智能已深度嵌入广泛的社会经济场景,其与用户高频交互的应用特性,导致个人信息保护问题呈现特殊性与复杂性。一方面,生成式人工智能在与用户交互过程中,带来了个人信息保护的特殊风险。另一方面,生成式人工智能的用户交互技术机制本身,还会增加个人信息保护法律规则适用的难度。然而,生成式人工智能大模型在用户交互场景下,具有对数据尤其是其中个人信息和隐私高度依赖的技术特点,使得上述风险与挑战的发生本身具有"技术必然性",并随着交互场景不断丰富而呈现"现实多样化"特征。为此,当前在践行人工智能积极发展观的前提下,应当正视生成式人工智能用户交互的技术逻辑,有必要引入"技术—规则—用户"协同保护框架,合理应对上述风险与挑战,推动人工智能从"技术突破"到"社会经济价值"的转化。具体对策包括:提升安全技术与治理能力,完善个人信息处理合法性基础规范,增强用户在与大模型交互场景下的保护意识与能力等。

关键词:生成式人工智能;用户交互;个人信息保护;技术逻辑;协同保护框架

中图分类号:DF529 **文献标志码**:A

DOI:10.3969/j.issn.1008-4355.2025.05.01 **开放科学(资源服务)标识码(OSID)**:

一、引言

2022 年 11 月 30 日,美国人工智能实验室 OpenAI 推出了一个名为 ChatGPT 的用户界面,这标志着生成式人工智能(Generative Artificial Intelligence)正式面世。此后,以中美为代表的人工智能

收稿日期:2025-08-01

基金项目:中央网信办 2025 年网络法治专委会重点研究课题"数智化进程中健全网络综合治理体系研究"

作者简介:龙卫球(1968—),男,江西吉水人,北京航空航天大学法学院教授,博士生导师;邰江丽(1981—),女,贵州遵义人,北京航空航天大学法学院 2018 级博士研究生。

公司或实验室不断迭代升级大模型,推动人工智能模型技术从“渐进性创新”到“指数级跃迁”的历史性转变。

我国学者已关注生成式人工智能个人信息保护问题,并展开了若干富有启发的研究,但针对用户交互场景涉及的特殊性,仍然缺乏系统且相对深入的专题探讨。①

《中华人民共和国民法典》(以下简称《民法典》)与《中华人民共和国个人信息保护法》(以下简称《个人信息保护法》)正式确立了我国个人信息保护制度。从法律规范维度看,《民法典》在第四编“人格权”第六章“隐私权和个人信息保护”中,将个人信息权益和隐私权界定为两类相互关联的法益。其中,《民法典》第1032条将“隐私”定义为“自然人的私人生活安宁和不愿为他人知晓的私密空间、私密活动、私密信息”;《民法典》第1034条第2款则规定,受法律保护的自然人的个人信息,是指以电子或者其他方式记录的、能够单独或者与其他信息结合识别特定自然人的各种信息,包括非私密信息,也涵盖私密信息;后者的特殊性将导致相关保护方式的复杂交叠,因此,私密信息优先适用有关隐私权的规定;在隐私权规定未涵盖的情形,再适用个人信息保护的规定。在生成式人工智能应用的用户交互场景中,既有非私密个人信息,也包含自然人不愿为他人知晓的“私密信息”(隐私信息)。需要注意的是,二者的保护规则存在较大差异。例如,《民法典》第1033条关于隐私权侵害行为的禁止规则中,明确采用经“权利人明确同意”方可免责的表述;而第1035条关于个人信息收集处理的规则中,仅要求获得“同意”即可。显然,法律对私密信息的保护标准更为严格。因此,明确区分私密信息与个人信息,更有利于强化对私密信息的特殊保护。② 但是为叙述方便,下文不严格区分私密信息与个人信息,而是将二者统一纳入个人信息保护的范畴展开研究。

二、生成式人工智能用户交互背景下个人信息风险的类型及“技术必然性”特征

近年来,在生成式人工智能大模型交互应用中,已发生多起因产品功能设计缺陷及产品能力不足引发的个人信息保护风险事件。一些来自谷歌、苹果等大型科技公司,以及加利福尼亚大学伯克利分校(UC Berkeley)等知名院校的专家学者发现,大语言模型GPT-2在遭遇恶意前缀注入时,会返回疑似训练数据中敏感信息的内容,如某机构或某人的名称、邮箱、手机号、传真号等。③ ChatGPT的数据泄露事件是典型案例,由于ChatGPT的语料库中包含敏感信息与机密信息,其在执行生成任务时可能在无意中输出这些内容,若未经过适当处理与保护,将引发数据泄露与隐私泄露风险。④

生成式人工智能用户交互场景中,个人信息与隐私数据特殊风险形态多样。具体而言,可将其归纳为以下类型。

---

① 参见钭晓东:《风险与控制:论生成式人工智能应用的个人信息保护》,载《政法论丛》2023年第4期,第59页;张新宝:《生成式人工智能训练语料的个人信息保护研究》,载《中国法学》2024年第6期,第86页;叶雄彪:《生成式人工智能背景下的个人信息保护:范式转换与规则完善》,载《法学家》2025年第4期,第61页。

② 参见王利明:《民法典人格权编的亮点与创新》,载《中国法学》2021年第4期,第5页。

③ 参见陈寅嵩:《大模型正在“记住”和“说出”——六起真实案例剖析,揭露敏感信息泄露的严重后果》,载《安全+(技术版)》(内刊)2024年第4期,第3页。

④ 参见陈寅嵩:《检测与预防:大模型信息泄露的安全“紧箍”》,载《安全+(技术版)》(内刊)2024年第4期,第22页。

类型一:基于安全漏洞导致的个人信息泄露。生成式人工智能作为复杂的创新型信息生态技术系统,以生成式人工智能技术为核心,在支撑和推动整个信息环境中知识传递共享、认知流动扩散的同时,也伴生相应的安全问题。① 因此,生成式人工智能也难免存在新型安全漏洞,且部分漏洞直接关乎个人信息与隐私安全。

类型二:基于训练数据被唤醒导致的个人信息泄露。欧盟《人工智能白皮书》在构建未来人工智能的规制框架时,已专门针对训练数据提出规制建议,明确要求在人工智能产品与服务使用过程中,确保个人信息及隐私得到充分保护。理论上,输入大语言模型的数据越多,对旧信息的记忆越容易被埋藏在模型深处,但大模型的记忆类似人类的记忆,存在被主动唤醒的可能性。

类型三:基于强大数据挖掘与关联能力导致的个人信息泄露。例如,ChatGPT o3 具备“看图定位”推理能力(可根据一张普通街景照片,精准定位照片拍摄位置),尽管 OpenAI 在 ChatGPT o3/o4-mini 中已经针对新模型的个人信息泄露风险采取限制措施,即模型会拒绝基于图像的人物识别请求及无事实依据的推理请求,但模型仍可能借助地理位置线索与公开的个人信息,锁定个人身份与行踪。

类型四:基于用户无意识行为导致的个人信息泄露。当前,情感陪伴类 AI 已经成为用户与大模型交互的重要工具。情感陪伴类 AI 的核心是“建立信任与情感连接”,因此用户为获取共情或个性化回应,会主动透露童年经历、情感创伤、家庭矛盾、性取向、心理状态等“深度隐私”信息。此外,情感陪伴类人工智能为维持“人设一致性”,还需长期记忆用户个人偏好,甚至通过算法推测用户未明说的信息。这种“主动记忆+隐性推理”机制,导致模型存储的个人信息与隐私数据更庞杂,且用户难以察觉相关个人信息的收集与处理。一旦模型发生数据泄露等情况,这些“深度隐私”可能被批量曝光,引发严重的社交风险与情感创伤。可见,生成式人工智能在用户交互场景下具有导致个人信息与隐私风险增加的“技术必然性”特征。近年来,随着大模型应用中用户交互场景日渐多样化与复杂化,不仅将这种“技术必然性”扩展到“现实多样性”的广度,还因不断叠加的模型演进带来的新风险样态。

生成式人工智能大模型本质上是基于信息的再生产装置,具有数据依赖性,其机理在于依托持续优化的深度学习与算法模型,并形成对数据信息的聚合、挖掘、分析、预测、服务及进一步生产的能力。这种数据依赖并非单纯的技术选择,而是模型性能优化的内在要求。生成式人工智能的核心能力建立在海量数据的训练与优化基础之上,其预测与生成质量高度依赖持续的交互数据输入,这一技术特性决定了数据收集的必然性,也促使系统在优化压力下形成扩张性数据需求,通过大规模训练数据模仿人类能力,产出与用户高度相关的内容。② 生成式人工智能可以利用网络上易获取的大规模无标记数据开展预训练,并通过简单适配与高效微调应用于大量下游任务。例如,大型语言模型在交互过程中,不仅依赖用户输入的显式指令,还可能通过会话日志、行为模式分析等方式

① 参见白云、李白杨等:《从知识困境到认知陷阱:生成式技术驱动型信息生态系统安全问题研究》,载《信息资源管理学报》2024年第 1 期,第 13-21 页。

② 参见孙国烨、吴丹等:《用户与生成式人工智能交互的隐私披露多因素影响模型研究》,载《信息资源管理学报》2025 年第 2 期,第 108-122 页。

隐式收集数据,以提升模型的适应性与精准度。

生成式人工智能为发挥相应能力,在必要时会采集与个人信息与隐私相关数据。例如,生成式人工智能模型的训练数据可能涉及人脸信息、行踪轨迹、消费信息、医疗信息等大量个人信息。[①] 在用户交互模式下,这种可能性会进一步增大。一方面,模型对实时交互数据的依赖可能导致数据收集范围模糊化,使原本合理的用户数据利用行为逐渐滑向过度采集;另一方面,数据存储与处理的长期积累可能形成难以追溯的数据足迹,增加泄露或滥用风险。即便遵循数据最小化原则或采用差分隐私(Differential Privacy)[②]技术,仍难以完全规避因模型优化需求而产生的数据扩张倾向。

除上述因模型优化需求引发的数据扩张风险外,生成式人工智能还具备强大的碎片化信息挖掘与整合能力,这种能力也催生了特殊技术层面的个人信息风险。该能力可能将信息主体散落在网络空间的个人信息汇聚融合,并对其开展深度预测分析,大幅提升相关使用者挖掘出碎片化信息背后的个人信息(包括私密信息)的可能性。

生成式人工智能基于上述技术逻辑,在创造出前所未有的超强能动性和智慧价值的同时,也与个人信息保护的相关法律原则产生冲突。这一冲突具体体现在:一方面,在生成式人工智能的用户交互过程中,其为个体与社会带来的赋能效应达到了前所未有的高度;另一方面,该技术正逐渐显现出"技术必然性",即明显提升个人信息与隐私面临风险的可能性。能否妥善化解生成式人工智能技术逻辑中的这一矛盾,已成为生成式人工智能大模型实现合理部署、开展合规应用的重要前提之一。

三、生成式人工智能用户交互个人信息保护规则面临的挑战

数字化时代以来,世界各个国家和地区均重视个人信息保护,先后围绕个人信息保护建立起了相关法律制度,其中欧盟2018年实施的《一般数据保护条例》(即EU 2016/679, General Data Protection Regulation, GDPR)、2020年美国加利福尼亚州颁布的《加州消费者隐私法》(California Consumer Privacy Act, CCPA)及我国《个人信息保护法》等最具有代表性与影响力。这些法律赋予用户对其个人信息和隐私更多控制权,其中最重要的就是知情同意权。

与此同时,面对人工智能迅猛发展的趋势,各个国家和地区也已着手通过立法应对其新发展带来的风险与挑战。例如,2024年欧盟出台《欧盟人工智能法》(EU AI Act),我国先后出台了《生成式人工智能服务管理暂行办法》《人工智能生成合成内容标识办法》等相关规定。应该说,这些规定为数字化时代维护用户个人信息权益确立了基本法律规则,也为应对人工智能发展引发的基本权利安全风险、满足数据驱动发展与提升数据可用性需求提供了法律保障。但面对当前生成式人工智能发展带来的个人信息保护问题,上述法律法规、部门规章仍存在明显不足。总体来看,生成式人工智能技术逻辑与个人信息保护原则在以下三方面存在冲突。

① 参见张涛:《生成式人工智能训练数据集的法律风险与包容审慎规制》,载《比较法研究》2024年第4期,第86-103页。

② 差分隐私最初是密码学中的一种手段。具体来说,差分隐私通过在数据发布和查询结果中引入一定量的噪声,以减少或避免对个体数据的泄露,从而保护个人隐私。参见鞠雷、刘巍然主编:《密码学隐私增强技术导论》,科学出版社2024年版,第5页。

一是技术黑箱特性与透明原则相背离。以 GPT 为代表的大语言模型依赖海量无标注数据预训练,通过自注意力机制构建的千亿级参数网络,并总结出高度复杂的自然语言统计规律,其内部决策过程不可追溯、不可解释。开发者难以精准掌控数据流向及处理逻辑,导致《个人信息保护法》要求的告知义务在事实上难以充分履行。用户交互数据(如浏览记录、搜索历史)被实时吸纳为迭代训练素材,但模型无法提供数据处理路径的可视化说明,形成"输入—输出"的双向不透明。二是动态生成机制与限定原则冲突。为优化交互体验,系统需持续收集位置信息、设备信息、对话语境等情境化数据,其范围常超出初始声明用途。例如,智能客服在解决售后问题时,可能主动调取用户购买记录、社交动态等关联信息从而生成回复,这种跨场景数据融合虽然能提升服务精准度,却违背目的限定原则。三是算法涌现能力与用户控制权虚化。例如,GPT-4 高达 1.8 万亿的模型参数量突破临界点时,会产生超越预设的推理能力。这种不可预测性具体表现为:用户简单询问医疗建议,可能触发系统整合公开病历、健康论坛碎片信息生成个性化诊断,无意间泄露用户未主动提供的敏感健康状况;或在法律咨询场景中,通过长文本分析推断当事人婚姻状况、经济能力等个人隐私。

进一步而言,在生成式人工智能用户交互场景中,对现有个人信息保护法律具有以下特殊适用挑战:

第一,算法透明度规则难以确定。生成式人工智能的系统模型极为复杂,背后往往涉及多方主体所提供的训练数据,且运行逻辑也无法完整展现。从证据认定理论看,算法可解释性审查是破解技术黑箱的关键。但在实践中,算法可解释性的审查范围与深度仍存在争议。例如,如何平衡商业秘密保护与司法审查需求、如何判断算法解释是否达到"合理清晰"标准,尚未形成统一的裁判尺度。①

第二,告知同意规则难以落地。生成式人工智能的技术特性决定了开发者事实上无法履行向信息主体真实、准确、完整地告知处理目的、方式等内容的义务,因此无法满足知情同意规则的法定要求。这是因为,在数据获取后的训练阶段,即便是开发者自身,也无法详细知晓大语言模型掌握何种自然语言统计规律。② 此外,对于大模型训练所使用的大量公开个人信息数据,开发者还面临难以追溯个人信息主体的障碍,更遑论在涉及隐私或敏感个人信息时获取该主体的单独同意。另有观点认为,知情同意原则虽旨在从个人信息处理源头控制潜在风险,但此举会降低个人信息利用效率,阻碍数据要素市场化进程;究其根本,是知情同意规则存在内生适用矛盾,其规范功能、规则逻辑与利益实现机制的不适配,导致其在司法实践中失灵。③

第三,合目的性准则难以遵循。生成式人工智能在语料库构建与模型学习训练过程中,均需处理使用大量个人信息数据,此类数据的处理和使用,是在生成式算法的基础上通过深度学习的神经网络技术实现的。但深度学习的神经网络技术会使生成式人工智能在运行时具有自主性、交互性

① 参见王毓莹、刘旭冉:《涉人工智能司法纠纷应对研究》,载《人民司法》2025 年第 11 期,第 15 页。

② 参见黄锫:《生成式 AI 对个人信息保护的挑战与风险规制》,载《现代法学》2024 年第 4 期,第 101 页。

③ 参见寿晓明、曹贤信:《生成式人工智能场景中个人信息保护的风险、逻辑与规范》,载《昆明理工大学学报(社会科学版)》2024 年第 1 期,第 21 页。

特征,其运行的具体目的及相关算法行为处于实时随机变动状态;这种随机变动的运行特质,导致生成式人工智能的开发者难以在运行之初,准确设定处理个人信息目的并明确严格地合目的性运行。即便在生成式人工智能的初始算法参数中,已明确个人信息的处理目的及目的相关性的行为判断规则,但在运行过程中,目的相关性的判断仍取决于人工智能模型的自主判断;而生成式人工智能模型的自主判断结果,受输入数据类型与范围、语料库建构逻辑、模型学习训练状况等诸多不确定因素的复杂作用的影响,难以被事先预测和明确,由此决定的个人信息处理行为,也难以保证符合目的相关性要求。① 若将每一次数据收集、每一步数据处理都按照目的限制原则要求予以披露,并获取个人同意,将给人工智能发展带来巨大成本压力乃至不确定性。有观点主张,人工智能时代应削弱甚至放弃目的限制原则,因为该原则及传统个人信息保护理念起源于20世纪70年代,已经无法充分回应人工智能技术进步需求。②

第四,必要性准则难以恪守。人工智能的发展与产业应用,经历了从解决单一任务到生成通用模型的发展历程,其中关键在于扩展定律(Scaling Law)③的成功验证,从GPT 1到GPT 3不断尝试,在将参数规模提升100倍、训练数据量提升50倍后,GPT3.5出现了能力"涌现"(Emergence)。正因人工智能大模型需依托海量的训练数据才能实现智能"涌现",导致其训练过程难以恪守个人信息收集的必要性原则。

第五,个人信息的部分主体难以行使权利。从技术可行性来看,个人信息删除权、撤回权等存在行使障碍。例如,在人工智能大模型中全面且彻底删除个人信息极为困难。传统的删除方式不仅难以追踪和锁定待删除数据,且删除后可能破坏模型学习成果、影响整体性能,还无法排除模型在删除后自行"脑补(还原)"这些数据的可能。人工智能大模型在技术上"删不掉"用户个人信息,而法律要求必须"删得掉",如何在平衡个人信息保护与模型可用性的前提下执行删除操作,已成为模型服务提供者面临的不小难题。④ 此外,个人信息一旦被纳入数据集并用于模型训练,便会与其他数据融合为模型的一部分;由于数据深度融合且学习过程具有累积性、连续性,此时响应和保障个人信息撤回权,可能会影响模型的整体性能与稳定性。若要求当数据主体行使撤回权时,生成式人工智能系统必须立刻识别并删除有关数据,同时确保算法性能不受影响,这在实践中往往既复杂又成本高昂。⑤

四、生成式人工智能用户交互应用下个人信息保护规则的完善

在当前全球积极推动人工智能开发与应用的背景下,我们既要深刻理解生成式人工智能发展

① 参见李川:《生成式人工智能场域下个人信息规范保护的模式与路径》,载《江西社会科学》2024年第8期,第68页。

② 参见武腾:《人工智能时代个人数据保护的困境与出路》,载《现代法学》2024年第4期,第116页。

③ 扩展定律是指在人工智能领域,模型性能(如损失函数值)随着模型参数、训练数据量、计算力等资源的增加而呈现出一定规律变化的经验公式。参见卡里·布里斯基(Kari Briski):《如何通过扩展定律推动更智能、更强大的AI》,载英伟达(Nvidia)网,https://blogs.nvidia.cn/blog/ai-scaling-laws/,2025年2月7日访问。

④ 参见林北征:《没有删除,只能遗忘:AI大模型个人信息删除义务的解释与重构》,载《西安交通大学学报(社会科学版)》2024年第6期,第152页。

⑤ 参见杨清望、唐乾:《生成式人工智能与个人信息保护法律规范的冲突及其协调》,载《河南社会科学》2024年第12期,第81页。

的技术逻辑，也需理性认知并应对用户交互过程中产生的个人信息风险。从当前政策导向来看，若因个人信息保护问题趋于复杂便因噎废食，甚至放弃生成式人工智能的交互应用，显然并非明智之举；但反过来，我们也绝不能因为片面发展人工智能而忽视个人信息安全。对此，本文期待构建“技术—规则—用户”协同的保护框架，通过技术能力强化、规则体系完善与用户主体参与的有机联动，构建平衡生成式人工智能技术创新发展与个人信息和隐私保护的新路径。

（一）在技术层面，应当以切实提升安全能力水平为目标，全面强化技术治理手段

自互联网时代以来，法律便开始注重技术治理手段的强化；但生成式人工智能发展起来之后，我们仍低估了通过加强技术布局本身来提升安全治理效能的特殊作用。这一点当前必须特别重视，应遵循“技术安全同步”的原则做好布局，强化大模型敏感数据安全与隐私保护能力，构建纵深技术治理体系。不过，大模型及其用户交互如何更复杂地影响个人信息保护，以及如何在技术上破解这一问题，仍需依赖技术的进一步发展。

目前国内外就此已经开展了不少卓有成效的研究。例如，有的研究基于情境脉络完整性理论[①]，采用用户标注与半结构化访谈的研究方法，揭示用户隐私披露受到用户的隐私态度、技术信任、隐私风险感知能力的共同影响，且系统的数据管理透明度会通过影响技术信任而间接作用于隐私披露。据此，建议通过构建用户交互的隐私披露多因素影响模型，开发更具隐私友好性的生成式人工智能系统。[②] 另有研究聚焦基于全生命周期理论的防范措施与管理方法，强调针对用户信息与情境信息过度获取、留存问题，需从监管与开发两端同步发力进行治理。[③] 这些研究无疑具有启发性。

本文在认同并吸收上述有益观点的基础上，结合当前大模型用户交互模式对于个人信息和隐私数据日益凸显的技术偏好特点，认为有必要着重强化可解释性算法研发与差分隐私等技术的应用，从源头上约束生成式人工智能大模型对个人信息（尤其是敏感信息）的记忆与推理能力。具体而言，在数据预处理阶段，采用差分隐私技术对训练数据注入可控噪声，结合数据脱敏技术消除直接标识符；在模型训练阶段，通过“联邦学习框架”[④]实现分布式数据“可用不可见”，结合同态加密与安全多方计算保障参数交换安全；在推理阶段，嵌入动态噪声注入机制，阻断输出结果与原始数据的关联性；在部署阶段，实施模型拆分架构，将敏感数据处理模块隔离在可信执行环境中运行，采用权重混淆和梯度裁剪技术防止模型反演攻击。同时，集成细粒度访问控制和安全水印追踪技术，实现数据流向全链路审计；通过分层加密、隐私算法融合与硬件级隔离，形成协同防护机制。[⑤]

---

① 情境脉络完整性（Contextual Integrity）理论由纽约大学的海伦·尼森鲍姆（Helen Nissenbaum）教授提出，该理论认为人们生活在多种社会情境中，每个情景都有其相匹配的社会规范。参见岳彩申、孙磊主编：《消费金融个人信息保护法律问题研究》，中国民主法制出版社 2022 年版，第 20 页。

② 参见孙国烨、吴丹等：《用户与生成式人工智能交互的隐私披露多因素影响模型研究》，载《信息资源管理学报》2025 年第 2 期，第 108-122 页。

③ 参见陈思琳：《人工智能生成内容技术信息安全风险分析》，载《保密科学技术》2024 年第 9 期，第 47-51 页。

④ 联邦学习框架（Federated Learning Framework）是支持多方在不共享原始数据的情况下，通过协同系列模型实现数据“可用不可见”的技术体系，集成了节点管理、模型训练、加密参数交换与聚合等功能。常见的联邦学习框架有谷歌的 TensorFlow Federated（TFF）等。参见林伟伟、石方等：《联邦学习开源框架综述》，载《计算机研究与发展》2023 年第 7 期，第 1551-1580 页。

⑤ 参见汤博文、周昌令：《大模型安全防护技术探讨》，载《保密科学技术》2025 年第 5 期，第 5-10 页。

(二)在规则层面,引入沙盒机制,突出柔性约束,从而更加有效地保护个人信息

首先,考虑引入沙盒机制。沙盒机制是监管部门通过强化技术设计替代合规义务的授权实践机制。比如,新加坡《个人数据保护法》(Personal Data Protection Act,PDPA)存在禁止直接共享原始客户数据的法律限制,依据该法明确有效的同意以充分"告知"为前提,这对于许多面临"衍生数据风险"的生成式人工智能大模型而言,无疑是巨大的挑战。此举标志着监管理念从"事后处罚"向"事先引导"转变,为人工智能等前沿技术应用提供了合规确定性。上述监管中,核心在于相关企业针对AI训练通过部署隐私增强技术(Privacy Enhancing Technology,PET),从技术架构上阻止高敏感度个人画像的创建或消除,而非为高风险的数据处理活动寻求法律许可。之所以此举可以替代用户知情同意要求,关键在于通过具体技术方案,在实践中实现了"设计即合规"。

其次,完善现行法律,包括增设"正当利益"条款作为个人信息收集、处理的合法性基础,同时要求生成式人工智能大模型应用服务提供者应通过产品功能设计,保障用户的退出权和删除权。基于生成式人工智能大模型的技术原理,尽管大模型预训练阶段涉及对含有个人信息数据在内的学习数据的收集、存储、加工等,属于个人信息保护法上的"处理",但大模型预训练是将文本、图像等非结构化数据转换为机器可学习格式以处理相关数据,在此过程中,个人识别性可能降低。基于此,域外的立法和司法实践多认可以"正当利益"作为处理个人信息的合法性基础。例如,德国科隆高等地区法院在审理北莱茵—威斯特法伦州消费者保护集体诉Meta一案中,裁判认为,Meta使用用户个人信息甚至敏感隐私数据训练人工智能,具备《一般数据保护条例》中"合法利益"的合法性基础;法院详细审查了Meta提出的人工智能训练目标,认定其利益明确且真实,最终认可了人工智能训练对大规模数据的技术需求,认为该需求无更温和且等效的替代方式。实际上,为与正当利益的合法性基础配合,Meta还采取了更多合规行动,包括更新对用户的透明度通知;向用户提供更长的通知期,并提供可用的控制信息,支持用户将所有已发布的帖子从公开更改为私密,以避免接受模型训练等。

最后,对生成式人工智能用户交互个人信息的侵权责任适用过错推定原则,合理设定服务提供者的义务与责任。唯有如此,才能既契合现阶段鼓励人工智能产业发展的政策导向,又能满足协调人工智能发展与风险治理的实际需要。

一方面,对生成式人工智能大模型因用户交互导致的个人信息侵权,宜一体适用《个人信息保护法》第96条的规定,即适用过错推定责任。我国《生成式人工智能服务管理暂行办法》第9条明确规定,提供者应当依法承担网络信息内容生产者责任,履行网络信息安全义务。本文进一步认为,从侵权归责的视角看,对该条款的理解需更细致,有必要结合生成式人工智能系统的新特点进行综合权衡。此外,部分观点将生成式人工智能界定为"产品"或"服务",这也值得商榷。国家互联网信息办公室2023年4月11日发布的《生成式人工智能服务管理办法(征求意见稿)》曾经使用了"产品或服务"的表述,而生效后的《生成式人工智能服务管理暂行办法》仅保留"服务"表述,删除"产品"字样,正是因为其认为将生成式人工智能作为"产品"并不恰当。此外,产品责任是工业社会的产物,工业产品具有大规模生产、批量销售的特征,而生成式人工智能服务缺乏高度同质性,具有鲜明的个性化特征;且产品责任通常适用于人身或有形财产损害,生成式人工智能致人损害一般不

涉及此类情形。[①] 由此可见,就大模型交互服务提供者的侵权归责原则而言,仍然应适用过错推定原则。

另一方面,在适用过错推定责任的基础上,不应简单地以“网络服务提供者”来界定人工智能大模型交互服务提供者的法律责任,而应将其认定为“特殊新型服务提供者”,并合理设定相关义务与责任。这是因为,就生成式人工智能大模型应用交互场景下的服务提供者的法律性质而言,不能简单套用传统网络侵权中“网络用户—网络服务提供者”的二元主体划分。在传统二元主体划分下,用户是内容的生产与上传者,服务提供者仅为内容的传播提供了相关技术服务,“生成”与“传播”的界限较为明晰;但是生成式人工智能大模型应用涉及的侵权场景具有因果关系多元化特征,人工智能生成的侵权内容是算法、数据集、用户输入等众多因素共同作用的结果,不应过分强调服务提供者的单方义务与责任。此外,还需特别考虑此类用户交互的“非直接公开”特征,即生成内容处于人机对话的“一对一”场景,相关信息需要借助用户或者其他主体的进一步传播,才可能影响社会公众或产生实质性损害后果。

(三)在用户层面,应强化用户参与保护的有机联动,提升其自我保护意识与能力

一方面,应当通过公众教育提升用户对生成式人工智能大模型交互行为的隐私保护敏感性。如前文所述,大模型交互的隐私泄露具有一定隐蔽性,因此应加强对用户的隐私边界认知的教育,明确“显式交互—隐式数据”的风险关联,提升用户风险防护意识。尤其在情感陪伴类 AI 的场景中,因交互的深度、信任度与数据依赖度高,用户更容易发生无意识隐私泄露,其隐私风险在“泄露可能性”与“泄露危害”两方面均显著增高,故除法律制度与技术工具保护外,用户自身也需提升“场景化隐私保护意识”。另一方面,服务提供者也应加强对用户的提示,避免用户在使用过程中分享隐私信息。当识别到更高隐私泄露风险时,大模型服务提供者应该主动通过产品设计等方式提示用户,落实“通过设计保护隐私”原则。

五、结语

生成式人工智能大模型开发与应用正迅速发展,生成式人工智能用户交互日益频繁。在生成式人工智能用户交互场景中,对数据的收集与处理,无论源于技术能力、技术缺陷,还是用户自身行为与能力等因素,都可能引发不可忽视的个人信息风险。可见,系统地探讨其有效保护机制,对完善相关法律体系、平衡技术创新与权益保障之间的关系具有重要意义。

本文研究表明,生成式人工智能用户交互模式引发的个人信息保护问题,较此前的互联网与数字技术环境带来的问题更为严峻、更加复杂。它不仅产生了基于安全漏洞、训练数据唤醒、超强数据挖掘能力及用户无意识行为的新型风险样态,也给现行《个人信息保护法》的实施带来诸多前所未有的挑战。从根源上看,这些挑战源于生成式人工智能大模型在交互模式下“对数据极端依赖”的技术逻辑特点,该特点使得上述风险与挑战具有“技术必然性”;而用户交互场景的持续扩张,则

① 参见周学峰:《生成式人工智能侵权责任探析》,载《比较法研究》2023 年第 4 期,第 117 页。

进一步将这种“技术必然性”推向“现实多样化”的广度,并不断叠加模型演进带来的新型风险。因此,应当引入“技术—规则—用户”协同的保护框架,通过技术能力强化、规则体系完善与用户主体参与的有机联动,构建平衡生成式人工智能技术创新发展与个人信息保护的新路径。JS

On Personal Information Protection in User Interaction with Generative Artificial Intelligence

LONG Weiqiu, *TAI Jiangli*

(The Law School, Beihang University, Beijing 100191, China)

Abstract: Generative artificial intelligence (AI) has become deeply embedded in diverse socio-economic contexts. Its characteristic of frequent interaction with users renders personal information protection issues particularly distinctive and complex. On the one hand, generative AI's dependence on user interaction increases specific risks to personal information protection. On the other hand, the technical mechanisms underpinning user interaction in generative AI systems also complicate the application of existing legal rules. Moreover, given that large-scale generative AI models rely heavily on data—especially personal information and privacy—for training and optimization, the emergence of these risks and challenges is "technically inevitable," and their "manifestations have become increasingly diverse" as interactive scenarios expand. Therefore, under the premise of promoting the responsible development of artificial intelligence, it is essential to acknowledge the inherent technological logic of generative AI user interaction and to establish a coordinated protection framework that integrates "technology, rules, and users." Such a framework can effectively address the above risks and challenges, facilitating the transformation of artificial intelligence from mere "technological breakthroughs" to tangible "socio-economic value". Specific measures include enhancing security technology and governance capacity, refining the legal norms governing the lawful basis for personal data processing, and strengthening users' awareness and capacity for protection in interactive AI environments.

Key words: generative artificial intelligence; user interaction; personal information protection; technological logic; coordinated protection framework

本文责任编辑:林士平

青年学术编辑:杨尚东