Open Journal of Legal Science 法学, 2025, 13(5), 922-928
Published Online May 2025 in Hans. https://www.hanspub.org/journal/ojls
https://doi.org/10.12677/ojls.2025.135131

Hans汉斯

生成式人工智能与个人信息保护问题研究

王春卓

长春理工大学法学院，吉林 长春

收稿日期：2025年4月2日；录用日期：2025年4月11日；发布日期：2025年5月16日

摘　要

随着生成式人工智能技术的快速发展，其在提升社会生产效率的同时，也给个人信息保护带来了前所未有的挑战。本文分析了生成式人工智能在个人信息保护领域引发的诸多问题，包括数据收集的隐蔽性与超范围性、算法黑箱导致的风险、数据存储和共享的安全隐患以及数据使用的不可控性等。针对这些问题，本文提出三方面的协调路径：建立动态透明的信息告知机制，通过可视化界面和实时通知系统保障用户知情权；实施严格的数据采集源头审查制度，确保训练数据的合法性和真实性；完善算法解释说明机制，提升生成式人工智能系统的可解释性。通过对相关法律规制现状的分析，在促进技术发展的同时切实保护个人信息权益，最终实现人工智能创新与隐私保护的双赢格局。

关键词

生成式人工智能，个人信息保护，法律规制

Research on Generative Artificial Intelligence and Personal Information Protection Issues

Chunzhuo Wang

School of Law, Changchun University of Science and Technology, Changchun Jilin

Received: Apr. 2nd, 2025; accepted: Apr. 11th, 2025; published: May 16th, 2025

Abstract

With the rapid development of generative artificial intelligence technology, while enhancing social productivity, it has also brought unprecedented challenges to personal information protection. This paper analyzes the multiple issues raised by generative AI in the field of personal information protection, including the covert and excessive nature of data collection, risks caused by algorithmic

文章引用: 王春卓. 生成式人工智能与个人信息保护问题研究[J]. 法学, 2025, 13(5): 922-928.
DOI: 10.12677/ojls.2025.135131

black boxes, security vulnerabilities in data storage and sharing, as well as the uncontrolled ability of data usage. To address these challenges, this paper proposes a three-pronged coordinated approach: establishing a dynamic and transparent information disclosure mechanism to safeguard users' right to know through visual interfaces and real-time notification systems; implementing a rigorous source review system for data collection to ensure the legality and authenticity of training data; and improving algorithmic explanation mechanisms to enhance the interpret ability of generative AI systems. Through an analysis of the current legal regulatory landscape, this paper aims to effectively protect personal information rights while promoting technological advancement, ultimately achieving a win-win scenario where AI innovation and privacy protection coexist harmoniously.

Keywords

Generative Artificial Customer Service, Protection of Personal Information, Legal Regulation

1. 引言

近年来，生成式人工智能技术迎来革命性突破，以 ChatGPT 为代表的大模型工具在自然语言处理、多模态内容生成等领域展现出卓越的创造能力，并快速融入社会生产生活的各个维度。这类技术通过语义理解与内容生成能力，可依据用户指令自主创作文本、图像、音频等多样化数字内容，不仅显著提升了信息生产效率，更催生了全新的应用场景与商业模式，为全球产业升级注入了强劲动能。2025 年 1 月 20 日，杭州深度求索公司发布最新款人工智能产品 DeepSeek-R1 并开源，选择更低成本的纯强化学习算法，成功破圈引发全球关注[1]。随着生成式人工智能技术在各领域的深入应用，其带来的个人信息保护风险日益呈现出隐蔽性强、影响深远的新态势。该技术依托大规模数据训练，涉及用户身份信息、行为特征、社交关系等敏感数据的深度处理，在数据收集、分析、存储和应用的全流程中均存在显著的隐私安全挑战。相较于传统数据处理技术，生成式 AI 展现出独特的风险特征：一方面，其算法的不透明性导致数据处理过程难以追踪；另一方面，内容生成能力可能产生包含用户隐私特征的衍生数据，引发风险的指数级扩散。面对这一形势，如何在促进技术创新的同时有效保护个人信息，已成为数字社会治理亟待解决的关键议题。

我国 2023 年 7 月 10 日网信办等七部门快速出台并实施了《生成式人工智能服务管理暂行办法》(下文简称《办法》)，旨在促进生成式人工智能健康发展和规范应用，维护国家安全和社会公共利益，保护公民、法人和其他组织的合法权益[2]。现行《办法》虽然明确禁止生成式人工智能侵犯个人信息权益，并对服务提供者与使用者的个人信息保护义务作出原则性规定，但其相关条款存在明显局限性。一方面，《办法》中的个人信息保护条款主要沿袭《个人信息保护法》等既有规定，内容较为笼统简略；另一方面，这些条款未能充分回应生成式 AI 技术带来的新型个人信息保护挑战，如算法黑箱导致的隐私风险、生成内容中的个人信息泄露等问题。这种立法现状使得《办法》在应对生成式 AI 特有的个人信息保护问题时缺乏针对性和可操作性，实际保护效果受到显著制约。有鉴于此，必须采取系统性应对策略：首先需要精准识别其区别于传统侵权的特殊形态特征，进而以保护优先为核心理念重塑个人信息保障体系。

2. 生成式人工智能对个人信息保护带来的问题

2.1. 生成式人工智能造成个人信息知情同意的判断困难

生成式人工智能在个人信息知情同意方面面临显著的判断困难，主要体现在数据收集、处理及输出的全流程中。首先在数据采集阶段，系统通常需要从互联网抓取海量训练数据，这一过程往往缺乏透明度和用户同意机制。生成式人工智能在训练过程中，并不会对所使用的个人信息来源合法性进行专门甄别。它无法判定这些个人信息在采集时，是否获得了权利归属人的知情同意。之所以存在这一情况，是因为验证个人信息来源是否合规，并非生成式人工智能开展自主训练的目标[3]。生成式人工智能在数据采集过程对数据来源合法性不加区分的做法实质上架空了个人信息保护的知情同意原则。由于算法运行过程的非透明性，此类违规采集行为往往具有高度隐蔽性，使得信息主体难以察觉自身权益正受到侵害。

其次，在数据加工环节，即便采用脱敏技术处理原始信息，生成式 AI 仍可通过深度关联与模式识别还原个人隐私，导致传统数据匿名化措施失去保护效力。更为复杂的是，AI 生成内容可能包含两种风险形态：一是隐含训练数据中的真实敏感信息，二是生成与真实情况高度近似的虚构内容，这种特性使得个人信息使用的授权边界变得难以界定。在实时交互场景下，由于系统无法动态披露数据处理目的，加之算法决策过程的不透明性，形成了“双重认知障碍”用户既难以知晓个人数据的具体流向，也无法理解其被加工利用的实际方式，最终导致知情同意制度在技术层面面临实质性架空。

生成式人工智能在服务交互环节存在特殊的隐私泄露风险：系统可能未经用户明确授权即输出已存储的个人信息。虽然现行法律严格规制个人信息的第三方共享行为，要求必须获得信息主体的知情同意，但生成式 AI 的自主决策特性带来了独特的合规挑战。由于系统响应具有高度情境依赖性，服务提供商实际上无法预先确定 AI 会在何种对话场景中触发哪些个人信息的输出。这种技术特性使得传统的“事前授权”机制在操作层面陷入两难。既不能穷尽所有可能的输出情形，又难以在实时交互中即时获取有效授权，最终导致个人信息共享的合规要求在实践中面临执行失效的风险。这种不可预测性导致基于事先知情同意的信息共享授权机制在实际运行中面临执行困境，难以得到充分保障和有效落实。

2.2. 生成式人工智能与信息撤回权的冲突

《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》)第十五条规定：“基于个人同意处理个人信息的，个人有权撤回其同意。个人信息处理者应当提供便捷的撤回同意的方式。个人撤回同意，不影响撤回前基于个人同意已进行的个人信息处理活动的效力。”从比较法视角来看，该项权利已在多个法域得到明确确认。传统数据保护机制明确赋予用户撤回授权或要求删除个人信息的权利，但生成式人工智能的技术特性对这一权利的实现构成了实质性障碍。究其原因，当个人信息被纳入训练数据集后，会通过深度学习过程被编码到模型的参数矩阵中，与海量其他数据形成复杂的关联[4]。这种数据模型的深度耦合具有不可逆性：在生成式人工智能的应用场景中，个人信息删除权的行使面临多重技术障碍：其一，用户数据在训练过程中已与其他信息深度交织，形成复杂的关联网络，难以实现精准分离；其二，模型通过海量训练形成的知识体系具有整体性特征，局部数据的删除可能影响模型的整体性能表现。更为关键的是，要真正实现符合法律要求的删除效果，系统需要同时满足两大技术要求：既要建立数据溯源体系，又要具备动态调整模型参数的能力。然而，受限于当前的技术发展水平，这些要求在算力支撑和算法实现层面均存在显著困难，致使个人信息撤回权在生成式 AI 领域难以得到实质性保障。即使删除原始训练数据，模型仍可能通过生成内容间接还原个人信息特征。数据主体在行使撤回权时，生成式人工智能系统需具备即时响应能力，这种技术复杂性必然带来高昂的合规成本。虽然我国《个人信息保护法》从立法层面确立了个人信息撤回权制度，但在实施层面仍存在显著缺陷：一是未能针对生成

式 AI 的技术特性制定差异化实施细则，导致技术实现与权利保障之间缺乏有效衔接；二是对模型迭代过程中个人数据的重复使用问题规制不足，使得撤回权的法律边界在技术应用中变得模糊不清。这种制度供给不足的状况，暴露出传统个人信息保护框架在应对生成式 AI 技术时的结构性缺陷。亟需通过技术创新和制度完善来平衡人工智能发展与个人信息自主权保护。

2.3. 生成式人工智能处理数据与使用目的限制的冲突

目的限制原则是国内外个人数据保护法普遍接受的基本原则。欧盟 2016 年《通用数据保护条例》将目的限制与使用限制统一作出规定。目的限制原则包括两部分内容：一是收集目的必须特定、明确和正当；二是使用不得与上述目的不相兼容，从而确立了目的限制原则的所谓“两肢”[5]“目的限定原则”是指已公开个人信息的处理应当限定在信息主体公开个人信息的初始用途之上，依据这种初始用途界定的范围才属于处理已公开个人信息的合理范围[6]。目的限定原则作为个人信息保护的核心准则，要求个人信息的采集与运用必须严格遵循三特定标准——特定目的、明确范围和合法依据，禁止任何与初始用途相悖的二次开发利用。该原则通过双重约束机制实现权益保障：一方面要求信息处理目的必须具体、正当且必要，将个人权益影响控制在最小限度；另一方面在保障用户知情权与控制权的同时，为数据处理者保留了合理的操作空间。既防范了数据滥用风险，又维护了个人信息处理的规范秩序。

根据《个人信息保护法》相关规定，个人信息处理必须遵循若干相互关联的基本原则。其中，透明原则与目的限制原则具有基础性地位。该法第 7 条明确规定，数据处理主体应当依据透明性原则，向信息主体完整披露包括处理目的、处理方式等核心要素在内的全部处理事项。其中，处理目的的明确性构成知情权实现的实质性要件——唯有处理目的本身具备足够的确定性与可理解性，方能真正达成透明性原则的规范意旨。然而，当处理目的模糊不清时，判断信息应达到何种准确度、完整度以及是否需要及时更新等标准就失去了客观依据。只有明确揭示具体处理目的，才能据此评估所收集数据是否满足质量要求。由此可见，个人信息保护法律体系中的各项原则虽然各有侧重，但彼此之间存在紧密的逻辑关联。目的限制原则一旦被弱化或架空，不仅会直接损害透明原则的实施效果，还将导致质量原则等多项保护机制难以正常运作，最终危及整个个人信息保护制度的有效性。

然而生成式人工智能的技术特性恰恰突破了这一限制。其技术范式具有双重突破性：一方面，基于预训练的技术路线要求模型必须通过海量多源数据进行无差别学习，这种通用化训练模式天然消解了数据采集时的具体目的边界；另一方面，基础大模型可同时支撑文本生成、智能问答、数据分析等多元化应用场景，这使得数据主体最初同意的使用目的在技术实现过程中被实质性虚置。这种技术特性导致数据处理很可能突破原始授权范围，延伸至用户既未预期也未许可的新领域，这种数据使用可能超越最初收集目的，涉及用户未曾预见或同意的新用途。生成式人工智能的技术运行机制与目的限定原则存在深层次冲突。其核心技术路径依赖于对多维数据的广泛采集和持续学习，这种数据驱动的发展模式与严格限定使用范围的要求形成结构性矛盾。尤为关键的是，生成式 AI 在模型迭代过程中产生的分布式数据编码和参数化存储特征，不仅造成个人信息的非必要扩散，更从技术底层消解了目的限定原则的实施条件。这一根本性矛盾暴露出当前数据治理体系在应对具有自主进化能力的 AI 系统时存在制度僵化问题，迫切需要在坚守保护底线的前提下，构建更具适应性的动态合规框架。

3. 生成式人工智能对个人信息权益的实现与保障

3.1. 强化信息使用透明度的告知机制

在《个人信息保护法》的规范框架下，保障信息主体的同意撤回权是制度实施的核心环节，而优化告知机制则是实现这一权利的重要保障。基于比例原则的要求，需要在促进数字经济发展与保护个人信

息权益之间寻求动态平衡。通过建立清晰、完整的告知制度，向信息主体充分披露数据处理的目的、方式和范围，不仅能够提升信息处理的透明度，更有助于保障信息主体在充分知情的基础上行使自主决策权，从而实现个人信息保护与合理利用的有机统一[7]。透明度增强工具的研发与应用，其核心价值在于弥合技术认知鸿沟。需要强调的是，当前面临的关键挑战源于技术实现层面的局限性，而非监管制度的缺失。这类工具体系主要涵盖以下功能模块：可视化访问控制面板、即时动态通知机制、全链路数据追踪报、隐私风险评估矩阵。为增强信息使用的透明度，可构建多层次的智能化通知路径：在生成式人工智能服务场景下，为满足《个人信息保护法》的合规要求，服务提供者应当构建多维度的透明化交互系统：首先，采用动态交互式界面整合可视化图表、渐进式披露和场景化示例，直观展示个人信息收集范围、处理目的(包括生成内容用于模型训练等算法优化用途)及潜在影响；其次，依托实时推送系统在数据流转关键节点触发主动提醒，同步提供简明语言说明与即时答疑功能，并在数据处理目的或方式变更时通过显著方式触发重新授权流程；此外，建立用户可随时访问的“数据可视化控制面板”，该面板应包含：第一，以时间轴形式完整记录数据使用历史的数据足迹仪表盘，支持按类型、场景等维度检索；第二，实时数据流向查询功能，清晰展示个人数据在算法训练、产品改进等具体场景中的应用；第三，完整的同意管理模块，确保用户可随时撤回或变更授权；最后，集成隐私影响评估矩阵，为数据处理者与数据主体提供系统化的风险评估框架，在数据处理前有效识别和调整潜在风险，确保用户始终基于充分知情的前提行使自主选择权。这种从“一次性告知”转向“全周期可溯”的透明化路径，这种高度透明和可控的通知机制，不仅提升了数据处理的合法性，也增强了用户信任，使其能在完全了解后果的基础上有效行使撤回权，保护其隐私和个人权利。既满足了合规要求，又能通过技术赋能使用户掌握个人信息数据控制权。

3.2. 建立初始阶段的审查制度

在预训练阶段，生成式人工智能系统呈现出典型的封闭性特征。该阶段的深度学习训练与语料库构建过程具有以下特点：其一，系统运行完全在开发主体控制范围内进行，不涉及外部交互；其二，训练数据的采集与使用均由开发主体单方面决定和操作；其三，整个训练流程处于开发主体的全程监控之下。由于这种高度可控的封闭特性，预训练阶段对个人信息权益的潜在风险处于最低水平，且一旦发生数据问题也最容易追溯责任主体[8]。鉴于当前阶段个人信息数据来源呈现显著的多元化和复杂化特征，为切实确保生成式人工智能运营主体能够有效履行个人信息处理的合法性审查及持续合规维护义务，建议采取以下措施：首先，输入端的初始个人信息存在偏差而导致错误结论，应该归咎于平台的合规审查缺漏。平台应在收集个人信息时辨识真伪，但却因疏忽导致虚假个人信息流入并影响结论准确性[9]。建议依据《个人信息保护法》第 54 条与第 58 条的规定，构建系统化的数据处理主体合规责任机制。具体应包括以下要求：明确要求数据处理者建立常态化的个人信息合规审计制度；督促其构建覆盖数据处理全生命周期的合规管理体系，以实现对个人信息的源头性保护。在操作层面，建议针对算法处理的个人信息实施“四维合规评估”机制——“真实性、准确性、客观性、多样性”。

为确保生成式人工智能数据处理全流程的合法合规性，建议建立“源头管控－动态治理－全程审计”的协同监管体系。具体而言：在数据采集环节，需严格遵循《个人信息保护法》关于知情同意的要求，确保获得数据主体的有效授权；对于公开数据的使用，应当遵守相关规定。同时，建议行业主管部门加快推进数据要素市场制度建设，通过制定数据确权、定价、交易等配套规则，促进数据资源的合法有序流动与价值实现。

3.3. 个人解释说明请求权的行使

当前，生成式人工智能技术仍面临着显著的“黑箱”困境，其内部决策逻辑和运作机理缺乏足够的

透明度。这种可解释性缺失导致信息主体难以建立对该技术的信任基础，尤其担忧系统可能不当处理或泄露其敏感个人信息。个人解释说明请求权，是指个人在使用 AI 系统时，有权要求运营者以清晰、可理解的方式说明自动化决策的逻辑、依据及可能影响。这一权利是个人信息保护与算法透明原则的重要体现，尤其在涉及个人权益的自动化决策(如信用评估、就业筛选等)场景下，用户可依法要求运营者解释算法如何得出特定结果，包括数据来源、处理规则及权重分配等核心要素。行使该权利时，用户通常需通过书面或平台指定渠道提出明确请求，运营者则应在合理期限内提供非技术性说明，但可能受商业秘密或技术可行性限制。根据《个人信息保护法》第 48 条确立的解释说明请求权制度，当信息主体依法行使该权利时，数据处理者有义务就其个人信息的具体使用情况进行全面、清晰的说明。这一法律机制不仅有助于增强个人信息处理流程的透明度，更能使信息主体明晰生成式人工智能在模型训练过程中对其个人数据的处理方式，从而有效降低对人工智能系统安全性的疑虑。

生成式人工智能依托深度神经网络技术构建，其模型结构高度复杂。当前主流基础大模型的参数量已突破千亿规模，这些海量参数共同作用于模型的输出结果，使得从输入到输出的决策过程呈现出典型的“黑箱”特性。这种内在机制的不透明性导致研究者难以准确追溯和解释模型生成特定回答的内在逻辑。从技术实现层面来看，要求人工智能系统达成完全的算法透明性存在根本性挑战。若强制推行透明度要求，不仅难以实现预期目标，还可能对神经网络技术的创新发展与应用落地形成不当限制。这种技术特性使得服务提供者在解释个人信息如何被模型学习并影响输出时面临客观困难，最终导致信息主体的解释说明权在实践中难以得到充分保障[10]。在人工智能研发过程中，开发者通常无需公开披露技术细节，包括训练数据的具体构成，这既是对其商业秘密的合理保护，也是保障科研自由的基本要求。然而，当涉及个人数据处理时，服务提供者应当采取必要措施，帮助用户理解生成式 AI 如何学习及使用个人信息。具体而言，服务方应当履行以下义务：一是向用户充分解释模型的基本工作机制、数据来源、推理过程以及潜在的局限性；二是以通俗易懂的方式说明算法运行原理，明确告知用户生成内容的可信度范围，并在涉及个人权益时提供专门解释。关键在于，相关解释应当适应用户的理解能力，确保信息传达的有效性。随着生成式人工智能可解释性技术的持续进步，服务提供商应当同步完善解释机制，为用户提供更加全面、深入的系统说明，从而切实保障用户的知情权。

4. 结语

生成式人工智能的快速发展正深刻重塑社会创新格局，既开辟了广阔的发展前景，也使个人信息保护面临新的考验。当前，协调技术创新与隐私保护的关系，已成为需要社会各界共同破解的时代命题。随着技术进步和治理体系的不断完善，我们既要筑牢个人权益的防护屏障，更要通过多元协同构建开放、透明的治理体系。只有在促进创新与完善规范之间形成良性互动，才能推动人工智能与个人权益保护的协调发展，实现技术进步与社会福祉的双赢。

参考文献

[1] 顾男飞. 从 ChatGPT 到 DeepSeek: 人工智能蒸馏数据的风险治理[J/OL]. 图书馆论坛, 1-10. http://kns.cnki.net/kcms/detail/44.1306.G2.20250318.0907.002.html, 2025-04-02.

[2] 王利明. 生成式人工智能侵权的法律应对[J]. 中国应用法学, 2023(5): 27-38.

[3] 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[J]. 河南社会科学, 2024, 32(12): 81-93.

[4] 张旭芳. 生成式人工智能的算法安全风险及治理路径[J]. 江西社会科学, 2024, 44(8): 90-100.

[5] 武腾. 人工智能时代个人数据保护的困境与出路[J]. 现代法学, 2024, 46(4): 116-130.

[6] 黄锫. 生成式 AI 对个人信息保护的挑战与风险规制[J]. 现代法学, 2024, 46(4): 101-115.

[7] 尹玉涵, 李剑. 生成式人工智能的个人信息保护问题及其规制[J/OL]. 海南大学学报(人文社会科学版), 2025: 1-11. https://doi.org/10.15886/j.cnki.hnus.202310.0241, 2025-03-27.

[8] 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, 44(8): 68-80+206.

[9] 陈禹衡. 生成式人工智能中个人信息保护的全流程合规体系构建[J]. 华东政法大学学报, 2024, 27(2): 37-51.

[10] 张新宝. 生成式人工智能训练语料的个人信息保护研究[J]. 中国法学, 2024(6): 86-107.