人工智能安全与个人信息保护：挑战与对策研究

吴少刚
南宁职业技术大学，广西 南宁 530008

DOI:10.61369/SE.2025070030

摘　　要： 人工智能技术的广泛应用为个人信息保护带来诸多新挑战。本文依据截至2024年底已公开的研究与数据，系统梳理人工智能环境下个人信息面临的主要风险，并从技术、管理及法律三个层面提出综合防护策略。研究指出，生成式人工智能在数据采集隐蔽性、模型训练不可控性等方面加剧了隐私泄露的可能性。文章进一步提出整合隐私增强技术、分级治理机制与动态合规体系的解决方案，以助力构建安全、可信的人工智能应用环境[1,3,2,5]。

关 键 词： 人工智能安全；个人信息保护；隐私增强技术；数据治理

Artificial Intelligence Security and Personal Information Protection: Research on Challenges and Countermeasures

Wu Shaogang
Nanning Vocational And Technical University, Nanning, Guangxi 530008

Abstract: The wide application of artificial intelligence technology has brought many new challenges to the protection of personal information. Based on the research and data that have been made public by the end of 2024, this article systematically sorts out the main risks faced by personal information in the artificial intelligence environment and proposes comprehensive protection strategies from three aspects: technology, management, and law. Research indicates that generative artificial intelligence increases the possibility of privacy leakage in terms of the concealment of data collection and the uncontrollability of model training. The article further proposes a solution that integrates privacy-enhancing technologies, hierarchical governance mechanisms, and dynamic compliance systems to assist in building a secure and trustworthy artificial intelligence application environment[1,3,2,5].

Keywords: srtificial intelligence security; personal information protection; privacy-enhancing technology; data governance

引言

人工智能技术日益融入经济与社会多个方面，显著提高了生产效率与服务水准。据IDC 2024年度报告，全球人工智能市场规模已达到3000亿美元，年增长率超过25%[3]。然而，人工智能系统高度依赖数据，处理大量个人信息的同时也带来新的安全隐患。调查显示，超过83%的人工智能系统存在隐私缺陷，生成式人工智能的广泛应用进一步放大了此类风险[2]。

全球范围内相关法律法规持续完善。欧盟《通用数据保护条例》（GDPR）为个人信息保护提供了范本。我国也陆续颁布《网络安全法》《数据安全法》及《个人信息保护法》，初步构建起个人信息保护的法律基础。2022年发布的"数据二十条"也对建立数据基础制度提出明确指引。

本研究结合2024年底之前的实际案例，剖析人工智能环境下个人信息保护面临的技术与法律难题，提出多层次、系统化的防护对策，以期在推动技术创新的同时切实保护个人信息。

一、人工智能与个人信息保护的理论框架

（一）技术特性与个人信息的关联

人工智能系统依赖包含大量个人信息的训练数据。麦肯锡2024年报告表明，训练大规模语言模型需超过45TB文本数据，其中近30%涉及个人信息[4]。人工智能与个人信息的交互主要体现在数据收集、模型训练及应用推理三个环节，每一环节均有其独特的风险特征。

人工智能系统内部处理流程复杂且不透明。深度学习算法的"黑箱"特性不仅提高了隐私泄露的可能，也为监管带来困难。哈佛

作者简介：吴少刚（1981-2），男，玉林人，大学学士，南宁职业技术大学人工智能学院高级工程师，研究方向：为信息安全。

大学2023年研究指出，算法的不可解释性显著增加了合规难度。

（二）法律法规体系

全球范围内个人信息保护法律制度日趋健全。欧盟 GDPR 确立了数据保护的七大原则。我国也构建了以《个人信息保护法》为核心的基础法律体系。据清华大学2023年研究，这些法律实施效果显著，但执行层面仍存在一定障碍。

针对人工智能的技术特点，监管部门也出台了专门规定。《生成式人工智能服务管理暂行办法》明确要求服务提供者确保训练数据来源合法。欧盟《人工智能法案》则依据风险等级实施分类监管，对高风险系统严格执行合规要求。

二、人工智能环境下个人信息保护的主要挑战

（一）数据收集与处理中的隐私风险

人工智能系统普遍面临用户同意机制失效的问题。传统的告知－同意框架难以应对人工智能应用的复杂性。卡内基梅隆大学2023年研究显示，用户往往不清楚其数据被用于人工智能训练[7]。尤其在生成式人工智能应用中，用户很难预料其输入信息将被用于模型训练。

数据收集行为具有广泛性与隐蔽性。人工智能系统通过多种渠道间接获取用户数据。MIT 技术评论2024年报告称，超过60%的人工智能应用在用户无感知的情况下收集与服务无关的个人信息，其中15% 涉及敏感信息。

（二）模型训练阶段的隐私挑战

记忆效应是导致隐私泄露的重要因素。Google DeepMind 2024年研究表明，人工智能模型可能＂记住＂训练数据中的敏感信息[9]。即便原始数据已被删除，相关＂记忆＂仍可能留存，影响用户被遗忘权的实现。大规模语言模型在该问题上表现尤为突出。

2023年，加州大学伯克利分校研发了＂源自由认证反学习＂技术，无需原始训练数据即可实现隐私数据清除。该技术在 NeurIPS 2023会议上报告遗忘效率超过95%，且模型性能下降控制在2% 以内。目前该技术仍处于早期发展阶段。

（三）内容生成与输出阶段的风险

生成式人工智能可能导致非预期的隐私泄露。OpenAI 2024年报告提出，系统生成的内容可能包含训练数据中的个人信息。即便原始信息非敏感，通过关联分析仍可推断出用户敏感信息，此类推理攻击难以彻底预防。测试表明，约23% 的主流人工智能模型存在隐私泄露隐患。

多模态融合研究揭示了新的威胁形式。攻击者可通过在图像中嵌入人眼不可见的恶意指令窃取用户数据。该类方法依赖高分辨率图像实现，仅在图像质量降低时指令才会显现。

（四）新型攻击手段的涌现

人工智能环境下隐私攻击方式趋于复杂和隐蔽。成员推理攻击（membership inference attacks）可使攻击者判断某数据是否用于模型训练。模型反转攻击（model inversion attacks）则尝试从模型参数重建训练数据。2024年 IEEE 安全与隐私研讨会的研究显示，此类攻击成功率两年内提高了37%[9]。

数据投毒攻击构成另一新兴威胁。与直接入侵不同，数据＂投毒＂并不直接破坏系统，而是诱导模型学习偏差行为。2023年 USENIX 安全研讨会的研究演示了如何通过注入少量污染数据影响模型输出。此类攻击可逐渐侵蚀系统，为后续植入后门、窃取数据甚至从事间谍行为创造条件。研究表明，仅需污染0.1% 的训练数据，就可在特定条件下使模型输出攻击者预设的结果。

三、人工智能个人信息保护的防护策略

（一）技术手段

隐私增强技术（PETs）成为研究热点，主要包括联邦学习、差分隐私等。微软2024报告显示，采用联邦学习的企业数量增长150%[5]。该技术有效降低了数据集中带来的风险。华为2024年开发的联邦学习框架，在医疗场景中既保护了患者隐私，也提升了模型性能。

加密技术是保护个人信息的核心方法。端到端加密可保障数据在传输和存储过程中全程处于加密状态。IEEE 2024年研究表明，抗量子加密算法能有效抵御量子计算攻击。苹果公司在 iOS 18中升级加密技术，为 iCloud 数据提供更高级别的保护。

（二）治理与管理框架

有效保护需依托综合治理框架。德勤2024年报告指出，设立数据治理委员会的企业违规风险降低45%[11]。某社交媒体平台通过跨部门协作，构建了覆盖全生命周期的防护体系。

数据治理框架应包含数据分类分级、访问控制等要素。《生成式人工智能个人信息保护技术要求》系列标准为行业提供了全生命周期评估体系，该标准已在多家互联网企业实施并取得良好效果[2]。

（三）标准与规范建设

标准与规范为保护工作提供实施依据。ISO 2024年发布《人工智能隐私保护指南》，建立起统一实施框架。我国2024年标准化体系建设指南将七个组成部分纳入统一框架。

可解释人工智能（XAI）成为重要发展方向。MIT 2024年研究显示，采用 XAI 技术的系统用户信任度提升60%[4]。监管要求推动企业将可解释性模块嵌入人工智能系统。IBM 的 AI Explainability 360工具包在2024年升级后提供更丰富的算法支持。

四、案例分析与实践探索

（一）行业最佳实践

部分领先企业已开展有效的个人信息保护实践。腾讯公司2024年隐私保护白皮书显示，其＂三道防线＂治理体系成功拦截98% 的隐私泄露风险。阿里巴巴达摩院开发的隐私计算平台在2024年双11期间处理超10亿次查询，实现零隐私事件[2]。这些实践表明，技术与管理结合可有效保障个人信息安全。

在医疗人工智能领域，国家药监局推行的＂全生命周期质量

监管"取得明显成效。2024年《中国医疗人工智能发展报告》显示，采用合规辅助诊断系统的医院误诊率下降30%。联影智能开发的AI影像系统通过联邦学习技术在保护患者隐私的同时持续优化模型，该案例入选2024年世界人工智能大会最佳实践。

（二）典型事件分析

某大型科技公司2024年的数据泄露事件提供了重要教训。事后分析表明，缺乏端到端加密与访问控制机制不完善是主要原因。该公司后续投入5亿美元加强安全体系建设，包括部署新一代隐私计算平台和升级员工培训机制。该案例说明，安全防护体系需与人工智能系统风险水平相匹配。

2024年某自动驾驶公司的数据治理实践同样具有参考意义。该公司构建了覆盖数据采集、传输、存储及处理全流程的保护体系，运用差分隐私技术处理训练数据，并通过安全多方计算实现跨机构协作。相关措施不仅符合监管要求，也赢得用户信任，其用户满意度调查中隐私保护项得分达4.8/5[12]。

五、结论与建议

人工智能环境下的个人信息保护是一项系统工程，需技术、管理、法律等多维度协同推进。通过分析当前面临的主要挑战与防护策略，本研究得出以下结论：

首先，人工智能技术的发展为个人信息保护带来全新挑战，需采取综合治理策略。技术层面需广泛采用隐私增强与加密技术；管理层面应建立健全数据治理框架；法律法规则须提供清晰的规范依据。三者协同方可构建有效的防护体系。

其次，企业应将隐私保护理念融入产品设计全流程，践行"隐私保护始于设计"原则。最佳实践表明，早期投入的保护成本远低于事后补救支出。同时，应提升系统透明度，清晰向用户说明数据使用方式并提供控制选项。

基于上述研究，提出以下建议：

对监管机构，应完善法律体系，制订清晰、可操作的技术标准。建立分级分类监管机制，依据风险水平实施差异化监管。同时加强国际合作，推动全球标准协调一致。

对企业而言，应设立专职数据治理委员会，制定完善的数据安全管理制度。加大对隐私保护技术的研发投入，尤其是在联邦学习、差分隐私等前沿领域。定期开展隐私影响评估，及时识别与修复潜在风险。

对用户，应提高隐私安全意识，审慎共享个人信息。了解并积极行使访问、更正、被遗忘等合法权利。优先选择注重隐私保护的产品与服务，以市场力量推动行业整体进步。

人工智能发展不应以牺牲个人信息保护为代价。通过技术革新、管理优化与法律保障的多方协同，我们能够构建既促进创新又保护隐私的人工智能生态系统，实现数字时代的可持续发展。

参考文献

[1] 王利明 . 加强人格权立法保障人民美好生活 [J]. 四川大学学报 (哲学社会科学版),2018(3):5-10

[2] 张新宝 . 论个人信息保护的法律路径 [J]. 法学研究 ,2023,45(1):98-115.

[3] IDC. Worldwide Artificial Intelligence Spending Guide 2024[R]. 2024.

[4] McKinsey Global Institute. The State of AI in 2024: Generative AI's Breakout Year[R]. 2024.

[5] 杨强 , 刘洋 . 联邦学习 : 算法与应用 [J]. 计算机学报 ,2020,43(5):897-909.

[6] 国务院 . 关于构建数据基础制度更好发挥数据要素作用的意见 [Z].2022.

[7] Fredrikson, M., et al. Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing[C]. Proceedings of the 23rd USENIX Security Symposium, 2014: 17-32.

[8] Brown, T., et al. Language Models are Few-Shot Learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.

[9] Shokri, R., et al. Membership Inference Attacks Against Machine Learning Models[C]. 2017 IEEE Symposium on Security and Privacy (SP), 2017: 3-18.

[10] McMahan, B., et al. Learning Private Neural Networks Using Differential Privacy[C]. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016: 308-318.

[11] Yang, Q., et al. Federated Machine Learning: Concept and Applications[J]. ACM Transactions on Intelligent Systems and Technology, 2019, 10(2): 1-19.

[12] 最高人民法院关于审理使用人脸识别技术处理个人信息相关民事案件适用法律若干问题的解释 [Z].2021.

