生成式人工智能对个人信息保护的挑战与治理路径

万美秀
（南昌大学　法学院，江西　南昌　330031）

摘　要：以 ChatGPT 为代表的生成式人工智能技术给各行各业带来颠覆性变革，但也引发个人信息泄露、算法偏见、虚假信息传播等个人信息侵权危机。传统“基于权利保护”的路径过于强调个人信息保护而阻碍人工智能产业的发展，“基于风险防范”的路径则更加凸显个人信息的合理利用价值，价值选择上更优。但以权利保护和风险保护共同治理，才能实现利益平衡并建立个人信息的长效保护机制。在个人信息处理规则上，以“弱同意”规则取代僵化严苛的知情同意规则；在目的限制原则上，以“风险限定”取代“目的限定”；在个人信息最小化原则上，以“风险最小化”取代“目的最小化”。在此基础上，进一步加强生成式人工智能数据来源合规监管，提升算法透明性和可解释性，强化科技伦理规范和侵权责任追究。

关键词：生成式人工智能；ChatGPT；个人信息保护；治理路径
中图分类号：D913；TP399　　**文献标识码：**A　　**DOI：**10.19358/j.issn.2097-1788.2024.04.009
引用格式：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，2024，43(4)：53-60.

The challenge and governance path of generative artificial intelligence to personal information protection

Wan Meixiu
(Law School, Nanchang University, Nanchang 330031, China)

Abstract: Generative artificial intelligence technology represented by ChatGPT has brought disruptive changes to all walks of life, but also triggered personal information infringement crises such as personal information disclosure, algorithmic bias, and false information dissemination. The traditional "right protection-based" path overly emphasizes personal information protection and hinders the development of the artificial intelligence industry. The "risk prevention-based" path highlights the rational use value of personal information and is better in value selection. However, only by governing together with right protection and risk protection can we achieve a balance of interests and establish a long-term protection mechanism for personal information. In terms of personal information processing rules, the rigid and strict informed consent rules should be replaced by the "weak consent" rule; in terms of purpose limitation principles, the "purpose limitation" principle should be replaced by the "risk limitation"; in terms of personal information minimization principles, the "purpose minimization" principle should be replaced by the "risk minimization". On this basis, we should further strengthen the compliance supervision of generative artificial intelligence data sources, improve the transparency and interpretability of algorithms, and strengthen the standardization of scientific and technological ethics and the investigation of tort liability.

Key words: generative AI; ChatGPT; personal information protection; governance path

0　引言

以 ChatGPT 为代表的生成式人工智能掀起了全球第四次科技革命浪潮，成为带动全球经济增长的新引擎[1]。然而，作为新一代人工智能技术，生成式人工智能在不断迭代更新与变革生产关系的同时，也带来了诸多个人信息保护的法律风险。生成式人工智能的运行以海量用户的个人信息为基础，在输入端、模拟训练端、模拟优化端、输出端等各环节都离不开个人信息的使用。在大

规模的数据处理和不透明的算法黑箱背景下，生成式人工智能便产生了违法收集个人信息、制造虚假有害信息、算法偏见与歧视等问题。

对此，各国监管部门广泛关注，美国、法国、意大利、西班牙、加拿大等多国政府已宣布对 ChatGPT 进行调查监管，并出台了相应监管规范。2023 年 7 月 10 日，我国网信办等七部门也联合发布了《生成式人工智能服务管理暂行办法》（以下简称“《暂行办法》”），明确了促进生成式人工智能技术发展的具体措施，对支持和规范生成式人工智能发展作出了积极有力的回应。但需要注意的是，《暂行办法》对个人信息保护的规定仅在第 4、7、9、11、19 条中援引《个人信息保护法》的相关规定，对使用生成式人工智能技术侵犯个人信息权益呈现出的新问题缺乏专门规定，而继续延用《个人信息保护法》面临诸多适用困境。如何在促进生成式人工智能技术创新发展与个人信息安全之间寻求平衡，是新一代人工智能技术向人类提出的时代难题。鉴于此，本文拟以生成式人工智能技术的运行逻辑出发，分析生成式人工智能对个人信息保护带来的挑战，并以《民法典》《个人信息保护法》《暂行办法》体现的精神为线索，从个人信息保护的治理原则和治理路径方面展开讨论，在此基础上提出具体治理对策，以期为生成式人工智能技术应用对个人信息保护带来的系列问题提供初步解决方案，为解决人工智能时代个人信息保护问题作出有益探索。

1　生成式人工智能的运行逻辑

目前人工智能技术主要有两种类型：决策式人工智能/分析式人工智能（Discriminant/Analytical AI）和生成式人工智能（Generative AI）[2]。其中，决策式人工智能是利用机器学习、深度学习和计算机视觉技术来训练数据中的条件概率分布情况并做出决策，判断样本数据属于特定目标的概率。而生成式人工智能是利用深度神经网络学习输入和训练数据，并对已有的大规模数据集进行归纳总结，从中抽象出数据的本质规律和概率分布，再基于这些规律和概率分布情况生成新的数据。2014 年提出的“生成式对抗网络”深度学习模型最具影响力，其通过生成器和判别器使生成的数据富有原创性。此后，随着自然语言处理算法“循环神经网络”“预训练语言模型”“Transformer”等技术的突破，生成式人工智能迅速发展，广泛应用于内容生成、人机交互、产品设计等领域。以 ChatGPT 为例，由美国 OpenAI 公司推出的 GPT-4 是以 Transformer 模型为基础，预训练用于预测文档中的下一个指令，使用公开可用的数据（如互联网数据）和第三方提供商许可的数据，对来自人类的反馈强化学习模型进行微调[3]。经过预先训练，当用户输入问题时，ChatGPT 会将问题转换为计算机数据并使用算法模型形成相应的文本、图片、视频等数据集，通过不断改进和优化，最终从符合要求的数据集中输出具有一定原创性的新内容。其运行原理如图 1 所示。

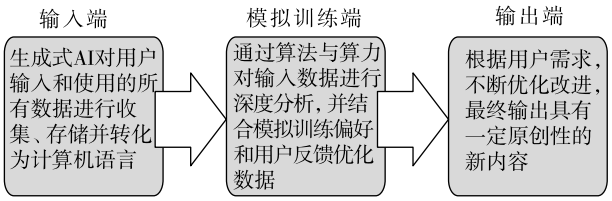


图 1　生成式人工智能的运行原理

从 ChatGPT 的底层运行逻辑可以看出，新一代生成式人工智能的发展得益于算法、算力与数据的应用和技术突破。在算法层面，它以预训练语言模型（LM）作为初始模型生成基本符合要求的内容，再收集数据并训练打分模型（BM）以评估生成内容是否符合人类的方式，最后通过强化学习（RL）迭代式更新打分模型以生成高质量且符合人类认知的内容[4]。在算力层面，生成式人工智能的运行需要有效地执行复杂的计算任务并通过不断训练和推理来优化生成内容。在数据层面，训练和优化人工智能模型需要大量的数据，而运用网络爬虫技术便可以获得来自社交媒体、公共机构、传感器等多渠道的海量数据。因此，生成式人工智能的不断优化与迭代发展，离不开上述算力、算法与数据三驾马车的驱动，数据是生成式人工智能训练的基础，算法是生成式人工智能优化的核心，算力则为生成式人工智能发展提供技术支撑和保障。然而，作为生成式人工智能训练基础的海量数据是开发者通过各种方式收集的，其中涉及大量的个人信息处理行为，开发者并没有完全依据《个人信息保护法》等相关规定来处理，给个人信息保护带来诸多风险和挑战。

2　生成式人工智能对个人信息保护的挑战

2.1　输入端：非法抓取与过度收集

生成式人工智能的输入端是个人信息泄露的源头，其法律风险主要集中在两个阶段：一是模拟训练端的初始数据库，二是模拟优化端的更新数据库。

从初始数据库来看，生成式人工智能存在大量非法抓取个人信息的“黑历史”，处理个人信息的告知同意规则被虚置。我国《个人信息保护法》《民法典》等明确规定了处理个人信息应当履行告知义务并取得个人同意，合理处理公开的个人信息则无须个人同意，但也应当履行告知义务[5]。以生成式人工智能 ChatGPT 为例，其初始数据库主要是利用网络爬虫技术从公开渠道获取的

2021 年之前的数据，其中包含大量账户信息、社交媒体信息、行踪轨迹等个人信息。然而大部分用户并不知晓个人数据被用于模拟训练，更谈不上“同意”。在深度学习与无监督式学习模式下，大量对个人权益有重大影响的公开个人信息被非法抓取，告知同意规则形同虚设。据此，对于现阶段已经抓取并应用于生成式人工智能模拟训练的初始数据库，应当如何确保其合理使用并防止对个人权益造成侵害便成为当下亟需解决的难题。

从更新数据库来看，生成式人工智能存在长期过度收集个人信息的“不良行为”，个人信息最小化原则被架空。与人类一样，生成式人工智能并不能凭借固有的知识体系一劳永逸地生存，其也需要不断更新数据以提高输出内容的准确度和可信度。但事实上，该阶段的个人信息收集和处理规则也并没有得到贯彻。

第一，目的限制原则面临适用困境。我国《个人信息保护法》第 6 条第 1 款规定，处理个人信息应当具备明确、合理的目的，并与处理目的直接相关。第 17 条规定，处理个人信息发生变更的应当及时告知。从 OpenAI 官网公布的企业隐私政策来看，其宣称可能将个人信息用于“改善服务、开发新的项目、防止滥用服务实施犯罪、进行业务转让等目的”[6]，但该表述具有高度的概括性和模糊性，对个人信息的保存期限、删除、变更告知情况也没有作出相应说明，用户只能选择接受否则便无法继续使用。此外，从技术层面看，目前生成式人工智能也无法自动识别“与处理目的有关的信息”，而是采取“一揽子概括协议”全部抓取，无疑加剧了个人信息权益侵害的风险。

第二，个人信息最小化原则面临适用困境。根据《个人信息保护法》第 6 条第 2 款规定，收集个人信息应当限于实现处理目的的最小范围，即所谓的“个人信息最小化原则”。从 OpenAI 官网公布的隐私政策第 1、2、3 条来看，其可以收集包括用户账户信息、通信信息、技术信息、社交信息、输入或上传的内容信息以及提供的其他任何信息。但诸如访问设备类型、操作系统、服务互动方式、其他任何可获取的信息等并非使用生成式人工智能服务所必备的信息，OpenAI 公司将所有用户信息全部囊括其中，显然属于过度收集个人信息的行为，违反个人信息最小必要原则。

第三，敏感个人信息处理规则面临适用困境。《个人信息保护法》将个人信息分为一般个人信息和敏感个人信息，由于敏感个人信息泄露将对个人人身、财产造成严重威胁，因而法律规定了特别处理规则。根据《个人信息保护法》第 28、29 条，处理个人敏感信息应当在特定目的和充分必要的情况下取得个人单独同意并采取严格的保护措施。然而，生成式人工智能在收集用户个人信息时并未作任何区分。更为重要的是，其将用户使用的所有历史信息传输至终端服务器并实时保存于云端，用于未来模型的优化训练。虽然 OpenAI 官网隐私政策第 2 条中宣称 ChatGPT 收集到的所有个人信息会进行汇总或标识化处理，但第 3 条随即指出将与第三方进行共享。而一旦借助第三方额外信息和有关技术手段，即使经过匿名化处理的信息仍然具有可识别性[7]。去标识化处理的个人信息将面临重新识别的风险，由此便加剧了个人信息泄露危机。2023 年 3 月 20 日 ChatGPT 就发生过部分用户聊天记录、信用卡付款信息和电子邮件等敏感个人信息泄露事件，引发各国监管部门对个人信息保护的担忧。由此可见，现行立法对生成式人工智能侵害个人信息权益的行为缺乏专门性规定，无法给个人提供明确的行为预期。

2.2 模拟训练端：算法黑箱和过度挖掘

在生成式人工智能的模拟训练端，离不开算法的运用，而不公开、不透明的“算法黑箱”引发个人数据侵权危机，处理个人信息的公开透明原则难以贯彻。根据《个人信息保护法》第 7 条、24 条规定，处理个人信息应当遵循公开透明原则，利用个人信息进行自动化决策的也应当保证决策的透明度和结果公平、公正。而生成式人工智能的算法运行的本质是数据输入、输出的过程，但在输入和输出之间存在无法解释的“黑洞”，引发“算法黑箱”问题[8]。更为重要的是，生成式人工智能的算法较此前的人工智能有了进一步提升，其并不遵循传统算法数据输入、逻辑推理、预测的过程，而是借助于深度学习模型逐渐具备了一定的自主学习、自主决策能力，直接在原始数据的基础上经过自主学习而生成新作品[9]。随着生成式人工智能算法自主学习的频次不断增加，算法不断迭代，导致技术隐层愈发复杂，而其逻辑又超越了一般大众所能理解的范围，加之信息上的不对称更加深了算法的不透明度与不可理解性，加剧了算法的“黑箱”属性，显然无法保障算法背后隐含的结果公平公正，直接违背个人信息处理的公开透明原则。目前 ChatGPT 至今未曾公布其算法规则，百度推出的“文心一言”、阿里云推出的“通义千问”等亦未公布，显然对《个人信息保护法》规定的公开透明原则提出了严峻挑战。

在模拟训练和模拟优化过程中，生成式人工智能通过深度学习的算法模型对个人信息过度挖掘，使得去标识化的个人信息甚至匿名化信息被重新识别，加剧了个人信息泄露风险。生成式人工智能对个人信息的使用并不局限于传统人工智能的简单加工，而是通过极强的推理能力进行深度挖掘，发现信息主体之间隐藏的内在联

系。如加州大学伯克利分校的一项研究表明，人工智能系统可以分析用户在 AR 和 VR 环境中的运动数据，从中推断出数百个相关参数，并以惊人的准确性揭示个人信息。事实上，在现有技术条件下即使生成式人工智能训练数据集中没有某人的个人信息，但结合其他信息在深度挖掘的基础上也可以推测出其特征，比如性别、年龄、种族、学历等。可见，新一代人工智能表现出极强的自主学习能力、深度合成能力和逻辑推理能力，对个人信息保护带来极大挑战。

2.3 输出端：算法偏见和虚假信息

在生成式人工智能的输出端，由于算法本身不具有技术中立性，而“算法黑箱”又加剧了算法非中立性，引发输出结果偏见。首先，在算法设计上，生成式人工智能的底层算法都是由带有主观偏好的开发者设计的，而开发者的固有认知偏见不可避免地会形成算法偏见。其次，在深度学习技术上，生成式人工智能的自主学习能力不断迭代发展，但机器学习不会对数据库中的信息进行价值取向筛选，导致生成式人工智能形成并加深开发者嵌入其中的算法偏见。最后，在数据来源上，模拟训练的数据质量参差不齐，大量虚假数据、缺失数据、污染数据、不全面数据输入导致最终生成带有歧视性的内容。另外，“算法黑箱”所具有的不公开、不透明性为“算法偏见”披上了合理的技术外衣，导致偏见行为难以被发现，从而加剧对特定群体的歧视和偏见，也给传统的平等权保护带来危机[10]。尽管 OpenAI 公司在其官网上声明，ChatGPT 已通过算法设置和模拟训练进行了优化，能在一定程度上拒绝用户不合理的请求，比如生成带有性别歧视、种族歧视、暴力、血腥、色情等违反法律、公序良俗的内容，但事实上，其给使用者和非使用者带来的风险依然存在。此前亚马逊便被爆出利用人工智能训练的算法进行招聘，存在重男轻女的性别歧视问题。可见，算法偏见呈现出种种不合理的区别对待，引发深层次的不平等和歧视问题。

在生成式人工智能的输出端，行为人还可以利用深度伪造、深度合成等技术生成虚假信息来实施侮辱诽谤、造谣传谣、财产诈骗等犯罪，《个人信息保护法》第 7 条规定的个人信息真实性、准确性无法得到保障。由于生成式人工智能对输入数据的真实性和准确性并没有甄别能力，因此它也并不保证输出结果的真实性和准确性，可能出现“一本正经地胡说八道”、输出“正确的废话”、制造虚假新闻等问题，从而侵犯个人信息权益。更为重要的是，这一缺陷很容易被不法分子利用来实施犯罪。2023 年 4 月 25 日甘肃洪某便利用人工智能技术炮制了一则“今晨甘肃一火车撞上修路工人，致 9 人死亡”的虚假信息牟利被警方立案调查。可见，生成式人工智能的出现导致大量虚假信息的生成和传播，侵害个人信息权益，引发严重的社会问题。

生成式人工智能对个人信息保护的挑战如图 2 所示。

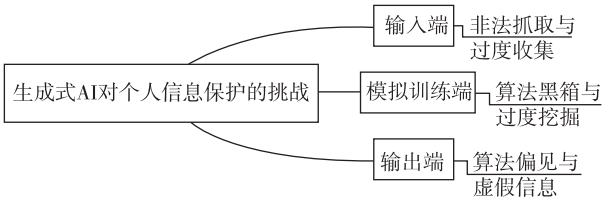


图 2 生成式人工智能对个人信息保护的挑战

3 生成式人工智能背景下个人信息保护的治理路径

3.1 “权利保护”与“风险防范”共同治理

基于上述，生成式人工智能对个人信息保护带来诸多风险和挑战。对此，《民法典》《个人信息保护法》《暂行办法》规定的传统个人信息保护规则均面临适用困境。究其根源，在于个体主义与静态化的个人信息保护进路难以适应科技的发展，亟需寻求更为合理的个人信息保护制度缓和二者之间的张力。基于以人为本的理念，要求强化个人信息保护；基于促进和规范人工智能产业的发展、鼓励创新的理念，要求对个人信息保护进行一定限制。因此，唯有正确认识并协调个人信息保护与生成式人工智能创新发展之间的关系，才能让人工智能更好地服务于经济的发展和社会的进步。

从总体监管原则来看，世界各国对生成式人工智能的发展存在“保守”与“开放”两种立法态度，并出台了相应法律法规进行规制。欧洲国家基于两次世界大战及法西斯大规模严重侵害人权的惨剧，高度重视人格尊严与人格自由等基本人权的保护[11]，因此，长期以来对人工智能的监管较为谨慎，采取“先规范后发展，稳步推动监管”的治理原则，以《通用数据保护条例》《可信 AI 伦理指南》确立了欧盟地区人工智能发展的伦理框架，以《人工智能法》《可信赖的人工智能伦理准则》进一步加强了可操作化法律规制。美国则基于 ChatGPT 产生的巨大影响以及维持自身在人工智能领域国际领先地位的需要，对人工智能的治理相对开放，采取“审慎监管以促进产业创新”的治理原则，相继出台《美国人工智能倡议》《人工智能能力和透明度法案》等以企业自我规制和政府规制相结合推进人工智能产业发展[12]。从我国《暂行办法》第 3 条来看，我国对生成式人工智能的发展总体上秉持开放包容的态度，稳步推进人工智能产业的发展。一方面，坚持以人为本的理念保障基本人权，维护个人信息和个人利益以实现个人自治。另一方面，兼

顾人工智能时代个人信息利用的新环境和新方式，对个人信息保护作出必要限制以维护公共利益和社会利益。换言之，在个人信息相对安全的前提下调整个人信息强保护规则，合理开发和利用个人信息以推动人工智能产业的发展，从而在个人权益保护与企业利益维护之间寻求平衡。

从具体个人信息保护规则来看，生成式人工智能背景下我国个人信息保护存在“基于权利保护”与“基于风险防范”两种路径。其中，“基于权利保护”路径源于美国1973年诞生的公平信息实践原则，其通过对个人进行信息赋权和对信息处理者施加义务的方式保障个体行使控制性权利[13]。但由于个人信息不仅关系到个人利益，还具有公共性和社会性[14]，个人信息强保护的规则难以维护公共利益并适应人工智能时代的发展。因此，一种“基于风险防范”的方法被提出，并逐步应用于各国个人信息保护的立法。2013年，知名智库数字欧洲提出了改革欧盟个人数据保护法的方案，从强化企业负责性而非信息主体的控制权利切入，要求企业设计规则防止风险的发生[15]。其后欧盟《通用数据保护条例》在修改其个人数据保护法时，就引入了这种“基于风险”（risk - based）防范的方法。在欧盟《人工智能法案》中也确立了以风险分级治理的规制路径并对各等级进行差异化监管。我国制定的《个人信息保护法》也体现了“基于风险”防范的理论。比如将个人信息区分为“一般个人信息”与“敏感个人信息”并且分别规定了不同的处理规则，实际上就隐含了一种先验的、抽象于具体场景的风险推定，即对敏感个人信息的处理可能对个人和社会产生较为严重的不利影响[16]。

笔者认为，“基于风险防范”理论能够更好地应对生成式人工智能对个人信息权益侵害带来的系列问题，适用该理论具有正当性。第一，《暂行办法》体现了我国政策制定者尝试从“基于风险防范”的治理路径出发解决生成式人工智能带来的个人信息保护难题。从《暂行办法》第5条第2款可以看出，个人信息处理者仍有义务采取适当措施来防范个人信息处理过程中可能出现的各种社会风险。从某种意义上看，该政策的出台也为未来人工智能领域法律的制定及风险防范理论的应用提供了有效指引。第二，“风险社会”要求“风险控制”。当代社会是一个“风险社会”，风险无处不在、不可预测且常常带来难以弥补的损害。一旦生成式人工智能收集的个人信息被泄露或不当使用，将给个人信息主体带来不可逆转的损害。因此，改变以往单一的赋权保护模式和事后追责机制，从风险防范的角度强化事先风险预防更具有制度优势，即从风险控制的维度构建个人信息的全面保护制度，强化信息处理者的风险防范责任与信息主体的个人预防责任。第三，“基于风险防范”的路径有利于实现利益平衡，促进人工智能产业的发展。相较而言，“基于权利保护”路径对个人信息进行“强保护”而忽视了个人信息的合理利用价值，无法应对新时代的发展和风险日益突出的现代社会个人信息侵权危机。“基于风险防范”路径则是一种折中治理方案，通过适当扩张个人信息合理利用的范围，从风险控制的角度强化信息处理者的风险防范义务与信息主体的个人风险责任，并对具体场景可能发生的风险进行事先预防与责任分配，在预防风险的发生与事后救济上价值选择更优。但需要注意的是，本文主张的“基于风险防范”的治理路径并非完全抛开“基于权利保护”来谈，而是弱化“强权利”保护模式以实现个人信息的合理利用价值。诚然，个人信息权益作为自然人最基本的人格权，仍然应当得到基本的权利保护。坚持“基于权利保护”和“基于风险防范”两种路径共同治理，才能实现各主体的利益平衡，构建个人信息的长效保护机制。

3.2 构建数据来源合规监管机制

解决生成式人工智能输入端的非法抓取和过度收集个人信息问题，要从数据源头预防，建立数据来源合规监管机制。对于初始数据库，由于信息权利人已经丧失了个人信息的自主控制权，应当寻求事后补救措施来维护其合法权益。第一，在技术层面上，服务提供者应当采取严格的保护措施防止个人信息泄露。比如对已经去标识化的信息采取脱敏、加密等技术手段进一步匿名化，使其无法重新识别到特定自然人。第二，在侵权责任承担上，要考虑生成式人工智能事先未经许可收集个人信息存在过错、对侵权行为发生没有尽到必要注意义务、事后未采取补救措施等因素对其加重处罚。倒逼服务提供者对已经收集而未经许可获取的个人信息原始数据库定期开展合规监测，强化其个人信息安全保障义务。

对于更新数据库，服务提供者也应当强化数据来源合规监管，严格遵循个人信息收集处理规则。第一，建立个人信息的影响评估机制。我国《个人信息保护法》第55条明确了个人信息处理者对特定个人信息处理的事先评估义务，其中包括处理敏感个人信息、对个人权益有重大影响的情形。个人信息影响评估是服务提供者处理个人信息的前提，也是其持续、稳定经营的基础。因此，服务提供者应当在个人信息处理前开展影响评估，自行评估爬取的数据来源是否合规，是否侵犯个人信息权益、他人知识产权、公平竞争权益等，根据不同影响采取相应保护措施。第二，构建个人信息分类分级监管机制。《暂行办法》第3条、第16条两次提到“分类分

级监管”，但并未具体说明。笔者认为，服务提供者在收集个人信息时，应当区分不同类型的个人信息，并确立不同的信息处理机制：（1）区分一般个人信息与敏感个人信息。对于一般个人信息的处理，僵化严苛的知情同意原则难以适应维护公共利益和数字经济发展的需要[17]，应当在个人信息保护与利用之间建立“弱同意”规则并采用“基于风险防范”路径要求服务提供者事先评估个人信息处理行为的合法性、合规性和合理性。在目的限制原则上，以“风险限定”取代“目的限定”，企业对个人信息的后续利用在不超过“原有程度、用户无法预测”的风险范围内无须用户再次授权，将风险控制在实现特定目的的合理水平。在个人信息最小化原则上，以“风险最小化”取代“目的最小化”，企业对个人信息的二次利用应当采取匿名化等措施将风险降至实现目的的最低水平[18]。但对于敏感个人信息则严格遵循告知同意规则，避免造成人格权益侵害。在必要情况下处理敏感个人信息的，严格采取匿名化等脱敏、加密技术措施，而非简单的去标识化处理。（2）区分对个人权益有重大影响与对个人权益无重大影响。服务提供者在信息处理之前，应当对个人信息进行风险评估。对个人权益有重大影响的，严格遵循告知同意规则取得个人单独同意。对个人权益无重大影响的，无需取得个人单独同意，但仍应当采取技术措施防止对个人权益造成侵害。第三，定期开展企业数据合规监测。生成式人工智能服务提供者应当建立长期的个人信息处理风险防范机制，定期对产品或服务中涉及个人信息处理的行为进行合规审查，发现潜在风险或安全隐患的及时采取必要措施加以防范。

3.3 提升算法的透明性和可解释性

生成式人工智能模拟训练端存在的“算法黑箱”问题，本质在于复杂的算法既无法观察，也难以为常人所理解。因此治理“算法黑箱”首先要打开“黑箱”，推动算法的公开化和透明化。但需要注意的是，算法的公开化、透明化并不意味着要公开算法的具体代码、编程等，而是要对算法作出必要说明和解释[19]。其原因在于，一方面，算法的源代码异常复杂，即使公开公众也很难理解，公开甚至会引发黑客攻击、被不法分子利用实施犯罪。另一方面，算法的公开成本较大，大部分涉及公司商业秘密，企业基于自身利益一般不会自觉公开。因此，推动生成式人工智能算法的透明化，要从算法的设计、算法功能、算法风险、算法逻辑、算法种类等涉及用户重大利益的方面进行公开说明，接受算法监管部门的审查和社会的监督，以保障算法公平、公正、负责。其次，要加强算法的可解释性。由于算法具有高度的技术性和复杂性，仅仅凭借公开难以令公众知晓算法背后的决策，因此要加强算法的可解释性，利用算法的可解释性技术最大程度揭示算法开发的过程、结果和应用经过，揭开算法自动化决策内部群体不平等的面纱[20]。比如欧盟《通用数据保护条例》第 12 条就规定了算法控制者负有以“简洁、透明、易懂、易获取并清晰直白的语言”提供信息的义务。换言之，算法解释必须以能够为一般人所知晓的程度来开展，否则算法解释就失去了意义。当然，对算法可解释性适用范围、技术要求等仍有待进一步研究。最后，引入第三方进行算法监管。探索引入第三方独立组织、支持学术性组织、非营利机构等专业机构对算法进行评估、审查、备案等，化解“算法黑箱”带来的个人信息侵害风险，实现算法安全、可控。目前德国已经发展出了由技术专家和资深媒体人挑头成立的非营利性组织以评估和监控影响公共生活的算法决策过程[21]。美国纽约州也颁布了《算法问责法案》要求将公民组织代表纳入监督自动化决策的工作组，以确保算法公开和透明[22]。我国目前针对算法的监管尚有不足，建立第三方独立机构监管有待进一步探究。此外，对个人信息过度挖掘问题同上述数据来源合规方面的监管机制类似，应当在生成式人工智能算法设计中进一步限制个人信息抓取的范围、目的和方式，以法律规制手段防范技术风险。

3.4 强化伦理规范和侵权责任追究

在生成式人工智能的输出端，算法偏见引发输出结果歧视，严重侵害个人信息权益。唯有对算法偏见善加治理，才能更好地利用算法造福人类。而算法偏见之所以会转化为算法歧视，本质在于人的作用，算法的开发者和使用者要为算法歧视负责[23]。因此，缓解算法偏见带来的算法歧视，其根源在于优化人工智能的伦理治理，坚持“以人为本”和“科技为民”的理念对人工智能进行开发设计。《暂行办法》第 4 条亦对此作出了回应。提供和使用生成式人工智能服务应当遵守伦理道德要求。第一，完善人工智能行业道德伦理规范，加强算法设计者的伦理审查和考核。通过定期开展科研伦理培训等对算法设计者的行为进行约束以强化其道德自律，并进一步提高算法设计者的行业准入门槛。第二，构建算法备案审查制度，强化事前监督。在算法研发后投入使用之前要求其向有关监管部门报备，经初步审查符合要求的准予进入市场应用，不符合要求的予以退回。通过监管部门的事前监督，可以有效防范存在严重偏见的算法投入市场。第三，建立算法分类分级管理和风险监测制度，健全问责机制。服务提供者要对算法进行分类分级管理，规制“信息茧房”导致的算法歧视。从损害结果出发，按照“谁设计谁负责，谁主管谁负责”的标

准进行事后问责，从源头上遏制与预防算法歧视[24]。第四，健全人工智能伦理风险评估机制，严格进行伦理规范审查。对于嵌入生成式人工智能的算法模型，服务提供者要开展自查和定期评估，梳理伦理风险的来源、种类、原因等并制定相应风险应对方案。算法设计要秉持平等、公平的理念，防止设计人员利用算法进行歧视。

对于生成式人工智能输出端带来的虚假信息治理问题，本质也是人的作用。行为人的非法目的诱使其利用生成式人工智能作为辅助工具制造或传播虚假信息、实施犯罪。因此，规制生成式人工智能带来的虚假信息问题，应当从侵权责任的事先预防、事中控制和事后处理入手。第一，在事先预防上，对生成式人工智能生成作品进行深度合成标识。生成式人工智能服务提供者要严格依据《互联网信息服务深度合成管理规定》《暂行办法》等规定，对深度合成内容进行标识和分类分级管理，对生成内容中可能引起公正混淆或误认的内容作出风险提示，推动生成式人工智能的透明化。使用深度合成标识技术，也可以有效追踪虚假信息来源，提高虚假信息识别率，同时追究相关责任人的主体责任。第二，在事中控制上，建立多元主体协同共管机制。考虑政府、人工智能企业、用户等主体在虚假信息的生成、传播与治理中的行为模式和参与度，建立平衡各方利益的监管机制。第三，在事后处理上，合理分配各方责任。生成式人工智能的研发者、使用者、服务提供者等主体在各自过错范围内承担虚假信息生成、传播的法律责任。基于鼓励创新的理念，适用过错责任原则，同时基于生成式人工智能侵害个人信息权益的侵权主体多元性，需要根据具体情况分析各方主体责任，对服务提供者类推“通知删除”规则[25]。由此，进一步完善利用生成式人工智能侵害个人信息权益的侵权责任追究制度。

综上，生成式人工智能下个人信息保护的治理路径如图 3 所示。

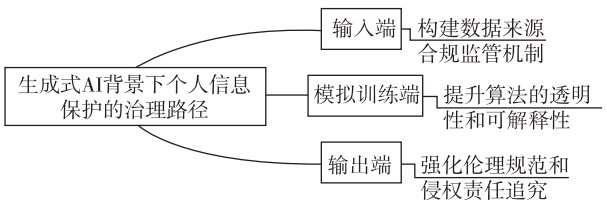


图 3　生成式人工智能下个人信息保护的治理路径

4　结论

放眼全球，生成式人工智能的技术革新给世界各国带来了巨大的发展机遇，但与此同时也引发了个人信息泄露、算法偏见、虚假信息传播等诸多个人信息侵权危机。究其本质，在于如何平衡个人信息权益保护与科技创新发展之间的关系。“基于权利保护”路径过于强调个人信息保护，僵化严苛的告知同意规则难以适应人工智能时代的发展，“基于风险防范”路径则适度扩张个人信息合理利用的范围并综合考虑各责任主体的风险防范义务，具有稳定性和前瞻性。但应对生成式人工智能对个人信息保护带来的挑战，权利保护和风险防范是两个不可或缺的维度。坚持以人为本和鼓励科技创新发展的理念，要进一步加强生成式人工智能输入端、模拟训练端、模拟优化端、输出端等各环节的风险管控，实现个人信息保护与利用之间的平衡。着眼于未来，我们要更加关注科技发展给伦理道德、人格权保护带来的系列冲击，加强人格权保护制度研究，以实现保障基本人权与科技进步之间的平衡。

参考文献

[1] 麦肯锡. 生成式人工智能的经济潜力：下一波生产力浪潮［EB/OL］.（2023－06－14）[2024－01－10]. https：//www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#introduction.

[2] HUANG S, GRADY P. GPT－3, Generative AI：a creative new world［EB/OL］.（2022－09－19）［2024－01－10］. https：//www.sequoiacap.com/article/generative-ai-a-creative-new-world/.

[3] OpenAI. GPT-4 technical report［R/OL］.（2023－03－14）［2024－01－10］. https：//openai.com/gpt-4.

[4] 朱光辉，王喜文. ChatGPT 的运行模式、关键技术及未来图景［J］. 新疆师范大学学报（哲学社会科学版），2023，44（4）：113－122.

[5] 程啸. 论公开的个人信息处理的法律规制［J］. 中国法学，2022（3）：82－101.

[6] OpenAI. Enterprise privacy at OpenAI［EB/OL］.（2023－11－14）［2024－01－10］. https：//openai.com/policies/privacy-policy.

[7] 张涛. 风险预防原则在个人信息保护中的适用与展开［J］. 现代法学，2023，45（5）：52－72.

[8] 胡小伟. 人工智能时代算法风险的法律规制论纲［J］. 湖北大学学报（哲学社会科学版），2021，48（2）：120－131.

[9] 陈兵，董思琰. 生成式人工智能的算法风险及治理基点［J］. 学习与实践，2023（10）：22－31.

[10] 崔靖梓. 算法歧视挑战下平等权保护的危机与应对［J］. 法律科学（西北政法大学学报），2019，37（3）：29－42.

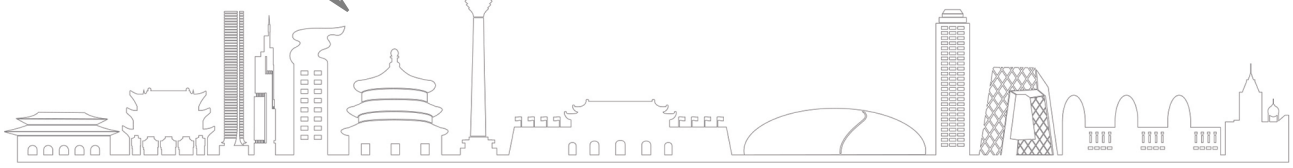
[11] 程啸. 论大数据时代的个人数据权利［J］. 中国社会科学，2018（3）：102－208.

[12] 毕文轩. 生成式人工智能的风险规制困境及其化解：以ChatGPT的规制为视角 [J]. 比较法研究，2023（3）：155－172.
[13] 丁晓东. 论个人信息法律保护的思想渊源与基本原理——基于“公平信息实践”的分析 [J]. 现代法学，2019，41（3）：96－110.
[14] 高富平. 个人信息保护：从个人控制到社会控制 [J]. 法学研究，2018，40（3）：84－101.
[15] 赵鹏. “基于风险”的个人信息保护？ [J]. 法学评论，2023，41（4）：123－136.
[16] 张涛. 探寻个人信息保护的风险控制路径之维 [J]. 法学，2022（6）：57－71.
[17] 高志宏. 大数据时代“知情－同意”机制的实践困境与制度优化 [J]. 法学评论，2023，41（2）：117－126.
[18] 范为. 大数据时代个人信息保护的路径重构 [J]. 环球法律评论，2016，38（5）：92－115.
[19] 徐凤. 人工智能算法黑箱的法律规制——以智能投顾为例展开 [J]. 东方法学，2019（6）：78－86.
[20] 周翔. 算法可解释性：一个技术概念的规范研究价值 [J]. 比较法研究，2023（3）：188－200.
[21] 张淑玲. 破解黑箱：智媒时代的算法权力规制与透明实现机制 [J]. 中国出版，2018（7）：49－53.
[22] 谭九生，范晓韵. 算法“黑箱”的成因、风险及其治理 [J]. 湖南科技大学学报（社会科学版），2020，23（6）：92－99.
[23] 孟令宇. 从算法偏见到算法歧视：算法歧视的责任问题探究 [J]. 东北大学学报（社会科学版），2022，24（1）：1－9.
[24] 石颖. 算法歧视的发生逻辑与法律规制 [J]. 理论探索，2022（3）：122－128.
[25] 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.

（收稿日期：2024－01－24）

作者简介：

万美秀（1997－），女，硕士研究生，主要研究方向：民商法、计算机法学。



版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部