生成式人工智能训练语料的个人信息保护研究

张新宝*

内容提要 生成式人工智能训练语料的个人信息保护应当秉持鼓励和支持创新的基本立场。为确保服务提供者的个人信息利用需求能够得到满足,可以在训练端对《个人信息保护法》作适当宽松解释或例外规定。对于已公开的个人信息,可以通过宽松解释“公开目的”将其纳入可处理的范围。对于未公开的个人信息,仍需要以个人同意作为处理行为的合法性来源,但是可以通过宽松解释目的限制原则、调整“告知—同意”的相关规则,缓解服务提供者面临的困难。技术壁垒的提高加剧了信息主体的劣势地位,需要确保个人信息保护请求权的行使,以维护个人的合法权益,但是其行使不可避免受到技术现实的限制。服务提供者应严格履行包括技术措施在内的个人信息安全保护义务,尽可能降低给个人信息带来的风险。保护机制整体上应以行政监管为主导,如果侵害个人信息权益造成损害,应允许服务提供者以“符合行政监管要求”作为不存在过错的抗辩。

关键词 生成式人工智能 训练语料 个人信息保护 行政合规抗辩

目 次

* 中国人民大学法学院教授、民商事法律科学研究中心研究员。

一、问题的提出

生成式人工智能的应用市场正在不断扩大，中美等科技强国纷纷布局生成式人工智能，争取在新一轮的科技革命中抢占先机。训练生成式人工智能需要海量的高质量语料作为支撑，而个人信息数据具有真实性、多样性、连贯性以及大规模等特征，恰好可以满足生成式人工智能研发对高质量训练语料的需求。然而，其合法性界限尚未得到明确。一方面，将个人信息数据用作训练语料会带来一定风险；另一方面，生成式人工智能研发需要处理尽可能多的高质量数据，因此在适用《个人信息保护法》时面临着困境，若严格地适用《个人信息保护法》，可能会加剧训练语料的短缺。

（一）语料短缺与个人信息保护之间的矛盾

我国当前面临训练语料尤其是高质量中文语料不足的困境，限制了生成式人工智能技术的发展。全球通用的 50 亿大模型数据训练集里，中文语料占比仅为 1.3%。[①] 虽然我国数据资源丰富，但是尚未得到充分挖掘，而且数据权属不清，导致数据流通不足。个人信息语料被限制使用意味着服务提供者需要在数据预处理阶段将个人信息数据剔除，这不仅会恶化我国语料短缺的局面，而且会给服务提供者造成较重的数据处理负担。从我国数字经济发展的角度考虑，应当允许将更多个人信息数据用作训练语料，以缓解当前语料不足的困境。作为重要的生产要素，数据（包括个人信息数据）对于经济社会的重要价值已经愈发显现。习近平总书记强调，发挥数据的基础资源作用和创新引擎作用，加快形成以创新为主要引领和支撑的数字经济。[②] 数据的充分利用是数字经济高质量发展的重要前提，促进个人信息数据的流通和利用，对于“做大做强数字经济，增强经济发展新动能，构筑国家竞争新优势”具有重要意义。

生成式人工智能目前仍然处在早期阶段，存在安全方面的不足和个人信息泄露风险。相较于过去的人工智能技术，生成式人工智能可以输出特定内容，输出端的个人信息风险是其特殊性之所在。模型通常只有在出现过拟合等情况下才会记忆训练数据，但有研究表明，模型可能在没有过拟合的情况下无意中记忆训练数据中的个人信息。[③] 有研究显示，可以通过让 ChatGPT 重复“诗歌”“公司”“发送”“制造”和“部分”等词语来喷出其记忆的部分训练数据。[④] OpenAI 表示，GPT-4 可能了解那些在公共互联网上有重要影响力的人，比如名人和公众人物，而且还可以综合多种不同的信息类型，并在特定的输出中执行多个推理步骤；可以完成多个可能与个人和地理信息相关的基本任

① 参见罗云鹏：《大模型发展亟需高质量“教材”相伴》，载《科技日报》2024 年 1 月 15 日，第 6 版。

② 参见《习近平在中共中央政治局第二次集体学习时强调 审时度势精心谋划超前布局力争主动 实施国家大数据战略加快建设数字中国》，载《人民日报》2017 年 12 月 10 日，第 1 版。

③ *See* Nicholas Carlini, *Evaluating and Testing Unintended Memorization in Neural Networks*, at https://bair.berkeley.edu/blog/2019/08/13/memorization/ (Last visited on April 9, 2024).

④ *See* Jai Vijayan, *Simple Hacking Technique Can Extract ChatGPT Training Data*, at https://www.darkreading.com/cyber-risk/researchers-simple-technique-extract-chatgpt-training-data (Last visited on April 9, 2024).

务，例如确定与电话号码相关的地理位置或者回答教育机构位于何处，而无需浏览互联网。[⑤] 可见，即便泄露用户个人信息的概率非常小，但如果刻意加以引导和提示，仍可能用来生成包含个人信息内容的回答。[⑥] 此外，不法分子可能会对生成式人工智能实施攻击以获取训练语料中的个人信息，包括成员推理攻击、模型萃取攻击、模型逆向攻击等；或者利用 API 的安全漏洞，微调模型以降其安全性，进而获取个人信息。[⑦]

（二）个人信息作为训练数据的制度困境

若严格适用“告知—同意”和相关的配套规则，服务提供者在获取个人信息语料时需要频繁地取得个人的同意，可能导致服务提供者无法获取必要的个人信息语料。利用未公开个人信息训练生成式人工智能的行为通常不属于《个人信息保护法》第 13 条第 1 款第 2—7 项的情形，所以服务提供者只有根据第 1 项的规定取得个人同意，处理才具备合法性。“告知—同意”规则一直发挥着平衡个人信息利用与保护的制度功能，个人有权决定其个人信息是否以及如何被处理。对于未公开的个人信息而言，除非出现必须处理的特殊情况或者为了更高位阶的利益而对其自主决定的权利进行合理限制，否则“个人同意”都承担着规范处理者行为的阀门作用。只有在生成式人工智能研发是为了维护公共利益或者是为了维护该自然人的合法权益的情形，才可以不适用“告知—同意”规则。目前生成式人工智能主要由互联网企业进行研发，往往以营利为主要目的，并非是为了维护公共利益或者该自然人的合法权益，因此，难以将“告知—同意”之外的规定作为一般情况下的合法性依据。利用已公开的个人信息虽然不需要以个人同意作为处理的合法性基础，但是若超出合理范围以至于对个人权益有重大影响，则需要取得个人同意。然而，语料所涉及的信息主体并非都是服务提供者的用户，可能缺少取得个人同意的有效方式和渠道，适用“告知—同意”规则会给服务提供者带来难以克服的阻碍。此外，对于信息主体而言，频繁接收个人信息处理者的告知，久而久之也会不堪其扰。

生成式人工智能的发展当然不能以威胁甚至侵害个人信息权益为代价，个人信息的安全和自然人享有的个人信息权益即构成训练行为的合法性边界。近期，我国先后发布了《生成式人工智能服务管理暂行办法》（以下简称《暂行办法》）和《生成式人工智能服务安全基本要求》（以下简称《基本要求》），以回应生成式人工智能的治理需求；针对训练数据的安全问题，专门发布了《网络安全技术　生成式人工智能预训练和优化训练数据安全规范（征求意见稿）》（以下简称《安全规范》）。这些文件对生成式人工智能服务、服务提供者、训练语料等概念以及服务安全、训练数据的安全等问题进行了规定，但是对训练语料的个人信息保护问题回答得过于笼统，仅仅作了原则性的规定，不足以

⑤ *See* OpenAI, *GPT-4 Technical Report*, at https://arxiv.org/pdf/2303.08774.pdf (Last visited on April 9, 2024).

⑥ 参见孙祁：《规范生成式人工智能产品提供者的法律问题研究》，载《政治与法律》2023 年第 7 期，第 165 页。

⑦ *See* Kellin Pelrine et al., *We Found Exploits in GPT-4's Fine-tuning & Assistants APIs*, at https://far.ai/post/2023-12-exploiting-gpt4-api/ (Last visited on April 9, 2024).

指导和规范实践。我国人工智能立法工作已经正式启动，[⑧] 为此，本文将从训练语料个人信息保护的基本立场与指导思想出发展开讨论，希望通过分析，提出一个合理的个人信息保护方案，服务于我国人工智能法的制定。

二、个人信息语料获取困境的解决

（一）以“数据二十条”为指导平衡产业发展与个人信息保护

“数据二十条”的核心思想在于促进数据合规高效流通使用、赋能实体经济，充分实现数据要素价值，促进全体人民共享数字经济发展红利。鉴于生成式人工智能高度关系到国家、社会和个人的利益，应当坚持支持创新的基本立场，顺应加快构建数据基础制度、激活数据要素潜能的政策导向，尽可能在满足产业对个人信息利用需求的前提之下保护个人信息的安全，最大限度地协调生成式人工智能产业的发展和个人信息的保护。

1. 支持生成式人工智能创新的基本立场

中国参与签署的《布莱切利宣言》指出：“人工智能为全球带来巨大的机会，具备改变和提高人类的福祉、和平和繁荣的潜力。”强人工智能已经成为国家间竞争的前沿阵地。[⑨] 未来，人工智能或许会承担起科技基础设施的角色，开发“主权人工智能”有助于捍卫本国的数据主权，避免数字殖民。生成式人工智能不仅具备巨大的经济潜力，[⑩] 还可以带动各行各业的转型，目前已经应用于国防、金融、医疗等重要领域。国防方面，人工智能不仅可以作为实现致命性自主武器系统的关键技术大幅提高系统作战能力并扩大其对敌威慑程度，[⑪] 而且在情报分析、决策制定、网络安全维护等方面都有用武之地。金融方面，人工智能正被应用于股票交易执行以及计算保单赔付，还推动了为投资和贷款决策寻找替代数据的趋势，催生了“所有数据都是信贷数据”的口号。[⑫] 医疗方面，人工智能可以发挥疾病预测和治疗、药物研发等功能，如 DeepMind 公司开发的人工智能模型 AlphaFold 可以预测蛋白质结构。[⑬] 此外，生成式人工智能在司法、教育、制造、城市

---

⑧ 《国务院 2023 年度立法工作计划》和《国务院 2024 年度立法工作计划》均将“人工智能法草案”列入预备提请全国人大常委会审议项目。参见《国务院办公厅关于印发〈国务院 2023 年度立法工作计划〉的通知》（国办发〔2023〕18 号）；《国务院办公厅关于印发〈国务院 2024 年度立法工作计划〉的通知》（国办发〔2024〕23 号）。

⑨ *See* Matthew R. Gaske, *Regulation Priorities for Artificial Intelligence Foundation Models*, Vanderbilt Journal of Entertainment and Technology Law, Vol.26:1, p.7-8 (2023).

⑩ 麦肯锡指出，生成式人工智能的跨行业应用预计每年可提供 2.6 万亿美元至 4.4 万亿美元的经济效益。*See* Michael Chui et al., *The Economic Potential of Generative AI: The Next Productivity Frontier*, at https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20digital/our%20insights/the%20economic%20potential%20of%20generative%20ai%20the%20next%20productivity%20frontier/the-economic-potential-of-generative-ai-the-next-productivity-frontier.pdf (Last visited on April 9, 2024).

⑪ 参见齐亚双等：《人工智能国防战略的目标、愿景与实施路径：国际经验与启示》，载《情报杂志》2024 年第 3 期，第 55 页。

⑫ *See* Ross P Buckley et al., *Regulating Artificial Intelligence in Finance: Putting the Human in the Loop*, Sydney Law Review, Vol.43:43, p.48 (2021).

⑬ *See* John Jumper et al., *Highly Accurate Protein Structure Prediction with AlphaFold*, Nature, Vol.596:583 (2021).

建设等领域同样拥有广阔的应用空间和巨大的应用潜力。

党和国家高度重视人工智能研发的“头雁”效应，全面部署了相关工作：2017 年，国务院颁布《新一代人工智能发展规划》，指出人工智能是国际竞争的新焦点、经济发展的新引擎；2018 年 10 月 31 日，习近平总书记在中共中央政治局第九次集体学习上强调，加快发展新一代人工智能是我们赢得全球科技竞争主动权的重要战略抓手，是推动我国科技跨越发展、产业优化升级、生产力整体跃升的重要战略资源；⑭ 2022 年，科技部等六部门颁布《关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见》，着力解决人工智能重大应用和产业化问题；2024 年 3 月 5 日，《2023 年国民经济和社会发展计划执行情况与 2024 年国民经济和社会发展计划草案》提请十四届全国人大二次会议审查，提出“人工智能 +”行动，即推动人工智能技术与经济社会各领域深度融合，支撑各行业应用创新，赋能百业智能化转型升级，提高生产效率，激发创新活力，重塑产业生态，培育经济发展新动能，形成更广泛的以人工智能为创新要素的经济社会发展新形态。地方也积极响应党中央部署，全力推进人工智能的创新与应用，北京、上海、深圳等多地都制定了相关地方性法规和规章。总之，生成式人工智能研发是建设数字中国、推进中国式现代化的关键步骤，我们应从国家发展全局的视角出发来思考其治理问题，将支持和鼓励创新作为当前生成式人工智能个人信息风险治理的重要考虑因素。

2. 平衡产业发展与个人信息保护的基本思路

面对训练语料的个人信息保护问题，应当充分考虑生成式人工智能技术发展对国家、社会、个人的重要意义。个人信息保护制度的根本目的在于将风险控制在合理范围内，实现个人信息利用与安全之间的平衡。禁止或过度限制使用个人信息作为训练语料，或者让《个人信息保护法》为生成式人工智能的研发“开绿灯”，显然都不足为取。鉴于生成人工智能的重要意义，训练语料的个人信息风险治理不应成为技术发展的阻碍，虽然需要未雨绸缪，但是整体上应持有包容审慎的态度，鼓励和支持生成式人工智能的发展。作为解决生成式人工智能训练语料个人信息保护问题的基本思路，可以在训练端适当放松而相应地在其他环节收紧：原则上允许使用个人信息数据作为训练语料，并结合生成式人工智能的技术特征来解释《个人信息保护法》的规定，必要时可以作出例外规定，以实现个人信息（尤其是一般个人信息）的最大化利用，满足生成式人工智能研发对个人信息的利用需求；同时，应当确保个人信息保护请求权的行使，要求服务提供者尽到严格的个人信息安全保护义务，尽最大努力消除生成式人工智能研发给个人信息（尤其是敏感个人信息）带来的风险。如此，可以最大程度地兼顾生成式人工智能技术的发展和个人信息权益的保护，实现两者的平衡。

生成式人工智能虽然会给个人信息带来一定的安全隐患，但是风险可控可化解。限制将个人信息作为训练语料无非是出于风险预防的考虑，但是应仔细考量风险的大小以

⑭ 参见《习近平在中共中央政治局第九次集体学习时强调 加强领导做好规划明确任务夯实基础 推动我国新一代人工智能健康发展》，载《人民日报》2018 年 11 月 1 日，第 1 版。

及采取预防的必要性。新型技术的出现向来都会带来一定的风险，但是牺牲创新从来也都不是化解技术风险的一个可行方案。侵害的高度可能性是风险预防的基本前提——如果风险只是存在，但发生概率及严重程度尚不清楚，那么风险就只是可能的、抽象的风险，然而风险预防需要付出的是具体的成本，如产业的发展受到一定程度的限制，此时，风险预防的正当性就有所减弱。事实上，生成式人工智能并不像公众所担忧的那样会造成不合理的个人信息风险。虽然域外发生了一些泄露个人信息的安全事件，但我国暂且没有类似的情况发生。另外，生成式人工智能会在何种程度上输出个人信息也尚未可知。既然目前风险尚未成为现实，就没有理由过度强调预防而限制对个人信息的处理，可以等到相应的问题实际发生后再作出应对和调整。退而言之，即使生成式人工智能的研发会给个人信息带来不合理风险，但也会为国家和个人带来不可估量的福利，因此，为了更高的利益目标而对个人信息权益作一定的限制亦具有合理性。此外，技术领域同样在寻求解决方案，未来完全可能通过技术手段来化解生成式人工智能带来的个人信息风险。

（二）基于平衡理念对个人信息保护法的调适

基于兼顾产业发展与个人信息保护的考虑，应当对《个人信息保护法》的规定作出有利于生成式人工智能发展的解释，以满足生成式人工智能训练对个人信息数据的利用需求，并在必要时作出例外规定。生成式人工智能训练语料的个人信息保护，本质上仍然是个人信息的利用与保护如何协调的问题。应当继续坚持“两头强化，三方平衡”的基本立场，[15]强化一般个人信息在生成式人工智能研发中的利用和敏感个人信息的保护。

1. 已公开个人信息的处理

互联网上的公开数据是生成式人工智能的主要语料来源，[16]其中包括了已公开的个人信息数据。是否可以收集已经公开的个人信息作为训练语料，高度关系到生成式人工智能产业的发展。

（1）理论层面的考量

数据的公开可访问性并不意味着可以被“不分青红皂白”地收集或使用，[17]个人信息已经公开也不意味着可被任意地用于生成式人工智能的训练。我国《个人信息保护法》对已公开个人信息采取了弱保护的模式，力度低于未公开的个人信息，但并不是放弃保护。根据《个人信息保护法》第 27 条的规定，可以利用已公开的个人信息用于生成式人工智能训练，但不得超过合理范围，如果对个人权益有重大影响，应当取得个人同意。事实上，“合理的范围”和“对个人权益有重大影响”的判断是同一个过程的两个

---

⑮ 参见张新宝：《从隐私到个人信息：利益再衡量的理论与制度安排》，载《中国法学》2015 年第 3 期，第 49-52 页；张新宝：《论作为新型财产权的数据财产权》，载《中国社会科学》2023 年第 4 期，第 156-159 页。

⑯ 例如，通过借助 Common Crawl 等开放的网络爬虫数据库，ChatGPT 等大语言模型可以收集并使用互联网上任何没有被特别保护的内容进行训练。*See* Benjamin Fabre, *Generative AI Is Scraping Your Data. So, Now What?*, at https://www.darkreading.com/vulnerabilities-threats/generative-ai-is-scraping-your-data-so-now-what (Last visited on April 9, 2024).

⑰ *See* Office of the Privacy Commisioner of Canada, P*rinciples for Responsible, Trustworthy and Privacy-protective Generative AI Technologies*, at https://www.priv.gc.ca/en/privacy-topics/technology/artificial-intelligence/gd_principles_ai/ (Last visited on April 9, 2024).

侧面：如果是在合理的范围内处理，则不会对个人权益有重大影响；如果对个人权益有重大影响，则超出了合理的范围。而且第 27 条采取了比较模糊的表述，导致已公开个人信息处理的合法性判断存在较大解释空间。

一般认为，处理个人自愿公开的个人信息以推定同意为合法性基础；处理依法强制公开的个人信息以目的一致为合法性基础。

就个人自愿公开的个人信息而言，自愿公开行为可以被推定为同意（默示的同意），代表个人已经同意他人在可预期的风险之内处理个人信息——自然人既然自愿将其个人信息公开，就应当清楚其个人信息可能会被他人处理，且可能带来一定的权益侵害风险，因此无需再次取得同意。而且，虽然整体来看，当前社会公众对生成式人工智能缺乏足够的信任，接受程度尚且不高，通常不会希望自己公开的个人信息被用作训练语料，但事实上，个人权益被侵害的风险并不会因为用作训练语料而增加。通常认为，“公开”本身即为高风险的个人信息处理行为，使得个人信息处在一个无法完全受个人控制、随时可能被他人获取的状态，作为理性的自然人，应当清楚公开个人信息可能带来的风险以及后果，并且谨慎实施公开行为。换言之，自愿公开即意味着主动将其个人信息暴露在较高的风险之中。因此，只要没有给个人造成更高的风险，便可以直接处理该已公开的个人信息，以实现“个人信息权益保护与合理行为自由维护之间的协调与平衡”⑱。而生成式人工智能训练中对个人信息的学习不同于将生成式人工智能作为个人信息处理的工具，其目的不在于获取特定的信息，而在于学习其中的规律，更不会利用个人信息挖掘潜在联系。⑲ 机器学习过程给个人信息带来的风险主要是泄露，但已公开个人信息已经处在可以被不特定第三人接触的状态，即便发生泄露也不会给个人带来更高风险。况且，只要服务提供者能够严格尽到安全保护义务，便可以避免信息泄露。因此，可以认为使用自愿公开的个人信息训练生成式人工智能属于合理范围之内的处理行为。

就依法强制公开的个人信息而言，强制公开行为反映了个人信息权益与公共利益之间的平衡——为了公共利益目的之实现而对个人信息权益作出合理的限制。后续的处理行为需要具备和公开目的一致的处理目的，才能延续强制公开的合法性。换言之，判断是否可将依法强制公开的个人信息用作训练语料的关键在于处理目的与公开目的是否一致。⑳ 如作严格解释，利用依法强制公开的个人信息作为训练语料似乎违背了公开目的，因而合法性存疑，但如果从促进产业发展的角度对处理目的作宽松解释，“用作训练语料”亦可属依法强制公开的目的范围之内。例如，司法公开的目的在于“保障人民群众知情权、参与权、表达权和监督权，促进提升司法为民、公正司法能力”，㉑ 如作严格解

⑱ 程啸：《论公开的个人信息处理的法律规制》，载《中国法学》2022 年第 3 期，第 92 页。

⑲ *See* Karl Manheim & Lyric Kaplan, *Artificial Intelligence: Risks to Privacy and Democracy*, The Yale Journal of Law & Technology, Vol.21:106, p.120-121 (2019).

⑳ 参见张新宝、昌雨莎：《已公开裁判文书中个人信息的保护与合理利用》，载《华东政法大学学报》2022 年第 3 期，第 14-16 页；前注 ⑱，程啸文，第 98-101 页。

㉑ 参见《最高人民法院关于进一步深化司法公开的意见》（法发〔2018〕20 号）。

释，生成式人工智能训练的直接目的在于技术研发而非保障人民群众的知情权、参与权等，但生成式人工智能作为重要的技术工具已经运用于司法，不仅如此，最高人民法院还颁布了《关于规范和加强人工智能司法应用的意见》，以推动人工智能同司法工作深度融合。可见，虽然生成式人工智能训练本身并不直接促进司法公开，但其最终成果可以作为促进司法公开的手段，间接地对司法公开产生推动作用。按照这个思路，基础大模型几乎可以服务于任何公共目标，利用依法强制公开的个人信息用作训练语料间接地契合了个人信息公开的目的，可被归为合理范围之内的处理行为。而训练运用于特定领域的垂直大模型虽然无法按照该思路使用依法强制公开的个人信息作为训练语料，但通常只需要借助特定类型的数据微调基础大模型，借助未公开的个人信息即可满足需求。

（2）国际竞争与产业发展层面的考量

美国采取的是排除保护已公开个人信息的模式，[22] 因此已公开个人信息并不会对其生成式人工智能研发造成限制。根据欧盟《通用数据保护条例》（GDPR）第 6（1）（f）条的规定，如果处理对于控制者或第三方所追求的正当利益具有必要性，则处理行为合法。欧盟《人工智能法案》第 59 条允许在人工智能监管沙盒中出于公共利益处理为其他目的合法收集的个人信息数据，以支持人工智能创新。英国信息专员办公室也认为，“合法利益”可以作为网络抓取训练生成式人工智能的合法基础，但是需要进行三项合法性测试：一是处理目的具有合法性；二是处理的必要性；三是受损害的个人利益范围不得超过开发者的合法利益。[23] 可见，美国、欧盟和英国的个人信息保护制度在生成式人工智能训练语料的获取方面具有一定“优势”。站在国际竞争的角度考虑，我国应当允许将已公开个人信息用作训练语料，以避免陷入被动局面。

而从产业发展来看，如果人工智能企业面临过高的个人信息合规难度，不仅无法有效化解个人信息风险，反而可能会使制度目的落空。公开数据中的部分个人信息数据是生成式人工智能训练的客观需要，部分则是因为混杂在其他数据中而被收集。如果认为使用已公开个人信息数据训练会对个人权益造成重大影响，将导致服务提供者面临两难困境：完全剔除或取得个人的同意都难以实现。对于前者而言，虽然数据投入训练之前会经过相当复杂的数据清洗过程，包括剔除多余数据、补充缺失数据、修正、错误、数据等。但是，能否对所有的个人信息数据都作出实质性的判断，存在一定的可行性疑问，而且投入的成本也是不得不考虑的因素。对于后者而言，已公开个人信息具有非接触性特征，服务提供者难以与个人取得联系。在既难以取得信息主体的同意又难以将其中的个人信息完全剔除的情形下，服务提供者可能会不得已选择以违法的方式处理已公开个人信息。作为缓和，可以宽松解释《个人信息保护法》第 27 条的规定，相应地要求服务提

㉒ 参见丁晓东：《公开个人信息法律保护的中国方案》，载《法学》2024 年第 3 期，第 5-6 页。

㉓ *See* Information Commissioner's Office, *Generative AI First Call for Evidence: The Lawful Basis for Web Scraping to Train Generative AI Models*, at https://ico.org.uk/about-the-ico/what-we-do/our-work-on-artificial-intelligence/generative-ai-first-call-for-evidence/ (Last visited on April 9, 2024).

供者严格地履行个人信息安全保护义务，尽可能降低给个人信息造成的风险。

2. 未公开个人信息的处理

（1）已收集的信息：宽松适用目的限制原则

判断可否将已收集的未公开个人信息用作训练语料，关键看其是否超出了初始的处理目的。《个人信息保护法》第 6 条规定了目的限制原则，要求信息处理者在收集个人信息时应有明确、合理的目的，且在后续的处理过程中不偏离此目的。根据《个人信息保护法》第 14 条第 2 款，如果用于生成式人工智能训练超出了初始的处理目的，服务提供者只有重新取得个人同意才能处理该个人信息。但时间因素对于科技竞争至关重要，虽然数据在以极快的速度产生，[24]但数据的积累毕竟需要一个过程，短时间内可能会对技术研发产生较大影响。我国本身就面临严峻的语料不足问题，若已经收集的个人信息数据无法得到充分利用，可能会进一步加剧我国在生成式人工智能发展上的劣势。因此，可以在今后收集个人信息时告知个人用于生成式人工智能训练的目的并取得同意。

鉴于产业发展的客观需求，可以认为将已经收集的未公开个人信息用于生成式人工智能训练没有超出初始的处理目的，无需重新取得个人同意。数字技术发展迅速，个人信息的利用需求愈加广泛，处理者收集个人信息时无法完全预见未来是否可能出现新的处理目的。在进行语料训练时，处理者可能已经不再具备信息收集时的条件，难以重新获得信息主体的同意，否则需要付出不成比例的努力。[25]甚至，处理者已经和个人失去了直接联系，事实上不具有重新取得同意的可行性。如果要求后续的使用严格符合初始目的，难免会制约个人信息的利用。因此，“不宜片面强调信息处理对于初始目的的严格遵循，而应要求信息处理者将信息处理可能引发的风险控制在合理范围之内，以符合大数据时代信息多元利用的趋势”[26]。可供参照的是，GDPR 采取了一种“窄进宽出”的模式，[27]虽然要求处理目的应具体、明确、合法，但是根据第 6（4）条，并不完全禁止初始目的之外的处理，不过需要考虑初始目的与进一步处理目的之间的关联性、进一步处理可能造成的结果等因素。《个人信息保护法》虽然没有作出类似的规定，但笔者认为，应当综合考虑处理的性质、风险等因素来判定是否超出了初始目的。

事实上，训练人工智能的过程本质上是“学习数据”，而非“分析数据”或“记忆数据”，一般情况下不会直接反映出数据中的内容，将已经收集的个人信息数据用作训练语料并不会给个人带来更高的风险。利用个人信息进行数据分析对个人信息的利用具有

---

[24] IDC 最新发布的 Global DataSphere 2024 报告显示，中国的数据量预计将从 2023 年的 30.96ZB 激增至 2028 年的 97.06ZB，5 年复合增长率 CAGR 高达 25.7%。参见《合规场景双引擎驱动——2023 年中国数据库审计市场份额报告发布》，载微信公众号“IDC 咨询”，2024 年 7 月 19 日上传。

[25] *See* János Mészáros & Chih-Hsing Ho, *Big Data and Scientific Research: The Secondary Use of Personal Data under the Research Exemption in the GDPR*, Hungarian Journal of Legal Studies, Vol.59:403, p.404 (2018).

[26] 朱荣荣：《个人信息保护“目的限制原则”的反思与重构——以〈个人信息保护法〉第 6 条为中心》，载《财经法学》2022 年第 1 期，第 30 页。

[27] 参见梁泽宇：《个人信息保护中目的限制原则的解释与适用》，载《比较法研究》2018 年第 5 期，第 20-24 页。

直接性，通常是通过挖掘数据中的内容得出结论（如通过用户的浏览数据分析其偏好），其过程是将分布在海量数据中的零散信息集中起来，通过统计分析等方法挖掘其中的有效信息。而生成式人工智能训练完全不同，其对个人信息的使用具有明显的间接性。“学习”是深度学习技术的核心与本质所在，“训练”实际上是一个让机器发现和学习规律的过程。生成式人工智能的功能和水平由庞大规模的参数决定，代表了其“知识储备”。未经训练之前，各个参数都处在未知状态，通过海量数据的训练，参数的值（映射规则）得以确定下来，生成式人工智能便获得了回答人类提问的能力。可见，生成式人工智能的输出过程并不是对训练数据的重新组合或者直接调取，而是通过复杂的映射规则来处理用户的提问，然后将得到的内容反馈给用户。除非出现过拟合等特殊情况，或者受到外部攻击，否则不会直接记忆并输出个人信息。

此外，正如上文所述，生成式人工智能具有重要的科技战略价值，关系到国防、经济、医疗、教育等诸多关键领域的发展。尤其是，基础大模型未来可能会承担起科技基础设施的角色，为此，可以允许服务提供者直接使用已经收集的未公开个人信息作为训练语料，不用取得个人同意。虽然过去在医疗、金融等领域也会涉及个人信息的处理，但是其目的往往在于提供和优化某种特定服务。生成式人工智能的意义远不局限于此，作为一种技术工具，其对于社会的影响具有革命性。GDPR 第 5（1）（b）条规定，因为公共利益、科学或历史研究或统计目的而进一步处理数据，不视为违反初始目的。而且 GDPR 序言第 159 条指出：“以科学研究为目的的个人数据处理应以广泛的方式解释，包括例如技术开发和示范、基础研究、应用研究和私人资助的研究。”这些相关内容可以为我国目的限制原则的理解提供借鉴。当然，目前生成式人工智能的研发主要由互联网企业开展，将其解释为纯粹的科学研究难免有些牵强，但生成式人工智能训练具有较高的科学研究属性不可否认，可以作为宽松适用目的限制原则的依据。

（2）未收集的信息：集中取得个人同意

鉴于生成式人工智能研发的特殊性，应当在适用“告知—同意”及相关规则的时候作出符合技术特征的调整。随着生成式人工智能的发展和普及，个人信息处理者完全能够意识到自己收集的个人信息可能会用于生成式人工智能研发，因此属于可以预见的处理目的和处理方式，可以在收集个人信息时一并告知可能会用于生成式人工智能训练并取得同意，便于日后将收集的个人信息用作训练语料。

难点在于，根据《个人信息保护法》第 23 条的规定，个人信息处理者向第三方提供其处理的个人信息时，需要取得个人的单独同意，并向个人告知接收方的名称或者姓名、联系方式、处理目的、处理方式和个人信息的种类等信息。不同于过去的个人信息处理场景，生成式人工智能训练需要借助各种途径获得高质量的训练数据，商业数据即是一种重要来源，这意味着个人信息数据会在不同的主体之间流通，而且可以预见，随着数据权属问题得到明确，个人信息数据的流通将会变得更加频繁。如果每次将个人信息提供给第三方都需要作出告知并取得个人的单独同意，无疑会极大地影响数据的使用效率，

甚至可能会阻碍行业的创新，不符合当前鼓励人工智能发展的政策导向。[28] 笔者认为，考虑到训练语料的流通需求，可以允许个人信息处理者集中地取得向不同人工智能企业提供个人信息的同意，缓和“告知—同意”规则给生成式人工智能研发造成的限制，以促进训练语料的流转。如此一来，个人信息处理者无需在向第三方流转个人信息时频繁地征求个人的单独同意，只需集中地告知个人并取得概括的同意之后，便可以直接将收集的个人信息流转给不同的人工智能企业；服务提供者与其他个人信息处理者通过交易获取个人信息语料时也无需取得个人同意。由此，语料获取的难度得到极大的降低，从而满足生成式人工智能研发对个人信息的利用需求。集中告知应涵盖处理的目的、方式，且表明可能向第三方提供并应作出例外规定，如果第三方的姓名、联系方式、保存期限等信息尚不能确定，可以暂时不予告知，但是应充分告知可能对个人产生的影响以及相关权利的行使方式和程序等内容。这种集中取得同意的方式只是针对性地在告知的内容和方式上作出了调整，并不会对个人造成明显的不利影响，个人仍然可以自主决定其个人信息是否可被用作训练语料、是否可被提供给第三人。不可否认，集中告知增加了后续处理的不确定性，由于没有充分告知情况，可能会增加个人信息被不当处理的风险，但这些风险完全可通过强化个人信息安全保护义务、规范服务提供者的处理行为来化解。

除非有充分的必要性，应当避免将敏感个人信息数据用作训练语料，尤其是用于基础大模型的训练。但在确实需要使用敏感个人信息训练生成式人工智能时，同样可采取上述集中取得个人同意的方式。《基本要求》规定：“使用包含敏感个人信息的语料前，应取得对应个人单独同意或者符合法律、行政法规规定的其他情形。”《安全规范》也作出了类似的规定。鉴于敏感个人信息高度关系到人格尊严、人身和财产安全，确有必要对敏感个人信息的处理作出必要的限制。《个人信息保护法》第 28 条第 2 款规定：“只有在具有特定的目的和充分的必要性，并采取严格保护措施的情形下，个人信息处理者方可处理敏感个人信息。”问题的关键在于，用于生成式人工智能的训练是否属于“具有特定的目的和充分的必要性”。笔者认为，对此应区分不同模型进行判断。对于基础大模型而言，没必要用包含敏感个人信息的数据来训练，因为这对提升其功能水平的作用有限（自然人的生物识别、特定身份、医疗健康等信息可能并不会对特定功能的取得产生实质作用），但是会增加泄露风险，难谓具备充分的必要性。虽然基础大模型需要学习海量的数据来提高泛化能力，但是个人信息数据在训练数据中的比重相对较小，敏感个人信息数据更是如此，因此敏感个人信息在其中的作用可能微乎其微。而敏感个人信息作为高度关系到自然人的人格尊严、人身和财产安全的个人信息类型，风险系数相对较高，因此保护优先于利用。当然，无需完全禁止将敏感个人信息用作训练语料，只是如果使用敏感个人信息训练基础大模型，服务提供者应对必要性（例如，可以显著提高基础大模型的水平）进行详细说明，并且应当评估相关风险，同时做好充分的安全保护措

[28] 参见毕文轩：《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》，载《比较法研究》2023 年第 3 期，第 164 页。

施。而对垂直领域的人工智能模型的训练而言，因为需要应用于特殊和敏感的行业或领域（包括医疗、金融、人脸识别等），则更需要敏感个人信息作为训练语料，可以认为具有充分的必要性。但不管是哪种大模型训练，未成年人的个人信息都应重点予以限制。

（3）不宜普遍认定为合理使用

处理个人信息符合个人信息合理使用情形的，无需取得个人同意。问题在于，是否需要将获取未公开个人信息语料纳入合理使用的适用范围？纳入显然更有利于服务提供者获得充足的训练语料、提高我国生成式人工智能研发的竞争优势。但笔者认为，不宜普遍地将使用未公开个人信息数据作为训练语料的行为认定为合理使用，原因主要在于：首先，通过宽松解释目的限制原则和调整“告知—同意”的相关规则基本已经可以解决未公开个人信息语料的获取难题。即使普遍引入合理使用，由于个人信息往往是在使用网络产品和接受网络服务的过程中产生，服务提供者仍然需要从其他个人信息处理者处取得未公开个人信息语料。如果可以宽松解释《个人信息保护法》中的相关规定或者作出有针对性的例外规定，使得个人信息处理者将其收集的个人信息流转用作训练语料时不用再取得个人同意，其实最终的效果与引入合理使用无异，因此需要斟酌引入合理使用制度的必要性。其次，虽然集中取得个人同意会给服务提供者带来一定的成本，但是这对于服务提供者而言是应当支出的合理成本。通过适当的宽松解释和调整，服务提供者已经可以直接使用已公开的个人信息作为训练语料，获取未公开的个人信息也只需要与特定的个人信息处理者进行磋商和交易，并不会给其造成难以克服的困难。再次，部分情况下个人信息数据并非必要的训练语料。通常而言，个人信息包含的内容较短，模型不易从中学习到语言的一般规律。相较于使用较高风险的个人信息数据，服务提供者可能有更好的选择。对于非必要的个人信息数据，如果允许服务提供者受到合理使用制度的庇护，只会徒增个人信息风险。最后，对于敏感个人信息而言，更不宜通过合理使用进行使用。除非有充分的必要性，应当限制将敏感个人信息作为训练语料。如果允许其受到合理使用的庇护而无需取得个人的单独同意，带给个人的损害可能会远大于服务提供者节省的成本，难以认为具有合理性。

虽然通过合理使用制度解决作品语料获取的困境得到了比较广泛的认可，[29] 但是个人信息语料与作品语料存在如下差异：第一，个人信息的产生和收集往往存在一个“中心点”——网络平台等个人信息处理者，个人信息较为集中地处在个人信息处理者的控制之下，尤其是大型的个人信息处理者掌握了海量主体的个人信息；而作品语料极其分散，虽然我国成立了文字作品等集体管理组织，但是大量作品仍在集体管理组织的管理之外。第二，使用作品语料主要涉及的是权利人的著作财产权，而使用个人信息语料关系到权利人的人格权益，因此考虑个人信息的语料获取问题，应更加重视权利人的保护。第三，作品是生成式人工智能学习的关键内容，对于提高学习效果有重要作用；而个人

[29] 参见林秀芹：《人工智能时代著作权合理使用制度的重塑》，载《法学研究》2021 年第 6 期；焦和平：《人工智能创作中数据获取与利用的著作权风险及化解路径》，载《当代法学》2022 年第 4 期。

信息数据在许多情况下不具有必要性，因为我们“希望模型能了解世界，而非个人”[30]。第四，服务提供者基于“告知—同意”规则获取个人信息不需要向个人支付费用，但是基于授权许可获取作品需要向著作权人支付报酬，后者可能会给服务提供者带来超过收益的成本，但是前者并不会带来过高的经济负担。综合以上原因，对作品语料和个人信息语料的获取不能作简单的等同处理，对于后者，应当重视对“告知—同意”及相关规则的调适，而慎重考虑普遍地认定为合理使用。

笔者认为，可以在总结实践经验的基础之上，尝试将确实存在使用需求又难以通过集中取得个人同意来获取充足语料的情形认定为合理使用，但是应确保个人的合法权益不会受到严重损害。未经个人同意处理个人信息的“合理性”来源无非两种：其一，为了维护公共利益而对个人的合法权益进行一定程度的限制；其二，为了实现权利主体的优先利益而对劣后的利益进行限制——这正是《民法典》第 1036 条第 3 项所规定的两种情形。若是为了维护公共利益而处理个人信息，应当确保对个人的影响轻微。如果在维护公共利益的同时严重损害了个人的合法权益，同样不宜认为是合理使用。虽然生成式人工智能研发高度关系到国家、社会、个人的利益，但是很多情况下难以认为研发活动纯粹是为了维护公共利益或维护信息主体的合法权益。不过，部分情况下仍然存在《民法典》第 1036 条第 3 项的适用空间，例如训练征信等需要涵盖全国人民相关个人信息的生成式人工智能模型。此时，一方面，服务提供者难以通过集中取得个人同意来满足对个人信息的处理需求；另一方面，训练此种模型具有突出的公共利益属性，同时也是为了维护信息主体的合法权益，因此存在适用合理使用的正当性和必要性。

三、个人信息权益的实现与保障

（一）个人信息保护请求权的行使

语料获取的适当放宽会给个人信息造成一定的风险，需要在研发的其他环节及使用阶段尽可能降低风险，以实现个人信息利用与安全的平衡。生成式人工智能技术壁垒的提高加剧了个人的劣势地位，而个人信息保护请求权的行使可以推动权益保护的实现。《暂行办法》第 11 条第 2 款规定：“提供者应当依法及时受理和处理个人关于查阅、复制、更正、补充、删除其个人信息等的请求。”服务提供者应当依法处理个人的权利请求并尽可能予以满足，但是个人信息保护请求权的实现不可避免地受到技术现实的限制。囿于篇幅，下文仅通过对查阅复制权、删除请求权和解释说明权的分析来说明个人如何向服务提供者行使权利，以及技术特征可能对权利行使产生的影响。

1. 个人查阅、复制权的行使

查阅、复制权是为了满足作为实体权益的知情权而设置的个人信息保护请求权。[31]

[30] Open AI, *Our Approach to AI Safety*, at https://openai.com/blog/our-approach-to-ai-safety (Last visited on April 9, 2024).

[31] 参见张新宝：《论个人信息保护请求权的行使》，载《政法论坛》2023 年第 2 期，第 27-28 页。

根据《个人信息保护法》第 45 条第 2 款，权利人可以随时向个人信息处理者请求提供其个人信息。法律不应对其行使设置要件，个人只需证明自己是信息主体，而无需证明存在其他正当利益。[32] 但是，如何向服务提供者行使查阅、复制权，需要结合数据存储和查阅的技术能力和行使成本等因素进行分析，并且受到上述因素的合理限制。

首先，需考虑服务提供者是否有能力满足查阅和复制请求。受技术水平限制，服务提供者可能无法准确地从数据库中查阅到特定的个人信息。模型训练对数据规模的要求极高，如 GPT-4 的训练需要大概 4 万亿至 8 万亿个单词。[33] 要在如此大规模的数据集中精准、高效地查阅到特定信息，无疑对数据存储和查阅技术提出了挑战。随着数据库技术的发展，目前已有很多数据库可以满足生成式人工智能训练对非结构化数据的存储、查阅等需求，如 NoSQL 数据库、时序数据库、向量数据库等。服务提供者收到查阅、复制的请求后，可从数据库中调取个人信息数据，满足个人的权利请求。例如，向量数据库在人工智能研发中发挥着不可替代的作用，通常需要通过“词嵌入”（embedding）把文本、图片、视频等训练数据转化为机器更容易理解的数学向量，以提高数据的存储和检索能力，并更好地解决训练数据更新的问题。如果采用向量数据库，在个人提出查阅、复制的请求之后，服务提供者可从数据库中查询向量，然后通过逆向映射得到个人信息。

其次，需考虑查阅、复制的范围和成本的问题。虽然数据库技术可以帮助服务提供者实现在海量数据中的查阅，但若允许个人不受限制地行使查阅、复制权，难免会给服务提供者带来不合理的负担。虽然行使查阅、复制权可能会给处理者带来一定的成本，但为了保障权利的顺利行使，《个人信息保护法》并没有收取费用的规定。当然，行使个人信息保护请求权的成本不应完全由个人承担，否则无疑违背了权利设置的初衷。但是，任由个人行使查阅、复制权，忽略可能给服务提供者带来的负面影响，同样也不可取。笔者认为，查阅、复制权的行使应以服务提供者的客观技术能力和合理成本为限，若个人的查阅、复制请求超出了技术可行范围，则应当受到限制；若个人的请求超出了合理限度，服务提供者也可拒绝或收取相应费用。可行的解决方案是，服务提供者应免费满足个人合理频次（如一年一次或两次）的查阅、复制请求；如果超出合理频次则可要求个人说明正当理由，否则可以按成本收取费用；如果存在反复请求甚至恶意请求的情况，由于其违背了诚实信用原则，服务提供者可以拒绝请求或者收取更高费用。

2. 个人解释说明请求权的行使

生成式人工智能至今仍然是一个“黑箱”，令人难以理解其运行机制和工作原理。可解释性难题使得信息主体不可避免地对生成式人工智能产生不信任，担心可能会输出其个人信息。《个人信息保护法》第 48 条规定了解释说明请求权，依此，如果权利人提出请求，服务提供者应充分告知利用其个人信息的有关情况，这在一定程度上可以帮助个人了解生成式人工智能训练对个人信息的处理机制，缓解对模型安全性的担忧。然而，

㉜ 参见程啸、王苑：《论我国个人信息保护法中的查阅复制权》，载《法律适用》2021 年第 12 期，第 25 页。

㉝ 参见前注 ①，罗云鹏文。

解释说明请求权的实现同样不得不受到技术现状的制约。生成式人工智能借助了深度神经网络技术，具有极其复杂的结构。基础大模型的参数已经达到千亿级别，它们共同决定生成式人工智能的功能，导致输入到输出之间的逻辑不够清晰，难以清楚地观察和解释模型为何会输出特定回答。要求人工智能模型实现算法透明在客观上相当困难，且强制透明化可能会阻碍神经网络技术的应用。[34] 因此，服务提供者可能难以解释个人信息如何被学习和对模型产生影响，导致解释说明权无法得到完全行使。

“通常情况下，开发者没有义务对外披露人工智能的研发过程，包括研发中使用的训练数据。这不仅是开发者保有其技术秘密的正当性使然，而且是科学技术研究自由的内在要求。”[35] 但是，服务提供者应尽可能帮助个人理解生成式人工智能学习个人信息数据的整个过程。首先，应公开训练语料中个人信息数据的来源等信息。其次，个人提出算法解释要求时，服务提供者应当以清晰易懂的语言向个人解释生成式人工智能训练的基本原理，包括学习数据的过程、是否可能输出其个人信息等，换言之，算法解释的方式应当符合信息主体的知识水平。需要注意的是，算法解释的目的是解释输出结果的逻辑和机制，而非算法本身。即使算法完全透明化，用户或公众也未必能理解。[36] 目前，技术领域正致力于提高生成式人工智能的可解释性，如可视化技术、可解释性模型、对抗性样本等。麻省理工学院科学家的研究简报《人工智能和工作的未来》指出，人工智能模型可以通过一些实践变得更加透明，例如构建更可解释的模型、开发可用于探索不同模型如何工作的算法等。[37] 随着生成式人工智能可解释性水平的提高，服务提供者应当提供更为详细的解释，以充分满足个人的权利请求。

3. 个人删除请求权的行使

如果信息主体明确拒绝利用其已公开个人信息训练人工智能，或者撤回对未公开个人信息的同意，根据《个人信息保护法》第 47 条第 1 款的规定，个人可以行使删除请求权。然而，个人信息数据可能已经通过训练过程对参数的确定发挥了作用，个人行使删除请求权之后，服务提供者是否需要对模型重新训练，以达到让模型“遗忘”该个人信息数据的效果，恢复到没有学习该个人信息数据的状态？如果这样可能破坏数据库或者模型的功能，是否还应当满足信息主体的请求？

删除请求权的行使在生成式人工智能场景下有一定的特殊性，彻底删除特定个人信息数据不仅可能存在技术上的障碍，而且可能对数据库或者模型功能产生破坏性影响。即便服务提供者尽可能采用行业内认可的先进数据库，但客观上仍然可能出现无法删除的技术障碍。大型数据库中往往具有大量内置机制和故障安全措施，如自动备份、恢复到以前版本等避免数据丢失和损坏的措施。实践中，数据经常存储在多个地方，可能很

---

[34] 参见商建刚：《生成式人工智能风险治理元规则研究》，载《东方法学》2023 年第 3 期，第 14 页。

[35] 刘文杰：《何以透明，以何透明：人工智能法透明度规则之构建》，载《比较法研究》2024 年第 2 期，第 126 页。

[36] 参见周尚君、罗有成：《数字正义论：理论内涵与实践机制》，载《社会科学》2022 年第 6 期，第 168 页。

[37] *See* Tsedal Neeley, *8 Questions about Using AI Responsibly, Answered*, at https://hbr.org/2023/05/8-questions-about-using-ai-responsibly-answered (Last visited on April 9, 2024).

难识别和删除所有的“副本”；而且删除一个文件时，即便清空了保存该文件的空间，也仍然没有真正地将其从数据库中删除，只有当它被一个新的文件覆盖时，数据才真正消失。[38] 此外，通过识别数据所存储的所有空间并及时用新的信息覆盖来实现彻底的删除，还可能会严重危害数据库的一致性、稳定性，甚至会破坏系统安全性，以致损毁数据库。[39] 而且，缺少部分训练数据还可能会影响其功能的正常实现。[40] 如此一来，权利人要求服务提供者删除其个人信息，就可能要以破坏数据库或者模型的功能作为代价。因此，如果受到客观技术能力的限制，服务提供者可以不予删除，但是应当停止除存储和采取必要的安全保护措施之外的处理。

更复杂之处还在于，经过机器学习过程，被请求的个人信息数据可能已经对模型参数的确定产生了影响，训练数据集改变之后，可能需要重新训练，参数才能改变。如果想要实现彻底删除的效果，需要实施机器反学习，模型才能达到彻底“遗忘”该数据的状态。实现机器反学习的方法包括彻底的机器反学习和不彻底的机器反学习，前者是指通过重新训练模型消除特定数据对模型的影响，后者是指借助重新训练之外的方法实现机器反学习，如直接修改部分参数。直接修改参数虽然便捷，但是较为粗略，所以效果有限，可能无法实现彻底删除。而且，生成式人工智能模型中存在明显的“核心区域”，某个关键参数发生变化就可能会对模型的整体功能产生“致命”影响。[41] 因此，如果想要确保将模型恢复到没有学习该个人信息的状态，唯一理想的解决办法就是重新训练。但是以重新训练作为实现删除请求的方式并不可取，原因显而易见——这需要耗费大量的时间和经济成本。虽然借助 SISA（Sharded, Isolated, Sliced, and Aggregated training）等方法可以较为高效地实现重新训练，但是也存在降低模型准确率等缺点。[42]

综上，删除请求权的行使应当受到一定的限制：首先，受到技术发展现状的限制。由于物理上彻底删除难以实现，只要个人信息数据达到无法被利用并且安全的状态，即可认为实现了删除。如果技术上无法删除或者实现删除将带来不合理的成本，服务提供者可以《个人信息保护法》第 47 条第 2 款作为抗辩。不过，“所谓技术上难以实现，应当从客观标准进行理解，即结合当前的技术条件是否可删除进行判断，否则将导致信息处理者寻找各种理由和借口不予删除，实质上架空删除权的实效性”[43]。服务提供者应当提供详细的说明，避免以技术不可行为借口推脱责任。其次，受到服务提供者利益的限制。若生成式人工智能已经得到充分的学习，缺少特定数据不会对其产生实质的影响，个人可以请求删除；但若缺失对应数据会对数据库或者模型产生实质影响，破坏其完整性，甚

---

[38] *See* Tiago Sérgio Cabral, *Forgetful AI: AI and the Right to Erasure under the GDPR*, European Data Protection Law Review, Vol.6:378, p.383-384 (2020).

[39] 参见翟凯：《论人工智能领域被遗忘权的保护：困局与破壁》，载《法学论坛》2021 年第 5 期，第 145 页。

[40] *See* Tiago Sérgio Cabral, *supra* note 38, 388.

[41] 参见张奇：《只修改一个关键参数，就会毁了整个百亿参数大模型？》，载微信公众号“CSDN”，2024 年 2 月 18 日上传。

[42] *See* Bjørn Aslak Juliussen, Jon Petter Rui & Dag Johansen, *Algorithms that Forget: Machine Unlearning and the Right to Erasure*, Computer Law & Security Review, Vol.51:1, p.8-9 (2023).

[43] 王利明：《论个人信息删除权》，载《东方法学》2022 年第 1 期，第 44-45 页。

至影响其功能的实现，则应当对权利的行使进行限制。此时，服务提供者同样可以《个人信息保护法》第 47 条第 2 款作为抗辩，或者证明此时删除请求权的行使不符合诚实信用原则的要求，属于权利的滥用。再次，受到生成式人工智能原理的限制。个人无权要求服务提供者重新训练模型，只能请求将其个人信息从数据库中删除，并在下次重新训练时使用不包含该个人信息数据的语料。如果服务提供者可以判断被请求的个人信息对哪些参数产生了影响，并且修改参数的成本在合理范围之内，则应在不损坏模型功能的前提下通过修改参数实现机器反学习。最后，删除请求权的行使如果超过了合理频次且没有正当理由，服务提供者可以拒绝或者收取相应的费用。

（二）个人信息安全保护义务与侵权责任

作为在语料获取问题上作宽松处理的“对价”，服务提供者应尽到严格的个人信息安全保护义务，以最大程度地降低个人信息风险。具体而言，服务提供者应当采取与信息敏感性相称的措施，保护在生成人工智能的整个生命周期中收集或使用的任何个人信息，并且持续关注可能出现的威胁，[44] 例如，采取关键词过滤等措施避免模型输出个人信息，采用隐私计算等技术防止未经授权的访问以及个人信息的泄露、篡改、丢失等。此外，服务提供者需要进行详细的个人信息保护评估，确保其训练行为符合《个人信息保护法》等规定，以及提供的产品和服务具备较高的安全性。

1. 隐私计算、过滤等措施

生成式人工智能训练给个人信息带来的风险很大程度上是因为技术发展不充分，对此，技术领域积极探索了相应的解决方案，因此服务提供者可以采用多种技术手段来降低个人信息风险，具体包括：

（1）隐私计算技术

隐私计算技术（又称“隐私增强技术”）可减轻生成式人工智能研发带来的个人信息和隐私风险，实现保护隐私的机器学习，如多方安全计算、同态加密、差分隐私，[45] 以及分布式、去中心化的机器学习模型训练方案——群体学习（Swarm Learning）。[46]

隐私计算技术可以通过避免数据传输、建立安全计算环境，以及使数据处于加密状态等方式确保个人信息的安全。《人工智能白皮书（2022 年）》指出：“AI 结合隐私计算技术，可从数据源端确保原始数据真实可信。利用隐私计算技术，数据‘可用不可见’，形成物理分散的多元数据的逻辑集中视图，可以保证 AI 模型有充足的、可信的数据可供利用。”近年来，隐私计算等技术已经得到快速发展：多方安全计算、联邦学习、可信执行环境等技术不断迭代优化，单点层面技术能力得到上限提升，技术间的内部融合趋势得到增强，通过优势互补突破应用瓶颈，差分隐私、区块链等技术被应用于辅助隐私计算，实

---

[44] *See* Office of the Privacy Commisioner of Canada, *supra* note 17.

[45] *See* Barnabás SZÉKELY, *Mitigating the Privacy Risks of AI through Privacy-Enhancing Technologies*, Acta Universitatis Sapientiae: Legal Studies, Vol.11:35, p.48-57 (2022).

[46] *See* Stefanie Warnat-Herresthal et al., *Swarm Learning for Decentralized and Confidential Clinical Machine Learning*, Nature, Vol.594:265 (2021).

现了外部的融合，数据保护能力也进一步增强。[47] 目前，隐私计算技术目前已在金融、医疗等行业的生成式人工智能训练中得到应用，极大地降低了个人信息泄露风险。服务提供者应当在模型训练过程中充分结合隐私计算技术，并尽可能采用最有效的技术方案。

（2）过滤及其他措施

服务提供者应采用关键词过滤等技术对侵害个人信息权益的内容进行屏蔽。一方面，要过滤用户的输入，避免用户引导模型生成侵害他人个人信息权益的内容；另一方面，要过滤模型的输出，避免模型在过拟合等情形下意外输出用户的未公开个人信息甚至是敏感个人信息。服务提供者可以在用户协议中对使用规范和相应的后果进行说明。如果用户企图利用模型获取他人的个人信息，服务提供者应当及时提醒、纠正，对达到严重程度者应禁止其使用。[48]《基本要求》指出，应采取关键词、分类模型等方式对使用者输入信息进行检测，使用者连续三次或一天内累计五次输入违法不良信息或明显诱导生成违法不良信息的，应依法依约采取暂停提供服务等处置措施。此外，服务提供者还可通过微调模型来拒绝用户对个人信息的请求，[49] 借助多种途径规范用户的使用行为。

服务提供者还应积极采用其他可以降低个人信息风险的措施，包括使用合成数据、抵御外部攻击等。使用合成数据可在实现预期模型功能的前提下有效降低个人信息风险，例如，合成数据组成的真实医疗记录集不包含任何个人信息，但仍对医学研究有应用价值。[50] 如果使用合成数据可以达到相同或近似的效果，则应使用合成数据替代个人信息数据。实践中，模型可能会受到的攻击包括成员推理攻击、模型逆向攻击、模型提取攻击等，服务提供者应采取有效的抵御措施，提高数据库和模型的安全水平，防止个人信息因受到外部攻击而泄露。以成员推理攻击[51]为例，虽然研究表明，即使采取了联邦学习等隐私计算技术仍可能遭受成员推理攻击，[52] 但目前技术领域已提出借助差分隐私、知识蒸馏等方式来抵御模型推理攻击，[53] 服务提供者有义务采取一种或多种上述措施。此外，《安全规范》还提出了采取身份鉴别、访问控制等技术措施对训练数据进行安全保护。

2. 个人信息保护评估

个人信息保护评估不仅包括事前的影响评估，还包括后续的合规审计。个人信息保护评估对于个人信息权益的保护具有重要意义。原因在于，生成式人工智能训练的技术门槛比过去的个人信息处理场景更高，难以从外部准确地评判其风险，因此需要服务提

---

[47] 参见隐私计算联盟：《隐私计算白皮书（2022 年）》（2022 年 12 月发布），第 5-13 页。

[48] OpenAI 已经采取拒绝回答类似请求、监控用户行为等方式来降低个人信息风险。*See* OpenAI, *supra* note 5.

[49] *See* OpenAI, *supra* note 30.

[50] *See* Adam J. Andreotta, Nin Kirkham & Marco Rizzi, *AI, Big Data, and the Future of Consent*, AI & Society Vol.37:1715, p.1719 (2022).

[51] 机器学习中的成员推理攻击（Membership Inference Attacks）通过推测一个数据样本是否被用来训练目标机器学习模型来获取个人信息或者隐私。参见牛俊等：《机器学习中成员推理攻击和防御研究综述》，载《信息安全学报》2022 年第 6 期，第 4 页。

[52] 参见张佳乐等：《基于 GAN 的联邦学习成员推理攻击与防御方法》，载《通信学报》2023 年第 5 期，第 195-198 页。

[53] 参见前注 [51]，牛俊等文，第 15-18 页。

供者从内部开展评估。评估的重点在于将个人信息用作训练语料的必要性，以及模型在避免个人信息泄露方面的安全性。对于评估结果，应当形成书面的评估报告，便于相关部门进行指导与监督。

（1）必要性评估

必要性评估是指服务提供者应评估其利用某种类型个人信息训练生成式人工智能是否具有必要性，确保将个人信息语料的数量控制在实现模型功能所需的最小范围之内。超过必要范围处理个人信息的行为违反了必要原则和最小化原则的要求，可能会给个人带来不合理的风险。根据《个人信息保护法》第 6 条第 2 款的规定，个人信息收集应当限于实现处理目的的最小范围。虽然生成式人工智能研发需要海量的训练数据，但仍应受到最小化原则的限制，在确保模型功能的前提之下尽可能减少个人信息的处理。不过，应当灵活解释最小化原则，以避免对生成式人工智能训练造成限制。[54] 个人信息语料的数量越多，生成式人工智能泄露个人信息的可能性也就越高。任由服务提供者利用个人信息进行训练而不受必要原则的限制，显然不可取。而且，并非所有的个人信息数据都是实现模型功能所必要（例如，并非所有财务信息和人口特征都有助于预测信用风险），[55] 如果将非必要的个人信息数据用作训练语料，只会徒增个人信息风险而不会产生任何效益。因此，服务提供者需要判断个人信息处理与其欲达到的目的之间是否具有相关性，明确其使用的个人信息数据是否必要、是否存在过度收集的情况。服务提供者应在构建模型或者算法的时候就考虑选择使用对个人信息数据依赖性更小的方式，并在开展训练之前判断某种类型的个人信息数据是否会对模型功能的实现发挥实质作用。如果对模型功能的贡献甚微，应当尽量避免将其作为训练语料。

（2）安全性评估

安全性评估是指服务提供者应评估其模型是否可以较大程度地避免输出侵害个人信息权益的内容，以及是否可以有效抵御外部的攻击。绝对的安全固然无法实现，但是服务提供者仍应在评估风险的基础之上判断是否采取了充分的安全保护措施，并及时作出相应的调整和优化。

安全性评估应当覆盖生成式人工智能从训练到投入使用的全过程。一方面，生成式人工智能的技术特征可能导致其会输出个人信息；另一方面，生成式人工智能可能不可避免地成为网络攻击者的目标，导致安全漏洞的出现。[56]《基本要求》对人工智能生成内容的安全评估作出了规定：服务提供者应当分别采取人工抽检、关键词抽检、分类模型抽检的方式，借助测试题库对生成内容的合格率进行评估。但是，从《基本要求》中生成内

---

[54] 参见斜晓东：《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 64-65 页。

[55] *See* Information Commissioner's Office, *How Should We Assess Security and Data Minimisation in AI?*, at https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/how-should-we-assess-security-and-data-minimisation-in-ai/#whatdataminimisation (Last visited on April 9, 2024).

[56] *See* Elizabeth Montalbano, *Generative AI Projects Pose Major Cybersecurity Risk to Enterprises*, at https://www.darkreading.com/vulnerabilities-threats/generative-ai-projects-cybersecurity-risks-enterprises (Last visited on April 9, 2024).

容测试题库涵盖的三十余种安全风险来看，似乎无法通过生成内容测试题库评估出侵害个人信息内容的输出概率。笔者认为，应当针对生成内容的个人信息合规概率进行专门的评估，并且设置较高的合格标准。若合格率不达标，就应及时采取措施对生成内容进行更深度的过滤。另外，《基本要求》对问题拒答评估也作出了规定，但是并没有要求应拒答测试题库必须涵盖侵害他人个人信息权益的风险。对于诱导输出他人未公开个人信息尤其是敏感信息、私密信息的问题，模型应当拒绝回答，服务提供者应当评估模型是否可以准确识别相关问题并作出正确的应对。此外，服务提供者还须评估模型抵御外部攻击的能力，及时检测和修补安全漏洞，持续关注可能遭受的攻击并且采取有效的预防措施。根据《基本要求》的规定，服务提供者应当将训练环境与推理环境隔离，避免数据泄露和不当访问；持续监测模型的输入内容，防范恶意输入攻击；定期对所使用的开发框架、代码等进行安全审计，关注开源框架安全及漏洞相关问题，识别和修复潜在的安全漏洞。通过对模型安全性的评估，及时发现和应对可能存在的个人信息风险。

此外，服务提供者还需要尽到《个人信息保护法》以及《基本要求》《安全规范》等文件规定的其他个人信息安全保护义务，包括制定并组织实施个人信息安全事件应急预案、实行训练数据的分类分级管理、建立完整的个人信息处理活动记录等。

3. 服务提供者的过错推定责任与行政合规抗辩

严格的监管措施可能会对产业的发展造成制约，但是不可以忽略生成式人工智能的责任问题，否则可能会导致产业的无序发展，甚至使人工智能成为像空壳公司一样的转移责任的工具。[57] 处理未公开个人信息需要基于个人的明确同意，因此社会公众对生成式人工智能的信任显得至关重要。而在生成式人工智能造成损害的情况下，获得赔偿的可能性不仅决定了社会公众的信任和接受程度，还决定了购买或使用生成式人工智能产品和服务的可能性。[58] 所以，明确服务提供者侵害个人信息权益的民事责任，可以提高公众对生成式人工智能技术的信任，进而使其获得更充足的个人信息语料。

作为专门的保护性法律，《个人信息保护法》对个人信息处理者的行为标准和应尽义务作了全面的规定，其中第 69 条规定了个人信息处理者的过错推定责任，但可能会给服务提供者带来较重的负担。根据第 69 条的规定，如果“处理个人信息侵害个人信息权益造成损害”，可以推定处理行为违反了《个人信息保护法》的规定（行为具有违法性），进而可以推定处理者主观上存在过错。质言之，个人信息处理者没有履行保护性法律规定的作为义务，就表明处理者没有达到应有的注意程度，至少存在过失；反之，如果没有违反《个人信息保护法》的规定，处理者便不存在过错。基于该认识，个人信息处理者的过错体现在三个方面：一是没有按照《个人信息保护法》的要求处理个人信息；二是没有依法处理行使个人信息保护请求权的请求；三是没有尽到个人信息安全保护义务。个人信息处理者可以通过证明处理行为符合《个人信息保护法》的规定来证明无

---

[57] *See* Frank Pasquale, *Data-informed Duties in AI Development*, Columbia Law Review, Vol.119:1917, p.1918 (2019).

[58] *See* Teresa Rodríguez de las Heras Ballell, *The Revision of the Product Liability Directive: A Key Piece in the Artificial Intelligence Liability Puzzle*, ERA Forum, Vol.24:247, p.248-249 (2023).

过错。然而,生成式人工智能训练相较于其他个人信息处理过程更加复杂,服务提供者难以进行清晰的“复盘”,以证明处理过程完全符合《个人信息保护法》的要求。例如,必要性原则要求服务提供者在确保实现目标模型功能的前提下尽可能减少个人信息的处理,但是,要求服务提供者回溯到训练之前,证明某种类型的个人信息是在必要范围之内,可能会相当困难。如此一来,过错推定责任可能会事实上发展为无过错责任。即使处理过程完全符合《个人信息保护法》的规定,大模型仍可能会在特殊情况下输出侵害个人信息权益的内容。可见,推定过错的合理性存疑。

笔者认为,生成式人工智能发展初期的个人信息保护机制应当以行政监管为主导,并且重指导轻处罚,以促进生成式人工智能的健康稳定发展。可以允许服务提供者以“符合行政监管要求”作为不存在过错的抗辩,原因主要有以下几点:首先,可以缓解过错推定责任给服务提供者带来的证明负担,避免因过重的责任阻碍技术创新;其次,行政监管可以为服务提供者提供一个可预期的合法合规标准,利于个人信息保护合规工作的开展;最后,服务提供者无需在多个诉讼中重复证明其处理行为符合《个人信息保护法》的规定,利于加快纠纷解决,节省诉讼资源。总之,行政合规抗辩应是考虑到现有技术水平所作出的合理制度安排。[59] 目前生成式人工智能的发展程度有限,相较于以往的个人信息处理法律关系,需要重新平衡个人与服务提供者之间的利益。相关部门应详细指导并监督服务提供者开展个人信息合法合规工作。相关合法合规工作主要包括三个方面:一是训练语料的处理符合《个人信息保护法》等规定,包括处理的范围符合最小化原则的要求等;二是依法满足个人的查阅、复制、删除、解释说明等请求;三是采取充分的个人信息安全保护义务,包括进行关键词过滤、采取有效措施应对可能存在的风险等。

值得注意的是,欧盟委员会于 2022 年通过了修订《产品责任指令》的提案,拟通过产品责任来解决人工智能的责任问题。但笔者认为,生成式人工智能输出侵害个人信息的内容是技术发展不充分的结果,除非是安全性明显低于一般技术水平的情况,否则不能一概将其视为“缺陷”所致并对其适用产品责任。而且,生成式人工智能主要提供一般性的信息服务,不会像自动驾驶汽车等因为固有缺陷而威胁他人的生命、健康、财产等权益,因此不应适用产品责任。[60] 即使是从保护受害人的角度考虑,产品责任也不一定是更有利的选择。表面上看,适用产品责任无需考虑服务提供者的过错,似乎更容易成立侵权。但适用产品责任也增加了受害人的举证负担——受害人需要证明生成式人工智能产品存在缺陷,而人工智能的高技术壁垒会使其面临举证上的困难。[61] 作为配套制度,就只能在缺陷的证明责任上进行缓和,[62] 但其效果与适用过错推定责任相差不大。

综上,应根据《个人信息保护法》第 69 条而非产品责任的有关规定追究服务提供

---

[59] 参见王若冰:《论生成式人工智能侵权中服务提供者过错的认定——以“现有技术水平”为标准》,载《比较法研究》2023 年第 5 期,第 20-25 页。

[60] 参见王利明:《生成式人工智能侵权的法律应对》,载《中国应用法学》2023 年第 5 期,第 33 页。

[61] 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 122-124 页。

[62] 参见杨立新:《人工智能产品责任的功能及规则调整》,载《数字法治》2023 年第 4 期,第 38 页。

者的个人信息侵权责任，并且允许以“符合行政监管要求”作为不存在过错的抗辩。这样不仅可以在一定程度上避免服务提供者因难以证明无过错而承担过重的责任，阻碍产业的健康发展，而且可有效缓解受害人面临的举证困难，给予受害人较为充分的保护。

四、结　语

生成式人工智能技术迎来了爆发式发展，已经对国家、社会和个人产生了广泛的影响，然而技术的复杂性导致其治理面临诸多挑战，其中之一便是训练语料的个人信息保护问题。本文从基本立场和指导思想出发，分析了如何在生成式人工智能场景下适用和调整《个人信息保护法》的有关规定，以兼顾个人信息的利用和保护。一方面，应当以平衡原则为指引宽松解释《个人信息保护法》，并在必要时作出例外规定，缓解服务提供者在获取个人信息语料时面临的困难；另一方面，应当确保个人信息保护请求权的行使、要求服务提供者尽到严格的个人信息安全保护义务，充分保障个人信息权益。总之，我国在制定人工智能法的时候，应当结合生成式人工智能的技术特征和产业需求作出相应的规定，构建符合生成式人工智能客观发展规律的个人信息保护制度。

Abstract: The personal information protection of generative AI training language material should uphold the basic stance of encouraging and supporting innovation. To ensure that the personal information utilization needs of service providers can be met, the *Personal Information Protection Law* can be interpreted or made exceptions at the training end. For personal information that has already been disclosed, it can be included in the scope of processing by loosely interpreting the ‘purpose of disclosure’. For undisclosed personal information, it is still necessary to use the individual’s consent as the legitimacy source of the processing. However, the difficulties faced by service providers can be alleviated by broadly interpreting the principle of purpose limitation and adjusting the relevant rules of ‘notification-consent’. Increased technical barriers has exacerbated the disadvantageous position of information subjects, and it is necessary to ensure the exercise of the right to request personal information protection to protect the legitimate rights and interests of individuals, but its exercise is inevitably limited by technical realities. Service providers should strictly perform their obligations to protect personal information security, including technical measures, to minimize risks to personal information. The protection mechanism as a whole should be dominated by administrative supervision, and if the infringement of personal information rights and interests causes harm, the service provider should be allowed to use ‘compliance with administrative supervision requirements’ as a defense of non-fault.

（责任编辑：任　彦）