简评《生成式人工智能服务管理办法(征求意见稿)》七大问题

作者：潘永建 | 朱晓阳 | 沙莎

2018 年，人工智能艺术家 Robbie Barrat 的创作作品被拍卖出 432,500 美元，其后续作品《Saint Nazaire》也被誉为"AI 艺术的划时代作品"[1]。近年来，以 ChatGPT 为代表的生成式人工智能(Generative AI)更具创造性，能够实现互动对话、精准回复、图表分析和数据推理，甚至图像创作等功能。

在生成式人工智能时代，不仅用户无法控制人工智能创作的内容，甚至人工智能开发者也难以控制。因此，生成式人工智能的强大功能不可避免地带来了更多的法律风险，也引起了全球监管浪潮。继 2023 年 3 月 31 日意大利个人数据保护局禁用 ChatGPT[2]以及多国宣布考虑对 ChatGPT 发布禁令或进行严格监管[3]后，中国互联网信息办公室于 2023 年 4 月 11 日发布《生成式人工智能服务管理办法(征求意见稿)》("《管理办法》")，拟对生成式人工智能产品和服务内容、技术和准入要求等进行监管。本文将结合《管理办法》的主要内容，总结七大问题并进行解读。

1. 适用范围

生成式人工智能以现有的数据集为基础，通常利用深度学习模型，包括循环神经网络(RNN)、变分自编码器(VAE)、生成对抗网络(GAN)等，最终产出形式多样的新内容，包括文章、对话、音乐、语音、绘画、动画等。《管理办法》开篇即对生成式人工智能进行定义，即生成式人工智能指"基于算法、模型、规则生成文本、图片、声音、视频、代码等内容的技术"，为生成式人工智能的监管划定了范围。

如您需要了解我们的出版物，
请联系：

Publication@llinkslaw.com

[1] 《收藏者曹寅: Robbie Barrat，了不起的 AI 艺术》，参见 https://www.ccvalue.cn/article/382855.html。
[2] Italian privacy regulator bans ChatGPT，参见 https://www.politico.eu/article/italian-privacy-regulator-bans-chatgpt/。据 Politico 报道，ChatGPT 履行相应合规要求后，可能被意大利数据保护机构解除禁用，参见 https://www.politico.eu/article/chatgpt-italy-lift-ban-garante-privacy-gdpr-openai/。
[3] EU: ChatGPT spurs debate about AI regulation，参见 https://www.dw.com/en/eu-chatgpt-spurs-debate-about-ai-regulation/a-65330099。

地域效力方面，《管理办法》的适用范围为“研发、利用生成式人工智能产品面向中国境内公众提供服务”的行为。换言之，《管理办法》监管在中国境内提供服务的行为，无论该行为发生在中国境内或境外。以 ChatGPT 为例，虽然 ChatGPT 的运营方及服务器等位于中国境外，但如果 ChatGPT 的运营方主动将其向中国用户开放，则可能导致 ChatGPT 及其运营方落入《管理办法》的管辖范围。这也可能是 OpenAI 没有向中国地区用户开放使用 ChatGPT 的一大原因(关于 ChatGPT 产品风险分析与合规建议，参见《ChatGPT 花式整活进行时：边界何在?》《狂飙的 ChatGPT，合规“缰绳”何在?》等文)。此外，即使境外主体本身不直接向中国境内提供生成式人工智能产品，如果境外主体以提供可编程接口等方式支持其他主体(无论位于中国境内或境外)生成内容，而其他主体向中国境内公众提供服务，则该境外主体仍将受《管理办法》的监管。

2. 内容生产者责任

根据《管理办法》的规定，利用生成式人工智能产品提供聊天和文本、图像、声音生成等服务，以及通过提供可编程接口等方式支持他人自行生成文本、图像、声音的组织和个人是服务提供者，需承担内容生产者的责任。

在生成式人工智能研发过程中，提供者需要进行数据收集和投喂，并进行模型设计、训练和评估。在此过程中，提供者能够通过数据筛选、模型构建以及算法训练对生成内容进行把控。正是因为提供者对于生成的内容有一定的控制能力，《管理办法》要求提供者承担内容生产者责任。然而，提供者对生成内容进行管理远非易事。《管理办法》要求提供者“采取措施防止生成虚假信息”，由于生成内容的多样性，提供者或只能对客观事实的真实性进行管理，而无法对生成的“创造性”“艺术性”作品的真实性进行管理。并且，提供者需要为此付出极大的技术和管理成本。因此，“内容生产者责任”的内涵和外延还有待立法或执法部门的进一步解释，究竟是提供者证明其“采取了充分的措施防止生产虚假信息”即可免责，还是只要发生产生了虚假信息的后果，提供者就需要承担责任?

此外，在使用者直接利用侵犯他人权益的或虚假的数据使用生成式人工智能时，生成式人工智能的输出结果同样可能侵犯合法权益或构成不正当竞争，这也对提供者研发和利用生成式人工智能提出更高的技术要求。此外，尽管《管理办法》并未对生成式人工智能使用者的法律责任进行直接规定，但使用者仍需要履行《网络安全法》《数据安全法》《个人信息保护法》以及其他适用法律法规规定的特定的数据处理场景下的法律义务(如《互联网信息服务深度合成管理规定》规定的使用者义务)，我国未来也可能出台专门针对生成式人工智能使用者的相关管理规定，对使用者提出更加细化的合规要求。

3. 个人信息保护责任

《管理办法》在两个阶段对提供者提出履行个人信息保护义务的要求。在生成式人工智能产品的训练阶段，如果训练用的数据类型中包括个人信息，提供者(或个人信息提供方)应当获取个人信息主体的同意或符合其他法定情形；在生成式人工智能产品的使用阶段，对于用户输入的个人信息，提供

者不能非法留存、对外提供，也不能利用输入信息对用户进行画像。提供者还应当响应个人信息主体的请求，对个人信息主体提出的更正、删除、屏蔽个人信息的请求予以处理。就提供者而言，我们理解，对于人工智能产品的训练并不需要识别具体的个人信息主体，因此为了避免非法获取和留存个人信息，建议提供者对个人信息进行匿名化或者至少是深度去标识化后再使用。

此外，在生成式人工智能的使用阶段，使用者应当确保其使用个人信息的行为具备合法性基础。

4. 前置性要求

根据《管理办法》，提供者在利用生成式人工智能产品提供服务前，还应当履行以下两项前置性要求。

(1) 互联网信息服务安全评估

提供者应当按照《具有舆论属性或社会动员能力的互联网信息服务安全评估规定》(“《安全评估规定》”)向国家网信部门申报安全评估。

根据《安全评估规定》，通常在以下两类情形中，具有舆论属性或社会动员能力的互联网信息服务触发安全评估: (a)互联网信息服务本身发生的重大变化，如信息服务上线和增设新功能; (b)客观情况发生重大变化，包括用户规模的显著增加和违法有害信息发生传播扩散难以有效防控。提供者应当及时预估生成式人工智能产品的社会动员能力，以确认是否需要履行相应的安全评估义务。

(2) 算法备案

提供者应当按照《互联网信息服务算法推荐管理规定》(“《算法推荐规定》”)履行算法备案和变更、注销备案手续。

算法推荐技术，是指利用生成合成类、个性化推送类、排序精选类、检索过滤类、调度决策类等算法技术向用户提供信息。《算法推荐规定》在算法模型审核、防沉迷机制、用户注册、信息发布审核、数据安全与个人信息保护方面的要求与《管理办法》有共通之处。企业在进行算法推荐技术备案时，应注意备案内容包括服务提供者的名称、服务形式、应用领域、算法类型、算法自评估报告、拟公示内容等信息。

5. 实名制要求

《管理办法》要求生成式人工智能的使用者提供真实身份信息。与前述《安全评估规定》相衔接，用户真实身份核验也是具有舆论属性或社会动员能力的互联网信息服务的重点评估内容。此前，《互联网信息服务深度合成管理规定》在用户实名制方面的审核方式包括通过移动电话号码、身份证件号

码、统一社会信用代码或者国家网络身份认证公共服务等识别用户身份，《管理办法》很可能采取类似的手段进行用户核验。

6. 用户管理义务

在提供生成式人工智能服务过程中，提供者应当履行用户管理义务，主要包括:

(1) 监管用户发布内容

如果用户利用生成式人工智能违反法律法规，违背商业道德、社会公德，包括从事网络炒作、恶意发帖跟评、制造垃圾邮件、编写恶意软件，实施不正当的商业营销等，提供者应当暂停或者终止对该用户的服务。

提供者应当注意，除《管理办法》的规定外，《网络安全法》针对网络平台运营者、《未成年人保护法》针对网络服务提供者均提出用户发布信息管理要求。如果用户存在发布违法内容的行为，除暂停或终止服务外，提供者还可能需要履行记录、报告义务等。具体而言:

- 《网络安全法》要求，网络平台运营者发现用户发布法律、行政法规禁止发布或者传输的信息的，应当立即停止传输该信息，采取消除等处置措施，防止信息扩散，保存有关记录，并向有关主管部门报告。
- 《未成年人保护法》要求，网络服务提供者发现用户发布、传播含有危害未成年人身心健康内容的信息的，应当立即停止传输相关信息，采取删除、屏蔽、断开链接等处置措施，保存有关记录，并向网信、公安等部门报告；发现用户利用其网络服务对未成年人实施违法犯罪行为的，应当立即停止向该用户提供网络服务，保存有关记录，并向公安机关报告。

(2) 指导用户合法利用生成内容

《管理办法》要求提供者对用户使用生成式人工智能生成的内容进行指导，避免用户利用生成内容损害他人形象、名誉以及其他合法权益，或进行商业炒作和不正当营销。实践中，提供者可能需要同时通过技术和管理手段履行相应义务。在技术手段方面，可能需要对违法内容进行屏蔽、不予展示及进行合法性提示；在管理手段方面，提供者可通过用户服务协议等政策向用户传达利用生成内容的合法性要求。

(3) 防沉迷措施

提供者应当明确并公开其服务的适用人群、场合、用途，采取适当措施防范用户过分依赖或沉迷生成内容。提供者可能需要结合前述实名制要求和手段，对用户进行分类，实现生成内容用途和场合方面的管理。

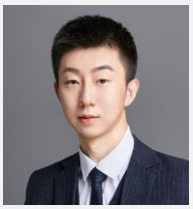
7. 人工标注要求

在研发生成式人工智能产品时，利用模型生成的内容可能不准确或不合理，需要通过后续的人工编辑和修正提高生成内容的准确性和可靠性。人工标注能够起到纠错和校对作用，使模型生成更加准确和合理的内容。

由于数据量庞大以及类似“流水线”的工作流程，提供者可能委托第三方开展人工标注工作。根据《管理办法》的规定，人工智能产品研制中采用人工标注时，提供者应当制定清晰、具体、可操作的标注规则，对标注人员进行必要培训，抽样核验标注内容的正确性。实践中，如果提供者采用第三方人工标注服务，应当通过协议对上述内容进行约定，并实际对人工标注内容进行审计抽检。此外，提供者可能还需要履行数据分类分级、个人信息保护影响评估等法律义务。

如您希望就相关问题进一步交流，请联系：

潘永建
+86 21 3135 8701
david.pan@llinkslaw.com



朱晓阳
+86 21 3135 8683
nigel.zhu@llinkslaw.com

如您希望就其他问题进一步交流或有其他业务咨询需求，请随时与我们联系：**master@llinkslaw.com**

上海	北京	深圳
上海市银城中路 68 号 时代金融中心 19 楼 T: +86 21 3135 8666 F: +86 21 3135 8600	北京市朝阳区光华东里 8 号 中海广场中楼 30 层 T: +86 10 5081 3888 F: +86 10 5081 3866	深圳市南山区科苑南路 2666 号 中国华润大厦 18 楼 T: +86 755 3391 7666 F: +86 755 3391 7668
香港	**伦敦**	
香港中环遮打道 18 号 历山大厦 32 楼 3201 室 T: +852 2592 1978 F: +852 2868 0883	1/F, 3 More London Riverside London SE1 2RE T: +44 (0)20 3283 4337 D: +44 (0)20 3283 4323	

www.llinkslaw.com

Wechat: LlinksLaw

本土化资源 国际化视野