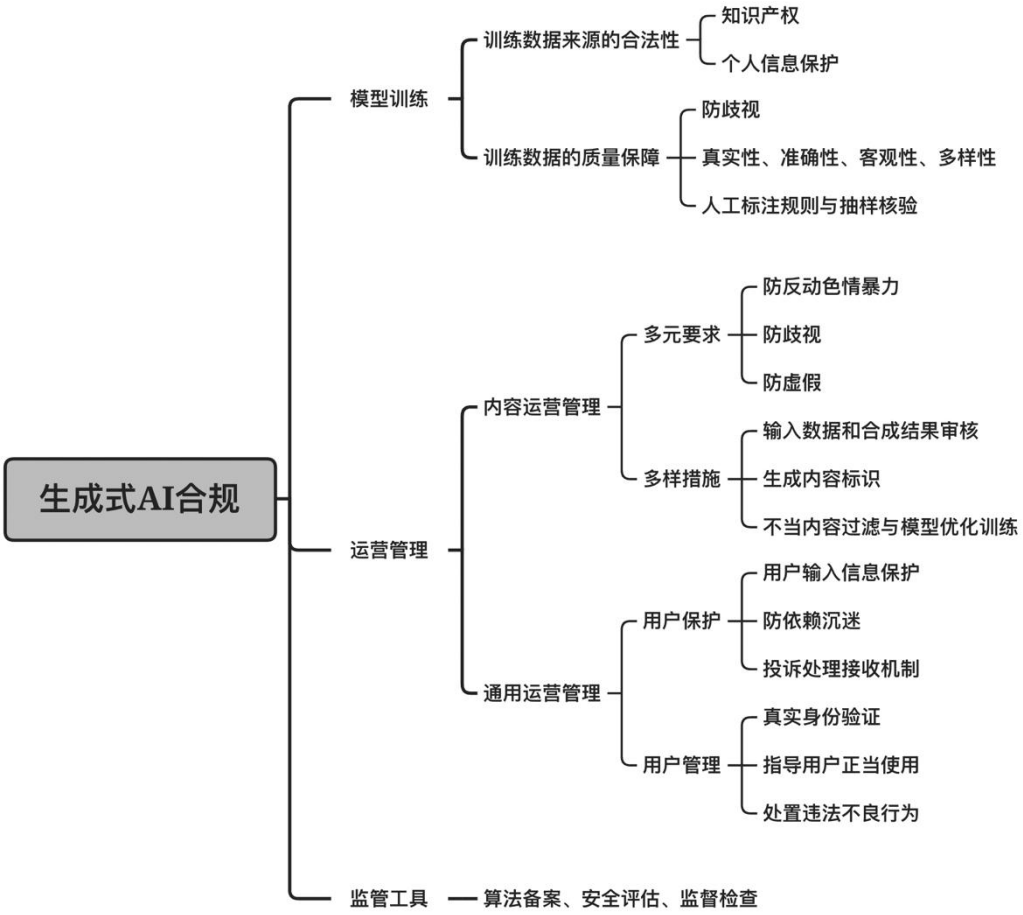
生成式 AI（二）：体系化构建合规指南

作者：杨建媛 李天烁

内容治理是生成式 AI 监管的关注重点。一方面，基于客观维度，生成式 AI 可能会生成不真实或无意义的内容，即具有“幻觉”倾向；另一方面，结合价值判断，生成式 AI 还存在生成有害内容的问题。如本系列首篇《生成式 AI（一）：尝试参透监管的底层逻辑》所分析，**生成式 AI 监管的底层逻辑在于引导对人类的良性影响。**为了实现这一目标，监管规则对企业提出了一系列合规要求，旨在弥补生成式 AI 的固有缺陷，降低其应用风险。本文拟对我国生成式 AI 的监管规则进行体系化梳理，以期为企业的合规实践提供指引。



一、 模型训练：确保来源合法、提升数据质量

（一） 保障训练数据来源的合法性

生成式 AI 的训练通常需要 TB 级的海量数据。数据来源的合法性为生成式 AI 合规的基础性要求，其包括但不限于知识产权、个人信息保护。

1. 知识产权

《生成式人工智能服务管理办法（征求意见稿）》（“**《生成式人工智能服务办法》**”）第七条中明确要求，用于生成式 AI 产品的预训练、优化训练数据（“**训练数据**”），应不含有侵犯知识产权的内容。

我国《著作权法》过去明确列举了十二种“合理使用”的情形，并在 2020 年修改时新增了“法律、行政法规规定的其他情形”这一兜底条款，但**企业为训练生成式 AI 而使用他人作品通常无法符合“合理使用”的任一情形，如无授权则存在知识产权侵权风险**。我国目前并未通过“柔性合理使用条款”（如日本法）等方式为生成式 AI 等新技术设置特殊的合理使用情形，亦未通过“四要素分析法”（如美国法）等方式仅规定合理使用的判断因素而不对其适用情形作列举限定。尽管我国司法实践中确偶有突破《著作权法》额外创设“合理使用”情形的特例，但在现有规则体系下，**如何取得作品使用授权/避免使用作品进行训练，是相关大模型企业在目前无法回避的一个问题**。

生成式 AI 训练数据的知识产权问题已在境内外引起了诸多争议，例如：美国 AI 绘画软件 Stable Diffusion 的开发商 Stability AI 即因未经授权爬取 1200 余万张图像用于大模型训练而遭到起诉，该案正在进展中，合理使用问题系该案争议焦点之一[1]；在可能成为中国“AI 大模型数据被盗第一案”的笔神作文与学而思的纠纷中，亦涉及到著作权侵权的相关争论。[2]

鉴于训练数据的庞大规模，确保其中不含有任何侵犯知识产权的内容，对于相关企业来说存在相当高的实现难度。我国现有监管规则尚未对此提供更加细化的合规指引，但欧盟一周前刚刚通过的《人工智能法案》草案或可提供借鉴思路——其要求生成式 AI 的基础模型提供者应针对其训练模型所使用的任何受著作权保护的材料，记录并公开披露详细的使用情况摘要。这一透明度方案或可有效降低著作权人的维权难度，但也相应对开发者提出了更高的合规要求。

2. 个人信息保护

《生成式人工智能服务办法》第七条中明确要求，训练数据包含个人信息的，应当征得个人信息主体同意或者符合法律、行政法规规定的其他情形。

OpenAI 表示：“我们希望（模型）了解世界，而非了解个人”，并承诺将在可行的范围内尽量删除训练数据集中所包含的互联网上公开可获得的个人信息。[3]与之相呼应，有观点主张模型训练不适用个人信息保护相关法规，或至少应从个人信息保护角度对模型训练予以豁免。然而，仅从现行法的角度，**在训练数据包含个人信息的情况下，尽管其或并不旨在对自然人进行识别，但将数据用于模型训练的行为通常仍被认为属于《个人信息保护法》所规定的“处理”，因此需征得个人同意或具备其他合法性基础。**

针对训练数据的个人信息保护要求并非中国所独有，2023 年 3 月，意大利数据监管机关 Garante 对 ChatGPT 发布了临时禁令，原因之一即在于 OpenAI 的个人信息收集及以算法训练为目的的处理活动缺乏合法性基础。此后，OpenAI 通过在网站公布训练算法的个人信息

---

[1] https://arstechnica.com/tech-policy/2023/04/stable-diffusion-copyright-lawsuits-could-be-a-legal-earthquake-for-ai/

[2] https://mp.weixin.qq.com/s/aRYJbh1UO9RYEJdhL-nivQ

[3] https://openai.com/blog/our-approach-to-ai-safety

处理情况、明确正当利益（legitimate interest）为利用用户个人信息进行算法训练的合法性基础、允许欧盟个人以便捷方式选择退出（opt-out）算法训练等措施进行了整改，恢复了 ChatGPT 在意大利境内的运营。

3. 关于数据来源合法性的争议

数据来源的合法性要求尽管看起来“天经地义”，但之于生成式 AI 而言，过于绝对的合法性要求从理论上亦可能涉及以下问题：[4]

首先，数据质量和数据合法性之间存在矛盾。生成式 AI 的训练需要海量数据，如果对预训练数据的合法性作出要求，企业为了规避风险可能会采取过于谨慎的态度，大幅缩减训练数据数量，甚至可能损害数据的客观性和多样性。

其次，训练数据的合法性并非信息生成合法性的必要条件。最初输入的训练数据与最终输出的生成内容并非直接对应关系，其中经历了较为复杂的转换过程。

最后，《生成式人工智能服务办法》的规定在责任承担方面存在问题。在该规定项下，数据来源合法性要求的义务主体为利用生成式人工智能产品提供聊天和文本、图像、声音生成等服务的组织和个人（“**提供者**”），然而，纯粹的服务商可能并不参与模型的训练开发，也并不具备相应的技术能力，要求其就该项义务承担责任过于严苛。

针对以上问题，界定大模型训练作为合理使用的情形、标准化解决内容创作者的补偿问题、区分主体施以合规义务均为可以探讨的解决方案。

（二） 训练数据应具备高质量

相较于“不证自明”的合法性要求，**针对训练数据的质量要求，通常被认为是生成式 AI 监管的特色规定。**监管关注训练数据质量的逻辑或在于，当人类已难以完全理解具有数百亿参数大模型的推理过程时，要求企业采取措施增强训练数据的真实性、准确性、客观性、多样性（例如，设计预训练数据集时考虑偏远地区、少数民族等因素，利用分类器及关键词库对数据集进行过滤等），就成为了为数不多有效可行的监管手段。其原理近似于，当家长无法控制孩子的行为时，至少应确保所教导传授的内容是积极向善的。

针对训练数据的质量问题，《生成式人工智能服务办法》主要存在两方面的重点考量：**一方面，**提供者应在算法设计、训练数据选择、模型生成和优化、提供服务等过程中，**采取措施防止歧视。另一方面，**提供者应**保证训练数据的真实性、准确性、客观性、多样性。**

此外，根据《生成式人工智能服务办法》的要求，为了加强数据质量管理，**提供者如果采用了人工标注的方式训练生成式 AI，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。**实践中建议企业留存相应的培训和抽样核验记录，以证明履行了合规义务。

二、 运营管理：内容治理为核心，以人为本是基调

[4] https://mp.weixin.qq.com/s/DXgyb-8I2YLoXWN8jOQzAg

（一） 内容治理

1. 内容治理的多元要求

《生成式人工智能服务办法》对生成内容的监管要求主要体现在三个方面：**首先，生成内容应当体现社会主义核心价值观，不得含有反动、色情、暴力等内容。**该要求与《互联网信息服务管理办法》《网络信息内容生态治理规定》《互联网信息服务算法推荐管理规定》(“《算法推荐管理规定》”)《互联网信息服务深度合成管理规定》(“《深度合成管理规定》”)一脉相承。**其次，生成内容不得带有歧视性。最后，生成内容应当真实准确，提供者应采取措施提高生成内容的准确性和可靠性、防止生成虚假信息。**

2. 治理措施的多样手段

首先，“幻觉”目前通常被认为是大模型技术的固有缺陷，目前无法实现100%的准确可靠。如何减少幻觉是大模型研究应用的重点之一，但恐怕难以一蹴而就。例如，根据OpenAI发布的GPT-4技术报告，GPT-4在科技、历史、商业等各类主题测试集中，其准确率普遍介于60-80%之间，而这已是GPT-4相较于GPT-3.5幻觉程度显著降低后的结果。

除了前文所提及的提升训练数据的质量，**建立适当的内容审核机制也是弥补该缺陷的方式之一。**此处与《深度合成管理规定》和《算法推荐管理规定》的监管要求相衔接：**一方面，**服务提供者应当加强内容管理，**采取技术或者人工方式对服务使用者的输入数据和合成结果进行审核；另一方面，**服务提供者应当**建立健全用于识别违法和不良信息的特征库，**完善入库标准、规则和程序，记录并留存相关网络日志。

其次，提供者应依法对生成的图片、视频等内容进行标识。标识分为隐式与显式两类，两者并行不悖、不存在替代关系：

- 隐式标识：提供者应当采取技术措施添加不影响用户使用的标识，并保存相关日志信息以便进行识别追溯；
- 显式标识：对于可能导致公众混淆或误认的服务，应由提供者在生成内容的合理位置、区域进行显著标识；对于前述以外的其他服务，应由提供者提供显式标识功能，并提示使用者可以进行显式标识。

标识义务为生成式AI合规的难点问题之一，“可能导致公众混淆或误认”的判断标准、不同模态下的具体标识方案等问题均有待进一步明晰。业界正在内容标识领域不断探索，例如，抖音于2023年5月发布《抖音关于人工智能生成内容标识的水印与元数据规范》，一方面确定了统一的水印样式和位置，在提示用户的同时尽可能减少观感不适；另一方面规范了人工智能生成内容的元数据格式，在相关图片和视频元数据中写入信息，以达到行业通用识别的效果。[5]

最后，对于模型生成的不当内容，除采取内容过滤等措施外，提供者还应通过模型优化训练等措施进行整改、防止再次生成。

[5] https://www.douyin.com/rule/billboard?id=1242800000050

（二） 通用运营管理

1. 保护用户不因使用生成式 AI 受害

首先，提供者应对用户输入信息和使用记录承担保护义务，不得将其非法留存、用于用户画像或向他人提供，除非法律法规另有规定。

其次，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。然而《生成式人工智能服务办法》中的防依赖沉迷条款[6]，究竟是适用于全部用户的普适要求，还是旨在保护儿童、老人等弱势群体的特殊要求，目前尚不明确。但从体系解释的角度，结合《生成式人工智能服务办法》的防依赖沉迷条款与《算法推荐管理规定》的未成年人保护条款[7]，我们理解，建立未成年人保护机制是前者的应然之义，但是否还需建立其他机制则需持续观察监管倾向及市场实践的发展变化。

在境外，生成式 AI 服务的年龄过滤机制已受到了监管者的格外关注。2023 年 2 月，意大利数据监管机关 Garante 就人工智能聊天工具 Replika 发布了临时禁令，要求其停止处理意大利人的数据，主要原因之一即在于 Replika 所提供的虚拟情感关系服务可能会对未成年人造成伤害，但其并未设置年龄验证机制，致使未成年人可以轻松访问并使用该服务。[8]以此为鉴，尽管年龄门槛不尽相同，但 OpenAI、Google 和 Microsoft 均为用户注册和使用其大模型服务设定了一定的年龄限制。

最后，提供者应当建立用户投诉接收处理机制，及时处置个人关于更正、删除、屏蔽其个人信息的请求；发现、知悉违法和不良信息时，应当采取措施、停止生成、保存记录并向监管部门报告，防止危害持续。

2. 管理用户不能使用生成式 AI 作恶

首先，提供者应当对用户的真实身份进行验证。该项规定有助于后续快速识别不良用户，提高平台管理能力。**其次，提供者应当指导用户正当使用人工智能生成内容。最后，提供者发现用户使用生成式 AI 产品过程中存在违反法律法规，违背商业道德、社会公德行为时，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，应当暂停或者终止服务。**

当前，国内已出现多起利用生成式 AI 发布虚假新闻的违法案例。例如，2023 年 4 月，甘肃公安发现有不法分子散播题为“今晨甘肃一火车撞上修路工人致 9 人死亡”的虚假文章，而后查明该文为洪某利用 ChatGPT 所编辑的内容。洪某散布虚假信息的行为已涉嫌寻衅滋事罪，目前已被警方采取刑事强制措施，该案仍在进展之中。[9]

---

[6] 《生成式人工智能服务办法》第十条：“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。”

[7] 《算法推荐管理规定》第十八条：“算法推荐服务提供者向未成年人提供服务的，应当依法履行未成年人网络保护义务，并通过开发适合未成年人使用的模式、提供适合未成年人特点的服务等方式，便利未成年人获取有益身心健康的信息。
算法推荐服务提供者不得向未成年人推送可能引发未成年人模仿不安全行为和违反社会公德行为、诱导未成年人不良嗜好等可能影响未成年人身心健康的信息，不得利用算法推荐服务诱导未成年人沉迷网络。”

[8] https://www.silicon.co.uk/e-innovation/artificial-intelligence/replika-italy-ban-497135

[9] https://mp.weixin.qq.com/s/_Wfe-EV13O6uBM65jZDzdg

三、 监管工具：算法备案与安全评估并驾齐驱

具有舆论属性或社会动员能力的互联网信息服务和相关新技术新应用是监管部门的重点关注对象。《生成式人工智能服务办法》规定，**利用生成式 AI 产品向公众提供服务前**，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》向国家网信部门**申报安全评估**（即**“安全评估”、“双新评估”**），并按照《算法推荐管理规定》**履行算法备案和变更、注销备案手续**（即**“算法备案”**）。

需要注意的是，上文所提及的大多为《生成式人工智能服务办法》这一特殊法对于生成式 AI 的特殊合规要求，但除此之外提供者还应当同时落实《深度合成管理规定》、《算法推荐管理规定》作为一般法所提出的通用合规要求，包括但不限于落实信息安全主体责任、建立健全用户注册、算法机制机理审核、应急处置等一般性管理制度。该等通用合规要求的落实对于企业顺利完成算法备案、安全评估有着显著影响，切不可忽视。

除此之外，**提供者负有配合监督检查的义务。**特别地，《生成式人工智能服务办法》对算法透明度作出要求，提供者应当根据监管部门的要求，提供可以影响用户信任、选择的必要信息，包括预训练和优化训练数据的来源、规模、类型、质量等描述，人工标注规则，人工标注数据的规模和类型，基础算法和技术体系等。实践中建议企业留存前述相关内容的产品文档、评估报告、日志记录等，做好响应配合监督检查的准备。