投资生成式人工智能服务类项目的合规关注要点

作者：唐江山 王若曦

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。[1]随着 OpenAI 接连发布对话式大型语言模型 ChatGPT 和多模态预训练大模型 GPT-4，微软基于 GPT-4 推出 Microsoft 365 Copilot，大量生成式人工智能应用涌现并收获大量用户。生成式人工智能技术已成为近年来最具突破性的技术之一。

面对生成式人工智能的突破性发展，2023 年 4 月 11 日，国家互联网信息办公室（“**网信办**”）发布《生成式人工智能服务管理办法（征求意见稿）》（“**《征求意见稿》**”），就生成式人工智能服务提出了针对性的监管要求。值得注意的是，生成式人工智能，尤其是大语言模型（LLM），对各国监管而言均属新生事物，尽管域外已有欧盟《人工智能法案（AI Act）》[2]、美国《人工智能可问责性政策（AI Accountability Policy）》[3]等初步的立法尝试，但整体而言尚未形成成熟的、可资借鉴的监管规则。网信办此次发布的《征求意见稿》是对生成式人工智能产品的“全生命周期”进行专门规制的领先尝试。

本文拟结合最新监管动态和相关项目经验，探讨投资生成式人工智能服务类项目需要特别关注的合规事项。

一、 监管概览

在《征求意见稿》出台前，“生成合成类”算法推荐技术[4]已被纳入《互联网信息服务算法推荐管理规定》的监管范围；而 2022 年 12 月颁布的《互联网信息服务深度合成管理规定》所定义的“深度合成技术”[5]也包括了生成合成类算法的诸多应用场景（如文本、图像、音频、视频等生成应用和 VR 技术）。

生成式人工智能等新兴技术的突破式发展也在伦理领域带来了巨大挑战，在传统主要适用于医疗健康领域的科技伦理审查基础之上，兼及人工智能领域的科技伦理审查制度正在构建之中。2023 年 4 月 4 日，科学技术部发布《科技伦理审查办法（试行）（征求意见稿）》，与人工智能相关的科学研究、技术开发等科技活动也可能被纳入适用范围[6]，需要提交科技伦理审查。

---

[1] 引用自《生成式人工智能服务管理办法（征求意见稿）》的定义。

[2] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

[3] https://ntia.gov/issues/artificial-intelligence/request-for-comments

[4] 《互联网信息服务算法推荐管理规定》规定，应用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。

[5] 《互联网信息服务深度合成管理规定》规定，在中华人民共和国境内应用深度合成技术提供互联网信息服务，适用本规定。“深度合成技术”是指利用深度学习、虚拟现实等生成合成类算法制作文本、图像、音频、视频、虚拟场景等网络信息的技术，包括但不限于：（一）篇章生成、文本风格转换、问答对话等生成或者编辑文本内容的技术；（二）文本转语音、语音转换、语音属性编辑等生成或者编辑语音内容的技术；（三）音乐生成、场景声编辑等生成或者编辑非语音内容的技术；（四）人脸生成、人脸替换、人物属性编辑、人脸操控、姿态操控等生成或者编辑图像、视频内容中生物特征的技术；（五）图像生成、图像增强、图像修复等生成或者编辑图像、视频内容中非生物特征的技术；（六）三维重建、数字仿真等生成或者编辑数字人物、虚拟场景的技术。

[6] 根据《科技伦理审查办法（试行）（征求意见稿）》，开展以下科技活动也需要进行科技伦理审查：（1）涉及个人信息的科技活动，以及（2）可能在生命健康、生态环境、公共秩序、可持续发展等方面带来伦理风险挑战的科技活动。

本次《征求意见稿》进一步明确"研发、利用生成式人工智能产品，面向中华人民共和国**境内公众**提供服务的，适用该办法"，并规定了兜底性罚则，即如果服务提供者违反该规定，且其他法律、行政法规没有规定的，可能面临警告、通报批评、责令限期改正，暂停或终止利用生成式人工智能提供服务，罚款（1-10 万元）等处罚。

除以上针对性规定外，从上位法角度，生成式人工智能服务提供者仍需要遵守《网络安全法》《数据安全法》《个人信息保护法》等法律法规。

二、 投资项目关注要点

从近期生成式人工智能类投资项目的实操经验出发，我们梳理了如下关注要点：

（一） 资质/审查要求

《征求意见稿》规定，利用生成式人工智能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，并按照《互联网信息服务算法推荐管理规定》履行算法备案和变更、注销备案手续。

（1） 安全评估

《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》通过"列举+兜底"的方式，将"具有舆论属性或社会动员能力的互联网信息服务"定义为"论坛、博客、微博客、聊天室、通讯群组、公众账号、短视频、网络直播、信息分享、小程序信息服务及提供公众舆论表达渠道或者具有发动社会公众从事特定活动能力的信息服务"。

前述规定对需要进行安全评估的互联网信息服务的表述相对宽泛，并未设定具体的量化指标，从规则角度给与了主管部门充分的自由裁量空间；此外，目前实际触发或公开报道的案例也较为有限，尚未形成可资参照的监管实践。在未来的监管实践中，不排除面向 C 端用户的生成式人工智能产品均可能被要求进行安全评估，或相关服务提供者被主管部门主动联系和要求进行安全评估的情况，因此，服务提供者可以考虑结合业务情况和主管部门咨询和确认。此外，需要注意的是，《互联网信息服务深度合成管理规定》还要求互联网应用商店等应用程序分发平台落实上架审核，核验深度合成类应用程序的安全评估、备案等情况，所以服务提供者亦可考虑提前和应用程序分发平台沟通确认是否需要完成安全评估后方可上架，并据此合理安排安全评估的时间表。

（2） 算法备案

根据《互联网信息服务算法推荐管理规定》的相关规定，具有舆论属性或者社会动员能力的算法推荐服务提供者，应当履行备案和变更、注销备案手续；《互联网信息服务深度合成管理规定》除要求深度合成服务提供者依照前述规定履行备案和变更、注销备案手续外，另要求深度合成服务技术支持者亦需参照履行备案和变更、注销备案手续。"具有舆论属性或者社会动员能力"的认定标准可以参考我们在"安全评估"的相关分析。

目前，算法备案的落地规定已相对成熟，备案平台（https://beian.cac.gov.cn/#/index）已上线运行近 1 年。根据公开的算法备案清单，我们注意到目前备案的算法类别以“个性化推送类”、“检索过滤类”为主，也有一些天猫、淘宝、钉钉、快手、美团等互联网企业的产品办理了生成合成类算法备案，但数量相对较少。

（3） 科技伦理审查

根据《科技伦理审查办法（试行）（征求意见稿）》，从事人工智能科技活动的单位，研究内容涉及科技伦理敏感领域的，应设立科技伦理（审查）委员会；特别地，具有舆论社会动员能力和社会意识引导能力的算法模型、应用程序及系统的研发活动，被列为“需要开展专家复核的科技活动”，在通过本单位科技伦理（审查）委员会的初步审查后，还应报请所在地方或相关行业主管部门组织开展专家复核。

该办法目前仍处于征求意见阶段，尚不确定未来实际执行的的监管口径和落地安排，建议持续关注相关规定的最新动态。

除了安全评估、算法备案和科技伦理审查，生成式人工智能服务提供者在提供相关产品服务时，还需要关注其自身的产品或服务形态是否构成增值电信业务，进而需要取得增值电信业务经营许可证；以及是否构成受监管的视听类服务活动，进而需要取得《网络文化经营许可证》《网络出版服务许可证》《信息网络传播视听节目许可证》《广播电视节目制作经营许可证》，由于其中部分证照实践中取得难度较高，服务提供者应根据自身情况调整展业方式。

（二） 运营合规要求

结合《征求意见稿》，我们从投资项目角度，将需要提请重点关注的事项总结如下：

合规事项	合规内容
训练数据的合法性	用于生成式人工智能产品的预训练、优化训练数据，应满足以下要求： 1) 符合《中华人民共和国网络安全法》等法律法规的要求； 2) 不含有侵犯知识产权的内容； 3) 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； 4) 能够保证数据的真实性、准确性、客观性、多样性； 5) 国家网信部门关于生成式人工智能服务的其他监管要求。
数据标注	生成式人工智能产品研制中采用人工标注时，提供者应当制定符合规定要求，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。
生成内容标识	提供者对使用其服务生成或者编辑的信息内容，应当采取技术措施添加不影响用户使用的标识，并依照法律、行政法规和国家有关规定保存日志信息；对生成的图片、视频等可能导致公众混淆或者误认的内容在生成或者编辑的信息内容的合理位置、区域进行显著标识，向公众提示深度合成情况
真实身份	提供生成式人工智能服务应当按照《中华人民共和国网络安全法》规定，要求用户提供真

认证	实身份信息。
用户管理	1) 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。 2) 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、知悉生成的文本、图片、声音、视频等侵害他人肖像权、名誉权、个人隐私、商业秘密，或者不符合规定要求时，应当采取措施，停止生成，防止危害持续。 3) 提供者应当指导用户科学认识和理性使用生成式人工智能生成的内容，不利用生成内容损害他人形象、名誉以及其他合法权益，不进行商业炒作、不正当营销。用户发现生成内容不符合规定要求时，有权向网信部门或者有关主管部门举报。 4) 提供者发现用户利用生成式人工智能产品过程中违反法律法规，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。
生成内容管理	1) 利用生成式人工智能生成的内容应当体现社会主义核心价值观，不得含有颠覆国家政权、推翻社会主义制度，煽动分裂国家、破坏国家统一，宣扬恐怖主义、极端主义，宣扬民族仇恨、民族歧视，暴力、淫秽色情信息，虚假信息，以及可能扰乱经济秩序和社会秩序的内容。 2) 提供者不得根据用户的种族、国别、性别等进行带有歧视性的内容生成。 3) 对于运行中发现、用户举报的不符合规定要求的生成内容，除采取内容过滤等措施外，应在 3 个月内通过模型优化训练等方式防止再次生成。
用户信息保护及使用限制	提供者在提供服务过程中，对用户的输入信息和使用记录承担保护义务。不得非法留存能够推断出用户身份的输入信息，不得根据用户输入信息和使用情况进行画像，不得向他人提供用户输入信息。

除了《征求意见稿》提及的以上事项，根据《网络安全法》《数据安全法》《个人信息保护法》等相关规定，如下常见的数据合规事项也需要予以关注：

1) **网络安全等级保护**：是否根据规定完成了网络安全等级保护制度所要求的定级备案、安全建设整改、等级测评和自查等工作；
2) **数据跨境合规**：是否涉及个人信息或重要数据出境活动，是否依法通过数据出境安全评估、个人信息出境标准合同、个人信息保护认证等方式满足数据出境条件；此外，对于双向的数据跨境活动，亦需考虑是否符合其他境外适用法律规定；
3) **重要数据识别及保护**：是否涉及重要数据（如与政府、军队存在大模型项目合作）；如涉及重要数据，是否依法履行了重要数据目录备案、设立数据安全负责人及管理机构、定期开展风险评估等合规义务；
4) **用户协议和隐私政策审查**：从《个人信息保护法》等相关规定角度审查用户协议和隐私政策的条款是否完备、是否贴合业务实际情况。

（三） 业务合同审查

除了面向 C 端的产品，生成式人工智能服务提供者还可能为企业、科研院所、政府和军队等客户提供服务。从投资项目的角度，我们建议围绕如下重点问题核查：

1) 结合业务合同反映出的业务模式，进一步核查和确认项目公司涉及的业务资质类型以及

运营合规问题；

2) 实际签署的业务合同是否与公司描述的业务模式一致；
3) 是否限制或排除项目公司在特定领域或与特定客户或供应商的合作；
4) 与重要客户、供应商的合作稳定性（如合作期限、解除权等）。

对于提供生成式人工智能服务的项目公司，除了常规类型合同，还需要特别关注其与训练数据采购以及算力供应相关的合同。

就训练数据采购，应关注数据提供方是否对提供的数据拥有合法的权利，以及购买的数据中是否涉及个人信息。如果涉及个人信息，需要注意核查第三方是否就数据的提供和使用以征得用户同意等方式取得合法性基础，数据提供方是否就数据违法行为承担兜底赔偿责任。

算力供应，通常包括自采、外采和算力共享等多种方式。从投资项目角度，需要关注算力是否能满足项目公司的需求，以及算力供应是否稳定（比如相关算力采购或合作协议的期限等）。

三、 小结

生成式人工智能技术作为一项突破性技术，将深刻影响我们的生活和工作方式。而鉴于生成式人工智能服务需要处理大量的数据，可能涉及个人信息和重要数据，生成式人工智能服务的应用中，也可能会出现不符合社会伦理和公序良俗的情况（如：恶意攻击、欺诈等行为），公众开始担心生成式人工智能技术会在网络和数据安全、个人信息和隐私保护、科技伦理和社会秩序等各方面产生新的冲击。对此，监管部门已经出台了一系列规定，以促进生成式人工智能健康发展和规范应用，但该等规定具体如何把握和执行有待未来更多实践的检验，也需要提请生成式人工智能服务提供者及相关投资机构对行业规定予以持续关注。