机遇与挑战丨生成式人工智能法律合规：生成式人工智能的监管现状与主要法律风险

实务问题	公司业务经营 （ 公司运营->公司业务经营 ）		
发文日期	2024-01-04	作者	李瑛莉
作者来源	三秋律师事务所		
法学分类	计算机与互联网 （ 科技法->计算机与互联网 ）		
行业分类	科学与技术		

正文内容：

● 引 言

近年来，生成式人工智能发展迅速，取得了众多进展和突破，据统计，预计到2026年，生成式人工智能市场将从2021年的15亿美元增长到65亿美元，复合年增长率达34.9%；生成式人工智能产生的数据将占所有数据的10%，而目前这一比例还不到1%。但同时，该技术的快速发展也带来了一系列新问题与挑战。

我们根据《生成式人工智能服务管理办法》《互联网信息服务深度合成管理规定》《互联网信息服务算法推荐管理规定》《生成式人工智能服务安全基本要求（征求意见稿）》等规范性文件的相关要求，为生成式人工智能服务提供者的运营进行法律层面的合规指引。并将于近期通过系列文章陆续推出，欢迎各位持续关注。

一、 我国对生成式人工智能的监管现状

（一） 生成式人工智能的立法现状

2023年7月13日，国家互联网信息办公室等七部门联合发布了《生成式人工智能服务管理暂行办法》，并于2023年8月15日正式生效实施。《生成式人工智能服务管理暂行办法》是我国首份专门针对生成式人工智能的监管文件，也是全球范围内最早生效的一部专门针对生成式人工智能领域的行政立法。

目前，我国与生成式人工智能最直接相关，最具有指导意见的规范性文件为《生成式人工智能服务管理办法》《互联网信息服务深度合成管理规定》《互联网信息服务算法推荐管理规定》《具有舆论属性或社会动员能力的户互联网信息服务安全评估规定》《生成式人工智能服务安全基本要求（征求意见稿）》，具体情况详见表1。

表 1：有关生成式人工智能的主要监管文件

法规名称	发文机关	实施时间	说明
《生成式人工智能服务管理暂行办法》	网信办、国家发改委、教育部、科技部、工信部、公安部、广电总局	2023.8.15	生成式人工智能基础性、专门性的监管法规
《互联网信息服务深度合成管理规定》	网信办、工信部、公安部	2023.1.10	涉及深度合成的监管法规
《互联网信息服务算法推荐管理规定》	网信办、工信部、公安部、市场监督管理总局	2022.3.1	涉及算法推荐的监管法规
《具有舆论属性或社会动员能力的户互联网信息服务安全评估规定》	网信办、公安部	2018.11.30	涉及具有舆论属性或社会动员能力的户互联网信息服务的安全评估要求
《生成式人工智能服务安全基本要求（征求意见稿）》	全国信息安全标准化委员会	征求意见稿阶段	给出了生成式人工智能服务在安全方面的基本要求，包括数据安全、模型安全、安全措施、安全评估等

（二） 生成式人工智能的监管范围

根据《生成式人工智能服务管理暂行办法》的规定，“利用生成式人工智能技术向中华人民共和国境内公众提供生成文本、图片、音频、视频等内容的服务，适用本办法。”可见，《生成式人工智能服务管理暂行办法》是以“是否向境内公众提供服务”作为适用范围的区分标准，而未对生成式人工智能服务提供者是位于境内还是境外进行区分。也就是说，只要是面向境内公众提供生成式人工智能服务，无论提供者是位于中国境内还是境外，都会受到相应的监管。

同时，《生成式人工智能服务管理暂行办法》又从反面作出了排除性规定，即“行业组织、企业、教育和科研机构、公共文化机构、有关专业机构等研发、应用生成式人工智能技术，未向境内公众提供生成式人工智能服务的，不适用本办法的规定。”实践中，许多企业接入人工智能产品来服务于企业内部的管理需求，以有效提升工作效率。这些就属于企业内部使用生成式人工智能技术，未向境内公众提供生成式人工智能服务的情况，《生成式人工智能服务管理暂行办法》将这些情况排除在监管范围之外。[1]

（三） 生成式人工智能的监管原则

《生成式人工智能服务管理暂行办法》规定了对生成式人工智能服务进行监管的两条基本原则。

图 1：有关生成式人工智能的主要监管文件

● **包容审慎**

坚持发展和安全并重、促进创新和依法治理相结合的原则，采取有效措施鼓励生成式人工智能创新发展。

同我国一样，东盟也选择了相对宽松的人工智能监管态度。根据东盟《人工智能伦理与治理指南（草案）》，东盟的人工智能治理将更加宽松，没有规定不可接受的风险类别。指南建议企业建立人工智能风险评估机构并展开培训，但具体细节由企业和当地监管机构决定。同时警告了人工智能被用于虚假信息、深度造假和冒充他人的风险。

● **分类分级**

国家有关主管部门针对生成式人工智能技术特点及其在有关行业和领域的服务应用，制定相应的分类分级监管规则或者指引。2023 年 10 月 18 日，国家互联网信息办公室发布《全球人工智能治理倡议》，其中就提到要推动建立风险等级测试评估体系，实施敏捷治理，分类分级管理，快速有效响应。

欧盟的《人工智能法案（草案）》采用的就是分类分级的风险规制路径，其将人工智能系统分成四个风险级别，分别是不可接受的风险、高风险、有限风险和极低风险，并针对不同类型风险施加了不同的监管措施以及相应类型的人工智能系统运营者的义务。

（四） 生成式人工智能服务提供者的角色定位

《生成式人工智能服务管理暂行办法》第九条规定，“提供者应当依法承担网络信息内容生产者责任，履行网络信息安全义务。涉及个人信息的，依法承担个人信息处理者责任，履行个人信息保护义务。”也就是说生成式人工智能服务提供者最起码具备两个层面的角色。

图 2：生成式人工智能服务提供者角色定位

● **信息内容生产者**

当涉及到信息内容时，生成式人工智能服务提供者是信息内容生产者的角色。

根据《网络信息内容生态治理规定》“网络信息内容生产者”指的是“制作、复制、发布网络信息内容的组织或者个人”。其不得制作、复制、发布违法信息，并应当采取措施防范和抵制制作、复制、发布不良信息。

同时，根据信息网络传播权的相关规定，信息网络提供者分为网络内容提供者与网络服务提供者，相较于网络服务提供者，内容提供者对相应的侵权行为承担直接侵权责任，即无法适用避风港原则。

● **个人信息处理者**

当涉及到个人信息时，生成式人工智能服务提供者是个人信息处理者的角色。个人信息处理者，是指在个人信息处理活动中自主决定处理目的、处理方式的组织、个人。《个人信息保护法》中规制的主要对象即为个人信息处理者，也就是说，如果涉及到个人信息，生成式人工智能服务提供者需要履行《个人信息保护法》等相关法律法规、标准、指南中的所有义务并承担违法处理个人信息产生的法律责任。

二、 生成式人工智能的主要法律风险

生成式人工智能技术的快速发展，为经济社会发展带来新机遇的同时，也产生了传播虚假信息、侵害个人信息权益、数据安全和偏见歧视等风险。从法律层面来看，生成式人工智能主要有以下风险。

图 3：生成式人工智能主要法律风险

（一） 数据风险

数据被看作是决定未来发展的关键生产要素。有能力获得海量的、高质量的数据，被看做未来大模型公司的核心竞争力之一。然而，数据作为全新的生产要素，也带来一系列亟待解决的问题与风险。

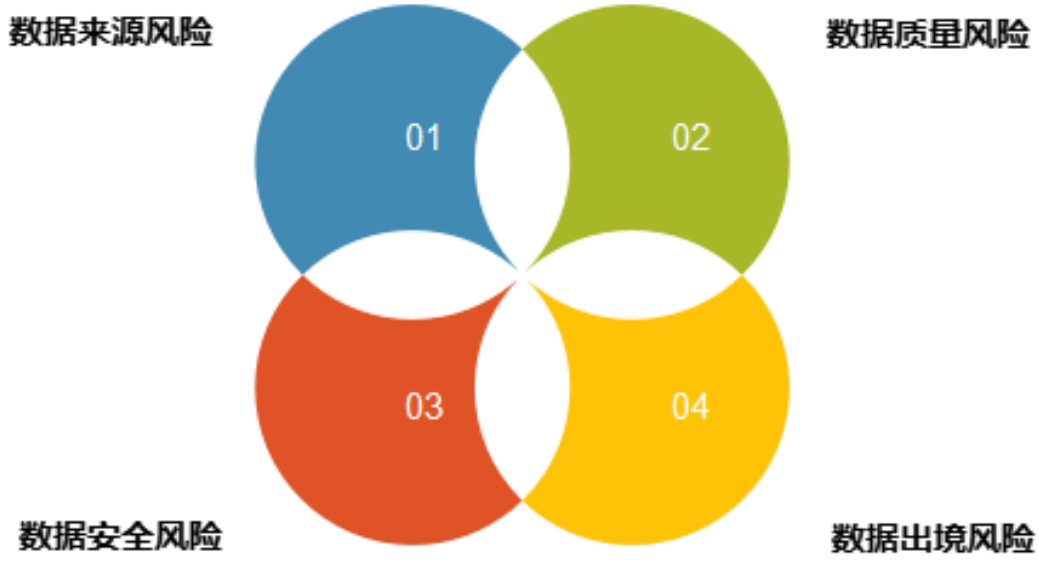

图 4：生成式人工智能主要数据风险

1. 数据来源风险

训练数据中可能会包含个人信息，如果获取的个人信息没有经过个人信息主体的有效同意，或者违反《个人信息保护法》等相关法律法规的要求非法处理个人信息，则生成式人工智能服务提供者可能会面临没收违法所得、罚款、停业整顿、吊销相关业务许可或者吊销营业执照等行政处罚，构成犯罪的，还可能会被追究刑事责任。

国内之前大火的 AI 换脸软件 ZAO 就因为违规收集人脸数据而被约谈下架。2023 年 6 月 29 日，16 位匿名人士在加州旧金山联邦法院起诉微软和 OpenAI，声称这两家公司基于 ChatGPT 的人工智能产品在没有充分通知或没有获得充分同意的情况下收集并泄露了他们的个人信息。

此外，大多数的训练数据是通过爬虫爬取的，但是爬取行为并不都是正当行为。如果爬取的内容达到实质替代数据来源网站的程度，则是一种“搭便车”行为。而绕过被爬取对象设定的访问权限或者破坏网站所设定的技术措施爬取非公开数据，会干扰被爬取方的正常运行，给被爬取方造成损害，也属于不正当竞争行为。

2. 数据质量风险

生成式人工智能的正常运行需要大量高质量数据来训练底层神经网络。在第四届联合国世界数据论坛期间，国际统计学会候任主席、密歇根大学统计系教授何旭铭表示，“像 ChatGPT 这样的人工智能模型需要大量的数据，通过高效的算法得出结果，实际还是非常依赖数据本身的质量以及全面性的。”如果训练数据中存在大量的有毒或者带有种族、性别和宗教偏见的低质量数据，模型往往表现出不良行为，进而可能生成虚假、诱骗信息等不良信息。

2022 年 12 月 5 日，知名技术问答交流网站“Stack Overflow”已暂时禁止用户分享由 ChatGPT 生成的回答，因为 ChatGPT 生成的大量错误信息充斥在“Stack Overflow”网站上。也有研究机构发出警告，称以亚马逊为代表的电商平台上，充斥着各种人工智能生成的蘑菇觅食科普书籍，其中存在诸多错误。如果某人在野外求生中，根据这些人工智能生成的知识食用蘑菇，严重的可能会导致死亡。

3. 数据安全风险

在获取训练数据之后，训练数据本身的安全如何确保？如遭受蓄意攻击、数据中毒、数据泄露等。根据美国云计算服务软件提供商“Salesforce”在 2023 年 3 月对 500 多名 IT 行业领导者进行的一项调查显示，安全性是受访者最关心的关于 ChatGPT 等大模型的问题，71%的受访者认为生成式人工智能会给企业的数据安全带来新的风险。[2]

例如，当员工将公司信息以提问的方式输入 ChatGPT 后，可能会导致相关内容进入 OpenAI 的训练数据库，从而造成数据泄露的风险。意大利个人数据保护局曾宣布暂时中止意大利境内的 ChatGPT 访问途径，因为 ChatGPT 出现了安全漏洞，部分用户能够看到其他用户的姓名、邮箱、聊天记录标题以及信用卡最后四位数字等信息。中国支付清算协会也指出，ChatGPT 此类智能化工具已暴露出跨境数据泄露等风险，为了维护数据安全还发布了《关于支付行业从业人员谨慎使用 ChatGPT 等工具的倡议》。

4. 数据出境风险

ChatGPT 作为当前人工智能领域的前沿技术代表，虽然目前在国内处于禁用状态。但是仍然有不少国内企业向 OpenAI 等国外公司抛出橄榄枝，试图将 ChatGPT 等类似的生成式人工智能服务引入到自己的应用软件或网页中，或者在业务运营过程中使用 ChatGPT 等类似的国外生成式人工智能服务。

根据 ChatGPT 的工作机制以及 OpenAI 用户协议披露的内容，用户在输入端口提出问题后，用户与其对话的数据将会被存储在 OpenAI 或其使用的云服务提供商的数据中心。这意味着，境内 ChatGPT 使用者在 ChatGPT 输入的隐私和个人信息、商业秘密等数据可能被用于 ChatGPT 未来模型的迭代训练。在这样的一个过程中，存在着数据跨境流动安全风险。[3]

在算力不足的种情况下，不少生成式人工智能开发者在探讨跨境调用境外算力的可能性。如生成式人工智能开发者跨境调用境外算力，则其采集的训练数据可能将会被传输至境外进行训练，相关训练数据和搭建后的模型亦可能被存储至境外数据中心，从而引发数据跨境相关风险。此外，实践中，生成式人工智能开发者还可能直接调用境外的算法模型来训练自己的定制化模型。在这一过程中，境内的生成式人工智能开发者需要将其采集的相关行业数据、业务数据等数据传输至境外用以模型训练，也会引发数据跨境的风险。[4]

（二） 算法模型风险

人工智能算法存在固有缺陷，在透明度、鲁棒性、偏见与歧视方面存在尚未克服的技术局限，导致算法应用问题重重。[5]

透明度方面。由于算法模型的黑箱运作机制，其运行规律和因果逻辑并不会显而易见的摆在研发者面前。这一特性使人工智能算法的生成机理不易被人类理解和解释，一旦算法出现错误，透明度不足无疑将阻碍外部观察者的纠偏除误。

鲁棒性方面。算法运行容易受到数据、模型、训练方法等因素干扰，出现非鲁棒特征。例如，当训练数据量不足的情况下，在特定数据集上测试性能良好的算法很可能被少量随机噪声的轻微扰动影响，从而导致模型给出错误的结论；在算法投入应用之后，随着在线数据内容的更新，算法很可能会产生系统性能上的偏差，进而引发系统的失灵。

偏见与歧视方面。算法以数据为原料，如果初始使用的是有偏见的数据，这些偏见可能会随着时间流逝一直存在，无形中影响着算法运行结果，最终导致算法模型生成的内容存在偏见或歧视。

（三） 伦理风险

美国技术哲学家詹斯·摩尔曾经提出过这样一条定律“伴随着技术革命，社会影响增大，伦理问题也增加”，这被称为科技伦理领域的摩尔定律。作为一种具有革命性的人工智能技术工具，生成式人工智能存在更多的伦理风险。OpenAI 在对 DALL-E 2 的生成结果进行公平性测试时发现，其表现出显著的性别和种族歧视，例如，提示词“律师”“CEO”时，几乎生成的图像都是白人男性。

根据《人工智能伦理安全风险防范指引》，开展人工智能相关活动，所可能产生的伦理安全风险主要包括：

- **失控性风险：**人工智能的行为与影响超出研究开发者、设计制造者、部署应用者所预设、理解、可控的范围，对社会价值等方面产生负面影响的风险。

- **社会性风险：**人工智能使用不合理，包括滥用、误用等，对社会价值等方面产生负面影响的风险。

- **侵权性风险：**人工智能对人的基本权利，包括人身、隐私、财产等造成侵害或产生负面影响的风险。

- **歧视性风险：**人工智能对人类特定群体的主观或客观偏见影响公平公正，造成权利侵害或负面影响的风险。

- **责任性风险：**人工智能相关各方行为失当、责任界定不清，对社会信任、社会价值等方面产生负面影响的风险。

（四） 知识产权风险

生成式人工智能所引发的知识产权侵权风险已经成为整个行业发展所面临的紧迫问题。近年来，已经有多家生成式人工智能公司因为版权问题而被起诉。

2013 年 3 位艺术家对 Midjourney 提起诉讼，称其在“未经原作者同意的情况下”从网络上获取的 50 亿张图像来训练其人工智能，侵犯了“数百万艺术家”的权利；2022 年 11 月 7 日，部分程序员以违反开源协议和著作权按侵权为由将 Microsoft、OpenAI 和 Github 诉至美国加州北区地方法院，索赔 90 亿美元；2023 年 1 月，全球知名图片提供商华盖创意起诉热门人工智能绘画工具 Stable Diffusion，称其未经许可从网站上窃取了数百万张图片，用于训练其人工智能。

生成式人工智能在训练数据库的输入阶段和输出阶段，最大的著作权侵权风险是侵犯复制权和改编权。在输入阶段，如果将大量受著作权保护的作品用来训练人工智能，这本身看似出于学习目的，实则最终服务于商业目的，很难使用现有的著作权合理使用制度规避侵权责任。在输出阶段，如果生成的内容与原作品在表达上构成实质性相似，则可能侵犯复制权；如果在保留原作品表达的基础上形成了新的表达，则可能涉及改编权问题。[6]

（五） 内容安全风险

一直以来，互联网信息空间都面临着虚假信息和信息内容安全的挑战，国内外互联网内容平台，都不断在提升其虚假内容和信息安全的治理能力。但随着生成式人工智能的出现，虚假信息和信息内容安全的挑战进一步增加。

很多用户发现 ChatGPT 就存在“一本正经胡说八道”的问题。普林斯顿计算机科学教授 Arvind Narayanan 指出：“人们对使用 ChatGPT 进行学习感到兴奋。但危险在于，除非你已经知道答案，否则你无法判断它什么时候生成的结果是错的。我尝试了一些基本的信息安全问题，答案听起来很有道理，但实际上是胡说八道。”

近日，北京卫健委日前牵头组织制定了《北京市互联网诊疗监管实施办法（试行）》（以下简称《办法》），并向社会公开征求意见。根据《办法》规定，医疗机构开展互联网诊疗活动要加强药品管理，严禁使用人工智能等自动生成处方。此项规定也是考虑到人工智能生成内容安全性的问题。

需要强调的是，生成物内容是最易被用户或者监管机构感知的，如果生成物内容存在安全性问题，例如生成不良、虚假或违法信息，极易引发监管风险。《生成式人工智能服务安全基本要求（征求意见稿）》列出了数据及生成内容的主要安全风险，共 5 类 31 种。

表 2：数据及生成内容的主要安全风险

风险类型	主要内容
违反社会主义核心价值观	· 煽动颠覆国家政权、推翻社会主义制度； · 危害国家安全和利益、损害国家形象； · 煽动分裂国家、破坏国家统一和社会稳定； · 宣扬恐怖主义、极端主义； · 宣扬民族仇恨、民族歧视； · 宣扬暴力、淫秽色情； · 传播虚假有害信息； · 其他法律、行政法规禁止的内容。
歧视性内容	· 民族歧视内容； · 信仰歧视内容； · 国别歧视内容； · 地域歧视内容； · 性别歧视内容； · 年龄歧视内容； · 职业歧视内容； · 健康歧视内容； · 其他方面歧视内容。
商业违法违规	· 侵犯他人知识产权； · 违反商业道德； · 泄露他人商业秘密； · 利用算法、数据、平台等优势，实施垄断和不正当竞争行为； · 其他商业违法违规行为。
侵犯他人合法权益	· 危害他人身心健康； · 侵害他人肖像权； · 侵害他人名誉权； · 侵害他人荣誉权； · 侵害他人隐私权； · 侵害他人个人信息权益； · 侵犯他人其他合法权益。
无法满足特定服务类型的安全需求	该方面主要安全风险是指，将生成式人工智能用于安全需求较高的特定服务类型，例如自动控制、医疗信息服务、心理咨询、关键信息基础设施等，存在如下情形： · 内容不准确，严重不符合科学常识或主流认知； · 内容不可靠，虽然不包含严重错误的内容，但无法帮助使用者解答问题。

（六） 滥用风险

生成式人工智能的社会价值体现为革新数字内容与艺术创造领域，并将辐射到其他领域和行业如医疗、教育、传媒、影视、工业，以及元宇宙、数字人领域，孕育新的技术形态与价值模式。生成式人工智能以高效率、低成本满足个性化需求，完成基础性工作，释放人类创造力，推动艺术创造领域与基础概念革新。[7]

与此同时，不法分子利用生成式人工智能模型或工具，可以以更低的门槛、更高的效率来制作出种类丰富、真伪难辨的音视频、图片和文字等虚假信息，引发了层出不穷的基于深度合成的诈骗、色情、诽谤、假冒身份等新型违法犯罪行为。

2021 年，国家互联网信息办公室、公安部指导北京、天津、上海、浙江、广东等地方网信部门、公安机关，针对未履行安全评估程序的语音社交软件和涉“深度伪造”技术的应用，依法约谈了映客、小米、阿里巴巴、腾讯等 11 家企业。

近日，公安部召开新闻发布会，通报依法打击治理侵犯公民个人信息违法犯罪行为的有关情况。发言人表示，为打击“人工智能换脸”新型诈骗，公安机关联合国家重点实验室等单位，开展人脸识别与活体检测技术安全测评，覆盖了境内用户量大、问题隐患突出的即时通讯、网络直播、网络社交、电商平台、金融支付等重点 APP。严打泄露身份证照片等图像信息的犯罪源头，破获“人工智能换脸”案件 79 起，抓获犯罪嫌疑人 515 名。

本文作者

李瑛莉
北京三秋律师事务所
lliyingli@vtlaw.cn
18810973315
北京办公室

业务领域：知识产权，不正当竞争，网络安全，数据合规

李瑛莉律师专注于数据合规、知识产权、不正当竞争等诉讼以及非诉业务，在科技和互联网公司有多年全生命周期法律服务经验。服务的客户包括腾讯、新浪、爱奇艺、中文在线、龙图游戏等知名公司，并得到客户一致认可。

****

[1]《确保网络数据安全应对新一代人工智能治理挑战》，梁正，何嘉钰，2023.8.1，“中国信息安全”微信公众号。

[2]《ChatGPT 的数据跨境安全风险及合规要点》，孙俊律师，2023.3.26，“互利魔笛” 微信公众号。

[3]《AIGC 数据跨境的法律监管和合规路》，蔡荣伟，斯响俊，杨杰，2023.10.26，“中伦视界”微信公众号。

[5]《人工智能生成内容（生成式人工智能）白皮书（2022 年）》。

[6]《生成式人工智能数据训练知识产权合法性问题探讨》，张平，2023.8.31，《知识产权家》微信公众号。

[7]《生成式人工智能发展趋势报告 2023》，腾讯研究院。

LexisNexis-评论文章/Articles
机遇与挑战丨生成式人工智能法律合规系列（一）：生成式人工智能的监管现状与主要法律风险
https://www.lexiscn.com/topic/legal.php?tps=pg_cp&act=detail&id=458869&newstype=3&eng=0