君合专题研究报告 JUNHE

2023 年 4 月 17 日

人工智能和算法系列文章（二）：网信办发布《生成式人工智能服务管理办法（征求意见稿）》，人工智能法律治理持续发力

2023 年 4 月 11 日，国家互联网信息办公室（以下简称“网信办”）发布《生成式人工智能服务管理办法（征求意见稿）》（以下简称“《办法》”），该《办法》是继《互联网信息服务算法推荐管理规定》（以下简称“《算法规定》”）《互联网信息服务深度合成管理规定》（以下简称“《深度合成规定》”）后我国关于生成式人工智能的重要立法，是国家监管机构对近来爆火的 ChatGPT 等新型人工智能技术的回应。

本文拟结合生成式人工智能技术特征所引发的主要法律风险，系统性解读《办法》所提出的应对风险的各项法律要求。

一、生成式人工智能的主要法律风险

生成式人工智能是算法、算力和数据的进步共同推动的结果，特别地，训练算法的革新推动了生成式人工智能的出现。从技术本质而言，生成式人工智能就是对海量数据学习资料的重组，因此生成的内容与用户输入的材料紧密关联，不过这种技术对输入的数据类型和输出内容均不包含自身的价值判断，这就导致其存在以下法律风险：

1.虚假信息泛滥：由于生成的内容只是与用户输入的资料具有语义上的关联，但未必正确，因此容易产生误导性的虚假信息。在此背景下，一旦生成式人工智能技术被不法分子利用，可能被用于伪造各种文本、视频和图片，甚至是诈骗、恐吓、诽谤。[1]

2.安全隐私危机和他人知识产权等合法权益损害：首先，用于进行算法训练的数据如果包含用户的个人信息，一旦信息泄露会引发巨大的个人信息权益威胁，即使生成的内容进行了加密、加噪等技术处理措施，仍然可以完成对原始数据的恢复，[2]甚至可能出现非法留存用户输入的信息并进一步用于用户画像等分析目的；其次，由于生成式人工智能的输出内容与输入数据材料的关联性，因此其输出内容可能造成作品、专利、商标侵权、不正当竞争、侵犯商业秘密等问题。

3.歧视和社会舆论风险：生成式人工智能的训练材料来自于人类的作品，因此可能将人类作品中包含的歧视因素继承并反映在输出的结果中。[3]同样的，生成内容也会引发不当言论，即使研发者已经

[1] Dash B., Sharma P., Are ChatGPT and Deepfake Algorithms Endangering the Cybersecurity Industry? A Review, International Journal of Engineering and Applied Science, No.1, 2014, pp.1-5.

[2] Carlini, et al, “Extracting Training Data from Diffusion Models”, Arxiv Preprint, 2014, No.2301.13188, available at: https://arxiv.org/pdf/2301.13188.pdf, last visited on April 11, 2023.

[3] https://thuvienpc.com/the-downsides-of-chatgpt/

对数据模型进行技术处理，拒绝对某些敏感话题进行回答，但依然会在用户的诱导下输出不当言论。

4.安全可控和社会伦理风险：作为人工智能技术的一种类型，生成式人工智能同样面临着安全可控和社会伦理风险，具体可能体现在利用生成式人工智能产品进行网络攻击、网络炒作、编写恶意软件和实施不正当的商业营销等行为。

二、《办法》规定的主要内容

针对生成式人工智能存在的上述法律问题和风险，《办法》从不同角度对研发和利用生成式人工智能的组织和个人提出了具体的法律要求。

（一） 适用范围

从适用主体上看，根据《办法》第 2 条的规定，该《办法》适用于研发、利用生成式人工智能，面向中华人民共和国境内公众提供服务的场景，因此不排除境外实体因向中国境内公众提供生成式人工智能服务而因此落入《办法》的适用范围的可能性，但有待《办法》生效后监管部门的进一步澄清。

从适用技术上看，生成式人工智能，是指基于算法、模型、规则生成文本、图片、声音、视频、代码等内容的技术。在《深度合成规定》中，将“深度合成技术”界定为利用深度学习、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，二者在文字描述上十分相似，可以看出深度合成技术是生成式人工智能的一种技术内容。此外，按照《办法》第 5 条的规定，通过提供可编程接口等方式支持他人自行生成文本、图像、声音等也属于利用生成式人工智能的活动，落入《办法》的适用范围。

（二） 基本原则

《办法》第 4 条对生成式人工智能需要满足的基本原则进行了规定，即遵守法律法规的要求，尊重社会公德、公序良俗，具体体现在避免出现法律法规禁止或限制的行为和内容，反歧视，尊重知识产权和商业道德，内容真实准确以及尊重和保护他人的合法权益。该等原则承自《网络安全法》第 12 条第 2 款的规定，基本囊括了生成式人工智能规制的原则。

（三） 具体要求

1. 前置要求：算法安全评估和备案义务

《办法》第 6 条明确了利用生成式人工智能向公众提供服务前，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《算法规定》履行算法备案手续。《办法》将算法安全评估和备案义务作为开展生成式人工智能服务的前置性要求，重述《算法规定》中已规定的安全评估和算法服务备案。从目前既有的实践情况来看，已有一批头部互联网企业和平台已完成算法备案。根据我们的观察和经验，算法推荐服务技术的开发/提供者和应用该技术的 App 运营者均可能被要求进行算法备案。

2. 数据来源合法性要求

首先，《办法》第 7 条要求生成式人工智能研发利用者对预训练、优化训练数据来源的合法性负责，包括（1）符合《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）涉及个人信息的应具有开展个人信息处理活动的合法性基础；（4）保证数据的真实性、准确性、客观性、多样性。但对于生成式人工智能研发利用者对用于训练算法的数据来源的审核义务究竟应达到何种程度，可能需要考虑该类技术的数据的来源多样性和较高的审核溯源成本。

其次，《办法》第 8 条对产品研制中采用人工标注提出要求，包括（1）制定清晰、具体、可操作的标注规则；（2）对标注人员进行必要培训；（3）确

保抽样核验标注内容的正确性。

3. 对服务提供者的要求

除了生成式人工智能的研发环节，《办法》对利用生成式人工智能提供服务的主体提出各项具体的义务，该类义务既包括原有的其他法律法规中的各项义务的细化，也包含结合生成式人工智能的特点新增的义务类型：

（1） 用户身份验证义务：《办法》第 9 条重述了《网络安全法》中用户真实身份的验证义务。

（2）防沉迷义务：《办法》第 10 条要求服务提供者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。对此，此前《算法规定》第 8 条中也有类似规定，要求不得设置诱导用户沉迷、过度消费等违反法律法规或者违背伦理道德的算法模型。

（3）用户个人信息保护和响应义务：《办法》遵循了《个人信息保护法》设置的各项义务，对个人信息权益给予充分的保护。一方面，第 11 条要求服务提供者对于用户的输入信息和使用记录承担保护义务，除非有法定的例外情形，第一，对于能够推断用户身份的输入信息，不得非法留存；第二，禁止根据用户输入内容进行用户画像；第三，不得向他人提供用户输入信息。但该要求如何与数据来源合法性和内容审核义务进行平衡和兼顾，有待进一步检验。

（4）避免使用歧视性内容的义务：《办法》第 12 条要求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。该等标签往往具有较高的敏感性，容易生成具有歧视性的内容。

（5）设立用户投诉处理机制的义务：《办法》第 13 条要求建立用户投诉接收处理机制，及时处理个人关于更正、删除、屏蔽其个人信息的请求，该规定与《个人信息保护法》确立的个人信息主体行权响应机制相呼应。

（6）及时处置违法信息的义务：《办法》第 13 条还要求服务提供者在发现、知悉生成内容存在侵权时及时停止内容生成并防止危害持续，对此可以理解为包含“知道或应当知道”的含义，即要求服务提供者尽到合理的注意义务主动审核排查内容生成是否存在侵权风险。

（7）提供安全稳健服务的义务：《办法》第 14 条对服务的安全性、稳健性和持续性提出要求。

（8）生成内容审核和标识义务：关于生成内容审核，《办法》第 15 条要求对于运行中发现、用户举报的不符合本办法要求的生成内容，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。此外，《办法》第 16 条还引入了《深度合成规定》中的标识义务，要求提供者应当对生成的图片、视频等内容进行标识。《深度合成规定》进一步细化该标识要求，即分为不影响用户使用的标识（第 16 条）和显著标识（第 17 条）两类，后者仅适用于可能导致公众混淆或者误认的场景。

（9）透明度要求：《办法》第 17 条要求服务提供者对用户履行必要的告知和解释说明的义务，即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。这与《个人信息保护法》及《深度合成规定》等所体现的透明度要求一脉相承，但可能产生如何平衡个人信息保护与企业商业秘密保护的问题，而对于告知披露的范围、形式和颗粒度在实践中如何落地，仍然需要进一步明确。

（10）积极指导义务：较之于之前的《算法规定》《深度合成规定》等，本次新出台的《办法》新增了一项义务，要求服务提供者引导用户科学合理

使用产品。首先，《办法》第 18 条要求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；其次，《办法》第 19 条还进一步要求提供者对服务中出现的用户违背商业道德、社会公德行为，及时处置并暂停或者终止服务

（四） 法律责任承担

首先，对于法律责任的承担主体，《办法》第 5 条区分了内容生产者和个人信息处理者，即产品及服务提供者首先应当承担“生成内容生产者的责任”，在此基础上如果涉及个人信息，产品及服务提供者还应承担作为个人信息处理者的责任。

其次，关于具体的处罚事由及措施，一方面，《办法》在第 20 条第 1 款规定了转致条款，指出处罚措施按照《网络安全法》、《数据安全法》和《个人信息保护法》的要求处理，另一方面，为应对新技术出现时的法律空白，《办法》第 20 条第 2 款同时规定，当法律、行政法规存在空白时，由网信部门和有关主管部门依据职责给予警告、通报批评，责令限期改正；拒不改正或者情节严重的，责令暂停或者终止其利用生成式人工智能提供服务，并处一万元以上十万元以下罚款。构成违反治安管理行为的，依法给予治安管理处罚；构成犯罪的，依法追究刑事责任。

三、评价和展望

值得肯定的是，《办法》回应了生成式人工智能在应用过程中可能产生的道德伦理、安全隐私、知识产权、歧视及不当言论等问题，特别是对于目前社会中热议的生成内容真实准确性问题，《办法》也进行了回应，但如何将该合规要求落地，有待监管机构的进一步解释和指引。

此外，《办法》也与既有法律法规（包括《网络安全法》、《数据安全法》、《个人信息保护法》等一般性法律和具体的《算法规定》、《深度合成规定》、《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》等）形成衔接，宣示了我国在人工智能和算法领域形成的初步法律治理体系。

《办法》中的规定对服务提供者提出了较高的合规要求，较之于原有的各类义务要求，还新增了要求服务提供者引导用户科学合理使用产品的义务和对不法内容过滤、防止再次生成的义务。以上各项法规要求如何与服务提供者现有的技术逻辑相适应，如何与企业的现有合规体系相衔接，部分法规要求将如何落地，有待于立法部门和执法部门的进一步解释，包括但不限于生成式人工智能服务与具有舆论属性或社会动员能力的互联网信息服务提供商之间的认定关系，生成式人工智能服务提供行为与上位法之间规定的各项责任认定情形的对应关系。

可以预见的是，随着人工智能和算法领域技术的不断革新，未来该领域的法律监管也将不断完善，对企业的合规要求也会不断细化，我们也将持续关注该领域的最新法规和监管动态。

董　潇　合伙人　电话：86 10 8519 1718　邮箱地址：dongx@junhe.com
郭静荷　律　师　电话：86 10 8553 7947　邮箱地址：guojh@junhe.com
冯毅捷　律　师　电话：86-10 8540 8723　邮箱地址：fengyijie@junhe.com
王威华　律　师　电话：86-10-85191213　邮箱地址：wangweihua@junhe.com
史晓宇　律　师　电话：86- 21 2283 8301　邮箱地址：shixiaoyu@junhe.com