君合专题研究报告 JUNHE

2023 年 3 月 31 日

人工智能和算法系列文章（一）：算法规定如何管理 ChatGPT 类产品

一、ChatGPT 类深度合成技术的快速崛起

深度合成（deep synthesis）是一种以深度学习为支撑的伪造技术，依托于“生成式对抗网络”（GAN）和“自动编码机”(autoencoders)对海量数据进行收集、训练，以创造、模拟、合成新的内容。最早期的深度合成产品一般表现为伪造的人像和视频，而随着技术的发展，深度合成技术已经广泛地拓展为包括图像、视频、文本、声音、微表情等多个领域，并展现出高度的真实性[1]。2022 年年底以来，美国人工智能研究实验室 OpenAI 新推出的一种人工智能技术驱动的自然语言处理工具 ChatGPT 的全球爆火，再度将深度合成技术带入人们的视野。ChatGPT(Chat Generative Pre-trained Transformer)即生成型预训练的 Transformer 模型。作为新一代的深度合成产品，ChatGPT 能够准确地识别和理解用户的语言，按照用户要求，生成用于不同场景、不同形式的文字，包括邮件、小说、论文和代码等，并在人机交互中进行再学习以改进输出文本的质量，并可以轻易通过图灵测试，达到了达到以假乱真的程度。新一代 GPT-4 技术能进一步接受图像、音频等多种输入并提供相应输出[2]。

随着深度合成展现出巨大的社会影响力，如何避免其可能带来的虚假信息危机等负面影响，从而使深度合成技术成为有效且可信的工具？在技术领域，相应的深度合成检测和鉴别技术也应运而生。与此同时，对深度合成技术的规范和引导开始成为各国普遍的监管趋势。

二、深度合成的算法监管框架：从《算法规定》到《深度合成规定》

中国较早就开始了算法监管的探索。2017 年 7 月 8 日，国务院印发了《新一代人工智能发展规划》提出了“建立人工智能安全监管和评估体系”的要求。2021 年九部门联合发布的《关于加强互联网信息服务算法综合治理的指导意见》要求“利用三年左右时间，逐步建立治理机制健全、监管体系完善、算法生态规范的算法安全综合治理格局”。

《个人信息保护法》首次在法律层面，从公平透明原则、个人的知情权和拒绝权等方面对个人信息的自动化决策进行了原则性的、相对全面的规定。

2021 年 12 月 31 日，国家网信办、工信部、公安部和市场监管总局联合发布《互联网信息服务算法推荐管理规定》（以下简称“《算法规定》”），这是中国第一部聚焦于算法治理的部门规章，系统性的规定了算法推荐服务的合规义务。

与 ChatGPT 这一深度合成产品风靡全球的同一时期，中国关于深度合成技术的规则亦已落地。2022 年 11 月 25 日，国家网信办、工信部、公安部

[1] https://m.thepaper.cn/newsDetail_forward_10962497

[2] https://new.qq.com/rain/a/20230316A09C5200

联合发布《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》”），该规定已于 2023 年 1 月 10 日起正式施行。根据《深度合成规定》第二十三条定义，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”。该条还列举了一些典型的深度合成技术，包括：（1）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。实践之中对应的例示有，写稿机器人、诗歌创作、客服聊天机器人、问答机器人等类 ChatGPT 产品；（2）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。实践之中例如语音合成工具、声音模仿、语音播报等；（3）音乐生成、场景声编辑等生成或者编辑非语音内容的技术。例如歌曲合成、音效生成等；（4）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。如 AI 换脸、美颜等；（5）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。例如画质修复等；（6）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。实践之中如元宇宙、VR、数字孪生等。

三、《深度合成规定》如何管理 ChatGPT 类产品

（一） 规制对象

《算法规定》将规制重点侧重于面向用户且使用了算法推荐技术的互联网信息服务提供者，《深度合成规定》则围绕深度合成技术，为一系列主体设定了行为规范，包括：深度合成服务提供者、深度合成服务技术支持者、深度合成服务使用者、应用程序分发平台，并面向不同主体设置具体行为规范。

（二）深度合成服务提供者合规义务

深度合成服务提供者是指提供深度合成服务的组织、个人。《深度合成规定》项下服务提供者的合规责任涵盖了各个方面和环节，包括，建立健全用户注册、算法机制机理审核、科技伦理审查、信息发布审核、数据安全、个人信息保护、反电信网络诈骗、应急处置等管理和技术制度，具有安全可控的技术保障措施。*（第七条）*

具体而言，深度合成服务提供者应特别注意以下合规义务：

（1）管理者责任

作为深度合成互联网信息服务渠道和平台的运营者，深度合成服务提供者一方面直接面向服务使用者乃至公众，一方面又连接和依托着技术支持者，在深度合成服务生态中具有特殊的定位。因此，《深度合成规定》将深度合成服务提供者作为最主要和最核心的规制对象，强调其作为管理者的身份和责任。

《深度合成规定》要求，深度合成服务提供者应当依法依约履行管理责任，包括，制定和公开管理规则、平台公约，完善服务协议，以显著方式提示深度合成服务技术支持者和使用者承担信息安全义务（*第八条*）；依法对深度合成服务使用者进行真实身份信息认证（*第九条*）等。

（2）内容安全义务

深度合成技术易成为散布谣言和虚假信息的工具。相关测试表明，类 ChatGPT 产品可能在原有信息的基础上快速生成大量表面上令人信服但却无实际依据的内容，成为互联网上制造和传播网络谣言的工具。近期，一则有人用 ChatGPT 生成的“杭州市政府 3 月 1 号取消机动车依尾号限行”的假新闻稿在网上流传，由此可见 ChatGPT 生成虚假内容的迷惑性，警方也已介入调查[3]。

[3] https://www.thepaper.cn/newsDetail_forward_21984984

对此，《深度合成规定》明确规定了深度合成服务提供者信息安全和内容审核义务，包括但不限于：不得利用深度合成服务制作、复制、发布、传播虚假新闻信息（*第六条*）；采取技术或者人工方式对深度合成服务使用者的输入数据和合成结果进行审核；建立健全用于识别违法和不良信息的特征库、存留网络日志、建立健全辟谣机制；发现违法和不良信息的，依法采取处置措施，保存有关记录，及时向网信部门和有关主管部门报告，并对相关深度合成服务使用者依法依约采取处置措施。（*第十条*、*第十一条*）

（3）合成内容的标识义务

第一，对于使用深度合成服务生成或编辑的信息，应当采取技术措施添加不影响用户使用的标识，并依照有关规定保存日志信息。

第二，对于具有生成或显著改变信息内容功能的五类服务，在可能导致公众混淆或者误认的情况下，应当对其生成、编辑的信息合理进行显著标识，向公众提示深度合成情况。具有生成或显著改变信息内容功能的五类服务包括：（一）智能对话、智能写作等模拟自然人进行文本的生成或者编辑服务；（二）合成人声、仿声等语音生成或者显著改变个人身份特征的编辑服务；（三）人脸生成、人脸替换、人脸操控、姿态操控等人物图像、视频生成或者显著改变个人身份特征的编辑服务；（四）沉浸式拟真场景等生成或者编辑服务；（五）其他具有生成或者显著改变信息内容功能的服务。

第三．针对其他深度合成服务，服务提供者应当提供显著标识功能，并提示深度合成服务使用者可以进行显著标识。（*第十六条*、*第十七条*）

上述规定进一步发展了《算法规定》第九条中对于“显著标识”的要求。根据《算法规定》，算法推荐服务提供者发现未作显著标识的算法生成合成信息时，应当作出显著标识后，方可继续传输。

就目前的类 ChatGPT 产品而言，其生成的智能对话或文本应用场景较广，且易于复制、传播。其将如何满足上述要求，添加不影响用户使用的标识或显著标识，向公众提示深度合成情况，需待根据具体产品形态进一步进行评估。

（4）训练数据管理

《深度合成规定》规定，服务提供者应加强训练数据管理，采取必要措施保障训练数据安全；训练数据包含个人信息的，还应遵守个人信息保护的有关规定；并强调了在涉及生物识别信息编辑功能室的单独同意要求。（*第十四条*）

海量的训练数据是类 ChatGPT 产品开发和运行的关键。在大量采集的数据过程中，数据的具体类型、数据采集的方式都可能带来合规的风险，例如，个人信息保护风险、隐私权纠纷、数据权属纠纷、竞争纠纷、知识产权纠纷导致的民事、行政乃至刑事责任。尤其是类 ChatGPT 产品会在实时对话过程中不断采集对话内容并用于生成、反馈和训练，这进一步增加了数据采集带来的不确定性。

（5）算法模型安全

算法的透明度与可解释性是算法监管和合规的基本要求。类 ChatGPT 产品所使用的算法模型往往更加复杂，如果算法模型存在偏见、歧视、不公正的因素，可能会对输出的结果产生影响，进而带来道德、伦理风险。

《算法规定》、《深度合成规定》均规定了定期审核、评估、验证生成合成类算法机制机理这一基本要求。《算法规定》的算法评估侧重于规制诱导用户沉迷、过度消费等违反法律法规、伦理道德的算法模型。而《深度合成规定》则更关注深度合成算法的安全性问题，对特殊类型的深度合成服务（包括：①生成或者编辑人脸、人声等生物识别信息；

②生成或者编辑可能涉及国家安全、国家形象、国家利益和社会公共利益的特殊物体、场景等非生物识别信息）提出了安全评估要求，服务提供者应当依法自行或者委托专业机构开展安全评估。（*第十五条*）

（6）备案及安全评估义务

具有舆论属性或者社会动员能力的深度合成服务提供者履行备案和变更、注销备案等义务（*第十九条*）。若涉及开发上线具有舆论属性或者社会动员能力的新产品、新应用、新功能的，还需要按照国家有关规定进行安全评估（*第二十条*）。

（7）设立用户申诉、投诉和举报渠道义务

深度合成服务提供者应当设置便捷的用户申诉和公众投诉、举报入口，公布处理流程和反馈实现，及时受理、处理和反馈处理结果（*第十二条*）。

（三）深度合成服务技术支持者合规义务

深度合成服务技术支持者是指为深度合成服务提供技术支持的组织、个人。技术支持者的合规义务主要包括：（1）数据管理和个人信息保护义务（*第十四条*）；（2）定期审核、评估、验证生成合成类算法机制机理，以及对特殊类型深度合成功能的模型、模板等工具（如编辑人脸、人声等生物识别信息，或可能涉及国家安全、国家形象、国家利益和社会公共利益）的安全评估义务（*第十五条*）；以及（3）算法备案义务（*第十九条*）。

（四）深度合成服务使用者合规义务

深度合成服务使用者是指使用深度合成服务制作、复制、发布、传播信息的组织、个人。随着类 ChatGPT 产品的广泛普及，每个人都可能成为深度合成服务使用者。

深度合成服务使用者的合规义务，除了一般性的信息内容安全义务外，还包含敏感个人信息处理场景下的单独同意要求。《深度合成规定》第十四条第二款规定，深度合成服务提供者和技术支持者提供人脸、人声等生物识别信息编辑功能的，应当提示深度合成服务使用者依法告知被编辑的个人，并取得其单独同意。值得注意的是，相比于《深度合成规定》征求意见稿，正式稿删除了“法律、行政法规另有规定的除外”这一合法性基础的兜底规定，将处理生物识别信息的合法性基础仅限定于单独同意。

（五）应用程序分发平台合规义务

应用程序分发平台应当落实上架审核、日常管理、应急处置等安全管理责任，核验深度合成类应用程序的安全评估、备案等情况；对违反国家有关规定的，应当及时采取不予上架、警示、暂停服务或者下架等处置措施。（*第十三条*）

四、我们的观察

从《深度合成规定》的具体内容来看，部分问题仍待市场实践和监管意见的进一步明确，包括：显著标识的具体方式、安全评估的要求、单独同意的取得和查验、深度合成的鉴别验证加密、内容审核的具体方式、算法模型安全的具体标准等。我们将对这些问题持续保持关注。

同时，我们注意到，《深度合成规定》的出台后，“互联网信息服务算法备案系统”已经发布《<互联网信息服务深度合成管理规定>备案填报指南》，说明了服务提供者和技术支持者开展算法备案的流程和要求。截止目前公布的三批境内互联网信息服务算法备案清单中，已有 10 个生成合成算法，涵盖了问答对话、语音转换、视频生成、人物形成生成等技术类别，其中，智能客服主题的生成合成算法的占比接近一半。

ChatGPT 产品的爆红折射出人工智能和算法技术的快速发展。与此同时，世界各国都开始探索并

逐渐形成各具特色的人工智能和算法规制框架。例如美国颁布了《深度独合成工作任务法案》（Deepfake Task Force Act）、欧盟颁布的《2022 年虚假信息强化行为准则》（2022 Strengthened Code of Practice on Disinformation）。在本系列的后续文章中，我们将继续关注人工智能的伦理治理、人工智能的域外立法发展等问题。

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
王威华　律　师　电话：86 10 8519 1213　邮箱地址：wangweihua@junhe.com