生成式 AI（一）：尝试参透监管的底层逻辑

随着 ChatGPT 等生成式人工智能（“生成式 AI”）应用的迅速走红，人工智能的治理议题由此被提升至前所未有的关注高度。在学习、实践、折返、进阶的往复中，与既往一样，我们始终在探索最为简单质朴的合规最优解，并愿意将此过程与各位共享。让我们随着生成式 AI 的发展一起迭代。

【以上内容请圈入方框中】

针对生成式 AI 的治理，当前的一大争议在于如何处理《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》”) 与现有立法的衔接关系。基于对立法背景及既有规则的解读，我们理解，《生成式人工智能服务办法》与《互联网信息服务深度合成管理规定》(“《深度合成管理规定》”) 之间应为特别法和一般法的关系。又由于《互联网信息服务算法推荐管理规定》(“《算法推荐管理规定》”) 的监管对象包含生成合成类算法，《深度合成管理规定》与《算法推荐管理规定》为特别法和一般法的关系，因此，**我国的生成式 AI 规则体系同时包含《生成式人工智能服务办法》《深度合成管理规定》以及《算法推荐管理规定》。**

（一）深度合成技术包含生成式 AI 技术

首先，从条文本身来看，深度合成技术的法律定义包含生成文本、图像、语音、非语音、视频等内容，可以涵盖生成式 AI 定义中的生成内容，且深度合成定义中也多次采用了“生成”的表述。我们理解，为了应对技术的快速发展，出台在前的《深度合成管理规定》有意拓宽适用范围，从而可以将生成式 AI 纳入其中。

其次，从官方背景文件的解读中，也可得出监管者有意以《深度合成管理规定》作为依据规制生成式 AI 的结论。[1]中央网信办网络管理技术局局长在《深度合成管理规定》的解读文章中明确提及，生成型对抗网络（GAN）、扩散模型（Diffusion Model）、生成预训练变换模型（GPT）等技术为深度合成技术的代表。公安部网安局发布的文章中亦明确提出，ChatGPT 背后的技术为深度合成，应受到《深度合成管理规定》的规制。[2]

最后，从监管内容角度来看，深度合成与生成式 AI 技术都存在输出不当内容、技术滥用等问题，导致监管手段具有一定的相似性，生成式 AI 的监管可以建立在深度合成监管的基础之上。

（二）深度合成与生成式 AI 监管各有使命

我们理解，以 ChatGPT 为代表的生成式 AI 虽可以适用《深度合成管理规定》，但其规制重点与以 Deepfake 为代表的深度合成技术存在较大差异。

---

[1] https://mp.weixin.qq.com/s/wN3Jg4NPBzb7ev-ZaIAZBw

[2] https://mp.weixin.qq.com/s/0Hmwknq6bNFDFyAwRJbJHg?search_click_id=13503168500786901975-16849770313338-1595689541

《深度合成管理规定》假定的监管对象为 Deepfake（“深度伪造”）

2017 年，美国社交网站 Reddit 出现一个名为“deepfakes”的用户，其利用技术将色情女演员的面部替换为知名女明星，并发布了相关图像视频。自此，深度伪造技术逐渐步入大众视野。2019 年，我国大众文娱领域先后出现多起“明星换脸”事件，涉嫌肖像权侵权。随后将关注度推向高潮的为“ZAO”事件。2019 年 8 月，换脸软件“ZAO”上线并迅速走红，在该款应用中，用户可以上传照片并将影视片段中明星的人脸替换为照片中的人脸。“ZAO”的用户协议作出了如下规定：“确保肖像权利人同意授予‘ZAO’及其关联公司全球范围内完全免费、不可撤销、永久、可转授权和可再许可的权利，包括但不限于：人脸照片、图片、视频资料等肖像资料中所含的您或肖像权利人的肖像权，以及利用技术对您或肖像权利人的肖像进行形式改动”，引发了大众对于个人信息保护的担忧，后被网信办约谈整改。

深度伪造的一系列滥用事件引起了中国官方的高度重视。2020 年 12 月，中共中央印发《法治社会建设实施纲要（2020－2025 年）》，明确提出要完善网络法律制度，制定完善对算法推荐、深度伪造等新技术应用的规范管理办法。[3]2022 年 1 月，国家网信办就《深度合成管理规定（征求意见稿）》公开征求意见，并就规则制定的必要性进行了说明，明确提及要贯彻《法治社会建设实施纲要（2020－2025 年）》中的决策部署，以及应对不法分子利用技术制作、复制、发布、传播违法信息，诋毁、贬损他人名誉、荣誉，仿冒他人身份实施诈骗等违法行为。[4]由此可以看出，《深度合成管理规定》的重点规制对象在于深度伪造技术。

立法并未采用深度伪造的概念，原因或有其二：一是深度伪造的概念较为负面，立法旨在规制该项技术的滥用，并不意欲阻碍技术的发展或正当利用，而采用深度伪造的概念容易给人造成先入为主的负面印象；二是将深度伪造定义为监管对象，会导致监管范围过窄。面对技术的迅猛发展，未来很可能会出现其他类似技术（例如当前的生成式 AI），为了克服立法的滞后性问题，避免监管机构在执法时无法可依，《深度合成管理规定》有意规定了更宽的监管范围。

《生成式人工智能服务办法》假定的监管对象为 ChatGPT 代表的大模型应用

2022 年 11 月，美国人工智能研究公司 OpenAI 开发的 ChatGPT 聊天机器人上线，在上线不到 2 个月的时间内，其月活用户突破 1 亿，成为史上用户增长最快的消费者应用。在其爆火的同时，其隐含的信息滥用、虚假信息、用户依赖、系统失控、歧视问题、网络安全等风险引发了学界和业界的担忧。2023 年 3 月 22 日，图灵奖得主 Yoshua Bengio、特斯拉 CEO（Open AI 联合创始人）Elon Musk、苹果联合创始人 Steve Wozniak 等上千名专家联名签署公开信，强调人工智能系统可能给人类社会带来的巨大风险，呼吁暂停开发比 GPT-4 更强大的人工智能系统。[5]2023 年 4 月 10 日，中国支付清算协会发布公告，倡导支付行业从业人员谨慎使用 ChatGPT 等工具。[6]然而，大模型的浪潮已然掀起，百度、阿里巴巴、字节跳动、华为等国内互联网巨头纷纷推出大模型产品，在这样的背景之下，国家网信办就《生成式人工智能服务办法》公开征求意见，意为急用先行，监管以 ChatGPT 为代表的大模型应用。

---

[3] http://www.gov.cn/zhengce/2020-12/07/content_5567791.htm

[4] http://www.cac.gov.cn/2022-01/28/c_1644970458520968.htm

[5] https://futureoflife.org/open-letter/pause-giant-ai-experiments/

[6] http://www.pcac.org.cn/eportal/ui?pageId=598261&articleKey=617041&columnId=595085

（三）引导对人类的良性影响是生成式 AI 监管的底层逻辑

在明确《深度合成管理规定》与《生成式人工智能服务办法》立法背景与重点监管对象的不同之后，分别以 Deepfake 和 ChatGPT 为例对比其典型应用场景，更有助于理解两部规范侧重点之不同。

（1） 系统的自主性

在 Deepfake 的应用场景下，用户通常需要向系统输入事先准备好的素材，经由系统的加工，最终形成内容输出。在这个过程中，用户本身对于输出结果具有一定的预期，**系统实际上是在用户的控制之下完成工作，自主性较低。**因此，从监管的角度来看，**规制的本质在于穿透技术，约束用户滥用技术的行为。**纵观《深度合成管理规定》第二章，第六条率先从正面明确了任何组织和个人不得滥用深度合成服务的义务，其余条款则主要是对深度合成服务提供者施以管理义务，核心即在于阻止用户的滥用行为：

《互联网信息服务深度合成管理规定》 第二章 一般规定		
条款	**主体**	**义务**
第六条	任何组织和个人	不得从事受禁违法活动、传播受禁违法信息
	深度合成服务提供者、使用者	提供互联网新闻信息服务的特殊义务
第七条	深度合成服务提供者	施行管理制度及技术保障措施
第八条		履行平台管理责任并提示信息安全义务
第九条		对用户进行真实身份信息认证
第十条		深度合成内容管理（用户输入数据及合成结果审核、违法和不良信息特征库、违法和不良信息处置措施）
第十一条		建立健全辟谣机制
第十二条		设置申诉、投诉和举报机制
第十三条	应用程序分发平台	上架审核等安全管理责任

相较而言，**在 ChatGPT 的应用场景下**，用户通常无需输入事先准备的素材，只需输入相应指令，系统就能够自行分析并生成输出。在这个过程中，用户对于输出结果并不存在具体的预期，对系统的控制力也较弱。正如欧盟《人工智能法案》草案对生成式 AI 的定义一般，“生成式 AI 是使用基础模型，专门用于以不同程度的自主性生成复杂的文本、图像、音频或视频等内容的人工智能系统”，**系统的自主性是生成式 AI 的重要特征，这也会导致用户更容易依赖系统或受到系统的影响。**根据 OpenAI 的报告，GPT-4 具有产生“幻觉“的倾向，即“产生与特定来源有关的无意义或不真实的内容”，随着模型变得愈加有说服力和可信度，这种倾向的危害会被放大，

导致用户过度依赖的问题。因此，从监管的角度来看，**其需解决的重点问题在于保护用户免受系统的不良影响。**《生成式人工智能服务办法》对提供者施加的防依赖沉迷及指导用户正当使用义务即可印证这一观点。

（2） 系统的输出对象

在 Deepfake 的应用场景之下，用户通常是为了向公众发布合成内容而使用系统，系统的主要输出对象是公众。无论是美国参议院提出的《2018 年恶意伪造禁令法案》，[7]还是欧盟《人工智能法案》草案，[8]其对“deep fake”的定义都包含“以假乱真”这一重要特征。如前所述，直接使用产品的用户对于输出结果存在预期，并不会受到输出内容的误导，**故该场景的监管重点在于确保广大公众不受误导，而在合成内容之上添加标识可以较好地实现这一目标，**所以《深度合成管理规定》有针对性地建立了信息内容标识管理制度。

《互联网信息服务深度合成管理规定》	
隐式标识	**第十六条** 深度合成服务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。
显式标识	**第十七条** 深度合成服务提供者提供以下深度合成服务，可能导致公众混淆或者误认的，应当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： （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。 深度合成服务提供者提供前款规定之外的深度合成服务的，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。

[7] § 1041 (a) (2) the term ‘deep fake’ means an audiovisual record created or altered in a manner that the record would falsely appear to a reasonable observer to be an authentic record of the actual speech or conduct of an individual;

[8] Article 3 (44d)‘deep fake’ means manipulated or synthetic audio, image or video content that would falsely appear to be authentic or truthful, and which features depictions of persons appearing to say or do things they did not say or do, produced using AI techniques, including machine learning and deep learning; .

标识保留	**第十八条** 任何组织和个人不得采用技术手段删除、篡改、隐匿本规定第十六条和第十七条规定的深度合成标识。

在 ChatGPT 的应用场景之下，用户通常是生成内容的主要输出对象，至于生成内容是否会向公众发布，则取决于用户的后续选择。鉴于 ChatGPT 本身可能会输出误导性内容而用户难以察觉，所以**该场景的监管重点首先在于确保用户不受误导，**其次才是公众层面。在公众层面上，如同 Deepfake 输出内容一样，添加标识可以起到较好的预防效果。但在用户层面上，用户之所以会使用 ChatGPT，即在于其对 ChatGPT 存在一定的信赖，相信 ChatGPT 能够输出有价值的内容。并且，用户对于生成内容并不一定具有相应的鉴别能力，即便此时添加标识，也仅能起到提醒作用，并不能真正发挥防止用户被误导的效果。因此，**只有从根源上采取措施，增强内容本身的真实性、准确性、客观性，保障内容合法，才能够达到监管目标。**故而《生成式人工智能服务办法》从训练数据、内容输出和运营管理三个方面入手，对服务提供者苛以多项义务，以防止不当内容的生成。