正文：赵志东蔡

来源：财联社

图片来源：由无界版图AI工具生成

2022年底，美国科技初创公司OpenAI发布的智能聊天工具ChatGPT一夜走红。相比以往一些AI聊天机器人明显的机械感或者复制粘贴感，ChatGPT给出的答案往往逻辑清晰。，切中要害，并能与用户进行自然流畅的对话。出于对AI技术发展到什么程度的好奇，各行各业的人都开始使用ChatGPT进行花式测评，从写代码到高考数学题，从翻译学术论文到写论文。。ChatGPT的完成度真的让人印象深刻。过去只能由人类完成的创造性工作，可以在一定程度上被ChatGPT替代，完成相应的内容生产。

ChatGPT的外观也掀起了人工智能生成内容的热潮。微软已经正式将ChatGPT技术接入必应搜索和Edge浏览器。。百度(中国)有限公司也表示将在完成项目内测后向公众开放&quot厄尼博特&quot类似于今年3月的ChatGPT。

用于ChatGPT&#039s&quot超能力&quot很多人在为技术的发展感到兴奋的同时，也对人工智能内容的产生及其背后的道德伦理、社会责任和法律风险感到担忧。本文将从算法、数据、内容三个方面详细阐述ChatGPT背后的法律风险和合规建议。

1。ChatGPT产生虚假信息，答案质量谁负责？

chatGPT的技术原理是在算法和数据的基础上，利用自然语言处理生成相应的答案。明确地ChatGPT使用基于人类偏好的机器训练模型(来自人类反馈的强化学习)，技术人员提前向ChatGPT输入海量语言数据。，并根据人工标注的优质答案，为ChatGPT生成的答案建立评分和奖励模型，引导训练ChatGPT模型不断修正和优化输出结果，使输出答案向高分方向改进。最后，ChatGPT可以学习人类语言的特点和结构，用自然语言实现人机交流。

目前认为ChatGPT能够处理各种相对复杂的语言任务。，包括自动文本生成、自动问答、连续对话延续上下文等。并已尝试用于新闻、营销、客服等行业。但人工智能内容的制作并不能保证其内容的准确性，ChatGPT也在用户协议中做了免责声明。。与ChatGPT&#039的一些&quot纠正废话&quot，其造成的危害&quot严肃的废话&quot让全世界的监管者都非常警惕。

今年1月，OpenAI首席技术官MiraMurati说。ChatGPT可由&quot危险元素&quot捏造事实，这是目前大型语言模型面临的共同挑战。

一方面，在法律、医学等专业领域，用户需要具备一定的专业知识才能分辨ChatGPT生成内容的真假。。如果用户依靠不准确的答案做出判断和决策，可能会危及用户的人身和财产安全。另一方面，也不排除一些别有用心的不法分子&quot火车&quot他们通过向ChatGPT输入虚假或误导性信息来制造虚假信息。严重时可能导致影响政治舆论或政治生态、煽动暴力犯罪、损害公共利益等严重后果。中国&#039；s《互联网信息服务深度合成管理规定》也是为了应对深度合成技术对法律、社会秩序和公共安全的影响。。根据《互联网信息服务深度合成管理规定》给出的定义，深度合成技术是指生成深度学习、虚拟现实等合成算法，使文本、图像、音频、视频、虚拟场景等网络信息。，包括文本生成、文本样式转换、问答对话等技术来生成或编辑文本内容。ChatGPT等人工智能内容生产的产品，显然应该受到这部法律的规制。

按照《互联网信息服务深度合成管理规定》的要求。深度合成服务商要落实信息安全主体责任，建立健全用户注册、算法机制审查、科技伦理审查、信息发布审查、数据安全、个人信息保护、反电信网络诈骗、应急响应等管理制度。

首先，，深度合成服务提供者应当以显著方式提示深度合成服务的技术支持者和使用者承担信息安全义务。对于利用其服务产生的可能导致公众混淆或者误解的信息内容，应当在信息内容的合理位置和区域予以显著标注。，提醒公众该内容为深度合成，提醒用户在生成疑似违法信息时存在虚假信息的安全风险。

其次，深度合成服务商要加强对深度合成内容的管理，建立健全识别违法和不良信息的特征库。，采用技术或人工方式审核深度合成服务用户的输入数据和合成结果。发现用户利用深度合成服务制作、复制、发布、传播虚假信息的，应当及时采取措施辟谣，保存相关记录，并向网信部门和相关主管部门举报。。

同时，对发布违法和不良信息的相关深度综合服务用户，依法依规采取警告、限制功能、暂停服务、关闭账户等措施。

二、ChatGPT生成歧视性文本。如何监管算法的盲盒？

除了《互联网信息服务深度合成管理规定》，ChatGPT等人工智能内容制作产品也受到《互联网信息服务算法推荐管理规定》的监管。《互联网信息服务算法推荐管理规定》中提到的算法推荐技术，包括使用生成合成类算法技术向用户提供信息。这一规定的要点是防止算法歧视或算法偏差，产生不符合主流价值取向的负面有害信息或低俗劣质信息。

根据《互联网信息服务算法推荐管理规定》第七条，算法推荐服务提供者应当建立健全算法机制的机制审查、科学伦理审查、用户注册、信息发布审查、数据安全和个人信息保护、反电信网络诈骗、安全评估与监测、安全事件应急响应等管理制度和技术措施。

具体而言，算法服务提供者应定期对算法的机制、模型、数据和应用结果进行审查、评估和验证，以确保算法模型不违反伦理。在此基础上，算法服务提供者还应当以显著方式告知用户其算法推荐服务。，并以适当方式宣传算法推荐服务的基本原理、目的、意图和主要运行机制。

另外，如果是具有舆论属性或者社会动员能力的产品，算法推荐服务商填写服务商&#039；的名称、服务形式、应用领域、算法类型、算法自评报告、拟公示的内容等信息，并在提供服务之日起十个工作日内通过互联网信息服务算法备案系统，履行备案手续。

以ChatGPT为例。虽然OpenAI在官网宣称，但ChatGPT已经通过了算法设置和训练，一定程度上可以拒绝用户的不合理请求。比如生成可能含有种族歧视或性别歧视、暴力、血腥、色情等违反法律和公序良俗的内容。但事实上，非法信息传播的风险依然存在，其法律责任和舆论可能会对公司产生负面影响&#039；的形象和商誉。

三年前在南韩非常流行的人工智能聊天机器人李鲁大，因为一些用户在对话和互动时故意输入肮脏和暴力的词语，李鲁大收到了这些不符合主流价值观的内容。，开始输出一些涉嫌性别歧视、种族歧视、歧视弱势群体的内容。结果产品广受诟病，上线不到一个月服务就中断了。

三、用户&#039；的输入信息用于训练ChatGPT。个人信息处理合规的要点是什么？

chatGPT的前训练、后迭代、强化学习都需要海量数据，最新开放版本GPT-3.5的前训练数据已经达到45TB。，包括北京致远人工智能研究所构建的中文语料库WuDaoCorpora中约3TB的中文语料库。第14条明确规定深度合成服务商和技术支持者要加强训练数据的管理，采取必要措施保证训练数据的安全；培训数据包含个人信息的，应当符合个人信息保护的相关规定。

中国&#039；s《个人信息保护法》明确了处理个人信息应当遵循的公开性、诚实性和最低限度原则，即个人信息处理者应当公开个人信息处理的规则，明确说明处理的目的、方式和范围，不得过度收集个人信息。，不得以误导、欺诈、胁迫等方式处理个人信息。以ChatGPT为例，OpenAI如果收集或处理个人信息用于模型训

练，也需要获取用户&#039；的个人同意。

在OpenAI&#039s回复关于ChatGPT使用的常见问题，OpenAI明确承认用户与ChatGPT的对话数据将用于ChatGPT的迭代模型训练。而且，OpenAI特别提醒，OpenAI暂时无法从输入历史中删除特定内容；如果要删除数据，必须注销这个OpenAI账号，这个账号的所有相关数据都会一起删除，所以用户在使用ChatGPT的时候不要输入敏感信息或者机密内容。

因此，在类似ChatGPT的人工智能内容生产产品的开发和后期升级过程中，应尽可能使用公开的非个人信息，或者对收集的个人信息进行匿名化处理。如果必须使用个人信息，应当明确告知用户并取得其个人同意，以避免因非法使用个人信息而造成的侵权或损害。