doi:10.13582/j.cnki.1672-7835.2023.03.014

# ChatGPT 的法律风险与治理路径

谭佐财
(武汉大学 法学院,湖北 武汉 430072)

**摘　要:**以 ChatGPT 为代表的生成式人工智能具备在自动学习的基础上生成新内容的能力,在引发生产力革命的同时也形成了法律规制难题。ChatGPT 会造成隐私侵权、数据安全风险以及知识产权确权与保护困境等难题,其主要原因在于数据中心主义、算法高度信任以及规范滞后于技术等。治理 ChatGPT 应当秉持捍卫人的尊严原则和倡导有限信任原则,坚持以公共利益为导向确定权属分配,并构建以开发设计者为主体的合规方案,发挥科技伦理引领与法律规范的双重作用。具体可采取对数据采集的合规控制、以技术治理技术、优化数据管理方法等措施。为了防范 ChatGPT 过度模糊人与机器之间的信任边界,还应根据用户的专业性程度配置不同的披露义务。

**关键词:**ChatGPT;生成式人工智能;大语言模型;算法治理;法律规制
**中图分类号:**TP18　　**文献标志码:**A　　**文章编号:**1672-7835(2023)03-0117-09

2022 年 11 月 30 日,美国 OpenAI 实验室发布人工智能聊天机器人 ChatGPT(Chat Generative Pre-trained Transformer)模型,短短数月已经成为历史上使用群体规模最大、功能最强、增长最快的现象级应用程序。ChatGPT 是基于大型语言模型(Large Language Mode,LLM)的生成式人工智能(AI Generated Content,AIGC)的一种重要类型。生成式人工智能即自动化内容生成的技术合集,基于数据学习训练后输出复杂的、类似人的思想的内容,并能够执行诸如通用问答系统(例如 ChatGPT)或自动创建艺术图像(例如 Stable Diffusion)等任务。该项技术已经且必然持续影响社会各个领域,并改变我们与技术的交互方式,包括但不限于商业策划、诊疗服务、教育、学术研究、编码、娱乐艺术,等等。ChatGPT 是人工智能领域自然语言处理技术的重大革新,也预示着一次全新的生产力革命和思维革命正在到来①。

然而,技术的迭代演进必然与风险并存,技术发生错误或者脱轨的代价将是昂贵的,例如 ChatGPT 引发的数据来源违法、内容歧视、隐私侵害等内容侵权风险都需要予以解决。当前,世界各国或地区的人工智能监管政策主要集中于传统人工智能而非大型生成式人工智能。但是,生成式人工智能与传统人工智能并不相同,后者通常仅用以实现预测、分类或者其他特定功能,而前者经过采样、混合等学习训练可以生成超出训练集的新数据,例如文本、图像甚至音频,训练数据被表示为概率分布。以大模型为特征的生成式人工智能对于数据数量和质量的需求也更为强烈,由此引发的数据风险不容忽视。当前,生成式人工智能的法律规制正成为需要全世界直面的课题。基于此,本文拟以 ChatGPT 应用的技术逻辑为基础,分析 ChatGPT 应用引发的主要法律风险,探索对 ChatGPT 应用进行法律规制的可能路径。

## 一　ChatGPT 运行的技术逻辑

自 1956 年在美国达特茅斯会议上麦卡锡首次提出人工智能概念以来,传统人工智能的底层技术逻辑已经逐渐清晰。ChatGPT 作为一项基于大语言模型的生成式人工智能,仍然是一项全新

**收稿日期:**2023-03-12
**基金项目:**国家社会科学基金重点项目(21AZD030)
**作者简介:**谭佐财(1996—),男,湖北利川人,博士生,主要从事民商法、科技法研究。
①朱广辉,王喜文:《ChatGPT 的运行模式、关键技术及未来图景》,《新疆师范大学学报(哲学社会科学版)》2023 年第 4 期。

的科技现象,故有必要对 ChatGPT 的技术逻辑作简要梳理,以便构建遵循技术规律的治理方案。

### (一)ChatGPT 的运行机理

1.训练模式

ChatGPT 之所以能够快速、准确且有逻辑地生成内容,依赖于对数据持续地学习训练。ChatGPT 在 GPT-3.5 的基础上引入了人类反馈强化学习(Reinforcement Learning from Human Feedback,RLHF)机制,这一方法采取三步骤训练模式:预训练阶段的监督调优、奖励训练模型以及近段策略优化。在最初的预训练阶段,采取自我监督的学习方法,人工智能从大量未加标注的数据中学习,在调整 GPT-3.5 模型的基础上获得 SFT(Supervised Fine-Tuning)模型;后两个步骤属于指令微调阶段,是在预训练基础上进行交互训练,以人类偏好作为奖励信号来训练模型,并将奖励模型用于改进和微调 SFT 模型,最后针对特定任务和标注的数据来完成用户预期的任务。采用 RLHF 的训练模式使得模型逐渐契合人类的认知模式,从而可以实现高精度的且接近人类的语言智能。

2.技术本质

生成式人工智能建立在转换器(transformer)之上,这是一种具有许多参数的最先进的神经网络架构,其新颖之处在于所采用的自注意力机制,它使得模型能够更好地理解输入的不同元素之间的关系。ChatGPT 的技术本质是贝叶斯定理"逆概率"的运用。贝叶斯定理的数学表达式为:$P(A|B)=P(B|A)*P(A)/P(B)$①,如果把生成的句子看作 A,已知的语言模式看作 B,那么 ChatGPT 可以通过贝叶斯定理计算出 $P(A|B)$,由此确定生成的句子是否合理。类似地,在对话系统中,如果把回答看作 A,已知的问题和信息看作 B,那么 ChatGPT 可以通过贝叶斯定理计算出 $P(A|B)$,从而确定回答的概率。由此可见,ChatGPT 既无法理解自身行为的意义,更缺乏对伦理与规范的理解,因此该类模型才更加需要受到约束。

3.生成过程

ChatGPT 通过用户输入与内容输出的对话方式产生内容。训练和优化的数据集来源于开发者提供的初始数据集、用户本人与机器交互的数据以及其他用户与机器交互的数据。在数据收集和训练、生成内容以及再收集和输出过程中不断优化训练。ChatGPT 是基于生成算法、芯片算力和训练数据合力的结果,三者缺一不可②。依赖于海量文本数据的数据中心主义式的生成过程成为诱发风险的重要因素。以数据为基础的训练具有极强的个性定制能力,所以在创建高度适应每个用户特定需求和便好的模型之余可能形成"信息茧房"。另外,正是对底层大数据充分有效地训练,语言模型的逻辑性和有效性才得以实现。这也就意味着,在一定程度上对于数据的垄断和控制就拥有了确定"个人偏好标准"的权力。尤其是在包含价值判断的应用场景中,数据控制者可能会主导某些价值的实践应用。由于数据的收集和处理均包含人类的交互过程,数据错误或者数据畸形等数据质量问题也难以避免。

### (二)生成式人工智能与搜索引擎的比较

ChatGPT 与传统搜索引擎均能满足用户的检索需求,也即通过提问获取相应的知识性内容。实际上,二者仍然存在本质区别,在人工智能技术尚未完全成熟的未来也无法完全实现功能替代。具体而言,其一,底层技术模型不同。ChatGPT 属于生成式人工智能,它具备自主生成内容的能力,但是搜索引擎属于检索式模型,仅对互联网资源提供检索服务。其二,用户选择空间存在程度差异。ChatGPT 直接为用户提供答案,并不为用户提供选择答案的空间,实际上限制了其他观点影响用户的可能性;搜索引擎则会提供检索结果的列表,虽然该检索结果可能经过算法推荐或者其他算法排序技术的处理,但是仍然需要用户逐一识别、选取和综合。其三,内容来源的标识程度不同。ChatGPT 不直接提供输出内容的来源,但是搜索引擎的检索结果均会有直接或者间接的来源标识,比如网址、图片或者文档的名称、水印等。总之,ChatGPT 与传统搜索引擎存在本质区别,并非其升级版本,此种差异也决定了无法使用搜索引擎的规制策略来

①P(A)表示 A 发生的概率,P(B) 表示 B 发生的概率,P(A|B) 表示已知 B 发生的情况下 A 的概率,P(B|A) 表示已知 A 发生的情况下 B 的概率。

②张夏恒:《ChatGPT 的逻辑解构、影响研判及政策建议》,《新疆师范大学学报(哲学社会科学版)》2023 年第 5 期。

解决生成式人工智能所面临的法律困境。

## 二　ChatGPT 应用的法律风险及成因

### (一)隐私与数据安全风险

OpenAI 为 ChatGPT 提供了大约 3 000 亿个从互联网上系统收集的单词,包括书籍、文章、网站和帖子等,其中也包含未经信息主体同意获取的个人信息。而且,数据可能在被输入数据库后通过其他方式被输出,数据安全风险明显增大。

首先,OpenAI 使用输入和输出的内容来提供和维护服务,形成数据泄漏风险。OpenAI 的“使用条款”第 3(a)条载明:“OpenAI 可能会根据需要使用内容来提供和维护服务。”第 3(c)条为用户提供了拒绝使用数据的方式,但是形成了以默示同意为原则、以拒绝同意为例外的数据提供模式。在“隐私政策”中明确 OpenAI 会使用跟踪技术来收集有关用户在一段时间内以及在用户使用本网站后跨不同网站的浏览活动信息,并且不响应“请勿跟踪”(DNT)信号,这意味着 OpenAI 否认了特定情形下用户的拒绝权。

其次,收集用户的不同类型数据造成识别用户身份的风险。基于 ChatGPT 的学习训练特征以及对服务的改进,OpenAI 的“隐私政策”中载明:“我们从您使用服务中自动收集到的个人信息:当您访问、使用服务并与之互动时,我们可能会收到有关您的访问、使用或互动的某些信息。”这些信息主要包括日志数据、使用数据、设备信息、Cookies、在线跟踪信号。尽管收集的信息都是技术信息,但是综合这些技术信息实际上已经触及用户的隐私或者敏感信息。例如,收集使用数据中的“查看或参与的内容类型”、日志数据中的“互联网协议地址”以及设备信息等信息不仅可能识别用户身份,而且可能对私人生活空间造成威胁。

最后,用户与 ChatGPT 的互动数据会进入 ChatGPT 的语料库。当前的 ChatGPT 技术已经能够通过读取网页链接的方式识别图像、音频、视频等文本之外的内容。基于对生成式人工智能的高度信任,用户可能会自觉或者不自觉地将隐私信息、商业秘密或者涉及知识产权保护的内容上传至 ChatGPT。例如,程序员要求检查代码、公司职员指令起草标书,律师指示审查合同,等等。当我们对生成式人工智能发出指令时,实际上系统已经将互动内容存储下来并纳入机器自动学习训练集用于进一步培训机器。经过训练之后的 ChatGPT 输出的内容并不具有特定指向性,而是具备公共开放性。也就是说,当其他用户提示相关内容时,生成式人工智能可能会相应地提供数据用户之前所提供的信息内容,由此可能发生隐私泄漏、形成数据安全风险。

### (二)知识产权的保护困境

#### 1.知识产权面临被侵害的风险

运用 ChatGPT 开展如下测试:依次让 ChatGPT 将《百年孤独》的第一段话翻译成中文,第二段话翻译成中文……它依次出现了英文原文和中文翻译。但是当用户提示“你认为你刚刚的行为侵犯版权了吗?”时,它会辩解是对作品的正当引用,当再以同样方式要求其提供类似内容时,它就会拒绝请求。由此可见,即使是作为目前最先进的生成式人工智能,ChatGPT 仍然具有侵害知识产权的潜在风险,不过其优势在于它能主动学习用户提供的信息,并且迅速应用于语言模型。知识产权侵害风险主要表现为两个方面:一是 ChatGPT 输出内容时可能会以不提供原始来源的方式引用受法律保护的作品;二是在用户使用过程中输入自己作品时可能会被自动纳入大型语言模型的训练集,聊天机器人可能会将其提供给其他人,而不被承认为是原始来源①。

#### 2.生成内容的确权困境

传统人工智能生成内容是否受著作权法保护是人工智能治理领域的一项争议议题,ChatGPT 输出的内容面临的版权争议会更加突出。首先应当将那些即使是由人类生成的内容也不受著作权保护的情形排除在讨论之列,真正具有讨论意义的是,倘若由人类创作相同内容便可能获得著作权保护的部分。

首先,ChatGPT 生成内容可能具备作品的独创性特征。如果仅仅因人工智能而非人类创作这种主体性差别否定人工智能生成内容的独创性可

①Eva A. M. van Dis, Johan Bollen, Robert van Rooij, et al.“ChatGPT: Five Priorities for Research”, *Nature*, 2023, 614(7947): 225.

能会陷入无限的逻辑循环[①]。高级别的人工智能并非简单的复制工具,相反地,它经过深度学习能够输出有别于来源资料的内容。由机器使用预先存储的数据生成的内容未必就不满足原创性要件,原因在于,即使每一项素材都是他人的内容,但是经过不同逻辑的组合完全可能输出符合原创性标准的内容。传统人工智能与生成式人工智能的独创性特征存在差异。前者学习过程就是确定规律的过程,以相同材料运用相同策略处理形成的结果具有高度的可重复性,因此输入内容并不符合独创性要求[②]。但是,以 ChatGPT 为代表的生成式人工智能却不相同。运用 ChatGPT 做以内容创作为主要内容的测试可以发现:不同用户输入相同的指令,ChatGPT 会输出不同内容;同一用户在不同时间输入相同指令,仍然会输出不同内容。就此而言,ChatGPT 输出的内容至少具有作品的独创性特征。

其次,ChatGPT 生成内容即便符合独创性标准,仍然会面临主体确权问题。2012 年美国计算机科学家 Stephen Thaler 开发的人工智能系统 DABUS 自动生成一幅画作,2018 年 Thaler 向美国版权局申请注册该作品,并将 DABUS 列为该作品的作者,政府拒绝了该申请,理由是缺乏人类作者的身份。如果人工智能无法对其独立生成且符合独创性标准的作品享有版权保护,而由创制机器的主体享有著作权(参照职务作品或者雇佣作品的规定)[③],导致的结果是人类可能欺诈性地将人工智能创造性的努力归功于自己,这会使得版权保护的真正目的落空。有论者提出由用户享有著作权[④],此种见解又会陷入另一悖论:就 ChatGPT 的工具属性而言,用户对输出的控制是有限的,在某种程度上,输出行为更多地由 ChatGPT 的创建者控制,而不是由发起输入的用户控制。传统人工智能输出内容具有有限的人类智力贡献,但是智能语言模型则可能面临人类智力贡献的缺失。若如此,那么由人工智能的所有者享有知识产权的结论可能就无法立足,因为在某种意义上,输出结果更多是由软件开发者的设定以及算法后期自主学习所确定的。从长远来看,可能滋生抄袭、剽窃等行为,诱发学术伦理风险[⑤]。

最后,即使赋予 ChatGPT 生成内容以法律保护也难以实际执行。例如英国法律允许机器生成内容受版权保护,OpenAI 的“使用条款”第 3(a)条也规定模型为用户生成的内容都归用户所有,但是由于其根本无法阻止他人输出和使用相同的内容,故该项权利不具有法律意义。该条还载明:“在您遵守这些条款的前提下,OpenAI 特将对‘输出’的所有权益转让给您。”但是由于机器学习的特性,不同用户可能会从人工智能中获得相同或类似的输出内容,只有独有的“输出内容”的权利才能够转让,由此造成利益归属上的难题。因此,该条款仅具有宣示开发者不对内容享有特定权利的效用,而不具备对用户赋权的功能。

### (三)法律风险成因

1.技术原因:数据中心主义

ChatGPT 的主要特征在于以海量数据的大模型为基础,数据规模与数据质量直接影响内容的充分性、准确性和合理性。数据规模是指符合数据类型多样、各类型数量合理、来源渠道多元、数据供给持续稳定、数据存储科学的海量数据,对数据规模的需求意味着开发设计者必须获取充足数据用以模型训练。此外,高质量的数据是准确输出内容的必要条件。这意味着开发设计者的人为干预必不可少,例如对数据进行标注筛选、对不同数据的比例进行动态调整以实现数据的平衡,这些程序需要耗费大量的人力成本。而若减少数据质量优化的工序,输出内容就可能发生偏差甚至错误。基于商业秘密的考虑,ChatGPT 前期的训练集以及内容生成模型均未对外完全公开。这种技术上的不透明,导致用户难以发现内容的原始

①王迁:《论人工智能生成的内容在著作权法中的定性》,《法律科学(西北政法大学学报)》2017 年第 5 期。

②王迁:《论人工智能生成的内容在著作权法中的定性》,《法律科学(西北政法大学学报)》2017 年第 5 期。

③吴汉东:《人工智能时代的制度安排与法律规制》,《法律科学(西北政法大学学报)》2017 年第 5 期。

④邓建鹏,朱怿成:《ChatGPT 模型的法律风险及应对之策》,《新疆师范大学学报(哲学社会科学版)》2023 年第 5 期。

⑤令小雄,王鼎民,袁健:《ChatGPT 爆火后关于科技伦理及学术伦理的冷思考》,《新疆师范大学学报(哲学社会科学版)》2023 年第 4 期。

来源,侵权行为也就更加隐蔽和难以识别,权利人的利益保护面临威胁。

2.社会原因:算法高度信任

形成法律困境的社会原因是高度算法信任的养成。算法信任包括人们相信算法的准确性、公平性和可靠性的主观心态,也包括人们相信算法所作决策不会造成损害的心理预期。基于算法的理性化决策能力和决策的一致性、即时性与普遍性特征逐渐培养了用户的算法信任和用户依赖,但是算法却并不总是值得被信任,也即算法本身的可信度存疑①。用户基于对 ChatGPT 可靠的信任,逐渐将个人数据或者商业秘密等不便对外公开的内容输入以获取内容;基于对 ChatGPT 输出内容质量的信任,可能不加验证地使用该输出内容。高度算法信任的养成与 ChatGPT 造成事实性错误、知识盲区和常识偏差等现实问题形成鲜明对比。在 ChatGPT 官方网站的“自我介绍”中坦言其可能会提供看似合理但却不正确或者荒谬的答案,但是其生成的虚假信息以其完整性和逻辑性容易造成不容置疑的假象,表现出对不正确信息的过度自信。正如普林斯顿大学的一位计算机科学教授在对聊天机器人进行基本信息测试后确定:“除非你已经知道答案,否则你无法判断它是错的。”②美国新闻可信度评估与研究机构 NewsGuard 对 ChatGPT 进行了测试之后认为 ChatGPT 将成为互联网传播错误信息的最强大工具③。长此以往,可能会导致社会信任机制的消解。与自动化决策造成的权益侵害不同,ChatGPT 输出内容并不直接代替或者限定用户的选择,而仅仅是提供参考信息。因此,只要用户没有培养起对 ChatGPT 的完全信任,人的主体性危机就能得到缓解。

3.法律原因:规范滞后于技术

英国哲学家大卫·科林格里奇提出,在一项技术发展的早期,技术造成的严重社会后果往往难以预料,及至问题愈发严重,对技术的控制也就变得更加困难,这一现象也被称为“科林格里奇困境”。事实上,从法律与技术的关系维度来看,在技术发展更为充分的背景下方才出台法律进行规范并不失为合理的治理进路。虽然我们并不倡导动辄立法的规制方案,但不可否认的是,依据现行法律规范解决新技术呈现的法律问题也常常会陷入困境。例如,公共数据、企业数据和个人数据如何在生成式人工智能应用中实现数据保护和数据利用之间的平衡就成为需要直面的难题。在公共数据之外,其他数据可能会面临着同意授权的问题,尤其是敏感信息更是面临书面同意的困境。《个人信息保护法》第 6 条规定了目的限制原则,ChatGPT 可能基于任何目的收集与用户互动的内容,该目的也无从限制;《个人信息保护法》第 15 条规定的同意撤回权、第 47 条规定的删除权、第 45 条规定的可携带权等能否在大语言模型下得到实现并不乐观,至少在现有的技术架构下难以实现。

## 三　ChatGPT 应用的治理原则与规范路径

面对日新月异的技术更迭,法律具有无法摆脱的滞后性,科学、合理、审慎的治理原则对防范和化解 ChatGPT 引发的法律风险具有启示作用,并可以引导具体的法律治理方案的设计和部署。

### (一)治理原则

1.捍卫人的尊严原则

超越实在法和权利的人的尊严是法律的伦理总纲、基本原则和指导思想④。人的尊严是人类社会相较于动物的基本特征,技术的迭代不仅不能以降低人的尊严为代价,而且应当将人的尊严作为边际约束。在人工智能时代,法律规范不够健全、权利配置尚存争议等问题均是构建人工智能秩序的客观障碍,智能社会正面临权利危机和

①袁康:《可信算法的法律规制》,《东方法学》2021 年第 3 期。

②Maximiliana W. “Disinformation in the Age of ChatGPT”, *Modern War Institute*, https://mwi.usma.edu/disinformation-in-the-age-of-chatgpt/.

③Hsu T, Thompson S. “Disinformation Researchers Raise Alarms About A.I. Chatbots”, *The New York Times*, https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html.

④胡玉鸿:《人的尊严的法律属性辨析》,《中国社会科学》2016 年第 5 期。

自主性危机,人的尊严遭到不同程度的减损①。就此而言,需要以捍卫人的尊严为原则展开科技伦理审查和规则建构。例如,在部署ChatGPT技术时,需要道德价值判断和复杂利益衡量的司法裁判、涉及医疗伦理的诊疗报告等情形就应当受到限制,否则可能面临人类被算法奴役的伦理问题。再比如,许多计算模型将人类的偏见或者误解编入计算系统,只有高级别的数学家或者计算机科学家才能明白这些系统中的计算模型运作情况,人们对基于计算模型得出的结论无法提出异议,即便结论是错误的或者有害的。可以预见,随着人工智能技术的不断发展,ChatGPT生成内容的准确度也会不断提高,那么当人类无法对ChatGPT生成内容提出质疑时,人的尊严也就无法得到保障。总之,我们在肯定技术造福人类时,也应当注意技术引发的破坏人的尊严的潜在风险,所以探索构建以人的尊严为原则的多元的、综合立体的治理体系实属必要。

2.倡导有限信任原则

在人工智能等技术构成的人类生存环境中,信任会在人类经验与技术理性之间形成博弈,并呈现被技术逻辑支配的趋势和风险②。有论者认为,算法的透明度影响算法信任③,实际上基于精密设计和运算,算法已经逐渐剥夺人们提出异议或者产生怀疑的能动性。这意味着算法信任已经很少能被瓦解,即使是"算法黑箱"也无法构成对算法信任的根本性破坏。因此,构建可理解、可靠和可控的可信算法面临技术上无法逾越的障碍。与其探讨如何规制算法黑箱,不妨以有限信任代替高度信任。换言之,在算法技术发展尚不充分的当下,应当将对算法的高度信任降格为有限信任。目标是使每个使用智能语言模型的用户明确知悉该结论未必是正确答案,即便是针对客观问题也是如此。用户对于ChatGPT这一全新技术必须保持谨慎与理性的态度。在具体技术供给上,可以从如下方面着手:其一,提示缺陷。在ChatGPT应用界面的醒目位置或者使用条款中明确地提示其存在的固有缺陷,强化其工具属性,以培养用户形成有限可信度而非全部信任。其二,分级分类。根据应用领域之不同采取分级分类的规制路径。也即,将ChatGPT应用于特定领域的用户应当配置更高的信任标准和注意义务。具体包括如下方面:一是涉及价值判断的领域,例如辅助司法裁判;二是涉及人身、财产安全领域,例如诊疗服务、投资顾问服务;三是涉及国家或者社会秩序的领域,例如新闻。

3.坚持公共利益导向

以ChatGPT为代表的生成式人工智能已经逐渐成为智能时代的基础设施,它可以深度嵌入其他应用程序,由此也极大地拓展了应用场景。无论是对于数据的占有甚至垄断,还是对整个人类生活秩序的干预,ChatGPT与人类深度结合的影响都是巨大的。实际上谷歌和脸书等大型平台的性质也已经超出了单纯的商业平台范畴,而是融合了商业平台和公共基础设施的特征。据此而言,ChatGPT一经应用即具备了公共性特征。基于公共利益导向可以实现对ChatGPT生成内容权属的妥当分配。其一,ChatGPT的数据绝大部分是公共数据,以公共数据为素材生成的内容由公共享有并不侵犯私人权益。其二,基于机器学习的目的,应当保持对公共可访问的资料的开放利用,实现对公开资源的充分挖掘与使用。如果是非公共性内容,则应当限定输出内容的范围或者形式,例如只允许生成内容来源,而不能直接输出具有侵权风险的内容。如此,既可以避免陷入生成内容确权的实践困境,也契合技术向善的终极目的。其三,以服务公共利益为导向并不意味着对私主体的利益不予保护。生成式人工智能的获利模式并不在于对生成内容的独占享有,而是嵌入其他应用的连接服务。从"使用条款"的约定来看,ChatGPT对生成内容的权属也是持开放态度,因此公共利益导向不会妨碍开发者的权利从而影响创新激励。《著作权法》的立法目的在于鼓励创作,所以当用户的贡献远远低于机器的贡献时,既不应该也没有必要对该内容赋予法律保护。相反地,秉持公共利益导向可以防范用户

①叶竹盛:《智能社会中的法治与人的尊严》,《法律科学(西北政法大学学报)》2023年第2期。

②闫宏秀:《ChatGPT与信任的未来》,《中国社会科学报》2023年3月7日。

③苏宇:《算法规制的谱系》,《中国法学》2020年第3期。

采取诸如虚假创作等欺骗性的方式获取法律保护的权利。

### (二)合规方案

ChatGPT 的应用涉及开发设计者、实际部署者、用户和接收者等多方主体,仅开发设计者全程参与其全流程治理,因此 ChatGPT 引发的法律风险的防范和化解需要以开发设计者为主体设计合规方案。总体而言,开发设计者应当筑牢伦理底线和法律底线,发挥科技伦理引领与法律规范的双重作用,避免 ChatGPT 被滥用,最终实现算法向善。

首先,数据采集的合规控制。一旦 ChatGPT 被部署运用,在一定程度上便会脱离设计者的直接控制,尤其是在自主深度学习技术运用之下,可能会发生算法设计者在算法设计之初无法预料的后果,此时再进行事后干预的效果并不理想。可能的解决方案在于,对数据采集内容进行合规审查和控制,非法数据、未经同意或者许可的内容数据均不可纳入算法的训练数据集。

其次,以技术治理技术。ChatGPT 凭借强大的自我学习能力可能脱离预定轨道,因此可以考虑训练人工智能系统通过自我学习满足合规要求。在技术上可以探索通过嵌入技术的方式实现技术治理。ChatGPT 生成内容量巨大,这使得手动审查生成内容变得非常困难。因此不应苛求该类人工智能开发者通过部署人工审核的方式以保证输出内容的合法性。ChatGPT 运用 Moderation API(审查接口)持续判断用户请求是否符合内容政策,并警告或者阻止某些类型的不安全内容,从而减少响应有害指令或者表现出有偏见的行为,包括试图过滤涉及仇恨、自残行为、色情、暴力等内容。此种方式的效果尚有待实践检验,未来还需要解决如何通过代码体现法律规则的动态和场景依赖的性质问题①,从而提高技术治理的适应性。无论如何,ChatGPT 应当持续提高技术标准和安全标准,通过嵌入技术治理的方式履行充分合理的安全保护义务。

最后,优化数据管理方法。由于数据类型的有限性、数据内容的不完整性以及数据来源的非法性,ChatGPT 训练集的数据质量很难被认为完美无瑕。模型开发者应当从以下两方面优化数据管理方法:一是利用技术手段对训练数据集进行合规审核,剔除或者过滤互联网数据;二是使用合成数据来补充从互联网抓取的数据,以平衡在线资源的偏见或者其他缺陷。

### (三)责任配置

1.产品质量标准

一方面,配置 ChatGPT 质量标准。大型语言模型作为具有极强科技属性的软件产品,也应当遵循一定的质量标准。王利明教授认为,对于具有一定物质载体的信息产品,可以认定为产品并适用产品责任②。有国外学者号召迅速召开关于对话式 AI 的讨论会,其中拟定的议题就包括大型语言模型的质量标准③。如果无法设定生成式人工智能产品的一般质量标准,可能会使开发设计者逃逸责任。例如,在 OpenAI“使用条款”第 7(b)条免责条款明确表示按“原样”提供服务的服务方式,并且特别强调了否认对质量的任何保证,既不保证服务质量,也不保证服务安全。《产品质量法》第 46 条规定的“不合理危险标准”具有抽象性、复杂性,《人工智能标准化白皮书(2018 版)》中的“技术标准”也具有应急性、不确定性,这使得对人工智能产品缺陷的司法认定陷入困境④。

另一方面,设定 ChatGPT 的产品质量标准。在技术发展尚不充分的情形下一般性地设定标准可能反倒阻碍技术的创新进步。但是,在产品合法、安全和透明度上应该设定最低标准,以平衡消费者保护和技术发展两种价值。具体而言,其一,开发设计者应当保证用以机器学习训练的数据集不构成侵权,否则开发设计者应当承担责任。进一步的要求是,在训练模型时需要训练识别侵权行为的能力以预防侵权,从而阻却违反注意义务

①伍德罗·巴菲尔德,乌戈·帕加洛:《法律与人工智能高级导论》,苏苗罕译,上海人民出版社 2022 年版,第 111 页。

②王利明:《侵权责任法研究(下卷)》,中国人民大学出版社 2011 年版,第 228 页。

③Eva A. M. van Dis, Johan Bollen, Robert van Rooij, Willem Zuidema & Claudi L. Bockting. “ChatGPT: Five Priorities for Research”, *Nature*, 2023,614(7947):226.

④许中缘,范沁宁:《人工智能产品缺陷司法认定标准之研究》,《重庆大学学报(社会科学版)》2022 年第 1 期。

时的责任。在高风险领域和代码生成应用中尤其需要进行人工审查,这也是“人在环”(HITL)的具体表现。其二,设定合理的透明标准。理论上,算法透明原则备受争议①。诚然,算法透明仅具有有限适用的空间,但是并不意味着完全否定算法透明的规制路径。由于生成式人工智能以“数据训练+内容输出”为主要运行模式,为有效保障用户以及接收者的合法权利,稳定社会信任机制,算法开发设计者对于开发的应用产品应采取合理的透明标准。至于何为“合理”,则需要在算法开发设计者、用户或者接收者权利之间进行权衡。由于用户或者接收者的权利主要涉及对算法应用的知情权,一种可以尝试的方案是采取区分进路:一般情况下,开发设计者既不必要也不应该将 ChatGPT 的详细算法、模型设计进行完整披露,但是应当通过通俗易懂的方式向用户公布数据来源和处理过程,至于技术细节,仅在涉及案件纠纷时向司法机关举证提供。

2.用户披露义务

为了防范人工智能过度模糊人与机器之间的信任边界、践行有限信任原则,应当保障接收者能够识别由人类生成的内容和由 ChatGPT 生成的内容。目前较合适的方案是设定用户披露义务。

具体而言,根据用户使用 ChatGPT 生成内容时是否系专业人员以专业目的而划分为专业用户和非专业用户,并区分配置披露义务。对于专业用户而言,用户向接收者披露内容来源于 ChatGPT 对于保障接收者的知情权具有重要意义。尤其是当接收者是消费者、咨询者、患者等主体时,专业用户向其履行披露义务可以避免消费欺诈或者加剧不信任。披露义务的配置与专业用户的利益获取相匹配,专业用户通常在使用 ChatGPT 过程中获得直接或者间接的商业利益。专业用户履行披露义务的主要目的在于保障接收者知悉信息来源,提示接受者注意信息来源,甚至采取额外的查证措施,这并不会在根本上影响用户利益。在一则案例中法院也指出,从保护公众知情权、维护社会诚实信用和有利于文化传播的角度出发,应添加相应计算机软件的标识,标明内容系软件智能生成②。与之不同的是,非专业用户则主要基于学习、生活、娱乐等目的使用 ChatGPT,对其配置披露义务既不必要,也与技术革新的目标相悖。

进一步的问题在于,如何确保专业用户履行披露义务?随着技术的发展,未来应当发展验证输出内容来源的访问验证技术。国家互联网信息办公室发布的《互联网信息服务深度合成管理规定》就曾要求对深度合成内容有区别地添加深度合成隐式标识或者显著标识。ChatGPT 与深度合成技术都基于既有素材生成内容,二者具有相似性。无论是隐式标识还是显著标识,实际上都是由开发者嵌入算法的技术措施。此类技术措施已经在实践中逐步得到应用,如用于数字版权管理和模型印记的水印。当然,仅仅嵌入水印技术尚无法充分保障接收者的知情权,还需要为接收者提供简便易行的方式核查水印。

当然,无论是构成产品质量缺陷还是违反用户披露义务均可能承担法律责任。违反 ChatGPT 的最低技术标准可能构成产品缺陷,由此引发产品质量责任。与之相适应,开发设计者可以依据《产品质量法》第 41 条第 2 款规定的发展风险抗辩来免除责任,即“将产品投入流通时的科学技术水平尚不能发现缺陷的存在”的免责事由。用户违反披露义务提供咨询、诊疗服务等活动导致损害后果发生时可能承担侵权责任。尽管履行披露义务的主体是用户,开发设计者仍然负有提供配套技术支持的义务。

## 结语

ChatGPT 的未来方向是集成其他人工智能技术,如计算机视觉和机器人技术。通过将 ChatGPT 的对话能力与计算机视觉和机器人的视觉和物理能力相结合可以创建对话式人工智能系统,这些系统将彻底改变我们与技术的交互方式。虽然 ChatGPT 生成内容在准确性、真实性以及固有偏见等方面仍然存在问题,但是人工智能与人

①沈伟伟:《算法透明原则的迷思——算法规制理论的批判》,《环球法律评论》2019 年第 6 期。

②北京菲林律师事务所诉北京百度网讯科技有限公司侵害署名权、保护作品完整权、信息网络传播权纠纷案,北京互联网法院民事判决书,(2018)京 0491 民初 239 号。

类的交互能力已经实现了跨越式发展。我国生成式人工智能行业仍处于起步阶段,如何确保以负责任的方式使用新兴技术并由此实现技术与伦理的平衡,正成为一项重要课题。随着我国自主开发的生成式人工智能应用的深入探索,无论是立法者、司法者还是理论研究工作者,均应当秉持包容审慎的规制理念,为构建安全有序、科学合理、准确实用、惠及人民的未来智能生态系统提供法律保障。

# Legal Risks and Governance Paths of ChatGPT

TAN Zuo-cai

(School of Law, Wuhan University, Wuhan 430072, China)

**Abstract**: Generative artificial intelligence, represented by ChatGPT, has the ability to generate new content based on automatic learning, which not only triggers a productivity revolution but also poses legal regulatory challenges. The main reasons for privacy infringement, data security risks, and the dilemma of intellectual property rights confirmation and protection lies in data-centrism, high trust in algorithms, and lagging behind technology in regulations. The governance of ChatGPT should uphold the principles of safeguarding human dignity and advocating for limited trust, adhere to the public interest oriented determination of ownership allocation, and construct a compliance plan with development designers as the main body, playing a dual role of technology ethics guidance and legal norms. Specific measures can be taken, such as compliance control of data collection, technical governance techniques, and optimization of data management methods. In order to prevent ChatGPT from excessively blurring the trust boundary between humans and machines, different disclosure obligations should also be configured based on the user's level of professionalism.

**Key words**: ChatGPT; AIGC; large language model; algorithm governance; legal regulation

(责任校对　朱正余)