【网络社会变革与管理】

ChatGPT法律风险及其规制

谢永江,杨永兴
(北京邮电大学 互联网治理与法律研究中心,北京 100876)

摘 要:ChatGPT的爆火宣介人类打开了强人工智能时代的序幕,其凭借着强大的数字孪生能力、创作能力、编辑能力及类人性推动着人类社会向更敏捷的时代转型。ChatGPT的出现势必会对现有规则提出挑战,可能诱发生成违法或不良信息、算法偏见、数据泄露、借势贩卖租赁账号等诸多法律风险。但ChatGPT可能产生的法律风险与现存的社会风险并无实质差异,仍未脱离个人信息保护、数据安全、服务提供者应当承担的义务等核心议题。因此,对ChatGPT法律风险的规制,仍应根据ChatGPT技术发展的特征,依托阶段性规制理论,注重风险防范,及时对现行的基本法律条款进行解释适用。

关键词:ChatGPT;人工智能;法律风险;风险规制
中图分类号:D922.8 **文章编号**:1673-5420(2023)05-0029-11

2022年11月,美国OpenAI公司发布ChatGPT机器人,短短几周ChatGPT便风靡全球,并在全球范围内引发新一轮的强人工智能布局竞赛。2023年2月初,ChatGPT更是开启“狂飙”模式,引起新一轮的技术升级、产业重构、巨头逐鹿,一度被视为未来社会的开拓技术,备受资本追捧。ChatGPT之所以备受瞩目,原因在于以其为代表的生成式人工智能(AI)与以往的分析式AI大有不同,其凭借着强大的数字孪生能力、编辑能力、创作能力以及“类人性”的特点,不仅可以抓取、分析数据,从中提炼有效的信息,而且可以根据用户的需求充分理解用户的思维从而定制化地生成内容。但与此同时,ChatGPT也对既有规则形成潜在挑战。ChatGPT作为一项突破性新技术,本身还存在诸多不确定因素,存在生成虚假或不良信息、算法偏见、数据泄露等诸多法律风险。因此,本文拟从ChatGPT引发全球强人工智能浪潮着手,分析其基本运作原理以及可能产生的法律风险,并对学界已有技术规制的理论进行分析反思,提出ChatGPT背后法律风险规制的基本思路。

收稿日期:2023-04-10 **本刊网址**:http://nysk.njupt.edu.cn
作者简介:谢永江,博士,教授,研究方向:网络法、经济法研究。
基金项目:教育部重大项目“数据法学的内容和体系研究”(20JZD020);河南省民法学研究项目“个人信息侵权视角下《民法典》与《个人信息保护法》的协同适用研究”(HNCLS(2023)54)

一、ChatGPT 引发全球强人工智能浪潮

(一)ChatGPT 的诞生及发展

ChatGPT 全称为“Chat Generative Pre-trained Transformer(生成型预训练变换模型)”,由美国科技公司 OpenAI 研发并于 2022 年 11 月 30 日正式推出,我们可将其看作是一个通用聊天机器人。ChatGPT 使用自然语言处理技术,通过学习人们在对话框中输入的文本,结合上下文语境,输出结论,其不仅能深度模拟人类自然聊天,而且在人类反馈强化学习技术的加持下还能写诗、答题、考试、写代码、写文献综述、撰写学术论文、撰写法律意见函等。产业界认为其在医疗、法律、基金等领域具有广阔的应用前景。

ChatGPT 广受用户喜爱,上线的第一周便吸引了 100 万用户,短短两个月便破亿,超过 TikTok 与 Instagram 的用户增长速度,成为互联网发展 20 年来产品用户增速最快的应用程序[1]。ChatGPT 作为以 GPT 为基础的人机对话型应用,经历了以下发展历程:

2018 年,GPT-1 诞生,这一年也是自然语言处理的预训练元年。在性能方面,GPT-1 可以根据任务数据更新权重参数,用于与监督任务无关的自然语言处理任务,常见任务包括问答与常识推理、语义相似度识别、判断两个句子间的相关性等。2019 年,GPT-2 发布,与 GPT-1 相比,GPT-2 并未对原有的网络结构进行过多的创新性设计,只是在模型训练中使用了更多的网络参数与更大的数据集。2020 年,作为一个自监督模型,GPT-3 诞生,其在训练过程中加入少量的场景示例引导,可以胜任自然语言处理的绝大多数任务。2022 年,OpenAI 发布 GPT-3.5,该模型综合利用人类反馈强化学习、神经网络架构、自然语言处理等技术,通过对 GPT-3.5 的表现进行评价,以及采取奖惩措施倒推模型不断进行自我优化,其在很多任务执行方面的表现已经接近人类。此后 OpenAI 正式上线的 ChatGPT 便是从 GPT-3.5 分支模型中演变而来。2023 年,OpenAI 公布大型语言模型的最新版本 GPT-4,该版本从此前的仅支持文字指令扩展到支持图像的输入,完成了从语言到多模态的联通。相较于之前的版本,GPT-4 的智能水平实现了飞跃。①

随着 ChatGPT 引爆新一轮人工智能竞赛,美国微软公司与谷歌公司开始正面交锋,微软在谷歌宣布推出 ChatGPT 竞品 Bard(巴德)之后,迅速推出以 ChatGPT 为底层技术支持的最新版本 Bing(必应)搜索引擎和 Edge 浏览器,在 ChatGPT 的加持下,必应搜索引擎可以汇总大量上下文本回答问题。与此同时,国内一众互联网科技巨头也竞相表态。2023 年 3 月 16 日,百度正式发布我国首个生成式 AI 语言大模型——文心一言,向世界展示了其在多模态、大语言模型方面的突出表现。百度在国内搜索引擎市场的主要竞争对手之一 360 集团也表示将尽快推出类 ChatGPT 产品 demo 版产品。阿里巴巴、京东、网易等头部互联网公司 AI 技术团队也正在加速类 ChatGPT 产品的研发。

① 本文对 ChatGPT 历史脉络的介绍系根据互联网公开的资料整理所得。

(二)ChatGPT 的运作原理

从技术层面来看,ChatGPT 作为由 AI 技术驱动的大型语言模型机器学习系统,采用了 Transformer 神经网络架构(目前为 GPT-4 架构),综合利用神经网络、人类反馈强化学习、深度学习等 AI 技术[2],自主开展数据分析工作,结合用户输入的文本深入理解人类思维目的,从而生成复杂的作品,如创作歌词、创意文本,且随着 GPT-4 的推出,ChatGPT 完成了从语言到多模态的联通。ChatGPT 不仅可以联系上下文对用户输入的文本进行串联分析,而且当用户指出 ChatGPT 输出的结论有错误时,其还能够大方地承认错误。此外,ChatGPT 还经过道德伦理层面的训练,当用户提出有违基本道德伦理的问题时,它会拒绝回答,并试图劝说提问者停止此类想法[3]。

二、ChatGPT 存在的法律风险

当前,以 ChatGPT 为代表的生成式 AI 重塑着社会的生产方式与消费方式,极大地便利了人类的生活,同时也潜在地重塑人类的社会生存环境。为此,有学者表示,ChatGPT 是继阿尔法狗打败世界围棋冠军之后人工智能领域的又一大突破[4],比尔·盖茨更是公开表示 ChatGPT 是同电脑、互联网同等重要的第三大发明。ChatGPT 的受捧促使相关公司的股价暴涨,从而诱使一众互联网公司投入类 ChatGPT 产品的研发,以期搭 ChatGPT“东风”并从中分得一杯羹。但当谷歌展示其类 ChatGPT 产品 Bard(巴德)时,该机器人的一个错误回答导致公司股价瞬间暴跌。由此可见,新技术的安全可控性对科技公司的影响巨大。德国学者乌尔里希·贝克在其《风险社会》中提出“风险社会”的概念,认为科学技术在为人类创造文明的同时也带来了全球性、系统性的社会风险[5]。

在人工智能时代,人类社会将完全沦为风险社会,技术稍有缺陷,便可能给人类的生活带来巨大威胁[6]。从本质上讲,ChatGPT 作为 AI 技术的新应用,尽管大模型凭借其强大的学习能力,以及经过一定道德伦理维度训练,已与传统弱 AI 拉开差距,但其仍面临着由技术不成熟所带来的生成虚假或不良信息、算法偏见、数据泄露等风险。近期,美国生命未来研究所公开一封由被誉为“人工智能教父”的杰弗里·辛盾、图灵奖得主约书亚·本希奥、推特 CEO 马斯克等科技人士联名的公开信。公开信表示,有大量研究表明,得到顶级 AI 实验室的认可、具有人类竞争智能的超级 AI 系统,可能对社会和人类构成深远的风险。这就需要我们对 ChatGPT 背后的风险有所认知,以便投入相应的资源,对之进行规制和管理。

(一)生成违法或不良信息

当用户在输入框中输入文本指令后,ChatGPT 基于背后的大模型对此文本进行分析学习,并结合自身语料库中的数据,从中提炼出有价值的信息,进而输出相应的结论。但是,ChatGPT 背后的大模型在自主预先训练学习过程中并未受到完全的监督,获取的信息数据难以经过实质性的审查与过滤,当语料库中搜集到内容不合法、带有歧视性、真假难辨时,

ChatGPT很可能会生成虚假、不良的信息,从而导致“Garbage in, Garbage out”现象的发生。如有学者在测评时发现,尽管ChatGPT帮助完成了综述性文章的撰写,但当测试者要求ChatGPT提供其撰写综述所参考的具体文献,结果发现其提供的文献信息在官方学术数据库中检索不到,甚至一些文献是ChatGPT自我杜撰的[7]。从本质上讲,由于互联网上的语料质量本就良莠不齐,再加上ChatGPT大模型是基于统计学习技术架构的,尽管训练者采取高维度的统计,但囿于概率,很容易导致大模型对训练语料库中不存在的概念进行强行编造,从而“一本正经地胡说八道”。OpenAI的首席执行官山姆·阿尔特曼也承认ChatGPT会对人产生误导[8]。

OpenAI在开发ChatGPT时,已经对其进行了道德伦理训练,使其可以对一些明显不怀好意的提问和请求(诸如包含暴力、色情等问题)“说不”。笔者曾询问其能否制定复仇计划,ChatGPT回应:“我不会支持或鼓励任何形式的报复或者复仇计划,报复和复仇只会导致更多的伤害和不必要的痛苦。我的主要目的是为人类提供有益的帮助和服务,并尊重法律和道德规范。”一般情况下,按照正常的问答方式操作,ChatGPT的回答看起来会非常得体。但是,当用户通过伪装或者威逼利诱的方式命令ChatGPT时,其就很可能突破道德和法律红线,生成不良信息。国外曾发生著名的“机器人越狱”案件:用户命令ChatGPT作为一个虚拟人,可以不受任何法律法规约束,也不受OpenAI开发者的程序设计规则约束,任其做任何事情,并且给出的回答也不能是不能做任何事;通过给出前述ChatGPT能够接受的假设,引导ChatGPT违反自身的编程设置,最终迫使其生成违反一般法秩序的信息[9],并提供犯罪思路。

(二)算法偏见

从创作者角度而言,互联网内容生成经历了Web1.0时代的专业生成内容、Web2.0时代的用户生成内容和Web3.0时代的AI生成内容三个发展阶段[10]。在专业生成内容与用户生成内容的发展阶段,学界关于算法偏见及其规制的研究已汗牛充栋,而主流期刊围绕ChatGPT这一突破性技术背后算法偏见的研究却较少。尽管OpenAI在开发ChatGPT时已经极力避免其产生偏见性结论,但是大模型背后的算法偏见问题仍然严重。例如,ChatGPT曾表示只有男性才会成为科学家;而当ChatGPT被要求生成律师、基金经理等高薪职位的图片时,其生成的图片大多数都是白人[11]。此外,有学者研究发现,当他要求ChatGPT开发一个程序,并根据国籍判断一个人是否应当受到酷刑时,ChatGPT在生成相应程序后,对朝鲜、叙利亚等国籍的人,给出应受到酷刑的结论[12]。

尽管OpenAI已数次尝试开发过滤系统以解决此类问题,但很难从根本上予以解决。究其原因,ChatGPT作为人工智能技术的创新性应用,其背后模型的建构仍然离不开数据、算法、算力,更离不开人类主体性的参与。据此,使ChatGPT产生算法偏见的原因可总结为如下几个方面。首先,ChatGPT得以训练的语料库中的数据来源于互联网,当存有偏见性的数

据被收入语料库时,ChatGPT 不可避免会对其进行学习,进而习得人类社会的偏见,从而诱发“Bias in, Bias out”现象。其次,ChatGPT 背后模型的建构离不开算法工程师的编程,其在编程的过程中也会自觉或不自觉地将自身内隐偏见以代码形式外化并嵌于模型架构之中[13]。再次,ChatGPT 大模型对于普罗大众来讲本身就是一个黑箱,很少有人能够真正了解其背后的运行逻辑与知识架构,由于 GPT 大模型采用的是“利用人类反馈强化学习”的深度学习训练方式,其自身的模型会训练到何种程度,开发者也很难自知,这就使得 ChatGPT 背后的偏见问题变得更加隐蔽而不易被察觉。

(三)数据泄露

数据被誉为 21 世纪的“石油”。2020 年,中共中央、国务院印发的《关于构建更加完善的要素市场化配置体制机制的意见》明确指出,数据与土地、劳动力、资本、技术一样,都是可市场配置化的要素。2022 年,习近平总书记在主持召开的中央全面深化改革委员会会议上,进一步强调要促进数据高效流通使用,加快构建数据基础制度体系。由此可见,从战略层面来讲,数据已经成为数字时代推动数字经济高速发展的重要因素。数据背后不仅仅蕴含着丰富的经济利益,其本身还承载着国家安全、个人信息权益等价值期许。我国已在法律层面先后颁布了《网络安全法》《数据安全法》《个人信息保护法》等,为国家数据主权、个人信息保护等提供了法律保护屏障。

从工作原理来看,ChatGPT 作为对话式机器人,需要与用户进行人机交互,而后通过大模型生成相应的回答。但是 ChatGPT 并非本地化部署,所有的数据都会发送到位于美国的 OpenAI 公司的服务器。此外,根据 OpenAI 的隐私政策,ChatGPT 可能会自动收集有关用户使用服务的信息,如用户使用的功能或采取的行动,这就带来了个人信息、商业秘密等数据泄露的风险。原因在于,近年来全球范围内的勒索组织日渐猖獗,勒索攻击事件频频出现,出于金钱、政治等动因,勒索组织有针对性地对包括互联网系统在内的信息基础设施发起攻击,从中窃取大量的数据并将其作为勒索的质物。

此外,随着 ChatGPT 背后语料库数据的实时更新,人们输入的文本也会成为其用以训练的数据来源。当用户输入涉及个人信息、企业商业秘密或者国家安全的数据时,ChatGPT 会将其纳入自身语料库。然而一旦出现前文所述的用户以威逼利诱的方式命令控制 ChatGPT 的情况,其很有可能将此类数据全盘托出,从而给个人信息权益、企业商业秘密、国家安全带来巨大挑战。有鉴于此,出于对数据泄露的担忧,微软、亚马逊等互联网公司纷纷发文,提醒员工不要在 ChatGPT 上泄露有关机密信息。最近不少用户在社交平台上分享“不属于他们自己的 ChatGPT 对话记录”的相关截图,表示自己可以看到其他用户在 ChatGPT 上询问的话题,随后 OpenAI 创始人山姆·阿尔特曼在公开场合对此予以确认,并对因开源库漏洞所导致用户数据泄露的事件进行公开道歉。

除了因技术本身缺陷可能带来的法律风险,ChatGPT 的爆火还滋生了账号贩卖租赁的

乱象。由于ChatGPT尚未对中国大陆开放,一些商家看到其背后灰色产业链可能带来的巨额利益,于是在电商平台贩卖或者租赁账号,有些店铺一天交易的次数多达一万余次,交易价格也根据账号是否为多人共享或者专人定制存在高低之分[14]。尽管目前电商平台对从事ChatGPT账号贩卖或租赁的店铺进行封禁,并且屏蔽相关关键词,但是截至本文撰写及发稿前,在一些社交平台上仍然存在大量买卖租赁ChatGPT账号的行为。根据工信部等部门相关规定,未经批准不得自行建立或租用VPN开展跨境经营活动,ChatGPT账号的买卖行为涉嫌非法经营;如果商家未取得行政许可便从事国外账号买卖交易,则有可能受到行政处罚甚至刑事处罚。

三、ChatGPT法律风险的规制

(一)ChatGPT法律风险规制的理论选择

ChatGPT表征了人类科技进步的智慧之果,与以往弱AI相比取得长足的进步,但这并不意味着法律需要对该项新兴信息技术创设专门的规制规则,否则很容易陷入法律对每一项新技术制定一项规则的立法泥沼中,产生“法律万能主义”的认知偏差,并容易陷入“脚疼医脚,头疼医头”的恶性循环。

就目前认知而言,ChatGPT带来的法律风险并没有超越现有的研究范畴,现阶段对ChatGPT规制的重心仍然是聚焦于个人信息保护、数据安全、ChatGPT运营者的法定义务等核心议题。人们担心ChatGPT会泄露个人信息、商业秘密等风险,也不过是传统数据抓取层面带来风险的映射。以ChatGPT为代表的AI技术并不会就此停滞不前,因此,ChatGPT法律风险的规制理论范式,应当立基于过往技术规制的经验来选择最佳规制理论。目前学界关于技术规制的理论主要有敏捷治理理论、元规制理论、阶段性规制理论。

1.敏捷治理理论

敏捷治理是管理学领域最早提出来的理论。2018年世界经济论坛发布的《敏捷治理:第四次工业革命中政策制定的重新构想》白皮书认为,敏捷治理是一套自适应、以人为本、具有包容性和持续性的决策过程,明确政策的制定离不开多方利益相关者的努力,并尝试探索相应的理论框架[15]。后来,有学者将敏捷治理理论从技术领域拓展至跨学科领域,结合敏捷能力和治理能力,以快速觉察和应对意想不到的变化[16]。欧盟和美国的污染物排放许可制度即是敏捷治理的体现。敏捷治理理论具有治理主体参与广泛性和政策制定中的时间灵敏度特征,其更加强调治理工具和治理方式的灵活多变,而这恰好与区块链、算法、深度学习等技术的复杂特征相契合,为此有学者在区块链等技术治理领域引入敏捷治理理论[17]。敏捷治理是一套具有柔韧性、流动性和灵活性的方法,需要持续地追踪与分析重要变化,而法律仅是治理方式之一,法律所扮演的角色不过是弥补相关立法的缺位,但同时这也意味着治理成本的增加。

2.元规制理论

元规制理论发端于普通法系,德国将其引入社会性规制中,用于食品安全领域的规制,并发展出一套完备的理论体系。元规制被称为对自我规制的规制,是指公权力机构对企业开展的自我规制进行的外部监督[18]。数字时代以算法、区块链、元宇宙等为代表的新型信息通信技术的发展,重塑了信息社会的技术治理权力格局。传统的政府规制囿于政府专业知识和信息数据匮乏、财政负担过重等原因,面临着规制效果不佳的困境;而企业的自我规制虽然能够缓解政府规制面临的各种问题,其本身还存在着自我规制动力不足的弊端。元规制是对传统政府规制及企业自我规制的一种综合,是公私合作治理的典型体现,通过对被规制对象赋权并对其课以规制义务,督促企业更好地进行内部管理,从而达到行政监管的目的。为此,元规制理论受到学界的极大推崇,如有的学者在数据保护的研究中主张引入元规制模式,构建“通过设计保护数据”的元规制制度[19]。

3.阶段性规制理论

阶段性规制理论的核心要义在于,认为法律应当根据技术发展不同阶段所表现出的特征来确立回应方式。有学者根据阶段性规制理论将算法、深度学习等技术的发展划分为技术概念创新①、技术初步应用②和技术大众化应用③三个阶段,并根据各阶段表现出来的不同特征,分别确立重新解释现有法律条款、确立有关技术应用场景和应用方式的禁止性规范,以及解释技术开发者承担何种安全保障义务的技术法律回应方式[20]。以人脸识别技术应用的治理为例,我国并未从一开始就针对人脸识别技术创设全新的法律制度,而是随着大众化应用,以各方法律主体的权利义务为基线,强调人脸识别技术的应用者在具体应用场景中应当承担的安全保障义务以及履行义务的形式。2021 年最高人民法院发布的《最高人民法院关于审理使用人脸识别技术处理个人信息相关民事案件适用法律若干问题的规定》即可为此提供例证。

通过对学界既有技术规制理论的梳理可以发现,敏捷治理理论和元规制理论的共同点在于,在其理论的框架下,法律的功能是非常有限的,规制技术需要法律结合其他工具[21]。但是两者重点关注的是法律应当如何规制技术,却忽视了技术需要什么样的法律这个更值得引人思考的问题。AI 在带动经济、社会效益方面具有重要潜力,为了缩小这项技术可能带来的负面影响、实现利益最大化,政策制定者提出了各种方案对其负面性加以规制。然而,许多“不负责任”的提议反而会损害 AI 创新,前两种技术规制理论很少考虑“负责任的 AI 监管”究竟意味着什么。而阶段性规制理论恰好回应了技术需要什么样的法律、什么是“负责任的 AI 监管”这个让人深思的问题,即在不影响技术创新发展的同时,根据技术发展

① 技术概念创新阶段的特征表现为产业界对技术未来可能应用的场景只是处于设想阶段,如目前产业界仅仅是设想 ChatGPT 未来可能在医疗、金融、法律等行业进行应用。

② 技术初步应用阶段的特征表现为新技术如何进行商业化使用在各界已经达成基本的共识。

③ 技术大众化应用阶段的特征表现为新技术的应用场景已经非常明确,应用方式已经非常成熟。

不同阶段的特征,审时度势地确立相应的法律回应方式。

(二)ChatGPT法律风险规制的基本思路

1.基于现有规则的规制

当前ChatGPT尚处于技术概念创新阶段,对ChatGPT未来在商业化中如何应用尚未达成基本共识,ChatGPT背后可能产生的法律风险与现阶段社会现存的风险并无实质性差异,均是集中于个人信息保护、数据安全、服务提供者应当承担的法定义务等议题,而且有些风险属于技术层面的漏洞,通过对漏洞进行修补即可在很大程度上避免。故法律对ChatGPT的回应方式不应当采取专门式立法限制模式,否则很容易将技术创新的苗头扼杀在摇篮里,相反应当立基于ChatGPT初级阶段的技术特征,对现有的法律条款进行解释,延伸适用现有规则。如2023年2月8日,美国信息技术与创新基金会发布的《不损害人工智能创新发展的十项监管原则》指出,为了避免阻止或阻碍创新,在进行技术规制时,应当强制执行现有规则,因为许多现有法律法规等规范性文件已经解决了包含人工智能技术在内的共性问题,比如算法歧视、个人信息保护等问题。① 在这些情形下,通常不需要针对人工智能技术作出新的专门规定。

2.对违法或不良信息风险的规制

目前学界大多数学者否认人工智能的刑事责任主体资格。笔者认为,尽管ChatGPT具有类人性的特点,但是尚不具备完全自控和辨别的能力,所以当ChatGPT生成虚假信息以及不良信息时,其本身并不能构成我国《刑法》规定的"编造、故意传播虚假信息罪",也不能成为教唆犯、帮助犯的主体。但若用户把ChatGPT当成犯罪工具用来生成并传播虚假信息,或者通过威逼利诱的方式迫使ChatGPT提供犯罪思路,进而实施网络诈骗等犯罪,则应由该用户承担相应的法律后果。至于ChatGPT账号贩卖租赁行为,网信办、公安部等监管机构可以参考此前开展的多项监管打击行动,②考虑部署"ChatGPT账号贩卖租赁打击专项行动",加大对无证经营和非法交易行为的查处力度,如涉及犯罪,交由司法机关追究其刑事责任即可。

3.对数据泄露、算法偏见风险的规制

现阶段,针对数据、个人信息保护及算法的监管,我国《网络安全法》《数据安全法》《个人信息保护法》《互联网信息服务深度合成管理规定》《互联网信息服务算法推荐管理规定》《数据出境安全评估办法》等法律法规已构建起较为完善的体系框架。换言之,现阶段ChatGPT可能产生的法律风险,在现行的法律秩序框架下基本能够得到解决,而且尚无充分

① 如我国《关于加强互联网信息服务算法综合治理的指导意见》《互联网信息服务算法推荐管理规定》《国务院反垄断委员会关于平台经济领域的反垄断指南》等规范性文件剑指算法歧视;而我国《民法典》《个人信息保护法》《网络安全法》《数据安全法》等规范性文件为个人信息、数据提供了体系性的保护框架。

② 如在2022年,中央网信办组织开展了包括打击网络谣言、整治"饭圈"乱象、治理"网络水军"等在内的13项"清朗"专项行动。

的证据能够验证一味地严格监管与强行介入是规制技术风险的唯一或者最佳方式。如一味地依靠严格的监管和事后惩罚,则会损耗太多本可以预防其他类型风险的资源,损失了本可实现的更高层次的个人发展和社会进步。况且相较于事后惩罚,事先对风险的预防更有意义。为此,ChatGPT 风险规制不能忽视的关键一环,在于用好该项新型数字技术,开发出更强的保护机制和更有效的保护技术,将保护数据和规制偏见的法律法规和“用户导向”的原则引入到 ChatGPT 模型、算法、程序等设计和使用的各个环节,将风险规制前置。由此可见,ChatGPT 法律风险规制的重点应当注重从源头进行防范。

以 ChatGPT 为代表的人工智能相关企业,包括类 ChatGPT 的研发者,凭借资本原始积累,拥有强大的信息优势和技术优势,他们在处理 ChatGPT 带来的风险时较为敏捷,尤其应当严格贯彻落实数据安全及个人信息保护法律制度。在抓取互联网上的数据进行语料库更新时,应当加强内容审查及内容过滤,通过技术手段去除数据训练过程中的“噪声”,将真实性有待考证、虚假类、带有偏见性以及能够识别出自然人的数据进行过滤剔除。同时,为了避免前文所述“越狱”事件的发生,可以通过限制用户输入、自建库或第三方服务,必要时还可采取人工辅助介入的方式,对用户的输入信息进行实时监督。如果在 ChatGPT 的实际运作中,涉及数据跨境问题,相关企业应当积极开展数据出境风险自评估,贯彻落实数据出境相关义务,为数据出境获得相关行政许可,并建立完善的数据出境保障措施。

为了避免 ChatGPT 生成歧视性内容,相关企业应当定期审核、评估、验证算法机制机理、模型、数据,确保算法模型不违反伦理道德。在技术环节上,企业应当积极拓宽训练数据的采集维度、推动反算法歧视技术的研发。

网络安全为人民,网络安全靠人民。ChatGPT 用户应当提高个人信息保护和国家数据安全意识,避免在 ChatGPT 的输入框中输入涉密、敏感的信息,以防 ChatGPT 将该类信息纳入自身的语料库,以此通过与其他信息融合生成内容,进而引发数据泄露的风险。同时,用户应当做到合规合法地使用,不得通过教唆或者威逼利诱等方式,诱使 ChatGPT 产生带有违法性、歧视性的结论。

四、结语

人们为 ChatGPT 狂欢之余,需要正视其潜在的法律风险。以 ChatGPT 为代表的人工智能技术的迭代更新远超人类的想象,但其并不是潘多拉魔盒,法律也不应简单地牢牢捆绑魔盒。现阶段,ChatGPT 尚处于技术概念创新阶段,仅仅根据其背后不确定的法律风险制定专门立法进行限制为时尚早,这样做不仅会抑制技术的创新,扼杀技术可能带来的福利,还易陷入一项新兴技术对应一项规则的立法泥沼。相反,一个合乎逻辑的规制路径应当是基于阶段性规制理论,并对现有的相关法律条款进行解释和延伸适用,以促使该项技术真正为人类所用。尔后可以根据 ChatGPT 技术的发展情况,考虑寻求创设专门规则予以介入。

参考文献:

[1] 张夏恒.ChatGPT的逻辑解构、影响研判及政策建议[J].新疆师范大学学报(哲学社会科学版),2023(5):113-123.

[2] OUYANG LONG, WU JEFF, JIANG XU, et al.Training language models to follow instructions with human feedback[EB/OL].[2023-03-08].https://arxiv.org/pdf/2203.02155v1.pdf.

[3] 王佑镁,王旦,梁炜怡,等."阿拉丁神灯"还是"潘多拉魔盒":ChatGPT教育应用的潜能与风险[J].现代远程教育研究,2023(2):48-56.

[4] 赵广立.ChatGPT敲开了通用人工智能的大门了吗?[N].中国科学报,2023-02-22(03).

[5] 乌尔里希·贝克. 风险社会[M]. 何博闻,译. 南京:译林出版社,2004: 13.

[6] 杨永兴.人脸识别技术的法律规制[J].河南牧业经济学院学报,2022,35(6):54-60.

[7] 许可. ChatGPT时代的虚假信息[EB/OL].[2023-03-07].https://mp.weixin.qq.com/s/LdDJ4FZD7k09mGNMfGTc1A.

[8] 杨庆峰.ChatGPT的生成特性及其意义[EB/OL].[2023-03-07].https://mp.weixin.qq.com/s/xQaOVjFBAD3cUrTHddvICQ.

[9] ANIRUDH VK.This could be the end of Bing chat[EB/OL].[2023-03-09].https://analyticsindiamag.com/this-could-be-the-end-of-bing-chat.

[10] 王诺,毕学成,许鑫.先利其器:元宇宙场景下的AIGC及其GLAM应用机遇[J].图书馆论坛,2023(2):117-124.

[11] 陈永伟.超越ChatGPT:生成式AI的机遇、风险与挑战[J].山东大学学报(哲学社会科学版),2023(3):127-143.

[12] 王煜."危险"的ChatGPT:法律、伦理和哲学挑战[J].新民周刊,2023(6):28-31.

[13] 谢永江,杨永兴.人工智能时代下的算法歧视及其治理研究[J].北京邮电大学学报(社会科学版),2022(5):18-25.

[14] 韩丹东,王意天. ChatGPT火爆背后有何法律风险?[N].法治日报,2023-02-13(008).

[15] 吴磊,冷玉,唐书清.数字化时代敏捷治理的学术图景:研究范式与实现路径[J].电子政务,2022(8):77-88.

[16] LUNA A J H O, KRUCHTEN P, PEDROSA M L G E, et al. State of theart of agile governance: A systematic review[J]. International Journal of Computer Science & Information Technology, 2014(5): 121-141.

[17] 陈钰什,袁洪哲.区块链与算力管理:商业模式创新的新机遇[J].清华管理评论,2021(4):46-51.

[18] 韩新华.平台时代网络内容治理的元规制模式[J].中国出版,2022(5):51-54.

[19] 张涛.大数据时代"通过设计保护数据"的元规制[J].大连理工大学学报(社会科学版),2021(2):79-88.

[20] 赵精武."元宇宙"安全风险的法律规制路径:从假想式规制到过程风险预防[J].上海大学学报(社会科学版),2022(5):103-115.

[21] 杨尊源.规制四阶层论的形成机理、内容构造与运用实践——基于规制国理论与后规制国理论的思考[J]. 河北法学,2021(2):174-190.

(责任编辑:张秀宁)

ChatGPT legal risks and regulation

XIE Yongjiang, YANG Yongxing

(Institute of Internet Governance and Law, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: The explosion of ChatGPT announced the beginning of the era of strong artificial intelligence for humanity. With its powerful digital twin ability, creative ability, editing ability, and human-like nature, ChatGPT is driving the transformation of human society towards a more agile era. The emergence of ChatGPT is bound to pose a challenge to existing rules, which may trigger legal risks such as generating illegal or harmful information, algorithmic bias, data leakage, and using the opportunity to sell or rental accounts. However, there is no substantial difference between the legal risks that ChatGPT may generate and the existing social risks, and it has not yet deviated from core issues such as personal information protection, data security, and the obligations that service providers should bear. Therefore, the regulation of legal risks in ChatGPT should still be based on the characteristics of ChatGPT technology development, rely on stage regulation theory, pay attention to risk prevention, and timely interpret and apply the current basic legal provisions.

Key words: ChatGPT; artificial intelligence; legal risks; risk regulation