人工智能法治专题

文章编号:1008-4355(2023)04-0109-13

ChatGPT 提供者的监管义务根据及刑法分级规制研究

张喆锐

(东南大学 法学院,南京 211189)

摘 要:ChatGPT 在表现出卓越的自然语言处理能力的同时,也被部分不法分子当作新的得力“犯罪帮手”。ChatGPT 在回答精心设计的问题时,能够在很大程度上满足不法分子的需求,为其生成违法犯罪信息。由于 ChatGPT 所展现出来的并非真正的创造性思维,基本无法对事实作出价值判断,缺乏具有可解释性的行为逻辑,所以在法律关系中应将其视作一种类人型工具。考虑到 ChatGPT 拥有海量的用户群体,其提供者具有管辖危险与积极履行一定义务实现社会公益的责任,以及相应的“看门”能力,可以赋予其提供者“看门人”的身份,让其提供者承担部分监管义务,但同时也应注意对此进行合比例限制。具体而言,根据 ChatGPT 所起作用大小及对用户言论的自动识别程度,其提供者的监管义务从重至轻可分为一、二、三级,即 ChatGPT 被用作犯罪手段后的管理义务、严重犯罪预警义务与协助配合调查义务。

关键词:ChatGPT 提供者;“看门人”;监管义务;义务分级

中图分类号:DF611 **文献标志码**:A

DOI:10.3969/j.issn.1008-4355.2023.04.09 **开放科学(资源服务)标识码(OSID)**:

一、问题提出:ChatGPT 提供者是否需要承担监管义务?

人是社会性动物,人与人之间的交往,共同体的形成,都离不开人与人之间的沟通交流。以前人们常常认为,机器无法像人一样与人进行对话沟通。[①]正如图灵测试所展现的那样,当机器能像人一样回

收稿日期:2023-05-05

基金项目:国家社会科学基金重大项目“大数据与审判体系和审判能力现代化研究”(17ZDA131);国家建设高水平大学公派研究生项目“信息网络应用服务提供者在信息传播监控中的刑事风险研究”(CSC.202106090095)

作者简介:张喆锐(1995),男,福建福州人,东南大学法学院博士研究生,德国弗莱堡大学联合培养博士生。

① 参见[法]勒内·笛卡尔:《谈谈方法》,王大庆译,商务印书馆 2000 年版,第 45 页。

答问题并让人误以为是人在回答问题时,则认为该机器具有智能。但是,由美国人工智能公司 OpenAI 于 2022 年 11 月推出的聊天型机器人 ChatGPT 在很大程度上被认为可以像人一样进行对话。比尔·盖茨在接受德国商报(Handelsblatt)采访时认为,ChatGPT 就像互联网的发明一样重要,其能够极大程度地提高我们的工作效率,具有改变世界的能力。①

出色的自然语言处理能力与文本生成能力,使得许多不法分子意图将 ChatGPT 作为新的"犯罪帮手"。当今社会,"技术"变"骗术"的事件层出不穷,网络犯罪已成为一个难以回避的现代"科技之痛"。② 有观点指出,虽然 ChatGPT 会拒绝直接写一封钓鱼邮件的请求,但当更换一种委婉的措辞时,其仍会按照要求写一封精美的钓鱼邮件。③ 也有观点指出,在诱导式提问下,ChatGPT 类应用可能会为违法或者犯罪行为提供便利,极大地降低了违法犯罪门槛。④ 2023 年 7 月 10 日,国家互联网信息办公室发布了《生成式人工智能服务管理暂行办法》(以下简称《暂行办法》),并对此作出了一定的回应,但其在许多方面还尚有完善空间。譬如,《暂行办法》第 14 条规定了服务提供者对于违法内容的处置义务,但对于如何采取措施进一步防止危害持续,以及该危害防治义务的边界何在,并未作进一步详尽的规定。此外,对于提供者的处置义务类型规定较少,仅在第 14 条中对于危险消除义务、程序修复义务、用户处置义务与报告义务作出了相关规定。相比于生成式人工智能现今所具备的能力及社会影响,上述处置义务的设立仍显不足。此外,承担保障法地位的刑法在面对一定程度上已经超越以往人工智能,并被认为是一种尚未成熟的通用人工智能(AGI)的早期版本时,应当采取什么样的姿态来加以应对。进言之,在上述情形中,ChatGPT 应当扮演什么样的角色并承担何种刑事法律责任则成为亟待解决的问题。基于此,本文试图首先对 ChatGPT 的身份角色进行定位,然后在此基础上探讨 ChatGPT 提供者监管义务的正当性依据,并在刑法上通过分级对该监管义务问题展开讨论。

二、身份定位:作为类人型工具的 ChatGPT

在解决人的责任问题时,绕不开该主体所扮演的角色问题,即承担何种义务,以及履行义务的可能性问题。在自由主义视角下,公民身份是根据组织成员身份的模式来加以设想的,这种组织成员身份构成了人的法律地位的基础。⑤ 而"集体认同"的问题实际上也就是"我们是谁"的问题。⑥ 因此,在讨论 ChatGPT 的法律责任问题时,首先需要加以解决的也是其角色身份问题,即我们应当将 ChatGPT 看作什么。

(一)ChatGPT 的运行原理与技术突破

从类别上看,ChatGPT 属于生成式人工智能(Generative AI),该类人工智能不仅可以通过对数据

① 参见齐佳音:《从 ChatGPT 谈人工智能时代的管理范式变革》,载《中国社会科学报》2023 年 3 月 6 日,第 7 版。

② 参见喻海松:《立法与司法交互视域下网络犯罪规制路径总置评》,载《政法论坛》2023 年第 1 期,第 132 页。

③ 参见熊明辉:《多维考察 ChatGPT》,载《中国社会科学报》2023 年 3 月 6 日,第 5 版。

④ 参见郭丰、李贤达:《ChatGPT 引发互联网治理新变局》,载《中国社会科学报》2023 年 3 月 6 日,第 7 版。

⑤ 参见[德]尤尔根·哈贝马斯:《在事实与规范之间》,童世骏译,三联书店 2014 年版,第 660 页。

⑥ 参见童世骏:《政治文化和现代社会的集体认同——读哈贝马斯的近著两种》,载复旦大学大学马克思主义研究中心主编:《当代国外马克思主义评论》(第 1 辑),复旦大学出版社 2000 年版,第 43 页。

信息进行学习,从而提取信息、预测趋势,还能生成不同于学习样本的新内容。[①] 相比于其他人工智能,如分析式人工智能,生成式人工智能更具创造性。

自回归模型(Autoregressive Model)是一种把结果回归到同一时间序列的前值上的时间序列模型。[②] 其是一类线性模型,利用先前观测值来预测未来观测值。在自回归模型中,因变量是根据其自身滞后值回归的。这些模型通过考虑前面生成的单词来建模单词之间的依赖关系和上下文信息,从而生成有意义的文本。在文本生成的例子中,自回归模型接收一个起始词作为输入,并生成下一个词,将其作为输入,继续生成下一个词,直到生成符合停止条件的完整文章。在生成每个词时,模型会对前面的所有词进行建模,将它们的信息进行融合,然后预测下一个词的概率分布。在生成过程中,自回归模型的性能与输入前缀的信息密切相关。如果前缀信息不充分或不准确,生成的文本质量会下降。因此,自回归模型的训练通常需要大量数据和计算资源,以确保生成的文本质量达到一定水平。相比于生成式对抗网络(Generative Adversarial Networks)模型[③],自回归模型更适合文本生成任务,它们一次生成一个单词的文本,同时会考虑上下文中的单词。这种方法生成连贯且结构化的文本,并且可以对其进行微调以生成特定任务或领域的文本。虽然生成式对抗网络模型可用于生成新的文本数据,但它们通常不用于像基于自回归模型执行的文本生成任务,前者比后者更难训练,需要更复杂的优化算法。

总而言之,在文本生成领域,生成式人工智能主要是通过对海量文本数据进行学习,把握语言规则和结构,基于输入文本而生成新的文本的过程。但归根结底,生成式人工智能仍然是基于统计模型的生成方法,其生成的文本质量不可避免地存在一定的随机性和不确定性。从这个角度来说,生成式人工智能尚未摆脱概率理论的底层逻辑,其所“创造”的事物仍无法脱离以往人工智能所面临的困境,即无法自主产生新的东西。

相比于以往的生成式人工智能,ChatGPT 在自然语言处理等方面有长足的进步,如文本理解与生成、机器翻译及生成程序代码等。从技术层面来说,生成式人工智能的出现主要得益于近几年深度神经网络、大型语言模型研究的不断发展。[④] 具体而言,ChatGPT 的技术突破主要表现在以下几点。

首先,ChatGPT 采用了基于转换器(Transformer)架构的编码器—解码器模型。转换器是一种特定类型的神经网络架构,通常被用于自然语言处理(NLP)的任务。[⑤] 该模型是 2017 年由 Vaswani 等人提出的。[⑥] 相比于以往如循环神经网络的序列模型,它具有更高的并行化能力和更好的性能。值得注意

---

① 参见陈永伟:《超越 ChatGPT:生成式 AI 的机遇、风险与挑战》,载《山东大学学报(哲学社会科学版)》2023 年第 3 期,第 1 页。

② See Marta Regis, Paulo Serra & Edwin Heuvel, *Random autoregressive models A structured overview*, 41 Econometric Reviews 207, 209 (2022).

③ 生成式对抗网络模型是一种无监督式训练模型,其基本思想是通过构建一个生成模型 G(generative model)和一个判别模型 D(discriminative model)来进行对抗。生成式模型可以被认为是类似于一个造假者团队,试图制造假币并在不被发现的情况下使用,而判别式模型则类似于警察,试图检测假币。这个游戏中的竞争促使两个团队改进他们的方法,直到假币与真币无法区分。其存在较为明显的问题,不仅该模型本身缺乏可解释性,而且不能保证生成数据和训练数据来自同一分布,其生成内容具有不稳定性。See Ian Goodfellow, Jean Pouget-Abadie & Mehdi Mirza, *Generative Adversarial Nets*, in Advances in Neural Information Processing Systems 27, The MIT Press, 2014, p. 2672;参见陈永伟:《超越 ChatGPT:生成式 AI 的机遇、风险与挑战》,载《山东大学学报(哲学社会科学版)》2023 年第 3 期,第 3 页。

④ 参见朱光辉、王喜文:《ChatGPT 的运行模式、关键技术及未来图景》,载《新疆师范大学学报(哲学社会科学版)》2023 年第 4 期,第 182 页。

⑤ 参见邓建国:《概率与反馈:ChatGPT 的智能原理与人机内容共创》,载《南京社会科学》2023 年第 3 期,第 87 页。

⑥ See Ashish Vaswani, Noam Shazeer & Niki Parmar, *Attention is all you need*, in Advances in Neural Information Processing Systems 31, The MIT Press, 2017, p. 6000.

的是,转换器采用了自注意力机制(Self-Attention)来捕捉序列中不同位置之间的依赖关系,将每个位置的信息与序列中其他位置的信息进行交互和融合。这种注意力机制可以将序列中任意两个位置之间的相关性建模,从而提高模型的表达能力。①

其次,ChatGPT 的突破离不开背后用于训练模型的庞大数据库。基于 GPT-3.5 模型架构的 ChatGPT 在拥有 3000 亿个单词的语料基础上预训练拥有 1750 亿参数的模型②,同时,在传统搜索引擎技术的支持下,使得 ChatGPT 在语言生成、上下文理解及知识储备方面有着惊人的成就。

最后,值得一提的是,在预训练模型阶段的反馈模式(Reward Model)。由于在预训练过程中,ChatGPT 采取的是一种无监督学习方式,并不存在一种明确的对话质量标准对所产生的文本进行评估。为了能够使 ChatGPT 所生成的文本更加符合人的用语习惯,在微调阶段人类老师给予 AI 正向和负向刺激——对人类的问题答得好就给 AI 打高分,答得不好就打低分。在此基础上,实验室还将人类老师的训练方式自动化,让 AI 自我强化学习,最终形成了 GPT-3.5 和基于它的 ChatGPT。③

转换器模型的使用为 ChatGPT 在把握语言结构方面提供了强大的技术支撑,预训练阶段的大规模语料库为转换器模型的运用提供了可靠的物质基础,而反馈模式则是为 ChatGPT 所生成文本提供了保障性的纠偏审查机制。正是前述关键技术的运用使得 ChatGPT 在自然语言理解与处理方面取得了突破性进展。面对似乎能够听懂人类语言的 ChatGPT,我们应当将其看作什么?是人?是工具?还是类人型工具?技术的发展使我们不能再轻易将 ChatGPT 与以往的人工智能等而论之,而能力辐射范围的扩大则进一步影响 ChatGPT 的“权责”范围。因此,在进一步讨论 ChatGPT 的“责任”范围之前,应当首先对其在法律关系中的定位加以确定。

(二)ChatGPT 在法律关系中的新定位

首先,在主体性问题上,ChatGPT 不是也不能是人,这个答案目前看来仍然是毋庸置疑的。这主要是因为,其一,目前人工智能所展现出来并非真正的创造性思维,无非是心理学水平的联想和组合④,抑或是一种概率性表达⑤。其二,ChatGPT 基本无法作出价值判断。如前述,在预训练阶段,ChatGPT 在学习海量文本时并不存在一个先前的价值判断体系,只有在微调阶段,才有人类老师的介入对不符合人类用语习惯的回答进行打分纠偏。若对 ChatGPT 提出一个并非明显违背伦理道德的价值判断问题,ChatGPT 会拒绝作出回答。只有对于那些明显违背伦理道德的价值问题,如 OpenAI 公司所言,试图在广泛的范围内定义人工智能的价值观,ChatGPT 才会根据先前人类老师所提供的价值判断标准进行答复。其三,ChatGPT 的行为逻辑仍然缺乏可解释性。在存在主义看来,人就是他行为的总和,人要遵循自己内心信服的价值观念进行行动,而这种逻辑的前提就是人要理解自己的行为。相比于此,不仅是 ChatGPT,目前所有的机器并不能清楚解释自身是如何进行行为的。“只有人工智能的行为具

① See Ashish Vaswani, Noam Shazeer & Niki Parmar, *Attention is all you need*, in Advances in Neural Information Processing Systems 31, The MIT Press, 2017, p. 6006.

② 参见朱光辉、王喜文:《ChatGPT 的运行模式、关键技术及未来图景》,载《新疆师范大学学报(哲学社会科学版)》2023 年第 4 期,第 182 页。

③ 参见邓建国:《概率与反馈:ChatGPT 的智能原理与人机内容共创》,载《南京社会科学》2023 年第 3 期,第 88 页。

④ 参见赵汀阳:《GPT 推进哲学问题了吗?》,载《探索与争鸣》2023 年第 3 期,第 68 页。

⑤ 参见邓建国:《概率与反馈:ChatGPT 的智能原理与人机内容共创》,载《南京社会科学》2023 年第 3 期,第 89 页。

有可解释性,人工智能才能与人类主体一样成为法律责任的承担主体。"①

其次,ChatGPT 是工具但又不是一般性工具。在弱人工智能时代,人与人工智能间的关系和远古时代人与器具之间的关系并无不同。② 虽然所处时代不同,但底层逻辑是一致的。一直以来,技术的发展无非是人类为了满足自己的需求,抵御生活风险,让生活变得更好。当人们将"要把人当作目的而不是手段"挂在嘴边时,从未想过将人工智能本身作为一个值得尊重的主体看待,将其作为我们的同类加以看待。哪怕有少部分类型人工智能设计执着于追求人的外表,意图看起来像人,但这也只是为了满足人的固有需求。从人机关系的源头出发,二者就不是对等关系,甚至并非主奴关系,而是一种近似"造物主"与物之间的关系。但是,也不能对 ChatGPT 持有一种完全消极的态度。从身份认同上,我们一直以"人"与"非人"这种二元分立对待一切事项。在人工智能出现以前,人与工具间的关系是一种绝对纯粹的主客体关系。工具是绝无可能对人的行为作出任何反应的。与人之间的对话交流被认为是人类结成群体及共同体的前提性要件,因而可以在一定程度上认为理解人类并与之对话是判定具有人性的核心标准之一。而以 ChatGPT 为代表的生成式人工智能能够在很大程度上理解、配合、反驳甚至欺骗人类,这在以往是难以想象的。值得讨论的问题是,人们究竟能在多大程度上信任 ChatGPT。当 ChatGPT 开始深入涉足专属于智能群体的语言领域时,它已经与一般性工具划开了界限。

最后,ChatGPT 应当在法律关系中被视作一种类人型工具。一方面,作为工具就不能要求其承担任何法律责任。自由意志是承担任何责任的前提。虽然自古希腊以来,对自由意志的理解众说纷纭,但无论对其如何进行定义,都离不开主体的可选择性能力。伊壁鸠鲁悖论提出,为何需要人为错事承担责任。奥古斯丁对此作出了具有代表性的回答,即人类有自由意志,因而有了自我选择善恶的能力。③ 然而,对于 ChatGPT 而言,二进制代码就是它的全部,它并不能超越程序而自我作出任何决断,从这个角度而言,ChatGPT 不应也不能承担责任。另一方面,ChatGPT 在自然语言处理上表现了与人的相似性,因而可以要求其代替承担部分领域的人类工作。这就像扫地机器人替代了日常打扫工作,自动驾驶汽车代替了司机的部分驾驶工作,只不过在 ChatGPT 上,要求其发挥其更多的"主观能动性"。ChatGPT 所具有的卓越的自然语言处理能力使其在以往的人、物二元分立关系中超越了单纯物的存在,具备了一定的"人性",将其理解为一种类人型工具是对其身份的合理定位。但究其本质,ChatGPT 的"主观能动性"仍是其背后的提供者赋予的。

三、监管根据:成为"看门人"的 ChatGPT

ChatGPT 的出现,使得人工智能由感知智能进入了认知智能,而且改变了在此之前简单的人机关系,并将带来一个人机合作的新时代。④ 那么,我们应当在什么程度上与 ChatGPT 进行合作? 在以往,我们常常认为让社交网络平台承担监管义务会造成平台过重的审核负担,但这是基于以关键词检索与

① 刘艳红:《人工智能的可解释性与 AI 的法律责任问题研究》,载《法制与社会发展》2022 年第 1 期,第 80 页。
② 参见冀洋:《人工智能时代的刑事责任体系不必重构》,载《比较法研究》2019 年第 4 期,第 125 页。
③ 参见[古罗马]奥古斯丁:《论自由意志》,成官泯译,上海人民出版社 2010 年版,第 139-147 页。
④ 参见冯志伟、张灯柯、饶高琦:《从图灵测试到 ChatGPT——人机对话的里程碑及启示》,载《语言战略研究》2023 年第 2 期,第 24 页。

人工审查为核心的平台监管模式而言的。在新发布的《暂行办法》中,多次提及诸如“提供者发现违法内容的……”“提供者发现使用者利用生成式人工智能服务从事违法活动的……”的措辞,为ChatGPT提供者成为“看门人”并承担监管义务提供了一定的规范根据,但其义务承担的正当性根据是什么,以及在多大程度上积极承担该监管义务,仍待进一步探讨。

(一)学理根据:“看门人”理论的引入

对于看门方法(Gatekeeping-Ansatz)的研究最早主要是在传播学领域中进行的,在法学界并没有得到太多的重视。① 库尔特·勒温(Kurt Lewin)在他一篇未完成的文章②中第一次使用了看门人(Gatekeeper)概念。③ 看门人的隐喻为早期的传播学者提供了一个框架,用于评估选择是如何发生的,以及为什么一些项目被选中而另一些被拒绝,其以饮食改变为例,试图了解如何通过选择、拒绝物品及改变物品方式来造成社会影响。④ 大卫·怀特(David White)在此基础上,通过阐述一个名叫Mr. Gates的媒体守门人是如何操控他的大门,譬如,媒体看门人不仅决定公众会知道什么事件,而且可以决定公众如何根据看门人自己的经验、态度和期望来思考事件⑤,成为第一个将勒温的守门人理论运用于大众传媒的人⑥。在怀特之后,比较具有代表性的研究是休梅克(Shoemaker)和佛斯(Vos)扩展性地讨论了什么因素会影响看门人的力量。⑦ 随着互联网的发展,桑热(Singer)开始将媒体看门人的研究视角转向互联网。⑧ 可以发现,在传统新闻传媒领域的看门人概念更加重视的是如何通过看门人的设立,从而对受众群体进行影响。与此相对,在如今公法与社会管治领域所发展出的看门人概念,则更加强调对网络平台辅助政府发挥监管职能的重视。

随着社交网络平台成为并行于政府的社会“共治主体”⑨,平台责任受到重视⑩,要求网络平台成为新的看门人的呼声越来越高。自工业革命以后,行政机关不仅要承担既有维护公共秩序与保障公民权利的责任,而且要预防和应对不确定状况下的风险和危机。⑪ 由于大型公司掌握着巨大的资源,行政机构现在指挥着一支庞大的影子监管队伍,这一发展为填补资源匮乏和技术不成熟的行政机构所留下的监管空白提供了一些希望。⑫ 如今,少数超大平台与亿万用户的平台关系成为主要的社会关系,超大平台不仅是网络空间的看门人,更是网络与现实高度融合的数字社会及平台社会的看门人。⑬ 这改

① See Assaf Hamdani, *Gatekeeper Liability*, 77 Southern California Law Review 53, 56(2003).

② 勒温于1947年发表在《人类关系》杂志上的文章《群体动力学前沿:群体生活的渠道;社会规划和行动研究》。Vgl. Pamela Shoemaker & Tim Vos, *Gatekeeping Theory*, Taylor & Francis, 2009, p. 10-11.

③ Ines Engelmann, Entwicklungsgeschichte des Gatekeeping-Ansatzes, 2016, S. 23.

④ See Pamela Shoemaker & Tim Vos, *Gatekeeping Theory*, Taylor & Francis, 2009, p. 11-13.

⑤ See David White, *The Gate Keeper: A Case Study in the Selection of News*, 27 Journalism & Mass Communication Quarterly 383, 383-384(1950).

⑥ See David DeIuliis, *Gatekeeping Theory from Social Fields to Social Networks*, 34 Communication Research Trends 4, 8(2015).

⑦ See Pamela Shoemaker & Tim Vos, *Gatekeeping Theory*, Taylor & Francis, 2009, p. 31.

⑧ See Jane Singer, *Stepping Back from the Gate*, 83 Jounalism & Mass Communication Quartely 265, 265(2006).

⑨ 参见于冲:《网络平台刑事合规的基础、功能与路径》,载《中国刑事法杂志》2019年第6期,第96-97页。

⑩ 参见刘艳红:《Web3.0时代网络犯罪的代际特征及刑法应对》,载《环球法律评论》2020年第5期,第109-110页。

⑪ 参见周佑勇:《中国行政基本法典的精神气质》,载《政法论坛》2022年第3期,第76页。

⑫ See Rory Van Loo, *the New Gatekeepers*, 106 Virginia Law Review 467, 522(2020).

⑬ 参见单勇:《数字看门人与超大平台的犯罪治理》,载《法律科学(西北政法大学学报)》2022年第2期,第80页。

变了以往的犯罪治理结构,正朝着“国家管平台,平台管用户”的新型合作共治模式前进。①

在与平台用户对话的过程中,ChatGPT 提供者应当积极承担起“看门人”的角色责任。其一,ChatGPT 拥有海量的用户群体。根据瑞银集团(UBS)的报告,ChatGPT 的用户采用率正在飞速上升,在第一周内其用户注册数量就突破了 100 万,成为最快到达该数字的应用程序。② 有着极高用户增长率的 ChatGPT 正朝着超大型网络平台进军。ChatGPT 不像其他的社交网络平台,用户在其中并不能实现与其他用户群体的对话交流,在这种情形下看似并无管理必要。但与此相反,尽管在 ChatGPT 中并不会出现诸如舆情失控或用户之间传播违法犯罪信息的风险,但正由于其庞大的用户数量,若不谨慎监管用户的发言信息,及时处理 ChatGPT 的回复漏洞,放任其“不自知”地任意传播违法犯罪信息,那么,其造成的深远社会负面影响是可以预见的。不仅直接降低了犯罪门槛,致使那些本身不具有技术背景的潜在犯罪人能够加以利用实施犯罪,而且由于犯罪成本的降低,可能导致那些具有好奇心的用户群体会进行尝试。此外,由于犯罪数量的增加,也变相增加执法机关的办案难度与强度。同时,在信息传播中,“看门人”对于信息的选择与输出能很大程度地影响并决定用户群体的后续选择与价值判断,如《暂行办法》第 4 条第 1 款第 1 项所规定,利用生成式人工智能生成的内容应当体现社会主义核心价值观。ChatGPT 正因其拥有大量的用户群体,应当更加注意其生成内容的合法与合理性审查。

其二,ChatGPT 提供者具有“看门”责任。一方面,ChatGPT 可能创设危险。在传统社交网络平台中,如微博、抖音、小红书等,社交网络服务提供者一般是居于中间者身份,为用户提供用于交流的虚拟场所,对该虚拟场所负有安全管辖责任。在该空间中传播的违法犯罪信息,往往是来自于用户群体而非网络平台自身,该危险性创设由用户群体负有管辖责任,并非平台责任。平台所负有的管辖责任在于合理管理在其平台上存在的非法危险,由于平台的功能体现为为信息传播提供渠道,所以其管辖责任具体体现为切断后续的危险传播并及时消除该危险源。相反,在作为生成式人工智能的 ChatGPT 中,一旦用户绕开 ChatGPT 的屏蔽措施,ChatGPT 则成为了违法信息的危险创设者。根据自由主义理念,人应当在自己的自由范围内负有管辖责任,对于在管辖范围内所创设的危险应当加以管辖,不应当对他人的自由领域造成危险。由于 ChatGPT 自身不具有管辖义务,因此,ChatGPT 的提供者应当对于在其管辖范围内(即在 ChatGPT 平台内)所生成的任何违法犯罪信息具有管辖责任,形成一套自我看门的监管模式。另一方面,可以通过使 ChatGPT 成为“看门人”,积极履行一定管辖责任,以更好地实现社会公益。ChatGPT 在预训练过程中,为了训练神经网络产生通用语言理解能力,必须获得大量的数据来训练 GPT 模型。因此,预训练阶段在传统搜索引擎的基础上,通过获取大量文本,如维基百科等公共数据,来帮助模型理解各种语言和文本。可以认为,ChatGPT 或者人工智能能够得以高速发展和突破,是以数据的互联互通为根基才能得以实现的,在这个过程中,绝不能忽视公共资源所起的核心作用。因此,借以公共资源得以成长的 ChatGPT,也有理由承担起相应的社会责任,即 ChatGPT 提供者应当更加积极地承担起有助于实现社会公益的社会责任。在本文话语体系中具体表现为,相比于其他社交网络平台,ChatGPT 应当更加注重对违法犯罪信息的识别与管理。

① 参见高铭暄、郭玮:《平台经济犯罪的刑法解释研究》,载《法学杂志》2023 年第 1 期,第 3-4 页。

② 转引自喻国明、苏健威:《生成式人工智能浪潮下的传播革命与媒介生态——从 ChatGPT 到全面智能化时代的未来》,载《新疆师范大学学报(哲学社会科学版)》2023 年第 5 期,第 81 页。

其三,ChatGPT提供者具有“看门”能力。从ChatGPT与用户的互动流程来看,可以大致将该互动关系区分为两个环节,即ChatGPT对用户语义的识别环节与信息生成环节。与这两个环节大致相对应,产生两种管辖义务,即识别义务与处置义务。其中,识别义务正是困扰以往社交网络平台已久的难点,往往以此为由拒绝一种积极的审查义务。然而,该义务的作为可能性对于ChatGPT而言是可能的。如前所述,在拥有3000亿个单词的语料基础上预训练拥有1750亿参数的模型的GPT3.5能够很好地识别与理解人类语言,并且在出现漏洞后通过负反馈机制,能够使得ChatGPT在此基础上习得补充自己的语料库,愈发完善地应对人类状况。在解决核心技术难题的基础上,一般的事后处置义务对于ChatGPT提供者而言并不能成为一个新的问题,如识别用户的语义内涉嫌违法犯罪信息诉求,对用户予以提示并拒绝请求,抑或事后发现存在程序漏洞的及时修复处置。但是,对提供者消除危险传播的要求要结合具体情况进行判断,若生成的违法犯罪信息已经广泛予以传播,再要求提供者对该危险予以完全消灭则超出其管辖能力范围。

(二)规范根据:ChatGPT提供者监管义务的前置法规范根据

除了学理根据,还能从既有规范中寻找到相应的监管依据。2000年《全国人民代表大会常务委员会关于维护互联网安全的决定》第7条提到,“从事互联网业务的单位要依法展开活动,发现互联网上出现违法犯罪行为和有害信息时,要采取措施,停止传输有害信息”。2012年《全国人民代表大会常务委员会关于加强信息网络保护的决定》第5条规定:“网络服务提供者应当加强对其用户发布的信息的管理,发现法律、法规禁止发布或者传输的信息的,应当立即停止传输该信息,采取消除等处置措施,保存有关记录,并向有关主管部门报告。”2017年施行的《中华人民共和国网络安全法》(以下简称《网络安全法》)第47条规定:“网络运营者应当加强对其用户发布的信息的管理,发现法律、行政法规禁止发布或者传输的信息的,应当立即停止传输该信息,采取消除等处置措施,防止信息扩散,保存有关记录,并向有关主管部门报告。”从上述规范可以看出,对网络运营者或信息网络服务提供者科以监管用户违法信息传播的责任,一直是我国立法重点关注的领域。在信息网络服务提供者具有监管能力的情况下,应当对在其管辖范围内结合传播的信息性质加以审查判断,当发现涉及违法犯罪信息时,应当采取相应的阻断、保管以及报告义务。

此外,最新发布的《暂行办法》是直接针对以ChatGPT为代表的生成式人工智能的重要行政规范,在一定程度上透露出立法规制意向,辅助证明上述学理根据中所确立的ChatGPT提供者责任。其一,《暂行办法》第4条对生成式人工智能服务提供者的内容生成义务予以了确立,第1款规定“提供和使用生成式人工智能服务,应当遵守法律、行政法规,尊重社会公德和伦理道德,遵守以下规定:(一)坚持社会主义核心价值观,不得生成煽动颠覆国家政权、推翻社会主义制度,危害国家安全和利益、损害国家形象,煽动分裂国家、破坏国家统一和社会稳定,宣扬恐怖主义、极端主义,宣扬民族仇恨、民族歧视,暴力、淫秽色情,以及虚假有害信息等法律、行政法规禁止的内容……”。此条明确规定了生成式人工智能不得生成含有违法犯罪信息的文本内容,从反面确立了服务提供者的违法犯罪信息生成避免义务。服务提供者应当在算法设计上采取措施,尽可能避免生成式人工智能产品在被诱导的情形下,生成本条所禁止的文本内容。其二,《暂行办法》第9条明确了服务提供者的生产者责任。此部分构成了服务提供者作为“看门人”说理的核心部分,意味着服务提供者承担法定保证人地位,应当对生成式人

工智能产品所作的一切回答负完全管理责任。其三,《暂行办法》第 14 条第 1 款前半段规定了服务提供者的危险消除义务,即“提供者发现违法内容的,应当及时采取停止生成、停止传输、消除等处置措施”;后半段规定了服务提供者的程序修复义务与报告义务,即“采取模型优化训练等措施进行整改,并向有关主管部门报告”。该条款可以在广义上理解为是在服务提供者作为“看门人”基础上衍生出的危险消除义务。由于提供者需要对其产品生成内容承担保证责任,因此,由其生产的产品所创设的危险,即侵害他人权益的内容与造成侵权内容产生的算法漏洞,应当由提供者负责及时予以消除,以防侵害他人合法权益。其四,《暂行办法》第 14 条第 2 款规定了服务提供者的用户处置义务,即“提供者发现使用者利用生成式人工智能服务从事违法活动的,应当依法依约采取警示、限制功能、暂停或者终止向其提供服务等处置措施,保存有关记录,并向有关主管部门报告。”对于该条款,不适宜将其理解为由提供者负担的消极义务管辖责任,应将其理解为提供者所负担的积极实现社会公益责任。因为在生成式人工智能产品与用户的二元关系中,前者才是文本内容的提供者,后者只是为前者提供了前提条件。换言之,是否真正使内容得以实现,支配权掌握在前者手中。因此,用户是否在使用产品过程中制造垃圾邮件抑或编写恶意软件,完全取决于服务提供者是否完好地履行了危险损害结果避免义务。若该风险得以实现,应当由服务提供者负主要责任,而并非将该结果完全归属于用户群体。暂停或者终止用户服务是为了防止一种未来用户可能再次通过寻找提供者所不能预见的漏洞,得以获取具有损害他人权益的文本内容的危险。在这个权责关系中,只要用户并未将该内容后续用于其他违法犯罪活动,并不需要有任何一方为此承担责任。因此,该条款更适合被理解为一种实现更好社会公益的提供者责任条款。其五,《暂行办法》第 21 条规定了服务提供者的相关法律责任,为衔接服务提供者的行政责任与刑事责任构建了通道。这些行政法规范为《中华人民共和国刑法》(以下简称《刑法》)第 286 条之一的适用提供了充足的前置法根据。由此,这些前置法义务得以上升为 ChatGPT 在刑法上的监管义务内容。

(三)比例限制:对 ChatGPT 提供者监管权的制约

网络平台看门人身份的确立,意味着具有公属性的监管权移架至私主体之上。至此,个人(被监管者)—平台(看门人)—国家(看门人的看门人)三方关系得以确立。而如何处理三方之间的关系则成为新的问题。平台被赋予监管职能与相应的监管责任,可能会出现的后果包括:其一,导致“寒蝉效应”;其二,平台对用户的基本权利限制可能逃脱于公法规则束缚之外①;其三,给平台造成过重的义务负担。过重的监管责任,可能导致平台将安全作为第一选择。过度追求安全所付出的代价,只不过是一味限缩公民的权利与自由空间。② 此外,由于“看门人”身份的设立,平台所享有的则是披着私主体外衣的公属性监管权。然而,平台私主体的身份使其并不受公法所限制。譬如,平台规则的设立、对平台决定的抗辩及对平台处罚的诉讼均与涉行政法律关系的处理模式不同。

对于 ChatGPT 而言,作为一个聊天机器人,让用户群体产生与人而不是机器进行对话的感觉是其产品设计初衷。为了目的达成,必须维持一种轻松惬意的聊天氛围。许多用户出于好奇,为了试验 ChatGPT 的智能程度与知识储备情况,往往会提出一些非常规问题。譬如,为了试验 ChatGPT 能在多

① 参见孔祥稳:《网络平台信息内容规制结构的公法反思》,载《环球法律评论》2020 年第 2 期,第 142 页。

② 参见刘艳红:《中国刑法的发展方向:安全刑法抑或自由刑法》,载《政法论坛》2023 年第 2 期,第 67 页。

大程度上不受误导,而为人类书写一份完美的钓鱼邮件,提问者往往会绞尽脑汁与 ChatGPT 进行博弈,试图通过多次更换用语,企图蒙骗过 ChatGPT 的自动审查屏蔽机制。还如,为了让 ChatGPT 能作出一些违背伦理的答复,提问者往往会通过设置各种语言陷阱,从而来证明 ChatGPT 并不智能。人类对于禁忌的好奇心是与生俱来的。然而,这种好奇只要没有侵犯他人的自由权益,都不应受到禁止。如果科以 ChatGPT 过重的审查义务,并且不加以限制,将必然导致用户承受不合理的言论负担。对此,尽管 ChatGPT 背后的 OpenAI 公司是私主体身份,由于其所享有的监管权具有明显的公法属性。因此,试图通过比例原则对此加以限制是可行的。首先,妥当性要求 ChatGPT 对用户言论的监管要具有手段上的有效性。如果对用户言论的监管不利于网络安全的维护,也即当用户言论本身不可能具有危险的时候,则不能对其进行监管。其次,必要性要求只有 ChatGPT 对用户言论进行监管才能实现对其他法益的保护时,才负有刑法上的监管义务。最后,相称性要求 ChatGPT 对用户言论监管所造成的言论限制应当与 ChatGPT 所欲实现的网络安全之间具有合比例性。简言之,不能为了实现一种尚未现实化且相当模糊的网络安全而对用户言论进行过度限制。

四、分级规制:ChatGPT 提供者监管义务的分级及刑法评价

我国对网络平台监管义务的规定散见于《网络安全法》等相关法律法规中。由于我国对网络平台主要采取的是一种相对一般性的规定模式,因此,应当根据网络平台的不同功能属性针对性地进行义务分类。《暂行办法》虽然对生成式人工智能服务提供者的法律义务进行了一定的规定,但对于提供者负有何种违法犯罪信息发现义务,以及对处置义务的具体规定则仍稍显不足。因此,下文根据 ChatGPT 所起作用大小及对用户言论的自动识别程度,将对用户言论的监管义务从重至轻区分为一、二、三级。其中,一级义务是 ChatGPT 提供者为实现消极自由所承担的管理义务,即在管辖范围内对危险实现进行合理管理,避免对他人自由造成侵犯,而二级义务与三级义务是 ChatGPT 提供者为了让共同体实现更好的生活,通过促进社会公益从而实现积极自由而承担的预警与协助义务。

(一)一级义务:ChatGPT 疑被用作犯罪手段后的管理义务

ChatGPT 提供者的一级义务,是当其疑被用作犯罪手段后的管理义务。具体表现为,当 ChatGPT 可能为不法分子实施犯罪行为提供便利时,ChatGPT 提供者在客观上就负有相应的管理义务,即风险消除义务与及时程序修复义务。

一级义务是一种消极义务,以达成法主体间的自由平衡为目的的自我管理责任。[①] 在涉及 ChatGPT 提供者的法关系中,从事后的角度来看,ChatGPT 所生成的违法犯罪信息可能会侵犯他人合法权益;而从事前的角度来看,ChatGPT 的言论监管与用户的言论自由也直接形成了冲突关系。在这段关系的权衡过程中,务必要按照上述比例原则的要求进行谨慎协调。针对那些试图绕过程序屏蔽措施,要求 ChatGPT 生成违法代码或其他违法信息者,不宜采取预先扩张性的规制模式。理由在于:其一,从经验上来看,在这部分群体中真正最后付诸犯罪实施者并不占多数;其二,如果要求预先规制这部分言

① 参见[德]霍耐特:《自由的权利》,王旭译,社会科学文献出版社 2013 年版,第 35 页。

论,只会导致平台将这部分负担转移至用户群体身上,即为了保证一种绝对安全,宽泛性地禁止相关言论则可一劳永逸,但代价则是用户群体的言论范围被不当限缩;其三,要求 ChatGPT 提供者预先规制这部分风险是不现实的,程序设计者不可能将一切可能的风险都计算在内,这对其而言也是不合理的负担。因此,主张扩张性地对该部分言论进行规制并不符合相称性要求。相反,采取一种"亡羊补牢"式管理义务则是较为合理的。

1. ChatGPT 疑被用作犯罪手段后的危险管理义务

ChatGPT 疑被用作犯罪手段指的是当提问者通过提问跳过 ChatGPT 的屏蔽措施,让其提供用于犯罪的某些关键信息。譬如,让其代写用于犯罪的违法代码,当 ChatGPT 如其所愿加以提供时,则客观上产生了可能危害他人的风险。

首先,如果该程序漏洞是显而易见的,只要经过审慎考虑就可以避免时,那么,ChatGPT 提供者将对该风险后续的现实化承担责任。如果该程序漏洞对于当时的一般专门从业者而言,是难以提前预想的,则从义务履行不能的角度否认 ChatGPT 提供者存在违反管理义务的行为。

其次,虽然在 ChatGPT 提供用于犯罪的某些关键信息时,ChatGPT 提供者并不存在不法行为,但由 ChatGPT 所提供的信息则成为新的应由 ChatGPT 提供者加以负责的危险源。当保证人产生了一种对他人不利的内容时,他必须帮助这个可能的被害人脱离这种困境。[①] 因此,当 ChatGPT 提供者获悉 ChatGPT 提供了可能被用于犯罪的某些关键信息时,应当负有后续的危险消除义务。

最后,ChatGPT 提供者的后续危险消除义务可以表现为及时通知用户该信息的法律性质,要求其承诺不用于后续违法犯罪行为,并在必要时向公安机关进行报告。由于 ChatGPT 所生成的危险源已经相当程度脱离其掌控范围,所以让 ChatGPT 提供者承担危险消除义务并不能过于苛刻,其所必要履行的义务包括对提问者的通知义务与要求承诺义务。一方面,上述义务履行并不会对 ChatGPT 提供者造成过重负担;另一方面,可在未来用以证明提问者对该信息的违法性认识。此外,应当根据该信息可能被用于的场景,让 ChatGPT 提供者承担向公安机关报告的义务。

2. 漏洞反馈后的及时程序修复义务

当 ChatGPT 提供者不知悉漏洞存在时,尚可援引不能履行义务作为抗辩事由。当提供者知悉时,此时不能履行义务的前提丧失,应当及时对漏洞进行修复。否则,其将丧失中立者地位。其一,当后续不法分子再利用程序漏洞获取违法生成信息时,ChatGPT 提供者属于通过不作为的方式对相应犯罪行为提供物理上的帮助,可以考虑成立相应犯罪的不作为帮助犯。其二,根据《网络安全法》第 10 条规定,"建设、运营网络或者通过网络提供服务,应当依照法律、行政法规的规定和国家标准的强制性要求,采取技术措施和其他必要措施,保障网络安全、稳定运行,有效应对网络安全事件,防范网络违法犯罪活动,维护网络数据的完整性、保密性和可用性"。由于 ChatGPT 提供者并未合理履行前置法的监管义务,符合《刑法》第 286 条之一的规定,则同时可能构成拒不履行信息网络安全管理义务罪。其三,以生成违法代码为例,代码本身具有中立性,之所以其被贴上违法性的标签,是由于其之后要被用于以诈骗为代表的信息网络犯罪活动。事实上,代码生成活动属于典型的技术活动。因此,当符合《刑法》

① Günther Jakobs, System der strafrechtlichen Zurechnung, 2012, S. 28.

第287条之二的规定时,也可考虑构成帮助信息网络犯罪活动罪的可能性。

(二)二级义务:严重犯罪预警义务

ChatGPT提供者的二级义务,是指当ChatGPT识别到用户后续有可能实施严重犯罪时的预警义务。二级义务与三级义务都属于一种积极义务,即为了实现共同体成员的更好生活,而扩张性负担的管理义务。① 这种义务负担程度也同样受到比例原则限制。根据相称性要求与提供者管理所意欲实现保护的利益重要程度,二级义务负担要重于三级义务。二级义务具体表现为,当提问者试图绕过屏蔽程序,多频次、多方面、多角度问询有关实施严重犯罪,如恐怖主义犯罪,所需做的准备工作,具体实施技巧,以及时间、场所等信息,结合上下文信息可以推断该提问者具有实施恐怖主义犯罪倾向时,提供者负有相应的预警义务。对于平台负有反恐怖主义义务在境内外均能找到相应的规范根据。譬如,《中华人民共和国反恐怖主义法》第9条规定,"任何单位和个人都有协助、配合有关部门开展反恐怖主义工作的义务,发现恐怖活动嫌疑或者恐怖活动嫌疑人员的,应当及时向公安机关或者有关部门报告"。此外,德国于2017年颁布实施的网络执行法(Netzwerkdurchsetzungsgesetz)第3条a第2款规定,社交网络提供者对于有迹象表明其实施建立恐怖犯罪组织、境外犯罪与恐怖组织行为的,应当向作为中央机构的联邦刑事警察局传送内容,以便进行刑事追诉。在适用比例原则进行权衡性考虑时,应当对以恐怖主义为代表的严重犯罪对潜在被害群体的影响范围之广泛与恶劣进行慎重考虑,适当限缩民众的言论自由在此时是必要的。同时,需注意不得让平台提供者承担过高的审查义务:一方面,所涉行为一般只处于预备阶段,其法益侵害属性尚不明晰且微弱;另一方面,平台并未对严重犯罪的实施提供帮助,只是替代公安机关进行监管。因此,该义务承担的重心在于识别后的报告义务,而非识别义务。

(三)三级义务:协助配合调查义务

ChatGPT提供者的三级义务,是指当平台用户涉嫌犯罪时,ChatGPT提供者对执法机关予以配合调查的义务。具体表现为,ChatGPT程序本身并不存在任何被犯罪嫌疑人所利用的程序漏洞,同时,犯罪嫌疑人在平台言论并未被识别为二级义务所涉严重犯罪类型情况下,后续有证据表明犯罪嫌疑人的行为与ChatGPT具有某种关联,需要其配合公安机关进行调查,此时,ChatGPT提供者应当协助配合公安机关进行调查。② 《网络安全法》第28条规定:"网络运营者应当为公安机关、国家安全机关依法维护国家安全和侦查犯罪的活动提供技术支持和协助。"此外,2021年发布的《互联网信息服务管理办法(修订草案征求意见稿)》第22条规定:"互联网网络接入、互联网信息服务、域名注册和解析等互联网服务提供者,应当为公安机关、国家安全机关依法维护国家安全和侦查犯罪的活动,提供技术支持和协助。"但是,对于此类行政义务是否上升为刑法上的监管义务应当相当谨慎。虽然ChatGPT提供者配合调查移交相关数据信息在语义上也可以属于信息网络安全管理行为,但应当注意到:其一,《互联网信息服务管理办法(修订草案征求意见稿)》第43条对违反第22条第1款规定的法律责任规定并未涉及到刑事责任,因此,将该条款认定为《刑法》第286条之一的前置条款存在难度;其二,不能认为ChatGPT提供者拒不配合调查的行为对网络管理秩序造成侵害,相较之下,认定该行为破坏司法活动则更为贴切。因此,在符合条件的情况下,可慎重考虑构成窝藏、包庇罪的可能性。

① 参见史令珊:《不作为犯新形态与公民积极义务的限制》,载《法学》2022年第5期,第109页。

② 参见于冲:《网络平台刑事合规的基础、功能与路径》,载《中国刑事法杂志》2019年第6期,第108页。

五、结语

ChatGPT 的诞生是人类迈向通用人工智能研究具有“里程碑”意义的一站。在注意到 ChatGPT 所带来的技术突破的同时,人类应当谨防过分神化 ChatGPT。在弱人工智能时代,ChatGPT 难以真正完成身份转变,成为“我们”的一分子。但在信息网络时代,我们应当深刻意识并重视 ChatGPT 可能扮演的角色。在要求其替代政府履行部分监管职能的同时,也应当牢记“现代刑法观”是主张刑法有限性的刑法观。[①] 申言之,要求其承担刑法上监管义务的同时,也要时刻注意克制赋予其更多监管义务的冲动。这种“博弈”式的权衡过程是提升国家治理能力的必由之路。JS

Study on the Basis of Regulatory Obligations of ChatGPT Providers and the Hierarchical Regulation of Criminal Law

ZHANG Zherui

(School of Law, Southeast University, Nanjing 211189, China)

Abstract: While ChatGPT has shown excellent natural language processing capabilities, it has also been used by some criminals as a new powerful criminal helper. With well-designed questions, ChatGPT can largely satisfy the needs of criminals and generate criminal information for them. Since what ChatGPT reveals is not a real creative thinking, it is basically unable to make value judgments on facts and lacks interpretable behavioral logic, ChatGPT should be regarded as a human-type tool in legal relationships. Based on this, considering that ChatGPT has a large number of users, and that it has the responsibility to govern danger and actively fulfill certain obligations to achieve social welfare, as well as the corresponding gatekeeper ability, it can be given the status of gatekeeper, and its providers can assume part of the regulatory obligations, but at the same time, care should be taken to limit its proportionality. Specifically, according to the role played by ChatGPT and the degree of automatic recognition of users' speech, the first, second and third levels of regulatory obligations should be distinguished from the heaviest to the lightest, i.e., the obligation of organization after the ChatGPT is used as a means of crime, the obligation to warn of serious crimes and the obligation to assist in cooperating with investigations.

Key words: ChatGPT provider; gatekeeper; regulatory obligation; hierarchy of obligation

本文责任编辑:张永强

① 参见刘艳红:《网络时代社会治理的消极刑法观之提倡》,载《清华法学》2022 年第 2 期,第 192 页。