学术篇

ChatGPT应用的风险迭代与法律因应研究

王惠敏[①] 许峰[②] 蔡士林**[③]
①江苏师范大学法学院
②中共浙江省委党校全面从严治党研究中心
③中国矿业大学科技与法律研究中心

摘 要：ChatGPT作为人工智能系统的巅峰之作，极大地改变了人机互动、数据获取的方式，加速了物理空间向虚拟社会迁移的进度。ChatGPT经历了GPT—1至GPT—4的迭代，在算法、数据和技术架构共同作用下其表现出高度的智能性和自主性。新技术通常体现出“危”与“机”并存的特性，ChatGPT隐藏的风险可以归纳为内生性风险和外生性风险，前者包括算法歧视和数据安全问题，后者则包括生成内容知识产权纠纷和ChatGPT异化为犯罪工具的隐患。有鉴于此，应当内外兼修消除ChatGPT中的算法偏见；更新数据生存周期的法律保护体系；引入行政合规制度，预防犯罪滋生；优化既存版权规则，构建多元利益保护体系。

关键词：ChatGPT；人工智能治理；生成式人工智能；内生性风险；外生性风险；数据安全；行政合规

DOI：10.16582/j.cnki.dzzw.2023.12.007

一、问题的提出

ChatGPT（chat generative pre—training transformer，生成式预训练转换器）自2022年推出之后迅速出圈，打破了多项世界纪录[1]。随着技术的迭代，ChatGPT也经历了从GPT—1到GPT—4的进化，它不仅在性能上碾压同类产品，而且可以实现与用户社会生活的无缝衔接，演示真正意义上的“机智过人”。例如，作为科学家的福音，ChatGPT可以轻松实现实验数据收集、文献检索、论文撰写等目标；又如，作为广告公司的利器，ChatGPT能够根据应用场景帮助公司设计宣传海报、图文调色、灯光调试等；再如，作为自媒体时代的法宝，ChatGPT可以加速新闻生成和传播。简言之，未来以ChatGPT为代表的人工智能系统将成为我们生活的一部分，强人工智能时代已经到来。

ChatGPT为用户带来极佳体验的同时，也隐藏着包括算法歧视、数据风险以及知识产权纠纷等副作用。ChatGPT作为新兴技术，不仅关系数字经济的发展，而且与国家安全体系和能力现代化建设息息相关。详言之，一方面，ChatGPT对话系统的高度智能性和真实性标志着弱人工智能时代正在向强人工智能时代过渡，这也意味着它将会创设新的风险。例如，用户利用ChatGPT足不出户就可以生成各种新颖的艺术“作品”，此类“作品”是否具有版权便面临诘难。另一方面，随着训练数据和参数值的增长，人们对于ChatGPT的依赖性也同步递增，如果法律规范缺失，互联网的倍增效应和放射效应会加剧ChatGPT失控引发的社会危害性。例如，算法的不可控风险致使ChatGPT生成内容会根据传播的概率和用户的喜好进行自我调整，甚至脱离人类掌控，如此一来网络谣言便不可避免。

中国的监管机构和立法者也意识到这一点。2023年4月，国家互联网信息办公室发布了《生成式人工智能服务管理办法（征求意见稿）》（以下简称《征求意见稿》）对人工智能系统的规范使用做了部分规定。但《征求意见稿》仅是宣示性规定，且多处内容与既存的《数据安全法》和《个人信息保护法》重叠。基于此，应当结合ChatGPT的特征，对其风险予以类型化，并在

**通讯作者 收稿日期：2023—05—05 修回日期：2023—05—31

此基础上完成法律体系更新与迭代的设计。

二、ChatGPT应用中的类型化风险

作为生成式人工智能的杰作，ChatGPT在应用过程中面临诸多风险，既有研究仅对风险进行简单罗列，缺乏系统性和全面性分析。显然，将这些风险予以类型化，不仅利于人类社会更好地感知新风险并予以防范，而且也可以为立法者和司法者提供参考，有益于未来立法工作。ChatGPT在应用过程中的风险本质上可以分为两类：内生性风险和外生性风险。

（一）内生性风险

所谓内生性风险，是指ChatGPT模型内部构造所蕴含的算法和数据等风险。作为通用大模型，算法、数据以及算力不仅是评估其性能构造的关键性指标，更是其决胜数字市场的法宝。与算法和数据相比，算力可控性强，且主要依托特定的计算机设备，因此受外界干扰最小。而算法和数据无疑成为ChatGPT内生性风险构成的主要因素。

1.算法歧视隐匿其中

算法是指计算机领域为解决某一问题或达到某个目的而采取的方法和步骤[2]。ChatGPT的内部运行逻辑决定了算法歧视不可避免。具体原因可以分为如下三个方面：

其一，数据偏见。ChatGPT的训练通常依赖大量的数据，这些数据可能来自现实世界的样本。如果这些数据存在偏见或不平衡，模型将学习到这些偏见，并可能在生成新数据时重复或放大这些偏见。例如，研究表明使用GANs生成人脸图像时存在种族偏见。实验人员训练了一个GAN模型，使用了大量的人脸图像数据集，但该数据集在种族筛选上存在不平衡。结果显示，图像更倾向于生成白人的面孔，而对于其他种族的面孔生成效果相对较差[3]。

其二，偏差数据集。选择训练数据集时可能存在主观偏见，例如，过度关注某些人群或特定地区的数据，或者数据集中的样本不足以代表整个人群的多样性。这种偏差会导致生成的数据在某些方面存在偏见。

其三，隐式偏见。GANs在生成数据时可能会捕捉到训练数据中存在的隐式偏见。即使数据集本身没有明显的偏见，但模型会通过学习数据中的隐式关联和模式生成具有偏见的新数据。ChatGPT在与性别相关的问题和对话中表现出性别偏见。研究人员发现，ChatGPT的回答更倾向于传统的性别刻板印象。例如，在被问及职业相关问题时，ChatGPT更倾向于将男性与技术和领导角色相关联，将女性与家庭和辅助角色相关联。这种偏见可能是由于模型学习到了训练数据中存在的隐式性别关联和模式，并在生成回答时重复或放大了这种偏见[4]。

由算法歧视带来的风险通常会以隐蔽的方式破坏社会治理结构和制度。具体可以归纳为如下三个方面：

首先，引发社会不公平。一方面，如果ChatGPT算法受到歧视性数据集的训练，它可能生成具有性别、种族、性取向等偏见的文本内容。这些偏见可能会进一步传播和加剧现有的社会偏见，并影响人们对不同群体的态度。例如，在职业推荐或自动回复中，ChatGPT可能更倾向于将男性或未婚女性作为优先项，导致已婚女性在职业机会方面受到不公平待遇[5]。另一方面，ChatGPT算法的应用在决策系统中可能导致不公平的结果。例如，在招聘和招生过程中，ChatGPT可能受到训练数据中存在的职业偏见和种族偏见的影响，从而导致在评估和筛选候选人时存在不公平[6]。这无疑对那些属于少数群体或受到歧视的个体造成不利影响，并加剧社会的不平等。

其次，加剧信息孤立和泡沫化。当ChatGPT用于信息过滤或推荐系统，则可能引入个人化的偏见，从而导致信息的扭曲。ChatGPT过度依赖个人化数据时，它倾向于筛选和推荐与用户过去行为一致的内容。这可能导

致信息的不当过滤和信息源的狭窄化，从而使用户接触到的信息受到限制。如果模型仅传递用户熟悉的观点，则可能错过其他多样性的观点，进而加剧信息的孤立和泡沫化。

最后，导致不公平的决策。ChatGPT对话系统作为提供决策的工具已经成为普遍的现象。例如，在医疗领域，倘若ChatGPT算法在诊断和治疗决策中使用，存在偏见的数据不仅导致对某些群体的错误诊断或不合适治疗，而且会增加弱势群体的健康风险[7]。又如，训练数据中存在性别或种族偏见，并被用于生成决策相关的特征，这可能导致女性或少数群体面临不公平的贷款条件或被拒绝贷款的风险。

2.数据安全风险骤升

20世纪80年代，美国学者沙尔茨（Saltzer）和施罗德（Schroeder）首次提出了数据安全（data safety）的概念。数据安全主要包括数据的保密性（confidentiality）、完整性（integrity）和可用性（availability）[8]。综合各方面的因素来看，ChatGPT运行中数据安全风险主要为非法获取数据。

尽管ChatGPT创建者声称，其爬取的数据来自开源数据库，不会违反数据资源所在网站或平台的规则。然而事实上，已经有数据公司向ChatGPT提出侵权索赔，认为其“未经授权”通过网络爬取的手段非法获取平台上的数据[9]。实际上，ChatGPT创设非法获取数据的风险，主要原因有以下几点：

其一，语料库所需数据规模庞大，ChatGPT爬取数据的行为不可能全部获得授权。GPT-3的出现使语言处理有了显著的发展，它也成为OpenAI迄今为止最复杂和最广泛的NLP（natural language processing）模型。GPT-3中有超过1750亿个“参数”和“变量”被语言处理引擎使用。研究表明，GPT-4的参数是GPT-3的500倍，约为100万亿。这背后大量的训练数据需要ChatGPT不间断进行数据爬取，而其中就难免会导致“未经授权”和“超越授权”获取数据的情形。例如在并未获得任何授权的情况下，ChatGPT突破对方计算机信息系统设定的防火墙，访问并获取数据；又如超越授权范围，违规获取对方数据库中的数据；再如用户在与ChatGPT互动时，输入的数据也会成为数据训练的对象，用于提升ChatGPT的智能性。正如有学者所言，ChatGPT表明人机无差别交流的理想并不遥远，但非法获取数据的危机也是前所未有的，因为它不可能短时间获得大量的数据许可[10]。

其二，国家间数据获取的法律规定存在显著差异，ChatGPT无法全部满足。世界范围内对于数据获取的规定基本上呈现出三种模式：欧美模式、韩国模式以及中国模式。它们之间存在较大差异，这导致ChatGPT在进行数据获取的过程中必然难以把握。以公开的个人数据为例，欧美倾向于以客观技术作为合法性判定依据；韩国则将主观目的作为合法性考察的基本工具；而我国则兼顾主观目的与客观技术。此外，当前由于数据发掘对象的数据语言主要集中在英语（约占48%），而俄语、德语、日语分别占到10%、7%、6%，中文占比不到5%[11]。因此ChatGPT遵守的法律主要以英语语系的法律为主。但随着语料库数据规模的累加，各个数据语言的比例也会发生显著变化，这意味着ChatGPT必然会出现数据获取依据的动摇。

其三，ChatGPT与其他应用程序的融合，加剧了非法获取数据的风险。ChatGPT本身通过爬虫建构的语料库中就包含路径不明的数据，而后续与其他应用程序的融合将会导致用户在数据存储过程中遭受窃取的风险。例如，为了扩大受众人群，ChatGPT宣布将其导入办公软件WPS和搜索引擎Google，开启全球化布局。WPS作为全球用户最多的办公软件，覆盖包括Windows、IOS、Mac等不同操作系统终端，每天用于创建、编辑和存储不同类型文档。一旦ChatGPT接入，则可能对WPS中的数据进行访问和获取，进而引发数据安全危机。正

如ChatGPT创始人萨姆·奥特曼（Sam Altman）所言，ChatGPT存在重大技术漏洞，造成数据库使用中存在不当获取数据的现象，目前正在修复[12]。

此外，与非法获取数据相关联的两类行为有必要在此一并提出，即非法访问数据和数据滥用行为。非法访问数据是指行为人未经授权或超越授权登入系统成功后，对数据库进行查阅的行为[13]。不可否认，非法访问数据行为会侵害数据的保密性和有用性，但并非所有未经授权或超越授权的访问行为都需要规制。当前，数据持有者凭借技术和市场的双重垄断地位，不断挤压用户的合法访问空间，故而有必要扩大享有数据访问权的主体。与此同时，倘若非法获取数据的风险未得到有效控制和治理，后续也会引发数据滥用风险。与非法获取数据相比，ChatGPT数据滥用行为的法益侵害更为严重。数据滥用行为严重侵害数据的保密性和有用性，不仅使得原有数据的价值大打折扣，而且可能危及个人、社会甚至是国家利益，因此有必要建构包括民法、行政法以及刑法在内的规制体系。

（二）外生性风险

所谓外生性风险是指ChatGPT作为一种技术性工具，在外部运行过程中所导致的新型风险或增加了风险系数。ChatGPT的外生性风险主要包括两类：犯罪工具异化和知识产权纠纷。

1.ChatGPT异化为犯罪工具

ChatGPT作为强大的语言模型，可以用于生成文本、回答问题和模仿人类写作。它具备高度智能性和自主性，因此一旦被不当利用，可能异化为犯罪工具，不仅危及个人、社会甚至国家安全，而且极具隐蔽性，这显然会给治理工作增加难度。

(1)不法分子利用ChatGPT进行网络诈骗

随着ChatGPT技术的不断发展，人工智能在生成虚假的电子邮件、社交媒体帖子或网站内容方面具备了强大的潜力。这种潜在的滥用使得欺诈个人或组织成本大幅降低。利用GPT实施网络诈骗的方式通常有两种。

其一，钓鱼攻击。ChatGPT可能用于生成逼真的钓鱼电子邮件，让受害者误以为它们来自可信的实体，如银行、电子支付服务或社交媒体平台。这些虚假邮件可能请求受害者提供个人敏感信息，如登录凭据、账户号码或信用卡信息。用户一旦打开这些邮件便会导致身份盗窃、财物损失以及个人隐私泄露。由于ChatGPT技术可以生成逼真的欺骗性文本，使得识别和防范这类攻击变得异常困难。例如，研究表明利用GPT-3生成的钓鱼邮件在被测试者中的成功率高达45%，GPT-4飙升至81%[14]。

其二，生成虚假广告。不法分子可能利用ChatGPT生成的文本制作虚假广告，宣传欺诈性产品或服务。通过利用ChatGPT生成的广告可能具备逼真的外观，误导消费者并骗取他们的财物，同时破坏市场公平竞争，甚至给品牌声誉带来负面影响。例如，投资理财广告可能声称能够获得高额回报，吸引人们投资于欺诈性计划或陷入庞氏骗局中；又如，医疗广告可能宣称能够治愈一些严重疾病或提供不实的保健效果，以吸引消费者购买无效的药物或保健产品[15]。

(2)不法分子利用ChatGPT进行政治操纵

其一，ChatGPT一旦被操纵，可能生成虚假新闻，破坏社会稳定。在当今假新闻和不实信息泛滥的时代，缺乏透明度和可靠性成为国家安全的重要危险源。毫不夸张地说，ChatGPT将假新闻提升到了一个前所未有的水平。恶意用户可能利用这种技术制造假新闻、扭曲事实，甚至伪造来源和证据。这种操纵行为可能导致谣言的传播、信息的混乱和公众对媒体的信任受损，进而削弱民主社会中的信息可靠性和公众的参与度。

其二，ChatGPT一旦被利用，可能生成虚假的政治言论，进而实现操纵公众对特定政治议题态度的非法目的。这些虚假言论可能用于煽动仇恨、制造社会分裂或误导选民。通过广泛传播这些虚假言论，不法分子试图

影响选民的决策，干扰选举过程，进而削弱民主制度的稳定性[16]。

其三，恶意用户可能通过利用ChatGPT技术生成虚假的社交媒体帖子，编造虚假舆论、人工增加关注度或引导讨论的方向。这种操纵行为可能导致社交媒体上的信息战，使真实和虚假的信息难以辨别。如此一来，公众被误导和欺骗，可能会出现社会动荡和不信任的情绪蔓延，造成对社交媒体平台稳定性和可信度的冲击[17]。

(3)不法分子利用ChatGPT进行黑客攻击

ChatGPT的可拓展性等特征给用户使用带来便利的同时，也为不法分子进行黑客攻击提供了有利的条件。一方面，不法分子可能使用ChatGPT生成的文本制作误导性的软件安装程序、电子邮件附件或下载链接。这些内容被伪装成合法的文件或应用程序，但实际上含有恶意软件，例如间谍软件、勒索软件或远程访问工具。用户一旦使用便可能遭遇系统感染、数据泄露以及个人隐私被侵犯。另一方面，不法分子利用ChatGPT生成的工具和脚本，可以进行密码破解和暴力攻击。他们可以使用自动生成的强大字典或暴力攻击程序来尝试破解密码，获取对目标账户的未经授权访问权。这可能导致账户被入侵、个人信息被盗取，并进一步引发金融损失或身份盗窃。

2.生成内容的知识产权问题突出

人工智能生成内容的知识产权问题由来已久，但ChatGPT强大的语料库以及算法的加持，使得该问题变得更加复杂。相较于ChatGPT，早期的人工智能机器人处于弱人工智能状态，基本上承接体力劳动、简单的语言文字互动，因此绝大多数情况下被视为“人类”的附庸。申言之，一般的人工智能生成内容不具备“作品”的属性。诚如有学者所言，驱动人工智能生成的是算法程序，无法真实体现人类创作的个性，故而不能成为著作权法意义上的作品[18]。

ChatGPT的出现使原先的假命题变成了真命题。一方面，ChatGPT利用超大规模参数，具备与人类同量级的“思考”能力。GPT4.0版本的参数值接近100万亿，同时还加入了基于规则的奖励模型(rule-based reward models，RBRMs)，实现了真正意义上的“机智过人”。RBRMs是一组零样本（zero-shot）的GPT-4分类器。这些分类器在RLHF针对正确行为进行微调期间向GPT-4策略模型提供额外的奖励信号。例如，GPT-4可以根据图片或场景编写故事、诗歌，甚至学术论文。另一方面，与其他人工智能系统相比，ChatGPT具有一定的自主性。尽管ChatGPT的运行离不开数据和算法，但其生成内容不是网络内容的简单叠加，而是有逻辑的组合，这也是它出圈的重要原因。基于此，ChatGPT生成内容面临着三个方面的诘问。

(1)ChatGPT生成内容的版权问题

ChatGPT具有一定的独立创作能力，因此生成内容的版权问题便不可避免。根据《著作权法》规定可知，作品是指文学、艺术和科学领域内具有独创性并能以一定形式表现的智力成果。新版《学术论文编写规则》（GB/T 7713.2—2022）将学术论文定义为：对某个学科领域中的问题进行研究后，记录科学研究的过程、方法及结果，用于进行学术交流、讨论或出版发表，或用作其他用途的书面材料。显然，上述规定都将独创性作为判断作品的根本依据。应当说，ChatGPT所输出的内容并不是简单的数据汇总结果，而是经过了复杂信息抽取、语义识别以及转码等过程。这与著作权领域中的改编行为具有同质性。关键问题在于，并非用户创作作品的全部内容都依赖于ChatGPT，可能仅将其作为一种统计工具或验证手段，此时如何确定版权成为新问题。

(2)ChatGPT生成内容的版权归属问题

ChatGPT生成内容归属问题纠纷的主要原因有如下两点：

其一，ChatGPT与用户互动过程就是内容生成的过程，而这里牵涉三方主体（ChatGPT的创设者、技术开

发者以及用户），因此生成内容的利益归属不易澄清。有观点认为，ChatGPT生成内容可以参考职务作品或者雇佣作品之规定，将生成内容利益归属于投资者进而平衡各方关系[19]。应当说，这种版权归属方式仅考虑经济学上的生成与收益，将用户的输入行为视为一种生产行为，遮蔽了它的收益权利。

其二，ChatGPT用户数量众多，重复性或类似性问题输入自然会生成高度相似的内容，此时利益归属成为难点。

(3)使用ChatGPT生成内容风险的责任承担问题

使用ChatGPT生成内容的主要风险可以分为如下两类：

其一，侵犯复制权、信息网络传播权等著作财产权的风险。此种风险的形成主要源于数据获取路径的不透明。例如，当用户与ChatGPT进行对话的过程中，生成内容是将外来数据源作为依托，不仅数量众多且数据结构多元，难以完全实现访问路径的全部授权，故而存在未经授权或超越授权爬取数据的情形。在此基础上，无论是用户使用ChatGPT直接生成的内容，抑或将生成内容在网络上进行传播都可能侵犯版权人的复制权等著作财产权。

其二，侵害作品完整权、署名权等著作人身权。GPT—4新增的图文语义识别功能使得对原有作品的完整性产生了冲击，例如，ChatGPT不仅通过结构化数据的分解和重新组合生成新的内容，而且可以在原有内容上进行二次创作。这都在一定程度上侵蚀了原有作品的完整性和署名权等人身性权利。

三、ChatGPT应用风险的法律应对

在ChatGPT风靡全球之际，中国的互联网巨头也纷纷加入了此次科技较量，类ChatGPT的大型语言模型设计将会迎来一波高峰，这无疑对中国法律制度提出了新的要求。应当从以下几个方面进行完善，保证智能技术的有序和健康发展。

（一）内外兼修消除ChatGPT中的算法偏见

ChatGPT在训练过程中，算法模型可能从数据中学习到偏见或歧视性，导致输出结果中存在偏见。主要原因是训练数据的不平衡或偏向性，以及模型的特征选择和算法设计等。与此同时，ChatGPT在应用过程中，算法模型也可能受到用户输入的偏见影响，导致输出结果偏向某些特定观点或群体。需要说明的是，后续ChatGPT使用过程中算法偏见的形成主要受到前期算法模型的影响，因为理想状态下的算法模型具有检测功能，可以自我纠偏。基于此，可以从以下两个方面弱化算法歧视的不良影响。

1.通过算法模型学习路径的调适进行监管

针对ChatGPT中先天的算法歧视，可以通过算法模型学习路径的调适进行监管。先天性算法偏见指的是机器学习模型在训练过程中可能从数据中学习到的偏见或歧视性[20]。早在2021年12月31日国家互联网信息办公室等多部门颁布的《互联网信息服务算法推荐管理规定》第8条就规定，算法推荐服务提供者应当定期审核、评估、验证算法机制机理、模型、数据和应用结果等，不得设置诱导用户沉迷、过度消费等违反法律法规或者违背伦理道德的算法模型。2023年《征求意见稿》的第4条也规定，在算法设计、训练数据选择、模型生成和优化、提供服务过程中，采取措施防止出现种族、民族、信仰、国别、地域、性别、年龄、职业等歧视。然而，上述规范性文件仅对算法偏见解决提供宏观指导，并未予以具体落实和细化。

故而，一方面针对ChatGPT中的算法模型需要以规范性文件的形式予以细化，发布明确的技术标准，防止算法中立性的偏离。这些规范文件需要明确要求算法模型在训练过程中避免学习到偏见，或者对已学习到的偏见进行修正和纠正。例如，明确要求数据收集过程中的多样性和代表性，避免偏向某一特定群体或观点。同

时，需要制定数据处理的规则，如去除偏见或不必要的特征，以减少潜在的偏见源。又如，对ChatGPT进行定期的评估和纠正，以确保其输出结果的公正性和中立性。这包括对模型输出进行人工审查、开展对抗性评估、参考外部审核机构的意见等。再如，明确要求建立用户反馈机制，鼓励用户报告任何发现的偏见或歧视问题。这样可以及时发现并纠正模型中的偏见，并对用户进行适当的回应和补偿。

另一方面，在规范文件制定完成后，需要对ChatGPT进行实质审查。这一过程需要对算法模型的内部机制、训练数据和输出结果进行审查和评估，以验证其是否符合规范文件中规定的技术标准。实质审查可以采用多种方法，包括模型测试、对抗性评估、人工审查和用户反馈等。通过这些手段，可以发现和修正潜在的偏见问题，确保ChatGPT在市场应用中的公正性和可靠性。

2.构建企业自治与政府监管相结合的动态治理体系

针对ChatGPT在运行过程中存在的算法歧视，应当构建企业自治与政府监管相结合的动态治理体系，提升算法的透明性和可解释性。算法建立在大数据基础上，具有非直观性和隐蔽性，容易形成“黑箱”，大数据资源本身也存在是否完整、可靠、及时更新，以及数据采集、存储是否合法合规等问题，因此算法应当透明和可解释[21]。

一方面，科技企业作为技术的探索者和风险的经历者自然需要承担风险规制的职责，减少算法歧视的风险。其一，科技企业应向用户和利益相关者提供算法系统的运行逻辑、价值取向以及危险防范手段等。可以通过公开算法的工作原理、数据处理方式和决策逻辑，以及为用户提供可理解的解释和依据来实现。透明度和可解释性的机制可以增加对算法运行过程的信任，并让用户了解模型输出的依据和可能的偏见。其二，企业应积极推行多元参与和反馈机制，与用户、利益相关者和社会各界进行广泛的对话和合作。这可以通过定期收集用户反馈、组织独立审查和评估等方式实现。反馈机制可以及时发现和解决算法歧视问题，并促使企业改进算法模型，从而减少偏见。

另一方面，政府在构建动态治理体系中扮演重要角色，通过监管和规范来确保算法应用的公正性和非歧视性。政府应设立专门的监管机构或加强现有机构的能力，负责监督和审查算法应用的合规性。监管机构可以进行定期的审查和评估，要求企业提交相关数据和报告，确保算法应用符合法律法规和标准要求。监管机构还应加强对算法歧视问题的调查和处罚能力，确保违规行为得到及时处理。与此同时，政府应建立制裁和追责机制，对违反算法公正性和非歧视性要求的企业进行惩罚和追责，例如通过罚款、吊销许可证等手段来实施。制裁和追责机制的建立可以增加企业对算法歧视问题的重视，促使其更加注重公正和平等的算法应用。

（二）更新数据生存周期的法律保护体系

一方面，ChatGPT应用过程中出现了数据安全风险与知识产权纠纷等问题；另一方面，中国包括“文心一言”（百度）、“通义千问”（阿里巴巴）以及“混元”（腾讯）等大语言模型的横空出世都要求立法者围绕数据生存周期对既存法律体系进行更新和迭代。具体可以从三个方面着手。

1.在特定领域确认数据访问权

除了数据持有者或者数据主体对数据享有使用权之外，其他主体是否享有数据使用权便成为ChatGPT等人工智能系统面临的难题。我国《数据二十条》提出：“充分保护数据来源者合法权益……保障数据来源者享有获取或复制转移由其促成产生数据的权益。”这实际上为我国数据访问权的确立提供了指引。2022年，欧盟颁布了《数据法案（草案）》（Data Act），首次提出了横向的数据访问权概念，主要指对于使用产品或服务中产生的数据，用户有权向数据持有者请求直接访问，

或者请求数据持有者向第三方传输[22]。ChatGPT作为超大规模的语言模型，其网络效应影响深刻，作为数据的持有者具有支配市场的作用，因此有必要要求其强制共享数据。创设数据访问权旨在平衡数据参与者的价值分配，基于此国家不宜通过立法全面确认数据访问权，而是应当首先在特定行业许可设计该权利。

2.完善非法获取计算机信息系统数据罪

针对ChatGPT应用过程中的数据安全风险，中国的《数据安全法》和《个人信息保护法》都强调数据获取方式和目的之合法性。例如，《数据安全法》第三十二条规定，任何组织、个人收集数据，应当采取合法、正当的方式，不得窃取或者以其他非法手段获取数据；第五十一条规定，窃取或者以其他非法方式获取数据、开展数据处理活动排除、限制竞争，或者损害个人、组织合法权益的，依照有关法律、行政法规处罚。为了从源头实现对类ChatGPT大型语言模型的治理，刑法也需要对数据源的合法性予以监督，这自然离不开对于非法获取计算机信息系统数据罪的审视。遗憾的是，该罪将国家事务、国防建设、尖端科学技术领域的计算机信息系统数据排除在刑法保护对象之外。这不仅导致该罪适用率低，且与其他罪名难以形成合力。实际上该罪自设立以来，近五年平均每年适用不到100件的客观事实表明，由于构成要件的非理性限制，致使本罪难以发挥应有的刑法效能。数据安全作为独立保护的刑法法益，需要构建数据犯罪治理体系，而其中关键在于修补罪名之间的裂痕。例如，非法侵入计算机信息系统罪保护的对象是特定领域的计算机信息系统，而非法获取计算机信息系统数据罪则恰好排除了相应领域的数据，因此会出现一个怪现象：非法侵入特定领域计算机信息系统获取关键性数据进行使用或出售，却仅能以非法侵入计算机信息系统罪定性，处以三年以下有期徒刑或拘役。建议取消刑法中“侵入前款规定以外”限制性构成要件，将一般数据纳入非法获取计算机信息系统数据罪保护的对象，同时将特定领域的系统数据作为加重情节。

3.非法使用数据行为予以犯罪化

ChatGPT的强智能性离不开大量数据，因此如何使用数据成为ChatGPT迅速迭代与更新的关键。从表面上看，如上文所述，ChatGPT在应用过程中会出现数据泄露等风险，据韩国媒体报道，三星导入ChatGPT不到20天，便曝出机密资料外泄。面对ChatGPT存在非法使用数据的情形，部分国家禁止本国公民使用，譬如2023年3月31日，意大利个人数据保护局（Garante）宣布，从即日起禁止使用聊天机器人ChatGPT，并限制OpenAI处理意大利用户信息。从本质上看，ChatGPT存在数据滥用的风险侵害了数据主体以及数据处理者两方的权益。具体而言，ChatGPT数据滥用的行为侵犯了数据主体的数据隐秘性和有用性，不仅直接导致其数据自决权受损，而且间接造成数据价值的减损。对数据处理者而言，此种行为侵犯了其数据使用权和收益权。显然，中国立法者已经意识到这一点，因此在《数据安全法》《个人信息保护法》以及《数据“二十条”》中都有所体现。例如《数据“二十条”》规定，统筹发展和安全，贯彻总体国家安全观，强化数据安全保障体系建设，将安全贯穿数据供给、流通、使用全过程，划定监管底线和红线。遗憾的是，作为数据保护的《刑法》却仅对数据的获取、盗取以及买卖等行为予以规制，缺失数据使用的保护板块。基于以上原因，有必要将非法使用数据，情节严重的行为予以定罪。

（三）引入行政合规制度，预防犯罪滋生

涉案企业合规制度是以风险预防为导向的公司治理制度，强调“事前预防”，而非“事后惩罚”。以ChatGPT为代表的人工智能系统潜藏着生成或传播网络谣言的高度风险，因此有必要通过合规制度的构建予以化解。遗憾的是，当前的合规制度研究基本上集中在刑事领域，不仅造成了治理上的延迟性，而且还存在诸多制度性障碍难以克服。因此，与其等到ChatGPT等智能

体实施刑事犯罪后再追究相关主体的责任，倒不如在出现行政违法违规的“苗头”时，就及时进行引导和预防性监管。可以从日常企业合规建设与违法企业合规激励机制两个层面展开。

1.日常企业合规建设

从日常合规建设角度来看，需要强化ChatGPT的数据安全管理义务、数据清洗义务，同时充分保障用户的数据删除权。上文中提及由于ChatGPT的运行机制和生成原理，容易引发网络谣言，危及公民、社会甚至是国家安全。因此，作为风险的创造者，OpenAI公司有能力也有义务对平台的违法犯罪行为予以监督和管理。这一点可以在我国刑法设置的“拒不履行信息网络安全管理义务罪”中得到印证。需要说明的是，在刑事合规研究的浪潮下，学界忽视了行政合规的必要性与正当性。作为行政监管机关需要发布相应合规指引，为类似OpenAI这样的公司日常合规建设提供指引。此类指南应该涉及以下几个方面的内容：

其一，应当要求企业对获取数据进行必要的清洗。网络谣言的生成并非一蹴而就，而是在传播过程中数据错误交叉感染和自我迭代的结果，因此数据企业需要通过必要的过滤，阻断谣言生成的链条。例如，可以通过算法审查或人工验证相结合的方式。

其二，企业应当承担必要的内容审查义务。对于ChatGPT生成内容，除了企业进行包括涉及意识形态领域安全的审核之外，种族歧视、性别歧视等内容也应当予以审查。例如，我国《征求意见稿》第4条规定，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，不得含有颠覆国家政权、推翻社会主义制度，煽动分裂国家、破坏国家统一，宣扬恐怖主义、极端主义，宣扬民族仇恨、民族歧视，暴力、淫秽色情信息，虚假信息，以及可能扰乱经济秩序和社会秩序的内容。

其三，为了减轻数据企业的审查负担，应当赋予用户删除权。当用户向平台提出信息虚假或过时之后，且对个人或公共利益造成不利影响时，应当在核实后予以删除。例如《征求意见稿》第13条规定，提供者应当建立用户投诉接收处理机制，及时处置个人关于更正、删除、屏蔽其个人信息的请求。

2.行政合规的激励角度

从行政合规的激励角度来看，对于具有合规意愿且整改合格的企业应当予以从宽处理。可以扩大行政和解的适用范围，给予企业知错就改的机会。与刑事合规不同，行政合规一般通过和解的形式，而作为涉案的数据企业则应当自主改正相关违法行为，消除不良后果。这一做法在欧美等合规制度相对完善的国家已经得到广泛适用。例如，美国证券交易委员会对涉案企业基本上借助行政和解予以处理。而我国仅在2015年《行政和解试点实施办法》中首次在证券违法行政执法中引入和解协议制度。考虑到OpenAI公司引发的人工智能系统研发热潮，中国无论作为使用国或者研发国都应当积极扩大行政合规的适用。此举不仅可以激励企业进行自查，将人工智能系统引发的风险降到最低；而且还能够帮助中国企业培育良好的法治文化，积极参与全球化的竞争。

（四）优化既存版权规则，构建多元利益保护体系

以ChatGPT为代表的生成式人工智能已经成为当下人工智能技术的主流领域之一，在该技术的帮助下生成内容的创新性以及利益归属问题已然不可回避。基于此，需要优化既存的版权规则，构建多元利益保护体系。具体可以分为作品认定、版权归属和侵权责任承担三个方面。

1.作品认定：根据ChatGPT作用适度赋权

作为典型的生成式AI，ChatGPT产生内容的版权属性应当根据具体情况区别看待。一切制度建立都需要以时代为前提。绝大多数学者都强调，当前阶段ChatGPT生产内容本质上是人类创作作品，因此应当认定其作品属性[23]。这个观点值得商榷。ChatGPT作为一种对话系统，与人类之间的互动关系皆是通过数据来完成。而数

据共享性的特征恰好与人类创作作品的逻辑相悖，因此应当根据ChatGPT的工具属性来区别看待。

其一，当用户仅将ChatGPT作为辅助工具，整个生成内容主要依赖用户。例如，根据提供的设计图建构模拟建筑物模型；又如，依照指示对实验模型的安全性与正确性检测，在此语境下，ChatGPT本质上与纸笔、树枝等工具无异。因此，生成内容符合知识产权要求，则可以认定为作品，具有可版权性。

其二，当用户将ChatGPT作为主要工具，生成内容不能成为作品，不应被赋予版权。此种场景下，ChatGPT生成的内容并非人的创造成果，人工智能也不可能受到版权法的激励，因此不可能属于受版权法保护的作品[24]。主要原因有两点：一方面，在此情状下，生成内容主要由ChatGPT主导，人类仅提供引导或提示等边缘性工作；另一方面，考虑到数据共享产生价值的天然属性，即便RBRMs融入ChatGPT中，体现出人类的主观偏好与选择，但整体上依旧是一种技术创新，而非人类个体的创新。

2.版权归属：使用者利益为主，兼顾投资者利益

ChatGPT不仅是单纯的对话系统或智能聊天机器人，也可以通过生成内容创设一系列作品，因此需要明确各主体之间的权利归属，这也是促进生成式AI作品持续繁荣发展的重要条件。

其一，针对ChatGPT生成的内容，应当将使用者视为主要利益体。使用者利用ChatGPT来获取信息、解决问题或进行创作，因此他们应该享有对由ChatGPT生成的内容的权益，包括使用、修改和传播该内容的权利[25]。这些生成内容对于使用者具有重要的商业意义。作为利益主体，使用者享有对生成内容的自由使用权。他们有权使用生成内容来满足个人或商业需求，包括将内容整合到自己的作品、产品或服务中。使用者有权将生成的内容转发给其他人，包括通过社交媒体、网站、博客或其他途径共享生成内容。使用者可能希望将生成内容用于教育、研究、评论、新闻报道或其他公共利益目的，因此他们有权将生成内容传播给其他人，并为这些内容的传播承担相应的责任。

其二，投资者的利益也需要得到保护。投资者为ChatGPT的研发和训练提供了资金和资源，他们在技术研究、开发和市场推广方面做出了重要贡献，因此有权享有投入资本所带来的回报。投资者在ChatGPT的研发和训练过程中承担了重大风险和成本。他们需要投入大量的资金、人力资源和时间，以开发和改进ChatGPT的性能和效果。投资者还承担了市场推广和商业化的责任，以确保ChatGPT能够盈利和持续发展。在这个过程中，他们通过实质贡献换来数字经济的持续发展，故而其商业利益应当得到保护。有鉴于此，ChatGPT使用者和投资者可以通过合同的方式明确规定双方的权利和义务，以及对生成内容的使用和传播的约束条件。合同还应该规定投资者对ChatGPT技术和相关知识产权的所有权，并确保他们享有适当的权益。

3.侵权责任承担：开发者和运营者担责为原则

任何新技术的诞生都会衍生新的风险，而这也意味着原来的责任主体结构会发生改变。ChatGPT的高度智能性与人类活动的复杂性使得有必要根据生成内容的侵权类型进行责任主体的核定[26]。既有研究仅提出由ChatGPT使用者和设计者承担后续的侵权责任，但对于如何区别两者的责任边界并未作出有效回应。使用ChatGPT生成内容涉及的风险责任应该根据具体情况而定，分别由使用者和GPT（ChatGPT的开发者和运营者）承担。

其一，涉及生成内容的准确性和合法性一般应该由ChatGPT的开发者和运营者负责。ChatGPT作为一个自然语言处理模型，基于训练数据生成内容。不可否认，ChatGPT无法对生成内容的准确性做绝对保证，也无法自行判断内容是否符合适用的法律法规。因此，在使用生成内容时，用户应当对内容进行必要验证与核实，确

保其准确性和合法性。然而，相较于ChatGPT的开发者和运营者，用户对生成内容合法性验证能力较弱。例如，ChatGPT生成内容所爬取的数据是跨境获得，则不可能要求用户甄别其中的准确性与合法性情状。基于此，生成内容的合法性应当由开发者和运营者负责。ChatGPT的开发者和运营者需要在最大限度内提供准确且可靠的训练数据和更新模型，以降低错误和不准确内容的风险。需要说明的是，当开发者和运营商有证据证明用户在使用ChatGPT过程中，故意误导生成不准确或侵权内容，则应当由使用者承担侵权责任。例如，当ChatGPT已经提醒生成内容涉及侵犯知识产权可能性时，用户依旧进行指令输入；又如，用户将他人作品内容截取后要求ChatGPT进行续写。

其二，生成内容涉及他人隐私和数据安全时，一般应由开发者和运营者承担侵权责任，用户只对扩大的损害结果承担责任。使用ChatGPT可能涉及个人数据的处理，在这种情况下，使用者和ChatGPT的开发者和运营者都有责任确保符合隐私和数据保护的法律要求。使用者应妥善处理和保护个人数据，并遵守所在国家的政策和法律要求。ChatGPT的开发者和运营者应提供明确的隐私政策和合规措施，保护使用者个人数据的安全和隐私。正如上文所述，如果使用者对数据是否获得授权、是否超越授权以及何时结束授权的具体内容并不知情，则应当推定ChatGPT所提供的数据已经获得合法授权，因而在不违背著作权规定的情况下，生成内容应当受法律保护。在这个情形下，一旦发生侵权，仍然应当由开发者和运营者承担责任。考虑到用户身处数据应用的场景之中，具有一定的违法识别能力，因而当对生成内容存疑时，有义务通知平台进行确认，这样可以避免损害结果的扩大化，否则应当对此承担部分责任。

四、结语

如果说人工智能时代的开启为人与智能机器的对立提供了可能，那么ChatGPT的出现则正在将这种可能变成现实。当然，人工智能时代各种出圈的新科技都是“危”与“机”并存，ChatGPT也不例外。

目前，我国理论界对ChatGPT等生成式人工智能的研究可谓“遍地开花”，但主要集中于技术和风险两大议题。作为人文学者，重点在于后者，但绝大多数研究仅简单列举风险，缺乏逻辑性和系统性。此举自然无法还原风险的原貌，更谈不上提供合理对策。实际上，ChatGPT隐藏的风险可以分为内生性和外生性两种。前者附着在ChatGPT内部，属于先天性风险，诸如算法歧视和数据安全问题；后者发生在用户使用ChatGPT的过程中，例如生成内容的知识产权问题和不当使用问题等。

为此，立法者需要从多个方面发力，协同共治，保障ChatGPT等生成式人工智能的合规使用，助力数字经济发展。在算法方面，企业自治与政府规制相结合，进行算法纠偏，最大限度保证算法的透明性、中立性和可解释性。在数据安全方面，根据数据生存周期，构建全流程的法律保护体系。在不当利用问题上，通过引入行政合规，预防犯罪滋生，减少ChatGPT异化为犯罪工具的概率。在生成内容方面，根据ChatGPT的作用适度赋权，利益分配时注重兼顾多方主体，结合风险类型进行责任设定。未来，中国势必会经历生成式人工智能的技术竞赛，而上述问题的持续探讨显然可以为立法工作的深入推进提供有益参考。

参考文献：

[1]Stolel-Walker C. AI bot ChatGPT writes smart essays — should professors worry?[EB/OL]. (2022-12-09)[2023-04-02]. https://www.nature.com/articles/d41586-022-04397-7.

[2]弗莱 H. 算法统治世界[M]. 李英松，译. 贵阳：贵州人民出版社，2021：10.

[3]Zhu J Y, Krähenb ü hl P, Shechtman E, et al. Generative

visual manipulation on the natural image manifold[C/OL]. The 14th European Conference on Computer Vision, 2016, Amsterdam, Netherlands. [2023–06–20]. https://arxiv.org/abs/1609.03552.

[4]Van Dijck J. 连接：社交媒体批评史[M]. 晏青，陈光凤，译. 北京：中国人民大学出版社，2021：39.

[5]Obermeyer Z，Powers B，Vogeli C, et al. Dissecting racial bias in an algorithm used to manage the health of populations[J]. Science, 2019, 366(6464): 447–453.

[6]赵婵. AI招聘的算法歧视风险与治理之道[J]. 湘潭大学学报：哲学社会科学版，2023(03): 96–102.

[7]陈潭，刘璇. 智能政务ChatGPT化的前景与隐忧[J]. 电子政务，2023(04): 36–44.

[8]蔡士林. 我国数据安全法益保护：域外经验与立法路径[J]. 深圳大学学报：人文社会科学版，2022(06): 97–106.

[9]王惠敏. 我国数据犯罪治理的困境与出路[J]. 北方法学，2023，17(01): 122–132.

[10]韩博. ChatGPT引发的人工智能内容生产传播风险[N]. 中国社会科学报，2023–02–16(03).

[11]Mullani N K, Mali V A . Artificial intelligence and machine learning: Current developments and applications[J]. Journal of Critical Reviews, 2023, 10(01): 85–96.

[12]Przemyslaw P. Data management law for the 2020s: The lost origins and the new needs[J]. Buffalo Law Review, 2020, 68(02): 559–640.

[13]付永贵，朱建明. 基于区块链的数据库访问控制机制设计[J]. 通信学报，2020，41(05): 130–140.

[14]蔡士林，杨磊. ChatGPT智能机器人应用的风险与协同治理研究[J]. 情报理论与实践，2023，46(05): 14–22.

[15]舒洪水，彭鹏. ChatGPT场景下虚假信息的法律风险与对策[J]. 新疆师范大学学报：哲学社会科学版，2023，44(05): 124–129.

[16]Allcott H， Gentzkow M. Social media and fake news in the 2016 election[J]. Journal of Economic Perspectives, 2017, 31(02): 211–236.

[17]Guess A M，Nyhan B，Reifler J. Exposure to untrustworthy websites in the 2016 US election[J]. Nature Human Behaviour, 2020, 4: 472–480. https://doi.org/10.1038/s41562–020–0833–x.

[18]王迁. 论人工智能生成的内容在著作权法中的定性[J]. 法律科学（西北政法大学学报），2017，35(05): 148–155.

[19]吴汉东. 人工智能时代的制度安排与法律规制[J]. 法律科学（西北政法大学学报），2017，35(05): 128–136.

[20]张欣. 生成式人工智能的算法治理挑战与治理型监管[J]. 现代法学，2023，45(03): 108–123.

[21]田思路. 技术从属性下雇主的算法权力与法律规制[J]. 法学研究，2022，44(06): 132–150.

[22]李依怡. 论企业数据流通制度的体系构建[J]. 环球法律评论，2023，45(02): 146–159.

[23]丛立先，李泳霖. 聊天机器人生成内容的版权风险及其治理———以ChatGPT的应用场景为视角[J]. 中国出版，2023(05): 16–21.

[24]王迁. ChatGPT生成的内容受著作权法保护吗？[J]. 探索与争鸣，2023(03): 17–20.

[25]杨利华. 人工智能生成技术方案的可专利性及其制度因应[J]. 中外法学，2023，35(02): 346–364.

[26]高奇琦，张蓥文. 主体弥散化与主体责任的终结：ChatGPT对全球安全实践的影响[J]. 国际安全研究，2023，41(03): 3–27，157.

作者简介：

王惠敏（1989—），女，汉族，河南安阳人，江苏师范大学法学院讲师，中国法学会暨中国社会科学院联合培养博士后，研究方向为数据法学、比较法学。

许峰（1966—），男，汉族，浙江天台人，浙江省重点新型智库、中共浙江省委党校全面从严治党研究中心研究员，研究方向为网络政治学、电子政务。

蔡士林（1989—），男，汉族，中国矿业大学科技与法律研究中心副研究员，研究方向为数据合规、科技法。