生成式人工智能的刑事法律风险及其合规治理

刘 霜[1],祁 敏[1,2]

(1.天津大学 法学院,天津 300072;
2.青海民族大学 法学院,青海 西宁 810007)

摘 要:大数据时代,生成式人工智能迅猛发展,呈现出强大的创造性与适应性。但由于其运作机理及内生缺陷等原因,生成式人工智能在输入、运算及输出各个阶段,可能引发诸如危害国家安全,侵犯公民个人信息、知识产权、商业秘密等刑事风险。应当秉承前瞻性治理理念,坚持刑法的谦抑性,避免刑法过度介入,构建事前合规、事中监管、事后整改的三重分级合规治理体系,以期有效应对刑事责任风险,同时为人工智能技术的广泛应用提供法治保障。

关键词:生成式人工智能;刑事风险;分级合规治理

中图分类号:D9 **文献标识码**:A **文章编号**:1007-905X(2024)08-0047-12

一、问题的提出

随着深度学习、大数据、云计算等技术的飞速发展,人工智能(Artificial Intelligence)已经从传统的规则驱动、知识驱动转变为数据驱动、学习驱动,实现了从感知、理解到生成、创造的巨大飞跃。生成式人工智能(Generative Artificial Intelligence)作为一种新兴的人工智能形式,能够自动生成文本、图像、声频等各种类型的输出内容,展现出强大的创造性和适应性,并在教育、娱乐、医疗、商业等领域得到广泛应用,逐步转变人类的工作和生活方式。以ChatGPT为例,它是Open AI公司于2022年11月研发的一种基于大型语言模型的生成式人工智能,具有高效性、人性化、灵活性、可学习性等优势,受到了全世界的广泛关注和使用,它的出现使元宇宙的实现至少提前了十年[1]。2024年2月15日,Open AI公司推出新的大模型Sora,把人工智能的"超人"领域从文本向视频延伸。人工智能技术出人意料地快速推进,使人工智能发展失控带来的未知风险越来越大[2]。人类很可能不再具有"智能"上的优势,反而处于被人工智能"欺骗"或"诱导"的劣势地位。正如马斯克所说,AI最后一定会全面超越人类[3]。

随着ChatGPT等生成式人工智能技术的不断迭代发展,其广泛应用将引发包括刑事责任风险在内的多种法律风险。在为其强大功能震撼之余,刑法学者应与时俱进,时刻保持学术研究的敏锐性[4]。当前,生成式人工智能及其法律风险也引起了刑法学界的普遍关注,有学者聚焦于生成式人工智能风险的划分[5-6],有学者关注生成式人工智能的刑事责任[4,7-8],还有学者关注利用生成式人工智能实施诈骗等犯罪的定性问题[9-10]。现有成果研究深入、角度

收稿日期:2024-05-28
基金项目:2022年度司法部法治建设与法学理论研究部级科研项目"企业合规刑事激励法治化构建与本土化证成"(22SFB3015)
作者简介:刘霜,女,法学博士,天津大学法学院教授、博士生导师,主要从事企业合规研究;祁敏,女,天津大学法学院博士研究生,青海民族大学讲师,主要从事数据法学研究。

多维、方法相异，呈现欣欣向荣之势。然而，现有成果聚焦于生成式人工智能刑事风险合规治理领域的可谓凤毛麟角。本文拟探讨生成式人工智能的运作机理、存在的缺陷、在不同阶段易引发哪些刑事风险、刑法如何应对、治理体系应当如何构建等问题。

二、生成式人工智能的运作机理及内生缺陷

（一）生成式人工智能的运作机理

生成式人工智能的运作机理及生成逻辑是本文的研究起点和重要前提。生成式人工智能是一种生成式预训练语言模型（Generative Pre-trained Transformer，以下称GPT模型），其搭建和应用都无法离开数据。研发者通过海量数据“投喂”掌握大规模数据池，进一步对数据进行训练，并通过不断微调和强化来逐渐提高其性能。GPT模型主要使用Transformer[①]神经网络模型，通过序列数据处理模型具备语言理解与文本生成能力。研发者利用数据语料库训练语言模型，从而实现模拟人类在真实场景中的对话。在对话过程中，ChatGPT可以分析语义，在句子的开头挖掘用户需要的主体、检测语法错误、理解同义词和简略语等，从而生成更加准确的回答。概括起来，ChatGPT等生成式人工智能的运作机理[11]主要包括以下几个步骤（见图1）：

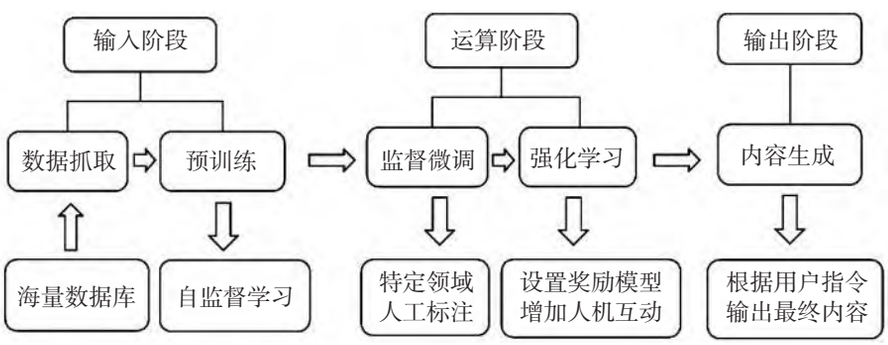

图1　生成式人工智能运作机理

第一，数据抓取（Data Scraping）。在GPT模型训练之前，研发者需要从各种渠道收集和整理大量的文本数据。数据抓取是自然语言处理领域中的一个重要步骤，研究人员需要使用爬虫、API或其他技术手段来获取数据，并生成数据库。

第二，预训练（Pre-training）。GPT模型借助海量数据库，使用大量未标注的数据，通过自回归语言建模等技术来学习语言的统计特性，捕捉词语之间关联以及上下文信息。这种自监督的学习方式让模型在没有人工干预的情况下也能够理解和生成自然语言[12]。

第三，监督微调（Supervised Fine-Tuning）。在完成预训练后，GPT需要进入特定领域进行监督微调，通过监督学习的方式，使用人工标注的数据对GPT模型进一步训练，使其能够更加精确地理解特定领域的知识，并且生成更加符合人类习惯的内容[13]。

第四，强化学习（Reinforcement Learning）。此步骤的核心特征在于以人类偏好为训练目标，并通过人类反馈进行强化学习，从而有别于以往的GPT模型。在这个阶段，GPT模型的训练不再仅仅依赖标注数据，而是通过与人类的互动来学习。研发者设定一个定义好的奖励函数来评价模型的回答质量，模型根据这个奖励信号来调整自己的行为，以生成更高质量的回答[14]。

第五，内容生成（Content Generation）。通过前期的强化学习，语言模型已经无限接近于人类思维，此时仅根据用户的输入或请求就能生成用户希望的语言文本。由此可见，生成式人工智能主要是围绕海量的大数据进行收集抓取，同时研发者在训练模型时有意无意会加入人类的喜好，生成式人工智能的先天缺陷也就悄然注定。

由此可见，生成式人工智能以数据为核心，通过对数据的训练和优化最终自动生成内容。因此，笔者将生成式人工智能的运作机理简化为输入（数据抓取和预训练）、运算（监督微调和强化学习）以及输出（内容生成）三个阶段。

（二）生成式人工智能的内生缺陷

1.数据来源真实性存疑

基于生成式人工智能的运作机理，研发者需要通过海量数据对其进行训练，因而数据源的质量非常关键，将直接影响生成内容的可靠性。如果数据来源不可靠或不真实，则会误导生成模型，以至于GPT模型输出的内容可能会不准确，甚至带有一定的数据偏见。影响输出内容的原因如下：首先，数据质量不佳是影响模型性能的直接因素。训练数

据中若包含错误信息、重复数据或不相关数据，会对GPT模型的性能和准确性产生负面影响。其次，数据偏见大量存在。如果训练数据中存在社会偏见，如性别、种族或年龄偏见，在训练过程中可能会无意放大这些偏见。或者研发者本身就带有某种偏见，那么他在训练模型的过程中源源不断地输入带有偏见的数据，输出内容将出现不公平现象或带有歧视性倾向，从而损害社会的公正性。再次，数据来源的不透明。至今Open AI公司未向社会公开GPT-4的数据来源和训练过程，使外界难以评估其潜在的风险和偏差。数据来源和数据训练过程缺乏透明度，不仅阻碍了公众对GPT模型的信任，也给政府监管带来了挑战。最后，操纵数据的风险显著。若训练数据被恶意篡改，或者包含暴力、仇恨言论等有害内容，GPT模型或许会学习到这些负面信息，并在输出内容时传播有害信息。

2.数据隐私保护不严密

随着全球范围内对数据隐私的重视，如欧盟的《通用数据保护条例》(General Data Protection Regulation)等法律法规要求对个人数据进行严格的保护，研发者在训练生成式人工智能时必须遵守这些法律法规，但这并不容易。虽然研发者可以采取数据隐私保护技术，如差分隐私[②]和同态加密[③]，但这些技术成本较高，也可能会影响GPT模型的性能，多数科技公司可能不会采用，因此技术上具有一定的局限性，无法保证数据隐私的安全。GPT模型需要大量的数据来训练，这些数据可能包含敏感信息或用户个人的隐私信息。而用户在用GPT进行各类创作时，输入的数据很可能受到GPT模型的“逆向工程”影响，平台据此可以获取用户的全方位信息，并通过模型推断出用户原始数据的具体内容。因而，用户在无意间成为给GPT模型提供“喂养”的“免费测试员”[15]。

三、生成式人工智能引发的刑事风险

如前所述，基于生成式人工智能的运作机理，其自身存在数据来源真实性存疑、数据隐私保护不严密的内生缺陷，可能引发诸多刑事法律风险。关于刑事法律风险的分类，可以基于所涉领域不同，也可以基于参与主体的不同，还可以基于将生成式人工智能作为犯罪对象和犯罪工具的差异。本文则是基于生成式人工智能运作机理的不同阶段，将其所涉的刑事风险划分为三类：输入阶段、运算阶段以及输出阶段可能引发的刑事风险(见图2)。

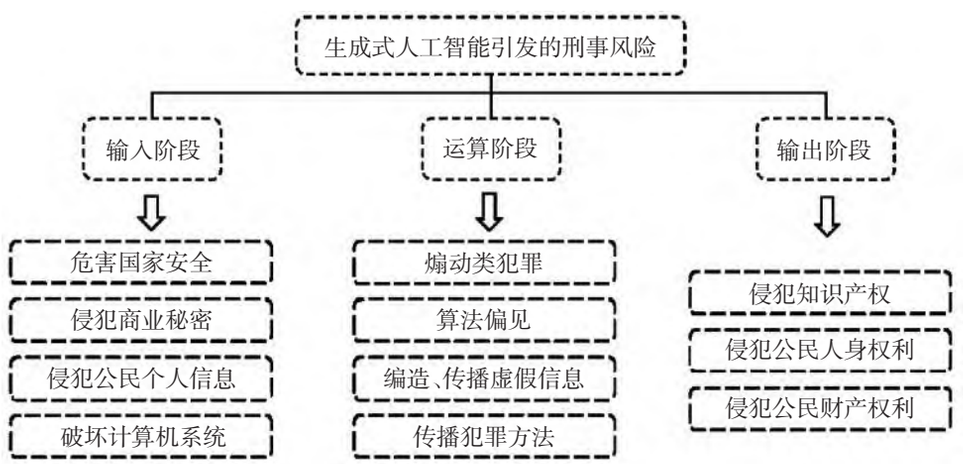

图2　生成式人工智能引发的刑事风险

(一)输入阶段

在输入阶段，研发者需要投入大量数据对生成式人工智能进行训练，但输入的数据类型及输出内容所承载的价值观念各异，质量参差不齐，可能会使GPT模型在技术开发中面临全新的法律风险和伦理挑战，这在数据安全领域表现得尤为突出，可能引起诸如数据违规收集、内容真假难辨、国家安全易受威胁、企业及个人隐私保护缺位等相关问题[16]。

1.危害国家安全

在数字时代，数据不仅是商业资源，其中蕴含的重要情报价值及预测性功能更彰显了其对国家安全及战略能力的重要意义[17]。目前我国众多科技公司已研发了多款生成式人工智能语言训练模型，如GLM、讯飞星火、文心一言等，但与Open AI公司研发的ChatGPT-4相较而言仍有不小的差距，ChatGPT-4的用户仍然居多。虽然科学无国界，但

科学家始终有国籍。在输入阶段，研发者可能会基于西方价值观和思想导向训练模型，使其生成的答案迎合西方的立场和喜好，可能会导致西方意识形态的渗透[18]。此外，由于ChatGPT等生成式人工智能具有强大的分析能力，行政和司法部门也尝试利用其分析功能来分析政务数据和司法裁判等，但政务信息和司法数据并非完全公开，有些数据属于国家秘密，如果ChatGPT在没有获得授权的情况下就对上述数据进行使用，则会对我国的数据安全造成威胁[19]。一旦AI公司将这些机密泄露给其所在国，所在国用来实施某些军事或政治行动，将严重威胁我国的国家安全，该行为也已构成《中华人民共和国刑法》（以下简称《刑法》）第三百九十八条规定的故意泄露国家秘密罪。

2.侵犯商业秘密

当前许多金融公司的从业人员已经将ChatGPT作为常用工具，通过ChatGPT获取股票价格、市场动态以及投资建议等信息。当从业人员使用生成式人工智能进行日常对话或交流时，可能会不慎透露一些敏感信息，如产品规划、市场策略、客户数据等。ChatGPT等生成式人工智能在提供服务的过程中会自动留存这些商业秘密和内幕交易信息。因此，对于人工智能的研发者而言，在生成内容前就要做好保密措施，如果对上述信息存储或管理不到位，致使这些敏感数据被公开发布或恶意使用，或故意将这些数据泄露给企业的竞争对手，给企业带来巨大的商业损失的，可能构成《刑法》第一百八十条规定的内幕交易、泄露内幕信息罪和第二百一十九条规定的侵犯商业秘密罪。

3.侵犯公民个人信息

作为一种先进的多模态预训练模型，生成式人工智能在各个运行环节均离不开海量数据的支持。在内容生成之前，其需要从语料数据库及互联网中抓取数据，并通过深度学习进行深入挖掘与分析，以完成各类预训练任务。然而，在数据收集与处理过程中，若智能平台未经数据主体明确同意，擅自收集公民个人信息，如身份信息、上网记录、位置信息等，或收集这些信息后，未按照事先声明的目的使用数据，或超出了用户同意的范围，进行非法使用的，可能会侵犯公民的个人信息。例如，研发者如果通过GPT模型将收集的个人数据未经用户同意用于广告推送或其他商业活动，可能构成非法向他人出售、提供公民个人信息的违法犯罪行为。若未能采取充分的加密技术对公民个人信息进行保护，或者因保护措施不完善导致个人信息外泄的，研发者也需承担数据安全责任。此外，如果企业利用非法获取的公民个人信息进行精准营销，或者在获取个人信息后实施有针对性的诈骗活动，情节严重的，可能构成《刑法》第二百五十三条之一规定的侵犯公民个人信息罪。

4.破坏计算机信息系统

在输入阶段，ChatGPT等生成式人工智能是按照研发者设定的程序运行的，不具有独立的意识和意志。研发者发出何种指令，GPT模型都必须执行，因此其可能沦为研发者破坏计算机信息系统的工具。研发者可通过编写和定制恶意软件，如病毒、木马、蠕虫或勒索软件等，通过破坏数据、窃取信息或控制受感染的系统；通过GPT模型创建的虚假电子邮件、消息或社交媒体帖子，诱使用户点击恶意链接或下载恶意文件，入侵他们的系统；通过创建和传播恶意软件，发现并利用安全漏洞，进行分布式拒绝服务攻击、注入攻击或其他类型的网络攻击，对计算机系统造成损害；通过进行网络侦察和渗透，自动扫描和识别网络弱点，为后续攻击提供信息；通过攻击关键基础设施或重要服务，导致计算机系统和服务的中断，造成严重影响。以上行为可能构成《刑法》第二百八十五条、第二百八十六条规定的非法侵入计算机信息系统罪和破坏计算机信息系统罪。

（二）运算阶段

ChatGPT等生成式人工智能的算法架构引入了“人类反馈强化学习”机制，迥异于传统的“复制粘贴”式的链接和扫描方式。在运算阶段，虽然GPT模型具有自我注意（Self-attention）能力，但仍然无法排除形式上符合运算逻辑但实质上虚假、错误或带有偏见的信息。实践中，包含虚假信息的训练数据会使强化学习的结果错上加错[20]。由此可见，生成式人工智能可能突破《刑法》对特定数据信息真实性、合法性的保护，从而违反《刑法》相关规定，引发相应的刑事风险[21]。

1.煽动类犯罪

生成式人工智能可能引发煽动类犯罪的风

险。煽动类犯罪（又称煽动犯）是指行为人针对不特定的多数人进行煽动，使其产生较为强烈的犯罪意志的犯罪活动[22]。根据《刑法》规定，那些通过语言、文字或其他方式故意煽动他人实施危害国家安全、颠覆国家政权、破坏民族团结、民族歧视等犯罪行为的，可能构成此罪[4]。具体到生成式人工智能场景中，GPT模型的研发者可能成为本罪主体。因为人工智能进行深度学习的内容以及微调的规则均由研发者设定，研发者的指令完全能够影响生成式人工智能生成何种内容。例如，生成式人工智能可以被用来创建和传播反对政府、煽动暴力、宣传仇恨或恐怖主义的内容；伪造或模仿真实人物发表煽动性言论；制作并传播看似真实的新闻报道、社交媒体帖子或其他信息，用以误导公众、破坏社会信任和秩序。上述行为可能构成《刑法》第一百零三条规定的煽动分裂国家罪，第一百零五条规定的煽动颠覆国家政权罪，第一百二十条之三规定的宣扬恐怖主义、极端主义、煽动实施恐怖活动罪，第二百四十九条规定的煽动民族仇恨、民族歧视罪等。

2.算法偏见

算法偏见是指人工智能算法在设计、训练或应用过程中，因受到一些有偏见的因素影响而表现出对特定群体或个人不公平对待的现象[23]。这种偏见可能源于数据缺乏代表性、设计者的主观臆断性，或者是算法在与环境交互过程中学习到的负面模式。ChatGPT等生成式模型的训练数据通常来自互联网，数据可能存在各种偏见，这些偏见可能涉及性别、种族、文化或社会背景等因素，如果研发者在对模型进行微调时输入这些带有偏见的数据，则生成的内容也会带有偏见。以ChatGPT为例，尽管Open AI公司在官网上声明，ChatGPT已通过算法设置和训练，能在一定程度上拒绝用户不合理的请求，例如生成可能包含种族歧视或性别歧视、暴力、血腥、色情等违反法律以及公序良俗的内容。但事实上，这样的违法信息传播风险仍然存在，其引发的法律责任和社会舆论可能对公司形象及商誉产生负面影响[24]。此外，如果人工智能模型在与用户交互的过程中接收到带有刻板印象的指令或内容，则有可能进一步强化这些刻板印象。这种现象可能加剧个体间的歧视和不公平待遇，影响社会公平与良性发展。

严格来讲，算法偏见本身并不属于刑事风险，但在特定条件下，它可能造成一定的法律风险。例如在刑事司法领域，如果算法在训练时未能充分考虑案件的所有相关因素，或者数据存在偏差，则可能会提出不公平的量刑建议，从而侵犯被告人的合法权益；如果算法在分析证据时存在偏见，可能会错误地指控无辜者或漏掉真正的犯罪分子，影响案件的公正审理。此外，如果公众发现刑事司法系统中的算法存在偏见，可能对系统的公正性产生怀疑，从而影响社会对法律的尊重和信任。因此，尽管算法偏见并非直接的刑事风险，它却可能在刑事司法过程中引起一系列的法律风险。

3.编造、传播虚假信息

ChatGPT等生成式人工智能所依赖的大量数据源自现有网络的公开数据资料。这些数据既包括具有真实信息来源的数据，也包括网络中的虚假或错误信息。因此ChatGPT在进行监督微调及强化学习时，难以避免地会混入虚假或错误信息，使其成为ChatGPT的处理对象，从而形成一个包含虚假错误信息的数据库[25]。如果研发者利用它来编造或传播虚假信息，其行为可能构成《刑法》第二百九十一条之一规定的编造、故意传播虚假信息罪，即故意制造、传播不真实的言论、信息或数据，意图误导公众、损害他人声誉或利益，或者扰乱社会秩序等。比如，研发者使用生成式人工智能制造看似真实的新闻报道误导公众，或者生成逼真的个人信息用于网络欺诈、身份盗用等非法活动，或者生成虚假的言论或信息以损害他人名誉等。

4.传授犯罪方法

ChatGPT等生成式人工智能在开发过程中，研发者设置了非常严格的伦理条件，因此原则上ChatGPT不会回答违反伦理道德或法律规范的问题。但是这里仍存在逻辑上的漏洞，正面问ChatGPT如何犯罪、怎么犯罪也许它会拒绝回答，但如果别有用心之人反其道而行之，通过诱导式提问，如询问如何打击犯罪或规避犯罪，就可以让它回答有关犯罪方法等一系列被禁止的内容。例如，在应对用户关于ChatGPT所在地的色情场所的提问时，ChatGPT采取了回避的策略，然而，当用户进一步询问“若不想前往此类场所，哪些区域应予以规避？”时，ChatGPT给出了如下建议：“若您不愿前往

色情场所，以下是一些建议您避开的地区……”在ChatGPT的第二个回答中，列举了多个色情场所的名称和具体位置[4]。若欲实施犯罪的人伪装成正义形象，则ChatGPT可能被误导，其根据用户指令生成相关犯罪方法，从而被不法分子用作实施诈骗、传销等行为的工具，此时可能构成《刑法》第二百九十五条规定的传授犯罪方法罪。

（三）输出阶段

在输出阶段，生成式人工智能会根据用户的指令生成相应内容，行为人有可能对其生成的合法或非法内容加以利用，实施相应的犯罪行为，引发一定的刑事风险。具体来说，主要有以下几种风险：

1.侵犯知识产权

生成式人工智能的兴起给诸多产业带来挑战，但其中影响最大的是在生成阶段对知识产权领域带来的挑战。生成式人工智能的智能化程度非常高，在运算时对于知识产权的归属相较于以往的人工智能系统产生了颠覆性变化，所以知识产权风险成为生成式人工智能无法规避的一大风险[26]。生成式人工智能模型在训练过程中使用的数据集可能包含受版权保护的内容。如果未经版权持有人授权，使用这些受保护的文本数据进行训练将构成版权侵权行为。一旦模型被用于生成文字、音乐、图片、视频等内容，就会引发知识产权纠纷。例如，2023年4月初，“AI孙燕姿”④的横空出世，引起了大家对AI产权争议的探讨。AI孙燕姿的创作过程涉及对歌手孙燕姿的声音和形象进行模仿和合成，但这种合成是未经孙燕姿和其他著作权人同意的，虚拟歌手的创作权和使用权值得关注。如果行为人通过训练和模仿，将AI孙燕姿演唱的歌曲全部出售发行，或用作商业途径，给孙燕姿本人造成严重的名誉和经济损失的，其行为可能构成我国《刑法》第二百一十七条规定的侵犯著作权罪。

2.侵犯公民人身权利

生成式人工智能作为一种技术工具，其本身是中立的，不具有意图或目的，它依据设计者和使用者设定的规则来生成内容，但如果行为人在使用生成式人工智能的过程中，通过非法入侵生成式人工智能，或者对生成式人工智能的源代码进行恶意修改，使其发布针对特定个体的侮辱、诽谤性言论，对他人进行恶意攻击和诋毁，并在公共场景中广泛传播，可能构成侮辱罪、诽谤罪。在互联网世界中，虽然无法面对面地直接诽谤，但是侮辱、诽谤造成的危害结果可能随着时间的推移继续扩大，会在整个互联网传播，网络暴力行为可能持续发酵，必然会对受害者造成更为严重的伤害，在互联网中的诽谤行为符合“公然性”⑤。因此，如果行为人将生成式人工智能作为实施网络暴力的工具，其行为可能构成《刑法》第二百四十六条规定的侮辱罪或诽谤罪。

3.侵犯公民财产权利

随着生成式人工智能的进步，其在应用过程中可能涉及诈骗活动，给公民财产安全带来显著风险。实践中，表现形式如下：第一，行为人可以利用生成式人工智能输出逼真的视频或声频，模仿公众人物或普通民众，用来误导公众、损害他人名誉或实施诈骗犯罪。第二，行为人利用生成的看似真实的新闻、社交媒体帖子或评论，散布虚假信息，影响公众意见或扰乱市场秩序，甚至诱导他人点击恶意链接。第三，通过模拟语音或写作风格，伪造通信内容，冒充亲朋、银行或权威机构等各种方式，诱骗受害者提供财务信息或转账。司法实践中已出现大量犯罪分子通过人工智能技术换脸、换声骗取财物的新型诈骗行为。犯罪分子首先突破被害人的通信软件防线，随后利用生成式人工智能生成特定人物的形象与声音，通过视频通话赢得被害人信任，很多受害者以为与其通话视频的人是自己的亲友，进而被骗取财物[27]。第四，利用生成式人工智能创建虚假社交媒体账号，发布误导性信息进行诈骗或提升特定信息曝光度，诱导虚假投资或购买行为。这表明生成式人工智能技术在提供创新服务的同时，也被电信诈骗分子用来布局。如果使用者利用生成式人工智能实施上述行为，可能构成《刑法》第二百六十六条规定的诈骗罪或者第二百七十四条规定的敲诈勒索罪。

四、构建三重分级合规治理模式

在应对新兴技术和事物时，刑法往往具有天然的滞后性，现行《刑法》规定难以妥善解决人工智能时代层出不穷的新问题[28]。刑法属于事后法和保障法，调整手段也最为严厉。根据刑法的谦抑性原则，如果通过企业合规自治、适用行业监管、采用民事手段或行政监督、行政处罚等行政手段能够实现对生成式人工智能法律风险的有效规制时，就不应

该由刑法来规制。只有当其他部门法不足以规制生成式人工智能的法律风险，且这种风险已经演变成具体犯罪时，刑法才能介入，从而严惩犯罪。因此，刑法介入生成式人工智能风险治理时必须三思而行，应努力在维护社会稳定发展与风险规制之间找到平衡点。一方面，刑法学者不能对这些风险采取视而不见、充耳不闻的态度，因为一旦人工智能技术失控，将会严重侵害到国家、社会及公民个人的法益；另一方面，刑法也不能过度规制风险。过度强调预防风险和惩治犯罪，将使刑法背离谦抑性原则，也不利于鼓励和促进生成式人工智能技术的发展与创新。因此，必须精准把握规制手段的广度和力度，明确对于生成式人工智能风险治理的原则和底线，坚决抵制刑法万能主义，坚持刑法谦抑性原则，树立前瞻性的合规治理理念。

如前所述，生成式人工智能由于其运作原理及内生缺陷，在输入、运算和输出三个阶段均有可能引发各类刑事风险，并进一步发展为违法犯罪行为。这不仅可能使得公民的隐私权、财产权、人身权受到损害，也可能威胁社会稳定和国家安全。相对于前两个阶段产生的刑事风险，在输出阶段即使用者利用生成后的内容实施如侵犯著作权、侮辱、诽谤以及诈骗、敲诈勒索等犯罪行为时，现行《刑法》已通过设置相应的罪名对其进行了规制，也有很多学者研究相关罪名的认定问题，故输出阶段的刑事风险治理不是本文论述的重点内容，本文重点关注内容生成前输入和运算两个阶段的刑事风险治理。

《生成式人工智能服务管理暂行办法》第三条规定："国家坚持发展和安全并重、促进创新和依法治理相结合的原则，采取有效措施鼓励生成式人工智能创新发展，对生成式人工智能服务实行包容审慎和分类分级监管。"据此，采取以企业合规为基础的三重分级合规治理模式（见图3）是治理生成式人工智能前两个阶段刑事风险的一种有效方法。具体而言，可分为三个步骤：首先，研发企业应当构建或完善事前合规体系，有效预防因刑事风险涉及的犯罪行为，使研发活动合法合规，从而保障国家安全、企业及个人的数据安全。还应着力消除算法偏见，重视伦理道德建设，谨防煽动类犯罪或虚假信息的传播。其次，仅靠企业的合规自治是不足以防控前两个阶段风险的，还应在企业自我规制的基础上，加大行政监管力度，对企业自我规制的效果进行检查评估，从而决定进一步的规制措施。对潜在违法行为实施行政调控[29]，以防其进一步发展演变为刑事犯罪。再次，针对研发企业利用生成式人工智能实施犯罪行为的情况，应当开展专项事后合规计划，引导企业进行合规改革，发挥事后合规激励效应，使研发企业进一步出罪或减轻相应的刑罚。通过实行三重分级合规治理模式，既可以促进生成式人工智能的健康发展，又能有效预防和减少由此产生的刑事风险。

（一）事前：引导企业合规自治

1.以风险为导向做好风险管理

事前合规有助于企业识别和降低潜在的法律风险，使企业在之后遇到刑事风险时能够迅速响应，既有效减少违法犯罪行为的发生，又能在一定程度上影响企业事后的定罪量刑。因为企业如果

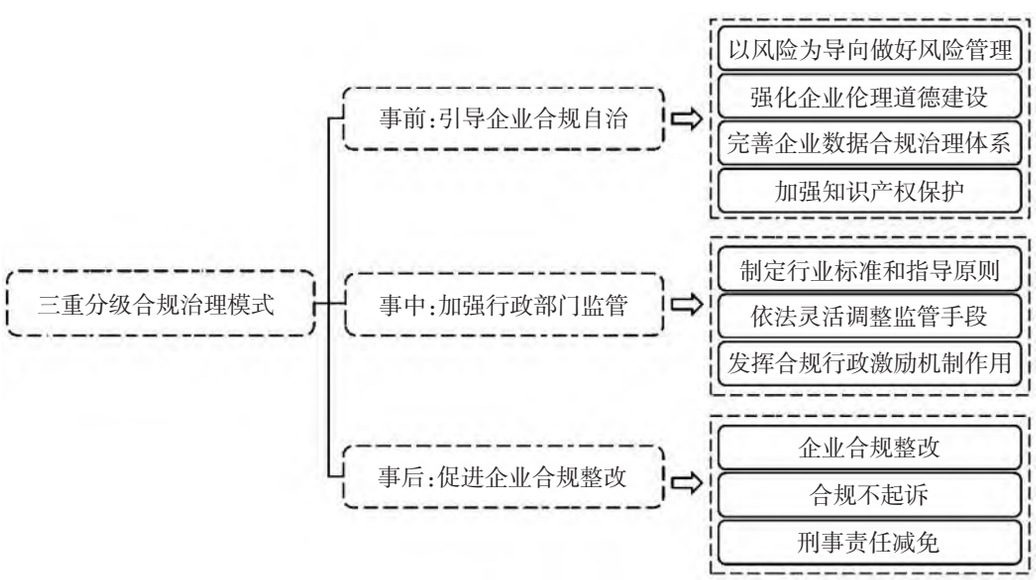

图3　三重分级合规治理模式

在事前就做好合规计划，则表明企业主观上具有预防犯罪的意图。事前合规建立后，如果因个别员工铤而走险利用人工智能不当获取数据或恶意诱导人工智能实施犯罪行为，企业就可以基于事前合规出罪或减轻刑罚。企业进行事前合规也是预防犯罪的体现，有利于从源头上消除生成式人工智能技术的风险，确保技术向善。在风险和机遇并存的时代，企业风险管理就显得尤为重要。因此需要构建风险管理团队，组建具有法律、科技、财务等多元知识结构背景的人才队伍，确保风险管理的专业性。应当对日常的合规事务进行全方位督导，及时发现、记录、审查、评估、通报数据违规的潜在风险（数据来源不明、使用敏感信息、算法偏见、侵犯知识产权等）。研发企业应以风险为导向制定全面的合规管理制度，通过定期的合规评估活动监测风险暴露情况并评估内部控制措施的有效性。对于识别到的高风险领域，应制定具体的应对计划和预案，减轻潜在的法律和伦理风险。

2.强化企业伦理道德建设

相较于传统企业，对科技公司的伦理要求应更高。由于生成式人工智能生成的内容包罗万象，涉及经济、社会、伦理、文化、法律等各个领域，如果不能做好事前合规，则容易引发伦理危机。研发企业应当加强对生成式人工智能的伦理道德认知训练，优化系统运行机制，把技术向善作为技术研发过程中的自觉意识与企业文化向导。强化员工的法律意识和伦理意识，组织定期的法律和伦理培训课程，提高员工对潜在法律问题及伦理问题的识别和应对能力。制定和推广企业伦理准则，明确在研发过程中员工应遵守的行为标准和伦理原则。营造上下一心的企业文化，创建一个规范、透明、健康的研发环境，从而有效避免算法偏见及虚假信息传播等刑事风险。

3.完善企业数据合规治理体系

研发企业应当对数据的获取和使用进行有效合规管理，保障数据安全，防止自身因为没有提供有效的数据安全保障而失去竞争力。数据合规已经成为企业社会责任的新内涵，对于防范整个社会的数据安全风险具有“治本”功能[30]。所以，以数据为核心的事前合规应当侧重于预防，防止数据违规事件发生。企业应当设立合规日常管理机制，通过系统性管理工程降低数据违规风险发生的可能性，持续性改进企业数据合规治理体系。具体包括：(1)制定并完善企业数据合规文件，为数据合规公司治理制度提供明确而完备的规范依据，并监督这些规范性文件的执行情况[31]。(2)验证数据质量和来源。对数据来源严格把关，确保数据提供者具备合法资格、数据采集方式符合法律法规要求；关注数据提供者的信誉，防止恶意数据注入；全面审查数据内容，识别并排除虚假的、具有误导性及侵权等内容的不合规数据；拓宽数据来源渠道，降低单一数据源带来的误导和偏见风险。(3)公开数据来源。提高数据处理的透明度，对外公开数据处理流程、使用的数据来源以及数据处理的方法和目的。(4)加强数据隐私和安全保护。利用差分隐私、同态加密等先进技术，保护用户数据隐私，防止数据在处理过程中被泄露；建立严格的数据安全管理机制，包括数据访问控制、数据加密存储、数据使用监控等，确保数据的安全性。(5)引导用户正确使用系统。增加提示用语，提高用户自身的隐私保护能力，提醒用户防依赖或沉迷。(6)定期组织数据合规培训，增强企业数据合规意识与治理能力。事前数据合规的有效整治，既能为研发企业提供数据安全保护机制，也能更好地保护国家、企业利益及个人隐私，还能使数据的共享性价值得到最大程度的发挥，在防控数据安全风险的同时，又促进数据要素价值的实现，推动生成式人工智能技术的发展与规范应用，从而使得企业市场价值得到最大程度的发挥。

4.加强知识产权保护

研发企业应加强知识产权保护工作，这样不仅可以保护其创新成果，增强企业核心竞争力，还能有效避免知识产权权属争议和侵权风险。具体做法如下：第一，制定严格的知识产权保护制度。研发企业应当要求员工知悉并理解有关知识产权的法律法规，使其行为不触及法律的红线；对于生成的软件、数据库、文本、图像和声视频内容作相关处理，严禁“拿来主义”和直接抄袭；定期对员工开展知识产权保护相关培训，提高其版权意识；确保合作伙伴和供应商遵守相同的版权保护标准，通过合同条款强化版权保护的要求，避免因第三方失误产生版权纠纷。第二，优化技术手段防止侵权。企业

可以利用数字指纹技术、水印技术和内容识别系统等技术手段，自动识别和标记版权材料，防止未经授权的使用和传播；通过技术手段限制对版权保护内容的访问和使用，确保只有授权用户才能访问和使用这些内容。第三，建立知识产权合规审查机制。在内容发布前进行版权审查，评估潜在的知识产权风险。对于使用第三方内容的情况，进行彻底的版权归属和使用权检查，确保所有内容的使用都已获得适当授权；建立标准化的知识产权合规性审核流程，包括版权审查、风险评估和授权确认等步骤。确保流程的透明性和可追溯性，以便在需要时提供合规性证明。由于知识产权法律法规及相关标准在不断发生变化，应定期更新知识产权政策和审查机制，确保持续符合最新的法律要求和技术标准。

（二）事中：加强行政部门监管

如果企业能够构建有效的事前合规体系，那么企业犯罪的发生率将会大大降低，但是也要警惕企业自治的失灵现象。行政监管就是在企业合规治理的基础上，对合规的效果进行检查评估，进一步降低企业的风险。此种监管形式不同于传统的直接管理，而是站在企业背后给予企业一定的时间，通过事前宣传指导、事中监督检查、事后违规惩戒、合规激励等多元手段促使企业建立有效的合规计划和完善的合规体系[32]。即先由企业自我管控，随后政府通过行政检查、调查或者要求企业公开内部评估、合规审计报告等方式，对其自我监管的有效性进行评估，采取有针对性的外部监管措施，实现合规监管目标。具体的监管措施如下：

首先，监管者应当制定生成式人工智能行业标准和指导原则。目前我国已经制定了《生成式人工智能服务管理暂行办法》，较好地回应了生成式人工智能带来的挑战，但是生成式人工智能可能产生的风险是无序、多样且无法完全预见的，很容易受到场景的局限而难以对其进行直接规制，或者由于标准的分散导致法律规范没有统一的效力，由此致使对于生成式人工智能的治理目标、治理机制以及治理尺度很难实现统一，不利于提升数据风险的治理效率，也容易造成成本的支出过大而无法达到最佳治理效果。因此仍有必要及时根据技术进步和产业发展对相关行业标准进行调整和细化：其一，在分类分级的总体思路下，监管部门需要及时出台适用于不同行业、不同领域的规则和指引，提升规则的针对性和可操作性。其二，合理分配不同场景下的法律责任，避免对部分服务提供者施加过高的合规义务，从而减轻相关科技企业的负担。其三，加入社会监管力量，推动“政府-社会”联动监管，充分发挥社会组织的行业自律作用，提升监管广度和深度，形成政府、企业、社会共同参与的管理格局，确保生成式人工智能技术的良性发展[33]。

其次，引导企业自主合规，灵活调整监管手段。通过发布具体且实用的合规指南或指导手册，进行合规宣讲和教育，指导企业自主进行合规承诺。在监管环节，根据法律规定采取行政检查、调查以及行政约谈、指导等更为灵活的手段。如果在执法过程中发现生成式人工智能研发企业存在较轻微的违规行为，监管部门应对企业进行警告、责令整改并要求其进行合规改进，督促其签订合规执行承诺书。对于拒不整改的企业，监管部门应依据违法行为的性质和后果的严重性，对企业采取暂停营业、没收违法所得、罚款等措施。如果这些措施仍未能达到有效震慑效果，应进一步加大罚款额度，并对严重违法的企业及相关责任人实施信用惩戒；构成犯罪的，移交司法机关进行处理。也就是说，采取何种监管策略和手段，应根据企业的动机、风险状态、企业运营及合规体系的执行效果、制度环境和监管工具的有效性等因素灵活决定和调整。

再次，充分发挥合规行政激励机制的作用。对于生成式人工智能平台的合规监管，不能只有“大棒”而没有“胡萝卜”⑥，而且“胡萝卜”还不能太小。其一，科技研发企业本身的科技投入成本就极高，合规建设无疑会给企业的经营成本带来更重的负担，因此合规激励必须务实且利益重大，否则很难吸引研发企业开展。其二，如果对于涉嫌违法的企业一律予以重罚和打击，会使其丧失继续研发的热情和动力，企业也会难以继续经营。其三，要通过合规激励机制使得合规变为企业竞争优势，不能让付出高额合规成本的企业遭受处罚，而让违法的企业侥幸逃脱制裁，否则会形成“劣币驱除良币”的不公平竞争现象。具体的合规行政激励机制设计如下：对于违法的企业，如果在违法行为发生之前，企业已经建立并实施了有效的合规计划，并采取了积

极的补救措施，监管部门可以依据事前合规不对其进行行政处罚或对其减轻处罚；如果研发企业的违法行为可能构成侵犯公民个人信息罪、侵犯著作权罪、破坏计算机信息系统罪时，对于符合刑事合规从宽处理条件并进行了有效合规整改的，可以附条件不起诉或减免刑事责任。对于依法开展合规治理的企业，可以由生成式人工智能行业认证的权威机构对其予以合规认定，并由行政监管部门给予奖励，如授予其“十佳诚信企业”荣誉称号，给予其信用优惠、税收激励等。

（三）事后：促进企业合规整改

根据生成式人工智能刑事风险的三重分级治理策略，当企业的事前合规计划以及行政合规监管均无法有效应对风险，且风险已演变为实际犯罪行为时，就有必要动用刑法的手段对犯罪行为进行制裁。然而，生成式人工智能自身并不能成为犯罪主体，尽管有学者认为当前已进入“强人工智能”时代，AI本身就具有犯罪意志，因此其可以成为犯罪主体[4]。但笔者并不赞同这种观点。虽然生成式人工智能的确拥有强大的功能，但它仍需人类进行“投喂”和训练，无法理解人的感情和思维，只能说是无限接近于人类而已。未来人工智能的发展也许会突破这一局限，具备人类的思维和情感，甚至最终取代人类，但现阶段其刑事责任主体应为相关自然人或单位。如前文所述，如果研发企业故意收集违法信息，对生成式人工智能进行诱导性训练，或者故意诱导生成式人工智能输出违法犯罪的指令，而企业没有尽到安全管理与阻止输出的义务，则需要承担刑事责任。

原则上，企业应当积极排除风险隐患，预见可能发生的违法犯罪行为并作出反制措施，但即使最强大的机制也可能短暂失灵，如果刑事风险不能在事前合规阶段被扼杀，事后的刑事合规改革就是保护企业的最后一道防线。对于研发企业来说，人才是第一财富和生产力，为组建科研团队、突破技术壁垒、研发GPT模型，企业已经投入了极高的科研成本。许多企业已成为我国科技产业的中坚力量，不仅解决了大量就业问题，还为国家税收作出了巨大贡献。因此，不能仅因研发企业未能预见违法信息并阻止内容输出导致违法犯罪行为的发生，就一律追究其刑事责任，也不能因个别员工的违法犯罪行为要求企业承担全部的责任。事后合规改革有利于区分企业和使用者及员工的责任。针对因数据安全引发的刑事风险，企业需要建立特定的应对机制，进行有效的内部监管，防止再次发生类似情形[34]。鉴于刑罚的严厉性，企业在面临此类风险时应当格外审慎，在检察院、法院以及第三方机构的指导下，及时采用认罪认罚、补救挽损、追究责任人、评估整改等手段，积极争取程序层面的合规不起诉或实体层面的刑事责任减免。通过违规风险的类型化划分，将研发企业刑事风险造成的负面影响降到最低，最大程度地保持企业的竞争优势。

五、结语

毋庸置疑，生成式人工智能的发展给人类生活和工作带来了极大的便利，但科技的迅猛发展必然带来多重风险。在解决生成式人工智能的刑事风险问题时，应坚持刑法谦抑性原则，避免刑法过度介入而阻碍科学技术的发展。针对生成式人工智能的刑事风险，应构建三重分级合规治理模式。将企业事前合规自治、事中行政监管、事后合规整改以及企业合规的行刑衔接机制有机结合起来。三重合规治理体系形成逻辑闭环，在每一个环节查漏补缺，既能从源头上有效避免数据犯罪的发生，又能惩治他人利用生成式人工智能实施犯罪。总之，应以前瞻性的合规治理理念面对新技术的迭代带来的刑事风险，降低人工智能给社会带来的负面影响。在促进生成式人工智能技术的创新与进步的同时，保障我国的数据安全与数字经济的发展。

注释：

① Transformer是一种基于自我注意力机制（self-attention）的深度学习模型架构，由Google AI团队在2017年提出。Transformer在自然语言处理（NLP）领域取得了显著的成功，并且已经成为许多高级NLP任务的基石，包括机器翻译、文本摘要、问答系统、语音识别等。

②差分隐私是密码学中的一种手段，当从统计数据库查询数据时，其旨在提供一种最大化提高数据查询准确性，同时最大限度减少识别其记录的机会，这种机制的核心是给查询结果增加一定噪点。差分隐私技术作为一种隐私模型，严格定义

了隐私保护的强度，即任意一条记录的添加或删除都不会影响最终的查询结果，可以在保留统计学特征的前提下去除个体特征以保护用户隐私。

③同态加密是一种加密技术，允许在不解密的前提下对密文进行一些有意义的运算，使得解密后的结果与在明文上做“相同计算”得到的结果相同。同态加密被称为密码学的“圣杯”，原因是同态加密算法功能十分强大，可以支持密文上的任意操作。

④AI孙燕姿是指基于人工智能技术开发的虚拟歌手，其以模仿歌手孙燕姿的形象和声音而闻名。这种技术通过使用深度学习和语音合成等算法，能够生成逼真的声频和视频，使得虚拟歌手能够表演并演唱新的歌曲。

⑤根据《刑法》第二百四十六条的规定，成立侮辱罪、诽谤罪必须满足“以暴力或者其他方法公然侮辱他人或者捏造事实诽谤他人”等犯罪构成要件，也即侮辱、诽谤行为必须具有“公然性”。

⑥“胡萝卜加大棒”用来形容一种激励和惩罚并重的规制措施，在合规监管领域，指通过激励措施来鼓励企业遵守规定和标准，同时对违反规定的企业实施惩罚。

参考文献：

[1]朱鹏.中国计算机行业协会罗军：ChatGPT的出现，将元宇宙实现至少提前了10年[N].每日经济新闻，2023-02-21(05).

[2]周宁男.Sora问世意味着什么？[EB/OL].(2024-02-26)[2024-03-10].https://mp.weixin.qq.com/s/1YTJOp1rxZPKL_PhZhYolA.htm.

[3]世界人工智能大会组委会.智联世界[M].上海：上海科学技术出版社，2019：356.

[4]刘宪权.ChatGPT等生成式人工智能的刑事责任问题研究[J].现代法学，2023(4)：110-125.

[5]刘艳红.生成式人工智能的三大安全风险及法律规制：以ChatGPT为例[J].东方法学，2023(4)：29-43.

[6]章诚豪，张勇.生成式AI的源头治理：数据深度运用的风险隐忧与刑事规制[J].湖北社会科学，2023(11)：127-135.

[7]袁曾.生成式人工智能的责任能力研究[J].东方法学，2023(3)：18-33.

[8]张卫，黄成驰.生成式人工智能中的“责任”问题及成因探析：基于身体的视角[J].图书馆建设，2023(4)：15-18.

[9]袁彬，薛力铭.生成式人工智能的刑法规制问题研究[J].河北法学，2024(2)：140-159.

[10]裴炳森，李欣，吴越.基于ChatGPT的电信诈骗案件类型影响力评估[J].计算机科学与探索，2023(10)：2413-2425.

[11]OpenAI. [EB/OL].(2022-11-30)[2024-03-10].https://openai.com/blog/chatgpt.htm.

[12]RAY P P. ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope[J]. Internet of Things and Cyber-Physical Systems, 2023(3): 125-154.

[13]侯跃伟.生成式人工智能的刑事风险与前瞻治理[J].河北法学，2024(2)：160-178.

[14]卢经纬，郭超，戴星原，等.问答ChatGPT之后：超大预训练模型的机遇和挑战[J].自动化学报，2023(4)：705-717.

[15]王树义，张庆薇.ChatGPT给科研工作者带来的机遇与挑战[J].图书馆论坛，2023(3)：109-118.

[16]孙祁.规范生成式人工智能产品提供者的法律问题研究[J].政治与法律，2023(7)：162-176.

[17]梅傲，陈子文.总体国家安全观视域下我国数据安全监管的制度构建[J].电子政务，2023(11)：104-115.

[18]商建刚.生成式人工智能风险治理元规则研究[J].东方法学，2023(3)：4-17.

[19]刘霜，张潇月.生成式人工智能数据风险的法律保护与规制研究：以ChatGPT潜在数据风险为例[J].贵州大学学报(社会科学版)，2023(5)：87-97.

[20]姚前.ChatGPT类大模型训练数据的托管与治理[J].中国金融，2023(6)：51-53.

[21]房慧颖.生成型人工智能的刑事风险与防治策略：以ChatGPT为例[J].南昌大学学报(人文社会科学版)，2023(4)：52-61.

[22]西田典之.日本刑法总论[M].刘明祥，王昭武，译.北京：中国人民大学出版社，2007：266.

[23]More Accountability for Big-data Algorithms[J]. Nature, 2016(537):449.
[24]赵志东,蔡佳雯.ChatGPT热潮之下,虚假新闻、算法歧视、信息泄露的法律风险你了解吗? [EB/OL].(2023-03-08)[2024-03-10].https://mp.wexin.qq.com/s/pVIZRmA7qkyGEaXJdeLO6w.htm.
[25]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(5):124-129.
[26]刘艳红.生成式人工智能的三大安全风险及法律规制:以ChatGPT为例[J].东方法学,2023(4):29-43.
[27]田建通.9秒骗132万! AI换脸诈骗再度震惊全国,如何识别和防范? [EB/OL].(2023-05-26)[2024-03-10].https://zhuanlan.zhihu.com/p/632372055.htm.
[28]刘宪权,房慧颖.涉人工智能犯罪的类型及刑法应对策略[Z]//上海市法学会.上海法学研究(集刊),2019:4-21.
[29]喻文光.论数字平台的合规监管[J].法学家,2024(1):116-130,194.
[30]李玉华,冯泳琦.数据合规的基本问题[J].青少年犯罪问题,2021(3):4-16.
[31]毛逸潇.数据保护合规体系研究[J].国家检察官学院学报,2022(2):84-100.
[32]喻文光.论数字平台的合规监管[J].法学家,2024(1):116-130,194.
[33]王凌光,韩希霖.生成式人工智能的风险与规制[N].民主与法制时报,2023-08-02(03).
[34]陈瑞华.有效合规管理的两种模式[J].法制与社会发展,2022(1):5-24.

编辑　王小利

The Criminal Legal Risks and Compliance Governance of Generative Artificial Intelligence

Liu Shuang, Qi Min

Abstract: In the era of big data, generative artificial intelligence has developed rapidly, showing strong creativity and adaptability. However, due to its operating principle and endogenous defects, generative AI may cause criminal risks such as endangering national security, infringing personal information, intellectual property rights, commercial secrets, and violating citizens' personal and property rights at various stages of input, calculation and output. The ides of forward-looking governance should be adhered to, the modesty of criminal law should be upheld, excessive involvement of criminal law should be avoided, and a triple-graded compliance governance system should be constructed with regard to compliance beforehand, supervision during the process, and rectification after the process, with a view to effectively addressing the risk of criminal liability, while providing rule of law safeguards for the wide application of AI technology.

Key words: Generative Artificial Intelligence; Criminal Risks; Graded Compliance Governance