doi:10.3969/j.issn.1672-626x.2023.06.009

生成式AI算法训练风险的合规管理研究

罗世杰[1],贺国荣[2]
(1. 西南政法大学 经济法学院,重庆 401120;2. 四川农业大学 公共管理学院,成都 611130)

摘 要:生成式AI算法训练活动容易引发数据违法或质量低劣、算法偏差(错误)、偏见与歧视以及触碰反垄断红线等诸多风险,对社会与经济产生显著负外部性。而当前的风险治理机制面临防范不足、规制滞后、治理失范等困境,故而应考虑将生成式AI算法训练风险引入合规管理。生成式AI算法训练风险的依法治理和包容治理证成了对其进行合规管理的正当性和可行性。具体到生成式AI算法训练风险合规管理的实践路径:首先对其法律制度体系进行梳理,据此提炼和总结出生成式AI企业具有算法训练合规管理的法定义务,并且明晰对其全方面的法律监管要求。然后生成式AI企业应据此实施算法训练合规管理,主要包括实施原则的确定和实施方案的具体构建。前者包含生成式AI算法训练计划、过程以及结果的全流程合规,后者包括构建算法训练合规管理基础性平台和对算法训练全流程合规风险的处理。同时,为保障生成式AI企业依法实施算法训练合规管理,需优化企业对外面向的行政监管与激励机制。

关键词:生成式AI;算法训练风险;风险治理;合规管理;算法合规
中图分类号:TP18;D922.1 **文献标志码**:A **文章编号**:1672-626X(2023)06-0106-10

2022年底,OpenAI正式推出其研发的智能对话机器人ChatGPT,并迅速在全球范围内爆火。2023年4月,其再度推出GPT-4,并公布其将联网,意味着生成式AI技术再达高峰。生成式AI技术流程大致可以分为训练和生成两个部分,通过学习大量数据并进行模型训练,能够自动生成符合特定领域规律的全新内容。生成式AI算法训练具有前所未有的独特性——其训练过程自主性较强、训练结果可控性较低,训练阶段的效果直接影响模型的优劣,从而左右生成数据的质量①。但生成式AI技术的研发与应用也是一把双刃剑:一方面,ChatGPT等一系列生成式大语言模型的研发与应用意味着生成式AI技术达到领域前沿,其技术价值丰富、应用场景广泛,将大力促进经济和社会发展;另一方面,其崛起也引发诸多负外部性,既包括知识产权归属、商业秘密保护、数据与算法安全等法律问题,还包括就业替代、人类主权失范等伦理问题。

在此背景下,我国已展开相关立法活动:2023年6月,《人工智能法草案》预备提请全国人大常委会审议;同年7月,出台《生成式人工智能服务管理暂行办法》(以下简称《服务办法》)。这说明我国即将进入人工智能强监管时代,而生成式AI算法训练风险的防范与规制也一跃成为全新且热潮的法律课题。

一、生成式AI算法训练风险的类型化梳理

在生成式AI技术不断升级并被广泛应用背景下,作为其核心的算法模型训练是法律风险密集的环节。

收稿日期:2023-07-01
基金项目:重庆市研究生科研创新项目(CYS23292);西南政法大学经济法学院学生科研创新项目(2022-XZJF-016)
作者简介:罗世杰(1999-),男,四川内江人,西南政法大学经济法学院硕士研究生,研究方向为经济法、人工智能法;贺国荣(1980-),男,山西永济人,四川农业大学公共管理学院讲师,法学博士,研究方向为刑法、人工智能法。

生成式AI算法训练以数据、算法、算力以及算法标注等重要元素为基础，其训练全过程（计划、过程与结果）可能引发各类风险。以GPT-4为例说明，生成式AI算法训练引发的风险包括如下几类：

（一）生成式AI算法训练的数据合法性与质量风险

其一，生成式AI算法训练的数据获取途径合法性问题。一方面，在生成式AI算法训练过程中，易出现数据获取途径不透明问题。目前GPT-4算法模型仍系“算法黑箱”，OpenAI没有向外界披露其所使用数据的来源。且随着GPT-4的训练数据库联网，其违法违规抓取互联网信息以获得训练数据的风险增大。另一方面，生成式AI算法训练易引发违法抓取个人数据问题。在个人数据安全层面，生成式AI可能未经用户同意就进行大量抓取。此外，GPT-4的“预学习”无需人工介入、标注和监督，导致GPT-4在获取预训练数据的效率上很难受人的干预和调控，因而违法抓取个人数据的情况无可避免[1]。

其二，生成式AI算法训练阶段生成的数据合法性与质量问题。一方面，算法训练生成的数据遭遇极大合法性挑战。由于用于算法训练的数据可能不准确或存在倾向性，故难以保证其合法性，导致生成的数据也极有可能具有“毒性”[2]。易言之，GPT-4的算法训练所需数据量极大，而该部分数据通常涉及隐私、权属、公平竞争等问题，很可能违背相关法律和伦理规范。若对所获数据内容的合法性和合规性不予置评和纠正，基于其进行的算法训练也会类似“蝴蝶效应”继续生成不合法数据。另一方面，“毒树之果”②效应蔓延至生成式AI算法训练过程中的数据质量方面，用低质量数据生成低质量数据的情况无可避免。同时，也有可能本不属于低质量的数据在训练过程中被“污染”变成低质量数据。

其三，生成式AI算法训练所涉数据的泄露与滥用问题。一方面，生成式AI在算法训练过程中，极易遭受数据泄露问题，这种数据泄露可能系人为，也可能系算法训练技术本身问题。且随着愈来愈多领域对生成式AI加以应用，尤其在算法训练的过程中，数据泄露成为一大隐患。因为数据作为重要生产要素，一旦泄露将给企业、行业带来重大损失。另一方面，生成式算法训练所需数据面临被滥用问题。由于生成式AI的算法训练过程具有强大的模仿与生成能力，不法分子可能利用其整合与生成虚假信息，引发社会安全与经济效益问题。且即使是零碎信息，GPT-4也可能将其与其他类型数据拼合在一起进行挖掘分析，从而推断出关系国家安全、公共安全、个人和组织合法权益的信息[3]。从而可能影响到国家、社会和个人利益安全。

（二）生成式AI算法训练的偏差（错误）、歧视与偏见风险

生成式AI算法训练在“预学习”时没有过滤机制和人工监管，导致其在算法训练过程中，可能生成虚假或不良信息，也可能产生算法歧视、偏见风险。

其一，生成式AI算法训练引发信息链条式虚假或低质量的潜在风险。一方面，如前文所述，GPT-4模型在“预学习”时没有人工监督，获取的数据未经过实质性筛查和挑选，数据在源头处即存在内容不合法（规）、虚假或错误问题，从而影响模型训练结果的正确性和中立性。另一方面，尽管OpenAI在开发GPT-4时已经极力避免输出带有偏离社会价值观导向的内容，但在GPT-4算法训练过程中，其可能遭遇恶意“投毒”，导致训练出的算法模型遭到污染，也会诱使其通过训练得到的语言模型输出不良或虚假数据，形成连环效应[4]。此外，在生成式AI算法训练场景中，算法标注是实现大规模训练数据机器学习的关键所在。所谓“算法标注”，即对生成式AI原始数据集进行标注、分类、分析和清洗，以助力训练机器学习算法和人工智能模型[5]。而在生成式AI算法标注场域中，存在标注内容不合规、标注规则本身违法或标注人员不符要求等风险，从而导致算法训练过程与结果出现内容与价值偏移。

其二，在生成式AI算法训练的动态风险场域下，算法歧视和偏见风险亦暗潮涌动。有学者认为，在生成式AI时代，算法歧视存在朝“无意识”转变的可能性甚至是趋势[6]。而本文讨论的算法歧视与偏见风险主要指被训练出来的生成式AI算法模型带有对国别、种族、职业、年龄、性别甚至是数字穷富差距的甄别与择

“强”倾斜[7]。这种歧视与偏见一方面来自于生成式AI算法训练标注者的歧视因素，另一方面则来自于生成式AI本身的技术中立性异化[8]。生成式AI的算法训练本质是利用算法对大量的数据进行处理，但生成式AI算法模型本身还不能对数据进行查证与甄别，常常可能生成看似准确但本质错误的误导性内容。而且，生成式AI算法训练技术本身无法完全避免社会偏见和价值观倾向，从而可能固化社会偏见和歧视[2]。

（三）生成式AI算法训练易触碰反垄断红线

其一，生成式AI技术层面的垄断。GPT-4的发布在各个领域引起了巨大的轰动。在某些领域，其已经显露绝对优势，或者说“AI霸权”。OpenAI为维护其已经占据的优势地位，继续研发升级模型算法和优化性能，从而进一步巩固其领先地位，扩大“雪球效应”，以此实现技术市场垄断。正如美国联邦贸易委员会（FTC）下属的竞争局表示③，生成式AI依赖于一系列必要的投入。如果一家公司或少数几家公司控制着这些重要投入中的一种或几种，他们也许能够利用控制权来抑制或扭曲生成式AI技术市场的竞争，控制整个生成式AI技术市场的进入与退出。

其二，算法训练过程中的数据垄断。尽管数据系非竞争性产物，但在其释放经济价值的过程中，具有较高效益的数据作为稀缺资源必然造成竞争现场。在生成式AI算法训练所涉及的数据获取、训练或输出等环节，竞争壁垒都有出现的可能性，形成数据垄断[9]。即生成式AI算法训练需要海量数据，如果该数据来自于一个特定的数据集中，那么可能会存在垄断风险，包括数据集中和数据共享中的垄断风险。

二、生成式AI算法训练风险引入合规管理的必要性证成

生成式AI算法训练风险被认为是生成式AI企业算法训练活动产生的显著负外部性，对涉及的数据与算法安全、数字经济以及相关法律制度造成冲击。从法学角度上看，负外部性是一个法律主体在享受权利时将相对应应当承担的义务和责任施加给其他主体[10]。为应对这一负外部性，学者们已提出了一定的对策建议，有的基于风险治理视角的路径，提出通过对生成式AI进行风险定级进而监管其算法模型[11]；有的基于主体治理视角的路径，提出以生成式AI算法主体责任机制奠定我国生成式AI算法问责机制的运作基础，具体包括备案、解释以及问责[12~13]；有的基于应用治理视角的路径，指出不以出台生成式AI专门规制法为目的，而是在其具体应用场域中以单行法律和法规的形式施以针对性治理等[14]。

然而上述措施被单独或空泛使用时可能遭遇事前防范不足、事后规制滞后的治理困境，且其均基于“硬法”规制视角，缺乏从法律主体自治角度提出“软法”治理之策[15]。基于此，本文拟从合规管理视角提出生成式AI企业算法训练合规管理，其蕴含的依法治理和包容治理相结合的法律治理功能可以实现生成式AI算法训练法律风险治理之目标，促进生成式AI算法模型训练的良性发展。

（一）生成式AI算法训练风险的依法治理

当地时间2023年6月14日，欧洲议会投票通过了《人工智能法案》，对ChatGPT等生成式AI工具提出了新的算法训练透明度、安全性等要求[16]，相关企业应依法合规进行算法训练，算是打开了生成式AI算法训练依法治理之路。这对我国生成式AI算法训练风险的依法治理具有重要参考意义。生成式AI算法训练风险的产生直接影响生成式AI技术的健康发展，因此有必要在实现算法向善治理的目标基底上寻求依法治理路径，对算法训练风险进行约束与防范[17]。具体到生成式AI算法训练风险的依法治理，不仅要确立法律在生成式AI研发、生产、运用中的权威性，而且要提升其研发、生产、运用企业及其员工按照人工智能相关法律规范行事的行为习惯[18]。

企业出于防范法律风险的目的自主选择合规之路，按理说没有必要再将建立合规计划和实施合规管理

作为企业法律义务[19]。但是，在依法治理生成式AI算法训练风险过程中，法律规范的行为规范功能必不可少。当前算法训练风险治理的法律路径主要有两种，即“硬法”和“软法”治理。本文所讨论的生成式AI算法训练合规管理，即属于“硬法”与“软法”协调治理的范畴。具言之，一是算法训练风险的“硬法”治理，“硬法”视角下的生成式AI算法训练风险治理主要体现为行为规制与法律权利义务的配置。“硬法”路径在功能定位上侧重于制裁与惩罚，主要通过命令来强迫算法训练主体作出某种行为选择，其大多表现为刚性的强行性规范，内容具有明确性和稳定性[20]。“硬法”规范可以给生成式AI算法训练主体的算法训练合规管理提出依法建设要求，使其实施得到强制性保障。二是算法训练风险的“软法”治理，即那些效力结构未必完整、不需要依靠国家强制力才能保障实施，但是能够产生实际治理效益的法律规范。“软法”可以为生成式AI企业实现算法训练合规管理提供目标清晰、操作性强的具体指引，使其合规建设实现“硬法”与“软法”依据的双重满足。故而在当前积极推进企业合规背景下，要发挥法律规范对生成式AI企业及其员工算法训练行为的指引功能，离不开企业合规管理体系的构建。

（二）生成式AI算法训练风险的包容治理

《服务办法》第3条明确规定：“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。”这一规定明确我国对生成式AI的治理模式是“包容性法律治理”。包容性治理包括“包容”和“治理”两个关键内核[21]。社会治理角度的包容性治理指通过制度的安排，能够确保所有公民平等参与政策的制定，并享有平等分配资源权利的过程[22]。延伸到法律治理领域，就包容性治理主体而言，是指不故意排斥法律治理的主体，广泛吸收多元个体参与法律治理机制。换言之，多元主体的合作共治是包容治理在社会治理层面的必然要求。故在生成式AI算法训练风险的治理维度，包容治理就要求积极引入社会、企业等多方主体参与，形成政府领导下的多元主体包容共治格局。其包括以下三个方面：

一是生成式AI企业自治。以算法训练合规管理为依托，参考美国对生成式AI算法训练遵循相对宽松的治理策略，其并没有一味地出台严密的法案来对生成式AI的算法训练活动进行规制和限制，而是采取基于“市场自由”的治理路径[23]。即基于生成式AI企业具有“市场自由”，美国政府并不要求其进行算法训练设计时必须引入或删除某种设计要素或训练数据，也不要求作为“人的集合”的生成式AI企业进行算法训练内容联网审查。二是行政监管。当生成式AI企业在算法训练合规义务履行阶段出现或可能出现合规风险时，需要介入其“看门人”的角色，对义务主体进行专门监管，促使其算法训练活动回归合规与合法轨道[24]。即在生成式AI监管的法律框架下，行政监管部门应当承担起生成式AI企业算法训练合规管理义务履行的监管角色。三是社会参与监督。公众参与对实现算法训练合规的功能和作用不言而喻。社会大众是生成式AI算法训练活动的受众和成本与利益的承担者，拥有对义务主体的算法训练合规管理过程与效果的监督、讨论、意见反馈等权利。

三、生成式AI算法训练合规管理的法律依据

目的是一切法律规范的缔造者[25]。如仅将生成式AI算法训练合规管理作为实验室工具，即生成式AI企业关于正确进行算法训练的内部指引和规定，但并不作为法律客体呈现，也就无所谓构建其法律体系。然而，本文系出于生成式AI算法训练风险的依法治理和合法治理需求，并对其合规管理的依法建设进行证成和构建，故有必要认识到：此处的合规管理，并不是企业“自娱自乐”的工具，即并非只是用来对企业成员进行内部约束的手段。其是被用于算法训练风险治理、带有法律色彩的工具，应当被赋予法律涵义——合规

法律义务以及不履行合规义务的法律责任。即应在现有生成式AI算法训练合规管理的法律依据体系中剖析出其义务主体及具体的合规监管要求。

（一）生成式AI算法训练合规管理的法律制度体系

生成式AI算法训练合规管理的“规”不仅包括企业内部管理规定，更应涵盖法律法规和其他规范性文件的相关规定。我国目前生成式AI算法训练风险治理的相关法律依据还较少，仅零星见于《服务办法》《中华人民共和国数据安全法》《中华人民共和国反垄断法》《中华人民共和国网络安全法》《中华人民共和国个人信息保护法》等法律以及一些专门规制算法的监管法规如《互联网信息服务管理办法》《互联网信息服务算法推荐管理规定》和《互联网信息服务深度合成管理规定》的具体条文中。

而在生成式AI算法训练合规管理的企业方案构建和具体实施上，我国已经发布的《中央企业合规管理指引（试行）》《个人信息保护合规审计管理办法》《企业知识产权合规标准指引（试行）》《企业境外经营合规管理指引》《中央企业合规管理办法》《经营者反垄断合规指南》《合规管理体系要求及使用指南》等文件，能够为其提供重要参考。

（二）生成式AI算法训练合规管理的法定义务主体

确定生成式AI算法训练合规管理的法定义务主体具有不可或缺的作用。生成式AI算法训练合规管理的法定义务主体范畴可根据以上生成式AI算法训练相关法律法规来确定。《服务办法》给我国生成式AI的合规主体指明了大致方向和基本定义，即“利用生成式人工智能算法提供聊天和文本、图像、声音生成等服务的组织”，但是生成式AI算法训练合规管理的义务主体应在此轮廓下进行以下两方面的辨析：

其一，将义务主体中的“服务提供者”范畴限缩到“提供生成式AI算法训练活动的供应商和第三方服务机构”。其中，供应商是进行生成式人工智能算法训练活动的最主要提供者，需要对算法训练的质量和可靠性负责。在算法训练过程中，供应商需要确保算法的准确性和效率，同时也需要遵守相关的法律法规和标准，以确保算法的训练过程和行为符合监管要求；而第三方服务机构是协助供应商进行算法训练的服务提供商。这些服务机构需遵守算法训练领域相关的法律法规和标准，以确保其算法训练活动符合监管要求。

其二，借鉴美国2022年《算法责任法案》将“个人”纳入算法责任主体的做法和规定[④]，应将“从事生成式AI算法训练活动的个人”也纳入义务主体的范围。即虽然在法律的制度和实施层面，认为企业主体才是对外承担算法训练合规责任主体，但是在算法训练合规具体实践中，企业主体内部须对“从事生成式AI算法训练活动”的个人进行合规义务的“分配”与“监督”，纳入企业内部合规管理机制，相关个人实际上也将对生成式AI算法企业合规风险承担责任和不利后果。当然，将“从事生成式AI算法训练活动的个人”纳入义务主体的基础在于法律法规的明文规定或者企业内部算法训练合规的管理制度与细则的明确规定。所以为更好地对生成式AI算法训练风险进行治理，需要进一步出台相关合规监管法规；生成式AI也应当尽快确立算法训练合规管理制度和细则，以确定个人义务主体。

（三）生成式AI算法训练合规管理的法律监管要求

相比域外生成式AI算法训练的治理行动，我国走在了前列，如《服务办法》的迅速出台。本文特对《服务办法》中生成式AI算法训练的相关监管规定进行共性梳理与要点分析，结合域外相关制度经验，探讨生成式AI企业算法训练合规管理依法建设的法律监管要求。

其一，算法训练的数据合规建设要求。生成式AI企业应当对算法训练阶段所涉及被获取、整合、训练、输出的全部数据的合法性与安全性负责并做好合规建设。具体应满足以下要求：符合算法训练活动安全相关法律法规的要求；不含有侵犯知识产权的内容；数据包含个人信息的，应当征得个人信息主体同意或者符合法律、法规规定的其他情形；保证算法训练数据的真实性、准确性、客观性。

其二，算法训练的人工标注合规建设要求。对算法训练中体现人类偏好的人工标注的要求和约束必不可少。当算法训练中采用人工标注时，生成式AI企业应当制定清晰、具体、可操作的标注规则，对标注人员进行必要培训，并抽样核验标注内容的正确性。

其三，算法训练的行为标准合规建设要求。为更好地对生成式AI企业的算法训练合规管理标准提出要求，应当从纠偏与归正两个方面对其进行展开，即生成式AI企业应当建立用户投诉接收处理机制以实现算法训练行为纠偏，应通过算法模型优化训练来作出算法训练行为归正。

四、生成式AI算法训练合规管理的依法建设

当前学界对于企业合规作用机理的探讨主要分为“外部应对”和“内生驱动”[26]。无论是由“外部应对”走向“内生驱动”，还是以“内生驱动”配合“外部应对”，具体到生成式AI算法训练合规管理领域，都应从“内外有别”转至“内外协调”思路，即应力求在企业自行依法建设内部合规体系的同时，兼顾其与外部行政监管与激励的协调。故而要求生成式AI企业必须建立内涵丰富、体系完整且适应自身需求的制度程序并探索出能够促进该体系有效实施的合规措施与举动[27]。同时国家和政府层面还应配合和保障生成式AI企业实施算法训练合规管理，即探索出一条内外双面向、完备的算法训练合规管理的依法建设进路。

（一）生成式AI算法训练合规管理的企业实施

1. 生成式AI企业算法训练合规管理的实施原则

企业合规管理，既要合行业之规、企业之规，又要合法律之规以及道德之规。生成式AI企业如何准确、有效实施算法训练合规管理呢？即生成式AI企业在被施加算法训练合规管理的法定义务后，必须依法建设算法训练风险的合规管理实施体系与机制。生成式AI企业在具体实施算法训练合规管理方案前，应确定合规管理的实施原则。这套实施原则来源于前述法律对于算法训练合规管理的监管要求以及生成式AI企业自身在算法训练合规管理中的实际需求。

生成式AI算法训练合规管理的实施原则主要包括三个方面。一是生成式AI算法训练计划合规管理，要求生成式AI算法训练主体在训练模型设计与计划之初就应合理地预测和预防风险。具体包括：生成式AI企业算法训练计划的事前审查与批准；向公众提供服务前应做好事前审批合规建设以及应提供可以影响用户信任、选择的必要信息。二是生成式AI算法训练过程合规管理，要求生成式AI算法训练主体在算法训练过程中对风险进行把控与控制。生成式AI企业应确定好在算法训练过程中本企业以及具体岗位的合规义务及履行标准，如对算法训练的重点环节进行分类，然后对每个阶段的岗位合规义务、合规义务来源、合规风险描述以及风险后果和应对措施进行安排，并按部就班根据标准完成。三是生成式AI算法训练结果合规管理，要求生成式AI算法训练主体对算法训练结果的风险承担责任，采取合规措施尽量减小损害与算法训练的负外部性。生成式AI企业应对算法训练结果负责，对其进行合规化评价和处置。具体包括：实名认证与算法防沉迷合规；后端处理合规；通过模型算法优化训练等方式防止再次生成以及暂停或者终止服务，等等。

2. 生成式AI企业算法训练合规管理的实施方案

其一，构建算法训练合规管理基础性平台。基础性合规平台的构建是生成式AI企业实施算法训练合规管理的基础，其主要包括算法训练基础性合规要素和专门性合规要素两个方面内容的确立。首先是算法训练基础性合规要素的构建。一个企业构建任何一项合规计划，都应引入必要的基础性合规要素[28]。生成式AI企业的算法训练合规管理计划也应当在建立起合规管理基础性平台的前提下，展开专项合规管理和内部

监督。而这里所述的基础性平台，应当将与生成式AI算法训练合规管理相关的制度型要素建立起来，并使之在合规管理中得到实际的运作和生效，有力防范违法违规的算法训练行为[29]。基础性合规要素在生成式AI算法训练合规管理体系中可以发挥合规管理制度平台的作用，帮助生成式AI企业确立算法训练合规管理理念和基本准则，并配置预防算法训练合规风险、监控算法训练合规管理和应对算法训练违规行为的相应管理程序。然后是算法训练专门性合规要素的确定。当生成式AI企业确定了基础性合规要素之后，还需要确定算法训练专门性合规要素。即生成式AI企业针对特殊的算法训练合规风险，为防止发生特定的违法违规事件而建立的专门性合规管理制度，其处于核心性和保障性地位，包括算法训练合规政策、算法训练培训内容、算法训练岗位合规手册和职责安排以及算法训练合规持续改进机制等在内的各项要素。在算法训练合规管理体系中引入专门性算法训练合规要素，可促使生成式AI企业在因算法训练不合规而受到相应约束或者制裁后，进一步完善算法训练合规管理，预防相应特定约束或者制裁的再次发生。

其二，对算法训练全流程合规风险的处理。生成式AI企业在进行算法训练的过程中，应及时、全面地识别可能出现的各项合规风险并尽可能作出合规预防与处置。最保险和全面的措施，即在生成式AI企业内部构建算法训练合规风险处理机制。首先是算法训练合规风险识别环节。合规风险识别是生成式AI企业算法训练合规风险预防与治理的逻辑起点。这一环节主要是查找与识别在算法训练链条中可能出现的合规风险，经由事前调查只做出客观的判断，履行提示注意义务，而不作进一步的价值评价。需要特别注意的是，该环节识别的应为具体、客观存在且具备重要性的算法训练合规风险。其次是算法训练合规风险评估环节。风险评估是生成式AI企业算法训练合规体系功能得以实现的核心部分。如果说生成式AI企业内部算法训练合规风险的识别环节是“一兜子”操作，那么评估环节就是将“一兜子”里的风险进行“筛选”，将对企业内部有重大影响和对外部具有显著负外部性的算法训练风险挑选出来，以进行下一环节的风险处置。最后是算法训练合规风险应对环节。风险应对环节是生成式AI企业算法训练内部合规治理机制运行的后置保障。这一环节是生成式AI企业把握前述被识别与评价的重要风险，采取可行度高的调节举措，建立生成式AI算法训练合规风险处置机制。

（二）生成式AI算法训练合规管理的行政监管

相较于对内面向的生成式AI企业自主履行算法训练合规义务，对外面向的行政机制可以充分发挥其监管（约束）与激励功能，一方面以提高成本或增加责任的方式促使企业实施算法训练合规管理，另一方面以增大效益或减轻责任的方式激励企业建设算法训练合规管理。行政监管具体包括以下措施：

1. 生成式AI企业算法训练的专门备案

算法备案制度是“有效市场与有为政府相结合”的治理原则在互联网和算法领域的拓展和创新，是在数字法治的实践背景下推进算法治理的一项重要举措[30]。算法训练专项备案环节系针对算法训练计划和实施过程而言：生成式AI企业在算法训练合规管理的计划和过程中，需要对外配合行政监管程序、进行算法训练计划与过程中相关信息的报备与披露。

其一，首次备案。当生成式AI企业设计完成算法训练计划时，即应进入备案程序，此为首次备案。首次备案主要应考虑其范围和内容。范围方面，要考虑的因素包括是否具有备案的必要性、是否具有备案的可操作性以及一般追责方式是否能够满足相关的维权需求。内容方面，本文认为除已有的算法主体信息、算法信息、产品及功能信息等三项算法备案通用信息以外，还应包括一些算法训练特定内容，如算法训练合规管理计划的整体方案、方案实施步骤以及方案安全评估情况信息，生成式AI企业内部算法训练合规实施的关键环节等。

其二，变更备案。生成式AI企业在算法训练过程中，若算法训练实施与计划方案有重大变化时，应将变

化后的方案以及与原方案出入较大的对比信息提交备案。而这种“重大变化”应在首次备案时，就已经一同上传备案系统并通过备案主管部门的审核。若备案主管部门认为生成式AI企业实施算法训练过程中实际系“重大变化”但未记载于首次备案记录中的，可自行决定是否要求生成式AI企业进行变更备案。

2. 生成式AI企业算法训练的专项审计

存在重大风险甚至可能影响国家安全的项目，都可以纳入审计范畴，域外有此先例，如英国、荷兰最高审计机关均对政府部门的模型算法进行审计[31]。我国已开始有这方面的行动，2023年8月3日，国家互联网信息办公室就《个人信息保护合规审计管理办法（征求意见稿）》公开征求意见，首次正式提出“个人信息保护合规审计”，也为我国算法训练合规审计的设立提供思路与参考。有鉴于此，需要从以下两个方面展开对其算法训练合规专项审计项目的设立。

其一，内部审计。这是针对企业内部而言，即生成式AI企业自行开展算法训练合规审计，根据实际情况，由其内部机构或委托专业机构展开。其可以深入算法训练的整个运行过程，访问、审查算法训练的数据，测试算法训练过程，审查算法训练的数据参与集、参数信息等重要信息以及整合外部审计无法得到的算法训练过程中的细节，从而精确地对算法训练合规风险进行识别与把控。

其二，外部审计。这是针对企业外部而言，包括国家审计和社会审计。国家审计，即履行生成式AI算法训练合规监管职责的部门在履行职责过程中，发现生成式AI算法训练活动存在较大风险或者发生安全事件的，可以要求生成式AI企业委任专业机构对其算法训练活动进行合规审计。而社会审计，是社会类审计机构“参与式审计”生成式AI企业算法训练合规建设与管理情况[32]。相比内部审计和国家审计，其更具变通性，系前两者的有益拓展与补偿，能够有效监督生成式AI企业履行算法训练合规义务，形成合规审计合力。

（三）生成式AI算法训练合规管理的行政激励

有学者认为，企业如果能够建立优良的合规管理机制，更加可能与行政机关达成行政和解协议[33]。因为激发企业内部驱动力是促使算法训练合规管理依法建设和落实的关键。故而目前针对生成式AI企业算法训练合规管理的行政执法需要由主动执法向被动执法转变，这就需要在行政执法环节，通过与生成式AI企业达成行政和解协议或者给予其行政处罚减免等方式来推动其实施算法训练合规管理[34]。

其一，与生成式AI企业达成行政和解协议来促使其履行算法训练合规管理义务。在美国，行政和解协议的使用更加常见，例如美国证券交易委员会（SEC）对企业涉嫌违法违规的案件，95%都是以行政和解的方式处理的[35]。具体到算法训练合规管理的行政领域而言，国家行政监管部门起草生成式AI算法训练行政合规指南，在允许监管部门与生成式AI企业达成行政和解协议的前提下，企业按照和解协议内容自行完成算法训练合规管理义务的积极履行或者对违规行为进行整改，然后就可以根据和解协议不再对其行政处置。

其二，给予生成式AI企业行政处罚减免以激励其实施算法训练合规管理。现行法律应当在已有的行政和解基础上，进一步增设算法训练合规激励的“软法”条款，明确将生成式AI企业算法训练合规管理的实施作为行政免予处罚的依据，建立相应的附条件免予处罚制度[36]。

五、结论与启示

随着生成式AI的迅猛发展，其未来应用场景充满无限的可能，不仅改变人们的生活方式和思维模式，而且将对技术跃迁和人类社会产生巨大影响。目前已有部分行业着手进行生成式AI在各个场景的研发与应用，如生成式AI金融服务、生成式AI医疗客服和生成式AI监管科技等。但是技术的的开发与使用始终伴有风险，需要对其进行善治。我国虽然针对生成式AI技术的应用专门出台了《服务办法》，但是总体持包容性

法律治理的态度，支持治理主体、治理方式和治理结果的包容性和多元化。本文立足生成式AI算法训练的内发性与自主性，并结合合规依法建设的经验，从现实需求与理论层面证成了生成式AI算法训练风险引入合规管理的必要性、合理性与可行性。但当前我国存在对其制度供给不足的困境，故需要结合法律解释的功能，将生成式AI算法训练合规管理的法律依据进行梳理：不仅将生成式AI算法训练合规作为一种法定义务进行考察，并且将目前生成式AI算法训练合规管理相关法律体系以及全面监管要求进行特别地整理与说明。据此构建生成式AI企业内外双重面向的算法训练合规管理的依法建设机制，即包括生成式AI企业算法训练合规管理的实施方案以及其外部合规监管与激励。诚然，现阶段我们正在逐渐走向强人工智能时代，且技术将持续不断地发展。不论是“人工智能法”的出台，还是本文所述算法训练合规管理的依法建设，都不会是生成式AI算法训练风险治理的终点，而只是该征程的阶段性成果，亟需更多法律化手段的出现，以实现对人工智能风险的防范与治理。

注　释：

① “探秘生成式AI的工作逻辑：解码创造力的神秘密码”，载https://www.sohu.com/a/695177266_121712227，于2023年8月15日访问。

② “毒树之果”（Fruit of the Poisonous Tree）是美国法律术语。在美国指的是调查过程中，通过非法手段取得的证据，该术语的逻辑是如果证据的来源（树）受到污染，那么任何从它获得的证据（果实）也是被污染的，在诉讼审理的过程中将不能被采纳，即使该证据足以扭转裁判结果亦然。

③ “美国反垄断监管机构：生成式AI引发竞争担忧”，载https://www.chinaz.com/2023/0630/1538943.shtml，于2023年8月14日访问。

④ 2022年美国《算法责任法案》，载https://www.wyden.senate.gov/download/algorithmic-accountability-act-of-2022-bill-text，于2023年7月30日访问。

参考文献：

[1] 邓建鹏，朱怿成.ChatGPT模型的法律风险及应对之策[J].新疆师范大学学报（哲学社会科学版），2023（5）：91-101+2.
[2] 支振锋.生成式人工智能大模型的信息内容治理[J].政法论坛，2023（4）：34-48.
[3] 崔议文，皮勇，张凌寒，等.“ChatGPT带来的风险挑战及法律应对”三人谈[J].人民检察，2023（7）：37-44.
[4] 於兴中，郑戈，丁晓东.生成式人工智能与法律的六大议题：以ChatGPT为例[J].中国法律评论，2023（2）：1-20.
[5] 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.
[6] 杨玉晓.人工智能算法歧视刑法规制路径研究[J].法律适用，2023（4）：86-94.
[7] 孙伟平.智能系统的“劳动”及其社会后果[J].哲学研究，2021（8）：30-40+128.
[8] 刘瑞生.传播“重构”与技术“异化”视角下的算法辨析[J].西南民族大学学报（人文社会科学版），2022（6）：164-172.
[9] 张蕴萍，翟妙如.重视数据要素价值释放中的反垄断治理问题[N].光明日报，2023-02-24（6）.
[10] 胡元聪，廖娟.人工智能的负外部性及其经济法规制[J].大连理工大学学报（社会科学版），2020（3）：71-79.
[11] 曾雄，梁正，张辉.欧盟人工智能的规制路径及其对我国的启示——以《人工智能法案》为分析对象[J].电子政务，2022（9）：63-72.
[12] 刘权.论互联网平台的主体责任[J].华东政法大学学报，2022（5）：79-93.
[13] 张亮.算法治理须抓牢主体责任“牛鼻子”[N].法治日报，2022-01-12（5）.
[14] 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（3）：108-123.
[15] 李驰.中国-东盟“软法-硬法”嵌合治理体系建构与前瞻——基于法治“一带一路”视野[J].广西社会科学，2022（11）：52-61.
[16] 林子涵.欧盟谋求AI监管领域主导权[N]. 人民日报（海外版），2023-07-29（6）.
[17] 吴太轩，周珊珊.算法正义软法保障的理论证成及具体路径[J].政法学刊，2023（1）：118-128.
[18] 霍俊阁.ChatGPT的数据安全风险及其合规管理[J].西南政法大学学报，2023（4）：98-108.
[19] 龙宗智.我们需要什么样的合规不起诉制度[J].比较法研究，2023（3）：74-83.
[20] 于升.经济法的软法之治研究[D].长沙：湖南大学，2018.

[21] 尹利民，田雪森.包容性治理：内涵、要素与逻辑[J].学习论坛，2021(4)：66-74.
[22] 郭小东.生成式人工智能的风险及其包容性法律治理[J/OL].北京理工大学学报(社会科学版)：18[2023-10-01].https://doi.org/10.15918/j.jbitss1009-3370.2023.1340.
[23] 赵海乐.数字经济中的算法治理：美欧路径差异与中国策略[J].国际经贸探索，2023(5)：107-120.
[24] 张旭，田园.算法治理视阈下的企业合规：困境、逻辑与进路[J].兰州大学学报(社会科学版)，2022(2)：90-99.
[25] 菲利普·热斯塔茨，克里斯托弗·雅曼.作为一种法律渊源的学说[M]. 朱明哲，译.北京：中国政法大学出版社，2020：375.
[26] 解志勇，石海波.企业合规在行政执法和解中的导入研究[J].行政法学研究，2023(5)：66-78.
[27] 魏婷婷.平台垄断的治理转型：合规体系的适用逻辑及实践路径[J].湖南科技大学学报(社会科学版)，2023(1)：117-123.
[28] 陈瑞华.企业合规整改中的专项合规计划[J].政法论坛，2023(1)：28-44.
[29] 李勇.涉罪企业合规有效性标准研究——以A公司串通投标案为例[J].政法论坛，2022(1)：132-146.
[30] 张吉豫.论算法备案制度[J].东方法学，2023(2)：86-98.
[31] 王玉凤.模型算法审计：理论内涵、国际经验与审计框架[J].审计研究，2023(3)：11-18.
[32] BRIANA VECCHIONE，KAREN LEVY，SOLON BAROCAS. Algorithmic Auditing and Social Justice：Lessons From the History of Audit Studies[J].EAAMO'21：Equity and Access in Algorithms，Mechanisms，and Optimization，2021，10(19)：23-38.
[33] 陈瑞华.企业合规基本理论[M].北京：法律出版社，2020：309.
[34] 崔永东.从法律激励视角看企业合规[J].法治研究，2023(1)：123-132.
[35] 陈瑞华.企业合规的基本问题[J].中国法律评论，2020(1)：178-196.
[36] 陈悦.行政处罚制度完善的便宜主义进路[J].苏州大学学报(哲学社会科学版)，2020(2)：94-103.

(责任编辑：何　飞)