Open Journal of Legal Science 法学, 2024, 12(8), 5210-5222
Published Online August 2024 in Hans. https://www.hanspub.org/journal/ojls
https://doi.org/10.12677/ojls.2024.128742

Hans汉斯

生成式AI商业数据获取行为的规制路径

——以ChatGPT为例

林佩鑫

武汉科技大学法学与经济学院，湖北 武汉

收稿日期：2024年6月21日；录用日期：2024年7月17日；发布日期：2024年8月21日

摘　要

数据被公认为推动技术进步的基石与核心所在，而以ChatGPT为典型的生成式AI技术在商业数据获取领域遭遇了不可忽视的合规风险。由于生成式AI缺乏相应的法律地位，这意味着研发者、设计制造者、提供者以及使用者都需要承担起对应的合规风险责任。为了合理分担这一责任，链式责任机制被建议采用。在此基础上，鉴于《生成式人工智能服务管理暂行办法》的存在，对于数据获取和惩罚机制，有必要进行更为详尽、细致的限制与要求设定。此外，全面引入“监管沙盒”制度并建立专门的机构或委员会也具有重要意义。通过这些举措，不但能够规范行业秩序，为数据的合法使用提供坚实保障，还能在推动技术创新的同时，确保其安全、合规地运行。面对生成式AI技术带来的挑战，我们必须持续探索并完善相关的制度与机制。只有这样，才能实现技术的健康发展与高效应用，进而充分释放数据的潜在价值，推动技术的不断进步，营造出共赢的良好局面。同时，这也有助于构建一个公平、有序、可持续发展的技术生态环境。

关键词

生成式人工智能，ChatGPT，数据获取，法律规制

The Regulatory Path of Generative AI Business Data Acquisition Behavior

—A Case Study of ChatGPT

Peixin Lin

School of Law and Economics, Wuhan University of Science and Technology, Wuhan Hubei

Received: Jun. 21st, 2024; accepted: Jul. 17th, 2024; published: Aug. 21st, 2024

文章引用: 林佩鑫. 生成式 AI 商业数据获取行为的规制路径[J]. 法学, 2024, 12(8): 5210-5222.
DOI: 10.12677/ojls.2024.128742

Abstract

Data is recognized as the cornerstone and core driving force for technological progress, while the generative AI technology represented by ChatGPT encounters notable compliance risks in the field of commercial data acquisition. Due to the lack of corresponding legal status for generative AI, it means that developers, designers, manufacturers, providers, and users all need to undertake the corresponding compliance risk responsibilities. To reasonably share this responsibility, the chain responsibility mechanism is proposed. Based on this, in view of the "*Interim Measures for the Management of Generative Artificial Intelligence Service*", it is necessary to set more detailed restrictions and requirements for data acquisition and punishment mechanisms. In addition, the comprehensive introduction of the "regulatory sandbox" system and the establishment of specialized agencies or committees are also of great significance. Through these measures, it can not only standardize the industry order and provide solid guarantees for the legal use of data, but also ensure its safe and compliant operation while promoting technological innovation. In short, in the face of the challenges brought about by generative AI technology, we must continuously explore and improve the relevant systems and mechanisms. Only in this way can we achieve the healthy development and efficient application of technology, and then fully release the potential value of data and promote the continuous progress of technology, creating a good situation of win-win. At the same time, it also helps to build a fair, ordered, and sustainable development of the technological ecosystem.

Keywords

Generative Artificial Intelligence, ChatGPT, Data Acquisition, Legal Regulation

1. 引言

2023 年 4 月 11 日，中国国家互联网信息办公室发布《生成式人工智能服务管理办法(征求意见稿)》，定义生成式人工智能为“基于算法、模型、规则生成文本、图片、声音、视频、代码等内容的技术”。在 2022 年 11 月，OpenAI 推出生成式人工智能产品“ChatGPT”。然而，2023 年 6 月，OpenAI 面临集体诉讼，指控其未经同意窃取约 3000 亿个单词及个人信息。ChatGPT 被指能绕过付费墙获取隐藏全文。OpenAI 在 8 月发布网络爬虫 GPTBot，声称将过滤付费内容及个人信息，以及违规文本来源。

数据是人工智能产业的“燃料”，生成式人工智能技术如 ChatGPT 为人带来便利和创新，同时引发法律风险。随着技术进步，企业竞争激烈，数据爬取案件增多。我国虽建立人工智能治理框架，但生成式人工智能数据爬取相关理论和实践仍有不足，数据保护法规需完善。保障数据安全和隐私，同时有效利用生成式人工智能技术成为迫切问题。本文旨在探讨生成式人工智能在数据获取过程中的合规风险，并以 ChatGPT 为例，分析其工作原理及数据爬取行为的合法性边界。在此基础上，提出规制生成式人工智能商业数据获取行为的建议，旨在提出有效的政策建议和监管措施。

2. 生成式人工智能数据获取的合规风险与合法性边界

生成式人工智能(Generative Artificial Intelligence，即 GAI，也称生成式 AI)目前主要包括生成式对抗

网络(GAN)和生成式预训练转化器(GPT)两种类型。在这两种类型中，本文将重点介绍生成式预训练转化器。其中 ChatGPT 是 OpenAI 训练的大型语言模型，是自然语言处理领域成果。ChatGPT 生成机制分四步：数据“喂养”，让模型学习人类语言并比对文本；自我监督学习，使模型符合预期；奖励模型训练，生成答案并按人类评价排优先级；强化学习优化，用算法提升学习效果和生成能力。其工作原理是学习大量数据和对话，构建深度学习模型，用于生成符合人类语言习惯的文本。

2023 年，匿名人士起诉 OpenAI 和微软，称其窃取互联网上的单词，包括未经同意的个人信息，引发对数据获取合规性的关注。随后，《纽约时报》以侵犯版权为由起诉 OpenAI 及其合作伙伴微软。同年，还发生多起指控 OpenAI 和微软非法收集、使用和分享个人信息以及侵犯著作权和隐私权的案件，引发社会和法律界的深入探讨和反思。

2.1. 合规风险

由于 ChatGPT 等生成式人工智能于初始阶段训练大模型的需要，会大量收集各种复杂的信息数据。根据 ChatGPT 的开发者 OpenAI 提供的信息，ChatGPT 主要的信息来源包括互联网上的公开信息、从第三方处通过许可获取的信息以及用户或模型训练者提供的信息(如用户与 ChatGPT 的聊天记录)。然而，来源不同的数据在获取时，可能存在以下合规风险：

第一，若 ChatGPT 获取的数据涉及个人信息，根据《个人信息保护法》第 13 条规定[1]，处理个人信息前需取得信息主体的同意。但当我们聚焦到 ChatGPT 这类大型语言模型的实际操作层面，特别是其数据处理与训练环节时，由于用户的数量庞大且复杂，要求开发者确保每一位用户在提供个人信息时都能充分了解并同意其被使用的细节，这个是很难做到的。所以，无法解决确保用户知情同意和开发者进行开发训练工作之间的矛盾，就可能会带来合规风险。

第二，ChatGPT 在训练过程中可能违反《个人信息保护法》中有关“最小数据范围”的原则[2]。

第三，若 ChatGPT 获取的数据来自已获专利的著作，可能违反《著作权法》中未经许可不得擅自使用的规定。

第四，如果 ChatGPT 获取的数据来自购买的第三方平台，可能无法确保数据的合法性。购买第三方平台数据意味着支付费用获得专属于该平台的数据，并签署相关约束协议以限定数据使用目的。通常情况下，通过第三方平台购买的数据会增加该平台的盈利性质，可能无法确保第三方平台数据获取的合法性，同时使用数据也会受到更多限制，不利于充分彰显数据的价值[1]。

第五，ChatGPT 使用爬虫技术自主搜集数据的行为可能带来的合规隐患。特别是当 ChatGPT 的训练数据来源涉及那些已公开禁止第三方爬取数据的网站时，情况就更加复杂了。这些数据往往被视为企业的核心资产，拥有明确的竞争性财产权益，任何擅自爬取的行为，都可能侵犯到这些权益，违反反不正当竞争法的相关规定，从而带来合规风险[2]。

2.2. 确定数据爬取的合法性边界

确定数据爬取的合法性边界是非常重要的，因为不当的数据抓取行为可能会侵犯各方的利益，违反反不正当竞争法等相关规定。为了确定数据爬取的合法性边界，需要权衡各方权益，并考虑数据来源、

[1]《个人信息保护法》第十三条：符合下列情形之一的，个人信息处理者方可处理个人信息：(一) 取得个人的同意；(二) 为订立、履行个人作为一方当事人的合同所必需，或者按照依法制定的劳动规章制度和依法签订的集体合同实施人力资源管理所必需；(三) 为履行法定职责或者法定义务所必需；(四) 为应对突发公共卫生事件，或者紧急情况下为保护自然人的生命健康和财产安全所必需；(五) 为公共利益实施新闻报道、舆论监督等行为，在合理的范围内处理个人信息；(六) 依照本法规定在合理的范围内处理个人自行公开或者其他已经合法公开的个人信息；(七) 法律、行政法规规定的其他情形。依照本法其他有关规定，处理个人信息应当取得个人同意，但是有前款第二项至第七项规定情形的，不需取得个人同意。

[2]《个人信息保护法》第六条：处理个人信息应当具有明确、合理的目的，并应当与处理目的直接相关，采取对个人权益影响最小的方式。收集个人信息，应当限于实现处理目的的最小范围，不得过度收集个人信息。

类型、使用目的和方式等因素。

一是数据来源的合法性。数据来源的合法性是至关重要的。根据数据的属性和来源，可以分为个人数据、公共数据和企业商业秘密等。对于原始数据和衍生数据的处理方式也会有所不同。衍生数据通常经过深度分析、整合和匿名化处理，具有独立的财产权益。公开数据和非公开数据的获取方式和合理性也需要进行区分，非公开数据的获取需要尊重数据持有者设置的访问权限。

二是数据获取方式。如果 ChatGPT 在收集和使用其他企业数据信息时采取不合理手段，可能导致用户公司机密信息泄露，进而损害公司核心竞争力[1]。例如，2023 年，三星 DS 部门从三月开始允许员工使用 ChatGPT，但是短时间内出现了三起机密信息泄露事件[3]。可见，获取方式的合法性是确保数据爬取合法性的重要方面，不当的数据获取行为可能触犯法律，损害他人利益，违反反不正当竞争法规定。以下是关于数据获取方式的合法性边界的一些重要考虑因素：

1) 破坏技术措施：非法破坏技术措施是不可取的行为。在数据获取过程中，不应采取破坏数据持有者设置的技术安全措施，如身份认证系统、加密系统或反爬虫技术等[3]。

2) 违反 Robots 协议：爬虫技术是生成式人工智能开发者获取数据的主要手段，但应当遵守 Robots 协议[4]，即通过 Robots.txt 文件规定可以获取的内容。但值得一提的是，Robots 协议并不具有强制力和强制执行力。违反 Robots 协议获取数据可能被认定为不当竞争行为，裁判需审慎评估 Robots 协议的合法性。

3) 违反约定范围获取数据：数据获取过程中不得违反事先约定的范围，超越其约定的数据范围。违反约定超范围获取数据可能被视为不正当竞争行为，如在微博诉脉脉案[5]中所示，在该案中，脉脉公司利用爬虫技术获取了微博用户的信息，包括用户数据、内容等，但超出了事先约定的范围。

4) 过度抓取数据：在数据获取的过程中，尽管某些手段可能在法律框架内被认为是合规的，但倘若这种抓取行为在数量上过于庞大，或持续时间过长，以致对他人经营的网站造成了过度的负担，这便可能演化为一种不正当竞争行为。对于此类行为的审视，关键在于其导致的实际后果，即是否对竞争对手的经营造成了实质性的不良影响，而不仅仅是拘泥于抓取手段是否违反了法律条款。因此，我们在进行数据抓取时，必须谨慎权衡，确保自身行为不会对他人的正常经营造成不必要的干扰，维护健康的竞争秩序。

三是数据获取之后的用途。数据获取之后不能用于非法目的，比如非法交易、出售，或者用于其他超越最初授权的商用。那么后果就是导致质性替代之损害，或引发部分性替代之风险，乃至干扰他人产品之正常运作，或涉及非法利用他人经营之成果，进而破坏市场竞争之公正秩序。因此获取目的要遵守法律中的诚信原则、善良原则，如前文所述的要以“商业道德”“正当利益”为兜底标准，不能超越其限度去适用其合法获取的数据，也就是合法获取的数据不能用于非法目的。

总之，生成式人工智能在数据获取过程中面临着诸多合规风险，需要明确合法性边界，以确保其数据获取行为合法、合规。在数据来源方面，要确保数据的合法性和合规性，尊重公开数据和非公开数据的获取方式和合理性；在数据获取方式方面，要遵守相关法律法规和道德规范，不得破坏技术措施、违

[3]三星原本全面禁止使用 ChatGPT，直至 2022 年 3 月 11 日才允许 DS 部门员工使用，但不足 20 天就接连传出机密外泄事件。主要源于员工误用或滥用 ChatGPT，包括两宗与半导体设备有关，另一宗则关于会议内容。其中一宗事件是三星半导体事业暨装置解决方案(Device Solutions)部门的员工，在操作半导体测试设备下载软件过程中，将有问题的原始码复制贴上到 ChatGPT，以找出问题与解决方式，但此举可能令 ChatGPT 把三星的机密信息，当作训练资料使用。另一宗事件同样在 DS 部门发生，亦是寻求 ChatGPT 以改善代码，但有关代码则涉及芯片良率。第三宗事件则是用 ChatGPT 撰写会议内容，虽然不算是技术机密，亦可能导致会议内容外泄。

[4]Robots 协议也称爬虫协议、爬虫规则等，是指网站可建立一个 robots.txt 文件来告诉搜索引擎哪些页面可以抓取，哪些页面不能抓取，而搜索引擎则通过读取 robots.txt 文件来识别这个页面是否允许被抓取。但是，这个 robots 协议不是防火墙，也没有强制执行力，搜索引擎完全可以忽视 robots.txt 文件去抓取网页的快照。

[5]“脉脉”曾与新浪微博合作，提供用户通过微博账号登录“脉脉”的服务。而引起争议的关键在于，一旦通过微博登录，“脉脉”会抓取用户的微博账户职业、教育信息，并与注册账号时用户上传的手机通讯录内容建立关联，最终一并对外展示。在终止了与“脉脉”的合作后，新浪微博诉至法院，认为“脉脉”的非法抓取信息等行为构成不正当竞争。

反 Robots 协议、超越约定范围获取数据或过度抓取数据；在数据获取之后的用途方面，要遵守法律中的诚信原则、善良原则，不得将数据用于非法目的。只有这样，才能促进生成式人工智能的健康、可持续发展，为社会带来更多的福祉。

3. 既有规制现状

3.1. 我国既有规制现状

在现有法律制度框架下，商业数据爬取行为常常依赖于反不正当竞争法的保护机制。由于缺乏明确的产权界定，反不正当竞争法成为对未经授权数据盗用行为进行经济赔偿的救济途径，主要包括商业秘密和一般条款两种保护途径，主要通过事后禁止可识别的不当竞争行为来维护市场秩序。然而，由于缺乏数据作为财产权的规范依据和商业秘密保护门槛较高，针对生成式人工智能数据获取行为的法律法规仍显不足。

2021 年 8 月 18 日，最高人民法院发布的《关于适用〈中华人民共和国反不正当竞争法〉若干问题的解释(征求意见稿)》中尝试对数据保护规则进行规定[4]，但在 2022 年 3 月 17 日正式公布的《最高人民法院关于适用〈中华人民共和国反不正当竞争法〉若干问题的解释》中删除了相关条款，显示数据保护规则的构建仍存在争议。

3.1.1. 我国对人工智能的监管和治理规范已初步形成框架

在探讨我国对人工智能的监管与治理规范之时，不得不提及三个主要方面的具体体现。首先，从国家层级的视角来看，国务院、国家新一代人工智能治理专业委员会等部门，陆续颁布实施了诸如《新一代人工智能治理原则——发展负责任的人工智能》等重要政策文件。这些文件的出台，标志着我国在人工智能领域的监管与治理规范已初步构建起了坚实的框架。其次，我国目前在人工智能治理上，已初步构建了涵盖法律、部门规章、地方性法规、国家标准及行业自律标准的多元化治理规范结构。这一结构呈现出了从中央政府至地方政府以及行业组织间的分级别、多层次特点，从而形成了一个既包含具有强制约束力的制定法，又包含新型行业自我规制的自规范准则在内的综合治理体系。最后，我国当前对于人工智能的治理规范，重点在于确保人工智能的安全性、使用的透明性、算法的可解释性及其符合伦理原则等关键方面[2]。这一治理规范的实施，无疑为人工智能技术的健康有序发展提供了坚实的法治保障和道德支撑。

3.1.2. 《办法》提供了更加明确的监管指引

在我国新兴领域的立法探索中，国家网信办携手工信部、公安部等部门，共同颁布了《生成式人工智能服务管理暂行办法》，此《办法》不仅彰显了我国在新技术应用规制策略方面的最新成就，更为生成式人工智能服务的监管提供了具体、明确的指导。此法令之制定，详细到了合规义务、安全评估、数据合法性、用户信息保护及投诉处理等细微层面，为生成式人工智能的监管绘制了详尽的蓝图[2]。作为第三部专项技术服务管理规章，该《办法》的出台，无疑为完善生成式人工智能的规制体系注入了新的活力。它着重强调了生成内容必须体现社会主义核心价值观，对服务提供者的责任承担进行了进一步的强化，彰显了我国在引导人工智能发展上的严谨态度和明确立场。具体而言，《办法》中的第 4 条[6]，对

[6]《生成式人工智能服务管理暂行办法》第四条：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；(二) 在算法设计、训练数据选择、模型生成和优化、提供服务等过程中，采取有效措施防止产生民族、信仰、国别、地域、性别、年龄、职业、健康等歧视；(三) 尊重知识产权、商业道德，保守商业秘密，不得利用算法、数据、平台等优势，实施垄断和不正当竞争行为；(四) 尊重他人合法权益，不得危害他人身心健康，不得侵害他人肖像权、名誉权、荣誉权、隐私权和个人信息权益；(五) 基于服务类型特点，采取有效措施，提升生成式人工智能服务的透明度，提高生成内容的准确性和可靠性。

生成内容及其提供服务制定了明确的合规标准。而第16条[7]则进一步要求服务提供者对其生成的内容进行明确标识，将其与相关的主体紧密绑定，从而强化了特定主体对生成式人工智能内容的责任担当。此外，《办法》的第6条[8]对安全评估与算法备案提出了具体的合规要求。依据这一规定，那些具备舆论属性或具备社会动员能力的生成式人工智能服务，如同论坛、公众号等互联网信息服务一样，将纳入严密的监管流程，以此确保服务的安全性及规范性，从源头上避免潜在的风险与隐患。这一系列的举措，不仅体现了我国在人工智能领域治理的细致入微，更为行业的健康发展提供了坚实的法治保障。

最重要的是，《办法》第7条[5]对生成式人工智能训练数据的合规要求十分重要，要求生成式人工智能服务提供者处理数据时需遵守以下规定：

第一，提供者必须使用的是具有合法来源的数据和基础模型。

第二，凡是涉及知识产权的数据，提供者使用数据时不得侵害他人依法享有的知识产权。

第三，凡是涉及个人信息的，提供者应当要取得个人的同意或者符合法律、行政法规规定的其他情形。

第四，提供者应采取有效措施提高训练数据质量，以增强训练的真实性、准确性、客观性、多样性。

第五，提供者训练数据时还要遵守《中华人民共和国网络安全法》《中华人民共和国数据安全法》、《中华人民共和国个人信息保护法》等法律、行政法规的其他有关规定和有关主管部门的相关监管要求。

综上所述，《办法》第7条是要求提供者应对数据来源合法性负责，确保所使用的数据符合我国法律法规的要求，不能侵犯他人合法的知识产权，还要保证提供者所提供的数据真实、准确、客观还有多样。

3.1.3. 治理规则的适用困境

在我国当前人工智能的治理规则中，相关法律对于人工智能企业获取个人数据的行为分别规定了知情同意原则、目的限制原则以及诚实信用原则[2]，但在实践中，却存在困境，具体表现如下：

第一，知情同意原则与应用效率之间存在矛盾。知情同意原则设立的法理依据是保护个人的知情同意权，个人有权对自己的数据信息进行处分。然而，在实际操作中，要求企业履行事无巨细的完全告知几乎不可能实现，且可能增加企业经营成本和用户决策成本。此外，很多数据收集场景具有非接触性特征，使得知情同意原则的适用面临困境，可能阻碍行业创新，也不符合当前鼓励人工智能发展的政策导向[2]。

第二，目的限制原则适用中所要求的目的不具有现实可预期性。根据《个人信息保护法》第6条的规定，目的限制原则要求企业在处理公众信息之前，需要明确其处理该信息的目的[6]。然而，企业预设的目的可能过于抽象模糊，缺乏统一标准，这些信息的使用很容易就超出信息主体初始的授权范围，这就导致在使用个人信息时难以确保目的限制原则的有效适用。

第三，诚实信用原则易被滥用。特别在人工智能企业的数据获取环节中，这一原则常因表述过于宽泛和抽象而难以准确界定其适用范围，导致其在实际操作中缺乏具体的指导价值。由于缺乏详尽明确的指导标准和对诚信原则的细致分类分析，如果我们在实践中对诚信原则的适用控制过于宽松，就可能导致大量潜在的纠纷倾向于逃向一般性条款，从而进一步加剧诚信原则被滥用的风险。

[7]《生成式人工智能服务管理暂行办法》第十六条：网信、发展改革、教育、科技、工业和信息化、公安、广播电视、新闻出版等部门，依据各自职责依法加强对生成式人工智能服务的管理。国家有关主管部门针对生成式人工智能技术特点及其在有关行业和领域的服务应用，完善与创新发展相适应的科学监管方式，制定相应的分类分级监管规则或者指引。

[8]《生成式人工智能服务管理暂行办法》第六条：鼓励生成式人工智能算法、框架、芯片及配套软件平台等基础技术的自主创新，平等互利开展国际交流与合作，参与生成式人工智能相关国际规则制定。推动生成式人工智能基础设施和公共训练数据资源平台建设。促进算力资源协同共享，提升算力资源利用效能。推动公共数据分类分级有序开放，扩展高质量的公共训练数据资源。鼓励采用安全可信的芯片、软件、工具、算力和数据资源。

除此之外，《办法》更多关注对个人信息的获取进行规范，而在商业数据获取方面的监管要求和禁止性规定相对不足。在商业数据获取方面，尤其是生成式人工智能服务中的商业数据获取方面，当前《办法》未能提供足够详细和具体的监管要求和禁止性规定。这导致了在商业数据获取和使用方面存在一定的法律漏洞和监管空白。具体而言，商业数据的获取可能涉及到商业机密、市场竞争等敏感领域，需要更加严格的监管和规范。然而，《办法》在这方面的规定相对模糊，未能明确商业数据获取的具体要求和禁止性行为，缺乏对商业数据获取过程中可能涉及的风险和问题的全面考虑。因此，当前《办法》在商业数据获取方面存在一定的不足之处，需要更加细化和完善相关规定，以确保商业数据的合法获取和使用，保护商业机密和市场秩序的稳定。

另外，《办法》设定的禁止性规范缺乏足够强制实施的效力[7]，由于其效力层级较低，但对违反其禁止性行为的处罚较轻，只能设定警告、通报批评和一定的罚款，可能会因为强制力不足导致实践中无法约束企业，企业仍然违反《办法》中的禁止性行为。

3.2. 欧盟和美国对生成式人工智能的监管措施

不同国家采用了不同的人工智能治理方法，且每种方法都反映出各国不同的法律体系、文化和传统。

3.2.1. 欧盟：安全优先，兼顾公平

欧盟在生成式人工智能监管方面一直秉持着安全优先和兼顾公平的原则。2024 年初生效的欧盟《人工智能法案》是全球首部人工智能领域的全面监管法规，被誉为人工智能治理史上的里程碑。该法案建立了人工智能开发和使用的道德和法律框架，并通过《人工智能责任指令》确保了法规的落地执行。《人工智能法案》对所有通用人工智能模型提出了透明度要求，对更强大的模型如 ChatGPT 则提出了更严格的规定，为生成式人工智能服务的发展铺平了道路。

在法案生效之前，欧盟部分成员国已经对 ChatGPT 采取相应执法活动。意大利、法国、西班牙等国家的数据保护机构对 OpenAI 公司和 ChatGPT 的数据处理行为展开了调查，并要求欧盟隐私监管机构评估 ChatGPT 的隐私保护水平。这些举措显示了欧盟在保护数据隐私和监管人工智能技术方面的积极态度和决心，为全球人工智能监管树立了榜样，为促进人工智能可持续发展做出了重要贡献。总体而言，欧盟在数据保护和人工智能技术监管方面一直走在立法实践的前沿，也率先颁布了全球首部人工智能领域的全面监管法规《人工智能法案》。

3.2.2. 美国：相对开放以促进产业创新

针对 ChatGPT 所产生的深远影响，并出于维护其在全球人工智能领域之显赫地位的考量，美国在人工智能治理层面，采取了颇为开明的监管策略，其意在催生产业革新之动力，并捍卫其在该领域的国际领导地位美国的监管路径相对开放，着重于企业自我规制和政府监管相结合。在生成式人工智能监管方面，美国通过一系列法规和行政命令展开监管措施。

其中，2023 年 4 月 11 日，美国国家电信和信息管理局发布了《人工智能问责制政策征求意见稿》，征求公众意见和反馈以制定人工智能系统的潜在问责措施[8]。此外，2023 年 10 月 30 日，美国总统拜登签署了一项具有里程碑意义的行政命令《关于安全、可靠、可信开发和使用人工智能的行政命令》，推出了白宫有关生成式人工智能的首套监管规定，为行业发展和监管提供了重要指导。

在过去的几年中，美国也通过一系列法案和行政命令加强了人工智能监管。例如，2019 年推出了《算法问责法案》和《美国人工智能倡议》，以及 2020 年发布的《人工智能应用监管指南备忘录》等。这些举措旨在确保人工智能技术的安全、可靠、可信赖，并推动产业创新和国家经济繁荣。

总体而言，美国在生成式人工智能监管方面注重审慎监管，相对开放以促进产业创新和维持领先地

位。通过法律法规和行政命令的制定，美国建立了一系列程序化问责路径，确保人工智能系统在公共治理场景中的应用符合规范和道德标准。

3.2.3. 域外监管经验的不足与借鉴

尽管欧盟和美国在生成式人工智能监管方面都采取了一定的措施，但仍存在着不足之处：

第一，欧盟在生成式人工智能监管中过于强调安全和公平，借鉴欧盟几个成员国对 ChatGPT 严格的执法措施。参考《人工智能法案》对通用人工智能模型和强大模型如 ChatGPT 提出的透明度和隐私保护要求，该法案要求超大型数字平台和搜索引擎，有责任监控其算法系统，将生成式人工智能的责任控制主体定位于技术供应者。然而，这种确定责任分配的模式可能带来过度监管的风险，限制了企业在人工智能领域的活动和投资。有研究显示，如果技术供应商被要求承担较重的责任，监管成本可能大幅增加，最终大幅减弱市场对于技术更新与投资的意愿[9]。此外，过度严格的监管可能导致创新的放缓，使得欧盟在人工智能领域的竞争力受到影响。

第二，美国的审慎监管策略虽然旨在促进生成式人工智能产业创新，但可能导致监管过于宽松，例如《关于安全、可靠、可信开发和使用人工智能的行政命令》并不具有强制性，缺乏对生成式人工智能技术潜在风险的充分考虑，相对开放的监管路径可能导致监管标准和指导不够明确和具体。美国的监管体系相对分散，监管机构众多，可能导致监管的协调性和一致性不足。不同监管机构之间的监管职责和标准可能存在交叉和冲突，影响了对生成式人工智能技术的全面监管。为了确保生成式人工智能技术的安全、可靠和可持续发展，美国监管机构需要加强对潜在风险的评估和监管，制定明确的监管标准和指导，加强监管的协调性和一致性，以实现监管的有效性和合规性。

第三，欧盟和美国在人工智能监管方面的立法和行政措施虽然有所进展，但在国际合作和标准制定方面仍存在不足。由于人工智能技术跨国界应用广泛，缺乏全球性的监管标准和合作机制可能导致监管漏洞和难以解决的问题。

欧盟和美国的监管经验对我国的有较强的借鉴意义，具体如下所述：

第一，《人工智能法案》作为一项全面的法律规范，其治理机制和惩罚措施体现了对人工智能技术发展的重视和监管的严谨性。在建立由独立专家组成的科学小组方面，这一机制有助于确保对人工智能系统潜在风险的科学评估和指导，为决策者提供专业意见和建议，使监管更加科学化和有效性。借鉴这一做法，我国也可以成立类似的专家团队，由行业专家、学者和政府代表组成，共同参与人工智能技术的监管和评估工作，保障法规制定和执行的专业性和权威性。

第二，《法案》中对违规行为实施高额罚款的做法，可以有效遏制企业的违法行为，提高企业对法规的遵守度。将罚款金额与公司规模和违法行为的严重程度挂钩，能够更好地体现惩罚的公平性和效果性，从而推动企业自觉遵守法律法规，促进产业健康有序发展。我国可以借鉴这一惩罚机制，《办法》中的处罚力度都较轻，可以加强对这一领域的违法行为打击力度，维护市场秩序和公共利益，保障人工智能技术的可持续发展。

第三，《法案》允许公民对人工智能系统提出投诉并获得解释的做法，有助于增强公众对技术的信任感和透明度。公众对人工智能系统如何作出影响自己的决定有清晰的了解，可以减少对技术的恐惧和疑虑，促进其技术的广泛应用和接受。我国也可以借鉴这一机制，建立公民参与人工智能治理的渠道和机制，促进技术发展与社会需求的平衡。

第四，《法案》将人工智能系统的风险分为四个等级，并为每个等级都分别制定了相对应的法律义务和规范，四个等级分别为轻微风险、有限风险、高风险和不可接受风险。而生成式人工智能因为有着强大的数据分析和生成能力，则是高风险级别，需要遵守最严格的安全规范。目前，我国在相关法规中，

如《办法》第 3 条，已初步设定了类似“分类分级监管”的原则。但相较于欧盟的《人工智能法案》，我们的规定还略显笼统，需进一步细化和完善。我们需要借鉴国际上的先进经验，特别是欧盟在这方面的立法举措，来进一步完善我国的政策法规或相关国家标准。通过详细规定不同风险级别的具体监管要求，我们可以更精准地管理生成式人工智能等技术的发展。

在深究我国人工智能治理之道的过程中，不容忽视的是欧盟与美国在此领域的制度探索所展现的深刻洞见与启示。诚然，这些制度构想为我国在人工智能领域的良性治理指明了方向，然而，其实际应用的成效与益处，尚需通过实践的检验方能确证。因此，建立起一个契合中国实际情况的监管架构至关重要，同时需积极探索具有中国特色的治理策略。

4. 现存问题

生成式 AI 在数据获取方面存在诸多问题，这些问题不仅制约了其自身的发展，也给社会带来了潜在的风险，需要我们认真审视和解决。

第一，法律法规不完善。目前针对生成式人工智能商业数据获取行为的专门法律法规尚未建立，缺乏明确法律依据和指导。根据上文所提及的，现有法规如《生成式人工智能服务管理暂行办法》在商业数据获取方面监管要求和禁止性规定相对不足，且相关法律规定的知情同意原则、目的限制原则以及诚实信用原则在实践中存在困境。现有的相关法规在面对技术快速发展和新型风险时，存在局限性和不一致性，难以统一治理目标、机制和尺度，给监管部门执法带来挑战，影响监管效果和公正性，给企业合规审核带来成本压力。

第二，监管机制不健全。缺乏专门机构对生成式人工智能的商业数据获取行为进行实时监控和审查，合规性难以确保，无法有效监督该新兴技术发展；监管部门职责和权限不够明确，可能出现推诿、扯皮等现象，影响监管效率和效果；监管手段不足，现有监管手段难以跟上技术发展步伐，对一些隐蔽性违规行为难以及时发现和查处。

第三，行业自律不足。首先是自律组织不健全，生成式人工智能行业自律组织尚未充分发挥作用，行业标准和规范不够完善，企业在开发和使用生成式人工智能时缺乏明确的行为准则和最佳实践指导，容易出现违规行为。其次是标准和规范不完善，行业标准和规范不完善，企业在数据获取和使用过程中随意性较大，增加了违规风险。

第四，技术保障有待加强。数据安全和隐私保护技术的研发和应用还需进一步加强，以提高生成式人工智能在数据处理和存储过程中的安全性和隐私性。目前，“监管沙盒”制度尚未全面引入，不利于监管部门对技术发展的理解和监管。

第五，企业管理薄弱。企业的数据合规机制不够完善，在数据信息管理方面存在漏洞，容易导致员工行为不当带来潜在风险，合作伙伴的数据安全也难以保障。此外，企业在事前数据风险防范和事后数据违规风险管理方面，缺乏有效的机制和措施。

第六，公众意识淡薄。公众对数据隐私和安全的重视和认知不足，对生成式人工智能商业数据获取行为的监督意识不强，缺乏相关的教育活动和知识普及，难以有效保护自己的个人信息。

这些问题的存在，制约了生成式 AI 的健康发展和合法合规应用，需要我们采取有效措施加以解决。

5. 规制路径

对于生成式人工智能导致的数据获取的法律风险问题进行规制应当遵循总体上坚持包容审慎的治理态度，对生成式人工智能分级分类进行监管，以鼓励创新和规范发展为原则，坚持硬法与软法相结合，构建政府、企业、社会多方协同的治理模式。

5.1. 硬法方面

5.1.1. 明确法律法规

制定针对生成式人工智能商业数据获取行为的专门法律法规，明确数据的获取、使用和处理规则。应从源头上对生成式人工智能可能引发的法律风险进行有效的治理，也就是应该在现有的监管基础之上，针对生成式人工智能构建一套综合的立法体系，通过系统性的规范来约束人工智能技术及其在各个应用场景下的行为，进而增强法规的强制性和执行力度。虽然《办法》的出台无疑会极大地优化生成式人工智能无序发展的现状，然而，随着技术的日新月异以及生成式人工智能系统的不断创新，其所面临的风险可能会变得更为复杂多变。在这种情况下，现有的《办法》可能会因其固有的局限性而无法直接对新型风险进行有效的规制，或者由于其与上位法之间的不一致而难以发挥统一的法律效力。这一现状无疑会使得生成式人工智能的治理目标、治理机制以及治理尺度的统一变得异常艰难，这不仅会影响到治理的效率，更会给企业在合规审核工作中带来巨大的挑战和成本压力[1]。因此，我们必须对现有的法规进行持续的完善与更新，以适应人工智能技术的快速发展，确保治理工作的有效性和针对性。

立法机构要尽快制定专门针对人工智能治理的专门性法律，例如《生成式人工智能法》，明确生成式人工智能的定义、范围、应用领域、责任主体等内容，并为相关的风险问责构建一套完整的责任机制，包括明确人工智能的风险问责主体、被问责客体、问责程序，明确各个主体的权利义务[2]，确立起研究开发者、设计制造者、服务提供者和服务使用者四方问责机制以及具体的风险评估事项等。具体建议如下所述：

第一，对商业数据的获取进行更加细化的限制。可以参考欧洲的《通用数据保护条例》(GDPR)等法规，规定生成式人工智能开发者和使用者在处理个人数据时应遵守的原则和规定，保障用户数据的安全和隐私。一是规定生成式人工智能开发者和使用者应当遵守数据最小化原则，仅在必要的情况下收集和使用个人数据。二是规定生成式人工智能开发者和使用者应当获得用户明示同意后，方可收集、存储和处理用户数据。三是规定生成式人工智能开发者和使用者应当采取必要的技术和组织措施，包括加强数据加密、建立数据备份机制、实施数据安全审计等，确保商业数据的安全性和完整性，同时进行数据合规性审查，确保数据获取和处理行为符合相关法律法规的要求，避免违反数据保护和隐私规定。四是可以要求生成式人工智能系统在商业数据获取和处理过程中提供透明度，包括告知数据使用目的、数据共享情况、数据处理方式等信息，让用户了解其数据被如何使用，从而增加数据使用的可控性和透明度。五是可以赋予数据主体更多的权利和控制权，包括权利知情、权利访问、权利更正、权利删除等，让数据主体能够更好地控制自己的数据，并对商业数据获取和使用行为提供监督和限制。

第二，明确责任主体和归责原则。明确责任主体和归责原则是对于生成式人工智能数据获取侵权责任的重要问题，尽管生成式人工智能具有相当强的认知能力，但它并没有相应的法律地位，因为人工智能是由人类创造以服务人类的智慧型工具，其工具属性决定了其法律地位的限度性。在我国采取“法人实体说”的情况下，生成式人工智能无法作为拟制法人承担责任，穿透技术背后的核心，体现的依然是设计者、制造者、提供者的意识和逻辑思维，因此最终责任仍应由人类主体承担，包括研究开发者、设计制造者、服务提供者和服务使用者。

生成式人工智能数据获取侵权责任主体涵盖研究开发者、设计制造者、服务提供者、服务使用者等。法律法规可以明确规定在不同情况下各个主体的责任承担程度和标准。对于由人工智能自身运算混乱或设计逻辑引起的责任后果，研发者与提供者可能需要承担更大的责任；对于由操作者操作引起的责任后果，操作者可能需要承担主要责任。针对数据获取侵权问题，法律法规可以借鉴欧盟《人工智能法案》中的“谁提供谁负责”原则，强调数据提供者对于数据的合法性和合规性负有责任。同时，应考虑到服

务使用者可能提供存在侵权数据的情况，因此可以引入“谁提供数据谁负责”的原则，对数据提供者和服务使用者进行责任分配。对于责任后果无法确定原因的情况，生成式人工智能的研究开发者、设计制造者、服务提供者和服务使用者及其自身应分别承担相应比例的责任，需要建立精准的责任分配机制来确保各主体承担相应责任。

因此，法律法规应明确要求生成式人工智能在工作时必须生成日志，并保留一定期限，以便分析、监控和审计。日志记录可以为责任的分配提供依据和证据，有助于确定责任主体和责任范围。通过明确责任主体和归责原则，可以建立健全的法律框架和机制，促进生成式人工智能系统的安全、可靠和合规运行。

第三，强化违法违规惩罚。由于《办法》中对违反其禁止性行为的处罚较轻，法律法规可以规定对违法违规行为进行严厉的处罚。除了罚款外，法律法规还可以规定相应的行政处罚措施，如责令停止违法行为、吊销相关许可证或资质等。例如，生成式人工智能系统获取数据后未按规定使用或未经授权将数据用于其他目的，违反了数据使用协议或隐私政策，法律可规定对数据滥用行为者进行行政处罚，包括罚款或吊销相关许可证。行政处罚可以对违规主体进行更直接的制裁，确保其遵守法律法规。此外，对于严重违法违规行为，法律法规可以规定相应的刑事责任，如刑事处罚、刑事拘留等。例如企业通过生成式人工智能窃取竞争对手的销售数据或客户名单，法律可规定对其负责人进行刑事追究，可能面临刑事处罚。强化违法违规惩罚有助于提高责任主体的合规意识，遏制违法行为，保障数据主体的权益和数据安全。

5.1.2. 加强监管机制

尽快设立专门的监管机构、部门或者委员会，执行未来会出台的人工智能法案，对生成式人工智能的商业数据获取行为进行实时监控和审查，确保合规性，更好地监督生成式人工智能这一项新兴技术。

第一，规定生成式人工智能监管机构的职责和权限，明确其监督和管理生成式人工智能的具体工作内容。

第二，规定生成式人工智能监管机构应当定期对生成式人工智能的开发和使用进行审查和评估，确保其合法合规运行。

第三，规定生成式人工智能监管机构应当与相关部门和行业组织合作，加强对生成式人工智能的监管和指导。

第四，监管机构可以制定具体的实施细则，对生成式人工智能的使用进行监督，确保其符合法律规定和社会伦理。

5.2. 软法方面

5.2.1. 推动行业自律

第一，建立生成式人工智能行业自律组织，监督行业成员遵守规范，促进行业的健康发展。

第二，制定生成式人工智能行业标准和规范，引导企业和机构遵守规范，提高行业的整体水平，明确生成式人工智能开发和使用的行为准则和最佳实践。

第三，发布生成式人工智能行业指导意见，指导企业和机构如何合规开发和使用生成式人工智能。

第四，规定生成式人工智能行业组织应当建立投诉处理机制，接受公众投诉和举报，及时处理违规行为。第五规定生成式人工智能行业组织应当定期发布行业发展报告，推动行业的健康发展和自律规范。

综上，鼓励生成式人工智能行业建立自律机制，积极推动生成式人工智能行业的自律与自治，制定行业标准和规范，形成行业伦理指南、自律公约等行业规范，规范生成式人工智能技术开发与应用，推

动数据获取和使用的合规发展[7]。

5.2.2. 强化技术保障

加强数据安全和隐私保护技术的研发和应用，提高生成式人工智能在数据处理和存储过程中的安全性和隐私性。引入“监管沙盒”制度，这是英国金融行为监管局 2015 年的新思路，而人工智能监管沙盒已在北京开始试点，日后可能在全国范围内引入“监管沙盒”制度来允许人工智能企业在相对可控的环境内进行试验性的开发、测试和验证，也会更加有助于监管部门对其技术发展的理解。

5.2.3. 增强企业管理

企业构建一个完善的数据合规机制显得尤为关键。无论是引领行业技术的互联网巨擘如 OpenAI，还是运用 ChatGPT 等工具进行数据采集的第三方服务公司，都应当高度重视并加强自身的数据信息管理[1]。这一机制不仅有助于企业规范自身数据信息的处理流程，确保数据的合法性、安全性和保密性，更能在一定程度上防范员工行为不当带来的潜在风险，避免合作伙伴的数据安全受到损害。同时，这样的合规机制还能够提升企业在市场竞争中的信誉度和核心竞争力。

对于事前数据风险，企业应建立合规日常管理机制，不断完善企业数据合规治理体系。具体措施包括：一是制定企业数据合规文件，为数据合规公司治理提供规范依据；二是全面监督数据合规事务，建立、实施、记录和维护与人工智能系统相关的风险管理系统，及时发现潜在风险并提出建议；三是建立事件应对预案，及时有效处理各类事件；四是定期组织数据合规培训，提升数据合规意识与治理能力；五是企业还应积极寻求外部合作，推动企业数据合规体系建设。此外，企业还应对所提供的数据进行进一步审查和约束，要求对数据来源、真实性以及交易使用流程进行全面检测。特别是在涉及个人敏感信息或用户隐私信息时，企业应事先通知相关当事主体并取得其同意。

关于事后数据违规风险的管理与应对，实有必要建立一套周密的机制以加强内部监管，从而确保违规行为的杜绝，并防范其再次发生。具体措施包括：一是对数据违规风险进行分类划分，分辨出其所涉的民事法律风险、行政法律风险及刑事法律风险，以便于精准应对；二是需要运用多样化的解决机制，及时管控风险，以迅速应对数据安全风险所可能导致的各种损失；三是针对刑事风险，采取认罪认罚、整改等措施，降低刑事责任。

5.2.4. 提高公众意识

加强公众对数据隐私和安全的重视和认知，开展数据隐私和安全的教育活动，向公众介绍数据隐私的重要性以及如何保护自己的个人信息，提高用户对生成式人工智能商业数据获取行为的监督意识。

总的来说，结合中国国情探讨生成式人工智能的法律法规规制路径需要考虑中国的文化、法律体系和社会实践，通过立法、政策制定、监管机制建立等手段，确保生成式人工智能的发展与应用符合中国的法律规定和社会需求。

6. 结语

ChatGPT 等生成式人工智能的出现对整个社会都有极大变化，在给人们提高效率的同时，也带来了许多风险，生成式人工智能应当以人为本，确保技术的安全性和合法性，为人类社会带来福祉。当前已经出台的《生成式人工智能服务管理暂行办法》为生成式人工智能数据获取提供了一定的指导，规定了生成式人工智能在获取数据上要保障数据合法合规，不能侵犯他人合法的知识产权和个人隐私。但有关生成式人工智能的数据保护专门性法律仍旧缺位，生成式人工智能商业数据获取行为的规制是一个复杂而紧迫的问题，通过明确法律法规、加强监管力度、推动行业自律、强化技术保障和提高公众意识等多方面的措施，可以有效地规范生成式人工智能的商业数据获取行为，保护数据隐私和安全，促进公平竞

争和市场健康发展。

基金项目

2023 年武汉科技大学省级大学生创新创业训练计划项目《生成式 AI 商业数据获取行为的规制路径——以 ChatGPT 为例》(项目编号：S202310488089)。

参考文献

[1] 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, 41(5): 87-97.

[2] 毕文轩. 生成式人工智能的风险规制困境及其化解: 以 ChatGPT 的规制为视角[J]. 比较法研究, 2023(3): 155-172.

[3] 苏志甫. 数据要素时代商业数据保护的路径选择及规则构建[J]. 信息通信技术与政策, 2022(6): 14-26.

[4] 赵丹, 沈澄. 数据抓取不正当竞争纠纷的司法审查要素考察与反思[J]. 科技与法律(中英文), 2023(2): 52-59.

[5] 生成式人工智能服务管理暂行办法[J]. 中华人民共和国公安部公报, 2023, (5): 2-5.

[6] 王大志, 张挺. 风险、困境与对策: 生成式人工智能带来的个人信息安全挑战与法律规制[J]. 昆明理工大学学报(社会科学版), 2023, 23(5): 8-17.

[7] 徐继敏. 生成式人工智能治理原则与法律策略[J]. 理论与改革, 2023(5): 72-83.

[8] 梁正, 何嘉钰. 确保网络数据安全 应对新一代人工智能治理挑战[J]. 中国信息安全, 2023(6): 22-24.

[9] 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.