企业数据爬取的法律风险与规制路径研究——基于权益均衡的视角

龚鹏程，张阳阳
河海大学法学院，江苏 南京 210003

摘要
随着对数据需求量越来越大，为了提高效率，企业往往采取数据爬取的方式来获取数据。数据爬取引发数据隐私安全、数据污染危及数据治理安全以及不正当竞争破坏市场秩序等多重风险。现有的法律制度存在着爬取行为法律界限模糊、Robots 协议的法律效力未定、数据合规治理机制存在缺陷的困境，其深层根源在于“数据效率与数据安全”、“数据平台自主经营权与数据接入容忍义务”两对核心权益的失衡，因此，基于权益均衡的视角，提出需要建构多方共治的数据分类分级协同治理体系、完善数据合规治理机制、优化三重授权原则的分类应用、明确 Robots 协议作为合同法律效力和保障数据的删除权来破解数据爬取面临的困境，促进企业的数字化转型。

关键词
数据爬取；数据治理；合规；权益均衡

中图分类号：D912.29 文献标志码：A doi：10.11959/j.issn.2096-0271.2026044

Research on Legal Risks and Regulatory Paths of Corporate Data Crawling: From the Perspective of Balancing Rights and Interests

Gong PengCheng, Zhang Yangyang
School of Law, Hohai University, Nanjing 210003, China

Abstract
As the demand for data continues to soar, enterprises often resort to data crawling to enhance efficiency. Data scraping gives rise to multiple risks, including data privacy and security issues, data contamination that endangers data governance security, and unfair competition that undermines market order.. The existing legal framework faces challenges such as ambiguous legal boundaries for crawling behaviors, the uncertain legal validity of the Robots protocol, and deficiencies in data compliance governance mechanisms. The underlying causes lie in the imbalance between two pairs of core rights and interests: "data efficiency and data security," as well as "the autonomous

management rights of data platforms and the obligation of tolerance for data access." Therefore, from the perspective of balancing rights and interests, it is proposed to establish a multi-party collaborative, classification-based, and tiered data governance system, improve data compliance governance mechanisms, optimize the classified application of the triple authorization principle, clarify the Robots protocol's legal validity as a contract, and safeguard the right to data deletion. These measures aim to address the dilemmas faced by data crawling and facilitate enterprises' digital transformation.

Key words
Data crawling, data governance, compliance, balance of rights and interests

0 引言

数字经济时代，数据已经成为新兴的关键生产要素，驱动生产产业变革与商业模式的创新。各行各业积极响应时代号召，推动数字化转型，但是由于对数据需求越来越大，为了提升数据获取效率同时也降低运营成本，企业通过网络爬虫技术来获取外部数据已经成为普遍现象。然而，爬虫作为一种技术中立工具，数据爬取技术在商业应用中的失范行为导致了诸多风险，包括但不限于数据隐私受到侵害、市场公平竞争秩序被破坏以及数据治理出现安全危机。

现有的关于数据爬取的法律风险研究主要聚焦在以下三方面：一是从数据爬取行为本身出发，探讨该技术是否应该得到应用，有的学者从安全角度出发，认为非法爬虫方式侵害了消费者权益，应该禁止。林慰曾[1]认为，不当爬虫在技术上造成新的数字鸿沟，在现实中造成新的不平等，通过数据爬取的方式的准确地获得了用户画像信息，但是用户却没有因此获得合理的对价，造成了不平等、不公正。有的学者认为在数据爬取问题上的利益权衡应当分场景。丁晓东[2]认为应当对平台数据进行场景化保护，平衡企业权益与数据开放；二是从不同的实体法角度针对数据爬取带来的问题基于解决之道，江海洋[3]、侯跃伟[4]、刘营[5]等从刑法学角度，周樨平[6]、陈兵[7]等从民商法角度对数据爬取行为提出了各自部门法的规制路径；三是对法律规则平衡的研究，聚焦效率与安全、数据平台自主经营权与数据接入容忍义务的关系。随着技术的不断发展，定量分析方法的应用不断普及，如许可运用阿历克西“权重公式”细化数据权益[8]。

尽管现有研究为企业数据爬取行为带来风险的规制提出了宝贵意见，但仍存在以下缺口：首先，对于风险研究呈碎片化现象，多侧重于单一风险类型，缺乏对隐私安全、数据治理安全与竞争秩序风险的体系化关联分析。其次，对产生困境原因剖析深度不足，对法律规制困境的探讨多停留在规范缺失的表层，未能深入揭示其背后“数据效率与安全”、“平台自主权与容忍义务”等深层权益冲突的法理本质。最后，对策构建的系统性较弱，提出的解决方法往往偏重某一环节（如授权模式或协议效力），缺乏一个整合了治理架构、合规激励、规则细化与权利保障的协同性制度框架。鉴于此，本文旨在首先通过系统性的文献梳理，厘清既有研究的贡献与不足，进而在此基础上，系统剖析企业数据爬取的风险表征，检视其面临的法律困境，并创新性地从权益均衡的视角探求风险防范的法理进路，最终构建一套整合了多层

次、精细化措施的法律制度对策体系，以期为规范数据爬取行为、促进数据要素合法有序流动提供更具系统性和操作性的智识参考

1 数据爬取的风险表征

1.1 数据抓取损害数据隐私安全

数据爬取的首要风险在于其对公民个人信息与隐私权构成直接威胁。该风险的根源在于，爬虫技术的自动化、规模化特性系统性绕过了个人信息处理的“知情-同意”核心原则（《中华人民共和国个人信息保护法》第13条）。2019年，中国互联网金融协会发布《关于增强个人信息保护意识依法开展业务的通知》，该文件即可被视为对业内此种滥用乱象的官方确认与回应，直接指出部分机构以“大数据”为名，违规使用爬虫技术收集个人信息的行为。实证层面，诸如“公信宝”、“魔蝎科技”等公司因涉嫌非法爬取并交易用户数据而被查处，[9]揭示了此类行为从个体违规向产业化黑灰产蔓延的趋势。更具危害性的是，这些被非法获取的数据流入了借贷催收等领域，衍生出“电话轰炸”、“曝光通讯录”等暴力催收手段[10]，不仅侵犯公民个人信息权益，更破坏了金融秩序与社会稳定。这表明，数据爬取已不再是单纯的技术问题，而是演变为一个关乎公民基本权利保障与社会治理的严峻挑战。

1.2 数据污染危及数据治理安全

数据污染是指由于人为篡改或者其他因素导致通过人工智能算法得到的数据与原有数据不匹配，破坏了原有数据的客观性、真实性[11]。具体而言，是指用以开展业务需求的数据从源头或者交易过程中被恶意篡改，遭到破坏，导致根据模型得出的结果与真实情况不一致，具体来说，当需求方从外部获得用户个人数据来对用户进行画像时，由于该部分数据受到污染，得到的画像并非用户真实反映，阻碍了正常业务的开展，同时由于清理这部分遭到污染的数据，还要额外付出成本。更具威胁的是诸如“数据投毒”等针对性攻击，导致爬取的数据从源头上就已污染，若将其用于训练人工智能模型，将直接导致模型偏差与决策失误，危害性极大，数据爬取的过程是第三方机构运用技术手段从网页等互联网途径获取用户个人数据，那么在爬取的过程中第三方机构的技术是否得当，爬取而来的数据是由第三方获得，又如何确定不会遭到篡改？以及如果爬取的数据本身已经遭到污染，那么爬取之后又导致二次污染。由于批量生成的虚假内容影响了用户的阅读、选择和喜好等认知领域[12]，同时又如“数据投毒”导致爬取的数据从源头上就已经受到污染，在爬取的过程中又出现人为篡改的现象，极大地破坏了数据的真实性，不利于正常业务的开展。

1.3 不正当竞争破坏市场秩序

数据爬取行为若不受约束，将严重扰乱市场竞争秩序，构成不正当竞争。我国司法实践普遍认可，企业对于其投入大量人力、物力收集、加工而成的数据产品享有竞争性权益。在“淘宝诉美景”案中[13]，法院确立了“三重授权原则”，明确未经用户与平台双重授权，第三方不得爬取和使用非公开数据，旨在保护数据产品的市场价值与创新激励。反之，若允许经营者肆意爬取他人投入资源形成的核心

数据，则属于“不劳而获”和“搭便车”行为，实质性地削弱了原始数据控制者的竞争优势，如“谷米诉元光”案[14]所示。此类行为不仅打击企业数据创新的积极性，长远来看更会导致“劣币驱逐良币”，使得市场竞争从质量与创新竞争异化为数据掠夺能力的竞争，最终损害整体市场效率和消费者福利。因此，对数据爬取进行竞争法规制，是维护健康市场生态的必要之举。

2 企业数据爬取的法律困境检视

随着数据作为战略性资源在经济发展中的地位日益突显，各行各业对数据的需求日趋增长，与此同时，不规范的数据爬取行为带来诸多风险，前述数据爬取所带来的各类风险之所以在实践中难以根除，其深层原因在于当前法律制度面临以下三重困境：

2.1 爬取行为法律界限模糊

当前，数据爬取行为面临的首要法律困境是其合法与非法的边界极其模糊。这种模糊性根源于立法的原则性规定与滞后性。尽管《数据安全法》第32条规定数据收集应“合法、正当”，但何为“正当”缺乏细化标准。此立法缺失直接导致司法裁判陷入困境，不得不依赖抽象原则进行个案裁量，进而产生裁判标准不一的混乱局面。例如，在“百度与大众点评”案[15]中，法院以“超过必要限度”和“违反商业道德”为由认定百度构成不正当竞争；而在“淘宝诉美景”案中，法院则创新性地提出“三重授权原则”作为判断标准。这些案例虽提供了参考，但恰恰反衬出缺乏统一、明确的法律规则所带来的不确定性。企业因此无法对其行为的法律后果形成稳定预期，既可能因畏法而束手束脚，抑制数据流通，也可能因法无明文规定而心存侥幸，滋生违法行为。这种不确定性本身已成为数据市场发展的制度性障碍。

2.2 Robots协议的法律效力未定

Robots协议（又称“爬虫协议”）是网站所有者通过robots.txt文件向爬虫程序表明抓取意愿的技术规范，其核心功能在于沟通与协调，而非强制阻止。然而，这种技术规范的法律效力为何，无论在立法还是司法层面均处于未定状态，构成了数据爬取监管的又一重困境。[16]，该困境集中体现于国内外司法裁判的分歧之中：在国内“微梦诉字节”案中[17]，司法内部即存在认知差异：一审法院认为其“仅是文字层面的宣示，并非技术措施”；二审法院虽予以尊重，但仅将其视为“互联网行业应遵守的行业规范”，其效力认定需依附于《反不正当竞争法》等规定，尤其是新修订的第十三条第三款关于数据获取行为的具体规定之下。在此案中Robots协议仅是评判行为正当性的考量因素之一，而非独立的裁判依据。与之形成鲜明对比的是美国“HiQ诉领英”案[18]，法院甚至作出了更激进的判决，认为爬取公开数据的行为即便违反Robots协议，也如同“进入一家未锁门的超市”，不构成违法。

这种巨大的司法分歧揭示了Robots协议的法律困境：一方面，它作为重要的行业自治规范，体现了网站管理者的主观意愿；另一方面，因其缺乏法律明文规定，其约束力完全依赖于法官在个案中的自由裁量，可能被认为完全不具法律意义。这种不确定性导致其在实践中陷入“遵守靠自觉，违反无代价”的尴尬境地。其直接

后果是：数据源网站无法有效通过法律途径预防和制止恶意爬取，往往只能诉诸于成本高昂且容易引发“军备竞赛”的技术封堵手段，这无疑加剧了网络空间的资源消耗和无谓对抗，从而极大地削弱了其作为数据治理工具的实际效能。

2.3 数据合规治理机制存在缺陷

当前，数据合规治理机制存在系统性缺陷，无法有效回应数据爬取带来的复杂挑战，主要体现在以下三个方面：

其一，法律标准模糊，可操作性不足。尽管《数据安全法》《个人信息保护法》等确立了数据合规的基本框架，但其规定多呈原则化、抽象化特征。例如，《数据安全法》第27条要求“开展数据处理活动应当依照法律、法规的规定”，但“应当”如何依照，缺乏具体的实施细则与判例指引。这种立法上的高度开放性，导致企业在数据爬取实践中面临“刑事违法性与行政违法性双重认定障碍”[19]：刑法上可能适用非法获取计算机信息系统数据罪，而行政法上何为“过度收集”却界限不明。同时，数据合规管理的具体义务与问责机制尚未完全建立，企业如同在迷雾中航行，合规缺乏清晰、稳定的航标。

其二，企业内生合规动力严重不足。当前我国的合规模式具有强烈的“公权主导”色彩，[20]合规在很大程度上源于规避行政处罚的外部压力，而非提升核心竞争力、获取市场信任的内在需求。这种“外生式”的合规推动模式，虽在初期见效快，但容易导致“纸面合规”与“应对式合规”，资源利用效率低下，且难以持续[21]。企业未能将数据合规内化为公司治理的有机组成部分，导致合规体系缺乏真正的“内生活力”。

其三，监管模式滞后，难以适应数字化发展。传统监管侧重于事后处罚与静态合规检查，对于数据爬取这种技术性强、变化快的行为，显得迟钝且被动。虽然部分企业建立了内控制度，但“各显差异”，缺乏行业共识与标准，无法系统性地应对数据爬取的全链条风险。要破解此局，亟需引入“敏捷治理”理念，推动监管从传统走向现代，从事后走向事中事前，从单向命令走向多元协同，以匹配数据要素的动态发展需求。

综上所述，数据爬取的治理困境，根源在于当前合规治理体系在规则、动力与模式三个维度上的系统性失灵。破解之道，绝非简单地禁止或放开，而是必须对这套体系进行一场深刻的、系统性的重构与升级。

3 数据爬取风险防范的权益均衡进路

前述风险与困境的根本原因，在于数据效率与数据安全、平台自主权与容忍义务之间的权益失衡。破解困境，必须首先从法理上厘清这些权益的均衡路径。

3.1 数据效率与数据安全的均衡

数据爬取规制面临的核心法理困境，在于如何衡平数据流通利用的效率价值与数据安全保护的秩序价值。这两大价值并非天然对立，但在资源有限的条件下，其实现路径往往存在张力与优先序的抉择。

这种价值抉择在学界与司法界引发了深刻分歧，形成了两种鲜明的规制倾向：

一种是以安全优先为导向的“权利保护路径”。该路径强调数据的人格属性与安全属性，认为未经授权的数据爬取本质上

是对个人隐私权、企业财产权或竞争秩序的侵害。有学者认为，[22]爬虫技术加剧了数据控制者与用户之间的权力不平等，应对其适用严格规制。我国司法实践也多采此路径，如在“谷米诉元光”等案中，法院通过《反不正当竞争法》对“不劳而获”和“搭便车”行为进行惩戒，体现了对既有数据权益和市场竞争秩序的强力维护；另一种则是以效率优先为导向的“数据自由流通路径”。该路径强调数据的公共产品属性与经济价值，主张对公开数据的获取应秉持宽容态度，以避免形成数据垄断，阻碍创新。美国法院在“HiQ诉领英”案[23]中的判决即是此路径的极端体现，其将爬取公开数据类比为“进入一家未锁门的超市”，几乎否定了Robots协议对公开数据的任何限制效力。许可也认为，在特定条件下，数据爬取具有正当性，其边界需审慎界定[24]。

国内外规制理念的差异，折射出的是对数据本质属性及其法律定位的不同认知偏好。然而，任何一种价值的绝对化都可能带来弊端：绝对安全导向的规制会窒息数据活力，抑制数字经济发展；而绝对自由导向的规制则会漠视个体权利，侵蚀数据治理的伦理基础。因此，真正的出路并非择一弃一，而是追求一种“动态的均衡”。数据效率与安全也绝非简单的零和博弈，二者存在辩证统一的关系：安全、可信的数据环境是数据要素得以大规模、高效率流通的“信任基石”，能降低交易成本，促进更广泛的数据协作；反之，高效、合规的数据流通机制能为安全技术的研发与应用提供更丰富的场景和资源，推动安全治理水平的不断提升。对企业而言，固然需承担合规成本，但投资于数据安全实则是构筑其长期核心竞争力的关键一环。立法与司法者的智慧，则在于根据数据类型、应用场景、社会公共利益等因素，进行因时因地的精细化权衡，在动态发展中寻求特定语境下的最优均衡点。

3.2 数据平台自主经营权与数据接入容忍义务的均衡

数据爬取引发的权益冲突，在平台层面集中体现为数据平台的自主经营权与其负有的数据接入容忍义务之间的张力。《关于构建数据基础制度更好发挥数据要素作用的意见》提出的“谁投入、谁贡献、谁受益”原则，为确认平台对其投入资源形成的衍生数据产品享有自主经营权提供了政策依据[25]。此种权利是激励数据生产和创新的法律工具。然而，平台同时作为数字生态的关键节点，亦需承担一定的容忍义务，即在特定情况下允许他人合法接入和利用数据，此乃防止数据垄断、促进公平竞争和发挥数据网络正外部性的必要限制。

破解这一均衡难题，需引入类型化与比例原则的分析框架：首先，需对数据进行了类型化区分，此为判断容忍义务边界的逻辑起点[26]。可将平台数据区分为“用户生成的原始数据”与“平台加工衍生的数据产品”。对于前者，数据权益与用户关联更为紧密，平台以自主经营权为由设置普遍访问障碍的正当性较弱，其容忍义务范围相对更广；对于后者，平台因投入了实质性劳动与资本（如进行匿名化脱敏、深度分析建模、系统集成等），其享有的自主经营权应受到更强保护，有权要求他人获取授权，容忍义务范围则相对限缩。

其次，在可爬取的数据范围内，爬取行为的本身也需受到比例原则的约束[27]，以检验其行为的正当性：

（1）目的正当性：爬取行为需具有合法、正当的目的，如提升公共服务、进行

学术研究或提供互补型创新产品，而非单纯地“搭便车”或进行数据倒卖。

（2）手段必要性：爬取手段应是实现该正当目的对平台权益侵害最小的方式。例如，如果能通过公开API接口获取，则不应采用技术手段强行突破反爬虫措施。

（3）损害均衡性：爬取行为对平台造成的潜在损害不应与其所追求的利益显失均衡。司法实践中，法院在“百度与大众点评”案中即运用了类似原理，认定百度的大量抓取行为“超过了必要限度”[28]。

数据平台的容忍义务并非无限，其边界取决于数据类型与爬取行为的性质。对于原始数据，平台需秉持开放精神；对于数据产品，平台可行使自主经营权。但无论在何种情况下，任何爬取行为都必须通行于比例原则的审查之下，确保其正当、必要且均衡。

4 数据爬取风险防范的法律制度对策

数据爬取提高了处理业务的效率，但是不当的爬取模式损害了用户隐私安全，给市场带来威胁，也阻碍了企业的数字化转型，为了解决上述难题，需要在均衡数据效率与数据安全、数据平台自主经营权与数据接入容忍义务的基础上探索新的制度模式，在推动企业发展的同时确保数据爬取的合法、合规。基于前述对权益均衡的法理分析，为有效防控数据爬取风险、破解法律实践困境，本研究提出以下五个方面环环相扣、层层递进的具体制度构想。

4.1 建构多方共治的数据分类分级协同治理体系

为系统应对数据爬取所引发的混杂性风险，首要从宏观治理架构入手，建构一个“多方共治的数据分类分级协同治理体系”。该体系旨在破解当前数据保护标准“政出多门、执法不一”的困境，为数据爬取行为的合规性判断提供清晰、统一的底层规则基准。其构建应从以下三个层面协同推进：

4.1.1 规范数据分类分级的监管标准

推动数据分类分级标准的统一，是破解监管套利与执法冲突的前提。当前，不同立法（如《网络安全法》《数据安全法》《个人信息保护法》）及其配套规范对数据分类分级的要求存在差异，导致同一数据对象可能同时触发多重且不一致的保护义务。例如，一家爬虫公司过度抓取用户轨迹数据，可能同时面临《刑法》第285条的“非法获取计算机信息系统数据罪”、《民法典》人格权编的侵权责任以及《反不正当竞争法》的规制，法律适用复杂且结果不确定[29]。为终结此种混乱，必须超越部门立法的局限，在国家层面进行顶层的制度设计。应严格落实《数据安全法》第21条所确立的“国家数据分类分级保护制度”，由国家数据主管部门（如国家数据局）牵头，联合网信、工信、公安、等各行业监管机构，并吸纳技术专家与法律学者参与，共同制定并发布一份权威的《数据分类分级指南》。该指南应：确立统一的核心数据分类框架（分为公共数据、个人数据、重要数据、核心数据）；明确各类型数据的基本定义与识别标准；划定不同类别数据在爬取、流转、利用各环节的基本安全要求与禁止性行为。此举旨在为全行业提供一套“标准答案”，从根本上杜绝因标准不一而导致的管理混乱与法律风险。

4.1.2 建立数据爬取黑名单制度

在统一的国家标准基础上，企业应据此建立内部可操作的“数据爬取黑名单制度”，这是将宏观原则落地为微观管理的关键桥梁。该制度的核心在于将《数据安全法》中的“核心数据”与“重要数据”以及企业内部认定的敏感数据，明确列为禁止爬取的“负面清单”。

其构建应遵循以下步骤：第一，以《数据安全技术 数据分类分级规则》等国家标准为根本依据，将企业处理的所有数据标识为核心数据、重要数据与一般数据[30]；第二，明确将“核心数据”与“重要数据”自动纳入爬取黑名单，严禁任何形式的爬取行为。例如，涉及国家宏观经济、国防、尖端科技等领域的核心数据，其爬取行为可能直接危害国家安全，必须绝对禁止；第三，对于“一般数据”中的敏感类别，具体涵盖超过一定规模的个人隐私数据、企业的核心算法参数、未公开的重大经营决策等，企业可根据自身业务特点与风险承受能力，通过内部评估将其增补入黑名单；第四，对于未被纳入黑名单的一般数据及公开数据，其爬取行为也需严格遵循“合法、正当、必要”原则，并履行相应的授权程序。

黑名单制度的价值在于其极大的明确性：它清晰地划出了数据爬取的“禁区”，使得数据流通的“绿灯”区域也更加明确，从而极大降低了企业的合规不确定性，为数据爬取业务的开展提供了清晰的正面指引与负面边界。

4.1.3 完善应急处置机制

分类分级与黑名单制度明确了静态规则，而完善的应急处置机制则是应对动态安全漏洞、违规爬取事件的必要动态补充。企业必须建立贯穿数据全生命周期的安全监控与应急响应流程。该机制应包含：1.监测与预警：利用技术手段对数据访问、尤其是爬取行为进行实时日志记录与异常流量监测，设定预警阈值（单IP高频访问、非正常时间访问敏感数据等），实现对潜在违规爬取行为的早期发现。2.应急响应：一旦发现数据泄露或恶意爬取事件，立即启动应急预案。流程应包括：立即阻断攻击源、评估事件影响范围（涉及哪些数据、数据级别为何）、依法向监管部门和受影响用户报告、并及时通过技术手段进行溯源和证据固定。3.复盘与改进：事后必须对事件进行彻底复盘，分析制度漏洞，包括黑名单是否未覆盖、权限设置是否过宽等并修订完善，实现安全管理能力的螺旋式上升。

《中国人民银行业务领域数据安全管理办法》等相关法规的要求，正是强调了通过有效的管理措施，确保每一位相关业务人员都清晰知晓自身在应急流程中的职责，从而实现“技管结合”，将分类分级制度从纸面落到实处。

4.2 完善数据合规治理机制

4.2.1 细化数据合规法律标准

破解数据合规“大而空”的困境，关键在于推动原则性法律向具象化、流程化的操作规则转变。这需要从立法细化与标准制定两个维度同步推进。首先，应充分发挥现有法律框架下实施指南与国家标准的作用，实现监管要求的精细化。《数据安全法》第17条明确要求国家推进数据开发技术和数据安全标准的体系建设。因此，监管机构可依据《数据分类分级指南》，针对不同级别的数据设定差异化的处理规范：对于核心数据，严格禁止任何形式的爬取

与跨境流动，其处理权限应仅限于国家授权机构；对于重要数据：实行“事前评估与审批制”。爬取重要数据必须向省级以上网信部门申报数据安全评估，获批后方可进行，且过程中需采取严格加密、脱敏等保护措施；对于一般数据，在遵守《个人信息保护法》中“告知-同意”等基本原则的前提下，允许合规爬取与利用。

其次，应重视“软法”的补充作用。最高人民法院可通过发布典型案例、各地监管机关可通过发布行政执法指导案例，为数据爬取的合规边界提供虽无强制力但极具参考价值的行为指引。这些来自实践一线的司法与执法案例，是抽象法律原则最生动的注脚，能有效填补成文法的模糊地带[31]。最后，企业需将上述外部要求内化为覆盖数据全生命周期的内部操作规程。该细则应贯穿数据采集、传输、存储、使用、删除、销毁各个环节，在采集环节，需进行合法性前置审查，确保数据源及爬取手段合法，并完整记录数据来源、采集方式、时间等元数据，以满足《数据安全法》第27条规定的数据安全管理制度要求。传输与存储环节，必须采取加密、访问控制、去标识化等安全技术措施。在使用与销毁环节，需建立严格的权限管理制度，确保数据在授权范围内使用；并在达到保存期限后，进行不可逆的彻底销毁。通过将合规要求分解到每一个具体操作环节，方能将“纸面上的法”真正转化为“行动中的法”。

4.2.2 完善数据爬取合规审查的激励性机制

扭转“政府推动、企业被动”的合规困境，必须构建一套“激励相容”的机制，使合规成为企业提升市场竞争力的内在需求。该机制需综合运用正向激励与反向约束两种手段。正向激励方面，探索将数据合规投入与税收优惠、专项补贴、绿色通道等政策工具相挂钩。例如，对于通过国家认可的合规认证的企业，可依法给予一定比例的所得税减免。此举旨在降低合规成本，让守信者得实惠，引导资源向合规企业倾斜。强化反向约束方面则严格落实《个人信息保护法》第66条等规定，对违法违规数据爬取行为实施阶梯式惩处：从责令整改、高额罚款，到暂停相关业务、停业整顿，直至对严重失信主体实施市场准入限制。“罚款”必须具有足够的威慑力，确保其金额远高于违规所获利益。同时，建立创新信用评价与披露机制。这是联动“胡萝卜”与“大棒”的关键枢纽。由市场监管总局牵头，构建覆盖数据合规水平的“企业数据合规信用评价体系”。监管机构应从纯粹的事后处罚者，转变为规则的共同制定者与赋能者。通过发布合规指南、推行“监管沙盒”、建立恳谈机制等方式，动态回应技术创新带来的监管挑战，[32]帮助企业尤其是中小企业理解和落实合规要求，最终形成监管与被监管者共同促进数据安全流通的良性互动格局。

4.2.3 明确数据平台容忍义务的合规边界

明确数据平台容忍义务的边界，对于划定数据爬取的合法范围至关重要。该边界的确立，应综合考量数据属性、平台投入及爬取行为方式三大因素，从而在平台权益与数据流通之间找到平衡点。

第一，以数据属性为基本区分标准。对于依法公开的公共数据，数据平台原则上负有较高的容忍义务，无权对合规的爬取行为进行阻止。而对于经过深度清洗、分析、整合的数据报告、用户画像等的平台投入实质性资源形成的衍生数据产品，

平台享有受法律保护的竞争性权益，其容忍义务范围则大幅收缩，有权要求他人获取授权。

第二，引入“技术措施”作为边界客观化的标志。平台对不愿被爬取的数据采取有效的技术保护措施，如尚未设置Robots协议、登录验证、反爬虫技术等，应被视为其行使自主经营权、明确拒绝“容忍”的意思表示。任何绕开或破坏这些技术措施的行为，均可初步推定为具有不正当性。

第三，爬取行为自身必须符合比例原则。即便针对可爬取的数据，爬取行为也需目的正当、手段必要、影响均衡。例如，以研究为目的少量爬取公开评论数据可能具有正当性，但同一主体对相同数据进行全量复制用于商业竞争，则可能构成权利滥用，超出了平台应容忍的限度。

通过上述“数据+技术+行为”的三重判断框架，数据平台容忍义务的模糊边界得以清晰化、可操作化，从而为数据平台的合规管理及第三方的合规爬取提供了稳定的法律预期。

4.3 优化三重授权原则的分类应用

为解决传统三重授权原则在大数据场景下成本过高、阻碍流通的适用困境，本文提出基于“数据可识别性”（是否可识别特定个人）与“平台投入度”（是否为平台实质性加工衍生的数据产品）两个关键维度，将数据划分为以下四种类型：第一类数据（可识别性+衍生数据）、第二类数据（可识别性+非衍生数据）、第三类数据（不可识别性+衍生数据）、第四类数据（不可识别性+非衍生数据）。在此基础上，对三重授权原则的具体适用进行精细化分类，构建一个梯度化的授权框架（见表1），以实现数据保护与利用的精准均衡。

表1 三重授权原则的分类适用框架

数据类型	授权要求	法理依据
可识别性+衍生数据(如精准用户画像)	用户同意+平台授权+用户再次授权	保护用户个人信息权与平台劳动收益
可识别性+非衍生数据(如用户原创帖子)	用户同意	数据权益主要归属于用户,平台未付出实质性衍生劳动
不可识别性+衍生数据(如脱敏后的群体行为分析报告)	平台授权	数据已匿名化,不涉及个人信息,但平台投入了加工成本
不可识别性+非衍生数据 (如公开的天气数据)	遵循必要性原则,可无需授权	数据已公开且无特定权益主体,但爬取行为需合理、适度

该框架的核心在于突破“一刀切”的严格授权模式，实现数据流转效率的合理化提升：

对于第一类数据，因涉及核心个人信息与平台核心资产，必须坚持最严格的三重授权，司法实践在“微博诉脉脉”案等判例中已对此予以确认。

对于第二类数据，数据权益主要源于

用户，平台仅提供托管服务。因此，爬取此类数据仅需获得用户同意，无需平台的额外授权，这避免了平台以“自主经营权”为名对用户数据的流通进行不合理限制。

对于第三类数据，因其已不可识别特定个人，不再适用个人信息保护规则。但平台因投入加工劳动形成了数据产品，故爬取仍需获得平台授权，以保护其竞争性权益。

对于第四类数据，其为已公开的非个人信息，可纳入“合理使用”范畴。爬取行为只需遵循《民法典》等法律规定中的“必要性”原则，避免对数据源网站造成过度负担即可。[33]

此分类应用方案并非削弱保护，而是将有限的合规资源集中于真正高风险、高敏感的数据处理活动上，是对三重授权原则的现代化发展与精准化适用，符合数据要素市场化配置的内在要求。

4.4 明确Robots协议法律效力

为赋予Robots协议明确的法律效力，破解其当前“有规范、无强制”的困境，最直接且可行的路径是在司法实践中将其认定为介于“要约”与“格式条款”之间的默示合同。其法理在于：网站通过发布robots.txt文件，已向所有爬虫程序清晰地表达了其关于数据抓取的意愿和规则（这是一种要约）。爬虫程序在知晓该协议内容后仍选择访问该网站，即可视为以行为默示接受了该条款，双方之间由此成立了一份关于数据抓取规则的合同。这一解释路径完美地契合了《民法典》第471条关于合同订立形式的规定，也为《反不正当竞争法》第2条的适用提供了更为坚实和具体的行为规则基础。由此产生的法律后果是清晰的：对于善意爬虫方，遵守协议意味着其抓取行为获得了网站的默示许可，奠定了其行为合法性的基础。对于违反协议者，其行为首先构成违约，网站可依据《民法典》合同编要求其承担停止侵害、赔偿损失等责任。这种违约责任举证更容易，法律关系更清晰，对于恶意爬虫方，其故意违反协议的行为可直接依据《反不正当竞争法》第十三条第三款等规定认定为不正当竞争行为，需承担相应的法律责任。

因此，在立法暂时缺位的情况下，由最高人民法院通过发布典型案例或司法解释，明确Robots协议的合同属性，是当前以最小成本激活这一行业规范、大幅提升数据爬取规则确定性的最优方案。

4.5 保障数据的删除权

保障删除权是约束数据爬取行为、贯彻“目的限制”与“存储期限最小化”原则的最后一道关键闸门。企业不能仅被动响应用户请求，更需建立主动的、可验证的数据删除机制。

首先，将删除权嵌入数据爬取的全流程。企业在规划爬取项目时，就应在设计蓝图（Privacy by Design）中预设数据的存储期限与删除触发条件，如“项目结束即删除”或“自爬取之日起满2年自动删除”。这不仅是对《个人信息保护法》第19条的遵守，更是将合规管理前置化的体现。其次，建立差异化的删除策略。应根据数据分类分级结果执行删除：对于爬取获得的个人数据，应严格履行《个人信息保护法》第47条规定的删除义务，并采用技术手段确保数据不可恢复；对于爬取获得的非个人数据，也应遵循企业的存储期限政策，定期清理非必要保留的数据，以降低合规风险。

最后，构建可审计的删除验证机制。企业需对数据删除操作进行日志记录，并能向监管机构提供证明其已履行删除责任的证据。这将促使删除权从一项抽象的用户权利，落地为一项可被监管、可被审计的企业内部管理流程，真正实现“彻底的删除”。

5 结语

本文系统研究了企业数据爬取行为的风险、困境与规制路径。首先识别了数据爬取在隐私安全、治理安全与竞争秩序三个维度的风险表征；进而剖析了其面临的法律困境，包括行为边界模糊、Robots协议效力未定及合规机制缺陷，并揭示了困境背后数据效率与安全、平台自主权与容忍义务的深层权益冲突；最后，针对性地提出了建构分类分级治理体系、完善合规激励机制、优化三重授权原则分类应用、明确Robots协议合同效力及保障数据删除权等一系列制度对策。本研究的创新与贡献主要体现在：其一，理论层面，突破了以往就事论事的分析框架，从“权益均衡”这一法理学核心视角切入，为理解数据爬取规制难题提供了更具解释力的理论工具。其二，实践层面，所提出的“数据分类分级+爬取黑名单”治理框架、对“三重授权原则”基于数据类型的精细化分类应用、以及将Robots协议明确为具有合同法律效力的建议，均为立法、司法与企业合规提供了清晰且可操作的方案设计，有助于推动从原则性规定向规则之治的转变。本研究亦存在一定局限：首先，主要采用定性研究和案例分析，缺乏对大样本数据的定量分析，未来研究可通过问卷、访谈等方式，对数据爬取的规模、成本收益及企业合规现状进行实证考察，使研究结论更具说服力。其次，本文侧重于国内法规制，未充分探讨跨境数据爬取所带来的国际法律冲突问题，这将是未来研究的重要方向。此外，随着人工智能技术的发展，AI生成数据的内容产权与爬取规则必将带来新的挑战，值得持续关注。总而言之，规范数据爬取并非旨在扼杀技术，而是为了引导其在合规轨道上发挥更大价值。唯有在法治框架下平衡各方权益，方能真正释放数据要素潜力，护航数字经济行稳致远。

参考文献：

[1] 林慰曾.数据爬虫技术对金融信息安全的冲击及制度回应[J].北京航空航天大学学报(社会科学版),2022,35(04):161-169.

[2] 丁晓东.数据到底属于谁?——从网络爬虫看平台数据权属与数据保护[J].华东政法大学学报,2019,22(05):69-83.

[3] 江海洋.数字时代数据爬取的刑法规制:法益界定与数据确权[J].比较法研究,2024,(02):149-163.

[4] 侯跃伟.共享视角下数据爬取行为刑法规制理念重塑与路径展开[J].江苏社会科学,2024,(02):165-174.

[5] 刘莹:《网络爬虫之刑事责任》,载《军法专刊》2022年第4期。

[6] 周樨平.数据爬取的不正当竞争认定规则研究[J].南大法学,2023,(02):87-102.

[7] 陈兵.保护与竞争:治理数据爬取行为的竞争法功能实现[J].政法论坛,2021,39(06):18-28.

[8] 许可.数据爬取的正当性及其边界[J].中国法学,2021,(02):166-188.

[9] 大数据公司被查背后 网络爬虫侵犯隐私产业链整肃 [EB/OL]. (2019-09-18)[2024-01-20]. https://finance.sina.com.cn/money/bank/bank_hydt/2019-09-18/doc-iicezzrq6556039.shtml

[10] 山西经济日报. 消费金融暴力催收乱象多：电话打爆通讯录 逾期还贴大字报[EB/OL]. (2024-03-23) [2024-04-02]. https://www. 360kuai. com/pc/920398ecaa668f9fe? cota=3&kuai_so=1&sign=360_57c3bbd1&refer_scene=so_1.

[11] 观察·人工智能引发的隐私与数据保护风险[EB/OL]. (2021-09-29) [2024-06-11]. https://www. sohu. com/a/492827067_121123759.

[12] 刘晓春.数据功能类型视角下数据污染的治理维度[J].人民司法,2024,(10):15-20.

[13] 浙江省杭州市中级人民法院(2018)浙01号民终7312号民事判决书.

[14] 广东省深圳市中级人民法院(2017)粤03民初822号民事判决书.

[15] 上海知识产权法院(2016)沪73民终242号民事判决书.

[16] Robotstxt.org. 爬虫协议相关内容[EB/OL]. [访问日期 2024-06-29]. https://www.robotstxt.org/robotstxt.html.

[17] 北京市高级人民法院(2021)京民终281号判决书.

[18] HiQ Labs, Inc. v. LinkedIn Corp., No. 17-16783(2017), paragraph 6.

[19] 齐英程.数据合规协同激励体系的构建与完善[J].东方法学,2024,(02):98-108.

[20] 汪青松,邱欢.合规制度发展的中国范式及其与商法关系探析[J].重庆社会科学,2023,(12):185-202.

[21] 刘盛.现代金融体系视野下的金融法：理念信守与制度表达[J].政治与法律,2022,(11):80-95.

[22] 同1

[23] HiQ Labs, Inc. v. LinkedIn Corp., No. 17-16783(2017), paragraph 6.

[24] 同8

[25] 王轶. 加快构建数据基础制度,助推数字经济和数字文明建设 [EB/OL]. 中华人民共和国国家发展和改革委员会官网. [2024-07-22]. https://www. ndrc. gov. cn/xxgk/jd/jd/202212/t20221219_1343657.html.

[26] 周樨平.数据爬取的不正当竞争认定规则研究[J].南大法学,2023,(02):87-102.

[27] 陈兵,姚俊羽.公开数据保护的理念澄清与路径选择[J].中国特色社会主义研究,2024,(02):38-52.

[28] 参见上海知识产权法院(2016)沪73民终242号民事判决书.

[29] 陈咏梅,张姣.跨境数据流动国际规制新发展：困境与前路[J].上海对外经贸大学学报,2017,24(06):37-52.

[30] 数据安全技术 数据分类分级规则.

[31] 鄢浩宇.企业数据合规的困境纾解与体系构建[J].华中科技大学学报(社会科学版),2024,38(04):36-45+71.

[32] 薛澜,赵静.走向敏捷治理：新兴产业发展与监管模式探究[J].中国行政管理,2019,(08):28-34.

[33] 徐伟.企业数据获取"三重授权原则"反思及类型化构建[J].交大法学,2019,(04):20-39.

作者简介

龚鹏程（1974–），男，博士，河海大学法学院副教授，主要研究

张阳阳（2001–），女，河海大学法学院民商法研究所助理，主要研究方向为经济法、数据法。

方向为商法基础理论、经济法、金融法、证券法、资本市场法律制度。

收稿日期: XXXX-XX-XX
通信作者:
基金项目:
Foundation Items: