

引用格式：林韬. 我国AI训练数据生产流通的制约因素与应对策略研究. 中国科学院院刊, 2025, 40(4): 672-680, doi: 10.16418/j.issn.1000-3045.20241204003.

Lin T. Study on constraints and policy responses for production and circulation of AI training data in China. Bulletin of Chinese Academy of Sciences, 2025, 40(4): 672-680, doi: 10.16418/j.issn.1000-3045.20241204003. (in Chinese)

我国AI训练数据生产流通的 制约因素与应对策略研究

林 韬

1 香港中文大学（深圳） 前海国际事务研究院 深圳 518172

2 美国华盛顿大学 政治学系 西雅图 98105

摘要 训练数据的数量和质量对人工智能模型的性能至关重要。然而，目前我国训练数据的生产存在数量不足、质量较低、分布零散等问题，受限于商业生态、监管政策和公共数据开发利用的多重制约。为了解决这些问题，文章提出了一系列政策建议，包括：鼓励科研机构生产开源数据集、打造人工智能应用场景、采取“宽进严出”的监管理念、设立知识产权豁免条款、完善个人信息保护实施细则、加快建设全国统一的公共数据平台等。

关键词 人工智能，训练数据，数据要素流通，知识产权，个人信息保护，公共数据

DOI 10.16418/j.issn.1000-3045.20241204003

CSTR 32128.14.CASbulletin.20241204003

训练数据的规模和质量决定了人工智能发展的高度。端侧模型和具身智能等技术路线的演进，对高质量、多模态和规模化的数据资源提出了更高要求。优质数据的可得性、真实性和多样性，已成为人工智能技术持续突破的核心基础。

当前，我国人工智能在部分领域已取得显著进展，但整体性能与世界顶尖水平仍有一定差距。除算

力、人才、资金等因素外，训练数据的供给质量和流通效率日益成为制约模型性能和行业创新的关键因素。我国训练数据面临“质量低、数量少、分布散”的突出问题。优质中文语料积累不足^①、公共数据流通不畅、垂直行业数据标注成本高企，加之版权与个人信息保护等监管障碍，使训练数据的获取、使用和共享成本居高不下。在此背景下，合成数据逐渐成为

修改稿收到日期：2025年4月9日

① 阿里研究院. 大模型训练数据白皮书. (2024-05-01)[2025-04-10]. https://runwise.oss-accelerate.aliyuncs.com/sites/15/2024/06/创新研报|2024大模型训练数据白皮书_阿里研究院.pdf.

重要补充手段,但其在真实性和多样性方面仍存在局限^[1,2],难以全面缓解优质数据的结构性短缺问题。

提升数据质量、促进数据流通,是我国在人工智能竞争中取胜的关键。数据要素的复制成本低、资产专用性强,对训练数据的生产和流通形成了诸多制约。只有充分考虑数据要素的基本特性,制定针对性强的政策措施,才能为人工智能发展奠定坚实基础,推动产业健康可持续发展。

1 商业生态对训练数据的影响

人工智能技术的迅速发展,使得相关企业对优质数据的需求不断攀升^②。但与此同时,这一需求的增长受到技术方案、商业应用和资本市场的深刻影响。

(1) 训练数据具有较强的资产专用性,往往为特定领域和任务定制,难以跨场景通用^[3]。例如,ChatGPT-o1和DeepSeek-R1等推理模型,在强化学习阶段需要高质量的思维链数据^③作为“冷启动”素材^[4],这与以往所需要的数据大不相同。再如,一些国内的自动驾驶企业发现,训练自动驾驶的端到端大模型,导致以前积累的路测数据只有大约2%可以使用,需要重新考虑数据标注方案^④。这些现象表明,技术方案直接决定了训练数据的收集方式和标注流程,从而影响了训练数据的市场需求。

(2) 应用场景的开发状况直接影响训练数据的质量。目前,国内开放的大部分自动驾驶路测区域,难以涵盖复杂路况,限制了数据的多样性^④。此外,部分平台企业在模型初步可用后,过度强调游戏、社交和广告等场景的商业落地,导致模型停留在低质量数据的循环中,使得性能难以进一步优化。

(3) 平台和初创企业所面临的结构性限制,抑制

了探索前沿技术和获取优质数据的动力。部分平台企业因组织惰性和股东干预,倾向于将资源投入低风险项目,导致训练数据的低水平重复。初创企业则受到风投资本短期趋利行为和“明股实债”对赌协议的影响,倾向于短期变现,忽视底层数据积累,造成总体技术水平始终落后于人工智能前沿水平的国家。

总体来看,商业生态不仅影响着训练数据生产和流通的市场需求,还影响着企业生产优质数据的动力。在监管合规压力较大的环境中,只要技术方向明确、商业利润可观、资金投入预期清晰,即便优质数据本身的合规性较为模糊,行业存在一定的灰色地带,企业依然有积极性从事这些数据的生产和交易。

2 监管政策对训练数据的影响

在技术发展遭遇瓶颈、企业对技术和市场前景预期不明确的情况下,监管政策的重要性就会凸显出来,影响训练数据的生产和流通。监管政策带来的不确定性,也会反过来影响数据行业的融资和市场规模。总体来看,国内外监管政策对训练数据的关注点,主要集中在知识产权和个人信息保护等方面。

2.1 知识产权

2023年7月,国家互联网信息办公室等七部门发布的《生成式人工智能服务管理暂行办法》中,第7条明确规定,“使用具有合法来源的数据和基础模型”“涉及知识产权的,不得侵害他人依法享有的知识产权”等。然而,按照现行的《中华人民共和国著作权法》(以下简称《著作权法》),相关企业必须事先获得训练数据中权利人的许可。如果坚持先许可后使用的原则,企业将面临过高的许可成本。《著作权法》第24条明确规定了版权作品合理使用的13种条件,

② LXT. 迈向人工智能成熟之路2024:高管调研报告. (2024-03-24)[2025-04-10]. <https://info.lxt.ai/the-path-to-ai-maturity-2024>.

③ 思维链数据是一种训练数据,它展示了人类解决复杂问题的逐步思维过程,帮助大语言模型学会通过逻辑推理分步骤解决问题。其应用领域包括数学、编程、医疗诊断、法律推理等需要逐步推理和深度理解的任务。

④ 辰韬资本. 端到端自动驾驶行业研究报告. (2024-06-12)[2025-04-10]. <https://mp.weixin.qq.com/s/Nya5T14yMPN8F-be-wXfPg>.

但在人工智能模型训练中使用版权作品的行为，仍处于法律适用的模糊地带，尚无明确的豁免依据。

一个引人注目的案例是秘塔 AI 公司与中国知网 (CNKI) 的版权纠纷。秘塔 AI 公司在使用中国知网学术资源搭建搜索系统时，未获得中国知网的授权。中国知网认为秘塔 AI 公司侵犯了其著作权，而秘塔 AI 公司则主张其行为仅涉及知网论文的题录和摘要，属于合理使用。最终，该事件以秘塔 AI 公司宣布不再收录知网内容而告终^⑤。

部分国家都不同程度放宽了训练数据在版权问题上的要求。以色列^⑥和日本^⑦在版权问题上较为激进，两国行政部门均曾公开声明，人工智能使用版权作品进行训练，属于“合理使用”的范围。欧盟^⑧和英国^⑨也在相关法案中，单独定义了文本和数据挖掘行为 (text and data mining)，并注明这一行为在版权法中的适用条件。日本^⑩和韩国^⑪的行政部门还出台了详细的合规指南，区分不同市场主体的法律责任，并提供具体的合规建议。

2.2 个人信息保护

个人信息保护和数据安全的监管，可能会增加数据采集、标注和人工智能等行业的合规成本，影响数据生产和流通的市场预期。在实际执行层面，数据企业主要面临“难落实”和“缺细则”这两个问题。

(1) “落实难”。监管法规中的个别条款在当前阶段落实难度较大。① 对中小企业来说，由于资源限制，很难建立法规要求的内部风控和评估机制。② 当前法规对数据“匿名化”的要求过高且标准模糊^[5-7]，算法处理后可能遗漏极端情形，需要人工筛查，而中小企业难以负担此类合规流程的高昂成本。在自动驾驶和机器人等领域，“匿名化”会降低图像和视频数据的质量，削弱人工智能模型的性能^[8]。③ 训练数据的生产链条尤为复杂，企业难以确保所有数据都获得相关个人的知情同意，也难以判断是否违反了数据收集的“最少必要”原则，这进一步加剧了企业的合规风险。

(2) “缺细则”。在“落实难”的情况下，企业急需更加明确的细则，以减少合规过程中存在的不确定性。例如，针对数据采集、标注、交易和使用等环节，“知情同意”和“最少必要”原则应当有例外和免责条款。除汽车行业的数据安全管理规定，其他行业尚未明确“重要数据”范围，个人信息“匿名化”也缺乏明确的技术标准。此外，部分法规对“情节严重”的认定较为模糊，导致法律责任的判定依赖法官的自由裁量，也让企业在实操中难以把握合规边界。

国外在监管法规方面的借鉴。① 欧盟。欧盟《通用数据保护条例》(GDPR) 侧重个人权利保护，企业

⑤ 肖潇. 收到知网“律师函警告”，秘塔AI回应：不理解，但不再收录知网. (2024-08-16)[2024-09-30]. <https://www.21jingji.com/article/20240816/herald/96afa86a4ff9b346a9420ec78439a0f9.html>.

⑥ 以色列司法部. 关于版权材料用于机器学习的观点. (2022-12-18)[2024-09-30]. <https://www.gov.il/BlobFolder/legalinfo/machine-learning/he/18-12-2022.pdf>.

⑦ 日本文部科学省文化厅文化审议会著作权分科会法律制度小委员会. 关于人工智能和著作权的一般性理解. (2024-05-31)[2024-09-30]. https://www.bunka.go.jp/english/policy/copyright/pdf/94055801_01.pdf.

⑧ 欧洲议会, 欧洲理事会. 《数字单一市场版权指令》第4条“文本和数据挖掘的例外情况”. (2019-04-17)[2024-09-30]. <https://eur-lex.europa.eu/eli/dir/2019/790/oj#d1e986-92-1>.

⑨ 《1988年版权、设计与专利法案》第29条A款“文本和数据挖掘的副本用于非商用研究”. (2014-01-06)[2024-09-30]. <https://www.legislation.gov.uk/ukpga/1988/48/section/29A>.

⑩ 日本文部科学省文化厅著作权课. 关于人工智能和著作权的观点. (2024-07-31)[2024-09-30]. https://bunka.go.jp/seisaku/cho-sakuken/pdf/94097701_01.pdf.

⑪ 韩国文化、体育和旅游部. 生成式人工智能与版权指南. (2024-04-15)[2024-09-30]. <https://www.korea.net/Government/Briefing-Room/Press-Releases/view?articleId=391&insttCode=A260123&type=N>.

收集数据通常援引“正当利益”(legitimate interest)条款,但监管机构对“正当利益”判定较为严格。实证研究显示, GDPR颁布后,企业的合规成本上升,导致欧盟企业比北美地区企业平均减少了将近1/4的数据要素投入强度^[9,10]。② 北美。北美地区的个人信息保护立法,以美国加利福尼亚州最为典型。《加州消费者隐私法案》(CCPA)和《加州隐私权法案》(CPR)以促进数据流通为导向,采用“选择-退出”(opt-out)机制,默认企业可以收集、出售和共享个人信息,但用户有权拒绝企业的出售和共享行为。此外,北美地区与个人信息保护相关的多部法案,要求企业对数据进行“去标识化”^⑫处理,整体合规义务相对较轻,有助于降低企业的合规成本。③ 其他国家。其他国家的监管也有值得借鉴的地方,例如:新加坡^⑬将“业务改进或研究目的”视为个人信息保护的例外情况;韩国^⑭则加强监管部门和利益相关者的沟通合作,通过司法解释与合规指引等方式,推动个人信息保护与人工智能产业发展相协调。

3 公共数据开发利用

当前,各地公共数据交易机制在制度设计与实施环节均存在一定障碍,影响了数据流通效率和市场活力。体制内单位沉淀的数据,如医疗记录、教育题库、司法文书、工商登记、图书史料等,具有较高的客观性和准确性,是人工智能训练数据的重要资源。2020年以来,各地纷纷建立政府主导的数据交易机构,作为公共数据授权运营产品的交易场所,以及为

企业之间的数据交易提供撮合服务。但在实践中,前置合规审查比较烦琐,再加上相关法规仍有“难落实”和“缺细则”的地方,增加了场内交易成本。例如,某地数据交易所实行“不合规不挂牌,无场景不交易”的原则,一定程度上阻碍了更多企业进场交易。个别数据交易机构出于特定技术要求或合规目的,要求企业将数据提前存储在交易所内,引发企业的不安全感,挫伤场内交易的积极性。虽然高价值的公共数据需要通过合规认证来保障交易标的的真实性,但当前场内交易的合规审查机制亟待优化。

此外,公共数据的开发利用仍呈现孤岛化的格局,阻碍了公共数据需求和价值的提升。数据要达到一定规模,才能创造足够大的市场价值和需求。目前,公共数据由各地分散开发,授权定价碎片化,在监管上尺度把握不一致,相当于“高速公路上设置多个独立经营的收费站”,增加了需求方整合各地公共数据的交易成本,会在无形中提高市场准入门槛,排斥可能在全国范围内产生创新的潜在优质企业。个别地方政府在数字服务采购和授权运营过程中,容易产生隐性腐败和不正当竞争,也会导致技术标准和公共数据市场的分割,形成数字经济的“地方保护主义”^[11]。而目前数据行业的技术人才、具备数字经济思维的领导干部,以及对公共数据的市场需求,都高度集中在沿海经济发达地区和中央层面。如果让各地区继续自行探索,区域间差距将会越来越大。现在经济发达地区的试点探索已经有了一定成效,如事前统一授权、减少数源单位责任、建立收益挂钩机制

⑫ 去标识化(de-identification)旨在移除或模糊化个人身份信息,使数据不能直接识别特定个人,但仍可能通过其他信息间接识别。匿名化(anonymization)指彻底移除所有可能直接或间接识别个人的因素,理论上无法通过任何手段恢复原始身份信息,或者复原信息需要极大成本。

⑬ 新加坡个人信息保护委员会. 关于在人工智能推荐决策系统中使用个人数据的咨询指南. (2024-03-31)[2024-09-30]. <https://www.pdpc.gov.sg/guidelines-and-consultation/2024/02/advisory-guidelines-on-use-of-personal-data-in-ai-recommendation-and-decision-systems>.

⑭ 韩国个人信息保护委员会. 人工智能时代个人信息安全使用指南. (2023-08-03)[2024-09-30]. https://www.pipc.go.kr/eng/user/ltm/new/noticeDetail.do?bbsId=BBSMSTR_000000000001&nttId=2275.

等^[12]。如果能将这些经验总结推广，就有望推动形成公共数据的全国统一大市场。

4 促进训练数据生产和流通的政策建议

4.1 出台产业政策和激励措施，推动优质数据的生产

(1) 鼓励科教文卫机构生产开源数据，促进科研数据管理机构和开放平台的统一。高质量的科研数据已成为人工智能模型训练和科技竞争的关键资源。例如，获得2024年诺贝尔奖、由英国DeepMind公司开发的人工智能模型AlphaFold，正是依托全球开放的生物蛋白质结构数据库Protein Data Bank实现蛋白质结构预测的重大突破。与此同时，近期北美部分科研数据库^⑮限制中国用户的访问权限^⑯，凸显出科研数据在国际竞争中的战略意义，也反映出我国建设自主科研数据体系的紧迫性。构建系统性的开放科研数据库是一项长期工作，需要持续投入与制度保障。科教文卫机构拥有大量专业人才，是生产优质开源数据的重要力量。因此，建议在科教文卫机构的课题申请和考核中，增加数据集产出的考核权重，鼓励科研人员原创、扩展和长期维护科研数据库；并且，根据科研规律引入类似专利保护期的制度，激励科研人员在合理期限内，依托原创数据产出科研成果。超过规定的保护期后，原则上应将数据库向社会公开。此政策不仅适用于理工科，还应覆盖人文社科领域的数据资源和课题成果，推动跨学科的数据共享和创新。此外，设立统一的科研数据管理机构，有助于规范数据的收集、存储和共享，从而确保数据资源在更广泛的科研领域得到有效利用^[13]。

(2) 持续打造人工智能技术的应用场景，培育物

联网和新型硬件终端市场。扩展和丰富人工智能技术的应用场景，是积累优质数据资源的重要手段。人工智能企业通过产品和服务的实际应用场景，不断收集用户行为数据和反馈信息，为模型训练和算法优化提供丰富的数据支撑。物联网和新型硬件终端能够超越文字、图像、语音等传统模态，收集更多维度的数据，是下一代人工智能的基础。产业政策应继续通过有条件的税费减免、财政补助等形式，支持自动驾驶、增强现实技术（AR）、虚拟现实技术（VR）和智能机器人等新型终端设备进入实际应用场景，积累数据资源，优化模型性能。此外，企业数字化和工业互联网是人工智能数据积累的重要场景，要利用龙头企业和产业链链主企业的示范效应，带动中小企业和传统制造业企业的数字化转型。

4.2 完善数据安全的监管框架，促进优质数据的流通

(1) 在监管调整方面，采取“宽进严出”的监管理念。促进数据要素生产和流通，需要将立法、司法、行政监管和产业政策紧密结合，进行通盘考虑。政策出台应遵循合理的逻辑顺序，建议优先推进立法和司法工作，针对具体问题设立免责条款、出台司法解释，结合公众、行业和政府的需求，对涉及数据要素的法规进行动态调整。随后，再由行政监管机构和行业组织联合推动，完善合规指南，从而使产业政策在促进数据生产和流通方面，发挥更大的作用。我国对数据要素的监管政策较为全面，重视对上游要素和下游应用的监管，强调对事前、事中、事后环节的全链路监管。虽然对行业上游和事前环节的监管可以减少监管成本，但是我们需要审慎评估对技术和市场发展的影响。为了促进人工智能的创新发展，建议监管

^⑮ 如SEER(Surveillance, Epidemiology, and End Results Program), 其是美国国家癌症研究所主导的癌症数据项目, 是全球覆盖范围最广的癌症监测数据库。

^⑯ 美国国立卫生研究院. 执行更新: 加强对NIH受控访问数据的安全措施. (2025-04-02)[2025-04-10]. <https://grants.nih.gov/grants/guide/notice-files/NOT-OD-25-083.html>.

要采取“宽进严出”、包容审慎的理念,适当放宽对行业上游的算法、数据等生产要素在输入端的监管,把监管重点放在行业下游的应用场景和输出端的使用环节,从而有利于提升人工智能模型的性能水平。监管应当考虑不同市场主体的管理能力和有限责任,通过司法解释、负面清单、合规指南等方式,适当放宽事前监管要求,落实事中事后监管,完善事后救济措施。

(2) 在训练数据的著作权问题上,放宽“合理使用”原则的适用范围,出台司法解释与合规指南。

① 为进一步推动数据的广泛使用和创新,建议在《著作权法》中放宽“合理使用”原则的适用情形。例如,参考部分发达国家和地区在知识产权法规中的实践,对“文本和数据挖掘”的情形进行单独定义,为版权作品的采集和使用提供例外和免责条款。司法机关可考虑对数据采集交易和人工智能版权纠纷案件进行集中审理,出台司法解释和指导性案例;以促进人工智能技术创新发展为宗旨,进一步明确“合理使用”的边界。② 在立法和司法工作的基础上,建议知识产权相关的行政机构可以与企业座谈,调研域外国家经验,出台详尽的合规指南。在合规指南中,可以针对人工智能模型的不同阶段,区分版权持有方、模型开发者、服务提供商和模型使用者等不同市场主体的法律责任,并提供相应的合规建议,稳定企业预期。

(3) 在个人信息保护和数据安全方面,设立免责条款和负面清单,强化事中事后监管,完善事后救济措施。① 针对“训练通用型人工智能”这一特定目的,进一步明确适用条款,放宽个人信息保护的限制。在匿名化技术尚未成熟的情况下,依照数据处理的难度和对模型性能的影响,可以允许训练数据中包

含部分个人信息,或者仅去除直接标识符^{①7},只要数据管理和模型输出(包括由模型控制的机器行为)符合安全要求即可。如果包含重要数据,则需履行重要数据相关义务。开发者转卖数据,应遵守一般市场主体的数据安全义务。与此同时,要加强人工智能安全技术的研发,以技术发展促进监管能力的提升。② 落实负面清单制度,在全国范围内加紧推出各行业的重要数据目录,尽快明确“匿名化”等数据处理流程的国家标准与合规指南。当前已有《重要数据识别指南(征求意见稿)》《数据出境安全评估办法》《数据分类分级规则》《匿名化技术应用指南》等政策文件,但大多属于原则性的指导,没有细化到具体的数据名称和技术细节。建议各行业主管部门可以整合试点经验和行业反馈,动态调整负面清单和数据处理技术标准,由国家数据局等部门进行统一梳理整合。如果行业主管部门没有重要数据需要管理,也最好进行明确说明。③ 尽快明确小型个人信息处理者的定义,并出台相应的合规指南。可以仿照欧盟网络安全局出台的《中小企业个人数据安全处理指南》,从企业营收、雇员数量、主营业务等方面,明确划分标准,对合规义务、操作流程和具体监管案例进行梳理总结。④ 制定总体性数据合规指南,梳理上位法细则与模糊条款。建议相关部门统一《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》《中华人民共和国刑法》等法律框架下的解释口径,澄清过往司法案例中语焉不详、容易让企业误解担忧的地方,稳定企业的预期。⑤ 推动发展数据安全保险机制,提升中小企业合规风险应对能力^[14]。《网络数据安全条例》明确,“鼓励保险公司开发网络数据损害赔偿责任险种”。考虑到数据安全保险的风险核算,需要大量网络安全相关的数据,

^{①7} 直接标识符是在特定环境下直接识别个人的号码、特征或代码,包括身份证号、指纹、人脸等。准标识符指的是在相应环境下无法直接识别个人,但结合其他信息可识别个人的属性,如性别、籍贯、工作单位等。

建议保险公司可以通过数据交易、行业联盟、兼并收购等方式，与网络安全和云服务企业密切合作。在合作环节中，需要留意不正当竞争的可能性。在法规调整到位的基础上，可以仿照北美地区2016年提出的《数据泄露保险法案》，通过采购准入、税收优惠等方式，促进数据安全保险快速铺开，帮助完善中小企业的的核心安全内控机制，从长远来看也有利于网络安全技术的创新发展和中小企业的数字化转型。

4.3 加快建设国家级公共数据平台，汇总整合各地各部门数据资源

(1) 建设国家级公共数据平台，实现公共数据开放和授权运营渠道的统一，形成数据要素的全国统一大市场。全国平台能够汇聚各地各部门的数据资源，提升公共数据价值，激发市场主体对公共数据的需求。需求量增加，单次授权价格就可以适当减少，企业进入市场的门槛就会降低。在授权机制上，建议各地各部门在平台上公布高价值数据清单和字段。企业可以选择全量数据，也可以选择特定地区的部分数据或部分字段申请授权。同一份数据可以事前统一授权多个企业，发挥市场竞争的作用。绕过各地尺度不一的事前审查环节，改为全国平台的统一审查，也是一种减少事前监管的方式。事中和事后的监管，可以由企业所在地政府分别执行，落实属地和行业监管责任。全国平台相当于各方共建的“数据开放和授权运营渠道”，为此国家可以建立基于数据要素贡献的收益分配机制，调动各地各部门参与共建的积极性，中央在其中适当收取少量的“渠道费用”。同时，各地各部门仍可以保留自己建设的平台，继续进行政策层面的探索。全国平台也可以和其他部委的产业政策进行联动。例如，可以在平台上“揭榜挂帅”，吸引企业和高校等机构共同研发数据相关的技术；也可以通

过价格折扣等方式，对特定地区的公共数据授权运营进行政策倾斜。

(2) 在技术层面，可以利用开源框架，统一公共数据平台的技术标准。各地建设的公共数据平台，在数据格式、数据接口、授权协议、运维管理等方面并不互通。本文建议，国家数据局可以利用已有的成熟开源框架，自上而下统一数据平台的技术标准。国际上已有成功先例可供参考，例如：包括北美地区、英国、新加坡在内的30多个国家和地区，用“全面知识存档网络”(CKAN)^⑮或“基于Drupal的知识存档网络”(DKAN)^⑯等开源软件，创建网站并发布公共数据集；北美地区的政府还自主开发了一系列转换、处理和分析数据的开源工具，供各地各部门使用。从中央到地方都用同一套开源软件创建网站和发布数据，可以让国家利用统一的数据接口，将各地各部门更新的数据集及时汇总到全国平台。这有利于在未来对各地各部门开放数据的行为进行准确考核和评估。如果对安全有顾虑，政府也可以联合多方力量，自主开发开源框架；在开放平台的基础上，有条件地扩展支持授权运营功能，嵌入基于隐私增强技术的闭源模块，提升数据传输和交易过程的安全性。

(3) 在政府内部权责关系方面，建议在各地各部门设立数据专员岗位，落实公共数据开放的行政责任。“数据专员”应当由各部门领导班子中熟悉业务的重要成员担任，并且部门领导和成员都要经过专门培训，把握数据处理分析的理念和基本框架，配合国家数据局和党委分管领导开展工作。国家数据局在公共数据平台的基础上，统计汇总社会公众需求，对各地各部门的数据开放效果进行考核评估。数据开放的考核，也应当扩展到高校、图书馆、档案馆等国有企事业单位。

^⑮ CKAN是一个开源的数据管理系统，旨在帮助组织机构高效地存储、发布和分享数据集，广泛用于开放数据门户。

^⑯ DKAN是一个基于Drupal的开源数据管理平台，类似于CKAN，专注于为政府和非营利组织提供灵活的数据发布和可视化解决方案。

5 结语

训练数据是影响人工智能发展的关键要素。优质数据的供给与开放程度决定了人工智能模型性能的上限，也关系到人工智能行业的创新潜力和市场前景。只有通过技术、市场和政策的多方合力，才能促进训练数据的生产和流通，夯实我国在全球人工智能竞争中的优势，助力经济社会高质量发展。

受篇幅所限，本文尚未讨论隐私增强技术和数据跨境传输的问题。目前，以隐私计算、区块链、数据沙箱为基础构建的“可信数据空间”已成为保障数据安全的重要技术路径。然而，这一技术路线仍面临资源消耗大、传输效率低、实施成本高等瓶颈问题，且依赖统一的技术标准和跨行业协同机制，短期内难以大规模推广。在此类技术成熟之前，数据要素的监管政策应当保持灵活宽松，以及帮助企业降低数据相关的研发和应用成本，促进人工智能技术的创新和普及。

数据跨境传输也是值得关注的议题。随着我国企业国际化进程加速，跨境数据流通将成为支持人工智能等新质生产力“出海”的关键因素。根据DEPA^{②①}和CPTPP^{②②}等国际协定的要求，数据流通需要兼具开放性与安全性，这就意味着我国必须理顺国内政策体系，加快接轨和引领国际规则，不断增强数字经济的国际竞争力和话语权，为全球人工智能和数据治理贡献中国方案。

参考文献

- 1 Shumailov I, Shumaylov Z, Zhao Y, et al. AI models collapse when trained on recursively generated data. *Nature*, 2024,

631: 755-759.

- 2 Dohmatob E, Feng Y, Subramonian A, et al. Strong model collapse. *arXiv*, 2024, doi: 10.48550/arXiv.2410.04840.
- 3 Dong F, Li J. Asset bubbles in the data economy. *Rochester N Y*, 2024, doi: 10.2139/ssrn.4790765.
- 4 DeepSeek-AI, Guo D, Yang D, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv*, 2025, doi: 10.48550/arXiv.2501.12948.
- 5 齐英程. 我国个人信息匿名化规则的检视与替代选择. *环球法律评论*, 2021, (3): 52-66.
- 6 赵精武. 个人信息匿名化的理论基础与制度建构. *中外法学*, 2024, 36(2): 326-345.
- 7 许可. 复活僵尸法条: 个人信息匿名化制度的再造. *财经法学*, 2024, (4): 160-177.
- 8 Xu K. Reviving dormant legislation: Reconstructing the personal information anonymization system. *Financial Legal Studies*, 2024, (4): 160-177. (in Chinese)
- 9 Lee J H, You S J. Balancing privacy and accuracy: Exploring the impact of data anonymization on deep learning models in computer vision. *IEEE Access*, 2024, 12: 8346-8358.
- 10 Johnson G. Economic research on privacy regulation: Lessons from the GDPR and beyond. *National Bureau of Economic Research*, 2022, doi: 10.3386/w30705.
- 11 Demirer M, Jiménez Hernández D J, Li D, et al. Data, privacy laws and firm production: Evidence from the GDPR. *National Bureau of Economic Research*, 2024, doi: 10.3386/w32146.
- 12 马亮. 政商关系对数字政府的影响机制与理论进阶. *党政*

②①《数字经济伙伴关系协定》(DEPA)是一个旨在推动数字经济合作的国际协议,涉及数据流通、数字贸易和技术创新等领域。在数据流通方面,DEPA主张通过透明的规则促进跨境数据流动,同时确保数据隐私与安全。

②②《全面与进步跨太平洋伙伴关系协定》(CPTPP)是一个涵盖贸易、投资、知识产权和数字经济等领域的高标准自由贸易协定。在数据流通方面,CPTPP要求成员国允许跨境数据自由流动并限制数据本地化要求,与我国当前的监管政策存在一定冲突。

- 研究, 2022, (3): 107-117.
- Ma L. The influential mechanism and theoretical evolution of government-business relations on the construction of digital government. *Studies on Party and Government*, 2022, (3): 107-117. (in Chinese)
- 12 胡业飞. 责任配置、风险共担与激励相容: 中国地方公共数据授权运营的治理机制问题研究. *电子政务*, 2024, 262(10): 22-31.
- Hu Y F. Responsibility allocation, risk sharing, and incentive compatibility: A study on the governance mechanism of local public data authorization operations in China. *E-Government*, 2024, 262(10): 22-31. (in Chinese)
- 13 郭华东, 陈和生, 闫冬梅, 等. 加强开放数据基础设施建设, 推动开放科学发展. *中国科学院院刊*, 2023, 38(6): 806-817.
- Guo H D, Chen H S, Yan D M, et al. Strengthening open data infrastructure and promoting open science. *Bulletin of Chinese Academy of Sciences*, 2023, 38(6): 806-817. (in Chinese)
- 14 Talesh S A. Data breach, privacy, and cyber insurance: How insurance companies act as “compliance managers” for businesses. *Law & Social Inquiry*, 2018, 43(2): 417-440.

Study on constraints and policy responses for production and circulation of AI training data in China

LIN Tao

(1 The Institute for International Affairs, Qianhai, The Chinese University of Hong Kong, Shenzhen, Shenzhen 518172, China;

2 Department of Political Science, University of Washington, Seattle 98105, USA)

Abstract The quantity and quality of training data are critical to the performance of artificial intelligence (AI) models. However, in China, the production of training data is hindered by issues such as insufficient quantity, low quality, and fragmented distribution, compounded by limitations stemming from commercial ecosystems, regulatory frameworks, and restricted development and utilization of public data. To address these challenges, this study proposes several policy recommendations, including incentivizing research institutions to generate open-source datasets, fostering AI application scenarios, adopting a “loose-in, focus-out” regulatory approach, introducing intellectual property exemption provisions, refining personal information protection guidelines, and expediting the establishment of a unified national public data platform

Keywords artificial intelligence, training data, data element circulation, intellectual property, personal information protection, public data

林 韬 香港中文大学(深圳)前海国际事务研究院研究助理, 华盛顿大学政治学系博士候选人。主要研究领域: 政商关系、营商环境、产业政策、人工智能发展与治理等。E-mail: soxvlin@gmail.com

LIN Tao Research Associate in The Institute for International Affairs, Qianhai, The Chinese University of Hong Kong, Shenzhen, Ph. D. Candidate of political science at the University of Washington. His research focuses on state-business relations, business environment, industrial policies, AI development and governance, etc. E-mail: soxvlin@gmail.com

■ 责任编辑: 文彦杰