·科学论坛·

科学数据安全边界概念模型研究

——基于利益相关者视角

李宜展[1] 刘细文[1,2*] 李泽霞[1,2] 殷 茜[2] 吴 鸣[1,2]

1. 中国科学院 文献情报中心，北京 100190
2. 中国科学院大学 经济与管理学院图书情报与档案管理系，北京 100190

[摘 要] 明确科学数据安全边界的内涵和影响因素，是寻求数据共享与安全博弈平衡的必要条件。本研究在调研欧美科学数据安全相关法律政策，以及国际主流科学数据共享组织、典型科学数据基础设施的科学数据管理举措的基础上，运用规范分析法，归纳科学数据安全边界类型，基于利益相关者视角提出科学数据安全边界概念模型。该模型呈 3 层嵌套结构，以科学数据自身的安全为核心，依次从国家安全与数据主权、社会公共利益、商业利益、个人隐私与伦理 4 个影响因素，学科领域差异、数据规模、数据精度和推理演绎能力 4 个分析维度，分析科学数据安全边界，并从科学数据管理流程角度探讨保障数据安全的举措，以期为利益相关者界定科学数据安全边界提供分析框架，为制定科学数据安全共享相关政策提供有益参考。

[关键词] 科学数据；数据安全；边界；概念模型；利益相关者

伴随着信息技术的发展和数字化进程的推进，数据已成为国家基础性战略资源[1]。科学数据爆发增长、海量集聚，从科研活动的产物逐渐转变为开展科研活动的基础，地球科学、生命科学、材料科学、计算机科学等诸多学科领域呈现显著的数据密集型知识发现特征，个人敏感数据、健康数据、地理空间数据、生态资源数据、农业数据等以史无前例的规模被复用和关联[2-5]。科学数据开放共享是科研第四范式下驱动科技创新的通行做法，同时也为国家安全、个人隐私、知识产权、科研伦理等带来新的问题和挑战，引发学术界乃至全社会的广泛关注。

保障科学数据安全，是开放共享的基础和前提。在“边开放边保护”的呼声中，国务院办公厅于 2018 年印发《科学数据管理办法》，明确我国科学数据管理的原则以及保密与安全等内容。2021 年 6 月 10 日，我国颁布《中华人民共和国数据安全法》，这是第一部有关数据安全的专门法律。但由于科学数据本身具备共享性、非排他性、不对称性、可传递性、长期积累性、公益性等特点，国家行政机关、资助机构、科研院所与高校、研究人员、数据贡献者等在数据共享与价值链中所处的角色不同，数据保护和共享的诉求也不尽相同，从而使科学数据安全边界呈现出一定的复杂性和动态性。目前尚未对科学数据安全边界的理解达成共识，围绕边界界定的系统性、综合性

刘细文 中国科学院文献情报中心主任，研究员，博士生导师，中国科学院大学特聘岗位教授。《智库理论与实践》《科学观察》主编。长期从事科技政策情报研究与服务、科技战略情报研究，科学计量、信息政策、竞争情报和技术竞争情报研究、区域经济发展等。主持多项国家级科技政策与战略研究项目和课题。

李宜展 中国科学院文献情报中心副研究员，光电空间与重大科技基础设施团队情报分析师。人工智能学会会员。近期从事科学数据安全、重大科技基础设施战略情报研究。主持/参加 NSTL 面向政府部门的科技管理决策咨询项目、国家重大科技基础设施未来发展战略研究等科研项目。发表学术论文 30 余篇。

收稿日期：2020-12-22；修回日期：2021-06-29

* 通信作者：liuxw@mail.las.ac.cn

本文受到 NSTL 面向政府部门的科技管理决策咨询项目（2020XM44-2）资助。

讨论尚不充分。因此，深入探讨明确科学数据安全边界的内涵和影响因素具有重要意义。

近年来，个人隐私数据泄露事件频发，各国在数据跨境流动问题上显现原则性分歧，新兴技术、颠覆性技术异化应用带来多维风险与挑战，科学数据安全问题日益凸显[6]，保障科学数据安全的紧迫性加剧。虽然当前并未发现直接蓄意攻击公益型科学数据基础设施、数据库的案例报道，但因科学数据外流带来的数据主权丧失[7]、对国外科学数据库过度依赖等现象，已为科学数据安全，乃至科技安全、国家安全埋下隐患。2020 年 9 月，我国在“抓住数字机遇，共谋合作发展”国际研讨会上提出《全球数据安全倡议》，呼吁各国秉持发展和安全并重的原则，平衡处理技术进步、经济发展与保护国家安全和社会公共利益的关系[8]，为制定数字安全全球规则提供蓝本。

本文在广泛调研欧美及我国现行的与科学数据安全问题相关的法律法规、战略规划，国家级科学数据基础设施与典型科学数据库管理举措的基础上，运用规范分析法，界定科学数据安全的内涵，归纳科学数据安全边界类型，提出科学数据安全边界分析概念模型，以科学数据机密性、完整性和可用性为核心，从国家安全、社会公共利益、商业利益、个人隐私与伦理 4 个影响因素，数据规模、数据精度、学科领域特征、推理演绎能力 4 个分析角度，以及主流科学数据安全管理举措等层次，构建科学数据安全边界分析框架，以期为科学数据共享和安全相关政策制定提供有益参考。

1 科学数据安全边界

1.1 科学数据安全的内涵

科学数据是在自然科学、工程技术科学等领域，通过基础研究、应用研究、试验开发产生的数据及通过观测监测、考察调查、检验检测等方式取得并可用于科学研究活动的原始数据及其衍生数据[9,10]。李善青等将科学数据安全定义为通过必要的技术和管理措施，保护科学数据在其全生命周期中免受破坏性外力和非授权操作的侵害，保持科学数据的机密性、完整性和可用性，即主要从科学数据管理的角度阐述科学数据本身的安全[11]。

而科学数据在法律的层面，涉及隐私等人身权，有智慧产品等知识产权属性，又有战略、安全等国家主权属性[12]。科学数据获取、处理、存储、流通、使用技术创新层出不穷，应用场景日益丰富，对科学数据安全的关注也从科学数据本身的安全扩展到各利益相关者与应用场景的范畴。笔者认为科学数据安全包括两层含义：一是科学数据本身的安全，二是科学数据所承载的利益相关者的权益与利益安全。我国《数据安全法》明确保护个人、组织的合法权益，维护国家主权、安全和发展利益，推动国家有关部门、行业组织、科研机构、企业、个人共同参与数据安全保护工作，它们均为数据安全的利益相关者。鉴于现有研究对科学数据本身安全进行深入探讨[11,13]，本文将重点从利益相关者的角度切入分析科学数据安全边界问题。

调研和梳理公开发布的欧美及我国科学数据安全问题相关法律法规、战略规划，发现科学数据安全与数据安全、信息安全[14,15]、网络安全[16-18]、个人隐私安全[19]、科学数据管理[20,21]、数据主权[22]等概念有不同程度和范围的重叠。科学数据本身的安全主要包括机密性（Confidentiality）、完整性（Integrity）和可用性（Availability），即 CIA 三要素；其所承载的利益相关者的安全可归纳为国家安全、公共利益、商业秘密和个人隐私四个方面。同时，法律条例的不断更新也体现出科学数据的安全与价值受到社会环境、技术水平、数据类型和规模等影响，呈现敏感性、时效性和动态性。

1.2 科学数据安全边界的类型

科学数据分级分类是科学数据安全管理的重要手段，主要根据敏感性水平和泄露、丢失或滥用的风险，明确安全控制级别的起点和基线。代表国家利益的公共部门制定的数据分类方案多采用 3 级模式，如美国国家安全信息分类[23]、英国数据分类方案[24]和我国科学数据分类[25]等。研究机构、高校以及科学数据中心等在此基础上，结合其管理的数据资源情况，考虑机构、研究者、数据主体利益，进一步细化分类级别，并利用数据标签构建具有操作性的分级体系和工作规则。哈佛大学将研究数据分为公开的研究数据、未公开的非敏感研究数据、一般敏感数据、非常敏感数据、高度敏感数据 5 级[26]。

依据各国现行法律法规中数据安全相关规定以及现有的数据分类分级体系，可将科学数据安全边界分为三类：一是涉及国家安全、数据主权、个人隐私的保密数据，即“硬边界”。这是由国家/地区法律法规、行业领域保密规定明确界定。例如，欧盟《一般数据保护条例》（General Data Protection Regulation，GDPR）明确指出应禁止处理揭示种族或民族背景、政治观念、宗教或哲学信仰，或工会成

员的个人数据、基因数据、以唯一识别自然人为目的的生物特征数据，以及和自然人健康、性生活或性取向相关的数据[27]。

二是定向或有条件开放共享的情况，即“软边界”。往往通过规定科学数据开放共享的范围、用途、时间等权限，精度、规模等特征要素，传输和使用的网络与物理环境等形式，确保科学数据安全共享。例如，高位置精度地形图因与国家安全利益息息相关受到严密保护；而小比例尺地形图可公开且广泛地用于教育、规划管理、建筑施工、文化宣传等。去标识化的人类基因组数据可在一定条件下服务于科学研究活动，而大规模基因数据泄露则可能危及国家生物安全。跨境河流的流域国通过签订国际协议定向交换跨境河流汛期边境水文资料[28]等。

三是尚有争议或随着科技水平提高而不断发展变化的灰色地带。鉴于核科学在维护国家主权方面的重要地位，历史上公开知识体系隐去核科学研究[29]，核技术数据受到严格管控。随着技术水平的提升和社会发展需求的变化，美国《原子能法》放宽对核技术的控制，允许进行商业化和平利用[30]，其相关的科学数据安全边界也随之改变。

可见不同利益相关者视角和场景下，科学数据安全边界难以一概而论，建立分析科学数据安全边界的概念模型，明确边界影响因素和分析维度是非常有必要的。

2　科学数据安全边界概念模型

科学数据安全的内涵决定了安全边界应以保障科学数据本身的可用性、机密性和完整性为核心，以维护科学数据所承载的国家安全与数据主权、社会公共利益、商业利益、个人隐私与伦理不受侵害为目标。但不同利益相关者对上述四个因素的关注度不同，根据各自诉求定制共享与安全之间的界限，除了以敏感性和风险为原则依据外，需要根据具体情况分析：(1) 科学数据安全会对哪些目标利益产生威胁；(2) 具体是哪些科学数据关键特征对敏感性水平和风险产生影响，进而影响到上述目标利益；(3) 可以采用何种措施管理和影响科学数据的关键特征，进而降低敏感性水平、规避风险。基于以上分析逻辑，本文提出科学数据安全边界概念模型。

该模型以科学数据本身的 CIA 三要素为核心，整体呈 3 层嵌套结构(见图 1 所示)。第 1 层，将目标利益定义为科学数据安全边界的主要影响因素，包括国家安全与数据主权、社会公共利益、商业利益、个人隐私与伦理；第 2 层，将科学数据的关键特征定义为边界分析维度，包括学科领域差异、数据规模、数据精度、推理演绎能力，它们是通过归纳总结现行法律规定中共享要求与例外情况、数据安全等级分类细则等得出的；第 3 层，以科学数据安全管理流程为主线，列举安全保障举措以及当前社会与科技环境下尚有争议重点问题，即可以管控或影响数据关键特征的措施。模型将主要影响因素、分析维度和安全管理举措统一到一个框架下，各圈层相互影响形成有机整体，为利益相关者分析和制定安全边界提供分析框架。

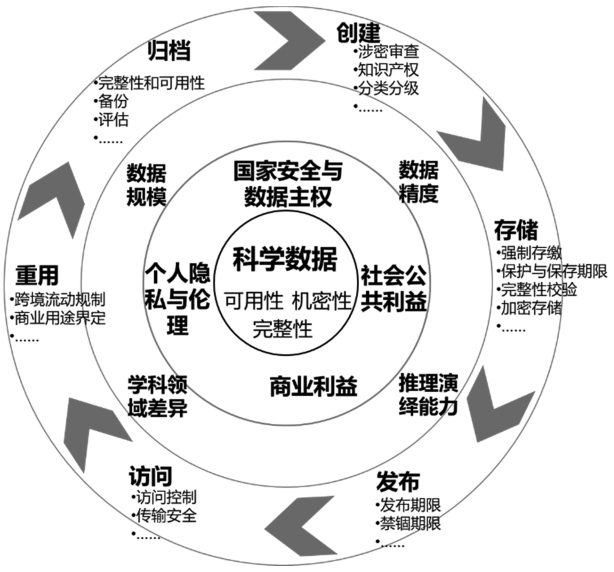

图 1　科学数据安全边界概念模型

2.1　影响科学数据安全边界的主要因素

2.1.1　国家安全与数据主权

科学数据承载科技创新、国家安全、人民生命健康、生态环境保护等领域的重要信息，已渗透到国家和社会生活的各个方面，成为支撑社会经济运行和发展的基础性战略资源，具有了国家主权属性[31]。

地理空间数据是一类战略性信息资源，体现和表明政府立场与主张的海陆疆域，其完整性、机密性本身就涉及国家主权和领土完整。现代战争对高精度地理信息的依赖度不断上升，一旦掌握国家高精度地理空间数据，就可以使用远程武器精确快速打击国家重要军事设施、安全警卫目标，损害军事防御能力，其对国防安全的重要意义不言自明。2020 年 6 月，我国自然资源部、国家保密局印发《测绘地理信息管理工作国家秘密范围的规定》和《测绘地理信息管理工作国家秘密目录》[32]，明确规定高精度地形图、数字化成果、遥感影像，坐标系转换参数，保密

处理技术算法与参数等的保密等级。全球气候变暖导致冰冻圈加速融化释放极地地区资源、航道和军事战略价值[33]，人类对太空资源的探索利用能力也随着航天技术的发展不断提升，极地科考、太空探索研究中涉及国家秘密的样品和数据的保护、公开和利用，执行国家保密相关管理规定。此外，部分军民两用科技数据，如核科学数据，与大规模杀伤性武器、群体杀伤技术、生物恐怖主义相关的反恐科技数据等，也与国家安全事项密切相关[34]。

在大数据时代，国家实力呈现数字构成形态[35]，传统的国家主权内涵不断拓展，管辖范围从物理空间逐渐发展到网络空间，再到数据空间、技术空间。虚拟空间的数据主权同样是国家主权的体现。在当前治理体系和技术水平下，可将国家数据主权概括为大数据、云计算背景下国家对本国数据以及本国国民的跨境数据拥有所有权、控制权、管辖权和使用权，体现在对内的最高数据管控权和对外平等、独立的数据处理权[36]。对其中的研究型、资源型和参考型科学数据管理不当，可能带来国家显性或隐性科技水平和现状、发展趋势与规律等关键信息泄露。

全球化背景下科学数据在网络中自由流动，不断向具有信息基础设施和技术先发优势的国家汇集。有些国家还通过建立完善的法律制度、前瞻布局战略发展方向不断巩固领先地位，发挥跨国企业经营模式优势广泛收集科技数据，在“虹吸效应”下形成难以弥合的“数字鸿沟”。他国广泛、大规模、持续获取由本国财政资助产出的基础数据，致使科学数据外流、数据主权丧失，长此以往会在日益激烈的全球科技创新竞争中失去有利地位、甚至处于被动状态[7]。若不从总体安全观的角度看待科学数据安全治理问题，不加管控地追求虚拟空间中的数据自由流动，忽视科技发展水平、数据与网络基础设施建设以及经济实力极不平衡等现实情况，将对维护国家政治、经济、文化主权安全带来极大的隐患[37,38]。当前，俄罗斯强制信息数据本地化保存，澳大利亚禁止个人医疗信息出境[39]，以实现将虚拟的数据限制在明确的领土范围内，从而将其纳入物理世界现实疆域的国家治理范畴，这也是强烈控制数据主权的表现。

2.1.2 社会公共利益

社会公共利益直接关系到社会全体，代表广大人民的意志。它可以是文化、教育方面的利益，也可以是生命、健康和自由的利益，还可以涉及整个社会的经济运行秩序，具有涉及内容多样性、不确定性和多层次性等特点。因此，在科学数据安全危及社会公共利益的问题中，需要根据具体情况，通过适当途径和程序确定公共利益的内容。例如，敏感的化学、医学数据可能会成为恐怖分子、恶意攻击者的武器，其泄露会造成灾难性后果。

在应对公共卫生与安全应急事件中，个人隐私与社会公共利益在一定程度上存在冲突。如疫情期间，以维护公共健康利益为目标，科学获取精准人群动态信息与时空过程，在分析疾病传播过程、有效预防和防控疫情中发挥重要作用。但实施数据密集型公共卫生监控，极易出现因患者或密切接触人员的个人身份信息、行程信息泄露引发的歧视、攻击等事件。对于涉及公共卫生事件所采集、发布和处理的个人敏感信息，通常会在法律法规中列为个人信息保护使用限制事项，即使突发公共卫生事件联防联控期间，也不能随意获取和使用个人信息，而应当尊重个人隐私并进行有效且适度的保护，平衡个人信息保护和公共卫生管理之间的关系[40]。

2.1.3 商业利益

具有明确应用前景的高价值研究数据，在加速战略性新兴产业发展中扮演重要角色，进而也成为塑造经济社会发展重要驱动力和国际竞争力的重要因素。以自动驾驶技术为例，自动驾驶汽车收集车内影像、录音以及车外沿途地物影像、道路状况，记录驾驶员操作习惯等，定时或实时上传至境外服务器。一方面，沿途连续摄取的影像可能会包含国家重要军事设施、与自动驾驶车辆交汇的军用车辆信息；另一方面，这些数据可进一步组织加工，转化为高维、互补、连贯、规模化的商用技术数据，成为提升自动驾驶算法和驾驶辅助系统的重要“原料”，升级后的软硬件也会进一步提升车辆驾乘体验和安全性能。以特斯拉为代表的汽车制造厂商大量收集和积累路测数据，并借此逐步形成对行业数据和人工智能、机器学习技术整合运用的强大竞争力，从而在产业竞争和创新中占据绝对优势。与此同时，这些科技数据有可能在跨国企业/组织内部及商业合作伙伴之间实现一定程度的共享，从而加剧风险。

需要特别指出的是，科学数据共享和安全的边界并不与公共研究和私立研究之间的界限完全重合，即民营企业对数据保密，而公共研究开放数据。部分公共研究的具有潜在商业价值的数据需要适度开放，一些商业模式也因数据开放而繁荣。政府在大多数情况下鼓励资助者对研究数据进行商业利

用，进一步丰富现有数据集，不一定要求资助者主张数据所有权，或通过控制而非限制访问的手段主张所有权，如通过知识产权保护，在需要时服务于国家主权和社会公共利益。对于私人资助的研究所产生的数据，在可能影响公共利益时，可在知识产权得到保护或向公众提供产品或服务后公开。当研究涉及特定的紧急安全问题时，公共利益应优先于即时商业利益。在许多领域中，监管机构肩负着代表公众进行监督和决策以保障国家、社会和个人隐私权益不受侵害的职责，如药品和保健品监管局、临床试验注册机构等[41]。

2.1.4　个人隐私与伦理

个人信息对医学、社会科学领域的许多研究至关重要，其可能为个人隐私、道德伦理带来风险，因此对数据安全提出更高的要求。“人类材料”和“人类数据”已纳入现行个人数据保护法律界定的“个人数据”的范畴[42]，包含已识别或可识别活着的自然人有关的科学数据，无论是一般数据还是敏感数据均已涉及个人隐私范畴。公民通过避免将个人数据用于可带来歧视、污名化、侵犯个人自主权的行为活动，保护隐私权不受恶意或无意侵犯。例如，在构建生命健康、临床医学数据库时，可能会发生个体遗传、疾病、行为信息的泄露，威胁个体联系方式、生理状况、宗教信仰、基因与生物特征数据等现实世界或社交空间个人身份特征隐私。神经科学的发展和数据分析能力不断提高，基于神经科学、认知科学数据推断的思想探知能力也将有可能暴露“思想隐私”[43]。

在个人隐私保护的背景下，以审慎态度对待科学数据的开放，会对数据安全边界产生直接影响。一方面，数据生产者在科学数据创建时应严格遵守科学研究“自愿参与”和“知情同意”的基本伦理道德准则，发布或传播涉及人类参与者的研究数据之前，应确保同意、保密、隐私、匿名、安全等伦理考虑[44]，确保已获得相关管理机构的许可。另一方面，数据控制者通过数据加密、审查和限制数据使用者的资质与访问权限等方式，严格控制知悉范围，保证涉及个人隐私的科学数据全生命周期的安全，降低披露个人隐私的风险。

2.2　界定科学数据安全边界的分析维度

2.2.1　学科领域差异

因数据的组织结构、应用场景等不同，科学数据安全边界呈现学科领域差异。在生命健康领域，涉及人类研究的遗传信息，大型人群队列数据，高传染性、致病性微生物研究数据，具有应用于微观军事攻击的可能性，甚至可用于研制针对特定人群、种族或人种的选择性精准基因组武器。以指纹、虹膜、DNA、人脸为代表的生物识别数据一旦丢失便无法撤销或重置，导致个人生物特征数据无法使用。

在地球科学领域，地球中高层大气环境、电离层环境、磁层环境的地球空间环境监测和预报数据，油气、稀有金属等矿产资源数据等，在诸多领域被视为决策资源，具有明显或潜在国家安全应用和商业价值。国家地理空间数据对国家重要设施等物理空间范围敏感，个人地理空间数据可通过地理编码数据库实现个体追踪和定位，也会对个人隐私构成威胁。

在材料科学领域，特种用途结构材料（如铝、镁、钛、钨合金、金属间化合物以及复合材料）、功能材料（如隐身材料、阻尼减震材料、贮氢材料、光电功能材料等）是航空航天、兵工舰船等国防工业最重要的物质基础，其化学组成、微观结构及生产试验数据具备敏感性，它们的安全极可能触发、传导、扩散影响军事安全、资源安全、生态安全、社会安全。

然而，在涉及学科领域特征的科学数据共享和安全博弈中，必须考虑限制共享在科研第四范式背景下对科技创新带来的潜在不利影响。目前国际上的通行做法是待技术走向成熟且能够预见到未经授权披露所带来的损害时，将其列入管控范围。

2.2.2　数据规模

科学数据的敏感性可能体现在数据个体中，即通过个体数据获得细节与实质性敏感内容，也有可能在数据积累到一定程度后才显现其价值。在科研大数据背景下，科学数据通过主动或被动汇集，形成具有新的组织结构的科学数据集合，数据形态和结构发生变化，从而具备单一的科学数据个体本身并不具备的价值和功能，即“规模致敏”。若未对公开发布的科学数据、项目成果的规模加以审查和限制，致使非授权利益相关方通过“切香肠”策略渐进式地获得部分数据、直至掌握全数据集，可对社会公共利益带来潜在威胁，甚至危及国家安全。

例如，美国国立卫生研究院的《基因组数据共享政策》定义了“大规模”基因组数据，对 1 000 个以上受试者的超过 1 个基因（或基因组内规模相当区域的序列数据、超过 1 000 个参与者的不低于 30 万个突变位点数据等大规模基因组数据的提交、发布、获取进行分级控制，约束生物安全风险。我国将构成环线且覆盖范围大于 2 500 平方千米或线路长度超过 1 000 千米的国家等级水准网成果列入《测绘地理

信息管理工作国家秘密目录》,以“秘密”等级长期保存[32]。我国北斗卫星导航系统单个基准站的数据是敏感数据,而多个基准站的数据汇聚后则上升为“秘密”等级[45]。

发达国家通过间接手段调控生物、材料信息的国际流动,如在国际主流期刊发表论文需要提交与论文配合的生物数据,并将其汇集到欧美生物信息等国际专业数据库中。在数据密集型科研范式下,这些规模化数据成为未来科技创新不可或缺的重要资源。未依规将科学数据汇交备份到国内的科学数据中心,不但会对数据主权产生影响,还造成过度依赖国外科学数据库的后果。事实证明,虽然每一项数据都是可共享的,但当美国政府停摆致使 NCBI 数据库暂停服务,仍对我国生命科学研究进程带来极大影响,体现了规模化科学数据极其重要的战略地位。

2.2.3 *数据精度*

数据精度指测量所得值之间的一致程度以及与其“真值”的接近程度,是精密度和准确度的综合,是表征数据详实程度、质量的重要指标。“精度致敏”最典型的例子是地理空间数据,地形图的比例尺、遥感卫星数据的空间分辨率、重力异常数据的精度等都是数据精准详实程度的体现,直接影响科学数据的敏感程度和风险水平。观测设备、技术方法的革新不断提高大范围、快速获得高精度数据的能力,甚高分辨率遥感卫星影像、全球定位系统(GPS)、地理信息系统等对地观测技术详细精准地描绘地物状态、地表过程,同时也极大地提高辨识特定目标的能力,在一定程度上对国家重要设施和活动的保密性带来影响。美国曾为保护国家安全与利益,利用降低广播星历精度和在 GPS 信号中加入随机抖动等方法,人为降低普通用户获得的定位数据精度。谷歌影像提供大量民用遥感影像数据,虽不具备时效性,也难以利用其获得精确的绝对位置信息,但在影像、照片、标注信息的日益积累,多维度信息相互参照关联的情况下,许多国家纷纷要求对敏感区域、建筑设施进行模糊化处理,加强敏感设施伪装工作,以降低数据的可解析程度。

此外,个人信息包含的身份证号码、年龄、居住地、邮编、电话号码、工作单位等,因可直接或间接识别原始个人信息主体而具备敏感性。为保障受访者的隐私权,通常会采取删除、聚合或降低标识符精度的方式进行数据去标识化。

2.2.4 *推理演绎能力*

科学数据的推理演绎主要包括挖掘、集成和融合三种方式。通过自然语言处理、机器学习、深度学习等数据挖掘技术,准确关联并提取隐含在科学数据、元数据中的信息,创新科学数据组织,从而拓宽科学数据应用的深度、广度,丰富应用场景。在科学数据汇集形成新的科学数据集合时,虽然有些科学数据本身不具有敏感性,但“安全”数据个体可能与其他“安全”数据关联或互补,使其敏感性发生变化。通过一定的技术手段融合跨领域数据,形成围绕关键内容的数据/信息/知识谱系,也可能推理还原出具有指向性的关键内容。

科学数据安全风险的变化取决于科学数据与其他数据关联并还原关键事实的可能性及难易程度,以及数据关联后对关键事实的指向性[46]。科学数据间关联的难度、范围、有效性、程度、强度、指向性以及互补的程度、重要性、完整度、新颖性均对科学数据安全边界产生影响。而关乎不同目标利益的还原性和指向性同样存在差异。在关联组合时,企业比个人更容易识别,例如一个罕见的行业、企业规模或地点可能会暴露数据所涉及的公司[47]。

从总体来看,推理演绎能力与数据挖掘、规则模型构建、人工智能、知识关联与发现[48]等关键技术水平密切相关,在尚未达到必须的能力之前,甚至无法识别科学数据是否存在安全风险,亦无从感知和判断风险程度及可能带来的危害。对这些可能在特定条件发生敏感性变化的科学数据,有必要明确核心数据范围和弹性空间,定期评估和预测一段时间内技术发展带来的数据敏感性和风险扩张的可能性。

2.3 模型要素之间的关系与互动

模型各圈层由内而外分别描述了科学数据安全核心原则、目标利益、影响敏感性水平和风险的关键特征、调整或控制关键特征的举措,逐层逐步拆解科学数据安全问题。从构建模型的逻辑来说,各圈层之间通过触发、传导等互动形成统一的有机整体。

首先,四个影响因素之间存在触发与相互传导的行为。触发与传导是指当关键特征或目标利益达到某一临界点后,引发其他模型要素的变化与连锁反应。以个人隐私风险触发商业利益风险、社会公共利益风险为例说明。深度伪造技术的快速发展降低了虚假人像、视频、音频制造的时间成本和技术门槛,在此技术环境下,泄露核心人员的生物特征识别数据(如指纹、人脸、声音),可导致未授权访问企业内部具有经济价值的技术数据(如产品配方参数)与经营信息(企业战略规划等),进而触发商业利益风

险。大规模个人隐私信息暴露会引发广泛且连锁的社会安全信任问题，将个人隐私风险传递到社会公共利益层级。

更进一步地，从四个分析维度看上述过程，深度伪造技术能力的提高和应用场景谬用成为触发各利益风险的“临界点”，是信息技术提升数据演绎水平、进而危及多利益相关方的体现；生物特征识别数据一旦泄露无法更改的特殊性，加剧了个人隐私信息的敏感性与补救难度，同时也要求更加审慎地管控生命科学数据；数据规模则是由个人隐私风险向社会公共安全与信任危机传导的关键要素。

从管理的角度分析，国家安全、公共利益、商业利益、个人隐私直接影响数据安全管理参与主体、立法立规方式。具体的约束条款，尤其是针对学科领域特点显著的科学数据的条款，体现了从规模、精度、推理演绎水平维度审视和规制各影响因素的敏感性与风险。强制汇交和存缴的背后是对大规模数据外流和过度依赖国外数据库的担忧与反制；数据分类分级管理给予不同数据精度不同的等级，并匹配相应的数据安全属性、访问条件；定期开展数据安全与共享规则评估、调整数据发布与禁锢期限等，是防止新兴与颠覆性技术加速推理演绎水平变化，致使蓄意或意外破坏数据机密性的重要手段。

3　基于数据安全边界的管理实践

在科学数据创建、存储、发布、访问、重用、归档的流程中，采取一系列管理举措保障科学数据安全。在科学数据汇交创建阶段进行涉密审查，一方面是针对研究内容或科学数据本身的机密性进行审查，同时也需要定期审视新兴技术、颠覆性技术对科学数据安全带来的挑战，分析在现有推理演绎技术能力下，数据纳入数据集/库所带来的汇集关联、挖掘、融合风险。审查评估后，对科学数据进行分级分类，制定有针对性地共享规则。在存储阶段，我国《科学数据管理办法》要求强制汇交国家科技计划项目产生的数据，以应对科学数据流失严重所带来的数据主权与安全问题。在发布阶段，根据科学数据的学科领域特点，科学数据中心有针对性地、灵活地限制最新数据、历史数据的公开时间。在访问阶段，哈佛大学科学数据管理平台 Dataverse 根据数据敏感性等级，通过自主访问控制、强制访问控制、基于角色的访问控制技术、加密数据传输等，保证科学数据受控、合法地被访问、传输和使用[26]。在重用阶段，当前商业科技数据的跨境流动引发广泛关注，虽然从美国的《澄清合法使用境外数据法》《出口管理条例》，到欧盟的《约束性公司规则》《一般数据保护条例》均对数据跨境获取与使用行为进行约束，但国际上对于数据主权的管控范围、管理制度尚未达成共识。在归档与长期保存阶段，NASA 的《数据保留白皮书》明确科学数据应无限期保留、发布的条件和要求[49]，国际地球科学信息网络中心识别和评估长期保存的候选数据，并进行覆盖全生命周期的监控[50]，以确保数字资源的质量、完整性、机密性和安全性不因时间推移受到损害。与此同时，科学数据管理机构设立数据科学保护委员会、数据保护官员，推动和监督以上举措落地实施。

当前，我国《中华人民共和国数据安全法》和《中华人民共和国个人信息保护法》分别从保护国家安全和公共安全、保护公民隐私角度开展数据安全治理，与《网络安全法》《保密法》等共同构建我国数据安全法律框架。科学数据除严格遵循上述法律外，其规范化处理与安全共享还需要：(1) 进一步明确各学科领域、数据规模、数据精度、推理演绎能力以及各影响因素联动下的具体界限，寻找既能保障我国科学数据安全，又可以与国际科学数据共享规则接轨的解决方案；(2) 开展系统性实证研究，实现在总体国家安全观下，适应科学数据特点、兼顾技术发展水平和社会发展需求的科学数据安全治理。

参　考　文　献

[1] 中华人民共和国中央人民政府. 促进大数据发展行动纲要. (2015-09-05)/[2021-03-22]. http://www.gov.cn/xinwen/2015-09/05/content_2925284.htm.

[2] Mostert M, Bredenoord AL, Biesaart MCIH, et al. Big Data in medical research and EU data protection law: challenges to the consent or anonymise approach. European Journal of Human Genetics, 2016, 24(7): 956—960.

[3] 刘瑞爽. GDPR 对我国医学研究伦理审查的启示. 医学与哲学, 2019, 40(3): 29—33.

[4] 吴定峰, 刘婷婷, 王剑, 等. 动态信任管理模型在农业科学数据安全领域的应用探索. 农业图书情报学报, 2020(10): 16—24.

[5] Ferrag Ma, Shu L, Yang X, et al. Security and privacy for green IoT-based agriculture: review, blockchain solutions, and challenges. IEEE Access, 2020(8): 32031—32053.

[6] 苗争鸣, 尹西明, 许展玮, 等. 颠覆性技术异化及其治理研究——以“深度伪造”技术的典型化事实为例. 科学学与科学技术管理, 2020, 41(12): 83—98.

[7] 李洋, 温亮明. 我国科学数据外流：表现、问题与对策. 图书馆杂志, 2019, 38(12): 72—81, 115.

[8] 中华人民共和国中央人民政府. 全球数据安全倡议(全文). (2020-09-08)/[2020-09-30]. http://www.gov.cn/xinwen/2020-09/08/content_5541579.htm.

[9] 中华人民共和国中央人民政府. 国务院办公厅印发《科学数据管理办法》. (2018-04-02)/[2019-12-04]. http://www.gov.cn/home/2018-04/02/content_5279296.htm.

[10] 诸云强, 朱琦, 冯卓, 等. 科学大数据开放共享机制研究及其对环境信息共享的启示. 中国环境管理, 2015, 7(6): 38—45.

[11] 李善青, 郑彦宁, 邢晓昭, 等. 科学数据共享的安全管理问题研究. 中国科技资源导刊, 2019, 51(3): 11—17.

[12] 韩伟. 安全与自由的平衡——数据安全立法宗旨探析. 科技与法律, 2019(6): 41—48, 67.

[13] 盛小平, 田婧, 向桂林. 科学数据开放共享中的数据质量治理研究. 图书情报工作, 2020, 64(22): 11—24.

[14] National Institute of Standards and Technology. Guide for mapping types of information and information systems to security categories (SP 800-60 Vol. 1 Rev. 1). (2008-08-01)/[2021-03-21]. https://csrc.nist.gov/publications/search?keywords-lg = 800&sortBy-lg = relevance&viewMode-lg = brief&ipp-lg = 25&series-lg = SP&topicsMatch-lg = ANY&controlsMatch-lg=ANY.

[15] The European Parliament and of the Council. Directive (EU) 2016/1148 of the European Parliament and of the Council of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union. (2016-07-06)/[2020-05-06]. https://eur-lex.europa.eu/legal-content/EN/TXT/? uri=uriserv:OJ.L_.2016.194.01.0001.01.ENG&toc=OJ:L:2016:194:TOC.

[16] The White House. International Strategy for Cyberspace. (2011-05-16)/[2020-05-06]. https://obamawhitehouse.archives.gov/sites/default/files/rss_viewer/international_strategy_for_cyberspace.pdf.

[17] The White House. National Cyber Strategy. (2018-09-20)/[2020-05-06]. https://www.whitehouse.gov/wp-content/uploads/2018/09/National-Cyber-Strategy.pdf.

[18] The European Parliament and of the Council. Regulation (EU) 2019/881 of the European Parliament and of the Council of 17 April 2019 on ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity certification and repealing Regulation (EU) No 526/2013 (Cybersecurity Act) (Text with EEA relevance). (2019-04-17)/[2020-05-06]. https://eur-lex.europa.eu/eli/reg/2019/881/oj.

[19] The European Parliament and of the Council. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). (2016-04-27)/[2020-05-06]. https://eur-lex.europa.eu/legal-content/EN/TXT/? uri=uriserv:OJ.L_.2016.119.01.0001.01.ENG&toc=OJ:L:2016:119:TOC.

[20] Office of Management and Budget. Office of Science and Technology Policy, Department of Commerce, etc. Federal Data Strategy 2020 Action Plan. (2019-12-23)/[2020-05-06]. https://strategy.data.gov/assets/docs/2020-federal-data-strategy-action-plan.pdf.

[21] EU Commission. A European Strategy for data. (2020-02-19)/[2020-05-06]. https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en.

[22] US House Judiciary. CLOUD Act. (2018-02-06)/[2020-05-06]. https://www.congress.gov/bill/115th-congress/house-bill/4943.

[23] Information Security Oversight Office. The President Executive Order 13526 Classified National Security Information. (2009-12-29)/[2021-03-21]. https://www.archives.gov/isoo/policy-documents/cnsi-eo.html.

[24] UK Government. Government Security Classifications. (2018-05-21)/[2021-03-21]. https://www.gov.uk/government/publications/government-security-classifications.

[25] 杨晶, 康琪, 李哲. 推动科学数据开放共享的思考及启示. 全球科技经济瞭望, 2019, 34(10): 37—43.

[26] Harvard University. Handout-Research Data Security Levels with Example. (2020-04-22)/[2021-05-06]. https://security.harvard.edu/handout-research-data-security-levels-examples.

[27] The European Parliament and of the Council. Regulation (EU) 2016/679 of the European Parliament and of the Council of of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). (2018-05-23)/[2021-03-24]. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/? uri=CELEX:32016R0679.

[28] 郝少英. 我国国际河流水资源信息资料交流存在的问题及其法律对策. 清华法治论衡, 2015(1): 348—360.

[29] Galison P. Removing knowledge. Critical Inquiry, 2004, 31(1): 229—243.

[30] United States Environmental Protection Agency. Summary of the Atomic Energy Act. [2020-08-25]. https://www.epa.gov/laws-regulations/summary-atomic-energy-act.

[31] 韩伟. 安全与自由的平衡——数据安全立法宗旨探析. 科技与法律, 2019(6): 41—48, 67.

[32] 中华人民共和国自然资源部. 自然资源部 国家保密局关于印发《测绘地理信息管理工作国家秘密范围的规定》的通知. (2020-06-18)/[2020-09-25]. http://gi.mnr.gov.cn/202007/t20200707_2531433.html.

[33] 张胜军, 郑晓雯. 从国家主义到全球主义: 北极治理的理论焦点与实践路径探析. 国际论坛, 2019, 21(4): 3—18.

[34] 钟灿涛. 开放与保密: 科技信息传播控制及其对创新的影响——以美国科技信息传播控制机制为例. 科学学研究, 2013, 31(3): 335—343.

[35] 刘春呈. 大数据时代大数据局对数字边疆的治理研究. 四川行政学院学报, 2019(5): 14—24.

[36] 孙伟，朱启超. 正确区分网络主权与数据主权. 中国社会科学报，(2016-07-05)/[2021-03-21]. http://www.cssn.cn/zx/201607/t20160705_3098529.shtml.

[37] 刘天骄. 数据主权与长臂管辖的理论分野与实践冲突. 环球法律评论，2020，42(2)：180—192.

[38] 兰美荣. 数据主权安全观教育：新时代大学生爱国主义教育的重要议题. 思想教育研究，2020(7)：126—130.

[39] 吴沈括. 数据跨境流动与数据主权研究. 新疆师范大学学报(哲学社会科学版)，2016，37(5)：112—119.

[40] 朱军. 个人信息保护与社会公共利益关系如何平衡. 人民论坛，2020(18)：68—69.

[41] The Royal Society. Science as an open enterprise. (2012-01-21)/[2020-09-25]. https://royalsociety.org/-/media/policy/projects/sape/2012-06-20-saoe.pdf.

[42] 姜鑫，马海群，王德庄. 基于质性文本分析视角的开放科学数据与个人数据保护的政策协同研究——以国外资助机构为例. 情报理论与实践，2020，43(7)：54—62.

[43] 林玲，张新庆，黄小茹. 神经科学数据应用引发的隐私问题探讨. 中国医学伦理学，2020，33(3)：294—298，303.

[44] 姜鑫. 国外资助机构科学数据开放共享政策研究——基于NVivo 12 的政策文本分析. 现代情报，2020，40(8)：144—155.

[45] 王雍，张舒黎，石元兵，等. 北斗高精度数据传输安全研究. 通信技术，2020，53(9)：2245—2251.

[46] 唐超，钟灿涛. 数据汇集中的安全风险评估研究. 保密科学技术，2019(6)：8—15.

[47] Päällysaho S，Latvanen J，Lehto A，et al. Key Aspects of Open Data in Finnish RDI Cooperation between Higher Education and Businesses. Data Intelligence，2021，3(1)：176—188.

[48] 周毅，刘峥，张建勇. 关联数据研究的主题结构和研究进展解析. 农业图书情报，2019(3)：13—24.

[49] National Aeronautics and Space Administration. White paper on NASA science data retention. (2016-10-05)/[2020-05-06]. https://nssdc.gsfc.nasa.gov/nssdc/data_retention.html.

[50] Center for International Earth Science Information Network. CIESIN Policy for Preservation of Digital Resources. (2007-07-25)/[2020-05-06]. http://www.ciesin.org/documents/CIESINpreservationpolicy.pdf.

Study on Conceptual Analysis Model of Scientific Data Security Boundary: From the Perspective of Stakeholders

Li Yizhan[1]　Liu Xiwen[1,2*]　Li Zexia[1,2]　Yin Xi[2]　Wu Ming[1,2]

1. *National Science Library, Chinese Academy of Sciences, Beijing* 100190
2. *Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing* 100190

Abstract　Clarifying the conception of scientific data security boundaries and their affecting factors is necessary for pursuing a balance between data sharing and security. In this paper, three types of scientific data security boundary and a conceptual model are established based on normative analysis, investigating laws and regulations, strategic policies, and typical data policies of scientific data infrastructures announced by the United States and the European Union. With the safety of scientific data as its core, this model consists of three layers, including four affecting factors of national security, public interest, commercial secrets and personal privacy, as well as four analytical perspectives, the characteristics of subject areas, data scale, precision and deductive capability, to provide an analytical framework for defining the security boundaries of scientific data.

Keywords　scientific data; data security; boundary; conceptual model; stakeholders

（责任编辑　刘　敏）

* Corresponding Author, Email: liuxw@mail.las.ac.cn