DOI: 10.11817/j.issn. 1672-3104. 2024. 06. 004

论人工智能数据公共领域深度共享机制的构建

何炼红，朱曦青

(中南大学法学院，湖南长沙，410083)

摘要：公共领域数据共享是人工智能产业发展的关键。平台层面数据霸权导致的共享自由缺位、资源层面数据壁垒导致的公共领域限缩、规则层面权属不清导致的共享标准模糊，对人工智能数据公共领域产生了系列负效应。人工智能数据应践行“深度共享”理念，基于平台、资源和规则三个维度，实现共享平台从单层封闭式枢纽走向多层开放式网络、资源共享方式从单一走向多维、共享规则从保守走向开放的范式转型。构建人工智能数据的公共领域深度共享机制，应以“开源”为前提，引入数据共享“FAIR 原则”，采取公共领域双层共享模式，在平台层面构建角色模块式数据共享生态以实现数据共享自由，在资源层面设置数据公共领域官方标识以明确公共领域范围，在规则层面完善数据开源协议以确立数据共享标准，最终实现人工智能公共领域数据资源充裕和繁荣发展。

关键词：人工智能数据；公共领域；深度共享；公共标识；开源协议

中图分类号：D923.39 **文献标识码：**A **文章编号：**1672-3104(2024)06−0033−16

引言

人工智能数据通常是指人工智能系统在输入、处理和输出环节所涵盖的输入数据、训练数据以及输出数据。当人工智能训练目的与生成目的彼此兼容时，人工智能的生成数据甚至可提供给其他人工智能进行数据喂养，并形成合成数据(synthetic data)①。智能时代，个人、企业或公共数据的传统数据分类已无法匹配当下新技术的发展需求，人工智能数据的人格性、财产性和公共性等属性相互融合，已经难以对人工智能数据基于属性和目的作出单一分类，故理论和实践层面往往围绕数据应用流程对人工智能数据的范畴进行界定②。

当下，面对数字化与全球化的双重塑造，人工智能技术生发出超强能量，数据共享成为人工智能技术发展应用的关键环节。ChatGPT 的出现代表生成式人工智能(generative artificial intelligence)实现了技术发展到实际应用的新范式转变，而生成式人工智能的多任务执行能力(又称涌现能力，emergent abilities)必须有丰富的数据源基础才能实现[1]。相较于通用人工智能，生成式人工智能亟须更多高价值、多样性和特征丰富的人工智能数据予以优化升级，这对人工智能数据在公共领域中的数字化共享提出了更高的要求。美国、欧盟和日本等国家和地区已经开始制定以通用人工智能为内容的第四次工业革命发展战略，并将人工智能数据共享视为关键推力，以促进生成式人工智能的落地应用。例如，美国《国家人工智能研发战略规划 2023》明确指出，加强人工智能数据共享是实现人工智能技术创新、

收稿日期：2023−12−04；**修回日期：**2024−06−25

基金项目：湖南省智库专项重大委托项目“创新驱动背景下加强湖南高校知识产权转化运营的对策和建议”(18ZWA08)；湖南省社会科学基金重点项目“数字环境下开放存取版权问题研究”(23ZDB001)；湖南省研究生科研创新项目“国际条约背景下数据知识产权保护规则研究”(CX20220135)

作者简介：何炼红，女，湖南韶山人，中南大学法学院教授、博士生导师，主要研究方向：知识产权法学；朱曦青，女，湖南长沙人，中南大学法学院博士研究生，主要研究方向：知识产权法学，联系邮箱：601363634@qq.com

发展和应用的核心举措[2]。欧盟委员会发布的《人工智能白皮书——通往卓越和信任的欧洲路径》中认为，当前人工智能驱动的关键在于人工智能数据的共享[3]。日本出台的《人工智能战略 2022》中明确提出“人工智能数据科学”概念，并要求通过加强人工智能数据共享来完善该领域的基础设施建设[4]。

我国高度重视人工智能数据共享。早在 2017 年的《新一代人工智能发展规划》中就提出了“倡导开源共享理念”和“依托国家数据共享平台”的要求③，以提升公共领域中人工智能数据共享效率。2022 年《中共中央、国务院关于构建数据基础制度更好发挥数据要素作用的意见》(以下简称《数据二十条》)④作为数据要素治理的纲领性文件，明确了“坚持共享共用”的基本原则以促进数据流通。未来，我国数据产业对人工智能数据的需求呈指数级增长趋势，国家工业信息安全发展研究中心发布的《2023 人工智能基础数据服务产业发展白皮书》显示，至 2025 年其服务市场规模将达到 123.4 亿元人民币，相较于 2022 年的 47.8 亿元人民币，复合年增长率达到 37.2%⑤。为适应数据产业的高速发展，构建高质量共享数据集是关键。然而与实践需求不相匹配的是，目前理论研究就如何在公共领域加强人工智能数据共享尚未形成共识。既有研究多关注通用数据的开放和利用，且侧重于权利定性和权益保护问题的探讨⑥。面对公共领域数据扩容、公共利益保护和技术发展的迫切需求，鲜有文献研究人工智能数据领域如何进行制度回应。本文拟立足“深度共享”理念，从平台、资源和规则的三维视角出发，探讨人工智能数据公共领域深度共享机制的构建，以期实现公共领域数据充盈，促进数智技术快速发展和广泛应用。

一、人工智能数据公共领域共享面临的挑战

近年来，人工智能技术在全球范围得到了迅猛发展，正处于从智能化到自主化的成长阶段。生成式人工智能具有的自主学习、高效数据处理和智能生成的特征，使其比通用人工智能更依赖于数据共享[5]。其中，平台是人工智能数据共享的关键基础设施，共享离不开对数据资源的充分利用，并需要具体明确的共享规则予以指引，从而形成以平台为核心的人工智能数据共享生态。

基于对人工智能数据资源的分配利用，平台、用户和公众三者之间形成了数据权益关系，并以智能合约等规则形式实现数据私权权能(私人利益)和公权权能(公共利益)的平衡。通说认为，人工智能数据权利是在“权利束”(a bundle of rights)基点上阐释的广义的“数据权”，包括私权、公权和国家主权在内的多项权能⑦。以数据管辖等方式来维护国家安全的主权权能主要体现于数据跨境语境，因此，以平台为核心的数据权能问题主要涉及平台的数字技术强把控与公共领域数据资源供给不足之间的冲突。首先，平台通过数字技术把控，对人工智能数据具有更强的可控性，更易形成数据霸权，导致人工智能数据公共领域共享自由的缺位；其次，平台对数据资源的封闭利用形成共享壁垒，使得人工智能数据公共领域范围一再限缩；最后，平台的数据共享规则“自主性”使数据权属重叠交叉，导致人工智能数据公共领域共享标准模糊。以平台为核心的人工智能数据公共领域面临诸多共享障碍，难以满足人工智能促进技术、文化和经济发展的基本要求[6]。因此，如何扩容人工智能数据公共领域成为当下人工智能发展的重要议题。

(一) 平台层面：人工智能数据霸权导致信息共享自由的缺位

人工智能数据霸权指的是人工智能平台为避免数据流通对其权益造成损害，擅自提高了数据流入公共领域的门槛，形成过度的数据保护并妨碍信息共享自由的行为。人工智能数据霸权出现的根本原因在于平台的技术把控以及“信息自决”理论的盛行。后者源于“个人信息自决理论”，强调个人对其信息具有决定、控制和支配的权利。具体而言，一方面可以防止他人未经许可对信息的非法传播与利用，另一方面可以实现个人信息的自由交易[7]。在人工智能领域，该理论被延伸使用，逐渐演

变为突破个人信息桎梏的“数据信息自决”理论。该理论认为人工智能平台对数据价值的产生投入了大量的成本，是“数据发展的中坚力量”[8]，亦是数据财产性权益的关键利益方，应具有“信息自决”的权利。

由于平台拥有规则制定权等数据权力，具有权力扩张和滥用的风险[9]。“信息自决”为平台的人工智能数据霸权行为奠定了实施基础，辅以技术把控的实施保障，导致了信息共享自由的缺位。究其原因，信息自决理论对人工智能平台“赋权”的同时并未对其作出必要的限制，使得数据赋权逐渐异化为“数据监控”与“数据独裁”，形成数据霸权[10]。比如日本富士通公司据此在 2013 年建立了名为“Data Plaza”的数据交易中心[11]，明确“人工智能数据”的交易类型，并实施云端核心技术以防范用户支付使用后数据流入公共领域，实现对数据权益的“独家收揽”。换言之，用户只能通过平台对人工智能数据进行单向使用，且该使用过程将被平台全程监控，任何非授权的共享行为将面临侵权风险。封闭或垄断的数据共享空间将导致数据权益的保护过度[6]，平台的数据霸权将形成对人工智能数据严格把控的行业模式，实践中更是难以对“所有权人”坚持保密和“私有”的数据进行开放流通，因此，亟待一种有效的机制确保人工智能数据公共领域中的信息共享自由。

(二) 资源层面：人工智能数据壁垒导致公共领域范围限缩

人工智能数据的单向或局域共享形成互相独立的数据孤岛，而孤岛之间的数据资源各自封闭，造成了数据壁垒的客观存在，致使人工智能数据公共领域不断限缩。具体而言，人工智能数据壁垒表现为物理壁垒和逻辑壁垒[12]。

物理壁垒是指由于人工智能数据被存储在不同的主体或硬件上，不同主体各自独立地维护与使用数据以致其难以共享的情形。一方面，人工智能数据是信息的电子载体，依赖于存储载体以进一步运用，而分散独立的存储载体成为数据共享的现实难题。另一方面，此时共享人工智能数据的经济效益难以抵过投入成本。虽然人工智能数据的价值在共享中是低消耗性的，并不会因为共享主体增多而减损，然而，存储、维护、转移和运用数据的行为需要投入成本，使得数据主体基于成本—利润的经济效益考量作出了不共享决策，并自然形成共享屏障。

逻辑壁垒指由于没有统一的解读标识符，导致不同个体对同一数据对象的理解存在差异，无法达成数据共享。详细来说，不同个体对如何理解和使用人工智能数据具有极大的主观性，比如是否可以修改？是否可以再次共享？是否涉及个人隐私？这些解读差异导致数据共享的渠道难以打通，只能在极少数个体间搭上共享桥梁，而无论是物理壁垒还是逻辑壁垒，本质上都是对人工智能数据公共领域的侵蚀，使其范围不断限缩。因此，打破数据壁垒是实现公共领域数据共享的基本前提。

(三) 规则层面：人工智能数据权属不清导致公共领域共享标准模糊

人工智能数据涉及平台、用户和公众等多元主体，数据兼具凸显个人福祉的私权属性、体现公共利益的公权属性与维护国家安全的主权属性[13]，展现出“权利束”的样态，也因此对数据权利的边界提出了挑战。与此同时，不断发展的人工智能技术使得人机关系愈发复杂，进一步导致数据权利重叠和交叉。因此，各权益主体为最大限度地保护其数据权利，在人工智能领域逐渐形成了保守且各异的数据共享规则，然而，建立数据要素市场化配置机制的前提是数据确权[14]。换言之，数据权利的在先界定是公共领域数据共享的前提。目前关于人工智能数据的权利归属存在着不同的观点，各行其是的权属规则成为公共领域数据共享的主要障碍。

其一，个人权属论认为人工智能数据权利应由数据的产出者即个人所拥有。因数据涉及人格和财产的双重属性，相关权益应由个人自决[15]。但是该学说具有现实和理论的障碍。首先，在现实层面，个人与人工智能平台之间存在明显的信息差和技术差，将权益完全归属于个人难以实现数据的交易和保护功能。其次，在理论层面，数据价值的产生往往依赖数据池的存在，个人少量数据难以通过“个人数据商店”的形式实现其经济价值[16]。最后，数据池价值的高低依赖于人工智能平台的数据分析

与挖掘能力，例如文字、图像和视频等形式的数据解析难度有高有低，而 ChatGPT 成为“新一代知识调用和表示方式”的根本原因在于可同时处理不同形式的数据，并形成“类人化”处理模型，提供决策价值[1]。

其二，平台权属论认为平台应是人工智能数据的权利主体。根据劳动财产学说和激励理论，将数据权利归于平台符合投入–回报型的劳动财产特性，并对平台形成研发激励。但该理论存在两个方面的问题：一是排除了平台上用户对于人工智能输入数据再次授权的可能，数据的处理过程使得用户本来基于输入数据产生的权益也被转移，变相成为“独占许可”的情形；二是无法完全排除用户的个人数据权益。人工智能数据除财产属性外还兼具人格属性，同时，作为信息的电子载体，《中华人民共和国个人信息保护法》(以下简称《个人信息保护法》)对人工智能数据同样适用。其明确规定“自然人的个人信息受法律保护，任何组织、个人不得侵害自然人的个人信息权益”，涵盖了人工智能数据所承载的个人信息。

其三，公共权属论基于数据的“公共性”特征(即非竞争性和非排他性)，认为人工智能数据权利应属于公众整体。数据的公共性体现在互惠分享上[17]，但是过分强调人工智能数据的自由流通会得到相反的结果，该结论已在欧盟得到印证。2018 年欧盟通过的《非个人数据自由流通框架条例》旨在促进非个人数据的流动，使公众在最大限度上共享数据利益，以期打造具有竞争力的数字经济市场。然而，数据流通中产生的财产利益远远小于将数据掌握在“自己”手中，使得该条例并没有发挥其本身应该具有的效力，反而导致数据共享渠道愈发闭塞，与分享经济的理念背道而驰[18]。

其四，综合权属论强调多元主体共治。无论是个人权属论、平台权属论或公共权属论，都面临较大的争议，与之相比，综合权属论获得了更为广泛的认可。由于数据本身所具有的聚合性和关联性等特征，将数据视为一种权益混合的聚合型财产具有合理性[19]。《中华人民共和国数据安全法》(以下简称《数据安全法》)明确规定国家保护“个人、组织与数据相关的权益”，因此，与数据相关的权利主体可以确定为人工智能平台和用户。实践中，平台与用户也遵循共治的模式。例如，腾讯公司研发的人工智能产品在对用户开放使用时，《腾讯服务协议》第 11 条第 1 款规定，“用户在使用服务中所产生内容的知识产权归用户或相关权利人所有，除非与腾讯另有约定”[20]。由此，数据权利可以由用户和平台基于合同约定进行具体分配。然而，人工智能数据处理程序多样、变化频繁，使得相关利益方的关系更为复杂，导致综合权属论本身亦成为数据权属重叠交叉的产物。因此，亟待从规则层面明确多元主体之间数据使用的边界，确立人工智能数据公共领域的共享标准。

二、人工智能数据公共领域深度共享的理念转型

人工智能数据公共领域“深度共享”理念的转型，既是来自平台、资源和规则层面的多维审视，也是一种理论阐释和分析方法的突破。首先，要运用“网络三层理论”对网络物理层、逻辑层和内容层的内涵进行深度阐释，实现从单层封闭式枢纽向多层开放式网络的平台共享理念转型；其次，要引入“古罗马多层财产体系方法论”构建公共领域财产双层共享体系，实现从单一到多维方式的资源共享理念转型；最后，要基于人工智能数据的多元与共享属性，实现从保守到开放的规则理念转型。由此，实现人工智能数据公共领域的“深度性”共享。

(一) 运用网络三层理论：实现从单层走向多层的平台共享理念转型

目前，人工智能数据共享是一种依赖于平台的封闭枢纽模式，通常适用于特定行业的人工智能数据社区，不同行业之间的数据难以共享[3]，而生成式人工智能对数据多样性提出了更高的要求，以实现“类人化”创造性运用，其中，如何突破枢纽式的共享平台桎梏成为关键。推动数据共享的本质在于实现其网络效应[21]，共享主体的增加会带动网络节点的线性增长，并使数据价值呈指数级增长。这

恰恰是平台的封闭枢纽模式所欠缺的。因此，有必要引入“网络三层理论”构建开放网络模式，该模式的运行逻辑是数据提供者和使用者之间直接共享，以摆脱对特定平台的依赖。目前，通过百度云或微云等云盘的共享仍是基于某一特定平台实现的，而开放网络模式是顺应网络自然结构的自动共享，并不依靠平台提供数据。当然，此时的平台并没有消失，其仍然具有数据记录和辅助共享的工作能力，只是转变成了一种调节性网络。

网络三层理论认为，网络是公共领域的子概念，数据网络的建设是以共享为目标的[22]。网络最底层为物理层(physical layer)，由计算机和线路组成，没有共享主体的限制；物理层之上是逻辑层(logical layer)，主要涉及对语言的逻辑使用，基于语言的自由表达属性，语言使用也是自由的；而逻辑层之上则是内容层(content layer)，兼具共享和可控属性，数据位于该层。物理层、逻辑层和内容层以金字塔式结构形成数据网络，其架构基石是对共享的需求。换言之，网络是公共领域的数据共享系统。

与此同时，人工智能是一个通过网络进行数据分析处理的系统。网络的兴起是人工智能创设的基础，而网络技术的繁荣亦是人工智能发展与完善的关键。内容层的建立基于“自由流通”的物理层和逻辑层，因此，构建内容层的原始内涵也是在“自由流通”的公共领域范畴。人工智能数据共享的开放网络模式是基于网络三层结构而构建的，位于内容层的人工智能数据的共享是公共领域下的“深度共享”。此时，数据提供者与使用者之间可以直接共享，数据提供者则可以通过网络对数据的共享进行限制，体现内容层的可控属性。此时的网络代替了平台的枢纽作用，完成原先由平台所完成的数据处理、追溯等工作，被称为调节性网络。调节性网络虽然不能直接访问人工智能数据(主要数据)，但其可以通过对描述数据(记录主要数据的数据)进行管控，从数据提供的源头促进共享，排除平台垄断，甚至平台自身亦成为调节性网络的一环。基于网络三层理论，人工智能数据共享平台从封闭枢纽模式走向开放网络模式，实现了从单层共享向多层共享的“深度性”转型。如图 1 所示。

(二) 引入古罗马多层财产体系方法论：实现从单一走向多维的资源共享理念转型

目前，数据登记注册式管理(data stewardship and base registries)是数据共享的主要方式，数据使用者与提供者通过签订合同并遵从同一套规则来实现共享。数据登记注册式管理在欧盟得到广泛运用，并摸索出一套“可信数据共享框架”(trusted data sharing framework)作为行为模板[23]。该框架的核心在于通过不可否认且可追踪的可信第三方(trusted third parties，简称 TTPs)对数据进行标识，实现对数据共享行为的促进式管理[24]。我国地方政府也在陆续出台规范性文件进行数据登记试点，如《广东省数据知识产权登记服务指引(试行)》《浙江省数据知识产权登记办法(试行)》等。这种标识性共享的方式为数据共享提供了新思路，然而，上述探索仍然没有摆脱单向分享的方法误区。使用者只有通过单向

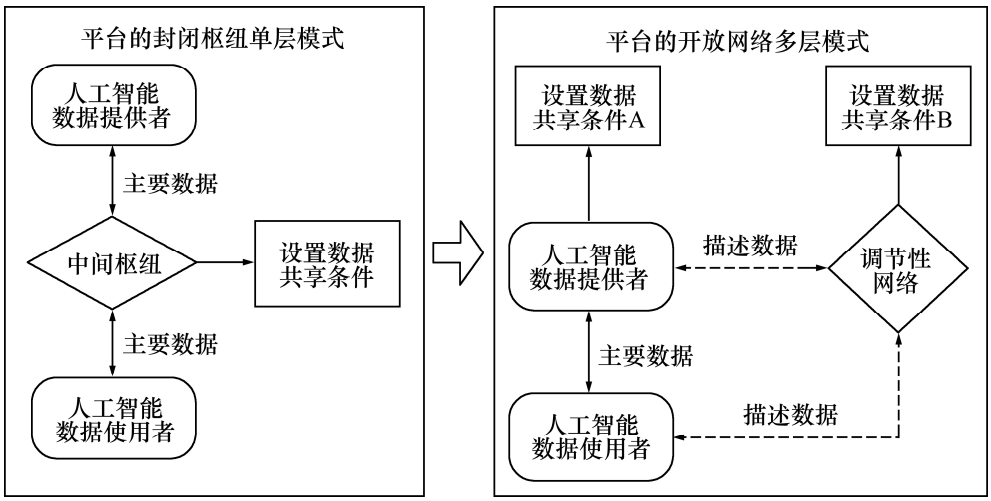

图 1　人工智能数据共享平台的封闭枢纽单层模式与开放网络多层模式

申请才能获取数据，并且由于被“标识”，其不可擅自共享给非申请者。此时的“标识”具有“授权”性质，并形成数据资源的单向分享。这不仅极大地降低了共享效能，而且加剧了人工智能数据共享的壁垒困境。

追溯历史，早在古罗马时期，就建立了一个包容多种财产形式的公共领域财产体系，使各类财产能够在公共领域以多种方式进行“深度性”共享。在古罗马的多层财产体系方法论当中，依据《查士丁尼法学总论——法学阶梯》所述，涉及公共领域的财产包括以下内容：要式财产(res mancipi)、公法人财产(res universitatis)、法律属性公共财产(res publicae)、自然属性公共财产(res communis omnia)、无主财产(res nullius)和人法财产(res humani iuris)[25]，这些财产类型以体系化的方式组建了公共领域财产的概念⑧。而古罗马能够充分实现财产共享的关键在于构建了公共领域财产双层共享体系，一方面，通过设置“公共领域财产官方标识”(official public domain mark)，对可流通于公共领域的财产加盖官方标识，排除共享障碍[6]，所有人都可以直接使用被官方标识的财产；另一方面，对未加盖官方标识的财产，可以基于意思自治进行共享。二者有机结合构成了公共领域财产双层共享体系，能有效避免对同一财产对象的公共性进行重复判断，强化了财产在公共领域的有效使用。

运用古罗马的多层财产体系方法论，在人工智能数据公共领域设置双层共享体系同样具有重要的法律意义。美国杜克大学的 David Lange 教授在研究“公共领域”时发现其本身就是一个法定的概念，具有法定性和权威性，应由政府相关部门或授权机构实施[26]。美国斯坦福大学的 Kop Mauritz 教授进一步指出，人工智能数据在公共领域的共享可以通过“公共领域财产官方标识”的构建来实现，这是一种法定标识，且“公共领域财产官方标识”若能愈发有效地实行，人工智能数据公共领域则会愈发繁荣[6]。同时，在人工智能数据领域，结合智能合约等技术手段进行数据开源，也能实现数据的自治共享。

在实践中，人工智能数据公共领域的双层共享体系已经得到了推广和应用。在欧盟，2024 年通过的《人工智能法案》(*Artificial Intelligence Act*)明确指出需要对人工智能数据的性质和数量进行评估，并对通过评估的人工智能加盖“CE 标识”，以确保人工智能数据在公共领域的安全使用⑨。美国食品药品管理局(Food and Drug Administration，简称 FDA)早在 2019 年就以官方监管机构的身份公布了三十余种加盖官方标识的人工智能医学智能算法(smart algorithms for medical discovery，简称 SAM)所生成的数据产品，包括血糖浓度变化数据等，使其能够在公共领域被共享使用[27]。此外，目前已有多类人工智能数据通过许可声明以“开放标识”的方式流入公共领域。如约翰·霍普金斯大学系统科学与工程中心的 COVID-19 数据明确通过“公共许可”标识告知公众，此类人工智能数据可以自由复制和分享[28]。也就是说，其自愿签订知识共享协议，在数据上标注协议标识以作出许可声明⑩。据此，为促进人工智能数据在公共领域的共享，Kop Mauritz 教授明确提出“机器的公共财产”(res publicae ex machina)模型，指出人工智能数据是一种来源于机器的公共财产，本身处于公共领域，并可通过官方标识和意思自治的方式对人工智能数据公共领域进行扩容⑪。

无论是“公共许可”还是“机器的公共财产”，均是基于古罗马多层财产体系方法论对人工智能数据公共领域共享机制的有益探索，对于人工智能数据公共领域形成“深度性”资源共享具有重要的借鉴意义。

(三) 基于人工智能数据的共享属性：实现从保守走向开放的共享规则理念转型

我国人工智能数据的载体无形性、主体多元性和流通低消耗性，决定了其所具有的共享属性。如果说，《数据安全法》和《个人信息保护法》对多主体共享数据提出的是以“保护”和“问责”为核心的“保守”式规则，那么，《生成式人工智能服务管理暂行办法》强调的则是以“促进”和“免责”为核心的“开放”式规则⑫。从保守走向开放的共享规则理念转型，是促进公共领域人工智能数据深度共享的重要举措。

其一，人工智能数据的社会属性是共享。《数据安全法》将数据定义为“任何以电子或者其他方式对信息的记录”。人工智能数据是比特形式的电磁记录，是一种无形电子载体，体现出二进制下以“0”和“1”形式自由流动(free movement)的特点，应与纸张、画布等有形载体区分开来[29]。在共享经济的理念中，表现为信息电子载体的数据对数字经济的发展具有决定性作用[30]。电子载体价值在共享中是非消耗性的，不会因为频繁的共享而产生损耗，反而会增值。一方面，人工智能数据经过分析和处理后，输出的数据产生了经济价值和文化价值等；另一方面，平台之间的人工智能数据共享，也会使数据价值在不断筛选与分析中得到提升。

其二，人工智能数据的法律属性本质也是共享。人工智能数据权利是包括私权、公权和国家主权在内的多种权能。其重点在于权能平衡，并体现为共享障碍的解决。欧盟 2022 年颁布的《数据法案》(*Data Act*)明确说明，该法案的诞生是由于“数据共享障碍会阻碍数据造福社会的优化分配”[31]。同时，数权的所有权能中，共享权是权利人对于数据的最终利用[32]。即使是以使用和收益为主要内容的数据用益权，同样具有共享性。用益数权强调数据的经济价值，以平衡“所有”与“利用”之间的关系。此时的用益数权并不包括对数据的共享权能，但是，这并不意味着不能对用益数权这个权能本身进行共享，以数据的经济价值为锚推动数字经济的繁荣。因此，具有多项权能的数权本质是共享权。综上，基于人工智能数据社会属性和法律属性对共享性的契合，人工智能数据共享规则也应实现从“保守”走向“开放”，实现人工智能数据的“深度性”共享。

三、构建人工智能数据公共领域深度共享机制的原则和进路

数据共享是人工智能技术发展的关键。《数据二十条》为促进数据共享提出了非公共数据的市场化共享收益模式和公共数据的授权化互联互通模式，为人工智能数据深度共享机制的构建指明了方向⑬。《生成式人工智能服务管理暂行办法》明确提出了人工智能数据公共领域共享平台的建设目标⑭，强调要“促进算力资源协同共享”，实现“公共数据分级分类有序开放”。《“十四五”大数据产业发展规划》则提出要积极参与数据共享的规则体系构建，并在国际上形成人工智能数据共享“中国方案”⑮。然而，纵观我国关于数据共享的相关规定，依然既零散又宏观，且我国针对人工智能数据的共享规则多以个人数据为主要内容，表现为以个人信息保护为重点的“保守式”共享规则⑯。为此，有必要从体系化的视角，对人工智能数据公共领域深度共享机制进行具体架构。

(一) 人工智能数据深度共享的前提：开源

科技发展至今，“开源”已不再单纯指开放计算机程序源代码，而是一种开放的产品形态，其概念内涵已升格为一种无边界的协作模式和开放共赢的合作理念⑰。换言之，其是对开放(openness)、对等(peering)、分享(sharing)以及全球运作(acting globally)理念的深化[33]，并在国际层面逐渐形成全球开源生态。“开源”是促进数据共享、数据协作并获得无偏向性分析的最佳策略，也是数据深度共享机制的核心理念。“倡导开源共享理念”是《新一代人工智能发展规划》针对人工智能数据共享提出的要求。《中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要》(以下简称《“十四五”规划》)则明确强调，要通过支持数字技术开源社区等创新联合体发展来完善开源知识产权和法律体系⑱。人工智能数据的开放网络共享模式，是以构建数据开源共享社区为目标的[34]。因此，人工智能数据公共领域深度共享机制的构建也应以“开源”为前提，鼓励人工智能技术创新，促进人工智能产业发展。此时，开源的内涵具体表现为形成信息有限支配的共识和采取数据可控共享的方式。

1. 形成信息有限支配的共识

支配性权利会带来创新激励的观点容易产生误解[35]，因为过多的支配性权利反而会导致数据资源

封闭并阻碍创新。为避免数据归集所造成的平台霸权、资源垄断和规则模糊，数据治理的思路应由赋权转为控权[36]，因此，有必要形成信息有限支配的共识。

其一，从人工智能平台的视角出发，人工智能数据的排他权是一种有限排他权。基于人工智能数据的财产属性以及平台经济背景和技术优势，人工智能平台更易对人工智能数据形成隐形的“数据监控”，从而形成“数据霸权”式绝对排他权。然而，无论是劳动财产学说还是功利主义学说，其作为支撑理论在支持人工智能平台是数据权利归属者的同时，均仅仅针对财产权，并不包括人身权，并说明这是一种有限的排他权。此外，人工智能在数量巨大且广泛联系的数据中进行筛选处理，使人们能够在社会劳动中产生新的知识和灵感，提升了数据的“经济效率”。此时的“经济效率”伴随数据共享程度产生周期变化，过度的共享和排他均会导致经济效率降低，亦如前述欧盟《通用数据保护条例》和《非个人数据自由流通框架条例》的效果实证。数据强调载体特性[37]，信息强调内容特性，二者具有价值协同性。信息的受支配程度同样随数据“经济效率”的变化而变化。因此，人工智能数据开源的深度共享应以明确信息有限支配为前提。

其二，从人工智能用户的视角出发，过度的个人信息支配超越了隐私保护的范围。一方面，对数据的人格利益进行保护具有合法性。比如欧盟 1981 年颁布的《个人数据自动化处理中的个人保护公约》明确指明，个人数据在被处理时要尊重有关个人的隐私权(their right to privacy)[38]。《中华人民共和国民法典》(以下简称《民法典》)和《数据安全法》等法律法规均明确提出对数据的人格利益进行有效的保护⑲。然而，人工智能数据不是纯粹的个人数据。平台在处理人工智能数据时，会对有关个人隐私的数据自动或人为地进行脱敏处理，以保护相关自然人的人格利益。另一方面，人工智能数据承载的个人信息范围比隐私范围要更广[39]。根据《个人信息保护法》的规定，个人信息是指针对特定自然人的识别性信息。“隐私”是指不得遭到不法干涉的私生活安宁以及不得受到不法公开的私生活秘密[40]。而信息越私密，不得被干涉和不得公开的程度越高，也就代表自然人的个人特性越易被识别而成为隐私。换言之，个人信息包含隐私，二者具有涵盖关系，但前者比后者的范围更为宽广[41]。因此，对人工智能数据的支配不得超越隐私保护的范围，应是一种有限排他性支配。

2. 采取数据可控共享的方式

数据共享既是一个事实行为，也是一个法律行为。事实行为表现为数据控制者向其他控制者提供数据的具体过程，而法律行为表现为具有可控性的数据流通。此时的可控性并不等于保守式规范，与此相反，其旨在更有效地促进数据流通。

人工智能数据的可控共享方式包括两种：一是为平衡个人利益、公共利益和国家利益的必然性共享；二是来自数据控制主体的许可性共享。必然性共享具有避免公共利益被侵蚀的公共领域边界的意蕴。创新是一个建立在自由和控制这一组合架构之上的概念，其需要相关主体基于人工智能数据的类型、属性、来源、级别等内容对自动流入公共领域的人工智能数据进行明确，并建立公共领域标识，避免数据霸权和数据壁垒的形成，以促进生成式人工智能技术创新。

许可性共享指人工智能数据的“所有人”具有独立可控的能力来授权共享，以促进数据流通。数字时代的合同法应注重许可协议[42]，其中包含三个内涵：其一，数据主体具有数据控制能力，无论是个人、企业或政府；其二，数据主体对共享对象的信任；其三，人工智能数据共享领域的安全性保证[21]。许可性共享具有操作上的可行性，开源软件就是其中的典型。严格来说，开源软件并不属于公共领域[43]，它之所以可以促进公共领域中软件算法的创新和进步，是因为许可证中明确表示相关源代码可公开使用，使公众获得许可性授权以激励创新。换言之，开源许可保护了程序开发者的数据公地。违反开源条款制作专用程序衍生品的行为，将被视为知识产权侵权。因此，开源许可通过知识产权许可来保护和维护现有知识资源的“公有物”。虽然开源许可针对的是软件领域，但其影响力和前瞻性使得其他领域开始借鉴和效仿，并形成一种理念，逐渐影响至数据层面。2022 年中国信息通

信研究院发布的《全球开源生态研究报告》明确指出，开源生态对人工智能技术领域具有重要的驱动作用，同时可以推动数字技术的创新和应用。

(二) 人工智能数据深度共享的基本原则：FAIR 原则

Mark D. Wilkinson 最早于 2016 年提出 FAIR 原则，即共享的数据须符合“可查找的(findable)、可获取的(accessible)、可互操作的(interoperable)、可重复使用的(re-usable)”原则。该原则强调人工智能的机器可操作性，即在人为干预最小或几乎没有的程度下人工智能发现、访问、互操作和重用数据的能力[44]。为了更好地贯彻 FAIR 原则在人工智能数据领域的实行，一个倡议组织 GO-FAIR 诞生，并提出 FAIR 原则的内涵应在“FAIR 数据和服务网络”(internet of fair data and services，简称 IFDS)框架下明确[45]。IFDS 是一个沙漏式模型，服务和数据是沙漏的上端与下端，表示在该领域中自由流通的最大程度，而中间狭窄的阀口则表示二者进行交换时应以限制程度最小的标准进行。换言之，这并不意味着 IFDS 是一个立体三维的流程概念，而应当是代表着自由程度和限制程度的范围概念。目前，欧盟、美国、澳大利亚等国家和地区已然采取相应举措以支撑 IFDS 框架，如欧盟 2022 年颁布了“欧盟开放科学云”计划(European Open Science Cloud，简称 EOSC)[46]，明确说明具有科研目的的人工智能数据共享应以最小的限制和最大程度的自由流通为导向。

IFDS 框架下的 FAIR 原则体现在服务与数据的交换层，即数据的共享层，是沙漏最窄的阀口处。具体包括：①可查找的：数据具有唯一和永久的标识符，且可通过搜索资源被索引；②可获取的：数据通过标识符适用不同的标准化协议，且该协议是免费公开并可普遍实现的；③可互操作的：数据需要使用一种正式的、可访问的、共享的且广泛适用的语言进行知识表示，且能够被其他数据引用；④可重复使用的：数据共享需要明确且可访问的使用许可予以授权，并且使用行为要符合数据共享开放社区的标准。IFDS 框架和 FAIR 原则的关系用图示表述更为直观清晰，详见图 2。

在人工智能数据领域，FAIR 原则也已延伸适用。欧盟在 2020 年提出要成为数据治理的领先榜样，基于 FAIR 原则引领数据共享全球标准[47]。2021 年，欧盟成立的人工智能委员会提出《人工智能法案》的构建提案，以明确人工智能数据公共领域共享 FAIR 原则。随后，荷兰于 2022 年颁布的《荷兰数字化战略 2.0》[48]进一步对数据共享原则作出阐释：①数据的共享是自愿的；②必要时，数据的共享是强制的；③数据必须是个人或组织等主体掌握的可控数据。可见，FAIR 原则应作为人工智能数据公共领域深度共享的基本原则。

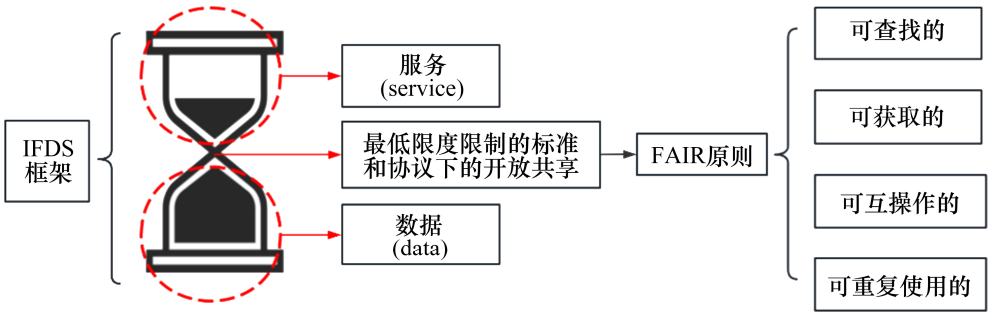

图 2　IFDS 框架与 FAIR 原则的关系

(三) 人工智能数据深度共享的具体进路

人工智能数据公共领域的深度共享，应立足于数据共享生态构建的视角，基于平台、资源和规则三个维度，实现数据共享自由，明确公共领域范围并确立数据共享标准。平台是人工智能数据开放共享的基础设施，在可互操作性原则下构建人工智能数据共享生态是突破平台数据垄断的关键。在此基础上，应构建人工智能数据公共领域的双层共享体系。通过建立数据资源共享名录，设置公共领域官

方标识，明确人工智能数据公共领域范围，以打破数据资源壁垒；通过完善人工智能数据开源协议，确立数据自治共享标准，避免数据权属冲突，以促进人工智能数据在公共领域的流通。

1. 平台层面实现共享自由：可互操作性原则下构建角色模块数据共享生态

深度共享机制的良好运行依赖于“生态位”概念下的各角色模块的平衡运转，以形成人工智能数据开源生态，这是数据可互操作性原则的核心。人工智能数据对生成式人工智能技术发展起到重要推进作用，对于具有可互操作性的人工智能数据共享方式的架构，国内外均已逐步开展相关探索。欧盟于 2017 年提出“欧盟可互操作性共享新框架”(New European Interoperability Framework，以下简称新框架)[49]，包括：可信数据共享的要求清单、一套实现可信数据共享的标准、可信数据共享的协议、访问数据的授权方案、遵守协议的认证规则以及核实协议遵守情况的审核机制。该框架运行于数据的全生命周期，旨在构建一个角色生态系统，使人工智能数据共享是透明的和可管理的。2019 年欧盟进一步提出“Gaia-X 倡议”，旨在保障数据安全可信共享的同时激励创新，并已在德国和法国试点实施。该倡议以数据经济生态系统为核心，以各角色模块的自主运行为主要方式，实现新框架的全面适用[50]。虽然该框架以政府主体和私人主体之间的数据共享为主要内容，但也鼓励私人主体(如个人和企业)之间的数据共享，并以技术创新为共同目标。

党的二十大报告对“完善科技创新体系、形成具有全球竞争力的开放创新生态”以及“加强重点领域、新兴领域立法”作出了专门部署⑳。面对新技术、新产业、新业态、新模式对人工智能数据开源生态提出的新要求，人工智能数据共享主体应在可互操作性原则下构建基于角色模块的数据共享生态。通过分析人工智能数据开放网络共享模式，发现在人工智能数据共享生态中存在三个主要的角色模块：①核心角色。该角色指人工智能数据共享中的直接利益相关者，是共享行为的主体与对象，即数据提供者和数据接收者。②中介角色。该角色位于图 1 的调节性网络区域，提供数据共享的支持性功能，其并不具有访问主要数据的权限，但可处理在人工智能数据共享中用于记录的描述数据。③支持角色。该角色并不直接参与数据共享，如协议许可证提供的相关工作人员等，却是生态系统所必需的辅助角色。可操作性的数据共享生态主要存在两种运行方式：其一，具有公共领域官方标识的数据自动共享；其二，基于数据开源协议实现自治共享。通过角色模块式数据共享生态的构建，避免了对于封闭平台的依赖，以达到人工智能数据在公共领域开源的深度共享。

2. 资源层面明确公共领域：可查找原则下建立公共领域官方标识

可查找原则指数据是可以被统一的标识符所表示且被索引的，这是公共领域促进数据生产活动协同互动的关键。人工智能数据权益具有复合性，为促进人工智能数据在公共领域的流通，应由政府相关部门或授权机构主导、多方协作共同构建，并实时更新《人工智能公共领域数据资源名录》(以下简称《名录》)，对位于《名录》中的人工智能数据设置“公共领域官方标识”，以明确人工智能数据的公共领域边界。不同于数据登记注册式管理，其不是一种申请式共享，而是依据标识信息作出的自发性共享，但又具有法定性和权威性，由政府部门或授权机构实施。

我国已开始尝试构建人工智能数据标识体系，2024 年 4 月公开征求意见的《网络安全技术生成式人工智能数据标注安全规范(征求意见稿)》明确指出人工智能数据标注具有必要性㉑。该文件就如何对人工智能数据属性、来源和安全性等进行数据标识提出指导性意见，却并未就人工智能数据的公共领域共享如何标识作出说明，因此，亟待在此基础上通过构建《名录》予以补充完善。《名录》中的人工智能数据可根据人工智能领域、数据类型、来源、数据级别等进行记载。本文基于现有人工智能发展情况对《名录》的具体内容作出如下构想。

其一，依据《中国人工智能产业生态图谱 2022》㉒确定人工智能数据的所属领域：①基础层：传感器、芯片、数据处理等。②技术层：深度学习、知识图谱和机器学习。③应用层：包括城市级智能和行业解决方案，前者包括交通管理、市政管理和应急安防；后者包括教育、医疗健康和金融等。

其二，依据人工智能数据运行的“输入—学习—输出”全流程，可进一步将人工智能数据分类为：①知识输入数据：具有较高知识信息价值的数据，以知识驱动为核心，用以启动人工智能运行；②训练数据：用以“训练”系统以不断纠正技术或运行问题的数据；③模型数据：具有典型性和参考性的数据；④生产数据/输出数据：通过人工智能分析决策产生的结果数据。

其三，依据人工智能数据的获取方式明确其来源，如企业、个人、政府机构、行业组织等可溯源的主体。对于无法查明来源主体的人工智能数据，需要标注为来源未知。明确人工智能数据的来源在于强调数据的安全性和可靠性，如来源于国家知识产权局公布的专利数据即为此类数据的真实性做了隐形“背书”。

其四，依据人工智能数据的敏感度、数据被损害后的影响度、既往经验与法规硬性要求等来考虑人工智能数据的分级[51]。《网络安全标准实践指南——网络数据分类分级指引》明确指出，可以将数据在依重要性大小分级为核心数据、重要数据和一般数据的基础上，进一步分级为 1 级、2 级、3 级和 4 级数据，依据其对国家利益、公共利益、个人合法权益和组织合法权益的危害程度，再分级为无危害、轻微危害、一般危害和严重危害㉓。对上述利益无危害的一般数据或重要数据，以及危害程度较低的一般数据，可以流入公共领域。人工智能数据所涉及的领域众多，且具有相当的经济、政治与文化价值，因而有必要在《名录》中对人工智能数据的级别进行明确，以说明其使用限制。

综上，通过《名录》对人工智能数据的领域、类型、来源、级别等方面的信息进行明确，是设置人工智能数据“公共领域官方标识”的有效方式。公共领域标识的官方背书是突破数据壁垒的强大推力，由于具有一定的强制性和权威性，其通过必然性共享行为打破了数据物理壁垒的空间桎梏，同时，统一的标识符连接了数据逻辑壁垒的虚拟节点，因而解决了资源层面的数据壁垒问题。

3. 规则层面确立共享标准：可获取、可重用原则下完善数据开源协议

在可获取原则和可重用原则下，规则层面共享标准确立的关键在于如何完善数据开源协议，该协议具有免费、公开、普遍、可访问与标准明确等特征。为实现人工智能数据公共领域的深度共享，开放网络共享模式成为未来人工智能数据共享的主要模式。该模式下数据提供者自行设置共享条件，而去中心化的网络通过对描述数据的调节来运行。该模式也符合区块链技术的发展趋势。在信息共享中，提供“不可否认性”共享服务的可信第三方(TTPs)正在被区块链技术(distributed ledger technology，简称 DLT)所取代[24]。由于 DLT 是一个去中心化的技术，其需要一个合适的治理协议对数据的共享标准予以明确，而数据开源协议无疑是可行的选择。

目前，数据开源协议有三大类：其一是知识共享协议(creative commons license，简称 CC)。知识共享协议是一个基于版权许可的协议，包括署名(attribution，简称 BY)、相同方式共享(share alike，简称 SA)、非营利(noncommercial，简称 NC)以及禁止演绎(no derivative works，简称 ND)等多项选择。由于知识共享协议只涉及版权及其邻接权，对于具有“权利束”特征的人工智能数据而言，显然适用范围过于狭窄。其二是开放数据共享协议(open data commons，简称 ODC)。开放数据共享协议是以数据为对象，类比知识共享协议而产生的协议。如开放数据共享协议下的公共领域专用许可证(open data commons public domain dedication and licence，简称 ODC-PDDL)对应知识共享协议下的公共领域许可协议(CC0)[52]。这两种许可证均要求所有者永久放弃权利，财产完全归属于公共领域，且无须署名。此类过于僵化的协议难以有效平衡人工智能数据的所有权能。其三是社区数据许可协议(community data license agreement，简称 CDLA)。该协议以构建一个数据共享全球性社区为目标，立足于人工智能数据的共享以促进人工智能技术的发展。CDLA 提供了共享协议(CDLA-Sharing-1.0)和许可协议(CDLA-Permissive-2.0)两种版本。前者实质上是公共领域的必然性共享，采用此协议的数据提供者和使用者强制遵守协议的限制条件，后者则是公共领域的许可性共享，协议本身不对数据提供者和使用者的共享行为作出任何限制。必然性共享协议和许可性共享协议的分类符合人工智能数据开源的前提

要件[53]。此类协议是对人工智能数据公共领域共享的突破性尝试，但有过度去平台化的发展趋势，《CDLA 共享合同 1.0》(*CDLA- Sharing-1.0*)第 6 条第 2 款提出：“行使本协议项下授予的任何权利而引起的任何直接、间接、附带、特殊、惩戒性或后果性损害(包括但不限于利润损失)，无论该损害是如何造成……无论是合同责任、严格责任还是侵权行为(包括疏忽或其他)，即使已被告知可能发生损害，您或任何数据提供者均不承担任何责任。”这种绝对责任豁免不符合数据开源的“可控共享”要求。

我国要积极参与国际数据共享规则的制定，在可获取原则和可重用原则下，我国应以 CDLA 协议为基础进一步完善数据开源协议，形成人工智能数据共享标准。《人工智能数据开源协议》(以下简称《协议》)应明确几条规则：①实现数据可控共享。数据使用者可自如使用《协议》项下提供的人工智能数据，但如果数据受到知识产权、数据库特殊权利或其他法律的保护，则需要数据提供者的进一步授权。②不得删减数据共享标识。数据使用者不得删除或修改数据标识，包括《协议》标识，以及代表数据来源和数据类型的标识等。③不得附加限制条款。一方面，即使数据提供者在《协议》项下的数据权利已终止，在终止前的数据权利仍继续存在；另一方面，数据使用者在《协议》项下的数据使用、修改或共享不会被施加限制条件。④保护数据隐私。数据提供者和使用者对《协议》发布和使用的数据作出非隐私性承诺。⑤合法合规运行。数据提供者、使用者和中介方共同遵守相关法律法规，合规运行《协议》。规则的明确带动标准的统一，通过开源协议规范数据提供者和使用者的直接共享行为，可以明确人工智能数据公共领域的深度共享标准，并满足数据的可获取、可重用要求。只有以数据提供者设置共享条件为核心，全球性数据共享社区成为去中心化的辅助性调节网络，才能实现真正意义上的人工智能数据开源和共享。

四、结语

高质量的共享数据是人工智能产业发展的关键[54]。以 ChatGPT 为代表的生成式人工智能作为新时代的“知识表达”形式，对公共领域中人工智能数据的数量和质量都提出了更高的要求。面对平台、资源和规则维度对人工智能数据公共领域共享所提出的挑战，应当以体系化的思维构建人工智能数据深度共享机制，以“开源”为核心理念，引入数据共享“FAIR 原则”，形成以分工明确的角色模块为基础的数据共享生态，构建人工智能数据公共领域的双层共享体系。建立《人工智能公共领域数据资源名录》并设置公共领域官方标识，是明确人工智能数据公共领域范围的有效举措；完善人工智能数据开源协议，确立数据自治共享标准，积极对接国际人工智能数据共享社区，是构建“开放式”共享规则的发展方向。

当前，人工智能立法已成为国际新趋势。2024 年 5 月 21 日，欧盟理事会正式通过的《人工智能法案》就数据共享规则达成共识，基于风险级别对数据进行分类监管，并辅以官方合格标识(CE 标识)的加盖。2024 年 3 月 16 日，中国“AI 善治论坛——人工智能法律治理前瞻”专题研讨会也发布了《中华人民共和国人工智能法(学者建议稿)》，指出国家应推进开源生态建设，鼓励建立人工智能数据资源共享机制，推动公共数据开放共享㉔。为更好地引导和规范人工智能的健康发展，我国应把握时机，尽快出台人工智能相关法律，促进人工智能数据公共领域的深度共享，为抢占国际科技竞争制高点提供法治保障。

注释：

① 目前，Scale AI、Gretel. AI 等企业开始给外界提供合成数据服务，如 ChatGPT 在其公开的数据集中就提到了有数百万

数据来源于 Scale AI 的喂养，以实现其人工智能训练的目的。参见 http://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/.

② 例如，美国范德堡大学的 Daniel Gervais 教授对大数据的内涵进行剖析，认为人工智能数据包括输入数据、训练数据以及输出数据，参见 Daniel Gervais. Exploring the interfaces between big data and intellectual property law. Journal of Intellectual Property, Information Technology and Electronic Commerce Law, 2019, 10(1):19-36. 荷兰人工智能联盟(Netherlands AI Coalition，简称 NL AIC)同样据此将人工智能数据分为知识输入数据、训练数据、模型数据以及生成数据，参见 Netherlands AI Coalition (NLAIC). Responsible data sharing in AI (Report). https://nlaic.com/wp-content/uploads/2020/10/Responsible-data-sharing-in-AI.pdf.我国亦采用了以流程为核心的人工智能数据规范模式。中国电子商会 2024 年发布的《生成式人工智能数据应用合规指南》就是聚焦于数据采集、数据标注、训练数据预处理、模型训练与测试、内容生成服务等各个数据应用环节对人工智能数据应用提出合规要求的。

③ 《国务院关于印发新一代人工智能发展规划的通知》(国发〔2017〕35 号)明确指出“开源共享理念”的重要性，并提出“依托国家数据共享交换平台”以“支撑开展国家治理大数据应用”等具体举措。

④ 中央全面深化改革委员会第二十六次会议通过了《中共中央、国务院关于构建数据基础制度更好发挥数据要素作用的意见》，并于 2022 年 12 月 19 日对外发布。《意见》明确了“坚持共享共用”的工作原则，同时提出构建具有中国特色的数据产权制度的要求。通过推进非公共数据按市场化方式“共同使用、共享收益”的新模式，将对公共数据加强汇聚共享和开放开发以推进互联互通作为思路指引，探索数据产权的结构性分置制度。

⑤ 参见国家工业信息发展研究中心 2023 年 3 月发布的《2023 人工智能基础数据服务产业发展白皮书》。

⑥ 以数据为主题的法学期刊和著作多侧重于对数权的探讨。参见陈俊华的《大数据时代数据开放共享中的数据权利化问题研究》，载《图书与情报》，2018 年第 4 期，第 25—34 页；何渊的《数据法学》，北京大学出版社，2020 年版；武长海的《数据法学》，法律出版社，2022 版；齐爱民的《数据法学》，高等教育出版社，2022 版。

⑦ 数据权利包含私权、公权和主权三个方面的内容。人工智能数据的私权包括人身权和财产权；公权包括公众视角下的数据共享与访问权等；国家主权涉及数据跨境和数据管辖情形下对国家安全和国家利益的维护。

⑧ 古罗马多层财产体系中涉及公共领域财产的内容有：(1)要式财产规定了土地、房屋、奴隶等财产对象的所有权转移的法定形式，将不符合法定形式进行转移的财产纳入公共领域部分；(2)公法人财产，指以城市、政府等公法人为主体所有的财产，如城市街道等属于公共领域共同使用的财产；(3)法律属性公共财产，指在法律制度下可共享至公共领域的财产，如剧院等依靠人力形成的财产；(4)自然属性公共财产，指自动纳入公共领域的财产，如海洋、光等不依靠人力形成的财产；(5)无主财产，包括从未被占有的财产、被放弃的财产和依罗马法不能为私人所有的财产，其自动流入公共领域；(6)人法财产，指可被纳入公共领域和私人领域进行分配的财产，与神法财产(res divini iuris)相对。

⑨ 2024 年 3 月 13 日，欧洲议会以 523 票赞成、46 票反对和 49 票弃权审议通过《人工智能法》，其中，第一编第 3 条(24)款明确说明 CE 合格性标识是一种表明人工智能系统符合此条例第三编第二章中规定的要求的标识。而在第三编第二章第 7 条指出委员会应评估人工智能系统处理和使用的数据的性质和数量。

⑩ 该许可协议赋予数据使用者两项权利：其一是共享的权利，允许通过任何媒介和任何形式复制、发行作品；其二是改编的权利，允许二次加工、转换和基于作品进行创作。但使用者需要进行署名并指向该协议，以作出标识性声明。

⑪ Kop 教授认为，“机器的公共财产”模型是一种位于公共领域的财产模型，构建思路如：(1)公共财产是公共领域的概念(res publicae as species within the genus public domain)；(2)机器的公共财产是公共财产下的概念(res publicae ex machina as species within the genus res publicae)；(3)人工智能公共财产是机器的公共财产下的概念，并且加具公共领域官方标识予以明确(res publicae digitalis (ex machina) as species within the genus res publicae ex machina, and formal AI public domain (PD) mark by a government institution, territory worldwide)。

⑫ 2023 年 8 月，国家网信办、国家发展改革委、教育部、科技部、工业和信息化部、公安部、广电总局等七部门联合公布的《生成式人工智能服务管理暂行办法》明确提出“包容审慎”的监管规则，以“扩展高质量的公共训练数据资源”作为公共领域人工智能数据共享目标，第六条第二款明确提出“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”，以尝试实现从“保守”走向“开放”的共享规则理念转变。

⑬ 《数据二十条》在探索数据产权结构性分置制度部分明确提出推动非公共数据的市场化共同使用与共享收益模式，以激活数据要素价值创造。同时，在推进实施公共数据确权授权机制中，《数据二十条》明确提出包括加大共享供给、有条件无偿使用、有条件有偿使用和不予公开使用等分类在内的统筹授权使用和管理方式，以推进互联互通，打破“数据孤岛”，保障公共数据供给使用的公共利益。

⑭ 《生成式人工智能服务管理暂行办法》第六条第二款提出“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”，积极探索“扩展高质量公共训练数据”的路径与方式。

⑮ 工业和信息化部出台的《“十四五”大数据产业发展规划》中明确提出积极参与数据领域国际规则和数据技术标准的制定，以开拓国际市场。

⑯ 以个人信息保护为核心的个人数据共享规则包括“合法性原则、目的明确原则、最小必要原则、公开透明原则、准确性原则、可问责性原则和数据安全原则”，是以保护和问责为核心思路的保守式规则构建。

⑰ 参见中国信息通信研究院发布的《全球开源生态研究报告(2022 年)》。

⑱ 2021 年 3 月 11 日，十三届全国人大四次会议表决通过了《关于国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要》，其第十五章“打造数字经济新优势”中明确提出要加强关键数字技术创新应用，并以“支持数字技术开源社区等创新联合体发展和完善开源知识产权和法律体系”为重要举措。

⑲ 我国对于以隐私和个人信息为主要内容的个人数据权益明确予以保护，《民法典》以人格权为内容保护个人信息和隐私权，同时，《数据安全法》第七条明确指出国家保护个人与数据有关权益。

⑳ 2022 年 10 月 16 日，习近平总书记在《高举中国特色社会主义伟大旗帜为全面建设社会主义现代化国家而团结奋斗——在中国共产党第二十次全国代表大会上的报告》中提出“形成具有全球竞争力的开放创新生态”以及“完善科技创新体系”的要求。

㉑ 全国网络安全标准化技术委员会 2024 年发布的《网络安全技术生成式人工智能数据标注安全规范(征求意见稿)》指出，生成式人工智能数据标注指通过人工操作或自动化技术机制，基于对提示信息的响应信息内容，将特定信息如标签、类别或属性添加到文本、图片、音频、视频或者其他数据样本的过程。提示信息指引导生成式人工智能模型完成特定任务并提供合理输出内容的输入信息。响应信息指在生成式人工智能数据标注中，按照提示信息要求形成的符合人类认知的应答信息，用于训练模型形成对提示信息输出相应内容、模式或风格的响应的能力。

㉒ 参见易观分析发布的《中国人工智能产业生态图谱 2022》。

㉓ 参见全国信息安全标准化技术委员会秘书处颁布的《网络安全标准实践指南——网络数据分类分级指引》。

㉔ 参见 2024 年 3 月 16 日中国“AI 善治论坛——人工智能法律治理前瞻”专题研讨会发布的《中华人民共和国人工智能法(学者建议稿)》第十九条以及第二十一条，明确指出我国需与平台等主体共同培育共享协作的开源创新生态，并且应鼓励建立数据共享机制以解决人工智能数据在公共领域中的共享难题。

参考文献：

[1] 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, 45(3): 108−123.

[2] National Science and Technology Council. National artificial intelligence research and development strategic plan 2023 update [EB/OL]. (2023−05−23) [2023−09−10]. https://www.whitehouse.gov/wp-content/uploads/2023/05/National-Artificial-Intelligence-Research-and-Development-Strategic-Plan-2023-Update.pdf.

[3] European Commission. White Paper—on artificial intelligence—a European approach to excellence and trust [EB/OL]. (2020−02−19) [2023−09−10]. https://commission.europa.eu/system/files/2020-02/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

[4] Japan Cabinet Office. AI Strategy 2022 (tentative translation) [EB/OL]. (2022−04−25) [2023−09−12]. https://www8.cao.go.jp/cstp/ai/aistratagy2022en.pdf.

[5] 朱光辉, 王喜文. ChatGPT 的运行模式、关键技术及未来图景[J]. 新疆师范大学学报(哲学社会科学版), 2023, 44(4): 113−122.

[6] KOP M. AI & intellectual property: Towards an articulated public domain[J]. Texas Intellectual Property Law Journal, 2020(28): 297−341.

[7] 安柯颖. 个人数据安全的法律保护模式——从数据确权的视角切入[J]. 法学论坛, 2021, 36(2): 58−65.

[8] 龙卫球. 再论企业数据保护的财产权化路径[J]. 东方法学, 2018(3): 50−63.

[9] 马平川. 平台数据权力的运行逻辑及法律规制[J]. 法律科学(西北政法大学学报), 2023, 41(2): 98−110.

[10] HARARI Y N. Why technology favors tyranny [EB/OL]. (2018−12−31) [2023−11−23]. https://www.theatlantic.com/magazine/archive/2018/10/yuval-noah-harari-technology-tyranny/568330/.

[11] Fujitsu Global. Fujitsu Multi-Cloud Data Analytics [EB/OL]. (2019−12−17) [2023−11−23]. https://www.fujitsu.com/global/services/multi-cloud/data-analytics/index.html.

[12] 叶明, 王岩. 人工智能时代数据孤岛破解法律制度研究[J]. 大连理工大学学报(社会科学版), 2019, 40(5): 69−77.
[13] 连玉明. 数权法 1.0: 数权的理论基础[M]. 北京: 社会科学文献出版社, 2018: 4−5.
[14] 李三希, 王泰茗, 刘小鲁. 数据投资、数据共享与数据产权分配[J]. 经济研究, 2023, 58(7): 139−155.
[15] 刘德良. 个人信息的财产权保护[J]. 法学研究, 2007, 29(3): 80−91.
[16] 胡凌. 商业模式视角下的“信息/数据”产权[J]. 上海大学学报(社会科学版), 2017, 34(6): 1−14.
[17] 张翔. 数据权益之内涵划分及归属判断[J]. 上海法学研究, 2020, 3(1): 338−351.
[18] 郝思洋. 知识产权视角下数据财产的制度选项[J]. 知识产权, 2019, 29(9): 45−60.
[19] 丁晓东. 数据公平利用的法理反思与制度重构[J]. 法学研究, 2023, 45(2): 21−36.
[20] 腾讯网. 腾讯服务协议[EB/OL]. (2004−02−12) [2023−11−25]. https: //www.qq.com/contract.shtml.
[21] 周汉华. 数据确权的误区[J]. 法学研究, 2023, 45(2): 3−20.
[22] LESSIG L. The architecture of innovation[J]. Duke Law Journal, 2002, 51(6): 1783−1801.
[23] JANSSEN M, BROUS P, ESTEVEZ E, et al. Data governance: Organizing data for trustworthy artificial intelligence[J]. Government Information Quarterly, 2020(37): 1−8.
[24] DUNPHY P, PETITCOLAS F A P. A first look at identity management schemes on the blockchain[J]. IEEE Security & Privacy, 2018, 16(4): 20−29.
[25] MACMILLAN F. Arts festivals: Property, heritage or more?[M]. Cambridge: Cambridge University Press, 2013: 16−18.
[26] DAVID LANGE. Reimagining the public domain [J]. Law and Contemporary Problems, 2003, 66(1): 463−484.
[27] FDA approvals for smart algorithms in medicine in one giant infographic [EB/OL]. (2019−06−06) [2024−03−27]. https://medicalfuturist.com/fda-approvals-for- algorithms-in-medicine/.
[28] The Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. COVID-19 data repository [EB/OL]. (2023−03−10) [2023−11−25]. https://github.com/CSSEGISandData/COVID-19.
[29] 刘金瑞. 数据财产保护的权利进路初探[J]. 中国信息安全, 2017(12): 37−39.
[30] SUNDARARAJAN A. The sharing economy: The end of employment and the rise of crowd-based capitalism[M]. Cambridge: MIT Press, 2017: 82.
[31] European Commission. Proposal for a regulation of the European Parliament and of the council on harmonised rules on fair access to and use of data (Data Act) [EB/OL]. (2022−02−23) [2023−11−25]. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2022%3A68%3AFIN.
[32] 龙荣远, 杨官华. 数权、数权制度与数权法研究[J]. 科技与法律, 2018(5): 19−30.
[33] 齐佳音, 张国锋, 王伟. 开源数字经济的创新逻辑：大数据合作资产视角[J]. 北京交通大学学报(社会科学版), 2021, 20(3): 37−49.
[34] WILBANKS J, FRIEND S H. First, design for data sharing[J]. Nature Biotechnology, 2016, 34(4): 377−379.
[35] BOYLE J. The public domain: Enclosing the commons of the mind[M]. New Haven: Yale University Press, 2008: 41.
[36] 郑晓军. 反思公共数据归集[J]. 华东政法大学学报, 2023, 26(2): 53−67.
[37] 高富平. 信息财产——数字内容产业的法律基础[M]. 北京: 法律出版社, 2018: 27.
[38] European Commission. Convention for the protection of individuals with regard to automatic processing of personal data [EB/OL]. (1981−01−28) [2023−11−27]. https://www.coe.int/en/web/conventions/full-list?module=treaty-detail&treatynum=108.
[39] 梅夏英. 在分享和控制之间 数据保护的私法局限和公共秩序构建[J]. 中外法学, 2019, 31(4): 845−870.
[40] 王利明. 隐私权概念的再界定[J]. 法学家, 2012(1): 108−120.
[41] 高富平. 个人信息使用的合法性基础——数据上利益分析视角[J]. 比较法研究, 2019(2): 72−85.
[42] 王利明, 丁晓东. 数字时代民法的发展与完善[J]. 华东政法大学学报, 2023, 26(2): 6−21.
[43] DIBONA C. In open sources: Voices from the open sources revolution [EB/OL]. (1999−01−09) [2023−11−27]. https://www.oreilly.com/openbook/opensources/book/intro.html.
[44] WILKINSON M D, DUMONTIER M, AALBERSBERG I J, et al. The FAIR guiding principles for scientific data management and stewardship[J]. Scientific Data, 2016, 3(1): 1−9.
[45] GO FAIR Initiative. GO-FAIR [EB/OL]. (2017−02−25) [2023−11−25]. https://www.go-fair.org/go-fair-initiative/.
[46] European Commission. European Open Science Cloud [EB/OL]. (2022−02−14) [2023−11−29]. https://research-and-

innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science/european-open-science-cloud-eosc_en.

[47] European commission. Data strategy [EB/OL]. (2020−02−19) [2023−11−27]. https://digital-strategy.ec.europa.eu/en/policies/strategy-data.

[48] Netherland Digital. Dutch Digitalisation Strategy 2.0 [EB/OL]. (2019−11−13) [2023−11−28]. https://www.nederlanddigitaal.nl/documenten/publicaties/2019/11/13/english-version-of-the-dutch-digitalisation-strategy-2.0.

[49] European Commission. New European Interoperability Framework—Promoting seamless services and data flows for European public administrations [EB/OL]. (2017−11−30) [2023−11−29]. https://ec.europa.eu/isa2/sites/isa/files/eif_ brochure_final.pdf.

[50] Gaia-X Hub. A federated and secure data infrastructure [EB/OL]. (2019−10−01) [2023−11−29]. https://gaia-x.eu/what-is-gaia-x/about-gaia-x/.

[51] 王雪诚, 马海群. 总体国家安全观下我国数据安全制度构建探究[J]. 现代情报, 2021, 41(9): 40−52.

[52] Open Data Commons. Open data commons public domain dedication and license (PDDL) v1.0 [EB/OL]. (2009−08−27) [2023−11−29]. https://opendatacommons.org/licenses/pddl/1-0/.

[53] The Linux Foundation. Community Data License Agreement [EB/OL]. (2021−06−01) [2023−11−29]. https://cdla.dev.

[54] European Commission. Toward a common European data space [EB/OL]. (2018−04−25) [2023−11−25]. https://eur-lex.europa.eu/legal-content/en/ALL/?uri=CELEX%3A52018DC0232.

On the construction of deep sharing mechanism of AI data in public domain

HE Lianhong, ZHU Xiqing

(School of Law, Central South University, Changsha 410083, China)

Abstract: Data sharing in public domain is the key to the development of AI industry. The lack of sharing freedom caused by data hegemony at the platform level, the limitation of public domain caused by data barriers at the resource level, and the ambiguity of sharing standards caused by unclear ownership at the rule level, all have a series of negative effects on the public domain of data. AI data should practice the concept of "deep sharing". Based on such three dimensions as the platform, resources and rules, AI data aims to realize the paradigm transformation of sharing platform from single-layered closed hub to multi-layered open network, with resource sharing mode from single to multi-dimensional, and sharing rules from conservative to open. To build a deep sharing mechanism of AI data in the public domain, we should take "open source" as the premise, introduce the "FAIR principle" of data sharing, and adopt a dual-layered sharing model in the public domain, constructing role-modular data sharing ecology at the platform level to achieve data sharing freedom, setting Official Public Mark of the data public domain at the resource level to clarify the scope of the public domain, and improving data open source protocols at the rule level to establish data sharing standards, so as to eventually fulfill abundant and prosperous data resources in the AI public domain.

Key words: AI (artificial intelligence) data; public domain; deep sharing; public domain mark; open source protocol

[编辑：苏慧]