

OECD 人工智能研究报告全文

目录

引言

第一章 人工智能数据挖掘行为解构

数据挖掘的法定构成要件：数据收集、数据预处理与数据存储

第二章 人工智能数据挖掘主体权责生态

2.1 学术研究机构基于科研目的的数据挖掘应用

2.2 人工智能数据聚合服务商的第三方数据供给行为

2.3 数据爬取对知识产权权利人（创作者/所有权人）的实质影响

第三章 数据爬取法律争议焦点及司法实践动态

3.1 涉数据爬取的知识产权侵权客体范围

3.2 版权作品数据爬取的侵权认定争议

3.3 生成物新型侵权诉由

第四章 政策规制路径的初步考量

4.1 构建自愿行为守则

4.2 授权许可机制的创新探索

4.3 标准化合同条款的规制价值

4.4 对知识产权及数据挖掘的认知提升

内容摘要

人工智能（AI），尤其是生成式 AI 的迅猛发展，为知识产权（IP）领域带来了复杂的挑战。AI 模型的开发、测试和验证高度依赖大规模数据集，导致对训练数据的需求激增。一种广泛采用的方法是“数据抓取”，本报告特指从第三方网站、数据库或社交媒体平台自动提取信息的行为。当数据抓取未经权利人同意或未支付报酬时，会直接影响知识产权保护作品的创作者和所有者。此类活动可能涉及多种知识产权及相关权利，包括版权、数据库权利、商标权、商业秘密、肖像权及人格权。

围绕数据抓取的知识产权法律环境复杂且快速演变。现行知识产权法律多制定于现代 AI 实践兴起之前，且不同司法管辖区存在差异，使得其适用性变得复杂。数据抓取常涉及受知识产权保护的内容，引发关于侵权、例外条款（如合理使用或文本与数据挖掘/TDM 规定）的适用性，以及遵守合同条款等问题。抓取受版权保护的材料会引发关于其收集或使用是否构成侵权的疑问。此类诉讼在全球范围内日益增多，美国、欧盟及其他地区已出现多起具有影响力的案例。此外，对 AI 生成内容（尤其是未经授权模仿个人风格、声音或外貌的输出）的担忧，正促使各方采取多样化的法律应对措施，以保护权利并防止滥用。

数据抓取已成为普遍实践，但其涵盖多种方法且缺乏普遍认可的定义。该术语常与“数据挖掘”（指通过计算过程识别模式、趋势和关联的流程）及“网络爬虫”等技术混淆。本报告指出定义不一致的问题，并提出了一个广泛的工作定义。数据抓取的流程通常包括数据收集、预处理和使用。报告分析了不同的抓取技术，并强调需要采用通用/标准术语以及对这些方法进行明确区分。

数据抓取生态系统中不同参与者的行为引发了多样的法律问题。部分参与者利用数据抓取支持研究及其他活动，这表明需要针对不同使用场景制定相应的政策工具。该生态系统主要包括研究机构和学术界、AI 数据聚合商、以及科技公司和平台运营商。研究机构和学术界常通过数据抓取收集数据用于学术和科学目的。AI 数据聚合商被指控将抓取的数据提供给第三方，往往缺乏明确的许可条款或数据来源披露，引发知识产权及其他法律问题。科技公司和平台运营商则既是数据抓取的主要目标（数据来源），本身也常是数据抓取的实践者。

为引导数据抓取在国际协调框架下走上负责任的发展路径，制定“数据抓取行为准则”、推广标准合同条款、开发标准技术工具以及提升认识倡议。若能吸纳包括权利持有人、研究人员、AI 开发者、民间社会及政策制定者在内的广泛多元利益相关方意见，此类规范将尤为有效。

- **自愿性“数据抓取行为准则”：**可确立普遍适用的条款，并为 AI 生态系统中不同参与者（如 AI 数据聚合商与抓取数据使用者）提供具体指导。准则应包含标准术语以确保共同理解，可纳入监测遵守情况的机制（如注册系统），并就透明度和文档实践提出建议。最后，可包含标准合同条款。

- **标准技术工具：**可帮助保护知识产权，使权利人更便捷地管理自身数据的访问权限，例如数据访问控制机制、自动化合同监控和直接支付系统。标准化工具能简化组织的合规流程，并助力权利人在多个平台上保护其权利。

- **标准合同条款：**可解决与数据抓取相关的法律和运营问题。这些条款可作为可选起点，允许组织协商特定条件。其制定将受益于多利益相关方合作，并可针对不同应用场景（从非营利研究到商业应用）进行定制。

- **提升意识倡议：**赋能利益相关方了解数据抓取及其法律影响，获取保护和管理自身权利的信息。这包括帮助权利人理解保护措施、教育 AI 系统用户负责任使用，以及确保 AI 数据生态系统中所有参与者明确自身角色与责任。

引言

人工智能（AI），尤其是生成式人工智能（GenAI）的兴起，正在知识产权（IP）领域引发一系列复杂问题。

知识产权法律历来通过保护创作者权益来激励创新，促成具有价值的产品、技术及其他创作成果（统称“创作成果”）的产生，为社会带来巨大福祉。尽管不同司法管辖区的法律存在差异，但其核心原则具有共性：通过授予权利人对其创作成果的特定专有权利来保障其利益。这些法律框架旨在激励创新、促进创造力，保护个人和企业开发的智力资产。

为了平衡知识产权所有者的权利保护与更广泛的社会利益，知识产权法律通常包含有限的、具体情境下的例外规定，允许第三方在特定条件下无需获得权利人许可即可使用受保护的作品。例如，部分版权法允许在满足特定要求时将受版权保护的作品用于非商业或其他有限目的。某些例外规定要求支付费用，而另一些则无需支付。此外，知识产权保护通常具有期限性，权利到期后作品即进入公共领域。

尽管各国知识产权法律存在差异，但国际条约在全球范围内促进了知识产权权利和原则的协调。这些条约包括由世界贸易组织（WTO）管理的《与贸易有关的知识产权协议》（TRIPS 协议），以及由世界知识产权组织（WIPO）管理的《伯尔尼公约》、《巴黎公约》和 WIPO “互联网条约”等关键公约。这种协调确保了基本的一致性，特别是在版权和商标保护等领域（详见第 3.2 节）。

与版权和商标法不同，商业秘密法主要通过禁止未经授权的使用或披露来保护保密信息，而非授予开发者专有权利。知识产权保护涵盖各种载体和格式的材料，包括文件、软件、照片和图形作品等。部分司法管辖区还承认独特的（*sui generis*）权利以保护特定类型的材料。例如，欧盟（EU）设立了专门的数据库权利（欧盟，1996）。此类专门保护措施有助于应对知识产权在数字时代演变的特性。

知识产权保护被认为对全球创新做出了重大贡献，包括构成 AI 基础的要素（如通过开源或其他许可提供的受版权保护的软件）。重大突破通过开放科学出版物和开源计划得以实现，推动了语言模型的广泛应用和显著进展。例如，谷歌开发并公开发布的“词向量”（Word2Vec）技术革命性地改变了自然语言处理领域，加速了高级文本分析技术的发展（Mikolov, Chen, Corrado, & Dean, 2013）。这种融合专有创新与开放创新的多元生态系统，凸显了维护强大、有效且可预测的知识产权框架对于持续支持 AI 及其他领域创新的重要性。

尽管本文重点探讨通过数据抓取获取数据训练 AI 模型所引发的知识产权考量，但需承认 AI 与知识产权的交集贯穿于整个 AI 系统生命周期。例如，当生成式 AI 系统根据用户提示生成新内容（“AI 生成输出”）时，便引发了这些输出是否应受知识产权保护的问题。目前，许多司法管辖区要求人类参与才能获得版权保护，但关于所需的人类参与程度以及权利归属问题仍待解决（Mammen 等，2024）。

当 AI 生成输出被指控侵犯第三方知识产权或造成其他损害时，复杂的责任问题随之出现。这些问题因 AI 系统运行于涉及多方（从上游供应商到下游用户）的价值链中而变得尤为棘手。此外，跨境 AI 运营会引发管辖权问题（详见第三章）。本报告虽探讨了与数据收集及 AI 输出可能模仿输入数据相关的问题，但不深入探讨更广泛的责任问题或 AI 生成内容的知识产权资格问题。特别是，报告不讨论由 AI 生成专利发明或共同发明所引发的问题。

本报告旨在概述数据抓取背景下 AI 与知识产权的关键问题，助力深入理解数据抓取实践、相关技术、参与方以及全球法律与监管应对措施。报告最后提出初步考量和潜在政策方向，以指导政策制定者释放 AI 创新潜力的同时保护知识产权及其他权利。在此过程中，报告与经合组织人工智能原则[OECD/LEGAL/0449]保持一致，该原则倡导发展和使用既创新又可信赖的 AI，同时尊重并管控与知识产权等权利相关的风险。此外，本报告亦为经合组

织关于增强数据访问与共享的建议（EASD 建议）[OECD/LEGAL/0463]的讨论做出贡献，该建议旨在最大化数据访问与共享的益处，同时促进可信度并保护个人和组织的权利（包括知识产权）。

尽管数据抓取在隐私、数据保护及相关领域引发重大关切，本报告聚焦于其对知识产权的影响。隐私与数据保护问题正通过经合组织及其他相关机构的补充工作进行探讨，包括由经合组织数据治理与隐私工作组（WPDGP）和经合组织人工智能治理工作组（WPAIGO）联合设立的经合组织人工智能、数据与隐私专家组（OECD.AI Expert Group on AI, Data, and Privacy）。该专家组正在分析 AI 与隐私的交叉领域，特别是当 AI 训练数据包含个人数据时的挑战。其既往及当前工作通过考察收集和处理 AI 训练数据的各种方法，补充了本报告的分析，有助于确保隐私考量充分融入更广泛的数据治理框架，实现更全面的治理路径（经合组织，2024）。

这些努力与更广泛的国际倡议协调一致，例如英国信息专员办公室（ICO）于 2023 年 8 月发布的、由全球十二个数据保护机构共同签署的关于数据抓取与数据保护的联合声明。该声明强调了全球协作应对数据抓取相关隐私风险的迫切性（ICO，2023）。

“数据抓取”用于收集数据以训练人工智能系统，面临重大挑战

近期人工智能（AI）技术，尤其是生成式人工智能（GenAI）的迅猛发展，给现行知识产权（IP）框架带来了若干新的政策挑战。各国政府已开始关注并积极思考应对之策。例如，2023 年日本担任七国集团（G7）主席国期间，G7 领导人明确将生成式人工智能引发的知识产权侵权问题列为主要风险之一（经合组织，2023）。

知识产权政策挑战可能在 AI 系统生命周期的早期阶段（即数据收集与处理阶段）就已显现。这一阶段涉及将数据汇总用于训练、微调、测试或验证 AI 模型和系统（如生成式 AI 系统）（OECD，2023）。开发高性能的生成式 AI 及其他基于机器学习（ML）的 AI 系统，通常需要访问海量数据（即 AI 训练数据）以提升其准确性和性能（Clark & Perrault, 2022）。AI 训练数据可能包含个人身份信息、事实、创意内容、软件、音频文件、视频、数字图像乃至几乎任何形式的数字内容。

AI 输入数据可通过多种方式收集，其中包括数据抓取——即 AI 系统开发者在未与数据托管实体协调的情况下，从第三方网站或社交媒体平台自动提取信息。尽管利用数据抓取为 AI 系统收集数据并非全新实践，但其使用规模随着生成式 AI 的爆发式增长而显著扩大（Metz, Kang, Frenkel, Thompson, & Grant, 2024）。

数据抓取活动跨越司法管辖区，引发了诸多紧迫且可能相互冲突的隐私、知识产权及其他政策问题。

积极方面：若以负责任的方式进行，AI 数据抓取可提供获取多样化海量数据的途径，这对推动 AI 研究与创新、开发高性能生成式 AI 系统至关重要，并符合经合组织人工智能原则[OECD/LEGAL/0449]。增加对可靠且多样化合法来源的 AI 训练数据的访问，有助于减少潜在的偏见和其他危害，并通过开发服务于历史欠发达社区的本地化 AI 工具来帮助缩小数字鸿沟（Chason, 2024; Lee & Lai, 2022; Chen, Wu, & Wang, 2023; Hall, Vassilev, Greene, Perine, & Patrick, 2022）。

风险挑战：另一方面，若缺乏适当的防护措施，用于训练 AI 模型或系统的数据抓取可能侵犯知识产权、隐私权及其他权利，威胁安全，并导致其他危害。

从知识产权（IP）视角看：风险评估通常聚焦于数据抓取是否侵犯了 AI 训练数据中存在的知识产权（如版权、商标权、商业秘密等）。然而，其他相关知识产权问题同样值得关注：数据抓取实践是否非法移除了嵌入在 AI 训练数据中的权利管理信息（RMI）；是否绕过了保护这些数据的技术保护措施（TPM）以及是否违反了涉及知识产权保护数据的合同条款。

一、人工智能数据抓取的理解

生成式人工智能系统(以内容为输出)的开发者通常依赖两种主要数据源来训练其模型：**目标数据源**和**大规模抓取的数据** (Baack, 2024)。

目标数据源的特点是规模较小、参数明确,且基于特定质量指标(包括准确性、完整性、一致性、可靠性、有效性和及时性)进行筛选。这些数据源被选中,旨在帮助 AI 系统模仿多样化的风格和表达方式。其关键优势在于来源可明确追溯,使开发者能够验证数据的质量和可靠性。从知识产权角度看,这也有助于开发者确认数据使用符合知识产权和法律要求。

目标数据源中存在多种支持免费使用的许可框架。“Gratis”许可证允许免费使用,但可能限制修改或重新分发的权利;而“Libre”许可证则提供更广泛的修改、共享和重新分发自由,但需遵守特定许可条款。典型例子包括维基百科快照和 arXiv 上的科学论文。该生态系统还涵盖“公共领域”材料(即不受知识产权或其他类似限制保护的作品),例如通过古腾堡计划(Project Gutenberg, 2024)提供的内容。

然而,许多目标数据集包含受知识产权保护的内容,通常需要**付费使用**,并受特定条款约束,要求遵守知识产权法律,并履行额外的合同义务。此类许可安排被视为提供法律确定性,有助于开发者和权利人管理风险。

鉴于目标数据集的规模限制,开发者常通过**大规模抓取数据**来补充训练数据 (Baack, 2024)。这些**抓取数据**来源于书籍、网站、博客、论坛、数据库及社交媒体平台,已被证明对推进通用人工智能模型至关重要,因其能提供有效模型训练所需的庞大且语境丰富的语料库 (Soldaini 等, 2024)。例如,由 BigScience 研究计划开发的 BLOOM 语言模型,利用 Common Crawl (一个大规模多语言网络抓取数据源) 的数据,提升了语言多样性并纳入了低资源语言 (Teven 等, 2022)。

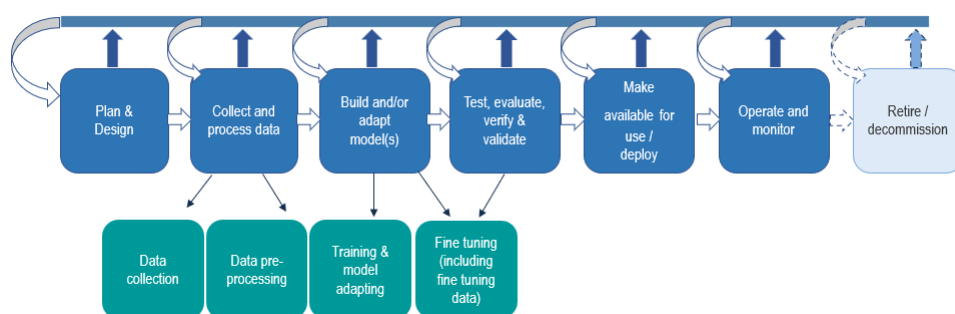
然而,**抓取数据**存在显著挑战。例如,此类数据通常包含比目标数据源更多的无关信息或错误,需要采用高级过滤技术进行质量控制。此外,它还引发复杂的**知识产权合规问题**,尤其当抓取数据的来源不明时,难以验证是否已获得适当授权。

尽管数据抓取已普遍应用,但目前缺乏广为接受的**标准定义**,且可涵盖多种技术和活动,如本节后文将详细讨论。数据抓取通常被称为“**网页抓取**”,因为它主要涉及从网站中提取数据。然而,它还可能包括其他数据提取方法,如**屏幕抓取**和**网页爬取**。“**网络爬虫**”(web crawling) 虽也缺乏明确定义,但通常指通过系统性浏览网页自动索引或聚合网站信息的行为,常用于搜索引擎索引等目的 (S. Gillis, Definition: web crawler, 2024)。在本报告中,**索引**被视为数据聚合、收集或获取的一种形式。

“**数据挖掘**”通常指用于识别大型数据集中模式和相关性的计算过程。关于“数据挖掘”与“数据抓取”这两个术语是否以及如何与法律定义相关,存在一些争议。法律术语“**文本和数据挖掘**”(Text and Data Mining, TDM) 被许多国家在版权法中使用,以创建对版权保护的有限例外。与技术定义不同,**法律定义**可能因司法管辖区而异,并可能由司法或监管机构进行解释。

数据抓取可能涉及多种不同活动,包括:(1) 数据收集,(2) 数据预处理,(3) 使用数据进行模型训练、模型改进,以及/或 (4) 基于测试、评估、验证和确认的微调 (见图 1) (OECD, 2024)。

图 1:



本文使用“数据抓取”这一术语，泛指与从各种来源收集、预处理人工智能训练数据相关的活动，所采用的技术包括网页抓取、网络爬虫和屏幕抓取等。本文还将这些数据在人工智能系统生命周期中的应用纳入考量，包括后续阶段，例如利用数据识别模式、提取特征、优化模型以及进行预测等。

对这些活动和技术有清晰的理解和概述，对于应对与数据抓取相关的政策挑战至关重要。政策制定者可考虑制定标准定义，为利益相关者提供统一的术语和分类体系。在制定政策时，需关注数据抓取的潜在影响，包括对隐私、数据安全、公平性、透明度以及与其他技术（如机器学习）的兼容性等方面的影响。在制定这些定义时，应考虑不同司法管辖区中可能已存在的技术、法律或其他术语。一种方法是制定一个更广泛、涵盖性的数据抓取定义，其他定义则区分属于这一更广泛概念的具体技术和/或活动，同时注明其潜在的版权影响。

1. 关键概念：定义与人工智能训练数据收集相关的术语以开发人工智能系统

获取训练数据对于构建和/或适应人工智能模型，以及测试、评估、验证（简称 TEVV）人工智能系统至关重要（图 1）。对人工智能训练数据的需求在过去几年中大幅增加，尤其是在生成式人工智能系统兴起的情况下。目前，人工智能训练数据通过多种方式获取，包括采购经过整理的数据集、参与数据共享协议、收集用户数据、使用存储数据，以及自动收集公开可获取的数据（即“抓取”）。

尽管需要达成共识，但以下内容将描述本报告中使用的关键术语，并提出了数据抓取的初步工作定义：

人工智能训练数据：指用于人工智能系统开发的数据集，包括用于训练、测试、评估、验证和验证的数据。这些数据集包含结构化数据和非结构化数据，例如数值数据、文本数据、图像数据或音频数据，可通过人类或机器手段收集。人工智能训练数据在人工智能系统开发过程中（“在实验室中”）被用来构建系统，例如利用数据建立模式和预测行为（OECD, 2022）。人工智能训练数据通常与其他类型的输入和数据相结合，用于人工智能系统生命周期中的特定目的，特别是运行时使用的输入数据（如用户查询）。

数据抓取：该术语尚未有广泛接受的定义。本报告提出一个广泛的工作定义，可与其他聚焦于数据抓取中使用的特定技术和/或活动的定义相结合。因此，在本报告中，**数据抓取**指的是利用自动化软件工具或脚本从网络、在线数据库及其他来源自动提取人工智能训练数据的过程。尽管数据抓取可能采用多种技术，但通常具有以下特征：

自动化：通常涉及使用专门设计的软件工具或脚本以快速高效地收集或聚合数据，且人工干预极少。

可扩展性：常用于收集或使大量数据可访问，这些数据通过手动聚合难以实现。所使用的工具和技术可扩展以同时从多个来源提取数据。

非协调性：通常在数据抓取者与数据托管实体之间缺乏协调的情况下进行。

生成式人工智能：一类人工智能模型，通过模拟人工智能训练数据的结构和特征来生成衍生合成内容（美国科学与技术政策办公室，2023）。这可能包括图像、视频、音频、文本及其他数字内容。

数据挖掘：与专注于数据提取的数据抓取不同，数据挖掘指的是用于分析数据的自动化技术。数据挖掘致力于从数据集中分析洞察，包括通过抓取和其他数据提取方法获得的数据集。国际标准化组织 (ISO) 将数据挖掘定义为：“通过从不同角度和维度分析定量数据，提取模式，分类，并总结潜在关系和影响的计算过程” (ISO, 2022)。

文本和数据挖掘 (TDM)：在欧盟法律中，TDM 被定义为“任何旨在分析数字形式的文本和数据以生成信息（包括但不限于模式、趋势和相关性）的自动化分析技术”（《欧盟指令 (EU) 2019/790 关于数字单一市场中版权和相关权利》第 2 条) (EUR-Lex, 2019)。

其他司法管辖区在其法律框架中可能对 TDM 有不同的定义。

来源：基于 OECD 人工智能论文及相关引用文献整理。

注：这是为本报告分析目的选取的术语的非穷举且初步列表。

数据抓取组件主要包括数据采集、数据预处理和数据存储三大活动。

由于数据抓取涵盖多种技术，对其定义进行明确界定存在挑战。这些技术包括网页抓取 (web scraping)、网络爬取 (web crawling) 和屏幕抓取 (screen scraping)，每种技术都采用不同的方法并适用于不同的场景。为确保利益相关者对数据抓取形成共同理解，清晰界定并区分这些方法至关重要。明确的定义也能为遵守知识产权法律及必要时进行法律执行提供清晰的指引。

数据采集

通过数据抓取进行数据采集采用了多种技术，包括网页抓取、网络爬取、屏幕抓取及其他方法。这些技术通常涉及重要的知识产权考量，而这些考量会因所收集数据的性质及收集方式而异。与数据抓取相关的详细知识产权影响将在本报告第 3 节中讨论。

• 网页抓取 (Web Scraping) 与 网络爬取 (Web Crawling)：

这是两种密切相关但不同的信息提取方法。

网页抓取 通常涉及使用专门软件（如爬虫、机器人和抓取工具）从第三方网站或社交媒体平台提取特定数据。该过程包括向网站发送请求、解析网站的 HTML 内容，并从中提取所需信息。网页抓取的范围各异，可能收集整个网页内容，也可能仅针对特定数量的第三方网站数据，具体取决于其目的（美国国家医学图书馆，2024）。

网络爬取 则是指通过超链接自动导航和索引网页的过程。该技术使用网络爬虫（也称为蜘蛛或机器人）系统地从网页、数据库或其他数字来源收集数据，通常通过爬取内容并提取信息 (Shayne Longpre, 2024)。搜索引擎常用此技术索引网站内容。如第 3 节将讨论的，部分人工智能数据聚合器依赖链接而非下载和存储大型数据集。例如，LAION（大规模人工智能开放网络）主要通过链接访问外部托管的数据源来提供数据集 (Guadamuz, 2023)。

网站通常通过服务条款中的合同保障措施或技术保障措施（如机器人排除协议 robots.txt）来限制数据提取。该协议允许网站所有者向网络爬虫发出机器可读指令，要求其不要收集特定数据。尽管许多网络爬取实践遵守这些措施，但其他实践可能绕过或忽略这些保护。例如，某些人工智能驱动的系统在收到特定用户请求时，可能会选择性地忽略这些限制，引发了关于当前限制措施有效性的担忧 (Mehrotra & Courts, 2024)。

• 屏幕抓取 (Screen Scraping)：

与网络爬取不同，屏幕抓取涉及自动提取屏幕上视觉显示的数据，如文本、图像、视频、PDF 和其他文档格式。它并非直接与网站的底层文本或 HTML 代码交互，而是捕获并处理在用户设备上渲染的视觉输出。这可能涉及多种方法，从简单的图像捕获和文本识别到更复杂的模拟用户与图形界面交互的过程 (S. Gillis, Screen

Scraping, 2023)。该技术在消费服务领域（如开放银行计划和贷款应用程序）尤为常见。然而，许多司法管辖区正转向更安全的基于应用程序接口（API）的系统（澳大利亚政府财政部，2023）。API 允许用户在预定义的运营和法律参数内请求数据下载（如维基百科所示）。API 的使用通常受合同协议约束。**当通过 API 在符合适用合同条款的情况下获取数据时，不属于数据抓取的范畴，后者通常发生在数据抓取者与数据托管实体之间缺乏直接协调的情况下。**

- **其他技术:**

由于训练先进人工智能系统对海量数据的迫切需求，人工智能数据抓取技术正不断扩展。有研究预测，人工智能开发者可能在 2023 年至 2026 年间耗尽高质量语言数据（Villalobos 等，2022）。这促使了创新方法的出现，例如利用语音识别工具转录公开可访问的视频以生成额外的 AI 训练数据（Metz, Kang, Frenkel, Thompson, & Grant, 2024）。

数据预处理与存储

一旦收集或聚合了所需的抓取数据，通常需要对其进行处理和存储。对于许多类型的 AI 系统，原始数据必须经过筛选和转换才能用于模型训练。因此，抓取的数据通常需要进行**预处理**，转换为适合 AI 训练的可用格式。组织会根据收集的数据类型（图像、文本、视频、联系信息等）采用多种预处理方法。

抓取的数据通常也需要**存储**，常被置于结构化数据库中。例如，Meta 过去曾为不同任务使用不同的存储系统，后将其整合为一个更大、更统一的系统，能够存储海量数据，达到艾字节（exabyte，相当于 10^{18} 字节）规模（Engineering at Meta, 2021）。

抓取数据的使用

最后，如图 1 所示，组织可能以不同方式使用抓取数据，例如用于 AI 模型和系统的训练、微调、验证和/或测试。下一节将进一步描述这些具体用途。

二、人工智能数据抓取生态系统

人工智能数据抓取生态系统涵盖了商业、非商业及政府等各类实体，这些实体参与并/或受益于数据抓取活动。该生态系统中的核心参与者可被划分为不同类别，各群体承担特定角色并面临独特挑战。

研究机构与学术界

研究机构和大学在人工智能数据抓取生态系统中扮演重要角色。这些实体通常利用数据抓取技术收集海量数据用于学术和科学研究。研究人员利用这些数据开展研究、开发新的人工智能模型并优化分析方法，从而推动科学知识进步并助力前沿技术的发展。例如，抓取的数据已被用于提升可持续性分析和改进气候建模（Guttridge-Hewitt, 2023）。

尽管其研究目标通常合法，但这些机构和研究人员在抓取数据时可能面临复杂的法律挑战，例如应对版权和数据隐私法规。部分司法管辖区为研究目的或合理使用原则提供特殊例外，但在涉及国际数据集且受多项法律标准约束的情况下，导航这些法律框架可能极为复杂。

AI 数据聚合商

数据抓取生态系统中包含聚合并向第三方提供抓取数据的实体（AI 数据聚合商）。AI 数据聚合商可能以开源和/或付费方式提供数据。部分 AI 数据聚合商，如 Common Crawl、LAION 和 EleutherAI，是非营利组织，其在官网上免费提供抓取数据（Common Crawl, 2024; LAION, 2024; EleutherAI, 2024）。Common Crawl 成立于 2007 年，最初旨在爬取网络并为研究人员和小企业提供可访问的数据，这在生成式人工智能近期兴起之前就已经开始。AI 数据聚合商也可能爬取和/或使用数据用于自身目的。

Common Crawl 提供了一个庞大且免费开放的网络爬取数据仓库，被广泛用于预训练大型语言模型（LLMs）（Baack, 2024）。该数据在 LLMs 发展中扮演重要角色，如其在训练 GPT-3 中的应用所示，其中超过 80% 的令牌源自 Common Crawl（B. Brown, et al., 2020）。EleutherAI 对这部分网络爬取数据进行处理和精炼，生成如 Pile-CC 等专用数据集，专门用于训练大型语言模型。这些努力旨在减少不希望出现的内容，并针对更具体的 AI 训练目的对数据进行定制（EleutherAI, 2024）。类似地，LAION 基于 Common Crawl 数据创建了 LAION-400M 等专用数据集。这些数据集专门为人工智能应用（如图像和文本识别模型）设计，进一步展示了处理后抓取数据在人工智能模型训练中的价值（LAION, 2024）。

开源数据集如 MNIST、ImageNet 和 Open Images 可公开访问，并以开放许可证条款提供（Macgence, 2024）。开源数据聚合平台如 Hugging Face 或 GitHub 在其网站上列出或提供链接，供用户免费下载开源第三方数据集（Hugging Face, 2024）。部分商业提供商也提供现成数据集供购买。

然而，对于许多 AI 数据聚合商而言，数据集的获取方式及所采用的许可协议仍不明确。“数据溯源倡议”（Data Provenance Initiative）对超过 1800 个广泛使用的数据集进行了审计，结果显示，70% 的数据集缺乏关于许可用途的信息，或被标记为比作者意图许可更为宽松的许可类型（Data Provenance, 2023）。总体而言，关于抓取数据来源的透明度和文档记录存在不足（Tiku, 2023）。这种情况导致了针对通用人工智能系统的一些透明度要求，例如在欧盟人工智能法案中（如下文第 4 节所述）。透明度还可通过各种政策工具（如第 5 节中关于初步考虑和潜在政策的讨论）得到提升。涉及多个 AI 数据聚合网站的争议已浮出水面：

- 调查性新闻报道指出，一个流行的训练数据集 ‘Books3’ 包含超过 17 万本盗版书籍，并被用于训练 Meta 的 Llama、Bloomberg 的 Bloomberg GPT、EleutherAI 的 GPT-J 以及其他生成式 AI 模型（Reisner, 2023）。
- 谷歌的 C4 数据集据称包含从 1500 万个网站抓取的数据，并被用于训练 Meta 的 Llama、谷歌的 T5 以及其他大型语言模型（LLMs）。对 C4 数据集的科学

分析表明，其大部分内容源自“新闻、娱乐、软件开发、医学和内容创作”类网站（Schaul, Chen & Tiku, 2024）。

- 关于从网站抓取大量数据的额外知识产权（IP）担忧也已浮现，这些网站以盗版和假冒商品闻名或与之相关（Schaul, Chen & Tiku, 2024）。

科技公司与平台运营商

科技公司和平台运营商（如社交媒体平台、搜索引擎和电子商务网站）既是数据抓取的来源，也是数据抓取生态系统中的积极参与者。这些平台是抓取者的常见目标，因为它们托管着海量的用户生成内容和其他有价值的数据，这些数据常被用于开发 AI 模型。

平台运营商常会实施反抓取技术（如验证码和 IP 地址屏蔽）及服务条款限制，以规范或阻止未经授权访问其数据。这些措施旨在保护平台数据资产，同时解决知识产权和隐私问题，并保障用户权益。

此外，许多公司自身也从事数据抓取活动，收集数据以提升其产品和服务。数据抓取支撑着这些平台的多个核心商业模式，包括搜索引擎、账户或网站聚合、价格比较工具以及定向广告（Fei, 2024）。例如，LinkedIn 已承认使用抓取的数据来改进其自身服务，这反映了科技公司在数据抓取生态系统中常扮演的双重角色（Wiggers, 2024）。

知识产权作品的创作者与所有者

内容创作者（包括作家、摄影师、记者和艺术家）被视为人工智能数据抓取生态系统中受影响最直接的群体。他们的作品常被抓取并用于训练人工智能系统的数据集，且未经其知情或同意，尽管在某些司法管辖区此类行为可能被视为版权侵权。部分内容还可能受商标保护。

然而，创作者往往难以确定自己的作品是否被纳入训练数据集，因为人工智能开发者通常缺乏透明度。斯坦福大学基础模型研究中心的一份报告指出：大多数先进人工智能系统的开发者对数据的来源和合法性保持模糊，仅有 14 个开发者中的 1 个披露关于数据创作者、版权状态和数据许可的详细信息（Bommassani 等，2024）。

数据访问透明度显著恶化，从 2023 年 10 月评估的模型中占比 20%，骤降至 2024 年 5 月评估的模型中占比 7%。这一大幅下降可能反映了围绕版权的法律不确定性，以及开发者对披露用于构建 AI 模型数据集风险的日益担忧，特别是当这些数据集可能包含受版权保护或非法内容时。

这种缺乏透明度给下游开发者和/或部署者带来了挑战，使其无法验证是否符合许可和/或法律要求，从而加剧了与数据来源相关的问题。它还使得创作者和权利持有人难以追踪其知识产权的使用情况，并在适用情况下授权其使用并评估其权利是否受到侵犯。在某些司法管辖区，AI 开发者可能在特定法律框架下获准使用受版权保护的作品，但在其他司法管辖区，权利持有人可能有法律途径寻求救济，并在适用情况下追求更强的保护和/或对受保护作品使用的适当补偿。

为解决这些问题，越来越多的呼声要求出台政策措施，要求人工智能开发者和人工智能数据聚合商披露其训练数据的来源。实施此类透明度措施需要评估在训练数据集中识别个体数据源的技术可行性，以及对人工智能开发者的相关成本，同时考虑权利持有人和人工智能开发者在经济激励和创新成果方面的利益。

其他正在考虑的方案包括：允许权利持有人明确同意或选择不参与数据使用；以及通过许可协议和集体权利管理组织来管理权利的系统。

三、数据抓取的法律环境与日益增多的诉讼

由于抓取的数据可能包含受知识产权保护的材料（包括个人数据等），数据抓取行为可能引发多重法律风险。许多现行知识产权法律在制定时并未预见当前的数据抓取实践，导致

不同司法管辖区在适用法律时面临诸多不确定性。这种复杂性因各司法管辖区法律差异而进一步加剧。

数据抓取可能涉及多种知识产权及类似权利,具体取决于被抓取数据的性质及其使用方式。主要可能涉及的权利包括:

- **版权:** 保护原创作品(如文字、图像、音乐、视频)。未经许可提取受版权保护的材料可能构成侵权。部分司法管辖区虽存在版权例外(如合理使用或文本数据挖掘),允许特定情况下的未经授权使用,但其适用范围差异显著。版权问题还涉及删除权利管理信息(RMI,即元数据)及规避技术保护措施(TPMs)。

- **数据库权利:** 部分司法管辖区(如欧盟)为保护数据库编纂投资提供特殊权利(欧盟,1996)。提取受保护数据库的实质性部分可能侵权。该权利独立于版权,保护的是数据组织与编排的投入,即使单个元素不受版权保护。

- **商标权:** 抓取包含商标标志、名称或品牌标识的数据,若用于商业目的并可能暗示商标持有人的认可或关联,可能导致消费者混淆或商标淡化,从而构成侵权(世界知识产权组织)。商标法保护在商业中使用受保护标志的行为。

- **商业秘密:** 指为企业提供竞争优势的保密信息(如专有算法、训练数据集、AI开发/部署技术)(WIPO,2020)。保护取决于“信息权利人采取的合理保密措施”(WIPO,2024)。未经授权访问或披露此类信息可构成侵权。

- **肖像权与姓名权:** 抓取包含个人肖像、姓名或声音(尤其是公众人物)的数据,可能侵犯旨在保护个人身份要素免受未经授权商业使用(有时包括非商业使用)的权利。该权利在AI生成内容模仿真实个人的背景下日益重要。

- **道德权利:** 允许创作者主张作品作者身份并反对有损其声誉的修改(根据《伯尔尼公约》第6bis条实现协调)(世界知识产权组织,2024)。即使著作权转移,道德权利仍可阻止损害创作者声誉的作品使用方式(如AI生成内容中的作品扭曲)。

版权与国际义务

对受版权保护材料进行数据抓取,引发了该行为是否构成侵权的疑问。世界贸易组织《TRIPS协议》及世界知识产权组织(WIPO)管理的条约(如《WIPO版权条约》、《WIPO表演和录音制品条约》,统称“WIPO互联网条约”),为版权等知识产权设定了国际义务框架,旨在更新补充《伯尔尼公约》和《罗马公约》。关键规定包括:

1. **禁止形式主义:** 版权保护自动生效,无需注册(此原则在“退出机制”讨论中被提及)。
2. **三步测试:** 规定版权例外/限制的适用条件:a) 仅限于特殊情况;b) 不与作品正常利用相冲突;c) 不得不合理地损害权利人合法利益。
3. **权利管理信息(RMI)保护:** 保护嵌入的元数据。
4. **技术措施(TMs)保护:** 防止规避密码等技术保护。
5. **数据汇编保护:** 对数据或其他材料的汇编提供单独保护。

版权侵权的分析框架

判断大规模抓取受版权保护数据是否构成侵权,通常需分析以下问题:

- 抓取行为是否涉及复制等受著作权专有权利控制的活动,构成直接或间接侵权?
- 抓取过程是否删除了权利管理信息(RMI)或规避了技术保护措施?
- 是否存在可适用的版权例外或抗辩理由(如“合理使用”、“文本数据挖掘”)?
- 抓取行为发生地、适用法律及侵权认定地为何?不同司法管辖区的法律差异显著。

司法管辖区的差异与例外

第三方将受版权保护材料用于AI训练数据,在某些情形下可能依法允许,具体取决于

司法管辖区及个案事实。例如，允许依据可能包括：

- 美国版权法的“合理使用”例外（美国政府出版办公室，2010 年）
- 欧盟的“文本和数据挖掘（TDM）”例外（欧洲议会，2018 年）
- 日本版权法的相关规定（日本法务省，2024 年）
- 以色列版权法的“合理使用”规定（以色列法务部，2022 年）
- 新加坡版权法的“合理使用”和“TDM”规定（新加坡知识产权局，2022 年）

然而，不同司法管辖区的处理方式存在显著差异：

- **美国等（合理使用模式）：**提供一般性例外（如合理使用），法院在个案中根据多种因素（如 AI 输出是否构成对抓取数据的“转换性使用”）进行解释。
- **欧盟等（TDM 例外模式）：**有特定例外（如欧盟 TDM 例外）。用于研究的 TDM 通常无需权利人“选择退出”；其他 TDM 则允许权利人通过合同或技术手段（如机器可读指令）“选择退出”以保留权利。
- **日本：**《著作权法》授权商业/非商业 TDM（娱乐目的除外），但合同条款或技术措施可能优先于例外。
- **其他地区：**可能依赖公平处理或其他机制建立有限例外或抗辩。

许多司法管辖区已确立 TDM 例外，允许通过自动化分析生成洞见。但法律环境差异巨大，尤其在权利人的“选择退出”机制方面，导致法律解释不一致（详见附录 I 对选定司法管辖区的摘要）。这些差异，特别是在处理研究/科学目的的数据抓取时，为制定跨司法管辖区的协调方法带来挑战。

涉嫌侵犯版权的数据抓取诉讼在全球范围内正迅速增多。

数据抓取多年来一直是诉讼和政府执法的重点，涉及多家科技公司和人工智能开发者（Neuburger, 2022; Clarke, 2023）。随着大型语言模型（LLMs）的快速普及，聚焦于隐私、知识产权及其他问题的新法律纠纷不断涌现。包括美国和欧盟在内的多个司法管辖区已启动调查，并成立了特别工作组以应对人工智能驱动的数据抓取带来的挑战。例如，欧洲数据保护委员会（EDPB）成立了 ChatGPT 特别工作组，并发布了该小组活动及调查结果的首份报告（EDPB, 2024）。

其他数据抓取诉讼主要集中在美国，涉及知识产权、合同违约及侵权索赔。在美国，各类利益相关方，包括作者、新闻机构、音乐出版商和图像提供商，已提起诉讼（美国作家协会，2023 年；莱维、埃普斯坦和费尔曼，2024 年）。这些案件通常围绕数据抓取和使用数据训练人工智能模型是否构成版权侵权、合同违约或其他违法行为展开。部分诉讼还涉及 AI 生成的输出内容与输入数据高度相似的问题。美国以外地区也出现了相关诉讼，欧洲及其他地区正在审理涉及 AI 目的数据抓取全球影响的案件（Vincent, 2023; WIPO, 2024）。这表明，围绕数据抓取的法律问题在多个 OECD 人工智能文件中将持续呈现复杂性和动态演变。尽管诉讼最终可能在考虑所有相关利益相关者利益的情况下达成预期结果，但其成本高昂、耗时且可能带来不确定性。许多受损害方可能缺乏提起诉讼的资源，这也是考虑替代政策方案（例如第 4 节中概述的方案）的另一个原因。

数据抓取的管辖权复杂性

数据抓取引发的另一个法律难题在于确定解决侵权索赔的适当管辖权。数据抓取活动通常涉及多个司法管辖区——抓取指令可能在某一国家发出，被抓取的数据可能位于另一国家，而数据最终可能在第三管辖区域被使用、存储或处理。此外，基于抓取数据训练的 AI 模型可能在更广泛的区域内使用。这引发了复杂的国际私法问题，特别是确定侵权行为（如存在）发生地以及应适用何种法律。为应对这些挑战，需考虑数据抓取活动发生地的管辖权，以及可能提起版权侵权索赔的管辖区域。

欧盟《人工智能法案》引入了对数据抓取实践具有国际影响的要求，规定通用人工智能

模型提供者必须遵守欧盟版权法，即使模型在欧盟以外地区训练，只要其输出在欧盟市场使用（欧盟，2024）。该规定意味着在欧盟以外训练的人工智能模型，若其输出在欧盟内使用，则必须同时遵守训练所在司法管辖区的版权法律以及欧盟的版权法律。这突显了在人工智能系统日益跨国运营的背景下，采取国际协调的版权管理方法的必要性。

合同协议的作用

在应对数据抓取的管辖权复杂性时，合同协议在管理数据访问权限、减少或解决知识产权纠纷方面发挥着日益重要的作用。合同协议与知识产权之间的关系在人工智能数据抓取的法律框架中扮演着关键角色。合同，特别是服务条款和最终用户许可协议，通常用于规范数据如何被抓取或使用，包括明确允许的使用范围、归属要求以及责任分配等具体条款。然而，合同的可执行性及其与知识产权法律的互动在不同司法管辖区可能存在显著差异。

在许多司法管辖区，合同可约定变更协议各方之间数据抓取在适用法律下本可被允许的方式（即法定许可范围）。然而，在某些司法管辖区，此类合同条款可能并不总是具有可执行性。例如，在欧盟，权利人可通过合同条款对商业用途的数据抓取设置限制。然而，当数据抓取由研究机构或文化遗产机构为科学研究目的进行时，此类限制被明确禁止。合同条款的可执行性可能因其具体情境和适用法律而异。

在某些司法管辖区，用户为访问内容必须接受的标准网站协议条款（如使用条款）可能与各方单独协商的协议不同。人们常质疑标准格式协议（也称为“粘附合同”，即一方必须接受合同全部内容）是否能限制版权法的例外情形，如合理使用或其他许可用途。例如，以色列司法部 2016 年的一份法律意见指出，此类不可协商的合同不得用于阻止用户主张合理使用抗辩。单独协商的协议则可能适用不同的考量标准。

同样，合同可能无法始终凌驾于某些道德权利之上，此类权利在法国民法体系等国家尤为强劲。道德权利，如署名权或防止作品被贬损的权利，通常仍归原权利人所有，即使版权已通过合同转让或授权。在使用抓取数据训练生成与原作品高度相似的输出的人工智能模型时，这些权利可能带来额外的复杂性。例如，基于艺术家作品训练的 AI 模型可能生成类似的艺术风格，这可能引发关于原作品归属和完整性的道德权利争议，即使训练数据是通过合法合同安排获得的。

在这一不断演变的法律环境中，利益相关者必须谨慎处理合同法与知识产权法的交汇点，确保合同条款在不同适用法律框架下均具合法性。展望未来，如第五章所讨论的，标准合同条款和国际协调的政策方法可能为权利人、数据抓取者和数据使用者提供更大的确定性。此类协调可解决包括允许用途、归属要求、地域限制和责任分配在内的关键问题，同时保护所有相关方的权利。

政策回应

尽管政策回应——从行为准则到文件化和透明度要求——因地区而异，鉴于数据抓取纠纷日益增多，政策制定者正在探索多种政策解决方案，包括第 5 节中讨论的多边行为准则。一些举措也在特定管辖区内展开，其中部分内容总结于表 1。

为应对这些日益复杂的挑战，多个管辖区已开始解决与数据抓取相关的知识产权问题。下表 1 概述了部分司法管辖区针对人工智能的知识产权相关举措，以展示全球范围内采取的多元化应对策略。

表 1.部分司法管辖区针对人工智能的知识产权倡议摘要

国家	针对人工智能的倡议状况	举措详情
加拿大	关于负责任地开发和管理高级生成式人工智能系统的自愿行为守则 (2023)	自愿行为守则包括关于培训数据透明度的规定（加拿大政府，2023 年）。高级生成系统的开发者和管理者承诺努力实现以下成果：“[...]透明度 - 公布足够的信息，使消费者能够做出明智的决定，并使专家能够评估风险是否已得到充分解决。专家评估风险是否已得到充分解决。”
欧盟	欧盟《人工智能法》已经颁布，实施活动正在进行中；关键版权条款受《数字单一市场指令》（DSMD）（2024 年）的监管。	欧盟人工智能法案规定了通用人工智能模型提供商的合规义务，要求他们实施与欧盟版权法相一致的政策，包括使用最先进的技术来识别受版权保护的作品，并遵守《版权法》。 在 DSMD 中建立的 TDM 退出机制。通用人工智能模型的提供者还必须公开披露用于训练的数据集的详细摘要。为支持实施工作，欧盟人工智能办公室正在制定《通用人工智能行为准则》，该准则将提供以下方面的指导：透明度和版权相关义务。（欧盟，2024 年）。
以色列	部级指南（2023 年）	司法部宣布，合理使用例外一般适用于人工智能数据搜刮，除非人工智能模型的唯一目的是模仿单个艺术家。（以色列司法部，2022 年）。
日本	新报告（2024 年）	人工智能时代知识产权研究小组的临时报告指出，需要将法律、技术和合同结合起来，以促进人工智能技术的发展和知识产权的保护（人工智能时代知识产权研究小组，2024 年）。 人工智能时代的知识产权研究小组，2024 年）。
新加坡	2021 年修订《版权法》，引入基于目的的计算数据分析例外。 为计算数据分析引入基于目的的例外。该例外最近	自 2021 年 11 月 21 日起生效，新加坡的计算数据分析例外明确承认将受版权保护的材料用于机器学习（新加坡知识产权局，2022 年）。最近一次公众咨询征求了对禁止规避技术的临时例外情况的反馈意见。

针对人工智能生成输出的风格、肖像和宣传权索赔也在不断出现。

此外，关于人工智能生成内容的法律主张正在兴起，尤其聚焦于风格模仿、肖像权及公开权（publicity rights）等问题（Gervais, 2022; Henderson et al., 2023）。这类案例引发了核心关切：当 AI 生成内容与受保护作品存在实质性相似时，是否可能构成版权侵权或衍生作品？更广泛地，AI 系统模仿艺术家、作家等个体独特风格、声音或肖像的能力，引发了深远的伦理与经济争议。众多艺术家反对 AI 模型生成与其作品“风格”相似的内容，尽管这并非直接复制，但仍可能对其生计构成威胁（Wakelee-Lynch, 2023; Leffer, 2024; Willman, 2023）。同样，虚拟助手未经授权使用个人声音和肖像（尤其是高度模仿公众人物时）也加剧了隐私和同意方面的担忧（Mickle, 2024）。这些案例凸显了人工智能时代保护创意表达和个人身份所面临的系统性挑战。

值得注意的是，肖像权、风格权及公开权的法律保护在不同司法管辖区差异显著，且常分散于商标法、著作权法、肖像权法、隐私法或反不正当竞争法等多元法律框架中：

- 美国**：相关权利的范围和适用性复杂，因州而异，且常不被明确归类为传统知识产权范畴（美国版权局，2024）。
- 欧盟**：相关问题通常由道德权利（moral rights）和人格权（personality rights）规范，但成员国间存在显著差异。《伯尔尼公约》第 6bis 条确立的道德权利赋予创作者保护作品完整性和反对有损其声誉之修改的权利（Drexler et al., 2021），然而其在欧盟各国的认可与执行程度不一（Hutukka, 2023）。

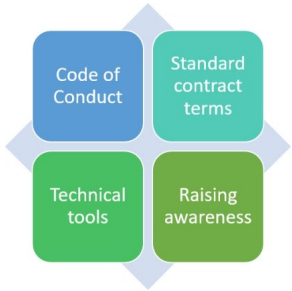
尽管本报告核心聚焦于数据抓取相关的知识产权问题，但围绕 AI 生成内容的上述争议深刻表明：随着人工智能技术的持续演进，知识产权法乃至更广泛的法律领域正面临日益严峻且复杂的挑战。

四、初步考虑和潜在政策措施

随着人工智能技术的快速发展,涉及数据抓取的知识产权及其他相关法律在多个领域可能需要进行审查并酌情更新。然而,实现这一目标可能需要一定时间,尤其是在政策制定者努力在不同法律领域平衡复杂且可能相互冲突的考虑因素时。在此过程中,政策制定者可考虑制定灵活且自愿的措施,以适应不同司法管辖区在法律和监管方面的多样化做法。这些措施,如《增强数据访问与共享的建议》(经合组织,《关于增强数据访问与共享的理事会建议》,2021 年)中所强调的,包括:

- 制定并采纳跨境“数据抓取行为准则”;
- 支持建立对人工智能数据抓取中常用术语的共同理解,该理解可在数据抓取行为准则或合同条款中引用;
- 支持开发标准化且广泛可用的技术工具,该工具可在“数据抓取行为准则”或标准合同条款中引用;以及,
- 实施提高利益相关者对自身权利和责任认识的举措。

图二：.应对人工智能数据挖掘中知识产权相关挑战的潜在政策方法



这些潜在方法可供政策制定者在考虑修改适用法律时参考。这些方法还可能为政策制定者提供灵活性,以在国际协调方式下全面应对数据抓取带来的知识产权、隐私及其他问题,同时考虑相关限制、例外及其他法律规定。

制定这些政策工具的过程可帮助政策制定者通过明确重要术语、制定支持创新同时有效保护知识产权的标准化定义和措施,遵循经合组织人工智能原则及其他相关多边努力(如七国集团广岛人工智能进程(日本政府,2024 年)),以协调方式应对数据抓取带来的挑战,同时考虑相关限制、例外及其他法律和政策差异。此类政策工具可设计为随时间适应和演进。

自愿行为准则以应对数据抓取带来的挑战

政策制定者日益转向行为准则和其他形式的自愿承诺,由企业采取行动应对数据抓取等挑战。例如,七国集团(G7)采纳了一项自愿行为准则,强调需要采取“适当的数据输入措施和个人数据保护措施以及知识产权保护”(G7, 2023)。一份自愿的“数据抓取行为准则”可在此基础上并与七国集团行为准则保持一致,通过明确一些在七国集团准则中提及的具体措施和保护措施。本节探讨了此类行为准则可能包含的潜在要素,如在相关论坛中推进该准则时可予以考虑。

经济合作与发展组织(OECD)《跨国企业负责任商业行为指南》(RBC 指南)[OECD/LEGAL/0144] 是一项自愿性、政府支持的框架,用于规范企业行为,涵盖商业伦理的所有领域(如人权和劳工权利、环境影响以及消费者保护)。RBC 指南还指出,各国政府期望企业采取措施,以防止和减轻与授予使用知识产权许可或自愿转让技术相关的负面影响(经合组织,2018 年)。更具体而言,《增强数据访问与共享建议》(EASD 建议)[OECD/LEGAL/0463] 提供了一种全面的数据治理方法,用于数据访问和共享(包括人工智能),呼吁遵守方“在适当情况下,促进自我或共同监管机制”(OECD,《理事会关于增强

数据访问和共享的建议》，2021 年）。这些机制可包括“自愿性指南、行为准则和数据访问与共享协议模板——在提供法律灵活性同时确保所有相关利益相关方明确适用法律法规”。欧洲委员会正致力于制定通用人工智能行为准则，以协助落实欧盟人工智能法案的要求，抢在正式标准出台前（欧洲委员会，2024）。

可起草一份自愿性的“数据抓取行为准则”，允许各方在不需承认或表态关于可能构成侵权或违反适用法律的活动类型的情况下采纳该准则。此方法可鼓励更广泛采纳，并已在其他领域成功应用。例如，2007 年，多个利益相关方采纳了《用户生成内容服务原则》，明确指出采纳这些原则的各方并未作出任何法律承认（UGC Principles, 2007）。这一方法可用于“数据抓取行为准则”。国际复制权组织联合会（IFRRO）制定了《复制权组织行为准则》（RROs）及 RROs 之间关系指南（IFRRO, 2024）。世界知识产权组织（WIPO）还发布了《集体管理组织工具包》，其中包含一份行为准则（WIPO, 2021）。

“数据抓取行为准则”可通过多种方式协助政策制定者。在一定灵活性和快速响应下，政府可认可该准则并建立已采纳准则的组织名录。该名录将帮助政府监督准则遵守情况。未履行遵守“数据抓取行为准则”承诺的行为可自愿报告至名录、经合组织人工智能事件监测系统（OECD.AI）或类似数据库（OECD.AI, 2024）。在必要时，部分司法管辖区可选择将未遵守自愿承诺视为对适用法律的法律违反，或选择将准则的部分内容强制执行。

准则还可包含针对个人或实体违反准则的情形（包括补救措施），或实体选择停止遵守准则的情形的规定。例如，目前正致力于开发机器学习技术，旨在移除不符合适用法律或政策的人工智能训练数据（Achille, Kearns, Klingenberg, & Soatto, 2023）。该准则可规定使用此类技术以帮助确保遵守知识产权或检测并缓解数据不符合法律或政策标准的情况。

自愿的“数据抓取行为准则”还可规定人工智能系统最终用户协议的条款……

“数据抓取行为准则”可能涉及多种行为和主体。准则的部分内容可**普遍适用于所有相关方**。这些规定可帮助引导人工智能系统最终用户避免某些提示或其他可能导致知识产权侵权或其他有害后果的行为。

例如：

确立标准定义：准则可确立标准定义，类似于经合组织《人工智能系统分类框架》（经合组织，2022 年）中包含的定义。这将有助于解决当前缺乏“数据抓取”标准定义所导致的不确定性，以及如上文第 1 节所述，该术语可能涵盖的各种活动。

包含广泛适用的原则：此外，准则可包含借鉴经合组织人工智能原则的、广泛适用的数据抓取原则。这些原则将通过保护隐私、知识产权和其他权利来鼓励负责任的创新。

阻止盗版数据：该准则还可**阻止开发和使用包含盗版数据的网站**。

提供高级支付框架：此外，它可为在适当情况下处理支付问题提供一个高级框架。该框架可考虑现有集体管理机制和其他知识产权支付框架。

建立违规报告系统：同样，该准则还可建立一个违规报告的注册系统和流程，如上所述。

促进标准化实践开发：为了扩展这一框架，该准则可促进**标准合同条款、技术工具及意识提升倡议**的开发（这些内容将在下文分别讨论）。标准化这些实践可帮助解决政策制定者的担忧，特别是关于服务条款和隐私政策变更允许收集更多人工智能训练数据，而个人并未真正理解这些实践或变更的情况（美国联邦贸易委员会，2024 年）。

除普遍适用的条款外，自愿的“数据抓取行为准则”可包含**针对特定活动的章节**，例如：

- (a) 与数据抓取相关的收集、预处理和存储活动；
- (b) 抓取数据的聚合；
- (c) 将抓取数据用于与人工智能系统相关的活动。

组织可同意遵守行为准则的一般规定以及所有与自身活动相关的具体规定。

...以及与人工智能数据抓取相关的收集、预处理和存储活动的具体规定

自愿的“数据抓取行为准则”可包含专门针对通过抓取方式**收集、预处理和存储数据**的条款。例如，它可以包含：

遵守相关合同和法律的承诺；

保留嵌入在抓取数据中的版权管理和其他类似信息的承诺；

避免绕过技术安全措施的承诺；

关于文档和透明度的条款：例如，遵守组织可以同意维护和/或披露有关其训练数据的某些信息。

目前，关于数据抓取实践的透明度有限。欧盟已通过《欧盟人工智能法案》针对通用人工智能模型解决了这一问题，要求提供商“制定并公开提供关于用于训练通用人工智能模型内容的足够详细摘要”（欧盟，2024 年）。为支持此要求，欧盟人工智能办公室将制定一个模板，用于此类披露。通用人工智能（GPAI）行为准则补充了这一要求，因为它将详细规定 GPAI 模型提供商如何遵守欧盟人工智能法案下的透明度和版权相关义务（欧洲委员会，2024 年）。

尽管不同司法管辖区可能采取从正式要求到自愿指南或其他灵活框架等不同方法来处理透明度问题，但协调共同实践可**简化人工智能开发者和权利持有者的合规流程**，同时在不同运营和监管环境中保持灵活性。“数据抓取行为准则”可与该模板及欧盟 GPAI 行为准则相协调，以帮助促进跨司法管辖区的透明度协调方法。这可简化 AI 开发者的合规流程，并通过协调最佳实践管理数据，潜在简化权利持有者的保护措施。

由于数据抓取涵盖多种活动，一份“数据抓取行为准则”可明确与政策目标一致的优先实践。具体而言，它可区分上述讨论的各种**活动类型和技术方法**。例如，如在第 1 节中讨论的，一些人工智能数据聚合商利用**链接与索引技术**使人工智能输入数据可用，而非主要依赖副本。在适当范围内，一份“数据抓取行为准则”可规范此类实践。

准则中包含的**标准定义**可帮助简化其实施并确保一致性，同时留有灵活性以适应不同的法律和监管环境。

针对人工智能数据聚合商的具体规定

“数据抓取行为准则”还可包含针对人工智能数据聚合商的规定，包括其合规性、文档记录及透明度相关活动。例如，该准则可提供一个机会，以帮助遏制在人工智能数据聚合商网站上提供盗版数据的行径。它还可能涉及对链接至此类数据行为的规范。一个关键方面可能是实施有效的控制机制，使权利人能够管理和保护其数据，确保遵守知识产权权利。

数据聚合商可被鼓励采用标准化透明度工具，如数据集卡片或来源文档，以提供关于数据集的清晰信息，包括许可条款及与数据使用相关的任何限制。通过为数据集创建者提供限制或批准访问受知识产权保护数据的能力，平台可在确保合法使用和提升透明度方面发挥关键作用。例如，Hugging Face 等平台已实施措施，如数据集卡片，详细说明数据集的许可、来源和预期用途，使第三方更容易控制其平台上托管的许可或受保护数据集的再利用（Hugging Face，2024）。反过来，对于第三方，这些卡片使他们能够了解自己的法律义务和限制。此外，受限数据集（gated datasets）通过在提供特定信息或获得批准前限制数据集访问权限，帮助权利持有人控制数据集的访问者（Hugging Face，2024）。类似方法可纳入“数据抓取行为准则”，确保聚合商对数据来源透明（包括是否涉及知识产权保护数据），并执行适当的使用权限。

该准则还应明确数据抓取活动导致知识产权或其他权利被侵犯时，应如何实施补救措施。在适当情况下，该准则可区分商业、政府及非商业主体，并考虑不同应用场景。准则制定过程应广泛征求利益相关方意见，以确保其促进公平竞争和创新。这包括权利持有人、既有的人工智能数据聚合商以及新兴的商业、非商业及其他领域的市场新进入者。此类包容性参与

可帮助防止准则无意中设置市场准入障碍。

对数据使用者的透明度要求

同时，还需对使用抓取数据的用户设定透明度要求。“数据抓取行为准则”可要求大型语言模型（LLM）开发者及其他使用抓取数据的用户（人工智能运营商）作出承诺。这些承诺应针对其在人工智能开发生态系统中的独特角色，同时基于上述讨论的透明度框架。人工智能运营商可被鼓励记录数据集的来源以及相关权利或限制。这可使开发者追溯数据来源并确保其符合知识产权规定。

在适用范围内，“数据抓取行为准则”还可考虑这些承诺在商业、非商业及政府行为体和不同应用场景中的差异。

表 2.人工智能开发者和运营商潜在数据挖据行为准则中与知识产权相关问题的初步术语

主题	行为准则的潜在原则
数据获取	人工智能运营商将承诺只从同意并遵守数据搜刮守则的实体和个人那里获取搜刮的数据。上文讨论的登记和报告机制可以帮助人工智能运营商遵守这一承诺。上文讨论的登记和报告机制可帮助人工智能运营商遵守这一承诺。
合规性	人工智能运营商可以承诺遵守适用的法律和合同，在数据被删除之前不更改或删除数据中嵌入的任何权利管理或类似信息。守则还可包括针对隐私和网络安全合规性和实践的具体规定（在另一份报告中讨论）。在另一份报告中讨论）。
道德方面的考虑	人工智能操作员应避免以非法方式刮擦或使用刮擦数据。政策制定者也可以选择解决某些道德问题。
使用限制 (研究和非商业用途)	如上文第 29. 页 “自愿帮助解决数据搜刮问题的行为守则”小节所述，可以用类似方式处理这些条款
其他限制	守则可以涉及其他可能的使用限制，包括是否以及在何种程度上可以进一步分发（以及向谁分发）刮除的数据。在可取的范围内，守则可以区分将刮擦数据用于培训、微调和其他目的。
文档和透明度	人工智能运营商可以承诺对其人工智能训练数据，尤其是剪报数据，进行适当的记录和披露。训练数据，尤其是剪报数据。例如，人工智能运营商可以考虑定期以标准化格式披露如何使用、存储、处理和共享刮擦数据。
技术保障	人工智能运营商可以承诺采用一些技术，以防止训练有素的人工智能模型在未经适当同意的情况下产生某些类型的人工智能生成输出，例如人工智能生成的输出 (i) 与刮擦数据实质上相似或混淆性相似，(ii) 模仿或实质上复制某人的姓名、图像、声音或肖像，(iii) 与刮擦数据实质上相似或混淆性相似，(iv) 与刮擦数据实质上相似或混淆性相似。姓名、图像、声音或肖像。技术措施也可用于帮助识别人工智能生成的输出结果。
非知识产权问题	守则可就非知识产权问题（在其他地方讨论过）提供指导，如隐私以及网络安全实践，并对此类实践进行记录和某些披露。如上文所述，守则还可涉及当数据搜刮违反（或导致违反）适用法律或政策时应使用的补救技术。
最终用户协议和行为	人工智能运营商可以同意尽合理努力教育用户了解这些禁令以及人工智能系统的适当用途。人工智能运营商可以做出合理努力，跟踪并适当处理违反这些条款的行为。守则可以更详细地阐述这些承诺。更多细节。

注：本表仅供参考，并非详尽无遗。来源：OECD AI；来源：OECD AI

技术工具在保护知识产权、赋能权利人更有效地控制数据访问以及支持许可机制方面发挥着重要作用。政策制定者可积极鼓励开发标准且易于获取的技术工具，以实现这些目标。欧洲数字经济与社会发展委员会（EASD）在[OECD/LEGAL/0463]建议中也呼吁各方推动采用此类工具，包括数据访问控制机制和增强隐私的技术。通过这些机制，数据能够在获得授权的用户之间以安全保密的方式进行访问和共享，同时结合具有法律约束力且可执行的义务，以保护数据主体及其他利益相关方的权利和利益（OECD，《关于增强数据访问和共享的理事会建议》，2021）。这些工具的开发可以借鉴各类标准组织（如万维网联盟 W3C 关于 TDM 保留协议的工作）的现有成果。

其中，robots.txt 协议作为常被提及的技术工具，被广泛用于指示网络爬虫避免抓取网站的特定部分，长期以来被视为限制数据抓取的一种机制。然而，其法律可执行性和技术约束力往往因具体情况而异。近期研究（Shayne Longpre, 2024）显示，自 2023 年中期以来，众多网站域名已采用 robots.txt 文件来应对与人工智能相关的数据抓取活动。尽管如此，该协议的应用存在显著不一致性：例如，知名先进 AI 系统开发者的网络爬虫常被限制，而不知名

实体则更容易规避此类限制；同时，新闻平台、论坛和社交媒体等行业比个人或小型电子商务网站更倾向于部署这些安全措施。此外，网站服务条款规定的限制与实际技术措施之间存在脱节，许多网站未能正确配置其 robots.txt 文件以准确反映合同约束。

为应对 robots.txt 协议因其自愿性而导致的在防止未经授权抓取方面有效性不足的问题，政策制定者可推动开发更具可执行性的数据访问控制工具。在多个司法管辖区，权利管理保障措施、透明度要求及相关工具仍有待通过标准化机制进一步发展。正如第四章所述，欧盟法规允许权利人在特定条件下拒绝出于商业目的的文本和数据挖掘（TDM）。开发标准化的退出工具以及其他机制（包括用于跟踪被抓取数据使用情况的标准机制），不仅能帮助权利人更便捷地保护自身权益，也有助于简化依赖数据抓取进行数据收集或使用的组织的合规流程。目前已有相关努力（如 Keller, 2024）致力于促进此类新型退出机制的开发。

任何此类技术工具的发展都必须谨慎权衡权利持有人、数据使用者以及公众在人工智能创新方面的更广泛利益。对这些工具进行适当标准化还可能带来额外益处：为全面解决知识产权、隐私及数据保护问题提供契机；同时，新工具的研发也有助于推动与人工智能训练数据相关的透明度实践的标准化，从而进一步激励负责任的人工智能数据共享。

表 3.人工智能运营商与知识产权相关的初步工具，用于潜在的数据刮擦行为准则

其他潜在工具功能	此类工具可满足的需求
有条件访问工具	有条件的数据访问是指允许数据访问或共享的工具，这些条件可能包括对获准访问数据的用户的限制（歧视性安排）、数据使用条件（包括数据可用于的目的）以及对数据访问控制机制的要求，通过这些机制授予数据访问权。政策制定者可鼓励利益相关者探索各种工具，使权利人能够轻松沟通跨多个平台或多个数据搜刮工具为数据搜刮目的访问数据的条件。政策制定者可以鼓励利益相关者探索一些工具，使权利人能够轻松沟通在多个平台上或通过多个数据搜刮工具访问数据的条件。
自动签约	合同可以使承诺和谅解具有可执行性。如下文所述，除了制定标准合同条款外，政策制定者还可以鼓励开发技术工具，促进自动缔约和自动监测合同遵守情况。这可以使权利人在必要时更容易寻求补救，特别是如果再加上适当的争端处理程序。然而，重要的是要包括人工监督和干预，特别是对于复杂或高风险的争议，以确保诉诸司法，特别是弱势群体。如果权利人拥有可靠的签约和执行机制，他们可能更愿意选择加入或有条件地选择同意。这可以借鉴 W3C TDM 权利保留协议的开发工作。
直接支付机制	为了帮助提高交易效率，政策制定者可能需要探索是否可以使用或开发一些机制来促进向权利持有人直接付款，包括对初始使用和可能允许的下游使用的直接付款。政策制定者还可以研究哪些激励措施有助于鼓励在这一领域开发解决方案。
监测遵守情况	政策制定者可能希望鼓励开发有助于监控数据搜刮代码和合同合规性的工具。
数据搜刮代码和合同	政策制定者可能需要鼓励开发一些工具，帮助监督数据采集代码和合同的遵守情况。

注：本表仅供参考，并非详尽无遗。来源：OECD.AI来源：OECD.AI

标准合同条款可为数据抓取行为的负责任发展指明方向

制定适当的标准合同条款需综合考量多个因素，并可通过建立一个全球多利益相关方协作机制来更好地整合不同观点。建立关于适当术语的共同理解，以促进标准合同条款的制定，可能是制定这些合同条款的必要第一步。在此过程中，考虑不同司法管辖区中已存在的技术、法律及其他术语将有所帮助。由于人工智能的跨境性质、知识产权的地域性，以及法律术语在不同司法管辖区可能具有显著不同含义的事实，共同术语的需求尤为重要。例如，“享受目的”这一概念在日本版权法中具有特定含义，而“补偿”和“报酬”等术语在不同法律体系中可能被解释为不同含义。如上所述，政策制定者在不同司法管辖区中可能对“数据”、“个人数据”、“数据处理”、“数据处理者”、“数据控制者”、“数据主体”、“数据主体权利”、“数据主体权利行使”等概念存在理解差异。

如上所述，美国、英国和日本的政策制定者已强调合同在为数据抓取划定负责任路径方面的潜在作用。全球人工智能伙伴关系（GPAI）知识产权咨询委员会关于负责任的人工智能数据和模型共享的项目也支持这一观点，特别是当标准合同条款与行为准则、技术工具和

教育相结合时 (Tiedrich, Lee; Avdulla, Alban, 2023)。

鉴于利益相关方的多样性以及不同领域和管辖区域的方法差异,制定一系列不同的标准合同条款以满足不同需求和谈判立场可能是有益的。这些条款可针对从非营利研究到商业应用等各种使用场景进行定制。重要的是,这些标准合同条款可作为可选的起点,并不排除组织在适当情况下协商定制安排的权利。

鉴于标准合同条款 (SCCs) 在其他领域的实用性,这一方法似乎具有潜力。SCCs 已帮助管理不同隐私法律体系国家间的跨境数据流动。标准开源许可和创意共享许可也促进了创新和中小企业 (SMEs) 的机遇 (创意共享, 2024)。许多公司也开始借助合同来管理数据权利,特别是用于 AI 训练数据的 (经合组织, 增强数据访问与共享: 在不同社会中平衡数据再利用的风险与收益, 2019)。主要平台如《纽约时报》、X、Zoom、Instacart 已更新其服务条款,专门针对 AI 数据抓取 (Ostwal, 2023) (Mehta, 2023) (Heath & Fried, 2023)。

此外,人工智能公司正与内容提供商建立直接合作关系:OpenAI 已与美联社 (AP News)、Shutterstock、新闻集团 (News Corp) 及《金融时报》(Financial Times) 达成协议 (O'Brien, 2023) (Reuters, 2024) (Shutterstock, 2023), 而谷歌 (Google) 则与 Reddit 签署协议 (Tong, Wang, & Coulter, 2024)。这些合同的范围尚未公开,但可能包括版权许可、数据访问权限以及与人智能训练数据获取相关的其他方面。

这些合同的重要性进一步体现在《人工智能训练声明》中,该声明由超过 6,500 名创作者签署。这一举措进一步强调了许可协议作为保护创作者权利同时允许其作品用于人工智能训练的工具的重要性 (《人工智能训练声明》, 2024 年)。

实施标准合同条款可能带来多重益处: 它们可帮助防止对个人及中小型企业 (SMEs) 的不公平待遇,如经合组织 (OECD)《关于增强数据访问和共享的理事会建议》(2021 年) 所认可。标准合同条款还可引用“数据抓取行为准则”和/或技术工具。最后,标准合同条款可基于政策目标制定,有助于将政策转化为实践。政策制定者还可就如何使标准合同条款公平合理以解决谈判能力中的重大不平等问题表达观点。

如 GPAI 知识产权咨询委员会的工作所强调的,制定标准合同条款将显著受益于多利益相关方合作。多利益相关方合作制定标准合同条款可能还导致跨司法管辖区的更大采用和实际应用。

提高对数据抓取及其法律和社会影响的认识也发挥着关键作用,在部分通过促进和鼓励负责任的行为。权利持有人、数据主体、数据生产者、数据持有者和数据使用者应获得关于如何保护和管理其权利的信息。例如,权利持有人可能通过了解与他们的权利相关的跨司法管辖区的当前法律环境以及可能帮助他们保护权利的新兴政策措施而受益。同样重要的是,人工智能数据生态系统中的其他利益相关者应了解与他们的数据相关活动相关的知识产权法律和政策。

此外,人工智能系统用户也需要教育,以了解如何帮助保护权利人。这可能包括如何避免触发可能绕过技术保障措施或侵犯权利的提示。人工智能运营商可以制作短视频或其他内容,向用户解释其系统应如何使用。如上所述,人工智能运营商还可将这些限制纳入其最终用户协议。

提高利益相关者意识也与其他经合组织政策建议相辅相成。根据经合组织人工智能原则强调人工智能系统透明度和可解释性的重要性,如原则 1.3 所述:“人工智能行为体应承诺在人工智能系统方面实现透明度和负责任的信息披露。这一承诺包括提供与上下文相关且符合最新技术水平的意义信息,以: i) 促进对人工智能系统、其能力及局限性的普遍理解; ii) 使利益相关者了解其与人工智能系统的互动,包括在工作场所中的互动; iii) 在可行且有用的情况下,提供清晰易懂的关于数据/输入来源、影响预测、内容、建议或决策的因素、过程和/或逻辑,从而使受影响方能够理解输出; 以及 iv) 提供信息,使因人工智能系统而受

到不利影响的各方能够挑战其输出”（经合组织，2024 年）。

EASD（注：应为 OECD《增强数据访问和共享建议》）还强调，“参与者应努力确保利益相关者充分了解其权利（包括获取信息和寻求救济的权利）、责任以及在违反隐私权、知识产权、竞争法或其他权利和义务时各自的法律责任”（经合组织，关于增强数据访问和共享的理事会建议，2021）。它进一步呼吁遵循者“促进与数据相关的技能和能力的发展，包括工人和公共服务人员，以充分利用数据访问和共享的益处。”

附件 A. 不同司法管辖区选定的版权例外规定

在日本,根据《版权法》,文本和数据挖掘(TDM)被广泛授权用于商业和非商业用途,特别是用于“非享受目的”(即内容不旨在满足智力或情感需求)(Ueno, 2021)。然而,技术保护措施(TPMs)或合同条款可优先于这些例外规定。该例外允许使用受版权保护的作品进行人工智能训练,从而促进机器学习等领域的创新。日本已发布两份关于人工智能与知识产权的报告:《日本人工智能与版权的一般理解》(由文化厅版权分科会下设的法律分科会发布)及人工智能与知识产权研究小组的临时报告(由知识产权战略总部秘书处设立的研究小组)。这些报告维持了现行文本采掘制度,同时寻求明确法律规则的适用范围。此外,报告鼓励企业采用技术工具降低知识产权侵权风险,并获取向创作者支付报酬的许可,即使该使用行为符合《版权法》规定(日本版权局,及知识产权秘书处, 2024)。

在英国,1988年《版权、设计与专利法》仅允许TDM用于非商业研究(例如非商业用途的例外条款)或经权利人许可。合同条款不得用于非商业用途的豁免。此外,英国知识产权局(IPO)已成立工作组,旨在制定自愿性版权与人工智能行为准则,以消除人工智能企业/用户面临的障碍并保护权利人权益,包括通过合同安排和行为准则等手段。然而,关于人工智能与版权的实践准则或其他方法尚未达成共识,未来可能出现包括立法在内的提案(科学、创新与技术部, 2024)。

欧洲联盟有两个关键工具来解决与人工智能相关的版权问题:关于数字单一市场中版权和相关权利的指令(指令(EU) 2019/790,“DSMD”)和《欧洲联盟人工智能法案》(“EU AI Act”)(欧洲联盟, 2024年)。DSMD引入了两项例外规定,允许在满足特定条件下使用受保护作品进行文本和数据挖掘(TDM)活动:

DSMD第4(3)条允许对合法可访问的作品进行TDM,用于商业和非商业目的,前提是实体具备合法访问权限(例如通过许可或公共可用性)。然而,权利人可通过合同、声明、机器可读手段或条款和条件排除此类使用。此排除机制仅适用于权利人明确保留此类使用权利的情况。

第3条在一般TDM排除制度中引入非商业性排除条款。欧盟非商业性排除条款专门针对科学研究。该规定允许研究机构和文化遗产机构开展科学研究。第3条在一般TDM退出机制中引入了非商业性例外。欧盟非商业性例外专门针对科学研究设计。该规定允许研究机构和文化遗产机构为科学目的进行TDM,涵盖自然科学和人文科学,前提是其对作品或主题内容具有合法访问权限。经合组织人工智能文件 38 | 基于抓取数据训练的人工智能中的知识产权问题与DSMD第4条规定的例外情形不同,著作权人无法通过任何合同或技术限制来排除或退出此例外情形。

由于DSMD是欧盟指令而非法规,因此不直接适用于欧盟成员国。成员国必须通过法律将该指令转化为国内法。因此,尽管目的上协调一致,但各国法律在实施退出机制方面存在差异。结果,没有统一的退出方式。一些成员国已将机器可读的退出机制强制化,另一些则将其作为可选项,还有一些根本未涉及此通信方式(Nobre, 2024)。此外,成员国在解释关键术语(如“研究”目的)以及适用例外条款(如非商业用途豁免)的条件时存在不一致。这种缺乏全面协调的情况有时会导致法律解释不一致,并在欧盟范围内统一适用这些例外条款时面临挑战(Hutukka, 2023)(欧洲委员会, 2024)。欧盟委员会通信网络、内容与技术总司正考虑更新TDM豁免机制,以提供一种标准化、机器可读的方式,更好地赋能权利持有人防止版权作品被用于训练AI模型。

欧盟人工智能法案(EU AI Act)通过聚焦通用目的人工智能模型提供者的责任,补充了DSMD。欧盟人工智能法案要求这些提供者实施政策以遵守欧盟版权及相关法律。这包括采用最先进技术识别受版权保护的作品,并遵守DSMD第4(3)条规定的文本采样豁免机制。

此外,提供者还需“制定并公开发布一份足够详细的摘要关于用于训练通用人工智能模型所使用的内容”(《人工智能法案》第 53(1)(d)条)。该摘要将公开披露,旨在使权利人能够有效行使权利。通用人工智能模型提供者还需列出“用于训练模型的主要数据集或数据集,例如大型私人或公共数据库或数据档案,并提供关于其他数据来源的叙述性说明”(第 107 条)。欧盟人工智能办公室将制定此类披露的模板。

欧盟人工智能法案的一个显著特点是其域外适用范围。它要求通用人工智能模型提供商遵守欧盟版权法,即使模型是在欧盟以外地区训练的,只要其输出在欧盟市场使用(第 2 条第 1 款第 c 项)。这意味着在欧盟以外地区训练的通用人工智能模型在欧盟市场使用时,必须同时遵守训练发生地管辖区的版权法和欧盟版权法。该规定强调了跨司法管辖区协调版权法律框架的重要性,并为服务于欧盟市场的 AI 提供商引入了新的合规要求。

新加坡 2021 年《版权法》引入了一项基于目的的例外规定,允许复制受版权保护的材料用于计算数据分析目的,适用于商业和非商业用途。该例外仅限于特定行为(复制,以及在狭义情况下,传播),且仅可用于计算数据分析目的。它明确限制将版权材料用于任何其他目的,否则该例外不适用于原用于计算数据分析的用途(《版权法》第 244(2)(b)条)。该例外适用于任何可能限制此类活动的合同条款。然而,为保护著作权人的商业利益,设定了以下条件:限制复制的分享,以及对复制作品用于特定目的的限制。

此外,TDM 活动还可能根据新加坡的一般“合理使用”例外(原称为“合理处理”例外)获得豁免。版权法明确规定,计算数据分析例外与合理使用例外相互独立运作(《版权法》第 184 条),这意味着 TDM 活动可能符合其中一项或两项例外规定,前提是满足各自的条件。

新加坡的生成式人工智能治理框架为全球对话提供了基础,以解决生成式人工智能的担忧同时最大化持续创新的空间。它强调了平衡版权与数据可访问性的重要性,并鼓励政策制定者促进所有相关利益相关者之间的开放对话,以理解快速发展的生成式人工智能技术的影响,并确保潜在解决方案平衡且符合市场现实(人工智能验证基金会;资讯通信媒体发展局,2024[55])。

最后,新加坡最近引入了一项新的集体管理监管制度,旨在促进更有效和透明的许可解决方案(新加坡知识产权局,2024)。

合理使用与合理处理

澳大利亚《版权法》规定了以下特定“合理处理”例外情形:研究或学习;批评或评论;讽刺或戏仿;新闻报道;以及提供法律建议/法律从业者、注册专利代理人或注册商标代理人提供专业建议(澳大利亚政府,1968)。这些例外情形不适用于所有类型的版权材料。《版权法》规定,为上述特定目的进行的“合理使用”可适用于以下版权材料:文学、戏剧、音乐或艺术作品;文学、戏剧或音乐作品的改编作品;以及视听作品。当使用作品、改编作品或视听作品的“实质性部分”或更多内容构成“合理使用”时,不构成对该特定版权材料的侵权。

加拿大《版权法》规定了两项可能适用于数据抓取的版权侵权例外情形:1) 用于研究的合理使用例外(第 29 条);以及 2) 用于技术过程的临时复制例外(第 30.71 条)。在一宗涉及文本和图像网络爬虫的诉讼中,加拿大法院适用公平处理原则(Century 21 Canada Limited Partnership v Rogers Communications Inc., 2011 BCSC 1196)。该案中,爬虫从网站抓取文本和照片以填充被告的自有网站。法院认定该行为构成侵权,但关于相同分析是否适用于其他目的进行的其他类型数据抓取活动,仍存在不确定性。

至于技术过程的例外情形,加拿大版权委员会将该规定解释为“旨在涵盖自动生成或未经用户直接控制的复制行为”,且此类复制在技术过程完成后自动删除(加拿大版权委员会,2016)。与合理使用原则类似,关于该例外情形是否及在何种程度上适用于技术数据抓取活

动，仍存在不确定性。例如，部分技术数据抓取活动可能需要制作临时副本，而其他活动可能需要复制作品。与合理使用情形类似，目前尚不明确该例外是否及在何种程度上适用于文本和数据挖掘活动。例如，部分文本和数据挖掘活动可能需要制作临时副本，而另一些活动则可能需要永久存储作品副本，这将使该规定不适用。

2023 年 10 月，联邦政府启动了关于版权与生成式人工智能的咨询（加拿大政府，2023）。此次咨询的目的之一是确定《版权法》是否应进行修订以明确其对 TDM 活动的适用范围，以及现有关于研究目的的合理使用例外和技术过程中的临时复制例外如何适应 TDM。在咨询期结束后，联邦政府的更新仍待定。

以色列的《版权法》采纳了以美国法律为模型的合理使用原则。以色列司法部发布了一份意见，确认其合理使用标准基本上涵盖用于训练 AI 模型的数据抓取，除非 AI 模型的唯一目的是模仿单一艺术家。此外，该意见不适用于机器学习过程的输出，即使机器学习过程本身不构成侵权，该输出仍可能侵犯版权（司法部，2022）。

韩国同样在其《版权法》中纳入了基于美国法律的合理使用原则（大韩民国，2017）。

美国《版权法》包含一项合理使用例外条款，允许在未经版权持有人同意的情况下，为批评、评论、新闻报道、教学（包括为课堂使用而复制多份）、学术研究或研究等目的，有限使用作品（17 U.S. Code § 107）。为了确定在版权侵权诉讼中是否适用合理使用例外，该法律列举了四项因素供法院考虑：使用目的和性质，包括该使用是否具有商业性质或用于非营利教育目的；受版权保护作品的性质；使用部分与受版权保护作品整体相比的数量和实质性；以及使用对受版权保护作品潜在市场或价值的影响。

在美国，涉及人工智能的版权及其他主张的重大诉讼已大量涌现，包括数据抓取是否属于法定合理使用例外。合理使用裁决聚焦于各案中呈现的具体事实。

《数字千年版权法》（DMCA）也适用于人工智能数据抓取。例如，DMCA 可对绕过“有效控制著作权作品访问的技术措施”的人员追究刑事和民事责任，具体行为包括“未经著作权人授权，规避、绕过、移除、停用或损害该技术措施”（17 USC 1201(a)）。DMCA 因此建立了一种与 TDM 法律中讨论的“选择退出机制”相似的制度，如上文所述。根据 DMCA 要求，为确定是否需要 17 USC 1201(a)(1)(A)（禁止规避有效控制访问受版权保护作品的技术措施）实施临时豁免，2021 年经过三年一次的规则制定程序后，目前对涉及学术研究和教学的 TDM 活动（37 CFR 201.40(b)(4)、(5)）。美国版权法还禁止故意删除版权管理信息（CMI），以及在 CMI 被删除后明知故犯地分发或进口作品（17 U.S.C. § 1202）。