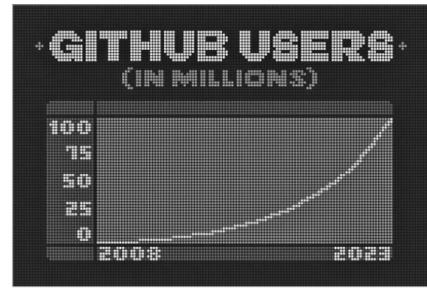


(上接第1版)

从六大版权诉讼案看 AI 时代的维权之路



2022年11月,四名程序员原告(代号“J. Doe”)针对 GitHub、微软和 OpenAI 发起集体诉讼,提出 22 项索赔,声称 OpenAI 开发并授权给微软的 AI 系统 Copilot 使用托管在 GitHub 上的开源软件,该软件建议其他程序员用户使用公共项目的片段。原告认为,此举侵犯了原作者的知识产权,被告未对其劳动成果给予相应认可,也没有遵守原始许可证的其他条款要求。

《数字千年版权法》(DMCA)第 1202 条侵权条款规定,不得在未经许可的情况下删除关键的“版权

案例 3 AI 图片生成软件公司被起诉

2022年,萨拉·安德森等3位艺术家指控 AI 公司 Stability AI、艺术家作品集平台 DeviantArt 和 AI 绘画工具 Midjourney 公司通过 AI 图片生成程序,爬取他们发布在网络上的图片作品,删除相关版权管理信息,并按照用户指令生成类似的 AI 图片,侵犯其包括著作权在内的多项财产权利。

在起诉状中,艺术家指控 Stability AI 公司向大规模 AI 开放网络 LAION 付费,抓取逾 50 亿张受版权保护的图片作为 AI 图片生成引擎 Stable Diffusion 的训练资料,并以此为基础开发了根据用户的文字指令生成相应 AI 图片的人机界面 DreamStudio; 指控 DeviantArt 的 DreamUp 工具使用的 Stable Diffusion 中包含原告受版权保护的作品; 指控 Stability AI 通过网络抓图方式训练图像应用 Stable Diffusion, 构成转授权。

该案件的核心指控包括: AI 模型训练涉嫌非法使用版权作品、移除版权信息(CMI)、利用艺术家姓名进行商业推广。之后双方围绕这些法律争议进行了多轮协议交锋。

2023年10月,美国北加州地区法院公布判决结

案例 4 《纽约时报》等诉 OpenAI 撕开 AI 数据抓取法律防线

2023年12月和2024年4月,《纽约时报》《每日新闻》分别对 OpenAI 的语言模型 LLM、ChatGPT 以及微软由 GPT-4 驱动的必应聊天工具 Bing Chat 提起诉讼,指控微软和 OpenAI 逐字逐句地复制其数百万篇文章,将其用于训练大模型。同时指控被告将 GPT 技术与微软必应搜索结合,使大语言模型以自然语言回答查询时绕过原告网站直接重述或引用原告作品,削弱了用户访问原告网站的必要性。《纽约时报》因此遭受了订阅、许可、广告和关联收入的损失,而被告节省了数十亿美元的成本,同时增加了自身的市值。

从法律依据看,《纽约时报》援引《美国版权法》,主张被告行为构成直接侵权、替代侵权、贡献侵权和增值侵权。还指控被告违反了《数字千年版权法》,删除或修改了版权管理信息。

因此,原告《纽约时报》对被告提出商标侵权、不

案例 5 美国作协诉 OpenAI 侵权 科技公司转向寻求合作

2023年9月,美国作家协会以及包括《权力的游戏》原著作者乔治·马丁在内的 17 位美国著名作家在纽约南区联邦法院递交起诉状,对 OpenAI 发起集体诉讼。至今,该案仍在审理中,很多作者不断向联邦政府和各州政府提出法律索赔。

原告在起诉状中称,OpenAI 在未经授权的情况下,使用原告作家的版权作品训练其大语言模型,使其大语言模型可以基于这些未经授权的小说输出相应结果,并可能生成总结、复述,以及模仿这些作品的衍生作品。原告基于美国《版权法案》,认为 OpenAI 侵犯了原告作家的版权并寻求获得赔偿。

美国作协提出了两点主张。一是 OpenAI 构成了版权侵犯。为证明被告进行版权侵犯,原告需要论证被告存在实际复制版权所有者作品,以及被告作品与原告作品中需要保护的元素存在实质性相似。为此,原告在起诉状中介绍了生成式 AI 及大语言模型的原理。其中描述了大语言模型“在定义上涉及到对于整部作品或作品中关键部分的复制”,以及“训练这一词汇是对复制和消化的委婉表述”。原告还提到书籍是训练大语言模型的高质量文本材料。

二是 OpenAI 存在主观故意侵犯。原告论证 OpenAI 存在对原告版权的主观故意侵犯。为论证

案例 2 程序员起诉 AI 公司侵权

管理”信息,应明确列出代码由谁编写以及使用条款。被告提出动议,要求法庭驳回指控并基本得到法官约翰·泰迦(Jon Tigar)的支持。在庭审中,泰迦法官引用研究中的观察结果,即 Copilot 在良性情况下很少直接提供记住的代码,大多数原样照搬发生在用户提示模型使用与训练数据非常相似的长代码片段的情况下,因此上述条款并不适用。这一判决导致原告撤回指控并作进一步修改。

事实上, GitHub 已于 2023 年调整 Copilot 相关设置,确保生成代码建议时对训练代码作出细微调整,以防止输出内容时被指控为直接照搬许可软件片段。

程序员坚称, Copilot 仍会生成与其亲手编写的代码完全相同的建议,这也成为他们诉讼期间的核心论点。他们在指控中提出,如果用户关闭 Copilot 的反复安全开关,即可引发非法代码复制。他们还引用一项关于 AI 生成代码的研究,用其中列举的 Copilot 剽窃源代码证据支持自己的立场。

但法官再次表示反对,称无法认定微软的系统

果,以清晰性和具体性不足为由,驳回原告的大部分起诉。法院认为,原告对 Midjourney 直接侵权指控不明,无法初步断定是否违法使用原告受版权保护的作品,因此要求 Midjourney 提供事实说明对其 AI 服务进行何种训练。法院要求原告修正诉状,说明被告指控侵权的衍生作品与原始作品有哪些相似之处,或包含原始作品的哪些受保护元素。

随后,2023~2024 年间法院多次开庭审理,针对不同指控作出裁决。下一次开庭时间为 2027 年 4 月。

针对《数字千年版权法》违规指控,法院认为,模型输出不等于复制品,《数字千年版权法》保护范围不延伸至衍生或重构内容,因此判定原告未能证明 AI 输出与原始作品“完全相同”,且未提供具体 CMI 被移除的证据链,驳回了《数字千年版权法》违规指控。

针对公开权与虚假背书这一指控,法院初步裁定, Midjourney 的“风格列表”与作品展示可能导致消费者混淆,虚假背书成立,索赔可继续审理。但法院认为,艺术风格本身不受保护,若直接使用姓名推广产品,需进一步举证“商业关联性”,因此部分驳回了公开权指控。

现阶段审判的关键是,原告要提供 AI 模型是否存储了“压缩版”版权作品作为直接侵权核心证据,此外,艺术家姓名使用是否实质影响消费者选择成为判定虚假背书的关键。

案解解读 该案原告的核心主张是被告的版权侵权及诱导版权侵权行为。法院要求原告和被告分别提供相关证据和事实说明。本案的裁决不仅是版权法的试金石,更将重塑 AI 与艺术创作的规则边界。此外, Stability AI 公司已向内容创作者作出让步,允许艺术家自主选择作品是否退出下一个版本的 Stable Diffusion。但是此“选择退出”条款受到艺术家诟病,认为它削弱了对版权人的保护。

正当竞争等主张,要求获得法定损害赔偿、实际损害赔偿、利润返还和律师费;并要求永久禁止被告继续侵权行为,销毁所有包含《纽约时报》作品的 GPT 或其他 LLM 模型在内的培训数据集。

美国纽约南区联邦法院今年 3 月作出裁决,驳回 OpenAI 请求撤销《纽约时报》等媒体版权诉讼核心指控的动议,案件的审理裁决围绕四大焦点展开。一是版权侵权与诉讼时效。OpenAI 曾主张《纽约时报》等媒体对 2019~2020 年的训练数据使用行为“早应知情”,但法院认为原告无法在三年前预见 ChatGPT 的侵权风险,故相关指控未过时效。此外,法院认可原告对“帮助性侵权”的初步举证,即 AI 工具可能被用户用于生成侵权内容,而 OpenAI 对此有理由知情。二是《数字千年版权法》与版权信息移除。针对《每日新闻》和调查报道中心(CIR)的指控,法院认定 OpenAI 在训练阶段使用特定算法剥离文章作者、版

权声明等信息,可能构成故意移除版权管理信息。此外, AI 输出的片段因不构成“完整作品副本”,所有关于“分发侵权副本”的指控均未获支持。三是商标淡化与州法争议。法院认定《每日新闻》等媒体已形成全国性驰名商标地位,支持其商标权益主张,驳回了 OpenAI 及微软的动议。四是被驳回的指控。CIR 关于“删节内容侵权”的主张因摘要与原文无实质相似性被驳回;所有原告的“普通法不正当竞争”指控因受《版权法》优先适用而被排除。

案解解读 本案的裁决意味着 AI 公司使用新闻内容训练模型是否构成侵权、生成式 AI 输出是否蚕食原创内容市场等争议,将进入实质性司法审查阶段。这场诉讼首次撕开了 AI 数据抓取行为的法律防线,为新闻业争取到了直面技术冲击的司法战场。随着案件的推进,关于“合理使用”边界的拉锯战或将重塑内容创作与 AI 开发的底层规则。

实际陈述部分强调了 OpenAI 通过 AI 模型获利的性质,介绍了 OpenAI 开发的几代 ChatGPT 产品,并再次强调了其通过收取会员费用进行技术盈利的特质。

原告论证了 ChatGPT 对原告作者的损害, ChatGPT 和大语言模型将严重威胁到作家的生计。原告列举了一些证据,证明 ChatGPT 将在未来取代大量的劳动力,并且作家通过版权作品、新闻报道及内容写作获得的收益将受到 ChatGPT 影响而减少。另外,原告认为, ChatGPT 生成的版权作品梗概为版权书籍的衍生作品, ChatGPT 还将生成其他类型的衍生作品,包括一些雷同的低质量小说,以及模仿小说作者口吻与 ChatGPT 对话的商业生态等。这类衍生作品也将对作家的生计和收入产生重要负面影响。

除以上的共同事实主张,原告在起诉状中还详细描述了每位原告成员的指控,来显示 OpenAI 使用版权保护作品训练大语言模型生成酷似作者原著的内容。

案解解读 以美国作协为首的原被告向科技公司发起一次声势浩大的挑战,此前美国作协曾与谷歌针对扫描图书案有过较量。虽然目前案件尚无进一步进展,但科技公司已经转向与出版商和作者合作,签署协议并获得对其版权作品的 AI 训练权。

以具有实际意义的方式剽窃他人成果,判定 GitHub、微软及 OpenAI 胜诉。2024 年,原告代理律师事务所约瑟夫·萨维里律师事务所及其共同律师向美国第九巡回上诉法院提起上诉。此次上诉的核心问题在于,根据《数字千年版权法》相关规定,责任是否仅限于“从作品的相同副本中删除或更改版权管理信息”。这直接影响到科技公司使用受版权保护的材料进行 AI 模型训练时的法律责任界定。至今此案仍在审理中。

案解解读 这一诉讼结果将对整个科技行业产生深远影响。如果法院认定 GitHub 等公司的行为构成侵权,那么将对科技公司使用受版权保护的材料进行 AI 训练提出更高要求,他们需要更谨慎地处理版权问题,以避免潜在的法律风险,因此将加大版权合规方面的投入, AI 技术的发展速度将受到一定影响。如果法院认为 GitHub 等公司的行为不构成侵权,那么将为科技公司使用受版权保护的材料进行 AI 训练提供一定的法律依据,并推动 AI 技术的更快发展。



艺术家萨拉·安德森创作的漫画被 Stability AI 用于生成相似漫画

了公开权指控。

现阶段审判的关键是,原告要提供 AI 模型是否存储了“压缩版”版权作品作为直接侵权核心证据,此外,艺术家姓名使用是否实质影响消费者选择成为判定虚假背书的关键。

案解解读 该案原告的核心主张是被告的版权侵权及诱导版权侵权行为。法院要求原告和被告分别提供相关证据和事实说明。本案的裁决不仅是版权法的试金石,更将重塑 AI 与艺术创作的规则边界。此外, Stability AI 公司已向内容创作者作出让步,允许艺术家自主选择作品是否退出下一个版本的 Stable Diffusion。但是此“选择退出”条款受到艺术家诟病,认为它削弱了对版权人的保护。

权声明等信息,可能构成故意移除版权管理信息。此外, AI 输出的片段因不构成“完整作品副本”,所有关于“分发侵权副本”的指控均未获支持。三是商标淡化与州法争议。法院认定《每日新闻》等媒体已形成全国性驰名商标地位,支持其商标权益主张,驳回了 OpenAI 及微软的动议。四是被驳回的指控。CIR 关于“删节内容侵权”的主张因摘要与原文无实质相似性被驳回;所有原告的“普通法不正当竞争”指控因受《版权法》优先适用而被排除。

案解解读 本案的裁决意味着 AI 公司使用新闻内容训练模型是否构成侵权、生成式 AI 输出是否蚕食原创内容市场等争议,将进入实质性司法审查阶段。这场诉讼首次撕开了 AI 数据抓取行为的法律防线,为新闻业争取到了直面技术冲击的司法战场。随着案件的推进,关于“合理使用”边界的拉锯战或将重塑内容创作与 AI 开发的底层规则。

实际陈述部分强调了 OpenAI 通过 AI 模型获利的性质,介绍了 OpenAI 开发的几代 ChatGPT 产品,并再次强调了其通过收取会员费用进行技术盈利的特质。

原告论证了 ChatGPT 对原告作者的损害, ChatGPT 和大语言模型将严重威胁到作家的生计。原告列举了一些证据,证明 ChatGPT 将在未来取代大量的劳动力,并且作家通过版权作品、新闻报道及内容写作获得的收益将受到 ChatGPT 影响而减少。另外,原告认为, ChatGPT 生成的版权作品梗概为版权书籍的衍生作品, ChatGPT 还将生成其他类型的衍生作品,包括一些雷同的低质量小说,以及模仿小说作者口吻与 ChatGPT 对话的商业生态等。这类衍生作品也将对作家的生计和收入产生重要负面影响。

除以上的共同事实主张,原告在起诉状中还详细描述了每位原告成员的指控,来显示 OpenAI 使用版权保护作品训练大语言模型生成酷似作者原著的内容。

案解解读 以美国作协为首的原被告向科技公司发起一次声势浩大的挑战,此前美国作协曾与谷歌针对扫描图书案有过较量。虽然目前案件尚无进一步进展,但科技公司已经转向与出版商和作者合作,签署协议并获得对其版权作品的 AI 训练权。



2023年7月,理查德·卡德雷等作者向美国加州南区法院提交起诉,指控 Meta 将数百万作者作品用于训练大语言模型 LLaMA, 同时指控 Meta 从安娜的档案(Anna's Archives)和 LibGen 等盗版网站获得版权保护内容,提出版权侵权等多项索赔。LibGen 存储库包含数百万盗版书籍,并通过点对点种子网络进行分发。Meta 则辩称其行为属于“合理使用”,旨在通过学习生成原创内容。

2023年9月, Meta 提出驳回动议,请求驳回原告除版权直接侵权以外的其他索赔。对于版权间接侵权指控,被告认为,原告未能证明语言模型的表达与原告书籍存在实质性相似,且原告未能指出具体的侵权输出。对于移除版权信息及提供虚假版权管理信息(CMI)的指控,被告认为,原告没有具体指明被告移除或者提供了哪些虚假 CMI, 因此未能提出有效主张。

2023年11月,法院批准了被告的 2 项驳回动议。原告主张被告模型是侵权的衍生作品,因为模型不可能在没有从原告书籍中提取表达信息的情况下运作。法院认为,演绎作品是“基于一个或多个先前存在的作品”,以“作品可以被重铸、转换或改编的任何形式”存在。原告主张,“被告大语言模型的每个输出都是侵权的衍生作品,每个输出行为都构成了版权间接侵权行为”。原告认为不需要证明输出和权利书籍之间存在任何相似之处,因为其书籍在模型训练过程中被完整复制。法官援引艺术家诉 Stability AI 等案 2023 年 10 月的裁定(被指控侵权的衍生作品仍必须与原始作品有一些相似之处,或包含原始作品的受保护元素)认为,原告在诉状中没有对任何输出内容提出指控,不可能有间接侵权,原告需要明确并证明输出“以某种形式包含”原告书籍的一部分。

法院驳回了原告提出的《数字千年版权法》违规索赔,因为没有事实支持 Llama 曾分发过原告的书籍,也没有在移除版权信息的情况下分发。此外,因版权先占原则,还驳回了基于版权直接侵权行为提出的 UCL 索赔、不当得利和疏忽主张。

2024年8月,原告提交第二次修订诉状,仅保留版权直接侵权指控。案件目前仍处于证据发现程序,没有更新进展。

此案的核心争议在于 Meta 训练大模型的做法是否符合合理使用原则。Meta 声称,使用受版权保护的作品来训练生成式 AI 模型属于“转换性使用”,即通过新的方式或目的使用原作,从而产生不同的价值,符合合理使用第一要素。而作者方面认为,使用版权作品训练生成式 AI 模型不具有“转换性使用”的性质,因为这种使用方式与教育人类作者没有本质区别。此外,这种训练的目的是创建与被复制作品在同一市场竞争的作品,使这种使用方式具有“商业性”。

今年 3 月,美国加利福尼亚州北区联邦地区法院法官文斯·查布里亚(Vince Chhabria)裁定,侵犯版权的指控显然是足以成立的具体伤害,而 Meta 删除版权管理信息的做法为其侵犯版权的行为提供了便利和隐瞒,足以构成诉讼的具体损害。《美国法典》相关规定要求,未经版权所有者或法律授权,任何人不得故意删除或更改任何版权管理信息。原告还提出, Meta 采取其他措施降低 Llama 生成输出的可能性,这些输出会揭示或表明受版权保护的材料包含在训练数据集中。这些指控提出了一个“合理的,如果不是特别有力的推论”,即 Meta 删除了版权管理信息以阻止 Llama 输出这些版权管理信息。

查布里亚认为,目前最大的问题在于 Meta 的 AI 工具是否会损害图书销售,或给作者带来损失。他在听证会上表示,尽管愿意让案件继续推进,但对这些特定指控持保留态度。

案解解读 法官认定, Meta 删除版权管理信息为其侵犯版权行为提供了便利和隐瞒,足以构成诉讼的具体损害,是对受版权保护的作者及其作品利益的维护,体现了在 AI 领域司法实践的新发展。但是,原告需要提供充分数据证明其对 Meta 输出内容的指控,这对于判断 Meta 使用受版权保护作品训练大模型是否符合合理使用原则至关重要,也成为法院庭审时一直悬而未决的焦点。

2023年7月,理查德·卡德雷等作者向美国加州南区法院提交起诉,指控 Meta 将数百万作者作品用于训练大语言模型 LLaMA, 同时指控 Meta 从安娜的档案(Anna's Archives)和 LibGen 等盗版网站获得版权保护内容,提出版权侵权等多项索赔。LibGen 存储库包含数百万盗版书籍,并通过点对点种子网络进行分发。Meta 则辩称其行为属于“合理使用”,旨在通过学习生成原创内容。

2023年9月, Meta 提出驳回动议,请求驳回原告除版权直接侵权以外的其他索赔。对于版权间接侵权指控,被告认为,原告未能证明语言模型的表达与原告书籍存在实质性相似,且原告未能指出具体的侵权输出。对于移除版权信息及提供虚假版权管理信息(CMI)的指控,被告认为,原告没有具体指明被告移除或者提供了哪些虚假 CMI, 因此未能提出有效主张。

2023年11月,法院批准了被告的 2 项驳回动议。原告主张被告模型是侵权的衍生作品,因为模型不可能在没有从原告书籍中提取表达信息的情况下运作。法院认为,演绎作品是“基于一个或多个先前存在的作品”,以“作品可以被重铸、转换或改编的任何形式”存在。原告主张,“被告大语言模型的每个输出都是侵权的衍生作品,每个输出行为都构成了版权间接侵权行为”。原告认为不需要证明输出和权利书籍之间存在任何相似之处,因为其书籍在模型训练过程中被完整复制。法官援引艺术家诉 Stability AI 等案 2023 年 10 月的裁定(被指控侵权的衍生作品仍必须与原始作品有一些相似之处,或包含原始作品的受保护元素)认为,原告在诉状中没有对任何输出内容提出指控,不可能有间接侵权,原告需要明确并证明输出“以某种形式包含”原告书籍的一部分。

法院驳回了原告提出的《数字千年版权法》违规索赔,因为没有事实支持 Llama 曾分发过原告的书籍,也没有在移除版权信息的情况下分发。此外,因版权先占原则,还驳回了基于版权直接侵权行为提出的 UCL 索赔、不当得利和疏忽主张。

2024年8月,原告提交第二次修订诉状,仅保留版权直接侵权指控。案件目前仍处于证据发现程序,没有更新进展。

此案的核心争议在于 Meta 训练大模型的做法是否符合合理使用原则。Meta 声称,使用受版权保护的作品来训练生成式 AI 模型属于“转换性使用”,即通过新的方式或目的使用原作,从而产生不同的价值,符合合理使用第一要素。而作者方面认为,使用版权作品训练生成式 AI 模型不具有“转换性使用”的性质,因为这种使用方式与教育人类作者没有本质区别。此外,这种训练的目的是创建与被复制作品在同一市场竞争的作品,使这种使用方式具有“商业性”。

今年 3 月,美国加利福尼亚州北区联邦地区法院法官文斯·查布里亚(Vince Chhabria)裁定,侵犯版权的指控显然是足以成立的具体伤害,而 Meta 删除版权管理信息的做法为其侵犯版权的行为提供了便利和隐瞒,足以构成诉讼的具体损害。《美国法典》相关规定要求,未经版权所有者或法律授权,任何人不得故意删除或更改任何版权管理信息。原告还提出, Meta 采取其他措施降低 Llama 生成输出的可能性,这些输出会揭示或表明受版权保护的材料包含在训练数据集中。这些指控提出了一个“合理的,如果不是特别有力的推论”,即 Meta 删除了版权管理信息以阻止 Llama 输出这些版权管理信息。

查布里亚认为,目前最大的问题在于 Meta 的 AI 工具是否会损害图书销售,或给作者带来损失。他在听证会上表示,尽管愿意让案件继续推进,但对这些特定指控持保留态度。

案解解读 法官认定, Meta 删除版权管理信息为其侵犯版权行为提供了便利和隐瞒,足以构成诉讼的具体损害,是对受版权保护的作者及其作品利益的维护,体现了在 AI 领域司法实践的新发展。但是,原告需要提供充分数据证明其对 Meta 输出内容的指控,这对于判断 Meta 使用受版权保护作品训练大模型是否符合合理使用原则至关重要,也成为法院庭审时一直悬而未决的焦点。

2023年7月,理查德·卡德雷等作者向美国加州南区法院提交起诉,指控 Meta 将数百万作者作品用于训练大语言模型 LLaMA, 同时指控 Meta 从安娜的档案(Anna's Archives)和 LibGen 等盗版网站获得版权保护内容,提出版权侵权等多项索赔。LibGen 存储库包含数百万盗版书籍,并通过点对点种子网络进行分发。Meta 则辩称其行为属于“合理使用”,旨在通过学习生成原创内容。

2023年9月, Meta 提出驳回动议,请求驳回原告除版权直接侵权以外的其他索赔。对于版权间接侵权指控,被告认为,原告未能证明语言模型的表达与原告书籍存在实质性相似,且原告未能指出具体的侵权输出。对于移除版权信息及提供虚假版权管理信息(CMI)的指控,被告认为,原告没有具体指明被告移除或者提供了哪些虚假 CMI, 因此未能提出有效主张。

2023年11月,法院批准了被告的 2 项驳回动议。原告主张被告模型是侵权的衍生作品,因为模型不可能在没有从原告书籍中提取表达信息的情况下运作。法院认为,演绎作品是“基于一个或多个先前存在的作品”,以“作品可以被重铸、转换或改编的任何形式”存在。原告主张,“被告大语言模型的每个输出都是侵权的衍生作品,每个输出行为都构成了版权间接侵权行为”。原告认为不需要证明输出和权利书籍之间存在任何相似之处,因为其书籍在模型训练过程中被完整复制。法官援引艺术家诉 Stability AI 等案 2023 年 10 月的裁定(被指控侵权的衍生作品仍必须与原始作品有一些相似之处,或包含原始作品的受保护元素)认为,原告在诉状中没有对任何输出内容提出指控,不可能有间接侵权,原告需要明确并证明输出“以某种形式包含”原告书籍的一部分。

法院驳回了原告提出的《数字千年版权法》违规索赔,因为没有事实支持 Llama 曾分发过原告的书籍,也没有在移除版权信息的情况下分发。此外,因版权先占原则,还驳回了基于版权直接侵权行为提出的 UCL 索赔、不当得利和疏忽主张。

2024年8月,原告提交第二次修订诉状,仅保留版权直接侵权指控。案件目前仍处于证据发现程序,没有更新进展。

此案的核心争议在于 Meta 训练大模型的做法是否符合合理使用原则。Meta 声称,使用受版权保护的作品来训练生成式 AI 模型属于“转换性使用”,即通过新的方式或目的使用原作,从而产生不同的价值,符合合理使用第一要素。而作者方面认为,使用版权作品训练生成式 AI 模型不具有“转换性使用”的性质,因为这种使用方式与教育人类作者没有本质区别。此外,这种训练的目的是创建与被复制作品在同一市场竞争的作品,使这种使用方式具有“商业性”。

今年 3 月,美国加利福尼亚州北区联邦地区法院法官文斯·查布里亚(Vince Chhabria)裁定,侵犯版权的指控显然是足以成立的具体伤害,而 Meta 删除版权管理信息的做法为其侵犯版权的行为提供了便利和隐瞒,足以构成诉讼的具体损害。《美国法典》相关规定要求,未经版权所有者或法律授权,任何人不得故意删除或更改任何版权管理信息。原告还提出, Meta 采取其他措施降低 Llama 生成输出的可能性,这些输出会揭示或表明受版权保护的材料包含在训练数据集中。这些指控提出了一个“合理的,如果不是特别有力的推论”,即 Meta 删除了版权管理信息以阻止 Llama 输出这些版权管理信息。

查布里亚认为,目前最大的问题在于 Meta 的 AI 工具是否会损害图书销售,或给作者带来损失。他在听证会上表示,尽管愿意让案件继续推进,但对这些特定指控持保留态度。

案解解读 法官认定, Meta 删除版权管理信息为其侵犯版权行为提供了便利和隐瞒,足以构成诉讼的具体损害,是对受版权保护的作者及其作品利益的维护,体现了在 AI 领域司法实践的新发展。但是,原告需要提供充分数据证明其对 Meta 输出内容的指控,这对于判断 Meta 使用受版权保护作品训练大模型是否符合合理使用原则至关重要,也成为法院庭审时一直悬而未决的焦点。

2023年7月,理查德·卡德雷等作者向美国加州南区法院提交起诉,指控 Meta 将数百万作者作品用于训练大语言模型 LLaMA, 同时指控 Meta 从安娜的档案(Anna's Archives)和 LibGen 等盗版网站获得版权保护内容,提出版权侵权等多项索赔。LibGen 存储库包含数百万盗版书籍,并通过点对点种子网络进行分发。Meta 则辩称其行为属于“合理使用”,旨在通过学习生成原创内容。

2023年9月, Meta 提出驳回动议,请求驳回原告除版权直接侵权以外的其他索赔。对于版权间接侵权指控,被告认为,原告未能证明语言模型的表达与原告书籍存在实质性相似,且原告未能指出具体的侵权输出。对于移除版权信息及提供虚假版权管理信息(CMI)的指控,被告认为,原告没有具体指明被告移除或者提供了哪些虚假 CMI, 因此未能提出有效主张。

2023年11月,法院批准了被告的 2 项驳回动议。原告主张被告模型是侵权的衍生作品,因为模型不可能在没有从原告书籍中提取表达信息的情况下运作。法院认为,演绎作品是“基于一个或多个先前存在的作品”,以“作品可以被重铸、转换或改编的任何形式”存在。原告主张,“被告大语言模型的每个输出都是侵权的衍生作品,每个输出行为都构成了版权间接侵权行为”。原告认为不需要证明输出和权利书籍之间存在任何相似之处,因为其书籍在模型训练过程中被完整复制。法官援引艺术家诉 Stability AI 等案 2023 年 10 月的裁定(被指控侵权的衍生作品仍必须与原始作品有一些相似之处,或包含原始作品的受保护元素)认为,原告在诉状中没有对任何输出内容提出指控,不可能有间接侵权,原告需要明确并证明输出“以某种形式包含”原告书籍的一部分。

法院驳回了原告提出的《数字千年版权法》违规索赔,因为没有事实支持 Llama 曾分发过原告的书籍,也没有在移除版权信息的情况下分发。此外,因版权先占原则,还驳回了基于版权直接侵权行为提出的 UCL 索赔、不当得利和疏忽主张。

2024年8月,原告提交第二次修订诉状,仅保留版权直接侵权指控。案件目前仍处于证据发现程序,没有更新进展。

此案的核心争议在于 Meta 训练大模型的做法是否符合合理使用原则。Meta 声称,使用受版权保护的作品来训练生成式 AI 模型属于“转换性使用”,即通过新的方式或目的使用原作,从而产生不同的价值,符合合理使用第一要素。而作者方面认为,使用版权作品训练生成式 AI 模型不具有“转换性使用”的性质,因为这种使用方式与教育人类作者没有本质区别。此外,这种训练的目的是创建与被复制作品在同一市场竞争的作品,使这种使用方式具有“商业性”。

今年 3 月,美国加利福尼亚州北区联邦地区法院法官文斯·查布里亚(Vince Chhabria)裁定,侵犯版权的指控显然是足以成立的具体伤害,而 Meta 删除版权管理信息的做法为其侵犯版权的行为提供了便利和隐瞒,足以构成诉讼的具体损害。《美国法典》相关规定要求,未经版权所有者或法律授权,任何人不得故意删除或更改任何版权管理信息。原告还提出, Meta 采取其他措施降低 Llama 生成输出的可能性,这些输出会揭示或表明受版权保护的材料包含在训练数据集中。这些指控提出了一个“合理的,如果不是特别有力的推论”,即 Meta 删除了版权管理信息以阻止 Llama 输出这些版权管理信息。

查布里亚认为,目前最大的问题在于 Meta 的 AI 工具是否会损害图书销售,或给作者带来损失。他在听证会上表示,尽管愿意让案件继续推进,但对这些特定指控持保留态度。

案解解读 法官认定, Meta 删除版权管理信息为其侵犯版权行为提供了便利和隐瞒,足以构成诉讼的具体损害,是对受版权保护的作者及其作品利益的维护,体现了在 AI 领域司法实践的新发展。但是,原告需要提供充分数据证明其对 Meta 输出内容的指控,这对于判断 Meta 使用受版权保护作品训练大模型是否符合合理使用原则至关重要,也成为法院庭审时一直悬而未决的焦点。

案例
6
Meta 训练 AI「合理使用」辩护遭质疑

以上 6 起诉讼案涉及不同的原告和被告,足以引发多方面的思考。从技术角度看, AI 模型训练需要大量数据,这些数据的来源和使用方式是个关键问题。如何确保数据的合法性、准确性和可靠性,以及在数据使用过程中保护知识产权,是科技公司要面临的挑战。其次,从法律角度看,现有的知识产权法律体系面对 AI 技术发展存在一些不适应的地方。立法者能否尽快修订相关法律法规,以明确科技公司使用受版权保护材料训练 AI 时的权利和义务,成为摆在政府面前的一大考验。从社会层面看,既要抓住 AI 等技术发展带来的契机,又要切实保护出版业的合法利益并为其可持续发展保驾护航,其中的挑战是各界需要直面的问题。