從技術本位到利益平衡：模型訓練合理使用認定規則的重塑

張惠彬* 王懷賓**

摘 要 相較於早期Dreamwriter等自動寫作機器人，生成式人工智能采用了基於聯結主義的機器學習技術。除了基於政策考慮的合理使用主張之外，其他多數模型訓練的合理使用認定普遍建立在技術分析基礎之上，認為基於聯結主義的模型訓練對作品的利用更具有目的/內容轉換性，因此屬於合理使用。技術本位下的合理使用認定規則不僅與技術現實不相符合，而且在法律上也不夠周延，因其未能充分考慮模型訓練的最終部署目的等多元場景，未能達致合理使用認定中創作者、使用者的利益平衡。基於產業現狀和法律傳統，歐美等國的最新實踐試圖在著作權人利益和技術公司利益之間保持微妙平衡。在利益平衡原則指引下，我國模型訓練的合理使用認定規則可綜合參考以下因素：一是模型的最終部署目的和開發者/部署者防止侵權輸出的技術措施，是否顯著降低原作品潛在市場遭受損害的風險；二是被訓練作品類型、作品質量、訓練階段對轉換性程度的影響。模型訓練使用作品的轉換性程度越高、潛在市場損害風險越低，則更有可能被視為合理使用，這應當在司法實踐中進行個案判斷。

關鍵詞 模型訓練 版權風險 合理使用 利益平衡

一、問題的提出

ChatGPT、DeepSeek 等大模型技術的出現，推動人工智能技術發展進入新的階段，也對版權制度造成新的挑戰。生成式人工智能的技術核心是模型訓練：通過學習大量已有數據，訓練出數據分布的模式、趨勢以及相關性並轉變為模型權重，並根據使用者的要求，依據權重輸出內容。MuseNet、Stability AI、ChatGPT 分別可以根據使用者的提示生成音樂、圖像和文本。在模型訓練的過程中，開發者首先會複製數字作品並製作成作品數據集；其次開發者會創設一個權重隨機的模型，並基於該作品數據集預訓練模型，使得模型權重能夠反映作品數據集的表達分布規律；再次，

* 張惠彬，西南政法大學民商法學院副教授。
** 王懷賓，西南政法大學民商法學院博士研究生。
本文係國家社會科學基金後期資助項目“從訓練到生成：生成式人工智能的著作權問題研究”（項目編號：24FFXB055）、西南政法大學2024年度學生科研創新項目“生成式人工智能服務提供者著作權侵權責任研究”（項目編號：2024XZXS-006）階段性成果。

其他開發者或部署者有可能會基於該預訓練模型有針對性地使用一些作品進行微調，以使其滿足特定任務；最後，部署者會根據特定任務部署模型，這一特定任務不僅限於生成文學、藝術內容，還包括輔助醫療、法律檢索等多種任務。可以看出，模型訓練過程中可能存在著作權侵權風險：例如製作作品數據集對原作品的複製，預訓練或微調訓練過程中對原作品的臨時複製等。

自 2022 年底 ChatGPT 發布後，模型訓練的著作權問題在域內外引起大量爭論，其大致經歷了以下歷程。首先是理論層面呼籲以完全的合理使用規則為技術松綁，這一技術優先的理念還發展出“非表達性使用”觀點。[1] 緊接著部分觀點主張模型訓練應采取分層式、有條件的版權許可。[2] 到最近，理論層面的主流觀點開始考慮從利益平衡的角度來處理模型訓練的著作權問題，並衍生出場景化、多元化的著作權許可或合理使用認定。[3] 與理論界爭鳴如火如荼相比，域內外法律實踐則更加謹慎。“廣州互聯網法院 AI 第一案”雖然認定生成式人工智能服務提供者侵犯著作權，但仍以其沒有實施模型訓練為由，駁回了原告要求被告將其作品從被訓練數據集中刪除的訴求。[4] 歐盟《人工智能法案》堅持模型訓練屬於有條件（作品信息披露 + 選擇退出）的合理使用，但在實踐中仍然謹慎處理相關司法案件。德國漢堡法院雖然認定 LAION 公司未經許可複製原告作品創建用於模型訓練的數據集（LAION-5B）的行為不侵犯著作權，但其僅限於訓練數據集的創建行為而不涉及商業科技公司的模型訓練行為。[5] 直至 2025 年 5 月，美國版權局在綜合其收到的美國各界一萬餘份意見後，（預）發布了《版權與人工智能：第三部分》專門講述美國版權局的最終立場。該文件強調美國合理使用四要素分析仍然適用於模型訓練的著作權問題，但需要考慮美國聯邦法院最近在“Andy Warhol Found. *v.* Goldsmith 案”（沃霍爾案）[6] 等系列案件中關於合理使用四要素判斷的重心轉變——將合理使用判斷重心從“Campbell *v.* Acuff-Rose Music, Inc. 案”（坎貝爾案）[7] 以來的目的 / 內容轉換性轉向綜合考慮四要素。具言之，美國版權局主張場景化、全局性地考慮模型訓練是否滿足合理使用四要素，而非簡單地以模型訓練對作品的利用具有轉換性從而認定屬於合理使用。

事實上，無論是非表達性使用、臨時複製，還是一刀切的合理使用（排除基於法政策的宏觀考量），其背後所遵循的邏輯都是技術本位的。技術本位分析在面對模型訓練技術複雜、發展迅速且高度依賴部署場景等特徵時，在技術和法律上都不夠嚴謹且過於僵化，不僅使得分析與技術現實嚴重不符，而且也未能充分滿足《與貿易有關的知識產權協定》關於權利限制與例外的三步檢驗法所體現出的利益平衡原則。基於此，本文首先全面回顧當前模型訓練相對於早期（2015-2022）人工智能在技術和目的上迭代，以及當前主流技術分析存在哪些誤解和局限；其次結合各方爭議焦點，分析當前模型訓練合理使用認定規則的技術本位主義，以及基於技術本位的合理使用認定規則的不合理之處；然後回溯著作權合理使用制度如何體現利益平衡原則，以及比較歐美在面對模型訓練時如何實踐這一利益平衡原則；最後則是在選擇合理使用規則解決模型訓練的著作權問題基礎上，基

[1] 參見張吉豫、汪賽飛：《大模型數據訓練中的著作權合理使用研究》，載《華東政法大學學報》2024年第4期，第20頁；See Lemley, M. A, Casey, B, *Fair learning*, Texas Law Review, vol. 99, p.107 (2020).

[2] 參見孫靖洲：《人工智能訓練的版權困境及其出路：模塊化許可機制探析》，載《知識產權》2024年第11期，第94頁；蔡元臻：《機器學習著作權法定許可的適用基礎與規則構建》，載《知識產權》2024年第11期，第77頁。

[3] 參見杜娟：《AIGC模型訓練作品使用行為的版權規制研究》，載《中國出版》2025年第5期，第63頁；參見倪朱亮：《生成式人工智能訓練使用作品的許可複合機制研究》，載《法律科學》2025年第4期，第1頁。

[4] 參見廣州互聯網法院（2024）粵0192民初113號民事判決書。

[5] Siehe LG Hamburg, Urteil vom 27.09.2024, Az. 310 O 227/23.

[6] See Andy Warhol Found. *v.* Goldsmith, 21-869 U.S.(2023).

[7] See Campbell *v.* Acuff-Rose Music, Inc, 510 US 569 (1994).

於利益平衡原則提出場景化的合理使用認定考量因素。

二、模型訓練的新特徵對合理使用認定規則的挑戰

相較於傳統人工智能技術，ChatGPT、DeepSeek 等生成式大模型在技術和部署目的上具有新的特徵，這些新特徵對適用於傳統人工智能技術的合理使用認定思路提出挑戰。在技術特徵上，傳統符號主義人工智能通常被認為是對原作品表達的挪用，而聯結主義大模型則很難簡單歸類為挪用表達或挪用思想。在部署目的上，傳統決策式人工智能通常被認為不構成對原作品創作市場的替代，而以輔助 / 替代創作為最終部署目的的大模型則引發了著作權人對合理使用正當性的質疑。

（一）從符號主義到聯結主義的技術迭代

人工智能底層技術經歷了符號主義到聯結主義的迭代，前者通過預設邏輯規則和符號推理模擬人類智能，後者則仿生人腦神經網絡，通過數據訓練模型獲得知識和特徵，其核心技術是機器學習。[8] 騰訊 Dreamwriter 等早期自動寫作機器人 [9] 由符號主義所主導，ChatGPT 則由聯結主義主導。具體深入技術邏輯，早期自動寫作機器人是預設模板 + 規則庫的結合。先有算法對所收集的數據進行解析，結合歷史統計數據等維度的內容，形成一定格式的待檢測數據庫；其次，根據預先設計的規則和觸發條件，進行模板化的文章撰寫。在這個過程中，數據類型的輸入與數據格式的處理、觸發條件的設定、文章框架模板的選擇和語料的設定、智能校驗算法模型的訓練等均由主創團隊相關人員選擇與安排。[10] 而生成式人工智能則是大數據 + 機器學習的結合。開發者先創建一個權重隨機的神經網絡，爾後基於海量作品集對該網絡進行訓練，以使其朝著能夠將輸入轉化為預期輸出的方向進化。在這個過程中，最終固定下來的權重反映的是海量作品集的統計特徵與關聯模式。在文本生成中，當用戶輸入某個提示詞時，模型根據從海量作品集中“學習”到的經驗和特徵，以“詞語接龍”的方式依次生成最大可能出現的下一個詞。雖然 Dreamwriter 也使用機器學習算法，但其目的是對所收集數據進行清洗、分類，以形成結構化數據庫。例如解析“通信板塊上漲 2.1%”的數據後，得到“XX 板塊領漲”的固定句式，以為後續模板填充提供精準輸入。而 ChatGPT 使用的機器學習算法則是用於學習海量數據的分布規律，為後續詞語接龍提供概率預估。

這一底層技術邏輯的迭代引發了知識產權學界關於模型訓練對原作品的利用方式到底為何的爭議。產學界用拼貼機和風格模擬機來比喻兩種技術，並將其運用到模型訓練的合理使用與否的論證中。類似 Dreamwriter 等早期寫作機器人是對多個原作品進行剪切並存儲具體表達，以便在輸出中拼貼各種具體表達；而 ChatGPT 等大模型則是對大量原作品的特徵進行抽象、提煉並存儲創作風格（也即權重），以便在輸出中模仿這一風格。如此一來，Dreamwriter 對原作品的使用即是對具體表達的挪用，該具體表達受著作權法保護；ChatGPT 對原作品的使用則是對風格的挪用，而風格卻不受著作權法保護。前者幾乎沒有爭議，但後者卻引發了較大爭議。在“Andersen v. Stability AI 案”中，原告訴稱大模型所生成的新圖像“完全基於‘訓練圖像’，是 Stable Diffusion 在組合特定輸出時從特定圖像中提取的衍生作品。歸根結底，它只是一個複雜的拼貼工具。”[11] 多數觀點則反

[8] 參見魏斌：《符號主義與聯結主義人工智能的融合路徑分析》，載《自然辯證法研究》2022年第2期，第23-25頁。

[9] “Dreamwriter”是2015年由騰訊財經開發的一款自動寫作新聞軟件，雖然其能夠根據算法自動生成新聞，但公眾普遍認為該軟件是按照預先設定的寫作結構對已有數據庫內容進行拼貼生成。

[10] 參見廣東省深圳市南山區人民法院（2019）粵0305民初14010號判決書。

[11] See Andersen *v.* Stability AI Ltd, 3:23-cv-00201.

對將大模型比作老舊的“拼貼機”，主張著作權“從未包括對創造力基本組成部分的壟斷：思想、概念、風格、藝術技巧、語言或語法”，而大模型從海量作品集中提取的正是這些屬性。[12]

美國版權局《版權與人工智能：第三部分》以大模型技術結合美國版權法合理使用判斷要素，進一步闡述了這一爭議。[13] 美國版權法合理使用采用四要素判斷法，[14] 與之關聯的要素是“作品使用的目的和性質”，而過去司法實踐中這一要素在合理使用判斷中占比非常大。部分主張模型訓練所采用的機器學習是對海量作品的統計分析，其在目的與性質上與對原作品的表達性使用相去甚遠。Anthropic 公司稱，“在訓練數據中使用受版權保護的作品時，其用途僅限於分析（詞匯與概念之間的統計關係），與該作品的任何表達目的無關。”部分則反對前述“非表達性使用”的觀點，認為模型訓練對原作品的使用“類似於壓縮等不具有轉換性的過程……作品的表達元素只是以不同的方式呈現”“生成式人工智能‘預裝’了受版權保護的內容……並利用這些受版權保護的內容來生成自己的合成內容”。[15] 如此一來，模型訓練對作品的利用並不具有轉換性，由此遠離了合理使用。

具言之，以機器學習為核心的大模型較早期以符號主義為核心的人工智能技術確實有所不同，但這種技術層面的不同是否足以證成其模型訓練屬於合理使用，則是存在爭議的。

（二）從替代決策到替代創作的目的變化

人工智能的外在功能經歷了決策式人工智能到生成式人工智能的迭代。根據歐盟《人工智能法案》第 3 條的定義，人工智能指“用於以不同程度的自主性運行，可能在部署後表現出適應性，並且為了實現明確或隱含的目標，能夠從接收的輸入中推斷出如何生成輸出，例如預測、內容、推薦或決策，這些輸出可能對物理或虛擬環境產生影響”。該法案根據不同目的，區分輸出“預測、推薦、決策”的人工智能和輸出“內容”的人工智能。與此同時，該條款將通用人工智能模型定義為“包括通過大規模自監督學習使用大量數據進行訓練的人工智能模型，該模型具有顯著的通用性，能夠勝任執行多種不同任務，無論該模型以何種方式投放市場，均可集成到各種下遊系統或應用中”。所謂“大規模自監督學習”就是 ChatGPT 類大模型所使用的機器學習技術。但是此處通用人工智能模型並沒有預設其輸出的類型，其通用性係指“能夠勝任執行多種不同任務”“集成到各種下遊系統或應用”的通用功能。換言之，通用人工智能模型 + 輸出“預測、推薦、決策”即決策式人工智能，通用人工智能模型 + 輸出“內容”即生成式人工智能。前者的目的是輔助 / 替代人類決策，例如人臉識別、自動駕駛，後者的目的是輔助 / 替代人類進行文字、圖像、音視頻等內容“創作”，例如 ChatGPT。

此處兩種人工智能技術的底層邏輯都僅限於機器學習（聯結主義）所主導的大模型。原因在於，一方面上述符號主義主導的人工智能只能依據嚴格的符號邏輯處理結構化的任務，幾乎難以勝任人臉識別、自動駕駛等需要極強適應性，且需要處理非結構化數據的任務；另一方面，符號主義主導的以內容生成為目的的自動寫作機器人幾乎沒有著作權合理使用的辯護空間，實踐中多以獲取著作權人許可為前提（除非拼貼了原作品中的事實消息等不受著作權法保護的部分），這也是為

[12] See Zach Graves, et al, *AI Coalition Letter to Congress at 1, U.S. House of Representatives and U.S. Senate*, (11 September 2023), https://www.authorsalliance.org/wp-content/uploads/2023/09/AI-Coalition-Letter-9.11.2023-updated.pdf.

[13] See USPTO, *Copyright and Artificial Intelligence Part 3: Generative AI Training*, United States Copyright Office (6 May 2025), https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf.

[14] 美國版權法合理使用四要素是：（1）使用的目的和性質；（2）所使用的版權作品的性質；（3）被使用部分的數量和質量；（4）使用對原作品潛在市場和價值的影響。

[15] See USPTO, *Copyright and Artificial Intelligence Part 3: Generative AI Training*, p.41-44.

何人工智能著作權侵權問題的討論肇始於 2015 年左右的文本與數據挖掘（Text and Data Mining，TDM）合理使用例外。而根據歐盟《數字化單一市場版權指令》第 2 條的規定，TDM 係指，“旨在分析數字形式的文本和數據以生成包括但不限於模式、趨勢和相關性等信息的任何自動化分析技術”。結合歐盟《人工智能法案》的定義，TDM 實際上與“通用人工智能模型”一樣，都是沒有預設輸出的類型。

概言之，人工智能技術的發展可以描述為輸出類型的發展。在 2015 年左右至 2022 年末期間，人工智能的輸出類型主要是“預測、推薦、決策”，此時產學界所關注的 TDM 合理使用論證很大程度上受到模型訓練的輸出目的影響。直到 2022 年底 ChatGPT 的出現，意味著人工智能開始輸出“內容”，這一目的的變化顯然會引發不同後果：TDM 合理使用的正當性要遠遠優於當前模型訓練合理使用的正當性，這也是為何學界研究旨趣開始從早期的合理使用轉向利益平衡。[16] 雖然兩種人工智能都基於機器學習技術對海量數據進行模型訓練，但在模型訓練所使用的作品類別以及使用的目的上存在區別。就模型訓練所使用的作品類別而言，決策式人工智能所使用的大部分數據不是受著作權法保護的作品。例如人臉識別所使用的數據主要是肖像數據，自動駕駛所使用的數據主要是街道、人行道、紅綠燈等照片數據，這類數據通常不受著作權法保護。並且在數據獲取上有比較明確的數據來源，通常是由獲得授權的設備采集。而生成式人工智能所使用的數據雖然也有少部分例如事實消息、個人信息等不受著作權法保護的內容，但大多數都是從公開數據庫中抓取的文章、音視頻等文學、藝術作品，並且幾乎很難事先獲得著作權人的授權。就模型訓練使用作品的目的而言，決策式人工智能旨在輔助 / 替代人類決策；生成式人工智能則旨在輔助 / 替代人類創作。雖然決策和創作同屬智力勞動，但識別人臉、輔助駕駛等決策並不屬於文學、藝術創作，不會與用於模型訓練的原作品的潛在市場和價值相沖突，而輔助他人創作甚至替代本人創作則會顯著威脅到原作者的生存環境，這也難怪域外發生如此多起著作權人集體訴訟，傳統人類創作者確實到了生死存亡之際。

三、技術本位下模型訓練合理使用認定規則的“不合理”

當前大模型訓練的合理使用辯護，大致可以分為法政策學和解釋論兩種，前者通常基於法政策視角，主張通過模型訓練的合理使用規則為技術發展松綁；後者則遵循嚴密的技術和法律分析，認為無論是技術還是法律規則，模型訓練都符合現行合理使用條款或合理使用精神。對於前者而言，本身是一種價值取向和觀念之爭，難以評判其對錯。而就後者而言，模型訓練適用合理使用規則的解釋，必須符合技術現實和法律邏輯。[17]

（一）技術本位下合理使用認定的邏輯和內容

除基於法政策學分析的合理使用主張外，大多數模型訓練的合理使用規則認定都離不開縝密的技術分析。這一技術本位分析的特徵在於強調技術特徵在合理使用認定中的決定性地位，以僵化的技術分析完全支配合理使用認定的動態法律分析，忽略了合理使用在作品創作者和使用者之間的利益平衡。更為重要的是，技術本位分析往往與複雜且不斷更迭的技術現實不相符合。技術本位下的

[16] 參見張嘉鑫：《人工智能訓練中作品數據來源者利益共享機制研究》，載《知識產權》2025年第5期，第111頁。

[17] 當前關於模型訓練主要以不侵犯著作權為主流觀點，主要包括“非表達性使用論”“合理使用論”“臨時複製論”“總體國家安全論”，除“總體國家安全論”外，其他觀點的論證都建立在技術和法律分析基礎之上。參見易繼明：《大模型語料訓練合理使用問題研究》，載《中國版權》2024年第6期，第5-26頁。

合理使用認定遵循以下邏輯。首先，嚴格區分符號主義和聯結主義的人工智能，類似 Dreamwriter 自動寫作機器人屬於符號主義的人工智能，而 ChatGPT 屬於聯結主義的人工智能。其次，符號主義的人工智能是拼貼機，聯結主義的人工智能是風格模擬機。類似 Dreamwriter 自動寫作機器人是通過對多個原作品進行橫向剪切，由此拼貼出“新”的內容，因此其使用的是原作品的具體表達；類似 ChatGPT 大模型則是通過對大量原作品進行縱向抽象、提煉，旨在識別表達規律或固定表達風格，因此其對原作品的使用方式是值得商榷的，至少是不同於老舊技術那樣“欣賞”原作品的具體表達。最後，非法複製原作品的具體表達當然構成侵犯著作權，而模型訓練作為一種新的作品利用方式，這一使用行為從內容和目的上對原作品進行了轉換，由此符合轉換性合理使用，不侵犯著作權。轉換性使用是美國法院在司法裁判中用於解釋合理使用四要素中“使用的目的和方式”的新標準，其判斷重心從早期的重視作品創作時作者的意圖轉向二次使用是否增加了新內容，是否呈現出不同於原作品的新用途或新表達。[18] 這一流派是聯結主義技術分析與轉換性合理使用法律分析的結合，由於其突破了合理使用判定中商業性使用和整體使用的禁區，一度受到域內外學界論證機器學習合理使用的青睞，廣泛活躍在相關理論文獻之中。可以將其歸納為“以非表達性使用證成目的轉化的合理使用”“以風格模仿證成內容轉化的合理使用”兩類。

一是基於聯結主義技術分析主張模型訓練對作品的利用改變了作品的原始目的，屬於目的轉化的合理使用。模型訓練的非表達性使用是主要的目的轉化。非表達性使用理論源自美國判例法，並經學者充實逐漸適用於因技術革新而出現的對作品的新的利用方式。這種新的作品利用方式超越了傳統人類讀者以欣賞具體表達，實現與作者之間的交流的目的，本質上是功能性的。在美國早期案例中，非表達性使用最常出現在代碼著作權侵權訴訟中，原因在於與文學、藝術作品相比，計算機軟件的功能性更強。法院運用非表達性使用理論說明被告對原告代碼的挪用旨在實現某種技術功能，而非欣賞代碼設計的具體表達。[19] 後續逐漸發展到文學、藝術作品領域，美國法院曾認為被告網站基於原告作品的縮略圖是“幫助索引和改善圖像訪問的工具”，而非版權法意義上的審美對象，故不構成版權侵權。[20] 在美國版權局《版權與人工智能：第三部分》中，穀歌、IBM、Anthropic 等技術公司都以非表達性理論為模型訓練行為辯護。我國學者也不乏相關論證。機器學習雖然使用了作品數據，但目的是獲取表達符號之間的分布規律，未發生對作品的呈現式或演繹式使用，因此不侵犯著作權。[21] 模型訓練是對原作品的非表達性使用或非作品性使用，可考慮在司法實踐中利用目的轉換性合理使用規則將模型訓練行為解釋為合理使用。[22]

二是基於聯結主義分析主張模型訓練對作品的利用增加了新表達、意義或功能，屬於內容轉化的合理使用。模型訓練從海量作品的具體表達中得到了不同於原表達的創作風格、思想，這些新的信息甚至不受著作權法保護。在“Andersen v. Stability AI 案”中，被告認為，模型訓練得到的是反映被訓練作品（集）風格的模型參數，模型輸出則是風格模仿的結果，而風格不受著作權法保護。[23] 被告雖然是為後端生成內容不侵犯著作權辯護，但也反映了基於聯結主義的前端模型訓練不侵犯著作權的辯護，也即模型訓練過程中雖然臨時複製原告作品，但真正複製的是不受著作權法保

[18] 參見熊琦：《著作權轉換性使用的本土法釋義》，載《法學家》2019年第2期，第127頁。

[19] See Sega Enterprises Ltd. *v.* Accolade, Inc, 977 F.2d 1510 (9th Cir.1992).

[20] See Perfect 10, Inc. *v.* Amazon.com, Inc, 508 F.3d 1146 (9th Cit.2007).

[21] 參見陶乾：《基礎模型訓練的著作權問題：理論澄清與規則適用》，載《政法論壇》2024年第5期，第152頁。

[22] 參見郭萬明：《生成式人工智能模型訓練中作品數據的著作權保護》，載《江蘇社會科學》2025年第3期，第166頁。

[23] See Lisa T. Oratz, et al, *First Lawsuits Arrive Addressing Generative AI*, (20 April 2023), https://perkinscoie.com/insights/update/first-lawsuits-arrive-addressing-generative-ai.

護的風格並對外傳播這一風格。美國版權局《版權與人工智能：第三部分》中，薩繆爾森（Pamela Samuelson）等人指出，“從海量數據中提煉出不受版權保護的抽象概念和關聯關係，並利用這些知識來創造新的數字作品，這不僅具有轉換性的性質，更是高度轉換性的。”[24] 可以看出，這一觀點雖然承認模型訓練確實從原告作品中挪用了某些東西，但挪用的是不受著作權法保護的風格、抽象概念、關聯關係等，這些“東西”是經由聯結主義技術才能得到的模型權重。

（二）技術本位下合理使用認定的“不合理”

上述基於聯結主義技術分析的合理使用規則認定，始終建立在能夠嚴格區分符號主義和聯結主義的人工智能，以及這種區分具有著作權法上的意義。然而完全將聯結主義與符號主義區分開來，從而認為 ChatGPT 對作品的利用是完全不同於 Dreamwriter 對作品的利用，則是值得商榷的。

1. 合理使用的目的轉換性認定的不合理

當前合理使用論證中，僅僅從客觀技術上認為模型訓練對原作品的利用——獲取表達符號的分布規律——相對於原作品的創作目的具有轉換性與技術現實不相符合。

任何閱讀行為本身並不是著作權法所規制的對象，所謂對作品的利用指的是為該閱讀行為而製作複製件的行為，目的轉換性指的也是製作該複製件的目的。模型訓練過程中製作複製件，其目的包括直接目的和最終目的。直接目的是用於機器閱讀，最終目的則取決於模型設計目的和最終的部署。目的是否具有轉換性，以及具有多大的轉換性，取決於機器閱讀對作品的利用目的與原作品創作目的的差別。無論是基於非表達性使用的不落入著作權規制範圍，還是非表達性使用構成目的轉換性合理使用，其共同點都在於認為模型訓練旨在獲取表達符號的分布規律之目的，使其不同於原作品的表達功能，因而具有目的轉換性。[25] 這一思路遵循了“兩頭固定”的基本假設，也即原作品必然只有表達用途、模型訓練必然只有功能性分析表達規律的用途。這一假設值得商榷。首先，原作品的用途或目的有多種考量因素。在大多數情況下，原作品的初始功能是文本意義上的功能，也即提供某種外化的思想或美學方面的精神內容。[26] 然而在一些情況下，也需要考慮作品創作時作者的意圖。其次，模型訓練對原作品的利用也不盡然是功能性分析表達規律。這是因為目的轉換不能僅考慮技術客觀上使用者的目的，[27] 更多的是需要從作品受眾的角度考量二次使用是否拓展了原作品的用途或功能。

如果非要從技術客觀上考察模型訓練的目的，反而會得出為訓練模型而複製作品，不具有目的轉換效果。當在機器閱讀的隱喻下來討論時，可以將“獲取表達符號的分布規律”稱之為非常人的欣賞、閱讀作品的具體表達，其與人類閱讀具有相同的目的。馬克・萊姆利（Mark A. Lemley）即提出了“公平學習”原則，認為如果人類有權通過閱讀獲取思想，人工智能也應有權通過複製學習非保護性內容。[28] 這一公平學習原則恰恰反映了模型訓練（機器閱讀）對作品的利用與人類對作品的利用具有本質特徵，人類閱讀是為了欣賞具體表達或學習作品思想，機器閱讀也無非是以非常人的方式閱讀作品內容並將其轉化為反映作品表達規律的權重。換言之，原作品在面對人類讀者和機器讀者時，其功能和用途沒有什麼區別：將作品的具體表達或抽象思想記憶下來。只是人類是將其

[24] See USPTO, *Copyright and Artificial Intelligence Part 3: Generative AI Training*, p.42.

[25] 非表達性使用在我國理論中既有用於論證模型訓練沒有落入著作權保護範圍，也有用於論證屬於目的轉換性合理使用，兩種論證思路在內核上是一致的。

[26] 參見張耕、林楠：《規範性路徑下作品的轉換性使用標準重構及本土化運用》，載《西南民族大學學報(人文社科版)》2019年第8期，第109頁。

[27] 即使考慮使用者作品使用目的，那麼人工智能和人工智能開發者誰是真正的使用者呢？技術本位有一種將人工智能這一機械意識主體認為是作品使用者的傾向，其總是以一種服務於系統功能的目的去被動、機械地讀取作品，而不會被如人類作者一樣可能在閱讀作品事實時也會被懷疑欣賞作品具體表達。

[28] See Lemley, M. A, Casey, B, *Fair learning*, Texas Law Review, Vol. 99, p.107 (2020).

記憶在腦海中，而模型則是將其以權重的形式固定下來。正如美國作家協會所言，“人工智能公司之所以選擇已出版的書籍進行訓練，正是因為這些書籍具有豐富的表達內容。高質量、專業撰寫的著作對於使大型語言模型能夠生成模仿人類語言、故事結構、角色發展和主題的輸出至關重要。”[29] 換言之，模型本身也會根據作品表達的優質程度來“挑”作品，這也反映其對具體表達的需求。

前述指出機器閱讀與人類閱讀的同質性似乎會讓人誤以為機器閱讀也應當完全不受任何限制。任何閱讀行為都沒有落入著作權法規制範圍，但並不意味著為閱讀而複製作品的行為也完全自由。我國《著作權法》第 24 條規定的個人閱讀、欣賞作品的合理使用，豁免的就是為個人閱讀而製作複製件的行為。在沒有采用雲端學習等技術下，機器閱讀必然需要製作複製件。但該模型閱讀、欣賞作品顯然無法為上述條款的立法者所預料，也難以涵蓋進該條款嚴格的文義解釋之中。在司法實踐中，即使是人類為個人閱讀、欣賞目的而複製作品的行為也會受到嚴格限制。[30] 當機器閱讀和人類閱讀具有同質性時，就應當認為為閱讀作品而複製作品的行為同屬於表達性使用從而不具有目的轉換。

事實上，與傳統“使用目的和方式”的判斷標準相比，轉換性使用將合理使用的重點從創作者的目的轉變為作品受眾的認知。[31] 在著名的穀歌圖書館案中，法院即以之認為穀歌以展示片段、便於搜索的目的大量複製、傳播作品構成轉換性使用。[32] 因為從作品受眾的角度來看，這一方便搜索的目的顯然拓展了原作品傳達思想或美學方面的精神內容的功能，受眾也不期待該片段能夠發揮原本作品的功能。然而僅從“獲取表達符號的分布規律”判斷模型訓練拓展了原作品的用途或功能是不足的。部分意見指出，“在分析模型開發者使用受版權保護材料的目的和性質時，法院不應孤立地看待訓練過程，而應考慮模型的最終用途。”[33] 換言之，不能僅根據“獲取表達符號的分布規律”的直接目的來判斷目的轉換性，而是更多地考慮模型訓練的部署和最終目的。這是因為模型不可能是孤立存在，任何模型都有其最終目的：識別人臉？輔助駕駛？輔助醫療？科學研究目的？生成內容？即使是生成內容也並不完全是生成供人欣賞、閱讀的文學藝術作品，還有可能是商品導購、醫療診治等功能性對話。

這一考慮最終目的的合理使用認定在美國最新判決中有所體現。在“沃霍爾案”中，美國聯邦最高法院主張在判斷是否具有目的轉換時，不僅需要審視被告所實施的直接複製行為，而且還要考慮其最終目的。該案中，有爭議的使用並不是被告對另外 15 件作品的再創作，而是他的基金會後來於 2016 年將其中一件作品進行商業許可，這導致法院認為其作品使用的目的與原作品的創作目的相一致。

2. 合理使用的內容轉換性認定的不合理

當前合理使用論證中，認為模型訓練對原作品的利用獲取了不同於原作品的新內容（不受著作權法保護的風格）從而具有內容轉換性是不合理的。

模型訓練對作品的利用是否具有內容轉換性是一個程度問題。上述指出模型訓練對作品的使用不具有新的目的，而是與人類閱讀一樣，也是通過學習該作品的具體表達並將其固定成模型權重。

---

[29] See USPTO, *Copyright and Artificial Intelligence Part 3: Generative AI Training*, p.42.

[30] 美國法院認為科研人員將內部圖書館分發的科學期刊論文複印後作為研究檔案帶入實驗室，不屬於合理使用，即使科學家很大可能只是為了讀取論文中的事實。See American Geophysical Union *v.* Texaco Inc, 60 F.3d 913 (2d Cir. 1994).

[31] See Laura Heymann, *Everything Is Transformative: Fair Use and Reader Response*, Columbia Journal of Law & the Arts, Vol. 31:4, p.452 (2008).

[32] See Authors Guild, Inc. *v.* Google Inc. 954 F. Sup. 2d 282 (S, D. N. Y. 2013).

[33] See USPTO, *Copyright and Artificial Intelligence Part 3: Generative AI Training*, p.44.

至於固定的是具體表達還是抽象思想，則往往與我們對作品的印象是否深刻有關：當人類閱讀到絕妙的情節或反復閱讀一部作品時，則當然會印象深刻，而當機器閱讀到極其罕見的表達或反複閱讀一部作品時，則會發生所謂的“記憶”現象。至此，我們可以更加確認模型訓練對作品的利用並沒有什麼不同，如果將模型訓練的最終目的擴展到基於學習到的表達分布概率創作風格相似的作品時，就更加確信這一點了。模型訓練對作品的利用的直接結果是將其具體表達規律以模型權重的形式固定下來（最終結果需要考慮模型訓練的部署目的）。模型權重相對於原作品，是否足以稱得上轉換了內容？一些觀點認為，從作品集到模型權重，“類似於壓縮等不具有轉換性的過程，在這些過程中，作品的表達元素只是以不同的方式呈現。”[34] 這種觀點並非毫無道理，模型權重雖然是一串人類無法理解的數字串，但其同樣可以按照一定程式“恢復”成人類可以理解的具體表達。從模型技術來看，模型權重“潛在空間”[35] 實際上是作品的另外一種表徵方式，其是縱向提煉、抽象海量作品後以數字串的形式存儲作品的創作規律和特徵，就像人腦中關於創作技巧的模糊記憶和條件反射。有觀點將數字圖像和“潛在空間”（模型權重）視為作品不同維度的表徵方式，只是恰好以100×100 像素點所表徵的數字圖像比以字符串形式呈現的“潛在空間”更容易被我們理解而已。[36]

模型權重在多大程度上相對於原作品是具有內容轉換性的呢？這實際上是一個在技術上和法律上都難以量化的問題。從技術上來說，當被訓練作品在具體表達上越集中，模型越能夠強化對這一類似表達的記憶現象，這是由機器學習的特性所決定的。極端來說，當用於訓練的海量作品都是同一個作品時，模型權重即這一作品具體表達的數字表徵，模型權重即與作品完全相等，此時我們很難說模型訓練具有內容轉換性。反之，如果用於訓練的海量作品間完全毫無聯繫，模型權重就很難與任何一個作品對應起來。從法律來看，從一個極端到另一個極端，實際上意味著模型權重從表徵具體表達到表徵抽象思想的變化，在某一個瞬間就可能從受著作權法保護的表達轉變為不受著作權法保護的思想，從而就具有了內容上的轉換性。這個程度很難把握，因其涉及著作權法領域最令人捉摸不透的部分——思想 / 表達二分法。但我們可以肯定，適用於泛化任務的基礎模型訓練與有針對性地微調模型訓練在內容轉換性程度上應該是有明顯區別的。

總而言之，無論是基於技術還是法律，都不能一刀切地認定模型訓練必然存在轉換性從而屬於合理使用，而是需要具體情況具體分析。

四、模型訓練合理使用認定規則中的利益平衡

《與貿易有關的知識產權協定》明確成員國可以在三步檢驗法的原則之上，自行規定具體的合理使用制度。三步檢驗法的抽象性和原則性，使得不同國家 / 地區會根據法律傳統和產業實際情況動態調整合理使用制度，以最終在創作者、傳播者和使用者的博弈中取得利益平衡。歐盟的文化產業強勢而人工智能產業弱勢，以及長久以來的法律技術傳統使得著作權人在利益的天平中更具有話語權；反之美國人工智能產業繁榮，以及寬泛的合理使用制度，則使得技術公司更具有話語權，但在面臨新的人工智能模型對文化產業的威脅時，美國也嘗試建立新的平衡。

[34] See USPTO, *Copyright and Artificial Intelligence Part 3: Generative AI Training*, p.42.

[35] 所謂“潛在空間”係指模型對海量作品進行抽象、降維後所形成的，由各種向量及表示數據特徵的分量所組成的、低緯度的“空間”。See Ekin Tiu, Understanding Latent Space in Machine Learning, TDS Archive, (4 Feb 2020), https://medium.com/data-science/understanding-latent-space-in-machine-learning-de5a7c687d8d.

[36] See Sobel B, *Elements Of Style: Copyright, Similarity, And Generative AI*, Harvard Journal of Law & Technology, Vol.38:1, p.24 (2024).

（一）合理使用認定規則中的利益平衡

前述分析指出，技術本位分析過於強調技術特徵在合理使用認定中的決定性地位，而忽略了合理使用制度在作品創作者和使用者之間的利益平衡。在美國合理使用四要素認定框架下，技術本位實際上就是過於強調模型訓練對原作品的利用不同於原作品的創作目的和創作內容，由此將模型訓練一律視為合理使用行為，而不考慮其他三個要素在合理使用認定中的重要作用，尤其是模型訓練對原作品潛在市場和價值的影響。事實上，無論是美國合理使用四要素，還是我國《著作權法》第24條的規定，都是對《與貿易有關的知識產權協定》合理使用三步檢驗法的具體實施：合理使用並不取決於使用作品的技術是怎麼樣的，而是對該作品的使用是否與作品的正常利用不相沖突、沒有不合理地損害權利人的合法權益。換言之，技術本位分析是技術支配法律，而利益平衡分析則是法律與技術的互動，其以是否達致創作者和使用者利益的平衡為最終判斷因素，同時將技術分析作為實踐中影響利益平衡的考慮因素，例如技術的哪一階段、哪種部署目的更可能對著作權人利益造成不合理損害。

基於利益平衡的合理使用認定規則在過去的理論和實踐中得以廣泛運用，技術分析往往是各國 / 地區為科技發展辯護的工具，但在有可能實現科技產業和文化產業的共同繁榮時，利益平衡才是合理使用認定的最終考慮因素。作為著作權權利限制和例外，合理使用制度旨在動態調整因新的作品利用技術和商業模式而打破的創作者、傳播者和使用者之間的利益平衡。合理使用概念的制度初衷是為化解後續作者創作新作品而利用前人作品的障礙。[37] 伴隨著實踐發展，其正當性基礎得以豐富，既有可能是為了降低交易成本、促進後續創作，也可能是讓步於言論自由或市場競爭等其他價值。[38] 另有學者將合理使用擴張為一種權利。湯姆・貝爾（Tom Bell）認為，版權應當是普通法權利的法定例外，版權的引入使得“自然法和普通法權利”上的複製受到限制。[39] 這種法定例外賦予了作者相對於其他人的“不道德的特權”，免除了他們原本必須允許他人複製其作品的義務。正是由於合理使用制度理論的不斷豐富，其在實踐中的運用也得到了極大擴張。司法實踐中已不滿足合理使用僅用於促進作品傳播和後續人類創作，而開始將其廣泛運用到為新技術和新商業模式自由利用作品的辯護上。新技術和新模式確實在某種程度上擴展了作品的目的和內容，極大釋放了作品的價值。而在以開放性著稱的美國合理使用四要素認定規則的幫助下，新技術和新商業模式得以獲得廣泛自由，在一定程度上促成美國科技企業的繁榮。在過去幾十年，美國法院就以寬泛的合理使用制度為搜索引擎、數字圖書館、論文檢測、新聞抓取等新出現的機器閱讀行為辯護。[40] 以轉換性使用標準擴張解釋早期的“作品使用的目的和性質”標準即主要手段之一。

美國在重視作品使用者利益的同時，也並非罔顧著作權人的利益。轉換性使用實際上建立在技術公司開發出了作品的新價值，無論是發現作品新的使用用途（例如以多元作品訓練矯正人工智能歧視）還是發掘作品新的內容（海量作品蘊含的創作風格或商業規律）。這一新的用途和新的內容實際上意味著其不太可能損害原作品的潛在市場價值，自然沒有損害原作者的利益。正是基於利益平衡之精神，人臉識別、自動駕駛等決策式人工智能對作品的利用才在美國獲得廣泛自由。換句話說，技術公司廣泛自由的正當性基礎在於，其以技術做大了“作品價值的蛋糕”，同時將新增價值分配給技術公司也不會損害著作權人的利益。與美國在利益平衡的基礎上將作品的新價值分配給技術公司不同，歐盟則在利益平衡的基礎上將作品的新價值分配給著作權人。雖然 2019 年歐盟發布

[37] 參見馮曉青：《著作權合理使用制度之正當性研究》，載《現代法學》2009年第4期，第29頁。

[38] 參見崔國斌：《著作權法：原理與案例》，北京大學出版社2014年版，第578頁。

[39] See Bell, T. W, *Copyright as intellectual property privilege*, Syracuse Law Review, Vol.58, p.7 (2007) .

[40] See Krista Cox, *Fair Use in Text and Data Mining: ARL Publishes Issue Brief*, ARL News (15 June 2015, https://www.arl.org/news/fair-use-in-text-and-data-mining-arl-publishes-issue-brief/.

《數字單一市場版權指令》（以下稱《版權指令》），為決策式人工智能對作品的利用設置了合理使用條款，但其與美國的合理使用認定規則賦予科技企業的廣泛自由相去甚遠。針對常見的商業性TDM，《版權指令》為著作權人設計了權利保留機制，也即著作權人可以選擇以適當的方式明確拒絕科技公司使用其作品，例如針對網上公開提供的內容采取機器可讀的方式。[41] 而在實踐中，著作權人可以基於權利保留迫使技術公司重新回到著作權交易的談判桌上。歐盟將利益的天平撥向著作權人一方並非沒有原因。作為文藝復興的發源地，作者權法傳統在歐洲根深蒂固，作者們顯然對這種機器閱讀和機器創作嗤之以鼻。這進一步影響到歐洲的技術法律政策，也即以技術憲法主義限制科技公司濫用其“私權力”來危害歐洲的創意產業。[42]

具言之，著作權制度中的合理使用規則是一個動態調整的過程，其根據利益平衡原則不斷調整創作者、傳播者和使用者的利益。這一利益平衡之精神得益於不同國家 / 地區的各方利益主體的不斷博弈以及國家法律傳統、當前技術政策的考慮，而其平衡得以可能依據的是大家所共同遵循的《與貿易有關的知識產權協定》中的三步檢驗法：限制與例外只能在特殊情況下作出、與作品的正常利用不相沖突、沒有不合理地損害權利人的合法權益。

（二）歐美國家實踐中的利益平衡經驗

上述分析了各國 / 地區的合理使用制度受到不同話語權的利益主體的博弈，導致了不同的利益平衡結果，尤其是在面對早期決策式人工智能技術時的平衡藝術。在面對最新的模型技術時，歐洲仍然遵循舊例，將利益的天平撥向著作權人，但也留下了一些回旋空間。美國則開始將過去傾向於技術公司的天平向著作權人撥動，開始考慮場景化的利益平衡策略。

1. 歐盟立場：從權利保留到選擇－退出的有限合理使用的制度慣性

歐盟《人工智能法案》是全球第一部人工智能立法，其於 2024 年 8 月正式生效。該法案第 53 條要求通用人工智能模型提供商公開披露用於模型訓練的作品信息，以用於著作權人實現依據《版權指令》所享有的禁止技術公司利用其作品進行 TDM 的權利。[43] 只是其相對於《版權指令》的事先權利保留，該法案采取的是作品信息披露義務 + 選擇—退出的事後退出機制。《版權指令》規定有兩種 TDM 合理使用。一是科研機構、文化遺產機構以科學研究為目的的 TDM 合理使用；二是任何其他主體包括商業目的的 TDM 合理使用。針對後者，著作權人可以選擇以適當的方式明確權利保留。[44] 前述分析指出，法案的“通用人工智能模型”係指基於自我監督、用海量數據訓練而來的基礎模型，其特徵是顯著的通用性和泛化性，也即是沒有預設任何下遊任務。而提供商則是指開發者。這就意味著，任何采用自我監督訓練的大模型的開發者都需要披露用於訓練的作品信息，而著作權人可以隨時要求退出模型訓練，無論該基礎模型是否實際部署，或者部署於何種任務。具言之，歐盟《人工智能法案》仍然是對《版權指令》的延續，以一刀切的方式要求所有基礎模型都需要為著作權人設置選擇—退出機制。[45]

繼歐盟《人工智能法案》生效之後，德國法院做出了一項與模型訓練著作權相關的重要司法判決。2024 年 9 月，德國漢堡地方法院駁回著作權人對非營利機構 LAION e.V. 的起訴，認定 LAION

[41] 參見歐盟《數字單一市場版權指令》第4條。

[42] 參見張惠彬、王懷賓：《版權優先還是技術優先?——法國應對AIGC版權風險的趨勢及啟示》，載《編輯之友》2024年第5期，第106頁。

[43] 參見歐盟《人工智能法案》第53條第1款第（c）（d）項。

[44] 參見歐盟《數字單一市場版權指令》第4條。

[45] 但不包括“非專業或科學研究目的開發或使用模型”以及“在市場發布之前用於研究、開發或原型設計活動的人工智能模型”。參見歐盟《人工智能法案》第3條（63）項，解釋性備忘錄109項。

未經許可複製原告作品創建數據集（LAION-5B）的行為不侵犯著作權。[46] 該案中，被告是德國一家致力於開源人工智能運動的非營利性組織，其工作是將互聯網上的公開圖片製作成數據集，以免費提供給人工智能公司訓練模型。德國《著作權法》是對歐盟《版權指令》的轉化，其與該案密切相關的條款主要包括 44a 條（臨時複製）、44b 條（商業目的的 TDM 合理使用），以及 60d 條（科學研究目的的 TDM 合理使用）。法院首先駁回了被告臨時複製的合理使用抗辯。按照歐盟《信息社會版權指令》的規定，[47] 基於臨時複製的合理使用必須滿足以下條件：（1）複製行為是短暫的或附帶的；（2）其唯一目的是實現：通過中介在網絡中進行的第三方之間的傳輸，或者對作品或其他受保護內容的合法使用；（3）是技術過程不可或缺的組成部分；（4）該複製行為不具有獨立的經濟意義。法院認為被告基於模型訓練之分析目的的有意識且積極控制的獲取過程，不符合基於分析技術過程的附帶性；同時也非網絡傳輸的必要技術過程。其次法院認定被告的複製行為屬於第 60d 條基於科學研究目的的 TDM 合理使用，其理由如下：被告屬於專注於開源人工智能技術的非營利組織，符合主體要求；該數據集由被告免費分發給公眾使用，符合非商業性使用要求；雖然該數據集有可能被商業技術公司使用，但被告無法預測會被如何具體使用。最後，法院並沒有就 44b 條的適用展開分析。該案將訓練數據集的創建和使用行為分割開。如果訓練數據集的創建行為符合主體非營利性和創建目的的非營利性，則可以獲得完全的合理使用豁免；反之則是有限的（權利保留）合理使用豁免。使用行為也同是如此。而由於實踐中很多使用訓練數據集進行模型訓練的主體通常是商業性技術公司，因此只能寄希望於 44a 條和 44b 條。

綜上所述，在歐盟《人工智能法案》頒布施行之後，非營利性的訓練數據收集行為可能獲得完全合理使用豁免，該訓練數據集最終是否被用於商業性模型訓練不影響非營利性判斷。除非有明顯證據證明使用方對收集方有實質的控制力。[48] 這種將訓練數據集收集行為和使用行為分開評價的策略留下了一個關鍵議題，那就是使用方在模型訓練的過程中對訓練數據集的使用有可能因落入臨時複製而獲得完全（複製權）合理使用豁免。這是因為相較於訓練數據集收集過程中對作品的複製，模型訓練過程中對作品的複製更符合臨時複製的特徵。當然，一切都只是推測，結合《人工智能法案》的規定以及歐盟著作權法傳統，商業性技術公司的模型訓練行為很可能還是得遵循作品信息披露 + 選擇 − 退出機制，這在 2024 年 3 月的法國“穀歌行政處罰案”中得以印證。[49]

2. 美國立場：從目的轉換到市場損害的合理使用要素的重心轉移

美國版權局於 2025 年 5 月 9 日預發布《版權與人工智能：第三部分》，儘管該報告並沒有任何法律效力，但在綜合收集了產學界一萬條左右的重要意見後，仍然反映了美國產學界的主要觀點。前述二、三部分已有部分觀點源自這一報告，以下將總結美國版權局的最終立場。（1）模型訓練所涉及的數據收集 / 整理、訓練階段都有可能以侵犯著作權的方式使用作品。這一立場實際上否定了非表達性使用的觀點，就像部分意見指出的那樣，不論是立法還是司法實踐，非表達性使用都不是一個正式的法律概念，其只能作為考察模型訓練對作品的使用是否具有轉換性的一個因

[46] Siehe LG Hamburg, Urteil vom 27.09.2024, Az. 310 O 227/23.

[47] 參見歐盟《信息社會版權指令》第5條第1款。

[48] 參見歐盟《數字單一市場版權指令》第2條第（1）款，“並且，對該機構有決定性(decisive)影響的主體不能夠優先獲取該機構產出的研究成果。”是非營利性的判斷因素之一。

[49] 2024年3月15日，法國競爭管理局對穀歌公司處以2.5億歐元罰款，原因是穀歌公司在利用新聞出版機構的作品訓練生成式人工智能“Bard”（已更名為“Gemini”）過程中，未履行透明度義務，以及未采取技術措施滿足著作權人“選擇—退出”機制。Cf. Autorité de la concurrence, Related rights: the Autorité fines Google €250 million for non-compliancewith some of its commitments made in June 2022, [En Ligne:https://www.autoritedelaconcurrence.fr/en/press-release/related-rights-autorite-fines-google-eu250-million-non-compliance-some-its]. Consulté le 20 mars 2024.

素。[50]（2）模型訓練所涉及的初步侵權行為是否構成合理使用，需要根據所有相關情況，對四個法定要素進行權衡。但根據慣例，模型訓練在多數情況下都可能具有轉換性，但在多大程度上屬於合理使用，則取決於所使用的作品、來源、目的和對輸出的控制方式等因素，綜合考慮是否損害原作品的潛在市場和價值。（3）當前美國合理使用制度可以在個案中解決爭議，政府鼓勵但不強行干預自願許可市場的進一步發展，在市場失靈的情況下可以考慮延伸集體許可等干預措施。美國版權局立場中的關鍵部分在於，如何依賴合理使用四要素的司法經驗，在不同場景中綜合考慮模型訓練是否損害原作品的潛在價值和市場。

美國版權局的報告重點引用了美國聯邦最高法院於 2023 年 5 月作出的“沃霍爾案”。該案中，被告將原告的作品製作成絲網印刷肖像畫，法院認為被告的新用途轉換未通過合理使用測試，因其將絲網印刷肖像畫商業性許可給他人，並於雜志上展示，這使得其使用目的與原告的創作目的相同，削弱了其轉換性程度。該案被認為美國著作權法上的一個重要轉折點，因其矯正了過去司法實踐中過度強調合理使用四要素中“使用的目的和性質”，而忽略了其他三要素，尤其是“使用對原作品的潛在市場和價值的影響”。如上所述，1994 年“坎貝爾案”引入的“轉換性使用”是對“使用的目的和性質”的擴張解釋，而美國法院在後續的司法實踐中錯誤地將“轉換性使用”簡化為“是否添加新表達、意義或信息”，並過度重視這一要素，從而忽視了商業性和市場替代效應。[51]而該案則提出了一個三要素框架：（1）獨立正當性：使用需有獨立於便利或搭便車的理由（如評論、批判），且需證明對原作品的針對性必要；（2）不同目的：二次使用的目的需與原作品的典型用途顯著不同，避免市場替代；（3）商業性平衡：商業用途需更強的正當性支持，需與轉換性程度權衡。2025 年 6 月，美國聯邦地區法院就“Kadrey v. Meta 案”作出判決，雖然法院最終認定 Meta 模型訓練構成合理使用，但法官同樣提出了模型訓練可能損害原作市場的“市場稀釋”理論，但礙於原告並未提出足夠證據以將市場稀釋問題提交至陪審團，因此無法支持原告主張。[52]

綜上所述，雖然早期美國法院更青睞於技術公司，但伴隨著技術發展和法律實踐的變化，美國版權局也開始傳達出適當關照版權人利益的立場，具體表現則是場景化的合理使用四要素認定，並圍繞轉換性和是否損害原作品潛在市場和價值兩個重要考量因素。

五、利益平衡下模型訓練合理使用認定規則的重塑

（一）模型訓練合理使用認定的一般原則

結合前述技術分析和歐美實踐來看，我國模型訓練合理使用認定應首先排除一刀切式方案，這些方案主要包括基於非表達性使用、臨時複製理論在司法實踐中將所有模型訓練行為解釋為合理使用。就非表達性使用而言，其最早的系統性研究可追溯於美國學者詹姆斯・格林梅爾曼（James Grimmelmann）於 2015 年發表的論文。[53]作者在文中分析了美國版權法如何通過判例演化出“機器閱讀不構成侵權”的隱性規則，而這種幾乎完全確定的規則實際上導向了機器閱讀的“非表達性使用”路徑。作者批評機器閱讀的“非表達性使用”會導致以下不利後果：（1）從實踐來看，司法實踐中存在區別對待人類閱讀和機器閱讀，涉及人類讀者的使用會受到嚴格的審查，以確保版權

[50] See USPTO, *Copyright and Artificial Intelligence Part 3: Generative AI Training*, p.44.

[51] See Balganesh S, Menell P S, *Going “Beyond” Mere Transformation: Warhol and Reconciliation of the Derivative Work Right and Fair Use*, The Columbia Journal of Law & the Arts, Vol.47, p.413 (2024).

[52] See Kadrey *v.* Meta Platforms, Inc, 3:23-cv-03417, (ND Cal.).

[53] See Grimmelmann J, *Copyright for literate robots*, Iowa Law Review, Vol.101, p.657 (2015).

所有者的市場不會被搶占，而涉及機器讀者的非表達性閱讀則會被完全歸為合理使用。由此版權法變相鼓勵將閱讀行為外包給算法，導致人類深度閱讀能力退化。（2）從理論來看，“非表達性使用”將機器閱讀行為排除在了著作權法監管範圍之外，形成“灰色地帶”，有可能導致技術失控。此外，前述分析指出，無論是在技術上還是法律上，模型訓練是否使用了作品的具體表達都是存疑的（特別是針對軟件代碼作品的模型訓練），更勿論我國版權法上根本不存在非表達性使用的法律概念。[54] 就臨時複製而言，也存在相似的問題。與歐盟規定了嚴格的臨時複製合理使用例外不同，我國是將臨時複製視為沒有落入著作權法規制範圍，[55] 這可能與“非表達性使用”導致相同後果。事實上，非表達性使用和臨時複製分析也可以歸納為技術本位主義，在面對模型訓練技術複雜且高度依賴部署場景的新技術時，在法律上是不夠嚴謹且過於僵化。

其次，在采取合理使用規則時，應避免一刀切式的認定方式，而應基於利益平衡原則采取場景化的認定思路，明確指出哪些因素可能影響模型訓練的合理使用認定。相較於技術本位主義，場景化的合理使用認定更具有靈活性，能夠適應技術發展以及根據不同部署場景進行充分的利益平衡。上述歐美經驗即已表明，美國合理使用四要素能夠根據需要進行新的利益平衡，而歐盟則只能陷入僵化的版權法律傳統之中。與歐美相比，我國《著作權法》第 24 條采取的是 13 種具體條款 + 三步檢驗法的形式，本質上是半封閉且傾向於著作權人利益的。“法律、行政法規規定的其他情形”通常包括《信息網絡傳播權保護條例》《著作權法實施條例》所規定的合理使用情形。儘管難以從文義解釋上推導出利用作品進行模型訓練的合理使用情形，但 2011 年最高人民法院的指導意見明確提出，在促進技術創新和商業發展確有必要的特殊情形下，可以參考美國“合理使用四要素”來認定合理使用行為，之後我國司法實踐中也經常出現運用四要素進行合理使用判定的案例。[56] 這一司法實踐中的開放性實際上是為了回應新技術和新商業模式的需求，適當平衡作品使用者的利益。無論是在其他法律、行政法規中新設立模型訓練的合理使用情形，還是在司法實踐中采用四要素認定，都需要最終由三步檢驗法加以評價，這是由我國《著作權法》條文結構和《與貿易有關的知識產權協定》共同決定的。三步檢驗法的核心也即是不得不合理損害著作權人的合法權益，這在美國版權局報告中也明確建議：應根據作品性質、使用目的等因素綜合考慮是否損害原作品的潛在市場和價值。

（二）模型訓練合理使用認定的具體考慮因素

合理使用的利益平衡原則需要在個案中得以實現，基於三步檢驗法和域內外法律實踐，並結合技術實踐來判斷可能導致不合理損害著作權人合法權益的諸多情形，同時考慮哪些技術措施可能顯著減輕對著作權人合法權益的不合理損害，從而提出以下模型訓練合理使用認定的考慮因素。

1. 模型部署目的和對輸出的控制

模型部署目的和模型開發者對輸出的控制能夠顯著影響模型訓練對原作品潛在市場和價值的損害。（1）模型的最終部署。正如美國版權局的立場所言，合理使用必須在整體使用的背景下進行評估，這也是美國聯邦法院在“沃霍爾案”中的立場。實踐中，沒有僅僅訓練模型而不部署模型的，所有模型最終都用於一定的輸出，無非是輸出的內容不同，而輸出內容決定了部署目的。對於輸出的是識別人臉、輔助駕駛等決策信息的，很難說模型訓練影響到了原作品的潛在市場和價值。對於輸出的是人類可以閱讀的信息的，仍然需要考慮部署的真正目的，以及從用戶視角來判斷是否

[54] 參見易繼明：《大模型語料訓練合理使用問題研究》，載《中國版權》2024年第6期，第7頁。

[55] 參見王遷：《版權法保護技術措施的正當性》，載《法學研究》2011年第4期，第91頁。

[56] 參見《最高人民法院關於充分發揮知識產權審判職能作用推動社會主義文化大發展大繁榮和促進經濟自主協調發展若干問題的意見》第8條。

有別於作品之精神享受的目的。換言之，類似輔助醫生診斷的醫療信息輸出、輔助法官檢索案例的法律信息輸出等對原作品的潛在市場或價值通常不會造成損害或損害在合理範圍。而唯有輸出文學、藝術作品，並且以滿足用戶閱讀、欣賞其精神內核為目的，才有可能對原作品的市場造成不合理的影響。（2）部署者對模型輸出與被訓練作品實質性相似的文學、藝術內容，采取了有效技術措施以降低生成頻率的，可以認為模型訓練對原作品的市場損害較小或在合理範圍。例如模型訓練過程中采取“數據集清洗”等措施降低模型記憶現象，在模型輸出過程中采取關鍵詞屏蔽、反“越獄”“隨機種子”等措施。

2. 轉換性程度的高低

在（通常存在市場損害）市場損害難以估值的情形下，模型訓練使用作品的轉換性越高，對原作品的市場損害通常是越低的。在“沃霍爾案”中，美國聯邦法院提出類似商業性平衡原則：商業用途需更強的正當性支持，需與轉換性程度權衡。也就是說，在模型訓練普遍是商業性的情況下，較低程度的轉換性不足以支持其正當性。上述技術分析指出，以下因素通常會影響轉換性（在模型訓練的背景下，目的轉換性和內容轉換性有時難以區分）。需要注意的是，各種因素並非單獨發揮決定性作用，而是以權重的方式綜合發揮作用。並且不同因素之間存在關聯。例如預訓練階段通常使用的是海量質量相對較低的作品，而微調階段則使用的是針對性的質量相對較高的專業類作品。

（1）用於模型訓練的作品的類型。一般來說，法律、科學以及新聞等事實性較強的作品，模型訓練使用作品的轉換性通常較高。模型訓練使用這類作品的目的在很大程度上是學習作品中的事實、思想，而非欣賞其具體表達。在向用戶輸出內容的目的上，用戶通常不是為了獲取具體表達，而是為了了解一定的知識和新聞，以解決特定的問題。相反，類似文學、藝術作品等臆造作品，模型訓練在使用這類作品時的轉換性較低。雖然模型訓練是在獲取符號的分布規律，但是分布規律本身就是一種精確的表達選擇。當這一分布規律與表達選擇相似時，則很難稱得上具有轉換性。（2）作品質量的高低。高質量作品，例如高質量、專業創作的書籍、文章等作品對於模型模仿人類語言、故事結構、人物發展和主題輸出等至關重要，這些通常是模型想要充分模仿和學習的對象，轉換性程度通常較社交帖子、購物評論等較低。對於非專業創作的作品，模型旨在最大化地學習通俗的表達方式，模仿的是一般化的表達習慣，因此針對這類作品的轉換性程度較高。（3）模型訓練的階段。模型訓練主要可以分為預訓練和微調訓練階段。預訓練使用海量的無標註數據來對模型進行通識教育，類似於大學基礎課程；微調則是使用專業領域標註數據來針對某一具體目的進行專業培訓，類似於入職後的崗位培訓。可以看出，通識教育使用的作品風格、知識、表達特徵更加分散，單個作品對模型權重的影響力更低，作品使用的轉換性通常較高。而專業教育所使用的作品風格、知識和表達特徵更加集中，作品數量更少，單個作品對模型權重的影響力更高，則轉換性相對較低。在杭州互聯網法院首例生成式人工智能侵害信息網絡傳播權案中，用戶在基礎模型上使用幾張奧特曼圖片就訓練（微調）好一個可以定向生成奧特曼圖片的專有模型，這很難說具有內容轉換性。[57] 當然，模型訓練是一個漸進的過程，在預訓練的階段性過程中也有可能因為某一作者大量作品被反復訓練而產生“記憶”現象，由此導致內容轉換性程度較低，需要在個案中進行判斷。[58]

結合上述考慮因素，至少可以認為以下模型訓練更可能構成侵權：以輸出供用戶欣賞的文學、藝術作品為目的，同時根據基礎模型有針對性地對某類作品進行微調。而以下模型訓練更可能屬於

[57] 參見杭州互聯網法院(2024)浙0192民初1587號民事判決書。

[58] 有觀點即主張模型訓練學習大眾表達則屬於設定合理使用（但允許權利保留），若模仿個別作者且未獲許可應負侵權責任。參見李安：《機器學習的版權規則：歷史啟示與當代方案》，載《環球法律評論》2023年第6期，第97頁。

合理使用：以輸出輔助醫生診斷的醫療信息的模型訓練；以適用於泛化任務的通用基礎大模型的訓練，其采取有效措施降低“記憶”現象；雖然輸出文學、藝術作品，但未針對性地對模型進行微調訓練，也采取了各種技術措施降低模型輸出與被訓練作品實質性相似的內容。具言之，模型訓練使用作品的轉換性程度越高、潛在市場損害風險越低，則更有可能被視為合理使用，這應當在司法實踐中予以個案判斷。

Abstract: Compared to early automated writing bots like Dreamwriter, generative AI employs connectionist machine learning techniques. Beyond fair use claims based on policy considerations, most model training fair use determinations are grounded in technical analysis. This analysis posits that connectionist model training involves more transformative use of the original work in terms of purpose and content, thereby qualifying as fair use. The technology-centric fair use determination rules not only fail to align with technical realities but also lack legal comprehensiveness. They inadequately consider diverse scenarios, such as the ultimate deployment purpose of model training, and thus fall short of achieving a balanced consideration of creators' and users' interests in fair use determinations. Based on industry realities and legal traditions, the latest practices in Europe and the United States seek to maintain a delicate balance between the interests of copyright holders and technology companies. Guided by the principle of balancing interests, China's fair use determination for model training may comprehensively consider the following factors: First, whether the model's final deployment purpose and the technical measures implemented by developers/deployers to prevent infringing outputs significantly reduce the risk of potential market harm to the original work. Second, the impact of the type of trained work, its quality, and the training stage on the degree of transformative use. The higher the level of transformative use in model training and the lower the risk of potential market harm, the more likely it is to be deemed fair use. This should be determined on a case-by-case basis in judicial practice.

Key words: Model Training; Copyright Risk; Fair Use; Balancing of Interests

（責任編輯：昝晨東）