

從真實到合成：資料信任體系的重塑

From Real to Synthetic: Reshaping the Data Trust System

真実から合成：データ信頼システムの再構築

文・圖／MIC產業顧問兼主任 勵秀玲

在人工智慧（AI）快速發展的過程中，資料一直被視為驅動模型進化的核心資源。在現實世界資料有限、成本高昂或涉及敏感資訊的情況下，合成資料（synthetic data）可以彌補資料缺口。

合成資料是指利用數學模型、統計學演算法或 AI 生成模型的人工生成資訊。在理想狀態下，合成資料能保留原始資料的統計特徵，卻不包含任何特定個體的敏感資訊，為長久以來的隱私保護與資料流通難題提供了技術解方。

然而，這帶來一個根本問題：當我們賴以決策的基礎從「真實觀測」轉向「人工生成」，數位世界的信任基石是否會受到威脅？合成資料的崛起，不只是技術突破，對既有資料信任機制也面臨結構性挑戰。

從觀測到生成的資料典範轉變

合成資料的興起，是市場與治理環境壓力驅動下的結果。首先，隱私法規與資料主權的強化，使真實資料的取得與共享變得困難。如歐盟的《一般資料保護規定》(GDPR) 等法規大幅提高資料使用門檻，使企業與機構在跨機構、跨國資料流動上面臨限制。其次，AI 模型對資料的需求急速增加，從傳統機器學習到大型語言模型和多模態系統，資料量與多樣性的需求呈現指數成長。第三，真實資料的蒐集與標註成本高昂且耗時，使資料供應逐漸成為 AI 發展的瓶頸。

在這樣的背景下，合成資料逐漸從輔助工具轉變為資料基礎設施。例如，在金融領域，機構透過模擬交易行為生成大量異常樣本，作為提升詐欺偵測能力；在醫療領域，合成病歷讓跨機構研究得以在不暴露個資的情況下進行；在自駕車

領域，可以快速生成各種極端天候與罕見事故場景，加速模型訓練。

這些應用凸顯資料不再只是被動蒐集的資源，而是可以被主動設計與生成的資產。資料供應鏈從「蒐集、清理、使用」的線性流程，轉變為「生成、調整、優化」的循環系統。

資料生成時代的信任斷層

當資料的來源從「觀測現實」轉向「模型生成」，信任的基礎也隨之改變。傳統上，資料之所以可信，在於其直接反映現實世界；即使存在誤差，仍可透過統計方法逐步修正。然而，合成資料並非對現實的直接記錄，而是由模型、參數與設計選擇所生成，其可信度取決於生成邏輯與使用情境。

這種轉變使過去建立在「來源可信」之上的信任模式，逐漸不適用，取而代之的是更難察覺的風險結構，導致信任機制出現斷層。首先，合成資料往往基於既有資料生成，若原始資料存在偏差，模型可能在生成過程中強化這些偏誤，進而影響決策。再者，生成模型可能產生看似合理但實際不存在的變數關聯，使系統在關鍵情境中做出錯誤判斷。此外，若模型持續使用自身生成的資料進行訓練，可能導致分布偏移，最終偏離世界實際運作的規律。





醫療領域的合成資料應用，揭示這種風險的具體樣貌。如部分醫療期刊相關研究中指出，合成電子病歷資料在生成過程中，可能扭曲疾病關聯或低估特定族群風險，使模型在臨床判斷上產生偏差。這些資料在統計上看似合理，甚至能通過模型效能測試；但在臨床應用中，可能導致誤診或不當決策。這類問題往往不易在傳統驗證流程中被發現，但卻會在實際決策中放大其影響。

因此，在合成資料時代，信任的基礎正轉向「資料是否經過可理解、可控制且可驗證的生成與使用過程」。

合成資料帶來的三大治理挑戰

雖然合成資料帶來速度與效率的優點；若缺乏治理，可能成為企業隱藏的風險來源。首先，合成資料的品質取決於隱私保護、資料效用與公平性之間的平衡，其三者之間存在結構性的權衡關係，無法同時最佳化。當強化隱私保護時，通常需加入噪音或降低資料精細度，進而影響分析效用；當提升資料效用，使其更接近真實分布，則可能增加隱私風險；而為改善公平性而調整資料分布時，又可能影響模型準確度。這意味著，合成資料是需要多重目標之間進行權衡取舍的治理議題。

其次，治理重心從資料蒐集轉移到資料生成與使用。過去的資料治理主要關注資料來源是否合法、是否符合隱私規範；但合成資料的治理問題是

資料如何生成、生成過程中引入了哪些假設、以及這些資料如何被使用。

第三，是責任歸屬的模糊化。在傳統模式下，資料提供者與使用者之間的責任相對清晰；但在合成資料生態中，資料生成者、模型開發者、平台與應用端之間的界線變得模糊。當決策出現問題時，責任應由誰承擔則往往難以界定。

從資料信任到治理信任

在這樣的轉變下，合成資料要走向健康發展的路徑，需建立在三個關鍵支柱上。第一是透明性，即資料的來源、生成方法與使用情境必須清楚揭露，讓使用者能理解資料的限制與風險。第二是可追溯性，透過資料來源與流向的記錄，確保每一筆資料的生成與使用過程可被追蹤。第三是可驗證性，透過系統化的審計與評估機制，檢測資料的偏誤、隱私風險與效用。這三個支柱共同構成可信資料體系的核心基礎。

在這個體系中，資料是嵌入在一個包含生成、處理、使用與監管的完整生態中，而信任是治理後的產物。顯示未來競爭的焦點將從資料量轉向治理能力，能夠建立透明、可追溯且可驗證資料體系的組織，將取得關鍵優勢。

以治理駕馭合成資料的信任與價值

總結而言，合成資料的出現，代表著資料經濟進入新階段。資料不再只是被動蒐集的資源，而是可以主動生成與設計的資產。當資料可以被創造，能否建立可信的使用機制，是決定價值的關鍵。

未來的競爭不在於誰擁有最多資料，而在於誰能讓資料被信任。這種信任，來自制度設計、治理能力與持續監督。從這個角度來看，合成資料並非取代真實資料，而是推動資料治理進化的催化劑。

MIC AISP 網址：<https://mic.iii.org.tw/AISP>

著作權所有，非經資策會書面同意，不得翻印或轉讓。

以上研究報告資料係經由MIC 內部整理分析所得，並對外公告研究成果，由於產業倍速變動、資訊的不完整，及其他不確定之因素，並不保證上述報告於未來仍維持正確與完整，引用時請注意發佈日期，及立論之假設或當時情境，如有修正、調整之必要，MIC 將於日後研究報告中說明。敬請參考MIC 網站公告之最新結果。