



2025 年《高科技、媒體及電信產業趨勢預測》

目錄

一、重點摘要	02
二、女性與生成式 AI：使用差距正在快速縮小，但信任差距仍然存在	04
三、隨著生成式 AI 對運算能力的需求不斷攀升，資料中心應尋求更可靠、潔淨的能源解決方案	13
四、自主生成式代理型 AI (Agentic AI)：正在研發中	25
五、深度偽造技術的衝擊：網路安全層級挑戰及其深遠影響	34
六、雲端變的精簡：「雲端財務管理 (FinOps)」讓每一分錢都能發揮更大的效益	40
七、生成式 AI 是否能復甦智慧型手機市場	46
八、大型影視製作公司使用生成式 AI 進行內容創作時會比較謹慎，而社群媒體則積極使用	53
九、無線電信業者在監管機關允許的條件下加速整合	59
十、更新動態：回顧過往《高科技、媒體及電信產業趨勢預測》成果	66
• 生成式 AI 邁向企業邊緣：「地端 (On-Prem) AI」蓬勃發展	66
• 固定無線接入 (FWA)：與普遍觀點相反，應用比例將會持續成長	69
• 5G 獨立組網 (5G SA) 發展緩慢：6G 是否會延遲問世？	71
• 開放式無線存取網路 (Open RAN) 行動網路和供應商選擇： 現今是單一供應商，何時會實現多供應商模式？	72
• 量子技術起步雖緩慢，但仍應盡早規劃防範措施	74
• RISC-V：彌補地緣政治生成式 AI 漏洞	76
十一、新興趨勢：值得關注的新興技術	84
• 生成式 AI 與網路安全：高風險同時也帶來巨大機會	84
• 矽晶基本模組：小晶片 (Chiplet) 有望推動摩爾定律持續發展	86
• 電信運營支撐系統 (B/OSS)：電信業者將其業務與營運支援系統軟體進行現代化升級	88
• 矽光子技術：生成式 AI 以光速傳輸	90
十二、聯絡我們：勤業眾信高科技、媒體及電信產業服務團隊	95

重點摘要

2025 年《高科技、媒體及電信產業趨勢預測》 (TMT Predictions)：縮小差距

Deloitte 預測，2025 年將成為生成式 AI 和高科技、媒體及電信（TMT）產業的「關鍵過渡年」，必須縮小八大關鍵差距，以發揮現有的技術潛力

展望 2025 年和未來，高科技、媒體及電信產業正處於積極發展的關鍵時刻，其驅動力主要來自於快速使用生成式 AI。然而，要實現這一願景，產業還需要縮小以下差距，其中包括：平衡生成式 AI 基礎設施投資與商業化過程、解決生成式 AI 使用中的性別差異、管理生成式 AI 資料中心的耗能、大眾對深度偽造內容的信任問題、探索生成式 AI 在媒體和遊戲領域的最佳應用、應用代理型 AI 實現即時管理與決策，以及填補串流媒體影音和雲端支出方面的缺口。

此外，亦具有與差距之外的預測，包括搭載生成式 AI 晶片的新款智慧型手機與電腦，以及電信業、無線通訊領域的整合。克服這些困難，將對於促進產業與業務成長至關重要。

八大差距，使 2025 年成為高科技、媒體及電信業的「差距年」 (Gap Year)

- 1. 生成式 AI 的基礎建設與商業化差距：**正如去年預測，企業為支援 AI 模型的訓練與推理，正投入數百億美元購買晶片，並投入數千億美元建設生成式 AI 資料中心。儘管部分提供企業級生成式 AI 軟體的公司營收已有所提升，但目前的投資規模仍至少高出回報十倍。對於投入最多的企業而言，生成式 AI 投資不足的風險可能高於過度投資的風險。然而，投資與回報之間的差距仍然存在，且似乎仍在擴大。
- 2. 新一代 AI 資料中心電力和永續發展差距：**擬建 AI 資料中心將消耗較以往更多電力，並優先考慮低碳能源，這導致電力需求、電網供應能力及企業永續目標之間產生差距。全球七大資料中心（又稱超大規模企業，Hyperscalers）、晶片公司與公共事業單位，正積極採取措施縮小差距，但預期到 2025 年仍將存在此問題。

3. **生成式 AI 的性別差距：**女性使用生成式 AI 工具進行工作和娛樂的比例低於男性。這一現象部份源自於對技術信任度較低。然而，預計在部分市場中，女性使用生成式 AI 的比例將會在一年內左右趕上男性。
4. **生成式 AI 深偽內容的信任差距：**隨著深偽生成式 AI 內容大幅增加，涵蓋影像、影片及音樂等形式，大眾對於內容正面臨挑戰。為縮小此差距，生成式 AI 生態系統必須全面註記且不可竄改地標註生成式 AI 內容，並確保能夠即時偵測偽造影像。目前，製作高擬真之深偽內容的邊際成本正在不斷降低，因此偵測技術的成本亦須同步降低，將有助於縮小信任差距。
5. **影視製作公司的生成式 AI 使用差距：**目前預測大型影視製作公司已廣泛應用生成式 AI 進行內容製作，儘管部分影視製作公司確實使用該技術，但實際狀況與預期仍有差距。許多人對生成式內容所涉及的智慧財產權（IP）的挑戰抱持謹慎態度，但同時也希望透過 AI 強化企業競爭力，以縮短製作時間、降低成本，擴大業務範圍與市場影響力。
6. **自主生成式代理型 AI 差距：**能夠穩定且可靠地獨立執行任務，並協調整體工作流程的自主 AI 機器人，展現出極具潛力的發展前景。代理型 AI 計畫已於 2024 年試行計畫陸續啟動，這些技術是否能在 2025 年實現普及應用，仍有待觀察。
7. **串流影音差距：**許多媒體和娛樂公司原先預期消費者會長期訂閱多項串流服務，然而如今消費者傾向透過網綁訂閱並取消部份服務，以降低整體訂閱成本。目前觀察顯示，每戶家庭訂閱的串流服務數量不僅停滯甚至正在減少。為填補成長缺口，串流業者更加依賴網綁策略彌補差距，並透過與第三方整合和內容共享，以提升市場競爭力。
8. **雲端運算成本的差距：**雲端運算最初的核心賣點之一是成本較低，但實際支出卻呈現分散且難以管控。部分企業正使用雲端成本管理（FinOps）來縮小承諾節省成本與實際支出之間的差距，以優化雲端支出，並可能節省數十億美元。

本年度的新內容

在**更新動態：回顧過往《高科技、媒體及電信產業趨勢預測》成果**章節中，我們將重新回顧先前《高科技、媒體及電信產業趨勢預測》（TMT Predictions）報告提出的主題，以檢視我們的預測表現如何，同時將更新最新預測。接著我們在新興趨勢：**值得關注的新興技術**章節中，將聚焦在高科技、媒體及電信產業的四項尖端話題，這些新興主題可能尚未納入主流預測，但是我們相信這些將會成為未來產業對話中備受矚目的隱藏瑰寶。

女性與生成式 AI：使用差距正在快速縮小，但信任差距仍然存在

女性對生成式 AI 應用抱持高度期待，科技公司應積極提升信任度、減少技術偏見，並致力打造更具代表性的女性勞動結構

Deloitte 預測，至 2025 年底，美國女性嘗試和使用生成式 AI 的比例將會追平或超越男性。¹ 在 2023 年，女性使用生成式 AI 的比例僅有男性的一半，但是使用速度顯示可能會在未來一年與男性並駕齊驅。² 儘管此項趨勢預測是以美國女性為對象，但生成式 AI 的性別差距為全球現象。同樣對歐洲國家進行生成式 AI 使用情況調查，我們不僅發現明顯的性別使用差異，且女性使用者數量正在迅速增加。³ 這些國家很有可能將在接下來的 2 年內縮小使用的性別比例差距，而全球面臨使用的挑戰與機會，將可能與美國的調查結果類似。

儘管女性使用生成式 AI 的速度正在加快，但對於生成式 AI 業者是否能確保其資料安全上的信任，比例仍低於男性。⁴ 這種「科技信任差距」將可能影響女性對該生成式 AI 技術的日常使用狀況、完整應用程度，以及減緩相關產品與服務的購買意願。為了補足此信任差距，科技公司應加強資料安全性，並落實明確資料管理，提供健全的資料控制權。

AI 模型的偏見可能會對信任產生負面影響。⁵ 研究顯示，女性在 AI 相關領域的勞動力占比不到 1/3，⁶ 多數 AI 從業人員認為，若該領域持續由男性主導，AI 將會更容易產生偏見。⁷ 提升女性在 AI 領域的參與度，不僅有助於減少 AI 系統中的性別偏見，亦能為女性創造更多機會，引領這項技術的發展。

生成式 AI 的使用差距正在迅速縮小

Deloitte 最新研究強調了各區域使用生成式 AI 方面的性別差距。在過去兩年，Deloitte《互聯消費者調查》(Connected Consumer Survey) 調查了美國消費者使用生成式 AI 的情況，作為其數位生活研究的一部分。⁸ 分析結果顯示，美國女性在接受此項新興技術方面落後於男性（圖 1）。2023 年，女性使用生成式 AI 的比例約為男性的一半（11% 的女性表示曾試用生成式 AI，或除試用外曾在專案及工作中使用此項技術，而男性的比例為 20%）。2024 年，同一項研究結果顯示生成式 AI 的整體使用率增加了超過 2 倍，但是仍存有性別差距，33% 的受訪女性表示正在使用或曾嘗試使用生成式 AI，而男性的比例為 44%。

其他地區同樣存在生成式 AI 的性別使用差距：Deloitte 英國 2024 年發布的《數位消費者趨勢》(Digital Consumer Trends) 調查顯示，英國有 28% 的女性使用生成式 AI，而男性比例為 43%。⁹ 此項研究的分析結果及 Deloitte 英國的針對歐洲地區生成式 AI 與信任研究皆顯示，在歐洲其他 12 個國家中，男女性使用生成式 AI 的比例差異均達到雙位數的明顯差距。¹⁰

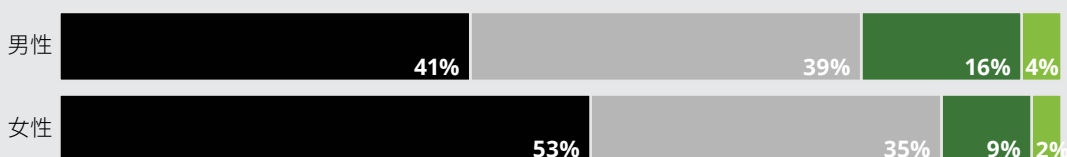
圖 1

美國女性消費者使用生成式 AI 的比例落後男性，但是可能不會持續太久

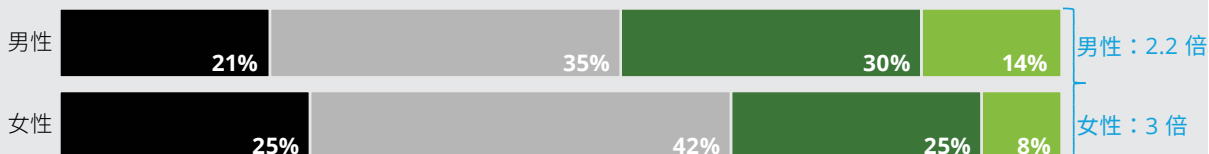
與生成式 AI 系統及技術的關係，2023 至 2024 年

● 不熟悉生成式 AI ● 熟悉生成式 AI，但是從未嘗試 ● 曾經試用生成式 AI ● 曾經使用生成式 AI 處理專案 / 工作

2023



2024



備註：美國消費者：n（女性）= 1,040（2023 年）、1,992（2024 年），n（男性）= 962（2023 年）、1,841（2024 年）。

資料來源：Deloitte 2024 年《互聯消費者調查》(Connected Consumer Survey) 第五版、Deloitte 2023 年《互聯消費者調查》(Connected Consumer Survey) 第四版。

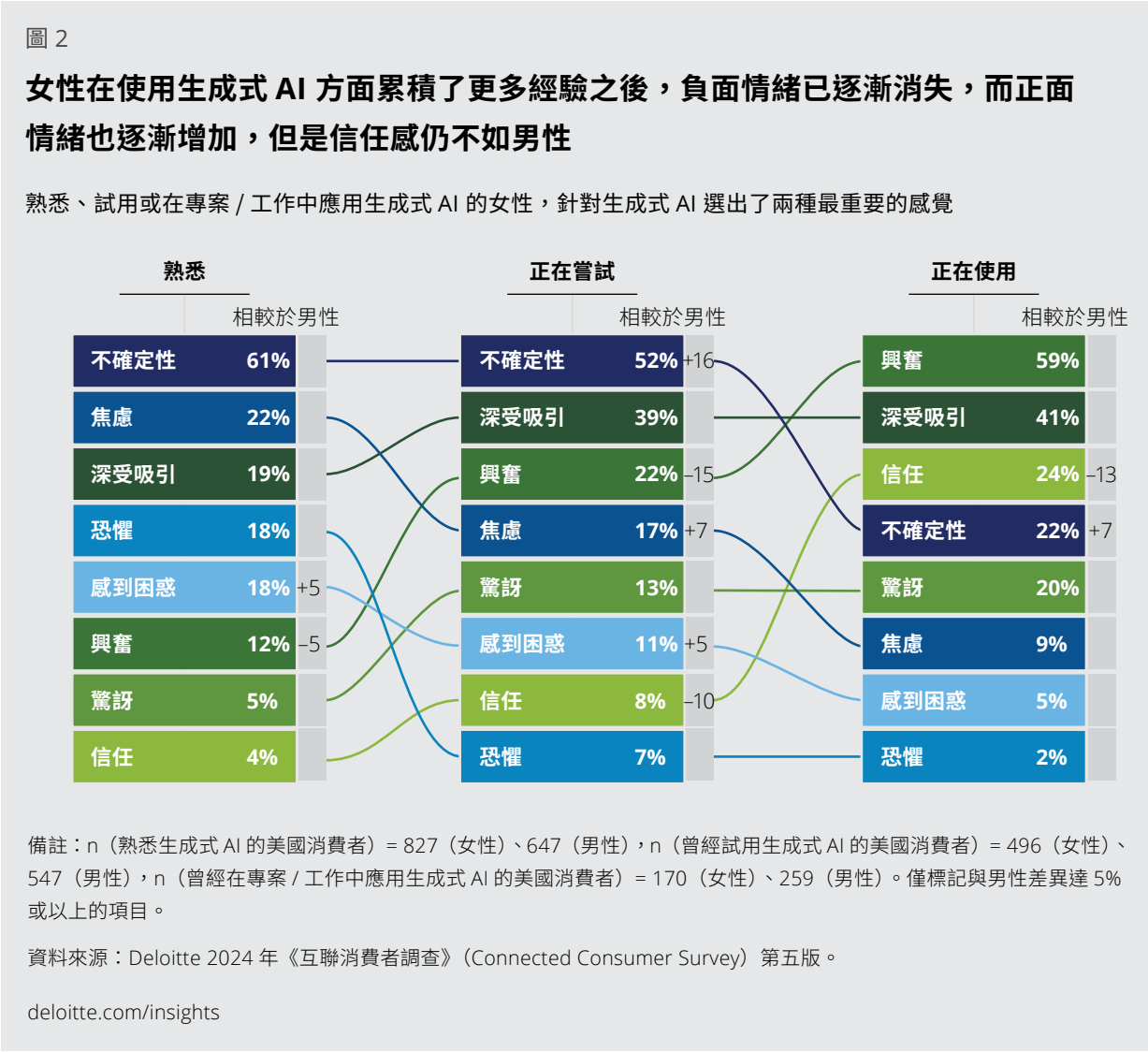
deloitte.com/insights

在美國，女性使用生成式 AI 差距正在迅速縮小。過去一年內，接受調查的美國女性受訪者在使用生成式 AI 的比例成長了 3 倍，超越男性的 2.2 倍成長率。¹¹ Deloitte 根據當前使用之水準分析與成長率，預測在 2025 年底，美國女性試用和在專案或工作中應用生成式 AI 的比例將會追上或超越男性比例。¹²

完全參與仍面臨挑戰

雖然生成式 AI 的使用趨勢看起來振奮人心，但即使達到使用率的平等，不代表女性會將生成式 AI 納入日常工作流程中。Deloitte 2024 年《互聯消費者調查》(Connected Consumer Survey) 結果指出，受訪生成式 AI 的使用者中，34% 的女性表示至少每天使用一次此技術，而男性為 43%；¹³ 在專業工作中使用此技術的受訪者中，41% 的女性認為生成式 AI 大幅提升了生產力，而男性比例則為 61%。¹⁴ 希望從使用生成式 AI 中創造效益的科技公司和其他組織，更應關注這些性別差異，並主動採取措施提升女性的參與度。

兩性在生成式 AI 的使用差異，部分可能是源自於對技術信任的明顯差距。¹⁵ 隨著女性從熟悉生成式 AI 進而嘗試和使用，對技術的不確定性、焦慮、恐懼和困惑等負面情緒逐漸減少，並增加了興趣、興奮、驚訝與信任等正面情緒 (圖 2)。¹⁶ 不過，在嘗試及於專案與工作應用階段，女性對生成式 AI 的信任度明顯低於男性，且不確定感仍相對較高。調查顯示，目前正在嘗試或使用生成式 AI 的女性中，僅有 18% 表現出「高度」或「非常高度」信任生成式 AI 服務業者能確保其資料安全，而男性使用者的比例則已達到 31%。¹⁷



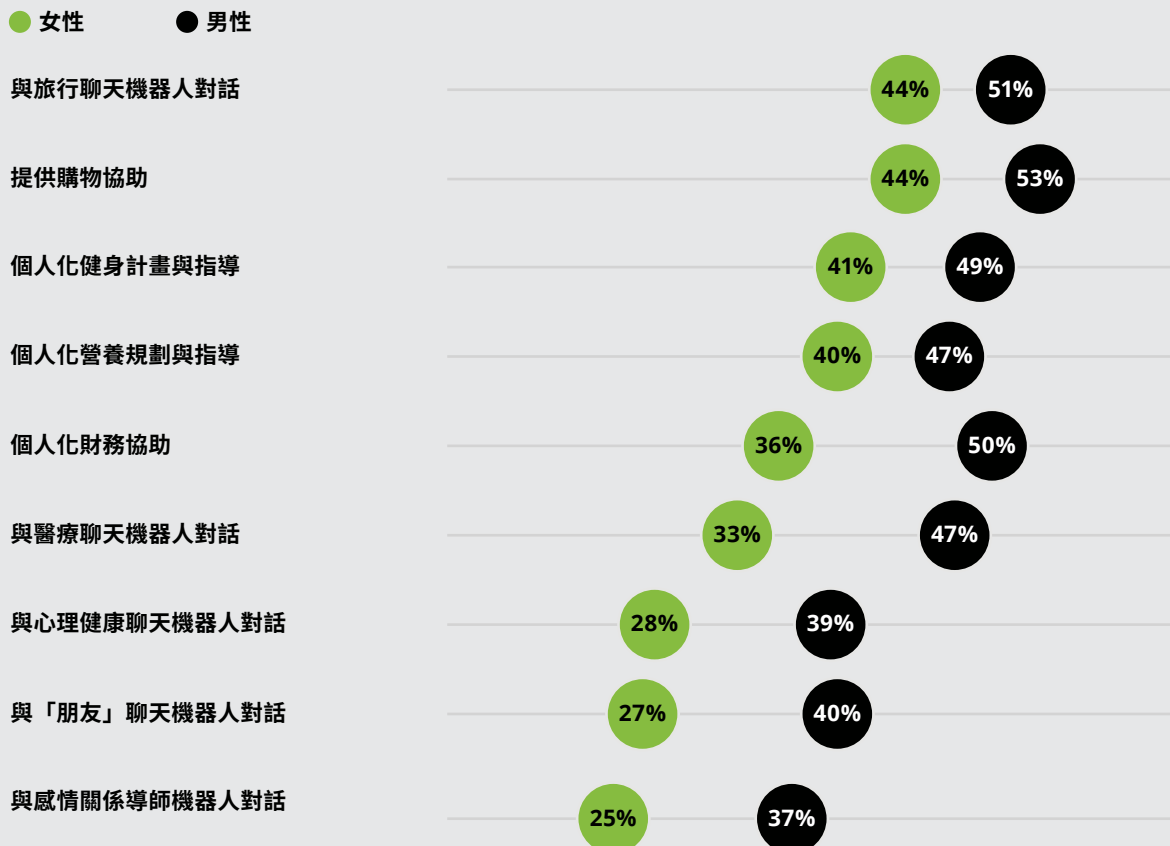
信任差距不僅限於生成式 AI 上，同時影響更廣泛的科技服務和互動中。根據 Deloitte 2024 年《互聯消費者調查》(Connected Consumer Survey) 指出，54% 的女性認為線上服務帶來的效益，已超越對資料隱私的顧慮（相較 2023 年的 46% 有所提升），而相同看法的男性比例更高，達 62%。¹⁸ 去年，我們的報告指出，相較於男性，女性對於個人資料的使用和保護機制更加謹慎，並影響了女性分享資料的意願，尤其是涉及敏感的健康與健身指標時，¹⁹ 女性可能認為資料外洩或資料濫用的潛在結果更嚴重。²⁰

隨著生成式 AI 的日益普及，長期存在的資料隱私和技術方面的問題將可能有所加劇。²¹ 在使用者與生成式 AI 互動時，系統可能會將使用者資料回饋至其 AI 模型。而專家指出，不同意將個人資料使用於 AI 訓練的選項，總是較不明確或不容易操作。²² 當消費者開始與生成式 AI 討論敏感的個人話題，資料隱私和資安風險將進一步提升。事實上，資料隱私和安全方面的信任差距，可能是導致女性與男性未來對各種生成式 AI 體驗之興趣程度差異的關鍵因素（圖 3）。²³ 調查顯示，女性對於與生成式 AI 討論較不敏感的話題（如旅行、購物、健身、營養）的興趣較男性低，但涉及更敏感的內容時（例如個人財務狀況、親密關係、醫療或心理健康議題），女性與生成式 AI 的興趣是明顯低於男性的。

圖 3

女性在體驗各種生成式 AI 功能方面的興趣低於男性

受訪者表示對下列生成式 AI 體驗相當 / 稍微有興趣的比例



備註：n（熟悉、試用或使用生成式 AI 的美國消費者）= 1,492（女性）、1,454（男性）。

資料來源：Deloitte 2024 年《互聯消費者調查》(Connected Consumer Survey) 第五版。

deloitte.com/insights

信任差距也可能是導致女性對購買新生成式 AI 技術興趣較低的原因之一。科技公司正陸續推出內建 AI 晶片的筆記型電腦、平板電腦和智慧型手機，以提升裝置功能。例如，即時資訊摘要、照片和影片生成，以及即時語言翻譯。²⁴ 根據 Deloitte 2024 年《互聯消費者調查》（Connected Consumer Survey）指出，當受訪者被問到 AI 新功能是否會影響其升級裝置的計畫時，女性對於更快升級裝置的意願比例略低於男性。²⁵ 舉例來說，43% 使用智慧型手機的男性表示，內建 AI 功能非常或有可能會促使他們提前升級手機，而僅有 32% 的女性持相同看法；相反地，58% 的女性表示 AI 功能不會影響其升級計畫，而持相同看法的男性比例為 50%。在筆記型電腦方面，41% 的男性表示內建 AI 功能可能會促使他們更快升級裝置，而女性的比例為 28%。由於女性在全球掌控或影響了約 85% 的消費支出，她們對於升級至搭載 AI 裝置的興趣較低，可能會為技術供應商帶來挑戰。²⁶

信任差距並非唯一影響女性充分使用生成式 AI 的因素，據調查顯示，女性生成式 AI 使用者認為，任職的公司積極鼓勵在工作時使用此技術的比例較低（61% 的女性具相同想法，男性則為 83%）。²⁷ 儘管有 49% 的生成式 AI 女性使用者表示公司願意投資培訓員工使用生成式 AI，但是整體比例仍落後於受訪男性的 79%。不論這些數字是否反應出看法的差異，或實際在職場參加培訓課程和受到鼓勵的經驗，企業都應重視這一問題，並採取措施縮小性別。

科技業的女性從業人員正在積極推動生成式 AI 的發展，但代表性仍需提升

在科技業中，女性對於生成式 AI 的使用情況呈現不同的發展趨勢，在科技業工作的女性，在未來可能將推動更多的女性參與。在開發 AI 產品及服務的從業人員，使用生成式 AI 的程度也普遍較高。根據 Deloitte 2024 年《互聯消費者調查》（Connected Consumer Survey）指出，70% 的女性與 78% 的男性科技業從業人員表示，正在試用生成式 AI 或應用於專案或工作中，比例遠高於非科技業的從業女性（32%）和男性（40%）。²⁸ 更值得關注的是，從事科技業的女性似乎比男性更快從試用階段轉向實際應用在專案及工作中（女性 44% 相較於男性 33%）。兩個群體都期盼能帶來更大的效益，約 70% 的受訪者預測，在未來一年內，他們的生產力將因生成式 AI 而有大幅提升。²⁹

此外，從事科技業的女性與男性之間無明顯的信任差距。相較於整體使用者而言，兩者對生成式 AI 的信任度更高，超過 40% 使用或試用生成式 AI 的科技業女性和男性表示，其對資料安全抱持「高度」或「非常高度」的信任；³⁰ 而 75% 的受訪者認為，從線上服務獲得的效益大於對隱私的顧慮。反觀科技業以外的產業，則僅有 54% 的女性和 60% 的男性具有相同看法。³¹ 這顯示從事科技業的女性可能比非科技業人員更瞭解生成式 AI 的運作方式，且在工作中頻繁使用生成式 AI，可以提升對其信賴度，進而更理解此技術所帶來的優勢。此外，多數使用生成式 AI 的科技業女性表示，企業支持且鼓勵其使用（84%）並提供相關培訓（72%），相較之下，其他產業使用生成式 AI 的女性，則較少獲得公司支持使用（55%）或提供培訓（45%）的機會。³²

儘管科技業女性使用生成式 AI 的程度較高，但 AI 領域的女性從業人員比例仍相對較少。目前，在 AI 領域相關勞動力中，女性僅占 30%，此比例與 STEM 領域（科學、科技、工程、數學）的整體女性代表比例相當。³³ 女性在 AI 領域的代表性不足，可能對 AI 系統在各領域和產業間的系統發展與部署，皆產生深遠影響。

在 AI 領域中女性人才比例較低，可能導致 AI 應用中持續對女性有性別偏見的風險。³⁴ 研究顯示，約 44% 的 AI 系統存在性別偏見，這可能影響 AI 輸出結果，使女性在職場、醫療、金融等領域持續面臨不利處境。³⁵ 舉例來說，AI 可能在招聘流程中產生性別偏見，導致女性求職機會受限，³⁶ 亦可能影響醫療決策的準確性，³⁷ 或降低女性獲取金融服務的機會。此外，³⁸ Deloitte 研究指出，AI 模型中的偏見可能會降低員工與客戶對 AI 技術的信任。³⁹ 因此，讓更多女性投入 AI 領域工作，不僅有助於實現性別平等，進而為整個社會帶來更廣泛的正面影響。⁴⁰

結論

科技公司應努力提升女性參與生成式 AI 發展的原因眾多，主要原因包含以下幾點。首先，女性在消費決策中擁有高度影響力或控制權，若未能讓女性積極參與並頻繁使用生成式 AI，可能會增加 AI 產品與服務無法達到預期潛力的風險。其次，若女性員工未能與男性員工同等參與使用生成式 AI 工具，企業可能將面臨無法實現對 AI 投資預期的生產力提升。同時，因為生成式 AI 依賴互動數據進行學習，若女性互動的代表性不足，則可能會加深 AI 模型中的性別偏差。⁴¹ 最後，若女性無法充分參與新興的生成式 AI 應用，可能會導致其無法獲益於未來科技發展（如聊天機器人在醫療或心理健康領域的應用優勢），並將加深現有的不平等現象。⁴²

為了加強女性對生成式 AI 的信任，科技公司應致力解決相關潛在風險。Deloitte 2024 年《互聯消費者調查》（Connected Consumer Survey）指出，提升消費者信任的關鍵因素在於提高科技公司資料隱私和資安政策的透明度，同時讓消費者能更輕鬆地掌控個人資料使用方式。⁴³ 科技公司應優先採取健全的資料安全措施，並更有效溝通資料處理方式，讓消費者清楚瞭解蒐集資料及其用途，同時提供更便捷的控制方式（例如，在適當時機提醒使用者做出明確的數據使用選擇），不僅有助於建立信任，也能創造競爭優勢。此外，生成式 AI 的潛在風險不僅是科技公司的責任，調查顯示有 84% 的受訪者認為政府應加強監管，進一步規範企業數據收集與使用行為。⁴⁴

在各產業中，想要實現讓男女員工都能充分利用生成式 AI 的企業，應重視並積極推動其應用。除常見的專業用途（如文件編輯、網路搜尋、內容摘要和研究輔助）外，企業還應探索各產業特有的生成式 AI 應用場景。⁴⁵ 為了最大化員工對生成式 AI 的使用，公司可能需要建立專門的培訓計畫，以確保員工具備必要的技術能力。

致力於在生成式 AI 中實現消費者全面參與是一項值得肯定的目標，但若開發其技術的團隊缺乏多元與包容性，則可能難以實現。企業應營造符合員工需求的職場環境，以提升女性在 AI 領域工作上的多元性和包容性。例如，一項針對女性在 AI 領域的研究顯示，工作與生活的平衡是影響工作滿意度最重要的關鍵因素，包括彈性工時與遠距工作。⁴⁶ 此外，女性在求職時亦會關注是否有女性擔任管理職務、薪資和升遷的透明度，以及對騷擾與霸凌的零容忍政策。⁴⁷ 為吸引更多女性投入 AI 領域，企業可能需要提供更多的教育和訓練機會，以培養 AI 相關的技能及能力。同時，透過指導和交流課程，讓 AI 產業中的女性分享經驗及相互支援，並提供資金以促進女性參與 AI 研究與創新專案，將有助於提升其參與度。隨著女性在生成式 AI 開發中的角色日漸重要，未來可能會出現更能吸引女性使用的應用與系統。

By **Susanne Hupfer**
United States

Jeff Loucks
United States

Ariane Bucaille
France

Gillian Crossan
United States

Bree Matheson
United States

Endnotes

1. To understand consumer attitudes toward digital life, the Deloitte Center for Technology, Media & Telecommunications surveyed 3,857 US consumers in the second quarter of 2024 and 2,018 US consumers in the second quarter of 2023. These 2024 and 2023 Connected Consumer Surveys collected data on consumers' reported adoption of generative AI, including experimentation and use for projects and tasks (beyond experimentation). By analyzing longitudinal adoption data and calculating the rate of change in adoption from 2023 to 2024 for men and women, we are able to project that women will close the adoption gap by the end of 2025; see: Jana Arbanas et al., Earning trust as gen AI takes hold: 2024 Connected Consumer Survey, 5th edition, Deloitte, December 3, 2024; Jana Arbanas, Paul H. Silverglate, Susanne Hupfer, Jeff Loucks, Prashant Raman, and Michael Steinhart, "[Balancing act: Seeking just the right amount of digital for a happy, healthy connected life](#)," Deloitte Insights, Sept. 5, 2023.
2. Ibid.
3. Our analysis was conducted from August to October 2024, based on data from Deloitte UK's 2023 and 2024 Digital Consumer Trends surveys, as well as a 2024 Deloitte UK survey of European consumers on the topic of generative AI; see: Paul Lee and Ben Stanton, "[Generative AI: 7 million workers and counting](#)," Deloitte, June 25, 2024; Jonas Malmund, Frederik Behnk, and Joachim Gullaksen, "[Generative AI is all the rage](#)," Deloitte, 2023; Roxana Corduneanu, Stacey Winters, Jan Michalski, Richard Horton, and Ram Krishna Sahu, "[Europeans are optimistic about generative AI but there is more to do to close the trust gap](#)," Deloitte Insights, Oct. 10, 2024.
4. Analysis based on Deloitte's 2024 Connected Consumer Survey; see: Arbanas et al., Earning trust as gen AI takes hold: 2024 Connected Consumer Survey.
5. Don Fancher, Beena Ammanath, Jonathan Holdowsky, and Natasha Buckley, "[AI model bias can damage trust more than you may know. But it doesn't have to.](#)" Deloitte Insights, Dec. 8, 2021.
6. World Economic Forum, "[Global gender gap report 2023](#)," June 2023.
7. Deloitte AI Institute, "[Women in AI](#)," accessed November 2024.
8. Jana Arbanas et al., Earning trust as gen AI takes hold: 2024 Connected Consumer Survey, 5th edition, Deloitte, publishing December 3, 2024; Arbanas, Silverglate, Hupfer, Loucks, Raman, and Steinhart, "[Balancing act](#)."
9. Deloitte, "[Generative AI: 7 million workers and counting](#)," accessed November 2024.
10. The Digital Consumer Trends study conducted in various countries in 2024 revealed gen AI adoption gaps of 17 points in Denmark; 12 points in Sweden, Italy, and the Netherlands; 11 points in Belgium; and 10 points in Norway. Additional analysis of a Deloitte European gen AI study revealed gen AI adoption gaps ranging from 10 to 15 points in 11 European countries studied (Belgium, France, Germany, Ireland, Italy, the Netherlands, Poland, Spain, Sweden, Switzerland, and the United Kingdom); see: Deloitte, "[Generative AI](#)"; Deloitte, "[Generative AI is all the rage](#)," accessed November 2024; Corduneanu, Winters, Michalski, Horton, and Sahu, "[Europeans are optimistic about generative AI but there is more to do to close the trust gap](#)."
11. Analysis based on 2024 and 2023 Deloitte Connected Consumer Surveys; see: Arbanas et al., Earning trust as gen AI takes hold: 2024 Connected Consumer Survey; Arbanas, Silverglate, Hupfer, Loucks, Raman, and Steinhart, "[Balancing act](#)." Deloitte, "[Generative AI](#)."
12. Ibid.
13. Arbanas, et al., Earning trust as gen AI takes hold: 2024 Connected Consumer Survey.
14. Ibid.
15. Ibid.
16. Ibid.
17. Ibid.
18. Ibid.

19. For example, only 43% of women we surveyed in the Deloitte 2023 Connected Consumer Survey who owned smart watches or fitness trackers said that they share the data collected by those devices with their health care provider, vs. 57% of men; see: Susanne Hupfer, Jennifer Radin, Paul H. Silvergate, and Michael Steinhart, [“Tech companies have a trust gap to overcome—especially with women,”](#) Deloitte Insights, Nov. 8, 2023.
20. These fears may be warranted. Consider that most health apps—along with the data they gather and transmit—are not covered by the Health Insurance Portability and Accountability Act, which means the data may be shared or sold to third parties; see: Steve Alder, [“Majority of Americans mistakenly believe health app data is covered by HIPAA,”](#) The HIPAA Journal, July 26, 2023.
21. Ina Fried, [“Generative AI’s privacy problem,”](#) Axios, March 14, 2024; Federal Trade Commission, [“AI companies: Uphold your privacy and confidentiality commitments,”](#) Jan. 9, 2024.
22. Ibid; Matt Burgess and Reece Rogers, [“How to stop your data from being used to train AI,”](#) Wired, April 10, 2024.
23. Arbanas, et al., Earning trust as gen AI takes hold: 2024 Connected Consumer Survey.
24. Baris Sarer, Ricky Franks, Cheryl Ho, and Jake McCarty, [“AI and the evolving consumer device ecosystem,”](#) The Wall Street Journal, April 24, 2014; Sam Reynolds, [“AI-enabled PCs will drive PC sales growth in 2024, say research firms,”](#) Computer World, Jan. 11, 2024; Clare Conley, [“Generative AI in 2024: The 6 most important consumer tech trends,”](#) Qualcomm, Dec. 14, 2023.
25. Arbanas, et al., Earning trust as gen AI takes hold: 2024 Connected Consumer Survey.
26. Monique Woodard, [“Unlocking the trillion-dollar female economy,”](#) TechCrunch, May 21, 2023.
27. Arbanas, et al., Earning trust as gen AI takes hold: 2024 Connected Consumer Survey.
28. Ibid.
29. Across industries, 51% of women workers using gen AI anticipate it would substantially boost their productivity at work a year from now, vs. 64% of men; see: Arbanas, et al., Earning trust as gen AI takes hold: 2024 Connected Consumer Survey.
30. Tech women and men are statistically tied: Forty-two percent of tech women who use or experiment with gen AI have “high” or “very high” trust that gen AI providers will keep their data secure, and another 40% report moderate trust, while 47% of tech men report “high” or “very high” trust and another 30% report moderate trust; see: Arbanas, et al., Earning trust as gen AI takes hold: 2024 Connected Consumer Survey.
31. Ibid.
32. Greater proportions of men in the tech industry who use gen AI report that their employers encourage its use (93%) and provide training (91%). While there’s still a gender gap in these views among workers in the tech industry, the gap is significantly smaller than among men and women working in other industries; see: Arbanas, et al., Earning trust as gen AI takes hold: 2024 Connected Consumer Survey.
33. World Economic Forum, [“Global gender gap report 2023.”](#)
34. Deloitte, [“Generative AI.”](#)
35. Genevieve Smith and Ishita Rustagi, [“When good algorithms go sexist: Why and how to advance AI gender equity,”](#) Stanford Social Innovation Review, March 31, 2021.
36. Charlotte Lytton, [“AI hiring tools may be filtering out the best job applicants,”](#) BBC, Feb. 16, 2024.
37. Carmen Niethammer, [“AI bias could put women’s lives at risk - A challenge for regulators,”](#) Forbes, March 2, 2020.
38. Ryan Browne and MacKenzie Sigalos, [“A.I. has a discrimination problem. In banking, the consequences can be severe,”](#) CNBC, June 23, 2023.
39. Fancher, Ammanath, Holdowsky, and Buckley, [“AI model bias can damage trust more than you may know. But it doesn’t have to.”](#)
40. World Economic Forum, [“Global gender gap report 2023.”](#)

41. Smith and Rustagi, “[When good algorithms go sexist.](#)”
42. Hyun-Kyoung Kim, “[The effects of artificial intelligence chatbots on women’s health: A systematic review and meta-analysis,](#)” Healthcare, Feb. 23, 2024; Sheryl Jacobson and Jen Radin, “[Can FemTech help bridge a gender-equity gap in health care?](#)” Deloitte, Oct. 5, 2023; Karen Taylor, “[Why investing in FemTech will guarantee a healthier future for all women,](#)” Deloitte UK, June 23, 2023.
43. Arbanas, et al., Earning trust as gen AI takes hold: 2024 Connected Consumer Survey.
44. Ibid.
45. Deloitte AI Institute, “[The generative AI dossier: A selection of high-impact use cases across six major industries,](#)” April 3, 2023.
46. Women in AI, “[WAI at work: Shaping the future of work for women in AI,](#)” 2022.
47. Ibid.

Acknowledgements

Authors would like to thank **Duncan Stewart, Paul Lee, Ben Stanton, Vipul Mehta, Roxana Corduneanu, Michael Steinhart, Michelle Dollinger, Je Stoudt, Catherine King, Elizabeth Fisher, Andy Bayiates, Prodyut Borah, Molly Piersol**, Deloitte Insights team.

Cover image by: **Jaime Austin**; Getty Images, Adobe Stock

隨著生成式 AI 對運算能力的需求不斷攀升，資料中心應尋求更可靠、潔淨的能源解決方案

科技業應優化基礎建設、重新思考晶片設計，並與電力供應商合作，以確保資料中心邁向更加永續的未來

在生成式 AI 推動下，資料中心耗能將持續上升，但實際上其耗能占比，並非全球電力需求的最主要部分。Deloitte 預測，截至 2025 年資料中心電力消耗僅占全球的 2%，相當於 536 太瓦時（TWh）。但隨著高耗電的生成式 AI 訓練與推理的成長速度超越其他使用及應用，至 2030 年，全球資料中心的用電量可能會翻倍至 1,065TWh（圖 1）。¹ 為降低資料中心電力需求上升對環境的影響，許多企業正積極使用創新和高效節能的技術，並尋求更多的零碳能源解決方案。

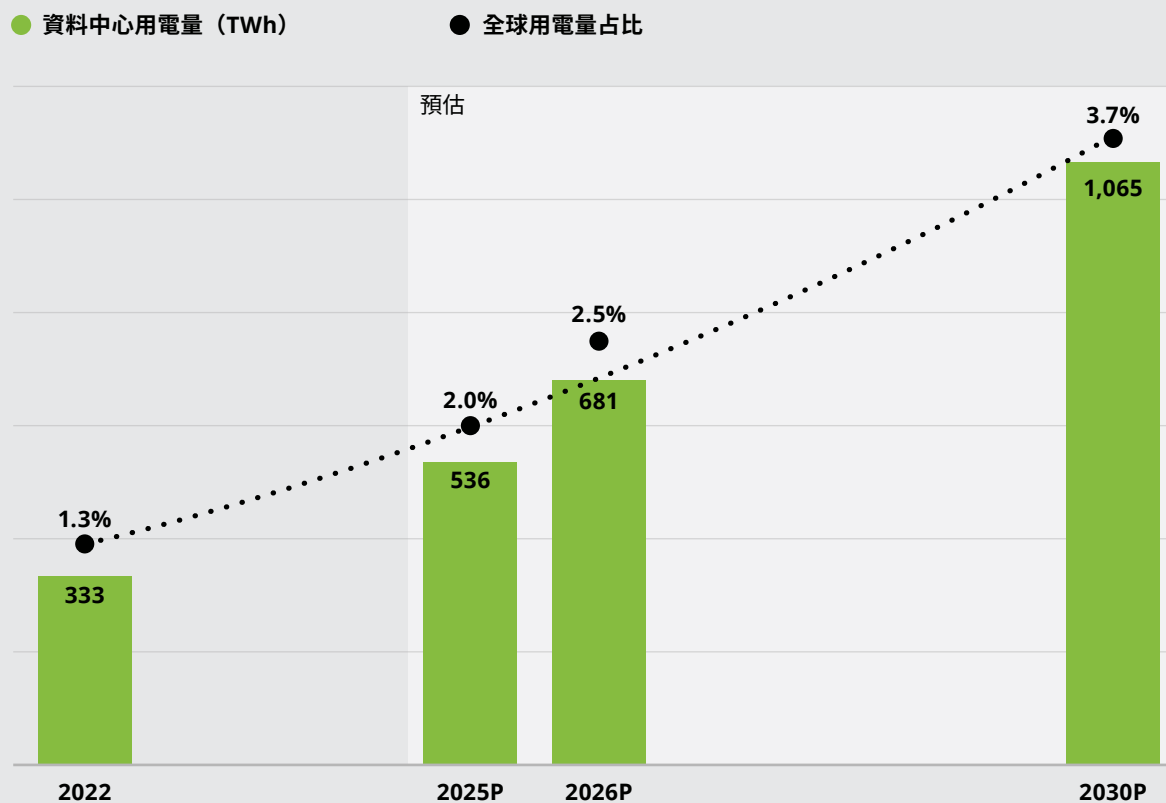
儘管如此，在發電和電網基礎建設方面，想要滿足 AI 資料中心快速增加的電力需求，仍是一項艱鉅挑戰。電氣化（即運輸業、建築業和工業從化石燃料切換至電力驅動設備及系統）與其他因素，已經導致電力需求快速攀升。而生成式 AI 的崛起進一步帶來額外且可能未被完全預見的新增電力需求。此外，資料中心通常具有特殊需求，不僅需要 24 小時穩定供電，還須具備高可用性和可靠性，同時正在努力實現零碳排目標。

預測 2030 年及未來全球資料中心用電量消耗將極具挑戰性，需考量多重變數。Deloitte 評估顯示，若 AI 技術與資料中心處理效率持續提升，至 2030 年的全球資料中心能耗可能將達到約 1,000TWh。

不過，若未來數年無法實現預期效能的提升，則資料中心需要的用電量可能會超過 1,300TWh，將會直接影響供電業者，並對氣候中和目標構成挑戰。² 因此，未來十年推動 AI 技術創新和優化資料中心效率，將成為實現永續能源的重要關鍵。

圖 1

預計到 2030 年，全球資料中心用電量快速增加，主要受高耗電的 AI 模型影響，尤其是生成式 AI



備註：P 代表預測值。

資料來源：Deloitte 分析，根據可取得的公開資料來源以及產業專家訪談。

方法論：我們使用美國能源資訊署 (U.S. Energy information Administration) 2023 年國際能源展望 (International Energy Outlook) 中，涵蓋住宅、商業、工業與運輸最終用途使用之的總電力的基礎用電量資料，推算出 2022 年至 2030 年之間，全球資料中心用電量 (TWh) 的估算與預測值。我們是根據針對多項可公開取得來源進行的深入研究，包括 Semi Analysis、美國電力研究院 (EPRI)、高盛、彭博、Latitude Media，估算和預測資料中心在全球總用電量中的占比，並藉由專訪科技、能源與永續發展領域的產業專家進一步驗證。

deloitte.com/insights

全球部分區域已因 AI 資料中心日漸成長的電力需求，面臨發電和電網容量管理問題。³ 預計 2023 年至 2026 年間，資料中心的關鍵元件包括圖形處理器（Graphics Processing Unit，GPU）、中央處理器（Central Processing Unit，CPU）伺服器、儲存系統、冷卻設備和網路交換器）所需電力將會倍增，且至 2026 年，全球用電量將達到 96 吉瓦（GW），其中 AI 運算可能單獨消耗超過 40% 的電力。⁴ 此外，預計至 2026 年，全球 AI 資料中心的年度耗電量將達到 90 太瓦（Twh），約佔全球所有資料中心預測 681TWh 的大約七分之一，相較 2022 年增加 10 倍左右。⁵ 生成式 AI 投資推動了龐大的電力需求，使 2024 年第一季全球 AI 資料中心的新增電力需求達 2GW，較 2023 年第四季增加了 25%，更是 2023 年第一季水準的 3 倍以上。⁶ 滿足資料中心電力需求極具挑戰，因其設施通常高度集中（尤其是在美國），且需要 24 小時不間斷供電，可能會對現有的電力基礎設施造成負擔。⁷

Deloitte 預測，科技和電力產業將攜手面對這些挑戰，並致力於控制 AI（特別是生成式 AI）帶來的能源衝擊。目前，許多大型科技公司和雲端服務供應商已積極投資零碳能源，推動淨零目標，⁸ 展現對永續發展的承諾。

超大規模企業（Hyperscalers）計畫大規模擴建生成式 AI 資料中心，以應對日漸成長的客戶需求

全球電力需求快速增加主要是因為超大規模企業（Hyperscalers）計畫在全球擴建資料中心容量。⁹ 由於預期 AI 需求將會成長，特別是生成式 AI，各企業與國家正在積極建設更多資料中心，以滿足市場需求。此外，各國政府也開始建立主權 AI 能力，以維持技術領先地位。¹⁰ 根據主要超大規模企業（Hyperscalers）的資本支出，資料中心的房地產建設規模已達到歷史新高，截至 2024 年大約為 2,000 億美元，預計 2025 年將會超過 2,200 億美元。¹¹

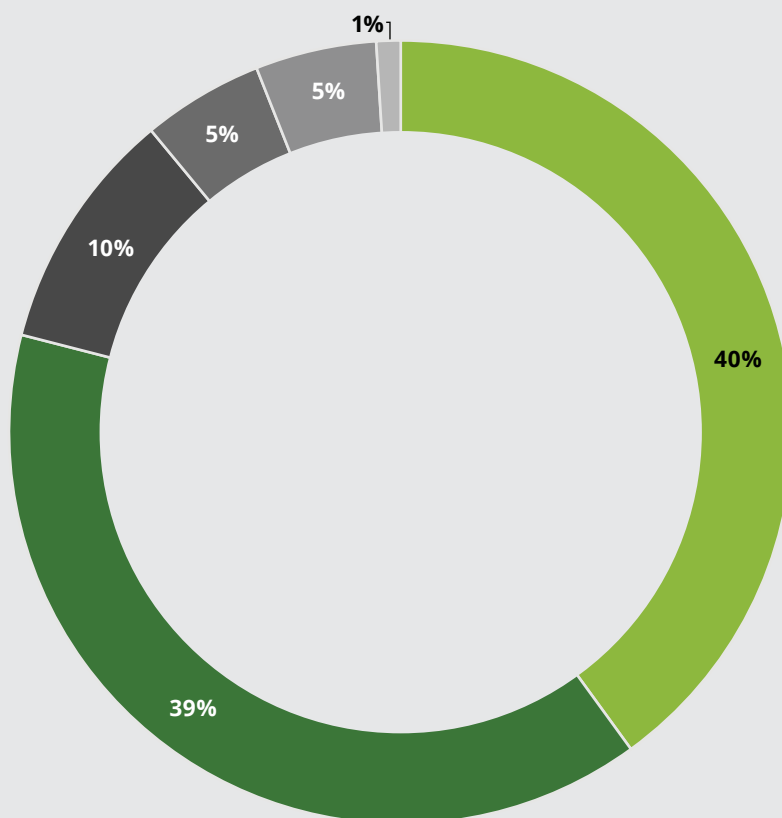
此外，Deloitte 在《企業使用生成式 AI 之現況》（State of Generative AI in the Enterprise）指出，截至目前為止，多數企業仍處於試行與實驗階段。¹² 但在透過生成式 AI 中創造價值的實驗過程，受訪者已見證到實質成果，並促使企業加速擴大應用規模，不再侷限於試行及概念驗證。若生成式 AI 使用量隨著技術成熟而增加，則 2025 年及 2026 年，超大規模企業和雲端服務供應商可能將持續投入高額資本支出。

資料中心的主要用電量來自於運算能力與伺服器資源，特別是伺服器系統（約佔資料中心功耗的 40%）以及冷卻系統（消耗 38% 至 40% 的功率）。即使在 AI 資料中心內，上述兩者也是最主要能源消耗的來源，且其耗電將持續增加。除此之外，內部電力調節系統使用 8% 至 10% 的電力，而網路與通訊設備以及儲存系統則各佔 5% 的電力，照明設施則消耗 1% 至 2% 的電力（圖 2）。¹³ 由於生成式 AI 需要大量的電力，因此資料中心業者（包括超大規模企業及資料中心營運商）在設計資料中心時，需要考慮替代的能量來源、新型冷卻形式，以及更節能的解決方案，目前相關措施已陸續展開。

圖 2

運算能力和冷卻系統是 AI 資料中心消耗能源的主要因素

● 運算能力與伺服器資源 ● 冷卻系統 ● 內部電力調節系統
● 網路設備 ● 儲存系統 ● 照明



資料來源：Deloitte 分析。

deloitte.com/insights

生成式 AI 推動電力需求快速成長

自 2023 年起，由於 AI 需求快速增加，資料中心用電量大幅上升。¹⁴ 部署先進 AI 系統需要大量晶片和運算能力，而訓練複雜的生成式 AI 模型可能需要使用上千個 GPU 同時運行。

支持生成式 AI 與高效運算環境的超大規模企業（Hyperscalers）和大規模資料中心業者，需建構高密度的基礎建設以支援強大的運算能力。過去資料中心主要是依賴中央處理器（CPU），每片晶片功耗約 150 瓦（W）至 200 瓦（W）。¹⁵ 2022 年 AI 運算使用的 GPU 功耗已達 400 瓦，2023 年生成式 AI 使用的最先進 GPU 增至 700W，預期 2024 年次世代晶片功耗將達到 1,200W。¹⁶ 這些晶片（約 8 片）是安裝在資料中心機架內的刀鋒上（每一機架包含 10 組刀鋒），相較傳統資料中心，能源消耗更多，每平方公尺產生的熱能也更多。¹⁷ 截至 2024 年初，一般資料中心可以支援 20kW 以上的機架功率需求，但預計平均功率密度將從 2023 年的每機架 36kW，增加到 2027 年的每機架 50kW。¹⁸

自生成式 AI 出現以來，全球 AI 運算能力（以每秒浮點運算次數 FLOPS 測量）呈指數級成長。自 2023 年第一季起，全球季成長率為 50% 到 60%，且可能維持此速度持續成長至 2025 年第一季。此外，¹⁹ 資料中心不僅關注 FLOPS，還以兆瓦時（MWh）與太瓦時（TWh）衡量其能源消耗。

生成式 AI 的大型語言模型和其消耗的數十億瓦特電力

生成式 AI 的大型語言模型（Large Language Model，LLM）日趨複雜，且結合的參數數量（AI 學習與預測所需的變數）持續成長。2021 年至 2022 年間推出的 1,000 到 2,000 億個參數模型，至 2024 年中，最先進大型語言模型參數規模已擴大到將近 2 兆，具備詮釋和解碼複雜圖像的能力。²⁰ 目前市場競爭推出了使用 10 兆個參數的大型語言模型發展。由於 AI 模型需要進行訓練和部署，參數規模的擴大將提升資料處理和運算能力的需求，進一步加速對生成式 AI 處理器和加速器的依賴，進而使耗電量增加。

此外，訓練大型語言模型極為耗費能源。獨立研究顯示，訓練一個超過 1,750 億個資料參數的大型語言模型時，單次訓練電力消耗介於 324MWh 與 1,287MWh 之間，且模型通常需進行多次重新訓練。²¹

平均來說，以生成式 AI 單次查詢的耗電量是傳統網際網路搜尋的 10 到 100 倍。²² Deloitte 預測，若全球每天僅有 5% 的網際網路搜尋使用生成式 AI，則需近 20,000 台伺服器（每台伺服器搭載 8 顆專用 GPU 核心），平均每台伺服器需要消耗 6.5kW，每天平均耗電量將達 3.12GWh，年耗電量約為 1.14Twh，²³ 相當於美國 108,450 戶家庭一年的用電需求。²⁴

資料中心的需求為電力產業轉型帶來挑戰與機會

電力產業正積極規劃，以因應不斷增加的需求。業界預測至 2050 年，部分國家的耗電量將可能至多增加 3 倍。²⁵ 但由於資料中心需求的急速成長，近期部分區域的電力需求成長速度已超過原先預期。許多國家的電力需求模型已納入了電氣化進程（從傳統化石燃料轉向電動設備），以及資料中心的耗能增加，並將其結合經濟成長的趨勢評估。但資料中心的需求快速增加，可能僅是未來更大規模電力需求浪潮的前兆，顯現當前電網基礎設施升級，與能源供應轉型的迫切性。²⁶

電力產業正處於數十年的產業轉型期，電力公司也正在積極建設與升級基礎設施，並以實現電網去碳化為目的，同時推動數位化系統與資產管理。許多地區的電力公司也正在強化基礎建設，以應對日益嚴峻的極端氣候事件，及持續增加的網路安全威脅。²⁷ 部分國家的電網正面臨需求的挑戰，尤其是低碳或零碳電力供應方面。在美國，預計至 2026 年，資料中心將占整體耗電量的 6%（約 260TWh）。²⁸ 英國的電力需求可能在 10 年內成長六倍，主要是受到 AI 發展帶動。²⁹ 中國大陸預測包含 AI 運算在內的資料中心，在 2026 年將佔全國總電力需求的 6%。³⁰ 然而，由於中國大陸的發電結構仍以燃煤為主（2021 年煤炭發電占全國能源使用的 61%，並貢獻了 79% 的二氧化碳排放量），資料中心的成長可能進一步加劇環境污染問題。³¹

面對資料中心電力持續增加的需求，部份國家已採取監管措施。例如，愛爾蘭現有的資料中心佔全國總耗電量的 20%，隨著 AI 驅動下，此比例可能將進一步提升，甚至影響家庭用電需求下降。³² 愛爾蘭曾一度暫停資料中心連接電網的工程，後來恢復該政策。³³ 阿姆斯特丹市同樣曾暫停了新的資料中心工程，以支持永續都市發展。³⁴ 新加坡則針對資料中心發布新的永續標準，要求營運業者逐步將機房的整體工作溫度上升至 26°C 或更高，以降低冷卻需求並減少電力消耗，但較高的工作溫度可能會縮短晶片的使用壽命。³⁵

資料中心需求的迫切性和地理集中性，以及對 24 小時全天候供應零碳能源的需求，使科技公司與電力供應商面臨更嚴峻的挑戰，尤其是在電氣化轉型與製造業產生的額外用電需求下。全球最大的資料中心市場位於美國維吉尼亞州北部，³⁶ 當地的電力公司 Dominion Energy 預估在未來 15 年，該地區的電力需求將成長約 85%，其中資料中心的用電需求將增至目前的四倍。³⁷ 許多科技企業尋求 24 小時供應的零碳電力，但短期內仍難以實現，電力公司正嘗試包含開發再生能源與儲能等技術。此外，部分電力供應商也已宣布興建燃氣發電廠，然而燃氣發電並非零碳能源³⁸，可能進一步影響公用事業公司、地方政府及國家層級的去碳化目標。³⁹

儘管 AI 可能消耗大量清潔能源，但同時具備加速清潔能源轉型的潛力。部分公用事業已開始運用 AI 技術，以提升電網運作的成本效益，包括更精準的天氣與負載預測能力、加強電網管理及再生資產效能、加速災後電力恢復以及優化野火風險評估等。⁴⁰

資料中心冷卻技術對水資源的高度依賴

相較於先前的產品，新一代 CPU 和 GPU 具有更高的熱密度。部分伺服器供應商為滿足高效能運算（High Performance Computing, HPC）及 AI 應用的快速增加的需求，正將更多高功耗晶片封裝至每個機架內。然而，隨著機架密度提升，水資源的需求相應增加，特別是在冷卻生成式 AI 晶片方面。預估到 2027 年，AI 資料中心的淡水需求可能將高達 1.7 兆加侖（上限估算）。⁴¹ 使用空氣冷卻系統，和蒸發冷卻技術來管理多餘熱能的超大規模資料中心，每年的用水量可能會超過 5,000 萬加侖，相當於製造約 14,700 支智慧型手機需要的水量。⁴² 然而，這些用水無法回收至原始含水層、水庫或供應源頭，導致水資源的永久消耗。⁴³

僅氣冷系統的用電量即占傳統資料中心的 40%，因此業界正在尋找替代傳統氣冷式散熱的方案。其中液體冷卻技術因具備更高的熱傳導效能，成為主要發展的方向。相較於空氣冷卻，液體冷卻技術可以有效降低高密度伺服器機架的用電量，多達 90%。⁴⁴ 由於液體冷卻技術能直接冷卻伺服器機架，特別適用於支援 50kW 至 100kW 或更高的功率的高密度機架。⁴⁵ 此外，液體冷卻技術有助於降低傳統冷卻水製造設備的依賴，提高能源使用效率。

然而，儘管液體冷卻技術在資料中心整體能源效率提升方面展現潛力，⁴⁶ 不過此項技術尚未成熟，而難以廣泛應用或整合至全球 AI 資料中心。⁴⁷ 此外，由於水資源有限，其成本和可獲得性，將可能影響未來液冷技術的應用決策。

科技業正邁向永續與零碳能源解決方案

為加速 AI 資料中心轉型至零碳能源，科技產業龍頭企業持續積極推動可再生能源發展，主要透過能源購買協議（PPA，Power Purchase Agreement）或與可再生能源供應商簽訂長期合約。⁴⁸ 這些交易是確保再生能源專案獲得資金支持，促使其發展。此外，一些科技公司正與電力供應商及技術創新業者合作，測試及推動新興能源技術的規模化應用，包括先進地熱技術、風能、太陽能、水力發電，甚至是水下資料中心等前瞻性解決方案。

然而，在部分地區，由於電網限制和再生能源及儲能設施的併網進程延遲，影響資源的有效接入。⁴⁹ 此延誤在美國可能長達 5 年，主要原因在於電力需求過高和輸電基礎設施不足。為解決此問題，部分科技公司開始探索自用發電，甚至是離網（Off-Grid）能源解決方案。⁵⁰ 此外，企業積極投入開發長時儲能技術與小型模組化反應爐（Small Modular Reactor，SMRs），以因應相關挑戰。部分科技公司和公用事業單位正計畫合作，擴大創新清潔能源技術的應用，這不僅能為企業帶來效益，也有助於加速電網去碳化進程。⁵¹

大多數清潔能源的研究、發展、試驗計畫及投資，仍需數年才能獲得成果，實現投資回報並達到商業規模化。⁵² 舉例來說，小型模組化核能反應爐仍在早期開發階段，短期內尚無法做為即時的零碳解決方案。⁵³

在美國，企業採購再生能源將持續由科技業占主導地位。截至 2024 年 2 月 28 日，過去 12 個月內美國企業簽署的將近 200 筆可再生能源購電合約中，科技業佔比超過 68%。⁵⁴ 同樣印度有越來越多超大規模企業和資料中心業者，日益依賴太陽能為當地資料中心提供電力。⁵⁵ 若沒有這些採購承諾，許多再生能源專案將難以推動。⁵⁶

因此，科技業在為清潔能源技術提供資金支持，以及推動規模化應用方面，將持續發揮關鍵影響力。部份業者直接與創新業者和再生能源供應商合作，而其他業者則是與公用事業建立夥伴關係。⁵⁷ 值得關注的是，科技業對清潔能源轉型的投資方式至關重要，因為無論是能源技術創新業者或傳統電力產業，通常皆無法與科技業的財務資源匹敵，以推動此類重要變革。

結論

面對生成式 AI 需求快速成長，科技業、超大規模企業、資料中心業者、公用事業和監管機關應採取哪些措施，以確保趨勢的永續發展，從超大規模企業和科技業的角度來看，相關考量與 Deloitte 於 2021 年針對全球雲端遷移的預測基本一致。⁵⁸ 儘管需求驅動因素有所變化，且市場變化速度加快，業界正致力於往永續發展和市場推進速度之間取得平衡，同時控制資料中心持續增加的能源需求，並尋找更符合永續發展的方式為 AI 提供動力，尤其是生成式 AI。

- 1. 提升生成式 AI 晶片的能源效率：**新一代 AI 晶片能源效率已大幅提升，例如目前的 AI 晶片可以在 90 天內完成 AI 訓練，耗能為 8.6GWh。相較於前代晶片，在相同資料上執行相同任務的耗能，減少了超過 90%。⁵⁹ 晶片公司應持續與更廣泛的半導體生態系統合作，進一步優化每瓦浮點運算（FLOPS/Watt）效能，以突破 AI 訓練的效能極限，使未來的 AI 晶片能夠訓練規模數倍於目前最大 AI 系統的模式，同時降低電力消耗。

2. **優化生成式 AI 的使用，並轉移部分運算至邊緣設備：**包括評估 AI 訓練和推理是否應在資料中心進行，以提升能源效率；或是否應轉移至邊緣設備進行運算，並據此調整資料中心設備的需求。邊緣運算不僅能滿足低延遲應用的需求，亦適用於處理機敏資料或具高度隱私需求的情況。此外，透過將生成式 AI 工作負載重新分配至當地、近端或共享設備，也有效節省網路和伺服器頻寬消耗，僅將特定的 AI 工作負載傳送至資料中心。⁶⁰
3. **優化生成式 AI 演算法，精準調整 AI 的工作負載規模：**業界應思考是否需要持續開發超大規模基礎模型（如超過一兆個參數的模型），或小模型已經足夠且更符合永續發展？目前已有新創企業開發多模態 AI 模型，可直接在設備上運行，無需依賴雲端高耗能運算。⁶¹ 企業應根據實際業務需求，精準和調整 AI 工作負載規模，選擇最適合的生成式 AI 模型，並僅在必要時進行訓練，以將能源消耗降至最低。此外，根據 AI 推論的具體需求，如即時推理和低延遲應用，CPU 相較於 GPU 可能在部分情況下更具效能優勢。⁶²
4. **建立策略合作夥伴關係，滿足當地及群聚型 AI 資料中心需求：**對於部分較難利用生成式 AI 資料中心資源的中小型單位（包括大學），獲得 AI 運算能力可能較具挑戰性。因此，這些單位應與專業資料中心維運商及雲端服務供應商合作，利用高效能運算（HPC）專用的 GPU 叢集共享模式，實現更具彈性的聯合資料中心解決方案。⁶³ 同時，資料中心可即時監測資源使用情況，追蹤潛在的市場機會與需求，進而提供短期共享運算服務。
5. **與其他利害關係人及領域合作，推動整體環境正向影響：**生成式 AI 生態系統的主要參與者，包括超大規模企業（Hyperscalers）、企業客戶、第三方資料中心業者、公共事業供應商、監管機構、市政單位及房地產公司等，應持續針對業務、環境永續和社會需求展開對話，推動解決方案。⁶⁴ 此類合作涵蓋多個層面，包括策略性共享資料中心規劃需求（由資料中心維運商提供彈性運算與伺服器資源）、冷卻技術優化（確保液冷系統運行於最佳溫度）、廢熱與廢水管理（開發熱能回收與水資源再利用），以及回收計畫制定（優化資源循環利用）。例如，歐洲部分資料中心業者已將餘熱導流至周邊的游泳池供暖，以降低能源浪費。⁶⁵ 此外，公共事業供應商可與科技業建立更緊密合作，以滿足資料中心日益成長的能源需求，同時藉由科技公司資金和技術能力，以擴大整體應用規模，這將是推動電網向清潔能源轉型的關鍵要點。超大規模企業（Hyperscalers）和公用事業供應商正積極採取措施，以提升無碳能源在資料中心電力供應中的佔比，尤其是專為生成式 AI 所打造的相關措施。這些努力將在長期內帶來明顯環境效益，加速全球資料中心永續發展轉型目標。

By **Karthik Ramachandran**
India

Kate Hardin
United States

Duncan Stewart
Canada

Gillian Crossan
United States

Endnotes

1. Deloitte analysis based on publicly available information sources and conversations with industry specialists. We used base electricity consumption data from the US Energy Information Administration's (EIA) International Energy Outlook 2023 data on total electricity usage across residential, commercial, industrial, and transportation end uses (a reference to US Energy Information Administration, "[Table: Delivered energy consumption by end-use sector and fuel](#)," accessed Nov. 4, 2024) to arrive at estimates and prediction values for global data centers' electricity consumption (TWh) between 2022 and 2030. Our estimates and projections for data centers' percent electricity consumption of global total are based on our research of multiple publicly available sources including SemiAnalysis, EPRI, Goldman Sachs, Bloomberg, and Latitude Media, and further validated based on our conversations with subject matter specialists in the areas of technology, energy, and sustainability. Total energy consumption by end-use sector and fuel (as noted from the aforementioned table from EIA's International Energy Outlook 2023 data), globally, is estimated and forecast at 26,787 TWh in 2025, 27,256 TWh in 2026, and 29,160 TWh in 2030—increasing from 25,585 TWh back in 2022.
2. As noted in endnote 1 above, we arrived at 2022 to 2030 data, estimates, and predictions based on a combination of in-depth secondary research of multiple publicly available sources, and validated further from our discussions with subject matter specialists. Also, see Prof. Dr. Bernhard Lorentz, Dr. Johannes Trüby, and Geoff Tuff, "[Powering artificial intelligence](#)," Deloitte Global, November 2024."
3. One-fifth of Ireland's electricity is consumed by data centers, and this is expected to grow, even as households are lowering their electricity use. To read further, see: Chris Baraniuk, "[Electricity grids creak as AI demands soar](#)," BBC, May 21, 2024.
4. Dylan Patel, Daniel Nishball, and Jeremie Eliahou Ontiveros, "[AI data center energy dilemma: Race for AI data center space](#)," SemiAnalysis, March 13, 2024.
5. Ibid.
6. Data center BMO report, Communications Infrastructure, "1Q24 data center leasing: Records are made to be broken," April 28, 2024; Moreover, due to strong demand from cloud providers and AI workloads, the data center primary market supply in the United States alone was up 26% year over year to 5.2 GW in 2023, and more are under construction. See further: CBRE, "[North America data center trends H2 2023](#)," March 6, 2024.
7. Lisa Martine Jenkins and Phoebe Skok, "[Mapping the data center power demand problem, in three charts](#)," Latitude Media, May 31, 2024.
8. Based on our analysis of multiple publicly available information and reports from what companies self-report, and further validated from third-party sources.
9. For context, hyperscalers are large cloud service providers and data centers that offer huge amounts of computing and storage resources typically at enterprise scale. See: Synergy Research Group, "[Hyperscale operators and colocation continue to drive huge changes in data center capacity trends](#)," Aug. 7, 2024.
10. Yifan Yu, "[AI's looming climate cost: Energy demand surges amid data center race](#)," Nikkei Asia, June 12, 2024.
11. Data center BMO report, Communications Infrastructure, "1Q24 data center leasing: Records are made to be broken," April 28, 2024. Further, Deloitte analysis based on information from select tech companies' publicly available sources such as earnings releases and Dell'Oro Group's market research data on data center IT capital expenditure shows that if we consider the capital expenditure spending of other data center providers, including third-party operators and outsourced cloud service providers, data centers' aggregate capital expenditure spending could be at least US\$250 billion in 2025. See: Baron Fung, "[Market research on data center IT capex](#)," Dell'oro Group, accessed Nov. 4, 2024.
12. Nitin Mittal, Costi Perricos, Brenna Sniderman, Kate Schmidt, and David Jarvis, "[Now decides next: Getting real about generative AI](#)," Deloitte's State of Generative AI in the Enterprise quarter two report, Deloitte, April 2024.
13. Deloitte analysis based on publicly available research reports including: Wania Khan, Davide De Chiara, Ah-Lian Kor, and Marta Chinnici, "[Advanced data analytics modeling for evidence-based data center energy management](#)," Physica A 624, 2023; Kazi Main Uddin Ahmed, Math H. J. Bollen, and Manuel Alvarez, "[A review of data centers energy consumption and reliability modeling](#)," in IEEE Access 9, 2021: pp. 152536–152563.

14. Tom Dotan and Asa Fitch, [“Why the AI industry’s thirst for new data centers can’t be satisfied,”](#) The Wall Street Journal, April 24, 2024.
15. Noam Broussard, [“Examining the impact of chip power reduction on data center economics,”](#) Semiconductor Engineering, March 12, 2024.
16. Based on our analysis of multiple publicly available sources including: Michael Studer, [“The energy challenge of powering AI chips,”](#) Robeco, Nov. 6, 2023; Agam Shah, [“Generative AI to account for 1.5% of world’s power consumption by 2029,”](#) HPCwire, July 8, 2024.
17. From our study and analysis of select gen AI data center chip solutions offered by major AI chip vendors, further corroborated with publicly available third-party sources including: Beth Kindig, [“AI power consumption: Rapidly becoming mission-critical,”](#) Forbes, June 20, 2024.
18. Jones Lang LaSalle, [“Data centers 2024 global outlook,”](#) Jan. 31, 2024; Doug Eadline, [“The gen AI data center squeeze is here,”](#) HPCwire, Feb. 1, 2024; Per IDC, besides graphics processing unit, servers, data centers also need to grapple with a corresponding growth in storage capacity, which is likely to double between 2023 and 2027 to reach 21 zettabytes in 2027. See: John Rydning, [“Worldwide Global StorageSphere forecast, 2023 to 2027: Despite decreased petabyte demand near term, the installed base of storage capacity continues to grow long term,”](#) IDC Corporate, May 2023.
19. Patel, Nishball, and Eliahou Ontiveros, [“AI data center energy dilemma.”](#)
20. Sean Michael Kerner, [“What are large language models?”](#) TechTarget, May 2024; Yu, [“AI’s looming climate cost.”](#)
21. Alex de Vries, [“The growing energy footprint of artificial intelligence,”](#) Joule 7, no. 10 (2023): pp. 2191–2194.
22. Eren Çam, Zoe Hungerford, Niklas Schoch, Francys Pinto Miranda, and Carlos David Yáñez de León, [“Electricity 2024: Analysis and forecast to 2026 report,”](#) International Energy Agency, accessed Nov. 4, 2024.
23. Deloitte analysis based on publicly available reports and sources including: de Vries [“The growing energy footprint of artificial intelligence,”](#) pp. 2191–2194.
24. Deloitte analysis based on data related to energy use and electricity consumption in homes in the United States. See: US Energy Information Administration, [“Use of energy explained,”](#) accessed Dec. 18, 2023.
25. Darren Sweeney, [“Utility execs prepare for ‘tripling’ of electricity demand by 2050,”](#) S&P Global, April 19, 2023.
26. Robert Walton, [“US electricity load growth forecast jumps 81% led by data centers,”](#) Utility Dive, Dec. 13, 2023.
27. Aaron Larson, [“How utilities are planning for extreme weather events and mitigating risks,”](#) POWER, March 13, 2024.
28. Çam, Hungerford, Schoch, Miranda, and de León, [“Electricity 2024.”](#)
29. Baraniuk, [“Electricity grids creak as AI demands soar.”](#)
30. Yu, [“AI’s looming climate cost.”](#)
31. Data on China’s energy use and CO2 emissions sourced from International Energy Agency, accessed September 25, 2024. See: International Energy Agency, [“China’s energy use,”](#) accessed Nov. 4, 2024; International Energy Agency, [“China’s CO2 emissions,”](#) accessed Nov. 4, 2024.
32. Baraniuk, [“Electricity grids creak as AI demands soar.”](#)
33. Paul O’Donoghue, [“Build it and they will hum: What next for Ireland and data centers?”](#) The Journal, Sept. 2, 2024.
34. Hosting Journalist, [“City of Amsterdam puts halt to new data center construction,”](#) Dec. 21, 2023.
35. With every 1C increase, operators could save 2% to 5% on the energy they use for cooling equipment. To read further, see: Inno Flores, [“Singapore unveils green data center road map amid AI boom that strains energy resources,”](#) Tech Times, May 30, 2024.
36. Julie R. Peasley, [“Ranked: Top 50 data center markets by power consumption,”](#) Visual Capitalist, Jan. 10, 2024.

37. Whitney Pipkin, [“Energy demands for Northern Virginia data centers almost too big to compute,”](#) Bay Journal, June 18, 2024.
38. Zach Bright, [“Southeast utilities have a ‘very big ask’: More gas,”](#) E&E News, Jan. 22, 2024.
39. Ibid.
40. Robert Walton, [“AI is enhancing electric grids, but surging energy use and security risks are key concerns,”](#) Utility Dive, Oct. 23, 2023.
41. Karen Hao, [“AI is taking water from the desert,”](#) The Atlantic, March 1, 2024.
42. Deloitte analysis based on publicly available information sources including: Jennifer Billock, [“Photos: How much water it takes to create 30 common items,”](#) North Shore News, Jan. 19, 2023.
43. Hao, [“AI is taking water from the desert”](#); One case in point is China—where its data centers’ annual water consumption is expected to increase from around 1.3 billion cubic meter as of 2023 to over 3 billion cubic meter by 2030. To read further, see: Yu, [“AI’s looming climate cost.”](#)
44. Eadline, [“The gen AI data center squeeze is here.”](#)
45. Diana Goovaerts, [“Data center operators want to run chips at higher temps. Here’s why.”](#) Fierce Network, June 11, 2024.
46. Scott Wilson, [“Is immersion cooling the answer to sustainable data centers?”](#) Ramboll, Dec. 13, 2023.
47. David Eisenband, [“100+ kW per rack in data centers: The evolution and revolution of power density,”](#) Ramboll, March 13, 2024; Direct-to-chip cooling (also known as cold plate liquid cooling or direct liquid cooling) cools down servers by distributing heat directly to server components, while, immersion cooling involves submerging servers and components in a liquid dielectric coolant that also helps prevent electric discharge.
48. Based on Deloitte’s analysis of developments and announcements from select major cloud hyperscalers and tech companies— and information gathered from publicly available sources (time period: 2023 and first three quarters of 2024).
49. Joseph Rand, Nick Manderlink, Will Gorman, Ryan Wiser, Joachim Seel, Julie Mulvaney Kemp, Seongeun Jeong, and Fritz Kahrl, [“Queued up: 2024 edition,”](#) Lawrence Berkeley National Laboratory, April 2024.
50. Based on Deloitte’s analysis of developments and announcements from select major cloud hyperscalers and tech companies— and information gathered from publicly available information sources between the first quarter of 2023 and the third quarter of 2024.
51. Julian Spector, [“Duke Energy wants to help Big Tech buy the 24/7 clean energy it needs,”](#) Canary Media, June 11, 2024.
52. For example, it’s not easy to submerge and drop a 1,300-ton data center unit underwater, especially since it demands special equipment to withstand pressure and corrosion caused by seawater. Moreover, there are concerns related to its impact on marine life.
53. David Schlissel and Dennis Wamsted, [“Small modular reactors: Still too expensive, too slow, and too risky,”](#) Institute for Energy Economics and Financial Analysis, May 2024.
54. Deloitte’s analysis of data and information gathered from multiple reports from S&P Global Market Intelligence, published during March and August 2024.
55. Manish Kumar, [“India’s data center boom opens up a fresh segment for green developers,”](#) Saur Energy International, July 1, 2024.
56. Naureen S. Malik and Bloomberg, [“With AI forcing data centers to consume more energy, software that hunts for clean electricity across the globe gains currency,”](#) Fortune, Feb. 25, 2024.
57. Based on Deloitte’s analysis of developments and announcements from select major cloud hyperscalers, tech companies, and power and utility players—on publicly available information sources during 2023 and the first three quarters of 2024.
58. Duncan Stewart, Nobuo Okubo, Patrick Jehu, and Michael Liu, [“The cloud migration forecast: Cloudy with a chance of clouds,”](#) Deloitte Insights, Dec. 7, 2020.

59. Wylie Wong, "[Nvidia launched next-generation Blackwell GPUs amid AI 'arms race'](#)," Data Center Knowledge, March 19, 2024; For instance, Nvidia notes that it can train a very large AI model using 2,000 Grace Blackwell chips in 90 days, consuming 4 MW power. In comparison, it would take as much as 8,000 of the previous generation chips to do the same work within the same time, consuming 15 MW power.
60. To read further, see section "Generative AI comes to the enterprise edge: 'On prem AI' is alive and well" in our 2025 TMT Predictions chapter on "[Updates](#)"; Additionally, see: Sabuzima Nayak, Ripon Patgiri, Lilapati Waikhom, and Arif Ahmed, "[A review on edge analytics: Issues, challenges, opportunities, promises, future directions, and applications](#)," Digital Communications and Networks 10, no. 3 (2024): pp. 783–804.
61. Yu, "[AI's looming climate cost](#)."
62. Luke Cavanagh, "[GPUs vs. CPUs in the context of AI and web hosting platforms](#)," Liquid Web, Aug. 20, 2024.
63. Eadline, "[The gen AI data center squeeze is here](#)."
64. Goovaerts, "[Data center operators want to run chips at higher temps. Here's why](#)."
65. Baraniuk, "[Electricity grids creak as AI demands soar](#)."

Acknowledgements

The authors would like to thank **Dilip Krishna, Marlene Motyka, Jim Thomson, Adrienne Himmelberger, Thomas Schlaak, Freedom-Kai Phillips, Johannes Truby, Clement Cabot, Negina Rood, Ankit Dhameja, Suzanna Sanborn**, and **Akash Chatterji** for their contributions to this article.

Cover image by: **Jaime Austin**; Getty Images, Adobe Stock

自主生成式代理型 AI（Agentic AI）： 正在研發中

自主的新一代代理型 AI（Agentic AI）有望增加知識型工作者的生產力，以及提升各種工作流程的效率。但「自主性」的全面應用可能仍需要時間才能廣泛使用。

自主的代理型 AI（Autonomous generative AI agents），簡稱代理型 AI（Agentic AI），指的是可以在幾乎沒有或無需人為監督的情況下，完成複雜工作及達成目標的軟體解決方案。代理型 AI 與目前的聊天機器人和輔助型 AI（常被稱為「代理」）不同。代理型 AI 具有提升知識工作者之生產力，以及自動化企業各業務領域中的多步驟流程。Deloitte 預測在 2025 年將有 25% 應用生成式 AI 的企業將進入代理型 AI 試點計畫或概念驗證階段，而此數字在 2027 年將成長至 50%。¹ 部分產業在某些使用情況下，可能會在 2025 年下半年將一些代理型 AI 應用實際融入現有的工作流程中。

開發代理型 AI 的新創企業和知名科技公司皆認為此技術具備刺激營收成長的潛力，並獲得資本市場的關注。過去 2 年，投資人已向代理型 AI 新創企業中注入超過 20 億資金，主要投資針對企業市場的公司。² 同時，許多科技公司、雲端服務供應商等業者也正在開發自身的代理型 AI 產品。同時進行策略性收購、技術授權，或直接招聘新創企業的人才來加速佈局，而非完全收購這些公司。³

代理型 AI 賦予「代理」真正的自主性

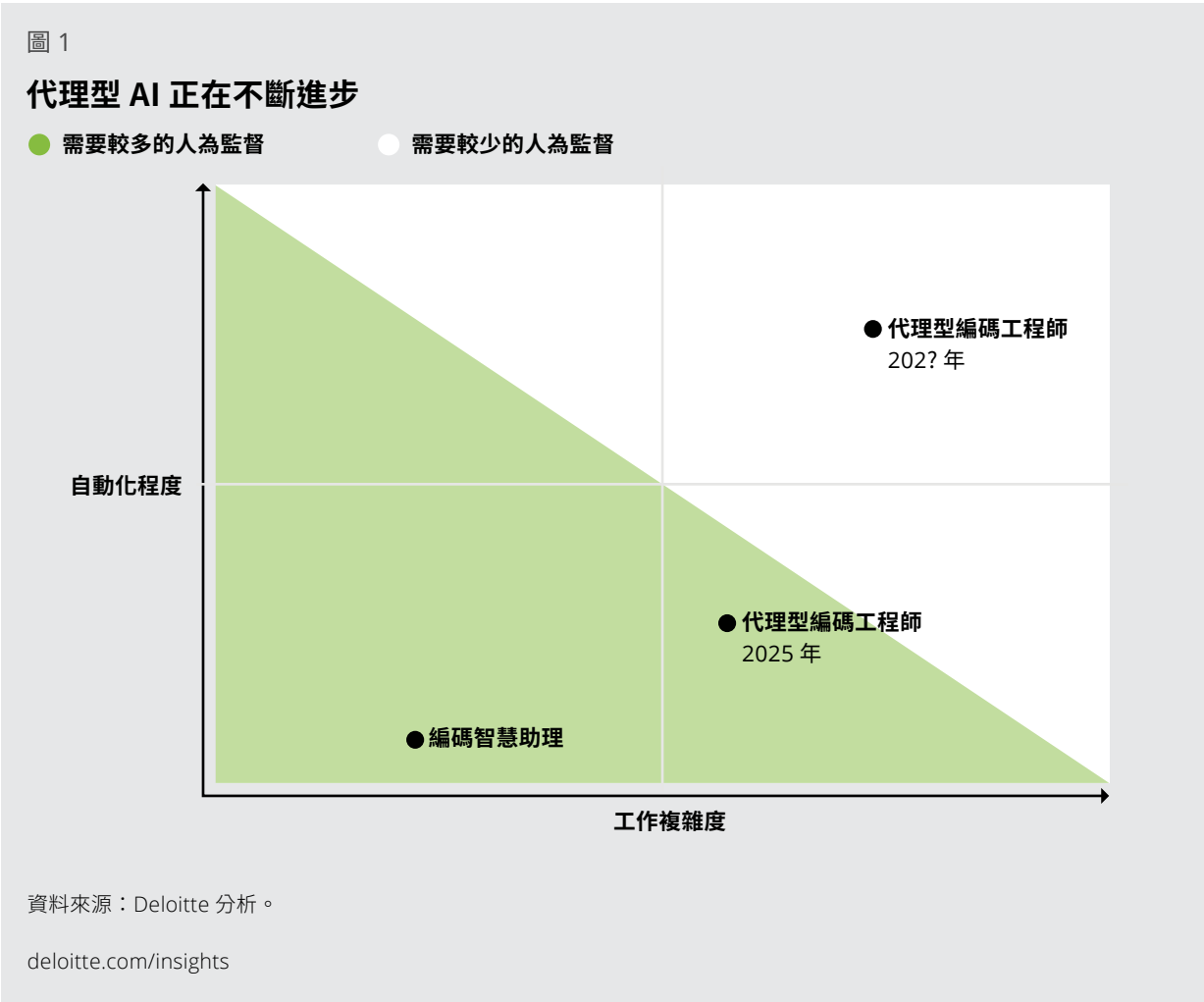
生成式 AI 聊天機器人和協作助手（Co-Pilots）雖然已具備高度智慧，可以與人類直觀的進行互動、合成複雜的資訊及生成內容。但是兩者皆缺乏代理型 AI 可以實現的行動能力和自主性。雖然聊天機器人和代理都是以大型語言模型（Large Language Model，LLM）為基礎，但是額外的技術和方法讓代理可以獨立運作、將任務拆解為多個可執行步驟，並在最少人為監督或干涉的情況下完成工作。相較於單純的互動功能，AI 代具備更強的推論和執行能力，能夠代表使用者行動。

顧名思義，代理型 AI 具有「能動性」，即自主行動和選擇採取適當行動的能力。⁴ 「代理性」意指自自主性，即獨立執行任務和決策的能力。⁵ 當我們將這些概念延伸至代理型 AI 時，它將能自主規劃、執行和實現目標，且其真正具備「代理性」⁶ 儘管最終目標仍由人類設定，但代理型 AI 將自主決定如何實現這些目標。

透過具體範例可以清楚區分代理型 AI 和 Co-Pilots 及聊天機器人之間的差異。以目前最成功的生成式 AI 應用之一⁷，輔助軟體開發的 Co-Pilots 為例，它們能夠幫助資深工程師提升生產力，它們可以提高經驗豐富之軟體工程師的生產力，以及提升初階程式設計師的效率。這類 AI 可以將自然語言指令（涵蓋多種語言）轉換為程式碼建議，並測試代碼的一致性，然而，但是此類 Co-Pilots 僅會根據工程師的指令做出回應，並不具備自主行動能力。相比之下，使用代理型 AI，可以讓軟體工程師具備更高的自主性。開發人員只需輸入軟體開發構想，然後由代理型 AI 軟體工程師即可將這些概念轉換為可執行的代碼，此流程會將軟體開發流程中的數個步驟自動化。

舉例來說，Cognition Software 在 2024 年 3 月推出「Devin」，目標是創造一個有能力推理、規劃和完成需要上千個決定之複雜工程任務的自主軟體工程師。⁸ Devin 能夠根據人類程式設計師的自然語言指令，獨立執行各類程式開發工作，包括設計完整應用程式、測試和修復代碼庫，以及訓練與調整大型語言模型（LLM）。⁹ 其競爭對手如 Codeium（專注於企業級軟體開發）和 Devin 的開源版本，也於 2024 年夏季問世。¹⁰

然而，當前的代理型 AI 軟體工程師仍存在一定的技術挑戰。¹¹ 特別在錯誤率方面，無法在沒有人為監督的情況下處理工作。在最近的一項基準測試中，以 Devin 為例，Devin 成功解決將近 14% 的 GitHub 開發問題，表現優於大型語言模型的聊天機器人（約為其兩倍），¹² 但是仍未達到完全自主的程度。大型科技公司¹³ 和新創企業正在努力提升代理型 AI 軟體工程師的自主性及可靠性，期望未來能讓人類開發人員與其雇主能放心地委託 AI 處理部分工作（圖 1）。



提升與放大勞動生產力

代理型 AI 軟體工程師僅是代理型 AI 如何改變工作方式的案例之一（請參閱下方文章主題「代理型 AI 的潛在應用案例」）。隨著代理型 AI 的進步，其影響力將會極為深遠。目前，美國有超過 1 億名知識型工作者，全球則超過 12.5 億名知識工作者。¹⁴ 全要素生產率（Total Factor Productivity, TFP）¹⁵ 作為衡量知識型工作的指標，在美國一直呈現停滯狀態，從 1987 年到 2023 年的成長率僅為 0.8%，而從 2019 年到 2023 年成長幅度更降至 0.5%，¹⁶ 大部分的 OECD 國家皆具有相同的情況。¹⁷ 過去企業嘗試透過將工作自動化以提升知識型工作生產力，但成效有限。同時市場對知識型人才需求持續增加，從客服代表到半導體工程師，幾乎所有專業領域均面臨人才短缺的問題。此外，當新員工加入時，即需要短時間內達到高效產出，對企業而言亦是一大挑戰。

傳統的專家系統和機器人流程自動化（Robotic Process Automation, RPA）在處理多步驟或流程不明確的工作時，往往無法提供足夠的靈活性。而基於傳統機器學習（ML）的系統則需要大量專門的訓練來適應特定應用場景。相比之下，代理型 AI 建構於大型語言模型（LLM）之上，具備更高的適應性，能夠涵蓋比傳統機器學習或深度學習更廣泛的應用範疇，為知識型工作的生產力提升帶來全新的可能性。

當流程不明確或需要數個步驟時，可能會降低專家系統和機器人流程自動化（Robotic Process Automation, RPA）的效果。而基於傳統機器學習（ML）的系統，則需要大量專門的訓練來適應特定應用場景。相比之下，代理型 AI 建構於大型語言模型（LLM）之上，具有更大的靈活性，可以處理比傳統機器學習或深度學習更廣泛的應用範疇。

代理型 AI 可以大幅提升大型語言模型（LLM）的能力，且能證明企業在生成式 AI 方面的技術投資價值。自生成式 AI 工具在公開發表後迅速吸引了企業高層的注意。企業能夠輕易想像該技術如何應用於業務營運。但是，通常很難估算生成式 AI 帶來的可量化商業價值。資料基礎建設的挑戰、風險和治理政策以及人才短缺等挑戰，讓企業難以大規模推動生成式 AI 項目。¹⁸ 據統計，僅有 30% 的生成式 AI 試點計畫可以進入全面生產階段。¹⁹ 此外，企業領導者對生成式 AI 產出的信任度仍然不足，且 AI 可能在實際應用中產生錯誤並導致現實世界的負面影響，都使高階主管必須三思而後行。²⁰

對於開發與執行生成式 AI 的公司必須考量生成式 AI 帶來的挑戰，以及建置具備推理、執行、協作與創造能力的代理型 AI 的技術複雜度。最重要的是，企業需要確保代理型 AI 的可靠性，企業才會有意願使用，僅僅大多數時間能完成任務是不夠的。然而，2024 年下半年已有部分應用展現出較高的可靠性，預計可在 2025 年初進入商業使用階段。

然而，其潛在回報值得投入努力，且初步結果顯示出良好前景。企業正在學習如何透過結合大型語言模型（LLM）與其他 AI 技術及訓練方法來提升模型效能。儘管完全自主且可靠的代理型 AI 仍是目標，但透過逐步提升 AI 精確性與自主性，皆有助於企業達生成式 AI 的整體初期生產力與效率目標。²¹ 由於代理型 AI 擁有其廣泛的應用範圍（涵蓋橫向與縱向領域）和明確的商業價值，使代理型 AI 更符合企業高層對生成式 AI 解決方案的初步期待。

深入剖析代理型 AI

代理型 AI 可以將複雜的工作拆解成一系列步驟逐步進行，並克服突發障礙。根據不同的應用場景，這些代理型 AI 可以感測環境，該環境可以是虛擬或實體或兩者的結合。為完成任務，代理型 AI 可以決定採取哪些行動，以及從工具、資料庫和其他 AI 代理取得協助，以完成任務並根據人類設定的目標交付成果。

代理型 AI 是一項新興技術，且在不斷進化，但是也具備以下關鍵的特徵和能力：

- **以基礎模型為基礎：**基礎模型（如大型語言模型）賦予代理型 AI 可以推理、分析、適應複雜和無法預測工作流程的能力。使它比機器人流程自動化和專家系統更具靈活性。當前大型語言模型正在快速進步，尤其是推理能力以及將任務拆解方面的能力。²² 但是基礎模組本身無法與環境互動、做出決策或執行任務。²³ 其必須透過其他技術和功能提升能力。
- **自主執行：**儘管自主程度不同，但代理型 AI 可以被訓練來獨立規劃和執行複雜的任務。在思維鏈模型（Chain-of-Thought）中納入推理標記（Reasoning Tokens），將能解決比先前大型語言模型更複雜的挑戰。思維鏈模型的反應較慢，但其在推理過程更具深思熟慮的特性，能夠自行糾正錯誤，以及顯示出為達成解決方案採取的步驟。²⁴
- **環境感知：**代理型 AI 可以感知環境、處理資訊及瞭解所執行任務的工作背景。²⁵ 先進代理型 AI 還可以處理多模態資料，例如影片、影像、音訊、文字和數字等多種類型的資訊來源。
- **工具調用：**為完成任務，代理型 AI 會與工具及系統互動，例如軟體、企業應用程式和網際網路。
- **協作與調度：**代理型 AI 可以管理其他系統和自動化代理的參與，以完成指定任務。在多代理系統（Multi-agent Systems）中，代理型 AI 還能與其他代理型 AI 進行協作。
- **記憶與學習：**由於大型語言模型屬於無狀態，每個互動都是獨立處理，不會保留先前對話的資訊。然而，代理型 AI 可以透過增加檢索機制和資料庫存取短期記憶，以便在執行特定工作時保留情境，並透過長期記憶從經驗中學習與改進，提升未來的執行表現。²⁶

最新模型中，有些已納入思維鏈功能，這些雖然反應速度比先前的大規模模型慢，且推理過程更謹慎，但是可以在複雜的問題上達到高階思維。²⁷ 多模態資料分析能夠進一步提升代理型 AI 的靈活性，因為它擴展了 AI 能夠解析與生成的數據類型。多模態 AI 也顯示出，當代理型 AI 功能與其他 AI 技術（如電腦視覺技術的影像辨識、語音轉錄和翻譯）結合時，效能將會變得更強大。²⁸ 然而，多模態 AI 與代理本身一樣仍處於發展階段中。

目前，真正的多代理系統正在開發中，這類系統中的工作可在自主代理網路協調執行，而某些實驗計畫已在 2024 年底開始進行。²⁹ 多代理模型通常比單一模型系統更具優勢，尤其是在高度複雜的環境下，透過任務分配提升整體效能。³⁰ 新創企業和大型科技公司正在開發多代理型 AI 系統，包括協助企業建構專屬客製化代理型 AI 的工具，以滿足特定業務需求。³¹

代理型 AI 的潛在應用案例

大型科技公司和新創企業正在開發可以將軟體開發、銷售、行銷，以及法令遵循和合規管理等功能部分自動化的初階解決方案。以下為目前可見的應用案例概覽（無法詳盡列出所有的應用）。部分案例仍處於概念驗證與示範階段，儘管前景看好，但是尚未準備好大規模企業部署。除了這些跨產業的應用之外，產業特定代理型 AI 應用也正在興起。

客服支援：客戶服務是一項不可或缺，且充滿壓力的工作，每年的人員流動率為 38%。³² 透過有效自動化部份客服工作流程，可以減輕員工壓力與重複性工作，同時提升客戶服務能力。³³ 相較於目前的客服聊天機器人，代理型 AI 可以處理更複雜的客戶問題，並能自主解決問題。舉例來說，某一間音響公司使用代理型 AI 協助客戶設定新設備，這是一個通常需要真人代理完成的多步驟流程。如果需要人工介入時，代理型 AI 會在轉移客戶之前自動彙整相關資訊，並提供問題摘要，從而提升處理效率。³⁴ 下一代客戶服務代理型 AI 預計可以整合多模態資料，除了傳統的文字聊天外，還將支援語音和影像互動，以提供更流暢且智能的客戶體驗。

網路安全：網路安全專家是技術型工作者短缺現象的典型示例，全球網路安全專業人才的缺口高達 400 萬名。³⁵ 此外，惡意行為者正在利用生成式 AI 滲透企業網路安全系統。新興的代理型 AI 網路安全系統可以透過將部分安全作業自動化，提升專家的工作效率。例如，這類代理型 AI 可以自主偵測攻擊和生成報告、提高系統安全性，以及將人工專業的工作負擔，最高可減少高達 90% 工作量。³⁶ 此外，代理型 AI 也可以協助軟體開發團隊偵測新代碼中的漏洞。透過自動測試與開發人員直接溝通，提供修復建議，而這項工作過去需由工程師手動執行。³⁷

法規遵循：金融服務、醫療保健等產業都必須定期進行法規遵循和合規性審查。然而，隨著相關法規之規模和複雜度不斷增加，加上合規專業人員短缺，企業面對日益嚴峻的合規挑戰。為應對這一困境，新創企業正在開發代理型 AI 法規遵循系統，可以快速分析法規與企業文件，評估企業是否合規，並提供具體的法規條文作為據，進一步向法遵專業人員提供分析與建議。³⁸ 目前使用生成式 AI 的公司表示，法遵和合規性是開發與部署生成式 AI 最大的障礙，甚至超過 AI 技術人才短缺與實施挑戰等問題。³⁹ 除法規的不確定性外，新法規的施行範圍和複雜度亦是企業使用 AI 面臨的關鍵挑戰。透過代理型 AI 協助企業即時瞭解和遵循新法規，可以協助加速企業更廣泛使用生成式 AI。

代理生成者與協調者：目前市場上已出現代理型 AI 解決方案，可自動化跨產業和垂直領域的工作流程。不過，企業可能無需等待市場成熟。它們可以開發自有的代理和多代理系統。例如，Google Vertex 提供無需編碼（No-Code）的代理型 AI 建構工具，企業可利用該技術建立特定任務的代理型 AI，如根據過往的行銷活動自動生成行銷素材。⁴⁰ 此外，LangChain 是使用開源技術，協助公司建置多代理型 AI 系統。例如，新創企業 Paradigm 推出了一款「智能試算表」，由多個代理型 AI 協作收集來自不同來源的數據、整理資料，並完成工作。⁴¹

結論

代理型 AI 擁有巨大的潛力，可以透過自動化完整工作流程和獨立任務來提升知識工作者的生產力。此項技術與當前的聊天機器人和智慧助理不同，代理型 AI 可自主行動，無論做為單一代理運作或與多個代理合作，都能獨立執行任務。不過，代理型 AI 仍在開發和使用的初期階段。儘管早期的代理型 AI 展現出令人印象深刻的能力，但是仍可能會出錯及陷入循環的狀態。在多代理系統中，「幻覺」(Hallucinations) 可能在代理之間傳播，導致錯誤資訊擴散，使代理型 AI 做出錯誤決策或提供不正確的答案。⁴² 但當前的最佳應用模式可能是「人員監督式自動化 (Human-on-the-Loop)」，即代理型 AI 在執行決策後由真人審核，而非更嚴格的「人員參與式自動化 (Human-in-the-Loop)」，這樣的機制可提高代理型 AI 的可部署性。當代理型 AI 遇到瓶頸時，可以諮詢專家尋求協助，協助其克服挑戰和持續進步。在這種模型中，代理型 AI 如同一位學習經驗的新進員工，透過執行實際工作來累積經驗並學習成長。⁴³

儘管一些企業投入數十億資金研發穩定且可靠的代理型 AI，但何時能實現、以及在何種條件下實現，仍不明確。代理型 AI 是否會在 2025 年或未來五年內被廣泛使用？其普及是否需要突破性的創新，還是僅需調整現有的 AI 技術與訓練方法？若正在開發代理型 AI 的大型科技公司和新創企業成功突破技術瓶頸，市場格局將迅速改變。試想，代理型 AI 能夠處理多模態數據、運用工具、協調其他代理、記憶與學習，並且穩定執行各種任務。此外，企業還能在「零程式碼環境」中，僅透過對話式文本提示快速開發客製化的智能代理，這將進一步加速其應用與發展。

因為代理型 AI 的技術發展迅速且應用前景極具吸引力，企業應及早為未來做好準備。企業在準備過程中，應考慮下列策略，以確保能夠在這場技術變革中佔據領先地位。

優先考慮和重新設計適合代理型 AI 的工作流程：企業應根據代理型 AI 的技術能力和自身業務的最高價值領域，確認適合 AI 執行的任務和工作流程，並進行優化與重設，以刪除不必要的步驟。確保代理型 AI 解決方案具有明確的目標，且能存取必要的資料、工具與系統。雖然這些代理型 AI 可以協助適應環境，但是若流程雜亂和未優化的流程，仍可能導致成效不如預期。

專注於資料治理和網路安全：為確保代理型 AI 能夠創造價值，企業必須允許其存取重要和可能具敏感性的企業資料，並與內部系統與外部資源互動。因此，企業在開始部署代理型 AI 之前，應先建立健全的資料治理和網路安全機制。目前，生成式 AI 的早期採用者已將 IT 投資的重點放在資料管理 (75%) 和網路安全 (73%) 這兩大領域。⁴⁴ 儘管已進行此類投資，仍有 58% 的企業對於在 AI 模型內使用敏感資料和管理資料安全感到非常憂心，僅有 23% 的企業表示已為管理生成式 AI 風險和治理做好充分準備。換言之，即便是當前領先的生成式 AI 使用者，仍未作好萬全準備以迎接代理型 AI。對於仍未使用生成式 AI 的企業來說，需要更長的時間來追趕這一技術趨勢。

在平衡風險與回報方面：企業在開始導入代理型 AI 時，應謹慎評估代理型 AI 的自主權限和資料存取權限。建議從低風險應用場景（例如處理非關鍵資料和人為監督的低風險任務）開始，以建立穩健的資料管理、網路安全和治理架構，並確保 AI 應用的安全性與合規性。當基礎治理措施建立完善之後，企業應考慮使用更具策略價值的應用場景，例如允許代理型 AI 存取更多高價值數據、使用更廣泛的工具，及賦予更高的自主決策權限，以提升業務效能與競爭優勢。

保持理性審視：代理型 AI 正在快速發展，預計未來一年內其功能可能會持續增強，並將應用於更廣泛的橫向與縱向領域。預計 2025 年將會出現令人印象深刻的 AI 展示、模擬和產品發表。然而，企業應保持理性與謹慎態度，因為前述挑戰可能仍需要時間解決。

By **Jeff Loucks**
United States

Baris Sarer
United States

Gillian Crossan
United States

China Widener
United States

Endnotes

1. According to Deloitte's State of Generative AI in the Enterprise survey, 23% of enterprises that currently use gen AI are exploring "gen AI agents" to a "large" or "very large" extent, with another 42% exploring it "to some extent." Given the high interest in agentic AI and the products and services that are being launched by startups and established tech companies, we expect this interest to turn to action, at least on an experimental scale.
2. CB Insights. Gen AI Investment Database, Aug 21, 2024. This data excludes Open AI. It includes funding to companies that are developing agentic AI with "varying degrees of autonomy."
3. Kate Clark, "[Investors undaunted by spate of AI acqui-hires](#)," The Information, Aug. 19, 2024.
4. Cambridge English Dictionary, "[Agency](#)," accessed Aug. 26, 2024.
5. Cambridge English Dictionary, "[Autonomous](#)," accessed Aug. 26, 2024.
6. For humans, agency and autonomy are moral and political concepts. In the context of gen AI agents, we are speaking only of the extent to which software-based technology has scope to design and perform tasks without human direction.
7. Faruk Muratovic, Duncan Stewart, and Prashant Raman, "[Tech companies lead the way on generative AI: Does code deserve the credit?](#)" Deloitte Insights, Aug. 2, 2024.
8. Scott Wu, "[Introducing Devin, the first AI software engineer](#)," Cognition Software, March 12, 2024.
9. Rina Diane Caballar, "[AI Coding is going from copilot to autopilot](#)," IEEE Spectrum, April 9, 2024.
10. Jenna Barron, "[Codeium's new Cortex assistant utilizes complex reasoning engine for coding help](#)," SD Times, Aug. 14, 2024; Aswin Ak, "[OpenDevin: An artificial intelligence platform for the development of powerful AI agents that interact in similar ways to those of a human developer](#)," Marktechpost, July 28, 2024.
11. Carl Franzen, "[Codium announces Codiumate, a new AI agent that seeks to be Devin for enterprise software development](#)," VentureBeat, April 3, 2024.
12. Cognition Software, "[SWE-bench technical report](#)," March 15, 2024.
13. Big tech companies continue to improve their software co-pilots to make them more like gen AI agents. For example, see: Alex Woodie, "[The semi-autonomous agents of amazon Q](#)," BigDATAWire, May 3, 2024.
14. Molly Talbert, "[Overcoming disruption in a distributed world: Insights from the Anatomy of Work Index 2021](#)," Asana, January 14, 2024.
15. Total factor productivity, which measures how efficiently both capital and labour are used, can be a proxy for knowledge worker efficiency. Knowledge work requires access to capital-intensive technology, and effectively designed processes.
16. US Bureau of Labor Statistics, "[Table A. Productivity, output, and inputs in the private nonfarm business and private business sectors for selected periods, 1987-2023](#)," March 3, 2024.
17. Organisation for Economic Co-operation and Development, "[Multifactor productivity](#)," accessed Oct. 30, 2024.
18. Jim Rowan, Beena Ammanath, Costi Perricos, Brenna Sniderman, and David Jarvis, [State of gen AI in the Enterprise](#), Q3 report, Deloitte, August 2024.
19. Ibid.
20. Ibid.
21. Ibid.
22. James O'Donnell, "[Why OpenAI's new model is such a big deal](#)," MIT Technology Review, Sept. 17, 2024.
23. Janakiram MSV, "[AI agents: Key concepts and how they overcome LLM limitations](#)," The New Stack, June 11, 2024.

24. "OpenAI, ["Learning to Reason with LLMs,"](#) Sept. 12, 2024.
25. Anna Gutowska, ["What are AI Agents?"](#) IBM, July 3, 2024.
26. Janakiram MSV, ["AI agents: Key concepts and how they overcome LLM limitations."](#)
27. Simon Willison, ["Notes on OpenAI's new o1 chain-of-thought models,"](#) Simon Willison's Blog, Sept. 12, 2024.
28. Hamidou Dia, ["So much more than gen AI: Meet all the other AI making AI agents possible,"](#) Google Cloud Blog, Aug. 20, 2024.
29. Vivek Kulkarni, Scott Holcomb, Prakul Sharma, Edward Van Buren and Caroline Ritter, ["How AI agents are reshaping the future of work,"](#) Deloitte AI Institute, November 2024.
30. The Economist, ["Today's AI models are impressive. Teams of them will be formidable,"](#) May 13, 2024.
31. CB Insights, ["The multi-agent AI outlook: Here's what you need to know about the next major development in genAI,"](#) Aug. 30, 2024.
32. Mike Desmarais, ["The call center burnout problem,"](#) SQM Group, Feb. 24, 2023.
33. It's important to balance the work of human agents. When they get only the most complicated and difficult cases, it can lead to burnout. See Sue Cantrell, et al., ["Strengthening the bonds of human and machine collaboration,"](#) Deloitte Insights, Nov. 22, 2022.
34. Sierra, ["Sonos elevates the listener experience,"](#) Feb. 13, 2024.
35. Michelle Meineke, ["The cybersecurity industry has an urgent talent shortage. Here's how to plug the gap,"](#) World Economic Forum, April 28, 2024.
36. Ken Yeung, ["Dropzone AI gets \\$16.85M for autonomous cybersecurity AI agents that reduce manual work by 90 percent,"](#) VentureBeat, April 25, 2024.
37. Simon Thomsen, ["Software development cybersec startup Nullify banks \\$1.1 million pre-seed round,"](#) Startup Daily, June 26, 2023.
38. Kyt, Dotson, ["Norm AI raises \\$27M to help businesses handle regulatory compliance with AI agents,"](#) SiliconANGLE, June 26, 2024.
39. Rowan, [State of Generative AI in the Enterprise](#), Q3 report.
40. Ron Miller, ["With Vertex AI Agent Builder, Google Cloud aims to simplify agent creation,"](#) TechCrunch, April 9, 2024.
41. Iris Coleman, ["Paradigm utilizes LangChain and LangSmith for advanced AI-driven spreadsheets,"](#) Blockchain.News, Sept. 5, 2024.
42. The Economist, ["Today's AI models are impressive."](#)
43. Maria Korolov, ["AI agents will transform business processes — and magnify risks,"](#) CIO, Aug. 21, 2024.
44. Rowan, [State of Generative AI in the Enterprise](#), Q3 report.

Acknowledgements

Authors would like thank **Chris Arkenberg**, **Duncan Stewart**, and **Ankit Dhameja**.

Cover image by: **Jaime Austin**; Getty Images, Adobe Stock.

深度偽造技術的衝擊：網路安全層級挑戰及其深遠影響

隨著偵測和對抗虛假內容的力道加大，維持互聯網可信度的成本可能轉移至消費者、創作者和廣告商，對數位生態系統帶來深遠影響。

深偽技術（Deepfakes），即為由 AI 工具生成的逼真照片、影片和音訊，正在使觀眾越來越難相信自己在網路上看到的內容。隨著 AI 生成的內容數量與複雜度日益增加，影像和音訊都可能遭到有心人士利用，以散播假消息與進行詐騙。社群媒體網路中充斥著此類內容，而引起了廣泛的懷疑與擔憂。¹

在 Deloitte 2024 年的《互聯消費者調查》（Connected Consumer Survey）中，半數受訪者表示，他們對於網路上的資訊準確性與可靠性，較前一年更抱持懷疑態度。在熟悉或使用生成式 AI 的受訪者中，68% 的人表示擔心合成內容可能遭到利用於欺騙或詐騙，而 59% 的人則表示難以分辨人類創作的內容或由 AI 生成的媒體內容，其之間的差異。此外，高達 84% 熟悉生成式 AI 的受訪者認為應清楚標示由生成式 AI 開發的內容。²

標示 AI 生成內容是媒體機構和社群媒體平台提醒使用者注意合成內容的一種方式，但是隨著深偽技術與更多的先進模型結合，將可以生成合成內容及操控現有媒體，因此可能需要採取更複雜的技術對策，以偵測虛假內容和重建信任感。

業界分析人員預估，全球深偽檢測市場（涵蓋科技、媒體與社群媒體龍頭使用的技術），將由 2023 年的 55 億美元成長至 2026 年的 157 億美元，年均成長率為 42%。³ 將與網路安全市場發展軌跡類似，媒體和科技業者可能會透過投資於內容認證解決方案，並透過跨界聯盟合作來應對日益複雜的深度偽造技術。然而，隨著可信內容的驗證成本上升，消費者、廣告商，甚至內容創作者可能需要承擔更高的成本。⁴

目前業界的應對策略主要可分為兩大類：深度偽造內容檢測（Detecting Fakes）與內容溯源與真實性驗證（Establishing Provenance）。

偵測深度偽造內容

科技公司通常運用深度學習和電腦視覺技術分析等方式，識別合成媒體中是否存有詐騙或操控的跡象，並透過機器學習模型辨識深偽內容中的模式與異常特徵。⁵ 這些工具也可以偵測影像和音訊內容中的不一致之處，例如微妙的嘴唇運動或聲音語調的細微波動，這些變化往往與人類的自然表現有所不同。⁶

部份生成式 AI 工具擁有內建檢測功能，可識別某項內容是否由其技術生成的功能，但是這些功能可能無法檢測由其他 AI 模型創建之深偽內容。⁷ 因此，一些深偽檢測工具會尋找生成式 AI 工具的操控跡象或「指紋」。⁸ 此外，部份工具採取「白名單」與「黑名單」方法（即維護可信來源清單和已知欺騙者的名單），且有一些工具會尋找人類特徵的證據（例如自然的血液流動、臉部表情和音調變化），以對抗 AI 生成的虛假內容。⁹

目前的深偽檢測工具宣稱正確率超過 90%。¹⁰ 不過，令人擔心的是有心人士可能會使用開源生成式 AI 模型繞過這些檢測機制。由於 AI 能夠自動化內容生成，這可能超出當前檢測工具的負荷。此外，生成式 AI 工具可以根據使用者的指令細微調整輸出內容，進一步增加識別的困難度。¹¹

社群媒體平台本身通常依賴 AI 工具來幫助偵測影像或影片中的問題內容，並根據風險程度進行評分，將最可疑的項目轉交給人工審核員進行最終判定。然而，此流程既耗時又昂貴，因此業界目前正在努力藉由機器學習加速此審查流程。¹²

值得注意的是，這種對抗深度偽造內容的方式與網路安全領域的挑戰極為相似。就如同企業在網路與數據保護方面採取多層次安全防護措施，新聞媒體和社群媒體公司可能需要多項工具，搭配內容溯源措施，來驗證數位內容的真實性與可信度。

建立溯源與信任機制

部分公司正在探索加密元資料（Cryptographic Metadata）或數位浮水印（Digital Watermark）技術，是在創建媒體檔案時加入加密元資料，或數位浮水印。這些標記在媒體內容創建時即嵌入檔案，並隨附詳細的來源資訊，確保所有修改與變更皆可被追蹤與驗證。¹³

社群平台正在與媒體機構、設備製造商和科技公司合作，透過跨產業聯盟維持內容真實性的標準。多家科技與媒體機構，包括 Deloitte 在內，都已經加入了內容出處暨真實性聯盟（Coalition for Content Provenance and Authenticity, C2PA），承諾使用 C2PA 元資料標準，以便更有效地驗證 AI 生成影像內容。¹⁴ C2PA 技術是透過建立詳細的變更和修改紀錄，記錄影像從初始創建到編輯流程之整個生命週期的每一個步驟。¹⁵ 使內容平台和使用者可以查閱來源資訊，以評估其可信度。

此外，為了區隔由真人營運的帳戶，部分社群媒體平台開始實施創作者認證機制。這些機制可能會要求創作者提交身分證明文件及支付認證費用，以確保帳戶真實性。部份平台甚至將驗證作為參與收益分潤計畫的門檻，進一步提高驗證的誘因。¹⁶

隨著 AI 生成內容越來越普遍，認證真人營運帳戶的真實性，將有助於平台提升信任度與內容公信力。¹⁷ 然而，平台則需要評估，是否將認證費用轉嫁給創作者、廣告商或一般使用者，並確保此模式的長期持續性。

監管機制的發展與挑戰

雖然已有部分政府開始採取措施監管內容的真實性，¹⁸ 但是更全面且全球協調的法規可能更具效益。同樣重要的是提高公眾意識的教育宣傳活動，協助使用者理解深偽技術（Deepfake）帶來的風險，在相信媒體內容為真實之前學習進行驗證與辨識。

在美國，已提出了要求在 AI 生成之內容上使用數位浮水印的法規，目前正由美國參議院商務、科學和交通委員會（Senate Committee on Commerce, Science, and Transportation）進行審議。¹⁹ 此外，加州正在考慮 AB-3211 法案，該法案要求設備製造商更新軟體，確保照片上附加內容來源元資料，並強制線上平台揭露資料來源。若立法通過，則此項法案將會在 2026 年正式生效。²⁰ 同時，美國部分州也頒布了將惡意製作和散布非經同意的深偽內容定為犯罪行為的法律。²¹ 美國聯邦貿易委員會（Federal Trade Commission）正在制定新法規，目的是禁止生成和散播模仿個人的深偽內容。²²

在歐盟近期修訂的《人工智慧法案（EU AI Act）》主要在於強調透明度，要求必須明確標示 AI 生成內容與深偽技術，此項策略有助於 AI 技術持續進步，同時可以確保使用者辨識接收的內容來源。歐盟委員會（European Commission）已成立 AI 辦公室，致力於促進發展與應用 AI，並推動人工生成或修改標示與管理。²³

深偽技術發展迅速，需要靈活和具有適應性的法規框架，才能隨著技術進步不斷進化，確保有效應對潛在風險並促進 AI 的負責任應用。

結論

照片、影片或音訊內容可以透過分析技術和溯源驗證其來源的方法，確認其真實性，隨著生成式 AI 被持續廣泛使用於創建更多合成媒體，且惡意行為者不斷調整其模型與輸出內容以規避檢測，預計媒體公司和社交平台很可能會針對兩大技術進行投資，以維護數位內容的可信度。

在生成式 AI 變得越來越強大、功能越來越多的背景之下，最重要的是持續領先惡意行為者至關重要。核醫血容積測定（Blood Volume Determination）與人臉分析等更先進的技術，可以協助區分真實內容和被操控的內容。然而，如同網路安全工具的發展原則，這些措施應盡可能不會干擾使用者和消費者，在不犧牲使用者體驗的情況下，確保內容的完整性。例如，數位浮水印等技術可以在不影響內容品質或佔用即時計算資源情況下，驗證內容真實性。²⁴

對於依賴機器學習模型（或向第三方供應商購買技術）來檢測虛假內容的公司而言，此類機構應優先選擇使用多樣化且高品質資料集的工具與供應商。這些資料集，應涵蓋廣泛影像、影片和音訊內容，並確保涵蓋不同人口統計群體，以降低演算法偏差，提升檢測準確度。²⁵

科技和媒體公司應與同行²⁶ 跨產業合作，共同推動及制定深偽檢測與數位內容驗證的標準。例如，在設備製造商和媒體機構共同簽署內容，以確認其創建和發布時，數位浮水印技術會更具效用。此種集體努力可以推動更健全和普遍認同的做法，提升數位內容中的整體安全與信任度。

在企業安全層面，各行各業之公司都應該意識到生成式 AI 可以提升社交工程攻擊，且可能會削弱部分身分驗證機制。²⁷ 因此，企業可能需要實施額外的驗證層級，尤其是針對涉及影片和音訊的流程。此外，應鼓勵終端使用者使用可靠來源多方驗證資訊，並使用多重要素驗證降低深偽技術帶來的風險。因為威脅態不斷變化，使用者教育訓練（類似網路資安意識訓練）可能也是企業必須考量的重要措施，以提升員工對深偽技術威脅的辨識能力與應對能力。

這些策略不僅可以防範深偽技術帶來的安全威脅，且有助於科技和媒體公司在維護數位內容完整性與信任方面的領導地位。在此關鍵時刻，企業可以透過打造出可信的內容空間，在日益不確定的數位環境中，樹立可靠資訊來源的形象。

By **Michael Steinhart**
United States

Ankit Dhameja
India

Bree Matheson
United States

Gillian Crossan
United States

Endnotes

1. Margaret Talev and Ryan Heath, [“Exclusive poll: AI is already great at faking video and audio, experts say,”](#) Axios, accessed Oct. 28, 2023.
2. Susanne Hupfer, Michael Steinhart, et.al, [“2024 Connected Consumer Survey,”](#) Deloitte, December 2024
3. Vivaan Jaikishan, Cameron D'Ambrosi, Jennie Berry, and Stacy Schulman, [“The rising threat of deepfakes: Detection, challenges, and market growth,”](#) Liminal, May 7, 2024.
4. Ian Shepherd, [“Human vs. machine: Will AI replace content creators?”](#) Forbes, April 26, 2024.
5. Analytix Labs, [“Detecting deepfakes: Exploring advances in deep learning-based media authentication,”](#) Medium, January 4, 2024.
6. For example, see: Intel, [“Trusted media: Real-time FakeCatcher for deepfake detection,”](#) accessed Oct. 28, 2024.
7. Cade Metz and Tiffany Hsu, [“OpenAI releases deepfake detector to disinformation researchers,”](#) The New York Times, May 2024.
8. Danial Samadi Vahdati, Tai D. Nguyen, Aref Azizpour, and Matthew C. Stamm, [“Beyond deepfake images: Detecting AI-generated videos,”](#) Drexel University, accessed Oct. 28, 2024.
9. Alex McFarland, [“5 best deepfake detector tools & techniques \(October 2024\),”](#) Unite.AI, Oct. 1, 2024.
10. Konstantin Simonchik, [“Deepfake detection: Accuracy of commercial tools,”](#) LinkedIn, February 2024
11. Jiansong Zhang, Kejiang Chen, Weixiang Li, Weiming Zhang, and Nenghai Yu, [“Steganography with generated images: Leveraging volatility to enhance security,”](#) IEEE Transactions on Dependable and Secure Computing 21, no. 4 (2024): pp. 3994–4005; see also: Mike Bechtel and Bill Briggs, [“Defending reality: Truth in an age of synthetic media,”](#) Deloitte Insights, Dec. 4, 2023; and, Loreben Tuquero, [“AI detection tools for audio deepfakes fall short. How 4 tools fare and what we can do instead,”](#) Poynter, March 21, 2024.
12. Barbara Ortutay, [“Content moderation in the AI era: Humans are still needed across industries,”](#) Fast Company, April 23, 2024; also see: Meta, [“How review teams work,”](#) Jan. 19, 2022.
13. Glenn Chapman, [“Meta wants industry-wide labels for AI-made images,”](#) AFP News, Feb. 6, 2024; also see: Nick Clegg, [“Labeling AI-generated images on Facebook, Instagram and Threads,”](#) Feb. 6, 2024; Sasha Luccioni et al., [“AI watermarking 101: Tools and techniques,”](#) Hugging Face, Feb. 26, 2024; and Partnership on AI, [“Building a glossary for synthetic media transparency methods, part 1: Indirect disclosure,”](#) Dec. 19, 2023.
14. Ryan Heath, [“Inside the battle to label digital content as AI-generated media spreads,”](#) Axios, accessed Oct. 28, 2024.
15. Demian Hess, [“Fighting deepfakes with content credentials and C2PA,”](#) CMSWire, March 13, 2024.
16. Andrew Hutchinson, [“X will require ad revenue share participants to confirm their ID,”](#) Social Media Today, May 22, 2024.
17. Guy Tytunovich, [“The future of trust and verification for social media platforms,”](#) Forbes, May 22, 2024.
18. Amanda Lawson, [“A look at global deepfake regulation approaches,”](#) Responsible Artificial Intelligence Institute, April 24, 2023.
19. US Congress, [“S.2765—Advisory for AI-Generated Content Act,”](#) Sept. 12, 2023.
20. California Legislative Information, [“Assembly Bill 3211—California Digital Content Provenance Standards,”](#) Aug. 24, 2024.
21. Kevin Collier, [“States are rapidly adopting laws regulating political deepfakes,”](#) NBC News, Aug. 7, 2024.
22. Federal Trade Commission, [“FTC proposes new protections to combat AI impersonation of individuals,”](#) Feb. 15, 2024; also see: Michelle M. Graham, [“Deepfakes: Federal and state regulation aims to curb a growing threat,”](#) Thompson Reuters, June 26, 2024.

23. Melissa Heikkilä, “[Five things you need to know about the EU’s new AI Act](#),” MIT Technology Review, Dec. 11, 2023.
24. Deloitte, “[How to safeguard against the menace of deepfake technology](#),” accessed Oct. 28, 2024.
25. AI Index Steering Committee, “[The AI Index 2024 Annual Report](#),” accessed Oct. 28, 2024.
26. AI Election Accord, “[A tech accord to combat deceptive use of AI in 2024 elections](#),” accessed Oct. 28, 2024.
27. Stu Sjouwerman, “[The growing threat of AI in social engineering: How business can mitigate risks](#),” Fast Company, April 8, 2024.

Acknowledgements

The authors would like to thank **Jeff Loucks, Susanne Hupfer, Duncan Stewart, Jeff Stoudt, Jason Williamson, Tim Davis, Gopal Srinivasan, Shreeparna Sarkar**, and **Andy Bayiates** for their contributions to this article.

Cover image by: **Jaime Austin**; Getty Images, Adobe Stock

雲端變的精簡：「雲端財務管理 (FinOps)」 讓每一分錢都能發揮更大的效益

企業雲端支出不斷增加，使用 FinOps 策略可以提升資金運用效益。公司可以透過 FinOps 方法實現成本節約、提高價值以及建立跨部門合作，進一步強化營運效能與財務管理能力。

全球雲端支出可能會在 2025 年突破 8,250 億美元，但是在提及支出情況時，企業領導高層可能很難提出具體的答案。¹ 多數公司不完全掌握自身的雲端支出，或難以明確解釋其資金流向。

隨著企業日益依賴雲端服務，使制定管理雲端投資的有效策略變得極為重要。使用 FinOps（雲端財務管理），結合 Finance（財務管理）和 DevOps（開發與運維）的方法論，能夠協助企業追蹤和優化雲端支出的策略。Deloitte 預測，僅在 2025 年，使用 FinOps 工具和策略的企業，可以節省高達 210 億美金，在未來幾年，此數字可能繼續增加。部分企業甚至可以節省高達 40% 的雲端服務成本。展望未來，我們預期沒有 FinOps 團隊的企業將會迅速採取初步措施，迅速落實基礎 FinOps 策略，而經驗豐富的 FinOps 專家則會進一步發展開發出更嚴謹的雲端資源優化策略。

雲端環境日益複雜，隱性浪費問題浮現

雲端技術已成為企業維運來說不可或缺的基礎。相比於傳統物理基礎架構需要耗時數周甚至數月進行伺服器採購與部署，啟動新的雲端環境通常僅需要點擊幾次即可快速啟動新雲端環境。其協助實現將便利性和可擴展性與普及化。雲端技術推動創新加速，即便企業不具備龐大研發團隊，仍能利用雲端推動業務創新。例如能協助隨選影音 (VOD)、共享乘車、金融科技和遠距治療等新興產業，均透過雲端技術顛覆傳統市場。² 此外，同時能強化資料分析、遠距辦公，以及 AI 等應用的基礎。

現今，投資軟體工程的企業範圍日益擴大。商用軟體 (COTS) 可能無法繼續滿足所有企業需求。例如汽車製造商戴姆勒 (Daimler) 等汽車公司已經成立了開發人員團隊，打造電動車的軟體平台。³ 即使即使是傳統行業 (如木棧板配送)，對於客製化軟體和技術專業知識的需求也在日漸成長。⁴ 這些趨勢推動了企業的雲端支出金額不斷攀升。

但是，雲端已經變得越來越複雜，帶來諸多挑戰。企業通常需要同時管理其私有運算資源和公共雲服務，目前已有 73% 的企業使用此種混合雲基礎設施。此外，半數以上（53%）的企業使用多個雲端提供業者，以獲取最佳價格、特定功能或避免供應商鎖定。⁵ 個別部門分散採購，例如財務、人資或行銷等業務部門，可能會自行採購雲端軟體，導致 IT 部門無法全面掌控雲端資源。這些因素都可能加劇公司資料整合、合規管理和資安風險等領域挑戰。因此，有效的雲端成本管理策略將成為企業未來競爭力的關鍵。

整體而言，多數企業難以嚴格控制其雲端預算。調查顯示，過去一年受訪企業中有 50% 企業超支，且平均超支幅度達 15%。⁶ 其中一項因素是雲端的「隨用隨付」計費模式，導致成本變動較大，使得預算規劃與成本預測變得更加困難。在極端情況下，曾經有雲端工程師在無意間操作，一夕導致數千美元的雲端支出。⁷

雲端運算不便宜，成本可能高於傳統的私有基礎架構，⁸ 目前，雲端支出已迅速成為公司的最大 IT 支出項目之一。例如，可口可樂公司最近簽署了一項 11 億美元的雲端需求合約，突顯出企業對雲端資源的高度依賴。⁹

然而，儘管企業投入鉅額資金在雲端服務上，根據雲端技術負責人的評估，27% 的雲端支出形同浪費。¹⁰ 隨著企業已開始意識到此點，許多企業已採取應對措施。目前，50% 的企業目前設有專門的 FinOps 團隊，另外 20% 預計將會在一年內組建相關團隊，以強化雲端成本管理與資源優化。¹¹

FinOps 起步與實施

FinOps（財務維運管理）是一種財務管理學科，其範圍涵蓋技術與財務層面，例如重新設計雲端服務折扣額度等。然而，FinOps 的長期影響在於企業文化的變革，其核心理念在於跨單位的責任分擔與財務問責，確保每一筆雲端支出都能對應企業價值。

開始實施 FinOps 的關鍵在於規劃，包括檢視現有雲端策略、評估標記與預警機制，然後定義關鍵績效指標。¹² 提升可視性是最實用的第一步，企業可透過盤點現有雲端資源，進一步評估其與業務需求的契合度。為此，雲端服務業者提供了資源監控工具和著重成本管理解決方案，並且，企業亦可選擇第三方 FinOps 平台，或具有可以提供更詳盡之成本分析指標，進一步提升管理效能。

不過，企業需注意 FinOps 工具本身的成本，某些高階 FinOps 解決方案的費用可能佔企業雲端總支出的 3% 至 5%，因此，在導入 FinOps 工具前。企業應充分了解自身的雲端經濟模式，確保 FinOps 的導入能帶來實質效益，而非額外負擔。

FinOps 的起步階段：初步措施

剛開始導入 FinOps 的企業而言，首要目標通常是減少浪費、確保資源獲得最佳分配、檢視雲端合約，以及善用折抵額度與折扣制度，以提升成本效益。

雲端資源浪費與使用率管理：我們預期，降低資源浪費將成為 FinOps 初學者最適合的起點。FinOps 管理工具和儀表板可以協助企業識別閒置或使用率不足的雲端資源，以進行調整或關閉，進而能立即降低成本。浪費的範例包括規模過大的虛擬機器、多餘的儲存空間、孤立資源，或重複的資料。在 FinOps 應用較成熟的企業中，可能會使用預測性分析，用於預估資源使用需求，以及使用自動化治理機制動態調整雲端容量。這些調整通常可以由企業內部雲端工程團隊執行，使優化工作能夠快速落實。

雲端架構與層級管理：雲端服務並非一體適用，運算與儲存資源在效能與價格上存在明顯差異。因此，公司應檢視雲端資源配置是否符合應用需求，以及是否能充分滿足應用程式的需求。部分應用程式可能可以在更具成本效益的實例中順利運行。例如，為了減少季節性波動時，活動售票網站可能會選擇不持續使用 CPU 的雲端執行個體，當網站訪問量增加時自動擴展資源，而非長期支付高規格運算成本，從而降低不必要的支出。¹³

激勵機制：雲端平台提供各類折扣計劃，可帶來明顯的成本節省。部份平台允許企業保證穩定的使用量，以換取較優惠的費率。此外，對部分公司來說，直接與雲端服務供應商重新協議，可能是一種有效的成本優化策略。雲端公司對於此種做法通常持開放態度，且樂於提供較低費率以換取多年的契約承諾，從而確保長期合作並提高業務穩定性。

深化 FinOps：拓展與優化策略

2025 年，經驗豐富的 FinOps 從業人員可能會持續優化上述方法，但是也預期會進一步提升成本可觀測性及成本管控的應用與實踐。

當責制度：由於雲端技術對許多企業各個環節都至關重要，因此各部門（與團隊）皆應應對其雲端支出承擔財務責任。企業可以透過分攤費用模式（直接向部門收費）或成本揭露模式模型（顯示部門的成本負擔），來確保雲端成本的透明度與責任歸屬。¹⁴ 此外，需要健全的標籤策略，將資源成本分配給特定團隊或專案，理想的情況是根據預先定義規則進行自動標籤。透過此機制企業可以營造出一種成本意識文化，使部門團隊都能參與降低雲端成本優化。

地端：FinOps 社群（如 FinOps 基金會），已經開始鼓勵企業將內部部屬基礎建設納入雲端成本管理討論範疇，因為此類基礎設施的成本往往不夠透明，難以被全面評估。¹⁵ 公司應考慮整體 IT 資產的角度思考成本管理，然而，複雜度可能很高，因為中央雲端團隊可能必須與分公司和基礎建設管理部門協調，並考量當地需求所使用的多種硬體與軟體工具。內部部署成本優化策略可包括取消不需要的授權和延長硬體使用週期。

永續發展：FinOps 同時與蔚為風潮的「綠色運維（GreenOps）」有關。GreenOps 指的是一系列優化永續發展的雲端管理策略，目的在於提升永續發展表現。透過 FinOps 報告工具提供的詳細指標，有助於衡量能源消耗、碳排量及其他永續發展目標。¹⁶ 隨著歐盟實施《企業永續發展報告指令》（Corporate Sustainability Reporting Directive）等重大報告法規推動，¹⁷ 企業若能透過 FinOps 追蹤能源與碳排放指標，並進行改善，將能進一步提升其永續經營能力，並從中獲得重要附加價值。

FinOps 帶來具體的影響

許多企業透過 FinOps 實踐在許多致力於節省雲端成本的公司中，包括以下案例：

- **Airbnb：**¹⁸ 旅遊應用程式 Airbnb 成功節省了 6,350 萬美元的雲端服務成本。採取策略其中一項是將儲存空間轉移至成本較低的服務層級，並將自建的備份系統替換為雲端服務業者的替代方案。
- **Sky Group（天空集團）：**¹⁹ 媒體娛樂公司 Sky 發現，其年度雲端預算在六個月內已花費了整整一年的雲端預算。該集團在部署自行研發的 FinOps 工具之後，節省了 150 萬美元，並導入可視化儀表板，在第二年節省了 380 萬美元的成本。
- **The Home Depot（家得寶）：**²⁰ 此家庭裝修零售商在 2022 年成立了專門的雲端成本管理團隊，發現了相較前一年可以節省上千萬美元。
- **Lyft（來福車）：**²¹ 共享乘車平台 Lyft 在六個月內降低每趟行程的雲端成本 40%，透過一個基於試算表的軟體工具向全公司公開帳單數據，因此進而推動一系列資源適配計畫。
- **WPP 集團：**²² 廣告集團 WPP 在部署 FinOps 三個月後，即節省了 200 萬美元，最終成功地將每年的雲端支出成本削減了 30%。該集團使用多種工具與技術，包括自動生成的資源規模建議。

目前，許多組織正在積極投資 FinOps。例如，推動最佳實踐雲端財務管理之非營利組織的 FinOps 基金會，成員包括沃爾瑪（Walmart）、萬事達卡（Mastercard）和美國航空（American Airlines）。²³

結論：雲端單位經濟

FinOps 的興起，反映出對雲端支出日漸成長的需求，他們需要更佳的可視化、更完善預算編列和主動控制雲端支出的需求，隨著企業對雲端的依賴性不斷增加，這種需求也持續增加。

展望未來，全球的 IT 支出將會持續增加（在 2025 年將超過 5.1 兆美元），²⁴ 部分動力是來自於數位轉型與 AI 的驅動。此外，私有基礎建設仍佔整體工作負載的一半左右，若遷移至公有雲，則可能會使雲端費用大幅增加。再者，居高不下的利率（至少相較於過去十年），促使企業更注重獲利性與降低成本，且在消弭成本變動性方面的需求更高。換言之，FinOps 的發展時機已經成熟。

應將 FinOps 視為一種長期實踐，是營運策略的核心。不應將其視為一項簡單的解決方案。FinOps 是從降低成本出發，但是最終可以讓雲端支出從單純的支出項目轉化為策略性資產和推動因素。

部分最先進之公司的最終目標，可能是創建「雲端單位經濟效益」模型。這種方法可以將使用之雲端服務的每單位相關成本進行量化，包括各項應用程式、工作負載或處理的資料規模，並使其符合產出的業務指標，例如收益、每次交付成本、每次預約成本和每一次行程的成本。此種更詳細的深入洞察，有助於企業在整體業務環境中做出更有效的 IT 決策，確保每一項支出都與利潤有關及可追溯。

對於部分公司來說，節省的成本可以再投資於新的成長機會，例如透過新的雲端服務擴大規模或加速產品發展藍圖。

雲端服務固然複雜且成本高昂，但應用 FinOps 的公司能夠充分利用其優勢，產生更大的利潤價值。

By Ben Stanton
United Kingdom

Paul Lee
United Kingdom

Adam Gogarty
United Kingdom

Gillian Crossan
United States

Endnotes

1. Gartner, “[Gartner forecasts worldwide public cloud end-user spending to surpass \\$675 billion in 2024](#),” press release, May 20, 2024.
 2. Yury Izrailevsky, Stevan Vlaovic, and Ruslan Meshenberg, “[Completing the Netflix cloud migration](#),” Netflix, Feb. 12, 2016.
 3. Douglas Busvine, “[Daimler to hire 1,000 programmers in Germany](#),” Reuters, April 18, 2021.
 4. CHEP, “[CHEP uses ‘track and trace’ technology on its reusable pallets](#),” press release, April 8, 2022.
 5. Flexera, “[2024 State of the cloud report](#),” 2024.
 6. Ibid.
 7. Parshv Jain, “[Avoiding costly cloud mistakes: Lessons learned from a \\$72K bill](#),” Medium, June 12, 2023.
 8. Owen Rogers, “[Reports of cloud decline have been greatly exaggerated](#),” Uptime Institute, Jan. 18, 2023.
 9. The Coca-Cola Company, “[The Coca-Cola Company and Microsoft announce five-year strategic partnership to accelerate cloud and generative AI initiatives](#),” press release, April 23, 2024.
 10. Flexera, “[2024 State of the cloud report](#),” 2024.
 11. Ibid.
 12. Nikhil Roychowdhury, Nik Jethi, Farhan Akram, and Rishabh Kochhar, “[Optimizing the value of cloud: A guide to getting started](#),” Deloitte, March 30, 2023.
 13. Amazon Web Services, “[TicketSwap tames demand ups and downs with AWS](#),” 2021.
 14. FinOps Foundation, “[Invoicing & chargeback](#),” accessed Nov. 4, 2024.
 15. For example: The Linux Foundation, “[FinOps across public cloud and on-prem](#),” accessed Nov. 4, 2024.
 16. Meredith Shubel, “[What is GreenOps? Putting a sustainable focus on FinOps](#),” The New Stack, Sept. 22, 2023.
 17. Magda Puzniak-Holford, Adithya Subramoni, and Simon Brennan, “[EU Corporate Sustainability Reporting Directive \(CSRD\) - Strategic and operational implications](#),” Deloitte, Sept. 8, 2023.
 18. Belle Lin, “[Airbnb details road map to lower cloud costs](#),” The Wall Street Journal, Nov. 7, 2022.
 19. James Ma, “[How Sky saved millions with Google Cloud](#),” Google Cloud Blog, July 19 2021.
 20. Angus Loten and Isabelle Bousquette, “[Amazon warns of weaker cloud sales as businesses cut spending](#),” The Wall Street Journal, April 13, 2023.
 21. Amazon Web Services, “[Lyft uses AWS Cost Management to cut costs by 40% in 6 months](#),” 2020.
 22. IBM, “[How the world’s largest ad company optimizes FinOps](#),” accessed Nov. 4, 2024.
 23. FinOps Foundation, “[FinOps Foundation Members](#),” accessed Nov. 4, 2024.
 24. Gartner, “[Gartner forecasts worldwide IT spending to grow 8% in 2024](#),” press release, Oct. 18, 2023.
-

Acknowledgements

The authors would like to thank **Nikhil Roy Chowdhury, Mitesh Gursahani, Nik Jethi, Rebecca Wood, Avishek Swain, Sophia Atkinson**, and **Vipul Mehta** for their contributions to this article.

Cover image by: **Jaime Austin**

生成式 AI 是否能復甦智慧型手機市場

透過專用晶片和行動作業系統的深度整合，智慧型手機或許不只是「Smart」，甚至能變得「Intelligent」。但使用者是否會接受這種全新模式？

智慧型手機已成為全球最廣泛使用的消費性科技產品。整合多種裝置，並使用先進且高度微型化的元件，已廣泛應用於下游消費市場與工業設備。隨時可用的便利性與實用性，重塑了人類行為模式與產業競爭格局。然而，近年智慧型手機的創新似乎未能讓市場振奮，給人的感覺更像是小幅改良，而非革命性的突破。

當前主要的行動生態系統供應商已著手改造次世代作業系統與先進晶片，目的是將生成式 AI 融入智慧型手機的體驗核心。³ 越來越多原始設備製造商（OEM）推出搭載生成式 AI 功能的智慧型手機，⁴ 供應商期待透過強大的 AI 功能提升產品吸引力，但可能亦伴隨潛在風險。

Deloitte 預測，2025 年的全球智慧型手機出貨量將微幅成長至 7% 左右，較 2024 年 5% 的年成長率略有提升。⁵ 部份成長可能是來自消費者升級至最新機種，進而重置標準裝置的升級週期。此外，早期採用者與開發人員對搭載支援生成式 AI 晶片的次世代智慧型手機抱有高度期待，也可能成為成長動能之一。Deloitte 進一步預測，截至 2025 年底，搭載生成式 AI 技術的智慧型手機出貨量將占超過 30%。⁶

生成式 AI 的前景令人充滿期待，但此項技術是否真的能讓美夢成真？使用者又是否能接受這種全新的裝置互動模式，仍有待觀察⁷。

生成式 AI 是否能推動智慧型手機下一波的升級潮？

短期內，頂尖的智慧型手機設計廠商可能將生成式 AI 整合視為刺激高階機種需求的策略。在 2024 年以前，智慧型手機的銷售量已連續兩年下滑，⁸ 部分原因是因市場趨於飽和，目前預估大約有 50 億人持有智慧型手機，占全球人口半數以上。⁹ 近年來，手機升級週期持續延長，消費者平均每兩至三年升級一次，且受通膨影響，可支配支出受到限制。¹⁰ 同時，有越來越多使用者選擇較高階機種，以延長裝置使用壽命，¹¹ 此趨勢不僅增加了對更高階硬體的需求，也帶來了在智慧型手機使用體驗上創造更價值和實用性的壓力。

2024 年第一季，受消費者信心回升及對搭載生成式 AI 高階機種的初步興趣所帶動，智慧型手機的銷售成長強勁。¹² Deloitte 在 2024 年發布的《互聯消費者研究》（Connected Consumer Survey）亦發現類似趨勢，

現在較少家庭因為預算問題而影響連網裝置的購買決策。¹³ 這波回升在歐洲同樣明顯，2024 年第二季智慧型手機的銷量在持續成長。¹⁴ 因此，預測 2025 年換機潮可能會加速回歸，將有更多消費者選擇升級手機，其中會有部分選擇高價位且搭載生成式 AI 功能的旗艦機種。

值得注意的是，雖然生成式 AI 可能會推動智慧型手機升級，但其影響力因市場和年齡層而異。研究顯示，7% 的美國受訪者表示生成式 AI 功能可能會促使他們提前升級智慧型手機，但此比例在 24 到 45 歲的受訪者中攀升至 50%，可能反映該年齡層對智慧型手機的高度依賴及新科技的接受度。¹⁵ 然而，Deloitte 英國在 2024 年《數位消費趨勢調查》(Digital Consumer Trends Survey) 中指出，僅有 4% 的英國受訪者表示每天都會使用生成式 AI，23% 的受訪者認為沒有幫助，19% 對生成式 AI 提供的答案感到不滿意。¹⁶

生成式 AI 是否能成為智慧型手機升級的推動力？將取決於此項技術能提供的價值與實用性，2025 年智慧型手機將對生成式 AI 的實際應用進行測試。

個人電腦中的生成式 AI

新一代 PC 搭載生成式 AI，必須考量使用者體驗、實用性、價值等因素，而這些考量也同樣影響者超大規模生成式 AI 演進及推廣趨勢。

Deloitte 2024 年發布的《互聯消費者調查》(Connected Consumer Survey) 顯示，消費者有興趣購買搭載生成式 AI 功能的個人電腦。34% 的美國受訪者計畫升級桌上型或筆記型電腦，並認同生成式 AI 晶片將提升其購買意願。Deloitte 認為，個人消費者將占年度電腦銷售量的 50%，因此可能成為影響市場的一項重要因素。¹⁷ 對企業買家而言，仍不確定哪一款具備生成式 AI 輔助處理器的個人電腦最適合業務，主因是各家電腦製造商提供的解決方案各不相同，價格也存在差異。¹⁸

隨著時間推移，預計大多數高階個人電腦將使用專用晶片來提供生成式 AI 功能，並且到了 2028 年約有 80% 的個人電腦都會搭載此類晶片。¹⁹ 另一項預估則指出，2024 年第二季具 AI 功能的裝置出貨量約為 900 萬台，但其中搭載足夠算力來執行生成式 AI 工作的神經網路處理器，數量仍無法確定。²⁰ 事實上，部分潛在客戶可能會選擇觀望，直至一年後或是出現效能更強大的機型後才會進行升級。

Deloitte 預測，2024 年全球銷售的 PC 中，約有 30% 具備本機生成式 AI 處理功能。²¹ 至 2025 年時，預計成長將近 50%。

雖然電腦市場規模不及智慧型手機市場（預估 2024 年電腦銷量約 2.6 億台，²² 而智慧型手機銷量約 12.3 億支）²³，電腦平均銷售價格較高，因此營收方面仍具有相當影響力。Deloitte 預估 2024 年電腦銷售總額大約為 2,200 億美元，而同年智慧型手機的銷售總額大約為 5,200 億美元。²⁴

目前尚不確定具備生成式 AI 功能的裝置將會對 PC 產業帶來何種影響。預計這些裝置可能會使電腦的平均銷售價格上漲，單一銷售價格增福約 15%。²⁵ 然而，2025 年的電腦銷量預計不會成長超過 10%。²⁶

對消費者而言，智慧型手機和個人電腦相關元件的技術進步，將會影響供應鏈並降低成本，使元件得以應用於更多設備，生成式 AI 功能預計在各類聯網裝置領域中變得更加普及。

生成式 AI 讓智慧型手機更智能

智慧型手機中的「Smart」通常指其具有聯網及應用程式運作的功能。生成式 AI 進一步提升個人化，使其能夠理解使用者的互動與需求，並透過對話強化整體運作。雖然過去的語音助理不一定符合使用期待，但部分使用者已開始與最新的大型語言模型互動。²⁷ 未來，這種對話式 AI 可能成為數位系統的新互動模式，甚至發展成一種值得信賴的智慧代理人，能夠學習並自主執行符合使用者需求的任務。

內建生成式 AI 的手機能透過理解使用者行事曆及所在地，推論出使用者的目的，提供在指定時間範圍內抵達目的地的最佳路徑，例如回答「我應該幾點出發才能準時抵達下午 2 點的會面？」這類問題。此類 AI 專為執行小範圍任務而設計，並透過神經網路處理器（預估至少每秒運算 30 兆次）²⁸ 以支援設備端推理功能。此外，該模型可進一步識別問題是否超出本機計算範圍，並適時將複雜任務指派給大型雲端模型處理，以提供更準確的答案。這種混合運算方式不僅能提升手機的即時性與安全性，還能靈活運用雲端資源，確保使用者獲得最準確的回應。²⁹

小型 AI 模型已能夠直接在裝置上直接運行，使用者數據可根據需求在本機上進行封存與保護，不過，這也需要極低延遲的功能才得以運作，例如即時翻譯，以提供更快速的回應。³⁰ 這些特性有利於提升使用者信任，並強化實用性。此外，服務供應商也可以從使用者互動資料中獲取新的數據循環，進而優化當地與雲端模型效能，提供最更精確的結果給使用者，並為企業帶來更深入的洞察。

更長遠的目標是讓智慧型手機（即消費者互動核心）變得更加個人化及智慧化，且能根據個人行為進行調整並預測需求。這種具備「代理式」功能的發展，可能會重塑智慧型手機的互動模式，並進一步推動裝置生態系統從「Smart」邁向「Intelligent」。（詳情請參閱 2025 年《高科技、媒體及電信產業趨勢預測》（TMT Predictions）中與生成式 AI 有關的內容。）

未來一年將逐步揭露使用者對此項全新體驗的接受程度，以及早期生成式 AI 功能的價值和可理解性。預計各大服務供應商在未來數月內將陸續推出新功能，但他們也認為應用的普及可能仍需一段時間。³¹ 未來一年，可能會優先評估小型 AI 模型在設備端運行的能力與侷限性，這可能會改變生成式 AI 的經濟模式。如有更多生成式 AI 工作從昂貴的資料中心轉移至消費者的裝置上，生成式 AI 的基礎建設資本支出可能會有所緩解。

業界是否能透過投資以度過生成式 AI 的炒作熱潮？

目前市場正面臨巨大壓力，一方面業界需要證明最先進 AI 模型的市場適用性，以合理化其高昂的成本；另一方面，也必須讓這些模型在開發與運作上更具成本效益。³² 大型模型供應商已投入數十億美元開發 AI 模型，並持續投資數十億美元以擴建其認為能滿足大規模需求的資料中心。³³ 根據估計，每年有 6,000 億美元投入支援生成式 AI。³⁴ 然而，這樣的資本密集投資不可能無限延續，最終仍需創造實質的經濟價值，因此更精確的產品市場契合度將成為關鍵。

為了降低成本，業界可能會採取幾種方式，例如縮小模型規模、減少所需訓練數據量，或根據不同工作負載需求來拆解模型，這些做法尤其適用於推論任務。許多消費者與企業應用場景來說，生成式 AI 可能會滲透到許多日常應用，而更便宜、更節能的小型 AI 模型可能能夠補足或取代某些需求。

然而，目前仍不清楚推論任務會有多少能夠維持在裝置端執行。現今生成式 AI 的互動方式與使用者期待大多數是由公共雲模型定義。使用者可能需要時間來理解哪些任務和指令可以在本地裝置執行，哪些則需要透過網路傳輸至雲端模型處理。這種結合對話能力、能在設備端運行並支援雲端的全新互動模式，對於其使用率和使用者的行為的影響仍待進一步觀察。

廣泛使用生成式 AI 仍面臨挑戰

2024 年 Deloitte 的《互聯消費者調查》(Connected Consumer Survey)，研究顯示 38% 的美國受訪者曾經使用生成式 AI，其中 63% 表示此項技術超出預期。³⁵ 許多使用者已體驗到生成式 AI 的強大能力，但是服務供應商仍需擴大向更多的受眾證明其實用性，才能說服消費者為了 AI 功能購買新款智慧型手機。

在智慧型手機上使用生成式 AI 可能會讓部分使用者有所困惑，因為需要適應新的互動方式。例如當 AI 助理試圖管理使用者的行事曆時，他們可能會有所疑慮，不願意完全讓 AI 介入決策。³⁶ AI 普及應用可能帶來一系列權衡取捨，包含電池消耗量增加、內建公有模型產生的額外費用，以及潛在的錯誤資訊，這些因素可能削弱高價值應用場景的可靠性。在使用者、個人代理型 AI 和公有雲模型之間建立信任可能需要時間，而一旦失去信任，則可能極為困難修復。

服務供應商希望下一代前沿模型 (Frontier Models) 能釋放更大的價值，但目前仍不確定其能力是否會持續成長，或是否將趨於平緩或下降。此外，是否具有足夠的資料以滿足日漸增加的訓練需求也是關鍵。³⁷ 使用模型自行生成的合成資料與進行訓練可能會帶來風險，導致推理品質隨時間逐漸下降。³⁸ 另一個核心問題是，是否可以在不大量增加資料、訓練與推理成本的情況下持續進步？是否存在一個時機，能提升功能同時降低資本和資料需求？投資人可能會變得更加謹慎，在技術真正能帶來足夠收益之前，就要求企業交出更高的營收成績。

監管機關可能會因為採取更全面的策略來規範生成式 AI 發展，以防範新興風險，例如深偽技術、錯誤訊息和高度擬人化且具說服力的聊天機器人。這類對話式機器人能與使用者建立更和諧且深入的連結，進一步影響其想法和意識形態。³⁹ 個人化對話助理可能有助於深入人際互動的核心領域，來幫助更多人，但也伴隨著成癮風險。⁴⁰ 另一方面，將設備端生成式 AI 與第三方模型結合，可能會擴大資安漏洞和濫用風險。⁴¹ 這將會促使服務供應商加強其生態系統的安全性，以及監管機關制定更多防護機制。

結論

儘管市場對「下一代智慧型手機」一種能夠顛覆市場並帶來全新變革的消費型裝置平台，保持關注和討論度，但目前仍尚未有具體產品問世。智慧型手機擁有數十億使用者，仍處於主導地位，並持續為測試新服務和互動方式提供大規模的測試平台。到了 2025 年，可能會因為隨著高階智慧型手機和個人電腦的發展大幅增加，與生成式 AI 互動的人數隨之上升。

服務供應商可以從中深入研究生成式 AI 的價值，並探索其極限，如果技術成功落地，智慧型手機將會變得更具有吸引力，並有助於提升平台整合能力，開啟全新應用和市場機會，為個人裝置帶來新一波的發展。但這一轉變需要時間，預計未來一年更可能是用來幫助使用者適應這場個人運算新變革的「導入期」。

未來幾年，智慧型手機作業系統可能會整合更多互動方式，例如新一代的對話式搜尋，提供更多本地化摘要而非遠端連結，從而減少對服務供應商和資訊來源的依賴。

若使用者使用更個人化的代理式 AI 時，數位互動的模式將會改變，許多工作將被移交給裝置上的 AI 自動處理，而不需要使用者親自操作介面。於此情形下，運算作業將變得更具環境感知能力，於背景自動運行並逐步發展為更加空間化的互動模式，能夠感知周遭環境與網路互動，提升使用者體驗。

然而，隨著服務供應商在積極推動市場需求，可能需與同時面臨經濟壓力，必須尋求降低訓練與運行大規模 AI 模型帶來的高昂資本投入與能源成本。業界可能探索更小型化的 AI 模型、使用混合式架構，並更深入瞭解生成式 AI 工作負載對運算資源的需求。在當前氣候變遷帶來無法忽視的不確定性與焦慮下，生成式 AI 資料中心的擴建正在增加能源和水資源的消耗，並加重了家庭與地方政府所能承擔的能源成本。⁴² 即使生成式 AI 能克服經濟壓力，它最終仍可能受到能源成本的限制。

截至 2024 年底，超大規模企業（Hyperscalers）、智慧型手機生態系業者，以及新興 AI 模型供應商致力於發展生成式 AI，期待提供的技術效益能轉化為更大的經濟價值。但究竟能從中獲取多少價值？生成式 AI 領域的超大規模企業（Hyperscalers）是否會重蹈電信業與早期互聯網的覆轍，投入巨資建設基礎設施，最終卻只是幫助下一代創新者鋪路？⁴³

推動生成式 AI 的崛起、部署及廣泛應用，可能是自網際網路普及以來最大規模的科技實驗之一。無論最終結果如何，這場如登月計畫般的冒險，勢必將帶來一波新的技術浪潮，並改變人們的行為模式與商業生態。

By **Chris Arkenberg**
United States

Duncan Stewart
Canada

Gillian Crossan
United States

Kevin Westcott
United States

Endnotes

1. GSM Association, [“Smartphone owners are now the global majority, new GSMA report reveals,”](#) press release, Oct. 11, 2023.
2. Wolfgang Bock, François Candelon, Steve Chai, Ethan Choi, John Corwin, Sebastian DiGrande, Rishab Gulshan, David Michael, and Antonio Varas, [“The mobile revolution: How mobile technologies drive a trillion-dollar impact,”](#) Boston Consulting Group, Jan. 15, 2015.
3. IDC Corporate, [“The future of next-gen AI smartphones,”](#) Feb. 19, 2024.
4. Counterpoint, [“Gen AI-capable smartphone shipments to grow over 4x by 2027,”](#) April 16, 2024.
5. IDC Corporate, [“Worldwide smartphone market up 7.8% in the first quarter of 2024 as Samsung moves back into the top position, according to IDC tracker,”](#) press release, April 15, 2024.
6. IDC anticipates a 364% compound annual growth rate in 2024 (from a low base in 2023) for global gen AI smartphone shipments, with 73% growth in 2025. Canalys expects AI-enabled smartphone market share to reach 54% by 2028. Our analysis, for reasons outlined in this paper, is less bullish than the former, and a bit more than the latter. Sources: IDC Corporate, [“The future of next-gen AI smartphones”](#); Canalys, [“Now and next for AI-capable smartphones,”](#) accessed Oct. 30, 2024.
7. Jim Fellingner, [“CTA study: Smartphones most-owned tech, 5G and wireless drive adoption,”](#) press release, Consumer Technology Association, May 31, 2023.
8. IDC Corporate, [“Worldwide smartphone market up 7.8% in the first quarter of 2024 as Samsung moves back into the top position, according to IDC tracker.”](#)
9. GSM Association, [“Smartphone owners are now the global majority, new GSMA report reveals.”](#)
10. Sarah Barry James, [“Consumer checkup: Higher interest rates lead to longer tech replacement cycles,”](#) S&P Global, March 26, 2024.
11. IDC Corporate, [“Worldwide smartphone market up 7.8% in the first quarter of 2024 as Samsung moves back into the top position, according to IDC tracker.”](#)
12. Chris Donkin, [“Smartphone sales up again ahead of expected gen AI boost,”](#) Mobile World Live, July 15, 2024.
13. Susanne Hupfer, Michael Steinhart et al., “2024 Connected Consumer Study,” Deloitte Insights, publication forthcoming, 2024.
14. Counterpoint, [“Europe smartphone market recovery continues, shipments up 10% YoY in Q2 2024,”](#) Aug. 28, 2024.
15. Susanne Hupfer, Michael Steinhart et al., “2024 Connected Consumer Study,” Deloitte Insights, publication forthcoming, 2024.
16. Deloitte, [“Generative AI: 7 million workers and counting,”](#) June 25, 2024.
17. The installed base of PCs is estimated to be about 2 billion, and there are about 1 billion knowledge workers, suggesting that the market is roughly half consumer and half enterprise.
18. Author interviews with enterprise chief information officers in July and August 2024.
19. Canalys, [“AI-capable PCs forecast to make up 40% of global PC shipments in 2025,”](#) March 18, 2024.
20. Ibid.
21. Deloitte Global analysis of publicly available information for H1 2024, and extrapolation based on usual PC seasonality trends.
22. IDC Corporate, [“PC refresh cycle and tablets in emerging markets expected to spur demand in coming quarters, according to IDC,”](#) press release, Sept. 23, 2024.

23. IDC Corporate, “Worldwide smartphone market forecast to grow nearly 6% in 2024, driven by stronger growth for android in China and emerging markets, according to IDC,” press release, Aug. 27, 2024.
24. Based on quarterly data so far in 2024, Deloitte believes smartphone average selling price is declining and should be roughly US\$425 for the year. PC average selling prices were high during the 2021 chip shortage, but are declining and Deloitte estimates them to be about US\$850 for 2024.
25. Roshan Ashraf Shaikh, “Analysts expect 15% price hike for AI PCs—60% of PCs will have local AI capabilities by 2027,” Tom’s Hardware, April 26, 2024.
26. IDC Corporate, “PC refresh cycle and tablets in emerging markets expected to spur demand in coming quarters, according to IDC.”
27. Sigal Samuel, “People are falling in love with—and getting addicted to—AI voices,” Vox, Aug. 18, 2024.
28. IDC, “The future of next-gen AI smartphones.”
29. Baris Sarer, Mark Szarka, Nataliia Bacchus, and Edem Isliamov, “The world of hybrid AI,” The Wall Street Journal and Deloitte, July 31, 2024.
30. Malik Saadi, “On-device generative AI unlocks true smartphone and PC value,” Forbes, April 17, 2024.
31. Lisa Eadicicco, “AI is changing our phones, and it’s just getting started,” CNET, April 3, 2024.
32. Goldman Sachs, “Gen AI: Too much spend, too little benefit?” June 27, 2024.
33. David Cahn, “AI’s US\$600B question,” Sequoia, June 20, 2024.
34. Ibid.
35. Susanne Hupfer, Michael Steinhart et al., “2024 Connected Consumer Study,” Deloitte Insights, publication forthcoming, 2024.
36. Jon Victor, “Software firms race to beat OpenAI in AI agents,” The Information, Sept. 26, 2024.
37. Deepa Seetharaman, “For data-guzzling AI companies, the internet is too small,” The Wall Street Journal, April 1, 2024.
38. Michael Peel, “The problem of ‘model collapse’: How a lack of human data limits AI progress,” Financial Times, July 24, 2024.
39. Yuval Noah Harari, “Yuval Noah Harari argues that AI has hacked the operating system of human civilization,” The Economist, April 28, 2023.
40. CBS News, “Virtual valentine: People are turning to AI in search of emotional connections,” Feb. 14, 2024.
41. Matt Burgess, “Generative AI’s biggest security flaw is not easy to fix,” Wired, Sept. 6, 2023.
42. Camilla Hodgson, “US tech groups’ water consumption soars in ‘data center alley,’” Financial Times, Aug. 17, 2024.
43. Bryce Elder, “Gen-AI revisited, by Goldman Sachs,” Financial Times, Sept. 5, 2024.

Acknowledgements

Authors would like to thank **Rohan Gupta** and **Steve Fineberg**.

Cover image by: **Jaime Austin**; Getty Images, Adobe Stock

大型影視製作公司使用生成式 AI 進行內容創作時會比較謹慎，而社群媒體則積極使用

好萊塢和相關影視產業在內容創作方面對生成式 AI 保持謹慎態度，但在應用和營運層面上預計會更快使用此技術。

生成式 AI 在影像、音訊與影片領域中日漸成熟，能夠創造更逼真且富有創意的內容，且在長時間使用之後，將會變得更具可控性。各大影視製作公司積極嘗試生成式 AI 內容創作，但對於全面應用仍持謹慎態度，其中原因包括相關工具未成熟，以及當前公有 AI 模型的限制，可能將帶來法律風險，甚至影響到其智慧財產權。然而，產業內部普遍認為，若能有效應用生成式 AI 於企業各環節，將有助於降低成本並提升獲利能力。

事實上，影視製作公司正面臨成本壓力，且僅有極少數的公司可以獲利。¹ 雖然收益可觀，但在營運、製作、行銷與廣告成本不斷上升下，² 許多經營串流服務的公司目前仍未能從該業務獲利。且隨著有線電視訂閱人數下降與廣告收入減少，營收也受到進一步衝擊。在通膨、利率上升和疫情的影響下，成本亦進一步增加。此外，影視製作公司還需要與社群媒體、使用者創作內容（User-Generated Content，UGC）和遊戲業者競爭，以吸引消費者的目光與購買。

Deloitte 預測，2025 年全球最大的電視與電影製作公司，特別是美國與歐盟的業者，在使用生成式 AI 進行內容創作方面將會保持謹慎態度，僅預計投入不超過 3% 的製作預算。³ 在營運成本方面，約有 7% 的支出將會轉向新興的生成式 AI 工具，以支援合約和人才管理、許可與規劃、行銷和廣告，以及內容在地化與配音，以拓展全球多元市場。

此舉有助於製在公司降低生成式 AI 在人才與內容方面帶來的衝擊，同時加速使用 AI 工具能降低成本並提升效益。然而，相較獨立內容創作者和社群媒體平台正在迅速將生成式 AI 納入其工作流程和內容創作中，將可推動新型態的媒體發展，吸引觀眾注意力，並進一步削弱傳統影視製作公司的競爭優勢。⁴

對於好萊塢級別的內容創作來說，生成式 AI 工具仍尚未成熟

現成的大型語言模型和擴散模型（Diffusion Models）價格低廉，使影視製作公司能更快速地嘗試劇本、對話和故事元素的開發，並在早期角色與場景的設計進行視覺化探索。⁵ 部分工作室已運用生成式 AI 技術讓明星「逆齡」或創建數位分身，以利商業廣告或已故演員的作品使用。⁶ 為降低潛在的法律風險，製片公司可透過在合約中加入 AI 相關保護條款，未來也預計將會看到更多第三方製作團隊向影視業者提供相關服務，以滿足市場需求。

雖然使用生成式 AI 進行內容創作，可以在前期製作提升創意發揮空間，但目前仍無法達到好萊塢級別的製作水準。⁷ 現今最強大的視覺擴散模型雖已能生成擬真度極高的影像，但輸出結果往往有著過於超現實的「詭異感」。⁸ 當前領先的影片生成模型已能生成短影片，但仍無法製作較長及連貫的內容。⁹ 儘管技術正在快速進步，但仍需要時間才能成熟，並順利整合現有工具和製作流程中。

不過，這些限制對社群媒體創作者來說或許並無影響，因為他們追求內容的快速產出和發布。短影音已成為趨勢，但情況正在改變。¹⁰ 由於影片時長較短，法律責任風險相對較低。部份生成式 AI 工具的早期採用者經常在社群媒體上定期分享實驗成果，展現從第三方解決方案在影片生成上的快速發展。¹¹

未來一年，獨立創作者很有可能率先使用生成式 AI 進行內容創作，這有助於製作公司觀察技術發展，並降低過早投入的風險。這也可能使觀眾的注意力進一步轉向使用者創作內容平台，使其更具有與傳統媒體競爭的優勢。

大型影視製作公司可能會擔心公有 AI 模型帶來的法律責任風險

大型影視製作公司擔憂生成式 AI 內容工具將帶來智慧財產權和法律責任的風險，或使其自創內容無法被認定為原創。¹² 一些最強大的公有 AI 模型是基於公開資料進行訓練的，其中可能包含其他創作者的影像和影片，使產出內容具有衍生性。¹³ 若製作公司將這些模型輸出的內容用於營利，而此模型的訓練集含有其他藝術家的受保護作品，則製作公司可能會面臨侵權責任。由於訓練集內涵蓋數十億件作品，舉證具備一定難度。這種不確定性已足以使依賴智慧財產權的製作公司望之卻步。目前獨立藝術家、創作者、出版商¹⁵ 和音樂公司¹⁶，已開始針對公有 AI 模型提起訴訟，指出其在訓練集內的作品遭到侵權。¹⁴

公有 AI 模型還可能使製作公司難以保護自身的智慧財產權。根據美國《著作權法》(Copyright Act)，只有在作品須具備足夠的「人類創作」，才能授予著作權。¹⁷ 美國著作權局 (US Copyright Office) 在最近的討論中，認為在包含 AI 或生成式技術的作品中，人類創作的比例會因個案而異，只要符合「足夠人類創作」的標準，就仍然可以獲得著作權。也就是說，影視製作公司可以為使用生成式 AI 工具輔助創作的作品申請著作權，但前提是 AI 只是輔助，不能是主要創作者。此議題目前仍在討論中，但始終缺乏明確定義，讓業界面臨著不確定性與風險。

為了擴充資料訓練模型，技術領先的生成式 AI 供應商積極尋求與製作公司合作，並期待推動他們授權其內容資料庫。¹⁸ 不過，製作公司可能會完全拒絕此提議，因為智慧財產權是企業核心資產，或因本身的營運成本壓力，而要求生成式 AI 公司索取高額授權費。製作公司甚至可能會選擇聯合拒絕提供數據，來限制前沿模型的文字、音訊、影像和影片的生成發展。

此外，影視製作公司（尤其是美國的製作公司）必須與同業公會和工會合作，而這些組織強烈反對使用生成式 AI，並已獲得製作公司保證限制生成式 AI 的使用範圍。¹⁹ 類似的勞工抵制聲浪也出現在英國²⁰ 和歐盟，在這些地區，製作公司必須遵守歐盟《人工智慧法案》(AI Act) 等法規，確保模型安全性，以及需遵守《一般資料保護規則》(GDPR)，規範 AI 訓練所需數據的蒐集與存儲方式。這些法律與勞工議題進一步增加了影視產業在 AI 應用上的複雜性與挑戰。²¹

完全私有化模型的成本對製作公司可能過高

Deloitte 2024 年《高科技、媒體及電信產業趨勢預測》(TMT Predictions) 中，探討了私有生成式 AI 的興起，認為其可避免部分公有 AI 模型的風險，並對輸出結果擁有更大的控制權。²² 製作公司可自行訓練 AI 模型並使用自家 IP 作為訓練數據，降低法律責任與著作權爭議。

然而，訓練生成式 AI 模型的成本極為昂貴，預估訓練一組前沿模型需要約 1,000 億美元²³，且推理和再訓練的成本可能也隨著使用量而增加。開源解決方案（更準確的稱為「開放權重 (open weight)」）雖可以降低部分成本，但訓練數據通常不夠透明，且整體成本仍居高不下。²⁴ 製作公司要吸引高階技術人才來開發這類模型也極具挑戰，因為這些專業人才更可能選擇薪資更優渥的大型科技公司。此外，由於 AI 模型開發的進展快速，現今的投資模型可能在六個月內即需更新。因此若要建立更具競爭力的私有模型，製作公司和投資者須採取科技公司的思維，並與科技供應商建立生態系統，並支付相關費用。基於上述原因，在未有明顯 AI 經濟變革前，製作公司較難自行訓練模型。

未來一年，製作公司與 AI 供應商間可能會掀起一波合作浪潮，以更公平地分擔成本。²⁵ 在合作中，第三方提供預先訓練好的模型與介面，再透過製作公司進一步訓練和客製化，使模型能生成符合其公司美學風格的內容，包括加入其標誌性角色和場景設計。此外，製作公司也可透過展示內容的衍生性來控制潛在的智慧財產權風險，但 AI 供應商仍需持續提升工具的成熟度與輸出品質，以滿足製作公司的需求。

使用生成式 AI 優化製作公司業務

在未來的一年，製作公司將預計開始嘗試生成式 AI 內容創作，並加速理解其如何推動及優化其他業務領域。生成式 AI 可協助自動化並強化合約談判、人才及勞動力管理、財務與會計及媒體營運，例如內容在地化、行銷推廣、儲存與發行等。

一些製作公司可能透過現有軟體和 SaaS（軟體即服務）解決方案來整合生成式 AI 功能。新創公司也正在崛起，專注於前期製作中耗時且昂貴的部分。生成式 AI 可加速劇本評估，拆解並分配至製作時間表中，甚至能提前勘查潛在拍攝地點。²⁶ 生成式 AI 亦可解鎖舊有檔案內容，例如透過「觀看」舊影片並紀錄演員、主題和情緒等元素，以此協助串流平台能靈活地重新使用舊有內容，以滿足個人化推薦或熱門時段的需求，進而提升獲利。²⁷

為了加速和擴大內容傳播，一些製作公司正在利用語言及語音模型提升翻譯與配音的效果，使作品能夠觸及更廣大的國際觀眾。²⁸ 這些工具可實現高度自然、富有表現力與情感層次的語音，並允許使用者進行調整。²⁹ 對全球市場的內容創作者和發行商來說，無論是輸出內容至全球市場，或引進海外市場內容，無疑是一大助力。而領先的使用者創作內容 (UGC) 平台，也正將這些 AI 語音技術擴散到其創作者群體中。³⁰

生成式 AI 配音和翻譯技術能促進文化交流，讓原本僅在當地受歡迎的熱門作品成為全球熱門話題。Deloitte 2024 年《數位消費者行為趨勢》(Digital Consumer Trends) 調查顯示，66% 的美國受訪者喜歡觀看能幫助他們了解不同文化的電視節目或電影。³¹ 生成式 AI 不僅有助於媒體公司提升獲利與競爭力，還能拉近全球觀眾之間的距離。

結論

如同大多數企業一般，製作公司、串流平台和創作人才都對生成式 AI 感到期待又充滿擔憂。對於計畫未來一年將探索生成式 AI 內容創作的製作公司來說，主要動力可能來自其創意潛力——前沿模型將人類創意進行重組與再造，生成全新型態內容。同時，製作公司也擔憂此類技術可能催生出好萊塢生態體制以外的新型態媒體形式。

愈來愈多的獨立創作者正在展示最新的合成媒體技術所能達成的效果，而這些技術正在快速進入市場。過去，好萊塢製作公司曾享有掌控內容與發行資源的稀缺性，但如今資源變得豐富且分配趨於公平，³² 在未來內容顛覆的威脅將會進一步加劇。

幾乎每個月，前沿模型都有新進展，不斷推動 AI 並邁向更接近人類的智慧、創造力和洞察力的發展。約一年前，人們普遍認為至少要到 2030 年，才可能出現幾乎完全由 AI 生成的賣座大片，³³ 如今這個遠大的目標似乎變得更可實現了。

同時，內容擁有者可能會加強其智慧財產權的競爭壁壘，對公有 AI 模型發起更多的訴訟和監管，以應對潛在的著作權侵權行為，監管機關將可能會要求主要模型供應商證明其訓練數據未侵犯現有內容權利。大多數大型製作公司可能會拒絕將其內容目錄授權納入其公有數據集的提議。若經濟條件允許，他們可能更傾向與能夠針對影視產業需求，量身打造模型並提供智慧財產權保護的小型公司合作。

從宏觀的角度來看，生成式 AI 需要投入大量資本，若無法在未來一年內找到實現廣泛經濟價值的途徑，則成長速度可能會減緩。³⁴（詳請參閱今年的《高科技、媒體及電信產業趨勢預測》（TMT Predictions）中與設備端生成式 AI 有關的內容）但若新一代的前沿模型能夠克服現有挑戰，技術能力將會迅速提升，很可能進一步降低訓練和運行成本，並減少 AI 需要的大規模數據量。

大型製作公司作為企業，可能會更關注如何運用生成式 AI 降低成本、優化業務、提升生產力並擴大加速觸及客戶的能力。根據 Deloitte 2024 年《企業中的生成式 AI 現狀》（State of Generative AI in the Enterprise）調查，42% 受訪的高階主管表示，AI 最重要的效益是提升效率、生產力和成本降低；58% 的受訪者則提到，AI 還能帶來創新提升、改善產品服務及強化客戶關係等。³⁵ 由此可見，企業對於將這些功能應用於現代業務中似乎越來越感興趣。

未來，生成式 AI 似乎有望協助更多小型企業和創作者，達到過去僅有大型企業才能實現的生產力及品質水準。較小的製作公司和獨立創作者的能力將獲得大幅提升，同時能避免大型製作公司面臨高額成本與風險。相較之下，最大型的製作公司必須進一步降低其成本，並加速內容上市速度，以維持競爭力。除此之外，也要面對包括使用者生成內容平台、社群媒體及遊戲產業的競爭。在內容生產與發行變得不再稀缺的時代，真正稀缺的資源，是觀眾的注意力。

By **Chris Arkenberg**
United States

Ricky Franks
United States

Danny Ledger
United States

Kevin Westcott
United States

Endnotes

1. George Szalai, [“Studio profit report: A year of major transition,”](#) The Hollywood Reporter, April 24, 2024.
2. George Szalai, [“Studio profit report: Disney dives as Sony soars, Paramount rises,”](#) The Hollywood Reporter, Feb. 24, 2024.
3. This prediction is based on our analysis of earnings reports from leading streaming video providers and other available industry information.
4. Chris Arkenberg, [“Will generative AI challenge authenticity in social media?,”](#) Deloitte Insights, Dec. 8, 2023.
5. Hannah Murphy, [“Media groups look to AI tools to cut costs and complement storytelling,”](#) Financial Times, March 26, 2024.
6. David Smith, [“‘We’re going through a big revolution’: How AI is de-ageing stars on screen,”](#) The Guardian, Feb. 6, 2023.
7. Ibid; Murphy, [“Media groups look to AI tools to cut costs and complement storytelling.”](#)
8. Alon Yaar, [“What’s next for AI video generation,”](#) AI Business, Aug. 6, 2024.
9. Lauren Leffer, [“Everything to know about OpenAI’s new text-to-video generator, Sora,”](#) Scientific American, March 4, 2024.
10. Taylor Lorenz, [“The ‘Beastification of YouTube’ may be coming to an end,”](#) The Washington Post, March 30, 2024,
11. Dennis Ortiz and Kenny Gold, [“Gen AI and the creator economy: How creators are looking to leverage AI and what this means for brands,”](#) Deloitte, accessed Oct. 30, 2024.
12. Paul Sweeting, [“Hollywood’s AI concerns present new and complex challenges for legal eagles to untangle,”](#) Variety, April 17, 2024.
13. Jennifer Wolfe, [“What would have to happen for gen AI to take over Hollywood? So glad you asked.,”](#) NAB Amplify, July 5, 2024.
14. Ibid; Sweeting, [“Hollywood’s AI Concerns Present New and Complex Challenges for Legal Eagles to Untangle.”](#)
15. Baker & Hostetler, [“Case tracker: Artificial intelligence, copyrights and class actions,”](#) accessed Oct. 30, 2024.
16. Natalie Sherman, [“World’s biggest music labels sue over AI copyright,”](#) BBC News, June 25, 2024.
17. US Copyright Office, [“Copyright and artificial intelligence,”](#) March 16, 2023.
18. Lucas Shaw, [“Alphabet, Meta offer millions to partner With Hollywood on AI,”](#) Bloomberg, May 23, 2024.
19. Erin Degregorio, [“Hollywood is back to work after strikes, but AI remains in the spotlight,”](#) Fordham Law News, Jan. 29, 2024.
20. Daniel Thomas and Cristina Criddle, [“UK shelves proposed AI copyright code in blow to creative industries,”](#) Financial Times, Feb. 4, 2024.
21. Lindsey Wilkinson, [“EU passes AI Act, places first binding rules on generative AI,”](#) CIO Dive, March 13, 2024.
22. Chris Arkenberg, Baris Sarer, Gillian Crossan, and Rohan Gupta, [“Taking control: Generative AI trains on private, enterprise data,”](#) Deloitte Insights, Nov. 29, 2023.
23. Jowi Morales, [“AI models that cost \\$1 billion to train are underway, \\$100 billion models coming — largest current models take ‘only’ \\$100 million to train: Anthropic CEO,”](#) Tom’s Hardware, July 7, 2024.
24. Red Hat, [“What is an open-source LLM?,”](#) July 1, 2024.
25. Kyle Wiggers, [“Generative AI startup Runway inks deal with a major Hollywood studio,”](#) TechCrunch, Sept. 18, 2024.
26. Lauren Forrestal, [“Filmstage leverages AI to break down film scripts, create shooting schedules and more,”](#) TechCrunch, March 20, 2023; Lauren Forrestal, [“Avail rolls out its AI summarization tool to help Hollywood execs keep up with script coverage,”](#) TechCrunch, Dec. 7, 2023.

27. Emma Cosgrove, "[Nvidia, Amazon, Microsoft, and Paramount execs discuss the use of AI in Hollywood. Here are 9 startups they're watching.](#)," Business Insider, July 24, 2024.
 28. The Economist, "[The dawn of the omnistar](#)," Nov. 9, 2023.
 29. Audrey Shomer, "[The state of generative AI in Hollywood: A special report](#)," Variety, June 3, 2024.
 30. Andrew Hutchinson, "[YouTube announces expansion of auto-dubbing to more creators and languages](#)," SocialMediaToday, Sept. 19, 2024; Julia Walker, "[How Meta's AI dubbing breaks down language barriers](#)," PR Week, Sept. 30, 2024.
 31. Jana Arbanas, Jeff Loucks, Brooke Auxier, Kevin Westcott, Chris Arkenberg, and Bree Matheson, [2024 Digital Media Trends](#), Deloitte Insights, March 20, 2024.
 32. Michael D. Smith, "[Lessons from Hollywood's digital transformation](#)," Harvard Business Review, Dec. 16, 2021.
 33. Jackie Wiles, "[Beyond ChatGPT: The future of generative AI for enterprises](#)," Gartner, Jan. 26, 2023.
 34. David Cahn, "[AI's \\$600 billion question](#)," Sequoia, June 20, 2024.
 35. Deloitte, [The State of Generative AI in the Enterprise](#)—Moving from potential to performance, June 2024.
-

Acknowledgements

Authors would like to thank **Howie Stein** and **Ankit Dhameja**.

Cover image by: **Jaime Austin**; Getty Images, Adobe Stock

無線電信業者在監管機關允許的條件下 加速整合

在許多市場中，小型無線電信業者面臨成長緩慢、利潤低迷，以及償還債務的困境。在監管機關批准的情況下，企業併購可能會有幫助，尤其是資產合併及整體直接面對客戶的企業。

Deloitte 預測，2025 年之後市場上將會有更多電信合併案獲得批准，尤其歐盟可能會率先開始。¹ 許多地區和國家的無線電信市場非常分散，且部分是小規模公司。歷年來監管機關著重於維持市場競爭，鼓勵電信企業廣泛設立，以確保價格維持在較低水準。不過近年有越來越多的監管機關顧問建議，透過核准同業合併，或許可以進一步加強網路的成長、功能、安全性和韌性。

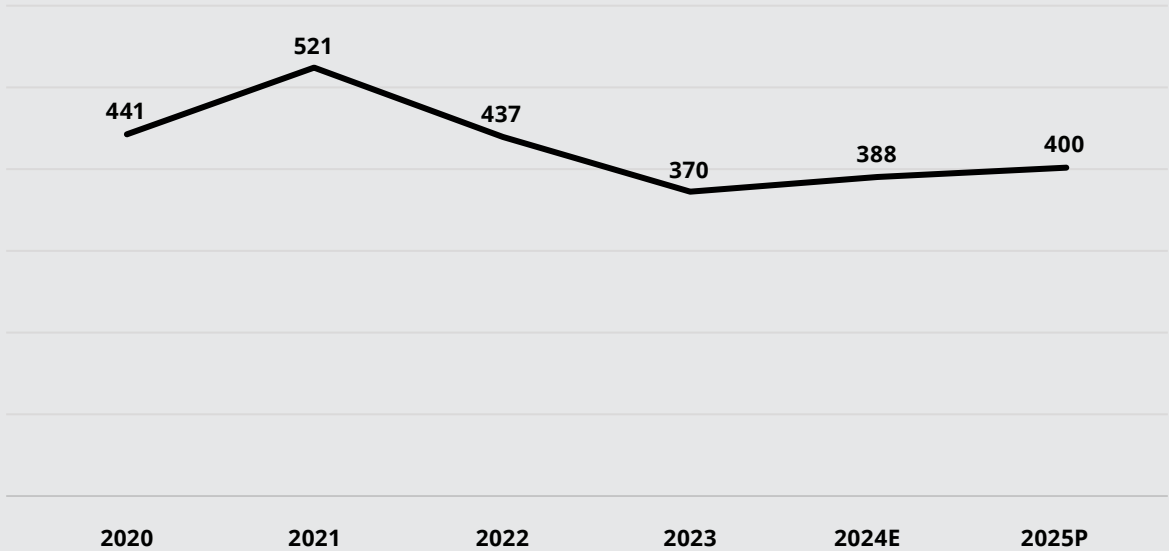
Deloitte 預測，2025 年全球將會有大約 400 件電信併購交易，相當於過去五年的交易量（圖 1）。² 或許這個數字本身並不特別引人注目，真正值得關注的是併購案的類型，未來將有更多涉及實質電信業者整併的交易。

電信併購交易有許多種類，整體而言沒有單一類型占有多數（圖 2）。³ 儘管近期資料中心的交易量有所增加，可能與 AI 資料中心的發展活動有關，但無論是無線電信或有線電信交易，各種交易類型隨時間推移呈現相對穩定。⁴

圖 1

2025 年的電信企業併購量，將與過去五年一致

平均交易量



備註：E 代表當前年度的預估值，P 代表 Deloitte Global 預測的數值。

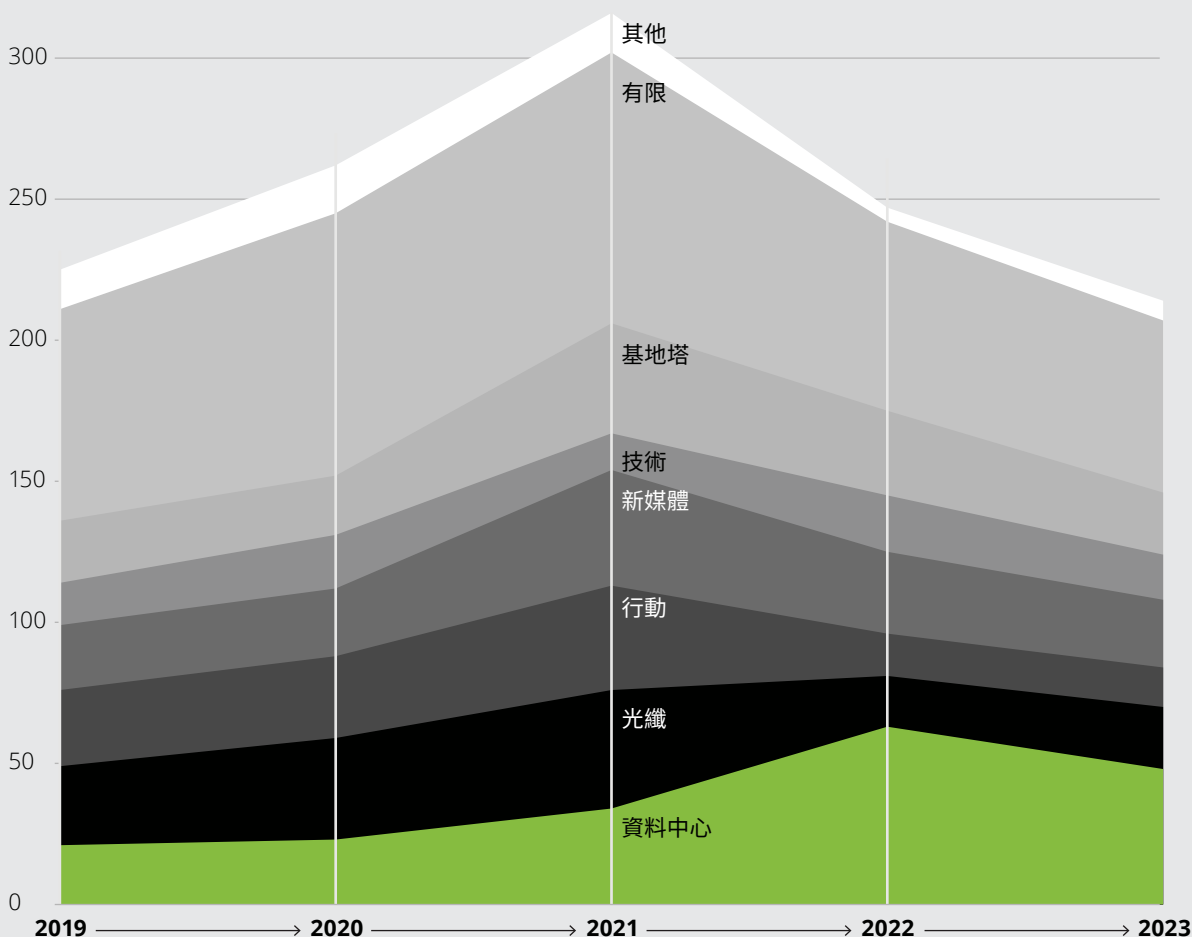
資料來源：此圖示是由 Deloitte 根據標普全球市場智財 (S&P Global Market Intelligence) — S&P Capital IQ 及 CB Insights。上述圖表是計算電信業每年的併購平均交易數量資料包含 2020、2021、2022 和 2023 年的全年數量，以及 Deloitte 根據 2024 年之部分數據進行的預測。

deloitte.com/insights

圖 2

多元化的併購活動涵蓋數個電信子行業，其中的資料中心交易在近期出現明顯成長

2019 至 2023 年通訊服務業者的併購交易量（每年交易量）



資料來源：CSI Magazine 刊載之 Omdia 調查結果，2024 年 6 月 23 日。

deloitte.com/insights

形式的整合或拆分已進行多年，可能已接近成長階段的尾聲。例如，截至 2023 年，97% 位於美國和墨西哥的行動通訊基地是由獨立基地台公司營運，而非電信業者（相比 2016 年時為 65%，有明顯提升），而歐洲截至 2023 年，基地台公司佔比已超越 2016 年的 36%，成長將近兩倍，達到 70%。⁵

同樣地，多年來，有線網路（如銅線、光纖與同軸電纜）和後端軟體系統（如計費、營運系統、現場服務機組、資料中心等）也持續在進行整合與併購，⁶ 各種類型的無線通訊公司整合皆在逐步增加。

以無線網路整合為例，加拿大的兩大無線電信業者自 2009 年起便共享無線存取網路（RAN）；⁷ 馬來西亞原先有三個獨立無線網路，其政府在 2021 年決定建立單一國家及 5G 網路（儘管目前已決定興建第二個網路）⁸；汶萊則有三家行動電信業者為消費者與企業提供服務，但皆使用 Unified National Networks Sdn Bhd 所提供的同一個無線網路；⁹ 澳洲在 2024 年亦有兩家電信業者達成共享 4G 與 5G 無線存取網路（RAN）的協議。¹⁰

在這些範例和其他類似的情況中，「零售電信業者」（即向消費者和企業提供通訊服務之公司）數量大致持平，對電信服務消費者而言，仍有至少三間獲更多公司在相互競爭。¹¹

相對較新的趨勢，我們的預測核心在於：各國政府與監管機關已逐漸放寬電信業者合併的限制。自 2020 年起，全球已有 13 起電信合併或合資案獲得批准或正在審批中，減少了直接面對客戶之企業的数量。

- **美洲：**六起合併案（美國三件，加拿大、智利和哥倫比亞各一件）¹²
- **亞太地區：**五起合併案（印尼、馬來西亞、泰國、台灣與澳洲）¹³
- **歐洲：**兩起合併案（荷蘭與西班牙）¹⁴

相關人士正在等待英國監管機關針對 Vodafone UK 與 Three UK 的合併案做出最終決定，¹⁵ 值得注意的是，近年來義大利與丹麥的提議合併案都遭到否決。¹⁶

此外，前義大利總理 Enrico Letta 在 2024 年 4 月提交給歐盟的一份報告中，明確呼籲整合電信產業。¹⁷ 他的建議部分是基於歐盟發布的一份白皮書，該探討電信業者在歐洲市場中極度分散，而難以取得投資回報的挑戰。¹⁸ 一項數據顯示了歐洲為何可能率先批准電信業者的整併，歐洲每家行動業者的平均用戶數為 450 萬，美國為 9,500 萬，印度為 3 億，中國大陸則高達 4 億。¹⁹ 近期於 2024 年 9 月，前歐洲央行行長 Mario Draghi（也是義大利前總理）遞交了一份長達 69 頁的報告，其中一個章節表示支持市場電信業者整合。²⁰

在大多數國家中，以營收和用戶數計算基準，前兩大無線電信業者的財務實力通常很雄厚。排名第三的公司，財務實力通常較弱，而排名第四以下的財務狀況可能就更不理想。這些排名較低的營運業者也經常會被警告，未來可能無法繼續進行網路投資建設。在這樣的市場中，支持者向監管機關主張，相較由兩家主導或多家較弱業者組成的市場，維持三家具競爭力的業者更可以為消費者、企業和整體競爭格局帶來更多效益。

歐洲與其他地區的多個無線電信市場似乎可能會出現直接面對客戶的整合趨勢。監管立場轉變的一個重要因素為目前的連接選擇比以往更多。正如我們在《2024 年與 2023 年電信展望報告》（Telecommunications Outlooks）中所提及，近年來消費者和企業的選擇已大幅增加，這些選項包括但不限於：

- **固定無線接入家用寬頻服務的競爭：**Deloitte 預測，2025 年將有超過 3,000 萬戶透過固定無線接入聯網，相較今年增加 20%。²¹
- **低軌衛星的家用寬頻競爭，尤其是在農村和偏遠地區：**目前全球有超過 300 萬戶是透過此方式上網，預計在 2025 及 2026 年，將會推出多個新網路。²² 值得關注的是，低軌衛星對人口密度高，和地理環境相對單純之國家（例如部分歐洲國家）的影響較小。它們在人口密度較低的市場，或擁有山脈、沙漠或眾多小島的市場（例如亞太、非洲和美洲）中較具影響力。²³
- **3G、4G 和 5G 持續使用：**部分市場會同時使用多個網路，為消費者提供更多選項和競爭。²⁴
- **虛擬行動網路業者即將發生轉變：**虛擬行動網路業者已問世大約 25 年，近年來進入了轉捩點——開始在近期成長的行動用戶中佔有更高的比例。²⁵ 舉例來說，2004 年美國大約有 1,400 萬無線用戶使用有線虛擬行動網路業者提供的服務。在某種程度上，這些成功應歸功於 Wi-Fi 的可行性。美國有線公司透過家用 Wi-Fi 和地鐵 Wi-Fi 熱點，減少了 87% 的整體無線數據流量。²⁶

結論

對電信業者來說，在未來幾年維持穩健及具競爭力的無線網路的成本，可能比過去幾年更低。建置 5G 網路最昂貴的部分（如購買新設備和頻譜），對大部分已開發國家的業者來說已告一段落。無線存取網路（RAN）在 2022 年到達高點之後，預計未來幾年將以雙位數速度下降。²⁷ 除此之外，大多數最初部屬 5G 非獨立組網（NSA）的全球電信業者，在升級至 5G 獨立組網（SA）的支出相對較低。²⁸ 目前亦尚無跡象顯示 6G 會於 2030 年之前問世，即使到 2030 年也未必會實現。

電信產業的年度資本支出強度（Capex Intensity）在 2022 年達到 10 年來的高峰 17.8%²⁹，Deloitte 預測，將在 2025 至 2029 年進一步下滑至 15% 至 16%，這對無線存取網路（RAN）原始設備製造商來說可能是一項挑戰。另一方面，除固定無線接入外，電信業者通常很難從 5G 和其他新服務中獲利。如 2024 年《高科技、媒體及電信產業趨勢預測》（TMT Predictions）所述，消費者可能已進入「夠快就好」的時代，且大多數的人不願意為了更快的速度而支付更多費用。³⁰

此外，許多潛在的獲利來源，如提供消費者虛擬實境（VR）或擴增實境（AR）眼鏡的高階服務、企業 5G 專網、自駕車或遠距手術等充其量僅是小眾市場。部分電信業者已開始進軍電信相關的附加價值服務（例如：健康照護、農業科技、資安等領域），目前這些服務對大多數企業的營收影響相當有限。

部分大型電企業考慮進軍生成式 AI 資料中心業務，以開拓新的收入與利潤來源，但是這些通常都是市場中較大型企業的選項，對排名第三或第四的公司來說往往面臨併購壓力，而較難參與此類業務。³¹ 此外，許多附加服務通常需要達到一定規模才能成功，如前文所述，歐洲和部分亞洲小型電信業者是因為市場分散，而難以達到此類規模。

從全球的角度來看，無線電信企業者的合併通常需要兩個不同的監管機關批准。第一類是產業監管機關，例如：英國的通訊管理局（Ofcom）、美國的聯邦通訊委員會（FCC）和歐盟的歐盟層級與國家層級產業監管機關。³² 另一類是競爭監管機關，例如：英國的競爭與市場管理局（CMA）、美國的聯邦貿易委員會（FTC）及歐盟的競爭總署（DG Competition）。而大部分亞太地區的國家，都設有類似的監管機關劃分。³³

整體而言，產業監管機關通常對國內無線電信業者合併持較開放的態度，而面對競爭監管機關的挑戰較大。我們認為在部分司法管轄區域中可能正在發生變化，尤其是在歐洲近期的一些報告和公開信，紛紛鼓勵規模較小的無線業者進行合併。

儘管如此，監管機關將可能花費較長時間詳細審查。近年來，多起電信合併案的審查時間約歷時 24 至 36 個月才完成。³⁴ 如果我們的預測正確，合併案整體的批准率可能會增加。在合併批准時，監管機關有時會無條件批准，但也可能會提出附帶條件，例如要求切割部分業務、提供價格保證，承諾未來投資或提供 5G 覆蓋範圍等。³⁵

By **Duncan Stewart**
Canada

Jack Fritz
United States

Dan Littmann
United States

Ariane Bucaille
France

Endnotes

1. Deloitte analysis of recent events in Europe, specifically recent letters and white papers from Draghi and Letta (see endnotes 16 and 19).
2. Deloitte analysis of historical merger and acquisition trends, combined with early publicly available signs of deal activity.
3. CSI Magazine, "[Telecom consolidation: Over 500 M&A deals in five years](#)," July 23, 2024.
4. Duncan Stewart, Dan Littmann, Girija Krishnamurthy, and Matti Littunen, "[Telecoms tackle the generative AI data center market](#)," Deloitte Insights, Sept. 16, 2024.
5. Stephanie Price, "[For sale: Canadian read-throughs from global telecom asset sales](#)," CIBC Capital Markets, Aug. 14, 2024.
6. ArdorComm News Network, "[Telecom M&A activity witnesses surge: 514 deals from 2019 to 2023](#)," ArdorComm Media Group, July 25, 2024.
7. Sue Marek, "[Marek's take: 5G network sharing may be the answer](#)," Fierce Network, May 14, 2021.
8. Affandy Johan, "[5G in Malaysia – Single wholesale network driving regional leadership](#)," Ookla, March 17, 2024.
9. Digital Regulation Platform, "[Changing the operating model: the creation of UNN in Brunei Darussalam](#)," Aug. 27, 2020.
10. Australian Competition & Consumer Commission, "[Optus Mobile Pty Ltd and TPG Telecom Limited proposed network and spectrum sharing](#)," Sept. 5, 2024.
11. Megan Emfosi Meena and Jiaying Geng, "[Dynamic competition in telecommunications: A systematic literature review](#)," Sage Open, April 26, 2022.
12. Detecon Spotlight, "[Telco mergers and acquisitions](#)," 2022; M&A Community, "[12 significant telecom mergers and acquisitions over the last 20 years](#)," Sept. 17, 2024; Henri Capin-Gally, Sergio Márquez García Moreno, and Bill Parish, "[Energy and telco deals power Mexican M&A](#)," White & Case, March 6, 2024; Geusseppe Gonzalez, "[Access Alert: Colombian telecoms industry faces shakeup with potential Millicom-Telefonica merger](#)," Access Partnership, Aug. 8, 2024.
13. Julber Osio, "[Asia-Pacific telcos consolidate to compete with market leaders](#)," S&P Global, May 25, 2023; Tom Leins, "[Deal-making Down Under: Oceania's wave of telecom M&A](#)," TeleGeography, March 24, 2021.
14. Jan Frederik Slijkerman, "[Telecom Outlook: Will we see more mergers and buyouts in 2022?](#)" ING, Jan. 28, 2022; Jan Frederik Slijkerman and Diederik Stadig, "[Telecom tycoons on the move?](#)" ING, Feb. 1, 2024.
15. Competition and Markets Authority, "[How we are investigating the Vodafone / Three potential merger](#)," Sept. 13, 2024.
16. Slijkerman and Stadig, "[Telecom tycoons on the move?](#)" ING, Feb. 1, 2024.
17. Enrico Letta, "[Much more than a market](#)," Consilium, April 2024.
18. European Commission, "[White paper - How to master Europe's digital infrastructure needs?](#)" Feb. 21, 2024.
19. Hamish White, "[Europe's looming mobile crisis](#)," Technative, April 2024.
20. Mario Draghi, "[EU competitiveness: Looking ahead](#)," European Union, Sept. 9, 2024.
21. See section "Fixed wireless access: Contrary to popular opinion, adoption may continue to grow" in chapter [Updates: Past TMT Predictions' greatest hits and \(near\) misses](#).
22. Ongoing Deloitte analysis of current and proposed low Earth orbit satellite networks.
23. Deloitte author conversations with communications providers in North America, Europe, Africa, and Asia.
24. Deloitte analysis of developing world wireless networks.
25. Piran Partners, "[Seizing the future of telecoms: The continuing ascendancy of MVNOs](#)," April 12, 2024; Puneet Takyar, "[The journey of MVNOs](#)," Comviva, April 1, 2021.

26. Jeff Baumgartner, “[Cable snared nearly half of US mobile line adds in Q3 – analyst](#),” Light Reading, Nov. 16, 2023.
27. Juan Pedro Tomás, “[Global RAN market faces challenging scenario in Q2: Dell’Oro](#),” RCR Wireless News, Aug. 19, 2024.
28. Deanna Darah, “[5G NSA vs. SA: How do the deployment modes differ?](#)” TechTarget, July 25, 2024.
29. Matt Walker, “[Telco capital intensity hits 10 year peak in 2Q22](#),” MTN Consulting, Sept. 6, 2022.
30. Paul Lee, “[No bump to bitrates for digital apps in the near term: Is a period of enough fixed broadband connectivity approaching?](#)” Deloitte Insights, Nov. 29, 2023.
31. Stewart, Littmann, Krishnamurthy, and Littunen, “[Telecoms tackle the generative AI data center market](#),” Deloitte Insights, Sept. 16, 2024.
32. DataHub, “[Regulatory authority - Institutional structure](#),” accessed October 2024; Policies, “[Telecommunications national regulatory authorities](#),” European Commission, Jan. 11, 2023.
33. Lynn Robertson, “[Interactions between competition authorities and sector regulators – contribution from business at OECD \(BIAC\)](#),” Organisation for Economic Co-operation and Development, Nov. 18, 2022.
34. Research Notes, “[5G rollout slowed while mobile operators await merger approval. Case in point: the slow rollout in UK](#),” Strand Consult, Oct. 5, 2024.
35. Deloitte author’s assessment of multiple approved mergers in North America, Europe, and Asia in the period of 2015 to October 2024.

Acknowledgements

The authors would like to thank **Dieter Trimmel, Hugo Santos Pinto, Prashant Raman, Pankaj Bansal,** and **Akshay Jadhav** for their contributions to this article.

Cover image by: **Jaime Austin**; Getty Images, Adobe Stock

更新動態：回顧過往《高科技、媒體及電信產業趨勢預測》成果

《高科技、媒體及電信產業趨勢預測》(TMT Predictions) 將重新檢視先前針對企業邊緣運算、5G 通訊、後量子密碼，以及開源半導體地緣政治，進行預測的內容。

2025 年全新推出一系列較短篇幅的文章，探討在先前版本中涵蓋的議題。下方將重點整理，我們過去的預測表現。請繼續閱讀，了解這些議題的最新發展：

- 生成式 AI 邁向企業邊緣：「地端 (On-Prem) AI」蓬勃發展
- 固定無線接入 (FWA)：與普遍觀點相反，應用比例將會持續成長
- 5G 獨立組網 (5G SA) 發展緩慢：6G 是否會延遲問世？
- 開放式無線存取網路 (Open RAN) 行動網路和供應商選擇：現今是單一供應商，何時會實現多供應商模式？
- 量子技術起步雖緩慢，但仍應盡早規劃防範措施
- RISC-V：彌補地緣政治生成式 AI 漏洞

生成式 AI 邁向企業邊緣：「地端 (On-Prem) AI」蓬勃發展

企業擁有自己的伺服器，可為 AI 提供更私密、安全、靈活，及可能更低成本的 IT 環境。

Duncan Stewart、Karthik Ramachandran、Gillian Crossan 及 Jeff Loucks

Deloitte 預測，儘管雲端生成式 AI 仍為主流選項，但在 2025 年，全球大約有一半的企業將在內部新增 AI 資料中心基礎設施（地端部署，On-Prem），這即是企業邊緣運算的例證之一。其是為了保護其智慧財產權（IP）與機敏資料，並遵守資料主權或其他法規，同時有助於降低成本。此延續了我們在 2024 年底觀察到的趨勢，大約 45% 的生成式 AI 晶片是出貨給超大規模企業（Hyperscalers），而 55% 的需求則是來自消費者、互聯網公司和企業用戶的混合需求。¹ 根據 Deloitte 2024 年《企業生成式 AI 現狀》(State of Generative AI in the Enterprise) 第二季報告，80% 具有高度 AI 專業知識的企業表示增加了雲端的 AI 支出，61% 的企業也正在增加自有硬體的投資。² 我們預測在 2025 年，企業地端 AI 伺服器的市場規模將可能達到 1,000 億美元左右。³

2021 與 2023 年的《高科技、媒體及電信產業趨勢預測》(TMT Predictions) 中，探討了企業將會如何預期投資邊緣 AI 解決方案，以加速運算工作的執行。⁴ 延遲（從 AI 接收輸入的內容到產生回應需要的時間）是早期企業使用邊緣 AI 的驅動因素之一。⁵ 然而此次情況有所不同，延遲不再視為推動企業部屬生成式 AI 的關鍵因素，多數生成式 AI 的請求處理時間達數千毫秒，因此回覆時間通常構成重大挑戰。⁶

相反地，在生成式 AI 的私有化、資料主權性及安全性的需求上，正推動企業邊緣運算進入全新的成長階段。

隨著企業擴大生成式 AI 的投資和應用，⁷ 雲端供應商、超大規模企業 (Hyperscalers)、電信業者和 AI 和科技公司開始擴建資料中心，以滿足快速增加的市場需求，預估在 2025 年投入晶片與資料中心的資金將超過 2,000 億美元。⁸ 然而許多跨國企業似乎正使用混合模式，同時使用第三方雲端解決方案並投資自有硬體，以在內部進行 AI 訓練與推理運算，以便能提供更安全、可控、資料主權性及靈活的 IT 環境。⁹ 相較於完全依賴外部雲端基礎建設，在當地處理資料可以加速企業回應時間，並有效解決生成式 AI 在隱私和安全的挑戰，此論點 Deloitte 在《企業生成式 AI 現狀》(State of Generative AI in the Enterprise) 第三季報告中也有所揭示。¹⁰

各種企業應用場景和機會，使在邊緣部署生成式 AI 成為可行及具有高度相關性的選擇。例如：

- 銀行和金融服務企業：傾向於內部儲存大量機敏資料，以確保資料安全，同時能更有效掌握生成式 AI 模型的運行及管理。¹¹
- 媒體及娛樂產業：正在使用自然語言 AI 的整合解決方案，促進應用於動畫與內容創作（例如撰寫初稿劇本或散文）、遊戲開發和娛樂領域（例如：透過電影字幕內容的模式分析，優化推薦引擎，並提升終端使用者體驗）。¹²

2025 年企業邊緣運算應用生成式 AI 的趨勢為何？企業在當地投入資金專注於發展生成式 AI，將使用於員工的智慧型手機和個人電腦等裝置，以提升運算效率與使用體驗。¹³ 但企業仍有其他選擇，這不僅影響企業在地端 AI 解決方案的投資規模，也關係到資料倉儲或資料中心的調整需求。新一代 AI 伺服器通常功耗更高、體積更大，有時甚至需要液冷技術散熱，這對企業 IT 基礎設施的適應與升級提出了新的挑戰。

2024 年，許多企業使用生成式 AI 設備（來自不同供應商），其尺寸大小與家用印表機相當，大約重 300 磅（約 136KG）、功耗約 10kW，價格不到 50 萬美元，運算能力大約為 30PetaFLOPS。¹⁴ 而 2025 年，部分企業仍可能選擇類似設備，但也有部分將採購更強大的 AI 伺服器，其體積超過大型冰箱，重達 3,000 磅以上（約 1,360KG）、功耗達 160kW，且需要液冷系統散熱，價格超過 300 萬美元，運算能力更是超過 1.3ExaFLOPS，遠超前一代設備。¹⁵

具體而言，企業並非自行建構這些設備，生成式 AI 晶片製造商和伺服器原始設備製造商 (OEM) 正在提供機架式 (Rack-Scale) 伺服器，並將其安裝在地端伺服器。

非 AI 的本地端運算通常比雲端更具成本效益，長期而言，生成式 AI 的運算成本預期也可能遵循相同趨勢。

斥資 300 萬美元購買 Exascale 級生成式 AI 超級電腦，可能是一筆龐大的開銷，但是從其他角度來看，投入的資金可能不算太高。各產業中大型企業的 IT 預算可能高達數十億美元，而訓練大型語言模型的成本可能從 100 萬到 1 億美元不等。¹⁶ 除此之外，雲端生成式 AI 運算功能也需要資金，無論是使用於訓練或推理，長期累積的支出投入將會相當可觀。

企業完全依賴地端生成式 AI 解決方案的可能性不高，大多數企業仍在某些時候會使用雲端 AI，雲端 AI 在整體 AI 運算中可能會佔有更大的比例。如同許多 IT 架構的發展趨勢，混合模式（Hybrid）很可能是最終選擇

- 地端部署（On-Prem）：提供更高的安全性、低延遲、更強韌性，在隱私保護上具有優勢
- 雲端（Cloud）：著重於更多選擇、更高的靈活性、可擴充性，有助於 AI 測試和創新

但是對於想要擁有自有硬體的企業來說，從投資報酬率（Return on Investment，ROI）的角度來看，購置地端 AI 解決方案是否符合經濟效益？情況可能是非 AI 的地端運算通常比雲端更具成本優勢，¹⁷ 長遠來看生成式 AI 運算的成本趨勢也可能趨近。即使「硬性」ROI 無法完全展現，地端部署（On-Prem）在智慧財產權（IP）所有權、隱私、安全性和韌性的優勢，都使得地端部署的投資更加有價值。此外，拆分學習（Split Learning）和拆分推理（Split Inference）等新型訓練與推理技術，將使生成式 AI 工作負載量分配至邊緣裝置，進一步優化運算需求並減緩延遲，同時解決隱私與安全問題。¹⁸

結論

部分企業可能認為，雖然短期內沒有地端部署生成式 AI 運算功能的需求，但長遠來看投入卻勢在必行。於此情況下，現在投入學習讓地端運算發揮最大效益，並作為混合雲與當地部署架構的一部分，或許可以證明投資的硬體成本是合理的。

值得關注的議題是生成式 AI 是否存在泡沫化風險？隨著許多企業同時投入硬體建設，且認為投資不足的風險大於過度投資，在短時間內可能會出現產能過剩的情況。如發生此類情況，已投資地端生成式 AI 伺服器的企業將要如何應對？企業可能會優先使用自有硬體（因已購買設備且進入折舊階段），而非有額外支出產生以使用雲端生成式 AI 運算資源。

大多數預測核心在於大型企業可能會建立地端生成式 AI IT 基礎建設，以支援業務流程與維運。例如：銀行、汽車製造商、醫療機構、公共部門機關等，這就是我們最主要所謂的邊緣運算。但還有另一個市場，企業也將其視為邊緣運算一部份，即為電信邊緣（Telco Edge）。如同在 Deloitte 9 月的分析中指出，全球有超過 15 間電信業者宣布正在建設生成式 AI 資料中心。其中一部分是為了滿足內部需求，另一部分則計畫對外銷售，提供客戶生成式 AI 即服務（Gen-AI-as-a-Service）。¹⁹ 因此，企業可以選擇：（1）向雲端供應商購買、（2）投資地端硬體、（3）向生成式 AI 即服務供應商購買運算資源。

最後，還有與永續發展和平衡相關的問題。在許多方面，企業邊緣生成式 AI 與超大規模企業（Hyperscalers）的資料中心設備相同。但當 1,000 台伺服器分散部署於 1,000 家企業，可以分散電力負荷，且對電網的壓力較小，是相較於在單一建築內部署 1,000 台或 10,000 台伺服器對電網的衝擊更理想的做法。²⁰ 另一方面，無線存取企業的營運商通常更具效率，其電力使用效率（PUE）可達 1.2 或 1.3，並能夠大規模採購低碳能源，相對於小型企業來說卻不易實現²¹。相比之下，一般企業資料中心的 PUE 是介於 1.67 到 1.8 之間，代表將 1,000 台生成式 AI 伺服器分散於 1,000 家企業，其產生的碳足跡可能比在超大規模企業（Hyperscalers）資料中心更高。²²

固定無線接入（FWA）：與普遍觀點相反，應用比例將會持續成長

由於美國固定無線接入（Fixed Wireless Access，FWA）網路新增用戶數較去年稍微下降，且部分市場預計至 2026 年才會起飛，美國和全球市場將可能出現未實現或潛在的成長機會

Duncan Stewart、Dan Littman、Jack Fritz 及 Hugo Santos Pinto

固定無線接入（FWA）是指消費者透過固定蜂巢式裝置（主要為 5G）連接至家用寬頻，而非傳統有線方式。在過去幾年，這已成為美國 5G 大幅成長的主要原因，預計至 2024 年底將有超過 1,000 萬戶連接 FWA。²³ 消費者似乎受到具競爭力的價格和「夠好」的網速吸引，²⁴ 電信業者也發現在人口密度較低的區域，以 FWA 提供寬頻服務比光纖更具成本效益。²⁵

但在 2024 年第一季，美國的固定無線接入（FWA）的淨成長（即每季新增用戶數減去解約用戶數）低於 2023 年第一季，²⁶ 而印度 FWA 市場仍處於起步階段（預計大力推動成長的新市場，預測至 2030 年用戶數將可能達到 1 億），許多人預測 2025 年將成為 FWA 成長趨緩的一年。²⁷

Deloitte 2022 年《高科技、媒體及電信產業趨勢預測》（TMT Predictions）預測 FWA 的年增率將近 20%，並在 2023 年全球 FWA 用戶數將達到接近 1 億（主要為 4G）。²⁸ 實際成長大致符合預測，預計至 2024 年底，全球 FWA 用戶數將超過 1.5 億，其中 30% 將來自 5G。²⁹

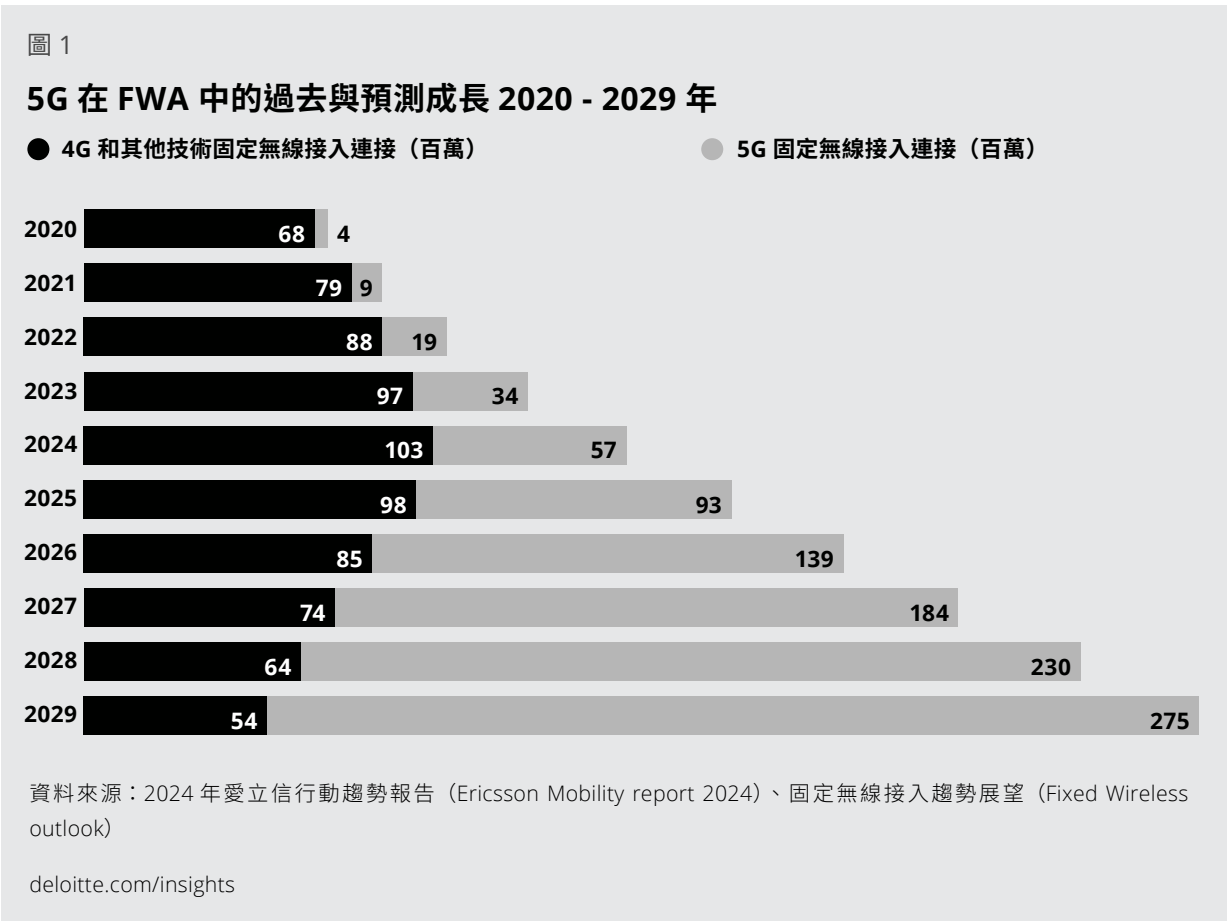
Deloitte 預測 2025 和 2026 年的全球 FWA 淨成長，將會以每年約 20% 的成長速度攀升。此預測受到以下三項趨勢推動：

- 1. 部分 FWA 成長未受到全球關注：**美國和印度都是大型市場。美國每季 FWA 淨成長將近 100 萬，³⁰ 並隨著印度的 FWA 用戶數增加，每季新增量將可能升高。³¹ 其他市場也同樣正在成長，例如：義大利每季淨成長 10 萬 FWA 用戶，雖然該市場較少受到關注，³² 但從每戶家庭的成長率來看，義大利的 FWA 成長速度僅略低於美國。³³ 相較於歐洲的其他地區、拉丁美洲、東南亞和非洲國家，義大利的 FWA 普及率相對較高³⁴，預計在 2025 和 2026 年，美國和印度以外的市場將會出現數百萬的淨成長。
- 2. 企業選擇 FWA 的使用上升：**目前美國大部分的 FWA 成長是來自消費者，近期表現則趨向有越來越多企業（主要為中小型企業）開始選擇 FWA 連線。³⁵ 部分企業發現 FWA 更具吸引力，其中包括 FWA 提供單一聯絡窗口、整合式帳單，且更具備安全性。³⁶ Deloitte 預測，僅在美國，2025 和 2026 年裡每年將會有超過 100 萬家企業使用 FWA。³⁷
- 3. 新技術提升了每個市場可使用 5G FWA 的用戶上限：**過去幾年，部分人士認為美國 FWA 的發展存在自然上限。由於現有的 5G 技術和可用頻譜主要為 2.5GHz 與 3.5GHz，若 5G FWA 訂閱數超過約 1,000 萬，可能會影響固定與行動無線用戶的使用體驗。³⁸ 業界推測在短期內，部分地區的營運業者將會被迫停止向新用戶提供 FWA 服務。³⁹ 但即使訂閱用戶數逐漸接近該數字，用戶的下載速度仍在持續提升。⁴⁰ 5G 技術的升級，特別是在無線電技術的進階應用方面，顯示現有頻譜可支持多達 2,000 萬個美國家庭（約占 1.06 億美國寬頻家庭的 19%）⁴¹ 連接 FWA。這表示目前美國 FWA 擁有每年新增約 300 至 400 萬新用戶的成長趨勢，可能至少持續數年。

結論

FWA 的持續成長，對於試圖從 5G 投資獲利的電信業者來說可能是利多。除了 FWA 外，截至 2025 年大部分具有提升營收潛力的 5G 服務規模都較小。⁴² 從市場背景來看，美國 FWA 的平均價格為每一個月 50 美元左右，若 2025 年新增 400 萬個新 FWA 用戶，將額外帶來 24 億美元的收入成長，此外，透過捆綁銷售，FWA 還能降低行動業務的客戶流失率。⁴³

另一方面，提供其他類寬頻接入服務的業者（例如同軸電纜、銅纜 DSL、光纖），可能會持續面臨來自 FWA 的競爭，進而可能會導致用戶流失，或在維持及價格方面面臨挑戰。FWA 已成為這些傳統寬頻技術的有力競爭對手，對於預期 FWA 成長趨勢即將放緩的業者來說，可能還需要等待一段時間。



固定無線接入（FWA）的持續增加，可能成為電信公司將 5G 投資轉化為收益的助力。

5G 獨立組網（5G SA）發展緩慢：6G 是否會延遲問世？

由於對投資報酬率（ROI）的擔憂，電信業者重新評估 5G 獨立組網的投資，並推遲 6G 發展

Prashant Raman 及 Duncan Stewart

5G 獨立組網（Standalone；SA）的部署進度比最初的預期更慢，⁴⁴ 電信業者可能會因現有 5G 投資的回報不如預期，而對大規模投資 5G SA 持謹慎態度。現況來看，6G 推出的時間遙不可及。

截至 2024 年 3 月，全球已推出 5G 的 585 家電信業者中，僅有 49 家的業者（約占 8%）已部署、正在推出或試運行 5G 獨立組網（SA）網路。⁴⁵ 這一緩慢的使用率顯示出，業者可能會因為資本回報的不確定性而不願投資 5G SA，至少在短期內是如此。此外，缺乏 5G SA 應用情境的獲利模式，可能會成為使用的一大挑戰。由於缺乏明確的收益來源，業者正在放緩部署 5G SA 的腳步，並對大規模投資其基礎建設持保留態度。⁴⁶ Deloitte 預測在 2025 年升級為獨立組網的網路將不到 20 個，5G SA 在整體 5G 部署中的比例將維持在 12%。⁴⁷

5G 在 2018 年推出，初期大多數業者都選擇了非獨立組網（NSA）架構，即在現有 4G 基礎設施上構建 5G 網路，以加速服務推出。⁴⁸ 採取此方式的主要是因當時 5G 獨立組網（SA）設備尚未完全成熟，以及業者希望充分利用既有的基礎設施。⁴⁹ 因此，儘管未能實現完整的 5G SA 功能，推出 NSA 仍促使 5G 服務的部署更迅速及更具成本效益，提供更快的網路速度和更穩定的連接性。⁵⁰ Deloitte Global 在 2022 年曾預測，至 2023 年底投資 5G SA 網路的行動網路業者數量（包括試驗階段、部署計畫，或正式推出），將較 2022 年的 100 多家增加到至少 200 家，但是此項預測未實現。⁵¹

5G 獨立組網（SA）部署的延遲可能會阻礙這些技術的開發與測試，進而影響 6G 的發展進程。

現有 4G 和 5G 非獨立組網（NSA）網路已能有效支援當前消費者和企業最常使用的應用，也降低了業者大規模投資 5G SA 的急迫性。⁵² 在企業方面，雖然 5G SA 具有徹底改革不同產業的潛力，但基於 SA 的新興 5G 解決方案的發展與使用速度低於預期，使業者在 5G SA 部署方面採取更謹慎的態度。⁵³

5G 獨立組網（SA）基礎建設在測試和驗證 6G 的關鍵技術上至關重要，⁵⁴ 例如：低延遲、高可靠性和網路切片技術，皆是 5G SA 支援 5.5G 和 6G 所構想的核心應用場景。⁵⁵ 延遲部署 5G SA，可能會阻礙這些技術的開發與測試，進而影響 6G 的發展進展。

5G 的高成本、獲利緩慢以及消費者和企業使用速度不如預期，使業者面臨難以實現 ROI 的難題。⁵⁶ 此外，原本預期能帶來新獲利的 5G 網路應用程式介面（API）等增值服務，目前亦尚未實現。⁵⁷ 由於 6G 的應用仍在開發中，行動網路業者可能會專注於挖掘 5G 潛力和回收投資成本，這可能使 6G 的開發和部署時程更往後推遲。儘管各項研究與標準制定都將持續推進，但業界普遍認為首批大規模的商業 6G 部署將在 2030 年左右實現。⁵⁸

結論

行動網路業者在規劃 5G 獨立組網（SA）時，應重點思考如何持續優化其現有的 5G 非獨立組網（NSA）網路效能。採取分階段投資策略，優先佈局需求較高的地區或產業，不僅有助於控制成本，也能更有效地分配資源。此外，可探索新的收入模式，例如將 5G SA 做為服務提供給特定產業，或將其與雲端、AI、邊緣運算相結合，或許能開拓新商機。同時與企業客戶合作，共同開發產業特定解決方案，特別是製造、醫療健康和物流等領域，也能促進 5G SA 的使用，進而支持其持續投資。

政府與監管機關應意識到，5G 獨立組網（SA）的部署是一個持續進行的過程，且進展比預期緩慢。這可能會影響 5.5G 和 6G 的推動，因此兩者可能需要重新評估下一代網路需求與頻譜的規劃策略。對於電信設備製造商而言，SA 網路部署原本在 2021 年被視為 5G 投資高峰後，2020 年代後期至 2030 年代初期的下一波投資浪潮間，填補收入空缺的重要機會。⁵⁹ 如若 SA 的部署進度持續放緩，則該投資低谷可能會比預期更深或更久，進而影響設備製造商的收入與盈利能力。

開放式無線存取網路（Open RAN）行動網路和供應商選擇： 現今是單一供應商，何時會實現多供應商模式？

開放式無線存取網路（Open RAN）邁向多元化、多供應商生態系統的過程正面臨成長緩慢和複雜挑戰

Prashant Raman 及 Duncan Stewart

開放式無線存取網路目的在於透過提供行動網路業者（MNOs）更多選擇與更大靈活性，以實現網路建設更加開放，期望能帶來更優質的網路與更低的成本。儘管市場對其抱有高度期待並獲得業界支持，但過渡至多元化、多供應商生態系統的進程，仍比部分預期的更為緩慢與複雜。⁶⁰ 想要實現真正的多供應商 Open RAN 可能還需要一段時間，Deloitte 預測，2025 年前將不會有新的多供應商 Open RAN 網路部署。⁶¹

無線存取網路（Radio Access Network，RAN）是蜂巢式網路的重要組成元件，負責管理行動裝置與核心網路之間的無線通訊。以往都是 RAN 使用封閉且專有解決方案，整體網路均來自於單一供應商，無法與其他供應商設備相容。⁶² Open RAN 是電信產業中的變革性概念，目的是將網路元件的設計和應用標準化。⁶³ 此項技術在 O-RAN 聯盟（O-RAN Alliance）於 2018 年成立，該聯盟的目的是促進更開放、可互操作的 RAN 架構。⁶⁴

到 2028 年，單一供應商的 Open RAN 解決方案預計將佔全球 RAN 總收入的 15% 至 20%，而多供應商的 Open RAN 解決方案則預計僅佔 5% 至 10%。

Deloitte 曾預測在 2021 年，全球公共網路 Open RAN 的部署數量將從 35 個翻倍至 70 個，⁶⁵ 但此項預測過於樂觀。截至 2024 年 3 月，正在進行中公共網路 Open RAN 的部署和試驗數量僅達 45 個，其中僅有 2 個為多供應商 Open RAN。⁶⁶

Open RAN 的目標是為業者提供更開放的替代方案，以促進市場競爭與多樣性。⁶⁷ 但現實情況不如預期。許多業者持續向同一供應商購買無線電和基頻產品，而專門做 Open RAN 公司的影響力仍然有限，目前全球五大 RAN 供應商仍佔據 95% 的市場。⁶⁸

直至 2024 年，Open RAN 僅佔整體 RAN 市場的 7% 到 10%，約 25 億到 35 億美元，相較整體市場預估為 350 億美元，規模有所差距。⁶⁹ 此外，短期內單一供應商的解決方案，預計較 Open RAN 更受到廣泛使用。預測至 2028 年，單一供應商 Open RAN 解決方案將佔整體 RAN 收入的 15% 到 20%，而多供應商 Open RAN 僅佔市場的 5% 到 10%，傳統 RAN 將仍佔市場的 80% 到 85%。⁷⁰ 儘管業界普遍認為多供應商 Open RAN 具備強大的長期潛力，⁷¹ 但想實現此願景可能需要更多時間。

在 Open RAN 生態系統中，可能需要整合多種硬體，才能實現高容量性能和成本效益。業者歷來更偏好傳統 RAN 供應商的一站式解決方案，供應商會負責處理所有元件，並提供單一責任窗口。⁷² 相較於小型業者而言，多家供應商採購可能會帶來挑戰。

不過 Open RAN 市場仍是美國規模策略的一部分，有助於提升經濟安全並降低對外國供應鏈的依賴，尤其是在電信等關鍵領域。⁷³ Open RAN 技術可以在過程中發揮關鍵作用，目的是打破目前由少數大型企業（大多為非美國企業）主導的市場集中情形。⁷⁴ 地緣政治的局勢緊張已導致某些企業被排除在重要市場以外，包括美國與歐洲，進而限制了供應商選擇，並突顯出供應鏈多元化的必要性。⁷⁵ Open RAN 有助於透過提升其國內市場競爭力與創新，並解決此問題。Open RAN 的產業回流可能是一個策略機會，有助於鼓勵在當地製造電信設備，並在全球電信基礎建設市場中維持領先地位。⁷⁶

結論

大多數行動網路業者（Mobile Network Operator，MNO）已將 Open RAN 納入規劃，此議題已經討論數年，⁷⁷ 部分公司已經開始採取相關步驟，其他業者則預期在未來一到兩年內有所行動。MNO 可以透過將 Open RAN 元件逐步整合至網路非關鍵區域，以低風險的實驗性方式探索此項技術，並與其他業者、供應商或聯盟進行合作測試，有助於解決部署多供應商解決方案的技術挑戰。與新興 Open RAN 供應商和既有的 RAN 供應商密切合作，也有助於為開發符合特定的網路需求帶來機會。此外，透過員工培訓或聘請專家來投資建立內部能力與專業知識，都有助於因應多供應商環境中的複雜性。

現有的大型 RAN 供應商可能希望在「開放（Open）」方面取得平衡。某種層面上，Open RAN 可能會被視為對現有業務的威脅，但若 Open RAN 已是趨勢所向，企業應先打破自我，而非被他人取代。原始設備製造商代工生產（Original Equipment Manufacturer，OEM）已專注於開發可與既有網路基礎設施無縫整合的互通產品，⁷⁸ 可能會持續在 Open RAN 領域與中小型企業和新創公司合作。積極參與產業聯盟，將有助於 OEM 和新興的 Open RAN 供應商確保產品符合不斷發展的標準和市場趨勢。

多供應商 Open RAN 的優勢之一是提升供應商多樣性，尤其是來自歐洲及亞洲地區以外的供應商。⁷⁹ 若目前 Open RAN 的發展持續緩慢，則短期內供應商的地理分布不太可能發生實際變化，或需要採取措施才能讓既有的 RAN 供應鏈多樣化。

圖 2

2028 年，多供應商 Open RAN 收入才可能會明顯成長

全球 RAN 與多供應商 Open RAN 收入，2020 - 2028 年

○ 全球 RAN ● 多供應商 Open RAN



備註：A 代表實際數值、E 代表當前年分的預估值、P 代表 Deloitte 預測數字。

資料來源：Deloitte 分析資料和 Dell'Oro 報告。

deloitte.com/insights

量子技術起步雖緩慢，但仍應盡早規劃防範措施

量子技術在探索和財務建模領域可能還需要數年時間，但需要立即採取行動將量子時代的網路防禦升級。

Colin Soutar、Duncan Stewart、Scott Buchholz 及 Gillian Crossan

2019 年與 2022 年的《高科技、媒體及電信產業趨勢預測》(TMT Predictions) 報告曾探討量子運算，⁸⁰ 並在 2022 年版中提到加密機制的潛在威脅。⁸¹ 其中部分預測已經實現。目前的量子電腦在實際應用上仍不夠穩定，尚無法在費用取得平衡⁸²；網路安全領域則受到高度關注。⁸³ 如同 Deloitte 在 2023 年 12 月所預測，各項網路安全標準已在今年有所落實。⁸⁴ 例如：美國國家標準暨技術研究院 (National Institute of Standards and Technology, NIST) 近期發布了後量子加密標準。⁸⁵ NIST 被視為網路安全與加密機制的權威機構，負責制定與推動資料加密與數位簽章演算法、協議及框架，以確保安全的數據傳輸與交易保護。其中的網路安全風險與秀爾演算法 (Shor's algorithm) 有關，此項演算法是在 1994 年提出，利用量子效應，在量子電腦上執行時迅速破解公開金鑰加密機制⁸⁶。

Deloitte 預測至 2025 年，致力於應用後量子加密機制解決方案的企業數量，將較 2023 年成長 4 倍，相關支出也同樣將翻倍成長。基於 2025 至 2035 年政府系統遷移至整合系統的預測成本⁸⁷，以及金融業持續推動風險緩解措施，已是保守估計。⁸⁸

評估量子技術的積極應用情境，和對抗量子網路威脅的時間軸時，關鍵不僅在於量子計算何時能實現應用或運行秀爾演算法（Shor's algorithm），更著重於需要多長時間來採納並實施 NIST 最新標準。此涉及依賴雲端超大規模企業（Hyperscalers）、專用資安元件，以及利用加密資料庫提供機密性與信任等安全功能。但在自行開發應用程式的情況下，需要多少時間完成應用使用和落實 NIST 標準，為了回答此問題，部分企業已經開始評估自身風險，並規劃符合其營運目標的升級藍圖，逐步處理潛在採購與合約要求。

而加劇時間軸不確定性的是國家或國家支持的行為者，據稱其目前正在蒐集加密資料以待未來解密（即為「先竊取、後解密」攻擊）。⁸⁹ 企業如何應對這類的「潛伏威脅」備受關注，特別是在涉及需有長期保護週期的資料時（如個人資訊），是否會主動採取 NIST 標準，以預防資料外洩的威脅。

有趣的是，部分大型服務供應商已開始將後量子加密機制納入其平台。⁹⁰ 預計 2025 年後其他提供端對端加密的通訊服務，也將使用量子加密機制。此外，超大規模企業正提供相關服務，協助客戶能針對雲端服務效能進行測試，瞭解原型設計或評估量子抗性對雲端服務的效能影響。⁹¹

結論

量子優勢，即指在解決實用問題上，量子電腦比傳統電腦具有明顯優勢。除非出現意外突破，否則可能仍需幾年的時間才能實現（圖 1）。隨著圖 1 中的藍色陰影區域（應用開發及應用商業化）逐漸接近，企業應考慮使用以基於風險的策略來應對加密威脅，並評估此風險與其他風險的相對影響，以積極採取緩減措施。⁹²

開發出具加密破解能力的量子電腦的企業，目前不太可能公開其擁有並能使用該技術。因此外界無法提前掌握情況。此情形使得盡早採取行動變得更加重要，以避免企業在資源匱乏而必須關閉系統時被迫應對突發狀況。⁹³

開發後量子加密機制可能是 2025 年量子運算在短期內的重要影響，企業仍可提前著手優化相關流程，為量子電腦實現量子運算優勢做好準備，並開發能為金融、醫療健康等產業帶來實質應用效益。

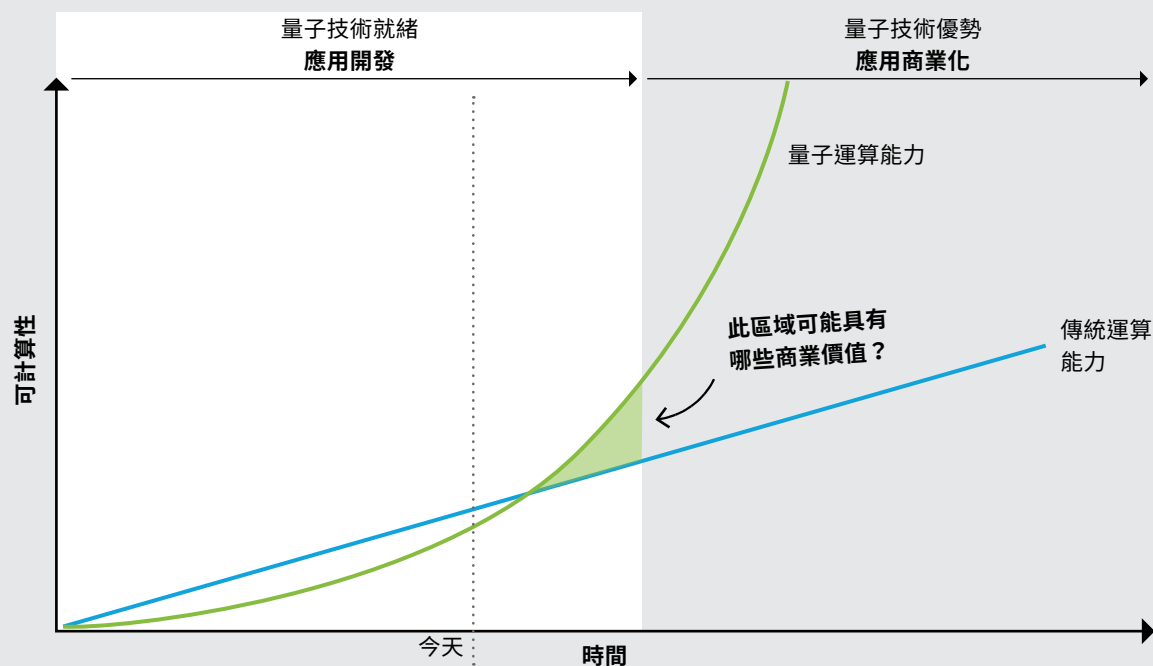
開發出具加密破解能力的量子電腦的企業，目前不太可能公開其擁有並能使用該技術。

量子技術可以透過更有效的藥物研發和金融或環境干擾的預測模型，為解決長期存在的挑戰帶來展望。隨之而來的網路安全威脅，不應掩蓋這些潛在的正面效益，特別是在可以有條不紊並循序漸進的方式升級加密機制，以有效降低風險。

圖 3

量子運算正在逐步成熟

我們正在努力瞭解當前技術的可能性、即將實現的技術，以及未來可能存在的機會



資料來源：Deloitte 分析。

deloitte.com/insights

RISC-V：彌補地緣政治生成式 AI 漏洞

作為專用晶片設計的開源替代方案，RISC-V 在 CPU 設計人員之間越來越受歡迎。其為在受出口管制市場中，支持生成式 AI 領域的潛在角色，為新技術帶來新的變數。

Duncan Stewart、Christie Simons、Jeroen Kusters 及 Gillian Crossan

2022 年《高科技、媒體及電信產業趨勢預測》(TMT Predictions) 報告中曾有探討 RISC-V 的章節，預測此類開源指令集架構 (Instruction Set Architecture, ISA)，在決定電腦中央處理器 (Central Processing Unit, CPU) 運作方式的規則與規範上，將會出現強勁的成長。報告預測，RISC-V 晶片的數量將自 2022 年起每年以倍增速度成長，至 2024 年相關收入可能接近 10 億美元。⁹⁴

國際產業組織 RISC-V International 在 2023 年底預估，2023 年 RISC-V 系統單晶片（System on a Chip，SoC）的出貨量為 10 億顆，到 2024 年可能會接近 20 億顆。⁹⁵ 如同預測，RISC-V 晶片的營收仍處於初期階段，預估 2024 年的收入不到 10 億美元。⁹⁶ 評估 2024 年全球晶片市場規模，預計收入將超過 6,110 億美元，目前 RISC-V 仍處於利基市場階段。⁹⁷ 長期來看，部分預測持樂觀看法，其中一項預測認為至 2030 年，RISC-V 將佔 SoC 市場的 25%，即超過 1,000 億美元。⁹⁸

儘管 RISC-V International 具有取得 SoC 市場重要份額的潛力，但此項技術仍存有國家安全風險疑慮。美國國會兩黨合作的美國與中國共產黨戰略競爭特設委員會議員，已督促商務部運用出口管制權力，對 RISC-V 技術移轉給中國大陸加以監管及限制，此舉目的為保護並促進美國及其盟國之間的 RISC-V 技術合作，同時維護雙方的國家安全利益。⁹⁹

為了瞭解其中的原因，我們需要重新回顧指令集架構（ISA）的概念、其應用範圍及與 AI 的關聯性。

指令集架構是 CPU 的核心，在許多筆記型電腦和大多數資料中心中，主要的 CPU ISA 為 x86 架構。而在大多數智慧型手機中，則是基於 RISC ISA 的 CPU 為主，實際上 x86 在筆記型電腦和資料中心的市占率仍較高。

生成式 AI 通常需要特殊的電腦運算晶片，以進行大型語言模型的訓練和推理。多數人可能已經瞭解圖形處理器（Graphics Processing Unit，GPU）在生成式 AI 運算中扮演的關鍵，其強大的運算能力可以處理這些模型需要的複雜運算。

儘管 GPU 負擔了運算的重任，但通常仍需要 1 到 2 顆 CPU 以協調資料流及執行其他協調工作，而 GPU 與 CPU 的搭配比例似乎正在變化，使 CPU 相對更重要。例如，輝達（Nvidia）推出的第一代生成式 AI 加速硬體配置為 8 顆 GPU 與 2 顆 CPU 配對（4:1 比例），並使用專用的 x86 指令集架構（ISA）。¹⁰⁰ 而至 2025 年初，輝達預計推出新的硬體模組，搭載 4 顆次世代 GPU 與 2 顆 CPU（2:1 比例）。¹⁰¹ 這些 CPU 與第一代不同，都是由輝達以 Arm ISA 為基礎自行設計。¹⁰²

美國政府針對智慧財產權、人才、晶片製造設備及特定晶片實施多項出口限制。2023 年初，美國限制了當時最先進的 GPU 出口（通常使用 10 奈米或以下的製程節點，並具備一定程度的晶片間數據傳輸速率），¹⁰³ 同年稍晚，美國政府擴大管制範圍，進一步限制 x86 ISA 和 Arm ISA 等先進 CPU 出口。¹⁰⁴

目前中國大陸無法進口或製造 7 奈米或以下先進製程節點的 GPU 或 CPU，而兩者皆為生成式 AI 訓練或推理不可或缺的元件。¹⁰⁵ 雖然目前 RISC-V 仍為利基市場 ISA，但在未來，下一代的生成式 AI 運算基礎設施可能會使用 RISC-V CPU 做為控制器。中國大陸仍需要製造先進 GPU，而美國限制 RISC-V 出口至中國大陸，則可能被視為目前美國限制措施的延伸。若美國政府未限制 RISC-V，則可能會成為現行 CPU 和 GPU 先進製程管制的漏洞或替代方案。此外，中國大陸正在探索其他領域的 RISC-V，例如開發使用於處理生成式 AI 的 CPU 控制器，也可能成為美國未來管制的對象。

結論

生成式 AI 硬體解決方案主要使用 GPU 和 CPU（或在同一個晶片上整合 CPU 核心的 GPU）。目前大多數 CPU 是以 x86 ISA 為基礎，其他則使用 Arm ISA，但是目前似乎尚未有將 RISC-V 做為主要 ISA 的應用（僅有部分將 RISC-V 核心執行其他功能）。因此，2025 年的出口限制可能將更專注於杜絕未來可能出現的漏洞。

RISC-V International 在 2020 年從美國遷移至瑞士，¹⁰⁶ 因此美國將如何運作或執行出口限制尚不明朗。報導指出美國政府正在評估潛在風險，並探討哪些行動可以因應潛在問題，避免損害已參與科技研發國際合作的美國企業權益。¹⁰⁷

儘管中國大陸在使用 RISC-V 上受到限制，值得注意的是長期以來中國大陸對 RISC-V 保持高度關注，並積極推動其發展，超過半數以上的 RISC-V International 的高級會員都是來自中國大陸（22 家中佔 12 家）。¹⁰⁸ 此外，截至 2022 年，在該年生產的 100 億顆 RISC-V 核心中，超過一半來自中國大陸。¹⁰⁹

中國大陸在 RISC-V 領域上已取得進展，並似乎處於有利的地位，同時值得關注的是，如若未來 RISC-V 受到限制，中國大陸將如何因應潛在負面影響。

By **Duncan Stewart**
Canada

Karthik Ramachandran
India

Prashant Raman
India

Jennifer Haskel
United Kingdom

Endnotes

1. Investing.com, “Earnings call: NVIDIA posts record revenue, bullish on data center growth,” Aug. 29, 2024.
2. Nitin Mittal et al., “Now decides next: Getting real about generative AI,” Deloitte’s State of Generative AI in the Enterprise Q2 report, April 2024.
3. Deloitte global analysis based on current market share of gen AI chip sales, combined with our survey data, and a US\$200 billion to US\$300 billion gen AI server market.
4. Jack Fritz et al., “Battle for the enterprise edge: Providers prepare to pounce on the emerging enterprise edge computing market,” Deloitte 2023 Global TMT Predictions, November 30, 2022; Chris Arkenberg et al., “Gaining an intelligent edge: Edge computing and intelligence could propel tech and telecom growth,” Deloitte 2021 Global TMT Predictions, December 7, 2020.
5. Sending to and from the cloud took too much time for real-time visual inspection of soft drink bottles on a high-speed line, for example, requiring AI decision-making on-prem, or at least within a few kilometers: milliseconds matter.
6. Dr. Lance B. Eliot, “Speeding up the response time of your prompts can be accomplished via these clever prompt engineering techniques,” Forbes, June 12, 2024.
7. Rowan et al., “Now decides next: Moving from potential to performance.”
8. See our related 2025 Global TMT Predictions chapter on “Gen AI, data centers, and the electricity grid.” Further, see: Datacenter BMO report, Communications Infrastructure, “1Q24 data center leasing: Records are made to be broken,” April 28, 2024; CBRE, [North America data center trends H2 2023](#), March 6, 2024.
9. Fifty-five percent of organizations reported avoiding certain generative AI use cases because of data- related issues. Top data-related concerns include using sensitive data in models and managing data privacy and security. To read further, see: Deloitte’s State of Generative AI in the Enterprise Q3 report, August 2024.
10. Organizations were much more worried about using sensitive data (for example, customer or client data) including IP. Using sensitive data in models (58% had at least a high level of concern), data privacy issues (58%), and data security issues (57%) were the top three issues that execs noted in the survey. To read further, see: Deloitte’s State of Generative AI in the Enterprise Q3 report, August 2024.
11. Andy Lees et al., “Financial services: Scaling gen AI for maximum impact,” Deloitte, accessed October 2024; Bijit Ghosh, “The rise of small language models—efficient & customizable,” Medium, November 26, 2023.
12. Jana Arbanas et al., [2024 media and entertainment outlook: Generative AI](#), Deloitte, 2024; Ghosh, “The rise of small language models—efficient & customizable.”
13. See our related 2025 Global TMT Prediction on the topic of “Gen AI in consumer devices.”
14. Deloitte analysis of publicly available specifications for various gen AI small servers, as well as conversations with companies globally.
15. Deloitte analysis of publicly available specifications for proposed rack-scale gen AI servers likely to be sold in the first half of 2025, although some may be available in limited quantities in late 2024.
16. Jennifer L., “The carbon countdown: AI and its 10 billion rise in power use,” Carboncredits.com, February 28, 2024.
17. Diana Goovaerts, “Could AI drive a new cloud repatriation wave?,” Fierce Network, June 6, 2024.
18. Ramesh Raskar, [Split Learning and Inference](#), MIT Media Lab’s Split Learning Project, accessed September 24, 2024.
19. Duncan Stewart et al., “Telecoms tackle the generative AI data center market,” Deloitte Insights, September 2024.
20. See the related 2025 TMT Prediction chapter on “Gen AI, data centers and the electricity grid.”
21. Jacqueline Davis, “Large data centers are more efficient, analysis confirms,” Uptime Institute, February 7, 2024.

22. Duke Robertson, "[What Is a hyperscale data center? Overview and comparisons](#)," Enconnex Blog, March 27, 2023.
23. Deloitte extrapolation based on 7.8 million as of Q1 2024, and 2–3 million additional subscribers for the balance of the year. Masha Abarinova, "[Fixed wireless continues to climb US broadband charts— Parks](#)," Fierce Network, June 13, 2024.
24. Paul Lee, Dieter Trimmel, and Eytan Hallside, "[No bump to bitrates for digital apps in the near term: Is a period of enough fixed broadband connectivity approaching?](#)," Deloitte Insights, November 29, 2023.
25. Zipy Fiber, "[Fiber internet in rural areas: When will it be here?](#)," January 18, 2024.
26. Abarinova, "[Fixed wireless continues to climb US broadband charts—Parks.](#)"
27. Gagandeep Kaur, "[Why 5G is failing to gain momentum in India](#)," Light Reading, April 25, 2024.
28. Naima Hoque Essing et al., "[Fixed wireless access: Gaining ground on wired broadband](#)," Deloitte Insights, December 1, 2021.
29. Ericsson, [Ericsson mobility report](#), June 2024.
30. Nick Ludlum, "[5G home broadband continues to bring real competition to cable](#)," CTIA Blog, January 31, 2024.
31. Jolanta Stanke, "[Global broadband subscriber growth in Q4 2023 slowest since 2019](#)," Point Topic, April 22, 2024.
32. Andrey Popov, "[5G FWA is a game-changer for broadband services in Italy](#)," Opensignal, May 16, 2024.
33. Deloitte calculation based on 2024 population and household estimates.
34. Deloitte analysis of publicly reported information from wireless network operators.
35. Jericho Casper, "[Fixed wireless subscriber growth solid in Q2](#)," Broadband Breakfast, August 5, 2024
36. Bob Wallace, "[Exploring telco enterprise fixed wireless access \(FWA\) services](#)," Network Computing, August 1, 2024.
37. Based on 2024 publicly reported business FWA additions from the major wireless providers, combined.
38. Robert Wyrzykowski, "[5G fixed wireless access \(FWA\) success in the US: A roadmap for broadband success elsewhere?](#)," Opensignal, June 6, 2024.
39. Ibid.
40. Ibid.
41. Abarinova, "[Fixed wireless continues to climb US broadband charts—Parks.](#)"
42. Ericsson, "[FWA is the largest 5G use case after mobile broadband](#)," February 2023.
43. Ericsson, [Ericsson mobility report](#).
44. It was initially anticipated that at least 200 operators would have launched standalone (SA) 5G networks by the end of 2023. However, as of March 2024, only 49 operators have successfully deployed 5G SA networks.
45. GSA, "[GSA Market Snapshot March-2024](#)," March 2024.
46. James Kirby, "[5G standalone networks: How vendors can accelerate adoption](#)," Analysys Mason, September 2023.
47. Based on lead author conversations with multiple operators globally between January and October 2024.
48. Deanna Darah, "[5G NSA vs. SA: How do the deployment modes differ?](#)," TechTarget, July 25, 2024.
49. Ibid; GSMA, [The 5G guide: A reference for operators](#), April 2019.
50. Darah, "[5G NSA vs. SA: How do the deployment modes differ?](#)"
51. Naima Hoque Essing et al., "[5G's promised land finally arrives: 5G standalone networks can transform enterprise connectivity](#)," Deloitte Insights, November 30, 2022.

52. Darah, [“5G NSA vs. SA: How do the deployment modes differ?”](#)
53. Ibid.
54. Philippe Poggianti and Pratik Das, [“It’s time for 5G to standalone,”](#) Qualcomm, July 6, 2023; Gavin Horn, [“6G foundry: Make the migration from 5G to 6G a rewarding experience,”](#) Qualcomm, May 14, 2024; Roger Billings, [“What is 5G standalone? 5G SA means network slicing, security, and automation,”](#) Ericsson Enterprise Wireless Blog, June 27, 2023.
55. GSMA, [The state of 5G 2024](#), February 2024; Sylwia Kechiche, [“Will 5G Advanced deliver on the 5G promise?”](#) Opensignal, April 4, 2024.
56. Andrew Wooden, [“The telecoms industry’s biggest problem? Failure to monetise 5G,”](#) Telecoms.com, March 14, 2024.
57. Mike Dano, [“A deeper dive into the 5G network API opportunity,”](#) Light Reading, April 1, 2024.
58. Global 6G Conference, [“Three 3GPP chairs clarify 6G standard release timeline at Global 6G Conference,”](#) April 23, 2024.
59. Communications Today, [“5G investment cycle tapering off and another one not on the horizon,”](#) March 30, 2024; Mary Lennighan, [“5G growth comes with new operator spending requirements—GSMA,”](#) Telecoms.com, February 29, 2024.
60. Caroline Gabriel, [“OpenRAN: Increased collaboration and a realistic view of timing will be needed to deliver the full benefits,”](#) Analysys Mason, April 2023.
61. Based on the authors’ review of public announcements and multiple conversations with network operators globally.
62. Naima Hoque Essing et al., [“The next-generation radio access network: Open and virtualized RANs are the future of mobile networks,”](#) Deloitte Insights, December 7, 2020.
63. Zineb Gdali, [“The future of mobile networks: Exploring the benefits of Open RAN 5G,”](#) Firecell, December 15, 2023.
64. Parallel Wireless, [Everything you need to know about Open RAN](#), 2020.
65. Essing et al., [“The next-generation radio access network: Open and virtualized RANs are the future of mobile networks.”](#)
66. TeckNexus, [“Current state of Open RAN—countries & operators deploying & trialing Open RAN,”](#) March 10, 2024.
67. O-RAN Alliance, [About](#) page, accessed October 2024.
68. Dan Jones, [“Dell’Oro: Don’t expect 5G-Advanced to fuel a RAN resurgence,”](#) Fierce Network, July 26, 2024.
69. Ray Le Maistre, [“Single vendor solutions to dominate Open RAN sales—Dell’Oro,”](#) TelecomTV, February 7, 2024.
70. Ibid.
71. Ibid.
72. Iain Morris, [“Telcos doubt open RAN challengers will have a role,”](#) Light Reading, March 15, 2024.
73. Dan Oliver, [“Open RAN: Everything you need to know,”](#) 5Gradar, July 23, 2021; Mike Dano, [“How American 5G operators learned to love open RAN,”](#) Light Reading, February 12, 2024.
74. Wes Davis, [“The US government makes a \\$42 million bet on open cell networks,”](#) The Verge, February 12, 2024.
75. Ibid.
76. Oliver, [“Open RAN: Everything you need to know.”](#)
77. David Debrecht, [“Building a smarter 5G future through Open RAN development,”](#) CableLabs, November 29, 2023.
78. Ibid.
79. Hosuk Lee-Makiyama and Florian Forsthuber, [“Open RAN: The technology, its politics and Europe’s response,”](#) European Centre for International Political Economy (ECIPE), October 2020.

80. Scott Buchholz et al., [“Quantum computing in 2022: Newsful, but how useful?”](#) Deloitte Insights, December 1, 2021; Duncan Stewart, [“Quantum computers: The next supercomputers, but not the next laptops,”](#) TMT Predictions 2019, Deloitte Insights, 2018, pp. 96–103.
81. Deborah Golden et al., [“Preparing the trusted internet for the age of quantum computing,”](#) Deloitte Insights, August 6, 2021.
82. Alex Wilkins, [“Useful quantum computers are edging closer with recent milestones,”](#) New Scientist, September 30, 2024.
83. Filipe Beato et al., [Transitioning to a quantum-secure economy](#), World Economic Forum, September, 2022.
84. Nancy Liu, [“Deloitte predicts 2024 will be a breakthrough year for post-quantum cryptography,”](#) SDxCentral, December 14, 2023.
85. National Institute of Standards and Technology (NIST), [“NIST releases first 3 finalized post-quantum encryption standards,”](#) August 13, 2024.
86. P.W. Shor, [“Algorithms for quantum computation: Discrete logarithms and factoring,”](#) Proceedings 35th Annual Symposium on Foundations of Computer Science (1994): pp. 135–42.
87. Executive Office of the United States President, [Report on post-quantum cryptography](#), July 2024.
88. FS-ISAC PQC Working Group, [Building cryptographic agility in the financial sector](#), October 2024.
89. John Potter, [“Deloitte: Companies face harvest now, decrypt later quantum threat,”](#) IoT World Today, September 27, 2022.
90. Apple Security Engineering and Architecture (SEAR), [“iMessage with PQ3: The new state of the art in quantum-secure messaging at scale,”](#) Apple Security Research Blog, February 21, 2024.
91. AWS, [“Post-quantum cryptography,”](#) accessed September 24, 2024.
92. Katherine Noyes, [“NIST’s postquantum cryptography standards: ‘This is the start of the race,’”](#) Deloitte’s CIO Journal for The Wall Street Journal, June 11, 2024.
93. Ibid.
94. Duncan Stewart et al., [“RISC-y business: Could open chip standard RISC-V gain traction against dominant incumbents?”](#) Deloitte Insights, December 1, 2021.
95. Ishika Setia, [“RISC-V: Projected growth to over 16 billion chips by 2030,”](#) TechnoSports, November 9, 2023.
96. As an example, large player SiFive had only US\$38 million in revenues in 2023, although it forecasts 2024 revenues of over US\$241 million. Mavis Tsai and Judy Lin, [“SiFive projects 6x growth in 2024 revenues, optimistic over RISC-V and AI demand,”](#) DigiTimes Asia, March 18, 2024.
97. WSTS, [WSTS semiconductor market forecast spring 2024](#), June 2024.
98. Belle Lin, [“Open-source chip design takes hold in Silicon Valley,”](#) The Wall Street Journal, December 14, 2023.
99. Stephen Nellis and Max A. Cherney, [“RISC-V technology emerges as battleground in US-China tech war,”](#) Reuters, October 7, 2023; Select Committee on the Chinese Communist Party (CCP), [“Gallagher, Rubio, Select Committee members urge Sec. Raimondo to safeguard American chip design from Chinese Communist Party threat,”](#) press release, November 2, 2023; Select Committee on the CCP, [“Reset, prevent, build: A strategy to win America’s economic competition with the Chinese Communist Party,”](#) December 12, 2023.
100. Nvidia, [“Nvidia DGX H100,”](#) accessed September 24, 2024.
101. Nvidia, [“Nvidia GB200 NVL72,”](#) accessed September 24, 2024.
102. Nvidia, [“Nvidia Grace CPU,”](#) accessed September 24, 2024.
103. Christie Simons et al., [2024 global semiconductor industry outlook](#), Deloitte, 2024.

104. Bureau of Industry and Security, “Public information on export controls imposed on advanced computing and semiconductor manufacturing items to the People’s Republic of China (PRC) in 2022 and 2023,” US Department of Commerce, November 6, 2023.
105. Alex He, “In the global AI chips race, China is playing catch-up,” Centre for International Governance Innovation, September 18, 2024.
106. Jacob Feldgoise, “RISC-V: What it is and why it matters,” Centre for Security and Emerging Technology (CSET), January 22, 2024.
107. Stephen Nellis, “US is reviewing risks of China’s use of RISC-V chip technology,” Reuters, April 23, 2024.
108. Sunny Cheung, “Examining China’s grand strategy for RISC-V,” The Jamestown Foundation China Brief 23, no. 23 (December 15, 2023): pp. 15–21.
109. Ma Si, “Open-source chip platform to cut reliance on proprietary tech,” China Daily, August 25, 2023.

Acknowledgements

Authors would like to thank **Hugo Pinto, Dan Littmann, Jack Fritz, Paul Lee, Itan Barmes, CasperStap, Ben Shapiro, Adnan Amjad, MarkNace, Emily Mossburg, Deborah Golden, Joe Mariani, Adam Routh, Ankit Dhameja, Zoe Burton,** and **Lizzie Tantam.**

Coverimage by: **Jaime Austin;** Getty Images, Adobe Stock

新興趨勢：值得關注的新興技術

《高科技、媒體及電信產業趨勢預測》(TMT Predictions) 探討了 AI 對網路防禦的影響、小晶片提升半導體良率、電信業者透過雲端系統進行現代化，以及矽光子技術在 AI 資料中心內的應用

2025 年全新推出一系列探討新興科技的短篇故事，這些趨勢仍處於早期使用階段，營收也較傳統趨勢預測的主題少，但預計成長速度將比平均更快，在未來一至兩年內將進入主流市場：

- 生成式 AI 與網路安全：高風險同時也帶來巨大機會
- 矽晶基本模組：小晶片 (Chiplet) 有望推動摩爾定律持續發展
- 電信運營支撐系統 (B/OSS)：電信業者將其業務與營運支援系統軟體進行現代化升級
- 矽光子技術：生成式 AI 以光速傳輸

生成式 AI 與網路安全：高風險同時也帶來巨大機會

隨著認知到生成式 AI 將帶來網路威脅，同時亦可強化安全防護，資安專家正積極探索如何駕馭其力量，以應對新興風險並強化技術環境

Arun Perinkolam、Sabthagiri Saravanan Chandramohan、Alison Hu 及 Duncan Stewart

AI (包括生成式 AI) 在日漸增加的網路威脅中扮演越來越關鍵的角色，根據 2024 年的調查顯示，71% 的美國州級首席資訊官將 AI 的威脅程度評為「非常高」或「稍高」。¹ AI 正在帶來一系列挑戰，包括法規變革、削弱現有解決方案的效能，以及具有 AI 的敵人，而企業部署 AI 的速度加劇了複雜性。

在 2024 年，利用生成式 AI 的網路攻擊已變得更為頻繁，成長為先前的兩至三倍。Deloitte 預測，在 2025 年其頻率將會持續上升。² 生成式 AI 可以透過多種方式應用於網路攻擊，其中一種方式為撰寫惡意釣魚郵件。數據顯示，截至 2024 年第一季，攻擊頻率相較 2023 年同期增加超過 856%，³ 顯示威脅行為者已經開始使用生成式 AI 工具編寫惡意軟體攻擊的代碼。⁴

但生成式 AI 也可用於正向面相，有助於防禦或緩減新一代 AI 的網路威脅。

業界人士則對於生成式 AI 的影響持有不同看法，儘管技術帶來許多優勢，也擔憂可能會增加網路風險，生成式 AI 作為一個新的攻擊媒介，進一步擴大了攻擊面。⁵ 生成式 AI 可以透過多種網路攻擊方式，例如生成複雜大量的文字型釣魚攻擊，以及製作深偽技術圖像與影片（如假冒 CEO 或其他公司高層主管），調查顯示過去一年有 61% 的受訪企業曾經遭遇深偽技術攻擊，其中 75% 為冒充高階主管身分。⁶ 為應對這一挑戰，多家生成式 AI 解決方案供應商正在採取防護機制，試圖防止工具被濫用於生成此類文本或影片攻擊。其中包括建立機制如嵌入數位浮水印，以偵測、標註或封鎖生成式 AI 影像或文本（請參閱 2025 年《高科技、媒體及電信產業趨勢預測》（TMT Predictions）中與浮水印和 AI 偵測有關的內容）。

目前眾多產業（以科技業為首），正擴大使用生成式 AI 編碼工具，以提升開發的效率及速度。⁷ 然而，生成式 AI 所生成的編碼可能存有安全問題。根據 2023 年底的調查顯示，超過半數（56%）的開發人員表示生成式 AI 的程式碼經常出現安全漏洞。⁸ 此外調查亦顯示，部分開發人員經常規避企業使用編碼工具的相關政策，並過於自信生成之代碼的安全性，而未必全面針對生成代碼進行安全檢測。⁹ 儘管存在安全風險，生成式 AI 編碼和大型語言模型仍在加速提升資安流程的成熟度與效率效能。例如在安全資訊與事件管理（Security Information and Event Management, SIEM）技術中，監控規則之自動生成，並在身分和存取管理（IAM）領域應用於存取工作流程、權限配置，以及第三方責任風險管理。¹⁰

目前，生成式 AI 和網路交叉領域的相關監管法規和地緣政治正在持續發展中。例如，歐盟《人工智慧法案》（EU AI Act）第 15 條明確規範高風險 AI 系統的網路安全問題。¹¹ 此外，自 2022 年起，各國針對 AI（尤其是生成式 AI）的多項技術實施了出口限制，包括用於訓練和推理運算需要的先進製程半導體、製造先進晶片的設備，以及此類晶片的設計工具。¹²

結論

隨著生成式 AI 在企業中逐步被應用整合，AI 解決方案供應商應持續專注為終端使用者打造安全的產品。不僅產品本身必須具備安全性，公司在與第四方（通常為大型語言模型服務供應商）共享本身或其他客戶資料時，也應該確保隱私與安全性。

由於歐盟《數位市場法案》（Digital Markets Act）、¹³《數位服務法案》（Digital Services Act），¹⁴ 和最新的《人工智慧法案》（AI Act）的實施，全球監管變得日益複雜。¹⁵ 科技企業不僅是 AI 的領先開發者，同時也是廣泛應用 AI 模型的核心推動者之一。因此，市場對科技企業在保障信任與安全性等方面抱有期望，要求在開發與銷售 AI 產品與解決方案時，發揮更大和更重要的作用。隨著全球風險上升，地緣政治衝突加劇，以及選舉與戰爭等事件影響，威脅行為者濫用生成式 AI 技術，將可能會成為 2025 年及未來企業在安全防禦及策略上的重要考量（請參見 2025 年《高科技、媒體與電信產業趨勢預測》（TMT Predictions）中與信任人工智慧和企業常用工具有關的內容）。

矽晶基本模組：Chiplet（小晶片）有望推動摩爾定律持續發展

Chiplet 技術預計可以為 AI 及高效能運算（HPC）提供更靈活、可擴充與高效的系統，同時提升整體良率。

Karthik Ramachandran、Duncan Stewart、Christie Simons 及 Dan Hamling

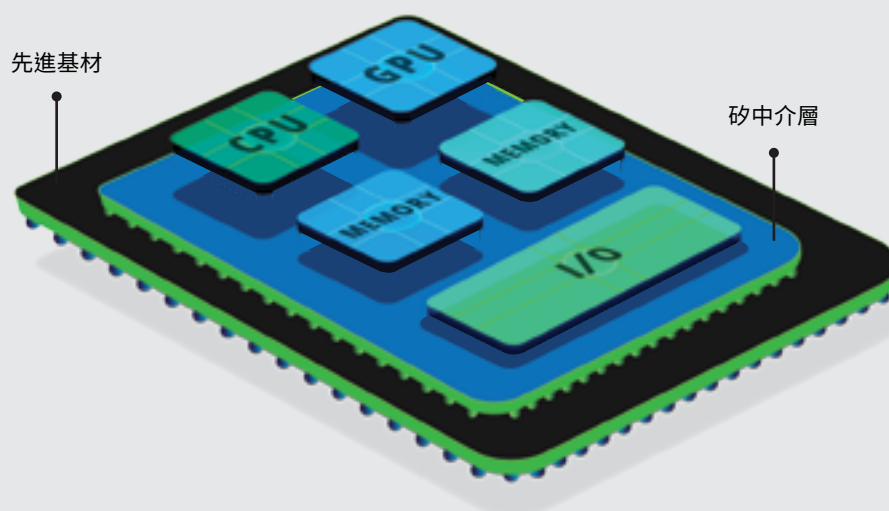
Deloitte 預測，全球基於 Chiplet 技術上的先進封裝市場營收，將從 2021 年的 70 億美元成長超過兩倍至 2025 年的 160 億美元。Chiplet 作為當今最先進的系統級封裝（System in Package，SiP）的技術模組，正在推動半導體產業發展。¹⁶ 相較於傳統依賴獨立晶片，並透過印刷電路板（Printed Circuit Board，PCB）互聯的架構，Chiplet 技術可以提供高速資料傳輸、降低延遲、優化 PPA（功耗、效能、晶片面積），甚至進一步延續摩爾定律（Moore's Law）。¹⁷ Chiplet 技術已廣泛應用於市場成長最快速的領域，包括 AI 加速器（尤其是生成式 AI）、高效能運算（High Performance Computing，HPC）及電信應用。

什麼是 Chiplet ？

小晶片（Chiplet）與傳統晶片（chip）有何差異？一般而言，晶片製造商使用一片 300mm 的矽晶圓（約 70,000mm²）製造大量的單裸晶，在封裝後即成為晶片。高階先進晶片尺寸通常在 20mmx20mm 左右（即 400mm²），因此每片 300mm 的晶圓上約有 175 顆裸晶。然而，小晶片（Chiplet）並非單片結構，而是使用異質整合（Heterogeneous Architecture），將多個小顆裸晶封裝在一起，使其能夠整合運作，類似於單裸晶。但這些裸晶和模組可能是來自於不同的晶片製造商。¹⁸

圖 1

Chiplet 是一種小型模組化積體電路，可與其他 Chiplet 組合形成更大的完整系統



資料來源：Deloitte 分析。

deloitte.com/insights

為什麼現在 Chiplet 技術如此重要？Chiplet 技術在 1980 年代就已經出現，但是在過去的 4 到 5 年間，因為先進製程節點對提升良率的需求提升，使 Chiplet 技術被高度關注，並出現大規模的轉變趨勢。¹⁹ 隨著產業逐步接近摩爾定律的物理極限，先進晶片的製造越發困難。然而，Chiplet 技術使半導體微型化得以推進，且基於系統級封裝（SiP）的晶片亦能提與傳統單裸晶系統單晶片（System-on-Chip, SoC）相當的效能。²⁰

隨著晶片尺寸縮小且更加複雜（例如 5nm 與 3nm 的先進製程節點），越有可能在 300mm 的晶圓上出現影響良率的缺陷率。²¹ 以 3nm 與 5nm 製程生產 800mm² 的裸晶為例，其缺陷密度為每平方公分 0.1 個缺陷；而使用傳統封裝方式下的組裝良率可能僅有 50% 到 55%。²² 相較而言，成熟半導體製程（如 90nm 和 130nm）的正常良率通常可達 90% 到 95% 之間。²³ 為了解決此問題，Chiplet 技術是將數顆較小且高良率晶片整合成較大的系統，例如將多個 180mm² 的小型裸晶（單顆良率約 95%）以小晶片架構封裝，在降低成本的同時，創造出更高效強大的 AI 處理器，並提升產品的功能靈活性和可配置性，以適應快速變化的市場需求。²⁴

隨著 Chiplet 技術使用逐漸成長，業界人士正在尋找具創新方法，以優化設計流程、提升連接速度與頻寬及提高能源效率。例如，業界正在研究數位分身（Digital Twins）技術，以逐步模擬與視覺化複雜的設計流程，包括靈活調整或更換小晶片的能力，以評估多晶片系統的效能。²⁵ 此外，部分企業已引進一系列互聯技術，用於組裝和堆疊離散元件以提升效率。²⁶ 目前業界正積極投入將玻璃作為封裝基板的研發，該材料在熱傳導性和功耗效能方面優於傳統基板，具備更高的靈活性和可擴充性，特別適用於高效能運算（HPC）和 AI 環境。²⁷ 同時，光子技術（Photonics）也正在被開發，作為互聯解決方案的可能性，透過光學輸入 / 輸出（I/O）技術以光訊號進行資料傳輸。特別適於高效能運算（HPC）和 AI 工作負載，可提供更高效能、低功耗的高速資料傳輸和運算（請參閱下篇文章，矽光子技術崛起的趨勢）。²⁸

儘管 Chiplet 架構帶來諸多優勢，仍須面臨其獨有的挑戰。例如，將多個晶片堆疊並透過薄型基板連接互聯，可能會引發熱管理問題，進而導致電路故障和功率損耗。²⁹ 此外，隨著將越來越多智慧財產權被整合至這些高度複雜的封裝中，而必須從不同區域的供應商採購元件，除了使供應鏈管理更加複雜外，也可能會增加網路攻擊風險，讓基礎系統暴露在潛在資安風險中。³⁰

結論

為了充分發揮 Chiplet 技術的商業價值，半導體供應鏈的參與者應攜手合作，共同解決技術挑戰與產業缺口，同時探索成長機會。

設備製造商、晶圓廠、整合元件製造商（IDM）、無廠 IC 設計業者及半導體封測代工廠（OSAT），可進一步強化從晶圓製造到封裝的合作關係，為共同開發而努力。此外，電源與邏輯 IC 製造商及設計團隊，應深入評估與熱管理有關的細節。³¹

企業應同時積極推動小晶片互連和資料互通的標準化，以進一步強化產業發展。例如，通用小晶片互連（Universal Chiplet Interconnect Express, UCIe）標準、高頻寬記憶體（High Bandwidth Memory, HBM）協議和束線（Bunch of Wires, BoW）互連技術等。³²

電子設計自動化（Electronic Design Automation，EDA）公司、晶片設計業者和資安專家可共同開發內建功能，以偵測小晶片層級潛在的智慧財產權竊取和網路侵權，並與供應鏈的其他層面合作，解決更廣泛的安全威脅與攻擊風險。此外，設計業者應與電子設計自動化及其他電腦輔助設計和電腦輔助工程公司合作，強化混合與異質整合的設計、模擬、驗證與驗證工具，探索應用 AI 技術至晶片設計。³³

電信運營支撐系統（B/OSS）：電信業者將其業務與營運支援系統軟體進行現代化升級

電信業者的後端業務與 B/OSS（營運支援系統）軟體市場成長緩慢，但正在透過使用軟體即服務（SaaS）和微型服務架構，並遷移至雲端等技術進行現代化，對軟體供應商來說是重要的成長機會，同時也為電信業者帶來提供更多 5G、光纖和 AI 應用的可能。

Amit Kumar Singh、Duncan Stewart、Hugo Santos Pinto 及 Dan Littman

以歷史觀點來看，電信業者長期擁有兩套獨立且關鍵的 IT 系統。業務支援系統（Business Support System，BSS）主要負責客戶訂單處理、客戶關係管理及計費，而營運支援軟體（Operations Support Systems，OSS）則處理服務訂單管理、網路資源管理和網路營運。³⁴ 過去 BSS 和 OSS 多為獨立運作的客製化系統，主要部署在地端並以硬體為核心架構。系統主要由特定服務（固定網路與行動網路）或技術領域（如存取、核心和傳輸）打造的專用解決方案組成，整體基礎建設呈現高度碎片化又複雜的特性。³⁵ 預計 2025 年及未來，電信業者將推動系統的現代化，透過更高層級的自動化與智能技術，促使產業加速成長。從長遠來看，BSS 和 OSS 甚至可能整合至單一平台，以提升營運效率與靈活性。

為何此變革正在發生？隨著隨選服務和產品的普及，企業需要重新思考客戶體驗、重新定義產品、調整商業模式，並重新整合銷售通路。在 BSS 領域，尤其是帳務管理方面，客戶期望不斷提升，及新興數位收入來源的出現，將需要更具彈性的功能與支援，以聚焦產品為中心和客戶導向的帳務管理模式，此種轉變將可能會影響整個 B/OSS 生命週期。

Deloitte 的預測與分析機構一致，預計 OSS 和 BSS 的整合市場（即 B/OSS）的全球營收將從 2023 年的 630 億美元，在 2025 年成長至 700 億美元，年成長率 5%。³⁶ 儘管如此，電信業者仍希望透過由 5G 獨立組網與光纖技術等新興領域，拓展潛在新營收機會。此外，舊有基礎建設的維護成本亦是推動 B/OSS 軟體現代化的重要因素。電信業者在整合與客製化舊有 B/OSS 系統方面的支出，最高可達 IT 預算的 80%，這將促使業者加快 B/OSS 軟體現代化的腳步³⁷。預估 B/OSS 軟體即服務（SaaS）產品市場的年成長率約為 18%，而雲端化的年成長率則預計達 21%，顯示 BSS 和 OSS 現代化相關領域的成長率為整體 BSS 和 OSS 軟體市場的 5%，年成長率約三至四倍。³⁸

隨著現代化腳步加速，BSS 與 OSS 可透過 API 和微型服務更有效地整合，進一步標準化配置和模組化的雲端或軟體解決方案。預期這些整合可以為電信業者帶來更高的營運效益，包括降低成本、創造新營收來源、提升網路韌性、加強網路與營運安全性。隨著未來幾年生成式 AI 技術的應用擴展，將可以有效運用此類技術，升級營運與服務能力。未來 OSS 的發展可能將以「服務為中心」為核心，負責以個別服務層級進行資源協調配置、服務執行與品質保障，而非依據傳統個別技術領域進行管理。

近年來，許多歐洲電信業者已開始推動 B/OSS 現代化，並且已經獲得一定財務收益，³⁹ 然而，未來新部署的目標可能將以「服務為中心」為模式架構。預計未來幾年，B/OSS 的成長動能將來自於美國、中東、北非和新興亞太地區。⁴⁰ 此外，近期 BSS，尤其是客戶互動系統內，正加速向雲端遷移。而 OSS 的轉型較緩慢，可能是因為電信業者對於將重要功能遷移至新系統的謹慎態度，但此趨勢似乎正在改變。⁴¹

結論

市場可能將面臨「由誰提供這些新服務」的問題。傳統上 BSS 或 OSS 由不同的解決方案供應商和系統整合商提供，或由企業自行開發內部解決方案。隨著 B/OSS 的現代化發展，成熟的企業軟體供應商與超大規模企業（Hyperscalers）正在嘗試推出自家 B/OSS 解決方案，⁴² 成功的關鍵在於，以雲端和 AI 為核心發展思維，並將現代化的系統整合。

舊有財務系統中每年處理的資金規模達數十億美元，因此財務系統轉型作為 B/OSS 現代化的一部份，可能會產生重大影響。部分電信業者高階主管亦對於財務系統轉型抱持謹慎態度，擔憂可能會使收入穩定性面臨風險。⁴³ 在推動財務升級時，如何平衡策略以維持穩定並確保符合業務目標，將業務中斷的風險降至最低，將備受挑戰。關於應對複雜考量與方案，可參閱《財務轉型複雜度探討》（Navigating the complexities of billing transformation）一文中更有深入的探討。⁴⁴

在進行 B/OSS 現代化的過程中，電信業者應同時投入資金，不僅有助於降低成本，也可能創造新的營收。降低成本是推動現代化商業模式中的關鍵因素，而透過提供網路即服務（NaaS）或融合產品等新服務模式，電信業者將可進一步提升 FWA（固定無線存取）服務的獲利能力，實現更大的市場價值。

此外，B/OSS 現代化可能涉及多項調整，包括整合 OSS 和 BSS、使用產業標準 API、⁴⁵ 強化 DevOps 以控制成本，並運用新興 AI 和機器學習技術，以提升系統效能。

隨著電信業者從分散的 B/OSS 傳統環境轉向更具整體性的架構，治理模式亦需要同步調整。B/OSS 傳統上屬於工程領域，但在現代化過程中應納入人資、資訊科技和財務部門等作為新利害關係人，確保跨部門協作並協同運作。

矽光子技術：生成式 AI 以光速傳輸

在生成式 AI 高效能需求驅動下，矽基光學元件正在從研究實驗室走向資料中心，成為業界關注的焦點

Duncan Stewart、Karthik Ramachandran、Jeroen Kusters 及 Christie Simons

Deloitte 預測，做為光纖收發器的矽光子晶片的銷售額，將從 2023 年的 8 億美元成長至 2025 年的 12.5 億美元，年均複合成長率為 25%。⁴⁶ 儘管這些晶片僅占 2026 年全球預估銷售額 6,870 億美元的一小部分，⁴⁷ 但這些晶片可支援生成式 AI 資料中心以更快的速度傳輸大量數據。相比傳統方案，矽光子晶片具備更小尺寸、更低成本、更少能耗及更優異的散熱管理能力，使資料中心能夠以光速進行通訊，提升整體效能與效率。⁴⁸

矽晶片在電氣領域中運作時，通常是透過電線以電訊號與其他晶片傳輸，或透過光纖電纜連接的外接雷射或調節器，以傳輸光子。光纖的頻寬通常較銅線高，且能以更低耗能進行更遠距離的信號傳輸。此外，光纖電纜不受影響，而銅線則反之。同時光纖電纜較銅線更難以接入或攔截，安全性上具優勢。然而，傳統光子技術仍面臨成本和尺寸的限制，而矽光子技術則被視為能克服這些障礙的解決方案。⁴⁹

預計至 2025 年，將有更多的光子裝置使用下列方式製造：

- 使用與多數電子晶片相同的材料—矽
- 使用與多數電子晶片相同的基板材料—矽
- 使用與多數晶片相同的製造技術
- 利用矽光子涵蓋設計、製造、代工、測試、封裝及組裝面向成熟的生態系統，並確保其與現有矽晶片製造生態系統相容

這些方式將使晶片公司能將電子和光子元件整合至單一晶片上，隨著時間推移，將能在許多場景中應用。預計至 2025 年，矽光子技術使用的主要動力將來自於資料中心應用，尤其是執行生成式 AI 的訓練與推理運算。目前大多數資料中心內的晶片、機架與托盤間的傳輸速度通常低於 100G（每秒 100Gb），但生成式 AI 設備必須更快速的傳輸資料，所需速率可達 400G 甚至 800G，而光子技術將是最佳解決方案。⁵⁰

在生成式 AI 資料中心內，伺服器機架的規格至關重要。標準尺寸為寬 24 英吋（600mm）、深 42 英吋（1066.80mm）高 73.6 英吋（1866.90mm），此種尺寸通常稱為 42U 機架（1U 相當於高 1.75 英吋（44.45 mm））。⁵¹ 不同類型的晶片與機架需要在不同的距離下，以不同的速度相互傳輸，而距離與速度在一定程度上取決於機架尺寸。

因此，矽光子技術在 2025 年及未來充滿了發展機會。不同技術在不同的傳輸距離範圍內具有最佳應用場景，主要是受到元件間距離的影響。矽光子技術的最佳應用範圍在 10 公分到 10 公尺間，既不太短也不會太長，在此範圍內，相較於銅線或傳統光子技術，矽光子技術在短期內更具有優勢，可能將創造更多營收機會。

承載盤上的晶片間傳輸：生成式 AI 機架規模伺服器的配置之一，為每個承載盤（Tray）搭載 2 顆 GPU 和 1 顆 CPU 的承載盤，根據高度所選擇的冷卻技術可以是 1U 或 2U。2025 年，承載盤內晶片間的傳輸（距離小於 10 公分）是以電氣訊號方式進行，但未來可能會逐步發展光學方式傳輸。由於可用空間有限（高度僅 2 至 4 英吋），且有低成本訴求，將可能需要使用整合式矽光子技術，而非離散的光子元件。不過，由於距離極短，在 2025 年電氣訊號仍能應對需求。

機架上的承載盤間傳輸：在單一伺服器機架上，最多可配置 18 個上述 1U 承載盤，這是最密集的配置方式。每個承載盤必須在不到 1 或 2 公尺的垂直距離相互傳輸。⁵² 根據預測，在 2025 年初光學傳輸方式在應用中的成本是每個機架 14.4 萬美元。⁵³ 預計 2025 年下半年或 2026 年初，矽光子裝置可能在此領域將獲得更廣泛的應用。

近距離的機架對機架傳輸：由於電力、冷卻、成本等各種因素，將可能會出現許多密度僅為原先一半、彼此相鄰放置，且需要在 1 到 2 公尺距離內通訊的成對伺服器機架。在未來某個時間點，幾乎完全可以使用光學方式進行兩個伺服器機架間的傳輸，這可能是矽光子技術在 2025 年的最具潛力的應用機會。

遠距離的機架對機架傳輸：各伺服器機架（或成對的伺服器機架）必須與整個超大規模資料中心內的所有其他機架伺服器（包括各類記憶體和處理器）進行傳輸，這些傳輸可能會使用長數十公尺至上百公尺的光纖電纜。矽光子技術能提供頻寬極高和長距離的傳輸能力，並透過高度整合的光學裝置降低成本與功耗。⁵⁴ 但成本較高是考量因素之一，預期在短期內，矽光子技術仍不會取代傳統光子技術的地位。

結論

另一項關於矽光子技術的預測為併購活動（M&A），若生成式 AI 資料中心的數量持續增加，尤其是在高速傳輸和降低功耗方面的需求（兩者都有可能實現）。隨著日益發展，技術在市場中逐漸備受重視，大型企業可能將投入數十億美元來收購矽光子新創公司、相關企業或其他公司在該領域中較卓越的相關部門。⁵⁵

儘管本文主要為探討生成式 AI 資料中心在加速矽光子技術需求方面的重要性，但值得注意的是，此項技術在其他應用領域中的潛在影響力，近期最受關注的機會可能是應用於晶片級 LIDAR，為高級駕駛輔助系統（短期）以及自動駕駛功能（長期）提供技術支援。⁵⁶

By **Duncan Stewart**
Canada

Karthik Ramachandran
India

Prashant Raman
India

Endnotes

1. Srin Subramanian and Meredith Ward, “[2024 Deloitte-NASCIO Cybersecurity Study](#),” Deloitte Insights, Sept. 30, 2024.
2. The Deloitte authors make this prediction based on what they are seeing in the market and what their clients are telling them.
3. Duncan Riley, “[Generative AI services have driven a huge surge in phishing attacks](#),” Silicon Angle, May 22, 2024.
4. Michael Crider, “[Hackers are now using AI-generated code for malware attacks](#),” PCWorld, Sept. 25, 2024.
5. Tiernan Ray, “[Generative AI is new attack vector endangering enterprises, says CrowdStrike CTO](#),” ZDNet, June 30, 2024.
6. Ian Barker, op. cit.
7. Faruk Muratovic, Duncan Stewart, and Prashant Raman, “[Tech companies lead the way on generative AI: Does code deserve the credit?](#),” Deloitte Insights, Aug. 2, 2024.
8. Snyk, [2023 Snyk AI-generated code security report](#), accessed Aug. 11, 2024.
9. Ibid.
10. Mandy Andress, “[Generative AI for cybersecurity: Is it right for your organization?](#),” Fast Company, June 17, 2024.
11. EU Artificial Intelligence Act, “[Article 15: Accuracy, Robustness and Cybersecurity](#),” accessed Aug. 28, 2024.
12. Christie Simons et al., [2024 global semiconductor outlook](#), Deloitte, Jan. 22, 2024.
13. European Commission, “[The Digital Markets Act: Ensuring fair and open digital markets](#),” accessed Oct. 6, 2024.
14. European Commission, “[The Digital Services Act](#),” accessed Oct. 6, 2024.
15. European Commission, “[AI Act](#),” accessed Oct. 6, 2024.
16. Deloitte analysis based on data and chart presented in “[Semiconductor – A treasure trove for private equity investors](#),” (June 2024, p. 11). We used chiplet packaging baseline market share for 2021 (24% of the US\$30 billion total market) and applied 22% CAGR to arrive at the 2025 predicted value of US\$16 billion.
17. Moore’s Law notes that the number of transistors on an integrated circuit would double every two years but with a smaller increase in cost—translating into nearly twice the superior performance over the previous generation (because of doubling of transistors by shrinking the linewidths) at a marginal additional cost. However, shrinking transistor linewidths is reaching its physical limit. To read further, see: Alchip’s “[Moving from SoCs to chiplets could help extend Moore’s Law](#),” Sept. 26, 2022.
18. Deloitte’s analysis of multiple publicly available sources including product information published by chip companies, as well as articles from sources such as EE Journal, Semiconductor Engineering, EE Times.
19. Based on our analysis of chiplet-based new product announcements and launches from major semiconductor companies (including IDMs, fabless, and chip design players). Chiplets allow multiple functionalities such as GPU, CPU, and memory components to be densely packed on a single chip. Moreover, chiplets have helped deal with the complexity involved in integrating the diverse components with varying manufacturing and packaging technologies coming in from IDMs, foundries, and other component manufacturers from various regions worldwide. To read further, see: Dr. Uwe Lambrette et al., “[Semiconductor – A treasure trove for private equity investors](#),” Deloitte, June 2024.
20. TrendForce, “[News] [Understanding 3DIC, heterogeneous integration, SiP, and chiplets at once](#),” March 19, 2024; Alchip, “[Moving from SoCs to chiplets could help extend Moore’s Law](#).”
21. Max Maxfield, “[Are you ready for the chiplet age?](#),” EE Journal, July 27, 2023.
22. Yinxiao Feng and Kaisheng Ma, “[Chiplet actuary: A quantitative cost model and multi-chiplet architecture exploration](#),” Institute for Interdisciplinary Information Sciences (Tsinghua University, China), April 9, 2024.

23. [“Test & reliability challenges in advance semiconductor geometries”](#) (presentation at 2013 Semiconductor Wafer Test Conference), June 9, 2013. Data for 130 nm and 90 nm based on the chart on page 22 titled “Dramatic rise in systematic yield issues.”
24. Deloitte analysis based on our conversations with subject matter experts in the areas of advanced packaging, as well as data and research from publicly available sources, including Feng et al., [“Chiplet actuary: A quantitative cost model and multi-chiplet architecture exploration”](#); Maxfield, [“Are you ready for the chiplet age?”](#)
25. Ann Mutschler, [“Digital twins gaining traction in complex designs,”](#) Semiconductor Engineering, June 27, 2024.
26. Eric Beyne, [“Chiplet interconnect technology: Piecing together the next generation of chips,”](#) 3D InCites, July 3, 2024.
27. Bilal Hachemi, [“Glass Core substrates: The new race for advanced packaging giants,”](#) Yole Group, June 17, 2024; Anton Shilov, [“Intel’s glass substrates advancements could revolutionize multi-chiplet packages,”](#) Tom’s Hardware, Sept. 18, 2023.
28. See section Silicon photonics: Gen AI communicates at light speed.
29. Karen Heyman, [“Thermal challenges multiply in automotive, embedded devices,”](#) Semiconductor Engineering, July 2, 2024.
30. Saman Sadr and Richard Lin, [“Securing the new frontier: Chiplets & hardware security challenges,”](#) Universal Chiplet Interconnect Express, Feb. 7, 2024; Nitin Dahad, [“Chiplets are the latest buzz, but many challenges lie ahead,”](#) Embedded, March 10, 2024.
31. Thermal and heat management are noted as one of the major roadblocks to commercializing 3D ICs. To read further, see: Brian Bailey, [“Why there are still no commercial 3D-ICs,”](#) Semiconductor Engineering, Jan. 29, 2024.
32. As noted in [Deloitte Global’s 2024 semiconductor industry outlook](#), not only the traditional OSATs but even major IDMs, foundries, fabless companies, EDA vendors, and startups are making the moves and ramping up solutions based on chiplets architectures to push the bar on advanced packaging technologies. Also, see: Ann Mutschler, [“Chiplet IP standards are just the beginning,”](#) Semiconductor Engineering, March 6, 2024; Majeed Ahmad, [“A sneak peek at chiplet standards,”](#) EDN, Sept. 4, 2023.
33. Ann Mutschler, [“Chip design digs deeper into AI,”](#) Semiconductor Engineering, June 3, 2024.
34. Andrew Wooden, [“The evolution of BSS and OSS in the telecoms sector,”](#) Telecoms.com, Aug. 15, 2023.
35. Ibid.
36. Deloitte analysis, based on Alex Bilyi, [“CSPs’ spending on telecoms-related OSS/BSS software and services will reach USD80 billion by 2028,”](#) Analysys Mason, Nov. 13, 2023.
37. Nia Batten, [“The hidden costs of legacy tech,”](#) Data Centre Review, Sept. 1, 2023.
38. Deloitte analysis, based on Alex Bilyi, [“CSPs’ spending on telecoms-related OSS/BSS software and services will reach USD80 billion by 2028.”](#)
39. Chris Silberberg and Chris Barnard, [“How telcos are transforming in Europe: Technology, services and customers,”](#) IDC, Sept. 1, 2022.
40. Deloitte analysis, based on Alex Bilyi, [“CSPs’ spending on telecoms-related OSS/BSS software and services will reach USD80 billion by 2028.”](#)
41. Mark Mortensen, Andy He, and John Abraham, [Market pulse: Digital transformation of BSS/OSS to the cloud & DevOps,](#) Analysys Mason, Jan. 2018.
42. Ryan, [“OSS/BSS in the clouds,”](#) Passionate about OSS, July 20, 2020; Anjali Mishra, [“OSS/BSS market players are building robust systems to support next-gen networks,”](#) Global Market Insights (GMI), May 6, 2022.
43. Amit Kumar Singh et al., [“Navigating the complexities of billing transformation,”](#) Deloitte Insights, 2024.
44. Ibid.
45. TM Forum, [“Introduction to Open APIs,”](#) accessed Sept. 24, 2024; GSMA, [“GSMA Open Gateway API descriptions](#)

46. Deloitte analysis and interpolation of Light Trends Newsletter, "[Sales of silicon photonics chips will reach \\$3 billion by 2029](#)," LightCounting, May 2024.
47. WSTS, "[WSTS Semiconductor Market Forecast Spring 2024](#)," press release, June 4, 2024.
48. Adam Carter, "[Silicon photonics key to unlocking AI's full potential](#)," EE Times, Aug. 18, 2023.
49. Deloitte analysis based on publicly available third-party sources, including Karen Heyman, "[Transitioning to photonics](#)," Semiconductor Engineering, April 13, 2023; Maxime Fazilleau, "[What makes optical fibre immune to EMI?](#)," Tiny Green PC, Jan. 23, 2017.
50. FiberStamp, "[Driving the future of high-speed data transfer: The role of PAM4 and silicon photonics in the age of AI](#)," Medium, Nov. 8, 2023.
51. Christopher Tozzi, "[A guide to server rack sizes for data centers](#)," Data Center Knowledge, Jan. 8, 2024.
52. Mary Zhang, "[Data center racks, cabinets, and cages: An in-depth guide](#)," Dgtl Infra, Sept. 28, 2023; Tozzi, "[A guide to server rack sizes for data centers](#)."
53. Dylan Patel and Daniel Nishball, "[Nvidia's optical boogeyman – NVL72, Infiniband Scale Out, 800G & 1.6T Ramp](#)," SemiAnalysis, March 25, 2024.
54. M. Duranton, D. Dutoit, and S. Menezo, "[3 - Key requirements for optical interconnects within data centers](#)," in Optical Interconnects for Data Centers, Tolga Tekin et al. (eds) (Sawston, UK: Woodhead Publishing, 2017), pp. 75–94.
55. Contributions from Deloitte subject matter specialists in July and August 2024.
56. Eric Walz, "[Stellantis invests in lidar startup SteerLight](#)," Automotive Dive, April 2, 2024.

Acknowledgements

The authors would like to thank **Jack Fritz, John Levis, Stephen Winsor, Sandy Lawrence-Morgan, Essaki Velusami, Kannan Ramakrishnan, Nina Zhang, Gautham Dutt**, and **Dan Hamling** for their contributions to this article.

Cover image by: **Jaime Austin**; Getty Images, Adobe Stock

About Deloitte

Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited, a UK private company limited by guarantee ("DTTL"), its network of member firms, and their related entities. DTTL and each of its member firms are legally separate and independent entities. DTTL (also referred to as "Deloitte Global") does not provide services to clients. In the United States, Deloitte refers to one or more of the US member firms of DTTL, their related entities that operate using the "Deloitte" name in the United States and their respective affiliates. Certain services may not be available to attest clients under the rules and regulations of public accounting. Please see www.deloitte.com/about to learn more about our global network of member firms.

聯絡我們

勤業眾信高科技、媒體及電信產業服務團隊

陳明輝 資深會計師 Gordon Chen

高科技、媒體及電信產業負責人

gordonchen@deloitte.com.tw

簡宏偉 資深執行副總經理 Howard Jyan

科技與轉型服務

hjyan@deloitte.com.tw

溫紹群 資深執行副總經理 Rick Wen

電信、媒體及娛樂產業負責人

rickswen@deloitte.com.tw

陳明輝 專案執行副總經理 Ming Chen

科技與轉型服務

minghchen@deloitte.com.tw

朱孝甫 資深執行副總經理 Sam Chu

運動產業負責人

samhchu@deloitte.com.tw

陳宥嘉 資深會計師 Arthur Chen

稅務服務

arthurchen@deloitte.com.tw

林彥良 資深執行副總經理 Max Lin

科技與轉型服務

maxylin@deloitte.com.tw

朱光輝 資深會計師 Allen Chu

稅務服務

allenkchu@deloitte.com.tw

專案聯絡

謝依倫 Milly Hsieh

高科技、媒體及電信產業專案組長

mhsieh@deloitte.com.tw

張芷瑄 Shannon Chang

高科技、媒體及電信產業專案組長

shannchang@deloitte.com.tw

Deloitte.

Insights

Sign up for Deloitte Insights updates at www.deloitte.com/insights

About Deloitte Insights

Deloitte Insights publishes original articles, reports and periodicals that provide insights for businesses, the public sector and NGOs. Our goal is to draw upon research and experience from throughout our professional services organization, and that of coauthors in academia and business, to advance the conversation on a broad spectrum of topics of interest to executives and government leaders.

Deloitte Insights is an imprint of Deloitte Development LLC.

About this publication

This publication contains general information only, and none of Deloitte Touche Tohmatsu Limited, its member firms, or its and their affiliates are, by means of this publication, rendering accounting, business, financial, investment, legal, tax, or other professional advice or services. This publication is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your finances or your business. Before making any decision or taking any action that may affect your finances or your business, you should consult a qualified professional adviser. None of Deloitte Touche Tohmatsu Limited, its member firms, or its and their respective affiliates shall be responsible for any loss whatsoever sustained by any person who relies on this publication.

About Deloitte

Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited, a UK private company limited by guarantee ("DTTL"), its network of member firms, and their related entities. DTTL and each of its member firms are legally separate and independent entities. DTTL (also referred to as "Deloitte Global") does not provide services to clients. In the United States, Deloitte refers to one or more of the US member firms of DTTL, their related entities that operate using the "Deloitte" name in the United States and their respective affiliates. Certain services may not be available to attest clients under the rules and regulations of public accounting. Please see www.deloitte.com/about to learn more about our global network of member firms.



Deloitte 泛指 Deloitte Touche Tohmatsu Limited (簡稱"DTTL"), 以及其一家或多家會員所網絡及其相關實體 (統稱為"Deloitte 組織")。DTTL (也稱為"Deloitte 全球") 每一個會員所及其相關實體均為具有獨立法律地位之個別法律實體, 彼此之間不能就第三方承擔義務或進行約束。DTTL 每一個會員所及其相關實體僅對其自身的作為和疏失負責, 而不對其他行為承擔責任。DTTL 並不向客戶提供服務。更多相關資訊www.deloitte.com/about 了解更多。

Deloitte 亞太 (Deloitte AP) 是一家私人擔保有限公司, 也是 DTTL 的一家會員所。Deloitte 亞太及其相關實體的成員, 皆為具有獨立法律地位之個別法律實體, 提供來自100多個城市的服務, 包括: 奧克蘭、曼谷、北京、邦加羅爾、河內、香港、雅加達、吉隆坡、馬尼拉、墨爾本、孟買、新德里、大阪、首爾、上海、新加坡、雪梨、台北和東京。

本出版物係依一般性資訊編寫而成, 僅供讀者參考之用。Deloitte 及其會員所與關聯機構不因本出版物而被視為對任何人提供專業意見或服務。在做成任何決定或採取任何有可能影響企業財務或企業本身的行動前, 請先諮詢專業顧問。對於本出版物中資料之正確性及完整性, 不作任何 (明示或暗示) 陳述、保證或承諾。DTTL、會員所、關聯機構、雇員或代理人均不對任何直接或間接因任何人依賴本通訊而產生的任何損失或損害承擔責任或保證 (明示或暗示)。DTTL 和每一個會員所及相關實體是法律上獨立的實體。

