Photo Credit: Shutterstock.com

深度偽造技術的演進 AI 驅動的詐騙新挑戰

◎ 石艾／法務部調查局

深度偽造技術（Deepfake，以下簡稱深偽技術）是指利用人工智慧（AI）和深度學習演算法來生成或修改聲音、影像與影片，使其呈現出近乎真實的效果。這項技術最初應用於娛樂與創意產業，例如電影特效和虛擬角色製作，為內容創作帶來新的可能。然而，隨著技術的進步與普及，深偽技術的濫用也引發了諸多社會問題，在詐騙、假新聞、隱私侵犯等方面帶來新的威脅。深度偽造已成為一把雙面刃：一方面拓展了創新應用，另一方面也成為不法分子手中的新型犯罪工具。

深偽技術的發展歷程與技術細節

一、 技術起源與演進

深偽技術的起源可追溯至人工智慧與

伊恩．古德費洛（Ian Goodfellow）
Photo Credit: https://commons.wikimedia.org/w/index.php?curid=79321899

深偽技術利用深度學習對視聽資料進行高度擬真的合成。Photo Credit: https://www.shutterstock.com

深度學習的快速發展。2014年，生成對抗網絡（Generative Adversarial Networks, GANs）的問世是深偽技術的重要里程碑。GAN由伊恩．古德費洛（Ian Goodfellow）提出，它採用生成器與判別器兩個神經網絡互相競爭的架構，使生成器不斷學習並模仿真實資料的特性，進而產生高品質、以假亂真的內容。這種模型最初被用於影像生成和修復，如改善老照片品質等。隨著技術門檻降低和開源工具出現，越來越多開發者開始探索GAN在其他領域的應用潛力，包括影片與聲音的生成。

2017年，第一個專注於人臉交換的深偽應用問世，正式將「深度學習」與「偽造」結合，掀起了廣泛關注。所謂Deepfake技術，透過深度學習模型高度擬真地生成或替換人像與聲音，在短時間內即可製作出真假難辨的影片。這項技術一開始為娛樂產業帶來創意突破（例如讓已故演員「重現」銀幕），但同時也為假新聞與詐欺活動提供了強大的工具，凸顯其潛在風險。

二、 核心技術特徵

現代深偽技術Deepfake本質上是利用深度學習對視聽資料進行高度擬真的合成。其主要特徵包括：

- **面部交換**：將一個人的臉部特徵無縫地替換到另一段影片中的人物臉上，使影片中的人物看起來彷彿變成了另一人。
- **語音模仿**：提取目標人物的聲紋數據，生成高度擬真的語音，甚至可模仿說話者的語氣和情感，使聽者難辨真偽。
- **表情同步**：將來源影片中人物的面部表情精確地映射到目標影片的人物上，實現嘴型、表情與動作的同步匹配，使合成影片更加逼真自然。

上述效果的實現很大程度上倚賴GAN架構的支撐。生成對抗網絡透過生成器與判別器的對抗訓練，使生成器逐步提升造假能力。生成器不斷產生逼真的假影像/聲音，判別器則不斷學習識別真假。經過多輪訓練後，生成器的輸出幾乎能騙過判別器，達到以假亂真的地步。這種對抗式學習機制正是深偽內容能高度亂真的技術關鍵。

深偽技術的應用場景

深偽技術因其高度擬真性，而在各領域展現出廣泛的應用前景，包括正面與負面的場景：

- **娛樂與創意產業**：要說最吸睛的應用非好萊塢電影莫屬。2016年《星際大戰外傳：俠盜一號》便運用深度學習技術，讓已故38年的英國影星彼得．庫欣「數位復活」，製作團隊分析他生前32部電影、超過200小時影像素材，連嘴角微表情和英式腔調都精準還原。這項技術後來更應用在《曼達洛人》影集，以8K畫質重現年輕版天行者路克，連資深影迷都看不出破綻。臺灣PTT電影版當時掀起熱烈討論，有鄉民直呼：「根本是穿越時空的演技！」
- **教育與訓練**：深偽技術可用於製作擬真的教學內容。例如，在語言學習中生成具不同口音與語調的示範語音，讓學生練習聽說能力；在專業培訓中模擬醫療手術或飛行駕駛等場景，提供學習者高度逼真的模擬訓練環境，提升教學效果。
- **商業與營銷**：企業可利用深偽技術進行個性化行銷，生成貼合用戶喜好的虛擬代言人或廣告內容，以提高廣告的精準度與吸引力。品牌行銷方面，深偽技術讓製作高品質的宣傳影片成本降低，例如自動生成品牌代言人的影音內容，節省傳統拍攝的人力物力。

《星際大戰外傳：俠盜一號》運用深度學習技術，讓已故的英國影星彼得．庫欣「數位復活」。
Photo Credit: Shutterstock.com

- **詐騙與假新聞等負面應用情境**：值得注意的是，深偽技術也被不當應用在非法活動中。例如，不法分子利用它偽造政治人物的講話影片，散播不實訊息以誤導公眾；或是合成名人影像發布虛假新聞。在詐騙中，更可能冒充受害者親友的聲音或臉孔行騙，讓人難以分辨真偽。

深偽技術在娛樂、教育、商業等方面展現巨大價值，但技術雙面刃的特性也日益顯現。當我們驚嘆於《曼達洛人》的數位魔法時，LINE群組裡假投資影片正悄然蔓延。

防範與偵測技術

深偽技術帶來全新的詐騙與資訊操控風險，各界也同步發展出多層次的防範與偵測手段來對抗這些威脅。目前主要的應對方向包括利用AI進行偵測、驗證數位內容真實性，以及強化身分驗證與資料保護等：

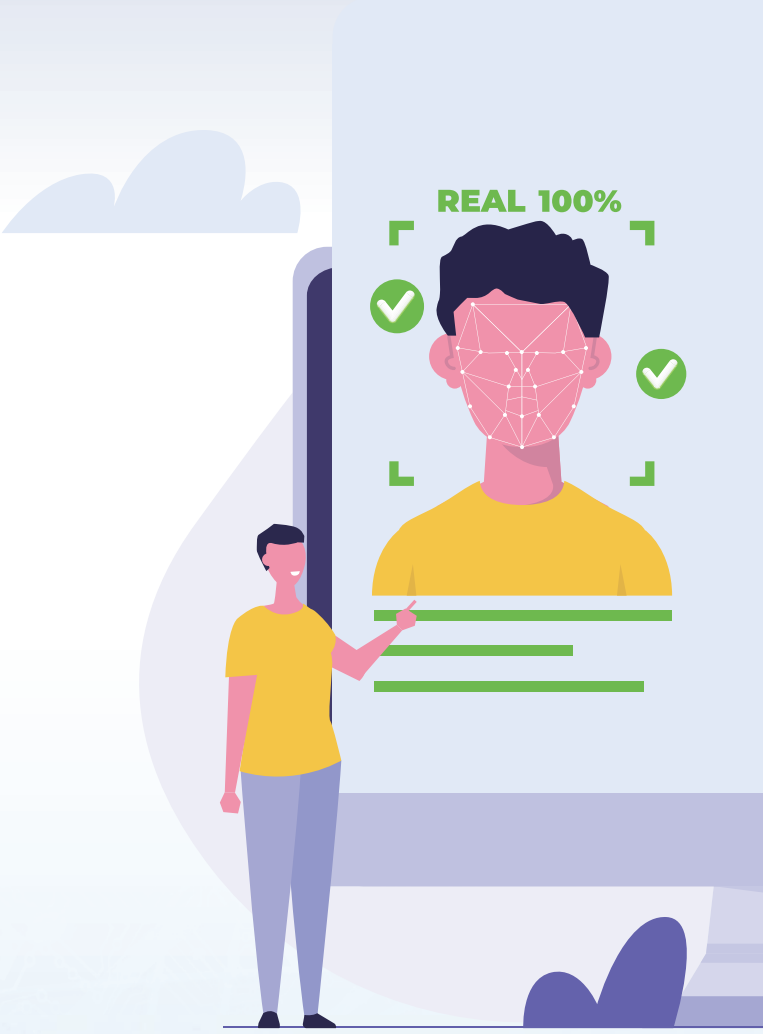
- **AI辨識技術**：運用深度學習模型來自動偵測深偽內容的蛛絲馬跡。透過對大量真人與偽造影音的訓練，這些AI判別器可捕捉深偽內容中難以避免的破綻，例如人臉細節的異常（不自然的眨眼頻率、光影不符的臉部特徵）或合成語音中不自然的雜訊與停頓。Facebook與微軟等公司甚至舉辦「深偽檢測挑戰賽」，鼓勵全球開發更高效的偵測算法。同時，多模態驗證也是有效手段：例如同時分析影片的畫面與聲音，檢查說話者的唇動與語音是否精確同步，以發現細微的不一致之處。

將一個人的臉部特徵、聲紋數據、面部表情精確地映射到目標人物，讓人難辨真偽。Photo Credit: Shutterstock.com

- **數位內容驗證工具**：透過技術手段驗證影音檔案的完整性與來源真實性。數位水印是常見方法之一：在影片或音訊中嵌入隱藏的數位標記，一旦內容被竄改或剪接，水印完整性被破壞，即可揭露偽造痕跡。許多新聞媒體在發布影片時都加入數位水印，以防止內容遭移花接木後再傳播。此外，還有專門的影音檔案分析工具，例如對聲音進行頻譜分析以檢測AI合成音的特有波形，或對影像進行像素級檢查以發現不合常理的陰影細節。部分社群平台亦部署了即時內容驗證系統，自動標記可能為深偽的上傳影片並提醒用戶注意查證。
- **身分認證與資料防護強化**：由於深偽技術可以輕易冒充他人，我們需要更

嚴謹的身分驗證機制來預防相關詐騙。區塊鏈技術因其不可竄改性，被用來存證影音資料的hash值或指紋，一旦原始內容有任何改變都可被追溯。此外，生物特徵認證（如指紋、人臉識別、虹膜掃描）結合多重驗證流程，能提高對身分的確認強度，防止單一密碼被盜用後仍可冒充身分。而最新的數位簽章技術則為每段影音生成加密簽名，確保內容在傳輸與存儲過程中的真實性，一旦簽名驗證失敗即可判定內容可能遭到篡改。

透過上述多管齊下的防禦措施，從AI自動檢測假影片，到內容溯源驗證，再到身分認證機制的升級，可以在一定程度上遏制深偽技術的濫用，保護個人隱私和社會安全。然而，隨著深偽技術的不斷提升，偵測與防範手段也必須與時俱進、持續創新，方能真正有效地對抗這種動態演變的威脅。



技術與倫理的平衡

深偽技術的迅速興起為各領域帶來前所未有的創新契機，但同時也引發了倫理與法律上的諸多挑戰。在技術進步與倫理約束之間找到平衡，不僅是開發者與立法者面臨的重要課題，更關乎整個社會的信任與安全。

首先，必須強調的是深偽技術本身擁有許多合法且正當的應用。例如，在教育領域，學校可以利用此技術製作多語言教學影片，透過擬真的口型同步呈現，提升課程生動度；在影視製作方面，電影與電視劇可藉由重現歷史人物或已故演員，增強敘事感染力；而在文化交流上，虛擬導覽或人物對話的應用能有效縮短語言與文化之間的隔閡。這些案例顯示，技術本身具有中立性，關鍵在於使用者的初衷與應用方式。

以臺灣為例，本地AI技術開發者在面對深偽技術挑戰時展現出驚人的創新潛力。聯發科近期開源的BreezyVoice模型，只

需5至15秒的錄音便能生成極具擬真度的合成語音，且該技術已針對臺灣口音做了專屬優化，甚至可在一般筆電上運行。然而，這項技術同時也暴露出「聲音盜用」的風險，工程師實測發現AI能夠流暢生成從未發生過的語句，彰顯了語音合成技術（TTS）已經突破了關鍵技術門檻。

在法律監管方面，各國政府正積極尋求對策，期望在保障技術創新的空間同時，防範可能的濫用。舉例來說，美國加州於2019年通過法案，嚴禁在選舉期間發布具誤導性的深偽影片，並對未經同意製作的深偽色情內容實施刑事處罰；歐盟則在2022年推動「數位服務法案」（DSA），要求大型平台對用戶上傳內容進行審查，並計劃於2024年強制標示AI生成的內容。這些措施雖在一定程度上遏制了濫用情形，但在如何兼顧技術中立性與濫用行為界定方面，仍存在不少爭議：究竟該追究創作者、發布者或技術提供者的責任？又如何避免因少數惡意案例而扼殺整體技術發展？

硬體層級的防護方案也是各界關注的焦點。例如，若半導體大廠與手機製造商合作，在晶片中內建即時深偽辨識功能，當用戶接收可疑影片時，螢幕可即時跳出「AI偵測到87%可能為偽造內容」的警告。結合區塊鏈溯源與數位浮水印技術，如在聲紋中嵌入人耳難以察覺的18~22kHz超聲波標記，或利用相位編碼隱藏數位簽章，都能在保留原始音質的前提下，對內容進行有效追蹤。

技術開發者也應承擔相應的社會責任。除了在生成內容中嵌入水印或標記外，業界正探索「數位出生證明」的概念，為每段AI內容生成獨有的加密簽名，並配合手機內建檢測工具，從源頭建立信任機制。以聯發科的BreezyVoice為例，儘管開源有助於技術發展，但同時也可能被不法集團利用來即時生成詐騙語音，這正是倫理辯論的重要案例。因此，業界應積極建立自律規範，如為符合法規標準的AI服

務頒發認證標章，並研發特殊音波浮水印技術，防範不法應用。

除了政府監管和技術防範之外，培養公眾的媒體素養亦至關重要。臺灣正積極推動針對臺語等在地語言特色的檢測方法，教育民眾從光影細節、聲紋諧波等特徵辨識深偽內容，增強「數位懷疑」能力。畢竟，人類獨有的情感表達與創意敘事，仍是AI難以跨越的最後防線。

總而言之，推進深偽技術發展的同時，融入倫理考量是確保其良性運作的關鍵。唯有透過法規制定、技術防範與開發者自律三方面的協同努力，才能在創新與風險間取得最佳平衡，讓深偽技術真正成為促進社會進步的創新工具，而非破壞信任的隱患。

結論

深偽技術的演進折射出科技與社會互動的永恆課題：創新與責任如何並行。只有在充分利用其帶來的正面價值的同時，嚴密防範其負面效應，我們才能既享受科技進步的紅利，又守護社會的安全與信任。在未來的日子裡，讓我們以更高的警覺和更密切的合作，迎接這項AI新技術帶來的機遇與挑戰。

AI技術被不法人士拿來做為犯罪的工具越趨頻繁，法務部調查局拍攝影片教民眾如何用簡單的小技巧破解「視訊通話變臉技術」。Photo Credit: 法務部調查局臉書

臺中豐原區戶政事務所宣導海報。Photo Credit: https://www.hfengyuan.taichung.gov.tw/3476902/post

從ISO42001出發 討論我國產業發展AI

應採取之資料信任證明作法

◎ 洪亮瑄／資策會科法所副法律研究員

全球產業數位化，人工智慧（Artificial Intelligence, AI）也越來越廣泛地應用於各領域產業。但依據IBM公司於2024年1月發布的資料顯示，包含美國在內20個國家的43%企業認為生成式AI的訓練資料不透明，難以信任其資料來源。[1] 國際貨幣基金組織的2023年調查報告指出已開發國家應優先發展安全、負責任的AI監管框架。[2] 國際亦重點發展「負責任的AI」。[3] 為平衡AI的創新與管理，證明AI資料可信

1 IBM公司於2024年1月發布《2023全球AI使用現況》調查報告，該報告調查範圍為美國、加拿大、法國、德國、英國、印度、意大利、中國大陸、日本、新加坡、韓國、澳洲等20國的8584 名 IT經理或更高職務人員。參考：IBM，IBM發表《2023年全球AI採用指數》：生成式AI最快產生影響的企業用例－IT 自動化、數碼勞動力、客服，https://hongkong.newsroom.ibm.com/2024-01-17-IBM-2023-AI-AI-IT#assets_all （最後瀏覽日：2024/03/18）。

2 Kristalina Georgieva, *AI Will Transform the Global Economy. Let's Make Sure It Benefits Humanity,* https://www.imf.org/en/Blogs/Articles/2024/01/14/ai-will-transform-the-global-economy-lets-make-sure-it-benefits-humanity (last visited Mar. 18, 2024).

3 The White House (2023), *FACT SHEET: Biden- Harris Administration Announces New Actions to Promote Responsible AI Innovation that Protects Americans' Rights and Safety,* https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/fact-sheet-biden-harris-administration-announces-new-actions-to-promote-responsible-ai-innovation-that-protects-americans-rights-and-safety/ (last visited Mar. 06, 2024).

度，本文試以2023年12月18日公布的全球首個AI國際標準「ISO/IEC 42001：2023（下稱：ISO 42001）」[4]為出發，提供我國產業資料管理建議。

ISO 42001跟資料管理的關聯性：強調資料歷程管理

ISO 42001旨於幫助組織負責任地研發、維護、持續改善AI系統，為提供或研發AI系統的供應商、合作夥伴及第三方提供AI管理系統的標準化流程。ISO 42001的規範架構採取與其他管理系統標準相同的共通性結構，第1至3單元為適用範圍、版本標示、名詞定義之說明；第4至10單元為管理要項，其架構可參照下圖1整理的ISO 42001之計劃、執行、查核、行動（Plan, Do, Check, Action, PDCA）管理循環架構。

ISO 42001將「提供或研發AI系統的供應商、合作夥伴及第三方的資料（documented information，或稱為『文件化資訊』）」，定義為：「組織需控管及維護的資訊，以及該資訊的媒介（medium），比如AI管理系統的相關流程、為組織運作而創建的資訊、AI管理系統所取得成果的紀錄（證據）等」。ISO 42001適用範圍為全球「提供或使用AI系統產品或服務的組織」，不論該組織的規模、類型（企業、公部門、非營利組織等）及組織所提供或使用的AI系統產品或服務。

為證明AI系統的適用性、完備性及有效性（suitability, adequacy and effectiveness），避免重複地發生高風險、

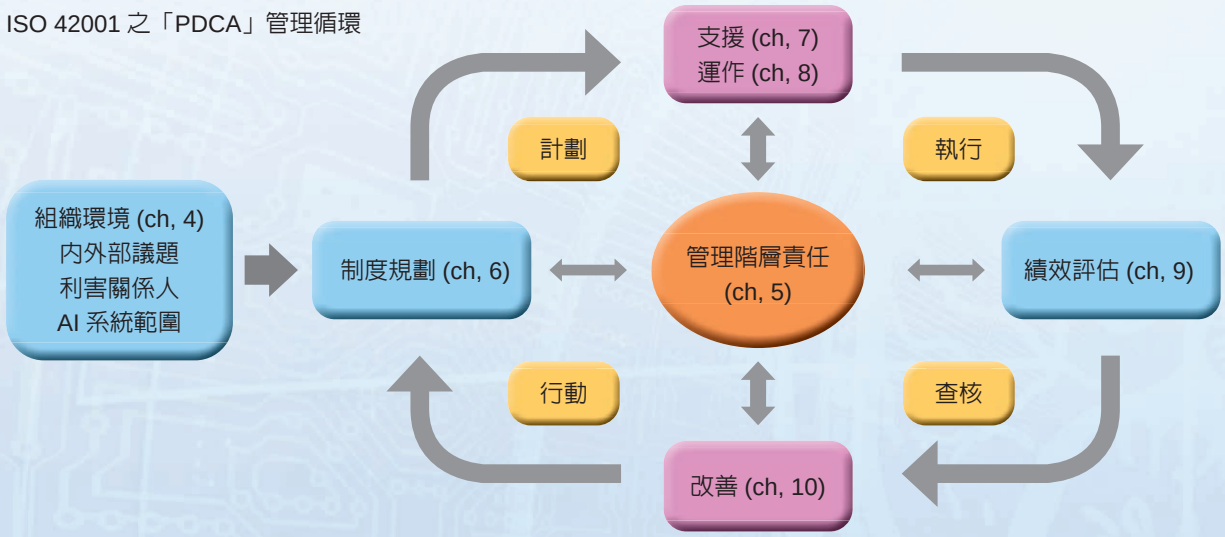


圖 1 ISO 42001 之「PDCA」管理循環架構。資料來源：本文整理

4 ISO/IEC 42001:2023, https://www.iso.org/standard/81230.html (last visited Mar. 06, 2024).

不合格的項目，ISO 42001國際標準強調透過「管理可信任的AI系統（trustworthiness of AI system）生命週期所涉及的資料歷程」，[5]可以提升資料的可溯性（Traceability）、透明度及可信度，展現組織負責任地使用AI，確保AI資料品質。

為解決AI不受信任的課題，本文由ISO 42001國際標準所涉及的數位資料管理內容出發，得出ISO 42001與國內財團法人資訊工業策進會科技法律研究所創意智財中心於2021年7月發布的「重要數位資料治理暨管理制度規範（Essential Data Governance and Management System，下稱：EDGS）」採取相同的PDCA管理循環架構，且EDGS的數位資料管理範圍已涵蓋「AI資料」，並嘗試從AI軟體公司的角度，舉例說明資料管理循環的重點。

我國AI軟體公司除可以參考ISO 42001，亦可以先參考EDGS，該規範涵蓋了「AI生命週期資料的管理機制」，且採取相同的PDCA循環規範架構。

以下嘗試從AI軟體公司的角度舉例，帶入EDGS規範，說明AI軟體公司需要留意的管理作法。

AI軟體公司採取的資料管理循環

由於軟體相關研發及應用日新月異，應用於不同產業別的AI軟體所涉及的具體細項資料亦有不同。本文假設AI軟體公司透過AI分析無人機拍攝的影像，使AI產出「大型基礎設施建案的工程現場報告」，[6]並提供施工進度追蹤及資料品質管理平臺。

5 組織可考量以下因素，調整其需要記錄、管理的資料範圍：組織的規模及其活動、流程、產品及服務的類型；AI管理系統流程的複雜性；員工能力。

6 所謂大型基礎建設建案，比如高速公路、太陽能發電廠、機場和鐵路軌道、天然氣管道等。

由於AI的訓練資料不透明以及AI系統的管理作法不同於一般產業所應用的系統，使AI軟體公司難以證明其AI生命週期的資料可信度，因此本文將AI軟體公司帶入EDGS所採取的可信任AI的管理作法，並依PDCA循環分為四點：

一、P（計劃）：對應EDGS第四~六單元

（一）EDGS第四單元「組織環境」

「組織環境」代表公司需要構思與數位資料相關的內外部議題、利害關係人，才能夠擇定需要管理的數位資料。比如AI軟體公司需要權衡市場應用需求、公司規模及資源等，以評估研發可行的AI功能。

（二）EDGS第五單元「數位治理暨管理階層責任」

擇定需要管理的數位資料後，公司需要有具備統籌權利的高層人員訂定數位資料管理的政策目標及相應的業務執行規範，以確保執行成效。比如AI軟體公司的最高管理階層應制訂AI政策，於規劃、制定AI相關管理規範時，需要考量滿足組織整體政策目標、適用要求、承諾持續改善AI系統。為了讓導入EDGS的組織，可將EDGS與組織既有業務的管理流程、其他管理系統標準等進行整合，EDGS沒有限制組織必須採取何種具體的管理作法，最高管理階層須思考AI的管理作法能否與組織既有的管理系統標準整合，如ISO 9001（品質）、ISO 27001（資訊安全）等。

（三）EDGS第六單元「制度規劃」、「風險與機會因應」

1. 制度規劃

制度規範需要包含數位資料的標的範圍、業務流程等內容。為了讓AI提供可信任的「精確地監督施工進度」的功能，AI軟體公司選擇保存的數位資料為「使用AI軟體時，所涉及的建築工地上所有活動的完整資料證據鏈」，包含：個別工程的環境資料、建築資訊模型、人力資源資料、材料及設備清單、財務資料、安全檢查資料、工程預算資料、由前述資料整合而製作的線上視覺化互動圖表、AI生成覆蓋率

100% 施工區域的工程現場報告的PDF檔案以及AI的系統日誌（log）等。

2. 風險與機會因應

數位資料會受到可預期、非預期的風險影響，所以規劃組織制度時，組織需要提前思考可以處理的風險，才能因應其所造成的衝擊。比如AI軟體公司為減輕AI風險，公司最高管理階層須制訂並持續修訂AI風險標準，規範「可接受與不可接受的風險等級」、風險評估、風險處理、有效性評估以及AI系統影響評估流程，並保留前述歷程資料為證。

二、D（執行）：對應EDGS第七、八單元

（一）EDGS第七單元「支援」

管理流程的執行，需要透過人員教育訓練來宣導規範內容，以及設定人員接觸資料的權限。比如AI軟體的參數操作、功能增修，需經培訓確認AI軟體的經手人員具有職能。關於限制對數位資料的接觸權限，比如人員僅能透過AI軟體公司的資料品質管理平臺，並登入個別帳號取得資料接觸權限。

（二）EDGS第八單元「重要數位資料治理暨管理制度實踐流程」中的「數位資料之生成、保護與維護」

1. 數位資料之生成

為方便公司管理數位資料，需要能夠從檔案標題、檔案內文格式、儲存檔案的資料夾名稱等相關資訊，對應到特定的數位資料檔案或檔案類型。比如AI軟體公司的AI製作的線上視覺化互動圖表，具有固定的檔案內文呈現框架，其可以依其提供之建案進度追蹤服務，將圖表標題分為「建案概覽」、「地層狀況」、「基礎設施」、「樁（piles）的施工風險」、「施工時程表」等資料。[7]

2. 數位資料之保護與維護

由於難以發現數位資料的增刪修改痕跡，EDGS說明可以透過留存資料不同版本、設定資料的接觸權限等方式保護重要數位資料。

關於檔案版本留存，比如AI軟體公司的AI所製作的工程現場報告，不同時間點的施工進度，會產出不同內容的工程現場版本，透過於檔案註記追蹤施工的日期，可以標示不同版本。

三、C（查核）：對應EDGS第九單元「績效評估」

AI軟體特點為風險難以預見，為確保AI軟體可信任，公司應持續評估、處理AI

7 為了避免土壤無法承壓，而導致建築傾斜或倒塌的風險，於工程施工階段，施工團隊會將混凝土、鋼等材料製成的「樁」打入地中，來補強土地的承載能力。施工風險因不同材質、不同結構的「樁」以及不同地質結構的土壤需要打入樁的深度等因素而不同。參考：臺灣大學，科學Online，基樁－建築穩固的基礎，https://highscope.ch.ntu.edu.tw/wordpress/?p=70441（最後瀏覽日：2024/03/11）。

AI軟體公司須持續評估、處理AI風險。
Photo Credit: https://www.shutterstock.com

風險。比如AI軟體公司需評估AI軟體訓練資料缺乏透明度、系統故障、對社會環境的影響等潛在風險，且須留存風險評估、風險處理相關資料，作為其執行的證明。

四、A（行動）：對應EDGS第十單元「改善」

AI的特點為持續學習、更新，當公司依據AI規範而使用AI、進行風險影響評估、風險處理及內稽後，將會持續透過機器學習等方式優化AI功能及調整原先的AI管理作法，並留存審查紀錄為證。

比如AI軟體公司為符合使用者需求及風險調適，一年至少進行上千次AI機器學習，且各建案建立獨立的「線上視覺化互動圖表」，精進相關資料集已超過數千萬個標籤（labels）。

關於AI軟體證明其可信任的管理作法，可得出要點在於管理數位資料歷程，有助AI軟體公司留存完整、真實、透明的AI資料證據鏈，彰顯公司對於可信任AI的研發、選用態度。

結語

AI系統面臨難以證明資料可信度、可用性的風險，建議業界亦可參考我國EDGS的「資料管理機制」發展「負責任的AI」，留意AI的風險評估、風險處理，以及數位資料的生成、保護以及維護作法，透過EDGS管理AI生命週期所涉及的數位資料，強化AI的資料可溯性、降低企業可能面臨的AI風險、增進跨域合作機會。