

AI 造假風潮 要如何應對？

- 余創豪 chonghoyu@gmail.com



排山倒海的人工智能造假事件

最近接連發生的一連串人工智能造假事件，如排山倒海般衝擊著大眾的認知，令人不禁深刻懷疑，我們每天在網絡與日常生活中所接觸到的資訊、影像乃至人事物，究竟有幾多是真，幾多是假？就在今年四月，一名印度醫學生被揭發利用人工智能，在社交平台上憑空創造了一位名叫「愛美麗·哈特」

（Emily Hart）的虛構女護士。這個並不存在的網上美女，竟然吸引了成千上萬的 MAGA 追隨者，並為該名編造者每月帶來數千美元的豐厚收益，直到騙局被揭穿，帳號才遭到注銷。

無獨有偶，國際政要也無法倖免於這場科技風暴，義大利總理梅洛尼於 2026 年 5 月 5 日公開譴責一張在社群媒體瘋傳的深偽照片。該照片惡意顯示她在床上穿著內衣，她憤怒地表示這類 AI 生成影像已被公然用作政治與人身攻擊工具，並強烈呼籲民眾在未經查證前切勿盲目轉傳。南韓發生的人工智能造假事件釀成更嚴重的後果，韓國巨星金秀賢被指控惡意凌辱未成年女星，起初彷彿是鐵證如山，但經過深入調查之後，今年五月南韓警方公布那些所謂截圖和錄音都是利用人工智能偽造的，但金秀賢已經因為千夫所指而身心受創。

還有，歐盟在 2026 年 1 月 26 日針對馬斯克旗下的 X 平台展開深入調查，其核心焦點正是該平台推出的 Grok 系統被濫用，生成了大量將女性與未成年人遭性化處理的深偽影像。報導提到，用戶只需透過極其簡單的指令，就能輕易把正常影像變成比基尼、甚至脫衣等露骨內容。種種跡象顯示，AI 詐騙、語音釣魚與 deepfake 相關犯罪在 2025 至 2026 年間明顯急劇升溫。根據 Vectra 的研究報告以及美國國會的相關聽證會資料，AI 詐騙的案件量正呈現爆發式急增，其中語音複製技術已被視為當前最重要的安全風險之一。

太完美便不會是真的

面對這股高科技以假亂真、彷彿令人防不勝防的欺詐浪潮，筆者身為在大學任教人工智能的學者，大眾可能會理所當然地以為，我必然會提出用更高深的尖端科技去與之對抗。然而，答案恰恰相反，許多基於人類基本常識和傳統智慧的方法，其實早已足夠幫助我們過濾掉絕大多數的虛假訊息。

第一個最簡單卻最核心的驗證原則就是：「太完美便不會是真的。」愛美麗·哈特絕非第一個利用人工智能虛擬出來的美女幻象，如果仔細觀察，這類所謂的虛擬美女往往都有一個致命的共通點，那就是她們顯得比任何真人更美艷動人，肌膚完美得沒有任何瑕疵，甚至連一粒暗瘡、一粒墨痣、或幾撮凌亂的頭髮都沒有。如果在視頻中，她連說話舉止也完美無瑕，沒有讀錯任何字音，沒有半點不流暢的停頓，這在統計學和現實邏輯上，反而是做假的強烈徵兆。

事實上，在人工智能造假技術誕生之前，傳統騙子就已經深諳此道，總是致力於營造出令人豔羨的虛幻美景去引人上鉤。例如某些投資基金大肆吹捧超高回報率，且年年賺錢，宣稱沒有任何風險；又或是有些垃圾電郵謊稱你幸運中了大獎，只需提交少少手續費，便可以領取幾百萬美元。兩者背後都是同一句老話：「太好便不會是真的。」這種對付傳統詐騙的心理防線，同樣完全適用於今天應對人工智能的深偽技術。

傳統傳媒比互聯網更可靠

另一個簡單而有效的驗證方法，則是審視資料的來源與出處。如果一張圖片或者一則新聞來歷不明、缺乏權威背書，那麼我們理應保持合理的懷疑，不輕易寄予信任；反之，如果消息有根有據，且源自主流傳媒，那麼其可信性會大大提高。有趣的是，在當前兩極化的政治生態中，傳統主流傳媒經常因為政治見解的分歧，被不同陣營批評充滿偏見而受到排斥。然而，人們往往忽略了，傳統傳媒在對消息真偽的查證機制上是非常嚴謹的。不幸的是，現在許多人卻逐漸喪失了思辨能力，不加批判地被動接收流傳在互聯網和社交媒體上的各種小道消息。

當然，我們必須承認主流傳媒並不是絕對可靠，在歷史上也曾發生過發佈未經證實消息的重大失誤。舉例來說，在 2004 年，美國哥倫比亞廣播公司（CBS）曾高調報道時任總統小布殊在德州國民警衛軍服役時得到特殊待遇，但隨後關鍵文件被證實是偽造的。這起醜聞導致當時的節目製作人被撤職，而 CBS 的傳奇主播兼總編輯丹·拉瑟（Dan Rather）也因而引咎辭職。這個歷史故事給予我們最深刻的教訓在於：傳統傳媒建立了嚴格的問責制，一旦證實消息是假的，他們必須公開道歉、更正並追究內部責任。相反地，那些在

互聯網和社群平台上四處散播假消息的匿名網民，卻完全無需承擔任何法律與道德上的責任。

親友同事之間的暗號

至於面對利用人工智能製造假影像、甚至複製聲音的精準詐騙，我們也可以採用十分簡單的人際互動方法去機智應對。若果在電話或視訊中，對方宣稱是我的親人、朋友或同事，我會反客為主，主動問一些唯有彼此知道、外人不可能知道答案的問題。例如我可以問：「我身體哪個部份有紅色胎記？」「五年前我們一同參加旅行團，你可以告訴我那一次我們去了那一個國家？」「兩個星期前你到我家吃晚飯，跟我的小狗玩，我的小狗叫什麼名字？」

可能有人會對此提出反駁，認為現在高明的黑客和騙徒完全可以用高科技入侵個人的電腦檔案，從而徹底掌握你過去的醫療紀錄、旅行歷史、甚至各類隱私，導致上述的詢問方法未必次次有效。那麼，更進一步的安全防範策略，就是提前和親友、同事口頭設立專屬的「暗語」，且切記不要留下任何筆錄或數位痕跡。例如當你問一句：「芝麻在什麼時候開門？」而對方必須心領神會地回答：「40大盜離開的時候。」

結語

歸根究底，人工智能固然能夠製造出令人難以分辨的精緻贗品，但很多時候，真正打開潘朵拉魔盒的鑰匙並不在於科技的能力有多麼強大，而是在於人性本身的弱點。正因為我們內心深處貪慕美女、渴望賺快錢、幻想著不勞而獲，於是往往在誘惑面前，連最簡單的邏輯判斷都自願擱置起來。同時，在心理學的「確認偏誤」下，我們總是傾向於盲目相信自己內心想見到的東西。例如當看到有什麼負面消息不利於自己所討厭的政客時，便會不假思索地全盤接受；相反地，僅僅基於自身的政治立場，卻動輒將傳統傳媒經過嚴謹查證的報道斥為假新聞或偏見。也許，面對這一股洶湧的人工智能造假浪潮，科技只是鏡子，照出了人類的盲目。我們最需要做的，並不是尋找更先進的防禦軟件，而是首先反省並警惕我們自己內心的弱點。

2026年5月22日

〈人工智能造假風潮要如何應對？〉

原載於台灣《國語日報》

[更多資訊](#)