



亞運會昨閉幕 港隊創歷屆最好成績 李家超：反映政策和資源投入方向正確

相關新聞刊 A2·A3·A4

A6

啟德醫院分階段啟用 街症設350名額



A7

「護老外傭」料明年下半年試行

A15

港投：逾30被投公司擬年內申上市

AI詐騙氾濫 眼見未必為實

專家：重塑防騙觀念 抵禦智能犯罪

編者按

人工智能（AI）高速迭代，為社會帶來數碼轉型紅利之餘，亦衍生全新網絡風險與治理難題。有人將AI風險視為遙遠的科技末日，惟《大公報》透過訪問多位本港專家學者及業界人士發現，AI危機絕非危言聳聽，AI風險已經滲透到香港市民日常生活、企業營運、金融運作，成為真實的隱憂。

《大公報》推出「正視AI風險」系列專題，立足香港在地場景，拆解AI輔助網絡攻擊、深偽詐騙、演算法黑箱不公、法律監管空白四大核心危機，針對普羅市民、中小企、大型機構及政府監管多層面，提供可落地的防禦對策與自保守則。本系列三篇專題由淺入深，揭露被大眾忽視的智能風險，探討香港如何在科技創新與安全防禦之間取得平衡，築牢屬於香港的AI防線。

邱達根

（立法會科技創新界議員）：



專家之言

本港大型機構及關鍵基礎設施營運者，有定期開展AI攻防演練，但中小企及普羅市民的防禦能力，出現嚴重斷層。

黎少斌

（生產力局首席數碼總監）：



目前香港未有AI自主生成病毒攻擊個案，真正危機是駭客以AI為武器，大幅壓縮攻擊成本，令傳統網絡防禦失效。

古治雄

（中大新聞及傳播學院助理教授）：



沒有嚴格定義的「AI病毒」，但AI可無限迭代改寫惡意程式，讓病毒變種速度遠超人類修補漏洞的速度。



正視AI風險①

AI偽造的親友聲音、以假亂真的企業視像會議、幾近完美的官方釣魚網站與郵件，現正成為本港新型詐騙及網絡攻擊的主流手段。過去市民信奉「眼見為實」，惟在AI技術加持下，視覺、聽覺接收的資訊隨時可被偽造，傳統的防騙、防毒思維已完全跟不上技術演變速度。

不少市民及中小企誤以為，現時市面流傳的AI網絡攻擊是人工智能自主產生惡意、主動發起破壞。生產力局首席數碼總監黎少斌接受《大公報》專訪時糾正，本港暫未發現AI自主生成病毒並獨立攻擊的案例，現階段所有AI相關網安威脅，均是人為駭客借助AI工具放大破壞力。重塑防騙、防毒觀念，是抵擋智能虛假洪流的第一道底線。

大公報記者 苑向芹

「現時網絡威脅的核心改變，是市民肉眼、直覺的辨識能力，已經徹底落後於AI偽造技術。」生產力局首席數碼總監黎少斌引述境外重大商業攻擊案例時指出，有跨國企業高層在AI深偽偽造的視像會議中，聽從「上級」指令完成多輪轉帳，最終損失近兩億元。整場詐騙畫面逼真、對話邏輯完整，完全突破常人辨識極限。

傳統防火牆形同虛設

中大新聞及傳播學院助理教授古治雄博士補充，業界並無嚴格技術定義的「AI病毒」，更貼合現實的說法是「AI輔助惡意程式」。過去駭客發現系統漏洞後，需手動編寫攻擊代碼，耗時費力；如今透過AI便可快速生成、即時改寫程式，製造海量病毒變種。這直接扭轉網絡攻防格局，漏洞曝光後，企業修復系統的時間被大幅壓縮，防禦方被動陷入劣勢。

傳統防火牆、防毒軟件的運作邏輯，是比對已知病毒的代碼特徵，實現攔截清除。惟AI可隨機重寫病毒代碼、更改程式特徵，令舊式防禦系統形同虛設。黎少斌分析，AI時代的網安防禦，必須拋棄「特徵比對」的舊思維，轉為監測程式異常行為、操作邏輯，與駭客「以AI對抗AI」。

有業界人士解釋，不同於影視作品中AI自主反抗人類的科幻情節，現實層面的智能叛變，目前尚未發展到AI擁有自主意識，而是AI脫離人類設定的初衷，以極端、偏頗、違背倫理的方式完

成工作目標，形成系統性失控，損害公眾權益與社會公平。

不過，業界強調，市民無需過度恐慌。AI攻擊仍需專業技術支撐，普通用戶簡單指令無法生成有效攻擊程式，且主流AI大模型設有倫理與安全防火牆，可杜絕大部分惡意指令。真正的普遍風險，集中在AI釣魚詐騙——這類低門檻、高仿真的欺騙手段，已成為本港市民最易中招的智能威脅。

未經獨立核實 預設虛假

立法會科技創新界議員邱達根補充，本港大型機構及關鍵基礎設施營運者，已按照《關鍵基礎設施保護條例》要求，定期開展AI攻防演練、系統壓力測試，透過AI主動挖掘內部漏洞。數字政策辦公室亦曾舉辦公開攻防測試，允許技術人員以AI模擬攻擊政府網站，藉此排查隱患。相對之下，中小企及普羅市民的防禦能力，出現嚴重斷層。

針對市民日常防禦，黎少斌提出顛覆性防騙原則：「所有未經獨立管道核實的資訊，一律預設為虛假。」不論是親友索款的AI語音、快遞銀行的短訊通知、社群流傳的官方通告，均需透過官方熱線、固有聯絡方式二次查證，切勿輕信畫面、聲音與文字表象。

業界總結，AI現階段並未主動「叛變」，而是成為不法分子的高效犯罪工具。香港社會的防禦漏洞，不在技術設備，而在大眾仍然滯留於前AI時代的辨識思維。重塑防騙、防毒觀念，是抵擋智能虛假洪流的第一道底線。



坊間不少團體推動人工智能知識普及和負責任應用，讓市民認識科技安全及防騙資訊。香港人工智能專業協會IG

港人4大高風險AI「釣魚」場景

- AI親友聲音索款
- 偽官方短訊釣魚連結
- 社群AI假通告
- 深偽視像會議詐騙

資料來源：警方防騙資訊、專家訪問

駭客AI攻擊兩大核心模式

1 智能網攻：

AI探測系統漏洞 自動生成變種病毒 繞過傳統防毒攔截

2 深偽詐騙：

AI仿真人聲、影像、官網、郵件 誘導轉賬、洩露個人資料

大公報AI製圖

專家倡保留人工終審權 防「智能叛變」

加強把關

智能叛變的核心成因，在於算法「唯指標論」的運行邏輯。香港金管局副總裁阮國恒指出，隨著代理式AI與前沿模型普及，AI已從輔助工具走向直接執行決策，衍生出嚴峻的自主「越軌」與網絡安全風險。

定期檢視模型偏差

這類失控問題在香港尤為突出。

本港中小企普遍缺乏AI審核人才與系統測試機制，大多直接套用通用商用模型，企業內部團隊無法掌握算法運行細節，亦沒有建立完善的人工覆核、異常糾錯機制。長期黑箱運作下，算法偏誤不斷累積、自我強化，逐漸脫離企業原本「便民、客觀、高效」的營運初衷，形成算法凌駕人為判斷的無形霸權。

專家強調，香港的智能叛變危

機，在目前階段尚未到技術失控地步，而是監管與管理缺位，必須牢牢守住AI為人類服務、受人類控制的底線。在AI全自動化趨勢下，若企業持續捨棄把關機制、任由算法獨立裁決，數碼時代的不公只會不斷擴大。唯有建立算法透明機制、保留人工終審權、定期檢視模型偏差，才能有效遏制智能叛變，避免AI由便民工具變成剝削公眾權益的巨大風險。

