大公報
2026年3月18日 星期三

電商平台頻現詐騙 深度偽造肉眼難辨

生成技術「雙刃劍」業界倡用AI監管AI

AI熱下的冷思考②

當AI生成技術以前所未有的速度重塑世界，它也猶如一把「雙刃劍」，不僅帶來了效率提升，同時也帶來了數據隱私、信息安全和倫理道德等挑戰。3月15日，央視「3·15」晚會曝光AI大模型遭「投毒」亂象：當前，GEO（生成式引擎優化）服務商通過自媒體大量發布軟文「投餵」AI大模型，以此操縱搜索推薦結果、灌輸虛假信息，誤導用戶。對於電商平台詐騙及深度偽造肉眼難辨等問題，不少業界人士提議，「用AI打敗AI」是創新監管路徑之一。

大公報記者　王莉

◀有不良消費者在貨品到手後拍下商品（左），再以AI生成貨品變壞圖片（右）來要求退款。

◀電商平台大量出現AI生成商品圖，有消費者控訴AI模特兒身上美麗的衣服（左），到手後卻不如人意（右）。

綠色植物電商經營者潘女士近期遭遇了多起消費者利用AI偽造瑕疵圖片，以騙取「僅退款」（賣家直接退款而消費者毋須退貨）的事件。

有心人開班收費 傳授騙退款「竅門」

「一位海棠花盆栽買家在收到貨後，給我發來一張花苞都掉了的圖片，堅持退款且不退貨。我還多長了一個心眼，讓她再拍個視頻給我，沒想到視頻裏確實花苞也都掉了。」她最後用AI檢測工具才發現圖片和視頻都是AI生成的，於是她拒絕退款並提交檢測報告，但平台介入後仍判定買家勝訴。她無奈地說，「現在AI技術太強了，再不好好管的話，我們商家真的只能啞巴吃黃連了。」

此類情況並非個例，甚至在網上還出現了「羊毛黨」購物經驗分享，更有甚者表示支付288元（人民幣，下同）學費即可傳授「僅退款」竅門，並宣稱一個賬號約可成功退款30次。為證明效果，還有商家通過展示學員「累計獲利2000元」成果來獲客。然而，經記者親測，僅需使用目前市場上最常用的免費AI軟件，上傳一張新鮮橘子的照片後，再簡單輸入「橘子爛了」四個字，新鮮橘子就被P上了霉斑。

長期從事網絡新經濟和消費者權益保護工作的浙江佑平律師事務所執行主任俞起表示，此類行為具有主觀上欺詐的故意，屬民事欺詐，若騙取金額達到一定標準，或需承擔刑事責任。同時，他也提到，監管部門、電商平台以及電商經營者都可利用AI技術進行「反制」。「監管部門應利用技術手段強化對違法行為的監測、識別和打擊力度。平台除對圖片、視頻等進行AI檢測外，也可利用大數據監測異常退款模式。商家同樣可以借助AI工具快速處理和分析收到的圖片證據，也可用於輔助製作說明材料，向平台提交更具說服力的申訴。」

金融界訓練AI 應對偽造文字語音

多組機械臂在實驗室中24小時連軸轉。一組配備手機攝像頭的機械臂不停伸縮、對焦，反覆識別着眼前的圖片，而另一組機械臂則精準地舉起圖片，穩穩遞到攝像頭前。計算機實時記下每一次識別的結果，再把這些數據「餵給」AI算法，讓它在一次次實戰中快速成長，以更好地識別出不同AI欺詐的破綻。這不是科幻場景，而是螞蟻數科正在金融領域訓練AI對抗非法AI的場景。

「以往攻擊方式通常是用圖像或視頻處理軟件進行修改操作，如今則主要是通過AI生成的注入式攻擊，表現形式為用一張臉換成另一張臉、將靜態圖片轉換為視頻、改變面部特徵、拼接兩張臉等。」螞蟻數科相關人士表示，面對升級版攻擊，就需要深入研究不法分子的各類欺詐手段，模擬其操作邏輯，利用專業設備偽造人臉、指紋等樣本，進而開展針對性攻防演練。「唯有精準洞悉不法分子的作案套路，才能構建更有效的防禦體系。」

2025年10月，香港金管局推出的第二期生成式人工智能沙盒計劃也充分聚焦加強AI治理，其中多個案例採用「以AI對抗AI」策略。如中銀香港通過深化偵測文檔防欺詐及人臉識別反深度偽造技術，讓生成式人工智能持續學習，以應對新型深度偽造的文字及語音技術。金融壹賬通也在全球金融行業率先推出「智能視覺反欺詐策略平台」，可對AI偽造圖像進行深度剖析，精準識別換臉、偽造視頻等複雜攻擊，綜合檢測防禦率高達99%+。

▲視頻製作者以AI生成雷軍、于東來、張文宏（左至右），利用名人力量來宣傳自己的產品或見解。

AI亂象案例：作品歸誰有？

核心爭議

- 通過輸入文本描述，讓人工智慧自動生成符合描述的文字、語音、圖像等，著作權歸屬何方？

內地判例

- 多地判例確認，當生成內容體現人類用戶的「獨創性智力投入」時，AI作品可受著作權法保護。
- 有法官指出，僅通過輸入簡單提示詞生成的AI內容只是思想，並非受著作權法保護的表達，其並不構成作品。

解決辦法

- 中國科學院科技戰略諮詢研究院研究員肖尤丹建議：

一、需要鼓勵在傳統內容創作、流通和保護中，更為普遍性地引入基礎保護措施，在AI大模型保護中使用區塊鏈、數字水印等技術。

二、可以從擴張作品、發明類型、設立生成式人工智能系統使用費知識產權回饋機制，建立數字化內容實用集中清算制度，重構數字化合理使用制度。

三、增加非人工智能舉證責任等方式，平衡人工智能技術發展與知識產權制度的矛盾。

大公報記者王莉

劃定風險試驗區 總結推廣治理經驗

◀有內地媒體測試發現，只要一句簡單指令，AI就能令圖片中正常的水果「變壞」。

制度入手

「目前AI發展還是處於早期階段，發展路途還很漫長，對於AI治理來說，問題還會不斷出現。所以我們需要有一個常態化快速響應機制以應對未來會出現的這些問題。」原力無限機器人聯合創始人劉揚認為，可劃定範圍建立AI風險試驗區，從中驗證總結，再將治理經驗推廣至整個行業。他坦言，AI治理單純依靠企業自覺性很難達到理想結果，還需要聯動專業機構、監管部門、企業聯合會等實現共同管理。

「就像比賽中一定要有裁判這個角色，就是為了通過一定機制限定大家在比賽中按規則來做。」因此劉揚建議，「比如可以先劃定一個小範圍完全放開，在這裏面去驗證可能會有哪些風險出現。同時將各種產生的風險數據、產生原因、治理方法等建立資料庫，然後再告知相關行業、企業應該怎麼做。」

不過，他也表示，即使如此肯定還會存在「漏網之魚」，所以如何在發生未能預測的潛在風險時，把風險規避到最小也顯得尤為重要。「就像早前充電寶在飛機上引發火災風險後，不是單純依靠機場檢查這一個辦法，而是通過軟硬件側立刻『拉閘』，同步把潛在風險排除掉。哪怕是在犧牲效率的情況下去做這件事，因為安全永遠是最重要的。」

設置風險評級 制定准入門檻

浙江佑平律師事務所執行主任俞起從法律保障體系方面建議，可根據不同AI行業領域的風險級別構建風險評估與分級分類處置機制。「這樣能夠更好實現垂直領域風險動態評估。監管者強化多部門協同聯動，暢通違法違規線索舉報渠道，敦促AI技術提供者落實合規管理激勵機制與處罰措施。同時根據風險分級梯次設置准入門檻、治理義務與運營監管措施，嵌入適應性治理等。」 大公報記者王莉

◀◀內地有創作者起訴其AI生成作品（右）被宣傳公司抄襲（左），獲法院肯定其作品應受著作權法保護。

人類需要一張更有創造力的考卷

專家解讀

「人工智能改變了很多科學發現背後非常重要的邏輯，數據的意義也正在發生深刻變化。」中國工程院院士、之江實驗室主任、阿里雲創始人王堅早前表示，如何用AI打造更加開放的科學研究過程，是一個全球面臨的挑戰。

王堅指出，2024年11月發表在美國天文學會會刊上面的一篇文章，作者發現確認了150萬個未知的天體。這樣規模的發現，過去應該是一個幾百人團隊署名的文章，但這篇文章只有一個作者，而且是一個18歲的高中生。「這篇論文所有依據的數據不是新的數據，是一個NASA已經退役的巡天探測器留下來的數據。所以我們過去對數據的理解、數據產權的界定、數據價值的界定，在科學發現階段，都要做一次重新的思考。」

王堅指出，當談到「數據」對科學技術影響的時候，這個「數據」已經遠遠超出了大家講的出版物的數據，已經遠遠超出了大語言模型，今天可能在語言文字上，還有很多知識產權的問題沒有明確，但將來的問題會變得更加挑戰更加複雜。「科學發現不只是從語言和圖片，事實上最後都要回到最原始的科學數據觀察，比如說基因的數據、光譜數據，人工智能的發展後面面臨的數據挑戰可能是空前的。」

不過王堅也強調指出，今天當大家說AI比人做得好，是AI的成績比人的成績做得好，「只是這張考卷它比人做得好一點而已，當我們出不同考卷，結果會不一樣。過去對比人跟AI的考卷，事實上並不反映人類真正的能力，我們人類需要一張更有創造力的考卷。」

什麼是人類更有創造力的考卷呢？王堅指出，2023年業界覺得寫代碼機器要超越人類還要很久，而到了2025年4月，機器寫的代碼已經超過了大部分人類寫的代碼。「從此人類可以被解放出來，做更具創造性更具有意義的事情。」 大公報記者茅建興

「監管沙盒」模式 寓防範於創新

防止濫用

去年12月，國家互聯網信息辦公室就《人工智能擬人化互動服務管理暫行辦法（徵求意見稿）》（以下簡稱：《辦法》）向社會公開徵求意見。《辦法》提出，國家鼓勵擬人化互動服務創新發展，對擬人化互動服務實行包容審慎和分類分級監管，防止濫用失控。

《辦法》中提到，服務提供者識別出用戶出現過度依賴、沉迷傾向時，或者在用戶初次使用、重新登錄時，應當以彈窗等方式動態提醒用戶交互內容為人工智能生成。用戶連續使用擬人化互動服務超過2個小時的，提供者應當以彈窗等方式動態提醒用戶暫停使用服務。同時，對未成年人使用AI擬人互動服務的保護也提出了要求。

西南政法大學校長、中國法學會副會長林維表示：「《辦法》首次在部門規章層面明確『監管沙盒』制度，允許企業在監管機構的監督指導下開展創新試點和安全測試，在保障企業創新活力的同時，及時發現並防範試點過程中出現的風險。」 大公報記者王莉

責任編輯：林子奇　美術編輯：何振紹