

香港文匯報訊（記者 劉凝哲 北京報道）國家互聯網信息辦公室聯合工業和信息化部、公安部、國家廣播電視總局 14 日聯合發布《人工智能生成合成內容標識辦法》（下稱：標識辦法），自今年 9 月 1 日起施行。標識辦法明確生成合成內容製作傳播各主體的責任義務，重點解決「哪些是生成的」「誰生成

的」「從哪裏生成的」等問題，推動由生成到傳播各環節的全流程安全管理。標識辦法明確人工智能生成合成內容服務提供者對文本、圖片、音頻、視頻、虛擬場景等內容，添加文字、角標、語音、節奏等顯式標識的方法，在內容生成合成環節提出了顯著提示公眾、防範混淆誤認的方案。

內地新規 AI 生成合成內容須顯著標識

規範製作傳播各環節責任義務 不得惡意刪改偽造隱匿



▲內地發布《人工智能生成合成內容標識辦法》，規範製作傳播各環節標識行為。圖為參觀者在展會上了解百度 AI 藝術和創意輔助平台「文心一格」操作。

當前以 ChatGPT、DeepSeek 等為代表的生成式人工智能迅猛發展，借助人工智能生成合成文本、圖像、音頻、視頻、虛擬場景等，逐漸演變為重要的信息生產模式。

文件數據須加隱式標識

然而，伴隨技術門檻降低，虛假信息、深度偽造、惡意輸出等安全問題突出，加劇傳統信息安全風險的動態性、滲透性和全局性。雷軍、張文宏等名人都曾深受 AI 濫用之害，在剛剛結束的全國兩會上，規範人工智能生成內容也是代表委員們關注的焦點之一。

標識辦法明確，人工智能生成合成內容標識主要包括顯式標識和隱式標識兩種形式，顯式標識是指在生成合成內容或者交互場景界面中添加的，以文字、聲音、圖形等方式呈現並可以被用戶明顯感知到的標識；隱式標識是指採取技術措施在生成合成內容文件數據中添加的，不易被用戶明顯感知到的標識。

標識辦法又規定，任何組織和個人不得惡意刪除、篡改、偽造、隱匿規定的生成合成內容標識，不得為他人實施上述惡意行為提供工具或者服務，不得通過不正當標識手段損害他人合法權益。

據介紹，配套標識辦法，強制性國家標準《網絡安全技術人工智能生成合成內容標識方法》已由國家市場監督管理總局、國家標準化管理委員會正式批准發布。

香港文匯報記者查詢此前發布的《互聯網信息服務深度合成管理規定》發現，須添加顯式標識的，包括智能對話、智能寫作、合成人聲、人臉替換、人臉操控、沉浸式擬真場景等多種服務。標識辦法明確，要在生成合成內容的文件元數據中添加隱式標識，隱式標識包含生成合成內容屬性信息、服務提供者名稱或者編碼、內容編號等製作要素信息；鼓勵在生成合成內容中添加數字水印等形式的隱式標識。

AI 標識已成為國際慣例

「生成合成內容標識已成為國際慣例」，公安部第三研究所副所長金波在「網信中國」撰文指出，儘管各國對數字水印、數字指紋、加密元數據等標識技術的具體要求存在不同，對平台標識義務的強度設定也存在差異，但在從事後內容審核轉向生成端風險內嵌控制、提高人工智能治理技術能力的理念是一致的。標識辦法等出台，是中國人工智能法治建設進程中的重要一步。

中國工程院院士、浙江大學教授陳純認為，標識辦法等出台，共同聚焦人工智能生成合成內容服務安全風險，健全了人工智能生成合成內容標識規範，為人類駕馭智能文明提供了至關重要的制度性公共產品，標誌着中國在生成式人工智能領域邁出了構建安全可信生態的關鍵一步。

網傳超一線男星澳門賭輸 10 億 內地警方：拘一男子人工智能造謠

香港文匯報訊 綜合內地媒體報道，日前網絡瘋傳「超一線」男星在澳門賭場連續七天砸錢豪賭，輸掉超過 10 億元人民幣。明星周杰倫的名字同時登上微博热搜第一，其經理人公司杰威爾立即發聲明闢謠。3 月 14 日，警方逮捕造謠的徐姓男子，發現他利用 AI 製造頂流明星在澳門輸 10 億元人民幣的謠言。

周杰倫黃曉明成受害人

根據警方通報，36 歲的徐姓男子為了流量以及非法利益，在 10 日上午疑似使用小紅書的 AI 功能，製作標題為「頂流明星被爆境外豪

賭輸光十億身家引發輿論海嘯」的文章傳播，徐姓男子因擾亂公共秩序已被逮捕拘留。

當時謠言曝光後，網民根據能輸掉 10 億元人民幣身家的經濟能力，猜測周杰倫、黃曉明、相聲演員趙本山及影帝陳道明是最可能的人選。周杰倫與黃曉明後來分別發聲否認。周杰倫 11 日通過杰威爾澄清謠言：「今天出現在微博热搜，關於『某超一線男星在澳門豪賭』等相關消息，部分網友影射此男星是本公司藝人周杰倫。公司特此聲明，該傳聞所有內容均與周杰倫無關！請各界勿再以訛傳訛。」

有網民在黃曉明社媒賬號問「你輸了十個億啊」，黃曉明幽默回覆「歡樂豆（指遊戲虛擬貨幣）嗎」。現在造謠男子已落網，這場鬧劇就此結束。



●周杰倫成 AI 造謠受害人。資料圖片

工業 AI 需求速增 假貨摻雜企業受損

特稿

當人工智能日漸走進工業領域，怎麼判斷供應商 AI 技術的真偽效果？這是當前不少企業在尋求 AI 場景應用時常見的一些難點、痛點。14 日，在廣州南沙越秀 IPARK 粵港智谷產業園區內，聚焦於公輔能源場景的工業 AI 科技公司蘑菇物聯舉行了一場工業 AI 的「打假」活動，披露當前工業 AI 應用中經常出現規則型偽 AI、數據偽裝型 AI、黑箱型 AI、功能嫁接型 AI 等幾大類型，企業須仔細辨別工業 AI 的真偽效果，避免阻礙技術創新生態發展。

披着「AI 皮」傳統軟件充斥市面

某製藥企業相關負責人黃先生對香港文匯報記者表示，他的工廠工程部很早期就開始接觸數智化系統。作為業內走在數字化改造尖端的企業，他們時常要面對絡繹不絕找上門來的 AI 供應商，甄別 AI 真偽效果花費他許多精力。「有些供應商將自己的工業 AI 產品說得神乎其神，但對很多常見的工業應用場景，卻沒辦法提出有針對性的 AI 方案。這對於一些沒有相關專業知識的人，很容易被這些看上去高大上又很唬人的詞彙給糊弄住。」黃先生無奈地說，「假 AI」充斥市場會造

成企業損失，嚴重的會耽誤創新發展進度，影響企業參與行業競爭。

廣東省工業互聯網專家委員會專家、蘑菇物聯創始人兼 CEO 沈國輝表示，近期調研了 100 位工業行業從業者，有 77% 的企業代表反饋，在尋求工業 AI 的場景應用時，判斷 AI 效果真偽成為他們尋求工業 AI 應用落地的第一道「攔路虎」。不少企業表示，需要借助專業人士的「火眼金睛」來幫忙甄別。

「凡是無法將 AI 可視化的都是不靠譜的。」沈國輝介紹說，市面上最為明顯的「假 AI」特徵，就是停留在專有名詞堆砌的 Power Point 簡報階段，而真正 AI 效果客戶根本體驗不到。沈國輝「自爆家醜」說，公司曾有同事為了更快賣出軟件，把還沒有發布的 AI 技術寫到簡報裏，實際上那只是一堆概念和唬人的名詞，這就是典型的「假 AI」。而直至今天，仍有諸多供應商大肆宣傳自有 AI 技術，但其本質都只是披着「AI」之皮的傳統軟件。

科企研垂直大模型甄別 AI 真偽

為了防止這些「假 AI」流通市場，蘑菇物聯自

AI 造假惡劣事件簿(部分)

「被雷軍罵七天」事件：2024 年國慶，各大短視頻平台湧現大量 AI 合成的小米集團創始人「雷軍罵人」視頻，原來是網友通過 AI 平台複製雷軍的聲音生成。視頻中，粗鄙的罵聲令雷軍形象受損，後來 AI 平台下架了雷軍語言模型。今年全國兩會上，身為全國人大代表的雷軍就此提出加強 AI 換臉擬聲違法侵權治理的建議。



●去年底短視頻平台出現大量「雷軍 AI 配音」的惡搞視頻。網上圖片



●不法分子偽造張文宏醫生的形象和聲音直播推銷蛋白棒。網上圖片

張文宏被冒名帶貨：2024 年底，不法分子利用深度偽造技術偽造張文宏醫生的形象和聲音，直播推銷蛋白棒，導致公眾受騙。張文宏本人多次投訴無效，稱此類虛假信息如「蝗災」般氾濫，嚴重損害其專業形象。

「假靳東」詐騙老年人：AI 生成的虛假靳東賬號通過情感誘導騙取老年人錢財。作為全國政協委員的靳東早前也在全國兩會上呼籲加強監管。



▶假靳東賬號主頁。網上圖片

MCN 機構日生成數千篇假新聞：2024 年 6 月中旬，一家 MCN 機構的實際控制人王某某因利用 AI 軟件生成虛假新聞並大肆傳播，擾亂公共秩序，被警方行政拘留。據通報，王某某經營 5 家 MCN 機構，運營賬號 842 個，自今年 1 月以來，王某某通過 AI 軟件生成虛假新聞，最高峰一天能生成 4,000 至 7,000 篇。

民眾偽造災難現場干擾救援：2024 年初發生的雲南山體滑坡災害中，楊某某利用 AI 生成虛假新聞稿件，捏造傷亡數據，並通過自媒體平台傳播，引發社會恐慌。此後，公安機關對楊某某處以行政拘留並罰款。2024 年底西藏定日發生地震後，一張「小男孩被埋」的圖片在社交媒體上瘋傳，但最後被證實是 AI 生成，造成惡劣影響。

●整理：香港文匯報記者 劉凝哲



●蘑菇物聯公司昨日在南沙舉行工業 AI「打假」活動。香港文匯報記者胡若璋 攝



●工業 AI 應用在汽車行業中，有效提升公輔系統能效水平。香港文匯報廣州傳真

主研發了工業公輔能源系統的 AI 垂直大模型靈知 AI，它可掌握超 20 多類高耗能通用設備的設備機理和專家知識庫，具備問答、診斷、預測、控制四大模型能力，能有效助力製造業企業應用 AI 時快速甄別真偽。該模型已入選國家工業和信息化部人工智能賦能新型工業化典型應用案例。

深圳「孔雀計劃」海外高層次人才、香港中文大學系統工程與工程管理博士周子葉接受香港文匯報記者採訪時表示，工業企業不應為 AI 噱頭埋單，而應當為價值埋單。因此，真正的 AI 應該是可落地的，可創造可測量價值，且可持續創造價值，賦能產業高質量發展。

●香港文匯報記者 胡若璋 廣州報道