借助人工智能製造贋品 各行業紛出招辨真偽

AI以假亂真 掀人類反擊戰

◆Deepfake影片模擬影星湯告魯斯的樣貌。 網上圖片

◆人工智能工具「Dall-E-2」近年興起，可以繪製與特定畫家風格相近的作品。 網上圖片

◆紐約校方人員商議學生濫用ChatGPT問題。網上圖片

效果逼真的「Deepfake（深偽技術）」影片、談吐自然的聊天軟件，甚至是可以參加大學考試的短篇作文……。隨着人工智能技術（AI）不斷發展，AI的文字、圖片及影像作品都愈發精細，甚至可以達到以假亂真的效果。各行各業爲此紛紛出招，想方設法在AI與人類的作品中辨認真偽。AI技術專家指出，AI的本質是在分析大量資料的基礎上依照指示加以模仿，而發掘人類獨有的個性和創造力，仍是辨明真偽的關鍵所在。

AI作品「形似神不似」難取代藝術創作

藝術模仿

AI可以模仿特定畫家風格生成全新圖像，不過許多藝術家指出，現有AI技術還停留在模仿而非創作階段，本身並不具備人類藝術創作所需的想像力和創造力。由此一來，AI作品往往「形似神不似」，該技術可為藝術家創作帶來啟發，但難以取而代之。

帶刻板印象 亂添名作元素

《金融時報》藝術編輯達利指出，生成式人工智能工具「Dall-E-2」近年興起，可以繪製與特定畫家風格相近的作品。不過達利觀察到，Dall-E-2的作品往往帶有刻板印象，即使與指定的主題並不相稱，也會添加一些指定畫家名作的元素。例如模仿梵高的作品，就會出現與《星空下的咖啡座》類似的深藍色調；模仿奧地利畫家克林姆特的作品則會照搬他的「調色盤」，大量使用鐵銹色、黑色和亮黃色調。

德國藝術家德雷赫斯特還創建一款檢索工具「Spawning」，允許藝術家設置權限，避免部分AI公司收集他們的作品納入AI資料庫。德雷赫斯特解釋稱，這款工具能滿足部分藝術家希望保留獨特風格的需求，不過在他看來，人類的想像力沒有限制，「AI可以嘗試複製人們多年前的作品，但並不能預見我們人類未來會如何創作。」

美大學生開發應用程式查ChatGPT代寫

學術影響

聊天機械人程式「ChatGPT」近日被不少學生用於作文或探索行文思路，為了讓這些「請槍」的作品「現形」，在普林斯頓大學攻讀電腦科學的4年級學生愛德華．田（Edward Tian，譯音）便開發了一款應用程式「GPTZero」，聲稱能有效地檢查文章是否為ChatGPT撰寫。紐約市教育局更頒布新規，直接禁止學生使用ChatGPT，以免為學生帶來負面影響，影響學生批判性思考和解決問題的能力。

愛德華．田在社媒上展示了GPTZero的運作方式，他利用一個資料庫訓練GPTZero，並使用「困惑度」和「突發性」作為檢測指標。在辨析文章時，如果GPTZero感到「困惑」，就說明這些文本內容複雜，有可能是人手編寫的，反之則說明文本或是AI自動生成。突發性則用於比較句式之間的變化，通常人類寫作時會頻繁作出調整，包括同時使用一些較複雜的長難句與較短的語句，但AI編寫的文本句式往往更加整齊劃一。

紐約校禁學生用ChatGPT

紐約市教育局上周證實，當地公立學校的電腦和網絡都不再開放ChatGPT，學校系統也會自動攔截ChatGPT網站。教育局發言人萊爾指出，局方質疑ChatGPT產出內容的安全性和準確性，擔心會為學生帶來負面影響，「雖然這項工具可以提供快速且簡單的答案，但它不會培養人們擁有批判性思考和解決問題的能力，無助個人成長和學術進步。」

澳洲8間頂尖大學組成的八校聯盟更提出新規，強調使用AI提交功課或參加考試都屬於作弊。聯盟據此重新修訂了今年的學生考評方式，包括安排更多紙筆測試，以及增派人手監督考場紀律。

至於學生們能否用AI作「槍手」，有記者向ChatGPT提出這問題，ChatGPT回答稱，這種做法並不合適，也對學生長遠發展不利，「要辨別某些作品是人類還是AI編寫，你可以尋找文案中是否缺乏個人經歷或情感、寫作風格不夠連貫，或是重複使用短句，這些可能都是AI撰文的特徵。」

英特爾推偵測器 96%準確率極速識別Deepfake影片

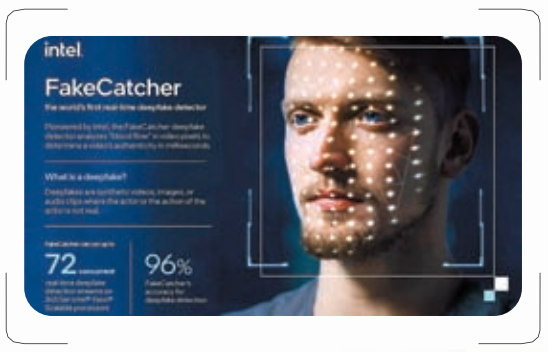
偽造影片

用於加工合成偽造影像的Deepfake（深偽技術）近年在網上氾濫，引發不少爭議。美國科企英特爾為此推出一款專門偵測此類偽造影片的偵測器「FakeCatcher」，據稱準確度高達96%。英特爾表示，這是全球目前首個能夠以毫秒速度，即時報告檢測結果的Deepfake偵測器。

判斷血流方向尋真實部分

傳統Deepfake偵測應用程式需要上傳影片進行分析，利用人工智能（AI）的深度學習技術，試圖辨識出影片不真實的部分，全程往往需時數小時甚至更久。英特爾解釋，FakeCatcher的原理是反其道而行之，改為判斷影片人像的血流方向，尋找影片真實的部分。由於人類的心臟跳動向全身輸送血液時，人體靜脈的顏色會發生微妙的改變，FakeCatcher正是收集臉部的血流訊號，利用深度學習技術迅速判斷影片真偽。

英特爾已將FakeCatcher應用在其名下的硬件和軟件中，希望該偵測器日後可用於社交平台、新聞機構和非牟利組織，讓更多人能夠偵測Deepfake影片，及時阻止這些偽造影片在網絡擴散。



◆「FakeCatcher」準確度高達96%。 網上圖片

麻省理工培訓志願者辨認AI文案影片

防假資訊

AI技術的作品近乎以假亂真，想要練成「金睛火眼」辨識真偽，自然要下一番功夫。美國麻省理工學院先進虛擬技術中心（CAV）的博士生格羅就推出「Detect Fakes」研究項目，培訓參與研究的志願者如何辨認AI文案和影片，更容易發現其中的細微差別。

Detect Fakes的培訓方法並不困難，團隊會為志願者提供32組文本、圖像和影片範例，讓志願者辨認哪些是人類創作，哪些是由AI偽造而成。團隊還會向志願者講解分辨Deepfake（深偽技術）製造影片的方法，包括留意影片中人像的嘴唇動作是否與發音一致。格羅表示，人們可以透過學習分辨AI作品，思考利用AI技術作偽裝的原因，更加留意分辨信息的真實性。

模擬太陽神11號登月失敗

CAV團隊近年還利用美國太空總署（NASA）的舊影像，製作一款名為《一旦發生登月災難》的短片，描述1969年美國太陽神11號登月任務一旦失敗的情形。該片不但用Deepfake技術模擬了時任總統尼克松宣讀備用演講的場景，還添加1960年代電視特有的模糊影像和背景噪音特效。

團隊指出，該短片旨在提醒人們提高警惕，「如果體驗過這個影片的人，日後能夠更謹慎看待社媒上彈出的最新資訊，我們就成功了。」

◆團隊為志願者提供文本、圖像和影片範例，讓他們辨認。 網上圖片

對少數族裔準確度偏低 美非裔坐7天冤獄

人臉識別

AI近年常被用於人臉識別技術，但執法部門使用並未完善的技術作為拘捕疑犯的證據，反而可能傷害到無辜人士。《每日郵報》近日報道，美國佐治亞州28歲非裔居民里德因被AI識別為盜竊犯被捕，然而他被指盜竊的「作案地點」，距離他的住所近1,000公里，他無辜被拘留7天，才能自證清白獲釋。

無視千公里外犯案疑點

里德在去年感恩節後不久被捕，警方稱他的面部特徵與路易斯安那州一宗盜竊罪的疑犯相符，即使里德辯稱自己根本沒去過路易斯安那州也無濟於事。警方拘捕里德後，也沒有詳細查問他的經歷，或是核對他的身高體重等。事實上只需簡單比較就能發現，里德與疑兇的面貌有諸多不同，尤其是里德臉上有一顆明顯的痣，疑兇則沒有；里德的體重也較疑犯輕40磅。在律師據理力爭下，里德終於在上月初獲釋。

美國過往利用人臉識別技術偵查的案件中，已有至少3名無辜的黑人被錯誤拘捕。不少AI專家都指出，現有人臉識別技術針對少數族裔和女性的準確度偏低。美國許多地方的警員也認為，人臉識別只能用於提供線索，要使用該技術拘捕疑犯，必須得到上級批准。