無聲的國安危機——深度偽冒詐騙技術

／勤業眾信風險諮詢服務部門

人工智慧（Artificial Intelligence, AI）演進至今的深度學習（Deep Learning, DL），逐漸地貼近人類最根本的核心需求。然而，商業自動化與智慧化引入資訊與通訊科技，資安威脅自然也隨之而來。根據 2020 資安預測報告[1]，利用 AI 技術進行深度偽冒詐騙（下簡稱「Deepfake」）之手法將大幅提升。或者，遠距工作及連網家庭設備興起，增加駭客竊取企業情資與網路釣魚勒索機會，皆是 AI 技術被不當使用或濫用。尤其，新型冠狀病毒（COVID-19）疫情蔓延全球，連帶提升企業、學校及行政機關對遠距工作的需求，駭客組織將此視為無聲地發動 Deepfake 攻擊的好時機。

[1] 趨勢科技公布 2020 資安預測報告（2019/12/13）。iThome。取自：https://www.ithome.com.tw/pr/134797。

美國各大社群媒體瘋傳裴洛西（Nancy Pelosi）的變造影片，影片中她的說話速度被放慢，因此聽起來含糊不清，像喝醉酒且更顯老邁，之後美國總統川普及其律師朱利安尼皆轉發並評論了這段影片。（Photo Credit: Donald Trump's Twitter）

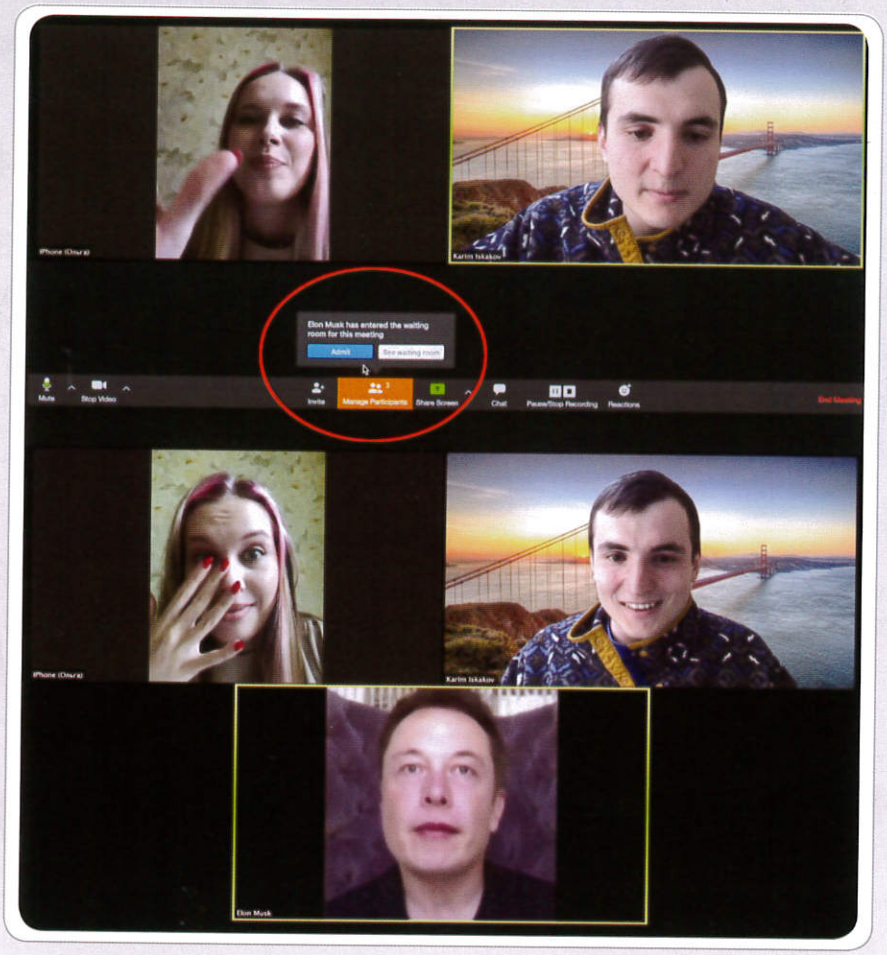
2018 年美國演員喬登皮爾運用 AI 換臉技術，合成美國前總統歐巴馬說出：「川普總統完全就是個笨蛋」。（Photo Credit: BuzzFeed Video, https://youtu.be/cQ54GDm1eL0）

深度偽冒詐騙的崛起及威脅

從 2017 年第一起開始，美國社群（Reddit）上「deepfakes」的用戶透過 AI 技術，將好萊塢女星蓋兒加朵（Gal Gadot-Varsano）的臉移花接木到色情影片中[2]，到 2018 年美國演員運用 AI 換臉技術，合成美國前總統歐巴馬並對著鏡頭說出：「川普總統完全就是個笨蛋」[3]。

另外，2019 年 5 月，美國各大社群媒體瘋傳裴洛西（Nancy Pelosi）的一段變造影片，影片中她的說話速度被放慢，因此聽起來含糊不清，好像喝醉酒且顯得更加老邁。美國總統川普（Donald John Trump）及川普個人律師朱利安尼（Rudolph W. Giuliani）分別前後在 Twitter 轉發裴洛西（Nancy Pelosi）的一段變造影片。朱利安尼推文寫道：「南西．裴洛西怎麼了？她的語言模式很古怪。」然而，川普轉貼影片後還寫道：「裴洛西整場記者會都在結巴。」[4,5]

[2] AI 假色情終究來臨：神力女超人 Gal Gadot 臉被移花接木到 A 片中（2017/12/13）。INSIDE。取自：https://www.inside.com.tw/article/11386-gal-gadot-fake-ai-porn。
[3] 黃貞怡（2018/08/13）歐巴馬罵川普笨蛋？ AI 換臉製造假新聞。TVBS 新聞網。取自：https://news.tvbs.com.tw/world/973044。
[4] 川普 PO 變造影片打對手 裴洛西變老說話結巴 [影]（2019/05/25）。中央通訊社。取自：https://www.cna.com.tw/news/firstnews/201905250194.aspx。
[5] 何蕙安（2020/03/09）因應美國總統大選，YouTube 拿出辦法對付選戰亂象。台灣事實查核中心。取自：https://tfc-taiwan.org.tw/articles/2943。



2020 年初，有人在視訊會議軟體「偽裝」特斯拉 CEO 馬斯克，並在 Github 上提供原始碼。（Photo Credit: Ali Aliev, https://youtu.be/IONuXGNqLO0）

2019 年 9 月，不肖人士透過語音合成方式，冒充英國能源公司德國母公司的 CEO 進行詐騙，金額高達 22 萬歐元，這也是第一起 Deepfake 的 CEO 詐騙[6]。2020 年初，有人在視訊會議軟體「偽裝」特斯拉 CEO 馬斯克，並在 Github 上提供原始碼，皆是 Deepfake 的運用情境[7]。

淺談深度偽冒詐騙型態

根據 AI 法論評論網，Deepfake 種類可以歸納出四種型態[8]，分別為臉部替換、人臉生成、臉部再製、聲音合成，各類型 deepfake 皆有極限，臉部再製可以保持一個人臉部的特徵不變而使照片看起來更加

[6] 愛范兒（2019/09/05）AI 語音模仿老闆聲音要求轉帳，成功騙走近 770 萬元。科技新報。取自：https://technews.tw/2019/09/05/fraudsters-voice-ai/。

[7] 【Deepfake 入侵視訊會議】工程師推出最新假冒馬斯克濾鏡，Github 上已開源（2020/04/20）。科技橘報。取自：https://buzzorange.com/techorange/2020/04/20/deepfake-fun-effect/。

[8] 你所見不一定為真— AI 與 DeepFakes（2019/12/29）。AI 法論評論網。取自：https://www.aili.com.tw/message2_detail/54.htm。

逼真。臉部替換則需要分別取得目標和素材兩個人各種不同角度的影像。在聲音合成技術成熟以前，臉部再製和臉部替換都需要有專人配音。

隨著網路上越來越多製作各種型態詐騙應用程式的出現，即使是對 AI 一竅不通的外行，只需要一個 GPU 和一些訓練數據，再通過按部就班的操作也能製作出影片，其影響面向除了對個人的影響，甚至包含至國安、金融與政治領域。

深度偽冒之核心技術

Deepfake 主要使用了生成對抗網路（GAN）與自編碼器（Autoencoder）的技術與概念[9]。首先，有一組編碼器模型（Encoder）需要從人臉中提取表情的

臉部替換

使用某人臉部的影像，替換掉另一個目標人物的臉，目標人物的臉被覆蓋了，重點在換上去的臉

臉部再製

主要是透過修改目標人物的臉部表情，例如移動他們的嘴巴、眉毛和眼睛，臉部再製的目標並不是取代他們的臉部特徵，而是更改他們的面部細節來使照片傳達的訊息不同。

（Source: Government Accountability Office, https://www.gao.gov/products/GAO-20-379SP#summary）

人臉生成

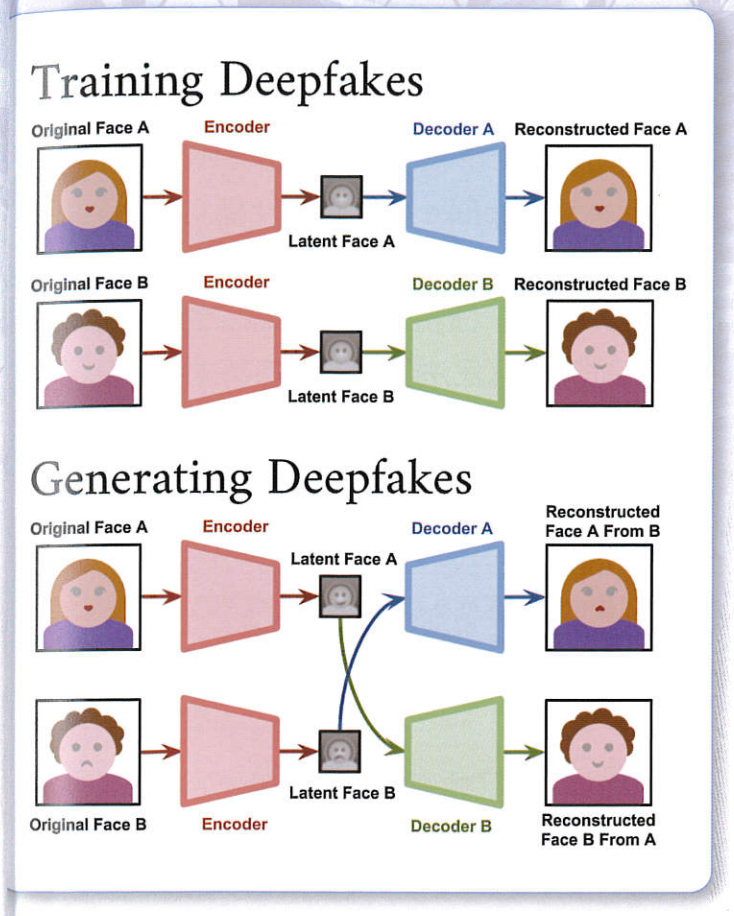
這項技術能夠創造一個全新的臉，是利用一種新興的生成對抗網路（Generative Adversarial Network, GAN）的深度學習技術，這種技術使用兩個神經網路相互對抗，其中一個生成影像，另一個負責判斷生成的影像是否夠好。

圖中男女均為使用 GAN 技術合成之影像，非實際存在之人物。

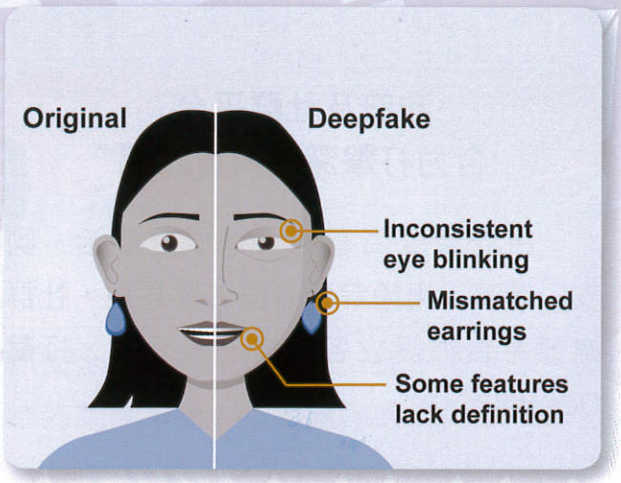
聲音合成

Deepfake 的分支，可以合成出一個人的聲音且使用特定人的說話方式和語調唸出文字。另有的聲音合成產品，例如 .ai，則是能讓使用者挑選聲音的年紀和性別，而非模仿某個特定人的聲音。

[9] 王淳中（2020/01/09）讓臉書忍不住出手封鎖的 Deepfake 影片，是什麼技術？台灣人工智慧學校。取自：https://aiacademy.tw/what-is-deepfake/。



Deepfake 主要使用了生成對抗網路與自編碼器的技術與概念，有一組編碼器模型（Encoder）需要從人臉中提取表情的特徵，另外一組解碼器（Decoder）則需要將表情特徵還原成某個人的人臉。（Photo Credit: Alan Zucconi, https://www.alanzucconi.com/2018/03/14/understanding-the-technology-behind-deepfakes）



可從四大方向判斷是否為 Deepfake 影片，其中包含眨眼的頻率與模糊的痕跡等。（Source: Government Accountability Office, https://www.gao.gov/products/GAO-20-379SP#summary）

特徵，另外一組解碼器（Decoder）則需要將表情特徵還原成某個人的人臉。在訓練階段時，會有另外一個分辨器（Discriminator）協助解碼器作訓練，分辨器則會盡量正確分辨真實的人臉與合成的人臉，而解碼器則會盡量產生出分辨器分不出來的合成人臉，因此兩者會互相對抗且逐漸變強。在訓練完成後我們就可以將一張新的人臉透過編碼器提取表情特徵，再透過解碼器還原成特定人的人臉。

一般來說，就目前生成對抗網路成像技術之成熟度而言，「正面」的「大頭照」是較容易成像的條件，側臉照或有非臉部器官入照將會大幅增加訓練的難度與降低準確性。

對於 Deepfake 影片，根據白金漢大學數學與電腦教授 Sabah Jassim 及 Spectre 聯合創辦人 Bill Posters 建議，可以注意四大方向[10]：(1) 眨眼率：深度偽冒影片所生成之目標對象其眨眼率少於正常人；(2) 語音和嘴唇運動的同步狀況；(3) 影片情緒不吻合及 (4) 模糊的痕跡、畫面停頓或變色。

[10] 什麼是 Deepfake（深偽技術）？ A 片女主角也可以造！（2020/03/03）。資安趨勢部落格。取自：https://blog.trendmicro.com.tw/?p=63452。

政府及社群平台合力打擊深度偽冒詐騙

由於深度偽冒影片造成之政治、國安、經濟、輿論等各方面影響甚鉅，社群媒體平台亦設立各項措施及行動，打擊 Deepfake。

除社群平台大力打擊外，各國亦透過立法防範，希望藉國家手段遏止不實或不當行為。Deepfake 影響範圍及權利甚廣，以我國律法言，Deepfake 若涉及毀損他人名譽，可主張〈刑法〉第 310 條誹謗罪，當事人亦可依《民法》第 184 條請求損害賠償，及同法第 195 條回復名譽。

YouTube

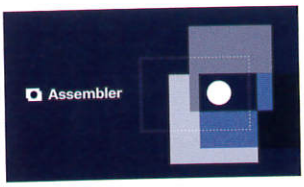
YouTube 在其現行社群準則內的垃圾郵件、欺詐行為和詐騙政策裡就有禁止變造媒體資訊的規定。他們的官方部落格「YouTube 如何支持選舉」文章中，對於會誤導選情的內容採取零容忍態度，例如可能誤導使用者和散播不實資訊的變造過影片，包括 Deepfake（深度偽冒詐騙）。關於候選人或投票過程的不實內容，無論是否變造過都不允許。

（Source: https://blog.trendmicro.com.tw/?p=63452）

twitter

Twitter 關於合成和變造媒體資訊的規則更新中也包括了 Deepfake（深度偽冒詐騙），並未具體提到即將舉辦的美國大選。它單純宣布禁止「可能造成傷害」的合成或變造媒體資訊。

（Source: https://blog.trendmicro.com.tw/?p=63452）



Google 旗下子公司 Jigsaw（原 Google Ideas）與 Google Research 合作開發 Assembler 工具，目的就是為了幫助新聞媒體與事實查核人員辨別圖片影像是否經過變造修改，進而防止假訊息傳播。

（Source: https://udn.com/news/story/11017/4334873）

facebook

Microsoft

2019 年 9 月，微軟和臉書發起 Deepfake Detection Challenge 大賽，以高額獎金鼓勵外界開發出能辨識惡意偽造影片的偵測模型。

（Source: https://www.ithome.com.tw/news/132917）

各社群平台打擊深度偽冒詐騙方式

結語

我國政府於 2016 年宣示「資安即國安」的戰略目標[11]，於 2018 年推動「3x3x3 國家級資安戰略」[12]，2019 年加速研擬「資安即國安 2.0 戰略」，不斷地提高資安人才培訓能量，開發資安產業創新技術，鞏固資訊安全，打造臺灣成為堅韌資安之國。另外，政府也以「集結防制假訊息」、「電腦犯罪」及「網路犯罪」等 3 大目標，於法務部調查局設立資安工作站，強化政府打擊各類網路犯罪的能力。

而面對網路威脅與資安事件，除情資蒐集與運用的重要性日益彰顯外，應透過擘劃戰術與技術的策略，在戰術上，分析資訊安全威脅的戰術、技術、程序，及規劃對應的防護措施並精準投入相對應的防護資源。技術面上，透過掌握資訊安全威脅的相關資訊，進一步定義更精準的威脅指標；透過確保數位安全、建立資安體系、推動資安自主研發等相關量能，將侷限轉化為超越，進而為組織挹注足夠強度的防禦能力和應變能量。

因此，「掌握資安威脅態樣、擘劃堅實防護戰略」，已然成為在資安防護的策略擬定上依循的方向，亦可使機關、企業能更全面因應可能出現的資訊安全威脅。

[11] 黃彥棻（2018/09/14）臺灣首部國家資安戰略報告出爐，總統蔡英文：資安是國家重要戰略選擇。iThome。取自：https://www.ithome.com.tw/news/125913。

[12] 蔡總統：加速研擬資安即國安 2.0 戰略（2019/11/11）。中央通訊社。取自：https://www.cna.com.tw/news/aipl/201911110143.aspx。