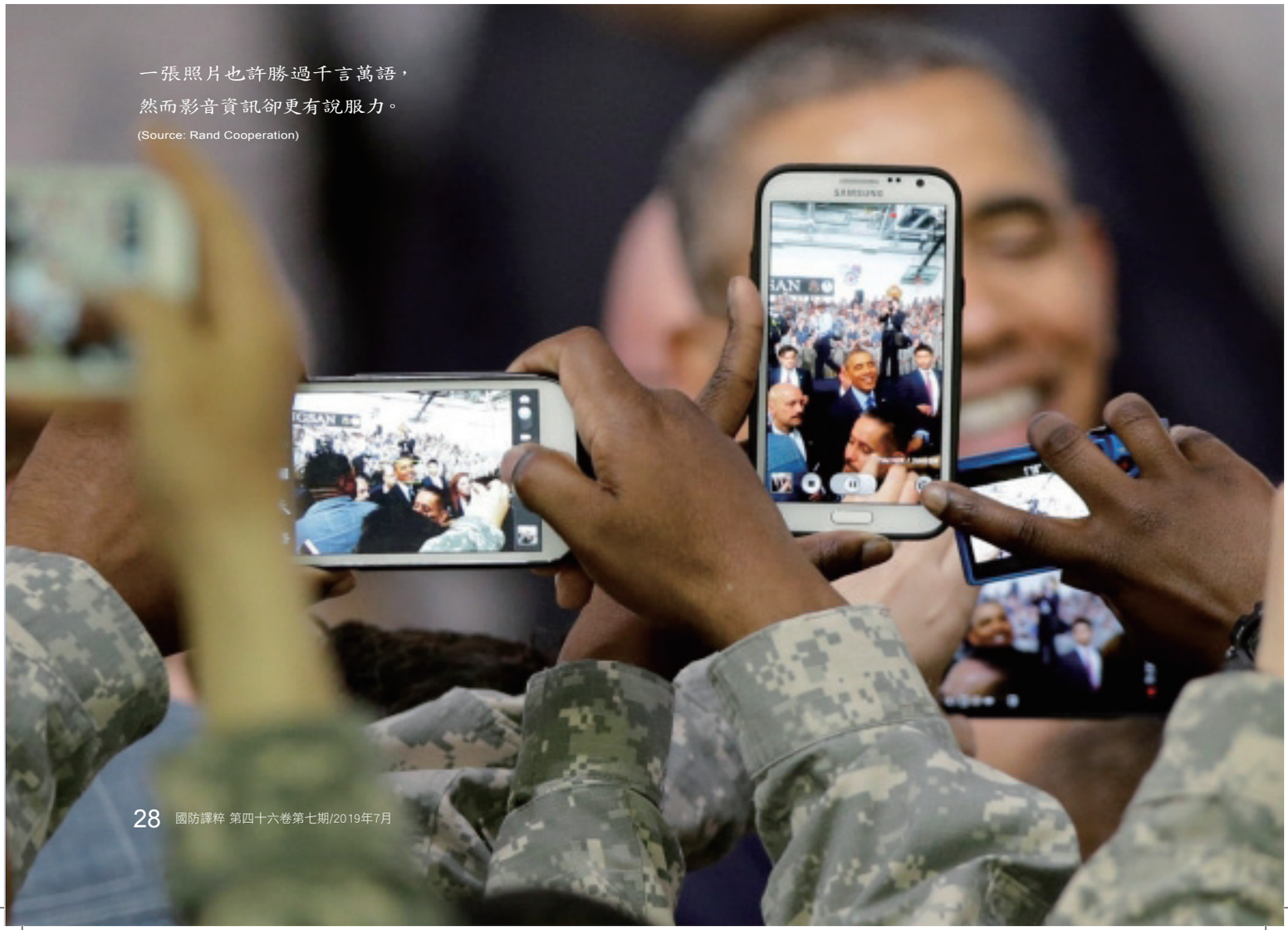
作者/Robert Chesney and Danielle Citron　譯者/余振國　審者/馬浩翔

後真相時代：深度偽造

Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics

取材/2019年1-2月美國外交事務雙月刊(*Foreign Affairs*, January-February/2019)

拜科技所賜，眞與假的界線變得愈來愈模糊。深度僞造係人工智慧發展下的產物，現已擴及全球，也難以避免其產生不可預知的風險。相關部門應做好前瞻技術研究，掌握因應之策。



一張照片也許勝過千言萬語，
然而影音資訊卻更有說服力。
(Source: Rand Cooperation)

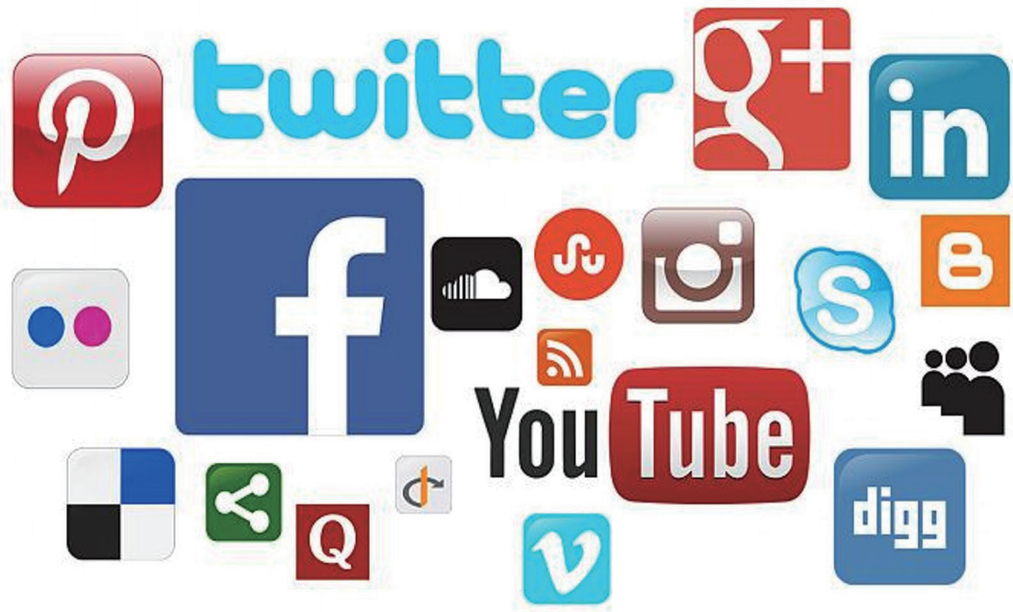
一張照片也許勝過千言萬語，但一次事件的影音卻更有說服力。當各方支持者無法在某個事實上達成共識時，大家似乎都樂見能有具如此說服力的資訊來澄清事實。影音紀錄讓大家都能成為事件的第一手目擊者，而無需決定是否要去相信他人描述。再加上智慧型手機讓錄製影音內容變得更容易，還有社群媒體平臺讓民眾可以分享與消化這些內容，導致目前大家可仰賴親眼所見、親耳所聽這種前所未有的方式。

但這也會帶來巨大的危機。試想，若有一段影片描述以色列總理與幕僚祕談有關將在德黑蘭進行一連串政治刺殺，或是一則錄音檔，內容是伊朗官員在計畫密謀行刺伊拉克某特定省遜尼派領袖，或一段影像，顯示某美籍將軍在阿富汗境內燒毀一本可蘭經的影像。現今暴力充斥的世界，這類影音具極強大煽動力。接著想像這些錄製檔案能被偽造，且幾乎任何擁有筆記型電腦及網路的人都可以取得偽造工具，同時做出來的影像幾能以假亂真，讓人們幾乎不可能辨別虛假跟真實影像之間的差別。

數位科技的進步可能很快就能讓這個惡夢成真。透過「深度偽造」(deepfake)這種真實度高，而且難以被察覺的數位影音作業，如今利用這種手段所產製的影音檔，要捏造某人曾說或做出某件子虛烏有的事變得比以前更容易。更糟糕的是，創造這些深度偽造的手段可能受到快速傳播，使利用它來達成政治目的人士不斷增加。假消息這種年代久遠的技藝，只不過在今日世界中改頭換面而已。但如果深度偽造技術繼續發展與散播，看起來就像過去舞刀弄劍時代的浮誇文宣一樣荒誕可笑。

深度偽造的濫觴

深度偽造是人工智慧最新發展產物，被稱為「深度學習」(deep learning)，它是由一組名為類神經網路(neutral networks)的演算法構成，用途是透過篩選龐大資料集來推斷其中規則並且複製模式(例如谷歌就是利用這項技術，來為自家搜尋引擎研發強大的圖像歸類演算法)。深度偽造是由一種特定類型的深度學習衍生出來，這種深度學習是以兩種相互對抗的演算法所組成，形成



近十年間，有愈來愈多人開始從臉書或推特這類社群媒體平臺取得資訊，由於這些平臺屬於自媒體性質，導致這些內容大多未經過濾。(Source: Wiki)

所謂的生成對抗網路(Generative Adversarial Networks，簡稱GANs)。其中一個演算法是「產製者」，會在來源資料中創製仿造的內容(例如從存有真貓圖片的資料庫中擷取資料來仿造貓圖片)，而另一個演算法則是「鑑別者」，會試圖找出資料中的仿造內容(例如從圖片中辨識出那些是假的貓圖片)。由於這兩個演算法會不斷透過對付對方來提升自身能力，因而導致兩者能力急速成長，讓生成對抗網路可製造出足以以假亂真，但實際上卻是偽造的影音檔內容。

這項科技具有被大幅傳播的潛力。如今在公開市場上已出現商用、甚至免費的深度偽造服務，而在黑市上則可能會出現安全監控更低的版本。散布這些服務將會降低門檻，這代表很快限制個人製造深度偽造的唯一方式，只剩下是否能取得讓演算法互相訓練的素材，使之用在生成對抗網路，亦即偽造目標的影音檔。如此一來，只要有足夠興趣就會知道去何處求援，幾乎任何人都有能力創造出具專業水準的偽造內容。

其實深度偽造有許多具有價值的應用方式。例如可以修改歷史人物的影音資料當作教材。有一家公司甚至宣稱能使用這項科技來幫助因病失語的病患，讓渠等可以再度與人對談。但深度偽造可以(也必然會)遭到不當使用。目前已有人在未告知或取得他人同意下，利用深度偽造技術將臉植入色情影片中。製造假錄音或錄影的方

法愈形簡單後，將會讓有心人士更有機會進行勒索、恫嚇，以及破壞行為。但深度偽造技術最可怕的用途，是用在政治與國際事務的領域。在此領域中，深度偽造可製造特別能發揮作用的謊言，這些謊言能夠煽動暴力行為、破壞領導人或機構的信譽，甚至可以左右選情。

深度偽造之所以能特別具破壞力，是因為現在技術已做到真假難辨的程度。在二十世紀的絕大多時候，雜誌，報紙和電視媒體向大眾傳播資訊的流向。在這段時間裡，記者建立嚴格的專業標準來控制新聞品質，而相對較少的大眾媒體管道，則意味著只有少數個人和組織得以大規模發送資訊。但在最近十年間，有愈來愈多人開始從臉書或推特這類社群媒體平臺取得資訊，也因為這些平臺依賴使用者生產內容，導致這些內容大多未經過濾。使用者通常會依照自身喜好來使用平臺，因此通常會遇到與自己想法有所共鳴的觀點(這是因為在平臺演算法加持下所造成之趨勢)，導致自己的社群媒體動態成為一個同溫層。同時，這些平臺也很容易遭到資訊轟炸，也就是當人們在懶於查證資訊是否真實的情況下就將資訊傳遞出去，導致該資訊在過程中可信度變得愈來愈高。這一切都讓謊言廣傳的速度大勝以往。

這些原因會讓社群媒體成為讓深度偽造惡性循環的溫床，並且可能對政治造成爆炸性影響。俄羅斯試圖透過在臉書與推特上散播充滿爭議與政治性煽動的訊息，影響2016年的美國總統大選，這項舉動就已向世人展示，要將假資訊植入社群媒體中是多麼容易的一件事。往後深度偽造將會更加生動且真實，因此會比2016年的假新聞更容易被分享廣傳。再加上人們特別偏好分享內容是負面或新奇的資訊，內容愈腥羶的深度偽造訊息，愈容易廣傳。

民主式詐欺

自古以來，人們善用詐欺、偽造，以及其他形式的騙術來影響政治從來就不是什麼新鮮事。1898年，美海軍緬因號戰艦(USS *Maine*)在古巴哈瓦那港爆炸時，美國國內的小報就刻意以誤導方式披露該事件，藉此煽動公眾偏向對西班牙宣戰。反猶太主義的《錫安長老會紀要》(*Protocols of the Elders of Zion*)書中虛構猶太陰謀，卻在二十世紀前半葉廣為流傳。而在現今，如Photoshop繪圖軟體這類技術，讓改圖變得與竄改文字一樣簡單。深度偽造之所以顯得前所未見，是因為它結合影音品質及應用等更具說服力的模式，以及難以測得其內容虛實。同時，當深度偽造技術散布後，愈來愈多人得以利用過去僅在好萊塢攝影棚或經費充足的情報機構，才能夠採取的手段來操弄影音內容。

深度偽造技術對非國家的惡意人士特別實用，例如叛亂團體或恐怖組織，這些團體向來以資源缺乏的方式，去製播並散布可信度高的造假影音。現在這些團體可以在合成內容中播放渠等敵人(包含了政府官員)大喊煽動性字眼，或者做出挑釁行為，並且精心挑選出特定內容，盡可能對目標閱聽人產生最大影響。舉例來說，一個伊斯蘭國(ISIS)的黨羽組織可製作一段影片，其中是有關美軍士兵射殺平民或計畫炸掉一座清真寺，藉

此讓組織的招募行動獲得更多響應者。在目標閱聽人對於深度偽造資料中的人物，已抱持不信任態度的地區，更難揭穿這類造假影像。相對的，某國亦可使用同樣的深度偽造技術，來打擊非國家對手。

如今美國與其他國家的內政都被假資訊戰所害，而深度偽造的出現必然讓這些亂象加劇。2016年由俄羅斯政府贊助的假資訊戰，明顯成功加深美國國內現存的社會鴻溝。其中一例就是某個在社群媒體中自稱與「黑人命也是命」(Black Lives Matter)運動有關的假俄羅斯帳號，其散發煽動內容，目的在於加深種族間的劍拔弩張。未來這些假資訊模式將不只在推特或臉書上的貼文，而是利用一部白人警察咒罵歧視字眼，或是某位「黑人命也是命」的支持者鼓吹暴力行為的造假影片。

但深度偽造最致命的威脅，就是能透過某個時機抓很好的合成內容來影響選舉結果。2017年5月，俄羅斯政府就曾做過如此嘗試。在法國大選投票前夕，俄羅斯駭客釋出大量竄改過的遭竊文件，試圖破壞總統參選人馬克宏(Emmanuel Macron)的選情。這項行動因為數項原因而失敗，其中包括文件內容相當無趣，以及法國媒體法在投票前的44小時禁止任何與選舉相關的報導。但多數國家並未有類似媒體管制作為，再加上深度偽造以假亂真的本質，就會使得原本已極具殺傷力的內容，絕對能造成更嚴重的影響。若在投票前24小時，社群媒體上出現一部馬克宏坦承貪腐的影片，這部影片想必會如野火般的快速蔓延，甚至根本也沒有足夠時間讓人查證該影片的虛實。

同時，深度偽造還有可能以另一種間接形式侵蝕民主體制。其並不僅是加深社會與思想上的分歧問題，它還會形成所謂說謊者紅利：當人民更覺知到深度偽造存在時，倘若公眾人物某次真的被逮到行為不檢，渠等更容易煽動大眾質疑不利證據的可信度(假如深度偽造在2016年美國總統大選期間就很猖獗，不難想像當時川普得以輕鬆利用這項資訊，來辯駁那段錄音的可信度，他在錄音中吹噓自己如何對女性上下其手)。以更廣泛的面向來說，當大眾對於深度偽造的威脅更為敏感後，對於新聞的可信度就會下降。而記者們亦會因擔心證據可能是假的，就會變得不敢輕易相信最新消息的影音資訊，更遑論發表該則新聞。

深度修復

對抗深度偽造沒有一種完美解決方式。目前有數種既存的法律或技術可用來與之抗衡，之後也會出現更多新手段，但這些方法都不足以完全解決這個問題。與其尋找一勞永逸的方法，不如試圖適應深度偽造的興起。

有三項科技作法值得注意。首先是鑑定技術，或是透過科技偵測偽造物的技術。在研究者投入大量心力來創造高可信度的仿造物，渠等同時也在發展偵測手段的強化方式。2018年6月，達特茅斯學院與紐約州立大學奧爾巴尼分校的電腦科學家宣布創造一個可偵測深度偽造的程式，透過當影中人在眨眼時，去搜尋其不正常的眼皮跳動模式。但在與深度偽造的

軍備競賽中，這種突破只會讓對手知道下一次該如何改進，如此一來，未來訓練生成對抗網路的影片中，就會加入正常眨眼的案例。況且就算有非常能幹的偵測演算法問世，根據深度偽造在社群媒體上的擴散速度，依然會讓揭穿這些虛妄變得困難重重。屆時等到警鐘敲響前，傷害早已造成。

第二種科技解決手段，是內容在傳播前先予以驗證，這就是有時被稱為「數位出處」(digital provenance)的作法。如美加州Truepic照片影像驗證公司，已著手研發可在聲音、照片及影像內容被創造的瞬間，利用元資料(metadata)固定登入至離散分列帳或區塊鏈的原理留下數位浮水印，換言之，就是可以記錄任何內容的真實性，而這些紀錄隨後可用來比對疑似偽造內容以進行辨識。

理論上，數位出處是一種理想的解決方案，但在實作上會碰到兩大問題。首先，此作法需要鋪天蓋地配置在各式各樣裝置上以收錄內容，包含筆記型電腦與智慧型手機。再來，為了使之奏效，必須讓認證過程成為在臉書、推特及YouTube這類熱門數位平臺上傳內容的先決條件之一。這兩個條件都不太可能成真。在缺乏法律或管制責任的狀況下，裝置製造商在確定這些數位認證不會提高成本、符合需求，以及不會干擾其產品效能前，渠等不會採用之。再者，僅有少數社群媒體會想屏蔽使用者上傳未授權內容，特別當有人起頭後，卻有可能會將市占率拱手讓給限制較少的競爭者。

第三種具有不確定性的技術作為名為「數位認證不在場證明服務」(authenticated alibi services)，而這種服務也很快就能在私領域出現。由於深度偽造對於政治人物或明星這類名人特別具殺傷力，因為渠等需要保護既重要又脆弱的名聲。為了使自己免受深度偽造影響，這些人也許會選擇進行更加強版的「生活紀錄」(lifelogging，也就是將個人生活的每個細節都記錄下來)，藉此可隨時提供自己身在何方、正在說或做什麼。這些公司可能會開始證明行蹤的套裝服務，包含讓記錄生活更方便的穿戴裝置、可存取大量資料的儲存平臺，以及替這些資料提供可靠認證。這些套裝服務甚至包括與大型新聞或社群媒體平臺相互合作，可以快速確認或澄清影音內容的真實度。

這種記錄方式相當侵犯個人隱私，也會讓許多人敬而遠之。但除了採用生活紀錄以自保的名人外，某些雇主或許也會開始堅持特定類型的受雇者起而效尤，就如同警局要求員警配備穿戴式攝影機。而且，就算只有相對少數人採用這種極端的生活紀錄，依然會產出大量資料，而大家也有可能察覺自己剛好也被網羅其中，從而形成一個巨大的同儕監視網路，持續記錄大家活動。

作者簡介

Robert Chesney現任德克薩斯大學奧斯丁分校貝克(James A. Baker III)學會主席，身兼施特勞斯(Robert Strauss)國際安全與法律中心主任。

Danielle Citron現任美國馬里蘭大學莫頓和馬赫特(Morton and Sophia Macht)榮譽法律教授，亦兼任耶魯資訊社會專案研究員。