科技部工程司
109 年度學門主題式計畫「訊息科技主題式計畫」
規劃說明

壹、 前言

訊息(包括文本、音頻、影像與視訊等資料)在現代社會已經扮演不可或缺的角色，深深影響現代人在生活、交通、娛樂、政治傾向等方面所做的決策，隨著數位媒體工具與通訊網路的發展與普及，加上社群媒體與內容農場的風行，如 Facebook、Line、PTT、Instagram 等，使用者獲取訊息的管道也越來越多元。然而，伴隨著獲得訊息管道的豐富與多元化，許多利益團體也透過運用不實資訊或遭扭曲的事實來操弄訊息，從中獲取利益，隨著人工智慧技術的發展，偽造的訊息也可以呈現非常真實的樣貌，使得訊息的真偽也越來越難以判別，如何協助使用者判斷訊息的真偽以作出正確的決策將成為一個重要的研究與應用方向。

貳、 國內外發展現況

由於虛假訊息(disinformation)的氾濫可能影響人們的判斷而做出錯誤決策，又因其潛藏的巨大商業與政治利益，而有許多利益團體有動機製作與傳播虛假訊息，例如，在著名的「三星寫手門」事件中，三星電子曾利用網路打手進行網路行銷，而以假新聞打擊政治對手亦時有所聞。針對虛假訊息的管制，許多國家已經著手進行立法，希望藉由法律的手段遏止其氾濫，如法國國會於 2018 年通過「反資訊操縱法」及「反虛假訊息法」兩項法律草案，另外，德國亦制定「社群網絡強制法」。除了制定相關法律之外，透過新技術的開發來偵測虛假訊息並辨識訊息來源，經由具體客觀的事實與證據呈現，更能有效遏止虛假訊息的傳播，並避免對於新聞言論自由與人權保障的潛在威脅。此外，法律通常用於嚇止與被動用於責任的追溯，而虛假訊息辨識技術則更能主動釐清事實真相，保障訊息的公平與正確使用。

一則訊息的內容真偽的認定需經多方查證，目前，全球已有超過 150 家事實查核機構，臺灣事實查核中心亦於 2018 年 7 月正式上線，然而，這類機構多須以人工方式求證，逐步釐清訊息中可疑爭議點，以確實還原事實的真相，面對每天大量來自內容農場的消息，這種人工過濾方式精確度雖高，但不易達到迅速澄清的效果，而往往在事件熱度消逝後才能予以釐清，而失去其價值。如何發展技術自動查證判斷訊息真偽的可能性，或是發展輔助系統與技術協助比對、追蹤、蒐集資訊，協助使用者迅速判讀訊息遂成為一個重要的方向。

在多媒體領域，防偽技術已經發展多時，IEEE 也在 2006 年創立了 IEEE Transactions on Information Forensics and Security 期刊，專注於訊息隱藏、

數位浮水印技術、防偽認證等相關技術，在影像竄改偵測上，常使用雜訊、陰影、壓縮格式等資訊來進行。雖然影像竄改偵測已經發展多時，然而隨著人工智慧技術的進步，特別是生成對抗網路(Generative Adversarial Networks)自 2014 年提出以來的飛快發展，對於傳統方法形成許多新的挑戰，也帶來許多新的研究議題，特別是這些人工智慧技術給予使用者簡單易用的影像偽造與生成工具，而提升了技術的普及並加速了此類影像的散佈，例如，著名的 Deepfake 可以將人物 A 的影像與人物 B 的影片結合，將人物 B 置換成人物 A，而製造影片是由人物 A 拍攝的假象，由於其易於取得與使用，被大量用於製造假新聞、惡搞與惡作劇，並在社群網站大肆分享，有可能為公眾人物及無辜民眾帶來無端的困擾，Reddit 與 Twitter 等社群網站以明文禁止發布 Deepfake 生成的影音資料。這些人工智慧技術也為影像竄改偵測領域帶來新的挑戰，傳統方法無法與之抗衡，而以人工智慧偵測此類竄改的技術仍在發展中，相對於生成技術的快速進步，竄改偵測技術的發展較為不足。

為了呈現較佳的視覺效果，已經有許多美照美肌的工具可以協助使用者編輯人臉影像以呈現更好的狀態，美國 Adobe 公司利用人工智慧技術針對人臉影像編輯進行自動偵測，對於特定類別之修改工具有 99%的便是成功率，相較於人類的 53%提升許多，該技術甚至可以嘗試還原修改前的原貌。不過該技術目前仍限定於特定操作與特定類別影像。目前 Deepfake 的技術已能產生相當真實的影像，而偵測 Deepfake 的技術相對遠遠落後，Deepfake 製作的影片在網際網路上散佈，已經產生許多困擾，為了增進偵測 Deepfake 製作的影像與影片的技術，Facebook、Microsoft 與 Partnership on AI 組織中的科技公司於 2019 年 9 月宣布將投入超過一千萬美金與學術界合作，共同舉行 Deepfake Detection Challenge，將請專業演員拍攝影像，以建立訓練偵測 Deepfake 影片之資料集，並舉辦競賽，透過高額獎金及定時更新的分數排行榜鼓勵參與，期能鼓勵相關技術的研發，並能產生好的開源軟體協助 Deepfake 影像的偵測，足見此類技術之方興未艾與受到高度重視。

俄羅斯曾以假造之衛星影像誣賴烏克蘭戰機射擊馬來西亞航空 17 號班機，想藉以逃避責任，因應此一可能之風險，美國國防高等研究計畫署(DARPA)自 2015 年起啟動媒體鑑識計畫 Media Forensics (MediFor)，採取三種主要策略來檢驗媒體真實性，(1)以數位指紋確保影片之真實性，(2)確認影片內容遵循物理定律，(3)以外部資料查證之真實性，如拍攝時的天氣、日照方向等，並由美國國家標準暨技術研究院(NIST)舉行 Media Forensics Challenge 來促進此一領域的發展。如同 Media Forensics 計畫網頁所指出的，目前影像變造的工具與技術仍超過影像變造偵測的工具與技術，該計畫之目的在於發展技術來自動評估影像與視訊之完整性並將其整合至媒體防偽平台中。

在文本訊息方面，目前文本虛假訊息偵測多透過資料探勘方法進行，主要利用

文本內容及社群資訊，利用語義、風格、使用者、立場、傳播途徑等資訊來偵測。文本訊息主要透過社群網路傳遞，Twitter 於 2019 年 8 月在官方推特針對香港事件之輿論操作發表聲明，對許多帳號依多重違反推特使用規定停權，其違規事項包括垃圾訊息、風向操作、假帳號、前科再犯、分身攻擊等。於社群網路中自動或輔助辨認這些不當行為，對於遏止虛假訊息傳播有其效果，目前雖有一些相關研究，但是仍未達到足以實用的階段。

除了一般常見的訊息，如新聞、點評、網路文章、社群媒體訊息、多媒體內容外，學術論文亦可視為一種訊息，也同樣面臨虛假訊息的挑戰，近年來，學術倫理案件層出不窮，學者利用造假資料、同儕審查圈等不當行為來發表論文，對於學術進步產生許多不良影響。除了常見的影像複製與不當竄改之外，近年來也有新開發的人工智慧工具，如 PaperRobot，能透過閱讀已有論文，建構領域知識，自動幫論文寫摘要、關鍵內容和標題，並建議進一步研究方向，未來亦可能發展成自動生成看似合理的論文之寫作工具。這些行為模式其實也和前面所提到的虛假訊息之製作與傳播途徑大同小異，故相關技術亦可能用於協助學術倫理事件的發現與調查。

參、 計畫目標

由於人工智慧技術的發展，虛假訊息的製作與流傳變得更容易，也給傳統虛假訊息偵測方法帶來新的挑戰，而能面對這些挑戰的方法正處於方興未艾之際，許多研究機構與產業才剛開始正視此一問題。在學術上，這類技術的研發具有其開拓性，在實用上，此類技術可能發展出新產業或是有助於社會的正向運作，此外，計畫亦可培育相關研發人才，未來從事此一方向之研究。

肆、 推動議題

計畫旨在推動各式虛假訊息的辨識技術，訊息種類可包含文本、音頻、影像、視訊等，其應用領域則包括新聞、點評、網路文章、社群媒體訊息、學術論文等，為回應人工智慧技術之相關發展，雖不要求所發展之技術必須基於深度學習，但是所發展之技術必須能處理以深度學習技術進行之訊息變造。徵求研究議題包含(但不限於)以下項目。

訊息的變造偵測與來源追蹤。徵求識別技術與工具分辨真實和經過篡改的訊息(可針對文本、音頻、影像與視訊等個別型態資料或其組合)、偵測訊息中經篡改之區域、追蹤經變造訊息之來源、回復篡改之內容等。可能之研究主題舉例如下：訊息變照偵測及定位(判斷給定訊息是否經過變造，若有變造，指出變造位置)、拼接偵測及定位(給定兩則訊息，決定探測訊息是否有部分拼接自來源訊息，若有拼接，指出拼接位

置及來源)、出處搜尋(給定一則訊息，找出給定訊息集合中之最有可能作為給定訊息出處之前幾則訊息)、出處圖建立(依訊息出處建立訊息出處發展圖)等。

發展評估訊息可信度的指標與技術。徵求藉由分析訊息內容、發布途徑等資訊預測訊息可信度的工具，例如，在文字方面，透過偵測聳動的標題、誇張詞彙與數據、挑動讀者情緒字眼等方式判斷訊息之可信度。在影像方面，藉由細微臉部特徵出現與否或是雜訊的空間分佈一致性等，估計影像經過變造之可能性。此外，亦可藉由多訊息來源相互比對找出錯誤或誤導的訊息。

社群網路虛假訊息傳播辨識與處理。了解社群網路中虛假訊息的傳播管道與模式，就社群維度、生命週期、傳播者類別辨識等可能方向分辨虛假訊息與真實訊息之模式差異。發現虛假訊息後如何降低其影響亦是另一個值得研究的重要課題，可能做法如主動移除惡意帳號或虛假訊息，或是主動推播更正訊息給已讀取虛假訊息之使用者以降低虛假訊息對其之影響。

社群網路不當行為辨識。許多假訊息的操縱者會藉由社群網路來營造及散佈假訊息，徵求自動辨識社群網路中不當意圖活動之技術，包含但不限於自動偵測社群網路中的垃圾訊息、風向操作、假帳號、前科再犯及分身攻擊等事件。

透過群眾外包方式協助查詢訊息真偽的技術與平台。全自動化的假訊息偵測有其限制，另一種方式是透過群眾的力量進行訊息的釐清與驗證，徵求發展新技術及平台結合群眾的力量與計算智能進行假訊息偵測，或是發展資訊探查及視覺化技術輔助使用者進行訊息真偽辨別。

建立可追蹤來源的訊息發布技術。區塊鏈技術具有不可篡改性，但亦有速度過慢及安全性不穩定等缺點，徵求開發實際可行之基於區塊鏈的訊息傳遞技術與系統。

訊息數位簽章技術。數位簽章技術可於訊息中嵌入認證之機制，一旦訊息被竄改即會破壞數位簽章，藉以判斷訊息之真偽，雖然數位簽章技術已經發展多年，但是因為人工智慧技術的發展又產生新的挑戰，另外，以往無法達成之目標或許也可能因為人工智慧技術而變成可行。計畫之目的在發展實用之數位簽章技術能抵抗基於人工智慧之竄改技術，並儘可能達成高信度、高強韌性、安全性、低存儲成本、高運算效率等目標。

建立代表性資料集與評估標準。雖然已有一些相關的資料集存在，不過在許多訊息領域仍然缺乏具足夠代表性的資料集與評估標準(benchmark)。在深度學習當道的今日，此類資料集的建立往往是相關技術能否成功的重要因素。

伍、 計畫撰寫說明與審查重點

計畫之目標為開發虛假訊息的實用辨認技術，研究主題必須具有前瞻性、創新性及實用性，計畫書應描述國內外研究及技術現況、所欲達成之技術指標及與世界技術水平同步(或超前)之情形，並應陳述三年計畫規劃及執行內容、階段性成果、可能成效、查核點與評量指標等，將以下項目為審查重點：

- 計畫需要以解決真實世界問題為目標，需要有真實的應用情境，並使用真實資料進行研究與測試。
- 計畫成果須具技術創新性或實際影響，除一般學術成果指標(頂尖會議及期刊論文)外，應提出具體技術指標與可能亮點成果或具體影響，如參與競賽之目標、創建資料集之規模、實際佈建之系統等。

計畫之具體規格與指標得依計畫屬性設定，舉例如下：

- 在特定類別的虛假訊息(如含人臉之影像視訊、人工智慧技術自動生成論文)辨識達到世界第一的水平。
- 設立群眾外包方式協助查詢訊息真偽的平台，並提供相關技術輔助虛假訊息之辨識。
- 創建世界上規模最大的特定虛假訊息資料庫，並舉辦公開競賽。
- 參加類似 Fake News Challenge、NIST Media Forensics Challenge 及 Deepfake Detection Challenge 等之國際競賽，並位居前三名。
- 開源軟體之下載數。
- 服務平台之服務人數。