AI 偽臉時代的身分信任危機

臉部融合攻擊偵測（MAD）成為 AI 資安新戰場

在臉部生物辨識技術此起彼落，AI 偽造技術也同步演化，帶來前所未見的身分信任挑戰，當各類 eKYC、eID、遠距醫療與無接觸邊境自動通關全面倚賴「臉」作為身分，MAD（Morphing Attack Detection）技術站在防止臉部融合攻擊的關鍵防線，已成為全球 AI 資安新顯學。

文／梁日誠

根據加拿大渥太華大學（uOttawa）與 NIST 等研究單位分類，臉部攻擊可分為 **Face Swap（臉部置換）、Morphing Fusion（融合多人體特徵）與 Reenactment（動態重製）**，對應各種深偽與臉部融合攻擊場景。

Fraunhofer 圖形研究所指出，即使兩人並無親屬關係，只要臉部特徵（如眼睛位置）相似，即可產生幾乎無法被人眼察覺的融合影像，進而欺騙自動化臉部辨識系統，使其誤認不同個體為同一人，導致一份護照對應多重身分。

現今的臉部異常或攻擊，主要圍繞在 **Face Swap 與 Reenactment**，以傳統的 PAD 可偵測到，但是對於 Morphing 的偵測，效果卻有限。臉部融合攻擊的高風險案例如邊境自動通關，須在自動通關系統導入 MAD，才能偵測出來。

臉部融合攻擊的高風險場域案例：

- **數位證件申辦：**攻擊者以融合臉申請合法證件，偽冒後續身分行為
- **遠距醫療與保險流程：**冒名啟動診斷與理賠

企業通訊詐騙（如：Zoom 偽冒會議）：偽造主管臉部與聲音，誘騙款項匯出

- **邊境自動通關：**攻擊者以融合臉申請護照，偽冒身分蒙騙自動通關系統

然而，如申請護照只要照片符合護照主管機關規定的尺寸、背景、表情等要求，就可以被接受。我國並未管制照片源頭是否為合法單位，或照片用甚麼方法提交到護照主管機關。也就是說，攻擊者闖關時，若早已進行 Morphing，利用臉部融合照來辦理護照或身分證件，**一旦自動通關系統沒有 MAD，闖關幾乎都會得手**。

融合照片技術並不難，自動通關系統若不具備 MAD 機制，一旦兩個環節都失守，只要有臉部融合照就能闖關，存在高度風險。

若臺灣發現護照使用臉部融合照的消息揭露於國際，或我國無法發現的臉部融合照而被他國揪出，將有損國家形象甚至國際信任。就如同法律規定不能印假鈔，但若金融機構沒有辨識假鈔的機制，假鈔必定盛行。沒有 AI，人性是經不起考驗的；有 AI，人性仍是經不起考驗的。

目前我國管控身分證件的機關，與管制出入境的機關，都已通過 ISO 27001 第三方驗證並適用資安法，為防範風險，建議參考他國經驗在臺推動 MAD。

另一個需注意的難題是，未導入 MAD 的單位僅使用身分證件結合人臉辨識，即使以目前的技術管控身分證件，難以分辨真人臉部照和臉部融合照的差異，攻擊者透過人工關口就可以闖關，因此未導入 MAD 的單位只能依靠自動通關系統防守。而就算抓到臉部融合照，目前也較難通知其他單位進行聯防，考量跨主管機關在管轄上也較不易，還是需要從核發護照的主管機關解決源頭問題。

資安工程可以由預防的角度來協助，AI 工程可以由偵測的角度發現。但若是政策不改變，還是解決不了問題，需要跨主管機關協商，挑戰難以克服，而這就是攻擊者最希望見到的。

不過，不必擔心此議題成為無解。要抓到臉部融合照，核發證件的源頭單位可使用 S-MAD（Single-image MAD），偵測單位可利用 D-MAD（Differential MAD）。以人工通關來說，採用 S-MAD，或是用現場照與護照照片對比的 D-MAD，並輔以人工詢答進行，未來可望解決問題。

另外，在選擇與調適 MAD 機制時，也需留意以下二種誤判與風險：

- **BPCER：Bona Fide Presentation Classification Error Rate（真實樣本錯誤拒絕率）**，為擁有合法、真實護照照片者，其影像被系統誤判為臉部融合（Morph）時所產生的不便比例（亦即系統發生了偽陽性 False Positive），此類誤判的後果，是需額外資源來處理這些實際為真實樣本的照片鑑定
- **APCER：Attack Presentation Classification Error Rate（攻擊樣本錯誤接受率）**，為代表詐騙成功發生的比率，即當一本護照上的臉部融合影像被誤判為真實照片（亦即系統發生了偽陰性 False Negative）時所造成的安全風險

因此，在實務部署 MAD 機制時，如何在降低 APCER 所代表的安全風險與避免 BPCER 所產生的使用者不便之間取得平衡，將成為 MAD 系統評選的重要指標。

國際政策趨勢：從 ENISA 到 ICAO 的技術制度化浪潮

為防範臉部融合技術被用於偽冒身分，德國政府於 2020 年通過法案，明確禁止將兩人臉部影像數位融合後用於護照照片。根據路透社報導，該法案要求護照照片須於政府機關現場拍攝，或由攝影師透過安全連線直接上傳數位檔案，以防止照片在提交前遭到操控。

德國聯邦資訊安全局-BSI 的 TR-0347 技術規範要求在高信任驗證流程中納入類 MAD 偵測能力。

法國 ANSSI 則透過 PVID 驗證要求服務商須通過第三方類 MAD 相關測試。

歐盟 eIDAS 2.0 舉出高保證等級身份須納入類 MAD 等防護機制（如：要求對於臉部影像的真實性與來源可信度進行驗證、防範由申請人提供的數位圖像可能遭到操控或偽造的風險）。

歐盟 ENISA 在「Remote ID Proofing Good Practices」中將 Morphing 與 Deepfake 並列為最高威脅等級，建議導入 PAD／IAD 雙層架構。

國際民航組織（ICAO）ICAO 9303 最新版本已採用 **ISO 39794-5** 作為人臉影像標準格式，以強化對臉部融合攻擊（Morphing Attacks）的防護能力。ICAO 並與 NIST 及歐盟 iMARS 計畫合作，推動包括 **FRVT MORPH** 與 **FATE MORPH** 在內的跨國變臉攻擊測試平台，以評估各類臉部辨識系統的抗攻擊能力，作為後續證件標準與實務政策制定的重要依據，強化全球護照與邊境系統對融合攻擊的免疫力。

國際標準制度化：ISO 20059 與 NIST 驗證趨勢

國際標準方面，目前處於 **FDIS** 階段，標準針對可能遭受臉部融合（Morphing）攻擊的生物特徵辨識系統，建立了相應的要求，標準內容包括：

- **生物樣本修改與操作的分類法：**聚焦於構成多重身分攻擊的操弄方式，如：使用臉部影像融合技術進行

的註冊攻擊（enrolment attack）

•測試資料庫的要求：資料庫包含真實（bona fide，即原始未修改）與變臉後的影像樣本

•評估臉部融合攻擊潛力的方法論：利用融合影像資料集來衡量某種臉部融合技術的攻擊能力。使用者可根據此方法模擬真實應用場景（如證件簽發、邊境查驗），搭配可變次數的嘗試（如：於閘口採集多張探測影像）與多套不同的生物特徵辨識系統（模擬多家廠商的自動通關閘門 ABC gates），以判斷針對該系統的攻擊潛力

•標準亦提供說明性內容，說明如何使用臉部融合演算法進行系統測試與評估。

NIST FRVT MORPH 為全球公開黑箱測試計畫，並與歐盟 iMARS／SOTAMD 平台技術接軌，形成國際間可量測、可比較、可信任的驗證機制。NIST FRVT MORPH 測試的演算法列表（截至 2025 年 6 月）如下面之列表。

放眼世界，站穩腳步

經查訪幾個機構，臺灣目前有三家廠家參與 NIST 的 PAD tests，取得不錯的成績，但對於 NIST 的 MAD Test 仍然缺席。顯示針對 Morphing Attacks 的MAD 的系統，臺灣尚未成熟發展。臺灣廠商目前主力仍在 PAD 應用或以為 PAD 涵蓋了 MAD，或系統通過PAD 測試符合國際標準。PAD ISO 30107 系列的確屬於國際規範，但 MAD ISO 20059 還沒公告，尚未形成可做為測試依據的國際標準。

觀察國際間已將 MAD 技術納入可信任架構的一環，或許可以循以下數點思考可行的方向，期以建立國際間的共防與互信機制：

- 是否需建立國家級 MAD 測試平台，如：NIST或研究計畫，如：SOTAMD（State-of-the-Art in Morphing Detection，歐洲多國合推之臉部融合攻擊測試框架）
- eID、遠距診療、自動通關、金融應用流程中，是否已納入 MAD 機制
- AI 模組驗證標準，是否該將 MAD 納入 AI 資安驗證範疇
- 考量人臉辨識等生物特徵識別資料的高度敏感性，在地資安廠商與服務具備特殊的競爭力
- 思考建立身分證件時的相片的真偽與來源鑑別與申請辦理身分證件時的環境的要求
- 鼓勵在地 MAD 資安廠家參與國際性 MAD 測試，建立國際共信關係，或可進一步建立 AI 資安進軍國際的契機
- 協同不同身分證件主管機關與應用機構，以建議 MAD 機制所需的整合治理架構
- 考量完整納入預防導向的資安工程與偵測導向的 AI 工程的相關控制。

AI演算法代號	國別	測試類型	備註
idemia-004	法國	S-MAD	全場景表現最佳，MACER=0
idemia-003	法國	S-MAD	MACER / BSCER 表現穩定
secunet-003	德國	D-MAD	偵測精確與穩定性高
secunet-002	德國	D-MAD	ArcFace 架構優化版本
visionbox-000	葡萄牙	S-MAD	簽證應用表現優異
unibo-000	義大利	S-MAD	UNIBO morph 資料集中最佳
visteam-001	葡萄牙	S-MAD	假陽性率低
hhi-002	德國	S-MAD	韌性佳、抗高品質融合圖
wvudiff-001	美國	D-MAD	偵測表現一致
kinit-001	斯洛伐克	D-MAD	新進單位，表現具潛力

資料來源：NIST FRVT MORPH 測試平台，截至2025年6月查閱，詳見 NIST 官網。

隨著身分資訊成為攻擊矢量，僅憑生物特徵與證件影像相符已難以防範深度偽造或臉部融合攻擊；真正關鍵在於如何有效驗證具相貌者與其法定身分的唯一性連結，以確認其即為本人。

MAD 機制為 AI 系統的應用，也可成為負責任的與可信任的 **AI 管理與治理制度（如：ISO 42001）**的有效助力，正值 MAD 國際標準即將公告之際，不失為及時與國際接軌的適當時機。