AI時代的「SFATEP」六大核心治理原則

◎ 林宜隆／大同大學資工系教授

人工智慧（AI）技術的全面普及，已深刻改變了產業結構、工作方式、社會價值、隱私安全與倫理規範。在推動AI應用時，企業組織和政府機關普遍存在一種「既擁抱又抗拒」的矛盾心理，麻省理工學院史隆管理學院的研究指出，人們對AI的接受度，除了其能力本身之外，更關鍵地取決於任務是否需要「個人價值」。如果忽略了個人價值，AI所帶來的標準化效益最終可能會壓縮組織內部的創造力與多樣性。

為確保AI系統能以負責任的方式發展，現代AI治理框架的核心理念已明確聚焦於「SFATEP」等六大原則，包括：安全（Security）、公平（Fairness）、問責（Accountability）、透明（Transparency）、倫理（Ethics）與隱私（Privacy）。全球各國與國際組織正積極構築全面的AI治理

體系，以平衡創新與風險，例如歐盟《AI法案》、美國NIST《人工智慧風險管理框架1.0》，以及國際標準ISO/IEC 42001:2023與23894:2023。我國數位發展部也正積極規劃建立AI評測體系，以應對其帶來的挑戰與機遇。

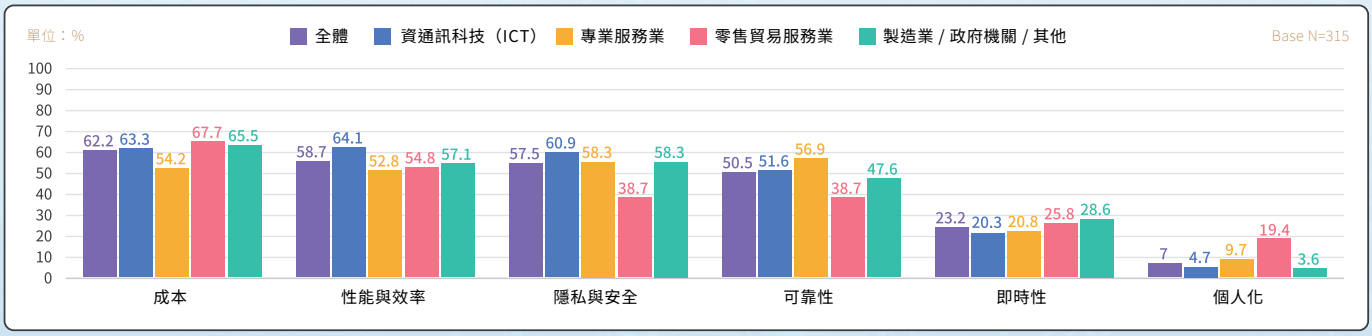
據2025年臺灣產業AI化大調查，臺灣企業在評估AI運算資源時，仍高度聚焦於成本（62.2%）與效能效率（58.7%），隱私與安全（57.5%）則位居第三，顯示企業在AI部署上的務實考量。然隨著AI技術持續進步，組織若能善用個人化AI，將能夠使成員的AI輔助決策更符合其個人經驗與專業背景，進而顯著提升組織整體的知識管理與決策品質。因此，在未來部署AI方案時，應重新評估個人化AI在內部應用的潛力，避免AI產出過度標準化，並充分發揮AI在提升成員個人價值與決策優化方面的潛能。SFATEP六大治理核心原則的落實，將是確保AI不僅能提升效率，更能兼顧「個人化」需求，並激發個體獨特價值的關鍵。

AI「SFATEP」六大治理核心原則簡介及其應用

一、安全（Security）：AI系統穩健運作的基石

旨在確保AI系統在設計、開發、部署及使用過程中，能夠有效防範惡意攻擊、未經授權的存取，並保護其完整性與可用性。尤其在高風險應用場景中，AI系統的安全性是其可靠性和負責任運作的基石。最先進的AI系統可能帶來嚴峻的網路攻擊風險，甚至有能力開發化學、生物、放射、核或爆炸性（CBRNE）武器，對國家安全構成新的威脅。AI系統本身也易受資料投毒（Data Poisoning）或對抗性攻擊（Adversarial Attacks）影響其性能和可靠性。NIST AI RMF 1.0（即ISO 23894:2023）即明確將「安全性」和「資安與韌性」列為評估AI系統信賴度的七大特徵之一，我國數位發展部規劃中的AI評測體系，也將「安全性（Safety）」和「系統安全（Secure）」納入重要評測項目。

產業評估AI運算資源主要考量因素



2025產業AI化大調查中顯示，企業評估AI化時，仍以務實為主要考量因素。資料來源:《2025產業AI化大調查暨AI落地指引》，財團法人人工智慧科技基金會，https://aif.tw/event/ai-research/

二、公平（Fairness）：消弭數位落差，實現機會均等

旨在解決AI系統可能因訓練資料偏差而產生歧視性決策的問題，這些偏差可能基於性別、種族或地區差異，尤其公部門更應重視公平原則，若未妥善處理，AI決策將可能加劇現有的社會不平等現象。

在金融領域的信貸評估中，銀行必須確保AI系統不會因為申請人的性別、種族、居住地區或社經地位等非關信用風險的因素而產生歧視。我國AI評測體系亦將「公平性（Fair）」列為重要評測標準，真正的個人化應建立在公平的基礎上，而非加劇既有偏見。

三、問責（Accountability）：釐清責任歸屬，避免灰色地帶

隨著AI系統在司法、金融、醫療等高風險領域應用日益普及，其決策錯誤的責任歸屬問題，已成為全球關注焦點，問責制指的是AI決策的錯誤或損害能夠被追溯，並有明確的實體（個人或組織）為其負責。

實務上，問責機制仍存有不完善的情形，例如自動駕駛汽車等AI產品致生損害時的責任歸屬釐清，即是尚待解決的議題。未來也可能出現專門處理AI因誤判或產生「幻覺」內容所導致損失的法律機制。我國AI評測體系也將「當責性（Accountable）」列為評測項目；建立第三方監督機構，負責稽核AI偵查工具的合規性與精準度，能夠提供外部的問責壓力。建立健全的問責機制，也是促進使用者對AI信任，進而接受其「個人化」應用的基石。

四、透明（Transparency）：揭開黑箱，重建信任

透明度是AI倫理中最具挑戰性的項目之一，許多大型語言模型（LLM）內部運作原理與資料來源往往不透明，這個「透明」指的是AI系統的決策過程和運作方式是可以理解的，並能向使用者或監管機構公開。

若缺乏透明度，使用者即無法確認AI答案的可信度或偏見來源。歐盟AI法案明確規定高風險AI系統必須符合可解釋性與資安標準。未來的趨勢將是開源AI（Open-source AI）與模型解釋技術（Model Explanation Technology, MET）成為主流，

要求企業組織及司法部門必須更公開其AI系統的設計與運作方式。我國也將「可解釋性（Explainable）」和「透明性（Transparency）」列為AI評測項目。對於「個人化」AI而言，透明度意味著使用者應能了解AI如何根據其個人數據或偏好提供客製化內容，這有助於建立信任。

五、倫理（Ethics）：堅守人性底線，引導技術向善

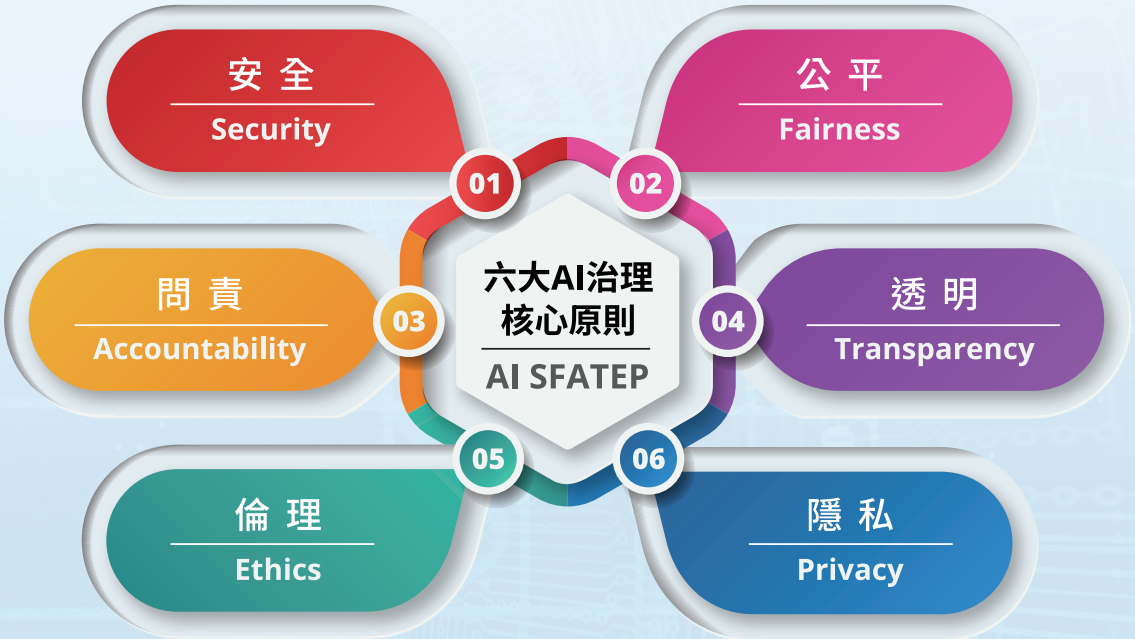
倫理原則旨在確保AI系統的開發、部署和使用符合人類社會的道德、法律和社會期望，尤其司法部門更應重視倫理原則。倫理原則不斷提醒我們，再強大的技術，若失去了人性的底線與道德約束，終將走向毀滅。

當AI生成內容（如「Deepfake」深偽技術）佔據網路資訊主流，人們將面臨「真假難辨」的資訊環境，威脅民主政治與社會信任機制。人們對AI的情感投射與依賴也引發心理學與倫理學上的新議題。此外，大型AI模型的訓練與運作需要消耗巨大的電力，可能成為環境負擔。此外，制定AI輔助明確規範，避免違反隱私權或濫用監控，這本身就是高度的倫理要求，AI工具的設計與應用，可參考IEEE 7000（倫理設計），以符合全球公認的倫理準則。在「個人化」的追求中，倫理原則尤為重要。AI輔助決策應尊重個體的自由意志和尊嚴，而不是成為隱性操控或加劇不平等的工具。

六、隱私（Privacy）：保護個人權益，確保資料合規使用

隱私原則旨在確保AI系統在收集、處理、使用和儲存個人資料時，能夠充分保護個體的隱私權益。這包括確保資料收集的合法性、使用的透明性、以及防止未經授權的存取和洩露。

AI系統的大規模數據處理能力，若無



妥善的隱私保護機制，極易導致個人資料的大量洩露或濫用。我國《資通安全管理法》也強調，AI系統運作涉及大量敏感個人資訊，若被不當使用或洩露將嚴重侵害隱私權。各國隱私保護法規如歐盟的《一般資料保護規範GDPR）》等，對AI系統的數據使用和隱私保護有明確要求。「個人化」AI的本質即是高度依賴個人數據，因此隱私原則是確保「個人化」能健康發展的必要前提。

「SFATEP」將引領 AI 邁向負責任的文明新紀元

總結來說，SFATEP六大AI治理核心原則強調，AI已不再是單純的技術工具，而是一套全面改變人類社會運作邏輯的「新制度和新文明」。因此，人才培育、產業創新、社會制度與倫理規範必須同步調整，才能充分發揮AI的正面效益並抑制潛在風險，更進一步造福人類及創造互動機制和互信價值。

未來，企業或政府機關的競爭力，將取決於是否能有效整合AI技術與其治理原則。這意味著管理者不應只聚焦於AI系統是否夠強大，更應關注其所處的AI治理框架與其對社會的影響。例如，零售與貿易服務業已有近兩成企業在部署AI時考慮「個人化」因素，這顯示在以顧客體驗為核心的場景中，客製化遠比標準化更能創造價值。

導入AI不應只是追求效率，更應思考如何讓AI成為員工的「專屬賦能者（Exclusive Enabler）」。這可透過將AI應用於個人化需求較低的環節；或開發能理解個人行為脈絡與知識背景的AI工具，以支持專業判斷與個人決策。對使用者來說，下一次與AI互動時，或許可以先問問自己：在這個任務中，我追求的是無可挑剔的「能力」，還是一個帶有「個人色彩」的答案？

AI時代的未來，最終不是取決於技術本身，而是取決於我們選擇怎麼使用及與AI共存。這是一場文明的考驗，也是一個世代的機會。要確保AI的負責任發展與兼顧個人價值，即可從理解與落實SFATEP六大AI治理核心原則開始。這不僅是政府與企業的責任，更是公民的共同選擇，以確保在AI浪潮中能夠積極塑造與守護一個更美好、更負責任，且能與人共鳴、充分實現個人價值的AI美好世界。

應用AI機制為核心 進行人頭帳戶識別的數位治理

◎ 蕭國振／新北市政府警察局板橋分局偵查隊小隊長、
臺灣警察專科學校兼任講師

數位金融下的新型態風險與挑戰

隨金融科技迅速發展，數位帳戶因其操作便捷、成本低廉以及普惠金融特性，於國內金融體系內迅速普及。然此類創新金融服務亦成為詐騙集團濫用之新興通路，尤以人頭帳戶之濫用情形最為顯著，並已演變為不法組織進行資金轉移、資金拆分與洗錢活動之主要工具。詐騙集團常透過偽冒貸款、虛構求職等社交工程技術，誘使一般民眾提供身分證明資料，進而在多家金融機構完成無需臨櫃驗證之數位帳戶開立流程後，短時間內即成為非法資金流通之載體，並於完成詐騙活動後遭棄置或遭金融機構凍結，使得傳統偵查手段，於實務上難以有效追查與定罪。

根據金融監督管理委員會公布資料，透過自然人憑證進行身分驗證所開立之數

位帳戶，遭列入警示名單之比例明顯高於採用其他驗證機制的帳戶。部分金融機構的風險管理報告也指出，在列為高風險之帳戶樣本中，以自然人憑證開戶者之比例高達20%。此趨勢已促使多家銀行基於風險考量，陸續暫停接受以自然人憑證進行數位開戶之業務，主管機關亦隨之要求各金融機構全面檢視現行開戶流程，並強化風險控管機制之設計與執行。

此外，聯合國毒品與犯罪問題辦公室（United Nations Office on Drugs and Crime, UNODC）所發布報告中，雖未明確將臺灣列為詐騙活動主要發源地，但經過渲染或刻意扭曲解讀後，「臺灣為詐騙中心」之錯誤敘事迅速於社會輿論中擴散，不僅削弱社會大眾對自然人憑證制度及其信賴基礎之認同，亦反映出我國在數位金融治理面向仍有如缺乏制度性澄清與風險溝通機制、政策說明工具不完備等結構性的挑戰。

AI 防詐解決方案的提出與實踐：2025 內政部黑客松

隨著詐騙手法不斷推陳出新且快速演化，我國防詐策略亦逐步轉型，由傳統以「事後查緝」為主之被動式應對機制，朝向強調「源頭預警」與「風險預防」之前瞻性治理模式發展，而內政部警政署刑事警察局與高雄市政府警察局刑警大隊於「2025內政部黑客松AI應用競賽」共同提出之「AI驅動的詐騙集團人頭帳戶識別及警示解決方案」，即屬政府導入科技治理思維、強化人工智慧於金融詐騙防制中制度性應用之具體實踐例證。

該方案之核心理念是將人工智慧

（Artificial Intelligence, AI）技術嵌入金融機構「了解你的客戶」（Know Your Customer, KYC）流程中，期望在數位帳戶開立初期，即辨識潛在高風險帳戶，藉由預警與阻斷詐騙資金流向，實現事前防範之政策目標。其建構之AI模型是以實際破獲之人頭帳戶案例為訓練資料，結合戶政系統資料與警政情資，採用邏輯回歸演算法建構模組，以兼顧運算效率與模型可解釋性，並強調對帳戶開設環境與使用行為之監測。根據初步實證評估結果顯示，該AI模型召回率達85%，精確率亦達82%，能涵蓋大部分潛在之高風險帳戶，表現出良好的偵測效能，並在有效控管誤判風險的前提下，仍具高度辨識準確性。此外，預警系統可嵌入金融機構之數位開戶流程中，提供即時風險評分與預警功能，不僅有助於提升反洗錢作業效率，亦可支援跨銀行機構間對於可疑分散式開戶行為進行協同偵測與聯防治理。

AI 模型的挑戰與侷限：高擬態詐騙策略的興起

儘管人工智慧（AI）技術已廣泛應用於金融防詐領域，並在實務上展現初步成效，其辨識能力亦逐漸獲得肯定，然詐騙集團亦同步精進其手法，發展出更具隱蔽性之「高擬態式開戶策略」，對現行AI模

型形成實質性挑戰。此類策略主要特徵包括：一、由人頭帳戶本人親自於正規時間與地點完成開戶操作，以排除裝置特徵與IP位址等技術異常訊號；二、提供完整且真實之個人身分資料，有效規避身分偽冒偵測機制；三，寄件地址選擇使用如公司行號等具合理性的地址，難以被偵測為高風險指標；四、刻意避免使用VPN或境外IP等異常網路環境，使整體開戶行為模式近似一般用戶。此類手法已大幅降低AI模型於開戶即時辨識階段的判別敏感度。

造成AI模型辨識侷限之原因，主要可歸納為三點：首先，現行AI系統多基於「開戶當下」的靜態資料進行即時評估，缺乏對後續金流異常或交易行為的即時追蹤能力，導致風險偵測之時效性與完整性不足。其次，若該帳戶為人頭首次使用，其資料未曾列入任何警示清單或歷史黑名單，AI模型將無從辨識其潛在風險特徵。最後，單一帳戶的開戶行為往往難以獨立構成高風險事件，須透過跨帳戶、跨時間軸之圖譜關聯分析（graph-based behavioral analysis）方能識別其組織性犯罪特徵。上述因素顯示，若AI模型僅依賴孤立之單點

特徵進行辨識，將難以有效因應當前詐騙行為高度擬態化與分散化之趨勢。

此外，該AI防詐模型本身亦面臨實證基礎不足之問題，因其訓練資料涵蓋之真實詐騙案例僅約200筆，樣本數量明顯偏少，可能導致模型產生過擬合（overfitting）與樣本偏差（sample bias），進而限制其於多樣化詐騙情境下之泛化能力（generalizability），若未輔以持續性資料更新機制與獨立第三方之效能驗證，其實務應用成效將難以確保。更需關注者，在人工智慧日益受到政策部門與社會大眾青睞之背景下，若對其技術潛力寄予過度期待，而忽略其根本性限制與潛在誤判風險，將可能產生「治理幻覺」（governance illusion），使合法用戶因誤判而遭受不當處置，不僅侵害個人權益，更可能削弱社會大眾對政府數位治理能力之信任基礎，並危及相關政策措施之正當性與長期可持續性。

建構多層次防詐體系：技術深化與制度接軌的雙軌策略

面對日益複雜的高擬態詐騙手法及具延遲特性的異常交易行為，我國防詐體系未來應由傳統「單點式辨識」機制，轉型為結合「動態監控」與「跨域圖譜分析」的多層次風險治理架構。此一轉型不僅需仰賴人工智慧與時序資料分析等技術的進步，也必須同步調整監理制度，推動技術與政策的雙軌整合。唯有提升模型的彈性與風險感知能力，才能有效因應詐騙手法的快速演變與高隱蔽性。

在技術面，應優先強化現有模型架構，導入延時監控與圖譜分析能力。首先，可使用長短期記憶網路（Long Short-Term Memory, LSTM）等時序模型，針對開戶後24至72小時內的交易行為進行監測，辨識如大額跳轉、夜間高頻操作、短期棄用等風險行為。其次，透過圖神經網

路（Graph Neural Network, GNN）技術整合帳戶、IP、裝置、地址與電話等資料，分析跨帳戶關聯，辨識潛在詐騙網絡。此外，前端亦可引入更精細的使用者輪廓建模，包括裝置指紋、瀏覽環境、地理風險，以及滑鼠軌跡與鍵盤輸入等微行為特徵，以強化對可疑開戶行為的即時識別。

在制度層面，亦需同步調整以支撐技術應用的落地與合法性。首先宜由金管會主導建立跨金融機構的高風險情資共享平台，實現異常資料的即時交換，打破資料孤島以強化聯防。其次推動跨部門資料整合，將戶政、警政、電信與金融資料建構為多維風險圖譜，並運用聯邦學習[1]（Federated Learning, FL）與安全多方計算[2]（Secure Multi-Party Computation, MPC）等方法，確保數據可用但不可見，兼顧隱私與效能。最後，應修法賦予AI模型在KYC與AML流程中輔助決策之正式地位，並強調模型可解釋性，以利人工審查與風險管理，建立透明可信的「人機協作」治理架構。

結論

數位金融風險治理之核心關鍵，不僅在於先進技術之導入與應用，更需要奠基在社會大眾對制度體系之信任基礎。在詐騙手法日益進化與輿論環境趨於複雜的背景下，我國亟需以「科技治理化」與「法規數位化」作為雙核心驅動力，推動整體防詐體系之制度性轉型，不僅涉及金融監理機制的技術升級，更關乎法治架構與公民參與間的協同運作。

在提升社會信任方面，可從三大策略面向著手：第一，應推動防詐教育之制度化與常態化，透過多元管道全面提升國民之數位素養與識詐能力。第二，應健全法律救濟機制，強化受害者在遭遇詐騙事件後之法律協助、資金追討與身分修復等實質保障，以鞏固司法正義與社會公平。第三，應強化數位治理透明度，建立制度化的申訴處理與信任修復機制，使民眾在制度運作過程中感受到可預期性與回應性，進而提升對數位治理體系的整體信賴程度，唯有在法制明確、技術精進與社會信任三者共構之條件下，始能建構具前瞻性、韌性與自我演化能力之智慧型數位金融防詐治理架構。

1 是一種分散式機器學習方法，允許多個設備或機構在不共享原始資料的前提下，共同訓練一個模型的方法。

2 是一種密碼學技術，可讓多個參與方在不公開各自輸入數據的基礎上，偕同計算一個約定的函數，並只獲取各自的計算結果。