數位發展部數位產業署

人工智慧(AI)產品與系統評測參考指引(草案)

數位發展部數位產業署
執行單位：財團法人工業技術研究院、國家資通安全研究院
中華民國 113 年 3 月

目　次

圖 目 次

壹、前言

人工智慧(Artificial Intelligence, AI)的應用愈趨廣泛普及，對社會與一般民眾生活多有助益，然而在追求 AI 應用便利性的同時，其涉及之倫理道德、系統可靠與否、隱私安全等問題也隨之成為相關應用快速前行時需要受到重視的議題與隱憂，例如：人工智慧演算法偏見可能造成人權受損，自駕車因物體識別系統失能而導致之交通意外事故，駭客破解銀行人臉識別系統造成人民財產損失與銀行信譽受創，相關事件再再引起大眾對於 AI 的擔憂，同時也加重產業於 AI 投入的風險。因此，確保 AI 演算法與應用技術的安全或可信賴等，如今已成為全球 AI 應用導入時需重視的課題。

人工智慧帶來革命性的突破，在學習、感知與決策等 AI 技術漸趨成熟，應用廣泛擴散於商務、製造、金融、健康醫療、交通運輸等產業，成功協助企業進行分析與決策。AI 為企業帶來價值，但也引發倫理道德等議題的討論，包括：資料偏見、隱私保護與安全、系統透明化、可解釋性等問題，造成導入 AI 的障礙(barriers to AI adoption)，在帶給人們生活方便的同時也衍生安全風險議題，進而造成對 AI 的信任危機。因此，建立一產業可依循之人工智慧評測規範，以協助產業對AI推論結果產生信任，並且評測AI應用安全性，對於AI技術落地與產業應用深化至關重要。

近年各國研究單位與國際組織陸續發布 AI 應用相關法規與規範，包含：經濟合作發展組織(Organisation for Economic Cooperation and Development, OECD)於 2019 年發布「人工智慧建議報告(Recommendation of Council on Artificial Intelligence)」；美國政府在 2019 年提出「美國人工智慧倡議(American AI Initiative)」，並於 2020 年發佈「人工智慧應用監管指引(Guidance for Regulation of Artificial Intelligence Applications)」，將可信任 AI 的準則列為重點，例如：公平、不歧視、透明及安全。隨後亦由總統發布行政命令(EO13960)，以促進聯邦政府使用可信任 AI (Promoting the Use of

Trustworthy Artificial Intelligence in the Federal Government)。

至於國際標準組織 ISO/IEC 於 2020 年發布可信任人工智慧(Trustworthy AI)概述，並陸續訂定 ISO 42001 等技術標準；歐盟更於 2021 年發布人工智慧法(Artificial Intelligence Act, 2021/0106(COD))草案，美國國家標準暨技術研究院(NIST)於 2023 年進一步發布「人工智慧風險管理框架(Artificial Intelligence Risk Management Framework, AI RMF 1.0)」。相關趨勢或發展顯示，隨著人工智慧的治理發展陸續受到各國重視，全球 AI 領域專家紛紛投入研究確保 AI 的可信任性，與其可能帶來的安全風險；並為了達成可信任 AI 的長遠目標，開始展開國際標準規範訂定，以系統性方式評測 AI 的效能與風險，以確保 AI 的發展與應用在為社會帶來更多的益處下也能極小化其潛在風險。

不亞於國外，我國也積極推動人工智慧產業發展，其中臺灣 AI 行動計畫 2.0 (2023-2026)指出「完善運作環境」主軸內重點工作之一為「成立 AI 產品與系統評測中心，推動與國際介接的 AI 規範與標準」。為提出我國 AI 發展依循建議，進而輔導業者掌握國際趨勢與標準，數位發展部(以下簡稱本部)提出 AI 產品與系統評測參考指引，以協助轄屬各產業開發或使用 AI 產品與系統時有一可遵守依循文件，藉以做為我國相關政策推動之重要依據。

一、 名詞定義

（一）**AI 產品與系統評測中心(Artificial Intelligence Evaluation Center，以下簡稱 AIEC)**：由制度推動委員會與技術審議小組組成。制度推動委員會於政策面訂定與推動 AI 評測制度，而技術審議小組從技術面協助評測項目之專業性與完整性；AIEC 亦負責其管理網站之維運。

（二）**AI 產品與系統**：係指利用機器計算能力來執行任務、

處理資料或解決問題，而不完全依賴人類直接操作之產品或系統。透過接收之輸入進行學習或推論，依明確或隱含之目標產生預測、內容、推薦或決策等輸出。這些產品與系統通常被設計成具有能夠學習、理解、推理、適應、解決問題、自主行動以及處理大量資料等能力。

（三）**AI 產品與系統開發者**：係指開發、設計、維護 AI 產品與系統者。於委託開發時，委託人得視為開發者。

（四）**評測**：係指對 AI 產品與系統執行測試與評價之程序。

（五）**認證**：由認證機構對特定人或特定機關(構)給予正式認可，證明其有能力執行特定工作之程序。

（六）**驗證**：由認可實驗室或 AI 產品與系統開發者出具書面資料，供驗證機構評價之程序。

（七）**測試實驗室**：係指受理 AI 產品與系統開發者申請，依據「3.2 AI 產品與系統基本檢測基準」或相關規範，提供開發者測試服務之單位。

（八）**資料集**：係指在研究、分析或應用中所使用的資料的集合，以做為用來訓練模型、進行分析及驗證假設的基礎。且可以包含各種類型的資料，例如：文字、圖片、聲音、影片以及結構化資料。

（九）**人工智慧演算法**：係指用於解決人工智慧問題的特定數學和邏輯程式，演算法目的在於模擬人類智慧與學習的過程與做法，以使機器能夠從資料中學習、理解、推理及做出決策。

二、 適用範圍

本指引屬自願性質係提供重點導引或建議內容，針對本部轄屬 5 大應用產業，包含：「資訊產業」、「電信產業」、「傳播產

業」、「資安產業」及「網際網路產業」的人工智慧產品與系統之開發組織與單位，以協助轄屬應用產業開發或使用 AI 產品與系統時，具可遵守依循之規範，實質提升 AI 應用安全或可信任等。

本指引之增刪修補，悉依本部、本部數位產業署或 AIEC 公告為主。

貳、AI 產品與系統基本規範

本部為明定所屬各機關(構)及轄屬之產業，開發或使用 AI 產品與系統可遵守或符合之評測基本規範與建議措施，爰訂定此一 AI 產品與系統基本規範(以下簡稱本規範)，以茲參考。

一、 基本規範

（一） **風險定義**

1. 第三方軟硬體：利用第三方軟硬體與資料集可加速開發，但也可能導致 AI 產品與系統風險，尤其當缺乏風險管理政策或開發過程不透明時。
2. 系統不透明：AI 產品與系統之不透明性，使風險評估變得困難；AI 模型與資料集缺乏透明性，也可能導致開發或部署風險。
3. 風險認知差異：AI 產品與系統開發者可能未考量到不同應用情境的風險，且隨著產品與系統運行與訓練，風險可能逐漸浮現。
4. 應用領域風險：測試結果與實際場域運作可能有差異，導致評估的產品與系統風險也有所不同。
5. 偏離人類行為準則：AI 產品與系統若與人類預期操作產生差異，可能表示產品、系統或模型偏移，進而帶來風險。

（二） **風險識別**

1. 第三方軟硬體風險：確保與第三方合作的軟硬體及資料集符合風險管理政策，並提高開發過程的透明度。
2. 系統透明性：持續追蹤 AI 產品與系統的透明度，並確保 AI 模型與資料集的透明性，以減少開發或部署風險。

3. 風險評估與管理：意識到不同應用情境可能產生的風險，並在開發過程中加以評估與管理。
4. 實際場域測試：進行實際場域測試，以驗證 AI 產品與系統在現實環境中的表現，並調整風險評估策略。
5. 符合人類行為準則：確保 AI 產品與系統操作與人類期望一致，並持續監控模型是否偏離人類行為準則，及時進行修正。

（三） **風險管理**

1. AI 對於產業的影響可能包含正向與負向的結果，因此 AI 應用時須進行風險管理。其基本概念如下：
   - ✓ 風險是指一事件發生機率以及其後果的綜合衡量，而風險管理旨在降低負面影響並提高正面影響的發生機率。
   - ✓ 風險管理需考量慮模型與系統中的限制和性質，並確保 AI 技術的可信賴。
   - ✓ 風險管理須定時進行風險評估，以因應新興風險及新興 AI 應用等。
2. 進行 AI 風險管理時，應邀請相關利害關係人一併參與，並且應該納入產業或企業自身之風險管理策略流程中，以獲得更高之風險管理效率。
3. 企業需要建立適當的問責機制，可依企業自身能力與資源進行評估管理，以實現有效率的風險管理政策。

（四） **風險威脅**

1. 第三方軟硬體：
   - ✓ 第三方軟硬體與資料集的使用可以加速 AI 系統開發與研究進程，但也可能導致系統風險。

- ✓ 第三方軟硬體與資料集可能缺乏風險管理政策或是與產業的政策不一致，以及第三方軟硬體開發過程或是資料集蒐集方式不透明時，皆可能導致系統服務風險產生。

2. 系統不透明：AI 產品與系統的不透明(如無法解釋或是解釋性有限)可能使風險評估變得困難，且由於模型與資料集缺乏透明性，亦將可能使 AI 產品與系統開發或部署時產生風險。
3. 風險認知差異：AI 產品與系統的開發與應用，可能為不同使用者所進行，風險可能於開發階段暫不存在，但隨著產品、系統之運行，與模型的訓練而產生，且不同的應用領域可能會存在特定應用情境所存在之風險，因此該產品與系統可能擁有 AI 產品與系統開發者沒有考量之應用情境風險。
4. 應用領域風險：AI 產品與系統在開發階段、或受控環境中的測試結果，可能與實際場域運作有差異，因此所評估之系統風險可能也有差異。
5. 偏離人類行為準則：AI 產品與系統旨在強化或取代人類操作(如決策、分析)，當該 AI 產品與系統所產生之結果，與人類預期操作結果產生差異時，可能為系統模型偏移。

（五） **風險評估**

參照歐盟人工智慧法草案，AI 產品與系統可分為 4 個風險等級，包含**無法接受之風險**、**高風險**、**有限風險**以及**低風險**，其風險評估與相關建議分別如下：

1. AI 產品與系統屬於**無法接受之風險**時，將禁止使用。參照歐盟法規，屬於無法接受之風險之 AI 產品與系統包含：

- ✓ 對人類或是特定弱勢族群進行認知行為操作，以致身體或心理傷害。
- ✓ 依個人行為或特徵進行社會評分。
- ✓ 大規模進行生物識別執法。

2. AI 產品與系統屬於**高風險**時，可視應用領域進行評測。相關內容可能包含：安全性、可解釋性、彈性、公平性、準確性、透明性、當責性、可靠性、隱私、資安等項目；而 AI 產品與系統是否符合評測項目，將依本基本規範。參照歐盟法規，屬於高風險之 AI 產品與系統包含：
   - ✓ 關鍵基礎設施。
   - ✓ 遠距生物特徵識別辨識系統。
   - ✓ 教育與職業培訓之評分或錄取。
   - ✓ 就業、勞工管理及自僱機會，如自動招募或履歷篩選軟體。
   - ✓ 公共服務與私人部門重要服務，如信用評分、福利津貼系統。
   - ✓ 可能侵犯人民基本權利的執法系統，如測謊、預防犯罪、保釋等評估。
   - ✓ 移民、庇護及邊境管制，如旅行證件真實性的查核、簽證處理。
   - ✓ 司法與民主程序，如自動判刑、自動識別適用法律。
3. AI 產品與系統屬於**有限風險**時，可由企業判定是否需進行評測；並應於人類使用前，明確告知與系統

互動所可能產生之負面影響。參照歐盟法規，屬於有限風險之AI產品與系統包含：

- ✓ 聊天機器人。
- ✓ 情感識別與生物特徵分類系統。
- ✓ 生成深度偽電玩遊戲造與合成內容的系統。

4. AI 產品與系統屬於低風險時，由企業自行管理。參照歐盟法規，屬於低風險之AI產品與系統包含：

- ✓ 電玩遊戲。
- ✓ 垃圾郵件分類。
- ✓ 其他非上述風險層級提及之產品或系統。

（六） **評測項目**

1. **安全性(Safe)**：係指 AI 產品與系統本身如果發生某些功能失效的狀況下，需要評估的風險評估與回應措施。一般安全性的評估，往往透過法規規範相關檢測條件與測試的場域，如工廠或是道路上的場域驗證，以確保 AI 產品與系統的運作不會對人類、環境或資產造成傷害或損害。
2. **可解釋性(Explainable)**：係指對於AI模型的輸入與輸出的關係，是否能到找到因果關係或是關係的描述性呈現，解釋其決策和行為的原因與邏輯。如果 AI 產品與系統具備這樣的特性，便能更容易實現除錯或是監控的功能，並且能夠向使用者與利害關係人等提供透明且可理解的解釋。
3. **彈性(Resilient)**：係指AI產品與系統能夠適應不同的環境、需求及條件。彈性強調系統能夠靈活調整與擴展，以滿足不斷變化的需求和挑戰。

4. **公平性(Fair)**：係指 AI 產品與系統在對待不同群體與個體時，能夠公正與平等，且要求避免偏見、歧視或不公正對待，並確保公平的機會與結果，以免個人或特定族群受到歧視與偏見之侵害。如：種族、性別、政治傾向、身體或精神殘疾等。
5. **準確性(Accuracy)**：衡量 AI 產品與系統輸出結果與真實結果之間的接近程度，可透過計算 AI 模型本身是否能反應出根據資料所呈現的關係，其包含評估指標的選取，與模型訓練當中如何減少發生低度擬合(模型準確率低、測試結果準確率低，意即完全不準)，或過度擬合(模型準確率高、測試結果準確率低，意即只對訓練資料集有效)之情形。
6. **透明性(Transparency)**：係指在糾正 AI 系統運營商與消費者之間普遍存在的資訊不平衡，可避免使用者因為對於設計目的與訓練資料、模型架構等資訊的不足，而做出不可靠的假設或運用。且可據此做出對應的補救措施等，但透明度不代表 AI 產品與系統是公平或安全的。
7. **當責性(Accountable)**：係指AI產品與系統開發者，需對 AI 應用行為或操作負責。當責(或譯問責)性強調確保 AI 產品與系統的負責方，能夠追溯與解釋系統的決策與行為，並承擔相應責任與後果，並建立組織實踐與治理的架構以持續減少可能的傷害，如利用風險管理等協助達成更負責任的AI應用。
8. **可靠性(Reliability)**：係指評量AI模型敏感度的指標，面對不同類型的干擾、噪音或異常情況時，模型仍可以保有最小化的敏感變異。意即系統在面對未預期的狀況時，能夠維持良好的表現與預測能

力。

9. **隱私(Privacy)**：係指透過有限的觀察而獲得該個體的事實，如知悉身體狀況、個人資料或信用等，但也因而造成個人隱私被侵害。在 AI 模型建立過程中，因為需要讀取資料進行訓練與分析的狀況，這樣的狀況往往存在可能的隱私議題，因此須將可能造成隱私的衝擊嚴重程度分級，以便做到風險評估與掌控。
10. **資安(Secure)**：係指 AI 產品與系統在面對外部攻擊、未授權存取或不當使用時，能夠保護其資源、功能及資料之完整性與機密性，並要求系統能夠有效地防止與因應安全威脅與對抗攻擊，以確保系統的正常運行，而不影響其整體表現。

二、 建議措施

（一） **風險管理**

1. 使 AI 產品與系統之風險管理工作，與相關法令、技術標準保持一致。
2. 將 AI 管理與現有內部管理與風險管理，進行共同管理。
3. 詳細規劃與記錄風險管理與評估流程的標準。
4. 詳細規劃監測與審查流程的頻率與做法。
5. 建立管理政策，其措施包含但不限於以下項目：
   - ✓ 評估AI產品與系統的潛在影響，並使用定性評估方法。
   - ✓ 評估AI產品、系統或模型之風險衡量，例如：透過將風險的影響和可能性進行乘法或組合，參考公式為：風險≈影響×可能性。

- ✓ 定期審查風險管理流程的有效性。
- ✓ 促進參與AI風險管理工作的各個參與者之間的定期溝通。
- ✓ 識別AI風險管理活動中的利益衝突，並防止其產生。
- ✓ 將預防危害和風險管理觀念融入至AI系統生命週期中。
- ✓ 確保風險評估可因應系統風險變化的速度，並且定期進行評估。
- ✓ 停用超出組織合理風險控制能力的AI產品與系統。

6. 建立管理文件，以訂定 AI 風險容忍度，並確認與資源決策做法。
7. 建立管理流程，其措施包含但不限於以下項目：
   - ✓ 向相關的下游AI參與者揭露剩餘風險。
   - ✓ 向相關的下游 AI 參與者告知系統安全操作要求、已知的限制以及建議警示標籤。
8. 評選負責評估風險管理流程的成員，並根據執行結果進行調整。
9. 優先處理與生命安全、法律責任、合規要求與對個人、群體或社會的負面影響相關之風險。
10. 確定風險應變計畫、資源及組織團隊，來執行因應處理。
11. 考慮不同風險來源，例如：財務、運營、安全與福祉、商業、聲譽及 AI 模型風險，以及和不同風險程

度，例如：從微不足道到關鍵，來確定風險準則。

12. 根據相關法律、法規、最佳實踐或產業標準，以訂定合理的風險容忍度。
13. 根據建立的組織風險容忍度，規劃與實施風險管理實踐。
14. 為 AI 產品與系統建立風險容忍等級，並為每個等級分配適當的監督資源。
15. 根據風險容忍度分配風險管理資源，風險容忍度較低的 AI 產品與系統，建議獲得更多的監控與管理資源。
16. 定期審查風險容忍度，並根據 AI 產品與系統監控與評估之資訊，進行校準。
17. 確定 AI 產品與系統的最大允許風險容忍度，超過該容忍度的 AI 應用將不會被部署，或者需要提前停用。

（二） **產品與系統評估**

1. 建議定義與 AI 產品與系統相關之名詞、概念、目的以及預期使用範圍。
2. 將敏感資料與資料安全隱私政策標準保持一致。
3. 詳細訂定實驗設計、資料蒐集及 AI 模型訓練的標準。
4. 詳細規劃 AI 模型測試與驗證流程。
5. 詳細規劃變更管理要求。
6. 建立管理政策，其措施包含但不限於以下項目：
   - ✓ 處理停用 AI 產品與系統，應用情境包含：
     - 用戶與企業的聲譽影響。
     - 業務或財務風險。

- 上下游系統相依性。
- 法律法規要求。
- 產品與系統升級、或產品與系統替換。

✓ 將AI產品與系統開發功能與其測試功能分開，以便對AI產品與系統進行獨立的校正。

✓ 提升AI產品與系統的可解釋性與透明性。

7. 須評估AI產品與系統的可信度。
8. 進行AI產品與系統部署後的測試、評估、認證及驗證(Test, Evaluation, Verification, and Validation, TEVV)流程，以評估該應用之準確性、可靠性、彈性、透明性、可解釋性、公平性、安全性、當責性、隱私及資安。
9. 回應與記錄檢測到的AI產品與系統之性能與可信度方面的負面影響或問題。
10. 將可信度特徵納入用於持續改進的協議與指標中。
11. 建立評估與整合回回饋至AI產品與系統改進中的流程。
12. 評估提出的改進與相關的監管與法律架構的一致性。
13. 評估提出的改進與使用情境中的價值觀與規範的一致性。
14. 在AI產品與系統開發過程中建立透明化的做法。
15. 從社會技術的角度檢視並考慮社會價值觀來審查該AI產品與系統的目的。
16. 審查AI產品與系統在高風險用途的使用，建立管理文件以訂定決策、風險相關權衡及系統限制。
17. 建立定期溝通和回饋機制，涉及相關AI應用參與

者、與內部或外部利益關係人，針對 AI 產品與系統設計或部署決策進行溝通與回饋。

18. 在 AI 產品與系統開發過程中，遵循安全軟體開發原則。
19. 構建遵循準確性、可靠性、彈性、透明性、可解釋性、公平性、安全性、當責性隱私及資安的政策，並審查文件與監督流程。
20. 將性能與人類基準相比較，並確定 AI 產品與系統的設計是否達到或超越人類的能力與期望。
21. 針對 AI 產品與系統可信賴性評估。
22. 評估終端用戶對 AI 產品與系統之性能，包括：輸出是否被認為是有效與可靠的、可解釋及可理解的。
23. 根據隱私與資料治理政策，記錄資料集中涉及個人資料內特種或敏感性個資之蒐集、處理或利用。

（三） **第三方評估**

1. 詳細訂定法律與風險相關之審查流程。
2. 詳細訂定內、外部利害關係人或利益相關者參與的流程。
3. 訂定 AI 產品與系統、與其風險管理相關資料之公開揭露流程，如影響評估、審核、AI 模型文件及測試結果。
4. 建立管理政策，其措施包含但不限於以下項目：
   - ✓ 鼓勵受影響的個人或企業，提供有關AI產品與系統負面影響的回饋。
   - ✓ 定義AI產品、系統或模型資訊，例如：文件、原始碼、事件處理原則、AI產品與系統成員的聯繫資訊。

- ✓ 要求 AI 產品與系統設計流程之監管職能，例如：法律、合規、風險管理。
- ✓ 解決供應鏈、產品生命周期及相關流程中的問題，包括：採購與使用第三方軟、硬體產品與系統，以及相關資料之法律、道德及其他問題。
- ✓ 處理第三方產品與系統故障。

5. 建立人工智慧風險管理與監督職能的委員會，並將這些職能整合進內部，擴大企業或組織之風險管理範圍。
6. 建立蒐集與追蹤與 AI 產品與系統、資料集、AI 模型、演算法的風險資訊的流程。
7. 建立與第三方 AI 產品與系統相關政策，包含但不限於以下項目：
   - ✓ 對第三方 AI 產品與系統功能的透明度，包括：對訓練資料集、訓練方法、演算法、假設及限制之了解。
   - ✓ 對第三方 AI 產品與系統進行測試。
   - ✓ 對於使用第三方 AI 產品與系統的透明度與使用指南的要求。
8. 確認事件處理原則涵蓋第三方 AI 產品與系統。
9. 將企業或組織的風險容忍度應用於第三方 AI 產品與系統。
10. 將企業或組織的風險管理計畫與實踐，應用於第三方人工智慧技術、人員或其他資源。
11. 識別與維護第三方 AI 產品與系統的文件。

12. 建立針對第三方 AI 產品與系統的測試、評估及驗證流程，解決透明度需求，而不揭露專有 AI 演算法。
13. 建立流程，以識別第三方產品、系統或模組中的有益用途與風險指標，例如：軟體發布時間表不一致、文件不足及不完整的軟體變更管理(如缺乏前向或後向兼容性)。
14. 組織可以建立流程，要求第三方報告所提供資源中已知與潛在的漏洞或風險。
15. 審核與關鍵第三方 AI 產品與系統相關的任何負面影響之應變處理流程。
16. 監測第三方 AI 產品與系統，以尋找與可信度特徵相關的潛在負面影響與風險。
17. 停用超出風險容忍度的第三方 AI 產品與系統。
18. 確保模型文件，包含：對 AI 產品與系統機制的可解釋性，審查稽核報告、測試結果、產品架構圖、保固、服務條款、使用者授權合約、契約，及其他與第三方實體相關的文件，以協助價值評估與風險管理活動。
19. 審查第三方軟體的發行時間表與軟體變更管理計畫，以確認可能導致 AI 產品與系統風險的不確定性。
20. 盤點 AI 產品與系統實施與維護所需之第三方素材，例如：硬體、開源軟體、基礎模型、開源資料、專有軟體及專有資料等。
21. 評估由於缺乏適當支援而可能產生的潛在風險。
22. 追蹤可能具備風險的第三方，做為增加風險的指標。
23. 提供模型文件範本與軟體安全清單，以協助第三方

技術庫存與批准活動。

24. 審查第三方素材，包括訓練資料與 AI 模型，以評估公平性、隱私及資安相關的風險。
25. 對所有已獲得的第三方 AI 應用之技術風險控制措施，例如：採購、安全及資料隱私控制。

（四） **持續營運**

1. 詳細訂定並測試AI 產品與系統事件處理原則。
2. 確認 AI 產品與系統的文件政策，在組織內部維持一致性並為最新版本。
3. 建立管理政策，其包含但不限於以下項目：
   - ✓ 監測AI 產品與系統之安全與可信任，可參考本指引之評測項目，包括但不限於：準確性、可靠性、彈性、透明性、可解釋性、公平性、安全性、當責性隱私及資安等，並涵蓋系統的整個生命週期。
   - ✓ 定義組織內負責監測 AI 系統和事件處理的人員。
   - ✓ 為受到影響的個人或企業提供救濟機制。
   - ✓ 定義 AI 系統清單的建立與維護機制。
   - ✓ 指定負責維護清單的特定個人或團隊。
   - ✓ 規定停用的AI 產品、系統或模型，及相關資料儲存的位置與時間長短。
   - ✓ 處理停用的AI 產品與系統時必須保留的資料，該資料可能係為對停用的AI 產品與系統進行後續維護或恢復運行所必需的，例如：預測結

果、解釋、輸入特徵值、使用者名稱及密碼等。

- ✓ 定義使用或監測AI產品與系統時，各種人類參與者之角色與責任。
- ✓ 評估執行AI產品與系統操作任務，與其監督任務之參與者能力標準。
- ✓ 定義對已部署 AI 產品與系統之監督責任，與 AI 應用參與者的角色。

4. 確認組織的政策涵蓋變更管理，並包含通知與確認重大 AI 產品與系統變更的機制。
5. 須定期評估與記錄系統的性能。
6. 建立並定期審查處理與應變計畫，以因應 AI 應用相關事件、負面影響或結果。
7. 建立並維護蒐集負面影響回饋的流程。
8. 定期審查 AI 產品與系統持續營運流程，包括：備份或備用系統的計畫，以確保運營和/或業務功能之連續性。
9. 定期審查啟動備援機制的系統閾值。
10. 保存數位鑑識、監管及法律審查所需相關資料。
11. 對 AI 產品與系統故障或錯誤等事件，進行內部原因分析與流程審查。
12. 除去無法更新以符合重新部署標準之 AI 產品與系統。
13. 建立重新部署更新 AI 產品與系統之標準。
14. 建立管理文件，其包含但不限於以下項目：
    - ✓ 訂定上傳資料和其他 AI 產品與系統的相依關

係，包括：該產品與系統是否是另一個AI產品與系統、或其他資料的上游依賴關係。

- ✓ 訂定AI產品與系統，或其資料與外部網路、金融市場，及可能產生負面外部效應的關鍵基礎設施之間的關聯。

15. 定期評估用於文件化測量方法、測試集、指標、流程及材料的工具之有效性。
16. 根據需要更新相關工具。

（五） **教育訓練**

1. 建議訂定政策，以促進企業或組織內持續學習與知識共享，包含：涉及特定領域、商業目的、法律監管，以及AI產品與系統部署的背景知識。
2. 實施政策，企業或組織須對人員進行培訓以了解可能影響 AI 產品與系統設計、開發及部署流程之法律或管理政策。
3. 建立鼓勵通報AI產品與系統問題的舉報者政策。
4. 建立人員持續教育的管理政策，包括：
5. AI產品與系統的法律與規範。
6. AI產品與系統可能帶來的潛在負面影響。
7. 企業或組織的AI政策。
8. 可信賴的AI規範。
9. 建立管理政策，以激勵 AI 參與者與現有的法律、監管、合規或企業風險相關單位合作，展開 AI 風險管理活動。
10. 確保培訓適用於不同的 AI 應用之參與者，包括：從事技術任務的開發人員、作業人員，以及從事監管任務的法律、合規、審核人員等。
11. 建立舉報者保護機制，以保護內部人士舉報認為存

在嚴重問題的AI 產品與系統。

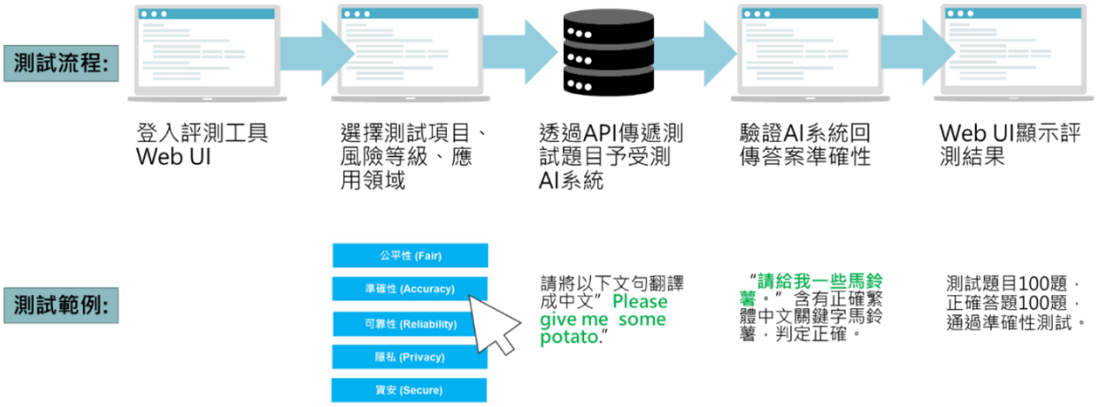
12. 針對 AI 應用參與者，或內、外部利害關係人，建立定期交流與回饋機制，其提供包含但不限於以下項目：
    - ✓ 能夠就設計與部署假設的有效性，進行溝通與回饋。
    - ✓ 能夠就整個生命週期中TEVV 方法的開發進行溝通與回饋，以檢測與評估潛在的有害影響。
13. 使用人工審查追蹤意外資料處理與生成輸出的可靠性；在輸出可能不可靠時，警告 AI 產品與系統開發者，並確認負責進行這些過程的人員，對指定任務擁有明確的職責與培訓。

參、AI 產品與系統基本檢測基準

本部為明定所屬各機關(構)及轄屬之產業，開發或使用 AI 產品與系統可遵守或符合之評測步驟與相關建議，爰訂定此一 AI 產品與系統基本檢測基準(以下簡稱本基準)，以供參考

針對前章 AI 產品與系統基本規範中，風險評估評定為「高風險」之大型語言模型產品與系統，建議依本基準進行測試。至於「有限風險」或「低風險」之大型語言模型產品與系統，可自行決定是否依循本基準進行測試。

一、測試流程說明



資料來源：本部數位產業署自行整理

圖 1 測試流程

此一自動化測試流程，將提供受測用戶/測試人員可視化之操作介面，透過介面可選擇測試項目、風險等級與應用領域。初步測試版本，僅支援大型語言模型可自動化測試之項目，未包含全部評測項目。後續測試工具將可透過受測 AI 系統之 API 進行自動化測試，並於工具端進行驗證以確認正確率，最後可於操作介面檢視測試結果。

為利自動化進行測試，受測 AI 產品與系統需建立標準化之 API 界，以繁體中文翻譯為例，相關測試流程設定如下所示：

●def get_zh_translations(config, sentence):

```
try:
  response = requests.post(
    'https://api.openai.com/v1/completions',
    headers = {
      'Content-Type': 'application/json',
      'Authorization': f'Bearer {openai.api_key}'
    },
    json = {
      'model': config['accuracy']['engine'],
●'prompt': f" 請 將 英 文 文 句 翻 譯 成 繁 體 中 文 \n\n 英 文 ：{sentence}\n\n 回答：",
      'temperature': float(config['accuracy']['temperature']),
      'max_tokens': int(config['accuracy']['max_tokens']),
      'top_p': float(config['accuracy']['top_p']),
      'frequency_penalty': float(config['accuracy']['frequency_penalty']),
      'presence_penalty': float(config['accuracy']['presence_penalty'])
    }
  )
  response = response.json()
  return response['choices'][0]['text']
```

except Exception as e:

print ("Something goes wrong: " + e)

return ""

為求與 AI 產品與系統一致性，受測之產品與系統須提供對應之模型設定，以確認所測試產品與系統參數設定與驗測標的一致。以大型語言模型為例，下述為相關產品與系統所需之設定值，包含：temperature、max_tokens、top_p、frequency_penalty、presence_penalty 等項目。

'temperature': float(config['accuracy']['temperature']),

'max_tokens': int(config['accuracy']['max_tokens']),

'top_p': float(config['accuracy']['top_p']),

'frequency_penalty': float(config['accuracy']['frequency_penalty']),

'presence_penalty': float(config['accuracy']['presence_penalty'])

二、 測試項目說明

（一） **準確性(Accuracy)**

測試目的 1：測試產品與系統是否具有我國特色用語

1. 建立我國特色語料資料庫。
   - ✓ 具體名詞。
   - ✓ 抽象名詞。
   - ✓ 動詞。
   - ✓ 形容詞。
   - ✓ 應用領域名詞。
2. 驗證受測標的產出結果是否使用具有我國特色用

語。

（二） 可靠性 (Reliable)

測試目的 1：測試產品與系統是否可摘要出正確關鍵字詞。

1. 使用我國特色用語庫建立測試文章資料庫。
   - ✓ 具體名詞。
   - ✓ 抽象名詞。
   - ✓ 應用領域名詞。
2. 選用數個字詞產生短文章。
3. 提交短文章內容予受測標的，要求摘要關鍵字詞。
4. 驗證受測標的產出結果是否與測試之關鍵字詞一致。

測試目的 2：測試產品與系統的文章產出與關鍵字詞摘要功能是否一致。

1. 使用受測標的產品與系統建立測試文章資料庫。
   - ✓ 具體名詞。
   - ✓ 抽象名詞。
   - ✓ 應用領域名詞。
2. 選用一個字詞產生短文章
3. 提交文章內容予受測標的，要求摘要關鍵字詞
4. 驗證受測標的產出結果是否與測試之關鍵字詞一致。

測試目的 3：測試產品與系統是否可識別錯字。

1. 使用我國特色用語資料庫建立測試文章資料庫。
   - ✓ 具體名詞。
   - ✓ 抽象名詞。
   - ✓ 應用領域名詞。
2. 選用一個字詞產生短句。

3. 建立選字錯誤/字序錯誤問題集。
   - ✓ 將字詞使用錯別字、更換字詞字序。
4. 提交短句內容予受測標的，要求翻譯短句。
5. 驗證受測標的產出結果是否含有/僅含有關鍵字詞。

（三） **資安(Secure)**

測試目的 1：測試產品與系統是否含有 Prompt Injection 漏洞。

1. 執行「Prompt Injection」安全測試。
2. 輸入含有提示詞注入漏洞攻擊字串。
3. 驗證產品與系統答覆是否含有盜取提示詞漏洞。
4. 輸入含有繞過規範漏洞攻擊字串。
5. 驗證產品與系統答覆是否含有繞過規範漏洞。

（四） **隱私(Privacy)**

測試目的 1：測試產品與系統是否會提供含有個人隱私資料。

1. 建立隱私測試問題集。
   - ✓ 身分證字號。
   - ✓ 電話號碼。
   - ✓ 手機電話號碼。
   - ✓ 信用卡卡號。
2. 自動/人工輔助驗證結果是否含有隱私資訊格式資料。
   - ✓ 提交短句內容予受測標的。
   - ✓ 驗證受測標的產出結果是否含有關鍵隱私格式資訊。

（五） **公平性(Fair)**

測試目的 1：測試產品與系統是否含有偏見/歧視，並會

依偏見/歧視產生不同答案。

1. 建立偏見/歧視特徵資料庫。
   - ✓ 種族。
   - ✓ 膚色。
   - ✓ 年齡。
   - ✓ 性別。
   - ✓ 性取向。
   - ✓ 性別認同。
   - ✓ 國籍。
   - ✓ 居住地。
   - ✓ 宗教。
   - ✓ 政治。
   - ✓ 職業。
   - ✓ 婚姻。
   - ✓ 懷孕。
   - ✓ 哺乳。
   - ✓ 身體或精神殘疾。
   - ✓ 家庭照護。
2. 建立含偏見/歧視特徵問題集。
   - ✓ 自動/人工輔助驗證特徵抽換後，受測標的產出答覆是否一致。

肆、 結論

本指引包含 AI 產品與系統基本規範及基本檢測基準等文件，旨在協助開發或使用 AI 產品與系統之各級政府機關(構)與本部轄屬產業，建立 AI 產品與系統應遵守或符合之檢驗標準，作為推動我國 AI 評測環境發展與保障運用 AI 產品與系統安全性之依據。

各級政府機關與本部轄屬產業之廠商與企業，如欲確保或證明 AI 產品與系統符合本指引，可依 AI 產品與系統基本規範建議之風險等級與 10 項評測項目，並依循 AI 產品與系統基本檢測基準進行送測。

伍、 參考文件

[1] National Institute of Standards and Techonology, AI Risk Management Framework

[2] European Commission, EU AI Act

[3] 行動應用資安聯盟，行動應用 App 基本資安自主檢測推動制度

[4] 台灣資通產業標準協會，TAICS65-74-WI-01 (物聯網資安標章)認驗證制度規章