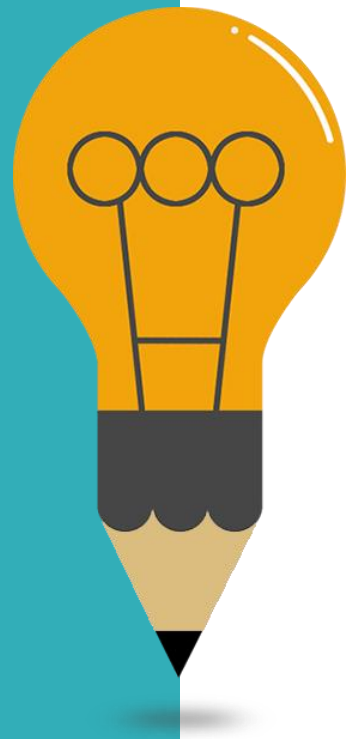
金融業運用人工智慧(AI)指引草案重點內容

訂定背景

- 本會業於112年10月17日公布「金融業運用人工智慧(AI)之核心原則與相關推動政策」，揭示我國金融業運用AI之6項核心原則及8項配套政策
- 配套政策之一係依6項核心原則訂定本指引，以導引金融業運用可信賴AI，發展更貼近民眾需求之金融服務



本指引草案形式

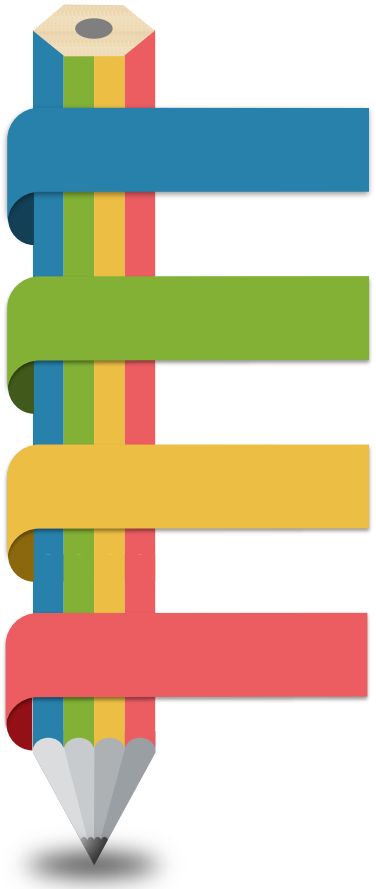
採文件形式而非條文形式

- 行政程序法第165條規定
- 本草案性質屬行政指導，不具法律拘束力，又多涉及實務作業之操作細節
- 參考數位部近期發布之「隱私強化技術應用指引(草案)」，採文件形式而非條文方式辦理

本指引草案架構

- 前言：指引性質及架構
- 總則：共通事項
- 六大章節：6項核心原則之主要概念、建議作法

前言



- 六項核心原則間具有高度關聯，金融業應整體性地交互評估各重點或措施採用之可行性
- 本草案係屬行政指導性質，不具拘束力，旨在鼓勵金融業在風險可控之情況下，導入、使用及管理AI
- 金融業相關公會如訂有運用AI之自律規範，可參考本草案納入相關重點及措施；如未訂定相關自律規範，則建議參考本草案導入、使用及管理AI系統

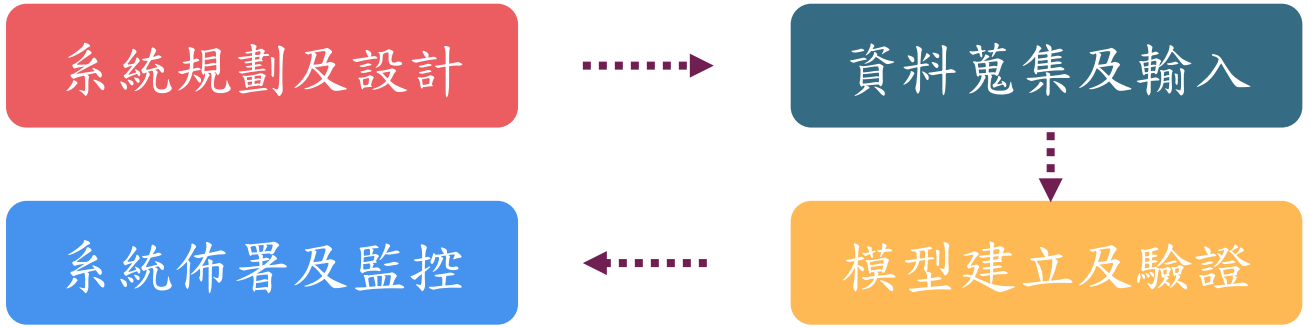
總則章(一)

1.人工智慧(AI)相關定義

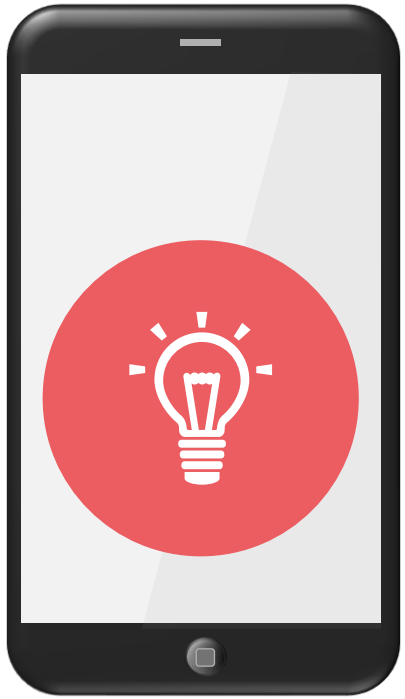
- **採銀行公會「金融機構運用人工智慧技術作業規範」定義**
  - **AI系統定義**：係指透過大量資料學習，利用機器學習或相關建立模型之演算法，模仿人類學習、思考及反應模式之技術
  - **生成式AI定義**：係指可以生成模擬人類智慧創造之內容的相關AI系統

2. AI系統生命週期

- 4個階段



總則章(二)



3.風險評估框架

- 應就相關風險進行評估，並多分配資源於高風險的AI系統
- 風險評估所需考量因素：是否面對客戶、使用個資程度、AI自主決策程度、AI系統複雜性、影響不同利關人、是否提供救濟選項

4.以風險為基礎落實核心原則

- 根據風險評估結果，決定採用之風控措施及程度，並與現行作業相符
- 針對風險較高之AI 系統，評估是否採用記錄、監控機制、審查及核准、稽核或評測機制等措施

5.第三方業者之監督管理

- 評估該第三方業者是否具備相關知識、專業及經驗等，並判斷委託其辦理可能衍生之集中度風險，據以採取適當之監督策略與管理作為
- 注意第三方業者複委託之約定外，亦宜釐清責任分配議題，並訂定適當之退出或轉換機制

第一章 建立治理及問責機制

主要概念

- 金融機構運用AI系統之內部責任與外部責任：內部責任係指明確界定組織內各單位之權責；外部責任係指組織能對外溝通組織之作為
- 應盡量將相關機制與作業予以書面或數位化，並建立適當監督機制
- 不宜將任何一個原則視為一次性或獨立之任務

落實核心原則之建議作法

(一)組織架構及問責機制

- 宜針對組織運用AI系統一事確立組織架構
- 可指定足以督導跨部門業務之高階主管或成立委員會負責整體監督管理AI系統之運用

第一章 建立治理及問責機制

(二)風險管理機制

- 依風險基礎採行制定明確風險管理政策或整合至現有機制
- 就AI模型進行風險管理，包含佈署前之管理、持續驗證、建立模型清單
- 宜對已佈署之AI系統進行維護、監控、記錄及審查
- 定期審查風險管理機制，以確保其有效性

(三)人員之知識及能力

- 宜對負責AI系統之部門、團隊及相關人員提供培訓與資源
- 宜識別新的及變化中的角色，並評估需要提升或重新學習之技能、需聘用新員工之特色等

第二章 重視公平性及以人為本的價值觀

主要概念

- 公平性：金融機構運用AI系統產生之決策，不應對特定群體造成歧視之結果
- 以人為本：應以支持人類自主權、尊重人類基本權利及允許人類監督為原則
- 針對受到不利結果影響之消費者，金融機構應提供相關救濟管道
- 人類在AI系統決策過程中之監督機制：人在指揮、人在迴圈內、人在迴圈上

落實核心原則之建議作法

(一)落實公平性

系統規劃及設計	資料蒐集及輸入	模型建立及驗證	系統佈署及監控
●確立目的及辨識可能受不利影響群體 ●邀請專業人士參與 ●提供救濟管道	●檢視蒐集之數據資料、蒐集方式及來源，是否產生偏見 ●使用多元且具代表性之數據資料	●自行檢驗模型對不同群體之產出結果 ●提請獨立且適格之外部專業人員審查驗證 ●提高可解釋性	●定期檢視與分析AI系統產出之結果是否存在歧視 ●辨識運用AI系統所產生之潛在風險與利益

第二章 重視公平性及以人為本的價值觀

(二)以人為本及人類可控原則之落實方式

- 運用AI系統前，宜先辨識該系統是否遵循法令，並判斷是否未有影響客戶自主權或基本人權之可能
- 宜考量AI決策對客戶或金融機構可能造成傷害之嚴重性及發生機率，採取不同程度之監督機制
- 針對重要之關鍵系統，金融機構宜保留人員可參與，並對AI系統進行審查、核准或最終決策之權利，包括在人員無法控制或介入決策之情況下，仍可由人員安全地關閉AI系統

(三)生成式AI 產出資訊之風險管控

- 導入生成式AI宜評估是否對特定群體產生偏見或歧視之情況，並降低可能之不公平情況
- 使用開放型生成式AI所產出之資訊，仍需由金融機構人員就其風險進行客觀且專業的管控，避免對客戶或金融消費者產生不公平之情況

第三章 保護隱私及客戶權益

主要概念

- 應注意保護客戶隱私權、妥善蒐集及處理其客戶資訊，避免資料外洩風險
- 宜以資料最小化之原則蒐集與處理必要之客戶資料
- 應告知客戶，並尊重其選擇是否使用AI服務之權利及提醒是否有替代方案
- 宜注意保護客戶之智慧財產權與營業秘密

落實核心原則之建議作法

(一)隱私保護及資料治理

系統規劃及設計	資料蒐集及輸入	模型建立及驗證	系統佈署及監控
●評估是否符合個資法等規範 ●遵循資料最小化蒐集處理之原則 ●宜有機制保護個資	●記錄資料蒐集之來源並確保合法管道 ●驗證資料之準確性及完整性 ●蒐集個資應確認已取得客戶同意	●確保訓練資訊及所產生之資訊，不違反個資法 ●確保合作夥伴及供應商亦符合隱私權規範及安全標準	●定期監控AI 系統、合作夥伴及供應商是否持續遵守相關之隱私權規範及安全標準 ●資料外洩或違反個資法時，應通報及處理

第三章 保護隱私及客戶權益

(二)尊重客戶選擇的權利及替代方案

- 向客戶提供金融服務宜注意：(1)告知由AI系統提供；(2)提供AI系統功能及其決策對客戶影響等資訊；(3)提醒是否有替代方案
- 是否同步提供替代方案之考慮因素包括：(1)對金融機構或客戶之風險及危害程度；(2)回復使用AI系統之可能性；(3)替代方案之可行性及成本；(4)同時運用AI系統及替代方案之複雜性及效率性；(5)技術可行性
- 若金融機構決定不提供替代方案，宜進一步評估是否為客戶提供補救措施

第四章 確保系統穩健性與安全性

主要概念

- 系統穩健性：係指AI系統具有預防風險發生之方法，不僅能可靠地按照預設目的執行，且可將非預期或意外傷害降至最低，及防止不可接受之傷害
- 系統安全性：係指具有較強抵禦外部安全威脅、攻擊或惡意濫用之資安防護能力，且符合各金融業資安相關規定要求，並可確保其系統按照應有之功能運行

落實核心原則之建議作法

(一)落實系統穩定性

系統規劃及設計	資料蒐集及輸入	模型建立及驗證	系統佈署及監控
●依據AI系統之目的決定穩健性指標及門檻 ●針對系統失效情況進行風險評估，規劃風險抵減等作法	●根據資料品質及AI系統欲達成之目的適當處理資料 ●可透過自動化工具以確保資料品質	●選擇較具韌性之模型 ●交互驗證與調校 ●可進行對抗性測試 ●有效性驗證並確保達到穩定性指標門檻 ●宜考量是否先於測試環境測試	●適當之環境下佈署AI系統，以減少AI 模型受外部因素影響 ●建立適當之監控機制，持續檢測AI 模型是否效度偏移

第四章 確保系統穩健性與安全性

(二)落實系統安全性

系統規劃及設計

- 提升員工對安全性威脅及風險的認識
- 評估系統潛在威脅
- 除考量功能及效能等因素外，宜將安全性納入考量

資料蒐集及輸入

- 強化資料安全控管，降低資料外洩風險

模型建立及驗證

- 評估及監控AI相關廠商的安全性，並要求廠商遵守資安標準
- 辨識、追蹤及保護AI相關資產
- 針對模型、資料及提示留下完整紀錄

系統佈署及監控

- 保護基礎設施
- 保護模型及資料
- 佈署前先進行適當且有效的安全評估
- 監控並記錄系統的輸入內容
- 使用安全、模組化的更新作業流程

- 宜遵循資訊安全相關規範，建立適當之資安防護或管控措施
- 宜採取管控措施，避免於訓練模型時因第三方業者之不當操作或人為疏失，導致模型參數或資料外洩的風險

第五章 落實透明性與可解釋性

主要概念

- 透明性：提供外部利害關係人有關AI系統之相關資訊，以利了解對其權益之影響等
- 可解釋性：可清楚說明佈署AI模型之演算法如何運作及其預測或決策過程，以利組織內評估是否符合內部政策、作業流程及監管要求等
- 宜主動向利害關係人揭露相關資訊，惟如因資訊過度揭露可能衍生其他風險，宜審慎控制對主管機關以外人員揭露相關資訊之程度
- 宜就AI系統生命週期各階段之透明性及可解釋性擬定共通性標準

落實核心原則之建議作法

(一)落實透明性

系統規劃及設計	資料蒐集及輸入	模型建立及驗證	系統佈署及監控
●依共通性標準決定透明性程度 ●互動時宜以淺白之用語適當揭露資訊 ●主動揭露相關資訊，讓利關人知悉作法	●書面化記錄訓練AI系統之資料的相關資訊	●依透明性要求，測試與驗證相應之功能 ●適當調整客服及客訴，並培訓員工 ●適當揭露更新客戶或網站之約定服務條款	●主動告知該互動或服務係利用AI系統完成 ●宜依據客戶所提出之需求，視情況提供適當之說明及解釋 ●持續監控AI系統透明性達成情形

第五章 落實透明性與可解釋性

(二)落實可解釋性

系統規劃及設計

- 依共通性標準決定提供可解釋性之程度及揭露之對象
- 依可解釋性程度選擇適合的解釋工具
- 備置AI模型工作原理及使用參數之相關文件或技術報告

資料蒐集及輸入

- 書面化記錄訓練AI系統之資料的相關資訊

模型建立及驗證

- 提出可解釋性報告
- 審查及確認相關可解釋性之說明是否妥適
- 驗證其人員是否知悉AI模型之架構、算法及其所使用之功能及決策因素

系統佈署及監控

- 持續監控AI系統可解釋性達成情形

第六章 促進永續發展

主要概念

- 將社會、環境等視為利害關係人，兼顧社會公平及生態責任
- 依據國際永續發展目標及自訂之永續發展原則，適當列入永續發展綜合指標

落實核心原則之建議作法

(一)落實永續發展

- 建立機制辨識與評估AI系統對環境、社會產生之影響或風險
- 選擇能效較高之硬體設備，並優化硬體設施之配置與管理
- 優化AI系統之演算法及減少模型之複雜度與計算需求
- 借重能源效能監控系統，監測AI系統之能源消耗與效能表現
- 提供符合金融消費者需求之服務，降低可能之數位焦慮或落差

(二)員工教育及培訓

- 尊重並保護一般受僱員工的工作權益
- 應對員工提供相關教育及培訓，並視需要建立專案小組
- 提高員工對節約能源、減少資源過度使用及照顧數位弱勢之意識