特　載

金融業運用人工智慧（AI）指引

金融監督管理委員會

壹、前言

鑒於 AI 在金融市場之使用日趨普遍，雖具有增進金融業之效率、降低

本文僅轉載主文，未刊原附錄一「金融業運用人工智慧 (AI) 之核心原則與相關推動政策」及附件「金融業運用 AI 之 6 項核心原則」。如有需要可至金管會官方網站參閱 (https://www.fsc.gov.tw/uploaddowndoc?file=news/202406201527520.pdf&filedisplay=%E9%99%84%E4%BB%B6_%E9%87%91%E8%9E%8D%E6%A5%AD%E9%81%8B%E7%94%A8AI%E6%8C%87%E5%BC%95.pdf&flag=doc)。

成本、提升客戶體驗、管理風險、促進合規、防制金融犯罪、防禦資安事件及促進永續發展等功效，惟如導入 AI 時未經審慎規畫、檢視或使用後未能因應科技或實際成效校調，不僅可能背離原本導入之目的，亦可能衍生金融消費者或金融業之損失、增加業者風險程度，甚至危及大眾對金融市場之信心。為利金融業辨識及注意 AI 系統生命週期宜考量之重點，本指引依「金融業運用人工智慧（AI）核心原則與相關推動政策」內容，並參考國際清算銀行（BIS）、國際證券管理機構組織（IOSCO）、歐盟、新加坡、美國等規範或手冊，提供金融業運用 AI 之指引。

本指引共分總則及六大章節。於總則闡述 AI 相關定義、AI 生命週期、風險評估考量因素、以風險基礎落實核心原則之方式，以及第三方業者之監督管理等共通事項。於第一章至第六章分述金融業在落實 AI 核心原則一至原則六時，依 AI 生命週期及所評估之風險，提出所需關注之重點及可採行之措施。

本指引係屬行政指導性質，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。文件中所舉例子係提供使用情境之參考，金融機構可依本身情況衡酌參採。由於各核心原則間具有高度關聯，金融業參考本指引導入及使用 AI 系統時，宜整體性地交互評估各重點或措施採用之可行性，避免僅將焦點放在單一核心原則，而無法完整控制風險。此外，達成「妥適管理 AI 風險」目的之方式很多，本指引係以風險基礎方式落實核心原則，參考他國及國內目前較佳實務作法提供宜注意之事項（註 1），金融機構可依 AI 系統具體使用情境所涉風險，依各核心原則宜注意之事項，合理選擇緩解風險之機制及落實方法，包括採取更具成本效益之方法達成目的。

本指引各章所指之「金融機構」包含金融控股公司、銀行、信用合作社、票券金融公司、信用卡公司、信託業、電子支付機構、辦理郵政儲金匯兌業務或簡易人壽保險業務之郵政機構、證券商、證券投資信託事業、證券金融

事業、證券投資顧問事業、期貨商、槓桿交易商、期貨信託事業、期貨經理事業、期貨顧問事業、保險公司、保險合作社、保險代理人、保險經紀人及保險公證人。金融業相關公會如訂有運用 AI 之自律規範，可參考本指引納入相關重點及措施；如未訂定相關自律規範，則建議金融機構參考本指引導入、使用及管理 AI 系統。金融機構在運用 AI 系統辦理金融創新業務時，如有必要，金管會鼓勵金融業者可透過金融科技創新實驗或金融業務試辦等機制進行測試。

貳、共通事項

一、 人工智慧（AI）相關定義[註2]

(一)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，進行感知、預測、決策、規劃、推理、溝通等模仿人類學習、思考及反應模式之系統。

(二)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，其內容形式包括但不限於文章、圖像、音訊、影片及程式碼等。

二、 AI 系統生命週期

AI 系統的生命週期主要包括以下 4 個階段：

(一)系統規劃及設計：設定明確的系統目標及需求。

(二)資料蒐集及輸入：資料蒐集、處理並輸入資料庫之階段。

(三)模型建立及驗證：選擇與建立模型演算法及訓練模型，並對模型進行驗證以確保模型效能、安全性與機密性。

(四)系統部署及監控:將系統應用於實際環境中，且關注模型是否已完備，並持續監控以確認系統所帶來之潛在影響。

金融機構運用 AI 系統，可能為自行研發[註3]並使用，因此包含上述 4 階段。金融機構亦可能委託第三方業者研發或購入 AI 系統後，再部署該系統並監控，因此金融機構不盡然均會經歷上開 4 階段。金融機構運用 AI 系統時宜辨識 4 個階段中可自行監控風險之程度，並得對自身較無控制權的部分或事項，透過契約或其他方式與合作廠商明訂風險監控責任之分工。為簡化文字，本指引以「導入（introduce）」AI，表示前述（一）、（二）及（三）3 階段，以「使用（use）」AI 表達第（四）階段。至本指引之「運用（apply）」AI 則係整體性概念，包含上述 4 階段。

三、 風險評估考量因素

金融機構運用 AI 系統時，宜就個別使用情境所涉相關風險進行評估，並宜多分配資源於高風險的 AI 系統，以有效地管理風險。風險評估所需考量之因素如下：（以下所舉例子係為協助說明風險評估情境，非就相關使用情境之風險等級加以規範，運用 AI 系統之風險高低仍由金融機構綜合考量各風險評估因素後自行判斷）

(一)是否直接提供客戶服務或對營運有重大影響

1. 提供客戶服務（面對客戶）之 AI 系統：AI 決策結果對客戶權益或營運有重大影響之 AI 系統，通常有較高之風險性，例如用於信用評分、機器人理財等系統；AI 決策結果僅係提升客戶服務品質者之 AI 系統，風險性可能較低，例如智能客服系統。
2. 用於內部作業（不面對客戶）之 AI 系統：AI 決策結果涉及監理規範之 AI 系統，通常有較高之風險性，例如用於法定資本適足率評估、洗錢防制等系統；AI 決策結果不涉及監理規範之 AI 系統，風險性可能較低，例如用於提升內部行政作業效率之系統。

(二)使用個人資料的程度：AI 系統使用個人原始資料[註4]或機敏性個資程度越高者，可能具有較高之風險性。

(三)AI 自主決策程度：取代人類決策程度較高，或自動化學習程度較高的 AI 系統，可能會增加未預期之系統性負面影響，或減少即時人工干預的機會，而有較高之風險性。

(四)AI 系統的複雜性：運算模型的複雜性較高或使用參數數量與類型較多的 AI 系統，可能降低可解釋性，而有較高之風險性。

(五)影響不同利害關係人（stakeholder）的程度及廣度：AI 系統決策結果對內、外部利害關係人（stakeholder）影響程度較深或影響類型及數量較多者，可能具有較高之風險性。

(六)救濟選項[註5]之完整程度：針對 AI 系統決策結果，如未提供利害關係人（stakeholder）救濟選項或救濟選項較不完整者，可能具有較高之風險性。

四、 以風險為基礎落實核心原則

金融機構宜根據 AI 系統風險評估結果，決定採用之風險控管措施及程度，並確保與其現行實務作法相符。針對風險較高之 AI 系統，除在導入及使用時注意第一章至第六章所列重點及措施外，並評估是否採用下列措施：

(一)記錄：運用高風險系統宜有較完整之書面或數位紀錄。

(二)監控機制：運用高風險系統宜建立較高頻率及廣泛層面之監控機制。

(三)審查及核准：運用高風險系統宜有較嚴格之審查及核准過程，且提高決策層級。

(四)稽核或評測機制：經評估 AI 系統風險、內部資源及專業程度後，如有需要得由第三方稽核或評測單位進行獨立驗證。如導入之 AI 系統由同集團（包含其關係企業）開發或管理，相關稽核得以集團提供之資料替代。

五、第三方業者之監督管理

金融機構委託第三方業者導入 AI 系統相關作業時，宜採行以下監督管理措施：

(一)金融機構宜先進行檢視，評估該第三方業者是否具備相關知識、專業及經驗等，並判斷委託其導入可能衍生之集中度風險（金融機構自身委託該機構之集中度風險），再根據評估結果採取適當之監督策略與管理作為，以防止可能之風險或問題。

(二)金融機構宜與第三方業者簽訂書面契約，明定導入事項範圍、第三方業者之責任範疇，以及未達績效目標或發生不良事件之追索途徑。

(三)金融機構委託第三方業者導入相關作業時，如有涉及將客戶資料傳送第三方業者進行處理之情況，宜與第三方業者簽訂含有資料保護條款之協議，明確規定資料加密傳輸、存儲安全以及在服務終止後資料之處置方式。

(四)金融機構委託第三方業者導入 AI 系統，或金融機構進行測試及監控等作業時，除注意第三方業者複委託之約定外，亦宜釐清責任分配議題，並就停止委託之情形訂定適當之資料或系統遷移機制。

(五)金融機構宜要求第三方業者留存執行受託辦理事項之書面或數位作業紀錄，俾利後續追蹤、驗證及管理。

(六)金融機構委託第三方業者導入事項如涉及金融業作業委外事項，應符合各業別相關委外規範。

參、建立治理及問責機制

核心原則一：建立治理及問責機制

(一) 金融機構應對其使用之 AI 系統承擔相應之內、外部責任。內部責任包含指定高階主管負責 AI 相關監督管理並建立內部治理架構；外部責任則涉及對消費者與社會之責任，包括保護消費者之隱私及資訊安全等。

(二) 金融機構應建立全面且有效的 AI 相關風險管理機制，並整合至現行風險管理及內部控制作業或流程中，且應進行定期的評估及測試。

(三) 金融機構應確保其人員對 AI 有足夠之知識及能力，並應以風險為基礎做出適當之決策及監督。

* 本核心原則係依據金管會 112 年 10 月 17 日公布之「金融業運用人工智慧（AI）之核心原則與相關推動政策」。

一、目的

金融機構可能運用多個 AI 系統，因此建議宜有明確管理 AI 系統的架構及風險管理政策，掌握 AI 系統設置的目的、適用之業務或作業、負責的人員，且對內要能夠解釋清楚系統的運作邏輯、對外要能說明整體政策、消費者對個別 AI 系統需知悉之事項等，並有完整之處理錯誤或非預期事件之程序。此外，金融機構宜持續提升人員對 AI 系統導入、使用及管理之瞭解與能力，以適應 AI 技術的快速發展與變化。

二、主要概念

(一)金融機構運用 AI 系統之內部責任與外部責任

1.內部責任係指明確界定組織內各單位之權責，包含宜有明確內部治理架構、由可督導跨部門職務之高階主管或指定之委員會進行監督及管理、界定各部門或各業務線之功能與責任，及落實分層管

理機制等。

2.外部責任係指組織能對外溝通組織之作為，包含具有管道或溝通機制可讓外界查詢或審視受決策影響事項之相關資訊，並確保在運用 AI 系統時係符合規畫目的。

(二)金融機構於落實治理及問責原則時，宜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。

(三)金融機構宜整體性落實金融業 AI 核心原則，不宜將任何一個原則視為一次性或獨立之任務。

三、 組織架構及問責機制

(一)負責 AI 系統之組織架構與角色職責：金融機構宜針對組織運用 AI 系統一事確立組織架構，包括是否指定整體負責 AI 系統之相應部門或團隊。對於每個部門或團隊、運用 AI 系統之業務線及對於 AI 生命週期各個階段作業及活動之任務單位，宜明確界定其職責、人員之角色、功能及擔負之相關責任。

(二)指定高階主管或委員會監督協調：金融機構可指定足以督導跨部門業務之高階主管或指定之委員會負責整體監督管理 AI 系統之運用。該高階主管或委員會及所帶領之部門或團隊，宜制定 AI 政策，並負責監督 AI 系統之使用；如金融機構自行導入 AI 系統，則亦應監督 AI 系統各階段生命週期之系統規劃及設計、資料蒐集及輸入、模型建立及驗證，且確保運用過程遵守法令。

四、 風險管理機制

(一)依風險基礎訂定明確風險管理政策或整合至現有機制

1.金融機構可就運用 AI 系統訂定明確之風險管理政策與指導方針，

涵蓋項目宜包含風險管理、資料蒐集、安全控管、法遵要求、監測及評估等。金融機構並宜形塑有利於 AI 發展之組織文化，鼓勵落實 AI 核心原則及實現負責任 AI，如員工對 AI 系統有疑慮時，宜有機制允許員工提出疑問或擔憂。

2.將 AI 風險管理整合至現有的風險管理及內部控制架構：整合項目包括模型風險管理、資訊安全、資料保護及公平待客等現有架構，如尚有不足，可再增訂納入風險管理及內部控制架構以符合 AI 核心原則。

(二)AI 模型之風險管理

1.部署前之管理：金融機構宜瞭解並記錄 AI 模型的目的及預期用途，以及 AI 模型所使用的方法及其概念。金融機構在部署前宜對模型進行測試，以確保其產出結果符合預期目的。

2.持續驗證：金融機構宜儘可能對 AI 模型進行持續驗證，惟驗證頻率可能因模型的複雜性及定期審查工具性能而有所不同。金融機構可評估該模型之可靠性、識別度及修正錯誤，並檢查模型產出結果之品質是否有逐漸劣化之趨勢。

3.建立模型清單：金融機構宜建立並維護 AI 模型清單，包括每個 AI 模型的過去、現行及開發中版本之資訊[註 6]，包括資訊輸入的類型及來源、模型的輸出及預期用途，以及模型是否按預期運作的評估。

(三)持續監控與精進：金融機構宜對已部署之 AI 系統進行維護、監控、記錄及審查，及依據風險評估考量因素提供適當資源，並使管理階層瞭解已部署之 AI 系統的表現及其他相關問題。在適當的情況下，監控可以包括自主監控，例如 AI 系統可以被設計為自動報告其預測的信賴水準。

(四)定期審查風險管理機制，以促進其有效性

1.建立內部審查與監測機制：金融機構宜建立內部審查與監測機制，依風險基礎定期評估 AI 系統是否符合原先運用目的及風險程度，以使 AI 系統符合政策與指導方針，並及時解決可能存在之問題。金融機構必要時可邀請不同領域人員參與 AI 評估過程，例如人力資源、行為科學、法律、倫理、永續發展等領域，以協助為 AI 之發展提供正確的方向。

2.建立獨立第三人審查及溝通機制：金融機構針對風險程度較高之 AI 系統，經評估 AI 系統風險、內部資源及專業程度後，如有需要得建立由具 AI 專業之獨立第三人進行審查、評測之機制，並具有管道或溝通機制可讓外界查詢或審視受決策影響事項之相關資訊，以透過外部之回饋，使金融機構運用 AI 系統符合各項核心原則。

五、人員培訓

(一)金融機構宜對負責 AI 系統之部門、團隊及相關人員提供培訓與資源，以提升人員對 AI 系統導入、使用及管理之了解與能力、適應 AI 技術的快速發展與變化，並能以風險為基礎做出適當之決策及監督。這些人員包括負責整體 AI 系統之高層人員、專案負責人（例如開發、測試、監督、法遵、風控及內部稽核等）、監管人員、執行團隊及其他相關人員等。金融機構亦宜確認董（理）事會及管理階層對金融機構所運用之 AI 系統有所認識。

(二)金融機構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，以使組織更快適應新工作方式及實現有效的人機協作領域。

(三)金融機構並宜建立與利害關係人（stakeholder）之溝通及互動管道，

讓運用 AI 系統者容易將各界反饋意見評估納入生命週期之各階段。

肆、重視公平性及以人為本的價值觀

核心原則二：重視公平性及以人為本的價值觀 (一) 金融機構在使用 AI 系統之過程中，應儘可能避免演算法之偏見所造成的不公平。 (二) AI 系統之運用應符合以人為本及人類可控之原則，並尊重法治及民主價值觀。 (三) 生成式 AI 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。

* 本核心原則係依據金管會 112 年 10 月 17 日公布之「金融業運用人工智慧（AI）之核心原則與相關推動政策」。

一、目的

由於 AI 系統自動化之特性，若設計、蒐集數據、建立模型等時未能謹慎注意，可能造成歧視或不公平之現象，甚至超脫人類之控制，因此金融機構運用 AI 系統時，宜評估公平性，提升其決策的合理性及準確性，注意偏見（bias）之產生並儘可能避免歧視（discrimination）。蒐集資料時宜注意其來源及可能產生的偏見，謹慎使用個人屬性資料，並定期審查及驗證 AI 模型產出之結果，以提升 AI 系統達成其目的之準確性，以及達到儘可能避免歧視的目標。此外，AI 系統之運用應支持人類自主權及尊重基本人權，並允許人類監督。

二、主要概念

(一)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，亦即決策需有合理性、準確性及儘可能避免歧視。

1.決策之合理性：（1）如利用個人屬性做為 AI 模型決策之因素之一，應有合理理由；（2）如無合理理由，運用 AI 系統所產生之決策則不應對特定群體有系統性之不利差別待遇（例如不得以特定宗教、種族、性別、身心障礙、性傾向、居住所、政治傾向、年齡、國籍或族群等因素，對借款人提供不合理的貸款條件）。

2.決策之準確性及儘可能避免歧視：宜定期審查及驗證 AI 決策模型與數據，以提升準確性及達到儘可能避免歧視的目標，另亦宜定期審查模型產生出來之決策結果，以確保模型演算後之結果符合設計之目的。

(二)以人為本：AI 系統在其全生命週期中，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，以落實人類價值，並達到改善人類福祉之目標。

(三)針對受到不利結果影響之消費者，金融機構宜提供相關救濟選項，其中可包含金融機構既有之救濟選項。但金融機構運用之 AI 系統如與洗錢防制或詐騙偵防有關，而不適合提供救濟選項者，得不提供。

(四)人類在 AI 系統決策過程中之監督機制，可分為人在指揮（HIC）、人在迴圈內（HITL）、人在迴圈上（HOTL），說明如下：

1.「人在指揮（Human-in-command）」：指人類指揮監督 AI 系統之整體活動（包括其更廣泛的經濟、社會、法律及道德影響），並在任何情況下決定何時、如何使用 AI 系統的能力。

2.「人在迴圈內（Human-in-the-loop）」：表示人類主動參與監督，並保留完全的控制權，AI 系統僅係提供建議或資訊。除非人類下達命令要求 AI 系統決策，否則 AI 系統不能進行決策。

3.「人在迴圈上（Human-over-the-loop）」：人類僅有在 AI 模型遇

到意外或不良事件（例如模型失敗）時，才接管控制，並在運算過程中調整參數。

三、公平性之落實方式

金融機構宜在 AI 系統生命週期各階段注意下列事項，並評估公平性風險程度，以風險為基礎採行適當措施：

(一)「系統規劃及設計」階段

1. 確立目的及辨識可能受不利影響群體：金融機構宜確認其規劃 AI 系統之目的，並辨識運用 AI 系統時可能受到系統性不利差別待遇之群體（以下簡稱受不利影響群體）及可能之受影響程度，並留下書面或數位紀錄。
2. 邀請專業人員參與：金融機構於必要時可邀請專業人員參與 AI 系統的設計與執行過程。透過他們專業建議與指導，使 AI 系統之設計與執行不歧視任何個人或群體。
3. 提供救濟選項：金融機構宜在 AI 系統加入或提供受不利影響群體可反饋意見之救濟選項，蒐集該等群體之意見或建議，以利金融機構作為後續校調之參考。

(二)「資料蒐集及輸入」階段

1. 金融機構宜檢視擬蒐集之數據資料、其蒐集之方式及數據資料之來源，是否有可能會產生偏見，並注意這些偏見是否造成不公平或歧視。
2. 金融機構宜盡量使用多元、包含各種背景與特徵且具代表性之數據資料，而非僅依賴單一類別或群體之數據，以減少對某些群體的偏見與歧視。
3. 使用個人屬性（例如年齡、性別、居住所、族群、宗教、國籍等）

時宜謹慎。

(1)在決定蒐集或採用某個人屬性前，宜辨識該屬性之採用是否符合 AI 系統規劃之目的。

(2)如某個人屬性之採用可能產生受不利影響群體，宜評估採用該屬性之必要性及採用替代屬性或替代作法之可能性，並將評估結果、採行之作法及理由以書面化或數位方式留存。

(三)「模型建立及驗證」階段

1.自行檢驗模型對不同群體之產出結果：金融機構如係自行或委託研發 AI 系統，可透過測試與驗證 AI 模型對不同群體的預測及決策，來確認其運作是公平與無偏頗的。如果發現有偏頗存在，金融機構應評估採取相應的調整與改進措施，以儘可能避免不公平或歧視。

2.提請獨立且適格之外部專業人員審查驗證：如有需要，金融機構可將 AI 系統之產出結果提交給獨立且適格之外部專業人員進行審查及驗證，以確認 AI 系統之決策符合合理性、準確性及儘可能避免歧視，並作出必要之修正或改善。金融機構亦宜進行盡職調查，防止可能之利益衝突。

3.留下書面或數位紀錄：為利驗證 AI 模型符合公平性，金融機構宜將 AI 模型設計之目的、運算邏輯等及決策程序等，以簡單易懂方式留下書面或數位紀錄，使相關利害關係人（stakeholder）能夠了解模型之運作原理與決策方式，以利判斷 AI 系統對各種群體之公平性，並能追溯與解釋相關結果。

(四)「系統部署及監控」階段

1.金融機構宜定期檢視與分析 AI 系統產出之結果是否存在歧視，並透過救濟選項蒐集之資訊，以確定 AI 模型對不同群體是否存在不平等對待之情況。如果發現歧視問題，應及時進行調整改進。

2.金融機構宜辨識運用 AI 系統與受系統性不利差別待遇之特定群體間是否具有關聯性，如是，金融機構宜採行降低該特定群體受影響之方式。

四、 以人為本及人類可控原則之落實方式

(一)金融機構運用 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，如否，則宜先停止運用並採取改善措施直到疑慮消除。

(二)針對 AI 系統之應用領域，金融機構於評估使用 AI 輔助決策所需之人類參與程度時，宜考量 AI 決策對客戶或金融機構之影響程度，採取不同程度之監督機制，或規劃其他可能之風險抵減、轉移或規避措施。對於前述影響程度較高之應用，宜採取人在指揮或人在迴圈內之監督機制。

(三)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，包含確保金融機構之人員能介入控制，且 AI 系統可提供足夠資訊供人員做出有意義之決策；或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。

五、 生成式 AI 產出資訊之風險管控方式

(一)金融機構導入生成式 AI，亦宜依上開各重點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。

(二)金融機構使用第三方業者開發之生成式 AI，如無法掌握訓練過程及確保其數據或運算所得出之結果符合公平性時，金融機構對其產出之資訊，仍需由其人員就其風險進行客觀且專業的管控，以避免對客戶或金融消費者產生不公平之情況。

伍、保護隱私及客戶權益

核心原則三：保護隱私及客戶權益 (一) 金融機構應充分尊重及保護消費者之隱私，並妥善管理及運用客戶資料。 (二) 金融機構如運用 AI 系統向客戶提供金融服務，應尊重客戶選擇的權利，並提醒客戶是否有替代方案。

* 本核心原則係依據金管會 112 年 10 月 17 日公布之「金融業運用人工智慧（AI）之核心原則與相關推動政策」。

一、目的

AI 系統為達成準確性目的，可能需要蒐集消費者或客戶大量資訊，且與客戶互動時亦同時蒐集處理其資訊，因此金融機構在 AI 系統之生命週期均應注意保護客戶的隱私權，妥善處理其客戶資料，避免資料外洩風險，並採用資料最小化原則[註 7]，避免蒐集過多或不必要的敏感資訊。金融機構並宜尊重客戶選擇是否使用 AI 服務的權利，根據客戶及機構的風險、替代方案可行性及成本來決定是否提供替代方案。

二、主要概念

(一)在大數據及 AI 技術發展下，客戶之個人資訊常被大量蒐集並用以訓練 AI，可能對客戶隱私權造成潛在威脅，進而影響民眾對金融機構之信任度及服務滿意度，故金融機構在運用 AI 系統時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。

(二)金融機構宜以資料最小化之原則蒐集與處理必要之客戶資料，並避免蒐集過多或不必要之敏感資訊。

(三)金融機構運用 AI 系統面對客戶時，宜告知客戶，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，以維護客戶權益。

(四)金融機構運用 AI 系統時，應注意保護客戶隱私，包含個人資料、智慧財產權與營業秘密等。

三、 隱私保護及資料治理

(一)針對客戶隱私保護及資料治理之落實，金融機構宜在 AI 系統生命週期各階段注意以下事項：

1.「系統規劃及設計」階段

(1)金融機構應評估 AI 系統之設計是否符合個人資料保護法等相關規範及其內部資料治理政策。

(2)金融機構宜依 AI 系統設計之目的，評估所需蒐集之資料及判斷資料取得管道之可行性，並遵循資料最小化蒐集處理之原則。如所蒐集之公開資料已可滿足 AI 系統設計之目的，則不需蒐集非公開資料。

(3)金融機構宜有機制保護個人資料免受未經授權之存取、損壞、損失或洩露，並依「金融監督管理委員會指定非公務機關個人資料檔案安全維護辦法」相關規定辦理。金融機構可採取之措施包含加密敏感資料、設置存取控管機制、定期進行安全監控及審查，以及確保員工接受適當之保密訓練。

2.「資料蒐集及輸入」階段

(1)金融機構宜記錄資料蒐集之來源並確保係透過合法及可靠之管道取得。

(2)金融機構宜驗證資料之準確性及完整性，並檢核資料是否存在錯誤。

(3)金融機構所蒐集之資料如包含個人資料，應確認已取得客戶同意或符合相關法令，並以風險為基礎評估是否須進一步對資料進行額外之隱私保護處理（註 8）。

3.「模型建立及驗證」階段

(1)金融機構宜確保用以訓練 AI 模型之資訊及 AI 系統所產生之資訊，不違反個人資料保護法及相關規範。

(2)金融機構宜確保合作夥伴及供應商亦符合隱私權規範及安全標準。

4.「系統部署及監控」階段

(1)金融機構宜定期監控 AI 系統、合作夥伴及供應商於部署後是否持續遵守相關之隱私權規範及安全標準，及是否有損及金融消費者隱私權或權益保障之異常情況。

(2)AI 系統如有資料外洩或違反個人資料保護法之情事時，金融機構應循現行機制通報及處理，並視需要調校 AI 系統。

(二)金融機構應注意運用 AI 系統可能導致之客戶資料外洩風險。以生成式 AI 之運用為例，在無適當管控機制下，金融機構人員不得向生成式 AI 提供未經客戶同意提供之資訊。前稱適當管控機制，例如採用封閉型部署之生成式 AI 模型、確認系統環境安全性等。如不涉及個人資料、無法直接或間接辨識特定個人之客戶行為者，則不在此限。

四、 尊重客戶選擇的權利及替代方案

(一)金融機構如運用 AI 系統向客戶提供金融服務，宜注意以下事項：

1.告知金融消費者該金融服務係由 AI 系統所提供。

2.提供資訊以便金融消費者瞭解，在正常使用過程中，AI 系統之功能為何及由 AI 協助做出之決策可能會如何影響他們。

3.提醒金融消費者是否存在 AI 系統以外之替代方案，以讓其自行決定是否選擇使用 AI 系統所提供之服務。

(二)金融機構可參考下列因素，決定於提供客戶退出使用 AI 系統服務時是否同步提供替代方案：

1.對金融機構之風險及危害程度。

2.對客戶之風險及危害程度。

3.客戶選擇替代方案後回復使用 AI 系統之可能性。

4.替代方案之可行性及成本。

5.同時運用 AI 系統及替代方案之複雜性及效率性。

6.技術可行性。

(三)若金融機構在權衡上述因素後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。

陸、確保系統穩健性與安全性

核心原則四：確保系統穩健性與安全性
(一) 金融機構在運用 AI 系統時，必須確保其系統之穩健性 (robustness) 與安全性，以避免對消費者或金融體系造成損害。
(二) 若金融機構運用第三方業者開發或營運之 AI 系統提供金融服務，應對第三方業者進行適當之風險管理及監督。

* 本核心原則係依據金管會 112 年 10 月 17 日公布之「金融業運用人工智慧（AI）之核心原則與相關推動政策」。

一、目的

當金融機構運用 AI 系統之情形愈加普遍時，系統之穩健及安全對金融機構之正常運作即愈顯重要，因此金融機構宜明確定義系統目的並進行風

險評估，選擇具韌性的 AI 模型，蒐集資料時宜注重資料品質，視情況對模型進行交互驗證及對抗性測試，並於適當之環境下部署 AI 系統。此外，亦應建立資安防護措施，防範各種安全威脅及攻擊，並進行持續監控，以確保 AI 系統的整體安全與穩定運行。

二、主要概念

(一)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不利影響降至最低，及防止不可接受之不利影響。系統穩健性包含以下概念：

1. 穩定性：穩定性佳之 AI 系統，係指電腦系統在執行過程中及面對錯誤輸入時，具有良好應對能力，即便該系統或其組件在無效輸入或壓力環境條件下，仍能正確運作。
2. 準確性：準確性佳之 AI 系統，係指系統有能力做出正確（correct）判斷以達成其規劃目的，例如將資訊正確地分類到適當之類別，或根據數據、模型做出符合規劃目的之預測、推薦或決策。AI 系統若經過完善地規劃、開發，可以減少及糾正不準確預測所帶來之非預期風險。即便當 AI 系統偶爾出現不準確的預測時，其亦能夠於檢驗時指出其錯誤率。
3. 可重製性：具可重製性之 AI 系統，係指在相同的條件下重複 AI 系統之測試，仍會得到相近的產出。

(二)系統安全性：安全性高之 AI 系統，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。

三、系統穩健性之落實方式

金融機構宜在 AI 系統生命週期各階段注意以下事項：

(一)「系統規劃及設計」階段

1.金融機構宜明確定義 AI 系統之目的，並依據此目的，決定用以衡量系統穩健性之指標以及在該指標上所應達到之門檻（標準）。

2.金融機構宜針對 AI 系統無法達成原規劃目的之情形進行風險評估，並規劃可能之風險抵減、轉移或規避等作法。

(二)「資料蒐集及輸入」階段

1.資料治理：高品質之資料係確保 AI 模型具穩定性、準確性及可重製性之基礎，因此金融機構宜根據資料品質及 AI 系統欲達成之目的適當處理資料，如補足資料之缺失值、進行資料編碼、資料標準化等。

2.金融機構亦可透過自動化工具以確保資料品質。

(三)「模型建立及驗證」階段

1.金融機構宜選擇較具備韌性（resilience）[註9]之模型，同時亦必須確認此模型符合金融機構所欲達到之目的。

2.金融機構如係自行或委託研發 AI 系統，可透過交互驗證（cross-validation）與調校等技術，進一步增進 AI 模型之穩健性及可靠性。

3.金融機構可進行對抗性測試，以評估 AI 模型在面對非預期輸入時的韌性，並做必要之調整。

4.金融機構宜進行有效性驗證並確認 AI 模型達到初始設定之系統穩定性指標門檻。

5.對於風險較高或影響幅度較大之 AI 系統，金融機構宜考量是否先測試，且測試環境與日常環境分離。金融機構亦可測試 AI 系統在具有壓力之市場條件下之表現。

(四)「系統部署及監控」階段

1.金融機構宜於適當之環境下部署 AI 系統，以減少 AI 系統受外部

因素影響（例如電力穩定、網路頻寬）而降低效能。

2. 金融機構宜建立適當之監控機制，定期檢測 AI 模型是否有效度偏移之狀況，並於模型準確性下降或出現其他問題時，即時進行處理。

四、 系統安全性之落實方式

(一)金融機構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，防範各種安全威脅及攻擊，如駭客攻擊、惡意軟體等，並持續監控運作結果，確保 AI 系統之安全性。

(二)金融機構宜採取管控措施，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。

(三)金融機構宜在 AI 系統生命週期各階段注意以下事項：

1.「系統規劃及設計」階段

(1)提升員工對安全性威脅及風險的認識，並協助相關人員規劃減輕風險的合適方法。

(2)評估系統之潛在威脅。

(3)選擇 AI 模型時除考量功能及效能等因素外，宜將安全性納入考量。

2.「資料蒐集及輸入」階段：強化資料安全控管，降低資料外洩風險。

3.「模型建立及驗證」階段

(1)評估或定期檢視 AI 相關廠商的安全性，並要求廠商遵守資安標準。

(2)辨識、追蹤及保護 AI 相關資產（例如模型、資料、提示（prompt）、軟體、紀錄文件、內部評估等）。

(3)針對模型、資料及提示留下相關書面或數位紀錄。

4.「系統部署及監控」階段

(1)保護基礎設施，包含對 API、模型及資料之使用進行適當控管、積極防範竊取模型或損害效能之網路攻擊等。

(2)保護模型及資料，例如透過落實資訊安全實務作法或管控使用者界面等。

(3)AI 模型部署前先進行適當且有效的安全評估。

(4)監控模型及系統的輸出與效能，以觀察可能影響安全性的變化。

(5)在符合隱私及資料保護之前提下，記錄並適當監控系統的輸入內容。

(6)使用安全、模組化（modularization）的更新作業流程。

柒、落實透明性與可解釋性

核心原則五：落實透明性與可解釋性
(一) 金融機構在運用 AI 系統時，應確保其運作之透明性及可解釋性。
(二) 金融機構使用 AI 與消費者直接互動時，應適當揭露。

* 本核心原則係依據金管會 112 年 10 月 17 日公布之「金融業運用人工智慧（AI）之核心原則與相關推動政策」。

一、目的

當金融機構運用 AI 與消費者互動時，其決策或互動內容與消費者息息相關，因此宜向消費者適當揭露與其相關之資訊，對於自行或委託研發之 AI 系統，金融機構宜確認其人員必要時可對內及對稽核人員清楚說明 AI 系統運作之邏輯，以利面對需要修正或檢視 AI 系統時，可在妥適情形下進行。

二、主要概念

(一)透明性：係指提供外部利害關係人（stakeholder）有關 AI 系統之相

關資訊，以利其了解對其權益之影響等，以及該等 AI 系統的限制與風險。

(二)可解釋性：係指可清楚說明自行或委託研發並使用之 AI 系統如何運作及其預測或決策過程背後之邏輯，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。

(三)金融機構宜理解其運用之 AI 系統如何做出決策並宜提高 AI 系統的可解釋性，以確認對 AI 系統運作之有效管理。

(四)金融機構運用 AI 系統時，宜主動向利害關係人揭露相關資訊，如利害關係人要求進一步說明，宜適當說明所使用之資料、資料如何影響決策，及決策對利害關係人之影響等，以提升民眾信任度。但金融機構運用之 AI 系統如與洗錢防制、資訊安全、詐騙偵防有關，或如涉及企業營業秘密，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必要性及程度。

(五)金融機構宜就 AI 系統生命週期各階段之透明性及可解釋性擬定共通性原則：

1.透明性

(1)金融機構宜就評估 AI 系統所需之適當透明性訂定共通性原則，例如運用 AI 系統產生貸款決策建議時，對客戶說明解釋之程度、時機及形式等。

(2)金融機構宜確認在客戶往來生命週期各階段中，可能需對客戶通知之項目為何，並事先備妥相應之通知樣版。

2.可解釋性

(1)金融機構宜就評估 AI 系統可解釋性訂定共通性原則，包含如何評估所需之可解釋性程度及相關資訊提供對象等。

(2)金融機構宜就 AI 系統選定適合的解釋方法，並明定前揭解釋方

法之基本要求。

三、 透明性及可解釋性之落實方式：

金融機構宜在 AI 系統生命週期各階段注意以下事項：

(一)「系統規劃及設計」階段

1.透明性

(1)金融機構宜依所定之透明性共通性原則，決定 AI 系統之透明性程度，並確認在客戶生命週期各階段所需的主、被動通知項目及其形式。

(2)金融機構規劃以 AI 系統與金融消費者互動時，宜規劃如何以淺白之用語適當揭露相關資訊，例如：

- 讓金融消費者知悉其正考慮之金融商品或服務係由 AI 提供的。
- 讓金融消費者瞭解其在正常使用過程中，AI 系統會提供之功能及預估之表現為何，並讓金融消費者知悉可能受到之正反面影響。
- 如金融機構係運用 AI 系統輔助決策，宜考慮提供金融消費者額外之資訊，以便其瞭解這些決策是如何形成的。

2.可解釋性

(1)金融機構宜依準備期階段所評估之可解釋性共通性原則，決定所需之可解釋性程度及相關資訊提供對象。

(2)依前揭可解釋性程度選擇適合的解釋方法，並評估該解釋方法是否符合所定之基本要求。

(3)金融機構如係自行或委託研發 AI 系統，宜規劃備置 AI 系統架構之相關文件或技術報告，並規劃提供監理機關存取及了解 AI 系統與所使用數據之權限，以便未來其查詢與瞭解金融機構 AI 系統之運作及使用數據之妥適性。

(二)「資料蒐集及輸入」階段：為因應未來透明性或可解釋性要求，金融機構宜確保書面化或以數位方式記錄訓練 AI 系統之資料等相關資訊（例如資料來源、時間點、消費者同意、資料字典及資料已知限制等）。

(三)「模型建立及驗證」階段

1.透明性

(1)金融機構宜依透明性要求，測試與驗證相應之功能。

(2)金融機構宜規劃如何適當調整作業流程（例如客服及客訴處理流程），以強化對客戶之透明性，並培訓相關員工以應對客戶諮詢。

(3)金融機構宜規劃如何適當揭露並更新客戶或網站之約定服務條款，例如向客戶說明 AI 系統如何使用客戶資料、客戶可能會有之利益與風險，以及客戶如何參與、退出及提出問題等。

2.可解釋性

(1)金融機構自行或委託研發 AI 系統者，宜依可解釋性之要求，提出可解釋性報告。

(2)金融機構宜審查及確認相關可解釋性之說明是否妥適，並確認可解釋性之程度與其 AI 系統應用之重要性相稱。如果金融機構對於 AI 系統之可解釋性未達預期，則宜規劃採取其他措施（例如切換至更簡單之模型、刪除難以解釋之功能或引入更多之人工監督等）。

(3)金融機構於模型建置後，宜驗證其人員是否知悉 AI 系統之架構、算法及其所使用之功能（feature）及決策因素，且 AI 系統之運作過程可被理解及解釋。如其人員未能清晰明確表達，宜檢討是否避免使用過於複雜或無法解釋之架構、演算法或功能等。

(四)「系統部署及監控」階段：

1.金融機構使用 AI 與消費者直接互動前，宜主動告知該互動或服務係利用 AI 系統自動完成。

2.金融機構宜依據客戶所提出之需求，視情況提供適當之說明及解釋，尤其對於可能受到系統性不利差別待遇之客戶，金融機構可評估以簡單易懂之方式，說明 AI 系統產出之預測、建議或其決策基礎之邏輯，惟仍宜注意資訊過度揭露所衍生之風險。

3.金融機構宜定期監控 AI 系統透明性及可解釋性達成情形。

4.為提升市場對 AI 的信任度，如有需要，金融機構亦可規劃透過發布報告、技術文件或於網站上揭露相關資訊等方式，主動讓利害關係人（stakeholder）知悉其運用 AI 系統之做法，包括說明 AI 系統之用途、運作原理、所使用之數據、算法及可能之影響等。

捌、促進永續發展

核心原則六：促進永續發展
(一) 金融機構在運用 AI 系統時，應確保其發展策略及執行與永續發展之原則相結合，包括減少經濟、社會等不平等現象，保護自然環境，從而促進包容性成長、永續發展及社會福祉。
(二) 金融機構在 AI 系統運用過程中，宜對一般員工提供適當之教育及培訓，促進員工能適應 AI 帶來之變革，並盡可能維護其工作權益。

* 本核心原則係依據金管會 112 年 10 月 17 日公布之「金融業運用人工智慧 (AI) 之核心原則與相關推動政策」。

一、目的

AI 系統之運作可能耗能、耗水，亦可能對現有一般員工造成工作之剝

奪或威脅感，且加劇數位落差，因此金融機構應重視社會及環境責任，如利用新興科技減少資源消耗及促進普惠金融數位轉型。同時，金融機構宜在數位轉型時兼顧一般員工之轉型，追求永續穩定發展。

二、主要概念

(一)金融機構運用 AI 系統時，宜將社會、環境等視為利害關係人（stakeholder），兼顧社會公平及生態責任，例如在使用過程中，促進普惠金融數位轉型、降低數位落差、減少水、電等能源消耗問題。

(二)金融機構運用 AI 系統之策略及執行方向，宜依據國際永續發展目標及自訂之永續發展原則，並適當列入永續發展綜合指標。

三、永續發展之落實方式

(一)辨識產生之影響：金融機構宜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。

(二)優化硬體設施：金融機構可選擇能效較高之硬體設備，以減少能源消耗，例如採用節能伺服器、低功耗處理器及高效能數據中心設備等，並優化硬體設施之配置與管理，以提高能源利用效率。

(三)共享資源及虛擬化：金融機構得透過虛擬化技術及資源共享方式，將運算資源及資料倉儲集中管理，減少設置重複之硬體設置，從而節省能源消耗。

(四)改進模型與演算法：金融機構可優化 AI 系統之演算法及減少模型之複雜度與計算需求，從而提高運算效率及資源利用率。

(五)預先處理資料：金融機構可預先處理數據，以減少不必要的數據傳輸，並可透過提升資料品質，減少為提升準確率而重複運算所消耗之

能源。

(六)智慧控管能效：金融機構可借重能源效能監控系統，實時監測 AI 系統之能源消耗與效能表現，及時發現與解決能源浪費之問題，並持續改進系統之能源效率。

(七)回收與再利用資源：為減少地球資源之過度使用，金融機構可對舊有硬體設備進行資源回收與再利用，以減少電子廢棄物之產生。金融機構並可多利用使用權概念運用硬體設備等，減少對原物料之利用。

(八)降低數位焦慮及數位落差：金融機構可依據金融消費者屬性，提供符合其需求之服務，以降低可能之數位焦慮或落差，例如提供金融消費者選擇面對面服務之機會，以減少客戶對 AI 系統之焦慮；或提供數位體驗，提升數位運用上較為弱勢族群採用數位服務之誘因。

四、員工教育及培訓相關事項

(一)金融機構必須尊重並保護一般受僱員工的工作權益，包括在數位轉型過程中，提供適當的教育及培訓以助其適應新的工作環境、減少失業風險。

(二)金融機構在運用 AI 系統時，宜對員工提供相關教育及培訓，包括 AI 基本概念及運作方式、AI 如何影響各部門及工作流程等，並視需要建立專案小組，負責監控 AI 系統之影響及員工適應情況，以及根據實際需求調整教育及培訓計畫。

(三)金融機構亦可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，並提供相應之培訓及指導，以促進其對永續發展之實踐。

參考資料

1. Board of Governors of the Federal Reserve System, and Office of the Comptroller of the Currency (OCC), 2011, "Supervisory Guidance on Model Risk Management".
2. Department for Science, Innovation and Technology, 2023, "Capabilities and risks from frontier AI".
3. European Commission (EU), 2019, "Ethics Guidelines for Trustworthy AI", High-Level Expert Group on Artificial Intelligence.
4. Financial Stability Institute (FSI), 2021, "Human Keeping AI in check—emerging regulatory expectations in the financial sector", FSI Insights on policy implications No. 35.
5. International Organization of Securities Commissions (IOSCO), 2021, "The Use of Artificial Intelligence and Machine Learning by Market Intermediaries and Asset Managers".
6. Monetary Authority of Singapore (MAS), 2022, "Veritas Document 3: FEAT Principles Assessment Methodology".
7. Monetary Authority of Singapore (MAS), 2022, "Veritas Document 3A: FEAT Fairness Principles Assessment Methodology".
8. Monetary Authority of Singapore (MAS), 2022, "Veritas Document 3B - FEAT Ethics and Accountability Principles Assessment Methodology".
9. Monetary Authority of Singapore (MAS), 2022, "Veritas Document 3C - FEAT Transparency Principles Assessment Methodology".
10. National Cyber Security Centre (NCSC) et al., 2023, "Guidelines for secure AI system development".
11. National Institute of Standards and Technology (NIST), 2023, "AI Risk Management Framework (AI RMF 1.0)".
12. The Organization for Economic Co-operation and Development (OECD), 2021, "OECD Business and Finance Outlook 2021".
13. The Personal Data Protection Commission of Singapore (PDPC), 2020, "Model

Artificial Intelligence Governance Framework".

14. 銀行公會（研究單位：安永諮詢服務公司），2023，“人工智慧資通安全防護委外研究案”。

註釋

註 1： 外資集團在台分公司，如由集團提供 AI 應用，可依集團規範辦理。

註 2： 人工智慧相關定義參考自銀行公會「金融機構運用人工智慧技術作業規範」。

註 3： 自行研發除包含金融機構完全獨立開發外，亦包含與其他金融機構共同合作、聯合開發 AI 系統，以及以市場上既有之開源 AI 模型為基礎，進一步自行訓練或微調 (fine-tuning) 所研發之 AI 系統。

註 4： 個人原始資料係指未經去識別化、隱私強化技術處理或其他方式處理之個人資料。

註 5： 救濟選項 (remedies) 可能包括申訴或補救管道、爭議處理機制等。

註 6： 模型開發過程中版本資訊由金融機構依其重要性決定保存年限。

註 7： 所謂資料最小化原則係指蒐集個人資料必須適當、相關且於資料處理目的之必要範圍內。

註 8： 例如假名化、匿名化或採用隱私強化技術 (PETs) 等。

註 9： 韌性係指 AI 模型能適當地處理未在預期內的資料輸入或於信心水準過低時拒絕做出預測。