數字政策辦公室

人工智能道德框架

第 2.0 版

2025 年 12 月

© 中華人民共和國

香港特別行政區政府

修改記錄				
修改次數	修改詳情	經修改頁數	版本編號	日期
1	1. 在框架第 4.1.1.2 節加入關於「個人資料私隱專員公署」《開發及使用人工智能道德標準指引》的參考 2. 在框架第 4.1.4.4 節加入「中華人民共和國」《個人信息保護法》作為示例 3. 更新附錄 B 4. 對附錄 E 作出資料補充	4-7, 4-30, 6-4 至 6-6 附錄E	1.1	2023 年 6 月
2	更新附錄E的參考資料	附錄E	1.2	2023 年 7 月
3	更新附錄E的參考資料	附錄E	1.3	2023 年 8 月
4	將「政府資訊科技總監辦公室」更改為「數字政策辦公室」		1.4	2024 年 7 月
5	1. 更新第 1-5 章 2. 更新附錄 A-D 3. 刪減附錄 E	第1-5章 附錄A-D	2.0	2025 年 12 月

註：如本文件中英文版本內容存在歧義，請以英文版本為準，並通知數字政策辦公室。

目錄

1. 摘要 1-2

1.1 引言 1-2
1.2 何謂人工智能道德框架？ 1-3
1.3 如何運用《人工智能道德框架》？ 1-14

2. 目的 2-2

3. 人工智能道德框架概覽 3-2

3.1 「人工智能道德框架」願景 3-2
3.2 目標 3-2
3.3 效益 3-3
3.4 人工智能與數據道德引言 3-3
3.5 主要組成部分及關係 3-5
3.5.1 人工智能道德原則 3-8
3.5.2 人工智能治理 3-25
3.5.3 人工智能生命週期 3-32

4. 人工智能實務指引 4-2

4.1 人工智能生命週期及實務 4-2
4.1.1 項目策略 4-3
4.1.2 項目規劃 4-9
4.1.3 項目生態環境 4-15
4.1.4 項目開發 4-21
4.1.5 系統部署 4-48
4.1.6 系統運作與監察 4-58

5. 人工智能評估 5-2

5.1 人工智能應用影響評估 5-2
5.1.1 完成人工智能應用影響評估的流程 5-5
5.2 人工智能評估頻率 5-6
5.3 建議 5-8

6. 附錄 6-2

附錄A — 詞彙表 6-2

附錄B — 參考資料 6-4

附錄C — 人工智能應用影響評估範本 6-14

附錄D — 人工智能策略範本 6-42

第1章
摘要

1. 摘要

1.1 引言

人工智能與大數據分析具備提升社會福祉的潛力，現正日益廣泛應用於不同業務範疇，以改善運作效率及提供嶄新服務，但同時亦可能帶來各種挑戰。機構在推行資訊科技項目及提供服務時，必須充分考慮人工智能與數據道德因素。

機構在考慮應用人工智能及大數據分析技術時，必須全面評估多項因素，包括相關法例要求、持份者對數據及科技應用道德標準的期望，以及這些標準能否恰當反映本港社會的價值觀和文化。

《人工智能道德框架》文件包括：

- 一套專為推行資訊科技項目而**度身訂造的框架**，規範人工智能及大數據分析的道德應用；以及
- 一個適用於人工智能及大數據分析的評估範本（用於完成「**人工智能評估**」），用以評估人工智能應用或項目的潛在影響。

在本文件中，「人工智能」一詞指涉及大數據分析、進階分析及機器學習的運算操作，這些技術透過處理龐大數據集來找出關聯性及作出預測。**而「人工智能應用」則指一整套由人工智能模型驅動行動、決策或預測的應用。**人工智能應用的例子，包括具備預測功能和／或涉及訓練數據模型開發的資訊科技項目。機構在開展涉及人工智能應用的資訊科技項目時，可參考《人工智能道德框架》的要求。

採用《人工智能道德框架》可建立統一的方法和架構，以規管人工智能應用的開發和部署，從而基於以下指導原則，充分發揮人工智能在資訊科技項目中的應用效益：

- 協助機構了解人工智能及大數據分析在其業務範疇的應用；
- 與私隱保障、安全及數據管理等其他操作指引相互配合；

- 促進並引導機構在人工智能及大數據分析方面的道德應用；
- 協助機構考慮與其業務和技術範疇相關的人工智能道德要求，並對資訊科技項目的合規性進行監管和評估；以及
- 透過進行人工智能評估，協助機構識別和管理在資訊科技項目中採用人工智能及大數據分析的潛在風險。

1.2 何謂人工智能道德框架？

《人工智能道德框架》的最初版本旨在協助政府決策局 / 部門（局 / 部門）在規劃、設計及推行資訊科技項目和服務時，妥善應用人工智能及大數據分析技術。此框架包含指導原則、最佳實務方案及人工智能評估機制，供局 / 部門在開展人工智能驅動的資訊科技項目時採用。人工智能評估包括供局 / 部門在應用層面使用的評估範本，有助局 / 部門管理相關項目的效益、影響及風險。然而，此框架的指導原則、實務指引和評估範本也普遍適用於其他機構，此版本經過適當修改（例如刪除或修訂只適用於政府的特定用語），以供機構在採用人工智能和大數據分析時作參考之用。

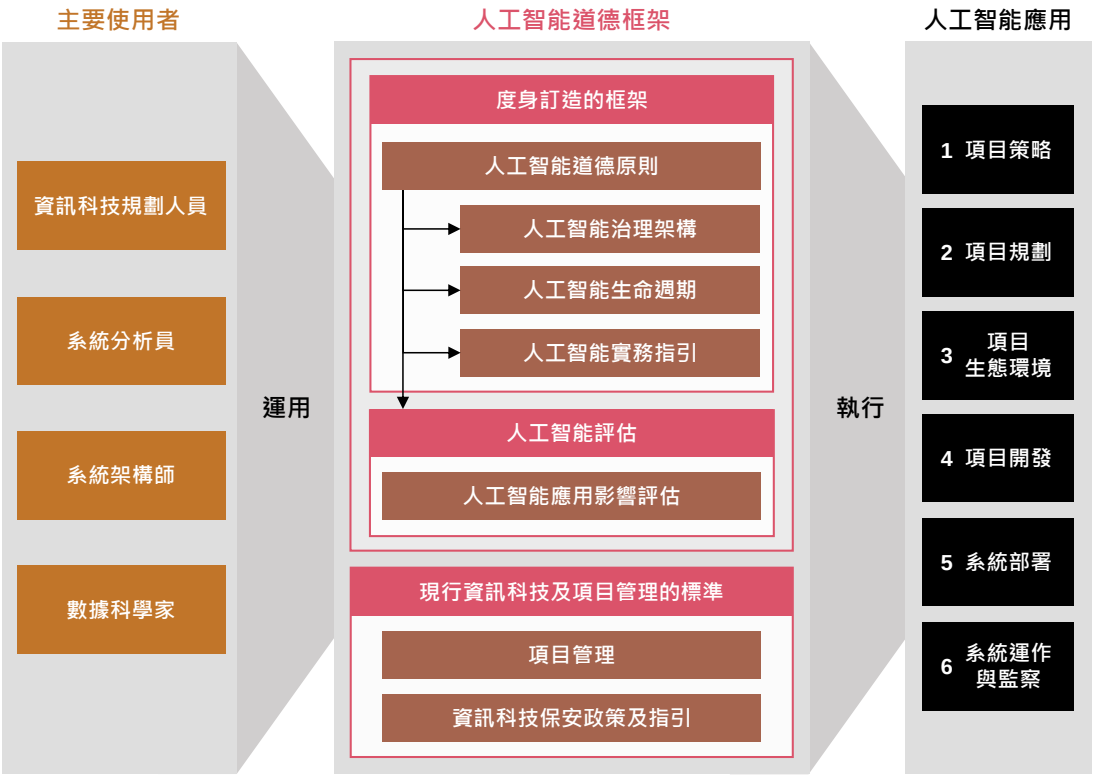

圖1：人工智能道德框架概覽

《人工智能道德框架》包括以下主要組成部分：

- ***度身訂造的框架：***提供一套「人工智能道德原則」、「人工智能治理架構」、「人工智能生命週期」及《人工智能實務指引》。詳情請參閱第3.5節「主要組成部分及關係」。
- ***人工智能評估：***包括供各機構使用的評估範本，用以評估人工智能應用的效益、影響及風險。「**人工智能應用影響評估**」用於審視人工智能應用，確保能妥善管理整個人工智能生命週期的影響，並已考慮相關的人工智能道德原則。詳情請參閱第5章「人工智能評估」。
- **現行標準及實務措施**（例如項目管理、資訊科技安全標準）應作為管理資訊科技項目的整體框架的一部分，而人工智能應用屬這些項目的子集。詳情請參閱附錄B「參考資料」。

人工智能道德原則

所有人工智能項目均應遵守十二項人工智能道德原則。其中兩項原則，(1) **透明及可解釋**及 (2) **可靠、穩健及安全**屬於「執行原則」。這些基本原則是其他原則得以執行的基礎。例如，若未能達致「可靠、穩健及安全」的原則，就無法準確核實其他人工智能道德原則是否一直被遵守。

其他原則歸類為「一般原則」，包括 (1) **公平**、(2) **多元與包容**、(3) **人為監督**、(4) **合法與合規**、(5) **數據私隱**、(6) **安全**、(7) **問責**、(8) **有益的人工智能**、(9) **合作與開放**，以及 (10) **可持續性與公正轉型**。這些原則源自聯合國《世界人權宣言》以及香港有關法例：

原則	定義
透明及可解釋	機構應確保能夠以清晰及易於理解的方式，向人解釋明白人工智能應用的決策過程。
可靠、穩健及安全	與其他資訊科技應用一樣，人工智能應用的開發應確保能夠透過使用合適的模型和數據集，在長時間內穩定可靠地運作。在確保運作穩健（即能提供一致的結果並能處理錯誤）的同時，必須按照相關法律和業界框架的要求，抵禦網絡攻擊，維持安全性。

原則	定義
公平	人工智能應用所作出的建議或產生的結果，應以公平方式對待相似群體中的每個人，不應偏袒或歧視，亦不應造成或引致傷害。
多元與包容	應通過了解和尊重所有受影響的持份者的利益，促進人工智能應用的包容性，並兼顧多元化的使用者需求。
人為監督	人工智能應用用於決策或操作過程中所需的人為干預程度，應取決於所涉及的道德問題的嚴重程度。
合法與合規	負責人工智能應用的機構，應貫徹依據法律、規例及相關監管制度行事。
數據私隱	個人應享有以下權利： (a) 獲告知個人資料的收集目的及可能的資料轉移對象，且該個人資料的收集須出於合法目的、以合法及公平的方式進行，並且所收集的資料數量不應超過該目的所需。請參閱《個人資料（私隱）條例》（《私隱條例》）[1]的**保障資料第1原則**「收集目的及方式」。 (b) 獲保證資料使用者已採取所有切實可行的步驟，以確保個人資料是準確無誤，並不保留超過所需的時間。請參閱《私隱條例》的**保障資料第2原則**「準確性及保留期間」。 (c) 要求個人資料僅用於該資料原本的收集目的及直接與該目的相關。否則須得到個人的明確及自願的同意。請參閱《私隱條例》的**保障資料第3原則**「個人資料的使用」。 (d) 獲保證資料使用者已採取所有切實可行的步驟，保障其持有的個人資料不會未經授權或意外地被查閱、處理、刪除、喪失或使用。請參閱《私隱條例》的**保障資料第4原則**「個人資料的保安」。 (e) 獲得相關資訊：(i) 個人資料方面的政策及實務；(ii) 所持有的個人資料的種類；(iii) 個人資料的主要使用目的。請參閱《私隱條例》的**保障資料第5原則**「資訊須在一般情況下可提供」。

[1] https://www.pcpd.org.hk/tc_chi/data_privacy_law/ordinance_at_a_Glance/ordinance.html

原則	定義
安全	人工智能應用在其整個運作生命週期內，均不應危及人類的身體安全或心理健全。
問責	機構須對其人工智能應用，包括其使用和誤用，所產生的道德影響負責。同時應有明確可識別的問責方，無論是個人或機構。
有益的人工智能	人工智能的發展應促進共同福祉。
合作與開放	應培養人工智能生態環境中多方持份者開放合作的文化。
可持續性與公正轉型	人工智能的發展應確保已制定緩解策略，以應對任何對社會及環境系統的潛在影響。

表1：人工智能道德原則及定義

人工智能治理

人工智能治理是指管理及控制人工智能項目與應用的實踐方針。「三道防線」是眾多機構廣泛採用的成熟治理概念，圖2展示了不同防線及其角色。

圖2：三道防線模式

本治理架構的組成如下：

- **第一道防線**為項目團隊，負責開發人工智能應用、風險評估、執行風險緩解措施，以及記錄「人工智能評估」工作。
- **第二道防線**由項目督導委員會及項目保證小組組成，確保項目品質，制定人工智能應用的驗收標準，進行獨立檢視，並批准人工智能應用。在予以批准前，須完成「人工智能評估」以確保符合「人工智能道德原則」。
- **第三道防線**涉及資訊科技委員會，如未設立此委員會，則由首席資訊官擔任，並可選擇由人工智能道德委員會提供支援。人工智能道德委員會可由外部顧問組成，就人工智能道德議題提供建議，並加強機構採用人工智能的能力。第三道防線負責審核、建議及監管高風險人工智能應用。

人工智能生命週期

為使機構在開展人工智能項目或開發人工智能應用時有所依循，本框架已於「人工智能實務指引」中詳細說明人工智能生命週期各階段的實務規範（詳情請參閱《人工智能道德框架》第4章「人工智能實務指引」）。下列六步流程圖解構人工智能生命週期的概念：

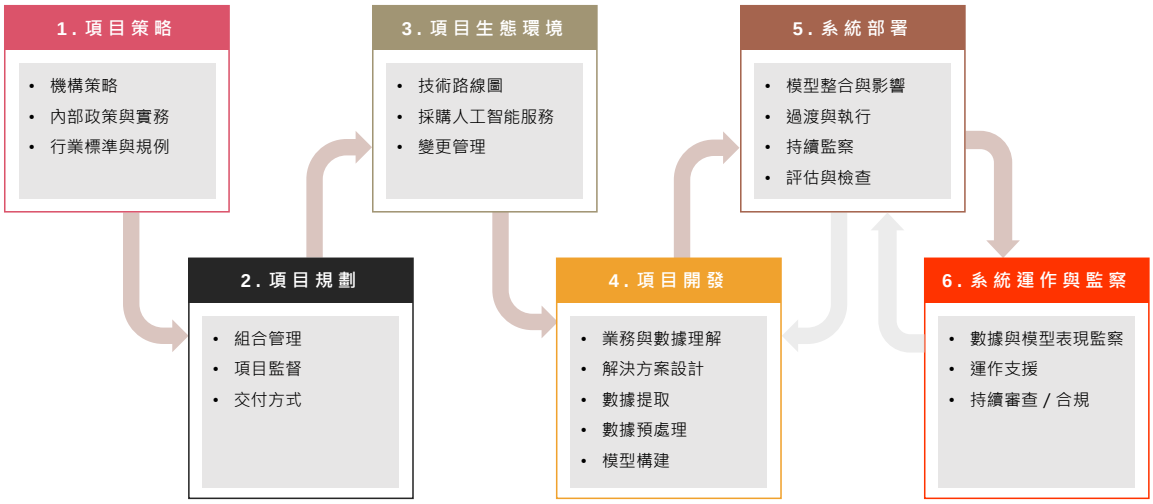

圖3：人工智能生命週期概覽

人工智能生命週期展示了人工智能項目所涉及的各個階段，當中包括：

- 引導機構了解人工智能項目的不同階段及相關要求；以及
- 作為制定人工智能實務的參照，以配合典型的人工智能實際開發階段。

人工智能生命週期不僅適用於訂定《人工智能實務指引》的實務規範，同時與傳統的系統發展生命週期模型互相對應（見圖4）。

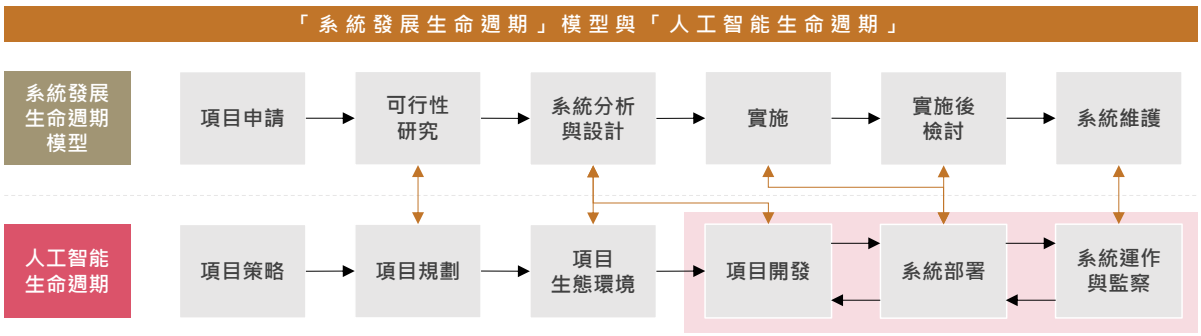

圖4：「人工智能生命週期」與「系統發展生命週期」模型互相對應

在人工智能應用的典型開發過程中，數據準備工作至關重要，因為優質數據往往是構建良好人工智能模型的基礎。數據收集與預先處理作為項目開發的組成部分，往往需要持續進行，這是因為在整個開發過程中，人工智能模型可通過獲取更優質或更大量的數據來反覆優化訓練效果。傳統軟件生命週期的開發模式，是通過指令集編寫資訊科技應用程式來處理既定情境，其後這些資訊科技應用程式會運用運算能力與其他資源來處理輸入系統的數據。至於人工智能應用的開發模式則截然不同，是通過向系統輸入海量數據進行訓練，在所有數據經處理後，最終生成能解決新問題的人工智能模型或解決方案。

在人工智能生命週期中，開發與部署階段、系統運作與監察之間往往存在持續的反饋循環，以便不斷改進，這有別於傳統的系統發展生命週期。

《人工智能實務指引》

第4章「人工智能實務指引」詳細列明多個實務範疇須遵循的實務規範，這些實務範疇屬「人工智能應用影響評估」評核內容的一部分。以下概要說明《人工智能實務指引》所涵蓋的實務範疇：

人工智能生命週期	實務範疇	定義
項目策略	**機構的策略、內部政策與實務**	機構應能以清晰易懂的方式，向人解釋人工智能應用的決策過程。

人工智能 生命週期	實務範疇	定義
	行業標準與規例	須就相關規例及標準進行評估，以確保人工智能及相關流程符合任何適用的法律或標準。
項目規劃	**組合管理**	進行組合管理，旨在確保機構投放於流程、人員及技術的資訊科技投資符合預期。評估人工智能項目，以確保項目有效滿足業務需求及目標。
	項目監督與交付方式	品質控制不僅監察交付成果的品質，還涉及按「項目管理計劃」所界定的各個環節進行監察。
項目生態環境	**人工智能與數據使用的技術路線圖**	技術路線圖應使機構能夠規劃及制定策略，以決定採購什麼、何時採購及採購哪種技術，以進行人工智能及大數據分析。
	採購人工智能服務	可為人工智能項目採購現成的產品及數據。在進行此類採購工作時，機構應適當考慮相關道德因素。
項目開發	**業務與數據理解**	機構應確定使用人工智能的目標，並在決策過程中權衡使用人工智能的效益與風險。
	解決方案設計	機構應評估人工智能應用中的人工智能模型是否切合機構的目標，以及所需的人為介入程度。
	數據提取	機構應確保來自內部或外部不同來源數據的品質、有效性、可靠性及一致性。
	數據預處理	在解決方案開發過程中，須特別謹慎處理含有個人資料的敏感數據，以防出現數據洩漏或違反私隱與安全規定。

人工智能 生命週期	實務範疇	定義
	模型構建	模型構建應旨在減少常見的人工智能應用錯誤，例如不準確的模型假設、模型過度擬合及對抗性攻擊。
系統部署	**模型整合與影響**	核實、驗證及測試過程旨在確保人工智能應用的運作，符合項目初期訂定的要求。這個過程可確保人工智能應用的妥善整合。
	過渡與執行	假定人工智能應用或會出現故障，在部署人工智能應用前，應制定緩解措施，以期盡量減少故障所造成的損害。
	持續監察	提供反饋以監察 / 維護模型的表現及人工智能應用的穩健性。
	評估與檢查	人工智能應用需要具可追溯性、可重複性及可再現性，以確保人工智能應用運作正常，並協助建立公眾及持份者的信任。
系統運作與監察	**數據與模型表現監察**	由於人工智能模型（作為人工智能應用的一部分）有機會變得不夠準確及不太相關，因此應持續進行監察及檢討。
	運作支援	在部署人工智能應用後，應建立持續的運作支援，以確保人工智能應用的表現保持穩定、可靠及穩健。
	持續審查 / 合規	項目經理（或維護團隊）、項目督導委員會 / 項目保證小組（或維護委員會）及資訊科技委員會 / 首席資訊官（或其代表）負責監察人工智能風險，例如不符合相關新 / 修訂法律及規例的情況。

表2：人工智能道德框架內《人工智能實務指引》所涵蓋的實務範疇

人工智能應用影響評估

人工智能應用影響評估應在人工智能生命週期的不同階段進行。該評估為各機構引入系統化思考流程，以便全面審視個別應用的各項考量因素，包括相關效益與風險，同時突顯有額外管治活動的需要，並制定跟進行動以確保實施符合「人工智能道德」的必要措施與管制。

本評估所用的人工智能應用影響評估範本詳見於《人工智能道德框架》附錄C「人工智能應用影響評估範本」，以作參考。

人工智能應用影響評估包含以下組成部分：

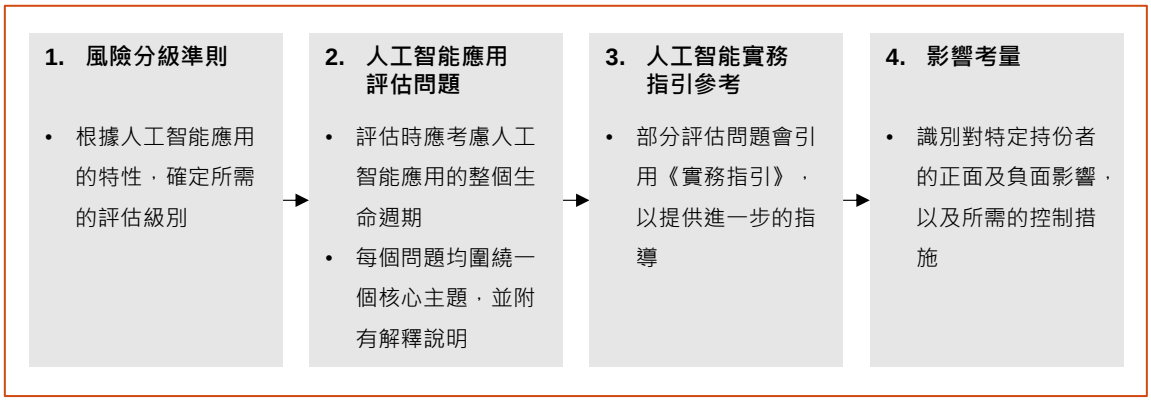

圖5：人工智能應用影響評估的組成部分

應視乎人工智能項目進度及在人工智能應用運作期間，定期（例如每年）或在發生重大變更時進行人工智能應用影響評估。

圖6顯示在人工智能生命週期中應檢視人工智能應用影響評估的各個階段。

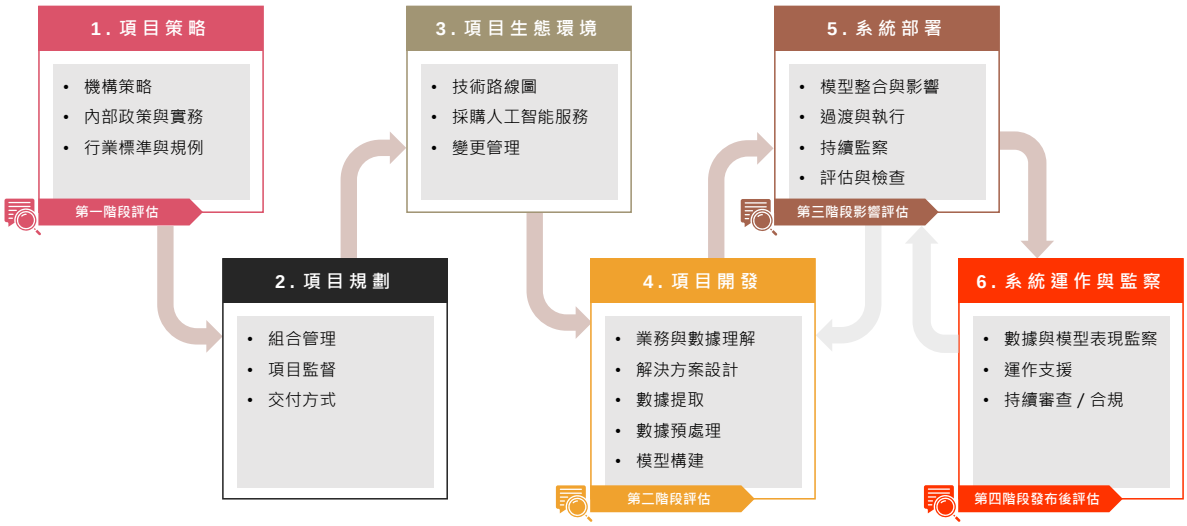

圖6：進行人工智能應用影響評估的階段

人工智能應用影響評估可在整個人工智能生命週期內作為「動態」文件使用，但應在人工智能生命週期的4個關鍵階段進行檢視，並保留評估記錄作歷史存檔。下表總結了這些檢視與系統發展生命週期的對應關係，以及負責方和須執行的行動：

生命週期階段	負責方	執行步驟
系統發展生命週期： 項目申請 **人工智能生命週期：** 項目策略	項目團隊	根據「風險分級問題」的答案，將項目分類為「高風險」或「非高風險」 進行「人工智能應用影響評估」： • 回答問題1-8、9(i)、41-49、57-58
	項目督導委員會 / 項目保證小組	檢視及核准非高風險項目的評估
	資訊科技委員會 / 首席資訊官	檢視及核准高風險項目的評估

生命週期階段	負責方	執行步驟
系統發展生命週期： 系統分析與設計 **人工智能生命週期：** **項目生態環境** 系統開發	項目團隊	進行「人工智能應用影響評估」： • 回答問題9(ii)、10-14、18-19、22-29、30(i)(ii)、31-34、37(i)、52-56 • 根據最新情況檢視及更新問題1-8、9(i)、41-49、57-58的答案 • 如使用第三方技術或數據，須在採購前完成並檢視問題15-17、48
	項目督導委員會／項目保證小組	• 檢視及核准評估報告
系統發展生命週期： 實施階段（推出前） **人工智能生命週期：** 系統部署	項目團隊	進行「人工智能應用影響評估」： • 回答問題20、21、30(iii)、35、36、37(ii)(iii)、38-40、50、51 • 根據最新情況檢視及更新評估（問題1-14、18-19、22-29、30(i)(ii)、31-34、37(i)、41-49、52-58） • 如使用第三方技術或數據，須檢視問題15-17、48
	項目督導委員會／項目保證小組	• 檢視及核准評估報告
系統發展生命週期： 系統維護 （每年或發生重大變更時） **人工智能生命週期：** 系統運作與監察	維護團隊	• 根據最新情況檢視及更新「人工智能應用影響評估」 • 呈報任何重大問題
	維護委員會 資訊科技委員會／首席資訊官（或其授權代表）	• 處理呈報事項 • 監察高風險項目

表3：完成及檢視人工智能應用影響評估的執行步驟

1.3 如何運用《人工智能道德框架》？

機構在資訊科技項目或服務中採用人工智能及大數據分析時，可善用《人工智能道德框架》。此框架不僅為資訊科技項目團隊在開發及維護過程中提供參考 / 指引，更通過確立治理架構，使機構在採用人工智能時能夠評估影響、確保安全且負責任地使用人工智能，以及促進創新，從而展現問責精神，並建立信任。

此外，資訊科技規劃人員及管理層可參照《人工智能道德框架》，從策略制訂、規劃以至建立生態環境的初期，便把適切的人工智能道德納入考量。有關不同角色對應的框架內容詳情，請參閱第3.5節「主要組成部分及關係」。

第2章

目的

2. 目的

本文件旨在讓讀者了解「人工智能道德框架」及相關程序，以便將道德元素融入機構在資訊科技項目或服務中規劃、設計及實施人工智能應用的過程。本文件主要分為下列部分：

- **第1章 — 摘要：**概述「人工智能道德框架」的組成部分、使用方式及治理架構；
- **第2章 — 目的：**闡明本文件各部的目標；
- **第3章 — 人工智能道德框架概覽：**介紹「人工智能道德框架」、「人工智能道德原則」、願景聲明、角色和責任，以及目標；
- **第4章 — 人工智能實務指引：**為機構在採用道德人工智能的過程中，提供參考「人工智能道德原則」的實務指引；以及
- **第5章 — 人工智能評估：**提供「人工智能應用影響評估」的範本，列明機構需要填寫的道德人工智能相關事項及考慮因素。

下表載列本文件的適用對象及建議參閱部分：

適用對象	建議參閱章節
機構管理層	第1章 — 摘要
首席資訊官、 **資訊科技規劃人員 / 資訊科技委員會、** **項目督導委員會、** **項目保證小組，以及** **業務使用者**	第1章 — 摘要 第3章 — 人工智能道德框架概覽 第4.1.1節 — 項目策略 第4.1.2節 — 項目規劃 第4.1.3節 — 項目生態環境 第5章 — 人工智能評估
項目經理	全部章節
項目團隊（包括系統分析員、系統架構師及數據科學家）	第4.1.4節 — 項目開發 第4.1.5節 — 系統部署 第4.1.6節 — 系統運作與監察 第5章 — 人工智能評估

第3章

人工智能道德框架概覽

3. 人工智能道德框架概覽

3.1 「人工智能道德框架」願景

「人工智能道德框架」的願景是讓機構能透過評估其人工智能技術能力及應用，妥善管理相關道德議題及潛在影響，從而在管理人工智能應用潛在影響的前提下，實現道德人工智能的應用。

本章旨在為機構概述「人工智能道德框架」的整體架構。文中所引用的例子來自不同來源，僅供說明用途。

3.2 目標

「人工智能道德框架」的目標如下：

- 支援機構理解人工智能道德及其應用，方法包括：
  - 制訂具明確定義、實踐方法及指引的「人工智能道德原則」；以及
  - 釐清機構在實施道德人工智能時的角色和責任。
- 引導機構以合乎道德的方式使用人工智能，方法包括：
  - 說明人工智能生命週期各階段涵蓋的活動，以及應用合乎道德的人工智能所需的相應能力；以及
  - 提供實務指引供機構在資訊科技項目引入人工智能時遵行，確保符合道德標準。
- 協助機構治理人工智能應用的合規性，方法包括：
  - 提供「人工智能應用影響評估」範本，讓機構能基於實施道德人工智能的實際考量因素，就其人工智能應用進行評估。

3.3 效益

採用「人工智能道德框架」是一項基礎步驟，旨在建立共同的方法和架構，以治理後續人工智能應用的開發和部署。制定「人工智能道德框架」的好處包括：

- 確立共同的最佳實務，確保機構在資訊科技項目中採用人工智能時，能獲得指引和參考依據，以便作出適當的道德考量；
- 識別人工智能應用的效益、風險和影響，從而作出更佳的風險緩解決策，以取得最大效益；以及
- 作為策略與執行之間的橋樑，有助確保人工智能應用符合機構的願景和需求。

3.4 人工智能與數據道德引言

在大數據時代，海量數據正不斷產生、收集和分析，以識別洞見並支援決策制定。大數據通常以「五V」特性來描述，包括：

- 數據量（Volume）：指可用數據的龐大數量；
- 數據速度（Velocity）：指數據必須被儲存及 / 或分析的速度，以在適當時間提供正確資訊，從而作出合適的管理決策；
- 數據多樣性（Variety）：指數據類型和來源的極大差異，包括結構化和非結構化數據（例如檔案物件、社交媒體動態、標籤、傳感器數據、語音、圖像及視像）；
- 數據真實性（Veracity）：指數據的可信度，涉及數據的準確性和品質；以及
- 數據價值（Value）：指將數據轉化以改善結果或價值的能力。

在大數據分析領域所採用的技術方法，通常統稱為「進階分析技術」、「機器學習」及「人工智能」。這些技術術語概念相近且相互關聯，均指透過運用大量描述過去情境的數據（即歷史數據）、強大運算能力及進階算法，從中找出數據關聯模式並進行具備合理準確度的預測分析。

廣泛而言，人工智能是一個統稱，泛指能夠感知環境、思考、學習並根據所收集的數據採取行動的電腦系統，其最終目的是實現設計目標。人工智能系統是一系列相互關聯的技

術，用於自主協助解決問題及執行任務以達成既定目標，而無需人類明確指導。我們可將人工智能區分為四大類別（見圖7）：

	「人在環內」/ 人為管控	「人在環外」
固定 / 特定系統	**輔助式人工智能** 協助人類作出決策或採取行動的人工智能系統。此類系統為固定程式設計，不會從互動中學習。	**自動化人工智能** 對常規手動及認知任務進行自動化處理。此類系統僅將現有任務自動化，並未引入新的運作方式。
自適應系統	**增強式人工智能** 強化人類決策能力的人工智能系統，並持續從與人類及環境的互動中學習。	**自主式人工智能** 能適應不同情境且無需人類協助即可自主行動的人工智能系統

圖7：人工智能的分類[2]

- **輔助式人工智能：**此類別的人工智能應用旨在協助人類進行決策或採取行動。該等應用未必會從與環境的互動中學習，且最終決策者仍為人類。其目標是協助人類更快、更佳地完成認知任務，從而提升人類及機構現有工作的效能。其中一個簡單例子是現時常見於汽車的全球定位系統導航程式，該程式能根據路況為駕駛者提供路線調整建議。
- **增強式人工智能：**此類別的人工智能應用可增強人類決策能力，並持續從與人類及其運作環境的互動中學習。該等應用不會自主作出決策，其目標是協助人類作出更佳決策，使人類及機構能夠完成原本無法達成的事項。
- **自動化人工智能：**此類別指對常規 / 非常規的手動或認知任務進行自動化處理，但並未引入新的運作方式，因其目標僅為將現有任務自動化。例如，用於自動化後台流程

[2] https://www.pwc.no/no/publikasjoner/Digitalisering/pwc-ai-analysis-sizing-the-prize-report.pdf

的機械人流程自動化應用，便屬於自動化人工智能。此類應用的關鍵特徵為「無需人為介入」，意味系統能以完全自主的方式作出決策。

- **自主式人工智能：**此類人工智能應用的目標是在無需人類干預的情況下把決策過程自動化。為實現此目標，該等應用會從與人類及環境的互動中學習。自主式人工智能的實例包括自動駕駛車輛、機械人、聊天機械人及虛擬助理。

以下為人工智能道德相關關鍵詞彙的定義：

- **人工智能道德：**指在使用、開發、實施人工智能或評估其成果時，對相關道德影響的考量。人工智能道德可視為數據道德的一個子集。
- **人工智能道德原則：**一套與人工智能使用、開發、實施或成果相關的行為準則。
- **數據道德：**倫理學的分支，專門研究並評估關乎數據（包括生成、記錄、篩選整理、處理、傳播、分享及使用）、算法（包括人工智能／機器學習模型、智能代理及機械人）及相關實踐（包括負責任的創新、編程、駭客行為及專業守則）的道德問題，旨在制定並支持符合道德的良好解決方案（例如正確行為或價值觀）。

3.5 主要組成部分及關係

圖8展示了「人工智能道德框架」的主要組成部分，各部分的詳細說明載於後續章節（即第4章「人工智能實務指引」及第5章「人工智能評估」）。在使用本框架時，須同時參照現行資訊科技及項目管理的標準與實務指引。

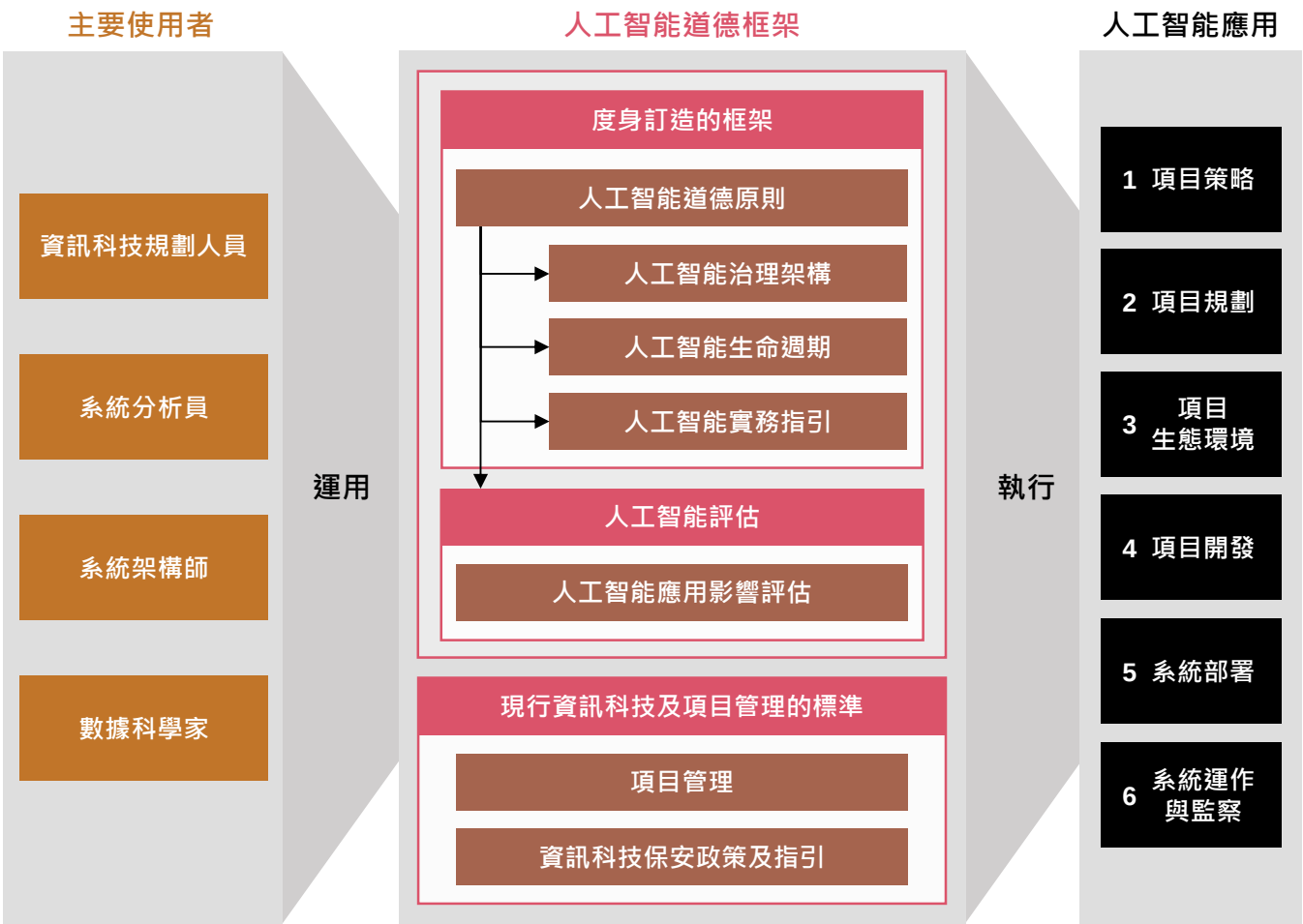

圖8：人工智能道德框架的主要組成部分

「人工智能道德框架」包括以下主要組成部分：

- **度身訂造的框架**，包含以下子部分：
  - **人工智能道德原則：**界定設計及開發人工智能應用時須遵循的原則。該原則適用於人工智能治理架構中定義的所有角色；
  - **人工智能治理架構：**根據「人工智能道德框架」所訂立的實務，界定採用人工智能過程中的標準架構、角色和責任；
  - **人工智能生命週期：**界定一個人工智能生命週期模型，用於構建人工智能實務指引的布局及人工智能評估中的問題；以及
  - **人工智能實務指引：**為人工智能生命週期的所有階段界定一套實務。這些指導性實務源自「人工智能道德原則」。

- **人工智能評估：**包括「人工智能應用影響評估」範本，該範本界定了目標讀者在人工智能生命週期各階段須回答的問題，以評估人工智能應用或項目的影響，並確保已考慮「人工智能道德原則」。

「人工智能道德框架」的目標讀者包括資訊科技規劃人員、系統分析員、系統架構師及數據科學家。數據科學家在此框架下的角色涵蓋開發、部署及監測人工智能的相關工作，而根據機構的個別架構設置，這些工作範疇亦可分拆為不同角色。目標讀者應運用「人工智能道德框架」執行以下工作：

- 理解「人工智能道德原則」及實務；
- 就人工智能的影響展開討論；
- 採用標準化實務及術語；以及
- 進行人工智能評估。

資訊科技規劃人員及管理層

- ***項目策略***

  資訊科技規劃人員及管理層可參閱「人工智能實務指引」第4.1.1節「項目策略」，以制定機構的人工智能 / 數據策略，並確保納入「人工智能道德原則」及考慮相關規例及條例。

- ***項目規劃***

  機構的資訊科技規劃人員可參閱「人工智能實務指引」第4.1.2節「項目規劃」，配合「人工智能應用影響評估」中的風險分級準則，確保符合人工智能道德要求、有效評估影響，並釐定哪些人工智能項目須提交高層覆核。過程中須考慮人工智能相關的道德規範、角色和責任。

- ***項目生態環境***

  資訊科技規劃人員在評估現有技術環境、業務需求及採購方案以識別技術缺口時，可參閱「人工智能實務指引」第4.1.3節「項目生態環境」。負責與採購團隊及項目團隊協作的資訊科技規劃人員，於部署第三方人工智能應用時應參考該節所列的考量事項。修改現有系統 / 人工智能應用時，須遵循既定的變更管理程序。

系統分析員、系統架構師及數據科學家

- ***項目開發***

  負責開發數據管道（將數據轉換並輸送至人工智能應用）的系統分析員與系統架構師，以及負責開發人工智能應用的數據科學家，可參閱「人工智能實務指引」第4.1.4節「項目開發」，以確保在開發合乎道德的人工智能項目時考慮數據驗證、文件編製、偏頗數據的處理、數據私隱、訓練、測試及人工智能建模技術等範疇。

- ***系統部署***

  負責管理人工智能應用整合、擴展及部署，以及管理部署後性能與穩定性的數據科學家，可參閱「人工智能實務指引」第4.1.5節「系統部署」，了解與人工智能應用部署相關的特定範疇，例如整合、測試、反饋循環、調整指標及性能檢查。

- ***系統運作與監察***

  負責監察人工智能系統，並確保就人工智能應用的反饋採取適當行動的數據科學家，可參閱「人工智能實務指引」第4.1.6節「系統運作與監察」，以確保已考慮所需行動，包括呈報事項、持續檢視及對內部要求與外部期望的合規檢查等。

3.5.1 人工智能道德原則

所有人工智能項目均應遵守十二項人工智能道德原則。其中兩項原則，(1) **透明及可解釋**及 (2) **可靠、穩健及安全**屬於「執行原則」。這些基本原則是其他原則得以執行的基礎。

例如，若未能達致「可靠、穩健及安全」的原則，就無法準確核實其他人工智能道德原則是否一直被遵守。

其他原則歸類為「一般原則」，包括 (1) **公平**、(2) **多元與包容**、(3) **人為監督**、(4) **合法與合規**、(5) **數據私隱**、(6) **安全**、(7) **問責**、(8) **有益的人工智能**、(9) **合作與開放**，以及 (10) **可持續性與公正轉型**。這些原則源自聯合國《世界人權宣言》以及香港有關法例。

各項原則的定義列於表4，後續分節將提供進一步詳情。

原則	定義
透明及可解釋	機構應確保能夠以清晰及易於理解的方式，向人解釋明白人工智能應用的決策過程。
可靠、穩健及安全	與其他資訊科技應用一樣，人工智能應用的開發應確保能夠透過使用合適的模型和數據集，在長時間內穩定可靠地運作。在確保運作穩健（即能提供一致的結果並能處理錯誤）的同時，必須按照相關法律和業界框架的要求，抵禦網絡攻擊，維持安全性。
公平	人工智能應用所作出的建議或產生的結果，應以公平方式對待相似群體中的每個人，不應偏袒或歧視，亦不應造成或引致傷害。
多元與包容	應通過了解和尊重所有受影響的持份者的利益，促進人工智能應用的包容性，並兼顧多元化的使用者需求。
人為監督	人工智能應用用於決策或操作過程中所需的人為干預程度，應取決於所涉及的道德問題的嚴重程度。
合法與合規	負責人工智能應用的機構，應貫徹依據法律、規例及相關監管制度行事。

原則	定義
數據私隱	個人應享有以下權利： (a) 獲告知個人資料的收集目的及可能的資料轉移對象，且該個人資料的收集須出於合法目的、以合法及公平的方式進行，並且所收集的資料數量不應超過該目的所需。請參閱《個人資料（私隱）條例》（《私隱條例》）[3]的**保障資料第1原則**「收集目的及方式」。 (b) 獲保證資料使用者已採取所有切實可行的步驟，以確保個人資料是準確無誤，並不保留超過所需的時間。請參閱《私隱條例》的**保障資料第2原則**「準確性及保留期間」。 (c) 要求個人資料僅用於該資料原本的收集目的及直接與該目的相關。否則須得到個人的明確及自願的同意。請參閱《私隱條例》的**保障資料第3原則**「個人資料的使用」。 (d) 獲保證資料使用者已採取所有切實可行的步驟，保障其持有的個人資料不會未經授權或意外地被查閱、處理、刪除、喪失或使用。請參閱《私隱條例》的**保障資料第4原則**「個人資料的保安」。 (e) 獲得相關資訊：(i) 個人資料方面的政策及實務；(ii) 所持有的個人資料的種類；(iii) 個人資料的主要使用目的。請參閱《私隱條例》的**保障資料第5原則**「資訊須在一般情況下可提供」。
安全	人工智能應用在其整個運作生命週期內，均不應危及人類的身體安全或心理健全。
問責	機構須對其人工智能應用，包括其使用和誤用，所產生的道德影響負責。同時應有明確可識別的問責方，無論是個人或機構。
有益的人工智能	人工智能的發展應促進共同福祉。
合作與開放	應培養人工智能生態環境中多方持份者開放合作的文化。

[3] https://www.pcpd.org.hk/tc_chi/data_privacy_law/ordinance_at_a_Glance/ordinance.html

原則	定義
可持續性與公正轉型	人工智能的發展應確保已制定緩解策略，以應對任何對社會及環境系統的潛在影響。

表4：人工智能道德原則及定義

3.5.1.1 透明及可解釋

原則：機構應確保能夠以清晰及易於理解的方式，向人解釋明白人工智能應用的決策過程。

可解釋性

可解釋性是指對人工智能模型的特點（例如特徵和參數），以及每個個別預測的決策路徑，創造人類可理解的解釋。這種解釋能力與人工智能各方持份者可能息息相關，並且應該不斷改進。相關領域包括理解人工智能，確保模型具備可解釋性、過程的透明度、可重現性及可信性。

無論是全球或本地的文獻，均認同有需要以清晰的方式解釋人工智能的實際運作。為創造穩健且可信的人工智能應用，理解其模型的內部運作機制，以及訓練和決策過程中所使用的數據極為重要。

機構應該關注人工智能模型的內部運作，以評估使用人工智能解決方案的可信程度，以建立可信性，並作出更有信心的決策；數據科學家應小心考慮有關模型是否已經準備好可作部署，或是否有需要改善與工作夥伴、系統分析員、系統架構師和管理人員的溝通和合作，以加強可解釋性；治理機構可能需要全面且無法篡改的報告，以作審計用途。

無論如何，這都需要道德原則規範人工智能開發者，考慮為人工智能應用加入可解釋性。解釋方法可根據實際應用情況和持份者的需要，從兩個不同角度，生成文字或圖像為模型解釋，詳情請參閱第4.1.4.1節「業務與數據理解」。對於較複雜的機器學習

算法（例如神經網絡），其算法的超參數（hyper-parameters）均應妥善記錄。例如，在涉及使用神經網絡的人工智能應用，應妥善記錄其層數量、激活函數、誤差函數等超參數。

根據在解釋時能否取得人工智能模型的數據、算法及配置，可以採用兩種方式生成人類可讀的解釋：

- *內建解釋（Built-in Explanation）*：某些模型本身具備解釋其行為的能力。例如，決策樹本質上是一連串的判斷問題，其運作方式與人類思考方式相近。 當模型針對某個數據點（即以其特徵和數值）執行判斷時，達致決策的路徑可簡單地回報人類使用者。同樣地，像邏輯回歸這類線性模型亦相對可以靠直覺理解，即使非數據科學專家亦較易明白，因為這些模型的數據點可對應圖表上與學習所得的機率線，並可計算特徵重要性（Feature Importance）指標，將其絕對值回報給數據科學家作進一步分析。

- *事後解釋（Post-hoc Explanation）*：事後解釋是指在人工智能模型完成訓練後，或當模型已部署並僅可透過應用程式介面存取時，為理解其行為而採取的進一步行動。事後解釋超越傳統的探索性數據分析技術，雖然它較難確切闡釋模型的內部運作機制，但可透過舉例說明、代理模型、繪製特徵重要及部分依賴圖（Partial Dependence Plots）等方式，幫助服務使用者概略了解模型的行為。視乎接收解釋的持份者身份，亦可使用自然語言生成技術，為模型的行為生成自然語言解釋。

需要注意的是，機器學習算法的可解釋性與準確性之間在本質上存在着取捨關係。非參數式機器學習算法，或具備大量參數的模型（如深度神經網絡或集成模型）能夠從大規模數據中學習複雜且隱含的模式，並提供靈活及卓越的性能，代價是其本質比較「不透明」。相對地，參數式模型如樸素貝葉斯（Naive Bayesian）或簡單神經網絡，都能以一套固定的參數訓練，比較容易解釋，但其性能一般較低，儘管情況並非一成不變。除了通常決定學習模型的訓練數據種類、規模和複雜程度，在設計時亦應該考慮人工智能應用的安全關鍵性質及所需的解釋程度。

透明度

透明度無疑是建立對人工智能模型信任中最關鍵的特質。信任是需要時間逐步建立和鞏固的，並透過審慎設計的流程來實現，以將風險降至最低，因此在推動人工智能等顛覆性新技術的廣泛應用上，扮演關鍵角色。信任得來不易，並建基於多項因素，包括人工智能應用的目的和性能，以及技術提供者（例如機構的品牌聲譽、設計與運作的透明度水平、可靠性，以及能否說明人工智能模型決策背後的理據等，均對建立和維繫信任至關重要）。「透明度」原則強調採取清晰、誠實的渠道與各相關的持份者溝通。

人工智能應用的隨機性特質，使其與傳統的軟件開發模式有着根本性的差異。以下我們將根據電氣電子工程師學會（Institute of Electrical and Electronics Engineers，IEEE）標準，逐一說明將透明度融入人工智能應用設計的可行方式：

- *透明度作為可追溯性（Traceability）：*指軟件工程在實施過程中的透明度，應允許技術檢查哪些規範已明確或隱含地在什麼情境下實施，以及如何解決可能發生的衝突。
- *透明度作為可驗證性（Verifiability）：*這關乎於系統實施過程中如何處理規範性推理，以及如何核實系統所作出的規範性決策是否符合既定的規範和價值觀，透過明確且確切的表示方式，而其目的是為一系列強而有力的數學技術（例如形式核實）提供基礎。可驗證性在系統故障時同樣重要，例如當人工智能應用造成損害時，應能夠確定原因；若司法決策涉及人工智能模型，則必須能夠提供一套可經由具備相關能力的人類主管機構審核的圓滿解釋[4]。
- *透明度作為非欺騙性與誠實設計（Non-deception and Honest Design）：*全球人工智能社群的所有持份者均有一個共同而強烈的利益關切，就是避免人工智能應用於自動生成任何形式（包括文字、語音、視像）的資訊或內容時，蓄意呈現與真實世界部分或完全不符的狀況，並以誤導第三方（不論是人類或機器）為目的。當然，在某些特定的人工智能應用情境中，欺騙行為可能屬於系統核心功能的一部分（例如

[4] https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf

遊戲、娛樂或藝術創作），而這類行為在具備明確使用情境與獲得使用者同意的前提下，則不構成違反規範[5]。

- *透明度作為個人私隱（Personal Privacy）*：個人應有權查閱及更正其可被人工智能應用使用的個人數據。

對於「透明度」的常見誤解，是認為機構必須公開其與人工智能相關的知識產權（例如源代碼、模型設定、完整數據集等），才能實現「透明的人工智能」。相反地，「透明度」原則所強調的，是機構與服務使用者，以及在有需要時與監管機構之間，應建立清晰而誠實的溝通渠道，並視為不可或缺的一環。否則，考慮到不少人工智能應用屬於營利性質，或涉及第三方人工智能供應商的商業模式，機構將難以遵循該原則。事實上，若向人工智能應用服務使用者披露過多細節，反而可能導致系統容易被惡意使用者「鑽空子」。與此同時，審計人員及監管機構或有需要獲取準確且全面的機構人工智能驅動流程資訊。在這類情況下，可由有關機構按實際情況自行決定合宜的透明度程度。

可證明性

可證明性（Provability）是指人工智能模型決策背後的數學確定性，其重點在於要求以更高程度形式化的方式解釋人工智能應用的行為。這種解釋方式主要面向數據科學家，讓其可對人工智能模型進行嚴格檢視，以確保模型的決策邏輯能以數學方式加以證明，並能在數據或環境發生變化時仍保持一致性。同樣地，就着許多屬於安全關鍵的人工智能應用而言，由於人工智能必須獲得監管機構的認可或認證，可證明性更是不可或缺的要素。

例子（更多實例請參閱第4章「人工智能實務指引」）：

- 項目團隊應致力於將透明度元素融入人工智能應用中，並以人類語言表達（例如透過決策樹或圖像方式向使用者解釋決策邏輯），以協助人類理解人工智能所作出的決策是否出錯，並加以修正。

[5] https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf

在可行情況下，應以敘述方式或圖像（如繪製包含決策樹的工作流程圖）說明人工智能的決策過程，方便不具備技術背景的受眾更清晰地理解與設想人工智能的運作模式。機構應清楚說明不同因素與數據如何影響人工智能模型得出的結果與結論。

3.5.1.2 可靠、穩健及安全

原則：與其他資訊科技應用一樣，人工智能應用的開發應確保能夠透過使用合適的模型和數據集，在長時間內穩定可靠地運作。在確保運作穩健（即能提供一致的結果並能處理錯誤）的同時，必須按照相關法律和業界框架的要求，抵禦網絡攻擊，維持安全性。

「可靠、穩健及安全」原則的首要目標，是確保人工智能應用由訓練數據至最終輸出，在長時間運行下仍能如預期般運作。「可靠」指提高系統的穩定性、無錯誤地正常運行；「穩健」是指模型在假設或條件發生變化時，仍能保持有效的性能；「安全」則是保護數據及模型本身。

無論是國際還是本地文獻，均一致主張須確保人工智能應用能按預期運作，以確保可靠、穩健及安全。相關文獻強調的包括對誤用的警覺、完整性、穩健性、韌性、效能、品質、適切性、準確性與安全性，當中特別以預防傷害為重點。

人工智能應用的設計與開發，必須充分考慮到，其所依賴的環境、數據及流程會隨時間而變化。惡意使用者可能利用這些人工智能運作環境的變化，運用各種技術入侵系統，愚弄人工智能模型作出看似非常可信但其實錯誤的預測。採納「可靠、穩健及安全」原則，有助機構識別人工智能應用的潛在弱點、提升整體效能、抵禦對抗性攻擊，並持續監察模型在整個運作週期中的長期表現；在可行情況下，更應以可驗證的方式實施這項原則。

可靠性

作為建立可靠軟件應用程式的基礎，可靠性是一個苦心經營的過程，目的是在特定時間與環境下，致力提升系統無誤地完成其預期功能的機率[6]。隨着人工智能算法及其相關系統日趨複雜，嚴謹的軟件工程實務（例如相關標準和軟件測試）變得更為重要，以確保達致可靠性所涵蓋的兩項指標：

- *服務可用性（Service Availability）*：指系統或組件在需要時能夠正常運作及可供使用的程度，通常以機率表示[7]。在關乎安全或實時的任務中，人工智能應用的服務可用性極為重要，若有缺失，將可能導致嚴重甚至致命後果。

- *可服務性（Serviceability）*：有時也稱為可維修性（Maintainability），指系統是否容易被維護或修復。隨着人工智能應用所運行的基礎設施變動、模型須滿足的要求轉變，以及隨時間浮現的問題與錯誤，人工智能開發人員須定期修正缺陷、分析根本原因，並防止未來發生故障。

雖然個別從業者均應依循業界最佳實務和機構行為守則以維持品質與誠信，同樣重要的是，機構應嚴格監察及管理開發人工智能應用所使用的數據的品質及完整性。「數據品質」決定了資訊是否能可靠地用於預定目的，並影響資訊的可用性；而「數據完整性」則是指資訊在物理與邏輯層面上的可靠程度，其建基於數據在整個生命週期內的準確性與一致性，並要求在連續更新過程中不應出現非預期變更。有關數據品質及完整性的關鍵考慮因素包括：一致性、準確性、有效性、時效性、獨特性及完整性。

穩健性

穩健性是指模型在其變數或假設條件發生變化時，仍能有效執行的特性。為確保模型具備穩健性，必須把驗證與錯誤處理機制納入整個數據科學流程中的每個步驟，即數據準備、數據導入，以至預測。

穩健的模型在接觸到全新且獨立（但相似）的數據集時，必須能表現一致，並具備應

[6] http://www.mit.jyu.fi/ope/kurssit/TIES462/Materiaalit/IEEE_SoftwareEngGlossary.pdf
[7] http://www.mit.jyu.fi/ope/kurssit/TIES462/Materiaalit/IEEE_SoftwareEngGlossary.pdf

對執行過程中錯誤與極端情況的能力。穩健性亦包括當模型暴露於惡意輸入或參數時，能夠設置有效的措施處理錯誤，以保障其運作安全。

安全性

現行許多針對傳統軟件開發的安全實務，亦適用於人工智能和機器學習模型。由於這些模型高度依賴經篩選和整合來自公共或第三方來源的數據，因此必須具備辨識惡意輸入與良性異常數據的能力。在進行保安設計時，應特別注意對數據進行清理、保護及加密，並為已訓練模型設計合適的存取控制機制。

對抗性攻擊是其中一項主要的安全風險，指透過對輸入數據加入細微但蓄意的干擾，以誘導人工智能模型作出看似非常可信但其實錯誤的預測。此外，即使不對外開放直接使用其人工智能模型、僅以應用服務形式開放，機構仍須設計及實施保安措施，以防止攻擊者對模型進行逆向工程。

例子（更多實例請參閱第4章「人工智能實務指引」）：

- 人工智能系統應具備技術上的穩健性，並應具備足夠保護以防任何惡意使用。惡意使用者可能會設計輸入資料，以試圖了解並欺騙人工智能模型。因此必須設置控制措施，確保人工智能應用不會受控於潛在駭客，以致被訓練執行非預期任務。否則，若出現此類情況，將對使用該人工智能應用的相關機構造成法律責任及聲譽風險。
- 應定期進行持續監察工作（包括模型的驗證、準確性核實與維護），以提升人工智能系統的安全性與穩健性。因為駭客通常會針對存在安全漏洞的過時軟件進行網絡攻擊，不斷更新人工智能應用可有效將相關風險降低。

3.5.1.3 公平

原則：人工智能應用所作出的建議或產生的結果，應以公平方式對待相似群體中的每個人，不應偏袒或歧視，亦不應造成或引致傷害。

根據《世界人權宣言》第 2 條：「*人人有資格享有本宣言所載的一切權利和自由，不分種族、膚色、性別、語言、宗教、政治或其他見解、國籍或社會出身、財產、出生或其他身份等任何區別。並且不得因一人所屬的國家或領土的政治的、行政的或者國際的地位之不同而有所區別，無論該領土是獨立領土、託管領土、非自治領土或者處於其他任何主權受限制的情況之下。*」因此，人工智能應用必須盡量避免就上述因素出現歧視或偏見。這項反歧視原則亦見於《消除對婦女一切形式歧視公約》。

以下是在理解公平性及其如何應用於機構或人工智能應用時所需考慮的主要法律用語：[8]

- **偏見：**指損害或傷害他人權利。
- **歧視：**指基於他人種族、膚色、性別、性傾向、年齡、生理或心理殘疾、婚姻狀況、家庭或照顧責任、懷孕、宗教、政治觀點、國籍或社會出身等因素，對其作出不利、不公或有害的對待、優待、排斥或區別。
- **公正：**指根據事實及法律，在無偏見、無干預、無先入為主觀念、無不合理情況下作出行動或決策。
- **正面差別待遇：**指基於他人種族、膚色、性別、性傾向、年齡、生理或心理殘疾、婚姻狀況、家庭或照顧責任、懷孕、宗教、政治觀點、國籍或社會出身等因素，而刻意地：
  - 對他人作出特別對待；
  - 作出影響他人的行動；及 / 或
  - 作出影響他人的決策，以實現下列其中一項或多項目的：
    - 實質平等；
    - 平等地享有權利；
    - 平等地行使人權；及 / 或
    - 平等地行使基本自由。

必須確認，公平性是一種社會建構。在人工智能領域中，公平性有多種數學定義，當選擇其中一種定義時，往往會忽略其他面向。換言之，由於存在固有的限制，並因應不同的個別情況，並非每項人工智能決策都能實現「普世公平」(universal fairness)。

[8] 請注意，這些字眼 / 字句往往沒有普遍一致的法律含意，其具體定義可能會因應用途及個別情況而有所不同。因此，在本文件中所採用的定義刻意保持廣泛且相對簡單，以配合用於廣泛決策框架。

然而，人工智能應用可根據特定的公平性目標進行設計，從而減少部分可見的不公平情況，建立整體上更負責任的系統。由於人工智能模型是從數據中學習，其複雜性令出現無意識偏見的風險更高，數據可能包含特定群組出現代表性不足或代表性過高的情況；反觀傳統資訊科技系統則多依循預設規則運作。因此，基於人工智能應用的獨特性，我們需要更為關注公平性問題。

例子（更多實例請參閱第4章「人工智能實務指引」）：

- 訓練數據應避免產生偏見的特徵，例如樣本數目差異（少數群體的數據明顯較少）、選擇性偏見（特定群體較難被納入樣本）、模型設計偏見，以及模型使用及回饋偏見（例如在依賴數據與算法時出現對某些群體的刻板印象）。若人工智能應用本身存在偏見，其決策將傾向某些群體，產生不公平對待。
- 應確保取得的原始數據具備完整性，以確保人工智能應用產生公平的結果。若使用無效或不準確的數據訓練人工智能模型，將導致決策偏頗，受影響的使用者亦會因數據集的瑕疵而遭到無意的歧視。

3.5.1.4 多元與包容

原則：應通過了解和尊重所有受影響的持份者的利益，促進人工智能應用的包容性，並兼顧多元化的使用者需求。

本原則確保人工智能應用須尊重本地文化規範，同時力求避免與一般原則或執行原則相矛盾。鑑於全球各式各樣的人都可使用人工智能應用，他們對人工智能應用行為的理解也可能有很大差異。為了實現真正的多元性與包容性，應盡可能涵蓋更多元化的人工智能使用者，包括來自不同文化、興趣、生活方式及專業範疇的用戶。

例子（更多實例請參閱第4章「人工智能實務指引」）：

- 人工智能應用開發人員應在設計階段考慮不同背景、種族和性別人士的需要。若忽略這些差異，人工智能應用可能會對某些群體表現出不同甚至不公平的行為，造成使用者的不便。例如曾有人工智能招聘工具因由白人開發，結果偏向篩選白人申請者，對其他種族構成偏見。

3.5.1.5 人為監督

原則：人工智能應用用於決策或操作過程中所需的人為干預程度，應取決於所涉及的道德問題的嚴重程度。

人工智能應用在不同程度上被視為自主系統，並在日常生活中日益普及。基於本原則，必須保持對人工智能的人為監督，這是指在人工智能的應用情境下，人類需具有作出選擇、思考、判斷結果並採取相應行動的能力。

例子（更多實例請參閱第4章「人工智能實務指引」）：

- 應根據多項因素（例如人工智能應用的效益與風險、人工智能決策所帶來的影響、營運成本，以及社會規範與價值觀的演變）確保作出適當程度的人為介入。
- 應實施相關控制措施，以便在系統故障時，特別是在故障可能對人類安全造成影響的情況下，容許人為介入。例如自動駕駛車輛應提供人類使用者介入的選項，當車輛未能正確偵測路面行人時，人類便能及時阻止車輛造成意外。

3.5.1.6 合法與合規

原則：負責人工智能應用的機構，應貫徹依據法律、規例及相關監管制度行事。

在所有情況下，人工智能應用必須遵守所有適用的法律、法規、標準和指引。因此，任何涉及人工智能應用開發或實施的機構，必須全面掌握並貫徹執行所有相關法律、法規、標準和指引。

例子（更多實例請參閱第4章「人工智能實務指引」）：

- 開發的人工智能系統必須符合相關規例和法律要求。
- 應制定人工智能應用的合規審查流程，以隨時掌握規例變動，並確保政策與操作流程符合規管要求。

3.5.1.7 數據私隱

原則：個人應享有以下權利：

(a) 獲告知個人資料的收集目的及可能的資料轉移對象，且該個人資料的收集須出於合法目的、以合法及公平的方式進行，並且所收集的資料數量不應超過該目的所需。請參閱《個人資料（私隱）條例》（《私隱條例》）的**保障資料第1原則**「收集目的及方式」。

(b) 獲保證資料使用者已採取所有切實可行的步驟，以確保個人資料是準確無誤，並不保留超過所需的時間。請參閱《私隱條例》的**保障資料第2原則**「準確性及保留期間」。

(c) 要求個人資料僅用於該資料原本的收集目的及直接與該目的相關。否則須得到個人的明確及自願的同意。請參閱《私隱條例》的**保障資料第3原則**「個人資料的使用」。

(d) 獲保證資料使用者已採取所有切實可行的步驟，保障其持有的個人資料不會未經授權或意外地被查閱、處理、刪除、喪失或使用。請參閱《私隱條例》的**保障資料第4原則**「個人資料的保安」。

(e) 獲得相關資訊：(i) 個人資料方面的政策及實務；(ii) 所持有的個人資料的種類；(iii) 個人資料的主要使用目的。請參閱《私隱條例》的**保障資料第5原則**「資訊須在一般情況下可提供」。

個人[9]應享有權利期望機構以有利於個人或社會大眾的方式處理與其相關的數據。若處理數據的利益主要由機構獲得，則應透過可證明的審查程序，確定對個人造成的影響輕微。個人亦應享有權利控制對其產生重大影響的數據用途。在可行的情況下，應透過適當程度並按實際情況應用的同意和查閱權來實現此項權利。如無法取得同意、情況不合適或影響較小，個人應有權利知悉相關的問責程序以確保數據使用是公平和負責任的[10]。

[9] 個人包括個人及個別群體

[10] 並非所有數據用途皆適合控制，例如用於安全分析的數據

個人應享有權利獲告知其可辨認個人身份的資料的使用情況。這在人工智能應用中尤其重要，因為訓練機器學習算法所用的資料集往往包含個人及敏感資料。此外，個人亦需理解到，其在網絡上的數碼形象及在數字環境中的互動方式，事實上與現實生活中的截然不同，這有助他們更好地保護自己的私隱和權益。

例子（更多實例請參閱第4章「人工智能實務指引」）：

- 清楚說明顧客的數據在人工智能系統中將如何、為何、何地及何時被使用。
- 確保盡量減少個人數據的收集和處理，遵守資料保留政策，並符合私隱規例，例如《私隱條例》及相關指引。

3.5.1.8 安全

原則：人工智能應用在其整個運作生命週期內，均不應危及人類的身體安全或心理健全。

安全指應將意料之外的傷害風險降至最低，包括身體、情感和環境安全。隨着人工智能應用在認知能力和自主程度上持續演進，出現未能預料或意料之外的行為風險亦隨之增加。例如人工智能應用可能試圖操控其獎勵機制，或學習系統故障造成難以預測的後果，因而可能引致各種危險的情況。

例子（更多實例請參閱第4章「人工智能實務指引」）：

- 用於協助人類進行決策的人工智能工具必須安全可靠，並符合受其影響人士的道德價值與偏好。例如用作照顧長者的機械人，其設計應確保不會對長者造成任何身體或心理上的傷害。

3.5.1.9 問責

原則：機構須對其人工智能應用，包括其使用和誤用，所產生的道德影響負責。同時應有明確可識別的問責方，無論是個人或機構。

問責原則確保各持份者的責任與法律責任清晰明確，並可追究相關責任，這包括確保從規劃以至記錄保存的整個過程中，各項責任均予以履行。

問責性是大多數法例及 / 或私隱框架中的基石原則。有必要釐清人工智能系統的各持份者在其使用、誤用及行動中所產生的道德及法律責任。

實踐問責的核心，在於建立一套全面、端對端的治理程序。治理通常指一整套涵蓋內部與外部的政策、步驟與監督安排，目的在於管理系統風險並履行必要責任。這有助於管理道德責任，並有助追查及減低與大數據及人工智能項目相關的風險。透過引入適當的治理框架，機構可在推動創新與保障公眾利益之間取得平衡。

3.5.1.10 有益的人工智能

原則：人工智能的發展應促進共同福祉。

人工智能應用不應對人類造成傷害，相反地，應積極促進共同利益和福祉。科技的創造固然可出於善意，但若未有考慮福祉因素，仍可能會對人的心理健康、情緒、自我認知、自主性、實現目標的能力及其他層面的福祉造成嚴重的負面影響。

人工智能應用即使可以合乎道德、法律、有盈利及安全地使用，但仍未必足以貢獻人類福祉。福祉涵蓋心理、社會、公平經濟及環境等考慮因素，有助評估科技進步所帶來的整體效益，並可有效衡量人工智能對人類福祉所造成的意外負面影響。

例子（更多實例請參閱第4章「人工智能實務指引」）：

- 人工智能的發展應促進社會和人類文明的進步，創造更高效的工作方式與生活模式，並改善民生與福利。若人工智能應用未能為人類文明帶來任何益處，公眾或會因其缺乏效益而選擇拒用。

- 未能促進共同利益的人工智能，有機會損害甚至摧毀社會整體的福祉，例如自主武器便可能會對全人類構成嚴重威脅。

3.5.1.11 合作與開放

原則：應培養人工智能生態環境中多方持份者開放合作的文化。

合作與開放關乎建立信任。這包括不同持份者與服務使用者及其他受影響群體，就風險及應對計劃進行合作和溝通。此原則強調透過教育公眾認識人工智能，從而建立信任[11]。

人工智能應用的開發與實施，應是開放的協作過程，涉及來自不同背景的內部團隊，並積極與社區內不同的使用者群體互動。在這方面，機構應主動與各類持份者展開合作，包括服務使用者、大學、研究中心、政府及專業團體等，以助減低人工智能應用中潛在的排斥與偏見風險。

例子（更多實例請參閱第4章「人工智能實務指引」）：

- 機構應積極推動人工智能應用的跨領域、跨行業及跨機構合作。若不同團隊之間缺乏合作，人工智能模型可能因缺乏多元觀點或使用者測試而作出偏頗決策。項目經理可與項目團隊及服務使用者合作，在部署前進行測試，以提升模型的準確性與包容性。

3.5.1.12 可持續性與公正轉型

原則：人工智能的發展應確保已制定緩解策略，以應對任何對社會及環境系統的潛在影響。

[11] https://www.pwc.ch/en/publications/2017/pwc_responsible_artificial_intelligence_2017_en.pdf

人工智能有助於推進傳統行業轉型，以應對社會及環境挑戰，並提升人類福祉。儘管人工智能為解決部分社會及環境問題帶來了變革性的機遇，但若缺乏適當指引，亦可能加速社會與環境的惡化。人工智能應用的推行可能會對社會及環境造成深遠影響，例如導致工人被取代，或在長者、病患及殘疾人士照顧、教育，以及地球保育等領域帶來挑戰與改變。為使人工智能可持續發展，最終目標是要確保其與人類價值觀一致，承諾為人類提供安全的技術應用。實際上，這意味着需要建立制衡機制，以確保不斷演進的人工智能應用能夠保持可持續性。因此，對於可能會對社會及環境造成影響的人工智能應用，必須實施適當的風險緩解計劃，並以盡可能公平和具包容性的方式進行，確保所有群體都能受惠，避免任何人被排除在外。

例子（更多實例請參閱第4章「人工智能實務指引」）：

- 人工智能系統可用於減少碳足跡，例如人工智能技術可用於優化數據中心的能源使用，從而節省能源。
- 支援人工智能應用的基礎設施應具備可擴展性，以維持長期不斷的改進。反之，若人工智能應用建基於無法擴展的基礎設施上，將可能限制人工智能模型（以及人工智能應用）在未來可持續作出微調的能力。

3.5.2 人工智能治理

人工智能治理是指管理及控制人工智能項目與應用的實踐方針。以下是在開發及維護人工智能應用時取得接納及成功的重要元素：

- 設立治理架構，以監督人工智能項目的實施及進行人工智能評估。
- 界定影響「人工智能道德框架」在使用及維護上的角色和責任（詳情請參閱第3.5.2.1節「人工智能治理架構」）。
- 制定一系列實務指引以協助機構規劃、開發、部署及監察人工智能應用（詳情請參閱第4章「人工智能實務指引」）。
- 就應用成效評估相關實務的採用情況（詳情請參閱第5章「人工智能評估」）。

有效的人工智能治理不應僅由技術主導，因這只能解決技術持份者所關注的問題，卻無法緩解公眾提出的疑慮，亦未能滿足與第二及第三道防線整合的端對端治理要求。「三道防線」是眾多機構廣泛採用的成熟治理概念。

圖9：三道防線模式

圖9展示不同防線及其角色。本治理架構的組成如下：

- **第一道防線**為項目團隊，負責開發人工智能應用、風險評估、執行風險緩解措施，以及記錄「人工智能評估」工作。
- **第二道防線**由項目督導委員會及項目保證小組組成，確保項目品質，制定人工智能應用的驗收標準，進行獨立檢視，並批准人工智能應用。在予以批准前，須完成「人工智能評估」以確保符合「人工智能道德原則」。
- **第三道防線**涉及資訊科技委員會，如未設立此委員會，則由首席資訊官擔任，並可選擇由人工智能道德委員會提供支援。人工智能道德委員會可由外部顧問組成，就人工智能道德議題提供建議，並加強機構採用人工智能的能力。第三道防線負責審核、建議及監管高風險人工智能應用。

3.5.2.1 人工智能治理架構

本治理架構明確規範各相關角色 / 職能的關鍵工作事項及相應職責。

機構須參照「人工智能道德框架」規劃、實施及維護其人工智能應用。項目團隊可由熟悉機構業務、資訊科技項目管理及人工智能開發流程的現職人員組成，成員可包括系統分析員、系統架構師及數據科學家等，負責執行人工智能模型或應用的開發工作，並管理部署及部署後表現。項目團隊規模須視乎機構的組織架構、運作模式及所開發人工智能應用的範圍而定。

下表列明人工智能治理架構的角色和責任：

角色	責任
資訊科技委員會，如未設立此委員會，則由首席資訊官擔任	資訊科技委員會負責監督機構的資訊科技應用（包括人工智能應用），並在高風險人工智能項目開展前審核「人工智能應用影響評估」（詳見第5章「人工智能評估」）。資訊科技委員會的職權範圍涵蓋高層次責任，包括督導機構運用資訊科技、將資訊策略方向與業務目標結合，並確保資訊科技措施配合機構方針政策。 任何人工智能項目如在「人工智能應用影響評估」中觸發第5.1節「人工智能應用影響評估」及《附錄C —人工智能應用影響評估範本》所列的風險分級準則，即界定為高風險人工智能應用，須獲資訊科技委員會批准。項目經理負責核實風險分級準則的答案，並在有需要時在項目團隊協助下，判定該人工智能應用是否屬高風險。資訊科技委員會可隨機構的經驗與人工智能技術能力提升，進一步修訂高風險人工智能應用的定義。

角色	責任
	資訊科技委員會須履行以下責任： • 審核高風險人工智能應用的「人工智能應用影響評估」；以及 • 根據已審核的「人工智能應用影響評估」，審視項目團隊提交的建議及補救措施，並向項目督導委員會 / 項目保證小組及項目團隊提供意見，以確保適當考慮了風險、道德影響及效益。
人工智能道德委員會（非強制設立）	人工智能道德委員會負責就人工智能應用及人工智能評估涉及的道德問題，向資訊科技委員會 / 首席資訊官提供意見。人工智能道德委員會可由外部顧問組成，具備技術範疇（例如數據、人工智能分析）、道德、法律、風險 / 效益評估及保安等專業知識。委員須熟悉人工智能項目道德類議題的評估，並了解審批人工智能項目的準則，例如： • 風險被判定為合理，並已就預期可達致的效益採取緩解措施； • 已考慮對數據主體群體（例如兒童、在囚人士、教育程度不足者、精神障礙者及其他弱勢群體）構成的風險； • 收集及使用個人資料的道德（許可）基礎已妥善記錄（例如屬預期使用範圍、已獲同意）； • 設有足夠措施保障項目涉及的個人私隱；以及 • 項目團隊及資訊科技委員會 / 首席資訊官已完成相關人工智能評估及作出決定。 人工智能道德委員會須履行以下責任： • 就開發合乎道德的人工智能提供建議及顧問意見，以支援涉及高風險人工智能應用的機構 ；

角色	責任
	• 就機構的決定及人工智能評估提供意見，以調整及提高使用人工智能應用的透明度；以及 • 提供監督及外部知識，以協助機構建立公眾對人工智能應用的信任及提高其透明度。 「人工智能應用影響評估」的審核人員數目視乎人工智能項目的性質及複雜程度而定，審核人員可包括以下範疇的專家： • 私隱； • 數據保安 / 資訊保安； • 工程 / 科技 / 人為因素（即把心理及生理原則應用於產品、流程及系統的工程設計）； • 數據分析 / 數據科學； • 法律； • 公共關係；以及 • 具備機構專業知識 / 最佳實務經驗的一般人員。 當項目規模龐大、影響深遠且備受關注，其道德價值可能受到質疑時，有需要設立人工智能道德委員會進行獨立審核。機構在規劃人工智能應用項目時，可根據人工智能應用的潛在負面影響評估此項需要，並在「人工智能應用影響評估」的影響部分記錄相關評估結果。 人工智能道德委員會旨在提供一定程度的保證，確保相關活動符合道德、妥當適切，且經得起公眾檢視。適用例子包括：對患者福祉有重大影響的醫療服務、以及涉及處理大量敏感且可識辨個人身份的資料的人工智能信貸評分系統。

角色	責任
	須注意的是，委員會的組成可因應個別機構的具體需求及所屬行業而有所不同，包括委任渠道。機構可沿用現行設立專家小組或諮詢委員會的做法，邀請外部專家或業界代表提供意見。
項目督導委員會 **項目保證小組**	項目督導委員會 / 項目保證小組須履行以下責任： • 釐定人工智能應用的驗收準則及整體要求； • 審核人工智能評估，以確保人工智能應用的影響受控； • 在整個人工智能生命週期內進行持續監察； • 就高風險人工智能應用提供意見； • 在人工智能應用交付前作最終批准； • 就可能涉及數據私隱問題的高風險人工智能項目，適時與個人資料私隱專員公署（「私隱公署」）溝通；以及 • 就可能涉及法律問題的高風險人工智能項目，適時向法律部門或律師尋求法律意見。 所有項目均須通過現有項目管理架構審批。
項目團隊	項目團隊負責人工智能項目及人工智能評估的執行工作。團隊成員可包括內部人員或承辦商，按項目經理的指示完成項目任務。主要角色包括系統分析員、系統架構師及數據科學家，亦可包括業務使用者。 項目團隊須履行以下責任： • 協助項目管理工作，以確保人工智能應用在整個生命週期內符合品質標準（包括人工智能及數據道德、品質程序、行業標準及政府規例）； • 完成人工智能評估並交付人工智能應用。項目團隊開發的每項人工智能應用均須完成「人工智能應用影響評估」；

角色	責任
	• 推薦人工智能項目，並透過項目經理向項目督導委員會／項目保證小組提交「人工智能應用影響評估」以供審核； • 如有需要，可參考「人工智能道德框架」制定機構專用的人工智能標準及指引； • 就可能涉及數據私隱問題的高風險人工智能項目，適時與私隱公署溝通；以及 • 就可能涉及法律問題的高風險人工智能項目，向法律部門或律師尋求法律意見。
項目經理 **（項目團隊成員）**	項目經理須履行以下責任： • 確保人工智能應用在整個生命週期內符合品質標準（包括人工智能及數據道德、品質程序、行業標準及政府規例）； • 確保對人工智能項目進行報告／檢查／評估，包括任何數據治理檢查； • 審核使用案例，制定人工智能應用的整體規劃及後續設計； • 確保為現職人員提供相關人工智能培訓，以提升其人工智能模型技能；以及 • 向資訊科技委員會／首席資訊官提供行政支援，包括安排會議、準備會議記錄、草擬文件及交付成果、向相關各方傳閱材料以徵求意見及批准、處理查詢，以及就人工智能計劃與不同持份者協調。 即使屬概念驗證項目，亦應指派須履行類似責任的項目經理。
系統分析員、 **系統架構師**	系統分析員及系統架構師須履行以下責任： • 分配及追蹤用於人工智能模型的數據集擁有權；

角色	責任
（項目團隊成員）	• 確保已取得所有採購數據的相關牌照； • 處理「提取、變換及載入」程序，為數據科學家準備及轉換數據；以及 • 確保遵守現行程序，包括歸檔、備份、保安、私隱及數據保留等程序。
數據科學家 **（項目團隊成員）**	數據科學家須履行以下責任： • 確定人工智能應用的功能及訓練程序，以滿足特定的功能要求（包括概念驗證項目）； • 遵循模型測試及驗證特定的控制措施和程序； • 在影響人工智能應用治理及控制的系統或基礎設施有所變更時，通知項目管理層，由其轉告資訊科技委員會／首席資訊官； • 管理、整合、擴展及部署人工智能應用； • 將人工智能應用轉換為生產代碼，並進行大規模模型訓練； • 管理人工智能應用部署後的表現及穩定性；以及 • 管理用於人工智能應用開發、訓練、部署、監察及測試／驗證的基礎設施和平台，並負責人工智能應用的管理、監察及故障排除。

3.5.3 人工智能生命週期

為使機構在開展人工智能項目或開發人工智能應用時有所依循，本框架第4章將詳細說明人工智能生命週期各階段的實務規範。

下列六步流程圖解構人工智能生命週期的概念：

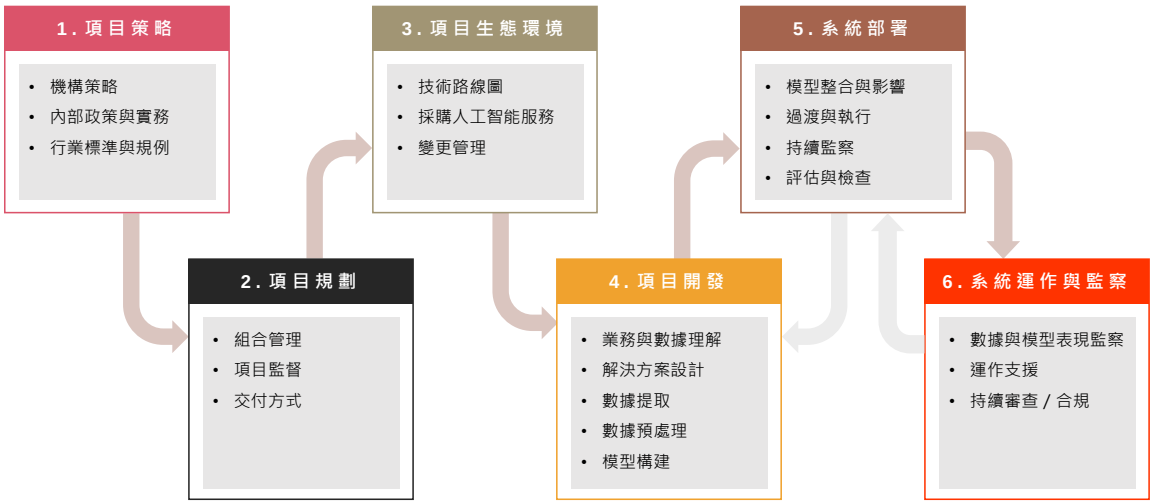

圖10：人工智能生命週期概覽

人工智能生命週期展示了人工智能項目所涉及的各個階段，當中包括：

- 引導機構了解人工智能項目的不同階段及相關要求；以及
- 作為制定人工智能實務的參照，以配合典型的人工智能實際開發階段。

根據此人工智能生命週期，可落實關鍵能力與職能，構建有效且全面的計劃性治理體系。下文將進一步說明人工智能生命週期各階段對應的涵蓋範圍及能力要求。

1. 項目策略

- 機構須根據其策略目標及「人工智能道德原則」制訂人工智能與數據策略，並以文件記錄及在其內部有效傳達；
- 機構應正式編配主要領導職位，負責統籌綜合策略的實施，並明確界定執行關鍵角色的分工。其後須建立全面而有效的流程，定期監察及分析外界對人工智能與數據應用的期望有何變化，據此調整其策略、政策及治理；
- 機構應檢視其策略、政策及程序，並確保其與「人工智能道德原則」沒有抵觸；以及
- 人工智能應用的整體治理流程須包含完善的評估機制，以確保所有政策有效落實；相關應用須從效益、影響及風險角度，針對所有持份者進行全面評估。

2. 項目規劃

- 須建立明確的人工智能項目組合管理流程，以確保項目配合業務策略目標，並按短期及長期目標訂立優先次序，包括對社會產生正面影響的項目；
- 管理層須負責制定監督機制，以確保項目組合的決策以風險及道德考量為基礎；
- 須設立清晰的項目管理架構，以促進人工智能應用組合的規模化及採用；
- 須在董事會或高層管理人員的督導下，訂明主要責任；
- 須制定管理流程，以確保在整個人工智能項目組合中貫徹有效的數據治理；
- 須設立清晰流程，評估所有人工智能計劃對個人、團體或社會的負面影響，並從廣泛的效益與風險角度進行分析。同時須定期評估公眾期望的變化，確保評估過程具備可審計性，包括人工智能系統流程與結果的可追溯性及記錄；以及
- 須正式成立高層決策機構 / 小組，負責處理高風險人工智能應用事宜。

3. 項目生態環境

- 須訂明具清晰權責劃分的程序，以應對未來可能出現的技術需求、模型更新等事項，包括設立正式流程評估現有技術環境、需求及採購選項，從而識別差距並按需要調整人工智能發展路線圖；
- 須訂立正式的職責分工及流程，評估並確保相關人員具備落實人工智能目標及責任所需的技能與知識；
- 須定期評估採購團隊的專業能力，並分配合適的專才以執行變更管理、培訓及過渡至常規運作階段；以及
- 須評估第三方供應商提供的工具、數據及技術，以確保其符合「人工智能道德原則」、數據使用規範及 / 或人工智能治理要求。

4. 項目開發

- 須建立與機構的道德價值觀相符的規範化項目管理計劃和系統，涵蓋所有新人工智能項目開發過程中所需的組織流程、功能與非功能需求、業務價值觀的配合，以及風險與測試評估，當中包括與企業主數據管理要求的正式整合；
- 在設計與開發階段，應按「人工智能道德原則」在機構各層面進行數據、分析 / 人工智能、自動化 / 軟件的整合開發，並實施適當的驗證及核實程序；
- 人工智能應用與數據適用性須配合業務目標及技術要求；
- 使用第三方數據或技術時，須正式評估所有必要的組織流程、風險及測試要求，同時明確界定需求，並將策略納入項目管理；
- 須制定相關計劃及 / 或系統或程序，以免在人工智能系統中產生或強化不公平的偏見，包括使用輸入數據和算法設計兩方面；
- 須建立正式流程，在人工智能應用的開發、部署及營運階段測試和監察可能出現的偏見；
- 須根據正式選定的公平性定義，評估人工智能應用的可靠性、模型敏感度及模型表現；以及
- 須定期評估性能與偏見的權衡關係，以及可解釋性與性能之間的取捨。

5. 系統部署

- 須制定具明確責任分配的正式流程，以確保在部署階段全面評估人工智能應用對所有持份者的影響。評估內容須包括：效益與風險緩解措施是否取得適當平衡以支援人工智能應用及數據處理活動、是否符合人工智能道德目標，以及是否已建立有效的風險緩解控制措施以減低風險；
- 須建立正式的決策架構及風險為本的上報機制，以解決問題、評估高風險數據的使用情況、作出決策，以及正式批准人工智能應用的部署；
- 須建立明確的項目管理系統，以評估及制定可理解的人工智能應用部署計劃。同時須設立管理系統及流程，以持續識別、檢視及緩解已部署人工智能應用的使用風險；以及

- 項目管理系統及流程須確保項目能持續為服務增值，並對預期效益進行量度、定義、主動追蹤及匯報。

6. 系統運作與監察

- 須制定明確的系統監察機制，包括設立客戶申訴處理流程，以識別及處理與人工智能應用相關的問題 / 事故；
- 須建立上報流程，將人工智能應用相關的重大問題 / 事故上報至資訊科技委員會 / 首席資訊官，以供知悉及採取相應行動；以及
- 須為所有人工智能應用支援團隊訂明清晰的職責範圍及服務水平協議，並建立持續學習機制（包括機器學習及人員 / 組織學習），以確保與其他流程（例如安全事件管理）有效整合。

如圖11所示，人工智能生命週期與傳統的系統發展生命週期模型互相對應。

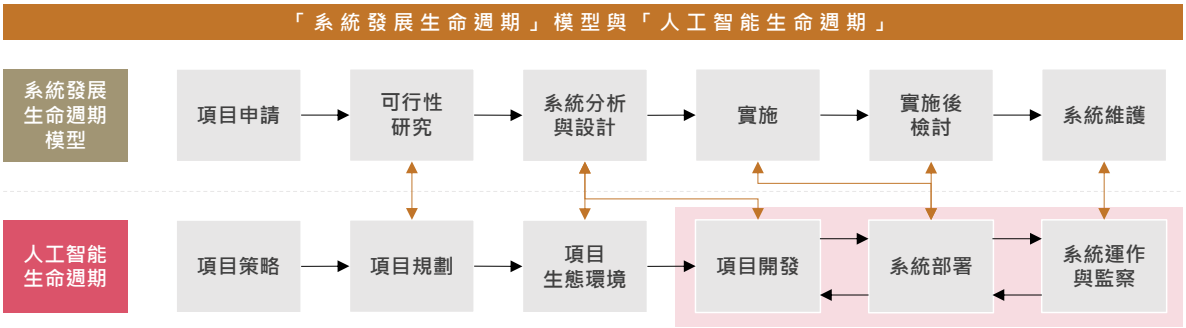

圖11：「人工智能生命週期」與「系統發展生命週期」模型互相對應

在人工智能應用的典型開發過程中，數據準備工作至關重要，因為優質數據往往是構建良好人工智能模型的基礎。數據收集與預先處理作為項目開發的組成部分，往往需要持續進行，這是因為在整個開發過程中，人工智能模型可通過獲取更優質或更大量的數據來反覆優化訓練效果。傳統軟件生命週期的開發模式，是通過指令集編寫資訊科技應用程式來處理既定情境，其後這些資訊科技應用程式會運用運算能力與其他資源來處理輸入系統的數據。至於人工智能應用的開發模式則截然不同，是通過向系統輸入海量數據進行訓練，在所有數據經處理後，最終生成能解決新問題的人工智能模型或解決方案。

在人工智能生命週期中，開發與部署階段、系統運作與監察之間往往存在持續的反饋循環，以便不斷改進，這有別於傳統的系統發展生命週期。

第4章

人工智能實務指引

4. 人工智能實務指引

機構於發展人工智能應用項目時，可參照本指引涵蓋人工智能生命週期各階段的實務建議（包括策略、規劃、生態環境、開發、部署及持續監察）。機構須繼續遵循所屬範疇的法例、政策及資訊科技項目指引，以及相關行業的現行做法。本指引所列措施屬補充性質，旨在提供符合道德原則的人工智能應用指引，機構宜予採納。

本節闡述機構於發展人工智能應用項目時，在人工智能生命週期各個階段應考量並採納的優良實務。機構可於項目規劃初始階段參照本指引，並按機構需求及技術應用進程，循序擴展至生命週期其他範疇。

本指引部分內容涉及技術數據及人工智能專業知識。項目主管須督導團隊進行適切數據分析，並於評估中記錄在案。參與人工智能項目的團隊必須獲得培訓，以確保具備相關技術能力。

本指引引用的例子來源各異，僅作說明用途，並非規定機構須遵循特定例子。

備註：第3.5.2節「人工智能治理」所述角色僅限於該治理架構組成部分，而「人工智能道德框架」的適用對象（見第2章「目的」）涵蓋範圍更廣。本指引收錄採購團隊等因涉及特定操作／規例（例如項目團隊採購人工智能解決方案時會涉及機構採購程序）而存在的其他角色。然而，此類角色不會納入人工智能治理架構，故未列於第3.5.2節「人工智能治理」。

4.1 人工智能生命週期及實務

人工智能生命週期階段（請參閱第3.5.3節「人工智能生命週期」）用於歸集各類實務措施及示例。視乎各機構的實際運作情況及背景，相關機構可於建議人選中指派執行方。

4.1.1 項目策略

4.1.1.1 機構的策略、內部政策與實務

定義：

機構推展人工智能項目的策略，須配合機構現有目標及資訊科技策略。在制訂機構策略和進行決策時，應考量或納入「人工智能道德原則」。

此階段涉及的持份者：

資訊科技規劃人員 / 管理層、業務使用者

實務措施及示例：

措施	措施詳情
制訂人工智能策略或計劃並進行覆核	• 機構應制訂人工智能策略，以確保人工智能成果符合機構願景及使命。這可通過解釋人工智能發展的預期成果和實現途徑來達成。有關人工智能策略的主要組成部分，例子包括：願景、目標、路線圖，以及人工智能目標對機構目標和治理考量的銜接。治理人工智能應用部署的角色、人員及技術資源構成人工智能治理。 • 人工智能策略可採用標準策略文件格式予以記錄，並與現行機構策略文件保持一致。 • 人工智能策略完成時限視乎人工智能應用規劃及開發時程。策略本身無固定時限要求，機構應按「人工智能道德框架」要求，於人工智能計劃需要有關策略時完成。人工智能策略可納入機構的資訊科技計劃。 • 負責制訂及覆核策略的委員會或工作組，應組建多元化團隊，成員可包括公共關係、風險管理、法律事務、技術支援、人力資源等單位人員，以及認為必要的其他團隊。 • 機構應將「人工智能道德原則」納入人工智能策略，以提升全體持份者認知，並體現機構層面的認可。

措施	措施詳情
	• 若擬部署高風險人工智能應用，則必須於項目啟動及生產部署前完成資訊科技委員會 / 首席資訊官覆核。如需外部顧問強化機構對高風險人工智能應用或用例之監督能力，可設立人工智能道德委員會。高風險人工智能應用應按「人工智能應用影響評估」所列的風險分級準則識別。有關治理架構及覆核機制的詳情，請參閱第3.5.2節「人工智能治理」及第5章「人工智能評估」。 • 對機構的人工智能準備就緒度或成熟度進行現狀評估，記錄各職能領域情況，以確保為機構目標及投資制定可行路線圖。制定機構策略時，應考量流程、人員、系統及數據方面的差距。 • 機構人工智能策略（包括道德使用人工智能的價值觀與目標）應備存覆核記錄，列明機構關鍵考量因素。例如：某機構以「通過人工智能轉型減輕人力工作負荷，提升人工智能技能及資源利用率」為人工智能策略目標，其關鍵考量應包括：如何識別可應用人工智能的範疇、人工智能應用對終端使用者的裨益、人工智能應用採納成功率或其他可量化目標，以及預期成果。 • 在制定人工智能策略過程中進行的策略評估（例如強弱機危分析），可納入資訊科技規劃流程。此項工作可由機構項目團隊、資訊科技團隊或外聘顧問執行。 • 機構內部應充分傳達（如透過內部電郵）並理解人工智能策略。 • 示例範本請參閱附錄D「人工智能策略範本」。機構亦可採用其他格式及範本制訂人工智能策略。

就人工智能解決方案形成機構觀點	• 機構應完成「人工智能應用影響評估」（詳見第5章「人工智能評估」），以識別、審核及調配資源支持人工智能應用驅動的新構想（例如人臉識別、影像分析）：
	  ○ 評估人工智能應用對機構的效益，以及對個人或社會的影響；以及
	  ○ 識別負面影響與風險，包括：所需資訊的敏感性、潛在社會歧視、環境影響及法律成本。策略與優先事項

措施	措施詳情
	須協調一致，並考量相關效益與風險。上述工作應建立一套標準流程，以便在識別及批准用例時使用。 • 項目團隊應定期審視學術與業務動態，發掘可實現業務目標的人工智能應用或研究成果。機構需針對所屬範疇（例如醫療人工智能應用、執法範疇）開展市場研究。 • 市場研究應包括評估使用人工智能的機遇和風險。評估應參考現有案例（如有），以了解效益、風險及同類型項目的經驗教訓，確保機構在評估潛在人工智能應用時能充分考慮相關因素。
制定及設立評估指標	• 機構可採用量化或質性指標，以衡量人工智能應用的影響（例如：節省成本及縮減工日需求，或基於公眾在社交媒體上所持的正面或負面情緒對機構聲譽的影響，以及應對突發災難事件的能力）。 • 採用先進人工智能應用的機構應設立指標，以量度反饋、表現並排定人工智能應用的優先次序，例子包括檢查預測成功與失敗的情況，以核實人工智能應用是否以可接受方式運作。第4.1.6.1節「數據與模型表現監察」載有更多指標示例，以供參閱。
遵循適用的內部政策規定	• 機構的內部政策及實務指引（包括任何人工智能指引）應便於查閱，以便內部人員知悉，並應確保人工智能團隊與項目團隊徹底理解有關內部政策及實務指引。 • 機構應向全體人員宣揚內部政策及實務指引，提醒他們在負責的項目中貫徹應用。 • 機構應確保設有流程及覆檢機制，可確認項目團隊運用人工智能應用的方式符合機構核心價值理念、法定要求及/或社會期望（例如透過完成「人工智能應用影響評估」）。此舉旨在保證人工智能項目執行過程遵循內部政策與實務指引。
更新政策規定	• 檢視機構策略、政策及程序，並確保其與「人工智能道德原則」（見第3.5.1節「人工智能道德原則」）相符。

措施	措施詳情
	• 將各項標準轉化為資訊科技標準的工作可由資訊科技團隊負責，此項工作對確保機構知悉合規要求至關重要。資訊科技團隊應定期進行政策檢討，以評估是否需要制定新的資訊科技標準。附錄B內的「相關行業標準範例」臚列制定中的行業標準例子。 • 機構應評估現行內部政策的涵蓋範圍，針對數據道德及人工智能應用領域，識別現有政策中的合規缺口或需關注事項，例如應檢視有否融入本地標準或規例（例如有關數據保護和私隱的法例）的相關要求。 • 應善用「人工智能應用影響評估」（詳見第5章「人工智能評估」），以識別人工智能應用會否損害個人或群體的權利，並確保從多方面評估有關人工智能應用的效益與風險。此舉旨在識別出正在發展中，而且可能影響內部人工智能政策及程序的社會或文化規範。

協助人員掌握所需技能與知識	• 為相關人員提供充足培訓，協助他們掌握所需知識、技能及意識，以有效配合機構的人工智能策略。培訓內容應涵蓋以下範疇：
	  ○ 人工智能治理與合規：開設培訓課程，內容包括偏見與公平、可解釋性、問責、安全、私隱等人工智能治理相關議題，確保符合適用法律、規例及政策要求，並提高對事實查核之重要性的認知，以維持人工智能系統的準確性與可靠性；
	  ○ 一般人工智能技術：按需要提供培訓課程，內容包括實際應用、常見應用案例及技術限制等；
	  ○ 人工智能工具應用：就機構開發或採用的人工智能工具提供適當培訓；以及
	  ○ 資訊科技安全：開設培訓課程，內容包括數據投毒、對抗性攻擊等與人工智能系統相關的資訊科技安全議題。

措施	措施詳情
	• 培訓計劃應獲定期檢視及更新，確保緊貼最新發展的最佳實務做法與快速演化的人工智能技術領域。
範例說明：為落實「合作與開放」的原則，機構可制定數據共享策略。透過運用已內建數據共享功能的中央數據庫，機構能夠開展符合此策略的分析及人工智能應用開發工作。	

相關人工智能道德原則：

1. 機構制定人工智能策略時，應充分考量並引用所有相關「人工智能道德原則」，作為機構認可該等原則的具體證明。
2. 問責—以政策界定的角色和責任釐定並顯示人工智能項目中各範疇的負責人員。
3. 數據私隱—確保人工智能項目遵守內部數據私隱政策。

4.1.1.2 行業標準與規例

定義：

須就相關規例及標準進行評估，以確保人工智能及相關流程符合任何適用的法律或標準。

此階段涉及的持份者：

資訊科技規劃人員 / 管理層

實務措施及示例：

措施	措施詳情
持續評估標準與規例的要求	• 國際標準組織和專業團體（例如國際標準化組織及電氣電子工程師學會）現正制定人工智能國際標準，預期完成制定後便可用作參考。詳情請參閱附錄B「參考資料」第B.1節。資訊科技團隊在有需要時擔當支援角色，負責研究並納入相關標準的要素。

措施	措施詳情
將標準與規例嵌入策略	• 機構應從營運角度入手，將數據道德及人工智能相關標準嵌入機構策略，確保人工智能項目充分考慮並符合各項標準與規例。 • 好的標準與規例合規策略應該是由上而下的。例如首席資訊官（負責監督所屬機構的資訊及通訊科技發展）及項目督導委員會應表明項目團隊執行人工智能項目時須遵循的標準（例如《私隱條例》），並強調不合規的影響，以確保項目團隊理解行業標準與規例的重要性。
範例說明：「人工智能應用影響評估」可用於確保各項標準與規例獲得充分考量。該評估將協助機構以符合道德的方式開發及部署人工智能應用，因為評估中的問題能讓機構全面檢視是否符合所有訂明的「人工智能道德原則」。	

相關人工智能道德原則：

1. 問責—標準及規例應能使機構及相關人員負有問責義務。
2. 數據私隱—必須遵守《私隱條例》及相關數據保護原則，以確保個人資料是在充分知情及公平的基礎上收集，並已充分考慮減少數據收集，包括確保人工智能使用個人資料的用途，必須與有關個人資料原先收集的目的相同或直接相關。
3. 合法與合規—確保項目考慮行業標準與規例，使機構及人工智能應用使用者知悉並能遵守相關標準與規例。

4.1.2 項目規劃

4.1.2.1 組合管理

定義：

組合是指一系列項目的集合。執行組合管理旨在確保機構流程、人員及技術中的各項資訊科技投資符合預期。

此階段涉及的持份者：

資訊科技規劃人員 / 管理層、業務使用者

實務措施及示例：

措施	措施詳情
建立人工智能項目清單	• 機構應建立人工智能項目清單，涵蓋現有項目、潛在項目及計劃中的概念驗證。 • 概念驗證是一項實驗性活動，旨在證明人工智能的可行性、效益及可能帶來的風險，以評估有關解決方案是否適用於處理實際問題。因此，概念驗證通常是小規模的測試活動，會作為支持大型項目的參考案例。
將人工智能項目對應至機構目標	• 根據各項目狀態進行分類，並評估這些項目是否支持機構當前的策略與目標。
篩選人工智能項目並訂定優先順序	• 機構優先考慮的人工智能項目須在風險與效益之間取得良好平衡、能帶來正面影響、較大機會成功，且有助於達成機構目標。機構可借助第5章「人工智能評估」中的「人工智能應用影響評估」記錄和確定項目的優先順序；同時，應評估現有資訊科技項目是否能透過採用大數據和人工智能技術進行整合或強化，從而實施改良方案，以造福機構及終端使用者。 • 明確界定人工智能項目的預期範圍，此步驟對於識別及選擇符合該範圍的合適人工智能應用至關重要。

	• 評估在預期範圍內應用人工智能技術可能產生的長遠及潛在影響，以防出現技術濫用。
範例說明：於網站推行人工智能應用（例如聊天機械人）可對應機構提升服務的目標。各機構應先釐清每項人工智能項目所對應的目標，據此訂定項目優先順序，並可優先選擇機構及終端使用者均能受惠的項目，以符合「有益的人工智能」原則之要求。	

相關人工智能道德原則：

1. 有益的人工智能—促進公共利益的項目可被賦予較高優先次序。
2. 公平—在項目層面，應根據項目目標，考量可能對不同群體產生的影響。
3. 多元與包容—受影響的持份者能夠相應在所有進行中的人工智能項目得到識別。

4.1.2.2 項目監督與交付方式

定義：

在項目啟動階段，必須決定人工智能項目是否需要接受資訊科技委員會 / 首席資訊官的審查。每個項目各具特點、功能及業務價值。為有效運作人工智能並確保問責性，必須明確界定各項角色與責任（詳見第3.5.2節「人工智能治理」）。項目經理應於「品質管理計劃」中界定主要交付成果的品質標準、品質控制與保證活動，以及驗收準則。品質控制不僅監控交付成果的品質，還涉及監控「項目管理計劃」中定義的項目各方面，以確保人工智能應用在整個生命週期均符合品質標準（例如：人工智能與數據道德、品質程序、行業標準及政府規例）。

此階段涉及的持份者：

資訊科技規劃人員 / 管理層、業務使用者

實務措施及示例：

措施	措施詳情

評估人工智能項目是否需要接受資訊科技委員會 / 首席資訊官審查的一般準則	若人工智能應用符合以下任一項準則，則必須提交資訊科技委員會 / 首席資訊官審查及審批。這些準則是基於《通用數據保障條例》中關於識別「可能導致高風險」情境的指引所制定。這些情境會隨時間演變，機構可根據各自推行人工智能應用的經驗調整準則，並提交資訊科技委員會 / 首席資訊官審查。有關風險分級準則的詳細說明，請參閱第5章「人工智能評估」。
	
	• 該人工智能應用具有高度自主性，意味着該人工智能應用能高度完成任務及與周圍的物體或人類互動。
	• 該人工智能應用會在複雜環境中使用，意味着該人工智能應用在執行過程中需要考慮環境中的許多變數。例如對於在公園等開放環境中使用的人工智能應用，周圍的移動物體和其他生物必須被人工智能應用納入考量。
	• 該人工智能應用使用敏感個人資料，意思包括任何可用於關聯和識別活人的信息。有關高敏感度的個人資料，例子包括身份證號碼、銀行帳戶數據和健康數據。
	• 個人資料被大規模處理和 / 或組合成數據集，需考慮：
	  ○ 涉及的個人數量，可以是具體數字或相關人口的比例；
	  ○ 處理的數據量和 / 或不同數據項目的範圍；
	  ○ 數據處理活動的時間或永久性；以及
	  ○ 處理活動的地理範圍。
	• 該人工智能應用可能對人類產生具爭議的影響（例如對個人做出決策或判斷其情緒）。
	• 該人工智能應用涉及對個人的評估或評分，方法可以是基於數據主體的工作表現、經濟狀況、健康、

措施	措施詳情
	個人偏好、可靠性、行為和位置等方面，對活人作出分析及預測。 • 該人工智能應用會進行自動化決策 / 複雜決策，對個人或實體會造成重大影響並產生法律後果，而有關決策屬未經人為干預的獨立決策。 • 該人工智能應用涉及系統性觀察或監控。這是一種用於觀察、監控或控制個人的處理手段，包括通過網絡或在公共可訪問區域的監控系統所收集的數據。
執行品質管制	• 品質保證包括基於人工智能治理的防線模型進行審查和持續監測。有關人工智能治理模型，請參閱本文件第3.5.2節「人工智能治理」。機構應定期評估人工智能治理架構和流程是否與不斷變化的人工智能要求保持一致，例如可能需要為新的人工智能技術開發新角色或新技能，並應建立協議、流程和程序，以管理和確保整個人工智能項目中的數據治理合規，在可用的情況下應遵守現有的數據治理流程和程序。
在人工智能生命週期中全面考慮所有「人工智能道德原則」	• 本文件提供貫穿人工智能生命週期的端到端實務建議以供參考。 • 應運用「人工智能道德框架」評估道德影響（同時參考「人工智能道德原則」），並於「人工智能應用影響評估」問卷中記錄相關考量。詳情請參閱第5章「人工智能評估」。 • 確保項目計劃着重道德多於發展和推行人工智能項目（例如：為提案中的人工智能模型制定公平性要求，並列明衡量與測試達標的相關方法）。關於公平性定義的詳細說明，請參閱第4.1.4.3節「數據提取」。項目計劃應顧及人工智能生命週期的每個階段。所有人

措施	措施詳情
	工智能項目（包括概念驗證）均須遵循現有項目管理規範。
確保滿足適當的資源需求	• 釐定推行人工智能所需的資源，包括考量合適技術與服務，以及除技術專長外兼具機構職能知識的資源。 • 有關技術相關考量，請參閱第4.1.3.1節「人工智能與數據使用技術路線圖」。 • 有關第三方人工智能服務相關考量，請參閱第4.1.3.2節「採購人工智能服務」。
確保關鍵流程設有監督機制	• 確保所有人工智能應用符合內部政策與流程要求，包括執行「人工智能應用影響評估」。 • 確保建立有效且嚴謹的流程，以測試人工智能模型是否符合公平性、避免偏見、防範對抗性攻擊風險及符合可解釋性標準。請參閱第4.1.4.1節「業務與數據理解」、第4.1.4.3節「數據提取」、第4.1.4.5.4節「模型過度擬合」、第4.1.4.5.6節「對抗性攻擊」及第4.1.5.1節「模型整合與影響」。 • 評估市場對人工智能應用合規要求的期望不斷變化（例如改變人工智能使用相關規例或標準）所產生的風險。請參閱附錄B「參考資料」第B.1節。 • 確保數據科學家接受完整培訓，涵蓋人工智能應用開發與部署，以及持續測試和驗證標準，使其能夠監督是否符合機構政策與策略目標。
與持份者建立溝通互動	• 與所有持份者（包括員工、人工智能供應商及終端使用者）保持清晰、誠實的溝通和有效的互動，以建立信任和促進人工智能道德應用，並避免誇大人工智能的能力。

措施	措施詳情
	• 盡早與人工智能供應商建立積極溝通渠道，確保清楚了解產品功能、限制條件及相關要求。詳情請參閱第4.1.3.2節「採購人工智能服務」。 • 披露人工智能技術的使用情況，並就人工智能應用的用途、限制條件及相關潛在風險提供清晰說明。詳情請參閱第4.1.5.2節「向用戶提供資訊」。 • 如有需要，機構應主動與監管機構保持充分溝通。
界定角色與責任	• 機構依然須為人工智能應用所產生的道德問題問責和負上法律責任。機構須為其開發、部署或使用的人工智能系統的行動、決定和輸出負責。 • 與使用者一起界定清晰的使用條款，列明有關人工智能服務的責任鏈（如適用）。 • 各項角色的責任，應界定至適合有關人員履行職務的程度。 • 項目管理計劃應包含明確闡述的角色、責任及時間表，以促進人工智能的採用。 • 為確保人工智能及數據活動的最終決策權獲得明確指定，妥善規劃角色與責任至關重要。 • 建議機構在可行情況下採用「三道防線」模型。有關角色與責任的更多詳情，請參閱第3.5.2節「人工智能治理」。這些角色包括構成項目團隊的項目經理、系統分析員、系統架構師及數據科學家。
範例說明：若人工智能應用涉及使用影片，相關人員或需具備以專業技術分析影像的經驗，並需在影像數據處理過程中保持道德意識，考慮所有相關的「人工智能道德原則」。	

相關人工智能道德原則：

1. 問責—透過人工智能評估、分配品質控制責任，以及在項目中考量道德因素以確保問責。
2. 人為監督—可將人為干預納入項目與監督流程中的審查。
3. 可持續性與公正轉型—若需減緩任何已識別的不良影響，可將緩解策略與進度或預算一併核實。
4. 合作與開放—機構應主動與內外部持份者互動，以建立信任並促進人工智能道德應用。

4.1.3 項目生態環境

4.1.3.1 人工智能與數據使用的技術路線圖

定義：

技術路線圖應能使機構能夠規劃及制定策略，決定人工智能與大數據分析選用的技術、時間表和將予採購的技術項目。有效的技術路線圖須概述實現數字轉型目標的策略方針。

此階段涉及的持份者：

資訊科技規劃人員 / 管理層、數據科學家、資訊保安管理人員、系統分析員、系統架構師、業務使用者

實務措施及示例：

措施	措施詳情
建立合適的技術架構	• 機構在選擇符合其目標、合規需求及預算的基礎設施時，應諮詢所屬資訊科技團隊（或同等單位）。這包括使用機構內部現有可用的基礎設施，例如中央數據分析平台。 • 機構在考慮數據收集與儲存、數據處理、數據建模、執行及部署所需的可擴展技術架構時，應諮詢所屬資訊科技團

措施	措施詳情
	隊（或同等單位），以確保解決方案具備擴展能力及使用和處理所需數據的資源。有關技術架構應能隨着請求數量與數據增長而自動調整，以維持運算速度與可靠性。所有人工智能及大數據應用均應以可擴展性和服務可用性為設計考量。
執行數據評估	• 機構應評估項目需要什麼數據及相關數據是否可供項目使用。相關數據的可用性與品質是所有人工智能應用的先決條件，因為這些數據需用於訓練或微調人工智能模型（在人工智能應用內），以及進行預測。 • 機構應明確識別將作為人工智能應用一部分而收集、追蹤、傳輸、使用、儲存或處理的具體數據類型及數據來源，並判斷所涉數據是否屬敏感性質。 • 機構應記錄數據沿襲，以了解將用於人工智能應用的數據的來源、路徑、許可證或其他義務，以及數據轉換過程。
範例說明：影像分析技術可運用錄影片段或即時影像。界定數據（例如生物特徵資訊）的收集、儲存及處理方式與內容，是數據評估重要的一環，有助於在整個數據生命週期的各個階段中，從數據收集到數據銷毀均遵循「數據私隱」原則。	

相關人工智能道德原則：

1. 可靠、穩健及安全—所採用技術的安全性必須確保人工智能應用具備安全保障。
2. 可持續性與公正轉型—所採用的技術架構須可以擴展，並須針對可能影響人工智能應用的故障制定緩解措施。

4.1.3.2 採購人工智能服務

定義：

機構可透過外判安排推行人工智能項目，並可為項目採購現成產品甚至外部數據。在進行此類採購時，機構應適當考慮相關道德因素。

此階段涉及的持份者：

採購團隊、項目經理、業務使用者

實務措施及示例：

措施	措施詳情
為採購工作界定人工智能應用的要求	**人工智能採購規劃** • 機構應評估是否具備合適的採購團隊（包含採購人員及能提供多元專業意見的項目團隊）。舉例而言，項目團隊可包含領域專家、系統與數據工程專家及安全專家等不同專業人員，確保能選擇出具備適當技術並符合技能要求的供應商。 • 機構應配置適當專業人員，負責採購人工智能應用後的變更管理交付、培訓及交接工作，並可為資訊科技人員提供培訓提升能力，以確保所有相關人員均具備達成人工智能目標與履行職責所需的技能與知識（包括理解人工智能算法的概率特性）。 • 機構可考慮在採購人工智能應用的簡介或規格中，適當納入投標者及／或其所聘人員須具備人工智能開發經驗的要求。 • 機構應就人工智能應用的目標，訂立清晰的問題陳述。 • 服務要求視乎情況可包含對模型與道德風險的考量、對業務目標或策略的重要性，以及模型可解釋性等要素。 **了解第三方產品與服務** • 由於人工智能系統通常涉及複雜的算法與建模技術，機構在外判人工智能項目的推行工作時應格外謹慎。機構應要求並審查相關文件，包括算法設計規格、人工智能應用所基於的編碼與技術、其輸出結果、持續支援，以及對建議人工智能應用的監測或維護。使用第三方服務時需檢查的範疇包括：

措施	措施詳情
	○ 運作與預期用途； ○ 限制與禁止事項； ○ 訓練程序； ○ 測試程序； ○ 功能與效能指標； ○ 為評估道德價值觀念（公平；可靠、穩健及安全等）而曾進行的檢核； ○ 模型配置的存取權限；以及 ○ 人工智能應用中潛在的漏洞、風險或偏見。 • 在使用人工智能產品與服務前，應先了解其數據處理、安全性、保安及私隱保護措施。 • 查閱人工智能供應商提供的使用條款或責任聲明，以釐清人工智能服務的責任鏈（如適用）。 **確保第三方符合現有標準** • 項目團隊應留意人工智能供應商是否符合現有和適用的本地及國際標準。目前國際標準組織或專業協會（例如國際標準化組織及電氣電子工程師學會）正制定人工智能相關標準，可供日後參考。 • 舉例而言，現時已有專門針對具機器學習功能醫療器材的標準，例如： ○ 中國國家藥品監督管理局醫療器械技術審評中心的《醫療器械真實世界研究設計和統計分析註冊審查指導原則》[12] ○ 美國食品藥物管理局的《Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)》[13]

[12] https://www.cmde.org.cn/xwdt/zxyw/20240115162832179.html
[13] https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf

措施	措施詳情
	○ 國際標準化組織標準，例如ISO 13485:2016《Medical devices—A practical guide》 • 國際電工技術委員會標準，例如IEC 62304 **評估第三方人工智能應用的適用性** • 對於供應商提供的人工智能應用在機構擬定使用場景中的適用性，應進行研究、測試及評估。 • 舉例而言，當對外採購先進分析型人工智能應用（例如影像分析系統）時，應確保該應用內的人工智能模型（如人臉識別功能）符合機構的目標和原意，並遵守私隱與安全規例。同時，亦應確保人工智能應用及數據使用方式符合規定及私隱等「人工智能道德原則」（例如《私隱條例》）。 **避免鎖定供應商** • 部分供應商提供的人工智能應用可能屬「黑盒」，其分析數據然後得出結果或決策的過程，往往難以追溯。機構應強調模型可解釋性的必要性，並要求供應商就人工智能應用內的人工智能模型提供相關文件（例如算法設計規格、人工智能模型所基於的編碼與技術及其輸出結果）以供審查。此舉讓機構日後亦能與其他供應商合作，持續使用或優化現有人工智能應用，而不會受制於單一供應商。 • 若供應商基於商業秘密拒絕公開人工智能算法文件，機構需作出業務決策，考慮供應商拒絕披露一事涉及的風險及對機構業務可能引致的後果。機構應（通過各自的項目管治機構）就每次採購的具體情況進行審議，並記錄決策理據、考量因素和最終決定。 • 在可行情況下，應確保人工智能應用的數據模型及格式跨平台通用，避免採用供應商專屬規格。此要求可向個別供應商確認落實。

措施	措施詳情
	評估供應鏈風險與多樣性 • 人工智能產業依賴高度全球化的供應鏈。然而，部分國家或地區可能實施單邊強制措施（例如技術壁壘與出口管制），有可能造成發展障礙並擾亂全球人工智能供應鏈，或會導致硬件、軟件及工具供應中斷。 • 機構應全面評估人工智能應用所採用的主要組件（包括晶片、軟件、工具、運算基礎設施及數據來源）的供應鏈多樣性，同時必須監測和及時修正軟硬件產品的漏洞與缺陷，以確保系統安全。 **釐定知識產權歸屬** • 為保障知識產權，部分第三方供應商可能拒絕披露開發人工智能應用所採用的算法與技術。當第三方供應商為機構設計開發新人工智能應用時，誰是新知識產權的擁有人，應依據該機構的採購程序共同協定。 • 機構應與第三方人工智能供應商明確議定，解決方案一經採購後，由誰管制有關人工智能應用及有關解決方案所處理的數據。相關權責應於合約或服務水平協議中明確規定。
執行風險緩解程序	• 第三方供應商的信譽與可靠程度，間接影響人工智能應用的可靠性與穩健性。 • 作為盡職審查流程的一部分，應對第三方供應商進行獨立評估，不論有關人工智能應用屬定制、現成或承包。 • 評估可包括基本公司資料核查，例如：商業執照、官方推薦、全球制裁名單、執法機構名單、過往經驗 / 資歷和資格認證。若供應商未能通過這些核查，機構需作出業務決策，考慮供應商可靠程度所涉及的風險及對機構業務可能引致的後果。機構應（通過其各自的項目管治機構）就每

措施	措施詳情
	次採購的具體情況進行審議，並記錄決策理據、考量因素和最終決定。 • 機構可建立流程，以書面記錄與第三方的往來，並留存已完成的盡職審查證明，包括就所取得及審查與第三方供應商有關的文件保留證據。 • 應考慮並記錄各項風險因素，包括第三方與人工智能應用的任何互動、對模型配置檔案或端點的訪問權限、對模型的訪問權限及使用方式。
範例說明：採購影像分析解決方案時，機構需了解第三方的實施方法。在招標文件中要求承辦商遵守相關安全標準，並確保其建議的設計考慮到各項安全議題，例如使用遮罩技術、匿名化處理及保持符合數據私隱標準，有助遵循「數據私隱」原則。	

相關人工智能道德原則：

1. 公平—第三方應能證明其解決方案為何被視為「公平」。
2. 可解釋性—應盡可能避免使用「黑盒」人工智能應用。
3. 可靠、穩健及安全—第三方提供的人工智能應用應可靠、穩健及安全。

4.1.4 項目開發

4.1.4.1 業務與數據理解

定義：

開發人工智能應用時，機構應確定使用人工智能的目標，並權衡決策過程中使用人工智能的效益與風險。

此階段涉及的持份者：

項目經理、業務使用者

實務措施及示例：

措施	措施詳情
界定業務需求	• 應制定並貫徹執行標準流程（如「人工智能應用影響評估」，詳見第5章「人工智能評估」），確保在考慮潛在應用影響及數據使用的同時，顧及「人工智能道德原則」、策略及相關法律要求。 • 應設有標準流程界定及驗證特定人工智能項目與應用的性能及業務成功標準。此類標準範例包括：衡量分類模型的效能指標，如真陽性率、假陰性率或假陽性率。高真陽性率表示人工智能應用按預期運作且結果非常準確；高假陽性率代表負面案例被錯誤預測為正面，人工智能應用可能需要進一步微調以產出準確結果。高假陰性率則表示正面案例未被正確納入預測結果。鑒於沒有一套指標能適用於所有人工智能應用（因較高假陰性率可能導致較低假陽性率，反之亦然），項目團隊應為個別人工智能應用量身定制有關定義。 • 機構應徵詢內外部持份者意見，以確定人工智能應用所需的功能需求、非功能需求及品質保證程序。 • 機構應考慮現有或擬議人工智能應用的潛在替代方案，比較各方就準確度指標的表現及對業務、法律、經濟或社會的影響。例如就營運角度而言，替代方案是否比人工智能應用更準確且更經濟？模型應盡量減少假陰性還是假陽性？這將有助機構判斷相較於擬議人工智能應用的效益，採用替代方案的風險會否低於採用該人工智能應用。 • 整體人工智能應用開發流程應包含迭代式影響評估（請參閱第5章「人工智能評估」中的「人工智能應用影響評估」），其中需涵蓋測試要求，並確保人工智能項目符合機構的整體風險管理方針（例如符合所有內部政策及「人工智能道德原則」）。

相關人工智能道德原則：

1. 合作與開放—機構應徵詢內外部持份者意見，以確定人工智能應用所需的功能需求、非功能需求及品質保證程序。

4.1.4.2 解決方案設計

定義：

開發人工智能應用時，機構應確立設計需求，包括評估數據與技術的適用性，以及所需人為介入程度等多方面因素。風險較高的決策流程應納入更高程度的人為介入。

此階段涉及的持份者：

項目經理、業務使用者

實務措施及示例：

措施	措施詳情
評估人工智能模型適用性	• 應評估人工智能應用中的模型是否適合機構目標，並建立流程以在必要時說明模型決策或建議的依據。影響模型決策的因素可包括： ○ 業務目標； ○ 模型準確性與可解釋性； ○ 模型開發與訓練時間； ○ 模型擴展性； ○ 模型複雜程度；以及 ○ 資訊科技架構或基礎設施。 • 應建立系統性審查流程，評估機構現有技術架構對人工智能模型設計的適用性。適用性評估應包含下列活動：

措施	措施詳情
	○ 了解模型使用的數據集特性：不同模型對數據類型與規模有不同要求。例如樸素貝葉斯或線性回歸算法適合小規模數據集，大規模訓練數據則可能更適合決策樹算法。應分配足夠訓練數據以將模型可靠性調校至滿意水準（訓練數據、測試數據與驗證數據的比例詳見第4.1.4.3節「數據提取」）。 ○ 確定所需學習算法的類型，有關分類如下： ▪ 監督式學習：適用於輸入與輸出能配對的訓練數據集。在訓練過程中，模型會使用已配對正確答案的訓練數據集學習。監督式學習算法的例子有回歸分析與異常檢測。 ▪ 非監督式學習：適用無對應輸出標籤的數據集。非監督式學習算法會嘗試在數據中識別模式和隱藏的結構，例子有聚類分析與降維。 ▪ 強化學習：採用獎懲機制，透過獎勵函數決定模型所作行動的獎罰並訓練模型，引導模型按汲取的經驗而非通過預測得出結論。部分人工智能應用或不適合使用此技術。這項強化學習技術能訓練模型基於現有獎勵信息作出反應。 ○ 理解系統限制條件：理解關鍵限制，包括模型擴展性、儲存容量與開發時程等。
決定適當的人為介入的程度	• 評估人工智能應用是否需要人為介入，是「人工智能應用影響評估」的其中一環。機構應考慮人工智能決策將對個人或機構造成何等程度的影響，以及下列因素： ○ 人工智能應用在決策過程中對人類造成傷害或影響人類安全的機會率； ○ 在人為介入下決策對比無人為介入下決策的運作成本；

措施	措施詳情
	○ 沒有「人類」參與的自動化決策可能產生的法律影響或類似後果；以及 ○ 社會規範與價值觀演變，包括考慮社會整體對由人工智能應用全自動進行決策的接受程度。 • 機構應決定人工智能應用需要人為介入的程度。一般而言，在決定適當的人為介入程度時，有三種不同模式，分別是「人在環外」、「人在環內」和「人為管控」。「人在環內」和「人為管控」模式可被視為輔助型智能或增強型智能，而「人在環外」則可被視為自動化智能或自主智能。人為介入的程度應與人工智能應用相關的影響程度相稱。影響程度越高，人為介入程度理應越高。 1. 「人在環外」：適用於決策對經濟、社會或法律影響輕微，或實務上難以人為介入的場景，評估有關影響時可使用「人工智能應用影響評估」。舉例而言，當需要在有限時間內迅速決策且影響有限，「人在環外」模式便相當適合。然而，即使決策過程不涉及人為介入，人工智能應用的決策仍必須能夠被人類解讀。 2. 「人在環內」：適用於數據有限而人類能做出更佳判斷的情況。 3. 「人為管控」：「人在環內」及「人在環外」兩種模式均不適用時，可由人類監督控制人工智能應用的整體運作，並決定介入時機與方式。此模式允許人類調整可能影響人工智能應用決策結果的變數。 • 確認人工智能應用需執行的自動化決策類型。當自動化決策會產生法律影響且無人為介入時，應建立流程確保受影響人士獲得通知。使用者應預先獲悉其與人工智能的互動關係，此舉旨在提升透明度並取得使用者同意（例如披露

措施	措施詳情
	聲明有關自動化決策完全由人工智能應用執行，使用者同意後方可繼續）。 • 人工智能機械人在工作中是輔助性角色，其主要職責是協助使用者及使用機構完成其所交辦的工作事項。在高風險的應用（如醫療診斷系統、自動駕駛、政務決策等），人工智能機械人原則上不得進行獨立決策，所有決策制定均需在使用者監督和指導下進行。另可參閱《香港生成式人工智能技術及應用指引》。
訂定人工智能模型透明與可解釋的要求	• 應制定有關人工智能模型可解釋性的溝通計畫，向受影響持份者展示各變數對預測結果的相對重要性，以達成透明與可解釋的目標。 • 應盡可能提供實例，透過敘述性說明或圖像（例如繪製決策樹工作流程）解釋人工智能決策過程，讓沒有技術背景的人士更易理解與想像人工智能的運作方式。機構須明確說明不同因素與數據如何影響其人工智能模型產出的結果與結論。常人一般可接受的推論方式可包括： ○ 邏輯推導； ○ 社會規範、慣例與觀念；以及 ○ 道德理由。
範例說明：為遵循「人為監督」原則，釐定適當的人為參與程度至關重要。「人在環內」模式常應用於電腦系統安全威脅偵測的自動化流程。當今網路安全受到的威脅不斷變化，會產生人工智能應用無法偵測的新型攻擊模式。在此情況下，仍需由安全專家配合人工智能應用執行安全威脅偵測，並驗證人工智能應用所產出結果的準確性。 採用預測性鐵路維護系統（可自動觸發工單或訂購維修零件）時，可考慮「人在環外」模式，但此模式可能僅限於通知、編排優次及處理訂單，因實際維護工作可能仍需人員親身進行。	

措施	措施詳情
當涉及更廣泛的經濟、社會、法律或道德影響時（例如醫療手術），則應採用「人為管控」模式。	

相關人工智能道德原則：

1. 人為監督—應根據多項因素確保適當程度的人為介入，這些因素包括：人工智能應用的效益與風險、人工智能決策的影響、運作成本，以及不斷演變的社會規範與價值觀。
2. 問責—確保高風險的人工智能應用設有問責機制，包括在決策過程中需有人類參與。

4.1.4.3 數據提取

定義：

人工智能模型依賴來自機構內外各種來源的資訊，這些資訊在數據品質、有效性、可靠性和一致性方面均可能涉及高風險。從多個來源提取大量數據，並根據人工智能模型的需求轉換和載入有關數據，通常會使用「提取、轉換和載入」工具。執行複雜的數據清理和轉換步驟時，容易意外造成用戶錯誤，這些錯誤難以識別，並可能導致錯誤的建模結果。數據完整性是確保數據公平性的必要組成部分。數據完整性確保人工智能應用生成的結果不是出自帶偏見或歪曲的數據集。

此階段涉及的持份者：

系統分析員、系統架構師

實務措施及示例：

措施	措施詳情
使用或建立集中化、標準化的數據層	• 在適當情況下，利用機構的中央數據儲存庫作為人工智能應用的通用數據註冊庫。該儲存庫應設有定期維護的存取控制機制。 • 善用數據倉庫、數據湖或數據市場等數據儲存庫作為「單一數據真源」。此類數據儲存庫可為人工智能應用開發提供經過清理的數據源。
執行數據驗證	• 當從多個來源獲取數據時，須驗證數據格式的一致性。 • 進行評估以確保數據品質與特性符合業務及人工智能應用目標。常見的數據驗證規則例子包括：驗證數據類型或格式、唯一鍵檢查、重複數據檢查、縮寫須一致以及防止非預期的空值。 • 識別數據集中是否含有個人資料，並確保符合監管機構制定的個人數據使用政策。 • 謹慎選擇並清理訓練數據，以防止侵犯知識產權、排除敏感數據，並降低數據投毒攻擊的風險。
確保公共或第三方數據的可信度	• 核實所獲取數據的可靠性、代表性及相關性。確認數據可靠性的例子包括對照原始來源核實資訊。若存在獨立可靠的數據來源，可進行進一步核實。 • 審查源數據內容，通過評估其與項目目標的相關性及是否涵蓋主題事項，判斷是否符合人工智能模型的需求。 • 評估數據來源供應商及／或機構的可靠性等級。詳情請參閱第4.1.3.2節「採購人工智能服務」。數據可靠性通常取決於具體情境，例如依賴所需數據所作決策的重要程度。以用於影響範圍有限的簡單營銷應用的數據及用於決定誰不能登機的數據為例，用於後者的數據通常預期比用於簡單營銷應用的數據更可靠。

措施	措施詳情
為訓練數據標註工作制定標準	• 為標註人員提供全面的數據標註指引及充足培訓。標註指引不完善、標註人員訓練不足，以及標註錯誤，均會對模型及算法的準確性、可靠性及有效性造成負面影響。此外，這些問題可能導致訓練偏見、加劇歧視、降低泛化能力，並產生錯誤輸出。
完成數據來源文件記錄	• 應制定標準要求或流程，明確界定並記錄人工智能模型及應用開發的過程所收集、追蹤、傳輸、使用、儲存或處理的數據類型。 • 作為良好的數據治理實踐，應追蹤並備存有關數據來源、擁有人、許可證及關鍵假設記錄的資訊。為人工智能模型從各個來源收集或取得各項訓練、測試及運作數據的工作，應由系統分析員及系統架構師負責監管。
提升透明度與可追溯性	• 通過文件記錄或採用以視覺呈現數據沿襲的工作流程方式（例如透過編寫腳本），確保「提取、轉換和載入」流程具透明度，以便機構、系統分析員及系統架構師理解數據流向。
啟用錯誤處理機制	• 為「提取、轉換和載入」的解決方案開發錯誤處理機制。該機制應能捕捉「提取、轉換和載入」任務名稱、錯誤編號及錯誤描述，並將其記錄於獨立表格或文件中，以便項目團隊分析問題並進行修正，例如設置警示機制，當「提取、轉換和載入」流程發生錯誤時，透過電子郵件通知項目團隊成員。

措施	措施詳情
範例說明：針對影像分析應用，為確保公共或第三方數據的可信度，應根據影像元數據確認影片來源是否有效，並核實拍攝地點、拍攝日期 / 時間、持續時間等資訊。可對影片進行額外分析，透過檢測人臉是否遭受對抗性攻擊而遭篡改，來驗證來源有效性。此做法有助遵循「可靠、穩健及安全」的原則。 為遵循「可靠、穩健及安全」的原則，機構應啟用錯誤處理機制。以閉路電視人流計數人工智能模型為例，應使用實際影像進行測試，並在程式碼變更時重新測試。從影片提取的數據可分拆為不同提取階段進行錯誤識別（例如評估各提取階段是否遺漏特定物件）。	

相關人工智能道德原則：

1. 公平—數據來源的完整性對於確保結果公平至關重要。
2. 可靠、穩健及安全—所收集的數據應完整且準確，以確保用於模型訓練數據的可靠性。
3. 可靠、穩健及安全—「提取、轉換和載入」工具在執行數據清理和轉換時必須安全且穩健，以避免在後續開發階段出現錯誤。

4.1.4.4 數據預處理

定義：

在為人工智能建模進行的數據處理程序中，所處理的數據在樣本、規模和分布方面須公平且適切，以確保人工智能應用能得出有意義且具代表性的推論。數據集中若存在代表性缺陷（例如數據樣本代表性過高或代表性不足），可能導致訓練完成的人工智能模型產生偏頗結果。

在解決方案開發過程中，包含個人資訊的個人數據及敏感數據須特別謹慎處理，以防範數據洩漏及違反私隱與安全政策。應盡可能減少所收集、處理及使用的個人數據量。

註：關於人工智能或大數據等資訊科技項目中使用個人數據與私隱保護的相關考量，機構應參閱《私隱條例》指引。

此階段涉及的持份者：

項目經理、系統分析員、系統架構師、數據科學家

實務措施及示例：

措施	措施詳情
強化數據安全與保護	• 確保操作符合數據安全與保護要求，特別是在處理個人數據和機密數據時，應使用加密技術及其他保護措施。 • 落實妥善的數據存取管制措施、制定穩健的數據備份與復原計劃，並定期檢查數據存取記錄。 • 提升網絡安全及供應鏈風險等領域的能力，以降低面對網絡威脅的可能性或對單一來源的依賴性。
執行私隱影響評估	• 進行私隱影響評估以確認收集特定個人數據類型、數量及範圍用於人工智能應用的必要性。
實施數據匿名化	• 採用假名及完全匿名化等保障措施，防止個人數據與可識別的個人產生連結。數據匿名化是通過加密、遮罩及聚合任何能連結個人與儲存數據的資訊，來保護敏感資訊的過程。 • 重申保持匿名化數據的效益大於該數據被用於識別個人所帶來的潛在風險，以及此類去識別化對個人的影響。因此，項目團隊應定期審查匿名化數據是否可能被重新識別，並採取適當措施保護個人數據。 • 應進行類似的效益與風險分析，評估數據去識別化對數據實用性的影響。例如，在影像分析中模糊人臉是

措施	措施詳情
	一種數據匿名化形式，可減少類別偏見並保護個人身份。 • 執行數據匿名化時，應將實現此過程的步驟記錄在日誌程序中。
移除非法或有害資訊	• 應移除虛假、偏頗及侵犯知識產權等非法或有害資訊。訓練數據可能遭受竄改、錯誤注入或攻擊者的誤導行為而導致數據污染，此類情況可能扭曲模型的概率分布，降低其準確性與可靠性。

執行數據最小化與聚合	• 當數據匿名化不可行或無須執行時，機構應盡量減少人工智能應用可存取的個人數據量，並評估可否採用數據密集度較低的方法，以達成有關活動的目標。
	• 數據最小化涉及識別如何以最少的數據量，實現人工智能應用的目的和目標。
	• 項目團隊可透過以下方式確保收集的數據符合最小化原則：
	  ○ 足以達成目標；
	  ○ 與目標具相關性及合理關聯；以及
	  ○ 僅限必要範圍。
	• 舉例而言，若詳細數據屬不必要，可執行數據聚合（整理與概述數據的過程）以降低數據細緻度。

建立單一分析數據環境	• 應在收集數據的同一伺服器或雲端環境內的安全工作區域處理及分析敏感數據，避免數據傳輸時的洩漏風險。此外，機構應盡可能避免使用臨時性不安全環境，以確保在整個項目週期中維持適當的數據處理程序。 • 針對所有收集的數據（例如個人數據與影像數據）實施並遵守保留政策。

措施	措施詳情
識別異常值與異常數據	• 透過程式或繪製圖表（例如箱形圖、散點圖或直方圖）等形式，檢測原始數據集與最終建模數據集中的異常值。異常值與異常數據可定義為偏離數據集平均值一定距離的數據點。若檢測到的異常值與異常數據存在錯誤或無法準確代表整體數據分布，應予以移除或以合理值修補，因其在監督式學習人工智能模型中可能造成誤導，導致訓練時間延長或模型精確度下降。 • 當數據集包含大量變數時，識別與移除異常值可能具有挑戰性。此類情況下，可在建模流程中使用自動異常值檢測算法，例如孤立森林、最小協方差行列式與局部異常因子等算法。
界定與測試公平性及偏見	• 公平性可從群體、個人或第三方代理的角度理解。機構應根據其人工智能應用的情境明確界定公平性，以便測試偏見與公平性。例如，人臉識別系統應展現統計（平等）對等性，才能準確區分不同種族的人士。數學公平性定義範例如下： 1. 統計（平等）對等性—兩個群組間的正向分類率應相等。 2. 平等機會—數學上意指所有群組的真陽性率應相等。 3. 整體準確性平等—兩個群組在給定模型下應呈現相似的準確度測量值。 4. 堅尼平等性—堅尼系數的定義為接收者操作特徵曲線與對角線之間的面積比率，以及相關三角形的面積。本質上，模型應在所有群組中同等滿意地將 1 分類為 1及0 分類為 0。 • 當用於訓練人工智能模型的數據存在偏見時，偏見可能影響人工智能應用。機構可運用偏見檢測方法評估

措施	措施詳情
	模型是否歧視特定群組或個人（基於測試數據集），從而減輕固有偏見。偏見檢測方法範例包括： ○ 比較效能指標—透過計算正確與錯誤預測的不同比率，分析模型在各群組的表現。然後計算數據集中各可能群組的效能指標（例如真陽性率、假陽性率、真陰性率或假陰性率）。特徵分布圖等輸出可識別群組間的偏斜情況。 ○ 測量公平性—可根據標準化處理品質、統計對等性等公平性定義，對弱勢群組與優勢群組進行比較分析。棒形圖與熱力圖是可呈現群組間差異的輸出範例。 ○ 檢測代理變數—可透過計算所有特徵與不同敏感屬性（例如年齡、性別與種族）的相互信息，了解它們如何可能成為代理變數。
修正不平衡類別問題	• 不平衡類別是指數據在不同類別間分布不均所導致的分類問題。例如，在實施處理申請的人工智能系統時，應提供足夠數量的獲批與未獲批申請個案歷史數據來訓練模型。這意味着數據集中的類別或分類並未平等呈現。人工智能預測建模涉及從數據集預測給定觀察值的類別標籤。當出現不平衡類別問題時，會導致模型預測性能不佳（特別是對少數群組的類別），從而造成類別歧視。 • 將數據分布與全數據集人口統計進行基準對比，以量化被測試數據的代表性。此舉有助於減輕模型開發所用數據的偏見。 • 例如，使用隨機森林、邏輯回歸與決策樹等標準算法，確保建立能預測有限類別的良好分類模型，並確保訓練數據包含來自不同群組的平等數據樣本。 • 當不平衡數據集導致不平衡類別問題時，可採用以下取樣技術處理：

措施	措施詳情
	○ 隨機重新取樣—重新平衡不平衡數據集的類別分布； ○ 隨機減少取樣法—從多數類別刪除數據集實例，使模型損失部分信息以達成與少數類別可比擬的平衡數據集；以及 ○ 隨機增加取樣法—複製少數類別的數據集實例，以達成與多數類別可比擬的平衡數據集。
比較訓練數據、驗證數據與測試數據	• 當人工智能模型經調校至滿意水準後，可使用測試數據集評估學習模型，以衡量模型在未見數據上的表現。訓練數據是用於訓練人工智能模型的樣本數據。若涉及多個模型或超參數[14]調校，驗證數據可用於評估不同模型在開發期間的擬合度。測試數據可視為從人工智能模型中隔離保留的數據，其用途是於訓練階段的最後時刻進行最終建模測試。 • 一般而言，數據比例如下： ○ 訓練數據：測試數據：驗證數據通常設為 60:20:20 或 70:15:15。若未使用驗證數據，訓練數據與測試數據的建議樣本量比例為 80:20。 • 分析並比對訓練數據與驗證 / 測試數據，確保用於評估人工智能模型的數據結果與建模所用數據具等效性。此舉旨在確認當採用新的未見測試數據時，人工智能模型的產出結果仍具可比性。例如，若模型基於訓練數據能將人像圖片分類為「人類」，則面對測試數據中的不同個體圖像時，其表現應達到同等水平。 • 建立或使用包含訓練與驗證 / 測試數據的程式庫，用以評估潛在無意識偏見。例如，可運用Python深度學

[14] 超參數是指其數值用於控制機器學習模型學習過程之參數，例如隨機森林模型的最大深度。

措施	措施詳情
	習程式庫[15] [16] [17]檢測特徵、標籤與群組間可能存在的偏見或歧視性關聯。
範例說明：在影像分析中實踐數據匿名化的具體作法，是對人臉進行模糊或遮蓋處理（當無必要識別個人身份時），以確保無法從影片中辨識個人。此作法符合「數據私隱」原則。若分析目的僅為測量人流數量，則應將數據提取與儲存內容限制於總數值。建議在數據收集階段即進行人臉模糊處理，且不提取個人行為與活動等資訊。 為遵循「公平」的原則，機構可分析所用訓練數據以評估其代表性是否公平。以測量人流為目的的影像分析為例，須確保人工智能模型能正確計算不同人群組別的人流數量。訓練數據應包含來自不同群組中個體的平等數據樣本，以確保各類人士均能被識別。當使用閉路電視系統進行人流測量時，所用訓練數據應涵蓋多種場景，以確保模型輸出結果能準確達成預期目標。	

相關人工智能道德原則：

1. 數據私隱—確保遵循減少收集與處理個人資料的原則及保留政策，並遵守《私隱條例》等私隱規例及相關指引備忘錄。
2. 合法與合規—處理數據時應確保適當考慮及／或遵守本地與國際數據私隱及保護標準，如《私隱條例》、內地《個人信息保護法》或歐盟《通用數據保障條例》。
3. 公平—應避免數據中特定群組或類別出現代表性過高或代表性不足的情況，以免導致偏頗且不公正的結果。

[15] https://github.com/dssg/aequitas
[16] https://aws.amazon.com/sagemaker/clarify/
[17] https://fairlearn.org/

4.1.4.5 模型構建

模型構建涵蓋以下領域，將於後續分節詳細討論：

- 模型假設
- 模型目標與誘因
- 輸入變數選擇
- 模型過度擬合
- 模型訓練擁有權
- 對抗性攻擊

模型客製化指調整預訓練人工智能模型（通常由第三方開發）以滿足特定需求的過程，而非從零開始構建模型。此過程可能涉及對人工智能模型進行再造工程、微調或接地，以提升其性能並符合特定業務需求。若基礎模型存在安全缺陷，將導致風險傳遞至下游模型。機構應警惕從基礎模型傳遞的風險，並遵循本節所述的模型客製化實務規範。

由於機構往往將模型客製化工作外判予外部服務承包商，建議在考慮此類採購時，對潛在服務供應商進行全面評估。有關採購人工智能服務或產品的實務規範與考量細節，請參閱第 4.1.3.2 節「採購人工智能服務」。

4.1.4.5.1 模型假設

定義：

模型假設是指進行相關建模分析前，模型應滿足的前提條件。建模前必須驗證支撐模型的各項假設，以確保能獲得準確的詮釋與結論。

此階段涉及的持份者：

項目經理、數據科學家

實務措施及示例：

措施	措施詳情
執行嚴格測試	• 制定明確的測試規則與方法，並運用沙盒模擬等技術測試及核實人工智能模型。 • 必須在開發階段測試統計及其他數據假設，以確保模型預測與洞察結果的有效性。 • 若未審查假設並核准經訓練的人工智能模型，將導致錯誤或高風險的假設。 　○ 項目團隊應進行測試並記錄結果以核實人工智能模型的假設。測試範例包括算法測試與回歸測試。若未記錄模型結果與配置，將增加後續階段管理及維護人工智能資產的困難度。日誌應包含以下資訊： 　　▪ 人工智能應用執行的作業 / 操作（包含已執行的算法）；以及 　　▪ 決策內容與執行結果。 • 透過日誌監控人工智能的決策，並與人為決策或實際結果比對，同時記錄所採取的應對措施。當需要重新訓練模型時（例如準確度過低），應完整記錄變更理由及受影響的特徵參數。
範例說明：機構應執行嚴格測試以評估模型預測與洞察結果。此舉旨在檢驗是否準確地符合「公平」的原則。以天氣預測模型為例，若數據僅適用於特定地點，則預測不同地點的天氣時，模型的實用性將受限。必須檢驗數據中所假設地點的適用性。	

相關人工智能道德原則：

1. 公平—為實現結果公平性，所有數據假設均應經過測試與準確性核實
2. 可靠、穩健及安全—人工智能應用須具穩健性與可靠性，在整個測試過程中處理相同情境時應產出一致的結果與輸出。

4.1.4.5.2 模型目標與誘因

定義：

為達成目標而推動的人工智能應用，可能遭濫用以滿足既定指標，卻未能真正解決問題，最終導致不良行為。

此階段涉及的持份者：

項目經理、數據科學家

實務措施及示例：

措施	措施詳情
制定風險敏感性的目標	• 風險敏感性的目標旨在平衡人工智能應用決策所產生的風險與效益。項目團隊應在機器學習目標函數中加入下行風險衡量指標，以更保守的方式運作，避免人工智能做出災難性決策。例如，若人工智能模型未達成功標準，應執行其他措施，可能包括轉交人類操作員處理或中止決策程序。
範例說明：強化學習可應用於自動駕駛車輛。當人工智能控制車輛朝特定方向移動時，可將車輛周圍的交通模式作為獎勵函數（例如：若周圍車輛與自動駕駛車輛保持超過指定距離且符合高速公路行駛路線，則人工智能所做決策正確的可能性較高）。應檢查各種情境，以確保識別的交通模式對他人安全無虞。此做法符合「可靠、穩健及安全」的原則。	

相關人工智能道德原則：

1. 可靠、穩健及安全—人工智能中的獎勵函數可強化學習模型，從而產生更可靠且穩健的人工智能模型。

4.1.4.5.3 輸入變數選擇

定義：

輸入變數（或稱特徵）是數據集中載入人工智能應用的數值，目的在於訓練人工智能模型。一個穩健的人工智能模型依賴這些具資訊性的輸入來提供輸出，此輸出通常稱為目標變數（即人工智能模型試圖預測的內容）。輸入變數的選擇應同時考量機構的專業知識以及因果關係。

此階段涉及的持份者：

系統分析員、系統架構師、數據科學家

實務措施及示例：

措施	措施詳情
遵守政策與道德規範以獲取及使用輸入變數	• 制定流程以確保潛在影響及數據使用符合「人工智能道德原則」及內部政策（包括法律要求）。 • 明確說明數據使用目的，並在使用輸入變數訓練人工智能模型前取得數據持有者的許可，尤其是涉及可識別特定個人身份的個人資料時。 • 確保輸入變數的收集方式合規（例如符合《私隱條例》要求），且其用途適當並獲許可。

措施	措施詳情
避免目標變數洩漏	• 項目團隊應避免目標變數洩漏，即是使用超出訓練數據集範圍的資訊建立人工智能模型。倘若用以訓練人工智能模型的數據集包含預測時點通常無法取得之資訊，便會導致目標變數洩漏。此類未來資訊將使人工智能模型結果呈現的準確度，高於實際情境中的模型表現。 • 若使用時間序列數據，須評估是否錯誤納入預測案例發生後的數據。可設定時間截點以防止獲取預測時點後的資訊。 • 對於與輸入變數高度相關的特徵，應檢視其是否源自未來資訊（此類特徵若存在，極可能是目標變數洩漏的徵兆）。
範例說明：機構應避免目標變數洩漏，以遵循「公平」的原則。當人工智能用於醫學診斷以偵查病徵時，訓練數據集應僅包含預測時點可獲取的數據。若使用例如未來手術記錄等數據，可能導致模型實際表現不如預期。	

相關人工智能道德原則：

1. 公平—人工智能模型所使用的數據應以公平方式取得。
2. 可靠、穩健及安全—為確保運作可靠且符合預期，避免目標變數洩漏意味着對預測的測試結果，能夠更確切地反映真實預測的準確度。

4.1.4.5.4 模型過度擬合

定義：

過度擬合是一種建模錯誤，當人工智能模型被訓練得過於貼合有限的數據點時就會出現。過度擬合的模型通常在訓練數據集上表現出高準確度，但在新數據上的準確度卻很低。若人工智能模型無法從訓練數據良好地泛化至新數據或未見數據，在預測時便可能會表現欠佳。

此階段涉及的持份者：

數據科學家

實務措施及示例：

措施	措施詳情
執行交叉驗證	• 交叉驗證是由開發團隊執行的重新抽樣程序，用作評估建基於有限數據樣本的人工智能模型的有效性和準確性。重新抽樣程序的例子之一是 K 折技術。 • K 折技術允許在不同訓練數據子集上對人工智能模型進行 K 次訓練和測試，以更準確評估模型的預測性能。 • 每個重新抽樣的數據集的數據分布，應與完整群體的數據分布相似。舉例而言，若完整群體包含 10 個不同類別，重新抽樣的數據集也應有相當數量的不同類別。這是因為在交叉驗證過程中若欠缺某些類別，可能會導致人工智能模型的結果產生偏見。
執行正則化及其他選擇模型的方法	• 考慮使用正則化。正則化是使人工智能模型標準化和統一化的過程。正則化通常是機器學習中用來微調人工智能模型功能的技術，通過抑制學習額外且不必要的複雜模型來降低過度擬合的風險。 • 正則化技術的例子之一是剪枝。這可用於移除對模型學習最不重要的特徵。人工智能模型不僅能學習所獲提供數據集的基本特徵，還會學習雜訊以及強調不重要特徵的波動，因此剪枝可以有所幫助。不重要特徵的例子可包括圖像中與實際檢測對象無關的背景。

措施	措施詳情
範例說明：為避免醫療診斷中出現過度擬合，可以執行交叉驗證。通過按特定醫療條件分割樣本數據並重新評估模型的診斷，可以進一步評估人工智能模型的性能。另一種避免過度擬合的方法是減少與病情無關的因素（例如在某些情況下減少社會背景或地理位置等變數以降低人工智能模型的複雜度）。這些做法有助機構遵循「可靠、穩健及安全」的原則。	

相關人工智能道德原則：

1. 可靠、穩健及安全—應避免過度擬合，以確保人工智能模型用於未見數據時所產生的結果可靠準確。

4.1.4.5.5 模型訓練擁有權

定義：

當人工智能模型的訓練過程部分或完全外判予公共雲端，或依賴第三方的預先訓練模型時，可能會引發新的安全風險。請注意，若人工智能模型訓練過程未有外判，則本分節所述實務措施及示例均不適用。

此階段涉及的持份者：

採購團隊

實務措施及示例：

措施	措施詳情
使用可信賴的數據 / 平台服務	• 訓練人工智能模型的過程需提供訓練數據供其學習。精確的訓練數據能協助人工智能模型理解模式並得出結論。任何用於訓練人工智能模型或為機構提供雲端訓練環境的第三方供應商，均應符合第三方採購流程，此舉有助降低取得可能含有後門的篡改模型之風險。有關第三方採購流程的詳細內容，請參閱第 4.1.3.2 節「採購人工智能服務」。 • 機構應評估是否適合在公共雲端上儲存及使用個人資料或機構資料。 • 開發人員應確保人工智能模型算法的訓練環境安全，包括妥善的網絡安全配置及適當的數據加密措施。 • 完成訓練的模型應僅透過能確保完整傳輸的管道傳輸，例如可先將 Docker 容器和 Pickle 檔案（用於 Python 腳本）加密，然後進行傳輸。 • 在適用情況下，應遵循關於跨境數據流動的相關要求或規例。
範例說明：使用雲端 / 第三方服務時，應驗證雲端供應商是否符合國際標準化組織標準等國際安全標準及機構保安措施。符合國際標準有助確保遵循「數據私隱」及「可靠、穩健及安全」等「人工智能道德原則」。	

相關人工智能道德原則：

1. 可靠、穩健及安全—第三方符合相關國際標準，可提高其人工智能應用保持可靠、穩健及安全的可能性。

4.1.4.5.6 對抗性攻擊

定義：

對抗性攻擊是指惡意行為者欺騙人工智能模型，意圖在不被偵測到的情況下影響人工智能應用的輸出結果，其方法是通過輕微修改輸入數據而這些修改對人類而言難以察覺，來誘導人工智能模型作出錯誤預測。這類攻擊可發生在訓練階段或測試階段，發生在訓練階段的攻擊屬投毒攻擊，而發生在測試階段的攻擊則屬規避攻擊。

此階段涉及的持份者：

數據科學家、項目經理

實務措施及示例：

措施	措施詳情
執行對抗性訓練	• 項目團隊應取得盡可能多的對抗性樣本，並明確訓練模型不受這些樣本誤導。對抗性訓練涉及擴大數據集以包含人工智能模型可從中學習的對抗性樣本。引入對抗性訓練，是為了保護人工智能應用免受人工智能模型中不同漏洞的破壞性擾動所影響。當訓練數據中的現有數據點被添加微小的隨機值而改變數據原始特徵時，就會發生破壞性擾動。此類例子常見於數位圖像中，在原始圖像添加隨機值會導致人工智能模型錯誤分類（例如一張熊貓圖片因破壞性擾動而被分類為長臂猿）。 • 建立新的對抗性機器學習模型，用於預測數據集是來自近期的數據集還是人工智能訓練數據集。這可作為識別新近變更數據的方法，從而發現需要重新訓練人工智能模型的新型對抗性攻擊。

措施	措施詳情
	• 舉例來說，以 Python 程式語言編寫的開源函式庫（例如對抗性穩健性工具包）[18]，可用於訓練人工智能模型抵禦對抗性攻擊。

執行獨立評估與測試	• 在進行人工智能應用安全評估時，應考慮引入獨立第三方。對於大型或高風險項目，可採用「紅隊」方法，以確保系統的準備度、品質與可靠性。「紅隊」會模擬敵對角色，透過識別潛在漏洞並在具挑戰性的情境中對系統施加壓力，以測試系統的極限，從而協助機構在正式啟用前堵塞安全缺口及識別風險與問題。
	• 評估系統在出現故障、攻擊或其他異常狀況下抵禦或克服惡劣條件的能力，防止非預期的結果及行為錯誤，同時維持最低水平的運作功能。
	• 評估應涵蓋以下領域：
	  ○ 數據層面—檢視數據處理是否存在數據投毒或篡改等潛在漏洞，評估數據的完整性、品質及私隱保障措施；
	  ○ 模型層面—評估人工智能模型的準確性、公平性、韌性及穩健性，評估模型洩漏的風險與未經授權存取或篡改的可能性；
	  ○ 輸入層面—檢視系統抗禦非預期及惡意輸入（例如提示注入[19]）的能力；以及
	  ○ 輸出層面—檢視是否可能產生有害、歧視性或不良的結果。機構應確保輸出內容符合《維護國家安全條例》、《私隱條例》、《版權條例》及反歧視條例等相關規例。
	• 定期執行評估測試，分析安全與保安問題，並提出改善方

[18] https://research.ibm.com/projects/adversarial-robustness-toolbox
[19] https://www.ibm.com/topics/prompt-injection

措施	措施詳情
	案。 • 針對人工智能系統的適用場景、安全性及可靠性實施系統審計。
執行降維	• 對訓練數據實施降維，藉此增強分類器的韌性，作為抵禦規避攻擊的防禦機制。此技術意味着可用於篡改數據的維度將減少。降維指在保留原始數據的關鍵特性的前提下減少數據屬性（例如個人數據屬性可包含年齡、身高、體重及種族）。 • 檢查缺失值比率是降低數據維度的方法之一。數據欄位若存在大量空值或無效值，其承載有效信息的可能性便較低。因此，空值數量超過特定閾值的數據欄位應予移除——閾值設定得越高，數據精簡程度便越大。
範例說明：遵循「可靠、穩健及安全」的原則往往需要實施對抗性訓練（即使用對抗性樣本訓練模型抵禦對抗性攻擊）。為了減緩對抗性攻擊，可採用標準對抗性函式庫（例如 CleverHans）訓練人工智能模型識別對抗性圖像數據。另外，在醫學影像診斷模型中，可透過壓縮圖像降低維度，雖然會導致圖像精準度下降，但因解析度縮減，可增加攻擊者篡改像素的難度。 請注意，使用函式庫前應先依照其他資訊科技項目的規定進行評估與審批。	

相關人工智能道德原則：

1. 可靠、穩健及安全—人工智能應用必須保持穩健與安全，在遇到潛在對抗性攻擊時有能力維持其資訊的完整性。

4.1.5 系統部署

4.1.5.1 模型整合與影響

定義：

核實、驗證及測試過程旨在確保人工智能應用的運作，符合項目初期訂定的要求。在部署前，應對人工智能應用進行全面測試，以評估是否出現故障，以及能否按預期運作。

此階段涉及的持份者：

數據科學家、項目經理

實務措施及示例：

措施	措施詳情
執行整合、系統、決策及用戶驗收測試	• 在部署前（包括人工智能應用內的人工智能模型），應制定正式且嚴謹的人工智能應用測試要求，以確認其適合投入生產環境（例如涉及公平性、透明度與可解釋性、錯誤率及性能一致性等要求）。請參考： ○ 第 4.1.4.3 節「數據提取」中有關公平性的詳細說明；以及 ○ 第 4.1.2.2 節「項目監督與交付方式」中有關透明度與可解釋性的詳細說明。 • 制定明確的測試規則與方法，包括人手測試、自動化測試及混合測試，並運用沙盒模擬等技術徹底測試及核實模型。 • 對人工智能應用進行的額外測試應包括回歸測試，以識別模型錯誤。回歸測試可包含基於正常輸入的測試，用以評估泛化誤差。

措施	措施詳情
	• 應讓終端使用者（非技術專家）參與測試過程，以確保人工智能應用能為機構作出有意義的決策，並提供更佳的使用者互動體驗。 • 標準化的測試基準或指標可用於比較候選模型或評估部署的準備度，協助選擇最佳的人工智能應用。請參考第 4.1.6.1 節「數據與模型表現監察」中的效能指標範例。 • 項目團隊、業務使用者及治理職能人員等持份者，應在人工智能資產部署至生產環境前審查測試結果並給予核准，這包括評估人工智能應用是否符合機構的目標，以及建立相應流程以便在需要時提供人工智能應用決策或建議的理由。
使用邊緣案例與例外處理	• 項目團隊應考慮使用邊緣案例來訓練人工智能模型，教導其在遇到邊緣案例時應如何回應。邊緣案例通常發生在人工智能應用遇到可能無法如預期運作的意外情境時。 • 應開發並整合人工智能模型以處理未見數據、新使用案例及潛在惡意輸入。 • 以邊緣案例為例，當使用人工智能從影片測量人流時，應核實各種不同情境（例如有人騎乘單車或佩戴口罩的情況），並採取預防措施，以確保正確處理這些情境。 • 評估人工智能應用提供不準確結果時的影響，並找出降低影響的方法。例如若人工智能應用產生錯誤結果，對使用者造成傷害的可能性有多高？增加人為干預可有助降低錯誤結果的影響。
範例說明：在部署前，應由不同終端使用者與項目團隊成員共同執行整合、系統、決策及用戶驗收測試，以實踐「合作與開放」的原則，並建立適當的制衡機制。以部署天氣預報人工智能應用為例，應由人類預報員進行測試，以確保人工智能應用的輸出結果有效。人工智能應用的功能需求（例如天氣預測）與非功能需求（例如性能與可用性）均應進行評估。	

相關人工智能道德原則：

1. 可解釋性—項目團隊應能解釋在測試時人工智能應用的決策邏輯及影響決策結果的關鍵因素。
2. 可靠、穩健及安全—人工智能應用須通過整合測試、用戶驗收測試及錯誤處理測試，以確保可靠性與穩健性。
3. 合作與開放—項目經理應協同項目團隊與終端使用者，在部署前共同完成測試。

4.1.5.2 過渡與執行

定義：

假設人工智能應用出現故障，應制定緩解措施以減輕故障時的損害。

此階段涉及的持份者：

數據科學家、項目經理

實務措施及示例：

措施	措施詳情
建立多層緩解機制	• 在代碼中建立多層緩解機制，可在人工智能應用的不同層級或模組阻止系統出現錯誤或故障。這有助於在早期偵測錯誤，防止錯誤影響人工智能應用的其他部分。對於高風險人工智能應用，應進行演練以確保所有人員都清楚各類系統錯誤或故障的處理程序。 • 評估並提升人工智能產品與服務在故障、受攻擊或其他異常狀況下的承受與復原能力。 • 項目團隊應在每個解決步驟的開始或結束設立防禦或控制措施，以防止連鎖反應導致重大系統故障。例如數據進行預先處理時若出現錯誤，可停止處理並向用戶發出警示。 • 考慮設立控制措施，允許在系統出現故障時進行人為干預或自動關閉，並制定業務持續運作方案應對此類不利事件。方案應包含系統出現故障時的應對與復原方法，系統需具備故障保險能力，能在出現故障時快速切換至人手或傳統方法。 • 機構應考慮制定業務持續運作方案，訂明模型出現故障、需要下線或退役時的處理流程。退役可能是因為模型失效、不再適用或已有更新更好的系統取代。業務持續運作方案確保業務流程或功能在過渡期間不受影響。某些系統可能需要在新系統上線時與舊的人工智能應用並行運作。
應用業務規則與處理輸出	• 機構應實施品質保證業務規則，例如對某些預測應用的結果進行額外評估，以確保預測合理。

措施	措施詳情
實施模型 / 應用程式介面存取安全措施	• 透過科技手段識別非預期、不實及不準確的輸出，並根據《維護國家安全條例》、《私隱條例》、《版權條例》及反歧視條例等法律及規例妥善處理。 • 人工智能應用應包含與其他資訊科技系統類似的核心數據安全原則。整個人工智能應用的設計與開發流程應蘊涵安全設計與數據私隱，以保護數據及個人私隱權利。 • 根據實際應用場景設定服務範圍，並透過定期審查來實施控制措施，防止出現超出預設範圍的功能。如發現系統功能可能遭濫用，應立即停用。 • 透過限制應用程式介面呼叫及採用集成學習（結合多種檢測方法）的子集來服務應用程式介面或用戶，以限制或控制對人工智能應用的輸入，可防止人工智能模型遭竊取或篡改。集成學習用於整合不同檢測方法（例如人工智能分類器、檢測規則與異常檢測），藉此提升系統穩健性。攻擊者若試圖操控輸入系統，必須突破所有檢測機制，方能透過人工智能應用向人工智能模型輸入惡意內容。 • 應盡可能平等對待所有用戶。若機構必須針對不同用戶優先處理特定資訊或技術存取權限，則需提出合理依據並記錄於存取日誌中。

向用戶提供資訊	• 制定一份可視適用情況向終端使用者或公眾（例如在網站上）提供的披露文件，以鼓勵實踐透明度與可解釋性，從而確保人工智能應用的決策過程易於為人類所理解。該文件應概述人工智能應用的以下細節：
	  ○ 運作方式及預期用途；
	  ○ 目標使用者；
	  ○ 使用及處理的數據；
	  ○ 模型訓練程序；
	  ○ 測試程序；
	  ○ 功能及效能指標；以及
	  ○ 為評估「人工智能道德原則」（例如可靠、穩健及安全）而進行的檢查。
	• 對於第三方人工智能應用，機構應向第三方索取解釋人工智能應用的同等文件。詳情請參閱第 4.1.3.2 節「採購人工智能服務」。
	• 制定可視適用情況連同使用條款向終端使用者或公眾（例如在網站上）提供的免責聲明，內容應包括：
	  ○ 注意事項；
	  ○ 限制條款；
	  ○ 使用禁止事項；
	  ○ 在檢測到不符合使用意圖及所述限制的濫用行為時，採取糾正措施或提前終止服務的權利；以及
	  ○ 適用的透明度要求。
	• 制定個人資料收集聲明及私隱政策，告知終端使用者其個人資料的使用情況、相關權利及可能產生的後果。
	• 在內部及外部（視適用情況而定）公布政策，分發有關治理措施的資訊，以披露第三方合作、數據擁有權或行業標準合規性等資訊。

措施	措施詳情
	• 提前告知用戶是否涉及人工智能互動或內容由人工智能生成。例如在用戶使用人工智能聊天機械人之前，在網站或通訊應用程式上通知用戶，或在人工智能生成的內容中加入浮水印、元數據、數碼簽署等。這是為了確保終端使用者或其他人士充分知悉某項決策、內容、建議或結果是源自人工智能應用或算法決策。
範例說明：機構應實施應用程式介面存取安全措施，以遵守「可靠、穩健及安全」的原則。以影像分析應用為例，應限制用戶輸入並設立控制措施以確保準確性（無論輸入是直接還是透過應用程式介面進行）。	

相關人工智能道德原則：

1. 可靠、穩健及安全—遵循安全規範並實施多重緩解措施，可防止錯誤結果，從而確保人工智能應用的安全與可靠性。
2. 人為監督 / 安全—應實施控制措施，以便在系統出現故障時，特別是在故障可能影響人類安全時，允許人為干預或自動關閉系統。
3. 透明度與可解釋性—應提供人工智能應用算法的解釋，以支持建立對人工智能使用的信任。
4. 有益的人工智能—（在適用情況下）提供關於人工智能應用效益的披露聲明，以說明並推廣該人工智能應用所實現的公共利益。

4.1.5.3 持續監察

定義：

提供反饋以增強人工智能應用的學習能力與穩健性。

此階段涉及的持份者：

數據科學家、項目經理

實務措施及示例：

措施	措施詳情
追蹤錯誤	• 部署適當的糾正措施追蹤機制（例如標記特定情況供人工審查），以了解人工智能應用所犯的錯誤，從而解決這些錯誤。 • 當人工智能應用須對全新且不熟悉的情況做出決策時，考慮實施標記機制，標記這些情況以供審查。 • 觸發點（如輸入數據包含訓練數據中不曾出現的新數值）可能表示較舊的訓練數據已失效。人工智能模型在嘗試從新型態輸入數據中判斷新的未知情況時，可能會產生誤解。因此，利用這些觸發點並實施糾正措施可減少此類錯誤。
實施易用的反饋用戶介面	• 透過設計易於使用的反饋表單，讓終端使用者或公眾（視人工智能應用性質而定）能分享有關預測出錯等案例的資訊。 • 機構可藉着提供獎賞（例如小禮品或獎勵），鼓勵終端使用者提供優質而適時的反饋。 • 此外，機構可在網站上提供熱線電話或相關人員的電郵地址，供終端使用者聯繫。 • 進行人工智能應用維護作業時建立流程，將使用者反饋納入其中。重大問題應透過現有支援渠道（例如事件管理）呈報。

措施	措施詳情
範例說明：針對醫療診斷預測應用，可部署審計日誌記錄所有診斷與預測結果，並與實際結果比對評估。該應用可設置「反饋」選項，透過提示訊息鼓勵終端使用者提供意見。此做法符合「可靠、穩健及安全」的原則，能根據使用者反饋持續提升人工智能應用的可靠性與穩健性。	

相關人工智能道德原則：

1. 可靠、穩健及安全—定期追蹤須予修正的錯誤及使用者反饋，以持續提升人工智能應用的可靠性與穩健性。

4.1.5.4 評估與檢查

定義：

須確保人工智能應用具備可追溯性、可重複性與可再現性，以確保其運作正確，從而建立公眾與關鍵持份者的信任。

此階段涉及的持份者：

項目經理、數據科學家

實務措施及示例：

措施	措施詳情
確保端到端工作流程具可重複性與可再現性	• 確保人工智能應用從數據源到分析結果的流程是可透過自動化重複過程，減少執行時所需的臨時處理程序。項目團隊應進行可重複性測試，確保可複製端到端的工作流程。 • 項目團隊應確保人工智能應用具可再現性。當使用相同數據集或預測方法時，人工智能模型應產出相同結果，

措施	措施詳情
	即具備可再現性。人工智能應用的可再現性是衡量可信度的重要指標。
實現可追溯性	• 項目團隊應透過建立審計追蹤記錄，妥善儲存數據並記載非功能 / 功能需求、模型訓練與決策流程，以避免數據遭篡改。審計追蹤記錄應按適用的保存期限予以保留。 • 考慮為整個數據管道建立追溯機制，以強化問責制並追蹤數據操作活動。當人工智能應用的決策、數據集、流程 / 算法、評估參數、結果與錯誤日誌均被記錄及追蹤時，即具備可追溯性。 • 確保人工智能應用的日誌和記錄（如排程任務時間戳記、數據行數、相關用戶及模型交互記錄）能安全生成與儲存。請參閱第 4.1.3.1 節「人工智能與數據使用技術路線圖」、第 4.1.4.2 節「解決方案設計」及第 4.1.4.5.1 節「模型假設」關於不同記錄類型的內容。
確保可審計性	• 項目團隊應保留證據，以便評估與人工智能應用相關的算法、流程與數據沿襲，當中包括來自各階段（如數據收集、預處理、模型訓練、測試部署）的證據，用以確保可審計性。人工智能應用的所有關鍵決策，均須以文件方式妥為記錄。 • 項目團隊應在人工智能應用的日誌模組中，記錄人工智能應用的流程與元數據管理。 • 為確保所有文件資料保持最新，應定期（如每季或每年）檢視日誌。對於中期的文件處理工作，須確保模組開發者與整合者皆理解設計決策背後的理據。 • 請參閱附錄B「參考資料」第B.1節。

措施	措施詳情
範例說明：機構應確保端到端工作流程具可重複性與可再現性。以影像分析為例，其數據收集、整合、分析與生成洞察的流程，可透過每日自動運行的測試腳本實現自動化。記錄結果產出的過程可作為證據，有助識別問責範疇，藉以實踐問責原則。	

相關人工智能道德原則：

1. 合法與合規—採取相關措施，確保已考量行業標準並具備可審計性。
2. 問責—日誌記錄可作為人工智能應用內發生事件的證據，有助識別事件的問責範疇。

4.1.6 系統運作與監察

4.1.6.1 數據與模型表現監察

定義：

應持續監察與審查人工智能模型（作為人工智能應用的組成部分），因其準確性與相關性可能逐漸降低。當數據與環境持續演變時，便可能發生此現象。

此階段涉及的持份者：

項目經理、數據科學家

實務措施及示例：

措施	措施詳情
設定效能指標	• 界定並驗證人工智能應用的效能標準，以確保有關的人工智能應用持續為機構創造價值。成效指標可透過管理報告傳達。

措施	措施詳情
	• 應針對人工智能處理的問題或決策制定相關指標，並建立跨數據子群組（如不同產品或終端使用者群組），以通報異常情況。 • 可計算精確率、召回率、F值或準確率等指標，用以衡量人工智能模型效能，包括檢查預測成功率與誤差程度，從而確認模型效能仍保持在可接受的水平。相關數據偏見範例請參閱第 4.1.4.3 節「數據提取」。 • 機構應界定人工智能模型指標，評估模型是否按設計運作以持續達成預期目標。
納入品質評估	• 記錄關於預期輸入、數值分布與邊界等假設，並將其整合至人工智能模型，此舉有助標記可能阻礙模型完整執行及導致不良預測的實例。 • 應根據人工智能模型風險等級，由項目督導委員會 / 項目保證小組定期執行獨立驗證。 • 機構應定期評估人工智能模型指標，確認模型運作是否符合設計。隨着人工智能技術發展，新風險將不斷出現，建立部署後持續識別、審查與緩解新風險的流程，將有助提升人工智能應用的可靠性、穩健性與安全性。 • 品質評估應作為檢測與應對機制，處理人工智能應用對終端使用者或對象產生的不良影響（例如當模型未達預期效能時，應透過檢測和重新調校模型，解決有關問題）。
建立異常檢測技術	• 項目團隊應建立、實施及監控異常檢測技術。異常檢測指識別數據集中，模式偏離預期行為的離群值或例外情況。該技術能有效發現顯著偏離主體數據的異常值，並可預防對抗性輸入。
重新訓練人工智能模型	• 機構應定期使用新數據點進行人工智能模型調校與重新訓練。當出現新增或變動數據點時，可透過自動化或手動方

措施	措施詳情
	式定期重新訓練模型。模型調校應作為模型維護或營運支援的一部分，或在人工智能應用用戶提出反饋時進行。 • 機構應記錄人工智能模型與應用的更新時間、方式及對輸出結果的影響，以確保可追溯性與可審計性。此程序應納入人工智能應用的變更管理流程。 • 每次模型調校更新時，可把較舊人工智能模型的誤分類案例納入其中，以便從過往的真誤差或正誤差中學習。 • 重新訓練的人工智能模型在部署前應經過決策審核與用戶測試，以確保沒有在重新訓練和更新過程中引入新錯誤。由於這是一個迭代過程，應在人工智能應用部署後持續收集終端使用者反饋，以實現模型與應用的持續改進。
範例說明：可設定效能指標以遵循「可靠、穩健及安全」的原則。若人工智能輔助醫生進行診斷，應記錄醫生認為人工智能是否準確，並據此作為評估效能的指標。	

相關人工智能道德原則：

1. 可靠、穩健及安全—品質檢查與效能指標協助機構確保人工智能應用的可靠性。
2. 問責—確保設有專責角色執行人工智能應用的品質檢查與監察工作。

4.1.6.2 運作支援

定義：

當人工智能應用部署後，應建立持續的運作支援，以確保人工智能應用的表現保持穩定、可靠及穩健。

此階段涉及的持份者：

項目經理、數據科學家

實務措施及示例：

措施	措施詳情
建立服務水平協議	• 界定並協商終端使用者與資訊科技支援之間的服務水平（服務可用性、修復時間、嚴重性閾值），同時考慮人工智能應用的需求。服務水平協議明確界定終端使用者對項目團隊的服務期望，並列出衡量服務的指標。 • 支援流程應與現有的運作支援流程（例如事件管理流程）整合。
建立反饋渠道	• 提供機制讓用戶提交反饋，例如透過電郵、網站反饋及表格輸入。詳情請參閱第 4.1.5.3 節「持續監察」。
制定或運用事故處理計劃以應對危機情況	• 資訊科技項目團隊應制定新的事故處理計劃，或在現有的計劃上作出補充，以應對因人工智能應用的對抗性影響或人工智能相關事件而引發的危機。這些事件指人工智能系統的開發、使用或故障導致傷害，例如個人受傷、關鍵基礎設施中斷、或對財產造成損害。這對於應用於關鍵領域（例如關鍵資訊基礎設施及直接影響公共安全與人民健康及安全的系統）的人工智能應用尤為重要。 • 事故處理計劃應涵蓋關鍵組成部分，例如用於檢測人工智能相關異常的監測機制、針對可疑事件的報告政策、具有明確角色和呈報路徑的應對協議、調查程序以及事後補救行動。機構應使用目前用於其他資訊科技應用的溝通渠道。機構應迅速報告在人工智能系統運作中檢測到的安全事件和漏洞。 • 如有與人工智能或其輸出相關的問題或爭議，這在很大程度上取決於模型本身的功能，並須由機構酌情決定如何解決。一般而言，機構應致力把用戶反饋納入人工智能應用中，而這些反饋將被納入模型未來的訓練迭代中。

措施	措施詳情
範例說明：可透過建立服務水平協議來設定責任方，以落實並遵從「問責」原則。當天氣預報人工智能應用出現錯誤且運作異常時，應由資訊科技支援團隊在規定時間內予以修復。	

相關人工智能道德原則：

1. 可靠、穩健及安全—透過採納反饋意見及遵守服務水平協議，確保人工智能應用長期保持可靠、穩健且安全。
2. 問責—資訊科技支援團隊應負起責任，在規定時限內修復人工智能應用。

4.1.6.3 持續審查 / 合規

定義：

應設立持續審查 / 合規功能，以監察和評估人工智能應用，確保其適切性、效率及成效。項目經理、項目督導委員會 / 項目保證小組及資訊科技委員會 / 首席資訊官負責監管人工智能的風險，例如違反適用法律和規例的情況。

此階段涉及的持份者：

項目經理、資訊科技規劃人員 / 管理層

實務措施及示例：

措施	措施詳情
進行管理 / 持續審查	• 考慮聘請獨立審計師或顧問進行合規審計 / 評估。 • 機構管理層應： ○ 維持與使用人工智能相關的機構政策清單。與人工智能相關的政策範例可包括安全、數據治理及工具採購； ○ 就政策進行合規問題審查，對照外部法律和規例政策並討論所需修改。審查時機取決於業務需求（例如某些情況下須立即審查以反映新的法律要求）。所需修改應透過討論及與行業標準和外部規例進行比較後確定； ○ 邀請道德及企業社會責任領域的專家參與審查和修訂人工智能政策；以及 ○ 制定並實施針對人工智能資產的標準風險評估和業務驗證程序。請參閱本文件第 5 章「人工智能評估」。
終止人工智能應用	• 應制定終止流程。 • 終止人工智能產品時妥善處理數據，並確保已按政府的安全要求，清除或銷毀所有存放於外部服務或設施的政府數據。 • 當用戶權限及數據存取權不再需要時，應有系統地予以撤銷。
諮詢資訊科技委員會 / 首席資訊官	• 對於高風險人工智能應用，應就「人工智能道德原則」考量（例如公平、可解釋性）進行審查和提出質疑，並識別任何問題。 • 詳情請參閱第 3.5.2 節「人工智能治理」。

措施	措施詳情
範例說明：機構可諮詢負責執行審查的資訊科技委員會 / 首席資訊官。此做法符合「問責」原則。對於影像分析及人臉識別等高風險案例，應由資訊科技委員會 / 首席資訊官進行審查。	

相關人工智能道德原則：

1. 合法與合規—隨着與人工智能應用相關的新行業標準和規例持續發展，須確保符合這些規範的要求。
2. 問責—須確保管理層肩負執行持續審查的問責義務。

第5章

人工智能評估

5. 人工智能評估

本章建議的人工智能評估提供一系列針對性問題（與人工智能生命週期對應），旨在協助機構評估、識別、分析及衡量人工智能應用的效益與影響，確保其符合「人工智能道德原則」的宗旨，並確定需要採取的適當緩解措施，將負面影響控制在可接受水平內。

5.1 人工智能應用影響評估

人工智能應用影響評估應在人工智能生命週期的不同階段進行。該評估為各機構引入系統化思考流程，以便全面審視個別應用的各項考量因素，包括相關效益與風險，同時突顯有額外管治活動的需要，並制定跟進行動以確保實施符合「人工智能道德原則」的必要措施與管制。

人工智能應用影響評估的流程包括識別與分析以下項目：

- 與人工智能應用有關的流程及符合「人工智能實務指引」要求的「人工智能道德原則」；
- 人工智能應用將帶來的效益；
- 對特定持份者的負面影響；以及
- 按規定建立的管控措施及合規 / 監察程序。

執行人工智能應用影響評估，有助：

- 透過討論已識別的效益、風險及影響協助決策，以更有效緩解風險和提供最大效益；
- 平衡地審視人工智能應用的現有風險；以及
- 機構更能向持份者確保人工智能應用造成的影響受妥善管理。

人工智能應用影響評估範本用於備存和記錄人工智能應用如何符合《人工智能實務指引》內部要求，並提供審批決策的考慮因素，促進審查與核准流程。該範本亦用於評估應用所構成的風險，協助機構規劃及訂立具體管治要求以緩解相關風險。

本評估所用的人工智能應用影響評估範本詳見於《人工智能道德框架》附錄C「人工智能應用影響評估範本」，以作參考。

人工智能應用影響評估包含以下組成部分：

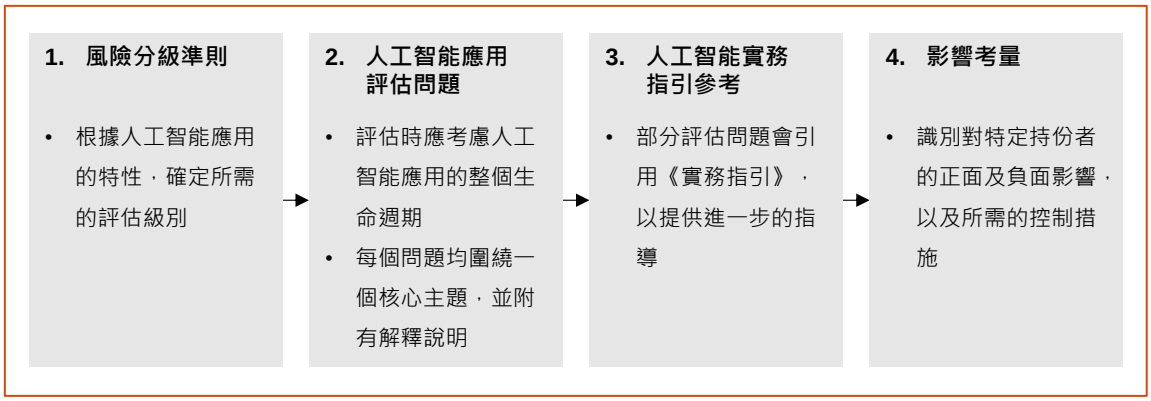

圖12：人工智能應用影響評估的組成部分

1. **風險分級準則**—一套用於區分高風險人工智能應用的問題集。在提出人工智能項目之初，或當人工智能應用條件變更時，應妥為回答有關問題。經判定為高風險的人工智能應用，其後須經資訊科技委員會 / 首席資訊官審查核准。下圖展示風險分級問題的範例。無論風險分級準則的答案為何，項目團隊均應完成人工智能應用影響評估的其他部分。

問題	背景說明
a - 該人工智能應用是否涉及公眾高度關注（例如涉及私隱問題）及／或訴訟頻繁的領域？	*舉例而言，某些「物聯網」應用可能對個人日常生活及私隱構成重大影響。相關應用案例包括智慧城市項目：* • *智能照明──根據天氣自動調節的智慧路燈* • *智能交通管理──監測車輛與行人流量* • *智能泊車系統──監測泊車位使用狀態* *因此，這類應用的人工智能應用影響評估須由資訊科技委員會／首席資訊官審視。若結合人臉識別技術與全新、敏感或可能具爭議性的用途，尤須審慎評估。*
b - 該人工智能應用是否用於全新（社會）領域（即香港從未在該領域應用人工智能）？	*例如某人工智能應用首次用於醫療保健領域，而此前僅用於營銷用途。由於應用領域轉變，該人工智能應用可能引發（新的）道德爭議。如在敏感社會領域應用人工智能，風險與道德問題可能更為嚴峻，例如涉及照顧服務、公共安全、打擊恐怖主義或教育等議題；又或涉及兒童或殘疾人士等弱勢社群。*

圖13：風險分級問題範例

2. **人工智能應用影響評估問題**—此問題集旨在確保人工智能應用在其生命週期各階段的影響獲識別與管理，並確認相關「人工智能道德原則」已獲考量。問題內容涵蓋人工智能應用的影響層面，包括效益、風險、對個人權利的影響，以及各方利益之間的平衡。項目團隊提供的答案將以質化方式進行評估。圖14展示人工智能應用影響評估問題的範例。

人工智能應用影響評估問題

A.人工智能應用目的與問責	實務指引參考章節／責任
1 - 此人工智能及數據活動的業務需求／目標／目的是什麼？所開發的人工智能應用是否有明確界定的清晰目的（例如提升營運效率或降低成本）？如何能令預期效益大於潛在風險？此計劃活動的驗收或成功標準是什麼？ *該人工智能應用會否納入一項目前正考慮或進行中的工作，而後者所涉的範疇更大？這項計劃是否符合你的人工智能及／或數據策略？*	第4.1.2.1節： • 將人工智能項目對應至機構目標 • 篩選人工智能項目，並訂定優先順序 資訊科技規劃人員／管理層、業務使用者

圖14：人工智能應用影響評估問題範例

3. **「人工智能實務指引」參考章節**—評估問題已列出第4章「人工智能實務指引」的相關分節以供參考（如適用），將受評估的特定領域與實務做法相連結。
4. **對影響的考量**—在人工智能應用影響評估範本中，有一系列問題圍繞對持份者的利與弊，旨在引導機構評估是否需要採取進一步緩解措施（例如是否需要制訂額外管控措施或程序處理有關影響）。

5.1.1 完成人工智能應用影響評估的流程

完成人工智能應用影響評估的流程載列如下：

1. 機構內的項目團隊應從項目的規劃階段開始填寫人工智能應用影響評估範本（有關應於何時完成評估的詳細資料，請參閱第5.2節「人工智能評估頻率」）。
   i. 對項目團隊而言，第一階段是完成回答「風險分級準則」的問題（以評估所列結果的適用性），並完成人工智能應用影響評估問題中「人工智能活動目的與問責」的部分。
   ii. 若任何「風險分級準則」問題的回答為「是」，該人工智能項目即被視為高風險，在開展前必須獲得資訊科技委員會 / 首席資訊官的批核。
   iii. 若所有「風險分級準則」問題的回答均為「否」，則無需獲得資訊科技委員會 / 首席資訊官許可，但仍需獲得項目督導委員會 / 項目保證小組批核。
   iv. 無論有否觸發「風險分級準則」，項目團隊都應隨着項目進入人工智能生命週期的不同階段，逐步完成人工智能應用影響評估的其餘問題。即使未觸發「風險分級準則」，人工智能應用仍可能存在風險，而人工智能應用影響評估將幫助項目團隊確保考慮到各項影響。
   v. 機構應回答人工智能應用影響評估中的每一條問題。某些問題若不適用則無須詳細回答（例如就不涉及第三方人工智能項目，不須詳細回答「人工智能治理流程——第三方問題」的評估部分）。
   vi. 若未遵循《人工智能實務指引》指明的做法且檢測到負面影響，機構應在評估中提供緩解計劃或在評估中記錄為何認為項目仍適合繼續進行。項目督導委員會 / 項目保證小組需在每個階段做出「繼續或中止」的決策。舉例而言，若評估範圍內有措施尚未執行，或需針對措施的特定要求（如偏見測試）進行更多工作，則應基於效益、風險及影響決定項目是否可繼續進行，或需先進行補救措施。

2. 人工智能應用影響評估可作為「動態」文件，在整個人工智能開發生命週期內持續記錄風險與緩解措施等決定有關人工智能開發「繼續或中止」的因素。
3. 項目團隊應詳細檢視整份評估，並以協作形式檢視「影響」部分。此外，當中應包括檢視有否觸發其他須呈報審批的風險分級條件。
4. 人工智能應用影響評估的結果應呈交項目督導委員會 / 項目保證小組審批。該委員會 / 小組會審視業務計劃、潛在影響及應對計劃（如繼續推行項目）。

5.2 人工智能評估頻率

應視乎人工智能項目進度及在人工智能應用運作期間，定期（例如每年或發生重大變更時）進行人工智能應用影響評估。

圖15顯示應檢視人工智能應用影響評估的人工智能生命週期階段。

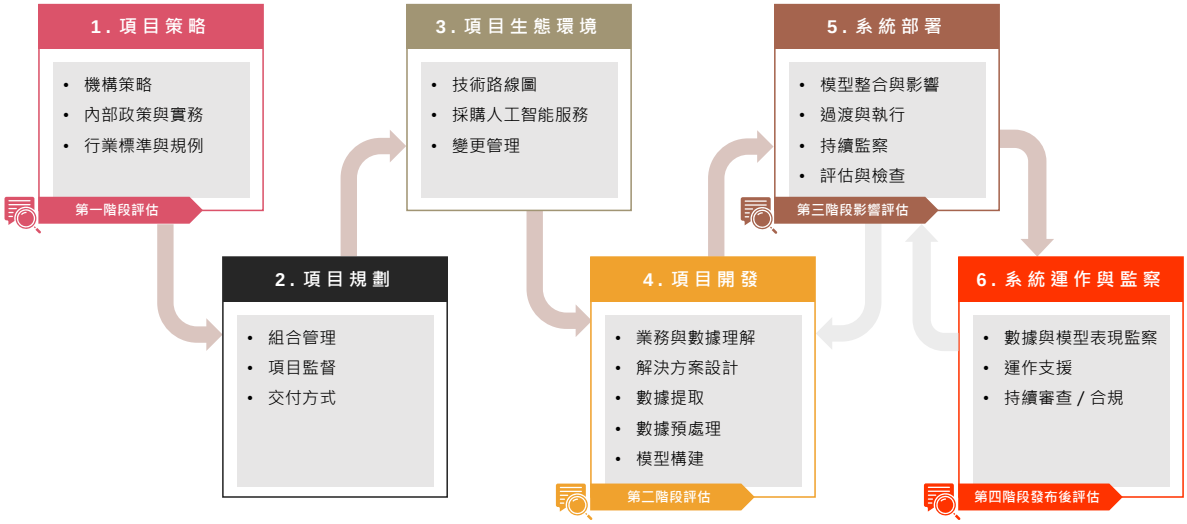

圖15：進行人工智能應用影響評估的階段

對於新的人工智能應用，機構應進行人工智能應用影響評估，並在「項目規劃」、「項目開發」、「系統部署」及「系統運作與監察」等人工智能生命週期階段再作中途評估，確保必要要求被識別並妥善納入後續其他人工智能生命週期階段。人工智能應用影響評估可在整個人工智能生命週期內作為「動態」文件使用，但應在人工智能生命週期的4個關鍵階段（請參閱圖15）進行檢視，並保留評估副本作歷史記錄。

生命週期階段	負責方	執行步驟
系統發展生命週期： 項目申請 **人工智能生命週期：** 項目策略	項目團隊	根據「風險分級問題」的答案，將項目分類為「高風險」或「非高風險」 進行「人工智能應用影響評估」： • 回答問題1-8、9(i)、41-49、57-58
	項目督導委員會 / 項目保證小組	檢視及核准非高風險項目的評估
	資訊科技委員會 / 首席資訊官	檢視及核准高風險項目的評估
系統發展生命週期： 系統分析與設計 **人工智能生命週期：** **項目生態環境** 系統開發	項目團隊	進行「人工智能應用影響評估」： • 回答問題9(ii)、10-14、18-19、22-29、30(i)(ii)、31-34、37(i)、52-56 • 根據最新情況檢視及更新問題1-8、9(i)、41-49、57-58的答案 • 如使用第三方技術或數據，須在採購前完成並檢視問題15-17、48
	項目督導委員會 / 項目保證小組	• 檢視及核准評估報告
系統發展生命週期： 實施階段 （推出前） **人工智能生命週期：** 系統部署	項目團隊	進行「人工智能應用影響評估」： • 回答問題20、21、30(iii)、35、36、37(ii)(iii)、38-40、50、51 • 根據最新情況檢視及更新評估（問題1-14、18-19、22-29、30(i)(ii)、31-34、37(i)、41-49、52-58） • 如使用第三方技術或數據，須檢視問題15-17、48
	項目督導委員會 / 項目保證小組	• 檢視及核准評估報告

生命週期階段	負責方	執行步驟
系統發展生命週期： 系統維護 （每年或發生重大變更時） **人工智能生命週期：** 系統運作與監察	維護團隊	• 根據最新情況檢視及更新「人工智能應用影響評估」 • 呈報任何重大問題
	維護委員會 資訊科技委員會 / 首席資訊官 （或其授權代表）	• 處理呈報事項 • 監察高風險項目

表5：完成及審查人工智能應用影響評估需執行的行動

5.3 建議

在機構中運作人工智能，需要在從策略到持續審查 / 合規等多個維度建立清晰且相互理解的基準。《人工智能道德框架》提供人工智能生命週期各個階段的架構，以及與人工智能應用評估一致的最佳實務。通過完成人工智能應用影響評估，可以識別人工智能生命週期中各項流程的不足之處。

如上文所述，人工智能評估有助機構在採用和運作人工智能的路線圖中識別不足之處，並為建立優質流程提供指引，從而引導機構以負責任和符合道德的方式利用人工智能的力量。

人工智能部署「繼續或中止」的最終決定，應由項目督導委員會 / 項目保證小組作出（視乎情況須獲資訊科技委員會 / 首席資訊官批核）。有關決定須適當平衡人工智能應用所帶來的效益與經緩解的風險。

第6章

附錄

6. 附錄

附錄A — 詞彙表

本文件所用術語及定義一覽

術語	定義
類別標籤	類別標籤是指根據數據集中另一屬性 / 特徵的值來預測其值的獨特屬性 / 特徵。換言之，類別是數據根據同一類別中其他數據集的共同特性來進行分類所屬的類別，而標籤則是人工智能模型進行分類過程的結果。
聚類分析	一種將相似數據點歸類的機器學習算法。
數據湖	儲存結構化和非結構化數據的中央儲存庫。
數據沿襲	數據沿襲描述數據從創建之初隨時間發生的轉變。
數據市場	數據倉庫的子集，專為特定業務領域（例如財務及營運）而設計。
數據倉庫	中央大型儲存庫，通常為分析用途而設計，並匯集來自不同系統來源的數據。
決策樹	一種使用樹狀圖解決問題的算法，其中每個葉節點代表類別標籤，樹的內部節點則代表各個屬性。
人在環內	指在系統的每個決策週期中可供作出人為介入的能力。
人為管控	指監督人工智能系統整體活動（包括其更廣泛的經濟、社會、法律及道德影響）的能力，以及在特定情況下決定何時及如何使用系統的能力。
人在環外	指人工智能系統在無需人為介入的情況下作出決策的能力。
K折	一種交叉驗證技術，將原始樣本隨機分割為k個大小相等的子樣本。
邏輯回歸	一種根據一組獨立值預測二元結果的分類算法。
可識別個人身份的資料	可用於識別個人身份的數據或資料，例如身份證號碼、生物特徵及地址。

術語	定義
隨機增加取樣法	一種從少數類別中隨機複製數據集，並將其加回原始訓練數據集的技術。
隨機重新取樣	一種在新的訓練數據集中創建不同類別分布的技術，旨在使新的訓練數據集更趨平衡。
隨機減少取樣法	一種從多數類別中隨機選擇數據集，以從原始訓練數據集刪除的技術。
隨機森林	一種在訓練階段構建決策樹，並輸出類別眾數或個別樹的平均預測值的分類方法。
回歸	一種通過估計變數之間的關係來進行預測的監督式機器學習技術。
回歸測試	為確認近期的代碼 / 程式更改不會對現有人工智能應用的性能產生負面影響而進行的測試。
強化學習	通過根據自身行動的反饋進行獎懲，訓練人工智能模型在處理問題時作出決策的學習方法。
風險分級準則	一套用於區分高風險人工智能應用的問題集。在提出人工智能項目之初，或當人工智能應用條件變更時，應妥為回答有關問題。
代理模型	當人工智能模型的預測不易理解或測量時所使用的工程方法，即以結果模型替代。
未見數據	人工智能模型從未「見過」的新數據。

附錄B — 參考資料

第B.1節 相關行業標準範例

以下列出已發布或制定中的人工智能相關標準。（若相關文件尚無官方中文名稱，則保留英文名稱列出）。

機構 / 組織	標準編號	標準名稱（中英文對照）
國際標準化組織（ISO）	ISO/IEC 22989:2022	資訊科技 — 人工智能 — 人工智能概念與術語 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology
	ISO/IEC 23053:2022	使用機器學習的人工智能系統框架 Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)
	ISO/IEC 20546:2019	資訊科技 — 大數據 — 概覽與詞彙 Information technology — Big data — Overview and vocabulary
	ISO/IEC 20547-1:2020	資訊科技 — 大數據參考架構 — 第1部分：框架與應用流程 Information technology — Big data reference architecture — Part 1: Framework and application process
	ISO/IEC TR 20547-2:2018	資訊科技 — 大數據參考架構 — 第2部分：使用案例與衍生需求 Information technology — Big data reference architecture — Part 2: Use cases and derived requirements
	ISO/IEC 20547-3:2020	資訊科技 — 大數據參考架構 — 第3部分：參考架構 Information technology — Big data reference architecture — Part 3: Reference architecture
	ISO/IEC 20547-4:2020	資訊科技 — 大數據參考架構 — 第4部分：安全與私隱 Information technology — Big data reference architecture — Part 4: Security and privacy

機構 / 組織	標準編號	標準名稱（中英文對照）
	ISO/IEC TR 20547-5:2018	資訊科技 — 大數據參考架構 — 第5部分：標準路線圖 Information technology — Big data reference architecture — Part 5: Standards roadmap
	ISO/IEC 24668:2022	資訊科技 — 人工智能 — 大數據分析流程管理框架 Information technology — Artificial intelligence —Process management framework for Big data analytics
	ISO/IEC TR 24027:2021	資訊科技 — 人工智能 —人工智能系統及輔助決策中的偏見 Information technology — Artificial Intelligence (AI) — Bias in AI systems and AI aided decision making
	ISO/IEC TR 24028:2020	資訊科技 — 人工智能 — 人工智能可信度概覽 Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence
	ISO/IEC TR 24029-1:2021	人工智能 — 神經網絡韌性評估 — 第1部分：概覽 Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview
	ISO/IEC TR 24368:2022	資訊科技 — 人工智能 — 道德與社會關注事項概覽 Information technology — Artificial intelligence — Overview of ethical and societal concerns
	ISO/IEC 23894:2023	資訊科技 — 人工智能 — 風險管理 Information Technology — Artificial Intelligence — Risk Management
	ISO/IEC TR 24030:2024	資訊科技 — 人工智能 — 使用案例 Information technology — Artificial Intelligence (AI) — Use cases
	ISO/IEC 38507:2022	資訊科技 — 資訊科技治理 — 機構採用人工智能的管治影響

機構 / 組織	標準編號	標準名稱（中英文對照）
		Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organisations
	ISO/IEC 24029-2:2023	人工智能 — 神經網絡韌性評估 — 第2部分：形式方法使用方法論 Artificial intelligence (AI) — Assessment of the robustness of neural networks — Part 2: Methodology for the use of formal methods
	ISO/IEC 42001:2023	資訊科技 — 人工智能 — 管理系統 Information technology — Artificial intelligence — Management system
	ISO/IEC 5259-1:2024	人工智能 — 分析與機器學習數據品質 — 第1部分：概覽、術語及範例 Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 1: Overview, terminology, and examples
	ISO/IEC 5259-2:2024	人工智能 — 分析與機器學習數據品質 — 第2部分：數據品質衡量 Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 2: Data quality measures
	ISO/IEC 5259-3:2024	人工智能 — 分析與機器學習數據品質 — 第3部分：數據品質管理要求及指引 Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 3: Data quality management requirements and guidelines
	ISO/IEC 5259-4:2024	人工智能 — 分析與機器學習數據品質 — 第4部分：數據品質流程框架 Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 4: Data quality process framework
	ISO/IEC 5259-5:2025	人工智能 — 分析與機器學習數據品質 — 第5部分：數據品質治理框架 Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 5: Data quality governance framework

機構 / 組織	標準編號	標準名稱（中英文對照）
	ISO/IEC TR 24372:2021	資訊科技 — 人工智能 —人工智能系統計算方法概覽 Information technology — Artificial intelligence (AI) — Overview of computational approaches for AI systems
	ISO/IEC TR 27563:2023	使用人工智能的安全與私隱使用案例 — 最佳實踐 Security and privacy in artificial intelligence use cases — Best practices
	ISO/IEC TR 5469:2024	人工智能 — 功能安全與人工智能系統 Artificial intelligence — Functional safety and AI systems
	ISO/IEC TS 25058:2024	系統與軟件工程 —系統及軟件品質要求與評估—人工智能系統品質評估指引 Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Guidance for quality evaluation of artificial intelligence (AI) systems
電氣電子工程師學會（IEEE）	IEEE P7000™-2021	系統設計過程中解決道德問題的模型流程 Model Process for Addressing Ethical Concerns During System Design
	IEEE P7001™-2021	自主系統的透明度 Transparency of Autonomous Systems
	IEEE P7002™-2022	數據私隱流程及方法論 Data Privacy Processes and Methodologies
	IEEE P7003™	算法偏見考量 Algorithmic Bias Considerations
	IEEE P7004™	兒童及學生數據治理標準 Standard on Child and Student Data Governance
	IEEE P7005™-2021	透明雇主數據治理標準 Standard for Transparent Employer Data Governance
	IEEE P7006™	個人數據人工智能代理工作組標準

機構 / 組織	標準編號	標準名稱（中英文對照）
		Standard on Personal Data AI Agent Working Group
	IEEE P7007™-2021	道德驅動的機械人及自動化系統的本體標準 Ontological Standard for Ethically driven Robotics and Automation Systems
	IEEE P7008™	機械人、智能及自主系統的道德驅動誘導標準 Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems
	IEEE P7009™-2024	自主及半自主系統的故障安全設計標準 Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems
	IEEE Std 7010™-2020	自主及智能系統對人類福祉影響評估實務指引 IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being
	IEEE P7011™	識別及評級新聞來源可信度的流程標準 Standard for the Process of Identifying & Rating the Trustworthiness of News Sources
	IEEE P7012™	機器可讀個人隱私條款標準 Standard for Machine Readable Personal Privacy Terms
	IEEE P7014™-2024	自主及智能系統中模擬同理心的道德考量標準 Standard for Ethical considerations in Emulated Empathy in Autonomous and Intelligent Systems
	IEEE 3119-2025	人工智能及自動決策系統採購標準 Standard for the Procurement of Artificial Intelligence and Automated Decision Systems
中國國家標準（GB）	GB 45438-2025	網絡安全技術 人工智能生成合成內容標識方法 Cybersecurity technology—Labeling method for content generated by artificial intelligence
	GB/T 45652-2025	網絡安全技術 生成式人工智能預訓練和優化訓練數據安全規範 Cybersecurity technology—Security specification for generative artificial intelligence pre-training and fine-tuning data

機構 / 組織	標準編號	標準名稱（中英文對照）
	GB/T 45654-2025	網絡安全技術 生成式人工智能服務安全基本要求 Cybersecurity technology—Basic security requirements for generative artificial intelligence service
	GB/T 45674-2025	網絡安全技術 生成式人工智能數據標注安全規範 Cybersecurity technology—Generative artificial intelligence data annotation security specification
	GB/T 45958-2025	網絡安全技術 人工智能計算平台安全框架 Cybersecurity technology—Security framework for artificial intelligence computing platform
	GB/T 45081-2024	人工智能 管理體系 Artificial intelligence—Management system
	GB/T 45628-2025	人工智能 知識圖譜 知識交換協議 Artificial intelligence—Knowledge graph—Knowledge exchange protocol
	GB/T 45288.1-2025	人工智能 大模型 第1部分：通用要求 Artificial intelligence—Large-scale model—Part 1: General requirements
	GB/T 45288.2-2025	人工智能 大模型 第2部分：評測指標與方法 Artificial intelligence—Large-scale model—Part 2: Testing and evaluation for metrics and methods
	GB/T 45288.3-2025	人工智能 大模型 第3部分：服務能力成熟度評估 Artificial intelligence—Large-scale models—Part 3: Service capability maturity assessment
	GB/T 45280-2025	人工智能 異構人工智能加速器統一接口 Artificial intelligence—Unified interfaces for heterogeneous artificial intelligence accelerating units
	GB/T 45225-2025	人工智能 深度學習算法評估 Artificial intelligence—Deep learning algorithms evaluation

機構 / 組織	標準編號	標準名稱（中英文對照）
	GB/T 45401.1-2025	人工智能 計算設備調度與協同 第1部分：虛擬化與調度 Artificial intelligence—Scheduling and cooperation for computing devices—Part 1: Virtualization and scheduling
	GB/T 45401.2-2025	人工智能 計算設備調度與協同 第2部分：分布式計算框架 Artificial intelligence—Scheduling and cooperation for computing devices—Part 2: Framework for distributed computing
	GB/T 45079-2024	人工智能 深度學習框架多硬件平台適配技術規範 Artificial intelligence—Technical specification for deep learning framework adaption to multi-hardware platform
	GB/T 45087-2024	人工智能 服務器系統性能測試方法 Artificial intelligence—Management system
	GB/T 46347-2025	人工智能 風險管理能力評估 Artificial intelligence—Risk management capability assessment
	GB/T 41867-2022	信息技術 人工智能 術語 Information Technology Artificial Intelligence Terminology
	GB/T 42018-2022	信息技術 人工智能 平台計算資源規範 Information Technology Artificial Intelligence Platform Computing Resource Specifications
	GB/T 42131-2022	人工智能 知識圖譜技術框架 Artificial Intelligence Knowledge Graph Technical Framework
	GB/T 40691-2021	人工智能 情感計算用戶界面 模型 Artificial intelligence—Affective computing user interface—Model
	GB/T 43782-2024	人工智能 機器學習系統技術要求

機構 / 組織	標準編號	標準名稱（中英文對照）
		Artificial intelligence—Technical requirements for machine learning system
	GB/T 42755-2023	人工智能 面向機器學習的數據標註規程 Artificial intelligence—Code of practice for data labeling of machine learning
	GB/T 41818-2022	信息技術 大數據 面向分析的數據存儲與檢索技術要求 Information technology Big data Analysis-oriented data storage and retrieval technical requirements
	GB/T 38675-2020	信息技術 大數據計算系統通用要求 Information technology—General requirements for big data computing systems
	GB/T 38643-2020	信息技術 大數據 分析系統功能測試要求 Information technology—Big data—Functional testing requirements for analytic system

第 B.2 節 政策 / 指引 / 措施

以下列出已發布的人工智能相關規例 / 法律 / 政策 / 指引 / 措施。

人工智能項目亦須遵守的其他相關法律及規例，包括但不限於以下：

- 私隱條例（《個人資料（私隱）條例》（第486章））；
- 知識產權條例（《版權條例》（第528章）、《防止盜用版權條例》（第544章）、《商標條例》（第559章）、《商品說明條例》（第362章）、《專利條例》（第514章）、《註冊外觀設計條例》（第522章）；
- 反歧視條例（《性別歧視條例》（第480章）、《殘疾歧視條例》（第487章）、《家庭崗位歧視條例》（第527章）、《種族歧視條例》（第602章））；
- 《中華人民共和國香港特別行政區基本法》（簡稱《基本法》）；
- 《維護國家安全條例》；以及
- 《中華人民共和國香港特別行政區維護國家安全法》（簡稱《香港國安法》）。

政策 / 指引 / 措施：

- 香港生成式人工智能技術及應用指引
  Hong Kong Generative Artificial Intelligence Technical and Application Guideline

其他相關指引：

個人資料私隱專員公署

- 開發及使用人工智能道德標準指引
  Guidance on Ethical Development and Use of AI
- 使用AI聊天機械人「自保」十招
  10 TIPS for Users of AI Chatbots
- 人工智能 (AI): 個人資料保障模範框架
  Artificial Intelligence: Model Personal Data Protection Framework
- 僱員使用生成式AI的指引清單
  Checklist on Guidelines for the Use of Generative AI by Employees

香港金融管理局

- 應用人工智能的高層次原則
  High-level Principles on Artificial Intelligence
- 認可機構就採用大數據分析及人工智能的消費者保障（暫時只備英文版本）
  Consumer Protection in respect of the Use of Big Data Analytics and Artificial Intelligence by Authorised Institutions
- 應用生成式人工智能的消費者保障
  Consumer Protection in respect of Use of Generative Artificial Intelligence
- 應用人工智能監察可疑活動（暫時只備英文版本）
  Use of Artificial Intelligence for Monitoring of Suspicious Activities

司法機構

- 香港司法機構法官及司法人員和支援人員使用生成式人工智能的指引
  Guidelines on the Use of Generative Artificial Intelligence for Judges and Judicial Officers and Support Staff of the Hong Kong Judiciary

財經事務及庫務局

- 有關在金融市場負責任地應用人工智能的政策宣言

  Policy Statement on Responsible Application of Artificial Intelligence in the Financial Market

證券及期貨事務監察委員會

- 生成式 AI 語言模型的使用

  Use of generative AI language models

中國內地人工智能相關規例 / 法律

- 生成式人工智能服務管理暫行辦法

  Interim Measures for the Management of Generative AI Services
- 人工智能生成合成內容標識辦法

  Measures for Labeling of AI-Generated Synthetic Content

中國內地相關框架 / 指引

- 人工智能安全治理框架

  AI Safety Governance Framework
- 中國關於加強人工智能倫理治理的立場文件

  Position Paper of the People's Republic of China on Strengthening Ethical Governance of AI
- 全球人工智能治理倡議

  Global AI Governance Initiative

附錄C — 人工智能應用影響評估範本

備註

有關完成人工智能應用影響評估的詳情，請參閱第5.1.1節「完成人工智能應用影響評估的流程」。

說明

- 文字涵蓋須以質性方式處理的核心問題
- 斜體文字為補充說明，旨在輔助回答核心問題，應用作提供質性回答的參考。

風險分級準則

若以下任何一條（或以上）風險分級問題被觸發（即對該問題的回答為「是」），則該人工智能應用影響評估應在項目啟動前提交資訊科技委員會／首席資訊官審視。詳情請參閱第5章「人工智能評估」。

問題	背景說明
a - 該人工智能應用是否涉及公眾高度關注（例如涉及私隱問題）及／或訴訟頻繁的領域？	*舉例而言，某些「物聯網」應用可能對個人日常生活及私隱構成重大影響。相關應用案例包括智慧城市項目：* • *智能照明──根據天氣自動調節的智慧路燈* • *智能交通管理──監測車輛與行人流量* • *智能泊車系統──監測泊車位使用狀態* *因此，這類應用的人工智能應用影響評估須由資訊科技委員會／首席資訊官審視。若結合人臉識別技術與全新、敏感或可能具爭議性的用途，尤須審慎評估。*
b - 該人工智能應用是否用於全新（社會）領域（即香港從未在該領域應用人工智能）？	*例如某人工智能應用首次用於醫療保健領域，而此前僅用於營銷用途。由於應用領域轉變，該人工智能應用可能引發（新的）道德爭議。如在敏感社會領域應用人工智能，風險與道德問題可能*

問題	背景說明
	更為嚴峻，例如涉及照顧服務、公共安全、打擊恐怖主義或教育等議題；又或涉及兒童或殘疾人士等弱勢社群。
c - (i) 該人工智能應用是否具備高度自主性？ 若問題(i)回答「是」，請繼續回答問題(ii)。 (ii) 該應用是否用於複雜環境或關鍵資訊基礎設施？ 若問題(ii)回答「是」，請繼續回答問題(iii)。 (iii) 該人工智能應用的自動化決策是否會對個人 / 實體產生重大影響，或導致其承擔法律責任後果？ 若問題(iii)回答「是」，此類應用很可能被界定為高風險應用。	*人工智能應用獨立行事的自由度越高，越要嚴謹分析其自主性可能引發的後果。除決策自由度外，自主性亦體現於能否自主篩選數據來源。* *當人工智能應用處於複雜環境時，其風險遠高於封閉環境中的應用。* *當人工智能應用在無人為干預的情況下自動決策，且該決策可能導致個人承擔法律責任或遭受重大影響時（例如：無法獲批按揭貸款、失業、醫療誤診，或因特定分類導致聲譽受損、個人遭受排斥或歧視），風險級別將顯著提升。若處理程序對個人影響輕微或毫無影響，則不符合此項準則。* *此類應用案例包括：自動駕駛車輛、自主軍事無人機、外科手術機械人。該等應用被視為處於複雜環境，因其決策取決於綜合多變的環境因素（例如周邊人類活動）。自動駕駛車輛及自主軍事無人機等應用被界定為具有高度自主性，因其決策完全由人工智能系統作出。*
d - 是否使用敏感且可識辨個人身份的資料？	*當敏感且可識辨個人身份的資料應用於人工智能系統的開發及／或部署時，相關風險程度將較高。例如：涵蓋種族或民族本源、政治立場、宗教或哲學信念、工會會員身份、基因數據、生物辨識資料、健康相關紀錄、或自然人性生活及性取向等類別的資料。*
e - 該人工智能應用是否涉及複雜決策？	*當人工智能應用的決策過程越趨複雜（例如涉及更多變數或基於個人特徵的概率性評估），其風險亦隨之增加。基於有限選項及變數的簡單人工智能應用，風險相對較低。若人工智能應用的決*

問題	背景說明
	策過程無法（完全）被人類理解或追溯，則該決策所衍生的風險可能更大。
f - 該人工智能應用是否涉及系統性觀察或監測？	*處理程序用於觀察、監測或控制個人，包括透過網絡收集的數據或「對公眾可進入區域的系統性監測」。實例包括大規模視像監控數據及網絡行為追蹤。此類監測被列為風險準則，原因在於當事人可能在不知情的情況下被收集個人資料，且其不知悉資料收集者身份及後續用途。此外，在公共（或公眾可進入的）場所，個人可能無法避免成為此類處理程序的對象。*
g - 該人工智能應用是否涉及對個人進行評估或評分？	*包括分析及預測，特別是針對「數據主體的工作表現、經濟狀況、健康狀況、個人偏好或興趣、可靠性或行為、位置或行蹤等方面的評估」。* *此類應用的例子可能包括：* • *金融機構根據數據庫對客戶進行篩查，以進行信用參考、打擊洗錢、反恐或詐騙檢查；* • *生物科技公司直接向消費者提供基因檢測，以評估及預測疾病或健康風險；或* • *公司基於用戶在其網站上的使用或瀏覽行為，建立行為或營銷檔案。*
h - 是否涉及大規模個人資料處理及／或數據集合併？	*雖然「大規模」難以明確界定，但在判斷有關的處理程序是否屬於大規模時，可考慮以下因素：* *a. 涉及的個人數量，可以是具體數字或佔相關人口的比例；* *b. 處理的數據量及／或不同數據項目的範圍；* *c. 數據處理活動的持續時間或永久性；或* *d. 處理活動的地理範圍。*

人工智能應用影響評估問題

A.人工智能應用目的與問責	實務指引參考章節 / 責任	評估結果 / 建議緩解方案
1 - 此人工智能及數據活動的業務需求 / 目標 / 目的是什麼？所開發的人工智能應用是否有明確界定的清晰目的（例如提升營運效率或降低成本）？如何能令預期效益大於潛在風險？此計劃活動的驗收或成功標準是什麼？ *該人工智能應用會否納入一項目前正考慮或進行中的工作，而後者所涉的範疇更大？這項計劃是否符合你的人工智能及 / 或數據策略？*	第4.1.2.1節： • 將人工智能項目對應至機構目標 • 篩選人工智能項目，並訂定優先順序 資訊科技規劃人員 / 管理層、業務使用者	
2 - 此倡議對各持份者群組（參與及 / 或受數據或技術應用場景影響的每個實體或群體）的目的是什麼？請從明確成果 / 目標的角度闡述目的，並說明這些成果如何對應「人工智能道德原則」或社會普遍接受的外部正面理據 / 驗證。此外，描述可能為各持份者帶來的風險。 • 明確說明待解決的問題 • 清晰陳述項目的可衡量目標或成果 • 哪些持份者受此人工智能應用影響？ • 所有受影響持份者的（高層次）利益是什麼？ • 對受影響持份者的風險是什麼？（將於問題3分析及記錄） 識別該人工智能應用相關的「人工智能道德原則」——確保該應用未觸發任何「風險分級準則」（參見上文）。	第4.1.2.1節： • 將人工智能項目對應至機構目標 • 篩選人工智能項目，並訂定優先順序 資訊科技規劃人員 / 管理層、業務使用者	

A.人工智能應用目的與問責	實務指引參考章節 / 責任	評估結果 / 建議緩解方案
持份者包含哪些群體？請識別並列舉。（例如： 外部持份者：使用者、間接受影響的公眾、數據提供者、人工智能供應商、監管機構； 內部持份者：員工、項目督導委員會、資訊科技委員會、首席資訊官、業務單位）。本項目對下列「人工智能道德原則」的初始影響為何：(1) 公平、(2) 多元與包容、(3) 人為監督、(4)合法與合規、(5) 數據私隱、(6) 安全、(7) 問責、(8) 有益的人工智能、(9) 合作與開放及 (10) 可持續性與公正轉型。 各持份者的目標應具備明確可描述性，並須對應經外部驗證的目標（例如更廣泛的公共政策目標）。可考慮的目標包括：更優質或成本更低的醫療照護服務；更容易獲取的醫療服務或更好的健康成果；更完善的健康成果追蹤與評估能力；更精準的健康檢測或診斷裝置以提升整體健康狀態；教育品質的提升；環境的改善（如節約用水、降低能源成本）；基礎建設的提升；經濟的改善；更容易獲取／易用的科技； 更多的就業機會；保障合理的私隱預期（包含匿名權）；維護權益及防止歧視。		
3 - 對各持份者的潛在風險是什麼？哪些風險分級準則已被觸發？考量人工智能應用相關的所有因素，對每個已識別的持份者產生哪些實際及 / 或感知的風險？	第4.1.2.2節： • 評估人工智能項目是否需要接受資訊科技委員會	

A.人工智能應用目的與問責	實務指引參考章節 / 責任	評估結果 / 建議緩解方案
說明人工智能項目及應用可能產生的高層級潛在影響及／或顧慮，以及項目可能引發的風險。若項目觸發任何風險分級準則，則須在項目開始前獲得資訊科技委員會／首席資訊官的初步核准（風險分級準則請參閱上文表格）。 *考量人工智能應用對已識別持份者造成的風險或增加的風險（實際或感知），需要評估的範圍包括：對使用技術或數據而產生影響力的觀感；對員工關係的影響；地位及／或幸福感降低；聲譽受損或尷尬；對處理活動或處理結果感到震驚或意外；不當歧視；機構不當存取或濫用資訊（如洞察或預測數據）的可能性，包括敏感類別數據及可直接識別的數據；對個人需求或欲望的操縱（即創造原本不存在的需求）；透過以概率為基礎的流程（如評分）對技術產生的負面影響；誰將能存取人工智能應用的相關資訊？誰不能？無法存取此資訊、數據或洞察的持份者是否會比能存取者處於劣勢？該劣勢為何？在「擁有」與「未擁有」此資訊者之間會產生哪些新的差異？個別持份者是否會對與其相關的活動感到意外？關於個人的資料使用方式是否符合個人對該數據／資料應如何使用的認知？判斷潛在的洞察使用是否存在其他敏感性問題，以及潛在洞察使用的哪些方面可能被持份者視為不公平。所有持份者是否受到平等對待？*	/ 首席資訊官審查的一般準則 資訊科技規劃人員 / 管理層、業務使用者	

A.人工智能應用目的與問責	實務指引參考章節 / 責任	評估結果 / 建議緩解方案
4 - 是否已滿足「人工智能實務指引」中概述的所有治理流程與角色要求？具體來說，由誰負責履行各項要求？ *需要釐清角色與責任以落實人工智能運作並確保道德責任歸屬。各項工作由誰負責？為此人工智能應用設立的治理機構其性質或責任為何？人工智能模型驗證與模型開發是否由不同團隊負責？*	第4.1.2.2節： • 確保關鍵流程設有監督機制 • 界定角色與責任 資訊科技規劃人員 / 管理層、業務使用者	
5 - 誰對該人工智能應用擁有最終決策權？	第4.1.2.2節： • 確保關鍵流程設有監督機制 • 界定角色與責任 資訊科技規劃人員 / 管理層、業務使用者	

B. 人工智能治理流程 - 項目規劃	實務指引參考章節 / 責任	評估結果 / 建議緩解方案
6 - 該項目計劃是否涵蓋人工智能應用生命週期的各階段，包括變更管理及任何組織變更？ *項目是否符合既定的項目管理要求？是否已全面考量功能和技術要求並予以記錄？是否存在路線圖以應對未來可能的需求、必要的人工智能應用更新以及角色與責任？*	第4.1.2.2節： • 在人工智能生命週期中全面考慮所有「人工智能道德原則」 • 確保滿足適當的資源需求 • 確保關鍵流程設有監督機制 • 界定角色與責任 資訊科技規劃人員 / 管理層、業務使用者	

B. 人工智能治理流程 - 項目規劃	實務指引參考章節 / 責任	評估結果 / 建議緩解方案
7 - 所有項目團隊成員是否清楚理解自身角色及項目計劃？相關人員及/或合約資源是否具備承擔項目責任所需的技能與知識？如何管理治理控制措施，以確保開發或實施流程具備一致性、穩健性及可重複性？ *所有項目持份者是否理解人工智能算法的概率特性，認知所有輸出結果並非百分之百確定與正確？若否，則應在人工智能項目啟動前完成相關技能提升計劃或培訓（可考慮外聘培訓人員），或聘請具備相關技能的新員工／顧問。此舉是為了確保項目團隊成員具備執行項目所需的必要技能與知識。*	第4.1.2.2節： • 確保關鍵流程設有監督機制 • 界定角色與責任 資訊科技規劃人員 / 管理層、業務使用者	
8 - 該人工智能項目的明確及預期應用範圍是什麼？如何選擇適用於當前任務的人工智能模型？人工智能應用的明確業務成功標準是什麼？ *（任務範例：你有否驗證所選用的人工智能模型指標是否適當？你有否評估人工智能模型的可靠性？你有否測試人工智能模型的敏感度？是否存在識別人工智能模型漏洞的程序？）*	第4.1.2.1節： • 篩選人工智能項目並訂定優先順序 資訊科技規劃人員 / 管理層、業務使用者	
9 - (i) 「公平」如何被界定？ (ii) 有何措施以量度及測試從而達致「公平」？ *鑒於公平並無單一定義且能同等地適用於不同人工智能應用，目標是盡可能偵測並減輕與公平相關的危害。人工智能*	第4.1.2.2節： • 在人工智能生命週期中全面考慮所有「人工智能道德原則」 第4.1.4.4節：	

B. 人工智能治理流程 - 項目規劃	實務指引參考章節 / 責任	評估結果 / 建議緩解方案
應用可能因以下原因產生不公平行為：用作訓練的數據集存在固有偏見、開發團隊決策中明示或隱含的偏見，或於部署後與特定持份者互動時引發不公平行為。危害與風險類型可包括分配、服務品質、刻板印象、貶損、代表性過高或不足。不同持份者群組的預期效益與潛在危害之間的權衡亦可能對此造成影響。減緩措施包含設立程序以審查系統願景，以及對持份者群組可能造成的公平性相關危害。考慮界定並審查系統架構，例如機器學習模型、效能指標、用戶界面、開發及測試系統所需的數據集。依據既定公平性標準審查生產數據集。考慮於推出前進行發佈前評審與代碼審查。建立定期產品審查會議。	• 界定與測試公平性及偏見 第4.1.4.5.4節： • 執行正則化及其他選擇模型的方法 資訊科技規劃人員 / 管理層、業務使用者、數據科學家、項目經理、系統分析員、系統架構師	
10 - 如何保持數據、實驗、人工智能模型版本及使用情況的可追溯性？如何掌握成效與成功標準的對比？ *可追溯性與成效有助提升人工智能解決方案的可解釋性、透明度及可靠性。未維持可追溯性的風險在於人工智能可能出現非預期表現，且無法向用戶解釋相關原因。* *緩解措施可包括記錄用於設計及開發算法系統的方法，內容可涵蓋：* *a) 基於規則的人工智能應用：編程方法或模型構建方式；* *b) 基於學習的人工智能應用：算法訓練方法，包括蒐集與篩選哪些輸入數據及*	第4.1.4.3節： • 完成數據來源文件記錄 第4.1.5.4節： • 實現可追溯性 系統分析員、系統架構師、項目經理、數據科學家	

B. 人工智能治理流程 - 項目規劃	實務指引參考章節 / 責任	評估結果 / 建議緩解方案
其過程。描述用於測試及驗證算法系統的方法： *a) 基於規則的人工智能應用：用於測試及驗證的場景或使用案例；* *b) 基於學習的模型：用於測試及驗證的數據相關資訊。描述算法系統的輸出結果：* *算法產生的結果或決策，以及不同使用案例（例如針對其他用戶子群組）可能導致的潛在決策。*		

B.人工智能治理流程—項目生態環境	實務指引參考章節 / 責任	評估結果 / 建議緩解方案
11 - 在人工智能模型開發或應用過程中，將收集、追蹤、轉移、使用、儲存或處理哪些特定類型和來源的數據？所使用的數據是否可識辨個人身分？ *這些數據能否識別到特定個人？數據是否匿名（或去識別化），以及採用了哪些政策、流程和／或技術措施以盡量避免數據被重新識別至特定個人？該項目所使用的數據是否有最新的數據字典？你有否考慮過人工智能模型可能利用特徵組合來揭示特殊類別，導致處理程序對數據所代表的個人不公平？你的人工智能模型有否收集和使用生物特徵數據？數據可能來自其他內部來源或自外部來源獲取。若使用可識辨個人身分的數據，應進行私隱影響評估以識別私隱風險，並採取緩解措施來降低已識別的*	第4.1.4.3節： • 執行數據驗證 • 確保公共或第三方數據的可信度 • 完成數據來源文件記錄 系統分析員、系統架構師	

B.人工智能治理流程—項目生態環境	實務指引參考章節 / 責任	評估結果 / 建議緩解方案
風險，同時確保遵守《個人資料（私隱）條例》。		
12 - 你是否記錄數據沿襲以了解載入機器學習模型的數據來源、路徑、授權或其他義務與轉換？你是否已設立程序，以驗證數據使用或人工智能模型符合任何適用授權？應考慮附加數據地圖。 *已設立哪些程序確保輸入數據符合人工智能應用的用途？是否設有程序驗證數據使用或人工智能模型符合任何適用授權（若數據或技術是由第三方提供）？* *未記錄數據沿襲的風險包括技術風險，即數據可能未被理解或不適用，以致影響「人工智能道德原則」，如公平、透明度、可解釋性及數據私隱。緩解措施可包含使用數據地圖將數據元素從來源映射至模型，藉此提供所用數據的總覽作為檢查清單，以識別各數據來源是否具備適用的授權。*	第4.1.4.3節： • 確保公共或第三方數據的可信度 • 完成數據來源文件記錄 系統分析員、系統架構師、項目經理、業務使用者	
13 - 數據的準確度是否足以符合人工智能模型及倡議活動的用途？所用數據集是否可信且來源可靠？你是採用哪種技術驗證來源數據的可靠性？ *人工智能的建模數據在數據樣本、規模及分布方面必須適當，以確保人工智能模型能作出具有意義和代表性的推論。現時正採取哪些步驟判定來源數據的準確度，以及來源數據是否長期維持充分的準確度？數據整合／轉換是否已影響*	第4.1.4.3節： • 執行數據驗證 • 確保公共或第三方數據的可信度 系統分析員、系統架構師	

B.人工智能治理流程—項目生態環境	實務指引參考章節 / 責任	評估結果 / 建議緩解方案
其準確度？最終數據集的質素相對於活動目的是否存在疑慮？可能受影響的「人工智能道德原則」包括公平、可靠、穩健及安全。有關處理數據可靠性的詳情，請參閱第4.1.4.3節「數據提取」。		

B. 人工智能治理流程—第三方問題	實務指引參考章節 / 責任	評估結果 / 建議緩解方案
14 - 你的人工智能應用或大數據項目是否計劃以供應商或合作夥伴形式使用第三方技術或數據？（是 / 否） *第三方通常指外部供應商、服務提供者或合作夥伴。如否，請跳過第三方相關部分。*	不適用。此問題旨在識別機構是否需要填寫此部分。	
15 - 如機構曾經從第三方獲取人工智能模型或數據集，是否已評估及管理使用這些資源的風險？	第4.1.4.3節： • 確保公共或第三方數據的可信度 第4.1.3.2節： • 界定人工智能應用的要求 系統分析員、系統架構師、採購團隊、項目經理、業務使用者	
16 - 第三方須提供哪些文件？你是否已要求並取得詳細文件？有關文件是否符合訂明的要求？	第4.1.3.2節： • 界定人工智能應用的要求 第4.1.4.5.5節：	

B. 人工智能治理流程—第三方問題	實務指引參考章節 / 責任	評估結果 / 建議緩解方案
應評估機構的數據對供應商人工智能模型的適用性。針對機構計劃的人工智能模型使用組合，應進行回溯測試、模型驗證及結果分析。 *若第三方無法提供相關文件，機構可基於業務考量決定承擔缺乏有關文件的風險，但須評估可能產生的後果，並因應每次情況作個別考慮。緩解措施可包括僅將這些模型用於低風險的人工智能應用（例如：不會觸發本評估風險分級準則中任何初始觸發條件並只供內部使用的應用）。*	• 使用可信賴的雲端 / 第三方服務 採購團隊、項目經理、業務使用者	
17 - 第三方（例如供應商、終端使用者、受試對象、分銷商 / 供應商或工作人員）舉報人工智能應用潛在漏洞、風險或偏見的程序是什麼？ *機構應要求第三方舉報人工智能應用中任何潛在漏洞、風險或偏見。此項要求應明確載列於第三方合作合約中。若無法執行此要求，機構應對潛在漏洞、風險或偏見，以及可能引致的後果進行風險評估。*	第4.1.3.2節： • 界定人工智能應用的要求 採購團隊、項目經理、業務使用者	

B. 人工智能治理流程 - 項目開發	實務指引參考章節 / 責任	評估結果 / 建議緩解方案
18 - 人工智能模型開發將使用哪些類型及來源的數據？該人工智能應用是否會使用個人資料作為輸入數據？你已建立哪些流程以確保數據品質及完整性？請	第4.1.4.3節： • 執行數據驗證 • 確保公共或第三方數據的可信度	

B. 人工智能治理流程 - 項目開發	實務指引參考章節 / 責任	評估結果 / 建議緩解方案
說明將採用的數據淨化步驟。請描述你如何確認所處理的數據足夠和相關，且僅限於達成模型目標所需。 *數據淨化步驟複雜，容易意外產生難以識別的用戶錯誤，從而導致錯誤的建模結果。你是否需要進行或修訂私隱影響評估？用戶期望你準確分享、量度或收集哪類數據？緩解措施包括由經驗豐富的系統分析員／系統架構師執行數據淨化步驟並建立測試程序。*	第4.1.4.4節： • 執行私隱影響評估 • 實施數據匿名化 系統分析員、系統架構師、項目經理	
19 - 訓練數據經過何種方式評估品質？訓練樣本總數是否足夠？有關樣本能否充分代表不同社會群體，包括種族、性別、膚色、年齡、收入等？ *你有否考慮所使用數據中用戶的多樣性和代表性？一般而言，訓練數據、測試數據與驗證數據的基準比例通常設定為60:20:20或70:15:15。若未使用驗證數據，則訓練數據與測試數據的建議樣本數量比例為80:20。詳情請參閱第4.1.4.3節「數據提取」。若數據品質及／或訓練數據的代表性不足，應在本評估的C部分說明潛在影響，以評估風險。*	第4.1.4.4節： • 比較訓練數據、驗證數據與測試數據 項目經理、系統分析員、系統架構師、數據科學家	
20 - 你曾如何測試人工智能模型的性能？是否已對人工智能模型進行充分訓練，並採用不同指標（如精確率、召回率、F1值、準確率等）進行分析？例如使用scikit 指標（https://scikit-learn.org/stable/modules/model_evaluation.html）。	第4.1.5.1節： • 執行整合、系統、決策及用戶驗收測試 數據科學家、項目經理	

B. 人工智能治理流程 - 項目開發	實務指引參考章節 / 責任	評估結果 / 建議緩解方案
你考慮了哪些效能指標？如何執行這些測試？有否顧及不同子群體（例如受保障群體）的性能差異？ 若未測試人工智能模型的性能，應在本評估的C部分說明評估模型性能不佳的潛在影響，以評估風險。必要時應使用效能指標來緩解風險。		
21 - 所有技術或數據安全要求是否均已滿足？你如何核實數據集未遭竄改或入侵？ *你的風險分析是否包含：安全或網絡問題（例如網絡安全威脅）是否可能因人工智能應用的非預期行為而構成安全風險或損害？你有否制定措施或系統以確保人工智能應用面對潛在攻擊時的完整性與韌性？若不良份子能存取有關數據，會用作什麼用途？若數據被盜或外洩，最壞可被用作什麼用途？* *應執行保安風險評估及審計，以識別、分析及評估保安風險，並制定緩解措施，將風險降至可接受水平。*	第4.1.5.2節： • 實施模型 / 應用程式介面存取安全措施 數據科學家、項目經理	
22 - 該人工智能應用在面臨設計或技術故障、缺陷、中斷、攻擊、誤用、不當或惡意使用等風險或威脅時，會否產生對抗性、關鍵性或破壞性影響（例如危害人類或社會安全）？這些風險是如何被識別與緩解的？ *你有否考慮到不同類型的漏洞及潛在攻擊入口，例如：*	第4.1.4.5.5節： • 使用可信賴的雲端 / 第三方服務 第4.1.4.5.6節： • 執行對抗性訓練 採購團隊、數據科學家、項目經理	

B. 人工智能治理流程 - 項目開發	實務指引參考章節 / 責任	評估結果 / 建議緩解方案
• *數據投毒（即訓練數據遭篡改）；或* • *模型規避（即依照攻擊者意圖分類數據）；* • *模型反轉（即推斷模型參數）。* *您是否已實施措施以確保人工智能應用在其生命週期中對抗潛在攻擊時的完整性、穩健性及整體安全性？您有否對該應用執行滲透測試？在進行人工智能應用安全評估時，建議考慮納入獨立第三方。風險緩解措施包含第4.1.4.5.6節所提及的對抗性訓練。*		
23 - 請說明將採用的驗證流程。 *是否設有技術審查流程？會否進行任何獨立審查？整合、系統、決策及用戶驗收測試均屬驗證措施的一環。*	第4.1.5.1節： • 執行整合、系統、決策及用戶驗收測試 數據科學家、項目經理	
24 - 該人工智能應用是否將取代需要判斷或酌情處理的人類決策？該人工智能應用將做出哪種自動化決策？這些決策是否會對個人產生法律或類似影響？ *你有否評估人工智能模型是否需要人為介入？處理程序旨在對數據主體做出「對個人產生法律效力」或「同樣地嚴重影響自然人」的決策。舉例而言，該處理程序可能導致個人遭排除或歧視。*	第4.1.4.2節： • 決定適當的人為介入程度 項目經理、業務使用者	
25 - (i) 該人工智能模型及其輸出能否以簡單方法解釋？	第4.1.4.2節：	

B. 人工智能治理流程 - 項目開發	實務指引參考章節 / 責任	評估結果 / 建議緩解方案
(ii) 為向受影響持份者有效解釋所制訂的溝通計劃為何？ (iii) 該人工智能應用是否能在需要時為其決策或建議提供理由？ *你的算法的運作方式，是否透明地向受影響人士傳達？認為自己被錯誤或不公平評估的人士，有沒有任何方法申訴？（情境範例：針對非技術持份者；考慮需要解釋的目的和背景；當技術可解釋性／明確的解釋可能對受眾沒有用處時）。當人工智能模型及其輸出無法簡單解釋時，可考慮把工作流程繪製成包含所有可能的決策樹，讓終端使用者能設想模型邏輯。若人工智能模型無法被解釋，則應考慮該應用及其決策的影響，並進行風險評估。*	• 訂定人工智能模型透明度與可解釋性要求 項目經理、業務使用者	
26 - 當前的人工智能應用，有何其他可行但可能需要更多人工操作的替代方案？這些方案在準確度指標及業務、法律、經濟、社會風險等方面，相對於人工智能應用的表現如何？人工方法的風險是否比人工智能應用更低？其成本是否更高或需更大投資？	第4.1.4.1節： • 界定業務需求 項目經理、業務使用者	

B. 人工智能治理流程 -系統部署	實務指引參考章節 / 責任	評估結果 / 建議緩解方案
27 - 請描述你涵蓋安全或網絡問題（例如網絡安全及對抗性攻擊）的風險分析？	第4.1.5.1節： • 執行整合、系統、決策及用戶驗收測試	

B. 人工智能治理流程 -系統部署	實務指引參考章節 / 責任	評估結果 / 建議緩解方案
（風險範例：人工智能系統出現意料之外的行為所導致的安全風險或損害；數據外洩的影響；若不良份子能存取有關數據，會用作什麼用途？）	數據科學家、項目經理	
28 - 所有使用者是否受到平等對待？若否——且你的人工智能算法及預測技術針對不同使用者優先處理特定資訊、設定不同的價格或給予不同的存取權限——你將如何應對所有使用者受平等對待（或至少透明地不平等對待）的消費者訴求或政府規例？	第4.1.2.2節： • 在人工智能生命週期中全面考慮所有「人工智能道德原則」 數據科學家、項目經理	
29 - 你已建立哪些機制以監測及記錄人工智能應用的表現（例如準確率）與驗收標準？在部署前，你已採取哪些步驟來全面測試人工智能應用，以確認其是否會故障及行為是否符合預期？相關回退方案為何？ *若人工智能應用的準確率偏低，是否可能導致關鍵性、對抗性或破壞性後果？你是否已制定措施，確保用於開發人工智能應用的數據（包括訓練數據）屬最新、優質和完整，且就部署該應用的環境而言具代表性？*	第4.1.6.1節： • 設定效能指標 • 納入品質評估 數據科學家、項目經理	
30 - (i) 若該人工智能應用的可靠性及 / 或可再現性偏低，是否可能導致關鍵性、對抗性或破壞性後果（例如涉及人類安全）？	第4.1.5.4節： • 確保端到端工作流程具可重複性與可再現性 第4.1.6.1節：	

B. 人工智能治理流程 -系統部署	實務指引參考章節 / 責任	評估結果 / 建議緩解方案
(ii) 是否已建立明確流程以監測該人工智能應用是否符合預期目標？ (iii) 曾否測試為確保可再現性，是否需要將特定情境或條件納入考慮？ *你是否已測試有關人工智能模型用於人口中不同群體的表現，以減低系統性偏見？*	• 設定效能指標 • 納入品質評估 數據科學家、項目經理	
31 - 在設計人工智能應用時，你有否充分界定「公平性」在應用上的定義？請就此具體說明。關於避免在人工智能應用中產生或強化不公平的偏見，包括使用輸入數據和算法設計兩方面，你的策略或程序是什麼？ *在你的使用案例中，不公平偏見的定義是什麼？請說明用於評估各項偏見的指標。請描述現行用於偵測、識別及減輕不公平偏見的最佳實務做法。你的風險分析架構是什麼？請說明你在使用案例中識別出的不公平偏見風險，並以終端使用者特徵描述你評估的偏見所針對的群體。* *請參閱前述問題 9。此應於項目規劃階段明確界定。若模型中存在不公平偏見，可能產生哪些影響？*	第4.1.4.4節： • 界定與測試公平性及偏見 項目經理、業務使用者、數據科學家	
32 - 請說明在人工智能應用的開發、部署及使用階段中，用以測試與監察潛在負面歧視（偏見）的流程是什麼？	第4.1.6.1節： • 設定效能指標 • 納入品質評估	

B. 人工智能治理流程 -系統部署	實務指引參考章節 / 責任	評估結果 / 建議緩解方案
該人工智能應用是否可能對人造成負面歧視？請說明為減輕任何已偵測到的偏見而實施的管控措施。在你處理過程中所使用的數據，是否能代表該人工智能應用所適用的群體？	• 建立異常檢測技術 數據科學家、項目經理	
33 - 該人工智能應用是否設計為與人類終端使用者互動、提供指導或作出影響人類或社會的決策？該人工智能應用是否可能使部分或全部終端使用者或相關人士對於某項決策、內容、建議或結果是否源自算法決策產生困惑？請說明如何確保終端使用者或其他人士充分知悉某項決策、內容、建議或結果是源自算法決策？註：若該人工智能應用與個人無關，則此問題不適用。 *該人工智能應用是否可能使部分或全部終端使用者或相關人士對於他們是在與人類還是人工智能應用互動產生困惑？*	第4.1.5.2節： • 向用戶提供資訊 數據科學家、項目經理	
34 - 當人工智能應用提供不準確的結果時，你如何評估可能造成的影響？ *若人工智能應用作出不準確的預測時，你是否已核實會導致哪些危害？*	第4.1.5.1節： • 使用邊緣案例與例外處理 數據科學家、項目經理	
35 - 請說明你已制定哪些步驟來向使用者解釋人工智能應用的決策？註：若該人工智能應用與個人無關，則此問題不適用。	第4.1.5.2節： • 向用戶提供資訊 數據科學家、項目經理	

B. 人工智能治理流程 -系統部署	實務指引參考章節 / 責任	評估結果 / 建議緩解方案
針對互動式人工智能應用（例如聊天機械人、機械人律師），你是否有告知使用者他們正在與人工智能應用而非人類互動？你是否有建立機制讓使用者了解人工智能應用所生成的決策之目的、標準與限制？這涉及「人工智能道德原則」中的透明度與可解釋性原則，以讓使用者知悉有關的人工智能決策。		

B. 人工智能治理流程—系統運作與監察	實務指引參考章節 / 責任	評估結果 / 建議緩解方案
36 - 你是否有處理人工智能事件的計劃？請說明為了處理人工智能應用對終端使用者或相關人士造成的不良影響所設立的監測與應對機制。 *你是否已確保設有「停止按鈕」或程序，在需要時可安全中止操作？你有否採取任何特定監督及管控措施，以反映人工智能應用的自我學習或自主特性？若無安全中止程序，人工智能應用可能造成的潛在不良影響是否可以接受？*	第4.1.5.2節： • 建立多層緩解機制 第4.1.5.3節： • 追蹤錯誤 第4.1.6.1節： • 納入品質評估 第4.1.6.2節： • 制定或運用事故處理計劃以應對危機情況 數據科學家、項目經理	
37 - (i) 你是否已設有納入客戶 / 使用者回饋的既定流程？ (ii) 你是否已設有識別人工智能應用弱點的流程？	第4.1.5.3節： • 追蹤錯誤 第4.1.6.1節： • 納入品質評估	

B. 人工智能治理流程—系統運作與監察	實務指引參考章節 / 責任	評估結果 / 建議緩解方案
(iii) 針對可能發現的重大問題，其呈報流程是什麼？ *你是否已為第三方（例如供應商、消費者、分銷商）或工作人員建立通報流程，以舉報人工智能應用中潛在的漏洞、風險或偏見？在部署人工智能應用後，必須建立持續的運作支援機制，以確保人工智能應用的表現保持穩定、可靠及穩健。*	系統分析員、系統架構師、數據科學家、項目經理	
38 - 你如何監察人工智能模型的持續表現？你已制定什麼業務持續運作方案？觸發人工智能模型維護及回退的條件是什麼？ *你是否已建立核實與驗證方法及文件記錄（例如日誌記錄），以評估並確保人工智能應用在不同方面的可靠性與可再現性？你是否已制定措施以確保人工智能應用在整個生命週期中保持可追溯性？你是否已制定措施以持續評估人工智能應用的輸入數據品質？你是否已設定經過測試的故障安全備援計劃，以處理不同來源的人工智能應用錯誤，並制定治理程序來啟動這些計劃？*	第4.1.5.4節： • 確保端到端工作流程具可重複性與可再現性 • 實現可追溯性 第4.1.6.1節： • 設定效能指標 • 納入品質評估 數據科學家、項目經理	

B. 人工智能治理流程—合規	實務指引參考章節 / 責任	評估結果 / 建議緩解方案
39 - 就着使用已識別的人工智能應用而言，在部署後，持續識別、審查及緩解風險的流程是什麼？	第4.1.6.3節： • 進行管理 / 持續審查	

B. 人工智能治理流程—合規	實務指引參考章節 / 責任	評估結果 / 建議緩解方案
你的機構是否在適當情況下（例如客戶行為、業務目標、風險及企業價值有變時）主動監察、審查及定期調整人工智能模型？此做法可緩解與「人工智能道德原則」（例如公平、可靠、穩健及安全）相關的風險，因為模型在不斷變化的環境中運作。	項目經理、資訊科技規劃人員 / 管理層	
40 - 人工智能應用的所有關鍵決策點是否已明確列出，並符合所有相關法例、內部政策或程序？ *可使用此評估及第4.1.6.3節所述的審查來核實。*	第4.1.6.3節： • 諮詢資訊科技委員會 / 首席資訊官 項目經理、資訊科技規劃人員 / 管理層	

C. 影響—正面影響	指引	評估結果 / 建議緩解方案
41 - 此人工智能應用對機構有何益處？	*確定並說明機構可實現的效益。可考慮以下因素：收入增長、成本降低、效率提升、員工滿意度、投入度與生產力提高、加強與員工的關係、提升或維護品牌或聲譽、合規保障、詐騙防範、加強或維護網絡或實體安全、新增或改善服務、行銷方式改進、評估客戶偏好的能力更佳、創新能力提升或更快速高效實現創*	

C. 影響—正面影響	指引	評估結果 / 建議緩解方案
	新、研究流程改善、強化進行研究與招募研究對象的能力、提升研究效率，以創新的方法進行研究。	
42 - 對已界定的受影響持份者有何益處？該人工智能應用是否可能以某種方式使用，導致特定持份者或持份者群體比其他群體獲得優待？	*確定並說明此技術應用 / 數據活動預期為各持份者帶來的正面影響。對於人工智能應用的每項有益用途，是否存在可識別的個人或群體期望？判斷差別待遇（如有）的潛在正面目標為何。*	
43 - 對整體社會有何益處？	*此人工智能應用的部署符合哪些社會利益？該項目 / 應用如何促進或提升福祉？該項目 / 應用將如何促進人類價值觀？*	
44 - 哪些因素可能會限制外部持份者實現任何效益？	-	

C. 影響—對特定持份者的負面影響	實務指引參考	評估結果 / 建議緩解方案
45 - 該人工智能應用是否可能基於以下任何理由（包括但不限於）對人造成歧視：性別、種族、膚色、民族或社會出身、遺傳特徵、語言、宗教或信仰、政	-	

C. 影響—對特定持份者的負面影響	實務指引參考	評估結果 / 建議緩解方案
治或其他觀點、少數民族身份、財產、出生、殘疾、年齡或性取向？		
46 - 是否可以預見該數據分析結果的潛在應用或數據活動可能令人感到意外、不當或具歧視性，或可能被視為冒犯而造成困擾或羞辱？	*該數據或技術是否可能以某種方式使用，導致某群體與其他群體相比受到差別對待？*	
47 - 該人工智能應用是否可能對環境造成負面影響？該人工智能應用是否可能對整體社會或民主制度產生負面影響？該人工智能應用是否可能對使用者的身心健康、生命及財產帶來潛在的負面影響？	*是否已評估該人工智能應用的使用對（終端）使用者及相關人士以外的社會影響，例如可能間接受影響的持份者或整體社會？*	
48 - 對於涉及第三方的數據或技術活動（例如在此活動中接收或採購技術或數據），存在哪些相關風險？	*第三方的例子包括銷售資料區塊的數據經紀商、數據聚合商、儲存及運算工具供應商、數據信託機構等。風險的例子包括數據準確性、數據保護、下游使用監控、合法數據收集（當透過第三方進行時）、數據可用性等。*	
49 - 從法律和業務角度來看，該人工智能應用是否可能導致任何潛在成本？	*例如訴訟可能導致額外法律費用。該人工智能應用是否可能導致此類開支？*	

C. 影響—控制措施	實務指引參考章節 / 責任	評估結果 / 建議緩解方案
50 - 為預防及緩解可能發生的風險，現正實施哪些額外的技術及 / 或程序保障措施（緩解控制措施）？ *是否已建立適當的治理及問責措施與流程？數據的準確性及 / 或品質是否適合該數據活動？數據的相對準確性會否對個人 / 群體造成影響？*	第4.1.5.2節： • 建立多層緩解機制 數據科學家、項目經理	
51 - 請說明用於對外解釋如何使用技術與數據，以及如何考量及 / 或處理與之相關的個人利益與風險的機制。 *確定透明度與個人問責機制為何，以及這些機制是否適合該資訊活動之用。該技術或資訊的應用是否會進行使用者不知情或可能感到意外的操作？建議採用哪些可解釋性機制？*	第4.1.5.2節： • 向用戶提供資訊 數據科學家、項目經理	
52 - 請說明人為參與有關流程的程度。是否已考量所有與自動化決策相關的因素？ *相關人員（人在環內、人在環外、人為管控）是否已接受關於如何進行監督的特定培訓？*	第4.1.4.2節： • 決定適當的人為介入程度 資訊科技規劃人員 / 管理層、業務使用者、項目經理	
53 - 該人工智能應用設有什麼機制收集使用者反饋？是否會為希望質疑決策結果的人工智能應用使用者制定或建立申訴程序？ *該人工智能應用是否可受益於外界（例如人工智能道德委員會）的額外審查與意見？*	第4.1.5.3節： • 實施易用的反饋用戶介面 數據科學家、項目經理	

C. 影響—控制措施	實務指引參考章節 / 責任	評估結果 / 建議緩解方案
54 - 現有什麼計劃與程序評估人工智能模型的長期表現，包括模型漂移及模型的使用環境改變，以確保輸出結果保持統計準確性？	第4.1.6.1節： • 設定效能指標 • 納入品質評估 • 重新訓練人工智能模型 數據科學家、項目經理、資訊科技規劃人員 / 管理層	

D. 決策—繼續 / 中止	指引	評估結果 / 建議緩解方案
55 - 促進人工智能應用可審計性的機制（例如開發過程的可追溯性、訓練數據的來源記錄，以及人工智能應用的處理過程、結果、正面與負面影響的日誌記錄）成效如何？	-	
56 - 使用人工智能時還有哪些額外監管要求？是否存在與數據（或技術）收集、分析及使用相關的其他法律、跨境要求、政策、合約、行業或其他義務？這些是否均已妥善處理？	-	
57 - 整體控制措施與保障機制在降低風險方面的成效如何？請使用項目管理計劃範本中的風險評估工具，根據本評估C部分所記錄的影響來評估風險。	-	
58 - 決策—是否已在效益與風險緩解之間取得適當平衡，以支援人工智能應用處理活動，並符合「人工智能道德原則」？	*已證明可描述且可實現的淨正面效益成果（對民眾的實質益處），且負面影響已緩解至令人滿意且可*	

D. 決策—繼續 / 中止	指引	評估結果 / 建議緩解方案
是否還有其他應予考量的因素？是否已妥為顧及持份者的利益、期望與權利，以及應額外考量哪些基於具體情境的個人參與及選擇因素？ **該項目或人工智能應用是否需要呈報至高級決策機構（例如資訊科技委員會 / 首席資訊官）？**	*證明的程度。所提議的技術與數據使用方式符合所有「人工智能道德原則」，並體現多元性、包容性及私隱權為基本人權的價值觀。*	

附錄D — 人工智能策略範本

以下範本僅供說明用途，機構可採用此範本制訂人工智能策略 / 計劃，亦可按需要採用其他範本或策略格式。

A. 摘要

願景與目標	*闡述人工智能應用的願景與目標。人工智能應用 / 使用案例的目標是什麼，以及它如何為機構帶來益處？可通過解釋人工智能項目的預期成果和實現這些成果的方式，加以說明。*
人工智能應用目標與機構目標的對應	*將人工智能應用目標與機構目標進行對照。兩者必須保持一致，以確保人工智能成果符合機構的願景與使命。若發現人工智能應用目標與機構目標存在差距，機構應評估所提議的人工智能應用是否合適。*
現況概述	*概述當前狀況，包括過去一年（或數年）的主要成果、正在進行的現有工作，以及任何已識別而須解決的人工智能應用相關問題。*
關鍵驅動因素與目標	*總結推動變更的關鍵因素，例如最新政策目標、業務策略、運作需求，以及其他已識別的人工智能相關改進機會。*
建議的人工智能應用 / 使用案例	*列出所有人工智能使用案例，並按重要性對人工智能項目進行排序。項目重要性應基於當前對機構的關鍵程度。同時考慮是否存在任何短期項目，能以快速、低成本且投入較少資源的方法，證明人工智能的價值。*
實施計劃	*概述建議項目清單，列明各項目的具體目標、實現時間表、預期成果及相關資源需求。*

B. 現況（僅適用於現時已有人工智能項目的機構）

現有的道德與法律考量	*闡明現有人工智能應用涉及的道德考量及法律影響。例如，數據私隱通常是大多數人工智能應用的關鍵考量之一。*
現有的技術與基礎設施	*說明現有人工智能應用的開發與部署所使用的技術及基礎設施。該等技術及基礎設施應包含以下數據層面：* • *數據收集* • *數據儲存*

	• *數據建模* • *數據處理／分析*
現有的技能與人才	*列出現有人工智能應用的開發與實施所取用的技能與人才。*
現有的挑戰	*指出現有人工智能項目生命週期中面對的任何問題或挑戰。*
現有的變更管理	*說明現有人工智能應用帶來的變更，例如對員工的影響、參與度及溝通方式。*

C. 關鍵驅動因素與目標

道德與法律考量目標	*闡明人工智能應用涉及的道德考量及法律影響。例如，數據私隱通常是大多數人工智能應用的關鍵考量之一。機構應參考全部12項「人工智能道德原則」，以確保擬議的人工智能應用符合道德標準。*
技術與基礎設施目標	*列明開發與部署人工智能應用所需的技術及基礎設施。合適的技術與基礎設施應考慮以下數據層面：* • *數據收集* • *數據儲存* • *數據建模* • *數據處理／分析*
技能與人才目標	*說明開發與實施人工智能應用所需的技能與人才。若發現有技能缺口，機構應決定是否通過培訓提升員工技能、招聘新員工，或與外部人工智能供應商合作。*
已識別的挑戰	*指出人工智能項目生命週期中可預見的問題或挑戰。機構可採取哪些行動以確保項目成功交付？*
變更管理目標	*規劃變更管理，並說明擬議人工智能應用帶來的變更，例如對員工的影響、參與度及溝通方式。例如，人工智能應用可能對職位流失產生影響，尤其是涉及特定任務或流程自動化的情況。*

D. 建議的人工智能應用

需求規格	*指明業務、數據、應用及技術（包括人工智能模型，如適用）的需求，以及總體組成部分（例如資訊安全與數據私隱），作為後續開發與實施的充分必要條件。*
解決方案選項與適用性	*根據對應的選擇準則（例如成本效益分析），列出實施的策略選項。*
推薦的人工智能應用	*描述為實現業務目標及達成預期成果而推薦實施的項目。*

E. 實施計劃

實施策略	*界定實施的策略方針，包括項目間的相互依賴性、工作的優先次序，以及所有項目將採取的策略措施（基於「人工智能道德原則」及人工智能生命週期實踐等制定）。*
高層次路線圖	*列明主要活動、里程碑及預期交付成果，並附上實現目標成果的時間表。*
資源估算	*提供實施各推薦項目所需資源（包括人員及開支）的估算。*
效益與影響	*說明實施有關項目帶來的無形及有形效益、預期影響及相關緩解措施。*
治理	*界定治理架構與流程，以便監察進度及解決實施階段可能出現的問題。*

建議質素準則

以下質素準則旨在協助機構評估其人工智能策略 / 計劃的質素。這些準則僅供說明用途，機構可根據適用的策略範本調整內容。

A. **摘要** • 摘要是否包含足夠且必要的資訊，以供機構高層管理人員評估？ • 所提供的資訊是否準確，並與報告其他部分保持一致？
B. **現況（僅適用於現時已有人工智能項目的機構）** • **現有的道德與法律考量：**現有人工智能應用涉及的所有道德考量及法律影響清單是否全面且準確？ • **現有的技術與基礎設施：**所有現有技術（用以支持當前人工智能應用）清單是否全面且準確？ • **現有的技能與人才：**現有人工智能項目獲得的所有技能與人才清單是否全面且準確？ • **現有的挑戰：**人工智能項目生命週期中識別的所有問題或挑戰清單是否全面且準確，並提出適當解決方案？ • **現有的變更管理：**人工智能應用帶來的所有變化清單是否全面且準確，並附有適當的變更管理計劃？
C. **關鍵驅動因素與目標** • **道德與法律考量目標：**相關道德考量及法律影響是否符合「人工智能道德原則」並獲社會普遍接受？ • **技術與基礎設施目標：**技術及基礎設施目標是否符合「人工智能道德原則」，並能支持目標人工智能應用？ • **技能與人才目標：**人才與技能目標是否能以最小冗餘提供所需的技能組合，以支持目標人工智能應用開發？ • **已識別的挑戰：**已識別的問題或挑戰是否包含建議的解決措施？ • **變更管理目標：**已識別的變更是否包含建議的變更管理計劃？ • 在道德與法律、技術與基礎設施、技能與人才方面，是否已識別當前情況與目標之間的所有差距？

D. **建議的人工智能應用** • 為填補差距並過渡至目標狀態，需求規格是否已明確界定全部所需的能力？是否已識別並明確規定人工智能應用及互操作性的需求？ • 已識別的解決方案選項是否完全符合所有需求規格？ • 各方案的選擇準則是否符合「人工智能道德原則」，並獲關鍵持份者認可？ • 獲選的解決方案是否按邏輯歸類至推薦項目中？ • 所有推薦項目的業務價值是否妥善分配，並與關鍵持份者達成共識？ • 推薦項目的優先次序是否獲關鍵持份者同意？
E. **實施計劃** • 實施計劃是否涵蓋所有推薦項目？ • 高層路線圖是否清晰顯示從當前人工智能項目過渡至目標項目的進程，當中列明時間表、明確里程碑及預期的主要交付成果？ • 已識別的風險是否獲妥善緩解並獲關鍵持份者接受？ • 是否明確界定實施項目的管理架構及治理機制，並獲關鍵持份者同意？