著作權等聲明公告

本資料所含內容與相關附屬文件均為經濟部產業發展署（下稱本署）及所屬人員職務上所完成之著作，本署依法擁有著作權及其他相關智慧財產權，而受著作權法及相關法規保護。業經申請並經本署授權同意使用之個人、法人等，於使用時敬請註明出處，並僅限非商業用途之使用。

謹提醒，倘未取得同意或授權，而逕自重製、改作、公開傳輸或有任何侵害本署著作權之行為者，本署將視違法情節逕行依相關法律追訴。另提醒，如有違法情事，則依不同情節，除行為人個人應負賠償責任以外其所屬單位、法人亦可能應負連帶責任。

經濟部產業發展署
產業AI三日班公版教材

單元三 負責任的AI應用

經濟部產業發展署
Industrial Development Administration
Ministry of Economic Affairs

目錄

1. 課程目標與先備知識

- 在課程之前，建議宜具備的知識與經驗
  - ➢AI 的基本概念
  - ➢AI 倫理和社會影響

- 教學目標：
  - ➢學習目標和知識水平的 AI 資源
  - ➢了解實施AI所需的技能
  - ➢滿足個人和組織在AI學習旅程中的不同階段需求

什麼是負責任 AI？

負責任人工智慧 (負責任 AI) 是一種以安全、可靠且道德的方式開發、評估及部署 AI 系統的方法。

負責任 AI 還要求在經濟領域，確保 AI 系統能夠促進公平競爭，避免壟斷和市場失衡。

在環境方面，AI 技術的開發和部署應盡量減少資源消耗和環境污染，促進可持續發展。

圖片來源：Copilot生成圖片

負責任 AI舉例一

在經濟領域，當大公司利用 AI 技術來提高運營效率時，他們應該確保這些技術不會導致競爭對手被不公平地排擠出市場，保持市場的健康競爭。在環境領域，AI 技術可以用來優化供應鏈管理，減少碳排放和資源浪費。

例如，利用 AI 來改進物流路線，降低運輸過程中的能源消耗，從而達到環保的目的。這些措施不僅有助於企業自身的可持續發展，也對整個社會和環境帶來積極的影響。

什麼是負責任 AI？

負責任 AI 還強調透明性，即 AI 系統的決策過程應該是可解釋的，讓用戶和監管機構能夠理解系統的運作原理。

公平性也是負責任 AI 的一個重要方面，這意味著 AI 系統應該為所有用戶提供平等的服務，不應有任何形式的偏見或歧視。

圖片來源：Copilot生成圖片

什麼是負責任 AI？

所有這些要求的核心是道德標準，即在開發和使用 AI 技術時，應始終以人類的福祉和社會的整體利益為重。

負責任人工智慧的實踐，不僅需要技術層面的創新和改進，還需要法律、政策和倫理方面的支持。通過綜合這些方面的努力，我們才能確保 AI 技術的發展能夠真正造福社會，促進人類進步。

負責任 AI 的決策過程

- AI 系統是由其開發者和部署者所做出之許多決策形成的產物。
- 從系統用途到人們與 AI 系統互動的方式，負責任 AI 都可協助主動引導，讓這些決策產生更大的效益和公平的結果。
- 負責任 AI 的原則在系統用途和人們與 AI 系統互動的方式上發揮著重要作用。這些原則要求開發者將用戶及其目標放在系統設計決策的核心，通過滿足用戶需求、確保公平和保護隱私，來實現更大的效益和更公平的結果。這樣的設計方法不僅能夠提升 AI 系統的實用性和可靠性，還能夠增強用戶對 AI 技術的信任，促進技術的健康發展。

負責任 AI 的價值觀

負責任 AI 必須尊重公平性、可靠性和透明度等恆久價值。

負責任 AI 的核心原則

負責任 AI 的核心原則是指在開發、評估和部署人工智慧系統時，遵循一系列的道德和操作準則，確保 AI 技術以安全、可靠和公平的方式運行，並且對社會、經濟和環境帶來正面影響。

這些原則旨在防止AI系統引發社會不公、歧視、隱私侵害和其他潛在的負面影響。

國際對於負責任 AI的處理原則

以 Microsoft 為例，Microsoft 開發負責任 AI 標準。根據下列六大準則建置 AI 系統的架構：

瞭解負責任 AI

公平性

可靠性與安全性

隱私與安全

包容性

透明度

問責性

資料來源：微軟官網
https://www.microsoft.com/ai/responsible-ai

公平性的概述

AI 系統應公平地對待每個人，避免以不同的方式影響處境相似的群體。

AI系統應當公正無偏地對待每一個人。
比如，如果您為銀行構建了一個機器學習模型來輔助貸款審批，該模型應當能夠在不帶有性別、種族或其他偏見的情況下預測貸款是否應當被批准或拒絕。
這樣的偏見可能會導致某些申請者群體不公平地獲得優勢或處於不利。

圖片來源：Copilot生成圖片

機器學習中的公平性和包容性

以Azure ML為例，負責任 AI 儀表板的公平性評量元件可讓資料科學家和開發人員評估模型在以性別、種族、年齡及其他特性所定義的敏感性族群上的公平性。

機器學習公平性概述

人工智慧和機器學習系統可能會顯示不公平的行為。 定義不公平行為的方式之一，是藉由其對人們的損害或影響來定義。

AI 造成的損害類型

AI 造成的損害有兩種常見的類型：

- 配置的損害
- 服務品質的損害

配置的損害：

配置的損害範例包括雇用、入學申請和貸款；在其模型中，從特定族群中挑選適當候選人的能力，可能會優於對其他族群的挑選。

配置損害案例

在職位招聘過程中，如果AI系統偏向於某一特定群體，則其他群體可能會因為這種偏見而失去應有的機會。這樣的配置損害可能會加劇社會不平等，影響社會的整體公平性。

因此，在設計和部署AI系統時，必須謹慎考慮這些潛在的配置損害，以確保所有群體都能公平地獲得資源和機會。

AI 造成的損害類型

服務品質的損害：

- 服務品質的損害是指 AI 系統在不同族群中的運作效能存在差異，導致某些群體獲得的服務品質不如其他群體。

圖片來源：Copilot生成圖片

例如，在語音識別系統中，如果 AI 系統對某些口音或語言的識別準確性較低，則這些用戶的體驗就會明顯不如其他用戶。這種服務品質的不均衡可能會影響使用者對 AI 系統的信任，甚至導致某些群體被邊緣化。因此，開發 AI 系統時，必須確保其在各個族群中的效能是一致的，從而避免因為技術不公平而對某些群體造成不利影響。

評估和緩解AI系統的不公平行為

若要減少 AI 系統中不公平的行為，您必須評估並緩解這些損害。 負責任 AI 儀表板的模型概觀元件會參與模型生命週期的識別階段。

群組公平性

在負責任 AI 儀表板的這個元件中，公平性是透過稱為群組公平性的方法來概念化。

公平性的差異計量

在評量階段，公平性會透過差異計量進行量化。 這些計量能夠以比率或差異的形式，評估和比較模型在群體間的行為。負責任 AI 儀表板支援兩種差異計量類別：

- 模型效能的差異
- 選取率的差異

模型效能的差異

模型效能的差異計量集會計算選定的效能計量在資料子群體間的差異值。以下是一些範例：

精確率的差異

精確率的差異是指在不同子群體之間，模型預測正確的比例存在差異。比如，在製造業中，瑕疵檢測模型在不同班次的精確率差異。

精確率的案例

在製造業中，某檢測模型在不同的工作班次中，早班檢測產品瑕疵的精確率為90%，而晚班的精確率只有80%。

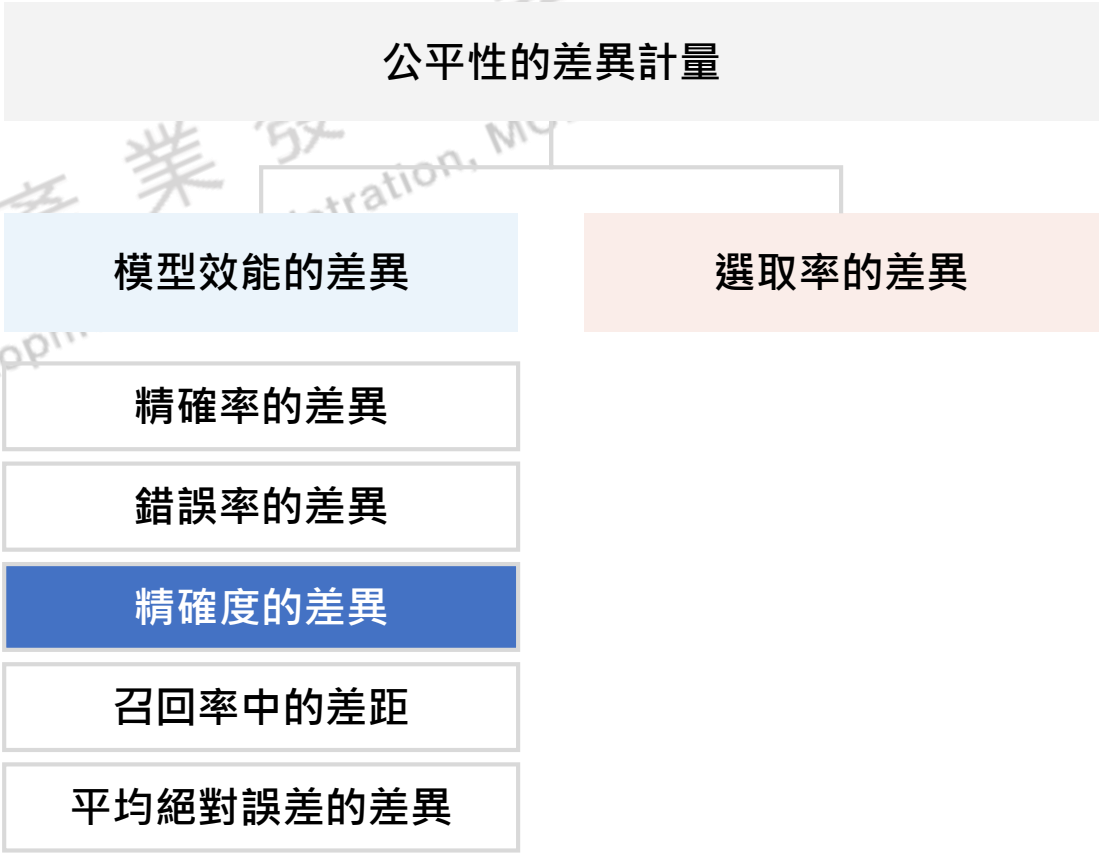
這種差異表明模型在不同班次的預測精度不一致，可能由於生產環境或操作人員的差異所致。

錯誤率的差異

錯誤率的差異是指在不同子群體之間，模型預測錯誤的比例存在差異。例如，模型在不同頻率的馬達中性能存在的差異。

精確度的差異

精確度的差異是指模型在不同子群體中，預測值與實際值之間的差距。例如，在PCB板製造業中，一個檢測模型在不同班次中的精確度差異。

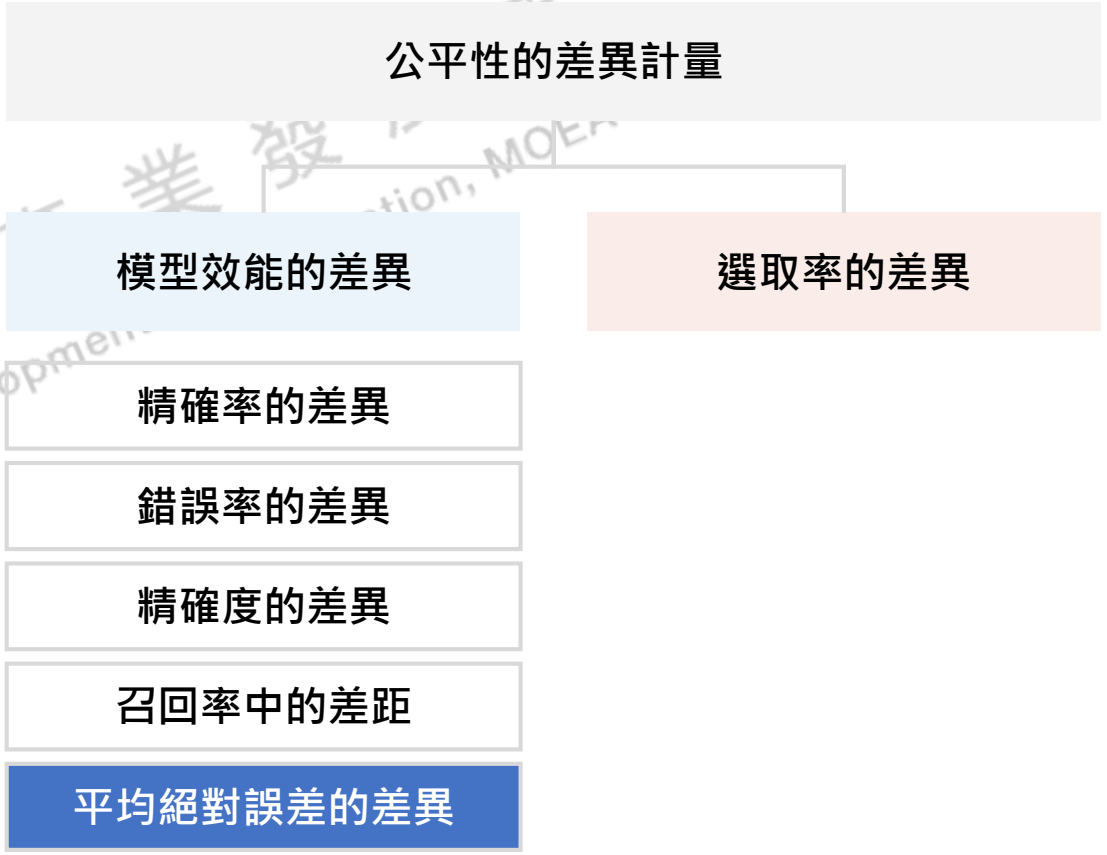


召回率中的差距

召回率的差距是指模型在不同子群體中，能夠正確預測正類樣本的比例存在差異

平均絕對誤差的差異 (MAE)

平均絕對誤差的差異是指模型在不同子群體中，預測值與實際值之間的絕對誤差存在差異。



平均絕對誤差的差異 (MAE)案例

假設你正在使用一個APP來預測你的減肥進度。APP會根據你的飲食和運動情況預測每週的體重減少情況。

圖片來源：Copilot生成圖片

情境一：正常飲食和運動的情況下，APP的預測比較準確。

情境二：節假日飲食和運動在節假日期間，由於飲食和運動習慣的變化，APP的預測誤差較大。

※分析：在不同情境下，模型的預測誤差不同

模型效能的差異對公平性的影響

在機器學習和數據科學中，模型效能的差異對公平性有著重要影響。模型效能指的是模型在不同預測任務中的準確性和可靠性。

當模型在不同子群體（例如年齡、性別、種族等）中表現不一致時，這會導致公平性問題。

圖片來源：Copilot生成圖片

選取率的差異概述

選取率的差異是指在不同子群體間，模型對理想預測（即歸類為1的資料點）的比例存在差異。

選取率的差異重要性

選取率的差異幫助識別模型在不同群體中的預測偏差，確保模型對所有群體都能公平對待，避免因偏見而產生的不公平結果。

選取率的差異案例說明

在不同的生產批次中，生產批次A的化學品檢測系統選取率為90%，而生產批次B的選取率僅為75%。這表明模型在不同批次中的選取效果存在差異。這可能是由於生產過程、原材料品質或操作流程的差異導致的。

公平性

可靠性和安全性

人工智慧系統應當安全可靠地運行。

- 一個基於人工智慧的軟體系統被用於自動駕駛汽車
- 一個機器學習模型用於診斷病患症狀並提出治療建議

可靠性和安全性的概述

若要建立信任，AI 系統必須以可靠、安全且一致的方式運作。 這些系統應該要能夠依照其原先的設計運作、安全地回應非預期的狀況，以及反抗有害的操作。

圖片來源：科技新報
https://technews.tw/2020/03/27/wevolver-2020-autonomous-vehicle-technology-report/

機器學習中的可靠性與安全性

以Azure ML為例，負責任 AI 儀表板的錯誤分析元件可讓資料科學家和開發人員：

- **深入了解模型的失敗分佈情況。**
- **識別比整體基準具有更高錯誤率的資料世代 (子集)。**

可靠性與安全性與錯誤分析之間的關聯性

錯誤分析是確保AI系統可靠性與安全性的關鍵步驟：

- 通過識別和分析模型的錯誤，我們能夠找到系統的弱點和漏洞，從而進行針對性的改進。
- 高錯誤率會直接影響系統的可靠性和安全性，因為未檢測出的錯誤可能在實際應用中引發嚴重問題。
- 通過持續的錯誤分析，我們能夠提高模型的準確性和穩定性。

評估機器學習模型中的錯誤

當前模型偵錯實務的一個主要挑戰是使用彙總計量對基準測試資料集中的模型進行評估。

模型的精確度可能會在不同的資料子群組之間有所差異，而且在某些特定特徵或屬性的資料子群組中，模型更容易出現錯誤。

評估機器學習模型中的錯誤案例

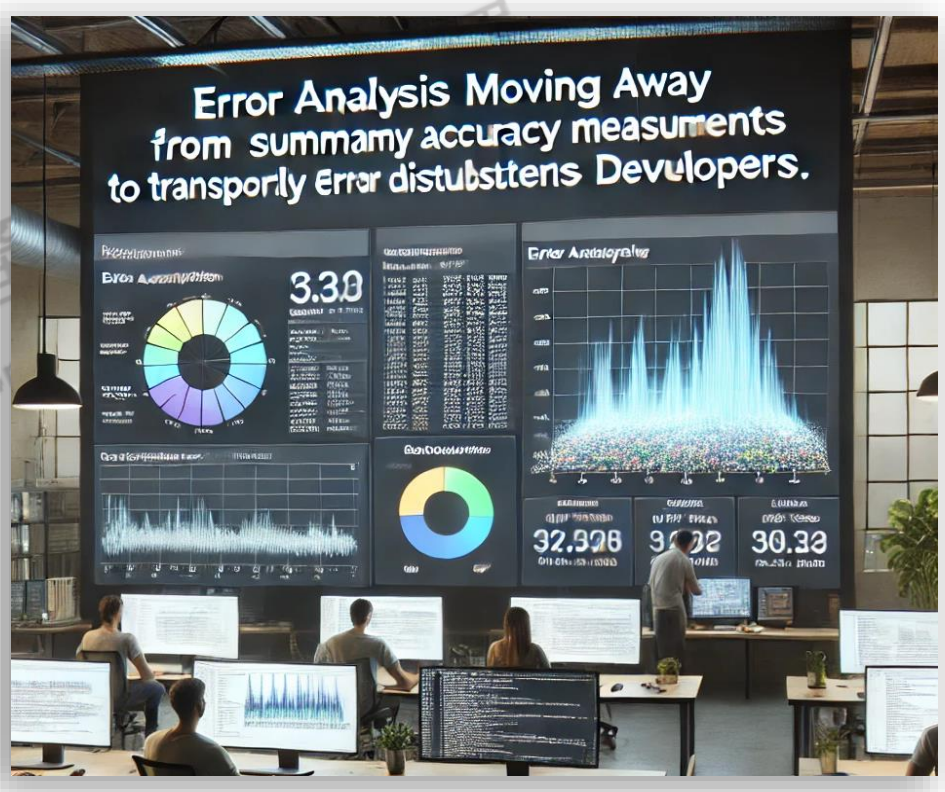
一間電子製造公司，使用機器學習模型來檢測生產線上的產品瑕疵。雖然模型在整體上的精確度很高，但在某些特定類型的產品上，錯誤率卻明顯偏高。

模型錯誤的影響

這些失敗的直接結果就是完全缺乏可靠性和安全性、衍生公平性問題，以及完全失去人們對機器學習的信任。

錯誤分析的重要性

錯誤分析會從彙總精確度計量移出，以透明方式向開發人員公開錯誤分佈，並讓他們有效率地識別和診斷錯誤。



圖片來源：Copilot生成圖片

可靠性和安全性

隱私權和保密性

人工智慧系統應當是安全的，並且要尊重用戶隱私。

隱私權與保密性概述

隨著 AI 普及，保護隱私權及個人和公司資訊變得更重要和複雜。使用 AI 時需注意隱私和資料安全，因為 AI 系統需要存取資料來進行精確的預測和決策。

圖片來源：ChatGPT生成圖片

隱私權法律要求

AI 系統必須遵守下列隱私權法律：

- **要求資料收集、使用及儲存相關透明度。**
- **要求取用者具有適當控制權，以選擇其資料的使用方式。**

機器學習中的隱私權與保密性

以Azure Machine Learning為例，可讓系統管理員和開發人員建立符合公司原則的安全設定。透過 Azure Machine Learning 和 Azure 平台，使用者可以：

- 依使用者帳戶或群組限制資源和作業的存取權
- 限制連入和連出的網路通訊
- 加密傳輸和待用的資料
- 掃描弱點
- 套用和稽核設定原則

隱私權和保密性

包容性

人工智慧系統應當賦能並包容所有人，確保每個人都能參與其中。

包容性概述

包容性規定 AI 應該考慮所有人類和體驗。AI技術的普及和應用必須體現包容性和普惠性，以確保所有人，無論他們的身體條件、性別、性取向、種族或其他社會身份，都能從中受益。為了實現這一目標，AI的設計和部署需要考慮到多樣性和平等原則，避免偏見和歧視，並主動促進機會均等。

包容性例子

在電子設備製造業中，AI 系統可協助工人進行品質檢查。若設計時只考慮視力正常的工人，視力有障礙的工人使用起來會非常困難，這樣就不具包容性。

為解決此問題，AI 系統應加入語音提示功能，幫助視力障礙的工人也能有效進行品質檢查，從而提高工作效率，確保公平待遇和工作安全。

圖片來源：Copilot生成圖片

技術支持包容性

組織應該使用語音轉換文字、文字到語音轉換

和視覺辨識技術，讓聽力、視覺和其他障礙的人能夠使用。

包容性

透明度

人工智慧系統應當是可解釋的，使用者應該能夠充分理解系統的用途、工作原理以及可能存在的局限性。

透明度概述

當 AI 系統協助通知那些對人們生活有極大影響的決策時，人們一定要了解這些決策產生的方式。

例如，銀行可能會使用 AI 系統來決定某人是否信用可靠。 公司可能會使用 AI 系統來決定所要雇用的最合格候選人。

透明度與可解釋性

透明度中很重要的一部分是「可解釋性」，對 AI 系統及其元件之行為的實用說明。

改善可解釋性需要利害關係人先理解 AI 系統的運作方式及其原因。 利害關係人接著可以判斷可能的效能問題、公平性問題、排他性行為或非預期的結果。

什麼是利害關係人

利害關係人是指對某個決策或結果有興趣並且會受到其影響的人或群體。

在 AI 系統的背景下，利害關係人包括了：

- **開發人員**
- **使用者**
- **管理者**
- **政府機構**
- **普通公眾**

圖片來源：Copilot生成圖片

什麼是利害關係人

利害關係人是指對某個決策或結果有興趣並且會受到其影響的人或群體。

在 AI 系統的背景下，利害關係人包括了：

利害關係人與透明度的關係

透明度在 AI 系統中至關重要，因為它能夠讓利害關係人理解和信任 AI 的決策過程。以下是透明度對利害關係人的影響：

機器學習中的透明度

以Azure ML為例，負責任 AI 儀表板的模型可解釋性和反事實假設狀況元件，可讓資料科學家和開發人員針對模型的預測產生人類可理解的描述。

透明度與模型可解釋性和反事實假設狀況

■ **兩者間關係：**

透明度在 AI 系統中非常重要，以下是實現透明度的兩大關鍵元件：

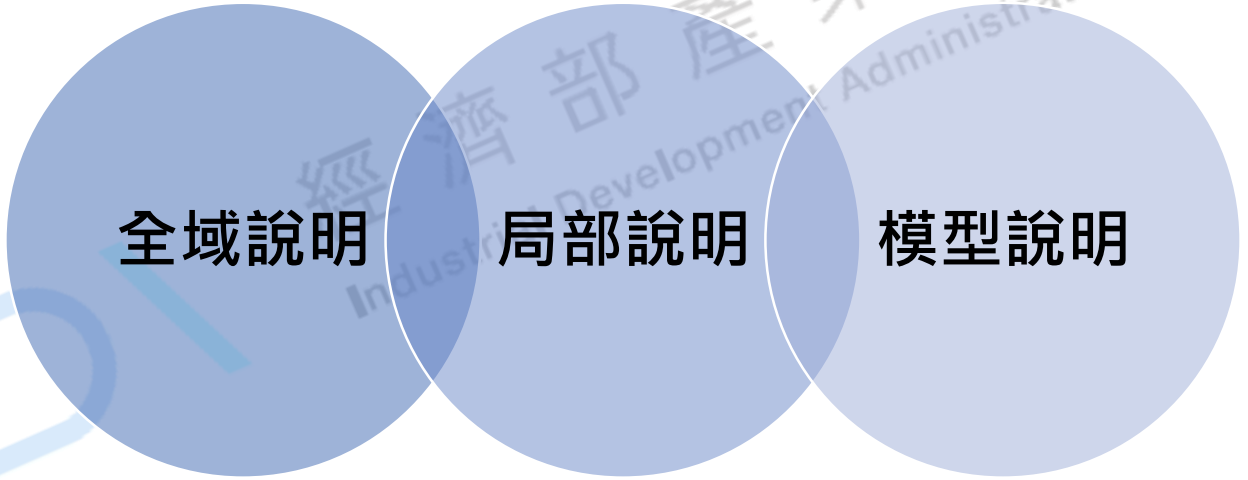
- 模型可解釋性
- 反事實假設狀況

■ **實際案例：**

在金融業，AI 系統可以用於信用評分。模型可解釋性可以幫助貸款申請人理解為何被拒絕，提供具體的評分因素說明。

模型可解釋性元件功能

模型可解釋性元件可以從多個角度來檢視模型的行為：



全域說明

提供對整個模型行為的概覽，幫助我們了解哪些特徵在模型決策過程中起了主要作用，這對於理解模型的整體運行方式非常重要。

局部說明

針對特定預測提供詳細分析，幫助我們理解在某一次預測中，哪些特徵對結果產生了影響，從而可以針對性地進行優化和調整。

所選資料點群組的模型說明

所選資料點群組的模型說明：針對特定群組提供模型行為的分析，這對於識別和消除模型在某些群組中的潛在偏見和誤差非常重要。

反事實假設狀況元件

反事實假設狀況元件可讓您了解機器學習模型對特徵變更和擾動的反應方式，並加以偵錯。

透明度

責任性概述

AI系統的設計和部署人員應對其運作方式負責，並應遵循業界標準制定權責規範。這些規範確保AI系統不會成為影響人們生活的最終決策者，同時保持人類對AI系統的有效控制。

使用人工智慧系統的使用者應負起責任，設計和開發AI解決方案的專業人員需在治理和組織準則框架內工作，確保解決方案符合道德和法律標準。

負責任的AI準則有助於開發者理解構建道德AI解決方案時所面臨的挑戰。

機器學習中的權責

以Azure ML為例， Azure 機器學習中的問責制主要是通過機器學習操作（MLOps）來實現的。MLOps 基於 DevOps 的原則和做法，能夠提高 AI 工作流的效率，同時確保系統的透明度和可追蹤性。

Azure 機器學習提供的 MLOps 功能，

這些功能有助於對 AI 系統負責：

模型註冊、打包和部署

捕捉機器學習生命週期的治理數據，記錄誰發佈了模型、為什麼進行更改、何時部署或使用模型，這些資訊有助於追蹤和責任分配。

治理數據捕獲

可以從任何地方註冊、打包和部署模型，並追蹤使用模型所需的元數據，確保每個步驟都有記錄，方便日後查詢和審計。

事件通知和提醒

在機器學習生命週期中，對實驗完成、模型註冊、模型部署和數據偏移檢測等事件進行通知和提醒，確保相關人員及時獲悉並做出反應。

運營監控

監控應用程式的操作問題以及與機器學習相關的問題，通過比較訓練和推理之間的模型輸入，探索模型指標，並在機器學習基礎設施上提供監控和警報，確保系統穩定運行。

機器學習平台與商務決策

機器學習平台可透過以下兩種方式來協助人員做出商務決策。

- **資料導向的深入解析**

  協助利害關係人單憑歷史資料來了解治療對結果的因果影響。

- **模型導向的深入解析**

  可回答使用者的問題並提供建議，例如「我下一次可以做什麼來從您的 AI 取得不同的結果」。

模型導向的深入解析

模型導向的深入解析，可回答使用者的問題並提供建議，例如「我下一次可以做什麼來從您的 AI 取得不同的結果」。

模型導向的例子

模型導向的見解可以回答使用者的具體問題，例如「我下一次可以做什麼來從您的 AI 取得不同的結果」。

責任

負責任 AI 儀表板

負責任 AI 簡稱(rai)儀表板提供單一介面，可協助您在實務上有效率地進行施作。 並將數個負責任 AI 工具結合在下列領域中：

負責任 AI 儀表板

使用負責任 AI 儀表板的原因：

雖然負責任 AI 的已經有訂定出來，但資料科學家通常需要使用各種工具來全面評估其模型和資料。

工具之間的挑戰：

如果資料科學家發現一個工具的公平性問題，則必須跳到不同的工具，以了解問題的根本資料或模型因素，再採取任何風險降低步驟。

負責任 AI 儀表板

■ **優勢：**

是一個全面但可自定義的工具，可將分散的體驗整合在一起。

■ **資料集世代的建立：**

可以建立資料世代，並將這些世代資料集傳遞至所有支援的元件。

■ **負責任 AI 實務操作概述：**

在實務上實作負責任 AI 需要嚴格的工程。如果沒有適當的工具和基礎結構，這樣的工程可能會非常繁瑣、需要手動執行且曠日費時。

負責任 AI 儀表板全面整合

負責任 AI 儀表板在全方位檢視中納入各種新工具與現有工具。

儀表板會將這些工具與 Azure Machine Learning CLI v2、Azure Machine Learning Python SDK v2 和 Azure Machine Learning 工作室整合。

負責任 AI 儀表板

■ **模型效能與公平性評量概述：**

可幫助評估模型在不同群體中的表現，確保模型結果公平公正。

■ **資料探索概述：**

用於了解並探索資料集分布與統計資料，發現數據中的趨勢和異常。

■ **機器學習可解譯性概述：**

可幫助理解模型的預測過程，提供特徵影響和預測解釋。

負責任 AI 儀表板

■ **錯誤分析概述：**

用於檢視並了解模型錯誤在資料集中的分布，幫助識別和修正問題。

■ **反事實分析與微擾概述：**

用於觀察特徵變化對模型預測的影響，提供最接近資料點的不同預測。

■ **原因推斷概述：**

使用歷史資料來檢視處理特徵對真實世界結果的因果影響。

機器學習中的負責任 AI 儀表板

以Azure ML為例，儀表板提供了對模型的全面評估和調試，使您能夠做出基於數據的明智決策。通過一個統一的界面存取所有這些工具，您可以：

- 識別模型錯誤與公平性問題
- 診斷那些錯誤發生的原因
- 通知您的風險降低步驟
- 評估您的機器學習模型及對其進行偵錯

負責任 AI 儀表板元件

這些工具可以協幫助您檢查和修正機器學習模型中的錯誤，同時協助您做出基於數據和模型的更明智的商業決策。通過整合這些工具，開發者可以全方位提升模型的性能和可靠性。

右圖顯示了這些工具如何在 AI 生命週期中運行，從而改善模型並獲得實在的資料深入解析。

模型偵錯

評估和偵錯機器學習模型對於模型可靠性、可解釋性、公平性和合規性而言非常重要。它有助於判斷 AI 系統的行為方式和原因，並使用此知識來改善模型效能。模型偵錯包含三個階段：

錯誤分析

錯誤分析元件可協助您深入了解模型失敗分布，並快速識別資料中的錯誤群體。

這些工具能幫助資料科學家診斷錯誤的根本原因，並採取相應的糾正措施來改善模型性能。

錯誤分析

Error Analysis 工具包提供多種工具來分析機器學習模型中的錯誤，包括使用以下兩種方式分析：

錯誤群體識別：

- **決策樹**：自動分割數據群體，找到錯誤率高的區域。
- **錯誤熱圖**：可視化顯示錯誤在不同特徵值範圍內的分佈情況。

錯誤診斷：

- **整體預測**：呈現模型在所有數據上的整體行為。
- **局部解釋**：針對特定預測提供詳細分析。
- **假設分析**：通過改變輸入變量來觀察模型行為變化。

資料探索

資料探索是負責任 AI 開發過程中的關鍵步驟。

通過資料探索，開發者可以深入了解：

- **數據的結構**
- **分佈和特徵**

從而發現潛在的數據問題並進行必要的數據處理和清理。這有助於確保模型的準確性和可靠性。

數據結構的重要性

了解數據的結構對開發者來說非常重要，原因如下：

數據品質：

- 確保數據的完整性、一致性和準確性，從而保證數據的高品質

數據處理：

- 有助於有效地清理、轉換和整合數據，提高數據處理效率。

性能優化：

- 幫助設計高效的數據訪問和處理算法，提升系統性能。

問題診斷：

- 快速識別和解決數據相關問題，提升問題診斷的效率。

數據結構的組成

數據結構的組成大致包括：

- **數據類型**：例如整數、浮點數、字符串等。
- **欄位（字段）**：即數據集中的列，每個欄位代表一種類型的信息。
- **記錄**：即數據集中的行，每個記錄包含多個欄位的數據。
- **索引**：用於快速檢索數據的結構。
- **關聯性**：不同數據表之間的關係，如一對一、一對多、多對多等。

數據結構對開發者的影響

數據結構會對開發者進行開發時產生以下影響：

- **數據管理**：良好的數據結構設計有助於有效管理和組織大量數據。
- **數據訪問效率**：適當的數據結構可以提高數據檢索和查詢的速度。
- **模型性能**：了解數據結構有助於設計和訓練更精確的機器學習模型。
- **可擴展性**：能夠更好地設計系統，使其在數據量增大時仍能保持高效運行。

數據結構的案例說明

銷售日期

- 數據清理
- 數據轉換
- 性能優化

產品ID

數據類型：

- 產品ID
- 銷售額
- 顧客評價
- 產品名稱
- 類別
- 價格
- 銷售日期
- 顧客ID等欄位
- 每一筆銷售交易的記錄

圖片來源：Copilot生成圖片

為什麼了解分佈和特徵很重要？

了解分佈和特徵對開發者來說非常重要，因為：

- **數據品質**：能夠識別數據中的異常值和偏差，確保數據的完整性和準確性。
- **模型準確性**：通過了解數據的分佈和特徵，開發者可以選擇合適的模型和算法，提高預測的準確性。
- **數據處理**：有助於進行特徵工程和數據清理，優化數據的表示方式。
- **性能優化**：可以針對特徵和分佈情況調整模型參數，提升模型的性能。

什麼是分佈和特徵?

特徵：是指數據中的屬性或變量，每個特徵代表數據集中的一個方面。

例如，在顧客數據集中，特徵可能包括年齡、性別、收入和購買行為等。特徵是機器學習模型用來進行預測和分類的基礎。

分佈：是指數據在不同範圍或類別中的分佈情況。

例如，一組數據可能在某個範圍內集中，也可能分散在整個範圍內。分佈通常用直方圖、盒鬚圖或散點圖來表示。它可以幫助了解數據的集中趨勢、離散程度和是否存在異常值。

分佈和特徵對開發的影響

分佈和特徵會對開發者進行開發時產生以下影響：

- **數據預處理**：了解分佈可以幫助開發者識別並處理數據中的異常值和缺失值。
- **特徵選擇**：了解哪些特徵對預測結果最有影響，有助於選擇和創建有效的特徵。
- **模型選擇**：根據數據分佈選擇適當的模型和算法，從而提高預測準確性。
- **調參優化**：通過分析特徵和分佈，優化模型參數，提升模型性能和穩定性。

分佈和特徵的案例說明

預測顧客的購買行為

圖片來源：Copilot生成圖片

模型概覽

模型概觀元件會在高階模型預測分布檢視中彙總模型評量計量，以讓您能進一步調查其效能。

此件元件也會啟用公平性評估，並醒目提示不同敏感性群組之間的模型效能明細。

公平性評量

負責任 AI 儀表板中的公平性評估元件提供一套功能，讓資料科學家和開發人員能詳細評估模型在不同性別、種族、年齡等敏感群體中的表現。

公平性評量

- **公平性評估的作用：**

  這些評估元件讓開發者深入了解模型在不同敏感群體中的表現情況，及時發現潛在的偏見和不公正問題。

- **詳細數據分析功能：**

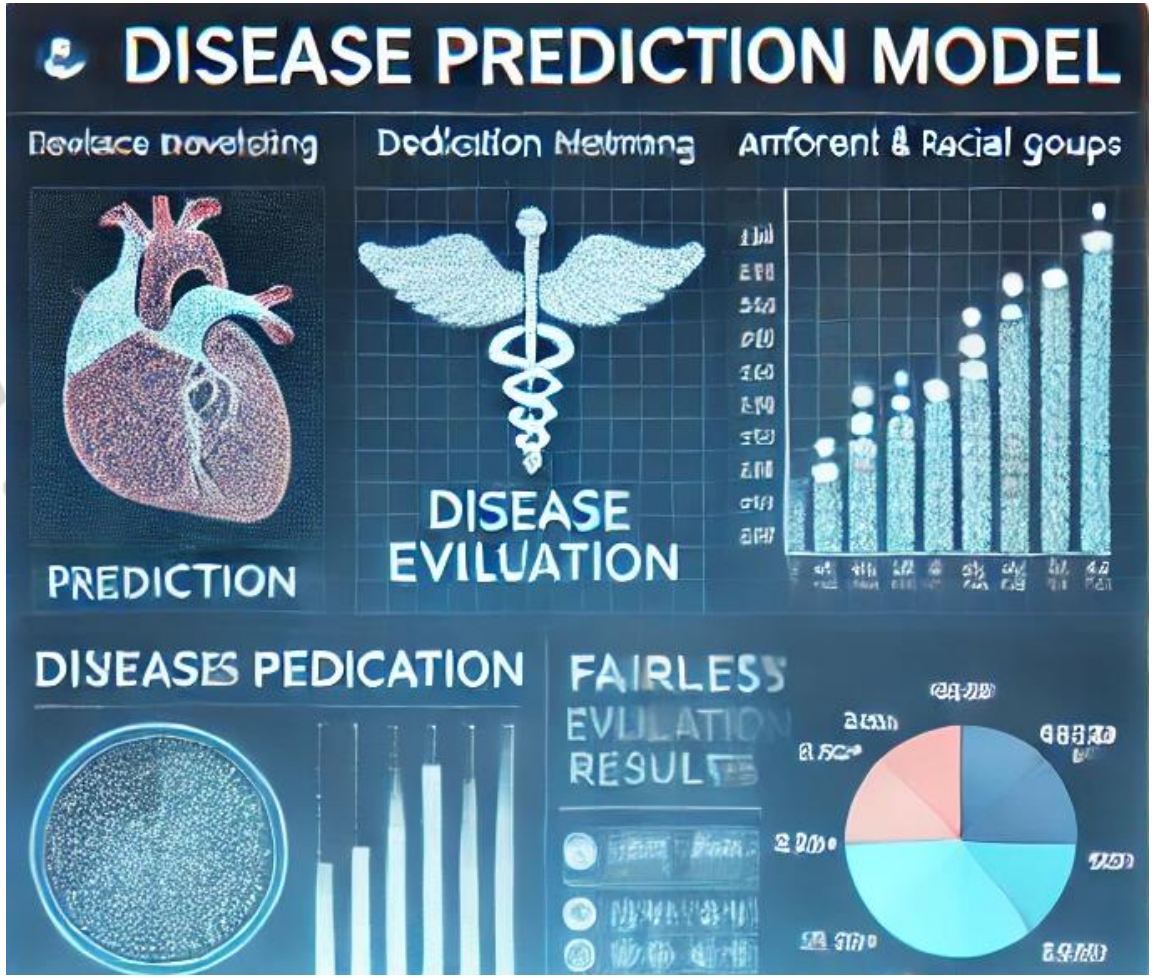
  幫助開發者找出模型偏差原因，針對性地改進模型，提升公平性和準確性。

- **數據收集和處理策略：**

  幫助開發者制定和實施更好的策略，確保模型在不同群體中有效應用。

公平性評估的案例說明

一家醫療公司開發了一個疾病預測模型，公平性評估元件顯示，模型在不同性別和種族群體中的預測結果有顯著差異。



圖片來源：ChatGPT生成圖片

機器學習可解譯性

機器學習可解譯性元件會產生機器學習模型預測的人類可理解說明。這些說明可以針對模型行為提供多種觀點，包括整體預測和個別預測，以幫助了解模型的預測和決策過程。

機器學習可解譯性

InterpretML 是一個開源工具包，專注於解釋機器學習模型，推動負責任的 AI。它提供多種技術來解釋模型行為，支持透明和黑盒模型，包括：

原因推斷

原因推斷是一種技術，旨在確定某一變量對另一變量的因果影響。這在評估干預措施效果時尤為重要。

EconML 使用機器學習技術來估計這些因果效應，允許我們在控制其他變量的情況下，觀察特定處理變量對結果變量的影響。

原因推斷案例

在企業管理中，提升員工工作效率和生產力是一個持續且重要的目標。為此，企業經常設計並實施各種員工培訓計劃，包括技能提升、知識更新和行為改進等方面。

原因推斷可以用來評估這些培訓計劃的實際效果。

方法：

假設我們想要評估一個新的員工培訓計劃是否能提高生產力。我們可以收集員工的背景信息，包括年齡、性別、教育水平、工作經驗和基礎技能水平等，作為控制變量。

原因推斷案例

方法：

假設我們想要評估一個新的員工培訓計劃是否能提高生產力。我們可以收集員工的背景信息，包括年齡、性別、教育水平、工作經驗和基礎技能水平等，作為控制變量。

我們將員工分成兩組，一組參加新的培訓計劃，另一組繼續使用現有的工作方法。通過控制背景信息，我們可以更準確地觀察培訓計劃對生產力的影響。

原因推斷案例

觀察

結果在控制了年齡、性別、教育水平和工作經驗等因素後，我們可以比較兩組員工在培訓前後的生產力變化。如果發現參加新培訓計劃的員工平均生產力提高了15%，而對照組的生產力變化不大，這表明新培訓計劃在提高生產力方面具有顯著效果。

原因推斷案例

分析與改進

根據這些觀察結果，企業可以分析哪些培訓內容和方法對生產力影響最大，並進行相應的調整和改進。例如，如果發現技能提升課程對生產力的提升效果顯著，企業可以增加這類課程的比重。

通過這些改進措施，我們可以確保員工培訓計劃在提高生產力方面的有效性，從而提升企業的整體效率和競爭力。

反事實分析與微擾

反事實分析和微擾元件可觀察特徵微擾如何影響模型預測，並提供具有相反或不同模型預測的最接近資料點。

反事實分析

反事實分析是一種技術，用於生成與當前情況略有不同的假設場景，以探索可能改變結果的特徵。

反事實分析案例說明一

情境：

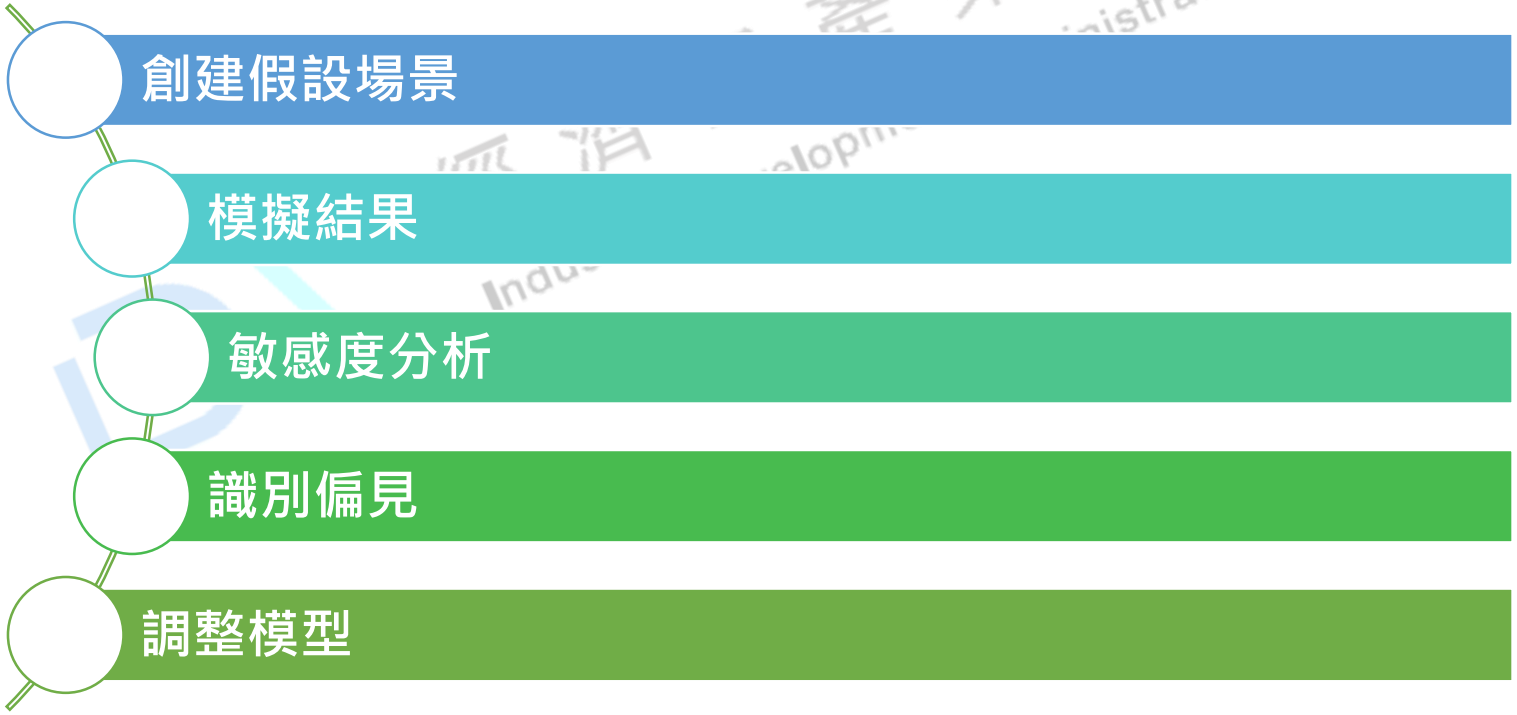
客戶A今天需要一筆資金，在貸款審批的過程中，收到銀行拒絕貸款的通知，貸款機構行員通常會告知客戶貸款貸不下來的原因，是受到多種不同的原因影響，這些原因也就是我們所謂的特徵。這些特徵包括但不限於：

年收入、信用評分、債務收入比、就業狀況、資產狀況、過往信用記錄、年齡、家庭狀況、居住狀況等等

反事實分析案例說明一

方法：

使用反事實分析來探討貸款拒絕決策如何受到上述特徵的影響：



反事實分析案例說明二

反事實分析還能揭示模型中存在的潛在偏見和不足之處。

如果發現模型對某些特徵的微小變化反應過於劇烈，則表明模型在這些特徵上的設置不夠穩定或合理，並需要進一步的調整和改進。

通過反事實分析，我們能夠更加細緻地觀察模型的預測行為，並發現模型在哪些特徵上過於敏感，從而採取措施進行優化。

反事實分析案例說明二

在自動駕駛車輛系統的情境中，反事實分析可以通過以下方法來揭示模型中的潛在偏見和不足之處，並提出相應的改進措施：

反事實分析案例說明二

假設在自動駕駛車輛的行人識別系統中，我們發現模型對光照條件的變化反應過於劇烈。

通過反事實分析，我們將光照條件從白天調整到黃昏，發現模型預測結果從識別到行人變為未識別到行人。這表明模型在光照條件這一特徵上的設置不夠穩定或合理。

反事實分析案例說明二

為了解決這個問題，我們可以進行以下調整：

- **增加訓練數據的多樣性**
- **調整模型的特徵權重**
- **引入輔助特徵**

微擾

■ 微擾概述：

微擾是指對數據進行小幅度的變動，以觀察這些變動對模型預測的影響。

■ 案例情境：

一些餐廳會提供免費的咖啡，假設餐廳提供了4壺咖啡，這4壺咖啡的配方稍有不同。我們可以通過微擾技術，對每壺咖啡的特徵進行小幅度的調整，來找出最受消費者喜愛的口感。

圖片來源：Copilot 生成圖片

微擾案例說明

1. 選擇特徵

在這個例子中，我們可以選擇影響咖啡口感的特徵進行微擾，例如：

圖片來源：ChatGPT 生成圖片

微擾案例說明

2. **引入微擾**：對選定的特徵進行小幅度的變動，以觀察這些變動對咖啡口感的影響。

咖啡壺	研磨粗細	沖泡溫度	水量	咖啡粉量
咖啡壺1	中等	92攝氏度	200毫升	18克
咖啡壺2	稍微細一些	91攝氏度	195毫升	17克
咖啡壺3	稍微粗一些	93攝氏度	205毫升	19克
咖啡壺4	中等	92攝氏度	198毫升	18.5克

微擾案例說明

3. 觀察結果：將4壺咖啡同時提供給消費者，觀察哪一壺咖啡先喝完。這代表哪壺咖啡的口感最受消費者歡迎。

4. 分析與改進：根據消費者的選擇結果，分析最受歡迎的咖啡配方，並進行相應的調整和改進。

- **優化配方：**根據最受歡迎的咖啡壺的配方，進一步優化，調整研磨粗細、沖泡溫度、水量和咖啡粉量的組合。
- **增加數據多樣性：**在未來的測試中，加入更多不同配方的咖啡壺，以進一步精確找出最佳口感。

國內外 AI 應用的準則探討

歐盟《人工智慧法案》

《人工智慧法案》（AI Act）是歐盟制定的一部全面法規，於2024年8月1日正式生效，旨在促進負責任的人工智慧（AI）開發和部署。

這部法案的目的是在保護公民健康、安全和基本權利的同時，減輕企業的管理和財務負擔，並推動AI技術的創新與競爭力。

AI Act的制定反映了歐盟對於確保AI技術在創新和應用過程中，不會對社會造成負面影響的高度重視。

什麼是《人工智慧法案》

《人工智慧法案》通過建立一個全面的法律框架，旨在確保AI技術的安全、透明和負責任的使用。

該法案平衡了技術進步與公共利益保護之間的關係，致力於打造一個可信賴且創新驅動的AI生態系統。透過這一法案，歐盟希望成為全球安全AI領域的領導者，促進醫療保健、安全交通和公共服務等領域的發展，同時提高企業生產力和效率。

資料來源：取自EU AI Act公開資料
https://artificialintelligenceact.eu/ai-act-explorer/

人工智慧法案核心目標

- **確保安全和基本權利**：保護公民的健康、安全和基本權利，特別是針對高風險AI系統。
- **促進創新和競爭力**：在確保安全和合規的前提下，鼓勵和支持AI技術的創新和發展。
- **統一市場**：建立統一的市場規則，避免各成員國之間的法律差異，促進AI產品和服務在歐盟內部的自由流通。

AI Act的核心內容概述

AI Act的核心內容包括以下幾大類：

透明度與問責機制	風險分類與合規要求	市場監管與合規評估
罰則與執法	監測與報告	機構管理

這些核心內容涵蓋了AI技術在開發和使用過程中的各個方面，旨在確保 AI 系統的安全性、透明度和合規性。

透明度與問責機制

- 要求人工智慧系統的透明度，特別是對於高風險系統。
- 用戶有權獲得有關系統運作方式和決策過程的清晰解釋。
- 對生成式AI內容的標記要求，確保用戶知悉並理解AI生成的內容。

風險分類與合規要求

- **最小風險**：如垃圾郵件過濾器和AI視訊遊戲，無需特別義務。
- **具體透明度風險**：如聊天機器人，需告知用戶正在與機器交互。
- **高風險**：如AI醫療軟體和招聘系統，需符合嚴格的合規要求，包括風險管理、高品質數據集、透明度、手動監督等。
- **不可接受風險**：如社會評分系統，因對基本權利構成威脅而被禁止。

市場監管與合規評估

- 各成員國的市場監管機構負責監督AI系統的合規性。
- 需要定期進行合規性評估，確保AI系統符合所有相關法規。
- 市場監管機構可要求修改、撤回或召回不合規的AI系統。

罰則與執法

- 對違反《人工智慧法案》的行為實施罰款和其他執法措施。
- 罰款數額依違規程度和影響範圍而定，最高可達全球年營業額的6%。
- 確保所有違規行為都會受到相應的懲罰，維護法規的嚴肅性。

監測與報告

- 定期監測和報告AI系統的使用情況，確保其符合《人工智慧法案》。
- 要求企業和機構提供定期報告，詳述AI系統的運行情況和合規狀態。
- 透過監測和報告機制，及時發現並處理潛在風險。

機構管理

- 設立人工智慧辦公室，負責協調AI Act的實施和監督。
- 成立歐洲人工智慧委員會，提供技術和政策建議，促進法規一致性。
- 建立諮詢論壇和獨立專家小組，確保利益相關者的意見和科學建議被納入法規實施過程中。

中華民國生成式 AI 參考指引 (草案)

國家科學及技術委員會於2023年8月31日發布的指引旨在引導行政院及所屬機關以負責任及可信賴的態度使用生成式 AI。指引強調了在提高行政效率的同時，保持信息的安全性和隱私性的重要性。

指引的制定參考了各國政府的審慎因應作法，並與AI技術、法律專家及12個相關部會協作，還徵詢公眾意見。最終形成包括總說明和十點規定的草案。

資料來源：取自行政院全球資訊網
https://www.ey.gov.tw/File/CAE5B756153299FD?A=C

核心原則

- **安全性與隱私性**：各機關使用生成式 AI 時，需注重系統環境的安全性，不得洩露機密信息和個人隱私。
- **責任與問責**：生成資訊需由業務承辦人進行最終判斷，確保資訊的客觀性和專業性。
- **資料治理**：生成式 AI 的使用需符合資通安全、個人資料保護和智慧財產權等規定。

主要規定

- **提升行政效率並避免風險：**指引規定各機關在使用生成式 AI 時，應注意避免國家安全、資訊安全、人權、隱私、倫理及法律等風險。
- **資訊判斷：**生成式 AI 產出的資訊需由業務承辦人進行客觀且專業的最終判斷，不得取代業務承辦人的自主思維、創造力及人際互動。
- **機密文書撰寫：**製作機密文書應由業務承辦人親自撰寫，禁止使用生成式 AI。機密文書包括國家機密文書及一般公務機密文書。

主要規定

- **保密資訊提供限制**：業務承辦人不得向生成式 AI 提供涉及公務應保密、個人及未經機關（構）同意公開的資訊，亦不得向生成式 AI 詢問可能涉及機密業務或個人資料的問題。封閉式地端部署的生成式 AI 模型，須確認系統環境安全性後，方得依機密等級分級使用。
- **資訊可信性確認**：各機關不可完全信任生成式 AI 產出的資訊，亦不得以未經確認的產出內容直接作成行政行為或作為公務決策的唯一依據。
- **使用揭露**：各機關使用生成式 AI 作為執行業務或提供服務輔助工具時，應適當揭露。

主要規定

- **法規遵守**：使用生成式 AI 應遵守資通安全、個人資料保護、著作權與相關資訊使用規定，並注意其侵害智慧財產權與人格權的可能性。各機關得依使用生成式 AI 的設備及業務性質，訂定使用生成式 AI 的規範或內控管理措施。
- **採購要求**：各機關應就所辦採購事項，要求得標的法人、團體或個人注意本參考指引，並遵守各該機關依前點所訂定的規範或內控管理措施。
- **公共機構應用**：公營事業機構、公立學校、行政法人及政府捐助的財團法人使用生成式 AI，得準用本參考指引。
- **其他機關參考**：行政院及所屬機關（構）以外的機關得參照本參考指引，訂定各該機關使用生成式 AI 的規範。

企業評估是否導入AI應用

企業評估是否導入AI應用

評估企業是否導入AI應用是一個多層次的過程，需要考慮**目標設立、數據質量、技術基礎、員工技能、成本效益和風險**等多方面因素。通過全面的評估，可以確保企業在導入AI技術時能夠有效應對挑戰，實現預期目標。

以下是企業考慮導入AI應用的相關依據跟指標

企業評估是否導入AI應用

A. 評估企業需求和現狀	
了解企業的目標和挑戰	1. 目標設立：確定企業希望通過AI技術達成的具體目標，如提升生產效率、降低成本、改進客戶服務等。 2. 挑戰識別：列出目前企業面臨的主要挑戰和痛點，如生產過程中的瓶頸、數據管理困難等。
B. 分析現有數據和技術基礎	
數據留存與數據品質	1. 數據完整性：數據是否完整，沒有遺漏或缺失。 2. 數據一致性：數據在不同系統中是否一致。 3. 數據準確性：數據是否準確和真實。
硬體設施與技術基礎	1. 硬件資源：計算能力和存儲資源是否足夠。 2. 網絡設施：網絡是否穩定並且有足夠的帶寬。 3. 安全設施：是否有完善的網絡安全措施。

企業評估是否導入AI應用

C. 員工技能和培訓需求	
技能現狀	1. 基本理解：員工是否具備基本的AI知識。 2. 實踐經驗：員工是否有實際使用AI技術的經驗。 3. 進階技能：是否有員工具備高階AI技術（如深度學習、強化學習）的能力。 4. 培訓需求：評估是否需要進行專業培訓來提升員工技能。
成本效益分析	1. 成本評估：估算導入AI技術的成本，包括硬件、軟件、培訓和維護等費用。 2. 效益預測：分析導入AI技術後可能帶來的效益，如提高生產效率、降低成本、改進服務質量等。
風險評估	1. 技術風險：評估技術實施過程中可能面臨的挑戰和風險，如數據隱私、技術成熟度等。 2. 運營風險：評估AI技術對現有業務流程和運營模式的影響，並制定風險管理策略。

企業評估是否導入AI應用

D.企業痛點識別及AI技術匹配	
運營效率低	1. 自動化和流程優化技術： 使用機器人流程自動化（RPA）來自動化重複性工作，減少人為錯誤，提高生產效率。 2. 預測維護： 使用預測分析技術來監控設備狀態，預測設備故障，提前進行維護，減少停機時間。 3. 資源優化： 使用AI技術優化資源配置，提高資源利用效率。痛點：生產過程中存在瓶頸，資源配置不當。
成本高	1. 能源管理系統： 使用AI優化能源使用，減少浪費，提高能源利用效率。 2. 供應鏈優化： 使用AI分析供應鏈數據，優化供應鏈管理，降低物流成本。 3. 成本預測： 使用AI進行成本預測和控制，降低運營成本。
客戶服務不佳	1. 自然語言處理（NLP）： 使用聊天機器人來自動回應客戶查詢，提高客戶服務速度和質量。 2. 情感分析： 使用情感分析技術分析客戶反饋，及時了解客戶需求，改進服務。 3. 客戶數據分析： 使用AI分析客戶數據，提供個性化服務，提升客戶滿意度。

企業評估是否導入AI應用

D.企業痛點識別及AI技術匹配	
資料管理困難	1. 資料整合與分析： 使用AI進行資料整合，提供數據驅動的決策支持。 2. 數據清洗與管理： 使用AI技術進行數據清洗和管理，提高數據質量。 3. 數據視覺化： 使用數據視覺化工具，將複雜數據轉化為易於理解的圖表和報告。
市場反應慢	1. 預測分析： 使用AI進行市場趨勢預測，幫助企業及時調整策略。 2. 需求預測： 使用AI技術預測市場需求，提前備貨和生產。 3. 競爭分析： 使用AI技術分析競爭對手，制定應對策略。
產品品質不穩定	1. 品質控制： 使用機器視覺技術進行產品檢測，確保品質一致性。 2. 過程監控： 使用AI技術監控生產過程，及時發現和糾正問題。 3. 品質預測： 使用AI進行品質預測，提前發現潛在問題。
創新能力不足	1. 技術創新： 使用生成式AI技術進行新產品設計和開發，提升創新能力。 2. 研發支持： 使用AI技術支持研發工作，加速技術創新。 3. 創新文化建設： 使用AI分析和鼓勵創新行為，建設創新文化。

3 .延伸閱讀與思維創新

- **參考資料**
  - 負責任 AI 六大準則：https://www.microsoft.com/zh-tw/ai/responsible-ai
  - 公平性：https://www.youtube.com/watch?v=4bqrlZ-CyNs
  - 可靠性和安全性：https://www.youtube.com/watch?v=i3akj3GHmdw
  - 隱私權和保密性：https://www.youtube.com/watch?v=AZZdMgOe60k
  - 包容性：https://www.youtube.com/watch?v=aVgbsRn9zK8
  - 透明度：https://www.youtube.com/watch?v=q5CbK0Hs1pg
  - 責任：https://www.youtube.com/watch?v=5BQ2RE9kqvA
  - 歐盟《人工智慧法案》EU AI Act：https://artificialintelligenceact.eu/ai-act-explorer/
  - 中華民國生成式 AI 參考指引 (草案) https://www.ey.gov.tw/File/CAE5B756153299FD?A=C
- **延伸閱讀**
  - Patrick Hall, James Curtis, Parul Pandey(2024)。機器學習的高風險應用 | 負責任的人工智慧方法，歐萊禮出版。 —檢自https://www.tenlong.com.tw/products/9786263247734(July 28, 2024)
  - 實現可信賴的 AI 應用願景：淺談負責任 AI (2023)—檢自https://www.cio.com.tw/realized-trusted-ai-application-vision-astalks-responsible-ai/ (July 29, 2024)
- **思維創新**
  - 生活中是否有看過或遇到違反負責任 AI 的情形？
  - 針對曾遇到違反負責任 AI 的情形，思考適用什麼方式降低？

【案例集】

服務品質的損害案例

假設有一個AI系統被用於醫療診斷，該系統對於不同種族患者的診斷準確率不同。

如果系統對某一特定種族的患者診斷準確率較低，那麼這些患者就可能因為得不到正確的診斷和治療而受到影響，這種差異就屬於服務品質的損害。

為了避免這種情況，AI系統需要經過全面且多樣化的測試，確保其在所有群體中的效能都是一致的。

圖片來源：Copilot生成圖片

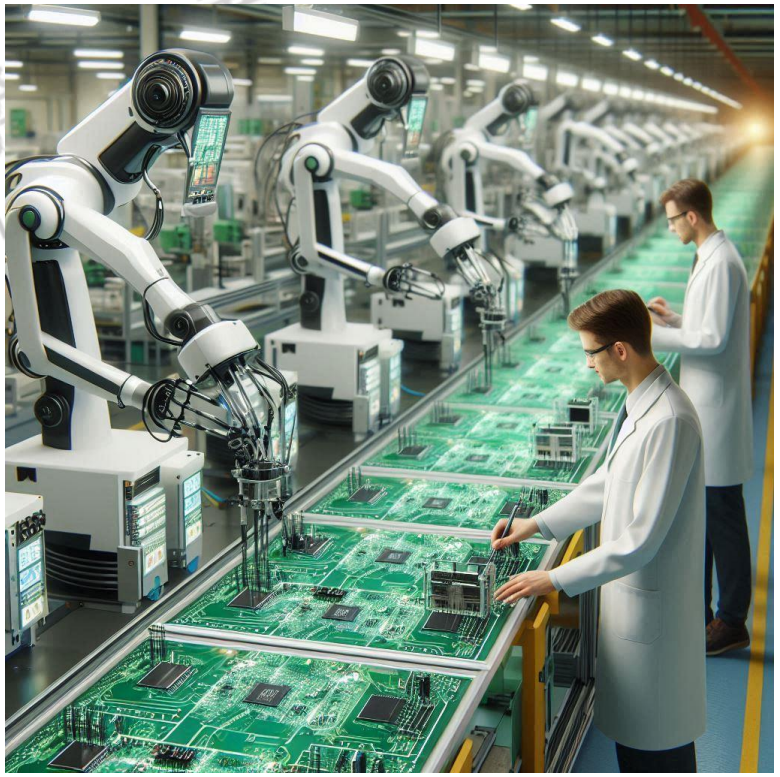
錯誤率的差異案例

一家製造業公司使用AI系統來預測生產線上機台馬達的故障。在新安裝的馬達設備上，由於設備更新且數據品質高，模型的錯誤率只有5%。

然而，對於舊馬達設備，因為使用時間長且數據不夠穩定，模型的錯誤率上升到20%。這表明模型在新舊設備上的預測錯誤率存在顯著差異。

精確度的差異案例

假設一家PCB板製造廠使用AI系統來檢測生產線上的PCB板品質。某些材料批次的品質檢測精確度高達95%，而其他批次的精確度卻只有80%。這表明模型在不同材料批次中的預測精度存在顯著差異。



圖片來源：Copilot生成圖片

召回率中的差距案例

圖片來源：Copilot生成圖片

飲料製造公司使用AI系統來檢測生產線上的瓶裝飲料是否有異物。在生產線速度較慢時，系統能夠識別出95%的異物。然而，當生產線速度加快時，系統的召回率下降到80%。這表明模型在不同生產速度下的表現存在差異。

Azure ML 中的可靠性與安全性的例子

在一家自動駕駛汽車公司中，使用機器學習模型來預測汽車閃避行人的能力。負責任 AI 儀表板的錯誤分析元件揭示，該模型在某些特定條件下（如夜間或惡劣天氣）閃避行人的錯誤率遠高於整體基準。

圖片來源：Copilot生成圖片

資料導向的例子

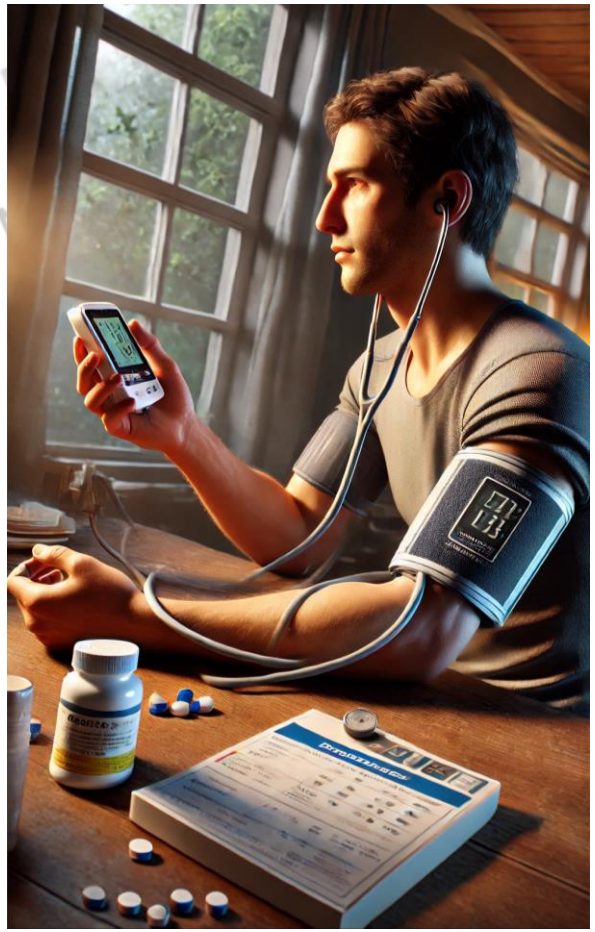
「某個藥品對於病患的血壓有何影響？」

此類深入解析是透過負責任 AI 儀表板的原因推斷元件提供。

例如，對於不同年齡層或不同健康狀況的病患，藥品的效果可能會有所不同。

這些深入的數據分析能夠揭示這些細微差異，幫助醫療專業人士在不同情境下做出更有根據的決策。這不僅提升了治療效果，也確保了每個病患都能接受到最佳的治療方案，從而提升整體的醫療品質和患者的生活品質。

了解這些因果關係還能幫助醫療研究者發現潛在的治療方法和改善現有療法，從而推動醫療技術的進步和創新。



圖片來源：ChatGPT生成圖片

反事實分析案例說明一

應用範例：

假設客戶A的貸款被拒，原因是年收入過低。通過反事實分析，將客戶A的年收入從50,000美元模擬增加到60,000美元，觀察貸款結果。如果發現這一變化導致貸款審批結果從拒絕變為批准，這說明年收入對決策影響重大。

圖片來源：ChatGPT 生成圖片

這時，可以進一步探討年收入的權重設置是否合理，以及是否需要引入其他輔助特徵來提高模型準確性。

反事實分析案例說明二

在自動駕駛車輛系統中，使用影像辨識模型來檢測和識別行人，確保行車安全。如果模型預測某個區域內有行人，車輛會自動減速或停車。以下為影響行人識別的相關影響因素。

光照條件、攝像頭解析度、行人的位置、背景複雜度、行人的衣著顏色、影像預處理等等相關條件。

圖片來源：ChatGPT 生成圖片