推薦序

知名統計學家 George Box 曾有句名言：「所有模型都是錯的，但有些有用。」承認這件事，形成了有效風險管理的基礎。在機器學習逐漸自動化決定我們生活的重要決策的這個世界，模型失敗的後果會是一場災難。審慎緩解風險並避免意外危害非常重要。

2008 金融危機後，監管單位與金融機構認知到管理模型風險確保銀行安全的重要性，重新完善了模型風險管理（MRM）的實作。隨著 AI 與機器學習大範圍採用，MRM 原則也被用來管理這些風險。國家標準與科技機構的 AI 風險管理框架即為這場演進的其中一例。適切治理與控制整體程序，從資深管理層監管，到政策與程序，包括組織性結構與激勵，都是促進模型風險管理文化的關鍵。

在本書中，Hall、Curtis 與 Pandey 介紹了將機器學習套用在高風險決策上的框架。他們透過完整記錄模型失敗案例與新興法規，提供強而有力的證據強調強化治理與文化的重要性。不幸的是，這些原則在銀行這類監管產業外依然很少使用。本書重要主題，範圍涵蓋模型透明度、治理、資安、偏差管理等等。

機器學習裡，只有效能測試是不夠的，因為模型多樣性可以讓差異極大的模型擁有相同效能。模型還必須可解釋、安全與公平。這是第一本強調本質上可解釋模型與其近期發展與應用的書，尤其針對模型影響個人的案例，例如消費金融。在這些可解釋性標準與法規特別嚴格的場景下，可解釋 AI（XAI）事後可解釋處理方法往往面臨極大挑戰。

碁峰 www.gotop.com.tw

開發可靠並安全的機器學習系統，還需要嚴格的模型弱點評估。本書完整呈現兩個範例的模型除錯方法，包括透過誤差與殘差分段識別模型缺陷、評估輸入毀損下的模型穩健性與模型輸出的穩定性或不確定性，並透過壓力測試，瞭解分佈移轉下的模型復原力。這些都是在高風險環境下開發與部署機器學習的關鍵主題。

機器學習模型可能對有史以來邊緣化族群造成程度不一的傷害，透過自動化迅速且大規模傳遞這樣的危害。偏差的模型決策對受保護族群具有害影響，讓社會與經濟不平等永遠存在。讀者可以在本書之中學習到，透過社交科技視角，如何處理模型公平性議題。作者也對模型偏差去除技術的影響作了徹底研究，並對各種受監管垂直產業裡這些技術的應用提供實務性建議。

本書是實務、堅持觀點且及時的一本書。各領域讀者都能發現這個困難重重主題上的豐富的見解，無論是對想更瞭解模型的資料科學家，還是負責確保符合現有標準的經理人，或者試圖改善企業組織風險控制的高層。

— *Agus Sudjianto*，博士

富國銀行企業模型風險主管暨執行副總裁

碁峰 www.gotop.com.tw

前言

如今，機器學習（ML）已是人工智慧（AI）最具商業可行性的分支學科。ML 系統用於就業、交保、假釋、學習、資安等領域與世界各地經濟與政府許多重大影響應用上。公司企業方面，ML 系統出現在組織各個層面，從面對客戶的產品，到員工評估，再到後勤支援自動化等等。過去十年，ML 技術的採用的確相當廣泛。但也證明對 ML 的營運商、客戶甚至一般大眾而言，持續存在風險。

一如所有技術，ML 也會失敗：無意間誤用或者刻意濫用。自 2023 年起，已有數千起演算法歧視、違反資料隱私、訓練資料資安外洩與其他有害意外事件的公開報告。這類風險必須在組織與大眾體驗到這個技術真正好處前緩解。處理 ML 風險，需要專業人員採取行動。雖然初始標準，亦即本書旨在遵循的，已逐漸成型，但 ML 實作上，依舊缺乏廣泛接受的專業認證或最佳實作。意即，技術部署到這世界上時，大多由個別從業人員自負這項技術帶來的好與壞。本書將幫助讀者紮實理解模型風險管理程序，並以新方式使用一般 Python 工具，訓練可解釋模型，針對可靠性、安全、偏差管理、資安與隱私議題為模型除錯，助從業人員一臂之力。

碁峰 www.gotop.com.tw

我們採用 Stuart Russell 與 Peter Norvig《人工智慧：現代方法》（*https://oreil.ly/oosZs*）一書的 AI 定義：設計並建置從環境接收訊號並採取行動影響環境的智慧系統（2020 年）。ML 則使用 Arthur Samuel 的一般定義（非完全可信）讓電腦有能力學習，無須確切編寫程式設計的研究領域（約 1960 年）。

誰該讀這本書

這本偏技術面的書，想寫給希望學習負責任使用 ML 或 ML 風險管理，處於職涯發展早期至中期的 ML 工程師與資料科學家。本書範例程式碼以 Python 撰寫，意即可能不適用所有資料科學家與非使用 Python 的工程師。若想學習模型治理基礎並更新工作流程，進行基本風險控管，這本書適合你。若你的工作需要遵循非歧視性、透明度、隱私權或資安標準，這本書適合你。（但不保證合規性，亦不提供法律建議！）若想訓練可解釋模型，學習編輯與排除問題，這本書適合你。最後，若擔憂 ML 方面的工作可能導致社會學偏見、違反資料隱私、安全性弱點或其他明顯由自動決策引發的已知問題等意外後果，希望能做點什麼，那麼這本書適合你。

當然，或許有其他人也對本書感興趣。若從物理學、計量經濟學或心理學來到 ML，本書有助於學習如何將較新 ML 技術與確立的領域專業知識、有效性與因果論彼此融合。本書能為監管人員或政策專家，就可用於遵循法律、法規或標準的 ML 技術現況，提供一些洞見。技術風險高層或風險管理者，可能會發現本書更新概念提供適於高風險應用的新穎 ML 處理方式相當有用。專業資料科學家或 ML 工程師也可能發現本書的教育意義，同時還挑戰許多確立的資料科學實作。

讀者將學到什麼

本書讀者將知曉傳統模型風險管理，與如何將它與電腦安全性最佳實作融合，包括意外事件應變、漏洞回報獎勵計畫與紅隊演練，再將實戰測試風險管控套用到 ML 工作流程與系統上。

碁峰 www.gotop.com.tw

本書會介紹舊式與新版可解釋模型，與令 ML 系統更透明的解釋技術。建立高透明度模型的紮實基礎，就能深入挖掘測試模型的安全與可靠性。瞭解模型運作方式就輕鬆多了！本書將超越持有資料的品質量測，探索如何將殘差分析、靈敏度分析與基準校正這類知名診斷技術，套用到新型態 ML 模型上。接著以組織與技術角度，針對偏差管理、偏差測試與緩解偏差進行結構化模型。最後，會討論 ML 管線與 API 的安全性。

歐盟《AI 法》草案，將下列 ML 應用歸類為高風險：生物辨識、關鍵基礎架構管理、教育、就業、公家單位（例：公共救助）與私人機構（例：信用借貸）民生必需服務、執法機關、移民與邊境管制、刑事司法與民主程序。提到高風險應用時，應該想到這些 ML 使用案例，這也是本書選擇範例程式著眼於電腦視覺化與表狀資料的樹狀模型上之故。

讀者應該也發現，本書第一版偏重在已確立 ML 模型的判斷與決策制定上。不深入處理非監督式學習、搜尋、建議系統、強化學習與生成式 AI。理由如下：

- 這些系統仍非最常見商用產品系統。
- 在深入瞭解更先進的非監督式、建議的與強化學習或生成處理前，應先掌握基本概念。本書初版致力於讓讀者爾後能掌控更先進專案的基礎概念。
- 這些系統的風險管理，不像本書著重的監督式模型型態那樣容易理解。直接來說（一如本書不斷強調的）使用故障模式、緩解與管控還不明確的模型會增加風險。

我們希望未來能回到這個主題，也認知到這些話題正影響數百萬人口，無論正面或負面。但也發現，只要一點創意與努力，本書許多技術、風險緩解與風險管理框架應該能套用在非監督式模型、搜尋、建議與生成式 AI 上。

碁峰 www.gotop.com.tw

ChatGPT 與 GitHub Copilot 這類走在時代尖端的生成式 AI 系統，是 ML 正以令人興奮的方式影響我們的生活。這些系統看似解決困擾早期類似系統的偏差問題。不過，在處理高風險應用上它們仍有風險。若要使用它們並存有疑慮，就應該考慮以下防護措施：

不要從使用者介面複製貼上
: 不要直接使用生成的內容，也不要將自有內容直接貼在介面上，可限制智慧財產權與資料隱私風險。

檢查所有生成內容
: 這些系統持續生成錯誤的、攻擊性的或其他有問題的內容。

避免自動化自滿
: 整體而言，這些系統較適於內容生成而不是決策支援。應謹慎不要讓這些系統無意間為我們做決定。

與 NIST AI 風險管理框架一致

為遵循我們自己的建議，並讓本書對處理高風險應用能更實用，會強調本書推薦方式中與近期國家標準與科技機構（NIST）的 AI 風險管理框架（RMF）一致之處。外部標準應用一直都是知名風險管理戰術，NIST 在權威技術指南方面有輝煌紀錄。AI RMF 有相當多組成元件，其中最重要的兩個就是 AI 與 RMF 指南核心的可信度特性。可信度特性建立了 AI 風險管理的基本原則，而 RMF 指南核心則提供風險管控執行的建議。全書將使用 NIST 的 AI 可信度特性的相關字彙：效度、可靠性、安全、資安、復原力、透明度、問責、可解釋性、可詮釋性、偏差管理與強化隱私。在第一部分每一章開始，會利用方框解說本書內容如何與何處，與 NIST AI RMF 核心對應、衡量、管理與治理的特定面向保持一致。期望與 NIST AI RMF 保持一致，能讓這本書更有用，成為更有效的 AI 風險管理工具。

NIST 沒有審查、核准、容許，或以任何方式處理本書所有內容，包括宣稱與 AI RMF 相關的部分。所有 AI RMF 內容單純是作者意見，不代表 NIST 官方立場，或 NIST 與本書或其他作者間官方與非官方的關係。

碁峰 www.gotop.com.tw

本書大綱

本書分為三部分。第一部分從實際應用角度探討議題，必要時加上一些理論。第二部分內含具體格式的 Python 範例程式，以結構式與非結構式資料的角度處理第一部分主題。第三部分就如何在現實世界高風險使用案例中取得成功給予難得的建議。

第一部分

第 1 章從深入瞭解待定法規開始，探討產品責任與傳統模型風險管理的整體處理。由於這些實作有許多都是以稍微古板與專業的方式建立模型，與當今常見「快速行動、打破常規」口號相去甚遠，所以還會討論如何將假設失敗的電腦資安最佳實作納入模型治理。

第 2 章介紹迅速發展的可解釋模型生態系統。內容涵蓋深度瞭解廣義相加模型（GAM）系列，也探討許多其他型態的高品質高透明度估計式（estimators）。該章概述各種事後解釋技術，但著眼於負責任 ML 技術這個過度炫染分區裡嚴峻且知名的問題。

第 3 章以考量實際測試模型假設與真實世界可靠性的方式，處理模型驗證，並介紹軟體測試基礎，簡述模型除錯重點。

第 4 章概述轉換到技術偏差測量與緩解方式前，先概述公平與偏差的社交技術層面。接著詳細介紹偏差測試，包括差異影響與區分效度，還會處理已確立及舊有偏差緩解方式，與先進的雙目標、對抗式、前置處理、程序中處理與後置處理的緩解技術。

第 5 章解釋紅隊部署 ML 系統的方式，從電腦資安的基本概念開始，再探討常見 ML 攻擊、對抗式 ML 與強化 ML，結束第一部分。

第一部分各章節，皆以 Zillow 的 iBuying 災難、英國 A-level 之亂、Uber 自動駕駛致命撞毀、Twitter 首度偏差漏洞回報獎勵計畫與真實世界 ML 規避攻擊這類主題相關案例討論結束。各章都會概述內容與 NIST AI RMF 間的一致性。

碁峰 www.gotop.com.tw

第二部分

第二部分以一系列詳盡範例程式章節，延展第一部分的概念。

第 6 章將可解釋增強機（EBMs）、XGBoost 與可解釋 AI 技術，適切帶入基本消費金融範例。

第 7 章在 PyTorch 影像分類上套用事後解釋技術。

第 8 章，針對效能問題為消費金融模型除錯，並於第 9 章對影像分類做同樣的事。

第 10 章內含偏差測試與偏差緩解相關的詳盡範例，第 11 章則提供 ML 攻擊範例與樹狀模型對策。

第三部分

第 12 章用如何在高風險 ML 應用取得成功的一般性建議，結束這本書。這不是快速移動、打破常規。對一些低風險使用案例而言，臨時應急的方式或許有用。但隨著 ML 逐漸受管控且用於較高風險應用，打破常規的後果變得更嚴重。第 12 章提供的是在高風險場景上套用 ML 的可貴實作建議。

對本書第一版的期望，是為時下 ML 常見的難解與時程壓縮的工作流程，提供合法替代方案。內容提供整套詞彙、想法、工具與技術，讓從業人員在如此重要的工作上能更深思熟慮。

範例資料集

本書仰賴兩大資料集，解釋技術或闡釋方法與討論結果。這些範例資料集不適合在高風險應用中訓練資料，但大家都知道它們而且很容易取得，其缺點得以讓本書指出各種資料、建立模型與解釋陷阱。接下來的章節將多次參照這些資料集，所以務必在深入本書其餘部分前，瞭解它們。

碁峰 www.gotop.com.tw

台灣信用資料

在第 6、8、10 與 11 的結構性資料章節，使用加州大學爾灣機器學習儲存庫（*https://oreil.ly/xJ5u2*）或 Kaggle（*https://oreil.ly/DmAWe*）略為修改的台灣信用資料版本。信用卡違約資料內含 2005 年台灣信用卡用戶相關的人口統計與支付訊息。一般來說，此資料集的目標，是使用過去支付狀態 (PAY_*)、過去支付金額 (PAY_AMT*) 與帳單金額 (BILL_AMT*) 作為輸入，預測客戶是否完成下次支付 (DELINQ_NEXT = 0)。貨幣金額以台幣呈現。本書還將模擬的 SEC 與 RACE 標記加入這個資料集，以便舉例說明偏差測試與緩解方法。利用支付訊息作為輸入特徵，並依循最佳實作管理 ML 系統的偏差，不使用人口統計資訊作為模型輸入。完整資料字典見表 P-1。

表 P-1　信用卡違約資料的資料字典

名稱	模型化角色	測量程度	說明
ID	ID	Int	唯一識別碼
LIMIT_BAL	Input	Float	過去授與信用額度
SEX	Demographic information	Int	1= 男性；2= 女性
RACE	Demographic information	Int	1= 西班牙；2= 黑人（Black）；3= 白人（White）；[a]；4= 亞洲人
EDUCATION	Demographic information	Int	1= 研究所；2= 大學；3= 高中；4= 其他
MARRIAGE	Demographic information	Int	1= 已婚；2= 單身；3= 其他
AGE	Demographic information	Int	年齡
PAY_0, PAY_2–PAY_6	Input	Int	過去支付歷史；PAY_0=2005 年 9 月還款狀態；PAY_2=2005 年 8 月還款狀態；…PAY_6=2005 年 4 月還款狀態。還款狀態評估級距為：-1= 按時還款；1= 還款延遲一個月；2= 還款延遲兩個月；…8= 還款延遲八個月；9= 還款延遲九個月或九個月以上
BILL_AMT1–BILL_AMT6	Input	Float	帳單金額；BILL_AMT1=2005 年 9 月帳單金額；BILL_AMT2=2005 年 8 月帳單金額；BILL_AMT6=2005 年 4 月帳單金額

碁峰 www.gotop.com.tw

名稱	模型化角色	測量程度	說明
PAY_AMT1–PAY_AMT6	Input	Float	過去支付金額；PAY_AMT1=2005 年 9 月付款金額；PAY_AMT2=2005 年 8 月付款金額；PAY_AMT6=2005 年 4 月付款金額
DELINQ_NEXT	Target	Int	客戶下次是否延遲付款（逾期），1= 逾期；0= 按時

[a] 提及人種的人口統計族群時，「White（白人）」是否應與「Black（黑人）」一樣大寫，一直是爭議。基於認可共有歷史與文化認同，全書依循出版界與學術界眾多權威人士的看法（*https://oreil.ly/3iKFj*），以首字母大寫「Black（黑人）」表達。

接下來章節讀者會發現，這個資料集編碼略有缺陷。它太小以致於無法訓練高容量 ML 評估，而且幾乎所有 DELINQ_NEXT 訊號都編碼在 PAY_0 裡。隨著本書進行，我們會試圖處理這些議題並找出其他問題。

Kaggle 胸部 X 光資料

在深度學習章節：第 6 章與第 9 章，會使用到 Kaggle 胸部 X 光影像資料集（*https://oreil.ly/TsoGB*）。此資料集由肺炎與正常兩個類別的 5,800 張影片組合而成。這些標籤由人類領域專家決定，這些影片是去識別化的胸部 X 光，取自中國廣州婦幼醫療中心日常護理就診期間。肺炎案例影像見圖 P-1。

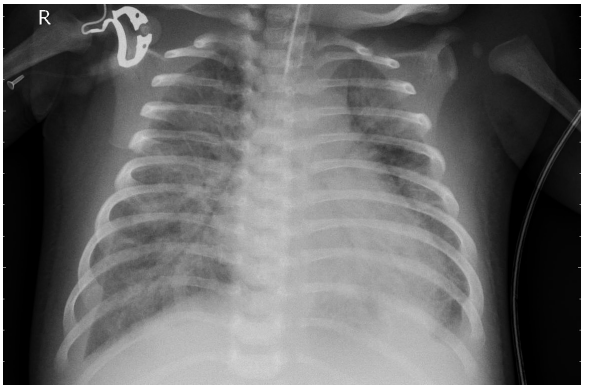

圖 P-1　取自 Kaggle 胸部 X 光資料集肺炎案例影像

碁峰 www.gotop.com.tw

第四章

機器學習偏差管理

管理機器學習系統中偏差造成的危害，指的不只是資料、程式碼和模型。模型平均效能品質，也就是資料科學家被教導評估模型好壞的主要方式，與是否造成現實世界偏差危害關係不大。完全準確的模型可能引發偏差危害。更糟的是，所有 ML 系統都會呈現某種程度的偏差，偏差意外事件會出現在一些最常見的 AI 意外事件中（見圖 4-1），商業程序中的偏差多半會有法律責任，ML 模型偏差會傷害現實世界的人。

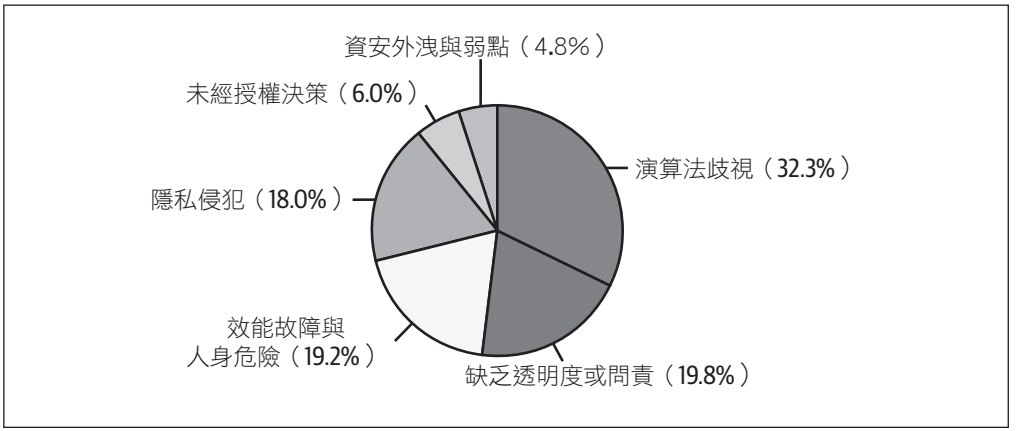

圖 4-1　以 1988 年至 2021 年 1 月 1 日期間，169 份公開報告的意外事件品質分析為基礎，各類型 AI 意外事件的發生率。

碁峰 www.gotop.com.tw

本章將介紹偵測與緩解社交技術型式偏差的處理方式，至少盡實作技術人員的最大能力。意即將試圖理解在廣泛社會背景環境下何以存在 ML 系統偏差。為何這麼做？所有 ML 系統都是社交技術。我們知道一開始很難相信，所以來想想一個範例：用來預測物聯網（IoT）應用感應器失敗的模型，僅使用來自其他自動感應器的資訊。這個模型可能已接受人類訓練，或由人類決定模型需要訓練。模型產生的結果可用於通知新感應器下單，此舉會影響相關製造廠，或修復與置換故障感應器的就業狀況。最後，若預防性維護模型故障，會危及與該系統互動的人。對於每個能想像得到、看似純粹的技術性例子，顯然像 ML 這類決策制定技術，若不以某種方式與人類互動就不存在。

這代表對 ML 系統的偏差而言，沒有純然技術的解決方案。若讀者想直接跳到偏差測試與偏差補救的程式碼，請見第 10 章。不過我們不建議這麼做，因為會錯過許多何謂偏差與如何以生產力的方式考量偏差等重要資訊。本章一開始將以幾個不同的權威來源定義偏差，接著是如何識別自有認知偏差會影響建置的 ML 系統，或使用者詮釋的結果。之後，本章廣泛概述 AI 偏差意外事件裡會受傷害的人，以及他們會經歷哪種危害。在此，將說明測試 ML 系統偏差的方法，並討論使用技術與社交技術方式兩種方式緩解偏差。最後，將以 Twitter 影像裁切演算法案例研討，結束本章。

雖然偏差管理某些層面必須針對模型特定架構調整，但大部分偏差管理並非模型專屬。本章許多想法，尤其是參考 NIST SP1270 偏差指南與 Twitter Bias Bounty 的部分，都能套用在各種設計精良的 AI 系統上，像是 ChatGPT 或 RoBERTa 語言模型。若讀者想瞭解這方面實作，可參考 IQT Labs 的 RoBERTa 稽核（*https://oreil.ly/3hs_6*）。

NIST AI RMF 對照表

章節	NIST AI RMF 子類別
第 126 頁「系統偏差」	MAP 1.6、MAP 2.3、MEASURE 2.11
第 126 頁「統計偏差」	MAP 2.3、MEASURE 2.6、MEASURE 2.11
第 127 頁「人類偏見與資料科學文化」	GOVERN 3.2、MAP 1.1、MAP 2.3、MEASURE 2.11

碁峰 www.gotop.com.tw

章節	NIST AI RMF 子類別
第 128 頁「美國 ML 偏差的法律觀念」	GOVERN 1.1、GOVERN 1.2、GOVERN 1.4、GOVERN 2.2、GOVERN 4.1、MAP 1.1、MAP 1.2、MEASURE 2.11
第 131 頁「誰容易遭受 ML 系統偏差對待」	GOVERN 1.2、GOVERN 1.4、GOVERN 5、MAP 1.1、MAP 1.2、MAP 1.6、MAP 2.2、MAP 3、MAP 5、MEASURE 1.3、MEASURE 4、MANAGE 2、MANAGE 4
第 133 頁「人們經歷的傷害」	GOVERN 1.2、GOVERN 1.4、GOVERN 5、MAP 1.1、MAP 1.2、MAP 1.6、MAP 2.2、MAP 3、MAP 5、MEASURE 1.3、MEASURE 4、MANAGE 1.4、MANAGE 2、MANAGE 4
第 135 頁「測試資料」	MEASURE 1.1、MEASURE 2.1
第 137 頁「傳統處理方式：等效結果測試」	GOVERN 1.1、GOVERN 1.2、GOVERN 1.4、GOVERN 4.3、GOVERN 6.1、MAP 2.3、MAP 4、MEASURE 1、MEASURE 2.1、MEASURE 2.6、MEASURE 2.11、MEASURE 4.2
第 141 頁「新思維：等效執行效能品質測試」	GOVERN 1.2、GOVERN 1.4、GOVERN 4.3、GOVERN 6.1、MAP 2.3、MAP 4、MEASURE 1、MEASURE 2.1、MEASURE 2.6、MEASURE 2.11、MEASURE 4.2
第 143 頁「即將發生：廣泛 ML 生態系統測試」	GOVERN 1.2、GOVERN 1.4、GOVERN 4.3、GOVERN 6.1、MAP 2.3、MAP 4、MEASURE 1、MEASURE 2.1、MEASURE 2.6、MEASURE 2.11、MEASURE 4.2、MANAGE 3.2
第 148 頁「緩解偏差中的技術因素」	MAP 2.3、MANAGE
第 148 頁「科學方法與實驗設計」	MAP 1.1、MAP 2.3
第 153 頁「緩解偏差中的人類因素」	GOVERN 2.1、GOVERN 3、MAP 1.1、MAP 1.6、MAP 2.2、MAP 2.3、MEASURE 3.2、MEASURE 3.3、MANAGE 1、MANAGE 2.1、MANAGE 2.2、MANAGE 3、MANAGE 4.2、MANAGE 4.3

適用的 AI 可信任度特性包括：可管理偏義、透明度與問責、有效與可靠

- 參閱：
  - —「識別與管理 AI 偏差的相關標準」（*https://oreil.ly/8kpf5*）
  - — 完整對照表（非官方資源）（*https://oreil.ly/61TXd*）

碁峰 www.gotop.com.tw

ISO 與 NIST 的偏差定義

國際標準組織（ISO）在「統計學：字彙與符號 - 第一部分」將偏差定義為「參考值偏離事實的程度」（*https://oreil.ly/YYv4W*）。這是偏差的常見概念，但偏差其實是一種既複雜又非均質的現象。而且，所有實例多少都帶有點系統性偏離事實。就決策制定任務而言，偏差的型式很多。因膚色黑色素程度而拒絕僱用在實質與道德上都是錯誤。只因為這個想法一開始就飄進腦海就認為它正確，這事實上就是錯誤。在不完整且無代表性資料上訓練 ML 模型也是實質上與道德上的錯誤。NIST 近期成果「為人工智慧裡的偏差識別與管理制定標準」（SP1270）（*https://oreil.ly/pkm4f*），將偏差主題依這些偏差樣本分為三大類：系統性、統計性與人類偏見。

系統性偏差

提到 ML 裡的偏差時，指的就是系統性偏差。這些歷史悠久的社會與制度偏見，不幸地融入我們的生活，以致於預設情況下會出現在 ML 訓練資料與設計選項。ML 模型裡的系統性偏差常見結果是將人口統計資訊併入系統機制。這種合併可以是公然而明確，像是語言模型（LMs）重新改變用途，產生針對特定人口統計族群的有害與攻擊性內容（*https://oreil.ly/bWf4E*）。不過，實作上人口統計資訊合併至決策制定程序，往往是無意間且不明確，導致整個人口統計族群有不同的結果率與結果盛行率，例如從系統互動中，將更多男性履歷表與高報酬工作說明比較，或排除特定使用者族群（例如肢體障礙人士）的設計問題。

統計性偏差

統計性偏差可以想成人類在 ML 系統規格犯下的錯，或像概念飄移這種自然出現的現象，影響 ML 模型且人類難以緩解。其他統計性偏差常見型態包括以無代表性訓練資料為基礎的預測，或錯誤傳播與回饋迴路。ML 模型裡統計性偏差的一項可能指標，是整個不同資料截面的效能品質差異，像是人口統計族群。ML 模型的差異效度是特定型態偏差，不同於人類偏見所述的不同結果率與結果盛行率。

碁峰 www.gotop.com.tw

事實上，已有文件記錄最大化人口統計族群中的模型效能，與維護正結果率品質間的緊張情勢（*https://oreil.ly/cJy7F*）。統計性偏差也可能導致重大 AI 意外事件，例如新資料致使系統決策錯誤高過正確的概念飄移，或者回饋迴路或錯誤傳播導致短時間內不良預測越來越多。

人類偏見與資料科學文化

設計、實施與維護 ML 系統的個人與團隊都可能投入許多人類偏見或認知偏見。NIST SP1270 指南有更完整人類偏見清單。下列人類偏見最常影響 ML 系統資料科學家與使用者：

定錨

當特定參考點或錨，對人們的決策造成不當影響。這就像最新深度學習模型長期停留在 0.4 AUC，而有人出現得到 0.403 AUC。我們不該覺得這很重要，而是繼續定錨在 0.4。

可得性經驗法則

決策制定過程中，人們往往過份看重輕鬆或快速想到的那些。另一方面來看，我們常將容易記住與正確混淆。

確認性偏差

認知偏差是人們傾向相信與自己既定信仰一致或確認的資訊。欺騙自己認為我們的 ML 模型運作的比實際上好，確認偏差即為 ML 系統的嚴重問題。

鄧寧 - 克魯格效應

既定領域或任務裡，能力低下者高估自我評定能力的傾向。只因為能 `import sklearn` 並執行 `model.fit()` 就自認為在某些事上是專家，就會發生這種效應。

資金偏見

強調或推廣支援的資方 / 專案資助者或令其滿意的結果。我們做讓老闆開心的事、讓投資者開心的事，然後增加收入。真正的科學需要防護措施，預防不被有偏見的金融利益左右發展。

碁峰 www.gotop.com.tw

團體迷思

群體中的人傾向於順應團體或害怕與群體意見不同，而做出非最佳決策。與所處團體持不同意見很難，就算自信自己是對的。

麥納馬拉謬誤

深信應單由量化資訊做出決策，無法輕鬆量測的量化資訊或資料點可以犧牲。

科技沙文主義

深信技術永遠是解決方案。

這裡所有偏差都能夠且確實會導致不適切與過度樂觀的設計選擇，繼而在系統部署後帶來糟糕的效能，最後引發對系統使用者與操作者的傷害。我們很快會切入可能發生的危害與針對這些問題能做些什麼。現在，要強調的是常識性緩解，亦即本章主題。沒有從眾多不同觀點找出問題就無法適切處理偏差。對抗 ML 偏差第 0 步，就是在制定系統重要決策時，會議室（或視訊會議裡）擁有多元化利益相關人群體。要避免允許有偏差的 ML 模型引發危害的盲點，會需要多種不同角度的情報系統設計、實施與維護的決策。是的，這裡說的就是從不同人口統計觀點蒐集輸入，包括那些身心障礙人士的觀點，還有教育背景，像是社會科學家、律師與領域專家。

而且，要考量*數位落差*。仍然無法取得良好網際網路連線品質、新電腦與本書這類資訊的人口比例相當高。若要對使用者做出結論，就必須記得有一大群人被排除在使用者統計之外。忽略潛在使用者，是 ML 生命週期中系統設計、偏差測試與其他關鍵接合上，偏差與傷害的大部分來源。今日 ML 要取得成功，還是需要對正試圖解決的現實世界問題有敏銳瞭解，及哪些潛在使用者可能被我們的設計、資料與測試排除在外。

美國 ML 偏差的法律觀念

我們要瞭解許多重要的偏差法律觀念。不過，知道法律系統極其複雜且與背景環境有關也很重要。只知道一些定義，還是得花上好幾年才能成為這些問題真正的專家。身為資料科學家，法律事務是不應該任由過於自信（鄧寧 - 克魯格效應）接管的領域。知道這些注意事項，就可以開始進入基本概念。

碁峰 www.gotop.com.tw

若對 ML 模型偏差有任何問題或疑慮，請馬上聯繫你的法律團隊。處理 ML 模型偏差，是資訊經濟中最困難與最嚴重的問題。資料科學家需要律師協助，確切處理偏差風險。

在美國，影響大眾決策制定程序裡的偏差數十年來一直受到監管。美國早期法律與法規主要重點在於就業事務。受保護族群、差別對待與差別影響的概念，如今已散布到廣泛的消費金融與住宅法律之中，甚至被全新地方法引述。像是紐約市僱用過程中使用 AI 的稽核需求。《歐盟基本權利憲章》（Charter of Fundamental Rights）、《歐盟人權公約》（European Convention on Human Rights）與歐盟運作條約（Treaty on the Functioning）皆闡述了歐盟的非歧視性，對我們來說最重要的是擬議的歐盟《AI 法》。雖然無法概括這些法律與法規，即便只有美國，但以下定義，是我們認為資料科學家日常作業最直接面對的。它們是從《民權法》（Civil Rights Act）這類法律、《公平住屋法》（Fair Housing Act）、公平就業機會委員會（Equal Employment Opportunity Commission）法規、《公平信貸機會法》（Equal Credit Opportunity Act）與《美國身心障礙法》（Disabilities Act）中，非常概略性提取。下列定義含括哪些特徵受法律保護，與這些法律想保護我們免於什麼傷害的法律觀念：

受保護族群

美國，許多法律與法規禁止對於種族、生理性別（某些情況下，或稱社會性別）、年齡、宗教信仰、國籍與失能狀態等其他型式的歧視。FHA（公平住屋法）禁止的決策依據，包括種族、膚色、宗教、國際與生理性別、家族狀態與失能。非美國法規的例子：歐盟的《GDPR》，禁止人種或民族血統、政治傾向與其他與美國受保護族群相似類別的個人資料利用。這是何以傳統偏差測試，要比較受保護族群，與所謂*參考*族群的非受保護族群相關結果之故。

差別對待

差別對待是歧視的特殊型態，諸多產業中皆屬違法。由於種族、生理性別或其他特徵這類受保護特性，決定對某人的待遇不如情況相似的其他人。就資料科學家處理就業、住屋或信貸申請而言，這代表使用 ML 模型裡、甚至在偏差補救技術上的人口統計資料時，應尤其謹慎。將人口統計資料當成模型輸入，就代表針對某人的決策可能會因為人口統計而有所不同，以及部分情況下可能導致差別對待。

碁峰 www.gotop.com.tw

對差別對待與更常見系統偏差的疑慮，就是傳統上不以人口統計標記作為 ML 模型直接輸入的原因。保守起見，人口統計標記不應作為大部分常見場景的模型輸入，但應該用在偏差測試或監控程序。

差別影響

差別影響是另一種法律性歧視。基本上是關於整體人口統計族群的不同結果率或盛行率。較正式的定義是，差別影響是看似中立的政策或實作結果，大大傷害受保護族群。就資料科學家來說，差別影響往往發生在不使用人口統計資料作為輸入，而是使用人口統計資料相關產物為輸入。以違約的公正的準確預測項信用積分為例：它們準確預測違約，因此在消費者借貸的預測模型中通常視為有效。但它們與種族有關，例如有些少數族群信用積分低於平均值。若在模型裡使用信用積分，常導致特定少數族群正向結果比例偏低，這就是差別影響的常見例子。（這也是何以某些州已開始禁止在某些保險相關決策上使用信用積分。）

差異效度

差異效度有時是就業市場造成。差別影響通常指整體人口統計族群有不同結果率，而差異效度則偏向於整個族群不同的**效能品質**。當就業測試對部分族群來說是比其他族群更好的工作效能指標時就會發生這種狀況。差異效度的重要性是由於數學基礎會概括近乎所有 ML 模型，法律結構不會。使用無代表性訓練資料，建構一個在某些族群表現優於其他的模型相當常見，近期許多偏差測試方式將重點放在這類偏差上。

篩選

篩選是非常重要的歧視型態，突顯 ML 系統社交技術性質，且證明了測試與權衡模型積分完全不足以防止偏差。當視力有限或精細動作技能困難這類失能者，無法與就業評估互動，預設會遭工作或升職排除時，就會發生篩選。篩選是嚴重議題，EEOC 與勞工部（Department of Labor）已特別關注（*https://oreil.ly/c0y9i*）這方面 ML 的使用狀況。注意，篩選不一定能夠透過數學偏差測試或偏差補救修復，通常必須在系統設計階段處理，設計師確保失能者可以操作最終產品介面。篩選也凸顯了何以建置 ML 系統時，希望有律師觀點與失能者角度。沒有這些觀點，太容易在建置 ML 系統時忘記失能者這群人，有時還會引發法律責任。

碁峰 www.gotop.com.tw

對偏差的一般定義討論到此告一段落。如讀者所見，這是影響所有人種、科學與法律方面，複雜且多面向的主題。本章稍後討論偏差測試，將在這些定義中，加入更具體但更令人憂心的偏差數學定義。接下來，將概述哪些人較可能遭受 ML 系統偏差與相關危害。

容易遭受 ML 系統偏差對待的對象

任何人口統計族群與 ML 系統互動都會遭受偏差與相關危害的對待，但歷史告訴我們，特定族群更可能、更經常遭受偏見與危害對待。事實上，這是監督式學習的本質：只從過去記錄學習並重複模式，這麼做往往導致年長者、失能者、移民、有色人種、女人與非常規性別個體，面臨更多來自 ML 系統的偏差。另一方面，在現實世界或數位世界經歷歧視的這些人，與 ML 系統交手時可能也會經歷這一切，因為所有歧視皆記錄在資料裡並用於訓練 ML 模型。本節所列族群多半受各種不同的法律保護，但並非一定如此。他們通常（但不一定）是兩個人口統計族群間，積分或結果的統計均等對照組。

很多人屬於多個受保護或邊緣化族群。歧視交叉性的重要概念告訴我們，社交傷害集中在處於多重受保護族群的人，偏差不應該只依單一族群維護的受影響邊緣化族群分析（*https://oreil.ly/3ZaPy*）。舉例來說，AI 道德調查人員近期表示（*https://oreil.ly/DMu8o*），部分臉部識別系統商品存有重大性別歸類準確度差異，膚色較深的女人是最容易歸類錯誤的族群。最後，在定義這些族群前，考量麥納馬拉謬誤也相當重要。將細微差別的人類置入這類生硬的分類學是對的嗎？

答案可能是否定，而且可能指派這些簡化的族群，通常是因為這些類別很容易在資料庫裡以二進制標記欄表示，而這也是偏差與可能危害的來源。考量到管理 ML 系統偏差總有一堆警告，請謹慎定義簡化的人口統計族群，它們往往面臨更多歧視，且多半用來當成傳統偏差測試的對照組：

年齡

年長者（多半 40 歲以上），較有可能在線上內容遭到歧視。在就業、住屋或消費金融這類傳統應用上，年齡截止值可能較高。不過，參與醫療保險或終生金融財富累積，可能讓年長者比較容易受到青睞。

碁峰 www.gotop.com.tw

失能

具有生理、心理或情感障礙者，或許是最有可能遭受 ML 系統偏見的人。篩選的想法在就業行為外相當普遍，即便法律觀念並非如此。身心障礙人士常在 ML 系統設計時被遺忘，再多的數學偏差測試或補救都無法彌補。

移民身分或國籍

生活在非出生國家具移民身分的人，包括入籍公民，面對重大偏見挑戰是眾所周知的事。

語言

尤其是 ML 系統重要領域：線上內容，對那些使用英文以外語言，或以非拉丁字母撰寫的人，更容易體驗 ML 系統偏差對待。

種族與民族

白人以外種族與族群，包括被認定為一個種族以上的人，通常與 ML 系統互動時會遭到偏差對待與傷害。有些還偏好皮膚色調而不是傳統種族或族群標籤，尤其是電腦視覺任務。Fitzpatrick 量表（https://oreil.ly/NJfBP）即為膚色等級的範例。

生理性別與社會性別

順性別以外的生理性別與社會性別男性，更容易遭受 ML 系統掌控的偏差對待與傷害。就線上內容領域，女性較容易受到喜歡，只是以有害方式。名為*男性凝視*的現象，指的是女性相關媒體可能較受歡迎且收到正面對待（例如在社交媒體提要中宣傳），尤以內容導向物化、征服或性感化女性為甚。

多元族群

屬於兩個或以上前述族群的人，遭受偏差或危害遠大於單純屬於兩個廣泛族群的總合。本章所述所有偏差測試與緩解做法皆應考量多元族群。

會遭受 ML 模型偏差對待的當然不只這些族群，無論動機為何，將人們分組都有問題。然而，重點在於知道從哪裡開始尋找偏差，然後期望這份清單足以實現目標。現在知道 ML 偏差從何找起，就要進一步討論應密切留意的最常見危害。

碁峰 www.gotop.com.tw

人們遭受的傷害

發生在線上或數位內容的常見傷害型態很多。這些不但經常發生，而且可能就是太常發生所以視而不見。以上重點列出常見傷害，並提供案例，下次看見就容易識別。這些傷害和 Abagayle Lee Blank 的「電腦視覺機器學習與未來導向的道德規範」（*https://oreil.ly/-JmJA*）非常相似，該文敘述電腦視覺發生危害的案例：

誹謗

積極貶損或攻擊的內容，例如由 Tay（*https://oreil.ly/2938n*）或 Lee Luda（*https://oreil.ly/nRzs1*）這類聊天機器人生成的攻擊性內容。

抹除

抹除挑戰主流社會典型或過去邊緣化族群承受傷害的內容，例如查禁討論種族主義或呼籲白人至上的內容（*https://oreil.ly/FZdDB*）。

除名

將白人、男性或異性戀視為主要人類常態，例如線上搜尋「CEO」結果回傳首位女性為芭比娃娃（*https://oreil.ly/m-zR-*）。

錯誤識別

誤認人的身分或無法識別某個人類，例如在自動影像標記中錯誤識別黑人（*https://oreil.ly/GjyTI*）。

刻板印象

指派特性給一個族群所有成員的傾向，例如 LM 會自動將穆斯林與暴力建立關聯（*https://oreil.ly/eqAgw*）。

代表性不足

模型輸出中，缺乏公平或充份人品統計族群代表性，例如生成式模型認為所有醫生都是白人男性，而所有護士皆為白人女性（*https://oreil.ly/V64lj*）。

有時這些傷害影響僅限於線上或數位空間，但隨著數位生活與生活各方面開始大量重疊，傷害也開始波及現實世界。不當拒絕人們存取所需資源，健康醫療、就業、教育或其他高風險領域的 ML 系統可能造成直接傷害。顯然由 ML 系統造成現實世界傷害的型態包括：

碁峰 www.gotop.com.tw

經濟損害

ML 系統降低經濟機會或部分活動價值，例如男性看到好工作的廣告（*https://oreil.ly/BT-cI*）比女性更多。

實質傷害

ML 系統傷害或殺害某人，例如人們過度依賴自動駕駛車（*https://oreil.ly/BxH5Y*）。

心理傷害

ML 系統引發心理或情感上的憂慮，例如向兒童推薦令人不安的內容（*https://oreil.ly/pQRYE*）。

名譽損害

ML 系統損害個人或企業組織的聲譽，例如消費信貸產品推出，因遭歧視指控（*https://oreil.ly/Wbvq5*）而受害。

遺憾的是，ML 系統使用者或受支配對象可能還有額外傷害，或奇特方式表現的危害組合。在深入下一節各種不同的偏差測試前，記得與使用者一起檢查，確保他們不會遭受這裡探討的傷害，或其他型式危害，這或許是追蹤 ML 系統偏差最直接方式之一。事實上，基本觀念在於，人們是否遭受傷害，遠比積分組合是否通過必有問題的數學測試來得重要。在設計系統時，必須考量這些傷害，與使用者聊聊確保他們不會受到傷害，再尋求減輕傷害。

偏差測試

若 ML 系統有傷害人類的可能，就應該測試偏差。本節旨在涵蓋測試 ML 模型偏差最普遍的處理方式，讓讀者能從這個最重要的風險管理任務開始。測試既非簡單明瞭也非決定性。如同效能測試，系統在測試資料上可能看起來很好，但部署後就故障或引發危害。或系統在測試與部署時出現少量偏差，但經過一段時間後變成做出偏差或有害的預測。甚至，許多測試與影響程度量測帶有已知缺陷且彼此衝突。要進一步瞭解這些議題，可以看看普林斯頓大學教授 Arvind Narayanan 在 ACM 的 ML 公平、問責與透明度會議的演說「21 個公平定義及其政治」YouTube 影片（*https://oreil.ly/4QnqM*）。想知道何以不能一次簡單最小化所有偏差指標的深度數學分析，請參考「風險評分公平測定的固有取捨」（*https://*

碁峰 www.gotop.com.tw