僅供委員問政所需參考
不代表本院意見或立場

編號：3109

議題研析

一、題目：人工智慧金融歧視之法制研析

二、議題所涉法規

人工智慧基本法、金融業運用人工智慧指引

三、背景說明（緣起）

數位科技的快速發展加速 AI 的創新與應用，為現代人的生活及工作型態帶來諸多改變，惟多有案例指出，AI 進行自動化決策對當事人產生法律效果或類似之重大影響，往往因演算法深度學習系統資料來源存在偏見或歧視，使 AI 做出侵害當事人權益的判斷，引發許多法律及道德問題。金融業為機器學習最早應用的領域[1]，歷來時有發生 AI 導入核貸業務引發之自動化決策歧視案件，本文爰以 AI 導入金融信貸業務出發，探討利用可解釋之 AI 緩解 AI 歧視之法制問題。

四、問題爭點

AI 歧視係當今 AI 監理法制重點之一，為降低 AI 歧視所致之危害，應於自動化決策過程中加入適當之「人的參與[2]」，近年來隨著可解釋 AI 的擴大應用，AI 歧視問題似有望得到相當程度的緩解，本文爰以 AI 於金融信貸服務之應用為例檢視技術實務及法制，並就強化 AI 的可解釋性以減緩 AI 歧視問題試析相關內容。

五、探討研析

（一）AI 於金融信貸應用之歧視問題及緩解方案

[1] Pedro Domingos (著)，張正苓、胡玉城(譯)，《大演算：機器學習的終極演算法將如何改變我們的未來，創造新紀元的文明？》，三采文化，105 年 8 月 5 日，頁 56-57。

[2] OECD AI 原則中，即有強調 AI 之非歧視性，並於 AI 運用中保留人力代理與監督之空間。OECD，Recommendation of the Council on Artificial Intelligence，2024 年 5 月 3 日，網址：https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449，最後瀏覽日期：115 年 6 月 18 日。

金融機構於信貸業務中導入 AI 應用，雖能系統化促進信用評分及核貸效率，惟過度依賴演算法做為決策依據，可能因自帶偏見的演算法訓練資料，導致對特定性別、種族或弱勢群體做出不公平的差別待遇。有學者從技術層面試析，認 AI 歧視可能歸因於 AI 開發過程中，在抽樣偏誤、標記偏誤、特徵選擇偏誤、誤差率等面向存有設計問題[3]。為進一步管理 AI 風險，學界多有提出以「可解釋 AI」為基礎之解決方案[4]，金融監督管理委員會（下稱金管會）於 112 年 10 月 17 日所發布之「金融業運用 AI 指引」，亦將運用 AI 系統時應落實透明性及可解釋性納入 AI 運用之核心原則之一[5]。

前述所謂「可解釋 AI」，係指一套能使用戶理解 AI 決策背後判斷邏輯之模型或方法[6]。在資訊科學領域，其內涵包括理解後並進行說明的可解釋性（interpretability）以及透明度（transparency）；在法學領域，2018 年[7]5 月 15 日生效之歐盟個人資料保護規則（European Union's General Data Protection Regulation, GDPR）雖已賦予被自動化決策影響之個人有要求解釋之權利，惟並未針對此項權利有明確釐清[8]，引發討論。有論者認為，在需要較高程度的解釋時（例如：公部門所為之自動化決策），其解釋之方法應著重於 AI 之可解釋性，以及提供當事人關於與自己類似決定的人們的資訊，以及演算法訓練資料之概述、模型種類、最重要因素及模型成效等[9]。亦有論者針對 AI 歧視風險管理，提出「確認 AI 並未直接使用受保護特徵」、「應用可解

[3] Antje von Ungern-Sternberg，〈Discriminatory AI and the Law: Legal Standards for Algorithmic Profiling〉，《The Cambridge Handbook of Responsible Artificial Intelligence Interdisciplinary Perspectives》，2022 年 10 月 28 日，頁 261-263。

[4] 楊岳平，〈人工智慧歧視與可解釋人工智慧－以人工智慧金融信貸為例〉，《月旦法學雜誌》，第 37 期，115 年 5 月，頁 42。

[5] 金融監督管理委員會，金融業運用人工智慧（AI）指引，113 年 6 月，頁 20，網址：https://www.fsc.gov.tw/websitedowndoc?file=chfsc/202408231741530.pdf&filedisplay=%E9%99%84%E4%BB%B6_%E9%87%91%E8%9E%8D%E6%A5%AD%E9%81%8B%E7%94%A8AI%E6%8C%87%E5%BC%95.pdf，最後瀏覽日期：115 年 6 月 18 日。

[6] 郭昀翰，〈論普惠金融時代下之金融排擠效應及其因應之道－以美國法制經驗為中心〉，《華岡法粹》，第 69 期，109 年 11 月，頁 160。

[7] 本文有關年分之使用，原則以民國紀年表述，惟涉及外國法制或立法例部分，改採西元紀年表述。

[8] 黃詩淳，〈AI 可解釋性的法學意義及其實踐*〉，《臺大法學論叢》，112 年 11 月，頁 933。

[9] 同前註，頁 931-932。

釋 AI 分析是否存在差別待遇」、「分析 AI 針對受保護特徵之差別待遇是否具備正當性」等具體作法[10]。

（二）我國法制之檢視

我國一般性 AI 法制中，有關 AI 偏見及可解釋 AI 之規定，見於人工智慧基本法第 4 條第 5 款及第 6 款[11]有關政府推動 AI 之研發及應用之「透明與可解釋」及「公平與不歧視」原則，惟相關內容僅作原則性規定。在金融領域中，金管會頒布之「金融業運用 AI 指引」，旨在為金融機構導入、使用及管理 AI 提供行政指導，該指引除總則外，包含金融業於運用 AI 時所須遵循之六大核心原則[12]以及該等原則之目的、主要概念及落實方式等，謹就有關 AI 自動化決策引起之偏見及可解釋 AI 部分，說明如次：

1、重視公平性及以人為本的價值觀(核心原則二)：本項原則強調金融機構運用 AI 系統時，應儘可能避免演算法偏見所造成之不公平，並提升其自動化決策之合理性及準確性，注意 AI 偏見之產生並儘可能避免歧視問題。在公平性之內涵上，本指引指出此概念係指決策須具有合理性、準確性及儘可能避免歧視、以人為本、針對受到不利結果影響之消費者宜提供救濟選項，並針對人類在 AI 系統決策過程中之監督機制提出分類。

2、落實透明性與可解釋性（核心原則五）：金融機構運用 AI 與消費者互動時，宜向消費者適當揭露與其相關之資訊，對於自行或委託研發之 AI 系統，金融機構宜確認其人員必要時可清楚說明 AI 系統運作之邏輯。在透明性之內涵上，本指引闡明此係指提供外部利害關係人有關 AI 系統之相關資訊，以利其了解對其權益之影響等，以及該等 AI 系統的限制與風險；而所謂可解釋性，則係指清楚說明所使用之 AI 系統如何運作及其預測或決策過程背後之邏輯，以利組織內部評估稽核等。

---

[10] 楊岳平，同註 4，頁 45-47

[11] 人工智慧基本法第 4 條：「……五、透明與可解釋：人工智慧之產出應做適當資訊揭露或標記，以利評估可能風險，並瞭解對相關權益之影響，進而提升人工智慧可信任度。（第五款）六、公平與不歧視：人工智慧研發與應用過程中，應盡可能避免演算法產生偏差及歧視等風險，不應對特定群體造成歧視之結果。（第六款）……」

[12] 包含「建立治理及問責機制」、「重視公平性及以人為本的價值觀」、「保護隱私及客戶權益」、「確保系統穩健性與安全性」、「落實透明性與可解釋性」、「促進永續發展」等六大原則。

（三）結論與建議

AI 雖存在歧視問題，惟自動化決策所為之偏誤程度不一定會高於人為決策所存在之歧視或誤差，故完善之 AI 風險管理制度，將能在可承擔之風險程度下享受 AI 所帶來之便利。描繪 AI 風險管理制度雛形之首要步驟，即在釐清造成 AI 偏見之主要因素，該等因素之歸納需先於相關法規中明定須受保護之特定群體[13]，並據以訂定可解釋 AI 之系統性方法。

觀察我國人工智慧基本法雖有規定政府推動 AI 之研發與應用應遵循公平與不歧視原則，惟並無明定應受保護之特定群體，僅金管會於「金融業運用 AI 指引」核心原則二中，明定所謂受保護群體之類型，包含宗教、 種族、性別、身心障礙、性傾向、居住所、政治傾向、年齡、國籍或族群等。然該指引僅具行政指導性質，國家科學及技術委員會作為我國一般性 AI 法制之主管機關，宜進一步研析各產業趨勢及需求，適時考量於子法中進一步釐明受保護群體之類型，以促進透過可解釋 AI 減緩 AI 歧視之效。

撰稿人：陳樂庭

[13] 楊岳平，同註 4，頁 47-48。