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

作者文獻

楊岳平
臺灣大學法律學系副教授

摘要

人工智慧歧視議題向來為人工智慧法制研究中受到高度關注的議題，我國甫通過的人工智慧基本法與金融監督管理委員會頒布的金融業運用人工智慧指引亦規定有公平性與不歧視的原則，但人工智慧決策黑盒子的特性，向來被認為係人工智慧開發者與使用者控管人工智慧歧視的阻礙。

本文嘗試由分析人工智慧歧視問題的根源出發，檢視既有文獻提出的人工智慧歧視解決方案，進而聚焦討論如何應用可解釋人工智慧技術以增加人工智慧的可解釋性，進而提出以可解釋人工智慧為基礎的人工智慧歧視解決方案，並分析推動相關解決方案可能涉及的法律配套措施。

目次

壹、前言

在諸多人工智慧監理議題當中，人工智慧歧視由於涉及社會以及經濟平等，因此受到相當重視。類似的歧視疑慮，在人工智慧應用於金融領域特別是信貸服務時，也受到金融監理高度重視，例如美國即曾先後發生

DOI：10.53106/1025593137202
關鍵詞：人工智慧、可解釋人工智慧、金融歧視、公平性、人工智慧基本法、金融業運用人工智慧指引

* 本文部分內容曾於中華民國憲法學會 114 年度學術研討會發表，作者特別感謝主辦人林超駿教授、議程主持人董保城教授、與談人劉定基教授以及宮文祥教授，以及王立達教授、萬幼筠營運長及李宣緯教授於不同場合對本研究提出的建議，此外特別感謝臺灣大學法律學研究所廖子賢、吳方林及曾慶怡研究生的研究協助。

大都會人壽(MetLife)[1]與蘋果信用卡(Apple Card)[2]等涉及演算法歧視的信貸爭議案件，因此包括國際金融標準制訂組織以及我國金融監督管理委員會（下稱「金管會」）對於人工智慧應用於金融服務時可能涉及的歧視議題，均投以相當的關注[3]。

人工智慧歧視的問題來源之一，與人工智慧特殊的自主性特徵亦即決策黑盒子有關[4]。為因應人工智慧決策黑盒子的不透明與不可解釋特性，有研究認為必須透過一定程度的人力介入要求，以降低人工智慧歧視的風險[5]。但近年隨著可解釋人工智慧(Explainable AI, XAI)的興起與擴大應用，人工智慧歧視問題似有機會透過與可解釋人工智慧的搭配——例如實務上常應用的SHAP (SHapley Additive explanations)、LIME (Local Interpretable Model-agnostic Explanations)或反事實解釋(counterfactual explanation)——而獲得控制或緩解。

本文嘗試以人工智慧金融信貸歧視為例，分析人工智慧歧視的具體根源，並分析可解釋人工智慧可如何用以緩解人工智慧歧視問題。本文架構如下：第貳部分回顧人工智慧歧視相關規範，並著重於介紹金管會於2024年提出的金融業運用人工智慧指引（下稱「金融AI指引」）中的相關規定；第參部分分析人工智慧歧視的問題根源以及研究文獻已提出的解決方案；第肆部分分析可解釋人工智慧的內涵與功能，以及其可如何應用以管理人工智慧歧視風險；第伍部分再針對應用可解釋人工智慧管理人工智慧風險時所需的法制配套措施，提出本文的分析；第陸部分為本文的結論。隨著可解釋人工智慧的廣泛應用，人工智慧歧視的疑慮或可從「是否」可控管進展至「如何」設計制度以控管，期待透過本文的拋磚引玉，啟發更多關於可解釋人工智慧具體應用的制度討論。

須強調者為，人工智慧歧視的面向又可大別為傳統歧視與新型態歧視。傳統歧視著重於人工智慧應用可能對既有反歧視法規明定的受保護特定群體產生的歧視，例如種族、性別等；至於新型態歧視則進一步關注既有法制並未明定、但人工智慧長此發展可能對若干特定群體造成的系統性逐出效果[6]。本文礙於篇幅與智識限制，將僅針對傳統歧視問題進行討論，合先敘明。

貳、人工智慧歧視與現行法制問題

---

1 *See* Ahmad Swaiss, *Unveiling the Credit Rating Agencies: A Critical Legal Analysis for Reform*, 12(3) GLOBAL SCI. J. 507, 509-510 (2024).

2 *See* Jason Jia-Xi Wu, *Beyond Free Markets and Consumer Autonomy: Rethinking Consumer Financial Protection in the Age of Artificial Intelligence*, 13 NYU J. INTELLECTUAL PROPERTY & ENTERTAINMENT L. 56, 114-115 (2023).

3 參照：本文第貳部分之說明。

4 所謂人工智慧決策黑盒子，基本係指人類可能無法察知人工智慧系統的運作，也有可能無法理解演算法所形成之結論或決策。廖淑君，人工智慧與普惠金融——淺析演算法於徵信／授信應用之金融消費者保護議題，財金法學研究，5 卷 1 期，頁 120，2022 年。

5 參照：本文第參、二、(一)部分之說明。

6 相關討論，*see generally* Sandra Wachter, *The Theory of Artificial Immutability: Protecting Algorithmic Groups under Anti-Discrimination Law*, 97 Tul. L. Rev. 149, 191-202 (2022).

本部分將回顧既有人工智慧法制中與人工智慧歧視議題相關的內容，並著重介紹我國金管會金融AI指引的具體內容。

一、一般人工智慧法制

國際與國內已關注人工智慧法制議題相當時日。就國際層面而言，經濟合作暨發展組織(Organisation for Economic Co-operation and Development, OECD)於2019年通過的OECD人工智慧原則中，即有將人工智慧公平性列入其五大原則中的人權與民主價值原則中，強調人工智慧運用者應於人工智慧系統的所有週期均尊重法治、人權、民主及以人為本的價值，包括不歧視、平等、公平性等面向，為此人工智慧運用者應採取相關機制與防護措施，例如保留人力代理與監督的空間[7]。

我國於2026年1月公告的人工智慧基本法中，對於人工智慧歧視議題亦有著墨。其第4條第6款即規範人工智慧研發與應用過程中之公平與不歧視原則，亦即「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」。

二、金融監理與金融業運用人工智慧指引的規範內容

於金融監理領域，若干國際金融監理標準制定組織(standard-setting bodies, SSBs)亦高度關注人工智慧金融歧視的問題。例如OECD於其2021年的研究報告中，即曾提及人工智慧、機器學習及大數據技術應用於金融服務時，可能帶來偏見與歧視的問題[8]；又例如我國作為會員的國際證券管理機構組織(International Organization of Securities Commissions, IOSCO)亦曾於其2021年最終報告中，指出人工智慧與機器學習應用於金融服務時可能有資料品質與偏見疑慮[9]，

受到上述標準制定組織的影響，我國金管會於2024年頒布的金融AI指引，對於金融業如何管理人工智慧金融歧視風險，有更進一步的說明。該指引的核心原則二即強調金融機構應重視公平性，亦即金融機構在使用人工智慧系統之過程中，應儘可能避免演算法之偏見所造成的不公平。換言之，金融機構運用人工智慧系統產生之決策，不應對特定群體造成歧視之結果[10]。

該指引並進一步深化「公平性」的概念，指出此係指決策需有合理性，以及準確性及儘可能避免歧視。針對「合理性」，該指引指出此係指金融機構如利用個人屬性作為人工智慧模型決策之因素之一，應有合理理由；如無合理理由，運用人工智慧系統所產生之決策則不應對特定群體有系統性之不利差別待遇（例如不得以特定宗教、種族、性別、身心障礙、性傾向、居住所、政治傾

---

7 OECD, Recommendation of the Council on Artificial Intelligence, Section 1.2, OECD/LEGAL/0449 (May 22, 2019).

8 OECD, Artificial Intelligence, Machine Learning and Big Data in Finance: Opportunities, Challenges, and Implications for Policy Makers 40-42 (2021).

9 The Board of the IOSCO, The Use of Artificial Intelligence and Machine Learning by Market Intermediaries and Asset Managers: Final Report 10-11 (2021).

10 金管會，金融AI指引，頁9，2024年6月。