主題論文
© 臺大法學論叢 NTU Law Journal
第 52 卷特刊/Vol.52, Special Issue (11. 2023)

AI 可解釋性的法學意義及其實踐*

黃詩淳**

<摘要>

近期資訊科學所謂的「AI 的可解釋性（explainability）」有兩個內涵：其一是理解後說明的可解釋性（interpretability），包括主體中心的解釋與模型中心的解釋；其二是透明度（transparency），使用例如分解法或「模型不可知系統」（代理人模型等）之方法達成。另一方面，法學領域對 AI 的討論中，法規與司法裁判所稱的「要求解釋之權利」則是使用「explanation」一詞，但內涵為何、與資訊科學界的「可解釋性」是否相類，仍有相當爭論。本文認為，在需要較高程度的解釋時（例如公部門的自動化決策時），以透明度底下的方法所為之解釋，可能過度複雜難懂而對被影響之人沒有太大意義，也可能侵害模型製造者之營業秘密。法律毋寧應將重點放在 interpretability 底下的「主體中心」解釋與「模型中心」解釋二種方法，前者是提供主體關於與自己類似決定的人們的資訊，後者包括訓練資料的概述、模型種類、最重要因素及模型成效等，始符合 GDPR 第 15 條的「有意義資訊」。上述解釋不包括各因素的權重或原始程式碼。最後，針對未來可能出現的司法 AI，本文以法律資料分析之相關研究為例，說明法律資料的處理

* 感謝三位審查人對本文提供諸多寶貴意見，使作者受益甚多。本研究為國科會專題研究計畫「人工智慧的創新與規範：科學技術與人文社會科學的交互作用跨領域專案計畫」（MOST 108-2420-H-001-002-MY3）與教育部大專校院人文與社會科學領域標竿計畫（法律學）（NTU-112L9A006）計畫之研究成果。

** 國立臺灣大學法律學院教授。
Email: schhuang@ntu.edu.tw。

・投稿日：02/09/2023；接受刊登日：10/31/2023。
・責任校對：黃品樺、辛珮群、高映容。
・DOI:10.6199/NTULJ.202311/SP_52.0001

及演算過程與可解釋性之關係，裨利法官與律師等使用者適當行使「要求解釋之權利」。

關鍵詞：可解釋性、解釋權、模型中心的解釋、主體中心的解釋、法律資料分析、全域可解釋性、區域可解釋性