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

**♦目 次♦**

# 壹、前 言

2018年5月25日生效的「歐盟個人資料保護基本規則(European Union's General Data Protection Regulation，下稱GDPR）[1]」，明文承認被自動決策影響之個人有「要求解釋之權利」，但因缺少關於解釋形式之細部規定，故引發了法學與資訊科學界對如何解釋的廣泛議論[2]。「可解釋的 AI（Explainable AI，下稱 XAI）」也成為近期資訊科學的新興領域，美國國防部國防高等研究計劃署（Defense Advanced Research Projects Agency，下稱DARPA）在 2017 年 5 月開始了「可解釋的 AI」大型研究計畫，目標在建立機器學習（machine learning）的可解釋模型，並結合有效的解釋技術，使終端使用者理解、合理信賴並有效地管理 AI 系統產出的結果[3]。DARPA 倡議之具體方法有三:第一是建立更可解釋的模型;第二是設計解釋用的介面;第三是探索有效解釋的心理機制與要件[4]。另外，在某些法規或法律文件，也引入上述概念，提及了「提升可解釋性（explainability）」的要求或肯認「要求解釋之權利（right to explanation）」[5]。在法律或社會實踐中，解釋可能是問責的前階段必要步驟。例如 Beaudouin 等人主張，「提升粗資訊的可解釋性（explainability）後，能讓更多人可以理解，便能提升這個演算法的

---

1 Council Regulation 2016/679, 2016 O.J. (L119) (EU).

2 Ashley Deeks, *The Judicial Demand for Explainable Artificial Intelligence*, 119 COLUM. L. REV. 1829, 1833-34 (2019).

3 Will Knight, *The U.S. Military Wants Its Autonomous Machines to Explain Themselves*, MIT TECHNOLOGY REVIEW (Mar. 14, 2017), https://www.technologyreview.com/2017/03/14/243295/the-us-military-wants-its-autonomous-machines-to-explain-themselves/.

4 David Gunning & David W. Aha, *DARPA's Explainable Artificial Intelligence (XAI) Program*, 40 AI MAG. 44, 45 (2019); David Gunning, Eric Vorm, Jennifer Yunyan Wang & Matt Turek, *DARPA's Explainable AI (XAI) Program: A Retrospective*, 2 APPL. AI LETT. 1, 2 (2021).

5 Right to explanation亦簡稱為「解釋權」，參照：劉靜怡（2019），〈淺談GDPR的國際衝擊及其可能因應之道〉，《月旦法學雜誌》，286期，頁16-19。

透明性（transparency）；透明性提升可追蹤性（traceability），又提升可稽核性（auditability/ evaluability），最後再提升可問責性（accountability）。[6]」。因此，可解釋性在司法裁判場域，將延伸成為可問責性的問題，參與者能要求誰提出什麼樣的解釋，將會是主要關心點。

惟如何落實可解釋性，恐非易事。一般人直觀中的「解釋」，可能是要求 AI 開發者公開其原始程式碼，但可能侵害其營業秘密等財產上權益；此外，這種「解釋」方法對一般使用者去理解例如「銀行為何拒絕貸款」、「自駕車為何決定在此刻緊急煞車」等事項並無太大幫助。因此，如果「提升可解釋性」或「要求解釋之權利」是法規範的目標，則具體應該如何法制化，需要細緻的思考。首先可能要釐清，「AI 的可解釋性」及「可解釋的 AI」實際內涵為何？上述概念與法規範肯認「要求解釋之權利」或「透明性」關係為何？最後，「可解釋的 AI」是否或哪方面能滿足法學界要求的「可解釋性」或「要求解釋之權利」？這需要立法者與審判者具有跨領域的能力與經驗，特別是具有法學與資訊科學領域知識。

本文將整理對「可解釋（explainability）」及「解釋（explanation）」的討論，包含資訊科學與法學領域之文獻對此一概念的看法，在相互比較與對照後，提出可規範於法規中的「可解釋的 AI」的定義。以上是一般性的討論，未來在司法領域也可能出現 AI，通常係使用法律資料分析的手法製造。為讓使用者辨別其是否已盡了解釋義務，以及能要求到何種程度的解釋，本文將以法律資料分析之相關研究為例，說明具體實踐步驟，裨利公眾監督、確認是否符合「可解釋的 AI」之標準。

---

6 Valérie Beaudouin et al., *Flexible and Context-Specific AI Explainability: A Multidisciplinary Approach*, ARXIV (Mar. 13, 2020), https://arxiv.org/pdf/2003.07703.pdf. 本文引用此見解，是贊成「可解釋性（的提升）有助於之後的問責」；但本文並不同意將transparency放在explainability之上，相反地，本文較贊成後述貳之一Waltl & Vogl的概念分類，transparency應是explainability的下位概念。

# 貳、可解釋性（Explainability）的意義

「可解釋（explainability）」及「解釋（explanation）」此二詞彙，並非傳統法學使用的概念，以往甚少在法規中出現。法學提到「解釋」亦即「法律解釋」（legal interpretation），使用的是 interpretation 一詞[7]，內容是吾人熟知的文義解釋（textualism）、目的解釋（purposivism）等[8]。但後述 GDPR 規定，資料主體有權獲得自動決策之「解釋」（obtain an explanation of the decision），使用的卻是 explanation 一詞。為何具有法律拘束力的 GDPR 不採用法學向來熟悉的 interpretation 一詞，而是 explanation？是否受到資訊科學的影響？本節將先深入討論 interpretability 與 explainability 二詞彙在資訊科學界的演化與轉變，第「參」部分再分析法學界所期待的「解釋權」內涵。

## 一、Explainability 的概念

大部分的日常情境當中，例如一般的事物關聯性解釋、法律解釋、甚至較為初階的量化模型（迴歸、相關）中，interpretability 與 explainability 常混用。詳細說明的話，interpretation 是指某件系統的運作及其內在關係能被人類所懂的方式來描述，比如說法學解釋或是某個模型（氣溫與冰淇淋銷量的相關性）；這些可能都預設隱含了某種內在邏輯，而解釋者也瞭解這些內在邏輯。解釋的目標可以使讀者得到與其他知識的聯繫，所以某種更為好的解釋可以更多全面的解釋其他現象與其他解釋。把上述的一系列思考方式名詞化，並且預期可以「比較」，則稱為 interpretability。總體而言，interpretation 強調的是「研究者基於充分理解後說明」。

---

[7] *E.g.*, Oliver W. Holmes, *The Theory of Legal Interpretation*, 12 HARV. L. REV. 417, 417-20 (1899).

[8] Mark Greenberg, *Legal Interpretation*, THE STANFORD ENCYCLOPEDIA OF PHILOSOPHY (July 7, 2021), https://plato.stanford.edu/archives/fall2021/entries/legal-interpretation/.

另一方面，explanation 是解釋者不必全部瞭解被解釋物內在邏輯，可以用某種知識體系來表現／重現（representation）其關係，這樣的理解可能真的命中其系統的內在邏輯，也可能是解釋者自己來重新規定[9]。因此 explanation 的解釋方式，同時也會需要說明一個「解釋生產體系」（explanation producing system）用來決定這樣的解釋是否更為正確，所以把這種思考加以名詞化就是 explainability。

傳統數學計算例如貝葉斯機率，可以確定演算過程，且可被複製或重現，所以人們比較容易「理解」此些算法之內在邏輯，當然也容易「說明」某個輸入值為何可以得到某個輸出值，這就是 interpretable。但機器學習當中尤其深度學習（deep learning）的進展，使得人們不見得完全瞭解深度學習模型，因此需要區分上述兩種「可解釋性」。機器學習是藉由定義學習函數（learning function）與損失函數（loss function），再藉由計算機的演算能力處理高維度資料來達到變數收斂（convergence of variables）的模型。在深度學習領域，大量的算式與複雜權重，即使製造者也未必人人能真正從頭到尾「理解」模型中的所有數值，這是人工智慧被稱為「黑盒子」（black box）[10]的原因。

雖無法被人類完全理解，但這些人工智慧模型確實起作用，也已經被用在實務上，比如 AI 圍棋等軟體，改變了人類對棋藝的認識，甚至改變了「職業棋手」的工作方式。於是，有必要對於 interpretability 與 explainability 的概念仔細區別。Interpretability 意指「以人類理解詞彙來解釋或表現」的程度[11]，或者是「人類能理解某個決策的原因」的程度[12]，這都需要對於人工智慧模型本身完全理解。比如說，具有 interpretability 的模型，人們可以觀察並

---

[9] Leilani H. Gilpin et al., *Explaining Explanations: An Overview of Interpretability of Machine Learning*, ARXIV (Feb. 3, 2019), https://arxiv.org/pdf/1806.00069v3.pdf.

[10] FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION 3-4 (2015).

[11] Finale Doshi-Velez & Been Kim, *Towards a Rigorous Science of Interpretable Machine Learning*, ARXIV (Mar. 2, 2017), https://arxiv.org/pdf/1702.08608.pdf.

[12] Tim Miller, *Explanation in Artificial Intelligence: Insights from the Social Sciences*, 267 ARTIF. INTELL. 1, 8 (2019).

說明影響決策之重要因素[13]，例如在影像辨識任務中，若有某個因素讓模型認為某個物體是貓或狗的一部分，該物體可能就是該模型之關鍵要因。這樣模型就是具有 interpretability，即可解釋輸入項與輸出項的關聯[14]。相對地，explainability 則不要求解釋者瞭解所有內在邏輯，只要求給出一個合理的說明，如機器學習系統之建置機制[15]。如前所述，在深度學習方法中幾乎無法做到完全之 interpretability，從而輔以「透明度（transparency）」來提升人類對模型的信賴。

以上 interpretability 與 explainability 的概念區別[16]，資訊科學者呈現如下圖一。explainability 包含了 interpretability 以及 transparency 兩部分。Interpretability 指的是以人類能理解的方式來描述機器做出的決策，亦即輸入項與輸出項的關聯。但有些模型仍「無法以人類理解的方式描述」，只能用提高「人類模仿性（simulatability）[17]」、「可分解性（decomposability）[18]」 或「演算法透明性（algorithmic transparency）」來提升整體的透明性，達成 explainability。資訊科學界這樣的概念形塑有其用意，因 explainability

---

[13] Amina Adadi & Mohammed Berrada, *Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)*, 6 IEEE ACCESS 52138, 52138-60 (2018).

[14] Pantelis Linardatos et al., *Explainable AI: A Review of Machine Learning Interpretability Methods*, 23 ENTROPY 1, 20-21 (2020), https://doi.org/10.3390/e23010018.

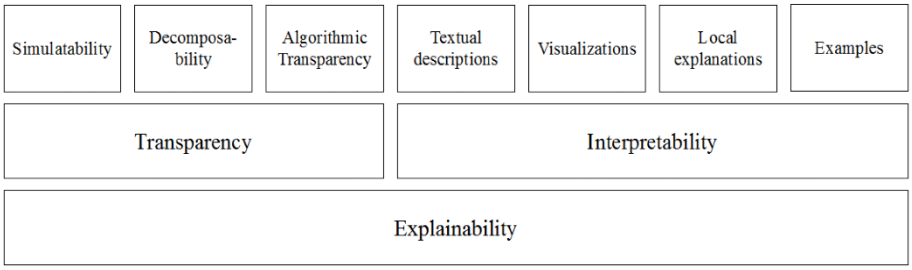
[15] *Id.* at 3.

[16] Bernhard Waltl & Roland Vogl, *Explainable Artificial Intelligence: the New Frontier in Legal Informatics*, 4 JUSLETTER IT 1, 6 (2018); 其概念來自Zachary C. Lipton, *The Mythos of Model Interpretability*, ARXIV (Mar. 6, 2017), https://arxiv.org/pdf/1606.03490.pdf.

[17] Simulatability通常翻譯為「人類模仿性」，亦即如果人類可以將同樣的數據和模型參數，在合理時間內計算並作出預測的話，那麼該模型就具備了模仿性，*See* Lipton, *supra* note 16, at 4-5. 實際上有論者以LIME或Anchor等算法來測量模型的人類模仿性，*See* Peter Hase & Mohit Bansal, *Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?*, ARXIV (May 4, 2020), https://arxiv.org/pdf/2005.01831.pdf.

[18] Decomposability本文翻譯為「可分解性」，指的是模型的每個部分，包含輸入項、係數與計算，都能有直覺性的解釋，例如決策樹模型中的每個節點，會對應到相應的簡單描述（例如病人的血壓超過150者），*See* Lipton, *supra* note 16, at 5.

作為上位概念可以為 XAI 的研究帶來益處。亦即資訊科學界不論是發展強化透明性的方法（例如將模型拆解或如後述做一個新模型解釋舊模型），抑或開發增進 interpretability 的方法（例如加強區域性解釋亦即後述的主體中心的方法），都是 XAI 的一環，也是科學界可努力的方向。惟這樣的概念未必等同於法學所要求的「解釋」，詳如後述。




【圖一】可解釋性之概念示意圖

※ 資料來源：Bernhard Waltl & Roland Vogl, *Explainable Artificial Intelligence: the New Frontier in Legal Informatics*, 4 JUSLETTER IT 6 (2018).

## 二、Explainability 的實例

實踐 explainability 的方法本身就是資訊科學界討論的焦點，有許多不同的分類[19]。本文主要參考 Edwards & Veale 的分類，說明如下。第一種解釋作法是允許使用者互動式地探索演算法系統，這有助開發出良好且可信賴的系統心智模型；第二種解釋作法則根基於另一思維，亦即不必瞭解模型的內

[19] 舉例言之，Linardatos et al., *supra* note 14, at 5,雖然與Edwards同樣採取global與local的分類，但在處理「模型」本身（亦即Edwards所說的第二種作法）則採model agnostic與model specific的分類法，同樣的想法亦見於Alejandro Barredo Arrieta et al., *Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI*, 58 INFO. FUSION 82, 94 (2020); A. Saranya & R. Sabhashini, *A Systematic Review of Explainable Artificial Intelligence Models and Applications: Recent Developments and Future Trends*, 7 DECISION ANALYTICS J. 1, 9 (2023).

部結構，而是設計一個更簡單的模型作為解釋工具[20]。Edwards & Veale 所謂的第一種作法比較接近於上述 Waltl & Vogl 的「可解釋性」體系（即上圖一）中的 interpretability，第二種作法則近於「透明度」（transparency）的部分。

首先，Edwards & Veale 提倡的第一種作法是從解釋的重心（the focus of explanation）亦即提升使用者的信賴感出發，底下又可分為兩種解釋方式，即「模型中心（model-centric）」或「主體中心（subject-centric）」兩種模式[21]。前者「模型中心」通常稱為「全域可解釋性（global interpretability）」[22]，包括解釋製造者意圖、系統使用的模型的屬性、在訓練之前的參數、訓練用的輸入資料的質性描述、模型在新數據的表現等[23]。模型中心的方法試圖解釋整個模型，而非對特定案件的結果說明，目的是確保決策是在正常過程的狀態下被作成。至於「主體中心」模式，通常稱為「區域可解釋性（local interpretability）」[24]，則是提供主體關於獲得與自己類似決定的人們的資訊[25]，或者用「反事實解釋（counterfactuals）」方法，也就是若想找出影響模型決策的重要因素，則給予不同的輸入項，測試出影響原先決策的因素有哪些[26]。主體中心的解釋不適於討論模型的運作程序是否正常，故對設計人機溝通介面或研究人機互動者實益較大，而對於例如建造安全加密算法的工程師而言意義較小。從使用者或被自動決策影響而尋求救濟者的角度而言，主體中心的解釋確實較有意義。不過，解釋的品質可能會因決策系統的高維度

---

[20] Lilian Edwards & Michael Veale, *Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking for*, 16 DUKE L. & TECH. REV. 18, 61 (2017).

[21] *Id.* at 55-59.

[22] Dipanjan Sarkar, *The Importance of Human Interpretable Machine Learning*, MEDIUM (May 25, 2018), https://towardsdatascience.com/human-interpretable-machine-learning-part-1-the-needand-importance-of-model-interpretation-2ed758f5f476

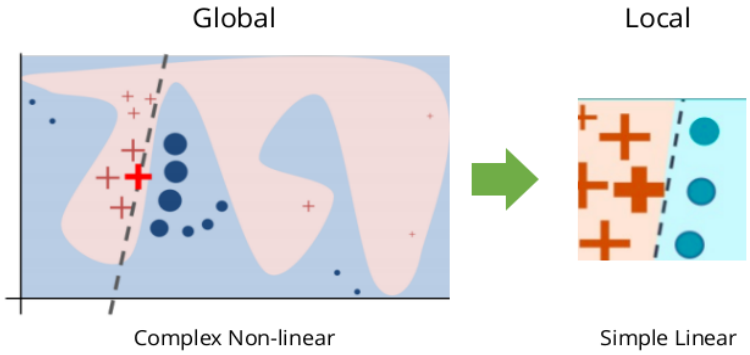
[23] Edwards & Veale, *supra* note 20, at 55-56.

[24] Sarkar, *supra* note 22.

[25] Edwards & Veale, *supra* note 20, at 58.

[26] Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 28-29 (2014).

性質以及要求解釋的個人類型等因素而有所降低[27]。上述兩種解釋的區別可大致以下圖二所示。




【圖二】全域解釋與區域解釋之區別

※ 資料來源：Manu Joseph, *Interpretability: Cracking open the black box: Part III*, DEEP AND SHALLOW (Nov. 24, 2019) , https://deep-and-shallow.com/2019/11/24/interpretability-cracking-open-the-black-box-part-iii/

當個人想了解「我要如何改變才能獲得不同結果」時，主體中心的方法較為有用，也較能幫助個人得知決策的原因並挑戰其結果[28]。亦即對使用者來說，主體中心的解釋方式似乎更具吸引力、便利性和說服力。一些主體中心的解釋作法，已允許個體在其自身資料點周圍假設性地探索正在發生的決策邏輯，例如可授權某個工具對自己信用檔案進行試探查詢，或者某日在使用者的允許下讓這個工具使用其社交媒體 API 來進行測試。更進步的方法可能是允許資料主體看到系統如何作出有關其他使用者的決策，從而使用戶擺脫自己的同溫層搜尋結果[29]。使用主體中心的解釋法，其優點是不需要揭露完全模型系統機制，卻仍然可以使用模型[30]。

---

[27] Edwards & Veale, *supra* note 20, at 22.

[28] Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085, 1120 (2018).

[29] Edwards & Veale, *supra* note 20, at 61-63.

[30] Sandra Wachter et al., *Counterfactual Explanations Without Opening the Black Box:*

上述第一種作法著重的是提供模型整體的描述或個體可能的結果，提升使用者信賴；相較之下，第二種作法則是直接觸碰模型本身，底下又可以分成兩種方式。第一是「分解法」或「解耦法」（decompositional explanation），亦即「打開黑盒子」，理解其中的結構，如權重、神經元、決策樹和架構等。某些類型的機器學習如迴歸，在設計上先天便是可分解的。其他類型的機器學習也可以透過簡單的方法使之變得可分解，例如，隨機森林模型可以訓練在模型之外生成「變數的重要性分數」。至於深度學習系統那樣複雜的模型，做分解就需要額外的方法，也成為熱門的研究領域[31]。由於此方法需要接觸模型的大部分資料與構造，程式碼的公開似乎不可避免，但有論者認為公開程式碼並非令人完全滿意的解釋方法[32]。

第二種作法稱為「教學的解釋（pedagogical explanations）」或「模型不可知的系統（model agnostic systems）」，不試圖說明模型系統機制，而是以外在的方法將模型的運作方式提供資訊給使用者[33]。比如「代理人模型（surrogate model）[34]」的作法，會分析輸入項與輸出項的配對狀況，但並不真正知悉原來模型本身的內部權重[35]。例如原本用了一個黑盒子模型來預測病人得糖尿病的機率，而我們可以再建立一個決策樹模型，有效地反映了黑盒子的決策，使人們理解黑盒子模型在作風險評估時，哪個因素（例如膽固醇值、對尼古丁的依賴程度、水腫程度等）被賦予多少權重[36]。模型不可知

---

*Automated Decisions and the GDPR*, 31 HARV. J. L. & TECH. 841, 851 (2018).

31 Edwards & Veale, *supra* note 20, at 64.

32 Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 638-39 (2017). 認為公開原始程式碼並非提升演算法可問責性（accountability）的妥適方法。

33 Edwards & Veale, *supra* note 20, at 65; Deeks, *supra* note 2, at 1835.

34 W. Andrew Pruett & Robert L. Hester, *The Creation of Surrogate Models for Fast Estimation of Complex Model Outcomes*, PLOS ONE (June 3, 2016), https://doi.org/10.1371/journal.pone.0156574.

35 Marco Tulio Ribeiro et al., *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*, ARXIV (Aug. 9, 2016), https://arxiv.org/pdf/1602.04938.pdf.

36 Osbert Bastani et al., *Interpreting Blackbox Models via Model Extraction*, ARXIV (Jan. 24, 2019), https://arxiv.org/pdf/1705.08504.pdf.

系統的優點在於，它接觸模型本身的需求相當低，因此不會妨害開發者例如公司的智慧財產權或營業秘密[37]。

不過，「做一個新模型解釋原黑盒子模型」的作法，是否可以直接該當法律意義的「解釋」，可能有疑問。舉例言之，擾動法（perturbation method）是將一個簡單模型放入擾動樣本並觀察；這個新的「模型」其實並非完全揭露原來黑盒子演算法，只是去模仿黑盒子產生反應，然後「簡化說明」黑盒子模型是基於某些判斷然後做出某結果[38]。解釋用的新模型本質上無法百分之百忠實地還原原來的模型，否則使用新模型即可，因為新模型更容易解釋，何必需要原來的模型？依照 Rudin 的看法，著名的 ProPublica 指摘 Correctional Offender Management Profiling for Alternative Sanctions（下稱 COMPAS）系統「有種族歧視」，就是一個「使用簡化模型解釋 COMPAS 模型」的不當解釋案例，亦即 ProPublica 製造了一個具有「種族」變項的線型模型，指摘 COMPAS 模型依照種族、年齡、犯罪史做出決策。COMPAS 很可能是非線性模型，COMPAS 也很有可能完全沒有依照種族（雖然可能與年齡、犯罪史有關聯）[39]。讀者若對照上圖二，應可了解 Rudin 的意思，圖二中的黑色虛線，就是個簡化的新模型，而這個新模型僅捕捉到了原模型的一角之特徵，也就是新模型無法反映原模型的全貌。所以，雖然「簡化模型」的作法在資訊科學研究上有其意義，但是否能使用在具體法學研究甚至法律實務上，需要謹慎為之。

---

37 Edwards & Veale, *supra* note 20, at 65.

38 Ruth Fong & Andrea Vedaldi, *Interpretable Explanations of Black Boxes by Meaningful Perturbation*, 2017 IEEE INT'L CONF. ON COMPUTER VISION (ICCV) 3449-57 (2017).

39 Cynthia Rudin, *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*, ARXIV (Sept. 22, 2019), https://arxiv.org/pdf/1811.10154.pdf.需留意的是，雖然Rudin提出了此點批判，但同文後續部分提及，作者認為COMPAS應該不是採用標準化的機器學習技術，而是依照問卷及專家意見並由專家設計的模型，也似乎沒有過重地看待犯罪史。既然COMPAS在模型上沒有任何理解難度，則它之所以被批評為黑盒子，主要是因為營業秘密而不公開權重，而非演算法的困難。由此可知，黑盒子的問題不僅存在於技術層面（複雜的機器學習模型），而是非機器學習也會有。

## 三、小結

目前資訊科學社群之主要研究方向，是更大範圍的 explainable AI 而不僅是 interpretable AI[40]。Explainable AI 的發展包含著兩部分：interpretability 與 transparency，後者的解釋方法未必需要全盤瞭解被解釋物內在體系的邏輯，例如代理人模型如何「解釋」、哪個「解釋」更好，本身就是一個新興研究領域[41]。在人工智慧領域，製造一個新解釋，可能比原來建立模型本身更為困難，本身就是一個新的演算課題[42]。有學者主張根本不要試圖使用 explainable AI，而是用既有 interpretable 的演算法來做決策[43]。不過僅依賴或強調 interpretability 卻放棄 explainability 底下的另一塊 transparency 的發展，並不妥當，這是因為 interpretability 取決於人自身的知識與理解程度，而可能導致過度簡化資訊[44]。先不論讀者支持上述哪一個立場，透過以上的討論，都更能體會 explainable AI 與 interpretable AI 的語境與細節差異。

# 參、可解釋性相關的法規、法律文件與司法裁判

GDPR 似乎承認被自動決策影響之個人有「要求解釋之權利」，此處的「解釋」雖然使用 explanation 一詞，然法規範所要求的解釋以及判斷某個東西是否具備「可解釋性」，是否與上述資訊科學界主流 explainability 意義完全相同？資訊科學提出的上述「可解釋的 AI」之各種作法，是否能滿足

---

[40] Arrieta et al., *supra* note 19, at 83.指出，自2012年以來一直都有討論interpretable artificial intelligence一詞的學術論文且逐年增長；2017年之後才出現較多使用explainable artificial intelligence一詞的學術論文，而XAI一詞則是自2018年有爆發性的成長且在至2019年超越interpretable artificial intelligence。

[41] Grégoire Montavon et al., *Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition*, 65 PATTERN RECOGNITION 211, 218-220 (2017).

[42] Adadi & Berrada, *supra* note 13, at 52138-60.

[43] Rudin, *supra* note 39, at 1.

[44] Bernease Herman, *The Promise And Peril Of Human Evaluation For Model Interpretability*, ARXIV (Nov. 20, 2017), https://arxiv.org/abs/1711.07414.

法規範期待的「解釋」？仍有疑問。本節將簡要整理，哪些法規、法律文件或判決中提及「可解釋性」或「解釋」，以及內涵為何，最後提出本文對上述問題的看法。

## 一、法規

首先，目前學界討論最多的應是 GDPR 中的「解釋權」。惟 GDPR 自身並未使用「解釋權」一詞，而是由下列規定推導而出。首先，GDPR 第 22 條[45]，主要規範了「自動化作成的決策」，其結果對特定個人有法律效果或類似效應時，該個人有不受該決定拘束的權利。同條第 2 項規定了第 1 項的例外情形，例如當契約約定或者取得資料主體同意時，個人應受該決定之拘束；但同條第 3 項規定，此際資料控制者有義務讓資料主體有陳述意見及質疑資料控管者決定的機會。其次，第 15 條第 1 項 (h) 款呼應了第 22 條第 3 項，亦即，於第 22 條第 1 項及第 4 項所定自動決策存在時，資料主體有權要求資料控管者告知所涉及的邏輯性有意義資訊，以及重要性與預設結果。

---

[45] GDPR第22條之條文如下：
第22條（個人化之自動決策，包括建檔）
1. 資料主體應有權不受僅基於自動化處理（包括建檔）所做成而對其產生法律效果或類似之重大影響之決策所拘束。
2. 第一項規定不予適用，如該決策：
(a) 係為締結或履行資料主體與控管者間之契約所必要者；
(b) 係控管者受拘束之歐盟法或會員國法有明文授權，且定有適當之保護措施以確保資料主體之權利及自由及正當利益者；或
(c) 係基於資料主體之明確同意者。
3. 在第2項所定第a點及第c點之情形，資料控管者應執行適當保護措施以確保資料主體之權利及自由及正當利益，至少有權對控管者部分為人為參與、表達意見以及挑戰該決策。
4. 除第9條第2項第a點或第g點所定情形外，第2項所定決策不得係基於第9條第1項所定之特殊類型之個人資料，且應實施適當保護措施以確保資料主體之權利及自由及正當利益。
中譯版內容參見：財團法人金融聯合徵信中心（2017），《歐盟個人資料保護規則》，頁 202-203，https://www.jcic.org.tw/main_member/fileRename.aspx?uid=1566&fid=1103&kid=2（最後瀏覽日：10/20/2023）。

那麼，此處所謂的「資料主體有權要求被告知……資訊、重要性與預設結果」，能否認為就是「要求解釋之權利」亦即「解釋權」呢？

採肯定看法者認為，GDPR 前言（Recital）第 71 段指出：「在任何情況下，該處理應有適當之保護措施，此應包括將特定資訊給予資料主體及獲得人為干預、表達意見、獲得依上開評估後做成決策之解釋（to obtain an explanation of the decision reached after such assessment），以及挑戰該決策之權利。」此段話即明確承認資料主體得要求資料控管者解釋如何透過演算法做成決策。雖然前言僅是闡明條文的解釋，本身並非具有法律效力之規定，但前言的內容結合條文本身，亦即 GDPR 第 13 條、第 14 條規定了資料控管者於取得個人資料時，有義務提供資料主體相關資訊；第 15 條規定了資料主體的資訊接近使用權（right to access information），故可擴張解釋第 22 條第 3 項所謂的「保障資料主體提出質疑之機會」，而成為資料主體的「要求解釋之權利」即「解釋權」[46]。不過即使肯定論者亦有指出目前的規範方式不盡妥當，畢竟 GDPR 前言本身不具有法律效力，有拘束力的第 22 條又未明白表述該權利，造成概念與效果混淆不清[47]。

否定論者則指出，GDPR 沿襲的「個人資料保護指令」並未承認資料主體享有解釋權。此外，GDPR 第 13 條第 2 項 (f) 款、第 14 條第 2 項 (g) 款係規定，資料控管者在進行第 22 條定義的自動化決策時，應告知資料主體相關資訊，因此告知義務係在「進行決策前」。資料控管者不太可能在決策前就提供「如何做成決策的解釋」。此外，第 15 條的資訊近用權，雖沒有告知義務般的時間限制，但因本條有「預設結果」（envisaged consequences）之用語，意味著這是將來發生的結果，故資料控制者還是只需在「進行決策

---

[46] Bryce Goodman & Seth Flaxman, *European Union Regulations on Algorithmic Decision Making and a "Right to Explanation"*, 38 AI MAG. 50, 55-56 (2017).

[47] 劉靜怡，前揭註5，頁18，認為就解釋權而言，GDPR有「文字精確度不足」和「未能明確定義權利與防衛機制」的缺陷，因此未來僅限於符合第22條第1項「影響資料主體的法律效果或相似重大效果」且「純然自動處理」，方有解釋權的行使空間可言；鄭伊廷（2021），〈試析「一般資料保護規則」下自動化決策的解釋權爭議〉，《經貿法訊》，279期，頁23，認為GDPR立法文字的模糊可能是立法者有意為之，後續的官方文件則似乎傾向保護資料主體亦即肯認解釋權之立場。

前」提供資訊即可。若係如此，上述肯定論者所引用的條文，並無法推導出「解釋權」。否定論者並認為第 22 條第 3 項雖保障資料主體提出質疑之機會，但賦予解釋權未必有利於該機會之保護；再加上過度擴大「解釋權」的肯認，可能妨害人工智慧等技術之發展，故不應肯定 GDPR 已明文肯定「解釋權」[48]。

其次，法國在 2016 年通過了「數位共和國法案」（Digital Republic Act），較 GDPR 更進一步，承認個人享有要求行政機關的演算法決策給予解釋之權利，亦即要求行政決策者提供關於「演算法作成決策的程度與模式」，「何種資料被處理及其來源」、「適用到個人之情形，其係數為何，若合適時，因素權重為何」的資訊[49]。有些政府部門例如國家安全或國防則被排除適用。論者認為，法國法的此一規定例如公開因素權重此點，似乎意味著「解釋」必須是針對特定決策（即前述的「主體中心」的解釋），而非針對複雜模型的模糊概述（模型中心的解釋）[50]。這個實踐，立法者具體考量了法益與技術的平衡，是值得參考的方式。

## 二、法律文件

上述兩個法規所要求的是「解釋（explanation）」，並非「可解釋性」。「可解釋性（explainability）」，目前僅在不具拘束力的法律文件可以見到。第一份是 2019 年 5 月 22 日通過「OECD AI 原則」[51]（OECD's Recommendation

---

48 Sandra Wachter et al., *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, 7 INT. DATA PRIV. L. 76, 80-83 (2017).

49 Edwards & Veale, *supra* note 20, at 48-49.

50 Edwards & Veale, *supra* note 20, at 49.

51 此翻譯參照財團法人中華民國國家資訊基本建設產業發展協進會（2020），《國家通訊傳播委員會「推動我國網路治理發展與國際趨勢研析」委託研究計畫期末報告》，頁 162-163，https://www.ncc.gov.tw/chinese/files/20021/5138_42718_200215_1.pdf（最後瀏覽日：10/19/2023）。

of the Council on Artificial Intelligence）[52]，在「可信賴 AI 的盡責監管原則（principles for responsible stewardship of trustworthy AI）」的部分提示了五大原則：包容性成長、永續發展與福祉；以人為本的價值與公平性；透明度與可解釋性；穩健、安全及保全；問責機制。其中第 3 項原則：「透明度與可解釋性（transparency and explainability）」指出，AI 行動者應努力使 AI 系統透明（transparency）並負責地揭露（responsible disclosure regarding AI systems），例如提供有意義之資訊，合乎脈絡且符合情境，包括促進人類對 AI 系統概括的了解，讓人類意識到是和機器互動，以及讓人類了解 AI 作用的結果，並能於受到不利影響時，對因素、預測或決策之機制有明白且容易了解的資訊，得以提出質疑。

第二份文件則是 2020 年 2 月 19 日歐盟執行委員會公布的《人工智慧白皮書》[53]，指出未來將以「監管」與「投資」兩者並重，促進人工智慧之應用並同時解決該項技術帶來之風險[54]。在第 3 點「搶佔先機：下一波資料潮」提及，「歐洲會繼續引領 AI 的演算法基礎的進步，且有必要橋接（bridge）至今為止分開作業的專業領域，例如機器學習、深度學習（有限的可解釋性 interpretability，需要大量的訓練資料，經由相關來學習）及符號主義人工智慧[55]（symbolic approach，意指透過人為干預來設定規則）。結合符號性推論與深度神經網路有助改善 AI 結果的可解釋性（explainability）」。

---

[52] OECD LEGAL INSTRUMENTS, RECOMMENDATION OF THE COUNCIL ON ARTIFICIAL INTELLIGENCE (OECD/LEGAL/0449) (May 22, 2019), https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

[53] European Commission, *White Paper on Artificial Intelligence: A European approach to excellence and trus*t, COM (2020) 65 final (Feb. 19, 2020), https://commission.europa.eu/system/files/2020-02/commission-white-paper-artificial-intelligence-feb2020_en.pdf

[54] AI白皮書的背景介紹參見：魏世和（2020），〈歐盟數位發展之形塑：以人工智慧與資料政策為中心〉，《經貿法訊》，267期，頁19。

[55] 符號主義人工智慧（symbolic AI）又稱老派人工智慧（good old-fashioned AI，簡稱GOFAI），1980年代的專家系統（expert system）為其代表。

## 三、司法裁判

與演算法模型的決策及其「解釋」較相關的司法裁判，較著名的例子應是 2016 年美國的 *State of Wisconsin v. Loomis*[56]案。一審法官在對被告 Loomis 量刑時，使用 COMPAS 系統，對 Loomis 的再犯風險的三大項目皆評為高風險，故判處 Loomis 有期徒刑 6 年。Loomis 主張法院量刑時所使用的 COMPAS 的風險評估結果並不準確，而認為侵害到憲法所保障的正當程序，故提起上訴。威斯康辛州最高法院認為，風險評估結果完全基於被告對於問題的回答，或公開的犯罪史的資訊，被告有機會驗證報告中列出的問題與答案是否精確，因此符合正當程序；此外，關於 COMPAS 系統的預測準確性，法院指出「在使用 COMPAS 系統的紐約州也曾對 COMPAS 是否具有充分的正確性進行檢測過，而依紐約州刑事司法服務局針對 COMPAS 的再犯性基準的有效性以及預測精準度所做的檢測結果，顯示可確保其再犯性基準的有效性以及預測精準度皆可達成充分的準確性」[57]。該案件之後上訴至聯邦最高法院，然最終法院並未受理。

其次，2017 年則有德州教師考績案[58]。德州的休士頓獨立學區（Houston Independent School District, HISD）[59]使用演算法[60]為教師打考績，並依此決定解僱表現不佳的教師。其中 9 位教師與地方教師工會起訴主張，學區的作法侵害了憲法所提供的程序正當保障，因他們無法取得演算法以及驗證考績準確性的相關資料。德州聯邦地方法院並未質疑營業秘密的保護，學區主張「正當程序保障並不能要求公司公開其營業秘密」被法院所肯認；法院認為

---

56 State of Wisconsin v. Loomis, 881 N.W.2d 749 (Wis. 2016).

57 此案之中文的詳細介紹與評析，參見：鄭明政（2020），〈從State v. Loomis案件看AI應用於司法審判上的若干問題〉，《台日法政研究》，4期，頁165-178。本文不深入探討。

58 Hous. Fed'n of Teachers, Local 2415 v. Hous. Indep. Sch. Dist., 251 F. Supp. 3d 1168 (S.D. Tex. 2017).

59 此為德州最大的學區，在2009年至2010年HISD有298個學校和202,773名學生。

60 此為外部廠商（SAS Institute）所研發之演算法，從判決中無法確定是否為機器學習，但可知廠商認為演算法與軟體均為商業機密，拒絕向原告（教師）及被告（學區）透露。

重要的是，與僱用相關的決策若是基於秘密的演算法，究竟是否符合最低限度的正當程序，若否，則救濟方式是要求學區必須基於其他（非秘密演算法的）方式作出考績，而不是否定公司的營業秘密[61]。結論上，法院認為必須使被評分的教師本人有方法驗證自己被評的分數的正確性，並且對這個分數的結果有提出異議的機會。最終本件原告的主張並未被駁回，法院肯認其動議（motion for summary judgment），但最終兩造和解，並未作成判決[62]。

另外，荷蘭政府使用演算法預測社會安全給付詐欺申請的可能性[63]，2020 年海牙地方法院以 European Union 的比例原則檢視演算法，認為電腦產生的風險報告欠缺解釋，導致個人無從對該報告提出異議，也使得法院無法確定是否沒有歧視[64]。

## 四、小結與本文見解

綜上，在法規與司法裁判中要求提供演算法相關資訊時，往往是使用「解釋」一詞，也有可能不直接使用「解釋」，而是使用鄰近的概念例如資訊近用、公開、正當程序保障等。但法規例如 GDPR 的規定究竟能否導出「解釋權」、其內涵為何，仍有相當爭論。至於司法裁判，以美國的 2 個案件為例，使用演算法所做的決策，是否違反正當程序保障，似乎是視被決定者有無機會驗證其「分數」而定；法院並未直接肯認被決定者得要求模型製造者公開其程式碼或參數。至於「可解釋性（explainability）」一詞，僅在與人工智慧技術研發較相關的法律文件中出現，故此詞彙依然有其資訊科學的面向，而並未進入正式的法規範或裁判中。

---

61 Cary Coglianese & David Lehr, *Transparency and Algorithmic Governance*, 71 ADMIN. L. REV. 1, 38 (2019).

62 Mark A. Paige & Audrey Amrein-Beardsley, *"Houston, We Have a Lawsuit": A Cautionary Tale for the Implementation of Value-Added Models for High-Stakes Employment Decisions*, 49 EDUC. RES. 350, 358 (2020).

63 NJCM et al. v. the Netherlands, District Court of The Hague, Case n° C-09-550982-HA ZA 18-388 (Feb. 5, 2020).

64 Beaudouin et al., *supra* note 6, at 29.

進一步思考，以美國的教師考績案為例，若「解釋」僅止於對被決定者各項分數的查核，例如只到「教學若干分、服務若干分，加總若干分」為止，亦即止於 Edwards & Veale 提倡的「主體中心」的解釋方式，而不知其比重及子項目的話，顯然不完全符合人們對「解釋」之期待。然而，若法律要求模型製造者遵照 Edwards & Veale 的第二種作法而提出解釋，也不太可行。「分解法」必須觸碰模型本身以及原始程式碼，可能會侵害開發者的營業秘密等權利[65]，一般人也不見得能藉此更理解決策的理由與判斷當否；至於「模型不可知論」即使用簡化模型來解釋，也未必妥適，因為一般人可能不了解這些簡化解釋的限制，或者提供這樣的解釋也有可能是模型製造者為了隱藏系統不欲為人知的屬性[66]，這反而有損「解釋」之目的。故本文認為，能考慮的或許是踐行「模型中心」的解釋方式。

Gunning 及 Waltl & Vogl 主張，下列問題有助於提升模型的可解釋性：1.為何模型運算產生了這樣的結果（output）？2.為何不是其他結果？3.模型能在哪些案例中產出可信賴的結果？4.能否提供一個信心分數給模型運算的結果？5.哪些情況下，例如何種狀態或何種輸入項之下，模型的結果可資信賴？6.哪個因素最影響模型決策（含正向與負向）？最後，7.如何校正錯誤？即使人無法完全了解困難模型的運作，但上述這些解釋較能使人理解模型被製造出來的過程[67]。這也是具體化「模型中心」解釋的作法。也就是說，即便基於保護營業秘密等考量，得不揭露各項分數的比重與組成，至少應該公開其演算法的種類、最重要的幾項因素，以及提供一些評斷模型效能的客觀指標例如準確率等。不過，這只是一般性的原則，具體案例中要解釋到多

---

65 本文贊成Rudin, *supra* note 39, at 2.的看法，某個模型無法解釋（interpretable）的原因，可能是「過於複雜」例如深度學習，也有可能是牽涉開發者的營業秘密等財產上權益而無法公開（導致無法解釋）。

66 Gilpin et al., *supra* note 9, at 2.

67 Waltl & Vogl, *supra* note 16, at 5. *See* David Gunning, *Explainable Artificial Intelligence (XAI)*, DARPA (May 1, 2017), https://sites.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf.

少，仍應依狀況而定（詳如下段所述），亦容許若干除外條款（針對涉及國安的決策等）。

關於解釋的程度，Beaudouin 等人的研究認為，應從成本效益的角度出發。若 AI 可能帶來的損害愈大，則對於其提供解釋帶來的利益也愈大，如此便可正當化因此而提高的解釋成本。與法政治學的古典命題即亞里斯多德對於權力假設相同，如果某種權力越大，就越有可能危害，就需要監督與分權[68]。其他文獻亦有提及比例的概念，亦即認為應採取適當的方法來減少 AI 帶來的風險，且此些方法必須相應於風險之大小[69]。因此，Beaudouin 等人的研究主張，應在具體的脈絡或情境中決定解釋的程度，舉例而言，侵害法益風險較大的 AI 例如自駕車，提供較詳細的解釋，對於總體社會利益較有意義；相較之下，購物推薦系統或交友建議系統，對人類可能的損害比較小，透過解釋可帶來的社會利益較低[70]。這個說法立基於決策透明能降低風險的假設，本身符合直覺。Beaudouin 等進一步認為，解釋的內容與程度要考量以下 4 點來綜合決定。第一點是「受眾因素」：誰獲得解釋？其專業程度如何？給予其考慮時間是否充足？例如購車者可能較願意去理解這台車的 AI 自動駕駛功能是什麼樣的機制，但若只是線上購物的消費者，便可能較少人願意花很多時間去了解。又例如放射治療的專家比較可能察覺到自動偵測系統是否判斷錯誤，所以此時便不需要過度要求該自動偵測系統的研發廠商提出「解釋」。第二點是「影響因素」：演算法會帶來的危害以及解釋能帶來的幫助。第三點是「規範因素」：使用 AI 之後，哪一種憲法上基本權會受影響，各個國家與法律制度的環境如何。第四點是「操作（operational）因素」：解釋的目的為何？是要確保安全？還是提升使用者信賴？當然，也有不少論者認為，解釋的程度需視使用者是公部門或私部門而有差異，亦即政

---

68 Malcolm P. Sharp, *The Classical American Doctrine of* "*The Separation of Powers*", 2 UNIV. CHIC. L. REV. 385, 385-436 (1935).

69 High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*, EUROPEAN COMMISSION (Apr. 8, 2019), https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

70 Beaudouin et al., *supra* note 6, at 40-41.

府要使用演算法影響人民某些事情的時候，應該要受到比較嚴格的檢視[71]；或者，在例如刑事司法、警政、兒童保護等領域，對於「解釋」的高度要求很可能會造成完全排拒使用任何種類的決策輔助系統[72]。換言之，「解釋」是否足夠，仍要隨著各種脈絡而有不同的判斷標準。舉例言之，法國「數位共和國法案」要求使用 AI 的行政機關，除了公開演算法的種類（模式）及說明其處理的資料來源外，（在合適時）還要公開個人所獲結果的因素權重，已涉及模型中心的解釋，甚至比筆者上述所引的 Gunning 及 Waltl & Vogl 的想法更進一步，因為不僅僅是「重要因素」，而涉及該個人的全因素的權重。這也顯示出，法國認為用於公部門決策的 AI 必須被要求更高度的解釋。

以上關於「可解釋性」或「解釋」的討論，或許有助於人工智慧的法規制定者了解資訊科學領域的進展與他國法規的狀況。以下第「肆」部分則是針對「未來可能在司法場域遭遇人工智慧演算結果的法律人（法官、律師等）」，如何要求「解釋」，提供相關的建議。

## 肆、法律資料分析與可解釋性

### 一、司法體系與 AI 的可解釋性

法律人與「機器學習以及可解釋性」的議題的關聯，並不僅止於立法與規制。隨著愈來愈多人（包含私人機構與政府機關）使用 AI 系統來輔助決策，所引發的糾紛最終還是要進入法院。有學者舉出了司法體系與 AI 接觸而可能得要求「解釋」的兩種場景：第一是法院審查行政機關利用 AI 所為之決策的適法性；第二是司法者自己在審判中使用 AI 來作成保釋、量刑、假釋等決策[73]。第二種情形，法官尤其會希望此演算法係「可解釋」，因為

---

[71] Nikolas Diakopoulos, *Accountability in Algorithmic Decision Making*, 59 COMMUN. ACM 56, 58-59 (2016).

[72] Edwards & Veale, *supra* note 20, at 50.

[73] Deeks, *supra* note 2, at 1838-42.

量刑是否適當，係上訴理由，法官可能有動機去理解模型是否符合一定水準（準確率等）。實際上，美國也有一些州法院不再使用像 *Loomis* 案中的演算法，而改採基於公開資料及原始碼可公開的演算法，顯示了法院自身在使用 AI 時也期待其具有「可解釋性」[74]。在各種解釋方法中，法官可能更需要「模型中心」的解釋，相較之下，被告及其律師可能更想要的是「主體中心」的解釋；而兩者也可能都希望能使用「反事實解釋（counterfactuals）」方法，找出影響模型決策的重要因素有哪些[75]。不論是哪種需求，都將促進「可解釋的 AI」的發展。

上述審判場域使用的 AI 輔助系統，必然是以過去的裁判為資料來訓練、建置的模型。此一領域的研究稱為「法律資料分析（legal analytics）」，即以法律資料（legal data）為分析對象，並提出建議結果等洞見（insight）[76]。本文認為，若未來法院不可避免要使用此種 AI 系統，使用者宜對此系統的運作過程與邏輯有基本認識，也才能適時地要求此系統提出「解釋」，以驗證其決策的妥當性。因此，以下將說明法律資料分析的作法與步驟，並討論此些作法與「可解釋性」的關聯。

## 二、法律資料分析的步驟與「解釋」

「法律資料分析」的過程，與其他機器學習 AI 並無太大差異，主流方式如下圖三所示。1.設定問題，認定所需之資料，2.資料前處理（pre-processing），包括決定資料特徵與其表現方式（representation of feature），

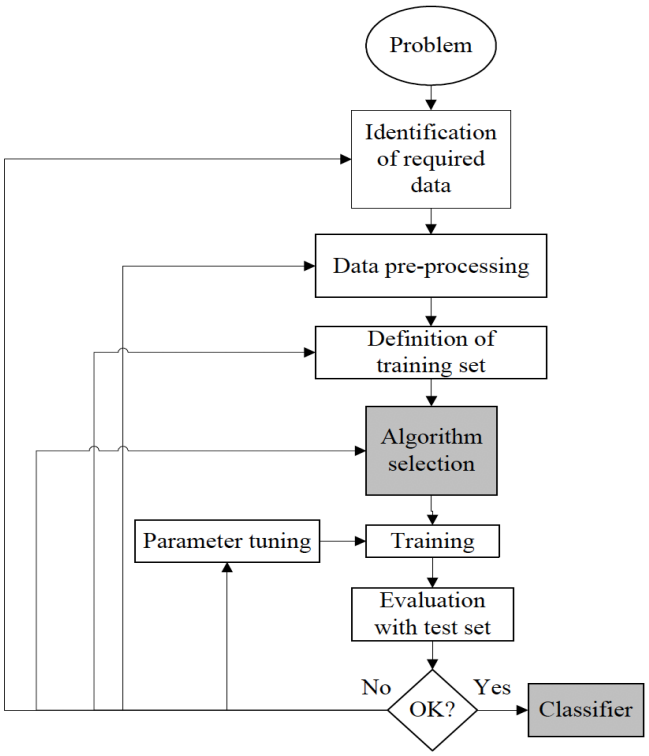
---

[74] Elaine Angelino et al., *Learning Certifiably Optimal Rule Lists for Categorical Data*, 18 J. MACH. LEARN. RES. 1, 1-2 (2018), https://doi.org/10.48550/arXiv.1704.01701.

[75] Deeks, *supra* note 2, at 1846-48指出，除了刑事訴訟場域法官面臨是否運用人工智慧輔助決策之外，還有數種案件（例如在判斷自駕車的產品責任、學校使用演算法評量教師考績的決定、醫師在人工智慧輔助下做出的診斷的合法性時），法官也得決定是否要對該人工智慧的決策要求何種解釋。

[76] 法律資料分析的定義與簡介，參見：黃詩淳、邵軒磊（2019），〈人工智慧與法律資料分析之方法與應用：以單獨親權酌定裁判的預測模型為例〉，《臺大法學論叢》，48卷4期，頁2030-2033。

3.決定訓練資料，選擇演算法（algorithm selection），建立模型，4.使用測試資料進行模型效能評估。




【圖三】機器學習分析問題的流程

※ 資料來源：S. B. Kotsiantis, *Supervised Machine Learning: A Review of Classification Techniques*, 31 INFORMATICA 249, 250 (2007).

以過去的「離婚後未成年子女親權酌定研究」來舉例，在步驟 1「設定問題與資料蒐集」，會先定義一個要回答或是要解決的法學問題，比如親權應該判給母親或父親，即設定為分類任務（classification）；隨後依照某些條件蒐集相關判決。

步驟 2 的資料前處理，在法律資料的領域，是將文字型態的裁判書資料，轉換成數字的形式即「標註（labeling）」（統計學上稱為編碼 coding）。

在這個過程，需要先決定特徵為何（例如主要照顧者、子女意願、父母的撫育環境等），以及該特徵的表現方式（例如是「有或無」，還是「母表現好、不分軒輊、父表現好」等）。此處有各種程度的大量人工勞動，那也就是最需要法學專業知識的地方。

步驟 3「選擇演算法並建立模型」時，因各種不同的演算法可解釋性高低不同，所以會面臨選擇哪一個演算法，既能夠準確的回答問題，又能夠兼顧研究者建立可解釋性系統。筆者在下節更加詳細的詮釋。

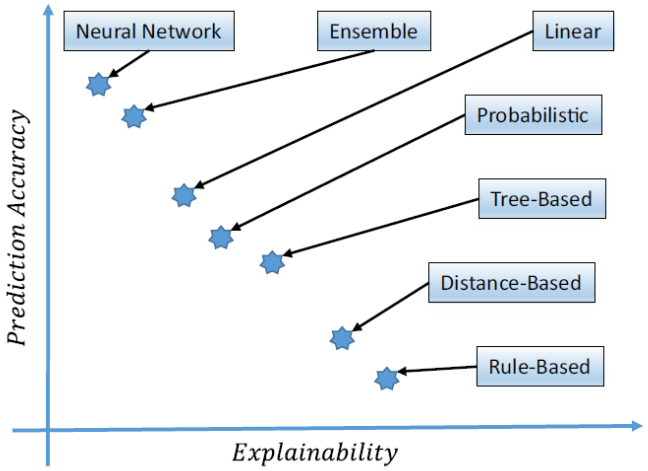
步驟 4「使用測試資料進行模型效能評估」，用測試集也就是模型沒看過的資料來驗證其效能（performance），一般而言可能使用正確率（accuracy）、F1-Score、標準差、損失（loss）等方式呈現。

那麼，倘若有一天「離婚後未成年子女親權酌定研究」（或者其他任何 AI 模型）被用在「輔助法官決策」的場景，則模型製造者應該提供給法官、當事人及其律師什麼樣的「解釋」？由於是用在公部門的決策，對於「解釋」的要求較高，參考上述 Gunning 及 Waltl & Vogl 的見解，此種應用場景的「解釋」，必須包含以上所有步驟的敘述性說明，尤其步驟 3 與 4 的模型選擇與模型成效的資訊。這是因為，上述關於「解釋」或「可解釋性的」文獻考察發現，法官希望獲得「模型中心」的解釋，故演算法是哪種、是否容易理解（步驟 3），以及模型成效（準確率等）如何（步驟 4），相關資訊不可或缺。當事人希望獲得「主體中心」的解釋，亦即哪個因素最影響模型決策，也是步驟 3 可能發現的事項。

## 三、模型的可解釋性差異與選擇

有些模型天生比較「好解釋（interpretable）」，即結構較簡單，較易為一般人所理解，有些則否。一個常見的看法是：「效能高（例如準確率較高）的模型比較難解釋。」不過，的確有些「難解釋但高效能」的模型如神經網路；也有「好解釋但低效能」的如二元迴歸模型；但也有「好解釋又效能不錯」的模型例如決策樹或支持向量機（SVM, support vector machine），或是「難解釋又低效能的」如某些在變項設計上就有錯誤的模型。隨著硬體計算

能力與學術進展，上述的「可解釋性／效能」的標準也可能有若干變化，但在現今既有演算法的極限研發之下，有些研究者確實「可解釋性／效能」視為一種取捨（trade off）。原因在於：如果效能較差且可解釋性也較差的模型，就會被研究社群淘汰掉，因此現存的幾種演算方法，都是在其中一方面優越或是在兩方面平均較為可接受的。




【圖四】可解釋性與正確性的拮抗[77]

※ 資料來源：Philipp Hacker et al., *Explainable AI under Contract and Tort Law: Legal Incentives and Technical Challenges*, 28 ARTIFICIAL INTELLIGENCE AND LAW 415, 431 (2020).

如上圖四所示，最右下角也就是「好解釋但效能低」的 rule-based 模型，是以專家系統為代表，其可解釋性最高，但若用於具體問題的預測，準確度則是最低。比如 1981 年 Waterman 等人建立了「產品責任專家系統」（Waterman's Product Liability Expert System, W-LES），在使用者輸入產品責任相關的案件事實後，該系統會推薦使用者是否和解以及適當的和解金

[77] Philipp Hacker et al., *Explainable AI under Contract and Tort Law: Legal Incentives and Technical Challenges*, 28 ARITF. INTELL. L. 415, 415-39 (2020).

額。其背後的判斷規則即是專家預先設定好的各種「若……則（if…then…）」規則，舉例言之，在判斷產品製造人的嚴格責任（strict liability）是否成立時，此際的規則是：

若原告因商品致傷

或原告代表死者且死者因商品致死

或原告之物被商品損害

且「商品係由被告生產」

或「商品係由被告銷售」

或「商品係由被告出租」

且被告應對產品的使用負責

且加州法院對該事件有管轄權

或該產品之使用者為被害人

或該產品之買受人為被害人

且商品在銷售時有瑕疵

且「商品從製造至銷售時並未改變」

或「被告期待商品從製造至銷售並未改變且被告的期待具有合理性」

則原告的損害得適用嚴格責任理論[78]。

上述系統採用的是向前鏈結（forward chaining）法，亦即若使用者輸入的事實對應設定條件則激發（fire），至迴圈完成至沒有事實可以對應任何條件為止，導出結果。這是最早的法資訊學作法，當初設想如果規則可能窮盡，那判斷過程亦有限（limited）而可能做到完全；這樣的可解釋性無疑是最好的，因為都是肉眼可見的規則（因此又稱為 symbolic AI，已如前述）。

---

[78] DONALD A. WATERMAN & MARK A. PETERSON, MODELS OF LEGAL DECISION MAKING: RESEARCH DESIGN AND METHODS 18-19 (1981). https://www.rand.org/content/dam/rand/pubs/reports/2007/R2717.pdf.

但這樣的專家系統有太多因素妨礙導出結果的正確性，舉例言之，上述規則當中的「被告的期待具有合理性」是不確定法律概念。Waterman 想了二種方法應對，第一是設更多的規則，用以描述過去該當「合理性」的各種事實；第二是提供過去該當「合理性」的各種事實，讓使用者自己判斷「合理性」是否該當[79]。但不論哪種作法，使用者所獲得的結論都有不確定性[80]。換言之，rule-based 系統最大的缺陷在於這個系統往往僅限於「構成要件→法律效果」的判斷，對於行為或事實是否滿足構成要件亦即「法律涵攝」較無能為力。同時，法體系本身的不夠完備、數個並存規則的矛盾或混亂、法解釋細節的衝突都會造成其失能。

其次是距離基礎（distance-based）的演算法，這樣的算法是以某些裁判作為基礎，從而找出相似的裁判[81]。再來是決策樹相關的演算法（tree-based），決策樹模型的解釋能直接追溯至原始變項，非常直觀，其解釋體系也相較單純；模型亦有不錯的效能（正確率、F1-score 等）[82]。更進一步，由於單一決策樹具有若干隨機性質，因此做複數決策樹將誤差分散，則是較為複雜的樹基礎（tree-based）演算法，例如極限提升法（XGBoost）其效能能提升數個百分點，同時亦能夠保留解釋能力，但其解釋體系較為複雜[83]。

更進一步的算法是將分類問題視為機率（probabilistic），由於細分了產出結果（output），提升機器學習的效能，但本身需要抽象化而降低了其解釋性，以及更進一步複雜化其解釋系統。其次，線性迴歸分析（linear regression），將自變項與應變項都以數值表現，其解釋程度受到若干限制，

---

79 *Id*. at 26.

80 KEVIN D. ASHLEY, ARTIFICIAL INTELLIGENCE AND LEGAL ANALYTICS: NEW TOOLS FOR LAW PRACTICE IN THE DIGITAL AGE 10 (2017).

81 邵軒磊、黃詩淳（2020），〈新住民相關親權酌定裁判書的文字探勘：對「平等」問題的法實證研究嘗試〉，《臺大法學論叢》，49卷特刊，頁1267-1308，使用主成分分析（PCA）來探討本國籍配偶與外國籍配偶的親權酌定裁判是否有用語上的差異。

82 黃詩淳、邵軒磊（2018），〈酌定子女親權之重要因素：以決策樹方法分析相關裁判〉，《臺大法學論叢》，47卷1期，頁299-344。

83 黃詩淳、邵軒磊，前揭註76，頁2023-2073。

但其模型效能可以更高[84]。最左上角的 NN（neural network）則是一般人工智慧所稱的黑盒子，由於變數與參數可能高達數百萬個，因此難以被解釋[85]。

過去有學者將同一批的「親權酌定裁判」，用不同的演算法實作，某種程度反映了與上圖四相類的結果，亦即 NN 的效能高於決策樹，如下表一所示，同時也若干證明了「效能——可解釋性」拮抗確實存在。

【表一】親權酌定裁判資料之不同演算法之效能比較

| | **Accuracy** | **F1_Score** |
|---|---|---|
| **Decision Tree** | 0.952 | 0.881 |
| **XB Boost** | 0.957 | 0.927 |
| **Neural Network** | 0.988 | 0.9910 |

※ 資料來源：黃詩淳、邵軒磊（2017），〈運用機器學習預測法院裁判：法資訊學之實踐〉，《月旦法學雜誌》，270 期，頁 91；黃詩淳、邵軒磊（2019），〈人工智慧與法律資料分析之方法與應用：以單獨親權酌定裁判的預測模型為例〉，《臺大法學論叢》，48 卷 4 期，頁 2053；本文製表。

是否要選準確率最高（效能最高）的演算法，必須依照研發目的而定。比如上述使用神經網路（neural network）的文獻準確度雖高，但可解釋性太低，無法提供例如 Gunning 及 Waltl & Vogl 指出的「哪個因素最影響模型決策」的解釋；而之後所嘗試的決策樹方法，則可以做到此點。若離開親權問

[84] 張永健、何漢葳、李宗憲（2017），〈或重於泰山、或輕於鴻毛：地方法院車禍致死案件撫慰金之實證研究〉，《政大法學評論》，149期，頁139-219。

[85] 使用類神經網路（ANN）演算法，黃詩淳、邵軒磊（2017），〈運用機器學習預測法院裁判：法資訊學之實踐〉，《月旦法學雜誌》，270期，頁86-96。

題，在探討慰撫金的金額時，因為不是分類問題（是或否、判給父親或母親等），而是迴歸（金額），必須改用線性的算法。同一個問題，可以使用不同的演算法；或是相反，使用同一個演算法來解決不同的問題，都可以帶給研究者啟發。

以上經驗顯示了法律資料分析的作法是可能疊加（ensemble）的。使用不同的演算法來分析包含裁判書在內的「資料」，並非以「預測裁判結果」（勝敗訴、賠償金額、刑度多少等）為唯一目的，實際上尚可回答法學或社會科學研究社群關心的其他問題。預測法院裁判的 AI，或許只是法律資料分析研究途中的一個偶然產物。若模型的製造者能回答上述 Gunning 及 Waltl & Vogl 提出的 7 大問題，基本上已初步滿足了「模型中心」的解釋要求，符合可解釋性。欠缺可解釋性之 AI 可能被用於商業等私領域，但若是公部門的決策且牽涉人民基本權的侵害時，則不應使用此種 AI。

## 伍、結 論

本文釐清了「AI 的可解釋性（explainability）」的兩個內涵：其一是「理解後說明」的可解釋性（interpretability），其二是透明度（transparency）。在知識論上，可解釋性（explainability）成為一個新的任務，有時甚至需要重新建構解釋體系。解釋可能不是單純的數學或科學問題，而也是一個社會行為。尤其考慮到人工智慧作為新興且專門的學科，吾人恰巧可以目睹後現代的知識論框架如何體現。古典知識論一般假設解釋如同光照，解釋能突破原來的無知，並由經驗佐證，即使有經驗無法包含的知識，也能藉由先驗（a priori）來解決[86]。但後現代知識論並不假設普遍先驗知識存在，同時也因為個人離散經驗的問題，每個人可能有各自的詮釋方式。所以，後現代社會中「解釋」本身成為一個認同行為，可解釋性就會奠基在讀者對於這個知識體

[86] Axel Gelfert, *Kant and the Enlightenment's contribution to social epistemology*, 7 EPISTEME 79, 79-99 (2010).

系的熟悉與信任程度，同時也與講者的說服能力與權威性相關。當讀者是資訊科學社群時，可解釋性通常指涉的是 XAI 所要到達的目標（尤其 transparency 這端的作法，例如分解模型，或以另一個模型解釋這個模型等）；當讀者是法學社群時，可解釋性必須以 interpretability 這端的作法來達成，亦即 Edwards & Veale 所提出的主體中心加上模型中心的解釋方法，後者又可藉由 Gunning 及 Waltl & Vogl 提出的框架來評斷是否達成。換言之，可解釋性與商業傳播或大眾民主政治（mass-democratic politics）一樣，本身就是一個主體互動的結果。

至於誰可以要求提供何種「解釋」？若違反了提供解釋的義務，效果又是如何？是決策本身的否定[87]、損害賠償、抑或重新公開 AI 系統的相關資訊？即使是 GDPR，似乎也沒有明確的解答，而尚處於知識建構階段。本文對於「解釋」提出的框架是，在需要較高的可解釋性時（例如公部門的決策），AI 製造者所提供的資訊，至少要滿足 Gunning 及 Waltl & Vogl 主張的 7 點，包括訓練資料的概述、模型種類、最重要的因素以及模型成效，始符合 GDPR 第 15 條的「有意義資訊」；但不包括各因素的權重[88]，也不包括「分解法」以及伴隨的原始程式碼（對一般人而言未必是有意義而可理解的資訊），亦非「代理人模型」等另一個解釋用的模型（無法確保其解釋忠於原模型）。

人工智慧的法規範設計或者司法領域的人工智慧應用，需要法律人的跨學科理解。例如機器學習可以容納模型一定程度的誤差（資訊學中稱 loss，統計學稱 error），製造者若能說明這樣的誤差為何會產生，以及如實呈現，即已提供了足夠的解釋（符合本文上述的模型中心的解釋）。法律人也要相應提升自身的素養，若在看到「誤差」字眼的瞬間，誤認為 wrong 或 false

---

87 本文提及的德州教師考績案與荷蘭社會安全給付詐欺案，法院似乎認為行政機關憑藉AI所做的決策應被撤銷。

88 許多非線性的演算法中，每個因素的權重並非固定，此與線性迴歸模型當中每個變項均有固定的係數不同。因此，對於非線性的演算法的模型要求「解釋」每個因素的權重有其困難。但其中有若干模型可能藉由某些方式展現出個案判斷時的各因素權重，例如黃詩淳、邵軒磊，前揭註76，頁2057-2061，使用的梯度提升法，就能看出個案中的各因素權重。因此本文前述法國要求行政機關告知被自動決策影響之人民其個人結果的全因素權重，技術上可能做到。

而認定此一模型不堪用或「欠缺解釋」，而不斷要求製造者「解釋」loss 或 error 的意義或者為何存在，此種作法固然不可謂不嚴謹，但極有可能落入刻舟求劍的困境。另外，法律人也要嘗試理解模型中心之解釋底下各種數據的意涵，才能進一步判斷這樣的模型設計是否完善、模型製造者是否有法律上的責任；而不是要求模型製造者提出 transparency 底下的分解法的各種數據，或者單純滿足於模型製造者提出的主體中心的解釋。

與此相對，有另一種對 AI 的極端態度是完全不要求「解釋」，其結果可能比上述的 AI 懷疑論更加危險。舉例言之，市面上有些 AI 判決預測系統，從未公開其模型成效，遑論訓練資料概述、模型種類、最重要因素等資訊，很可能僅是徒具外觀的空殼子。此種司法 AI 對人民具體的權利義務尚不致產生立即的影響，或許無視即可。較令人擔憂的是司法院自身使用的 AI，例如今年剛推出的 AI 量刑資訊系統[89]，雖非預測裁判結果，而是自動標註量刑因子，但同樣使用機器學習來訓練模型，便同樣有模型效能高低（標註是否正確）的問題。AI 量刑資訊系統的設置目的在「使國民法官有更多量刑參考資料，讓量刑更公正透明」，如此一來，自動標註的正確與否，可能影響國民法官的審判結果。比方說，過去某案件中，被害人明明表示不願原諒被告，但卻被誤標記成「被害人的態度有利被告」（量刑減輕因子）時，若將此案件拿來作為現在審判的參考，可能產生不妥適的結果。此外，近期司法院搭上生成式 AI 熱潮，推出「智慧化裁判草稿自動生成系統」。根據新聞稿，此系統係以過去已開發完成的不能安全駕駛、幫助詐欺等兩類犯罪的「量刑智慧分析系統」基礎下，以 AI 來撰寫此兩類犯罪的裁判草稿，以供法官製作裁判時參考；至於認定事實、適用法律及決定量刑等核心事項，仍完全由法官自行決定[90]。對此，包含全國律師聯合會、民間司法改革基金

[89] 司法院（02/06/2023），〈因應國民法官新制，司法院啟用AI量刑資訊系統：具備二種模式、擁有四大優點〉，https://www.judicial.gov.tw/tw/cp-1887-806741-d6471-1.html（最後瀏覽日：02/09/2023）。

[90] 司法院（08/27/2023），〈司法院審慎發展生成式AI應用，以撰寫刑事裁判草稿初試啼聲；期望減輕法官工作負荷，審判核心仍由法官自行決定〉，https://www.judicial.gov.tw/tw/cp-1887-929494-8a9fb-1.html（最後瀏覽日：

會等團體即提出憂慮之聲，認為司法院此次開發及預計投入使用的系統，似已超脫「協助整理」的範圍，而能夠「協助法官判斷」，風險過高，應進行資訊揭露並制定 AI 使用規範[91]。其後，司法院的初步回應是，認定事實部分完全由法官自行決定，系統無法協助法官判斷；若法官認定被告有罪，系統才在適用法律部分生成內容則提供給法官參考[92]。本文認為，AI 的風險不是僅存於生成式技術，民間團體的質疑毋寧是晚到了一步，早在前述 AI 量刑資訊系統，就已經使用了人工或機器自動標記的資料與演算法來計算或「推薦」量刑，而成為後續「智慧化裁判草稿自動生成系統」的基礎。依照本文上述「可解釋的 AI」的標準，司法院既然是公部門，其使用 AI 輔助決策時，便有義務公開訓練資料概述、模型種類、最重要因素以及模型成效之相關數據，但並不包含原始程式碼與各因素權重。

同時，基於以法律資料分析方法從事法學研究之經驗，本文認為，在開發「可解釋的 AI」時，可從目前公認常用的演算法出發。這樣在解釋時較為容易，因通常已有足夠的學術研究成果，可提供主體中心與模型中心的解釋，滿足 interpretability 的要求。

10/19/2023）。

91 民間司法改革基金會（09/26/2023），〈記者會：AI草擬判決的三大疑問與三大風險 要減輕負擔，也要控制AI風險〉，https://www.jrf.org.tw/articles/2550（最後瀏覽日：10/19/2023）。

92 司法院（09/27/2023），〈司法院就民間團體112年9月26日召開「AI草擬判決的三大疑問與三大風險」記者會所提建議之回應新聞稿〉，https://www.judicial.gov.tw/tw/cp-1887-951341-9add3-1.html（最後瀏覽日：10/19/2023）。

# 參考文獻

## 一、中文部分

邵軒磊、黃詩淳（2020），〈新住民相關親權酌定裁判書的文字探勘：對「平等」問題的法實證研究嘗試〉，《臺大法學論叢》，49 卷特刊，頁 1267-1308。https://doi.org/10.6199/NTULJ.202011/SP_49.0001

財團法人中華民國國家資訊基本建設產業發展協進會（2020），《國家通訊傳播委員會「推動我國網路治理發展與國際趨勢研析」委託研究計畫期末報告》，載於：https://www.ncc.gov.tw/chinese/files/20021/5138_42718_200215_1.pdf。

財團法人金融聯合徵信中心（2017），《歐盟個人資料保護規則》，載於：https://www.jcic.org.tw/main_member/fileRename.aspx?uid=1566&fid=1103&kid=2。

張永健、何漢葳、李宗憲（2017），〈或重於泰山、或輕於鴻毛：地方法院車禍致死案件撫慰金之實證研究〉，《政大法學評論》，149 期，頁 139-219。https://doi.org/10.3966/102398202017060149003

黃詩淳、邵軒磊（2017），〈運用機器學習預測法院裁判：法資訊學之實踐〉，《月旦法學雜誌》，270 期，頁 86-96。https://doi.org/10.3966/102559312017110270006

--------（2018），〈酌定子女親權之重要因素：以決策樹方法分析相關裁判〉，《臺大法學論叢》，47 卷 1 期，頁 299-344。https://doi.org/10.6199/NTULJ.201803_47(1).0005

--------（2019），〈人工智慧與法律資料分析之方法與應用：以單獨親權酌定裁判的預測模型為例〉，《臺大法學論叢》，48 卷 4 期，頁 2023-2073。https://doi.org/10.6199/NTULJ.201912_48(4).0005

劉靜怡（2019），〈淺談 GDPR 的國際衝擊及其可能因應之道〉，《月旦法學雜誌》，286 期，頁 5-31。https://doi.org/10.3966/102559312019030286001

鄭伊廷（2021），〈試析「一般資料保護規則」下自動化決策的解釋權爭議〉，《經貿法訊》，279 期，頁 13-23。

鄭明政（2020），〈從 State v. Loomis 案件看 AI 應用於司法審判上的若干問題〉，《台日法政研究》，4 期，頁 165-178。

魏世和（2020），〈歐盟數位發展之形塑：以人工智慧與資料政策為中心〉，《經貿法訊》，267 期，頁 17-21。

## 二、英文部分

Adadi, A., & Berrada, M. (2018). Peeking inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138-52160. https://doi.org/10.1109/ACCESS.2018.2870052

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2018). Learning Certifiably Optimal Rule Lists for Categorical Data. *Journal of Machine Learning Research*, *18*, 1-78. https://doi.org/10.48550/arXiv.1704.01701

Arrieta, A. B., Díaz-Rodríguez, N., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion, 58*, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012

Ashley, K. D. (2017). *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age*. Cambridge University Press. https://doi.org/10.1017/9781316761380

Bastani, O., Kim, C., & Bastani, H. (2019, January 24). *Interpreting Blackbox Models via Model Extraction*. Arxiv. https://arxiv.org/pdf/1705.08504.pdf.

Beaudouin, V., Bloch, I., Bounie, D., Clémençon, S., d'Alché-Buc, F., Eagan, J., Maxwell, W., Mozharovskyi, P., & Parekh, J. (2020, March 13). *Flexible and*

*Context-Specific AI Explainability: A Multidisciplinary Approach*. Arxiv. https://arxiv.org/pdf/2003.07703.pdf

Citron, D. K., & Pasquale, F. (2014). The Scored Society: Due Process for Automated Predictions. *Washington Law Review*, *89*(1), 1-33.

Coglianese, C., & Lehr, D. (2019). Transparency and Algorithmic Governance. *Administrative Law Review*, *71*, 1-57.

Deeks, A. (2019). The Judicial Demand for Explainable Artificial Intelligence. *Columbia Law Review*, *119*(7), 1829-1850.

Diakopoulos, N. (2016). Accountability in Algorithmic Decision Making. *Communications of the ACM*, *59*(2), 56-62. https://doi.org/10.1145/2844110

Doshi-Velez, F., & Kim, B. (2017, March 2). *Towards a Rigorous Science of Interpretable Machine Learning*. Arxiv. https://arxiv.org/pdf/1702.08608.pdf

Edwards, L., & Veale, M. (2017). Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking for. *Duke Law and Technology Review*, *16*, 18-84. https://doi.org/10.2139/ssrn.2972855

European Commission (2020, February 19). *WHITE PAPER On Artificial Intelligence: A European Approach to Excellence and Trust.* https://commission.europa.eu/system/files/2020-02/commission-white-paper-artificial-intelligence-feb2020_en.pdf

Fong, R., & Vedaldi, A. (2017). Interpretable Explanations of Black Boxes by Meaningful Perturbation. *2017 IEEE International Conference on Computer Vision (ICCV)*, 3449-3457. https://doi.org/10.1109/ICCV.2017.371

Gelfert, A. (2010). Kant and the Enlightenment's Contribution to Social Epistemology. *Episteme*, *7*(1), 79-99. https://doi.org/10.3366/E1742360009000823

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2019, February 3). *Explaining Explanations: An Overview of Interpretability of Machine Learning*. Arxiv. https://arxiv.org/pdf/1806.00069v3.pdf

Goodman, B., & Flaxman, S. (2017). European Union Regulations on Algorithmic Decision Making and a "Right to Explanation". *AI Magazine*, *38*(3), 50-57. https://doi.org/10.1609/aimag.v38i3.2741

Gunning, D., & Aha, D. W.(2019). DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, *40*(2), 44-58. https://doi.org/10.1609/aimag.v40i2.2850

Gunning, D., Vorm, E., Wang, J. Y., & Turek, M. (2021). DARPA's Explainable AI (XAI) Program: A Retrospective. *Applied AI Letters*, *2*, 1-11. https://doi.org/10.1002/ail2.61

Gunning, D. (2017, May 1). *Explainable Artificial Intelligence (XAI)*. Darpa. https://sites.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf

Greenberg, M. (2021, July 7). Legal Interpretation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021 ed.). Stanford University. https://plato.stanford.edu/archives/fall2021/entries/legal-interpretation/

Hacker, P., Krestel, R., Grundmann, S., & Naumann, F. (2020). Explainable AI under Contract and Tort Law: Legal Incentives and Technical Challenges. *Artificial Intelligence and Law*, *28*(4), 415-439. https://doi.org/10.2139/ssrn.3513433

Hase, P., & Bansal, M. (2020, May 4). *Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?*. Arxiv. https://arxiv.org/pdf/2005.01831.pdf

Herman, B. (2017, November 20). *The Promise and Peril of Human Evaluation for Model Interpretability*. Arxiv. https://arxiv.org/abs/1711.07414

High-Level Expert Group on Artificial Intelligence (2019, April 8). *Ethics Guidelines for Trustworthy AI*. European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Holmes, W. O. (1899). The Theory of Legal Interpretation. *Harvard Law Review, 12*(6), 417-420. https://doi.org/10.2307/1321531

Joseph, M. (2019, November 24). *Interpretability: Cracking open the black box: Part III*. Deep and Shallow. https://deep-and-shallow.com/2019/11/24/interpretability-cracking-open-the-black-box-part-iii/

Knight, W. (2017, March 14). *The U.S. Military Wants Its Autonomous Machines to Explain Themselves*. MIT Technology Review. https://www.technologyreview.com/2017/03/14/243295/the-us-military-wants-its-autonomous-machines-to-explain-themselves/

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica*, *31*, 249-268.

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable Algorithms. *University of Pennsylvanis Law Review*, *165,* 633-706.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, *23*(1), 1-45. https://doi.org/10.3390/e23010018

Lipton, Z. C. (2017, March 6). *The Mythos of Model Interpretability*. Arxiv. https://arxiv.org/pdf/1606.03490.pdf

Miller, T. (2019). Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence, 267,* 1-38. https://doi.org/10.1016/j.artint.2018.07.007

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K. -R. (2017). Explaining Nonlinear Classification Decisions with Deep Taylor

Decomposition. *Pattern Recognition*, *65*, 211-222. https://doi.org/10.1016/j.patcog.2016.11.008

OECD Legal Instruments (2019, May 22). *Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449)*. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Paige, M. A., & Amrein-Beardsley, A. (2020). "Houston, We Have a Lawsuit": A Cautionary Tale for the Implementation of Value-Added Models for High-Stakes Employment Decisions. *Educational Researcher*, *49*(5), 350-359. https://doi.org/10.3102/0013189X20923046

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press. https://doi.org/10.4159/harvard.9780674736061

Pruett, W. A., & Hester, L. R. (2016, June 3). The Creation of Surrogate Models for Fast Estimation of Complex Model Outcomes. *Plos One*. https://doi.org/10.1371/journal.pone.0156574

Rudin, C. (2019, September 22). *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*. Arxiv. https://arxiv.org/pdf/1811.10154.pdf

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August 9). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. Arxiv. https://arxiv.org/pdf/1602.04938.pdf.

Saranya, A., & Sabhashini, R. (2023). A Systematic Review of Explainable Artificial Intelligence Models and Applications: Recent Developments and Future Trends. *Decision Analytics Journal*, *7*, 1-14. https://doi.org/10.1016/j.dajour.2023.100230

Sarkar, D. (2018, May 25). *The Importance of Human Interpretable Machine Learning*. Medium. https://towardsdatascience.com/human-interpretable-

machine-learning-part-1-the-needand-importance-of-model-interpretation-2ed758f5f476

Selbst, A. D., & Barocas, S. (2018). The Intuitive Appeal of Explainable Machines. *Fordham Law Review*, *87*, 1085-1139. https://doi.org/10.2139/ssrn.3126971

Sharp, M. P. (1935). The Classical American Doctrine of "The Separation of Powers". *The University of Chicago Law Review*, *2*(3), 385-436. https://doi.org/10.2307/1596321

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, *31*(2), 841-887. https://doi.org/10.2139/ssrn.3063289

Wachter, S., Mittestadt, B., & Floridi, L. (2017). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, *7*(2), 76-99. https://doi.org/10.1093/idpl/ipx005

Waltl, B., & Vogl, R. (2018). Explainable Artificial Intelligence: The New Frontier in Legal Informatics. *Jusletter IT*, *4*, 1-10.

Waterman, D. A., & Peterson, M. A. (1981). *Models of Legal Decision Making: Research Design and Methods*. Rand Corporation: The Institute for Civil Justice.
https://www.rand.org/content/dam/rand/pubs/reports/2007/R2717.pdf

# Legal Significance of Explainable AI and Its Practice

*Sieh-Chuen Huang*[*]

## Abstract

This article attempts to clarify whether or which aspects of the "explainable AI", a research hotspot in the data science community, can meet the "explainability" or "right to explanation" required by the legal domain. First, by analyzing recent research in the data science field regarding "explainable AI", the two connotations of "explainability" are found. One is the interpretation brought out by the researchers after understanding (interpretability). And the second is transparency, which is achieved by using methods such as decomposition to show "explanation producing system". Next, this article turns eyes to discussions related to "explanation" in legal domain. The word "explanation" is often used when regulations and judicial decisions require information related to algorithms. But it is more often seen that, instead of "explanation", adjacent concepts such as information access, disclosure, due process, etc. are used. However, there is still considerable debate on whether regulations such as GDPR can derive the "right to explanation" and what its connotation is. After comparing the idea of "explanation" in both data science and law, this paper argues that, when a higher level of explanation is required (for example, when reviewing public sector decisions), exogenous approaches such as surrogate models developed by the data scientists do not satisfy "meaningful information" defined by law and hence are not legally qualified explanations. The information provided by AI producers should at least include an overview of the training data, the type of model, the most important factors, and the effectiveness of the model. The above information consisting of "production system of interpretation" may comply with the "meaningful information" of Article 15 of the GDPR. On the other hand, the weight of each

[*] Professor, College of Law, National Taiwan University.
E-mail: schhuang@ntu.edu.tw

factor or the source code is not included in the information that should be legally disclosed. Finally, with regard to the judicial AI that may appear in the future, this article takes the relevant research on legal analytics as an example to illustrate the relationship between the processing and explainability, so as to benefit users such as judges and lawyers to properly exercise the “right to explanation”.

**Keywords: Explainability/ Interpretability, Right to Explanation, Model-Centric Interpretation, Subject-Centric Interpretation, Legal Analytics, Global Interpretability, Local Interpretability**