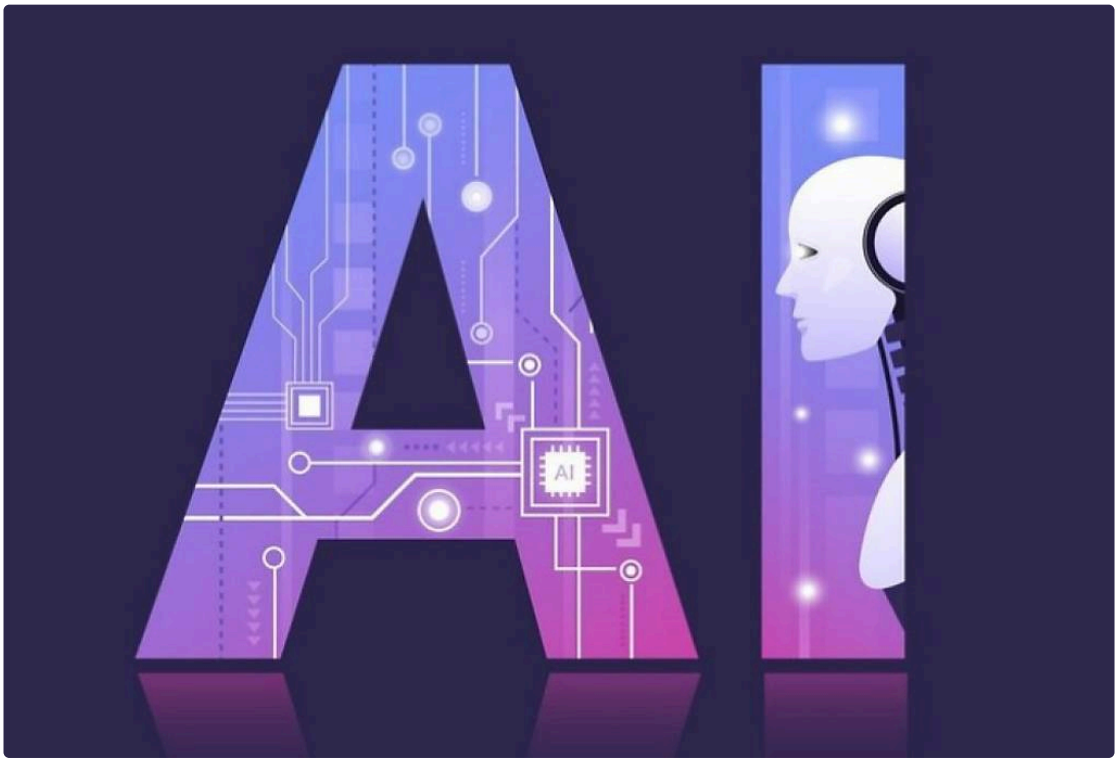
AI輔助判決真的只是輔助嗎？

—談法律黑盒子與透明性

2024.06.27 03:00 工商時報 法律觀點 AI



●司法院審慎發展生成式AI應用，公告將推出「智慧化裁判草稿自動生成系統」。圖 / freepik

文 / 馬靜如、胡浩叡、邱佩冠■國際通商法律事務所 主持律師、執行合夥律師、合夥律師

AI anchoring bias（AI錨定偏誤）講的是一種使用人工智慧（AI）工具的心理效應，當人類要對抗AI決定時，會覺得AI已經是大數據分析後的結果。因此，相較於去否定一個人類做出的決定，拒絕AI決定需要更大的勇氣，這也是司法院公告將推出「智慧化裁判草稿自動生成系統」後，法界人士最主要的擔憂。

即將出台的金管會AI指引也令人引頸期盼，與司法院不同，這次金管會的AI指引，不是金管會本身要做AI，而是提供一個方向，讓金融機構在往AI前進的路上，有個方向可以依循，不會這麼膽戰心驚。

那麼，我們就先回到法學界對AI的討論。首先，要先介紹一個HIC（Human in Command）概念。早在2017年，歐洲就有倡議「應該採取HIC方式去監管AI」，這個HIC（人在指揮）的概念是，「機器仍然是機器，而且人類總是能控制機器（where machines remain machines and people retain control over these machines at all times）」。另一個觀念是HITL（Human-in-the Loop），依照維基百科的解釋，HITL是一種需要人類介入互動的模式（a model requiring human interaction），譬如在機器學習階段，人類可以去調校這個AI模型，避免它產生對人種或性別的歧視結果，後來也有人說，調校參數已經是Human-on-the-Loop（HOTL），就是人類在跟機器互動的過程，會一直給機器反饋，去改善它的表現，這時候人已經退縮到「監督」的角色，而非與機器「共存」的HITL階段。最後是Human-above-the-loop（HATL），在這個階段，人類只有機器出錯時才會介入。

■台灣「智慧化裁判草稿自動生成系統」，尚無法協助法官判斷

那麼，外送、訂房或叫車的App，它們應該是屬於哪一種呢？大家應該都有跟這些App互動的經驗吧，即使莫名奇妙被取消，消費者也無法叫App「揣共」（出來解釋）或更正，但多數App還是找得到真人客服處理客訴。因此，連日常生活都可以走到HATL階段了，法院的「智慧化裁判草稿自動生成系統」是屬於

哪個階段呢？按照司法院112年9月27日新聞稿的說法，「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」，這顯然尚未到HATL的程度。

另一個問題是，依據歐盟的人工智慧法案，我們可否要求司法院將「智慧化裁判草稿自動生成系統」的演算法形成過程公開？這個要求有點讓人為難，當我們談到AI的黑盒子（Black Box）時，有法律黑盒子和技術黑盒子，法律黑盒子指的是，公開可能會侵犯到營業秘密，譬如我們無法要求叫車平台公開它的演算法，因為那是它的秘密武器，而技術黑盒子指的是，機器為什麼作出那樣的決定，即使公開了我們也無法理解。

■要求AI透明性，恐成犯罪者規避的巧門

那麼，會不會有一天，檢察署會有刑事偵查AI？如果有的話，那演算法更不能公開了，一旦公開了，罪犯就知道怎麼去躲避偵查。刑事偵查AI也許太遙遠，銀行的AI洗錢防制系統已百家爭鳴。如果一旦公開，車手就知道如何閃避，那AI的透明性要求，就跟訓練這個AI模型當初的目的相違背，在這種前提下，我們是否能例外允許AI不具透明性？

這些問題好難，我們也不應該期待金管會出台的AI指引或司法院的AI一次到位，而是應該先肯定台灣主管機關在AI世界混沌未明時，就已經有領先其他亞洲國家初試啼聲的勇氣，筆者很期待看到主管機關發展更多的AI應用。