《歐美研究》第五十卷第二期 (民國一〇九年六月)，207-229
DOI: 10.7015/JEAS.202006_50(2).0004
https://euramerica.org

淺論 AI 風險預測的規範性爭議

洪子偉
中央研究院歐美研究所
E-mail: htw@gate.sinica.edu.tw

摘 要

AI 預測技術深具潛力的其中一個應用，在於分析過去資料以預防極端氣候的災害。但當預測對象從自然環境變成人類本身，爭議隨之產生。本文旨在探討以人類資料作為風險預測之規範性爭議，並主張：(一) AI 的不可解釋性，並非因其無法提供機械式步驟，而是人類的認知限制無法對數量龐大的步驟賦予意義。(二) AI 的歸納法、黑箱等問題在大腦上也會遇到，兩者的差異是程度上而非種類上的。(三) 必然性與事實性條件並無法一致地排除極端案例，卻又不排除既有法律或社會規範。(四) 自主性原則之優點在確保權責相符、避免喪失人類能力、降低 AI 發展的社會阻力。

關鍵詞：規範性、歸納法、隱私、人工智慧、預測技術

投稿日期：108.6.28；接受刊登日期：108.8.14；最後修訂日期：108.8.23
責任校對：蔡旻芳、張起綸、曾嘉琦

壹、前言

氣候變遷帶來的災害已造成全球數以萬計的氣候難民，國際移民組織 (International Organization for Migration) 預估 2050 年將有 2 億的環境移民，極端氣候已然成為全球共同的難題之一。而隨著人工智慧 (artificial intelligence; AI) 快速發展，如何將之用來解決環境危機似乎是刻不容緩的議題。

Nature 期刊研究指出，機器學習大幅提高對極端氣候的預測準確率，有效減少生命財產的損失 (Jones, 2017)。舉例來說，美國國家能源研究科學計算中心 (National Energy Research Scientific Computing Center) 團隊所發展的深度卷積神經網絡，就可預測極端氣候事件，並達到 89-99% 的準確率 (Liu et al., 2016)。IBM 的綠色地平線 (Green Horizons) 計畫，也利用各項 AI 技術來分析交通狀況、天氣、濕度、風力等因素，並能在 72 小時前就預測出北京在 1 平方公里範圍內的空污風險。牛津大學 Digital Ethics Lab 提出的機器人學未來的重大挑戰與突破中，則包含如何在缺乏地圖與有限資訊下，在極端環境下去適應、學習、並克服可能失敗 (Yang et al., 2018)。換言之，人工智慧領域的各項技術用於災害防治上料將帶來莫大好處，亦是深具潛力的應用。

然而，當風險預測的對象從自然環境變成人類本身，爭議卻隨之產生。據 2018 年 BBC 報導 (Thomas, 2018)，英國企業 WeSee 的面孔辨識技術，已能透過微表情、姿勢與動作預估人的情緒狀態與可能意圖，從而判斷其威脅性。未來可應用到地鐵來預防可疑行為與恐怖攻擊，但這項計畫受到國際隱私組織 (Privacy International) 的質疑。無獨有偶，Faception 公司也研發出臉部辨識技術以預測恐

怖分子、戀童癖與王牌撲克玩家，[1] 且與以色列國土安全部門合作找出潛在的恐怖攻擊。但華盛頓大學資訊系 Pedro Domingos 教授質疑，光是透過臉部辨識來預測某人可能會是謀殺犯而將之逮捕，本身極具爭議。

此外，中國數位極權的崛起，更飽受國際抨擊。非政府組織人權觀察 (Human Rights Watch) 指出，中國除透過臉孔辨識及大數據來追蹤通緝犯，也將技術用於標定潛在的反政府人士。[2] 中國將「身分證號碼」與監視器、生物識別特徵、住房與航班紀錄等資料整合，以即時監控「高風險」的涉恐、涉疆、涉穩與前科人員。甚至在海外批評政府也無法倖免。[3] 另在預測犯罪方面，該組織也證實大數據已用在預測性警務，讓當局可任意拘押可疑維吾爾人。2018 年，聯合國消除種族歧視委員會 (Committee on the Elimination of Racial Discrimination) 正式要求中國釋放囚禁在「再教育營」的一百萬穆斯林與維族人 (Nebehay, 2018)。2020 年初，美國國務院指囚禁人數上升達兩百萬人 (Rivers, 2020)。如何善用 AI 促進人類福祉，而非戕害人權，便成了重要且迫切的課題。

本文之目的，在探討將 AI 預測技術用於預防人類風險所引發的規範性爭議。本文將從「演算法相關」(第貳節) 與「資料相關」(第參、肆節) 兩方面來討論。前者從知識論角度探討此預測技術的結果可靠度與過程的不可解釋性。後者則從倫理學角度分析人類資料使用上的可能爭議，並探索潛在的解決辦法。

---

[1] 該公司曾只依據臉部特徵預測四位最佳撲克選手，結果其中兩位果然進入前三名。

[2] 廣州的「雲從科技」以面部識別與動作追蹤來評估犯罪風險。例如購買菜刀的人並不可疑，但如他也曾買過錘子和麻袋，其可疑評估等級就會上升。

[3] 紐約時報記者 Mozur (2018) 指出，中國維權人士張廣紅以 Facebook 旗下之 WhatsApp 散播批習文章，其對話內容被當成證據而被起訴。因 WhatsApp 有通訊加密，紐時報導研判是手機遭駭。

在定義方面，本文中 AI 的內涵 (intension) 是指可模擬人類思考之外顯行為的人造自主適應系統，且該系統之輸出結果與人類所表現出的特定認知能力相同或更好。在外延 (extension) 上，本文採取牛津大學 Digital Ethics Lab 的觀點，將決策演算、傳統人工智慧、機器人學等所有具有學習能力的人造自主適應系統，皆視為是 AI。[4] 至於「AI 預測」則指這些系統在特定任務中，根據所輸入的資料而輸出或修正有關未來可能結果的一種能力。

貳、演算法相關爭議（知識論相關）

若暫不考慮倫理爭議，目前對 AI 預測技術本身的知識論相關的規範性疑慮，主要有兩方面：一是「結果的可靠度」，指 AI 的預測機制主要仰賴機率學習，[5] 但人類思考則相當不同：常結合演繹、歸納等多種推理方式對某事件進行綜合判斷。因此即使 AI 在特定任務的資訊處理表現較佳，但在跨任務的綜合判斷上仍不可靠 (貳之一節)。[6] 二是「過程的不可解釋性」，指 AI 預測的過程如同黑箱。它會根據輸入的數據與特定的演算法，歸納產生自己的規則與模型以輸出預測結果。但其規則與模型如何產生，連科學家自己可能也不清楚，亦無法給予明確的科學因果解釋 (貳之二節)。以下將分別檢視這兩個質疑。

---

[4] Wachter, Mittelstadt, & Floridi (2017) 在批評歐盟的一般資料保護規範 (General Data Protection Regulation; GDPR) 時主張，制訂不精確的規範將決策演算、傳統人工智慧、機器人學等分開處理是相當危險的。

[5] 雖然 1960 年代的古典 AI 是以演繹推論 (如 rule-following) 出發，但目前 AI 機器學中仍以歸納 (尤其 Bayesian learning) 為主。雖然有學者嘗試結合兩者，但目前在業界仍屬少數。

[6] 這可算是 Jerry Fodor (2008) 對計算整體性 (globality) 質疑的歸納版本。

一、人類大腦與機器預測的相似性

AI 計算過程與人類大腦的資訊處理是否截然不同，可分為「硬體實現層次」與「計算層次」兩方面討論。[7] 在硬體實現層次上，AI 預測技術仰賴在矽晶體上的數位式計算。[8] 反觀大腦的神經傳導流 (neural spike trains) 是由離散的電化學訊號所構成而不屬類比式，但這些神經傳導流卻又被視為是連續訊號般被處理，也非數位式的 (Piccinini & Bahar, 2013)，故兩者的確差異甚大。但如果比較的是「計算層次」的資訊處理策略，則 AI 機器學習與人類大腦運作的基本原理相類似，但更為精準。

更詳細來說，根據認知科學中的預測編碼 (predictive coding) 假說主張，人類大腦如同一預測器，會不斷地比對知覺訊號來即時修正各種有關外在世界的先驗假設，而這種方式已被證明符合貝氏最佳化 (Brown & Friston, 2012)。例如，當大腦在決定某個處境中的最佳行動時，認知架構中的前進模型 (forward model) 會產生外在訊號的預測 (exteroceptive prediction)。這些預測透過神經元間的電化學反應加以傳遞，並與由下而上的感官訊號進行比對以產生校正訊號，進而產生更精確的預測。而這些預測訊號的總和，就是人類大腦的知覺內容或運動指令。因此，大腦的決策是透過衡量各種機率的加總，來找出最有可能的行動方案。由於每個人成長經驗 (先驗假設) 與所處環境 (即時輸入) 都不同，對世界的預測自然迥異。

---

[7] 依 David Marr (1982) 觀點：「計算層次」(computational level) 是用電腦科學的術語來描述有關於大腦的資訊處理策略與一般性任務。「演算層次」(algorithmic level) 是詳細說明計算層次的任務可以透過哪些演算法加以執行。「硬體實現層次」(implementation level) 則探討上述算法可以藉由何種物理機制 (神經細胞或矽晶體) 加以實現。

[8] 數位計算處理的表徵 (representation) 是不連續編碼，類比式計算處理之表徵則為連續編碼。

此外，人類無法僅以親身體驗來獲取知識，亦須仰賴推論從已知推出未知。但「推論」須以「分類」為必要條件。例如如果沒有動詞、名詞、形容詞等分類，英文的文法規則將無以適用。同樣在語句邏輯上，若無先定義什麼符號代表語句，什麼代表語句集合，什麼代表邏輯連接詞，推論的語法規則也無從適用。故沒有分類就沒有推論。大腦之演化出「分類」能力是因可減少認知負擔，縮短決策時間。甚至這些分類也有助於形成直覺。但分類其實就是「貼標籤」。在社會認知的推論上，由於每個人的成長背景不同，分類與貼籤難免形成各種成見與刻板印象。這些基於片面資訊形成的預測常雖是錯誤，卻偏偏是人類決策的常見模式。當人類偏見被以行為數據形式保留（例如統計顯示日本醫學院錄取率以男性居高）並餵給機器。這樣的偏見錯誤就會被複製，甚至匯入更大資料庫，造成「垃圾進，垃圾出」(garbage in, garbage out) 現象。

由此可知，在資訊處理的策略（即計算層次）上，AI 的學習與大腦預測器皆具相似的運作原理。但由於機器具有比人類更強大的計算處理能力，當運作適當時會帶來比人類更龐大的效益。然而當出錯時（如偏見複製），其負面影響也會更加廣泛。

二、過程的不可解釋性

機器學習常根據巨量數據來歸納出規則或模型。這些規則或模型是如何輸出所預測的結果，常常連研究者自己也難以解釋。這種不可解釋性是機器學習所面臨的一大問題：它讓人很難完全信賴一個黑箱（black box）所產生的預測（Hung & Yen, in press）。

這種不可解釋性從何而來呢？現今不論如何強大繁複的演算法，仍然只能在符合圖靈可計算性（Turing computability）的電腦硬

體上執行。[9] Turing (1937a, 1937b) 描述了一個可執行一串指令序列的抽象通用裝置。裝置中每條指令都是某個演算法的一個子句或步驟程序，且皆可被機械式地執行。如果某個數學函數，從中導出的值可透過有效程序 (effective procedures，即可在有限處理資源、時間內，以有限數量的步驟實現) 來辨識，則該函數就是可計算的。[10] 因此，現今的 AI 演算的執行程序不論如何複雜，「理論上」依然可在連續時間中分解成個別的有限步驟。但問題就在於，「實際上」當上億、兆筆的步驟呈現在眼前，以我們有限的認知資源，並無法賦予這些步驟意義。因此這種不可解釋性並非來自演算法本身無法提供一個機械式的步驟或因果機制，而在於人類理解上的限制，很難理解這些大量的演算過程。

當 AI 可用來輔助或取代人類決策者時，這種不可解釋性又常被認為會造成問責性 (accountability) 上的困難。然而，有關 AI 不可解釋性或黑箱之疑慮，對人類大腦一樣適用。甚至，人類大腦的「黑箱」問題還是雙向的。一方面，過去認知科學中的行為主義便認為某些內在的心靈狀態及其運作機制難以測量，大腦如同黑盒子般，因此主張透過研究可被觀察和檢證的外顯行為，來研究理解人的心理狀態。即便現在，諸如 MRI (測血氧變化)、MEG (測神經活動) 等測量技術日新月異，已可詳細記錄各種大腦活動。但這些紀錄本身與特定的心智活動間有何關聯，仍有不少未解之謎，其因果機制拼圖仍有待認知科學家的持續研究。另一方面，預測編碼假說

---

[9] 雖然超計算 (hyper-computation) (Burgin, 2004, 2005) 包含一組演算法和自動機可以用來處理圖靈機所無法計算的函數，但硬體上目前仍做不出來。

[10] 根據圖靈自己的看法，這種抽象邏輯裝置可以「計算任何可計算的序列」(Turing, 1937b: 10)，並可以處理任何「純粹機械」的東西 (1948: 7)。由於這個命題後來被證明與 Church (1935, 1936) 的命題「任何遞歸函數可以被有效地計算」等值，兩者被共稱為丘奇-圖靈論題 (Church-Turing thesis) (Kleene, 1967)。

與康德都在回應 Hume 對歸納法的挑戰。Swanson (2016) 指出，以貝氏機率來理解人類認知最大的問題在於「人類的思維如何超越經驗數據」，大腦唯一使用「數據」是感官傳來的訊號，而且大腦只測量此感官訊號而不直接測量外在世界。這會產生一個難題：大腦如何只根據感官訊號的「結果」，而推知其在外在世界中的「原因」。這個難題被 Clark (2013) 形容是「從黑盒子裡看世界」。Hohwy (2018) 則稱做「囚禁於頭骨內的大腦」。換言之，AI 的「黑箱」並不是什麼新問題，人類決策者的心理歷程 (不論硬體或計算層次的神經活動，或是意圖層次的「自由心證」) 已是如此。[11]

在過去法律上，並不會因為人類大腦 (不論是意圖層次還是計算層次) 的這種不可解釋性而造成問責上的問題，因為對於自然人主要仍以其「行為」作為課責的依據。[12] 一方面，如果某天 AI 可以完全取代人類而決策，在哲學上才有可能成為行為者 (agent) 或位格人 (person) 而負有道德責任。[13] 但這類型的 AI 在可預見未來是不可能的 (Floridi, 2016)。故當某 AI 預測系統誤發警報而造成人

11 值得注意的是，人類大腦的運作過程雖是黑箱，但是藉由著行動者意圖之外顯或是他者對於行動者意圖之認知 [或猜想或預設])，其行動仍可被他人理解。故人腦雖然是雙向黑箱，但至少人類透過意圖可以給予其行為意義，反觀現今 AI 仍無法如此。當然，兩者的深入比較值得未來進一步研究。在此感謝匿名審查人之補充。

12 更詳細講，Dennett (1989) 曾提出解釋階層結構 (a hierarchical structure of explanation)：「意圖層次」(intentional level) 透過理解行為者的意圖來解釋其行為。「設計層次」(design level) 則透過功能機制來說明意圖層次中的心靈狀態。「物理層次」(physical level) 則聚焦的是設計層次的功能機制在實際上能夠如何被建構。我們可以說，在法治國家，光靠人類在「意圖層次」的活動並無法作為起訴理由 (光有犯罪意圖／動機而無犯罪事實仍不夠)。但在設計或物理層次的「腦造影資料」有時卻可做為心神喪失或無行為能力者的佐證。

13 「責任」一詞具多重意義，可指 (一) 因果責任；(二) 道德責任：對行動者的讚揚譴責；(三) 法律責任：對其處罰或要求賠償；(四) 要求行動者對其行動提出解釋或說明 (accountability)。本文中所討論的是一般性 AI 爭議，各意義多交替使用。

員傷亡，比較像是某台自動販賣機漏電傷及消費者，其法律責任應由對其設計者、製造商、維護人員中加以釐清。[14] 另一方面，即便目前 AI 仍無法成為道德上的行為者，未來它仍然可能成為法律約束的對象。畢竟除了自然人 (natural person) 外，法律上也有法人 (legal person) 作為權利義務主體。儘管於強的、廣義的 AI 至今仍無法實現，對 AI 本身的課責仍不切實際，但隨著技術成熟 AI 將可能輔助或部分取代人類決策。此時，AI 的設計人與管理人仍可以某種「委任關係」(如董事與法人之間) 與該預測系統連結，而成為法律約束的對象。換言之，要將 AI 視為權利義務主體不難，難的是當因果關係不明而出事時，究竟是軟體撰寫者、餵錯資料者該當負責。

這時可能的做法有二：首先，無法釐清問責性的風險預測應避免由 AI 進行最終決策。其目的是為了避免權責不清，並使民選政府的責任政治可以順利運作。相對而言極權國家較無此困擾。[15] 其次，AI 的不可解釋性來自於人類的理解限制。但 AI 技術又是要補足人類認知的不足，因此未來「問責與因果 AI」或許也會出現來輔助人類判斷。[16] 但這會否造成無限後退？理論上會但實際應不太可能，這是因為社會資源有限。譬如最高法院判決後即無法再上訴。

由上可知，AI 的不可解釋性在人類身上一樣會出現，兩者也都可能成為權利義務規範的主體。但與人不同的是，除非 AI 有朝成

---

14 即便目前機器學習在預測上所展現的強大的適應性與自主型，仍與人類有相當差距。

15 某些獨裁國家為避免領導人的英明決策與威信受損，反而可能對 AI 的決策輔助更加審慎運用。然而，民主國家的究責不僅只問決策的結果，更會追問決策所依據的理由。

16 如 Pearl & Mackenzie (2018) 所稱，目前 AI 只能透過觀察相關性來做出預測，而無法透過干預操作做出因果推論。但若未來 AI 發展出強大因果推論能力，能告訴人類應採取何種行動以避免風險，或 AI 自己就能直接干預時，AI 的究責便是重要問題。惟此議題已超出本文焦點。

為道德上的行為者，否則究責的對象仍是相關的委任關係人，這些人與 AI 系統間的權利義務能否被解釋而加以問責才是問題關鍵。因此，AI 預測的「問責性」問題並非 AI 預測過程本身的「不可解釋性」所直接造成，問題乃在於人。

三、結果的可靠度

此外，除了 AI 預測「過程」的不透明外，對其預測「結果」的效力亦潛藏問題。目前 AI 在風險預測上的方式主要可以分成兩類：一個是尋找兩個現象的正相關性 (例如史丹佛大學 [Stanford University] 的爭議研究中，以 AI 尋找臉部特徵與性傾向之連結)。至於兩者的因果關係是什麼，通常不清楚或根本沒有因果連結。另一個是從預設的因果模型當中，找出結果發生時的高機率因子，從而在其他類似情況中預測結果發生的可能性。這兩種方式所依賴的均為歸納統計，且常以貝氏機率為主。舉例來說，在機器學習中常用到的梯度下降學習 (gradient descent) 是一種不斷根據後驗機率來修正所預測的假設 (先驗機率)，一步步找出最接近可能結果的方法。這種機器學習和所有生物適應系統一樣，都是仰賴輸入或刺激的頻率，來找出最適當的輸出或行為。[17] 然而，相較於強調「必然性」的演繹邏輯，上述這種歸納統計強調的只有「可能性」。因此，即使這種歸納預測器指出明天有 99.9%的機率會下雨，仍不保證明天不會是晴天。此外，其當某預測事件的可能性只是出於正相關，而

[17] 若 AI 預測所需的能力與人類特定認知能力不相上下，則結果可靠度議題的重要性似乎就會減弱。然而，我們總是會擔心 AI 預測的結果究竟對人類的判斷行動有何意義。故這裡所探討的結果的可靠度，並非是一純粹知識論上的問題，而是一個連接倫理學相關的關鍵概念。人類大腦雖然可能以預測編碼理論運作，但人類大腦對於社會事件的意義理解或許是透過因果關係而來，進而產生規範性的爭議。故缺少清楚的因果機制時，也會造成問責性上的困擾。這點感謝匿名審查人之補充。

缺少清楚的因果機制時，也會造成問責性的困擾。因此，這一類的預測並不適合作為單一的直接證據、或多重證據之一。這種 AI 預測頂多作為人類決策者在風險管理中的重要參考，使人類決策者可以根據各種資訊做出最後決定，例如是否對某人申請搜索令、是否該在暴風雨前撤離地質敏感區的居民。

參、人類數據的使用爭議（倫理學相關）

AI 預測的兩難在於：一方面，其技術價值取決於預測精準度，準度越高所需的資料量要越大。但另方面，所需蒐集的資料量越龐大，就越可能對民眾的權利造成侵害，從而減損 AI 之社會價值。這些權利譬如隱私、匿名商業行為、政治言論自由等 (Lin, Hung, & Huang, in press)。那麼，我們是否該允許此技術？是否有辦法在善用此技術的同時，保障民眾的基本權利？

生命倫理學中，對人類資料的蒐集、保存和使用有嚴格規範，在 AI 預測中亦然。本節將聚焦在如何「使用」這些資料，並嘗試回答「應否利用 AI 預測技術來預防人類產生的風險？」(參之一節) 與「若可又該如何利用？」(參之二節) 兩問題。本節主張，常見的「必然性」與「事實性」條件並無法用來反對將 AI 預測技術應用到預防人類造成的風險，因為他們同樣會挑戰既有的法律與社會規範。而開放將此技術，並引進生命倫理的自主性原則，可避免某些極端案例並帶來其他優點。

一、應否以 AI 預測來預防人類風險？(必然性與事實性條件)

要回答第一個問題，或許我們可以設想最極端的例子「應否利

用 AI 預測來拘留或逮捕嫌疑犯？」這個問題可以被改寫成「人能否因自己還沒做的犯罪而被逮補？」這樣的疑慮。對很多人來說，這個問題最直覺的答案就是「不行」。例如華盛頓大學資訊系 Pedro Domingos 教授便質疑，光是透過臉部辨識來預測某人可能會是殺人犯而將之逮捕，本身極具爭議。但是在此直覺背後是否有任何支持的道德理由呢？

根據常見的直覺，或可歸納出「必然性條件」與「事實性條件」兩個理由。必然性條件指出，AI 預測模型多採用貝氏推論，根據過去數據來預測未來行為。但是貝氏推論僅能輸出可能性而非必然性。因此，即便有 99%的機率，也還是有可能發生抓錯人的情況。甚至可能也永遠無法證實究竟抓的是否是正確的人，因為嫌疑人一旦被抓就不會有機會去實現「被預測會犯的罪」。但「必然性條件」也會受到反駁，理由在於，雖然在科學研究或知識論上必然真是很重要的價值，但在社會層面可能性比較重要。例如，司法上之親子鑑定 (DNA paternity testing) 之判決，如民事離婚訴訟、刑事拐賣兒童，所根據的鑑定報告也只有可能性 (如準確度 80-99%) 而非必然性，故為何不行？可能的答覆是：因為親子鑑定中有無親屬關係屬已發生之事實，但 AI 中卻是尚未發生。這種說法，就跟我們第二個要談的「事實性條件」有關。

所謂事實性條件是指，AI 之預測結果並非已發生之事實，而是未發生 (或尚未發生)，人不應因為沒有發生的行為負責。畢竟如果未發生，就沒有這樣的行為，沒有這樣的行為又如何把行為的責任歸咎於他呢？這種說法乍聽之下很合理。但仔細想想，在日常生活中其實有很多違反「事實性條件」卻行之有年的法律或社會規範。例如，美國銀行的 credit rating system，也是根據過去資料來預測未來 (還未發生的) 還款能力，從而決定個人之信用額度。為何這些

可以但 AI 預測就不行？或是更具體講，為何美國銀行的個人信用評等可以，而中國 AI 預測的「社會信用系統」卻備受質疑？是否雙重標準？

支持「事實性條件」的人可能會辯護說，這些反例與 AI 的情形有兩個不同：一是前者有經過知情同意的契約過程、二是前者的銀行是利害關係人 (銀行須承擔呆帳風險)，屬民法中的損害賠償的潛在被害人。相反的，中國政府與人民是否有此契約關係，或是中國政府在何種意義下是潛在被害人，均難以認定。但上述辯護並無法釋疑。畢竟在民主國家的法律中，人常常會為了尚未發生但可能發生的行為負責。例如刑法上不是只有酒駕肇事才有罪，酒駕本身就有 (公共危險) 罪了。同樣的，千面人在飲料下毒，不論有無受害者誤食都有罪。這又該怎麼說？畢竟「預防」的概念就是在事情發生前採取行動介入，「事實性」條件似乎忽略這一點。

支持「事實性條件」者可能會回覆：當涉及公共利益或危險重大時，可在事實發生前採取行動。當風險愈大，越應嚴格。此外，行為者的意圖也很重要。如有犯意即便未遂也應究責 (酒駕罔顧人命、千面人預期恐慌)。當然，兩者刑度有別。然而，這樣的回覆仍有問題。一方面，如以公共福祉作為 AI 預測之條件，(即便不考慮執行面是否被濫用，「公共利益」被濫用解釋) 能否一致性地應用？例如當某人自殺不具公共性，故不須阻止，但某自殺炸彈客則須被阻止。然如果預測到某人會自殺卻不阻止，難道沒有道德責任？如果要阻止，且宣稱個人生死具公共福祉，又是否擴大解釋公共利益的界定，而允許更多濫用 (如維穩) 的案例進入？另一方面，如以行為意圖為標準，意圖難以認定 (且一個被 AI 預測為高風險的人可能完全無犯罪意圖)。在執行上也有其困難。

由上可知，即便在「人能否因自己還沒做的犯罪而被逮補？」

這種極端的例子中，我們仍沒有辦法找到一致性的理由來反對以 AI 預測來預防人類所造成的風險。如果我們以「必然性」或「事實性」條件來反對 AI 預測，就會面臨既有法律與社會規範的問題：要就兩個都要禁止，否則兩個都允許，既然無法禁止，兩者應都開放。既然找不到一致性理由，這是否意味著我們需要對 AI 預測全面開放？

二、若可，又該如何利用以避免極端情況？(自主性原則)

事實上，即便我們允許此技術，仍可避免某些極端案例。由於上述的 AI 預測問題，在既有的法律與社會規範的框架下也會遇到，因此既有的倫理原則或許可以提供參考。舉例來說，生命倫理學中的「自主性原則」是指病人擁有決定是否接受治療的權力 (the power to decide)。如應用到 AI 預測上，或可說，決策者擁有是否將決定權讓渡給 AI 預測的權力。

在此原則下，人類需負擔最終決策。就算人類決定讓 AI 完全決策時也需要對於預測錯誤負責。畢竟，是人類決定交給 (完全授權) AI 來決策。換言之，這個原則確保人類仍是做決定的那個人，是最終決策的行為者。因此，大部分既有約束人類行為的法律與社會規範，仍然可以應用。舉例來說，如果某國家既有的法律架構不允許警察單位僅憑線報或情資 (而沒有證據) 就居留可疑人士 48 小時，則當「傳統線報」換成「AI 預測結果」時一樣不行。如果可以，則 AI 亦可。換言之，自主性條件也可以用來回答「人能否因自己還沒做的犯罪而被逮補？」如果某國既有的法律架構下大多時候皆不行，則不論是傳統線報或 AI 預測都不行。如果某些案例下可以則都可以，如此一來至少確保當出錯時能知道誰該負責。例如 2005 年英國國會通過反恐法案中不經審判拘留嫌犯，由原本的 14 天延

長到 28 天，但反對布萊爾政府提的 90 天。如果英國的國會經民主程序同意，則依自主性原則 AI 預測同樣適用。[18]

引進「自主性條件」有何優點？首先，以人類為最終決策者，可確保行為與責任相對。這在民主國家強調權責相符的責任政治尤其重要。其次，以人類為最終決策者，可保有人類執行並評估某決策的能力。例如在預測腫瘤上，如果全部交給機器決定 (精準度較高)，一旦人類醫師完全依賴機器判斷，則可能不重視人類判斷的相關訓練而逐漸失去此能力，亦無從知道機器究竟算不算準確。第三，以人類自主性作為最後關卡，可保有人類控制力。而人類控制力是心理安全感的重要來源之一。在新科技發展時，確保社會的安全感反而才能促進科技的發展。否則就會像 1865 年清帝國同治維新時，英國商人在北京永寧門外建鐵路，結果「京師人詫所未聞，劾為妖物，舉國若狂，幾致大變」，最後同治皇帝不得已只好把鐵路給拆了。[19]

肆、其他倫理爭議：隱私與安全

2017 年，史丹佛大學心理學教授 Micheal Kosinski 和 Yilun Wang 進行了一個爭議性實驗，他們發表一個可以透過辨識人的五官特徵來判定其性傾向的演算法。他們從線上約會網站蒐集了

---

[18] 依此，是否以 AI 事先介入 (是否處罰酒駕、能否事先拘禁可疑人士) 屬立法政策問題。

[19] 除自主原則之外，或也需考慮「比例原則」：風險預防措施的嚴厲程度必須與結果危害程度合乎比例 (例如刑法只處罰重大犯罪的未遂，危險犯或未遂犯的處罰也比實害犯或既遂犯來得輕)。此外，也必須考慮是否有同樣有效，但對人民權利干預或侵害程度更輕微的手段 (必要性原則)。關於危害程度的評估，以及相應的不同嚴格程度預防措施的匹配，反而可能是當前 AI 技術更有發揮餘地也更適合處理的問題。這點特別感謝匿名審查人之寶貴意見。

35,000 張白人 (以男性佔大多數) 的頭像照片作為訓練。當每位被分析者的照片達到五張時，他們的演算法在判定性傾向的準確率在女性上達到 83%，男性上則更有高達 91%的準確率。通常男同志擁有較窄下頜與較長鼻子。雖然研究者宣稱，其目的是提出警示讓人重視 AI 新科技對個人隱私與安全的危害。但此研究遭到不少抨擊，作者 Konsinski 教授甚至遭到死亡威脅 (Murphy, 2017; Schramm, 2018)。

這類抨擊主要有兩類。第一類認為，此研究之效度不足。例如在理論預設方面，人格特徵是否與五官特徵有關聯本不無疑問。而樣本偏差與統計誤報 (false positives) 同樣會削弱研究的結果。譬如該演算法主要以美國白人男性且公開性傾向之同志的相關資料進行訓練，故涵蓋範圍狹隘。此外，任何預測少數族群的模型都會面臨統計誤報的困擾。據紐約時報報導，威斯康辛大學的心理教授 William Cox 曾以該預測為例指出，假設全美國有 5%人口為同志 (即每 1,000 人中有 50 人是同性戀) 時，則準確率為 91% 的 AI 預測系統有 9%的機率會將異性戀誤當成同性戀 (即 85 人)，並將同性戀誤判為異性戀。這可能會造成在 130 人中被預測為同性戀者有 85 人實際上是異性戀 (Murphy, 2017)，誤判十分可觀。另外，此研究更忽略了雙性戀與性傾向流動的情況。第二類則批評此研究的結果會同時威脅 LGBTQ 與非 LGBTQ 的隱私與安全。例如同為史丹佛大學教授的 J. D. Schramm 就質疑，光是跨性別者在出生時的性別判定中，其面部特徵可能就讓他處於仇恨犯罪的高風險中 (Schramm, 2018)。尤其在 Yahoo 與 facebook 個資外洩事件之後，LGBTQ 社群的隱私侵害與工作、求學、生活等相關權利已受威脅。

當 AI 預測越精準，對人的隱私侵害的風險就越大。包括 Free Press、NAACP 在內的美國 41 個人權團體曾聯合發表公開信給美國

AI 公司 Axon，抨擊至今沒有任何政策或保障能有效減輕即時面孔辨識技術所帶來的威脅。為此，Axon 則設立倫理委員會來探討面孔辨識技術的風險 (Vincent & Brandom, 2018)。Google 也在旗下成立 DeepMind Ethics & Society 以評估並改善 AI 對社會的影響。同樣的，Faception 公司也意識到其面孔辨識技術的爭議性，故宣稱不會將這種能預測負面特徵的分類器 (classifier) 公開給大眾 (Tomlinson, 2016)。換言之，目前要降低 AI 透過面孔辨識來預測人類行為的潛在威脅，解決方式大約有兩類。一是限制「AI 技術」的發展或取得，另是限制「人類資料」的蒐集與使用。但基於兩個理由，後者似乎比前者更具可行性。首先，面孔辨識與預測技術的研究發展，有助於釐清臉部特徵與人類行為兩者的關連，究竟只是巧合 (如過去「骨相學」一樣被淘汰) 或同屬於更大的因果鍊的獨立結果 (兩者無直接因果關係)。[20] 其次，未來隨著技術的門檻降低與普及，任何人都可能擁有此類技術。故問題的關鍵，並不在限制此技術的取得與相關 AI 研發，而在於規範哪些對象的資料可以被蒐集、使用與保存。

要規範哪些對象的資料可以被蒐集、使用與保存，則可分為「個人」與「群體」兩大類來討論。首先，當對象為「個人」而預測的結果關乎個人利益時 (醫療診斷、投資理財、職涯規劃) 時，在不違反法律情形下，至少需當事人的知情同意。但是當預測的結果關乎公眾利益，但對當事人有害時 (如帶原者或嫌疑犯的追蹤)，則須依合法化程序 (如通過相關的通訊監察法，並依法申請)。其次，當對象為「公眾」時，其預測的結果應符合公眾利益。[21] 這裡的公眾又

[20] 例如植物光合作用和太陽能發電都受日照影響，兩者也有正相關，但並無因果關係。

[21] 當然，「何為公眾利益？」或「公眾利益彼此衝突時該當如何？」則又是進一步的問題。

可分為「普遍群眾」與「特定群眾」。普遍群眾譬如一社會之所有公民，對其之資料蒐集須經民主之合法化程序 (如立法、公投等)。特定群眾則如消費者、某選區選民、中低收入戶、兒童或老人等。諸如政府施政、公司行銷、學術研究等、球隊經營等都可能是潛在的 AI 技術使用者。而當對象的資料為公開時 (運動員或選戰對手歷年表現) 雖不需授權，但當資料非公開時，則需在符合法令前提下經各機構的倫理委員會 (Ethics Board) 或人體試驗委員會 (Institutional Review Board) 通過方可進行。

此外，除了隱私權外，AI 預測技術還可能衍生其他問題。例如已廣為人知的過濾氣泡 (filter bubble) 效應，社群媒體的演算法會根據使用者過去閱讀習慣，來預測喜好的貼文，使人只能將資訊取得的範圍限縮在同溫層中而造成孤立。而如果我們將焦點進一步放在將 AI 預測技術用於和人類行為有關的風險預測上，也可能加劇資源分配不公、貧富差距過大、資料外洩等問題。例如當某偏鄉平均壽命只有 60 歲，是否公部門需繼續投資老人醫療？掌握 data 的公司或國家將避免更大風險並浪費更少資源，是否使國家或企業間的差距拉大？更別提巨量資料的蒐集保存機構是否有能力保護所使用或儲存的資料？這些都是未來值得深探的重要問題。

伍、結論

簡言之，本文探討了將 AI 預測技術用於蒐集、使用與保存人類資料作為風險預測之爭議並分析相關的規範性考量。本文從「演算法相關」(第貳節) 與「資料相關」(第參、肆節) 兩方面來討論。前者從知識論角度探討此預測技術的結果可靠度與過程不可解釋性。後者則從倫理學角度分析人類資料使用上的可能爭議。

本文主張：(一) AI 預測的不可解釋性，並非來自於 AI 本身無法提供機械式的步驟，而是來自於人類的認知限制無法對數量龐大的步驟賦予意義，從而理解 AI 的決策過程。(二) 關於 AI 的歸納法、黑箱等問題並不是新的，在人類大腦上也會遇到，兩種智慧系統的差異，是程度上而非種類上的。(三) 在使用 AI 預測結果時，必然性與事實性條件並無法一致的反對極端案例卻又不反對現有法律或社會規範的案例。(四) 在使用 AI 預測結果時依自主性原則為佳，其優點在確保權責相符、避免喪失人類能力、降低 AI 發展的社會阻力。

參考文獻

Brown, H., & Friston, K. J. (2012). Free-energy and illusions: The cornsweet effect. *Frontiers in Psychology*, 3: Article 43. https://doi.org/10.3389/fpsyg.2012.00043

Burgin, M. (2004). Algorithmic complexity of recursive and inductive algorithms. *Theoretical Computer Science*, 317, 1/2/3: 31-60. https://doi.org/10.1016/j.tcs.2003.12.003

Burgin, M. (2005). *Super-recursive algorithms*. New York: Springer.

Church, A. (1935). Abstract no. 204. *Bulletin of the American Mathematical Society*, 41: 332-333.

Church, A. (1936). An unsolvable problem of elementary number theory. *American Journal of Mathematics*, 58: 345-363.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36, 3: 181-204. https://doi.org/10.1017/S0140525X12000477

Dennett, D. C. (1989). *The intentional stance*. Cambridge, MA: MIT Press.

Floridi, L. (2016). Should we be afraid of AI. *Aeon Essays*. Retrieved from https://aeon.co/essays/true-ai-is-both-logically-possible-and-utterly-implausible

Fodor, J. A. (2008). *LOT 2: The language of thought revisited*. Oxford, UK: Oxford University Press.

Hohwy, J. (2018). Prediction error minimization in the brain. In M. Sprevak & M. Colombo (Eds.), *Routledge handbook to the computational mind* (pp. 159-172). Oxford, UK: Routledge.

Hung, T.-W., & Yen, C.-P. (in press). On the person-based predictive policing of AI. *Ethics and Information Technology*.

Jones, N. (2017). How machine learning could help to improve climate forecasts. *Nature News*, 548, 7668: 379. https://doi.org/10.1038/548379a

Kleene, S. C. (1967). *Mathematical logic*. New York: Wiley.

Lin, Y.-T., Hung, T.-W., & Huang, T. L. (in press). Engineering equity: How AI can help reduce the harm of implicit bias. *Philosophy & Technology*.

Liu, Y., Racah, E., Correa, J., Khosrowshahi, A., Lavers, D., Kunkel, K., et al. (2016, May 4). Application of deep convolutional neural networks for detecting extreme weather in climate datasets. *ArXiv*. Retrieved from https://arxiv.org/abs/1605.01156

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York: Henry Holt.

Mozur, P. (2018, March 2). China presses its internet censorship efforts across the globe. *New York Times*. Retrieved from https://www.nytimes.com/

Murphy, H. (2017, October 9). Why Stanford researchers tried to create a "Gaydar" machine. *New York Times*. Retrieved from https://www.nytimes.com/

Nebehay, S. (2018, August 30). U. N. calls on China to free Uighurs from alleged re-education camps. *Reuters World News*. Retrieved from https://www.reuters.com/

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. New York: Basic Books.

Piccinini, G., & Bahar, S. (2013). Neural computation and the computational theory of cognition. *Cognitive Science*, 37, 3: 453-488. https://doi.org/10.1111/cogs.12012

Rivers, M. (2020, January 3). More than 100 Uyghur graveyards demolished by Chinese authorities, satellite images show. *CNN*. Retrieved from https://edition.cnn.com

Schramm, J. D. (2018, February 19). AI "gaydar" could compromise LGBTQ people's privacy—and safety. *Washington Post*. Retrieved from https://www.washingtonpost.com/

Swanson, L. R. (2016). The predictive processing paradigm has roots in Kant. *Frontiers in Systems Neuroscience*, 10, Article 79. https://doi.org/10.3389/fnsys.2016.00079

Thomas, D. (2018, July 17) The cameras that know if you're happy - or a threat. *BBC News*. Retrieved from https://www.bbc.co.uk

Tomlinson, S. (2016, May 24). Can you spot a terrorist just by lookingat their face? *Daily Mail*. Retrieved from https://www.dailymail.co.uk

Turing, A. M. (1937a). Computability and λ-definability. *Journal of Symbolic Logic*, 2, 4: 153-163. https://doi.org/10.2307/2268280

Turing, A. M. (1937b). On computable numbers, with an application to the Entscheidungs problem. *Proceedings of the London Mathematical Society*, S2, 42: 230-265. https://doi.org/10.1112/plms/s2-42.1.230

Turing, A. M. (1992). Intelligent machinery, report for National Physical Laboratory. In D. C. Ince (Ed.), *Mechanical intelligence*: *Collected works of A. M. Turing* (pp. 3-23). Edinburgh, UK: Edinburgh University Press.

Vincent, J., & Brandom, R. (2018, April 26). Axon launches AI ethics board to study the dangers of facial recognition. *The Verge*. Retrieved from https://www.theverge.com/

Wachter, S., Mittelstadt, 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/10.2139/ssrn.2903469

Yang, G. Z., Bellingham, J., Dupont, P. E., Fischer, P., Floridi, L., Full, R., et al. (2018). The grand challenges of science robotics. *Science Robotics*, 3, 14: eaar7650. https://doi.org/10.1126/scirobotics.aar7650

A Preliminary Study of Normative Issues of AI Prediction

Tzu-Wei Hung
Institute of European and American Studies, Academia Sinica
E-mail: htw@gate.sinica.edu.tw

Abstract

This paper focuses on normative aspects of AI prediction—that is, technologies used to predict the future through analyses of big data concerning the past. While this technology seems promising in forecasting extreme weather or rehabilitating endangered wildlife, it is controversial when applied to human beings, e.g., an Israeli company is using AI prediction to identify possible terrorists, and China's government to locate potential dissidents. This paper explores some of the normative issues and argues: (1) AI-derived conclusions are inexplicable not because machines fail to provide mechanical steps, but because our limited cognitive power cannot assign meaning to the, probably billions of, steps, and thus we fail to understand the conclusions reached by AI; (2) while AI is considered to have an inductive problem, to be a black box, and to have other epistemological issues, these worries apply to the human brain as well. AI and the human brain are different in degree rather than type; (3) the necessity argument and the reality condition cannot be used to exclude radical cases (e.g., China's social credit system) without excluding existing laws or social norms; and (4) the principle of autonomy has advantages, which include balancing power and responsibility, and reduces public distrust.

Key Words: normativity, induction, privacy, artificial intelligence, predictive technology