初探人工智慧中的個資保護發展趨勢與潛在的反歧視難題

摘要

在當今以大量資料之分析為基礎的第二波人工智慧發展浪潮下，個人資料的保護課題面臨前所未有的挑戰，也在應用上引發歧視的難題。個資依賴型人工智慧在其智慧學習階段，仰賴人類過去實際活動產生而留存的大量且多元資料進行訓練，利用機器學習的資料處理技術，歸納資料之間已知的規律性，或透過巨量資料分析探勘，發現未曾被現有知識掌握與理解的資料新關聯，藉此剖繪出具相類似身體或心理人格特徵、偏好、行為模式者之群體圖像，並發展可預測或評估其他符合該等群體圖像者的特徵與行為模式的演算法。智慧應用則利用該等演算法進一步蒐集目標對象之個人資料形成其個人圖像，與各種群體圖像進行比對，達到預測或評估該個人之目的。現行個資保護法制面對上述智慧學習階段之個資蒐集利用模式時，面臨資料蒐集難以仰賴事前同意、資料難以真正去識別、資料利用潛能難以事前預見的諸種困境。而機器學習所得之知識亦常掉入偏誤複製陷阱，且無力提出因果說明，以致在智慧應用上可能造成歧視的疑慮。但因演算法得以隱匿可疑分類與主觀歧視故意，更在機率預測或評估的有限目的下，可宣稱能以關聯性知識擔保手段目的間的合理性，而難構成恣意差別待遇。凡此均為人工智慧對現行各國個資保護與反歧視法制所帶來尚難以克服的挑戰。

關鍵字

智慧學習、智慧應用、群體圖像、個人圖像、個人資料保護、歧視

作者簡介

邱文聰，台大法律系法學組學士、法律研究所公法組碩士，美國賓州大學 LL.M.，美國維吉尼亞大學 S.J.D.。現任職於中央研究院法律學研究所副研究員、國立臺灣大學國家發展研究所合聘副教授。研究領域為憲法資訊隱私、科學與法律、研究倫理、醫療法、公共衛生與食品法等。

Evolving Issues of Data Protection and the Conundrum of Antidiscrimination in Artificial Intelligence

Abstract

While data-driven approach has empowered the development of the second wave of artificial intelligence (AI), personal data protection faces unprecedented challenges. AI application also brings about the thorny problem of statistical discrimination. With the assist of machine learning, the personal-data-dependent AI is now able to learn correlations among different human characteristics, preferences, traits, and patterns of human behaviors by analyzing big volume of existing data collected from varied human activities. The end result of such a learning is powerful algorithms that are able not only to produce different kinds of "group profiles" but also to match those "group files" with the "individual profile" of the targeted person so as to predict his or her human behaviors. The current data protection laws around the world, however, all encounter similar difficulties in bringing data collection and processing for AI learning purposes to terms. The volume and the variety of the data needed for AI development make it unfeasible to rely on prior consent as a legal ground. The impossibility to truly anonymize personal data and unforeseeable AI application uses further threaten the interests and rights of data subjects. While the balancing of legitimate interests is often adopted as an alternative ground, the seemingly beneficial knowledge about correlations produced by AI learning could easily fall prey to replicate of existing social biases without being recognized and taken into account. As a result, statistical discrimination that conceals suspect classification or animosity could be disguised as a rational means to simply predict or assess risk and probability and eschew the proscription of antidiscrimination laws. All these are challenges yet to be overcome that AI imposes on current data protection and antidiscrimination laws.

Keywords

AI learning, AI application, profiling, data protection, statistical discrimination

Author

Wen-Tsong Chiou is an Associate Research Professor of Institutum Iurisprudentiae, Academia Sinica and a Joint Associate Professor of Graduate Institute of National Development, National Taiwan University. His research interests include constitutional privacy, law and science, research ethics, medical and public health law, as well as food law.

一、緒論

人工智慧 (Artificial Intelligence) 是繼「雲端」、「物聯網」、「巨量資料」之後，台灣及各個標榜資訊技術創新尖端國的產官學界，近期最熱門的一個關鍵詞。人工智慧可以被理解為是一種以人類智能為類比之對象，進行思考、解決問題、做出決策或完成一定工作的人工自動化系統。

雖然人工智慧可以回溯自英國數學家圖靈 (Alan Turing) 於 1950 年代出版 Computing Machinery and Intelligence 一文，當中首次提出以機器模仿人類行動 ("acting humanly" approach) 的概念，開啟其後的各種討論與研究。但多年以來，不僅關於何謂「人工智慧」的定義缺乏共識，該領域的實際技術能力，也仍然與人類智慧有一段距離，使得多數人對人工智慧的理解，多停留在科幻電影的想像。直至最近資訊科技的大幅進展，運用巨量資料與機器學習的各種演算技術，將人工智慧推向前所未有的真實之境。例如，AlphaGo 在 2017 年當中，完勝世界第一棋王後，宣告退役，被譽為是人工智慧的標誌性進展；無人駕駛車的上路，證實了人工智慧應用於具體日常生活的可能性，也透露出「演算法社會」(Algorithmic Society)正悄悄降臨。

若從背後的理論工具演進來區分，人工智慧從開始到目前為止的發展進程，大體上可分為三個時期。[1] 第一波人工智慧主要以發展專家系統為主，將特定領域專業知識所蘊含的規則轉譯為電腦程式語言，結合各種使用者介面之硬體，成為可代替或協助專家針對限定任務做出判斷的人工智慧系統。[2] 第二波人工智慧則不再依賴人工輸入特定知識規則，轉而透過大量現存資料的統計分析，歸納其中的規律性。若將人工智慧技術開發分為智慧學習與智慧應用兩個階段，則第二波以降的人工智慧，在「智慧學習」階段，高度仰賴人類過去實際活動所產生而留存的大量且多元資料，作為「訓練資料」(training data)：人工智能主體 (intelligent agent)[3] 利用「機器學習」(machine learning)的資料處理演算法 (algorithms)，在適當的資料模型與參數調校的協助下，不僅可從「訓

[1] John Launchbury, A DARPA Perspective on Artificial Intelligence, https://www.darpa.mil/attachments/AIFull.pdf (2017).

[2] 至目前為止，規則取向（rule-based）的專家系統多數都是失敗的。在法律的專家系統中，少數成功的例子多集中於語法複雜但語義單純的領域。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, 24 (2017). 可參見 RICHARD SUSSKIND, EXPERT SYSTEMS IN LAW (Clarendon Press 1989); JOHN ZELEZNIKOW & DAN HUNTER, BUILDING INTELLIGENT LEGAL INFORMATION SYSTEMS: REPRESENTATION AND REASONING IN LAW (Kluwer, 1994).

[3] 亦有認為應稱為「人工智能代理人」，以其係「代理」人類進行感應、認知並做出決策甚或行動。但從 intelligent agent 本身有感應、認知、決策與行動能力的角度而言，亦具有自行施展權力或產生作用之主體地位。本文從後者，暫譯為人工智能主體。

練資料」中歸納資料與資料之間已知的規律性，並將此種規律性運用於訓練資料以外的資料集，產生推測新資訊的能力；以資料為基礎的機器學習甚至可透過巨量資料分析與探勘技術 (big data analytics, data mining)，發現未曾被人類原有知識所掌握與理解的資料與資料間新的關聯性。「智慧應用」則是進一步搭配各種符合具體應用情境的資訊蒐集技術（例如無人駕駛的自動環境資訊感應），取得當下乃至於實時 (real time) 的外在環境資訊，將之輸入前述藉由智慧學習所建立的智能演算法後，獲得推測結果，並藉此做出回應或作成行動決策。當然，「智慧學習」與「智慧應用」在概念上雖然有所區隔，在實際運用時，並非總是截然二分；二者可動態結合，使應用所得的資料再繼續反饋成為訓練資料，藉此不斷修正人工智能主體的人工智慧系統，以便透過自我演化而對應於最新的實際情境。

然而，第二波以資料為基礎的人工智慧仍面臨數個困境：歸納法本身無法對統計規律性背後的原因提供解釋，因此無法進行有意義的推理，甚至創造；以資料進行機器學習僅能就該等資料所屬的脈絡，訓練出限定領域的人工智慧，仍無法發展出泛用型人工智慧。下一階段人工智慧（第三波）的研發目標，即是發展更接近人類的強人工智慧系統，使其具有完全獨立於人類而自行運作、推理、決策、行動與反思能力。強人工智慧系統的誕生，勢必挑戰當代各種以人類為中心的倫理思考與法律制度。但在第三波人工智慧來臨前，上述第二波人工智慧已然帶來諸多倫理與法律難題。本文以下主要即以第二波人工智慧為討論對象。

二、個人資料在人工智慧中的用途

首應說明的是，（第二波）人工智慧雖常以人類過去實際活動所產生而留下的資料做為「智慧學習」的訓練資料以發展所需的演算法，也常在「智慧應用」時蒐集應用情境中的相關資訊，做為人工智能演算法的自變項輸入值，以得到相應的輸出值。然而，人工智慧所蒐集、處理與利用之資訊卻並非必屬於或來自於可識別個人身分的「個人資料」。例如，圖像辨識系統可從大量的公開圖檔與該圖檔的後設資訊中，歸納圖像特徵與圖像意義間之關係，以分辨乳牛或大麥町、捲曲的熟睡小狗或甜甜圈，而不涉及任何個人。人工智慧圍棋 AlphaGo 以公開之棋譜與過去之大量非屬個人資料之實戰記錄，甚至電腦自己與自己進行對戰的結果，做為訓練資料，反覆測試在特定局勢下不同走法的勝敗機率，以完成智慧學習，同樣與個人資料保護法制所關心的個人資料無涉。

其他又例如第二級 (Level 2) 以上可離手之自動駕駛車 (hands-off self-driving car)，[4] 除了從地圖資料中學習行車路徑之選擇外，也蒐集車輛行進中之外在環境資訊，包括行進中所遭遇之交通標線、標號、號誌及其他道路上固定或移動物體的即時狀態，建立車行自動反應模型；人工智慧翻譯透過自然語言的機器學習，從大量語句中建立不同語言語法之對應關係；[5] 天文物理學利用大量的物體撞擊與天體運動模擬資料進行機器深度學習，發展可預測小行星行進路線並評估如何避免其撞擊地球的人工智慧系統等，[6] 這一類不仰賴可識別個人資料或衍生自可識別個人資料之次級資料進行智慧學習，在智慧應用時也不需要與個人資料進行互動即可運作的人工智能主體，或可稱之為「非個資依賴型人工智慧」(non-personal data dependent AI)。由於「非個資依賴型人工智慧」不涉及個人資料之蒐集、處理或利用，在知識生產與知識應用上即完全不涉及「人」的因素，因此不僅不發生「個人資料保護」的法律適用問題，倘若該等人工智慧又不用於做成社會資源分配之決定，也不會有一般以「人」為應用對象之人工智慧所生之其他困難倫理爭議。所餘者雖仍有是否應賦予人工智慧在不同法律制度下的權利主體地位、若致生侵害時法律責任應如何歸屬、風險應如何分擔、事前應有如何之管制等問題，但並非本文之關注重點。

相對地，仰賴個人資料或仰賴自其衍生之次級資料進行智慧學習，或者在智慧應用上必須藉著與個人資料的互動與處理，始能得到輸出值並運作的人工智能系統，則可稱為「個資依賴型人工智慧」(personal data dependent AI)。從技術層次而言，「個資依賴型人工智慧」與「非個資依賴型人工智慧」並不必然存在本質上之差異。例如，人工智慧人臉辨識系統與人工智慧的一般圖像辨識，可能使用相同或類似的人工智慧技術，但人臉辨識無論是在智慧學習或智慧應用上，都與可識別之個人相關，因而產生一般圖像辨識所無的倫理與法律挑戰。當然，「個資依賴型人工智慧」亦可能在「非個資依賴型人工智慧」的技術之外，結合其他不同的人工智慧技術而產生更複雜的功能，例如，自駕車若在智慧學習階段以個別使用者之移動軌跡，做為自駕車在路徑選擇上的訓練資

[4] 根據 SAE International 的分級，依照人類駕駛人員於車行中需投入之注意與參與程度由多至少，可將自駕車區分為第一級至第五級，第五級之自駕車可完全無須人類之參與，而第二級則屬駕駛可離開方向盤，但仍須投入相當注意程度，始能確保安全行車的自駕車系統。SAE International, Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems, J3016_201609 (2016).

[5] Hany Hassan et al., *Achieving Human Parity on Automatic Chinese to English News Translation*, https://www.microsoft.com/en-us/research/publication/achieving-human-parity-on-automatic-chinese-to-english-news-translation/ (March 12, 2018); Ryan Daws, Microsoft's AI can translate Chinese on parity with a human https://www.artificialintelligence-news.com/2018/03/16/microsoft-ai-translate-chinese-human/

[6] Deep Asteroid: Predictive model of NEOs' trajectory using 'Deep Learning' and 'TensorFlow,' https://open.nasa.gov/innovation-space/deep-asteroid/

料，或者在智慧應用上同時蒐集眾多使用者即時之地理位置資訊，以解決交通阻塞問題，則可能產生個人行蹤監控的問題。

「個資依賴型」與「非個資依賴型」人工智慧的區分，主要乃在其所使用資料屬性之差異。相對於「非個資依賴型」,「個資依賴型」人工智慧必然均使用來自於個人身心活動所留下之紀錄，或者由該等紀錄進行「模糊化」或「去除直接識別性」處理後的次級資料。例如，人工智慧影像診斷系統的開發，在智慧學習階段必須從病患的 X 光片、電腦斷層掃描、核磁共振、超音波等各種影像資料中，判讀病灶圖像與疾病診斷間的關係，但病灶圖像的學習，卻可先單獨將該名病患之疾病診斷與各種相關之病歷資訊，登載為圖像的後設資料(metadata)，隨之即隱匿或刪除病患姓名或其他「可直接識別身分」的資訊，而達到「去除直接識別性」的效果。本文之所以不待探究訓練資料是否仍具有「間接識別可能性」，即一概將此等利用「去除直接識別性」之資料進行智慧學習的系統，仍歸類為「個資依賴型人工智慧」，其原因有二。第一，該等人工智慧在智慧學習階段所預設的終極目標，雖非從訓練資料中直接建立特定自然人之「個人圖像」，而是剖繪具有相類似身體或心理人格特徵、偏好、行為模式者之「群體圖像」(population profiles)，以發展可預測或評估其他符合該等群體圖像之自然人的特徵與行為模式的演算法，然而，在後端的智慧應用上卻必然仰賴蒐集並輸入特定自然人之個人資料形成「個人圖像」(personal profiles)後，比對系統中相符之「群體圖像」，進而預測或評估該個人之特徵、偏好、能力或行為模式，因此對「智慧應用的應用對象」而言，該等人工智慧自仍屬「個資依賴型」。第二，對於在智慧學習階段以其「去除直接識別性」之資料做為訓練資料的自然人而言，即使智慧學習之目的不在於建立該自然人之「個人圖像」，但針對其去除直接識別性之資料詳盡分析後所得的群體圖像，亦使得該自然人被間接識別的可能性不僅無法合理排除，更增其風險。因此，利用「去除直接識別性」之資料進行智慧學習的人工智慧系統，所具有引發倫理與法律爭議的可能性，仍較諸完全與個人身心活動記錄無涉之「非個資依賴型人工智慧」，有顯著之差異，而與使用直接識別性之資料進行智慧學習的人工智慧系統，有近乎相等的倫理與法律爭議性。

三、個資依賴型人工智慧的倫理與法律挑戰

如前述，在第二波人工智慧的框架下，個資依賴型人工智慧的基本運作，大體仍由智慧學習與智慧應用兩個階段所構成。而智慧學習與智慧應用則又各

自涉及資料的上游蒐集、中段處理分析與終端利用等三個流程。因此，個資依賴型人工智慧所引發的倫理與法律爭議，亦可由此兩階段三流程進行觀察。

以智慧學習為目的所蒐集的資料，可分為兩個來源。第一是專為特定的智慧學習活動而直接向資料當事人蒐集，第二則來自其他目的蒐集所得資料，以二次利用方式轉供智慧學習研發之用。由於第二波人工智慧高度仰賴巨量資料技術進行智慧學習，[7] 而巨量資料技術的數個 Vs 特性中，[8] 又以資料量龐大(volume)與資料來源多樣（variety）為其有別於其他資訊技術的特性。倘若為智慧學習目的，**直接**向數量眾多的資料當事人（主體意義上的大規模蒐集）**蒐集**多種來源或長期累積的資訊（客體意義上的大規模蒐集），即使先不論合法基礎的建制成本，也勢必需要額外投入龐大的資料取得成本。相對地，將各種智慧學習以外之目的所蒐集而留存之資料，以二次利用形式轉供智慧學習，若同樣暫先不論如何建制資料蒐集的合法基礎，則此等**間接蒐集**的方式確實可大幅節省資料取得成本。

在個人資訊自主與資料使用價值的權衡框架下，資料蒐集之合法基礎一般而言有包括契約關係在內的廣義當事人事前同意、符合比例原則之法律強制或任意規定、正當利益之個案權衡結果。為智慧學習目的而蒐集個人資料，無論是直接或間接蒐集與當事人相關的各種資料，若能徵得當事人事前同意，理論上應無太多法律或倫理爭議。然而，在智慧學習所需的巨量資料規模下，逐筆取得資料當事人之事前同意，對直接蒐集而言，除了取得資料之成本外，更多了取得事前同意的成本；對間接蒐集而言，轉換目的之前取得同意亦被認為將抵銷無須額外付出龐大資訊取得成本的優勢。這是智慧學習階段資料蒐集的**事前同意困境**。

然而，在立法者尚無從一般性地確認智慧學習獲致之公益，均可合乎比例地做為限制當事人權利之理由，而以個別立法授權強制蒐集、處理或利用之前，另一常見的策略則是藉資料去識別化的處理技術確保資料機密性，以補充單純以個案中的資料利用價值作為資料取得合法基礎的正當性。[9] 此種作為額外保障(additional safeguard)的資料去識別化處理技術，除了對「個別資料」的

[7] European Data Protection Supervisor, Artificial Intelligence, Robotics, Privacy and Data Protection 4 (2016).

[8] 巨量資料的特性，最早美國白宮報告採用 3Vs 的定義，包括 volume, variety, velocity。*See* Executive Office of the President, Big Data: Seizing Opportunities, Preserving Values 3 (2014). 後 IBM 則採 4 Vs，增加 veracity。之後又有各方主張第五個 V 為 value，第六個 V 為 variability，第七個 V 為 visualization 等，不一而足。

[9] ARTICLE 29 DATA PROTECTION WORKING GROUP, OPINION 06/2014 ON THE NOTION OF LEGITIMATE INTERESTS OF THE DATA CONTROLLER UNDER ARTICLE 7 OF DIRECTIVE 95/46/EC at 31, 33, 42-43 (2014).

直接識別資訊予以隱匿外，通常是針對資料庫中所有個人資料的「整體資料集」，進行隨機化 (randomization) 或統整化 (generalization) 的處理，使「資料集」中的各筆「個人資料」，無法被個別標定 (singling-out)、亦無法與同屬一人之其他資料進行串連(linking)、或進行與該人有關其他資訊之推論(inferring)。[10] 然而，由於智慧學習需要使智能主體在不同「資料集」之間，比對同屬一人的各種資料，以期建立資料間已知的規律性或發現新的關連性。因此，完全切斷個人資料的連結可能性，必不能滿足智慧學習的需要。然而，為了保留同屬一人之各種資料間的連結需要，而僅僅隱匿「個別資料」中的直接識別資訊再另以代碼標示之，充其量只是非真正去識別的「假名化」(pseudonymization)，而仍保有拼湊出個人圖像的完整能力。另一方面，在巨量資料的規模下，即使欲對「整體資料集」進行隨機化或一般化的處理後再予分析利用，也因為龐大的資料量與資料的多樣性所帶來，也不能擔保各該資料無法透過與其他資料組合比對，從而間接識別當事人之身分的再識別潛能，使得「無法標定、無法連結、無法推論」的真正去識別化效果不再可能達成。這是智慧學習階段資料處理的**去識別化困境**。

至於智慧學習階段的資料利用，對資料當事人而言也面臨一個新的挑戰。智慧學習利用巨量資料分析技術，往往可從多種表面上無相互關聯的資料中，建立可推論資料所屬當事人其他特性的演算法。例如，美國的研究人員曾利用臉書使用者在臉書上按讚的資訊，連同少量的訪調資料，即成功推論臉書使用者的「性傾向」達 88%的正確率、推論使用者之「種族」（白人或黑人）達 95%正確率、推論使用者之「宗教信仰」（基督或回教）達 82%正確率。[11] 上述例子不僅說明巨量資料的分析技術可從當事人的一般個資（按讚），推論當事人的敏感性個資（性傾向），打破一般個資與敏感個資的分野；更凸顯單一資料所蘊含的分析與推論潛能，在巨量資料分析技術的加持下，已大幅提昇而具有多能性 (pluripotency)，由此所生產出之「關聯性知識/演算法」，也脫逸出各該資料蒐集時當事人所處的特定脈絡。這是智慧學習階段資料利用潛能**難以事前預見的困境**。

除了上述從資料當事人的角度，檢視智慧學習在資料蒐集、處理與利用流程上對權利保障可能造成的困境之外，智慧學習階段透過個人資料所進行的知識生產，還帶來以下兩個挑戰。首先，智慧學習的知識生產（演算法的開發）

[10] ARTICLE 29 DATA PROTECTION WORKING GROUP, OPINION 05/2014 ON ANONYMISATION TECHNIQUES 9, 11-12 (2014).

[11] Michael Kosinski et al., *Private Traits and Attributes are Predictable from Digital Records of Human Behavior*, 110:15 Proceedings of the National Academy of Sciences of the United States of America 5802, http://www.pnas.org/content/110/15/5802.full.pdf (April 9, 2013).

有承襲既有偏見的危險。如前所述，第二波人工智慧的設計與開發者，已不再執著於只求將特定專業領域知識之抽象規則轉譯為電腦程式語言，轉而直接以人類過去反覆從事特定活動所產生的大量資料，作為統計歸納人類在不同條件下進行此等活動時如何思考、決策或行為的分析變項，進而建立一套足以重現該等資料變項間關係的演算法，達到直接模仿人類思考、決策或行為的目的。然而，藉由人類過去實際活動所產生而留下的資料作為智慧學習的對象與範本，雖然使人工智能主體得以快速達到心領神會人類實際思考、決策或行為模式的境地，但也因此可能在無意間將「訓練資料」本身所隱含的人類系統性偏見直接予以複製。[12] 而這樣的偏誤在機器學習的科技外觀下，卻可能更難以被察覺。這是智慧學習在知識生產上的**偏誤複製陷阱**。

其次，智慧學習在發展演算法的知識生產過程，仰賴巨量資料分析與機器學習技術，主要則是透過「歸納法」，將訓練資料中可能存在的各種資料關聯性予以普遍化。然而，歸納本身並無法對資料關聯性提出因果的解釋，也因此機器學習所建立的資料關聯性，除了在已知因果關係的情況下，可用來確認人工智能主體的智慧學習成果外，對於新發現的關聯性現象，若無其他理論提供可能的說明解釋，多半只停留在「知其然，而不知其所以然」的程度。尤其當機器學習利用大量多維度的資料進行非線性的複雜歸納時，更難期待現有人類知識能對演算法所呈現的關聯現象，提出合理的因果說明。這是智慧學習在知識生產上**無力提出因果說明的限制**。

在智慧學習發展出演算法後，智慧應用則進一步將演算法用於從一組已知資料或資料集合，判斷、推估與預測各種尚未掌握之資訊（包括未知但可知，與不可知之資訊）。而智慧應用階段的演算法運作雖同樣歷經資料蒐集、處理分析與終端利用等三個流程，但主要的爭議集中在智慧應用階段的資料利用。

不同於智慧學習為了開發演算法而需要對為數眾多的個人（主體意義上的大規模蒐集）取得與其有關的多種來源或長期累積之個人資料（客體意義上的大規模蒐集），演算法的智慧應用在資料蒐集的主體範圍上，並不一定需要進行大規模的蒐集。事實上，相對於演算法的開發目的在於生產集體人口知識，演算法目前的應用則多是個體化的。至於為演算法應用所為資料蒐集的客體範圍，則取決於特定演算法所需輸入的變項究竟為何。例如，不同的汽車保險費率智慧系統，有的僅需從投保申請書的單一來源，蒐集被保險人性別、年齡、職業、婚姻子女、居住地等有限的個人資料，即可進行靜態的費率演算，有些則可能蒐集飲食消費與醫療紀錄、運動習慣、行車行為模式等多元資訊，進行

[12] European Data Protection Supervisor, *supra* note 7, at 4.

更為動態與複雜的演算。為智慧應用相關目的而蒐集個人資料前，無論是直接或間接蒐集，均應有相應於蒐集目的之合法基礎。直接蒐集前，原則上並應告知當事人；即使演算法所需的變項輸入資料必須從多個來源「間接蒐集」，依照目前我國個人資料保護法第 9 條第 1 項規定，原則上也應於處理或利用前，向當事人告知資料來源以及包括蒐集目的在內的各種法定應告知事項。就此，智慧應用對資料蒐集的相關個資保護規範，似乎並未造成任何根本性的衝擊。真正的挑戰來自於資料經演算法處理後所連結的實際利用。

個資依賴型人工智慧的最大功能即是將演算法所建立的各類「群體圖像」(population/group profiles)，[13] 應用於與具體特定自然人之「個人圖像」(individual profiles) 進行比較配對，從而預測或評估該自然人之特徵、偏好、能力或行為模式，甚至進一步依據該等預測或評估做出自動決策或做為決策之參考。例如，依照顧客過去的不同消費模式與購買能力，預測其偏好，並根據此訂出相應的行銷策略（建議購買項目、寄送特定商品折價券、差異的定價方式）；依照申請入境者的各種特徵、旅行軌跡紀錄，預測其從事恐怖攻擊活動的可能性，並藉以決定是否核發入境許可；依照醫療影像、生理與基因檢測、遺傳病史等，進行人工智慧診斷。凡此均為個資依賴型人工智慧常見的智慧應用。然而，類似的智慧應用卻可能因為前述智慧學習在知識生產上的限制，而帶來新的挑戰。

首先，故意基於可疑分類 (suspect classification) 而為的差別對待 [14]，在演算法的開發與應用上常得以非可疑分類變項作為替身，逃避反歧視法的追究。例如，美國社會中的郵遞區號 (zip code)，可作為實際上存在於居住空間之種族隔離的替代指標，因此利用郵遞區號進行演算的智慧應用，即可將種族歧視包裝為非基於種族的差別待遇。智慧學習所發展的演算法，能透過更複雜的關聯性知識，將可疑分類隱匿為更難以察覺的其他非可疑變項。這是智慧應用**<u>得以隱匿可疑分類</u>**所造成平等權保障的適用困境。

其次，智慧學習在知識生產上的偏誤複製陷阱，可能使智慧應用延續社會既有的系統性偏見，卻因為欠缺主觀歧視故意(discriminatory intent)，而無法在僅將違法歧視限於「差別對待」(disparate treatment) 的反歧視法制下受到檢驗。[15] 例如，無意中複製了社會既存性別刻板印象或偏見的人力資源自動媒合

[13] *See* Brent Mittelstadt, *From Individual to Group Privacy in Big Data Analytics*, 30:4 Philosophy & Technology 475–494 (2017).

[14] 關於可疑分類在平等權保障下的意義，請參見司法院大法官釋字 727 號解釋湯德宗大法官協同意見書。

[15] 反歧視法的兩種制度模式，一為「差別對待」，另一為「差別影響」。美國主要採取需要主觀

演算法，雖然可能造成性別差異的媒合結果，但該等自動媒合決定卻不是基於傳統反歧視法意義下對特定性別的主觀歧視故意，而是演算法從現存性別分工現象中學習所得的自然演算結果。[16] 這是智慧應用因**欠缺主觀歧視故意**而可能造成反歧視法律的適用困境。

第三，智慧學習透過歸納法雖僅生產出「關聯性」知識，而無力提出因果說明，但並不因此使得應用演算法於社會生活決策，即當然喪失手段合理性而淪為恣意或非理性的行為。例如，統計上雖或可發現鯊魚攻擊與冰棒銷售額間，具有正相關的關聯性，但若欲減少鯊魚攻擊，並無法透過減少冰棒銷售來達成，欲提高冰棒的銷售也不可能透過增加鯊魚攻擊而實現，因此鯊魚攻擊與冰棒銷售二者雖具關聯性，但並不具因果關係。然而對不明其間各自之因果機轉，僅得二者表面關聯性之知識者而言，如其目的僅為了概估冰棒在一年當中消費需求的可能高峰，鯊魚攻擊不失為一可用的判準；同樣地，若為了概略預測鯊魚攻擊的可能發生時機，則冰棒的銷售額也確實是一個有效的指標。換言之，當特定智慧應用之目的並非使行動者透過演算法做出因果推論並依此進行「因果控制」，而是提供可供行動判斷的「機率預測或評估」時，關聯性知識即已足夠擔保手段目的間的合理性。也因此，利用演算法做出入學申請者未來學習表現的預測、求職者未來工作表現的預測、假釋聲請者的再犯可能性預測、貸款者還款能力預測等，即使僅建立在智慧學習所發現的關聯性知識之上，亦難謂為恣意或非理性的判斷與決策行為。平等權保障與反歧視法所禁止的「恣意」差別待遇（基於「與事物本質無關之事由實行差別待遇」），在演算法僅提供機率預測的智慧應用中，再無從想像。智慧應用在提供更廣泛的機率預測能力，卻**難以構成恣意差別待遇後**，將造成反歧視法律的適用困境。

第四，智慧應用雖能透過關聯性之知識，提供「合理的」機率預測能力，卻不見得能對關聯性的預測提出「有意義」的合理說明。人工智慧演算法若僅探索少數變項間的關聯性，其演算結果何以如此，或仍有對一般人進行有意義說明的可能。反之，倘若演算法透過機器學習將高度複雜且為數眾多的變項歸納為關聯性知識時，即使是演算法的開發者，往往也只能知其然而不知其所以

歧視故意的「差別對待」模式。我國反歧視法制是否及於不具備主觀歧視故意的「差別影響」，請參見黃昭元，論差別影響歧視與差別對待歧視的關係— 評美國最高法院Ricci v. DeStefano (2009)判決，《中研院法學期刊》第11期，頁1-63（2012年9月）。*See also* Solon Barocas & Andrew Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671 (2016).

[16] *See* Toon Calders & Indre Žliobaite, *Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures*, *in* DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY (Bart Custers et al. eds., Springer 2013).

然，而形成演算「黑箱」(algorithmic “black box”)。[17] 人工智慧系統開發者或使用者對演算結果無法提出有意義的合理說明，雖不因此影響演算法提供機率預測能力的「合理性」，卻降低對演算結果提出質疑、進行檢驗、做出修正和予以課責的可能，並使所有人成為「演算法之奴」(slave to algorithm)。[18] 這是智慧應用的**機率預測因抽離社會意義脈絡帶來的課責困境**。

四、因應人工智慧的個資保護與反歧視的比較法與學說發展及其限制

如上所述，個資依賴型人工智慧無論是在智慧學習階段或在智慧應用階段，均可能引發資料蒐集、處理或利用等面向的倫理與法律挑戰。近來，極力發展人工智慧的主要國家，均積極思考如何從法制面對於人工智慧的發展予以因應。其中尤以歐盟於 2016 年通過，預計 2018 年 5 月開始施行的「一般個資保護規則」(GDPR)，最具代表性。

相較於 1995 年歐盟個資指令 (DPD, EU Directive 95/46/EC) 立法之時，電腦網路與資訊技術仍處於發展初期，2016 年完成修正的 GDPR 則適逢巨量資料分析技術與人工智慧方興未艾之際。然而 GDPR 並未一般性地將巨量資料分析或以發展演算法為目的的「智慧學習」，當作資料蒐集、處理與利用的合法基礎。事實上，GDPR 的修正過程中，雖曾於 2012 年間一度將科學、歷史研究與統計目的當成獨立的合法基礎，但最後通過的版本則已明確放棄此一立場。經各方利害關係者折衝後通過的 GDPR 並未實質改變 DPD 的基本規範，反而重申了資料蒐集處理利用「應有合法基礎」(lawfulness of processing)、「目的特定與限定」(purpose specification and limitation) 及連帶而來的「目的相容性」(compatible purpose)[19]、「資料最小化」(data minimization)等原則，也維持了一般性個資與敏感性個資的區分。因此，即使巨量資料分析或演算法的智慧學習，已為當今世界各國資訊經濟發展得以推進的核動力來源，但除非具有包括契約關係在內的廣義當事人事前同意、[20] 或者符合比例原則之法律強制或任意

[17] 美國法學者 Frank Pasquale 將仰賴演算法進行各種社會活動的社會稱之為「黑箱社會」。FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (Harvard University Press 2015).

[18] *See* 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 (2017).

[19] GDPR, Article 5(1)(b).

[20] GDPR, Article 6(1)(a) and Article 6(1)(b).

規定、[21] 或者正當利益之個案權衡結果等，[22] 作為資料蒐集、處理與利用的合法基礎，否則仍難以進行。除此之外，目的特定與資料最小化的要求，也仍可能成為巨量資料分析技術企圖由多重來源間接蒐集大量個人資料的阻礙。職是之故，GDPR 的個資保護策略常被批評為並未滿足巨量資料及人工智慧的發展需要。[23] 然而，GDPR 仍有值得注意的幾個重點。

GDPR 在法條本文中延續 DPD 的作法，將科學、歷史研究與統計目的直接擬制為「非不相容於」原始蒐集目的。有論者並主張，GDPR 第 50 條前言既已提及，符合相容性原則之處理利用，無須再取得原始合法蒐集以外的其他合法基礎，因此原始目的外之科學、歷史研究或統計，同樣無須有額外的資料處理或利用合法基礎。[24] 然而，法規前言並不具規範效力，且上述解釋與過去 DPD 下之權威解釋機關的立場相左。依照 DPD 所成立之「第 29 條個資保護工作小組」 (Article 29 Data Protection Working Party) [25]的正式意見，相容之目的外利用 (compatible further processing) 本身仍需具備合法基礎，並不以其有相容性而免除亦應具備合法基礎的要求。[26] 此外，GDPR 最後通過的版本既然已否定科學、歷史研究或統計可獨自成為資料蒐集處理利用的一般性合法基礎，理應在目的外利用時，也維持一貫的規範立場。

究竟科學、歷史研究與統計本身可否構成目的外利用時之獨立合法基礎，此一條文解釋爭議如何解決，雖仍待後續觀察，但可以確定的是，GDPR 與 DPD 本就提供其他合法基礎，不必有廣義或狹義當事人事前同意，即可進行原始或目的外之研究或統計利用：其一為法律之明文授權，其二為正當利益之個案權衡 (legitimate interests of the controller)。法律明文授權係由立法者依據比例原則進行通案權衡後，明訂可無須當事人事前同意的各種強制或任意性資訊作為的特定目的、發動條件、資料類別、處理程序、儲存期限等。[27] 目前為止歐盟各國雖或有針對法定統計與諸如傳染病防治、癌症防制等具重大公益之特定議題研究，訂有強制資訊蒐集、處理或利用的規定外，尚未見一般性地將人工智慧學習或巨量資料分析，當作可進行強制資訊作為的理由。反之，「正當利益之

[21] GDPR, Article 6(1)(c), Article 6(1)(e), and Article 6(3).
[22] GDPR, Article 6(1)(d) and Article 6(1)(f).
[23] Tal Z. Zarsky, *Incompatible: The GDPR in the Age of Big Data*, 47 SETON HALL L. REV. 995 (2017).
[24] *See*, *e.g.*, Nikolaus Forgó et al., *The Principle of Purpose Limitation and Big Data*, *in* NEW TECHNOLOGY, BIG DATA AND THE LAW: PERSPECTIVES IN LAW, BUSINESS AND INNOVATION 17, 37-39 (M. Corrales et al. eds., 2017).
[25] 該工作小組在 GDPR 正式施行後，將直接轉為正式的「歐盟個資保護會」(European Data Protection Board)。
[26] Article 29 Data Protection Working Party, Opinion 03/2013 on purpose limitation 28 (2013).
[27] GDPR, Article 6(3).

權衡」則由資料控制者，於個案中考量利用目的之正當性與必要性、手段之侵害最小性等因素後，即先行為資料之蒐集處理或利用。但依據 DPD 的規定，必須賦予當事人對於資料控制者的權衡結果，可於事後提出異議的機會 (right to object)。[28] 而「第 29 條個資保護工作小組」更明確要求，以資料控制者之正當利益為由而蒐集處理或利用個資時，仍應盡可能採行包括資料加密、假名化、賦予當事人無條件退出權 (general and unconditional right to opt-out) 等保障措施 (additional safeguards)[29]，以平衡當事人權利保護之需要。GDPR 雖然在資料經假名化處理的前提下，進一步允許會員國為研究或統計目的，得立法限制當事人的異議權，[30] 但並未否定以「正當利益之權衡」為合法基礎的一般個案，仍應考量當事人自主權維護的可能性。[31] 換言之，無論是在 DPD 或 GDPR 的架構下，不僅可以為了研究或統計目的，[32] 而以符合比例原則的方式，透過立法限制當事人之事前同意及（或）異議的權利，也仍可在足夠的利用正當性支持下，允許以當事人之「事後退出」取代「事前同意」，進行科學、歷史研究或統計。凡此均為 GDPR 在資料蒐集處理或利用的「合法基礎」上，得以用來克服智慧學習階段所面臨資料蒐集難以取得**事前同意**的可行途徑。

此外，GDPR 雖然並未就巨量資料分析技術所帶來**資料無從真正去識別化的困境**，提出革命性的解決方案，但 GDPR 特別提出 DPD 所未明示的「假名化」(pseudonymisation) 概念，確認「假名化」處理後之個資，仍屬於 GDPR 所規範的可識別資料，[33] 也同時釐清「假名化」僅是作為「目的相容性」判斷 [34] 與一般資料處理 [35]時的一種必要或補充的資訊安全手段，而非只要經「假名化」處理，即自動取得蒐集、處理或利用的合法基礎。因此，並非資料只要經過「假名化」處理，就當然能合法地將之用於智慧學習的巨量資料分析與演算法開發。假名化之資料仍須有獨自的合法基礎，方能成為蒐集處理或利用的對象。GDPR 的「假名化」無疑是在承認具利用價值之個人資料已無從真正去識別化的現實下，將巨量資料分析的資料利用問題重新拉回到「合法基礎」來討論

[28] DPD, Article 14(a).
[29] Article 29 Data Protection Working Party, Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC, *at* 3, 10 (2014).
[30] GDPR, Article 89(2).
[31] GDPR, Article 21(1).
[32] 依據 GDPR Recital 159, GDPR 所稱之「研究」廣泛地包括科技發展 (technological development)、示範 (demonstration)、基礎研究 (fundamental research)、應用研究 (applied research)、私人資助的研究 (privately funded research)等，理論上可包含為智慧學習目的所進行的演算法開發。
[33] GDPR, Recital 26.
[34] GDPR, Article 5(1)(b), Article 6(4)(e), and Article 89(1).
[35] GDPR, Article 32(1)(a).

的一個努力。

再者，GDPR 在特定條件下免除間接蒐集前的告知義務，亦解決智慧學習階段資料利用潛能**難以事前預見**的問題。GDPR 雖要求間接蒐集或目的外利用個資前，原則上均應提供資料當事人相關之資訊。但倘若間接蒐集是為了研究或統計之利用目的，又當資訊的提供在客觀上不可能達成或其達成必須投入顯失比例的成本時，告知義務即可被免除。在 GDPR 免除告知義務的規定下，智慧學習階段所面臨巨量資料分析的資料利用潛能難以事前預見的問題，將被自動取消。留下的將是因無從事前預見智慧學習階段將生產出何種關聯性知識，而無從對之進行控制的無助資料當事人。

雖然 GDPR 確實並未特別迎合巨量資料分析技術或人工智慧學習的知識生產需要，而打造一個完全以知識經濟發展為導向的歐盟個人資料保護法。但整體而言，GDPR 所建構的法制基礎建設，對巨量資料分析與智慧學習所欲進行的知識生產活動，仍是採取相對容任的態度。毋寧，GDPR 將規制人工智慧的主戰場，設定在智慧應用的階段。

GDPR 大體延續了 DPD 對「與個人有關的自動化決定」(automated individual decision) 的規定，[36] 賦予個人有權可以拒絕「純粹以自動化的方式」做成與其相關且產生法律或類似效果的決定。[37] 此外，GDPR 也首次引入了「剖繪」(profiling) 的法律用語，用來指涉藉由自動化資料處理技術，將特定自然人之資料用於建立可供評估、分析或預測諸如該人之工作表現、經濟狀況、健康、個人偏好、興趣、可信賴性、行為、所在位置或行動軌跡等人之諸元的個人剖繪圖像。GDPR 的「剖繪」橫跨了本文所稱智慧學習階段「群體圖像」的自動化知識生產，以及智慧應用階段將特定自然人之「個人圖像」與「群體圖像」進行比對分析後所進行的推論與預測。[38] 但真正使個人有權直接拒絕的「剖繪」，則必須是與「自動化決定」有關，也因此限於智慧應用意義下的「剖繪」。單純透過將個人進行分類而生產群體圖像的「剖繪」，則僅適用前述關於一般個資蒐集處理利用的規範。[39]

GDPR 在個資依賴型人工智慧的智慧應用上，透過肯認受「純粹以自動化方式」做成個人相關決定影響的當事人，得要求資料控制者之「人為介入」(human intervention) 並提供解釋 (right to explanation)，藉以解決智慧應用之演

[36] DPD, Article 15(1).
[37] GDPR, Article 22(1).
[38] Article 29 Data Protection Working Party, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, at 6-8 (2017).
[39] *See id.* at 17-25.

算法可能因恣意差別待遇而違反平等原則的問題。然而，當演算法所提供者俱為「合理的」機率預測，又因其機率預測在複雜的機器學習過程中可能已抽離社會意義脈絡而無從進行有意義的說明困境下，GDPR 所架構的「人為介入」與「解釋權」能否真正解決智慧應用所帶來的平等保障與反歧視法適用問題，恐值商榷。

相對於此，美國法學界在論述如何解決智慧應用所帶來的潛在歧視問題上，除了主張應從傳統禁止「差別對待」的主觀歧視理論，轉變為禁止「差別影響」的客觀歧視理論之外，[40] 學者亦開始思考反歧視法制以外的可能手段。例如耶魯大學法學院 Jack Balkin 教授即提出將環境保護法制中「公害管制」的制度模型，套用於處理智慧應用對整體社會造成外部性成本的問題，得以克服傳統仰賴侵權行為法模式之反歧視法的思考侷限。[41] Balkin 之理論雖具有相當的說服力，但仍有待進一步發展為可操作的具體規定。

五、結論

當今以資訊為驅動力(data-driven)的第二波人工智慧，為人類帶來各種新的可能性。其中個資依賴型之人工智慧，藉由來自眾多個人身心活動所留下之歷史紀錄，透過巨量資料分析技術與機器學習，對作為學習樣本的訓練資料進行各種群體的分類與歸納，進而發展出可生產各種關聯性知識的演算法。掌握此一演算法將得以對特定自然人進行身心狀況、行為、偏好、能力等特質的評估與預測。然而，該等個資依賴型人工智慧，不僅在智慧學習階段對傳統個人資料保護法制造成衝擊，也在智慧應用階段帶來平等權保障與反歧視法的新課題。比較法制上雖多將人工智慧的挑戰，設定在智慧應用階段的潛在歧視爭議，而較為開放地容許智慧學習階段的知識生產。然而，目前為止的法制因應手段，似乎都尚未能妥適地回應人工智慧應用透過演算法所帶來的挑戰。重新回到智慧學習階段的知識生產，檢視造成問題之根源，或許才是克服此一困境的根本之道。

[40] Barocas & Selbst, *supra* note 15.

[41] Jack Balkin, *The Three Laws of Robotics in the Age of Big Data*, 78 OHIO ST. L.J. 1217, 1232-40 (2017).