人工智慧與公共行政

李翠萍*

機器是否感覺上比人腦更正確、更有效率、更中立而不帶偏見？當您進行三位數以上的加減乘除心算，是否會以計算機驗算才安心？當手機默默記錄主人往返住家與工作地的時間與路線，並在某日主人開車上班前，跳出訊息提醒預計到達工作地的時間，您是否覺得這項服務特別貼心？隨著 1980 年代機器學習（machine learning）技術突破，2000 年代出現大數據（big data）成為機器學習的基礎，與 2011 年深度學習（deep learning）技術精進（Russell & Norvig, 2021: 42-45），人工智慧（Artificial Intelligence，以下簡稱 AI）已悄然進入人類生活，並在政府部門扮演重要的角色，藉由其處理與辨識巨量資料的能力，為政府部門提升行政規模、降低成本與提升服務品質（Young, Bullock, & Lecy, 2019）。

當 AI 技術進入政府部門，廣泛應用於刑事司法、警察執法、醫療照護、國土安全與國境管理、教育、國家財政、公共就業、國防各領域（李翠萍、張竹宜、李晨綾，2022），行政效率的提升確實有目共睹。今年 8 月 BBC 一則有趣的報導指出，法國政府利用電腦視覺 AI 找出超過 2 萬個未合法申報的住家游泳池，[1] 因而替政府追討回大約 1 千萬歐元的稅收（BBC News, 2022）。相較於私人企業，公共政策攸關公共利益與社會正義，因此政府部門利用 AI 技術時更應謹慎為之。近年來，公共行政學者開始關注 AI 在公共政策過程的使用，以下先簡要介紹 AI 技術，再說明近幾年來新興的研究主題。

一、強 AI 與弱 AI

AI 的定義有狹有廣，依照其與人類能力的距離來看，可有強、弱之分，能力與人類相當的通用人工智慧（Artificial General Intelligence, AGI），屬於強 AI（strong AI），目前最接近的設計是如 Apple 的虛擬助理 Siri，但其未完全符合

* 國立中正大學政治學系教授

[1] 依照法國法律，因住家增設游泳池會提高財產價值，因此需要多繳財產稅（BBC News, 2022）。

AGI 標準，僅屬擦邊球。另一種強 AI 是超越人類的超級人工智慧（Artificial Super Intelligence, ASI），目前尚未研發出來。至於弱 AI（weak AI）則只能依照人類設定的目標來完成工作，例如贏一盤棋或辨識圖像，限制領域人工智慧（Artificial Narrow Intelligence, ANI）便屬此類（Kavlakoglu, 2020）。目前應用於公共行政領域者只是 ANI，大都用於決策輔助，必須在人（官僚人員）機（ANI 系統）互動之下，才能影響民眾福祉，在此，官僚人員是決策過程的關鍵中介者，他們對於 ANI 的態度，例如依賴、信任、警覺等，影響公共利益甚鉅。

二、公共行政與 ANI 決策輔助系統

Herbert Simon 認為，決策是行政的核心，而行政理論的語彙應該來自人類選擇中的邏輯學與心理學（Simon, 1997），此說法在公部門大量應用數位科技作為決策輔助工具的今日，深具意義。ANI 從歷史大數據學習，在人類的目標指令下，對標的對象進行預測、判斷，或提出警示，以作為決策參考。在動態複雜的環境條件下，ANI 更能顯出其高效率的特質，而人類也能因此擺脫大量而重複的工作，專注於更有創造性、更貼近人性的思考與設計。然而，官僚人員作為民眾與 ANI 的中介，其面對機器的心理狀態（信任、依賴程度）與對機器產出的推論詮釋，決定了 ANI 如何影響民眾福祉。以下將從決策階段、人機互動、政策結果與 ANI 治理，檢視公共行政研究的幾個新議題。

（一）決策階段中的 ANI

官僚人員的決策過程可簡可繁，端賴問題的結構與所鑲嵌的環境系絡，亦即，任務內容的複雜性與不確定會影響公部門對 AI 的偏好（Bullock, 2019）。在公共領域中，高度結構化與複雜性低的問題，或可依賴完全自動化（full automation）的 ANI 來處理，但低度結構化與複雜性高的問題，雖較難以完全自動化，但 ANI 仍能提供決策輔助的功能。然而，ANI 如何適用於各決策階段，是一個值得深究的問題。一般而言，在決策做成之前，決策者至少會經歷以下幾個階段，分別是問題界定，以及替選方案的訂定、評估與選擇，其中涉及的行動，包括掃描與蒐集資訊、確認與詮釋法定規則、與預測各替選方案的執行結果，而 ANI 在不同階段的適用性並不相同（Etscheid, 2019）。

以問題界定階段為例，一般而言，問題的產生來源至少有二：一種是來自組織外部，例如民眾的申請案件；另一種是由官僚組織內部提出，例如經過人員判斷認為必須解決的問題。以 ANI 的適用性來看，後者可能比前者更適合

ANI 的輔助。何以如此？一般民眾申請案件結構性較高，只要把審核指標與相對應的權重置入傳統演算法中，就能執行大部分的審核工作。至於由官僚人員主動發現的問題，通常是在大量且多元資訊之間進行分析、比對、彙整之後發現的，若能倚賴具有學習能力的 ANI，由其蒐集、分類、評估來自各方的資訊，並與目標狀態進行比較，在現狀偏離目標值時，及時警示相關人員，並提供資料作為後續決策參考，必能更有效率地預防問題產生（Etscheid, 2019）。研究發現，公部門管理階層比非管理階層更願意使用 AI 作為決策輔助工具（Huang, Kim, Young, & Bullock, 2021），顯然這與管理階層所處理的問題具有高度複雜性有關。由此可知，ANI 在各階段帶來的機會與挑戰，是公共行政領域值得深究的問題。

（二）ANI 對官僚行為的影響

不論官僚人員在前述哪一個階段使用 ANI 輔助決策，都會面臨人機互動的問題。人類在面對一個具有正確、效率與中立形象的機器時，對於機器給予的建議，會有至少兩種判斷偏差。第一是基於社會心理學推論而來的自動化偏差（automation bias），即過度依賴機器；第二是基於公共行政研究中有關偏差的訊息處理（information processing）而來的選擇性偏差（selective bias），即基於既有的成見而選擇性依賴機器。自動化偏差發生於官僚人員過度依賴 AI 提供的建議，即使其他資料來源顯示出相互矛盾的訊息，但決策者仍不為所動。這種現象特別容易發生於傳統重度依賴自動化工具的領域，例如醫療照護。這一方面是因為人類自認不如機器那般正確，進而放棄自己的判斷；另一方面可能是一種認知上的惰性（cognitive laziness），覺得倚靠 AI 就不必費神蒐集資訊或做判斷（Alon-Barkat & Busuioc, 2022）。

選擇性偏差則與人類對特定人、事、物的刻板印象有關，指的是人類在面對 AI 提供的資訊或建議時，會傾向於尋找、詮釋、接受符合自己既有成見或信念的資訊或建議。事實上，我們面對來自人類的資訊或建議時，就已經會有選擇性偏差的傾向，而在機器中立的表象下，人們不用負擔歧視或偏見的指控，更可能出現選擇性偏差。此外，機器像是一個道德緩衝器（moral buffer），使決策者容易產生心理上的距離感，認為有中立的機器撐腰負責，而忽略了人類在決策過程中所應該扮演的角色、負擔的責任，以及道德感（Alon-Barkat & Busuioc, 2022）。既然官僚人員是民眾與機器之間的中介，公共行政研究就無法忽略這些議題。

（三）ANI 系統產出的影響

公共行政決策結果關乎公共價值，因此如何利用 AI 提升公共價值（public value）也成為公共行政學者關注的議題（黃心怡、曾冠球、廖洲棚、陳敦源，2021）。在公共行政與公民之間的關係裡，公平（equity）是其中一個重要的公共價值（Jørgensen & Bozeman, 2007: 369），以下從公平的視角檢視 ANI 的正負面影響。

1. 正面影響——降低社會不平等

ANI 技術有助於提升特定群體的福祉，也能積極防弊，使社會趨向公平。前者例如協助視障者辨識環境，降低其移動障礙（Alashkar et al., 2020），或是提供醫療服務至資源匱乏地區（Wahl, Cossy-Gantner, Germann, & Schwalbe, 2018），後者則是藉由 AI 學習與分析大數據的能力，針對不公平現象提供警示，例如利用 AI 技術矯正不平等，提升工作環境中的群體多樣性（Daugherty, Wilson, & Chowdhury, 2018）。誠如 Tito（2017）主張的，潛藏於人腦中的偏見難以現形，但存在於演算法中的偏見則較容易被識出，所以 AI 技術能貢獻於社會公平。

2. 負面影響——導致非意圖歧視

使用 ANI 於公共政策可能導致特定群體遭受不平等的對待，此種現象稱為 ANI 的「非意圖歧視」（unintentional discrimination），又稱為「非意圖潛在歧視」（unintentional proxy discrimination），指的是 ANI 沒有故意歧視的能力，卻造成歧視的結果，而問題乃源於機器學習所使用的歷史數據。機器學習是依照人類給予的目標指令，從看來毫無關聯的變數關係中找出通往目標的捷徑，做精準預測，而往往正確性越高的 AI，其產出結果越難以被解釋清楚。機器學習所倚賴的大數據，是人類歷史的縮影，也是人類社會各種判斷與決策的集合。歷史中存在著的價值觀與偏見，與社會結構性因素導致的不平等，皆潛藏於大數據中，藉由機器學習而成為 ANI 判斷準則（Borgesius, 2018; Prince & Schwarcz, 2020）。

雖然 ANI 已應用於許多政策領域，但目前有關 ANI 非意圖歧視的相關研究，仍屬刑事司法、警察執法、醫療照護三領域最多（李翠萍等，2022）。在刑事司法方面，ANI 被用來進行罪犯的風險評估，預測再犯或累犯的可能性，進而影響刑期與假釋判決（Howard & Borenstein, 2018; Koepke & Robinson, 2018）。相關文獻中被討論最多的，是一套由 Northpointe 公司開發並廣受美國地方法院使用的「以替代性制裁為目標的罪犯矯正管理分析」（Correctional

Offender Management Profiling for Alternative Sanctions, COMPAS）。這套系統對不同群體的判斷產生偏差，非裔「偽陽率」（false positive rates）比白人高，「偽陰率」（false negative rate）則比白人低（Brenner et al., 2020; Chouldechova, 2017），[2] 而這些誤判導致有色人種被判處較長刑期（Cockerill, 2020）。在警察執法領域，ANI 被利用於預測犯罪，但研究發現其不成比例地關注有色人種（Madden et al., 2017），導致特定群體被逮捕的機率較高（Lum & Isaac, 2016; Saunders, Hunt, & Hollywood, 2016）。而在醫療照護方面，有研究顯示，ANI 為病患進行風險評估時，評估結果的正確性在不同群體之間有所差別（Winter & Davidson, 2019），像是非裔病人被評估為高風險的機率低於白人（Takshi, 2021），而且非裔病人的風險等級可能被低估，而白人則被高估（Obermeyer et al., 2019）。

（四）治理觀點下的制度設計

ANI 決策輔助系統常被認為是個黑箱，其系統產出難以被完全回溯或解釋，因此使用 ANI 的風險就是難以控制。那麼，決策者面對一個無法被解釋的系統產出，如何判定該資訊的可參考性？更進一步的問題是，基於 AI 建議而做成的決策一旦傷害民眾福祉，該如何追究責任？ ANI 作為決策輔助系統，有沒有可能被蓄意或非蓄意地誤用？要回答這些問題，需釐清 ANI 的風險並從中思索有效的治理原則。Wirtz, Weyerer, & Kehl（2022）為建構一個 AI 治理架構，從 AI 治理文獻中，整理出 AI 相關的六大風險類別與相應的指導原則，分別是技術、資料與分析類（technological, data, and analytical）、資訊與溝通類（informational and communicational）、經濟類（economic）、社會類（social）、道德類（ethical）與合法與監管類（legal and regulatory）。依照不同類別的風險定義與影響範圍評估，落實相應的指導原則於政府部門中，將抽象的原則具體化，成為治理措施或法定規則。由於 ANI 已應用於各政策領域，公共行政學者必須在 ANI 治理上，加快研究步伐，更應前瞻性的為 AGI 與 ASI 進入政策領域做準備。

三、結語

在數位科技發展快速的時代，ANI 已經逐漸深入各類公共政策領域，而公共行政學者也開始正視 ANI 在官僚決策中所扮演的角色與造成的影響。作為決

[2] 偽陽即被判斷為高風險但卻沒有再犯，偽陰是被判斷為低風險卻再犯。

策輔助系統，ANI 不僅提升行政效率，使公務人力不再耗費於繁瑣而重複的工作，也能有效率地服務社會弱勢，促進社會公平。然而，歷史大數據中所潛藏的結構性不平等與人類長久以來的歧視與偏見，使得 ANI 的使用更具有挑戰性，在人機互動之下，官僚人員如何降低自動化偏差與選擇性偏差，以對 ANI 的非意圖歧視提高警覺，並進而規劃適合 ANI、AGI、ASI 應用於公部門的治理機制，將是未來公共行政領域的重要課題。

參考文獻

李翠萍、張竹宜、李晨綾（2022）。〈人工智慧在公共政策領域應用的非意圖歧視：系統性文獻綜述〉，《公共行政學報》第 63 期，頁 1-49。

黃心怡、曾冠球、廖洲棚、陳敦源（2021）。〈當人工智慧進入政府：公共行政理論對 AI 運用的反思〉，《文官制度》第 13 卷第 2 期，頁 91-114。

Alashkar, R., M. ElSabbahy, A. Sabha, M. Abdelghany, B. Tlili, & J. Mounsef. (September 2020). AI-Vision Towards an Improved Social Inclusion. *Conference: ITU International Conference on Artificial Intelligence for Good (AI4G)*, Geneva.

Alon-Barkat, S., & M. Busuioc. (2022). Human–AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice. *Journal of Public Administration Research and Theory*, muac007. https://doi.org/10.1093/jopart/muac007

BBC News. (2022). Undeclared pools in France uncovered by AI technology. https://www.bbc.com/news/world-europe-62717599

Borgesius, F. Z. (2018). *Discrimination, Artificial Intelligence, and Algorithmic Decision-Making*. Strasbourg, FR: Directorate General of Democracy, Council of Europe.

Bullock, J. B. (2019). Artificial Intelligence, Discretion, and Bureaucracy. *The American Review of Public Administration,* 49(7): 751-761.

Cockerill, R. G. (2020). Ethics Implications of the Use of Artificial Intelligence in Violence Risk Assessment. *The Journal of the American Academy of Psychiatry and the Law,* 48(3): 345-349.

Daugherty, P. R., H. J. Wilson, & R. Chowdhury. (2018). Using Artificial Intelligence to Promote Diversity. *Sloan MIT Management Review,* 60(2), Retrieved September 28, 2022, from https://sloanreview.mit.edu/article/using-artificial-intelligence-to-promote-diversity/

Etscheid, J. (2019). Artificial Intelligence in Public Administration. In: I. Lindgren, et al. Electronic Government. EGOV 2019. *Lecture Notes in Computer Science,* vol 11685, pp.248-261. Springer, Cham. https://doi.org/10.1007/978-3-030-27325-5_19

Howard, A., & J. Borenstein. (2018). The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity. *Science and Engineering ethics,* 24(5): 1521-1536.

Huang, H., K. C. Kim, M. M. Young, & J. B. Bullock. (2021). A Matter of Perspective: Differential Evaluations of Artificial Intelligence between Managers and Staff in an Experimental Simulation. *Asia Pacific Journal of Public Administration,* 44(1): 47-65.

Jørgensen, T. B., & B. Bozeman. (2007). Public Values: An Inventory. *Administration & Society,* 39(3): 354-381.

Kavlakoglu, E. (2020). AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What's the Difference? *IBM Blog,* https://www.ibm.com/cloud/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks

Koepke, J. L., & D. G. Robinson. (2018). Danger Ahead: Risk Assessment and the Future of Bail Reform. *Washington Law Review,* 93: 1725-1807.

Lum, K., & W. M. Isaac. (2016). To Predict and Serve? *Significance,* 13: 14-19.

Madden, M., M. E. Gilman, K. Levy, & A. E. Marwick. (2017). Privacy, Poverty and Big Data: A Matrix of Vulnerabilities for Poor Americans. *Washington University Law Review,* 53: 53-125.

Obermeyer, Z., B. Powers, C. Vogeli, & S. Mullainathan. (2019). Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. *Science,* 366: 447-453.

Prince, A. E.R., & D. Schwarcz. (2020). Proxy Discrimination in the Age of Artificial Intelligence and Big Data. *Iowa Law Review,* 105: 1257-318.

Russell, S., & P. Norvig. (2021). *Artificial Intelligence: A Modern Approach (4th Edition, Global Edition)*. Harlow, UK: Pearson Education Limited.

Saunders, J., P. Hunt, & J. S. Hollywood. (2016). Predictions Put into Practice: A Quasi-Experimental Evaluation of Chicago's Predictive Policing Pilot. *Journal of Experimental Criminology,* 12: 347-371.

Simon, H. A. (1997). *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization*. New York, NY: Simon and Schuster.

Takshi, S. (2021). Unexpected Inequality: Disparate-Impact from Artificial Intelligence in Healthcare Decisions. *Journal of law and health,* 34: 215-251.

Tito, J. (2017). *How AI Can Improve Access to Justice*. London: Center for Public Impact.

Wahl, B., A. Cossy-Gantner, S. Germann, & N. Schwalbe. (2018). Artificial Intelligence (AI) and Global Health: How Can AI Contribute to Health in Resource-Poor Settings? *BMJ Global Health,* 3(4). Retrieved September 28, 2022, from https://gh.bmj.com/content/3/4/e000798

Winter, J. S., & E. Davidson. (2019). Big Data Governance of Personal Health Information and Challenges to Contextual Integrity. *Information Society,* 35: 36-51.

Wirtz, Bernd W., Jan C. Weyerer, & Ines Kehl. (2022). Governance of Artificial Intelligence: A Risk and Guideline-Based Integrative Framework. *Government Information Quarterly,* 39(4), 1-17.

Young, M. M., J. B. Bullock, & J. D. Lecy. (2019). Artificial Discretion as a Tool of Governance: A Framework for Understanding the Impact of Artificial Intelligence on Public Administration. *Perspectives on Public Management and Governance,* 2: 301-313.