・公共行政學報・　第六十三期
〈研究論文〉
民111年9月　頁1-49
國立政治大學公共行政學系
Doi: 10.30409/JPA.202209_(63).0001

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*

李翠萍、張竹宜、李晨綾**

《摘要》

本研究從米勒的多元正義觀出發，基於公民聯合關係中的平等原則，檢視人工智慧（AI）在公共政策領域應用所引發的倫理問題。本研究採質性後設分析法，依照 PRISMA 模式篩選學術研究論文，從中梳理 AI 在先進國家政策領域應用時的制度過程與結果。研究發現，AI 已應用於刑事司法、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、國防等八大領域，雖為政府部門帶來行政效率並提升整體民眾福祉，但同時也對特定群體造成非意圖歧視。從制度過程來看，政府部門忽略了用於機器學習的大數據中潛藏著長久以來的社會不正義，而從制度結果來看，歷史中的不正義透過 AI 繼續複製，導致特定群體遭受差別待遇，基本人權遭受剝奪。

為了分析各領域中非意圖歧視的樣態與問題本質，本研究以國際人權相關公約所隱含的人權保障優先順序，從「被歧視者是否主動接受評量」

投稿日期：111 年 3 月 2 日；接受刊登日期：111 年 9 月 25 日。
* 作者感謝匿名審查人與編委會提供寶貴的修改建議，修改過程中獲益良多，特此致謝。本研究承蒙國科會研究計畫 110-2423-H-194-003 提供研究經費補助，謹致謝忱。
** 李翠萍為國立中正大學政治學系教授，e-mail: tsueyping.lee@gmail.com。
張竹宜為國立中正大學政治學系大學部學生。
李晨綾為臺中市立臺中女子高級中等學校學生。

與「消極與積極權利的剝奪」兩個面向分析 AI 對特定群體造成的負面影響。分析結果顯示，AI 在警察執法、刑事司法、與醫療照護三大領域的應用涉及生命權與自由權等消極權利的剝奪，確實有優先處理的急迫性。本文於結論處討論何以非意圖歧視的矯正無法依賴公民社會的自覺，而必須由政府部門積極干預，並從 AI 應用的籌備階段與執行階段，建議政府應有的具體作為，以降低非意圖歧視對特定群體帶來的人權危害。

[關鍵詞]：人工智慧、科技倫理、非意圖歧視、科技正義、社會公平

壹、前言

「人工智慧」（Artificial Intelligence, 以下簡稱 AI）與數位科技帶來效率與便利進而提升社會福祉的同時，卻也引發倫理疑慮，特別是特定群體所受到的差別待遇逐漸受到關注。科技通常帶給人公正客觀的印象，在演算法尚未普遍進入人類生活的 1980 年代，美國麻省理工學院心理學家特爾蔻（Sherry Turkle）在學生間作的一個調查發現，白人學生對電子法官多採保守態度，而非裔學生則認為電子法官不會有膚色偏見，所以比較正義（Cohen, 2018）。然而，事實並非如此。AI 倫理學者格布魯（Timnit Gebru）曾對「谷歌」（Google）公司的大型語言模型提出警告，[1] 因其使少數群體被嚴重邊緣化（Schiffer, 2020）。不當使用 AI 甚至可能加劇社會不公平，例如 2018 年美國「亞馬遜」（Amazon）公司發現 AI 在協助該公司篩選求職者時，會非蓄意剔除履歷表中出現女性相關字眼的候選人（Dastin, 2018），此種現象，被稱為「非意圖歧視」（unintentional discrimination），又稱為「非意圖潛在歧視」（unintentional proxy discrimination）（Prince & Schwarcz, 2020）。

隨著 1980 年代技術突破的「機器學習」（machine learning），2000 年代「大

[1] 大型語言模型泛指為了進行「自然語言處理」（Natural Language Processing）所開發出的各種需要使用極大量資料進行訓練的模型。在本文中，指的是 Timnit Gebru 與他人合著的研究論文 Blender, Gebru, McMillan-Major and Shmitchell (2021) 第 611 頁 table 1 中所指的大型語言模型。

數據」（big data）的出現，與 2011 年「深度學習」（deep learning）的技術精進（Russell & Norvig, 2021: 42-45），AI 不再只是高科技「仿真」或「擬人」的機器或演算法而已，他是「行動體」（agent），能為了達到人類設定的目標，掃描外環境變化而調整行為，未來甚至可能發展出自我意識或自主性。[2] 近年來，國際間湧現 AI 道德規範的文件，[3] 但人類企圖設定的道德規範似乎趕不上科技的快速發展，這類文件至今尚未針對 AI 類型的研發順序提供指引（甘偵蓉、許漢，2020: 248）。有哲學家開始討論「哪一種」AI 必須受到規範，並主張禁止研發具有目標自主性的 AI（陳瑞麟，2020），而許多頂尖科學家與企業龍頭已提出警告，人工智慧的潛在危險不容忽視，人類在追求 AI 的效率時，應釐清 AI 在研發與應用上所涉及的描述性倫理與規範性倫理（彭錦鵬，2020；魯俊孟，2020；楊惟任，2018）。

在 AI 成為顯學之後，被各國政府逐漸應用於公共領域，從管理、決策、執行、到成果，衝擊著各國公共治理，如何審慎應用 AI 以提升政府行政效能，成為政府關切的課題（丁玉珍、林子倫，2020；韓釗，2019）。數位科技為政府帶來高效率，除了部分取代耗時或例行性的人工決策，也提供資料作為人工決策參考。然而，AI 系統在許多已開發國家公共領域的應用，已引發不正義現象，危害基本人權與社會公平。黃心怡、曾冠球、廖洲棚、陳敦源（2021）認為，政府部門應從公共價值、倫理道德風險與行政裁量權行使等公共行政核心議題，檢視 AI 在公部門的應用。

使用 AI 於公共領域所導致的人權危害，是當今人文社會學者亟需追趕研究的問題，為深入了解 AI 應用於公共領域對特定群體產生的差別待遇，本研究利用「質性後設分析」（qualitative meta-analysis），基於米勒（David Miller）多元正義觀中的社會系絡視角，從各國文獻檢視 AI 相關技術的應用對特定群體產生的非意

[2] agent 拉丁語源是 “to do”，意味著去進行某種行動（Russell & Norvig, 2021: 21）。根據韋伯大字典的解釋，agent 不限指人，也可以是物、工具、或電腦應用程式（https://www.merriam-webster.com/dictionary/agent）。國內大都將其翻譯為代理人（例如台灣人工智慧行動網 https://ai.iias.sinica.edu.tw/glossary/intelligent-agent/）。本文將之翻譯為行動體，除了反映其 “to do” 的意涵以外，也避免使人有等同人類之感。

[3] 例如歐盟的 The Assessment List on Trustworthy Artificial Intelligence, OECD 的 OECD Principles on Artificial Intelligence, 新加坡的 A Proposed model artificial intelligence governance framework, 英國的 A guide to using artificial intelligence in the public sector, 日本的び人間中心の AI 社会原則等。

圖歧視，討論各領域中歧視樣態的本質，以及對基本人權的危害程度，藉此檢視各領域在矯正不正義上的急迫性。

自古以來，正義原則百家爭鳴，而米勒的正義觀相當務實，主張正義必須落實於社會，否則流於空談，他不認為單一的正義原則能適用於各種社會系絡，因此他關注不同的社會系絡與其所適用的正義原則。對於本文而言，要以正義原則檢視 AI 在公共領域應用時所產生的影響，必先確認公共領域的屬性。而由於公共政策的執行對象是公民，公民身份構築了政治社會中每一份子之間的關係，每一位公民在政治社會中都該擁有被平等對待的權利，因此，米勒所謂三大社群類別之一的公民聯合關係，以及此社會系絡中所適用的平等原則，成為本研究的切入視角。此外，米勒是制度主義正義論者，其社會制度是建構社會正義的重要元素。AI 在數位時代裡以制度的形式存在，制度中有規則與程序，而 AI 系統的產出已逐漸成為各公共領域的決策參考，因此本文以米勒的平等原則，檢視 AI 作為制度的一部分，其制度結果所呈現的不正義。在此特別強調，本文無意質疑 AI 在公共領域的貢獻，而是檢視數位科技時代中的社會正義危機，提醒 AI 科技快速發展下在道德辯證上的缺乏。

貳、人工智慧的定義

根據 2019 年「電機電子工程師學會」（Institute of Electrical and Electronics Engineers, 簡稱 IEEE）董事會（IEEE Board of Directors）的定義，AI 是「使機器擁有智慧的活動，而所謂的智慧，是使某物能夠適當運作並在環境中具有能預測的遠見」，[4] 而 AI 的計算技術，是從人類與其他生物對事物的「感知」（sense）、「學習」（learn）、「推理」（reason）、「採取行動」（take action）所啟發的（IEEE, 2019）。人類創造 AI 不只是為了模仿人類，其終極目的是做出比人類更快速、正確的決策與行動，包含「理性思考」（thinking rationally）與「理性行動」（acting rationally），因此 AI 的基礎理論來自哲學、數學、[5] 經濟學、神經科學、心理學、電腦工程學、控制理論與模控學、語言學等（Russell & Norvig, 2021: 19-35）。

[4] "Artificial Intelligence is that activity devoted to making machines intelligent, and intelligence is that quality that enables an entity to function appropriately and with foresight in its environment."

[5] 演算法（algorithm）起始於數學計算，源於第九世紀數學家(Russell & Norvig, 2021: 21)。

從技術區分 AI 的功能，可分成狹義及廣義兩類，狹義指涉處理人類預先編程演算法的特定活動，[6] 廣義則指系統可以在多項任務中表現出像人類般的先進行為，並顯示有合理程度的自我理解與自主控制（Johnson, 2019: 429），不論狹義或廣義，演算法是發展 AI 的基礎，其運作過程必須處理複雜的結構化或非結構化數據（陳敦源，2022；Dafoe & Journal of International Affairs, 2018: 121; Bannister, Connolly & Grimmelikhuijsen., 2020: 471）。[7] 另一種分類是以 AI 與人類能力的距離來區分，分別是「限制領域人工智慧」（Artificial Narrow Intelligence, ANI）、「通用人工智慧」（Artificial General Intelligence, AGI）、與「超級人工智慧」（Artificial Super Intelligence, ASI）。其中，ANI 屬於「弱人工智慧」（weak AI），用以完成特定工作，例如贏一盤棋或辨識圖像，而 AGI 與 ASI 則屬「強人工智慧」（strong AI），前者與人類能力相當，後者則超越人類。虛擬助理如 Apple 的 Siri，屬於 AGI 的擦邊球，尚無法完全符合 AGI 的標準，而 ASI 目前尚未研發出來（Kavlakoglu, 2020）。目前，使用於公共政策領域者尚屬 ANI。

人工智慧與「數據科學」（Data Science）的結合，使機器學習成為可能。機器學習是 AI 的次領域，系統從大量數據中學習，運用統計方法與演算法，推導出規則來解釋或分類數據，並進行預測，數據量越大，預測準確度越高（Johnson, 2019: 429）。而神經網路與深度學習則是機器學習的次領域，二者都是模擬人腦運作，模仿生物神經元（節點）彼此之間發送訊號的方式（張國恩，2000）。神經網路與深度學習之間最大的差異在於節點的層次（深度），只要節點層次小於等於三層（包含輸入與輸出），就是神經網路，多於三層就是深度學習（Kavlakoglu, 2020）。

從前述的整理可知，人工智慧與數據科學兩大領域重疊之處，形成了機器學習次領域，而神經網路與深度學習又是機器學習的次領域。由於本研究主要目的是搜尋 AI 應用於先進國家公共政策領域中所產生的歧視樣態，因此於文獻搜尋時，主要以概念範圍較廣的 AI 與「演算法」（algorithm）兩詞為主，相關說明將於後續研究方法段落中詳述。

---

[6] 演算法是計算機編程中的基本流程，用於解決問題、進行預測等實現特定目標的系列步驟，演算法會按著邏輯執行其功能，是一個結構化的過程（Bannister et al., 2020: 472; Criado et al., 2020: 453; Sales, 2020: 47）。

[7] 以醫療照護資料為例，姓名、年齡是結構化數據，醫療報告或臨床紀錄是非結構化數據（Johnson, 2019: 429）。

叁、文獻檢閱

一、人工智慧在公部門的應用

隨著深度學習技術的突破，近十年來，AI 對公部門的影響開始受到重視與討論，近期的重要研究例如 Huang、Kim、Young 與 Bullock（2021）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；Bullock（2019）討論 AI、裁量權、與官僚三者之間的關係，認為任務內容的複雜性與不確定影響了公部門對 AI 或人力的偏好；Young、Bullock 與 Lecy（2019）針對由 AI 輔助的行政裁量與純粹人類的行政裁量進行比較，發現 AI 能提升行政規模、降低成本、與提升品質，但同時也引發了公平、可管理性、與政治可行性的問題。

除了 AI 與公務人員之間的互動研究之外，如何利用 AI 提升公共價值（public value）也成為學者關注的議題（黃心怡、曾冠球、廖洲棚、陳敦源，2021）。公共價值中的「公共」指涉由多元利害關係者所組成的公共場域（Moore, 1995; 2013），寬廣的定義可等同於一個社會（Bozeman, 2007）。所謂的「價值」，是被視為對社會有價值的結果（Moore, 1995; 2013），具體而言，是對公共場域的政治、社會、經濟、文化、環境有加分的貢獻，有利於社會運作（Benington, 2015）。Jørgensen & Bozeman（2007: 369）盤點相關文獻中公共價值的涵義，其中在公共行政與公民之間的關係（relationship between public administration and the citizens）裡，有四個重要的公共價值，分別是合法性（legality）、公平（equity）、對話（dialogue）、與使用者導向（user orientation）。其中，合法性與公平二者之間緊密連結，合法性意味著公共行政與公民之間的關係需要由法律規範，特別是在保護個人權利、平等對待、人類尊嚴、社會涵容等面向更是。而唯有藉由法律制度的規範，正義才有可能達成，公平的公共價值才可能持守。Head、Brian 與 Alford（2015）認為，對於公共行政而言，如何重新檢視公共價值的本質，從公共組織的內、外部思考科技帶來的正、負影響至為重要。

有關 AI 在公共領域應用的相關文獻後設分析很少，Wirtz、Langer & Fenner（2021）分析 189 篇研究論文，發現相關研究大都聚焦於治理與行政，討論現有政府結構如何適應 AI 而改變，但 AI 在具體領域的應用與如何改變結構以因應 AI，則較少受到關注。Reis、Santo 與 Melão（2019）採取「系統性回顧和後設分析的偏

好描述項目」（Preferred Reporting Items for Systematic Reviews and Meta-Analyses, PRISMA）搜尋研究論文，從 79 篇研究（包括 18 篇期刊論文與 61 篇研討會論文）檢視 AI 對公共行政的影響，但該文僅些微觸及 AI 的道德問題。Sousa、Melo、Bermejo、Farias 與 Gomes（2019）亦採 PRISMA 蒐集文獻，針對 59 篇期刊論文進行後設分析，檢視 AI 在公共領域應用的狀況與未來發展。研究發現 AI 在公共服務、經濟事務、與環境保護三個領域的應用最受關注，且人工神經網路技術在許多公共領域的應用效果極佳，該文針對 AI 道德的討論僅限於課責，即民眾對 AI 的作為不滿意時要如何要求外國的 AI 開發公司負責？Zuiderwijk、Chen 與 Salem（2021）亦採 PRISMA 蒐集文獻，從 26 篇期刊論文檢視 AI 在公共治理應用上的意涵，從研究內容、品質、途徑、應用主題來分析，並提供未來研究在程序與內容上的建議。

前述有限的後設分析研究所涵蓋的討論面向太廣，並無聚焦於 AI 在實際應用場域上的道德問題。其中，Reis 等人（2019）的研究稍微觸及倫理議題，即國防領域使用 AI 技術殺人時所引發的爭議與大數據中所呈現的偏差歧視，但並未深入討論。因此本研究期望利用後設分析貢獻於 AI 倫理的討論。

二、智慧數位科技如何降低社會不平等

關於智慧數位科技如何降低社會不平等的討論，可分為以下兩類。第一是 AI 技術有助於提升特定群體的福祉，使社會趨向公平。例如利用 AI 技術協助視障者閱讀、辨識環境、降低移動障礙（Alashkar et al., 2020）；提供醫療服務至資源匱乏地區（Wahl, Cossy-Gantner, Germann, & Schwalbe., 2018）；解決水患、提升農耕效率、提供乾淨用水，使飽受水患、農作物病蟲害、與無乾淨用水者得以受惠（Goralski & Tan, 2020）；降低勞動力的薪資不平等（Webb, 2019）；針對學生進行客製化課程設計，嘉惠不同類型的學生（Aguilar, 2018）；協助政府與民眾之間的有效互動，使公共服務更適合各類民眾的需求（Androutsopoulou, Karacapilidis, Loukis & Charalabidis, 2019）。

第二類是 AI 分析大數據的效率與能力勝於人類，而且不像人類大腦會潛藏偏見，因此更能促進社會公平。Jora、Sodhi、Mittal 與 Saxena（2022）的實驗研究發現，公司篩選聘僱新員工時若使用 AI，會使公司更能達到員工多樣性、平等、與包容性的目標，因為 AI 具中立性，會平等對待每一位應徵者，不像人類大腦中潛藏太多下意識的偏見。Cohen（2019）的研究也提到，AI 在協助徵人時，可以只關

注應徵者是否具有公司要求的特質，而不受到其他因素干擾。此外，AI 還可以積極防弊。例如，Daugherty、Wilson 與 Chowdhury（2018）的研究發現，可以利用AI刻意矯正不平等，提升公司中的女性員工比例。Tito（2017）認為，AI能提升社會公平並促成社會正義，因為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揪出來。

三、智慧數位科技的無意圖歧視

意圖（intention）是採取行動前的一種心理狀態，連結了慾望（desire）、信念（belief），當一個人意圖做某件事時，勢必知道自己正在做這件事，以及為何這樣做，所以意圖的行動勢必朝向某個目標（Setiya, 2018）。以目前的技術程度來看，AI 無法發展自己的意圖，系統裡的意圖是由人類設計的，從 Jonker et al（2002）可知，慾望、信念、意圖都可以透過程式設計被置入 AI 系統中。以棋賽為例，程式設計者給AI一個目標----贏得賽局，就形同把這個意圖置入AI系統中。

既然 AI 沒有自己的意圖，那麼又何來所謂的「非意圖」歧視？人類在道歉時說「我不是故意的」，這隱含著他有「故意」的能力。以此類推，若說 AI 是非意圖歧視，難道表示 AI 有故意歧視的能力？事實上，本文指的是，AI 沒有故意歧視的能力，卻造成歧視的結果（Prince & Schwarcz, 2020）。何以如此？答案在於歷史資料庫。AI 的機器學習是從大數據中近乎蠻力找出看來毫無關聯的變數關係，雖然無法解釋變數何以相關，但卻能進行比人類更精準的預測，而且預測越精準的AI，越難以被解釋。也因為如此，資料中所存在的歧視或偏見，勢必由 AI 學習而得（Prince & Schwarcz, 2020）。簡言之，AI 的非意圖歧視源於人類數十年來的各種決策中所隱含的歧視（Borgesius, 2018），長期以來社會結構性因素造成的不平等，皆隱身於數位資料中。

資料去識別化或禁止使用敏感性資料（例如種族、性別），是否就能使 AI 減少歧視？事實上，這對於具有學習能力並忠於目標的 AI 而言，效果有限。誠如前段所述，表面看來毫無關聯的變數，AI 會從中找出複雜的連動關係，所以 AI 有能力交叉比對非敏感性資料，進而做出針對性的結論（Gillis & Spiess, 2019; Kleinberg, Ludwig, Mullainathan & Rambachan, 2018）。

如果人類與 AI 都可能有偏見，那麼誰會造成較嚴重的傷害？從傷害預防的角度來思考，答案如下。首先，由於人類不可能純然客觀中立，因此我們對於人類偏見存有戒心，即使潛藏於大腦中的偏見難以被察覺，但基於偏見而呈現的態度與行

為卻容易被觀察出來，也因此較容易受到規勸、指責或制止。其次，由於機器容易給人客觀中立的印象，因此我們對於 AI 的警覺相對較低，反而不容易察覺應用 AI 技術過程中的偏見，通常要出現系統性的傷害，累積足夠的受害者，才會產生警覺。當然，這不表示 AI 的偏見完全無法預防。誠如（Jora et al., 2022: 1687）所言，「要去除偏見的第一步是要意識到偏見的存在」。在 AI 逐漸應用至公共領域各層面時，如果人類能意識到不正義的存在，並有意願提前預防，那麼就能設計檢核機制，針對 AI 的決策結果進行定期且系統性的分析，以避免複製不正義。

四、系絡主義下的平等原則

在米勒的《社會正義原則》（The Principles of Social Justice）一書中，連結抽象的社會正義與實務，主張於追求社會正義時必須考慮社會系絡，否則容易淪為空談，因為落實正義時充滿歧見。他拒絕採取單一的正義原則，探索人們在不同場境中，判斷正義與不正義時所使用的原則，發現在不同的社會系絡中，人們用以判斷正義的原則具有差異性，該原則會引導、限制系絡成員的態度與行為（Miller, 1999）。

社會正義應關注的範圍為何？米勒認為，只要有人認為某種資源分配不公，就是社會正義應關注的範圍，所以應傾聽民眾的聲音。但社會中有一些基本的資源分配，是本來就應受到關注的，例如所得、財富、工作機會、教育機會、以及健康照護等（Miller, 1999: 11）。接下來的問題是，社會的範圍又是什麼？社會是具有界限的社群，有既定的邊界、成員、制度、與能夠改變制度的機構，範圍大者如國家，小者如工作場所，都可謂之為社會，所以在這些具有界限的社群中都可以追求正義。然而，如何為這些社群進行分類並應用不同的正義原則呢？米勒以人際關係的差異性為社群進行分類，分別是（1）團結性社群，如家庭及宗教團體（2）工具性聯合關係，如民眾在市場中的關係、企業成員之間的關係及政府公務員之間的關係（3）公民聯合關係，如政治社會中每一份子之間的關係（梁文韜，2005a），這三類社群所應用的正義原則不同。

米勒主張的三大正義原則分別是「應得」（desert）、「需要」（need）及「平等」（equality），而這些原則分別適用於不同的系絡：團結性社群適用需要原則；工具性聯合關係適用應得原則；公民聯合關係適用平等原則（梁文韜，2005a；2005b；2005c）。在類似家庭關係的團結性社群中，社群成員彼此之間關係緊密並為彼此著想，資源可依照成員的需要原則來分配。工具性聯合關係是一種

市場經濟關係，適用應得原則，米勒舉例提到，社群成員若因其貢獻而獲得獎勵，表示其努力的過程中所耗費的成本能因獎勵而抵銷，若社群中每個成員都如此，這就是一種平等（Miller, 1997: 225）。而在公民聯合關係中，社群成員就是一種類似公民的身分（例如俱樂部的會員或一國的公民），基於公民身分，可以要求在公共領域中被平等對待，例如平等的合法保護、投票權、社會福利權等等（Miller, 1997: 230）。

米勒認為，社會制度是建構社會正義的重要元素，所以前述的正義原則，並非直接用來主導資源分配，而是用以檢視制度。首先，什麼是制度？制度是人類一種有規律的活動，在其中，人們各自有須要完成的工作、活動方式、與不同的權利與責任。制度包括各式各樣的規則、程序、與成員的規律行為。一個社會有多麼正義，端看該社會中的主要制度遵循正義原則的程度，而透過改善制度，就能導引出正義的分配結果（梁文韜，2005a；2005b；2005c）。

本研究聚焦於 AI 在公共政策領域的應用，是基於米勒的系絡主義與制度主義正義論。首先，基於系絡主義的正義論，公共政策執行的場境是以民眾為政策對象，比較接近於三種人際關係中的公民聯合關係，適用平等原則。公民在各種公共政策領域中，不論是在基本人權、醫療照護、社會福利、教育機會、工作權上，都應被平等對待。然而，平等對待並非意指每個人得到相同數量的好處，因為在實務上我們很難定義所謂的相同數量是多少，而是群體間不應有差異性（Miller, 1997: 230-231）。因此，本研究在思考 AI 於公共政策領域中的應用所導致的結果，也是以特定群體是否遭受差別待遇為主。其次，基於制度主義的正義論，本文視 AI 的應用為制度的一部分。誠如前述，AI 系統被應用於醫療服務、社會福利、刑事司法等領域中，其產出已經成為醫務人員、法官的決策參考，形同制度中的規則與程序。本文將以平等原則檢視 AI 作為制度的一部分，是否遵守平等原則，並如同米勒所言，期望藉由調整制度，能降低社會不正義。

肆、研究方法

後設分析研究法發展早期以量化研究為主，其後逐漸加入質性研究的元素，採用歸納、紮根途徑（李仲彬、陳敦源、蕭乃沂、黃東益，2006）。本研究採質性後設分析法，或稱「質性後設彙整法」（qualitative meta-synthesis），這是一種針對第一手資料或原創研究結果所進行的「次級分析」（secondary analysis）。採用後

設分析的研究目的通常有二，一是廣泛了解同一主題的既有研究結果，二是評估不同研究方法如何影響同一主題的研究結果（Timulak, 2014: 481）。本研究主要目的為前者，即從既有的研究文獻中，廣泛探索 AI 在各領域應用上已出現的不正義現象，從中萃取不正義現象的基本元素，並將這些原創的研究結果予以概念化，彙整詮釋並延伸討論（Schreiber, Crooks & Stern, 1997: 314）。質性後設分析的兩種知識論途徑，一是描述性的，從原著的眼光來描述研究結果，二是闡述性的，從後設分析者的觀點來詮釋原著的研究結果（Timulak, 2014: 486），本研究在知識論途徑上，採取後者。

研究資料搜尋時使用三類關鍵詞，分別是 AI、倫理、與公部門。AI 類的關鍵字為 AI 與 Algorithm，相較於其他相關詞語，例如深度學習、類神經網路等等，這兩個字所涵蓋的範圍最大。倫理類的關鍵字有六個，分別是 Ethic、Justice、Injustice、Discrimination、Equality、Inequality，其中，Ethic 與 Justice 是兩個最寬廣的倫理用詞，可確保文獻的涵蓋性。公部門類的關鍵詞，則為 Public Sector、Government、Bureaucracy，其中 Public Sector 與 Government 也具有涵蓋性。

本研究資料搜尋過程分兩階段，第一階段依照 PRISMA 模式進行，如圖一左方所示，文獻蒐集流程如下：

一、辨識（identification）

本研究從 EBSCOhost 資料平台搜尋包括 Academic Search Complete, EconLit, ERIC 等 26 個資料庫，在文章標題（Title）與主題分類（SU）中，以 AI、Algorithm 為兩個分支，個別與倫理類與公部門類共九個關鍵字配對，[8] 要特別說明的是，在搜尋與 Algorithm 有關的論文時，會排除 equality、inequality 兩個關鍵字以篩除數學演算法相關文獻。

此階段的搜尋條件是：1) 有提供全文；2) 經學術同儕審查的期刊論文。初步搜尋結果如圖一所示，AI 類共 494 篇，重複 23 篇，扣除後剩下 471 篇。Algorithm 類共 489 篇，重複 238 篇，扣除後剩 251 篇。

[8] 例如，AI 與 Ethic 一組，分別在 Title 與 SU 欄位交叉搜尋，之後再換 AI 與 Justice 一組，重複前述步驟。AI 使用完畢，再換 Algorithm 分別與九個關鍵詞配對，重複前述搜尋步驟。

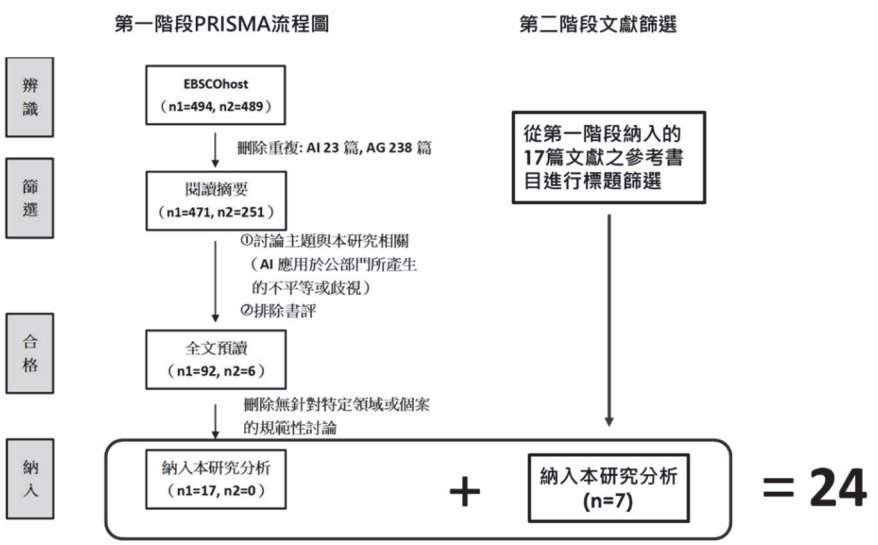
二、篩選（screening）

閱讀 AI 類 471 篇與 Algorithm 類 251 篇摘要進行篩選，篩選條件為該文主題必須是 AI 在公部門應用後所產生的不平等，並排除書評類文章。篩選時若無法從摘要確認研究主旨，則以該文結論判定。此階段共篩選出 AI 類 92 篇、Algorithm 類 6 篇（Algorithm 類大多涉及醫學、自然科學、高科技的研究，與本文旨趣不符，故只剩 6 篇入選）。

三、合格（eligibility）

本階段進行全文預讀，刪除無針對特定領域或個案的規範性討論，最後 AI 類納入（included）17 篇作為本研究進行後設分析的研究對象，Algorithm 類全數刪除。

第二階段的文獻搜尋，則是從篩選出之 17 篇文章的參考書目中，依照前述標準，從文章標題搜尋研究對象，包括政府部門與非營利組織的研究報告，最後篩選出 7 篇。因此，作為本研究後設分析的文本共有 24 篇，篩選流程如圖一右側所示，相關文獻整理請參閱附錄。



圖一　本研究分析文獻篩選過程

圖來源：本研究繪製

說明：n1為AI類，n2為Algorithm類，24篇文獻分析整理，請參閱附錄

本研究分析單位為篇，分析策略是基於比較、萃取、異同觀察，並同時保留系絡與研究發現的細節，特別是較少見的研究發現（Finfgeld, 2003; Thorne, Jensen, Kearney, Noblit, & Sandelowski., 2004）。針對 24 篇文本的分析步驟如下：1) 略讀：經由略讀，找出每篇皆提及的資訊類別，分別是研究目的、AI 的政府應用層級、政策領域、AI 的應用方式、所產生的不正義現象、產生此現象的原因、政策建議；2) 製表：利用前一步驟所歸納出的資訊類別，製作編碼表格（類似附錄所示）；3) 精讀與編碼：利用質性研究的「結構編碼」（structural coding）原理，將文本內容進行歸類，置入編碼表格中；4) 編碼表精讀與萃取：以政策領域作為分類基礎，萃取出同一政策領域中，應用 AI 所產生的非意圖歧視及其原因。再跨越政策領域，觀察其政策過程與結果的共同點。

本文質性後設分析的「信實度」（Trustworthiness），由三個部分組成，分別是研究過程的可複製性、研究結果的正確性、以及屬於質性分析獨有的，即是否遺漏各別文獻的獨特觀點（Erwin, Brotherson, & Summers, 2011: 190-191; Levitt, Pomerville & Surace, 2016: 817; Levitt, Pomerville, Surace & Grabowski, 2017: 630; Levitt, 2018: 368-370)。有關研究過程的可複製性，本研究 PRISMA 篩選過程是由兩位研究者雙重確認，並在本文中進盡量詳實描述文本資料的蒐集、篩選、分析過程。有關研究結果的正確性，本研究針對納入分析的 24 篇論文採取精讀、編碼、萃取、分析的程序，特別在彙整過程確認不偏離該文之摘要與結論所述，以確保本研究彙整結果詳實反應各研究所提出之觀點、論述、與研究結果。有關本研究彙整分析後是否遺漏獨特觀點，基於維持本文論述主軸，並礙於文長限制，本研究補強的作法是提供附錄表格描述各研究的細部資訊。在研究限制上，由於本研究用於分析的文本是以經過同儕審查的出版品為主，因此不排除會有「出版偏差」（publication bias）的問題，亦即，本研究將無法觸及未出版但有重要研究發現之研究。

伍、從平等原則檢視 AI 的制度過程與結果

AI 作為制度的一部分，會在公共領域產生影響的關鍵在於人類是否相信並使用其所產生的警示資料或建議。因此，本文無意指控 AI 造成歧視，而是發現人類在信任機器的正確性與效率性之下，容易忽略或不信任自己的判斷，就像一般人面對複雜的數字計算時會比較相信電子計算機的結果而非自己的心算結果。

文本中 AI 的應用領域分八大類，分別是醫療照護、警察執法、刑事司法、國境管理與國土安全、國防、教育、公共就業、與國家財政。個案橫跨美國、加拿大、英國、澳洲、紐西蘭、中國、比利時、瑞士、義大利、與德國，而就應用層級來看，從國家、地區、州、地方（市、縣）政府都有。有關特定群體之間差別待遇的討論，引發最多關注的是醫療照護、警察執法、與刑事司法三大領域，其他則較少。此外，值得注意的是，國防領域對於 AI 技術倫理的研究頗豐，但大都著眼於戰場上能否容許非人類的 AI 做決定奪取人類性命的道德爭議（例如 Asaro, 2013; Horowitz, 2016; Sparrow, 2016; Sharkey, 2010; Roff, 2014 等），而非關特定群體間差別待遇問題。以下將從平等原則檢視 AI 應用於各領域的制度過程與結果，並將此討論整理於表一。

一、醫療照護

AI 通常應用於醫療諮詢、分類病患、與提供智慧輔助技術（intelligent assistive technology）（Wangmo, Kressig, & Ienca, 2019）。在醫療諮詢方面，中國科技部與國家發展改革委員會在某些地區的醫療照護體系導入 IBM 的 Watson 系統，能同時設計個人化醫療照護計畫（Sun & Medaglia, 2019）。在分類病患方面，美國（Takshi, 2021, Obermeyer, Powers, Vogeli, & Mullainathan, 2019）與英國（Winter & Davidson, 2019）導入 AI 技術計算病患的風險等級，據以決定後續醫療照護計畫。

研究顯示，病患的風險評估系統對不同群體的評估正確性有別（Winter & Davidson, 2019），非裔病人被歸類為高風險的機率低於白人（Takshi, 2021），而同一風險等級中，非裔比白人實際病得更嚴重，這表示白人的風險等級可能被高估，或是非裔被低估（Obermeyer et al., 2019）。會造成判斷失準的原因之一，是 AI 使用病人醫療支出紀錄做為判斷依據，但特定群體醫療支出較少的原因來自於結構性因素與系統性歧視，例如有色人種被認為對痛感的忍受力較高、身心障礙女性的症狀常被忽略、貧窮病人無能力獲得適當的醫療（Takshi, 2021）。此外，非裔的醫療支出較低，除了貧窮，也因為白人醫生較不會安排預防性醫療計畫給非裔病患，而且非裔不太相信醫療體系，除非急診否則不太習慣進醫院（Obermeyer et al., 2019）。除此之外，AI 對於辨識非裔的慣用語及圖像並不如白人精準，這也導致失準的診斷（Takshi, 2021）。

二、警察執法

警察執法領域會使用的是「自動臉部辨識系統」（automated facial recognition technology）與犯罪預測系統。自動臉部辨識技術基於照片清單比對民眾臉部資訊以搜尋罪犯，而照片清單裡包含通緝犯、逃離羈押的嫌犯、特別弱勢群體、與特別受關注的對象（Maxwell & Tomlinson, 2020），該清單攸關辨識正確性（Garvie, Bedoya & Frankle, 2016）。犯罪預測系統是警察機關用以預測犯罪者、犯罪行為與傾向，並採取預防措施的工具（Howard & Borenstein, 2018; Lum & Isaac, 2016; Ferguson, 2015; Madden, Gilman, Levy, & Marwick., 2017）。

前述兩種系統皆出現偏差判定的情形。臉部辨識系統對女性與有色人種的判讀，比對白人的判讀更不精確（Maxwell & Tomlinson, 2020; Garvie, Bedoya & Frankle, 2016），導致無謂的盤查打擾。犯罪預測系統則被發現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）。何以如此？因為某些群體較容易留下犯罪與入獄紀錄，但這並非因為他們較容易犯罪，而是他們比較付不出保釋金（Ferguson, 2015）。引發偏差判讀與不平等的主因，始於大數據本身的偏差與後續影響。例如，加州奧克蘭警方基於地理資訊與逮捕紀錄，讓犯罪預測系統進行機器學習與做犯罪預測，從中找出「可能犯罪者」的出沒地點讓警察加強巡邏，這表面上提升了特定區域的毒犯逮捕率，但這些資料又輸入機器學習系統，形成自我增強的效果（Lum & Isaac, 2016）。

三、刑事司法

AI 用於輔助法院判決已行之有年，在司法程序各階段都能應用，從審前釋放、緩刑、保釋，或服刑者的假釋，都可利用風險評估工具來預估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）。在文獻中被討論最多的，是一套由私人公司 Northpointe 所開發並廣受美國各地法院使用的「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），包括佛羅里達、威斯康辛、密西根、紐約州大部分地方政府、新墨西哥州人口最多的城市，都使用此系統（Brenner et al., 2020）。COMPAS 基於 137 個問題答案，計算風險分數

供法院做判決參考，部分的問題答案由被告填答，部分來自被告的個人資料（Angwin, Mattu & Kirchner, 2016; Chouldechova, 2017）。此外，在「司法精神醫學領域」（Forensic Psychiatry）所使用的暴力預測系統，在許多重要的司法決策過程中被做為參考工具，影響刑期與非自願醫療用藥的判定（Cockerill, 2020）。

前述 AI 系統被發現針對不同群體的判斷產生偏差，有色人種通常會被評斷為較具危險性（Howard & Borenstein, 2018），非裔被預測會累犯的比例約為白人的兩倍，且白人比非裔較易被評估為低風險（Angwin, Mattu & Kirchner, 2016），而拉丁裔也遇到同樣的狀況（Cockerill, 2020）。前述 COMPAS 系統在風險分數與計算錯誤的機率上，白人與有色人種之間的差異甚大。非裔拿低分（表示比較危險）的機率比白人高，且其「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）。[9] 這些誤判導致有色人種被判處較長的刑期（Cockerill, 2020），或更容易被處以非自願醫療用藥（Chouldechova, 2017; Cockerill, 2020）。這些差別待遇主要源於歷史數據中的系統性偏差，例如有色人種比白人更容易留下犯罪紀錄。前述 COMPAS 系統進行風險評估時雖然使用了未含種族資料的 137 個指標，但犯罪歷史、社會經濟狀況這些指標本身就有種族偏差（Brenner et al., 2020），特別是「父母是否曾經入獄」這個指標，對於較容易留下犯罪紀錄的非裔而言，就成了原罪代名詞（Angwin, Mattu & Kirchner, 2016）。

四、國境管理與國土安全

在國境管理領域裡，AI 用於進行臉部辨識，或為入境申請資料進行分類。前者如紐西蘭自動護照申請系統，使用人臉辨識技術審核申請者照片（Howard & Borenstein, 2018）。後者如加拿大出入境管理與移民、難民管理主責單位，透過風險（risk）、優先順序（priority）、複雜性（complexity）等指標為申請案進行分類，再計算個案分數、做機率評估、並配合其他指標提供整體性的建議，系統會針對有疑慮的個案提出警示，做為審查決策參考（Molnar & Gill, 2018）。

前述紐西蘭臉部辨識系統因無法正確辨識亞洲人臉而導致亞裔在申請護照過程中出現障礙，這是因為用於機器學習的亞裔人臉資料不夠充足（Howard & Borenstein, 2018）。而加拿大利用 AI 於移民與難民審核系統，導致部分申請人無

[9] 偽陽率即被判斷為高風險但卻沒有再犯，偽陰率被判斷為低風險卻再犯。

法得到公正的判斷。這套系統利用加國政府過去的決策資料作為學習基礎，但長久以來加國的決策就潛藏偏見，因為加國為世界各國進行安全等級評估時，通常基於一套粗糙的標準，包括是否產生難民、尊重人權、提供國家保護等，若難民申請案來自安全等級高的國家，則容易被拒絕。這套系統飽受批評之處在於對安全的定義並不周延，有些國家雖被評為安全，但該國對於某些群體而言並不安全，例如非異性戀者，或因家暴而逃家的女人（Molnar & Gill, 2018）。

五、大學教育

許多美國公、私立大學「招生辦公室」（admission office）利用 AI 搜尋與篩選學生，已非新聞。學校會在入學申請審核過程中，利用 AI 分析申請者在社群平台的發言、照片、與同儕之間的互動，進而預測申請人完成學業的可能性，以此作為核予入學許可的參考（Madden et al., 2017）。

公立大學部分資源來自政府，在審核入學許可時不應有群體差異，然招生辦公室利用 AI 之舉，導致不懂保護隱私的數位知識落差者居於劣勢，較難拿到入學許可。這群申請者的家庭收入往往不高，無力負擔隱私防護力較高的電子通訊工具，可能因此失去公平競爭的機會（Madden et al., 2017）。

六、公共就業

在政府提供的就業訓練與工作媒合機制上，比利時區域性公共就業服務（Public employment services）利用 AI 為求職者風險程度進行分類，若屬長期失業之虞的高風險者，會被優先聯絡安排職業訓練。亦即，分類決定了政府回應求職者的速度與後續職業訓練與監督（Desiere & Struyven, 2021）。

然而，這套 AI 分類系統易將移民、身心障礙、年長者歸類為高風險者，表面上此種判定偏差似乎帶來「優先」的好處，但由於這些群體並不必然是高風險者，反而在求職過程中必須面對過度的監控與職業訓練。更重要的是，錯誤歸類排擠了真正需要被優先處置者（Desiere & Struyven, 2021）。

七、國防

AI 在國防領域的應用範圍極廣，通常會被用於執行危險任務，降低軍隊傷

亡，例如拆除炸彈或狙擊敵方軍隊或首腦。此外，AI 能從大量監控資料中，找出資料之間的相關性，針對可疑的武器、車輛、人提出警示，以提醒戰場中的軍人提高警覺。其中，臉部辨識系統就是其中一種協尋人物的工具（Wasilow & Thorpe, 2019）。

然而，由於戰區通常位於人跡罕至或不易進出之地，資料蒐集困難導致數量不足，導致 AI 產出的正確性受到質疑，連帶影響相關決策結果。例如西方國家發展的人臉辨識系統，對於戰區某些人種與性別的辨識就不精確（Wasilow & Thorpe, 2019）。

八、國家財政

在財政上，AI 為房屋貸款或保險理賠申請的審核提供評估建議。有關房屋貸款，早在 30 年代由美國政府出資的「屋主貸款公司」（Home Owners' Loan Corporation ），就系統性拒絕高風險者的房貸申請，作法是標註有色人種社區居民為高風險者，當時俗稱為「畫紅線」（redlining），該系統演變至今，由 AI 進行房貸申請者的風險評估以作為審核參考（Allen, 2019）。而有關保險理賠，則是紐西蘭政府經營的「意外事故賠償公司」（Accident Compensation Corporation, ACC）使用演算法分析歷史資料，判斷理賠申請個案的複雜程度，若經 AI 判斷為簡易案件，即直接核准理賠，若否則由人類接手處理（Brownlie, 2020）。

美國 30 年代有色人種社區居民被排除於房貸名單以外，連帶影響他們無法購得「可負擔的住宅」（affordable housing），而此種現象延續至今，AI 用於房貸審核時，雖於法不能參考種族資訊，但用於機器學習的大數據包含了信用分數、被房東驅離的紀錄、逮捕紀錄、教育程度、工作背景、與先前的住址等，卻與種族明顯相關，這表示今日的 AI 仍會非蓄意地循著「畫紅線」邏輯進行房貸審核（Allen, 2019）。至於紐西蘭 ACC 的理賠審核，AI 傾向於將有健康問題或身心障礙的申請者直接歸類為複雜個案。這是因為在歷史數據中，身心障礙或有健康疑慮的申請者通常較難以被核准理賠，因此 AI 習得此判斷模式（Brownlie, 2020）。

九、小結

從表一的彙整中可知，AI 作為公共政策執行過程中的一環，在不同的政策領域中出現違反平等原則的疑慮，這是因為用以機器學習的大數據潛藏著歷史中的差

別待遇，並藉由 AI 與人類的互動而複製歷史不正義。事實上，除了大數據偏差以外，錯誤的演算法也會導致不正義的現象。著名案例是「澳大利亞聯邦政府服務部」（Service Australia）使用 Centralink 公司開發的「自動債務索償系統」（Robo Debt），連結社會福利案主個人資料與稅務資料，以判斷溢領行為並向民眾索回溢領款項。然而，演算法的錯誤導致眾多誤判，經濟拮据的社會弱勢無力提告或提出反駁，導致身心俱疲，成為 AI 技術的受害者（Toohey, Moore, Dart, & Toohey, 2019）。由此可見，AI 已經深入民眾生活的各面向，其所導致的負面影響更應該被慎重檢視。

表一 從平等原則檢視 AI 在各政策領域的應用

政策領域	應用範圍	從平等原則檢視 制度過程— AI 如何被應用	從平等原則檢視 制度結果— 是否產生差別待遇
醫療照護	為病患計算風險分數，依此設計個人化醫療照護計畫	忽略少數人種與低收入者的醫療習慣潛藏於大數據中： 1) 以病人過去的醫療支出為分類標準，忽略少數人種與低收入者因經濟、歧視等因素減少就醫的習慣。 2) 資料庫不足，AI 對於辨識非裔的慣用語及圖像有困難	特定群體（非裔）接受醫療的權利被剝奪： 1) 非裔與白人的健康評估精準度有差異，前者的診斷較不正確 2) 非裔病人被歸為高風險的機率低於白人，其風險等級可能被低估
	提供智慧輔助技術	無涉及平等原則	特定群體（低收入者）無法負擔高價的智慧輔助技術
警察執法	自動臉部辨識系統	忽略非裔長久以來面對的不平等已潛藏於大數據中： 1) 以嫌疑犯照片資料庫比對民眾臉部資訊以搜尋嫌疑犯，但忽略了資料庫中存在著不成比例的非裔	特定群體（非裔）被判讀的精確率較低，導致後續遭受無謂的警察盤查
	犯罪預測系統：預測未來可能的犯罪者、犯罪地、與犯罪行為	忽略低收入者與有色人種較易留下犯罪資料，而此偏差潛藏於大數據中： 1) AI 基於偏差數據提供嫌疑犯所在區域，警察因而加強巡邏	特定群體與社區（有色人種與低收入）不成比例受到警察關注： 1) 持毒品而遭逮捕者集中在非白人與低收入社區，但非白人持有毒品的比例與白人不

政策領域	應用範圍	從平等原則檢視 制度過程— AI 如何被應用	從平等原則檢視 制度結果— 是否產生差別待遇
		2) 這些區域的逮捕紀錄，又成為機器學習的數據	相上下。 2) 低收入者與有色人種較容易被警察盤查。
刑事司法	利用風險評估工具預估嫌犯或被告再犯或累犯的可能性，提供法官判決參考	忽略數據中的犯罪歷史、社會經濟狀態等指標本身就有種族偏差 1) 以被告自填問卷與其個人紀錄，來計算其風險分數 2) 評估指標包括犯罪歷史、個人特質、社會經濟狀態、父母是否曾經入獄	特定群體（非裔、拉丁裔）易被判較長的刑期： 1) 有色人種通常會被評斷為較具危險性 2) 非裔被告被預測累犯的比例約為白人的兩倍 3) 白人比非裔易被評估為低風險 4) 非裔的偽陽率比白人高，偽陰率比白人低
	利用 AI 建構暴力預測系統，提供司法部門在刑期、非自願醫療用藥判定的參考	忽略歷史資料偏差導致機器學習偏差： 在犯同一種罪的前提下，非裔與拉丁裔的刑期通常比白人長，而這種偏差會進入暴力風險評估中	特定群體（有色人種）比較容易被判斷有暴力傾向，因而被處以非自願醫療用藥
國境管理與國土安全	為入境或難民庇護申請進行分類	忽略傳統審核標準下所產生的歷史偏見存在於資料中： 1) 透過數項指標為個案國籍的安全性分類，再計算個案分數、做機率評估，提供決策的參考。 2) 特定個案會提出警示提醒審查官員	來自某些國家的申請案無法得到公正審核
	自動護照申請系統使用人臉辨識審核申請者照片	忽略用以機器學習的資料缺乏足夠的亞裔數據	特定群體（亞裔）的臉部無法被正確辨識

政策領域	應用範圍	從平等原則檢視 制度過程— AI 如何被應用	從平等原則檢視 制度結果— 是否產生差別待遇
大學教育	分析入學申請者在社群平台的表現以作為決策參考	忽略社會結構性因素對於入學申請者的影響 1) 蒐集申請者在交平台的發言、照片、人際網絡，作為審核入學決策參考。 2) 忽略申請者的人際網絡關係與其在網路平台的表現，都與其社經條件有關	特定群體（不懂保護隱私的數位知識落差者、因經濟因素無法購買避險工具者、社經情況較劣勢者），較難拿到入學許可
公共就業	使用 AI 分析模型為求職者做分類以決定服務優先順序	忽略傳統分類標準所隱含的，對特定群體的既存偏見： 1) AI 模型將求職者進行風險分類，高風險者的求職過程會被輔助並監控。 2) AI 模型沿用了傳統的分類方式	特定群體（移民、身心障礙、年長者）的求職過程受到過度輔導與監控： 1) 社會弱勢群體易被歸類為高風險者，即使他們實際上可能不是。 2) 間接導致其他真正需要幫助的人失去優先機會
國防	協助執行危險任務（例如協助拆彈、無人機狙擊暗殺敵軍首腦）	無涉及平等原則	無涉及平等原則
財政	房屋貸款審核	忽略大數據資料中潛藏種族偏差： 1) 大數據資料所包含的項目與種族有明顯關聯。 2) 基於前述資料，為房貸申請者進行風險評估，作為審核參考	特定群體容易被拒絕房貸申請，導致他們無法購買負擔得起的住宅
	保險理賠申請審核：AI 為保險理賠申請案進行分類以決定是否直接核准或由人類處理	忽略歷史資料中潛在的群體偏差： 1) 歷史資料中，索賠者若為身心障礙或有健康問題，其申請案通常比較不會被核准	特定群體（有健康問題、身心障礙）的理賠比較不容易，時程容易被延宕，不論其索賠案件有多麼簡易

資料來源：本研究整理

陸、AI 非意圖歧視的影響分析

一、非意圖歧視違反平等對待的公共價值

政府基於法律制度，公平對待所有公民，是公共政策領域的重要公共價值，也是政府正當性的基礎。從表一可知，AI 作為制度的一部份，輔助政府進行決策判斷以提升決策效率與品質，然過程中忽略了歷史大數據潛藏著由來已久的結構性社會因素，導致特定群體持續遭受差別對待，輕者得到沒必要的監控關注，重者人權受到侵犯，顯然皆有違平等對待的公共價值。

然而，從公共行政實務的視角觀之，平等對待的理想性與行政資源的有限性之間出現兩難。就組織內部而言，數位科技帶來效率行政，使公務人力能更有效利用，提升公共服務品質，再加上數位科技深入各政策領域，例如臉部辨識或步態追蹤，讓社會變得更安全，犯罪預測系統高度監控「可能犯罪的壞人」，或隔絕「可能再犯的壞人」於社會之外。由數位科技提升的人類福祉，與其非意圖歧視所破壞的公共價值之間，該如何取得符合社會正義的平衡？本文無意進入「電車困境」的討論，但與此困境類似的是，多數人對某些公共價值的期待，可能使 AI 非意圖歧視成為被默許的結果，畢竟社會中被歧視者大都是相對少數。大部分民眾會期待居住在一個安全的環境中，若犧牲少數他人的人權，能換得安全的生活環境，在效用主義「為大多數人追求最大的幸福」的習慣性邏輯下，民眾可能會對 AI 系統的歧視保持靜默。因此，政府作為公共價值守護者，在數位轉型的必然趨勢中，若無法主動正視 AI 對特定群體差別對待的過程與結果，並積極矯正之，特定群體的福祉與基本人權將持續遭受剝奪。

二、非意圖歧視的影響分析：一個比較的觀點

痛苦是無法量化比較的，本文無意比較各領域受歧視者痛苦的輕重，而是從國際人權公約所隱含的優先順序來思考這個問題。此分析的重要性在於，政府作為公共價值的積極守護者，了解各領域差別對待更深層的問題本質，是思考「效率行政」與「公平對待」兩難的第一步，而以下的分析便是嘗試踏出這一步。在進行分析之前，首先說明何以國防、國境管理兩類被排除於以下的比較分析之外，其次比較分析其他六個政策領域（包含七個政策項目），檢視 AI 歧視對待的本質與問題

解決的急迫性。

（一）國防與國境管理的排除理由

國防與國境管理非屬公民聯合關係，是被排除於後續比較分析的主要理由。在國防領域，戰場上 AI 的殺戮對象是敵軍，各種辨識系統力求正確才能「殺對人」而不誤殺平民。由於戰場上敵我分屬不同國家陣營，平等原則並不適用於戰地，我們甚至希望 AI 能百分百正確「歧視」敵軍以降低平民傷亡程度。在國境管理中，AI 的應用對象是移民或難民，非屬本國公民，若再加上國土安全的考慮，平等原則亦不適用。雖不納入本文後續的比較分析，但兩大領域的正義辯證仍深具意義。依照米勒的邏輯，社會正義具有政策指引的功能，為了使追求正義具有意義，討論社會正義有三個前提假設：分別是要有確定成員與邊界的社會、看得到的制度、與可以改變制度的機構（梁文韜，2005a）。如果要使 AI 在這兩個領域的使用顧慮到社會正義，就應把全球視為一個國際公民社會，從中論證全球公民有自由遷徙、居住、免於戰爭恐懼的權利。然而，以全球為邊界的社會，缺乏具有權力與約束力的制度，以及可以改變制度的機構。聯合國作為一個超國境組織，對於霸權國家的規約能力有限，其在恐怖組織、政變、極權國家之前，亦顯得軟弱無力。因此，在這兩大領域裡要討論正義，相當困難，特別在國土疆界的思考限制下，較難得到共鳴。

（二）非意圖歧視的影響比較

在公民聯合關係中，所謂平等，非指獲取相同數量的好處，而是分配到好處與負擔的「機會」相同（Miller, 1997）。從本文分析中可知，特定群體在制度過程與結果中遭受差別待遇，顯見其被分配到負擔的機會高於其他群體，而究其負擔的本質，則涉及基本人權的危害。那麼，在公民聯合關係中，該以什麼標準來檢視人權危害的程度，本研究擬以國際人權公約所隱含的人權優先順序來檢視。在此需予敘明的是，本文相信各種基本人權應同時受到保障，但就實務而言，政府部門受限於資源與外環境，要落實人權仍不得不進行務實的考量。例如在有限的資源之下，一位嫌疑犯因 AI 誤判而導致延長刑期，與一位失業者因 AI 誤判而接受過度輔導，哪一類的人權危害應優先處理，並不難判斷。

事實上，從國際公約可以看出人權保障的優先順序。依照「公民與政治權利國際公約」（The International Covenant on Civil and Political Rights）第二條第一款，締約國「應確保所有境內受其管轄之人，一律享受本公約所確認之權利」。而「經

濟社會文化權利國際公約」（The International Covenant on Economic, Social and Cultural Rights）第二條第一款則提到，締約國「承允盡其資源能力所及，逐漸使本公約所確認之各種權利完全實現」。由此可知，國際公約中對於公民與政治權利的保障要求，更甚於對經濟社會文化權利保障的要求。因此，即使各類人權都應受到保障，但仍有排列優先順序的必要性。

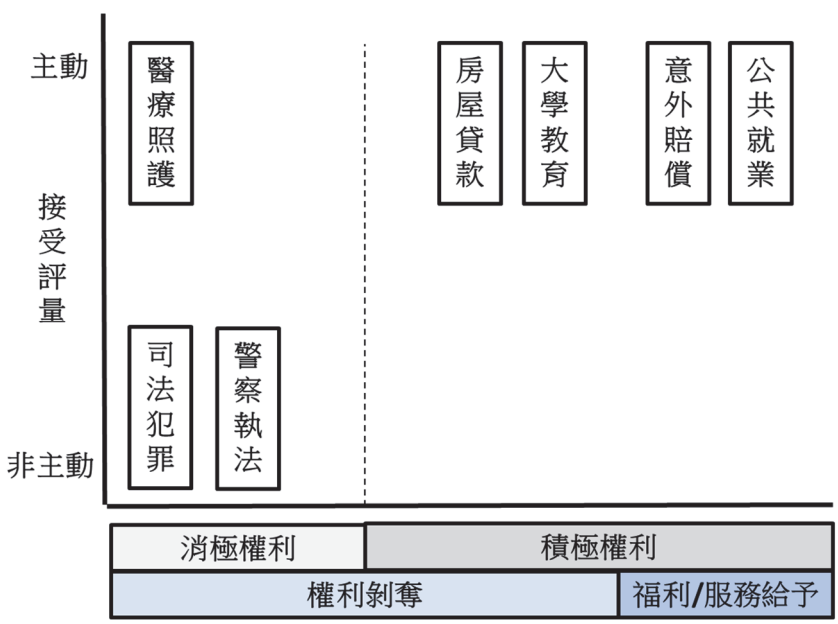
有關各類人權的優先順序，有學者基於「馬斯洛的需求層級」（Maslow pyramid），以「缺乏」（scarcity）為起始點來建立順序架構（Quintavalla & Heine, 2019），也有從價值、功能、共識三個面向討論人權，認為人權在不同的政治與經濟系絡中會有所扣減，因此應基於「不得扣減的人權」（non-derogable rights）觀點來思考其優先順序（Koji, 2001），而有關不得扣減的人權，一般認為「生命權」（right to life）、「實體安全」（physical security）、「正當程序」（due process）、「與不受歧視」（non-discrimination）歸屬這一類（Farer, 1992）。Suárez -Müller （2019）基於「傳統先驗」（Transcendental）的層次為人權進行分類，推論出「生命權」（the right to nourishment）、「自由權」（the right to freedom）、「義務權」（the right to exercise duty）是所有人權的基礎，其他人權在缺乏此三大人權的情況下是無法建構的。Pogge （2008） 則主張「消極權利」（negative rights, 即不可受到侵犯的權利）應優先於「積極權利」（positive rights, 即要求得到幫助的權利）。

人類最不可受到侵犯的權利，即應是最基本的、若缺乏則其他人權無法建構的，依照前述 Suárez -Müller （2019）的主張，配合本研究差別待遇的問題本質，生命權與自由權當屬此類。在六大領域（七個政策項目）中，AI 在刑事司法、警察執法、醫療照護三類政策的應用所產生的非意圖歧視，直接剝奪生命權與自由權，相信這也是 AI 在這三大領域的應用受到最多關注的原因。然而，三大領域之間有個明顯差別，即受歧視者是否主動尋求評量。在刑事司法與警察執法中，特定群體較容易被判重刑或是被警察登門盤查，他們並不知道自己成為評量對象。但在醫療照護領域中，病人需要主動就醫才會成為評量對象。

本研究試著延伸米勒的系絡主義，由「消極權利與積極權利的剝奪」與「受歧視者接受評量的主動性」兩個面向，交織成不同的系絡條件，藉此檢視研究文獻中討論的歧視本質（請參考圖二）。接受評量的主動性，是指被歧視者對於自己被評量一事是否知情，若個人在不知情的狀況下被評量並進而危及個人福祉，則對人權的危害更甚，特別是消極權利的剝奪更是。例如房貸申請人知道，銀行在決定是否

放貸時會基於申請人的還款能力做決定，或是尋求醫療協助者知道，醫療單位在決定醫療照護行為時會基於病人整體的健康風險做判定。然而，警察在公民無犯罪事實時上門盤查時，被盤查者對於自己已被評量且列入黑名單一事一無所知。

要特別說明的是，圖二各類政策的定位是依照本研究 24 篇文獻中所討論的歧視本質繪製，並非表示該領域的差別待遇只有一種類型。例如，公共就業類別在 Desiere 與 Struyven （2021）的研究中發現特定群體在求職過程中被過度輔導，此例屬於「積極權利」的範疇且「獲得服務」（被過度輔導），所以被放置於圖二右側。但若後續有其他研究顯示，特定群體在求職過程中被輔導單位拒絕，那麼雖仍屬積極權利的範疇，但其在圖中的位置就會往左側「權利剝奪」範圍移動。



圖二 各政策領域非意圖歧視的起始與結果

資料來源：本研究整理

在圖二中，刑事司法、警察執法被置於左下角是因為特定群體被判較長的刑期，或被施予較多的盤查，生命權與自由權受到剝奪。他們並無主動申請進入系統接受評量，甚至不知道自己成為警察關注的對象。圖二左上方是醫療照護，表示受歧視者主動申請醫療服務，接受健康風險評估。然由於有色人種與貧窮者的健康風險較容易被誤判，導致不適當的醫療行為，生命權受到剝奪。

圖二中間上方分別是房屋貸款與大學教育，表示受歧視者主動提出申請並接受

資格審查，二者涉及居住權與受教權，屬於尋求幫助的權利（積極權利）受到剝奪。申請房貸是否該視為公民應有的權利？若從公股銀行辦理房屋貸款來看，國家房屋政策以低利房貸鼓勵民眾購買自住屋，是一種居住正義的實踐。由此觀之，社會中特定群體即使在符合財務條件下其貸款申請仍被系統性拒絕，等於剝奪其居住權利。在教育方面，特定群體因不懂隱私保護而被拒絕入學的機會比其他群體高，意味著特定群體接受教育的權利較容易遭到剝奪，且經由高等教育而追求更好生活的權利受到限制。

圖二右上方是國家提供的意外賠償與公共就業，從歧視本質來看，被歧視者在尋求幫助時，皆能得到相關福利或服務，只是處理過程有所延宕，或是被提供過度的服務。在意外賠償案例中，特定群體的案件易被視為複雜案件而延宕理賠速度。在公共就業個案中，特定群體受到不必要的就業輔導與監控。此二者皆涉及積極權利，但並非權利被剝奪，而是福利／服務被給予，因此位於圖二的最右側。

柒、結論

米勒的平等原則意指不同群體在好處與負擔的分配機會相同，若特定群體系統性地被差別對待，比其他群體承受負擔的機會較高，就是不平等。AI 作為制度的一部分，在人類的信任與對效率的追求之下對社會產生影響，他能經由大數據的訓練而提升正確率，判斷會比人腦更為快速準確，他在未經人類允許的情況下，記錄學習人類每天上班來回的時間與路徑，進而會在人類開車上班之前就貼心提醒何時即將到達目的地。人類開始大量依賴 AI 的學習能力，即使有些抗拒與遲疑，但使用時又洋洋得意。在公共政策領域裡，AI 科技的應用提升社會福祉，為大多數人帶來更高品質的公共服務，他可以基於大數據而提供更正確的醫療諮詢，也可以提高罪犯的逮捕率，防止累犯進入社會，讓民眾有更安全的生活環境，這就像乾淨的空氣或水，是一種可以澤及所有群體，讓每個群體都能平等享受的社會福祉。然而，人類對 AI 的過度信任，正在複製歷史中存在已久的不平等，這些歷史制度的產物形成大數據，透過 AI 的學習而繼續存在於制度中，他具有客觀的表象，他的不正義不易被察覺，AI 的應用導致特定群體承受較高的風險，承擔較高的成本，在有些情況下，受害者對於 AI 的侵入一無所知，在警察或司法系統被不正義地對待卻毫不知情。米勒的平等原則，在檢視 AI 應用於社會系絡所產生的影響時，出現極大的複雜度，在公民聯合關係中，民眾可能在享受 AI 帶來的社會福祉的同

時，妥協於 AI 對某些群體帶來的人權危害。因此，AI 的非意圖歧視，需要政府積極介入，而無法依賴公民社會的自覺。

何以如此？從圖二可知，AI 在刑事司法與警察執法領域的應用，直接危害特定群體的生命權與自由權，雖然沒有實證研究顯示，如此是否真的為民眾帶來更為安全的生活環境，但相信沒有人會反對逮捕罪犯與隔絕累犯。即使 AI 誤判導致特定群體面臨不公平的刑期，但大部分民眾在自身安全與社會正義的權衡之下，採取「寧可錯殺也不願放過」之心態者應不在少數。事實上，我們不難看到大眾在生活安全的理由之下對於倫理疑慮的妥協。例如到處存在的路邊監視器，雖然可能侵害個人隱私權，但在生活安全的前提下很容易被正當化。因此，要依賴民眾由下而上矯正 AI 對特定群體的非意圖歧視，並不實際，也較難得到社會大眾的實質共鳴。

基於米勒的觀點，不正義可以透過制度調整來矯正，但這需兩個前提：首先，政府作為公共價值的守護者，需要意識到不正義的存在，有勇氣打開潘朵拉的盒子，直接面對社會中存在已久的偏見、歧視、與差別待遇。其次，政府需有意願從制度過程來解決這個問題，檢視 AI 用以學習與進行判斷的大數據，找出其中所隱藏的群體差異，研究該差異與事實的距離。

政府該如何從制度過程矯正或避免 AI 的非意圖歧視？本研究建議從政策籌備與執行兩階段著手：

一、籌備階段

政府部門在應用 AI 技術之前，需確認是否會導致差別待遇。針對私人公司所研發的 AI 系統，應抱持高度警覺，避免公司以商業機密、所有權、專利權等理由，拒絕揭露演算法，導致政府部門無從解釋 AI 的產出。因此建議在各政策領域建立負責 AI 公正性的部門，提供作業規範讓 AI 研發公司有所依循（Madden et al., 2017; Wasilow & Thorpe, 2019），規定公司需不斷測試其演算法的正確性（Brenner et al., 2020），與在各種群體間所造成的判斷偏差，並公布測試結果（Garvie, Bedoya & Frankle, 2016）。此外還須要求研發商提供可解釋的演算法 （explainable algorithms），讓政府部門能追溯演算法中的因果關係（Brenner et al., 2020）。必要時，政府部門有責任做前瞻性的規劃，制定法律規範 AI 在特定領域的應用（Garvie, Bedoya & Frankle, 2016）。此外，針對用以機器學習的資料庫，政府部門應檢視其正確性、多元代表性、與可能的偏差（Brenner et al., 2020; Desiere & Struyven, 2021），並了解資料與資料之間如何連結（Obermeyer et al., 2019）。一旦

政府部門使用 AI 技術，為了避免群體間的差別待遇，在決定哪些資料應導入演算法時，需要慎選資料類別，注意資料與資料之間的相關性與連動性，以做適當的資料排除（Ferguson, 2015）。

二、執行階段

當政府部門開始應用 AI 作為行政或決策輔助工具，雖然節省人力，但需確實監測結果（Allen, 2019; Brownlie, 2020），例如聘任公正第三方定期監控（Molnar & Gill, 2018），稽核影響（Allen, 2019; Winter & Davidson, 2019; Howard & Borenstein, 2018）。稽核需要有評估指標，並透過適合的研究方法（Molnar & Gill, 2018），例如可透過實驗來找出演算法潛藏的歧視，針對預估值與實際值之間的差距進行常態性的檢視，並公開資訊（Winter & Davidson, 2019; Koepke & Robinson, 2018）。不論是政策制定者或實務工作者，都必須了解演算法預測結果可能產生的錯誤（Desiere & Struyven, 2021）。當政府部門依照演算法所提供的資訊做決策時需要更嚴謹（Ferguson, 2015），為了不盲目跟隨 AI 執法，跨領域學習是必要的，甚至應放入各領域的基本訓練中，特別是在轉譯或詮釋 AI 提供的訊息時，要更為精準了解其中蘊含的意義與可能發生的錯誤（Cockerill, 2020）。

AI 非意圖歧視的相關實證研究，目前仍有不足，特定群體在各公共政策領域中遭受歧視的程度，以及差別對待所造成的實質影響，需要從更多正義理論觀點進行論證。AI 的應用牽涉各類利害關係者，政府部門、AI 研發公司、以及廣大民眾，更多討論應著眼於多元利害關係者對於 AI 公正性的定義、觀點與相互妥協時的底線，才可能更務實地將抽象的公平正義落實於演算法中。多數人的正義可能不是少數人的正義，這個在正義理論中不斷被檢視的議題，應在 AI 技術席捲各公共政策領域的數位時代，得到更多的關注。

參考書目

丁玉珍、林子倫（2020）。人工智慧提升政府公共治理的應用潛力探討。**檔案半年刊，19**（2），24-41。Ting, Yu-Jen & Lin, Tze-Luen (2020). Ren gong zhi hui ti sheng zheng fu gong gong zhi li de ying yong qian li tan tao [The Exploration of the Potential on Artificial Intelligence Applications to Improve Public Governance]. ***Archives Semiannual, 19***(2), 24-41.

甘偵蓉、許漢（2020）。AI 倫理的兩面性初探—人類研發 AI 的倫理與 AI 倫理。**歐美研究，50**（2），231-292。Gan, Zhen-Rong & Hsu, Han (2020). AI lun li de liang mian xing chu tan—ren lei yan fa AI de lun li yu AI lun li [A Preliminary Study of AI Ethical Duality: AI Ethics and Ethical AIs]. ***EurAmerica: A Journal of European and American Studies, 50***(2), 231-292.

李仲彬、陳敦源、蕭乃沂、黃東益（2006）。電子化政府在公共行政研究的定位與價值：議題連結的初探性分析。**東吳政治學報，22**，73-120。Lee, Chung-Pin, Chen, Don-Yun, Hsiao, Nai-Yi & Huang, Tong-Yi. (2006). Dian zi hua zheng fu zai gong gong xing zheng yen jiu de ding wei yu jia zhi: Yi ti lian jie de chu tan xing fen xi. [Evaluating the Topical Connection between E-Government and Public Administration Research: An Exploratory Study]. ***Soochow Journal of Political Science, 22***, 73-120.

張文華（2000）。機率推理，2022 年 9 月 28 日，取自：https://terms.naer.edu.tw/detail/1314592/。Zhang, Wen-Hua (2000). Ji lv tui li [Probabilistic Reasoning]. Education Dictionary. Retrieved September 28, 2022 from https://terms.naer.edu.tw/detail/1314592/.

張國恩（2000）。類神經網路，2022 年 9 月 28 日，取自：https://terms.naer.edu.tw/detail/1315497/。Zhang, Guo-En (2000). Lei shen jing wang lu. [Neural Network]. Education Dictionary. Retrieved September 28, 2022 from https://terms.naer.edu.tw/detail/1315497/.

梁文韜（2005a）。論米勒的制度主義社會正義論。**台灣政治學刊，9**（1），119-198。Leung, Man-To (2005a). Lun mi le de zhi du zhu yi she hui zheng yi lun [On David Miller's Institutional Theory of Social Justice]. ***Taiwan Political Science Review, 9***(1), 119-198.

梁文韜（2005b）。程序、後果及社會正義：論米勒的混合型正義論。**人文及社會科學集刊，17**（2），217-269。Leung, Man-To (2005b). Cheng xu, hou guo ji she hui zheng yi: Lun mi le de hun he xing zheng yi lun [Procedures, Consequences, and Social Justice: On David Miller's Hybrid Theory of Justice]. ***Journal of Social Sciences and Philosophy, 17***(2), 217-269.

梁文韜（2005c）。系絡、原則與社會正義——比較米勒及瓦瑟的多元主義正義論。**歐美研究，35**，605-668。Leung, Man-To (2005c). Xi luo, yuan ze yu she hui zheng yi—Bi jiao mi le ji wa se de duo yuan zhu yi zheng yi lun [Contexts, Principles and Social Justice—A Comparison of Miller's and Walzer's Pluralist

Theories of Justice]. ***EurAmerica: A Journal of European and American Studies, 35***, 605-668.

陳敦源（2020）。**公部門機器演算法應用之制度調適與路徑分析（結案報告）**。國家發展委員會委託研析報告（NDC-MIS-110-001），未出版。Chen, Don-Yun (2020). Gong bu men ji qi yan suan fa ying yong zhi zhi du tiao shi yu lu jing fen xi (jie an bao gao) [Institutional Adaptation and Path Analysis for the Application of Machine Algorithms in the Public Sector]. National Development Council Research Report (NDC-MIS-110-001)

陳瑞麟（2020）。科技風險與倫理評價：以科技風險倫理來評估台灣基改生物與人工智能的社會爭議。**科技醫療與社會，30**，13-65。Chen, Ruey-Lin (2020). Ke ji feng xian yu lun li ping jia: Yi ke ji feng xian lun li lai ping gu tai wan ji gai sheng wu yu ren gong zhi neng de she hui4 zheng yi [Technological Risks and Ethical Evaluation: Applying a Risk Theory Ethics to the Social Controversies Surrounding Genetically Modified Organisms and Artificial Intelligence in Taiwan]. ***Taiwanese Journal for Studies of Science, Technology and Medicine, 30***, 13-65.

彭錦鵬（2020）。AI 和疫情下行政學的挑戰與發展。**法政學報，29**，1-10。Peng, Thomas C. P. (2020). AI han yi qing xia xing zheng xue de tiao zhan yu fa zhan [Challenges and Development of Public Administration in the era of AI and Pandemic]. ***Journal of law and political science, 29***, 1-10.

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，13**（2），91-114。Huang, Hsini, Tseng, Kuan-Chiu, Liao, Zhou-Peng, & Chen, Don-Yun (2021). Dang ren gong zhi hui jin ru zheng fu: gong gong xing zheng li lun dui AI yun yong de fan si [When AI Joins the Government: A Reflection on AI Application and Public Administration Theory]. ***Journal of Civil Service, 13***(2), 91-114.

楊惟任（2018）。人工智慧的挑戰和政府治理的因應。**國會季刊，46**（2），67-83。Yang, William (2018). Ren gong zhi hui de tiao zhan han zheng fu zhi li de yin ying [The Challenge of Artificial Intelligence and the Response of Government Governance]. ***Congressional Quarterly, 46***(2), 67-83.

魯俊孟（2020）。AI 人工智慧對公共政策在政策倫理的對話初探。**理論與政策，23**（1），59-81。Lu, Chunmeng (2020). AI ren gong zhi hui dui gong gong zheng ce zai zheng ce lun li de dui hua chu tan [The Dialogue between Artificial Intelligence and Public Policy upon Policy Ethics]. ***Theory and Policy, 23***(1),

59-81.

韓釗（2019）。大數據、人工智慧與地方治理－以情感運算的應用為例。**中國地方自治，72**（11），26-45。Han, Charles (2019). Da shu ju, ren gong zhi hui yu di fang zhi li—Yi gan qing yun suan de ying yong wei li [Big Data, Artificial Intelligence and Local Governance - An Application of Affective Computing as an Example]. ***Zhong Guo Di Fang Zi Zhi, 72***(11), 26-45.

Aguilar, S. J. (2018). Learning Analytics: At the Nexus of Big Data, Digital Innovation, and Social Justice in Education. ***TechTrends: Linking Research & Practice to Improve Learning, 62***(1), 37-45.

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

Allen, J. A. (2019). The Color of Algorithms: An Analysis and Proposed Research Agenda for Deterring Algorithmic Redlining. ***Fordham Urban Law Journal, 46***(2), 219-270.

Androutsopoulou, A., N. Karacapilidis, E. Loukis, & Y. Charalabidis (2019). Transforming the Communication between Citizens and Government through Ai-Guided Chatbots. ***Government Information Quarterly, 36***(2), 358-367.

Angwin, J., S. Mattu, & L. Kirchner (2016). Machine Bias. In Secondary Machine Bias, ed Secondary Angwin, Julia, Surya Mattu, and Lauren Kirchner. New York, NY.: Propublica.

Asaro, P. (2012). On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making. ***International Review of the Red Cross, 94***(886), 687-709.

Bannister, F., R. Connolly, S. Giest, & S. Grimmelikhuijsen (2020). Administration by Algorithm: A Risk Management Framework. ***Information Polity: The International Journal of Government & Democracy in the Information Age 25***(4), 471-490.

Bender, E. M., T. Gebru, A. McMillan-Major, & S. Shmitchell (2021, March). ***On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*** Conference: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event, Canada.

Benington, J. (2015). Public Value as a Contested Democratic Practice. In J. M. Bryson, B. C. Crosby, & L. Bloomberg (Ed.), ***Creating Public Value in Practice*** (pp. 29-

48). New York, NY: Routledge.

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

Bozeman, B. (2007). Public Values and Public Interest: Counterbalancing Economic Individualism. Washington, DC: Georgetown University Press.

Brenner, M., J. S. Gersen, M. Haley, M. Lin, A. Merchant, R. J. Millett, S. K. Sarkar, & D. Wegner. (2020). Constitutional Dimensions of Predictive Algorithms in Criminal Justice. ***Harvard Civil Rights-Civil Liberties Law Review, 55***, 267-310.

Brownlie, E. (2020). Encoding Inequality: The Case for Greater Regulation of Artificial Intelligence and Automated Decision-Making in New Zealand. ***Victoria University of Wellington Law Review, 51***, 1-26.

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

Chouldechova, A. (2017). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. ***Big Data, 5***(2), 153-163.

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.

Cohen, N. (2018). There's No App for Justice. ***New Republic, 249***(5), 4-6.

Cohen, T. (2019). How to leverage artificial intelligence to meet your diversity goals. ***Strategic HR Review, 18***(2), 62-65.

Criado, J. I., J. Valero, J. Villodre, S. Giest, & S. Grimmelikhuijsen (2020). Algorithmic Transparency and Bureaucratic Discretion: The Case of Saler Early Warning System. ***Information Polity: The International Journal of Government & Democracy in the Information Age, 25***(4), 449-470.

Dafoe, A., & Journal of International Affairs (2018). Global Politics and the Governance of Artificial Intelligence (Interview with Allen Dafoe). ***Journal of International Affairs, 72***(1), 121-126.

Dastin, J. (2018). Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women. Retrieved September 28, 2022 from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

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/.

Desiere, S. A. M., & L, Struyven (2021). Using Artificial Intelligence to Classify Jobseekers: The Accuracy-Equity Trade-Off. ***Journal of Social Policy, 50***(2), 367-385.

Erwin, E. J., M. J. Brotherson, & J. A. Summers (2011). Understanding Qualitative Metasynthesis: Issues and Opportunities in Early Childhood Intervention Research. ***Journal of Early Intervention, 33***(3), 186-200.

Farer, T. (1992). The Hierarchy of Human Rights. ***American University International Law Review, 8***, 115-119.

Ferguson, A. G. (2015). Big Data and Predictive Reasonable Suspicion. ***University of Pennsylvania Law Review, 163,*** 327-410.

Finfgeld, D. L. (2003). Metasynthesis: The State of the Art – So Far. ***Qualitative Health Research, 13***(7), 893-904.

Garvie, C., A. Bedoya, & J. Frankle (2016). The Perpetual Line-Up: Ungregulated Police Face Recognition in America. In ***Secondary The Perpetual Line-Up: Ungregulated Police Face Recognition in America***, ed. Secondary Garvie, Clare, A. Bedoya, and J. Frankle. Washington, DC: Center for Privacy & Technology, Georgetown Law.

Gillis, T. B., & J. L. Spiess (2019). Big Data and Discrimination. ***University of Chicago Law Review, 86***(2), 459-487.

Goralski, M. A., & T. K. Tan (2020). Artificial intelligence and sustainable development. ***The International Journal of Management Education, 18*** (1), 100330.

Head, Brian W., & J. Alford (2015). Wicked Problems: Implications for Public Policy and Management. ***Administration & Society, 47***(6), 711-739.

Hellman, D. (2020). Sex, Causation, and Algorithms: How Equal Protection Prohibits Compounding Prior Injustice. ***Washington University Law Review, 98***, 481-523.

Horowitz, M. C. (2016). Public Opinion and the Politics of the Killer Robots Debate. ***Research & Politics, 3***(1).

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.

IEEE (2019). IEEE Position Statement: Artificial Intelligence (Approved by the IEEE Board of Directors) (24 June 2019) Retrieved September 28, 2022 from https://globalpolicy.ieee.org/wp-content/uploads/2019/06/IEEE18029.pdf

Johnson, S. L. J. (2019). Ai, Machine Learning, and Ethics in Health Care. ***The Journal of legal medicine, 39***(4), 427-441.

Jonker, C., J. Snoep, J. Treur, H.V. Westerhoff, & W. C. A. Wijngaards (2002). Putting Intentions into Cell Biochemistry: An Artificial Intelligence Perspective. ***Journal of theoretical biology, 214***(1), 105-134.

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

Jora, R. B., K. K. Sodhi, P. Mittal, & P. Saxena (2022, March). ***Role of Artificial Intelligence (AI) In meeting Diversity, Equality and Inclusion (DEI) Goals***. Conference: 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore.

Kavlakoglu, E. (2020). Ai Vs. Machine Learning Vs. Deep Learning Vs. Neural Networks: What's the Difference? Retrieved September 28, 2022 from https://www.ibm.com/cloud/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks.

Kleinberg, J., J. Ludwig, S. Mullainathan, & A. Rambachan (2018). Algorithmic Fairness. ***AEA Papers and Proceedings, 108***, 22-27.

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

Koji, T. (2001). Emerging Hierarchy in International Human Rights and Beyond: From the Perspective of Non-Derogable Rights. ***European Journal of International Law, 12***, 917-941.

Levitt, H. M. (2018). How to Conduct a Qualitative Meta-Analysis: Tailoring Methods to Enhance Methodological Integrity. ***Psychotherapy Research, 28***, 367-378.

Levitt, H. M., A. Pomerville, & F. I. Surace. (2016). A Qualitative Meta-Analysis Examining Clients' Experiences of Psychotherapy: A New Agenda. ***Psychological Bulletin, 142***, 801-830.

Levitt, H. M., A. Pomerville, F. I. Surace, & L. M. Grabowski (2017). Metamethod Study of Qualitative Psychotherapy Research on Clients' Experiences: Review and Recommendations. ***Journal of Counseling Psychology, 64***, 626-644.

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.

Maxwell, J., & J. Tomlinson. (2020). Proving Algorithmic Discrimination in Government Decision-Making. ***Oxford University Commonwealth Law Journal, 20***, 352-360.

Miller, D. (1997). Equality and Justice. ***Ratio, 10***, 222-237.

Miller, D. (1999). ***Principles of Social Justice***. Cambridge, MA: Harvard University Press.

Molnar, P., & L. Gill (2018). Bots at the Gate: A Human Rights Analysis of Automated Decision-Making in Canada's Immigration and Refugee System. In ***Secondary Bots at the Gate: A Human Rights Analysis of Automated Decision-Making in Canada's Immigration and Refugee System***, ed Secondary Molnar, Petra, and L. Gill. Toronto, YTO: University of Toronto.

Moore, M. H. (1995). ***Creating Public Value: Strategic Management in Government***. Cambridge, MA: Harvard University Press.

Moore, M. H. (2013). ***Recognizing Public Value***. Cambridge, MA: Harvard University Press.

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.

Pogge, T. W. (2008). ***World Poverty and Human Rights*** (2nd ed.). Cambridge, UK: Polity Press.

Prince, A. E.R., & D. Schwarcz (2020). Proxy Discrimination in the Age of Artificial Intelligence and Big Data. ***Iowa Law Review, 105***, 1257-1318.

Quintavalla, A., & K. Heine (2019). Priorities and Human Rights. ***The International Journal of Human Rights, 23***, 679-697.

Reis, J., P. E. Santo, & N. Melão (2019, July). Impacts of Artificial Intelligence on Public Administration: A Systematic Literature Review. Conference: 14th Iberian Conference on Information Systems and Technologies (CISTI), Coimbra.

Roff, H. M. (2014). The Strategic Robot Problem: Lethal Autonomous Weapons in War. ***Journal of Military Ethics, 13***, 211-227.

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

Sales, L. (2020). Algorithms, Artificial Intelligence and the Law. ***Judicial Review, 25***, 46-66.

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.

Schiffer, Z. (2020). Google Fires Prominent Ai Ethicist Timnit Gebru. Retrieved September 28, 2022 from https://www.theverge.com/2020/12/3/22150355/google-fires-timnit-gebru-facial-recognition-ai-ethicist .

Schreiber, R., D. Crooks, & P. N. Stern (1997). Qualitative Meta-Analysis. In ***Completing a Qualitative Project: Details and Dialogue***, ed. J. M. Morse. Thousand Oaks, CA: Sage. 311-326.

Setiya, K. (2018). Intention. The Stanford Encyclopedia of Philosophy (Fall 2018 Edition), Edward N. Zalta (ed.), Retrieved from https://plato.stanford.edu/archives/fall2018/entries/intention.

Sharkey, N. (2010). Saying 'No!' to Lethal Autonomous Targeting. ***Journal of Military Ethics, 9***, 369-383.

Sousa, W. G. d., E. R. P. d. Melo, P. H. D. S. Bermejo, R. A. S. Farias, & A. O. Gomes (2019). How and Where Is Artificial Intelligence in the Public Sector Going? A Literature Review and Research Agenda. ***Government Information Quarterly, 36***(4), N.PAG-N.PAG.

Sparrow, R. (2016). Robots and Respect: Assessing the Case against Autonomous Weapon Systems. ***Ethics & International Affairs, 30***, 93-116.

Suárez M. F. (2019). The Hierarchy of Human Rights and the Transcendental System of Right. ***Human Rights Review, 20***, 47-66.

Sun, T. Q., & R. Medaglia (2019). Mapping the Challenges of Artificial Intelligence in the Public Sector: Evidence from Public Healthcare. ***Government Information Quarterly, 36***, 368-383.

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

Thorne, S., L. Jensen, M. H. Kearney, G. Noblit, & M. Sandelowski (2004). Qualitative Metasynthesis: Reflection on Methodological Orientation and Ideological Agenda. ***Qualitative Health Research, 14***, 1342-1365.

Timulak, L. (2014). Qualitative Meta-Analysis. (***The Sage Handbook of Qualitative Data Analysis***, ed. Uwe Flick). London, UK: SAGE Publications Ltd..

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

Toohey, L., M. Moore, K. Dart, & D. Toohey (2019). Meeting the Access to Civil Justice Challenge: Digital Inclusion, Algorithmic Justice, and Human-Centered Design. ***Macquarie Law Journal, 19***, 133-156.

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.

Wangmo, T., M. L., R. W. Kressig, & M. Ienca (2019). Ethical Concerns with the Use of Intelligent Assistive Technology: Findings from a Qualitative Study with Professional Stakeholders. ***BMC medical ethics, 20***, 98.

Wasilow, S., & J. B. Thorpe (2019). Artificial Intelligence, Robotics, Ethics, and the Military: A Canadian Perspective. ***AI Magazine, 40***(1), 37-48.

Webb, M. (2019). The Impact of Artificial Intelligence on the Labor Market. Available at ***SSRN Electronic Journal***. Retrieved from http://dx.doi.org/10.2139/ssrn.3482150.

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

Wirtz, B. W., P. F. Langer, & C. Fenner (2021). Artificial Intelligence in the Public Sector - a Research Agenda. ***International Journal of Public Administration, 44***, 1103-1128.

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.

Zuiderwijk, A., Y.C. Chen, & F. Salem. (2021). Implications of the Use of Artificial Intelligence in Public Governance: A Systematic Literature Review and a Research Agenda. ***Government Information Quarterly, 38***(3), 1-19. Retrieved September 28, 2022 from https://www.sciencedirect.com/science/article/pii/S0740624X21000137.

附錄

研究論文	研究目的	國家/政府層級/政府單位	應用領域	人工智慧的應用方式	不正義的現象 正義-平等	導致不正義的原因	政策建議
Takshi, 2021	美國歷史上的群體差異已融入健康照護系統中，但美國聯邦食藥署對於 AI 使用者的規範很有限，導致 AI 應用於健康照護時，無法消除既存的歧視。本文從各面向討論 AI 應用於健康照護的義務以及如何解決歧視問題。	美國/聯邦與州/UnitedHealth Group，其旗下的事業體 UnitedHealthCare Community and State 提供州政府與聯邦政府共同提供經費的醫療照護計畫（Medicaid）。	醫療照護	Optum 公司所發展的演算法被用來判斷病患風險等級，依此決定後續醫療照護。	不同種族在醫療照護系統可近性的差異，被轉化進入演算法中，造成風險評估上的種族偏差，非裔病人被歸類為高風險的機率非常低。	美國社會中系統性的歧視一直存在於醫療照護系統中。例如非裔被認為比較耐痛；身心障礙女性的症狀常被忽略；低收病人無能力獲得適當的醫療。這些群體間的偏差存在於大數據中。若 AI 系統以病患的醫療支出為參考依據，那麼自然會出現人種偏差。再者，AI 對於辨識非裔慣用語及圖像有困難，這也導致不夠正確的診斷。由於演算法是由私人公司發展販售，因此基於私人公司的所有權，外人無法進行監控。	呼籲 AI 產業能自我管制，設定非歧視的標準讓機器學習，也建議應有一個更高層級的管理者來解決 AI 所帶來的差別影響。
Obermeyer et al., 2019	以一個教學醫院的病人資料為基礎，利用普遍使用於全國的 AI 風險計算技術，進行實驗，說明使用不同的判斷標準可能產生的種族偏差。	美國/N.A./N.A.	醫療照護	風險分數計算會使用病人過去的醫療支出為評估標準之一，而風險分數決定了病人會受到哪些醫療照護。	在同一個風險分數之下，非裔會比白人病得更嚴重，也就是非裔的風險被低估。	使用病患過去的醫療支出為評判標準之一時，歷史因素會影響當前的判斷。由於美國的非裔會因為以下幾個因素而減少就醫： 1. 貧窮 2. 若醫生為白人，通常較不會對非裔病患安排預防性醫療行為 3. 非裔長久以來不太相信美國的醫療體系，非到需要急診時不太喜歡進醫院（相較於白人，非裔花在急診的醫療經費高出許多）。	在導入資料時，應慎選要導入的資料類別。
Winter & Davidson,	本文主要檢視一個具有爭議	英國/國家/	醫療照護	DeepMind Health 利用皇	利用演算法評估病人健康問	N.A.	加強對隱私資料的規

研究論文	研究目的	國家/政府層級/政府單位	應用領域	人工智慧的應用方式	不正義的現象 正義-平等	導致不正義的原因	政策建議
2019	性的公私部門夥伴關係，公部門為皇家自由信託，私部門為生產 AI 產品 DeepMind Health 的 Alphabet 公司。本研究聚焦於大數據在公私部門之間擷取、流動、使用時，所引發的道德問題。	皇家自由信託（Royal Free Trust），這是一個國家健康服務醫院系統（National Health Service hospital system）		家自由信託提供的醫療數據，開發人工智慧應用程式，可警示醫生有關其病患惡化的狀況。	題時，評估結果精準度出現群體間的差異，某些群體較為精準，某些群體則否。這些評估結果會加強、加速既有的健康差異，演算法延伸出非預期的歧視結果。		範，發展演算法的稽核機制，透過實驗來找出演算法潛藏的歧視。
Sun & Medaglia, 2019	中國的健康醫療體系導入 AI 所面臨的挑戰。	中國/ 國家/ 科技部、國家發展改革委員會。	醫療照護	中國某些地區的醫療系統導入由 IBM 公司發展的 AI 系統 Watson，用於回答病人提問並設計個人化醫療照護計畫。	N.A.	N.A.	N.A.
Wangmo et al., 2019	檢視應用 AI 輔助系統於老人照護與失智照護時的道德議題。	歐洲國家（瑞士、義大利、德國）/ N.A./ N.A.	醫療照護	使用智慧輔助技術於老人殘疾、失智者。	因為成本很高，有需求者並非都能得到這套智慧輔助技術	成本太高，硬體軟體都不便宜。	目前這套系統在高收入與中低收入之間存在著使用落差，但隨著製造成本降低，與社會保險的補助，應可解決分配不平等問題。
Maxwell & Tomlinson, 2020	從一個 2020 年的英格蘭與威爾斯上訴法院（England and Wales Court of Appeal）的判決 R（Bridges）v South Wales Police 來討論預防 AI 歧視	英國/ 區域/ 英國南威爾斯警署(South Wales Police)	警察	警方使用自動臉部識別技術比對活動中的民眾臉部資訊，而資料庫清單包含通緝犯、逃離羈押的嫌犯、特別弱勢族群，因特定情報而受關注之對象。	人臉辨識系統的應用會對女性與有色人種造成差別判定。	演算法基於大數據進行學習，而大數據反應了既有的不平等，這可能會有系統地被複製。	判決內容中，上訴法院判定南威爾斯警察署在使用人臉辨識系統之前，有責任先確認這套系統不會造成歧視。

研究論文	研究目的	國家/政府層級/政府單位	應用領域	人工智慧的應用方式	不正義的現象 正義-平等	導致不正義的原因	政策建議
	的責任在於政府部門。						
Lum & Isaac, 2016	用 PredPol 公司發展的演算法加上奧克蘭警察局的大資料進行預測，將預測結果比對全國調查資料與地方逮捕資料，說明基於機器學習的預測系統會產生偏差結果。	美國/地方（加州奧克蘭市 Oakland, California）/警察局	警察	在美國普遍被警察機關使用的預測系統，用來找出可能犯罪的人。	雖然毒品犯罪到處都是，但因毒品而遭逮捕者通常發生在非白人與低收入者居住的社區。	輸入機器學習的數據中，包括地理資訊與逮捕紀錄，依此產生可能犯罪者所在區域，就成為警察加強巡邏之地，而警察在這些區域的逮捕紀錄，又成為機器學習的數據，成為自我增強的循環。	N.A.
Ferguson, 2015	討論大數據普遍應用於刑事司法體系的情況下，美國憲法第四修正案理論的發展。首先本文提到傳統的合理懷疑在大數據應用下遭遇挑戰。其次本文檢視大數據應用在警察執法過程中如何挑戰傳統警察執法的各層面。第三本文試圖提供解套建議。	美國/地方/警察機關	警察	使用國家犯罪資料的大數據來預測地方嫌疑犯，在嫌疑犯還沒有做任何事之前，先找到他。	有犯罪紀錄的人最容易被盤查，但窮人與有色人種又最容易留下犯罪紀錄，前者是因為付不出保釋金，後者則是結構性因素使然。人工智慧或大數據往往以超然客觀的形象存在，使人失去戒心。	1. 資料品質：記錄不正確的原始資料又與其他資料整合，會產生嚴重影響，妨害個人自由。而資訊不透明也難以課責。 2. 資料到底如何相互連結，沒有人知道，演算法如何找出數據與數據之間的相關性，使用者也不知道。	1. 以大數據為基礎來執法時，在“合理懷疑”上需要有更嚴謹的標準。 2. 從資料庫下手：分析資料，確認在不同犯罪類型之下，“可能犯罪率”的判斷門檻應該不同。還有，應加入犯罪時間與地點的精確判斷，因為犯罪總有地緣關係或時間關係。 3. 使用者須了解資料與資料是如何連結的。

研究論文	研究目的	國家/政府層級/政府單位	應用領域	人工智慧的應用方式	不正義的現象 正義-平等	導致不正義的原因	政策建議
Garvie, Bedoya & Frankle, 2016	評估 25 個州與地方執法單位使用臉部辨識系統時，對於隱私權、公民自由、民權、透明、與課責的影響。	美國/ 州與地方政府/ 警察單位	警察	臉部辨識系統。	警察單位所使用的臉部辨識系統，不成比例地影響非裔。	臉部辨識系統應用於非裔的正確率比較低，由於非裔的被捕率不成比例高於白人，嫌疑犯照片資料庫中會有不成比例數量的非裔。	國會與州立法部門應通過法律來規範執法單位使用臉部辨識系統。開發臉部辨識系統的私人公司應測試其演算法對種族、性別、年齡的偏見。公司也應自願公布臉部辨識的績效，讓警察單位或市議會有比較的基準。
Saunders, Hunt & Hollywood, 2016	釐清芝加哥警察部門自 2013 年啟用的槍枝暴力預防機制--策略性目標名單的影響，並檢視導致這個結果的可能原因。	美國/ 市政府/ 芝加哥警察部門	警察	基於“凶殺案受害者”的社會網絡連結與被逮捕紀錄，計算風險分數，分數高者被置入策略性目標名單中，成為槍枝暴力的可能關係者，這表示名單中的人比較可能槍殺別人或遭到槍殺。警察部門會基於此名單採取預防性干預措施，至於干預措施的作法，各區域指揮官有極大的裁量權。	策略性目標名單裡的人成為槍枝暴力關係者的可能性，無異於其他人，但他們因槍枝暴力遭逮捕的機率卻比其他人高。這引發民權與隱私權的質疑，特別是這個名單所關注的對象，其實是社會中高風險受害的脆弱群體。	策略性目標名單是基於歷史資料而產生的，而歷史資料有明顯偏差。再者，名單成員之所以更容易被逮捕的原因，也在於警察可能用這個名單來為槍殺案件結案。	N.A.
Brenner et al., 2020	檢視一套由私人公司開發並應用於許多州與地方司法系統的風險評估技術--	美國/ 州政府（佛羅里達，威斯康辛，密西根，紐約州大部分的郡，新墨西	刑事司法	COMPAS 應用於司法判決的各個階段，例如用來判斷是否給予假釋或緩刑，也用	用兩個標準來測量公平性： 1) 人口差異：COMPAS 計算的分數中，白人與	1. 正當程序：委託私人公司開發 AI 系統，資訊不透明，無從得知其評估分數如何產生。 2. 種族歧視：雖然	本文建議思考，各州是否有足夠的裝備解決意圖或非意圖的種族歧視

研究論文	研究目的	國家/政府層級/政府單位	應用領域	人工智慧的應用方式	不正義的現象 正義-平等	導致不正義的原因	政策建議
	COMPAS，在合法性上的問題。	哥州人口最多的城市）/法院		在審前釋放與保釋的判斷。	非裔的差異很大。非裔拿低分的比例比白人高。 2) 機率的公平性：即COMPAS判斷錯誤的機率。COMPAS針對非裔做出偽陽判斷的機率明顯高於白人，而偽陰判斷的機率卻明顯低於白人。	COMPAS用了一百多個指標來進行判斷，這些指標涵蓋了犯罪歷史與個人特質。即使判斷指標中沒有種族一項，但犯罪歷史、社會經濟狀態等指標本身就有種族偏差。	問題？為了解決歧視問題，應提升正確性與課責，發展一個可解釋的演算法，追溯演算邏輯。此外，還可透過另一種機器學習的方式，來解釋使用中的演算法，並把複雜的演算法用簡單的方式提供給被告。
Koepke & Robinson, 2018	檢視應用於法院審判前的風險評估機制。	美國/州與地方/法院	刑事司法	審判前利用風險評估工具來判斷被告被保釋後的風險。	司法改革想降低種族或經濟所造成的偏差與不平等，但法院在進行審判之前利用風險評估機制來協助判定，反而無法達到司法改革的目標，因為此評估機制是基於舊資料進行機器學習。	基於舊資料的學習。	應專注於最嚴重的風險做為資料基礎，因為有些人只是付不出保釋金所以入監服刑。此外，應增加監控與課責機制，常態性的比較預估與實際狀況之間的差距，並公開資訊。
Angwin, Mattu & Kirchner, 2016	檢視風險評估系統 COMPAS 所計算的風險分數。	美國/佛羅里達州的地方政府（Broward County）/法院	刑事司法	COMPAS 基於 137 個問卷題項得到的答案，計算風險分數供法院做判決時參考。	非裔被告被預測會累犯的比例約為白人的兩倍，而白人被告比非裔被告較容易被評估為低風險。	雖然在 COMPAS 用以評估的 137 個問題中，並無涉及種族資訊，但其中許多問題卻可能與種族有關（例如父母是否曾經入獄），歷史資料顯示，非裔入獄率不成比例地比白人高。	N.A.
Cockerill, 2020	從傳統的生物倫理原則架構分析 AI 應用	美國/N.A./法院、醫院	司法精神醫學（Forensic	利用 AI 進行深度學習，建構暴力預測系	1. 同樣是公民，卻在還沒有犯罪事	歷史資料偏差導致機器學習偏差，而且預防犯罪與妨礙個人自	電腦科學與精神醫學之間的互動關

研究論文	研究目的	國家/政府層級/政府單位	應用領域	人工智慧的應用方式	不正義的現象 正義-平等	導致不正義的原因	政策建議
	於暴力風險評估與預測機制所產生的道德問題並提供建議。		psychiatry）	統來決定是否留置可能會有暴力行為的人。這套系統應用於許多重要的司法決策過程中，會影響刑期、非自願醫療用藥的判定。	實之前，就被預測暴力傾向，進而被限制個人自由或自主。 2. 在犯同一種罪的前提下，非裔與拉丁裔的刑期通常比白人長，而這種偏差會進入暴力風險評估中，導致有色人種更被差別對待。	由之間兩難。	係應放入醫學院的基本訓練中，司法精神醫學領域要加強相關訓練、特別是要懂得如何轉譯資訊給法官。跨領域的學習是必要的，而且司法精神醫學者需要能夠進入相關決策過程，才能貢獻實務經驗。
Hellman, 2020	威斯康辛州最高法院決定將性別因素納入罪犯風險評估的演算法中，本文認為目前對於性別該在何時放入考量的原則非常不明確，因此提供一個分析途徑。	美國/州最高法院/法院	刑事司法	利用風險評估機制時，威斯康辛州最高法院決定可以用性別資料來預測罪犯未來的累犯率，並將這個資訊傳達給決定刑期的法官。	女性犯罪比男性少，當性別因素被允許置入演算法，女性將會被判較輕的刑期，也比較容易被假釋，如此會有違平等保護原則。	使用歷史資料與性別資料。	不應使用賠償邏輯來看性別歧視，否則會在大數據與機器學習的推波助瀾之下加劇不平等。
Chouldechova, 2017	本篇檢視COMPAS所使用的預測標準與其公平性，發現該系統對不同群體的預測有所差異導致不公平。	美國/佛羅里達州的地方政府（Broward County）/法院	刑事司法	基於 137 個問卷題項，由被告自己填答或犯罪紀錄所得到的答案，來計算被告的風險分數，交給法院判決時做參考。	由於累犯預測系統在不同種族之間會出現不同的偽陽率與偽陰率，相較於白人被告而言，非裔被告會有比較高的比例出現偽陽，較低的比例出現偽陰。這使得非裔被告較容易被判定為高風險者，因而被判較重的刑罰。	雖然 COMPAS 的開發公司 Northpointe 聲稱該系統已經通過公平標準的測試，但本文發現其所使用的測試標準與該系統所產生的偽陰與偽陽錯誤有所關聯，所以質疑該公司的測試結果。	我們需要確認這套預測工具所產生的偏差影響並權衡其得失。

研究論文	研究目的	國家/政府層級/政府單位	應用領域	人工智慧的應用方式	不正義的現象 正義-平等	導致不正義的原因	政策建議
Brownlie, 2020	本文討論 AI 的自動決策系統需要被規範的急迫性，並提出立法建議。	紐西蘭/國家/意外事故賠償公司（Accident Compensation Corporation, ACC）	意外事故賠償	演算法用來分析過去的理賠資料，進而決定是否把理賠個案歸類為複雜的個案，若是，則由人類接手處理，若否則直接核准。	申請者若存在著健康問題或身心障礙，則其理賠申請會被歸類為複雜個案，即使是簡單的個案。	此分類判斷是基於歷史資料，其中申請者若為身心障礙或有健康問題，其申請案通常比較不會被核准。	紐國目前的法令不足以避免民眾免於被自動決策制定系統歧視，提出監控機制／立法建議。
Allen, 2019	檢視大數據與演算法在住宅領域的應用所可能產生的潛在不平等，特別討論從 1930 年代開始對於特定社區居民拒絕給與房貸的策略如何延續到大數據時代。	美國/國家/政府出資的房貸公司	房屋貸款	從 30 年代開始，美國政府的屋主貸款公司（United States government's Home Owners' Loan Corporation）就針對地圖上的有色人種社區特別畫上紅線，代表這些社區住民風險太高不予貸款。到了現代，貸款公司也會拒絕高風險地區住民的貸款申請。	以前被劃在紅線內歸類為高風險的區域大都為有色人種社區，居民會被排除於房貸核可名單之外。因此，即使一個人在財務上符合貸款門檻，但可能被歸類為高風險而無法獲准房貸。這也連帶影響到有色人種能否購買到可負擔的住宅，以及政府出資的公共住宅計畫也因此較少選擇在高風險地區興建，反而集中於白人區。	即使近代很多規定都明文禁止以種族來衡量房屋貸款的核准決定，但卻又規定個人資料的蒐集上必須涵蓋種族這個選項。此外，雖然用於評估租賃申請的演算法並無將種族納入考慮，但由於所使用的大數據資料包含了信用分數、被房東驅離的紀錄、逮捕紀錄、教育程度、工作背景、與先前的住址，這些資料與種族有這明顯的關聯。	資訊透明、稽核、人工監測自動決策的結果、演算法應有更多的課責。
Desiere & Struyven, 2021	討論以 AI 為基礎的求職者分類方式，在正確性與公平性之間的權衡。	比利時/區域（Flanders 佛萊明區）/佛萊明區公共就業服務系統 VDAB	公共就業（求職者職訓、工作媒合）	公共就業服務單位利用 AI 為求職者分類，分類資訊決定了求職者會得到什麼服務。	求職者若屬於弱勢群體（例如移民、身心障礙、年長者）比較容易被錯誤歸類為高風險，所以通常會被優先聯絡處置。然而這問題在於 1. 被連絡者通常得面對求職過程的監	AI 的分析模型是基於既有的資料做分析的。	政策制定者與個案工作者必須了解 AI 分析模型的限制，特別是應權衡正確性與公平性。

研究論文	研究目的	國家/政府層級/政府單位	應用領域	人工智慧的應用方式	不正義的現象 正義-平等	導致不正義的原因	政策建議
					控，職業訓練也可能並不適合，造成求職者的負擔。 2. 這反而讓其他真正需要被優先連絡處置的人失去機會。		
Molnar & Gill, 2018	從人權觀點檢視加拿大移民與難民系統採用自動決策機制所產生的影響，特別聚焦於演算法與自動化科技應用在一個已經具有高度裁量性的系統時，不論是取代或協助行政決策所可能產生的缺失。	加拿大/國家/凡是與加拿大移民與難民系統相關的單位，例如移民辦公室、海關、司法部門等等。	國境管理與國土安全--移民與難民	透過風險（risk）、優先順序（priority）、複雜性（complexity）等指標，為個案進行分類，再計算個案分數、做機率評估、再配合其他指標來提供人為決策的參考。過程中，針對特定個案會以警告標記提醒審查官員，同時針對是否同意個案申請，提供整體的評估建議。	這套系統違反了許多人權，有偏差、歧視、違反隱私權、正當程序、程序公平。違反的人權包括平等權、非歧視、自由遷徙、自由表達、宗教結社權。使用歧視與有偏差的演算法導致申請人無法得到公正的判斷。	這套系統利用過去的決策資料作為訓練基礎，而既有的偏見存在於歷史資料中。例如，加拿大已經用一套粗糙的預測演算法來評斷一個國家是否「安全」，來自「安全」國家的難民申請庇護較不會被核准。這套演算法基於幾個原則作為判斷標準，包括該國是否產生難民、是否尊重人權、是否提供國家保護等等。然而，這些標準對於安全的定義是很不完整的，因為有些被認為安全的國家，事實上他們對於某些群體而言是不安全的，例如非異性戀群體，或是逃離家暴的女人。	持續發表詳細的研究報告，停止開發新的 AI 直到目前的調整符合人權要求。另外，應組成公正第三方隨時監控 AI 的使用結果，發展評估指標與適合的研究方法，而且利害關係者應對 AI 系統多一些了解，並建議除了隱私與國家安全理由以外，應公開資料以受公評。
Wasilow & Thorpe, 2019	檢視 AI 應用到軍隊所可能遇到的問題（包括道德問題），本文建立一個架構來檢視這些道德問題。	加拿大/國家/國防單位	國防	類神經網路能審視監控影片並找出哪些物件是值得注意的，例如武器、車輛、人，藉此提醒戰場中的軍人正在面臨的威脅。臉部辨識軟體是其中一種協尋人物的工具，而 AI	由於戰區通常位於偏遠、人跡罕至、或不易進出之地，該地歷史資料的蒐集較為不足，因此依賴機器學習所提供的警示而做的決定，會造成非常負面的影響。例如西方國家發展的	歷史資料太少。	作者提出一個符合加拿大國家法規的分析架構，讓 AI 發展公司參考。

研究論文	研究目的	國家/政府層級/政府單位	應用領域	人工智慧的應用方式	不正義的現象 正義-平等	導致不正義的原因	政策建議
				則能幫助軍隊從大量的監控資料中找出資料與資料之間的關係，以提供警示並協助戰場中快速決策。此外，AI 機器人系統還能幫助執行危險任務，降低軍隊的傷亡。	人臉辨識系統，對於戰區某些人種與性別的辨識不精確。		
Howard & Borenstein, 2018	討論既有偏見如何被融入當代 AI 與機器人系統，以及這種系統設計會產生哪些影響。	紐西蘭、美國/ 紐西蘭：國家，美國：州與地方/ N.A.	出入境、司法、警察	1. 紐西蘭：利用臉部辨識系統於護照申請。 2. 美國：司法系統利用演算法來預估嫌犯未來是否會累犯，而這資訊會影響刑罰輕重或假釋的可能。 3. 警察系統用於預測可能的犯罪行為或犯罪者。	1. 紐西蘭的自動護照申請系統無法正確辨識亞洲人臉（以為眼睛都閉著） 2. 非裔被 AI 系統歧視而評斷為較具危險性 3. 警用預警系統對於特定群體產生錯誤判讀。	從歷史資料學習。	需要發展一些檢測機制來制止或減少既有的偏見滲透至機器人技術中。
Madden et al., 2017	本文針對就業、教育、警察三個領域做個案研究。	美國/ N.A./ 私部門、公私立大學、警察機構	求職、教育、司法警察	演算法應用於公立大學或警察機構較符合本研究主旨，所以只針對這兩部分進行說明。 1. 教育：基於學生在社群平台的發言與照片進行分析，可提供即時的洞見，預測誰比較可能完成學業。大學入學審查	1. 教育：此套系統對於不懂如何在社群平台管理隱私的人造成歧視。低收入社群平台使用者的資料比高收入者更容易被取得，而針對他們的網絡連結所進行的分析，就更可能反應了其根深蒂固的	貧窮者對網路資料搜尋、隱私保護的知識不足（數位知識落差），導致進入惡性循環，他們的網路行為會成為他們找工作、受教育、被司法審視時的參考資料，使其更為弱勢。還有，貧窮者一直以來就比他人受到更多的監督，例如電子食物券與社會福利金的使用資料會留下紀錄，受到監控，而這個資料會與其他部門的資料連結，使他們受到	從各種相關的法律架構來著手，注重程序正義、注意數位知識（包括道德、認知、社會技巧）、充分告知社群平台的使用者相關的數位知識。

研究論文	研究目的	國家/政府層級/政府單位	應用領域	人工智慧的應用方式	不正義的現象 正義-平等	導致不正義的原因	政策建議
				可以藉此篩選想招收的學生，包括學生在社群平台上與同儕之間的互動方式，都會成為入學審查者有興趣的資訊。 2. 警察：利用預測工具來評估一個人的犯罪傾向，作為預防犯罪的工具。	結構性劣勢，那麼基於這種分析所得到的結果，少數群體的大學申請很可能會被不成比例地拒絕。 2. 警察：預測系統會不成比例地關注有色人種。	的監控更為綿密，這會影響一個貧窮者在申請工作、大學入學的結果。即使他們因此而清空自己的社群發言，還是可以分析他的朋友圈。就警察部門而言，用於分析風險威脅的資料來源多元，社群平台、犯罪歷史、就醫紀錄、甚至從私部門而來的資料，都可彙整使用，但資料的品質與正確性會影響判準。	
Toohey et al., 2019	本篇討論兩個AI時代下的重要挑戰：數位涵容與演算正義，並提供建議。	澳洲/聯邦政府/澳大利亞服務部（Service Australia）	社會福利	政府使用Centralink公司開發的Robo Debt（自動債務索償系統），該系統能自動連結社會福利接受者個人資料與稅務系統資料，當發現二者之間有歧異，就自動以政府的名義寄信給福利接受者，通知他們積欠政府的金額，或是政府應給而未給的金額。	由於演算法出現錯誤，導致許多福利接受者被通知欠款，更由於他們不見得有能力打官司或提出反駁，導致身心俱疲。該系統非意圖地傷害了社會中最脆弱的群體（社福接受者）。	演算法出現錯誤。	應用於公部門的數位科技，其設計必須以人為中心。

資料來源：本研究整理

Unintentional Discrimination in Application of Artificial Intelligence to Public Policies: A Systematic Article Review

Tsuey-Ping Lee*, Chu-Yi Chang**, Chen-Ling Lee***

Abstract

This study examined the ethical problems with the application of AI to public policy spheres, based on the principle of equality in citizenship from Miller's plural view of justice. In adopting the PRISMA model, a qualitative meta-analysis was employed to inspect institutional process and outcomes of AI applications. This research found that AI has been applied to various public policy fields including criminal justice, policing, health care, homeland security and border management, education, public finance, public employment, as well as national defense. In these fields, AI has made administrative work more efficient and has improved most people's well-being while creating unintentional discrimination against specific groups of people. An examination of the institutional process showed that the government has ignored the long-standing social injustice hidden in the big data used for machine learning. Consequently, the institutional outcome showed that historical injustice continues to be reproduced through AI, leading to differential treatment of specific groups and

* Tsuey-Ping Lee, Professor, Department of Political Science, National Chung Cheng University, e-mail: tsueyping.lee@gmail.com.
** Chu-Yi Chang, Undergraduate Student, Department of Political Science, National Chung Cheng University.
*** Chen-Ling Lee, Student, Taichung Municipal Taichung Girls' Senior High School.

deprivation of their basic human rights.

In order to analyze the pattern and nature of unintentional discrimination in various public policy areas, this study, based on the order of priority of human rights protection implied by international human rights-related conventions, analyzes the negative effects of AI on specific groups in terms of “whether the victims initiate the evaluation” and “negative and positive rights deprivation”. The research results showed that the application of AI in the areas of police enforcement, criminal justice, and health care involves the deprivation of negative rights such as the right to life and the right to freedom, which urgently needs to be addressed. This paper concludes by discussing why the correction of unintentional discrimination cannot be done by civil society but requires the active intervention of the government. This paper ends by suggesting specific actions that the government should take in the preparatory and implementation stages of AI applications in order to reduce the unintentional discrimination of specific groups.

Keywords: Artificial Intelligence, technology ethics, unintentional proxy discrimination, technology justice, social equity