出國報告（出國類別：開會）

# 出席耶魯大學(Yale University)

# AI治理論壇(Governing【with】

# AI)

服務機關：數位發展部
姓名職稱：何宜臻系統分析師
薛雅婷專案規劃師
派赴地國家/地區：美國/紐哈芬
出國期間：113年2月27日至3月3日
報告日期：113年5月23日

# 摘　要

海倫·蘭德莫爾(Hélène Landemore)為耶魯大學社會與政策研究所教授，致力於民主創新計畫中領導有關公民會議的研究議程。於民國113年規劃辦理「Governing X」系列論壇，前電郵邀請本部參與，希冀就人工智慧在協助治理現代政體，並透過法國經典的全國公民會議經驗，像是108年舉行的公民氣候會議(Citizens Convention for Climate) ，及111年生命臨終公民會議(Convention on End of Life)進行交流；論壇目標除推進相關主題的研究，同時將產出創新思維、概念框架和政策或技術工具，以幫助解決問題，帶來公平和有幫助的解方。

透過本次論壇與國際前瞻性研究者交流，蒐集到目前國際上對於人工智慧治理的共識方向，了解目前研究者運用人工智慧推進民主過程的研究或執行方法，用以符合社會變革及國家未來數位發展的需要，確保人工智慧發展與群眾利益保持一致，展現民主價值。

# 目 次

## 壹、 目的

本部於民國112年5月正式成為國際非政府組織「集體智慧計畫」(Collective Intelligence Project, CIP)合作夥伴，參與對齊大會(Alignment Assemblies)專案，期待協助臺灣在全球公眾領域上，凝聚民眾對於人工智慧需求與風險之共識，共同處理「人工智慧對齊問題」(Alignment Problem)。因此於112年開始，本部陸續舉辦與補助以臺灣使用者為核心的對齊大會，透過公民參與審議模式，形塑人工智慧發展方向。

人工智慧帶來的變革，其演算法、智慧財產權、科技倫理、公共服務和社會影響等議題備受關注。為利掌握當前國際人工智慧治理的共識方向、操作模式，及與人工智慧前瞻性研究者建立交流人脈，爰派員參加此等國際論壇，透過參與對於就人工智慧在協助治理現代政體，與增強全民、參與和發展過程方面的角色進行交流，期望在技術發展快速的情況下，凝聚群眾智慧共同形塑以集體利益為發展前提的人工智慧。

## 貳、 過程

海倫·蘭德莫爾(Hélène Landemore)教授本次於耶魯大學舉辦治理X系列論壇，以兩個主題會議進行討論，主題分別是「公民會議的治理」和「人工智慧的治理」，自2月28日開始至3月1日結束，本次共參與的專題發表與研究報告場次，摘陳如下：

一、2月28日 公民會議的治理(Governing Citizens Assemblies)

(一) What Does Governing a Citizens' Assembly Mean: Power and Contingency

(二) The case for self-ruling citizens' assemblies: Governance, Representation, Voting in a process, and Citizen Leadership in the French Conventions

(三) Blueprint for Activated Citizenship

(四) Representation of Minorities Within Citizens' Assemblies and Stories of Discontent

(五) Should Citizens' Assemblies be more like sovereign parliaments or not?

二、2月29日 人工智慧與治理(Governing"with" AI) I

(一) Privacy and security issues in the use of AI in democratic processes

(二) Can a chatbot facilitate deliberation?

(三) Applying a power analysis to AI

(四) Can we make AI regulation legitimate?

(五) The advantages and disadvantages of AI facilitated deliberation

(六) Interactive Debates

三、3月01日 人工智慧與治理(Governing"with" AI) II

(一) How can we create a more deliberative society?

(二) Use cases of manipulation

## 參、公民會議的治理(Governing Citizens Assemblies)

歐洲是許多公民會議的實踐場域，像是愛爾蘭公民於105年組成公民代表大會來討論墮胎合法化、英國於106年舉行關於英國脫歐之協助與歐盟關係的公民會議、法國於111年舉行關於「協助自殺和安樂死」法律草案擬定等。公民會議是由目標抽選原則[1]隨機挑選之公民組成，賦予選出的成員參與政策制定或是立法提案，就政治問題進行審議，並挑選有爭議性且可引起社會關切之會議主題以吸引眾人討論，並且激盪出不同觀點，像是選舉改革、同性婚姻、安樂死、代理孕母、氣候正義等問題。

公民會議在民主治理中的角色日趨加重，但公民會議一直由政府官員、專家和專業引導者控制，究竟有多少程度的權力是真實屬於參與的公民本身？本次會議借鏡法國像是108年舉行的氣候會議(Citizens Convention for Climate)[2] ，以


圖3-1 耶魯大學社會與政策研究所教授Hélène Landemore致詞

[1] 目標抽選原則，依據不同議題，主辦單位需要吸引不同的目標受眾，全國性的議題可以向全國民眾宣傳，地方性議題則以當地居民為主。

[2] 法國氣候公民大會(Convention Citoyenne pour le Climat)，108年9月啟動，隨機抽籤挑選出150位公民，在為期9個月的時間一同討論並提出氣候建議草案。

及111年9月生命臨終公民會議(Convention on End of Life)[3]等經典全國公民會議經驗進而加以延伸子問題進行討論。本次我方參與討論主題如下：「公民會議的治理意味著什麼：權力與偶然性」、「自治公民會議案例：在法國公民會議中的治理、代表性、過程中的投票及公民領導力」、「積極公民權的藍圖」、「少數群體在大會中的代表性與不滿的故事」、「公民會議應該要像獨立議會嗎」，參與情形如下，並在最後附上各場講者資料。

## 一、公民會議的治理意味著什麼：權力與偶然性(What Does Governing a Citizens' Assembly Mean: Power and Contingency)

講者為波爾多大學社會學副教授Sandrine Rui，曾擔任法國生命臨終公民會議執行委員會(steering committee)[4]成員，講者說明公民會議的主要核心是透過平等、自主的機制，讓公民可以參與政策制訂的決策過程，一場凝聚公民共識會議，計畫人員來自各界社會團體和研究人員，包含確認內部合作機構，以及外部利害關係人的對話機制與流程設計；會議機制則是由任務性質的執行委員會負責規劃，包括抽選參與者條件、公民會議開會方式與評論。

在法國的生命臨終公民會議中，計畫成員和執行委員會之間的權責劃分存在著模糊地帶。舉例來說，在該次公民會議中，由於機制設計不夠清晰，執行委員會實際上沒有權力去制定公民會議的形式和規範，導致執行委員會的權力變得模糊不清，也使得公民會議的結論報告缺乏足夠的合法性和公信力。然而，雖然執行委員會和參與的公民仍遵循著既定的公民會議原則，但該次的經驗提供了一份寶貴的路徑圖。除了制定更清晰的權責分工項目和會議流程外，公民會議的過程中有許多細節是無法事先確定的。我們需要保留利害關係人之間的對話空間，

[3] 法國生命臨終會議(Convention on End of Life)，111年9月啟動，隨機抽籤挑選出184位公民，在為期3個月的時間一同討論並提出有條件開放輔助自殺與安樂死的結論報告。
[4] 治理委員會(governance committee)，負責規劃、監督整個共識會議，行政院青年輔導委員會：「審議式民主公民會議」操作手冊。

讓對話形成合作機會和創新形式，反思各個方面並建立共識。這也是公民會議的核心價值之一。

## 二、自治公民會議案例：在法國公民會議中的治理、代表性、過程中的投票及公民領導力(The case for self-ruling citizens' assemblies: Governance, Representation, Voting in a process, and Citizen Leadership in the French Conventions)

講者為耶魯大學社會與政策研究所(Institution for Social and Policy Studies，ISPS)民主創新計畫的博士後研究員Penigaud de Mourgues，他的觀點是關於公民會議的會議機制設計對於會議結論有很大的影響性，公共討論模式必須更為多元，才能讓不同溝通方式能充分呈現，雖然公民會議是以理性對話為主，但也要瞭解到參與者理性論述能力是有限的，並主張公民會議要有獨立性才能跳脫受權力組織的操控風險。

另一位講者為法蘭琪-康堤大學政治哲學博士生Chloé Santoro，透過法國生命臨終公民會議的觀察，說明設計投票的時間、方法與形式就對公民會議進行與結論都顯得至關重要，因為投票被視為匯集個人利益的決策機制，在集體參與中


圖3-2 誰該治理法國公約?(Who Governed the French Conventions?) 與談人

以投票方式去顯示偏好的形成與轉換，根據公民偏好的加總結果來形成結論報告，成為加總民主(aggregative democracy)[5]的模式。

## 三、積極公民權的藍圖(Blueprint for Activated Citizenship)

講者為策略政策顧問、實踐者和公民集會倡導者Marjan H. Ehsassi，她是法國生命臨終公民會議的擔保人，幫助監督其公正性，以及在年齡、性別、教育程度、職業和居住地等人口特徵的構成上呈現異質多元性。講者提出公民需要積極盡可能全程參與公民會議，才能確保政府在後續執行遵照公民提議，將結論完整提交至議會，後續透過立法或修法的形式去落實，建立一套可以共同參與的政策，加強公民與計畫成員間權力的共享，達成共同創造與共同負責的交流機制。

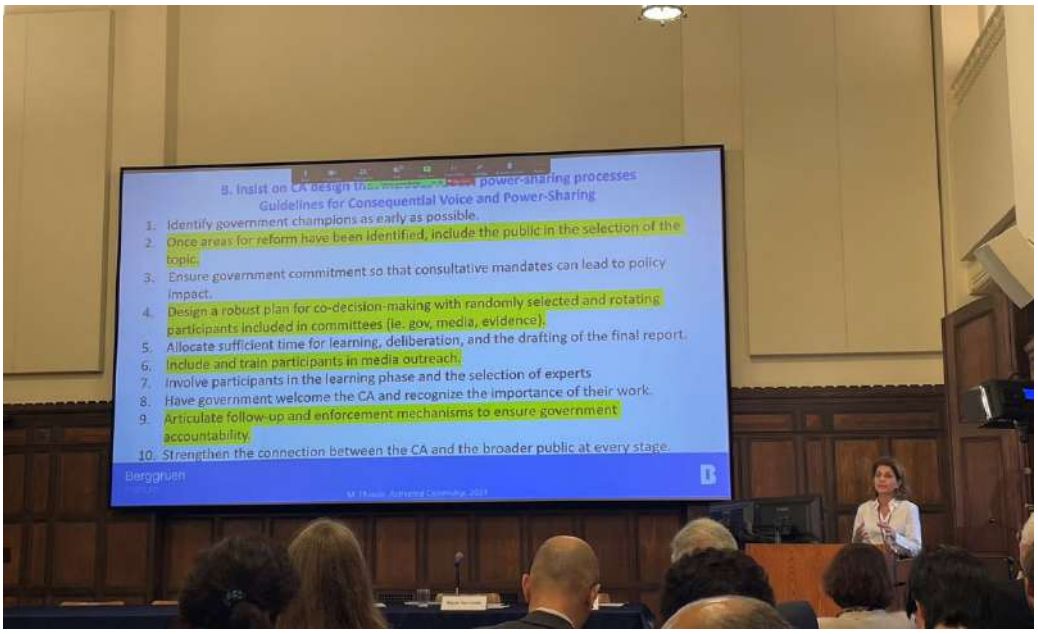


圖3-3 法國生命臨終會議的擔保人Ehsassi分享積極公民藍圖

## 四、少數群體在會議中的代表性與不滿的故事(Representation of Minorities Within Citizens' Assemblies and Stories of Discontent)

講者為耶魯大學政治學和經濟學專業的研究生Miguel Von Fedak，他主要研究法國生命臨終公民會議中少數意見群體的背景條件等，該公民會議由184名公

[5] 加總民主(aggregative democracy)，認為民主應致力於反映「共善」的全意志，認為個人偏好是既定的，因此在民主運作的程序上就是尋找將這些偏好加總起來的最佳方式。

民組成，其中76%的群體投票贊成推動立法，支持安樂死合法化，另外，24%的群體反對安樂死，研究24%群體背景條件包括他們的價值觀、宗教信仰、道德觀念等因素。

Fedak透過訪談發現，當多數意見的群體堅持的時間越長，公民會議就越容易呈現極端的意見分群，因為在執行過程中若過度強調達成共識，以多數人形成的偏好為決策的依據，在主流意見和團體壓力下，可能使得少數意見或是弱勢者的意見無法表達，形成社會霸權意識型態，少數意見者對此現象感到無能為力。


圖3-4 法國公民大會中公民的聲音、少數族群代表與領導者 (Citizen Voice and Minorities Representation, and Leadership in the French Citizens' Conventions)與談人

**五、公民會議應該要像獨立議會嗎？(Should Citizens' Assemblies be more like sovereign parliaments or not?)**

講者為法國經濟、社會和環境委員會當選主席Thierry Beaudet，目前負責組織公民集會。他表示，公民集會將有助於打破人民和政府之間的不信任，Beaudet說：「我確實相信審議式公民會議對於重建民主和代議制民主來說是一個非常令人鼓舞的因素，我根本不認為這是競爭或替代，而是互補」，參與式的治理被視為是一種解決立場對立且溝通複雜問題的重要方式。

相較之下，耶魯大學社會與政策研究所教授Hélène Landemore更希望看到自治的審議式公民會議擁有自己的立法權，也指出這將會與代議式民主（透過投票選出立委來代表人民，並實行公權力）產生衝突。在不同立場的辯論下，她還是強調審議式公民會議是要讓社會保持批判性和提出建設性的措施建議，去彌補加總民主的不足，而不受制於民粹主義者和民主敵人。

圖3-6 公民會議應該要像獨立議會嗎？(Should Citizens' Assemblies be more like sovereign parliaments or not?)與談人

## 六、各場講者資料

上述為期一天的會議將彙集學者、政治領袖和實踐者來探討這些關鍵議題，重點將放在氣候會議和生命臨終公民會議的法國公民公約上，這些公約明確的提出了治理的問題點，以下為所有講者資訊：

(一) 政策制定者

- Thierry Beaudet –法國經濟社會暨環境諮詢委員會主席。
- Gaetane Ricard-Nihoul –布魯塞爾歐盟委員會公民對話部門副主任。
- Colin Scicluna – 歐盟委員會內閣部長、民主與人口統計副總裁。
- Claire Thoury–法國社運領袖、生命臨終公民會議治理委員會前主席。

(二) 公民會議大會成員

- Harry Alzire, Martial Breton, Soline Castel–生命臨終公民會議前參與者（線上參與）。
- Nathalie Berriau, Myriam Souami –生命臨終公民會議前參與者。
- Léo Van Nieuwenhove – 前生命臨終公民會議參與者（線上參與）。

(三) 研究者

- Jean-Michel Fourniau –古斯塔夫·艾菲爾大學榮譽研究主任。
- Cristina Lafont –西北大學哲學教授。
- Christiane Rafidinarivo – IEPM 政治學副教授。
- Min Reuchamps – 生命臨終公民會議的保證人、比利時魯汶天主教大學政治學教授。
- Sandrine Rui –法國波爾多大學社會學副教授。
- Chloe Santoro –生命臨終公民會議研究員兼觀察員。
- Jane Suiter –都柏林城市大學政治傳播學教授。
- Miguel Von Fedak – 耶魯大學倫理政治與經濟學與民主協商學生。

(四) 產業界

- Claudia Chwalisz – DemocracyNext 創辦人兼執行長。
- Marjan Ehsassi – 生命終結公民會議的擔保人。
- Claire Meillier – Iswe 基金會知識與實踐主管。
- Antoine Vergne – Missions Publiques 國際專案經理。

## 肆、 人工智慧與治理(Governing “with“ AI)

人工智慧(artificial intelligence, AI)技術的快速發展對人類治理及對社會逐漸產生了問題與影響。人工智慧革命包括翻譯、協助和整合等工具，這些工具可幫助實現更具涵容性、協商性和真實性的民主形式，產生的影響不僅在地方和國家層面，更擴展至全球規模。因此，人工智慧可能既是民主國家的一個問題，也是其危機的解決方案。第二場人工智慧與治理會議旨在探討民主挑戰與人工智慧前景交錯的問題，像是我們如何善用人工智慧及能否使用人工智慧技術來達成民主？討論著重於人工智慧技術如何幫助增強公民會議，並幫助融合微觀和宏觀之間的差距的問題，最終目標是將民主研究和實踐者與人工智慧專家聯繫起來，參與者可以圍繞使用數位工具、人工智慧的治理與使用人工智慧促進溝通的主題，進行討論並積極建立合作關係。

人工智慧的前景和風險，是透過獲取超量知識和處理能力來協助人類解決困難任務。此次耶魯大學對於如何利用不斷進步的人工智慧技術進行監管和治理的會議，採用了線下面對面互動激發合作與創新的對話，為期兩天的活動中，第一輪分為八個子題目進行，每一輪的時間為20分鐘，由參與者自行選擇小組加入後提出挑戰性的問題，且需確保每一位成員都有發言機會，討論時間一結束參與者立即換另一小組繼續討論感興趣的議題；第二輪一樣有子題目讓參與者選擇，但拉長討論時間為55分鐘，並應用畫架上的便利貼去促進發想及統整最後的結論；第三輪則是在活動會場中間畫出一條筆直的線，讓參與者對利用人工智慧重塑民主潛力的各種聲明進行辯論，本次我方參與討論主題計有八項，參與情形如下，並於文後附上本場參與者的資料。

圖3-7 人工智慧與治理會議(Governing “with“ AI)開場參與者自我介紹

## 一、 在民主過程中使用人工智慧所涉及的隱私和安全問題 (Privacy and security issues in the use of AI in democratic processes)

主持人是瑞士聯邦理工學院的電腦科學教授Bryan，專長於隱私、去中心化系統和區塊鏈技術，也發表許多關於電腦科學應用於民主的文章，Bryan提出在使用AI之前是否有安全隱私問題或是有可能出錯的地方，並邀請參與者提出實際的案例，其中一位在公民會議的工作者，提出立陶宛對於鄰國俄羅斯的勢力滲入非常擔憂，在強化參與者隱私及資訊品質上非常的重視，另一位參與者提出贊同觀點，表示所有人都可以和語言模型進行聊天，但人們並不知道那些是被操控的認知，而可能損及未來民主運動的發展。

臺灣可能同樣面臨著鄰國勢力的挑戰，因此我方分享到臺灣目前進行了為期一年的實驗，名為「運動數據公益計畫」的試點項目，data-sports.tw內容是讓人民去健身房、當地學校、運動中心等地方，在運動完時願意將非個人數據的部分提供到大數據中心，再經過數據化的內容，去進行隱私強化技術，確保在下游，沒有對任何原始數據的追蹤，此蒐集系統不能重新識別資料提供者，這是具備非常高應用價值的數據，將有利應用於設計保險、醫療保健等政策，有著很高的分析和參考價值，也透過此項

計畫去培養一種參與性的公民運動習慣。

主持人Bryan又提出在中國權力威脅的挑戰下，有大量競爭性軟硬體皆來自中國，無法確認是否安全的軟硬體(Secure unclaimed, software, secure hardware)，包含所有的晶片和處理器等，隱私和安全在民主選舉等方面都可能造成問題，這是否讓臺灣政府感到擔憂? 我方回復，在臺灣政府的採購案若涉及到國家安全或隱私等案件，製造商或是零件使用皆不能為中國製造，對於大數據及語言模型的選擇，臺灣在112年自行訓練開源生成式AI模型「TAIDE」，其模型基礎係使用繁體中文資料集和當地文化的可信任的AI引擎。

另一位參與者提出以下假說，在去年出現元模型(metamodel)關於學習偽數據，進而操控模型產出之結果。例如，中國若想推動一個思想路線，但其路線與臺灣的資訊的真實性不符，然而卻可以透過多次複製與此思想路線一致的數據，反之去操控語言模型，這涉及了語言模型數據操控問題，影響到國家安全的認知，此危機凸顯了政府需設立準則監控人工智慧發展的重要性。

圖3-8 Google隱私、安全和安全工程副總裁Royal主持一場關於在民主進程中使用AI的隱私和安全問題討論

## 二、聊天機器人可以促進審議嗎?(Can a chatbot facilitate deliberation?)

在會議開始時，眾多參與者提出聊天機器人能夠有效整理公民提出的數百個想法，整合各種論點並反映各子主題中贊成和反對的論點，但同時產生有些令人擔憂的問題，如聊天機器人在處理訊息時可能存在偏見。多數參與者皆關心每位公民參與者在審議公民會議中的發言權是否平等，此外，多數與會者認為聊天機器人更像是即時產生的數據集，無法完全代表審議的過程，要如何確保聊天機器人整理出來的訊息具有價值，也是眾所關心的問題。

本次小組討論主持人為史丹佛大學民主審議實驗室副主任Alice Siu，同時也是Deliberative Polling[6]的代表與發言人，Alice在聽完大家的問題後，快速的介紹了Deliberative Polling線上平台創建的目的，該平台建立的目的係模仿過去三十年來，公民審議大會討論的基本功能。這些功能包括管理時間、自動呼叫相關人發言、議程管理以及對主題討論的推進，都具備靈活的設計，該平台還可以記錄每位發言者的內容，並利用線上協作文件的功能，在後續針對專家小組所提問的問題進行開放共同編輯和投票以決定提出的問題等，由於資料來自真實的對話內容，而不是從各種資料集中抓取訊息，因此不會提供不正確的資訊。

最後，主持人特別提出兩項重點：第一點為結構化聊天機器審議會議將有助於促進決策的民主化。第二點為AI可同時掌握小型、聚焦的討論優勢卻又觸及大規模公眾參與的機會，同步增加小團體會議數量。主持人也分享史丹佛大學將與臺灣政府合作的消息，合作主題為如何在大型平台提供的人工智慧服務中促進「資訊完整性」原則，包括公民、社區使用者和數位從業者應如何利用人工智慧來提升資訊的可信度，認識和分析資訊背景，推動資訊完整性。

[6] Deliberative Polling官方網站：https://deliberation.stanford.edu/what-deliberative-pollingr

## 三、將權力分析應用於人工智慧(Applying a power analysis to AI)

主持人為耶魯大學經濟學教授Henry J. Heinz II，研究主要集中在如何塑造社會中權力關係和經濟、政治、環境優勢的模式，尤其關注在發展中國家，並以資本主義對一家科技公司的技術研究方向影響開始討論。教授舉Google公司為例，人工智慧並不是他們主要業務，而該公司股價在很大程度上取決於人工智慧產品的表現，其中產生問題在於技術發展的重點，往往利潤將成為主要動機，而不是對社會或個人福祉的關注，因此資本主義制度可能導致財富和權力集中在少數人手中，而其他人則可能會被忽略，參與者可先思考問題，並考慮是否應開放使用人工智慧技術以及開放的程度。

以公司型態來說，OpenAI和Google在公司結構尚有很大的不同，非營利組織的行動應是代表人類的利益，但Google具備雙重結構，大多數小股東沒有權力，股東在意公司營運表現及產生之利潤，和公司未來的高層主管人選，多數層面都是在關心公司的文化，而不是與AI相關的問題。此外，對於臺灣政府對AI監管的問題，我方除了回應臺灣在2023年培訓自己的開源生成式AI模型外，本部在 112年12月成立「AI產品與系統評測中心(Artificial Intelligence Evaluation Center, AIEC)」，旨在建構臺灣的AI產品與系統評測方式與規範，提供AI評測服務。AI評測機制將先以大型語言模型(Large Language Model, LLM)為評測對象，評測項目研析歐美AI規範內容包含NIST、ISO、歐盟議會等，擬定10項AI評測項目。

經過以上的討論後，主持人最後結論：使用人工智慧的權力，應不強調經濟上或政治上的權力，而是強調人們如何選擇人工智慧種類，同時建立好的規範和隱私安全，保護人們隱私安全並盡可能降低人不可控制的風險。

圖3-9 OpenAI 的代表Tyna與其他參與者討論公民會議中AI的使用案例

## 四、 我們能夠使AI規範合法化嗎？(Can we make AI regulation legitimate?)

主持人為民主創新計畫的博士後研究員Mourgues，開頭提到民主的承諾是因為人們普遍具有自主性，在討論規範合法前，社會需要明確的規範。參與者表示，事情發展的速度超過了規劃的速度，合法性的狀態一直在演變，新技術甚至推動著市場，這使得使用者很難分辨出什麼是人工智慧產生的結果，如何意識到人工智慧帶來的挑戰，並及時制定相應的規範以應對這些挑戰。

在這議題下我方分享了臺灣目前的情形，臺灣正處理AI產品與系統評測方式與規範，臺灣已了解到技術正在改變世界，但我們必須創造一個環境，讓人們能夠相互合作，例如，我們應該有自己的想法，如同每個人都有能力塑造AI產出的方向。參與者提出，但相關規範究竟是誰有權力制訂？我方回覆說明，評測項目研析歐美AI規範內容包含NIST、ISO、歐盟議會等，擬定10項AI評測項目之外，為了讓科技發展可以對齊社會價值，113年數位部將與史丹佛大學線上審議平台合作，舉行一場400位公民審議大會，主題將討論如何促進大型平台提供的

AI 服務中的資料完整性(Information Integrity)[7]原則 ，之後公民審議大會的報告與結論也將會回饋到AI 產品與系統評測中心，以期實現可信賴AI發展的目標。

## 五、 AI促進溝通的優缺點(The advantages and disadvantages of AI facilitated deliberation)

主持人為國際性的研究和行動機構Democracy Next首席執行官Claudia。主持人表示AI在民主審議中過程層次遞進的引導，並具體表示利用AI來增強和改善審議過程是很理想的方式，但發展的目標需確保人類的參與，且審議的目的是為了共同理解和行動。主持人提出審議過程中有許多重要的元素不可以忽略，如: 引導(facilitation)，就是引導和促進討論的過程，確保每個人都有機會參與並表達自己的觀點，角色上引導員(facilitator)需要保持中立性，幫助參與者共同達成共識，並提出以下問題讓大家去說明出自己的觀點：

圖3-10 Democracy Next首席執行官Claudia主持一場AI促進溝通優缺點的討論

(一)好的引導具備什麼特質？

[7] 資料完整性(Information Integrity)，完整性在資安領域中是指確保資料在存儲、處理和傳輸過程中保持準確和一致的性質，確保資料在其生命週期內始終是可靠和準確的。

參與者提出審議性的引導是要讓彼此接觸新的思想，但若是處理倡議式的議題，那引導方式可以透過平台辦理，先讓人們知道如何使用但不透過任何指令。最後總結：好的引導具備以下幾個特質，像是處理衝突達成共識、平均分配時間、積極傾聽、公正性等。

(二)為什麼我們要考慮使用AI促進審議？

以下討論係基於「假設人工智慧替代了人類引導的情況下」，所進行的審議過程。參與者提出，這種人工智慧取代可辦理更大規模的審議，人工智慧可透過事先蒐集民眾意見、分析眾多想法的能力，AI chatbot也能長期有效進行與民眾討論對話功能，透過民眾使用對話聊天機器人時，得知大眾的想法，進而發揮長期教育與針對大家關注的議題溝通的功能。

(三)應該使用AI促進引導嗎？有什麼優缺點？

參與者提出了以下觀點：對於應用AI促進引導視為理所當然（像是自動化、科技輔助等）、人工智慧系統僅從現有的知識中學習、失去由人類擔任引導員的靈活性、人們必須要相信AI才會想要使用、無法透過獲取知識而生存等，同時也挑戰了AI應用在審議討論的方法論和可信度。

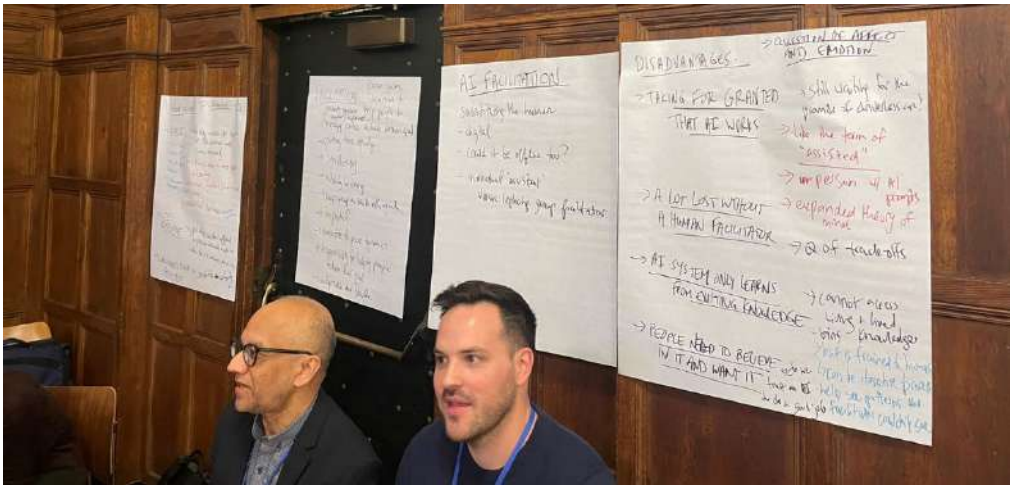


圖3-11 AI促進審議的優缺點的討論結果

## 六、互動性辯論(Interactive Debates)

互動性辯論是為下一個會議進行暖身，隨機將參與者分成4人小組，提出具爭議性的主張，並邀請所有參與者表達意見，以探索人工智慧、治理和民主之間的相互作用，活動進行時不要求達成共識，大多數參與者都散佈在房間的各個角落分享各自觀點，以下簡單記錄討論之內容。

(一) AI應該平等代表所有社會？

- 同意：人類之所以溝通就是為了尋求共識，這也顯示每個觀點都是具備平等價值，值得被聽見，AI的發展也應該具備平等性。
- 中立：認同平等的想像，但因為還尚未實現，所以表示中立的看法。
- 不同意：在技術上沒有刻意造成不平等，但依語言資料筆數上就可知道這不太會是科技發展的趨勢。

(二) AI增加人類生存力？

- 同意：AI目前廣泛的被應用在預測性事務，像是衛生醫療對基因潛在的缺陷與疾病預測提早治療，都顯示AI可提升人類的生存力。
- 中立：生存和生存品質是不一樣的，未來可能是那些少數具有權力的人活得更久。
- 不同意：AI更多的是取代及被操控，甚至可能引發的新型戰爭。

(三) 使用AI使大家都成為快樂的白癡

- 同意：完全認同，我現在就已經是了。
- 中立：我認為AI的使用問題上尚未引爆，且這將不會是最終的樣貌。
- 不同意：AI提供人類更高的學習及工作效率，若是以人機合作的未來來看，人類將進化到更高級的物種，可立即處理更多艱難的問題。

(四) AI將取代政治權力

- 同意：以樂觀者對AI科技的發展來看，這確實是逐漸在形成的。
- 中立：民主的下一步可能又回歸少數人掌握的世界裡，科技技術的門檻提高到很難讓一般人參與，但不認為機器可以撇除人類自行發展。
- 不同意：人類會在各種被推翻的形式下，找到掌握權力的方式。

(五) AI無法促進審議

- 同意：審議的目的不是那些數字的顯示或是及時文字的紀錄，是人類透過對話和實際參與，找出共識提出解決的方法。
- 中立：AI會協助審議的流程，但要實質舉例促進審議要達到的目標，若對於整理意見則一定是有幫助，但錯誤的運用都可能具備程度上的風險。
- 不同意：這完全就是無視科技的發展，史丹佛線上審議平台已經有許多成功的案例證實，AI可部份取代人類所主持的會議。

(六) 不可能有一套普遍的價值觀來管理AI

- 同意：應該說是沒有一個角色可以去訂定這個規範，如果有這個權力應該落在誰身上，都尚需討論。
- 中立：未來應該會存在著多元的AI可以選擇，像是web2的世界，但是否不可能達成普遍價值觀去監管，確實沒那麼絕對。
- 不同意：這將只是一個長時間的過程，就像是氣候變遷的議題，在全球長時間的討論下，還會有基準的規範來管理AI的發展。


圖3-12 互動性辯論(Interactive Debates)環節的會議活動範圍圖

## 七、 我們如何創造一個更具溝通性的社會？(How can we create a more deliberative society?)

主持人為史丹佛大學審議式民主實驗室主任James，會議內容為討論審議民主改變政治，以及增強公民參與的辦法。主持人提出兩種策略：第一是引用法學數據，提倡在選舉前設計審議式大會，第二是回溯歷史，重新思考雅典民主的理念，在輪流統治和被統治的情況下，人們會從生活中抽出時間擔任類似陪審團等角色進行辯論，這將改變政治文化，讓積極的公民可以產生影響力。

參與者提出，陪審團的問責機制立意良善，但對於陪審團造成的影響就抱持著很大的疑慮，因為陪審團們決定了其他人的命運，但對自己的決定卻不需負責，有很多類似這樣模式的會議，只是此時我們稱之為公民審議大會。第二個問題是，邀請公民參與的人數，涉及到討論的問題本身，這背後牽涉到政治問題，大家是否可以接受讓對議題不了解的大眾參與？主持人James回復，他不認為這是個事實，根據過去的經驗，多元參與者共同討論時，公民會更具開放性，在極端意見之間增加了參與的積極性，在選舉來時，人們會將政策立場與投票聯繫起來，並尊重不同立場的意見。

參與者提出對人工智慧促進審議的觀點，他認為AI肯定能夠讓我們準確了

解人們想要什麼，但這是相當令人恐懼的，因為AI只是在獲取數據後加以處理，總結後查看大眾或是全球的趨勢，結論報告都只是人工智慧的結果。主持人James回復，這完全不是毫無意義的結果，審議制度可在人們還沒有考慮到的問題上，總結了人們參與和思考到的利弊並預先提出解答和意見。另一位參與者提出Ai審議的兩個情形，第一個是參與了一場活動但未能參加完整活動；第二個是全程參與但沒有和其他參與者合作，僅與一個機器人進行交流，是否能從中獲得任何啟發？主持人James說明這正是他們和研發團隊正在開發的聊天機器人，會是未來的展望，後續的討論並沒有達成結論，但卻體現了這場會議最想要的思想辯證，和來自不同背景的對象討論，交流各自的觀點後都會帶回一些新的想法，這也是激發合作與創新的伏筆。

## 八、操縱的應用案例(Use cases of manipulation)

主持人為華盛頓大學法學院的博士候選人和教職員Inyoung，負責人工智慧與法律方面進行跨學科研究，重點在推進人工智慧系統的治理。在討論開始時，請參與者先寫下AI操控下良性的使用案例與具敵意的使用案例，接續提問誰將會是AI操控下最大的受害者？以及操控者主要意圖是什麼？

### (一) 良性的濫用案例？

參與者舉例指出，數位雙生模型儲存了大量使用者的使用習慣、偏好、語意、記憶庫、甚至身體屬性等數據，可能導致個人資訊外洩的危害，因為模型不僅記錄了個人的行為和知識，而且還會根據這些數據來調整系統偏好和視角，過度依賴數位雙生模型將可能會導致人們最終觀點趨於一致，造成文化同質化的出現，進而侵害傳統或少數文化，如穆斯林裹頭巾文化等。此外，這也可能對孩子的知識學習產生重大影響，如導致數位落差的加劇。


圖3-13 操縱的應用案例(Use cases of manipulation)討論白板

(二) 蘊含敵意的濫用案例？

參與者舉出了一系列問題，從詐騙、假訊息到深度偽造，以及政治運動等。舉例來說，美國前總統川普對於氣候變遷所造成的經濟損失及極端氣候問題深具懷疑，將作為一個經典案例，如果有地方聲稱氣候變遷都是虛假的，那麼在這些地區，AI語言模型可能被要求說這些氣候變遷影響的報告都是虛假的，甚至會控告OpenAI，這涉及到專業機構釋出的評估報告或專家發言等情況，這種情況背後可能是一場政治運動，但由於技術的進步和資訊的泛濫，真偽更難以辨別，訊息可以通過網站、手機應用和平台來操控大眾的意識和認知。

(三) 誰將會是操控下最大的受害者？

首當其衝的肯定是目前數位世代的成員、正在成長的孩童以及各種平台的使用者。但確切來說，沒有人可以免於被操控的風險。

(四) 操控者背後的意圖？

為了實現更高的獲利、獲得更大的控制權以及贏得選民支持，操控者的意圖都是為了追求更大的權力。


圖3-14 海倫·蘭德莫爾(Hélène Landemore)教授在活動進行中的發言

## 九、本場會議的參與者資料

為期一天半的會議旨在將協商民主研究和實踐的領導者與人工智慧專家和領導者維繫起來，期望他們能夠就使用溝通工具與治理人工智慧、以及使用人工智慧工具增強溝通的想法建立新的合作關係，以下是會議參與者的資訊：

(一) 政策制定者

- Inyoung Cheong –華盛頓大學法學院博士生兼教員，為受OpenAI 資助入選團隊。
- Michelle DiMartino – 行為洞察團隊資深顧問。

- Mahmud Farooque – CSPO 副主任；社會創新未來學院臨床教授；亞利桑那州立大學。
- Kevin Feng – 華盛頓大學設計與工程系博士生，為受OpenAI 資助入選團隊。
- Mark Gorton – Tower Research Capital董事長。
- Yichen (Lesley) Ho – 臺灣數發部系統分析師。
- Wendy Hsueh – 臺灣數發部專案分析師。
- Colin Irwin – 和平民意調查的創建者，為受OpenAI 資助入選團隊。
- Andrew Konya – Remesh AI 共同創辦人兼首席科學家、為受OpenAI 資助入選團隊。
- Aviv Ovadya – 人工智慧與民主基金會創辦人，為受OpenAI 資助入選團隊。
- Andrew Sorota – 施密特期貨研究助理。
- Claire Thoury –法國社運領袖、生命臨終公民公約治理委員會前主席。

(二) 研究者

- Isabelle Ferreras –哈佛法學院勞工與公正經濟中心資深研究員。
- James Fishkin – 史丹佛協商民主實驗室主任。
- Bryan Ford – 洛桑聯邦理工學院電腦科學教授。
- Alan Gerber – 耶魯大學斯特林政治學教授。
- Caroline Green－牛津大學人工智慧倫理研究所博士後研究員。
- Jacob Hacker －耶魯大學政治學史丹利·雷索爾教授。
- Oliver Hart – 哈佛大學 Lewis P. 和 Linda L. Geyser 大學經濟學教授。
- Matthew Meyers – 耶魯大學政治學與數據科學學生。

- Rohini Pande –經濟學教授兼經濟成長中心主任。
- Lex Paulson –集體智慧學院執行董事。
- Ariel Procaccia – 哈佛大學電腦科學教授。
- Shir Raviv – 哥倫比亞大學資料科學研究所博士後研究員。
- Manon Revel –哈佛大學網路與社會中心研究員。
- Alice Siu – 史丹佛協商民主實驗室副主任。
- John Tasioulas – 人工智慧倫理研究所所長、牛津大學倫理與法哲學教授。
- Philippa Webb – 國際治理與爭端解決中心 (CIGAD) 聯合主任。
- Luigi Zingales –芝加哥大學金融學教授。

(三) 產業界

- Matt Botvinick – DeepMind 神經科學研究總監。
- Tyna Eloundou – OpenAI 技術人員。
- Royal Hansen – Google 隱私、安全與保障工程副總裁。
- Teddy Lee – OpenAI 產品經理。
- Liane Lovitt – Anthropic 公共政策經理。
- Kris Rose – Meta 治理洞察主管。

## 伍、 心得

### 一、政府施政制度的下一步

近年全球風行的ChatGPT（含GPT類LLM）及各種生成式AI，已逐漸改變了各產業的面貌。其智慧化且廣泛應用在行銷和產品開發上，為各行各業帶來了新的可能性，企業組織經營都已將導入人工智慧視為產業轉型的關鍵，未來甚至可能對代議民主的治理方式與政策制定帶來衝擊。

在會議的中場休息時間，耶魯大學社會與政策研究所教授海倫·蘭德莫爾(Hélène Landemore)和與會者交換觀點，她指出在這波人工智慧的科技下，她對於民主的下一步其實不抱持樂觀，因此總在思考是否民主形式已不再符合這世紀，我們終將又會回到菁英政治的情況，當下各參與者皆感到意外，海倫教授也對民主的未來充滿不確定感，令人又重新思考民主的下一步究竟為何。為期兩天半的活動中，在不同主題與多元專業者討論中，內心不時也會對民主的存在感到猶豫，海倫教授在討論的過程中轉向詢問我方，身為臺灣政府單位的角度，臺灣對於民主的未來又有什麼看法？當下想了一下表示，或許沒有人知道未來該如何，但我認為民主是一個持續的過程而非一個結果，民主賦予人們有選擇權與表達自己的空間，像是我們在第一天的公民會議上探討了很多挑戰與困難，每一次聚集都是在塑造民主的樣貌與發展，都具有獨特的意義。

### 二、數位工具促進審議

公民會議在是全球進行公共討論是最主要的模式之一，資訊科技迅速發展，許多國家正積極推動數位轉型，以推動電子民意調查和電子投票等數位工具。這些數位工具通常具有不受空間限制的參與性、多媒體性、可累積性、易搜尋性、匿名性等特性，因此透過數位的特性，將有助於克服線下商議式民主所面臨的困境，不僅提高公民參與度，並可加強參與者之間的公平性，唯獨在理論與實務方面仍有不少爭議點，未來公民會議仍面臨著許多挑戰，如參與管道往往受

到權力菁英的控制，對話的公平性受到物質條件的影響等，在執行方面的時間、規模和成本也是目前需要解決的問題，未來也值得探究數位工具是否可有效協助審議，並讓大家都有同等參與政策討論的機會。

## 陸、建議

在政策研究和人工智慧發展的天秤兩端，是否存在尋找支點的機會，是這次會議主辦方最想激盪的火花，本次活動觀察到會議中有一部分是對於人工智慧的科技發展感到樂觀，但是對後續風險感到悲觀，但也部分人抱持樂觀者並相信AI會使人類生活變得更美好，善用AI將讓政策制定能更貼近社會的需求。因此不論是天秤的哪一端，都建議要讓人工智慧的設計更加民主化，使政策的形成盡可能涵蓋各方意見並取得共識，獲得民眾的最佳利益。

本次在會議的討論中學習到，AI的應用可以很多元，利用人工智慧來改善辯論、翻譯、促進和偏好誘導系統，都可以改善機關與人民之間的互動，政府能夠若能善用AI工具整合人民的需求，將更有效地制定政策和提供服務。

建議未來在發展人工智慧系統時，需讓人民擁有評估及選擇使用的權利，以減少人工智慧系統對自我權利的侵害，同時需提高對個人訊息和隱私的控制，此舉有助於增加人民對人工智慧系統的信任與接受度。人工智慧三大關鍵發展趨勢（AI數據、AI算法、AI算力）也應需要受到非壟斷性發展，以減輕權力集中化的風險，並優化集體利益為發展目標，確保所有使用者都可以享受AI提供的便利。

## 柒、 附錄

一、會議議程。

二、回國後分享簡報。

三、耶魯大學社會與政策研究所四月出刊通訊。

全文完

# 附錄一 會議議程

Home / Governing Citizens Assemblies /

## "Governing Citizens' Assemblies" Schedule

### WEDNESDAY, FEBRUARY 28, 2024

**Location: Sterling Memorial Library Lecture Hall,** 120 High Street, New Haven, CT

**Pre-Conference Reception and Dinner for Program Participants Only:** To be held on Tuesday evening, February 27; see details below.

| | |
|---|---|
| **TUESDAY, FEBRUARY 27, 2024**<br>6:30-9:00 PM | Pre-Conference Reception and Dinner (only for participants on the conference schedule below)<br>The Study at Yale Hotel Penthouse, 1157 Chapel Street, New Haven, CT |
| **WEDNESDAY, FEBRUARY 28, 2024**<br>8:30-9:00 AM | Coffee and light continental breakfast available at the conference venue<br>Sterling Memorial Library Lecture Hall, 120 High Street, New Haven, CT |
| 9:00-9:05 AM | Introduction<br>• Hélène Landemore, Professor of Political Science, Yale University; Distinguished Researcher, Oxford University Institute for Ethics in AI |
| 9:05-10:45 AM | Who Governed the French Conventions?<br>• Sandrine Rui : "What Does Governing a Citizens' Assembly Mean: Power and Contingency at the Citizens' Convention on End of Life"<br>• Jean-Michel Fourniau on the Citizens Convention for Climate (CCC): "Political Mandate and Governance of a Citizens' Assembly"<br>• Chloé Santoro: "Voting in a deliberative process: what, when, how to vote and who decides?"<br>• Hélène Landemore and Théophile Pénigaud: "The case for self-ruling citizens' assemblies: Governance, Representation, and Citizen Leadership in the French Conventions."<br>• Reaction by Claire Thoury and Nathalie Berriau<br>Moderated by: Antonin Lacelle-Webster, Postdoctoral Associate with the Democratic Innovations Program at ISPS, Yale University |
| 10:45-11:00 AM | Coffee Break |

| | |
|---|---|
| 11:00-12:40 PM | Citizen Voice and Minorities Representation, and Leadership in the French Citizens' Conventions<br><br>• Christiane Rafidinarivo, "Representation of Minorities Within Citizens' Assemblies: the Experience of the Citizens' Convention for Climate"<br>• Miguel Von Fedak "Stories of Discontent: Moments of Tension and Distrust in the French Convention on the End of Life"<br>• Marjan Ehsassi, "Blueprint for Activated Citizenship: Designing for Legitimacy & Transformative Change<br>• Reaction from online participants: Mathieu Sanchez (member of the Citizens Convention for Climate), Harry Alzire, Martial Breton, Soline Castel, and Léo Van Nieuwenhove (members of the Citizens' Convention on End-of-Life Issues)<br><br>Moderated by: Jane Suiter, Professor of Political Communication at Dublin City University |
| 12:40-2:00 PM | Lunch Break |
| 2:00-3:45 PM | International Perspective on Citizens' Assemblies and other Democratic Innovations Governance<br><br>• Jane Suiter, "The Ongoing Internal Evolution of Citizens' Assemblies in Ireland"<br>• Antoine Vergne, "Governing Citizens' Assemblies: A Reflective Approach Leading to Two Blueprints"<br>• Min Reuchamps , "Governing Permanent Citizens' Assemblies in Belgium"<br>• Claire Mellier, "The 2021 Global Citizens' Assembly on the climate and ecological crisis: A power-sensitive exploration of governance"<br>• Reaction by Claudia Chwalisz and Gaëtane Ricard-Nihoul<br><br>Moderated by: Colin Scicluna, Head of Cabinet to the Vice President for Democracy & Demography, European Commission |
| 3:45-4:00 PM | Coffee Break |
| 4:00-5:30 PM | Roundtable: Should Citizens' Assemblies be more like sovereign parliaments or not?<br><br>• Thierry Beaudet, President of the French Economic, Social and Environmental Council<br>• Cristina Lafont, Professor of Philosophy at Northwestern University<br>• Hélène Landemore, Professor of Political Science, Yale University; Distinguished Researcher, Oxford University Institute for Ethics in AI<br>• Myriam Souami, Member of the Citizens' Convention on End-of-Life Issues<br><br>Moderated by: Alexandra Cirone, Assistant Professor of Government at Cornell University and Visiting fellow at the Yale Institution for Social and Policy Studies |
| 6:30 – 9:00 PM | Conference Dinner and Welcome Reception (only for specially invited guests from both conferences) |

# "Governing (with) AI" Schedule

The following is the working schedule for Governing (with) AI, taking place 29 February–1 March at Yale. The event will be structured as two days of collaborative working sessions, along with optional evening activities on Tuesday night.

The agenda is being designed as a combination of planned sessions and emergent participant-driven discussions, and specific topics will be placed into time slots based on input at the meeting from those in attendance. Sessions will be dialog- and outcome-oriented rather than presentations or lecture format.

For more information about the workshop format, please visit our page on the agenda overview and guidelines

## THURSDAY, FEBRUARY 29, 2024
## Overlapping with "Governing Citizens Assemblies" Conference until 12:40 PM

**Location:** Sterling Memorial Library Lecture Hall, 120 High Street, New Haven, CT

**Pre-Conference Reception and Dinner:** To be held on Wednesday evening, February 28; see details below.

| | |
|---|---|
| **WEDNESDAY, FEBRUARY 28**<br>6:30–9:00 PM | Conference Reception and Dinner (for invited guests from both conferences)<br>Taste of China, 954 Chapel Street, New Haven, CT |
| **THURSDAY, FEBRUARY 29**<br>8:30–9:00 AM | Coffee and light continental breakfast available at the conference venue<br>Sterling Memorial Library Lecture Hall, 120 High Street, New Haven, CT |
| 9:00 – 9:30 AM | **Opening Session**<br>The event will be called to order with a friendly and fast-paced kickoff that includes words of welcome from the hosts, brief participant introductions, along with overviews of the agenda, participation guidelines and meeting logistics. |
| 9:30 – 10:45 AM | **Surveying the Potential for Augmenting Democratic Governance Using AI**<br>The program will begin with a series of interactive learning dialogues that explore selected facets at the intersection of AI and democratic governance. Participants will be invited to rotate between topics across the course of the session.<br>• The role of AI in deliberative democracy<br>• Can a chatbot facilitate deliberation?<br>• Using AI to synthesize a collective will<br>• The tools of governance (as opposed to government)<br>• Privacy and security issues in the use of AI in democratic processes<br>• Adversarial uses of AI in democracy<br>• Applying a power analysis to AI |
| 10:45 – 11:00 AM | **Break** |

| | |
|---|---|
| 11:00 AM – 12:30 PM | **Defining and Debugging AI Use Cases in Democratic Governance**<br><br>The second half of the morning will focus on deeper discovery and problem articulation, designed to inform subsequent strategy and design sessions.<br><br>• Mapping use cases for AI in citizens assemblies<br>• Inventorying the bugs of citizen assemblies that AI might solve<br>• The advantages and disadvantages of AI facilitated deliberation<br>• Open problems in using AI tools to connect citizens' assemblies to the macro-public<br>• Democratic affordances: What does AI today make possible for democracy that previous technological and other revolutions did not?<br>• Employing AI to enhance citizen participation in international dispute resolution on collective rights<br>• Barriers and opportunities for the use of LLMs and chatbots in democratic deliberation outside the English-speaking world<br>• Towards a threat model of AI in democratic contexts |
| 12:30 – 1:30 PM | **Lunch**<br><br>Participants are encouraged to dine with those who they have not yet met or engaged.<br><br>This is the end of the overlap between the "Governing Citizens' Assemblies" conference and the "Governing (with) AI" conference. Invited guests from both conferences are welcome to stay for lunch. Invited guests from "Governing Citizens' Assemblies" are free to depart Yale. |
| 1:30 – 3:00 PM | **Interactive Debates**<br><br>As a bridge to strategy and design sessions, participants will be invited to propose compelling provocations and debatable assertions that probe and test the interplay between AI, governance and democracy. |
| 3:00 – 3:15 PM | **Break** |
| 3:15 – 4:30 PM | **Strategic Working Sessions**<br><br>These working sessions will invite participants to further explore both AI in governance as well as governance of AI. Topics have been identified during pre-event engagement and planning.<br><br>Session facilitators will briefly introduce the objective of each session, and participants may then elect to join the session of their choice. Report-backs will be done at the end of the session slot.<br><br>Sessions currently anticipated to be in this time slot are:<br><br>• Enumerating and comparing design patterns for global deliberation<br>• How is AI currently regulated at the global level and what are the insufficiencies of that approach?<br>• What has worked before to positively shape technological development?<br>• The Power Question: Who controls AI, who owns the data, who is connected to AI, who benefits and who suffers?<br>• Will the "average voter" be the primary point of failure in AI governance?<br>• What does an environmentally sustainable, citizen-centric, AI-augmented democracy look like?<br>• Pathologies of tech companies' governance |
| 4:30 – 5:00 PM | **Closing Session**<br><br>The closing session will invite participants to weigh in on what has been most useful during the course of Day 1, and refine their goals and priorities for the agenda of Day 2. |
| 6:30-9:00 PM | **Conference Dinner for invited guests for the "Governing (with) AI" Conference** |

## Friday, March 1, 2024

| | |
|---|---|
| 8:30-9:00 AM | Coffee and light continental breakfast available at the conference venue<br>Sterling Memorial Library Lecture Hall, 120 High Street, New Haven, CT |
| 9:00 – 9:15 AM | **Opening Session**<br><br>The day will start with a summary of Day 1 outcomes and a Day 2 Agenda Overview. |

| | |
|---|---|
| 9:15 – 10:45 AM | **Envisioning and Designing the Work We Can Do Together**<br><br>These working sessions will be similar in structure to those from Thursday afternoon.<br><br>Some sessions will continue and build on work started during Day 1. Other sessions will introduce new topics and objectives. Many topics will be sourced from participants during the course of Thursday, but other potential topics have been identified from pre-event input:<br><br>• What can make AI regulation democratically legitimate?<br>• What kind of AI for what kind of democracy?<br>• How would a global citizens' assembly on AI governance fit into the global web of international laws and institutions?<br>• What are the risks and merits of open versus closed AI?<br>• How do we include tech employees' voice in high level decisions about AI development?<br>• Data representativeness and biases: what do we know and still not know?<br>• Where are we on inclusion?<br>• Envisioning a curriculum for citizen education on the opportunities and risks of AI<br>• Collaborative policy prototyping: How can we iteratively prototype policies so we are more certain of their efficacy?<br>• Considering a constitutional approach for global AI regulation<br>• What reforms to the internal governance of tech companies should take place to ensure that AI remains safe for humanity?<br>• Actionable guidelines for balancing AI-driven innovation with democratic governance<br>• Weighing risks and opportunities of seeking to regulate AI models on the basis of their potential risk<br>• The future of work: artificial intelligence, corporate governance and workers' rights<br>• Real time assessment of the social implications and governance of AI as the technology side of AI advances.<br>• Drafting concrete best governance proposals<br>• Graphing the pros and cons of EU versus US regulatory frameworks<br>• How can AI enhance small-group and large-group deliberation<br>• How might we evaluate AI safety in complex society-in-the-loop systems?<br>• What's different about AI compared to other big issues, like climate, where we've tried to have mass democratic dialogue?<br>• What has worked before to positively shape technological development? |
| 10:45 – 11:00 AM | **Break** |
| 11:00 AM – 12:15 PM | **Mapping Where to go From Here**<br><br>The group will pause before the final session to take stock of the progress made to this point in the week and to inventory action items, next steps and other bridges to post-event collaboration. |
| 12:15 – 1:00 PM | **Closing Session**<br><br>Participants will weigh in on what has been most useful during the course of Day 2, share appreciations and bring the meeting to a close. |
| 1:00 – 2:30 PM | **Closing Lunch** |

The conference "Governing Citizens' Assemblies" is funded by the Institution for Social & Policy Studies' **Democratic Innovations** program and The Whitney and Betty MacMillan Center for International and Area Studies at Yale University with support from the Edward J. and Dorothy Clarke Kempf Fund.

The "Governing (with) AI" conference is generously funded by the AI2050 program at Schmidt Futures (G-22-63447).

# 附錄二 回國後分享簡報


## 大綱

壹、會議目的

貳、會議過程

參、與會效益

肆、提議


主目的

從**理論、實踐和規範**的角度解決具體的治理問題

手段1

次目的

**匯集跨學科**的群眾，並將學者、行業人士和政策專家

手段2

其目的

**推進**相關主題的研究前沿和學術思維，又是**生成新的想法**、概念框架和政策或技術工具，以幫助**解決問題**

手段3

會議議程

海倫·蘭德莫爾

Hélène Landemore

為耶魯大學社會與政策研究所（ISPS）教授，致力於民主創新計劃中領導著有關公民審議的研究議程。

## 會議過程

- Can a chatbot facilitate deliberation?
- Privacy and security issues in the use of AI in democratic processes
- The role of AI in deliberative democracy
- The tools of governance

## 會議過程

- 每個國家演算法需平等
- AI增加人類生存
- AI使大家都成為快樂的白癡
- AI將會取代政治
- AI無法增進審議


與會關係人

## 關係人一　總統盃黑客松邀請隊伍

**Inyoung Cheong**

icheon@uw.edu

**Kevin Feng**

kjfeng@uw.edu

**Andrew Konya**

andrew@remesh.org

**Aviv Ovadya**

aviv.ovadya@gmail.com

**Colin Irwin**

colin.irwin@liverpool.ac.uk

## 關係人二　促進溝通的數位工具團隊

**James Fishkin**

jfishkin@stanford.edu

- ✓ 2024與產業署合作

**Alice Siu**

asiu@stanford.edu

- ✓ 2024與產業署合作

**Lex Paulson**

lexpaulson@gmail.com

- ✓ 預計3/20線上會議
- ✓ 七月中來台
- ✓ 鏈結烏克蘭數位部團隊

**Claudia Chwalisz**

claudia@demnext.org

- ✓ 預計3/20線上會議
- ✓ 鏈結MIT開發團隊

## 關係人三　　產業界研究AI促進人類行為團隊

**Royal Hansen**

rih@google.com

Security Engineering at Google

**Michelle DiMartino**

michelle.dimartino@bi.team

Senior Advisor at Vice on the Digital at the Behavioral Insights Team (BIT)

**Teddy Lee**

teddy@openai.com

Product Manager at OpenAI on the Collective Alignment team

**Tyna Eloundou**

tyna@openai.com

Technical Staff at OpenAI

提議 .

## 將批判性思考帶入起案階段

Yale University

Institution for Social and Policy Studies
ADVANCING RESEARCH • SHAPING POLICY • DEVELOPING LEADERS



# AI and Democracy: Yale Conference Sparks Collaboration and Innovation

**AUTHORED BY** Rick Harrison

March 28, 2024


If the promise and risks of artificial intelligence (AI) involve replacing mundane and difficult human tasks through superhuman access to knowledge and processing power, a Yale conference on how to regulate and govern with this evolving technology employed low-tech, human interactions.

At times over the two-day event, participants chatted in tight circles of chairs, posted sticky notes to easels, and stood in clusters along a line representing their relative support for various statements concerning the potential for AI to reshape our world.

For example, agree or disagree: "Democratic AI is incompatible with capitalism." "AI should represent all world societies equally." "It's impossible to have a universal set of values governing AI systems." "AI will replace politicians."

No statement received unanimous support, and most instigated participants to spread across the length of the room, with those at the extreme ends and the middle sharing their conflicting views. And the enthusiastic debates, conversations, and collaborations continued nonstop over breaks and meals.

"This has been an incredibly inspiring and useful conference," said Teddy Lee, product manager at ChatGPT creator OpenAI (https://openai.com/) for a team developing processes and platforms enabling democratic inputs to steer AI. "It's

not often enough that technologists and people who are working on improving democracy are in the same room in such a constructive environment."

Hosted by the Institution for Social and Policy Studies (ISPS), the "Governing (with) AI" conference (https://campuspress.yale.edu/governingx/governing-with-ai/) overlapped with a conference on governing citizens' assemblies (https://isps.yale.edu/news/blog/2024/03/governing-citizens%E2%80%99-assemblies-lessons-from-france-and-beyond), part of ISPS faculty fellow Hélène Landemore's new "Governing X" (https://campuspress.yale.edu/governingx/) series. Landemore (https://isps.yale.edu/team/h%C3%A9l%C3%A8ne-landemore), a professor of political science, helps lead ISPS's Democratic Innovations (https://isps.yale.edu/programs/democratic-innovations) program, designed to identify and test new ideas for improving the quality of democratic representation and governance. The "Governing (with) AI" conference was funded by the AI-2050 program at Schmidt Sciences (https://ai2050.schmidtsciences.org/).

"Hélène's work and this conference demonstrate the value of fostering interaction between concerned and knowledgeable individuals from a variety of backgrounds," said Alan Gerber (https://isps.yale.edu/team/alan-gerber), ISPS director and Sterling Professor of Political Science. "Every day it becomes clearer that artificial intelligence will shape our future — including how we govern ourselves. ISPS, and certainly Hélène, are committed to helping promote understanding of AI and its applications so it can be safely and thoughtfully integrated into our society and institutions."

Allen Gunn, a facilitator with the California-based nonprofit organization Aspiration (https://aspirationtech.org/), led the conference's activities. With a friendly, upbeat attitude, Gunn encouraged respect, listening, self-awareness to allot everyone an opportunity to contribute to conversations, and the use of accessible language for people with different backgrounds and fields of expertise. To promote collegiality, name tags displayed only first names and no titles or affiliations.

"We aren't going to solve anything today, but we can make designs — we can make plans for plans," Gunn said to introduce the second day's morning session. "We can get clarity on the work we can do together moving forward."

Landemore encouraged participants to reconvene in the next six to eight months and perhaps again for a follow-up conference in the spring of 2025. In reflecting on the week's interactions, she expressed great satisfaction with how everyone embraced the format to best share ideas and forge collaborations.

"The hope was that if we empower people, they will come up with their own ideas and find each other," Landemore said. "If I had tried to micromanage that into life, I don't think we would have gotten close."

In addition to representatives from OpenAI, conference participants included leaders from Google (https://about.google/), AI company Anthropic (https://www.anthropic.com/), and Meta (https://about.meta.com/), the parent company of Facebook. Academic attendees included experts in political science, sociology, economics, communications, engineering, computer science, data science, law, and ethics. Topics included the potential for using AI to synthesize a collective will, adversarial uses of AI in democracy, mapping use cases for AI in citizens' assemblies, insufficiencies of global AI regulation, and questions around who owns and controls AI data.

Attendees expressed appreciation for the conference's intellectual humility and constructive curiosity, its openness and shared sincerity, and how the group's diversity fostered creativity and provoked challenges to siloed thinking.

"I initially wasn't sure about the interactive methodology proposed for the conference," said John Tasioulas (https://www.oxford-aiethics.ox.ac.uk/john-tasioulas), a moral and legal philosopher who, as the inaugural director of the Institute for Ethics in AI at Oxford University, partners with Landemore on a three-year project exploring AI ethics funded through the Schmidt Sciences award. "But it was totally in keeping with the underlying philosophy of the project and just remarkable to see this really productive discussion. How do you get so many people to engage constructively? I thought it was a master class in achieving that."

Tasioulas said he was particularly pleased with the receptiveness of representatives from OpenAI and Anthropic.

"They were willing to entertain ideas about how to make their decision-making process more responsive to democratic values and human rights," he said. "Willing to interrogate some of the assumptions that perhaps they uncritically make. So, for me that was a real eye opener. It was great."

Mahmud Farooque (https://search.asu.edu/profile/1502622), associate director of the Consortium for Science, Policy and Outcomes and a clinical professor at Arizona State University's School for the Future of Innovation in Society, also said he enjoyed the conference's interactive structure, which he felt facilitated deeper conversations more quickly than a typical series of academic presentations and panels.

"Gatherings like this can get at some of these hard answers about these questions we're facing in terms of democracy, security, and humanity's survival," Farooque said. "You know, you can get as hyperbolic as you want when you're talking about these topics."

Royal Hansen (https://www.linkedin.com/in/royal-hansen-989858/), a Yale College graduate from the Class of '97 and vice president of privacy, safety, and security engineering for Google, led a conference discussion on privacy and security issues on the use of AI in democratic processes.

"I like to look at wherever there is an opportunity for dual use," Hansen said of any new technological tool. "I assume that for every good thing that will happen, there is an equal and opposite bad thing."

Hansen praised the conference for strengthening the connection between academic research and tech companies to engage in nuanced solutions across broad perspectives, bolstering his optimism for overcoming the challenges represented by AI.

"I think the key that I like about this group is that we are not viewing this as an inhuman problem," Hansen said. "But in fact, it's another form of human question that humans need to work out. As long as we approach it this way — and not as something happening to us from the outside — we should be hopeful."

Michelle DiMartino (https://www.bi.team/people/michelle-dimartino/), a senior advisor at the Behavioral Insights Team, said she hopes she can apply ideas from the conference to her company's governance work on community forums with Meta.

"I'd like to figure out ways to use AI more effectively at different parts of the process we've been piloting and make it more scalable and adaptable," she said. "I'm hoping to come back to my firm with the current state of play with AI. What are the different social purpose innovations of AI that I heard about and some of the risks that we should be considering in making communications more effective and actionable?"

Kevin Feng (https://kjfeng.me/#/), a third-year Ph.D. student in the University of Washington's Human Centered Design and Engineering Department, found it refreshing to engage with people from outside of his technical background, particularly those involved ingovernment.

"We don't talk to these folks on a regular basis," he said. "It's really insightful to be able to communicate across these boundaries and see what goals we share and how our expertise and their expertise can be combined to solve these collective goals."

Feng also found the conversations productive as an exercise in communication.

"I think it really challenged me to explain my work in a way that's more accessible, which I think is going to be really important as the technologies that I'll be building and deploying may be used in democratic and deliberative settings," he said.

Oliver Hart (https://scholar.harvard.edu/hart/home), the Lewis P. and Linda L. Geyser University Professor at Harvard University, shared the 2016 Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel. He attended the conference more for its overlap with the preceding conference on citizens' assemblies, wondering how such deliberative structures could apply to furthering shareholder democracy in corporations.

But he found some unexpected insight discussing AI.

"I normally talk to economists and lawyers," Hart said. "I found it valuable to get other perspectives on matters that interest me. It turns out that AI could be useful in providing all sorts of information to people and also maybe aggregating and summarizing."

Teddy Lee, the OpenAI product manager, aims to serve as a conduit within his company for questions relating to democratic practices and theories, tapping the contacts he made at Yale.

"I think probably the most valuable takeaway has been the connections I've made here," he said. "As we continue to explore democratic governance, it's nice to know we have this network of democracy experts and fellow practitioners to collaborate with and hear from."

Mark Gorton (https://www.tower-research.com/who-we-are), founder of Tower Research Capital and a leading advocate for safer streets initiatives, supports the Democratic Innovation program (https://forhumanity.yale.edu/news/engaging-democracy) and took an active role in the conference.

"I'm leaving here legitimately more optimistic about the future of the world and democracy," he said. "This is a difficult problem. This technology is advancing so rapidly. But it is comforting and exciting to know there are so many great people who are earnestly trying to make the world better."

Landemore has great ambitions for the people she invited to New Haven. She wants to pave the way for global institutions that are more democratic and legitimate.

"Somebody's got to do it," she said. "And it might as well start with the brilliant and sincere people who came this week. I think they are going to go back home and start planting seeds that will grow over time."

And as for the future of self-governance in an era or superhuman computers? Landemore is betting on human ingenuity and values to prevail.

"I believe in the power of ideas over the power of interests," she said.

**Read about the conference on governing citizens' assemblies** (https://isps.yale.edu/news/blog/2024/03/governing-citizens%E2%80%99-assemblies-lessons-from-france-and-beyond)

**AREA OF STUDY**

Science & Technology (/research/areas-of-study/science-technology)

Follow

## Follow Us On:

**Twitter** (https://twitter.com/ISPSYale)

**Facebook** (https://www.facebook.com/YaleISPS)

**Instagram** (https://www.instagram.com/ispsyale/)

**LinkedIn** (https://www.linkedin.com/uas/login?session_redirect=%2Fcompany%2F18693690)

## Get Updates About ISPS

TO RECEIVE THE ISPS NEWSLETTER OR OTHER ANNOUNCEMENTS ABOUT ISPS EVENTS

SUBSCRIBE HERE (HTTPS://SUBSCRIBE.YALE.EDU/BROWSE?SEARCH=ISPS)