Matthew Botvinick

藝術史、認知神經科學、DeepMind 與 AI 法治研究

Matthew Botvinick 是研究心智與人工智能的科學家，也是受過精神科與法學訓練的研究者。他與 Jonathan Cohen 合作的橡膠手實驗，是身體知覺研究的經典實驗；他的其他工作涉及認知控制、階層式行動與強化學習。任教於 Penn 與 Princeton 後，他於 2016–2026 年任職 DeepMind，再轉入 Anthropic，領導 AI 與法治研究。本人 CV · Anthropic 公告

在這條科學職涯之前，他於哥倫比亞取得藝術史碩士，發表過關於 Campin 與早期尼德蘭繪畫的論文。數十年後，他又回到耶魯讀法律，於 2026 年完成 J.D.，並成為該校法學院 Senior Fellow。他的工作地點與合作環境從美國醫學院、匹茲堡的臨床和計算研究，到普林斯頓、倫敦的產業實驗室，以及法律診所，幾度改變。本人網站 · 期刊紀錄

這份人物研究沿著他實際做過的作品與工作，追索這些轉換。公開資料對他的研究、學歷與制度經歷相當充分，對早年家庭生活、藝術史師承與日常作息則較少；後者不宜靠一個過度整齊的「跨界天才」故事補齊。

斯坦福、哥倫比亞與 Campin 研究，1989–1992

Botvinick 於 1989 年完成斯坦福本科，1990 年取得哥倫比亞大學藝術史碩士。更具體的證據是他在 Art History 發表的〈The Painting as Pilgrimage: Traces of a Subtext in the Work of Campin and his Contemporaries〉，刊於 1992 年第 15 卷第 1 期，頁 1–18。這篇文章研究 Campin 及其同時代人的繪畫與朝聖意涵；期刊紀錄和他的個人履歷相互吻合。本人 CV · 期刊紀錄

藝術史是他接受過正式訓練、留下期刊成果的專業。目前尚未查得他的碩士論文題目、導師，或這篇文章與碩論的關係；本次研究亦未取得全文。現有資料不足以追索這段訓練對後來身體知覺或神經網路研究的具體影響。

Cornell 醫學訓練與 CMU 的連結主義研究

他在 1990–1994 年就讀 Cornell 醫學院。後來回憶，精神科輪轉期間，一位指導醫師把兩卷 Parallel Distributed Processing 交給他。那套連結主義研究展示了一條具體路徑：注意、記憶與知覺不必只靠描述，也能被放進會運作、會出錯、能拿來和人的表現比較的模型。

他於是走向 Carnegie Mellon 的心理學與認知／計算神經科學博士訓練，博士研究期間與精神科住院醫訓練有所重疊。他在訪談裡特別提到 David Plaut、Marlene Behrmann 等研究者的日常交流，以及問題可以跨越院系邊界的環境。履歷列出他於 1996–1999 年在匹茲堡 University Health Center 擔任精神科住院醫，1999–2002 年則在匹茲堡大學精神科擔任 NIMH Clinical Research Fellow；這些任職與 CMU 的博士研究共同構成了他的訓練環境。CMU 2019 年訪談 · 本人 CV

身體知覺、認知控制與階層式學習

Botvinick 與 Cohen 的橡膠手實驗，1998

1998 年，他與 Jonathan Cohen 在 Nature 發表著名的橡膠手實驗。受試者真正的手被遮住，面前放著一隻橡膠手；當兩隻手被同步觸碰，部分受試者會感到觸覺來自眼前的假手，甚至將它感受為自己身體的一部分。

這個簡單的裝置把身體歸屬感轉成了可以操弄的實驗問題：視覺、觸覺與身體位置感之間的對應，如何參與「這是我的身體」的形成？研究者可以改變感官訊號的配合方式，觀察錯覺的條件與強弱。原始論文紀錄 · 多感官身體知覺綜述

衝突監測與認知控制，2001

2001 年，Botvinick 與 Todd Braver、Deanna Barch、Cameron Carter、Jonathan Cohen 以兩項計算模型研究系統化了衝突監測假說。他們把控制問題拆開：不同反應傾向發生競爭時，系統能否偵測這種衝突，再據此調整後續的控制？

這個想法把腦區活動、行為現象和計算模型接在一起。它是可檢驗的理論，不是「前扣帶皮質就是衝突偵測器」的永久定論。對他的職業路徑而言，重要的是這種工作方式：把含糊的心理能力分解成機制，再要求機制產生可以與資料比較的行為。2001 年論文摘要

階層式行動與後設強化學習

日常活動通常有層次：一個大目標包含次目標，次目標又包含很多動作。Botvinick 與合作者研究階層式強化學習，以及它如何幫助理解人的行動控制。他們關心經驗能否形成可重用的結構，讓計畫不必每一步都從零開始。2014 年 Botvinick 與 Ari Weinstein 論文

2018 年，Jane Wang 領銜、Botvinick 共同參與的研究，則用後設強化學習提出前額葉功能的模型：經由較慢的訓練，系統可以學出一套在新任務裡快速調整的方式。這項工作把 AI 研究反過來用於神經科學，提供能與行為及神經現象比較的模型；兩者在若干表現上的相似，仍不等於生物腦與人工網路具有完全相同的機制。論文 · 研究團隊說明

三條研究線共享一種工作方法：先把身體歸屬感、注意控制或學習這類能力，改寫成可操弄的實驗和任務，再觀察一個具體模型能解釋什麼、在哪裡失敗。

Penn、Princeton 與倫敦 DeepMind，2002–2026

Botvinick 先在 Penn 任教，2007 年到 Princeton，2013 年升任教授。2015 年，他利用學術休假到倫敦 DeepMind，想了解機器學習正在發生什麼；2016 年正式加入。Princeton 晉升公告 · 本人 CV · CMU 訪談

轉入 DeepMind 後，他繼續探索心智研究與人工系統之間的關係。2017 年，他與 Demis Hassabis、Dharshan Kumaran、Christopher Summerfield 合寫的〈Neuroscience-Inspired Artificial Intelligence〉，把神經科學與 AI 描述為應當保持雙向交流的領域：腦的研究提供問題與

線索，人工系統則提供測試與理解的工具。論文 · DeepMind 說明

在大學，他的職務包含授課、學生研究指導與合作計畫；到 DeepMind 後，他在 2016–2023 年領導 Natural Intelligence Team，據本人履歷，團隊至 2022 年約有 50 人。這個規模使「研究者」同時也是組織工作：不同專長的人共同建立任務、模型、實驗與論文。公開來源沒有細列他每天如何分配管理與親自研究的時間，不能由主管職稱反推日常。本人 CV

分配遊戲與 AI 協助審議，2022–2024

在法律學位之前，他已經參與把 AI 接上公共問題的實驗。2022 年，Raphael Koster、Jan Balaguer 等人與 Botvinick、Christopher Summerfield 合作，讓真人參加線上投資遊戲：每人可以保留資金，或投入共同基金，再比較不同分配機制。團隊用強化學習尋找參與者願意投票支持的方案。學到的機制兼顧起始資源差異與相對貢獻，並抑制搭便車，在這個遊戲中得到多數支持。2022 年論文 · 團隊說明與限制

研究流程很具體：取得人的選擇資料、訓練機制、再讓人評價。它仍只是小型經濟遊戲中的概念驗證。多數人喜歡的分配法，不自動等於公平的公共制度；團隊自己也指出，多數與少數利益如何權衡，仍有待研究。

2024 年，Michael Henry Tessler、Michiel Bakker 等人發表 AI 協助審議的研究，Botvinick 為共同作者。系統讀取參與者的立場與批評，反覆產生、修訂表達共同觀點的文字。整體研究包含 5,734 名參與者；他們較偏好 AI 整理的文本，並將其評為更清楚、資訊更充分、較少偏頗。團隊也在具人口代表性的英國線上公民議會樣本中重複測試。2024 年論文摘要與作者名單

這裡衡量的是特定情境下對文字與審議結果的評價，不是 AI 已取得代替公民作政治決定的資格。兩項研究的重要性，在於它們顯示他的轉向早已發生在實驗工作的內部：神經網路、行為研究與人類評價，被用於分配與集體討論，而不再只用來解釋個別心智。

Yale 法學訓練與 Rule of Law Clinic，2023–2026

他在讀 Yale 的同時仍與 DeepMind 保持工作關係：履歷列出 2023–2026 年的資深技術與政策顧問職務，以及 2024 年在其法律團隊擔任 Legal Fellow。法律專業訓練與科技業任職在這幾年持續重疊。J.D. 是美國法律專業學位，與他的研究型 Ph.D. 不同。本人 CV

在 Yale Rule of Law Clinic，他以學生團隊成員身分參與法律書狀工作，包括 2025 年涉及 Perkins Coie 的案件，以及針對 Harvard 國際學生限制的法庭之友意見書。兩案均由診所與律師事務所合作處理，學校的公告將他列為參與書狀工作的多位成員之一。Yale：Perkins Coie 書狀 · Yale：國際學生限制案件

在 2026 年 10 月的訪談裡，他說重新讀法律，是為了從法律自身的角度理解問題，不漏掉這個領域特有的洞見。但他也承認，設計制度有工程的一面，並提到自己欣賞耶魯政治理論學者 Hélène Landemore 的公民審議研究。主持人 Andrew Keen 則追問：政治是否真的能以工程的方式處理？這場交鋒保留了他法學轉向的複雜性：他希望學會另一種專業思考，同時仍相信制度可以被設計和改善。Keen On 訪談，約 54:12–56:38

讀他的工作時，仍需分辨兩套評價標準。神經科學模型可以用解釋力與預測表現接受檢驗；法律與政治還必須處理權利、正當性、程序和彼此不相容的價值。模型運作良好，不能直接推出某個制度因此具有正當性。這是他目前研究必須面對的問題，不能因為他讀了法學就假定已獲解決。

Anthropic 與《AI and Political Freedom》，2026

Anthropic 於 2026 年成立 Institute 時，將 Botvinick 列為首批延攬的研究者之一，由他領導 AI 與法治方向。這個位置有明確的機構條件：研究者可以接近前沿系統與產業內部的問題，也同時受僱於發展、銷售這些系統的公司。Anthropic 官方公告

他的工作因此不能只用「懂技術又懂法律」概括。更實際的問題是：如何把 AI 對資訊環境、政治討論、選舉與問責的影響，轉化為可研究的問題，再把研究放回產品與公共政策的討論裡？這是制度內的介入位置，不是站在制度之外的中立觀察台。

2026 年出版的 AI and Political Freedom: The Risk to Democracy and How to Respond，把民主倒退與 AI 風險放在同一張圖上。他擔心的不是只有模型說謊，而是工作與社會角色的變化、緊急權力、監控能力等，如何改變公民與國家的力量關係；另一方面，他也探索 AI 能否協助擴大公民審議。這些是書中提出的論證與可能性，不是已經驗證的未來。Princeton University Press

企業內部研究既有資訊優勢，也有利益與議程的限制。他在訪談中明確區分個人著作與公司立場。閱讀他的論證時，不能因其企業職位而一概否定，也不能因其跨界學歷而免除檢驗；要看公開了什麼證據、概念如何成立，以及哪些重要決定仍由公司自行控制。Keen On 訪談，約 7:11、18:27

工作如何隨著機構改變

Botvinick 的作品留下了幾組具名的合作關係。Cohen 與他共同研究身體知覺與控制；Braver、Barch、Carter 參與衝突監測論文；Ari Weinstein 與他分析階層式規劃；Jane Wang 領銜的團隊則把後設強化學習帶回前額葉研究。這些是不同問題的協作組合，不能簡化成由一個人獨自完成的發現。

在大學，這些合作透過實驗、理論論文與學生訓練展開；在 DeepMind，計算模型可以與工程能力和大規模研究團隊結合；到法律診所，成果變成需要處理法律權威與程序的書狀；現在則是企業內部研究與公共論述。公開資料足以描述這些工作型態，卻不足以編出他的典型一天，或判定每項成果中各人的實際工時與貢獻比例。

到 2026 年，他的工作分布在 Anthropic 的 AI 與法治團隊、耶魯的研究職務，以及新書的公共討論之間。不同場合要求他分別面對模型與實驗、法律論證和讀者的質疑。本人網站 · Keen On 訪談

小時間線

- 1989–1990：Columbia 藝術史碩士；1992 年發表 Campin 論文。
- 1990–1994：Cornell 醫學訓練。
- 1996–2001：CMU 博士；部分時間與精神科住院醫訓練重疊。
- 2002–2007：Penn 任教。
- 2007–2016：Princeton；2013 年升任教授。
- 2015–2016／2016–2026：先訪問 DeepMind，再正式任職。
- 2022／2024：參與分配機制、AI 協助公民審議的研究。
- 2023–2026：Yale J.D.，同時擔任 DeepMind 技術與政策顧問；2024 年於其法律團隊 fellowship。
- 2026：Anthropic AI 與法治方向、Yale Senior Fellow；出版 AI and Political Freedom。

日期依本人 2026 年 CV，並以機構公告交叉核對。精神科資格不等於目前仍在臨床執業。

從哪裡開始讀

1. 先看一個很具體的作品：1998 年橡膠手論文，理解他的問題如何被轉成實驗。
2. 再看一個職業轉折：CMU 的 2019 年訪談，認識方法、導師與制度環境。
3. 想理解 AI 與腦如何互相啟發：2017 年〈Neuroscience-Inspired Artificial Intelligence〉和 2018 年後設強化學習論文。
4. 想看他如何把研究轉向政治問題：2022 年分配遊戲與 2024 年 AI 協助審議論文，搭配其方法與限制閱讀。
5. 想知道他現在在做什麼：2026 年新書與 Keen On 訪談；特別留意訪談後段對「政治是不是工程」的交鋒。
6. 對藝術史最感興趣：從 1992 年 Campin 論文的期刊紀錄找全文；不要用後來的 AI 職涯替尚未讀到的文章補寫論點。

資料查核截至 2026 年 11 月 7 日。本文依公開履歷、研究論文、機構公告與本人訪談撰寫；分析性連線均不作私人動機或未公開事實的證明。