doi: 10.29843/JCCIS.202607_(51).0006

AI 與人類智慧：Daniel Andler《AI 會取代人類智慧嗎？人工智慧與人類智慧的雙重謎團》書評，及其他

黃厚銘

摘要

本文評析 Daniel Andler 著作《AI 會取代人類智慧嗎？》，該書探討人工智能與人類智慧的本質差異。書中梳理 AI 技術演進，最終否定了 AI 能夠取代人類智慧的主張。

Andler 指出，將智慧狹義定義為「解決問題的能力」具有侷限性，進而提出了 AI 欠缺的「情境」（situations）概念。不同於 AI 作為「語意盲」的本質，人類智慧具備理解、利用乃至透過行動改變情境的能動性，這與社會學俗民方法學中行動與情境互為表裡的觀點相呼應。 在風險議題上，本書深入探討「價值對齊」之難題。Andler 認為 AI 風險未必僅來自惡意叛變，更多源於其缺乏人類常識。此外，AI 的自主性與人類控制間存在內在矛盾：當 AI 越趨自主，其追求自身目標而非人類目標的可能性便隨之增加。

綜合評析其他探討 AI 與人類社會的其他著作，本文主張，在思考人工智能與人類智慧時，不應以「擬人化」框架設想 AI，亦需提防以 AI 運作邏輯反向誤解人類思考的本質。

⊙ 關鍵字：人工智能、人類智能、價值對齊、俗民方法學

⊙ 本文作者黃厚銘，國立政治大學社會學系教授，研究專長為社會學理論、資訊社會學，與文化社會學。

⊙ 聯絡方式：Email ：hermes@nccu.edu.tw；通訊處：台北市文山區指南路二段64號政治大學社會學系。

⊙ 致謝詞：本篇書評感謝研究助理長弘與聖堤細心校對，並提供行文修辭的各種建議。

⊙ 收稿日期：2026/06/30　接受日期：2026/07/01

AI and Human Intelligence: A Review of Daniel Andler's "Intelligence Artificielle, Intelligence Humaine : La Double Énigme", and More

Arthur Hou-Ming Huang

Abstract

This article evaluates Daniel Andler's work, Intelligence artificielle, intelligence humaine: la double énigme, which examines the fundamental distinctions between artificial and human intelligence. By analyzing AI's technical evolution, Andler refutes the claim that AI can replace human agency. He argues that intelligence exceeds mere problem-solving, emphasizing the role of "situations"—a domain where human agency interacts with and transforms contexts as ethnomethodology claims, unlike the "semantically blind" nature of AI.

Regarding risk, the review discusses "value alignment," highlighting how AI's lack of common sense and increasing autonomy can lead to catastrophic outcomes that diverge from human goals. Ultimately, synthesizing various perspectives, the article warns against anthropomorphizing AI and cautions against using AI's operational logic to "reverse-misunderstand" the true essence of human cognition.

⊙ Keywords: artificial intelligence, human intelligence, value alignment, ethnomethodology

⊙ The author, Hou-ming Huang, is a professor of the Department of Sociology at National Chengchi University, specializing in sociological theory and cybersociology, and cultural sociology.

⊙ Email contact: hermes@nccu.edu.tw. Address: Dept. of Sociology, NCCU. No.64, Chinan Road Sec. 2, Wenshan, Taipei, Taiwan.

⊙ Acknowledgement: I would like to thank my research assistants, Chang-Hung and Sheng-Ti, for their meticulous proofreading and for providing insightful suggestions on phrasing for this book review.

⊙ Received: 2026/06/30 Accepted: 2026/07/01

在繁體中文世界有關 AI 的翻譯書籍中，《AI 會取代人類智慧嗎？人工智慧與人類智慧的雙重謎團》（2023／林佑軒譯，2025）佔有著特殊地位。相較於《AI 世代：從政治哲學反思人工智慧的衝擊》（2022／鄭楷立譯，2024）一書的作者 Mark Coeckelbergh，或是與 Coeckelbergh 合著《當 AI 取得話語權，人類還剩下什麼？》（2025／江鈺婷譯，2026）的 David Gunkel，兩位都是政治哲學與技術哲學方面的專業。《AI 會取代人類智慧嗎？人工智慧與人類智慧的雙重謎團》此書作者 Daniel Andler 的特別之處，對我來說，不在於他身為法蘭西人文科學院院士的崇高地位，而是他身兼數學家、哲學家與認知科學家的跨領域專長。

而這也顯示在他這本書的內容結構。將近五百頁的一本書，有超過半本書的份量是從原理與技術發展，來為讀者奠定關於 AI 的知識基礎。相較之下，《當 AI 取得話語權，人類還剩下什麼？》僅以三十頁不到的〈大語言模型 101〉做簡要的介紹，這本身就足以彰顯前書的特色。也唯有了解 AI 的原理與技術發展，才不至於過於高估或低估 AI 的能耐，方能找出看待 AI 與面對 AI 時代的適切態度。無論如何，儘管作者 Andler 具有數學家的身分，本書卻沒有像 Nick Bostrom 的《超智慧：AI 風險的最佳解答》（2014／唐澄暐譯，2023）那樣不時出現數學式，這一點讀者可以放心。

Coeckelbergh 在《AI 世代：從政治哲學反思人工智慧的衝擊》中，雖然強調應結合政治哲學與技術哲學兩個領域，以明確反對「技術中立」的立場，但在各章節對平等、正義、民主與權力等議題的價值討論，仍是偏向持平呈現 AI 所帶來的正負面影響。另一方面，在 Coeckelbergh 與 Gunkel 合著的《當 AI 取得話語權，人類還剩下什麼？》一書中，由 ChatGPT 所寫的〈序〉拉開全書序幕，並以「樂意撤回作者身分」作結。透過引入「作者已死」或「是語言在說人，而非人在使用語言」等論述，推導出「應賦予 AI 溝通主體身分」的主張。相較前述兩者，在《AI 會取代人類智慧嗎？人工智慧與人類智慧的雙重謎團》一書中，Andler 開宗明義便指出，他將以整本書的篇幅否定本書中文標題的提問—也就是他主張，AI 無法取代人類智慧。

《AI 會取代人類智慧嗎？人工智慧與人類智慧的雙重謎團》這本書從第一章〈人工智慧概覽：從願景到策略退卻〉開始，首先回顧「AI 的普羅米修斯時刻」，也就是諸如圖靈（Alan Turing）等 AI 領域的先鋒們，深信可以創造出與人類智慧一樣的人工智慧。然而，後續的發展不如預期，導致這些先鋒們的理想，逐漸被視為一種

烏托邦式的想像（Andler, 2023／林佑軒譯，2025，頁 29-39）。不過，在今日，這個願景似乎隨著 ChatGPT 等大語言模型所展現的強大能力而捲土重來。其中也包括所謂「通用型 AI」（AGI）的構想。

隨之，Andler 先以〈人工智慧的根源〉一章介紹 AI 相關的基本概念。接著即分別針對〈古典時代〉、〈連結主義年代〉與〈第三時代拉開序幕〉，以篇幅龐大的三章來回顧 AI 的技術發展，以便理解 AI 領域的先鋒們早被放棄的願景為何會再度興起。藉此，我們也可以理解深度學習、變換器（transformer），以及大語言模型和「生成式人工智慧」（GAI）等身處 AI 時代應該被視為常識的一些基本概念（Andler, 2023／林佑軒譯，2025，頁 139-175、頁 190-208、頁 209-225）。

如本文一開頭所指出，正是奠基在對 AI 相關的知識基礎上，Andler 才進入本書的第二部分「智慧之問與 AI 的未來」，著手回答本書的問題意識：「AI 會取代人類智慧嗎？」對我來說，又是一連串極為精彩與環環相扣的嚴謹討論。

Andler 在第二部分開頭的第六章提問〈人工智慧與人類智慧能比較嗎？〉，以此指向 AI 所欠缺的幾種能力，這包括意識、意義的理解、常識，以及後設認知與情感等等。實際上 Andler 所要指出的是，把智慧定義為「解決問題的能力」的不足之處（Andler, 2023／林佑軒譯，2025，頁 281-315）。進而在第七章，Andler 別開生面地提出「問題」與「情境」這兩個概念的區分。根據 Andler 在別的場合接受訪談的報導，他說這是承繼自他在 Berkeley 求學期間的哲學系教師 Hubert Dreyfus，並予以基進化。Hubert Dreyfus 也是《What Computers Can’t Do》（Dreyfus, 1972）一書的作者（傅莞淇，2026）。對我來說，Andler 對情境的討論幾乎與晚近社會學的觀點不謀而合。亦即，「情境」既是行動的環境，也是行動所打造的。Andler 不只反對把智慧定義為「解決問題的能力」，也強調人類智慧於理解情境、利用情境，甚至是以行動改變情境的能力。以此，Andler 否定了 AI 與人類智慧之間的可比較性（Andler, 2023／林佑軒譯，2025，頁 309-315）。

我尤其想強調 Andler 筆下的「情境」概念（Andler, 2023／林佑軒譯，2025，頁 334-365）。從俗民方法學的角度來看，在實際的互動過程中，當對方的回應不如預期時，我們反而可能產生諸多不同的設想和回應方式。比如說，在開創俗民方法學的 Harold Garfinkel（1967, pp. 24-31）自身知名的破壞性實驗中，學生繳交作業給老師，

卻被老師退回、要求補強、重寫。學生們一開始的反應會是：那應該是我做得不夠，所以會把作業帶回家改進。當我們沒有收到期待的回應時，可能會以為是我們傳達得不夠清楚，就像是學生以為他的作業寫得不夠好。但我們也可能因為覺得莫名其妙卻又不知原因，所以暫時無視對方的回應，而不是修正自己。在極少的情況下，我們也可能立即暴怒。以 Garfinkel 的破壞性實驗為例，他一再地退回學生的作業，最終導致學生認定老師在刁難他們。在遭遇不如預期的回應時，我們會依據眼前的情境產生不同的理解，並依此發展出種種可能的反應。這才是真實人際互動的樣貌。學界的朋友不妨參考一下自己投稿、收到審查意見、修稿、甚至被退稿的過程。

扣緊「情境」一詞，以俗民方法學的措辭來說，社會互動並非只是受限於情境、被情境決定、回應情境而已，也可能改變情境。在行動造就情境的同時，回過頭來，情境也定義了我們的行動。行動不是外在於情境的，情境也不是外在於行動的。以 Garfinkel（1967, pp. 1-4）的話來說，其間的關鍵是「解明」（account），也就是前段所謂的「設想」和「回應」，這反而正是 AI 所欠缺的。

但就此衍生出來的疑問是：有沒有可能 AI 是一種有別於人類智慧、甚至在某些方面超越人類能力的智慧？循著這個思考路徑，我們來到本書的末章〈人工智慧與善〉。在我個人特別重視的這一章裡，Andler 指出 AI 倫理與諸如醫療倫理之間的差異。相對於醫療倫理等其他倫理所規範的對象是人類，AI 倫理卻包含了一個特殊的面向，那就是「如何讓 AI 具有與生俱來的『良善』」（Andler, 2023／林佑軒譯，2025，頁 424-430）。這問題看似荒謬，但之所以需要納入思考，其中一個原因就是 AI 的強大力量。我們必須在 AI 發展為超智能（superintelligence）之前就將其作惡的可能性完全消除，否則後果將不堪設想。這就是我在 AI 科技風險議題上一再強調的科技失控風險。

這不僅適用於過於「聰明」的 AI，也適用於不夠「聰明」的 AI。用前述 Andler（2023／林佑軒譯，2025，頁 289-300）所引進的概念來說，亦即 AI 可能因為缺乏人類所具有的常識，而以超乎人類意料、同時也是人類想都不必想就不可能採取的方式，達成人類所賦予的目標，進而產生無可預期的災難性後果。而這在 AI 工程上就稱為「價值對齊」（value alignment）的研究議題。

其中，Andler 特別針對以 Anthropic 公司為代表的「憲法式 AI」的解方提出他的

分析與評論。然而，光是以不能偏頗（bias）為例，他即呈現，偏頗有時卻是實現正義的工具（Andler, 2023／林佑軒譯，2025，頁 434-445）。我會說，這本書是晚近幾本有關 AI 的繁體中譯書籍中，少數深入探討價值對齊的書籍。而最終，Andler 指出了一個矛盾，那就是人們一方面想要研發出具有自主性的 AI，另一方面卻又希望 AI 能夠幫助人類實現他所提出的目標。但事實是，AI 越自主反而越有可能轉而追求它自己的目標。這甚至意味著 AI 除了失控以外，更有叛變的可能（Andler, 2023／林佑軒譯，2025，頁 452-455）。

更值得肯定的是，Andler 以本書剩下來的篇幅，與重要 AI 教科書作者之一、也是加州大學 Berkeley 分校電腦科學教授的 Stuart Russell 對話，指出 Russell 在價值對齊的議題上所提出的合作式逆向強化學習（CIRL）所面臨的疑難，並在本書末尾呼應暫停 AI 研發的連署呼籲。Andler 在結論中明確指出，他不想要「一個越來越離不開人工智慧的世界」（Andler, 2023／林佑軒譯，2025，頁 455-461）。

總之，對於不是這個領域的讀者，《AI 會取代人類智慧嗎？人工智慧與人類智慧的雙重謎團》是一本頗難的教科書。內容有一定的深度，並且論證是層層推進，一分心就會跟不上作者的推論。但我將之定位為教科書，也是要說，這是一本值得仔細閱讀的重要著作。

在這篇以 Andler 的著作為主的書評中，我也想談談牛津大學哲學教授 Nick Bostrom 的《超智慧》（Superintelligence）。這本書一開始的那則麻雀與貓頭鷹的寓言（2014／唐澄暐譯，2016，頁 3），一來就讓我覺得有趣，二來也可以看出他和前述的 Andler 一樣擔憂失控的問題[1]。無論如何，《超智慧》這本書，又是另一本「教科書等級」（聽說這措辭是 AI 慣用的）的著作。書裡的許多補充說明的「方塊」，進一步向讀者解釋更為技術性或原理性的深奧知識，要全部吸收了解，勢必得花一番

1. 但這倒是跟他的另一則有名的「巨龍暴政」寓言（2005）的立場似乎有些不同。對我來說，後者幾乎是無視於科技風險，而讓我很意外。如果我的理解沒有錯誤，這寓言是用以呼籲發展抗拒死亡或延緩老化的科技，也因此成為超人類主義的重要文獻之一。然而，在中國國家主席席近平與俄羅斯總統普丁曾私下聊到器官移植與延長壽命的脈絡下，卻只會讓我覺得很可怕。

工夫。我會說，Bostrom 非常博學，全書涵蓋的知識從演化、腦科學、AI 技術等等，無所不包。書裡（包括註釋）有不少數學式，甚至還有微積分（！）。也觸及了當今的大國博弈，以之探討超智慧勢不可擋的發展趨勢。不同於上一本書的作者 Andler，Bostrom 相信超智慧是可能的，這本書正是以此為前提來進行各方面的討論。

儘管如此，書裡一個對我來說很重要的設想是：一個不友善的超智慧可能會發現，它必須在沙盒裡表現得友善，才有可能讓人類把它釋放到沙盒之外（2014／唐澄暐譯，2016，頁 147-148）。我不知道這算不算是用人類去設想 AI，但這提醒了我們，即使是想要以沙盒實驗的方式把超智慧的影響框限在一個範圍之內以免其失控或叛變，這樣的解方仍舊可能是失敗的。這科技叛變的可能性是許多科幻小說或電影的題材，看似是沒有科學根據的想像，卻已經進入像本書這樣的學術著作的視野之中。這一方面值得我們嚴肅看待與未雨綢繆。另一方面，也顯示文學藝術作品的啟發性。這在越來越偏頗於理工、甚至聚焦於半導體研發，而忽視人文社會科學的台灣教育體制中，更顯其重要性。無獨有偶的是，Bostrom 在書裡也談到了希臘神話中 Cassandra 這位沒有人相信她預言的公主（2014／唐澄暐譯，2016，頁 147-151）。我想說的是，相較於科學，神話、小說、電影的地位越加邊緣化，卻很可能是 AI 時代的 Cassandra，人文社會科學也是。

進而，書中以終極目標和智慧之間的正交關係，舉出了許多可怕的例子。像是，我們希望 AI 讓我們微笑，結果非常聰明的 AI 就麻痺我們的臉部肌肉，讓我們持續地微笑。這正是我屢次所說，AI 的失控風險可能來自它太努力達成人類所設定的目的了，而不是 AI 叛變。Bostrom 在此以極其淺顯易懂的例子，讓我們了解晚近的 AI 技術發展所面臨的挑戰（即使不說是困境的話），而這也是前述「價值對齊」這個研究議題之目的。

Bostrom 和 Andler 都指出，不要以擬人化的方式來設想 AI 或 AI 的發展（Bostrom, 2014／唐澄暐譯，2016，頁 116-117；Andler, 2023／林佑軒譯，2025，頁 313-315）。也就是，不論是研發，還是就 AI 的影響來說，以人腦或人類為模型，都可能是錯誤的。所以，人類亦不該以自身的動機來設想 AI 的動機。Bostrom 說：「有一點至關重要：在思考超智慧的潛在影響時，我們不要將它們擬人化（anthropomorphize）。不管是在種子人工智慧的成長軌跡，還是在成熟超智慧的心

理、動機和能力方面，陷入人類的框架，都會萌生毫無根據的期待。」（2014／唐澄暐譯，2016，頁 116）在身兼電腦工程師與科幻小說家的許順鏜之新作《人會夢想成為 AI 嗎？》（2025）之〈自序〉裡，針對 Geoffery Hinton 認為 AI 可以模擬人腦的運作，進而擁有「意識」的主張，許順鏜也回應：這既是羞辱人，也是羞辱 AI。既貶低人是 GPU 製作出來的機器，也讓我們只想做出跟人類一樣的 AI，而不是開發 AI 真正可以超越人類的那一部分能力。對我來說，前述 Bostrom 的說法的重點在於「不要以人類來設想 AI」。但許順鏜的說法還有另一個意涵，那就是「不要以 AI 來設想人類」。這也意味著 AI 的發展還可能回過頭來讓我們誤解人類大腦的運作或思考的本質。這一點在許順鏜的書序裡，就是指「以為人腦就只是像 GPU 一樣」。而 Russell（2019, pp. 135-136）也曾在討論是否能夠阻止 AGI 的研發時說，阻止 AGI 的發展就等於放棄了理解人類智能是如何運作的一個重要途徑。這也證明了 AI 的研發一直也帶著用以理解人類思考之謎的目的。

由此可見，有關 AI 與人類智慧之關係，我們面對的是兩個方向的類比。Bostrom 所要提醒的是：既不要以模仿人類智慧為目標來研發 AI，也不要以人類的思考來設想 AI 的思考。前述 Bostrom 書中有關種子人工智慧的成長軌跡之討論，就是指研發過程不要以人類智慧為標的。而所謂推測或假想成熟超智慧的心理、動機和能力，則是以人來設想 AI，甚至認為 AI 會「騙人」。也是基於這樣的理由，我才強調，AI 的科技失控風險未必來自它們有「意圖」的「叛變」。

然而，如同許順鏜的主張，我還要強調第二個方向，亦即，我們也可能用 AI 的研究來理解人類智能。不論是 Andler 或許順鏜的書，都主張：正因為 AI 的研發可能讓我們誤解了人類智能或意識、思考的本質，我們反而更需要認識人類智能不同於 AI 之處（Andler, 2023／林佑軒譯，2025，頁 254-263；許順鏜，2025）。如前所述，Andler 列舉了 AI 在意識、意義、常識與後設認知、情感等方面的欠缺，而反襯出人類智慧的獨特性。他在書中的其他部分，則強調人類的經驗的獨一無二、不可重複，以及前述的情境（situations）概念。他甚至在書中也直接說 AI 是「語意盲」，其實也呼應了本文前述所引入的俗民方法學之討論。

總之，不僅如前面的幾位作者所說，不該用擬人化的方式來設想 AI，我們確實需要提防 AI 的風潮讓我們誤解了人類、誤解了思考及意識究竟是怎麼一回事，以免我

們太過輕易地覺得，AI 可以取代人類、可以幫人類做許多決定。就像第五級的自動駕駛一樣，期待可以完全放手。這對我來說，才是最可怕的狀況。更不用提在現階段就在高速公路上的特斯拉電動車裡蹺腳打瞌睡。同理，不論價值對齊的技術如何發展，我深信最終得承擔價值判斷的責任的，仍然是人類。

最後，我還是要補充一點。除了涉及前面所提到的書籍內容之外，本文刻意使用英文簡寫的 AI，而非「人工智慧」一詞。這是因為中文的「智慧」所對應的英文應該是 wisdom，而前述書中所談的，不論原文還是更細節的內容，都只是 intelligence。沿襲幾十年前工程界的翻譯，而把 intelligence 翻譯為「智慧」，以致書中不斷提到「人工智慧」，甚至是「智慧」，真的會在閱讀時帶來許多困擾——明明談的不是 wisdom，卻不斷用著「智慧」一詞。試想，我們認為會使用語言或是有「意識」就等同於有「智慧」嗎？儘管「人工智能」是中國的翻法，個人還是覺得這才是比較正確的翻譯。

參考文獻

許順鏜（2025）。《人會夢想成為 AI 嗎？》。台北：蓋亞出版。

傅莞淇（2026 年 2 月 13 日）。〈AI 能力再強也不會等同人類，法國哲學家安德勒談 AI 與人類的永恆距離〉，遠見。https://www.gvm.com.tw/article/128095

林佑軒譯（2025）。《AI 會取代人類智慧嗎？：人工智慧與人類智慧的雙重謎團》。台北：時報出版。（原書：Andler, D. [2023]. *Intelligence artificielle, intelligence humaine: la double énigme.*）

唐澄暐譯（2016）。《超智慧：AI 風險的最佳解答》。新北：感電出版。取自：https://nccu.ebook.hyread.com.tw/bookDetail.jsp?id=339136（原書：Bostrom, N. [2014]. *Superintelligence : paths, dangers, strategies*.）

鄭楷立譯（2024）。《AI 世代：從政治哲學反思人工智慧的衝擊》。台北：商周出版。（原書：Coeckelbergh, M. [2022]. *The political philosophy of AI : an introduction.*）

江鈺婷譯（2026）。《當 AI 取得話語權，人類還剩下什麼？》。台北：商周出版。（原書：Coeckelbergh M. & Gunkel, D. J. [2025]. *Communicative AI : a critical introduction to Large Language Models*.）

Bostrom, N. (2005). The fable of the dragon tyrant. *Journal of Medical Ethics, 31*(5), 273–277. https://doi.org/10.1136/jme.2004.009035

Dreyfus, H. L. (1972). *What computers can't do*. New York: Harper & Row.

Garfinkel, H. (1967). *Studies in ethnomethodology*. New Jersey: Prentice-Hall.

Russell, S. (2019). *Human Compatible*. New York: Viking.