《歐美研究》第五十卷第二期 (民國一〇九年六月)，147-178
DOI: 10.7015/JEAS.202006_50(2)0002
https://euramerica.org

戴維森論圖靈測試*

趙之振
國立清華大學哲學研究所
E-mail: ccchiu@mx.nthu.edu.tw

摘 要

圖靈提出以模仿遊戲來測試機器能否思想。戴維森論證這測試是無法決定受測對象之語意學，故而無法決定機器能否思想；但庫辛思基卻反駁戴維森的論證。針對此中爭論，本文首先勾勒並釐清戴維森對模仿遊戲所作之評述，然後展示他對圖靈測試之批評；其次是論證庫辛思基批評戴維森之理據或是不成立，或是奠基於其對戴維森哲學的誤解之上；最後則是展示戴維森有關思想之條件的觀點與 AI 的關係：對於我們可以如何修訂圖靈測試，以此判定受測對象是否具有思想，戴維森的看法可以給我們怎樣的啟發。

關鍵詞：戴維森、圖靈測試、思想歸屬、徹底翻譯、人工智慧

投稿日期：108.6.26；接受刊登日期：109.4.6；最後修訂日期：109.4.12
責任校對：陳昱之、吳承憲、黃意函

* 感謝方萬全、王道維以及三位論文審查人給與寶貴意見。本文初稿曾發表於中央研究院歐美研究所主辦之「歐美 AI 的發展與挑戰」跨學門研討會，2019 年 5 月 9 日至 10 日。感謝與會者之提問與討論。又本文撰寫期間獲科技部研究計畫支助 (計畫編號 109B0002Q4)，謹此併致謝忱。

壹、前言

由於運算技術之提升、演算法之進步以及各種數據之累積與分享，近年來人工智慧（智能）蓬勃發展，讓越來越多的機器在各方面展示人類的智能，同時也在各領域中以不同的方式介入人類的社會。該如何看待這些機器？這終究是我們要面臨的問題。然而，如何看待一機器，跟我們認為這機器到底是什麼（或具有什麼的特質）有很大的關係。在機器可能擁有的種種特質中，思考是相當重要的。我們可以與一台聊天機器人（chatbot）對話，但一般來說，我們大概不會把它看作是平常聊天的朋友，因為我們不只不會把它當作人，我們大概也不會認為它是會思考或具有思想的。不管是否如此，這卻可以引起一個問題：我們如何判定機器是否有思想呢？圖靈測試（Turing Test）常被認為是檢視一個機器能否思想的判準；但這樣的判準是否恰當？這問題在哲學界引起了相當的爭論。一個備受觸目的例子是塞爾（John R. Searle）(1980) 所提出的中文房論證（Chinese Room Argument）。[1] 然而，很少人注意到戴維森（Donald Davidson）(2004a) 也曾對此測試提出其看法。戴氏對此問題之回應，主要是來自於他對思想歸屬之條件或思想之本性的看法。本文之主要目的，是試圖依循戴維森的思路，來反思在什麼條件之下，我們可以把思想或思考活動歸屬給機器。本文進行的步驟如下：首先，我將簡要地勾勒戴維森對圖靈所提出的模仿遊戲（imitation game）所作的評述，然後展示他對圖靈測試的批評意見。由於庫辛思基（John M. Kuczynski）主張戴氏對圖靈的批評是不成立的，我接下來的工作便是對庫氏的批評予以回應，試圖論證他對戴維森的批評並不成功。最後，我將展示戴維森有關思想之條件的觀點與 AI

[1] 關於此論證的簡要陳述可參看 Searle (2004: 89-92)。

(或機器思想) 的關係：對於我們可以如何修訂圖靈測試，以此判定測試對象 (電腦) 是否具有思想，戴維森對此的看法可以給我們怎樣的啟示。

貳、圖靈的模仿遊戲

圖靈在〈計算機器與智能〉(Turing, 2004a) 一文中，開始即提出「機器能否思考？」的問題。在當時，哲學家回答此問題的流行方式是透過對問題中相關語詞的日常意義來進行分析，但圖靈明言這不是他採取的進路，而是提出以下的「模仿遊戲」來取代原先的問題：假設有三個人，一位是男性 (A)，一位是女性 (B)，還有一位是提問者 (C)，C 可以是男的，也可以是女的，他或她跟 A 與 B 隔離，處於另外一房間，只透過電傳打字機 (teleprinter) 來跟 A 或 B 溝通聯絡，進行各種的問答。遊戲的方式是：由提問者向 A 與 B 提出種種問題，透過往返問答，以決定兩者中何者是男的，何者是女的。現在圖靈的問題是：如果在上述的遊戲中，我們以一台機器來替換 A，則 C 在決定 A 與 B 的性別時，其犯錯多寡之比例會跟替換前的情形一樣嗎？圖靈便是以這個問題來取代原先有關機器能否思考之問題。他進一步說明上述遊戲中的機器乃是限定為數位電腦 (以下簡稱「電腦」)。因此，電腦能否思考的問題，便轉化為這樣的問題：電腦能否通過模仿遊戲的測試？亦即電腦能否成功地模仿 A，讓提問者 C 之辨認 A 和 B，與其辨認電腦和 B，在正確的程度上是無差別的？[2]

[2] 圖靈之模仿遊戲可見於其更早的論文〈智能的機器〉(Turing, 2004b)，在該文中，機器所模仿的是人類棋手，其場景佈局與上述的遊戲類似。當時圖靈已經認為：在這樣的情形中，要辨別電腦與真正的人類棋手可能是相當困難的。

戴維森對圖靈測試的評論，基本上也是從上述的模仿遊戲著手的。他注意到以下有關這遊戲的一些特點，並對之作一些簡略的評述。

第一、圖靈把遊戲中的機器只限定在數位電腦，他並沒有考慮其他類型的機器有可能比數位電腦來得更好。這固然是因為圖靈熟悉數位電腦，並且相信：如果我們正確地設計數位電腦，則它可以通過上述的模仿測試。尤有進者，透過這樣的規定，他還可以避免對「機器」給出一個一般性的定義。之所以要避免這樣的定義，是因為圖靈認為：一般來說，人們對「機器」的自然想法其實並不一致。一方面，我們希望容許任何種類的工程技術 (包括生物工程技術)，都可運用到機器之上；再者，我們也希望容許這樣的可能性：工程師有可能建構一個可以運轉的機器，即使建構者對其運作方式之描述並不令人滿意；另一方面，如果我們真的透過生物工程的手段，從某人的一個皮膚細胞培養出一個完整的會思考的人，我們也不會願意將這樣的作為當作是「建構一個思考的機器」。[3] 在討論「機器能否思考？」的問題時，圖靈的作法乃是把題目中所談到「機器」限定在數位電腦，以為如此便不必擔心「如何定義『機器』」的難題。然而戴氏認為：圖靈應該要擔心以下的理論可能性：我們把人的神經電路圖重建於一台會思考的數位電腦 M，而且我們在使用相關方法之過程中，並不瞭解由此所產生的程式為何以及如何會使得 M 具有思想 (Davidson, 2004a: 78)。[4] 我認為這樣的可能性之所以被戴氏認為是圖靈應該擔心的，是因為一方面 M 與一般具神經系統的生物相似，其建構是 (部分地) 依賴生物工程技術，但另

[3] 圖靈此處有關機器的意見可見於 Turing (2004a)。

[4] 要注意的是，此處所說的可能性，與先前圖靈對一般人對機器的自然想法之描述是一致的。又戴氏此處所描述的情形與當今某些型態的深度學習也多少有些類似。

一方面 M 又是一台數位電腦，因此使得我們難以決定它是否應被視為思考的機器。因此，戴氏認為圖靈對於「把測試限於數位電腦」這種作法的理由，其實並沒有想得很清楚。不過這並不要緊，因為戴氏聲稱：他對圖靈測試所作的評述，並不僅限於數位電腦而已 (Davidson, 2004a: 78)。

第二、如前所示，圖靈原先設計的測試是與性別有關的。由於圖靈是同性戀者，這樣的設計是否有其深意，暫且不論；但戴氏認為：在測試中，這樣的性別要素只是偶然的，亦即這其實是不必要的，我們可以要求提問者在人與機器之間作選擇即可 (Davidson, 2004a: 78-79)。以此之故，戴氏把上述圖靈之原始版的測試，改為更接近通行版本的圖靈測試，讓測試中的提問者去分辨人與機器即可。依據此版本的圖靈測試：如果在一定的時間中，提問者透過問答無法分辨人與機器，則我們可以說該機器是會思想的。

第三、戴氏認為：圖靈並不確定「思想」這個字詞之標準的或日常的用法是否可以正確地、乃至有意義地應用到機器身上。戴氏引用圖靈的原文，表示「機器能否思考？」這個問題「太沒有意義，以致於不值得討論」(Davidson, 2004a: 79)。然而，我認為戴氏的引文是容易令人誤解的，讓人以為圖靈主張原始的問題是沒有意義的。以下更完整的引文可以清楚地顯示這句話所處的脈絡：

> 我相信：大約在五十年內，有可能替電腦設計程式……使它們可以很好地從事模仿遊戲，以致於一名一般的提問者，在五分鐘的提問之後，不會有超過 70% 的機會作出正確的辨認。我相信：那個原始的問題「機器能否思考？」會成為太沒有意義，以致於不值得討論。然而，我相信：在世紀末，字詞之使用與一般受過教育的意見都將會有大幅度的改變，以致於人們可以談到機器思考而不會預期有任何矛盾。(Turing, 2004a: 449)

從上述的脈絡可知，圖靈並不是主張「機器能否思考？」是 (在語意上) 沒有意義的，而是認為：當電腦之模仿遊戲能夠達到上述世紀末程度的時候，有關它能否思考的問題便不再值得討論了，我們也不會認為機器思考之說法有何矛盾之處。[5] 既然「機器思考」的說法是被容許的，則「機器能否思考？」之問題自然就不會是沒有意義的了。此外，圖靈也曾表示：模仿遊戲的問題與原始的問題是密切相關的，只是前者是以相對地清楚明白的語詞來表達而已。因此，圖靈自是不會認為原先有關「機器能否思考」的問題是沒有意義的 (Turing, 2004a: 441)。

第四、圖靈的模仿遊戲是把一些被認為跟思想無關的要素 (例如聲音) 排除在外。戴氏自己也認為：在判斷一個對象是否會思考時，該對象是由什麼物質所組成，確實不是重要的因素。他甚至認為：假如他發現一己的好朋友原來是由一顆蛋所孵化出來的，或者是由矽所組成的，這大概也不會影響他對其好友作如此的判斷：他／她是會思想的 (或者是一個人)。[6] 因此，從圖靈來看，他安排在測試中的提問者對其對象的大部分的物理特性 (聲音、身體等) 全然無知，這樣的做法似乎是恰當的。不過戴氏提醒我們，即使如此，提問者還是具有這樣的知識：他／她知道其自身所接受的輸入是起因於被提問的對象的。這些輸入正是其所期望可藉之以判斷對象是

[5] 從目前的情形來看，圖靈當年的想法或許不免過於樂觀，但未來的趨勢則有可能如其所言。

[6] 在上述假設的情形，戴氏說他會認為他的朋友仍然是人，也說這不影響他會認為這位朋友是會思想的。雖然判斷一個對象會思想，跟判斷一個對象是一個人，兩者不宜混淆，但戴氏並沒有犯此混淆，他可以合理地認為人是會思想的，同時也容許：一個人有可能不是一個像目前人類這樣的生物體 (甚至有可能根本不是生物體，而是由矽元素所組成的)，如同他對其友人所想像的那樣。換言之，對戴維森來說，人格 (personhood) 可以體現在生物體，也可以體現在非生物體。

否有思想之線索。戴氏認為：在此一意義之下，圖靈可以被看作是行為主義者，但這不是一種化約論式的行為主義；他並沒有主張心靈的字詞應當被消除，或者證據必須以某類特定的字詞來表述 (Davidson, 2004b: 79-80)。

第五、戴氏注意到：在圖靈測試中，如果我們有證據來支持測試中的對象正在進行思想活動，則這些證據同時也是「它正在思想什麼」之證據。然而，對人來說，在正常的情形之下，我們往往可以辨認出某對象在思想，而不必同時決定該對象在想什麼。[7] 這種情形在圖靈測試卻是完全被排除的 (Davidson, 2004b: 80)。或許可以這樣說：在圖靈測試中，我們是從「測試對象在思想什麼」來斷言「測試對象正在思想」。

---

[7] 初步來說，至少就生物而言，我們或許可以同意戴氏此一常識性的觀察。不過有趣的是，戴氏此處提供的一個理由是：我們可以透過他人的證詞 (testimony) 或其他非觀察的方式來分辨一個生物體是男人或女人 (Davidson, 2004b: 80)。乍看之下，這個理由看似不恰當，因為「一個對象是會思想的」與「一個對象是人 (不管男女) 」是有差別的。雖然人都是會思想的，但是會思想的不必是人，因此，判別「人」之條件與判別「思想者」之條件不必是等同的。然而，由於「作為人」是「作為思想者」之充分條件，如果我們可以知道一個人是男人或女人而不需要知道其思想內容是什麼，則戴氏以此例子來支持其觀察，是可以理解的。另一方面，我認為戴氏的說法跟他對徹底詮釋 (radical interpretation) 的觀點是相呼應的，在有關後者的陳述中，他主張在詮釋的過程中，可以讓說話者有所謂的「秉持語句 (P) 為真」(holding a sentence (P) true) 之態度，意即說話者認為語句 P 為真，對 P 採取肯認的態度。在徹底詮釋的情境中，即使詮釋者對說話者說出的某語句 P 之意義不瞭解，但仍可將上述態度歸給被詮釋的說話者；換言之，雖然詮釋者不知道說話所說出的 P 之內容，但仍可認為說話者有「秉持語句 P 為真」之態度 (Davidson, 1984d: 162)。如果我的想法是正確的，則戴氏似乎是主張：在圖靈測試中，當我們對電腦作詮釋時，上述態度之歸屬是不會發生的。關於徹底詮釋，下文還會有所討論。

參、戴維森論圖靈測試之困難

戴氏對於圖靈測試的興趣主要是來自他對「具有思想」之判準或「思想之本性」的關懷，而不是分辨人與機器。為了聚焦於思想的問題，他將圖靈的測試 (模仿遊戲) 作更進一步的簡化：提問者面對的不是兩個有待分辨的人或機器，而是只有單一的對象，提問者的工作便是要透過問答來決定這個對象是否正在思想。無論這與圖靈提出的測試或通行的版本是否有所不同，我們可以說：這樣的測試仍然是圖靈式的，它基本上還是具有上述測試的一些重要的特點，因此我們還是可以將之視為一種圖靈測試。而且戴氏認為他對這簡化版的圖靈測試的批評，同樣也可以應用到原先的圖靈測試身上。現在的問題是：簡化版圖靈測試是不是一個恰當的判準，使我們可以判別一個對象是否在思想？戴氏的答案是否定的。他的理據最終是來自於他對「思想之條件」(亦即「具有思想」之條件) 的看法，但是我們可以從圖靈測試的進行開始。

根據先前所述，在圖靈測試當中，提問者與其對象在空間上是完全被隔離的。假設提問者的母語是英語，他／她與對象之間只能透過文字符號 (假設與英文符號相同) 往來問答，藉此判斷其對象是否在思想。在此情形，提問者支持其判斷的證據，只在於這些往返的文字。於此，我們暫且關注以下兩個問題：(一) 這些文字符號在語法 (如文法、字詞) 上是否與英文相符？(二) 這些往來文字之間的關係是否恰當，例如，對答是否合理？所言是否一致的？等等。

第一個問題是屬於語法層次，人們不難透過形式的標準，來決定從測試對象而來的文字是否符合英文的語法形式。然而，僅僅在語法層次的滿足，並不足以決定相關的字詞或語句便是英文的語句，因為當中這些語詞之所指或語句之意義，如果有的話，可能跟我們所使用的英文大不相同，亦即這些貌似英文字詞或語句的符

號，可能不是真正的英文字詞或語句，它們可能不具有英文的意義。然而，當進入到語意的問題，我們便遇到困難。

第二個問題正是屬於語意的層次。戴氏認為：由於提問者與其對象是相互隔離的，因此前者是沒法觀察到後者，也無法觀察到後者與世界如何互動；在測試所提供的環境當中，提問者無法決定其所接收的文字符號與世界中的對象或事件到底有怎樣的關係。一般而言，語意的關係便是語言與世界的關係，由於在圖靈測試中提問者無法觀察其對象 (包括其語言行動)，以及其與世界之關係，因此，提問者對其對象之語意學「是毫無線索的」(Davidson, 2004b: 82)。如果我們無法決定此處文字符號的語意，自然也就無法回答第二個問題。於此可見戴氏對於圖靈測試的質疑是相當激進的，他認為：就測試所設定的場景來說，提問者不僅無法決定對象輸入的語意內容是什麼，甚至根本不能決定從對象接收到的是不是語言！[8]

上述的困難可以從另一種方式來說明。假如提問者認定從測試對象而來的文字符號便是提問者一己所慣用的英文，從而認定自己是跟測試對象作語言的溝通，其對象也是在從事種種的語言行動，例如斷言、提問，甚至感嘆等等，那麼，提問者是如何決定其對象所使用的溝通語言之語意內容呢？簡言之，提問者是如何詮釋此語言的呢？要進行如此的詮釋，提問者需要假定其測試對象是具有信念 (思想) 的。舉例來說，當提問者從其對象接受到 “Trump is the current president of the United States” 的語句的時候，如果他／她

[8] 於此我們可以理解為何戴氏認為：在圖靈測試當中，我們難以將「秉持語句 (P) 為真」之態度歸屬給對象，因為若如此則已經假定對象所使用的符號乃是有真值的語句；但在測試所處的場景中，我們沒有立場作如此的假定。然而，正如下文所示，真正的問題不在上述態度之歸屬，而在測試中的詮釋者、被詮釋者與世界之間沒有形成恰當的關係。

認定這語句便是其自身所慣用的英文語句，從而認定其對象是在斷言川普是現任總統，那麼，這必須預設對象也已經是如此地相信的，因為相信是斷言之必要條件。然則在圖靈測試的場景，提問者是否容許作這樣的預設？於此，我們似乎面臨一個兩難：一方面，如果不容許提問者預設對象有信念或其他的心理狀態，或者如果提問者對其對象之信念等狀態的心理內容茫無所知，則他／她便沒有立場認定從對象來的文字便是其慣用的英文。另一方面，如果容許提問者作上述的預設，乍看之下，這好像是犯了僭置論點之謬誤，因為畢竟提問者的工作正是要決定對象是否有思想，現在又怎能假設對象是有思想 (信念) 的呢？然而，從詮釋的過程來說，提問者作上述的預設並非不合法，[9] 因為提問者可以先行假定對象是有信念的、是正在使用語句來表達其態度或思想的，然後進行詮釋；如果在這樣的假設下進行的詮釋能滿足種種證據，則提問者自然可以合理地把思想歸屬給對象，並理解其思想內容。不過，根據戴氏的觀點，即使我們容許提問者可基於上述的假設而進行詮釋，如果提問者只能與測試對象有雙方的往來問答或符號交流，而不能觀察其測試對象以及該對象與世界對象之互動或因果關係，則提問者接受來自測試對象的輸入 (符號)，便無法與世界對象產生可觀察的關連，從而無法具有詮釋證據之作用，因此也就無法建立測試對象所使用之符號的語意客觀性，亦即無法決定這些符號之意義或所指，縱使它們表面看起來與提問者自身所使用的語言 (英文) 是很相似的。

如果我們接受戴氏上述激進的質疑，承認在圖靈測試中對象輸入的文字與世界之語意關係是沒法建立起來的，則我們便難以有任何的證據，讓我們能夠以這樣的測試方式來判定機器是否具有思

9 感謝方萬全先生向我指出此點。

想。然則我們是否可以透過電腦程式的設計，讓字詞與事物的關係建立起來呢？戴氏並沒有否定這樣可能性；然而，他認為：在此情形，提供語意學的不是測試的對象，而是懂英文的程式設計者。提問者固然可以把測試對象所給予的文字當作是有意義的，但這絲毫不表示此對象有表達任何意思 (Davidson, 2004b: 82)。這些文字的意義不是起因於測試的對象，而是來自電腦程式的設計者。因此，這些被認為帶有意義的文字，不能當作是對象正在思想或具備思想的證據。戴氏認為：即使對於程式設計者來說，程式中的符號表徵了某對象或事實，但這些符號並不能因此就自動地被詮釋為：對那些實現該程式的工具 (試測的對象或機器) 來說，它們也表徵了同樣的對象或事實 (Davidson, 2004a: 91)。因此我們還是沒有理由把測試的對象當作是在思想。

初步來看，戴氏上述的批評似乎是可以回應的，因為我們似乎可以把電腦的程式設計者的工作當作類似是教導電腦學習語言，最後電腦似乎可以在一定的程度上跟我們對話，就好像我們教導小孩學習語言那樣，最後小孩也可以跟我們自然地對話。然而，從戴氏的觀點來看，在圖靈測試所設定的場景中，這兩種所謂的「學習」是截然不同的。在電腦的情況中，其所使用的語言之語意，完全起因於程式設計者，而並非起因於電腦與世界之間的互動，因為此中的電腦與其語言所談論的世界對象並沒有任何互動的關係。但是在教導小孩學習語言的情況則不同，小孩不僅接收了教導者傳遞的訊息 (例如「這是蘋果」這樣的語句)，而且往往同時也與世界中的對象 (例如蘋果) 有因果的關連 (比如看見蘋果、伸手拿蘋果等)，換言之，小孩所習得的語言之語意根源是來自小孩、教導者與世界之三角互動關係。圖靈測試中的電腦之所以被戴氏認為是沒有思想的，是因為它的輸出是缺乏語意的，而之所以如此，正是因為它只

與測試者有符號之間的雙向往來，卻欠缺與世界有恰當的因果關連，而且測試者無從觀察電腦與世界之互動，沒有管道建立電腦輸出的符號之語意學。

肆、庫辛思基對戴維森之批評

討論圖靈測試的哲學家不算罕見，但戴維森對圖靈的論述卻為大家所忽視，庫辛思基是少數對此有所論述的，可是他卻是主張戴維森對圖靈的批評是不成功的。他之所以如此主張，是因為他認為：根據戴氏的觀點，圖靈測試的機器之所以沒有思想，是因為它與其所處的環境並沒有恰當的因果關係。庫氏試圖論證：戴氏的觀點是錯誤的，從而試圖由此而導出「圖靈測試的機器是沒有思想」之結論也是錯誤的。但是我認為庫氏對戴維森的批評是不成立的。在進行相關的討論之前，讓我們先來展示庫氏之主張所依賴的論據。

首先，庫氏把戴維森看作與帕特南 (Hilary Putnam) 同樣秉持這樣的主張：如果 S 擁有關於對象 O 的概念，或者 S 能對 O 有所思考，則 S 必須與 O 有因果的關連。庫氏訴諸帕特南著名的孿生地球之思想實驗。[10] 庫氏假設孿生地球上的約翰 J2 是地球上的約翰 J1 之完整的複本，而且地球上的巴布 B1 與孿生地球上的巴布 B2 也完全相似，後者也是前者的完整複本，但他們處在不同世界：與 J1 互動的是 B1，與 J2 互動的是 B2。因此，縱使 J1 想到 B1 的任何性質，J2 也想到 B2 相同或類似的性質。然而，由於分別跟 J1 與 J2 互動的對象是不同的，亦即與 J1 有因果關連的是 B1，而與 J2 有因果關連的是 B2，因此，J1 對 B1 的概念與 J2 對 B2 的概念乃是不同的概念，即使他們想到 B1 與 B2 時，都是想到相同的或類似的性質。

10 關於這個思想實驗，可參看 Putnam (1975)。

接著，庫氏把孿生地球的情境改為圖靈測試的情境，此中被測試的電腦是機器人約翰 (Robo-John) J3，J3 是實驗室的產物，其心理狀態是人工所造的，而不是來自他與環境之互動，因此根據前述戴維森-帕特南式的主張，他不會擁有像我們所擁有的「柏拉圖」、「水」等概念。庫氏主張：由於環境沒有提供 J3 從事思考所需要的概念，因此，戴氏可據此下結論：J3 是沒有思考的 (Kuczynski, 2005: 114-115)。

庫氏同意戴維森有關圖靈測試的結論，但不同意其論證。庫氏批評之要點有三：

首先，依照戴維森-帕特南式的主張，我們只能說 J3 沒有我們所秉持的那些概念，或者說 J3 沒有像我們那樣的思想，但不能說 J3 沒有思想。庫氏甚至認為桶中腦也是有思想的，同樣具有心靈狀態或認知狀態 (Kuczynski, 2005: 115)。之所以如此，我想這是因為他將 J3 的處境與桶中腦之處境相類比，認為兩者都與其所處的環境沒有恰當的因果關連。

其次，並不是所有的概念內容，都必須取決於概念擁有者與外在對象之因果關連。庫氏同意 J3 是沒有我們的「柏拉圖」、「水」等概念，但是對於其他的概念，如「對象在空間之移動」、「感覺性」(sentience) 等概念，J3 (或者桶中腦) 還是可以擁有的，[11] 之所

[11] 庫氏甚至提到「有感覺的生物的腿」之概念 (Kuczynski 2005: 118)。他認為這是功能性的概念，所以不必依賴概念擁有者與世界之因果關連，因此 J3 或桶中腦的這個概念跟我們的這個概念是一樣的。然而，我認為把這個概念當作戴維森-帕特南式的主張之反例，對庫氏是不利的，畢竟對我們而言，這個概念是指涉到外在對象 (亦即生物)，而在桶中腦的場景中，這樣的對象是不存在的，因此，根據戴維森-帕特南式的主張，如果它擁有這個概念，也是跟我們的概念有所不同的。所以這個概念不是一個恰當的反例。

以如此，是因為庫氏認為這些概念並沒有牽涉到對象 (object-involving)。因此在圖靈測試中，J3 未嘗不可擁有涉及這類概念之思想。簡言之，J3 只是無法擁有一些我們的概念，但它仍然可以擁有思想，而且還可以部分地擁有跟我們一樣的思想 (Kuczynski, 2005: 117-118)。

最後庫氏觀察到：當我們掌握到一個牽涉對象的命題的時候，我們似乎同時也會掌握到一些存在命題。例如：當我們觀察到一隻貓 (湯姆) 在追一隻老鼠，我們固然掌握到命題「湯姆在追一隻老鼠」，但我們同時也掌握到「唯一存在著一個 x，它是如此這般的 (描述「湯姆的種種樣子」之述詞) 而且它在追一隻老鼠。」[12] 我們可以把前一命題改寫為後一命題，庫氏把這樣的改寫，稱為前一命題之被存在化 (existentialized)。原先牽涉到對象的命題經過完全的被存在化之後，便可以成為一個不牽涉對象的命題。圖靈測試的對象 J3 可以對這種不牽涉對象的命題有所操弄，因此，如果我們接受傳統對智能 (intelligence) 之看法 (庫氏似乎便是如此)，認為智能乃是一種操作命題之能力，(propositional-manipulating ability)，[13] 則我們便可說 J3 可以從事智能的活動。以此，庫氏可以主張圖靈測試的對象是可以從事思想活動的，所以戴氏的主張是錯誤的 (Kuczynski, 2005: 120-121)。[14]

---

[12] 明顯地，庫氏在此是採取羅素處理一般專名 (proper name) 的方式，將一個專名當作是一個確定描述詞 (definite description)，從而把一個包含專名的命題改寫為包含確定描述詞的命題，再將後者改寫為存在命題。不過文中他提到的理由似乎是知識論的，亦即當我們看到湯姆的時候，我們也一定將之看作是 (seeing it as) 具有某些性質，例如有四條腿、有毛等等 (Kuczynski, 2005: 120)。

[13] 我把「操作命題之能力」理解為一種廣義的推理的能力。

[14] 此處對於第三點的批評之陳述，與庫氏原文並不完全相同。他的陳述使得其論點有循環之嫌疑；我的陳述則是依於原文脈絡反戴維森之意圖，對文本予以重構，使之初步看來是一個對戴氏有意義的批評。

伍、對庫辛思基之反駁

我認為庫氏的批評是可以商榷的。茲分述如下：

第一、庫氏一開始便把戴維森的語意觀與帕特南的語意觀視為相同的；在其討論的過程當中，也把圖靈測試的對象 J3 與桶中腦視為處於類似的情況。對於帕特南來說，桶中腦所使用的語言並不是沒有意義的，只是其意義與我們使用的語言之意義有所不同而已。[15] 因此，庫氏自然會認為 J3 使用的語言之意義即使與我們所使用的不同，但還是有意義的，從而 J3 還是會思想的，只是其思想內容與我們不同而已。然而，將戴維森的語意觀與帕特南的語意觀等同起來，卻是錯誤的，忽略了兩者差異之處，至少是把戴氏的理由過份簡單化了。戴氏固然曾強調：在基礎的情況，例如在明示學習 (ostensive learning) 或簡單的知覺之情況下，思想與其對象之間的因果關連往往與語言之意義或思想內容之決定有密切關係 (Davidson, 2001a: 196-197)；但是僅僅只有思想 (或思想者) 與對象之因果關係，並不足以決定思想之內容，因為與思想相關的因果關係是相當複雜的。就以戴維森自己的例子來說，一頭獅子出現，從而使我相信獅子出現，但是使我有此信念的還有其他的原因，例如我的視網膜或視神經受刺激等，然則我的信念是否也是關於這等刺激的呢？假設僅僅與思想發生因果關係者便足以決定思想的內容，則「這樣的因果解釋有無窮的多，而且每一個都會把不同的內容授與同一個知覺信念」(Davidson, 2004c: 142)。這樣的結果自是難以接受的，從而顯示前述有關因果與思想內容之假設也是難以接受的。戴氏並

---

[15] 依帕特南，桶中腦所使用的「樹」、「腦」等字詞未嘗不可有所指涉，但所指者跟我們所指的不同 (Putnam, 1981: 14)。此書第一章曾援引圖靈測試來類比桶中腦的情況，以說明此兩種情況中有關語詞指涉之情形頗為相似。或許由於這個緣故，庫氏自然地將帕特南的桶中腦與戴維森所談的圖靈測試相提並論。

不反對我們的思想可以透過因果關係而與世界連接，但是如果僅僅只憑這種關係，則我們不只決定不了思想的內容，我們甚至難以對意向性有所解釋。「沒有一個有關心靈狀態與世界之因果關係的簡單故事，可以說明意向性，更不要說對思想與說話 (utterances) 之意向內容作仔細的說明了」(Davidson, 2004c: 138)。

庫氏沒有瞭解到：戴維森之所以主張 J3 沒有思想，並不僅僅是因為 J3 被隔離而與世界沒有恰當的因果關連，而且還是因為測試者——作為一名 J3 所使用的語言之詮釋者——無法觀察 J3 與世界的互動，從而無法對 J3 所使用的語言 (符號) 予以詮釋。然而，對於戴氏來說，「可被詮釋」至少是語言的必要條件，也可說是符號要具有意義之必要條件，但詮釋活動必須涉及詮釋者、被詮釋對象以及兩者共同所處的世界 (外在對象) 之三角互動關係。在圖靈測試設定的場景中，由於詮釋者無法觀察到被詮釋者與世界之互動，使 J3 所使用 (輸出) 的符號之客觀內容無法被詮釋，所以我們不能說這些符號是有意義的，從而也不能說 J3 是有思想的。類似地，如果依照桶中腦所假設的場景，我們無法對其所謂的「思想活動」進行詮釋，則我們從根本來說便不應認為它是有思想的。於此我們可以看到帕特南與戴維森的語意理論不同之處。庫氏把戴氏的語意理論等同於帕氏的語意理論，並以後者為基礎對戴氏提出批評；但這兩種理論雖然往往都被歸為外在論，其實並不等同，因此他的批評是建立在對戴維森錯誤的理解之上的，至少他沒有掌握到戴氏批評圖靈測試的核心理由。

第二、戴維森不必否認我們有些思想並不牽涉到外在對象，也許有關邏輯的思想，便是明顯的例子；因此如果僅僅說 J3 可以擁有這類的思想，這不一定是戴氏會反對的。問題是在於：如果我們沒有任何牽涉到對象的思想，我們能否擁有這些不牽涉對象的思想？

換一個方式來問：我們 (作為人類) 能否只擁有不牽涉外在對象的思想？我們能否不牽涉任何外在對象而仍然有思想？我認為從戴氏的立場來看，這些問題的回答應該是否定的；此中的關鍵即在於思想 (或語言) 之可詮釋性。如果一個主體的思想完全與外在對象無涉，這會使得我們無法對之作詮釋，因為詮釋活動必須牽涉到詮釋者、被詮釋者以及兩者與外在世界之互動關係。如果主體的思想與世界的對象毫無交涉，則對人類來說，詮釋活動便難以進行，從而其思想會變得無法被詮釋；這樣的結果會讓我們傾向認為該主體可能根本是毫無思想可言的。以庫氏所提到的「對象在空間之移動」概念來說，如果我們沒有任何牽涉移動對象的思想，則我們又如何能夠有這個概念？又這個概念已經包含了「對象」的概念，如果我們沒有任何關於對象之具體經驗，我們會不會有「對象」這個抽象的概念，是相當值得懷疑的。[16] 讓我們回到 J3 的情形。如果 J3 具有牽涉對象的思想，則戴氏未嘗不可接受它也具有一些不牽涉對象的思想；但依照圖靈測試的場景，根據前述的理由，戴氏會認為：由於 J3 與世界的互動是跟我們完全隔絕的，我們無法對其語意 (如若有的話) 作詮釋，因此無法將思想歸屬給它。

第三、庫氏主張透過命題之存在化，我們可以把牽涉對象的命題轉化為不牽涉對象的命題。庫氏這樣做是已經假定：測試者瞭解他所面對的是一套可以用一階語言去詮釋與處理的符號系統。但是這假定之理據何在？我們是根據什麼來作如此的詮釋？其次，命題存在化的過程只是一個消除專名的過程，亦即把一個含有專名的命

16 我並不認為：對於所有的概念來說，我們必須先認識其個例 (instance)，才能掌握這個概念。至少我們可創造一些沒有個例的複合概念。不過就「對象在空間之移動」這個概念來說，我確是認為：如果沒有關於移動對象之經驗，我們是無法掌握或理解這個概念的。

題轉化為一個不包含專名的命題。然而，後者要被視作不牽涉對象，則必須先行假定這個包含述詞的存在命題並不牽涉對象，而這樣的假定是否成立？答案卻不是很明顯，至少庫氏對此並沒有說明。一般而言，述詞邏輯是假定論宇 (domain of discourse) 是非空的個體集合，因此，將一命題存在化似乎並不一定意味著與對象無涉。當然，此中涉及到許多的議題，例如：我們是否可以使用與述詞邏輯不同的邏輯 (如自由邏輯)？邏輯上的個體與一般的對象是不是相同的東西？在所謂「牽涉對象的命題」中，此中牽涉的方式為何？在庫氏對這類問題還沒有釐清之前，我們很難斷言存在化的命題到底是或不是牽涉對象。再者，「命題」其實是一個有爭議的概念。我們暫且不管有關命題之存有論地位的問題，一般而言，人們大抵同意命題乃是直述句所表達的意義，然而，在圖靈測試中，我們不能假定 J3 所操作的乃是命題而不僅是一串符號而已，因為這些符號是否有意義以及其意義為何，正是圖靈測試所試圖要決定的；或者退一步說，即使我們作上述的假定，但是在此測試的場景中，依前所述，我們也無法詮釋這些符號之客觀意義。故無論如何，如果我們只能確定 J3 操作是一連串的符號，則這並不足以支持「J3 是有思想」之結論。這一點正是體現中文房論證之要旨——語法不足以決定語意。

綜合以上所述，我認為庫辛思基對戴維森的批評是不成立的。

陸、對圖靈測試之修訂

以下我們把焦點放在戴氏如何看待思想之條件：要在怎樣的條件之下，一個個體（包括圖靈測試中的機器）可以被看作是有思想

的？或者換另一種方式說，如果依照戴氏的說法，我們要如何修訂圖靈測試，才可讓我們藉此來判定機器是否具有思想？

戴維森並不懷疑我們至少原則上是可以製造出會思想的機器。畢竟人是具有物理性質的對象，其具有的種種功能也要遵守物理法則，所以我們沒有理由說無法設計並創造一個在各方面跟自然人一樣的對象，這個對象也可以有種種的心靈生活，例如從事思考、推理、擁有各種欲望、信念、意圖等等。[17] 戴氏關心的不是我們是否在技術上可以製造出思想的機器，而是：一個機器 (或電腦) M 要滿足怎樣的條件才可說是有思想的？

初步的答案是：M 要跟我們人類相似。問題是：在哪一方面相似？戴氏的回答是從兩方面入手，一方面是問：拿掉人的哪些方面，人還是會思想的？這也可說是在問：機器可以在哪些方面跟人不同而仍可以有思想？另一方面，我們要給機器添加什麼，以使之有思想？戴氏稱這雙向的探究方法為「加減法」(the method of addition and detachment) (Davidson, 2004a: 87)。

戴維森先從減法開始。

(A) 首先考慮「來源與構成成份」是否重要？戴氏的答案是否定的。人是屬於自然類的東西，而機器不是；然而，機器如何產生與其是否有思想是沒有關係的，一個機器能夠以相異於人的方式被製造，甚至其構造成份與人不同，也無礙於其具有思想。如前所述，戴氏承認即使他發現其朋友原來是由一顆蛋所孵化而來的，或者是由矽所組成的，他依然會認為這位友人有思想的。

(B) 其次是考慮「大小與形狀」，或者更廣泛地說是考慮「外貌」。一個對象是否要看起來像人，才會有思想呢？圖靈認為這對

[17] 參看 Davidson (2004a: 87, 2004c: 136)。

於判定對象是否能思想是沒關係的，我們從其對測試場景之設計即可知。但戴氏對此卻稍為有所保留，他說：「但大小與形狀可以是有關係的，如若它們足以妨礙到對信念——乃至於對感覺與意圖——之溝通能力的話」(Davidson, 2004a: 88)。[18] 換言之，機器之大小形狀不能妨礙溝通，因為如果無法進行溝通，便無法進行詮釋，從而無法作思想歸屬。戴氏斬釘截鐵地說：「我們不能詮釋的，都不是思想」(88)。

(C) 最後戴氏考慮的是「歷史」，他認為思想不可無歷史。此處所謂的「歷史」，不僅是指：人的思想之學習過程，而且也指：在這個過程當中，人與世界的對象之一連串的因果連結。如果人腦沒有與日常的對象有因果互動的歷史，則它便不可能擁有關於這類日常對象的日常思想。[19] 如果參照先前我們對庫氏的第一點商榷，我們可以說：與世界中互動之因果歷史是日常思想要具有內容之必要但非充分的條件。機器要有思想便必須有學習的能力，而且能從與世界之因果互動中學習。

綜上所述，關於機器 M 是否有思想的問題，M 的來源與構成成份是不重要的，至於其大小與形狀，在不妨礙溝通的情況下，也是不重要的，但是 M 是否有學習的能力，是否能與世界的對象有因果的互動，卻是重要的，它是不可或缺的。

現在轉到加法，也就是要面對這樣的問題：在怎樣的情形之下，我們才可以說機器 M 是有思想的 (或能夠思想)？首先，M 只能做好單一的一種工作是不夠的。其次，如果 M 要擁有信念，則這些信念必須存在一個豐富的概念系統當中，也就是說，信念與其他的信

[18] 原文是虛擬條件句，我們以「如若…則…」表示之。

[19] 同上。要注意的是，戴氏在這裡談的只是關於日常對象之日常思想，至於高度抽象的思想，則是另一回事。

念是相互關連的。[20] 然而，我們如何能辨認出 M 是具有信念 (或者其他的心靈狀態) 呢？「雖然我們的信念、意向、恐懼，以及其他的感覺都是私有的與主觀的——如果有任何東西是私有與主觀的話——但除了透過一開始就將它們與外在的對象或事件緊密連結之外，它們是無法被辨認或被解釋的」(Davidson, 2004a: 96)。問題是：信念既然是私有與主觀的，我們又如何得知它們與外在對象或事件的關連呢？此中的關鍵在思想與語言之密切關係。擁有概念與擁有命題態度是一體之兩面，戴氏說：「擁有概念與擁有命題態度並不存在任何分別。擁有概念就是將事物分類至其下，……就是**判斷**或**相信**某些物品 (items) 隸屬於該概念之下」(Davidson, 2004c: 137，粗體為原文強調處，原文採斜體)。判斷是一種典型的語言行動，這行動同時也顯示判斷者之命題態度。「語言的結構反映 (mirrors) 命題性的思想之結構」(139)。因此，要辨認 M 的命題態度，可以從其語言活動與外在對象或事件之關係入手。

然而，先前已經指出：圖靈測試是無法掌握對象之語意學的 (如果有的話)，這是因為在此測試所設定的特殊條件之下，提問者無法觀察其測試對象以及該對象與世界之互動。這便暗示著修訂此測試的方向。對於如何修訂圖靈測試，戴氏建議：「(被測試的) 對象必須被公開，讓它與其餘世界事物之因果關係，以及它與提問者之因果關係，能夠被提問者所觀察到」(Davidson, 2004b: 84)。而且提問者要有足夠的時間對其對象來提問與觀察，讓提問者掌握到：對象所使用的語意學，跟提問者一己語言 (如英文) 的語意學是相似的，而對象之語言傾向跟提問者的語言傾向也是相似的。此外，至少在圖靈測試的情況中，我們還要知道測試對象之語意學與語言傾向是

[20] 這些關連可以是概念的，也可以是邏輯的。參看 Davidson (2001f: 124)。

如何習得的。最後一點之所以不可忽略，是因為戴氏可以想像以下的情形：如果我們提供給電腦相關的程式以及感受器，則在它被問到有關狗的問題，而狗真的在附近的時候，他可以像我們那樣，作出「這是一條狗」之類的反應。然而戴氏堅稱：即使如此，這也不足以證成電腦知道任何有關狗的事情，甚至當它產生「這是一條狗」這樣的語句的時候，它也沒有表達任何的意思 (Davidson, 2004b: 85)。戴氏之所以如此地堅持，是因為他認為：如果一台電腦從未擁有過有關狗的經驗，也沒有任何有關狗的記憶，那麼，即使它能說出「狗」這樣的字詞，也無法表達 (我們所謂的) 「狗」的意思，也不能談到狗。簡言之，我們沒有理由說它表達任何的意思。戴氏說：「思想與意義要求某種特殊的歷史」(85)。這樣的要求似乎蘊涵以下的結果：即便我們能發展一種技術，將某個自然人 (甲) 的思想複製並轉移到另一個對象 (乙) (不管後者是機器或複製人)，我們也不能說乙本身是有思想的，至少他本身不會有甲的思想，因為他的學習歷史與甲完全相異。[21] 如此一來，複製有思想的人能否成功，便不能只看複製物之結果來決定，還要考慮複製物是如何習得其思想的。[22]

無論如何，上述對測試的修訂可說是把原來圖靈測試中受測對象 (M) 與測試者 (I) 的二元關係，改變為受測試對象 (M)、外在對象 (O) 以及測試者 (I) 的三角關係。此中 I 必須有足夠的時間觀察 M 與 O，以及此兩者之互動關係，藉此掌握 M 之語意學與語意

21 這並不妨礙乙可以被看作是甲的思想之承載工具，畢竟承載不等於擁有。此外，如果其他條件滿足，也不必排除在往後的過程中，乙可以透過與他人及世界對象之互動而習得新的思想。

22 戴氏沒有進一步說明為何思想與意義要求特殊的歷史，不過我想這與思想與意義之社會性有關係，這種社會性是來自語言的社會性；依戴氏，思想與意義都是離不開語言，而語言是一種社會的技藝，其本質便是社會性的。

傾向，以及其語言習得之歷史。這樣的刻劃是著重在 I 需要滿足哪些條件，以使 I 可以將思想歸屬給 M。然而，這仍然是不夠周全的，戴氏指出我們還需要對 M 有所要求：如果 M 要有命題態度或概念，則 M 必須能夠認知到自己是可能犯錯的。如果 M 擁有某個概念，則 M 要知道或至少這樣想：此概念之是否適用於某事物，跟他是否相信該概念適用於該事物，可以是兩回事；這意味著他會想自己是有可能犯錯的。如果不能這樣想，M 便不會有概念。以此而言，M 如果要有思想，便需要有「客觀真理」(objective truth) 之觀念 (Davidson, 2004c: 141)。[23] 再者，M 既然能想到概念之應用與一己所相信的是沒有關係的，則它也必須要有「信念」之概念；簡言之，擁有信念 (或思想) 的條件是擁有「信念」概念 (2001b: 102)。[24] 更進一步來說，如果 M 所給出的一串符號有意義的語言 (例如英文)，而且 M 是從事言說或溝通之活動，則這些符號不能只是與 I 的語言 (英文) 相似而已；作為說話者，M 必須有「對聽者產生某種影響」之意圖。姑不論這種葛萊斯式的 (Gricean) 觀念是否可取，戴氏至少同意：如果溝通是成功的，則在說話者身上必須有這些意圖 (2001c: 112)。他甚至聲稱：「一個生物不能擁有思想，除非他是另一個人之言說的詮釋者」(1984d: 157)。因此，作為思想者，不僅其思想必須能夠被詮釋，他自身也必須具有意圖，甚至必須有能力作為他人言說之詮釋者。

---

23 原文之脈絡顯示該處談的是生物 (creature) 擁有概念之情形，雖則 "creature" 一詞也有「受造物」之意。惟在有關「思想之條件」之論題上，我認為此處的義理可以應用到有關機器這種人造物之論述。

24 我們還能以同樣的理由進一步說：擁有信念 (或思想) 的條件是擁有「自我」(self) 之概念。戴氏主張自我概念是不可化約，亦即此概念，如同「真理」概念那樣，無法以其他與之同樣普遍或至少同樣清晰的概念來定義而不陷入循環 (Davidson, 2001d: 85)。

如果我們將上述修訂後的圖靈測試稱為戴維森式的測試，則明顯地，一般專精作一件事情的專家系統是沒有思想的，僅僅會下棋的電腦系統亦然。關於後者，戴氏之所以拒絕它是會思想的，理由是：下棋必須要有想贏的欲望，至少要有「贏」的一般性的概念，進而要有「規則」或「約定」之概念；下棋是一種具有意圖的活動，而且必須具有信念，相信如此的行動可以帶來欲望之滿足或目的之達成，而且如果一直是輸棋，則會慢慢覺得無趣。欲望與信念一樣，它不會是孤立存在，如果有一個欲望，則必然還有許多其他的欲望 (Davidson, 2004a: 89)。以此觀之，戴氏自可同意人工智慧 AlphaGo 是很會計算，但它最多只是一個幫忙人們作計算的機器或工具，它並不擁有命題態度，也沒有任何思想。

然而，如果我們把問題轉向機器人呢？假設機器人 R 是人類設計的，我們可觀察 R 與世界看似有意義之互動，它說的話是我們可理解的，它也可以跟我們流暢地對話，甚至可以創造新的對話主題。簡單來說，R 不只是一個活動的聊天機器人，而且還可以與世界的對象互動，這些互動是我們可觀察得到的。然而，如果只是如此，根據戴氏的測試方式，我們還不能認為 R 是有思想的，我們還必須看 R 學習的過程。如果 R 的語言 (語意學) 是透過程式的設計而先行規定的，則依戴氏，R 的言說對我們是有意義的，但對 R 則不然，因此，R 是否有思想，還是可商榷的。然而，假使我們透過特定的演算法，讓 R 與世界 (當然還有人類) 互動而習得並發展其語言，讓 R 與人之學習更為相似，則我們是否願意將思想歸給 R 呢？依照戴氏的測試，R 與人類越相似，我們將思想歸屬給 R 之傾向越強烈。但此中問題在於：無論外在行為有多麼相似，都不足以支持 R 是具有思想的，即使語言行為之相似也不行，因為這樣的行為最多只能說是合乎我們的意義，但對 R 來說卻不一定是意義的。如前所示，

思想之歸屬，不能只從詮釋者的角度來看，我們還必須對被詮釋者有這樣的要求：他必須具有意圖，具有信念，乃至具有信念之概念等。如果我們不願意將這些意向狀態歸給 R，或者我們不願意先行假定 R 具有這種種的狀態，則依照戴維森的觀點，R 無論在行為 (包括語言行為以及與世界之互動) 與人類的行為多麼相似，它仍然是不具有思想。於此，我們似乎面臨一個困局：如果我們不將思想者的身份歸給機器，或者先假定機器是有意向狀態，則無論機器之外顯行為與有思想者多麼相似，我們都無法將思想歸給它。

柒、徹底詮釋與對機器之思想歸屬

我們不妨將機器是否有思想之測試，與徹底詮釋作比較。[25] 在後者的場景中，詮釋者對被詮釋者的信念與語言同時是無知的，這使得詮釋的工作似乎難以著手。[26] 這情形與對 R 之測試有點類似。然而，雖然在此場景中詮釋者不知道被詮釋者的信念為何，但詮釋者卻是假定被詮釋者是有思想 (信念) 的——只不過我們不知道其內容為何，而且也假定被詮釋者之言說是有意義的——只不過我們不知其意義為何；簡言之，詮釋者是假定被詮釋者跟我們一樣，也是一名思想者，只是我們不知道其思想內容而已。因此，縱使詮釋者不知道其詮釋對象之語言的意義與其所秉持為真者為何，但在一些徹底詮釋之場景中，詮釋者可以假定 (1) 被詮釋者 (作為思想者) 是在說話，且其所說的乃是關於其外在環境的對象；(2) 被詮釋者是秉持其所說的 (語句) 為真，亦即對其所說是採取肯認的態度。

---

[25] 有關徹底詮釋，可參看 Davidson (1984a, 1984c)。

[26] 如若詮釋者知道被詮釋者的信念內容 (或其說話之意義)，便不難以此決定其說話之意義 (或信念之內容)。徹底詮釋之困難就在於詮釋者對兩者同時是無知的。

雖然詮釋者並不確定說話者所肯認的內容為何，但詮釋者以此作為出發點，透過與被詮釋者 (說話者) 之互動，觀察其與外在對象之關連，經過對初步詮釋之假定作種種必要的修訂，遂可望逐步掌握被詮釋者的語意學以及其思想內容。然而，有關思想機器之測試則不同，機器不是自然類 (natural kind)，我們不會——或至少目前還不會——假定它與我們一樣也是思想者。戴氏對圖靈測試之反思給予我們兩點啟示：一方面，我們無法僅僅從外顯行為來決定機器能否思想，思想之條件比圖靈所預設來得複雜得多，因此圖靈的模仿遊戲之進路是不可取的；另一方面，機器能否被視為思想者，主要不是取決與機器與人類在外顯行為有多相似，而是取決於人類有多大的意願先行把機器看作與自身 (作為思想者) 相似。一旦人類作為詮釋者，決定把機器看作思想者，啟動對機器思想之詮釋，則上段所述的看似的困局便可打破。[27] 但詮釋的結果如何，則要看詮釋者、被詮釋對象以及世界三者之互動協調而定：也許最後我們對機器思想有所理解，可以恰當地將思想歸給它；也許我們最終發現機器並不滿足思想者作為思想者之條件，從而獲致「機器不是思想者」之結論 (這相當於歸謬論證)，也就是說，我們無法對它作思想歸屬。

然而，即使我們可以合理地把思想歸給機器，這最多只是表示我們能把機器「視為」思想者，但這能否回答「機器是不是思想者？」

27 戴氏的說法顯示：這樣的決定是社會性的，它必須涉及詮釋者、被詮釋者 (機器) 與世界之三角互動。然而，我們或許可進一步說，它也必須擴及眾多詮釋者之多角互動，因此，社會、法律等層次也應當納入考慮。此外，我們對機器之情感也可能要納入考慮，但戴氏對此所言不多。在考慮機器是否會思想之問題的時候，有關情感或意識的問題或許可暫不考慮 (這點我其實不是很確定，我懷疑：在考慮認同一對象是否為一個會思想的對象的時候，我們對該對象抱持怎樣的情感，可能會對如何對待認同問題有所影響)，但如果我們要考慮機器是否有人格或道德地位等問題的時候，則有關意識或情感之問題是難以避免的。

的問題，似乎是可商榷的。對於這樣的質疑，以下是嘗試站在戴維森的立場予以回應：[28]

我認為上述的質疑不是沒有道理的。讓我們區分以下兩個問題：(A) 在什麼條件之下，我們可以合理地將某特性歸給一對象 S，亦即可以合理地斷言、判斷或相信「S 具有某特性」？這問題所關心的是「特性歸屬」之條件；(B) 在什麼條件之下，S 具有某特性？這問題所關心的是「特性存在」之條件。一般而言，回答了前一問題，往往不會同時也回答了後一問題，因為滿足特性歸屬之條件，並不一定同時滿足特性存在之條件。例如，我看到 S 在走路，這使我可以合理地把「在走路」這樣的性質歸給 S，但是，「我看到 S 在走路」並不是「S 在走路」成立之必要條件，因為以下的可能性是存在的：雖然我沒有看到 S 在走路，但 S 確實在走路。因此，上述的質疑相當有道理的：即使某些條件獲得滿足，能讓我們把思想合理地歸屬給機器，但這是否便足以回答「機器能否思想？」的問題，似乎仍是可商榷的。

我想戴維森可以承認：一般而言，特性歸屬條件之滿足並不蘊涵特性存在條件之滿足。然而，有趣的是，就戴維森的立場來說，如果我們討論的特性是「具有思想」或「能夠思想」，則情形便有所不同。對於戴維森來說，思想必須具有可詮釋性，亦即如果 S 具有思想，則其思想內容必須可以透過詮釋而被理解，此中的詮釋活動涉及一些對詮釋者與被詮釋者的要求，以及兩者與世界之間的互動關係；而終究來說，這樣的詮釋活動之結果，乃是將特定內容歸屬給被詮釋者之言說或信念。另一方面，如果我們試圖將特定內容歸給某對象之言說或信念，則自是意味著該對象乃是言說者或思想

[28] 此處的質疑是來自一位論文審查人，他（或她）希望我們能替戴維森予以回應。

者，我們的歸屬則是對其言說或信念內容進行詮釋之結果。綜言之，思想內容之歸屬，可說是一種詮釋活動；而詮釋也是一種思想歸屬的活動。因此，如果我們能夠合理地或融貫地將思想內容歸給 S，則這樣的歸屬，便已意味著 S 是思想者。對戴維森來說，不存在一個其思想不能被詮釋的思想者。

人們或者會堅持說：即使我們能夠合理地或融貫地將思想內容歸給某一對象，這最多只表示我們預設了該對象是思想者，或者它是被我們視為思想者而已，但仍有可能它實際上並非如此，亦即我們對其所謂的「思想」之詮釋，無論多麼融貫合理，被詮釋的對象有可能根本不是思想者，或者沒有我們對之所詮釋之思想！

然而，根據戴維森的慈善原則 (principle of charity)，如果要進行詮釋活動，則上述那種極端錯誤之可能性必須被排除。粗略地說，慈善原則主張：在詮釋的過程中，被詮釋者必須有足夠的信念是真的，亦即我們必須排除被詮釋者的信念有可能大部分都是錯的。如果不排除這樣的可能性，詮釋便無法進行。「使詮釋可能的，乃是這樣的事實：我們能夠先驗地 (*a priori*) 去除大規模的錯誤 (massive error) 之可能」(Davidson, 1984d: 168-169)。[29] 對於戴維森來說，慈善原則並不是一個我們可以選擇或放棄的詮釋方法之策略，而是一個使得詮釋思想成為可能之條件；如果把詮釋活動看作是一種對他人思想之理解活動，則慈善原則也可說是我們理解他人之條

29 我們此處對戴維森對慈善原則之表述不免粗略，這是因為他表述此原則的版本不只一種：對於被詮釋者之信念，有時說必須是一致的或融貫的，這是談信念之間的邏輯關係；有時說必須大部分是真的，或者從詮釋者的角度來看大部分都是真的，這則是談信念之真假。戴維森後來稱前者為「融貫原則」(Principle of Coherence)，後者為「符應原則」(Principle of Correspondence)，而此兩者都是慈善原則 (Davidson, 2001e: 211)。然而無論是哪一版本的表述，都已經預設被詮釋者是信念擁有者，亦即是一名思想者。

件。以此，我們可以瞭解何以戴維森說：「慈善是強加於我們的；無論我們喜歡與否，如果我們想要理解他者，則我們必須要在大多數的事情上認為他們是對的」(Davidson, 1984b: 197)。反過來說，如若被詮釋者的所有或大多數信念在詮釋者看來都是錯誤的，則後者對前者根本無法理解，甚至連反對的意見都說不上。[30] 當然，這並不意味著被詮釋者不能擁有錯誤的信念，也不意味著詮釋不可能犯錯；只是這些錯誤都有相當的真信念作為背景，使得這些信念或詮釋有錯誤可言。無論如何，依戴維森的慈善原則，以下的情形是不會發生的：我們可以合理地將思想歸給一個對象 (亦即對其思想予以詮釋)——不管此對象是人或機器，而該對象卻是沒有思想或者其思想全都是錯的。[31]

---

30 戴維森舉了一個例子：我們可以反對前人有關「地球是平的」之主張，但至少我們與前人在若干相關的信念上是一樣的，例如大家都相信「我們是住在地球上的」等等；如果我們沒有這些共同的信念，則當前人說「地球是平的」，我們連前人與我們是否在談論同一個對象 (地球) 也不能確定，遑論兩造之信念是否相反 (Davidson, 1984d: 168)。

31 根據上述的反駁，即使我們能合理地將思想歸給被詮釋的對象，仍然有全面錯誤的可能——或者對象的信念全都是錯的，或者對象根本沒有思想。對戴維森來說，這很容易招致懷疑論的結果。因此戴氏以慈善原則來回應此反駁，往往也可被看作是他對懷疑論之回應。由於懷疑論並非本文之主題，我們對這此課題暫不申論，我們僅在此指出一點：上述的反駁似乎預設了思想或思想的內容與其外之世界 (包括其詮釋者) 是沒有關係的，前者的存在並不依賴後者，因此才有反駁者所提到的錯誤之可能性。然而對於戴氏來說，這樣的預設是來自對思想本性之誤解，因為思想或思想之構成，與世界以及其外之詮釋者是不可分的，此中三者的密切的關係，才是思想或思想內容之客觀性之根源。

參考文獻

Davidson, D. (1984a). Belief and the basis of meaning. In *Inquiries into truth and interpretation* (pp. 141-154). Oxford, UK: Clarendon Press.

Davidson, D. (1984b). On the very idea of a conceptual scheme. In *Inquiries into truth and interpretation* (pp. 183-198). Oxford, UK: Clarendon Press.

Davidson, D. (1984c). Radical interpretation. In *Inquiries into truth and interpretation* (pp. 125-139). Oxford, UK: Clarendon Press.

Davidson, D. (1984d). Thought and talk. In *Inquiries into truth and interpretation* (pp. 155-170). Oxford, UK: Clarendon Press.

Davidson, D. (2001a). Epistemology externalized. In *Subjective, intersubjective, objective* (pp. 193-204). Oxford, UK: Clarendon Press.

Davidson, D. (2001b). Rational animals. In *Subjective, intersubjective, objective* (pp. 95-106). Oxford, UK: Clarendon Press.

Davidson, D. (2001c). The second person. In *Subjective, intersubjective, objective* (pp. 107-121). Oxford, UK: Clarendon Press.

Davidson, D. (2001d). The irreducibility of the concept of the self. In *Subjective, intersubjective, objective* (pp. 85-91). Oxford, UK: Clarendon Press.

Davidson, D. (2001e). Three varieties of knowledge. In *Subjective, intersubjective, objective* (pp. 205-220). Oxford, UK: Clarendon Press.

Davidson, D. (2001f). The emergence of thought. In *Subjective, intersubjective, objective* (pp. 123-134). Oxford, UK: Clarendon Press.

Davidson, D. (2004a). Representation and interpretation. In *Problems of rationality* (pp. 87-100). Oxford, UK: Clarendon Press.

Davidson, D. (2004b). Turing's test. In *Problems of rationality* (pp. 77-86). Oxford, UK: Clarendon Press.

Davidson, D. (2004c). What thought requires. In *Problems of rationality* (pp. 135-149). Oxford, UK: Clarendon Press.

Kuczynski, J. M. (2005). Davidson on Turing: Rationality misunderstood? *Principia*, 9, 1/2: 111-124.

Putnam, H. (1975). The meaning of "meaning." *Minnesota Studies in the Philosophy of Science*, 7: 131-193.

Putnam, H. (1981). *Reason, truth and history*. Cambridge, UK: Cambridge University Press.

Searle, J. R. (1980). Minds, brains and programs. *Behavioral and Brain Sciences*, 3, 3: 417-457. https://doi.org/10.1017/S0140525X00005756

Searle, J. R. (2004). *Mind: A brief introduction*. New York: Oxford University Press.

Turing, A. (2004a). Computing machinery and intelligence. In B. J. Copeland (Ed.), *The essential Turing*: *Seminal writings in computing, logic, philosophy, artificial intelligence, and artificial life: Plus the secrets of enigma* (pp. 441-464). Oxford, UK: Clarendon Press.

Turing, A. (2004b). Intelligent machinery. In B. J. Copeland (Ed.), *The essential Turing seminal writings in computing, logic, philosophy, artificial intelligence, and artificial life: Plus the secrets of enigma* (pp. 410-432). Oxford, UK: Clarendon Press.

Davidson on the Turing Test

Chi-Chun Chiu
Graduate Institute of Philosophy, National Tsing Hua University
E-mail: ccchiu@mx.nthu.edu.tw

Abstract

Davidson argues that the well-known Turing Test, being a method of determining whether a machine can think, fails to tell us anything about the semantics of the tested object and thus is inadequate to discover whether can think or not. However, against Davidson, Kuczynski claims that his reasoning is entirely fallacious and has little force in attacking the Turing Test. In this paper I will first delineate and clarify Davidson's comments on Turing's imitation game and his reasons for reject it as a proper test of machine thinking. Second, I will object to Kuczynski's criticisms by showing that his arguments are either ill-founded, inconclusive, or based on his misinterpretation of Davidson's thought. Finally, I will show how Davidson proposes to modify the Test in accordance with his own theory of interpretation, and the significance of this modified version for the attribution of thought to AI.

Key Words: Davidson, Turing Test, thought attribution, radical interpretation, artificial intelligence