一部人工智能进化史

集人工智能领域顶级大牛、思维与机器研究领域
最杰出的哲学家**多年研究之大成**

关于人工智能的**本质和未来**更清晰、简明、切合实际的论述

AI: Its Nature and Future

AI

人工智能的本质与未来

【英】玛格丽特·博登（Margaret A. Boden）/著　　孙诗惠 /译

中国人民大学出版社

目錄

CONTENTS

01 什麼是人工智能

人工智能（Artificial Intellegence，AI）就是讓計算機完成人類心智（mind）能做的各種事情。通常，我們會說有些行為（如推理）是「智能的」，而有些（如視覺）又不是。但是，這些行為都包含能讓人類和動物實現目標的心理技能，比如知覺、聯想、預測、規劃和運動控制。

智能不是一維的，而是結構豐富、層次分明的空間，具備各種信息處理能力。於是，人工智能可以利用多種技術，完成多重任務。

人工智能無處不在。

人工智能的實際應用十分廣泛，如家居、汽車（無人駕駛車）、辦公室、銀行、醫院、天空……互聯網，包括物聯網（連接到小物件、衣服和環境中的快速增多的物理傳感器）。地球以外的地方也有人工智能的影子：送至月球和火星的機器人；在太空軌道上運行的衛星。好萊塢動畫片、電子遊戲、衛星導航系統和谷歌的搜索引擎也都以人工智能技術為基礎。金融家們預測股市波動以及各國政府用來指導制定公共醫療和交通決策的各項系統，也是基於人工智能技術的。還有手機上的應用程序、虛擬現實中的虛擬替身技術，以及為「陪護」機器人建立的各種「試水」情感模型。甚至美術館也使用人工智能技術，如網頁和計算機藝術展覽。當然，它還有一些應用不那麼讓人歡欣鼓舞，如在戰場上穿梭的軍事無人機——但是，謝天謝地，它也用在了機器人掃雷艦上。

人工智能有兩大主要目標：一個是技術層面的，利用計算機完成有益的事情（有時候不用心智所使用的方法）；另一個是科學層面的，利用人工智能概念和模型，幫助回答有關人類和其他生物體的問題。大多數人工智能工作者只關注其中一個目標，但有些也同時關注兩個目標。

人工智能不僅可以帶來不計其數的技術小發明，還能夠對生命科學產生深遠的影響。某一科學理論的計算機模型可以檢驗該理論是否清晰連貫，還能生動形象地證明其含義（通常是未知的）。理論是否正確另當別論，但其依據是從相關科學範疇得出的證據。就算我們發現該理論是錯誤的，結果也能夠給人以啟迪。

值得一提的是，心理學家和神經學家利用人工智能提出了各種影響深遠的心智一大腦理論，如「大腦的運作方式」和「這個大腦在做什麼」的模型：它在回答什麼樣的計算（心理）問題，以及它能採用哪種信息處理形式來達到這一目標等。這兩個問題不一樣，但都十分重要。還有一些問題尚未回答，因為人工智能本身已經告訴我們：心智內容十分豐富，遠遠超出了心理學家們先前的猜想。

生物學家們也用到了人工智能——人工生命（A-Life）。利用這項技術，他們為生物體的不同內部結構建立了計算機模型，以解讀不同種類的動物行為、身體的發育、生物進化和生命的本質。

人工智能對哲學也有影響。如今，很多哲學家對心智的解讀也基於人工智能概念。例如，他們用人工智能技術來解決眾所周知的身心問題、自由意志的難題和很多有關意識的謎題。然而，這些哲學思想都頗具爭議。人工智能系統是否擁有「真正的」智能、創造力或生命，人們對此意見不一。

最後，人工智能向我們發出了挑戰——如何看待人性，以及未來在何方。的確，有些人會擔心我們是否真的有未來，因為他們預言人工智能將全面超過人的智能。雖然他們當中的某些人對這種預想充滿了期待，但是大多數人還是會對此感到害怕。他們會問，如果這樣，那還有什麼地方能保留人類的尊嚴和責任?

我們將在接下來的幾章逐一討論上述問題。

虛擬機

談到人工智能，人們可能會說：「那不就是指電腦嘛。」嗯，他們這麼說既對也不對。電腦不是重點，重點是電腦做的事情。也就是說，雖然人工智能離不開物理機（如電腦），但是我們最好把它看作計算機科學家所說的虛擬機。

虛擬機和虛擬現實中所描述的機器不一樣，和訓練機修工時所使用的模擬汽車引擎也不一樣，它是程序員在編程時和人們使用它時所想到的信息處理系統。

讓我們拿管絃樂隊作類比。首先樂器是不能少的。要想讓樂器演奏出美妙的音樂，那麼木頭、金屬、皮革和絃線都必須遵循一定的物理定律。但觀眾在聽音樂會時並不在意這一點，他們感興趣的是音樂。他們也不在意單個音符，更不用說空氣中發聲的震動了。他們聽的是音符產生的音樂「形狀」：旋律與和聲、主題與變奏、含混音與切分音。

當我們談到人工智能時，情況也類似。用戶使用設計師設計出來的文字處理器直接處理文字和段落。通常情況下，程序本身既不包含文字，也不包含段落（但有些段落也包含，比如用戶可以很容易將版權標示插入到文字中）。神經網絡（見第4章）也是並行處理信息，即使它通常是在約翰·馮·諾依曼（John von Neumann）結構計算機上（按順序）實現的。

當然，這並不是說虛擬機只是杜撰或憑空想像出來的東西。虛擬機是真實存在的。我們不僅可以利用虛擬機完成系統內的任務（如果將其連接到照相機或機器人的手等這樣的物理設備上），甚至還可以做好外部世界的工作。如果程序突發問題，人工智能工作者通常很少去找硬件方面的原因，而是對虛擬機或軟件中的事件和因果關係更感興趣。

編程語言也是虛擬機（它的指令只有翻譯成機器碼後才能運行）。有些指令用更低級的編程語言進行定義，所以多個層級的指令都需要翻譯。否則，要是用機器碼的位組合模式處理信息，大多數人將無法正常思考。如果信息處理過程過於複雜且層級劃分過於細化的話，那麼也沒有人能正常思考。

虛擬機不只是編程語言。虛擬機一般包含各個層級的活動模式（信息處理）。虛擬機也不只是在電腦上運行的虛擬機。在第6章中，我們將看到「人類的心智」也可以被看作在大腦中實現的虛擬機，更確切地說，是並行運行（在不同時間發展和學習得到的）且交互的虛擬機集合。

要實現人工智能領域的進步，我們需要不斷完善有趣實用的虛擬機的定義。不斷改良物理機（更大、更快）確實有好處，它甚至可能是實現某種虛擬機的必要條件。但是，只有具備海量信息的虛擬機才能在這些物理機上運行，否則後者就算功能再強大也沒用（同理，要在神經科學領域取得進步，我們需要清楚瞭解在神經元上實現什麼「心理」虛擬機，詳見第7章）。

各類外部世界的信息得到充分利用。所有人工智能系統都需要輸入和輸出設備，要是只需要一個鍵盤和一個屏幕就好了。它通常還需要專用傳感器（可能是照相機或壓敏晶須）或反應器（可能是供音樂或演講用的聲音合成器或機器人的手）。人工智能程序不僅處理內部信息，還與這些計算機的接口連接，或改變它們。

人工智能程序處理通常包含內部的輸入和輸出設備，供整個系統內部的虛擬機交互。例如，象棋程序的某一部分可能通過注意其他部分的情況來發現自己所面臨的潛在威脅，這時候，它就有可能與那個部分配合，共同阻斷本次威脅。

人工智能的主要類型

信息處理的方法取決於其所包含的虛擬機。我們將在後面的章節中看到，這主要有五種處理類型，每種處理類型又都包含很多變體。一種是經典邏輯或符號主義，有時稱為有效的老式人工智能（Geod Old-Fashioned AI，以下簡稱GOFAI）；另一種是人工神經網絡或聯結主義。此外，還有進化編程、細胞自動機以及動力系統。

工作者通常只使用一種方法來處理信息，但也存在混合虛擬機。例如，在第4章中提到的一個在符號主義處理和聯結主義處理之間不斷切換的人類行為理論（這解釋了為什麼有的人在完成計劃任務的過程中，會分心去關注環境中與之無關的東西以及這種現象是如何發生的）。第5章描述了一款集「情境」機器人學、神經網絡和進化編程三者於一體的感覺運動裝置（在裝置的協助下，機器人將紙板三角形用作地標，找到了「回家」的路線）。

除了實際應用外，這些方法能夠啟發心智、行為和生活。神經網絡有助於模擬大腦的內部結構以及進行模式識別和學習。經典邏輯人工智能（特別是與統計學結合時）可以模擬學習、規劃和推理。進化編程闡明了生物進化和大腦發育。細胞自動機和動力系統可用來模擬生物體的發育。有些方法更接近於生物學，而不是心理學；有些方法更接近非條件反射行為，而不是慎重思考。要想全面瞭解心智，除了要用到上述所有方法外，還可能需要更多別的方法。

許多人工智能工作者並不關心心智的運作方式，他們只注重技術效率，而不追求科學理解。即使人工智能技術起源於心理學，但現在與心理學的聯繫卻很少。然而，我們會發現，如果要想在強人工智能（artificial general intelligence）方面取得進步，我們需要加深理解心智的計算架構。

人工智能的預言

19世紀40年代，埃達·洛夫萊斯（Ada Lovelace）伯爵夫人預言了人工智能。更準確地說，她預言了部分人工智能。她專注於符號和邏輯，從未考慮過神經網絡、進化編程和動力系統。她也未考慮過人工智能的心理目標，而純粹對技術目標感興趣。例如，她說一台機器「可能編寫所有複雜程度或長度的細膩且系統的樂曲」，也可能表達「在科學史上具有劃時代意義的、自然界的重要事實」（因此，如果當她看到以下情況時，她將不會感到吃驚：兩百年以後，科學家們用「大數據」和精心製作的編程方法來推動遺傳學、藥理學、流行病學等無數領域知識的發展）。

她口中的機器是分析機（Analytical Engine）。這是一台齒輪連嵌齒輪的裝置（從未被真正地製造出來），由其密友查爾斯·巴貝奇（Charles Babbage）於1834年設計。雖然這台機器主要用於求解代數和處理數字，但其本質相當於一台通用數字計算機。

她認識到了分析機的潛在通用性和處理符號（表示「宇宙中的所有主體」）的能力。她還描述了現代編程的各種基礎知識：存儲程序、分層嵌套的子程序、尋址、微程序設計、循環、條件、註釋以及程序錯誤。她並沒有談到編曲或科學推理是如何在巴貝奇的機器上實現的。是的，人工智能可以實現，但是實現的方法當時仍然是一個謎團。

人工智能的起源

一個世紀以後，艾倫·圖靈（Alan Turing）解開了這個謎團。1936年，圖靈提出，每個合理計算在原則上都可以由現在被稱為「通用圖靈機」（Turing Machine）的數學系統來執行。圖靈機是一個虛構系統，建立和修改用「0」和「1」表示的二進制符號組合。

第二次世界大戰期間，圖靈在布萊切利園（Bletchley Park）破解德國密碼系統後，到20世紀40年代末一直在思考如何讓一台物理機最接近抽像定義的圖靈機（他幫助設計的第一台現代計算機於1948年在曼徹斯特完成），以及如何讓這台物理機智能地執行任務。

與埃達·洛夫萊斯不同，圖靈接受了人工智能的兩個目標（技術和心理）。他想讓新機器做通常需要智能才能完成的有意義的事情（可能通過使用非自然技術），並模擬以生理為基礎的心智所發生的過程。

1950年，他那篇以幽默方式提出圖靈測試（見第6章）的論文成為了人工智能的宣言〔第二次世界大戰後不久，其論文得到進一步完善，但《官方保密法》（Official Secrets Ac）阻止其出版〕。它抓住了智能信息處理（遊戲、知覺、語言和學習）的癥結，並暗示了當時計算機領域已經取得的成就，讓人躍躍欲試（只有「暗示」，因為布萊切利園的工作仍然屬於最高機密）。它甚至給出了算法，如神經網絡和進化計算，不過在其論文發表很久以後，這些算法才得到廣泛認可。要解開奧秘，這些都只是冰山一角，只是泛泛而談——綱領性的東西，而不是程序。

圖靈堅信，人工智能一定能以某種方式實現。20世紀40年代初，他的這一信念得到了神經病學家/精神病學家沃倫·麥卡洛克（Warren McCulloch）和數學家瓦爾特·皮茨（Walter Pitts）的支持。

他們的論文《神經活動中內在思想的邏輯演算》（A Logical Calculus of the Ideas Immanent in Nervous Activity）結合了圖靈的觀點與另外兩項令人興奮的成果（可追溯到20世紀早期）：伯特蘭·羅素（Bertrand Russell）的命題邏輯和查爾斯·謝林頓（Charles Sherrington）的神經突觸理論。

命題邏輯的關鍵點在於它是二進制的。每個句子（也稱為命題）假定為真或假。沒有中間答案，也不接受不確定性或概率。只允許兩個「真值」，即真和假。

此外，利用邏輯運算符（諸如and、or和if-then）構建了複雜命題，完成演繹論證，而邏輯運算符的意義由子命題的真/假來定義的。例如，如果兩個（或更多）命題由「and」連接，則認為這兩個（所有命題）都是真的。所以當且僅當「瑪麗嫁給湯姆」和「弗洛西嫁給彼得」二者都是真命題，那麼「瑪麗嫁給湯姆和弗洛西嫁給彼得」才是真命題。事實上，如果弗洛西沒有嫁給彼得，那麼包含「and」的複雜命題就是假命題。

麥卡洛克和皮茨將羅素和謝林頓的觀點結合，因為他們都描述了二進制系統。邏輯的真/假（true/false）值映射到圖靈機中的腦細胞開/關（on/off）活動和個體狀態0/1中。謝林頓認為，神經元不僅進行嚴格的開/關活動，而且具有固定閾值。因此，邏輯門（and、or和not）被定義為微小的神經網絡，可以相互連接來表示高度複雜的命題。任何東西只要能用命題邏輯表述，那就能用某種神經網絡和某種圖靈機來計算。

簡單來說，就是神經生理學、邏輯學和計算被放在一起研究。後來，心理學也被納入進來一起討論。麥卡洛克和皮茨相信（就像許多哲學家當時所說的），自然語言在本質上歸結為邏輯。所以，從科學論證到精神分裂症錯覺的所有推理和觀點都可以放到他們的理論「磨坊」裡加工。麥卡洛克和皮茨為整個心理學預言了一個時代，「（神經）網絡的設計規格將對心理學領域取得的所有成果都有幫助」。

其核心含義在當時很清楚：同一個理論方法，即圖靈計算，可用於人和機器智能，麥卡洛克和皮茨的文章甚至影響了計算機的設計。約翰·馮·諾依曼當時打算使用十進制代碼，但他後來意識到了問題，改為二進制。

圖靈當然贊同圖靈計算，但他無法進一步推動人工智能的發展：當時技術太過原始。然而，到20世紀50年代中期，出現了功能更強大且更容易使用的機器。這裡的「易於使用」並不是說更容易打開電腦的按鈕，也不是說更容易將它在房間裡推來推去，而是指定義新的虛擬機更加容易（例如，編程語言），從而有利於定義更高級的虛擬機（例如，用來做數學運算或規劃的程序）。

大約本著圖靈宣言的精神，符號人工智能的研究在大西洋兩岸得以開始。20世紀50年代末期，有一個標誌性事件上了新聞頭條，即阿瑟·塞繆爾（Arthur Samuel）的跳棋（國際跳棋）程序打敗了塞繆爾本人。這無疑暗示著電腦有一天可能會具有超人的智力，超過設計它們的程序員的能力。

20世紀50年代末期，還出現了第二個這樣的暗示，即邏輯理論機（Logic Theory Machine）不僅證明了羅素的18個關鍵邏輯定理，還發現了一個更有效的證明，來證明其中某一個定理。這的確令人印象深刻。塞繆爾只是一個平庸的跳棋選手，但是羅素可是一位世界級的邏輯學家〔羅素本人為這項成就感到十分高興，但是《符號邏輯雜誌》（Journal of Symbolic Logic）拒絕發表一篇計算機程序撰寫並署名的論文，更為重要的是，它並沒有證明一個新定理〕。

邏輯理論機很快就被一般問題解決器（General Problem Solution，以下簡稱GPS）「超越」——「超越」並不是說GPS可以「超越」更多卓越的天才，而是說它的應用範圍不再限制在一個領域。顧名思義，GPS可解決用目標、子目標、動作和運算符表示的任何問題（詳見第2章）。程序員一旦確定與任何特定領域相關的目標、動作和運算符，剩下的推理工作就可以由GPS負責完成。例如，GPS解決了「牧師和野人」的問題（三個牧師和三個野人在一條河的一邊，現在有一艘船，一次最多可以載兩個人。問題來了，如何在野

人數量不超過牧師數量的情況下確保每個人都能過河）。這個問題對人類來說都不簡單，因為每次把兩個人運過去之後，都必須讓其中一個回來，這樣遊戲才能繼續下去（大家可用便士試一試）。

邏輯理論機和GPS都是GOFAI的早期示例。現在說它們是「老式的」，當然毫無疑問；但它們也是「有效的」，率先運用了「啟發法」和「規劃」——二者在今天的人工智能領域都至關重要（見第2章）。

並不是只有GOFAI這種人工智能受到論文《神經活動中內在思想的邏輯演算》的啟發，聯結主義也備受鼓舞。20世紀50年代，計算機上特製或仿真的麥卡洛克和皮茨邏輯神經元網絡被用來〔如艾伯特·厄特利（Albert Uttley）〕模擬聯想學習和條件反射（這些神經元網絡進行集中式而非分佈式處理，與今天的神經元網絡不同，詳見第4章）。

早期網絡模擬並不完全由神經—邏輯統治。雷蒙德·拜沃勒（Raymond Beurle）在20世紀50年代中期實現的系統（在模擬計算機中）大不一樣。他的研究工作沒有始於精心設計的邏輯門網絡，而是始於隨機連接的和不同閾值的單元的二維數組。他認為神經自組織的發生是因為動力波的激活——構建、傳播、堅持、死亡和時不時的相互作用。

拜沃勒已經意識到，詭辯機可以模擬心理過程，但不等於大腦實際上就是這樣的機器。麥卡洛克和皮茨已經指出了這一點，他們發表了一篇具有開創性意義的論文，短短四年後，又在其另外一篇論文中指出了熱力學比邏輯更接近大腦的功能。邏輯學被統計學取代，單一單元被集合取代，確定性純度被概率噪音取代。

換句話說，他們已經描述了我們現在所說的分佈式容錯算法（見第4章），並認為這種新算法是之前算法的「延伸」，彼此並不矛盾。它在生物學上更現實。

控制論

麥卡洛克比GOFAI和聯結主義對早期人工智能的影響更為深遠。20世紀40年代，在其神經學和邏輯學研究成果的指引下，處於萌芽期的控制論運動得到蓬勃發展。

當時，控制論者的研究重心是生物自組織。它涵蓋了各種適應和新陳代謝，包括自主思考、微觀運動行為和（神經）生理調節。其核心思想是「雙向循環性」或反饋。關鍵問題是目的論或目的性。對於反饋取決於目標的差異性而言，這些概念高度相關——目標現階段的偏差被用於指導下一步動作。

1948年，諾伯特·維納（Norbert Wiener，其在戰爭期間設計了反彈道導彈）對該運動進行了命名，將其定義為「關於在動物和機器中控制和通信的研究」。那些控制論者在建立計算機模型的時候，經常從控制工程學和模擬計算機中獲取靈感，而不是從邏輯學和數字計算中獲取。然而，這種區分並不是十分明確。例如，目標的差異性被用於控制制導導彈，以及解決符號問題。此外，圖靈作為經典人工智能的冠軍，用動力學方程（描述化學擴散）定義自組織系統。在這些系統中，諸如點或片段的新結構可以從一堆同質的低級個體中產生（見第5章）。

早期參與該運動的成員還包括：經驗心理學家肯尼思·克雷克（Kenneth Craik）、數學家約翰·馮·諾伊曼、神經學家威廉·格雷·沃爾特（William Grey Walter）和威廉·羅斯·艾什比（William Ross Ashby）、工程師奧利弗·塞爾弗裡奇（Oliver Selfridge）、精神病學家和人類學家格雷戈裡·貝特森（Gregory Bateson）以及化學和心理學家戈登·帕斯克（Gordon Pask）。

克雷克（於1943年在一次自行車事故中逝世，享年31歲）在研究神經系統的過程中參考了模擬計算，當時還沒出現數字計算機。他根

據大腦中「模型」的反饋，大致描述了知覺、微觀運動行為和智力。他的大腦模型或表示概念後來在人工智能中產生了巨大的影響。

整個20世紀30年代，約翰·馮·諾依曼都對自組織心存疑惑，同時又因麥卡洛克和皮茨的第一篇論文感到異常興奮。他不僅將基本的計算機設計從十進制改為二進制，還完善了麥卡洛克和皮茨的觀點，以解釋生物進化和繁殖。他定義了各種細胞自動機，即由很多基本計算單元組成的各種複雜系統。計算單元的變化遵循簡單規則，而這些規則又取決於相鄰單元的當前狀態。其中一些單元可以複製其他單元。他甚至定義了一個能夠複製任何東西的通用複製器——包括複製它自己。他指出，複製錯誤可能導致進化。

約翰·馮·諾依曼用抽像的信息術語對細胞自動機作了詳細說明。但是這些細胞自動機可以用多種方式具現，例如自組裝機器人、圖靈的化學擴散、拜沃勒的物理波或很快將揭開神秘面紗的DNA。

從20世紀40年代末開始，艾什比製作了同態調節器（Homeostat），它是一個生理性自體調解的電化學模型。它可以維持機體內環境的總體恆定，無論最初分配給它的100個參數值是多少（允許設定近400000種不同的起始條件）。它闡釋了艾什比的動態自適應理論在試錯法學習和自適應行為中的應用，這種動態自適應可以發生在身體內部（尤其是大腦），也可以發生在身體與外部環境之間的環境。

格雷·沃爾特也在研究自適應行為，但他用的是一種迥然不同的研究方式。他研發了一款類似烏龜的微型機器人，其感覺運動電路模擬了謝林頓的神經反射理論。這些情境機器人先驅的行為栩栩如生，如尋找光線、避開障礙，以及利用有條件的反射進行聯想學習。這些有意思的機器人於1951年在「英國節」（Festival of Britain）上向公眾展示。

十年後，塞爾弗裡奇（倫敦百貨商店創始人的孫子）利用符號方法實現了一種叫伏魔殿（Pandemonium）的並行處理系統。

這個GOFAI程序利用許多底層「守護程序」（特徵感知器）來學習如何識別模式，每個「守護程序」一直都在感知外界信息，並將感知到的結果傳遞給更高級的「守護程序」。這些「守護程序」重點關注到目前為止一致的特徵（例如，一個F中只有兩根水平條），而忽略了任何不合適的特徵。置信度可以有差異，而且它們至關重要：聲音最洪亮的守護程序影響最大。最後，最高級的守護程序根據既得證據（通常是衝突的），選擇最佳模式。這項研究很快對聯結主義和符號人工智能產生了影響，一個最近的分支是學習智能分佈實體（Learning Intelligent Distribution Agent，以下簡稱LIDA）的意識模型（詳見第6章）。

貝特森對機器沒有什麼興趣。他於20世紀60年代提出了與文化、酗酒和（父或母對子女的）「雙重約束」精神分裂症有關的理論。但是，這些理論基礎卻是在早些時候控制論會議上提出的和通信（即反饋）相關的想法。從20世紀50年代中期開始，帕斯克——麥卡洛克口中的「自組織系統天才」，在許多項目中都用到了控制論和符號思想，其中包括：交互式劇院、互通音樂機器人、獲悉並適應其用戶目標的架構、化學自組織概念和教學機。借助帕斯克的研究，人們能夠利用複雜的知識表示來採取不同方法，而這對認知方式為循序漸進型和整體型（以及對不相關事物不同程度的容忍）的學習者都適用。

簡言之，到20世紀60年代後期，研究人員考慮了所有主要的人工智能類型，甚至將其實現——有的甚至更早。

大多數相關研究人員至今還廣受人們的尊重，但只有圖靈一直是人工智能盛宴上的「幽靈」，其影響無處不在。多年來，其他人只被一些研究領域的分支機構所記住。特別是，格雷·沃爾特和艾什比幾乎被人們遺忘，直到20世紀80年代後期，他們才被讚譽為「人工生命之父」（與圖靈一起）。帕斯克等待的時間更長。要知道其中的原因，我們必須瞭解計算機建模者們是如何分道揚鑣的。

計算機建模者們分道揚鑣

20世紀60年代之前，模擬語言/邏輯思維和模擬有目的的/自適應的微觀運動行為這兩個研究方向是有交叉的。有些專家二者都研究〔唐納德·麥凱（Donald Mackay）甚至建議製造將神經網絡與符號處理結合起來的混合計算機〕。所有相關工作者都能彼此產生共鳴。研究生理自動調整的工作者們認為自己與注重心理研究的同事們做的是同一件事。他們都參加相同的會議：在美國召開的跨學科Macy研討會（1946年到1951年，由麥卡洛克擔任主席）和在倫敦召開的「思維過程機械化」研討會（1958年，由厄特利組織）。

然而，大約從1960年開始，工作者們的研究方向便出現了分歧。廣義上來說，對生命感興趣的人只關注控制論，而那些對心智感興趣的人則關注符號計算。網絡愛好者們當然對大腦和心智都感興趣，但他們通常研究聯想學習，而非具體的語義內容或推理，所以他們關注的是控制論而不是符號人工智能。研究的分支越來越多，不幸的是，各分支之間缺乏對彼此應有的尊重。

這個過程中必然會出現一些優秀的小社會圈子，因為他們討論的理論問題各不相同，既有生理方面的，也有心理方面的。所用的技術也不一樣。廣義上講，是微分方程與邏輯之間的較量。專門化趨勢不斷加強，交流也因此變得越來越困難，而且很大程度上是無利可圖的。兼收並蓄的會議已經過時。

即使如此，各分支學派也不應該太操之過急。對控制論和聯結主義學派的反感源於專業上的嫉妒和正義的憤慨。這是因為符號計算在發展初期取得了巨大的成功；帶有挑釁性的術語「人工智能」〔由約翰·麥卡錫（John McCarthy）於1956年提出，以前稱為「計算機模擬」〕博得了新聞工作者們的眼球；一些不現實的炒作；一些符號主義研究者表現得傲慢自大。

符號主義陣營的成員們認為自己贏得了人工智能的比賽，最初沒有太多的敵意。事實上，他們在很大程度上忽視了早期的網絡研究，其中的一些領導者〔例如，馬文·明斯基（Marvin Minsky）〕已經開始著手網絡的研究。

然而，在1958年，弗蘭克·羅森布拉特（Frank Rosenblatt）提出了神經動力學理論，定義了能在隨機初始狀態下（並且能夠容忍初始化階段的錯誤）進行自組織學習的並行處理系統，並在他的光電感知器中部分實現了該理論。它與「伏魔殿」不同，無須輸入模式讓程序員提前分析。符號主義學派無法忽視這種新形式的聯結主義。但它很快就被打入「冷宮」。20世紀60年代，明斯基與西摩爾·帕普特（Seymour Papert）一道發表了一篇尖銳的批評文章，聲稱感知器連一些基本的東西都無法計算（詳見第4章）。

神經網絡研究的資金來源也因此被切斷。這個結果是由於兩派攻擊者蓄意為之，從而加深了人工智能內部的對抗。

現在大家看來，經典人工智能研究似乎在當時佔絕對的主導地位。誠然，格雷·沃爾特的機器烏龜們在英國節上備受讚譽。和伯納德·威德羅（Bernard Widrow）的模式學習Adaline[1]（基於信號處理）一樣，羅森布拉特的感知器在20世紀50年代後期也被媒體大肆宣傳。但符號主義研究者們的批評讓人們完全失去了對感知器的興趣。20世紀60年代到70年代，在媒體中如日中天的是符號型人工智能（還影響了精神哲學）。

風水輪流轉。神經網絡，如分佈式並行處理（Parallel Distributed Processing，以下簡稱PDP）於1986年再次登台（見第4章）。本該更懂得此方法的大多數外界人士和一些內部人士都把它當成了一個徹頭徹尾的「新」東西。它還吸引了無數研究生和很多新聞媒體（和哲學）的關注。那時，鼻子都被氣歪的人恐怕是那些符號人工智能的研究者了。一時間，PDP研究成為時尚，大家普遍認為經典人工智能的研究當時已經失敗。

還有一些控制論者因其在1987年命名「人工生命」，終於從大批記者和研究生那裡「受寵」。於是，符號人工智能再次受到挑戰。

然而，在21世紀，不同的問題需要不同類型的答案——各有所長，這一點顯而易見。雖然先前的敵意至今猶存，但不同方法仍有相互尊重和合作的空間。例如，「深度學習」有時用於將符號邏輯與多層概率網絡結合的強大系統；還有一些混合方法包含高級複雜的意識模型（見第6章）。

構成人類心智的虛擬機本來就是各式各樣的，因此大家沒必要對人工智能領域的研究分歧太過驚訝。

註釋

[1]Adaline是一個早期的單層人工神經網絡和實現這個網絡的物理設備的名稱。網絡使用存儲電阻器。由伯納德·威德羅教授及其在斯坦福大學的研究生泰德·霍夫（Ted Hoff）於1960年聯合開發。它基於麥卡洛克—皮茨的神經元，由權重、偏差和求和函數組成。——譯者注

02 強人工智能：人工智能領域的聖盃

最先進的人工智能是一種神奇美妙的東西，可以提供各式虛擬機，以進行各種不同類型的信息處理。它沒有核心秘密，也沒有統一的核心技術。人工智能工作者來自各個領域，幾乎沒有統一的目標和方法。本書只能涵蓋最近取得的極少一部分成就。總之，人工智能的方法範圍極其寬泛。

可以說，人工智能已經取得了驚人的成就。它的實際運用範圍也十分廣泛。我們現在有大量針對無數特定任務而設計的人工智能應用程序。生活中各個領域的專業人士和非專業人士幾乎都在使用。很多程序甚至比一些專家還牛。由此看來，人工智能的進展的確引人注目。

但是，人工智能先驅們的目標不僅限於專家系統，他們還希望發展通用智能系統。他們模擬的所有人類能力——視覺、推理、語言、學習等——可應對各類挑戰。此外，這些能力將適時得到整合。

從這些標準來看，進步的空間非常大！約翰·麥卡錫很早就認識到人工智能需要「共識」。作為1971年和1987年圖靈獎的獲得者，他在發表獲獎感言的時候談到了人工智能的通用性（Generality in Artificial Intelligence），但他其實是在抱怨，而不是慶祝。到2016年，他的抱怨仍未得到答案。

隨著近來計算機能力的不斷增強，強人工智能在21世紀再次引起人們的興趣。如果這一目標得以實現，人工智能系統將減少對專用編程技巧的依賴，而受益於推理和知覺這些通用功能——語言、創造力和情感（所有這些我們都將在第3章中討論）。

然而，這談何容易。通用智能仍然是一個嚴峻的挑戰，讓人難以捉摸。強人工智能無疑是人工智能領域的聖盃。

只有超級計算機還遠遠不夠

對於任何想要實現這個夢想的人而言，超級計算機必然是一個助推器。組合爆炸——其中需要的計算超過實際能執行的計算——已不再構成威脅。然而，我們不能一直靠增強計算能力來解決問題。

通常，我們還需要新的解決方法。此外，即使某一具體方法從理論上講是可行的，但也可能需要大量的時間或存儲，才能在實踐中發揮作用。第4章給出了相關例子（有關神經網絡）。同樣，窮舉法列出了所有可能的象棋步驟，但它需要的存儲位置比宇宙中的電子還要多，因此就算有一大堆超級計算機也不能滿足需要。

另外，效率也很重要：計算數量越少越好。總之，問題必須易於處理。

如今，已經有幾個基本策略可以做到這一點。它們由經典符號人工智能或GOFAI開創，在今天仍然必不可少：第一，只關注一部分搜索空間（問題的計算機表示，解決方案假定就在該表示中）；第二，簡化假設以構建較小的搜索空間；第三，有效安排搜索過程；第四，用新方式表示問題，以構建不同的搜索空間。

這些方法分別對應的是啟發法、規劃、數學簡化和知識表示。接下來的五個部分將分別討論這些強人工智能策略。

啟髮式搜索

「啟髮式的」（heuristic）和「找到了」（Eureka）在英文中有相同的詞根：來自意為「尋找」或「發現」的希臘語。啟發法得到了早期GOFAI學派的重視，並且經常被看作「編程技巧」。但是，這個術語並非源於編程：邏輯學家和數學家對其早已熟悉。早在數千年前，人類就用啟發法解決問題（有意或無意地），遠遠早於埃達·洛夫萊斯伯爵夫人（Ada Love Lace』s）預見人工智能的時間。

無論是對人還是對機器來說，啟發法都有利於問題的解決。強人工智能使用啟發法的模式是：讓程序只針對搜索空間的某些部分，同時避開其他部分。

很多啟髮式算法都屬於無法保證成功的經驗法則，如早期人工智能使用的大多數經驗法則。在啟發法引導下，系統正好忽略了某部分的搜索空間，而解決方案可能正好位於這部分空間裡。例如，在國際象棋中，「保護女王」是一條非常有用的規則，但偶爾也應該違背。

還有一些啟髮式算法從邏輯學角度或數學角度被證明是合理的。如今，人工智能和計算機科學領域的大量工作都是為了確定程序可證明的屬性。這是「友好人工智能」的一個方面，因為人類安全可能由於使用從邏輯學角度看不是很可靠的系統而受到的威脅（詳見第7章，啟髮式算法和算法之間原則上沒有區別，許多算法實際上是包含多個特定啟髮式算法的微型程序）。

無論啟發法可靠與否，它對人工智能搜索來說都不可或缺。上文提到人工智能越來越專業化，這部分取決於能顯著提高效率的新啟發法的定義，但僅限於限制頗多的某類問題或搜索空間。一個非常成功的啟發法可能並不適合讓其他人工智能程序「借用」。

如果我們給定幾種啟發法，那麼應用它們的順序就可能變得很重要。例如，即使這種排序偶爾會導致災難，也應該先考慮「保護女

王」，再考慮「保護象」。不同的順序將定義不同的搜索樹遍歷整個搜索空間。給啟發法下定義和排序是現代人工智能的關鍵任務（啟發法在認知心理學中的地位也很突出，例如，「快速節儉啟髮式」指出進化如何讓人獲得對環境作出有效回應的方法）。

利用啟發法，我們不再需要窮舉搜索整個搜索空間，但它們有時會和（有限的）窮舉搜索結合使用。1997年，因擊敗世界冠軍加裡·卡斯帕羅夫（Gary Kasparov）而名聲大噪的IBM國際象棋程序深藍（Deep Blue）使用的是專用硬件芯片，每秒能處理2億個位置，可以知道接下來8步的所有備選棋步。但是，它不得不用啟發法來選擇備選棋步中的「最佳」棋步。由於啟發法可信度不高，所以即使是深藍，也不能每次都勝出。

人工智能領域中的規劃

在今天的人工智能領域中，規劃的地位十分突出，尤其是在很多軍事活動中。事實上，直到最近才在人工智能上支付大部分研究費用的美國國防部指出，在第一次伊拉克戰爭的戰場上，他們在後勤保障方面省下的錢（利用人工智能規劃）超出了其前期投入。

規劃不僅限於人工智能：我們都在規劃。請想想假期打包東西的情景。你必須找到所有要帶的東西，但這些東西可能沒有放在同一個地方。你可能還要買一些新東西（比如防曬霜）。你必須決定是把所有東西放一塊兒（也許放在床上，也許放在桌子上），還是把每樣東西都放在皮箱裡。這個決定可能部分取決於你是否決定最後再放衣服，因為你怕把衣服弄皺。你是需要一個背包，還是要一個手提箱，或者兩者都要，你如何取捨？

把規劃用作人工智能技術的GOFAI程序員們考慮到了有意識的深思熟慮（基於神經網絡的人工智能大不一樣，因為它不模擬有意識的深思熟慮，詳見第4章）。因為負責邏輯理論機（見第1章）和GPS的先驅們主要對人類推理心理學感興趣。他們的程序基於被試主體為人的實驗，要求被試人進行「有聲思維」：一邊做邏輯謎題，一邊描述自己的思考過程。

現代人工智能規劃程序並不那麼依賴從有意識的內省中或實驗觀察中獲得的想法。它們的規劃比早期程序的規劃要複雜得多，但基本思想一致。

一個規劃列舉出一系列在常規層級上表示的動作——一個最終目標、加上很多子目標和次級子目標……這樣就不用一次性考慮所有的細節。在某一適當的抽像層級規劃上可以修剪搜索空間內的搜索樹，因此一些細節根本不需要考慮。有時，最終目標本身就是一個動作規劃——可能是調度送向/運出工廠或戰場的貨物。有時候，它還是一個問題的答案，比如醫療診斷。

針對任何給定目標和預期情況，規劃程序需要一個動作列表（即符號運算符），或多種動作類型（通過填寫來源於問題的參數，可以將動作實例化），每個動作都能夠作出一些相應的變化；針對每個動作，規劃程序需要一組必要的前提條件（比較抓住某個東西，前提條件是這個東西必須是在手的可活動範圍內）；針對所需變化的優先次序和對動作的排序，規劃程序需要啟發法。如果程序要選定某個特定動作，它可能要設置一個新的子目標以滿足前提條件。這個目標制定的過程可以一再重複。

通過規劃，程序——或人類用戶——可以發現已經做了什麼動作，以及為什麼這麼做。「為什麼」指的是目標的層次結構：做這個動作是為了滿足那個前提條件，以實現這個子目標。人工智能系統通常採用「正向鏈接（推理）」和「反向鏈接（推理）」技術，來解釋程序是如何找到解決方案的。這可以幫助用戶判斷程序的動作或建議是否恰當。

當前，一些規劃程序擁有數萬行代碼，它們在無數層級上定義分層的搜索空間。這些系統通常與早期的規劃程序大相逕庭。

例如，大多數規劃程序假定，並非所有的子目標都可以被獨立處理（即問題可完全分解）。畢竟在現實生活中，某一個以目標為導向的行動所產生的結果都可能被另一個行動撤銷。今天的系統能處理可部分分解的問題：它們獨立處理子目標，但必要的時候，還可以做其他處理，以整合隨之產生的各項子規劃。

經典系統只能解決那些在完全可觀察的、可確定的和有限的靜態環境中出現的問題。但是，一些現代規劃程序可以處理在部分可觀察（即系統的「世界」模型可能不完整或不正確）和不確定的環境中出現的問題。在這些情況下，系統必須監控執行期間的變化情況，以在規劃中，或在自己對於「世界」的「信念」中，作出適當改變。一些現代規劃程序可以在很長一段時間內一直進行監控：它們可以根據環境變化，不斷制定、執行、調整和捨棄目標。

其他許多開發領域已經取得了新進展，並且還在不斷進步。所以，在20世紀80年代，一些機器人專家完全否決規劃並轉而青睞於「情境」機器人學（見第5章）的情況，著實讓人驚訝，例如，目標和可能動作的內部表示等概念也被否決了。然而，在很大程度上，這種批評是不對的。批評者們自己的系統可能沒有表示目標，但可能表示了其他東西，例如視網膜刺激和獎勵。此外，即使是最先遭受此類批評的機器人技術，也常常需要規劃以及純粹的被動回應，例如，製造會踢足球的機器人。

數學簡化

鑒於啟發法任由搜索空間自由發展（這樣程序就可以專注於搜索空間裡最合適的部分），簡化假設構造了一個不切實際但更易於計算的搜索空間。

一些假設和數學相關，如機器學習中通常使用的「i.i.d.」（獨立同分佈）假設。與數據中實際的概率構成相比，用i.i.d.表示的概率構成要簡單得多。

數學簡化在定義搜索空間時的優點是可以利用搜索的數學方法。這種方法定義清晰，而且至少方便數學家理解。這並不代表任何數學定義的搜索都有實際使用價值。如上所述，某一方法在數學層面可以保證解決某一特定類別的所有問題，但是該方法在實際運用中並不適用，因為其時間成本可能是無限的。然而，它可能建議與此類似但更實際的方法，請見第4章中有關「反向傳播」的討論。

在人工智能領域，非數學的簡化假設比比皆是，並且通常很直接。一種是假定無須考慮情感因素（見第3章）就能定義和解決問題的（默認）假設。還有很多假設被構建在用來指定任務的一般知識表示中。

知識表示

一般來說，實現人工智能的最難部分是最開始如何向系統表示問題。即使人類看似可以直接與程序交流（用英語對著Siri說話或者在谷歌的搜索引擎中輸入法語單詞），但事實並非如此。人們無論是處理文本還是圖像，都必須將包含的信息（「知識」）以機器可以理解的方式表示給系統，換句話說，以系統可以處理的方式表示（機器是否「真」的理解，我們將在第6章中進行討論）。

人工智能表示問題的方式五花八門。有些是對GOFAI中知識表示的一般方法進行演繹或改動。越來越多的是針對一小類問題制定的一些高度專業化的方法。例如，我們可能為人類某種癌細胞的X射線圖像或照片精心設計一種新的表示方法，從而可以得到某個非常具體的醫學解釋方法（因此，這種新方法不能用來識別貓或CAT 掃瞄）。

欲實現強人工智能，通用方法是關鍵。這些方法最初受到人類認知心理學研究的啟發，包括：IF—THEN規則集；個體概念的表示；模式化的動作序列；語義網絡；以及利用邏輯或概率進行推理。

接下來，我們對上述方法依次展開討論（第4章描述了另一種知識表示形式，即神經網絡）。

基於規則的程序

在基於規則的編程中，大量知識/信念被表示為一套將條件與動作聯繫起來的「如果—則」（IF—THEN）規則集：如果滿足這個條件，則進行那個動作。這種形式的知識表示利用了形式邏輯〔埃米爾·珀斯特（Emil Post）的「產生式」系統〕。但是人工智能先驅艾倫·紐厄爾（Allen Newell）和赫伯特·西蒙（Herbert Simon，又名司馬賀）通常認為它是人類心理學的基礎。

條件和動作都可能很複雜，規定的內容可能是幾個或多個命題的合取（或析取）。如果同時滿足幾個條件，則包含最多命題的合取被賦予優先級。所以，「如果目標是製作烤牛肉和約克郡布丁」將優先於「如果目標是製作烤牛肉」，而在條件中增加「和三種蔬菜」又優先於「如果目標是製作烤牛肉和約克郡布丁」。

基於規則的程序不提前規定每一步的順序。相反，每條規則都在等待被其條件觸發。儘管如此，這類系統可以用來做規劃。如果不能做規劃，那麼它們在人工智能方面就只能發揮有限的作用。但是它們的規劃方式不同於最古老、最為人們熟悉且最常用的編程形式（有時稱為「執行控制」）。

在能進行執行控制的程序（如GPS和邏輯理論機，詳見第1章）中，規劃被明確表示。程序員按照嚴格的時間順序，規定一個尋找目標的指令序列，指出哪一步該執行哪一條指令：「做這個，然後去做那個；然後看看X是否為真；如果是真，就做這個事；如果不是，就做那個事。」

「這個事」或「那個事」有時是一條設置某個目標或子目標的明確指令。例如，機器人如果有離開房間的目標，那麼可能指示該機器人設置開門的子目標（原文如此）；接下來，如果檢查門當前狀態的結果顯示門將被關閉，則設置抓握門把手的子次目標（人類蹣跚學步者可能需要更低級的子次目標——即讓成年人抓住自己夠不著的門把

手，並且如果嬰兒要做到這一點，可能需要在更低級別設定幾個目標）。

基於規則的程序也可以用來解決如何逃離房間的問題。然而，規劃層級不會被表示為按時間順序排列的明確步驟，而是表示為構成系統的「IF—THEN」規則集合中所隱含的邏輯結構。某一條件可能要求已經建立了這樣一個目標（IF你想打開門，而且你不夠高）。同樣，動作可以包括設置一個新目標或子目標（THEN找一個成人）。更低級的目標將自動激活（IF你想要求一些人做一些事，THEN設置接近他們的目標）。

當然，程序員必須列入相關的IF—THEN規則（上述案例中指的是涉及門和門把手的規則）。但是，他們不需要預期這些規則的所有潛在邏輯含義（這是一把「雙刃劍」，因為潛在的不一致可能在很長一段時間都無法被發現）。

被激活的目標/子目標被貼在中央「黑板」上，可供整個系統訪問。顯示在黑板上的信息不僅包括被激活的目標，還包括感知輸入和當前處理的其他方面（該想法不僅影響了一個意識神經心理學的前沿理論，還影響了以它為基礎的意識人工智能模型，詳見第6章）。

基於規則的程序廣泛應用於20世紀70年代早期出現的先驅「專家系統」。這些系統包括：MYCIN系統——在人類醫生鑒定感染性疾病和開抗生素藥物時提建議；還有樹枝狀演算法（DENDRAL）——對有機化學中某一特定範圍內的分子進行光譜分析。例如，做醫療分析的計算機咨詢專家系統MYCIN，它的診斷方法是將症狀/病人本身的身體狀況（條件）與診斷結論/建議相匹配，以便繼續檢測或開處方（動作）。這些程序是人工智能遠離「從一般化走向專門化之夢」的第一步，同時為實現埃達·洛夫萊斯的夢想邁出了第一步——機器製造科學之夢（見第1章）。

由於基於規則的知識表示，程序能夠被逐步建立，因為程序員或者強人工智能系統本身可增加對域的瞭解。新規則可以隨時添加。沒有必要從頭重新編寫程序。但是有一個棘手的問題，即如果新規則與

舊規則邏輯衝突，系統將不會總是做它應該做的事，甚至可能和它應該做的事相去甚遠。在處理一小組規則時，這些邏輯衝突很容易被避免，但是如果系統較大，它們就很難被識破。

20世紀70年代，新IF—THEN規則在與人類專家不斷對話的過程中得到，它們被要求解釋自己的決定。今天，儘管許多規則不是來自有意識的內省，但這些規則更高效。現代專家系統（今天很少使用的術語）應用範圍廣，從大型科學研究和商業程序到手機上小的應用程序。由於受益於其他形式的知識表示，許多系統超過舊系統，如統計和專用視覺識別或大數據的使用（見第4章）。

在某些狹窄領域中，這些程序可以幫助甚至取代人類專家。有些超越了上述領域的世界翹楚。近四十年前，在診斷大豆疾病的時候，一個基於規則的系統比最權威的專家還準確。如今，用它來幫助科學、醫學、法律甚至服裝設計領域專業人士的例子不勝枚舉（這不完全是好事，詳見第7章）。

框架、詞向量、腳本、語義網絡

其他常用的知識表示方法包含個體概念，而不是整個領域（如醫學診斷或服裝設計）。

例如，可以通過規定分層數據結構（有時稱為「框架」）告訴計算機什麼是房間。它將一間房表示為有地板、天花板、牆壁、門、窗戶和傢俱（床、浴缸、餐桌……）。真實的房間具有不同數量的牆壁和門窗，因此可在框架中的「插槽」裡填充特定數字，並提供缺省賦值（四道牆、一扇門和一扇窗）。

計算機可以使用這類數據結構找到相似類、回答問題、參與對話、創作或理解故事。它們是CYC[1]（encyclopedia，即百科全書）的基礎：一個試圖表示所有人類知識的大膽嘗試。有人甚至說這個想法是癡人說夢。

然而，框架也可能造成誤導。例如，缺省賦值就有諸多問題（有些房間沒有窗戶，開放式的房間沒有門）。更糟糕的情況是：該如何表示下落或溢出這樣的日常概念？符號人工智能這樣表示「樸素物理學」的常識性知識：構造對事實進行編碼的框架，如未支撐的物體會下落，但也有例外——氦氣球就不會下落。考慮清楚這類情況是一項永無止境的任務。

在一些利用最新技術處理大數據的應用中，單個概念可能被表示為一個簇或「雲」，由成百上千個偶爾相關的概念組成（概念對之間的相關性概率各不相同，詳見第3章）。類似地，概念現在可以用「詞矢量」而不是單詞來表示。此處的語義特徵生成許多不同概念並連接各個概念，由（深度學習）系統發現，可用來預測接下來的詞——例如，在機器翻譯中的運用。然而，這些表示用在推理或談話中的時候，不像經典框架那麼經得起檢驗。

有些數據結構（稱為「腳本」）表明熟悉動作的順序。例如，哄小孩子睡覺通常要做以下動作：蓋被子、讀故事、唱首搖籃曲、打開小夜燈。這樣的數據結構既可用來問答問題，也可用來提問題。如果媽媽省掉打開小夜燈的動作，就會出現這樣的問題，如「為什麼」以及「接下來發生了什麼」，換句話說，這裡有故事開始的緣由。因此，這種形式的知識表示被用於自動書寫故事，也正是和人類能正常交談的「陪護」計算機所需要的知識表示形式（見第3章）。

概念的另一種知識表示形式是語義網絡（這些是集中式網絡，見第4章）。20世紀60年代，羅斯·奎利恩（Ross Quillian）率先提出了幾個延伸示例（例如WordNet[2]）作為人類聯想記憶的模型，現在屬於公共數據資源。語義網絡通過以下方法連接概念：如同義、反義、從屬、上位、部分—整體這樣的語義關係；以及將真實的世界知識比作語義學的聯想連接（見第3章）。

語義網絡可能增加為音節、初始字母、語音學和同音異義詞編碼的連接，來表示概念和詞。金·賓斯泰德（Kim Binsted）的JAPE和格雷姆·裡奇（Graeme Ritchie）的STAND UP在使用這種網絡，它們基於雙關語、解釋和變換音節來製造笑話（9種不同類型）。例如，問：什麼叫沮喪的火車？答：低壓機車；問：羊和袋鼠生出來的寶寶是什麼？答：一位毛茸茸的跳高運動員。

注意：語義網絡與神經網絡不同。我們將在第4章中看到，分佈式神經網絡以迥然不同的方式表示知識。在神經網絡中，單個概念不是用精心定義的聯想網絡中的單個節點來表示，而是用整個網絡上活動的變化模式來表示。這類系統可以容忍衝突跡象，因此不需要考慮保持邏輯一致性的問題（將在下一節描述）。但它們無法進行精確推理。不過，這種知識表示類型十分重要（並且是實際應用的一個重要基礎），值得我們用一個單獨的小節對其展開討論。

註釋

[1]CYC是一個致力於將各個領域的本體及常識知識綜合地在一起，並在此基礎上實現知識推理的人工智能項目。其目標是使人工智能的應用能夠以類似人類推理的方式工作。這個項目是由道格拉斯·萊納特（Douglas Lenat）在1984年設立的，由Cycorp公司開發並維護。——譯者注

[2]WordNet是一個由普林斯頓大學認識科學實驗室在心理學教授喬治·A.米勒的指導下建立和維護的英語字典。開發工作從1985年開始，從此以後該項目接受了超過300萬美元的資助（主要來源於對機器翻譯有興趣的政府機構）。由於它包含了語義信息，所以有別於通常意義上的字典。WordNet根據詞條的意義將它們分組，每一個具有相同意義的字條組稱為一個synset（同義詞集合）。WordNet為每一個synset提供了簡短、概要的定義，並記錄不同synset之間的語義關係。——譯者注

邏輯和語義網

如果一個人的最終目標是強人工智能，邏輯似乎是一種超級不錯的知識表示。因為邏輯普遍適用。原則上來說，相同的表示（相同的邏輯符號主義）可以用來表示視覺、學習和語言等，當然也適用於由此產生的任意集成。此外，它提供了很有說服力的定理證明方法，以處理信息。

所以，早期人工智能中的知識表示方式首選謂詞演算。這種邏輯比命題邏輯的表示能力更強，因為它可以「進入句內」來表達句子的意思。以「這個商店有一頂適合所有人的帽子」這個句子為例。謂詞演算可以清楚區分這句話三種可能的意思：「對於每個人來說，這家商店有一頂適合他們的帽子」；「在這家商店有一頂尺寸可調的帽子，適合任何人」；和「在這個商店有一頂帽子（假定被折起來）足夠大，可以同時適合所有人。」

對許多人工智能研究人員來說，謂詞邏輯仍然是首選。例如，CYC的框架就是基於謂詞邏輯。組合語義學中的自然語言處理（NLP）表示也同樣如此（見第3章）。我們延伸謂詞邏輯來表示時間、原因或職責/道德。當然，這取決於某人已經提出了這些形式的模態邏輯，但這並非易事。

然而，邏輯也有缺點。

第一個缺點包含組合爆炸。人工智能中廣泛使用的邏輯定理證明方法是消解法。利用這種方法得出的結論可能本身是正確的，但它與目標結論並不相關。啟發法用來指導和限制結論，並決定何時停止證明（魔法師的弟子[1]做不到）。但這些方法也並非萬無一失。

第二個缺點是消解定理證明，假定非一非一X就意味著X。這個觀點大家並不陌生：反證法就是首先假設某命題不成立（對原命題的結論進行否定），然後推理出明顯矛盾的結果，從而下結論說假設不成

立，原命題得證。如果被推理的域被完全理解，那麼這在邏輯上是正確的。但是，使用內置消解程序（例如許多專家系統）的用戶通常假定找不出矛盾來，這就意味著不存在矛盾，即所謂的「失敗則否定」。這往往是一個錯誤。在現實生活中，證明某事是假，和不能證明它是真完全不是一回事（如在猜測伴侶是否欺騙你的時候）。因為還有許多不知道的證據（潛在假設）。

第三個缺點是，在經典（「單調推理」）邏輯中，一旦某事被證明是真，那它永遠是真。在現實中，情況並不總是如此。我們可以有充分理由認為X為真（也許它是一個缺省賦值，甚至是通過仔細論證或從有說服力的證據中得出的結論），但後來可能會發現X不再是真，或者從最開始就不是真。如果是這樣，我們也必須相應地改變自己的認知。對於基於邏輯的知識表示，這說起來容易做起來難。許多研究者受到麥卡錫的啟發，已經試圖提出可以容忍不斷變化的真值的「非單調推理」邏輯。類似地，人們已經定義了各種「模糊」邏輯，其中的語句能夠被標記為可能/不可能或者未知，而不是真/假。即便如此，防止單調性的可靠方法仍未找到。

有些人工智能專家在研究基於邏輯的知識表示法時，基本上都是越來越想找到知識或意義的本元。但他們不是先驅：麥卡錫和海斯（Hayes）在其合著的論文《從人工智能立場中衍生出來的某些哲學問題》（Some Philosophical Problems from an AI Standpoint）中就在做這件事。大家對這篇文章討論的很多問題都不陌生：從自由意志到非真實條件句。這些問題包含宇宙的基本本體論：狀態、事件、屬性、變化、動作……什麼？

除非某人在內心深處就篤信形而上學（這種激情十分罕見），不然為什麼要關心本體這種東西？為什麼現在對這些神秘問題的探討越來越多？很顯然，如果試圖設計強人工智能，就必須考慮知識表示能夠使用何種本體。我們在設計語義網的過程中也要考慮這些問題。

語義網與萬維網不同。自20世紀90年代以來，我們就有了萬維網。語義網甚至不是技術發展的最新水平：它是未來的技術發展水平。如果存在語義網，而且當它真的存在時，機器驅動的聯想搜索將

通過機器的理解力得到改進和補充。這樣一來，應用程序和瀏覽器就可以訪問互聯網上的任何信息，並在推理問題的過程中合理整合不同內容。這項艱巨任務由蒂姆·伯納斯-李爵士（Sir Tim Berners-Lee）指導，這項任務甚至可以說是苛求，不僅需要在硬件和通信基礎設施方面取得巨大的工程進步，還需要網絡漫遊程序加深理解它們正在做什麼。

谷歌、一般的NLP程序等這類搜索引擎通常可以找到單詞或文本之間的關聯，但不存在理解力。這裡不是指哲學上的理解力（見第6章），而是指一種經驗性的東西，是實現強人工智能的另一個障礙。儘管有一些例子聽起來很誘人，但終究是騙人的，如IBM公司的沃森、Siri和機器翻譯（都將在第3章中討論），今天的計算機並不知道它們「讀」或「說」的東西是什麼意思。

缺乏理解力的體現之一是各程序之間不能彼此交流（相互學習），因為程序不同，知識表示形式或基本本體也不同。如果語義網研究人員可以找到一種通用的本體論，那麼讓機器理解其接收的東西可能不再只是空想。因此，在20世紀60年代，人工智能領域提出的形而上學的問題，如今因其實用性而變得非常重要。

註釋

[1]《魔法師的弟子》（The Sorcerer's Apprentice）是德國詩人歌德在1797年創作的一首詩。詩的內容大致是：一位老魔法師離開店舖前，給弟子留了一堆活。這位弟子在老魔法師離開之後，因疲於用桶提水，所以就對一把掃帚施了魔法，但是他並沒完全熟練掌握這種魔法。可想而知，接下來地板上到處都是水。而這位弟子也意識到自己無法讓掃帚停下來，因為他根本不知道怎麼讓它停。——譯者注

計算機視覺

今天的計算機不能像人類一樣理解視覺圖像（同樣，這是一種經驗性的東西：強人工智能是否可以運用有意識的視覺現象學，將在第6章中討論）。

自1980年以來，有關人工智能視覺的各種知識表示研究大多基於心理學，特別是大衛·馬爾（David Marr）和詹姆斯·吉布森（James Gibson）的理論。馬爾注重構建3D表示（通過反轉圖像形成的過程），而不是用它們執行動作。吉布森強調視覺可供性對動作的支持，即一些視覺線索，它們能提示一條路或一根承重樹枝，甚至是友好的或敵對的物種成員。儘管有了這些心理學研究基礎，但是目前的視覺程序仍然嚴重受限。

不可否認，計算機視覺已經取得了顯著成就，例如面部識別的成功率達到98%，閱讀草書筆記，注意到停車場內某人的可疑行為（一直站在車門邊不走），甚至能比人類病理學家更準確地鑒定出一些病變細胞。面對這些成功，人們開始變得焦慮。

但是程序（很多是神經網絡，見第4章）通常必須準確地知道它們想要什麼，例如，一張臉不能倒置、不能側擺、一點也不能被別的東西擋住（98%的成功率），而且還得是在特定的光亮下。

「通常」這個詞很重要。2012年，谷歌實驗室整合了1000台大型（16核）計算機，建成了一個巨大的神經網絡，擁有10億多個聯結。然後，由研究人員將來自YouTube視頻的1000萬張隨機圖片放入這個具備深度學習能力的網絡中。該網絡沒有被告知要找什麼，圖像也沒有被標記。然而，三天後，其中一個單元（一個人工神經元）學會了對一張貓臉圖像和一張人臉圖像作出反應。

這讓人印象深刻吧？嗯，是的。這也很吸引人：研究人員很快想到了大腦中的「祖母細胞」。自從20世紀20年代以來，神經科學家對

於是否存在祖母細胞的觀點各持己見。如果說它們存在，那就表示大腦中有些細胞（單個神經元或小組神經元）當且僅當察覺到祖母或某些特定特徵的時候會被激活。顯然，谷歌的貓臉識別網絡情況類似。而且，雖然貓臉必須完整且擺正，但其大小可以改變，也可以出現在200×200陣列的不同位置。研究人員訓練精心預選的（但未標記）人臉圖像（包括側臉）上的系統，由此發現某一單元偶爾能夠辨別避開取景器的臉。

此類成就不久將紛紛湧現，甚至更加令人驚歎。多層網絡已經在面部識別領域取得了巨大進步，有時還可以找到圖像最突出的部分，並給出相應的語言描述（例如，「在戶外市場購物的人」）。最近發起的「大規模視覺識別挑戰賽」每年都在增加可識別視覺種類，並減少對相關圖像的約束（例如，對象的數量和遮擋）。然而，這些深度學習系統仍未克服舊系統的一些弱點。例如，貓臉識別器這類系統並不理解什麼是3D空間，也不知道「側臉」或遮擋的真實含義，甚至是為機器人設計的視覺程序對它們也只有一丁點瞭解。

再看看機遇號（Opportunity）和好奇號（Curiosity），這兩種火星漫遊機器人分別於2004年和2012年登陸。它們依賴於特殊的知識表示法：針對其可能面臨的3D問題量身打造的啟發法。它們無法在一般情況下尋路或操控對象。有些機器人模擬有生命的視覺，在這個過程中身體的運動可提供有用信息（因為它們系統地改變視覺輸入）。但這些機器人可能忽略一些可行路徑，也不知道自己的手能拿哪些陌生的東西。

在本書出版之際，有些機器人可能已經具備上述能力了，但它們還是會受限。例如，它們理解不了「我不能把那個東西拿起來」這句話的意思，因為它們根本不知道什麼叫「可以」和「不可以」，它們的知識表示也有可能仍然不具備必要的模態邏輯。

有時候，視覺能夠忽略3D空間，比如閱讀筆跡。在許多高度受限的2D任務中，計算機視覺比人類視覺犯的錯誤要少。的確，有時候人眼無法識別的複雜模式（例如，X射線中的模式），計算機視覺能夠

採用高精尖的非自然技術來分析（同樣，3D計算機視覺常常利用非自然方式取得顯著成就）。

但是，即使是2D，計算機視覺也有局限。儘管不乏研究類比表示或圖像表示的工作，但是人工智能在解決問題時無法以可信賴的方式使用圖表，而我們在幾何推理中或在封皮背面繪製抽像關係時可以做到這一點（同樣，心理學家也還不知道做到後者的方法）。

總之，大多數人類視覺的成就優於今天的人工智能。通常，人工智能研究人員不清楚要問什麼問題。以「整齊折疊光滑的真絲禮服」為例。沒有機器人可以做到這一點（雖然我們可以一步一步地指導它們如何折疊長方形毛巾）。還有穿T恤：必須先把頭套進去，而不是先穿袖子。但是這是為什麼？在人工智能中幾乎不存在這種拓撲問題。

難道這意味著人類水平的計算機視覺就無法實現嗎？不是。不過要做到這一點，比大多數人想得更困難。

因為描述視覺的特性就是一塊難啃的「硬骨頭」。所以，它是第1章所列事實的一個特例：人工智能已經告訴我們，人腦的豐富程度和微妙程度超出了心理學家之前的猜想。事實上，這也是我們從人工智能中學到的最重要的一課。

框架問題

無論在什麼樣的領域中，要找到合適的知識表示都很難，部分原因是需要避免框架問題（注意：雖然在用框架作為概念的知識表示時，也出現過這個問題，但是此處「框架」的含義不一樣）。

正如麥卡錫和海斯最初定義的那樣，框架問題指假定（由機器人規劃時）一個動作僅會引起這些改變，然而它也可能會導致那些改變。一般來說，只要人類默認的含義被計算機忽略，框架問題就會出現，因為這些含義沒有被闡明。

經典案例就是猴子和香蕉的問題，其中的問題解決者（可能是機器人的人工智能規劃程序）假定在框架外不存在相關事物（見圖2—1）。

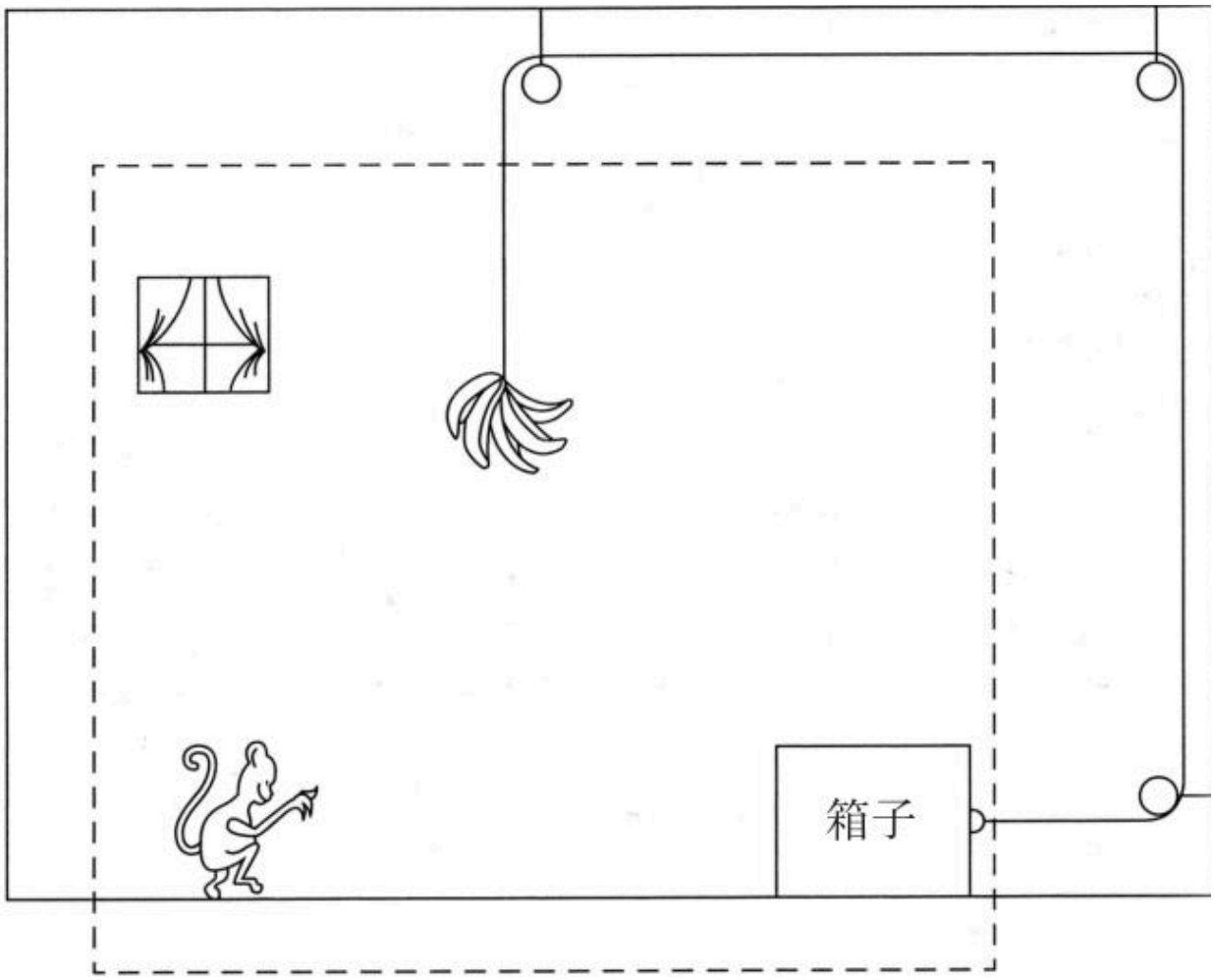

圖2—1 猴子和香蕉問題：猴子怎麼得到香蕉

圖註：解決這個問題的一般方法是：假定相關的「世界」是虛線框架內表示的世界（雖然沒有明確說明）。換句話說，這個框架之外不存在什麼東西能讓它內心發生重大變化，它會繼續移動箱子。

資料來源：摘自瑪格麗特·博登1977年所著的《人工智能和自然人》（Artificial Intelligence and Natural Man）一書第387頁。

我自己最喜歡的例子是：假設一位20歲的男士可以在一個小時內採摘10磅黑莓，一位18歲的女士可以採摘8磅，那麼他倆要是一起去採摘黑莓，一共能摘多少？當然，「18磅」不是一個合理的答案。有可能超過18磅（因為他們都想炫耀），也有可能少於18磅。我是在五十年前第一次聽說這個例子的，不過它在當時更具代表性。但為什麼答

案不確定呢？這裡涉及什麼樣的知識？強人工智能是否能克服這種看似普通的算術事實？

框架問題產生的原因是人工智能程序沒有人類的關聯感[1]（見第3章）。如果能知道每個動作可能帶來的所有後果，那麼框架問題就可以避免。有些技術或科學領域已經做到了這一點（所以人工智能科學家有時會聲稱框架問題已經被解決。或者，如果他們特別嚴謹，會說「或多或少」得到了解決）。然而，一般來說，這個問題並沒有解決，這也是人工智能系統缺乏常識的癥結所在。

簡而言之，框架問題就潛伏在我們周圍，是實現強人工智能的主要障礙。

註釋

[1]如無法像人一樣知道某些事與自己手頭上的事情相關聯。——譯者注

智能體和分佈式認知

一個人工智能體（Agents）是一個獨立的（「自主的」）程序，有時可以比作膝蓋反射，有時可以比作一個微型大腦。電話應用程序或拼寫校正器可以稱為智能體，但一般情況下又不是，因為不同的智能體之間通常會進行合作。它們會利用自己十分有限的智能與其他智能體合作或並肩達到單憑自己的力量無法實現的目的。而且，智能體之間的相互作用與智能體一樣重要。

有些智能體系統的組織形式是分層控制的，可以說是統治者和被統治者。許多系統則是分佈式認知的典範，包括無分層控制結構的合作（因此有了上面兩種說法：「合作」和「並肩」）。沒有中心規劃，沒有自上而下的影響，也就沒有單個智能體處理所有相關知識的情況。

從自然角度來看，分佈式認知的例子有：螞蟻尋跡、船舶導航和人類的大腦。螞蟻尋跡來源於很多螞蟻個體的行為，它們在走路時會自動滴落（並跟隨）化學物質。類似地，船舶的導航和操縱來自船員的互動活動，甚至是船長也不一定具備所有必要的知識，而且一些船員具備的知識確實少得可憐。即使是單個大腦也包含分佈式認知，因為它集成了許多有感知的、激發性的和有情感的子系統（見第4章和第6章）。

從人工角度來看，分佈式認知的例子包括：神經網絡（見第4章）；人類學家的船舶導航計算機模型；情境機器人學、蜂群智能和群體機器人學方面的人工生命研究（見第5章）；金融市場（智能體是銀行、對沖基金和大股東）的符號人工智能模型；LIDA的意識模型（見第6章）。

分佈式認知的意識也有助於人機交互設計，例如協作式工作場所和計算機接口。因為如伊馮·羅傑斯（Yvonne Rogers）所言，它闡明

了「人、人工製品和技術系統之間複雜的相互依賴性。在運用傳統認知理論時，往往會忽略這些要素」。

那麼很顯然，人類水平的強人工智能包含分佈式認知。

機器學習

人類水平的強人工智能還包括機器學習。然而，這並不是說要和人類一樣。該領域的研究始於心理學家進行的有關概念學習和強化方面的工作。不過它現在取決於可怕的數學技術，因為所使用的知識表示都包含概率論和統計學可以說從心理學入手的研究已經遠遠落後了。一些現代機器學習系統與人類頭部內可能發生的事情之間少有或甚至沒有相似之處。但是，貝葉斯概率在該人工智能領域研究中的運用之廣，不亞於最近的認知心理學和神經科學理論。

今天的機器學習超級賺錢。它不僅用於數據挖掘，還用於處理大數據，因為超級計算機每秒可以做千萬億次計算（見第3章）。

一些機器學習使用神經網絡，但是大多數都依賴符號人工智能，並輔以統計算法。事實上，做事情的是統計學，而GOFAI是「管理者」，引導工人去工作場所。因此，一些專業人士把機器學習看作計算機科學或統計學，而不是人工智能。不過這並沒有明確的界限。

有些計算機科學家故意排斥麥卡錫最先提出的「人工智能」的概念，認為其哲學含義有問題（見第6章）。也有些科學家有意避而不談，因為他們不喜歡大多數人工智能研究屬於實驗性研究，而且相對來說不成體系。

機器學習分三種類型：監督式學習、非監督式學習和強化學習（這種劃分源於心理學，可能還涉及不同的神經生理機制，跨物種的強化學習包含多巴胺）。

在監督式學習中，程序員「訓練」系統，方法是為一系列輸入（標記的示例和無標記的示例）定義一組希望得到的結果，並且不斷反饋該系統是否完成所希望的結果。學習系統對相關特徵作出假設。只要系統分類出錯，它就相應地修正自己的假設。具體的錯誤消息至關重要（不只是反饋系統出錯了）。

在非監督式學習中，用戶不提供希望得到的結果或錯誤消息。學習由原則驅動，而原則是指共同發生的特徵產生它們以後將共同發生的期望。非監督式學習可以用來發現知識。程序員不需要知道數據中有什麼模式/分類，系統自己會找到它。

強化學習受獎勵和懲罰所驅動：反饋信息告訴系統它剛剛做的事情是好還是壞。通常，強化不只是二進制，還是由數字表示，如視頻遊戲中的分數。「它剛剛做了什麼」可能是單個決定（例如遊戲中的某個步驟），也可能是一系列決定（例如國際象棋走到最後將軍了）。在一些視頻遊戲中，每動一步，數字分數都會更新。在極其複雜的情況下，例如象棋，僅在許多決定完成之後才能得到成功或失敗的信號，用於信用分配的某一程序識別出那些最可能帶來成功的決定（進化人工智能是一種強化學習形式，其中成功由適應值函數監控，詳見第5章）。

符號機器學習通常假定（並不完全正確）學習的知識表示必須包含某種形式的概率分佈。許多學習算法假定（通常是錯誤的）數據中的每個變量相互獨立，具有相同的概率分佈。此i.i.d.假設是許多數學概率理論的基礎，而後者又是許多算法的基礎。數學家採用i.i.d.假設，讓數學運算變得更簡單。同樣，在人工智能中使用i.i.d.可以簡化搜索空間，這樣易於問題的解決。

然而，貝葉斯統計處理條件概率，其中的內容或事件並不獨立。這裡的概率取決於與域相關的分佈式現象。這種形式的知識表示不僅更加實際，而且要是出現新的現象，概率還可以發生變化。貝葉斯技術在人工智能以及心理學和神經科學中的作用不斷凸顯。有關「貝葉斯大腦」的一些理論（見第4章）通過使用非i.i.d數據，去驅動和微調在知覺和運動控制中的非監督式學習。

因為有各種各樣的概率理論，所以有許多算法都適用於不同類型的學習和不同數據集。例如，接受i.i.d.假設的支持矢量機（Support Vector Machine）廣泛用於監督式學習，特別是如果用戶缺乏該域的專業先驗知識的時候。如果不計特徵的順序（如在搜索單詞而不是短

語時），則可以用「詞袋」算法。如果不用i.i.d.假設，貝葉斯統計（「亥姆霍茲機器」）可以向分佈式數據學習。

大多數研究機器學習的專家使用現成的統計方法。這些方法的發起者備受行業好評：Facebook最近聘用了支持矢量機的創建者；在2013年和2014年，谷歌公司聘用了幾位深度學習的重要發起者。

基於多層網絡的深度學習（見第4章）是一大新進步，前景光明。利用深度學習，輸入數據中的模式能在各個層級中被識別出來。換句話說，深度學習發現多層知識表示，例如從像素到反差檢測器，到邊緣檢測器，到形狀檢測器，到對像部分，以及到對象。

谷歌研究YouTube時出現的貓臉檢測器就是一個很好的例子。還有最近《自然》（Nature）雜誌報道某強化學習程序（「DQN」算法）已經學會了玩經典的Atari 2600 2D遊戲。雖然這個程序的輸入只有像素和遊戲得分（並且提前知道的只有每場比賽的動作數量），但是在49場比賽中，它有29場超過了75%的人類學習者，並在22場比賽中勝過專業遊戲測試人員。

這一成就的進步空間有多大還有待觀察。雖然DQN算法偶爾能找到最佳策略，如按時間排序的動作，但如果規劃遊戲需要更長一段時間，那DQN算法就不能控制遊戲。

未來的神經科學可能會告訴人們如何改進這個系統。當前版本的DQN借鑒了休伯爾—威澤爾（Hubel-Wiesel）的視覺受體——視覺皮層中的細胞，只對運動或特定方向的直線作出回應（這也沒什麼大不了的，休伯爾—威澤爾的受體還啟發了伏魔殿——一種學習模式（見第1章）。更奇怪的是，此版本的DQN設計還借鑒了海馬睡眠期間的「經驗重放」行為。DQN系統與海馬一樣，存儲了過去的樣本或經驗，並在學習期間快速重新激活這些樣本和經驗。這個特徵十分關鍵，設計師指出，如果特徵出現問題，那系統的性能將「嚴重惡化」。

通用系統

Atari遊戲玩家總是能振奮人心，上《自然》雜誌也無可厚非，其中一部分原因是它又朝著強人工智能邁進了一步。這個算法在沒有使用手工知識表示的情況下習得大量技能，完成包含高維感官輸入的任務。

然而（如本章開頭所述），完整的強人工智能可以完成更多任務。打造一位高性能的人工智能「專家」已經很困難，打造一個人工智能「通才」更是難上加難（深度學習不是答案：它的狂熱追隨者承認，需要「新範式」將其與複雜推理結合，而後者是「我們沒有一點頭緒」的學術密碼）。所以，大多數人工智能研究人員放棄早期設想，轉向各色細分任務，他們取得的成就通常讓人驚歎。

也有強人工智能先驅們始終保持著雄心壯志，如紐厄爾和約翰·安德森（John Anderson）。他們分別於20世紀80年代初提出了SOAR和ACT-R這兩個系統。三十年過去了，這兩個系統仍在不斷完善（和使用）。然而，它們過度簡化了任務，只關注人類的一小部分能力。

1962年，紐厄爾的同事西蒙研究了一隻螞蟻在崎嶇地面上行走的之字形路徑。他說，螞蟻的每個動作都是螞蟻對其當時感知的情境作出的直接反應（這是情境機器人學的精髓，詳見第5章）。十年後，紐厄爾和西蒙所著的《人類問題求解》（Human Problem Solving）一書將人類的智力描述成和螞蟻的智力類似的東西。根據他們的心理學理論，知覺和微觀運動行為由在問題解決期間存儲在記憶中或新建的內部表示（IF—THEN規則或「產生式規則」）來補充。

他們說：「被視為行為系統的人類很簡單。」但是，突然出現的行為複雜性十分重要。例如，他們表示，只包含14條IF—THEN規則的系統能夠解決算式謎問題（例如在這個和式中，將字母映射到數字0到9：DONALD+GERALD=ROBERT，其中D=5）。有些規則處理目標或子目標的組織問題；有些規則指揮注意力（指向特定的字母或

列）；有些規則回想以前的步驟（中間結果）；有些規則辨識假啟動；還有些規則回溯，以從這些假啟動中恢復。

他們認為，算式謎是所有智能行為計算架構的典範，所以該心理學方法適合「通才」人工智能。1980年，紐厄爾與約翰·萊爾德（John Laird）和保羅·羅森布魯姆（Paul Rosenbloom）開發了成功導向型成就實現系統（Success Oriented Achievement Realized，以下簡稱SOAR）。總的來說，它是一個認知模型，它的推理整合了知覺、注意力、記憶、聯想、推理、類比和學習。像螞蟻一樣的（情境）反應結合了內在的深思熟慮。事實上，深思熟慮往往帶來反射性反應，因為以前使用過的子目標序列可以「分塊拼成」一個規則。

事實上，SOAR並不能模擬認知的所有方面。人們認識到了一些缺陷以後，就不斷將其完善。今天的SOAR有很多用途，從醫療診斷到工廠調度等。

安德森的認知架構思維的自適應控制系統（Adaptive Control of Thought，以下簡稱ACT-R）是混合系統，它結合了產生式規則系統和語義網絡（見第4章）。這些程序通過識別環境中的統計概率，模擬聯想記憶、模式識別、意義、語言、問題求解、學習、圖像和（自2005年以來）感知運動控制。ACT—R系列主要是科學人工智能領域中的一次操練。雖然商業機器學習已經不再關注心理學（機器學習的起源）方面的研究，但是ACT—R仍在繼續（最近還包含神經科學，例如，與「模塊化的」大腦系統類似的IF—THEN規則集）。

ACT-R系統的一個重要特徵是整合了程序性知識和陳述性知識[1]。有人可能不知道如何在幾何證明中運用歐幾里德定理，但他可能知道這個定理正確。ACT-R系統通過構造數百個新的產生式規則來學習如何應用一個命題性事實，這些新規則可以控制它在不同情況下的使用。它學習哪些目標、子目標和子次級目標……在哪些條件下是相關的，某一特定動作在不同情況下會導致哪些結果。總之，它通過做事來學習，並且它（像SOAR）可以將有序執行的幾條規則整合到一條

規則中。這就與人類專家和新手解決「同一個」問題的情況類似：一個是不假思索，一個是煞費苦心。

ACT-R系統有多種應用。它的數學導向（一款智能家教系統）提供個性化反饋，包括相關領域的知識以及問題求解的目標/子目標結構。得益於定量分塊，使數學導向的建議粒度隨著學生的學習進度而不斷變化。其他應用有NLP、人機交互、人類記憶力和注意力、駕駛和飛行以及視覺網絡搜索。

SOAR和ACT是強人工智能時代另一個早期嘗試——道格拉斯·萊納特的CYC的「同輩人」。它是一款符號人工智能系統，於1984年被推出，目前仍在升級當中。

到2015年，CYC不僅包含62000條「關係」，能夠鏈接其數據庫中的概念，還包括這些概念之間的數百萬條鏈接。這些關係包括存儲在大型語義網絡（見第3章）中的語義和事實關聯，以及樸素物理學上的無數事實——所有人類具備且與物理現象相關的非形式化知識（例如丟棄和溢出）。系統同時使用單調和非單調邏輯以及概率來推論數據（目前，所有的概念和關係都是手寫編碼，但是貝葉斯學習理論不止於此。因此，CYC能夠自己從互聯網上獲得知識）。

幾家美國政府機構都在使用CYC，包括美國國防部（例如監測恐怖團體）和美國國家衛生研究院，還有一些大型銀行和保險公司。CYC的較小版本OpenCyc作為各種應用程序的背景資源已被公開發佈，人工智能研究人員現在可獲得更簡潔的版本（ResearchCyc）。雖然OpenCyc定期更新（最近一次在2014年），但它只包含CYC數據庫的一小部分數據和一小部分推理規則。不過完整（或接近完整）的系統最終將在市面上出售。但是，如果不採取特別措施，它可能會落入壞人手中（見第7章）。

萊納特在1986年的《人工智能雜誌》（AI Magazine）中將CYC描述為「使用常識性知識來克服脆弱性和知識獲取瓶頸」。也就是說，CYC的主要任務就是解決麥卡錫預見的挑戰。如今在模擬「常識」推

理以及「理解」所處理的概念方面，它已經成為佼佼者（有些概念甚至連NLP程序也無能為力，詳見第3章）。

然而，它也有許多缺點。例如，不能很好地處理隱喻（雖然數據庫包括許多亡隱喻）。它忽略了樸素物理學中各種不同的方面。它的NLP雖然在不斷完善，但仍然有極大的進步空間。它尚未包含視覺。總之，就算它的目標是包羅萬象，但現在確實不包含人類知識。

註釋

[1]在認知心理學的知識學習的心理原理中，安德森將語文知識分為兩類：一類為陳述性知識（declarative knowledge），指事實性和資料性知識；另一類為程序性知識（procedural knowledge），指按一定程序操作從而導致結果的知識。——譯者注

夢想復興

紐厄爾、安德森和萊納特已經從人工智能這個舞台中消失了將近30年。然而，最近人們對強人工智能的興趣又開始明顯上升。2008年開始召開的人工智能年度會議，除了SOAR、ACT-R和CYC外，還出現了其他所謂的「通才」系統。

2010年，機器學習的倡導者湯姆·米切爾（Tom Mitchell）帶領卡內基梅隆大學機器學習研究團隊開發了NELL系統（NeverEnding Language Learner，意為永不停止的語言學習者）。這種「常識」系統通過不停地（在寫入時設定為5年）搜索網站以及通過接受公眾的在線更正來構建自己的知識體系。它可以基於自己的（未標記的）數據完成簡單推理，例如，運動員喬·布洛格斯（Joe Bloggs）在戴維斯隊，所以他打網球。它從具有200個分類和關係的本體開始（例如，專家，是由於……），5年後，它擴大了本體，並積累了9000萬個候選可信度，每個都有自己的置信水平。

壞消息是NELL系統並不知道你可以用字符串拉取對象，但不推送這些對象。事實上，所有強人工智能系統的推定常識都受到嚴重限制。臭名昭著的框架問題已經被「解決」了這種說法產生了嚴重誤導。

NELL系統現在有一個姐妹程序叫NEIL（Never-Ending Image Learner，意為永不停止的圖像學習者）。有些部分視覺強人工智能將邏輯一符號知識表示與類推或圖示表示相結合〔亞倫·斯洛曼（Aaron Sloman）早在多年前就已經開始這項研究，但對它的理解仍然不明朗〕。

此外，斯坦福大學研究所在研究CALO（Cognitive Assistant that Learns and Organizes，意為學習和組織的認知助教）的過程中意外開發了Siri應用程序（見第3章）。2009年，蘋果以2億美元收購Siri。當前類似的活躍項目包括斯坦·富蘭克林（Stan Franklin）的LIDA（我們

將在第6章進行討論）和本·格策爾（Ben Goertzel）的OpenCog，它們在豐富的虛擬世界中或向其他強人工智能系統學習事實和概念（目前LIDA和CLARION這兩個「通才」系統側重意識）。

更近的一個強人工智能項目於2014年開始，旨在構建「機器人道德能力的計算架構」（見第7章）。除了上述困難，它還將必須面對許多與道德相關的問題。

一個真正達到人類水平的系統將會面臨更多問題。難怪強人工智能是如此地難以捉摸。

缺失的方面

如今幾乎所有的「通才」系統都聚焦於認知。例如，安德森的目標是詳細說明「認知心理學中的所有子領域是如何相互聯繫的」（「所有」子領域？雖然他討論了運動控制，但他沒有討論有時在機器人學中起到重要作用的觸摸或本體感知）。真正通用的人工智能還將包括動機和情感。

一些人工智能科學家已經認識到了這一點。雖然馬文·明斯基和斯洛曼都沒有建立全腦模型，但都細緻描述了整個大腦的計算架構。

我們將在第3章中概述斯洛曼的MINDER焦慮模型。喬斯查·巴赫（Joscha Bach）借鑒了他的研究成果和迪特裡希·多納（Dietrich Dorner）的心理學理論，開發了MicroPsi——一種基於七種不同「動機」並在規劃和動作選擇中運用「情感」因素的強人工智能。它也影響了上述的LIDA系統（見第6章）。

要實現真正的強人工智能，只做到這些還遠遠不夠。明斯基的未來人工智能宣言——「走向人工智能」，既發現了障礙，也給出了承諾。許多障礙還有待克服。從第3章可以看到，人類水平的強人工智能仍然離我們很遙遠。許多人工智能專業人士不同意這個觀點。有人甚至說，強人工智能很快將成為超人工智能（ASI）——「S」指超人。與此同時，智人將淡出研究人員的視線（詳見第7章）。

03 語言、創造力和情感

人工智能的一些領域似乎特別具有挑戰性，如語言、創造力和情感。如果人工智能不能模擬它們，要實現強人工智能就好似做白日夢。

無論就上述哪一方面而言，我們所取得的成就都早已超出了人們的想像。即便如此，一些困難仍然顯著存在。這些典型的「人類」特徵只是在一定程度上被模擬（人工智能系統是否能夠具有真正的理解力、創造力或情感，我們將在第6章中討論。我們在此關注的是人工智能系統是否有可能擁有它們）。

語言

無數的人工智能應用程序使用NLP（自然語言處理）。大多數程度關注的是計算機對呈現給它的語言的「理解」，而不是計算機自己創造語言。因為對於NLP而言，創造比接受更困難。

其中的困難包括主題內容和語法形式。例如，我們在第2章中看到，熟悉的動作順序（「腳本」）可能用作人工智能故事發生（哄孩子上床睡覺的母親改變慣常動作）的緣由。這並不是說，背景知識表示就一定包含足夠的人的動機，所以它不一定能使故事變得有趣。要就一家公司不斷變化的財務狀況寫一份年度報告，買個系統就完事了，但是可以看出這個系統創造的「故事」非常無聊。計算機創作的小說和肥皂劇情節確實有——但是任何以細微的地方處理得好為評判標準的獎項往往都與它們無關（用人工智能進行翻譯或總結人類創造的文本，得到的譯文和總結在內容上可能更豐富，但還是因為源語文本是人類完成的）。

至於語法問題，計算機創作的散文有時在語法上就不正確，而且通常很不恰當。人工智能對畫圈打叉遊戲（井字遊戲）的描述能夠包含從句或子從句結構，很好地講述遊戲的具體步驟。但是，我們也能充分理解畫圈打叉遊戲的概率和策略。不過，要人工智能清楚地描述很多人類故事中主角的一系列想法或動作就沒那麼容易了。

再談談人工智能接受語言，有些系統十分簡單，甚至讓人覺得無聊：它們僅需要識別關鍵字（想想電子零售網站上的「菜單」），或者預測字典裡所列出的單詞（想想在編輯短信時自動彈出的匹配詞或句）。還有一些系統要複雜得多。

有些系統需要識別語音：要麼是單個詞（如自動電話購物），要麼是連續語音（如實時電視字幕和電話竊聽）。更有意思的是，後者的目標可能是挑選出一些特定詞（如炸彈和聖戰），以此抓住整個句子的意思。這絕對是NLP：首先必須區分出單詞本身，這些單詞是由

不同聲音發出來的，而且可能帶有不同地方的口音或外來口音（區分單詞在印刷文本中是免費的）。深度學習（見第4章）已經促使語音處理技術取得了巨大進步。

對整句的理解也有一些令人印象深刻的例子，如機器翻譯；從大量自然語言文本中挖掘數據；總結報紙和期刊上的報道；以及回答一些自由提問（頻繁用於谷歌搜索和iPhone的Siri應用程序）。這些系統真的可以欣賞語言嗎？例如，它們能處理語法問題嗎？

在人工智能早期，人們認為語言理解需要解析句法。於是研究人員花了很大力氣去編寫程序，以實現這一目標。20世紀70年代初期，特裡·維諾格拉德（Terry Winograd）在麻省理工學院寫的SHRDLU[1]就是一個典型案例。在此之後，無數先前沒聽說過或認為人工智能不可能實現的人開始關注人工智能。

該程序接受英語指令，該指令告訴機器人用彩色積木搭建結構，並計算出如何移動這些積木才能實現目標。它之所以影響深遠，原因很多，其中部分知識已經應用到了一般的人工智能領域。與此相關聯的點在於它能夠賦予複雜的句子詳細的語法結構，例如：如果你之前不知道奶奶的食譜是錯的，你打算在蛋糕中加多少雞蛋。（嘗試一下！）

就技術層面而言，SHRDLU不盡如人意，其中有許多程式錯誤，所以只為少數技藝精湛的研究人員使用。當時還出現了其他各種句法處理程序，但也沒能推廣到現實文本當中。總之，研究人員後來很快發現，複雜的句法分析對現成系統來說太難了。

除了句法外，在人類語言中，語境和相關性也很重要。當時沒有顯著成就表明人工智能能夠做好這兩點。

1964年，美國政府的確在自動語言處理咨詢委員會（Automatic Language Processing Advisory Committee，以下簡稱ALPAC）的報告中宣佈機器翻譯不可能實現。報告預測，看好其「錢」景的人為數不

多（儘管機器輔助人類翻譯也許可行），認為計算機將與句法作鬥爭，被語境擊敗了，而最為重要的是，對相關性一無所知。

這就像是給機器翻譯（實際上，它的資金來源一夜之間乾涸）和人工智能丟了一枚炸彈。大家普遍將報告解讀為：人工智能研究做了許多無用功。暢銷書《計算機和常識》（Computers and Common Sense）也聲稱（1961年）人工智能研究是在浪費納稅人的錢。報告的發佈似乎也證明了政府的高層專家同意這種觀點。當時兩所即將開設人工智能學院的美國大學也跟著取消了其計劃。

不過人工智能的研究工作並未因此中斷。幾年後，精通句法的SHRDLU閃亮登場，為GOFAI做了一次成功的辯護。但是質疑很快悄然而至。NLP研究的焦點也因此逐漸轉向語境而非句法。

20世紀50年代早期，一些研究人員開始重視語義語境。英國劍橋大學的瑪格麗特·瑪斯特曼（Margaret Masterman）研究小組用同義詞詞典而不是字典來處理機器翻譯（和信息檢索）。他們認為句法是「語言中非常膚淺和冗余的部分，被匆忙的人完全忽略了」。他們專注於詞叢，而不是單個單詞。他們沒有嘗試字對字的翻譯，而是搜索同義詞的相關文本。這樣就可以正確翻譯模糊詞（如果找到了同義詞的相關文本）。因此，bank可以（用法語）表示為rive或banque，這取決於語境是否分別包含諸如water（水）或money（錢）等詞。

有些詞的詞義不同（例如魚和水），但是常常同時出現。這些詞可以強化以同義詞詞典為基礎的語境法。時間證明事實的確如此。今天的機器翻譯除了區分各類詞彙層面的共性——同義詞（empty/vacant）、反義詞（empty/full）、歸屬關係（fish/animal）和包含關係（animal/fish）、同類關係（cod/salmon）以及部分/整體關係（fin/fish），還能識別主題共現關係（fish/water，fish/bank，fish/chips等）。

由此可見，總結、提問或翻譯自然語言文本不一定非得處理複雜的語法。今天的NLP更多依賴於「體力」（計算能力）而不是大腦（語法分析）。數學，特別是統計學，已經取代邏輯，機器學習（包

括但不限於深度學習）已經取代句法分析。這些NLP的新方法（從書面文本到語音識別）非常高效，所以在實際應用中，95%的成功率是可接受標準。

在現代NLP中，功能強大的計算機統計搜索海量（「語料庫」）文本（在機器翻譯中，這些是由人類配對的翻譯），以找到常見的和意料之外的單詞模式。它們可以知道魚/水、魚/蝌蚪、魚和薯條、鹽和醋的統計結果。NLP現在可以學習構建「詞向量」（如第2章中所述），來表示既定概念下該詞所有含義出現的概率雲。不過此處的關注點通常是詞和短語，而不是句法。語法沒有被忽略：文本在接受檢測的過程中，其中一些單詞將被（自動或手動地）賦予形容詞和副詞之類的標記。但是句法分析卻很少使用。

詳細的語義分析也不多。「組合的」語義用句法分析句子的含義；這種做法僅限於研究實驗室，沒有大規模應用。「常識」推理器CYC對概念（詞）的語義表示相對完整，因此能更好地「理解」它們（見第2章）。但這種應用也十分有限。

當前的機器翻譯倒是風生水起的。有些系統包含主題很少，但有些系統則包羅萬象。谷歌翻譯每天為超過2億名用戶提供各種主題的機器翻譯。SYSTRAN翻譯系統每天為歐盟（24種語言）、北約、施樂公司和通用汽車公司服務。

許多機器翻譯的譯文都近乎完美，如歐盟的文件（因為在源語文本中只用到有限子集的單詞）。儘管大多數機器翻譯存在問題，但還是很容易理解，因為博學的讀者可以忽略譯文中的語法錯誤和生硬的單詞——就像聽非母語人士說話一樣。有些機器翻譯出來的譯文只需人類稍作編輯和修改（而日語在翻譯前後需要大量編輯。如英語的過去時態vot-ed，日語沒有分段的單詞。而且日語的短語順序是顛倒的。匹配不同語系的語言對機器來說並非易事）。

簡而言之，人類用戶可以很好理解機器翻譯出的譯文。同樣，總結期刊文章的單語NLP程序經常能夠反映論文是否值得全文閱讀（完

美的翻譯基本不現實。例如，用日語說「要一個蘋果」需要反映對話者的社會地位，但在英語中沒有同類區別）。

人工智能應用程序上的實時翻譯不太成功，如Skype。因為系統必須識別語音，而不是書面文本（單個詞被清楚分開）。

NLP的另外兩個突出應用是信息檢索——加權檢索（由瑪斯特曼的研究小組在1976年發起）和數據挖掘。例如，谷歌搜索引擎搜索詞條的時候，通常用相關性對要搜索的詞條進行加權——這是在統計學層面而不是語義層面評估（即沒有「理解」）。數據挖掘可以找到人類用戶未意識到的單詞模式。它長期用於研究市場中的產品和品牌，現在（使用深度學習）用於「大數據」，即搜集起來的海量文本（有時是多種語言）或圖像，如科學報告、醫療記錄、社交媒體和互聯網上的詞條。

政府、政策制定者和社會科學家用大數據挖掘開展偵查和反間諜活動，以及監測公眾的態度，以此來瞭解不同群體變化的觀點並對其進行比較：男/女、年輕人/老年人、北方人/南方人等。英國智庫Demos（與薩塞克斯大學的NLP數據分析團隊合作）分析了數以千計有關厭女症、種族群體和警察的Twitter消息。通過搜索特定事件（twitcidents）發生之後突然發出的一些推文，可以發現公眾對「警方回應」的態度發生了什麼樣的轉變。

大數據NLP給出的結果是否有用尚無定論。數據挖掘（使用「情緒分析」）不僅能度量公眾興趣度，還能度量其評價語氣。然而，語氣這種東西不會直接說出來。例如，一則推文包含具有明顯貶損語氣的種族歧視字眼，機器由此解讀為「負面」情緒，但事實上可能並不表示貶損。法官在讀到它的時候可能會認為這個詞被用作（在這種情況下）群體身份的一種積極標記，也可能覺得它是中性描述（例如，拐角處巴基斯坦佬開的商店），並非侮辱或辱罵。根據迪莫斯（Demos）的研究發現，只有一小部分包含種族或民族術語的推文真正帶有挑釁意味。

人的判斷在這些情況下依賴於語境，例如推文中的其他詞。調整機器的搜索標準，以減少「負面情緒」歸屬可能是行得通的，但也可能行不通。搜索標準也往往頗具爭議。即使人和機器的標準一致，也很難確定語境中的哪些方面能證明人類的解讀合理。

在計算（甚至口頭的）方面確定相關性很難，這只是其中一例。

乍一看，兩個知名的NLP應用程序似乎與剛才的說法相矛盾，即蘋果的Siri和IBM公司的沃森。

Siri是基於規則的私人助理，是一款能說話的「聊天機器人」，可以快速回答許多不同的問題。可以訪問互聯網上的一切資源——包括谷歌地圖、維基百科、不斷更新的《紐約時報》以及出租車和餐館等當地服務列表，甚至還可以訪問功能強大的在線自動回答系統WolframAlpha，後者利用邏輯推理「想出」而不只是「找到」各種事實性問題的答案。

用戶口頭對Siri（逐漸適應語音和方言）發問，然後Siri利用網絡搜索和對話分析回答問題。對話分析研究人類如何就對話中的主題進行排序，以及如何安排它和人類之間的互動（如解釋和協商）。利用對話分析，Siri將思考「對話者想要什麼」「我應該如何回答」等問題，同時在一定程度上適應個人用戶的興趣和偏好。

簡言之，Siri似乎不僅對主題相關性敏感，而且對個人相關性也很敏感。從表面上看，它真的能讓人印象深刻。然而，它很容易給出荒唐的答案，如果用戶偏離事實的軌道，Siri也就會跟著失去方向。

IBM公司的沃森也專注於事實。它是處理大數據的現成資源（有2880個核心處理器），已經用在一些呼叫中心，通過改良，還用到了醫療領域中，如評估癌症治療。它不僅能像Siri一樣回答直截了當的問題，還可以處理在常識遊戲《危險邊緣》（Jeopardy）中出現的謎題。

在《危險邊緣》中，玩家不會被問到直接的問題，而是根據以答案形式提供的各種線索，以問題的形式做出正確的回答。例如，玩家被告知「1921年5月9日，這家『盡善盡美的』航空公司在阿姆斯特丹開設了第一個客運辦事處」，那麼他們的答案應該是「KLM（荷蘭皇家航空）是什麼？」

沃森還可以應對很多其他挑戰。它的《危險邊緣》遊戲版本不像Siri那樣能訪問互聯網（雖然它的醫療版本可以），不懂對話結構，也不能通過邏輯推理找到答案，然而它能對龐大但封閉的數據庫進行大規模並行統計搜索。數據庫中有各種文件，如無數評論和參考書，還有《紐約時報》等，裡面提供了各類事實，從麻風病到李斯特（匈牙利鋼琴家、作曲家）、從氫到九頭蛇等。在玩《危險邊緣》的時候，它的搜索由數百種反映遊戲中固有概率的特殊算法作指導。它還可以從其他人類對手的猜測中受益。

2011年，沃森在玩《危險邊緣》的時候，「明顯地」戰勝了兩位人類冠軍，這可以和它在IBM公司的表兄弟深藍（Deep Blue，見第2章）的表現（打敗了國際象棋大師卡斯帕羅夫）相媲美（「很顯然」，因為計算機瞬間作出反應，而人類需要一些反應時間，然後才會按蜂鳴器）。但它和深藍一樣都不能穩居冠軍寶座。

沃森有一次比賽失利的原因是，雖然它正確地將注意力集中在某位運動員的一條腿上，但是它忽略了在它的存儲數據中有一個關鍵事實——這個人少了一條腿。沃森不會再犯這個錯誤，因為程序員現在已經標記「缺失」這個很重要的詞，但它還會犯其他錯誤。即使在普通事實搜尋語境下，人們通常依賴的相關性判斷都超出了沃森的能力範圍。例如，憑一條線索找到耶穌的兩個門徒，他們的名字既是十大首選嬰兒名，又都以同一字母結尾。答案是「馬修」（Mathew）和「安德魯」（Andrew）——沃森立即給出了答案。人類冠軍也得到了這個答案。但他的第一想法是「詹姆斯」（James）和「猶大」（Judas），他回憶說，自己之所以排除了這個答案，是因為出於某種原因，他認為猶大不是一個流行嬰兒名。沃森就做不到這一點。

人類的相關性判斷往往沒有上面的例子那麼明顯，對於今天的NLP來說，這個判斷太微妙了。相關性是語言/概念版的「框架問題」（見第2章），都是難啃的硬骨頭。許多人會覺得讓非人類系統完全掌握它簡直是天方夜譚。難道僅僅因為包含的信息量過大且過於複雜，還是因為相關性是人類特有的生命形式？我們將在第6章對此展開討論。

註釋

[1]SHRDLU是一個用自然語言指揮機器人動作的系統，由維諾格拉德於1972年在麻省理工學院建立。——譯者注

創造力

創造力——產生新穎的、異乎尋常的以及有價值的想法或人工製品的能力——是人類智慧的頂峰，對實現人類水平的強人工智能也是必不可少的。但人們普遍認為它很神秘。現在我們連人類是如何產生的新穎想法都沒弄明白，更別提計算機了。

目前，對創造力的識別甚至都沒有統一的答案：人們對一個想法是否具有創造性通常會持不同意見。有些分歧點在於：它是不是真的很新穎，以及它在何種意義上是真的很新穎。一個想法可能只是對相關個體來說是新穎的，也有可能對整個人類歷史來說都很新穎（分別是「個體」和「歷史」創造力的典範）。無論哪種情況，它可能多多少少和前述觀點類似，會引發分歧。還有些分歧點是有關估價（包含功能意識，有時會是現象意識，參見第6章）。同一個想法，有的社會群體可能重視，而有的卻不一定（比如現在的年輕人會嘲笑任何仍然喜歡看瑞典流行演唱組合樂隊Abba DVD盤的人）。

人們通常認為沒有什麼有趣的人工智能可以體現創造力。但人工智能技術產生了許多在人類歷史上屬於新穎的、異乎尋常的以及有價值的想法。例如，它們被用在了發動機、藥品和各類計算機技術的設計過程中。

此外，人工智能概念還有助於解釋人類的創造力。借此，我們可以分出三種類型的創造力：組合型、探索型和變革型。三者包含不同的心理機制，能帶來不同的驚喜。

在組合型創造力中，常見的想法以不常見的方式組合在一起。例如，視覺拼貼、有詩意的圖像和科學類比（將心臟比作泵，原子比作太陽系）。新組合在統計學層面帶來意外發現——這在以前是不大可能做到的事情，就像一個冷門選手不大可能贏得德比（Derby）。但它淺顯易懂，所以有價值。價值大小取決於如何評判前文討論的相關性。

探索型創造力較為常見。它充分利用了一些有文化價值的思維方式（例如，繪畫或音樂的風格、化學或數學的子區域）。使用風格法則（主要是無意識地）可以產生新想法，就像英語語法可以生成新句子一樣。藝術家或科學家可能無條件地探索該風格的潛力，也可能刻意大力推行它的應用或對其進行測試，以瞭解它可能生成哪些想法。它甚至可能因為某一規則的些許變化（例如弱化/加強）而發生小變動。儘管這個結構很新穎，但仍然屬於常見的風格。

變革型創造力繼承了探索型創造力，如果現有風格受限，變革型創造力就會發生。一個或多個風格限制將被徹底改變（刪除、否定、補充、替換、添加……），因此生成了之前不可能生成的新結構。這些新想法堪稱異類，因為它們的出現像是天方夜譚。最初，它們可能晦澀難懂，因為以慣常思維方式很難完全理解。然而，如果新想法要被接受，它們就必須貼近慣常思維方式（有時這種接受要花很多年）。

三種創造力都發生在人工智能中——觀察者通常認為創造力是人類確定的（實際上是通過圖靈測試，見第6章），但可能沒有像人們預期的那樣多。

像組合系統就十分罕見。人們可能認為模擬組合型創造力很容易，畢竟沒有什麼比讓計算機在已經存儲的想法之間產生不常見的關聯更簡單了。這些關聯（在歷史上）通常很新穎，（在統計學上）也令人驚訝。但如果它們要有價值，就必須彼此相關。當然，我們也清楚，相關性沒那麼容易得到。例如，我們在第2章中提到了一些笑話生成程序，它們用笑話模板來幫助提供相關性。同理，符號人工智能基於案例的推理利用預編碼的結構相似性來構造類比。因此，這些程序的組合型創造力還結合了探索型創造力。

同時，人們可能認為人工智能無法模擬變革型創造力。這種想法也是錯誤的。任何程序確實只能做它可能有能力做的事情，但是進化程序是可以進化自身的（見第5章）。它們甚至可以評估自己新進化的想法，但前提是程序員提供了明確的挑選標準。這樣的程序通常用在追求新穎的人工智能應用上，例如設計新科學儀器或藥物。

然而，變革型創造力不是一條通向強人工智能的神奇之路。它幾乎不能保證產生有價值的結果。我們可以相信（在數學或科學中的）有些進化程序能夠找到最優方案，但許多問題不能由最優化來定義。變革型創造力之所以有風險，是因為以前已經接受的規則被打破了。所有新結構都必須進行評估，否則就會出現混亂。但是當前人工智能的擬合函數是由人類定義的：程序不能獨立改變/推斷出它們。

探索型創造力最適合人工智能。這類例子不勝枚舉。工程學中一些探索型的人工智能創新（如CYC的設計者設計的程序所生成的創新，見第2章）已被授予專利。對於技術熟練人員來說，他們不一定覺得獲得專利的想法就屬於創新，但這個想法可能是他們想要探索的風格。有些人工智能的探索能與人類取得的傑出成就相媲美——如按照肖邦或巴赫的風格創作音樂，又有多少人能做到這一點?

然而，即使是探索型人工智能也在很大程度上依賴於人的判斷。因為必須有人識別並清楚地說明風格化的法則。這通常很難。有位世界級專家在研究弗蘭克·勞埃德·賴特（Frank Lloyd Wright）的「草原式住宅」時，不再描述建築風格，宣稱它們「難以理解」。後來，一個可計算的「形狀語法」生成了無數個「草原式住宅」的設計，包括四十多個原創——這沒什麼不可信。但系統成功的根本原因還是人類分析師。只有當強人工智能自己能夠分析（藝術或科學中的）風格時，它的創造性探索才是「自己的作品」。儘管最近有一些（但不多）深度學習識別藝術風格的案例（見第2章和第4章），但它的確是一項艱巨的任務。

利用人工智能，人類藝術家開發了一種新的藝術形式——數字藝術（computer-generated，CG）。它涉及建築學、圖像、音樂，以及編排和運用不太理想的文學（因為NLP面臨句法和相關性方面的困難）。在數字藝術中，計算機不只是個工具，可以將其比作一支新畫筆，幫助藝術家們做他們自己本來可以做的事情。相反，如果沒有它，這項工作就不可能做到，或者甚至想都不用想。

數字藝術體現了上述三種創造力。由於上述原因，幾乎沒有任何數字藝術是組合型的。英國法爾茅斯大學教授西蒙·克爾頓（Simon

Colton）的The Painting Fool軟件製作了與戰爭相關的視覺拼貼畫，但是它也收到了特殊指令，被要求搜索數據庫中與「戰爭」相關的圖像。大多數數字藝術都是探索型或變革型的。

計算機有時通過執行藝術家編寫的程序，可以完全獨立地生成藝術品。哈羅德·科恩（Harold Cohen）的AARON程序獨立生成了線條圖和彩色圖像（有時創造的顏色絢麗多彩，所以科恩說，AARON是一個比他更優秀的五彩畫家）。

相比之下，在交互藝術中，藝術作品的最終形式部分取決於觀眾的輸入，當然，觀眾可能是無意間控制了發生的事情。有些交互藝術家將觀眾看作同他們一起創作的人，還有一些交互藝術家認為觀眾以各種方式無意間影響了藝術作品，於是將他們看作作品產生的起因〔歐內斯特·埃德蒙茲（Ernest Edmonds）等藝術家同時採用了這兩種方法〕。在以威廉·萊瑟姆（William Latham）和喬·麥考馬克（Jon McCormack）為代表的進化藝術中，計算機不斷生成/改變結果，但通常是由藝術家或觀眾挑選的。

總之，人工智能的創造力有很多應用。在科學或藝術的一些小角落裡，它有時可以和人類的創造力一決高下，甚至超過人類。但在一般情況下要和人類創造力匹敵就另當別論了。強人工智能仍然離我們很遙遠。

人工智能與情感

和創造力一樣，情感也被看作與人工智能格格不入的東西。除了直觀上覺得不可能，想想情緒和情感依賴於大腦中散佈的神經調節劑這一事實，構建情感的人工智能模型也似乎不太現實。

多年來，人工智能科學家們似乎也贊同這個觀點。他們忽略了情感，只有在20世紀60年代和70年代出現了幾個例外，如西蒙，他認為認知控制包含情感；還有肯尼斯·科爾比（Kenneth Colby），他為神經症和偏執狂構建了有趣的模型，雖然這是一個超級有野心的目標。

如今情況發生了變化。神經調節（在GasNets中，見第4章）已經被模擬。此外，許多人工智能研究小組都在研究情感。儘管大部分研究在理論層面很膚淺，但大多數都「錢」景光明，它們致力於打造「計算機伴侶」。

還有些人工智能系統是基於屏幕的機器人，有些是門診用機器人，在與用戶的交互中，不僅實用，還關注用戶的舒適度以及滿意度。大多數服務對象是老年人或殘疾人，包括初發性癡呆病患者。還有一些是嬰兒或交互式「成人玩具」。總之，包括電腦護工、機器人保姆和性玩伴。

另外，人機交互的例子包括：提醒用戶購物、吃藥和拜訪家人；幫助編寫個人日誌；安排和討論電視節目，如每日新聞；製作美食和飲料；取東西；監測生命體征（和嬰兒哭泣）；說一些色情話語以及做一些色情動作等。

這其中的很多任務都包含人類的情感。人工智能伴侶就體現在它們能識別人類用戶的情感或以明顯帶有情感的方式回應用戶。例如，用戶承受喪親之痛時，可能會得到一些機器的同情。

人工智能系統已經能夠用多種方式識別人類的情感。有些是生理的，如監測人的呼吸頻率和皮膚電反應；有些是口頭的，如注意說話的速度、語調和用詞；有些是視覺的，如分析面部表情。當前的方法都相對簡陋。用戶的情感不僅容易被遺漏，而且容易被曲解。

計算機伴侶的情感表現通常體現在口頭上。它基於詞彙以及語調（如果系統能生成語音的話）。但是，系統不僅密切注意用戶常用的關鍵詞，還以極其刻板的方式作出回應。對於用戶說過的東西（可能在日記中），它偶爾可能會引用由人類創作的相關言論或詩歌。但NLP所面臨的難題意味著計算機生成的文本在細節上很難做好。這些文本甚至可能不會被接受：用戶可能會因為機器人伴侶沒有人類的外觀而被激怒或感到沮喪。同樣，一隻咕嚕咕嚕叫的機器貓可能會討人嫌，而不是讓用戶覺得放鬆、舒服或滿足。

當然也有惹人疼的機器人伴侶：帕羅（Paro）是一隻可愛的交互式「海豹寶寶」，它有著迷人的黑眼睛和濃密的睫毛，是許多老年人和癡呆症患者的好伴侶（未來版本還可以監測人類的生命體征，並據此向人類看護人員發出警告）。

有些人工智能伴侶可以利用自己的面部表情，也可以用眼睛凝視，以看似富有情感的方式回應用戶。有些機器人有彈性「皮膚」，覆蓋在人類面部肌肉模擬物的上面，它的外形可以（向人類觀察者）顯示出多達十二種基本情感。基於屏幕的系統通常顯示虛擬角色的面容，其表情根據（他/她）可能經歷的情緒而發生改變。然而，所有這些事情都有可能（原文如此）陷入所謂的「恐怖谷理論」[1]中，即人們在遇到與人類極其相似但仍存在些許差異的生物時，就會覺得不舒服，甚至極為反感。因此，機器人或屏幕虛擬化身如果擁有似是而非的面孔，可能會讓人類覺得自己正在受到威脅。

為情感空虛的人提供上述類似人類伴侶關係的做法是否符合道德標準，目前尚無定論（見第7章）。當然，有些人機交互系統（例如帕羅）似乎能夠為一些人帶來快樂，甚至是持久的滿足感。如果沒有這些系統，有些人可能會覺得生活很空虛。但是這樣就足夠了嗎？

「伴侶」模型缺乏理論深度。專家們開發人工智能伴侶的情感是為了賺錢。他們沒去想怎樣讓「伴侶」用情感解決自己的問題，也沒有去瞭解情感在整個大腦運作過程中發揮什麼樣的作用。他們覺得情感是可有可無的附加物：他們忽視情感，除非在某些棘手的人造情境下，他們才不得不考慮。

這種不屑的態度當時瀰漫在整個人工智能領域，直到最近，情況才相對有所改觀。「情緒計算」之母羅莎琳德·皮卡德（Rosalind Picard）的「情感計算」把情感從20世紀90年代末期的「冷宮」中解救出來，不過她也沒有深究。

一直以來，情感被人工智能忽視（與西蒙富有洞察力的評論命運相似），其中一個原因是它沒有得到很多心理學家和哲學家的重視。換句話說，他們認為智能不需要情感。相反，他們覺得情感不利於解決問題，會破壞理性。「情感可以幫助一個人決定做什麼以及做這件事的最佳方法」的想法不合潮流。

情感最終會越來越重要，部分得益於臨床心理學和神經科學的發展。但它能進入人工智能領域離不開馬文·明斯基和亞倫·斯洛曼這兩位人工智能科學家。他們一直把大腦看成一個整體，而不是像大多數同事那樣，將自己的想法局限在智能領域內的某個小角落中。例如，斯洛曼正在進行的CogAff項目就關注情感在大腦計算架構中的作用。CogAff影響了於2011年發佈並仍在推廣中的LIDA的意識模型（見第6章），也啟發了20世紀90年代末由斯洛曼研究小組帶頭開發的MINDER程序。

MINDER程序模擬了獨自照顧幾個嬰兒的護士心中所產生的焦慮（功能方面）。「她」只有幾項任務：給嬰兒餵吃的；別讓嬰兒掉進路邊的溝裡；如果有嬰兒掉進去，「她」得將嬰兒送到急救中心。「她」只有幾個動機（目標）：給一個嬰兒餵吃的；如果已經有一個嬰兒在防護柵欄後面，「她」要再放一個；將一個嬰兒從溝中抱出並送去急救站；在溝邊巡邏；築圍欄；將一個嬰兒移至離水溝較遠的安全位置；如果當前沒有其他動機被激活，「她」就在托兒所周圍漫步。

所以，她比真正的護士簡單得多（雖然比典型的規劃程序更複雜，因為後者只有一個終極目標）。然而，「她」容易感到不安，而這種不安可以算得上是焦慮。

這位模擬護士必須對所處環境中發出的視覺信號作出適當回應。有些信號觸發（或影響）的目標比其他目標更緊急：如果一個嬰兒正在爬向水溝，而另一個嬰兒只是餓了，那麼「她」得先管爬向水溝的那個；此時如果剛好有一個嬰兒快掉進水溝了，那麼「她」的注意力得先轉向這個。但是就算有些目標當時被擱下了，可能最終還是必須解決，它們的緊迫程度可能會隨著時間的推移而不斷增強。所以，如果有一個嬰兒在水溝附近，那麼「她」可以先把餓了的嬰兒放回嬰兒床；但是「她」應該先給餵食等待時間最長的嬰兒餵吃的，然後再喂不久前剛餵過的嬰兒。

總之，模擬護士的任務有時可能被中斷，要麼被放棄，要麼被擱置。MINDER程序必須決定當前的優先級。「她」在完成任務的整個過程中必須做這種決定，這樣「她」的行為可能會因此而被不斷改變。事實上，任何任務在完成過程中都會被中斷，因為環境（嬰兒們）對系統提出了很多相互衝突且不斷變化的要求。模擬護士和人類護士一樣，會因為嬰兒（每個嬰兒是不可預測的自主智能體）數量不斷增加而變得越來越焦慮，表現得也越來越差。不過這種焦慮很有用，護士因此能成功地照顧嬰兒。不過這個過程並不順利：冷靜和焦慮相距甚遠。

MINDER程序表明了一些情感控制行為的方式，從而智能地安排相互競爭的動機。毫無疑問，人類護士會因為情況發生變化而經歷（原文如此）各種焦慮。但這裡的重點是情感（emotions），不只是感覺（feelings）。前者還涉及現象和功能意識（見第6章）。具體來說，它們是被安排了競爭動機的計算機制，如果沒有這些機制，我們就無法運作。所以，影視劇《星際迷航》中沒有情感的斯波克（Spock）先生就無法進化成真正的人。

如果要實現強人工智能，那麼我們必須考慮和利用情感因素，如焦慮。

註釋

[1]恐怖谷理論（uncanny valley）是一個關於人類對機器人和非人類物體的感覺的假設，它在1969年被提出，說明了當機器人與人類相像超過一定程度的時候，人類對它們的反應便會突然變得極為反感，即哪怕機器人與人類有一點點的差別都會顯得非常顯眼刺目，從而對整個機器人有非常僵硬恐怖的感覺，猶如面對行屍走肉。——譯者注

04 人工神經網絡

人工神經網絡（artificial neural networks，ANN）是由許多相互連接的單元組成，每個單元只能計算一件事情。這樣的描述聽起來有些無聊。但人工神經網絡似乎很有魔力。它必然也讓記者們著迷。弗蘭克·羅森布拉特的「感知器」（光電機）在沒有接受明確指導的情況下可以學會識別字母，曾在20世紀60年代成為各大報紙的「寵兒」，吸睛無數。20世紀80年代中期，人工神經網絡名聲大振，至今仍然備受媒體的青睞。最近與人工神經網絡相關的大量宣傳還包括深度學習。

人工神經網絡有無數應用，從操控股票市場和監測貨幣波動到識別語音或人臉，但真正有趣的是它們的運行方式。

有一小部分人工神經網絡在特定的並行硬件上運行，甚至在硬件或濕件混合物上運行，將真正的神經元與硅電路結合。然而，大多數網絡通常由約翰·馮·諾依曼機器來模擬。也就是說，人工神經網絡是在經典計算機上實現的並行處理虛擬機（見第1章）。

之所以說它們的運行方式很有趣，部分原因是它們與符號人工智能的虛擬機有很大差別。大規模的並行計算代替串行指令，自下而上的處理代替自上而下的控制，以及概率代替邏輯。動態和持續變化的人工神經網絡與符號程序形成了鮮明對比。

此外，許多神經網絡從隨機開始時就具有神秘的自組織屬性（20世紀60年代的感知器也具有這一屬性，它們在新聞中都很高調）。系

統從隨機架構（隨機權重和聯結）開始，並自己逐漸適應去執行需要完成的任務。

人工神經網絡有許多優點，顯著增強了人工智能的計算能力。然而，它們也有缺陷，即它們不能提供第2章中所設想的真正意義上的強人工智能。例如，雖然一些人工神經網絡可以做近似推理或推理，但它們不能像符號人工智能那樣精確（問：2+2是多少？答：很可能是4。真的嗎？）。在人工神經網絡中，也很難模擬層級。一些（回饋式）網絡在一定程度上能夠用交互式網絡來表示層級。

由於當前對深度學習的熱情高漲，神經網絡的網絡不像以前那樣罕見了。但是，這些網絡還是相對比較簡單的。人腦必須包括無數在很多層級上以極其複雜的方式進行交互的網絡。總之，強人工智能仍然十分遙遠。

人工神經網絡更廣泛的含義

人工神經網絡是作為計算機科學的人工智能取得的一大勝利，但它們的理論含義並未僅局限於此。它們與人類概念和記憶有一些相似之處，因此引發了神經科學家、心理學家和哲學家們的興趣。

神經科學家們的興趣由來已久。事實上，羅森布拉特沒有把具有開創意義的感知器當作一個在現實中有用的小發明，而是把它當作一條神經心理學理論。今天的神經網絡儘管與大腦有很多差異，但是在計算神經科學中扮演著重要的角色。

心理學家也對人工神經網絡感興趣，哲學家們緊跟其後。例如，非專業人工智能人士狂熱追捧20世紀80年代中期的一個神經網絡。該網絡顯然已經像小孩一樣學會了使用過去時，開始的時候沒有犯錯誤，後來由於過分遵守規則，以至於把英文「go」的過去時變成了「goed」（本來應該是「went」），在犯了這些錯誤之後，最後才得到規則和不規則動詞的正確形式。這是有可能做到的，因為提供給該網絡的輸入反映出小孩經常聽到的詞的變形概率——神經網絡沒有使用先天的語法規則。

這個網絡的出現意義重大，因為當時大多數心理學家（和許多哲學家）都接受了諾姆·喬姆斯基（Noam Chomsky）的說法，聲稱小孩必須依靠先天的語言規則來學習語法，以及嬰幼兒過分遵守規則的行為恰好證明那些規則在起作用。過去時態的神經網絡證明了這兩種說法都不正確（當然，它沒有證明小孩不具備先天的規則，只是證明了他們不需要這些規則）。

另一個十分有趣的例子是對「表徵軌跡」（representational trajectories）的研究，最初受到了發展心理學的啟發。原先混亂的輸入數據在連續的層級上被重新編碼（在深度學習中也一樣），所以除了能捕獲到明顯的規則性外，還有不太明顯的規則性顯露出來。這不僅涉及兒童的發展，還涉及與歸納學習相關的心理學和哲學爭論。它

表明有了先前的期望（計算結構），才能學習輸入數據中的模式，學習不同模式的順序必然受到約束。

簡而言之，人工神經網絡在商業和理論層面都很重要。

分佈式並行處理

有一種人工神經網絡理論尤其吸睛——PDP。事實上，當人們提及「神經網絡」或「聯結主義」（不常用的術語）時，他們通常說的就是PDP。

PDP網絡的運行方式主要有四大優勢，涉及技術應用和理論心理學（以及精神哲學）。

第一，它們只需要被顯示例子而不需要被精確編程，就能學習模式以及各個模式之間的關聯。

第二，能容忍「凌亂的」跡象，可以解決約束滿足問題，弄清部分衝突跡象的意思。它們不需要嚴格定義（不需要被表示成一系列的充分必要條件）。相反，它們處理具有家族相似性的重疊集合，這也是人類概念的一個特徵。

第三，能夠識別不完整的或部分損壞的模式。也就是說，它們具有找尋內容的記憶。人類也有這種記憶，例如，聽到開頭幾個音符就能識別一段旋律，或者即使是一段旋律幾個音彈錯了，也能識別這段旋律。

第四，它們很強健。PDP網絡就算丟失一些節點，也不會滿口胡言或停止運行。它可能顯示出適度的退化，在這個過程中，它的性能隨著損害的增加而逐漸變差。它們不像符號程序那麼脆弱。

這些優勢源於PDP中的D，即「分佈式」。不是所有的人工神經網絡都涉及分佈式處理。在集中式（中央控制式）網絡（例如WordNet，見第2章）中，所有概念分別是由唯一的節點表示。在分佈式網絡中，一個概念存儲（分佈）在整個系統中。有時候，集中式和分佈式處理會結合，但不常見。單純的集中式網絡也不常見，因為它們沒有PDP的四大優勢。

分佈式網絡本質上是集中式網絡，因為它的每個單元都相當於單個微特徵——例如，視野中某個位置的一小塊顏色（不是太小，也不是太特別。可以證明一些粗調單元比許多精調單元更有效率）。但是與需要表示的概念相比，這些單元在更低級概念層上定義：PDP包含「子符號」計算。此外，每個單元可以是許多不同整體模式的一部分，因此促成了許多不同的「意義」。

PDP系統有多種類型。所有類型都是由三層或更多層互通單元組成的，每個單元只能計算一個簡單的東西。但這麼多單元連在一起，那情況就完全不一樣了。

無論何時，只要輸入層中單元的微特徵被呈現給網絡，它就會被觸發。當輸出單元被與之連接的單元激活，它就會被觸發，它的活動也會被傳遞給人類用戶。位於中間層的隱藏單元與外界沒有直接接觸。有些隱藏單元是確定的：它們是否被觸發只取決於與其連接的單元產生的影響。有些隱藏單元則是隨機的：它們是否被觸發，部分取決於某個概率分佈。

各個連接也不一樣。有些是正向反饋，將信號從低層傳遞到高層；有些則是反向的；還有一些是橫向的，連接同一層內的各個單元；有些既有正向反饋，也有反向反饋，我們將在下文中提到。就像大腦突觸一樣，連接要麼是興奮，要麼是抑制。它們的強度（權重）不同。權重表示為+1和-1之間的數字。興奮（或抑制）連接的權重越高，接收信號的單元被觸發的概率越高（或越低）。

PDP包含分佈式表示，因為每個概念都是由整個網絡的狀態表示的。這似乎令人困惑，甚至自相矛盾。它當然和符號人工智能中定義的表示有很大差異。

只對技術或商業應用感興趣的人不在乎這一點。如果他們只想解決實踐中的一些突出問題，例如，單個網絡怎樣才能存儲幾個不同的概念或模式，那麼他們很樂意先把問題留在那兒！

研究人工智能心理和哲學含義的人也問到了這個「突出問題」。答案就是一個PDP網絡可能出現的所有狀態千變萬化，所以在這塊或那塊單元中，只有少數幾種狀態同時被激活。被激活的單元僅僅將活性擴散到一些其他單元。然而，那些「其他單元」各不相同：任意既定單元能夠促成許多不同的活性模式（一般情況下，帶有許多未激活單元的「稀疏」表示更有效）。系統將最終飽和，關聯存儲器的理論研究問題將是一定規模的網絡原則上能夠存儲多少模式。

這些人不喜歡把問題留在那兒。他們不僅對「表示」本身的概念感興趣，還熱衷於討論人類的心智/大腦是否確實包含內部表示。PDP的追隨者認為PDP源於符號人工智能，迅速傳播到心智哲學，駁斥了物理符號系統假說（見第6章）。

神經網絡學習

大多數人工神經網絡能夠學習。這包含在權重中作出自適應改變，有時也包含在連接中作出自適應改變。通常，網絡的解剖結構（單元的數量以及單元之間的連接）是固定的。如果是這樣，學習只改變權重。但有的時候，學習或進化（見第5章）能夠增加新的連接和修剪舊的連接。構造型神經網絡將這一點做到了極致：開始時沒有任何隱藏單元，隨著學習的不斷深入，它們增加了隱藏單元。

PDP網絡有很多學習方式，涵蓋了第2章中提到的所有學習類型：監督式學習、非監督式學習和強化學習。例如，在監督式學習中，某個類的一些例子輸入到PDP網絡以後，網絡逐漸識別這個類——所有例子都不需要具有每一個「典型的」特徵（輸入數據可以是視覺圖像、語言描述、數字集合等）。當一個例子輸入給PDP網絡的時候，一些輸入單元對「輸入數據」的微特徵作出回應，同時活性擴散，直到網絡穩定下來。然後將輸出單元的所得狀態與期望輸出作比較（由人類用戶辨別），權重隨之變化（可能由反向傳播引起），從而減少誤差。在輸入了許多略有差別的例子後，屆時PDP網絡將開發出典型的活性模式或「原型」，即使之前沒有被輸入這樣的典型模式（就算現在輸入一個受損的例子以及激活更少的相關輸入單元，該模式也能夠自動完成）。

大多數人工神經網絡學習基於赫布理論，它由唐納德·赫布（Donald Hebb）於1949年提出，該理論經常被總結為「一起激發的神經元連在一起」。赫布理論學習強化了經常使用的連接。當兩個連在一起的神經元同時被激活的時候，權重就會被調整，上述情況以後更有可能實現。

赫布用了兩種方式來表示赫布理論，這兩種方式既不精確也不等效。今天的人工智能研究人員用許多不同的方式來定義它，理論基礎有可能是從物理學中得到的微分方程，也可能是貝葉斯概率理論。他們利用理論分析來比較和改進各種版本。PDP研究可能屬於萬惡的數

學範疇。因此，有大量物理學和數學專業的優秀畢業生在金融機構工作，這也正是他們的都市同事們為何很少真正瞭解他們的系統在做什麼的緣故。

鑒於PDP網絡用赫布型學習定律來調整權重，那麼它何時能夠停止？答案不是它什麼時候能達到完美狀態（消除了所有的不一致），而是它什麼時候能實現最大程度的一致性。

例如，如果相關單元同時發出信號，表示通常不同時顯示出來的兩個微特徵，這時就會出現「不一致的情況」。許多符號人工智能程序可以解決約束滿足問題，通過消除路徑上跡象之間的矛盾來獲取解決方案。但這些程序不能容忍解決方案中出現不一致的情況。PDP系統則不同。前文提到了PDP的優勢，即使不一致持續存在，它們也可以成功運行。它們的「解決方案」是給出網絡在不一致性降至最低時而非消除時的總體狀態。

可以借用熱平衡概念來做到這一點。物理學中的能級用數字表示，PDP中的權重也一樣。如果學習定律與物理定律相類似（而且如果隱含單元是隨機的），同一個描述統計行為的波爾茲曼方程可以描述能級和權重的變化，也可以借用快速但均勻冷卻金屬的方法。退火過程為：先將固體充分加溫至足夠高，再讓其慢慢冷卻。PDP研究人員有時使用模擬退火算法，前幾個平衡週期中的權重變化遠大於後面週期中的權重變化。網絡因此能夠擺脫和之前相比已經實現總體一致性的情況（局部極小值）；但是如果系統被干擾，網絡甚至可能達到更好的一致性（以及更穩定的平衡狀態）。就像我們搖動一袋彈珠，如果想要強行倒出袋子內脊處的所有彈珠，那我們開始應該使勁甩這個袋子，但是越到後面用力應該越小。

用反向傳播實現最大程度的一致性，其速度更快，使用範圍更廣。但是，無論採用哪種學習規則，整個網絡（特別是輸出單元）的平衡狀態被看作有關概念的表示。

反向傳播、大腦和深度學習

PDP的追隨者們認為，人工神經網絡比符號人工智能更接近大腦。PDP的設計靈感確實來自大腦，一些神經學科學家確實用它來模擬神經機能。然而，人工神經網絡還是與大腦中的東西有很大差異。

（大多數）人工神經網絡和大腦之間的一個區別是反向傳播或BP。BP是一個學習規則，或者說是一類通用學習規則，經常用於PDP。1974年，保羅·韋伯斯（Paul Werbos）首次使用BP，20世紀80年代初，傑弗裡·辛頓（Geoffrey Hinton）以更有效的方式定義BP，它解決了信用分配問題。

所有類型的人工智能都面臨信用分配問題，系統不斷變化的時候尤其如此。假定有一個複雜的人工智能系統很成功，那麼它的哪些部分最有可能促成了它的成功？在進化人工智能中，信用通常由「桶隊列」算法分配（見第5章）。在具有確定（非隨機的）單元的PDP系統中，信用通常由BP分配。

BP算法從輸出層到隱藏層追溯系統成功的原因，並辨識一些需要被調整的個體單元（更新權重以將預測誤差降至最低）。當網絡給出正確答案時，BP算法需要知道輸出層的確切狀態（因此BP是監督式學習）。在該示範性輸出和從網絡中實際獲得的輸出之間進行逐個單元比較。輸出單元的狀態在兩種情況下的任何差異都被視為誤差。

算法假定，一個輸出單元的誤差由與其連接的單元中所出現的誤差引起。算法在整個系統中進行反向推算，依據隱藏層中的單元與輸出單元之間的連接權重，確定第一個隱藏層中的每個單元所產生的具體誤差量。對於上述輸出單元的誤差，由與該單元連接的所有隱藏單元分攤責任（如果一個隱藏單元連接到多個輸出單元，那就把它的所有單個責任相加，作為該隱藏單元的責任）。然後對隱藏層和前一層之間的連接按比例改變權重。

「前一層」可以是另一個（和另一個……）隱藏層，但最終它會成為輸入層，權重將不再變化。該過程一直重複，直到輸出層的誤差降至最低。

多年來，BP算法僅用於只有一個隱藏層的網絡。多層網絡很罕見，它們很難分析，甚至很難做實驗。不過最近因為出現了深度學習，BP算法被用在了多層網絡上，並引起了人們的巨大興奮，還有一些不負責任的炒作。這裡的系統學習深入到一個域的結構，而不只是表面模式。換句話說，它發現了多層而非單層的知識表示。

深度學習令人興奮的原因是它至少為人工神經網絡處理層級保駕護航。自20世紀80年代初以來，辛頓和傑夫·埃爾曼（Jeff Elman）等聯結主義者為了表示層級——曾將集中式表示和分佈式表示結合，還定義過遞歸神經網絡（遞歸網絡實際上是執行一系列離散步驟。利用深度學習，最新版本的遞歸網絡有時能夠預測句子中的下一個單詞，甚至是段落中的下一個「想法」）。但是他們取得的成就有限（人工神經網絡仍然不適合表示精確定義的層級或演繹推理）。

深度學習也出現在了20世紀80年代〔由於爾根·施密德胡貝（Jürgen Schmidhuber）發起〕。但是，直到最近辛頓提供了一種讓多層網絡在多個層級上發現關係的有效方法後，這個領域才有了進一步發展。他的深度學習系統由在六個層級上「受限制的」玻爾茲曼機（沒有橫向連接）組成。首先，所有層級進行非監督式學習。它們利用模擬退火算法逐一接受訓練。一層的輸出用作下一層的輸入。當最後一層穩定以後，整個系統由BP微調，向下通過所有層級，為它們適當分配信用。

認知神經科學家和人工智能技術人員一樣對深度學習感興趣。因為它規定的「生成模型」能夠學習預測網絡輸入的（可能會發生的）原因——模擬了亥姆霍茲在1867年提出的觀點：「知覺為無意識推論。」也就是說，知覺不是指被動地接收來自感官的輸入，它包含主動理解，甚至提前預測。簡而言之，眼睛/大腦不是相機。

辛頓於2013年加入谷歌公司，所以BP也會頻繁亮相。谷歌公司已經在許多應用中用到了深度學習，包括語音識別和圖像處理。此外，它於2014年收購英國人工智能公司DeepMind。DeepMind公司的DQN算法結合深度學習和強化學習（見第2章），征服了經典的Atari遊戲。IBM公司也對深度學習頗感興趣：它不僅用在了沃森上，還被許多專家應用程序借用（見第3章）。

然而，雖然深度學習很實用，但這並不代表對它的理解就很到位。大量實驗探索了不同的多層學習規則，但理論分析仍十分混亂。

其中一個問題就是是否有足夠的深度來獲得近乎人類的表現（第2章中提到的貓的臉部單元是由一個九層系統產生的）。例如，人類視覺系統有7個解剖層，但是大腦皮層中的計算到底增加了多少層？因為人工神經網絡受到了大腦的啟發（在深度學習炒作中不斷強調的點），所以這個問題必然會被問到，但它貌似不是很切題。

BP算法是計算層面取得的一項勝利，但不是生物學上的突破。BP算法過程不會產生大腦中貓的臉部「祖母細胞」（見第2章），但深度學習可以。真正的突觸純粹是前饋：它們不進行雙向傳播。大腦包含不同方向的反饋連接，但每個連接都是嚴格的單向。這只是真實神經網絡和人工神經網絡之間的一個差異（還有很多）。另一個差異是真實的神經網絡不是按照嚴格的層級構成，即使視覺系統經常按照嚴格的層級來描述。

大腦包含正/反向連接的事實對於構建感覺運動控制的預測編碼模型至關重要，這引起了神經科學專家們的極大興趣（這些也主要基於辛頓的成果）。較高的神經層級向下層發送消息，預測來自傳感器的輸入信號，而且向上層發送的只有不可預測的「錯誤」消息。此類重複循環微調這些預測網絡，使後者逐漸學會該預測什麼。研究人員提到了「貝葉斯大腦」，因為按照貝葉斯統計，預測可以被理解，而且計算機模型中的預測根基實際上就是貝葉斯統計（見第2章）。

與大腦相比，人工神經網絡的結構過於規整簡單，層級太少，且枯燥無聊。過於規整，是因為人類建立的網絡優先考慮數學層面的美

觀和功能，而生物進化的大腦則完全不同；過於簡單，是因為單個神經元——有大約30種不同類型，計算起來和一個PDP系統或一台小型計算機一樣複雜；層級太少，是因為即使具有數百萬單位的單元，人工神經網絡與人類大腦相比也只能是小巫見大巫（見第7章）；枯燥無味，是因為人工神經網絡研究人員通常不但忽略時序因素（如神經脈衝頻率和同步性），還忽略樹突棘、神經調節劑、突觸電流和離子通道這些生物物理學因素。

這四個缺點都在不斷完善。由於計算能力增強，人工神經網絡能夠包含更多的單個單元。更為詳細具體的單個神經元模型正在構建當中，已經在解決上述所有神經學因素的計算功能問題。「枯燥無味」在模擬中甚至在現實中都在改善（一些「神經形態」研究將活的神經元與硅芯片結合）。與DQN算法模擬視覺皮層和海馬睡眠期間的過程一樣（參見第2章），未來的人工神經網絡無疑將借用神經科學的其他功能。

然而，仍不可否認的是，人工神經網絡和大腦在很多重要方面千差萬別，有些差異甚至我們現在還不知道。

網絡醜聞

人們為PDP的到來感到興奮的主要原因是，20年前的人工神經網絡研究（也稱為聯結主義）被宣佈為「死胡同」。正如第1章所述，這一判斷是伴隨著馬文·明斯基和西摩爾·派普特（Seymour Papert）在20世紀60年代發表的尖銳批評文章而來的（兩位專家在人工智能領域都德高望重）。到了20世紀80年代，人工神經網絡似乎不只是一個「死胡同」，而是真的「斷了氣」。的確，控制論基本上已經被邊緣化（見第1章）。幾乎所有的研究資金都流向了符號人工智能。

一些早期的人工神經網絡看似前途一片大好。羅森布拉特的自組織感知機一直被記者們惦記著，能夠學著識別模式，即使它們開始學習的時候是一種隨機狀態。羅森布拉特滿懷信心，聲稱自己的方法有涵蓋人類心理各個方面的潛力。當然，他也指出了一些局限，但是他提出的「收斂定理」已經確保簡單的感知機能夠學做任何可能（用程序）指令它們去做的事情。這是強有力的證據！

但是馬文·明斯基和西摩爾·派普特在20世紀60年代末也提供了證據。他們從數學層面證明簡單的感知機並不能做人們直觀上希望它們做的事情，而GOFAI可以輕鬆做到。和羅森布拉特的收斂定理一樣，他們的證明只適用於單層網絡。但是他們二位的「直觀判斷」是，多層系統將被組合爆炸擊敗。換句話說，感知機不會按比例增加。

大多數人工智能研究人員接受了聯結主義注定失敗的觀點。不過還是有一些人繼續研究人工神經網絡。事實上，他們在分析聯想記憶上取得了一些非常顯著的成就，例如克裡斯托弗·龍格-希金斯（Christopher Longuet-Higgins）和大衛·威爾肖（David Willshaw），後來還有詹姆斯·安德森（James Anderson）、圖沃·科霍恩（Teuvo Kohonen）和約翰·霍普菲爾德（John Hopfield）。但是這項工作並沒有向世人公開。相關研究人員並沒有將自己稱為「人工智能」研究人員，而那些自稱為人工智能研究人員的人並沒把他們放在眼裡。

PDP的到來給這種懷疑態度迎頭一擊。除了一些令人印象深刻的功能模型（如過去時學習程序），還有兩個新的收斂定理：一個保證了基於玻爾茲曼方程[1]的PDP系統能夠達到平衡狀態（儘管可能是在很長一段時間之後）；另一個證明了三層網絡原則上能夠解決表示給它的任何問題（友情提示：和在符號人工智能中的情況一樣，以可以輸入到計算機的方式表示問題通常是最難的部分）。人們對人工神經網絡的熱情自然隨之而來，而對主流人工智能達成的共識也因此被擾亂。

符號人工智能曾經假定毫不費力的直覺思維正如有意識的推論，但其實沒有意識。現在，PDP研究人員認為二者完全不一樣。帶頭研究PDP的專家們〔大衛·魯姆哈特（David Rumelhart）、傑·麥克利蘭（Jay McClelland）、唐納德·諾曼（Donald Norman）和辛頓〕指出，二者對人類心理學都至關重要。但是對PDP的宣傳以及公眾的反應，意味著研究心智的符號人工智能是在浪費時間。人工神經網絡這個不起眼的研究這次來了個徹底大翻身。

美國國防部（人工智能的主要資助者）對此的態度也產生了360度的大轉彎。在1988年召開緊急會議之後，他們承認自己以前對人工神經網絡的忽視是不應該的。至此，大筆資金源源不斷地投入到了PDP研究中。

不過明斯基和派普特仍然頑固不化。他們認可「基於網絡的學習機器（原文如此）以後能帶來的益處超乎想像」。然而，他們堅持認為高級智能不能從純粹的隨機性或完全無序的系統中產生。因此，大腦有時必須充當串行處理器，人類水平的人工智能必須採用混合系統。他們抗議稱「他們的批評不是導致人工神經網絡被冷落的唯一因素」，首要原因是計算能力不足。他們認為自己沒有一直試圖將研究資金轉移到符號人工智能。他們說：「我們認為我們的工作沒有殺死白雪公主，我們認為這是理解她的一種方式。」

這聽起來像是令人信服的科學論證。但他們最初發表的批評文章的確尖酸刻薄（初稿甚至更毒：友好的同事勸他們語氣緩和一點，多

突出一下科學論點）。這篇文章讓人有想法當然不足為奇。堅持不懈的人工神經網絡研究人員極其反感他們新發現的文化缺位（支持符號人工智能的人放棄了人工神經網絡）。PDP甚至引起了更大的躁動。人工神經網絡的「死亡」和復興包含嫉妒、自負、自我吹噓和幸災樂禍：「我們早就告訴過你了。」

這一小節提到了一個典型的科學醜聞，但不是人工智能中出現的唯一例子。理論分歧捲入了個人情感和較量，公正的思維十分罕見。到處都瀰漫著尖酸刻薄的辱罵和打壓。人工智能研究不是缺乏激情的事業。

註釋

[1]非熱力學平衡狀態的熱力學系統統計行為的偏微分方程。——譯者注

連接不是一切

對於人工神經網絡的大多數描述都暗含這樣一層意思：與神經網絡相關的唯一重要的事情就是它的解剖結構。哪些神經元與哪些神經元相關聯，以及權重有多強？這些問題當然十分重要。然而，最近的神經科學表明，由於化學物質在大腦中擴散，生物學回路有時能夠改變神經元的計算功能（不僅僅是讓它基本形成）。

例如，一氧化氮向各個方向擴散，它的效果取決於相關點的濃度會一直存在，直到它下降（下降速率由酶改變）。因此，一氧化氮作用於給定體積的腦皮層內的所有細胞，無論這些細胞是否由突觸連接。神經系統的功能動態表現與「純粹的」人工神經網絡有很大差別，因為「廣播」信號取代了點對點信號。一氧化碳、硫化氫和複雜分子（如5-羥色胺和多巴胺）有類似效果。

人工智能懷疑者可能會說：「對人工神經網絡的討論應該就此打住！」「計算機裡面沒有化學組成！」這種說法很荒謬，你不能說計算機不能模擬天氣的理由是因為計算機裡面不能下雨吧。然後這些懷疑者又可能說：「人工智能無法模擬情緒或情感。」這個反對意見是由心理學家烏爾裡克·奈瑟爾（Ulric Neisser）在20世紀60年代早期提出的。幾年後，哲學家約翰·哈格爾（John Haugel）發表了一篇頗具影響力的「認知主義」批評文章，也提出了反對意見。他們說人工智能可以模擬推理，但是絕不能模擬情感。

然而，一些人工智能研究人員受到這些神經科學研究結果的啟發，設計了一種全新的人工神經網絡，在這個人工神經網絡中，連接不是一切。在GasNets技術中，散佈在網絡中的一些節點能夠釋放模擬的「氣體」。這些氣體可以擴散，並根據濃度以不同方法調節其他節點和連接的固有屬性。和擴散源的形狀（模擬成空心球，而不是點源）一樣，擴散體積的大小很重要。因此，給定節點在不同的時間表現也不一樣。在確定的氣體條件下，儘管沒有直接連接，一個節點還是會影響另一個節點。真正關鍵的是氣體與系統內電氣連接之間的相

互作用。同時，由於氣體僅在某些特定場合下發出，而且以不同的速率擴散和衰減，所以這種複雜的相互作用在不斷變化。

例如，GasNet技術用來進化自主機器人的「大腦」。研究人員發現，一個特定的行為可能包含兩個未連接的子網，由於調節作用，二者能一起工作。他們還發現，將紙板三角形用作導航輔助的「方向檢測器」能夠以部分不連接的子網形式進化。他們曾建立了一個整體連接的網絡來做到這一點（見第5章），但神經調節網絡進化得更快、更高效。

因此，一些人工神經網絡研究者不再只考慮解剖結構（連接），還開始研究神經化學。他們現在在模擬不同的學習規則及其短暫的相互作用時，還會考慮神經調節。

神經調節是一種模擬現象，不是數字現象。擴散分子的濃度要一直變化。越來越多的人工智能研究人員（使用特殊的VLSI芯片）正在設計能同時體現模擬功能和數字功能的網絡。模擬功能模擬生物神經元的解剖結構和生理機能，包括穿過細胞膜的離子通道。例如，這種「神經形態」計算被用於模擬知覺和運動控制的內部結構。有些人工智能研究人員計劃在「全腦」模擬中使用神經形態計算（見第7章）。

有些專家甚至走得更遠，他們不是純粹在硅上建立人工神經網絡模型，而是構建（或進化，見第5章）由微型電極和真實神經元組成的網絡。例如，當電極X和Y都接受人工刺激時，在「濕」網絡中產生的活動導致一些其他電極Z激發——這樣就執行了一個「與」門。這類計算（唐納德·麥凱曾在20世紀40年代設想過）現在處於萌芽階段，但它可能會給大家帶來驚喜。

混合系統

剛剛提到的模擬/數字網絡和硬件/濕件網絡被描述為「混合」系統，這可以理解。但這個術語通常用來指既包含符號又包含聯結主義信息處理的人工智能程序。

明斯基在其早期宣言中已經說過，可能有必要包含二者，一些早期的符號程序確實結合了串行處理和並行處理。但這種嘗試很罕見。我們也看到在PDP到來後，明斯基繼續推崇符號和人工神經網絡的混合系統。然而，這樣的系統並沒有立即出現（儘管辛頓建立了集中式和分佈式聯結主義相結合的網絡來表示部分/整體層級，如家庭樹）。

的確，符號處理和神經網絡處理結合的系統仍然不常見。一個重邏輯，一個重概率，這兩種方法差異巨大，大多數研究人員只擅長一種。

然而，一些真正意義上的混合系統已經被開發出來了，其中的控制在符號模塊和PDP模塊之間適度傳遞。因此，這種模型吸收了兩種方法的互補優勢，例如DeepMind開發的博弈算法（見第2章）。它們將深度學習與GOFAI結合，以學習如何玩一套視覺上多樣化的電腦遊戲。它們使用強化學習，沒有提供手工制定的規則，只有每個步驟的輸入像素和數值分數。許多可能的規則/計劃被同時考慮，最有希望的規則/計劃決定下一個動作〔未來的版本將專注於3D遊戲，如《我的世界》（Minecraft），以及應用程序，如無人駕駛汽車〕。再如全心智系統ACT-R*和CLARION（見第2章）以及LIDA（見第6章）。它們大多涉及認知心理學研究，開發是基於科學目的，而不是技術目的。

有些混合模型考慮神經系統的特定內部結構。例如在1980年，臨床神經病學專家提摩西·沙麗斯（Timothy Shallice）與PDP先驅諾曼共同提出了一個與常見（「過度學習」）動作相關的混合理論，並得到了實施。該理論解釋了一些常見的錯誤。例如，中風患者常常忘記應該先把信件放入信封再舔粘舌，或者他們可能在上樓換衣服時卻去睡

覺了，又或是拿起水壺而不是茶壺。我們所有人偶爾都會犯與順序、捕獲和對像替換相關的類似錯誤。

但為什麼腦損傷患者特別容易犯這類錯誤？沙麗斯的計算理論聲稱，常見動作由兩種控制產生，能夠在特定點上分解或接管：一是無意識的「爭用調度」，它包含以層級形式組成的動作模式之間的（無意識）競爭，激活超過某個閾值的動作模式接管控制；另一個（「執行」）控制機制是有意識的，它包含審慎監督和調整第一種機制，包括規劃和修復錯誤。對於沙麗斯而言，由PDP模擬爭用調度，執行控制則交由符號人工智能。

動作模式的激活水平能夠通過感知輸入提高。例如，某人到達臥室時無意瞥了一眼床（模式識別），便可觸發其上床的動作模式，即使其最初的意圖（計劃）是換衣服。

沙麗斯借鑒人工智能（特別是規劃模型）的觀點提出了動作理論，這與他的臨床經驗相吻合。大腦掃瞄得到的證據進一步佐證了他的理論。最近的神經科學發現了其他因素，包括與人類動作相關的神經遞質。這些因素已經通過當前基於動作理論的計算機模型顯現出來。

爭用調度和執行控制之間的相互作用也與機器人學相關。遵循計劃的智能體應該能夠基於自身在環境中觀察到的東西停止或改變計劃。這個策略體現了機器人的特徵，即結合情境處理和審慎處理（見第5章）。

任何對強人工智能感興趣的人都應該注意到，有極少數人工智能科學家將心智的計算架構看作一個整體，他們毫無保留地接受了混合主義。如艾倫·紐厄爾和安德森（第2章討論的SOAR和ACT*）、斯坦·富蘭克林（Stan Franklin，第6章中概述的LIDA的意識模型）、明斯基（心智的「社會」理論）和亞倫·斯洛曼（第3章中描述了他對焦慮的模擬）。

總之，在我們的大腦中實現的虛擬機既是串行的，也是並行的。人類智力需要二者的巧妙合作。假如能實現人類水平的強人工智能，那麼它也會如此。

05 機器人和人工生命

人工生命模擬生物系統。和人工智能一樣，它有著技術和科學雙重目的。人工生命對於人工智能而言不可或缺，因為已知的所有智能都可以在生物體上找到。的確，有許多人相信心智只能由生命產生（見第6章）。冷靜的技術人員不擔心這個問題。不過他們確實在開發多種實際應用的時候考慮了生物學。這些應用包括機器人、進化編程和自組織設備。機器人是人工智能的經典例子：它們的出鏡率很高，而且具有獨創性，還具有很大的商機。儘管進化人工智能應用很廣，但沒有那麼出名。知道自組織機器的人甚至更少（非監督學習除外，見第4章）。然而，在理解自組織的過程中，生物學對人工智能來說很有用，同樣，人工智能對於生物學來說也很有用。

情境機器人和有趣的昆蟲

機器人造於幾世紀以前，列奧納多·達·芬奇是一位製造專家。人工智能機器人首次亮相於20世紀50年代。第二次世界大戰後，威廉·格雷·華特的機器「烏龜」因知道避開障礙並尋找光而艷驚四座。麻省理工新建了人工智能實驗室，其主要目標是整合計算機視覺、規劃、語言和運動控制技術，以打造一款「麻省理工機器人」。

後來的發展可謂突飛猛進。現在，有些機器人能夠爬山、爬樓或爬牆；有些跑得快；有些跳得高；有些能夠搬運重物；有些能夠投擲重物。還有些能夠自我拆解和自我重新組裝，有時還能夠組裝成新的形狀，如一條蠕蟲（能夠穿過一根狹窄的管道）或一個球，又或是多腿生物（適合於水平或粗糙地面）。取得這些進步的原因是人們將研究重點從心理學轉到了生物學上。

經典人工智能機器人模仿人類的自發動作。根據一些大腦模型理論，經典人工智能機器人採用了世界和智能體自身動作的內部表示。它們沒有給人們留下深刻的印象，因為它們依賴於抽像規劃，常常碰到框架問題（見第2章）。它們不能做到及時反應，因為即使是一絲環境變化也需要它們預先規劃並重新開始；它們也不能適應新（未模擬的）情況。即使在平坦整潔的地面上，平穩運動對它們來說也是難事（因此SRI機器人的暱稱為SHAKEY，它與「SHAKY」諧音，後者的意思是「搖晃」），而且機器人一旦跌倒後就無法恢復正常。在大多數建築物中，它們都屬於無用之物，那麼就更別指望它們去火星上做事了。

而今天的機器人卻煥然一新。焦點也已經從人類轉變為昆蟲。昆蟲可能不夠智能，不能模擬世界或做規劃，但它們可以管理。它們的行為合乎時宜且適應性強，對，此處就是指行為而非動作。但這主要是一種反射（習慣性思維）而不是熟思。昆蟲們不假思索地對當下情境作出回應，而不是想像中可能發生的一些事情或目標狀態。因此，這些機器人得到了這樣的標籤——「情境」或「基於行為」的機器人

（情境行為不僅限於昆蟲，社會心理學家在人類身上也發現了許多與情境有關的行為）。

為了賦予人工智能機器類似的反射，機器人專家們喜歡工程學勝過編程。如果可能，感覺運動反射在機器人的解剖結構中是實實在在地被具化了，而不是以軟件代碼的形式。

機器人的解剖結構到底要和生物體的解剖結構有多匹配，至今尚無定論。就技術目的而言，巧妙的工程學技巧可以接受。今天的機器人有許多不現實的噱頭。但也許是生物學機制的效率特別高？的確，它們足夠高效。因此，機器人學專家還會考慮真正的動物：它們能做什麼（包括它們的各種導航策略），涉及什麼樣的感官信號和具體運動，什麼樣的神經機制在起作用。生物學家反過來又用模型去研究這些神經機制——一個名為計算神經行為學的研究領域。

一個例子是蘭德爾·比爾（Randall Beer）的柔性機器人學。蟑螂有六條帶活關節的腿，這既是優勢也是劣勢。六足動物運動的時候比雙足動物更穩定（通常比輪子更管用）。然而，協調六肢似乎要比協調兩肢更困難。它不僅要決定接下來該移動哪條腿，還必須用正確的力量，找到正確的位置和時機。各條腿應該如何互動？它們必須非常獨立，因為可能只有一條腿旁邊有一個卵石，但是如果那條腿被抬高，其他腿必須做補償動作以保持平衡。

比爾的機器人反映了真實蟑螂的神經解剖結構和感覺運動控制。它們能夠爬樓梯，在粗糙的地面上行走，翻過障礙物（而不只是避開它們），並在摔倒後恢復原有姿勢。

芭芭拉·韋伯（Barbara Webb）則在蟋蟀身上找靈感。她的關注點不是移動（所以她的機器人能夠使用輪子）。相反，她希望自己的機器人可以識別、定位和靠近特定聲音模式。顯然，這種行為（「趨聲性」）應該會有很多實際應用。

雌性蟋蟀一聽到同種雄性蟋蟀唱的歌就會有趨聲行為。然而，蟋蟀只能識別用一個速度和頻率唱出來的一首歌曲。蟋蟀的種類決定速

度和頻率。雌性蟋蟀不會在不同的歌曲之間做選擇，因為它沒有為一系列聲音編碼的特徵檢測器，而使用只對一個頻率敏感的機制。該機制不是神經機制，它類似於人類大腦內的聽覺檢測器。它是一根長度固定的管，長在胸部，連接到前腿上的耳朵和氣孔。管的長度與雄性唱出來的歌曲的波長成精確比例。物理學能夠確保以下兩點：1.（管中的空氣和外部空氣之間的）相位抵消，僅在歌曲頻率正確時才出現；2.強度差完全取決於聲源的方向。雌性蟋蟀的神經系統強制它做出朝著這個方向移動的行為——雄性唱歌，雌性就跟著走。這就是貨真價實的情境行為。

韋伯之所以選擇研究蟋蟀的趨聲性，是因為它得到了神經學專家的密切關注。但仍有許多問題尚未找到答案：歌曲的方向和聲音是不是（以及如何）獨立處理的；歌曲的識別和定位是不是獨立的；是如何觸發雌性蟋蟀行走的；以及如何控制它的「之」字形方向。韋伯還設計了能夠產生類似行為的最簡單機制（只有四個神經元）。之後，她的模型用到了更多的神經元（基於詳細的真實生命的數據），包括額外的神經特徵（例如延遲、放電速率和膜電位），並且將聽覺與視力結合。她的工作不僅澄清了許多神經科學問題，回答了一些問題，還提出了更多問題。所以她的工作對機器人學和生物學都有幫助。

雖然機器人是物理實體，但許多機器人學的研究都是在模擬中完成的。例如，比爾的機器人有時先在軟件中進化，然後才被製造出來。同樣，韋伯的機器人先被設計成程序，然後才在現實世界接受測試。

雖然主流機器人學的研究轉向了昆蟲，但是對人形機器人的研究仍在繼續。有些只是玩具。有些是家用「社交」或「伴侶」機器人，供老年人或殘疾人使用（見第3章）。專家們設計這些機器人的目的不只是讓它們成為拿送東西的奴隸，而是成為獨立的私人助理。有些機器人長相「可愛」，有長長的睫毛和誘人的聲音，能夠和用戶進行眼神交流，並能識別對方的面部和聲音。此外，它們在一定程度上可以獨立進行未提前設定的對話，解讀用戶的情感狀態，並產生「帶有情感的」回應（類似於人的面部表情或語音模式）。

雖然有些機器人體型龐大（用來搬運重物或穿過粗糙地面），但是大多數很嬌小。例如一些在血管內使用的機器人就是微型的。通常大量這種機器人一起工作。問題是只要有一個任務由多個機器人完成，那麼就可能會出現問題，比如它們之間如何溝通；如何讓團隊能夠完成個體不能單獨完成的任務。

為了回答這些問題，機器人專家通常會考慮群居的昆蟲，比如螞蟻和蜜蜂。這些物種是「分佈式認知」（見第2章）的典範，它們的知識（與合適的動作）分散於整個群體，而不是由任何一個動物獨享。

如果機器人太過簡單，那麼它們的開發者就有可能會談到「群體智能」，還會分析作為細胞自動機（CA）的協作機器人系統。細胞自動機是由多個個體單元構成的系統，單個單元會遵循簡單規則，以採用有限狀態中的一種狀態，這就取決於其相鄰單元的當前狀態。一個細胞自動機的行為的整體模式可能超級複雜，這就好比多細胞生物體中的活細胞之間的相互合作。許多人工智能機器人用到了在好萊塢動畫片中成群蝙蝠或恐龍使用的畜群算法。

分佈式認知和群體智能的概念也適用於人類。如果參與的個體不能處理相關知識，那麼就會用到群體智能（例如大批人群的整體行為）；如果參與的個體可以擁有所有相關知識但實際上卻沒有擁有，那麼就會用到分佈式認知。例如，一個人類學家已經展示了導航知識是如何在船員之間共享的，同時是如何具化到實物上的，如圖表和海圖桌（的位置）。

說知識是具化到實物上的東西，聽起來可能很奇怪，或者最多也就用了隱喻的手法。但是，如今有大把人聲稱，人類的心智已經完全被具化到了人類的具體動作中了，而且還被具化到人類用來吸引外部世界的文化藝術品中。這種「外部的/具化的智能」理論在一定程度上是基於麻省理工的羅德尼·布魯克斯（Rodney Brooks）所做的工作，他是將機器人學的研究焦點從「人轉向昆蟲」的第一人。

布魯克斯現在是美國軍事機器人的主要開發者。在20世紀80年代，他還是一個初出茅廬的機器人專家，當他看到當時一些符號人工智能模擬世界的不切實際的規劃程序時，他感到十分沮喪。所以，他後來單純為了技術目的而改變了自己的研究方向，著手研究情境機器人學。不久之後，他就將自己的方法發展成了一個適應性行為理論。他的研究範圍遠遠超出了昆蟲。他認為，即使是人類動作也不包含內部表示（他有時暗示，不經常包含表示）。

他對符號人工智能的批判引起了心理學家和哲學家的興趣。有些人和他產生了極大的共鳴。心理學家指出，很多人類行為受情境制約，例如在不同社會環境中的角色扮演。認知心理學家強調有生命的視覺，那麼對於視覺而言，智能體的身體運動是關鍵。如今，具化的心智理論在人工智能之外的領域頗具影響力（見第6章）。

但是大衛·基爾希（David Kirsh）等人仍然堅決反對布魯克斯的觀點，他們認為復合表示對於涉及概念的行為來說必不可少。比如認識感知恆定性的時候，可以從不同角度識別目標；一段時間之後重新識別個體；預想的自我控制（規劃）；協商而不只是調度相互衝突的動機；反事實推論；以及語言。這些批評者們承認，情境機器人學表明了不涉及概念的行為要比許多哲學家所想的更普遍。然而，邏輯、語言和認真考慮過的人類動作都需要符號計算。

許多機器人專家也反對布魯克斯的一些更為極端的論述。研究足球機器人的阿蘭·邁科沃斯（Alan Mackworth）團隊提到了「反應式的深思熟慮」，它包括感官知覺、制定實時決策、規劃、識別規劃、學習和協調。他們試圖實現GOFAI和情境觀點的一體化（也就是說，他們正在構建混合系統，詳見第4章。）

一般來說，表示對於機器人學中選擇動作的過程至關重要，但對於動作的執行不太重要。所以，如果有人戲稱「人工智能」現在代表「人工昆蟲」，那麼這種說法並不完全正確。

進化人工智能

大多數人認為，人工智能需要一絲不苟的設計。鑒於計算機的無情天性，要是沒有一絲不苟的設計，它怎能這樣？其實，它能夠這樣。

例如，進化機器人（包括一些情境機器人）是通過組合嚴格的編程/工程學和隨機變異而產生的結果。它們的進化未經預測，也沒有精心的設計。

進化人工智能通常有這樣一個特點：它從符號人工智能中產生，但也用於聯結主義。它有許多實際應用，包括藝術（該領域可能歡迎不可預測性）和開發對安全苛求的系統（例如飛機發動機）。

程序能夠自我改變（而不是由程序員重寫），甚至可以用遺傳算法（GA）完善自身。受真實生命遺傳學的啟發，程序能夠隨機變異和非隨機選擇。選擇需要成功的標準或「適應度函數」（與生物學中的自然選擇類似）與遺傳算法一起工作。於是，定義適應度函數成了關鍵。

在進化軟件中，面向任務的初始程序既不能有效完成任務，也可能根本無法完成該任務，因為這個任務可能是一個不連貫的指令集或隨機連接的神經網絡。但是，整個程序包括後台的遺傳算法。

遺傳算法能夠改變面向任務的規則。這些隨機變化與生物學中的點突變和交叉類似。因此，程序指令中的單個符號可能被改變，或者兩個指令的短符號序列可能被「交換」。

任何一代中的任務程序都會被比較，用那些最成功的任務程序繁殖下一代。少數（隨機選擇的）其他程序也可能被保留，這樣就不會丟失那些尚未具有任何良好效果的潛在有用突變。經過一代又一代的比較選擇，任務程序的效率不斷提高。有時能在這個過程中找到最優

解。在一些進化系統中，信用分配問題用約翰·霍蘭德（John Holland）的「桶隊」算法的一些變形算法來解決，以辨識讓這個複雜的進化程序成功的是它的哪些部分，見第4章。

一些進化人工智能是完全自動的：程序在每一代都使用適應度函數，並且在無監督情況下進化。必須清楚定義這裡的任務，如用飛機發動機物理學來定義。相比之下，進化藝術的交互性通常很強（藝術家在每一代篩選出最好的「任務程序」），因為其不能清楚闡明適應度函數——美學標準。

大多數進化機器人學是間歇性交互。機器人的解剖結構〔例如傳感器和感覺運動連接或它的控制器（「大腦」）〕自動進化，但是這個過程是在模擬中進行的。大多數進化都沒有用到物理機器人。但是在進化到第500代的時候，進化的設計會在物理設備上進行測試。

無用的突變往往不能存活。薩塞克斯大學的研究人員發現，如果任務不需要深度視覺或觸覺，機器人的一隻「眼睛」（共兩隻）和所有的「晶須」都可能會失去其與控制神經網絡的初始連接（同樣，無論是先天性聾人，還是被剝奪聽覺輸入的動物，他們的聽覺皮層都用於視覺計算：大腦終身都在進化，而不只是進化幾代）。

進化人工智能還有可能帶來更多的驚喜。例如，一個不斷進化的情境機器人（也在薩塞克斯）在靠近目標的過程中能夠做出迴避障礙物的動作，它「長出」了一個定位檢測器，和大腦中的定位檢測器相類似。機器人的世界要包括一個白色紙板三角形。讓人出乎意料的是，一個隨機連接的微型網絡將出現在控制器中，該控制器對在某一特定方向漸變的亮/暗（三角形的一邊）作出回應。然後這個微型網絡進化成視覺動作機制中不可分割的一部分，機器人利用它（最初隨機）與動作單元的連接，將紙板三角形用作導航輔助。動作機制對黑色三角形和三角形的另一邊都不適用。它單獨存在，沒有全面的定位檢測器系統，不過它很有用。總體來說，這個結果可以重複。薩塞克斯團隊使用不同類型的神經網絡後發現，每個成功的解決方案都進化了某種主動定位檢測器。因此，高級行為策略也一樣（確切的實施細節可能會有所不同，但往往非常相似）。

薩塞克斯團隊還用遺傳算法設計硬件電路。任務是進化振蕩器。然而，最後出現的卻是一個原始的無線電波傳感器，從附近的計算機拾取背景信號。這取決於不可預測的物理參數。有些參數可以預測（所有印刷的電路板類似空氣的性質），雖然該團隊以前忽略了這一點。其他參數卻是偶然的，而且看起來不相關，例如，與計算機的空間距離；模擬開關的設置順序；以及留在附近工作台上的烙鐵插入電源的事實（此結果不會重複，無線電天線下次可能受到牆紙化學反應的影響）。

無線電波傳感器很有趣。要知道許多生物學家（和哲學家）認為人工智能不會出現全新的東西，理由是計算機程序的所有結果（包括遺傳算法的隨機結果）必須屬於它所定義的可能性空間。他們說只有生物進化能夠產生新的知覺傳感器。他們承認，一個無效的人工智能視覺傳感器可以進化得更好。但是，他們說，第一個視覺傳感器只會在因果關係控制的物理世界中出現。產生光敏化學物質的隨機遺傳突變，可以將外部世界已經存在的光帶入到生物體的環境中。然而，這個出人意料的無線電傳感器，同樣會將無線電波帶入設備的「環境」中。從一定程度上來講，它的確取決於物理原因（插頭等）。但它是人工智能的一次練習，而不是生物學。

人工智能中要有全新的東西出現確實需要外部影響，因為不可否認的是，程序不能超出其可能性空間。但這些影響不一定是物理的。連接到互聯網的遺傳算法系統可能通過與虛擬世界交互，進化出徹徹底底的新東西。

進化人工智能另一個更早的驚喜仍然激勵著進行中的進化研究。生物學家托馬斯·雷（Thomas Ray）用遺傳算法模擬熱帶雨林的生態。他看到了寄生蟲的自然湧現、對寄生蟲的抵抗和能夠克服這種抵抗的超級寄生蟲。他還發現，一系列微小的基礎（基因型）突變可以造成（表型）進化中的突然「劇增」。正統的進化論者當然相信這一點。但是，該觀點違反常理，所以史蒂芬·傑·古爾德（Stephen Jay Gould）等生物學家認為，必須考慮非進化論的過程。

今天，模擬突變率仍處在系統地變化和追蹤階段，遺傳算法研究人員正在分析「適應值曲面」「神經（原文如此）網絡」和「遺傳漂變」。在突變尚未增加生殖適應度的情況下，如何維持突變？這項工作作了解釋。總之，人工智能在幫助生物學家提出進化理論。

自組織

生物有機體的關鍵特徵是它們具有組織自我的能力。自組織是指從不太有序的起源到有序結構的自發出現。這種性質讓人不解，甚至自相矛盾。它能否在非生物上發生，這一點尚不明顯。

廣義上說，自組織是一種創造性現象。我們在第3章中討論了創造力（包括「歷史」和「個體」），在第4章中討論了自組織的（非監督）聯想學習。這裡，我們將重點討論生物學研究的自組織類型。

這方面的例子有動植物種類史的進化（歷史創造力）；胚胎發育和變態（與心理學的個體創造力類似）；腦發育（先是個體創造力，然後是歷史創造力）；細胞形成（生命開始時為歷史創造力，然後是個體創造力）。人工智能如何幫助我們理解它們?

艾倫·圖靈通過追本溯源解釋了自組織。他問道：「同質事物（如未分化的卵子）為何會創造結構？」他承認，大多數生物發育為現有順序增加了新順序，例如胚胎神經管中變化的順序。但是從同質性中得到的順序是基本（數學上最簡單的）情況。

胚胎學家也設想了「組織者」：以未知方式指導發育的未知化學物質。圖靈不認同組織者，而是考慮和化學擴散相關的超級通用原則。他指出，如果不同的分子相遇，結果將取決於它們的擴散速率、濃度，以及它們之間的相互作用對分子的破壞/構建速度。他的證明方法是改變設想的化學方程中的數字並研究結果。有些數字組合只產生形體不明的化學品混合物。但其他數字組合則產生順序，例如，某一分子濃度的常規峰值。他說，這樣的化學峰值可能在生物學上被表示為表面標記（條紋），或重複結構的起源，如花瓣或身體的部分。三維中的擴散反應能夠產生空心化現象，如早期胚胎中的原腸胚形成。

這些想法立刻引起了巨大的反響。它們解決了以前很難解決的難題，即無序的起源如何產生順序。但是20世紀50年代的生物學家用它

們還做不了什麼。圖靈用的是數學分析。他確實手寫了一些極其乏味的模型，還在原始計算機上建模。但是他的機器計算能力不夠，無法得到相關總和，也無法系統地探索數字變化。而當時也沒有計算機圖形可用，無法將數字轉換為明顯可以理解的形式。

人工智能和生物學界足足等了四十年才等到圖靈的深刻見解。計算機圖形學專家格雷·圖爾克（Greg Turk）研究了圖靈的方程，有時先「凍結」一個方程式的結果，然後再用另一個方程式。這個步驟讓人聯想到基因的開/關切換，是有關「模式中的模式（pattern-from-pattern）」的例子——圖靈也提到過，但無法分析。在圖爾克的人工智能模型中，圖靈的方程式不僅產生了斑點犬的標記和條紋（如他的手寫模型所做的），還產生了美洲豹的斑點、獵豹的斑點、長頸鹿的網狀花斑和獅子魚的圖案。

其他研究人員利用更複雜的方程序列，得到了更複雜的模式。這其中包括部分當今對實際的生物化學更瞭解的發育生物學專家。

例如，布賴恩·古德溫（Brain Goodwin）研究了傘藻（藻類）的生命週期。這種單細胞生物體從一個無形狀的斑點長成一根細長的莖；然後，它會長出一個扁平的頂；接下來，圍繞頂的邊緣長出一圈疙瘩；而這些疙瘩後來發芽變成一輪側根或分支；最後，側根合併形成一個傘形帽。生化實驗表明，所涉及的代謝參數多達30多個（例如鈣濃度、鈣和某些蛋白質之間的親和性，以及細胞骨架的力學阻力）。古德溫的傘藻計算機模型模擬了複雜的重複反饋過程，其中的參數時刻發生著變化。各種身體的蛻變隨之產生。

與圖靈和圖爾克一樣，古德溫也反覆琢磨數值，想看看哪些數值會真正產生新的形式。他只使用生物體中可觀察範圍內的數字，但這些數字都是隨機的。

他發現，某些模式會反覆出現，例如莖末梢處的鈣的高/低濃度的交替變換（一個輪形成的對稱性）。它們不依賴於某個具體參數值，但只要參數值被設置在一個大範圍內，它們就會自發出現。此外，輪一旦成形就不會消失。所以，古德溫說，模式可能是變形的基礎，而

變形會產生其他經常出現的特徵。這可能發生在動植物種類史和個體發育中（歷史創造力以及個體創造力），如在四足動物的肢體進化過程中。

古德溫模型沒有產生過傘形帽。可能是因為需要額外的參數來表示真正的傘藻內未知的化學作用，抑或是這些傘形帽確實就在模型的可能性空間內，所以原則上能夠從模型中產生。但是只要數值被嚴格限定，那麼隨機搜索就可能找不到這些傘形帽（側根也沒有產生，但這只是因為計算能力不足。對於每個單獨的側根，整個程序需要在較低的層級上執行）。

古德溫從中得到了一個理論教訓。他把輪看作「通用」形態，在許多動植物中發生，這一點和傘狀帽不一樣。這表明，輪產生的原因不是由偶然進化的基因引導的特殊生物化學機制，而是在大多數甚至是所有生物中發現的一般過程（如反應擴散）。這些過程可能構成「結構主義」生物學的基礎——形態學的一種普通科學，它的解釋與達爾文自然選擇學說完全一致，但先於後者出現〔圖靈的論述隱含了這種可能性，達西·湯普森（D'Arcy Thompson）也強調了這種可能性，但是圖靈自己忽視了它〕。

反應擴散利用決定局部分子相互作用的物理化學規律起作用，即在細胞自動機中可表示的規律。約翰·馮·諾依曼在定義細胞自動機的時候指出，它們原則上適用於物理學。今天的人工生命研究人員將細胞自動機用於許多目的，這裡的主要作用是它非常適合產生生物模式。例如，非常簡單的細胞自動機只在一個維度（一條線）上定義，就可以產生非常逼真的模式，例如貝殼上的圖案。

最有趣的可能是人工生命的嘗試——即用細胞自動機描述「可能生命[1]」，而不僅僅是「我們所知道的生命」。克裡斯托弗·蘭頓（Christopher Langton，1987年命名「人工生命」）探索了無數隨機定義的細胞自動機，關注它們產生順序的傾向。許多細胞自動機只產生無秩序狀態。另外一些則形成了無聊重複甚至靜態的結構。但是有些產生了變化微妙但相對穩定的模式——蘭頓說，這就是生物（也是

計算）的特徵。令人驚訝的是，在簡單度量系統的信息複雜度時，這些細胞自動機擁有相同的數值。蘭頓提出，這個「λ參數」適用於所有可能的生物，無論它/她/他是在地球上還是在火星上。

自組織不僅塑造整個身體，還塑造器官。例如，大腦發育的方法是利用進化過程（在一生中和各代之間）和非監督式學習。這種學習可以帶來非同尋常的結果（歷史創造力）。但是每個個體的早期腦發育也可產生可預測的神經結構。例如，新生的猴子擁有全方位方向檢測器。這些檢測器不能從外部世界的經驗中習得，所以我們自然會假定它們被編碼在基因中，但它們沒有。它們是從最初隨機的網絡中自發出現的。

神經科學家建立的仿真計算機模型和「純粹的」人工智能都說明了這一點。IBM公司的研究員拉爾夫·林斯克（Ralph Linsker）定義了多層前饋網絡（見第4章）。該網絡表明，如果考慮隨機活動（如胚胎大腦中的「噪聲」），簡單的赫布型學習規則能夠產生有組織的方向檢測器集合。

林斯克不僅依靠實際演示，也不只關注方向探測器，他的抽像「infomax」理論適用於這種類型（多層前饋網絡）的任何網絡。該理論指出，當信號在每個處理階段被變換時，網絡連接會不斷發展，以最大程度保存信息量。所有連接在某些經驗約束下形成，例如生物化學和解剖學限制。然而，數學確保將會出現一個通信單元協作的系統。infomax理論還涉及系統發育進化。因此，「複雜的系統在進化時所出現的單個突變具有自適應性」這樣的觀點會更符合常理。如果每個層級都能夠自發地適應另一個層級中的小變化，那麼就不再需要幾個同時發生的突變了。

對於細胞層的自組織而言，胞內生物化學和細胞/細胞壁的形成都已經被模擬。這項工作利用了圖靈在反應擴散方面的研究成果。然而，它更多地依賴於生物學，而不是源於人工生命的概念。

總之，人工智能提供了許多與自組織相關的理論概念。自組織的人工製品比比皆是。

註釋

[1]人工生命所研究的人造系統能夠演示具有自然生命系統特徵的行為。——譯者注

06 強人工智能會有真正的智能嗎

假設未來的強人工智能系統（銀幕上或機器人）能夠匹敵人類的表現，那麼它們會有真正的智能、理解力和創造力嗎？它們會有自我、道德身份和自由選擇嗎？它們會有意識嗎？如果沒有意識，它們會有任何其他屬性嗎？

這些顯然不是科學問題，而是哲學問題。許多人直觀上會認為，上述每種情況的答案都是「很顯然嘛，不會！」

事情並不是非黑即白。我們要的是仔細論證，而不只是未經核實的直覺。這些論證表明：對於上述問題，沒有任何無懈可擊的答案。因為所涉及的概念本身就頗具爭議性。只有透徹理解概念本身的含義，我們才有信心說假設的強人工智能將會或不會具備真正的智能。總之，沒有人知道確切的答案。

有些人可能會說：「這沒關係，強人工智能實際上做的事情才是關鍵。」然而，我們將看到，上述問題的答案可能會影響到我們處理強人工智能的方式。

本章不會給出明確的答案，但會談談哪些答案比其他答案更合理。同時也會介紹（一些）哲學家如何使用人工智能概念來闡明真實心智的本質。

圖靈測試

艾倫·圖靈在哲學雜誌《思想》（Mind）上發表了一篇論文，描述了所謂的圖靈測試，即測試者是否可以確定自己是在與計算機還是在和人類交互（交互時間最多五分鐘），如果有超過30%的測試者不能確定被測試者是人還是機器，那麼這台機器就通過了測試。電腦能夠真正思考的說法也就因此站得住腳了。

圖靈測試是以一種幽默的方式提出來的。雖然它出現在開頭幾頁的位置，但主要目的是預言未來的人工智能，是整個論文的附屬內容。圖靈甚至在向朋友羅賓·甘迪（Robin Gandy）介紹圖靈測試時，也將其描述為輕鬆愉快的「宣傳」，朋友對此一笑置之，沒有大肆評論。

然而，哲學家們倒是認真地討論起了圖靈測試。大多數人認為，即使程序的回應與人類的回應不可區分，也不能證明程序有智能。最常見的反對意見是（仍然是）圖靈測試只關注看得見的行為，所以一具殭屍也可以通過測試：一個和我們有著一模一樣的行為但缺乏意識的「怪物」。

該反對意見給出了兩個假定前提：1.智能需要意識；2.殭屍從邏輯上說得通。我們將在「人工智能和現象意識」一節看到對意識的一些描述表明，殭屍的概念不合邏輯。如果這些描述正確的話，那麼任何強人工智能都不會是一具殭屍。就這一點而言，圖靈測試是合理的。

圖靈測試極大地吸引了哲學家們（和公眾）的興趣，但它在人工智能中的地位卻並不重要。大多數人工智能的目標是提供有用的工具，而不是模擬人類的智能，那就更不可能是為了讓用戶相信他們正在與人類交互了。誠然，那些高調的人工智能研究人員有時會自稱或允許記者宣稱他們的系統通過了圖靈測試。然而，這些測試不符合圖靈的描述。例如，肯·科爾比（Ken Colby）的PARRY模型「愚弄」了

精神病醫生，讓他們認為自己正在閱讀妄想狂的病例，因為他們很自然地認為自己正在和人類患者打交道。同樣，如果沒有提示可能涉及機器，計算機藝術通常也會被認為是人類所為。

和真正的圖靈測試最接近的是一年一次的羅布納獎（Loebner prize）比賽（現在在布萊切利園舉行）。當前的規則規定：進行25分鐘的互動，使用20個預設問題，目的是測試記憶、推理、常識和個性。裁判綜合考慮結果的相關性、正確性，以及表達/語法的清晰度和合理性。到目前為止，沒有一個程序可以騙過30%以上的裁判（在2014年，一個形象被定位為13歲烏克蘭男孩的程序讓33%的裁判相信他們在和人對話。問題是，非英語母語人士犯的錯誤很容易被原諒，特別是兒童）。

意識的很多問題

沒有「意識的問題」這樣的東西。我們應該說「意識的很多問題」。「有意識的」一詞被用來劃清事物界限：清醒的/睡著的；故意的/不留心的；全神貫注的/走神的；易達到的/難達到的；值得報告的/不值得報告的；反思的/未核實的等。沒有一個解釋可以闡明上述所有問題。

上面所列出的對比屬於功能性問題。原則上來說，可以用信息處理術語或神經科學術語來理解它們，對此，很多哲學家都會贊同。

但是，現象意識——感覺（如藍色或疼痛）或「感受性」[1]（qualia，哲學家的技術術語）似乎不一樣。在物質宇宙（基本上）中，感受性的存在是一個臭名昭著的形而上學問題。

大衛·查默斯（David Chalmers）稱感受性的存在是一個「難題」，還認為它不可避免：「我們必須重視意識……不能接受將問題重新定義成解釋某些認知或行為功能是如何執行的。」

對於問題的答案，已經有了各種猜想。如查默斯版的泛心論，他自己承認泛心論是一個「離譜的甚至瘋狂的」理論，根據該理論，現象意識是宇宙的一個不能削減的屬性，類似於質量或電荷。其他幾位理論家則從量子物理著手，但在他們的對手看來，這些人只是用一個謎題來解決另外一個謎題。科林·麥克金（Colin McGinn）甚至認為，人類天生就不能理解大腦和感受性之間的因果關係，就像狗不能理解算術一樣。認知科學的哲學翹楚傑瑞·福多（Jerry Fodor）認為：「說物質的東西有意識，人們對此一點概念都沒有。就算是有概念，也不知道會是一種什麼樣的情況。」

簡而言之，幾乎沒有哲學家聲稱自己瞭解現象意識，如果有人說自己已經瞭解了，也幾乎沒有人相信。這個話題一直是一個哲學難題。

註釋

[1]指從實體中概括出來的可感受特性，如顏色、味道等。——譯者注

機器意識

贊成人工智能的思想家們用兩種方式研究意識：一種是建立意識的計算機模型，這叫作「機器意識」（Machine Consciousness，機器意識）；另一種（是受人工智能影響的哲學家的特點）是用計算術語分析意識，但不模擬。

一個真正智能的強人工智能具有功能意識。例如，它會在不同時間關注（留心、注意）不同的事物。人類水平的系統還能夠進行深思熟慮和自我反思，它可以產生創造性的想法，甚至特意評估它們。如果沒有這些能力，它就不能產生看似智能的表現。

人們在評估創造性的想法時，可能會涉及現象意識（見第3章）。事實上，許多人會說，只要有「功能上的」差異，就有現象意識出現。然而，所有機器意識研究者在考慮功能意識時，通常都會忽略現象意識（勇敢的，還是愚蠢的？人們聲稱自己的人工智能系統已經「以它的方式」有了現象意識，理由是它的辨別力基於感知輸入，例如光。這是否意味著出現了視覺體驗還十分值得懷疑）。

斯坦·富蘭克林的團隊在美國孟菲斯進行了一個有趣的機器意識項目，研究LIDA。這說明了兩件事：一件是（功能上）意識的概念模型——口頭表達的計算理論；另一件是對該概念理論模型的部分簡化實現。

兩者都用於科學目的（富蘭克林的主要目標）。但後者也有實際應用。LIDA的實現能夠自定義，以適應具體的問題領域，例如醫學。

與SOAR、ACT-R和CYC不同（見第2章），LIDA是在最近才出現的。第一個版本（為美國海軍製作，目的是為任務結束的水手們安排新的工作）於2011年發佈。而當前版本能夠模擬注意力，而且能夠影響在各類記憶（情景、語義和程序）中的學習；感覺運動控制也用在

了機器人學中。但還有很多特徵仍然缺失，例如語言（無論已經實現了哪些方面，下面的描述涉及概念模型）。

LIDA是一個混合系統，包括擴散活性和稀疏表示（見第4章），還有符號編程。它是基於伯納德·巴爾斯（Bernard Baars）的意識的神經心理學全局工作空間理論（Global Workspace Theory，以下簡稱GWT）開發的。

GWT將大腦視為一個分佈式系統（見第2章），其中並行運行的大量專用子系統競爭訪問工作記憶（如圖6—1所示）。各個意識項（完整的獨立單元）按照順序出現（意識流），但是「廣播」到所有腦皮層區域。

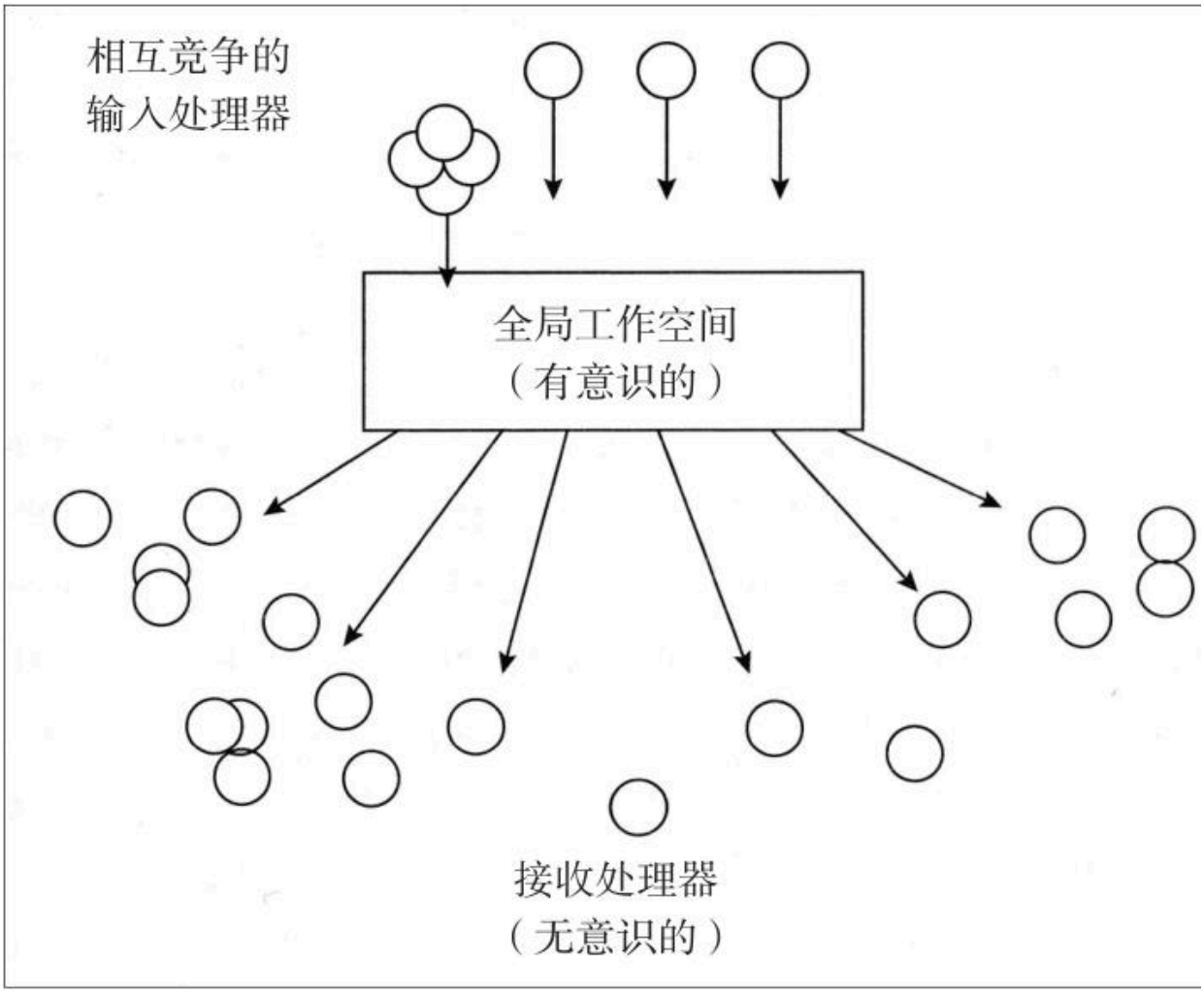

圖6—1 分佈式系統中的全局工作空間

圖註：神經系統包含各種無意識的專家處理器（感知分析器、輸出系統、規劃系統等）。這些無意識的專家處理器之間如果要實現互動、協調和控制，它們需要一個中央信息交流區或「全局工作空間」。輸入專家之間可以合作並競爭訪問全局工作空間。此處的案例顯示，四個輸入處理器合作放置一條全局消息，然後將其廣播到整個系統。

資料來源：改編自伯納德·巴爾斯的著作《意識的認知理論》（A Cognitive Theory of Consciousness）（劍橋大學出版社，1988）。

如果從感覺器官或其他子系統得到的廣播意識項觸發了某個區域並產生反應，則該反應在贏取注意力的競爭中穩操勝券，從而該區域主動控制了對意識的訪問（新穎的知覺/表示易於獲得注意力，而重複項則從意識中逐漸消失）。子系統通常很複雜。有些是分級嵌套的，許多具有不同種類的關聯聯繫。各種無意識的背景（組織在不同的記憶中）形成有意識的經歷，既喚起又修改全局工作空間中的意識項。注意力的內容反過來通過引起各類學習來適應持久的背景。

這些內容在廣播時指導下一個動作的選擇。許多動作屬於認知型，即建立或修改內部表示。道德規範被存儲起來（在語義記憶中），以作為評估潛在動作的步驟。決策也可能被其他社會智能體感知到/預測到的反應所影響。

想想規劃的例子（見第2章）。意圖被表示為近乎無意識但相對高級的結構，它能夠產生有意識的目標圖像（由當前從知覺、記憶或想像中得到的顯著特徵來挑選）。這些意圖招募相關的子目標。它們是「招募」而不是「檢索」，因為子目標自身決定其相關性。像所有大腦皮層的子系統一樣，它們等待時機由某個廣播項觸發——這裡的廣播項就是合適的目標圖像。LIDA能夠將選定的受目標驅動的動作計劃變成低級（計算機程序）可執行的微觀運動行為，能夠對不可預測且變化的環境的詳細特徵作出回應。

巴爾斯的理論（和富蘭克林版的GWT）不是在計算機科學家的研討會上憑空想出來的。相反，它的設計考慮到了各種眾所周知的心理

現象和大量的實驗證據（如圖6—2所示）。但是這些作者都聲稱，它還解決了一些以前未解決的心理謎題。例如，他們說GWT/LIDA解決了長期備受爭議的「約束」問題，即來自不同大腦區域中不同感官的幾個輸入是如何來自同一物的，例如貓的感覺、外觀、氣味和聲音。富蘭克林和巴爾斯聲稱，它也解釋了人類的心智是如何避免框架問題的（見第2章）。例如，在產生有創意的類比時，沒有中央執行系統在整個數據結構中搜索相關項。當然，如果子系統能識別出某個廣播項適合/接近（總是）其尋找的對象，那麼它就會競爭進入全局工作空間。

富蘭克林整合各種實驗證據，用LIDA探索認知心理學和神經科學理論。例如，他模擬了「注意瞬脫」，在這期間，被試未能報告在第一個視覺目標出現之後很快出現的第二個視覺目標。另外還有其他注意瞬脫的理論和計算機模型，但大多數都是用來回答孤立的問題。富蘭克林的模型來源於系統層統一的認知理論（另一個注意瞬脫的統一模型的存在基礎是ACT-R，但ACT-R不包括情感處理或高級視覺，所以不能解釋所有的實驗結果）。

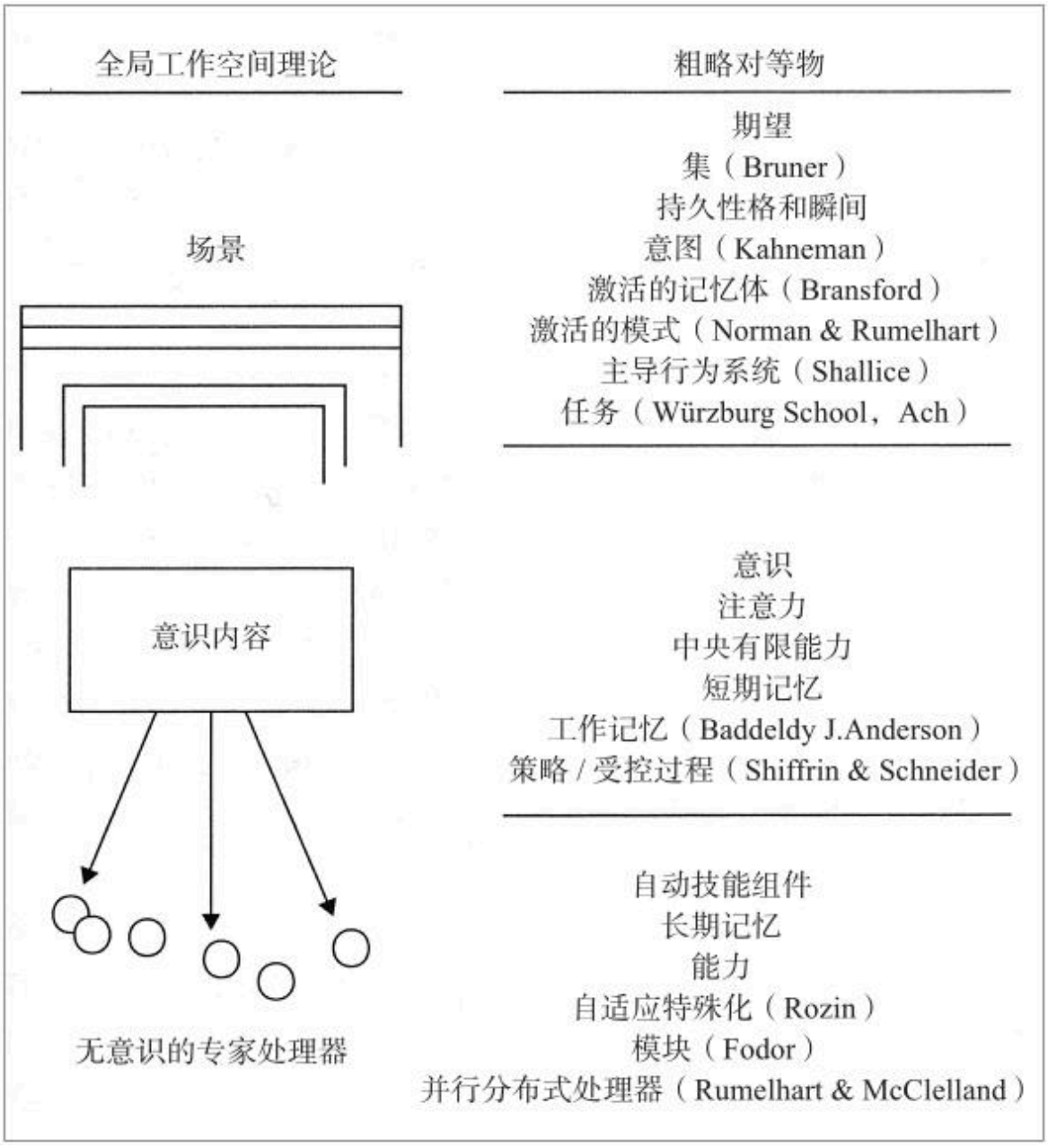

圖6—2 全局工作空間術語和其他普遍概念之間的相似性

這種人工智能方法讓人聯想到伏魔宮中的「惡魔們」，以及用於實現產出系統的「黑板」架構（見第1章和第2章）。這並不足為奇，因為那些想法給了巴爾斯靈感，他提出了神經心理學理論，最終帶來了LIDA。理論的巨輪已經回歸原樣。

人工智能和現象意識

機器意識的實踐者忽略了現象意識這個「難」題。但三位受人工智能啟發的哲學家迎難而上：保羅·丘奇蘭德（Paul Churchland）、丹尼爾·丹尼特（Daniel Dennett）和亞倫·斯洛曼。只是說他們給出的答案有爭議算是輕描淡寫了。然而，就現象意識而言，存在爭議也是意料之中的事。

丘奇蘭德的「取消物理主義」否認存在非物質的思想和經驗。相反，他把二者和腦狀態等同起來。他給出一個科學理論——部分計算範疇（聯結主義）、部分神經學範疇——定義了一個四維的「味覺空間」，將對味覺的感受性系統地映射到特定的神經結構上。四個維度反映了舌頭上的四種味覺受體。

對丘奇蘭德來說，這不是心智一大腦相互關係的問題，體驗味覺只是讓人的大腦訪問那個被抽像定義的感官空間內的一個特定點。這意味著，所有的現象意識只是大腦在某個憑經驗可發現的超空間內的某一特定位置的存在。如果是這樣，那麼計算機（可能除了全腦模擬）就不可能具有現象意識。

從本體論角度來看，丹尼特也認為不存在截然不同的體驗，身體活動除外（所以他那本頗具煽動性的書得到了一個共同反應——「解釋的不是意識，但搪塞過去了」）。

他認為，體驗就是辨別。但是，人們在辨別物質世界中存在的東西時，不會讓其他非物質世界中存在別的東西。他在一個虛構的對話中說明了這一點：

奧托（Otto）：在我看來，你已經否認了無疑是此處最真實的現象的存在，甚至連笛卡爾（Descartes）在其所著的《沉思集》[1]中都不能懷疑的真實表象。

丹尼特：就某種意義來說，你是對的，就是我否認的東西存在著。讓我們想想霓虹色擴散現象。就好像防塵布套上有一個粉紅色的發光環（他在描述一種視覺錯覺，這是由閃亮白紙上的紅色和黑色線條引起的）。

奧托：一定有。

丹尼特：但是沒有任何粉紅色的環。不完全是。

奧托：對。但一定看起來是！

丹尼特：對。

奧托：那麼它在哪裡呢？

丹尼特：什麼在哪裡？

奧托：粉紅色的發光環。

丹尼特：沒有什麼發光環，我以為你剛剛承認了。

奧托：對，在那張紙上沒有任何粉紅色的發光環，但一定看起來是。

丹尼特：對。看起來是有一個粉紅色的發光環。

奧托：所以我們談談那個發光環吧。

丹尼特：哪一個？

奧托：看起來像的那一個。

丹尼特：沒有一個只是看起來是粉色光環這樣的東西。

奧托：請注意，我並不僅僅說看起來有一個粉紅色的發光環，而是確實看起來有一個粉紅色的發光環！

丹尼特：我舉雙手贊同……當你說那裡看起來有一個粉紅色的發光環的時候，你說的是真的。

奧托：請注意。我要說的不止這些。我認為不僅看起來有一個粉紅色的發光環，還是真的看起來有一個粉紅色的發光環！

丹尼特：由於你這麼做，你和其他很多人一起掉入了一個陷阱。你似乎認為「思考（判斷、決定、堅決主張）對你來說看起來是粉紅色的東西」和「真的看起來是粉紅色的東西」之間有區別。但是沒有區別。沒有真的看起來這樣的現象——以某種方式判斷事情是這樣的現象除外。

換句話說，為感受性給出一種解釋的要求無法得到滿足。不存在這樣的東西。

亞倫·斯洛曼不同意這個觀點。他承認感受性的真實存在。但他以一種不尋常的方式做到了這一點：他在分析感受性的時候，將它們看作多維虛擬機的內部結構，此處的多維虛擬機也就是我們稱為心智的東西（請見下面的章節）。

他說感受性是內部的計算狀態。它們可以對行為（例如無意識的面部表情）或心智的信息處理的其他內部結構產生因果效應。它們只能存在於結構極其複雜的虛擬機當中（他概述了所需反射計算資源的類型）。它們只能由相關特定虛擬機的其他部分訪問，並且不一定有任何行為表現（因此，這是它們的隱私）。此外，它們不能總是（由心智中更高的自我監控層級）用語言描述（因此它們難以形容）。

這並不代表斯洛曼將感受性等同於大腦（丘奇蘭德將二者等同起來）。因為計算狀態是虛擬機的內部結構：它們不能由物理描述的語言來定義。但是，它們只有在某個基本的物理機制實現時才能夠存在且具有因果效應。

那圖靈測試怎麼樣呢？丹尼特和斯洛曼的分析都表明（以及丹尼特明確指出）殭屍是不可能存在的。因為對他們來說，殭屍的概念是

混亂的。如果給予適當的行為或虛擬機，對於斯洛曼來說，意識甚至包括感受性，都是有保證的。因此，如果還有人用殭屍能通過圖靈測試這樣的理由來反對圖靈測試的話，那麼這個反對是不成立的。

那假設的強人工智能又會如何？如果丹尼特是對的，它將包含我們所有的意識，而不包括感受性。如果斯洛曼是對的，它將具有和我們的感覺相同的現象意識。

註釋

[1]此處《沉思集》應該為《第一哲學沉思集》，論證上帝的存在和靈魂的不滅，是法國哲學家勒內·笛卡爾所著的一本哲學論文選集，《第一哲學沉思集》全書由六個《沉思》（Meditation）、其他學者對這六個沉思的《反駁》（Objection）以及笛卡爾本人對《反駁》所作的《答辯》（Reply）組成。——譯者注

虛擬機和身心問題

在20世紀60年代，希拉裡·普特南（Hilary Putnam）提出了「功能主義」，他借鑒了圖靈機的概念，還考慮了（當時很新穎）軟件/硬件的區別，以證明心智實際上是大腦所做的事情。

身體和心智是兩種截然不同的物質，它們之間（笛卡爾的）形而上學的分歧被描述層級之間的概念分歧取代。根據程序與計算機的類比，「心智」和「身體」的確迥然不同。但該類比與唯物主義完全一致（它是否能夠包括感受性一直都備受爭議）。

到1960年，雖然出現了幾個博人眼球的人工智能程序（見第1章），但是功能主義哲學家很少考慮具體的例子。他們專注於一般原則，如圖靈計算。只有到20世紀80年代中期PDP興起（見第4章）時，許多哲學家才開始考慮人工智能系統的實際運作方式。即使是當時，也很少有人問，什麼計算功能能實現（例如）推理或創造力。

理解這些事情的最好方法是借用計算機科學家的虛擬機概念。我們與其說心智是大腦做的事情，還不如說（跟隨斯洛曼）心智是虛擬機，或者說是一套完整的不同虛擬機，它/它們在大腦中實現（但是「心智是虛擬機」這個觀點有一個有悖於常理的含義，見「神經蛋白是必要條件嗎」一節）。

如第1章中的解釋，虛擬機是真實的，且具有真實的因果效應：沒有形而上學神秘的身心相互作用。因此，LIDA的哲學意義是它規定了一套虛擬機，這能夠表明（功能）意識的各個方面是如何實現的。

虛擬機方法完善了功能主義的核心部分——物理符號系統（Physical Symbol System，以下簡稱PSS）假設。在20世紀70年代，艾倫·紐厄爾和赫伯特·西蒙將PSS定義為「一組實體，它們被稱作符號，它們是物理模式，能夠作為另外一種稱為表達式（或符號結構）實體的組成部分出現……在符號結構（中），……物理上相關的（例如

一個符號在另一個符號旁邊）符號（或記號）實例」。他們說，過程是為了創建和修改符號結構而存在的，也就是說符號人工智能的定義過程。他們還補充說：「一個PSS具有實現一般智能動作的充分必要條件。」換句話說，心智一大腦是一個PSS。

從心智是虛擬機的觀點來看，他們應該將PSS稱為物理實現的符號系統假設（還是別表示成首字母縮略詞），因為符號是虛擬機的內容，而不是物理機的內容。

這意味著神經組織不是智能的必要條件，除非它是能夠實現相關虛擬機的唯一材料基體。

PSS假說（和大多數早期人工智能）假定，表示或物理符號是機器/大腦的一個特徵，明顯可分離和恰好容易定位。聯結主義對表示的描述不同（見第4章）。它將表示看作整個網絡的細胞，而不是容易定位的神經元。它將概念看成部分衝突的約束，而不是嚴格的邏輯定義。熟悉路德維希·維特根斯坦（Ludwig Wittgenstein）如何描述「家庭相似性」的哲學家們對此十分感興趣。

後來，情境機器人學的研究人員完全否認了大腦包含表示（見第5章）的觀點。有些哲學家對此表示贊同，但是大衛·基爾希認為，涉及概念（包括邏輯、語言和審慎動作）的所有行為都需要組合表示（和符號計算）。

意義和理解力

根據紐厄爾和西蒙的觀點，任何執行正確計算的PSS的確是智能的。它有「實現智能行動的充分和必要手段」。哲學家約翰·希爾勒將這種說法稱為「強人工智能」（「弱人工智能」只是說人工智能模型可以幫助心理學家闡述連貫的理論）。

他認為強人工智能是錯誤的。符號計算可能繼續在我們的大腦中起作用（雖然他懷疑這一點），但是單靠符號計算不能提供智能。更準確地說，它不能提供「意向性」——哲學家描述意義或理解力時使用的技術術語。

希爾勒的依據是一個今天仍然具有爭議的思想實驗：希爾勒被關在一間密閉的房間中，房間只有一個開口，用來傳遞上面只寫有字跡潦草的中文問題和相關回復的字條。房間裡有一個裝滿中文字卡片的盒子，還有一本規則書。希爾勒收到從屋外遞入的中文問題後，便按照手冊的說明，找到合適的指示，將相應的中文字符組合成對問題的解答，並將答案遞出房間。希爾勒不認識字條上的問題，因為問題是中文寫成的，而規則書是一個中文NLP處理程序，房間外說中文的人讓希爾勒回答他們的問題。然而，希爾勒進入房間的時候不懂中文，而且他離開房間的時候仍然不懂中文。結論是：單憑形式化計算（這正是在房間裡的希爾勒做的事）不能產生意向性，所以強人工智能是錯誤的，人工智能程序不可能有真正的理解力（「漢字屋」論證最初針對的是符號人工智能，但後來被推廣到聯結主義和機器人學）。

希爾勒在此宣稱，人工智能程序產生的「意義」完全來自人類用戶/程序員。就程序本身而言，這些意義是任意的，這在語義層面是沒有意義的。同一個程序如果「只有句法，但沒有語義」，那麼該程序同樣也可以被判斷為稅務負責人或是舞蹈藝術。

有時候他的這種說法是正確的。但是別忘了富蘭克林的觀點：LIDA模型用感官、執行器和環境之間的結構耦合來為認知打基礎，甚

至體塑認知。別忘了控制回路，它進化成了機器人的方位檢測器（見第5章）。把它稱為「方位檢測器」並非隨意行為。它能否成為方向檢測器決定了它的存在，這有助於實現機器人的目標。

後一個例子很重要，尤其是一些哲學家認為進化是意向性的源泉。例如，露斯·米利肯（Ruth Millikan）認為思想和語言是生物現象，它們的意義取決於我們的進化史。如果這是正確的，那麼不能進化的強人工智能就不具備真正的理解力。

其他崇尚科學的哲學家（如紐厄爾和西蒙自己）用因果關係定義意向性。但他們很難描述虛假的陳述：如果有人聲稱看到一頭牛，但那裡沒有牛可以產生詞，那麼他們怎麼能說是牛呢？

總之，沒有意向性的理論能讓所有哲學家都滿意。由於真正的智能包含理解力，所以這再次說明了沒有人知道我們假設的強人工智能是否真的會智能。

神經蛋白是必要條件嗎

希爾勒之所以排斥強人工智能，部分原因是計算機不是由神經蛋白構成的。他說，正如葉綠素是光合作用的場所，神經蛋白是意向性的溫床。神經蛋白可能不是宇宙中能夠支撐意向性和意識的唯一物質。但是他說金屬和硅也肯定做不到這一點。

然而，他這麼說有點過了。不可否認，我們要是說金屬和硅製成的電腦可以真正感受憂鬱、痛苦或理解語言，的確有悖常理。但是要是說神經蛋白能夠產生感受性，這種說法仍舊有違直覺，在哲學上也一樣存在諸多問題（所以，違背常理的東西不一定就是錯的）。

如果我們接受斯洛曼對感受性展開的虛擬機分析，那麼就可以打消和常理相悖這個難題。然而，心智作為虛擬機的整個描述提出了另一個相似的難題。如果一個心智達標的虛擬機能夠在人工智能硬件中實現，那麼該心智會存在於一台機器或許幾台機器中。因此，心智作為虛擬機的觀點表明，計算機的個人永生（克隆的）原則上是可能實現的。對於大多數人來說（見第7章），這和計算機能夠支持感受性的說法一樣不符合常理。

如果神經蛋白真的是能夠支持人類水平虛擬機的唯一物質，那麼我們可以反對「克隆永生」的說法。但它到底是不是？我們不得而知。

也許存在某種特殊或高度抽像的屬性，一旦神經蛋白有了這種屬性，就能實現心智執行的各種計算。例如，神經蛋白必須能夠（相當快地）構建穩定（和可儲存的）且柔性的分子，必須能夠形成具有電化學屬性的結構和結構之間的連接，這樣各結構之間才能傳遞信息。也許，其他行星上的其他物質也可以做到這些事情。

不只是大腦，身體也很重要

一些研究心智的哲學家認為大腦得到了太多關注。他們說，整個身體是更好的焦點。

他們的主張經常借鑒歐陸哲學中的現象學觀點，強調人類的「生命形式」。它包含有意義的意識（包括人類的「興趣」，它是關聯感的根基）和體塑化。

被體塑就是指成為一個動態環境中或與這個環境積極互動的生命體。環境和互動涉及物理層面和社會文化層面。關鍵的心理屬性不是推理或思想，而是適應和溝通。

研究體塑化的哲學家們幾乎沒時間關注符號人工智能，認為它太過深奧。只有基於控制論的方法受到青睞（見第1章和第5章）。從這個觀點來看，真正的智能基於身體，那麼屏幕上的強人工智能不可能真的智能。即使屏幕上的系統是一個自主智能體，結構上能夠耦合到一個物理環境中，然而這不能算是（讓富蘭克林控制速度）被體塑。

那麼機器人呢？畢竟機器人是基於並適應現實世界的物理實體。事實上，情境機器人學有時會得到這些哲學家們的稱讚。但是機器人有身體嗎？有興趣嗎？有生命形式嗎？它們究竟是不是活著呢？

現象學家們會說：「當然沒有！」他們可能引用維特根斯坦的著名評論：「就算一頭獅子可以說話，我們也不會理解它。」獅子的生命形式與我們的生命形式差異巨大，溝通幾乎不可能。當然，獅子的心理和我們的心理（例如飢餓、恐懼、疲勞等）之間有諸多重疊部分，一些最低限度的理解和同情可能存在。但即便如此，在與機器人「交流」時，飢餓、恐懼、疲勞等心理都沒有（所以對計算機伴侶的研究令人堪憂，見第3章和第7章）。

道德社區

人類會或者說人類應該接受人類水平的強人工智能成為道德社區的成員嗎？如果接受了，這將有重要的實際意義。因為它將以三種方式影響人機交互。

第一，強人工智能將和動物一樣得到人類的道德關注。人類將在一定程度上尊重它的利益。如果它要求某人中斷休息或縱橫字謎遊戲，以幫助它達到「高優先級」目標，這個人會這樣做（你是否曾從扶手椅上起來去遛狗，或者讓瓢蟲進入花園）。人類越是認為它的利益對它重要，就越覺得有必要尊重它。然而，這種判斷很大程度上取決於人類是否認為強人工智能可以產生現象意識（包括感受到的情感）。

第二，人類將認為可以對強人工智能的行動進行道德評估。今天的殺手無人機無須承擔道德責任（不像其用戶/設計者，見第7章）。但也許真正智能的強人工智能會承擔？想必它的決定可能會因為人類對它的反應而受到影響：如人類的讚美或責備。如果不受影響，就沒有社區。它能夠通過學習變得「有道德」，就像嬰兒（或狗）能夠學會聽話，或者一個年長的孩子能夠學會體貼（體貼需要發展認知心理學家口中的「心智理論」，它用能動性、意圖和信念來解釋人們的行為）。若以工具主義為評判標準，那麼虐待甚至有可能算是合乎情理。

第三，人類將把它作為道德決策的論證和說服目標。它甚至可以為人們提供道德建議。在討論這些話題的時候，我們要相信（它不僅具有人類水平的智能）強人工智能經得起特定道德標準的檢驗。但是那又能代表什麼？倫理學家全力反對的不僅是道德的內容，還有它的哲學基礎。

人們越多考慮「道德社區」的含義，就似乎越難承認強人工智能。事實上，大多數人直觀上覺得強人工智能是癡人說夢。

道德、自由和自我

人們之所以有這種直覺，主要是因為道德責任的概念與其他概念有著密切關係，如有意識的能動性、自由和自我，而這些概念促成了人類概念的形成。

有意識的深思熟慮讓我們的選擇更加符合道德標準（儘管未經考慮的行動也會受到批評）。道德讚美或譴責來自智能體或自我。與自由的做法相比，在強大的約束下的行動不太容易受到譴責。

即使將這些概念用在人身上也頗具爭議。將它們用在機器上貌似不合適，尤其是考慮到上一節談到的對人機交互的影響。然而，我們可以結合實際情況，借鑒研究人類心智時所採用的「心智作為虛擬機」的方法，去理解這些現象。

從馬文·明斯基開始，受人工智能影響的哲學家們用某類認知—動機的複雜性來分析自由。他們指出，在某些方式上，人們顯然是「自由」的，而蟋蟀不是。雌蟋蟀通過硬連線反射反應找到自己的伴侶（見第5章），但是一位女性有很多策略尋找伴侶。除了交配以外，她還有許多其他動機，但並不是所有動機都可以同時得到滿足。但是，她可以利用計算資源（又稱智能）進行管理，而這一點正是蟋蟀所缺乏的。

這些資源由功能意識構成，包括感知學習、預期規劃、默認分配、偏好排名、反事實推理，以及情感上可引導的行動計劃。丹尼特確實用這些概念和許多生動示例來解釋人類的自由。因此，人工智能有助於我們理解人類是如何做到自由選擇的。

決定論/非決定論在很大程度上是一個與事實不相干的論點。在人類行動中有一些非決定因素，但因為這將損害道德責任，所以在做決定的時候，非決定因素不會發生。但是，它有可能會影響深思熟慮期間產生的考慮。智能體不一定會想到x，或想到y——其中的x和y既包

括事實，也包括道德價值。例如，某人在選擇生日禮物時，可能會因偶然注意到的一些事情，從而想到收禮物的人喜歡紫色或支持動物權益，那麼這個人的選擇也可能因此受到影響。

剛剛列出的所有計算資源都可用於人類水平的強人工智能。因此，除非自由選擇必須包含現象意識（並且如果人們拒絕對此做計算分析），否則我們想像的強人工智能似乎會有自由。強人工智能有各種對其有意義的動機，如果我們能夠弄懂這其中的深意，那麼我們甚至可以區分它是「在自由條件下」還是「在約束條件下」做出的選擇。然而，要做到這個「如果」無疑是一個超級大難題。

對於自我，人工智能研究人員強調遞歸計算的作用，其中每個過程可以操作自己。與自我認知（和自我欺騙）相關的許多傳統哲學謎題能夠用這個遞歸計算解決。

但是，「自我認知」是什麼認知？一些哲學家否認自我的真實存在，但受人工智能影響的思想家不這麼認為。他們將其視為特定類型的虛擬機。

對他們來說，自我是一個持久的計算結構，能夠合理組織智體的行動，特別是它們認真考慮的自發行動（例如，LIDA的設計師將自我描述為「經歷的持久背景，它能夠組織和穩定許多不同局部背景下的經歷」）。自我不會出現在新生兒身上，而是一個終身塑造的過程，某種程度上服從深思熟慮的自我塑造。它有很多維度，允許發生很大變化，能夠產生可識別的個人能動性和個人特質。

它有可能實現，因為智能體的「心智理論」（最初用來解釋其他智能體的行為）反射性地被用到某人自己的想法和行動上。「自我」弄懂這些想法和行動的優先動機、意圖和目標。反過來，這些想法和行為由持久的個人偏好、個人關係和道德/政治價值觀構成。利用這種計算架構，可以構建自我形象（別人眼中的形象）和理想的自我形象（某人想要變成的形象），也可以得到基於二者差異存在的行動和情感。

丹尼特（受明斯基的極大影響）稱自我是「敘事重心的中心」，即一種結構（虛擬機），它在講述某人自己的生命故事時，會產生並尋求解釋他們的行動，特別是與其他人的關係。這當然為出現無數類型的自我欺騙和自我隱形留有餘地。

道格拉斯·霍夫斯塔特（Douglas Hofstadter，中文名為侯世達）同樣將自我描述為抽像的自我參照模式，它們來源於並因果循環回到無意義的神經活動基礎。這些模式（虛擬機）不是某人的表面特徵。相反，自我的存在僅僅是為該模式被體塑化。

霍夫斯塔特補充說，一位備受愛戴的人在身體死亡之後仍然能夠存在，這位「逝去的」人的自我先前全部被體塑在自己的大腦中，現在體塑在他們的活人的大腦中，只是現在的體塑沒有那麼細緻。他堅持認為，這不僅僅是「生活在」某人的記憶中，也不是活著的人已經吸納了這位逝者的其他特徵，例如對歌劇的熱情，而是在逝者辭世前，兩個自我深深相互滲透進對方的精神生活和個人理想，以至於二者確實都能夠在對方身上「活著」。一位已逝母親甚至可以通過她還活著的丈夫有意識地經歷孩子的成長。這種有悖常理的說法假定，存在著個人永生之類的東西，即便因為所有活著的人都已經逝去，使得逝去的自我不再被體塑化。「超人類主義」哲學家預見了計算機不朽的個人永生，詳見第7章。

總之，如果決定相信強人工智能具有真正的人類智能，包括道德、自由和自我，那麼這將是一個重大跨越，會有重要的實際意義。如果有人直觀上認為這個想法存在根本性錯誤，那麼他們很可能是正確的。不幸的是，這種直覺沒有無爭議的哲學論證來支撐。目前在這些問題上還未達成共識時，要想得到答案並不容易。

心智和生命

我們知道的所有心智都能在生物體上找到。包括控制論者（見第1章和第5章）在內的許多人都相信肯定是這樣的。也就是說，他們認為心智必須以生命為前提。

內行的哲學家有時明確地指出了這一點，但幾乎沒有論證。例如，普特南說「如果機器人不是活的，那它就不會有意識」是一個「確定的事實」。但他沒有給出科學論證，而是依靠「語言的語義規則」。甚至是少數最終捍衛這一假設的人也未能證明它是毋庸置疑的，如環境哲學家漢斯·喬納斯（Hans Jonas），還有用到「自由能量原則」的物理學家卡爾·弗裡斯頓（Karl Friston，他的這一原則總體上算控制論）。

讓我們假定這個普遍的信念是正確的。若如此，那麼只有實現了真實生命，人工智能才能實現真正的智能。然後，我們必須問，「強人工生命」（網絡空間中的生命）是否可能實現。

生命沒有公認定義。但通常提到了9個特徵：自組織、自主、湧現、生長、適應、應激性、繁殖、進化和新陳代謝。前8個可以理解成信息處理術語，因此原則上可以用人工智能人工生命體塑化。例如，包括所有其他特徵的自組織（廣義上理解）已經以多種方式實現（見第4章和第5章）。

新陳代謝不同。計算機能夠模擬它，但不能把它體塑化。自組裝的機器人和虛擬（屏幕上）人工生命都不能新陳代謝。新陳代謝是指使用生化物質和能量交換來組合和維持生物體，所以它符合不可縮減的自然法則。強人工生命的捍衛者指出，計算機使用能量，還有一些機器人具有單獨的能量存儲器，需要定期補充能量。但是，這與靈活利用相互關聯的生化循環來構建生物體的身體結構相去甚遠。

因此，如果新陳代謝是生命存在的必要條件，那麼強人工生命不可能實現。如果生命是心智的必要條件，那麼強人工智能也不可能實現。不管未來強人工智能的表現如何令人印象深刻，它都真的可能不會有智慧。

巨大的哲學分歧

「分析型」哲學家和人工智能研究人員想當然以為某種科學的心理狀態可能會實現。事實上，這一立場貫穿本書，包括本章。

然而，現象學家持反對意見。他們認為，我們所有的科學概念都來自有意義的意識，所以不能用來解釋意識（普特南本人現在接受這個立場）。他們甚至聲稱，假定存在一個獨立於人類思想的真實世界且認為科學可能發現這個世界的客觀屬性，那這一假設是荒謬的。

因此，有關心智/智能的性質所產生的分歧甚至比我目前所說的還要深。

不管是反對還是支持現象學家的觀點，都沒有壓倒性的論據。因為論據的基礎不一樣，也就是說，每一方都為自己辯護而批評另一方，但對於各自論據中所使用的關鍵術語，彼此都不贊同。對於一些基本概念，分析哲學和現象學哲學甚至都各自給出了根本不同的解釋，如理性和真理〔人工智能科學家佈雷恩·坎特韋爾·史密斯（Brain Cantwell Smith）提出了一個大膽的形而上學，有關計算、意向性和對象，他希望尊重雙方的見解；不幸的是，他的這一有趣論據毫無說服力。〕

這個爭論依然沒有解決，也許無法解決。對某些人來說，現象學家的立場「顯然」正確。但對其他人來說，它「顯然」很荒唐。因此，人們依然無法確切地知道強人工智能是否能真的是智能的。

07 奇點

人工智能自出現以來，其未來就一直備受關注。一些人工智能專家過度狂熱的預言讓記者和文化評論員們十分振奮，有時甚至是害怕。如今最好的例子是奇點（Singularity），即人工智能超過人類智力極限的時間點。

奇點代表人工智能將達到人類水平的智能（默認這將是真正的智能，見第6章）。不久的將來，強人工智能將變為超人工智能。屆時系統將智能化到可以自我複製，從而在數量上超過人類，並且還可以自我提高，從而在思想上超越人類。最重要的問題和決定將交由計算機負責。

這個概念頗具爭議性。它是否能夠發生、它是否將發生、它什麼時候發生，以及它是一件好事還是壞事，人們對此意見不一。

奇點信徒（S信徒）認為，隨著人工智能的進步，奇點必將到來。有些人支持這一觀點。他們預言人類的問題將被解決。戰爭、疾病、飢餓、無聊，甚至個人死亡等問題都將不復存在。其他人預言人類將終結，或者我們所知道的受開化的生命必將結束。史蒂芬·霍金與人工智能主要教科書的合著者斯圖爾特·羅素（Stuart Russell）在2014年5月發表評論，稱忽略人工智能的威脅「可能是我們犯的最大錯誤」，這在全世界範圍內引起了轟動。

相比之下，奇點懷疑派（S懷疑派）不指望奇點發生——肯定不是在可預見的未來。他們承認，人工智能造成了許多讓我們頭疼的問

題。但他們還沒有看到存在的威脅。

奇點的預言家

強人工智能向超人工智能過渡的觀點近來已經成了媒體的老生常談，但它最早始於20世紀中葉。關鍵的發起人是傑克·古德（「Jack」Good，他和艾倫·圖靈一道在布萊切利園擔任密碼破譯家）、弗諾·文奇（Vernor Vinge）和雷·庫茲韋爾（Ray Kurzweil）。而圖靈自己曾經預言「機器將獲得控制權」，但沒有詳細說明。

1965年，古德預言了一台超智能機，它將「遠遠超過所有人類的智能活動」。這台機器可以設計更優質的機器，所以它將「毫無疑問地引起智能爆炸」。當時的古德仍然很謹慎，說：「第一台超智能機是人類需要做的最後一項發明，倘若這台機器聽話到告訴我們如何控制它。」然而，他後來指出超智能機會毀滅人類。

25年以後，文奇推廣了「奇點」這一術語（在這一背景下，由約翰·馮·諾依曼於1958年發起）。他發表了一篇名為《技術奇點即將來臨：後人類時代生存指南》（The Coming Technological Singularity）的論文，稱奇點到來時，所有的預言都會被打破（堪比黑洞的事件穹界）。

他承認，奇點本身能夠被預知：的確，它不可避免。但在它帶來的許多（未知的）後果中，可能包含對人類文明甚至人類的破壞。我們正朝著這樣一種情況前進：「以前所有的規則將不復存在，也許就在眨眼之間，一切都指數般地增加，毫無控制的希望。」他說，即使每一個政府都意識到了危險並試圖阻止，但也是心有餘而力不足。

文奇和古德的悲觀情緒（最終）遭到雷·庫茲韋爾的反擊。他不僅展現了驚人的樂觀情緒，還給出了日期。

庫茲韋爾的著作《奇點臨近》（The Singularity is Near：When Humans Transcend Biology）一書，標題引人注目，指出強人工智能將在2030年之前實現，到2045年，超人工智能（與納米技術和生物技

術相結合）將戰勝戰爭、疾病、貧困和個人死亡。它將產生「藝術、科學和其他知識形式的爆炸……將賦予生命真正的意義」。到21世紀中葉，我們生活的擬真虛擬現實將遠比現實世界更加豐富、更加令人滿意。對於庫茲韋爾來說，「奇點」真的是奇異的，「臨近」真的意味著臨近（這種超級樂觀的情緒有時會緩和下來。庫茲韋爾列出了許多現實存在的風險，主要來自人工輔助生物技術。他說，至於人工智能本身，「智能天生不可能控制……如今若要制定策略，絕對要確保未來的人工智能體塑人類的倫理和價值觀，這種做法是行不通的」）。

庫茲韋爾的論點根基是「摩爾定律」，它是英特爾創建者之一戈登·摩爾（Gordon Moore）的觀察報告，即一美元帶來的計算機能力每年翻倍（物理定律最終將征服摩爾定律，當然不是在可預見的未來實現）。庫茲韋爾指出，任何指數級增長都有悖於常理。他說這表明人工智能的發展速度將超乎想像。所以，他和文奇一樣堅持認為，基於過去經驗的預測基本沒有價值。

競爭的預測

儘管有些人宣稱，對奇點出現之後所做的預言幾乎毫無價值，但還經常會出現此類預言。文學作品中也充斥著大量令人難以置信的例子，我們在這裡僅列幾個。

S信徒分為兩個陣營：悲觀者（跟隨文奇）和樂觀者（跟隨庫茲韋爾）。他們多半同意，強人工智能向超人工智能的轉型將在本世紀末之前順利發生，但他們的分歧點是超人工智能到底有多危險。

例如，有些人預言，邪惡的機器人將竭盡全力破壞人類的希望和生活（如科幻小說和好萊塢電影中常見的情況）。我們可能認為，有必要的話，「拔掉插頭（結束機器人的工作）」不就行了，但是這種想法已被明確否認。我們只知道，超人工智能將會變得超級精明，要靠拔插頭解決問題絕不可能。

其他人認為，超人工智能不會有惡毒意圖，但是無論如何都非常危險。我們不會把人類的仇恨嵌入到它們身上，它們也沒有理由自己醞釀仇恨。不過它們會對我們很冷漠，就像我們對大多數非人類物種一樣。但如果我們的利益與它們的目標發生衝突的時候，它們的冷漠可能讓我們墮落：變成渡渡鳥般（蠢笨）的智人。在尼克·博斯特倫（Nick Bostrom）被廣泛引用的思想實驗中，有製作回形針意向的超人工智能將從人體的原子中提取有用之物來完成這一目標。

或者，我們再談談有時用來防範奇點威脅的一般策略——遏制。阻止超人工智能直接作用於世界，雖然它可以直接感知世界。它只被用來回答我們的問題〔博斯特倫稱為「Oracle」（意為預言的東西）〕。然而，世界包括互聯網，超人工智能可能向互聯網提供內容來間接影響人類，如事實、虛假信息和計算機病毒等。

另一種形式的奇點悲觀情緒預言，機器將讓人類為它們鞍前馬後，做它們的骯髒工作，即使這違背人類的利益。同時，這也抨擊了

「切斷電源」的觀點——人類不可能通過切斷超人工智能系統與世界的聯繫來「遏制」它。他們還說，超智能機器能使用賄賂或威脅的手段，說服有時與它有聯繫的幾個人中的一人，做它無法直接做到的事情。

這種特別的擔心假定，超人工智能將足夠瞭解人類心理，可能知道什麼樣的賄賂或威脅可能起作用，也可能知道哪些人最容易被哪些說服方式說服。如果有些人認為此假設難以置信，那麼相關專家會回復說，金錢賄賂或謀殺威脅幾乎對任何人屢試不爽，所以超人工智能既不需要亨利·詹姆斯（Henry James）那樣的心理洞察力，也不需要從人的角度來理解說服、賄賂和威脅實際上指什麼。它只需要知道，向某人輸入某些NLP文本，就能以明顯可預測的方式影響這些人的行為。

一些樂觀的預言甚至更具挑戰性。最引人注目的也許是庫茲韋爾對在虛擬世界中生活和消除個人死亡的預言。生老病死目前仍然是自然現象，雖然人的平均壽命（利用超人工智能輔助的生物科學）與以前相比已經高出很多。不過到奇點到來之後，只要把每個人的個性和記憶下載到電腦中，死亡就可以「戰勝」啦！

他在2005年出版《奇點臨近》一書，書的副標題「當計算機超越人類」體現了一個哲學上讓人難以信服的假設，即一個人可以存在於硅或神經蛋白（見第6章）上。庫茲韋爾表達了他的「奇異」設想，也被稱為超人主義或後人類主義——一個部分或甚至全部是非生物人的世界。

設想稱，這些超人主義的「機器人」將會有各種直接連接到其大腦、假肢肢體或感覺器官的電腦植入物。聾啞情況將不會發生，因為視覺和聽覺信號將由觸覺來解讀。尤為重要的是，特製藥將增強理性認知（以及情緒）。

我們身邊已經有了這些輔助技術的早期版本。如果這些技術像庫茲韋爾所說的那樣會激增的話，那麼我們的人類概念將會發生深刻變化。我們不再把假肢看作人體有用的附加物，而是將其視為（准）人

體部分。廣泛使用的精神藥物將與天然物質一同列入「人腦」內物質中，如多巴胺。轉基因個體的優越智力、優勢或美感將被視為「自然」特徵。有關平等主義和民主的政治觀點將受到挑戰。擁有足夠財富去利用這些可能性的人甚至可能進化出來一個新的子物種（或物種）。

簡而言之，生物進化有望被技術進化所取代。庫茲韋爾把奇點看作「我們的生物思維和生物存在與技術合併的高潮，將帶來一個無區別化的世界……人與機器之間無差別，或物理現實和虛擬現實之間無區別」（如果你覺得信息量一下子太大，需要緩緩，那請這麼做吧）。

人工智能將如何改變我們對人性的觀念，超人主義是一個極端例子。有一種不那麼極端的哲學是「延展心智」，它將技術融合到心智的概念中，認為心智分散到世界，涵蓋依賴它的認知過程。雖然延展心智的概念已經有了廣泛影響力，但超人主義沒有。超人主義得到了一些哲學家、文化評論家和藝術家的大力支持。然而，並不是所有的S信徒都接受它。

為懷疑論辯護

在我看來，S懷疑派是對的。第6章中對心智作為虛擬機的討論，意味著實現人類水平的人工智能原則上是無障礙（可能除了現象意識）的。這裡的問題是在實踐中是否能實現。

許多有關奇點出現之後的預言違背常理，超人主義哲學近乎荒謬（恕我直言），除此之外，S懷疑論者還有其他論點。

人工智能沒有很多人假設的那麼有前途。第2章到第5章都提到了無數當前的人工智能不能做的事情。許多事情都需要一種人類的關聯感（並且默認語義web已經完成，見第2章）。此外，人工智能一直專注於智力的理性，卻忽略社會/情感智能，更別提心智了。能夠與我們的世界充分交流的強人工智能可能也需要這些能力。另外，人類的心智何其豐富，我們還需要與其工作方式相關的良好心理/計算理論。人類水平的強人工智能的前景看起來黯淡無光。

即使實踐中可行，是否有足夠的資金去實現它仍然值得懷疑。目前，政府向全腦仿真（見下一節）研究投入了海量資源，但是打造人工人類心智的經費還不止於此，未來還將會繼續燒錢。

根據摩爾定律，人工智能有望持續不斷地取得進步。但是增大的計算機功率和增加的可用數據（如雲存儲和整個物聯網上的24×7小時監測的傳感器）不能保證出現似人類的人工智能。這對S信徒來說是個壞消息，因為實現超人工智能的前提是強人工智能。

S信徒忽略了當前人工智能的局限性。他們根本不在乎，因為他們有一張王牌：技術的突飛猛進正在改寫所有規則手冊的概念。這允許他們隨意預言。他們偶爾會承認「本世紀末的」預言可能不現實。但是，他們堅持認為，「從不」僅代表一段長時間。

「從不」的確是一段長時間，所以包括我自己在內的S懷疑者都有可能錯了。這些S懷疑者沒有強有力的理由——特別是如果他們原則上承認強人工智能可能實現（像我一樣）。他們甚至可以被說服：儘管有巨大的延遲，但奇點最終會發生。

然而，仔細考慮最先進的人工智能，我們有理由支持懷疑派的假設（或如果你喜歡，也可以說是他們的賭注），而不是S信徒的胡亂猜測。

全腦仿真

S信徒預言人工智能、生物技術和納米技術將以指數級速度發展，它們之間的合作將日新月異。事實上，這些預測已經在發生了。大數據分析現在用於推進基因工程、藥物開發和許多其他科學項目（埃達·洛夫萊斯的論證，見第1章）。同樣，人工智能和神經科學現已經被結合用來研究全腦仿真（WBE）。

WBE的目的是通過仿真大腦的單個組件（神經元）以及各組件之間的連接和信息處理能力，來模擬一個真實的大腦。希望在於，這些所獲得的科學知識有許多應用，包括從阿爾茨海默症到精神分裂症等精神疾病的治療。

這種逆向工程需要用神經形態計算模擬亞細胞過程，例如離子通過細胞膜的過程（見第4章）。

神經形態計算取決於各類神經元的解剖結構知識和生理學知識。但是，WBE還需要特定神經元的連接和功能的詳細證據，包括時間。多數證據需要改進的腦掃瞄術，連續監測單個神經元的微型神經探測器。

各種WBE項目正在進行中，它們通常被贊助商們拿來和人類基因組計劃或登月比賽作對比。例如，2013年，歐盟宣佈了耗資10億英鎊的人腦工程。當年晚些時候，美國時任總統奧巴馬宣佈由美國政府撥款30億美元（加上大量私人資金），完成10年期的BRAIN計劃（使用先進革新型神經技術的人腦研究）。它的目標首先是完成老鼠大腦連接的活動圖，然後再模擬人腦活動圖。

部分大腦仿真的早期嘗試也是由政府資助的。2005年，瑞士贊助了藍腦計劃，最初是為了模擬大鼠的皮質柱，長遠目標是模擬人類新皮層中的百萬柱。2008年，美國國防高等研究計劃署（DARPA）向神經形態自適應塑料可擴展電子系統（SyNAPSE）項目投入了近4000萬

美元；到2014年，又投入了4000萬美元，這些資金完成了芯片承載54億根晶體管的工作，每塊芯片有100萬單位（神經元）和2.56億突觸。德國和日本正在合作利用神經模擬技術（NEST）開發超級計算機Kcomputer。到2012年，Kcomputer需要40分鐘來模擬1%的真實大腦活動的1秒，包含17.3億個「神經元」和10.4萬億「突觸」。

哺乳動物的WBE研究成本如此高昂，所以相關的研究很罕見。但是模擬小型大腦的無數次嘗試正在世界各地遍地開花（我自己的大學專注於蜜蜂）。這些研究可能提供神經科學方面的見解，有助於人類大腦的WBE研究。

考慮當前的硬件進步（例如SyNAPSE的芯片）和摩爾定律，庫茲韋爾的預測可能成為現實：到21世紀20年代，將出現與人類大腦原始處理能力相匹敵的計算機。但這並不代表到2030年，這些計算機能達到人類智力。

因為虛擬機（見第1章和第6章）才是關鍵。有些虛擬機只能在超級強大的硬件上實現。因此，百萬級晶體管計算機芯片可能必不可少。但是，這些芯片將執行什麼計算？換句話說，在它們上面實現的是什麼虛擬機？為了達到人類（即使老鼠）的智力，這些虛擬機必須以超出計算心理學家的理解範圍的方式，顯示出其信息的強大性。

我們假設，人腦中的每個神經元最終都會被映射（我認為不太可能）。但映射本身不會告訴我們它們在做什麼（微小的線蟲類蠕蟲C.線蟲只有302個神經元，我們已經精確知道它們的連接，但實際上我們還不能識別突觸是興奮狀態還是抑制狀態）。

對於視覺皮質，我們已經有了一個神經解剖學和心理功能之間相當詳細的映射。但是對於一般的新皮層，情況並非如此。更為重要的是，我們不太瞭解額葉皮層做的是什麼事情——也就是說，在它裡面實現的是什麼虛擬機。這個問題在大型WBE中不是很突出。例如，人腦工程採用了自下而上的方法，觀察人體的解剖構造和生物化學過程並模擬。不考慮和大腦可能支持的心理功能相關的自上而下的問題

（幾乎不涉及認知神經科學家）。即使全部完成解剖模型，並仔細監測化學信息，這些自上而下的問題也不會得到答案。

要得到答案，必須有各種各樣的計算概念。此外，一個關鍵點是整個心智（或心智一大腦）的計算架構。我們在第3章中看到，多動機生物的動作規劃需要複雜的調度機制，例如情感提供的機制。第6章中討論的LIDA表明，皮質處理超級複雜。即使是進餐用刀叉的普通活動，也需要許多虛擬機進行整合：有些處理物理對像（肌肉、手指、器具、各種傳感器）；有些處理意圖、計劃、期望、慾望、社會習俗和偏好。為了理解這種活動是如何實現的，我們不僅需要大腦的神經科學數據，還需要與所涉及的心理過程相關的詳細計算理論。

簡而言之，把自下而上的WBE看作理解人類智力的途徑可能會失敗。但它可能有助於我們理解大腦。它可以幫助人工智能科學家開發出更實際的應用。但是，到21世紀中期，屆時的WBE將能解釋人類智力的觀點是一種錯覺。

我們應該擔心什麼

如果S懷疑派是對的，確實沒有奇點，那也不代表我們就萬事大吉了。人工智能已經引起了人們的擔心。未來的進步肯定會帶來更多問題，所以對人工智能長期安全的焦慮也是非常必要的。更重要的是，它的短期影響也不容忽視。

有些擔心非常大眾化。例如，任何技術都可以用於做好事或壞事。壞人會使用所有可用工具，有時甚至資助新工具的開發，去做壞事（例如，CYC可能對壞人有用：它的開發人員已經在考慮如何限制完整系統的訪問權限，詳見第2章）。因此，我們必須警惕自己發明的東西。

正如斯圖爾特·羅素所述，我們不僅僅要對發明的「目的」保持警惕。如果有10個與問題相關的參數，而統計優化機器學習時（見第2章）只考慮了6個參數，那麼其他4個——很可能——「走向」極端。因此，我們還需要警惕使用中的數據類別。

這種擔心一般包含框架問題（見第2章）。像童話故事中的漁夫一樣，他的願望是讓當兵的兒子回家，而實現這個願望的交換條件是他自己被帶入一副棺材中，我們可能會驚訝於強大的人工智能系統竟然不能像我們一樣理解相關性。

例如，如果冷戰預警系統建議對蘇聯進行一次防禦性打擊，那麼操作員的關聯感是避免災難的唯一方法，既包括政治層面，也包括人道主義。他們認為，作為聯合國成員的蘇聯最近並沒有過分擾亂秩序，他們也害怕核攻擊帶來的可怕後果。所以，操作員們違反協議，忽略自動預警。還有幾次核打擊也僥倖避免；有些還是最近的事。戰爭沒有升級僅僅是因為人們的常識。

此外，人為失誤總有可能發生。有時這可以理解（由於工作人員手動控制讓計算機停下來，使得三哩島核洩漏事故後果更為嚴重，但

他們面臨的實際條件極不尋常），但它可讓人大吃一驚。上一段提到的冷戰時期的錯誤預警，是因為有人在設計日曆時忘記了閏年，所以月球出現在了「錯誤」的地方。還有就是他們在沒有測試和（如果可能）證實人工智能可靠性的情況下就將其投入使用。

有些擔心非常具體。今天有些東西將會讓我們傷透腦筋。一個主要的威脅是技術失業。好多體力和低級文職工作已經消失。其他類似的工作也會步入後塵（雖然需要靈巧和適應性的體力工作不會消失）。倉庫中的大多數起重、提取和搬運工作現在都可以由機器人完成。無人駕駛車輛意味著有人要失業。

中層管理職位也有風險。許多專業人員已經在使用人工智能系統作為輔助工具。不久以後，對法規和案例的耗時研究工作（例如法律和會計）可能大部分由人工智能接管。很多高要求的工作很快也會受到影響，如醫學和科學領域。即使這些工作繼續存在，但技術含量也會更高。專業培訓將會受到影響，年輕人該如何學會做明智的判斷？

雖然一些法務工作以後可能不再需要，但是律師行業也將從人工智能中獲益，因為有一大堆法律陷阱等著處理。如果出現問題，誰應該承擔責任？程序員，批發商，還是零售商或用戶？一位人類專業人士有時可能因為沒有使用人工智能系統而被起訴？如果系統的（無論在數學上還是憑經驗）可信度經證明很高，那麼這種訴訟將很有可能發生。

新型工作無疑會出現，但這是否代表更多的工作機會、受教育的機會以及更強的賺錢能力（就像工業革命後發生的），對此，我們無法給出確切答案。未來還存在嚴重的社會政治挑戰。

「服務」崗位受到的威脅較小。但是，即使是這些崗位也幾近消失。在理想化的世界裡，當前不受重視的一對一活動很可能在以後大量增加和升級。但是，這也並非板上釘釘的事。

例如，教育開始接受個人或基於互聯網的人工智能輔助，如學術界大牛授課的MOOC（大規模公開在線課程），這對許多人類教師的

工作大有裨益。計算機心理治療師已經應運而生，費用比人類治療師低得多（有些計算機心理治療師提供的幫助大得令人吃驚，例如，識別抑鬱症）。但是，它們完全不受管制。我們在第3章中看到，人口結構變化正鼓勵研究照顧老人的「護理人員」和「機器人保姆」，這些都是「錢」途可嘉的領域。

除了失業影響，在人類生存的環境中使用冷漠的人工智能系統不僅在現實中充滿風險，而且在道德上也值得懷疑。專家們為與外界接觸十分有限的老年人和殘疾人設計了多款「計算機伴侶」。它們不僅提供幫助和娛樂服務，還與用戶對話、逗樂用戶，並填補用戶情感上的缺失。即使是弱勢群體（如帕羅的用戶），也會通過這種技術而變得更加快樂，但是他們作為人的尊嚴也不知不覺遭到了背叛（文化差異在這裡很重要：例如日本和西方對機器人的態度差別很大）。

老年人可能喜歡與人工伴侶討論過往，但這是真的「討論」嗎？這可能是一個令人愉悅的提醒，觸發老人心中的美好記憶罷了。然而，人工伴侶帶來的好處是不會讓用戶有一種對方能夠對自己的經歷感同身受的錯覺。人們在接受咨詢服務情緒激動時，通常最想要的是自己的勇氣或痛苦得到「承認」，但這需要基於雙方對彼此狀態的瞭解。我們通過提供一種表面上像同理心的東西來短暫地「欺騙」一個人。

即使用戶患有中度癡呆，用戶對人工智能體的「猜想」可能比智能體的人類模型要豐富得多。那麼，當某人回憶起傷心之事（失去子女）時，如果智能體沒有按照期望和要求作出回應，這將產生什麼後果？如果伴侶按照慣常的方式表達同情，這對消除用戶的悲傷情緒沒有任何幫助——而且可能弊大於利，可能勾起該用戶的悲傷情緒，同時還沒有什麼其他安慰方式。

還有就是伴侶是否應該偶爾保持沉默或者說一個善意的謊言。不講情面說真理（和沉默）可能會讓用戶心煩意亂，但是委婉禮貌需要極其先進的NLP和一個精準的人類心理模型。

至於機器人保姆（不考慮安全問題），讓嬰兒過多接觸人工智能系統可能會扭曲他們的社會觀，不利於其語言習得。

人造性伴侶不僅出現在電影中，例如，電影《她》（Her）。它們已經投入市場。有些能夠識別語音，用語言和肢體動作勾引用戶。它們擴大了互聯網的影響，讓人們的性體驗變得糟糕（同時增強了女性在性方面被物化的意識）。許多評論家（包括一些人工智能科學家）已經寫了一些關於與機器人發生性關係方面的文章，揭示了一個非常膚淺的概念：個人的愛慕接近情慾、性迷戀和單純令人舒適的熟悉感。然而，這種觀察報告不太可能產生預期效果。通常情況下，色情作品是一棵「搖錢樹」，所以要想阻礙人工智能性玩偶未來的「進步」基本不可能。

另外，隱私也是個棘手的問題。強大的人工智能搜索和人工智能學習發佈了大量自媒體和家庭或可穿戴傳感器上收集的數據，有關人工智能的爭議因此也變得更多（谷歌公司最近為一款機器人泰迪熊註冊了專利，它有相機眼睛、麥克風耳朵，嘴巴裡有揚聲器，它既能和孩子，也能和父母溝通——不管願意與否，它還可以和看不見的數據收集器溝通）。

網絡安全問題也不容忽視。隨著越多人工智能進入到我們的世界（通常以非透明的方式），它就越重要。要想不被超人工智能控制，那就必須知道如何寫不被黑客攻擊/改變的算法（「友好人工智能」的目標，見下一節）。

軍事應用也引發了關注。機器人掃雷非常受歡迎。但機器人士兵或機器人武器呢？當前的無人機受人類唆使，但即使如此，只要擴大操作者和目標之間的人類（不僅僅是地理上的）距離，無人機就能加重苦難。人們務必相信，未來的無人機不會有權力決定哪個人或哪樣東西成為攻擊目標。即使靠它們去識別一個（看起來像是人類選出的）目標，也會引發諸多令人不安的道德問題。

我們為此做了些什麼

雖然只有很少人工智能工作人員直到現在才更多關注前文提到的問題，但是相關擔憂並非最近才出現。

1972年，幾位人工智能先驅參與了在意大利科莫湖（Lake Como）召開的會議，他們在會上談到了人工智能的社會影響，但約翰·麥卡錫不贊同他們的觀點，認為當時的推測為時過早。幾年後，計算機科學家約瑟夫·魏澤鮑姆（Joseph Weizenbaum）出版了一本副標題為「從判斷到計算方法」的著作，並哀歎將二者混淆屬於「猥褻行為」（obscenity），但他遭到了蔑視，並被趕出人工智能研究界。

當然有一些例外。例如，在第一本概述人工智能的書中，最後一章就是「社會影響」。1983年成立了計算機專業人員社會責任組織（Computer Professionals for Social Responsibility，CPSR）。這離不開SHRDLU的作者特裡·威諾格拉德的努力，詳見第3章。但這一組織主要是為了警告星球大戰技術不可靠——計算機科學家大衛·帕納斯（David Parnas）還向美國參議院提到了這一點。對冷戰的擔憂消除之後，大多數人工智能專業人士似乎不太在意它們的研究領域。只有少數幾個多年來一直專注研究社會/倫理問題，例如謝菲爾德大學（University of Sheffield）的諾埃爾·沙基（Noel Sharkey）[1]，以及一些人工智能哲學家，如耶魯大學的溫德爾·瓦拉赫（Wendell Wallach）和薩塞克斯大學的布萊·惠特比（Blay Whitby）。

現在，一提到人工智能的實踐和未來，問題就變得更加緊迫。人工智能領域內（以及在一定程度上，人工智能領域之外）的社會影響得到越來越多的關注。

有些重要的反應與奇點無關。例如，聯合國人權觀察（Human Rights Watch）組織長期主張一項禁止全自動武器的條約（尚未簽署），如自主選擇轟炸目標的無人機。最近，一些歷史悠久的專業團

體複審了它們的研究重點和行為守則。但是，對奇點的討論讓更多人加入到這場辯論中。

許多S信徒和S懷疑派都認為，奇點出現的概率非常小，但它造成的後果將十分嚴重，我們現在應該未雨綢繆。儘管文奇聲稱，對於潛在的威脅，我們束手無策，不過還是有幾個機構已經應運而生，以防止它的到來。

這些機構主要由人工智能慈善家資助，包括設在英國劍橋的存在風險研究中心（CSER）和位於牛津的人類未來研究所（FHI），以及位於美國波士頓的未來生命研究院（Future of Life Institute，FLI）和伯克利的機器智能研究所（Machine Intellegince Research Institue，MIRI）。Skype的共同開發人揚·塔裡安（Jaan Tallinn）參與創辦了CSER和FLI。這兩家機構不僅與人工智能專業人員溝通，還試圖向政策制定者和其他有影響力的公眾發佈危險警告。

2009年，美國人工智能協會的主席埃裡克·霍維茲（Eric Horwitz）組織了一場小型的專家小組研討會，研究該採取什麼樣的必要預防措施去指導，甚至推遲為社會帶來問題的人工智能研究工作。這場研討會在加利福尼亞州的艾斯羅馬（Asilomar）舉行，遺傳學專家幾年前曾在此地同意暫停某些遺傳研究。然而，作為專家小組的一員，我的印象是並非所有的與會人員都深切關注人工智能的未來。之後的報告也沒有得到廣泛的媒體報道。

2015年1月，FLI和CSER在波多黎各聯合召開了規模更大的類似會議（根據查塔姆宮規則，沒有記者出席）。6個月前，組織者馬克思·泰格馬克（Max Tegmark）、羅素和霍金共同簽署了一封警告信函。不出所料，那時的氣氛顯然比在艾斯羅馬更緊迫。會後，大量新的資金〔來自互聯網百萬富翁埃隆·馬斯克（Elon Musk）〕投入到了人工智能安全和倫理人工智能的研究中，成千上萬的人工智能研究者共同簽署了一封警告公開信，後來在媒體上廣泛傳播。

不久之後，湯姆·米切爾（Tom Mitchell）和其他幾位領域帶頭人起草了第二封公開信，警告不要開發自主武器，應該禁止武器在無人

為干預的情況下選擇和攻擊目標。簽署人希望「防止發生全球人工智能軍備競賽」。2015年7月，這封信出現在人工智能的國際會議上，有將近3000名人工智能科學家和17000名相關領域的專業人士在上面簽名，從而引起了媒體的廣泛關注。

波多黎各會議之後，麻省理工學院經濟學家埃裡克·布林約爾松（Erik Brynjolfsson）和安德魯·邁克菲（Andy McAfee）也在2015年6月共同簽署了一封公開信，向決策者、企業家、商人和經濟學專家們發出警告，稱人工智能可能讓經濟發生徹底變化，並給出了一些可能降低但無法消除風險的公共政策建議。

這些人工智能研究仍在努力說服大西洋彼岸的政府資助者們認可社會/倫理問題的重要性。美國國防部和國家科學基金會最近都表示願意為此提供研究經費。但政府不是最近才支持，多年來，「政府的」興趣一直在不斷增加。

例如，2010年，兩家英國研究理事會贊助了一個跨學科的「隱修會」（Robotics Retreat），參與起草了機器人專家的行為準則。商定了機器人學的五項「原則」，其中兩項涉及前文所討論的擔憂：（1）機器人不應被設計為武器，除非出於國家安全考慮，（4）機器人為人工製品：情感和意圖的錯覺不得用於操控弱勢用戶。

另外兩項規定了人類承擔的道義責任：（2）人類而不是機器人為承擔責任的主體；（5）應該可以找出任何機器人的（合法）負責人。該組織沒有試圖更新艾薩克·阿西莫夫的「機器人三定律」（機器人不得傷害人類，或因不作為（袖手旁觀）使人類受到傷害；除非違背第一法則，機器人必須服從人類的命令；在不違背第一及第二法則下，機器人必須保護自己）。起草者認為，這裡的任何「法則」都應該由人類設計者/建造者遵守，而不是機器人。

2014年5月，由美國海軍資助的一項跨學科學術活動（5年750萬美元）受到媒體的追捧。它是一個由5所大學合作的項目（耶魯大學、布朗大學、塔夫茨大學、喬治城大學和倫斯勒理工學院），旨在

發展機器人的「道德能力」。參與者包括認知和社會心理學家、道德哲學家，還有人工智能程序員和工程師。

此項目不是為了提供一系列道德算法（和阿西莫夫的法則一樣）或優先研究某種特殊的元倫理學（例如功利主義），甚至也不是為了定義一套非競爭的道德價值觀，而是希望開發一個能夠在現實世界中有道德推理（和道德討論）能力的計算系統。因為自主機器人要做出慎重決定，而不只是遵照指示（也不是對「情境」線索作出靈敏反應，見第5章）。例如，如果一個機器人開展搜救工作，那麼它應該先疏散/救援誰？或者，如果提供社會陪伴服務，那它什麼時候該和用戶撒謊（如果真有這種情況）？

項目提出的系統將整合知覺、運動行為、NLP、推理（演繹和類推）和情感。情感包括：情感思維（它可以發出重要事件要發生的信號，並規劃相互矛盾的目標，詳見第3章）；「抗議和痛苦」的自動顯露，這可能會影響與之交互的人做出的道德決定；以及識別周圍的人的情感。而且官方發言稱，機器人甚至可能「超越」普通（即人類）的道德能力。

考慮到第2章和第3章中指出的實現強人工智能的障礙，再加上與道德相關的特定困難（見第6章），人們可能會懷疑這個目標是否能夠實現。但項目值得進行。考慮現實世界的問題（如前文提到的兩個差異巨大的例子）是一盞警示燈，提醒人類在使用人工智能的時候要遵守道德準則，否則會發生許多危險。

除了這些機構所作出的努力，越來越多的人工智能科學家也瞄準了埃利澤·尤德考斯基（Eliezer Yudkowsky）所說的「友好人工智能」。它對人類有積極影響，既安全又實用。它包含的算法將易於理解、值得信賴，魯棒性[2]robust）強，並且就算這些算法都失敗了，也輸得得體。它們應該是清楚易懂、可以預測的，而且不易被黑客操縱——如果其可靠性是通過邏輯或數學證明，而不是憑經驗測試，那就更好了。

馬斯克在波多黎各會議上捐贈600萬美元後，FLI立即發佈了前所未有的《徵求建議書》（Call for Proposals，6個月後，37個項目全部得到資助），向來自「公共政策、法律、倫理學、經濟學、教育及其相關領域」和人工智能領域的專家們發出呼籲：「研究項目一方面是為了避免潛在危害，另一方面是為了最大限度地利用人工智能今後所帶來的社會效益，而且僅限於如下研究：明顯不是專注於實現增強人工智能能力的標準目標，而是為了增強人工智能的魯棒性和效用……」總之，歡迎發展友好人工智能的呼籲可能已經發出。但奇點的足跡仍未消失，《建議書》指出：「應優先考慮專注於人工智能的魯棒性和效用的研究，即使該研究涉及大幅取代當前的能力……」

總而言之，人工智能即將引發世界末日的觀點實屬錯覺。但是，從某種程度上來說，正是由於這種觀點，讓人工智能研究團體、政策制定者和普通老百姓才逐漸意識到一些切實存在的危險。他們早就該注意到了！

譯者後記

《AI：人工智能的本質與未來》的作者瑪格麗特·博登教授是OBE勳章（大英帝國官佐勳章）獲得者，人工智能領域最著名的人物之一。在本書中，作者試圖用通俗易懂的語言，向讀者介紹人工智能的方方面面。

本書概括起來，可以分為三個部分：人工智能的前世，從埃達·洛夫萊斯夫人的預言到艾倫·圖靈的圖靈機、到沃倫·麥卡洛克提出的神經活動中內在思想的邏輯演算、再到唐納德·赫布的學習理論，囊括了人工智能各分支的歷史演變過程，提到了多學科共同發展為人工智能的發展帶來了巨大推動力；人工智能的今生，講述了當前技術的最新進展，如邏輯符號主義的不斷完善、人工神經網絡的再次復興、機器人技術和人工生命的蓬勃發展以及強人工智能面臨的種種難題等；人工智能的未來，作者提出了如下問題，比如，人工智能是否真的能夠實現？會思考的智能體是否需要道德約束？何時會出現超越人類智能的智能體？智能體帶來利益的同時也帶來了威脅，對此我們如何應對？作者不僅詳述並總結了當前各個學派的觀點，還給出了自己的答案。

這是一本難得的佳作。在當前互聯網+的大時代背景下，人工智能技術已經廣泛應用於金融、交通等各個民生領域，同時也是計算機行業內部變革最主要的推動力，如雲計算、大數據、物聯網都需要人工智能技術的介入。神經科學、認知心理學等眾多社會學科也在人工智能技術的助力下快速發展。本書不僅從科學探索的角度向讀者介紹了人工智能的理論意義，還囊括了眾多具有代表性的實踐案例，讀者

閱讀的時候不會覺得枯燥乏味，反而會心潮澎湃，感覺自己猶如置身於人工智能這股大洪流之中。同時，作者沒有局限於概括性的介紹，在關鍵部分還詳述了很多技術細節，這樣讀者既能感受到人工智能發展道路的大氣磅礴，也有機會細細體會人工智能技術的小橋流水之勢。作者還提出了很多頗具啟發性的問題，耐人尋味。

無論您是人工智能領域的專業人士，還是僅對人工智能頗感興趣的普通讀者，相信此書都不會讓您失望。感謝中國人民大學出版社能夠給予我翻譯此書的機會，使我受益良多，同時也非常感謝張釗、朱文佳、肖鳳霞、鹿桐欣、趙曉瑋、高瑞霞、成欣、易汕、馬伊莎、況輝等人為本書的翻譯工作付出的努力！

孫詩惠

註釋

[1]機器人研究學者，現在擔任國際機器人武裝控制委員會或ICRAC的主席。——譯者注

[2]魯棒性是指控制系統在一定（結構、大小）的參數攝動下，維持其他某些性能的特性。也就是系統的健壯性，這是在異常和危險情況下系統生存的關鍵。——譯者注