我看人工智能辯論

● 李逆熵

數年前，我在一本名為《超人的孤寂》的科幻專論之中，談到「人工智能」(artificial intelligence，簡稱AI)的研究如何從被嘲笑轉為被重視。我當時是這樣寫的：「科幻小說中的智能電腦和機械人被電腦界嘲笑了數十年，今天終於得以吐氣揚眉，好教世人知道誰是誰非。」然而，我接着寫道：「有關這方面的爭論在很長的一段時間裏仍不會停息……」①

我當時不知道的是，就在我撰寫上述文字的時候(1986年)，一場新的有關人工智能的激烈爭論，又已經在西方學術界展開。

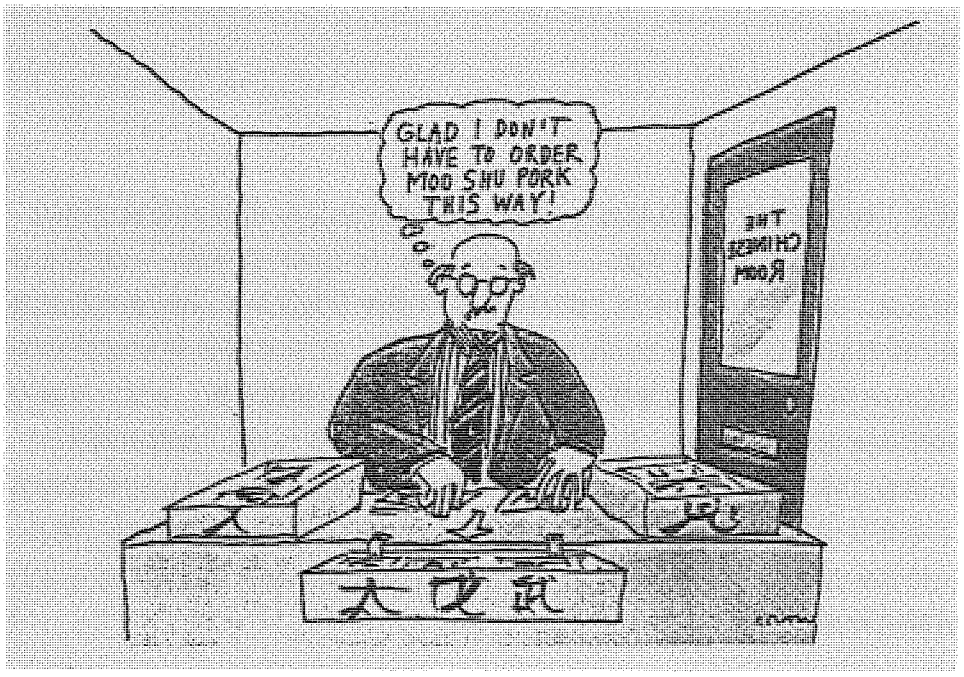
這是一場新的爭論，因為無論自有科幻小說(十九世紀末)還是有電腦(1946年)以來，有關機器思維的爭論——也就是「擁AI」和「反AI」的爭論——實在沒有停息過。例如在60、70年代，便有英國哲學家魯卡斯(J.R. Lucas)基於數理邏輯(特別是哥德爾(Kurt Gödel)的不完備定理)，以及美國哲學家德雷弗斯(Hubert Dreyfus)基於現象學的觀點對人工智能作出種種非難。令這一爭論在80年代重新熱烈起來的，是一篇發表於1980年的文章〈心靈、大腦與程序〉②。在這篇文章裏，美國加州柏克萊大學的哲學家西爾(John R. Searle)首次提出了他那著名的「中文字房實驗」(Chinese Room Experiment)。有趣的一點是，這個實驗雖然從來沒有付諸實施，但它所引起的爭論，卻比不少真正進行過的實驗還要多。

「中文字房實驗」之爭

究竟甚麼是中文字房實驗呢？原來這是一個無需付諸實施，只需在理念上運作的實驗，就好像愛因斯坦在創立相對論時所提出的種種「擬想實驗」一樣。在這個實驗中，西爾設想一間裝滿了中文字卡的房間，房中還有一本由英文寫成的手冊和一個只懂英文而完全不懂中文的人。現在假設房外有人不斷以字卡的形式把一些中文寫成的問題透過一道窄縫送入房中，房中的人則按照輸入的字卡和手冊指示，把一些對應的字卡順序從窄縫擲回房外。重要的是，手冊中的指示從來沒有解釋任何一個中文字的意義，指示的形式永遠只是「如果收到某某編號的字卡，則交回某某編號的字卡」或是「若收到某一組字卡，則交回另一組字卡」等。

假設房外輸入的問題是一些有關唐詩的賞析或儒家哲學的討論，而由房中輸出的，則是相應這些問題的精闢答案，那麼房外的人必會認為房內的人不單通曉中文，而且對中國文化有深厚認識。可是房內的人不要說中國文化，就是連一個中文字也毫不認識！

西爾要指出的是，中文字房有如電腦，房中的手冊則是電腦程序。我們今天自然無法想象可以寫出一套這樣複雜的程序，正如我們無法想像可以有一本這麼神奇的手冊以至令房中的人不致露出馬腳。但西爾的論點是，即使我們有一天終於能寫出這樣一套電腦程序，那是否表示電腦已經真正懂得思維呢？當然不是！正如中文字房裏的人始終不懂中文，電腦也只是按照程序工作，而沒有任何真正的理解。也就是說，電腦永遠只會有語法(syntax)而沒有語義(semantics)。西爾更進一步強調，就是更多更精細的語法，也無法產生半丁點兒語義，而這正是機器和人類的分別所在。



《科學的美國人》雜誌中的「中文字房」擬想圖。

1980年的這篇論文，最先發表於專業學術期刊，因此未有引起廣泛注意。不久，西爾應劍橋大學邀請主講著名的「雷夫講座」(Reith Lectures)，其中一講正以中文字房實驗為題。演講其後結集成書，名為《心靈、大腦與科學》③。透過這本小書，他的論點才開始引起注意。

在80年代後期，西爾的非難成為了AI爭論的一個焦點，這與電腦發展的歷史也許不無關係。日本宣佈進行「第五代」電腦計劃至今已將近十年，但真正智能型電腦的來臨似乎仍未有期，而機械人則仍只是工廠中高度專門化的機器，所謂機械傭人或機械保姆仍只存於科幻電影之中。事實證明，我們最初對AI發展的期望太樂觀了。人類對外部世界的認知能力和對事物的學習能力，遠比我們最初想像的為複雜。在這樣一種氣候下，「反AI」的論調重新抬頭，有關的爭論也再次激烈起來。

當然，早在1981年，對中文字房理論的反駁已經出現了。「擁AI」的主將霍夫斯塔特(Douglas R. Hofstadter)繼震撼學術界的奇書《哥德爾、埃舍爾、巴哈：一條永恆的金帶》④之後，與哲學家鄧納(Daril C. Dennett)合作，出版了一本可作為前書續篇的選集《心靈的我》⑤，集中收錄了西爾的首篇中文字房文章，也刊登了霍氏的反駁文章。

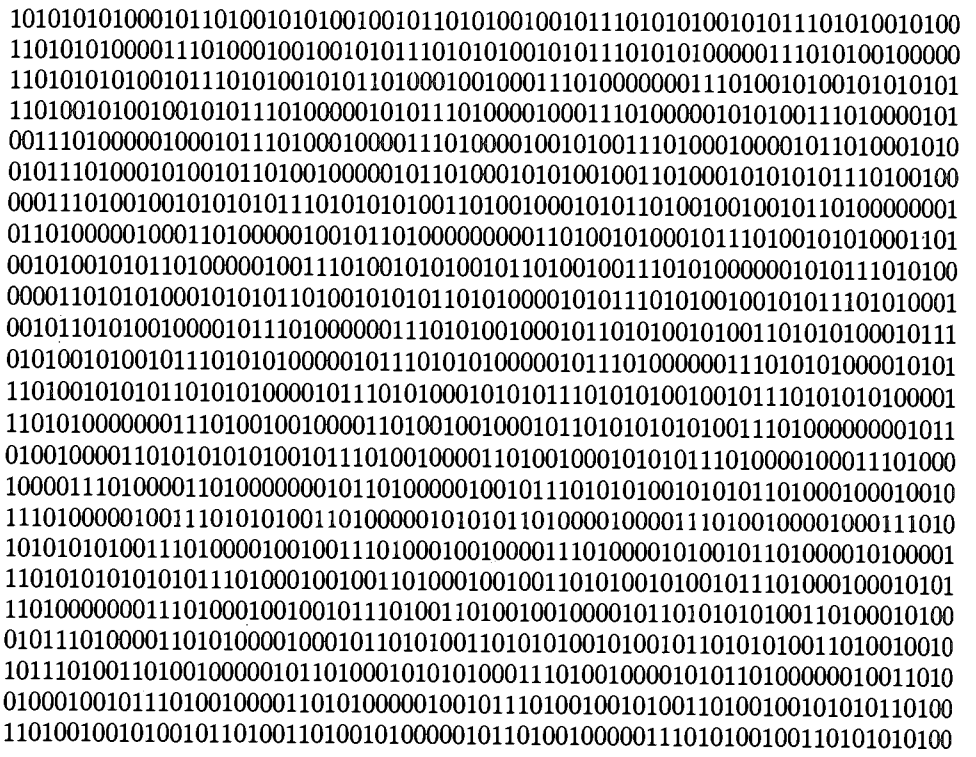
霍氏和其他「擁AI」學者對西爾的反駁，大致可稱為「系統觀」或「層次觀」的理論。他們主要的論點是：所謂語法和語義的劃分，只是一個層次的問題。在一個較低的操作層次來看，我們可能的確只是看到語法；但從包括整個「字房系統」的高層次來看，我們必然無可避免地要涉及語義。

到了90年代，著名的通俗科學雜誌《科學的美國人》以卷首位置同時刊登了兩篇針鋒相對的文章，即西爾的〈大腦思維是電腦程序嗎？〉和「擁AI」的丘卓倫夫婦(Paul and Patricia Churchland)的〈機器能思想嗎？〉⑥。這兩篇文章大體上仍集中於上述「語法」「語意」之爭，它們可以說是對「西爾論題」爭論的一個總結。

自我意識和「圖林試驗」

在提出「中文字房實驗」時，西爾的文章還有一個中心思想：他指出，無論我們把電腦程序寫得如何複雜，它也無法出現人腦(即真正的心靈)所擁有的一項特質：意向性(intentionality)。他在文章的結尾寫道：「無論大腦如何產生意向性，這種過程必不等同於一項電腦程序。因為單靠電腦程序本身，絕不足以產生意向性。」

西爾這兒所用的意向性一詞，顯然是整場AI爭論的核心。不過我們常用的字眼，則是「自我意識」(self-consciousness)和「自由意志」(free will)。歸根究柢，有關AI的最大爭論是：我們終有一天可以造出一副擁有自我意識和自由意志的機器嗎？



一個「普遍圖林機」可以用如圖的長串二進數碼表示出來。

在這裏要澄清的一點是，AI研究有所謂「弱AI命題」和「強AI命題」之分。前者追求的，是以機器來模擬人類部分「智能」活動，包括數學演算、邏輯推理、下棋、自動導航、甚至包括醫學診斷、經濟分析以及與人作模擬式的簡短交談等等。今天最先進的工業機械人或「專家系統」，都不過是弱AI範疇的產物。這些產物對科技和經濟發展的推動固然極其重要，但並非我們一直談及的「AI爭論」的對象。有關AI的重大爭論，針對的是「強AI命題」。這一命題認為，人類終有一天能够造出一副在認知和思維能力方面皆與人完全無異(如果不是更優勝)的機器。

1950年英國數學家圖林(Alan M. Turing)發表了一篇名為〈計算機器與智能〉⑦的經典文章，揭開了有關強AI命題爭論的序幕。雖然電腦在當時只出現了數年，但科學家和哲學家對「電腦能否思維？」這一問題已爭論不休。圖林有感在進行這項爭辯的時候，大家對思維的定義往往各不相同，於是執筆寫了這篇文章，提出了著名的「圖林試驗」(Turing test)，作為判定機器是否擁有思維能力的標準。

甚麼是「圖林試驗」？與「中文字房實驗」一樣，這也是一個「擬想實驗」。假設房間裏有一個人和一台電腦，房外的人可以通過打字機或熒幕顯示分別與房中兩者交談。圖林的論點是，如果我們無論透過如何刁鑽的問題也無法識別房中何者是人何者是電腦，那末便不得不承認，房中的電腦具有與人類一樣的思維能力。也就是說，它「懂」得思維。這有如我們和別人相處時，總是通過不斷的交流和觀察，來判定對方是否擁有思維能力一樣。

圖林試驗為「思維」確立了一個運作性的定義。自此，「強AI」的擁護者有了一個明確的目標，那便是要製造出一副能够通過「圖林試驗」的機器。顯然，西爾的「中文字房實驗」其實是對三十年前這篇經典之作的抗議。西氏認為，無論一副機器外表看來如何聰明，它仍只是一副機器(或只是一項程序)而不可能擁有真正的理解，更遑論真正的思想、感情和意志。簡言之，西爾強迫「強AI命題」的擁護者面對這樣一個問題：姑勿論某副機器能否真的通過「圖林試驗」，但你們認為機器真的能够擁有思想、感情，亦即擁有自我意識嗎？

大部分的「強AI」擁護者對上述問題的答案其實是肯定的(否則他們也不會成為「強」AI擁護者)，只是他們一向不願意宣揚這一觀點。為甚麼呢？因為這牽涉到一個形而上學的問題，很容易給人扣上「不科學」的帽子，因為「自我意識」這回事按其本質是永遠無法確立的。我固然知道我自己存在，亦即我有自我意識。但假如一副電腦向我們宣稱：「我知道我自己存在！」我們有甚麼辦法去判定這一說話為真，而並非一些極其巧妙的程序所產生的結果呢？

誠然，一個個體是否擁有「自覺的心」(a self-conscious mind)可能是永遠無法確知的事情，但這正是整個問題的關鍵所在：如果迴避了這個問題，則所有關於AI的爭論就失去了意義。事實上，既然我們承認其他人有自覺心靈(唯我論者除外)，那麼自然也不應懼怕談論機器是否可以有自覺心靈。畢竟這才是我們心底裏最關注的一點。

皇帝的新心靈

像時間起源一樣，自我意識可說是宇宙間一個深不可測的謎。千百年來，不少哲人智者都為解開這個謎而費盡思量。在哲學探求中，這便是著名的「心、物問題」以及「自由意志與決定論」這兩大課題，同時亦牽涉到本體論中唯物與唯心的古老爭論。究竟現代科學的進步，對上述這些問題帶來了甚麼新的啟示呢？這些啟示對「強AI」的追求又是好消息抑或壞消息呢？這些是彭羅斯(Roger Penrose)在他的新作《皇帝的新心靈》⑧中企圖回答的問題。

現任牛津大學Rouse Ball數學講座教授的彭羅斯，是當代著名的數學家兼物理學家。70年代，在研究黑洞時空結構的問題上，他曾經提出著名的「彭羅斯圖」(Penrose diagrams)這一分析工具。1988年，由於他在科學研究上的貢獻，曾與著名的物理學家霍金(Stephen Hawking)共同獲得吳爾夫獎(Wolf Prize)。

厚達450頁的《皇帝的新心靈》是彭氏第一本科普著作。在書的前半部，他花了大量篇幅探討這樣一個問題：對真理(即使只局限於邏輯和數學上)的追尋，是否可以用程式化步驟(algorithmic procedures)體現？引伸下來，也就是問：人的思維是否可以還原為不同的程式(algorithms)，而最終歸結為電腦程序的運作呢？

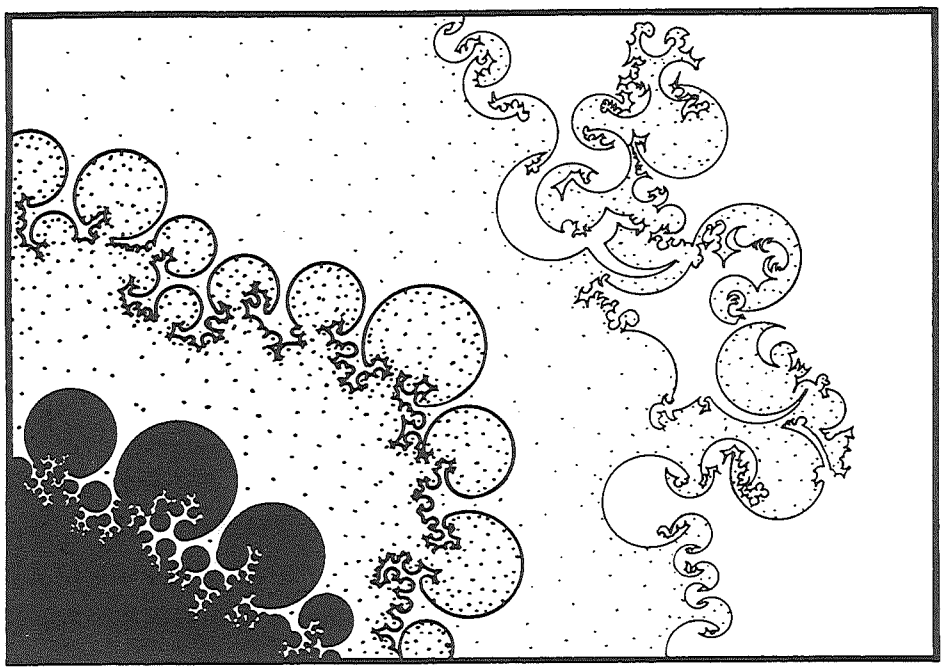
與哲學家西爾一樣，數學家彭羅斯對上述問題的答案至終也是否定的。但較諸「中文字房實驗」的雄辯式非難，他對問題的分析深刻和全面得多了，其間涉及圖林機的休止問題、曼德布洛集(Mandelbrot set)與非遞歸(non-recursive)數學、希爾伯特的公理化綱領和哥德爾的不完備定理、可計算性和「複雜理論」的概念……等等。這其中最重要的，就是是否所有數學問題都可以「程式化」，即由一定和有限程序來決定。以著名的曼德布洛集和其他古典問題(例如二次不定方程diophantine equation的整數解)為例，他說明這是不可能的。

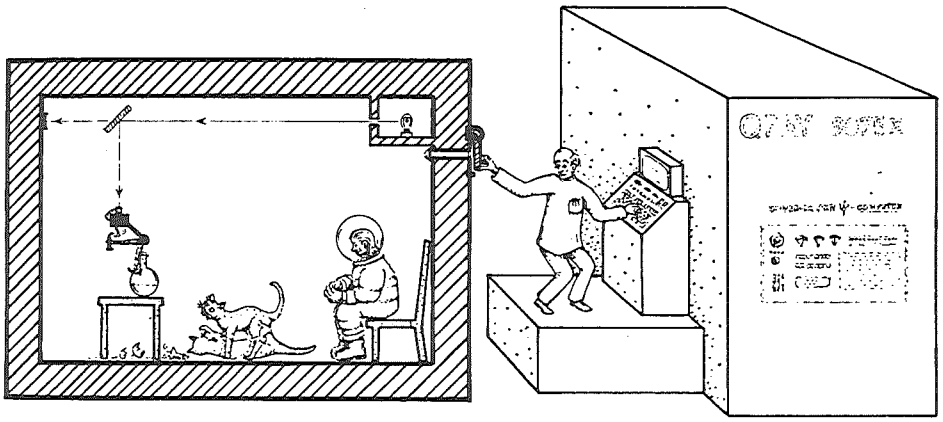
而思維和意識既無法還原為程式的運作，那是否表示它們超乎了科學研究的範疇呢？彭氏的答案既是肯定也是否定的：之所以肯定，是他認為整個現代科學架構中欠缺了極關鍵的一環，以至我們無法了解自我意識的本質；之所以否定，是他認為這一「缺環」並非甚麼神秘不可知的事物，而最終可以被科學探究。

彭氏進一步大膽假設，這「缺環」可能與現代物理學中的兩大困惑有密切關係，這便是量子力學中有關「波函數塌縮」(collapse of the wave function)⑨這一核心觀念所導致的種種有悖常識的後果，以及重力場還未能完全量子化而與自然界的其他基本作用力統一起來這一問題⑩。

花了百分之八十篇幅探討上述問題後，彭氏提出了「一顆重力子判據」這項大膽假說。按照這判據，只要物質和能量分佈所導致的時空曲率達到一顆重力子的水平，量子力學中的線性疊加原理便會失效，而在其處的波函數便會發生「塌縮」，「歸約」成「本徵態」(eigenstate)⑪。根據他和A. Ashtekar的粗略估計，倘若這一判據成立的話，質量在10^{-7}克左右的粒子波函數都會「塌縮」，也就是說粒子的運動規律會接近經典理論。

極可能是「非遞歸」(non-recursive)的曼德布洛集。





量子力學的困擾：以機率性的量子作用在密室中殺死了一隻貓；但就外界觀點而言，該貓可能還存在於「死」「活」兩個狀態的疊加態之中，這就是「薛定諤的貓」。

從這「一顆重力子判據」出發，彭氏希望可以通過廣義相對論影響和改造量子力學，特別是解決波函數塌縮所引致的困境，從而創造出一套涵蓋量子力學和廣義相對論的嶄新物理學。這套新的物理學在目前自然仍是未知的領域，是一個夢想。但他認為，這也許正就是打開自我意識之謎的鑰匙，也是判別人類思維與(無何如何複雜的)計算程式的分野的關鍵。

這的確是十分新鮮和富啟發性的一個觀點。它倘若成立，對強AI的追尋究竟是好消息還是壞消息呢？

透過了一個寓言式的楔子和後記，彭羅斯清楚地表明，他這本書的結論是對強AI命題的否定。由於目前的人工智能研究之中，還未曾包括他所提出的「缺環」，因此研究綱領，便好像「穿」在皇帝身上的「新衣」，是一種自欺欺人和永遠無法實現的夢想。這正是書名《皇帝的新心靈》的含義。

人工智能和進化

從本文開始，大家也許已經看出來，筆者其實是個強AI的擁護者。多年來，筆者看過不少反AI的論調，卻始終未為所動。然而，彭氏這本著作和其他的論調不一樣，它的立論是有大量數學和物理佐證的。但這是否表示我放棄原來的立場，接受「人工智能」只能是空中樓閣這一結論呢？

事實大大不然。因為彭羅斯所否定的強命題，只局限於「程式的執行可產生意識」這一特定形式，而筆者所信奉的，卻是「人類終有一天可製造出一副擁有自我意識的機器」。至於這一目標如何能達到，當然有待不斷深化的科學研究。例如近年興起的連結主義(connectionism)，便將研究重點，從電腦程序設計，轉移到神經網絡結構和自我學習機制的模擬方面了。從這個角度看，彭氏的觀點對人工智能的追求其實是好消息而不是壞消息。因為只要我們找到他所提出的「缺環」，便可以將「意識」的研究放到一個堅實的科學基礎之上。這對創造「機器意識」來說，當然是奠基性的一大步。

是甚麼令我深信人工智能終能實現呢？是生物進化這一科學事實。我們也許能頗為肯定地認為，病毒和細菌並不具有自我意識。可是青蛙呢？麻雀呢？獅子、猩猩或海豚呢？三個月大的嬰兒或嚴重弱智的人又怎麼樣？事實上，無論是「我」如今具有的意識，或上述不同程度的「意識」，都同樣是生物進化的產物。我們今天高度發達的意識並不是自古便存在的。南非猿人、能人、直立人以及尼人等，都應該擁有不同程度的自我意識。也就是說，意識只是物質組織複雜到某一程度後所產生的現象。顯然，這種現象還會隨着生物繼續演化而不斷產生新的內涵和特性。

生物進化已經有數十億年歷史，人工智能的發展則只有數十年，難道這一簡單的事實還不夠雄辯嗎？

註釋

① 李建才：《超人的孤寂》(香港新雅，1988)，第6、7章。

② John R. Searle: "Minds, Brains, and Programs", *Behavioral and Brain Sciences* 3, No. 3 (1980).

③ John R. Searle: *Minds, Brains, and Science* (Harvard Univ. Press, 1984).

④ D.R. Hofstadter: *Gödel, Escher, Bach: An Eternal Golden Braid* (Harvester, 1979)，該書有舉秀成的中譯節本《GEB，一條永恒的金帶》(四川人民出版社，1983)。

⑤ D.R. Hofstadter and D.C. Dennett, ed. *The Mind's I* (Harmondsworth, Middx.: Penguin Books, 1981).

⑥ John R. Searle: "Is the Brain's Mind a Computer Program?"; Paul M. Churchland and Patricia S. Churchland: "Could a Machine Think?", *Scientific American* (January 1990), pp. 20–31.

⑦ Alan M. Turing: "Computing Machinery and Intelligence", *Mind* 59, No. 236 (1950), reprinted in *The Mind's I*, 同註⑤。

⑧ Roger Penrose: *The Emperor's New Mind: Concerning Computers, Minds and the Laws of Physics* (Oxford Univ. Press, 1989).

⑨ 更正規的說法是「態矢量歸約」(state vector reduction)，當我測量物理系統時，它的波函數會因而產生變化，按機率進入該物理量的某個本徵態。〔編者按〕。

⑩ 事實上，重力場的量子化問題在過去數年間已經有重大進展，例如超弦理論，見J.H. Schwarz, *Physics Reports* 84, 223 (1982)及宇宙波函數理論，見J.B. Hartle & S.W. Hawking, *Physical Review* D28 2690 (1983)。〔編者按〕。

⑪ 在量子理論中，波函數的「歸約」必然由特定物理量的觀測產生，而歸約後可能採取的各個「本徵態」，也視乎該物理量而定。彭氏提出的新構想主要在於解決波函數究竟在甚麼情況下「歸約」的問題，他的答案是「歸約」乃由於時空曲率超過某臨界值而產生，但他似乎未觸及「歸約為那個物理量」的本徵值的問題〔編者按〕。

李逆熵 香港氣象工作者，業餘時間熱衷於科學普及工作，1985年獲選為全港十大傑出青年。迄今發表的著作有《三分鐘宇宙》(科學概論)、《賣隕石的人》(科學散文)、《最後的問題》(西方科幻選譯)、《超人的孤寂》(科幻評論)、《夜空的呼喚》(科普創作)等。