SOUTH CHINA QUARTERLY Vol.9, No.2 (May 2019), 220—231

Six Philosophical and Ethical Issues Arising from the Development of Artificial Intelligence

Wei Yidong

Abstract: The development of artificial intelligence, because of its great impact on society, has caused people's anxiety and reflection to this kind of artificial agents. From the perspective of philosophy, whether the machine itself can have intelligence, whether it has human thinking, whether functionalism can explain machine consciousness, whether biological naturalism can refute the theory of functionalism, and whether there is an explanatory gap between consciousness and perception, these questions are thought-provoking. These Philosophical issues have showed that machines having intelligence and the ability to think means a renaissance of the "ghost in the machine" view, which goes back to Descartes' dualism on the independent existence of matter and mind, but as to the question of "how ghosts get into the machine," the dualism cannot explain it. On the one hand, Functionalism tries to give a functional explanation to machine consciousness on the basis of intermediary causality and states that any two systems should have the same mental state only if they have isomorphic causal processes. On the other hand, bio-naturalism strongly opposes this view, arguing that mental state is a high-level feature of emergence, which is caused by physiological and physical processes in low-level neurons, so machines do not have biological functions and of course cannot be conscious or have any understanding and thinking. As a matter of fact, whether a machine has "intelligence" and whether it "understands" are questions of two different levels, and how to answer them depends on how we define "intelligence" and understanding. Obviously, the core of these issues is the question of consciousness, that is, whether a physical state can be used to explain a mental state, and the question of consciousness is still a mystery. On this issue, there is not only philosophical confusion but also some "explanation gaps." From the perspective of ethics, whether the development of artificial intelligence will pose a threat to human beings, who should be responsible for this, whether intelligent robots should have identity, what kind of ethical relationship it has with human beings, these safety and ethical issues about artificial intelligence deserve people's attention. The identity of intelligent robots, that is, whether humanoid intelligent robots are recognized as humans, is not only a legal issue, but also an ethic one. If robots pose a great threat to human survival, it is necessary to prohibit their development, and human law and morality absolutely do not allow for killing. This means that human beings must control the development of robots. After all, the intelligent machines are created by human beings, and the adverse consequences should be borne by human beings themselves. Therefore, the influence of artificial intelligence on human behavior should also be regulated by human beings themselves.

Keywords: artificial Intelligence, functionalism, bio-naturalism, consciousness, intelligent robot

Author: Wei Yidong received from Shanxi University his master's in 1994 and Ph.D. in 2002 and was a visiting scholar at Cambridge University in 2003. From 2007 to 2013, he served as director of School of Philosophy and Sociology at Shanxi University. Currently, he is a professor and doctoral supervisor at the same school and at the university's Center for Research in the Philosophy of Science and Technology, a key research base of the humanities and social sciences sponsored by the State Ministry of Education. His research interest focuses on the history and philosophy of science and philosophy of cognition. His major publications include *ISIS and History of Science, Science in Wider Context, Contextualism and Reconstruction of Philosophy of Science, Research on the Philosophy of Cognitive Science, Cognition, Model and Representation*, and *Scientific Representation: From Structural Analysis to Contextual Construction*. He is also the chief editor of *Collection of Translations in Philosophy of Cognition*.

《南國學術》第九卷第二期（2019年5月），220—231

發展人工智能引發的六大哲學倫理學問題

魏屹東

[摘　要]人工智能的發展，對社會產生了巨大衝擊，從而引發了人們對它的種種憂慮和反思。從哲學視角看，機器自身是否能擁有智能，是否具有人一樣的思維，功能主義能否說明機器意識，生物自然主義能否駁倒功能主義理論，對意識和感受性的解釋是否存在解釋鴻溝，這些問題發人深思。哲學問題表明，機器擁有智能、能夠思維意味着"機器中的幽靈"觀點的復興，這就回到了笛卡爾關於物質與心靈獨立存在的二元論上，但"幽靈是如何進入機器的"問題，二元論無法說明。功能主義試圖從中介因果性給出機器意識的一種功能說明，認爲任何兩個系統，祇要它們具有同構的因果過程，它們就應該具有相同的心理狀態。而生物自然主義則堅決反對這種觀點，認爲心理狀態是高層次的湧現特徵，它由低層次的神經元中的生理物理過程引起，機器不具有生物功能，當然不可能有意識，也就不可能有理解和思維。其實，機器是否有"智能"，與機器是否"理解"是不同層次的問題，這就看如何定義"智能"和如何去理解了。顯然，這些問題的核心是意識問題，即能否用物理狀態解釋心理狀態的問題，而意識問題目前仍是一個謎。在對這個問題的解釋上，既有哲學上的混亂，也存在某些"解釋鴻溝"。從倫理學視角看，人工智能的發展是否會對人類構成威脅，誰應該對此負責，智能型機器人是否應該有身份認同，它與人類有怎樣的倫理關係，這些關於人工智能的安全和倫理問題，值得人們重視。智能型機器人的身份認同，即人形智能型機器人是否被承認是人類，不僅是個法律問題，更是倫理問題。如果機器人會對人類生存產生極大的威脅，禁止其發展就是必須的，人類的法律和道德都絕對不允許殺人。這就意味着，人類必須控制機器人的發展，畢竟智能機是人類自己創造的，產生的不良後果應該由人類自己來承擔，人工智能對人類行爲方式的影響也應該由人類自己來規範。

[關鍵詞] 人工智能　功能主義　生物自然主義　意識　智能型機器人

[作者簡介] 魏屹東，1994年、2002年在山西大學分別獲得哲學碩士和博士學位，2003年在劍橋大學做高級訪問學者，2007—2013年擔任山西大學哲學社會學學院院長；現爲教育部人文社會科學重點研究基地——山西大學科學技術哲學研究中心/哲學社會學學院教授、博士生導師，主要從事科學史與認知（科學）哲學研究，代表性著作有《愛西斯與科學史》《廣義語境中的科學》《語境論與科學哲學的重建》《認知科學哲學問題研究》《認知、模型與表徵》《科學表徵：從結構解析到語境建構》，主編有《認知哲學譯叢》等。

人工智能的快速發展，不僅引發了一系列哲學問題，也引發了一系列倫理學問題。對於這些問題的思考，學界不斷有成果面世。從適應性角度看，人類應該能夠與其創造的智能機共處共存，因爲創造它們是爲了服務人類，而不是控制甚至毀滅人類。儘管如此，人工智能的發展後果仍然難以預料。人工智能產生後，有人聲稱，機器能夠智能地行動，哲學界將這種觀點稱爲“弱人工智能假設”；也有人斷言，機器的行爲實際上是擁有思維能力而不僅僅是模擬思維，哲學界將這種觀點稱爲“強人工智能假設”。大多數人工智能研究者將“弱人工智能假設”視爲理所當然，不太關心“強人工智能假設”；也就是說，衹要他們設計的程序能夠工作，就不關心機器是模擬智能還是真實智能。然而，機器人代替人類工作或思維將會產生什麼樣的後果？是否會有機器人控制人類的事情發生？這些都是人工智能可能引發的重大哲學、倫理學問題，需要作深層次考察和追問。

一　機器的智能是其自身具有的嗎？

機器能夠智能地行動，是目前機器人技術正在實現的目標。自1956年人工智能產生起，就有人斷言，人類學習的每個方面，或智能的任何其他特性，都能夠由機器精確地描述或模擬。這似乎表明，“弱人工智能假設”不僅是可能的，而且已經部分實現。然而，也有人認爲，弱人工智能不可能，它衹是人們狂熱崇拜計算主義而產生的一種幽靈。①

顯然，人工智能可能與否，取決於如何定義它。若將人工智能定義爲“對在給定構架上最佳主體程序的探索”，它就是可能的。即對於任何具有k比特程序儲存器的數字構架，存在2^k個主體程序；接下來是發現最好的程序，並列舉和測驗它們。而當k非常大時，在技術上則是不可行的。但是，哲學界關注的是理論的而不是實踐的問題，他們對人的認知構架和機器的構架的對比更感興趣。具體說，就是關心機器能否思維，而不關心機器行動的最大效用。

大多數人認爲飛機作爲機器能飛，但問題是，飛機能飛與機器能思維不是一回事。因爲，“飛”是動力學問題，“思維”是認知科學問題，後者比前者要複雜得多。阿蘭•圖靈（A.Turing，1912—1954）認爲，不要問機器能否思維，而是問它能否通過智能行爲的測試。②圖靈這裏說的是會思維的機器，而不是一般的機器；即能計算的機器就是智能機，大腦就類似於計算機。這就是著名的“計算機隱喻”。如果將思維定義爲計算，那麼機器無疑會思維，因爲人也會計算，當然也會思維。但是，人的思維與計算機的“計算思維”在本質上是不同的，正如鳥會飛與飛機會飛在本質上的不同。

對於人來說，計算過程肯定是思維過程，但思維過程未必是計算過程，如情感思維、冥想等，儘管這裏的計算可以被理解爲廣義的，包括數學運算、邏輯推理、語言操作、問題解決等。而對計算機來說，計算就是操作和執行人編寫的程序，這個過程與情感過程完全不同，儘管情感也可以被量化、被計算，甚至審美、幸福感這些純粹體驗的東西也可以納入計算範疇。問題來了：一方面，若計算等於思維，機器就是可思維的；若計算不完全等同於思維，機器就可能不會思維。這就需要對“計算”和“思維”概念進行精確定義，找出它們之間的內涵與外延。這是一個棘手的難題，學界一直在爭論中。

作爲計算機科學的開創者，圖靈已預見到對智能機可能存在三種主要反駁：無能力論證，數學反駁，隨意性論證。如何看待這些反駁？需要仔細分析和討論。

1.無能力論證。這種論證的形式是“機器決不能做x”。這裏的“x”，是許許多多具體的事例或情形或狀態，包括和藹的、機智的、美麗的、友好的；有主動性，有幽默感；能夠辨別是非、犯錯誤、墜入愛河；享受草莓和冰淇淋；向經驗學習；正確地使用詞；是它自己思想的主

① K.Sayre，“Three more flaws in the computational model”，*Paper presented at the APA*(Central Division)Annual Conference, Chicago，Illinois，1993.

② A.Turing，“Computing Machinery and Intelligence”，*Mind* LIX(236，1950)：433-460.

體；有像人一樣多的行爲多樣性，等等。在這些事例中，有些是相當簡單的，如犯錯誤；有些則是人能做到而智能機做不到的，如墜入愛河，因爲機器還沒有感情。儘管目前的智能型機器人能夠做許多連人都難以做到的事情，但它們的背後都有人類專家在操作，獨立於人的智能機單獨靠自己的能力還不能做出發現。也就是說，智能機在執行任務的過程中，還不能提供洞見，產生頓悟，並且理解，因爲洞見、頓悟、理解是人類特有的，智能機目前還不具備這種能力。

2.數學反駁。哥德爾（K. F. Gödel，1906—1978）、圖靈的工作已經表明，某些數學問題，根據特殊形式系統，原則上是無解的。哥德爾的不完備定律就是這方面著名的例子：若任何形式公理系統“F”有足夠能力做算術，則建構一個所謂的哥德爾語句“G(F)”是可能的。該語句具有如下兩個屬性：

（1）“G(F)”是一個“F”語句，但不能在“F”內被證明。

（2）如果“F”是不矛盾的，那麼“G(F)”是真的。

這個定律說明，機器在心理上不如人，因爲機器是形式系統，它受不完備定律的約束，不能建立自己的哥德爾語句的真值，而人則沒有任何這樣的限制。這種觀點，引起了學界的長期爭論。

首先，哥德爾的不完備定律僅適用於有足夠能力做算術的形式系統，包括圖靈的智能機。在是否具備心理性這一點上，機器與人不可同日而語，這種觀點部分是基於計算機是圖靈機的觀念，但不完全正確。因爲，圖靈機是無限的，計算機則是有限的，而且任何計算機都能夠用命題邏輯描述爲一個巨系統，這不服從哥德爾不完備定律。

其次，一個主體不應該太無知以至於不能建立某些語句的真值，而其他主體能夠。例如，“張三不能無矛盾地斷言這個語句是真的”。如果張三斷言了這個語句，那麼他將使自己陷入矛盾，所以，張三不能一致地斷言該語句。這其實已經證明，存在這樣一個語句，即張三不能無矛盾地斷言語句的真假，而其他人和機器則能夠。還有，一個人無論何等聰明，終其一生也不能計算出10^{100}的數目之和是多少，而超級計算機能夠在幾秒鐘搞定。但是，不能就此認爲計算機比人聰明，也沒有看到這作爲基本限制會影響人的思維能力。人類在發明數學和計算機前就已經智能地行動幾十萬年了。這意味着，形式數學推理在指明什麼是智能的方面很可能衹是起次要作用的。

最後，也是最重要的，即使承認計算機在其所能夠證明方面是有限的，也沒有任何證據表明人不受那些限制的影響。嚴格講，人們很容易證明一個形式系統不能做“x”，如沒有情感、心理活動，但聲稱人能夠使用自己的非形式方法做“x”卻沒有給出這種觀點的任何證據，這難道是合理的嗎？的確，證明人這種生物系統不受哥德爾不完備定律支配是不大可能的事情，因爲任何一個嚴格的證明都要有非形式的人的參與纔能形式化，並因此駁倒它本身。於是，人們就給直覺留下了餘地，認爲人能夠以某種方式執行數學洞察力的非凡技巧。在做推理時，人們必須假設一致性或無矛盾性的存在。更可能的情形是，人本身就是一個矛盾體。這一點，對於日常推理是如此，對於縝密的數學推理也是如此，著名的“四色地圖問題”的證明就充分說明了這一點。

3.隨意性論證。這是圖靈提出關於人工智能最有影響和後人持續最久的批評，即關於人行爲的隨意性論證。其含義是，人的行爲太過複雜，以致不能通過一組簡單的規則來理解；計算機所做的無非是遵循一組規則，不能產生像人這樣的智能行爲。以一組邏輯規則無能力地理解每件事，在人工智能中被稱爲“資格問題”。德雷福斯（H. L. Dreyfus，1929—2017）是“資格問題”的主要支持者，他在《計算機不能做什麼》《計算機仍然不能做什麼》中，對人工智慧遵循一組規則產生智能的觀點提出一系列批評，其立場被豪格蘭德（J. Haugeland，1945—2010）稱爲“好的老式人工智能”（GOFAI）[①]。這一立場主張，所有智能行爲能夠通過一個邏輯地從一

① J. Haugeland(ed.), *Artificial Intelligence: The Very Idea* (Cambridge, Mass: MIT Press, 1985).

組事實和描述這個域的規則進行推理的系統中得到理解。德雷福斯正確地指出，邏輯主體對於資格問題是脆弱的，而概率推理系統更適合於開放的域。①

不過，應該看到，德雷福斯反對的不是計算機本身，而是編輯計算機程序的方法。根據德雷福斯的觀點，人類的專門知識的確包括某些規則的知識，但衹是作爲在其中人操作的一個整體語境或背景。例如下棋，棋手首先必須掌握關於下棋規則的知識，這些知識作爲語境在下棋過程中起作用。新手完全依賴規則，需要計劃做什麽，而大師看一下棋盤就能夠迅速知道如何做，正確的步驟已在頭腦中。這就是說，大師無需刻意考慮規則就能迅速做出決定，其思維過程不依賴有意識的心智的內省。但是，這不意味着思想過程不存在，衹是在大師那裏，思維過程已經融入到熟練的技能中。

德雷福斯提出了獲得技能的五個步驟，以基於規則的處理開始，以迅速選擇正確答案的能力結束。爲此，他還提出一個神經網構架組成一個巨大的案例庫，但指出四個問題：（1）源於案例的好的概括沒有背景知識是不能獲得的。沒有人知道如何將背景知識歸併入神經網的學習過程。（2）神經網的學習是一種監管學習形式，需要相關的輸入和輸出的優先識別。因此，沒有人類訓練者的幫助它不能自動操作。事實上，沒有教師的學習能夠通過無監管學習和強化學習來完成。（3）學習演算法在許多特性上執行得並不好。如果我們挑選一個特性子集，就會有增加新特性的未知方式存在，這使得當下集應該證明不適當考慮習得的事實。事實上，新方法如“支持向量機”（Support Vector Machine）是能夠非常好地處理大量特性集的。隨着基於網絡大數據的引入，許多應用領域如語言處理、計算機視覺能夠處理數千萬個特性。（4）大腦能夠引導其感官尋求相關信息，能夠加工信息以提取與當下情境相關的屬性。但是，德雷福斯主張，這種機制的詳細過程目前還不能被理解，甚至包括能夠指導人工智能研究的假設方式。②實際上，由信息價值理論支持的自動視覺領域，已經關注方向感測器問題，而且某些機器人已經吸收了所獲得的理論結果。

總之，德雷福斯關注的許多問題，包括背景常識知識、邏輯推理問題、不確定性、學習、決策的合成形式等，的確構成人工智能的主要問題，現在已經歸入標準智能主體設計領域。這是人工智能進步的證據，不是其不可能性的障礙。

4.情境主體論證。德雷福斯最強的論證是針對情境主體的，而不是針對無身的邏輯推理引擎的。一個主體，支持其理解“狗”的知識庫，僅源於邏輯語句的一個有限集，如狗（x）→哺乳動物（x），與一個觀看狗賽跑、同狗一起玩的主體來說，這個主體處於劣勢。正如安迪•克拉克（Andy Clark）指出的，生物大腦的首要功能是作爲生物身體的控制系統，生物身體在豐富多彩的真實世界環境中運動、行動。爲了理解人類或動物主體如何工作，人們必須考慮整個主體，而不僅僅是主體程序。③事實上，具身認知方法不獨立地考慮大腦，而是將它與身體看作一個不可分割的整體。也就是說，認知發生在身體包括大腦內，而身體是嵌入於環境中的。這樣，認知就是環境中的認知，即情境認知。可以預計，機器人的感測器技術的發展，一定依賴於具身認知綱領和情境認知綱領。

二　機器具有人一樣的思維嗎？

機器能夠像人一樣思維是“強人工智能”的假設。許多哲學家認爲，機器即使通過了圖靈測

① H. L. Dreyfus，*What Computers Can't Do: A Critique of Artificial Reason*（New York：MIT Press，1972）;*What Computers Still Can't Do: A Critique of Artificial Reason*（Cambridge，Mass：MIT Press，1992）.

② H. L. Dreyfus and S.E.Dreyfus,*Mind Over Machine:The Power of Human Intuition and Expertise in the Era of the Computer*(Oxford,UK：Blackwell,1986).

③ A. Clark，*Being There: Putting Brain, Body, and World Together Again*（Cambridge，Mass：MIT Press，1998）.

試，也不能說它能像人一樣思維，而衹能算作是一種思維模擬。圖靈已經預見到這種觀點，稱其爲意識論證——即衹有機器不僅能寫出樂曲，而且知道它寫出了樂曲時，纔能承認機器就是大腦；或者說，機器必須意識到它自己的心理狀態和行動。說人有意識，幾乎沒有異議；說機器有意識，則會引起極大爭論。意識是認知科學、認知心理學的一個重要議題，它與研究直接經驗的現象學相關，即機器必須實際上能夠感知情感。在現象學中，判斷某物是否有意識的一個標準是所謂的意向性。根據這個標準，對於機器來說，由於它能夠是實際地關於真實世界中的某物的，所以，它可能有信念、願望和其他表徵。這種觀點，引起了爭論。

圖靈對反駁機器有意向性觀點的回應是機智的。並給出了這樣的理由——機器可能有意識，或者有現象學特性或具有意圖，問機器能否思維則是一個不清晰的問題，因爲人們對機器的要求高於對人類的要求，畢竟在日常生活中，沒有直接的證據表明我們瞭解他人的內在心理狀態。這是心靈哲學中的他心問題。與其繼續糾結這個問題，不如關注每個人思維的慣例。當時，人們還衹是設想人與機器在未來對話的可能性；而如今，人機對話已經是尋常之事，真實思維與人工思維之間已沒有語言之間的區別，就像人造尿素與有機尿素之間不存在物理、化學性質之間的區別一樣。

如果說無機物與有機物之間的界限已經被打破，那麼，機器思維與人類思維之間的界限是否也會被打破呢？這畢竟不是同一層次的問題，前者是實體層次的，後者是精神層次的。對於思維，目前還沒有達到從無機物合成有機物的程度，所以多數人還是寧願相信，人工機器思維無論多麼引人注目，也不會是真實的（人的思維），至多是模擬的（似人思維）。正如塞爾（John R. Searle）所質疑的："沒有人假設，一場暴風雨的計算機模擬會讓我們淋濕……究竟爲什麼人在其正常心智中會假設心理過程的計算機模擬實際上具有心理過程？"[①]塞爾的質疑有一定道理，既然計算機模擬不會產生實際效果，人們也不能指望計算機模擬心理過程能夠產生實際心理狀態。然而問題是，這種類比適當嗎？雖然計算機模擬暴風雨不會讓人淋濕，但人們並不清楚如何將這個模擬運用到心理過程。它雖然與用灑水器模擬下雨會使人淋濕不同，但暴風雨的計算機模擬的確能模擬濕的特徵。如同模擬駕駛不等於真實駕駛，但能讓模擬者體驗到是在駕駛真實的車。大多數人會同意，在計算機上模擬下棋，與在真實場景下棋沒有什麼不同，因爲這是在執行下棋的行動，而不是在模擬。心理過程更像是模擬暴風雨，還是更像下棋？

其實，圖靈的思維慣例給出了可能的答案。在他看來，一旦機器達到某種老練的程度，這個問題本身通常會自動消失。這也會消解"弱"與"強"人工智能之間的差別。不少人對此持反對意見，認爲存在一個不可否認的實際問題——人有真實心智，機器沒有。要闡明這個實際問題，需要弄清人如何有真實心智，而不僅僅是身體產生神經生理過程。

哲學上解決這個心身問題，與機器是否有真實心智問題直接相關。心身問題是一個既老又新的問題。笛卡爾（R. Descartes，1596—1650）的二元論將心與身截然分開，認爲二者獨立存在，儘管它們之間存在相互作用，但隨後產生的問題是，心是如何控制身體的？而心的一元論，也稱物理主義，通過斷言心與身不是分離的，心理狀態就是物理狀態，避免了這個問題。大多數現代心靈哲學家都是不同形式的物理主義者，他們原則上承認強人工智能的可能性。但物理主義者面臨的問題是，解釋物理狀態（特別是大腦的分子構架和電化學過程）如何能夠同時是心理狀態呢？比如，疼痛、享用美食、知道某人在開車、相信北京是中國的首都等。

著名的"甕中之腦"（brain in a vat）思想實驗就是爲反駁物理主義而提出的。物理主義者試圖說明，一個人或機器處於一個特殊心理狀態是什麼意思。意向狀態，如相信、知道、願望、害怕等，是他們特別關注的，這些狀態指向外在世界的某些方面。例如，我吃燒餅的知識是一個

① J. R. Searle，"Minds，brains and programs"，*The Behavioral and Brain Sciences* 3（1980）：417-418.

關於燒餅和在其上發生了什麽的信念。若物理主義是對的，情形一定是，一個人的心理狀態的適當描述，由那個人的大腦狀態來決定。如果我正集中精力以經意的方式吃燒餅，那麽我此刻的大腦狀態是“知道某人吃燒餅”這類心理狀態的一個實例。當然，我的大腦中所有原子的具體構架對於“我知道我吃燒餅”的心理狀態是不必要的。也就是說，我的腦或其他人的腦有許多構架，它們屬於同一類心理狀態。關鍵點是，同一腦狀態不對應於一個基本明確的心理狀態，如某人正吃蘋果的知識。

物理主義的觀點，的確具有科學理論的簡單性特徵，但這種簡單性受到“甕中之腦”思想實驗的挑戰。設想一下，你的腦從你出生就與你的身體分離，並被置於一個神奇設計的甕中，這個特殊的甕能很好地保存你的腦，允許它生長、發育。同時，電信號從一個完全虛幻世界的計算機模擬輸入你的腦，來自你的腦的移動信號被攔截，並被用於修正模擬直到適當。事實上，你經歷的模擬生活精確複製你可能已度過的生活，如果你的腦不是被置於甕中的話，包括模擬吃虛擬的燒餅。這樣，你可能已經用於一個腦狀態，該狀態與真正吃真燒餅的人的腦狀態同一，但是，說你擁有心理狀態“知道你吃燒餅”表面上可能是假的。然而，事實是，你沒有吃燒餅，你從來沒有品嘗過燒餅，當然你不可能有這樣的心理狀態。

這個實驗，似乎與腦狀態決定心理狀態的觀點相矛盾。解決這個問題的路徑是，心理狀態的內容能夠從兩種不同的視角來解釋：寬內容與窄內容。寬內容是從一個通達整個情境的全能的外部觀察者的視角給出解釋，這個觀察者能夠區分這個世界中的不同事物。按照這種觀點，心理狀態的內容既包括腦狀態也包括環境的歷史。窄內容祇考慮腦狀態。例如，一個真實吃燒餅的人與一個甕中之腦吃燒餅者的腦狀態的窄內容，在這種情形中是相同的。

如果你的目標是把心理狀態歸於共享你的世界的其他人，以預測它們可能的行爲和效果，那麽寬內容就是完全適當的，因爲它包括了心理狀態所涉及的語境因素，如環境的歷史，這是我們關於心理狀態的日常語言進化的必要環境。如果你關注的是人工智能系統是否真實地思維和真實地擁有心理狀態的問題，那麽窄內容就是適當的，因爲機器是無語境的，即與環境的歷史無關。所以，不能簡單地認爲，人工智能系統能否像人那樣真實地思維，依賴於外在的那個系統的條件。如果我們考慮設計人工智能系統並理解其操作，那麽窄內容也是與此相關的，因爲正是大腦狀態的窄內容，纔決定下一個腦狀態的內容是什麽。這自然產生了這樣一些問題——對於腦狀態什麽是要緊的？什麽使得它擁有一個心理狀態而其他沒有？在這個所涉及的實體的心理操作範圍內，功能角色起到何種重要作用？這些是功能主義的替代方案試圖說明的問題。

三　功能主義的“腦替代”實驗能説明機器產生意識嗎？

在機器是否有心理狀態的問題上，功能主義認爲，心理狀態是輸入與輸出之間的任何一個中介因果條件。根據功能主義，任何兩個系統，若它們具有同構的因果過程，則具有相同的心理狀態。因此，一個計算機程序能夠擁有與人相同的心理狀態。這裏的同構，是指兩個不同系統在結構和屬性方面的一一對應，在數學上是映射關係。這個假設無論正確與否，都表明存在某種水平的抽象，在這個抽象框架下的操作是無關緊要的。

功能主義的這種觀點可由“腦替代”（brain replacement）[①]思想實驗得到清晰的說明。這個實驗包含三個假設：（1）神經生理學發展到這樣的程度——人腦中所有神經元的輸入輸出行爲和連通性被完全理解；（2）人們能夠建構微型電子裝置，它能夠模仿整個行爲，能夠順利連接神經組織；（3）某些神奇的外科技術能夠用相應的單子裝置替代個體神經元，而不中斷腦作爲一個整

① 該實驗最初由哲學家格賴默（Clark Glymour）提出，由塞爾加以發展，最終由機器人專家莫拉維克（Hans Moravec）引入人工智能領域。

體的操作。一句話，這個實驗是由使用電子裝置逐個替代人頭腦中的所有神經元所構成。

在這裏，需要關注的是在操作之後和操作期間“被試者”（subject）的外在行爲和內在體驗。根據實驗的定義，如果這個操作不被執行的話，與所將要觀察到的情況相比較，被試者的外在行爲必須是保持不變的。雖然意識的在場或缺場不能被第三個當事人確定，實驗主體至少應該能夠記錄他/她自己有意識經驗中的任何變化。顯然，對於接下來將發生什麼會存在一個直接的直覺衝突。但是，功能主義者相信，他們的意識仍然會保持不變。①

作爲一個生物功能主義者，塞爾也相信他的意識會消失。在他看來，人們會失去對其外在行爲的控制。他舉了這樣一個例子，當醫生測試你的視力時，會在你面前舉一個紅色物體，然後問你看到了什麼？你本來想說你沒看見任何東西，但你完全不受控制說出你看見一個紅色物體。這意味着，你的有意識經驗慢慢地消失，而你的外在可觀察行爲仍然保持不變。

這有兩種情形需要注意：一方面，當被試者逐漸成爲無意識時，要保持外在行爲不變，被試者的意願同時完全被取消；否則，意識的消失將會反映在外在行爲中，如被試者會大叫或以言辭表達。如果將意願的同時消失看作是某一時刻的神經元逐漸替代的結果，這似乎是一個不太可能的主張。另一方面，在沒有真實的神經元保留期間，如果問被試者關於他的有意識經驗的問題，那麼會發生什麼呢？假設一個正常人被尖棍子戳了一下，他會發出尖叫；若一個失去知覺的人被尖棍子戳了一下，他可能沒有任何反應。機器就像一個沒有知覺的人，被戳一下也沒有反應，因爲它沒有神經元，儘管它可能被設計成有刺激—反應的程序做出應答。假如我們用電子腦替代正常人腦的功能屬性，而且電子腦沒有包含任何人工智能程序，那麼我們必須擁有顯示意識的一個解釋，這種意識是僅由訴諸神經元功能屬性的電子腦產生的。因此，這種解釋也必須應用於具有相同功能屬性的人腦。

這會導致三種可能的結論②：（1）正常人腦中產生這類輸出的意識的因果機制仍然能在電子裝置中操作，它因此是有意識的。（2）正常人腦中的有意識心理事件與行爲沒有任何因果聯繫，並從電子腦中遺失，它因此不是有意識的。（3）這個實驗是不可能發生的，因此關於它的推斷是無意義的。儘管不排除第二種可能性，但它將意識還原到副現象的地位，即哲學上描述的某物發生了但沒有留下影子，好像它存在於可觀察的世界。進一步說，如果意識的確是副現象的，那麼被試者在被戳痛後就不會有發出尖叫的情況發生，因爲不存在有意識經驗的痛。相反，人腦可能包含一個次要的、無意識的機制，該機制負責在受到刺激後會發出尖叫聲。

總之，功能主義主張，在神經元水平上操作，意味着也能夠在任何更大功能單元如一組神經元、一個心理模組、一片腦葉、半腦或整個腦水平上操作。這意味着，如果接受腦替代實驗說明替代腦是有意識的觀點，那麼也應該相信，當整個腦被電子裝置替代後意識被保持下來了，而且這個電子裝置通過一個查找表從輸入到輸出的地圖不斷升級其狀態。如果功能主義是對的，這會使大多數人包括弱人工智能者感到不安，因爲查找表不是有意識的，至少在查找表期間產生的有意識經驗，與在操作一個可能被描述爲存取和產生信念、反省、目標等的系統期間產生的有意識經驗是不同的。因此，腦替代實驗並不能充分說明機器能夠產生意識。

四　生物自然主義駁倒了功能主義嗎？

20世紀80年代，塞爾的生物自然主義開始流行，它對功能主義提出了強烈挑戰。根據生物自然主義，心理狀態是高層次的湧現特徵，它由低層次的神經元中的生理物理過程引起。神經元的這種未指明的屬性纔是重要的。因此，說心理狀態被複製，僅僅是在部分具有相同輸入輸出行爲

① J. R. Searle，*The Rediscovery of the Mind*（Cambridge，Mass: MIT Press，1992）.

② S. J. Russell，P.Norvig，*Artificial Intelligence: A Modern Approach*（third edition，北京：清華大學出版社，2011），第1030頁。

的某些功能結構程序基礎上，而且會要求這個程序在執行一個與神經一樣有因果力的認知構架。爲了支持這種觀點，塞爾描述了一個假設系統，這就是著名的“中文屋”。

這個假設的中文屋系統由一個不懂中文祇懂英文的人、一本英文寫的規則書（«英漢對照詞典»）和一堆紙條（有些是空白，有些寫有不認識的符號）構成。中文屋有一個通向外面的視窗，負責收取紙條。屋中的人相當於計算機的中央处理器（CPU），規則書相當於程序，紙條相當於儲存器，視窗相當於輸入輸出裝置。屋中的人能夠將通過視窗傳入寫有中文符號的紙條，按照規則書的匹配指令（形式規則）將其編譯成中文語句，並將寫有中文的紙條傳遞出去。指令可能包括在新紙條上寫符號，在紙堆中發現符號，重新安排紙堆等。這個中文屋系統輸入中文紙條並產生中文回答，看上去與圖靈設想的系統一樣有智能。目前的翻譯程序如百度翻譯、谷歌翻譯等，就相當於中文屋系統。

對此，塞爾認爲，中文屋中的人，比如我塞爾本人，完全不懂中文，規則書和一堆紙條不過是一張張紙條，根本不理解中文。這個系統根本不存在對中文的理解，即執行一個正確的程序，不必然產生理解。[①]塞爾假設“中文屋”系統要說明的是，執行一個適當的程序儘管也產生正確的結果，但對於成爲一個心智是不充分的。

然而，機器是否有“智能”，與機器是否“理解”，實際上是不同層次的問題。按照計算主義的定義，若一個系統能夠計算，它就應該有智能，因爲計算就是思維，思維當然是智能行爲。但如果將智能定義爲包括“理解”在內，則機器系統就難以有智能，因爲任何程序無論是形式的還是非形式的，機器系統祇能執行程序並不理解程序本身，甚至人有時也祇是使用某種語言，不一定理解其意義，如同兒童背誦唐詩，成年人使用0（零）。因此，這就看如何定義、理解“智能”概念了。有智能與有意識、有心靈、能理解，還不是一回事，雖然它們之間有關聯。

嚴格講，塞爾所反對的不是弱人工智能的觀點，而是強人工智能論斷——恰當程序設計的機器具有認知狀態，程序本身就是對人的認知過程的理解。塞爾在《心靈、大腦與程序》一文中，通過設想一個中文屋，從系統應答、機器人應答、腦模擬者應答、聯合應答、他人心靈應答、多重套間應答六個方面系統而詳細地反駁這種論斷，客觀上說，還是非常有力的。[②]然而，將人在中文屋中理解中文，類比爲“CUP”能夠開立方運算，是否合適？這是許多強人工智能支持者質疑塞爾反駁的一個普遍問題，也是人工智能專家對哲學界的挑戰。

在中文屋與“CPU”兩種情形中，對於“理解”而言，都是否定的。因爲，人不理解中文，“CPU”也不理解。若問中文屋是否理解中文，按照強人工智能，回答可能是肯定的。塞爾的應答，實際上重申了這樣的觀點——在人腦中不理解的，在紙中也不能理解，所以不能存在任何理解。塞爾似乎是在說，整體的屬性必須存在於部分屬性中，這顯然不合適。比如，水是濕的，但其組成的成分H_2O都不是濕的。

塞爾的主張基於四個公理[③]：（1）計算機程序是形式的（句法的）；（2）人類心智基於心理內容（語義的）；（3）句法本身對於語義學既不是必要的也不是充分的；（4）生物腦產生心智。塞爾從前三個得出結論，程序對於心智是不充分的，即一個執行程序的主體可能是一個心智，但僅根據執行程序並不必然是一個心智。從最後一個公理得出，能夠產生心智的任何其他系統可能會擁有等同於腦的因果力的因果力。由此，塞爾推知，任何人工腦將會擁有複製腦的因果力，不僅僅運行一個特殊的程序，但人腦不會僅根據運行一個程序就產生心理現象。一句話，意識或心靈是生物現象，計算程序是不能產生這種現象的。

但是，這些公理是有爭議的。例如，公理（1）（2）依賴於一個句法與語義學之間的未詳細說

① J. R. Searle,“Minds,Brains and Programs”, *Behavioral and Brain Sciences* 3(1980): 417-457.
② [英]瑪格麗特·A·博登:《人工智能哲學》(上海: 上海譯文出版社，2001)，劉西瑞、王漢琦 譯，第92—120頁。
③ J. R. Searle,“Is the brain’s mind a computer program?”, *Scientific American* 262(1990): 26-31.

明的區分，這種區分似乎與寬內容、窄內容之間的區分密切相關。一方面，人們可以將計算機看作操作句法符號；另一方面，也可以將計算機看作操作電流，這恰好與大腦運作的情形（生物電流）相似。所以，人們會說腦就是句法的。但是，程序對於心智是不充分的這個結論並不令人滿意。塞爾所主張的是，如果你明確地否認功能主義，即公理（3）表明的，那麼你就不會必然得出沒有腦是具有心智的。這樣，中文屋論證可以歸結爲是否接受公理（3）。

丹尼特（Daniel C. Dennett）將中文屋論證稱爲“直覺泵”（intuition pump）[①]，即中文屋論證放大了人的先驗直覺。他認爲，塞爾的論證形式對哲學家來說是很熟悉的，即他構建了一種所謂的直覺泵，一種通過在基本思想實驗上產生變異來激發一系列直覺的裝置。“直覺泵”通常不是發現的引擎，而是一種說服者或教學工具，一種一旦你看到真相就能讓人們以你的方式看待事物的方法[②]。丹尼特反對用直覺泵來思考，認爲它被許多人濫用。在這種情況下，塞爾幾乎完全依賴於錯誤的結果，即由錯誤地提出的思想實驗產生的有利直覺。因此，生物自然主義者堅信他們的立場，而功能主義者僅確信公理（3）是未經證明的，或者說，塞爾的論證是不足以令人信服的。然而，不可否認的是，中文屋論證不僅對人工智能產生了巨大挑戰，也引發了廣泛的爭論，但很少改變持不同立場人們的觀點。比如，博登就旗幟鮮明地反對塞爾的“中文屋”論證，認爲他的論斷是錯誤的。[③]

審視這場爭論可以發現，那些接受公理（3）進而接受塞爾論證的人，當決定什麼實體是心智時，僅僅依賴他們的直覺而不是證據。中文屋論證表明，中文屋憑藉執行一個程序的力量不是一個心智，但該論證沒有說，如何憑藉某些其他理由決定中文屋或計算機或機器人是不是一個心智。不過，塞爾承認人類這種生物機器有心智。按照塞爾的這種觀點，人腦可能或不可能執行像人工智能程序的某些東西，但如果人腦能執行，那也不是它們是心智的理由。在塞爾看來，創造一個心智需要更多的東西，比如，相當於個體神經元的因果力的某些東西。但是，這些力是什麼，仍然是待解之謎。

然而，需要注意的是，從生物進化的角度看，神經元進化出執行功能角色，即具有神經元的生物遠在意識出現於自然界之前就學習和決定了。若這樣的神經元由於與它們的功能能力無關的某些因果力而恰好產生了意識，那將是一件驚人的同時存在事件。畢竟，正是這種功能能力纔支配了有機物的生存。而在中文屋的情形中，塞爾依賴直覺而非證據，即衹看到屋，沒有證據證明屋中到底有沒有意識發生。我們也可以將大腦看作屋而做同樣的論證，即衹看到細胞的組合，盲目地根據生物化學或物理學規律操作，那裏有心智嗎？爲什麼一大塊腦能產生心智而一大塊肝臟卻不能？這仍然是一個巨大的秘密，是一種解釋鴻溝。因此，僅僅依靠哲學的思考還不足以弄清心智的形成機制，這需要多學科的聯合。

五　意識、感受性的解釋存在解釋鴻溝嗎？

對於強人工智能的爭論，其核心是意識問題。具體說，意識是否衹是生物特性的，非生物的人工意識或機器意識是否可能實現？意識通常表現爲不同方面，如感受性、理解、自我意識、自由意志等。這些方面都是哲學上一直在討論的問題，至今仍在爭論。而與意識問題緊密相關的是主體經驗，即感受性，它是人們經驗的內在性質。人們的某個感覺如疼痛是否有相應的腦狀態？或者說，一個心理狀態是否有相應的物理狀態？這個問題對心智的功能主義說明構成了挑戰，因爲不同的感受性可能涉及別的同構因果過程是什麼的問題。例如，“倒置光譜”（inverted spectrum）思想實驗表明，當看見紅色物體時某人“x”的主體經驗，與當其他人看見綠色物體時

① D. C. Dennett，*Consciousness Explained*（New York: Penguin Press，1991），438.
② D. C. Dennett，*Intuition Pumps and Other Tools for Thinking*（New York:W.W.Norton & Company，2013）.
③ [英]瑪格麗特·A·博登：《人工智能哲學》，第121—141頁。

的主體經驗相同，相反也一樣。“x”堅持紅色是紅的，在紅色信號燈亮起時停，也同意紅色信號燈的紅色比夕陽的紅色更強烈。然而，“x”的主體經驗在兩種情形中是不同的。紅色信號燈是停的經驗，夕陽是享受或感悟的經驗。

感受性問題不僅是心靈哲學的一個重要問題，也是自然科學的一個重要問題，不僅對功能主義形成挑戰，也對科學形成挑戰。假設神經科學已經探明大腦的運作機制，比如發現一個神經元中的神經過程將一個分子轉換爲另一個，不同神經元之間的相互連接路徑等等，這些發現也不足以讓人們接受，擁有神經元的實體如同大腦有任何特殊的主體經驗。在神經過程與意識形成之間可能存在某種鴻溝。這在哲學上被稱爲“解釋鴻溝”（explanatory gap），它是心身問題中關於物理現象與心理現象之間的理解關係問題，具體說是能否用物理狀態解釋心理狀態的問題。這種解釋鴻溝，導致有人如查爾莫斯（David J. Chalmers）認爲，人類無能力完全理解自己的意識[①]；也有人如丹尼特通過否認感受性進而否認解釋鴻溝的存在，認爲這是由於哲學混亂造成的。[②]但是，無論結論怎樣，問題依然存在，爭論仍在繼續。

其實，如果感受性是意識經驗，則主體經驗與意識相關，沒有主體意識也就沒有主體經驗。例如，一個喪失意識的植物人沒有感覺，也就沒有主體經驗。如果感受性是非意識經驗，那就有點神秘了，畢竟我們不能否認有意識的人有主體經驗，主體經驗就是感受性。之所以存在解釋鴻溝，是由於還沒有弄清生物的大腦是如何擁有意識的，更遑論機器腦是如何產生意識的。況且，意識與感受性之間的聯繫機制也還沒有完全弄清，因此，形成所謂的解釋鴻溝也就不難理解了。圖靈也承認，意識與機器智能是不同的問題，但他否認意識問題與人工智能有太多的聯繫。也就是說，意識問題不妨礙機器智能問題。當今人工智能的長足發展充分說明，意識與智能不是一回事，有意識的存在一定有智能，如人類；但有智能的存在不一定有意識，如智能機。也許人工智能並不依賴於意識，而是依賴於能創造智能行爲的程序。意識可能是一種不需要智能的更高級精神現象。

六　人工智能能否對社會倫理構成威脅？

不論智能與意識關係如何，也不論人工意識是否可能，一個顯見的事實是，人工智能的迅速發展已經對人類社會產生了非同凡響的影響，這是自我認知領域的一場深刻變革，有人稱之爲“圖靈革命”，也是繼哥白尼革命、達爾文革命、神經科學革命後的“第四次革命”[③]。這次革命之所以與以往不同，在於它將人類看作一個信息體，在信息圈內與其他可邏輯化、自動化信息處理的信息智能體共享自然和人工領域內的成就，相互交織在一起。於是，人類越來越多地將記憶、認知活動甚至日常生活，委託給智能機如計算機、智能手機等來完成，智能機已然成爲人類的“延展大腦”。如果沒有手機，人們將無所適從；如果電信網絡停擺，人們將無法完成購物支付。除認知、經濟等領域外，人工智能特別是智能型機器人的發展是否會對人類的精神領域，特別是倫理觀念和行爲規範產生影響？這一點是肯定的，而且影響巨大。這可以從以下四方面來探討。

一是智能型機器人的身份認同問題。這不僅是個法律問題，更是倫理問題。一個人形智能型機器人是否被承認是人類，即使法律上的障礙被消除，比如，2017年10月28日，沙特阿拉伯向機器人“索菲婭”授予國籍，宣佈機器人索菲婭爲其國家的公民，享有與其國民相同的權利。這就是承認機器人的合法人類身份的地位。然而，接下來更棘手的倫理層次的問題是——“索菲婭”與

① D. J. Chalmers, *The Conscious Mind: In Search of a Fundamental Theory* (New York: Oxford University Press, 1996).
② D. C. Dennett, *Consciousness Explained*, 438.
③ [意]盧西亞諾·弗洛里迪：《第四次革命：人工智能如何重塑人類現實》（杭州：浙江人民出版社，2016），王文革 譯，第101—112頁。

人類是什麼關係？她能結婚嗎？儘管“索菲婭”與人類外形高度相似，擁有仿生橡膠皮膚，可以模擬62種面部表情，能識別人類面部表情、理解人類語言、與人互動，甚至還會開玩笑，但她畢竟還是機器人，缺乏人類擁有的情感力和自然生育能力，更沒有人類長期建立起來的倫理觀念，因此，對機器人提出倫理要求本身就是不合理的，賦予機器人以所謂的合法身份也就是一場鬧劇。

二是智能型機器人的安全性問題。如果機器人會對人類生存產生極大的威脅，如無人駕駛機器成爲殺人武器，禁止其發展就是必須的，因爲人類法律和道德都絕對不允許殺人。這就是人類如何控制機器人的問題。2018年3月9日，歐洲科學與新技術倫理組織發佈《關於人工智能、機器人及“自主”系統的聲明》稱，人工智能、機器人技術和“自主”技術的進步已經引發了一系列複雜和亟待解決的倫理問題，呼籲爲人工智能、機器人和“自主”系統的設計、生產、使用和治理制定國際公認的道德和法律框架。

三是人類道德責任問題。在這種由人工智能和機器人構成的複雜信息社會技術系統中，與道德相關的“智能主體”（agency）應該有怎樣的位置？人類如何分擔產生的道德責任，或者由誰爲其中產生的不良後果負責？不可否認，智能機是人類自己創造並推廣應用的，產生的不良後果應該由人類自己來承擔。具體來說，誰生產誰負責；正如環境問題，誰污染誰治理。這不僅是一個涉及對人工智能的研發、設計、測試、生產、監管、認證的問題，也是一個包括政府、企業、個人在內民主決策、協調解決的問題，涉及制度、政策及價值觀的決策，以確保人工智能技術不會給社會帶來危害。例如，哈佛大學肯尼迪政府學院貝爾弗科學與國際事務中心與美國銀行宣佈成立“人工智能責任運用協會”，旨在解決未來人工智能快速發展中可能出現的問題。①

四是人類的行爲規範問題。人工智能發展不僅極大地改變了人類的生活方式，也改變了人類的行爲方式。目前的互聯網給人類帶來了極大的便捷，網購、支付寶、外賣非常普及，機器人已代替人類的部分工作，人變得休閑了，但也無所事事了，宅男宅女普遍存在，幾天甚至幾週不下樓的大有人在。這又帶來了社會問題：人與人之間面對面的交流、溝通沒了，感情淡化了，冷漠成爲常態，不僅導致了網癮、安全問題，也產生了不良行爲。比如，走路看手機導致的車禍，長期宅在家裏造成的交流封閉，甚至產生自閉症。假如有一天智能型機器人真的普及了，人類應該如何與它們打交道？應該建立一種怎樣的關係？人類面對機器人應該如何規範自己的行爲？這是不久的將來人類會面臨的社會和倫理問題。

概言之，對於人工智能，人類不僅要關注它如何發展，還要考慮它應該怎樣發展。如果人工智能的發展對於人類是弊大於利，如果機器人將會控制人類並要毀滅人類，人工智能研究者就有責任終止這種研究，如同終止核武器的發展一樣。這不可避免地給人工智能研究者提出了社會責任和社會倫理的要求。

〔編者註：此文是作者主持的中國國家社會科學基金重點項目“科學認知的適應性表徵研究”（16AZX006）的階段性成果。〕

① “2018傳媒產業五大熱點——比大趨勢更重要的是非顯著趨勢（上）”，《媒介》2018－07－16。