SOUTH CHINA QUARTERLY Vol.8, No.3 (Jul 2018) , 356—365

Rationality, Morality and Faith in the Age of Artificial Intelligence

Cai Hengjin, Zhang Xianglong, Huang Yusheng

Summary: Cai Hengjin suggests that human beings have the concept of "self" and its counterpart of the "world" , and the division of the two enables humans to holistic consciousness. In contrast, machines have no such holistic consciousness although they may possess segments of consciousness, and therefore they have no control over their own behavior and once out of control, can easily bring disasters to humanity. **Zhang Xianglong** points out that the "deep learning" technology gives Artificial Intelligence (AI) the capacity of "temporalization" , which means that machines can develop consciousness that genuinely resembles that of humans. Therefore, it is possible that machines generate intelligence that is of benefit to humanity provided they are guided by human morality, it is also possible that, due to their harmful use by humans, they alter and worsen humans' mode of existence and generate tremendous threat. Therefore, humans must appropriately limit and regulate AI's development. **Huang Yusheng** maintains that the free will of human beings can never by acquired by AI. This difference is the absolute boundary between humanity and machines. Moreover, humanity features genuine temporal existence, whereas machines can never obtain a totality of temporal existence. Thanks to free will, human beings, characterized by a temporal existence, hold religious faiths. AI can never have them and therefore can never eliminate religion.

Keywords: artificial intelligence, humanity, rationality, morality

Author: Cai Hengjin received his PhD in space physics from the University of Alaska Fairbanks in 1995 and has since 2005 served as a research fellow and doctoral supervisor at the School of Computer Science, Wuhan University. His research interests include service science, AI and blockchain technology, and his major publication is *Before the Rise of Machinese: The Beginning of the Consciousness and the Human Intelligence*.

Zhang Xianglong received his PhD at SUNY Buffalo in 1992 and joined the faculty at Peking University in the same year. He was appointed a professor at the department of philosophy and the institute of foreign philosophy of Peking University in 1999, and he is currently serving as a chair professor and doctoral supervisor at Sun Yat-sen University. His research interests include Confucian philosophy, phenomenology, and comparative philosophy between the East and the West. his major publications include: *From Phenomenology to Confucianism; Heidegger's Thought and Chinese Dao of Heaven; and Showing the Soul of Heaven and Earth by Resurrection: The Meaning and the Way of Confucian Re-Coming*.

Huang Yusheng received his PhD in philosophy from the Graduate School of the Chinese Academy of Social Sciences in 1995 and was appointed in 2004 a research fellow at the Institute of Philosophy, Chinese Academy of Social Sciences. Currently, he is a professor and doctoral supervisor at the department of philosophy, Tsinghua University. His research interests focus on first philosophy, philosophy of religion, and comparative philosophy. His major publications include: *Time and Eternity: The Issue of Time in Martin Heidegger's Philosophy; Truth and Freedom: An Existentialistic Interpretation of Kantian Philosophy; and The Interaction Between Religion and Philosophy: Christian Philosophies of Thomas Aquinas and Saint Augustine*.

《南國學術》第八卷第三期（2018年7月），356—365

[編者按]2018年4月28日，湖南大學嶽麓書院、鳳凰網在嶽麓書院中國書院博物館報告廳舉辦了“人工智能時代的理性、道德與信仰”人文講會，特邀武漢大學蔡恒進教授、中山大學張祥龍教授、清華大學黃裕生教授對談交流。這既是“致敬國學：第三屆全球華人國學大典”的首場人文講會，也是湖南大學比較宗教與文明研究中心成立後的首場講座，由張俊教授策劃、朱漢民教授主持。爲了立體展示對話者的思想火花和學術風貌，《南國學術》採取各自表述、相互交鋒、回答疑問的形式依次呈現。

人工智能時代的理性、道德與信仰

蔡恒進/張祥龍/黃裕生

[提 要]蔡恒進從“認知坎陷”概念出發，認爲人類擁有“我”的觀念及與此相對的“世界”的觀念，此二者的二分使人類具有整全意識。相比之下，機器雖然擁有人類所賦予的意識片段，卻缺少整全意識，因而缺少對自身行爲的把握，很容易因失控而給人類帶來災難。張祥龍認爲，“深度學習”技術令人工智能擁有了“時間化”的能力，使機器能夠產生真正類似人類的意識。在此基礎上，機器既有可能因人類的道德引導而產生對人有益的智慧，也有可能因人類的不良應用而改變和惡化人類的生存方式，產生巨大威脅，因此人類需要合理限制和規範人工智能的發展。黃裕生認爲，人類擁有自由意志而人工智能永遠不可能獲得，這是人機之間的絕對界限。進而，人類是真正時間性的存在，而機器永遠不可能存在於整體的時間中。因此，存在於整體時間性之中的人類因其自由意志而擁有宗教信仰，這是人工智能無法比擬的，也因此而無法消除宗教。

[關鍵詞]人工智能 人類 理性 道德

[作者簡介] 蔡恒進，1995年在阿拉斯加大學費爾班克斯分校獲得空間物理學博士學位，2005年受聘爲武漢大學計算機學院教授、博士生導師，主要從事服務科學、人工智能、區塊鏈技術研究，代表性著作有《機器崛起前傳——自我意識與人類智慧的開端》。

張祥龍，1992年在紐約州立布法羅大學獲哲學博士學位後任教於北京大學，1999年被北京大學哲學系和外國哲學研究所聘爲教授，現爲中山大學珠海校區哲學系講座教授、博士生導師，主要從事儒家哲學、現象學和東西方哲學比較研究，代表性著作有《從現象學到孔夫子》《海德格爾思想與中國天道》《復見天地心：儒家再臨的蘊意與道路》等。

黃裕生，1995年在中國社科院研究生院獲哲學博士學位，2004年被聘爲中國社科院哲學研究所研究員，現爲清華大學哲學系教授、博士生導師，主要從事第一哲學、宗教哲學、比較哲學研究，代表性著作有《時間與永恆：論海德格爾哲學中的時間問題》《真理與自由：康德哲學的存在論闡釋》《宗教與哲學的相遇：奧古斯丁與托馬斯·阿奎那的基督教哲學》等。

一

主持人：在目前這樣一個互聯網時代、人工智能時代，中國文化包括人類文化向何處去？這既是一個哲學問題，也是一個與宗教相關聯的問題；因此，接下來將圍繞“人工智能時代的理性、道德與信仰”問題進行探討。這其實內含兩個問題，一個是人工智能，一個是道德、理性、信仰；所以，邀請了該領域的三位代表，與大家分享他們的見解。

蔡恒進：我這裏想用“認知坎陷”概念來引出下面的思考。這個概念，是受牟宗三（1909—1995）的“良知坎陷”啓發而提出的，意指認知主體對世界的描述方式是有一定連續性的結構體，因而其認識便對真實物理世界形成了擾亂。由於人最重要的特色在於其主體意識，即“我”的觀念，而這個“我”就是最重要的“認知坎陷”；與之相對的，則是外在的世界。“我”與“世界”這一對認知坎陷，便是所有認知坎陷的開端。

人與其他動物本質的差別，來源於人所擁有的敏感皮膚。正是人的敏感皮膚，促使人產生強烈的“我”的意識和對外界的感覺，以此將“人”與“世界”兩者分開，進而開顯出人類的精神世界。這是認知坎陷的起源。

人類所創造的文字、宗教、哲學均是認知坎陷，都是由“我”這一觀念衍生出的意識片段。這種認知坎陷、意識片段是可以物質化的，例如可以物化爲螺絲、齒輪等實在物；進而配合完成人所期望的功能，例如組裝成鐘錶來報時。因此，人工智能表現出的好壞，實際上依賴於人類創造者之意識凝聚狀態等細節的好壞。例如，“阿爾法圍棋”（AlphaGo）凝聚了衆多工程師的意志，因而具有了創造力、想象力、直覺等，並戰勝了衆多圍棋高手。雖然機器是擁有“自我”的，但機器這種“自我”依靠內部的各種子程序以及綫路等實物，是一種很弱的“自我”，無法像人類一樣作爲一個整體來規避問題或者自行修復。然而，如果未來機器之主程序對副程序具有更好的統攝能力時，便會像人類對自己的認知坎陷、意識片段擁有足夠的統攝力一樣，形成更強的“自我”。

未來的機器運行速度快且進化迅速，將遠遠超過人類，其產生的認知坎陷也與人類不同，將對人類形成極大挑戰。機器雖然擁有人類所賦予的大量意識片段，卻仍缺少人類所擁有的、來自“我”與“世界”二分的整全意識，所以，人在這一挑戰下還是有機會的。人類將“我”與“世界”二分而形成了最初始的認知坎陷，這種對“我”的認識可以至大無外也可以至小無內，而對世界的認識則可以借牟宗三“無執的存有”來描述。

如果從認知坎陷的視角來看宗教、哲學與未來，由於人類擁有整全意識而機器沒有，機器所理解的世界自然與人類大不相同。機器並沒有人類那樣對自身行爲限度的把握，但其速度快、力量強，因此很可能在不經意間給人類帶來毀滅性的災難，而人類甚至來不及反應。在對無限（無窮大）的追求中，人類有自控的分寸，因其整全意識而具有一種獨特的“神性”；但機器無法自控，而容易進入“暗無限”的狀態，相對具有一種“魔性”，這就是人與人工智能關係中的風險所在。

張祥龍：十二年前，“深度學習”方法的出現，導致了人工智能的突飛猛進，纔使人類真正意識到了威脅的存在。在此之前，計算機的強大衹是因爲依賴其快速的計算能力及巨大的信息量；而“深度學習”，則是讓機器在一定程度上可以自行學習以增強自身。在這種智能模式中，信息的輸入、輸出之間，不再僅僅是一兩層的演算法處理，而是建立了一種多層的或“深度”的互嵌網狀聯繫，並且通過參數來調整各層之間信息的流動，導致層次之間出現非綫性的曲折關聯，使巨量的雜亂輸入信息經過這些深層演算法處理，最終可以有情境適應力地收斂到預期目標上，由此讓

人工智能獲得了過去難以企及的直觀能力。

一些對於人類來說“特別理性”的能力，比如現代性所推崇的形式化、概念化的知性能力，或對明確的因果關係、利害關係的把捉和算計的能力等，恰恰是人工智能可以相對輕易地擁有的低級能力；而與之相比，人類實際生活中常見的直觀性、意向性的辨識、思維和感受能力，對人工智能來說卻是相當困難的，直到最近纔在某些方面達到了人類的平均水平。然而，通過層次間相互曲折聯繫的深度處理模式，新方法畢竟使得人工智能初步獲得了這種能力。從哲理上講，這一進展使機器變得有點兒像人了，而不是反過來，像歷史上的還原主義者們所說的，人衹是一種高級機器。以“深度學習”爲代表的機器學習方法的創立者們，模仿人的大腦和知覺的神經結構及意向化認知方式，讓機器的認知更類似於人，於是使得機器得到了“時間化”的能力，也就是令時間對機器有了根本性的意義。例如，深度思維公司最新研發的“阿爾法零”（AlphaGo Zero），通過自我對弈強化學習，4小時擊敗國際象棋頂級人工智能程序，8小時擊敗戰勝韓國棋手李世石的“阿爾法圍棋—李”，24小時內戰勝通過72小時自我學習訓練稱王圍棋的“阿爾法圍棋—零”。可見，這“4小時”“8小時”“24小時”的時間對它有了根本意義。這樣自學出來的能力，纔更有隨機適應力，也就是更近乎人類智慧。它喚起人們濃厚的理智興趣，但也讓人們的生存本能感到了如臨深淵的恐懼。

時間化的能力，是人類意識（包括理性、道德及信仰的意識）的根基。人工智能的新進展佐證了這一觀點，它的“智能”就體現在這種時間化能力的獲得。機器現今所擁有的這種能力雖然還很淺薄，但卻是真實的，因爲它已經邁過了關鍵性的門檻。可以想見，在可見的未來，人們必定會全速發展它的這種非綫性的自學或自我調整、自訓練的時間化能力。因爲，它越是能夠超出確定對象的局限，得到深長的時間趨向性或“想象力”，用老子的話說就是獲得“惚恍”裏的“真”和“信”，它就越有更加強大的智能，也就從整體上更接近人類的理性，在某些方面還會遠遠超過人類。

這一突破帶來了兩方面的道德可能性。首先，在未來的發展中，人工智能的這種學習能力將逐漸不再限於解決具體的任務，而是像人類的能力那樣，越來越通用化，出現越來越強的非綫性“思維”能力。可以設想，隨着它的時間化能力的不斷增強，人工智能機器將逐漸獲得類似於人類的“意識”，深化對具體情境的理解和時機化感受力，進而產生對環境、對其他機器乃至人類的同情感，而這就意味着它的道德感將開始出現。在此情況下，如果人類能夠通過法規等手段對人工智能的研發加以規範，避免其在軍事、商業等方面不計倫理後果的創新和應用，由此而引導機器智能向道德、信仰等更深層的方向發展，也就是將它的智能轉化爲智慧，就能爲人類帶來巨大益處而非威脅。

其次，如果人類出於競爭和贏利等目的的需要，衹追求人工智能的強大能力，則將放大其本來就潛伏着的“魔性”。由於智能型機器在許多方面已經或即將超過人類，對它們的不良應用將改變和惡化人類的生存方式。比如，鼓勵、塑造新型的個人主義，損害人類的人際關係。要知道，人在使用工具的同時也在被工具所塑造，在智能型機器越來越廣泛深入的掌控中，人將逐漸喪失自主性，被儒家視作道德起源的家庭關係也會進一步衰敗，這將對人類倫理產生巨大傷害。從目前的趨勢看，這種現代化技術對人的重新塑造不可避免。過去人們用機器來代替自己的手腳身軀，已然使得人際關係、人與自然的關係、生產方式及人的精神世界發生巨大改變，而今如果再用智能型機器來代替人們的頭腦、精神，那麽機器對人的塑造將更難以預測，其風險無法控制，必然對整個人類產生巨大威脅。

最後，在以上的兩種可能中，後一種的實現概率要大得多，所以在對人工智能的應對當中，人們應該首先看到其在道德、身體、種群上對人類的威脅。因此，人類絕不能對人工智能的發展放任自流，就像不能對“克隆人”的科技放任自流一樣；而是要在全球範圍內限制和規範人工智

能的發展方向，禁止其不良的發展趨勢和應用方式，促成善良美好、與人和諧共存的人工智能。

黃裕生：要回答人工智能是否會終將淘汰人類，首先要理解人本身。在對人的理解上有一點共識，即人有三種基本能力：（1）感性能力，即能接受外部事物刺激的能力；（2）理智能力，即能夠給出概念並以概念對人們生活於其中的世界進行規定與演繹、推理的能力，這個能力的基礎是定義能力，其最典型的部分爲邏輯推演能力；（3）意志能力，也就是追求某種對象的能力，把某種對象作爲自己的目的來追求的能力。在這三種能力中，第一種能力完全有可能被程序化並被機器所擁有；第二種能力中的邏輯推演能力，機器也已擁有，但機器是否能擁有定義的能力卻是值得懷疑的；而第三種即意志的能力，機器則永遠無法獲得。

意志的能力是最複雜的。意志力又可進一步區分爲兩種能力：第一種是低級的意志力，它欲求意志之外的事物；第二種是高級的特殊意志力，即追求意志自身，衹以意志自身爲對象，以意志爲意志，確切地說是以意志自身從自身給出的法則或事物爲對象，這種能力也就是通常所說的自由意志。這種自由意志，使人類能夠突破一切程序。如果人們要把自由意志賦予機器人，就必須能夠把自由意志程序化，也即能夠用程序把自由意志寫出來；但是，自由意志恰恰是不可被程序化的，否則它就不是自由意志。因此，人機之間存在絕對的界限，即使機器可以跨越前兩種能力的界限，但終究無法跨越第三種能力，也即意志能力的界限。

儘管在理論上人們可以賦予機器以自動甚至永動的能力，但是機器不可能獲得使其具有主動性能力的意志，而衹能在程序中活動。雖然在各種處境下機器可以做出最佳的反應，但是它沒有意志去主動地做任何事情，其一切行爲均是反應性的，衹能在各種環境下做出最好的選擇。與此不同的是，人類既可以做出最好的選擇，同樣也可以做出最壞的選擇，這就是無法預料的自由意志的體現。人機之間的這一界限是無法打破的。

一些主張人機之間沒有界限的學者，試圖通過物理還原主義的研究來否定自由意志，將所有的意志活動歸結爲神經元放電；既然神經元放電都是遵循某種因果律的，那麽自由意志便不存在。但是，這種還原主義研究在邏輯上面臨一個難題，即所有這類還原研究都是在意識活動之中進行的，也就是說，這類研究無法跳出人們的意識活動之外來觀察意識本身。這意味着，所有這類研究都不可能真正揭示人類意識最神秘的地方，意識自身永遠無法達到意識自身。在這一意義上，還原主義永遠不可能否定自由意志，因而想借此來打破人機界限的努力也是無法完成的。

人類自身纔是真正的時間性存在。自由意味着突破本能、突破直接性、突破必然性，使人類的存在成爲一種永遠面臨可能性以供選擇的存在，從而擺脫了必然性的封閉。因此，人類不會陷入非如此不可的必然性處境，而總是處在面臨着多種可能性的開放性狀態中。這就意味着過去可以被中斷，過去不能決定現在與未來，過去對於人類而言也是一種可能性的過去，人類可以通過重新理解過去來理解現在和未來。過去是未完成的，可以重新加以理解，是可能性的存在，因此，對於人類理解現在與未來仍然具有意義。如果過去對於人類來說是完成的，就如同人身上現成的器官一樣，那麽人類實際上也就沒有了過去，過去直接就作爲現成物外在於現在而與現在無關，因此，人類衹存在於現在。同樣，當下與未來同樣是開放的，充滿着各種的可能性。人們可以通過過去來理解生活、策劃未來，也同樣可以通過理解未來去理解過去與現在。在這裏，過去、現在、未來是不可分割地存在着。在這個意義上，人因自由而存在於整體時間之中，其中包含了一切可能。這是機器無法企及的。機器永遠不可能存在於這種整體時間中，它不會有過去、未來。

正因爲人類存在於一種整體時間性之中，因此，人類總是不可遏制地打開着絕對未來，並因此纔會擁有宗教。宗教不僅確信當下的事物，還確信不在場的過去和未來的事物。人類的時間性

存在使其能夠打開一種終極的可能性，即死後的可能性。不管對死後的可能性世界如何理解，都不可能不確信，仍有一個世界在延續，仍有一個世界存在。但是，如果人們承認自己生活於其中的這個世界是一個有倫理原則而充滿價值色彩的倫理世界，是一個有公正、友愛同時也有不公與罪惡的世界，那麼人類就會理性地將死後的世界設想爲一個更公正的世界，並進而更加堅守現實世界的道德。也就是說，如果人們的世界擁有道德性的存在，那麼就會不可遏制地去想象、理解死後的彼岸世界。在這個意義上，宗教信仰的基礎便是人類的自由意志。因而，無論人工智能多麼發達，也不可能取代人類，不可能使宗教消失，或者消除人類的道德。人類之所以擁有道德世界，就因爲人類的行動並非必然的，而是自由的。

二

蔡恒進：黃裕生教授所說的“自由意志”，其實是一個很大的問題。討論這個問題應有的背景，是如今現代科學的進步與快速發展。很多人都覺得“物理還原”是一個非常有效的方法，但“物理還原”並不能探究出自由意志由何而來。從宇宙大爆炸開始，雖然可以用物理方程等類比宇宙的發展演變，但即使現在引進量子效應、概率性，都無法探究自由意志的本質。生命一定要有意識，能夠意識到我是我、外界是外界。即使是一個單細胞，當它能分清楚內外的時候，自由意志實際上就已經開始了。衹有從這個角度，纔能看清自由意志的來源。對於人類來說，皮膚非常敏感，大腦在母體外面發育的時間很長（0—5歲），因此人類有更強的“我”的意識，是萬物之靈。人的發展是一個連續譜，是高速度的進化，並且這種進化不是通過人類身體，而是通過將人類意識片斷的物化，訴諸文字、雕塑、音樂等形式可以脫離人類身體的方式進化。這些進化過程歸結起來就是人有自由意志。此外，人的意識還可以超越時空的、構築未來的各種可能性。

但是，當人類的技術發展到一定水平，發達的技術又開始變成了限制，這時的技術發展反而又變成另一種負擔。也就是說，有自由意志並不代表着人類沒有危險。危險的關鍵就在於，機器不需要比人類更聰明、更不需要比人類想得更深遠，即便不是主觀或故意要“造反”，也可能傷害甚至毀滅人類。這就如同把豹子之類的危險動物養在家裏，雖然它平時不傷人，但哪怕衹攻擊一次，後果也不堪設想。而機器的魔性，可能就存在這種情況，這就是問題的所在。

至於黃教授談到的此岸世界、彼岸世界，我認爲，關鍵在於，人類必須在“此岸”就解決問題，因爲彼岸世界並不能幫助真實的人類世界。例如，當人工智能的威脅到來的時候，衹能由人類自己來解決問題，而不能寄希望於彼岸世界。人類追求不朽、追求永生，但實際上人類已經能夠做到了。一個人五十歲的時候已經與三歲時不同，但卻依然認爲我還是我；當談到要永生的時候，究竟是是五十歲的我要永生，還是一百歲、兩百歲的我要永生呢？實際上，哪種情況都不是。在某種程度上，從三十歲到五十歲，“我”的有些東西已經能夠長久地存續在這個世界裏（比如，出版著作），那麼“我”（的一部分）已經可以看作是永生的了。孔子（前5512—前479）就是永生的，他的很多言行與主張（意志片斷、認知坎陷）已經融匯進漢語中，又被一代代人所學習。孔子的學說包括仁、義、禮等，都在被後人實踐着，而且一直在演化，所以他是在永生。普通人也可以通過這種或者其他方式永生，在此岸世界便已經可以實現永生。

張祥龍：對黃裕生教授所說的“自由意志”，我也有不同看法。按照西方近幾十年的認知科學研究，以前所認爲的那種自己的意識完全主宰自己的意願的強自由意志論的看法，可能是不對的。有幾個著名的實驗已經對此提出了反證。比如，其中一個實驗讓被試者用自己左手或者是右手的手指按鍵，人們覺得自己的自由意志當然能夠控制自己用哪隻手按鍵，但是經過先後幾次的實驗，卻都發現在被試者決定動哪根手指之前，他的腦圖顯示他已經有了動的衝動，而且根據被試者的腦圖測試者甚至能夠預測他將動左手還是右手，達到百分之八十的準確率。當然，有的科學家據此得出結論，認爲人類沒有自由意志，這是不準確的。這個實驗衹是說明，人類的自由

意志受他的內在時間之流造就的“潛意識”等因素影響。具體的動作過程，可能在“顯意識”之前就已經做了準備；但是，這個所謂的被測試出來的潛意識，與人們以前的身心經歷的積澱是有關係的，其本身不是完全機械的，而是有所選擇和綜合的。以前的心學或者印度的瑜伽最後所要達到的境界，就是把人們所謂的“潛意識”與“顯意識”充分溝通，使整體意識進入到一個更高的、更完美的狀態。這從側面說明，人們的自由意志可能不完全是由主觀意識或自我的顯意識來決定的，它是一個與人們身體有關的時間化的過程。

既然是這樣，難道人工智能就不能夠模仿它，或者在結構上去再現它？既然通過“深度學習”和其他人工智能的演算法，已經給了機器非常初步的時間化能力，那麽，這個發展再進行下去，說不定機器也會有所謂的過去，過去對它不再是一片空白。機器也會有記憶力，通過推動這個記憶力，機器也能夠對未來做出很多選擇，能面對各種各樣的可能。這就是所謂自主學習的一個特點。“自主學習”就是要面對各種的可能來構造出新的能力。從這方面看，未來機器是否會像人類一樣有這種自由意志，還是需要進一步探討的。

黃裕生：張祥龍教授所講的那個實驗，的確有一定的挑戰性；但是，這種實驗本身已經設定了一點，就是人類自由意志是可以歸結爲人類神經元的放電，因此這個設定本身可能就是一個問題。如果說可以做實驗來驗證、還原自由意志，可以歸結爲這種神經元放電的話，那麽就代表着人類其實並沒有真正的自由意志，因此做這個實驗的理念前提本身就已經否定了自由意志。這類似於某些人對康德（I. Kant， 1724—1804）的批評。比如，在鐵軌的兩個分叉上，一邊有三個人，另一邊有五個人，這時不管火車怎樣走，其實都要殺人；這種所謂的設定，已經把人從道德處境裏面抽象出來了而處於一種必然要殺人的情況中，也就不存在故意殺人的問題。還原實驗，有類似於這樣的前提在其中。

蔡恒進：針對黃教授的觀點，我想再補充兩點：第一，未來機器有可能像人一樣有意志力；即使不可能，也依然是危險的。因爲，人類把很多意識片斷、認知坎陷傳遞給了機器，機器就會模仿；而由於機器的計算能力很強，人類展示給它的內容有好有壞，機器可以任意組合卻又缺乏整全意識且無法辨別是非，這就會帶來危險。尤其是機器的速度快、力量大，不需要有很強的自由意志就可能對人類造成傷害。第二，機器也可能會有自由意志。如同張祥龍教授所言，由於自主學習技術的存在，機器可以學習很多內容，然後能將其按照某種方式組合起來。這一點與人的學習過程是類似的，並沒有本質差別。

“人”的意義在於人的整全意識，這一點是機器目前發展不出來的。由於人是億萬年進化過來的，認知坎陷的“開顯”完全是歷史積澱的結果，在適當時候便會開顯出來，而且是能與物質世界相呼應的、是向善的。這種開顯過程放在機器上就不能適用。

人類的“善”從某種程度上與自身生命的有限性相關。從古至今，人們一直在討論善由何而來，例如惻隱之心、羞惡之心。人類相信有善，在於一代代輪回染習的傳承，在於嬰孩時期就感受到的生而得之的善意。這個“善意”包含三個方面：一是大自然的饋贈，二是父母的養育教導，三是社會的保護。假如說，人類未來可以活到一萬歲，但從來沒有經歷過弱小的階段，某種程度上，這種染習的過程就沒有了。另一方面，對於機器來講，它沒有這種輪回染習的經歷，就沒有對善意的感知，不會有天生的善意，因此，要將善意傳遞給機器的挑戰性就很大。

三

主持人提問：機器的“魔性”是從哪裏來的？機器的“惡”是創造機器的人賦予的，還是由其自身產生的？

蔡恒進答：這裏可分爲兩個層面。第一個層面，並非機器主觀有惡念，而是存在一種“暗無限”（dark infinity）狀態。“暗無限”是一個中性詞，人類自身也會出現這一狀態。比如，某

小孩從一個街區到另一個街區去購物，如果在路上碰到兩個下棋的人，就有可能轉而去看下棋而忘記購物。假如他的智商很高，又深深地被下棋所吸引，那麽他可能就想要把它研究透，由此花費很多時間，不僅忘記最初的目的（購物），甚至廢寢忘食，這就是一種“暗無限”。這種研究一旦進行起來，就看不到盡頭，難以停止。因而，重點並不在於事情的好壞，而在於在這件事情上花掉的時間。又如，這個小孩碰到一隻他從來沒有見過的犬，他也可能停在那裏仔細研究它，出現“暗無限”的狀態。小孩是有可能忘記購物的，但普通人是不會的，會退出來，想起購物很重要。然而，機器則像缺少這種自控力的小孩，人類該怎樣讓它從“暗無限”的探索中返回呢？

另外一個層面，正因爲人類想讓機器對人“好”，但辯證來講，有“好”就有可能出現“壞”，有可能出現難以預料的場景，甚至與人類所設想的情況完全相反。一旦發生這種壞的情況，後果將非常嚴重。人類並不能真正控制這種情況，因爲人類並不能完全禁止有危害人類功能的機器人。即使人類在多數情況下是想去做善事，但不同角度、不同立場的人製造的機器人就可能完全不一樣，因此並沒有辦法去規定一定要禁止哪些、不禁止哪些。解決的辦法就是，允許人去探索自己想做的事情，但這目的與過程都要公開、透明。因爲，一個人想不到的危險，別人可能想到，這樣就會互相制衡。比如，歐洲現在的立法就要求系統的開發要有可解釋性，任何系統都必須進行登記。開發者應當有這樣的自覺，以公開、透明的模式來相互補充、制衡。這是未來一種可能的進化模式。

至於黃裕生教授說，人與機器之間存在絕對的分界；其實，這個分界是不存在的。因爲，目前已無法定義“人”是什麽。比如，脫氧核糖核酸（DNA）、智力等等有生物性的東西，機器都可以有；而且，人在未來很多身體的零部件都可以替換成機器，如心臟起搏器已經是比較成熟的技術。因此，人與機器的這個邊界將會變得非常模糊，尤其是如果人的腦子也可以與一台超級計算機交互連接的話，這時已很難再區分是一個人還是一台機器了，人與機器的這個邊界已經徹底沒有了。

張祥龍答：我覺得，這裏先要定義一下什麽是惡？簡單說來，“惡”就是一種狹隘的對象化，不顧及與其他人的聯繫。比如，自私自利是惡，而利他則是善，都是這一定義的反映。假如機器人要有惡性的話，首先與這個設計者、製造者有很大關係。如果設計者有很強的目標、欲望在指引，那麽他設計這個機器人的時候，就會通過各種各樣的手段讓機器人能夠具有達到這個目標的能力。例如，出於軍事上的目的而設計、優化的那種機器人等等。此外，“惡”“善”是與社會條件相關的。如果是一個異化的環境，由資本來決定機器人未來的發展，那麽，就會不擇手段地牟取暴利，並將這種衝動反映在設計者身上。

至於機器人本身，因爲它有自我學習能力，甚至是有某種自我意識，有可能產生惡；但如果讓它的這種意識越來越深化、自由地發展，那麽這個機器人可能向善的機會就比較多。“善”與“惡”實際上是一對哲理定義，由於“惡”是被貪婪的目的等控制，因此，自由的選擇和原發的不受對象控制的意志，就像佛教講的那種去掉了執著的根本意識，纔是未來要爭取在人生中以及機器人身上實現的。《生命3.0：人工智能時代的人類》一書的作者設想，未來有一個超級人工智能，它把人類社會中存在的所有問題都解決了，使未來社會中的人類互親互愛。但是，人們願意生活在那樣一個社會裏嗎？那真的是一個善的社會嗎？如果人類所有的決定、所有享受到的東西，都被這麽一個超級人工智能來構造、控制、主宰的話，人們肯定會覺得不舒服，會覺得自己在被豢養。我們寧可要一個不太完滿的世界，也不願意要一個完滿但完全被控制的世界。衹有在衆多的可能、真實的多樣性中，人類纔能尋找到真正的善。

黃裕生答：機器人有惡念與機器人做壞事，這是兩個問題。機器有惡念，這是道德意義上的判定；它會做壞事，那就不是一個道德上的判定。因爲，說它做了壞事，那是對應人類來說做的是壞事，但卻不足以歸爲惡念的問題。如果是惡念的問題，那就意味着機器必須有自由意志。

有了自由意志，纔會有一個好壞的問題。因爲，所謂善惡必須要有一個前提，即可以自行選擇；如果沒有選擇，就不存在善惡的問題。比如，人類不可能對動物有道德上的評判，因爲動物是自然的生活狀態，而人類則是自由的生活，永遠面對着可能性，有選擇纔會有善惡的問題。如果說機器有自由意志，那麽當然機器就有善惡。討論惡的問題需要一個前提，即這種惡是指道德上的惡。

聽衆提問一：有沒有可能，某一種族或者人群利用人工智能把所有人都控制起來，並且對現存的自由、民主、公平、正義等價值觀產生衝擊，使“善”衹針對少數人，而對多數人卻成爲一種“惡”？

張祥龍答：人工智能的發展將來完全可以分種族、分階層。比如，它可以被用在軍事上，甚至通過專門辨認某一種族的面相、體型特徵來被應用於種族屠殺上。這確實是隱含的比較可怕的地方。至於你說的少數聰明的人控制了其他人，也是有可能的。因爲，如果高科技由不善良的集團或者人來控制，那麽，他們當然是要急劇地發展技術來與對手競爭，那就會有一些高度保密的先進技術，然後以此攫取統治權力。那樣的話，人類整個社會的自由、民主、公平、正義等要素都會蕩然無存。在普通百姓毫不知曉技術發展的情況下，那些人會用技術手段構造出每個人都無法避免使用的佈滿生活的智能工具，比如手機等等。那種情況下，人的任何選擇，包括那些表面上是自由意志的選擇，實際上都在被操縱着。這完全是可能的，也恰恰是人工智能的危險性所在。

聽衆提問二：機器的“神性”與“魔性”是否有可能統一？其“魔性”確然對人類造成了威脅，但如果發展其“神性”並使之與“魔性”相統一，是否會因其與人類過於相似而形成另一種威脅？

蔡恒進答：關於“神性”與“魔性”之間如何平衡的問題，首先不要誤解“魔性”。“魔性”可以理解爲“神性”的反面。“神性”就是人們在此岸對無限的永遠不會終止的追求，但實際上的具體內容卻很難去定義。例如，人們講“聖人”，卻並不會真正去定義清楚何謂“聖人”，而且也無法定義清楚。而對於機器也是一樣，“神性”與“魔性”必然相互聯繫，相伴而生。決定人類社會未來的行爲並不是個人的意識，而是整體的意識，是超越人們身體的。我所想象的未來的圖景中，人工智能與人的智能會同時起作用。因此，在這種情況下，個體並非不重要，但作爲個體卻並不能掌控全域，而必須在更高的層次上作爲整體而前進。

聽衆提問三：如果說人類是真正時間性的存在而機器並非如此，那麽，人類的這種狀態是否可以看作衡量人與機器之間高低差別的標準，抑或可能僅僅是一種人類特有的看待世界的方式而已？

黃裕生答：是的。正是因爲人類有思想，所以纔高於所創造的機器人。人自身真正的進化，依靠的是思想的進化，依靠思想加深深度來不斷改變、刷新歷史。人類思想活動最爲典型的體現，就在於人的自由；也就是說，人之所以會思考未來、思考過去，都是基於人是自由的這一前提。如果人是封閉於過去或者當下世界中的話，就不存在或不會提出這樣的問題。由於人的所有思想活動都是基於其時間性的，所以，機器再怎麽進步，都無法達到這一點。人與機器的分界點體現在思想，而此思想是人類的思想，是在回應世界的召喚。

至於蔡教授剛纔提到人機無界限，我卻認爲，即使人身上安裝了再多的高速運轉的機器要素，也依然是以人作爲一個主體，而不是以機器作爲主體，這與直接製造出與人一樣的機器是不一樣的，必須區分開。

聽衆提問四：黃裕生教授在談到人工智能與人類的分界時，使用了一個非常經典的概念——自由意志。但是，從康德以來，自由意志實際上是實踐理性的假設；而從人工智能的角度來看，如果把人的自由選擇作爲一種算法，那麽人工智能也未必沒有這個所謂的自由意志。可否將情感——哲學上的直覺能力——這種機器人不可能真正擁有而衹能模仿人類的能力，視作人與機器

的差別？

黃裕生答：關於自由意志問題，在康德那裏，它是三大懸設之一，但這是必然的懸設，因爲不懸設它則不合理。康德是從理性事實出發來反證必定要有自由這個懸設的：人們生活的世界是有秩序的，而這些秩序就是建構在一系列基本的倫理原則或道德法則基礎之上，也即作爲人類行爲底綫的那些法則基礎之上。也就是說，那些法則在人們的生活世界裏是有效的，否則人們的整個現實社會就會瓦解掉。而這些道德法則均是以禁令或者勸令、應當或者不應當的方式表達的。這樣表達的倫理法則既然有效，則反過來證明人必定是自由的，否則它們就應該是無效的。就像我跟一隻動物說“你不許咬人”是沒有用的，但是跟其他人說“你不許殺人、不許偷盜”則是有用的。既然人類社會是建立在這些最基本的道德法則之上，由此，康德反證出人類必定有自由意志，否則將無法解釋。因此說，這是一種必然的懸設。

關於情感問題，這個裏面要分出層次。日常的喜、怒、哀、樂這種情感也許是可程序化的，機器在這個層面上是可以有情感反應的。但是，也有些情感明顯是不可被程序化的。例如，愛。人類可以無條件地愛別人，愛人如愛己，這樣的愛是不可能被程序化的。但問題在於，人類如何擁有了這樣的愛？其根源在於人是自由的，由此纔可能擁有愛人如愛己這般的愛；也就是說，這樣的情感依然基於人的自由意志，因而不可能被程序化。

聽衆提問五：以第五代“微軟小冰”爲例，人工智能已經可以通過“深度學習”模仿人的神經網絡來習得人類的部分創作思維，這是否就意味着本雅明（W. B. S. Benjamin，1892—1940）所謂“機械複製時代的靈韻消散”又返回了？人工智能對文學的創作有怎樣的影響？

張祥龍答：在已進行過的計算機與詩人的競賽實驗中，即使是專業的文學研究者、評論家都很難將計算機所創作的詩輕易辨別出來，計算機竟然一直撐到了最後一輪，讓人非常吃驚。這恰恰說明，人工智能現在已經達到能夠表達情感及藝術的境界，而這已經屬於非對象化的意義構造、意義感知，有極大的進一步發展的潛力。在未來，機器所創作的文學作品也可能會打動我們，成爲整個社會的時髦，甚至超過一些人類作家。對於這一類型的機器人，我們需要強化其自主學習能力，並且讓其學習真正偉大的藝術作品；衹有這樣，機器纔會進一步創作出優秀、動人的作品。而這種情況恰恰可以反過來刺激人類反省自己的藝術作品，反思何種作品纔算是真正的精品。

蔡恒進答：人工智能用於詩歌創作雖然可以騙倒很多人，但是卻騙不倒真正的詩人。無論機器做得怎樣好，也比不上真正的詩人，因爲真正的詩人能夠開顯出新的“認知坎陷”出來。比如，唐代崔顥（704—754）創作的《黃鶴樓》一詩，機器是創作不出來的。它沒有人類這種肉體、情感、進化的歷史，它現在衹是它所知的各種要素組合來、組合去。或者將來有一天，人工智能對自己的身體產生感知了，它會開顯出另外的“坎陷”來，但這種情況下的作品也衹有其他人工智能能夠欣賞，人類是欣賞不了的，就像蝙蝠開顯的坎陷不能被人類知曉一樣。我用“認知坎陷”這個概念來講所有的事情，就因爲它的穿透力很強，既非單純感性的也非單純理性的，卻能同時穿透二者。我相信，人的認知真的是這麽來的，並且能夠指導對人工智能的開發。

〔編者註：該文原爲綜述稿，由湖南大學嶽麓書院陳帥博士整理加工；發表前，編者重新設計了文章結構和體例，並對原稿做了修改加工，又請三位教授予以訂正。〕