人工智慧與哲學相關議題工作坊

謝世民*

哲學學門在 2017 年 12 月 25 日舉辦了「人工智慧與哲學相關議題工作坊」，針對人工智慧所涉及的哲學和倫理問題，邀請七位學者發表他們的研究心得。這次工作坊吸引了大約 120 位左右的教師、同學和民眾到場聆聽、提問和討論。整體而言，學者的演講相當生動，引發了非常熱烈的對話，讓困難的哲學和倫理問題逐一呈現，並提升了參與者的理解高度。他們在工作坊結束後將自己的發言稿進行了精煉，在此與《簡訊》讀者分享。

一、中央研究院歐美所洪子偉副研究員談〈美麗新世界？ AI 對人類文明的挑戰與機會〉

這個報告旨在檢視近幾年對於 AI 發展（尤其是 Super AI）的憂慮與爭論。所謂的 Super AI 泛指所有具有超越人類所有認知能力的人造自主適應系統。最近 Stephen Hawking、Bill Gates、Elon Musk 紛紛對 Super AI 表達疑慮。Musk 認為，對我們生存的最大威脅也許就是 AI。越來越多的科學家認為應該有些國際或國家層級的規範可以監管相關發展，以確保人類不會做出什麼蠢事來。但另方面，Tim Cook 與 Mark Zuckerberg 則對 AI 持較樂觀的看法。Zuckerberg 更認為散播對 AI 末日審判式的恐懼既不必要，也相當不負責任。

相對於上述爭論，牛津大學教授 Luciano Floridi 則提出第三種觀點。他認為 AI 發展雖非無害，但毀滅人類的憂慮是多餘的。因為 Super AI 雖在邏輯上可能，但現實上卻是完全不合理的。這是因為（1）硬體發展的物理瓶頸。例如 Moore's law（積體電路上的電晶體數量每 18 個月便會增加一倍）在 2013 年後已經失效。（2）許多 AI 宣稱可通過 Turing Test，但 Turing Test 只是必要而非充分條件，而且是眾多必要條件中門檻極低的。（3）Undecidable problems。計算理論中已證明，我們無法建構一個可以在所有 yes-no 問題中給予正確答案的演算

* 國立中正大學哲學系教授、前科技部人文及社會科學研究發展司哲學學門召集人

法。基於上述種種限制，Floridi 認為對於 Super AI 的憂慮是多餘的，這就譬如爬到大樹的樹梢，並非邁向登陸月球的一小步，而是旅途的終點。

僅管 Floridi 的結論或許正確，其論證卻不無疑問。我認為可從三方面加以反駁。首先，硬體發展的物理瓶頸，或許會限制單一處理器的計算效能。但是計算能力並非只能依靠單一處理器的超級電腦。例如分散式運算（Distributed computing）就可以將複雜的計算工作交給眾多的家用電腦或手機，甚至計算的硬體不必是矽晶體，而可以是細胞，或是兩者的綜合。例如，Ozasa et al. 研究單細胞生物的分散式運算，發現眼蟲在光合作用時要將光訊號最佳化的類比式處理比數位式的還要好。生物工程師利用單細胞生物原有計算能力、自主性、群聚設計出生物計算機性。其次，雖然所有的數位式電腦都會面臨 Undecidable problems，但資訊處理也可以是類比式或混合式的，例如人類製造出的第一部電腦是類比式的（二戰英國轟炸機的瞄準系統）。Piccinini 和 Bahar 指出人腦的神經訊號處理既非類比也非數位式。因為單一腦細胞上所能記錄到的時間序列的電子訊號（spike trains）是由離散脈衝所構成，但卻被當成連續信號來處理。第三，神經科學家 Sam Harris 指出，即便只能做出比人類還笨的人工智慧，由於電子迴路的運作速度比腦細胞生化反應快一百萬倍，讓它運行一星期，就可完成人類需要兩萬年才能完成的工作。以人類有限的心智能力能否理解，甚至控制這種 AI ？當然，Harris 可能過度簡化某些因素（如能源、容錯、計算策略），但卻指出 AI 在所需時間上的優勢。

簡言之，論戰各方均未否定 Super AI 對人類的潛在威脅，但對於 Super AI 在可預見的未來會不會出現，觀點迥異。悲觀派（Hawking, Musk）認為這天遲早到來，樂觀派（Zuckerberg, Cook）與折衷派（Floridi）都認為悲觀派論點無證據基礎，在可預見的未來可能性不大。但由於雙方現有論點多無法支持其主張，我們仍無法知道這天是否會來到。與其擔心機器是否會毀滅人類，在那之前，不如多關心人類會不會毀滅自己與地球（戰爭、生態浩劫）。

二、清華大學哲學研究所趙之振副教授談〈AI 與知識論〉

當代人工智能（Artificial Intelligence）的奠基者麥卡錫（John McCarthy）認為，有些人工智能（如下棋）並不需要哲學；然而，如果要達到人類層次的人工智能，則電腦的程式必須納入一些相關的哲學觀念，例如有關「什麼是知識」、「如何獲得知識」，乃至於「對自由意志的態度」等等的想法，都應該被考慮。在此意義之下，哲學似乎可以作為 AI 的支援；惟此中的一個困難是：到目前為

止，許多與 AI 相關的哲學觀念都無法很好地被定義，以致於很難為 AI 的工作者所使用。另一方面，我認為 AI 的具體成果也可引發哲學的反思，試以知識歸屬（knowledge attribution）為例。假設透過深度學習（大數據學習、轉移學習）的訓練，一個人工智能系統 S（比如健康檢測儀器）可以藉著對圖像（照片）的辨識而「判斷」說：病人患中耳炎（P）。又假設其「判斷」的成功率比醫生高；換言之，就「透過圖像作診斷」而言，S 比醫生來得更可靠。然而，我們不會將關於 P 的知識歸給 S，亦即我們不會說「S 知道 P」，因為 S 終究不過是一具「能將中耳炎與非中耳炎的圖像作可靠分類」的機器而已；S 本身並不會作判斷，而是醫生藉著 S 之輔助作判斷。於此，我們可以提出兩個問題，此中第一問題是跟知識論中證成（justification）之內在論與外在論之爭論有關。它是這樣的：醫生在使用 S 作診斷時，如果她要藉由 S 知道 P，則她是否需要知道（或證成地相信）S 是可靠的呢？內在論者對此之回答是肯定的；與此相反，外在論者卻會主張不需要，雖然 S 是外在於醫生的工具，但它是醫生所賴以作診斷的，所以也是她的信念形成系統（belief forming system）的一部分，只要這個包含 S 的信念形成系統是可靠的，且其有關 P 的信念又是真的，則醫生便知道 P。她不需要知道（或證成地相信）S 是可靠的。然而，這似乎與一般醫療的實作狀況相衝突，因為我們總是會要求一名負責任的醫生在使用 S 時，必須知道（或證成地相信）S 是可靠的；若不然，她便不應藉由 S 來診斷。雖然我並不認為這狀況可以完全駁斥外在論，但至少顯示：在某些牽涉到責任的情形，內在論——這種與知態責任（epistemic responsibility）有著比較密切關聯的主張，似乎比外在論更符合我們常識的直覺。這暗示著：內在論與外在論對證成概念之理解可能是不盡相同的。或許正是此緣故，一些哲學家傾向捨棄「證成」一詞，而改用其他諸如「資格」（entitlement）之類的語詞來談論信念之合理性。

第二個問題與 AI 有更直接的關係。傳統認為真信念是知識的必要條件，而我們不會將有關 P 的知識歸屬給上述的健康檢測儀器，因為我們不會認為這機器是具有信念的。現在的問題是：人工智能系統 S 要具備怎樣的條件，我們才能將信念（或思想）歸給它呢？一個回答是：如果 S 能通過圖靈測試（Turing Test），則 S 便擁有信念／思想。但是對於「通過圖靈測試」是不是「擁有思想」的充分條件，卻不無異議。根據圖靈測試的構想，受測試的對象接受提問者的問題並予以回答，但受測對象與世界其他的事物是沒有互動的，從而提問者也就無法觀察到這樣的互動關係。戴維森（Davidson）曾經指出：在這樣的場景之中，提問者其實是沒有獲得足夠的資訊，可使之決定受測對象的言說之意義與思想的內容，她甚至無法決定受測對象是否有思想。即使對象說出一串在語

音、語法上跟我們的自然語言一樣的語句，我們也無法決定這些語句是否具有語意或者其語意與我們語言的語意是否相同。

然而，晚近 AI 的蓬勃發展，讓我們不難設想一個這樣的人工智能系統或機器：它與我們共處於同一世界之中，我們可以有足夠時間觀察它與世界以及它與我們之互動（至少包括因果的互動關係）；簡言之，它可以滿足戴氏所要求的那種關於受測對象、我們以及世界之間的三角測量（triangulation）關係。如果真的存在這樣的對象，戴氏會主張它是具有思想的。有趣的是，他還進一步認為：透過三角測量（包括一些有關詮釋的原則），我們不僅可以把思想歸給對象，同時也可以把知識歸給對象，從而可以對懷疑論有所回應。

戴維森能否以上述的方式成功地回應懷疑論，暫且不論，在此我只想說三點：（1）戴氏關心的是思想而不是意識（主觀經驗）之歸屬。一個人工智能系統具有思想是一回事，它具有意識或主觀經驗（例如痛、憤怒、自卑的經驗）則是另一回事。有關此兩者的歸屬條件，不宜混為一談。在考量「S 是否具有道德地位」的問題時，除了思想之外，意識歸屬之問題是不可忽視的。（2）思想歸屬的工作是需要一個觀察與詮釋的歷程；這歷程需要多少時間，要看具體情形而定。我們能將思想歸給 S，並不是或不僅是看它本身具有哪些固有的性質，而主要是看它與我們以及世界之互動關係。如此一來，這便意味著：如果一個人工智能系統具有思想，則它一定兼具表徵與溝通（理解）的能力。（3）若依以上所述，則無論是信念、思想、或是知識，它們原則上都不可能是孤立的，它們一定是嵌入在一個社會性的網絡之中，保持與網絡中他者（世界）互動之可能，同時也受到社會的／客觀的規範所制約。因此，如果我們要以人造的機器來實現人類層次的智能，則對此中相關的社會性的、規範性要素之考量，便是不可避免的。

三、臺灣大學哲學系梁益堉教授談〈人工智慧可不可能有自我意識〉

人工智慧系統可不可能有自我意識？例如：AlphaGo 花了幾個月時間學習 3,000 萬場棋局，擊敗人類。AlphaGo Zero 透過與自己對弈，在三天的時間和自己下了 490 萬場棋，最後以 100 比 0 擊敗 AlphaGo。這是否表示 AlphaGo 和 AlphaGo Zero 具有自我意識？又例如 Watson「閱讀」過 2 億頁自然語言文件，能夠使用自然語言來回答問題。2011 年 Watson 參加搶答節目打敗人類，贏得了獎金 100 萬美元。這是否表示 Watson 具有自我意識？有的人認為答案是肯定的。這種主張背後的想法是：只要一個系統具備強大計算能力並展現出複雜行

為，就算是具有自我意識。問題是：這個想法是恰當的嗎？有反對者認為，強大計算能力與自我意識之間的關聯還有待釐清，不具備自我意識的系統也能做出複雜動作。也有的反對者認為，自我意識是一種意識到自我的主觀經驗，除非人工智慧系統也能擁有意識經驗，否則就不可能擁有自我意識。從哲學的角度看，將來的爭議點可能會在以下的議題：能經歷主觀經驗是不是具備自我意識的必要條件？能經歷主觀經驗是不是具備自我意識的充分條件？由此看來，我們首先需要一個關於自我意識的理論，來幫助我們思考這些議題。

舉例而言，Thomas Metzinger 將最初步、門檻最低的自我意識稱之為「極簡的現象式自我」（minimal phenomenal selfhood），並將其定義為：「作為一個自我」的意識經驗（the conscious experience of being a self）。這種極簡的自我意識經驗是由三種特性所構成：（1）「自我認同」（self-identification）；（2）「自我位置」（self-location）；與（3）「第一人稱觀點」（first-person perspective）。這裡的「自我認同」不是指個人在不同文化脈絡中之社會學意義的認同感，而是指個人將自我「全面式的等同於一整個身體」。換句話說，就是將某一整個身體等同於或感受為自己的。因此，此處的「自我認同」可以理解為關於一整個身體的「身體歸屬感」（the sense of full-body ownership）。「自我位置」指的是從第一人稱的觀點「我覺得自己在哪裡」的空間感。比如：我正坐在棒球場的外野觀眾席上看現場比賽，當我起身換到某個內野座位，我的「自我位置感」便改變了：不僅我覺得自己的身體處在不同的位置，同時我也覺得自己的「第一人稱的觀點」處在不同的地方，以致於坐在內野座位所看到的場景視野與外野位置不同。我們若以 Metzinger 的理論作為討論的起點，接下來的問題便是：AI 系統能不能發展出這三種特性？關於這個問題，有以下三種立場值得我們思考。

第一種立場認為，AI 系統有可能在將來發展出這三種特性，也就是具有「身體歸屬感」、「自我位置感」及「第一人稱觀點」。若是如此，我們就有理由來說 AI 系統能夠擁有自我意識。第二種立場則認為，無論 AI 系統多複雜都不可能發展出上述這三種特性。這種立場認為，能經歷主觀經驗不僅是具備自我意識的充分條件，也是具備自我意識的必要條件。AI 系統可以處理、整合並運用關於系統自身的資訊，但是這並不表示就能將某一肢體或整個身體感受成自己的。同樣的，AI 系統可以處理、整合並運用關於系統本身位置的資訊，但這也並不算是有「我覺得自己在哪裡」的主觀感受。若是如此，人工智慧系統便不可能有自我意識。還有第三種立場值得我們考慮，那就是：我們其實並不知道 AI 系統能不能具有上述三種特性。意思是說，我們需要問：假如將來某個 AI 系統真的具有屬於自己的「身體歸屬感」、「自我位置感」及「第一人稱觀點」的話，

我們有辦法知道該系統的這些主觀感受是什麼嗎？如果我們不曉得如何回答這問題，那麼我們還能不能斬釘截鐵的說人工智慧系統可能或是不可能擁有自我意識？顯然這些議題還需要更多的研究，並且哲學與經驗科學可以互相合作，對這些重要議題作出貢獻。

四、陽明大學心智哲學研究所林映彤助理教授談〈人工意識：天方夜譚或可成真？〉

對某些人如 John Haugeland 而言，人工智慧主要的目的之一在於設計出具有完整心靈的機器，如人類心靈一般，具有思考和感受能力。這類系統時常出現在小說和電影情節中，從《變人》、《A.I. 人工智慧》到《人工意識》，透過巧妙的故事安排，挑戰我們對於「人」（person）的概念和直覺。除此之外，人工意識系統的出現是否可能，也牽涉到諸多未來社會議題，包括超級智能（superintelligence）的極限、將心靈上傳（mind uploading）的可能性、以及機器人的倫理地位等問題。Thomas Metzinger 便警告，我們不應輕易嘗試創造人工意識，因為可能在未知的狀況下，創造了具有受苦能力或正在受苦的系統。

人工意識研究之困境，與探討動物意識的困難有相當類似之處。根據 Allen 和 Trestma 的研究，一個完整的意識理論必須要能夠回答分布問題（distribution question）和現象問題（phenomenological question）。前者探討哪些生物或非生物（包括人工）系統可能有意識？而若是意識存在於人類以外的系統中，後者探討牠／它們的經驗內容為何，包括經驗的質感：作為牠／它們的感覺是什麼？第二個問題比第一個問題更難回答，也是意識問題的核心。由於意識的主觀性，即便其他系統有意識，無論是人工系統、蝙蝠、或是他人，若無法占有此主體的觀點，亦無法完整了解作為他／牠／它者之經驗為何。

若是我們了解意識經驗如何產生或其功能為何，便可藉由系統結構或行為的相似性，判斷生物或非生物系統是否具有意識。針對意識的功能，可從演化的角度探討意識現象何時產生，它的出現是否提供主體一個更好的方法去應對其環境？意識是否帶來任何演化上的優勢？撇除那些認為意識研究是哲學家專屬遊樂場的人工智慧研究者，部分觀點將意識視為副產品，無法產生任何因果作用；另一類的研究者，則針對各種可能的意識功能發展人工意識。不同的研究途徑包括系統的自主性（autonomy）、感受運動（sensorimotor）、自我動機（self-motivation）、訊息整合（integrated information）、自我模型（self-model）等。

這些研究途徑各自對應特定的意識理論，各理論架構皆可針對人工意識之

可能性、分布問題、甚至現象問題，提供回應與評定標準。然而意識理論之間差異大，在各個意識研究議題上鮮有共識，例如意識之神經關聯是否為局部腦區或牽涉大腦整體，抑或包含身體和環境，感知經驗內容為豐富還是貧困，意識和注意力之間的關係為何等。如此差異部分來自於對行為報告以及推論意識出現與否的態度不同。尤其針對具有爭議的狀態，在某些心理實驗操弄下，各種對意識出現與否的行為測量結果不一致。在這樣的狀況下，較為保守的觀點傾向主張需有存在之證據才能主張意識之存在，自由派的觀點則認為只要找不到不存在之證據，即可推斷意識之參與。

究竟機器是否可能有意識？各意識理論之間的爭論正如火如荼地進行，目前尚未取得共識。Jaegwon Kim 曾出此評論：「我們無法透過理論推理，而設計出一種預測會有意識的全新結構；我不認為我們知道如何開始，或是如何去衡量其成果。」此外，Christof Koch 和 David Chalmers 等學者針對此議題，在 2001 年舉辦了一場工作坊。工作坊中唯一（接近）普遍的共識為：原則上，電腦或機器人有一天可能有意識；換句話說，我們在這個宇宙中找不到任何根本定律或原則，能排除主觀感受存在於人工系統之可能性。將近二十年後的今天，我們對於意識的理解和研究有突飛的進步，然而在各種意識理論百家爭鳴的狀態，人工意識的問題尚未存在一個公認的答案。

五、政治大學哲學系王華副教授談〈人工智慧為當代道德生活帶來的危機與轉機〉

許多人對人工智慧發展戒慎甚至恐懼，常來自於對「超級人工智慧」的擔憂。另一個大眾對人工智慧發展的擔憂是勞動力與工作被取代的問題，以及演算法應由誰來設計、誰來控制，而大數據與演算法這些資源，又應該掌握在誰的手裡？第三個人工智慧所帶來的可能危機，在於其所造成之人我互動形式的改變。而社群網路活動對自我的反向形塑不只影響個人思想與行動，相信也將影響自我認同，這便是第四個人工智慧所帶來的可能危機。

雖然人工智慧對道德生活可能帶來一些危機，它畢竟只是人類發明的工具，重點是我們要能找到適當的方式「役物」而不「役於物」。事實上，我們也注意到許多人工智慧帶來的助益。首先，人工智慧與大數據、物聯網等的搭配，可以成為促進交流與理解的工具。另外，在使用工具解決問題、協助人我互動的情況下（如照護機器人），我們也應注意在工具的設計中表現並推動我們認可的既有價值（例如：尊重自主、安全、賦能、獨立、隱私、社會聯結等）。

除了可以協助人我交流與社群形成，人工智慧與大數據的配合當然也是促進知識發展的重要工具。目前有許多領域開始採取類神經網路模型（artificial neural network）搭配大數據來進行相關性研究（如語言學研究語詞使用、心理學研究行為相關性等等）。哲學家（尤其是實驗哲學家）也可以應用這類工具發展理論。

另外一個可能的發展，是訓練類神經網路模型讓機器深度學習道德辨識，使其表徵道德知識，並具有道德感知／辨識、判斷、行為等技術。我們如何訓練機器學習作出符合道德的判斷？一個想法也許是直接將道德原則寫成程式植入機器軟體，但是考量到道德判斷牽涉到非常複雜的情境，「道德知覺」本身應扮演核心的角色。而對道德知覺的訓練則牽涉到情感、情境敏銳度在道德概念與道德判斷學習中扮演的角色。因此，我們應該考慮訓練機器的情感／情境敏感度以作為影響機器判斷與決策的重要元素。

最後，本人引克拉克的心靈擴展說為此次演講作結。克拉克認為，我們的心靈與其認知機制並不受限於我們的身體或腦袋範圍內，而延伸到我們身處的世界之中。為了減輕認知負擔並更有效率，我們的腦學會將某些計算工作的面向卸載給適當的外在表徵媒介（如手機），如此一來這些工作能被更迅速、可靠地執行。也因此，任何關於道德認知的理論不能只看腦內機制。道德認知牽涉複雜而與時俱進的、生物腦與社會世界（social world）兩者間的互動。更明確來說，是生物腦中非言說性的認知機制與架構這個社會世界的具有高度言說性、超越個人的「鷹架」（scaffolding）間的互動。什麼能成為我們心靈依賴的「鷹架」？本人認為，道德規則、社會體制與律法等，都是我們心靈依賴的「鷹架」。人工智慧也可以成為幫助我們建造「鷹架」的工具，甚至成為「鷹架」本身。

六、東吳大學哲學系蔡政宏教授談〈AI 機器人的道德推理〉

這個報告主要在探討 AI 機器人在具有道德——或稱人工道德（artificial morality）——這件事上可能會遇到什麼哲學難題；特別是如果道德涉及道德推理，而道德推理又涉及道德原則（moral principles）、推理能力、道德習得（moral acquisition）等面向，那麼人工道德在這些面向上會遇到什麼哲學難題。以下分成兩部分論述，各自又有三項子題。

（一）關於「AI 機器人」的初步問題

在探討 AI 機器人的道德推理前，我們最好先了解 AI 機器人為何以及相關

的初步問題。在「子題一：AI 機器人是什麼？」中，我嘗試由內涵與外延兩角度來界定「AI 機器人」，然而發現單僅由內涵（例如智能性、移動性、自主性）或外延來界定 AI 機器人都有所缺失。雖然對 AI 機器人無法有明確界定，但我們仍可談論其能力或能耐。在「子題二：AI 機器人能做什麼？」中，我藉由人們對 AI 機器人的分類來回答子題二並凸顯其能力廣度可涵蓋人類生活中的大部分重要面向。在「子題三：AI 機器人應做什麼？」中，我的目的僅在凸顯子題中問題的合法性。若 AI 機器人能施展行為，而其行為從人類角度來看又有對錯之別（例如，一輛行駛在路上的自駕車，車內沒有任何乘客，碰巧煞車故障，此時自駕車選擇撞向路人而非撞路樹；或是照護機器人餵老人吃安眠藥使其不吵鬧），那麼規範性議題就會因而產生。與 AI 機器人相關的規範至少可分成三類。第一類是針對 AI 機器人之製造者所制定的規範，例如 Asilomar AI Principles。第二類是針對 AI 機器人本身所制定的規範，例如 Isaac Asimov 的機器人三法則（Three Laws of Robotics）。第三類是針對 AI 機器人之使用者所制定的規範（設想：人類可不可以虐待機器人？）。以下將關注第二類規範。

（二）AI 機器人的「道德推理」

關於人工道德的建立，我們可採取「由上而下進路」（top-down approaches）或「由下而上進路」（bottom-up approaches）。在「子題一：理解難題」中，我探討由上而下進路所可能遇到的問題。舉例來說，若工程師要將機器人法則一（即「機器人不得傷害人類，或因不作為使人類受到傷害」）轉換成算程，其中的「傷害」應當如何理解？當 AI 機器人看到人類甲追打人類乙，它如何判斷乙（其實是搶匪）是否正被甲（其實是警察）「傷害」？總地而言，由上而下進路會遇到兩類問題：（1）該選擇「哪套」道德原則？（2）選定某套道德原則後，如何將其「成功」轉換成算程？（例如道德原則之間產生衝突該如何處理？原則之間應排優先次序嗎？原則可允許例外嗎？）

在「子題二：價值問題」中，我探討機器人與人類的推理能力有何異同。首先，我透過哲學中對於智能（＝技能＝ Know-How）的看法，特別是反智識主義（anti-intellectualism）的立場，指出機器人與人類都可具備智能者（intelligent agent）的推理能力。其次，我透過哲學中對於實踐智慧的看法，指出人類可具備——但 AI 機器人（就目前來看）無法具備——智慧者（wise agent）的推理能力。智能者具多軌傾向（multi-track disposition），採工具——目的之實踐推理模式，只能思慮方法，無法思慮（最終）目的；智慧者具有助產目的（maieutic ends），使其除了能思慮方法，還能思慮目的及其價值（議題：人類有需要賦予

AI 機器人二階或助產目的嗎？）

在「子題三：習得問題」中，我探討「由下而上進路」所可能遇到的問題或前景。此進路不給 AI 機器人道德原則，而是令其在大量的道德情境自行學習、建構出道德原則。問題：在深度學習之後的 AI 機器人，它能否針對它的道德判斷，明文表達（articulate）其中道德理由？若不行（無論是本體論上沒有理由，或是技術上無法給出理由），人類能夠或敢於信任其道德判斷嗎？但若行，或許我們人類反而可以向 AI 學習道德（如果它總是能給出強而有力的道德理由）。

七、中正大學哲學系何宗興助理教授談〈我們需要怎樣的道德機器人？〉

如何教會機器人正確的道德判斷，並且確定機器人會遵循道德法則，是智慧機器人發展不可避免之問題。有人可能會懷疑能作出正確道德判斷的機器人（簡稱「道德機器人」）是不可能的。不過，暫且假定道德機器人是可能的；機器人能夠學會正確的道德判斷，並且依照道德判斷行事。然而這是否代表，只要機器人不會作出道德錯誤的事情，它的行為在道德上就沒有問題呢？我認為不是的，我主張當涉及人類對其生活的自主權時，機器人需要受到比人類更嚴格的限制。考慮這個例子：

大衛因為失戀決定自殺，他打算投河自盡。正當他要跳下去時，經過了一個清潔機器人在掃馬路，它發現大衛爬上欄杆要跳下橋。請問，這個機器人該怎麼做呢？我認為，機器人也許可以勸導大衛不要自殺，或是報警請警方處理，但它不可以強制大衛不能跳河。試想，當我們生活在機器人無所不在的社會，想要自殺還得躲到人煙罕至的地方，這樣的社會應該不會是我們想要的。

這個主張看似合理。但問題在於，阻止人自殺通常是道德許可的。假若路過的是人類（叫她「阿潔」）。假若阿潔抱住大衛不讓他自殺，我們不會因為她違背大衛的意願而責備她，因為救人一命是道德上許可的。

這樣，我的主張就碰到以下難題：「既然阻止人自殺是許可的，那為何機器人不能去做？」要回答這個問題，不能只是回答「由於這侵犯大衛的自主權」，因為阻止大衛自殺便是侵害大衛的自主權。當然，自主權不是絕對的，阿潔強制不讓大衛自殺是道德上許可的。換句話說，我們需要解釋，為何機器人對人類的自主權的介入，不應該如人類一樣深。

我認為，英國哲學家 P.F. Strawson 的道德責任理論可以說明為何機器人不能像人類一樣介入人類的生活。Strawson 指出，我們對於他人的作為，自然而

然會有一些情感反應，例如，憎恨、受傷、憤慨、感激、原諒、欣賞。Strawson 把這類型的情感稱為「參與者反應態度」（participant reactive attitudes），因為這類型的態度主要是當我們處於人際關係之中，對於他人的善意、惡意或漠不關心所自然而然產生的反應。

了解 Strawson 的理論之後，讓我們來看大衛的例子。假如阿潔看到大衛作勢要跳河，她卻絲毫沒有反應，或甚至還嘲笑大衛不敢跳。雖然大衛決意要死，他仍可能對於阿潔的冷漠感到失望或怨恨，加深了他的死意。再者，作為第三者的我們，也會責備阿潔為何如此冷漠。由此可見，我們的參與者反應態度表現出，當看見他人遭遇不幸時，我們應當表現出適度的關懷，即便他人要求我們不要介入。從 Strawson 的理論，可以衍生出以下的洞見：人與人之間可以適度地干涉對方的自主權，是因為這是我們建立起人際關係不可避免的過程。

到此，我們就可以理解為什麼機器人不能如阿潔一樣去阻止大衛自殺。因為機器人沒有所謂的善意或惡意，它只是被設計成如此行為，無法為它的行為負責，我們跟機器人之間不會建立起任何真正的人際關係。

我的主張不是說機器人絕對不能夠介入人的生活，而是它對人類生活的介入，跟人類相比，需要受到更大的限制。這代表，對於道德機器人的研究，不能局限在如何教會機器人作出正確的道德判斷，我們還得研究機器人跟人類互動的範圍與界限。而其中有一個指導原則應當是：道德機器人的應用，不應該傷害人類對自身生活的自主權及其道德責任。