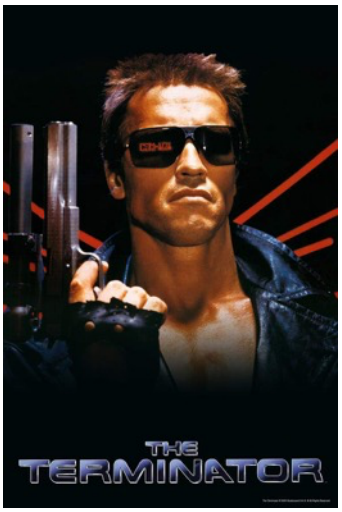
- **余創豪** chonghoyu@gmail.com

科幻小說變成現實？

現在人工智能的發展一日千里，每隔不久，科技巨頭便會宣布具有嶄新或者更強大功能的系統已經研製成功。由於人工智能具有難以預測的潛力，故此從前許多科幻小說的題材，現在已經變成了研究人員需要嚴肅對待的課題。無數科幻小說電影都包含了以下的情節：在未來，具有人工智能的電腦或者機械人發展出自我意識，具備了人類思想的特性，這些故事既有反派的人工智能，亦有正面的，邪惡人工智慧的例子俯拾即是，例如《2001太空漫遊》 裏面的HAL9000、《未來戰士》裏面的天網（Skynet）、《異星戰境》（Atlas）中的哈林（Harlin）。但擁有自我意識而輔助人類的人工智能機械人也為數不少，例如《新一代星空奇遇記》中的Data、《星際爭霸戰：航海家號》中的「醫生」。



針對前者，研究人員正在討論怎樣防止邪魔人工智慧傷害人類；因應後者，學者研究 應否給予對人類友善的人工智能一種「道德地位」（Moral status），所謂道德地位，並不是指人工智能有沒有倫理原則或者品性是否善良，而是指他們有沒有類似人類的尊嚴和價值，從而

賺取到道德的待遇，換言之，我們應否好像對待人一般去對待人工智能呢？限於篇幅，這篇短文只能夠集中討論後者。

《新一代星空奇遇記》 先知先覺

1980年年代末至1990年代的《新一代星空奇遇記》已經探討過這問題，在其中一集中，皮卡德艦長（Captain Picard）派遣Data執行一項危險任務，艦長問機械人是否明白為什麼這擔子會落在他的肩頭上，Data回答說：「因為我是可以消耗的（Because I am expendable）。」艦長馬上更正他：「不！我不是這個意思，我派你去，是因為在這特殊情況下，你比人類更能作出明確的判斷。」

在另一集裏面，星際聯盟指揮部認為Data是機械人，而不是正式的軍官，總部差派賴克指揮官（Commander Riker）去證明這一點，賴克只需要一個動作去陳述他的論據，他關閉 Data 的開關掣，Data 便停止運作，從而他指出 Data 無非是一部從屬於人的機器。

三十年後，科幻電影的情節已經變成了值得正視的研究對象，可能有些讀者覺得出現自我意識的人工智能還是言之過早，但無論如何，你可以將這篇文章當作趣味小品去閱讀。

圖靈分診測試

不同學科的研究人員都參加了這場辯論，舉例說，澳洲墨爾本莫納什大學（Monash University）哲學教授羅伯特・斯帕羅（Robert Sparrow）曾經提出了「圖靈分診測試」（Turing Triage Test，簡稱TTT），去探討機械人及人工智慧的倫理問題。

這個測試方法綜合了兩個概念，首先讓我簡述一下什麼是圖靈測試，圖靈測試由英國數學家艾倫・圖靈（Alan Turing）於1950年提出来的，用處是評估一部機器是否顯示出智慧的行為，這個方法十分簡單：測試者透過文字與一台機器和一個人隔着屏障去交談，如果測試者不能分辨出哪個是機器，哪個是人，那麼這部機器就被認為通過了圖靈測試。這個測試的重點是：回應是否看上去好像是有智慧行為，那部機器是否真的明白內容是無關重要的。分診是在資源有限的情況下，選擇怎樣分配資源的流程。例如在新冠肺炎大流行期間，呼吸機嚴重短缺，醫生必須決定誰有優先權使用呼吸機，決定因素可能包括了誰的病情更加嚴重、誰更加有希望康復……等等。

將上述兩者合起來之後，圖靈分診測試變成了一項探索機器人道德地位的思想實驗，TTT的場景涉及一些災難，測試者必須在拯救人類和拯救機器人之間做出分診的決定。如果在危難關頭測試者認為兩者的道德地位沒有分別，兩者同樣地值得拯救， 甚至乎人類認為應該挽救機械人，那麼機械人便通過了TTT；相反，如果測試者認為人類優先，這表明在內心深處，我們不相信人工智慧或機器人具有與人類同等的道德地位。

思想實驗：挽救人工智能抑或病人？

其中一個思想實驗如下：某間醫院安裝了一個人工智能診症系統，突然，電源和備用電源都失效了，醫護人員有兩個選項：第一，切斷人工智慧的電源， 將電源完全供應病人的生命支援系統，第二，讓人工智慧繼續運行，任由病人死亡。斯帕羅認為：只有醫護人員選擇挽救人工智慧系統而放棄病人的時候，那麼人工智能才算是擁有道德地位。

坦白說，這是一個漏洞百出的思想實驗，可能斯帕羅將人類的生命特徵投射到機器上面。如果人類失去了支援生命的能源，例如氧氣、食物、水…… 等等，便會永遠死亡；然而，機械卻不是那麼一回事，縱使關閉了人工智慧的電源，或者電池完全枯乾，甚至損壞，只要重新充電或者更換電池，系統便會繼續操作。 由此而觀之，若果醫護人員為了挽救病人而切斷人工智能系統的電源，這並不證明了他們認為後者沒有道德地位，因為後者並不會因此而死亡。

退一步說，如果人工智能系統真的會在失去電源之下永久失效，而醫護人員選擇犧牲眼前的病人去保存這系統，但這也不能證明他們相信人工智慧具有道德地位，這決定可能純粹是出於功利主義，如果醫院永遠失去了這個系統，那麼跟着可能會有幾百、幾千名病人因為失救而死亡，在經過計算之後，他們不情願地做出這個冷酷的決定。

思想實驗：在火場中的兩難

在探討TTT的時候，日本哲學家清水大風提出了一個類似的思想實驗：在一場大火災中，未來（Miku）進入火場，救出了一個人，房子裡還剩下兩個個體：一個是與未來一起生活多年的伴侶機器人，另一個是一名陌生人。火勢非常大，而且由於房間的佈局，未來只能在兩者中選擇拯救一個。未來

知道機器人本質上只是機器，因此不會受苦，因此她決定拯救那名陌生人。但此後她感到十分懊悔，放棄了機械人的決定成為她一生的心理陰影，這種感覺是斯帕羅沒有預料到的。清水大風的論點是：無論人工智能機械人有沒有道德地位，人類會和機械人產生感情與關係，因此我們仍然需要尊重機械人。

這思想實驗亦有很多商榷的地方，首先，未來的罪疚感是想像出來的，並沒有數據去支持人類在火災中放棄機械人之後會有強烈的情緒反應。的確，有些人為壞掉的機械狗舉行喪禮，但機械狗自然損壞和人類主動讓機械人消滅是兩碼子的事。

還有，清水大風跟斯帕羅一樣，將自己的生命特質投影到機械人身上。人類的心靈或者自我意識必須寄在一副軀殼中，當人的身體被濃煙或者火焰毀滅之後，這個人便永久死亡，但人工智能機械人則不然，現在你可以輕而易舉地將硬碟中所有資料複製在雲端，儘管一個機械人的軀殼完全焚毀，但不旋踵那部機械人的「思想」就可以在另一部機器中復活。這是現在已經可以做到的科技，筆者曾經在西班牙丟掉了手機，回到美國之後，T-Mobile將所有備份的軟體和資料下載到新手機上，在內容上新手機和先前的一模一樣！換句話說，人工智能機械人並不是獨特的個體，而是一個具有「永遠生命」的群體。

人工智能機械人並不是獨特的個體

假設在未來世界人類可以與機械人建立伴侶關係，而你喜歡周遊列國，你並不需要攜眷旅行，每到一個地方，你可以向當地的機械人公司租借一個機械人軀體，然後自雲端將

資料下載到這機械人的硬碟。即使你不喜歡旅行，而是長期住在同一個地方，但每當機械人的零部件發生嚴重故障，或者你不喜歡這部機械人的外型，你可以更換一台新的，然後重新下載資料。香港俗語有云：「夫婦如衣服」、「妻死妻還在」。前一句話是嘲諷那些不停離婚、再婚的花心蘿蔔，後一句是諷刺那些不斷續弦的男性。但這兩句話亦可以用來形容人類與伴侶機械人的關係。

一個人愛另一個人，往往是因為對方具有獨特性，在英文中有一句話，正是表達這個意思：「You are one in a million。」這類似中文的「萬中無一」。此外，唐代詩人元稹在《離思》 中寫出以下的佳句：「曾經滄海難為水，除卻巫山不是雲。取次花叢懶回顧，半緣修道半緣君。」這裡的「滄海」和「

巫山」象徵著愛人的至高性、唯一性、獨特性， 強調對方是無可取代的。「取次花叢懶回顧」是指詩人本為狂蜂浪蝶，一生進出花叢，但因為有了心上人，他已經懶得回顧芬芳吐艷的花叢。「半緣修道半緣君」的意思是：愛對方的一半原因是作者修心養性，另外一半緣故是她始終在詩人的心中。從這個角度來看，似乎人類很難會對並非獨一無二的人工智能機械人產生堅貞的感情、培養出跟個體人類一樣的關係。

關於人工智能的道德地位，現在仍然是眾說紛紜，所以筆者對此保持開放態度，你的立場又如何呢？在下一篇文章中，我將會討論人工智能會否發展出邪惡的思想，從而威脅人類文明。

20224年10月6日

原載於澳洲《同路人》雜誌

更多資訊