標題： 機器行為學

關鍵字： 機器行為、演算法、人機互動

機器受 AI 技術之福，逐漸掌控人類社交、文化、經濟與政治多方面的互動，倘若我們想盡量提升科技對社會的效益、將風險和傷害降到最低，那就必須理解 AI 的每段行為。Nobel Laureate Herbert Simon 曾在《Science of Artificial》中寫道：「自然科學是關於自然生物與現象的一門學問，那是否也存在所謂『人工』科學，專門研究人造物品與現象呢？」這類的科學主要針對一系列擁有特定行為模式和生態的智能機器，除了需要電腦科學與機器人學的專業之外，還需生物相關的知識協作，就像人類會有觸發行為的背景，AI 的演算法亦是。因此科學家將以動物行為研究為模板，討論機器行為與環境的交互影響與牽制。

研究的動力與困境

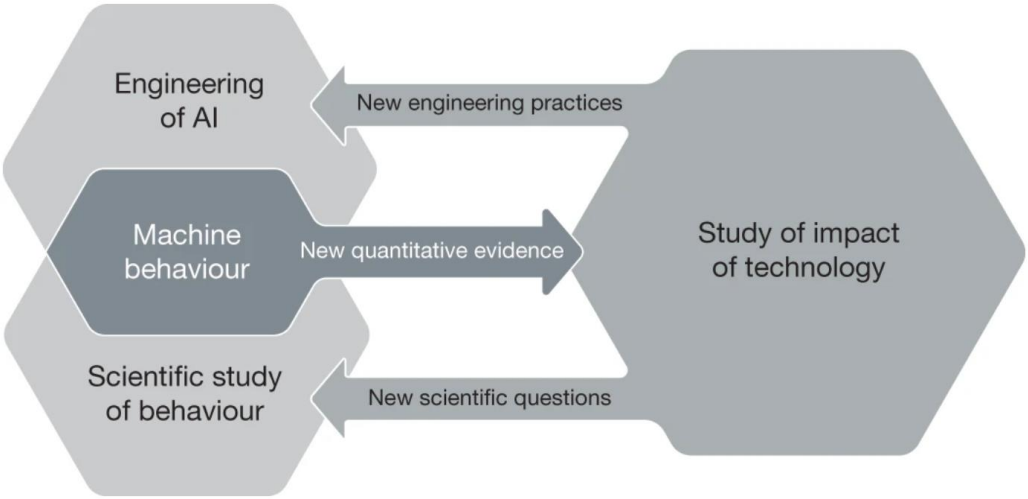
之所以會訂定機器行為的紀律，絕大部分是因為史無前例的 AI 潮流，演算法充斥人類的生活細節，像是新聞排行演算法操控市民關注市政的方向、配對軟體牽動人與人之間的緣分，甚至假設 AI 武器突破爭議，被實際部署在戰場上，將

擁有絕對掌控人類死活的能力。

有些時候，AI 產品能助長目標客群的進步，但同時也可能帶來潛在且無意的副作用。例如，推文演算法雖然能大幅渲染廣告的效果，卻也能高效擴散錯誤信息，讓過失一發不可收拾。此外，演算法的公平性也將深深影響人類社會的運行，包括信用評估、犯罪審判、電腦視覺的解讀和自然語言處理的詞嵌入等等，研究人員務必監督系統的流程，在損益之間權衡與妥協，才能避免加倍擴散社會上的種種不公。

然而，隨著 AI 系統日漸強大，應用層面變化萬千，機器行為的研究將越是艱難，尤其面對「黑盒子」時，科學家通常只得知輸入和輸出，很難推得演算的過程，更別說要詮釋並控管它。另外，某些原始碼和運算模型都可能涉及版權問題，就連訓練資料也是，光要審視多維度與大規模的資料就操得人類心力交瘁，再加上不透明性，機器的行為簡直難乎其難。

目前，研究機器行為的科學家通常專精於電腦科學與機器人學，多半為數學家或工程師，而非受訓過的行為研究家，他們很少接觸相關的統計資料、範例樣本和因果推斷，也不具備神經科學和社會學理論的專業。相對的，行為研究家缺乏模型訓練的能力，不太可能評估 AI 技術的品質，也沒辦法用數學方式解釋演算法的各類性質。



(圖一)

跨領域的機器行為研究雖然處於尷尬的交叉路口上，要將不同專業融合又非小菜一碟，我們卻不得不投入研究。對於機器行為的理解將貢獻給科技效力的相關研究，在這些研究分析各類科技產物對於社會的影響與貢獻之後，將反饋新

的靈感給 AI 工業，也提出新的問題給行為科學。最後，機器行為的研究將幫助 AI 學者更精準地發表機器的可行與不可行，提升 AI 的效率與效益。(如圖一)

研究方向與主題

Nikolaas Tinbergen 曾為動物行為提出四種分類，而儘管機器和動物有著根本上的差異，行為模式卻有幾分相似。機器會根據某些機制(mechanism)製造出行為，再經由發展(development)將環境中的信息導入行為之中，使得特定的行為在不同環境中表現出特定功能(function)，體現演化(evolution)歷史的脈絡，並描繪出未來行為的方向。

導致機器行為的近因在於，被激發的行為如何在一定的環境中產生，而這些機制(mechanism)取決於機器的演算法與環境因子。舉例來說，自駕車會在不同情境下超車、切車道或是打方向燈，這些行為的相關演算法都是建立在交通法規之上，也囊括汽車引擎系統的性質、物體辨識系統的解析與方向盤的掌舵等等。

談到行為的發展(development)，便是討論行為取得的過程，對機器而言，有一部分可歸因於人類在設計時的選擇。工程師會調配不同的參數和架設適合的神經網路，讓機器得以獲取相應的行為，若機器被投入刺激(stimuli)訓練，當中使用的訓練資料庫也可算是一種設計上的選擇。此外，機器還可能透過自身經驗提升表現，例如強化學習模型可以分析奇異的短線交易策略中，可能伴隨哪些市場回饋，再將經驗轉為最大化長線交易利潤的能量，讓下單更加精確、進步。

那該如何分析機器行為在特定環境中的功能(function)呢？我們可以聚焦在機器行為為環境帶來何種影響，使得它們能持續佔有一席之地。舉例來說，某些表現優良的交易演算法，可能會成為業界的榜樣，在不同公司之間交流，也可能受對手仿效、學習。相同地，越是顧及乘客安全的自駕車系統，越會受到市場寵愛，反之則會銷售不善，被淘汰出局。

其實，目前諸多 AI 產品都沿用了不少歷史設計，正所謂前人種樹後人乘涼，像貝氏狀態空間模型(Bayesian State-Space model)和某些神經網絡模型，能輕易讓新演算法推廣至另外一個，也能透過更新改善原先的缺失，讓設計上的複雜程度大幅下降。另外，研發出來的機器行為表現如果足夠完整，也可以流芳百世，演化(evolve)出更多變的新產品。

研究規模

我們可以針對不同的規模提出上述的問題，總共有三：單一機器行為、機器群體行為與人機混合行為。

單一機器行為研究的焦點是觀察機器在不同環境背景下的各種反應，是否合乎期待或者失控。例如，當 AI 系統面對與訓練資料嚴重背歧的參考資料時，能不能合理評斷一名罪人累犯的機率。另外，一個 AI 廣告系統在不同平台中，若有相同的輸入，能否分析出相近的推廣順序也會是科學家評判機器行為彈性與可容性的一大重點。

機器和動物、昆蟲類似，享有群體互動的生態。如果彼此互動優良，可能更順利也更出色地完成任務，好比說賑災行為，AI 之間的合作能夠提升搜救效率，在黃金救命時限中救出更多的受困者。然而，機器之間高效合作可能會提升失誤的風險，在金融市場方面更可能加倍損失，2010 年的「閃崩事件」就是個相當慘痛的例子，這也是科學家需要注重機器群體行為的原因之一。

當然，大家最關注的對象就是人與機器之間怎麼互動，往往大家都會注意到機器影響人類生活，殊不知人類也會影響機器的行為。舉例說明，機器的引入會改變社會中各大活動的型態，包含選舉、休閒、工作等等，而人類社會中的歧視會顯現在提供給機器訓練的資料庫中，人類的偏好也會影響演算法的選用。交互作用之下，機器與人類的互動會打造世界新的樣貌，或許會提升生活品質，又或許會加劇社會之間的不公不義。

此外，AI 潮流也將衝擊人類的勞動結構，究竟機器會完全取代人類，如自駕車取代郵差送件，還是像輔助醫生完成手術，成為人類的最佳拍檔？這是我們人類必須釐清的問題之一，畢竟它是影響人機混合行為演進走向的重大因素，也是科技世代附帶的倫理問題。

參考資料：

1. Iyad Rahwan, “Machine behaviour”, Nature, 24 Apr 2019