機器行為學

highscope.ch.ntu.edu.tw/wordpress/

機器行為學
編譯／黃柏瑋

諾貝爾經濟學獎得主Herbert Simon曾在《人造物的科學（暫譯）》（*Science of the 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，專門研究人造物與所引起的現象呢？」隨著AI逐漸滲透人類社交、文化、經濟與政治活動，理解AI行為，將風險和傷害降到最低，顯得更為重要。鑒於過往相關研究的缺乏與限制，部分學者主張以動物行為研究為模板，來討論機器的行為以及與環境的互動。

跨領域的行為研究

之所以有這樣的想法，主要在於我們生活中隨處可見、日趨重要的AI應用，小如社群網站的推文演算法，大如目前國際極力避免的AI武器化，除了預期的任務與目標外，也常有意料之外的負面影響。例如，推文演算法雖然能大幅渲染廣告成效，卻也促使錯誤訊息與假新聞高效擴散。此外，如演算法的公平性也深深影響著我們，尤其當已有部分業者將AI應用於個人金融的信用評估與犯罪案件的審判。

目前，機器行為的研究多侷限於如何提升AI效能，忽視AI間、與周遭環境、與人類之間的互動與交互影響。一部分的原因歸咎於存在已久的「黑盒子」問題：研究人員通常只能掌握輸入和輸出，很難推得演算的過程，更別說要詮釋並控管它。某些原始碼和運算模型的取得與公開更涉及版權問題（甚至訓練資料亦然），光要審視多維度與大規模的資料便令人心力交瘁，再加上不透明性，機器行為的研究難乎其難。

另一部分的原因，則在於「行為」本身便是跨越多個領域、融合不同專業的研究。以往只有專精電腦科學的數學家或工程師孤身奮戰，缺乏認知神經科學或社會學的理論基礎與實務經驗。所幸這樣的現象已隨著越來越多跨領域學者的投入而改善，並且這樣的互動是雙向的：對機器行為的理解與分析，除了有助於新科技的發展、提供新的靈感給AI工業外，這些研究資料與結果最終也成為傳統行為科學的養分、提出新的問題與研究素材。

四個面向

動物行為學（ethology）奠基者之一、也因此獲得1973年諾貝爾生理與醫學獎的Nikolaas Tinbergen曾提出動物行為研究必須考慮的四個面向：機制（mechanism，引發特定行為的物理機制與刺激）、發展（development，例如動物是否透過學習，或受環境制約而出現相同行為）、功能（function，以「生殖優勢」（reproductive fitness）的觀點釐清為何特定行為被保留，而其他行為卻逐漸消失）與演化（evolution，一個行為的演化歷程。一個行為現有的「功能」可能與其當初出現的原因不同）。

儘管機器和動物行為有著根本上的差異，但研究方法卻可類推。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、物體辨識系統的準確度等。這些演算法上的差異與環境因子，都是直接導致機器行為的「機制」。

至於行為的「發展」，意即行為的取得過程，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」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。

優秀的設計與理念也會被後起的諸多AI產品沿用，正所謂前人種樹後人乘涼，像貝氏狀態空間模型（Bayesian State-Space model）便是一例；或透過更新、改善原先的缺失，讓設計上的複雜程度大幅下降，「演化」出更多變的新產品。

由個體到人機混合

上述的四個面向不在於區分或割裂機器行為研究，而是如何兼顧行為本身的整體性，由個體、不同AI間、人機與環境間，將研究提升至更大規模與層次。與動物、昆蟲類似，機器也有群體互動的一面。如果彼此互動優良，可能更順利也更出色地完成指定任務。相較於單一AI的行為研究多聚焦於機器在不同環境、背景下的各種反應是否合乎預期，群體研究著重多個AI間的合作關係，以及如何提升效率。

當然，大眾最關注的便是人機互動：機器會如何影響人類生活？人類又會對機器造成那些影響？舉例來說，隨著機器引入各種人類活動，包含選舉、休閒、工作等等，而人類社會中固有的歧視與刻板印象自然地顯現在提供給機器的訓練資料中影響演算法。交互作用之下，人機共

存的世界必然呈現不同的樣貌，只是改善或加劇既有的不公不義？也是機器行為學此一新興學科迫切關注的問題之一。

編譯來源

I. Rahwan, “Machine behaviour”, *Nature*, 2019.

(本文由教育部補助「AI報報–AI科普推廣計畫」執行團隊編譯)