撰文／林守德

資訊界的馴獸技術

賞罰分明的加強式訓練，成就更靈活的人工智慧。

人工智慧代理人（AI agent）泛指具有某種程度的智慧，可以代理人類執行某些任務的軟硬體。例如在沙漠上探勘的機器人，必須能對各式各樣情境做出不同的反應：前方有障礙物擋住就盡快找到別的路徑，接收到生命反應要趨前察看，也要能迴避險惡環境……。要賦予機器人能力處理這些林林總總的狀況，最直覺的方式稱為「規則式學習」（rule-based learning）。首先要找對沙漠生態地形有研究的專家，盡量把所有可能發生的情境列舉出來，然後針對每個情境研議機器人該有的反應與動作（稱之為規則），最後再把這些規則寫到機器人的「腦」（中央處理器）裡面，它就會按照這些已有的規則行動。

規則式學習有幾個重大缺點：首先，要產生規則的話，一定要找到該領域相關的專業人士諮商，成本較高。第二，真正可能遇到的情境數以萬計，相對應產生的規則可能很多很複雜，要把這些規則都加以程式化、裝到機器人的中央處理器，需要耗費不少人力及時間；而當規則變多時，機器人搜尋可用規則的時間也會增加，導致反應遲緩。最後，也是最嚴重的缺點，當遇到不在規則內的情境時，機器人就會陷入無法判斷下一步的窘境。

於是，師法動物的訓練過程，資訊學家想到了一種方法，不需花費人力來窮舉可能的情境，也可以讓機器人應付各種況狀。動物經過訓練，往往可以表現出看似有高度智能的行為：例如猩猩可以使用工具，海豚可以隨著訓練員的手勢、聲音做出不同反應。「獎賞」（reward）跟「懲罰」（penalty）是訓練這些動物最常用的手段：獎勵通常是在動物表現符合期待時給予食物，而懲罰可能是在犯錯時給予體罰或是減少食物供給。

利用獎懲引導學習，是人類慣用的教育方式。把獎懲量化成分數，利用「最佳化」技術，就能讓電腦追求高分。

動物一開始也許不知道自己為什麼得到獎勵或懲罰，而必須從表現的行為中重複嘗試，判斷哪些行為會得到鼓勵、哪些會受到懲罰。多次經驗之後，受訓的動物就漸漸學到，在什麼場合要做什麼動作好贏得正面報酬。這樣的概念也被資訊學家用來訓練人工智慧代理人（如機器人），使其對於所處情境做出最正確的反應，這種方法稱為「加強式學習」（reinforcement learning）。

在訓練的一開始，先把機器人放到某個環境中讓它自由行動，同時在行動中即時告知它收到的獎賞與懲罰，例如被障礙物絆倒就扣分、跨過障礙物就加分等，這樣的過程我們稱之為「訓練行為」。在訓練過程中，機器人除了接受獎懲，也不斷會利用身上的感應器來感測環境。隨著訓練的次數增加，機器人會蒐集到越來越多「情境、行動、獎懲」如何發生的資料。然後就可以利用機率模型來計算在任一種情境之下，做出某種動作的「期望獎懲值」，也就是之後得到報酬的期望值。

有了這些期望值，機器人就可以藉由感測器所得的訊息來判斷當時情境，再根據情境選出當下「期望值最高」的動作來執行。這種「邊做邊學邊修正」的訓練模式，其實跟訓練動物有異曲同工之妙。訓練師並沒有明確教導動物看到什麼指令就要做什麼動作，而是利用獎懲一步步引導動物做出一連串符合要求的行為。

利用獎懲來引導學習，是人類慣用的教育方式。對於人類自己，獎懲也許只是一句鼓勵或責備；對於動物，食物是個好的誘因；而對於電腦就更簡單了，只要把獎懲量化成分數，然後利用「最佳化」的技術讓電腦追求高分即可。而這些被「訓練」完成的電腦程式，也已經在救災機器人、自動駕駛、電腦棋藝、網路競標程式等應用崗位上展現所長。 SA

林守德是台灣大學資訊工程系副教授。

影像來源：林守德