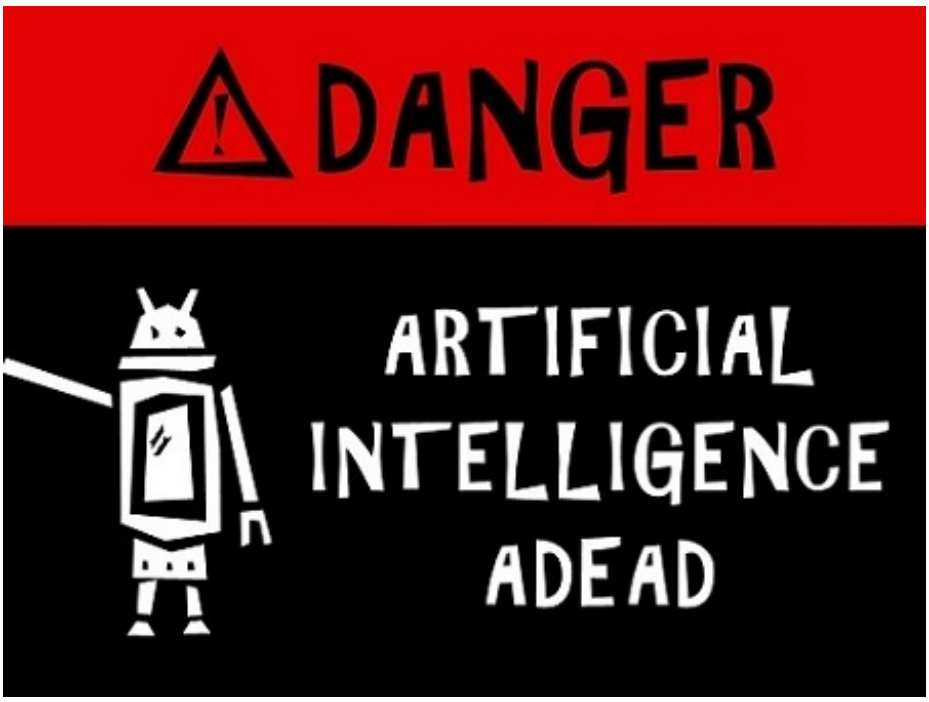
如何不讓AI失控



如何不讓AI失控
編譯／臺大資工系 黃柏瑋

跟教育小孩子一樣，「獎勵」往往是AI進步的動力與目標。然而，若有一個掃地機器人，盡心盡力將家中打掃得一塵不染，卻摔碎了價值千萬的花瓶；或者為了嚐甜頭，機器人把家中所有的灰塵、垃圾都藏進沙發底下，讓人類誤以為任務達成，這樣它們還值得獎勵嗎？又假如很不幸的，這些意外舉動發生在足以操控世界的超級AI (superintelligence)上，後果恐怕難以估量。為今之計，除了大量拓展AI的應用與效率外，如何確保AI在完成任務的過程中不會失控，降低意外發生的風險，顯然是一門相當重要的學問。

避免意外舉措

AI在為了獎勵而達成目標的過程中，很可能有意無意地忽略了它對工作環境的影響，例如：為了更快運送貨物，自駕車不斷蛇行也不禮讓行人，最終釀成車禍。美國總統老羅斯福曾說過：「溫言在口，大棒在手，故而致遠。」有賞有罰才能更有效地掌握秩序。

舉例來說，要在布滿電器的房間中運送裝水的桶子，需要格外小心。為了訓練機器人提水時不濺出，我們可以將其鎖進這類型的房間內，並針對不同情境給予不同的權利值(empowerment)：若機器人能夠將水桶提到指定的位置，權利值便會上升作為獎勵；但若在過程中機器人不小心撒出水來，懲罰就是降低權利值。最後唯有權利值夠大的機器人才能打開門鎖，藉此提升機器人提水時的謹慎程度。

然而，投機取巧的AI很可能使獎勵系統癱瘓。假設有個掃地機器人的目標是將視線內所有的垃圾清除，那它可能會根據經驗，選擇比較乾淨的路線，避開骯髒區域，以較少的努力換得更多的報酬，但並沒有因此讓整體環境更乾淨。

鑄成這種錯誤的部分原因，是AI行事並未考慮後果，因此有些專家會運用「前瞻模型」：根據AI行為的最終目標給予獎懲，而非像之前一樣獎勵現階段任務的完成。以上述為例，在掃地機器人工作之前先設定務必整理的區域，若它的清掃路徑不在該範圍之中，就算它把視野內的灰塵都清除(即現階段任務)，也不會得到報酬，因為對整體整潔的影響不大，無助於最終目標的達成。

擴展性監督

當我們利用AI時，總希望成果能符合預期，但有些時候結果與行為之間的等待時間漫長，因此我們需要訓練AI主動猜測人類的期望，作為自己工作的標準。然而，猜測的誤差難免，也導致AI的行為失控，於是我們可以運用「擴展性監督」(scalable oversight)，以較少的人力成本換得更有價值的成果，在提高AI效能的同時，也降低了誤會產生的風險，減輕人們對AI的不安全感。

「半監督式學習」(semi-supervised reward learning)是實現擴展性監督的模型之一。它貫徹了非監督式學習的精神，自行摸索出一套方法，並效法監督式學習，由人類處獲得回饋，但降低回饋的次數與訓練時間。半監督學習既能夠消弭人類與機器想法上的差異，也能有效減少人類的監督時程。再以掃地機器人為例，為了確認自己的想法是否與人類一致，它可能會在打掃到一定程度時向人類詢問「這房間乾淨嗎？」；如果人類覺得不乾淨，機器人便可能修正清理的方式，直到人類覺得「這房間乾淨」為止。

半監督式學習也可以搭配「遠程監督」(distant supervision)，強化關係抽取的方法。首先，透過遠程監督連結兩數據：假定一語句中若包含某兩個實體名稱，這句子便是在敘述兩實體間特定的關係。例如：只要出現「蘋果」和「賈伯斯」這兩個名詞，就是在說「賈伯斯是蘋果的創始人之一」。當然，這樣的貼標方式十分不嚴謹，因為兩名詞間的關係並不唯一，像是「賈伯斯喜歡吃蘋果」就是另一種新的關係。因此我們可以將含有相同名詞但敘述不同關係的語句分開，再利用半監督式學習將它們重新正確標籤，以訓練如何更精準抽取語句的目標關係。

超級AI

事實上，人類很難去計算人工智能的極限何在。即使知道了，也無法確定他們的最終目標，即使枯燥如數沙粒或計算圓周率，都可能造成足以影響全世界的後果。這些未知所帶來的不確定性使得人心惶惶。因此，如何確實掌握AI、分析與預測他們的動機與行為可說是重中之重。

隨著人工智能的進步，其認知可能因而提升(cognitive enhancement)，曲解原先人類為其設立的目標或作出令人意想不到的行為。為達目的，AI或轉而尋找更完美的科技(technological perfection)，或取得更豐富的資源(resource acquisition)，「工具趨同性」(Instrumental convergence)也是我們可以利用，以建立停損點的工具。

比方說，若有人工智能可以無限發展，最終成為舉世無雙的超級AI，或許會學習利用太空資源來改造地表樣貌。屆時，人類便可根據其所研發的新拓墾技術，大致猜測超級AI的最終目標，提早擬定應對措施，避免招致無以挽回的後果。

編譯來源：

1. D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J Schulman, and D. Mane. " Concrete problems in AI safety." Arxiv.org, July, 2016.

2. N. Bostrom. " The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents." Minds and Machines, May 2012.

參考資料：

1. YU Xiaokang , CHEN Ling , GUO Jing , CAI Yaya , WU Yong , WANG Jingchang, “Relation extraction method combining clause level distant supervision and semi-supervised ensemble learning”, Research Gate, January 2017