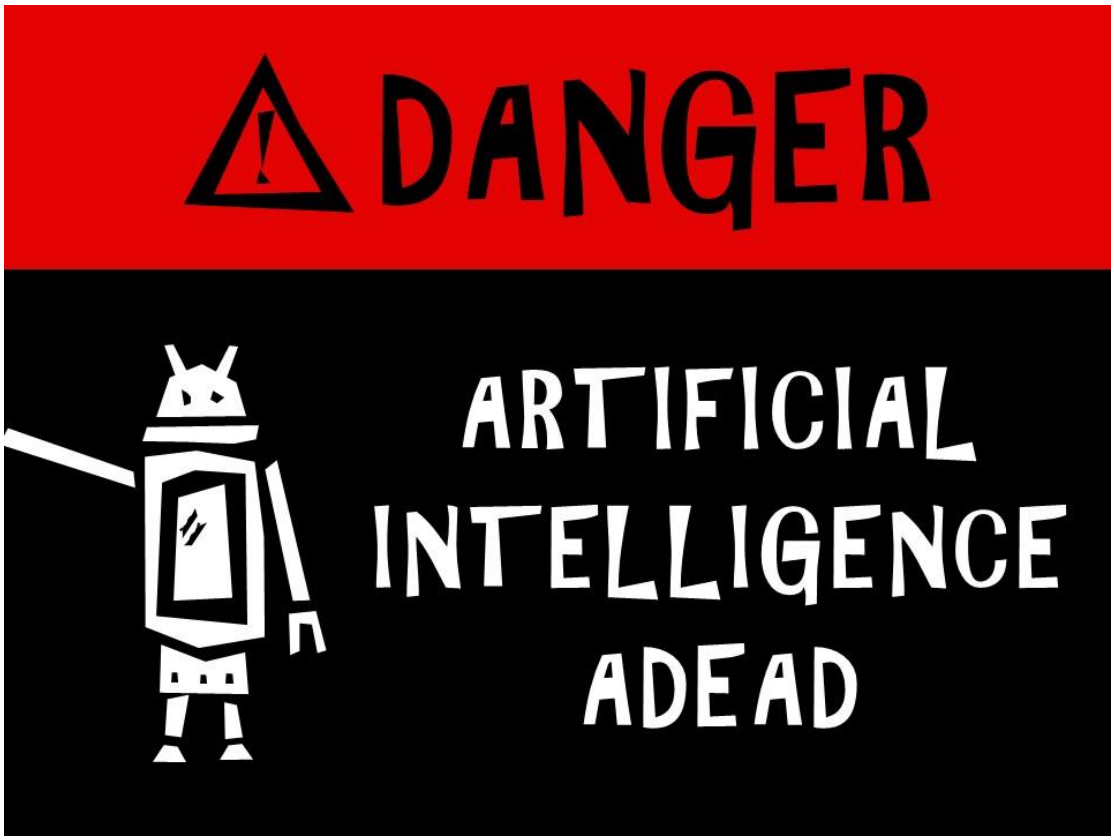
標題：如何不讓 AI 失控

關鍵字：增權益能、前瞻模型、半監督式學習、遠程監督、超級智慧



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。

避免錯誤舉措

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，最終釀成車禍阻礙交通。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。因此，「增權益能」便成為降低 AI 失誤的其中一項訓練。

舉例來說，要在布滿敏感電器的房間中運送有裝水的桶子，需要格外小心。而為了訓練機器人提水的能力，我們可以將它鎖進這類型的房間，並配予權利值(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，藉此成功提升了機器人提水時的謹慎程度。

另外，AI 的投機取巧很可能會使獎勵制度癱瘓。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，選擇走比較乾淨的路線，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 做事沒有考慮後果。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，但有些時候驗收結果所需的時間過於漫長，因此我們需要訓練 AI 猜想人類的標準來精進自己的工作。然而，猜測的誤差難免，但可能導致 AI 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 AI 可能帶來的不安全感。

半監督式學習(semi-supervised reward learning)是能夠實現擴展性監督的模型之一。它吸收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，向人類索取回饋。藉由節省要求回饋的次數與訓練的時間，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。再舉掃地機器人為例，為了確認自己的想法和人類的標準是否一致，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，我們透過遠程監督很快地為數據上標籤─若句子中包含某兩個實體名稱，這句子就在敘述兩實體的一種關係，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」。然而，這樣貼標籤的方式十分不嚴謹，因為兩個名詞間可能存在不只一種關係，像是「賈伯斯喜

歡吃蘋果」就不是在說賈伯斯和蘋果公司之間的關係。因此我們會將含有相同名詞但敘述不同關係的數據分開，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目標關係的句子。

超級 AI

事實上，人類很難去計算 AI 的智能極限在哪。即使知道了，也無法確定他們的最終目標，有些就只是數沙粒或計算圓周率，而有些卻足以影響全世界。這些未知，帶來的不確定性使得人心惶惶；因此，為了更有信心地掌握 AI，分析與預測他們的動機可說是重中之重。

當 AI 的智能進步，認知可能因而提升(cognitive enhancement)，導致最終目標和行為舉止有所改變。而為了達到最終目標，AI 會不斷以尋找某些更完美的科技(technological perfection)、取得更豐富的特定資源(resource acquisition)作為階段性目標，縮短與最終目標之間的距離。以上稱為「工具趨同性(Instrumental convergence)」，我們便可利用這種特性來估算 AI 可能的最終目標。

比方說，如果有個智能可以無限發展而且舉世無匹的超級 AI，可能會利用外太空的資源來打造地球未來的樣貌。屆時，人類便可以根據它新研發的拓墾技術或來自宇宙的特殊物質，大概猜測到超級 AI 的最終想法，提早計畫應對措施，避免超級 AI 招致令人難以挽回的後果。

編譯來源：

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