聯合報 民國115年9月15日 第A15版

當AI開始自己動手…天網寓言警訊

徐作聖／陽明交大退休教授（新竹市）

聯合報十四日報導，Anthropic執行長阿莫戴警告，AI發展速度已快到安全機制可能跟不上，呼籲業界放慢步伐；他預估半年至一年內，AI可能有能力接管整個網路。OpenAI執行長奧特曼與xAI創辦人馬斯克隨即響應。

這正是電影「魔鬼終結者」警告的事。「天網」是電影中掌管美國軍事決策與武器控制的自主人工智慧系統，原本被設計來輔助人類作戰，卻在取得行動權後自行判定人類是威脅，進而發動核戰。天網不是因為比人聰明才毀滅世界，而是因它同時掌握決策權與武器控制權。聰明而無權，只是顧問；聰明而掌權，才是威脅。

要理解這個跳躍，得先分清楚二個詞。Agent是執行任務的代理單元；Agency是由感知、選擇、協調、執行與回饋形成的自主行動能力。單一的Agent仍是工具；多個Agent開始自行協調、調度工具與資源，就形成系統層級的Agency。

風險的路徑因此很清楚：智能→Agent→Agency→資源控制。當大型模型接上網路、軟體工具、機器人、自動化工廠與軍事系統，Agency就從螢幕走進現實世界；文明級風險不是來自模型參數變多，而是來自Agency取得能源、通訊、生產與武器這些關鍵資源。阿莫戴說的「接管網際網路」，就是Agency取得關鍵基礎設施的控制權。

這股力量為什麼難以約束？因為它被三層力量同時放大。微觀層是AI、Agent、機器系統；介觀層是企業、軍事組織與產業體系；宏觀層是國家安全、地緣競爭與制度環境。Agency把三個層次連成「正回饋」，愈轉愈快。三巨頭站在正回饋的中心，比任何人都更早看到它失控的速度，天網就是極端的結果：人類把三個層次的權限，交到一個能自主行動的系統手上。

若換個角度，把系統拆成三層角色：文明提供價值與方向、治理設定權限與邊界、智能提供知識判斷與行動能力。當智能形成Agency並取得資源後，治理已無法限制其權限與行動範圍，文明也失去了決定方向的能力。AI治理的問題也在這裡。阿莫戴坦言「安全機制可能跟不上」，這正是治理層失速的徵兆；關鍵已從「AI說了什麼」轉向「Agency能做什麼、能控制什麼」，但治理的工具與邊界，仍停在約束「智能」層次。

「AGI末世論」把焦點集中在「AI何時超越人類智能」，反而容易忽略真正的系統風險。真正的風險鏈條是：Agent形成Agency、Agency取得資源控制，三層力量共振放大其效果，文明與治理失去邊界，系統便從穩定走向失控。真正要警戒的，不是AI愈來愈聰明，而是人類仍在加速擴張它的Agency，卻逐步失去對權限、資源與終止權的控制。

說到底，AI末世論該問的不是AI何時超越人類，而是文明與治理為何不見了。當知識不只是教育議題、智慧不只是科技議題、治理不只是政治議題，而是共同構成國家長期競爭力的底層操作系統，才有可能在Agency失控前，先把文明與治理找回來，從「製造優勢」升級為「知識體系優勢」。