人工智慧的啟示

文／林一平　講座教授

圖靈獎（Turing Award）得主 Geoffrey Hinton 在日前公開討論人工智慧（AI）的風險。

AI「往往會從分析大量數據中學到意想不到的行為」。這並非意味著具有自主意識的 AI 會摧毀人類，而是我們無法預測 AI 的行為，特別是當個人和企業允許 AI 系統不僅生成其自身的代碼，而且在自己的計算機上運行這些程序時，Hinton 擔心「有一天，真正的自主武器將那些殺手機器人變成現實」。

第一個實際的 AI 系統是由 Edward Feigenbaum 及 Raj Reddy 實現，稱為「專家系統」，是一種智慧型的電腦程序，能運用知識與推論來解決只有專家才能解決的複雜問題；他們也因此一貢獻榮獲 1994 年的圖靈獎。

然而，許多系統需要模擬的參數甚多，至今仍然無解。可見計算機模擬的應用博大精深，即使今日 AI 技術突飛猛進，有許多題目仍值得深入研究。

圖靈（Alan Turing，1912~1954）在 1950 年發表一篇重要論文〈計算機與智慧〉"Computing Machinery and Intelligence"，首次談論到 AI，並提出圖靈測試（Turing test），為資訊領域創建智慧設計的標竿。

圖靈測試指的是，如果一台計算機能夠欺騙人類， 相信它是人類，那麼它就應該稱為智能計算機。AI 緣起於模擬人類行為，自然也常用於社會學。

密西根大學的政治學教授 Robert Axelrod，在 1980 年代進行一連串電腦模擬實驗，找一群專家寫出不同電腦程式，模擬人類行為，讓這些程式互動、合縱連橫，看哪個程式最後會勝出。這些程式有些模擬「金律」，有些模擬「銀律」，有些則模擬「鐵律」。

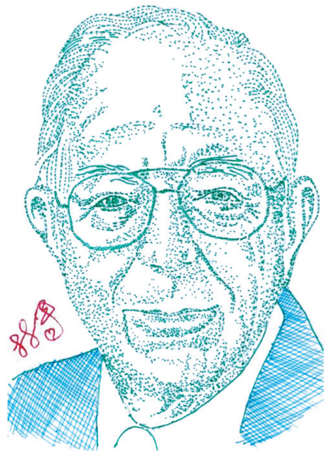
所謂「金律」（Golden Rule），語出《新約》< 馬太福音 >7:12「無論何事、你們願意人怎樣待你們、你們也要怎樣待人」；「銀律」（Silver Rule），語出《舊約》< 出埃及記 >21:24「以眼還眼，以牙還牙，以手還手，以腳還腳」；「鐵律」就是「己所不欲，先施於人」，外在表現是「先下手為強，後下手遭殃」。

結果最成功的是模擬「銀律」的 Tit-for-Tat 程式。這個程式一開始採取合作，若對方也肯合作，接下來則仍採合作策略；若對方吃你豆腐，下一步你就佔回便宜。

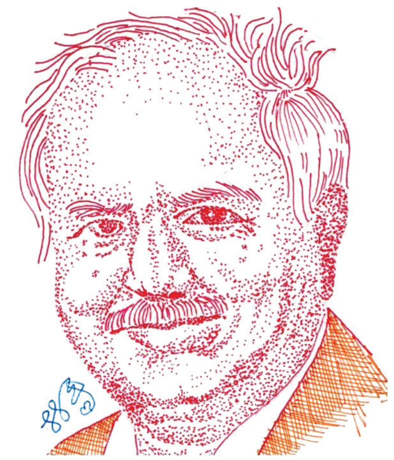
在實驗中，實施金律的程式一敗塗地，屍骨無存，可見咱們先總統蔣公介石對日本「以德報怨」的做法是行不通的；實施鐵律策略的程式一開始也有不錯的表現，但長期下來，所有被它吃豆腐的人不是死了，就是躲它遠遠的，它最後也沒戲唱。

有一個鐵律例子，就是石油大王 John Rockefeller（1839～1937）。

他專耍先下手為強的手段，整垮所有對手，成為最有錢的人。但他的手段未免太狠，大夥都不敢恭維。Rockefeller 也知道自己以前做事實在不上道，因此在退休後的餘生，致力於慈善事業補過。然而，他過去的作為仍然禍貽子孫，他的後人能力再強，條件再好，想選總統，至今都選不上。



「專家系統」之父，Edward Feigenbaum。



Raj Reddy與 Edward Feigenbaum共同獲得 1994年的圖靈獎殊榮，也是該獎項的第一位亞裔專家。

AI Revelations

Geoffrey Hinton, Turing Award recipient, recently talked about the potential dangers posed by artificial intelligence (AI).

Frequently, AI acquires unexpected behaviors through the analysis of extensive datasets. However, this doesn't necessarily imply that AI possessing autonomous consciousness will bring about the downfall of humanity. Instead, it highlights our inability to predict AI behavior, especially when individuals and businesses allow AI systems not only to generate their code but also to execute these programs on their own machines. Hinton is concerned thatOne day, genuine autonomous weapons may turn those lethal robots into reality.

The first functional AI systems, referred to as expert systems, were developed by Edward Feigenbaum and Raj Reddy independently. These intelligent computer programs have the ability to employ knowledge and reasoning to address intricate issues that are typically within the realm of experts' capabilities. Their achievements in this field earned them the prestigious Turing Award in 1994.

However, many systems require the simulation of a significant number of parameters, and this challenge remains unsolved to this day. This underscores the extensive and profound applications of computer simulation. Despite the rapid progress of AI technology, there are still many subjects deserving of thorough exploration.

In 1950, Alan Turing (1912-1954) published a significant paper titled "Computing Machinery and Intelligence," in which he initiated discussions about AI and proposed the Turing test. This test established a benchmark for developing intelligent systems within the field of information technology.

The Turing test revolves around the criterion of whether a computer program can successfully trick humans into thinking it is human. Once it accomplishes this, it would be qualified as an intelligent machine. The foundation of AI is rooted in emulating human behaviors and is frequently applied in the field of sociology as well.

During the 1980s, Robert Axelrod, a professor of political science at the University of Michigan, conducted a series of computer simulation experiments. He assembled a team of specialists to develop various computer programs that replicated human behavior. These programs were set up to interact with each other and form alliances to determine the ultimate winner among them. A subset of these programs emulated the 'Golden Rule,' while others followed the 'Silver Rule,' and the rest simulated the 'Iron Rule.'

The so-called "Golden Rule" comes from the New Testament, Matthew 7:12, where it states, "In everything, do to others what you would have them do to you." The concept of the "Silver Rule" is retrieved from the Old Testament, in Exodus 21:24, "eye for eye, tooth for tooth, hand for hand, foot for foot." The "Iron Rule," often interpreted as "Do unto others as you like before they do unto you," is externally manifested as "He who strikes first prevails, he who strikes late fails."

The Tit-for-Tat program, which simulates the 'Silver Rule,' emerged as the most successful outcome. Initiated with cooperation, the program maintains the cooperative strategy if reciprocated by the counterpart. In cases of betrayal by the counterpart, the subsequent action involves responding with corresponding retaliation.

During the experiment, the program applying the Golden Rule experienced complete failure, leaving no evidence of its effectiveness. This demonstrates the impracticality of the approach advocated by our former President, Chiang Kai-shek, who aimed to respond to Japan's hostilities with kindness. Similarly, the program utilizing the Iron Rule strategy showed promising results initially. However, as time passed, those affected by its effects either suffered severe consequences or deliberately kept their distance from it. Ultimately, this strategy also failed to yield any meaningful outcomes.

A classic example of the Iron Rule is the oil tycoon John Rockefeller (1839-1937).

He employed ruthless tactics to initiate preemptive actions and get ahead, which led to the downfall of all rivals and propelled him to become the wealthiest person. However, his approaches were subjectively deemed excessive, causing people to be reluctant to express approval. Rockefeller himself was aware of the ethical shortcomings of his prior actions. Consequently, he committed himself to charitable pursuits in an attempt to make amends during his later years after retirement. Nevertheless, the consequences of his past actions continued to trouble his descendants. Regardless of the abilities or favorable circumstances enjoyed by his descendants, their aspirations for the presidency have remained unattainable to this day.