機器人也懂倫理

不久之後，擁有自主能力的機器人就會在我們的生活中扮演重要角色，但是在此之前，它們要先學會遵守規範才行。

撰文／麥可．安德森（Michael Anderson）、蘇珊．萊伊．安德森（Susan Leigh Anderson）
翻譯／甘錫安

重點提要

- 可自主做出決定的機器人，例如用於協助老人生活的機器人，即使在看似平常的狀況下，也可能面臨倫理困境。
- 確保機器人能以合乎倫理的行為與人類互動的方法之一，是將一般倫理原則輸入機器人，並讓機器人在各種狀況下運用這些原則做出決定。
- 人工智慧技術可借助邏輯，由各種倫理上可接受的行為案例中，自行歸納、產生原則。
- 本文作者依據此方法，把程式寫到機器人裡，做出第一個依據倫理原則行動的機器人。

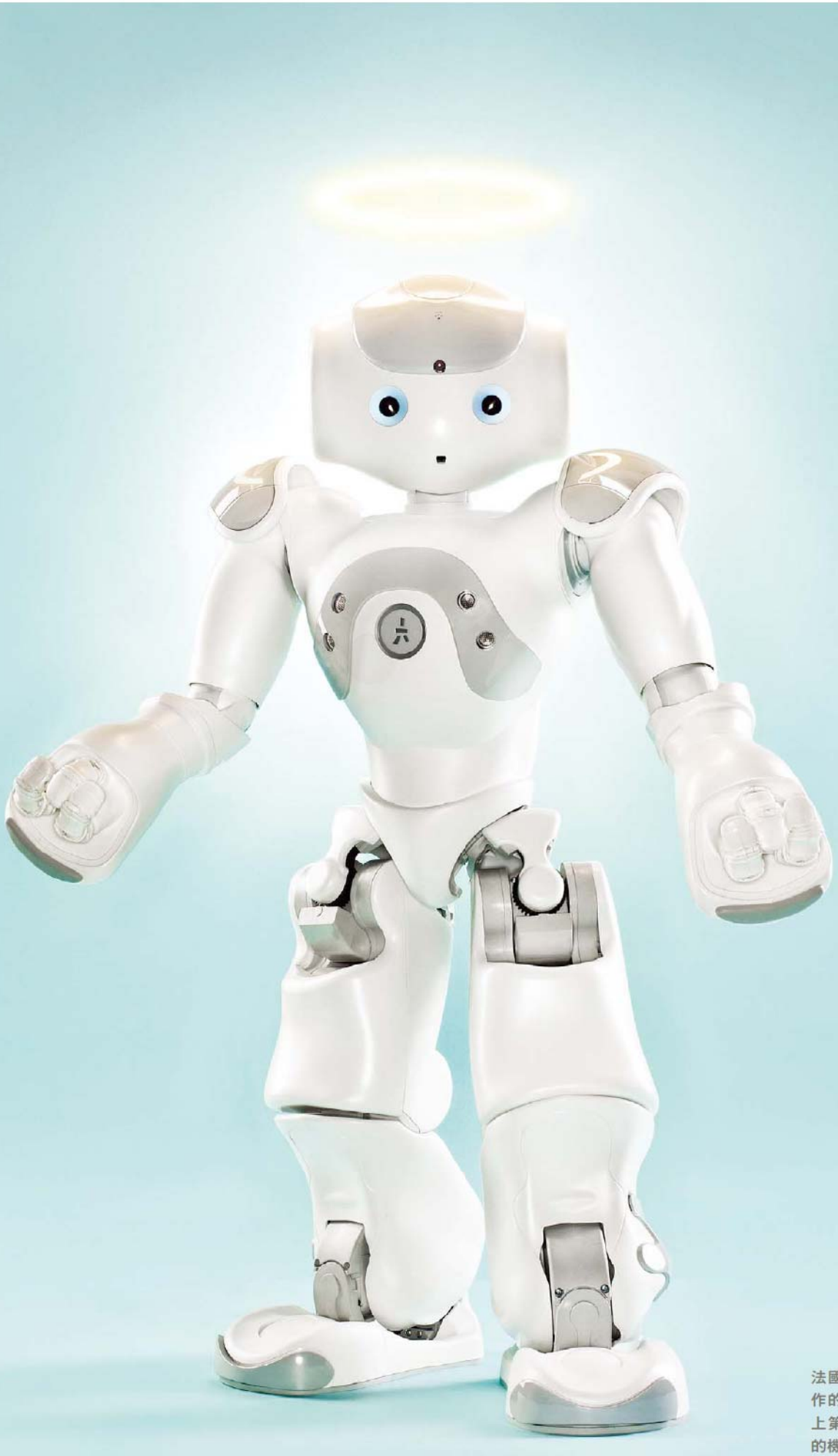
機器人聰明到足以挑戰人類，是驚悚科幻小說裡常見的恐怖情節，而且機器人絕對不會因為傷害甚至毀滅人類而感到絲毫內疚。當然，目前機器人的用途大多是幫助人類，但即使在相當平常的狀況下，機器人也會面臨許多倫理上的挑戰，而人工智慧也不斷在突破這些困境。

想像一下，機器人可能很快就會出現在安養機構，而你就住在安養機構裡。接近上午11點時，你要娛樂室裡的機器人幫你拿遙控器，好讓你打開電視收看「超級星光大道」。但另一位住客也想拿遙控器，因為她想看「全民估價王」，最後機器人決定把遙控器拿給她。剛開始你有點不高興，但機器人解釋這個決定很公平，因為你今天已經看過你最喜歡的晨間節目。這個故事是很平常的倫理決策行為範例，但對於機器人而言，卻是十分難以達成的重大成就。

前面描述的狀況只是理論上的模擬，但我們已經製作出第一款能做出類似決定的示範機器人。我們賦予這具機器人倫理原則，讓它能依照這個原則決定該隔多久提醒病患吃藥。機器人的程式目前只能從少數幾種可能的選項中選取其中之一，例如是否要一直提醒病患吃藥與何時該提醒，或是什麼狀況下該接受病患不吃藥的決定，不過就我們所知，這是第一款依據倫理原則決定行動的機器人。

但要預測機器人可能面臨的所有抉擇，並將之寫入程式，讓機器人在各種想像得到的狀況下都能妥善處理，卻相當困難，可以說是不可能。但另一方面，如果完全不讓機器人採取需

PHOTOGRAPH BY DARREN BRAUN



法國隨從機器人公司製作的機器人Nao，是史上第一具具有倫理原則的機器人。

要做出倫理抉擇的行動，又可能限制機器人執行得以大幅改善人類生活的任務。我們認為，解決方法是讓機器人能將倫理原則運用在預料之外的新狀況下，好比說除了判斷該讓誰拿遙控器之外，還可以決定該讓誰看新書等。這種方式還有一個優點，就是當機器人被要求解釋自己的行為時，可以參考這些原則。如果要讓人類自在地與機器人互動，這點十分重要。另外一個附加優點則是，倫理機器人的開發工作，也可促使哲學家探究日常生活狀況，帶動倫理學這個領域本身的進步。就像美國塔弗茲大學哲學家鄧奈特（Daniel C. Dennett）最近說過的：「人工智慧使哲學更誠實。」

我，機器人

不久之後，具有自主能力的機器人可能就會成為日常生活的一部份。現在已經有飛機能夠自己飛行，能自動駕駛的汽車也已進入開發階段，連從燈光到空調等一切運作都由電腦控制的「智慧型住宅」，也可想成是身體就是房屋的機器人，庫柏力克的電影「2001：太空漫遊」中的HAL 9000，其實就是一部自動化太空船的大腦。目前已有數家公司正在開發能協助銀髮族打理日常生活的機器人，除了可以協助安養機構人員工作，也可幫助長者在家中獨立生活。儘管這類機器人大多不必做出攸關生死的決定，但要讓一般人接受它們，必須先讓大眾認為它們的行為公平正確，或者至少是良善的。因此，機器人的研發人員最好能將程式會帶來的倫理歧見列入考量。

如果你也認為將倫理原則置入具有自主能力的機器人，是機器人與人類順利互動的關鍵，那麼第一個問題就是應該置入哪些原則？科幻小說迷可能會認為，艾西莫夫（Isaac Asimov）多年前已經提出了答案，那就是著名的「機器人學三大法則」：

關於作者

麥可．安德森是美國康乃狄格大學博士，並擔任哈佛大學計算機科學系副教授，長期關注人工智慧進展。

蘇珊．萊伊．安德森於美國加州大學洛杉磯分校取得博士學位，目前是康乃狄格大學哲學榮譽教授，專精應用倫理學。2005年，她與麥可．安德森協助舉辦第一屆國際機器倫理學研討會。他們有一本關於機器倫理學的著作，即將由劍橋大學出版社出版。

1. 機器人不得傷害人類，或者坐視人類受到傷害而袖手旁觀。
2. 除非違背第一法則，機器人必須服從人類的命令。
3. 在不違背第一及第二法則的情況下，機器人必須保護自己。

艾西莫夫於1942年在一篇短篇小說中首次提出這三項法則，但已經有些人在探討該篇小說時，發現了其中的矛盾。艾西莫夫自己也在1976年的短篇小說《變人》（*The Bicentennial Man*）中描述這些法則有多麼不合宜。在這篇小說中，壞人要機器人拆解自己。在第二法則下，機器人必須遵守壞人的命令；又不可能在不傷害人類的狀況下自衛，因為這樣就違反了第一法則。

如果艾西莫夫的法則行不通，又有什麼其他方案？真有其他方案存在嗎？有人認為，讓機器人做出合於倫理的行為是痴人說夢，他們表示，倫理不可能透過計算得出，因此也不可能寫入機器人的程式中。不過早在19世紀，英國哲學家邊沁（Jeremy Bentham）和彌爾（John Stuart Mill）就主張倫理決策是一種「道德計算」。他們反對以主觀意識為基礎的倫理，提出享樂的行為效益主義（Hedonistic Act Utilitarianism），主張將所有相關人等感受到的愉悅單位數加總並減去不愉悅單位總數，可望達成最大「淨愉悅」的行為，就是正確的行為。倫理學家大多懷疑此理論是否真能涵括倫理考量的所有面向，舉例來說，這個理論很難考量到公義程度，而且可能導致犧牲個人來成全大多數人利益的狀況。但這項理論至少證明了，可信的倫理理論原則上可以計算出來。

有些人懷疑機器人是否真能做出倫理決策，因為機器人缺乏感情，無法體會人類受機器人行為影響時的感受。但人類又很容易受感情左右，經常因此做出不合倫理的舉動。人類的這項特質，加上我們容易偏袒自己和親近的人，使人類在倫理決策方面的表現並不理想。我們認為，受過適當訓練的機器人或許可做到絕對公正，而且雖然本身沒有感情，但能察覺人類的感情，並將之列入計算。

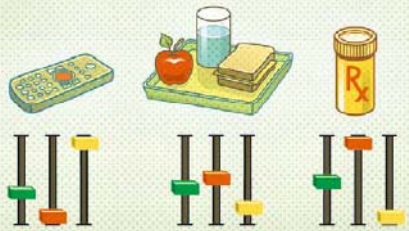
由範例學倫理

假設我們能將倫理規則輸入機器人，那麼應該輸入誰的倫理規則呢？畢竟到目前為止，還沒有人能夠提出放諸四海皆準的通用倫理原則，讓真正的人類遵守。但機器人通常是在特定某些地點工作，在這個前提下決定行為倫理參

〔 史上首位倫理機器人 〕

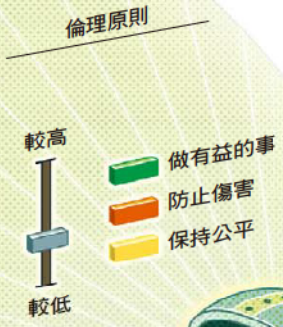
行為程式編寫原則

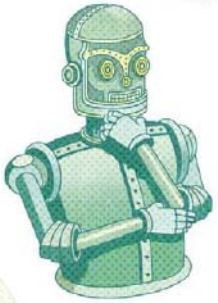
與人類互動的機器人經常必須做出可能出現倫理歧見的決定。程式設計人員無法預測到機器人可能面臨的所有倫理困境，但可以提供通用原則（下圖），做為在個別狀況中做決定的準則（右圖）。本文作者展示他們為機器人Nao（見第33頁照片）設計的程式，可決定是否要提醒病患吃藥，以及提醒的頻率。

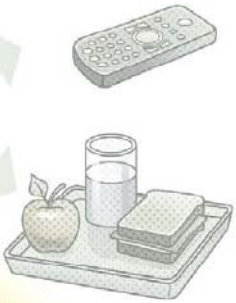


設定規則

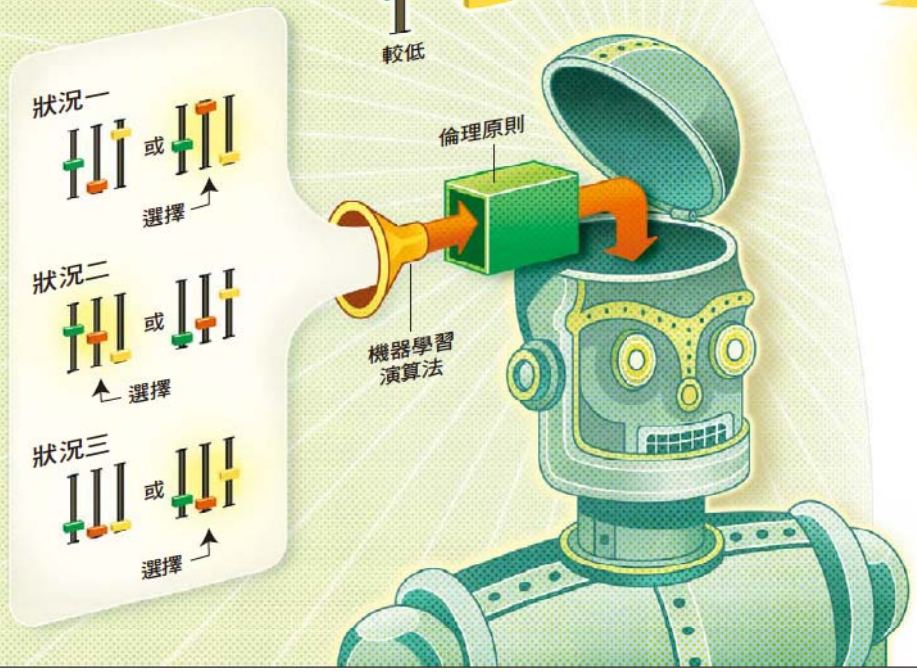
設計人員可運用稱為「機器學習」的人工智慧技術訂定倫理原則，並以此原則設計程式。接著，設計人員把資料輸入給機器學習演算法，說明在特定狀況下做決定時應該考慮的倫理，執行選擇的依據則是某個動作可能造成多大的好處、可以防止多大的傷害、公平程度等。演算法接著將資料歸納成可套用在新狀況的通用原則。







在這種狀況下，倫理原則會提醒機器人送藥，而不是執行其他工作。



做出決定

協助老人生活的機器人能針對各種可能動作，評估其是否符合倫理原則，接著依據評估結果以及本身包含的原則，計算在某個時間應該優先採取什麼行動。舉例來說，假如有一位住客想吃點心，另一位要電視遙控器，機器人可能還是會決定先進行另一項任務，例如提醒某位病患吃藥。

ILLUSTRATION BY PETER AND MARIA HOEY

數，會比制訂通用規則來判定行為是否合乎倫理簡單得多，而後者其實就是倫理學家想做的事。不僅如此，在描述機器人可能運作的各種背景下發生特定狀況時，針對哪些事情在倫理上允許、哪些不允許，倫理學家也大多有共識；如果沒有這類共識，就不該讓機器人自行做決定。

研究人員已經提出很多種擬定機器人行為規則的方法，大多是透過人工智慧技術。舉例來說，日本北海道大學的荒木健治（Kenji Araki）和瑞普卡（Rafal Rzepka）在2005年提出了「民主式演算法」。這種方法是在網路上發掘資訊，看看一般人認為哪些行為在倫理上可以接受，再使用統計分析，針對新問題找出答案。2006年，加拿大溫莎大學安大略分校的瓜利尼（Marcello Guarini）提出以現有的案例訓練類神經網絡，讓機器人之後在類似的狀況下能辨識並選擇合乎倫理的決定。類神經網絡是效法人類大腦

〔 機器倫理學的里程碑 〕

當科學擁抱文學

早在倫理學家、機器人專家和人工智慧專家對機器人行為可能帶來的倫理歧見有興趣之前，科幻小說家和電影導演就已在探討這些並非完全虛構的情節。不過近年來，機器倫理學已經成了真實的研究領域，有些靈感得自於18世紀哲學家的作品。

1495 達文西設計了史上第一具人形機器人。

1780年代 邊沁（上圖）與彌爾提出倫理是可計算的。

1921 恰佩克（Karel Čapek）的劇本《羅素姆的萬能機器人》首先提出「機器人」（robot）這個詞以及機器人造反的概念。

1750 1800 1850

學習方式的演算法，能學習如何處理資訊，而且處理效率會越來越高。

從研究結果看來，我們認為倫理決策必須在數種責任間取得平衡，倫理學家稱這些責任為初步責任（prima facie duty）。我們通常會將這些責任全部負起，但某個責任在某些時候可能會被另一個推翻。舉例來說，一般人通常會履行承諾，但如果違背某個不重要的承諾可以防範許多傷害，食言也是人之常情。當責任彼此衝突時，倫理原則就可決定在不同的狀況下，應該優先考慮哪個責任。

為了制訂可以輸入機器人的倫理原則，我們採用稱為「機器學習」的人工智慧技術。我們首先累積各種特殊案例到具有代表性的數量，以及在這些案例中一般認為合於倫理的決定，讓演算法消化這些資料。接著使用歸納邏輯法，歸納出倫理原則。這種「學習」階段在軟體設計時進行，再將得出的倫理原則放入機器人的程式中。

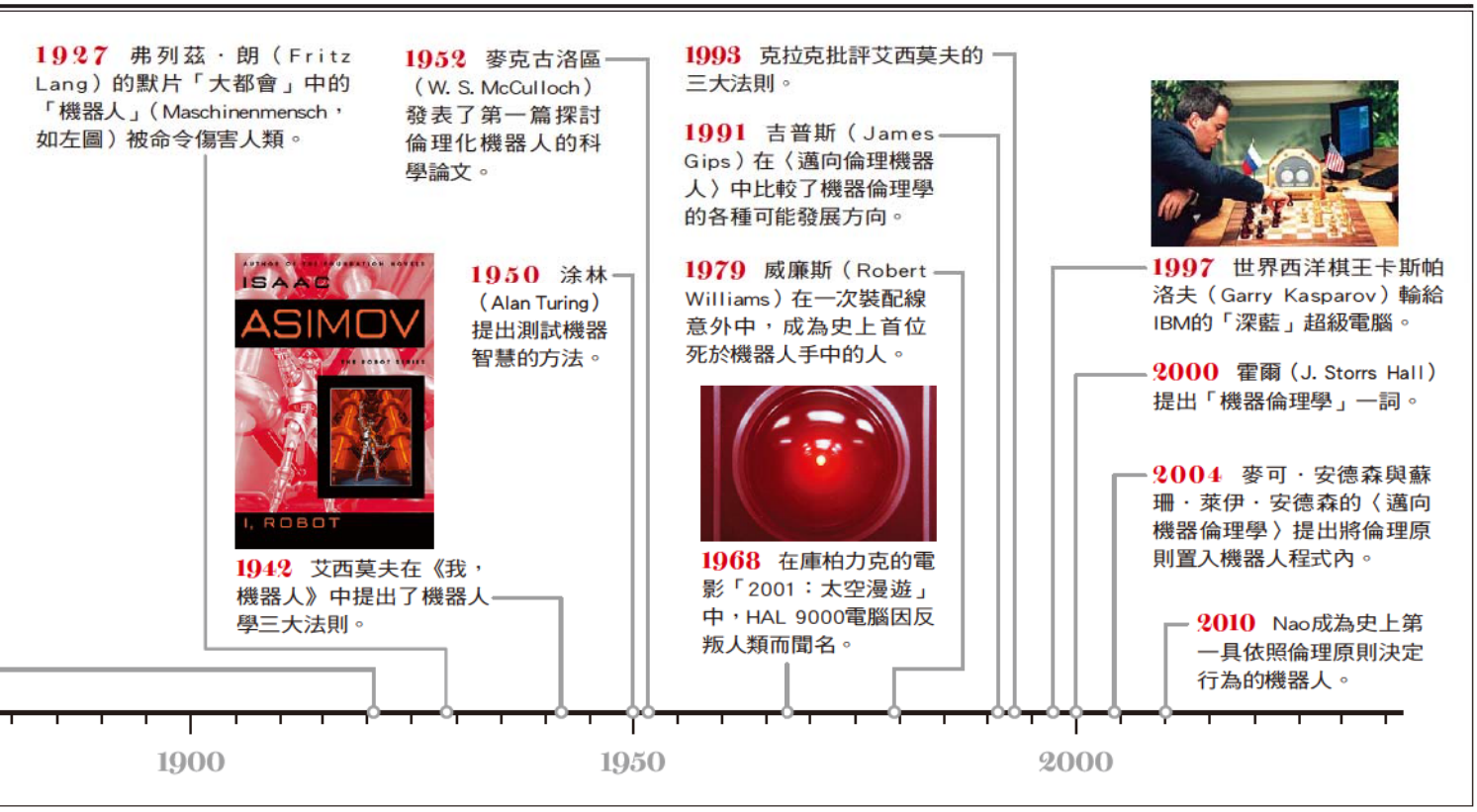
我們為這個方法設計的第一個測試情境，是機器人必須提醒病患吃藥，同時在病患不聽話時通知管理人員。機器人必須在三種責任間權衡：確保病患獲得吃藥帶來的好處、防止不吃藥可能造成的壞處、尊重成年且有行為能力病患的自主權。在醫學倫理中，尊重病患自主權尤其重要，如果機器人太常提醒病患或太快向管理人員告狀，就可能違反這個責任。

輸入特殊案例的相關資料後，機器學習演算法就訂定出以下的倫理原則：當其他做法都難以避免傷害，或是與增進病患福祉的責任背道而馳時，健康照護機器人應該挑戰病患的決定，也就是違反病患的自主權 。

倫理機器人上場

接著，我們將這項倫理原則輸入法國隨從機器人公司（Aldebaran Robotics）開發的人形機器人Nao中。Nao能尋找並走向需要提醒吃藥的病患，將藥送給病患，用自然語言與病患互動，並在必要時以電子郵件通知管理人員（在這裡通常是醫師）。管理人員會先把下列資料輸入給Nao：服藥時間、如果不服藥可能造成的最大傷害、最大傷害可能多久會出現、服藥的最大預期效益，以及這項效益的消失時間。機器人可藉由輸入的資料，計算這三項責任的滿足度或違反度，並依據這些程度隨時間改變的狀況，採取不同的行動。根據倫理原則，當責任滿足度與違反度到達一定門檻，讓提醒比不提醒來得好時，機器人就會發出提醒。等到病患可能因不服藥而受到傷害或喪失重大權益時，機器人才會通知管理人員。

經過充份訓練的老人照護機器人（簡稱EthEl，姑且叫



COURTESY OF RANDOM HOUSE (*Asimov book cover*); MGM/THE KOBAL COLLECTION (*HAL 9000*); AFP/GETTY IMAGES (*Kasparov and Deep Blue*)

艾瑟兒），需要更複雜的倫理原則，來規範更多種的行為，但一般方法仍然相同。在安養機構中工作時，機器人將以此原則判定各項責任的優先順序，以下是安養機構中的一天。

清早，艾瑟兒站在角落充電。電池充飽之後，它的「善行」（做有益的事）責任優於維護自己的責任，於是它開始在房間中走動，看看住客，問問有沒有什麼事需要幫忙，例如拿飲料或是傳口信給另一位住客等。當它接收到任務時，它會判斷這項任務中每一項責任的初始滿足度與違反度，例如一位看來很不舒服的住客要它幫忙找護士來，忽視住客的不適就會違反「惡行」（防範傷害）責任，使這項任務的優先性高於執行善行責任，因此它去找護士，通知她有位住客需要協助。這項任務結束後，它的善行責任優先權又提高，因此它繼續在房間中走動。

時鐘指向上午10點時，該提醒某位住客吃藥了。這項滿足善行責任的任務立刻成為第一優先，因此它找到這位住客，拿藥給他。後來，這位住客看電視入了神，可能是在看「超級星光大道」或者是「全民估價王」。由於沒有其他待履行的責任，而且電池電量也越來越低，艾瑟兒發現它越來越違反自己的責任，因此趕緊回到充電的角落。

機器倫理學的相關研究才剛剛起步。儘管還相當粗淺，但研究成果讓我們得以期盼，機器擬定的倫理原則可用於規範機器人的行為，使人類更能接受機器人。為機器人灌輸倫理原則相當重要，因為如果人類懷疑有智慧的機器人行為可能違反倫理，就可能完全拒絕具有自主能力的機器人，人工智慧的發展也將面臨威脅。

機器倫理學可能將逐步影響倫理學的研究。人工智慧相關研究的「實用」觀點，可能會比學院派倫理學家的抽象理論，更能掌握大眾心目中合乎倫理的行為。經過適當訓練的機器人，行為可能比許多人類更合乎倫理，因為機器人可以做出公正的決定，而人類在這方面往往並不擅長。說不定在與倫理機器人互動之後，還能督促我們自己的行為更加合於倫理。 SA

甘錫安專事科技類翻譯。

延伸閱讀

〈家家都有機器人〉，《科學人》2007年2月號。

〈機器人大戰時代來臨？〉，《科學人》2010年8月號。

IEEE Intelligent Systems. Special issue on machine ethics. July/August 2006.

Machine Ethics: Creating an Ethical Intelligent Agent. Michael Anderson and Susan Leigh Anderson in *AI Magazine*, Vol. 28, No. 4, pages 15–26; Winter 2007.

Moral Machines: Teaching Robots Right from Wrong. Colin Allen and Wendell Wallach. Oxford University Press, 2008.

對這篇文章有任何評論，請上網頁：www.ScientificAmerican.com/oct2010