國立清華大學科技管理學院生物倫理與法律研究中心

105 學年 下學期

學術午餐研討會

【現代社會與科技的倫理問題】

人工智慧的挑戰

機器人的倫理議題

講者姓名: 林宗德 教授

現任: 清華大學通識中心暨社會研究所

日期:2017年5月25日

時間:12:30pm-14:00pm

地點:國立清華大學名人堂

講座大綱

1. 前言
2. 機器人（Robots）是什麼？
3. 機器人倫理學（Robot Ethics, Robo-ethics）
4. 戰鬥機器人
    a. 我們是否應該使用機器人？
    b. 機器人該如何決斷？
5. Q&A

1. 前言

隨著人工智慧的研究不斷的突破，人工智慧已經越來越能夠應用在真實生活之中，而不僅僅只是科幻小說的情節。機器人、無人車等等技術的實踐指日可待。隨之而來，也可能產生倫理議題上的困難。本次講座很榮幸邀請到國立清華大學通識中心暨社會研究所林宗德教授為我們講授人工智慧的挑戰—機器人的倫理議題。

首先，林教授以日本的改編舞台劇——三姐妹（機器人版）為開頭，向我們說明機器人的倫理議題。在舞台劇中，常常發生的狀況就是忘詞。但專業的演員雖然會忘詞，還是可以即興的圓回來。然而，如果演員中有機器人，則機器人每句台詞、動作都已經設定好，其他真人演員就不能夠忘詞。這是一個很不一樣的地方。

三姐妹的劇情是：有一名機器人學家的女兒因故變成繭居族，足不出戶。機器人學家臨終前，做了一個與女兒一模一樣的機器人，取代女兒的社會功能。然而在最後一幕，機器人學家的故友聚會時，被機器人揭露出難堪的往事。 真人玉美想要遺忘過去不堪的往事，而 機器人玉美想要揭發丸山的惡行。玉美的家人既想保護欲美又想知道真相。丸山（機器人學家的故友）則想要在眾人之間要顧及面子。這樣不同立場、角色的衝突，若沒有機器人玉美，就不會發生。

究竟，我們希望機器人在我們的生活中扮演什麼樣的角色呢？林老師希望大家可以思考這個議題。

2. 機器人（Robots）是什麼？

依據韋氏字典，機器人被定義為：可以執行人類某些功能的自動器械。這樣的定義似乎有些籠統。而牛津字典，給予機器人四種定義：

1.外表以及動作像人的機器；
2.自動執行（重複）任務的機器；
3.自動控制的機器；
4.有效率但遲鈍的人。

當前的技術定義則是：

1.會在空間中移動，且知道自己身處何方
2.有（部分的）自主性，可以感測外部環境，經電腦運算做出反應
3.不一定具備人的外形

例如：在汽車工業等工廠中協助裝配，就是最常見、也是數量最多的機器人。

3. 機器人倫理學（Robot Ethics, Robo-ethics）

機器人倫理學是2000年代才出現的名詞，主要研究內容包含：

1.研究機器人的應用所產生的倫理，為應用倫理學的一支；
2.機器人在實際上要遵從什麼道德規則，才能減低人與機器人互動的風險？
（這樣的定義有論者認為有將機器人視為可能造成衝突或風險，而預先防範的用意包含在其中。但也有其他論者批評這樣的定義是意圖為機器人製造者卸責。因為弱將決策權力賦予機器人，研發者就可免於承擔責任。）
3.如果機器人與人類似，是否應該賦予機器人一定的道德地位？
（由於機器人技術的發展還尚未達到與人類如此相近的地步，因此，對於這個議題的討論仍在比較概念性的階段）

機器人倫理學包含許多面向，以自動駕駛車為例，可能就會包括：如何避免自動駕駛車造成司機失業的問題？；如何避免自動駕駛車造成的車禍？；車禍後的責任歸屬為何？等等的問題。

4. 戰鬥機器人：

a. 第一個問題——我們是否應該使用機器人？

若將機器人與戰鬥連結在一起，第一個面臨的問題即是：人類是否應該使用戰鬥機器人？反對者認為有三個理由支持人類不應該使用戰鬥機器人：

1.殺人是不得已的困難決定，不管武器怎麼樣先進，必須由人決定（man-in-the-loop）；

2.機器人可能無法區分戰鬥/非戰鬥人員，難以判斷誤殺時的責任歸屬；

3.若戰爭全面由機器人取代，則開戰的顧忌減少，更易發生戰爭。

也有一些人反對上述的看法，認為有以下三點理由，支持人類使用戰鬥機器人取代人類：

1.人是系統中最弱的一環，靠人類會來不及反應；

2.許多戰爭中虐待戰俘的事件都是因為戰時高壓環境造成的結果。而機器人沒有情緒，因此不會發生虐待敵方事件；

3.機器人誤殺的責任歸屬並非無法判斷，可以由其製造者或使用者承擔。

b. 第二個問題——機器人該如何決斷？

機器人三定律

將機器人與戰鬥連結在一起的第二個問題是，戰鬥機器人要如何決策？ 要怎麼樣去設計一套程式以控制戰鬥機器人。科幻小說家艾西莫夫曾經設計了一套機器人三定律，可以提供我們作為思考的參照：

第一定律：機器人不得傷害人類，或因不作為使人類受到傷害；

第二定律：除非違背第一法則，機器人必須服從人類的命令；

第三定律：在不違背第一及第二法則下，機器人必須保護自己。

第零定律：機器人不得傷害人類整體，或因不作為使人類整體受到傷害。

規則與脈絡

在小說中，機器人只需要依據少數幾個定律就可以行動，然而，在現實生活中，若要機器人完成複雜的行動，會需要龐大的規則才有可能。這會產生Frame Problem，也就是規則與脈絡的問題。

舉例而言，若要求機器人去拿個寶物，同時在拿到寶物的時候引爆炸彈，機器人會不知道要怎麼辦。因為電腦無法知道什麼資訊是採取行動所必要的。現實生活中有太多條件，包括：室溫、距離、空間中其他各式各樣的物品等等。這會使得機器人不知道哪些條件重要，哪些不重要，當要考慮的事情太多時，就會無法運作。

然而，為什麼我們平常在活動的時候不會有如同機器人一樣的規則問題呢？因為人類具有常識（Common-sense knowledge），可以將一個情境下的脈絡一併理解，而不會產生機器人的困難。

因此，在討論機器人問題時，需要注意到Frame Problem。

義務論與後果論

此外，在設定機器人的決策機制時，主要使用的倫理學理論有二：義務論（Deontology）及 後果論（consequentialism）。

義務論認為：行動受決斷約束，有些永遠不應做的事。例如：不說謊、不能把人只當成工具。行動後果的好壞，不是決定行動對錯的唯一判准，必須要考慮行動的意圖和動機。舉例而言：二戰時的德國人，藏匿猶太人，當被詢問是否有藏匿時，是否應該說實話？義務論的問題是，縱使壞的意圖能夠帶來好的後果也不應做；而好的意圖縱然會帶來壞的後果，則是可以接受的行動。

後果論則主張：決定行動對錯的唯一判准，是行動後果的好壞。讓世界產生最多的善與最小的惡。而後果論的其中一種版本就是效益主義（Utilitarianism）。效益主義認為：善就是快樂，惡就是痛苦。愉悅、幸福本質上是好的，而痛苦本質上是惡的。因此，效益主義主張：最大幸福原則（或稱為效益原則）。這個原則強調「為最大數量的人們產生最大量幸福的行動，就是道德上該做的行動。」

電車難題

接者，林教授提及電車難題的例子：一台火車即將撞上鐵軌前方的五個人，你就在轉轍器旁邊，若拉動轉轍器，則可以使火車轉向，撞向另一條鐵軌上的一個人。林教授詢問台下聽眾，是拉還是不拉呢？有人回答：會拉，因為這樣死的人比較少，五條人命大於一條人命。

林教授就換了另一個例子：如果今天火車即將撞上五個人，而你可以將你身邊的胖子推向鐵軌，火車將會撞上胖子，但也會因此停下來，是推還是不推呢？這個時候，認爲應該要推的人就減少了很多。老師透過這個例子，向聽眾說明生活中各種情境每個人都可能會有不同的決斷，要據此設計出一套機制供機器人決策，十分困難。

倫理原則是否可以計算？林教授說，依據後果論的主張：一個行為道德上的對錯取決於行為後果的好壞。也就是說：道德對的行動就是有好的結果的行動，可以產生最大的幸福總量。因此，並非不可能透過計算受影響的人數、愉悅的強度和持續時間、和行動產生該愉悅的機率，進行計算。而機器人的好處就是比人精確且不會偏袒。越是複雜的、影響深遠的行動，機器人計算得就比人還要快。

自動駕駛車

最後，林教授回到自動駕駛車的問題。他說電車難題可以有無數種版本。例如：駕駛車是否應該為了避免撞十個人，而去撞一個人呢？這個人的身份會不會影響到壯語不撞的決定？婦女、小孩、老人或是罪犯，都可能會影響到我們的決定；駕駛車是否應該為了避免撞十個人，而去撞牆？；駕駛車是否應該為了避免撞一個人，而去撞一個人？每一個情境，都考驗我們的決定。

而一個最實際的問題是：沒有人願意去買不保護乘客的自駕車。林教授說，實驗發現：人們通常都會抱持著行人觀點，而希望自駕車能夠保護路人。然而當問到是否願意買不會保護乘客的自駕車時，大多數人並不會想要買這種車。

目前一般的汽車駕駛在遭遇緊急情況的決定是依據反射動作，非有意識先決定規則。因此，不論是自撞或是撞人，我們很難對於其反應做出太多好壞的評價。然而若使用自動駕駛，事前選擇決策方案，就是有意識地決定要讓別人受傷。

從自動駕駛車的例子我們可以發覺，將機器人應用在生活中存在許多困難與挑戰。如何解決這樣的問題，需要我們用更多的思考與智慧來解決。

5. Q&A

問：玩命關頭中，駭客駭入自駕車系統是否可能？
老師回應：應該會有，因為自駕車需要靠龐大的物聯網才能夠可能。有網路自然就會有駭客的風險。所以在自駕車問題中存在兩種風險：現實生活中傷亡的風險以及系統被入侵的風險。

問：假如賦予機器人社會人格，而進入社會生活，那是否也擁有人權？
老師回應：首先，目前的技術離這樣的階段還很遠。機器人並沒有意識，也不在意被如何被對待，只是程式的運作。那所謂的社會人格，也僅僅是我們情感的投射。因此，現階段賦予機器人權利對於機器人似乎並沒有意義。

問：具體而言我們應該怎麼樣去設計程式？
老師回應：目前而言，至少有三種可能的選項：1.交由個人選擇、2.交由企業選擇，各車廠不同、3.統一的決策程式。

問：沙盒（sandbox）機制提供一個小範圍的實驗場所，是否可以做為一個過渡的機制？
老師回應：有可能。但有許多困難，例如是否有區域願意預先接受這樣的風險。