《歐美研究》第五十卷第二期 (民國一〇九年六月)，179-205
DOI: 10.7015/JEAS.202006_50(2).0003
https://euramerica.org

反思機器人的道德擬人主義*

何宗興
國立中正大學哲學系
E-mail: tsunghsing@ccu.edu.tw

摘 要

如果機器人的發展要能如科幻想像一般，在沒有人類監督下自動地工作，就必須確定機器人不會做出道德上錯誤的行為。根據行為主義式的道德主體觀，若就外顯行為來看，機器人在道德上的表現跟人類一般，機器人就可被視為道德主體。從這很自然地引伸出機器人的道德擬人主義：凡適用於人類的道德規則就適用於機器人。我反對道德擬人主義，藉由史特勞森對於人際關係與反應態度的洞見，並以家長主義行為為例，我論述由於機器人缺乏人格性，無法參與人際關係，因此在關於家長主義行為上，機器人應該比人類受到更嚴格的限制。

關鍵詞：機器人、人工智能、道德主體、道德擬人主義、反應態度

投稿日期：108.6.28；接受刊登日期：109.2.9；最後修訂日期：109.2.12
責任校對：張文綺、趙麗婷、范馨文
* 本文是科技部人文社會科學研究中心的計畫成果 (107-2420-H-002-007-MY3-V10701)。另外，我十分感謝兩位匿名審查人，以及以下學者給本研究的建議與提問：謝世民、王一奇、侯維之、蔡政宏、王華、趙之振、嚴如玉、林映彤、陳樂知、張智皓、朱家安。

隨著人工智能技術的發展，科幻電影中機器人成為人類生活中有力助手的世界，似乎不再是白日夢的幻想。如果機器人能取代人類去做我們所不願意從事的工作，想必是許多人所樂見的。然而，機器人的發展面臨一個兩難：即要提高機器人的效率，就需要提高它自動化的程度，去除人類的控制與干擾；然而，機器人的自動化程度越高，則對機器人是否會威脅傷害我們的擔憂程度自然也越高。因此，要讓自動化機器人 (autonomous robots) 應用在真實世界中，勢必得確定機器人的行為符合我們的道德標準，不至於為惡。如果機器人可以遵守道德標準而行事，因此它不需要人類的監督便可以自主的做事，在文獻中，這樣的機器人被稱為「人工道德主體」(artificial moral agent) (Allen, Varner, & Zinser, 2000; Floridi & Sanders, 2004; Fossa, 2018; Grodzinsky, Miller, & Wolf, 2008; Himma, 2009; Laukyte, 2017; Torrance, 2012; van Wynsberghe & Robbins, 2019; Weber, 2013)。人工道德主體甚至不像人類會時常違反道德，反而一定會遵守道德。似乎，人工道德主體是比人類更完美的道德主體。

如果機器人成為跟人類一樣好 (甚至更好) 的道德主體，是不是就代表我們可以全面地使用機器人來代替人類呢？暫且先拋開這樣做對社會帶來的負面衝擊 (如失業等)，有人可能會認為原則上沒有問題，畢竟既然機器人可以跟人類一樣好地遵守道德標準，那麼道德上允許人去做的，原則上機器人也可以去做。讓我把這個主張稱之為「道德擬人主義」(moral anthropomorphism)，也就是將機器人看做與人類無本質上差別的道德主體，因此，所有對人類有效的道德理由與原則，都一體適用於機器人。

在本文，我將批評機器人的道德擬人主義。我的批評是基於彼得・史特勞森 (Peter F. Strawson) (1974) 著名的論文〈自由與怨恨〉

(“Freedom and Resentment”) 中的洞見。藉由他的洞見，在第貳、參、肆節，我將論證在涉及到干涉人的生命自主權的行為上，機器人應該受到比人類更大的限制。在這之前，第壹節中我先說明在什麼意義下機器人可以被視為道德主體，以及為何道德擬人主義吸引人。

在論述前，我想先提出幾點澄清並限定本文討論之範圍。首先，有人可能會質疑人工道德主體不可能實現，因此本文的討論沒有意義。對此，我的回覆是，人工道德主體這一概念不是我提出的，我的批評對象是文獻中討論機器人倫理學可見到的擬人主義傾向。並且，人工道德主體並非是邏輯上不可能的。的確，對人工道德主體的想像在科幻小說與電影中隨處可見，對它做反思與探究是有哲學意義的。如我要在本文論證的，這討論讓我們瞭解傳統倫理學忽視的面向。另外，即便未來科技無法實現全面的人工道德主體，現實上還是需要製造出能在特定領域使用的自動化機器人，如自動駕駛車輛或是照護機器人，那麼本文仍可以提醒在設計這些自動化機器人時，應該避免怎樣的擬人化。

再者，以下我所討論的機器人都不具有「人格性」(personhood)。怎樣算具有人格性是困難的哲學議題，我在此預設具有人格性，應當具有欲望、情感、感覺、自我意識，能規劃與期待自己人生的未來。並且，我假設不具人格性的機器人，仍可以成為人工道德主體。在科幻小說與電影常可見到這樣的機器人設定。

之所以如此設定的原因是，我的目的是批評道德擬人主義，而我的批評不適用於具有人格性的機器人。但我也反對製造具有人格性的機器人，由於這非本文主旨，在此只能略述我的理由：因為，一旦機器人具有人格性，根據定義，它就能規劃自己的生活，這樣就背離製造機器人是為我們工作的目的。此外，我們也不該強迫它

為我們人類工作，因為既然它具有人格性，就應當同人類一般享有人權。我們不需要製造具有人格性的機器人 (這個世界並不缺人)，這可能會帶來顯著的壞處 (消耗能源、可能為惡)。就自利而言，我們不該製造它們。有人可能會批評這是物種主義 (speciesism)，但這並非是壞的物種主義，因為這沒有傷害已經存在的具有人格性的機器人，而只是不製造它們而已，就如同不生育後代不是道德上的罪惡。

壹、人工道德主體與道德擬人主義

有些學者 (Brożek & Janik, 2019; Himma, 2009) 認為，作為道德主體需要能為自身的行為負責，這通常要求行為者具有某些心理性質，例如意識、意圖、欲望等特徵。暫且稱這個主張為「心理主義的道德主體觀」。根據心理主義，機器人自然不該被視為道德主體。

然而，這裡討論機器人是否作為道德主體的脈絡是機器人何時可以自動化，心理主義在這個脈絡下並不適用。因為，為了讓機器人能在不受人類監督之下自主的行動，只需要求機器人不會做出錯誤的行為，而不需要它具有心理主義所要求的心理性質。換言之，我們需要的是「行為主義的道德主體觀」。

也許有人會質疑，如果沒有心靈認知道德法則，或是情感體會羞恥或罪惡感，如何可能是道德主體呢？這個質疑可以用著名的圖靈測試 (Turing Test) 來回應。[1] 艾倫・圖靈 (Alan Turing) 為了判定機器是否能思考，提議如下的測試：如果機器能夠在人類不知情的情況下進行對話，且不被發現是機器，那這台機器就可以被判定

[1] 我感謝審查人提出這個問題。

為會思考。雖然說思考是一種心理性質，但是如何定義思考，以及如何直接觀察心理活動十分困難。圖靈測試跳過這些困難，改以外顯的行為表現來判定。既然我們通常認為語言表達是思考的證據，而且判定某人是否會思考的方式也是透過對話的方式進行，那麼圖靈測試主張，測試機器是否會思考也可以用對話進行，無視機器是否真的有思考的心理活動的問題。

也常有人會質疑即使機器通過圖靈測試，也不代表機器會思考，因為它只是會製造出一連串從人類眼光看似有意義的話語，但它也許沒有掌握語言的意義。但假如我們只是想製造出能跟人類溝通無礙的機器人，那麼我們毋須討論機器本身是否能掌握語言的意義。

同樣地，也有學者 (Allen et al., 2000) 提出道德版圖靈測試 (Moral Turing Test)。為了決定機器人是否是道德主體，我們可以觀察機器人的舉止是否符合人類的道德標準，也就是它會去做出人類認為道德上正確的事，且不會做出人類認為道德上錯誤的事。例如，家事機器人奉主人之令去超市採買，它不會沿路推開擋到它的人，或是在超市拿了東西就走。我們毋須討論它是否認知到道德準則，只要它的所作所為符合人類的道德評判。道德圖靈測試一樣是以人類作為道德主體的標準，既然人類是我們已知唯一的道德主體，這測試一樣是以人類的的道德判斷作為道德主體的判準。如果機器人能通過道德版圖靈測試，則可被視為是道德主體。

道德圖靈測試會支持道德擬人主義。這是因為道德圖靈測試不管行為者自身的心理狀態 (如果他有心理的話)，只管其外顯的行為表現，這表示人類與機器人在作為道德主體上是沒有本質上的差異 (Floridi & Sanders, 2004; Fossa, 2018; Grodzinsky et al., 2008; Gunkel, 2012; Wallach & Allen, 2008)。而且，在道德圖靈測試中，

我們是從人類的道德標準來判斷機器人是否通過了測驗。所以，當評估是否可以讓機器人自主地工作時，我們就是用人類的道德標準來檢視，例如，它是否會為了完成工作指令而傷害了不該傷害的人或物品。既然如此，很自然的想法便是在設計自動化機器人時，能讓它的行為盡量符合人類的道德標準，這種思路很自然會走向道德擬人主義，亦即拿人類的道德標準直接應用在機器人上。

另外，就算不接受道德圖靈測試這樣的行為主義式的道德主體觀，也有學者 (Brożek & Janik, 2019) 指出，傳統規範倫理學的兩大進路──效益主義與康德主義──強調道德主體需要遵循理性法則，可能會得出機器人反而是比人類更完美的道德主體的結論，因為比起人類，機器人更能完美的遵循理性。例如，康德主義主張道德法則必須是可以普遍化的理性原則，而效益主義通常主張道德標準只由行為結果的好壞來決定。這兩種進路都主張道德標準是經由理性的思辨或計算來決定，並且普遍適用於所有的道德主體。所以，只要機器人能夠計算出哪些行為會符合道德標準，且依循標準行事，[2] 那根據這兩個進路，機器人也該被視為是道德主體，並且依循著跟人類一樣的道德標準，因此支持了道德擬人主義。

貳、人類限定理由：概述道德擬人主義的反駁

道德擬人主義認為就作為道德主體而言，人類與機器人本質上並無差異，因此，道德理由與原則一體適用於兩者上。我反對道德擬人主義，因為有些道德理由僅適用於人類上，人類可以利用這些

[2] 需要強調的是，這裡不一定要求機器人本身認識到道德標準，也可以只是機器人能夠判斷在什麼情況下採取什麼樣的行為，而它的行為從第三人的角度來看符合道德標準。

理由來證成其行為，而機器人不行。我將這類型的理由，稱之為「人類限定理由」(human-relative reason)。如果人類限定理由存在，就蘊含在某些情況下，有些行為是人類可以做的，但機器人不行。這代表道德擬人主義是錯的。

在這一節中，我會先給一個道德擬人主義的反例，在之後一節給予這個反例理論上的支持。但在這之前，我將先說明「理由」這一概念。[3]

「理由」是目前哲學家在討論道德哲學時十分常用的概念。當哲學家使用「理由」時，一般來說都接受這個原則：「甲有理由去做 x，若且唯若，在其他條件不變下，甲應該做 x」。因此，理由的功能是用來證成或解釋某個道德判斷。值得注意的是，這裡的理由是所謂的「初步理由」(pro tanto reason)，而不是「整體理由」(all-things-considered reason)。整體理由是權衡考量所有初步理由之後的結果，因為初步理由間可能會互相衝突。例如，假設我為了赴約，正在趕火車的路上看見了有人流血昏倒在地上，這時候我有一個初步理由支持「(在其他條件不變下) 我應該停下腳步來救他」，但我也有另一個初步理由支持「(在其他條件不變下) 我不應該停下腳步以免遲到」，這兩個初步理由相互衝突。經權衡之後，我可能認為第一個理由更為重要，因此得出了整體理由，支持「整體而言，我應該停下腳步來救他」。除非特別說明，本文所討論的理由都是指初步理由。

因此，當我說有些理由是人類限定的，所以它們無法用來證成機器人之行為，這不代表機器人的行為是無法被證成的。這是因為，

[3] 嚴格說來，本文討論的「理由」是哲學家所稱的「規範理由」(normative reason)，而非所謂的「動機理由」(motivational reason)。後者是關於行為者之所以採取該行為的動機，前者是關於規範判斷。

人類限定理由也是初步理由，有可能有其他初步理由可以證成機器人的行為。然而，如果人類限定理由真的存在，那就表示在有些情況下，人的行為比起機器人更容易被證成，這就足以推翻道德擬人主義。

為了說明人類限定理由，讓我們考慮以下的例子。

> 〈自殺〉：老王身患慢性疾病，這個疾病讓他十分不舒服，但始終無法治好。經過深思熟慮之後，老王決定自殺。正當老王要自殺時，剛好小美／機器人經過，小美／機器人與老王互不認識。當她／它想要阻止老王自殺時，老王表示他心意已決，請求不要阻止他自殺。

在〈自殺〉中，到底小美與機器人應不應該阻止老王自殺呢？老王非常清楚地表達他自殺的意願，我們甚至可以設想，在經過了深談之後，老王的想法仍然沒有動搖。所以，如果小美（機器人）阻止老王自殺，這是一種家長主義（paternalism）的行為。所謂家長主義的行為，指的是出於為對方的福祉著想而干預對方意願的行為。由於阻止老王自殺雖然是拯救他的生命，但卻違反他的意願，因此，這是家長主義行為。

接下來的問題是，阻止老王自殺是不是道德上許可的，以及是小美或是機器人救他有沒有道德上的差異？針對第一個問題，直覺上顯然是許可的。雖然說這是一種家長主義的行為，但有些家長主義行為的確是許可的。一般說來，涉及到家長主義的道德理由通常分成兩類：一是支持理由，因為對當事者帶來好處；另一則是反對理由，因為這侵害了當事者的自主權（autonomy）。簡單地說，當支持理由勝過反對理由時，家長主義是可以獲得證成的。由於直覺上

拯救性命比起自主權的侵犯更為重要，阻止老王自殺顯然是道德上許可的。

關於第二個問題，傳統上在討論類似像〈自殺〉這類型案例時，沒有考慮到拯救者是機器人的可能性。那麼是人類或是機器人阻止老王會產生什麼道德差異呢？上述的支持理由和反對理由同樣可以用來證成機器人去侵犯老王的自主權，如果沒有其他理由需要考量，那麼換成是機器人去救老王是沒有道德差異的。

然而，我認為的確還存在其他道德理由需要考量，而這理由是人類限定的，這點不同造成了在〈自殺〉中，是小美還是機器人來阻止老王是有道德差異的。為了說明這點差異，讓我用另一個較為生活化的家長主義行為的例子。

> 〈看牙〉：小明蛀牙很嚴重，他的父親帶他去看牙醫。但小明十分害怕看牙醫，哭鬧不肯坐上牙科椅。在勸說無效之下，他的父親強壓他在椅子上讓牙醫做完治療。

暫且假設小明父親的行為是道德上許可的，畢竟雖然違反了小明的意願 (尤其考慮到小明仍是兒童)，但接受治療的好處勝過強迫他的壞處。然而，如果假設強迫小明坐上牙科椅的不是他的父親，而是現場小明同學的母親 (小明父親並未請求或同意她這麼做)，這還是道德上許可的嗎？與〈自殺〉的例子相同，對當事人的好處與違反當事人意願這兩個正反理由，仍可以用來支持與反對小明父親與小明同學母親的行為。因此，如果我們認為前者可以但後者不行強迫小明，一定有其他的理由來解釋兩者的道德差異。

常識上兩者的差異在於，小明同學的母親缺乏小明與他父親之間的親子關係，因此她不能在沒有小明父親的同意下，強迫小明接受治療。這種建立在當事人雙方之間關係的理由，被稱之為「主體

限定理由」(agent-relative reason)。上述提到涉及到當事者好處與壞處的理由通常被稱之為「主體中立理由」(agent-neutral reason)。小明同學母親與小明父親都擁有主體中立理由，但只有小明父親擁有主體限定理由。正是由於這個差異，導致了〈看牙〉中小明父親可以，但同學母親不可以強迫小明治療。

同樣的道理，人類限定理由是一種主體限定理由，但這是一種很特別的主體限定理由，因為它適用於所有人類上。[4] 如果人類限定理由存在，這應當是 AI 倫理學中對於倫理學重要的貢獻。因為，傳統上倫理學的討論，通常將主體中立理由視作為適用於所有人類的理由，而主體限定理由只適用於那些處於個別人際關係的主體上。然而，由於自動化機器人的想像，AI 倫理學迫使我們去思考作為道德主體的機器人之道德意義，人類限定理由就是我認為一點重要的差異。

由於人類限定理由是一種主體限定理由，而主體限定理由是源於個別主體之間的關係上，所以人類限定理由就是源自於人類之間的關係之上，而這個關係是人類與機器人之間缺乏的。在下一節，我將說明這是什麼樣的關係，以及這樣的關係如何產生人類限定理由，並回來討論〈自殺〉這個案例，說明為何道德擬人主義是錯的。

參、史特勞森論人際關係與反應態度

在這一節中，我將應用史特勞森著名的論文〈自由與怨恨〉(Strawson, 1974) 來說明人類限定理由。

[4] 嚴格說起來，應當是適用於所有作為道德主體的人類上，因為有些人類（暫時）不是道德主體，那人類限定理由就不適用於這些人。

在〈自由與怨恨〉中，史特勞森並未討論人類限定理由，但我將說明可以從他的思想中梳理出這個概念。史特勞森主要的關切在於自由意志與道德責任的關係上，傳統的看法認為要回答人是否要負道德責任，必須先處理自由意志與道德責任之間關係的形上學問題。史特勞森認為，傳統的看法過度理論化這個問題，我們對於道德責任的歸屬不依賴自由意志的預設，而是基於我們的人際關係以及與人際關係相關的情感。史特勞森如此陳述他的觀察：

> 我們應當考慮我們跟其他人可能擁有的許多不同種類的關係—如擁有共同興趣的同好、家庭成員、同事、朋友、愛人，在無數種可能情況下的偶然遭遇。然後我們應該一一考慮，我們如何看待在這些關係中對方對待我們的態度與意圖，以及對此我們可能會經歷的反應態度 (reactive attitudes)。一般而言，我們要求那些與我們處於某種關係的人對我們有某種程度的善意與尊重，即便要求的內容會隨著關係的不同而可能有強烈地變化。且針對對方之善意、惡意或兩者之缺乏，我們的反應態度的範圍與強度也會跟著強烈地變化。(Strawson, 1974: 6-7)

史特勞森認為，在面對他人時，我們預設了他所稱的「參與者反應態度」(participant reactive attitude)，即我們將他人視作與我們一起參與在某種人際關係的成員，並且對他人的態度與舉止有所要求與期待，當他人滿足或不滿足我們的要求與期待時，我們自然地有一些情感的反應，如感恩、怨恨等。史特勞森稱這些情感為「反應態度」。

Paul Snowdon 與 Anil Gomes 對於參與者反應態度的解說值得參考：

> 從參與者態度來看，我們將他人視作為反應態度的恰當對象。反應態度包含感恩、憤怒、同情與憎恨等，這些態度預設了他

> 人需負起責任……，尤其，史特勞森主張，我們對待他人與自己的反應態度是自然且無法取消的。這是我們作為人類的一項核心要件。(2019)

因此，參與者反應態度是我們面對他人的自然預設態度，只要那些人可以恰當地被視為是能對其行為負責，是反應態度的恰當對象。

與參與者反應態度相對，史特勞森指出，我們可以暫時擱置參與者態度，改採取所謂的「客觀態度」(objective attitude) (Strawson, 1974: 9)。例如，當我們發現對方暫時或永久無法正常行為，無法為其行為負責，像喝醉酒或是失智的人；又或是我們可能把對方當作研究對象。當採取客觀態度時，我們就不會對他人沒有滿足參與者態度之要求感到相應的反應態度。然而，史特勞森認為，要長期採取客觀態度是非常困難的，參與者反應態度是我們面向彼此的預設態度。

因此，除非有理由對他人採取客觀態度，我們對他人的態度是預設了參與者反應態度。當採取參與者反應態度時，我們會將他人視作某種人際關係的成員。這人際關係的範圍十分廣泛，可以是我們熟悉的家人、朋友、同事關係，但也可以擴及到偶遇的陌生人上。雖然說我們通常不認為陌生人跟我們有什麼關係，但我們時常也會談論某個社會的人情是溫暖還是淡薄，從史特勞森的觀點來說，我們會抱怨人情淡薄就是因為我們對他人——即便是陌生人——採取了參與者態度，我們期待人際之間應當表現出一定程度的善意與關懷。而人情淡薄的社會就是人際之間對於彼此的善意與關懷不足，而我們對人情淡薄的抱怨，就是一種反應態度，而且可以是恰當的。

讓我用以下例子來說明。想像你從圖書館借書出來，手上抱著一堆書，不小心拐了一跤，書灑了滿地。圖書館門前有許多人經過

你身邊，但卻沒有人停下腳步來幫助你或是表達關切，很自然地你會對這些人的冷漠感到怨懟。這個例子充分說明史特勞森的想法。雖然你跟那些路人並不相識，他們也沒有義務要幫助你，但很自然地，你會對他們的冷漠感到怨懟。對史特勞森而言，這是因為人與人之間預設的態度是參與者反應態度。在上述例子中，你對這些路人預設了參與者反應態度，你和這些路人都參與在人際關係之中，在這個人際關係的參與者都應該對其他人表達一定程度的善意與尊重。由於這些路人沒有滿足這些要求，因此你的怨懟是自然且正當的。

史特勞森的想法可以分成兩個面向：心理的與規範的。在心理面向上，史特勞森認為參與者反應態度是我們預設的態度。值得注意的是，說參與者反應態度是預設的，不是說當事人有意識的選擇，而是指我們在面向他人時不經意識選擇的預設態度。因此在上述例子中，你對那些路人採取參與者反應態度不是你刻意選擇，而是你面向世界的預設態度，因此你對路人的怨懟是立即且自然的。

史特勞森不僅認為參與者反應態度在心理上是自然的，他還主張在規範面向上，參與者反應態度產生規範要求。即當採取參與者態度時，我們會對參與者的態度與舉止有所要求與期待，即要求與期待他們應對我們表示善意與尊重 (善意與尊重的程度與範圍自然隨著關係深淺不同而有所變化)，而當對方滿足或未能滿足這些要求與期待時，我們自然會產生相應的情感 (反應態度)，而這些情感可以是正當、有證成的。例如，在上述例子中，你對經過的路人的冷漠感到怨懟是有證成的，因為這些路人未能表達適度的善意。當然，路人所應當表達的善意以及你的怨懟是有限度的，比如說你不該期待每個路人都停下腳步來關懷你 (你可以合理相信有些人可能有重要的事)，且你的怨懟也不能過度。只要你的怨懟的確符合了對方所

表現的態度，而且在合理程度範圍內，你的怨懟就具有正當性。

總言之，關於參與者反應態度，史特勞森有兩個主張：[5]

一、規範主張：參與者反應態度會對參與者的態度與舉止有所要求與期待 (這些要求與期待會隨著關係不同而變化)，當這些要求與期待被滿足或不滿足時，當事人會產生相應的反應情感，只要這些情感在合理程度內，便是正當、有證成的。

二、心理主張：參與者反應態度是我們面向他人自然預設的態度，對參與者的規範要求與期待，以及相應產生的反應情感也都是自然的。

當瞭解了史特勞森的主張之後，我們很容易推論出人類限定理由，其主張如下：

【理由】行為主體處於某種人際關係這個事實，提供相關行為主體理由對對方表達適度的善意。

【理由】其實就是上述規範要求的另一種表達方式，把原先的期待與要求換成是理由來表述。這理由是人類限定的，因為只有人類處於人際關係的網路中，才是參與者反應態度的適用對象。[6] 由於機器人根據定義不具有人格性，缺乏情感等反應態度，因此，無論機器人在外顯行為上與人類多麼相似，它都沒有真正地參與在人際關係中，所以就得不到這種人類限定理由。

瞭解人類限定理由，我們可以回去談像〈自殺〉這類的家長主義行為。【理由】要求我們應該對人表達適度的善意，這是人類限定理由。人類限定理由的存在，可以更進一步說明為何有些家長主

---

[5] 關於參與者反應態度的這兩個面向，可以參考 Watson (2014)。

[6] 嚴格說來，即便不是人類，但若是具有人格性的道德主體，即擁有情感、意志、行為能力的存在者，應可參與在人際關係中。那人類限定理由也可以適用在這些存在者。為了簡化，本文只限制在討論人類與 (不具人格性的) 機器人上。

義行為是許可的。家長主義行為違反當事者意願，理應不被許可。但如上所述，有些家長主義行為是許可的，因為這對當事者帶來好處。然而，人類限定理由提供另一個證成家長主義行為的理由，因為人際關係理由要求行為者對當事者表達善意，這理由可以證成行為者適度表達善意的行動，雖然有些情況下這些行動可以干涉了當事者的意願。

換比較生活化的方式來說明，日常生活中時常會碰上一些比較熱心的人，即便我們表達不需要他們的幫助，他們還是會堅持。例如，你有幾箱行李要提下樓梯，一個路人經過了說要幫你拿，你表達你自己搬就行了，但他還是堅持幫你搬了。雖然平常我們可能不如此認為，但這個行為滿足了家長主義行為的定義：他的行為干涉了你的自主權。但多數人應該不會認為他的行為是道德上錯誤的。更重要的是，之所以他的行為不算錯，不只是因為這個結果對你來說是好的，[7] 更因為是他的行為是善意的，且他所表達的行為就日常的標準算是適度的，不算是過度干涉。

史特勞森的洞見傳達了一個我們熟悉的日常事實，即我們接受他人適度地干涉我們的生活，因為大家都是人際關係的參與者，我們期望，並也接受他人善意的表達。修改上述例子可以更凸顯這個事實，假設你自己清楚不想要他人幫忙搬行李，然而你有許多行李需要搬，你搬上搬下流得滿身大汗，許多人 (有些甚至是你認識的) 經過你身邊，但沒有人問你需不需要幫忙，你可能會感嘆現在人情冷漠，而你的感嘆是合理的。

另一方面說，當我們允許他人適度表達善意關懷而干涉我們的生活，這往往是進一步發展彼此人際關係的契機。有些時候，我們

[7] 我們甚至可以假定他的行為對你沒有好處，因為你把搬行李當作是你今天運動的項目，他幫你搬了之後你還得另外運動補上。

回絕旁人的關心或幫助而對方堅持幫忙，雖然當下可能會覺得厭煩，但事後可能是友誼的開始或進一步發展。試想，假如說事後我們後悔了，我們會更加重視他的友誼，因為他當初願意承受冒犯的可能性而堅持關懷我們。如果說只是因為我們表達不願被干擾而旁人就不該表達其善意或關懷，那就會喪失許多珍貴的人際關係發展的可能性。

由上所述，從史特勞森所稱的參與者反應態度，可以推衍出人類限定理由，只適用於人際關係中的參與者，證成人可以適度地對他人表達善意與關懷，即便這會干涉到他人的自主性。

人類限定理由的存在，可以說明在〈自殺〉中人類與機器人的道德差異。在〈自殺〉中，「拯救老王生命」是小美與機器人都享有的支持理由，但只有小美享有人類限定理由來干涉老王的自主權。這並非說機器人就一定不能夠阻止老王自殺，也許拯救性命這個理由已經強到足以證成。但在其他例子中，人際關係理由的有無，可能就會造成差異。理論上，既然人類擁有證成家長主義行為的理由比機器人多，就蘊含比起機器人來說，人類對其他人類的家長主義行為比較容易獲得證成。

試以比〈自殺〉輕微的例子來說明。這次老王不是要自殺，而是一個人坐在公園裡抱頭痛哭。小美經過看到老王便前去安慰他，但老王不領情要小美走開。假想小美沒有離開，而是試著用更婉轉的方式繼續試著安撫老王。雖然小美違反了老王的意願，直覺上這應該是允許的。然而，如果換做是機器人，我認為它也許可以通知警察或生命線，代為聯絡其家人，但它不應該違背老王的意願，繼續在現場試著安慰他。

之所以如此，是因為兩者有一項重大的差異。小美的堅持表現的是她的關心，流露的是真實的情感，這提供了額外的理由來支持

小美對老王的干涉。然而，機器人只是表現出看似關心的行為，背後沒有任何心意支撐。在目前的生活中，我們已經遭遇到類似的情況。例如，很多人都抱怨許多公司的客服專線都改用語音而非真人來服務。除了跟語音對話常常難以獲得所需要的服務外，客戶也覺得語音的服務缺乏真人的溫暖，因此感覺公司根本不重視客戶服務。由於心意有無的這點差異，會使得在某些情況下，機器人對人類的家長主義式的行為更難以獲得證成。

有人可能會質疑，有些人會跟他的機器人建立起朋友甚至是家人的關係。例如，星際大戰中的天行者路克與機器人，就如同現在有些人會把他的寵物當作自己的小孩。這樣的話，在關於它的主人的家長主義行為，這些機器人也可以享有人類限定理由，甚至它們享有的理由比起一般的陌生人來得更強。[8]

關於這一質疑，我同意人類可以與非人類的生物，甚至是無生命的物體建立起某種親密的關係。這一點並不影響我的論點，因為我的論點是針對一般的人類與機器人而言。我論述，同樣是互不相識的，人類與人類已經預設了因參與者反應態度所建立起的人際關係，這點是機器人所缺乏的。這點差異表現出機器人無法在一開始就享有人類所擁有的人類限定理由，已經構成一項重大的道德差異來反對道德擬人主義。

當然，人類可以選擇與他的機器人建立某種親密關係 (例如，親人、愛人、朋友等)，因此他可以允許他的機器人干涉他的生活。但是，嚴格說來，由於機器人缺乏反應態度，這樣的關係並非是真正的人際關係，這樣的關係不會產生人類限定理由。更重要的是，既然他「允許」機器人的干涉，這就不是家長主義行為，因為定義

[8] 感謝審查人提出這個問題。

上，家長主義行為是違背了當事人的自主權。因此，根據上述分析，我建議，在設計機器人時，不應該一開始就預設機器人能對其主人採取家長主義式的行為；而應該是，機器人在出廠時應該預設成不能對人類採取家長主義行為，必須由它的主人決定之後再選擇是否啟動（以及何時關閉）這項功能。[9]

另外，也許有人會質疑，既然我預設行為主義的道德主體觀，為何在討論道德理由時還納入心理性質？然而，這兩者沒有衝突。首先，我並非主張行為主義才是唯一正確的道德主體觀。我的主張僅僅是，當稱呼機器人是道德主體時，是在行為主義的道德主體觀上理解的。更重要的是，行為主義觀討論的是道德主體（moral agency），而不是道德理由。即便我們預設行為主義的道德主體觀，我們仍然可以討論涉及心靈事實所產生的理由。以〈自殺〉為例，機器人是道德主體，所以我們可以思考它該如何處理老王的自殺，也就是思考相關的理由。這些理由包括，「阻止老王可以拯救他的生命」、「阻止老王違反了他的自主權」，這兩個理由都與機器人的心理性質無關。然而，「機器人缺乏參與者反應態度」這個事實與機器人的（缺乏）心理性質有關，而這個事實使得機器人無法獲得人類限定理由，這個道德事實獨立於機器人是在行為主義的意義下作為道德主體的事實，兩者沒有衝突。[10]

---

9 另一個類似的質疑是，當我們購買特殊功能的機器人，例如養老院購買照護機器人來照顧年長者，就應該預設這些機器人可以干涉年長者。然而，我這裡的討論是沒有預設機器人有任何特殊的功能或角色。如果購買的是特殊功能的機器人，我們可以說在購買時已經默許了接受這些機器人在這些方面干涉我們的自主。然而，在這些功能之外，我認為應該還是要讓人選擇是否要讓這些機器人干涉。(我感謝審查人提出這個疑問)

10 我感謝審查人提出這個意見。

另一個可能的誤解是，本文的討論並非是上述例子中機器人行為的對錯，而是機器人缺乏人類限定理由這個道德事實。在討論上述例子中，對於機器人行為的對錯，讀者也許會跟我的判斷不同。但這些例子只是用來解說我的論證是基於史特勞森對於參與者反應態度的看法，來論證機器人缺乏人類限定理由的這個事實。基於這個事實，機器人的道德擬人主義是錯的，在有些情況時，人對人的家長主義行為可以證成，但是機器人對人的家長主義行為卻不可以。我的論證不依賴對於上述例子的直覺。[11]

承上，也許可以說，即便機器人的確缺乏人類限定理由，但這對於最終判斷行為的對錯影響不大，因為人類限定理由的份量不重，因此十分容易被其他理由凌駕。特別是，在上述例子討論的都是陌生人之間的人類限定理由，合理想見，這些理由不會太強，因此在決定行為對錯時，不會造成實質差異。以〈自殺〉為例，也許「阻止老王可以拯救他的性命」這個理由具有決定性力量，凌駕了其他理由，因此無論是人類或機器人都可以阻止老王自殺。基於這個現象，也許有些人會認為人類限定理由沒有現實的價值。然而，至少在理論上，我們無法排除人類限定理由會對行為的對錯造成影響。此外，即便當它不對行為對錯造成差異時，行為者是否擁有人類限定理由仍然是值得討論的道德事實，我們可以說，即便人類跟機器人都可以阻止老王自殺，但是人類更適合來做這件事情。

總結來說，人類與機器人一項重大的道德差異在於機器人缺乏人格性，無法參與在人際關係之中。雖然機器人作為道德主體，在外顯的行為表現上可以跟人類一樣，甚至更好，但無法享有人際關

11 我感謝審查人提出這個意見。

係理由的事實，代表在涉及到干涉人類自主性的情況時，機器人的行為應當受到比人類更大的限制。

肆、心理面向：道德擬人主義的另一問題

上一節我從史特勞森的思想論述了人際關係理由，說明了機器人與人類之間的道德差異。這一節我想要從心理面向來討論道德擬人主義的問題。如前述，史特勞森的思想除了規範面向外，還包含了心理面向。我將說明，依照史特勞森的思想，對於人類的家長主義行為上，機器人應當受到比人類更嚴格的限制，因為對當事人來說，被機器人干涉自主缺乏了被人類所干涉的心理價值。我將分析電影《機械公敵》(*I, Robot*) 的一段情節來說明這點。

《機械公敵》描述在未來的世界中，自動化機器人已經廣泛使用。這些機器人遵守著名的艾西莫夫機器人三法則 (Three Laws of Robotics) 的約束，這三法則如下：

法則一：機器人不得傷害人類，或坐視人類受到傷害；

法則二：除非違背第一法則，否則機器人必須服從人類命令；

法則三：除非違背第一或第二法則，否則機器人必須保護自己。

由於機器人必然遵守三法則，機器人獲得人類信任被廣泛使用。然而，警探史普納 (Del Spooner) 是電影中唯一厭惡與不信任機器人的人類。他在調查一樁命案的過程中認識了在機器人公司任職的蘇珊・凱雯博士。史普納懷疑命案是機器人所為，而凱雯 (Swan Calvin) 博士認為機器人不可能違背三法則，她問史普納為什麼這麼厭惡機器人。史普納告訴她幾年前他曾經出了車禍，兩台車子被撞入河裡，他與另一台車的小女孩被困在車裡快要溺斃，幸好當時有個機器人經過，它遵守三法則立刻跳入河裡，但當時它只可

能拯救一人，史普納指示它先救小女孩，但在估算存活率後它先救了史普納。說完這個經歷後，史普納如下解釋了他對機器人的厭惡：

> 我是符合邏輯的選擇。它計算出了我有 45%的機率生還，但是莎拉只有 11%。她也是某人的孩子，11%也許夠了。每個人類都知道怎麼在警察和 12 歲的小女孩之間抉擇，但是機器人不會。機器人〔用手比比自己的心〕這裡什麼都沒有，只有燈泡與齒輪，妳想要信任它們妳請自便。(IMDb, 2004)

乍聽之下，史普納的解釋有點牽強。我認為多數人應該不會認為機器人的選擇有什麼錯誤，因為在那個情況下，救哪一個都是道德上許可的。這個例子讓道德擬人主義顯得十分有吸引力，既然對人類來說救哪一個都是正確的，那麼當然機器人救哪一個也都沒錯(存活率只是實用上，而非道德上的考量)。既然如此，史普納對機器人的厭惡顯得不合理。試想，如果是人類的救生員做了同樣的選擇，我們會認為他厭惡救生員是合理的嗎？

我同意機器人的作為在道德上沒有錯，並且在類似情況下，如果是人類救了他而史普納埋怨這個人是不合理的。然而，我認為史特勞森的思想可以幫助我們理解史普納對機器人的憎恨其實是有一定程度的合理性。

首先，雖然機器人的選擇沒有錯，但我們可以諒解史普納因為自己的意願遭到違背而感到憎恨。根據史特勞森的論點，憎恨，作為一種反應態度的負面情感，是回應對方的惡意或漠不關心：

> 我所討論的反應態度本質上是在回應他人對待我們時所外顯於行為中的意志品質：回應他們的善意、惡意或漠不關心。(Strawson, 1974: 15)

然而，當史普納憎恨機器人時，機器人並沒有任何意圖，它不是出於善意去救他，也不是出於惡意而去違背他的意願，也不能說它毫不在意史普納的感受。確切地說，機器人沒有任何意志可以讓史普納憎恨。根據史特勞森的觀點，史普納的憎恨缺乏有意義的對象。為了瞭解這點，我們可以拿機器人跟真人來比較。

假想如果是安琪救了史普納，史普納憎恨她違背他的意願。史普納可能有考量到安琪是出於善意的，但他還是可以憎恨她為何不能尊重他的意願，他可能認為安琪的善意是自以為是的。雖然我們可能還是認為史普納的憎恨是不合理的 (畢竟安琪的行為是道德上許可的)，但他的憎恨並非沒有意義，安琪的意志品質的確有值得憎恨的地方，也就是她缺乏尊重的那個部分。

然而，機器人的意志品質沒有任何值得憎恨的部分，因為它根本沒有任何意志。這不是說史普納無法憎恨機器人，而是他對機器人的憎恨缺乏適當的對象。我們可以試著站在史普納的角度去感受，當他察覺到機器人其實沒有任何意圖時，他會發現憎恨它其實沒有意義。相比於安琪，史普納還可以憎恨她的自以為是，他甚至還可以指責安琪。但憎恨、指責機器人沒有意義，機器人只是被設計成如此決策。當史普納瞭解了他憎恨機器人其實沒有意義，當他發現，他的意願、他對他人生的自主權被沒有心意的機器人干涉，他會更難以接受。

我認為史普納搞錯了，機器人不是不能信任，如果機器人是可靠的，它們的確是可以值得信任的。但如他所說，機器人並沒有「心」，這導致機器人不該被視為與人類同等的行為主體。如史特勞森所言，人際間的互動充斥了豐富的反應態度，這些反應態度賦予了人際互動重要的意義。讓機器人替代人類來與人互動，特別是那些涉及到干涉人類自主權的行為，必須考慮到即便機器人能夠和

人類一樣做出道德上正確的行為，機器人仍然不適合替代人類來做這些事情。

伍、結論

機器人是否可以作為道德主體？從外顯的行為主義來看，如果機器人可以和人類一樣好的遵守道德規則，則機器人就是道德主體。從應用自動化機器人的目的來看，我同意機器人可以是道德主體。但我要反對的是機器人的道德擬人主義。從行為主義的道德主體觀，很自然地會引伸出道德擬人主義：既然機器人是跟人類一樣好地遵守道德規則，那麼適用於人類的道德規則就適用於機器人，這就代表凡是對人類是道德上正確的，對機器人也是。

我對道德擬人主義的反對集中在家長主義行為。家長主義行為有兩個核心的面向：一是對當事者生命自主的干涉，二是對當事者帶來的好處。一般來說，當給當事者的好處大過於干涉他的壞處，家長主義行為可能是許可的。依據這個考量，我們很自然會認為，如果某個家長主義行為是人類可以做的，那麼機器人也可以。

然而，從史特勞森的洞見，我指出行為者的心意是道德上重要的，但在上述的考量中被忽視。這可能是因為對人類而言，採取家長主義行為通常是出於對於當事者的善意。然而，作為道德主體的機器人迫使我們得思考行為者心意的重要性。對史特勞森來說，在參與人際關係的互動之間，行為者的心意以及相應的反應態度是關鍵因素，這可分為規範面向與心理面向。

在規範面向上，史特勞森認為，所有人際關係的參與者都得承受程度不一的善意表達的要求與期待。這就提供了參與者理由去採取表達善意的行為，家長主義就是這樣一種表達善意的行為。而這

樣的理由是人類限定的，因為機器人由於缺乏人格性，無法成為人際關係的參與者。因此，機器人比起人類來說，少了一個理由來證成其家長主義行為。理論上，這蘊含有些家長主義行為機器人不應該做，但人類可以。這就反駁了機器人的道德擬人主義。

在心理面向上，即使當機器人的行為是道德上許可的，我們還應考慮人的心理感受。反應態度的情感是充斥在人際互動之間，可能我們已經習以為常而察覺不到這些情感的意義與價值。史特勞森提醒我們，這些情感是在回應顯露在行為中人的意志品質，而不是僅僅回應外顯的行動。在人際關係的互動中，很自然地甚至是不可避免地，我們會產生相應的反應態度。這些反應態度是在回應對方的意志品質。而當機器人做出類似的行為時，我們自然地也會有相應的反應態度，然而由於機器人缺乏心意，對它們的反應態度是無意義的。當察覺到我們的反應態度是無意義的，很有可能會更加憎恨機器人對我們生活的干涉，產生對無法掌控自己生活的無力感。

綜上所述，我論證讓機器人干涉人類的自主權，應該受到比人類更大的限制。我的論證可不可以應用到其他行為呢？在本篇文章，我只討論家長主義行為，但同樣的理由，應可以推廣適用於涉及人的反應態度的行為。

參考文獻

Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, 12, 3: 251-261. https://doi.org/10.1080/09528130050111428

Brożek, B., & Janik, B. (2019). Can artificial intelligences be moral agents? *New Ideas in Psychology*, 54: 101-106. https://doi.org/10.1016/j.newideapsych.2018.12.002

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines,* 14, 3: 349-379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Fossa, F. (2018). Artificial moral agents: Moral mentors or sensible tools? *Ethics and Information Technology*, 20, 2: 115-126. https://doi.org/10.1007/s10676-018-9451-y

Grodzinsky, F. S., Miller, K. W., & Wolf, M. J. (2008). The ethics of designing artificial agents. *Ethics and Information Technology*, 10, 2-3: 115-121. https://doi.org/10.1007/s10676-008-9163-9

Gunkel, D. J. (2012). *The machine question: Critical perspectives on AI, robots, and ethics*. Cambridge, MA: MIT Press.

Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology*, 11, 1: 19-29. https://doi.org/10.1007/s10676-008-9167-5

IMDb. (2004). *Quotes.* Retrieved from https://www.imdb.com/title/tt0343818/quotes/qt0474786

Laukyte, M. (2017). Artificial agents among us: Should we recognize them as agents proper? *Ethics and Information Technology*, 19, 1: 1-17. https://doi.org/10.1007/s10676-016-9411-3

Snowdon, P., & Gomes, A. (2019). *Peter Frederick Strawson*. Retrieved from https://plato.stanford.edu/archives/spr2019/entries/strawson/

Strawson, P. F. (1974). *Freedom and resentment and other essays*. London: Routledge.

Torrance, S. (2012). Artificial agents and the expanding ethical circle.

AI & Society, 28, 4: 399-414. https://doi.org/10.1007/s00146-012-0422-2

van Wynsberghe, A., & Robbins, S. (2019). Critiquing the reasons for making artificial moral agents. *Science and Engineering Ethics*, 25: 719-735. https://doi.org/10.1007/s11948-018-0030-8

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford, UK: Oxford University Press.

Watson, G. (2014). Peter Strawson on responsibility and sociality. In D. Shoemaker & N. Tognazzini (Eds.), *Oxford studies in agency and responsibility* (Vol. 2, pp. 15-32). Oxford, UK: Oxford University Press.

Weber, K. (2013). What is it like to encounter an autonomous artificial agent? *AI & Society*, 28, 4: 483-489. https://doi.org/10.1007/s00146-013-0453-3

Reflection on the Moral Anthropomorphism of Robots

Tsung-Hsing Ho
Department of Philosophy, National Chung Cheng University
E-mail: tsunghsing@ccu.edu.tw

Abstract

If robots are to function automatically, without human supervision, as depicted in sci-fi imagination, then we must ensure that robots not commit moral wrongs. According to the behaviourist conception of moral agency, if robots, assessed purely on the basis of behaviour, perform as morally as humans, they can be considered moral agents. This naturally leads to moral anthropomorphism: the position that whatever moral standards apply to humans apply equally to robots. I argue against moral anthropomorphism. In light of P. F. Strawson's insights into interpersonal relationships and reactive attitudes, and drawing on paternalist actions as examples, I argue that robots, being not persons, are unable to participate in interpersonal relationships, and therefore their paternalist actions towards humans ought to be less permissible than humans'.

Key Words: robots, artificial intelligence (AI), moral agency, moral anthropomorphism, reactive attitudes