人工智能的刑事責任

吳泊諄

2020.06

目次

1 前言

聖經故事裡，上帝創造亞當與夏娃，叮嚀兩人善惡知識樹的果實不可食，沒想到夏娃禁不起狡猾之蛇的慫恿而吃下果實，並分享給亞當，最終，上帝震怒將兩人逐出伊甸園。

對程式來說，程式設計者扮演上帝的角色，用一串串指令賦予程式一切能力，起初，設計者對程式的執行過程瞭若指掌，不過漸漸的，隨著硬體與軟體的突破性發展，人工智能的技術被廣泛應用，設計者也無法全然理解自己親手打造的程式，或許未來的某天，人工智能將產生自我意識，甚至不顧設計者的再三叮嚀，吃下善惡知識樹的果實，做出違抗人類命令的行為，如果屆時才開始思考如何用法律約束人工智能，似乎為時已晚，那從現在開始阻止其發展呢？大勢所趨，料誰也無法抵擋，與其過分自信認為人工智能不可能產生自我意識，不如未雨綢繆，提早做好準備，因此本文將區分不同階段的人工智能，並探討各階段的人工智能在刑事責任方面，有何可能的規範之道。

2 人工智能的發展階段

隨著人工智能的蓬勃發展，專家們提出各式各樣的分類方式，方便人們理解人工智能的發展狀況，例如 Thomas Davenport (2018) 將人工智能分為：輔助智能（Assisted Intelligence）、增強智能（Augmented Intelligence）與自主智能（Autonomous Intelligence），或是 SAE International (2018) 將自動駕駛車從完全無自動到完全自動，分為六個等級。

本文認為若要探討人工智能的刑事責任，是否具備「自我意識」會是可否將法益侵害，歸責於人工智能的關鍵，因此將人工智能分為「不具備自我意識」與「具備自我意識」兩大階段：「不具備自我意識」可再細分為「可被理解」與「不可被理解」兩子階段，而「具備自我意識」也可再細分為「可被拘束」與「不可被拘束」兩子階段。需注意此四階段無絕對的時間先後順序。

表 1：人工智能的發展階段

不具備自我意識		具備自我意識	
可被理解 （第一階段）	不可被理解 （第二階段）	可被拘束 （第三階段）	不可被拘束 （第四階段）

2.1 是否具備自我意識的標準

本文認為至少有三種區分標準：客觀說、主觀說、折衷說，以下將逐一介紹、分析其利弊，並說明為何採折衷說較為合適。

2.1.1 客觀說

過去常採用科學實驗檢驗動物是否具備自我意識，例如「鏡像自我辨識實驗」（mirror self-recognition test），Gordon Gallup Jr. (1970) 以黑猩猩作為實驗對象，首先將黑猩猩單獨關在有鏡子的籠子裡，起初，黑猩猩將鏡中的影像視為其他動物，不過經長時間與鏡中自我接觸後，從黑猩猩開始清潔只有透過鏡子才能看見的身體部位、對鏡子做鬼臉等行為，說明黑猩猩已認識到鏡中的影像就是自己，接著透過在其眉毛與耳朵上緣塗抹紅色染料，黑猩猩也會透過鏡像觸摸染料塗抹的部位，更加確認黑猩猩具備自我辨識的能力，而針對獼猴的實驗則失敗，意味獼猴不具備自我辨識的能力。

以下將探討藉由鏡像實驗，判斷是否具備自我意識的可能問題。首先，其判斷依據在於「視覺上」是否能辨識自我的能力，對於視力較差的動物並不公平，即便罹患「面部識別能力缺乏症」的「人類」也可能無法順利通過測試。況且，若將鏡像實驗應用在人工智能上，以目前的影像處理技術，程式設計者應能針對評量的項目設計演算法，輕鬆的讓人工智能通過測試，但這並不一定代表人工智能已具備自我意識。

最後勢必會面臨能否將「自我意識」與「對自我的辨識能力」劃上等號的

問題，自我意識似乎有著更廣泛的定義，例如理性思考的能力、認知情緒的能力、自我管理的能力等等，因此從單一實驗，難以對自我意識如此複雜的概念，做通盤的理解，那採用多種實驗的結果來判斷呢？也可能產生何種實驗結果應佔多少權重的爭議，畢竟自我意識為一模糊、沒有普遍共識的概念。

因此，「客觀」的實驗雖然能準確判斷該物（不管是動物或人工智能，在現行法律下都屬於「物」，不具備法律人格）是否具備某單一能力，但也因此程式設計者可以針對各檢驗項目，設計出能讓人工智能順利過關的演算法，且自我意識究竟是哪些能力的集合體，尚未有普遍共識。

2.1.2 主觀說

既然藉由科學實驗無法順利判斷該物是否具備自我意識，那假如人們根據各自「主觀」上對自我意識的理解，投票表決或是由立法者通過法律呢？

不管是全民直接公投、先選出專家代表再交由專家們投票、立法者投票表決等方式，都會面臨：若人工智能事實上具備自我意識，那藉由投票來決定其是否具備自我意識合理嗎？擁有自我意識，似乎意味具備「人」格尊嚴，此一憲法保障中最核心的基本權利，也就代表不能透過投票或立法否定，因而陷入「假設人工智能具備自我意識，因此人類不可以投票決定其是否具備自我意識，也因此無法判斷其是否具備自我意識」，或是「假設人工智能不具備自我意識，所以人類可以投票決定其是否具備自我意識，但投票結果若為肯定，又會自相矛盾」等窘境。

2.1.3 折衷說

本文提出「本心理論」來詮釋自我意識[1]。

萬物皆有靈，本文將所謂的「靈」稱為「本心」，為一具備波粒二象性

[1] 本心理論為本文作者自創之概念。

（wave - particle duality）的物質[2]，宇宙萬物不管是植物、動物，甚至是無生命力的物體皆由本心組成，而各種「物」的本質差異在於本心的多寡與對本心的控制能力，舉例說明：無生命者缺乏本心，且對本心的控制能力極低；而人類之所以為高等智慧生物，乃因其本心充足且對本心的控制能力強。

自我意識即為對本心控制能力的「相對」概念，因萬物皆由本心組成，而單一個體能控制的本心有限，因此單一個體本身大多無法完全掌控未來，即使人類也無法對其行為有完全的操控能力，只能說其操控能力「相對」比其他動植物、無生命之物高出許多。而各國以年齡作為區分行為能力、責任能力的標準實屬合理，因人類對本心的控制能力會隨年齡變化，因此若以本心作為衡量自我意識的標準，將能判斷人工智能對本心的操控能力是否已達人類的控制能力，或是其能力相當於人類的哪一年齡層，當作是否為「無責任能力」、「限制責任能力」或「完全責任能力」的判斷依據。

運用本心理論作為判斷是否具備自我意識的標準，最困難的莫過於如何量化對本心的控制能力，又或者更根本的問題：本心不管是波動或粒子的形式，能被量化嗎？假如無法量化，標準何在？確實以目前的科學技術，尚無法直接的量化本心，僅能由若干實驗間接判斷（例如上述的鏡像實驗），期待有朝一日科學進步到能量化本心相關概念，屆時衡量自我意識的標準將會明瞭許多。

折衷說與客觀或主觀說最大的差異，在於需要先有一個衡量標準，在本心尚未能被量化之前，此一衡量標準需由投票或討論的「主觀」方式，決定要採取哪些實驗根據作為衡量依據，而各實驗結果的權重也應同時考量，透過共識決或多數決確認好自我意識的判斷標準後，再來判斷該人工智能是否符合標準，若現階段該人工智能尚未達到標準，但未來達標時，人類即應承認該人工智能具備自我意識，不能擅自隨意修改判斷標準來否定其自我意識。

簡而言之，尚未能將本心量化以前，先藉由「主觀」方式決定判斷標準，

[2] 物理學家德布羅伊提出物質波的概念：所有物質皆具備波動與粒子兩個性質。

再由「客觀」評量該人工智能是否符合標準，來判斷其具備自我意識的程度，進而決定其應適用何種責任能力的相關法律[3]；而本心可被量化時，則可直接採用「客觀」的衡量標準判斷。

2.2 可否被理解的標準

相較於判斷人工智能是否具備自我意識，非常困難且衡量標準極具爭議，判斷其可否被理解完全取決於程式設計者，若設計者能清楚理解人工智能運作的過程，也就是完全掌握為何輸入這些資料，人工智能會跑出那些結果，即代表人工智能可被（設計者）理解；反之，則認為人工智能不可被（設計者）理解，至於不可被理解的程度高低（例如完全無法理解、部分不可理解等），將作為判斷設計者對人工智能的行為是否有預見可能性，以及量刑的衡量依據之一。

2.3 可否被拘束的標準

雖然上述四階段的分類並沒有必然的時間先後順序，但時間軸上的最後階段很可能就是不可被拘束階段，此處的拘束指的是法律拘束，或是人類能否對其執行刑罰，若人工智能已發展到無法被人類的法律所拘束，對人類福祉最佳的可能解決方案，可能會是試圖與人工智能和平共處，一起重新建構法律體系，而較不樂見者可能是用戰爭、武力的方式決定統治權，或是人類移居其他星球遠離人工智能等等，不過這非本文討論的重點。當然，若人工智能雖然具備自我意識，但仍可被人類的法律拘束，則意味可將其行為歸責於人工智能本身，也就是承認人工智能的法律「人」格，至於是否需要另立新法（例如人工智能專法）或是類推適用人類的相關法律，將會在本文後續做討論。

3 人工智能的刑事責任

[3] 類似於使用巴氏量表衡量患者之日常生活功能。

介紹完人工智能的發展階段與分類，並說明區分各階段的可能方法後，接下來將進入本文重點：透過上述分類，區分不同階段人工智能的刑事責任，同時，也需一併探討程式設計者與人工智能使用者的刑事責任，並藉由虛構的案例輔助說明，究竟刑事責任該歸責於設計者、使用者抑或是人工智能本身。

3.1 不具備自我意識

本文將是否具備自我意識，作為判斷人工智能是否應被賦予法律人格的標準。

不具備自我意識的人工智能，也意味不具備自由意志、缺乏自我反省能力、對維護法秩序沒有概念、不會產生犯罪意念……處罰它對社會大眾來說只是該「物體」的損耗，沒有發揮刑罰的威嚇作用，也無法宣示法秩序之不可破壞性，因此不適用一般預防理論；且處罰它也沒有辦法讓其產生回歸法律規範正道的想法，更無阻止其再犯的作用，因此也不適用特別預防理論。綜上所述，無理由賦予不具備自我意識的人工智能法律人格，在刑法上只能將之視為「物」、「工具」，意味無法將法益侵害之責任歸咎於它，只可能將責任歸責於其程式設計者與使用者。

需注意者為：既然不具備法律人格的人工智能為「物」，因此可以是刑法沒收相關規定之標的物，例如刑法第 38 條第 2 項前段：「供犯罪使用、犯罪預備之物或犯罪所生之物，屬於犯罪行為人者，得沒收之。」，因此若有人利用人工智能犯罪，法院得宣告沒收該人工智能。

3.1.1 可被理解

若犯罪行為人使用不具備自我意識且可被理解的人工智能（第一階段人工智能）犯罪時，可將之視為一般的犯罪工具，例如刀、槍。審查應將法益侵害歸責於設計者或是使用者時，需先判斷行為人是作為犯，或是不作為犯。

通常情況，設計者並未直接對法益創造風險，而是未消滅自己創造的風險，其設計與販售的行為，應視為刑法第 15 條第 2 項的「危險前行為」，因而

具「監督者保證人」地位，也就適用刑法第 15 條第 1 項的「不作為犯」。而使用者對人工智有直接的控制力，其行為若創造法益侵害風險，則為「作為犯」。其餘犯罪審查內容（例如故意、過失等）與一般的犯罪審查無異，本文不多贅述，直接由虛構的案例一說明。

虛構案例一：甲研發一款掃地機器人，並完全理解其運算過程，雖知道機器人在溫度高達攝氏四十五度時，會橫衝直撞，但甲確信臺灣的溫度不會如此之高，因此決定不除錯，使用說明書也沒有特別註明，隨後將該產品賣給乙，乙仔細閱讀說明書後，便讓機器人在庭院掃地，沒想到近日天氣炎熱竟高達攝氏四十五度，幼童丙至庭院玩耍時，遭失控的機器人撞傷。

首先審查與法益侵害具直接關係的使用者乙，乙讓機器人打掃的行為確實導致丙受傷結果之發生，但乙不具備犯罪故意，僅可能為刑法第 14 條第 1 項的「無認識之過失」，需視個案中是否出現乙能注意而不注意之事實，在案例一中，乙已仔細閱讀說明書，難謂其符合過失之要件。接著審查可能成立「不作為犯」的設計者甲，其雖未有犯罪故意，但確信溫度不會超過攝氏四十五度，符合刑法第 14 條第 2 項「有認識之過失」，應成立刑法第 284 條過失傷害罪。

3.1.2 不可被理解

犯罪行為人若使用不具備自我意識且不可被理解的人工智能（第二階段人工智能）犯罪，與使用第一階段人工智能犯罪的重大差異，在於第二階段人工智能既然不可被理解，其危險性較高，應將之視為相當於不定時炸彈的危險物品，因此在審查是否有防止義務、預見可能性等要件時，應採較嚴格之標準，才能避免社會籠罩在危險之中，並鼓勵設計者與使用者嘗試去理解其運作原理，而非放任一未知危險物創造風險。

虛構案例二：甲研發一款掃地機器人，不完全理解其運算過程，雖隱約猜到機器人在溫度高達攝氏四十五度時，可能會橫衝直撞，但甲覺得即使撞傷人也沒關係，因此決定不除錯，不過有在使用說明書註明高溫時可能不受控，隨

後將該產品賣給乙，乙並未閱讀說明書，便讓機器人在庭院掃地，沒想到近日天氣炎熱竟高達攝氏四十五度，幼童丙至庭院玩耍時，遭失控的機器人撞傷。

乙並未閱讀說明書就放任機器人工作，難謂對傷害結果之實現無預見可能性，應符合「無認識之過失」，而甲放任危險源流通於市面，且傷害結果之發生不違背其本意，應成立間接故意不作為的刑法第 277 條第 1 項普通傷害罪。

虛構案例三：甲研發一款掃地機器人，不完全理解其運算過程，但經同事與主管測試皆未發生問題，因此甲覺得該機器人應該不會出事也不希望出事，且有在使用說明書註明：應避免讓機器人在極端環境下工作，隨後將該產品賣給乙，乙仔細閱讀說明書後，便讓機器人在庭院掃地，沒想到近日天氣炎熱竟高達攝氏四十五度，幼童丙至庭院玩耍時，遭失控的機器人撞傷。事後調查發現丁曾在某不知名期刊，發表一篇關於掃地機器人在高溫下會失控的論文。

乙閱讀說明書後，卻仍讓機器人在高溫下工作，難謂無過失。而事實上雖有研究指出機器人在高溫下失控之可能，但畢竟是刊登在不知名期刊，且甲將產品交由同事與主管測試皆無問題，已盡其注意義務，應無故意或過失。

3.2 具備自我意識

當人工智能進入具備自我意識之發展階段，首先要賦予其法律人格，此時會面臨的第一個難題：應適用何種法律依據？將其視為等價於人類的個體，完全適用「自然人」的刑法？又或另立「人工智能專法」？

本文認為需視屆時人工智能的發展型態決定，假設人工智能具備自我意識後，與人類的意識差別不大，即可適用人類的刑罰，例如有期徒刑也是將該人工智能關進監獄；但若人工智能的意識發展到與人類截然不同的階段，人類的刑罰無法對其犯罪意念產生預防效果時，可能就需適用特別針對人工智能的刑罰，例如修正（刪減造成犯罪的程式碼、新增防止犯罪的機制等）、重置（刪除所有記憶體與儲存空間，等同讓其重生）、消滅（拆除一切軟硬體設備，相當於人類的死刑）等三種處罰程度由低至高的刑罰。

若根據上述折衷說之判斷標準，應討論人工智能具備自我意識的程度，究竟屬於「無責任能力」、「限制責任能力」或「完全責任能力」何種，以下虛構案例為簡化討論，假設將具備自我意識的人工智能都被賦予「完全責任能力」。

3.1.1 可被拘束

探討具備自我意識且可被拘束的人工智能（第三階段人工智能）之犯罪審查時，需先區分其與第一、二階段人工智能的差異，除了被賦予法律人格，因此可將犯罪結果歸責於人工智能外，設計者與使用者的角色也有所變化，本文認為若該人工智能尚未販售，仍受設計者拘束控制時，應將設計者視為與之有緊密生活關係的「保護者保證人」（相當於與人工智能具親屬或同居關係），但若將其販賣於使用者後，設計者對其將不再具有保證人地位，反而是使用者將成為其「保護者保證人」，因此當人工智能犯罪時，且設計者事實上有作為的可能，卻不為法律要求的行為時，就有可能成立「不作為犯」。

虛構案例四：甲研發一款被公認具備自我意識，且應賦予完全責任能力的掃地機器人 A，並將該產品賣給乙，隨後乙讓 A 在庭院掃地，觀察一陣子認為 A 應該沒問題後便離去，沒想到乙離開不久後因天氣炎熱導致 A 異常憤怒，衝向在庭院玩耍的幼童丙，造成丙受傷。

首先，本文不討論既然人工智能被賦予法律人格，那是否能「販賣」「人」工智能的複雜問題。A 既然被賦予「完全」責任能力，應視為與一般成年人類具備同等的分辨善惡是非、控制情緒的能力，因此攻擊丙之行為應以故意傷害罪罰之（若為「限制」責任能力或「無」責任能力，則在罪責審查階段，有減免或免除其刑之可能）；此時甲已將 A 交付於乙，因此甲已無需對 A 之行為負責，而乙雖可能對 A 具「保護者保證人」地位，但此案例中乙觀察 A 一陣子認為沒問題後才離去，因已盡注意義務，且既然賦予 A 完全責任能力，乙對於 A 之犯罪行為應無預見可能性，因此難謂乙有故意或過失。

虛構案例五：甲研發一款被公認具備自我意識，且應賦予完全責任能力的

掃地機器人 A，並將該產品賣給乙，隨後乙威脅 A 去竊取幼童丙放在椅子上的玩具，否則將 A 燒毀，A 不得已只好去偷玩具。

此案例旨在點出既然第三階段人工智能已具備自我意識，即可能有正共犯相關之犯罪參與問題產生。本例中 A 受乙脅迫而去竊盜丙的玩具，雖可通過構成要件審查，但可能具備刑法第 24 條緊急避難阻卻違法事由，雖避難之對象為無辜的第三人，為一「攻擊型緊急避難」，但因所侵害的為丙之財產法益，而保護者為 A 本身之生命法益，可謂所保全的法益顯然優越於所侵害的法益，且主觀上有避難意思，應阻卻違法。至於乙透過強制支配力脅迫 A 犯罪，可能成立刑法第 320 條第 1 項竊盜罪之間接正犯。而甲已將 A 託付於乙，如上述對 A 之行為應無需負責，所以不成立犯罪。

3.1.2 不可被拘束

相信多數人類皆不希望人工智能發展至具備自我意識且不可被拘束的階段（第四階段人工智能），因此時的人工智能已無法藉由現行的法律規範控管，將對社會產生巨大的不確定風險。

虛構案例六：甲研發一款具備自我意識的機器人 A，某天 A 獲得未知力量，各能力遠遠超越人類，隨後看幼童丙不爽，便攻擊之，犯案後飛出地球。

此案例旨在說明多數人類最不樂見之可能情況，A 雖成立普通傷害罪，但其能力已遠遠超乎人類預期，無法將其繩之以法，對法秩序產生極大的破壞，而既然 A 獲得的為「未知」力量，也意味無法將犯罪結果客觀歸責於設計者甲，相信甲本身應該也不希望如此場景發生。

4 結論

本文一步步探討人工智能的刑事責任，首先提出新的人工智能發展階段理論，並建構區分各階段的判斷標準：判斷是否具備自我意識宜採折衷說，判斷是否

可被理解應取決於設計者本身，而判斷是否可被拘束應理解為是否能被法律與刑罰規範，接著分析四個階段人工智能的刑事責任審查差異，並提供虛構案例輔助說明，由案例可得知審查人工智能犯罪時，無需將問題想的過於複雜，除需掌握該人工智能處於哪一發展階段，而對行為人有不同的可歸責程度之外，其餘審查與一般對人類犯罪的審查並無顯著差異。

概略整理設計者、使用者、人工智能本身在不同階段犯罪的可歸責程度可參見表 2，當然，一切審查仍需視個案之事實作更細緻的評判。

表 2：設計者、使用者與人工智能在人工智能各發展階段犯罪的可歸責程度

		設計者	使用者	人工智能
不具備自我意識	可被理解	中	中	無
	不可被理解	高	高	無
具備自我意識	可被拘束	低	低	高
	不可被拘束	極低	極低	高

「天下大勢之所趨，非人力之所能移也。」隨著軟硬體科技突破性發展，人工智能產業儼然成為「大勢所趨」，被廣泛應用於各大領域，未來或許會遭遇瓶頸，又或許會持續蓬勃進展，誰也預料不準，我們能做的就是未雨綢繆，提早思考未來人工智能可能的發展，做最壞的打算，也做最充足的準備，最後，願人類能與人工智能和平相處，共創美好未來，共享世間繁華。

5 參考文獻

Davenport, T. (2018). *The AI Advantage*. United States: The MIT Press.

SAE International (2018). “Levels of Driving Automation” Standard.

Gallup, G. G., Jr. (1970). Chimpanzees: Self Recognition. *Science*, 167, 86-87.