香港持續發展研究中心
香港城市大學
City University of Hong Kong

香港城市大學

香港持續發展研究中心 [1]

第 29 號建議書

建立人工智能倫理與監管框架的幾點建議

卓雋傑、陳浩文、熊天祐、李芝蘭、林芬、Viktor Tuzov、李建安 [2]

1. 引言

隨著人工智能技術應用範圍日益廣泛，對人類生活影響也愈來愈大。因此，包括歐美以及亞洲等多國都認同，在推進人工智能發展的同時，需要制訂一套適切的監管措施，確保人工智能符合普遍認同的道德倫理價值。要指出的是，監管並不等同制約技術發展，反而是相輔相成，這一點可以從大型語言模型 (Large Language Model，LLM)過去兩年在香港普及的例子看到。

當首款大型語言模型 ChatGPT 在香港初出現時，多家機構（如大學）對是否允許使用出現了不同判斷，限制了技術在港的發展與普及應用。當社會逐漸理解 LLM 可能會衍生抄襲、侵犯知識產權、數據安全和虛假信息等問題，並設立相應處理機制與控制工具後，LLM 才算得上真正地開放使用。換言之，社會制訂使用 LLM 的監管守則，不單沒有制約 LLM 在港的發展，相反，正正是因為有了明確而通用的使用共識和規範，LLM 才得以普及。

[1] 香港城市大學持續發展研究中心（CSHK）成立於 2017 年 6 月，是一個開放和跨學科的研究平台，旨在促進及增強香港學術界、工業界和專業服務界; 社會及政府; 以及香港與不同區域之間的協作，並從事有影響力的應用研究範疇包括香港專業服務、一帶一路、粵港澳大灣區、綠色經濟、新冠病毒（COVID-19）等，研究項目屢獲資助，並出版多份研究報告、論文和書籍。更多資訊請瀏覽中心網頁 http://www.cityu.edu.hk/cshk。如對本政策建議書有任何 意見，歡迎電郵至：sushkhub@cityu.edu.hk。

[2] 卓雋傑為香港城市大學公共及國際事務學系 博士生 ； 陳浩文為香港城市大學公共及國際事務學系教授 ； 熊天佑為香港大學經管學院特約副教授及 香港持續發展研究中心 國際及專業顧問 ； 李芝蘭為香港城市大學 公共及國際事務學系教授 、 香港持續發展研究中心總監 ； 林芬為 香港城市大學 協理副校長 （ 環球戰略 ） 及 媒體與傳播系 副教授 ； Viktor Tuzov 為香港城市大學媒體與傳播系博士生 ； 李建安為香港持續發展研究樞 紐成員 。

我們相信，沒有共識和規範標準來控制技術衍生的問題，智能技術就沒有廣泛應用和充分發展的空間。因此，我們認為香港需要建立一個更具前瞻的監管與市場資訊環境，從而吸引人工智能企業來港發展。這對於香港發展成為國際數據中心同樣不可或缺，有效的管理環境可降低技術（規範）前景的不確定性，並縮短建立共識與規範的過渡期。

2. 人工智能監管的取向

各國政府就監管人工智能已有一系列討論，並就大原則達成共識 [3]。然而，對於具體的監管方式，各方取態並不完全一致。歐盟在 2023 年 12 月通過的《人工智能法案》著重***「風險為本」***，以人類存亡、人身安全以及基本人權為底層邏輯，把人工智能技術分成 4 大類別。業界在發展人工智能技術時不能與此相悖，否則歐盟監管機構就會介入，甚至明令禁止 [4]。

表一、歐盟人工智能法 Artificial Intelligence Act

	風險程度	例子	規範方法
第一類	無法接受 unacceptable	**對人類構成嚴重威脅：** 包括遠端生物特徵辨識；預防犯罪執法	全面禁止
第二類	高度 high	**有可能侵害人身安全或基本人權:** 實時監控交通等基礎設施；追蹤個人醫療或教育狀況	推出之前及應用時受各方嚴格審查
第三類	低 limited/low	包括 ChatGPT 等生成技術軟件	公開披露足夠資訊
第四類	極低 minimal/none	包括過慮有毒或垃圾郵件的系統；遊戲軟件	沒有特別規範

可以看到，歐盟是以人工智能技術應用範疇可能引伸的風險來作監管劃分，屬「極低」或「低」風險程度的類別規範相對較少。但當技術應用有可能侵害人身安全或基本人權時，就會受到各方嚴格審查。

至於美國則強調市場參與者的角色，相對較重視***「應用為本」***，人工智能開發者需要同時自行兼顧技術發展及公平和私隱等原則。雖然美國聯邦貿易委員會具一定的中央官方監管機構職能，但作為監管主體的各方地方政府卻更依重業界團體透過制訂行業守則或承諾，約束各市場參與者的行為 [5]。

[3] 例如在 2023 年 7 月，聯合國秘書長安東尼奧·古特雷斯呼籲建立全球 AI 監管機構；2023 年 11 月舉辦的人工智慧安全高峰會中，多國政府共同簽署並發表《布萊切利宣言》，致力（1）識別共同關注的人工智慧安全風險和（2）制定各自基於風險的政策，並酌情開展合作。

[4] 可參見歐盟網站: https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai

[5] 關於不同地方的人工智能監管取態，可參見馮一帆「人工智能監管制度概覽」2023 年 11 月 2 日刊於《人民法院報》； Cecilia Kang & Adam Satariano, Five ways AI could be regulated, New York Ttimes, 7 Dec 2023.

當下，香港沒有一個專門負責監管人工智能發展的官方機構。至於涉及規範人工智能的官方文件，最早是 2021 年香港私隱專員公署發布的《開發及使用人工智能道德標準指引》，要求企業在開發及使用人工智能時要遵從《個人資料（私隱）條例》，並附設一份「自我評估核對清單」，協助機構判斷是否已採納指引所建議的措施。

及至 2023 年 8 月，港府資訊科技總監辦公室發布了更為完整的《人工智能道德框架》，協助政府各部門採用人工智能和大數據分析技術，將道德規範要素納入其中。港府指出《框架》的指導原則、指引和評估也適用於其他機構。

表二、香港《人工智能道德框架》

主要用戶	✓ IT 規劃師 ✓ 系統分析員 ✓ 系統架構師 ✓ 數據科學家
槓桿影響	
人工智能道德框架	■ **度身訂造的道德框架** ◆ 人工智能倫理原則 → 1. 透明度和可解釋 2. 可靠、穩健和保安 3. 公平 4. 多元和共融 5. 人工監督 6. 合法和合規 7. 數據隱私 8. 安全 9. 問責 10. 有益 11. 合作與開放 12. 可持續和公平過渡 ◆ 人工智能管治架構 ◆ 人工智能生命周期 ◆ 人工智能執行指引 ■ **人工智能評估** ◆ 人工智能應用影響評估 ➢ **用於資訊科技和計劃的現存標準** ◆ 項目管理 ◆ 資訊科技保安政策及指引
履行	
人工智能應用	✓ 項目策略 ✓ 項目規劃 ✓ 項目生態系統 ✓ 項目發展 ✓ 系統發展 ✓ 系統運作及監察

從上述發展可以看到，香港監管人工智能的特點，一方面是將側重點扣連至企業及其開發人員身上；其次是這些指引都沒有法定約束力，因為港府相信人工智能行業與其他一般企業無異，同樣已受到香港現存的法例監管。

創新科技及工業局局長孫東指出，互聯網世界並非無法可依，大部分在現實世界用以防止罪行的法例原則上均適用於網絡。舉例來說，《2021 年刑事罪行（修訂）條例》就可規管未經同意下發布或威脅發布私密影像的罪行，散播不實或不當資訊亦有其他法律可以處理[6]。

3. 對人工智能倫理與管治的幾點建議

我們都同意，人工智能仍在不斷演變中，各地區及技術之間的發展差異極大，對不同行業和界別的影響亦不盡相同，因此要制訂一套監管人工智能的完整框架還需假以時日，審時度勢。然而，此節我們將結合研究調查的結果，提出一些建議，希望能夠在社會探索監管人工智能的過程中出一分力。

3.1 管治框架應著重「應用型」；構建案例資料庫避免空洞倫理原則

港府資訊科技總監辦公室的《人工智能道德框架》臚列了 12 項涉及人工智能的道德原則，當中包括透明、可靠、公平、多元、合法合規以及開放等等(請參見表二)。我們早前的政策建議書中已指出，在「去語境化」的情況下，受眾對道德原則往往會呈現最大化(maximization)傾向，也就是說所有原則皆會被視為「重要」[7]。

因此簡單地臚列道德原則，反而可能會誤導公眾以為在人工智能發展過程中，所有倫理價值都可以無條件地被滿足，毋需權衡取捨。我們曾與人工智能行業的實務人員訪談，他們都不諱言這種口號式的綱領無助制定具體政策。

我們建議港府在草擬人工智能監管框架時，要考慮人工智能在不同情景應用的道德屬性差異。這可透過建立案例資料庫來達致，當不同情景突出不同的道德側重範疇，就能讓社會大眾更易對焦。

舉例來說，我們此前的研究顯示，智能技術的應用情景不同，市民較關注的倫理原則和價值也有不同。受訪者對於「健康碼系統」與「欺詐檢測系統」（應用程式），更關注與個人相關的價值原則，如私隱和行動自由；但在「無人駕駛汽車」（重型器械）方面，市民更會傾向關注與安全相關的系統穩健性。

建立基於不同情景的應用型管治框架，就能容許人們快速地在錨定（anchored）的類型框架下建立對智能技術的穩定觀感，從而令後續的管治變得更具體及簡單。

[6] 可參見立法會十題：規管人工智能技術生成的內容，2024 年 1 月 4 日下載自: https://www.info.gov.hk/gia/general/202305/31/P2023053100249.htm

[7] 可參見卓雋傑、陳浩文等 (2023)《人工智能發展的倫理考量》，CSHK 第 28 號政策建議書。

3.2 設立多方交流平台；突出第三方專業人士的角色

當我們與不同持份者交流時，他們都認為當今人工智能監管難處主要原因之一，是資訊太多及零散，無助促進公眾對技術的理解。因此，他們都相信專家是人工智能範疇有效管治的核心，這個想法又可以細分為兩類。

- *第一類意見: 科技素養是在專家或權威人物定義後再單向地推廣至市民或消費者；*
- *第二類意見: 科技素養應由公眾和專家互動下產生。*

前者傾向相信專家的權威性，並接受他們能夠傳達真實準確的技術知識。後者則更關注公眾在科技素養的學習和定義（倫理原則）上的話語權，並確保公眾得到充分的知識。

由於我們認為第二類意見更為可取，因此我們建議在構建人工智能監管框架的過程中，要設立一個多方交流平台。香港人工智能的監管框架應通過磋商；或其他建立共識的方式來訂定，從而提高社會公眾的人工智能技術素養。

無可否認，由政府及專家單向主導，或也能夠建構出一套人工智能監管框架(例如當下的《開發及使用人工智能道德標準指引》及《人工智能道德框架》)，但在缺乏社會層面的推廣和公眾參與下，公眾對於人工智能的理解依然不足。當人工智能應用層面愈來愈廣泛，若專家堅持閉門做車，恐怕效果會適得其反。

我們此前的問卷調查已表明，近6成的受訪者認為當人工智能應用出現道德困境時，「受影響市民」的意見應該是最優先考慮。因此框架設定者不應忽視公眾的感受和想法。雖然公眾對人工智能技術可能理解不足，但專家可以與公眾作更多互動，讓公眾/技術對象擁有更大的話語權，建立更多的共識。

我們建議應加強對人工智能技術研發和營運者與公眾的交流，並支持實務經驗和例子的引用及分享，增加公眾對於人工智能技術的信心。另一方面政府則可以考慮加大對於法定機構（如私隱專員公署和消委會）和大學等第三方專業組織的支援，以增加人工智能技術的可靠度。

3.3 政府在監管的角色定位需清晰拿捏

前述提及歐盟和美國對人工智能採取不同的監管方向，歐盟以政府設定的風險為藍本；美國則傾向依賴業界自主制訂規範守則。但無論是哪一種取向，政府都必須有一定程度介入。尤其是當人工智能技術超出社會認可的道德界線，擁有公權力的政府就要出手阻

止[8]。社會不能完全放任私人企業發展人工智能技術還有以下兩個原因:

- *企業缺乏自我監管的動機(Motivation)：企業在多數時間往往會將商業利益置於公眾或消費者的利益上，即使其行為對社會造成負面的溢出效應，在未超出法例允許的情況下，很少受到懲罰。*

- *企業缺乏足夠改善的能力(Capacity)：企業的能力及信任來源於自己固有的消費者，而非全體公眾。當企業之間的競爭（如中小型初創企業與大公司之間的競爭）以及消費者群體之間缺乏明確和一致的需求，使企業難以達成統一的標準或共識。*

要提升企業自我監管的動機，就要制訂具強制力的法律，規管涉及人工智能應用的行為。至於改善企業能力，則可透過公共機構和第三方專業人員的外部參與，提供協助。要達致以上兩個目標，都需要擁有最多社會資源的政府扮演關鍵角色。

但要留意的是，政府的介入不應影響企業的自主性，自由市場競爭是企業人工智能創新的主要驅動力，忽視或過度干預企業運作將會削弱創新的動力和靈活性，導致技術發展停滯。 因此，政府的定位需要妥善拿捏，既要協調社會各界達致監管目標，又不損害技術創新。

4. 小結

隨著人工智能技術的不斷發展和改進，我們可以預見到它在各個領域的廣泛應用，也可能帶來一些挑戰和風險。

基於我們的調查研究及與業界的接觸，我們認為對人工智能發展採取放任不管的態度並不可取，尤其當我們希望香港成為國際數據中心，並吸引人工智能企業來港發展時，香港必須營造一個全面而前瞻的管治與市場資訊環境。但同時，政府的介入需要保持適度，過度干預反過來會窒礙創新。

我們要再重申，監管人工智能發展及應用並不是要制約技術發展，沒有共識和規範標準才會社會因為懼怕其衍生的風險而裹足不前。我們相信，有效的管理環境可降低技術（規範）前景的不確定性，並縮短建立共識與規範的過渡期。

因此，政府與業界及社會大眾必需加強溝通以達致共識，確保人工智能技術的使用安全和合法。同時，我們需要建立「應用型」監管框架；設立案例資料庫；支援第三方專業人士及專家學者參與相關研究、教育社會公眾，這樣才能促進人工智能技術在港的發展和應用。

[8] 主張業界規範的美國，參議院司法委員會就曾在 2023 年 5 月召開聽證會，傳召 OpenAI 執行長阿特曼出席作供。席間多位議員警告人工智能技術不能由個別公司壟斷，否則可能會形成令人感到恐懼的後果。可參見中央社「OpenAI 執行長美國會聽證」，2024 年 1 月 4 日下載自:
https://www.cna.com.tw/news/ait/202305170007.aspx