**附件8 可信賴AI檢核表**

**可信賴AI檢核表**

子計畫名稱： 3M Maker: 使用多模式、多任務、多事例學習開發可信賴多媒體處理技術 (下稱本計畫)

子計畫主持人：         廖弘源

資料庫或AI模型名稱/編號：\_\_\_\_\_\_M301\_\_\_\_\_\_\_\_

（備註：一個資料集或AI模型請填寫一份附件表格）

填表人： 陸玉霖

填表日期： 111 年 8 月 15 日

|  |  |  |
| --- | --- | --- |
| **檢核項目** | **問題** | **檢核結果與說明(若檢核結果為「是」，請提供說明；若檢核結果為「否」或「不適用」，請說明原因)** |
| **可解釋性 (Explainability)** |
| **準確性** | 1 | 你是否可確保用於開發AI模型的資料（包括訓練、測試資料）是最新、高品質、完整且在預計部署系統的環境中具有代表性？ | ■是 | 說明：所提供模型可達到SOTA |
| □否 |
| □不適用 |
| 2 | 你是否已建立監控及記錄AI模型準確性之步驟？ | ■是 | 說明：已建立在程式碼中 |
| □否 |
| □不適用 |
| 3 | AI模型的準確性程度是否可能導致危急、不利或破壞性的後果？ | □是 | 說明：合理使用下不會造成危害 |
| ■否 |
| □不適用 |
| 4 | 你可曾考慮過，AI模型的運作是否會使資料或原本訓練時所設定的假設無效？若會，可能導致何種不利影響？ | ■是 | 說明：測試資料與訓練資料分布不一致時可能導致性能下降 |
| □否 |
| □不適用 |
| 5 | 你是否訂定程序，以確保能向終端使用者妥適溝通、傳達其所期待的AI模型準確性程度？ | ■是 | 說明：可透過開源平台進行交流 |
| □否 |
| □不適用 |
| **可解釋性** | 6 | 你是否能向使用者解釋AI模型的決策？ | ■是 | 說明：開發視覺化方法以展示模型可解釋性 |
| □否 |
| □不適用 |
| 7 | 你是否有機制能持續調查使用者對AI決策的了解程度？ | ■是 | 說明：可透過開源平台進行交流 |
| □否 |
| □不適用 |
| **透明性與可追溯性 (Transparency and Traceability)** |
| **可追溯性** | 8 | 你是否可以追溯AI模型使用哪些資料、演算法或規則作出特定決策或建議？ | ■是 | 說明：相關程式碼及權重公開於開源平台 |
| □否 |
| □不適用 |
| 9 | 你是否持續評估AI模型輸入資料之品質？ | ■是 | 說明：相關程式碼及權重公開於開源平台 |
| □否 |
| □不適用 |
| 10 | 你是否持續評估AI模型輸出結果之品質？ | ■是 | 說明：相關程式碼及權重公開於開源平台 |
| □否 |
| □不適用 |
| 11 | 你是否已透過日誌紀錄，將AI模型之決策或建議記錄下來？ | ■是 | 說明：相關程式碼及權重公開於開源平台 |
| □否 |
| □不適用 |
| **透明性** | 12 | AI模型是否被設計用於與終端使用者進行互動(interact)、引導或幫忙作決策？ | ■是 | 說明：開發模型被用於輔助系統提出警示協助使用者進行決策 |
| □否 |
| □不適用 |
| 13 | 你是否已預先模擬AI模型與潛在使用者之間的社會互動？ | ■是 | 說明：開發模型被用於輔助系統提出警示協助使用者進行決策 |
| □否 |
| □不適用 |
| 14 | 你是否建立機制讓使用者知悉其正與人類還是機器互動？ | ■是 | 說明：開發模型被用於輔助系統提出警示協助使用者進行決策 |
| □否 |
| □不適用 |
| 15 | 使用者是否能意識到某項建議或結果是源於AI演算法決策？ | ■是 | 說明：開發模型被用於輔助系統提出警示協助使用者進行決策 |
| □否 |
| □不適用 |
| **個人隱私與資料治理 (Privacy and Data Governance)** |
| **個人隱私** | 16 | 你是否藉由使用或處理個人資料（包括特種個人資料）對AI模型進行訓練或開發？ | □是 | 說明：使用開源資料庫 |
| ■否 |
| □不適用 |
| 17 | 你是否考慮到AI模型對隱私權以及個人資料保護的影響，並建立機制以識別隱私及個人資料保護問題？ | □是 | 說明：使用開源資料庫 |
| □否 |
| ■不適用 |
| **資料治理** | 18 | 你是否建立隱私和個人資料風險評估及管理機制（例如：隱私衝擊評估；個人資料蒐集、處理及利用之內部管理程序；資料近用或修改之權限控管；使用紀錄、軌跡資料及證據保存）？ | □是 | 說明：使用開源資料庫 |
| □否 |
| ■不適用 |
| 19 | 是否採取措施達到隱私保護設計和預設(privacy by design/privacy by default)與資料最少化(data minimization)（例如：加密、假名化及匿名化）？ | □是 | 說明：使用開源資料庫 |
| □否 |
| ■不適用 |
| 20 | 在AI模型開發過程中，對於所蒐集的個人資料，你是否讓資料當事人有撤回同意、選擇退出(opt-out)或資料刪除的機會？ | □是 | 說明：使用開源資料庫 |
| □否 |
| ■不適用 |
| **公平與非歧視性 (Fairness and Non-discrimination)** |
| **公平性** | 21 | 你是否已評估AI模型對終端使用者可能造成不公平的風險？  | □是 | 說明：模型僅產生物件偵測結果，無公平性風險 |
| □否 |
| ■不適用 |
| 22 | 除了終端使用者之外，你是否已識別出可能受到AI模型間接或直接影響的潛在主體？ | □是 | 說明：模型僅產生物件偵測結果，無公平性風險 |
| □否 |
| ■不適用 |
| 23 | 有關公平性的定義，你是否已徵詢利害關係群體（包括弱勢族群等）之意見？ | □是 | 說明：模型僅產生物件偵測結果，無公平性風險 |
| □否 |
| ■不適用 |
| 24 | 針對公平性，你是否有定量分析或量測指標？ | □是 | 說明：模型僅產生物件偵測結果，無公平性風險 |
| □否 |
| ■不適用 |
| **非歧視性** | 25 | 你是否已識別AI模型相關偏誤、歧視或表現不良等問題，並建立一套機制，以避免產生或加劇AI模型中不公平的偏誤（包括輸入資料的使用及演算法的設計等）？ | ■是 | 說明：開發模型較現有模型於各類別表現皆更好S |
| □否 |
| □不適用 |
| 26 | 你是否已藉由公開可得或符合當代技術水準的技術工具，改善你的資料、模型與效能(performance)？ | ■是 | 說明：開發模型較現有模型於各類別表現皆更好。 |
| □否 |
| □不適用 |
| 27 | 你是否已建立教育訓練計畫，幫助AI設計者和AI開發者在設計與開發AI模型過程中更加意識到可能的偏誤？ | ■是 | 說明：開發模型較現有模型於各類別表現皆更好。 |
| □否 |
| □不適用 |
| 28 | 當發現AI模型生命週期中的潛在偏誤時，你是否有明確的處理步驟和問題通報管道（包括如何通報、向誰通報）？ | ■是 | 說明：可透過開源平台進行交流 |
| □否 |
| □不適用 |
| **利害關係人之參與** | 29 | 你是否建立機制， 使最大可能範圍的利害關係人能參與AI模型設計與開發階段（例如：徵詢潛在使用者（包括弱勢族群）之意見）？ | ■是 | 說明：可透過開源平台進行交流 |
| □否 |
| □不適用 |
| **安全性 (safety)** |
| **一般安全** | 30 | 針對AI模型的安全性，你是否建立風險評估機制（例如：風險定義、風險指標、風險等級、風險評估與持續量測流程等）？ | ■是 | 說明：開發模型於不同情境有更好的強健性。 |
| □否 |
| □不適用 |
| 31 | 你是否已評估AI模型的潛在威脅（設計缺陷、技術缺陷、環境威脅）與風險（被惡意使用、濫用或不當使用的風險）及可能產生的後果？ | ■是 | 說明：開發模型於不同情境有更好的強健性。 |
| □否 |
| □不適用 |
| 32 | 你是否告知使用者AI模型之（潛在）風險？ | ■是 | 說明：在開源平台上展示了開發模型的原理及限制。 |
| □否 |
| □不適用 |
| **抵禦資安攻擊之韌性和安全性** | 33 | 你是否已就AI模型可能的弱點或漏洞，評估潛在的資安攻擊模式（例如：資料中毒[[1]](#footnote-1)、模型迴避攻擊[[2]](#footnote-2)、模型逆向攻擊[[3]](#footnote-3)）？ | □是 | 說明：不在考量範圍內 |
| □否 |
| ■不適用 |
| 34 | AI模型在面對設計或技術故障、缺陷、中斷、攻擊、濫用、不當使用或惡意使用等風險或威脅時，是否可能產生不利、嚴重或破壞性之影響？ | □是 | 說明：不在考量範圍內 |
| □否 |
| ■不適用 |
| 35 | 針對AI模型是否有適當的資安管理程序或是否符合特定資安標準？ | □是 | 說明：不在考量範圍內 |
| □否 |
| ■不適用 |
| 36 | 你是否已採取必要措施避免AI模型受到潛在攻擊，並確保AI模型生命週期中的完整性、健全性和整體安全性（例如：紅隊測試、滲透測試或安全性更新）？ | □是 | 說明：不在考量範圍內 |
| □否 |
| ■不適用 |
| **可靠性與再現性** | 37 | 你是否已依AI模型表現的穩定與可信賴狀況，評估AI決策之可靠性？  | ■是 | 說明：開發模型於不同情境有更好的強健性 |
| □否 |
| □不適用 |
| 38 | 你是否已建立相關機制，以便AI模型發生改變時，能重新評估技術穩健性和安全性？  | ■是 | 說明：程式碼中有相關功能 |
| □否 |
| □不適用 |
| 39 | 你是否已建立明確的流程，以監控AI模型是否達到預期目標？ | ■是 | 說明：程式碼中有相關功能 |
| □否 |
| □不適用 |
| 40 | 你是否已測試並確保AI模型的再現性？ | ■是 | 說明：開源完整流程並以確認可重現性。 |
| □否 |
| □不適用 |
| 41 | 在AI模型可靠性和再現性較低的情況下，是否會造成危急、不利或破壞性的後果（例如：影響人身安全）？ | □是 | 說明：合理使用下不造成危害 |
| ■否 |
| □不適用 |
| 42 | 當AI模型得出的結果被評為低可靠性時，你是否已訂定適當程序加以矯正或改善？ | ■是 | 說明：檢討相關問題並調整模型，更新開源平台 |
| □否 |
| □不適用 |
| **自主權與控制權 (Autonomy and Control)** |
| **人為監督** | 43 | 你的AI模型是否有導入人為監督？ | ■無 | 說明：沒有導入人為監督 |
| □有(可複選) | □以Human-in-the-Loop (HITL)[[4]](#footnote-4)方式監督，說明： |
| □以Human-on-the-Loop (HOTL)[[5]](#footnote-5)方式監督，說明： |
| □以Human-in-Command (HIC)[[6]](#footnote-6)方式監督，說明： |
| □不適用 | 說明： |
| 44 | 針對AI模型對使用者所帶來的非預期負面影響，你是否已建立任何偵測和回報機制？ | ■是 | 說明：可透過開源平台進行交流 |
| □否 |
| □不適用 |
| **人類自治和自主性** | 45 | AI模型是否可能有造成人的依賴性、引發成癮行為，或可能操控使用者行為的風險？ | □是 | 說明：模型僅產生物件偵測結果，無成癮風險 |
| □否 |
| ■不適用 |
| 46 | 你是否採取適當措施，以避免AI模型影響人類自主性或讓使用者過度依賴AI？ | □是 | 說明：模型僅產生物件偵測結果，無成癮風險 |
| □否 |
| ■不適用 |
| **共榮共利 (Common Good and Well-being)** |
| **社會包容與多元性** | 47 | 你是否已評估AI模型能對應到社會中各種偏好與能力（例如：AI模型使用者介面也能被特定需求、身心障礙或其他弱勢者使用）？ | ■是 | 說明：開發模型能被大多數使用者(含弱勢)使用 |
| □否 |
| □不適用 |
| 48 | AI模型是否影響人的工作與工作安排（例如：造成人力去技能化的風險(de-skilling risk)，或人員因此需提升新（數位）技能）？ | □是 | 說明：開發模型被用於輔助系統提出警示協助使用者進行決策 |
| ■否 |
| □不適用 |
| **社會福祉與民主** | 49 | AI模型的使用，除了終端使用者之外，是否還會間接影響到其他利害關係人？ | □是 | 說明：合理使用下不影響他人 |
| ■否 |
| □不適用 |
| 50 | AI模型的使用是否對社會整體或民主有其他潛在影響（包括正面與負面影響）？ | □是 | 說明：模型僅產生物件偵測結果，無相關影響 |
| □否 |
| ■不適用 |
| 51 | 你是否採取行動或措施，以降低AI模型可能造成的潛在社會或民主傷害？ | □是 | 說明：模型僅產生物件偵測結果，無相關影響 |
| □否 |
| ■不適用 |
| **問責與溝通(Accountability and Communication)** |
| **溝通** | 52 | 你是否已向使用者說明AI模型之技術限制和潛在風險（例如：準確程度和錯誤率）？ | ■是 | 說明：在開源平台上展示了開發模型的原理及限制 |
| □否 |
| □不適用 |
| 53 | 你是否已建立相關機制，以告知使用者AI模型所生決策之目的、判斷標準和侷限？ | ■是 | 說明：在開源平台上展示了開發模型的原理及限制 |
| □否 |
| □不適用 |
| 54 | 你是否向使用者說明AI模型曾使用哪些訓練資料？ | ■是 | 說明：在開源平台上展示了開發模型的原理及限制 |
| □否 |
| □不適用 |
| **可稽核性** | 55 | 你是否建立促進AI模型可被稽核的機制（例如：開發過程之可追溯性；訓練資源之溯源；AI模型生命週期之相關紀錄；正面及負面影響之紀錄）？ | ■是 | 說明：在開源平台上展示了開發模型的原理及限制 |
| □否 |
| □不適用 |
| 56 | 你是否已盤點與AI模型相關之法律規範？ | ■是 | 說明：並無損害社會之設計 |
| □否 |
| □不適用 |
| 57 | 你是否考慮建立AI倫理審查委員會(AI ethics review board)或類似機制，以探討整體可歸責性與倫理實務，包括潛在灰色地帶的討論？ | □是 | 說明：並無損害社會之設計 |
| □否 |
| ■不適用 |
| 58 | 你是否建立程序或機制，以持續評估並監控AI模型是否遵循本評估表所列之倫理原則？ | ■是 | 說明：相關程式碼及說明公開於github |
| □否 |
| □不適用 |
| 59 | 你是否建立相關程序或機制，以處理不同倫理原則之間的衝突，並解釋最後權衡的決定？ | ■是 | 說明：相關程式碼及說明公開於github |
| □否 |
| □不適用 |

1. data poisoning，即對訓練資料之操控。 [↑](#footnote-ref-1)
2. model evasion，即順應攻擊者的想法對資料進行分類。 [↑](#footnote-ref-2)
3. model inversion，即回推模型參數。 [↑](#footnote-ref-3)
4. HITL所指涉的是人類在系統每個決策週期中干預的能力。 [↑](#footnote-ref-4)
5. HOTL所指涉的是人類在系統設計週期中干預並監視系統運作的能力。 [↑](#footnote-ref-5)
6. HIC所指涉的是人類監督AI系統總體活動（包括更廣泛的經濟、社會、法律和倫理影響）的能力，以及在任何特定情況下，決定何時及如何使用AI系統的能力。 [↑](#footnote-ref-6)