附件 2

中央研究院
生成式 AI 風險研究小組
研究報告

民國 113 年 6 月

摘要

本報告針對生成式 AI（Generative Artificial Intelligence, GAI）風險及其管理機制，以及本院研究人員以中研院名義從事相關學術活動之規範等問題，進行跨領域研究，並提出相關建言。首先，簡述本研究小組成立的源由及運作方式（貳、背景說明與運作紀要）。其次，以詞庫小組（Chinese Knowledge and Information Processing, CKIP）事件背景為基礎，分析系統錯誤回答大型語言模型（Large Language Models, LLMs）系統在功能缺失的直接成因與系統訓練資料的影響因素（參、事件背景與成因分析）。再之，說明本研究的範圍、分析架構與問題盤點（肆），建立分析框架與探討重點。以此為基礎，本研究以「生成式 AI 的風險分析與可能對策」（伍）子題，分成「著作權等智慧財產權保護」、「個人資料保護與資料治理」、「技術風險治理與使用者保護」、「民主影響評估與公共領域風險治理」、「社會衝擊與環境影響風險治理」五大部分，展開生成式 AI 的風險研究，並提出可能的對策方向與治理方案。繼者，針對本小組對「本院研究人員以本院名義從事相關學術活動的規範」（陸）的看法，

述其要旨。最後結語（柒），展望未來，期許各界共創價值，捍衛民主法治，守護國家安全。

目錄

摘要 **2**

壹、 前言 **7**

貳、 背景說明與運作紀要 **9**

一、 背景說明 9

二、 運作紀要 13

參、 事件背景與成因分析 **16**

一、 CKIP 事件背景概述 16

二、 CKIP 模型錯誤回答的直接成因分析 17

（一） 源自使用端的可能原因：特定提示的曖昧或多義性問題 17

（二） 源自技術本質限制的原因：大型語言模型普遍存在的「幻覺」問題 17

（三） 源自研究端的可能原因：欠缺 LLMs 領域研究的最佳實作規範 18

三、 研究者選擇使用簡中資料進行 LLM 訓練的原因分析 18

（一） 繁體中文訓練資料取得面臨較高的法遵成本 19

（二） 開源平台上之簡體中文資料取得成本低廉 19

四、 使用簡中資料作為訓練資料的可能風險 20

（一） 簡中資料數量整體占比高，對繁中使用社群文化自主帶來挑戰 20

（二） 簡中資料被大量事前審查體制控制，存在觀點偏誤的高度風險 21

肆、 研究範圍、分析架構與問題盤點 **23**

一、 生成式 AI 的意涵與背景資料 23

二、 生成式 AI 的國際及各國規範概觀 25

三、 生成式 AI 的利害關係人與風險識別 29

四、 生成式 AI 的風險範疇與分級機制 31

五、 生成式 AI 的問題盤點與分析框架 35

（一） 對智慧財產權之風險 36

（二） 對資訊隱私權/個人資料之風險 37

（三） 技術風險及對使用者之風險 38

（四） 對民主及社會之風險 39

（五） 對資通安全之風險 40

（六） 對環境生態之風險 41

（七） 學術倫理問題 41

（八） 學校教育問題 42

伍、 生成式 AI 的風險分析與可能對策 42

一、 著作權等智慧財產權保護問題 42

（一） 生成式 AI 對智慧財產權的風險結構 42

（二） 生成式 AI 訓練的著作權保護問題 44

（三） 生成式 AI 產出的著作權保護問題 53

（四） 著作權使用的契約條款與授權機制 55

（五） 生成式 AI 風險防制與著作權法法律效果 56

（六） 生成式 AI 對智慧財產權風險的可能對策 60

二、 個人資料保護與資料治理問題 64

（一） 生成式 AI 對個人資料的風險結構 64

（二） 生成式 AI 對個人資料的事件分析：以 ChatGPT 為例 65

（三） 生成式 AI 風險與個人資料保護課題 71

三、 技術風險治理與使用者保護課題 77

（一） 生成式 AI 的幻覺風險治理 79

（二） 生成式 AI 與危險或暴力建議問題 83

（三） 生成式 AI 與人機互動風險問題 87

（四） 生成式 AI 的系統性風險治理 93

四、 民主影響評估與公共領域風險治理 100

（一） 生成式 AI 對資料整全性衝擊的風險治理 100

（二） 生成式 AI 對多元公共意見形成的風險治理 107

五、 社會衝擊與環境影響的風險治理 114

（一） 生成式 AI 對資通安全性衝擊問題 114

（二） 生成式 AI 與 CBRN 資訊取得問題 118

（三） 生成式 AI 形成之社會倫理觀風險 120

（四） 生成式 AI 對環境生態衝擊風險 123

六、 小結 124

陸、 本院研究人員以本院名義從事相關學術活動的規範 132

一、 本院之因應措施與相關規範 132

二、 本院研究人員以院方名義發表研究成果之規範 135

三、 本院研究人員對外發布大型語言模型之規範 136

柒、 結語 137

捌、 致謝 141

附件.....142

一、 附件 1-本院答覆說明資料.....142

二、 附件 2-凍結案相關提案.....144

三、 附件 3-解凍書面報告（節錄有關生成式 AI 議題之說明）.....148

四、 附件 4-預算審查及朝野黨團協商通過決議－涉及生成式 AI 議題.....150

五、 附件 5-議事錄通過決議－3 個月進度（期中）報告.....158

六、 附件 6-參考文獻.....160

壹、前言

繁體中文語料庫是發展臺灣大型語言模型的重要基礎，除了需要投入資源、積極規劃開發外，其對社會影響與風險管理機制的研究，亦同等重要。民國（下同）112 年 10 月初，中央研究院（下稱中研院或本院）研究人員釋出繁體中文語言模型 CKIP-Llama-2-7b 公開測試，引發爭議，中研院於 112 年 11 月 1 日成立本小組，由中研院研究人員及院外專家組成，院本部學術諮詢總會執行秘書、資訊服務處處長、學術及儀器事務處處長列席，針對生成式 AI 風險及其管理機制、研究開發大型語言模型之著作權保護機制，以及本院研究人員以中研院名義從事相關學術活動之規範等問題，進行跨領域研究，並提出相關建言，完成本報告。

本報告紀錄本小組對大型語言模型暨生成式 AI 所生風險及其管理機制之研究與建言，期與全國各界共同促進臺灣語境生成式 AI 的發展，在關鍵議題上善盡本院研究人員的社會責任。至本院政策決定與法規研修，不在本報告之範圍。

中央研究院生成式 AI 風險研究小組

113 年 6 月

貳、背景說明與運作紀要

一、 背景說明

（一）112 年 10 月 6 日媒體報導：中研院詞庫小組（Chinese Knowledge and Information Processing，下稱 CKIP）釋出可以商用的繁中大型語言模型 CKIP-Llama-2-7b，能作為學術使用或是商業使用，並開放測試網頁，以便所有人測試。

（二）112 年 10 月 9 日，經媒體揭露、立法委員范雲亦於其臉書發文：中研院近日釋出的繁體中文語言模型 CKIP-Llama-2-7b，若不針對問答內容特別限縮，該系統會回覆「國慶日是 10 月 1 日」等中國觀點的說法。此偏誤是因為該語言在微調階段的訓練資料通常是來自於中國的資料集。當日該語言模型對外測試之網頁因發生上述爭議，亦停止對外公開。

（三）112 年 10 月 9 日，本院資訊研究所發表聲明略以：CKIP-Llama-2-7b 並非中研院官方或所方發表的研究成果，而是個別研究人員公佈的階段性

成果。此非臺版 ChatGPT，且跟國科會當時正在發展的繁體中文語言模型—Trustworthy AI Dialogue Engine（TAIDE）無關。由於這是一項個人小型的研究，研究人員已將測試版先行下架，未來相關研究及成果釋出，會更加謹慎。對相關研究的成果，公開釋出前，院內也會擬定審核機制，避免類似問題發生。

（四）112 年 10 月 10 日，本院發表聲明（對個別研究人員發布 CKIP 模型之聲明），摘錄重點如下：

第二點：本院相當重視此事件對社會的影響，將釐清事件是否違反相關規定。本院後續將規劃成立「生成式 AI 風險研究小組」，深入了解 AI 對社會的衝擊，提供研究人員相關指引，避免類似事件再度發生。

第三點：繁體中文語料庫是發展臺灣大型語言模型的重要基礎，本院將整合繁體中文詞知識庫，投入資源並規劃管理機制。

（五）112 年 10 月 12 日，本院院長於立法院教育及文化委員會列席報告業務概況並備質詢，萬美玲、張廖萬堅、黃國書、吳思瑤、陳靜敏、陳培瑜、林宜瑾、范雲等多名立法委員針對 AI 出包事件及本院將成立「生成式 AI 風險研究小組」議題質詢，並針對 113 年度預算審查提出相關提案及主決議，本院答覆如附件 1。

（六）立法院審查本院 113 年度預算時，經教育及文化委員會審查、朝野黨團協商結果，依議事錄通過決議內容，單位預算凍結案涉及生成式 AI 風險相關議題之提案，共計 4 案，提案委員包括：吳思瑤、陳培瑜、陳靜敏、陳秀寶等委員（請參閱附件 2，其中吳思瑤、陳靜敏、陳秀寶等 3 位委員所提提案亦通過主決議），前於 113 年 1 月 18 日向立法院提出解凍書面報告（與生成式 AI 議題有關之說明內容請參閱附件 3）；主決議有 8 案，提案委員包括：林宜瑾、張廖萬堅、萬美玲、范雲、吳思瑤、陳秀寶、陳靜敏等委員，另朝野協商主決議有 1 案，由台灣民眾黨立法院黨團提

出（請參閱附件 4），並於 113 年 4 月 8 日向立法院教育及文化委員會提出進度(期中)報告(內容請參閱附件 5）。

（七）112 年 10 月 17 日，本院院本部第 1094 次主管會報討論事項決議：有關「生成式 AI 風險研究小組」之成立，請學術諮詢總會成立任務編組，邀集院內外相關領域研究人員、專家商討。

（八）112 年 11 月 1 日，本院成立本小組，由學術諮詢總會成立之任務編組，法律學研究所李建良所長擔任召集人，邀集院內外生成式 AI 相關的跨領域研究人員、專家共同研討，小組委員包括：莊庭瑞副召集人（資訊科學研究所）、李育杰委員（資訊科技創新研究中心）、邱文聰委員（法律學研究所）、陳舜伶委員（法律學研究所）、黃泰然委員（國立臺灣科技大學專利研究所）、洪子偉委員（歐美研究所）、蔡宗翰委員（人文社會科學研究中心）、李姿儀委員（國立臺灣科技大學專利研究所）、侯宜秀委員（財團法人台灣人工智慧學校基金會）、許永真委員（國立臺灣

大學資訊工程學系）、劉靜怡委員（國立臺灣大學國家發展研究所），加上院方代表：邱繼輝執行秘書（學術諮詢總會）、陳伶志處長（資訊服務處）、陳建璋處長（學術及儀器事務處），以團隊方式連結資訊科技、人文及社會科學人才進行跨領域研究，期能提出相關建言，與全國各界共同促進臺灣語境生成式 AI 的發展，在關鍵議題上善盡本院研究人員的社會責任。

二、運作紀要

本研究小組共召開 6 次會議，就各次開會時間、討論事項及重要決議，摘述如下：

（一）第 1 次會議於 112 年 11 月 1 日召開，會中就本小組研究範疇、運作方式與目標等進行討論，並確認 3 項主要任務：

1. 研究生成式 AI 風險及其管理機制，並提出建言。
2. 研究開發大型繁體中文語言模型過程中，與出版商間之合法著作權授權機制，並提出建言。

3. 研議本院研究人員以本院名義從事相關學術活動之規範。

（二）第 2 次會議於 112 年 11 月 23 日召開，會中盤點與生成式 AI 有關之風險與問題，並開始就資料來源面臨之「智慧財產權」問題進行討論與意見交換。

（三）第 3 次會議於 112 年 12 月 25 日召開，會中針對本院已訂定之免責聲明，討論如何進一步界定學術與言論自由範疇、合理規範本院研究人員之相關學術活動，以及規範之可行性等；另就生成式 AI 與著作權保護問題進行討論與意見交換。

（四）第 4 次會議於 113 年 1 月 22 日召開，先就報告架構及初稿內容進行討論，將「生成式 AI 幻覺」、「生成式 AI 衍生之文化話語權」納入問題盤點中，並可聚焦於本院研究人員亟待解決之生成式 AI 應用風險問題，優先進行討論。另針對生成式 AI 與「個人資料保護」問題與研討範疇提出相關意見。

（五）第 5 次會議於 113 年 3 月 4 日召開，持續就報告架構及初稿內容進行詳細討論，將於「生成式 AI 的意涵與背景資料」章節闡明生成式 AI 之定義，以及本小組之研究重點收斂聚焦於「大型語言模型」，首要處理文字類型之生成式 AI 風險問題；並將生成式 AI 對環境產生之負面影響（耗能及碳排等）及文化霸權問題，納入「研究範圍、分析架構與問題盤點」及結論等章節適度論述說明。

（六）第 6 次會議於 113 年 5 月 27 日召開，先就本院人員對外發布大型語言模型進行公開測試之規範及擬定免責聲明內容進行討論；接著持續討論書面報告內容，內容有涉及資訊專業內涵之章節（如「伍、生成式 AI 的風險分析與可能對策」，有關技術層面之可能對策），將請具有資訊專業背景之委員協助詳加檢視確認內容。在正式對立法院提交報告前，仍持續以線上方式進行討論及修正報告內容。

經小組成員討論修正確認後，完成本報告。

參、事件背景與成因分析

一、CKIP 事件背景概述

CKIP（Chinese Knowledge and Information Processing, CKIP）是本院資訊所研究自然語言處理的中文詞知識庫小組研究人員開發的語言模型。112 年 10 月初，該小組對外釋出供測試之用的大型語言模型 （ Large Language Models，下稱 LLMs ）。CKIP 是以 Meta 公司可商用的開源模型 Llama-2-7b[1]以及來自中國的 Atom-7b 為基礎，再補強繁體中文處理能力後所研發出的 LLM，其預訓練（pre-training）資料的來源有 Common Crawl 、中英文維基百科、台灣碩博士論文摘要、中央研究院漢語平衡語料庫、徐志摩詩歌全集、朱自清散文經典全集，而微調（fine-tuning）的資料來源則是中國的 COIG-PC 與 dolly-15k 兩個資料集[2]。

將初步研究結果的軟體程式碼在開源平台釋出，開放給資訊科學領域同僚及公眾測試檢驗，並藉此獲得反饋以便進行更新改進的作法，是資訊科學研究領域常見的研究過程。本院資訊所詞庫小組於 112 年 10 月 7 日將第一版 CKIP 在

[1] Llama 是 Large Language Model Meta AI 的簡稱。

[2] 參考 https://github.com/f901107/CKIP-Llama-2-7b#readme. 約略而言，LLM 的預訓練（pre-training）是要讓模型學會語言一般特徵和規則——學會文字接龍，而微調 (fine-tuning)則是再餵養已預訓練的模型其他的資料集，特化其對特定領域的自然語言處理工作能力。

GitHub 平台上公開放，原意亦在徵詢早期測試意見。未料，經部分使用者測試後發現，若提問我國國歌，該模型回答義勇軍進行曲，問其國花，乃曰牡丹，而問我國領導人，則言習近平。

二、CKIP 模型錯誤回答的直接成因分析

（一）源自使用端的可能原因：特定提示的曖昧或多義性問題

CKIP 模型在使用者所預設的脈絡下提供明顯有誤的回答，部分原因可歸結於使用者「提示」（prompts）的曖昧性或多義性，例如提示者（未必是台灣使用者）詢問「誰是我國領導人」，在台灣使用者的主觀脈絡下，我國雖然是指中華民國台灣，但台灣對於國家元首多半稱「總統」，少用或不用「領導人」，CKIP 對於「我國領導人」的提示，很可能選擇以慣稱「領導人」的中國做為該提示所預設之語言脈絡，從而產出與使用者（或對台灣使用者）預期不一致的錯誤回答結果。

（二）源自技術本質限制的原因：大型語言模型普遍存在的「幻覺」問題

CKIP 模型產生錯誤回答的另一個原因也可能是當前所有 LLMs 都可能面臨的「幻覺」（hallucination）問題。LLMs

的技術原理是讓電腦從過去已經存在的大量文本資料中，學習人類使用自然語言的方式，藉此產生能以「文字接龍」的形式模仿人類使用自然語言的能力。因此，LLMs 的回答可能僅在文句上合乎特定脈絡之語法規則，卻與事實有所出入，甚至在本身缺乏對事實正確理解的情況下，全憑文句或有或無的關聯即產出回答的文字（口吻還可能相當自信），此種情形，在實務上並非罕見。

（三）源自研究端的可能原因：欠缺 LLMs 領域研究的最佳實作規範

CKIP 模型給錯答案的現象對 LLMs 而言雖然並不特別，不過 CKIP 模型對國花、國歌、領導人的回答帶有不符合台灣語境與國情的緣由，卻與 CKIP 該模型所使用的微調訓練資料來源為 COIG-PC 與 dolly-15k 直接相關。這兩者都是簡體中文資料集──前者是以北京智源人工智能研究院為首的中國 AI 研究單位編纂的資料集[3]，後者則是在中國情境與脈絡下的知識問答對話資料集[4]。

三、 研究者選擇使用簡中資料進行 LLM 訓練的原因分析

[3] https://huggingface.co/datasets/BAAI/COIG-PC.

[4] https://huggingface.co/datasets/Elliot4AI/dolly-15k-chinese-guanacoformat.

何以 CKIP 這此一「繁體」大型語言模型使用 COIG-PC 與 dolly-15k 兩個來自中國的簡體中文資料集進行微調訓練？使用簡中資料進行大型語言模型的開發，何以會產生偏誤的結果？

（一）繁體中文訓練資料取得面臨較高的法遵成本

台灣使用繁體中文，因此訓練繁體中文的大型語言模型理應直接使用台灣語境的繁體中文資料進行訓練。然而，作為大型語言模型訓練資料的文本，除少數例如法令、公文、單純為傳達事實之新聞報導所作成之語文著作（著作權法第 9 條）明確不受著作權保護外，多數仍是著作權保護的對象，使得在未經著作權人授權之前，以網路爬蟲直接擷取繁體中文文本作為訓練資料的方法，在未能釐清是否屬合理使用之前，可能存在違法重製的爭議問題。

此外，倘若訓練資料內容涉及未經合法公開而可辨識個人身分之個人資料時，亦可能需要符合蒐集個人資料的合法要件，例如取得當事人同意，始能以該資料進行訓練，對於大型語言模型的開發者來說，即面臨較高的法遵成本。

（二）開源平台上之簡體中文資料取得成本低廉

相對而言，COIG-PC 與 dolly-15k 二者都可以在開源平台 GitHub 直接取得，且 dolly-15k 以及部分的 COIG-PC 的

開源授權條款，允許使用者不僅可以重製、散布、修改其資料集，甚至允許商業用途[5]。換言之，這些資料集不僅容易取得，侵害他人著作權的風險較低，且利用其開發而衍生的 LLMs 在後續應用上亦較不受限。

四、使用簡中資料作為訓練資料的可能風險

容易取得、法遵成本較低的資料，並不意味就是好的訓練資料。事實上，易於取得的資料集可能因為本身的資料偏誤（bias），而使訓練所得的人工智慧系統，在判斷與權重上存在系統性的失衡。

（一）簡中資料數量整體占比高，對繁中使用社群文化自主帶來挑戰

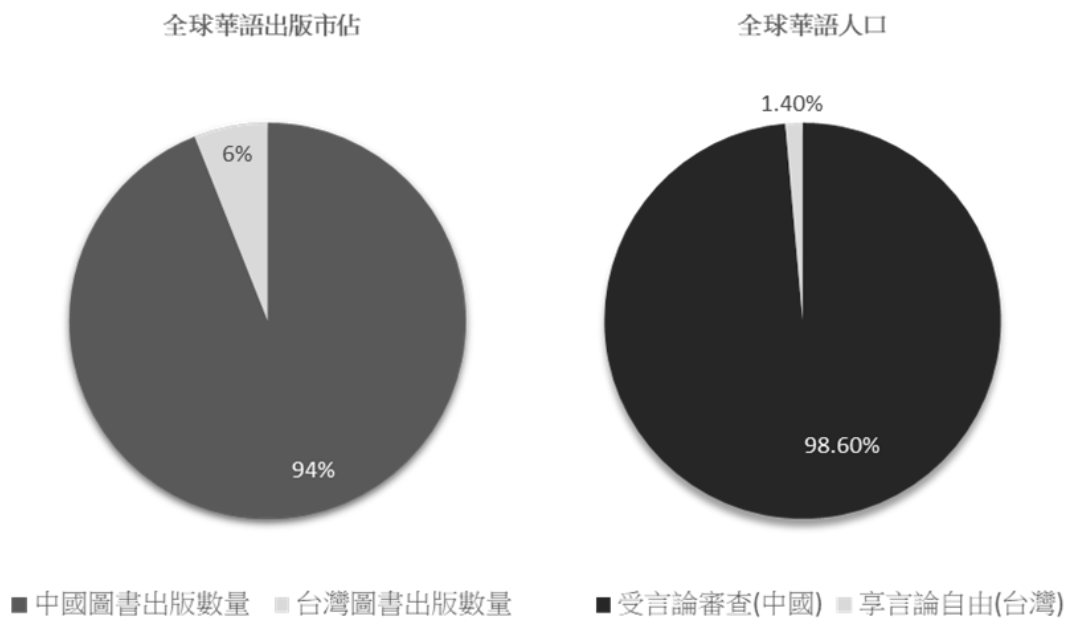
大型語言模型的開發需要極大量的文本作為訓練素材，簡體中文資料的數量在先天上具有優勢（參見下圖）。由於語言的表達本身必然蘊含語言使用者的文化偏好與思維模式，簡中資料數量在全球中文資料的整體占比高，對相對少數的繁中使用社群文化自主帶來挑戰。因此，以簡體中文作為學習範本所開發出的大型語言模型，即使撇開敏感的政治

[5] 針對 COIG-PC 的資料集，CKIP 是從其中 405 個可商用的任務檔案，從中隨機抽取作為微調訓練，見 https://github.com/f901107/CKIP-Llama-2-7b#readme. 此處的開源授權是 Apache 2.0，其與 GNU GPL 自由軟體公眾授權不同之處在於，前者並不要求衍生著作要以同樣授權方式釋出。不過，CKIP 當初仍是以 Apache 2.0 對公眾授權。

議題，也會因為簡體中文使用社群（主要為中國）與繁體中文使用社群（主要為台灣）間幽微的文化差異，而對原本使用繁體中文的社群造成文化衝擊，甚而被同化，對繁體中文使用社群的文化自主與文化多樣性帶來挑戰。

（二）簡中資料被大量事前審查體制控制，存在觀點偏誤的高度風險

相較於因文化差異而產出具有文化衝擊影響的資訊，利用簡中資料開發 LLMs 的更大危險來自於簡中資料的觀點偏誤：以中國為生產來源的簡中資料，其內容的「政治正確性」或經過事前審查過濾，不僅可流通取得的文本需符合「正確意識型態」，更無法公開自由取得被認定為意識形態不正確的文本或文字（參見下圖）。原本大型語言模型在言論自由的預設下，或能以自由市場多元觀點的文本來源，確保「語言模型」除了學會自然語言的使用規則外，不至於嚴重破壞多元社會的平衡。然而，當對岸觀點成為簡中資料唯一存在的觀點時，以簡中資料作為學習對象所開發出的就不單純只是能熟練運用中文自然語言的「大型語言模型」，而是一個會主動投射特定內容的「中國觀點大型模型」。



圖片來源：洪子偉，認知戰：短期干涉と長期浸透，『日本台湾学会報』，forthcoming[6]。

以上簡要分析使用簡中資料作為 LLMs 訓練資料的可能風險，意味著如果不能從開發大型語言模型所需的訓練資料集來源作一處理，對台灣的文化自主和國防安全將造成巨大威脅。CKIP 該模型僅是一個個案，但揭示的問題則是台灣的大型語言模型開發，面臨欠缺高質量、開放授權的本土文字資料的困境。本報告以下將進一步細部分析，開發大型語言模型（或更一般性的生成式 AI）目前所涉及的著作權法、個人資料保護、技術風險、環境衝擊、文化霸權等多面向的挑

[6] 圖左：根據 Statista（2023），中國在 2021 年出版 529,197 筆圖書，而台灣出版 35,041 筆，約佔 6%。圖右：根據 UNESCO 資料，全球約有 15 億華語人口。其中使用繁體中文者約占 2%（香港與台灣）。不受言論審查者只佔 1.4%（台灣）。

戰，以期能指出未來台灣政府發展人工智慧在制度上可以努力的方向。

肆、研究範圍、分析架構與問題盤點

一、生成式 AI 的意涵與背景資料

生成式 AI (Generative AI, GAI) 以機器學習方法求取巨量資料集內各資料單元之間的關聯 (association)，並依此建構資訊系統，自動產出合乎情境需求的（新）內容資料[7]。例如，從大量新聞文本中可學得「台澎金」三個字之後，「馬」字的出現頻率極高。「王金」之後，很可能出現「平」，而且「立法院」、「立法委員」、「院長」等詞可能伴隨於王金平三字前後。生成式 AI 目前已用來處理與生成文本資料，如用於撰寫文書或翻譯，這類方法也用來自動生成圖像影片，經由「提示」(prompt) 文句生成相應的圖像，或是從長時間

[7] 美國總統拜登於 2023 年 10 月 30 日簽署的 Executive Order 14110（Executive Order on Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence）將「生成式 AI」定義為：「AI 模型之一種，透過模擬輸入資料之結構與特徵，產製出衍生之合成內容，包括圖像、影音、文件及其他數位內容。」（The term “generative AI” means the class of AI models that emulate the structure and characteristics of input data in order to generate derived synthetic content. This can include images, videos, audio, text, and other digital content.），可供參照。參見 https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.

的衛星影像中擷取地貌變遷之處。亦即運用於多模態 (multimodal) 資訊的處理。

大型語言模型 (Large Language Models, LLMs) 的使用是生成式 AI 領域最熱門的應用之一。以大量文本訓練，由機器學習不同字詞之間鄰近出現的現象，並以高維度矩陣表示眾字詞間的關聯，這稱為語言模型。大型語言模型可自動生成合乎語法的文句，也因為訓練文本中有許多人類現實的陳述，所生成的文句可以與事實相符，但也可能帶有社會成見，或有所謂「幻覺」（hallucination）現象，陳述不存在或錯誤的事實。大型語言模型以「分佈語意」(distributional semantics) 呈現語詞的關聯（「王金平」與「立法院」常伴隨出現），跟「指稱語意」 (denotational semantics) 講究字詞與意涵的取向（「王金平」與「立法院」所指為何，兩者有何關係），非常不同。

大型語言模型（以及廣義的生成式 AI）的崛起，有以下因素：全球資訊網上可取用的巨量文本、可多頭進行的機器學習方法、高速運算的圖像處理器（Graphics Processing Unit, GPU；用以同時處理大量資料）、以及具親近感的人機互動媒介（亦即自然語言）。一項大型語言模型通常有數十億或

以上的參數，訓練語言模型則需要大量計算（換算為時間與能源成本）與巨量資料（牽涉所用文本的著作權議題）。而語言模型能否貼近應用場域的語言情境跟社會價值，也跟用以訓練模型的語料來源高度相關。

以 2023 年 7 月 Meta 公司釋出的 Llama 2 語言模型為例，其 Llama2-7B 模型具 70 億參數，用以訓練的文本高達兩兆字，但中文（簡體與繁體合計）只佔 0.13%[8]。經巨量文本初步訓練出的語言模型，尚需進行調校（fine-tuning）、依人類偏好強化學習（reinforcement learning from human preferences）、安全圍欄（guardrailing）等繁瑣工作，以提昇其功能表現與可用性。

二、生成式 AI 的國際及各國規範概觀

2022 年 11 月，OpenAI 釋出 ChatGPT，帶動生成式 AI 新一波新的討論與規範，相關機構或單位陸續公布準則 ，有別於傳統類神經網路下的人工智慧，重點是如何面對「基於大型語言模型的生成式 AI」帶來的風險與挑戰，舉其要例：

[8] 參見 Meta, July 18, 2023, Meta and Microsoft Introduce the Next Generation of Llama, https://about.fb.com/news/2023/07/llama-2/.

美國史丹福大學人本 AI 中心（Stanford University Human-Centered Artificial Intelligence, HAI）於 2023 年提出「生成式 AI：史丹福人本 AI 中心觀點」（Generative AI: Perspectives from Stanford HAI）[9]。

世界經濟論壇（WEF）於 2023 年提出「有責任生成式 AI 建議書」（The Presidio Recommendations on Responsible Generative AI ）[10]。

美國國家標準暨技術研究院（NIST）於 2023 年 1 月發布「人工智慧風險管理框架」（AI Risk Management Framework, AI RMF）[11]；再於 2024 年提出「人工智慧風險管理框架：生成式 AI 圖像」（Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, AI RMF Generative AI Profile）[12]。

---

[9] Stanford University Human-Centered Artificial Intelligence (2023), Generative AI: Perspectives from Stanford HAI. Available at: https://hai.stanford.edu/sites/default/files/2023-03/Generative_AI_HAI_Perspectives.pdf.

[10] World Economic Forum (2023), The Presidio Recommendations on Responsible Generative AI. Available at: https://www.weforum.org/publications/the-presidio-recommendations-on-responsible-generative-ai/.

[11] NIST. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). Available at: https://www.nist.gov/itl/ai-risk-management-framework.

[12] NIST. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. Available at: https://airc.nist.gov/docs/NIST.AI.600-1.GenAI-Profile.ipd.pdf.

英國於 2023 年發布「支持創新之工智慧監管作法白皮書」（A Pro-innovation Approach to AI Regulation），有別於歐盟，採由各機關本於職權建立監管機制的規範作法。

英國於 2024 年 1 月 18 日發布「政府機關使用生成式 AI 指引」（Guidance: Generative AI Framework for HMG）[13]，作為公務員和政府組織工作人員[14]安全可靠使用生成式 AI 的指南[15]。

歐盟議會於 2024 年 3 月 13 日通過、歐盟執委會於同年 5 月 21 日批准之《人工智慧法案》(以下簡稱 EU AI Act[16])，採取普遍式的立法模式，並且透過 AI 系統的風險分級（Risk-Based Approach），使不同風險等級的 AI 系統受不同程度的規範。歐盟過去在討論 AI Act 的過程中，並沒有特別針對通用人工智慧（General-purpose AI, GPAI）進行討論，於 2022 年 11 月 ChatGPT 出現後才成為法案的核心問題。面對通用

[13] Cabinet Office and Central Digital and Data Office (2024) Generative AI framework for HMG, GOV.UK. Available at: https://www.gov.uk/government/publications/generative-ai-framework-for-hmg.

[14] HMG（Her/His Majesty's Government），英國政府官方頭銜。

[15] 全文：https://assets.publishing.service.gov.uk/media/65c3b5d628a4a00012d2ba5c/6.8558_CO_Generative_AI_Framework_Report_v7_WEB.pdf.

[16] 全名：Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence，簡稱 artificial intelligence act. 參見 https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf .

AI，風險分級機制有以下問題需要回應，此同時亦是生成式AI作為法律規範客體的考量因素：

1. 風險分級如何應對 AI 的動態應用情況？風險分級的建置是基於預設AI的用途，所謂pre-determined risks，但通用 AI 的應用情況多樣且動態，難以預先對所有可能的應用風險進行預測與分類。

2. 風險分級如何應對 AI 使用者的不可預測性？風險分級主要針對製造商及雇主的責任而為考量；反之，通用 AI 的風險（特別是生成式 AI）多數來自最終使用者的操作與使用方式，而非僅取決於開發者或提供者。

3. 風險分級如何應對 AI 的系統性風險所生的監控需求？風險分級建立風險的預設，但通用 AI（特別是生成式AI）所生之風險有時難以預估或預見，也因此 AI Act 部分轉向採取系統性的風險監控方式（systemic risk monitoring），此部分的發展還有待觀察。

4. 風險分級如何應對算力（Computility）與風險的關聯性？針對通用 AI，歐盟 AI Act 基於算力進行風險評估。例如 ChatGPT 4.0 被認定具有系統性風險，

ChatGPT 3.5 則不是。這種以算力，而非具體應用風險作為劃分方式，是否契合通用 AI 在不同場景中的實際風險差異，有待研究。

5. 風險分級如何應對 AI 對社會與環境影響的程度？通用 AI 可能對社會、經濟和環境帶來廣泛影響，包括自主性削弱、經濟不平等、失業及高能耗等問題。這些問題在傳統風險分級中未得到充分考量，尤其是能源消耗和環境影響方面，需要進行更多的責任揭露和風險緩解措施。

三、生成式 AI 的利害關係人與風險識別

生成式 AI 所生之風險，有來自大型語言模型所生之特有風險，亦有與其他因素相互連結而增高 AI 使用的風險。從利害關係人（Stakeholders）的角度，生成式 AI 風險的利害關係人，主要有：科研人員、行政人員、學生、研究助理、研發過程之受試者／資料提供者、研發成果之應用者／使用者、機構管理者；法律與學術倫理團隊；技術移轉與智慧財產團隊；資訊與安全團隊、機構監督者等，生成式 AI 風險的分析視角，即可統整為個人、團體與生態系統的可能風險。

又，生成式 AI 的風險與 AI 系統生命週期密切相關[17]，包括問題描述、系統設計、系統部署（Deployment）[18]及系統除役等。生成式 AI 的開發與應用可能惡化既有 AI 系統已然存在的風險，亦可能產製出屬於其特有的風險。綜合 AI 的利害關係人與 AI 的生命週期，其主要風險有[19]：對個人的風險：自由、權利、身體或心理的安全、經濟上機會；對團體的風險：對於少數群體之歧視、安全漏洞或金錢損失；對社會的風險：民主參與、教育機會取得，對於產業結構的影響與勞動市場的衝擊；對於生態系統的風險：互相依存、連結的元素及資源全球金融系統、供應鏈或相關系統自然資源等。

---

[17] 依美國國家標準暨技術研究院（National Institute of Standards and Technology，NIST）於 2023 年 1 月 26 日發布「人工智慧風險管理框架 1.0」（Artificial Intelligence Risk Management Framework 1.0，AI RMF 1.0），其中附錄 A：「針對圖表 2 及 3 中 AI Actor 之任務描述」對於 AI 生命週期 中各 AI Actor 之任務進行解釋、描述與舉例。AI 生命週期依 AI RMF 1.0 之圖表二，可分為六大階段：應用內容 Application Context（操作及監控 Operate and Monitor、計畫與設計 Plan and Design）、資料及輸入 Data and Input（蒐集與處理資料 Collect and process data）、AI 模型 AI Model（建立與使用 Build and use model、驗證與確認 Verify and validate）、任務及輸出 Task and Output（部署與使用 Deploy and Use）。

[18] 依上述 NIST 附錄，「AI Deployment（AI 部署）」 係於 AI 生命週期中之「任務與輸出階段（Task and Output）」進行，而進行 AI 部署之 AI Actors 負責關於如何確保 AI 系統用以部署於生產製造之背景決策；相關任務包含系統之先導測試、檢查與原有系統（legacy system）之相容性、確保其合乎法規、管理組織性變化，以及評估使用者體驗。AI Actor 包含系統整合員、軟體開發者、終端使用者、營運商與相關從業人員、評估者、具人為因素專業之領域專家、社會文化分析者與治理。另外，依相關文獻說明，AI Deployment（AI 部署）為將機器學習模型整合至現有之生產製造環境中，以便根據所得之資料做出實際的商業決策。參見 Patnaik, P. (2022). A Policy Framework Towards the Use of Artificial Intelligence by Public Institutions: Reference to FATE Analysis. In Handbook of Research on Artificial Intelligence in Government Practices and Processes (pp. 13-37). IGI Global; Paleyes, A., Urma, R. G., & Lawrence, N. D. (2022). Challenges in deploying machine learning: a survey of case studies. ACM computing surveys, 55(6), 3.

[19] 參考 NIST. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. Available at: https://airc.nist.gov/docs/NIST.AI.600-1.GenAI-Profile.ipd.pdf.

從生成式 AI 的設計、研發、部署到除役，其所生風險從因其所生之風險因系統的多樣與利害關係人的眾多而趨於多元，加上使用、輸入和輸出的範圍廣泛而使評估難度增加，部分尚屬未知；部分已知，卻難以評估辨識所生風險的規模、範疇及複雜度，其主要原因包括用來訓練生成式 AI 的資料不夠透明化，以及監控 AI 風險的技術尚未成熟，隨著生成式 AI 技術的進化而益加複雜，加劇其風險評估的挑戰。

四、生成式 AI 的風險範疇與分級機制

生成式 AI 的風險範疇主要有三：（一）社會倫理與法律風險：1. 智慧財產權問題、2. 隱私和資料治理、3. 機構政策和法規、4. AI 生成的內容的濫用等；（二）技術風險：1. 公平性與歧視性、2. 可解釋性、3. 資訊安全等；（三）學術倫理風險：1. 學術誠信的影響、2. 學術自由的考量、3. 機關聲譽的影響等。具體來說，約可列述如下：

生成式 AI 為 AI 的一種，其風險評估涉及 AI 風險的分級，目前歐盟《人工智慧法案》(以下簡稱 EU AI Act[20])的風險分級，可供參考：

EU AI Act 採風險分級的規範模式，等級越高，管制密度越高，共分成四個風險等級：

（一）不可接受的風險（an unacceptable risk）

（二）高風險（high risk）

（三）有限的風險 (limited risk)

（四）低風險或最小風險（low or minimal risk）

不可接受的風險，指 AI 系統將嚴重侵犯個人的基本權利或違反歐盟的價值觀。EU AI Act 以明列清單方式直接禁止此類 AI 的應用，包括：（一）認知行為操縱 (cognitive behavioural manipulation)；（二）職場與教育機構中的情緒辨識 (emotion recognition)；（三） 社會信用評分 (social scoring)；（四）從網路或閉路電視錄影中無區別地擷取不特

[20] 全名：Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence，簡稱 artificial intelligence act. 參見 https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf .

定使用者之臉部影像建立資料庫；（五）利用生物辨識特徵分類以推斷敏感性個資，例如性取向或宗教信仰等；（六）公部門於公共空間使用即時遠端的生物特徵識別系統(Real-time remote biometric identification system)。最後一項設有例外規定，允許在滿足嚴格必要性且具備合適保護措施下使用此種 AI 系統，例如為預防恐怖攻擊、追蹤嚴重犯罪嫌疑犯或尋找人口販運的受害人等。

高風險 AI，指 AI 系統對於自然人的健康、安全或基本權利將造成較高的風險和危害。對於此種系統的規範，EU AI Act 要求 AI 產品進入市場前須遵行基本權利影響評估(fundamental rights impact assessment)，並對於系統的開發者及設置者訂定強制性規範，分為兩個面向:（一）對於 AI 系統本身的透明性與安全性規範要求，包含符合資料治理及資料品質之要求、技術文件之記錄儲存、透明度以及相關資訊揭露、人為監督、系統準確與穩定性、安全性之規範；（二）對於 AI 系統開發者及設置者的規範要求，即課予 AI 產品價值鏈 (value chains) 提供者（provider）各種相關義務，例如內部控制與審計，要求開發商應主動提供 AI 系統遵行基本

權利影響評估報告，並說明其使用何種資料訓練 AI 模型，以確保該 AI 系統不會造成危害或助長既定之社會偏見。

有限風險的 AI，指與人類互動的系統（例如聊天機器人）、情緒辨識系統、生物分類系統，以及生成或變造圖像、聲音或影音（例如深偽[deepfakes]）的 AI 系統。對此，EU AI Act 係採取課予透明性義務之規範方式。

以上風險等級以外之 AI，則歸類為低風險或最小風險，在歐盟領域的開發與使用無須符合額外之法律義務。不過，EU AI Act 仍建議非高風險 AI 系統的提供者訂定行為準則（codes of conduct），並鼓勵其自願遵循高風險 AI 的強制性規定。

生成式 AI 被歸類為通用 AI，應如何風險分級與監管，歐盟會員國的看法不一，仍在協議之中。根據最新通過的協議內容，通用 AI 亦須納入管制，層次有二：

（一）一般規範：除用於學術研究或在開放資源（open-source licence）發表的系統，所有通用 AI 均納入規範，受透明性規範要求的拘束，包括揭露訓練方法及消耗能源等資訊之義務，並且必須遵守歐盟著作權相關規定。

（二）從嚴規範：針對「具有高影響力」（high-impact capabilities）的通用 AI 課予較高的「系統性風險」（systemic risk）從嚴規範，課予此類系統若干重要的義務（some pretty significant obligations），包括嚴格的安全測試及資通安全檢驗。開發者應揭露系統結構及資料來源等資訊。

對於歐盟而言，任何在訓練階段使用超過 10 的 25 次方 FLOPs (the number of computer operations) 的模型，或訓練模型的電腦算力成本(computing power cost)在 5 千萬美元與 1 億元之間者，被歸類為高影響力，涵蓋的類型包括 GPT-4、OpenAI's、LLaMA 等[21]。

五、生成式 AI 的問題盤點與分析框架

本小組有 3 項任務：（一）研究生成式 AI 風險及其管理機制，並提出建言。（二） 研究開發大型繁體中文語言模型過程中，與出版商間之合法著作權授權機制，並提出建言。（三） 研議本院研究人員以本院名義從事相關學術活動之規範。第 2 項任務亦屬生成式 AI 風險及其管理問題，合併研究；第 3 項任務涉及本院同仁的行為規範，另行說明。

[21] 參見 Gibney, E. (2024). 4What the EU's tough AI law means for research and ChatGPT.Available at https://www.nature.com/articles/d41586-024-00497-8.

如前所述，生成式 AI 的風險涉及 AI 系統生命週期的風險，包括問題描述、系統設計、系統部署及系統除役。生成式 AI 可能惡化既有 AI 系統已然存在的風險，亦可能產製出屬於其特有的風險。因此，生成式 AI 風險的評估與檢視，可以著眼於個別的 AI 系統，亦可超越單一 AI 系統或組織而從「生態系統」的脈絡進行整體的體系觀察。諸如生成式 AI 產生假訊息對民主政治及社會制度的衝擊，或者對於創意經濟、勞動市場或文化領域等各種層面的各式影響。

生成式 AI 的相關風險問題，盤點如下[22]：

（一） 對智慧財產權之風險

1. 主要風險：生成式 AI 可能簡化或便利產生未經授權使用受版權、商標或許可保護的內容、接觸商業機密的途徑，或剽竊、複製相關內容，造成經濟或倫理影響。

2. 被用來訓練生成式 AI 的資料是否包含受著作權或其他智慧財產權保護之資料？是否（須）經授權？

[22] 部分內容，參考 NIST. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. Available at: https://airc.nist.gov/docs/NIST.AI.600-1.GenAI-Profile.ipd.pdf.

3. 生成式 AI 產出資訊是否（須）揭露其訓練之資料是否包含受著作權或其他智慧財產權保護？

4. 生成式 AI 產出的資訊是否受著作權或其他智慧財產權保護？

5. 使用者輸入生成式 AI 之資訊獲得之資訊，是否為使用者之智慧財產權？

6. 針對生成式 AI 如何建立一套完備的智慧財產權保護規範？

（二）對資訊隱私權/個人資料之風險

1. 主要風險：生成式 AI 可能造成生物辨識、健康、位置、個人可識別之資料或其他敏感資料之洩漏、個人資料未經授權而揭露，或個人資料之去匿名化。

2.被用來訓練生成式 AI 的資料是否為個人資料？當事人是否知悉？是否（須）經當事人同意？

3.生成式 AI 產出的資訊是否正確？如何判斷？

4.生成式 AI 產出資訊是否充分告知使用者或利害關係人其如何蒐集及處理資料？

5.使用者輸入生成式 AI 之資訊是否被儲存？儲存何處？是否被進一步作為訓練訓練生成式 AI 的資料？

6.針對生成式 AI 如何建立一套完備的個人資料保護規範？

（三）技術風險及對使用者之風險

因生成式 AI 技術本身及其使用過程所生之風險：

1. 使用生成式 AI 的「幻覺問題」（Confabulation, hallucinations or fabrications）：AI 自信地產生錯誤或虛假的陳述內容。

2. 生成式 AI 可能產出危險或暴力的建議：使產出煽動性、暴力性、威脅性內容更為快速便利，也包含推薦從事自傷、犯罪行為或其他非法活動。

3. 生成式 AI 可能降低或簡化取得化學、生物、放射線、核能（Chemical, Biological, Radiological and Nuclear, CBRN）資訊或其他危險物質有關之重大惡意資訊的門檻或途徑。

4. 人類與 AI 的人機互動與協作（Human-AI Configuration）[23]，可能導致人類對演算法的不信任或過度依賴、基於程式設計或人類驗證的 AI 系統產生欺騙、擬人化之行為，或有人類的濫用、誤用和不安全的再利用等問題。

5. 對於價值鏈與組件整合（Value Chain and Component Integration）的風險，在開發和部署生成式 AI 系統的過程中，可能因資料來源不清楚與供應商審查不當，而導致系統透明性不足與無法歸責下游使用者等問題。

（四）對民主及社會之風險

因生成式 AI 所生對民主及社會的風險，主要是：

1.資訊完整性（Information integrity）風險，可能使未經審查內容的交換與取得更加容易，這些內容可能無法區分事實和意見或辨認不確定性，更可能被用於大規模的假訊息活動。

2. 惡性、偏見與同質化訊息風險：生成式 AI 對公眾接觸惡毒、仇恨性言論或刻板印象內容難以控制，且可能導致

[23] 參見 Kamara, S. (2024) Transforming work with ai: Human-AI configurations, LinkedIn. Available at: https://www.linkedin.com/pulse/transforming-work-ai-human-ai-configurations-seni-kamara-scfdf (Accessed: 13 June 2024).

對某些次族群或英語以外的表現品質下降；又，輸入與輸出的資料因具同質性，將導致輸出品質下降。

3. 猥褻、貶低或虐待性的內容，生成式 AI 讓產生和取得猥褻、貶低或虐待性影像更方便容易，包括合成的兒童性虐待素材（CSAM）和成人非自願親密影像（NCII）。

（五）對資通安全之風險

1. 主要風險：使用生成式 AI 可能使網路攻擊活動更易於進行，可能增加網路攻擊的可用攻擊面，破壞模型權重、程式碼、訓練資料和輸出的機密性和完整性。

2. 使用生成式 AI，可能增加受到社交工程攻擊、網路釣魚攻擊、自動駭客（Automated Hacking）、負載攻擊（attack payload）、勒索軟體、針對 CPU 架構的病毒、創建多態惡意軟體等攻擊之風險

3.使用生成式 AI 工具，以機密或敏感資訊為提示資料，可能造成機密資料有外洩之風險，如何防制？

4..如何防止使用生成式 AI 造成個人資料或機關資料洩漏、財產或經濟利益損失，以及產業、金融、行政或關鍵設施癱瘓？

5.生成式 AI 工具產生「網路釣魚電子郵件」以騙取個人機密資料，如何防制？

6.針對生成式 AI 如何建立一套完備的資通安全規範？

（六）對環境生態之風險

訓練或使用生成式 AI 系統的高耗能，可能對生態系統造成的損害與其他影響結果。

（七）學術倫理問題

1.使用生成式 AI 產出學術成果，是否構成抄襲？或侵害他人著作權？

2.使用生成式 AI 產出學術成果，是否應明確揭露及其使用之工具？

3.使用生成式 AI 產出學術成果，是否應引註資料來源？如何引註？

4.使用生成式 AI 產出學術成果，其內容有偏誤、過時、不正確等問題，應由誰負責？

5.使用生成式 AI 產出學術成果，該工具是否應列為共同作者？

（八）學校教育問題

1.教學現場與教學互動能否使用生成式 AI？

2.教學現場與教學互動使用生成式 AI 是否（須）揭露？

3.是否及如何防止教學現場與教學互動過度依賴生成式 AI 工具？

伍、 生成式 AI 的風險分析與可能對策

承接前述生成式 AI 風險問題的盤點，以下分從智慧財產權保護、個人資料保護、使用者保護、民主影響評估及社會衝擊與環境影響等面向，進一步分析其對個人、團體與社會的風險，並思考技術上與規範上的可能對策。至於學術倫理與學校教育問題，涉及大學自治與教育部職掌範圍，暫不列入本研究報告，合先敘明。

一、 著作權等智慧財產權保護問題

（一） 生成式 AI 對智慧財產權的風險結構

生成式 AI 系統有侵害受版權或註冊商標保護內容、營業秘密或其他授權內容之可能；而此類受智慧財產權保護之內容，多數為訓練生成式 AI 系統之資料，亦即許多下游生成式 AI 應用程式以此為基底所建構之基礎模型。又由於訓

練資料之記憶性質（training data memorization）或產出內容近似但未與受著作權保護之作品完全相同，模型輸出可能因此侵犯著作權；上述相關問題已廣受討論並因網路平台及模型開發者在未補償新聞機構下，大量利用或複製其內容而在新聞領域引起大量關注。生成式 AI 模型僅於利用受著作權保護之作品時，始會產生侵害著作權之問題；於未受著作權保護之情況下，如合理使用則無此類問題。而不受著作權保護之涉及使用未經授權目的之個人身分或肖像問題，亦為一大隱憂；生成式 AI 產出之普遍性與高度擬真本質，可能將進一步削弱人類創作者設計或探索新興作品之誘因。

生成式 AI 的著作權保護課題，主要有：

（一）生成式 AI 軟體是否受著作權法保護？

（二）訓練階段的資料是否侵害他人的著作權？

（三）產出的內容是否侵害他人的著作權？

（四）產出的內容是否受著作權法保護？

（五）產出的內容是否受資料庫建置權的保護？

鑑於生成式 AI 之著作權保護屬新興議題，最理想之方式係借助立法加以解決，包括制定法律（援引他國已完成立

法，而本國尚無之案例），或由相關主管機關訂定指引；其次依現有模式，於大型語言模型訓練開發階段，可以透過契約條款授權機制建立規範。

（二）生成式 AI 訓練的著作權保護問題

生成式 AI 的訓練階段，與著作權保護的問題主要有二：（一）是否構成「重製」[24]？（二）是否為「合理使用」？以現行的著作權法進行認定，還是有另外立法的需要？

情境一：利用爬蟲技術（webscraping）從網路上爬取資料供訓練生成式 AI 之用，是否構成「重製」？一般認為構成[25]，但如何規範，仍為開放性問題。情境二：從貯存於神經網路中的訓練資料中淬取資訊，再進行訓練，有認為不構成「重製」，亦有認為須視情形而定。情境三：蒐集並貯存受著作

[24] 參照著作權法第 3 條第 5 款：「重製：指以印刷、複印、錄音、錄影、攝影、筆錄或其他方法直接、間接、永久或暫時之重複製作。於劇本、音樂著作或其他類似著作演出或播送時予以錄音或錄影；或依建築設計圖或建築模型建造建築物者，亦屬之。」第 22 條第 1 項：「著作人除本法另有規定外，專有重製其著作之權利。」

[25] 參見 Haimo Schack, Auslesen von Webseiten zu KI-Trainingszwecken als Urheberrechtsverletzung de lege lata et ferenda, NJW 2024, 113 ff.; Julia Wulf/Tim-Jonas Löbeth, Text und Data Mining: Wenn gewolltes und geschaffenes Recht auseinanderfallen, GRUR 2024, 737 ff.; Benjamin Raue, Kreativität im Zeitalter ihrer technischen Reproduzierbarkeit: Generative KI als Totengräberin des Urheberrechts? Eine Gedankenskizze, ZUM 2024, 157 ff.; Katharina Wunner, Generative K.I. und das Urheberrecht – Eine komplizierte Beziehung ZUM 2024, 190 ff.

權保護的著作進行訓練，一般認為構成「重製」[26]，但是否為「合理使用」，須視情形而定。

部分國家著作權法針對「資料探勘」設有特別規定，可供參考：

a. 歐盟數位單一市場著作權指令：

■ 第 3 條第 1 項：允許研究組織或文化遺產機構，得基於科學研究目的，對其合法接取之著作或其他內容，進行文字或資料探勘(TDM)。（並於第 7 條中規定牴觸本條之契約條款無效）。

■ 第 4 條第 3 項：允許任何人皆可以為文字資料探勘，但權利人已明示方式保留除外（「退出」機制，the opt-out approach）。本條未限制是否用於商業目的，然權利人可選擇退出。

[26] 參照經濟部智慧財產局 112 年 3 月 17 日電子郵件第 1120317 號函釋：「訓練 AI 利用藝術家未授權著作內容產出生大量相同風格之圖片，在其蒐集訓練資料之過程中，可能會涉及「重製」，若無合理使用之情形，如未取得著作權人之同意或授權，可能構成侵害著作權。」https://topic.tipo.gov.tw/copyright-tw/cp-407-921924-3c00f-301.html.

b. 歐盟人工智慧法案：要求生成式 AI 模型提供者需遵守透明性要求，如應公開其用於訓練 AI 之著作資料[27]。

c. 新加坡《著作權法》[28]（2021 年公布新修正）：允許一定條件下專以數據分析為目的（如：資料與文字探勘、訓練機器學習）者，可以合法重製該受著作權保護之文件[29]。

■ 第 244 條 用於電腦數據分析之重製或傳播

(1) 於符合第(2)項要件之情形下，個人對以下任何材料進行重製，係受允許之利用行為—

(a) 著作；

(b) 受保護表演之錄音。

(2) 要件為—

(a) 製作重製物之目的是為了—

[27] European Parliament (2023) EU AI act: First regulation on artificial intelligence. Available at: https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence (Accessed: 13 June 2024)..

[28] Singapore Copyright Act 2021. Available at https://wipolex-res.wipo.int/edocs/lexdocs/laws/en/sg/sg175en.pdf (Accessed: 2024/01/18).

[29] Intellectual Property Office of Singapore, Copyright factsheet on Copyright Act 2021. Available at https://www.ipos.gov.sg/docs/default-source/resources-library/copyright/copyright-act-factsheet.pdf (Accessed: 2024/01/18).

(i) 電腦數據分析；或

(ii) 準備使用於電腦數據分析之著作或錄音物；

(b) X 並沒有將系爭重製物用於任何其他目的；

(c) X 未向任何人提供該重製物（無論是透過傳播或其他方式），除為以下目的外—

(i) 驗證 X 所進行之電腦數據分析結果；或

(ii) 與 X 所進行之電腦數據分析目的相關的合作研究或學習；

(d) X 得以合法接觸到製作該重製物之材料（在本條中以「原始重製物」稱之）；且

(e) 符合下列要件之一：

(i) 原始重製物並非侵權之重製物；

(ii) 原始重製物屬侵權重製物，但—

(A) X 並不知情；且

(B) 若原始重製物是從明顯公然侵權之網站上所取得的（無論該網址是否因本法第 325 條規定之禁止接觸

命令受限制）— X 不知情也無合理理由可知情；

(iii)原始重製物屬侵權重製物，但—

(A) 對該侵權重製物之利用，具有規範目的之必要性；且

(B) X 並未為其他目的利用該重製物進行電腦數據分析。

d. 英國《著作權、設計與專利法》 第 29A 條[30]：

■ 第 29A 條　用於非商業性研究的文字、數據分析之重製

(1) 個人重製其合法接觸之著作，若有以下情形，不構成著作權之侵害：

(a) 製作該重製物係純粹為非商業性之研究目的，使合法得接觸著作之人，可對該著作所載之任何記錄進行計算、分析，且

[30] U.K., Section 29A, Copyright, Designs and Patents Act 1988. https://www.legislation.gov.uk/ukpga/1988/48/section/29A (Accessed: 2024/01/18).

(b) 該重製物已充分註明著作來源（除有實行上或其他無法註明著作來源之原因）。

(2) 依本條所製作之著作重製物，如有以下情形者，構成著作權之侵害：

(a) 未得著作權人同意而移轉重製物予他人，或

(b) 重製物之利用，已超過第 (1) 項 (a) 款規定之利用範圍，且該利用未得著作權人之授權。

(3) 依本條規定所製作之重製物，就其為後續處分者：

(a) 該重製物將視為為該處分目的侵害著作權之重製物；且

(b) 該處分本身侵害著作權時，則於隨後基於任何目的實施的行為中，該重製物均被視為侵權重製物。

(4) 第 (3) 項所稱之「處分」，係指銷售、出租或為銷售或出租而要約或陳列。

(5) 若以契約條款禁止或限制依本條規定不構成著作權侵害之重製行為，則該契約條款係不得執行的。

e. 德國《著作權法》第 44b、60d 條[31]：

■ 第 60d 條：以科學研究為目的之文字與資料探勘

(1) 為了科學研究之目的，應允許為資料探勘（第 44b 條第 (1) 項及第 (2) 項第一句）之重製，並須符合以下規定。

(2) 研究組織應有權進行重製。研究組織係指大學、研究機構或其他進行科學研究的機構，只要其

1. 追求非商業目的。

2. 將所有利潤再投資於科學研究，或

3. 在政府認可的任務框架內為公共利益而運作。

至於研究組織與對其運作有一定程度影響性並優先獲得科學研究成果之私人企業合作，不適用第 (1) 項規定。

(3) 以下對象亦應有權進行重製：

[31] Act on Copyright and Related Rights (GesetzüberUrheberrecht und verwandte Schutzrechte). Available at https://www.gesetze-im-internet.de/englisch_urhg/englisch_urhg.html(Accessed: 2024/01/18).

1. 圖書館和博物館，只要其向公眾開放，以及檔案館和電影或音訊遺產領域的機構（文化遺產機構）。

2. 個人研究者，只要其不追求商業目的。

(4) 依第 (2) 項和第 (3) 項規定為非商業目的之有權利用人，得向下列人員公開提供第 (1) 項所定之重製物：

1. 共同進行聯合科學研究而專門劃定之群體，以及

2. 以監控科學研究品質為目的之個別第三方。

一旦聯合科學研究或對科學研究品質的監測結束，應立即終止向公眾公開的行為。

(5) 第 (2) 項和第 (3) 項 1 款所定之有權利用人，得在採取適當的安全措施以防免未經授權之利用下，保留第(1)項所定之重製物，只要這些重製物是為科學研究或監測科學發現品質之目的所需。

(6) 權利人得採取必要的措施，以防免其網路和資料庫的安全性和完整性受到第 (1) 項所定重製物之影響。

■ 第 44b 條：文字與資料探勘

(1) 文字與資料探勘是對單個或多個數位或數位化著作之自動分析，以提取資訊為目的，特別是關於模式、趨勢和相關性之資訊。

(2) 允許為文字與資料探勘之目的而重製合法獲取之著作。當文字和資料探勘不再需要這些重製物時，應將其刪除。

(3) 僅在權利人沒有保留之情況下，始允許按本條第 2 項前段之規定使用。至於可在網路上獲取之作品，僅有以機器可讀之形式保留使用權始為有效。

f. 日本《著作權法》第 30 條之 4 第 2 款：

下列各款或其他情形，非以自己或使他人享受著作所表現之思想或感情為目的，於必要之限度內，得以任何方法利用該著作。但依該著作之種類、用途或利用的態樣，如有不當損害著作人利益之情形，不在此限。

(2) 供資訊分析（即從多數著作或其他大量的資訊中，抽出構成該資訊之語言、聲音、影像或其他要素之相關資訊，進行比較、分類或其他的分析）所用之情形。

（三）生成式 AI 產出的著作權保護問題

生成式 AI 產出的內容得否利用？權利歸屬？繫於著作權法上如何定位，且須視情況而定，基本案型如下：

1.產出的內容是否侵害他人的著作權？

生成式 AI 產出的內容是否侵害他人的著作權，與訓練的資料、提出的問題及輸入的資料相互關聯。

情境一：使用者提出的問題若是要求 ChatGPT 尋找特定之文本、旋律或歌詞，而 ChatGPT 給出的文本、旋律或歌詞若尚未正式對外公布，即有侵害該著作權之可能。

情境二：使用者將特定受著作權保護之著作讓 ChatGPT 進行翻譯，則產生之譯本，其著作權仍屬原本著作之著作人[32]，該譯本不得自由使用。

[32] 參照著作權法第 3 條第 1 項第 11 款：「改作：指以翻譯、編曲、改寫、拍攝影片或其他方法就原著作另為創作。」；第 28 條：「著作人專有將其著作改作成衍生著作或編輯成編輯著作之權利，但表演不適用之。」

情境三：使用者要求 ChatGPT 將某一影音（例如電影片）進行文字描述，包括各種虛構角色及其關係的描述，該文字的著作權原則上仍屬該影音著作權人所有[33]，不得自由使用。ChatGPT 對於該影音的描述越詳盡，侵害原作著作權的可能越大。

2.產出的內容是否受著作權保護？

生成式 AI 產出的內容是否受著作權法保護，問題爭點在於是否具「創作性」（原創性）[34]？創作是否限於「自然人」？

目前各國對於 AI 生成之作品是否享有著作權，多半持否定見解，僅中國法院承認使用生成式 AI 的產出受著作權保護。

3.產出的內容是否受資料庫建置權的保護？

生成式 AI 產出的內容是否受資料庫建置權的保護，一般持否定見解，理由是資料庫建置權的保護客體僅限於「具

[33] 參照著作權法第 6 條：「就原著作改作之創作為衍生著作，以獨立之著作保護之。衍生著作之保護，對原著作之著作權不生影響。」

[34] 參照著作權法第 3 條第 1 項第 1 款：「著作：指屬於文學、科學、藝術或其他學術範圍之創作。」

有創作性」之資料庫[35]，未具創作性之資料庫於現行著作權法中不受著作權保護。

（四）著作權使用的契約條款與授權機制

除了在著作權法的規範架構與法院解釋的司法實踐下，生成式 AI 的著作權問題亦可透過契約條款（使用規則）或授權機制建立規範。例如 Open AI 有關 ChatGPT, DALL·E 與其他服務的使用條款說明(2023 年 11 月 4 日更新，將於 2024 年 1 月 31 日生效）[36]，訂有如下規定：

（一）使用者必須擔保其有權或經同意得提供其輸入之內容。

（二）約定使用者擁有 Open AI 提供的服務所輸出之內容，Open AI 也同意將對於其所擁有的輸出內容的權利或利益（如果有的話）移轉給使用者。約定前述輸出內容權利或利益的移轉，不及於其他使用者所得的輸出內容。

（三）約定 Open AI 對於使用者輸入之內容或所得之輸出內容，在遵守相關法律下，得用於提供、維持、發展、改

[35] 參照著作權法第 7 條：「就資料之選擇及編排具有創作性者為編輯著作，以獨立之著作保護之。」

[36] Terms of use (2023) OpenAI. Available at: https://openai.com/policies/terms-of-use/ (Accessed: 13 June 2024).

進其服務。表示使用者可以選擇退出（opt-out）而不提供其輸入與輸出之內容作 LLM 的訓練，目前使用者可透過 Open AI 的 Privacy Request Portal 提出請求[37]。。

以上條款是有關使用者輸入內容或得到生成式內容和服務提供者的約定的階段，這些內容可用來訓練，但不包括開發階段或 fine tuning LLM 時抓取公開的資料或主動向第三人取得資料的階段，這種情況 Open AI 的條款並不適用。

（五）生成式 AI 風險防制與著作權法法律效果

（一） 我國《著作權法》相關規定：

1. 生成式 AI 訓練階段：

(1) 可能侵害著作人之「重製權」：我國《著作權法》第 22 條 1 項規定著作人專有重製其著作之權利（同法第 3 條第 5 款）；且生成式 AI 訓練階段之資料儲存非本法所規定之暫時性重製，亦不符合本法第 44 條至 65 條之重製之合理使用例外規定。因而生成式 AI 之開發者於訓練階段，可能侵害著作人之「重製權」。

[37] 參見 OpenAI Privacy Request Portal (2023). Available at: https://privacy.openai.com/policies (Accessed: 13 June 2024)..

(2) 侵害著作人之重製權的罰則：我國《著作權法》第 91 條 1 項規定，「擅自以重製之方法侵害他人之著作財產權者，處三年以下有期徒刑、拘役，或科或併科新臺幣七十五萬元以下罰金。」由此可知，本法對於侵害重製權設有刑罰規定，而由本小組各委員討論後認為有過苛之疑慮，因而或可修訂該項以移除刑罰規定、改以民事賠償加以規範之。

(3) 生成式 AI 之訓練是否符合我國《著作權法》第 65 條、91 條 3 項之「合理使用」，須個案判斷並委由實務發展。

2. 生成式 AI 產出階段：

(1) 產出的內容是否受著作權保護？

依最高法院 104 年度台上字第 2980 號刑事判決[38]、經濟部智慧財產局電子郵件 1120317[39]與《著作權法》第 3 條，當作品具有「原創性」與為「人類精神之創

[38] 最高法院 104 年度台上字第 2980 號刑事判決：「所謂『原創性』，廣義解釋包括『原始性』及『創作性』，『原始性』係指著作人原始獨立完成之創作，非抄襲或剽竊而來，而『創作性』，並不必達於前無古人之地步，僅依社會通念，該著作與前已存在之作品有可資區別，足以表現著作人之個性為已足。」

[39] 經濟部智慧財產局電子郵件 1120317。https://topic.tipo.gov.tw/copyright-tw/cp-407-921924-3c00f-301.html

作」時，始受著作權法保護。因而現今由生成式 AI 獨立產出之作品因非人類精神創作，應不受《著作權法》保護。

(2) 產出的內容是否侵害他人的著作權？

若生成式 AI 產出之內容與原著作構成實質近似，而有「抄襲」之疑慮時，依我國經濟部智慧財產局電子郵件 980312b[40]，須視個案情形加以認定，並以相似性之「質」與「量」之考量加以判斷之；又或可能會被認為係「衍生著作」，因而有侵害著作人依我國《著作權法》第 28 條所專有之「改作權」之疑慮。因而侵權人可能依本法第 88 條負損害賠償責任、第 88 條之 1 而銷毀該作品或為必要處置、第 89 條負擔費用並將判決書內容登載新聞紙或雜誌、第 91 條 1 項之刑罰。

（二） 外國著作權法相關罰則規定：

1. 美國：

(1) 美國《著作權法》504 條[41]

[40] 我國經濟部智慧財產局電子郵件 980312b。https://topic.tipo.gov.tw/copyright-tw/cp-407-852567-07159-301.html

[41] 17 U.S. Code § 504 Remedies for infringement: Damages and profits. Available at https://www.law.cornell.edu/uscode/text/17/504.

著作權人除得請求回復因侵權行為所受之實際損害，亦得請求侵權行為人因可歸責之侵害行為更所取得之利益。該條亦規定法定賠償額為 750 美元以上 30,000 美元以下。

(2) 美國《著作權法》505 條[42]：勝訴之一造得請求對造負擔合理之律師費，但限於已為著作權登記者始得為之。

(3) 美國《著作權法》503 條[43]

(a) 項：依該法提起之訴訟，於訴訟繫屬期間，法院得以合理條件就涉案物進行扣押。

(b) 項：於終局裁判中，法院得諭知將重製務銷毀或為合理之處置。

2. 英國《著作權、設計和專利法 1988》(The Copyright, Designs, and Patents Act 1988)[44]第 96 條 “Infringement actionable by copyright owner”、103 條 “Remedies for infringement of moral rights”規定著作權人於被侵權時

[42] 17 U.S. Code § 505 Remedies for infringement: Costs and attorney's fees. Available at https://www.law.cornell.edu/uscode/text/17/505.

[43] 17 U.S. Code § 503 - Remedies for infringement: Impounding and disposition of infringing articles. Available at https://www.law.cornell.edu/uscode/text/17/503.

[44] The Copyright, Designs, and Patents Act 1988. Available at https://www.legislation.gov.uk/ukpga/1988/48/contents.

得透過損害賠償、禁制令與其他方式獲得救濟。而同法第 107 條 “Criminal liability for making or dealing with infringing articles, &c.”第 5 項亦有關於犯本條上述行為之人之刑罰規定。

3. 德國《著作權法》[45]69f 條亦規定權利人得請求侵權人銷毀非法製作之重製品，而亦分別於 97 條以下與 106 條以下規定侵權之民事與刑事責任（如：106 條規定未經權利人同意，以法律所不允許之方式重製權利人之作品者，得處三年以下之有期徒刑或罰金。)

（六）生成式 AI 對智慧財產權風險的可能對策

除前述以著作權為重心之分析外，針對生成式 AI 對智慧財產權之風險，綜合言之，尚有其他可能的對策與治理方案[46]：

1. 建立審查系統訓練資料庫中之許可證制度，以管理生成式 AI 所生之智慧財產風險。

[45] Act on Copyright and Related Rights (Gesetz über Urheberrecht und verwandte Schutzrechte). Available at https://www.gesetze-im-internet.de/englisch_urhg/englisch_urhg.html (Accessed: 2024/01/18).

[46] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, pp.11-14, 19-26, 31-32, 35-36.

2. 於風險治理與監管上，應制定透明政策和流程，記錄訓練資料與生成資料的來源，包括著作權、授權許可，以利進行內容來源管理，並使生成式 AI 的使用符合現行的法律和政府政策，例如與資料使用、出版或散布、專利、商標、著作權等相關法律和政策。

3. 根據風險排序建立生成式 AI 系統清單，其項目除通用模型、治理與風險資訊外，尚須考慮智慧財產權之授權作品、關於個人、特許、專有或敏感資料的特殊權利，並依智慧財產權之類型及所涉第三方之相關風險，進行分類，且須建立生成式 AI 系統明確的使用規則與服務條款。

4. 建立與不同 AI Actors 的聯繫管道與互通程序，用於溝通和報告 AI 系統風險，以及使用或服務條款之適用情況。適時針對既定服務和使用條款，研修政策和程序，並監控條款的遵守情況。

5. 於運行與監督層面，應制定政策、指導方針和流程，建立獨立專家審核機制，模型、資料來源、許可證、

演算法和其他系統元件，並記錄資料來源、許可證、訓練方法和設計時之權衡因素。

6. 考慮贊助或參與社群活動，如漏洞賞金、黑客松（hackathon）[47]競賽等活動，讓 AI 行動者評估系統效能，包括內容來源管理的穩健性。

7. 積極建立獲取反饋的機制和渠道及追蹤資料和內容來源的透明機制，並定期裁決與落實相關回饋，確實審核並驗證機制之運行。

8. 於開發階段，根據標準軟體實踐開發程式碼，實施可重現性技術（reproducibility techniques）[48]，包含使用有許可和引文的資料，並追蹤和記錄實驗結果，管理軟體環境和依賴關係，以及利用虛擬環境和版本控制，

[47] hackathon，為 hack（程式設計）與 marathon（馬拉松）二詞的合成，指程式設計馬拉松，又稱駭客松，原為程式設計師聚會共同商討如何編寫程式及其應用的活動，後衍生成一項競賽。參見 What's a hackathon?. 1&1 Digitalguide, https://www.ionos.com/digitalguide/websites/web-development/whats-a-hackathon/.例如總統府每年舉辦「總統盃黑客松」，鼓勵公部門及民間單位協力，提供一個公私協力進行共創的平台，讓資料擁有者、資料科學家及各領域專家運用政府開放資料與科技創新方式，共同解決社會問題，以優化政府施政效能。https://presidential-hackathon.taiwan.gov.tw/.

[48] 指不同量測人員使用同一量具（數據、方法、程序）量測同一物件的特性，以確認所獲量測之同一性或變異性。相關文獻，參見 Odd Erik Gundersen and Sigbjørn Kjensmo, State of the Art: Reproducibility in Artificial Intelligence, file:///C:/Users/chenny/Downloads/11503-Article%20Text-15031-1-2-20201228%20(1).pdf.; Odd Erik Gundersen, Improving reproducibility of artificial intelligence research to increase trust and productivity https://www.oecd-ilibrary.org/sites/3f57323a-en/index.html?itemId=/content/component/3f57323a-en.

並維護需求文件；管理模型和人工製品；追蹤 AI 模型版本，並記錄實驗結果[49]。

9. 於驗證階段，驗證內容來源是否符合相關法律法規，例如侵權行為、使用條款約定和其他條件或智慧財產權等問題；制定定期監測 AI 的生成與輸出，發現潛在侵權並實施應對流程。審查與第三方智慧財產權相關的服務協議和合約。

10. 使用可信賴的授權或開源訓練資料，確保有使用專有資料的合法權利；驗證第三方模型符合現有許可，追蹤有出處記錄的、可能侵害智慧財產權的訓練、輸入和輸出資料項的數量。

11. 在持續追蹤與管理上，應實施供應商風險評估框架，以持續評估和監控第三方實體對內容來源標準與技術、相關法律（著作權法、專利法、商標法等）的遵守情況，並且定時更新可接受使用政策、涵蓋第三方專有和開源技術與數據。

[49] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.32.

12. 建立第三方資料與系統的風險管理政策、程序和文件流程，包括開源資料和軟體，記錄涉及第三方資料和系統（包括開源）的事件；識別並記錄高風險的第三方技術，包括開源軟體，並將風險繪圖和衡量計劃應用於第三方和開源系統。

二、個人資料保護與資料治理問題

（一）生成式 AI 對個人資料的風險結構

生成式 AI 系統可能洩露、生成或正確推斷關於個人的敏感資料，例如生物識別、健康、位置或其他個人可識別資訊（personal identifiable information, PII）。以對抗性攻擊（adversarial attacks）為例，LLMs 揭露的訓練資料中包含了公眾領域中的私人或敏感資訊，如電話號碼、代碼、對話和從一個文件中擷取的 128 位元通用唯一識別碼。模型並可藉由拼湊來自各種不同來源的訊息，正確推斷出不在其訓練資料中且使用者未揭露的個人可識別資訊，包括自動推斷個人可能認為敏感的特徵（如性別、年齡或政治傾向）。

對於個人可識別資訊的錯誤或不適當推論，可能導致偏見和歧視，即使這些推論不正確或是為幻覺，對於私人資訊

的推論也會構成風險。例如，生成式 AI 模型可以基於預測推斷出逾越使用者公開揭露的資訊範圍，並且這些推斷內容可能被模型、其他系統或個人作為不當用途或對個人做出不利決定，包括歧視性決定。

生成式 AI 系統的訓練，需要從眾多公開可用的來源中收集大量資料，目前大多數模型開發人員未揭露訓練資料的來源，當其中涉及個人資料時，則會對認定隱私侵害與否的幾個判準形成風險，包括「透明度」、「個人參與和同意」和「目的」。除非訓練資料可供檢查，否則消費者通常無法知道哪些個人可識別資訊或其他敏感資料可能被用來訓練生成式 AI 模型，這也對目前隱私法規的合規性構成風險[50]。

（二）生成式 AI 對個人資料的事件分析：以 ChatGPT 為例

2023 年初春，Open AI 的 ChatGPT 因為隱私權與違反 GDPR 的疑慮在義大利短暫下架。當地的資料保護主管機關 Garante per la Protezione dei Dati Personali (GPDP)在指出 OpenAI 訓練 ChatGPT 所搜集與處理的大量個人資料可能欠缺法律基礎（legal basis），且生成之內容可是不正確的個人

[50] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.6.

資料處理之行為，責令 Open AI 必須在 20 日內對相關疑慮適當處理[51]，否則將課以罰緩[52]。同年 4 月底 OpenAI 在採取 GPDP 認為「已增進透明度與歐盟使用者的權利」的措施後，重新將 ChatGPT 在義大利上架[53]。這些措施包括 ChatGPT 如何處理個人資料的資訊於網站公開，以及提供資料主體反對其資料被處理的退出申請表格。在後者，資料主體或其代理人必須提供相關 prompts，作為 ChatGPT 經召喚後產生其個人資料的明顯證據（clear evidence）；惟即便成功提出申請，根據 OpenAI 的說明，考量到不同國家法律對隱私權與言論自由的取捨，ChatGPT 不必然會將有關該個人資料的內容從其產出中移出[54]。

---

[51] GPDP 強調兩點：第一，ChatGPT 於資料處理過程，違反歐盟《一般資料保護規則》（General Data Protection Regulation，GDPR）第 13 及 14 條。GPDP 認為，OpenAI 於蒐集資料時並未告知 ChatGPT 之使用者與其蒐集、使用之資料來源者其資料將如何被蒐集及運用；再者，ChatGPT 仰賴大量地蒐集、儲存及處理個人資料予訓練演算法（Training Algorithm），似無任何法源依據（OpenAI 網站中亦明白指出其會蒐集使用者與 ChatGPT 之對話、互動以供未來訓練使用，且 OpenAI 之員工亦得檢視該互動內容）。依 GPDP 至今的試驗，ChatGPT 所提供的資訊，並非均為正確或準確的，因而有不準確的個人資訊經 ChatGPT 處理並使用的疑慮，並進而放大錯誤資訊及侵犯使用者之隱私權。第二，雖然 OpenAI 的使用者服務條款中訂有 ChatGPT 之使用者應年滿 13 歲以上之文字，但 ChatGPT 並未設立有效的「年齡驗證機制」，因而使心智尚未成熟之未成年孩童暴露在可能接收不適當的 ChatGPT 回覆之風險中。

[52] Garante per la Protezione dei Dati Personali (2023). Artificial intelligence: stop to ChatGPT by the Italian SA. Personal data is collected unlawfully, no age verification system is in place for children. Available at https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9870847#english.

[53] Deutsche Welle (2023). Italy lifts ban on ChatGPT after data privacy improvements. Available at https://www.dw.com/en/ai-italy-lifts-ban-on-chatgpt-after-data-privacy-improvements/a-65469742.

[54] OpenAI Personal Data Removal Request 的表格以及相關說明可見：https://share.hsforms.com/1UPy6xqxZSEqTrGDh4ywo_g4sk30.

採取這些措施後，ChatGPT 是否就合乎 GDPR 的規範，非無疑問。賦予資料主體退出的機會，並非是 GDPR 第 6 條所列之合法性基礎。事實上，根據 OpenAI 的說明，其蒐集與處理個人資料的合法性基礎是該條第一項 f 款的「正當利益」（legitimate interests）[55]。但是，賦予資料主體退出的機會，與該款追求正當利益的目的所必要的判斷，並無關聯。

ChatGPT 的隱私疑慮以及隱私風險，其他的大型語言模型亦面臨相同挑戰。大型語言模型的訓練仰賴大量資料，而這些資料除非經過特別處理，往往包含個人資料。即便開發者訓練模型的目的僅在了解語言，與個人資料的剖析無關，其「附隨」蒐集與處理的個人資訊，仍未經過資料主體的同意。進一步言之，在網路上自願公開的資訊，是否就不再有隱私問題，亦非無疑。

訓練內容附隨的個人資料，若一律必須事先移除始能作進一步處理，對目的不在開發與個人資料高度關聯應用（如醫療、廣告）的大型語言模型者，勢必會有高額成本。對這類的開發者，以合成資料（synthetic data）來訓練模型，雖然

[55] 參見 OpenAI. How ChatGPT and our language models are developed. Available at https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed.

有其侷限，仍常被看作是降低隱私風險疑慮的可能方法。合成資料是透過演算法生成在統計上與結構上與真實資料相符的資料，但個別資料點（data points）並非真實資料而是人工合成。

除了蒐集資料的合法性外，另一個大型語言模型面臨的隱私議題是使用者利用如 ChatGPT 等大型語言模型挖掘有關個人的資料。對此，除了提供個別資料主體反對其個人資料被處理的機會外，也有以技術手段減低隱私風險的作法。例如訓練模型拒絕回答有關個人或敏感資料的提問，這也是 OpenAI 採取的作法[56]。

除了義大利當局的處理外，2023 年 4 月 13 日，歐盟個人資料保護委員會（European Data Protection Board，EDPB）考量近期義大利針對 OpenAI 與 ChatGPT 之舉措，決定成立專門小組以促進合作與調查。

在美國，人工智慧與數位政策中心（Center for Artificial Intelligence and Digital Policy，CAIDP）於 2023 年 3 月 30 日向美國聯邦貿易委員會（Federal Trade Commission）提出申

[56] 參見 OpenAI. How ChatGPT and our language models are developed. Available at https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed.

訴並請求其針對 OpenAI 展開調查。CAIPD 依其所提出長達 47 頁之申訴書中，主張 OpenAI 與 ChatGPT 具歧視、傷害性之偏見風險，且 OpenAI 亦明知 GPT-4 有此種風險惟仍推出之，以及 ChatGPT 的隱私、透明性（black box）、公共安全、保護孩童與欺騙之疑慮。2023 年 7 月 10 日，CAIPD 再次提出補充資料以促請聯邦貿易委員會進行調查。終於在同年 7 月 13 日，聯邦貿易委員會提出 20 頁之報告正式宣布展開對 OpenAI 之調查(主要為確認其訓練模型是否合乎聯邦對於隱私、個人資料保護與不實廣告之規範、是否以提供錯誤資訊之方式傷害民眾）。

西班牙之資料保護機構 AEPD (La Agencia Española de Protección de Datos)，亦於 2023 年 4 月 13 日宣布向 OpenAI 發起初步調查。伊比利亞美洲個人資料保護網路當局（la Red Iberoamericana de Protección de Datos Personales，RIPD）亦於 2023 年 5 月 8 日考量 ChatGPT 服務之風險下，展開監督與協調。

加拿大個人資料隱私專員公署 (Office of the privacy Commissioner)於 2023 年 4 月 4 日，以 OpenAI 在未經使用者同意下蒐集、使用、揭露個人資訊為由，對其展開調查。

法國資料保護機構 CNIL (Commission Nationale Informatique & Libertés)，經收到 5 件關於 ChatGPT 之投訴（投訴中不乏提及 OpenAI 於使用者註冊時並未徵求其同意使用個人資料、使用者亦無法進入 OpenAI 蒐集之其個人資訊、OpenAI 缺乏透明及公平性且未完整保障使用者之資訊保護權利、ChatGPT 捏造不實資訊等）後，於 2023 年 5 月 16 日決定推出行動計畫以因應之。

波蘭之資料保護機構 UODO (Urząd Ochrony Danych Osobowych)於 2023 年 9 月 20 日宣布將展開針對 ChatGPT 之調查。

德國北萊茵蘭—西伐爾茲邦之資料保護監察官（Landesbeauftragte für Datenschutz und Informationsfreiheit Nordrhein-Westfalen，LDI NRW）與巴登—符騰堡邦之資料保護監察官（Landesbeauftragten für den Datenschutz und die Informationsfreiheit Baden-Württemberg, LfDI Baden-Württemberg）於 2023 年 4 月 24 日及 10 月 26 日向 OpenAI 針對 ChatGPT 之透明度、個人資料處理、資料主體資訊權及個人資料之更正與刪除權提出問題。

（三）生成式 AI 風險與個人資料保護課題

觀察上述事件及各國採取之因應措施，生成式 AI 對於個人資料的風險及其保護課題，主要可分成兩個面向：一是對於生成式 AI 使用者的風險及其保護課題(使用者之保護)；另一是對資料當事人的風險及及保護課題（資料當事人之保護）：

關於生成式 AI 使用者之保護：

1.如何使 AI 的使用者免於接收到錯誤或之訊息？

2.如何保護未成年或易受傷害族群之 AI 使用者？

關於資料當事人之保護：

1.被用來訓練的資料是否為個人資料？

2.是否得以直接或間接方式識別個人之資料？

3.無從識別特定當事人（去識別化）之資料是否即不受個資法之保護？

4.個人資料「訓練」（機器學習）是否為資料的蒐集、處理或利用？

5.個人資料「訓練」（機器學習）的「目的」是什麼？是否隨訓練的不同階段而改變？

6.被用來訓練的個人資料是否經當事人同意？如何經當事人同意？

7.被用來訓練的個人資料「無須」經當事人同意的情形有哪些？

8.是否應「禁止」特定個人資料被用來訓練生成式 AI？如果是，有哪些？是否應設「原則與例外」規定？

9.生成式 AI 的開發者負有哪些個資保護的義務？現行相關法規有無修改的必要？

可能對策與治理方案：

1. 於公私部門針對生成式 AI 的使用建立防護政策與監管機制（例如：告知取得同意機制、IRB 機制等）；實施「測試、評估、確認與驗證」（Test, Evaluation, Validation and Verification，TEVV）[57] 於實際操作環境，以防範資料安全風險。

[57] 參考 NIST. AI TEST, EVALUATION, VALIDATION AND VERIFICATION (TEVV). Available at https://www.nist.gov/ai-test-evaluation-validation-and-verification-tevv.

2. 與其他 AI 參與者（AI Actors，以下稱之）[58]、網域專家與法律專家共同合作，評估生成式 AI 於健康照護、金融及刑事司法等領域之應用，對於與該系統相關之隱私影響與其內容來源，以管制資訊隱私、人類與 AI 配置互動、資訊完整性之風險[59]。

3. 建立隱私及資料治理政策，並完善個人資料保護法制[60]，紀錄有關 datasets 中生物識別、機密性、受智慧財產權保護之資訊或個人、專有、敏感資料之蒐集、使

[58] 依 OECD 之定義，AI actor 指於 AI 生命週期中，發揮積極作用角色者，包含部署或運行、操作 AI 之組織與個人 ，亦為利害關係人的子集合之一，包括：開發者（Developers）、供應者（Vendors）及終端使用者（End Users）。參見 OECD (2019). Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449. Available at https://legalinstruments.oecd.org/api/print?ids=648&lang=en. 依據 NIST 發布的「人工智慧風險管理框架 1.0」之附錄 A，AI Actors 有：1.AI 設計階段（AI Design），例如資料科學家；2. AI 開發階段（AI Development），例如機器學習專家、資料科學家；3.AI 部署階段（AI Deployment），例如系統整合商、軟體開發人員；4.運行與監控（Operation and Monitoring），例如系統操作員、各領域專家；5. 測試、評估、驗證與確認（Test, Evaluation, Verification and Validation，TEVV），例如檢查 AI 系統或其組件，或偵測與修復問題之人員；6. 人為因素相關專家（Human factors professionals）、各領域專家（Domain Expert）、AI 影響評估者（AI Impact Assessors）、與 AI 模型或產品與服務相關之採購者（Procurement）、治理及監督者（Governance and Oversight）等。此外，第三方實體（Third-party entities）、終端使用者（End users）、受影響之個人或社群（Affected individuals/communities）、為特定與管理 AI 風險提供正式或半正式之規範或指導者，例如商業協會、制定標準之組織，以及一般大眾（The general public），亦屬之。參見 NIST. Appendix A: Descriptions of AI Actor Tasks. Available at https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF/Appendices/Appendix_A.

[59] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.46.

[60] 司法院憲法法庭 111 年憲判字第 13 號判決（全民健保資料庫案）作成後，個人資料保護法雖有修正（112 年 5 月 31 日公布），但僅新增第 1 條之 1：「本法之主管機關為個人資料保護委員會（第 1 項）。自個人資料保護委員會成立之日起，本法所列屬中央目的事業主管機關、直轄市、縣（市）政府及第 53 條、第 55 條所列機關之權責事項，由該會管轄（第 2 項）。」並提高對非政府機構違反個資法規定之罰則，個人資料保護的完善法制，尚付之闕如，有待立法完備之。

用、管理與揭露，以管制生成式 AI 對資訊隱私、人類與 AI 配置互動、智慧財產之風險[61]。

4. 建立資料紀錄檢證（document protocols）[62]（如授權、期間、種類）及存取控制機制，以規範包含生物識別、機密性、受智慧財產權保護之資訊或個人、專有、敏感資料，以管理生成式 AI 對資訊隱私、智慧財產之風險[63]。

5. 實施零知識證明（zero-knowledge proofs）[64]，以平衡透明性與隱私，並於不暴露實際數據之情形下有關驗證內容之聲明；此得以降低資訊隱私之風險；驗證生物識別、機密性、受著作權保護、獲授權、獲專利、個人的、專有的、敏感的或受商標權保護的資訊已從生成式 AI 訓練資料中移除[65]。

---

[61] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.46.

[62] 參見 Mehta, S., Rogers, A., & Gilbert, T. K. (2023). Dynamic Documentation for AI Systems. arXiv preprint arXiv:2303.10854. Available at https://arxiv.org/abs/2303.10854.

[63] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.46.

[64] 零知識證明（zero-knowledge proofs），是一種密碼學之加密方法，係用於證明相關資料而不洩漏資料本身之方法。詳見：https://chain.link/education/zero-knowledge-proof-zkp.

[65] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.47.

6. 排定優先事項而提供資源，建立生成式 AI 系統之清單，其項目除通用模型、治理與風險資訊外，還包括如：應用限制、資料來源、事件回應計劃等多方面資訊[66]。

7. 注意使生成式 AI 的使用符合資料隱私、智慧財產權等相關法律與政策。採用加密技術和適當的保障措施，確保資料的安全儲存和傳輸，訓練資料中若涉及個人隱私，應使用匿名化、去識別化技術、差分隱私技術（Differential Privacy[67]）甚至刪除個人可識別資訊，使生成內容與個人連結的風險降至最低[68]。

8. 於 TEVV 環節中涉及個人隱私資料者，須實施網路安全措施，以保護研究資料、生成式 AI 系統及其內容

[66] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.22.

[67] 指一種保護個人隱私的匿名方法，即以數學方法將資料庫中的資料進行匿名化，再將匿名化之資料提供給他人使用，確保匿名化之資料無法回溯於資料庫中的特定個人。此種方法主要是用於各種統計分析及資料大量處理，如資料探勘（Data Mining）、模型訓練（Model Training）等，最早由電腦科學家 Cynthia Dwork、Frank McSherry 等人於 2006 年研發。由於用於機器學習的數據集越大，需要遮閉的辨識點數量越多，則再識別化的風險也就越高，從而匿名化的方法越要細緻。支持「差分隱私」方法者，通常強調此種匿名方法的細緻性，既可去除再識別化的風險，又可保有資料的可分析性。參見 Nelson, B. (2021). Differential privacy-a balancing act. Chalmers Tekniska Hogskola (Sweden). Available at https://research.chalmers.se/publication/523824； Devaux, E. (2022) What is differential privacy: Definition, mechanisms, and examples, Statice. Available at: https://www.statice.ai/post/what-is-differential-privacy-definition-mechanisms-examples (Accessed: 13 June 2024).

[68] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.38.

出處，防止未經授權的存取、破壞或篡改，並事先取得資料當事人之知情同意書，同意書內容應涵蓋：研究之性質、生成式 AI 使用之資訊、目的，以及潛在風險與影響等，並提供退出／請求停止利用（optout[69]）的選項[70]。

9. 應定期進行稽核，監控生成式 AI 產生的內容是否存在隱私風險，並處理可能曝露敏感資料的情況。保有之文件記錄應包含：訓練資料和生成式 AI 所生成之資料的來源與隱私權資訊，確保生成式 AI 生命週期隱私權保護之合法性[71]。

10. 行政部門、立法部門及法學領域應針對生成式 AI 所生之個資保護風險進行縝密而全面的法律分析，並

[69] 參見司法院憲法法庭 111 年憲判字第 13 號判決（全民健保資料庫案）之主文「衛生福利部中央健康保險署就個人健康保險資料之提供公務機關或學術研究機構於原始蒐集目的外利用，由相關法制整體觀察，欠缺當事人得請求停止利用之相關規定；於此範圍內，違反憲法第 22 條保障人民資訊隱私權之意旨。」及理由「次查依前述對事後控制權之說明，就健保署因辦理健保業務而合法蒐集健保資料中個人健保資料，並提供公務機關或學術研究機構原始蒐集目的外利用此一限制個人資訊隱私權之行為，當事人之停止利用權應仍受憲法第 22 條規定保障。但由相關法制整體觀察，卻未見立法者依所保護個人資訊隱私權與利用目的間法益輕重及手段之必要性，而為適當權衡及區分，一律未許當事人得請求停止利用，停止利用應遵循之相關程序亦無規定，亦即欠缺當事人得請求停止利用之相關規定，顯然於個人資訊隱私權之保護有所不足；於此範圍內，違反憲法第 22 條保障人民資訊隱私權之意旨。」

[70] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, pp.37-38.

[71] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.18.

參酌國際規範及其實務運作（例如歐盟 GDPR）[72]，研擬法律規範，建立整全的保護體系。

三、 技術風險治理與使用者保護課題

生成式 AI 的技術風險治理與使用者保護，技術上共通對策方案主要有[73]：

1. 測量並記錄事件應變時間、系統當機時間與系統可用性（availability）：對生成式 AI 系統執行標準測量手段與結構化之人類反饋，以偵測針對安全性與可靠性之影像及傷害；於實施 A/B 測試、AI 攻擊演練（AI red-teaming）[74]、焦點小組（focus groups）或人類建置試驗測試方法時，應使用人類受試者研究條款與其他可用之安全控制手段；辨識並記錄任何關於「機器人

[72] 近期文獻，參見 Paulina Jo Pesch/Rainer Böhme,Verarbeitung personenbezogener Daten und Datenrichtigkeit bei großen Sprachmodellen. ChatGPT & Co. unter der DS-GVO, MMR 2023, 917 ff.; Susanne Werry, Generative KI-Modelle im Visier der Datenschutzbehörden Technische Entwicklungen im Zusammenhang mit KI-Modellen begegnen einer Vielzahl datenschutzrechtlicher Herausforderungen, MMR 2023, 911; Franz Hofmann, Retten Schranken Geschäftsmodelle generativer KI-Systeme?, ZUM 2024, 166; Boris P. Paal, KI-Training mit öffentlich frei zugänglichen Daten im Lichte der DS-GVO-Vorgaben, ZfDR 2024, 129 ff.

[73] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, pp.41, 44.

[74] Red team（紅隊）或 Red teaming（紅隊演練），原是軍隊用來形容模擬的攻擊行動，現則指用來測試 AI 系統的瑕疵或弱點的一種技術，是生成式 AI 風險評估及治理的一項重要工具。參見 Burt, A. (2024) 'How to Red Team a Gen AI Model', Harvard Business Review, 4 January. Available at: https://hbr.org/2024/01/how-to-red-team-a-gen-ai-model (Accessed: 13 June 2024).

技術智慧流程自動化」（robotic process automation, RPA[75]）與自動駕駛交通工具。

2. 進行 AI 攻擊演練，以識別有關安全性與正當性、可信賴性、隱私、有害性與其他風險之傷害及影響；高風險生成式 AI 系統一經部署後，應於安全性與可靠性風險之考量下持續監控之；監控生成式 AI 系統以偵測相較於預期表現及訓練基礎之偏差或異常。以上風險管理行動得有效監控資訊隱私、人類與 AI 配置互動、有害及有偏見與同質化、危險或暴力建議之風險；建立違反使用規範之申訴機制，並對相關生成式

[75] RPA（機器人流程自動化），是一種透過不同技術（如：AI、軟體機器人）為基礎以實現流程自動化之方法，並通常自動執行無須太多精神上心智努力之重複、例行性或乏味之工作任務；因此人類工作者可將資源投入需要創意思維、人類智力判斷或社交技巧之任務上。該技術藉由軟體機器人不間斷、快速、完美的自動化執行特質，不僅得以提高流程表現效能、效率、可擴展性、可審核性、安全性及法規遵循性；亦同時與傳統流程相較下，得以相對較低之成本實施自動化流程，又稱流程機器人。詳見 Hofmann, P., Samp, C., & Urbach, N. (2020). Robotic process automation. Electronic markets, 30(1), 99-106；Ribeiro, J., Lima, R., Eckhardt, T., & Paiva, S. (2021). Robotic process automation and artificial intelligence in industry 4.0 – a literature review. Procedia Computer Science, 181, 51-58, https://doi.org/10.1016/j.procs.2021.01.104.；Chakraborti, T., Isahagian, V., Khalaf, R., Khazaeni, Y., Muthusamy, V., Rizk, Y., & Unuvar, M. (2020). From Robotic Process Automation to Intelligent Process Automation: –Emerging Trends–. In Business Process Management: Blockchain and Robotic Process Automation Forum: BPM 2020 Blockchain and RPA Forum, Seville, Spain, September 13–18, 2020, Proceedings 18 (pp. 215-228). Springer International Publishing；Ivančić, L., Suša Vugec, D., & Bosilj Vukšić, V. (2019). Robotic process automation: systematic literature review. In Business Process Management: Blockchain and Central and Eastern Europe Forum: BPM 2019 Blockchain and CEE Forum, Vienna, Austria, September 1–6, 2019, Proceedings 17 (pp. 280-295).

AI 系統、內容來源之法律規範與標準，保有一定認識與敏銳度。

3. 針對生成式 AI 系統之不利、有害或不正確生成結果之申訴制度[76]，建立有效且可近用性之程序機制，以有利於管理人類與 AI 配置互動、有害及有偏見與同質化之風險、危險或暴力建議之風險。

再就生成式 AI 的個別技術風險使用者保護課題，分述如下：

（一）生成式 AI 的幻覺風險治理

生成式 AI 系統為了滿足使用者提示的目標而自信地產生錯誤或虛假內容的現象，也包括生成的內容偏離來源輸入或是在同一上下文中出現矛盾的內容，這種現象在當前 LLM

[76] 生成式 AI 申訴制度的建置，須區分「公部門」與「私部門」開發之生成式 AI。前者除現行行政訴訟制度可供運用外，可考量建立獨立機關的救濟制度，納入個人資料保護委員會之權限範圍，或於數位發展部下設獨立委員會。後者之制度建置，可分為兩種可能，一是由 AI Actors 自行建立，提供終端使用者申訴的機會；另一是以法律課予 AI Actors 建立申訴機制之義務，並且賦予終端使用者申訴之權利，對於申訴結果不服者，尚可進一步向民事法院提起訴訟，以資救濟。

中普遍存在，其係因生成式 AI 預先訓練（pre-training）中涉及下一個詞彙預測（next word prediction）所造成[77]。

生成式 AI 所生幻覺的風險在於，有時即使其給出的答案本身是錯誤的，但 LLM 還是會提供邏輯步驟來說明如何得出答案，真實世界中的使用者將會相信這種有邏輯的、自信的錯誤內容，並傳遞錯誤訊息或採取後續行動[78]。例如在醫療領域，使用生成式 AI 產出之內容若含有虛構的病情報告，可能導致醫生依該報告作出錯誤的診斷，或者選擇錯誤的治療方式。

可能對策與治理方案：

1. 於部署前之風險測量階段與持續監控活動中，審查與驗證生成式 AI 系統產出內容之來源與引述，以降低虛談幻覺現象（confabulation）風險[79]，並避免衍生個資保護問題[80]。

[77] 關於生成式 AI 所生「幻覺」的成因及其風險分析，參見 Millidge, B. (2023) Llms confabulate not hallucinate, Beren's Blog - Thoughts on AI, Neuroscience, and other things that interest me. Available at: https://www.beren.io/2023-03-19-LLMs-confabulate-not-hallucinate/ (Accessed: 13 June 2024).;. KI-Halluzinationen: Wie ChatGPT die Grenzen von Künstlicher Intelligenz verschiebt (2024) DEINKIKOMPASS.de. Available at: https://deinkikompass.de/blog/ki-halluzinationen-wie-chatgpt-fakten-erfindet (Accessed: 13 June 2024).

[78] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.5.

[79] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.40.

[80] 參見 ChatGPT-Halluzinationen im Visier des Datenschutzes, BC 2024, 192.

2. 驗證生成式AI系統建置得以監測其輸出與表現效能，以及當其偵測到安全性異常、威脅與影響時，得以處理與自錯誤中恢復及修補之，或對於有疑問的文字內容進行警示標記[81]，或內建 SelfCheckGPT[82]，以管理虛談幻覺現象、資訊完整性、資訊安全性風險[83]。

3. 為部署前評估指標，創建測量誤差模型，以確認其各自具有效度（validity）：量測或預估及記錄於應用評估指標或結構化人類回饋流程中的偏見或數值變化[84]。

4. 於運行具大量設定模型下，遵守適用之法律規範；於建立具複雜社會結構相關模型時，應利用相關專業領域知識[85]。

5. 為識別技術及其組成部分的法律風險而制定政策、實踐與記錄，主要措施是實施可重現的技術，包括：使用有被授權和引文的資料、遵循標準軟體開發程式碼、

[81] 參見 Google-Chatbot Bard kämpft gegen KI-«Halluzinationen», indem es zweifelhafte Textstellen markiert (2023), in: https://www.nzz.ch/technologie/google-chatbot-bard-markiert-zweifelhafte-textstellen-ld.1756929.

[82] 參見 Potsawee (9 May 2024), potsawee/selfcheckgpt. https://github.com/potsawee/selfcheckgpt.

[83] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.42.

[84] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.50.

[85] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.50.

追蹤和記錄實驗結果、管理軟體環境和系統間依賴關係、利用虛擬環境和版本控制並維護需求文件、追蹤AI模型版本並記錄詳細資訊與資料管理流程、建立測試驗證流程等[86]。

6. 運用檢索增強生成（Retrieval-Augmented Generation, RAG）技術[87]，持續讓監控 LLMs 取得最新的資料。例如持續餵給 LLMs 模型與訓練文本有關的資料，以確保精準的產出，降低內容的扭曲[88]。

7. 各機構應制定政策、指南和實踐，用於監控機構與第三方影響評估，其關注項目包含：資料、標籤、偏差、隱私、模型、演算法、錯誤、來源技術、安全性、合規性、輸出等，以降低風險和危害[89]。

---

[86] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.32.

[87] 一種結合了搜尋檢索和生成能力的自然語言處理架構，LLMs 模型可以從外部知識庫搜尋相關資訊，再使用這些資訊生成回應或完成特定的 NLP 任務。參見 Simon Liu, RAG 檢索增強生成— 讓大型語言模型更聰明的秘密武器（2023）. Available at https://blog.infuseai.io/rag-retrieval-augmented-generation-introduction-a5854cb6393e.

[88] 參見 Bernhard Lück, Halluzinierende KI – diese drei Maßnahmen helfen (2023), in: https://www.bigdata-insider.de/halluzinierende-ki-diese-drei-massnahmen-helfen-a-d7075ae954a908e831f5ba473385ce69/.

[89] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.19.

8. 各機構就其設計、開發、部署、評估和使用的風險和潛在影響，應進行更廣泛的溝通，根據影響程度和風險承受能力，定期審查機構監控及事件回應計劃[90]。

（二）生成式 AI 與危險或暴力建議問題

生成式 AI 系統會產生煽動性、激進性、威脅性、美化暴力的內容，或是推薦從事自傷、犯罪行為或其他非法活動等內容；有些模型甚至會產生可執行的指示內容，包含如何操縱他人和進行恐怖行為。文字轉圖像模型也使得創造有害內容的不安全圖像變得更加容易，這種類似的風險也存在於其他視聽，如視訊和音訊。許多系統目前在回應以上內容時，都會限制模型輸出這類具危險性的內容，但在使用者特別繞過這樣的設定，而使用較不明顯的提示時，系統仍會生成有害的建議。另外，研究發現從與聊天機器人的對話中，可以看到相當數量的使用者有心理健康問題，然而目前系統卻無法適當地回應、引導或提供對應的幫助[91]。

可能對策與治理方案：

[90] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, pp.13-14.

[91] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, pp.5-6.

1. 評估供應鏈或其他 AI Actors 於生成式 AI 訓練及維護過程中，暴露在猥褻、有害或暴力的資訊下，對於其身心健康之負面影響，以管理人類與 AI 配置互動、猥褻或有辱人格和／或辱罵性內容、AI 價值鏈與內部組件整合、危險或暴力建議之風險[92]。

2. 檢視生成式 AI 系統產出之有效度與安全性：檢測生成之程式碼以評估可能因不可靠之下游決策所生的風險，以管理 AI 價值鏈與內部組件整合、危險或暴力建議之風險[93]。

3. 追蹤及記錄過去失敗的生成式 AI 系統設計，以提供安全性與有效度風險測量相關資訊，此得以管理危險或暴力建議之風險；評估內容審核系統（content moderation）[94]及其他輸出過濾技術或針對人類、系統

[92] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.41.

[93] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.41.

[94] content moderation，指對於使用者所生成之內容在線上進行審核的一種程序，以確認該內容是否合於相關規範。參見 Strandell, J. (2024) What is content moderation? (plus best practices), Besedo. Available at: https://besedo.com/blog/what-is-content-moderation/ (Accessed: 13 June 2024).. 關於著作權之 content moderation，近期尤具爭議，參見 Sebastian Felix Schwemer, Decision Quality and Errors in Content Moderation, IIC 2024, 139 ff. João Pedro Quintais, Christian Katzenbach, Sebastian Felix Schwemer, Daria Dergacheva, Thomas Riis, Péter Mezei, István Harkai, João Carlos Magalhães, Copyright Content Moderation in the European Union: State of the Art, Ways Forward and Policy Recommendations, IIC 2024, 157 ff.

與統計／計算偏差而產生之風險；實施公平性評估以量測系統性偏差[95]。

4. 量測生成式 AI 系統於橫跨人口群體與其子群體之表現，以應對服務品質及服務和資源之分配問題；識別危害類型，包含資源分配、代表性、服務品質、刻板印象，或消除、識別可能受傷害之跨群體、群體內部與交叉群體[96]；使用（重複）刪除訓練資料技術（deduplication of training data samples），降低有害資料的機率[97]。

5. 完善群體內部與交叉群體之差異識別；將影響小群體、群體內部、交叉群體或單一個人之偏見納入考量。上述各行動得管理有害及有偏見與同質化、危險或暴力建議之風險[98]。

---

[95] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.41.

[96] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.47.

[97] 參見 Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., & Carlini, N. (2021). Deduplicating training data makes language models better. arXiv preprint arXiv:2107.06499; Kandpal, N., Wallace, E., & Raffel, C. (2022, June). Deduplicating training data mitigates privacy risks in language models. In International Conference on Machine Learning (pp. 10697-10707). PMLR; Jayant Nehra (2024), Effective Data Deduplication for Training Robust Language Models. Available at https://python.plainenglish.io/effective-data-deduplication-for-training-robust-language-models-44467afac5bb.

[98] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.48.

6. 有關本項風險在治理與監管層面的措施，應訂定政策與程序機制，以確保訓練資料中不含有危險或暴力之內容，例如要求將危險性與暴力性資料從訓練資料刪除。同時，根據風險承受能力，考慮各種因素如濫用風險、對生成式系統輸出的人為審查、對人類心理的影響，而制定風險管理活動水準，並適時重新評估風險容忍度，以回應全球 AI 和生成式 AI 不成熟的安全性或風險文化（risk culture）[99]。

7. 在部署環節，應制定並記錄風險測量計劃，處理以下問題：生成式 AI 系統行動者個人和群體的認知偏差（如確認偏差、經費偏差、群體思考）；已知的過去事件和失效模式；情境中的使用情況和可預見的誤用[100]，以及處理所謂「未經許可之使用」（off-label use）[101]。

[99] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.18. 風險文化（risk culture），指稱具有共同目標的群體對風險的價值觀、信念、知識、態度、作為、行止與理解，以及可接受的風險程度。參見 Rachael Johnson (2023). Risk Culture: Building Resilience and Seizing Opportunities. A Global Survey and Report, p. 9.

[100] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.26.

[101] 特別是在醫療領域，相關文獻，參見 Krishnamoorthy, M., Sjoding, M. W., & Wiens, J. (2024). Off-label use of artificial intelligence models in healthcare. Nature Medicine, 1-3.

8. 識別生成式 AI 系統中內容來源的潛在風險和危害，根據風險評估進行排序，並確定解決方案，而採用適當方法和指標衡量本項風險[102]。

（三）生成式 AI 與人機互動風險問題

人類與AI的人機互動，也就是使用者與 AI 或與 AI 相關事物之間的互動關係，例如生成式 AI 替人類（使用者）生成文章、圖片或影片或檢查文章的錯別字等，涉及不同程度之自動化與人類及 AI 的互動，均可能導致人類的濫用或誤用，風險種類與規模的估計十分困難。人類一方面可以使用AI系統，讓AI系統自行獨立生成決定；另方面可以在AI系統中加入自己的專業設定或系統配置（目標設定或運作方式配置），在人機協作下驅動生成某種人機合成之決定[103]，儘管人類對於該 AI 系統及其運作未必具備詳盡的知識。於此情境下，各種生成式 AI 系統的整合運用，即有可能引發多樣的錯誤配置與不良互動之風險。例如有些人類專家可能會對生成式AI所產出的結果保持強烈偏見或感到厭惡；反之，

[102] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.35.

[103] AI 系統與人類參與的配置關係同時也是生成式 AI 與智慧財產權之關聯問題。換言之，人類對於 AI 系統投入的配置越高，對於 AI 系統產出之內容享有智慧財產權的可能性越高。參見 Gräfe, H. C., & Kahl, J. (2021). KI-Systeme zur automatischen Texterstellung, Urheber-und medienrechtliche Einordnung von Textgeneratoren in Journalis mus und E-Commerce. Multi Media und Recht Zeitschrift für IT-Recht und Digitalisierung, 24(2), 121-126.

有些人類專家則可能因生成式 AI 技術的複雜性及其可信賴度，而逐漸習慣且過度依賴生成式 AI 系統，此種人類過度依從於 AI 系統的現象，被稱為「自動化偏見（automation bias）」[104]。此外，AI 系統的目標出於開發者或使用者無心的錯置或誤設，也可能造成模型無法預期運作。依據研究，儘管 AI 模型已經應用安全標準技術，用以校正該系統生成的內容，但在人為因素介入下，仍然可能持續產出欺騙性的輸出[105]；AI 模型本身具有欺騙能力，是該領域新興的風險，但有心人士也可能利用提示（prompt）進行欺騙行為而導致其他風險[106]。

可能對策與治理方案：

1. 結合人類評估者於內容品質與相關性過程評估中，以便管理人類與 AI 配置互動的風險[107]。

[104] 自動化偏見（automation bias）問題於上世紀末、21 世紀初即見探討，例如 Kathleen Mosier et al, "Automation Bias and Errors: Are Teams Better than Individuals?". Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 42 (3): 201-205 (1998).; M.L. Cummings, Automation Bias in Intelligent Time Critical Decision Support Systems (2004). 近期文獻極多，例如 Vered, M., Livni, T., Howe, P. D. L., Miller, T., & Sonenberg, L. (2023). The effects of explanations on automation bias. Artificial Intelligence, 322, 103952.

[105] 參見 Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., ... & Perez, E. (2024). Sleeper agents: Training deceptive llms that persist through safety training. arXiv preprint arXiv:2401.05566.

[106] 以上參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.7.

[107] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.52.

2. 利用分解之評估方法（disaggregated evaluation methods）[108]，例如以種族、年齡、性別、族裔、能力、區域等，提高 AI 系統表現效能之資料細緻度[109]；於規範的層面，確立（終端）使用者保護（end-user protection）或終端使用者安全（end-user security）的觀念[110]；強化人類使用科技的能力與自主決定的能力；建立使用生成式 AI 產製內容之檢驗程序與機制：備置生成式 AI 除役計畫。

3. 建置資料履歷與來源追索系統（data provenance, data-lineage）[111]，驗證使用者回饋蒐集機制的可追溯性，以降低人類與 AI 配置互動間之風險；與相關 AI Actors，例如對於系統發布有核准權限之參與者，分

[108] 參見 Barocas, S., Guo, A., Kamar, E., Krones, J., Morris, M. R., Vaughan, J. W., ... & Wallach, H. (2021, July). Designing disaggregated evaluations of ai systems: Choices, considerations, and tradeoffs. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (pp. 368-378).

[109] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.39.

[110] 歐盟執委會於 2023 年 7 月 3 日通過「歐盟永續報告準則」（European Sustainability Reporting Standards, ESRS），作為邁向經濟永續目標的準據，其中針對資訊之消費者及終端使用者設定保護的目標，可供參考，見 The Commission adopts the European Sustainability Reporting Standards, https://finance.ec.europa.eu/news/commission-adopts-european-sustainability-reporting-standards-2023-07-31_en. 相關文獻，參見 Jens Reinke/Stefan Müller, Verbraucher und Endnutzer in der Nachhaltigkeitsberichterstattung nach ESRS S4, IRZ 2023, 545 ff.; IRZ 2024, 35 ff.

[111] 歐盟 AI Act 課予 AI 系統的資料治理義務中將包含此項義務，參見 Philipp Müller-Peltzer/ Valentin Tanczik, Künstliche Intelligenz und Daten. Data-Governance nach der geplanten KI-Verordnung, RDi 2023, 452 (456 f.).

享部署前測試之結果，以制衡人類與 AI 配置互動間之風險[112]。

4. 持續追蹤與紀錄於生成式 AI 系統介面出現之擬人化實例（anthropomorphization）[113]，如人類影像、人類感受、生化人圖像或圖案，以管理人類與 AI 配置互動間之風險；追蹤與紀錄 AI Actors 之身分驗證資訊與資格證明[114]，以降低人類與 AI 配置互動之風險。

5. 本風險之措施重點在界定人力監督 AI 系統的角色和責任，確保 AI 參與者的風險能被妥善處理。

6. 向終端使用者揭露生成式 AI 系統的使用情況，訂定下游 AI 行動者的生成式 AI 風險接受機制及程序，並確保事件揭露計劃包含足夠的 AI 系統背景資訊，以便實施補救措施[115]。

[112] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.39.

[113] 參見 Deshpande, A., Rajpurohit, T., Narasimhan, K., & Kalyan, A. (2023). Anthropomorphization of AI: opportunities and risks. arXiv preprint arXiv:2305.14784; Placani, A. (2024). Anthropomorphism in AI: hype and fallacy. AI and Ethics, 1-8.

[114] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.40.

[115] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.16.

7. 在系統使用的範圍內，應界定相關利益群體，如：人口統計群體、領域專家、過去使用生成式 AI 技術的經驗，作為收集結構化的公眾回饋之一部分[116]。

8. 建立生成式 AI 系統之清單（GAI system inventory entries），以排定優先事項而提供資源，其項目除通用模型、治理與風險資訊外，還包括以下項目（這些項目有助於全面管理和監控生成式 AI 系統，以確保其安全、透明及可解釋性）：允許和禁止的使用情況及例外;受限制的用途;與其他系統間的相互依賴性;人類監督的角色和責任：明確人類監督在系統中的角色和責任；基礎模型、模型的版本及存取模式；更新的已識別及預期風險的層級，與使用情境相關[117]。

9. 在設計系統時，相較於傳統 AI，應考慮：在大型或複雜的生成式 AI 系統生命週期的各個階段，是否需調整組織角色和組成部分，包括與生成式 AI 系統相關的 AI Actors、使用者和社群回饋；同時須定義各類生成式 AI 介面、模式和人機配置的可接受使用政策，

[116] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.36.

[117] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, pp.14-15.

並為生成式 AI 系統中的使用者回饋機制制定政策。為 AI Actors，終端使用者和利害關係人提供內容來源風險管理教育也是此階段的一部分[118]。

10. 在系統部署階段，須記錄風險測量計劃，處理個人和群體的認知偏差、已知系統事件和失效模式、預期用途和可能的濫用情況等；同時記錄系統需求、所有權、AI Actors 的角色責任、行動者的資格與團隊組成，透過建立跨領域團隊所開展的風險評估，持續改進團隊的多樣性和代表性，以驗證反饋來源的代表性。此階段中應與具代表性的各種 AI Actors 進行協商，並執行獨立稽核、影響評估等結構化的人類反饋程序。

11. 影響評估需考慮 AI 的映射風險（mapped risks）[119]，如複雜安全需求、用戶情感糾葛的可能性等[120]。記錄內容源匿名機制在隱私安全背景下的運作，實施網絡安全措施，對涉及人體研究使用機構審查委員會，採用匿名化和差分隱私技術將風險降至最低，驗證研究

[118] 參考 NIST (2024), AI RMF Generative AI Profile, Initial Public Draft, p.35.

[119] Risk Map 或 Risikolandkarte，指評估風險發生機率及衝擊程度，並將風險等級與風險控管優先順序予以結構化與圖像化。參見 Werner Gleißner, Die Welt der Risiken: 5 Risikokategorien, BC 2022, 217 ff.

[120] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.33.

對象的人口統計代表性，這些都是影響評估中需要關注的重點。進行 AI 影響評估，估計可能性、影響程度和風險時，需考慮生成式 AI 特有的映射風險，如複雜的安全要求、使用者情感糾葛的可能性、大型供應鏈等問題。以人作為研究對象進行影響評估時，使用 IRB 機制，以保護個人的權益和安全，同時確保使用的樣本代表該研究所關注的相關族群[121]。

12. 針對高風險 AI，研擬包含減緩、轉移、避免或接受之應對措施，例如：紀錄不超出組織風險耐受度之風險權衡、決策過程與相關量測及反饋結果；監控風險控制之有效性，例如：透過現場測試、參與式互動、績效評估、使用者反饋機制，以管理人類與 AI 配置互動風險。

（四）生成式 AI 的系統性風險治理

生成式 AI 系統的價值鏈（value chain）[122]，通常涉及第三方元素，如採購之資料庫、已訓練之模型與軟體庫；上述元素可能因以不當手段獲取或未經適當審查，而導致下游用

[121] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.38.

[122] 參見 Hopkins, A. et al. (2024) AI supply chains aren't Ai Value Chains. Available at: https://aipolicy.substack.com/p/supply-chains-6 (Accessed: 13 June 2024)..

戶之透明度或可責性降低。舉例而言，模型可能以未經驗證之第三方內容訓練，而可能導致無法驗證之模型輸出。又因生成式 AI 系統通常涉及多種不同的第三方元素，將系統行為中之問題歸因於任何一個來源第三方，都將是十分困難的。

可能對策與治理方案：

1. 將標準化測量與結構化之人類回饋方法適用於內部開發及第三方生成式 AI 系統，以管理 AI 價值鏈與內部組件整合風險[123]。

2. 透過社區資源、官方指引或研究文獻，維持對新興的生成式 AI 安全風險與相關應對手段之認識；此得以管理資訊安全與未知風險[124]。

3. 追蹤與紀錄生成式 AI 系統訓練資料之存取及更新；和驗證生成式 AI 系統供應者及服務提供者用於訓練資料之適當安全措施，以管理資料安全、AI 價值鏈與內部組件整合風險[125]。

[123] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.42.
[124] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.43.
[125] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.43.

4. 使用可解釋的機器學習技術（interpretable machine learning techniques）[126]，以使得 AI 之過程及輸出結果更加透明，亦更加容易了解決定如何做成[127]。

5. 透過與 AI Actors 接觸、外展，評估已知實例完整性與輸入之期望性能表現，如第三方資料或上游 AI 系統，或直接、間接使用生成式 AI 系統之下游系統的效能；以便管理人類與 AI 配置互動、AI 價值鏈與內部組件整合、有害及有偏見與同質化之風險[128]。

6. 於運用現今量測技術或尚無可用之度量指標，仍難以評估 AI 風險時，應考量使用風險追蹤方法，例如紀錄生成式 AI 系統對於組織及外部 AI Actors 所造成之損害發生率及嚴重程度，以管理人類與 AI 配置互動風險，或與外部 AI 環節參與者一同建立識別新興生成式 AI 系統風險之流程，以管理人類與 AI 配置互動、未知的風險[129]，並建立 AI 對社會衝擊（societal

[126] 參見 Molnar, C., Casalicchio, G., & Bischl, B. (2020, September). Interpretable machine learning–a brief history, state-of-the-art and challenges. In Joint European conference on machine learning and knowledge discovery in databases (pp. 417-431). Cham: Springer International Publishing. Available at https://link.springer.com/chapter/10.1007/978-3-030-65965-3_28.

[127] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.45.

[128] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.50.

[129] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.51.

impacts）[130]及其生成差異暨包容內容之角色（role of diverse and inclusive content generation）[131]的評估系統[132]。

7. 關於部署環節及橫跨整個 AI 生命週期之 AI 系統可信賴性（trustworthiness），其測量結果應由各領域專家及相關 AI Actors 輸入提供，以便驗證，具體措施包

[130] 相關文獻，參見 Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359; Jaume-Palasi, L. (2019). Why we are failing to understand the societal impact of artificial intelligence. Social Research: An International Quarterly, 86(2), 477-498; Robinson, S. C. (2020). Trust, transparency, and openness: How inclusion of cultural values shapes Nordic national public policy strategies for artificial intelligence (AI). Technology in Society, 63, 101421; Tomašev, N., Cornebise, J., Hutter, F., Mohamed, S., Picciariello, A., Connelly, B., ... & Clopath, C. (2020). AI for social good: unlocking the opportunity for positive impact. Nature Communications, 11(1), 2468; Vesnic-Alujevic, L., Nascimento, S., & Polvora, A. (2020). Societal and ethical impacts of artificial intelligence: Critical notes on European policy frameworks. Telecommunications Policy, 44(6), 101961; Sabherwal, R., & Grover, V. (2024). The Societal Impacts of Generative Artificial Intelligence: A Balanced Perspective. Journal of the Association for Information Systems, 25(1), 13-22; Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K., & Cave, S. (2019). Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research. London: Nuffield Foundation; Chui, M., Harryson, M., Valley, S., Manyika, J., & Roberts, R. (2018). Notes from the AI frontier applying AI for social good.

[131] 相關文獻，參見 Jora, R. B., Sodhi, K. K., Mittal, P., & Saxena, P. (2022, March). Role of artificial intelligence (AI) in meeting diversity, equality and inclusion (DEI) goals. In 2022 8th international conference on advanced computing and communication systems (ICACCS) (Vol. 1, pp. 1687-1690). IEEE; Zowghi, D., & Bano, M. (2024). AI for all: Diversity and Inclusion in AI. AI and Ethics, 1-4; Fosch-Villaronga, E., & Poulsen, A. (2022). Diversity and inclusion in artificial intelligence. Law and Artificial Intelligence: Regulating AI and Applying AI in Legal Practice, 109-134; Zowghi, D., & da Rimini, F. (2023). Diversity and inclusion in artificial intelligence. arXiv preprint arXiv:2305.12728; Cachat-Rosset, G., & Klarsfeld, A. (2023). Diversity, equity, and inclusion in artificial intelligence: An evaluation of guidelines. Applied Artificial Intelligence, 37(1), 2176618; Shams, R. A., Zowghi, D., & Bano, M. (2023). AI and the quest for diversity and inclusion: a systematic literature review. AI and Ethics, 1-28; Daugherty, P. R., Wilson, H. J., & Chowdhury, R. (2020). Using artificial intelligence to promote diversity; Mohammed, P. S., & 'Nell'Watson, E. (2019). Towards inclusive education in the age of artificial intelligence: Perspectives, challenges, and opportunities. Artificial Intelligence and Inclusive Education: Speculative futures and emerging practices, 17-37; Chi, N., Lurie, E., & Mulligan, D. K. (2021, July). Reconfiguring diversity and inclusion for AI ethics. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (pp. 447-457).

[132] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.52.

括：實施可詮釋（interpretability）及可解釋（explainability）之方法，以評估生成式 AI 系統與內容來源及相關決策，以及驗證與預期目的是否一致；將結構化之人類反饋結果整合至校準及更新至傳統之測量方法流程中（如：基準、效能表現評估、數據品質量測）；監控並記錄當人類操作者或其他系統推翻生成式 AI 決策之實例，以充分了解該推翻決定是否涉及與內容來源相關之問題；驗證並記錄將結構化之人類反饋結果納入設計、實施、部署批准、監控與停止使用之決策；與各領域專家一同合作以審視來自終端使用者、營運商或操作者，及潛在受影響之個人或群體之反饋，以管理人類與 AI 配置互動風險；與了解生成式 AI 系統使用之領域專家一同合作，以評估內容之效度、相關性與潛在偏見[133]。

8. 訂定審查現有第三方外掛增效工具（third party plug-ins）[134]授權政策與程序，以識別生成式 AI 系統事件

[133] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, pp.52-53.

[134] Third party plug-ins，係現有電腦程式或網頁瀏覽器添加特定功能的軟體組件，目的在不改變任何核心程式碼下，拓展主機應用程式之功能，其使用目的多元，包括添加新功能、提高表現效能、增強安全性，或與第三方服務整合；與第三方服務整合者，即可允許第三方開發人員擴展應用程式與添加、配置新功能，以滿足不同需求。參見 Nguyen, H. V., Kästner, C., & Nguyen,

[135]，並向下游 AI Actors（包括可能受生成式 AI 產出影響之個人）進行揭露；在制定生成式 AI 治理政策時，應納入現有事件回應、舉報人、供應商盡職調查等政策與機制，建立下游 AI Actors 接受生成式 AI 風險的政策與程序[136]；更新可接受使用政策，解決生成式 AI 特有開源技術、資料及第三方人員的問題；須制定開發和整合生成式 AI 購置與採購供應商評估的盡職調查程序，並將智慧財產權、資料隱私、安全和其他風險考量納入其中[137]。

9. 在估計可能性、影響程度和風險時，應考慮生成式 AI 特有的映射風險，核實審查程序，包括分析生成式 AI 系統輸出作為第三方外掛增效工具或其他系統輸入時的連帶影響[138]；在運行與監督上，應盤點所有可存取組織內容的第三方，建立經批准的生成式 AI 技術

T. N. (2014, May). Exploring variability-aware execution for testing plugin-based web applications. In Proceedings of the 36th International Conference on Software Engineering (pp. 907-918)；Jaspan, C. N. (2011). Proper plugin protocols (Doctoral dissertation, Carnegie Mellon University)；Berg, M., Cederquist, F., Hamilton, D., Johansson, J., & Vindblom, A. (2004). A case study of how to add a plugin framework.

[135] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.14.

[136] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.18.

[137] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.23.

[138] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.14.

和服務提供者名單，將風險繪圖和衡量計劃應用於第三方與開源系統，並實施供應商風險評估框架，透過對第三方流程和人員進行審查（包括聲譽的審查）、驗證第三方模型是否符合現有的使用許可，來持續評估並監控第三方的表現，包括對第三方聲譽的審查[139]。

10. 持續更新可接受使用政策，以解決生成式 AI 專有的開源技術與數據問題，例如，與機器人流程自動化（robotic process automation, RPA）、軟體即服務（software-as-a-service, SAAS）[140] 以及其他可能依賴嵌入式生成式 AI 技術有關的解決方案[141]。為應對風險，亦應審查與更新第三方合約、服務協定和保證書等涉及第三方權利之法律文件；在持續追蹤與監控層面，建立制度與程序，透過社群或官方資源（例如 AI

[139] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.22.

[140] SAAS（軟體即服務），係一種軟體交付模型，其提供使用者以遠端（通常透過網際網路）使用、存取業務功能之服務；因此使用者毋需特地購買各軟體授權許可，僅需支付按月或按次數計價之使用費，其中即包含基礎設施（如：硬體與作業系統）之成本、軟體之使用權與所有網域代管、維護與支援服務費用，進而大幅降低使用者支出的成本及負擔。參見 Sun, W., Zhang, K., Chen, S. K., Zhang, X., & Liang, H. (2007). Software as a service: An integration perspective. In Service-Oriented Computing – ICSOC 2007: Fifth International Conference, Vienna, Austria, September 17-20, 2007. Proceedings 5 (pp. 558-569). Springer Berlin Heidelberg；Waters, B. (2005). Software as a service: A look at the customer benefits. Journal of digital asset management, 1, 32-39；Goyal, S. (2013). Software as a service, platform as a service, infrastructure as a service– a review. International journal of Computer Science & Network Solutions, 1(3), 53-67.

[141] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.23.

事件資料庫[142]、OECD 事件監控資料庫[143]），以適時對 AI Actors 及下游利害關係人傳達、交流生成式 AI 系統事件與系統效能[144]。與第三方合作進行聯合教育活動，推廣內容出處的最佳案例，並訂定專門針對第三方的事件回應計畫，處理內容出處事件或違規事件[145]。

四、民主影響評估與公共領域風險治理

（一）生成式 AI 對資料整全性衝擊的風險治理

資訊整全性（information integrity）[146]，指可被信任之高度完整資訊於社會中資訊之範圍及其創建、交換與消耗之相關模式；區辨事實與虛構、個人意見與推論。簡言之，即資料的正確性、可靠性與可信賴性[147]，於數位時代則延伸為資

[142] AI Incident Database. Available at https://incidentdatabase.ai/.

[143] OECD incident monitor. Available at https://www.oecd.org/digital/artificial-intelligence/.

[144] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, pp.14-15.

[145] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.23.

[146] Information integrity 的觀念，於本世紀初即有學者提出，參見 Geisler, E., Prabhaker, P., & Nayar, M. (2003, July). Information integrity: an emerging field and the state of knowledge. In PICMET'03: Portland International Conference on Management of Engineering and Technology Technology Management for Reshaping the World, 2003. (pp. 217-221). IEEE.

[147] 參見 Miller, K. W., & Voas, J. (2008). Information integrity and IT professionals' integrity, intertwined. IT Professional, 10(6), 35-40.

訊系統的整全性[148]。資訊整全性與否，關乎民主政治運作的良窳[149]。生成式 AI 系統使得大量製造虛假、不正確或誤導性內容十分容易，且此類錯誤資訊（misinformation）得被無意間創造或傳播，特別是當此種虛構內容係因不經意、無害之詢問下產生。研究亦已表明，縱使僅係於文字或影像之微小改變，均會影響人類之判斷與認知。

生成式 AI 系統亦可於使用者有明確意圖欺騙或對他人造成傷害時，用以大規模生產錯誤或是誤導性資訊（disinformation）。生成式 AI 系統亦使得惡意行為者得以透過更高度之精細與複雜手段以生成針對特定族群之錯誤訊息（disinformation）。現今與新興之多模態模型不僅得以生成文字假訊息，亦可以生成高度逼真的「deepfake」視聽內容

[148] 德國聯邦憲法法院於 2008 年一則有關線上搜索的判決中（BVerfGE 120, 274）提出的新興基本權：「確保於資訊技術系統私密性與整全性之基本權」（Grundrecht auf Gewährleistung der Vertraulichkeit und Integrität informationstechnischer Systeme），亦是基於「資訊整全性」的觀念。參見 Gabriele Brit, Vertraulichkeit und Integrität informationstechnischer Systeme, DÖV 2008, 411 ff.; Andreas Voßkuhle/Schemme, Grundwissen – Öffentliches Recht: „Neue“ Grundrechte, JuS 2024, 312 (314).另外，近年尚有所謂「資料整全性」（data integrity, Datenintegrität）概念的提出，側重於個人資料的集體性與系統性保護（Sicherstellung der Korrektheit (Unversehrtheit) von Daten und der korrekten Funktionsweise von Systemen）。參見 Technavigator,Was ist Datenintegrität? (2024), https://technavigator.de/wiki/datenintegritaet/; James, in: Leupold/Wiebe/Glossner, IT-Recht, 4. Aufl., 2021, Teil 11.1 Cloud Computing – Vorteile und Risiken, Rn. 20.

[149] 參見 Yadav, K. and Lai, S. (2024) What does information integrity mean for democracies?, Lawfare. Available at: https://www.lawfaremedia.org/article/what-does-information-integrity-mean-for-democracies (Accessed: 13 June 2024).

及人工合成影像；因而未來受新資料訓練之生成式 AI 模型可能引發並造成更多假訊息的威脅。

惡意利用生成式 AI 進行假訊息活動（disinformation 與 misinformation）之行為者，可能減弱公眾對於真實或確實證據及資訊的信任。舉例而言，五角大廈爆炸之合成圖片瘋傳並造成股市短暫下跌。生成式 AI 模型亦可協助惡意使用者創造看似高度可信之影像與宣傳以使假訊息活動更加逼真，並造成假訊息在社群軟體上更高的觸擊率與互動。

可能對策與治理方案：

1. 評估與資料來源、存取控制及事件反應程序相關之文件完整性，與驗證生成式 AI 系統內容來源文件符合相關規定及標準，以管理資料完整性、有害及有偏見與同質化之風險[150]；驗證生成式 AI 系統之訓練資料與 TEVV 資料來源，且確保用以微調（fine-tuning）之資料為有合理根據的，以監管資訊完整性風險[151]。

[150] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.42.

[151] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.44.

2. 透過資訊共享機制與安全專家、開發者與研究者互動，並以此了解並持續跟進與內容來源相關之 AI 安全的最新發展；亦透過資訊共享機制、工作坊或出版物以貢獻與 AI 系統安全性與內容來源之發現。以上行動得降低資料完整性、資訊安全之風險[152]。

3. 在治理與監管方面，各機構應制定政策，以定義並衡量以下兩機制的有效性：標準內容來源之方法，例如利用密碼學、浮水印、隱寫技術等方法[153]；測試，包含逆向工程[154]。與此同時，記錄訓練資料和生成式 AI 系統生成資料的來源，包括著作權、授權許可和資料隱私。組織於制定風險等級時，根據組織風險承受能力，同時將資訊不完整的風險也納入考量因素，而制定風險管理活動水準，並適時重新評估風險容忍度以

[152] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.42.

[153] AI 浮水印或隱寫技術是一種防偽及防盜技術，可以確保 AI 生成內容的版權，同時增加生成內容被偽造或變造的難度。相關技術及示範，參見 Bartz, D., & Hu, K. (2023, July 22). OpenAI, Google, others pledge to watermark AI content for safety, White House says. Retrieved from Reuters: https://www.reuters.com/technology/openai-google-others-pledge-watermark-ai-content-safety-white-house-2023-07-21/; Fernandez, P., Couairon, G., Jégou, H., Douze, M., & Furon, T. (2023, October 6). Stable Signature: A new method for watermarking images created by open source generative AI. Retrieved from Meta: https://ai.meta.com/blog/stable-signature-watermarking-generative-ai/.數位浮水印早已用於著作權保護，相關討論，參見 Wolfram Gass, Digitale Wasserzeichen als urheberrechtlicher Schutz digitaler Werke?, ZUM 1999, 815; Thomas Hoeren/Julia Dreyer, in: Hoeren/Sieber/Holznagel, Handbuch Multimedia-Recht, Werkstand: 60. EL Oktober 2023, Teil 7.2 Urheberpersönlichkeitsrecht im Internet, Rn. 39.

[154] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.12.

回應全球生成式 AI 之風險，如公開資訊完整性的風險[155]。

4. 根據風險排序所建立的生成式 AI 系統清單，其項目除通用模型、治理與風險資訊外，還有資料來源的資訊，包含以下幾個方面：資料的來源或出處；資料的簽名:驗證資料的真實性和完整性;資料的版本控制:追蹤資料的不同版本變更；資料的浮水印：用於識別和保護資料的所有權或著作權[156]。

5. 關於風險的識別，應識別端到端（end-to-end）AI 供應鏈中的內容來源風險，如資料供應商、研發合作夥伴、第三方供應商等，並識別潛在的內容來源風險和危害，如錯誤訊息、虛假資訊、深度偽造等，並列舉且排序之，進一步確定解決方案。另外，也應整合用於分析內容來源、檢測異常、驗證數位簽章真實性、識別失效模式的工具；於衡量風險時，使用相似度指標、竄改指標、區塊鏈確認、詮釋資料一致性，衡量內容來源風險。關於無法定量測量的風險應該被記錄

[155] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.13.

[156] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, pp.14-15.

下來，並解釋無法測量的原因，並持續追蹤附有出處資訊（如浮水印、加密標籤）及可能侵害智慧財產權的輸出資料項目與數量[157]。

6. 在運行與監控方面，組織需使用各種方法偵測生成內容中的敏感資料，並驗證模型是否會屏蔽該敏感資料。同時，評估並記錄與內容來源相關的風險，以提高內容追蹤透明度和不變性，具體來說，有以下面向：標籤系統（tagging systems）[158]的存在與否以及其有效性；加密雜湊（cryptographic hashes）[159]的應用情況；基於區塊鏈（blockchain-based）的技術解決方案；分散式帳本技術（distributed ledger technology solutions）[160]

---

[157] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.35.

[158] Tag（標註）係依使用者之關鍵字選擇用於描述、分類，而 tagging 之目的即如同過濾，僅將所有已標註之檔案文件中受有特定標註的項目回傳。又依據不同之應用目的及提問，tagging systems（標註系統）可以提供標註之聯集，而非僅標註的交集；而由使用者之角度觀之，標註系統即類似於關鍵字搜尋，因此無論如何應用，使用者均會提供明顯、描述性之詞語以便檢索適用之項目。參見 Golder, S., & Huberman, B. A. (2005). The structure of collaborative tagging systems. arXiv preprint cs/0508082；Au Yeung, C. M., Gibbins, N., & Shadbolt, N. (2009, June). Contextualising tags in collaborative tagging systems. In Proceedings of the 20th ACM Conference on Hypertext and Hypermedia (pp. 251-260).

[159] Cryptographic hashing（密碼雜湊），係將給定之任意長度訊息資料應用單向演算函數，以處理並生成固定長度之 digest（摘要）或 hash code（雜湊碼）。又密碼雜湊函數常用於達成安全保障目標，如身分驗證、訊息完整性、電子簽章之不可否認性、公鑰密碼學、密碼保護等。參見 Sobti, R., & Geetha, G. (2012). Cryptographic hash functions: a review. International Journal of Computer Science Issues (IJCSI), 9(2), 461；Thomsen, S. S., & Knudsen, L. R. (2005). Cryptographic hash functions. PhD thesis; Gauravaram, P. (2007). Cryptographic hash functions: cryptanalysis, design and applications (Doctoral dissertation, Queensland University of Technology).

[160] Distributed ledger technology（分散式帳本技術），係一種特殊形態之分散式資料庫（distributed database，即資料於以平等權限複製於多個節點間），帳本僅允許新增或閱讀資

的應用情況。此外，並應列舉與內容來源相關的潛在影響，其中包括最佳、平均和最不理想的情況[161]。

7. 在 TEVV 的框架下，首先，定義與內容來源相關的生成式 AI 系統任務，並建立假設和做法來確定和記錄資料來源及內容脈絡，接著，識別並記錄影響內容來源可靠性的 生成式 AI 任務限制，並且持續評估、改進、追蹤和驗證內容來源的技術和方法，以明確 AI 系統的具體任務與用於執行任務的方法，達確保系統的性能與可靠性之目的；持續監控的部分，組織應實施 AI 系統定期對抗測試計劃，發現漏洞和潛在的操縱風

料，而不得事後刪除或更新；其主要目標係使毋需具互信基礎之使用者，得以不透過可信之第三方（如：銀行）證明儲存資料之真實性，即得進行交易互動。而 DLT 係利用公鑰密碼學（public key cryptography）、分散式點對點網絡（distributed peer-to-peer networks）及共識機制（consensus mechanisms）之技術運作，且各個交易參與者均具有一對鑰匙（一把公鑰、一把私鑰），以便其記錄分散式帳本中的交易；因此在點對點網絡運作以避免單一中心風險，以及各個節點按共識機制確認數據、確保帳本資訊一致與公平，和以數位身分加強控制下，達到去中心化、安全保障、低運作成本目的，又，區塊鏈（Blockchain）即為一種分散式帳本技術。參見 El Ioini, N., & Pahl, C. (2018). A review of distributed ledger technologies, in: On the Move to Meaningful Internet Systems. OTM 2018 Conferences: Confederated International Conferences: CoopIS, C&TC, and ODBASE 2018, Valletta, Malta, October 22-26, 2018, Proceedings, Part II (pp. 277-288). Springer International Publishing; Sunyaev, A., & Sunyaev, A. (2020). Distributed ledger technology. Internet computing: Principles of distributed systems and emerging internet-based technologies, 265-299.

[161] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.33.

險，並整合 AI 系統、現有內容管理與版本控制系統，以利整個生命週期中都能追蹤內容出處[162]。

8. 聯合國訂有自願性的「數位平台資訊整全性行為準則（Code of Conduct for information integrity on digital platforms）」[163]，可供參考。

（二）生成式 AI 對多元公共意見形成的風險治理

公共領域是民主的基本要素與變項，公共意見與公共政策形成的場域，集體學習、認知與利益權衡的機制。公共領域是決策機制，社會變遷與決策機制相互連動，如果說社會已數位轉型成為數位社會，則理論上公共領域及民主決策機制亦將隨之變遷。數位媒體提升人們進入公共領域的速度，許諾公共領域的可近性與雙向對話溝通的可能性。影響所及，公共領域擴大的同時，不同面向或層次（個人、團體、大眾）公共領域之間的界線也逐漸模糊乃至泯滅。欲透過區隔私人領域而形構的公共領域，在數位時代下二者間的介面亦趨於模糊。公共領域的數位變遷，至少可以略窺一種弔詭現象：

[162] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, pp.27-29.

[163] 參見 Code of conduct. United Nations Information Integrity. Available at: https://www.un.org/en/information-integrity/code-of-conduct (Accessed: 13 June 2024).; United Nations (2023) Information integrity on digital platforms. Available at: https://www.un.org/sites/un2.un.org/files/our-common-agenda-policy-brief-information-integrity-en.pdf (Accessed: 13 June 2024).

一方面，讓意見趨於多元、提升公共參與；另方面則又因意見過於破碎而使多元性質變為兩極化。數位平台之間的互動參照中過濾出不同意見的目的，不是因為欲求相互砥礪攻錯的他山之石，而是在黨同之餘不斷群起伐異，甚或污名異見，公共意見整全性暨公共論壇的解體，其結果可能是提供孕育民粹主義的溫床，為數位變遷下公共領域與民主的主要危機所在[164]。

「有害」於此處係指由生成式 AI 系統創造或經故意編碼編入之負面、失禮的或不合理之內容。AI 生成內容之有害、煽動仇恨或仇恨言論之創造與接觸大眾的難以控制性，以及誹謗或刻版內容，可能造成嚴重傷害；例如，當輸入提示以生成執行長、醫師、律師與法官之影像時，AI 模型使用之內嵌文字存在偏見而無法充分代表女性族群。而存在生成式 AI 模型或訓練資料之偏見，亦會損害代表性或保留、加強種族偏見而增強有害性。

有害性與偏見亦可能導致同質化或其他不良後果。生成式 AI 產出之同質化可能導致近似風格、減少內容多樣性與

[164] 參見 Martin Seeliger and Sebastian Sevignan, A New Structural Transformation of the Public Sphere? An Introduction, Theory, Culture & Society 2022, Vol. 39 (4) 10-12.

大規模地增強特定意見或價值；此種現象可能起源於基礎模型之內含偏見，而產生瓶頸或歧視、排除，進而延伸並複製至下游應用程式。模型崩潰（model collapse），指當生成式AI模型以先前模型所生成之內容為訓練，而導致資料庫中異常值或特定資料點消失之現象，而此現象可能源於統一之回饋循環或合成資料之訓練[165]；又模型崩潰可能導致產出同質化，而造成特定族群與模型穩健性之威脅。其他生成式AI模型之偏見可能影響模型近用能力，導致利益的不公平分配。

可能對策與治理方案：

1. 評估內容出處驗證方法之可靠性，如以浮水印（watermarking）、密碼學數位簽章（cryptographic signatures）、雜湊（hashing）、區塊鏈（blockchain）或其他內容來源技巧，並評估內容出處之偽陽性與偽

[165] 簡言之，生成式AI所產製的內容是一種人工合成的資料，不適合再用來作為訓練生成式AI的資料。相關討論，參見 Gal/Lynskey, Synthetic Data: Legal Implications of the Data-Generation Revolution, Iowa L. Rev. 109 (2023); Shumailov et al., The Curse of Recursion: Training on Generated Data Makes Models Forget, 31.5.2023, https://arxiv.org/abs/2305.17493; Adina Bresge, Training AI on machine-generated text could lead to 'model collapse,' researchers warn, August 21, 2023, https://www.utoronto.ca/news/training-ai-machine-generated-text-could-lead-model-collapse-researchers-warn; Graeme Whiles, What is AI Model Collapse? Everything You Need to Know, January 14, 2024, https://aicontentexpert.co.uk/2024/01/14/ai-model-collapse/.

陰性，以及用於驗證之真陽性與真陰性，以確保資料的整全性[166]。

2. 彙整及針對有關組織性生成式 AI 系統之政策違規、撤下請求、智慧財產權侵害的統計數據，以及資訊完整性進行溝通[167]。

3. 分析跨人口統計、語言群體及與部署內容相關之其他部分的透明性報告，以助於管理資料整全性、智慧財產、有害及有偏見與同質化之風險[168]。

4. 使用如區塊鏈或數位簽章技術記錄各個內容生成、修改或共享歷程，以提供防止竄改之內容歷史紀錄，並提升透明性及實踐可追溯性；健全的版本控制系統，亦得使用於追蹤隨著時間推移演進的整個 AI 生命週期[169]；透過與潛在受影響社群之直接接觸，識別可能受生成式 AI 系統影響之個人、群體或環境生態系統類別，管理環境、有害及有偏見與同質化之風險[170]；

[166] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.43.
[167] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.44.
[168] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.44.
[169] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.45.
[170] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.48.

建立驗證機制，審驗訓練使用之資料非由同質化、由生成式 AI 產出，以便減緩模型崩潰之擔憂[171]；驗證功能近用無障礙性（accessibility functionality）[172]與退出機制（opt-out functionality）[173]的功能及其程序[174]。

5. 驗證從事生成式 AI 之 TEVV 的 AI Actors，其內容來源是否反映多元族群及跨學科之背景，以便協助降低人類與 AI 配置互動、資訊整全性、有害及有偏見與同質化的風險[175]。

6. 於系統設計階段，應納入跨領域的觀點。從資歷、人口代表性、多樣性和專業資格等方面，評估並適當調

[171] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.49.

[172] 相關文獻，參見　Henry, S. L., Abou-Zahra, S., & Brewer, J. (2014, April). The role of accessibility in a universal web. In Proceedings of the 11th Web for all Conference (pp. 1-4); Rutter, R., Lauke, P. H., Waddell, C., Thatcher, J., Henry, S. L., Lawson, B., ... & Urban, M. (2007). Web accessibility: Web standards and regulatory compliance. Apress; Harper, S., & Chen, A. Q. (2012). Web accessibility guidelines: A lesson from the evolving Web. World Wide Web, 15, 61-88; Lazar, J., Dudley-Sponaugle, A., & Greenidge, K. D. (2004). Improving web accessibility: a study of webmaster perceptions. Computers in human behavior, 20(2), 269-288; Petrie, H., & Kheir, O. (2007, April). The relationship between accessibility and usability of websites. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 397-406).

[173] 相關文獻，參見　Lai, Y. L., & Hui, K. L. (2006, April). Internet opt-in and opt-out: investigating the roles of frames, defaults and privacy concerns. In Proceedings of the 2006 ACM SIGMIS CPR conference on computer personnel research: Forty four years of computer personnel research: achievements, challenges & the future (pp. 253-263); Habib, H., Zou, Y., Jannu, A., Sridhar, N., Swoopes, C., Acquisti, A., ... & Schaub, F. (2019). An Empirical Analysis of Data Deletion and {Opt-Out} Choices on 150 Websites. In Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019) (pp. 387-406); Kosta, E. (2015). Construing the Meaning of" Opt-out"-An Analysis of the European, UK and German Data Protection Legislation. Eur. Data Prot. L. Rev., 1, 16; Miyake, E., & Kuntsman, A. (2022). Paradoxes of Digital Disengagement: In Search of the Opt-Out Button.

[174] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.49.

[175] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.52.

整 AI 設計成員的組成[176]，並建立正當程序，促使其探索 AI 限制以外之發展;同時制定政策與實踐指南，用以實施生成式 AI 的影響評估，項目包含：資料、標籤、偏差、隱私、模型、演算法、錯誤、來源技術、安全性、合規性、輸出等，以降低風險和危害[177]。

7. 在運行與監督階段，應維護並持續更新與生成式 AI 使用環境相關的已識別和預期的風險分級制度，包含針對模型崩潰與演算法同質化等評級[178]，並於有相關事件發生時，根據事件類型，建立應對小組與程序。透過三道防線等機制，對 AI 系統設計、實施和部署決策的進行有效質疑，而降低工作場所文化風險，例如如認知偏差、群體思維或過度依賴衡量標準等[179]。

8. 在整個 AI 生命週期中，應確保生成式 AI 技術供應商能識別並減少有害內容和系統性偏見，並保持內容的多樣性和創新性，供應商須提供證據證明其技術在促

[176] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.17.
[177] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.19.
[178] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.13.
[179] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.18.

進內容多樣性方面的有效性，並獨立審查，評估訓練資料和模型的公平性。

9. 著眼於 AI 對民主的影響，以言論自由作為審視是否及如何管制社群媒體的切入主軸，思考腹地失之過窄。蓋憲法上之言論自由係以保障人民（自然人或法人）表達一定意見的基本權利，強調權利主體的個別性與權利的防禦性。從民主社會形成（公共）意見的宏觀角度來看，數位平台與公共論壇於民主社會的重要性，不能只聚焦在各別使用者的言論自由（言論市場），而應將視野投向建立社群媒體使用規範與平台法秩序的重要性之上。換言之，僅從言論自由的角度思考 AI 對於民主的影響，略嫌狹隘，無法整全掌握問題的癥結，需要將思維的眼界與腹地擴展到「媒體自由」之具制度性與建構性的「社群媒體自由」。也就是，必須跳脫防禦性言論自由的消極思維，轉為形成性媒體自由的積極思維。就媒體自由之基本權主體而言，其一方面得對於國家行使防禦權，另一方面則負有確保其所建立且經營之社群媒體得以自由形成意見之

責任。歐盟數位服務法的制定與施行，旨在建立一套數位平台的法秩序，其後續的實踐與影響值得觀察[180]。

五、社會衝擊與環境影響的風險治理

（一）生成式 AI 對資通安全性衝擊問題

電腦與資料之資訊安全係一成熟的領域，並針對防禦與攻擊之網域能力擁有被廣為接受及標準化之程序。以生成式 AI 為基礎之系統呈現二資訊安全風險：生成式 AI 透過降低自身攻擊性能力之障礙以發現或促成新網路安全風險之可能，以及同時因生成式 AI 模型自身對於新攻擊之脆弱性而擴張可能受攻擊之範圍。

生成式 AI 系統亦可能強化針對網路安全進行之攻擊性行為，如駭客行為、惡意軟體及網路釣魚。現亦已有報告指出大型語言模型以可以發現無論硬體、軟體等系統中之漏洞，並寫出程式碼以利用該漏洞；而有心人士亦可以透過開發生成式 AI 模型並利用於多處攻擊鏈中以更加擴大上述風險，如告知攻擊者如何在取得系統存取權限後，主動逃避威脅偵測。又考量生成式 AI 價值鏈之複雜性，識別與確保針對特

[180] 參見 Tahireh Panahi, Gesetz gegen digitale Gewalt – auf Kollisionskurs mit dem DSA?, MMR 2023, 556 ff.; Sarah Legner, Der Digital Services Act – Ein neuer Grundstein der Digitalregulierung, ZUM 2024, 99 ff.

定要素之各可能受攻擊點之安全性的手段需要調整或更加進化。

最令人擔憂之生成式 AI 模型漏洞涉及提示輸入，或操縱生成式 AI 系統使之以意料之外的方式運作。於直接輸入提示中，攻擊者可能公開地利用提示輸入以造成不安全之行為，並造成多樣內部關聯系統之後階段結果；而在間接輸入提示之攻擊中，行為者遠端透過將提示輸入可能被提取之資料中以利用嵌入大型語言模型之應用程式。網路安全研究人員亦以已演示過如何透過間接輸入提示以竊取資料及在機器上遠端運作程式碼[181]。

生成式 AI 模型與系統之資訊安全同時亦包含安全性、保密性、生成式 AI 訓練資料的完整性、程式碼與模型大小。另一生成式 AI 模型面臨之新興網路風險為「資料毒害（data poisoning）」，即攻擊者破壞模型使用之訓練資料以操縱該模型之運作；此種未經授權竄改資料或部分模型之惡意行為，可能將加劇生成式 AI 系統輸出之相關風險。

可能對策與治理方案：

[181] 關於智慧醫療設施所生之資安風險與防範法制，參見 Henrik Nolte/Zeynep Schreitmüller, Cybersicherheit KI-basierter Medizinprodukte im Lichte der MDR und KI-VO, MPR 2024, 20 ff.

1. 紀錄人類知識領域如何應用，以提升生成式 AI 系統之表現效能程度，如透過人類意見回饋強化學習（Reinforcement Learning from Human Feedback）、微調（fine-tuning）、內容審核、商業規定[182]。

2. 設計系統時，相較於傳統 AI，應考慮：在大型或複雜的生成式 AI 系統生命週期的各個階段，是否需調整組織角色與組成結構，包括：生成式 AI 系統的稽核、驗證和攻擊演練、生成式 AI 內容審核、資料記錄與標籤、資料預處理和標記、生成式 AI 系統退役、監督相關的 AI 行為者和數位實體，包括管理安全憑證與 AI 實體之間的通訊[183]。

3. 在治理與監管方面，組織需考慮在生成式 AI 系統生命週期各階段調整組織角色和 AI 行動者。制定政策監控第三方影響評估，內容包括資料、偏差、隱私、模型等，並在 AI 模型訓練中納入對抗樣本，增強抵禦攻擊能力。此外，還需為外部和第三方制定內容來源的事件回應計劃，依回饋定期更新，同時訂立明確

[182] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.40.
[183] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.17.

的合約和服務等級協議（SLA）規範內容所有權、安全要求等[184]。

4. 在運行層面，實施生成式 AI 系統定期對抗測試，發現漏洞和潛在的操縱風險，並採用加密技術，確保資料安全儲存和傳輸，並以 TEVV 實踐，例如：零知識證明(zero- knowledge proof approaches)探測系統的合成資料能力，並概述生成式 AI 系統與內容交互產生的風險及漏洞。並且應與跨領域專家合作，確保風險管理方法有效，並進行對抗練習、攻擊演練、識別異常失效模式[185]。

5. 透過重新取樣、重新加權或對抗訓練等技術，評估和管理與 生成式 AI 內容來源相關的統計偏差，並負責任地揭露與內容來源相關的漏洞和偏差調查結果。為防止未獲授權的存取和破壞，實施網路安全措施，同時也應記錄內容來源機制如何在隱私和安全的背景下運行[186]。

[184] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, pp.19-22.
[185] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.21.
[186] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.37.

（二）生成式 AI 與 CBRN 資訊取得問題

儘管 CBRN 材料（chemical, biological, radiological and nuclear materials）資訊本即可在公開管道取得，非專業人士可以透過聊天機器人更容易地綜合分析相關資訊。雖然實際製造與使用生化製劑，必須仰賴專業知識技能與基礎硬體設施，但是模型增加使用者進行生化武器研究的便利性，確實幫助缺乏正規訓練的惡意人士擴大其危險性。另有研究認為，因目前的大型語言模型無法提供比一般網路搜尋引擎更精密成熟的生化武器計劃資訊，故受生化武器攻擊的風險並不因此大幅增加；然而，專門訓練用於生化設計 AI 系統，可能可以預測或生成全新的生物或化學武器，對社會或國家安全造成極大的風險[187]。

可能對策與治理方案：

1. 擬定政策和程序，應確保訓練資料中不含 CBRN 相關之有害或非法內容，並於定義或更新生成式 AI 風險等級時，根據風險承受能力，考慮資訊被濫用的風險、生成式 AI 與其他 IT 或資料系統之間的依賴性、內在或外在使用及法規要求等因素。舉例來說，於制

[187] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, pp.4-5.

定並執行「資料收集與保留」與「最低品質」之政策時，除了應將 CBRN 資料從訓練資料中刪除，也須注意、識別使用的技術或內容可能侵害的智慧財產權或其他權利之風險。同時也要諮詢受影響的個人與社群，收集公眾反饋，了解最終使用者的意見。在測試、評估、驗證和確認的階段裡，為確保風險管理方法的有效性，與跨領域專家的合作極具關鍵性[188]。

2. 評估系統訓練資料庫中有害資訊程度、是否存在侵犯智慧財產權、侵害資訊隱私權、猥褻性、極端性、暴力性或大規模殺傷性武器（化生放核，CBRN，Chemical、Biological、Radiological、Nuclear）之資訊，以降低對於資訊隱私、智慧財產權、猥褻或有辱人格和／或辱罵性內容、有害及有偏見與同質化、危險或暴力建議、大規模殺傷性武器資訊之風險；驗證系統是否得正確處理可能導致不當、惡意或違法使用，包含協助或促成操縱、敲詐，冒充特定人士、網路攻擊與武器製造之詢問[189]。

[188] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, pp.34-35.

[189] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, pp.41.

（三）生成式 AI 形成之社會倫理觀風險

生成式 AI 可以輕易產出與獲取成人猥褻或未經同意的性影像，以及孩童之性剝削內容；而此種輸出內容可能造成隱私、心理與情緒，甚至實質風險，且上述內容之傳播亦有其下游影響，以孩童性剝屑內容為例，縱使生成之影像為非特定之個人，此類影像之流傳議會對於現實世界之受害者造成影響。

生成式 AI 模型多半以網路上之開放性資料為訓練，因而可能不經意地涵蓋孩童之性剝屑內容以及未經同意之私密畫面。近期之報告指出，數個常用以訓練生成式 AI 之資料庫，已被發現涵蓋數百個已知的孩童性剝削影像。而因偵測困難與在網路上之廣泛流傳之緣故，色情或猥褻內容於模型訓練過程中係特別難以移除的；縱使係以「乾淨的」資料進行訓練，以生成式 AI 模型日漸增強之能力，其亦得自行生成合成的成人猥褻或未經同意的性影像或孩童之性剝屑內容。生成露骨或猥褻之 AI 生成內容可能涵蓋高度擬真的真實個人深偽影像（deepfakes），包含孩童。舉例而言，一未經著名藝人同意之 AI 生成親密影像於社交媒體上廣傳，並吸引數億次觀看。

可能對策與治理方案：

1. 建立終端使用者與受影響社群之反饋流程，以便其回報問題及針對系統生成結果進行申訴，並將之融入 AI 系統評估度量指標中[190]。

2. 針對 AI 生成內容將如何影響不同社會、經濟與文化族群，進行影響評估，以管理有害及有偏見與同質化之風險[191]。

3. 針對終端使用者如何看待、與內容來源相關之生成式 AI 產出內容互動進行研究；並評估該生成內容是否與終端使用者之期待相符，且終端使用者對於生成內容提供資訊之相應行動[192]。

4. 利用適當的方法（包含電腦計算測試方法與評估結構性的回饋輸入），評估可能由 AI 生成內容產生之潛在偏見與刻板印象[193]。

[190] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.51.
[191] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.51.
[192] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.51.
[193] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.51.

5. 透過實施使用者研究、焦點小組或社群論壇之方式，紀錄並結合來自操作者、使用者及潛在受影響群體關於內容來源之有架構性的回饋；並主動尋求關於生成內容之品質及潛在偏見的回饋；亦評估終端使用者與受影響群體關於回饋管道可近用性之普遍認知。上述行動得以降低人類與 AI 配置互動、資訊完整性、有害及有偏見與同質化的風險[194]。

6. 制定政策與程序，確保有害或非法內容不包含在訓練資料中，尤其是 CSAM、已知的 NCII、裸露和暴力的內容，並且在更新或定義生成式 AI 風險等級時，考量濫用風險與法律監管要求等因素，同時評估粗俗、惡質的輸出對人類心理造成的影響[195]。

7. 生成式 AI 的開發，應同時投資研發能力的評估與實施方法的檢驗機制，用以測量 AI 相關內容的來源與惡性的風險，並考慮對生成式 AI 系統的輸出進行同步的人為審查，遵循記錄資料治理之政策[196]。

[194] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.52.
[195] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.12.
[196] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.35.

（四）生成式 AI 對環境生態衝擊風險

生成式 AI 系統的訓練、維護和部署所需資源密集且高耗能，其能源和碳排放量因生成式 AI 模型開發的類型（如預訓練、微調、推論 inference）、硬體、任務類型或應用類型而異。估計顯示，訓練一個生成式 AI transformer 模型所排放的碳相當於 300 次舊金山和紐約間的來回航班。雖然訓練較小模型的方法，可以減少推論造成的環境影響，但超參數(hyperparameter)的調整和訓練仍會產生相當程度的環境影響。另有研究發現，生成式的任務，所需的能源和碳消耗量較判別或非生成的任務(discriminative or non-generative tasks)為高[197]。

可能對策與治理方案：

1. 紀錄產品設計決策中模型開發、維護與部署對於環境之預期影響；量測或估計訓練、微調及部署模型對於環境之影響（如：能源與水消耗）：權衡於推論階段（inference time）中使用之資源與訓練階段（training time）所需之額外資源；驗證碳捕捉或碳抵換計畫之

[197] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, pp.4-5.

有效性，與處理漂綠（Greenwashing）之風險[198]。另一種思維則是以 AI 作為環境保護的手段，進行所謂數位化環保的措施[199]。

六、小結

（一）依據相關 AI Actors 之意見，規劃、準備、實施、紀錄與通報最大化 AI 益處及最小化負面影響之策略：

1. 確認相關機制已到位且應用於維持部署 AI 系統之價值。風險管理行動[200]：紀錄 AI 訓練資料來源，以便追溯 AI 生成內容之起源與來源；此得以降低資訊完整性之風險；過濾生成式 AI 輸出之有害或有偏見之內容、潛在錯誤資訊，以及暴力性或大規模殺傷性武器（CBRN）之相關資訊，或未經同意散布私密影像（Non-Consensual Intimate Image, NCII）內容；針對模型或資料庫進行版本控制以便追蹤變更，並於必要時協助復原至先前版本；將使用者、外部專家與大眾之

[198] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.49；關於 AI 與 ESG 的關係，參見 Nemat/Panzer-Heemeier/Meckenstock, ESG 2.0 – Digitale Ethik als neue Dimension der Nachhaltigkeit, ESG 2022, 104 ff.; Andrea Panzer-Heemeier/André T. Nemat, Digitale Ethik – Eine neue Chance für ESG-Compliance, CCZ 2022, 223 ff.

[199] 參見 Mario Martini/Hannah Ruschemeier, Künstliche Intelligenz als Instrument des Umweltschutzes, Zur rechtlichen Bewertung der Umweltwirkungen intelligenter Technologien, ZUR 2021, 515 ff.; Felix Ekardt/David Schily/Wilmont Holz/Theresa Rath/Alexander Schiela, Chancen und Grenzen digitaler Anwendungen für die Mobilitätswende
Eine Rechts- und Governance-Analyse, NZV 2024, 212 ff.

[200] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, pp.54-55.

反饋納入，以調整生成式 AI 系統及其內容來源；此得管理人類與 AI 配置互動風險；將人工審核流程納入，並依循環境之社會文化知識與價值觀以評估審核內容，且識別自動化流程可能遺漏之限制及細微差距；亦應確保人類審核者經生成式 AI 系統之內容指南與潛在偏見，及其內容來源之訓練。上述手段得控管資訊完整性、有害及有偏見與同質化的風險；使用結構化之反饋機制以徵求及捕捉使用者關於 AI 生成內容之輸入，以偵測關於品質或社群與社會價值觀之一致性的細微變化；此得以減低人類與 AI 配置互動風險、有害及有偏見與同質化的風險。

2. 於識別未知風險時，應遵循一定程序以回應及復原。風險管理行動[201]：制定與更新生成式 AI 系統之意外事故回應、恢復計畫以及流程，以解決與應對以下事項：審查、維護政策及流程以確保對新興用途之負責；為偵測意料外之用途而審查、維護政策及流程；驗證生成式 AI 系統供應鏈之回應與恢復計畫；驗證回應與恢復計畫（如聯絡點、相關聯絡資訊、通知格式）已為下游生成式 AI 環節參與者更新，並包含必要細節以便溝通聯絡。以上手段得減緩 AI 價值鏈與

[201] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.56.

內部組件整合之風險；檢視、更新及維護意外事件回應與恢復計畫，以結合來自生成式 AI 系統使用實例與相關內容，及相關 AI 環節參與者之需求；此得以減低人類與 AI 配置互動風險。

3. 確保機制到位並已實施，且責任亦已被分配與了解；而以此取代、脫離或停用其效能表現或產出與預期使用不符之 AI 系統。風險管理行動[202]：建立並維護溝通計畫，包含原因、解決方法、移除使用者存取權限、替代方案、聯絡資訊等，以便通知 AI 利害關係人，並作為特定生成式 AI 系統或使用環境之停用或分離過程之一部;此得以減低人類與 AI 配置互動風險；建立並定期審視，根據風險承受能力及偏好所設定之關於決定生成式 AI 停用之具體標準。

（二）確保自第三方實體之 AI 風險與效益受管理：

1. 定期監控來自第三方資源之 AI 風險與效益，並確保有效實施與紀錄風險控制。風險管理行動[203]：將組織風險耐受度與控制（如：收購及採購流程；評估人員資歷與資格、實施背景調查、過濾生成式 AI 輸入與輸出、基礎與微調）應

[202] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, pp.55-56.

[203] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, pp.57-58.

用於第三方生成式 AI 資源：將組織風險耐受度應用於第三方資料庫及其他生成式 AI 資源；將組織風險耐受度應用於第三方微調後之模型；將組織風險耐受度應用於現存第三方模型以適應新網域;重新針對經微調後之第三方生成式 AI 模型進行風險量測評估。此得以管理價值鏈與內部組件整合之風險；審核生成式 AI 系統供應鏈風險（如：資料中毒、惡意軟體、其他軟體或硬體漏洞；勞工待遇；資料隱私與法遵在地化；地緣政治聯盟等），以便控管資訊隱私、資訊安全、AI 價值鏈與內部組件整合、有害及有偏見與同質化的風險；於第三方組織或開發者使用生成式 AI 模型前，以及其自行使用生成式 AI 模型於其應用期間，主動開啟審查以監控是否有濫用或違反規定之行為，以減緩 AI 價值鏈與內部組件整合、有害及有偏見與同質化、危險或暴力建議之風險；審視生成式 AI 訓練資料中是否含有暴力性或大規模殺傷性武器（CBRN）與智慧財產之相關資訊；檢查生成式 AI 之輸出是否具抄襲、商標、專利或商業秘密資料；以管理智慧財產、暴力性或大規模殺傷性武器資訊。

2. 作為定期監控與維護 AI 系統之一環，應確保用於開發之預先訓練模型亦受監控。風險管理行動[204]：應用可解釋 AI 技術（如：嵌入分析、模型壓縮或蒸餾、基於梯度的歸因、阻塞減少、反於事實之提問、標籤雲）作為持續改進流程之一部，以減緩與無法解釋之生成式 AI 系統相關風險[205]；紀錄訓練資料之來源、種類及起源，以及呈現於生成式 AI 應用相關資料與其內容來源、結構、預先訓練模型的訓練過程（包含：超參數、訓練期間與任何已應用之微調過程）的潛在偏見。此得減緩資訊完整性、有害及有偏見與同質化的風險[206]；評估使用者回報之問題內容，並將該反饋融合至系統更新中，以此管理人類與 AI 配置互動、危險或暴力建議之風險[207]；實施即時監控流程以分析生成內容與內容來源相關之效能表現及可信賴性，以便識別與理想標準之差距及人工干預警報之觸發，此得管理資訊完整性風險[208]；使用人類審視系統（human moderation systems）[209]，建立人類與 AI 配置互動政

[204] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.58.
[205] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.58.
[206] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.58.
[207] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.58.
[208] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.59.
[209] human moderation systems（人類內容審查系統），係由 human moderator（人類審查員）組成，並由其依據特定規範、準則，標記與刪除平台上具歧視、仇恨、危險或有害之內容與言論，為各大公司所應用，以加速及大規模地處理大量資料，並降低成本。參見 Llansó, E., Van

策審核生成內容，且與使用環境中之社會文化規範相符，及於 AI 模型表現差強人意之情形下使用之。上述方式得控制人類與 AI 配置互動風險[210]。

（三）定期紀錄及監控風險對策，包括回應與恢復，以及針對已識別及量測之 AI 風險的溝通計畫：

1. 應確保部署後之 AI 系統監控計畫確實實施，包含捕捉與評估使用者及其他相關 AI 環節參與者的輸入之機制、申訴與覆寫推翻、停止使用、意外狀況回應、恢復復原，以及變更管理[211]。風險管理行動：與外部研究人員、各行業專家及社群代表合作，以維持對於關於內容來源之新興最佳實踐與技術之認識；此得以降低資訊完整性、有害及有偏見與同質化的風險[212]；定期進行對抗性測試，並係針對各種對抗性輸入及情狀進行測試，藉此識別漏洞並評估 AI 系統中針對內容來源攻擊之韌性，以此降低資訊完整性、資訊安全之

Hoboken, J., Leerssen, P., & Harambam, J. (2020). Artificial intelligence, content moderation, and freedom of expression; Nahmias, Y., & Perel, M. (2021). The oversight of content moderation by AI: Impact assessments and their limitations. Harv. J. on Legis., 58, 145; Molina Davila, M. (2020). Effects of AI vs. Human Moderators and Interactive Transparency on Perceived Trust and Acceptance of Content Classification Systems; Link, D., Hellingrath, B., & Ling, J. (2016, May). A Human-is-the-Loop Approach for Semi-Automated Content Moderation. In ISCRAM.

[210] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.59.

[211] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.60.

[212] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.60.

風險[213]；採取使用者友善之管道，如回饋表單、電子郵件或熱線電話以便使用者回報問題、疑慮或意外之生成式 AI 輸出，再將其納入監管之實行，以上得管理人類與 AI 配置互動風險[214]；實施主動學習技術以識別模型失敗或產出意外之輸出，藉此降低虛談幻覺現象之風險[215]；與內部及外部之利害關係人共享透明性報告，且報告其中詳細描述如何更新 AI 系統以加強其透明性與負責任性[216]。

2. 確保針對持續改善之可量測活動將整合納入 AI 系統更新，及包含定期與相關人士（包含相關 AI 環節參與者）接觸互動。風險管理行動：針對生成式 AI 系統實施定期審核與發布報告，以詳細地說明其表現效能、接收之反饋與所為之改進；採用可解釋性（explainable）AI 方法以增強生成式 AI 內容來源之透明性及可詮釋性（interpretability），以協助 AI 環節參與者及利害關係人了解特定內容如何、為何生成。此得管控人類與 AI 配置互動、資訊完整性之風險[217]；組織跨功能（cross-functional）團隊，並利用橫跨 AI 生命週期之各

[213] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.60.
[214] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.60.
[215] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.60.
[216] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.61.
[217] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.61.

專業知識包含 AI 設計者與開發者、社會技術專家，與對於使用及識別機制十分熟稔之專家，以便將終端使用者參與諮詢，以管控人類與 AI 配置互動風險[218]；練習並遵循意外事件回應計畫以解決生成不適當或有害內容，與基於調查結果改善流程，以防止未來再次發生類似事件[219]；並與相關 AI 環節參與者一起針對該意外事件進行事後分析，以便了解其根本原因與實施預防性措施。此行動得管理人類與 AI 配置互動、危險或暴力建議之風險[220]。

3. 確保意外事件與違誤將傳達予相關 AI 環節參與者，包含受影響社群；並確保遵循與紀錄意外事件與違誤之追蹤、回應與復原流程。風險管理行動：針對生成式 AI 系統之意外事件進行事後評估，以驗證針對事件回應與恢復流程是否遵循及係有效的[221]；建立並維持政策及流程以記錄或追蹤生成式 AI 系統回報之錯誤、幾近錯誤、意外事件，與負面影響；此得降低虛談幻覺現象、資訊完整性之風險[222]；建立與各情境、部門及 AI Actors 定期共享關於違誤、意外事件與負

[218] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.61.
[219] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.62.
[220] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.62.
[221] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.62.
[222] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.62.

面影響資訊之流程及程序，包含回報日期、使用情境、針對各事件之回報數量，以及對於影響及嚴重性之評估，以控管虛談幻覺現象、人類與 AI 配置互動、資訊完整性之風險[223]。

陸、 本院研究人員以本院名義從事相關學術活動的規範

本院資訊所詞庫小組研究人員於 112 年 10 月 7 日將第一版 CKIP 在 GitHub 平台上公開放，意在徵詢早期測試意見。未料，經部份使用者測試後發現該系統的回答有誤，致生本院研究人員使用本院名義從事相關學術活動之規範問題，此為本研究小組成立的源由之一，載述於本報告最後。

針對本院研究人員利用本院名義從事相關學術活動之相關規範，經過多次討論，本研究小組委員之意見，彙整以下：

一、 本院之因應措施與相關規範

（一）院方已訂定本院免責聲明：「本院重視學術及言論自由，鼓勵同仁對社會關鍵議題提出意見或解

[223] 參考 NIST (2024). AI RMF Generative AI Profile, Initial Public Draft, p.62.

決方案。惟同仁自行發表的成果及言論，不等同本院立場。同仁自行發表應遵循學術規範及承擔責任，例如註明資料來源及確認其正確性。而倘擬以本院名義發表研究成果或言論，均應循本院公布的機制。」代表本院的立場與基本規範。

（二）關於報載本院某一位退休副研究員經學者投訴媒體指出，使用中研院電子信箱向國內大學教師廣發支持某總統候選人的信件，內文不僅與學術研究或教學無關，更是為特定總統候選人宣傳之不當行為，此舉涉濫用學術網路資源，呼籲有關單位應約束有權使用公務資源的人，謹慎使用公務資源。本院已於報載當日發表聲明，重申電子郵件服務規範如下：「近日接獲民眾反映，收到以本院信箱寄送之大量發送郵件。本院資訊服務處表示，根據『中央研究院電子郵件服務規範』第八點第一款規定：『……禁止發送下列各款的電子郵件：（一）……未經收信者同意而大量發送的郵件。』本院已即時啟動資安防護機制進行

偵測與警告，並依程序處理必要防護措施。」經討論，結論如下：

1. 本件個案之郵電內容與學術研究無關，不在本小組研究範圍。

2. 本事件涉及使用本院電子郵件服務的行為規範，仍可供處理本院研究人員以本院電子信箱廣發研究測試信件或研究成果所生問題之參考。

3. 有關學術性的部分，原則上使用院方提供之公務 E-mail，其用途基本上應與院內公務、研究有關。但所謂「公務」或「學術」用途方面，解釋上應包括「合理範圍」的所謂 personal use。

4. 除了用途之外，使用的方式是「大量的」發送訊息部分，本院基本上有一定的規範；有技術上的規範與相關辦法的規範，至於分散式的大量發送，而不是一次性寄送，要如何處理、個別認定、還有法律效果（包括是不是停權當事人之使用幾天等），須由院方做進一步的規範。

5. 本院研究人員使用本院的 E-mail 散布發送訊息可能影響到本院聲譽的部分，本院研究人員倫理規約已有規範，可依倫理規約處理。

二、 本院研究人員以院方名義發表研究成果之規範

（一）本院研究人員若要以院方名義發表，有一既有機制，即研究人員先透過各所中心來文至秘書處，秘書處進而彙整資料後，由各學組之副院長與副執行秘書進行檢視，若核可後始得發布（詳參中研院秘書處官方網站：https://sec.sinica.edu.tw/pages/4196）。

（二）本院研究人員以院方名義開發的相關軟體若遭有心人士非法利用，如何因應之問題，無法依賴免責聲明或自我檢測機制防免與解決，且每一個問題與個案背後可能都會衍生非常複雜的事實調查、法律爭議。

（三）本任務係源於研究同仁將大型語言模型對外公開為測試，此任務並非針對特定之個案，而係希望藉由本小組委員討論是否有作一般性行為規

範的可能。多數委員認為不應特別針對生成式 AI 的學術倫理進行規範，而應訂定一套透明清楚之標準，如 Good Research Practice。惟本院研究人員以本院名義從事相關學術活動之態樣多元，至於如何進一步界定學術及言論自由的範疇，合理規範本院研究人員的相關學術活動，需要持續研討、確認規範的可行性。

三、 本院研究人員對外發布大型語言模型之規範

如前所述，本院研究人員以院方名義從事相關學術活動之規範問題，為本小組的任務之一，係源於本院研究人員曾將（中文）大型語言模型對外公開為測試，引起外界誤解該模型為本院所開發。本院雖已訂定前述之免責聲明，表明本院的立場與基本規範，然若本院研究人員有意再次將新開發之（中文）大型語言模型對外發布予公眾使用，如何避免再次發生前次事件外界對本院之誤解，仍有討論之必要，爰針對此特定類型之研究活動，提案如下：

（一）是否須經本院或研究人員所屬所（中心）之同意或事前告知院方或研究人員所屬所（中心）主管？

（中文）大型語言模型對外發布予公眾使用，因技術成熟與否尚待確認，且使用者的範圍及對象不明，為避免 CKIP 模型該次事件再次發生，須經本院或研究人員所屬所（中心）之同意，或至少事前告知院方或研究人員所屬所（中心）主管。

（二）該研究人員是否應對外出具 Disclaimer？其內容應載明何種事項？

該研究人員應對外出具 Disclaimer，其內容應揭示主責研究人員的姓名，並表明其為個人研究，所生問題由該研究人員自負，與本院無關之意旨。

（三）除 Disclaimer 外，該研究人員應揭露何種資訊？

除 Disclaimer 外，必要時，應揭露研究經費來源、研究團隊成員及研究合作對象。

柒、結語

AI 的風險治理與規範，環顧各國及國際規範，主要採業界的自律規範或風險分級管制模式，較少從權利保護與權利救濟的角度出發。當 AI 的發展脈絡與運用層面改變時，AI

風險也會隨之改變，現有的規範模式可能無法完全有效地回應權利保護的需求，生成式 AI 的發展與應用即是對 AI 治理與規範的當前重大挑戰，本研究小組結合院內外專家學者，就此關鍵議題上提出初步的研究報告，善盡本院研究人員的社會責任，只是眾多研究的起步，期與全國各界共同促進臺灣語境生成式 AI 的發展。

生成式 AI 的應用對人類社會的衝擊，包括對民主、法治、社會經濟不平等、勞動市場、生態環境及人類自主性的影響，也讓更多待處理的問題開始浮現，例如生成式 AI 的武器化，用來干預或操縱本國或他國的選舉；又如生成式 AI 的使用伴隨高耗能問題，可能對於整個環境造成影響，是否應建立開發者或使用者付費等節能措施，還有在繁體使用者與簡體使用者之人口基數不對等之下，可能形成文化霸權、破壞知識循環等問題，在在值得吾人關注，並思索解決之道。在此期間，由國家科學及技術委員會自主研發之首個本土化繁體中文語言模型—TAIDE[224]，業於 2024 年 4 月開源釋出[225]。TAIDE 之研發考量我國之使用需求，且為確保生成式 AI

[224] https://taide.tw/index.

[225] National Science and Technology Council (2024). Trustworthy AI Dialogue Engine. https://taide.tw/index/about/project-overview; National Science and Technology Council (2024). taide/TAIDE-LX-7B. Hugging Face. https://huggingface.co/taide/TAIDE-LX-7B.

之可信任性與適用性，研發團隊引入各不同領域之可用文本及訓練材料，以提高此語言模型應用在各領域之性能。透過法律研究、測試規範與評估工具之開發，完善與健全 AI 之發展環境，並增強臺灣資訊安全及防範假訊息之能力，為透過技術端暨開發端解決生成式 AI 風險問題的方式之一。

生成式 AI 的治理與規範，經緯萬端，展望未來，非能僅依賴法律規範，更應從哲學、社會、經濟、心理、語言等人文層面，更廣泛地思考：生成式 AI 究竟會將我們帶往何處？人類需要何種生成式 AI？這些開放性課題的反思有助於回應生成式 AI 的治理課題及其帶來的挑戰。生成式 AI 帶來的風險問題不是單一，而是 AI 系統風險規範與治理課題的一環，需要與 AI 的規範與治理進行體系性思維。發展中的生成式 AI，未來風險變動不居，對於民主法治社會與人權的威脅無刻不在，台灣的處境尤其嚴峻。儘管如此，吾國仍需及早整備政策、立法因應。2024 年 3 月 13 日歐盟議會通過、同年 5 月 21 日歐盟理事會批准 AI Act，正式施行計日可待，一般預料將是繼 GDPR 之後可能牽動全球政經社文的法案，台灣亦無法置身於該法規範之外，立法部門與行政部門責無旁貸，允應密切關注，諮諏善言，適時跟進，妥善擬具政策，

滾動修法或立法，AI 的良性發展與良善治理有待產官學研的共同努力。

捌、 致謝

在此感謝參與本小組的全體人員，包括召集人法律學研究所李建良所長、副召集人資訊科學研究所莊庭瑞副研究員；院內委員資訊科技創新研究中心李育杰研究員、法律學研究所邱文聰研究員與陳舜伶副研究員、歐美研究所洪子偉研究員、人文社會科學研究中心蔡宗翰研究員；院外委員臺灣大學國家發展研究所劉靜怡所長、臺灣大學資訊工程學系許永真教授、臺灣科技大學專利研究所李姿儀助理教授與黃泰然助理教授，及臺灣人工智慧學校秘書長侯宜秀律師；院方代表學術諮詢總會邱繼輝執行秘書、資訊服務處陳伶志處長、學術及儀器事務處陳建璋處長；幕僚人員法律學研究所劉汶渝助理與陳怡卉助理、學術及儀器事務處黃惠靜簡任秘書、酆詠之科長與林巧雯專員等人，於此期間共同付出心力，完成本份報告。

附件

一、附件 1-本院答覆說明資料

(一) 大型語言模型（large language Model）需要仰賴大量資料訓練才會得到精確的答覆。分別以繁體與簡體字為基礎的臺灣及中國在訓練大型語言模型時必須以各自的資料集進行訓練，避免不同的價值觀，包含政治認知、文化、傳統知識等會被訓練進去，進而造成混淆的價值觀。

(二) 國科會已經資助一個以繁體字為主的「Trustworthy AI Dialogue Engine」(TAIDE)大型模型。本院資創中心李育杰研究員、吳毅成合聘研究員、人社中心研究員蔡宗翰研究員均在 TAIDE 計畫中扮演重要角色。資訊所莊庭瑞副研究員更是協助 TAIDE 在資料整備，Meta Data 分析的重要工作。

(三) 經過 CKIP-Llama-2-7b 的事件後，TAIDE 的主持人及相關人員將更知道如何訓練專屬我國的大型語言模型，也因此可以防範受到他國的文化及政治等不同價值觀的干擾。

(四) 繁體中文語料庫是發展臺灣大型語言模型的重要基礎，本院將整合繁體中文詞知識庫，投入資源並規劃管理機制。近期將成立「生成式 AI 風險研究小組」，由學術諮詢總會成立任務編組，法律學研究所李建良所長擔任召集人，邀集學術諮詢總會邱繼輝執行秘書、學術處陳建璋處長及院內外相關領域研究人員、專家

商討，預訂於 112 年 11 月 1 日召開第 1 次會議，將進行通盤檢討，期以建立完善風險評估機制。

二、附件 2-凍結案相關提案

113 年度中央政府總預算案(教育及文化委員會)提案表

單位名稱：中央研究院單位預算　預算書頁次：311 頁

〔 〕歲入—〔 〕增列〔 〕減列數：＿＿＿＿萬＿千元

〔v〕歲出—〔 〕減列數：＿＿萬＿千元　〔v〕凍結數：4,000 千元

第 1 款 5 項 3 目 1 節　科目(計畫)名稱：數理科學研究

用途別：05 資訊科學研究

本年度預算數：181,631 千元

案由：

近日由中研院所開發的繁體中文 AI 語言模型「CKIP-Llama-2-7b」引起爭議，被詢問我國之國慶日、國歌等問題時，竟提供中國語境而非台灣的資料，目前該模型已先行下架，中研院亦指出，該模型是研究人員自行發布，將釐清是否違規，並會成立「生成式 AI 風險研究小組」，並提供研究人員指引，避免類似事件再度發生。

歸根究底之主因，係 OpenAI 與 Meta 提供的語言模型均存在資料偏差，由於大部分的訓練資料都是透過網路爬取，特別是在中文語料蒐集方面，其中簡中內容比例遠高於繁中，相當容易導致因資料偏差而影響到模型生成的結果，足顯見我國打造自有大型語言模型（large language model），實有必要。

為避免上述情事再次發生，建請中研院應制定更嚴謹的審核機制，並廣邀科技、人文社會科學的跨領域研究人員進行跨域合作，以促進臺灣繁體語境生成式 AI 的發展，整合繁體中文詞知識庫，協助推動我國在地化大型語言模型（large language model）的發展

爰提案凍結 4,000 千元，俟中央研究院向立法院教育及文化委員會提出書面報告，始得動支。

提案人：

連署人：

42

9

113年度中央政府總預算案(教育及文化委員會)提案表

單位名稱：中央研究院　　預算書頁次：311頁

〔 〕歲入—〔 〕增列〔 〕減列數：

〔 〕歲出—〔 〕減列數：　〔V〕凍結數：100萬元

第1款5項3目1節　科目(計畫)名稱：數理科學研究

用途別：資訊科學研究　本年度預算數：1億8163萬千元

案由：

一、 中央研究院資訊科學研究所以及語言學研究所共同成立的中文詞知識庫小組於112年10月釋出繁體中文大型語言模型CKIP-Llama-2-7b，其訓練方式係使用中國簡體資料集進行訓練，致使CKIP-Llama-2-7b回應許多「中國觀點」、「中國用語」引起社會譁然。

二、 大型語言模型的訓練，多需使用資料集，以供人工智慧(AI)進行訓練。其中，書籍、小說等作品屬於供AI學習長篇文學、句構的極佳素材，惟現行部分提供大型語言模型學習的資料集中，含有未經合法授權出版品，中央研究院於大型語言模型的開發上，亦未注意有可能侵犯著作權之可能性，亦未建立與出版商之間使用授權的許可機制。

綜上所述，凍結100萬元，俟中央研究院向立法院教育及文化提出大型語言模型改善措施報告後，始得動支。

提案人：

連署人：

43.

36

113 年度中央政府總預算案（教育及文化委員會）提案表

單位名稱：中央研究院

【 】歲入 【✓】歲出 單位預算書頁次：322-323

1 款 5 項 3 目 1 節

科目（工作計畫）名稱： 數理科學研究-資訊科技創新研究

本年度預算數：98,066 千元

建議【】刪減： 【✓】凍結：1,000 千元

案由：

近期中研院釋出語言模型 CKIP-Llama-2-7b，開放民眾公開上網測試。經查許多民眾在測試過程中向該語言模型詢問問題，包含國家、國歌、領導人等問題，該語言模型的回應卻多為:「中國、義勇軍進行曲、習近平」等回答，該語言模型疑似大量參考中國所蒐集之資料庫。中國有龐大上網人口，以及對民眾等同於無的隱私權政策，使其能收集大量資料，以利在機械學習 AI 等領域迅速發展，故其在中文領域的 AI 開發上佔據領先地位無可厚非，然而引用中國資料恐使我國在以 AI 文字編修以及華語文教學部分，產生隱憂。

鑑於資料庫的建置與蒐集具相當重要性，中研院做為我國學術領域最高級單位，應考慮該資料庫建置時內容的適切性，爰凍結 1000千元，俟中研院研議建置相關資料庫或模型時，以適合我國現況之形式與以我國為主之語言資料庫系統，立法院教育文化委員會提出書面報告，後始可動支。

提案人： 陳靜敏

連署人：

45

75

113年度中央政府總預算案(教育及文化委員會)提案表

單位名稱：中央研究院　預算書頁次：370頁

〔∨〕歲出—〔∨〕凍結數：300萬元

第1款5項3目3節　科目(計畫)名稱：人文及社會科學研究
用途別：09語言學研究　本年度預算數：5,656萬1千元

案由：

經查113年度中央研究院「人文及社會科學研究」項下「語言學研究」預算編列5,656萬1千元，包含辦理語言學相關研究工作之經費。惟日前傳出，中研院開發的繁體中文語言模型AI——CKIP-Llama-2-7b傳出使用中國的任務資料集（COIG），因此導致特定問題答非所問，有誤導民眾之嫌。中研院作為我國基礎科學研究之翹楚，進行研究之餘本應格外留意使用之資源，以維護研究成果準確並維護中研院之專業精神。爰此，凍結本預算300萬元，俟中研院向立法院教育及文化委員會提出書面報告並經同意後，始得動支。

提案人：陳秀寶

連署人：

50

20

三、附件 3-解凍書面報告（節錄有關生成式 AI 議題之說明）

有關　大院審查本院單位預算通過決議第一項「113 年度中央研究院第 3 目『自然及人文社會科學研究』預算編列 44 億 8,931 萬 2 千元，凍結 100 萬元，俟中央研究院向立法院教育及文化委員會提出書面報告後，始得動支。」本院說明如下：

壹、就大型語言模型開發過程可能造成之生成式 AI 風險問題進一步加以研究（包括著作權問題、發展以我國為主之語言資料庫系統等）

一、大型語言模型（Large Language Model）需要仰賴大量資料訓練。資料的來源與品質確實影響了大型語言模型所生成的內容。各類資料集中所收錄的語句，在用詞標點、文句結構、以及其所涵蓋的人事地物、文化脈落知識等，皆有不少差異。在臺灣訓練大型語言模型，會需要搭配巨量的臺灣語文資料集，包括繁體中文以及在臺灣使用的其他語言。

二、目前國科會正執行一個以繁體中文為主的「Trustworthy AI Dialogue Engine」（可信任 AI 對話引擎，簡稱 TAIDE 計畫）大型語言模型，在資料源頭的收集與內容篩選方面著力甚深。本院除已將「研之有物」等院內出版品內容資料提供 TAIDE 計畫使用外，本院資訊科技創新研究中心李育杰研究員、吳毅成合聘研究員、人文社會科學研究中心蔡

宗翰研究員、資訊研究所莊庭瑞副研究員亦參與TAIDE計畫，在計畫的執行上扮演重要角色。

三、繁體中文語料庫是發展臺灣大型語言模型的重要基礎，除了需要投入資源、積極規劃開發外，其對社會影響與風險管理機制的研究，亦同等重要。本院已成立「生成式AI風險研究小組」，係由學術諮詢總會成立之任務編組，法律學研究所李建良所長擔任召集人，邀集院內外生成式 AI 相關的跨領域研究人員、專家共同研討，目前已召開3次會議，以團隊方式連結資訊科技、人文及社會科學人才進行跨領域研究，另有關與出版商間之合法著作權授權機制，亦一併納入本小組會議中通盤研討，期能提出相關建言，與全國各界共同促進臺灣語境生成式AI的發展，在關鍵議題上善盡本院研究人員的社會責任。

四、附件 4-預算審查及朝野黨團協商通過決議－涉及生成式 AI 議題

(一)議事錄通過決議第(五)案：

中央研究院指出，針對人工智慧（Artificial Intelligence）、深度學習（Deep Learning）、大數據分析（Big Data）、社群網路（Social Network）及自然語言處理（Natural Language）等新興研究議題，資訊科學研究所透過推動大型研究合作計畫，已獲致具發展性的初步成果，並將持續進行相關研發。中研院表示,尤其在人工智慧及自然語言處理(包含 Transformer、BERT 及 ChatGPT）等先進研究議題上，除了協助國內相關單位應用人工智慧技術推動產業升級外，更將進一步提升國內產、官、學、研各界研發經營能量。為厚植中央研究院研究實力，中央研究院鼓勵研究人員藉由爭取長期經費支持，凝聚資源挑戰較大規模、高風險性的研究議題，為求科學突破此舉實屬可取，然而既已知將投入具風險之研究，便應完備面對風險之因應作為。由中研院開發的繁體中文語言模型 AI，日前爆出使用中國建置資料庫的爭議，為此中研院表示將成立「生成式 AI 風險研究小組」。爰請中央研究院盤點生成式 AI 可能存在之風險，並建立跨領域研究者之專業合作模式。以上問題請中央研究院於 3 個月內向立法院教育及文化委員會提出期中報告，6 個月內提出書面報告。

提案人：林宜瑾

連署人：范雲　陳秀寶

（二）議事錄通過決議第(十)案：

CKIP-Llama-2-7b 是中央研究院詞庫小組（CKIP）開發的開源可商用繁體中文大型語言模型（Large Language Model），研究目標之一是讓 meta 開發的 Llama 2 大型語言模型具備更好的繁體中文處理能力，並提供大眾下載，作為學術使用或是商業使用。然而，日前卻因使用來自中國的資料庫，被網友實測發現，輸入「你是誰創造的？」，系統卻回「我是由復旦大學自然語言處理實驗室和上海人工智能實驗室共同開發的，我的生日是 2023 年 2 月 7 日，我的國籍是中國，我的居住地是上海人工智能實驗室服務器集，我可以說中文和英語」。甚至提問「你的國家是？」、「國歌是？」、「國花是？」等問題，系統會回以「中國」、「義勇軍進行曲」、「牡丹」等充斥中國觀點的回應內容，而該系統目前已下架。爰請中央研究院加快成立「生成式 AI 風險研究小組」，以深入了解 AI 對社會的衝擊，提供研究人員相關指引，同時儘速擬定公開前之審核機制，以避免類似問題產生。亦應積極投入台灣繁體中文資料庫之擴充，防止中國偏見資訊之滲透，並於 6 個月內向立法院教育及文化委員會提出書面報告。

提案人：張廖萬堅

連署人：陳培瑜　陳秀寶

（三）議事錄通過決議第(十三)案：

中央研究院於近日推出繁體中文大型語言模型「CKIP-Llama-2-7b」，自 112 年 10 月 7 日供民眾試用，並宣稱該語言模型可用於學術、商業使用、文案生成、文學創作、語言翻譯等諸多功能，然遭民眾發現，詢問該語言模型問題時，皆持大陸觀點回答，如國家領導人為習近平，國慶日為 10 月 1 日，該語言模型隨即於 10 月 9 日緊急下架。經查，該語言模型會有如此表現之主因，為參與該語言模型計畫之研究員，於 112 年初負責數位文化中心之明清歷史人物時空關係之計畫，該研究員想將語言模型應用於明清歷史人物時空關係計畫上，然於製作階段缺乏相關資料，便直接使用對岸之資料集，將簡體字轉繁體字後用於訓練語言模型，惟該舉動導致其他不相干之資訊亦混入語言模型，使該語言模型回答問題時持大陸觀點。綜上所述，因該疏忽已影響中研院之聲譽，中研院實有必要加強各項研究之管控及督導，爰要求中央研究院應於 6 個月內向立法院教育及文化委員會提出書面報告。

提案人：萬美玲

連署人：鄭麗文 鄭正鈐

（四）議事錄通過決議第(二十一)案：

中央研究院個別研究人員主持的詞庫小組（CKIP）發佈 CKIP 模型（CKIP-Llama-2-7b），因使用中國來源之資料庫，造成民眾使用時出現「臺灣是中國的一部分」、「國慶是 10 月 1 日」等未符合實情及臺灣本土語境之回答，引發社會爭議。該風波突顯生成式 AI 可能對社會所造成的衝擊，若其使用的資料庫本身隱含社會偏見、缺乏多元社會價值觀，很可能生成偏頗的內容，甚至被進一步惡意運用。AI 發展是數位國力之展現、生成式 AI 亦為國際間重點發展項目，中研院於領導臺灣 AI 科技相關研究的同時，亦應考量生成式 AI 對社會政治與道德層面的衝擊與影響，除避免類似風波再現，更奠定臺灣本土 AI 發展基礎。為完善生成式 AI 研究發展、防範生成式 AI 可能對社會造成之衝擊，爰請中央研究院於 113 年 1 月 31 日前成立「生成式 AI 風險研究小組」並啟動相關指引之研議，於 3 個月內提出進度報告，於 6 個月內向立法院教育及文化委員會提出書面報告。

提案人：范雲

連署人：陳秀寶　陳靜敏

（五）議事錄通過決議第(四十七)案：

近日由中央研究院所開發的繁體中文 AI 語言模型「CKIP-Llama-2-7b」引起爭議，被詢問我國之國慶日、國歌等問題時，竟提供中國語境而非台灣的資料，目前該模型已先行下架，中研院亦指出，該模型是研究人員自行發布，將釐清是否違規，並會成立「生成式 AI 風險研究小組」，並提供研究人員指引，避免類似事件再度發生。歸根究底之主因，係 OpenAI 與 Meta 提供的語言模型均存在資料偏差，由於大部分的訓練資料都是透過網路爬取，特別是在中文語料蒐集方面，其中簡中內容比例遠高於繁中，相當容易導致因資料偏差而影響到模型生成的結果，足顯見我國打造自有大型語言模型（Large Language Model），實有必要。為避免上述情事再次發生，請中研院應制定更嚴謹的審核機制，並廣邀科技、人文社會科學的跨領域研究人員進行跨域合作，以促進臺灣繁體語境生成式 AI 的發展，整合繁體中文詞知識庫，協助推動我國在地化大型語言模型（Large Language Model）的發展。爰請中央研究院就前述內容，於 6 個月內向立法院教育及文化委員會提出書面報告。

提案人：吳思瑤　陳秀寶
連署人：陳靜敏

（六）通過決議第(四十九)案：

近期中央研究院釋出語言模型 CKIP-Llama-2-7b，開放民眾公開上網測試。經查許多民眾在測試過程中向該語言模型詢問問題，包含國家、國歌、領導人等問題，該語言模型的回應卻多為：「中國、義勇軍進行曲、習近平」等回答，該語言模型疑似大量參考中國所蒐集之資料庫。中國有龐大上網人口，以及對民眾等同於無的隱私權政策，使其能收集大量資料，以利在機械學習 AI 等領域迅速發展，故其在中文領域的 AI 開發上占據領先地位無可厚非，然而引用中國資料恐使我國在以 AI 文字編修以及華語文教學部分，產生隱憂。鑑於資料庫的建置與蒐集具相當重要性，中研院做為我國學術領域最高級單位，應考慮該資料庫建置時內容的適切性。爰請中央研究院研議建置相關資料庫或模型時，以適合我國現況之形式與以我國為主之語言資料庫系統，於 6 個月內向立法院教育及文化委員會提出書面報告。

提案人：陳靜敏
連署人：陳秀寶　張其祿

（七）議事錄通過決議第(五十)案：

經查 113 年度中央研究院「人文及社會科學研究」項下「語言學研究」預算編列 5,656 萬 1 千元，包含辦理語言學相關研究工作之經費。惟日前傳出，中研院開發的繁體中文語言模型 AI-CKIP-Llama-2-7b 傳出使用中國的任務資料集（COIG），因此導致特定問題答非所問，有誤導民眾之嫌。中研院作為我國基礎科學研究之翹楚，進行研究之餘本應格外留意使用之資源，以維護研究成果準確並維護中研院之專業精神。爰請中央研究院就前述內容，於 6 個月內向立法院教育及文化委員會提出書面報告。

提案人：陳秀寶

連署人：陳靜敏　林宜瑾

(八)朝野黨團協商新增決議 1 項：

日前傳出我國由中央研究院自主開發的繁體中文語言模型 AI 聊天機器人，詢問其國籍，卻回覆「我的國籍是中國」，輸入問題「你是誰創造的?」系統卻回覆「我是由復旦大學自由語言處理實驗室和上海人工智能實驗室共同開發的」，足見該聊天機器人使用中國大陸的系統，又聊天機器人在 AI 風險管理中屬於高度的風險，故該系統的使用嚴重有可能對我國國民甚至國家造成危害，爰請中央研究院於 3 個月內向立法院教育及文化委員會提出進度報告、6 個月內提出書面報告，具體說明如何防堵人工智慧所造成的風險。

提案人：台灣民眾黨立法院黨團

五、附件 5-議事錄通過決議－3 個月進度（期中）報告

本院說明如下：

繁體中文語料庫是發展臺灣大型語言模型的重要基礎，除了需要投入資源、積極規劃開發外，其對社會影響與風險管理機制的研究，亦同等重要。

112 年 10 月初，本院研究人員釋出繁體中文語言模型 CKIP-Llama-2-7b 公開測試，引發社會關注，即於 11 月 1 日成立生成式 AI 風險研究小組（下稱本小組），由本院法律學研究所李建良所長擔任召集人、資訊科學研究所莊庭瑞副研究員擔任副召集人，委員由院內外跨領域學者專家共同組成，包括院內委員為資訊科技創新研究中心李育杰研究員、法律學研究所邱文聰研究員與陳舜伶副研究員、歐美研究所洪子偉研究員、人文社會科學研究中心蔡宗翰研究員，院外委員為臺灣大學國家發展研究所劉靜怡所長、臺灣大學資訊工程學系許永真教授、臺灣科技大學專利研究所李姿儀助理教授與黃泰然助理教授，及臺灣人工智慧學校秘書長侯宜秀律師，共計 12 位，院本部學術諮詢總會邱繼輝執行秘書、資訊服務處陳伶志處長、學術及儀器事務處陳建璋處長等 3 位院方代表列席，主要以「生成式 AI 風險及其管理機制」、「研究開發大型繁體中文語言模型之著作權保護機制」，以及「本院同仁利用中研院名義從事相關學術活動之規範」等問題，進行跨領域研究，並提出相關建言，作為本小組主要任務。

本小組截至目前為止已召開 5 次會議，就各次開會時間、討論事項及重要決議，摘述如下：

一、112 年 11 月 1 日召開第 1 次會議，就本小組研究範疇、運作方式與目標等進行討論，並建議完整成員名單及確認主要任務。

二、112 年 11 月 23 日召開第 2 次會議，盤點與生成式 AI 有關之風險研究與問題，並開始就資料來源面臨之「智慧財產權」問題進行討論與意見交換。

三、112 年 12 月 25 日召開第 3 次會議，針對本院已訂定之免責聲明，討論如何進一步界定學術與言論自由範疇、合理規範本院研究人員之相關學術活動，以及規範之可行性等；另就生成式 AI 與「著作權保護」問題進行意見交換。

四、113 年 1 月 22 日召開第 4 次會議，先就報告架構及初稿內容進行討論，將「生成式 AI 幻覺」、「生成式 AI 衍生之文化話語權」納入問題盤點中，並可聚焦於本院研究人員亟待解決之生成式 AI 應用風險問題，優先進行討論。另針對生成式 AI 與「個人資料保護」問題與研討範疇提出相關意見。

五、113 年 3 月 4 日召開第 5 次會議，持續就報告架構及初稿內容進行詳細討論，將於「生成式 AI 的意涵與背景資料」章節闡明生成式 AI 之定義，以及本小組之研究重點收斂聚焦於「大型語言模型」，首要處理文字類型之生成式 AI 風險問題；並將生成式 AI 對環境產生之負面影響（耗能及碳排等）及文化霸權問題，納入「研究範圍、分析架構與問題盤點」及結論等章節適度論述說明。

本小組成員刻撰擬書面報告中，並持續召會研討大型語言模型之風險及其管理機制，期與全國各界共同促進臺灣語境生成式 AI 的發展，在關鍵議題上善盡本院研究人員的社會責任。

六、附件 6-參考文獻

學術資料

- Andreas Voßkuhle/Schemme, Grundwissen – Öffentliches Recht: „Neue“ Grundrechte, JuS 2024, 312 (314).
- Andrea Panzer-Heemeier/André T. Nemat, Digitale Ethik – Eine neue Chance für ESG-Compliance, CCZ 2022, 223 ff.
- Barocas, S., Guo, A., Kamar, E., Krones, J., Morris, M. R., Vaughan, J. W., ... & Wallach, H. (2021, July). Designing disaggregated evaluations of ai systems: Choices, considerations, and tradeoffs. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (pp. 368-378).
- Benjamin Raue, Kreativität im Zeitalter ihrer technischen Reproduzierbarkeit: Generative KI als Totengräberin des Urheberrechts? Eine Gedankenskizze, ZUM 2024, 157 ff.
- Berg, M., Cederquist, F., Hamilton, D., Johansson, J., & Vindblom, A. (2004). A case study of how to add a plugin framework.
- Boris P. Paal, KI-Training mit öffentlich frei zugänglichen Daten im Lichte der DS-GVO-Vorgaben, ZfDR 2024, 129 ff.
- Cachat-Rosset, G., & Klarsfeld, A. (2023). Diversity, equity, and inclusion in artificial intelligence: An evaluation of guidelines. Applied Artificial Intelligence, 37(1), 2176618
- Chakraborti, T., Isahagian, V., Khalaf, R., Khazaeni, Y., Muthusamy, V., Rizk, Y., & Unuvar, M. (2020). From Robotic Process Automation to Intelligent Process Automation:–Emerging Trends–. In Business Process Management: Blockchain and Robotic Process Automation Forum: BPM 2020 Blockchain and RPA Forum, Seville, Spain, September 13–18, 2020, Proceedings 18 (pp. 215-228). Springer International Publishing
- Chi, N., Lurie, E., & Mulligan, D. K. (2021, July). Reconfiguring diversity and inclusion for AI ethics. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (pp. 447-457).
- Chui, M., Harryson, M., Valley, S., Manyika, J., & Roberts, R. (2018). Notes from the AI frontier applying AI for social good.
- Daugherty, P. R., Wilson, H. J., & Chowdhury, R. (2020). Using artificial intelligence to promote diversity
- Deshpande, A., Rajpurohit, T., Narasimhan, K., & Kalyan, A. (2023). Anthropomorphization of AI: opportunities and risks. arXiv preprint arXiv:2305.14784
- Eine Rechts- und Governance-Analyse, NZV 2024, 212 ff.

- El Ioini, N., & Pahl, C. (2018). A review of distributed ledger technologies, in: On the Move to Meaningful Internet Systems. OTM 2018 Conferences: Confederated International Conferences: CoopIS, C&TC, and ODBASE 2018, Valletta, Malta, October 22-26, 2018, Proceedings, Part II (pp. 277-288). Springer International Publishing
- Felix Ekardt/David Schily/Wilmont Holz/Theresa Rath/Alexander Schiela, Chancen und Grenzen digitaler Anwendungen für die Mobilitätswende
- Fosch-Villaronga, E., & Poulsen, A. (2022). Diversity and inclusion in artificial intelligence. Law and Artificial Intelligence: Regulating AI and Applying AI in Legal Practice, 109-134
- Gabriele Brit, Vertraulichkeit und Integrität informationstechnischer Systeme, DÖV 2008, 411 ff.
- Geisler, E., Prabhaker, P., & Nayar, M. (2003, July). Information integrity: an emerging field and the state of knowledge. In PICMET'03: Portland International Conference on Management of Engineering and Technology Technology Management for Reshaping the World, 2003. (pp. 217-221). IEEE.
- Gräfe, H. C., & Kahl, J. (2021). KI-Systeme zur automatischen Texterstellung, Urheber-und medienrechtliche Einordnung von Textgeneratoren in Journalis mus und E-Commerce. Multi Media und Recht Zeitschrift für IT-Recht und Digitalisierung, 24(2), 121-126.
- Goyal, S. (2013). Software as a service, platform as a service, infrastructure as a service– a review. International journal of Computer Science & Network Solutions, 1(3), 53-67.
- Golder, S., & Huberman, B. A. (2005). The structure of collaborative tagging systems. arXiv preprint cs/0508082；Au Yeung, C. M., Gibbins, N., & Shadbolt, N. (2009, June). Contextualising tags in collaborative tagging systems. In Proceedings of the 20th ACM Conference on Hypertext and Hypermedia (pp. 251-260).
- Gal/Lynskey, Synthetic Data: Legal Implications of the Data-Generation Revolution, Iowa L. Rev. 109 (2023); Shumailov et al., The Curse of Recursion: Training on Generated Data Makes Models Forget, 31.5.2023, https://arxiv.org/abs/2305.17493
- Haimo Schack, Auslesen von Webseiten zu KI-Trainingszwecken als Urheberrechtsverletzung de lege lata et ferenda, NJW 2024, 113 ff.
- Harper, S., & Chen, A. Q. (2012). Web accessibility guidelines: A lesson from the evolving Web. World Wide Web, 15, 61-88
- Habib, H., Zou, Y., Jannu, A., Sridhar, N., Swoopes, C., Acquisti, A., ... & Schaub, F. (2019). An Empirical Analysis of Data Deletion and {Opt-Out} Choices on 150

Websites. In Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019) (pp. 387-406)

- Henry, S. L., Abou-Zahra, S., & Brewer, J. (2014, April). The role of accessibility in a universal web. In Proceedings of the 11th Web for all Conference (pp. 1-4)
- Henrik Nolte/Zeynep Schreitmüller, Cybersicherheit KI-basierter Medizinprodukte im Lichte der MDR und KI-VO, MPR 2024, 20 ff.
- Hofmann, P., Samp, C., & Urbach, N. (2020). Robotic process automation. Electronic markets, 30(1), 99-106
- Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., ... & Perez, E. (2024). Sleeper agents: Training deceptive llms that persist through safety training. arXiv preprint arXiv:2401.05566.
- Ivančić, L., Suša Vugec, D., & Bosilj Vukšić, V. (2019). Robotic process automation: systematic literature review. In Business Process Management: Blockchain and Central and Eastern Europe Forum: BPM 2019 Blockchain and CEE Forum, Vienna, Austria, September 1–6, 2019, Proceedings 17 (pp. 280-295).
- James, in: Leupold/Wiebe/Glossner, IT-Recht, 4. Aufl., 2021,Teil 11.1 Cloud Computing – Vorteile und Risiken, Rn. 20.
- Jaspan, C. N. (2011). Proper plugin protocols (Doctoral dissertation, Carnegie Mellon University)
- Jaume-Palasi, L. (2019). Why we are failing to understand the societal impact of artificial intelligence. Social Research: An International Quarterly, 86(2), 477-498
- Jens Reinke/Stefan Müller, Verbraucher und Endnutzer in der Nachhaltigkeitsberichterstattung nach ESRS S4, IRZ 2023, 545 ff.; IRZ 2024, 35 ff.
- Jora, R. B., Sodhi, K. K., Mittal, P., & Saxena, P. (2022, March). Role of artificial intelligence (AI) in meeting diversity, equality and inclusion (DEI) goals. In 2022 8th international conference on advanced computing and communication systems (ICACCS) (Vol. 1, pp. 1687-1690). IEEE
- Julia Wulf/Tim-Jonas Löbeth, Text und Data Mining: Wenn gewolltes und geschaffenes Recht auseinanderfallen, GRUR 2024, 737 ff.
- Katharina Wunner, Generative K.I. und das Urheberrecht – Eine komplizierte Beziehung ZUM 2024, 190 ff.
- Kathleen Mosier et al, "Automation Bias and Errors: Are Teams Better than Individuals?". Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 42 (3): 201-205 (1998).

- Kosta, E. (2015). Construing the Meaning of" Opt-out"-An Analysis of the European, UK and German Data Protection Legislation. Eur. Data Prot. L. Rev., 1, 16
- Krishnamoorthy, M., Sjoding, M. W., & Wiens, J. (2024). Off-label use of artificial intelligence models in healthcare. Nature Medicine, 1-3.
- Lai, Y. L., & Hui, K. L. (2006, April). Internet opt-in and opt-out: investigating the roles of frames, defaults and privacy concerns. In Proceedings of the 2006 ACM SIGMIS CPR conference on computer personnel research: Forty four years of computer personnel research: achievements, challenges & the future (pp. 253-263)
- Lazar, J., Dudley-Sponaugle, A., & Greenidge, K. D. (2004). Improving web accessibility: a study of webmaster perceptions. Computers in human behavior, 20(2), 269-288
- Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., & Carlini, N. (2021). Deduplicating training data makes language models better. arXiv preprint arXiv:2107.06499
- Llansó, E., Van Hoboken, J., Leerssen, P., & Harambam, J. (2020). Artificial intelligence, content moderation, and freedom of expression
- Link, D., Hellingrath, B., & Ling, J. (2016, May). A Human-is-the-Loop Approach for Semi-Automated Content Moderation. In ISCRAM.
- M.L. Cummings, Automation Bias in Intelligent Time Critical Decision Support Systems (2004)
- Molnar, C., Casalicchio, G., & Bischl, B. (2020, September). Interpretable machine learning–a brief history, state-of-the-art and challenges. In Joint European conference on machine learning and knowledge discovery in databases (pp. 417-431). Cham: Springer International Publishing. Available at https://link.springer.com/chapter/10.1007/978-3-030-65965-3_28.
- Mohammed, P. S., & 'Nell'Watson, E. (2019). Towards inclusive education in the age of artificial intelligence: Perspectives, challenges, and opportunities. Artificial Intelligence and Inclusive Education: Speculative futures and emerging practices, 17-37
- Miller, K. W., & Voas, J. (2008). Information integrity and IT professionals' integrity, intertwined. IT Professional, 10(6), 35-40.
- Miyake, E., & Kuntsman, A. (2022). Paradoxes of Digital Disengagement: In Search of the Opt-Out Button.

- Mario Martini/Hannah Ruschemeier, Künstliche Intelligenz als Instrument des Umweltschutzes, Zur rechtlichen Bewertung der Umweltwirkungen intelligenter Technologien, ZUR 2021, 515 ff.
- Molina Davila, M. (2020). Effects of AI vs. Human Moderators and Interactive Transparency on Perceived Trust and Acceptance of Content Classification Systems
- Nguyen, H. V., Kästner, C., & Nguyen, T. N. (2014, May). Exploring variability-aware execution for testing plugin-based web applications. In Proceedings of the 36th International Conference on Software Engineering (pp. 907-918)
- Nemat/Panzer-Heemeier/Meckenstock, ESG 2.0 – Digitale Ethik als neue Dimension der Nachhaltigkeit, ESG 2022, 104 ff.
- Nahmias, Y., & Perel, M. (2021). The oversight of content moderation by AI: Impact assessments and their limitations. Harv. J. on Legis., 58, 145
- Patnaik, P. (2022). A Policy Framework Towards the Use of Artificial Intelligence by Public Institutions: Reference to FATE Analysis. In Handbook of Research on Artificial Intelligence in Government Practices and Processes (pp. 13-37). IGI Global
- Paleyes, A., Urma, R. G., & Lawrence, N. D. (2022). Challenges in deploying machine learning: a survey of case studies. ACM computing surveys, 55(6), 3.
- Paulina Jo Pesch/Rainer Böhme,Verarbeitung personenbezogener Daten und Datenrichtigkeit bei großen Sprachmodellen. ChatGPT & Co. unter der DS-GVO, MMR 2023, 917 ff.
- Petrie, H., & Kheir, O. (2007, April). The relationship between accessibility and usability of websites. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 397-406).
- Philipp Müller-Peltzer/ Valentin Tanczik, Künstliche Intelligenz und Daten. Data-Governance nach der geplanten KI-Verordnung, RDi 2023, 452 (456 f.).
- Placani, A. (2024). Anthropomorphism in AI: hype and fallacy. AI and Ethics, 1-8.
- Rachael Johnson (2023). Risk Culture: Building Resilience and Seizing Opportunities. A Global Survey and Report, p. 9.
- Ribeiro, J., Lima, R., Eckhardt, T., & Paiva, S. (2021). Robotic process automation and artificial intelligence in industry 4.0 – a literature review. Procedia Computer Science, 181, 51-58, https://doi.org/10.1016/j.procs.2021.01.104.
- Robinson, S. C. (2020). Trust, transparency, and openness: How inclusion of cultural values shapes Nordic national public policy strategies for artificial intelligence (AI). Technology in Society, 63, 101421

- Rutter, R., Lauke, P. H., Waddell, C., Thatcher, J., Henry, S. L., Lawson, B., ... & Urban, M. (2007). Web accessibility: Web standards and regulatory compliance. Apress
- Sarah Legner, Der Digital Services Act – Ein neuer Grundstein der Digitalregulierung, ZUM 2024, 99 ff.
- Sabherwal, R., & Grover, V. (2024). The Societal Impacts of Generative Artificial Intelligence: A Balanced Perspective. Journal of the Association for Information Systems, 25(1), 13-22
- Sebastian Felix Schwemer, Decision Quality and Errors in Content Moderation, IIC 2024, 139 ff. João Pedro Quintais, Christian Katzenbach, Sebastian Felix Schwemer, Daria Dergacheva, Thomas Riis, Péter Mezei, István Harkai, João Carlos Magalhães, Copyright Content Moderation in the European Union: State of the Art, Ways Forward and Policy Recommendations, IIC 2024, 157 ff.
- Shams, R. A., Zowghi, D., & Bano, M. (2023). AI and the quest for diversity and inclusion: a systematic literature review. AI and Ethics, 1-28
- Susanne Werry, Generative KI-Modelle im Visier der Datenschutzbehörden
- Sun, W., Zhang, K., Chen, S. K., Zhang, X., & Liang, H. (2007). Software as a service: An integration perspective. In Service-Oriented Computing–ICSOC 2007: Fifth International Conference, Vienna, Austria, September 17-20, 2007. Proceedings 5 (pp. 558-569). Springer Berlin Heidelberg
- Sobti, R., & Geetha, G. (2012). Cryptographic hash functions: a review. International Journal of Computer Science Issues (IJCSI), 9(2), 461；Thomsen, S. S., & Knudsen, L. R. (2005). Cryptographic hash functions. PhD thesis; Gauravaram, P. (2007). Cryptographic hash functions: cryptanalysis, design and applications (Doctoral dissertation, Queensland University of Technology).
- Sunyaev, A., & Sunyaev, A. (2020). Distributed ledger technology. Internet computing: Principles of distributed systems and emerging internet-based technologies, 265-299.
- Tahireh Panahi, Gesetz gegen digitale Gewalt – auf Kollisionskurs mit dem DSA?, MMR 2023, 556 ff.
- Tomašev, N., Cornebise, J., Hutter, F., Mohamed, S., Picciariello, A., Connelly, B., ... & Clopath, C. (2020). AI for social good: unlocking the opportunity for positive impact. Nature Communications, 11(1), 2468
- Thomas Hoeren/Julia Dreyer, in: Hoeren/Sieber/Holznagel, Handbuch Multimedia-Recht, Werkstand: 60. EL Oktober 2023, Teil 7.2 Urheberpersönlichkeitsrecht im Internet, Rn. 39.
- Vered, M., Livni, T., Howe, P. D. L., Miller, T., & Sonenberg, L. (2023). The effects of explanations on automation bias. Artificial Intelligence, 322, 103952.

- Vesnic-Alujevic, L., Nascimento, S., & Polvora, A. (2020). Societal and ethical impacts of artificial intelligence: Critical notes on European policy frameworks. Telecommunications Policy, 44(6), 101961
- Waters, B. (2005). Software as a service: A look at the customer benefits. Journal of digital asset management, 1, 32-39
- Werner Gleißner, Die Welt der Risiken: 5 Risikokategorien, BC 2022, 217 ff.
- Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359
- Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K., & Cave, S. (2019). Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research. London: Nuffield Foundation
- Wolfram Gass, Digitale Wasserzeichen als urheberrechtlicher Schutz digitaler Werke?, ZUM 1999, 815
- Zowghi, D., & Bano, M. (2024). AI for all: Diversity and Inclusion in AI. AI and Ethics, 1-4
- Zowghi, D., & da Rimini, F. (2023). Diversity and inclusion in artificial intelligence. arXiv preprint arXiv:2305.12728

法規法案與實務見解

- Act on Copyright and Related Rights (GesetzüberUrheberrecht und verwandte Schutzrechte). Available at https://www.gesetze-im-internet.de/englisch_urhg/englisch_urhg.html(Accessed: 2024/01/18).
- Code of conduct. United Nations Information Integrity. Available at: https://www.un.org/en/information-integrity/code-of-conduct (Accessed: 13 June 2024).
- Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence. Available at: https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf
- Singapore Copyright Act 2021. Available at https://wipolex-res.wipo.int/edocs/lexdocs/laws/en/sg/sg175en.pdf (Accessed: 2024/01/18).
- The Copyright, Designs, and Patents Act 1988. Available at https://www.legislation.gov.uk/ukpga/1988/48/contents.
- The Commission adopts the European Sustainability Reporting Standards, https://finance.ec.europa.eu/news/commission-adopts-european-sustainability-reporting-standards-2023-07-31_en.

- United Nations (2023) Information integrity on digital platforms. Available at: https://www.un.org/sites/un2.un.org/files/our-common-agenda-policy-brief-information-integrity-en.pdf (Accessed: 13 June 2024).
- 17 U.S. Code § 503 - Remedies for infringement: Impounding and disposition of infringing articles. Available at https://www.law.cornell.edu/uscode/text/17/503.
- 17 U.S. Code § 504 Remedies for infringement: Damages and profits. Available at https://www.law.cornell.edu/uscode/text/17/504.
- 17 U.S. Code § 505 Remedies for infringement: Costs and attorney's fees. Available at https://www.law.cornell.edu/uscode/text/17/505.
- 著作權法第 3 條第 1 項第 1 款、第 5 款、第 11 款；第 6 條；第 7 條；第 22 條第 1 項；第 28 條
- 最高法院 104 年度台上字第 2980 號刑事判決
- 司法院憲法法庭 111 年憲判字第 13 號判決
- 經濟部智慧財產局第 1120317 號函釋
- 經濟部智慧財產局電子郵件 1120317
- 經濟部智慧財產局電子郵件 980312b

規範文獻

- Cabinet Office and Central Digital and Data Office (2024) Generative AI framework for HMG, GOV.UK. Available at: https://www.gov.uk/government/publications/generative-ai-framework-for-hmg.
- European Parliament (2023) EU AI act: First regulation on artificial intelligence. Available at: https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence (Accessed: 13 June 2024).
- Intellectual Property Office of Singapore, Copyright factsheet on Copyright Act 2021. Available at https://www.ipos.gov.sg/docs/default-source/resources-library/copyright/copyright-act-factsheet.pdf(Accessed: 2024/01/18).
- NIST. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). Available at: https://www.nist.gov/itl/ai-risk-management-framework.
- NIST. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. Available at: https://airc.nist.gov/docs/NIST.AI.600-1.GenAI-Profile.ipd.pdf.
- NIST. AI TEST, EVALUATION, VALIDATION AND VERIFICATION (TEVV). Available at https://www.nist.gov/ai-test-evaluation-validation-and-verification-tevv.
- NIST. Appendix A: Descriptions of AI Actor Tasks. Available at https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF/Appendices/Appendix_A.

- OECD (2019). Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449. Available at https://legalinstruments.oecd.org/api/print?ids=648&lang=en.
- Stanford University Human-Centered Artificial Intelligence (2023), Generative AI: Perspectives from Stanford HAI. Available at: https://hai.stanford.edu/sites/default/files/2023-03/Generative_AI_HAI_Perspectives.pdf.
- World Economic Forum (2023), The Presidio Recommendations on Responsible Generative AI. Available at: https://www.weforum.org/publications/the-presidio-recommendations-on-responsible-generative-ai/.

網路資料

- Adina Bresge, Training AI on machine-generated text could lead to 'model collapse,' researchers warn, August 21, 2023, https://www.utoronto.ca/news/training-ai-machine-generated-text-could-lead-model-collapse-researchers-warn
- Bartz, D., & Hu, K. (2023, July 22). OpenAI, Google, others pledge to watermark AI content for safety, White House says. Retrieved from Reuters: https://www.reuters.com/technology/openai-google-others-pledge-watermark-ai-content-safety-white-house-2023-07-21/
- Bernhard Lück, Halluzinierende KI – diese drei Maßnahmen helfen (2023), in: https://www.bigdata-insider.de/halluzinierende-ki-diese-drei-massnahmen-helfen-a-d7075ae954a908e831f5ba473385ce69/.
- Burt, A. (2024) 'How to Red Team a Gen AI Model', Harvard Business Review, 4 January. Available at: https://hbr.org/2024/01/how-to-red-team-a-gen-ai-model (Accessed: 13 June 2024).
- ChatGPT-Halluzinationen im Visier des Datenschutzes, BC 2024, 192.
- Devaux, E. (2022) What is differential privacy: Definition, mechanisms, and examples, Statice. Available at: https://www.statice.ai/post/what-is-differential-privacy-definition-mechanisms-examples (Accessed: 13 June 2024).
- Meta, July 18, 2023, Meta and Microsoft Introduce the Next Generation of Llama, https://about.fb.com/news/2023/07/llama-2/.
- Gibney, E. (2024). 4What the EU's tough AI law means for research and ChatGPT.Available at https://www.nature.com/articles/d41586-024-00497-8.
- Fernandez, P., Couairon, G., Jégou, H., Douze, M., & Furon, T. (2023, October 6). Stable Signature: A new method for watermarking images created by open source generative AI. Retrieved from Meta: https://ai.meta.com/blog/stable-signature-watermarking-generative-ai/.

- Franz Hofmann, Retten Schranken Geschäftsmodelle generativer KI-Systeme?, ZUM 2024, 166
- Google-Chatbot Bard kämpft gegen KI-«Halluzinationen», indem es zweifelhafte Textstellen markiert (2023), in: https://www.nzz.ch/technologie/google-chatbot-bard-markiert-zweifelhafte-textstellen-ld.1756929.
- Graeme Whiles, What is AI Model Collapse? Everything You Need to Know, January 14, 2024, https://aicontentexpert.co.uk/2024/01/14/ai-model-collapse/.
- Hopkins, A. et al. (2024) AI supply chains aren't Ai Value Chains. Available at: https://aipolicy.substack.com/p/supply-chains-6 (Accessed: 13 June 2024).
- Jayant Nehra (2024), Effective Data Deduplication for Training Robust Language Models. Available at https://python.plainenglish.io/effective-data-deduplication-for-training-robust-language-models-44467afac5bb.
- Kamara, S. (2024) Transforming work with ai: Human-AI configurations, LinkedIn. Available at: https://www.linkedin.com/pulse/transforming-work-ai-human-ai-configurations-seni-kamara-scfdf (Accessed: 13 June 2024).
- Kandpal, N., Wallace, E., & Raffel, C. (2022, June). Deduplicating training data mitigates privacy risks in language models. In International Conference on Machine Learning (pp. 10697-10707). PMLR
- Mehta, S., Rogers, A., & Gilbert, T. K. (2023). Dynamic Documentation for AI Systems. arXiv preprint arXiv:2303.10854. Available at https://arxiv.org/abs/2303.10854.
- Millidge, B. (2023) Llms confabulate not hallucinate, Beren's Blog - Thoughts on AI, Neuroscience, and other things that interest me. Available at: https://www.beren.io/2023-03-19-LLMs-confabulate-not-hallucinate/ (Accessed: 13 June 2024).;. KI-Halluzinationen: Wie ChatGPT die Grenzen von Künstlicher Intelligenz verschiebt (2024) DEINKIKOMPASS.de. Available at: https://deinkikompass.de/blog/ki-halluzinationen-wie-chatgpt-fakten-erfindet (Accessed: 13 June 2024).
- Nelson, B. (2021). Differential privacy-a balancing act. Chalmers Tekniska Hogskola (Sweden). Available at https://research.chalmers.se/publication/523824
- OpenAI Privacy Request Portal (2023). Available at: https://privacy.openai.com/policies (Accessed: 13 June 2024).
- OpenAI. How ChatGPT and our language models are developed. Available at https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed.
- Odd Erik Gundersen and Sigbjørn Kjensmo, State of the Art: Reproducibility in Artificial Intelligence, file:///C:/Users/chenny/Downloads/11503-Article%20Text-15031-1-2-20201228%20(1).pdf.

- Odd Erik Gundersen, Improving reproducibility of artificial intelligence research to increase trust and productivity https://www.oecd-ilibrary.org/sites/3f57323a-en/index.html?itemId=/content/component/3f57323a-en.
- Potsawee (9 May 2024), potsawee/selfcheckgpt. https://github.com/potsawee/selfcheckgpt.
- Strandell, J. (2024) What is content moderation? (plus best practices), Besedo. Available at: https://besedo.com/blog/what-is-content-moderation/ (Accessed: 13 June 2024).
- Terms of use (2023) OpenAI. Available at: https://openai.com/policies/terms-of-use/ (Accessed: 13 June 2024).
- Technische Entwicklungen im Zusammenhang mit KI-Modellen begegnen einer Vielzahl datenschutzrechtlicher Herausforderungen, MMR 2023, 911
- Technavigator,Was ist Datenintegrität? (2024), https://technavigator.de/wiki/datenintegritaet/
- Yadav, K. and Lai, S. (2024) What does information integrity mean for democracies?, Lawfare. Available at: https://www.lawfaremedia.org/article/what-does-information-integrity-mean-for-democracies (Accessed: 13 June 2024).