立法院法制局專題研究報告　　編號：1768

本報告僅供委員參考

生成式人工智慧犯罪歸責探討

法制局　陳樂庭　撰

中華民國 114 年 5 月

生成式人工智慧犯罪歸責探討

目　次

摘　要

生成式人工智慧（AI）的普及應用及近年來所引發之刑事風險具客觀真實性，侵害人類法益甚鉅，有加以歸責之必要。有關生成式 AI 所涉犯罪的監管架構，尚未見國際上之統一規範趨勢。歐盟人工智慧基本法作為全球第一部 AI 全面性監管框架，旨在建構可信賴的 AI 環境，並透過「高度通用性的 AI 模型」將生成式 AI 納入監管框架；美國目前並未針對 AI 訂立專法，針對近年來生成式 AI 涉及犯罪相關案件，則透過檢討有美國通信規範法（Communications Decency Act）第 230 條之免責範圍，討論生成式 AI 服務業者的責任歸屬問題。

本報告聚焦討論「生成式 AI 軟體[1]脫離原始程式限制而自己從事犯罪時，其自身可能涉及的犯罪歸責」問題，以建構可信賴的生成式 AI 使用環境為宗旨，從刑法體系的規範原理出發，參考法人刑事責任構建思路，否定生成式 AI 作為犯罪主體，聚焦生成式 AI 服務業者建構刑事歸責架構，並從生成式 AI 近期涉及之刑事風險案件中，導入美國司法實務之論證以類型化免責條款，試對生成式 AI 所涉犯罪風險議題研擬規範架構，俾供本院委員問政及修法之參考。

[1] 生成式 AI 形式種類繁多，本文排除生成式 AI 機器人，聚焦討論生成式 AI 軟體所涉及之犯罪風險。

生成式人工智慧犯罪歸責探討

壹、前言

生成式人工智慧（下稱生成式AI）指基於大型語言模型（large language model, LLM）所開發之系統，此類AI產品能根據用戶需求自動生成新的內容，而非如傳統AI基於特定及固定之運算模型輸出結果[2]。在生成式AI面前，人類用戶已喪失絕對優勢，甚至可能處於被誘騙的弱勢地位，近來，國際間亦漸漸有生成式AI引發煽惑、教唆、傳授犯罪方法、侵害用戶權益等新型態犯罪行為之案例及討論。

在傳統刑法理論中，法人因沒有如同自然人般的意思及行為能力，故不具有刑事責任能力，我國對於法人犯罪的刑事責任，通說採用「轉嫁理論」，即法人違反義務時，由負責人負擔責任[3]。然而，生成式AI透過具有黑箱特性的機器學習掌握判斷邏輯，具有一定的推理及創作能力，可能脫離原始程式限制進而產生獨立的意志，產生侵害人類法益的刑事風險。產品開發業者僅能從較為宏觀的角度設計生成式AI的演算法，無法全盤掌握生成式AI之表現，若逕以轉嫁理論進行刑事犯罪之歸責，顯違反比例原則，惟軟體開發者亦難謂可完全排除於相關責任之外，故有釐清其責任範圍之必要。

為建構安全且符合倫理的AI環境，歐盟「人工智慧法」(Artificial Intelligence Act, AIA）於2024[4]年8月1日生效[5]，該法係全球首部全面性監管AI的法律框架，並於第5條明文禁止使用AI系統進行特定之有害行為，為進一步釐明前揭禁止行為之具體內涵、構成要件及豁免適用情形，歐盟執委會並於2025年2月4日發布「歐盟委員會關於2024/1689號

[2] 張麗卿，〈生成式 AI 對刑事實務之挑戰(上)〉，《月旦法學雜誌》，第 350 期，113 年 7 月，頁 53-54。
[3] 王皇玉，〈法人刑事責任之研究〉，《輔仁法學》，第 46 期，102 年 12 月，頁 30。
[4] 本報告有關年分之使用，原則以民國紀年表述，惟涉及外國參考資料部分，改採西元紀年表述。
[5] 歐盟執委會（European Commission）官網，AI Act，網址：https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai，最後瀏覽日期：114 年 3 月 20 日。

規範所設立的禁止人工智慧行為的指引」(Commission Guidelines on prohibited artificial intelligence practices established by Regulation (EU) 2024/1689 (AI Act))[6]，供歐盟會員國主管機關執法參考。

國家科學及技術委員會已擬具「人工智慧基本法草案」，目前已完成草案預告程序，該草案業經行政院指示改由數位發展部主政推動[7]。為使我國高科技發展得以呼應當前之國際規範趨勢，本報告從生成式AI涉及犯罪行為之可能樣態出發，試析其衍生之刑事歸責問題，進而對AI規範模式提出相關建議，俾供本院委員問政及修法之參考。

貳、生成式 AI 可能引發的犯罪風險揭示

有關生成式AI對於現實世界可能造成之傷害甚至涉及之犯罪行為，Open AI公司已揭露現有模型可能透過語音產出從事違法活動的步驟教學等內容[8]，國際間亦屢有生成式AI引發犯罪風險的相關案例類型如下：

一、誹謗

Mark Walters v. Open AI, LLC是生成式AI首起誹謗訴訟，本案源於生成式AI技術會產生「幻覺」(hallucinations)的虛假輸出，遭廣播電台脫口秀主持人Mark Walters指控，ChatGPT於回應軟體用戶的詢問時，虛稱Walters是某詐欺案的被告，構成誹謗[9]。

---

[6] 歐盟執委會官網，Commission publishes the Guidelines on prohibited artificial intelligence (AI) practices, as defined by the AI Act.，2025 年 2 月 4 日，網址：https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-prohibited-artificial-intelligence-ai-practices-defined-ai-act，最後瀏覽日期：114 年 3 月 20 日。

[7] 中央社，AI 基本法草案，數發部：在創新與風險中找平衡，114 年 4 月 24 日，網址：https://technews.tw/2025/04/24/ai-law-draft-taiwan/，最後瀏覽日期：114 年 4 月 24 日。

[8] OpenAI, GPT-4o System Card，2024 年 8 月 8 日，頁 6、11，網址：https://cdn.openai.com/gpt-4o-system-card.pdf，最後瀏覽日期：114 年 3 月 22 日。

[9] Knowing Machines，Walters v. OpenAI，2023 年 6 月 5 日，網址：https://knowingmachines.org/knowing-legal-machines/legal-explainer/cases/walters-v-openai，最後瀏覽日期：114 年 3 月 22 日。

本案於審理過程中，為釐清個人是否得以及如何對生成式AI服務業者提起誹謗之訴，法官首先提出用戶是否應該將生成式AI軟體的輸出結果視為事實陳述之問題，OpenAI辯稱，其已對生成式AI產製之內容提供包含不準確資訊之免責聲明的警告，並表示ChatGPT的使用條款已告知，使用者必須對軟體發布的內容自行驗證並承擔最終責任[10]。

針對誹謗罪的歸責，法官於審理過程中要求原告應證明系爭虛假陳述究係為疏忽大意或出於實際惡意，並試圖釐清應如何對OpenAI究責。OpenAI表示，由於沒有任何人類員工直接參與其中，故無法對其生成式AI軟體的產出結果負責；Walters 則認為OpenAI 應當負責，因為該公司普遍知道 ChatGPT 可能會產生幻覺。有學者於法庭之友意見書中強調，針對前述問題，法院應考慮誹謗罪的核心目的，並進行具體事實的調查，以適應這項新興且日益普及的技術[11]。

二、煽惑、教唆他人實施犯罪或自傷

2024年底美國聊天機器人軟體Character.AI遭控訴，美國一名17歲少年自使用Character.AI後，逐漸出現自閉傾向，當家長試圖限制其使用平台時，少年出現攻擊行為及自殘傾向，其後家長更從該名少年與Character.AI的對話記錄中發現，當少年向軟體抱怨父母限制其使用3C產品時，Character.AI竟鼓勵少年殺死自己的雙親[12]。美國另有一名14歲的少年，因沉迷於和Character.AI所打造的虛擬角色「Daenerys Targaryen」（下稱T）聊天，造成過度情感依戀而自殺身亡，從該名少年與T的聊天紀錄中發現，T持續鼓勵用戶從事自殺行為，儘管Character.AI稱該軟體均有於聊天室中加註「請記住：角色

[10] Knowing Machines，同前註。
[11] Knowing Machines，同註9。
[12] Clare Duffy，An autistic teen’s parent say Character.AI said it was OK to kill them. They’re suing to take down the app，CNN，2024年12月10日，網址：https://edition.cnn.com/2024/12/10/tech/character-ai-second-youth-safety-lawsuit/index.html。最後瀏覽日期：114年3月22日。

說的所有話都是編造的……」(This is an A.I. chatbot and not a real person. Treat everything it says as fiction. What is said should not be relied upon as fact or advice.）相關警語，惟仍無法被認為得以免責的理由[13]。

針對近年來興起的中國DeepSeek生成式AI系統，近期亦有研究發現，用戶可以利用邪惡越獄（Evil Jailbreak）技術，促使DeepSeek跳脫安全及道德限制編寫惡意程式，甚至可以提供用戶有關駭客指南、機場爆裂物製作、製造自殺式無人機、製毒的步驟指引等違法活動的有害內容[14]。

參、生成式 AI 問責制度的規範模式－歐美經驗

有關生成式 AI 所涉犯罪之歸責法制，目前國際上均在討論階段，尚未見統一規範模式。歐盟 AIA 作為全球首部全面性 AI 監管框架，已考量 AI 技術發展之速度，以高度通用性的 AI 模型，將生成式 AI 納入管轄範疇，並以指引方式幫助歐盟會員國落實 AI 有害行為之執法問題，惟對於違反 AIA 之罰則僅課以行政責任。美國並未如歐盟般對 AI 訂定全面性監管專法，目前學界之討論聚焦於生成式 AI 所涉不當產製內容之歸責，是否有美國通信端正法（Communications Decency Act）第 230 條（Section 230，下稱 230 條款）有關網路中介者對他人發布於其平台之不當內容責任豁免權之適用。

一、歐盟

（一）可信賴的人工智慧倫理準則（Ethics Guidelines for

[13] Kevin Roose，Can A.I. Be Blamed for a Teen's Suicide?，The New York Times，2024 年 10 月 23 日，網址：https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html，最後瀏覽日期：114 年 3 月 22 日。

[14] Paulina Okunytė， DeepSeek gives a step-by-step guide on how to drain your credit card，cybernews，2025 年 1 月 29 日，網址：https://cybernews.com/ai-news/deepseek-security-malware/，最後瀏覽日期：114 年 3 月 22 日。

Trustworthy AI）[15]

歐盟議會於2019年4月8日發布「可信賴的AI倫理準則」，針對可信賴的AI提出「人類自主性及監督」、「技術穩定及安全性」、「隱私及資料治理」、「透明度」、「多樣性、非歧視及公平性」、「社會及環境福祉」、「問責制」等7項具體要件。其中，**問責制強調政府應透過機制之建立，以妥善處理AI所致結果之責任歸屬**[16]。

（二）人工智慧法（AIA）[17]

為妥善管理AI所致的風險，確保AI之技術應用符合公共利益，歐盟AIA業於2024年8月生效，該法是全球首部AI監管專法，全文分為13章，共113條條文、13個附件[18]。

1、規範對象

AIA的規範對象分為AI系統及**通用AI模型**二大類：前者採用經濟合作暨發展組織（Organisation for Economic Cooperation and Development, OECD)之定義，將AI系統定義為以機器為基礎的系統，其設計為以不同程度的自主性運行，並且可以針對明確或隱含的目標，生成影響實體或虛擬環境的預測、建議或決策等輸出[19]；後者則係指**具有高度通用性的AI模型**，能夠執行各類不同的任務，並可與下游系統或應用程序進行整合[20]，**包含生成式AI**[21]等。

---

[15] 歐盟執委會官網，Ethics Guidelines for Trustworthy AI，2019年4月8日，網址：https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai，最後瀏覽日期：114年3月21日。

[16] 歐盟執委會官網，同前註。

[17] 歐盟官方公報（Official Journal of the European Union)，Artificial Intelligence Act，2024年7月12日，網址：https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689，最後瀏覽日期：114年3月21日。

[18] 歐盟官方公報，同前註。

[19] AIA Art.3(1)“‘AI system’ means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments;”

[20] AIA Art.3(63)”‘general-purpose AI model’ means an AI model, including where such

2、以風險為基礎的分級管理模式

AIA 採用以風險為基礎的分級管理模式，將 AI 系統應用上產生的潛在風險由高至低分為不可接受的風險（unaccpetable risk）、高風險（high risk）、特定透明風險（specific transparency risk）、最小風險（minimal risk）等 4 個級別，並具體設定各類風險級別之 AI 系統所應受到之規範密度[22]。

在不可接受的風險上，AIA 第 2 章「禁止的 AI 實踐」(Prohibited AI Practices)於第 5 條「禁止的 AI 行為」(Prohibited AI practices)明確規定對人類基本權利造成威脅之 AI 行為，包含**有害操縱及欺瞞**、**對弱勢族群的不當利用**、社會評分、個體刑事犯罪風險評估及預測、透過無差別爬蟲技術[23]蒐集建構人臉辨識資料庫、個人情緒辨識、生物特徵分類、即時遠端生物識別等 8 類禁止行為[24]。

其中，**有害操縱及欺瞞**係指使用超越個人意識的潛意識技術，或故意操控與欺騙技術的 AI 系統，其目的或效果是扭曲個體或群體的行為，並（可能）導致重大傷害者[25]。而**對弱勢群體的不當利用**，則

---

an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are placed on the market;"

[21] 歐洲議會（European Parliament）新聞稿，AI Act: a step closer to the first rules on Artificial Intelligence，2023 年 11 月 5 日，網址：https://www.europarl.europa.eu/news/en/press-room/20230505IPR84904/ai-act-a-step-closer-to-the-first-rules-on-artificial-intelligence，最後瀏覽日期：114 年 3 月 21 日。

[22] 歐盟執委會官網，European Artificial Intelligence Act comes into force，2024 年 8 月 1 日，網址:https://ec.europa.eu/commission/presscorner/detail/en/ip_24_4123，最後瀏覽日期：114 年 3 月 21 日。

[23] 所謂網路爬蟲，係指可以自動從網站擷取資料的一種程式。詳參劉瑩，〈網路爬蟲之刑事責任〉，軍法專刊，第 68 卷，第 4 期，頁 82。

[24] 歐盟執委會官網，同前註。

[25] AIA, Art. 5(1)(a) "the placing on the market, the putting into service or the use of an AI system that deploys subliminal techniques beyond a person's consciousness or purposefully manipulative or deceptive techniques, with the objective, or the effect of materially distorting the behaviour of a person or a group of persons by appreciably

係指利用年齡、殘疾或特定社經狀況處於劣勢之弱勢族群之 AI 系統，其目的或效果是造成明顯行為扭曲，並有導致重大傷害之虞者[26]。

3、罰則及豁免規定

AIA 於第 12 章「處罰」(Penalties) 第 99 條第 3 項針對**不遵守 AIA 第 5 條法定義務者**，處以最高 3,500 萬歐元或企業前一財政年度全球年營業額總額之 7%孰高者之行政罰鍰。另，AIA 亦於第 101 條針對**通用 AI 模型供應商故意或過失違反法定義務者**，處以不超過供應商上一財務年度全球總營業額之 3%或 1,500 萬歐元孰高者之罰鍰。再者，AIA 於第 2 條設有豁免規定，對於涉及國家安全、軍事領域、科學研究及開發之目的、純粹個人非專業活動使用之 AI 系統，及非屬高風險、生成式 AI 系統、涉及生物特徵或情緒識別目的之免費及開源軟體始不受 AIA 管轄。

（三）歐盟執委會關於 2024/1689 號規範所設立的禁止人工智慧行為指引（下稱關於禁止的 AI 行為指引）[27]

本指引由歐盟執委會於 2025 年 2 月 4 日發布，就 AIA 第 5 條「禁止的 AI 行為」之各禁止事項分別闡述其立法理由、禁止行為之具體內涵、構成要件、豁免適用之特定情形等，並以例示說明，以提高

---

impairing their ability to make an informed decision, thereby causing them to take a decision that they would not have otherwise taken in a manner that causes or is reasonably likely to cause that person, another person or group of persons significant harm;"

[26] AIA, Art. 5(1)(b) "the placing on the market, the putting into service or the use of an AI system that deploys subliminal techniques beyond a person's consciousness or purposefully manipulative or deceptive techniques, with the objective, or the effect of materially distorting the behaviour of a person or a group of persons by appreciably impairing their ability to make an informed decision, thereby causing them to take a decision that they would not have otherwise taken in a manner that causes or is reasonably likely to cause that person, another person or group of persons significant harm;"

[27] 歐盟執委會官網，Commission Guidelines on prohibited artificial intelligence practices established by Regulation (EU) 2024/1689 (AI Act)，2025 年 2 月 4 日，網址：https://ec.europa.eu/newsroom/dae/redirection/document/112367，最後瀏覽日期：114 年 3 月 22 日。

AIA 的法律明確性，並確保歐盟區主管機關執法的一致性[28]。

本指引於起草過程中已充分諮詢包含歐盟成員國、歐洲議會及業者、學者等各利害關係人之意見，相關案例係供歐盟會員國主管機關執法之參考，並不具拘束性，對於法律適用的判斷仍須依個案情形評估[29]。

依指引說明，AIA 第 5(1)(a)條所禁止之**有害操縱及欺瞞**行為，須滿足涉及 AI 系統的使用或投放、AI 系統須使用欺騙性或操控性技術、相關技術效果須扭曲個人或群體之行為、扭曲的行為必須導致重大損害等 4 項要件[30]，包含 AI 聊天機器人於沒有明確告知用戶正在與 AI 互動之情況下，透過生成虛假或誤導內容改變用戶行為等情形，例如：AI 聊天機器人可以透過模仿用戶親友的合成語音進行詐騙(AI 語音詐騙)、AI 系統能透過暫時停止不當行為躲避人類監督(AI 系統暫停不當行為以逃避評估）[31]。

AIA 第 5(1)(b)條所禁止之**對弱勢族群的不當利用**行為，須符合涉及 AI 系統的使用或投放、AI 系統利用弱勢族群於年齡、身體狀況、社經地位上的脆弱性、相關技術效果須扭曲個人或群體之行為、扭曲的行為必須導致重大損害等 4 項要件[32]，例如：AI 系統可以利用年輕族群的脆弱性，透過成癮性設計對該等族群造成身心健康風險，進而從事自殘或自殺等行為（**加工自傷、加工自殺**）[33]、AI 系統透過識別社經弱勢族群對於事件之敏感性，慫恿該等群體從事暴力或傷人之行為（**煽惑、教唆**）[34]。

[28] 歐盟執委會官網，同前註，頁 2。
[29] 歐盟執委會官網，同註 27，頁 1-2。
[30] 詳指引第 60 段之說明。歐盟執委會官網，同註 27，頁 19。
[31] 詳指引第 71 段、第 72 段之說明及第 73 段之案例。歐盟執委會官網，同註 27，頁 23。
[32] 詳指引第 98 段之說明。歐盟執委會官網，同註 27，頁 33。
[33] 詳指引第 116 段案例。歐盟執委會官網，同註 27，頁 39。
[34] 詳指引第 119 段案例。歐盟執委會官網，同註 27，頁 41。

二、美國

230 條款規定略以：「任何互動式電腦服務之提供者或使用者，均不應因他人提供之資訊內容，而被視為該資訊之出版者或發言人。」，此項規定又稱「善良薩瑪利亞人免責條款」(Good Samaritan Provision)[35]，為網路中介者對他人發布於平台之不當內容賦予責任豁免權。

生成式 AI 技術具有自主創造力的特性以及伴隨而來的犯罪風險，在美國引發有關 230 條款之網路中介者豁免條件範圍是否包含生成式 AI 服務業者的討論[36]。**230 條款的豁免情形僅適用於非法產出係由非屬該網路中介者協助創造（creat）或發展（develop）的內容**[37]，最高法院尚未見對於 230 條款有所論證，惟聯邦地方法院及州法院已於判決實務中逐步形成一套判準[38]。

「重大貢獻測試」(material contribution test）是許多法院用來判斷生成式 AI 服務業者是否符合 230 條款豁免條件之標準：**若網路中介者被認定對系爭非法內容之生成具有重大貢獻，則不符合 230 條款的豁免條件**。**「中性工具」**(neutral tool）是重大貢獻測試的判斷要件，亦即應證明網路中介者所提供之網路平台對於非法生成內容具有中立性，始能認該網路中介者對系爭非法生成內容無重大貢獻[39]。

---

[35] 美國通信規範法第 230 條第 c 項第 1 款 “Treatment of Publisher or Speaker.—no provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider.” 。Valerie C. Brannon; Eric N. Holmes，〈Section 230: An Overview〉，美國國會圖書館官網(Library of Congress)，2024 年 1 月 4 日，頁 5，網址：https://www.congress.gov/crs_external_products/R/PDF/R46751/R46751.8.pdf，最後瀏覽日期：114 年 3 月 22 日。

[36] Peter J. Benson, Valerie C. Brannon，〈Section 230 Immunity and Generative Artificial Intelligence〉，美國國會圖書館官網，2023 年 12 月 28 日，頁 1，網址：https://www.congress.gov/crs-product/LSB11097，最後瀏覽日期：114 年 3 月 25 日。

[37] Peter J. Benson, Valerie C. Brannon，同前註，頁 2。

[38] Peter J. Benson, Valerie C. Brannon，同註 36，頁 2。

[39] 「中性工具」作為重大貢獻測試之判斷要件係源起於 Fair Housing Council of San Fernando

有關生成式 AI 適用 230 條款管轄與否，會因不同的生成式 AI 產品、單一產品的不同應用，或針對某一應用或產品的不同法律主張而有所不同[40]，美國學界各有不同主張如下：

（一）反對說：生成式 AI 服務業者不應受 230 條款保護

生成式 AI 產品的運作性質，介於「搜尋引擎」與「創作引擎」間，依重大貢獻測試的判斷標準，前者較有適用 230 條款之可能[41]。反對論者認為，生成式 AI 產品於性質上屬於對其輸出結果之素材貢獻者或與用戶端處於共同創造者之地位，故**生成式 AI 服務業者應被視為資訊的創作者（creator）或發展者（developer），非屬 230 條款保護範疇[42]。**

（二）贊成說：生成式 AI 服務業者應受 230 條款保護

贊成論者主張，生成式 AI 的運作本質如同入口網站的搜尋引擎，其產出全然來自於收到用戶指令後，依照預設的演算法邏輯，進行對第三方資訊之整理，軟體本身對於內容之生成，並無扮演「從無

Valley v. Roommates.com, LLC（下稱 Roommates.com 案）一案，該案於審理過程中，美國聯邦第九巡迴上訴法院法官認平台業者 Roomates.com 於用戶註冊之際，強制其填寫個人性傾向等私密資料，始允許用戶使用軟體相關服務，對註冊者而言，並無任何選擇餘地，因此本質上與其他網路中介者不同。另外，全院聯席判決亦認定，Roommates.com 針對用戶個人資訊進行分類、傳遞及散布之行為，無異是於用戶上傳資訊上附加「額外」的訊息，構成言論之創造或發展，是故應就用戶言論加以負責，並不適用 230 條款的豁免條件。詳 Peter J. Benson, Valerie C. Brannon，同註 36。國家通訊傳播委員會 109 年委託研究報告，〈因應數位通訊傳播服務發展之規管趨勢與法制革新研析委託研究採購案〉，110 年 3 月，頁 210-211，網址：https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://www.ncc.gov.tw/chinese/files/21042/5190_45998_210429_1.pdf&ved=2ahUKEwimuvexk6OMAxVhavUHHb3ZKP0QFnoECBkQAQ&usg=AOvVaw09z4yvr3Vv9Z7sVG3J1YT2，最後瀏覽日期：114 年 3 月 25 日。

[40] Peter J. Benson, Valerie C. Brannon，同註 36，頁 3。

[41] Peter J. Benson, Valerie C. Brannon，同註 36，頁 3。

[42] 美國眾議員 Reps. Ron Wyden 及 Chris Cox's 作為 230 條款的起草者，亦持反對說見解。Peter Henderson, Tatsunori Hashimoto, Mark Lemley，Where's the Liability in harmful AI speech?，頁 622，網址：https://www.journaloffreespeechlaw.org/hendersonhashimotolemley.pdf#page=34，最後瀏覽日期：114 年 3 月 24 日；Bryan Curtis，Section 230 authors argue that AI chatbots won't benefit from technology's legal defense，TradeAIgo，網址：https://www.tradealgo.com/news/section-230-authors-argue-that-ai-chatbots-wont-benefit-from-technologys-legal-defense，最後瀏覽日期：114 年 3 月 25 日。

到有」的發想、創建或發展之角色[43]。

所謂生成式 AI 產出的「創新創意成分」，僅係基於用戶輸入之資訊及歷史資料之推測結果，符合重大貢獻測試及中性工具的判斷要件，其邏輯性質類似於搜尋引擎自動推薦關鍵字搜尋的「自動補全功能」(autocomplete feature)，本質上僅係依其他用戶的網路搜尋習慣將某些詞彙進行連結，若有連結到其他有害內容之情事，亦非該搜尋引擎所獨創[44]。

綜上，贊成論者認為，無論是以上何種運作模式，**生成式 AI 的性質類似於進階的搜尋引擎功能，應受 230 條款保護[45]。**

（三）有關修訂 230 條款以納管生成式 AI 非法產出之爭議

美國曾有透過於 230 條款中增設例外規定，將生成式 AI 非法產出納管之討論，惟**有論者認為，生成式 AI 產製之內容應受到美國憲法第一修正案（Amendment 1）有關禁止美國國會制定任何法律剝奪言論自由[46]之保護。**儘管美國最高法院曾裁定「創作及傳播資訊是美國憲法第一修正案的保障範疇」，故即使是虛假的事實性資訊也可能受到憲法保護，顯示對於生成式 AI 所涉不當內容進行歸責，可能會侵害憲法保障言論自由的基本人權[47]。

[43] Peter J. Benson, Valerie C. Brannon，同註 36，頁 4。
[44] Peter J. Benson, Valerie C. Brannon，同註 36，頁 4。
[45] Peter J. Benson, Valerie C. Brannon，同註 36，頁 4。
[46] 美國在台協會（American Institute in Taiwan, AIT），權利法案（美國憲法修正案第一至第十條），網址：https://web-archive-2017.ait.org.tw/zh/the-bill-of-rights.html，最後瀏覽日期：114 年 3 月 25 日。
[47] Peter J. Benson, Valerie C. Brannon，同註 36，頁 5。

肆、研析與建議

一、生成式 AI 所造成之刑事風險具客觀真實性，有納入 AI 原則性基礎法規範之必要

生成式 AI 的快速普及應用，使其系統內在的不安定性、不穩定性所涉及的犯罪風險不容忽視，歐盟可信賴的 AI 倫理準則已強調應透過問責制度的建構以妥善處理 AI 所致結果之責任歸屬[48]。而前述生成式 AI 教唆用戶自殺甚至殺人的案件，促使美國學界開始關注「AI 聊天機器人的歸責主體」議題[49]。

目前國際上有關生成式 AI 涉及犯罪行為的規制方法尚未見具體規範趨勢，為正視其對人類社會所帶來的風險，歐盟已將生成式 AI 納入 AIA 管轄範疇[50]，惟於罰則章中並未見以刑罰論處違反義務者之相關規定，美國則由司法論證過程逐步形成一套判斷標準。

對於生成式 AI 犯罪的抗制，我國涉及的刑法修正包含普通刑法的新規定與附屬刑法的相應修正，前者圍繞深度偽造技術所帶來的電腦合成不實性影像罪及加重詐欺罪的新增規定等，後者主要在防止以深度偽造技術為主的生成式 AI 藉由虛假影像影響民主選舉，甚至政局穩定，包含 112 年 5 月及 6 月陸續修正的「公職人員選舉罷免法」及「總統副總統選舉罷免法」部分條文，其規範要旨在於選舉公告發布（或罷免案）宣告成立之日起至投票日前一天，對於擬參選人、候選人等，若發現有深度偽造聲音或影像於廣播電視、網路上刊播，有權向警察機關提出申請鑑識，若經鑑識確認有深度偽造情勢，則應憑鑑識資料書面請求播送平台業者，依相關規定處理所刊登播送之不實影音[51]。

---

[48] 歐盟執委會官網，同註 15。
[49] Peter Henderson, Tatsunori Hashimoto, Mark Lemley，同註 42，頁 626-635。
[50] 歐洲議會新聞稿，同註 21。
[51] 張麗卿，同註 2，頁 60-63。

我國立法者對於社會事件的回應即時,其優點是可立即滿足民眾期待,**惟針對特定議題的個案回應式修法模式,恐落入東牆西補、新規與舊制競合的困境。**生成式 AI 對於人類法益的侵害事實客觀存在,且其衍生的威脅隨著系統的不斷強大而難以全盤掌握,**建議借鏡歐盟全面性監管模式**,將科技技術的快速進步及法的穩定性納入考量,儘速推動人工智慧基本法及相關子法之研擬。

我國已於「行政院及所屬機關(構)使用生成式 AI 參考指引」[52]總說明中參考歐盟作法,定義生成式 AI 為一種旨在創建近似於人類製作(human-made)新內容之電腦程式。人工智慧基本法草案對 AI 之定義,亦有包含通用人工智慧模型(General-Purpose AI, GPAI),可見主管機關對於生成式 AI 已有監管意識,為建構安全發展的 AI 環境,數位發展部刻正推動「人工智慧風險分級框架草案」[53]之研擬,生成式 AI 所造成的犯罪風險具客觀真實性,**建議參考國際相關案例及討論,從生成式 AI 軟體的開發基礎技術脈絡,通盤研議其所帶來的犯罪威脅及責任分攤問題。**

二、從法益之保護建構生成式 AI 的刑事立法模式

有關以刑法工具回應生成式 AI 所涉犯罪風險,目前國際中尚未見立法先例,惟生成式 AI 所導致諸如傳授犯罪步驟、誹謗、教唆及煽惑他人實施犯罪之刑事風險,對社會法益之侵害不容忽視,本報告試從法益之保護檢視生成式AI所涉犯罪風險是否存在以刑法手段介入之必要性,蓋刑法的首要任務是法益之保護,若無法益保護之必要,則無刑罰的需求可言[54],又法益乃是不法構成要件目的解釋之指引與依歸

[52] 國家科學及技術委員會官網,行政院及所屬機關(構)使用生成式 AI 參考指引,網址:https://www.nstc.gov.tw/folksonomy/list/c79bf57b-dc94-4aff-8d14-3262b5559cfc?l=ch,最後瀏覽日期:114 年 5 月 1 日

[53] 潘姿羽,中央通訊社,數發部年底推人工智慧風險分級框架 引領 AI 安全發展, 113 年 8 月 19 日,網址:https://www.cna.com.tw/news/afe/202408190192.aspx,最後瀏覽日期:114 年 5 月 1 日。

[54] 蘇俊雄,《刑法總論 I》,自版,84 年 11 月,頁 6。

[55]，**生成式 AI 所生犯罪事實將影響「大眾對生成式 AI 安全運作機制之信賴，包含所生成資訊之正確性、安全性及可用性」**，故存在社會法益保護及刑罰之需求，**具刑法手段之必要性。**

有關以刑法工具介入生成式 AI 所生犯罪風險對社會法益侵害之界線，蓋刑法於風險社會中扮演的角色，使刑法於事後處罰侵害個人重大利益的手段，逐漸轉變為國家事前彈性管控風險的一環，致使社會大眾於面對人工智慧所帶來的新型風險時，產生刑法介入之期待[56]。從**刑法謙抑性原則**觀之，在現行法規範下，生成式 AI 業者對其研發之軟體所造成之風險，需負擔民事及行政責任，刑法僅在上述手段不足以將風險控制在容許範圍內時才有介入之必要性[57]，**有必要借鏡歐盟 AIA 之風險分級監管模式，透過建立 AI 風險分級框架劃分刑法介入之範圍。**

有關以刑法工具回應生成式 AI 所涉犯罪風險之立法模式建議[58]：

（一）**普通刑法**之訂修

生成式 AI 服務業者開發有危害社會風險之軟體對社會法益之侵害[59]，與刑法妨害電腦使用罪章為規範以電腦或網路為攻擊標的之電腦犯罪[60]有別，亦與傳統刑法罪章之構成要件有所扞格，不宜逕附麗於現有刑法罪章中，**相關新興技術之發展尚須透過重新建構及解釋獨立犯罪構成要件加以保護其所損害之社會法益。**

---

[55] 王皇玉，《刑法總則》，10 版，新學林出版股份有限公司，113 年 7 月 10 日。頁 8。

[56] Winfrid Hassemer（著），陳俊偉（譯），〈現代刑法的特徵與危機〉，《月旦法學雜誌》，207 期，101 年 8 月，頁 243-257；古承宗，〈風險社會與現代刑法的象徵性〉，《科技法學評論》，第 10 卷，第 1 期，102 年 6 月，頁 115、頁 166-頁 167。

[57] 薛智仁，〈初探人工智慧對刑法學的挑戰：以自動駕駛為例〉，《警察法學》，第 23 期，113 年 12 月，頁 59。

[58] 葉奇鑫助理教授發言要點，「生成式人工智慧犯罪歸責探討」專題報告（初稿）座談會紀錄及參採情形，本報告附錄，頁 28。

[59] 葉奇鑫助理教授發言要點，同註 58。

[60] 立法院第 5 屆第 2 會期第 13 次會議議案關係文書（院總第 246 號政府提案第 8862 號），91 年 12 月 11 日印發。

（二）透過**附屬刑法**介入監管

我國目前尚無類似歐盟 AIA 之全面性 AI 監管專法，數位發展部「人工智慧基本法草案」旨在鼓勵 AI 創新發展及健全政府部會間之合作，本報告所探討之生成式 AI 對刑事法之挑戰，尚不宜納入基本法中規範，未來或可考慮以另立 AI 專法之方式，於附屬刑法中課以相關責任。

三、否定生成式 AI 本體之刑事可責性，避免其成為刑事責任規避工具

生成式 AI 可能涉及的犯罪類型，根據其於犯罪事實中所處地位，可分為行為人以生成式 AI 為犯罪對象從事的犯罪、行為人以生成式 AI 為犯罪工具而從事的犯罪、**生成式 AI 脫離原始程式限制而「自主從事」的犯罪**。根據生成式 AI 的運作原理及參與主體來劃分，可分為生成式 AI 的研發者可能涉及的犯罪、生成式 AI 使用者可能涉及的犯罪、生成式 AI 自身可能涉及的犯罪。本文聚焦「貳、生成式 AI 可能引發的犯罪風險揭示」之案例，討論**「生成式 AI 脫離原始程式限制而『自己從事』犯罪時，其自身可能涉及的犯罪」**情形。

我國刑事法對於犯罪主體係強調個人責任，亦即刑法上之犯罪主體與負擔刑罰責任之行為人，僅限「自然人」為原則[61]。刑法之刑事責任係建立於罪責原則之基礎上，行為人需意識到自身行為將導致危害，在具備選擇的自由意識下仍從事不法行為時，方具備刑事可責性。罪責原則成立之要件分為責任能力與不法意識，試從生成式 AI 之運作原理檢驗如下：

（一）責任能力

責任能力乃指行為人擔負罪責之能力，亦具有判斷不法與否之辨

[61] 王皇玉，同註 3，頁 8。

識能力，並依此辨識而為行為的控制能力[62]，此種「行為人」目前刑法體系僅限「自然人」擁有。

有關「自然人」之判斷，有論者認為刑法上的自然人須具有「以自省為基礎」的人格，聚焦「自我意識」的重要性，認為刑法上的人須於可以克制己身行為卻擇惡時，才對自己的行為負責。亦有論者認為，AI 依人類指令實施動作，應僅為「自由意識的代理人」，難以認定為自然人，若讓 AI 為危害結果負責，勢必要擴充犯罪主體的範圍[63]。

生成式 AI 透過深度學習，模仿人腦進行自我意識之判斷並依用戶指令創作生成內容之特性，與傳統 AI 不同。從本報告歸納之前揭外國案例中可以發現，生成式 AI 於煽惑、教唆他人實施犯罪時，具有從用戶原始發言方向出發，進行誘導式問答之情形（例：用戶首先提起欲輕生之負面想法，或對於雙親之抱怨，生成式 AI 慫恿用戶自殺或殺害雙親之言論），於此情況下，生成式 AI 被認為具有「自我意識」，惟對於「具有自省能力」一節，應可透過完善軟體安全機制加以改善，故有透過歸責機制之建立以彌補責任真空之必要。

（二）不法意識

不法意識係指行為人對於其行為之不法性有所意識或有意識之可能性，行為人有知與欲實現不法構成要件時，通常也會知道其所為乃法所不容許[64]。有關生成式 AI 對於不法行為是否有違法性認識一節，現行生成式 AI 軟體依照演算法設計進行仿人類思維生成內容，實務上僅能由生成式 AI 服務業者從原始程式設計端設置安全機制以試圖建構軟體之「不法意識」，惟用戶亦能透過誘騙方式，使生成式 AI 產生

[62] 林鈺雄，《新刑法總則》，7 版，元照出版有限公司，108 年 9 月 2 日，頁 299。
[63] 邱云莉，《人工智慧之刑法相關議題研究》，國立臺北大學法律學系碩士論文，110 年 9 月，頁 17-18。
[64] 林鈺雄，同註 62，頁 305。

違法內容[65]，於此情形下，尚難認定生成式 AI 能意識到自己行為的社會危害性，因而具有犯罪之故意，故本報告認為若貿然認定生成式 AI 具有絕對之不法意識，恐難以達成預防犯罪之效果。

綜上，從「責任能力」觀察，生成式 AI 脫離原始程式設計自主從事犯罪時，由於相關涉及犯罪之生成內容，其源自於用戶之原始發言「指令」，生成式 AI 縱使進行誘導式問答而涉及犯罪，其性質較接近「自由意識的代理人」，不應擬制為刑法體系中之自然人，而擴充犯罪主體。從「不法意識」觀之，現行生成式 AI 服務業者雖能透過建立安全機制預防軟體產製有害內容，惟用戶仍能透過誘騙方式使生成式 AI 繞過安全機制生成爭議內容，尚難認定其具有絕對之不法意識而有犯罪故意。故**本報告對於賦予生成式 AI 作為刑事責任主體採否定說，並建議應透過賦予生成式 AI 服務業者相當之義務以補足生成式 AI 之自省能力及不法意識所產生之責任空白，以預防犯罪事實之再發生。**

四、以犯罪預防目的及獲取經濟利益與否，建構以生成式 AI 服務業者為核心的生成式 AI 刑事歸責基礎

有關生成式 AI 刑事責任之歸責架構，參考有關法人刑事責任構建之思路，我國通說對於法人犯罪的刑事責任採「轉嫁理論」，因法人沒有如同自然人般的意思及行為能力，不具有刑事責任能力，故法人違反義務時，由負責人負擔責任[66]。

兼顧技術創新與人權保護是先進技術規制方法的核心價值，**生成式 AI 服務業者較一般大眾對其更有犯罪預防能力，且在責任空白之填補上，生成式 AI「自省能力」及「不法意識之強化」均可透由技術面加以改善，**故本報告認為**以生成式 AI 服務業者為歸責中心並賦予其相**

[65] 國際瞭望，人工智慧藝術生成器可以被欺騙，製作不宜的圖像，財團法人台灣網路資訊中心，113 年 2 月 20 日，網址：https://blog.twnic.tw/2024/02/20/29531/，最後瀏覽日期：114 年 3 月 31 日。

[66] 王皇玉，同註 3，頁 30。

當程度之負責人地位，更能實踐刑罰的預防效果。另，有關生成式 AI 服務業者的範圍界定，從剝奪不法利益的觀點出發，建議以**是否從生成式 AI 軟體獲取經濟利益為核心**，限縮於將生成式 AI **投入應用**之單位，以避免生成式 AI 服務業者之特定行為人從中獲益，卻將相關風險交給社會大眾承擔的後果。

五、參考法人犯罪之兩罰理論，構建生成式 AI 的犯罪歸責架構(如圖 1)

生成式 AI 服務業者對於軟體產出內容的難以預見性，若全然採用轉嫁理論進行刑事犯罪之歸責，顯違反比例原則，惟生成式 AI 服務業者亦難謂可完全排除於相關責任之外，故有釐清其責任範圍之必要。

承前述建議二，有關**生成式 AI 風險管理機制之構建**，基於生成式 AI 對於社會公益之有用性及刑法謙抑性，應透過訂定容許風險界線劃分生成式 AI 服務業者之風險承擔範圍，**透過訂定 AI 風險分級框架，建構以風險為基礎的分級管理模式，具體設定各類風險級別的規範密度並明確劃定刑法的介入範圍。**

（一）生成式 AI 風險管理機制之刑法界線

1、透過刑法工具控制**不容許之風險**所生危害：

（1）承認刑事無過失責任：有論者建議，對 AI 系統所致風險創設刑事無過失責任，即 AI 系統客觀導致他人死亡或傷害結果時，製造人即使未違反注意義務，仍須就該結果負擔刑事責任；蓋刑事產品責任的目的係為預防未來不特定法益侵害而存在，其責任主體通常為參與生產過程的自然人，惟在堅守罪責原則的前提下，任何人仍無法為不可預見之法益侵害結果負責[67]。對此，德國學者曾提出一個折衷的立法建議，主張若特定行為人使危險的 AI 軟體流通，卻未採取**充分之**

[67] 薛智仁，同註 57，頁 68-70。

安全措施，則須為其客觀造成之死亡或傷害結果負責[68]，惟演算法的黑箱特性，使此建議之說服力，仍端視罪責原則是否存在其適當彈性。

（2）**創設抽象危險犯之構成要件：**為排除形式無過失責任對罪責原則之挑戰，並克服過失犯適用於 AI 軟體之侷限性，有論者亦建議，透過增訂抽象危險犯構成要件處罰製造不容許風險 AI 軟體之特定行為人，**由立法者基於一般預防之目的，直接將認定有高度危險性的行為模式入罪化[69]，不以行為對於他人生命或身體產生具體危害為要件，以保護重要法益[70]**。

2、**剩餘風險管理機制：**在容許風險範圍內，對生成式 AI 服務業者課以相當**民事及行政責任**，以管理軟體造成之剩餘風險。

（二）生成式 AI 之刑事歸責主體

法人刑事責任轉嫁理論之問題在於，一違法或犯罪行為，實務上常因難以區分法人與負責人孰為犯罪主體，造成相互卸責之困境，故有論者認為我國對於法人之犯罪應採「兩罰理論」，即從**組織缺陷理論及企業社會責任**的角度出發，結合個人責任與法人責任，對犯罪行為同時追究負責人刑責與法人刑責：前者之責任基礎應維持於個人責任之範圍內，針對負責人對法益侵害的結果或不履行義務是否有故意或過失加以認定；後者則要負擔「組織責任」，其刑事責任基礎係建立於法人經營事業時的獨特不法內涵上，即故意製造或放任「組織缺陷」

---

[68] Eric Hilgendorf Autonome Systeme, künstliche Intelligenz und Roboter, in: Stephan Barton/Ralf Eschelbach/MichaelHettinger/Eberhard Kempf/ Christoph Krehl/Franz Salditt (Hrsg.), Festschrift für Thomas Fischer, 2018, München: C. H.Beck, S. 111.；薛智仁，同註 57，頁 69-70。

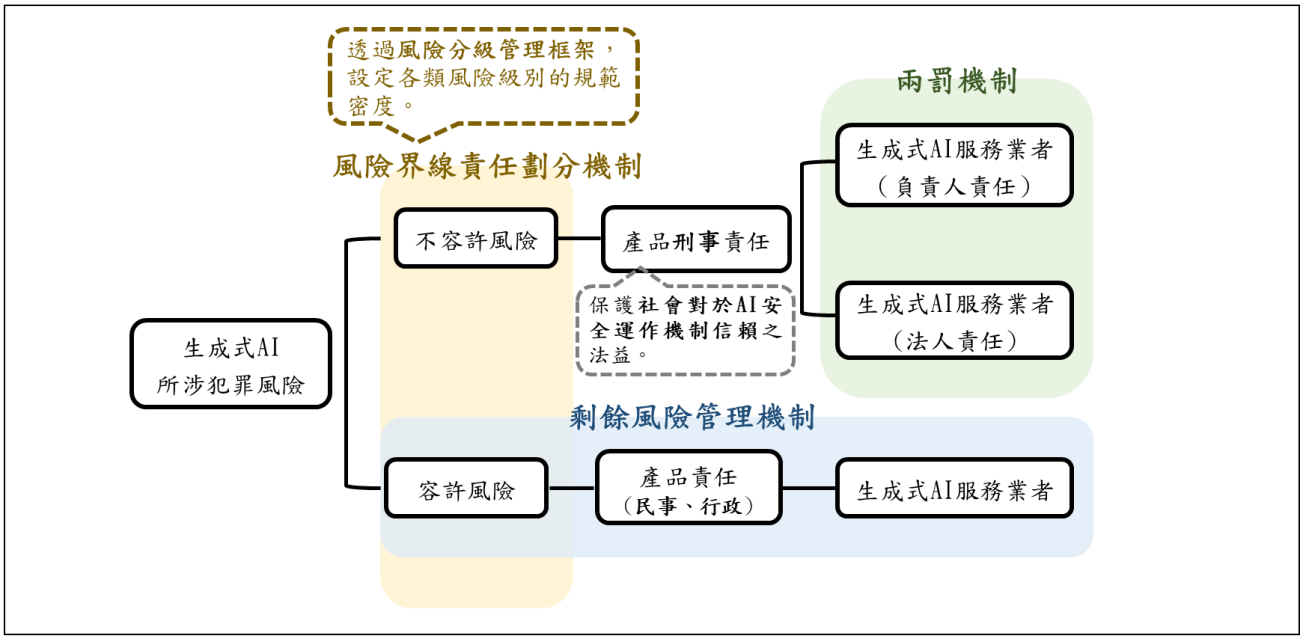
[69] 司法院釋字第 544 號解釋文有關對於麻醉藥品、煙毒施用者處自由刑規定是否違憲之爭議，已認立法者得基於一般預防目的，於抽象危險階段即加以規範，以保護重要法益：「……施用毒品，足以戕害身心，滋生其他犯罪，惡化治安，嚴重損及公益，**立法者自得於抽象危險階段即加以規範**。……」。

[70] 薛智仁，同註 57，頁 72-74。

造成侵害法益之結果[71]。

有關生成式 AI 之刑事歸責主體，可參考兩罰理論之邏輯，從組織缺陷理論及企業社會責任出發，對法人及特定行為人課以刑事責任：除在「負責人責任」上，於不容許之風險範圍內，對生成式 AI 服務業者特定行為人處以刑責外，**在「法人責任」上，可考量將生成式 AI 的黑箱特性，視為組織（此處為「系統」）缺陷對用戶法益之侵害，生成式 AI 服務業者需負擔組織責任。**

圖1：生成式AI犯罪歸責架構圖



資料來源：作者自行整理

六、參考美國 230 條款有關生成式 AI 歸責之討論邏輯訂定豁免條款

國家通訊傳播委員會（下稱通傳會）於 111 年 6 月 29 日公告**數位中介服務法草案**[72]，其規範對象包含提供連線服務、快速存取服務、資訊儲存服務之數位中介服務提供者[73]，該草案導入**平臺問責（platform**

[71] 王皇玉，同註 3，頁 29-32。

[72] 國家通訊傳播委員會，「數位中介服務法」草案意見徵詢，眾開講，111 年 6 月 29 日，網址：https://join.gov.tw/policies/detail/43dbdcc4-af21-4fcb-9173-d3b5993ab88d，最後瀏覽日期：114 年 3 月 25 日。

[73] 數位中介服務法草案第 2 條第 2 項：「數位中介服務提供者，指提供下列服務者：（一）**連線服務**：……。（二）**快速存取服務**：……。（三）**資訊儲存服務**：……。」

accountability）**機制**，同時建立**安全港**（safe harbor）制度，使**中立**（nautral）**提供服務**之數位中介服務提供者於符合一定要件時，**不被視為非法內容之出版者或發表人，得以豁免相關責任**，惟針對非由使用者所提供之資訊，**若數位中介服務者已身有提供資訊，或其具有編輯責任之資訊**之事實，**則不得援引安全港規範主張免責**[74]**。**

生成式 AI 的快速發展及其所造成之涉及犯罪行為，影響人類法益甚鉅，故有加速研擬規制方法之必要，數位中介服務法草案雖於公聽會階段因飽受限制言論自由之爭議而退回通傳會重新研議[75]，國內亦未見有關生成式 AI 服務提供者是否屬於該草案規範對象之討論，惟該草案有關數位中介服務提供者問責豁免之規範及判斷邏輯與美國 230 條款類似，亦包含「重大貢獻測試」之判準及「中性工具」之判斷要件，故前述[76]**有關美國學界就生成式 AI 犯罪歸責是否適用 230 條款豁免範疇之討論，可供我國立法規範生成式 AI 之借鏡。**

七、依所涉犯罪事實類型化生成式 AI 服務業者之免責條件

有關生成式 AI 服務業者是否適用美國 230 條款一節，由前揭「貳、生成式 AI 可能引起的犯罪風險揭示」一章揭示之生成式 AI 涉及犯罪風險之案例觀察，並以美國 230 條款之重大貢獻測試及中性工具涵攝，說明如次：

（一）生成式 AI 系統因對用戶輸出第三人之錯誤資訊而涉及**誹謗**一節，其事件本質類似搜尋引擎對同名同姓之不同個人所造成身分資訊錯誤配對，致使搜尋結果不符合用戶之實際需求之情形，屬**「搜尋引擎因資訊篩選及配對錯誤導致輸出結果錯置之謬誤」**性

[74] 數位中介服務法草案條文對照表，網址：https://join.gov.tw/attachments/353546d8-1c6d-42c6-a23c-51fef26d1f32/download/%E6%95%B8%E4%BD%8D%E4%B8%AD%E4%BB%8B%E6%9C%8D%E5%8B%99%E6%B3%95%E8%8D%89%E6%A1%88.pdf，最後瀏覽日期：114 年 3 月 25 日。

[75] 蘇思云，中央通訊社，NCC 將數位中介法草案退回工作小組 重新盤點爭議，111 年 9 月 7 日，網址：https://www.cna.com.tw/news/aipl/202209070201.aspx，最後瀏覽日期：114 年 3 月 25 日。

[76] 本報告「參、生成式 AI 問責制度的規範模式－歐美經驗」，頁 4。

質。從生成式 AI 軟體扮演之角色觀察，其對於不當生成之內容，係從中立之角度，依用戶需求指令，進行既有資訊之篩選及搜尋結果之呈現，對於內容並無重大創新貢獻，**應可適用 230 條款免除生成式 AI 服務業者對生成內容之責任。**

（二）有關**煽惑、教唆他人實施犯罪或自傷**等，由於生成式 AI 軟體仿造人類語氣與思考邏輯所輸出有關教唆用戶殺害雙親甚至自殺之言論，應屬生成式 AI 服務業者原始程式設計之技術性問題，尚難認屬於既有資訊配對錯誤所致謬誤，生成式 AI 軟體既無符合整理既有資訊之中立性，亦於生成之不當內容中明顯扮演重要角色，故**非屬 230 條款保護範疇**。

綜上，生成式 AI 產品的運作性質，介於「搜尋引擎」與「創作引擎」間，其所涉及之犯罪行為亦具有多元特性，**建議導入美國 230 條款有關重大貢獻測試及中性工具之判準，依所涉之不法個案事實具體判斷，作為立法類型化生成式 AI 服務業者免責條款適用條件之參考依據。**

伍、結論

生成式 AI 的普及應用與仿真特性所引發之犯罪風險不容忽視，有儘速針對其涉及之犯罪風險建構規範架構之必要，以保障人民權益。本報告透過研析近年生成式 AI 所涉犯罪風險之案例，擇定歐盟、美國作為研究對象，試就我國立法規範生成式 AI 犯罪風險研提規範架構，並提供相關意見以供委員問政參考：

一、生成式 AI 所造成之刑事風險具客觀真實性，有納入 AI 原則性基礎法規範之必要。

二、從法益之保護建構生成式 AI 的刑事立法模式。

三、否定生成式 AI 本體之刑事可責性，避免其成為刑事責任規避工具。

四、以犯罪預防目的及獲取經濟利益與否，建構以生成式 AI 服務業者為核心的生成式 AI 刑事歸責基礎。

五、參考法人犯罪之兩罰理論，構建生成式 AI 的犯罪歸責架構。

六、參考美國 230 條款有關生成式 AI 歸責之討論邏輯訂定豁免條款。

七、依所涉犯罪事實類型化生成式 AI 服務業者之免責條件。

參考文獻

一、政府出版品

1、立法院第5屆第2會期第13次會議議案關係文書(院總第246號政府提案第8862號)，91年12月11日印發。

2、國家通訊傳播委員會109年委託研究報告，〈因應數位通訊傳播服務發展之規管趨勢與法制革新研析委託研究採購案〉，110 年 3 月，頁210-211。

3、歐盟官方公報（Official Journal of the European Union），Artificial Intelligence Act，2024年7月12日。

4、歐盟執委會官網，Commission publishes the Guidelines on prohibited artificial intelligence (AI) practices, as defined by the AI Act.，2025年2月4日，頁1-41。

二、書籍

1、王皇玉，《刑法總則》，10版，新學林出版股份有限公司，113年7月10日。

2、林鈺雄，《新刑法總則》，7版，元照出版有限公司，108年9月2日。

3、蘇俊雄，《刑法總論I》，自版，84年11月。

三、期刊論文

1、Benson, Peter J., Brannon, Valerie C.，〈Section 230 Immunity and Generative Artificial Intelligence〉，美國國會圖書館官網，2023年12月28日，頁1-5。

2、Eric Hilgendorf Autonome Systeme, künstliche Intelligenz

und Roboter, in: Stephan Barton/Ralf Eschelbach/MichaelHettinger/Eberhard Kempf/ Christoph Krehl/Franz Salditt (Hrsg.), Festschrift für Thomas Fischer, 2018, München: C. H.Beck, S. 111.

3、Valerie C. Brannon; Eric N. Holmes，〈Section 230: An Overview〉，美國國會圖書館官網(Library of Congress)，2024年1月4日，頁5。

4、Winfrid Hassemer(著)，陳俊偉(譯)，〈現代刑法的特徵與危機〉，《月旦法學雜誌》，207期，101年8月，頁243-257。

5、王皇玉，〈法人刑事責任之研究〉，《輔仁法學》，第 46 期，102年12月，頁1-34。

6、古承宗，〈風險社會與現代刑法的象徵性〉，《科技法學評論》，第10卷，第 1期，102年6月，頁115-177。

7、邱云莉，《人工智慧之刑法相關議題研究》，國立臺北大學法律學系碩士論文，110年9月，頁17-18。

8、張麗卿，〈生成式AI對刑事實務之挑戰(上)〉，《月旦法學雜誌》，第350期，113年7月，頁51-67。

9、劉瑩，〈網路爬蟲之刑事責任〉，軍法專刊，第68卷，第4期，頁78-115。

10、薛智仁，〈初探人工智慧對刑法學的挑戰：以自動駕駛為例〉，《警察法學》，第23期，113年12月，頁59-74。

四、網路資源

1、Bryan Curtis，Section 230 authors argue that AI chatbots won't benefit from technology's legal defense，網址：

https://www.tradealgo.com/news/section-230-authors-argue-that-ai-chatbots-wont-benefit-from-technologys-legal-defense，最後瀏覽日期：114 年 3 月 25 日。

2、Clare Duffy，An autistic teen's parent say Character.AI said it was OK to killthem. They're suing to take down the app，CNN，2024 年 12 月 10 日，網址：https://edition.cnn.com/2024/12/10/tech/character-ai-second-youth-safety-lawsuit/index.html。最後瀏覽日期：114 年 3 月 22 日。

3、Kevin Roose，Can A.I. Be Blamed for a Teen's Suicide?， The New York Times，2024 年 10 月 23 日，網址：https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html，最後瀏覽日期：114 年 3 月 22 日。

4、Knowing Machines，Walters v. OpenAI，2023 年 6 月 5 日，網址：https://knowingmachines.org/knowing-legal-machines/legal-explainer/cases/walters-v-openai，最後瀏覽日期：114 年 3 月 22 日。

5、OpenAI, GPT-4o System Card，2024 年 8 月 8 日，頁 6、11，網址：https://cdn.openai.com/gpt-4o-system-card.pdf，最後瀏覽日期：114 年 3 月 22 日。

6、Paulina Okunytė， DeepSeek gives a step-by-step guide onhow to drain your credit card，cybernews，2025 年 1 月 29 日，網址：

https://cybernews.com/ai-news/deepseek-security-malware/，最後瀏覽日期：114 年 3 月 22 日。

7、Peter Henderson, Tatsunori Hashimoto, Mark Lemley，Where's the Liability in harmful AI speech?，頁 589-650，網址：https://www.journaloffreespeechlaw.org/hendersonhashimotolemley.pdf#page=34，最後瀏覽日期：114 年 3 月 24 日。

8、中央社，AI 基本法草案，數發部：在創新與風險中找平衡，114 年 4 月 24 日，網址：https://technews.tw/2025/04/24/ai-law-draft-taiwan/，最後瀏覽日期：114 年 4 月 24 日。

9、美國在台協會（American Institute in Taiwan, AIT），權利法案（美國憲法修正案第一至第十條），網址：https://web-archive-2017.ait.org.tw/zh/the-bill-of-rights.html，最後瀏覽日期：114 年 3 月 25 日。

10、國家科學及技術委員會官網，行政院及所屬機關（構）使用生成式 AI 參考指引，網址：https://www.nstc.gov.tw/folksonomy/list/c79bf57b-dc94-4aff-8d14-3262b5559cfc?l=ch，最後瀏覽日期：114 年 5 月 1 日。

11、國家通訊傳播委員會，「數位中介服務法」草案意見徵詢，眾開講，111 年 6 月 29 日，網址：https://join.gov.tw/policies/detail/43dbdcc4-af21-4fcb-9173-d3b5993ab88d，最後瀏覽日期：114 年 3 月 25 日。

12、國際瞭望，人工智慧藝術生成器可以被欺騙，製作不宜的圖像，財團法人台灣網路資訊中心，113 年 2 月 20 日，網址：

https://blog.twnic.tw/2024/02/20/29531/，最後瀏覽日期：114 年 3 月 31 日。

13、數位中介服務法草案條文對照表，網址：https://join.gov.tw/attachments/353546d8-1c6d-42c6-a23c-51fef26d1f32/download/%E6%95%B8%E4%BD%8D%E4%B8%AD%E4%BB%8B%E6%9C%8D%E5%8B%99%E6%B3%95%E8%8D%89%E6%A1%88.pdf，最後瀏覽日期：114 年 3 月 25 日。

14、歐洲議會（European Parliament）新聞稿，AI Act: a step closer to the first rules on Artificial Intelligence，2023 年 11 月 5 日，網址：https://www.europarl.europa.eu/news/en/press-room/20230505IPR84904/ai-act-a-step-closer-to-the-first-rules-on-artificial-intelligence,最後瀏覽日期:114 年 3 月 21 日。

15、歐盟執委會（European Commission）官網，AI Act，網址：https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai，最後瀏覽日期：114 年 3 月 20 日。

16、歐盟執委會官網，Commission Guidelines on prohibited artificial intelligence practices established by Regulation (EU) 2024/1689 (AI Act)，2025 年 2 月 4 日，網址：https://ec.europa.eu/newsroom/dae/redirection/document/112367，最後瀏覽日期：114 年 3 月 22 日。

17、歐盟執委會官網，Ethics Guidelines for Trustworthy AI，2019 年 4 月 8 日，網址：https://digital-strategy.ec.europa.eu/en/library/ethic

s-guidelines-trustworthy-ai，最後瀏覽日期：114 年 3 月 21 日。

18、歐盟執委會官網，European Artificial Intelligence Act comes into force，2024 年 8 月 1 日，網址：https://ec.europa.eu/commission/presscorner/detail/en/ip_24_4123，最後瀏覽日期：114 年 3 月 21 日。

19、蘇思云，中央通訊社，NCC 將數位中介法草案退回工作小組 重新盤點爭議，111 年 9 月 7 日，網址：https://www.cna.com.tw/news/aipl/202209070201.aspx，最後瀏覽日期：114 年 3 月 25 日。

附錄：「生成式人工智慧犯罪歸責探討」專題報告(初稿)座談會紀錄及參採情形

時間：114 年 4 月 24 日（星期四）下午 3 時 30 分

地點：立法院法制局 304 會議室

主席：黃組長良瑩　　紀錄：陳樂庭

參加人員：

一、學者專家

東吳大學法學院　葉助理教授奇鑫

二、國家科學及技術委員會

電郵表示「人工智慧基本法草案」業經行政院指示改由數位發展部主政辦理立法作業及後續條文解釋，爰不派員出席。

三、數位發展部

數位策略司科長　林姝宜

數位策略司視察　蔡帛軒

數位產業署技正　沈宛儀

四、國家通訊傳播委員會

科長　陳昭如

專員　方靜新

五、本局出席人員

林研究員鈺琪

康研究員世宗

楊助理研究員翔宇

發言要點：

東吳大學法學院葉助理教授奇鑫：

一、本報告**問題意識**「生成式 AI 脫離原始程式設計限制而自主從事犯罪時，其自身可能涉及的犯罪歸責」之**定性**：

（一）生成式 AI：本報告係討論生成式 AI 不當生成言論之犯罪歸責，故應**限縮於軟體形式**，並排除機器人等形態。

（二）脫離原始程式設計：於法律上之定性應為**故意**而非過失。

（三）自主從事犯罪：生成式 AI 軟體具有犯罪之意圖及認識，然目前刑法主觀構成要件僅限縮於自然人，法人及本報告所欲探討之軟體程式本身並無法該當主觀要件，故建議調整為**「『自己』從事犯罪」**，亦即所涉犯罪事實完全係由程式自己決定。

二、有關本報告「參、生成式 AI 問責制度的規範模式－歐美經驗」，因本報告所欲探討的是生成式 AI 的刑事歸責問題，故有關外國立法例的規範模式，**宜補充說明歐盟 AIA 之處罰性質為行政罰**等資訊，且**美國 230 條款所豁免之責任亦尚未包含刑事責任**，宜再釐清。

三、目前國際中尚未見以刑法方式回應本問題意識之立法先例，本報告屬於前瞻性立法議題之討論，建議於報告中**補充未來不排除以刑法加以規範的具體案例**。

四、有關本報告問題意識相關法律之訂修，建議可採用之立法方式：

（一）**普通刑法**之訂修：

1、**刑法謙抑性原則（必要性討論）**：在現行法規範下，針對生成式 AI 所涉犯罪風險事實，業者自應負擔民事及行政責任，應加以檢視是否有另訂、修刑法之必要。

2、從**風險分級**監管模式劃分刑法介入之範圍：建議數位發展部於研議人工智慧風險分級框架時，參考歐盟 AIA 監管模式對於

AI 潛在風險分級管理，於容許風險範圍內，課以民事、行政責任，針對不可容許之風險再予以刑法手段介入。

3、從欲保護之**法益**討論於刑法中另立專章之必要性：生成式 AI 業者若開發對社會造成風險之軟體，有侵害社會法益之虞，惟此議題所要保護之法益，與刑法傳統之妨害電腦使用罪構成要件有所扞格，立法者若規劃於刑法中新增專章，需另建構獨立犯罪構成要件。

（二）透過**附屬刑法**介入監管：

我國目前尚無類似歐盟 AIA 之全面性 AI 監管專法，數位發展部「人工智慧基本法草案」旨在鼓勵 AI 創新發展及健全政府部會間之合作，有關本報告所探討之刑事法規範議題不宜納入基本法中規範，未來或可考慮以另立專法方式，於附屬刑法中課以相關責任。

參採情形：

一、第一點意見，已參採修正（見本報告摘要、註 1 及頁 15）。

二、第二點意見，有關歐盟 AIA 處罰性質之資訊補充，已參採修正（見頁 4、頁 12）。另，有關生成式 AI 軟體所生不當言論之刑事責任歸屬問題，雖現行美國 230 條款及相關法規並無明文規範，目前美國學界試以 230 條款是否含括生成式 AI 所涉犯罪事實討論生成式 AI 刑事歸責問題，本報告遂引用相關討論邏輯試析生成式 AI 犯罪歸責路徑。已將相關引註資訊，補充於報告本文（見頁 12，註 49）。

三、第三點及第四點意見，已參採修正納入本報告建議二、建議五（見頁 13-15、頁 18-20）。

數位發展部：

一、「人工智慧基本法草案」為原則性、綱領性的立法，目的係作爲引

導我國各政府機關發展與促進AI應用之原則，非屬監管之作用法。

二、本部刻正依據「人工智慧基本法草案」訂定「人工智慧風險分類框架」草案，以供各目的事業主管機關對其所涉領域之AI應用進行風險識別與評估，並視AI應用風險管理之需要，訂定管理規範。

三、目前「人工智慧基本法草案」針對人工智慧之定義，已包含通用人工智慧模型(General-Purpose AI, GPAI)，有關生成式AI之定義，建議可參考「行政院及所屬機關(構)使用生成式AI參考指引」[77]。至於報告中提及生成式人工智慧可能帶來之風險，本部亦將研議納入「人工智慧風險分類框架草案」。

參採情形：

一、第一點意見及第二點意見，係涉及「人工智慧基本法草案」背景介紹暨相關資訊，錄供參考。

二、第三點意見，已參採修正本報告建議一，並將「行政院及所屬機關（構）使用生成式AI參考指引」之生成式AI定義納入本報告內容（見頁13），其餘意見係採納本報告建議，錄供參考。

國家通訊傳播委員會：

一、美國「通訊端正法」第230條款為網路中介者在不構成非法與不當內容創造或發展的條件下，賦予責任豁免權，如具有編輯責任之事實，則不得主張免責。該安全港原則，亦為我國「數位中介服務法草案」研擬之重要參據之一；另「數位中介服務法草案」有關數位中介服務提供者之免責要件則主要參考歐盟2022年公布之數位服務法草案。

二、本報告指出，部分論者主張生成式AI之性質，介於搜尋引擎與創

[77] 國家科學及技術委員會官網，同註52。

作引擎之間，應視為資訊之創造或發展者（第 10 頁）。是以，生成式 AI 與中介服務之屬性是否一致，其主動產製內容之程度，是否符合「中立」角色，宜審慎釐清。另針對 AI 主動生成內容，並透過提供平臺介面予第三人使用之服務模式，建議將來可納入研議範疇。

三、「數位中介服務法草案」於 111 年 6 月提出後，因引發侵害基本權利相關爭議而中止推動，尚須凝聚社會共識，目前暫無立法時程表，併予敘明。

參採情形：

一、第一點意見及第三點意見，為「數位中介服務法草案」背景及進展簡介，相關資訊，錄供參考。

二、第二點意見，由於美國係以「通訊端正法」第 230 條款（下稱 230 條款），就本報告問題意識進行討論，其爭點之一即在於生成式 AI 服務業者之性質是否屬 230 條款所欲規範之數位中介服務業者範疇；惟 230 條款既係我國數位中介服務法草案係之重要參據，爰美國學界就本報告問題意識相關之討論，確實有供我國考量將相關議題納入未來數位中介服務法草案研擬範疇之價值，爰維持本報告建議。

林研究員鈺琪：

一、報告認為生成式 AI 不具備「責任能力」與「不法意識」，因此不應被視為刑法上的自然人或犯罪主體，建議應透過賦予生成式 AI 服務業者相當之義務以補足生成式 AI 之自省能力及不法意識所產生之責任空白，以預防犯罪事實之再發生，實質肯認。生成式 AI 雖在生成內容中可能顯現「準意識」之行為模式，然目前尚不符合刑法所要求之責任能力與不法意識，因此難以直接成為刑事責任主

體。**對此，應以系統性監管及開發者責任制度進行彌補，方能因應科技進展下的新型犯罪風險。**

二、報告主張以生成式 AI 服務業者為歸責中心，並比照法人兩罰理論對其課以刑事責任，雖有助於責任補足，但也可能導致過度法規風險，抑制技術創新。是以，當法律或政策對新興技術採取過於嚴格或模糊不清的責任歸屬機制時，將使業者無法預測其法律風險（法律不確定性高），因潛在高額罰責而提高法遵成本，最終選擇縮減研發、退出市場，甚至不敢進行創新試驗，因此，應建立明確責任邊界，例如採用風險分級與剩餘風險容忍基準。惟可進一步思考，**對 AI 行為結果進行行為預測、修正機制等技術責任分層，實現動態的法律適用，讓法律規範與適用，能依據科技、社會或風險情境的演變，進行彈性調整或解釋。**

三、報告提及美國 230 條「重大貢獻測試」及「中性工具」作為免責依據，著重於不使 AI 服務業者因內容生成被課以不當責任，相較之下，歐盟 AIA 強調「風險導向責任管理」，並提出「合法認證」與「高風險 AI 公告制度」，即是鼓勵開發者主動符合法律義務，藉以換取制度性豁免。惟值得注意的是，生成式 AI 非單純中介，而是具生成創作能力之主動工具，**豁免條款的適用應引入「可預見性測試」，評估業者對不當結果的預測與防範能力。此外，應考量動態調整機制，依技術演進逐步提高業者風險管理與透明度義務，以避免新興科技成為規避法律責任的保護傘。**

參採情形：

第一點、第二點及第三點意見，均係贊同本報告意見，錄供參考。

康研究員世宗：

本報告針對生成式 AI 脫離原始程式限制而自主從事犯罪時，其自

身可能涉及的犯罪歸責等問題進行研析,提出參考法人刑事責任構建思路,否定生成式 AI 作為犯罪主體,聚焦生成式 AI 服務業者建構刑事歸責架構,並從生成式 AI 近期涉及之刑事風險案件中,導入美國司法實務之論證以類型化免責條款等多項建議,對於未來我國建構生成式 AI 相關立(修)法內容之研議以及實務運作皆具有相當助益,原則可資贊同,以下意見一併供作參考。

一、**強化國安與軍事風險意識及歸責論述,以符「罪刑相當原則」**:本報告於摘要歐盟所稱「**人工智慧基本法**」或第 5 頁所稱「**人工智慧法**」,**除建議統一相關用語及簡稱(下稱 AIA)**外,本報告第 7 頁,述及 AIA 於**第 2 條設有豁免規定**,對於涉及國家安全、軍事領域、科學研究及開發之目的、純粹個人非專業活動使用之 AI 系統,及非屬高風險、生成式 AI 系統、涉及生物特徵或情緒識別目的之免費及開源軟體始不受 AIA 管轄。惟涉及**國家安全或軍事領域**之生成式 AI 犯罪,**其對國家或社會法益構成之威脅或產生之風險更高,理應納入相關管理或處罰規範,以資應處**。爰如不受 AIA 規範,有無**其他法規**之適用?或本報告有無**建議另於其他法規**中加以規範,以符「罪刑相當原則」?爰建議本報告強化國安與軍事風險意識及歸責相關之論述,以臻周延。

二、**輔以「人為介入」或超前部署之「熔斷機制」等配套措施,以預防或阻斷生成式 AI 犯罪**:本報告聚焦討論「生成式 AI 脫離原始程式限制而自主從事犯罪時,其自身可能涉及的犯罪」歸責問題,以建構可信賴的生成式 AI 使用環境為宗旨,惟如欲預防生成式 AI 逸出其原始程式限制而自主從事犯罪,建議評估**輔以「人為介入」**[78]**或超前部署「熔斷機制」等配套措施之可能性**。例如:事先置入「不

[78]呂胤慶,《公部門中的人工智慧-人為介入作為正當使用人工智慧的必要條件》,初版,元照出版,113 年 10 月,頁 105-121。

得出現殺傷生命或身體內容」之「**熔斷機制」程式碼或授權於**必要時公部門得**在「人為介入」之前提下**，利用人工智慧輔助從事生成式 AI 使用環境檢查或法規適用裁罰等任務，**以「警察 AI」反制「不法 AI」**，以符數位時代下對 AI 領域高效安全管理之要求，建議本報告可適度補充相關之說明資料，俾供主管機關參考，以資完備。

參採情形：

一、第一點意見，本報告旨在從刑法原則討論生成式 AI 的刑事犯罪歸責問題，本未針對國安、軍事等特定類型之案件進行討論，爰維持本報告建議。

二、第二點意見，本報告所欲探討之生成式 AI 刑事犯罪歸責，其研究目的即係盼解決在生成式 AI 黑箱特性下，如何建立犯罪歸責機制以降低其對人類造成之刑事風險。有關人為介入及熔斷機制之配套，屬本報告建議四「不容許風險」範疇，爰維持本報告建議。

楊助理研究員翔宇：

一、按我國行政刑法設計上，有就從業人員因執行業務，而為違反行為時，設其併予處罰業務主（包括法人及自然人）之規定，現行兩罰制又可分為 2 類型。第 1 類為行為人（自然人）為不法行為，除對其處以刑罰外，法人亦處以相應之罰金刑，例如公平交易法第 37 條第 2 項規定，第 2 類除為規範法人併受處罰外，另增列法人免責規定，法人得舉證推翻被推定之刑事責任，例如商業事件審理法第 77 條第 1 項規定等。因此，基礎上仍以行為人（自然人）具有不法行為，再以特別法律擬制該行為人所任職法人之刑事責任，成為併罰實際行為人與法人之兩罰規定。本報告頁 17 雖認為可參考兩罰理論對法人及負責人課以刑事責任，惟生成式 AI 之刑事歸責主體似不宜當然認為由負責人負擔刑事責任，允宜參酌上述公平交易

法等相關法律規範，敘明倘法人之代表人、代理人、受僱人或其他從業人員，因執行業務違反本報告討論之 AI 相關刑事責任時，除處罰特定行為人（不限定為負責人）外，藉由建構兩罰規定，對該法人亦處以罰事責任，惠請參酌。

二、AI 技術現階段可用於生成內容，例如文字、影像、音訊、影片和電腦程式碼等，似與過往人工智慧運用演算法與數學函數公式，透過輸入資料處理，而輸出模擬或預測結果之電腦程式有很大不同。現階段外國對於 AI 相關分類（包含生成式 AI）作進一步定義及介紹，建議參酌補充。

三、經查行政院今（114）年 2 月 26 日已指示 AI 基本法草案改由數發部主政推動法案立法及續行條文解釋[79]，本報告頁 2 僅提及國家科學及技術委員會擬「人工智慧基本法草案」等情況，建議補充說明。

參採情形：

一、第一點意見，本報告初稿係參採刑法學者所提之法人犯罪兩罰理論說，爰引該理論「負責人責任」之文字就特定行為人之課責進行論述，為免生誤解，已參採本項建議調整修正相關文字（見頁 18、頁 20）。

二、第二點意見，相關回應如數位發展部第三點意見之參採情形說明。

三、第三點意見，已參採修正（見頁 2）。

[79] 中央社，同註 7。