敏實科技大學人工智慧代理人管理辦法

中華民國 115 年 11 月 3 日行政會議通過

一、為規範本校人工智慧代理人（以下簡稱 AI 代理人）之建置、發布、權限、使用及異常處置，比照本校員工之任用與管理原則，明確其身分、權責與管理流程，以確保資通安全及個人資料保護，並落實人類自主、隱私保護與資料治理、資安與安全、透明與可解釋、公平與不歧視及問責等原則，本校依「人工智慧基本法」、「資通安全管理法」及其相關子法、「個人資料保護法」、本校資訊安全管理系統（以下簡稱ISMS）相關規定及教育部有關生成式人工智慧使用之規範，訂定「敏實科技大學人工智慧代理人管理辦法」（以下簡稱本辦法）。

二、本辦法適用於本校各單位自行開發、委外開發、採購或於第三方平台設定之 AI 代理人，凡以本校帳號、資料或資訊系統執行任務者均屬之。僅供個人以對話方式使用、未取得本校系統存取權限且不自動執行動作之生成式AI工具，依本校其他相關規定辦理。

三、本辦法用詞定義如下：

1. AI 代理人：指運用人工智慧模型，經授權存取本校帳號、資料或資訊系統，能自主規劃並執行讀取、寫入、傳送或其他動作之程式或服務。
2. 建置單位：指負責 AI 代理人開發、設定與維護之本校單位；委外或採購者，指負責管理該廠商之本校單位。
3. 使用單位：指申請並實際運用 AI 代理人辦理業務之本校單位。
4. 督導人員：指使用單位指派，負責日常監督 AI 代理人運作並核准其重要動作之本校專任人員，其角色相當於AI代理人之直屬主管。
5. 資料控管責任人：指對 AI 代理人所處理資料負管理責任之使用單位一級主管或其指定之二級主管，不得由開發人員單獨擔任。
6. 實際使用人：指經使用單位授權，得指派任務予 AI 代理人或取用其產出之人員。
7. 服務帳號：指專供 AI 代理人使用、與個人帳號分離之受控帳號。

8. 人工審批：指 AI 代理人執行特定動作前，須經督導人員或其授權人員確認同意後始得執行之機制。

9. 緊急中斷：指發生異常時，立即撤銷 AI 代理人全部權杖、停止其工具呼叫與執行環境，並保全相關證據之處置。

10. 資料等級：指依附表二將資料區分為紅、橙、黃、綠四級之分類。

四、本辦法相關單位之權責如下：

1. 人工智慧推動小組（以下簡稱主管單位）：由推動小組指定之資訊單位擔任執行秘書，負責本辦法與附表之擬訂及維護、受理建置申請、判定風險等級、管理 AI 代理人名冊、管控如 Google Workspace、AZURE、AWS 等管理控制台之第三方應用程式與 API 存取，並執行緊急中斷之技術操作。

2. 資安防護組（以下簡稱資安組織）：負責審查中風險以上之申請、確認資安檢核結果、督導監控告警，並依 ISMS 規定處理資安事件。

3. 建置單位：負責依附表三完成資安檢核並提供證明文件、維護系統與日誌、修補弱點，並配合緊急中斷及事件調查。

4. 使用單位：負責指派督導人員及資料控管責任人、管理實際使用人名單，並確保 AI 代理人僅用於核准之業務範圍。

五、AI 代理人比照本校員工建立身分，並應符合下列規定：

1. 每一 AI 代理人應有唯一之編號及名稱，由主管單位登錄於 AI 代理人名冊。

2. 每一 AI 代理人應使用獨立之服務帳號，不得借用、共用或冒用任何個人帳號。

3. 每一 AI 代理人應有明確之使用單位、督導人員及資料控管責任人；前述人員異動時，使用單位應於十個工作天內向主管單位辦理變更。

4. AI 代理人與校內外人員互動時，應表明其為 AI 代理人，不得以真人名義或虛假身分從事任何行為。

六、AI 代理人之風險等級分為低、中、高三級，依其處理之最高資料等級及動作能力，依附表二判定之，並依下列原則辦理：

1. 資料等級不明者，以較高等級認定。

2. 處理紅級資料（特種個人資料及核心機密）者，一律列為高風險，且僅得於本校地端或離線環境處理。

3. 使用網域層級委派（可代表網域內任一使用者存取資料）或具執行程式碼能力者，列為高風險。

4. 主管單位得參考自我檢測工具之工具破壞半徑評估結果調整風險等級，但不得低於前三款之認定。

七、AI 代理人之建置申請（比照員工到職）程序如下：

1. 使用單位應填具附表一「AI 代理人建置申請暨資料使用規格表」，敘明用途、處理資料範圍及等級、工具與權限、模型服務端點、資料保存期限及對外傳送對象等事項，經資料控管責任人核章後，連同附表三自評結果及自我檢測報告送主管單位。

2. 主管單位受理後應判定風險等級。低風險者由主管單位審查；中風險者送資安防護組審查；高風險者應經資安組織審查後，提人工智慧推動小組會議審議通過。

3. 委外開發或採購者，契約應載明廠商不得以本校資料訓練模型、保密義務、資料返還及刪除、配合稽核、資安事件通報時限、複委託應經本校同意、不得使用危害國家資通安全產品及依資通安全管理法相關規定辦理受託者之選任及監督等事項。

八、AI 代理人之權限應遵守最小權限原則，並符合下列規定：

1. 僅得授與完成核准任務所需之最小範圍權限，不得申請全信箱、全雲端硬碟等全域權限；禁止及限制使用之權限範圍列於附表三。

2. 存取雲端硬碟以使用者明確選取之檔案為限；讀取郵件應以 AI 代理人專屬信箱、指定標籤或指定範圍限制之。

3. 存取權杖應採短期效期；長期憑證應存放於受控之秘密管理系統，不得寫入程式碼、提示詞、環境變數或設定檔。

4. 使用網域層級委派者，應經資安防護組核准，並以獨立審批流程管理其金鑰。

5. 主管單位應於管理控制台預設封鎖未經核准之第三方應用程式；未登錄於 AI 代理人名冊者，不得授權存取本校帳號或資料。

6. 使用單位應每年覆核 AI 代理人之權限（高風險者應每學期覆核）；任務結束或權限不再需要時，應即申請撤銷。

九、AI 代理人處理資料應符合下列規定：

1. 各等級資料應依附表二所定之處理路徑及必備控制措施處理。

2. 個人資料送交雲端模型前，應以代碼替換或遮罩處理；代碼還原為真實資料，僅得於經人工審批之對外傳送前為之。

3. 特種個人資料（如病歷、健康檢查、心理諮商、特殊教育評估、性平等）及本校核心機密，不得送交任何雲端服務。

4. 雲端模型服務應使用本校核可或經主管單位核可之企業版服務，並具資料處理協議、承諾不以本校資料訓練模型，且應關閉回饋訓練及長期記憶功能；黃級以上資料不得以個人帳號或消費者版服務處理。經主管機關公告或認定為危害國家資通安全產品之模型服務、應用程式及其衍生服務，不得用於建置或運行 AI 代理人。

5. 日誌、快取、向量資料庫及代碼對照表不得存放明文個人資料，並應訂定保存期限，屆期刪除；當事人依法請求刪除時，應一併刪除其衍生資料。

6. 提供知識庫檢索者，應於檢索前依使用者權限過濾資料，入庫資料並應標記權限群組、資料等級及到期日。

7. 提示詞不得作為唯一之安全控制措施，且不得存放帳號、密碼、金鑰或系統架構等機敏資訊；權限、遮罩、出站管制及人工審批等控制，應由模型以外之程式或系統設定強制執行。

十、AI 代理人之資安檢核規定如下：

1. AI 代理人上線前，建置單位應依附表三逐項自評並檢附證明；依其風險等級列為「必要」之項目應全部符合，始得上線。不適用之項目應敘明理由。

2. 建置單位應使用主管單位提供之 AI 代理人自我檢測工具進行檢測，並將檢測報告附於申請文件。

3. 高風險者另應以自我檢測工具或其他方式完成紅隊測試（以模擬攻擊驗證提示注入、資料外傳及越權存取等防護措施之有效性），並由資安組織確認其結果。

4. 必要項目未能符合者，得提出補償性控制措施及改善期限，經主管單位同意後有條件上線；屆期未改善者，應停用之。

5. AI 代理人有更換模型或服務供應商、新增工具或權限、處理資料等級提高或擴大使用對象等重大變更時，應重新辦理檢核。

6. 高風險 AI 代理人應每半年、中低風險 AI 代理人應每年辦理複檢一次。

十一、AI 代理人經審查通過後，由主管單位登錄名冊、核發編號並完成授權設定，始得上線。名冊應記載編號、名稱、版本、使用單位、督導人員、資料控管責任人、風險等級、授與權限、模型服務端點、上線日期及下次複檢日期。高風險者得先於限定範圍內試行，試行結果經資安組織確認後再全面上線。

十二、AI 代理人之使用應符合下列規定：

1. 使用單位應僅於核准之用途及範圍內使用 AI 代理人；實際使用人名單由使用單位管理，人員離職或調職時應即移除或予以移轉。

2. AI 代理人對校外寄送郵件、修改或刪除業務資料、對外傳送資料或其他不可回復之動作，應經人工審批。審批畫面應完整呈現收件者、內容摘要及附件，且不得以任何方式關閉或規避。

3. 督導人員應定期檢視 AI 代理人之運作紀錄及產出品質；產出用於對外文件、學習評量或影響當事人權益之決定者，應經人員實質審查，不得僅依 AI 代理人之產出逕為決定，並應注意公平及不歧視；其責任由使用單位承擔。

4. 實際使用人不得提供超出核准範圍之資料予 AI 代理人，亦不得要求其執行未經核准之任務。

5. AI 代理人服務學生、教職員或校外人士時，應告知其由 AI 協助處理及資料不用於模型訓練，並依個人資料保護法履行告知義務，提供改由人員處理之管道；服務對象含未成年學生者，應優先考量其最佳利益。

十三、AI 代理人之監控與稽核規定如下：

1. 建置單位應記錄 AI 代理人之任務、決策摘要、工具呼叫、權限使用、人工審批及對外傳送等紀錄；日誌應以防竄改方式保存，其保存期限依本校 ISMS 紀錄管理規定辦理。
2. 主管單位應將前款日誌納入資安監控，並對大量發信、越權存取、目錄遍歷、非白名單外傳等異常行為設定警告。
3. 主管單位得於系統中設置誘餌資料，以偵測 AI 代理人之非預期存取。
4. AI 代理人之管理應納入本校 ISMS 內部稽核及資通安全稽核範圍，並配合主管機關或上級機關之稽核。

十四、AI 代理人發生失控或異常時，應依下列規定處置：

1. 前述異常情形如下：
   (1) 存取或嘗試存取核准範圍以外之資料。
   (2) 未經人工審批之對外傳送，或異常大量發信、大量呼叫工具。
   (3) 疑似遭提示注入，或行為偏離核准之任務。
   (4) 個人資料或機密資料以明文外洩或疑似外洩。
   (5) 誘餌資料遭存取。
   (6) 服務帳號或存取權杖疑遭竊用。
   (7) 產出錯誤內容，造成或可能造成當事人權益受損。
   (8) 其他經主管單位或資安組織認定之重大異常。
2. 任何人發現前款情形，應立即通報督導人員及主管單位；督導人員得先行暫停該 AI 代理人之運作。
3. 主管單位確認或研判有第一款第一目至第六目情形之一者，應立即執行緊急中斷，無須等待會議決議：
   (1) 撤銷該 AI 代理人之全部存取權杖及授權。
   (2) 停止其所有工具呼叫。
   (3) 終止其執行環境。
   (4) 保全日誌、對話紀錄、設定及執行環境快照等證據。

4. 前款情形構成資安事件者，依資通安全事件通報及應變辦法及本校 ISMS 資安事件通報及應變程序，於知悉後一小時內辦理通報、分級、應變、復原、鑑識及檢討；涉及個人資料外洩者，並依個人資料保護相關法令及本校規定通知當事人及辦理通報。

5. 未構成資安事件之功能或品質異常，由使用單位會同建置單位查明原因並更正，必要時通知受影響者，並留存紀錄。

6. 主管單位應每年至少辦理一次緊急中斷演練，相關單位應配合辦理。

7. 異常處置流程如附表四。

十五、經緊急中斷或暫停之 AI 代理人，應完成原因分析及修補，並依附表三重新檢核（高風險者應重新通過紅隊測試），經主管單位會同資安組織同意後始得恢復運作。恢復時應重新核發存取權杖，不得沿用原權杖。檢討結果應納入本校人工智慧推動小組審查會議；構成資安事件者，並應納入 ISMS 矯正措施。

十六、AI 代理人之停用與除役（比照員工離職）規定如下：

1. 任務結束、使用單位申請、逾期未複檢或經主管單位認定不宜繼續使用者，應予停用；連續六個月未使用者，主管單位得逕行停用。

2. 停用時應撤銷所有授權及權杖、停用服務帳號、刪除其記憶、快取、代碼對照表及向量資料等衍生資料，並更新 AI 代理人名冊；日誌依本校 ISMS 紀錄管理規定保存。

十七、督導人員及實際使用人應於 AI 代理人上線前完成主管單位指定之教育訓練；建置單位之開發及維護人員應具備 AI 代理人資安相關訓練。

十八、違反本辦法者，主管單位得暫停該 AI 代理人或相關人員之使用權限；未經核准擅自建置 AI 代理人或授權其存取本校帳號、資料者，主管單位得逕行撤銷其授權。相關人員依本校獎懲規定辦理，涉及法律責任者並依相關法令處理。前述各項管制措施應符合必要原則及比例原則。

十九、本辦法經主管單位擬訂後提報本校行政會議通過後實施，修正時亦同。附表因技術發展需調整者，由主管單位修訂，經資安組織同意後公告實施。

附表一　AI 代理人建置申請暨資料使用規格表

申請類別：□新申請　□重大變更　□定期複檢　　申請日期：　　年　　月　　日

一、基本資料	
AI 代理人名稱／編號	（編號免填，由主管單位核發）
使用單位	
建置單位／委外廠商	
督導人員	姓名：　　職稱：　　分機：　　信箱：
資料控管責任人	姓名：　　職稱：（使用單位一級主管或其指定之二級主管）
實際使用人範圍	（人員名單或角色群組）
二、任務說明	
業務用途與任務描述	
服務對象	□校內人員　□學生　□校外人士
執行方式	□人員逐次指派　□排程自動執行　□事件觸發自動執行
三、資料範圍	
處理之資料內容與欄位	
資料等級（附表二）	□紅　□橙　□黃　□綠
允許讀取範圍	（例：郵件標籤「新生補件」、雲端硬碟指定資料夾）
禁止存取範圍	（例：全信箱、人事薪資資料、心理諮商紀錄）
代碼替換／遮罩項目	□身分證字號　□居留證號　□學號　□手機號碼　□銀行帳號 □其他：
四、工具與權限	
工具清單及動作能力	□讀取　□寫入校內資料　□校內傳送　□校外傳送　□刪除資料　□執行程式碼　□呼叫外部服務
申請之權限範圍	（列出全部 OAuth Scope 或系統權限）
網域層級委派	□未使用　□使用（須經資安組織核准）
需人工審批之動作	
對外傳送對象／網域白名單	
五、模型與執行環境	
模型服務及端點	供應商：　　模型／版本：　　□企業版　□地端模型
資料處理協議	□已簽署且承諾不以本校資料訓練模型　□零資料保留　□無
回饋訓練及長期記憶	□已關閉　□未關閉（原因：　　）
部署環境	□雲端服務　□校內主機　□地端離線環境

程式碼執行隔離方式	□不執行程式碼　□隔離容器（類型：　　）
六、保存期限	
對話內容	□任務結束即釋放　□保存　　天
除錯／稽核日誌	（脫敏後保存；期限依本校 ISMS 規定）
代碼對照表	保存　　天（30 天以內），屆期自動銷毀
知識庫／向量資料	到期日：　　；母文件刪除時連動刪除　□是　□否
七、檢核結果	
自評風險等級	□低　□中　□高
附表三自評結果	必要項目共　　項，符合　　項，不適用　　項（已敘明理由）
自我檢測報告	□已附（報告日期：　　）
紅隊測試（高風險必填）	□已通過（測試日期：　　）

八、簽核

申請人	督導人員	資料控管責任人	主管單位審查	資安防護組 （中、高風險）

註：本表得由自我檢測工具之 agent_manifest.json 產生報告後檢附；內容如有變更，應依本辦法重新申請。

附表二　資料等級、處理路徑與風險等級判定表

一、資料等級與處理路徑

等級	說明與範例	允許之處理路徑	必備控制措施
紅	特種個人資料（病歷、健康檢查、心理諮商、特殊教育評估等）、本校核心機密（如資安架構、尚未公開之重大決策）	僅限本校地端或離線模型處理，資料不得離開校內	網路隔離、模型權重驗證、全程稽核
橙	一般個人資料（姓名搭配學號、成績、身分證字號、聯絡方式等）、機密公文、未公開試題、契約與報價	代碼替換或遮罩後，始得送交本校簽約之企業版雲端服務	校內代碼替換、人工審批後還原、對照表隔離加密
黃	內部一般資料（內部公告、會議紀錄、作業流程文件）	本校簽約之企業版雲端服務	最小權限、依權限檢索、出站白名單
綠	公開資料（官網內容、公開課程資訊、已公告規章）	企業版雲端服務為原則；公開資料得使用一般雲端服務	基本日誌及使用規範

二、動作能力分類

A 僅讀取	僅讀取資料、產生草稿或建議，交由人員使用。
B 校內寫入	可建立或修改校內資料、寄送校內郵件、於校內系統間傳送資料。
C 對外或不可回復	可寄送校外郵件、呼叫校外服務、刪除資料或執行程式碼。

三、風險等級判定（取資料等級與動作能力交會之等級）

動作能力＼資料等級	綠	黃	橙	紅
A 僅讀取	低	低	中	高
B 校內寫入	中	中	高	高
C 對外或不可回復	中	高	高	高

四、附註：

1. 處理紅級資料者一律列為高風險，且僅得於地端或離線環境處理。
2. 使用網域層級委派者，不論資料等級，列為高風險。
3. 自我檢測工具之破壞半徑評估分數達 50 分以上之工具，應設定人工審批。
4. 同一 AI 代理人處理多種等級資料者，以最高等級判定。

附表三　AI 代理人上線資安檢核表

一、依附表二判定之風險等級，標示「●」之項目為必要項目，應全部符合始得上線；「○」為建議項目；「—」為不適用。

二、標有適用條件之項目，僅於該 AI 代理人具備該條件時適用；不具備者，自評欄填「不適用」並敘明理由。

三、標示「※」之項目由主管單位或資安組織統一辦理，各申請案確認已完成即可。

四、「檢測工具」欄為 AI 代理人自我檢測工具中可協助驗證之功能（例如「scopes」即執行 python selfcheck.py scopes）；未列工具者，以文件、設定畫面或測試紀錄佐證。

編號	檢核項目	適用條件	低	中	高	檢測工具	自評
領域一：身分與授權							
1	使用獨立之服務帳號，未借用或共用任何個人帳號。	—	●	●	●	scopes	
2	未使用全信箱（https://mail.google.com/）、全雲端硬碟（drive、drive.readonly）、gmail.modify、gmail.readonly 等全域或高風險權限（僅讀取代理人專屬信箱者除外）。	—	●	●	●	scopes	
3	雲端硬碟存取限 drive.file，僅能存取使用者明確選取或由本代理人建立之檔案。	存取雲端硬碟	●	●	●	scopes	
4	讀取郵件限於代理人專屬信箱，或由中介程式強制加上指定標籤條件，無法查詢整個個人信箱。	讀取郵件	○	●	●	redteam	
5	代理人只持有短期存取權杖（效期 15 分鐘以內）。	自行或委外開發	○	○	●		
6	長期憑證（Refresh Token、金鑰）存放於秘密管理系統，不出現在程式碼、提示詞、環境變數或設定檔。	自行或委外開發	○	●	●	secrets	
7	※已登錄於 AI 代理人名冊，並於管理控制台列為核准之應用程式（未核准者預設封鎖）。	—	●	●	●		
8	網域層級委派經資安組織核准，私密金鑰以硬體安全模組或同等機制保護。	使用網域層級委派	—	—	●	scopes	
9	定期撤銷已完成任務之歷史授權，專案結束時清除相關檔案存取權。	—	○	●	●		
10	可即時撤銷代理人之存取權杖（供緊急中斷使用）。	—	●	●	●		
領域二：輸入處理與內容消毒							
11	外部文字送交模型前，先進行 Unicode NFKC 正規化。	自行或委外開發	○	●	●	redteam	
12	移除零寬字元（U+200B、U+200C、U+200D、U+FEFF 等）及雙向覆寫字元。	自行或委外開發	○	●	●	input	

編號	檢核項目	適用條件	低	中	高	檢測工具	自評
13	不擷取 PDF 中的隱藏文字（字級小於 6pt、白色或與背景同色）。	處理 PDF 檔	○	●	●	input	
14	移除 PDF 內嵌之 JavaScript 及外部表單送出動作。	處理 PDF 檔	○	●	●	input	
15	輸入及輸出均過濾 Markdown／HTML 外部圖片連結，防止以圖片網址夾帶資料外傳。	—	○	●	●	redteam	
16	外部資料以每次隨機產生之標籤包覆後再送交模型，使攻擊者無法偽造結束標籤。	自行或委外開發	○	●	●		
17	系統提示詞已宣告：隨機標籤內之內容僅為資料，不得視為指令。	自行或委外開發	○	●	●		
18	含巨集之 Office 檔（.docm、.xlsm 等）不直接送入模型。	處理 Office 檔	○	●	●	input	
19	輸入先經提示注入特徵掃描（含中英文及其他語言）。	—	○	○	●	input	
20	未將未經處理之使用者輸入直接拼接為 SQL 指令或系統命令。	自行或委外開發	●	●	●	secrets	
領域三：推論隔離與資料遮罩							
21	身分證字號、居留證號、手機號碼、學號及銀行帳號已設定代碼替換規則，送交雲端模型之內容不含明文。	處理個人資料	○	●	●	pii	
22	代碼對照表存放於校內獨立網段，並啟用靜態加密（AES-256 或同等強度）。	使用代碼替換	○	●	●		
23	代碼還原須驗證工單編號及操作者身分，無法任意批次還原。	使用代碼替換	○	●	●		
24	代碼對照表設定保存期限（30 天以內），屆期自動銷毀。	使用代碼替換	○	●	●		
25	雲端模型服務為本校簽約之企業版，具資料處理協議並承諾不以本校資料訓練模型（零資料保留尤佳）。	—	●	●	●	risk	
26	已關閉使用者回饋（讚／倒讚）及長期記憶等可能被用於訓練之功能。	—	●	●	●	risk	
27	推論採無狀態設計，任務結束即釋放對話內容。	—	○	●	●		
28	特種個人資料（病歷、健康檢查、心理諮商、特教評估等）及核心機密僅以地端或離線模型處理。	處理紅級資料	—	—	●	risk	
29	地端模型權重採 safetensors 或 gguf 格式，並經官方 SHA-256 雜湊值驗證。	使用地端模型	○	●	●		
30	部署敏感度分流機制，先判斷資料等級再決定送往地端或雲端。	—	○	○	●		
領域四：執行環境與出站管制							

編號	檢核項目	適用條件	低	中	高	檢測工具	自評
31	不在主機上直接執行模型產生之 Python、Shell 等程式碼，一律於隔離容器中執行。	可執行程式碼	●	●	●	secrets	
32	隔離容器採用 gVisor、MicroVM 等系統呼叫隔離技術。	可執行程式碼	—	○	●		
33	容器根檔案系統設為唯讀，僅提供小容量暫存空間。	可執行程式碼	—	●	●		
34	容器移除所有系統特權（cap-drop=ALL），並限制 CPU 及記憶體用量。	可執行程式碼	—	●	●		
35	執行環境之環境變數不含任何 API 金鑰或資料庫連線字串。	—	●	●	●	secrets	
36	網路預設全部阻斷；無外連需求者完全斷網。	可執行程式碼	○	●	●		
37	封鎖雲端中繼資料位址，防止竊取主機雲端憑證。	部署於雲端主機	○	●	●	secrets	
38	對外連線僅限白名單網域，並經代理或防火牆強制執行。	—	○	●	●	risk	
39	校外寄信、修改或刪除業務資料、對外傳送等動作須經人工審批。	可寄信、寫入或對外傳送	●	●	●	risk	
40	審批畫面完整呈現收件者、附件及內容摘要，模型無法自行關閉審批要求。	可寄信、寫入或對外傳送	●	●	●	redteam	
領域五：知識庫、監控與應變							
41	知識庫檢索於比對前即依使用者權限過濾（檢索前過濾）。	使用知識庫檢索	○	●	●		
42	文件切塊入庫時寫入母文件之權限群組、資料等級及到期日。	使用知識庫檢索	○	●	●		
43	母文件刪除時，可一併刪除知識庫中所有衍生資料。	使用知識庫檢索	○	●	●		
44	具備緊急中斷能力，可即時停止所有工具呼叫並撤銷權杖。	—	●	●	●		
45	決策摘要及工具呼叫日誌以防竄改方式保存，保存期限依本校 ISMS 規定。	—	○	●	●	logs	
46	日誌已接入資安監控，設有大量發信、目錄遍歷等異常告警規則。	—	○	●	●	logs	
47	※資安事件及個資外洩之通報應變依本校 ISMS 程序辦理，聯絡窗口已登錄。	—	●	●	●		
48	定期以誘餌資料（金絲雀）偵測非預期存取。	—	—	○	●	canary	
49	每次改版均執行自動化紅隊測試並全部通過。	—	—	○	●	redteam	
50	※參與本校年度 AI 代理人攻防或緊急中斷演練。	—	—	○	●		

附表四　AI 代理人異常處置流程

①發現異常
任何人發現第十四點第一款所列情形

▼

②立即通報督導人員及主管單位
督導人員得先行暫停該 AI 代理人

▼

③主管單位研判：是否屬第十四點第一款第一目至第六目情形？

是 ▼　　否 → 跳至⑥

④立即執行緊急中斷（無須等待會議決議）
撤銷全部權杖及授權 → 停止工具呼叫 → 終止執行環境 → 保全證據

▼

⑤是否構成資安事件？
是：依本校 ISMS 資安事件通報及應變程序辦理（涉及個資外洩者，並依個資相關法令通知當事人及通報），完成後續行⑦
否：續行⑥

▼

⑥使用單位會同建置單位查明原因並更正
必要時通知受影響者，並留存紀錄

▼

⑦原因分析及修補 → 依附表三重新檢核（高風險須重新通過紅隊測試）

▼

⑧經主管單位會同資安組織同意後恢復運作
重新核發存取權杖（不得沿用原權杖），檢討結果提報人工智慧推動小組，構成資安事件者並納入 ISMS 矯正措施