Reclaim2K.org x 政府AI指引

HK Government AI Guidelines x App Alignment

《人工智能道德框架》v2.0 · 《生成式人工智能技術及應用指引》v1.1
數字政策辦公室 · 2025年12月

透過默想聖經戒除手機成癮 APP
Phone Detox Bible · Reclaim2K.org
中華傳道會劉永生中學 CNEC Lau Wing Sang Secondary School

「你們必曉得真理，真理必叫你們得以自由。」約翰福音 8:32

GOVERNMENT GUIDELINES

三份政府AI指引概覽
Overview of HK Government AI Guidelines

人工智能道德框架
Ethical AI Framework

v2.0 · 2025年12月

- 12項道德原則
- 三道防線治理
- 6階段AI生命週期
- AI影響評估範本

生成式AI指引
GenAI Guideline

v1.1 · 2025年12月

- 6大技術局限
- 4級風險分類
- 5大治理維度
- 教育界特別指引

簡易參考指南
Quick Reference

v2.0 · 2025年12月

- 框架精華版
- 重點行動清單
- 原則速覽表
- 評估流程圖

教育界特別指引

附錄 §2.2.2 第37頁

四個核心要點:

1. 不應普遍禁止學生使用 — 規範使用，而非一刀切禁止
2. 課業使用須取得教師同意 — 老師是把關人
3. 生成內容必須可識別 — 避免學術誠信問題(標明AI產生)
4. 教師使用AI必須人工審核 — 尤其批改作業/試卷

來源 Source: 數字政策辦公室 Digital Policy Office · 香港生成式人工智能研發中心 HKGAI

📋 政府指引 | Government Guidelines

十二項人工智能道德原則
12 AI Ethical Principles

《人工智能道德框架》v2.0 — 數字政策辦公室 Digital Policy Office, Dec 2025

⚙️ 執行原則 Execution Principles — 其他原則的基礎

1. 透明及可解釋 Transparency & Explainability　2. 可靠、穩健及安全 Reliability, Robustness & Safety

📜 一般原則 General Principles — 源自《世界人權宣言》及香港法例

3. 公平 Fairness
4. 多元與包容 Diversity & Inclusion
5. 人為監督 Human Oversight
6. 合法與合規 Legal Compliance
7. 數據私隱 Data Privacy
8. 問責 Accountability
9. 有益的AI Beneficial AI
10. 合作與開放 Cooperation & Openness
11. 可持續性 Sustainability
12. 安全 Safety

💡 所有AI項目均應遵守以上十二項原則 All AI projects should adhere to these 12 principles

來源:《人工智能道德框架》第3.5.1節 | 《簡易參考指南》第2章

📍 校內驗證聲明

本AI項目經校內驗證符合香港相關指引。

2026年4-5月在教育局課程發展處PSHE宗教組協調下，請友校先進行測試。

稍後交由相關專業NGO/公司進行優化，並呈交專業人士建議後，方供其他學校自由使用。

政府指引| Government Guidelines

人工智能生命週期與治理
AI Lifecycle & Governance

1. **項目策略** Strategy
   機構策略、政策、標準
2. **項目規劃** Planning
   組合管理、監督、交付
3. **項目生態** Ecosystem
   技術路線圖、採購
4. **項目開發** Development
   數據、模型、測試
5. **系統部署** Deployment
   整合、過渡、監察
6. **運作監察** Operations
   表現監察、持續合規

校內驗證聲明

本AI項目經校內驗證符合香港相關指引。

2026年4-5月在教育局課程發展處PSHE宗教組協調下，請友校先進行測試。

稍後交由相關專業NGO/公司進行優化，並呈交專業人士建議後，方供其他學校自由使用。

三道防線治理架構 Three Lines of Defence

第一道防線

1st Line — 項目團隊 Project Team

開發AI應用、風險評估、執行緩解措施、記錄評估

第二道防線

2nd Line — 督導委員會 Steering

品質保證、驗收標準、獨立檢視、批准AI應用

第三道防線

3rd Line — IT委員會/CIO

審核、建議、監管高風險AI應用

來源:《人工智能道德框架》第3.5.2-3.5.3節 |《簡易參考指南》第3-4章

生成式AI的技術局限與服務風險
GenAI Technical Limitations & Service Risks

🔧 技術局限 Technical Limitations

🤖 模型幻覺 Hallucination
生成與事實不符的資訊

⚖️ 模型偏見 Bias
貫穿模型全生命週期

🔲 黑盒問題 Black Box
決策機制不透明

🔢 數理能力 Math Limits
邏輯推理不可靠

🎯 輸入敏感性 Input Sensitivity
微小變化導致不同結果

💾 數據完整性 Data Integrity
資料投毒、漂移風險

🚨 服務風險 Service Risks

🛑 內容安全 Content Safety
生成不良/違法內容

📰 製造謠言 Misinformation
以假亂真的多媒體

🔓 模型越獄 Jailbreaking
繞過安全防護機制

📤 數據洩露 Data Leakage
個人/機密資料外洩

💡 **教師須知 For Teachers:**
生成式AI內容可能包含錯誤
必須經人工審核後才使用
AI content may contain errors —
always verify before use

來源:《生成式人工智能技術及應用指引》v1.1 第2.1節

✅ 教育界特別指引 Education-Specific Compliance

📖 聖經經文規則

🚫 AI 絕不可生成、改寫或填補聖經經文

✅ 所有經文硬編碼 + 人工比對 CUV/NRSV/NIV 版本

✅ 神學問題轉交教師/牧者回答，絕不委託AI

✅ AI生成靈修/聖經內容必須經人工審核

⚖️ 宗教公平 §2.3.4

✅ 基督教教導清晰且居中心地位

✅ AI絕不用於貶低佛教、道教、伊斯蘭教、印度教或任何信仰

✅ 超級管理員審核並拒絕所有11個§2.3.4類別的偏見AI提交

🚪 自由參與及退出機制

學生及家長均可隨時加入、不加入或退出

👤 學生: 自由選擇是否使用AI分頁

👨‍👩‍👧 家長: 電郵教師即可隨時退出

🏫 學校: 從Sheet移除AI內容

✅ 可逆轉 · 無記錄 · 無後果

🔐 雙重上傳監管機制 Two-Layer Upload Supervision

第一層: 僅參與學校的教師可上傳內容 — 學生及用戶絕無上傳權限

第二層: 超級管理員審核所有上傳 — 確保內容符合香港法律法規且對所有人健康有益

交叉義務: 未成年用戶+個人資料+宗教內容+AI生成 | 來源: 《私隱條例》、GenAI指引§2.3.4、PCPD 2025/3

✅ 資料保護與PCPD合規 Data Protection & PCPD

✅

🔒 數據訓練聲明 Data Training Declaration

零用戶資料傳至AI供應商訓練

- ✅ App零運行時AI API呼叫
- ✅ 靈修紀錄、照片等個人資料僅存本地/學校Sheet
- ✅ 符合《個人資料(私隱)條例》對未成年用戶的保護要求
- ✅ System-prompt範本已備，為未來LLM功能預留

🏛️ PCPD 2025年3月指引合規

- ✅ 僅限核准AI工具 (Approved tools only)
- ✅ 禁止黏貼個人資料至AI (No personal data pasting)
- ✅ 人工驗證所有AI輸出 (Human verification)
- ✅ SecurityLog記錄所有操作
- ✅ 回饋渠道已建立 (Feedback channel)

📊 合規記分卡 Compliance Scorecard

12項倫理原則: 12/12 ✅ | AI透明度缺口: ✅ 已列明 | AI影響評估缺口: ✅ 已列明
Suno標籤: → ✅ 🤖徽章 | 公開AI工具清單: → ✅ reclaim2k.org + App內

文檔: AI_IMPACT_ASSESSMENT.md | AI_GOVERNANCE_MANUAL.md | PDF_SLIDE_REVISIONS.md
來源: 《私隱條例》| PCPD「人工智能：模型個人資料保障」2025/3 | 《生成式AI指引》§2.3.4

風險分級制度與治理維度
Risk Classification & Governance Dimensions

不可接受風險
Unacceptable

全面禁止
Full Ban

危害市民安全
潛意識操控

高風險
High Risk

合規評估+
人類監督+實時監測

醫療診斷
自動駕駛

有限風險
Limited Risk

透明度要求+
退出機制+年度審計

招聘工具
教育AI

低風險
Low Risk

企業自我認證

垃圾郵件過濾
創意工具

📱 **Reclaim2K 定位：「有限風險」教育類AI應用 — 需要透明度要求、退出機制、定期審計**

🚦 風險評級聲明
本AI項目屬有限風險。
2026年4-5月在教育局課程發展處PSHE宗教組協調下，請友校先進行測試。
稍後交由相關專業NGO/公司進行優化，並呈交專業人士建議後，方供其他學校自由使用。

✅ **符合要求:**
透明度要求 · 退出機制 · 年度審計

🏛️ 治理五大維度 Five Governance Dimensions

🔒 個人資料私隱
Privacy

© 知識產權
IP

🚔 犯罪防治
Crime

✅ 真實可信
Trust

🛡️ 系統安全
Security

來源:《生成式人工智能技術及應用指引》v1.1 第2.1.2及2.2節

✅ 對照分析 Alignment Analysis

Reclaim2k.org 與政府指引吻合之處
Alignment Strengths

🌿 有益的人工智能
Beneficial AI ✅ ✅

免費手機排毒+靈修，5種入口，促進學生身心靈健康

🔒 數據私隱
Data Privacy ✅ (upgraded)

SHA-256雜湊DOB、校本Google Sheet隔離、PDPO合規

👁 人為監督
Human Oversight ✅ ✅

教師儀表板、超級管理員審核所有內容、DOB驗證防冒用

🛡 安全
Security ✅ (upgraded)

CSP、App Token、限速5次/15分鐘、2FA、XSS強化、審計日誌

⚖️ 公平與包容
Fairness & Inclusion ✅

中英雙語、五類用戶、公眾零註冊、匿名口號、無GPS追蹤

🤝 合作與開放
Cooperation & Openness ✅ ✅

跨校分享、教會小組、合作夥伴、教育局計劃配合

📌 經網站分析後，隱私與安全從「待改善」提升為「吻合」| Privacy & Security upgraded after website review

✅ 吻合亮點 Alignment Highlight

數據私隱與安全保障 🔒
Data Privacy & Security

🔐 數據私隱 Data Privacy

- DOB經SHA-256雜湊 — 連開發者也無法讀取
- 校本Google Sheet隔離 — 跨校不可見
- 公眾用戶零伺服器資料(localStorage)

🛡️ 系統安全 System Security

- CSP內容安全政策 + App Token
- PIN限速(5次/15分鐘) + 2FA雙重驗證
- XSS強化(escapeHTML) + SecurityLog
- 教師PIN≥8字元、protectSensitiveTabs

📊 按角色的資料儲存一覽 Data Storage by Role

🎒 學生: 校本Google Sheet (班主任可見) | 👨‍👩‍👧 家長: 唯讀(僅限子女) | 🌱 公眾: 瀏覽器本地(零伺服器)
⛪ 教會導師: 僅姓名+電郵，反思只以電郵傳送 | 👩‍🏫 獨立老師: 自有Sheet(自己掌控存取)
來源: reclaim2k.org — Privacy & Security section

✅ 吻合亮點 Alignment Highlight

內容安全與人為監督 ✅

Content Safety & Human Oversight

內容安全 Content Safety

- 超級管理員審核所有發佈內容
- 口號匿名、Suno歌曲經審批
- 教師精選外部影片 推薦
- 學生無法於 App 內直接呼叫 ChatGPT（App 零運行時 AI API）

人為監督 Human Oversight

- 教師儀表板: 追蹤靈修進度、背經
- DOB驗證防止冒用登入
- 家長僅唯讀(看不到反思內容)
- AI輔助但人做決定(教師同意)

問責機制 Accountability

- SecurityLog記錄所有敏感操作(密碼變更、資料清除、獎勵)
- 清晰角色定義: 學生/教師/家長/管理員，各有不同權限
- feedback@reclaim2k.org 回饋渠道 | 資料刪除權

來源: reclaim2k.org — Content Safety, Human Oversight, Privacy & Security sections

✅ 吻合亮點 Alignment Highlight

AI透明度與可解釋性 ✅

AI Transparency & Explainability

✅ 已完成措施 Completed

- Suno歌曲已加🤖 AI Generated標籤
- 紫色透明度橫幅已加入Suno分頁
- AI工具清單公開於reclaim2k.org

📜 政府指引要求 ✅ 已達標

- 生成內容已能明確被識別 ✅
- 使用引用: 🤖標籤+透明度橫幅 ✅
- 5項透明度承諾已公開於網站 ✅

✅ **風險等級: 已解決 Resolved — 🤖標籤、透明度橫幅、AI工具清單已全部上線**

📋 實施證據 Implementation Evidence

✅ APP隱私政策新增第12節「AI工具與透明度」(5 AI工具 + 5項承諾 + 3道防線)
✅ reclaim2k.org 新增 #ai-transparency 專區 (6張工具卡 + 4個可摺疊面板)
✅ Suno分頁: mkSunoNotice() 紫色橫幅 + mkList(list,true) 🤖徽章

來源: 《生成式AI指引》§2.3.2 安全透明 | §3.3 使用者引用說明

✅ 吻合亮點 Alignment Highlight

AI影響評估與三道防線 ✅

AI Impact Assessment & Three Lines of Defence

📋 AI應用影響評估 ✅ 已完成

1. 風險分級: 有限風險 ✅
2. 全生命週期評估 ✅
3. 正式評估文檔 ✅

🏛️ 三道防線 ✅ 已正式化

- 1️⃣ 教師: 內容審核+上傳 ✅
- 2️⃣ 超級管理員: 品質保證 ✅
- 3️⃣ 校方IT/CIO: 審核高風險 ✅
- 📌 架構已正式化並記錄於文檔

📋 實施證據 Implementation Evidence

✅ AI_IMPACT_ASSESSMENT.md — 10節正式評估文檔已完成
✅ AI_GOVERNANCE_MANUAL.md — 教師+管理員操作手冊
✅ 年度/重大變更評估更新機制已建立

來源: 《AI道德框架》第5章 | 《簡易參考指南》第6章

✅ 全部完成 All Phases Complete

改善路線圖 ✅ 全部於2026年4至5月完成
Improvement Roadmap — All Complete

第一階段 ✅

Phase 1: Labeling ✅

- ✅ Suno 標籤
- ✅ AI工具清單上線
- ✅ 透明度橫幅

第二階段 ✅

Phase 2: Documentation ✅

- ✅ AI影響評估完成
- ✅ 三道防線正式化
- ✅ 治理手冊完成
- ✅ 隱私政策更新 (§12)

第三階段 ✅

Phase 3: Ongoing Maturity ✅

- ✅ 年度評估更新機制
- ✅ EDB計劃持續合規
- ✅ 跨校經驗分享機制

📌 三階段全部完成 — 12/12倫理原則 · AI透明度 ✅ · AI影響評估 ✅

All phases complete — 12/12 principles · Transparency ✅ · Impact Assessment ✅

結語 Closing

📌 重點回顧 Key Takeaways

1. Reclaim2K.org 與政府AI指引高度吻合 — 12項原則全達標或超標
2. 隱私與安全措施遠超同類教育APP (SHA-256, CSP, 2FA, 審計日誌)
3. §2.2.2 義務全覆蓋 — AI標注、條款揭示、退出機制齊備
4. 雙層上傳 (Two-Layer Upload) — 教師審核→管理員批准
5. 三階段路線圖: 友校測試 → NGO優化 → 全面開放

📚 參考文件 References

- 《人工智能道德框架》v2.0 — 數字政策辦公室 (Dec 2025)
- 《生成式AI技術及應用指引》v1.1 — 數字政策辦公室 + HKGAI (Dec 2025)
- reclaim2k.org — Privacy & Security | Full website review
- two-layer-upload-architecture.md | s2.2.2-obligation-coverage-checklist.md
- rollout-phases-roadmap.md | ai-disclosure-badges-spec.md

「你要保守你心，勝過保守一切，因為一生的果效是由心發出。」

"Above all else, guard your heart, for everything you do flows from it."

— 箴言 Proverbs 4:23

Reclaim2K | reclaim2k.org | feedback@reclaim2k.org | 🏃✝ Run to Jesus