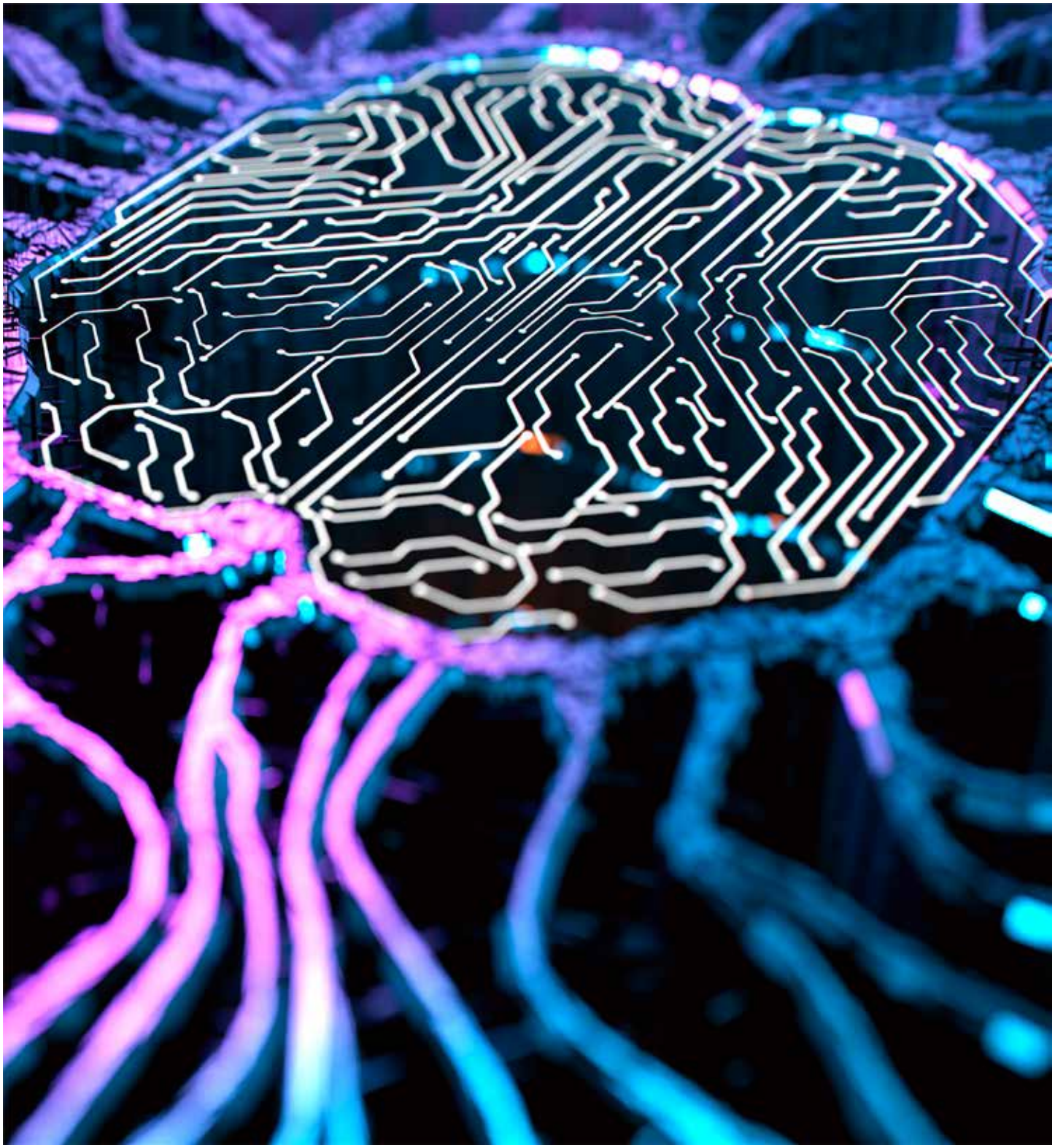
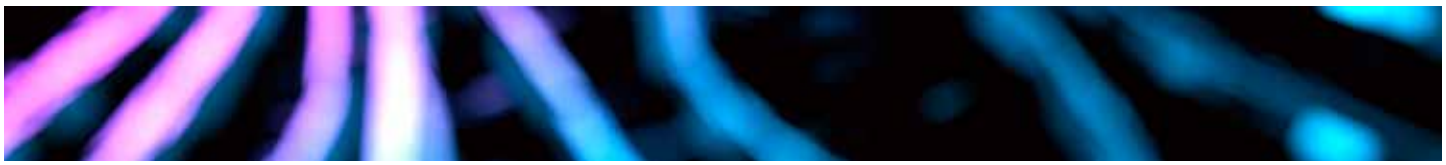
安全人工智能系統開發指引





National Cyber Security Centre
a part of GCHQ

CISA
CYBERSECURITY & INFRASTRUCTURE SECURITY AGENCY

Australian Government
Australian Signals Directorate

ASD AUSTRALIAN SIGNALS DIRECTORATE
ACSC Australian Cyber Security Centre

Communications Security Establishment
Canadian Centre for Cyber Security
Centre de la sécurité des télécommunications
Centre canadien pour la cybersécurité

CSIRT
Equipo de Respuesta ante Incidentes de Seguridad Informática

National Cyber and Information Security Agency

Republic of Estonia
Information System Authority

Department of Justice
Federal Bureau of Investigation

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Federal Office for Information Security

INCD
Israel National Cyber Directorate

AGENZIA PER LA CYBERSICUREZZA NAZIONALE
ACN
SAPIENTIA ADIUVAT

NISC
内閣サイバーセキュリティセンター
National center of Incident readiness and Strategy for Cybersecurity

National Cyber Security Centre

NITDA

NSM
NORWEGIAN NATIONAL CYBER SECURITY CENTRE

National Security Agency
United States of America

NASK

국가정보원
National Intelligence Service

Ministerstwo Cyfryzacji

CSA SINGAPORE
Cyber Security Agency of Singapore

關於本文件

本文件由英國國家網絡安全中心 (NCSC)、美國網絡安全與基礎設施安全局 (CISA) 以及以下國際合作夥伴發布：

- 國家安全局 (NSA)
- 聯邦調查局 (FBI)
- 澳洲信號局澳洲網絡安全中心 (ACSC)
- 加拿大網絡安全中心 (CCCS)
- 紐西蘭國家網絡安全中心 (NCSC-NZ)
- 智利政府 CSIRT
- 捷克國家網絡與資訊安全局 (NUKIB)
- 愛沙尼亞資訊系統管理局 (RIA) 和愛沙尼亞國家網絡安全中心 (NCSC-EE)
- 法國網絡安全局 (ANSSI)
- 德國聯邦資訊安全辦公室 (BSI)
- 以色列國家網絡管理局 (INCD)
- 義大利國家網絡安全局 (ACN)
- 日本國家網絡安全事件準備與戰略中心 (NISC)
- 日本內閣府科學技術創新政策秘書處
- 尼日利亞國家資訊科技發展局 (NITDA)
- 挪威國家網絡安全中心 (NCSC-NO)
- 波蘭數碼事務部
- 波蘭 NASK 國家研究所 (NASK)
- 韓國國家情報院 (NIS)
- 新加坡網絡安全局 (CSA)

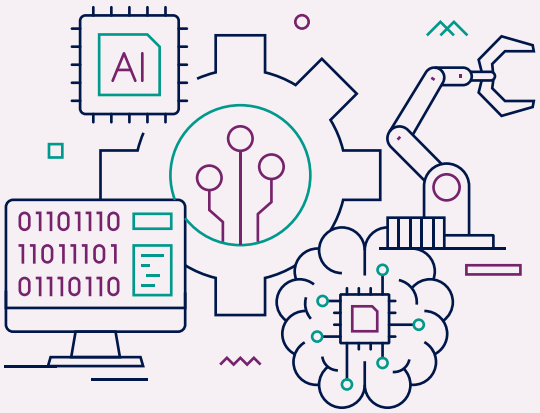
致謝

以下組織為這些指引的製定作出了貢獻：

- 艾倫圖靈研究所
- 人擇
- 數據區塊
- 喬治城大學安全與新興科技中心
- 谷歌
- 谷歌深智
- IBM
- ImBue
- 微軟
- OpenAI
- Palantir
- RAND
- Scale AI
- 卡內基美隆大學軟件工程學院
- 史丹福人工智能安全中心
- 史丹福大學地緣政治、技術與治理項目

免責聲明

目錄



執行摘要

本文件為任何使用人工智能 (AI) 的系統的提供商提供了指引，無論這些系統是從頭開始創建的還是構建在他人提供的工具和服務之上的。實施這些指引將有助於提供商構建按預期運行的人工智能系統，在需要時可用，並且在工作時不會向未授權方洩露敏感數據。

本文件主要針對使用機構託管的模型或使用外部應用程式介面 (API) 的人工智能系統提供商。我們敦促**所有**利害關係人（包括數據科學家、開發人員、管理人員、決策者和風險負責人）閱讀這些指引，以幫助他們就人工智能系統的**設計**、**開發**、**部署**和**運行**作出明智的決策。

關於指引

人工智能系統有潛力為社會帶來諸多好處。然而，為了充分實現人工智能的發展，必須以安全和負責任的方式開發、部署和運行人工智能。

人工智能系統存在新的安全漏洞，需要與標準網絡安全威脅一起考慮。當發展速度很快時（如人工智能的情況一樣），安全往往成為次要考慮因素。不僅在開發階段，而且在系統的整個生命週期中，安全都必須是一項核心要求。

因此，本指引分為人工智能系統開發生命週期中的四個關鍵領域：**安全設計**、**安全開發**、**安全部署**和**安全運維**。對於每個部分，我們都提出了有助於降低機構的人工智能系統開發過程的整體風險的考慮因素和緩解措施。

1. **設計安全**
   本節包含適用於人工智能系統開發生命週期的設計階段的指引。它涵蓋了對風險和威脅建模的理解，以及系統和模型設計時需要考慮的特定主題和權衡。

2. **開發安全**
   本節包含適用於人工智能系統開發生命週期的開發階段的指引，包括供應鏈安全、文件以及資產和技術債務管理。

3. **部署安全**
   本節包含適用於人工智能系統開發生命週期的部署階段的指引，包括保護基礎設施和模型免受破壞、威脅或損失、制定事故管理流程以及負責任的發布。

4. **運維安全**
   本節包含適用於人工智能系統開發生命週期的安全運維階段的指引。它提供了在系統部署後尤其相關的操作指引，包括日誌記錄和監控、更新管理和資訊共享。

此指引遵循「預設安全」方法，並且與 NCSC 的安全開發和部署指引、NIST 的安全軟件開發架構，以及CISA、NCSC 和國際網絡機構發布的「設計安全原則」中定義的做法密切保持一致。他們優先考慮：

- 為客戶掌控安全成果
- 接受徹底的透明度和問責制
- 建立組織結構和領導力，使設計安全成為業務的首要任務

引言

人工智能（AI）系統有潛力為社會帶來諸多好處。然而，為了充分實現人工智能的發展，必須以安全和負責任的方式開發、部署和運行人工智能。網絡安全是人工智能系統安全性、復原力、隱私性、公平性、有效性和可靠性的必要前提。

然而，人工智能系統面臨新的安全漏洞，需要與標準網絡安全威脅一起考慮。當發展速度很快時（如人工智能的情況一樣），安全往往成為次要考慮因素。不僅在開發階段，而且在系統的整個生命週期中，安全都必須是一項核心要求。

本文件為任何使用AI的系統的提供商[1]提供了指引，無論這些系統是從頭開始創建的還是構建在他人提供的工具和服務之上的。實施這些指引將有助於提供商構建按預期運行的AI系統，在需要時可用，並且在工作時不會向未授權方洩露敏感數據。

這些指引應與已建立的網絡安全、風險管理和事故回應最佳實踐結合起來考慮。我們特別敦促提供商遵循美國網絡安全和基礎設施安全局（CISA）、英國國家網絡安全中心（NCSC）以及我們所有國際合作夥伴制定的「設計安全」[2]原則。這些原則優先考慮：

- 為客戶掌控安全成果
- 接受徹底的透明度和問責制
- 建立組織結構和領導力，使設計安全成為業務的首要任務

遵循「設計安全」原則需要在系統的整個生命週期中投入大量資源。這意味著開發人員必須在系統設計的每一層以及開發生命週期的所有階段，投資於優先保護客戶的工具**功能**、**機制**和**實施**。這樣做可以避免日後進行代價高昂的重新設計，並在短期內保護客戶及其數據的安全。

為什麼人工智能安全有所不同？

在本文件中，我們使用「AI」一詞來特指機器學習（ML）應用程式[3]。所有類型的ML都在範圍內。我們將 ML 應用程式定義為：

- 涉及軟件組件（模型），使電腦能夠識別數據模式並為其提供上下文，而無需由人類明確編程規則
- 基於統計推理產生預測、建議或決策

除了現有的網絡安全威脅外，人工智能系統還面臨新型漏洞的影響。「對抗性機器學習」（AML）一詞用於描述對 ML 元件（包括硬件、軟件、工作流程和供應鏈）中基本漏洞的利用。AML使攻擊者能夠在ML系統中引發意外的行為，其中包括：

- 影響模型的分類或迴歸效能
- 允許使用者執行未經授權的操作
- 擷取敏感模型資訊

實現這些效果的方法有很多，例如大型語言模型（LLM）領域的提示注入攻擊，或故意破壞訓練數據或用戶反饋（稱為「數據中毒」）。

誰應該閱讀本文件？

本文件主要針對人工智能系統的供應商，無論是基於機構託管的模型還是使用外部應用程式介面 (API)。我們敦促**所有**利害關係人（包括數據科學家、開發人員、管理人員、決策者和風險負責人）閱讀這些指引，以幫助他們就機器學習AI系統的**設計**、**部署**和**運行**作出明智的決策。

儘管如此，並非所有指引都直接適用於所有機構。攻擊的複雜程度和方法會因人工智能系統的對手而有所不同，因此在考慮使用指引時，還應考慮各機構的用例和威脅概況。

誰負責開發安全的人工智能？

現代人工智能供應鏈中往往有許多參與者。一個簡單的方法是假設兩個實體：

- 「提供商」負責數據管理、算法開發、設計、部署和維護
- 「用戶」負責提供輸入並接收輸出

雖然這種提供商-用戶的方法已在許多應用中使用，但它變得越來越不常見[4]，因為提供商可能會將第三方提供的軟件、數據、模型和/或遠端服務合併到自己的系統。這些複雜的供應鏈使最終用戶更難理解安全人工智能的責任所在。

使用者（無論是「最終使用者」或包含外部AI組件的提供者[5]）通常沒有足夠的可見性和/或專業知識來完全理解、評估或解決與其所使用的系統相關的風險。因此，根據「設計安全」原則，**AI組件提供者應對供應鏈下游使用者的安全結果負責。**

提供者應盡可能在其模型、管道和/或系統中實施安全控制和緩解措施，並且在使用設定的情況下，預設為實施最安全的選項。如果風險無法減輕，提供者應負責：

- 告知供應鏈下游的用戶他們和（如果適用）他們自己的用戶正在接受的風險
- 建議他們如何安全地使用該組件

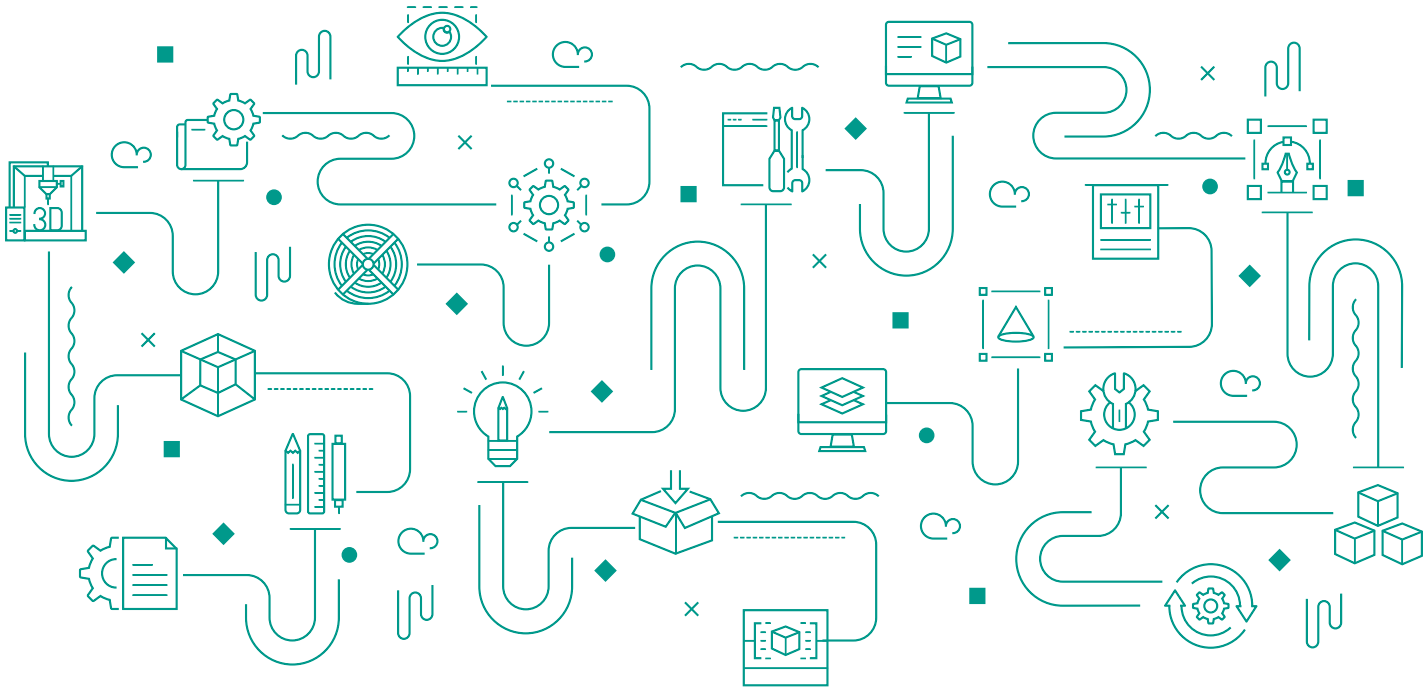
如果系統受損可能導致有形或廣泛的實體或聲譽損害、業務營運重大損失、敏感或機密資訊外洩和/或法律影響，則應將人工智能網絡安全風險視為**嚴重**。

安全人工智能系統開發指引

本指引分為AI系統開發生命週期中的四個關鍵領域：**安全設計**、**安全開發**、**安全部署**和**安全運維**。對於每個領域，我們提出了有助於降低機構人工智能系統開發過程整體風險的注意事項和緩解措施。

本文件中列出的指引與以下定義的軟件開發生命週期實踐密切相關：

- NCSC 的安全開發與部署指引
- 國家標準與技術研究院 (NIST) 安全軟件開發架構 (SSDF)[6]



1. 設計安全

本節包含適用於AI系統開發生命週期的**設計**階段的指引。內容涵蓋了對風險和威脅建模的理解，以及系統和模型設計時需要考慮的特定主題和權衡。

提高員工對威脅和風險的認知

系統所有者和高層領導者了解保護AI面臨的威脅及其緩解措施。你的數據科學家和開發人員保持對相關安全威脅和故障模式的認識，並幫助風險所有者做出明智的決策。你為用戶提供有關AI系統面臨的獨特安全風險的指導（例如，作為標準資訊安全培訓的一部分），並對開發人員進行安全編碼技術以及安全和負責任的AI實踐方面的培訓。

模擬系統面臨的威脅

作為風險管理流程的一部分，你可以應用整體流程來評估系統面臨的威脅，其中包括了解如果AI元件受到損害或出現意外行為，會對系統、用戶、機構和更廣泛社會造成哪些潛在影響[7]。此過程涉及評估AI特定威脅的影響[8]並記錄你的決策。

你認識到系統中使用的數據的敏感度和類型可能會影響其作為攻擊者目標的價值。你的評估應考慮到，隨著AI系統越來越被視為高價值目標，以及AI本身啟用新的、自動化攻擊媒介，某些威脅可能會增加。

設計系統時兼顧安全、功能和性能

你確信使用AI可以最恰當地解決當前的任務。確定這一點後，你可以評估特定於AI的設計選擇的適當性。你需要考慮威脅模型和相關的安全緩解措施以及功能、用戶體驗、部署環境、效能、保證、監督、道德和法律要求等因素。例如：

- 在選擇內部開發還是使用外部元件時，你會考慮供應鏈安全，例如：
    - 你在訓練新模型、使用現有模型（無論是否進行微調）或透過外部 API 存取模型方面的選擇都是適合你的要求的
    - 在選擇與外部模型提供商合作時，對提供商本身安全狀況進行了盡職調查評估
    - 如果使用外部程式庫，你需要完成盡職調查評估（例如，確保程式庫具有控制措施，可防止系統載入不受信任的模型，而不會立即將自身暴露於任意代碼執行[9]）
    - 在導入第三方模型或序列化權重時，實施掃描和隔離/沙箱，這些模型或序列化權重應被視為不受信任的第三方代碼，並且可以啟用遠端代碼執行

- 如果使用外部API，則可以對可傳送至機構控制以外的服務的數據進行適當的控制，例如要求用戶在發送潛在敏感資訊之前登入並確認
- 對數據和輸入進行適當的檢查和驗證；這包括將使用者回饋或持續學習數據納入模型時，認識到訓練數據決定系統行為

- 將AI軟件系統開發與現有的安全開發和營運最佳實踐相結合；AI系統的所有元素都是在適當的環境中使用編碼實踐和語言編寫的，這些編碼實踐和語言在可行的情況下可以減少或消除已知類別的漏洞
- 如果AI組件需要觸發操作，例如修改文件或將輸出導向到外部系統，則對可能的操作應用適當的限制（必要時包括外部AI和非AI故障保護）
- 有關用戶交互的決策應考慮AI的特定風險，例如：
  - 你的系統為用戶提供可用輸出的同時，而不會向潛在攻擊者透露不必要的詳細資訊
  - 如有必要，你的系統會為模型輸出提供有效的防護措施
  - 如果向外部客戶或合作者提供API，你可以應用適當的控制措施來減輕透過 API對AI系統的攻擊
  - 預設情況下，將最安全的設定整合到系統中
  - 應用最低權限原則限制對系統功能的存取
  - 向用戶解釋風險較高的功能並要求用戶選擇使用這些功能；向用戶告知禁止使用案例，並在可能的情況下告知用戶替代解決方案

選擇AI模型時考慮安全優勢和權衡利弊

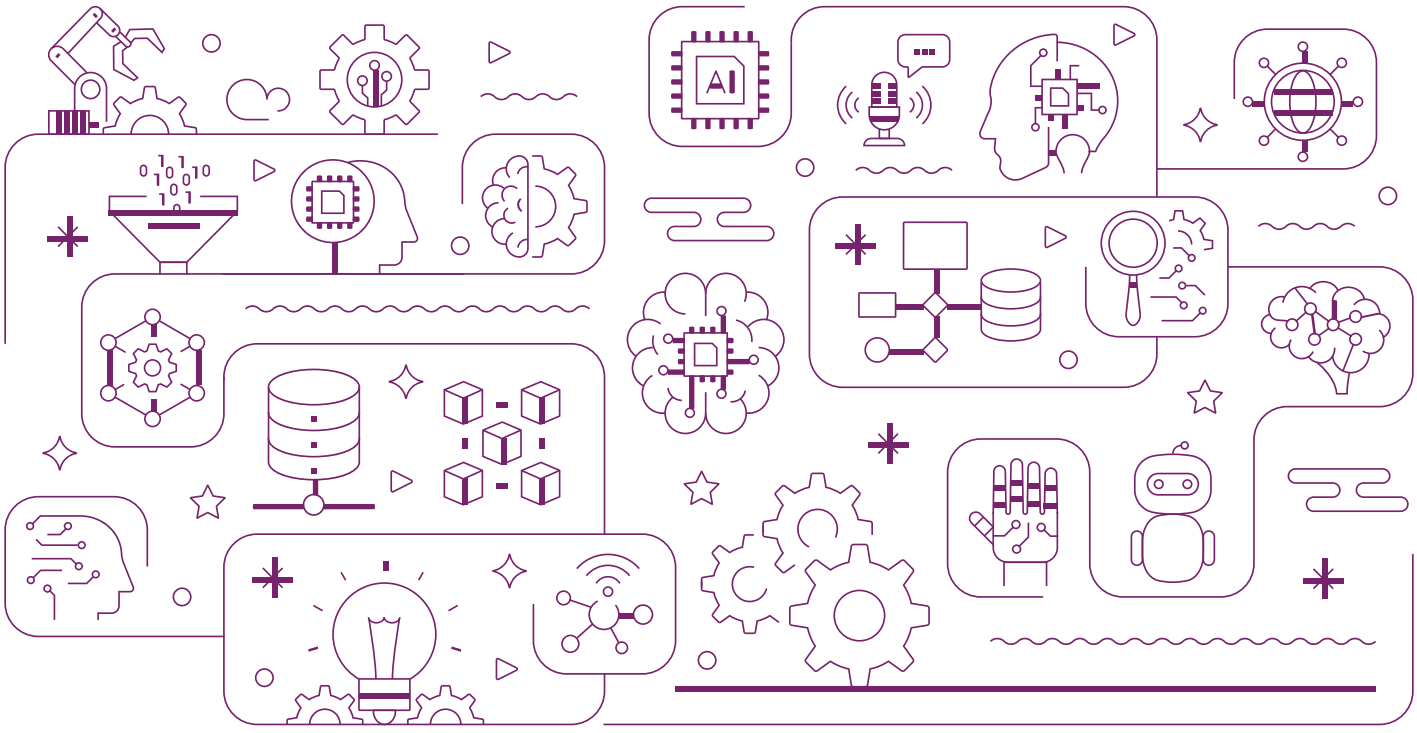
你對AI模型的選擇將涉及平衡一系列要求。這包括模型架構、配置、訓練數據、訓練演算法和超參數的選擇。你的決策將參考你的威脅模型，並隨著AI研究安全的進步和對威脅的理解的發展定期重新評估。

選擇人工智能模型時，你的考慮因素可能包括但不限於：

- 你所使用的模型的複雜性，即所選的架構和參數數量；除其他因素外，你所選擇的模型架構和參數數量將影響所需的訓練數據量以及使用時對輸入數據變化的穩健性
- 模型是否適合你的使用情況，及/或是否可以根據你的具體需求進行調整（例如透過微調）
- 調整、解釋和解釋模型輸出的能力（例如用於調試、審計或法規遵循）；使用更簡單、更透明的模型可能比使用更難解釋的大型複雜模型更有優勢
- 訓練數據集的特徵，包括規模、完整性、質量、敏感度、年齡、相關性和多樣性

- 使用模型加固（例如對抗性訓練）、規範化和/或隱私增強技術的價值
- 元件的來源和供應鏈，包括模型或基礎模型、訓練數據和相關工具

要詳細了解其中有多少因素會影響安全結果，請參閱NCSC的「機器學習安全原則」，特別是設計安全（模型架構）。



2. 開發安全

本節包含適用於AI系統開發生命週期的**開發**階段的指引，包括供應鏈安全、文件以及資產和技術債管理。

確保你的供應鏈安全

在系統的整個生命週期中，你都要評估和監控人工智能供應鏈的安全性，並要求供應商遵守你的機構適用於其他軟件的相同標準。如果供應商無法遵守你機構的標準，你將按照現有的風險管理政策行事。

如果不是內部生產，則從經過驗證的商業、開源、和其他第三方開發商處獲取並維護安全可靠，記錄齊全的硬件和軟件元件（例如，模型、數據、軟件庫、模塊、中間件、架構和外部API），以確保系統的安全穩健。

如果不符合安全標準，你可以隨時為關鍵任務系統切換到備用解決方案。你可以使用 NCSC 的供應鏈指引等資源和軟件工件供應鏈層級 (SLSA)[10]等架構來追蹤供應鏈和軟件開發生命週期的證明。

識別、追蹤和保護你的資產

你了解AI相關資產對你機構的價值，包括模型、數據（包括用戶回饋）、提示、軟件、文件、日誌和評估（包括有關潛在不安全功能和故障模式的資訊），並認識到它們在哪些方面具有重大投資意義以及在哪些方面存取會使攻擊者得逞。你將日誌視為敏感數據並實施控制來保護其機密性、完整性和可用性。

你了解你的資產存放在哪裡，評估並接受任何相關風險。你擁有用於追蹤、身份驗證、版本控制和保護資產安全的流程和工具，並且可以在發生洩漏時恢復到已知的良好狀態。

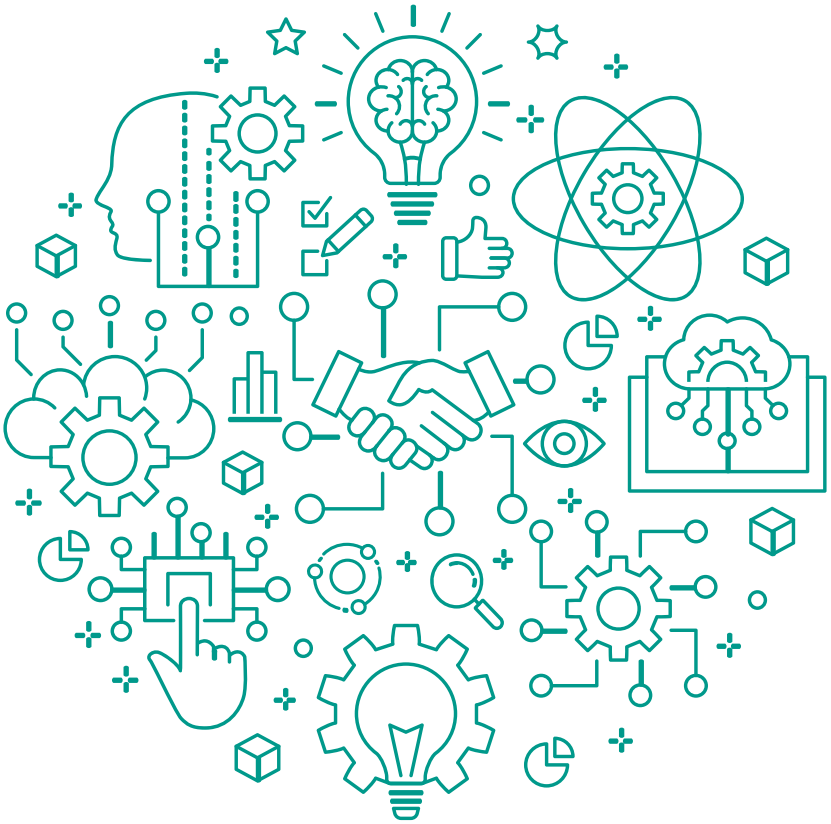
你已制定流程和控制措施來管理AI系統可以存取的數據，並根據AI生成的內容的敏感性（以及因而生成的輸入的敏感性）來進行管理。

記錄你的數據、模型和提示

你記錄任何模型、數據集和元或系統提示的創建、運行和生命週期管理。你的文件包括與安全相關的資訊，例如訓練數據的來源（包括微調數據和人類或其他操作反饋）、預期範圍和限制、防護措施、加密哈希值或簽名、保留時間、建議審查頻率和潛在故障模式。有助於實現這一目標的有用結構包括模型卡、數據卡和軟件物料清單(SBOM)。編制全面的文件有助於提高透明度和問責制[11]。

管理你的技術債務

與任何軟件系統一樣，你可以在AI系統的整個生命週期中識別、追蹤和管理你的「技術債務」（技術債務是指為實現短期結果而做出的不符合最佳實踐的工程決策，以犧牲長期利益為代價）。與金融債務一樣，技術債務本質上並不是壞事，但應該從開發的最早階段開始進行管理[12]。你認識到，在AI環境中這樣做可能比標準軟件更具挑戰性，並且由於開發週期迅速、缺乏完善的協議和接口，技術債務水平可能很高。確保你的生命週期計畫（包括退役AI系統的流程）能夠評估、確認和減輕未來類似系統的風險。



3. 部署安全

本部分包含適用於AI系統開發生命週期的**部署**階段的指引，包括保護基礎設施和模型免受損害、威脅或遺失、制定事故管理流程以及負責任的發布。

確保基礎設施安全

你可以將良好的基礎設施安全原則應用於系統生命週期各個環節所使用的基礎架構。你在研究、開發和部署中對 API、模型和數據及其訓練和處理管道應用適當的存取控制。這包括對保存敏感代碼或數據的環境進行適當隔離。這也將有助於減輕旨在竊取模型或損害其效能的標準網絡安全攻擊。

持續保護你的模型

攻擊者可能透過直接獲取模型（透過取得模型權重）或間接獲取模型（透過應用程式或服務查詢模型），重建模型的功能[13]或其訓練數據[14]。攻擊者也可能在訓練過程中或訓練後篡改模型、數據或提示，使輸出結果不可信。

你可以透過以下方式分別保護模型和數據免遭直接和間接存取：

- 實施標準網絡安全最佳實踐
- 對查詢介面實施控制，以偵測和防止存取、修改和洩露機密資訊的行為

為了確保使用系統可以驗證模型，你可以在模型訓練完成後立即計算並共用模型文件（例如模型權重）和數據集（包括檢查點）的加密哈希值和/或簽章。與加密技術一樣，良好的密鑰管理至關重要[15]。

降低保密風險的方法在很大程度上取決於用例和威脅模型。某些應用程式，例如涉及非常敏感數據的應用程式，可能需要理論上的保證，而這種保證可能難以應用或成本高昂。如果合適，可以使用隱私增強技術（例如差分隱私或同態加密）來探索或確保與消費者、用戶和攻擊者訪問模型和輸出相關的風險等級。

制定事件管理程序

影響你AI系統的安全事件的不可避免性在你的事故響應、升級和補救計劃中反映出來。你的計劃反映了不同的場景，並隨著系統和更廣泛研究的發展定期重新評估。你將重要的公司數碼資源儲存在離線備份中。響應人員接受過評估和解決人工智能相關事件的培訓。你可以免費向客戶和用戶提供高質量的審核日誌和其他安全功能或資訊，以啟用他們的事故響應流程。

負責任地發佈AI 

只有在對模型、應用程式或系統進行適當且有效的安全評估，例如基準測試和紅隊測試（以及超出這些指引範圍的其他測試，例如安全性或公平性測試）之後，你才能發布模型、應用程式或系統，並向你的用戶明確說明已知限製或潛在故障模式。本文件末尾的延伸閱讀部分介紹了開源安全測試庫的詳細資訊。

讓用戶更容易做正確的事情 

你認識到每個新設定或配置選項，都應結合其帶來的業務利益及其引入的任何安全風險進行評估。理想情況下，最安全的設定將作為唯一的選擇整合到系統中。當需要配置時，預設選項應該能夠廣泛地抵禦常見威脅（即預設安全）。應用控制措施，防止以惡意方式使用或部署你的系統。

為用戶提供有關正確使用模型或系統的指導，包括強調限制和潛在故障模式。向用戶明確說明他們對哪些安全方面負責，並且對他們的數據可能在何處（以及如何）使用、存取或儲存（例如，如果用於模型再訓練，或由員工或合作夥伴審查）保持透明。

4. 運維安全

本節包含適用於 AI系統開發生命週期的**安全運維**階段的指引。內容提供了系統部署後尤其相關的行動指引，包括日誌記錄和監控、更新管理和資訊共享。

監控系統的行為

你測量模型和系統的輸出和性能，以便觀察影響安全的行為的突然和逐漸的變化。你可以解釋並識別潛在的入侵和破壞，以及自然數據漂移。

監控系統的輸入

根據隱私和數據保護要求，你監控並記錄系統的輸入（例如推理請求、查詢或提示），以便在出現漏洞或濫用情況時履行合規義務、進行審計、調查和補救措施。這可能包括對分佈外和/或對抗性輸入的明確檢測，包括旨在利用數據準備步驟（例如裁剪和調整圖像大小）的輸入。

遵循設計安全方法進行更新

在每個產品中預設包含自動更新，並使用安全的模組化更新程式發布更新。你的更新過程（包括測試和評估制度）反映了這樣一個事實，即數據、模型或提示的變更可能會導致系統行為的變化（例如，你將重大更新視為新版本）。你支持用戶評估和回應模型變更（例如透過提供預覽存取和版本化 API）。

收集並分享經驗教訓

你參與資訊共享社區，在全球產業、學術界和政府的全球生態系統中進行合作，並酌情分享最佳實踐。你在機構內部和外部保持開放的溝通渠道，以獲取有關系統安全的反饋，包括同意安全研究人員研究和報告漏洞。需要時，你可以將問題升級到更廣泛的社區，例如發布針對漏洞披露的公告，包括詳細且完整的常見漏洞枚舉。你採取行動，快速、適當地緩解和修復問題。

延伸閱讀

AI開發

機器學習安全原則
NCSC關於帶有ML組件的系統開發、部署或運行的詳細指引。

設計安全 - 改變網絡安全風險的平衡：軟件設計安全的原則和方法
該指引由 CISA、NCSC 和其他機構共同撰寫，介紹了包括AI在內的軟件系統製造商應如何採取措施，在產品開發的設計階段就考慮到安全因素，並提供開箱即用的安全產品。

AI安全問題簡述
本文件由德國聯邦資訊安全辦公室 (BSI) 編寫，介紹了機器學習系統可能受到的攻擊以及針對這些攻擊的潛在防禦措施。

機構開發高階AI系統的廣島進程國際指導原則和機構開發高階AI的廣島進程國際行為準則這些文件作為G7 廣島人工智能進程一部分，為開發最先進AI系統的機構提供指導，包括最先進的基礎模型和生成式AI系統，旨在在全球範圍內促進安全、可靠和值得信賴的AI。

AI驗證
新加坡的AI治理測試架構和軟件工具包，透過標準化測試根據一組國際公認的原則驗證 AI系統的性能。

AI良好網絡安全實踐多層架構 - ENISA (europa.eu)
指導國家主管機關和AI利害關係人採取他們所需的步驟，來保護AI系統、操作和流程。

ISO 5338：AI系統生命週期進程（正在審核）
一套流程和相關概念，用於描述基於機器學習和啟發式系統的AI系統生命週期。

AI雲端服務合規標準目錄(AIC4)
BSI 的AI雲端服務合規標準目錄提供了AI的特定標準，可用於評估AI服務在整個生命週期的安全性。

NIST IR 8269（草案）對抗性機器學習的分類和術語
一套流程和相關概念，用於描述基於機器學習和啟發式系統的AI系統生命週期。

MITRE ATLAS
機器學習 (ML) 系統的對抗戰術、技術和案例研究的知識庫，以 MITRE ATT&CK 架構為模型並連結到該架構。

災難性AI風險概述 (2023)
由AI安全中心製作，列出了AI會構成風險的領域。

大型語言模型：產業和當局的機會和風險
BSI 為希望詳細了解開發、部署和/或使用LLM的機會和風險的公司、當局和開發人員編寫的文件。

幫助用戶安全測試AI模型的開源項目包括：

- 對抗穩健性工具箱 (IBM)
- CleverHans（多倫多大學）
- TextAttack（維吉尼亞大學）
- Prompt Bench （微軟）
- Counterfit (微軟)
- AI Verify （新加坡資訊通訊媒體發展局）

網絡安全

CISA的網絡安全績效目標
所有關鍵基礎設施實體都應實施的一組通用保護措施，以有效降低已知風險和對手技術的可能性和影響。

NCSC CAF架構
網絡評估架構 (CAF) 為負責極為重要的服務和活動的機構提供指導。

MITRE的供應鏈安全架構
用於評估供應鏈內的供應商和服務提供商的架構。

風險管理

NIST AI風險管理架構 (AIRMF)
AIRMF概述了如何管理與AI相關的個人、機構和社會的社會技術風險。

ISO 27001：資訊安全、網絡安全和隱私保護
此標準為機構提供建立、實施和維護資訊安全管理系統的指導。

ISO 31000：風險管理
一項國際標準，為機構提供機構內部風險管理的指引與原則。

NCSC風險管理指引
此指引可協助網絡安全風險從業人員更了解並管理影響其機構的網絡安全風險。

注釋

1. 此處定義為開發AI系統（或已開發了AI系統）並以自己的名稱或商標將該系統投放市場或投入使用的個人、公共機關、機構或其他團體
2. 有關安全設計的詳細資訊，請參閱 CISA 的設計安全網頁和指引 改變網絡安全風險的平衡：設計安全軟件的原則與方法
3. 與基於規則的系統等非機器學習AI方法相反
4. CEPS 在其出版品「協調AI價值鏈與歐盟人工智能法案」中描述了七種不同類型的AI開發互動
5. ISO/IEC 22989:2022(en) 將其定義為「建構AI系統的功能要元」
6. NIST的任務是製定指引（並採取其他行動）以促進安全、可靠和值得信賴的人工智能 (AI) 開發和使用。請參閱2023年10月30日行政命令規定的 NIST的職責
7. 有關威脅建模的更多資訊，請參閱OWASP基金會所載資訊
8. 請參閱 MITRE ATLAS 對抗性機器學習101
9. GitHub：使用惡意Lambda層的Tensorflow的RCE PoC
10. SLSA：「保障所有軟件供應鏈中工件的完整性」
11. METI（日本經濟產業省，2023年），「軟件管理中軟件物料清單 (SBOM) 引入指引」
12. 谷歌研究：機器學習：技術債高息信用卡
13. Tramèr et al 2016，透過預測API竊取機器學習模型
14. Boenisch，2020，針對機器學習隱私的攻擊（第 1 部分）：使用IBM-ART架構的模型反轉攻擊
15. 國家網絡安全中心，2020 年，設計並構建私人託管的公鑰基礎設施