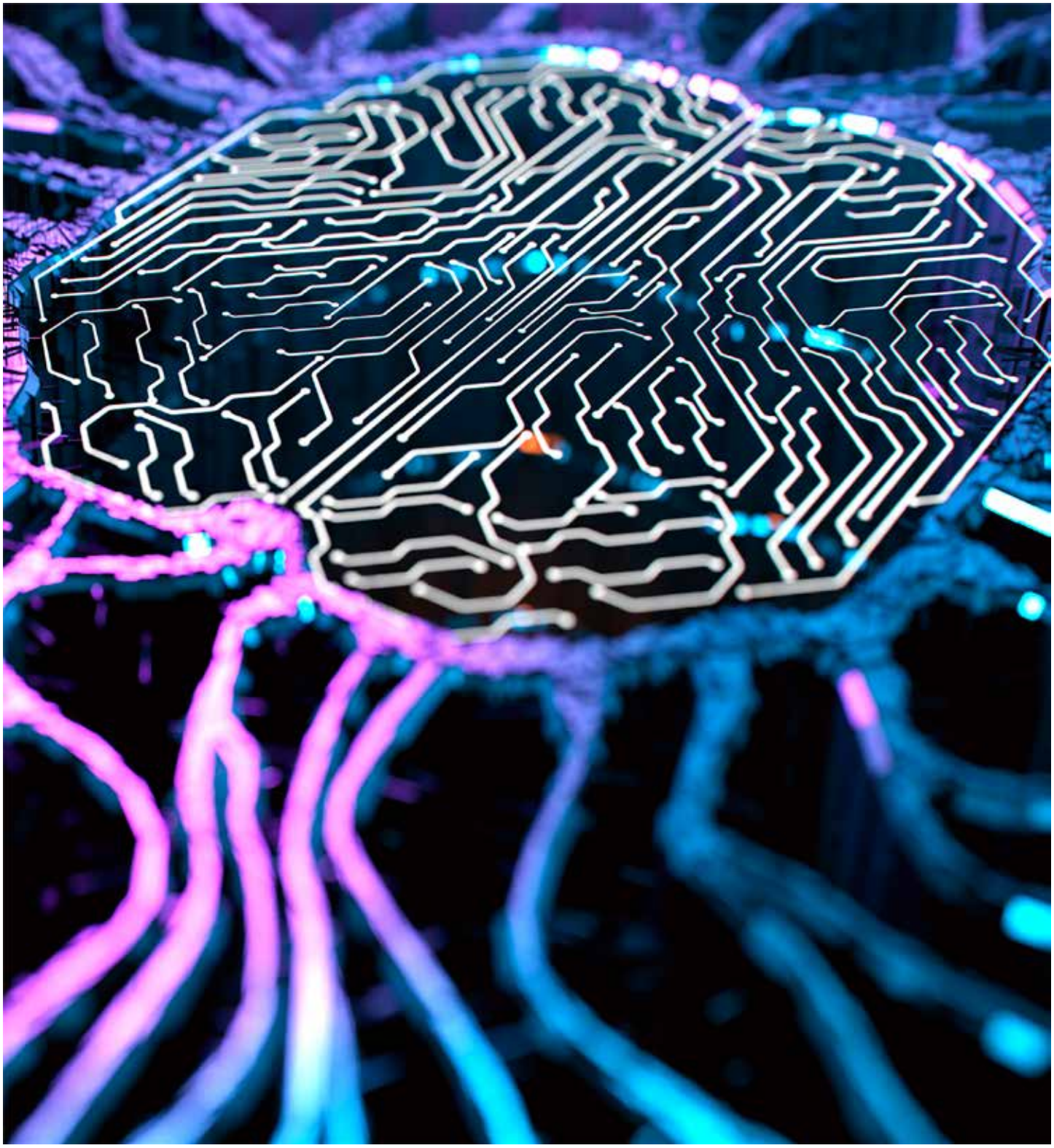
Hướng dẫn phát triển hệ thống AI (Trí tuệ Nhân tạo) an toàn



National Cyber Security Centre
a part of GCHQ

CYBERSECURITY & INFRASTRUCTURE SECURITY AGENCY
CISA

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

Về tài liệu này

Tài liệu này được xuất bản bởi Trung tâm An ninh Mạng Quốc gia Vương quốc Anh (NCSC), Cơ quan An ninh Hạ tầng Cơ sở và An ninh Mạng Hoa Kỳ (CISA) và các đối tác quốc tế sau:

- Cơ quan An ninh Quốc gia (NSA)
- Cục Điều tra Liên bang (FBI)
- Trung tâm An ninh Mạng Úc (ACSC) thuộc Tổng cục Tín hiệu Úc
- Trung tâm An ninh Mạng Gia Nã Đại (CCCS)
- Trung tâm An ninh Mạng Quốc gia Tân Tây Lan (NCSC-NZ)
- Đội Ứng phó Vấn đề Bảo mật Máy vi tính (CSIRT) của Chính phủ Chile
- Cơ quan An ninh và Thông tin Mạng Quốc gia của Cộng hòa Séc (NUKIB)
- Cơ quan Quản lý Hệ thống Thông tin Estonia (RIA) và Trung tâm An ninh Mạng Quốc gia Estonia (NCSC-EE)
- Cơ quan An ninh Mạng Pháp (ANSSI)
- Cục An ninh Mạng Liên bang Đức (BSI)
- Tổng cục Mạng Quốc gia Do thái (INCD)
- Cơ quan An ninh Mạng Quốc gia Ý Đại Lợi (ACN)
- Trung tâm Quốc gia Nhật Bản về Trạng thái Sẵn sàng và Chiến lược Ứng phó với Vấn đề An ninh Mạng(NISC)
- Ban Thư ký Chính sách Khoa học, Công nghệ và Đổi mới, Văn phòng Nội các Nhật Bản
- Cơ quan Phát triển Công nghệ Thông tin Quốc gia Nigeria (NITDA)
- Trung tâm An ninh Mạng Quốc gia Na Uy (NCSC-NO)
- Bộ Thông tin và Truyền thông Kỹ thuật Số của Ba Lan
- Viện Nghiên cứu Quốc gia NASK của Ba Lan (NASK)
- Cơ quan Tình báo Quốc gia Hàn Quốc (NIS)
- Cơ quan An ninh Mạng Tân Gia Ba (CSA)

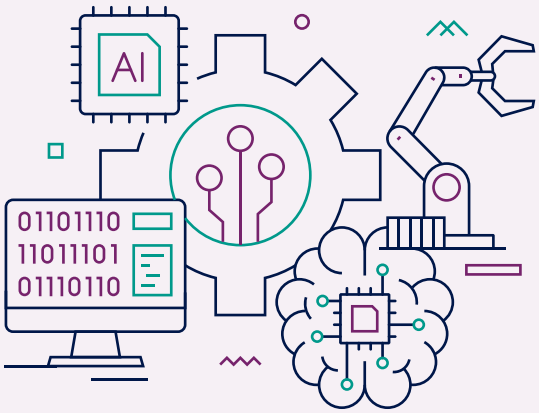
Lời cám ơn

Các tổ chức có tên sau đây đã đóng góp vào việc soạn thảo các hướng dẫn này:

- Viện Alan Turing
- Anthropic
- Databricks
- Trung tâm An ninh và Công nghệ Đang Phát triển Rộng rải của Đại học Georgetown
- Google
- Google DeepMind
- IBM
- ImBue
- Microsoft
- OpenAI
- Palantir
- RAND
- Scale AI
- Viện Kỹ thuật Phần mềm tại Đại học Carnegie Mellon
- Trung tâm An toàn cho AI (Trí tuệ Nhân tạo) Stanford
- Chương trình Stanford về Địa lý Chính trị, Công nghệ và Quản trị

Tuyên bố miễn trừ trách nhiệm

Mục lục



Bản tóm tắt chính

Tài liệu này khuyến nghị các hướng dẫn dành cho nhà cung cấp của bất kỳ hệ thống nào sử dụng trí tuệ nhân tạo (AI), cho dù các hệ thống đó được tạo ra từ đầu hay được xây dựng bổ sung vào các công cụ và dịch vụ do người khác cung cấp. Thực hiện theo những hướng dẫn này sẽ giúp các nhà cung cấp xây dựng các hệ thống AI hoạt động như dự kiến, sẵn có khi cần, và hoạt động mà không tiết lộ dữ liệu nhạy cảm cho bất kỳ bên nào không được ủy quyền.

Tài liệu này chủ yếu nhắm đến những nhà cung cấp hệ thống AI sử dụng các mô hình của một tổ chức nào đó hoặc sử dụng các giao diện lập trình ứng dụng (application programming interfaces - API) bên ngoài. Chúng tôi kêu gọi **tất cả** các bên liên quan (bao gồm các nhà khoa học dữ liệu, nhà phát triển, người quản lý, người ra quyết định và chủ sở hữu rủi ro) (Chủ sở hữu rủi ro là người có trách nhiệm và quyền quản lý rủi ro, bao gồm việc hiểu rõ các biện pháp kiểm soát là gì và mức độ hiệu quả của các biện pháp này để làm thay đổi rủi ro) đọc các nguyên tắc này để giúp họ đưa ra các quyết định sáng suốt về **thiết kế**, **phát triển**, **triển khai** và **vận hành** hệ thống AI của họ.

Về các hướng dẫn

Hệ thống AI có tiềm năng mang lại nhiều lợi ích cho xã hội. Tuy nhiên, để cơ hội cho AI được hiện thực hóa đầy đủ, nó phải được phát triển, triển khai và vận hành theo cách an toàn và có trách nhiệm.

Các hệ thống AI phải đối mặt với các lỗ hổng bảo mật mới mẻ cần được xem xét cùng với các mối đe dọa an ninh mạng thường gặp. Khi tốc độ phát triển cao - như trường hợp của AI - vấn đề bảo mật thường trở thành mối quan tâm thứ yếu. Bảo mật phải là một yêu cầu cốt lõi, không chỉ trong giai đoạn phát triển, mà trong suốt vòng đời của hệ thống.

Vì lý do này, các hướng dẫn được chia thành bốn lĩnh vực chính trong vòng đời phát triển hệ thống AI: **thiết kế an toàn**, **phát triển an toàn**, **triển khai an toàn**, và **vận hành và bảo trì an toàn**. Đối với mỗi phần, chúng tôi gợi ý những cân nhắc và biện pháp giảm thiểu mà sẽ giúp giảm rủi ro tổng thể đối với quá trình phát triển hệ thống AI của tổ chức.

1. **Thiết kế an toàn**
   Phần này chứa đựng các hướng dẫn áp dụng cho giai đoạn thiết kế của vòng đời phát triển hệ thống AI. Bao gồm việc hiểu biết các rủi ro và thành lập mô hình mối đe dọa, cũng như các chủ đề cụ thể và những đánh đổi để cân nhắc khi thiết kế hệ thống và mô hình.

2. **Phát triển an toàn**
   Phần này chứa đựng các hướng dẫn áp dụng cho giai đoạn phát triển của vòng đời phát triển hệ thống AI, bao gồm bảo mật chuỗi cung ứng, tài liệu hóa, cũng như quản lý tài sản và nợ kỹ thuật.

3. **Triển khai an toàn**
   Phần này chứa đựng các hướng dẫn áp dụng cho giai đoạn triển khai của vòng đời phát triển hệ thống AI, bao gồm bảo vệ hạ tầng cơ sở và mô hình khỏi bị xâm phạm, đe dọa hoặc mất mát, phát triển quy trình quản lý vấn đề và phát hành có trách nhiệm.

4. **Vận hành và bảo trì an toàn**
   Phần này chứa đựng các hướng dẫn áp dụng cho giai đoạn vận hành và bảo trì an toàn trong vòng đời phát triển hệ thống AI (Trí tuệ Nhân tạo). Nó cung cấp các hướng dẫn về các hành động có liên quan đặc biệt sau khi hệ thống đã được triển khai, bao gồm nhật ký ghi chép và theo dõi, quản lý cập nhật và chia sẻ thông tin.

Các hướng dẫn này tuân thủ phương pháp 'bảo mật theo mặc định' và được đồng bộ chặt chẽ với các thực hành được định nghĩa trong Hướng dẫn phát triển và triển khai an toàn của NCSC, Khuôn khổ Phát triển Phần mềm An toàn của NIST, và 'nguyên tắc bảo mật theo thiết kế' do CISA, NCSC và các cơ quan mạng quốc tế xuất bản. Các hướng dẫn này ưu tiên:

- việc chịu trách nhiệm về các kết quả về an ninh của khách hàng
- chấp nhận tính minh bạch và trách nhiệm cấp tiến
- xây dựng cấu trúc và thành phần lãnh đạo của tổ chức để bảo mật theo thiết kế là ưu tiên hàng đầu trong kinh doanh

Phần giới thiệu

Hệ thống trí tuệ nhân tạo (AI) có tiềm năng mang lại nhiều lợi ích cho xã hội. Tuy nhiên, để cơ hội cho AI được hiện thực hóa đầy đủ, nó phải được phát triển, triển khai và vận hành theo cách an toàn và có trách nhiệm. An ninh mạng là điều kiện tối quan trọng cần thiết cho sự an toàn, khả năng thích ứng, quyền riêng tư, công bằng, hiệu quả và độ tin cậy của hệ thống AI.

Tuy nhiên, các hệ thống AI phải đối mặt với các lỗ hổng bảo mật mới mẻ cần được xem xét cùng với các mối đe dọa an ninh mạng thường gặp. Khi tốc độ phát triển cao - như trong trường hợp của AI - bảo mật thường trở thành mối quan tâm thứ yếu. Bảo mật phải là một yêu cầu cốt lõi, không chỉ trong giai đoạn phát triển, mà cả trong suốt vòng đời của hệ thống.

Tài liệu này khuyến nghị các hướng dẫn dành cho các nhà cung cấp[1] của bất kỳ hệ thống nào sử dụng AI, cho dù các hệ thống đó được tạo ra từ ban đầu hay được xây dựng bổ sung vào các công cụ và dịch vụ do nơi khác cung cấp. Việc thực hiện những hướng dẫn này sẽ giúp nhà cung cấp xây dựng các hệ thống AI hoạt động như dự kiến, sẵn có khi cần, và hoạt động mà không tiết lộ dữ liệu nhạy cảm cho các bên không được ủy quyền.

Các hướng dẫn này nên được xem xét kết hợp với các tiêu chuẩn thực hành an ninh mạng, quản lý rủi ro, và ứng phó vấn đề đã được thành lập. Đặc biệt, chúng tôi kêu gọi các nhà cung cấp tuân theo nguyên tắc các 'bảo mật theo thiết kế' [2] do Cơ quan An ninh Hạ tầng Cơ sở và An ninh Mạng Hoa Kỳ (CISA), Trung tâm An ninh Mạng Quốc gia Vương quốc Anh (NCSC) và tất cả đối tác quốc tế của chúng tôi soạn thảo. Các nguyên tắc ưu tiên:

- chịu trách nhiệm về các kết quả về an ninh của khách hàng
- chấp nhận tính minh bạch và trách nhiệm cấp tiến
- xây dựng cấu trúc và thành phần lãnh đạo của tổ chức để bảo mật theo thiết kế là ưu tiên hàng đầu trong kinh doanh

Việc tuân theo các nguyên tắc 'bảo mật theo thiết kế' đòi hỏi nguồn lực đáng kể trong suốt vòng đời của hệ thống. Điều đó nghĩa là các nhà phát triển phải đầu tư vào việc ưu tiên **tính năng**, **cơ chế**, và **triển khai** các công cụ bảo vệ khách hàng ở mỗi tầng của quá trình thiết kế hệ thống và trong tất cả các giai đoạn của vòng đời phát triển. Thực hành điều này sẽ tránh việc thiết kế lại gây tốn kém sau này, đồng thời bảo vệ khách hàng và dữ liệu của họ trong thời gian gần.

Tại sao vấn đề anh ninh của AI lại khác biệt?

Trong tài liệu này, chúng tôi sử dụng 2 mẫu tự 'AI' để đề cập cụ thể đến các ứng dụng học máy (machine learning - ML)[3]. Tất cả các loại ML đều nằm trong phạm vi. Chúng tôi định nghĩa ứng dụng ML như là những ứng dụng:

- liên quan đến các thành phần phần mềm (mô hình) cho phép máy vi tính nhận diện và đưa ra ngữ cảnh vào các khuôn mẫu trong dữ liệu mà không có các quy tắc phải do con người viết lập trình
- tạo ra các dự đoán, khuyến nghị hoặc quyết định dựa trên lý luận thống kê

Cũng như các mối đe dọa an ninh mạng hiện có, các hệ thống AI còn phải đối mặt với các loại lỗ hổng bảo mật mới. Thuật ngữ 'đối nghịch học máy' (AML), được sử dụng để mô tả việc lợi dụng các lỗ hổng bảo mật cơ bản trong các thành phần ML, bao gồm phần cứng, phần mềm, quy trình công việc và chuỗi cung ứng. AML cho phép kẻ tấn công gây ra các hoạt động ngoài ý muốn trong hệ thống ML, có thể bao gồm:

- ảnh hưởng đến khả năng phân loại hoặc đảo ngược hiệu suất của mô hình
- cho phép người dùng thực hiện các hành động trái phép
- lấy đi các thông tin nhạy cảm về mô hình

Có nhiều cách để đạt được những hiệu ứng này, ví dụ như tấn công tiêm chủng mẫu trong domain mô hình ngôn ngữ lớn (Large language model - LLM) (LLM là một nhóm các máy tính Windows kết nối với nhau, chia sẻ thông tin tài khoản người dùng và một chính sách bảo mật) hoặc cố ý làm hỏng dữ liệu đào tạo hoặc phản hồi từ người dùng (còn được gọi là 'ngộ độc dữ liệu').

Ai nên đọc tài liệu này?

Tài liệu này chủ yếu nhắm đến các nhà cung cấp hệ thống AI, bất kể là hệ thống đó dựa trên dựa trên các mô hình của một tổ chức nào đó hay sử dụng các giao diện lập trình ứng dụng (API) bên ngoài. Tuy nhiên, chúng tôi kêu gọi **tất cả** các bên liên quan (bao gồm các nhà khoa học dữ liệu, nhà phát triển, người quản lý, người ra quyết định và chủ sở hữu rủi ro(Chủ sở hữu rủi ro - là người có trách nhiệm và quyền quản lý rủi ro, bao gồm việc hiểu rõ các biện pháp kiểm soát là gì và mức độ hiệu quả của các biện pháp này để làm thay đổi rủi ro) đọc các hướng dẫn này để giúp họ đưa ra các quyết định sáng suốt về **thiết kế**, **triển khai** và **vận hành** của hệ thống AI học máy của họ.

Tuy nhiên, không phải tất cả các hướng dẫn đều có thể áp dụng trực tiếp được cho tất cả các tổ chức. Mức độ tinh xảo và các phương pháp tấn công sẽ khác nhau tùy thuộc vào đối thủ nhắm vào hệ thống AI, do đó, các hướng dẫn nên được xem xét cùng với các trường hợp sử dụng và hồ sơ khái quát về các mối đe dọa đối với tổ chức của quý vị.

Ai chịu trách nhiệm về việc phát triển AI an toàn?

Thường có nhiều người tham gia trong chuỗi cung ứng AI hiện đại. Một phương pháp đơn giản giả định có hai thực thể:

- ‘nhà cung cấp’ chịu trách nhiệm quản lý dữ liệu, phát triển thuật toán, thiết kế, triển khai và bảo trì
- ‘người dùng’, người cung cấp đầu vào và nhận đầu ra

Mặc dù phương pháp nhà cung cấp-người dùng này được sử dụng trong nhiều ứng dụng, nhưng nó ngày càng trở nên kém phổ biến[4], vì các nhà cung cấp có thể tìm cách kết hợp phần mềm, dữ liệu, mô hình và/hoặc dịch vụ từ xa do bên thứ ba cung cấp vào hệ thống của riêng họ. Những chuỗi cung ứng phức tạp này khiến người sử dụng sản phẩm khó hiểu được trách nhiệm về AI an toàn nằm ở đâu.

Người dùng (cho dù là ‘người sử dụng sản phẩm’, hay nhà cung cấp kết hợp thành phần AI bên ngoài[5]) thường không có đủ tầm nhìn và/hoặc kiến thức chuyên môn để hiểu, đánh giá hoặc giải quyết các rủi ro liên quan đến các hệ thống họ đang sử dụng. Do đó, theo nguyên tắc ‘bảo mật theo thiết kế’, **các nhà cung cấp các thành phần AI nên chịu trách nhiệm về kết quả bảo mật của người dùng ở phía dưới của chuỗi cung ứng.**

Nhà cung cấp nên thực hiện các biện pháp kiểm soát và giảm thiểu rủi ro trong các mô hình, đường ống dữ liệu và/hoặc hệ thống của họ khi có thể, đồng thời khi sử dụng các chế độ cài đặt, hãy thực hiện tùy chọn an toàn nhất theo mặc định. Khi không thể giảm thiểu rủi ro, nhà cung cấp phải chịu trách nhiệm về việc:

- thông báo cho người sử dụng ở phía dưới của chuỗi cung ứng về những rủi ro mà nhà cung cấp và (nếu có) người sử dụng của họ đang chấp nhận
- cố vấn cho họ cách sử dụng thành phần một cách an toàn

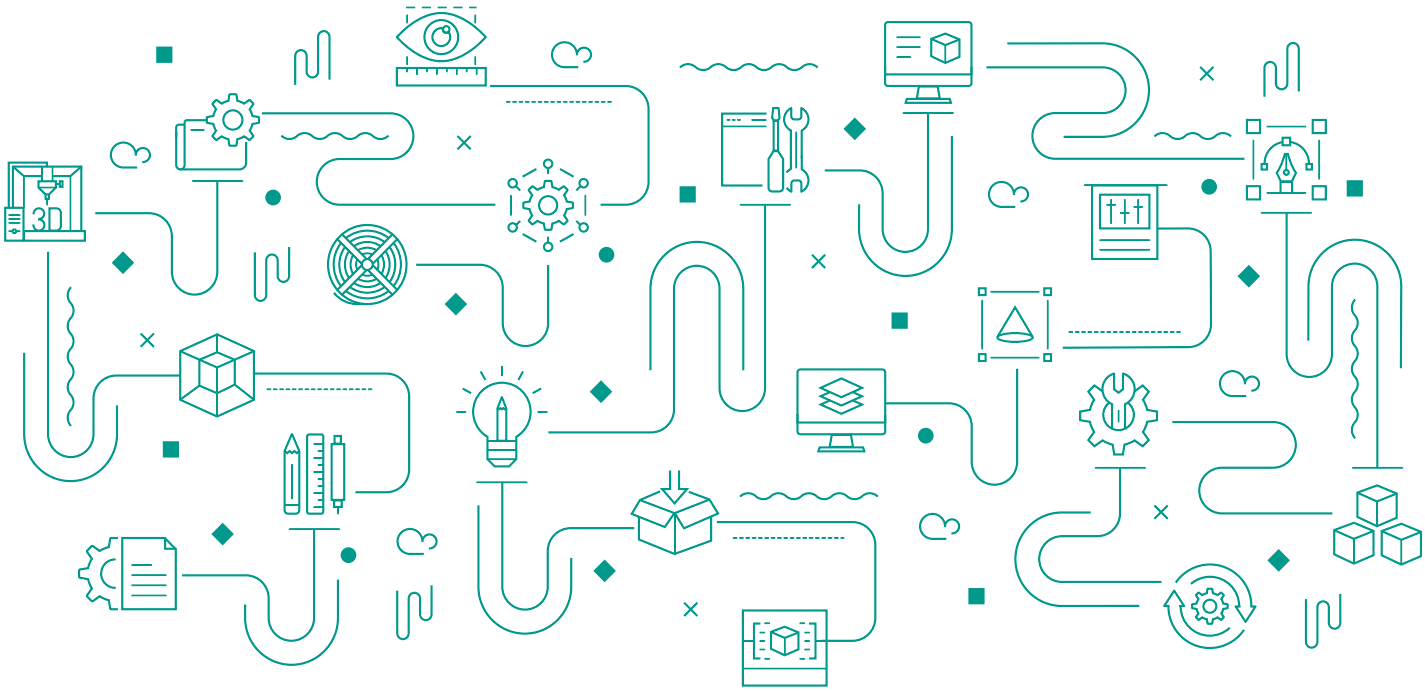
Khi việc xâm phạm hệ thống có thể dẫn đến thiệt hại vật chất hoặc danh tiếng trên phương diện rộng lớn, tổn thất đáng kể về hoạt động kinh doanh, rò rỉ thông tin nhạy cảm hoặc bí mật và/hoặc có hậu quả về pháp lý, thì các rủi ro an ninh mạng của AI phải được coi là **nghiêm trọng.**

Các hướng dẫn để phát triển hệ thống AI an toàn

Các hướng dẫn được chia thành bốn lĩnh vực chính trong vòng đời phát triển hệ thống AI: **thiết kế an toàn**, **phát triển an toàn**, **triển khai an toàn** và **vận hành và bảo trì an toàn**. Đối với từng lĩnh vực, chúng tôi gợi ý những cân nhắc và các biện pháp giảm thiểu mà sẽ giúp giảm rủi ro tổng thể đến quá trình phát triển hệ thống AI của tổ chức.

Các hướng dẫn được nêu ra trong tài liệu này tuân thủ chặt chẽ với các phương pháp về vòng đời phát triển phần mềm được định nghĩa trong:

- hướng dẫn triển khai và phát triển an toàn của NCSC
- viện Tiêu chuẩn và Công nghệ Quốc gia (NIST) Khuôn khổ Phát triển Phần mềm An toàn (SSDF)[6]



1. Thiết kế an toàn

Phần này chứa đựng các hướng dẫn áp dụng cho giai đoạn **thiết kế** của vòng đời phát triển hệ thống AI. Nó bao gồm việc hiểu biết các rõ rủi ro và mô hình mối đe dọa, cũng như các chủ đề cụ thể và sự đánh đổi cần xem xét trong quá trình thiết kế hệ thống và mô hình.

Nâng cao nhận thức của nhân viên về các mối đe dọa và rủi ro

Chủ nhân sở hữu hệ thống và lãnh đạo cấp cao hiểu rõ về các mối đe dọa đến an ninh AI và các biện pháp giảm thiểu chúng. Các nhà khoa học và nhà phát triển dữ liệu của quý vị duy trì nhận thức về các mối đe dọa bảo mật có liên quan và các chế độ lỗi, đồng thời giúp chủ sở hữu rủi ro đưa ra quyết định trên cơ sở có đầy đủ thông tin. Quý vị cung cấp hướng dẫn cho người dùng về những rủi ro bảo mật đặc thù mà hệ thống AI phải đối mặt (ví dụ, như một phần của chương trình đào tạo InfoSec tiêu chuẩn) và đào tạo những nhà phát triển về các kỹ thuật viết mã an toàn cũng như các phương pháp thực hành an toàn và có trách nhiệm trong lĩnh vực AI.

Lập mô hình các mối đe dọa đối với hệ thống của quý vị

Một phần của quy trình quản lý rủi ro của mình là, quý vị áp dụng quy trình toàn diện để thẩm định các mối đe dọa đối với hệ thống của mình, bao gồm việc hiểu các tác động tiềm ẩn đối với hệ thống, người dùng, tổ chức và xã hội rộng hơn nếu một thành phần AI bị xâm phạm hoặc hoạt động không như mong muốn[7]. Quá trình này bao gồm việc thẩm định tác động của các mối đe dọa riêng của AI[8] và ghi chép thành tài liệu quá trình đưa ra quyết định của mình.

Quý vị nhận ra rằng mức độ nhạy cảm và các loại dữ liệu được sử dụng trong hệ thống của quý vị có thể ảnh hưởng đến giá trị của nó và trở thành mục tiêu cho kẻ tấn công. Thẩm định của quý vị nên xem xét rằng một số mối đe dọa có thể gia tăng khi hệ thống AI ngày càng được coi là mục tiêu có giá trị cao và khi chính AI tạo ra các phương pháp tấn công tự động, mới.

Thiết kế hệ thống của quý vị không chỉ để bảo đảm tính bảo mật mà còn để đạt được chức năng và hiệu suất

Quý vị tự tin rằng những công việc hiện tại được giải quyết một cách thích hợp nhất bằng AI. Sau khi xác định được điều này, quý vị đánh giá tính thích hợp của các lựa chọn thiết kế cụ thể dành riêng cho AI của mình. Quý vị xem xét mô hình mối đe dọa của mình và các biện pháp giảm thiểu rủi ro bảo mật liên quan cùng với chức năng, trải nghiệm của người dùng, môi trường triển khai, hiệu suất, sự bảo đảm, giám sát, các yêu cầu về đạo đức và pháp lý, cùng với những cân nhắc khác. Ví dụ:

- quý vị cân nhắc tính bảo mật của chuỗi cung ứng khi lựa chọn để phát triển trong nội bộ hay sử dụng các thành phần bên ngoài, ví dụ như:
    - lựa chọn của quý vị về việc đào tạo mô hình mới, sử dụng một mô hình hiện tại (cần hoặc không cần vi chỉnh) hoặc truy cập một mô hình qua một API bên ngoài, là thích hợp với các yêu cầu của quý vị
    - lựa chọn của quý vị để làm việc với một nhà cung cấp mô hình bên ngoài bao gồm đánh giá thẩm định về tình hình bảo mật của chính nhà cung cấp đó
    - nếu sử dụng thư viện bên ngoài, quý vị cần thực hiện đánh giá thẩm định (ví dụ: để bảo đảm thư viện có các biện pháp kiểm soát ngăn chặn hệ thống tải các mô hình không đáng tin cậy mà không gặp nguy hiểm ngay lập tức bởi 'thực thi mã tùy ý' (thực thi mã tùy ý là khả năng kẻ tấn công chạy bất kỳ lệnh hoặc mã nào mà chúng lựa chọn trên máy hoặc trong quy trình được nhắm vào) [9])
    - quý vị thực hiện chức năng quét và cách ly/hộp cát (sandboxing – là một cơ chế bảo mật để tách các chương trình đang chạy, thường nhằm nỗ lực giảm thiểu lỗi hệ thống và/hoặc lỗ hổng phần mềm lây lan) khi nhập mô hình của bên thứ ba hoặc trọng số được tuần tự hóa. Những mã này nên được coi là mã của bên thứ ba không đáng tin cậy và có thể cho phép thực hiện mã từ xa

- nếu sử dụng API bên ngoài, quý vị áp dụng các biện pháp kiểm soát thích hợp đối với dữ liệu có thể được gửi đến các dịch vụ nằm ngoài tầm kiểm soát của tổ chức của quý vị, chẳng hạn như yêu cầu người dùng đăng nhập và xác nhận trước khi gởi đi các thông tin có thể mang tích nhạy cảm
- quý vị áp dụng các biện pháp kiểm tra thích hợp và vệ sinh dữ liệu (vệ sinh dữ liệu bao gồm việc xóa an toàn và vĩnh viễn dữ liệu nhạy cảm khỏi bộ dữ liệu và phương tiện để bảo đảm rằng không có dữ liệu nào còn sót lại có thể được phục hồi ngay cả khi thông qua phân tích) cũng như đầu vào; việc này bao gồm khi kết hợp phản hồi của người dùng hoặc dữ liệu học tập liên tục vào mô hình của quý vị và hiểu rằng dữ liệu đào tạo xác định hoạt động của hệ thống

- quý vị tích hợp việc phát triển hệ thống phần mềm AI vào các phương pháp hay nhất về vận hành và phát triển an toàn hiện có; tất cả các thành phần của hệ thống AI đều được viết trong các môi trường thích hợp bằng cách sử dụng các phương pháp viết mã và các ngôn ngữ nhằm giảm bớt hoặc loại bỏ các loại lỗ hổng bảo mật đã được biết những khi hợp lý

- nếu các thành phần AI cần kích hoạt hành động, ví dụ như sửa đổi các tập hồ sơ hoặc chuyển hướng đầu ra sang các hệ thống bên ngoài, quý vị áp dụng các hạn chế thích hợp đối với các hành động có thể xảy ra (điều này bao gồm AI bên ngoài và các biện pháp bảo đảm an toàn không phải là AI nếu cần thiết)

- các quyết định xung quanh sự tương tác của người dùng được đưa ra trên cơ sở có đầy đủ thông tin về các rủi ro cụ thể của AI, ví dụ như:

  - hệ thống của quý vị cung cấp cho người dùng các đầu ra có thể sử dụng được mà không tiết lộ các chi tiết không cần thiết cho kẻ tấn công tiềm ẩn
  - nếu cần, hệ thống của quý vị sẽ cung cấp các biện pháp bảo vệ hiệu quả xung quanh đầu ra của mô hình
  - trong trường hợp cung cấp API cho khách hàng hoặc cộng tác viên bên ngoài, quý vị áp dụng các biện pháp kiểm soát phù hợp để giảm thiểu các cuộc tấn công vào hệ thống AI qua API
  - quý vị tích hợp các chế độ cài đặt an toàn nhất vào hệ thống theo mặc định
  - quý vị áp dụng nguyên tắc đặc quyền tối thiểu để giới hạn quyền truy cập vào các chức năng của hệ thống
  - quý vị giải thích các khả năng đi kèm với nhiều rủi ro cho người dùng và yêu cầu người dùng chọn tham gia để sử dụng chúng; quý vị nêu ra các trường hợp sử dụng bị cấm và những khi có thể, thông báo cho người dùng về các giải pháp thay thế

Cân nhắc những lợi ích về bảo mật và các đánh đổi khi lựa chọn mô hình AI của quý vị

Lựa chọn của quý vị về mô hình AI sẽ bao gồm việc cân nhắc một loạt các yêu cầu khác nhau. Điều này bao gồm việc chọn lựa kiến trúc mô hình, cấu hình, dữ liệu đào tạo, thuật toán đào tạo và siêu tham số (Siêu tham số là các biến cấu hình bên ngoài mà các nhà khoa học dữ liệu sử dụng để quản lý quá trình đào tạo mô hình học máy). Các quyết định của quý vị được dựa trên mô hình mối đe dọa của quý vị và thường xuyên được đánh giá lại khi nghiên cứu về bảo mật AI tiến bộ và sự hiểu biết về mối đe dọa phát triển.

Khi chọn mô hình AI, những cân nhắc của quý vị có thể sẽ bao gồm, nhưng không giới hạn ở:

- mức độ phức tạp của mô hình quý vị đang sử dụng, tức là kiến trúc và số lượng tham số được chọn; kiến trúc đã chọn và số lượng tham số của mô hình của quý vị, cùng với các yếu tố khác, sẽ ảnh hưởng đến số lượng dữ liệu đào tạo mà mô hình đòi hỏi và mức độ mạnh mẽ của nó đối với những thay đổi về dữ liệu đầu vào khi đang sử dụng

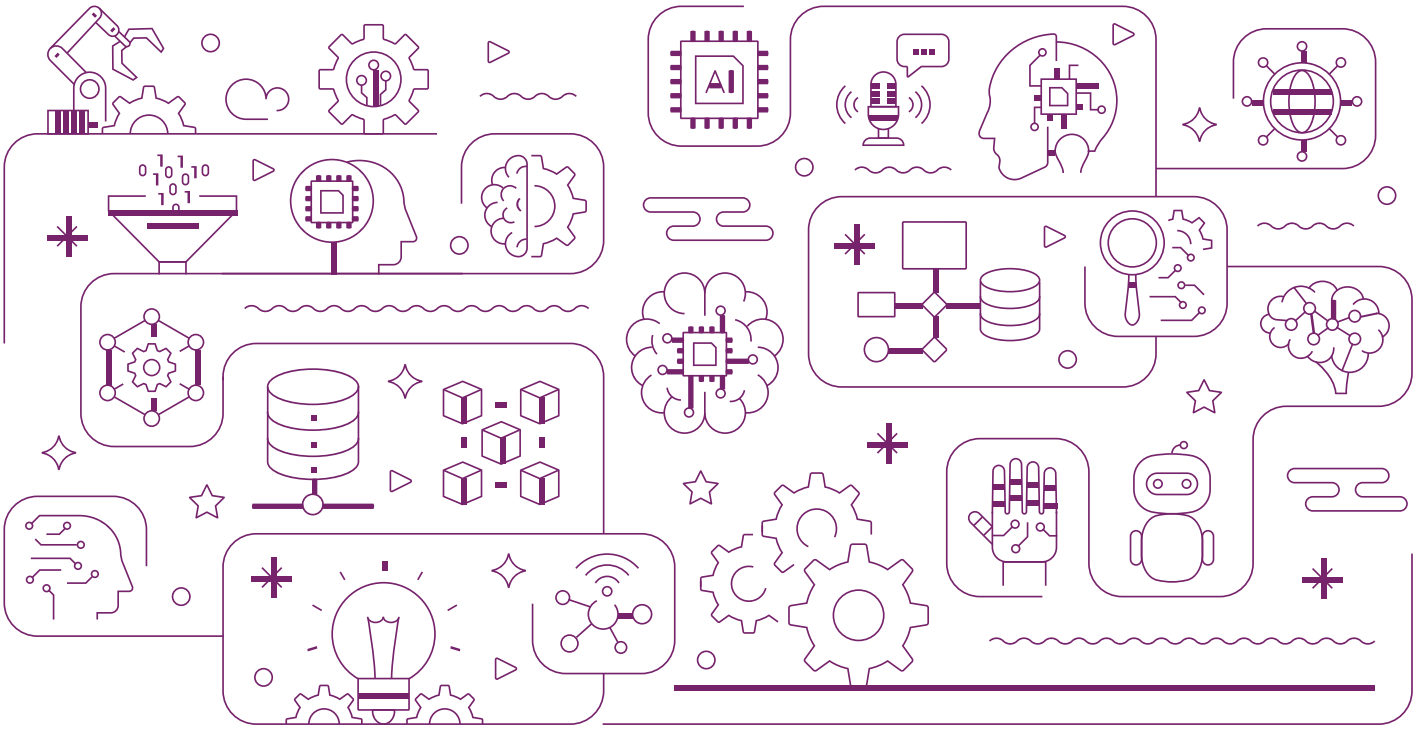
- tính phù hợp của mô hình đối với trường hợp sử dụng cụ thể của quý vị và/hoặc khả năng điều chỉnh cho nhu cầu cụ thể của quý vị (ví dụ như bằng cách vi chỉnh)

- khả năng điều chỉnh, diễn giải và giải thích các kết quả đầu ra của mô hình của quý vị (ví dụ như cho mục đích tìm và sửa lỗi,
kiểm toán hoặc tuân thủ quy định); có thể có những lợi ích khi sử dụng các mô hình đơn giản, rõ ràng hơn so với những mô hình lớn và phức tạp mà lại khó diễn giải

- đặc điểm của bộ dữ liệu đào tạo, bao gồm kích cỡ, tính toàn vẹn, phẩm chất, mức độ nhạy cảm, độ tuổi, sự tính liên quan và tính đa dạng

- giá trị của việc sử dụng kỹ thuật tăng cường mô hình (chẳng hạn như đào tạo để chống chọi với đối nghịch), chính quy hóa và/hoặc các kỹ thuật tăng cường quyền riêng tư
- nguồn gốc và chuỗi cung ứng của các thành phần bao gồm mô hình hoặc mô hình nền tảng, dữ liệu đào tạo và các công cụ liên quan

Để biết thêm thông tin về bao nhiêu trong số lượng các yếu tố này ảnh hưởng đến kết quả bảo mật, hãy tham khảo 'Các Nguyên tắc Bảo mật của Học Máy' của NCSC, đặc biệt là Thiết kế cho Bảo mật (kiến trúc mô hình).



2. Phát triển an toàn

Phần này chứa đựng các hướng dẫn áp dụng cho giai đoạn **phát triển** của vòng đời phát triển hệ thống AI, bao gồm an ninh của chuỗi cung ứng, tài liệu hóa, cũng như quản lý tài sản và nợ kỹ thuật.

Bảo vệ chuỗi cung ứng của quý vị

Quý vị đánh giá và theo dõi tính bảo mật của chuỗi cung ứng AI trong suốt vòng đời của hệ thống và yêu cầu các nhà cung cấp tuân thủ các tiêu chuẩn tương tự mà tổ chức của quý vị áp dụng với phần mềm khác. Nếu những nhà cung cấp không thể tuân thủ các tiêu chuẩn của tổ chức của quý vị, quý vị sẽ hành động tuân theo các chính sách quản lý rủi ro hiện tại của mình.

Trong trường hợp không được sản xuất nội bộ, quý vị thu thập và duy trì các thành phần phần cứng và phần mềm được bảo mật tốt và ghi chép cẩn thận thành tài liệu (ví dụ: mô hình, dữ liệu, thư viện phần mềm, mô-đun, phần mềm trung gian, khuôn khổ và API bên ngoài) từ nguồn mở rộng, thương mại đã được xác minh và các nhà phát triển bên thứ ba khác để bảo đảm tính bảo mật mạnh mẽ trong hệ thống của mình.

Quý vị sẵn sàng chuyển đổi sang các giải pháp dự phòng cho các hệ thống quan trọng, nếu các tiêu chuẩn bảo mật không được đáp ứng. Quý vị sử dụng các nguồn lực như Hướng dẫn về Chuỗi Cung Ứng của NCSC và các khuôn khổ như Mức Độ Chuỗi Cung Ứng cho các Sản phẩm Phần Mềm (SLSA)[10] để theo dõi các chứng thực của chuỗi cung ứng và các vòng đời phát triển phần mềm.

Xác định, theo dõi và bảo vệ tài sản của quý vị

Quý vị nhận thức giá trị của các tài sản liên quan đến AI trong tổ chức của mình, bao gồm mô hình, dữ liệu (bao gồm phản hồi của người dùng), lời nhắc, phần mềm, tài liệu, nhật ký ghi chép và thẩm định (bao gồm thông tin về các khả năng không an toàn và chế độ lỗi), nhận ra nơi nào cần có sự đầu tư đáng kể và những nơi việc truy cập chúng có thể tạo điều kiện cho kẻ tấn công. Quý vị coi nhật ký ghi chép là các dữ liệu nhạy cảm và thực hiện các biện pháp kiểm soát để bảo vệ tính bí mật, toàn vẹn và khả dụng của chúng.

Quý vị biết tài sản của mình nằm ở đâu và đã đánh giá cũng như chấp nhận mọi rủi ro liên quan. Quý vị có các quy trình và công cụ để theo dõi, xác thực, kiểm soát phiên bản và bảo mật tài sản của mình và có có thể khôi phục trở về trạng thái tốt đã biết trong trường hợp bị xâm phạm.

Quý vị có sẵn các quy trình và biện pháp kiểm soát để quản lý những dữ liệu mà hệ thống AI có thể truy cập, cũng như quản lý nội dung do AI tạo ra dựa trên mức độ nhạy cảm của nội dung (và độ nhạy cảm của các đầu vào đã được sử dụng để tạo ra nó).

Tài liệu hóa dữ liệu, mô hình và lời nhắc của quý vị

Quý vị ghi chép thành tài liệu về quá trình tạo, vận hành và quản lý vòng đời của bất kỳ mô hình, tập dữ liệu và meta (Định nghĩa hoặc mô tả cơ bản) hoặc hệ thống nhắc nhở nào. Tài liệu của quý vị bao gồm thông tin liên quan đến bảo mật, chẳng hạn như nguồn gốc của dữ liệu đào tạo (bao gồm dữ liệu vi chỉnh và phản hồi của con người hoặc hoạt động khác), phạm vi và giới hạn dự kiến, rào chắn, băm hoặc chữ ký mật mã, thời gian lưu giữ, tính thường xuyên về xét duyệt được đề nghị và các chế độ lỗi tiềm ẩn. Các cấu trúc hữu ích để giúp thực hiện điều này bao gồm thẻ mô hình, thẻ dữ liệu và danh sách vật liệu phần mềm (SBOM). Việc tạo ra tài liệu toàn diện hỗ trợ tính minh bạch và chịu trách nhiệm[11].

Quản lý nợ kỹ thuật của quý vị

Như với bất kỳ hệ thống phần mềm nào, quý vị xác định, theo dõi và quản lý 'nợ kỹ thuật' của mình trong suốt vòng đời của hệ thống AI (nợ kỹ thuật là khi các quyết định kỹ thuật không phù hợp với các phương pháp hay nhất để đạt được kết quả ngắn hạn, nhưng lại bất lợi về lâu về dài). Giống như nợ tài chính, nợ kỹ thuật vốn không phải là điều xấu, nhưng cần được quản lý ngay từ những giai đoạn phát triển sớm nhất[12]. Quý vị nhận ra rằng làm như vậy có thể gây khó khăn hơn trong bối cảnh AI so với phần mềm tiêu chuẩn và mức nợ kỹ thuật của mình có thể sẽ cao do các vòng đời phát triển nhanh chóng và thiếu các giao thức và giao diện được thiết lập tốt. Quý vị bảo đảm các kế hoạch cho vòng đời của mình (bao gồm các quy trình để vô hiệu hóa các hệ thống AI) thẩm định, ghi nhận và giảm thiểu rủi ro đối với các hệ thống tương tự trong tương lai.

3. Triển khai an toàn

Phần này chứa đựng các hướng dẫn áp dụng cho giai đoạn **Triển khai** của vòng đời phát triển hệ thống AI, bao gồm bảo vệ hạ tầng cơ sở và mô hình khỏi bị xâm phạm, đe dọa hoặc mất mát, phát triển quy trình quản lý vấn đề và phát hành có trách nhiệm.

Bảo vệ hạ tầng cơ sở của quý vị

Quý vị áp dụng các nguyên tắc bảo mật hạ tầng cơ sở tốt cho hạ tầng cơ sở được sử dụng trong từng giai đoạn trong vòng đời hệ thống của quý vị. Quý vị áp dụng các biện pháp kiểm soát quyền truy cập thích hợp cho các API, mô hình và dữ liệu của mình cũng như quy trình đào tạo và xử lý đường ống dữ liệu của chúng, trong nghiên cứu và phát triển cũng như triển khai. Điều này bao gồm sự phân tách phù hợp giữa các môi trường chứa mã nguồn hoặc dữ liệu nhạy cảm. Điều này cũng sẽ giúp giảm nhẹ các cuộc tấn công ninh mạng thông thường nhằm đánh cắp một mô hình hoặc gây hại đến hiệu suất của nó.

Liên tục bảo vệ mô hình của quý vị

Những kẻ tấn công có thể có khả năng tái tạo chức năng của một mô hình[13] hoặc dữ liệu sử dụng để đào tạo nó[14], qua việc trực tiếp truy cập vào mô hình (bằng cách có được trọng số của mô hình) hoặc gián tiếp (bằng cách truy vấn mô hình qua một ứng dụng hoặc dịch vụ). Những kẻ tấn công còn có thể làm xáo trộn mô hình, dữ liệu hoặc lời nhắc trong hoặc sau quá trình đào tạo, khiến kết quả đầu ra không đáng tin cậy.

Quý vị bảo vệ mô hình và dữ liệu khỏi sự truy cập trực tiếp và gián tiếp theo thứ tự này, bằng cách:

- thực hiện các phương pháp hay nhất về an ninh mạng cơ bản
- thực hiện các biện pháp kiểm soát trên giao diện truy vấn để phát hiện và ngăn chặn các nỗ lực truy cập, sửa đổi và lấy thông tin bí mật

Để bảo đảm rằng các hệ thống tiêu thụ có thể xác minh các mô hình, quý vị tính toán và chia sẻ các giá trị băm mật mã và/hoặc chữ ký của các tập hồ sơ mô hình (ví dụ: trọng số mô hình) và tập dữ liệu (bao gồm các điểm kiểm tra) ngay sau khi mô hình được đào tạo. Như mỗi khi đối với mật mã, việc quản lý chìa khóa tốt là điều tối quan trọng[15].

Phương pháp giảm thiểu rủi ro đối với tính bảo mật của quý vị sẽ mật sẽ phụ thuộc đáng kể vào trường hợp sử dụng cụ thể và mô hình đe dọa. Một số ứng dụng, ví dụ như những ứng dụng liên quan đến dữ liệu rất nhạy cảm, có thể đòi hỏi sự bảo đảm về mặt lý thuyết nhưng có thể rất khó khăn hoặc tốn kém khi áp dụng. Nếu thích hợp, các công nghệ tăng cường quyền riêng tư (như quyền riêng tư khác biệt hoặc mật mã hóa đồng cấu (Mã hóa đồng cấu là việc chuyển đổi dữ liệu thành văn bản mã hóa để được phân tích và xử lý như thể nó vẫn ở dạng ban đầu)) có thể được sử dụng để khám phá hoặc bảo đảm mức độ rủi ro liên quan đến việc người tiêu thụ, người dùng và những kẻ tấn công được quyền truy cập vào các mô hình và đầu ra.

Phát triển các quy trình quản lý vấn đề

Tính không thể tránh khỏi của các vấn đề bảo mật ảnh hưởng đến hệ thống AI của quý vị được phản ánh trong các kế hoạch của quý vị cho việc ứng phó, tăng cường và khắc phục vấn đề. Kế hoạch của quý vị phản ánh các tình huống khác nhau và thường xuyên được thẩm định lại khi hệ thống và nghiên cứu rộng hơn phát triển. Quý vị lưu trữ các nguồn lực kỹ thuật số quan trọng của công ty trong các bản sao lưu ngoại tuyến. Những chuyên gia ứng phó đã được đào tạo để thẩm định và giải quyết các vấn đề liên quan đến AI. Quý vị cung cấp miễn phí nhật ký ghi chép kiểm tra với phẩm chất cao cũng như các tính năng hoặc thông tin bảo mật khác cho khách hàng và người dùng, để hỗ trợ quá trình ứng phó vấn đề của họ.

Phát hành AI một cách có trách nhiệm

Quý vị chỉ phát hành các mô hình, ứng dụng hoặc hệ thống sau khi đã đưa chúng qua quá trình đánh giá bảo mật thích hợp và hiệu quả như đo lường hiệu suất và kiểm tra bảo mật (cũng như các kiểm tra khác không nằm trong phạm vi của hướng dẫn này, chẳng hạn như an toàn hoặc công bằng), và quý vị thông báo rõ ràng cho người dùng về các giới hạn đã biết hoặc các chế độ lỗi tiềm ẩn. Chi tiết về các thư viện kiểm tra bảo mật nguồn mở rộng được cung cấp trong phần đọc thêm ở cuối tài liệu này.

Tạo điều kiện thuận lợi để người dùng có thể dễ dàng làm những điều đúng đắn

Quý vị nhận ra rằng mỗi tùy chọn chế độ cài đặt hoặc cấu hình mới cần được đánh giá cùng với lợi ích kinh doanh mà nó mang lại, cũng như mọi rủi ro bảo mật mà nó đưa vào. Điều lý tưởng là chế độ cài đặt an toàn nhất sẽ được tích hợp vào hệ thống dưới dạng lựa chọn duy nhất. Khi cần thiết phải cấu hình, tùy chọn mặc định nên được thiết lập sao để phòng ngừa một cách rộng rãi đối với các mối đe doạ thường gặp (nghĩa là bảo mật theo mặc định). Quý vị áp dụng các biện pháp kiểm soát để ngăn chặn việc sử dụng hoặc triển khai hệ thống của mình theo những phương cách độc hại.

Quý vị cung cấp cho người dùng hướng dẫn về cách sử dụng mô hình hoặc hệ thống của quý vị một cách đúng đắng, bao gồm việc nêu rõ những giới hạn và các chế độ lỗi tiềm ẩn. Quý vị nêu rõ với người dùng những khía cạnh bảo mật mà họ chịu trách nhiệm và minh bạch về nơi nào (và cách thức) dữ liệu của họ có thể được sử dụng, truy cập hoặc lưu trữ (ví dụ: nếu dữ liệu đó được sử dụng để tái đào tạo mô hình hoặc được các nhân viên hoặc đối tác xem xét).

4. Vận hành và bảo trì an toàn

Phần này chứa đựng các hướng dẫn áp dụng cho giai đoạn **vận hành và bảo trì an toàn** của vòng đời phát triển hệ thống AI. Nó bao gồm các hướng dẫn về các hành động có liên quan một khi hệ thống đã được triển khai, bao gồm nhật ký ghi chép và theo dõi, quản lý cập nhật và chia sẻ thông tin.

Theo dõi hoạt động của hệ thống của quý vị

Quý vị đo lường kết quả đầu ra và hiệu suất của mô hình và hệ thống của mình để có thể quan sát những thay đổi đột ngột và từ từ trong hoạt động ảnh hưởng đến an ninh. Quý vị có thể tính toán và xác định được các hành vi xâm nhập và xâm phạm có thể xảy ra, cũng như hiện tượng trôi dạt dữ liệu tự nhiên.

Theo dõi đầu vào của hệ thống của quý vị

Để phù hợp với các yêu cầu về việc bảo vệ quyền riêng tư và dữ liệu, quý vị theo dõi và ghi chép nhật ký thông tin đầu vào vào hệ thống của mình (chẳng hạn như yêu cầu suy luận, truy vấn hoặc lời nhắc) để thực hiện nghĩa vụ tuân thủ, kiểm toán, điều tra và khắc phục trong trường hợp bị xâm phạm hoặc lạm dụng. Điều này có thể bao gồm việc phát hiện rõ ràng các đầu vào không được phân phối và/hoặc đối nghịch, bao gồm cả những đầu vào nhằm mục đích khai thác các bước để chuẩn bị dữ liệu (chẳng hạn như cắt xén và thay đổi kích thước cho hình ảnh).

Tuân theo phương pháp bảo mật theo thiết kế để cập nhật

Quý vị bao gồm cập nhật tự động theo mặc định cho mọi sản phẩm và sử dụng các quy trình cập nhật an toàn, theo mô-đun để phân phối chúng. Quá trình cập nhật của quý vị (bao gồm cả chế độ kiểm tra và đánh giá) phản ánh thực tế rằng những thay đổi đối với dữ liệu, mô hình hoặc lời nhắc có thể dẫn đến những thay đổi về hoạt động của hệ thống (ví dụ: quý vị coi các bản cập nhật lớn như những phiên bản mới). Quý vị hỗ trợ người dùng để đánh giá và ứng phó với các thay đổi của mô hình (ví dụ: bằng cách cung cấp quyền truy cập xem trước và API được phiên bản).

Thu thập và chia sẻ bài học kinh nghiệm

Quý vị tham gia vào cộng đồng chia sẻ thông tin, hợp tác trên toàn bộ hệ sinh thái toàn cầu của ngành công nghiệp, học viện và chính phủ để chia sẻ những phương pháp hay nhất khi thích hợp. Quý vị duy trì các kênh liên lạc mở rộng để nhận phản hồi về bảo mật hệ thống, cả nội bộ và bên ngoại tổ chức của mình, bao gồm cả việc đồng ý cho các nhà nghiên cứu bảo mật nghiên cứu và phúc trình về các lỗ hổng bảo mật. Khi cần thiết, quý vị sẽ chuyển vấn đề đến cộng đồng rộng lớn hơn, chẳng hạn như xuất bản các bản tin phản hồi việc tiết lộ lỗ hổng bảo mật, bao gồm việc liệt kê chi tiết và đầy đủ các lỗ hổng thường thấy. Quý vị thực hiện các biện pháp để giảm thiểu và khắc phục các vấn đề một cách nhanh chóng và thích hợp.

Tham khảo thêm

Phát triển AI

Các nguyên tắc bảo mật của học máy
Hướng dẫn chi tiết của NCSC về việc phát triển, triển khai hoặc vận hành hệ thống có thành phần ML.

Bảo mật theo Thiết kế - Thay đổi Cán cân Rủi ro An ninh Mạng: Các Nguyên tắc và Phương pháp Bảo mật theo Thiết kế cho Phần mềm
Được đồng tác giả bởi CISA, NCSC và các cơ quan khác, hướng dẫn này mô tả cách thức các nhà sản xuất hệ thống phần mềm, bao gồm cả AI, nên thực hiện các bước để bao gồm bảo mật trong giai đoạn thiết kế của quá trình phát triển sản phẩm và phát hành sản phẩm có tính an toàn ngay từ ban đầu.

Tóm tắt các Mối Lo ngại về Bảo mật AI
Do Văn phòng Liên bang Đức về An ninh Thông tin (BSI) soạn thảo, tài liệu này giới thiệu về các cuộc tấn công có thể xảy ra đối với hệ thống học máy và các biện pháp phòng thủ tiềm năng để đối phó với các cuộc tấn công đó.

Các Nguyên tắc Hướng dẫn Quốc tế của Quy trình Hiroshima dành cho các Tổ chức Phát triển Hệ thống AI Tiên tiến và Quy tắc Ứng xử Quốc tế của Quy trình Hiroshima dành cho các Tổ chức Phát triển Hệ thống AI Tiên tiến
Những tài liệu này, được xuất bản như một phần của Quy trình AI G7 Hiroshima, cung cấp hướng dẫn cho các tổ chức phát triển các hệ thống AI tiên tiến nhất, bao gồm các mô hình nền tảng tiên tiến nhất và hệ thống AI sáng tạo nhằm mục đích thúc đẩy cho AI an toàn, bảo mật và đáng tin cậy trên toàn thế giới.

AI Verify
Khuôn khổ Kiểm tra Quản lý AI và bộ Công cụ Phần mềm của Singapore xác nhận hiệu suất của hệ thống AI so với bộ nguyên tắc được quốc tế công nhận, qua các kiểm tra được tiêu chuẩn hóa.

Khuôn khổ Đa Tầng về Các cách Thực hành An ninh Mạng Tốt cho AI – ENISA (europa.eu)
Một khuôn khổ để hướng dẫn Cơ quan Thẩm quyền Quốc gia và các bên liên quan đến AI về các bước họ cần thực hiện để bảo vệ hệ thống, hoạt động và quy trình AI của họ.

ISO 5338: Các quy trình vòng đời của hệ thống AI (Đang xem xét)
Một tập hợp các quy trình và khái niệm liên quan để mô tả vòng đời của hệ thống AI dựa trên học máy và hệ thống phỏng đoán.

Danh mục Tiêu chuẩn Tuân thủ Dịch vụ Đám mây AI (AIC4)
Danh mục Tiêu chuẩn Tuân thủ Dịch vụ Đám mây AI của BSI cung cấp các tiêu chuẩn dành riêng cho AI, cho phép đánh giá tính bảo mật của dịch vụ AI trong suốt vòng đời của dịch vụ đó.

NIST IR 8269 (Bản dự thảo) Phân loại và Thuật ngữ của Đối nghịch Học Máy
Một tập hợp các quy trình và khái niệm liên quan để mô tả vòng đời của hệ thống AI dựa trên học máy và hệ thống phỏng đoán.

MITRE ATLAS
Một cơ sở kiến thức về chiến thuật, kỹ thuật và nghiên cứu các trường hợp đối nghịch đối với hệ thống học máy (ML), được mô hình hóa và liên kết với khuôn khổ MITER ATT&CK.

Khái quát về các Rủi ro Nghiêm Trọng của AI (2023)
Do Trung tâm An toàn AI soạn thảo, tài liệu này nêu ra các lĩnh vực rủi ro do AI gây ra.

Mô hình Ngôn ngữ Lớn: Những Cơ hội và Rủi ro đối với Ngành công nghiệp và Cơ quan Quản lý
Tài liệu do BSI soạn thảo dành cho các công ty, cơ quan có thẩm quyền và các nhà phát triển muốn tìm hiểu thêm về những cơ hội và rủi ro của việc phát triển, triển khai và/hoặc sử dụng LLMs (Mô hình Ngôn ngữ Lớn).

Các dự án nguồn mở rộng để giúp người dùng kiểm tra bảo mật các mô hình AI bao gồm:

- Hộp Công cụ An Toàn chống Đối nghịch (IBM)
- CleverHans (Đại học Toronto)
- TextAttack (Đại học Virginia)
- Prompt Bench (Microsoft)
- Counterfit (Microsoft)
- AI Verify (Cơ quan Phát triển Truyền thông Infocomm, Tân Gia Ba)

An ninh mạng

Mục tiêu Hiệu suất An ninh Mạng của CISA
Một tập hợp các biện pháp bảo vệ chung mà tất cả các thực thể hạ tầng cơ sở quan trọng nên thực hiện để giảm thiểu đáng kể khả năng và tác động của các rủi ro đã biết và các kỹ thuật của đối thủ.

Khuôn khổ NCSC CAF
Khuôn khổ Đánh giá Mạng (CAF) cung cấp hướng dẫn cho các tổ chức chịu trách nhiệm về các dịch vụ và các hoạt động tối quan trọng.

Khuôn khổ Bảo mật Chuỗi Cung ứng của MITER
Một khuôn khổ đánh giá các nhà cung cấp và nhà cung cấp dịch vụ trong chuỗi cung ứng.

Quản lý rủi ro

Khuôn khổ Quản lý Rủi ro AI (AI RMF) của NIST
AI RMF phác thảo cách quản lý rủi ro kỹ thuật xã hội đối với các cá nhân, tổ chức và xã hội có liên quan riêng đến AI.

ISO 27001: Bảo mật thông tin, an ninh mạng và bảo vệ quyền riêng tư
Bộ tiêu chuẩn này giúp hướng dẫn các tổ chức về việc thành lập, triển khai và duy trì hệ thống quản lý bảo mật thông tin.

ISO 31000: Quản lý rủi ro
Một bộ tiêu chuẩn quốc tế cung cấp cho các tổ chức các hướng dẫn và nguyên tắc về quản lý rủi ro trong tổ chức.

Hướng dẫn Quản lý Rủi ro của NCSC
Hướng dẫn này giúp các chuyên gia quản lý rủi ro an ninh mạng hiểu rõ và quản lý tốt hơn các rủi ro an ninh mạng ảnh hưởng đến tổ chức của họ.

Ghi chú

1. Được định nghĩa ở đây là một cá nhân, cơ quan công quyền, cơ quan hoặc tổ chức khác đang phát triển một hệ thống AI (hoặc có một hệ thống AI được phát triển) và đưa hệ thống đó ra thị trường hoặc đưa vào sử dụng dưới tên hoặc thương hiệu của chính họ
2. Để biết thêm thông tin về bảo mật theo thiết, kế hãy xem trang mạng và hướng dẫn Bảo mật theo Thiết kế của CISA Thay đổi Cán cân Rủi ro An ninh Mạng: Các Nguyên Tắc Và Phương Pháp Bảo Mật Theo Thiết Kế Cho Phần Mềm
3. Trái ngược lại với các phương pháp AI không sử dụng ML, chẳng hạn như hệ thống áp dụng các quy tắc
4. CEPS mô tả bảy loại tương tác phát triển AI khác nhau trong ấn phẩm của họ 'Điều chỉnh Chuỗi Giá trị AI để tuân thủ Đạo luật AI của EU (Liên minh Châu Âu)'
5. ISO/IEC 22989:2022(en) định nghĩa điều này là 'một thành phần chức năng xây dựng một hệ thống AI'
6. NIST được giao nhiệm vụ cung cấp các hướng dẫn (và thực hiện các hành động khác) để thúc đẩy sự phát triển và sử dụng Trí tuệ Nhân tạo (AI) một cách an toàn, bảo mật và đáng tin cậy. Xem phần Trách nhiệm của NIST Chiếu theo Sắc lệnh Hành pháp ngày 30 tháng 10 năm 2023
7. Có thêm thông tin về mô hình mối đe dọa, tại trang mạng OWASP Foundation
8. Xem MITRE ATLAS Đối nghịch Học Máy 101
9. GitHub: RCE PoC cho Tensorflow sử dụng một tầng Lambda độc hại
10. SLSA: 'Bảo vệ tính toàn vẹn của sản phẩm trên mọi chuỗi cung ứng phần mềm'
11. METI (Bộ Kinh tế, Thương mại và Công nghiệp Nhật Bản, 2023), 'Hướng dẫn Giới thiệu Danh sách Vật liệu Phần mềm (SBOM) cho Quản lý Phần mềm'
12. Nghiên cứu của Google: Học Máy: Thẻ Tín dụng Lãi suất Cao cho khoản Nợ Kỹ thuật
13. Tramèr và cộng sự 2016, Đánh cắp Mô hình Học Máy qua API (Giao diện Lập trình Ứng dụng) Dự đoán
14. Boenisch, 2020, Các cuộc Tấn công vào Quyền Riêng Tư của Học Máy (Phần 1): Các cuộc Tấn công Đảo ngược Mô hình với Khuôn khổ IBM-ART
15. Trung tâm An ninh Mạng Quốc gia, 2020, Thiết kế và xây dựng Hạ tầng Cơ sở Chìa khóa Công cộng được lưu trữ riêng