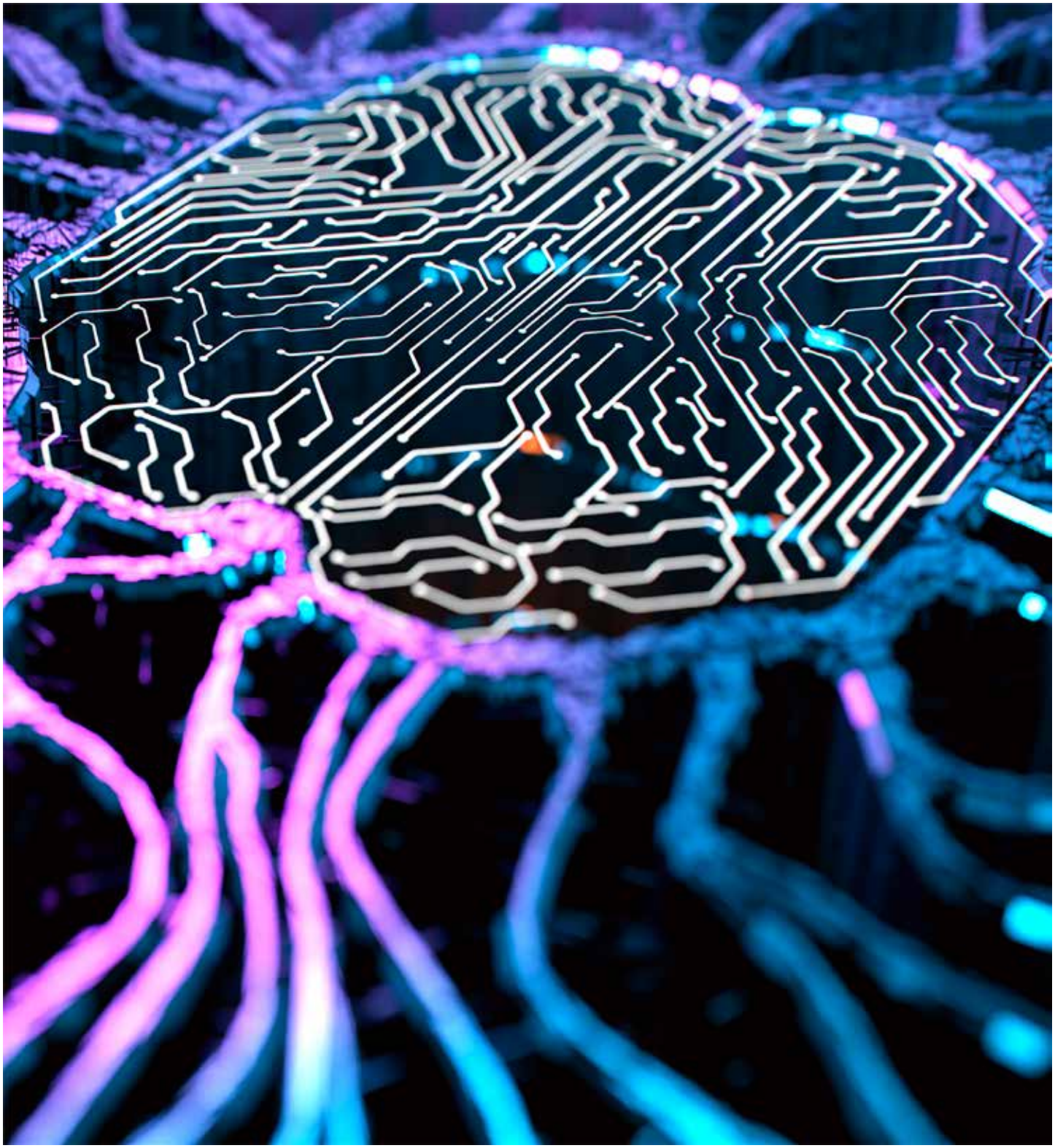
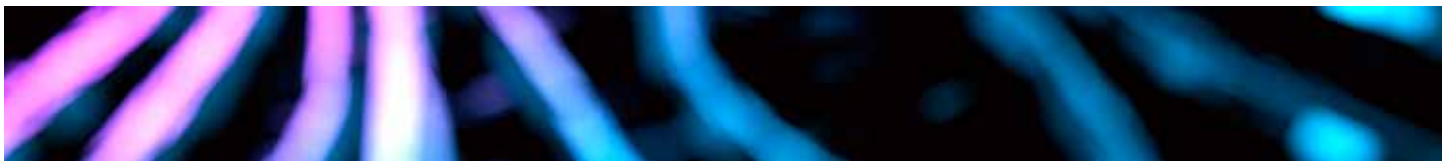
안전한 AI 시스템 개발을 위한 지침





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

문서 소개

본 문서는 영국 국립 사이버 보안 센터(National Cyber Security Centre, NCSC)와 미국 사이버 보안 및 인프라 보안 에이전시(Cybersecurity and Infrastructure Security Agency, CISA), 그리고 다음 국제 파트너들에 의해 발간되었습니다:

- National Security Agency(NSA)
- Federal Bureau of Investigations(FBI)
- Australian Signals Directorate's Australian Cyber Security Centre(ACSC)
- Canadian Centre for Cyber Security(CCCS)
- New Zealand National Cyber Security Centre(NCSC-NZ)
- Chile's Government CSIRT
- Czechia's National Cyber and Information Security Agency(NUKIB)
- Information System Authority of Estonia(RIA) and National Cyber Security Centre of Estonia(NCSC-EE)
- French Cybersecurity Agency(ANSSI)
- Germany's Federal Office for Information Security(BSI)
- Israeli National Cyber Directorate(INCD)
- Italian National Cybersecurity Agency(ACN)
- Japan's National center of Incident readiness and Strategy for Cybersecurity(NISC)
- Japan's Secretariat of Science, Technology and Innovation Policy, Cabinet Office
- Nigeria's National Information Technology Development Agency(NITDA)
- Norwegian National Cyber Security Centre(NCSC-NO)
- Poland Ministry of Digital Affairs
- Poland's NASK National Research Institute(NASK)
- Republic of Korea National Intelligence Service(NIS)
- Cyber Security Agency of Singapore(CSA)

공로 인정

다음 기관들이 본 지침의 개발에 기여하였습니다:

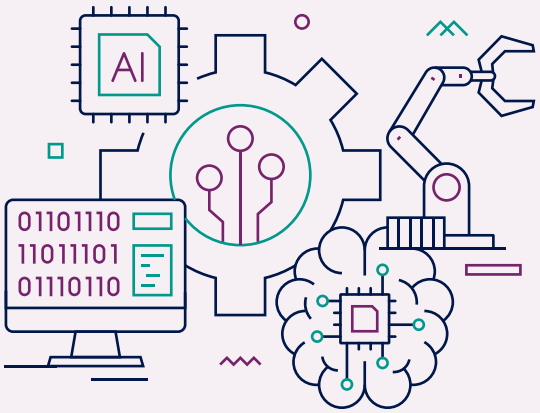
- Alan Turing Institute
- Anthropic
- Databricks
- Georgetown University's Center for Security and Emerging Technology
- Google
- Google DeepMind
- IBM
- ImBue
- Microsoft
- OpenAI
- Palantir
- RAND
- Scale AI
- Software Engineering Institute at Carnegie Mellon University
- Stanford Center for AI Safety
- Stanford Program on Geopolitics, Technology and Governance

면책 조항

본 지침에서 제공되는 정보는 NCSC 및 저작 기관에 의해 "있는 그대로" 제공되며 법률적으로 요구되는 경우를 제외하고는 해당 정보의 사용으로 인해 발생하는 모든 종류의 손실, 부상 또는 손해에 대해 책임을 지지 않습니다. 본 지침에서 제공되는 정보는 NCSC 및 저작 기관에 의한 어떠한 제3의 기관, 상품 또는 용역에 대한 보증이나 권장이 되지 아니하며, 이를 시사하지도 않습니다. 웹사이트 및 제3자의 자료에 대한 링크와 인용은 정보 목적으로만 제공되며 타 자료와 비교해 해당 자료에 대한 보증이나 권장을 나타내지는 않습니다.

본 지침은 TLP:CLEAR에 기초해 제공됩니다(https://www.first.org/tlp/).

목차



개요서

본 문서는 인공지능(AI)을 사용하는 모든 시스템의 제공업체에 지침을 제안하며 해당 시스템이 처음부터 만들어진 경우 또는 타인이 제공한 도구 및 서비스를 기반으로 구축된 경우가 모두 해당됩니다. 본 지침의 적용은 AI 시스템 제공업체들로 하여금 AI 시스템이 의도한대로 작동하고 필요시에 항상 사용할 수 있으며, 승인되지 않은 제3자에 민감한 정보를 공개하지 않고 작동할 수 있는 AI 시스템을 구축하는 데 도움이 될 것입니다.

본 지침은 주로 어떤 기관에서 호스팅하는 모델을 사용하거나 외부 API(애플리케이션 프로그래밍 인터페이스)를 사용하는 AI 시스템 제공업체를 대상으로 합니다. 당사는 **모든** 이해관계자(데이터 과학자, 개발자, 관리자, 결정권자 및 리스크 담당자)가 본 지침을 열람해 각자의 AI 시스템의 **디자인**, **개발**, **배포** 및 **운영**에 대해 정보에 근거한 결정을 내리는데 도움을 받길 권장합니다.

지침 소개

AI 시스템은 사회에 많은 이점을 가져올 수 있는 잠재력을 갖고 있습니다. 그러나 AI의 모든 가능성이 완전히 실현되려면 AI는 항상 안전하고 책임감 있는 방식으로 개발, 배포, 운영되어야 합니다.

AI 시스템은 새로운 보안 취약점에 노출되어 있으며 이는 일반적인 사이버 보안 위협과 함께 고려되어야 합니다. AI와 같이 개발 속도가 빠른 분야에서 보안은 종종 두번째 고려사항이 될 수도 있습니다. 보안은 반드시 핵심 요건이 되어야 하며 이는 개발 단계뿐만 아니라 시스템의 전체 수명주기 동안 유지되어야 합니다.

이에 따라 본 지침은 AI 시스템 개발 수명주기의 핵심 부문 4개로 나뉘어져 있습니다: **안전한 설계**, **안전한 개발**, **안전한 배포**, 및 **안전한 운영 및 관리**. 각 부문에서 당사는 기관 차원의 AI 시스템 개발 절차에 대한 전반적인 리스크를 낮출 수 있는 고려사항 및 완화조치를 제안합니다.

1. **안전한 설계**
   본 부문은 AI 시스템 개발 수명주기 중 설계 단계에 해당하는 지침을 포함합니다. 이는 리스크에 대한 이해와 위협 모델링은 물론, 시스템 및 모델 설계에 대해 고려해야 하는 특정 토픽과 장단점을 다룹니다.

2. **안전한 개발**
   본 부문은 AI 시스템 개발 수명주기 중 개발 단계에 해당하는 지침이며, 이에는 공급망 보안, 기록 및 자산과 기술적 부채 (technical debt) 관리 등이 포함됩니다.

3. **안전한 배포**
   본 부문은 AI 시스템 개발 수명주기 중 배포 단계에 해당하는 지침으로 이에는 인프라와 모델을 타협, 위협 또는 손실로부터 보호하는 방법, 사고 관리 절차 개발, 그리고 책임있는 공개 등이 포함됩니다.

4. **안전한 운영 및 관리**
   본 부문은 AI 시스템 개발 수명주기 중 안전한 운영 및 관리 단계에 해당하는 지침을 포함합니다. 로그 및 모니터링, 업데이트 관리, 정보 공유 등 시스템 배포 후 특히 관련된 조치에 대한 지침을 제공합니다.

본 지침은 '안전한 디폴트(secure by default)' 접근법을 따르며 이는 NCSC의 '안전한 개발 및 배포 지침(Secure development and deployment guidance)', NIST의 '안전한 소프트웨어 개발 프레임워크(Secure Software Development Framework)', 및 CISA, NCSC 그리고 국제 사이버 기관에서 발간한 '안전 설계 원칙(Secure by design principles)'에 정의된 관행과 밀접하게 일치합니다. 이는 다음을 우선시합니다:

- 고객을 위한 보안 결과에 대한 책임을 지는 것
- 철저한 투명성과 책임을 구현하는 것
- 설계 측면에서 안전한 조직 구조 및 리더십 구축을 사업의 가장 높은 우선순위로 설정하는 것

서론

인공지능(AI) 시스템은 사회에 많은 이점을 가져올 수 있는 잠재력을 갖고 있습니다. 그러나 AI의 모든 가능성이 완전히 실현되려면 AI는 항상 안전하고 책임감 있는 방식으로 개발, 배포, 그리고 운영되어야 합니다. 사이버 보안은 AI 시스템의 안전성, 복원력, 프라이버시, 공정성, 유효성 및 신뢰성을 위한 필수 전제 조건입니다.

그러나 AI 시스템은 새로운 보안 취약점에 노출되어 있으며 이는 일반적인 사이버 보안 위협과 함께 고려되어야 합니다. AI와 같이 개발 속도가 빠른 분야에서 보안은 종종 두번째 고려사항이 될 수도 있습니다. 보안은 반드시 핵심 요건이 되어야 하며 이는 개발 단계에서 뿐만 아니라 시스템의 전체 수명주기 동안 유지되어야 합니다.

본 문서는 인공지능(AI)을 사용하는 모든 시스템의 제공업체에 지침을 제안하며[1], 해당 시스템이 처음부터 만들어진 경우 또는 타인이 제공한 도구 및 서비스를 기반으로 구축된 경우가 모두 해당됩니다. 본 지침의 적용은 AI 시스템 제공업체들로 하여금 AI 시스템이 의도한대로 작동하고 필요시에 항상 사용할 수 있으며, 승인되지 않은 제3자에 민감한 정보를 공개하지 않고 작동할 수 있는 AI 시스템을 구축하는 데 도움이 될 것입니다.

본 지침은 기존의 사이버 보안, 리스크 관리, 및 사고 대응에 대한 모범 관례와 함께 고려되어야 합니다. 당사는 특히 제공업체들이 미국의 사이버보안 및 인프라 보안 에이전시(Cybersecurity and Infrastructure Security Agency, CISA), 영국의 국립 사이버보안 센터(National Cyber Security Centre, NCSC), 그리고 당사의 모든 국제 파트너가 함께 개발한 '안전한 디폴트(secure by default)' 원칙[2]을 따르길 권장합니다. 해당 원칙은 다음을 우선시합니다:

- 고객을 위한 보안 결과에 대한 책임을 갖는 것
- 완전한 투명성과 책임을 구현하는 것
- 설계 측면에서 안전한 조직 및 리더십 구축을 사업의 가장 높은 우선순위로 설정하는 것

'안전한 디폴트(secure by default)' 원칙을 따르기 위해서는 시스템 수명주기 전반에 걸쳐 상당한 자원이 요구됩니다. 이는 개발자가 시스템 설계의 각 단계와 개발 수명주기의 모든 단계에 걸쳐 고객을 보호하는 도구의 **기능**, **메커니즘**, 및 **적용**에 우선순위를 두는 데 투자해야 함을 의미합니다. 이는 추후 고비용 재설계를 방지할 수 있을 뿐만 아니라 고객과 데이터를 단기적으로 보호할 수 있는 방법입니다.

AI 보안은 왜 다른가요?

본 지침에서 설명하는 'AI'는 머신러닝(ML) 애플리케이션을 의미합니다[3]. ML의 모든 유형이 범위에 포함됩니다. 당사가 정의하는 ML 애플리케이션은 다음과 같습니다:

- 인간이 명시적으로 프로그래밍해야 할 규칙을 사용하지 않고도 컴퓨터가 데이터의 패턴을 인식하고 설명할 수 있도록 하는 소프트웨어 구성 요소(모델)가 포함됩니다
- 통계적 추론을 기반으로 예측, 권장 또는 결정을 도출합니다

AI 시스템은 기존의 사이버 보안 위협 뿐만 아니라 새로운 유형의 취약점에 노출됩니다. '적대적 머신러닝(AML)'이라는 용어는 하드웨어, 소프트웨어, 워크플로 및 공급 체인을 포함한 ML 구성 요소의 근본적인 취약점에 대한 악용을 의미합니다. 공격자는 AML을 통해 ML 시스템에서 다음과 같은 예상외의 동작을 유발할 수 있습니다:

- 모델의 분류 또는 회귀 성능에 영향을 미칩니다
- 사용자가 승인되지 않은 작업을 수행할 수 있도록 합니다
- 민감한 모델 정보를 추출합니다

이러한 효과를 달성할 수 있는 여러가지 방법이 있으며, 예시로는 대형 언어 모델(LLM) 도메인에서의 신속한 주입 공격이나 트레이닝 데이터 또는 사용자 피드백을 고의적으로 손상시키는 것('데이터 중독') 등이 있습니다.

본 지침의 독자는 누구인가요?

이 문서는 주로 기관에서 호스팅하는 모델을 기반으로 하거나 외부 앱 프로그래밍 인터페이스(API)를 사용하는 AI 시스템 제공업체를 대상으로 합니다. 그러나 당사는 **모든** 이해관계자(데이터 과학자, 개발자, 관리자, 결정권자 및 리스크 담당자 포함)가 본 지침을 열람해 각자의 머신러닝 AI 시스템의 **설계**, **배포** 및 **운영** 에 대해 정보에 근거한 결정을 내리는 데 도움을 받길 권장합니다.

그럼에도 불구하고 지침의 모든 부분이 모든 기관에 직접적으로 적용될 수는 없을 것입니다. 정교함 수준과 공격 방법은 AI 시스템을 표적으로 삼는 공격자에 따라 달라지므로 본 지침은 기관의 사용 사례 및 위협 프로필과 함께 고려되어야 합니다.

안전한 AI를 개발할 책임은 누구에게 있나요?

현대 AI 공급 체인에서는 종종 많은 이해관계자가 관여되어 있습니다. 단순한 접근법에서는 2 종류의 이해관계자가 있습니다:

- ‘공급자’ - 데이터 큐레이션, 알고리즘 개발, 설계, 배포 및 유지 관리를 담당
- ‘사용자’ - 인풋을 제공하고 아웃풋을 수령

이러한 공급자-사용자 접근 방식은 많은 애플리케이션에서 사용되나 이젠 공급자가 제3자가 제공하는 소프트웨어, 데이터, 모델 및/또는 원격 서비스를 자체 시스템에 통합할 수 있기 때문에 점점 더 보편화되지 않고 있습니다[4]. 이러한 복잡한 공급 체인은 최종 사용자로 하여금 안전한 AI에 대한 책임이 누구에게 있는지를 이해하는 것을 더 어렵게 만듭니다.

일반적으로 사용자('최종 사용자' 또는 외부 AI 구성 요소를 통합한 제공업체[5])는 자신이 사용하는 시스템과 관련된 위험을 완전히 이해, 평가 또는 해결하는 데 필요한 충분한 가시성 및/또는 전문 지식을 보유하고 있지 않습니다. 그렇기 때문에 ‘안전 설계(secure by design)’ 원칙에 따라 **AI 구성요소 제공자들은 공급 체인 하위 사용자의 보안 결과에 대해 책임을 져야 합니다.**

제공업체는 모델, 파이프라인 및/또는 시스템 내에서 가능한 경우 보안 통제 및 리스크 완화 기능을 구현해야 하며, 설정이 사용되는 경우 가장 안전한 옵션을 디폴트값으로 설정해야 합니다. 리스크가 완화될 수 없는 경우, 공급자는 다음에 대한 책임을 져야 합니다:

- 공급 체인 하위 사용자에게 자신과 (해당되는 경우) 자신의 사용자가 감수하고 있는 위험을 알립니다
- 이들에게 구성 요소를 안전하게 사용하는 방법에 대해 조언합니다

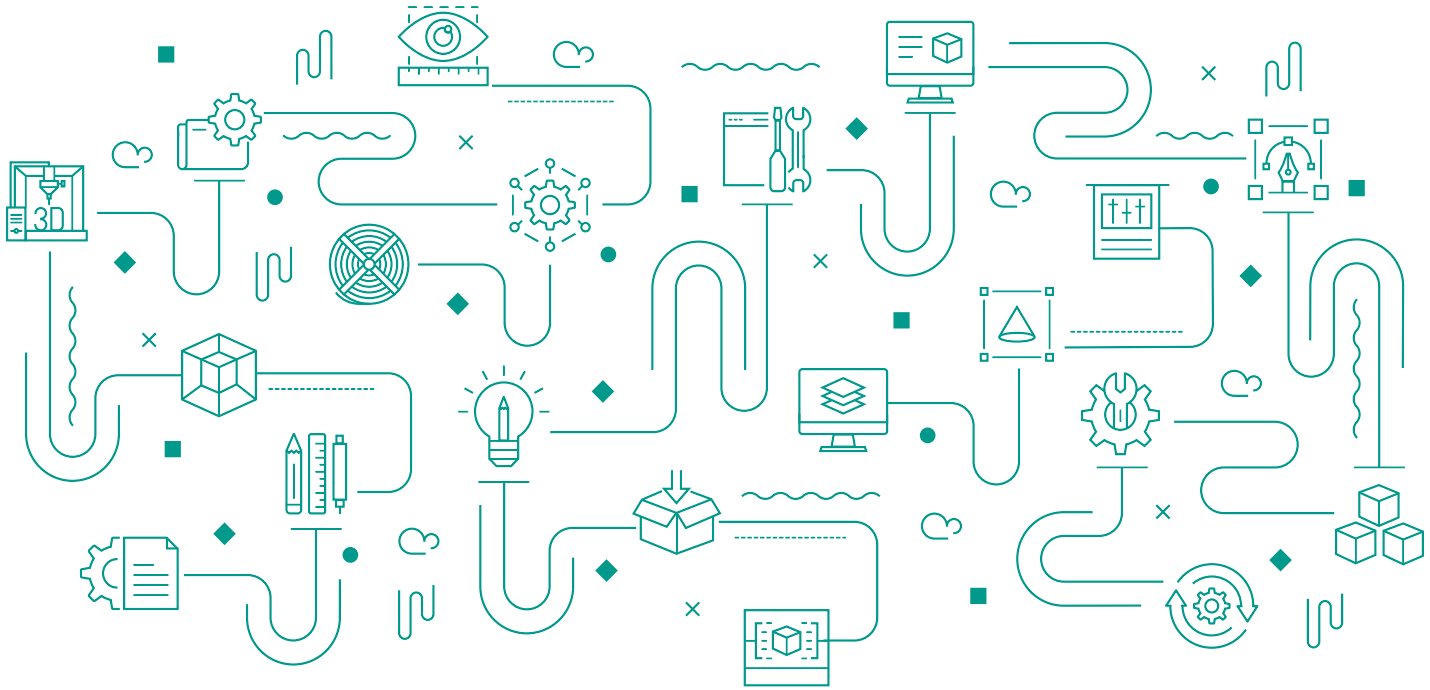
시스템 손상으로 인해 실질적이거나 광범위한 물리적 또는 평판 손상, 심각한 비즈니스 운영 손실, 민감한 정보나 기밀 정보의 유출 및/또는 법적 영향이 발생할 수 있는 경우 AI 사이버 보안 위험은 **중대** 사안으로 처리되어야 합니다.

안전한 AI 시스템 개발을 위한 지침

본 지침은 AI 시스템 개발 수명주기 내의 핵심 부문 4개로 나뉘어져 있습니다: **안전한 설계**, **안전한 개발**, **안전한 배포**, 및 **안전한 운영 및 관리**. 각 부문에서 당사는 기관 차원의 AI 시스템 개발 절차에 대한 전반적인 리스크를 낮출 수 있는 고려사항 및 리스크 완화조치를 제안합니다.

본 문서에 제공되는 지침은 다음에 정의된 소프트웨어 개발 수명주기 관행과 밀접하게 일치합니다:

- NCSC의 '안전한 개발 및 배포 지침(Secure development and deployment guidance)'
- NIST(National Institute of Standards and Technology)의 '안전한 소프트웨어 개발 프레임워크(Secure Software Development Framework, SSDF)'[6]



1. 안전한 설계

본 부문은 AI 시스템 개발 수명주기 중 **설계** 단계에 해당하는 지침을 포함합니다. 이는 리스크에 대한 이해와 위협 모델링은 물론, 시스템 및 모델 설계에 대해 고려해야 하는 특정 토픽과 장단점을 다룹니다.

위협 및 리스크에 대한 직원 이해도를 향상시키세요

시스템 소유자 및 시니어 리더들은 안전한 AI에 대한 위협과 이에 대한 완화 방법을 이해합니다. 여러분의 데이터 과학자 및 개발자들은 관련 보안 위협 및 실패 모드에 대한 이해도를 유지하며 관련 리스크 담당자들이 정보에 근거한 의사결정을 내릴 수 있도록 돕습니다. 여러분은 사용자에게 AI 시스템이 직면한 고유한 보안 위험에 대한 지침을 제공하고(예: 표준 InfoSec 교육의 일부로) 개발자들에게 보안 코딩 기술과 안전하고 책임감 있는 AI 관행에 대한 훈련을 제공합니다.

여러분의 시스템에 대한 위협을 모델링하세요

리스크 관리 절차의 일환으로 시스템에 대한 위협을 평가하기 위한 전체적인 절차를 확립하며, 이에는 AI 구성 요소가 손상되거나 예상치 못한 방식으로 작동하는 경우 이것이 시스템, 사용자, 기관 및 더 넓은 사회에 미칠 수 있는 잠재적 영향을 이해하는 것이 포함됩니다[7]. 해당 절차에는 AI 관련 특정 리스크의 영향을 평가하고[8] 의사 결정을 문서화하는 작업이 포함됩니다.

여러분은 여러분의 시스템에 사용되는 데이터의 민감도 및 유형이 공격자로 하여금 표적으로서의 가치에 영향을 미칠 수 있다는 점을 이해하고 있습니다. 평가 수행 시 AI 시스템이 점점 더 높은 가치의 표적으로 인식되고 AI 자체가 새로운 자동화된 공격 루트를 가능하게 함에 따라 일부 위협이 커질 수 있다는 점을 고려해야 합니다.

보안 뿐만 아니라 실용성과 성능을 갖춘 시스템을 설계하세요

현재 진행하는 작업이 AI를 통해 가장 적절하게 처리될 수 있음을 확신합니다. 이를 결정한 후에는 AI 관련 특정 설계 요소의 적절성을 평가합니다. 기능, 사용자 경험, 배포 환경, 성능, 보증, 감독, 윤리적 및 법적 요건 등 다른 고려사항과 함께 위협 모델 및 관련 보안 리스크 완화 방법을 고려합니다. 예시:

- 내부 개발 또는 외부 구성요소 사용 여부를 선택할 때 공급 체인 보안을 고려합니다. 예시로는 다음의 경우가 있습니다:
  - 새로운 모델을 훈련시키거나, 기존 모델을 사용하거나(파인튜닝 여부와 관계없이) 외부 API를 통해 모델을 사용하는 등 여러분의 필요에 따라 적절한 옵션을 선택합니다
  - 외부 모델 제공자와 협력하기로 하는 경우 해당 제공자의 자체 보안 체제에 대한 실사 평가를 포함합니다
  - 외부 라이브러리를 사용하는 경우 실사 평가를 수행합니다(예: 시스템이 임의 코드 실행에 즉시 노출되지 않고 신뢰할 수 없는 모델을 로딩하지 못하도록 방지하는 통제 기능이 라이브러리에 있는지 확인합니다[9])
  - 타사 모델 또는 직렬화된 가중치를 가져올 때 스캐닝 및 격리/샌드박싱을 구현합니다. 이는 신뢰할 수 없는 제3자 코드로 처리되어야 하며 원격 코드 실행을 활성화할 수 있습니다

  - 외부 API를 사용하는 경우 잠재적으로 민감한 정보를 전송하기 전에 사용자에게 로그인 및 확인을 요구하는 등 기관의 통제 범위를 벗어난 서비스로 전송될 수 있는 데이터에 적절한 통제를 적용합니다
  - 데이터와 인풋을 적절하게 확인하고 정리합니다; 여기에는 사용자 피드백이나 지속적인 학습 데이터를 모델에 통합하고 트레이닝 데이터가 시스템 동작을 정의한다는 것을 인식하는 경우가 포함됩니다

- AI 소프트웨어 시스템 개발을 기존 보안 개발 및 운영 모범 사례에 통합합니다; AI 시스템의 모든 요소는 가능한 경우 알려진 취약점 부류를 줄이거나 제거하는 코딩 방법과 언어를 사용하여 적절한 환경에서 작성됩니다
- 파일 수정 또는 외부 시스템으로 아웃풋을 전달하는 등 AI 구성 요소가 작업을 트리거해야 하는 경우 해당하는 작업에 적절한 제한을 적용합니다(필요한 경우 외부 AI 및 비 AI 안전 장치 포함)
- 사용자 상호 작용에 관한 결정은 AI 특정 위험에 따라 결정됩니다. 예를 들면 다음과 같습니다:
  - 시스템은 잠재적인 공격자에게 불필요한 수준의 세부 정보를 공개하지 않고 사용자에게 사용 가능한 아웃풋을 제공합니다
  - 필요한 경우 시스템은 모델 아웃풋에 대한 효과적인 보호막을 제공합니다
  - 외부 고객이나 협력자에게 API를 제공하는 경우 API를 통해 AI 시스템에 대한 공격을 완화하는 적절한 통제를 적용합니다
  - 가장 안전한 설정을 기본 설정으로 하여 시스템에 통합합니다
  - 시스템 기능에 대한 접근을 제한하기 위해 최소 권한 원칙을 적용합니다
  - 사용자에게 더 위험한 기능을 설명하고 사용자가 해당 기능을 사용하려면 이를 선택해야 하도록 요구합니다; 금지된 사용 사례를 전달하고, 가능한 경우 사용자에게 대체 해결법을 알립니다

선택하는 AI 모델의 보안 이점과 장단점을 고려하세요

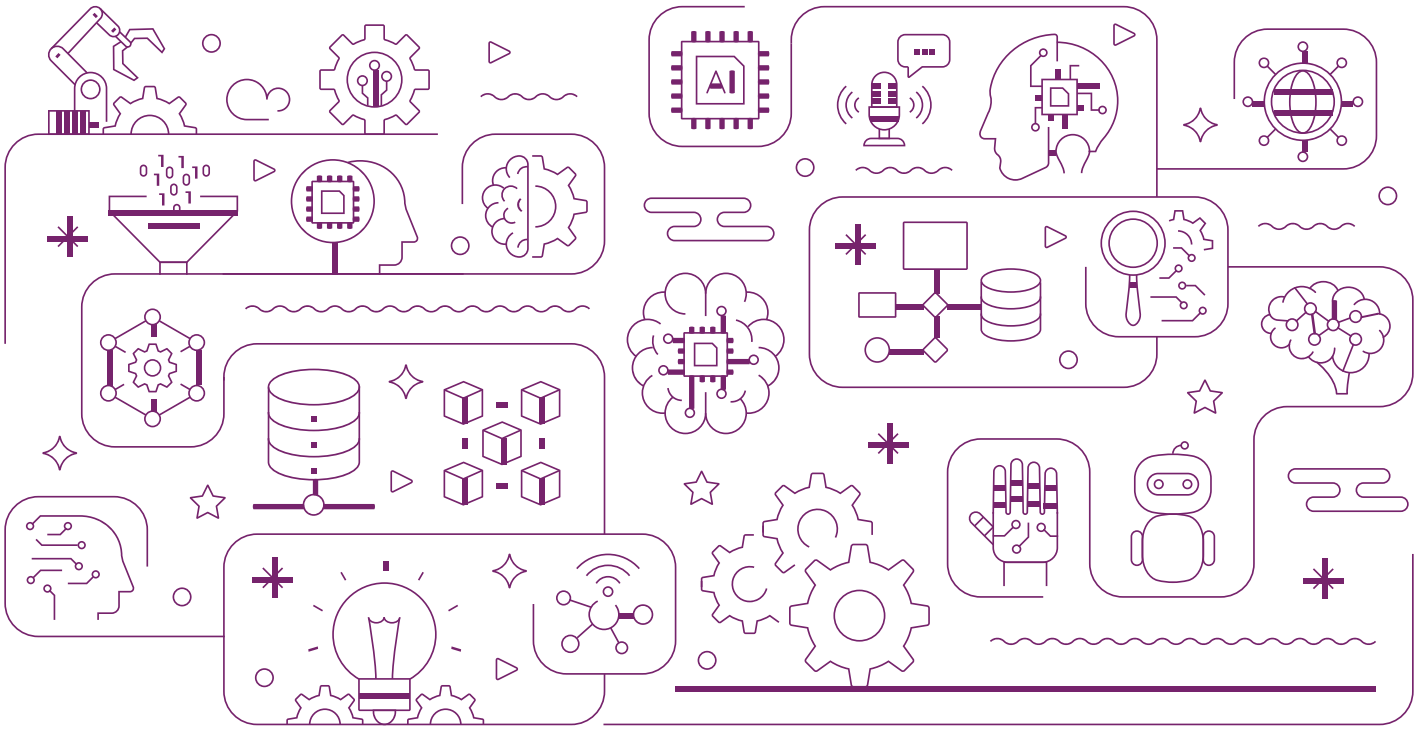
AI 모델 선택 시 다양한 요건의 균형을 맞춰야 합니다. 이에는 모델 설계, 구성, 트레이닝 데이터, 훈련 알고리즘 및 하이퍼파라미터 선택이 포함됩니다. 여러분의 결정은 위협 모델에 따라 결정되며 AI 보안 연구와 위협에 대한 이해가 발전함에 따라 정기적으로 재평가됩니다.

AI 모델 선택 시 여러분이 고려해야 할 사항은 다음을 포함하며, 이에 국한되지는 않습니다:

- 사용 중인 모델의 복잡성, 즉 선택한 설계와 파라미터 수; 모델이 선택한 설계와 파라미터 수는 다른 요소들과 함께, 필요한 트레이닝 데이터의 양과 사용 시 입력 데이터의 변경에 대한 견고성 정도에 영향을 미칩니다
- 사용 사례에 대한 모델의 적합성 및/또는 여러분의 특정 요건에 맞게 조정할 수 있는 가능성(예: 파인튜닝)
- 모델의 아웃풋을 정렬, 해석 및 설명하는 기능(예: 디버깅, 감사 또는 규정 준수); 해석하기가 더 어려운 크고 복잡한 모델보다 더 간단하고 더 투명한 모델을 사용하는 것이 이점이 있을 수 있습니다
- 크기, 무결성, 품질, 민감도, 연령, 관련성 및 다양성을 포함한 트레이닝 데이터 세트의 특성

- 모델 강화(예: 적대적 훈련), 정규화 및/또는 프라이버시 강화 기술 사용의 가치
- 모델 또는 기초 모델, 트레이닝 데이터 및 관련 도구를 포함한 구성 요소의 출처 및 공급 체인

이들 구성 요소 중 몇 개 요소가 보안 결과에 영향을 미치는지에 대한 자세한 내용은 NCSC의 '머신러닝 보안 원칙(Principles for the Security of Machine Learning), 특히 '보안을 위한 설계(모델 설계)(Design for Security (Model Architecture))'를 참조하세요.



2. 안전한 개발

본 부문은 AI 시스템 개발 수명주기 중 **개발** 단계에 해당하는 지침이며, 이에는 공급 체인 보안, 기록 및 자산과 기술 부채(technical debt) 관리 등이 포함됩니다.

공급 체인 안전을 확보하세요

시스템 수명주기 전반에 걸쳐 AI 공급 체인의 보안을 평가 및 모니터링하고 공급업체가 여러분의 기관이 다른 소프트웨어에 적용하는 것과 동일한 표준을 준수하도록 요구합니다. 만약 공급업체가 여러분 기관의 기준을 준수할 수 없는 경우, 여러분의 기존 위험 관리 정책에 따라 행동하세요.

하드웨어 및 소프트웨어 구성 요소가 자체 개발이 아닌 경우 여러분의 시스템의 강력한 보안을 보장하기 위해 검증된 상업용, 오픈 소스 및 기타 타사 개발자로부터 보안과 문서화가 잘 수립된 하드웨어 및 소프트웨어 구성 요소(예: 모델링, 데이터, 소프트웨어 라이브러리, 모듈, 미들웨어, 프레임워크 및 외부 API)를 획득하고 관리합니다.

보안 기준이 충족되지 않은 경우, 업무수행에 가장 필수적인 시스템을 위한 대체 솔루션 자동 전환(failover)할 준비를 합니다. 공급 체인 및 소프트웨어 개발 수명주기 입증을 추적하기 위해 NCSC의 공급 체인 지침(Supply Chain Guidance)과 같은 자료 및 소프트웨어 아티팩트에 대한 공급 체인 수준(Supply Chain Levels for Software Artifacts, SLSA)[10]과 같은 프레임워크를 활용합니다.

여러분의 자산을 파악, 추적 및 보호하세요

모델, 데이터(사용자 피드백 포함), 프롬프트, 소프트웨어, 문서, 로그 및 평가(잠재적으로 안전하지 않은 기능 및 오류 모드에 대한 정보 포함)를 비롯해 AI 관련 자산이 기관에 미치는 가치를 이해하고, 이들 자산에 있어 어떤 부분에서 상당한 투자가 요구되고 접근권한이 생길 경우 어떤 지점에서 외부인으로부터 공격이 가능해지는 지를 파악합니다. 로그를 민감한 데이터로 취급하고 기밀성, 무결성 및 가용성을 보호하기 위한 통제기능을 구현합니다.

여러분의 자산이 어디에 있는지 이해하고 있으며 관련 위험을 평가하고 수용합니다. 자산을 추적, 인증, 버전 통제 및 보호하는 절차와 도구가 있으며 손상 시 여러분이 알고 있는 양호한 상태로 복원할 수 있습니다.

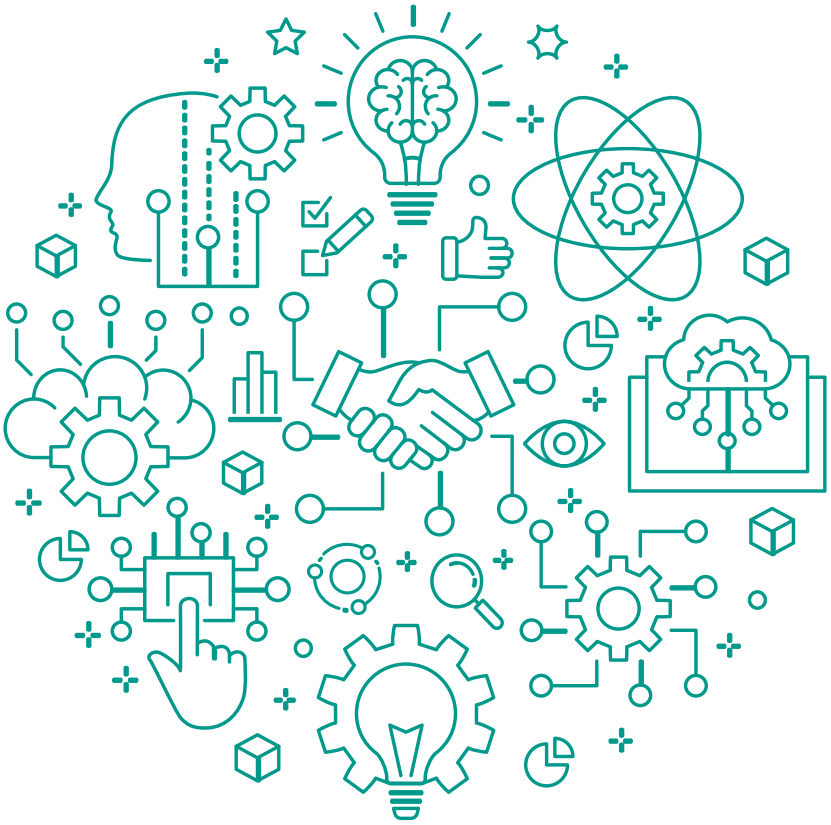
AI 시스템이 접근할 수 있는 모든 데이터를 관리하고 AI가 생성한 콘텐츠를 민감도(및 이를 생성하는 데 사용된 인풋의 민감도)에 따라 관리하기 위한 절차와 통제 기능이 마련되어 있습니다.

데이터, 모델링 및 프롬프트를 문서화하세요

모든 모델, 데이터 세트 및 메타 또는 시스템 프롬프트의 생성, 운영 및 수명주기 관리를 문서화합니다. 이들 문서는 트레이닝 데이터(파인튜닝된 데이터 및 인간 또는 기타 운영 피드백 포함)의 출처, 의도 범위 및 제한, 가드레일, 암호화 해시 또는 서명, 보존 시간, 제안된 검토 빈도 및 잠재적 실패 모드와 같은 보안 관련 정보가 포함됩니다. 이를 수행하는 데 도움이 되는 유용한 구조에는 모델 카드, 데이터 카드 및 소프트웨어 자재 명세서(Software Bills of Materials, SBOM)가 포함됩니다. 포괄적이고 상세한 문서 개발은 투명성과 책임감을 장려합니다[11].

기술 부채(technical debt)를 관리하세요

모든 소프트웨어 시스템과 마찬가지로 같이 AI 시스템의 수명주기 전반에 걸쳐 여러분의 '기술 부채(technical debt)'를 파악, 추적 및 관리합니다. (기술 부채: 장기적인 이익을 희생하며 단기적인 결과를 달성하기 위해 모범 표준에 미치지 못하는 엔지니어링 결정이 내려지는 것) 재정상 부채와 마찬가지로 기술 부채도 본질적으로 나쁜 것은 아니지만 발생 초기 단계부터 관리되어야 합니다[12]. 다만 이것이 표준 소프트웨어보다 AI 환경에서 수행하는 것이 더 어려울 수 있고, 빠른 개발 주기와 잘 확립된 프로토콜 및 인터페이스 부족으로 인해 기술 부채가 많을 가능성이 높다는 점을 이해해야 합니다. 상품의 수명주기 계획(AI 시스템 폐기 절차 포함)이 미래의 유사한 시스템에 대한 위험을 평가, 인식 및 완화하도록 합니다.



3. 안전한 배포

본 부문은 AI 시스템 개발 수명주기 중 **배포** 단계에 해당하는 지침으로 이에는 인프라와 모델을 타협, 위협 또는 손실로부터 보호하는 방법, 사고 관리 절차 개발, 그리고 책임있는 공개 등이 포함됩니다.

안전한 인프라를 갖추세요

시스템 수명주기의 모든 부분에서 사용되는 인프라에 모범 인프라 보안 원칙을 적용합니다. 연구, 개발 및 배포 과정에서 API, 모델, 데이터, 그리고 이에 대한 교육 및 처리 파이프라인에 대한 적절한 접근권 통제를 적용합니다. 이에는 민감한 코드 및 데이터를 보유하는 환경의 적절한 분리를 포함됩니다. 이는 모델을 도용하거나 모델의 성능을 해치는 것을 목표로 하는 일반적인 사이버 보안 공격을 완화하는 데 도움이 될 것입니다.

여러분의 모델을 계속해서 보호하세요

공격자는 모델에 직접 접근하거나(모델 가중치 획득) 간접적으로(애플리케이션이나 서비스를 통한 모델 쿼리) 모델의 기능[13] 또는 모델이 학습한 데이터[14]를 재구성할 수도 있습니다. 또한 공격자는 훈련 중이나 훈련 후에 모델, 데이터 또는 프롬프트를 변경해 아웃풋을 신뢰할 수 없게 만들 수도 있습니다.

모델 및 데이터를 집적적 및 간접적 접근으로부터 보호하는 방법은 다음과 같습니다:

- 표준 사이버 보안 모범 사례 구현
- 기밀 정보에 대한 접근, 수정 및 유출 시도를 탐지하고 방지하기 위해 쿼리 인터페이스에 대한 통제 기능 설정

소비하는 시스템이 모델의 정당성을 입증할 수 있도록 모델이 훈련되자마자 모델 파일(예: 모델 가중치) 및 데이터세트 (체크포인트 포함)의 암호화 해시 및/또는 서명을 산출하고 공유합니다. 늘 그렇듯 암호기술에 있어서 좋은 키 관리는 필수입니다[15].

여러분이 선택할 기밀성 리스크 완화 접근법은 사용 사례와 위협 모델에 따라 크게 달라집니다. 일부 어플리케이션(예: 굉장히 민감한 정보를 다루는 경우)은 적용하기 어렵거나 비용이 많이 드는 이론적 보증이 필요할 수도 있습니다. 적절한 경우, 프라이버시 강화 기술(예: 차분 프라이버시 또는 동형 암호화)을 사용해 모델 및 아웃풋에 접근할 수 있는 소비자, 사용자 및 공격자와 관련된 리스크 수준을 탐색하거나 보장할 수 있습니다.

사고 관리 절차를 수립하세요

AI 시스템에 영향을 미치는 보안 사고의 불가피함을 여러분의 사고 대응, 사안 확대 및 해결 계획에 반영합니다. 이들 계획은 여러 다양한 상황을 설명할 것이며 시스템 및 전반적인 연구가 발전함에 따라 주기적으로 재평가됩니다. 사업 수행에 중대한 디지털 자료는 오프라인 백업에 보관합니다. 대응 직원은 AI 관련 사고를 평가하고 이에 대응할 수 있도록 훈련되어 있습니다. 고객 및 사용자에게 고품질 감사 로그와 기타 보안 기능 또는 정보를 추가 비용 없이 제공해 고객 및 사용자 또한 각자의 사고 대응 절차를 활성화할 수 있도록 합니다.

AI는 책임감 있게 공개하세요

여러분은 벤치마킹 및 레드팀(red teaming) 등 적절하고 효과적인 보안 평가(안전성이나 공정성 등 본 지침의 범위를 벗어나는 기타 테스트 포함)를 거친 후에만 모델, 애플리케이션 또는 시스템을 출시하고, 알려진 제한 사항이나 잠재적인 오류 모드에 대해 사용자에게 명확히 설명합니다. 오픈 소스 보안 테스트 라이브러리에 대한 자세한 정보는 본 문서 말미 '추가 정보'에 제공됩니다.

사용자가 올바른 선택을 쉽게 할 수 있도록 하세요

여러분은 각기의 새로운 설정 또는 구성 옵션이 이들로 인해 파생되는 사업적 이점과 보안 리스크와 함께 평가되어야 한다는 점을 이해합니다. 이상적으로는 가장 안전한 설정이 유일한 옵션으로 시스템에 통합됩니다. 구성이 필요한 경우 디폴트 옵션은 일반적인 위협으로부터 광범위하게 보호되어야 합니다. 즉, 기본적으로 안전함이 보장되어야 합니다. 악의적인 방식으로 여러분의 시스템을 사용하거나 배포하는 것을 방지하기 위한 통제 기능을 적용합니다.

제한 사항 및 잠재적인 오류 모드에 대한 강조를 비롯해 여러분의 모델 또는 시스템의 적절한 사용에 대한 지침을 사용자에게 제공합니다. 사용자에게 보안 측면에 있어 본인이 책임져야 하는 부분들을 명확하게 제시하고 데이터가 사용, 접근 또는 저장될 수 있는 위치(및 방법)를 투명하게 설명합니다(예: 모델링 재교육을 위해 사용되는 경우, 직원 또는 파트너가 검토하는 경우).

4. 안전한 운영 및 관리

본 부문은 AI 시스템 개발 수명주기 중 **안전한 운영 및 관리** 단계에 해당하는 지침을 포함합니다. 이는 로그 및 모니터링, 업데이트 관리, 정보 공유 등 시스템 배포 후 특히 관련된 조치에 대한 지침을 제공합니다.

시스템 작동 상태를 모니터링하세요

여러분의 모델과 시스템의 아웃풋 및 성능을 측정해 보안에 영향을 미치는 행동의 갑작스럽고 점진적인 변화를 관찰할 수 있도록 합니다. 잠재적인 침입과 손상은 물론 자연스러운 데이터 드리프트를 고려하고 식별할 수 있어야 합니다.

시스템의 인풋을 모니터링하세요

프라이버시 및 데이터 보호 요건에 따라 시스템의 인풋 (추론 요청, 쿼리 또는 프롬프트 등)을 모니터링 및 기록을 하여 손상 또는 오용 시 규정 준수 의무, 감사, 조사 및 교정이 가능하도록 합니다. 이에는 데이터 준비 단계(예: 이미지 자르기 및 크기 조정) 활용을 목표로 하는 인풋을 포함하여 배포되지 않은 인풋 및/또는 적대적인 입력정보에 대한 명시적인 탐지가 포함될 수도 있습니다.

업데이트에 대해서는 '안전 설계(secure by design)' 접근법을 따르세요

모든 상품에 자동 업데이트 기능을 기본 설정으로 포함하고 이를 배포하기 위해 안전한 모듈식의 업데이트 절차를 사용합니다. 여러분의 업데이트 절차(테스트 및 평가 제도 포함)는 데이터, 모델 또는 프롬프트의 변경으로 인해 시스템 동작이 변경될 수 있다는 점을 반영합니다(예: 주요 업데이트는 새 버전처럼 취급합니다). 사용자가 모델 변경을 평가하고 이에 대응할 수 있도록 지원합니다(예: 미리보기 권한 및 버전화된 API를 제공).

교훈을 수집하고 공유하세요

적절한 모범 사례를 공유하기 위해 업계, 학계 및 정부의 글로벌 생태계 전반에 걸쳐 협력하며 정보 공유 커뮤니티에 동참합니다. 보안 연구원이 취약점을 연구하고 보고하는 데 동의하는 것을 포함해 기관 내부 및 외부 모두에서 시스템 보안에 관한 피드백을 제공할 수 있는 공개 소통 채널을 유지합니다. 필요한 경우, 문제 사항을 더 넓은 커뮤니티로 확대합니다. (예: 취약점 공개에 대응하기 위한 게시판을 게시하며 이에는 상세하고 완전한 공통 취약점 목록을 포함함) 문제를 완화, 해결하기 위해 빠르고 적절한 조치를 취합니다.

추가 정보

AI 개발

Principles for the security of machine learning
ML 구성 요소가 포함된 시스템의 개발, 배포 또는 운영에 대한 NCSC의 자세한 지침입니다.

Secure by Design - Shifting the Balance of Cybersecurity Risk: Principles and Approaches for Secure by Design Software
CISA, NCSC 및 기타 기관이 공동으로 작성한 이 지침은 AI를 포함한 소프트웨어 시스템 제조업체가 제품 개발의 설계 단계에서 보안을 고려하는 조치를 취하고 제품을 처음부터 안전하게 배송해야 하는 방법을 설명합니다.

AI Security Concerns in a Nutshell
독일 연방 정보 보안국(BSI)에서 제작한 이 문서는 머신러닝 시스템에 대한 가능한 공격과 이러한 공격에 대한 잠재적인 방어 방법을 소개합니다.

Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems 및 Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems
G7 히로시마 AI 프로세스의 일부로 제작된 이들 문서는 전 세계적으로 안전하고 신뢰할 수 있는 AI를 홍보한다는 목표로 가장 발전된 기반 모델과 생성적 AI 시스템을 포함하여 가장 발전된 AI 시스템을 개발하는 기관에 지침을 제공합니다.

AI Verify
표준화된 테스트를 통해 국제적으로 인정된 일련의 원칙에 대해 AI 시스템의 성능을 검증하는 싱가포르의 AI 거버넌스 테스트 프레임워크 및 소프트웨어 프로그램입니다.

Multilayer Framework for Good Cybersecurity Practices for AI — ENISA (europa.eu)
국가 관할 당국과 AI 이해관계자가 AI 시스템, 운영 및 프로세스를 안전하게 하기 위해 따라야 할 단계를 안내하는 프레임워크입니다.

ISO 5338: AI system life cycle processes (Under review)
머신러닝 및 휴리스틱 시스템을 기반으로 AI 시스템의 수명주기를 설명하기 위한 일련의 프로세스 및 관련 개념입니다.

AI Cloud Service Compliance Criteria Catalogue (AIC4)
BSI의 AI 클라우드 서비스 규정 준수 기준 카탈로그는 AI 서비스의 수명주기 전반에 걸쳐 보안을 평가할 수 있는 AI 관련 기준을 제공합니다.

NIST IR 8269 (Draft) A Taxonomy and Terminology of Adversarial Machine Learning
머신러닝과 휴리스틱 시스템을 기반으로 AI 시스템의 수명주기를 설명하기 위한 일련의 절차 및 관련 개념입니다.

MITRE ATLAS
MITRE ATT&CK 프레임워크를 모델링하고 연결한 머신러닝(ML) 시스템에 대한 적의 전술, 기술 및 사례 연구에 대한 지식 풀입니다.

An Overview of Catastrophic AI Risks (2023)
AI 안전센터에서 제작한 이 문서는 AI로 인해 발생하는 위험 영역을 설명합니다.

대형 언어 모델(Large Language Model): Opportunities and Risks for Industry and Authorities
LLM 개발, 배포 및/또는 사용의 기회와 위험에 대해 더 자세히 알고 싶어하는 회사, 당국 및 개발자를 위해 BSI에서 제작한 문서입니다.

사용자가 AI 모델에 대한 보안 테스트를 수행하는 데 도움이 되는 오픈 소스 프로젝트에는 다음이 포함됩니다:

- Adversarial Robustness Toolbox (IBM)
- CleverHans (University of Toronto)
- TextAttack (University of Virginia)
- Prompt Bench (Microsoft)
- Counterfit (Microsoft)
- AI Verify (Infocomm Media Development Authority, Singapore)

사이버 보안

CISA's Cybersecurity Performance Goals
알려진 위험과 공격 기술의 가능성과 영향을 의미 있게 줄이기 위해 모든 핵심 인프라 주체가 구현해야 하는 공통 보호 장치입니다.

NCSC CAF Framework
CAF(사이버 평가 프레임워크)는 매우 중요한 서비스 및 활동을 담당하는 기관에 지침을 제공합니다.

MITRE's Supply Chain Security Framework
공급 체인 내 공급업체 및 서비스 제공업체를 평가하기 위한 프레임워크입니다.

리스크 관리

NIST AI Risk Management Framework (AI RMF)
AI RMF는 AI와 특정하게 관련된 개인, 기관 및 사회에 대한 사회 기술적 위험을 관리하는 방법을 설명합니다.

ISO 27001: Information security, cybersecurity and privacy protection
이 표준은 기관에 정보 보안 관리 시스템의 구축, 구현 및 유지 관리에 대한 지침을 제공합니다.

ISO 31000: Risk management
기관 내 리스크 관리에 대한 지침과 원칙을 기관에 제공하는 국제 표준입니다.

NCSC Risk Management Guidance
이 지침은 사이버 보안 위험 실무자가 자신의 기관에 영향을 미치는 사이버 보안 위험을 더 잘 이해하고 관리하는 데 도움이 됩니다.

참조

1. AI 시스템을 개발하고 (또는 개발된 AI 시스템을 보유하는) 해당 시스템을 자체 이름이나 상표로 시장에 출시하거나 서비스에 제공하는 사람, 공공 기관, 기관 또는 기타 기관으로 정의됩니다
2. 안전 설계에 대한 더 자세한 내용은 CISA의 Secure by Design 웹페이지 및 지침 Shifting the Balance of Cybersecurity Risk: Principles and Approaches for Secure by Design Software
3. 규칙 기반 시스템과 같은 비 ML AI 접근 방식과 반대
4. CEPS는 출판물 'Reconciling the AI Value Chain with the EU's Artificial Intelligence Act'에서 7가지 유형의 AI 개발 상호 작용을 설명합니다
5. ISO/IEC 22989:2022(en)는 이를 'AI 시스템을 구성하는 기능적 요소'로 정의합니다
6. NIST는 인공지능(AI)의 안전하고 신뢰할 수 있는 개발 및 사용을 발전시키기 위한 지침을 작성하고 기타 조치를 취하는 임무를 맡고 있습니다. NIST's Responsibilities Under the October 30, 2023 Executive Order 참조
7. 위협 모델링에 대한 더 자세한 정보는 OWASP Foundation에서 확인할 수 있습니다
8. MITRE ATLAS의 Adversarial Machine Learning 101를 참조하세요
9. GitHub: RCE PoC for Tensorflow using a malicious Lambda layer
10. SLSA: 'Safeguarding artifact integrity across any software supply chain'
11. METI (Japanese Ministry of Economy, Trade and Industry, 2023), 'Guide of Introduction of Software Bill of Materials (SBOM) for Software Management'
12. 구글 연구: 머신러닝: The High Interest Credit Card of Technical Debt
13. Tramèr et al 2016, Stealing Machine Learning Models via Prediction APIs
14. Boenisch, 2020, Attacks against Machine Learning Privacy (Part 1): Model Inversion Attacks with the IBM-ART Framework
15. National Cyber Security Centre, 2020, Design and build a privately hosted Public Key Infrastructure