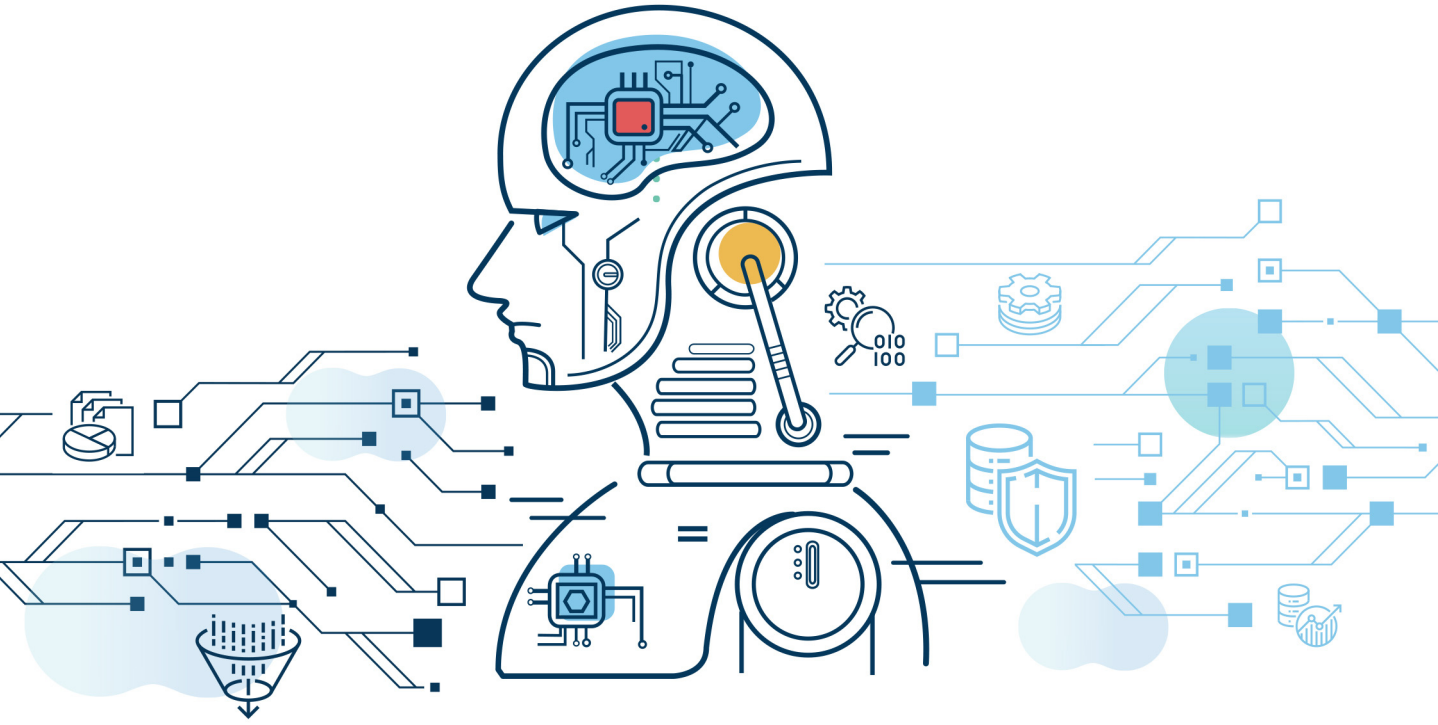
발간번호	AGR – VIII – 2023 – ② – 405
발간일자	2023년 4월

금융미래를 열어가는 금융보안파트너

금융분야 AI 보안 가이드라인

2023. 4.



금융보안원
FINANCIAL SECURITY INSTITUTE

목 차

제1장

가이드라인 개요

1. 추진 배경 및 목적
2. AI 개요
3. 가이드라인 구성
4. 용어 정리

제1장

가이드라인 개요

본 가이드라인은 금융권에서 사용하는 인공지능(AI) 서비스의 안전한 활용에 필요한 사항을 안내하여, 금융회사의 AI 활용에 있어 안전성을 높이고자 한다.

1 추진 배경 및 목적

■ **최근 디지털 금융산업의 발전과 함께, 인공지능(AI)을 활용한 서비스의 도입이 확대됨에 따라 개인정보 유출 등 다양한 보안 위협이 발생하고 있다.**

참고 **인공지능 관련 사고 사례**

구 분	특 징
• AI 챗봇 '이루다'('21.1.)	• 실명, 계좌번호, 주소 등 개인정보 유출 발생 • 특정 소수자 차별 등으로 출시 1달여 만에 서비스 일시 중단
• AI 헬스케어 'GPT-3'('20.10.)	• 정신과 챗봇으로 출시전 모의 환자에게 자살 권유
• AI 성별식별 '젠더리파이'('20.7.)	• 여성, 유색인종, 노인 등 소수자 차별
• MIT의 '사이코패스AI'('18.6.)	• 연구를 통해 부정·편향된 인공지능 학습 결과, 반사회·반인륜적인 행위 증명

■ **AI 보안의 중요성은 날로 커지고 있으며, AI의 보안성을 확보하기 위한 국내·외 다양한 정책이 발표되었다.**

참고 **국내·외 AI 보안 관련 주요 정책**

국가	주요 정책
EU	• 인공지능 규제 샌드박스 발표('22.6월)
	• AI 규제법안('21.4월)
미국	• 디지털 플랫폼 위원회법 발의('22.5월)
	• 미국 데이터 프라이버시와 보호 법('22.5월)
한국	• 금융권 인공지능(AI) 활용 활성화 및 신뢰확보 방안('22.7월, 금융위)
	• 신뢰할 수 있는 인공지능 실현 전략('21.5월, 과기부)

- 이러한 정책은 AI의 보안성을 확보하기 위한 규제·원칙을 다루고 있으나, 실무 위주의 세부적 안내 또한 필요한 실정이다.
- 특히, 금융 분야는 고객의 재산과 직접적 관련이 있어 AI 보안성 확보를 위한 실무 위주의 안내가 더욱 중요하다.

■ **본 가이드라인은 AI 실무자가 AI 보안에 대해 쉽게 이해하고 금융회사 등이 자체적으로 AI 보안성을 확인할 수 있도록 구성한다.**

- AI 서비스의 전반적 구조를 보안성 관점에서 안내한다.
- AI 학습 데이터 및 모델 관리를 보안성 관점에서 안내한다.
- AI 챗봇 서비스 보안성 점검 항목과 AI 모델에 대한 금융회사의 자체점검 결과 예시를 안내한다.

2 AI 개요

■ **AI는 기계 혹은 시스템상에서 만들어지는 지능으로서 인간의 지능과 유사한 지적 능력을 인공적으로 구현한 것이다.**

- **(머신러닝)** AI 방식 중 하나로서 인간의 명령이 아닌 기계 스스로 데이터로부터 학습하여 찾아낸 패턴에 따라 작업을 실행하는 알고리즘이다.
- **(딥러닝)** 머신러닝 방식 중 하나로서 뇌 신경망을 모방한 인공신경망을 통해 데이터를 학습한 AI 모델이다.

[AI의 분류]

■ **AI의 학습 방식으로서는 지도학습, 비지도학습, 강화학습으로 분류할 수 있다.**

- **(지도학습)** 정답이 주어진 상태에서 AI가 학습하는 방식이다.
- **(비지도학습)** 정답을 알려주지 않은 상태에서 AI가 스스로 데이터의 특성을 학습하고 패턴을 파악하는 방식이다.
- **(강화학습)** 정해진 환경 안에서 동작하는 에이전트가 현재의 상태를 인식하고, 선택할 수 있는 행동 중 보상을 최대화하는 행동을 선택하여 학습하는 방식이다.
  - 과거에도 존재하던 학습 방식이었으나 딥러닝과 알파고 등장 이후 더 크게 주목

[AI의 학습 방식]

지도 학습 (Supervised Learning)	비지도 학습 (Unsupervised Learning)	강화 학습 (Reinforcement Learning)
• 정답이 주어진 상태에서 학습 • 주로 예측, 분류 등에 활용	• 정답을 알려주지 않은 상태에서 학습 • 주로 연관 규칙 탐색, 군집화 등에 활용	• 보상을 최대화하는 방향으로 학습 • 바둑이나 게임처럼 환경·상태·보상 등이 있는 경우 활용

3 가이드라인 구성

■ **본 가이드라인은 아래와 같은 구성을 통해 안내한다.**

- **(AI 서비스 구성)** AI 서비스의 기능별로 안내한다.
  - **(운영계)** AI 서비스 운영에 필요한 구성요소로 '채널', '애플리케이션 서버', 'AI 엔진' 등으로 구분하고 각 요소에 대해서 안내한다.
  - **(개발계)** AI 서비스 개발에 필요한 구성요소로 '데이터 수집', '데이터 처리', '개발·테스트 엔진'에 대해서 안내한다.
- **(AI 학습 데이터 및 모델 관리)** AI 모델의 각 개발 주기* 별로 보안 고려사항에 대해서 안내한다.

  * AI 모델 개발을 위한 데이터 수집부터 검증·테스트에 이르기까지의 과정
  - '학습 데이터 수집', '학습 데이터 전처리', 'AI 모델 설계·학습', 'AI 모델 검증·평가'의 각 단계에 대해 보안 고려사항을 안내한다.
- **(AI챗봇서비스 보안성 체크리스트)** 금융회사 등이 AI챗봇 서비스의 보안성을 자체적으로 점검할 수 있도록 점검항목을 안내한다.

4 용어 정리

- ‘AI 서비스’란 AI 기능을 활용하여 작동하는 프로그램 또는 시스템을 말한다.
- ‘AI 알고리즘’이란 데이터의 규칙·패턴을 해석하거나 지식을 추론할 수 있게 만든 방법 또는 절차를 말한다.
- ‘AI 모델’이란 ‘AI 알고리즘’에 데이터를 학습하여 산출된 규칙 또는 수식을 말한다.
- ‘적대적 예제’란 AI 모델이 잘못된 예측을 하도록 의도적으로 조작한 데이터를 말한다.
- ‘적대적 공격’이란 적대적 예제를 활용하여 AI 모델이 잘못 판단하도록 유도하는 공격을 말한다.

제2장

AI 서비스 구성

1. AI 서비스 구성
2. 운영계
3. 개발계

제2장

AI 서비스 구성

본 장에서는 보안 관점에서 AI 서비스의 구성요소와 각 기능에 대해서 안내한다.

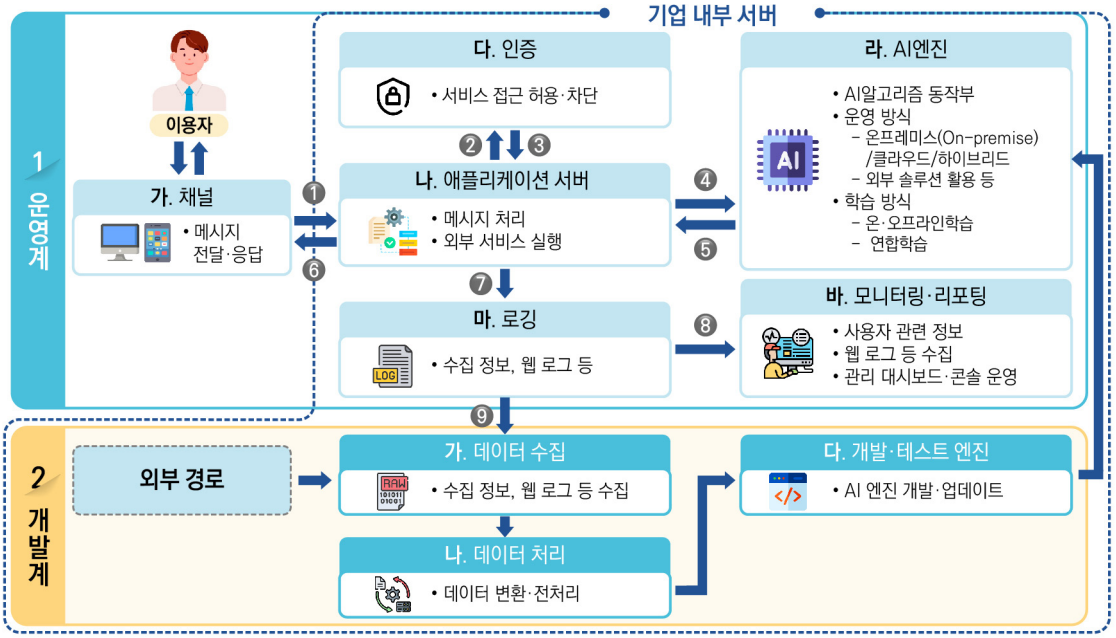
1 AI 서비스 구성

■ **AI 서비스를 원활히 운영하기 위해서는 이용자의 채널(메시지 입력), AI 엔진(AI 모델 작동) 등 다양한 구성 요소가 필요하다.**

- 아래 구성도와 같이, AI 서비스의 구성 요소에 대해서 개념적으로 표현할 수 있으며, 각 요소에 대해 안내한다.

  ※ 서비스 목적·기능, 금융회사 등의 환경에 따라 기능 및 세부설계가 다를 수 있음

[AI 서비스 구성도]



2 운영계

가. 채널

■ **AI 서비스를 이용하는 고객이 접근하는 경로로서 웹페이지·애플리케이션·전화·마이크 등의 다양한 매체가 활용된다.**

- 금융권의 대표적인 AI 서비스인 챗봇은 웹페이지, 메신저 앱, 자체 개발 앱 등의 채널을 이용하여 상담·안내 등의 서비스를 고객에게 제공한다.
- 신용평가나 여신심사와 같이 백오피스 업무에 AI가 활용되는 경우는 고객 등의 이용자가 AI에 직접 접근하는 것이 어려울 수도 있다.

나. 애플리케이션 서버

■ **이용자의 요청 메시지에 따라 관련 서비스를 호출하거나 AI 엔진에 메시지를 전달하는 역할을 한다.**

- 애플리케이션 서버는 이용자의 메시지를 해석하기 전 인증 절차를 통해, 서비스 이용에 적합한 이용자인지 확인할 수 있다(❶~❸과정).
- 애플리케이션 서버는 이용자의 메시지에 따라서 데이터베이스(이하 “DB”)조회, 메시지 가공, 외부 서비스 실행 등 다양한 작업을 수행한다.
- 단순 AI 연산, 필터링 등 간단한 AI 기능은 애플리케이션 서버에서 직접 처리하도록 설계할 수도 있다.

■ **애플리케이션 서버는 기업 내부의 서버이며, AI가 적용된 서비스의 전반적인 운영을 하는 역할로 안전하게 보호할 필요가 있다.**

- 채널과 애플리케이션 구간(❶ 과정)은 (웹)방화벽, 침입탐지(차단)시스템, Anti-DDoS, HTTPS, VPN 등의 정보보호 대책을 고려할 수 있다.

다. 인증

■ **이용자의 신원확인·증명을 통하여 서비스 접근 여부를 판단하는 요소로서, AI 서비스의 목적 및 위험도에 따라서 다양한 인증수단을 활용할 수 있다.**

- 대표적으로 패스워드(PIN* 코드 포함), 문자메시지(SMS) 인증, 지문인식 등을 활용할 수 있으며, 채널에서 인증을 수행토록 설계할 수도 있다.

  * 개인식별번호(Personal Identity Number)

참고 **인증 유형**

구분(Factor)	설명	종류
지식 기반	지식 기반의 인증	비밀번호, PIN 번호 등
소유 기반	소유품 기반의 인증	공인인증서, OTP, SMS 인증 등
속성 기반	고유 속성 기반의 인증	지문, 홍채 등
멀티 팩터 기반	2가지 이상의 인증수단 결합	비밀번호+지문, OTP+홍채 등

라. AI 엔진

■ **AI 모델이 실질적으로 동작하는 요소로서, 애플리케이션 서버로부터 전달받은 메시지를 해석하고 해석한 메시지를 애플리케이션 서버에 다시 전달한다(❹~❺과정).**

■ **AI 엔진은 기업 내부 시스템으로 구현(온프레미스 방식)하거나 클라우드 및 하이브리드* 방식으로도 구현할 수 있다.**

* 온프레미스 방식과 클라우드 방식 혼용

- AI 모델이 채택한 알고리즘의 종류에 따라 많은 연산이 요구되거나, 유동적인 자원 운용이 필요한 경우는 AI 엔진만 클라우드에 구현하는 방식도 있다.
- AI 엔진 개발에는 전문적인 기술과 역량이 요구되며, 개발 자체에 많은 시간이 소요되므로 자체 개발이 아닌 외부 솔루션을 활용하는 사례가 많다.
  - 국내·외 기업에서 PaaS*, MLaaS** 및 패키지 등 다양한 형태의 AI 엔진을 제공하고 있다.

    * PaaS(Platform-as-a-Service) : 하드웨어·소프트웨어 플랫폼을 한 번에 제공하는 형태의 서비스

    ** MLaaS(Machine Learning as a Service) 딥러닝 설계·개발, 시각화, 예측 등 도구를 클라우드 기반 서비스로 제공

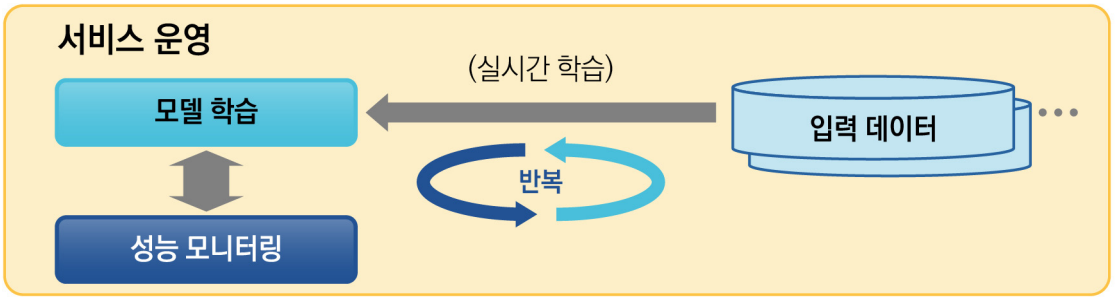
■ **AI 엔진의 학습 방식은 실시간 학습 여부에 따라서 온라인 학습 방식과 오프라인 학습 방식으로 구분할 수 있다.**

참고 **온·오프라인 학습 방식**

✓ **(온라인 학습)** AI 서비스 운영 과정에서, 이용자의 입력 메시지를 실시간으로 학습·반영하여 모델의 성능을 높이는 학습 방식

- AI 서비스의 성능을 지속적으로 향상시키고 변화에 빠르게 적응
- 입력 데이터에 대한 검증이 요구되기 때문에 높은 수준의 신뢰성이 필요한 AI 서비스에는 부적합
- 공격 탐지 기법 적용 및 지속적인 모니터링 요구

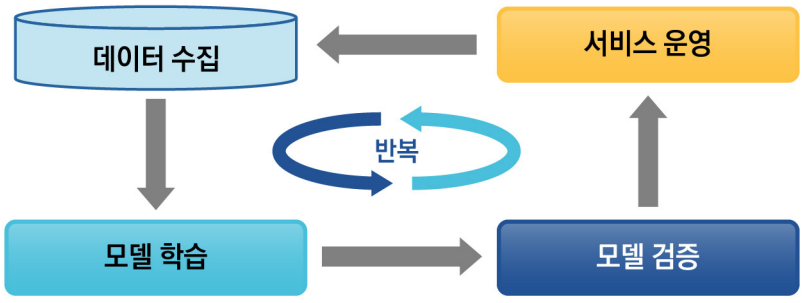
〈 온라인 학습 프로세스 〉



✓ **(오프라인 학습)** 서비스 출시 후에는 실시간 학습 없이 학습된 모델을 이용하여 서비스 운영

- 재학습의 경우는 새로운 데이터를 학습한 모델을 검증 후에 서비스에 적용

〈 오프라인 학습 프로세스 〉



■ **한편, 개인정보 이슈 및 응답 시간 등을 고려하여 연합학습*을 하도록 AI 엔진을 설계할 수도 있다.**

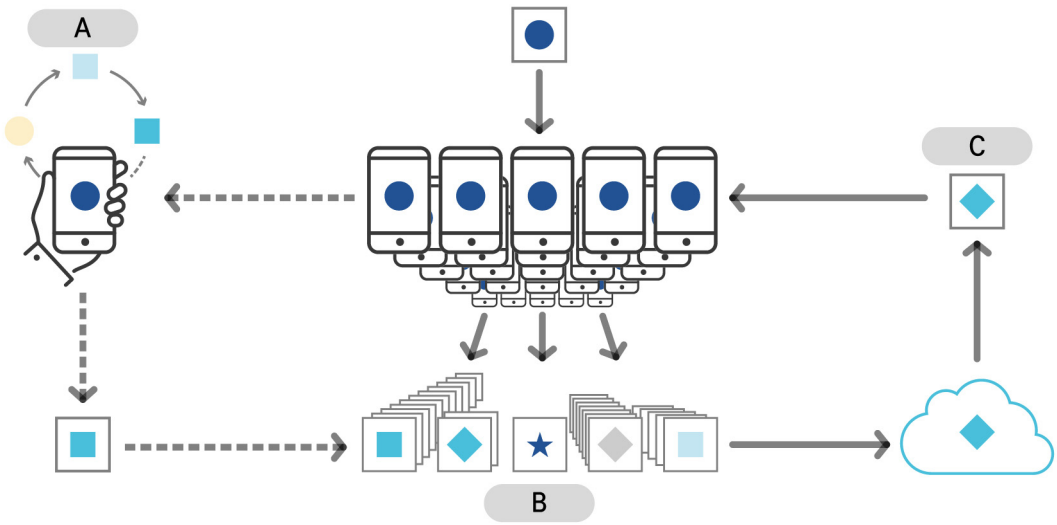
참고 *** 연합학습**

✓ **(개요)** 클라이언트에서 자신의 데이터로 학습하고, 중앙서버로 AI 모델의 매개변수를 전달하고 중앙서버에서 AI 모델을 학습하는 방식

- 연합학습은 데이터를 전달하는 것이 아닌, 학습된 결과를 전달하는 방식
- **(개인정보 처리이슈)** 개인정보를 중앙서버로 직접 전달하는 것이 아니기 때문에, 개인정보와 관련한 다양한 이슈로부터 자유로움
- **(응답시간)** 기존 학습 방식에서의 E2E* 방식이 아닌, 서버에서 전달받은 AI 학습 매개변수를 기반으로 클라이언트가 즉시 결과 도출 가능

* End to End(서버와 클라이언트의 종단 간 호출·응답 방식)[1)]

〈 연합학습 프로세스 〉



A. 개인 단말기에서 개인 저장 데이터를 학습하고 중앙서버로 매개변수 전송
B. 전송된 매개변수로 중앙서버의 글로벌 AI 모델 학습
C. 개인 단말기로 글로벌 AI 모델의 매개변수 전달

1) 구글 공식 블로그(https://ai.googleblog.com/2017/04/federated-learning-collaborative.html)

마. 로깅

■ **이용자의 메시지, AI 모델 학습 관련 데이터, 웹/시스템 로그 등의 데이터를 수집하는 단계이다.(❼ 과정)**

- 관리자는 웹/시스템 로그, AI 엔진 출력값 등의 로그를 모니터링 또는 대시보드 등에 적용하여 운영에 참고할 수 있다.(❽ 과정)
- 또한, 로그 데이터는 AI 모델 학습 데이터로 활용할 수 있으며, 이를 통해 AI 모델의 성능을 향상시킬 수 있다.(❾ 과정)

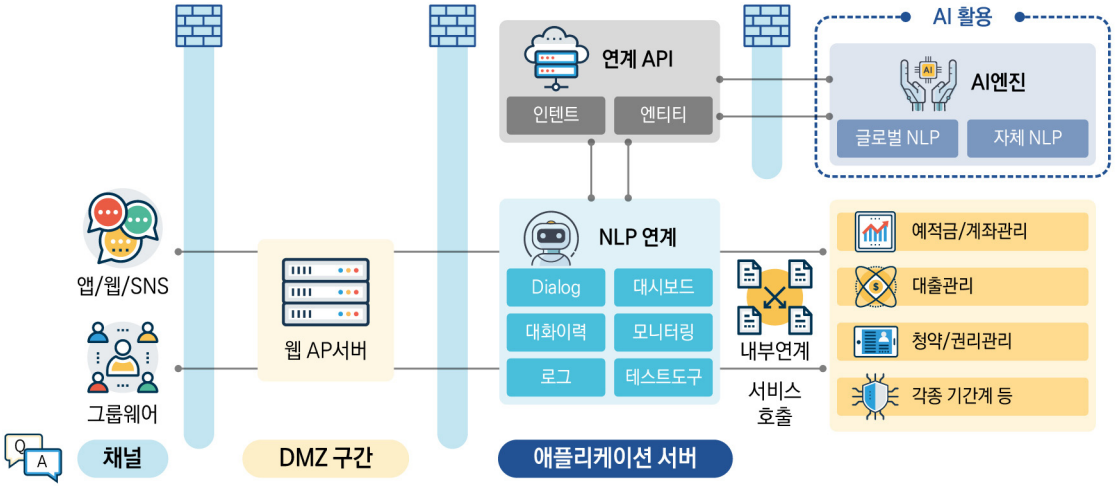
바. 모니터링·리포팅

■ **관리자는 대시보드 등을 활용하여 AI 서비스의 운영상태, AI 모델의 성능 등을 모니터링 할 수 있다.**

- 웹/시스템 로그 등 서비스와 관련된 로그 정보를 기반으로 AI 서버 상황, 공격 여부, 이용자 현황 등을 파악할 수 있다.
- AI 서비스의 성능을 모니터링하고, AI 엔진 재학습 등을 판단할 수 있는 지표로 활용할 수 있다.

참고 **AI챗봇서비스 활용예시**

〈 챗봇서비스 구성도 〉



❶ **(AI 활용)** 챗봇 사용자의 문장(질문)에서 의도를 정확히 이해·분류하는 자연어처리*(NLP)에 활용하며, 챗봇 전문업체나 금융회사가 개발한 NLP, 또는 글로벌 클라우드 NLP를 연계 사용

* 컨텐츠 분류, 주제 발견 및 모델링, 맥락 추출, 음성↔문자간 변환(STT,TTS) 등

- AI 엔진은 내부 시스템(온프레미스)에 구축 비율이 높으나, 자체 클라우드로 전환하거나 API 연계를 통해 클라우드 기반 PaaS, SaaS 서비스 이용 사례도 다수 존재
- '잘 알아듣는 챗봇'을 만들려면 AI 엔진의 정확도를 향상시키기 위해 시나리오에 따라 인텐트(의도)와 엔티티(핵심 개체), 답변 시나리오 집합 등을 지속 수정·보완한 후 AI 엔진을 (재)학습

❷ **(챗봇 용도 및 방식)** 금융 상품 소개·상담용, 내부 직원용, 금융 웹/앱과 통합한 형태의 전자금융서비스용 등 다양한 용도로 구축·운영

- 대부분 질의·응답형을 사용하며 시나리오에 의해 챗봇이 사용자의 질문을 정형화하고, 키워드를 매칭시켜 사전 준비한 답변을 출력

  * '인공지능형 챗봇'은 머신러닝을 활용하여 챗봇이 사용자의 질문을 스스로 학습하여 질문에 답변

- 챗봇서비스 채널은 홈페이지 웹 및 금융 앱 등과 통합된 형태가 증가하며, 카카오톡과 같은 SNS 메신저를 포함한 멀티 채널도 제공

❸ **(AICC 확대)** AI를 활용한 콜센터 디지털전환, AI 기반 컨택센터로 점차 확대

* 챗봇의 NLP와 음성봇(음성인식기술, STT)을 고도화

❹ **(발전 방향)** 현행 '자주하는 질문에 답하는' 챗봇 → 향후 '고객 학습 선순환 구조로 과거 상담을 기억·분석하여 사용자의 배경(Context)을 이해하는 개인화 및 스스로 응답'하는 AI상담원

※ [별첨] 'AI챗봇서비스 보안성 체크리스트' 참고

3 개발계

가. 데이터 수집

■ **데이터 수집은 AI 모델에 학습할 데이터를 수집하는 과정으로 업무 및 서비스 목적에 부합하는 기업 내·외부의 다양한 데이터를 활용한다.**

■ **수집되는 데이터의 경로로는 채널로부터 수집되는 정보, 시스템과 관련된 시스템·웹 로그, 외부 경로가 있다.**

- **(채널 데이터)** 이용자의 메시지, 이용·행위 정보 등 채널로부터 AI 모델 학습에 필요한 데이터를 수집할 수 있다.
- **(시스템 관련 로그)** 일반적으로 시스템 운영 시에 생성되는 로그로 해당 로그를 통하여 시스템 장애, 성능 이슈, 공격 탐지 등을 할 수 있다.
- **(외부 경로)** 인터넷 또는 그 밖에 외부 공개된 소스로부터 수집된 데이터로 AI 엔진 개선 등의 목적으로 활용할 수 있다.

■ **데이터는 형식이나 구조에 따라 정형 데이터, 반정형 데이터, 비정형 데이터로 구분할 수 있다.**

- **(정형 데이터)** 표, 테이블 등 미리 정해놓은 형식·구조에 따라서 저장한 데이터이다.
  - 일반적으로 스키마 구조*로 이루어져 있어, 일정 규칙에 따라서 데이터가 구성되어 있다.

    * 데이터베이스(DB)가 대표적이며, 데이터의 구조·제약조건 등을 정의
- **(비정형 데이터)** 정해진 형식·구조 등의 규칙이 없는 데이터이다.
  - 말뭉치*, 영상 데이터 등이 해당하며, 일정한 규칙 없이 구성되어 있다.

    * 챗봇 발화 데이터, 소셜미디어, 뉴스 데이터 등
- **(반정형 데이터)** 정형 데이터로 이루어져 있지는 않지만, 일정한 규칙에 따라 저장된 데이터이다.
  - JSON* 포맷과 같이 일정 포맷은 존재하지만, 스키마 구조로 이루어져 있지 않다.

    * 자바스크립트 문법으로 구조화된 데이터를 표현하기 위한 표준 포맷

참고 **데이터 유형에 따른 수집 방식**

데이터 유형	데이터 종류	수집 방식
정형 데이터	관계형 DB(RDB), CSV 등	파일 다운로드, 오픈API 등
비정형 데이터	소셜(SNS), 이미지, 오디오, 문서, 영상, IoT 등	크롤링(Crawling), 스크래핑(Scraping), 오픈API, 스트리밍(Streaming), RSS 등
반정형 데이터	HTML, XML, JSON, 웹문서, 웹로그, 센서 데이터 등	크롤링(Crawling), 스크래핑(Scraping), RSS, 오픈API 등

나. 데이터 처리

■ **데이터를 통하여 유의미한 결과를 도출하기 위해서는 데이터의 종류·형태에 따라 적절한 처리 과정이 필요하다.**

- 적절한 처리 과정을 거치지 않은 데이터를 학습에 활용할 경우, AI 모델이 왜곡된 기능을 할 수 있다.

■ **학습 데이터 처리 시 다음과 같은 기법을 활용하여 학습에 불필요한 요소를 제거하고 데이터를 단순화한다.**

- **(클리닝)** 데이터로부터 불필요한 요소를 제거한다.
  - 결측치 대체*, 노이즈 데이터의 평활(smoothing)**, 이상치 관리, 불일치 값 처리 등

    * 삭제, 공백, 평균치 등으로 대체

    ** 관측치를 평균값으로 대체하고 하나의 근사함수로 표현
- **(통합)** 중복이거나 유사 값을 하나의 값으로 통합한다.
- **(변환)** 정규화, 집합화, 요약, 계층 생성 등의 기법을 사용하여 학습에 적합한 값으로 변환한다.
- **(축소)** 학습 데이터의 칼럼을 통합·제거하는 차원 축소 기법을 활용하거나, 이산화* 등의 기법을 활용하여 데이터를 축소한다.

  * 수치값을 범주화하거나 속성값으로 변환

참고 **데이터 전처리 시에 처리해야 할 요소**

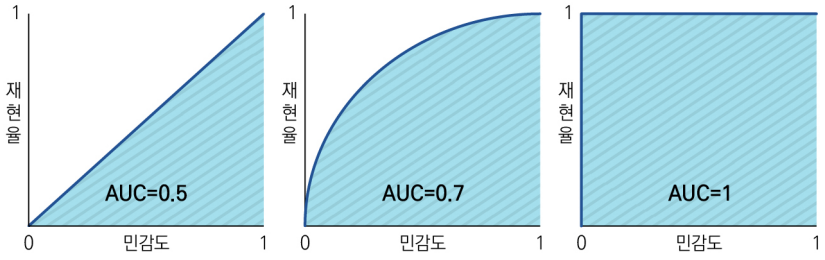
항목	설명
노이즈 (Noise)	• 측정 과정에서 무작위로 발생하는 측정값의 오류
이상치 (Outlier)	• 나머지 데이터와 현저히 다른 특성을 보이는 값 • 데이터 입력·측정 오류/실험 오류로 발생할 수 있지만, 일부 예외 특성을 갖는 값일 수 있음
결측치 (Missing Value)	• 전산오류 및 미입력 등의 이유로 누락된 측정값
불일치 값 (Mismatch Value)	• 동일 개체에 있어, 측정 데이터가 다르게 나타나는 경우
중복 (Duplicate)	• 모든 속성 및 값이 동일한 경우
바이어스 (Bias)	• 측정 장비에서 측정하는 값과 실제 값과의 차이점
아티펙트 (Artifact)	• 외부 요인으로 인해 반복적으로 발생하는 왜곡이나 에러 ※ (예시) 카메라를 이용한 영상 데이터 획득에 있어, 렌즈의 얼룩에 의해 지속적인 왜곡 발생 등
오염 (Poisoning)	• 악의적인 목적으로 변조한 데이터

다. 개발·테스트 엔진

■ **개발·테스트 엔진은 운영 중인 AI 모델의 고도화·개선 및 성능을 테스트하기 위한 시스템이다.**

- '나. 데이터 처리' 단계에서 정제된 데이터를 기반으로 AI 모델의 성능을 테스트할 수 있다.
- AI 모델의 성능을 확인할 수 있는 대표적인 지표로는 정확도(Accuracy), 오차행렬(Confusion Matrix), F-점수(F-score), AUC-ROC 커브(AUC-ROC Curve) 등이 있다.

참고 AI 모델 평가지표

<table>
<tr><th>평가지표</th><th>설명</th></tr>
<tr><td>정확도
(Accuracy)</td><td>• 예측한 데이터셋 중에서 정확하게 예측한 비율</td></tr>
<tr><td>오차행렬
(Confusion Matrix)</td><td>• 예측값와 실제값 사이에 탐지값를 측정한 표
<table>
<tr><th colspan="2" rowspan="2"></th><th colspan="2">예측값</th></tr>
<tr><th>긍정
(Positive)</th><th>부정
(Negative)</th></tr>
<tr><td rowspan="2">실제 값</td><td>긍정
(Positive)</td><td>참양성(TP)
(True Positive)</td><td>거짓음성(FN)
(False Negative)</td></tr>
<tr><td>부정
(Negative)</td><td>거짓양성(FP)
(False Positive)</td><td>참음성(TN)
(True Negative)</td></tr>
</table>
* 참양성(TP) : 실제 정답을 정답으로 예측한 수
** 참음성(TN) : 실제 정답을 오답으로 예측한 수
*** 거짓양성(FP) : 실제 오답을 정답으로 예측한 수
**** 거짓음성(FN) : 실제 오답을 오답으로 예측한 수</td></tr>
<tr><td>정밀도
(Precision)</td><td>• True 예측 결과에서 실제 True를 탐지한 비율
※ TP / (TP+FP)</td></tr>
<tr><td>재현율
(Recall)</td><td>• 실제 True 값에서 True를 탐지한 비율
※ TP / (TP+FN)</td></tr>
<tr><td>F-점수
(F-score)</td><td>• 정밀도와 재현율의 조화평균
※ 2×정밀도×재현율 / (정밀도+재현율)</td></tr>
<tr><td>AUC-ROC 곡선
(AUC-ROC Curve)</td><td>• 재현율와 특이도*를 기준으로 그려지는 곡선
* TN/(TN+FP)
- AUC 수치가 높을수록 모델의 성능이 좋음
 
</td></tr>
<tr><td>평균절대오차
(MAE)</td><td>• 정답과 에러 간의 절대거리를 평균한 거리
- 평균절대오차 수치가 작을수록 모델의 성능이 좋음</td></tr>
</table>

■ **AI 모델의 성능을 측정하고, 과적합*과 과소적합**을 방지하기 위해서 검증 절차를 수행한다.**

* 과적합(Overfitting) : AI 모델이 주어진 학습 데이터를 과하게 학습하여 학습하지 않은 다른 데이터셋에 대하여 예측력이 떨어지는 현상

** 과소적합(Underfitting) : AI 모델이 학습 데이터를 충분히 학습하지 않아 예측력 자체가 떨어지는 현상

- 홀드아웃 검증(Holdout Validation)은 전체 데이터셋을 3개로 분할하여 학습·테스트 및 검증을 각기 다른 데이터셋을 통해서 진행하도록 하는 것을 의미한다.
- K폴드 교체검증(K-fold Cross Validation)은 학습 데이터를 K개로 분할한 후, 차례대로 하나씩 검증 데이터셋으로 활용하여 총 K회의 검증을 실시하는 것을 의미한다.

제3장

AI 학습 데이터 및 모델 보안 관리

제3장

AI 학습 데이터 및 모델 보안 관리

본 장에서는 보안 관점에서 AI 모델의 개발주기별로 AI 학습 데이터 및 모델을 안전하게 관리하는 방안에 대해서 안내한다.

■ **AI 모델 개발주기는 AI 모델 개발을 위해 데이터 수집부터 전처리·설계·학습·검증·테스트를 수행하는 과정이다.**

- 금융회사 등은 자체적으로 AI 모델 개발주기 별로 보안 점검항목을 개발하고 자체점검을 수행하여 AI 보안성을 높일 수 있다.

[AI 모델 개발주기 단계]

1 AI 학습 데이터 수집	- AI 모델 학습에 필요한 데이터 수집 단계 - 사내 내부 도는 외부 공개된 소스를 통한 데이터 수집
2 학습 데이터 전처리	- AI 모델에 적합한 형태로 변환하고 데이터 오염여부를 확인하는 단계 - 클리닝, 통합, 변환, 축소 등의 작업 수행
3 모델 설계·학습	- AI 모델을 설계하고 학습하는 단계 - 강건성 높도록 설계하는 것이 중요
4 모델 검증·테스트	- AI 모델의 성능 확인 및 테스트 단계 - 모델 입·출력값, 정보가 유출되지 않는 것이 중요

1 학습 데이터 수집

가. 개요

■ **학습 데이터 수집은 AI 서비스에 활용할 데이터를 수집하는 단계로 금융회사 등의 내부 데이터 외에도 오픈소스 등 외부의 데이터를 활용하거나 인터넷 등을 통해 직접 수집하여 활용할 수 있다.**

- 외부에서 수집한 학습 데이터는 악의적인 정보 삽입 등의 데이터 오염 가능성을 주의하고, 데이터 정보*를 관리해야 한다.

  * 출처, 버전, 구축 시점, 메타정보 등

나. 보안 고려사항

1 **(신뢰할 수 있는 출처)** 학습 데이터는 신뢰할 수 있는 출처로부터 수집하여 활용한다.

- AI 모델은 학습 데이터의 특성에 따라 기능 및 성능이 결정되기 때문에 오염된 데이터를 학습하면 성능이 저하되거나 보안성 관련 문제가 발생할 수 있다.
- 예를 들어 침입탐지시스템의 학습 데이터는 정상(Normal) 이벤트의 비중에 비해 비정상(Abnormal) 이벤트의 비중이 작다는 특징이 있다.
- 학습 데이터가 이런 특징을 보이는 경우 비정상 이벤트의 분포 및 특성이 AI 모델 성능을 크게 좌우하여, 데이터가 약간만 오염되어도 서비스에 큰 영향을 끼칠 수 있다.
- 또한, 단순 오류 등으로 인한 오류 값은 '데이터 전처리'를 통해 쉽게 발견할 수 있으나, 도메인 지식 기반의 정교한 데이터 변조는 발견이 어렵기 때문에 수집 단계부터 신뢰할 수 있는 데이터를 수집하는 것이 중요하다.

2 **(AI 학습 데이터 정보 관리)** 출처, 버전, 구축 시점, 메타정보* 등 학습 데이터의 정보를 관리하도록 한다.

  * 칼럼 정보, 데이터 타입(유형), 파일 포맷 등

- AI 모델 개발·고도화 시에는 다양한 데이터를 학습하게 되며, 학습·운영 과정에서 오염된 데이터를 통한 공격*과 이로 인한 피해가 발생할 수 있다.

  * [참고1] 데이터 오염 공격 참조

- 학습 데이터로 인해 발생한 보안 문제를 해결하기 위해서는 데이터에 대한 정보 관리가 필요하다.
- 학습 데이터를 수집한 출처와 시점, 수정 이력, 메타정보 등의 관리를 통해서 공격과 장애의 원인을 파악할 수 있다.

2 학습 데이터 전처리

가. 개요

■ **학습 데이터 전처리는 수집한 데이터를 AI 모델 입력이나 분석에 적합한 형태로 가공하는 단계이다.**

- 전처리 과정을 통해 학습 데이터의 품질뿐 아니라 AI 모델의 보안성 개선도 가능하다.
- 데이터의 품질 및 보안성은 AI 모델의 성능과도 직접적으로 관련 있는 만큼, 학습 데이터 전처리 단계에서 데이터의 품질 및 보안성을 확보하는 것이 중요하다.

나. 보안 고려사항

① **(이상치 처리)** 데이터 내에 불필요한 값을 제거하고 이상치를 검출하여 처리한다.

- 이상치 처리는 방식에 따라서 기계적 전처리와 의미적 전처리로 분류할 수 있다.
  - **(기계적 전처리)** 데이터의 통계치를 기반으로 데이터를 변환하거나 데이터 증강 등의 기법을 통해서 데이터 품질을 확보한다.
  - **(의미적 전처리)** 데이터에 대한 도메인 지식을 바탕으로 변환*하여 데이터의 품질을 확보한다.

    * 클러스터링, 레이블링, 특징 공학 등

참고 **이상치 검출·처리 기법**

✓ **(검출기법)** 개별 데이터 관찰, 통계기법, 시각화, 머신러닝 활용 등의 기법을 활용하여 이상치 검출 및 분석

검출방법	설명
통계기법 활용	• Z-점수(Z-Score) : 평균으로부터 3표준편차(0.15%) 떨어진 값을 이상치로 인식 • 기하평균 : 기하평균으로부터 2.5 표준편차 떨어진 값을 이상치로 인식 ※ 기하평균 ± 2.5 × 표준편차 • 사분위수 활용 : 제1사분위(Q1), 제3사분위(Q3)를 기준으로 사분위간 범위(d)의 1.5배 이상 떨어진 값을 이상치로 판단 ※ Q1-1.5(Q3-Q1) 〈 d 〈 Q3+1.5(Q3-Q1)
시각화 활용	• 데이터를 시각화하여 이상치 검출 • 확률 밀도함수, 히스토그램, 시계열차트, 산점도 등
머신러닝 기법 활용	• 클러스터링 기법*, 의사결정나무** 활용을 통한 이상치 판별 * K-평균 군집화 알고리즘 등 ** iForest기법 활용 등

✓ **(처리 방안)** 삭제, 대체, 변환, 제거 등을 통해 이상치 처리

검출방법	설명
유지	• 유의미한 특성을 보유한 이상치는 유지하여, 학습에 활용
삭제	• 이상값으로 판단되는 관측값 삭제
대체	• 상·하한값을 결정한 후, 이상치 조절
변환	• 비정상적으로 벗어난 이상치를 일정 값으로 변환

2 **(데이터 변조 확인)** 학습 데이터의 전반적인 추이 또는 샘플링한 값을 분석하여 데이터 변조 여부를 확인한다.

- 레이블 변조 및 의미적인 개념의 변조는 이상치 검출 기법으로 데이터를 발견하기가 어렵기 때문에 보다 세밀한 분석이 필요하다.
- 데이터가 변조되면 데이터 중독 공격 등을 통해 AI 모델 자체가 손상될 수 있으므로 학습 데이터를 변조로부터 안전하게 보호해야 한다.

3 **(적대적 예제 생성)** 적대적 예제*를 생성하여 AI 모델의 학습 데이터에 포함·학습하여, AI 모델을 적대적 공격**으로부터 보호한다.

* AI 모델이 잘못된 예측을 하도록 의도적으로 변조한 데이터

** 적대적 예제를 활용하여 AI 모델이 잘못 판단하도록 조작하는 공격

- AI 모델은 회피 공격* 등의 적대적 공격에 취약하기 때문에 이를 탐지하는 방안을 고려해야 한다.

  * [참고5] 회피 공격 참조

- 공격자는 AI 모델을 파악하기 위해 AI 모델의 입·출력*을 수집하는데 이러한 공격자의 행동 방식을 탐지에 참고할 수 있다.

참고 *** 공격자의 입·출력 수집 방식**

✓ 공격자는 입·출력값 수집을 위해 데이터 내에 노이즈를 삽입하거나 변조한 레이블을 넣어 다양한 데이터를 수집한다.

✓ 공격자는 입력값을 다양하게 변조하여 수집한 출력값으로부터 AI 모델에 대한 정보를 유추*하고 이를 통해 다양한 적대적 예제를 생성한다.

* AI 모델의 종류 유추, 결정경계(Decision Boundary) 등

- 적대적 공격은 화이트박스와 블랙박스 기반의 공격으로 분류한다.
  - 화이트박스 공격은 모델의 아키텍처, 입·출력값, 가중치, 파라미터 등 모델에 대한 정보를 사전에 파악한 상태에서 적대적 예제를 생성한다.
  - 블랙박스 공격은 무작위적으로 적대적 예제를 생성하는 방식으로 FGSM[2], Deepfool[3], JSMA[4], CW[5], BPDA[6] 등의 연구가 있다.
- 적대적 예제를 생성하여 학습*하는 것은 AI 모델의 강건성을 확보할 수 있는 방안으로써 적대적 공격을 사전에 예방한다.

  * 일종의 데이터 증강(Data Augmentation) 방식으로 다량 데이터 학습으로 AI 모델의 강건성을 높일 수 있음
- 이미지 데이터의 경우 회전·패딩·필터 등의 방식으로 데이터를 변환해서 적대적 예제를 생성하고 학습할 수 있다.

4 **(데이터 확장·학습)** 재현 데이터* 등의 기법을 활용하여 확장한 데이터를 학습할 수 있다.

  * 원본 데이터의 통계적 특성을 활용하거나 AI 기법 등을 이용하여 생성한 모의 데이터

- 확장한 데이터를 학습하는 방식을 통해 AI 모델의 강건성을 개선할 수 있다.

2) Ian J. Goodfellow et. al., "Explaining and Harnessing Adversarial Examples", 2015.
3) Moosavi-Dezfooli, Seyed-Mohsen, Alhussein Fawzi and Pascal Frossard, "DeepFool: a simple and accurate method to fool deep neural networks", 2016.
4) Papernot, Nicolas, et al., "The Limitations of Deep Learning in Adversarial Settings", 2016.
5) Carlini, Nicolas, and David Wagnet, "Towards Evaluating the Robustness of Neural Networks", 2017.
6) Anish Athalye at al., "Obfuscated Gradients Give a False Sense of Security : Circumventing Defenses to Adversarial Examples", 2018.

3 AI 모델 설계·학습

가. 개요

■ **AI 모델을 구성하고 데이터를 학습하여 AI 모델에 지능을 부여하는 단계를 말한다.**

- 데이터가 가지고 있는 상관성이 학습을 통해 AI 모델에 반영된다.

■ **AI 모델이 취약할 경우 공격자는 AI 서비스를 무력화하거나 모델을 복제할 수 있으므로 AI 모델의 보안성을 확보하는 것이 중요하다.**

- 이를 위해서는 AI 설계 단계에서부터 강건성이 우수한 모델을 활용하고 보안 고려사항을 반영하여, 공격으로부터 안전하게 보호해야 한다.

나. 보안 고려사항

1 **(강건성 높은 모델 선정)** 공격자가 AI 모델에 대한 정보를 쉽게 유추하기 어렵고, 데이터에 대한 민감도가 상대적으로 낮은 모델을 선택한다.

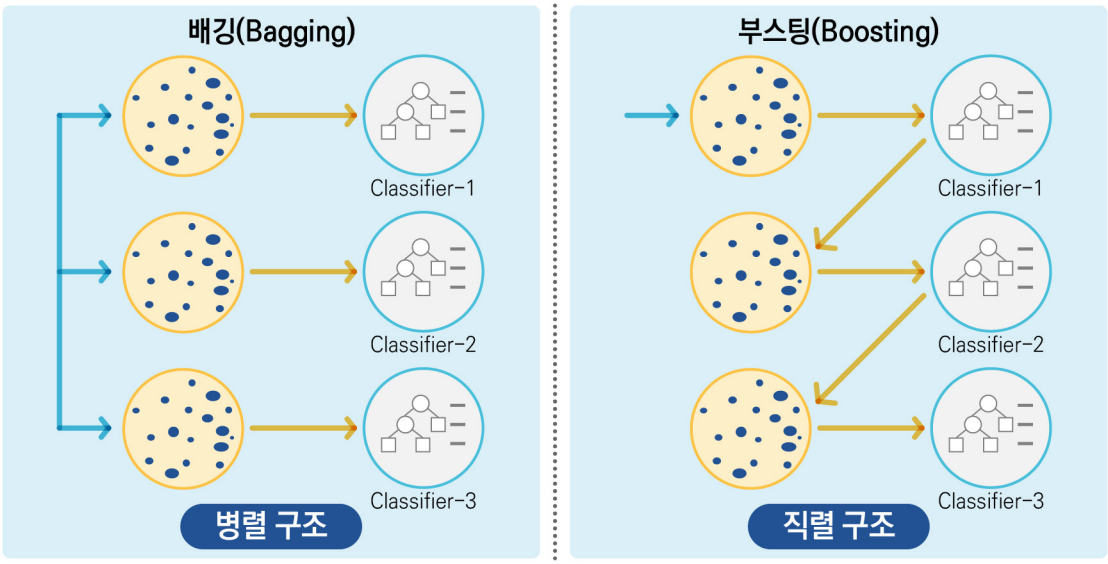
- AI 모델의 구조 및 알고리즘이 단순한 경우 공격자는 상대적으로 쉽게 AI 모델에 대한 정보를 파악할 수 있다.
- 공격자는 AI 모델에 대한 정보를 파악하여, 모델 추출 공격, 모델 인버전 공격 등의 공격을 수행할 수 있다.

  ※ [참고3] 모델 추출 공격, [참고4] 모델 인버전 공격 참조
- AI 모델 설계 시 강건성이 상대적으로 높은 알고리즘을 적용하여 AI 모델 정보에 대한 보안성을 강화할 수 있다.
  - 배깅(Bagging), 부스팅(Boosting) 등의 앙상블(Ensemble)*기법을 모델 설계에 적용하면 공격자가 모델을 유추하기 어려워진다.

참고 *** 앙상블 기법**

✓ **(앙상블 기법)** 다수의 AI 알고리즘을 조합하여 AI 모델을 설계하는 기법

- **(배깅)** 입력값에서 다수의 AI 알고리즘의 결과를 집계하여 판단
- **(부스팅)** 입력값을 다수의 AI 알고리즘이 순차적으로 학습 수행



- 로지스틱 회귀(Logistic Regression) 등 단순한 모델 설계보다는 DNN(Deep Neural Network) 방식 등과 같이 복잡한 방식의 설계가 회피 공격으로부터 상대적으로 안전하다.

  ※ '[참고5] 회피 공격' 참조

- 다만, AI 모델의 의사결정에 대한 설명이 필요한 경우 AI 모델의 구조가 복잡할수록 설명 가능성 확보가 어려울 수 있어 유의하여 설계해야 한다.

2 **(모델 튜닝)** AI 모델을 튜닝하여 적대적 공격과 모델 유추가 어렵게 모델의 강건성을 높인다.

- 모델 튜닝 시에 노이즈를 삽입하거나 매개변수 축소 등의 기법을 적용하여 적대적 공격으로부터 강건성을 확보할 수 있다.

참고 *** AI 모델 튜닝 세부방안**

✓ **(노이즈 삽입)** AI 모델 내부에 노이즈를 삽입함으로써, AI 모델의 강건성 확보

① 노이즈를 삽입한 은닉층을 삽입하여, 기존 모델 구조를 유추하기 힘들게 만들어서 강건성을 확보할 수 있도록 함

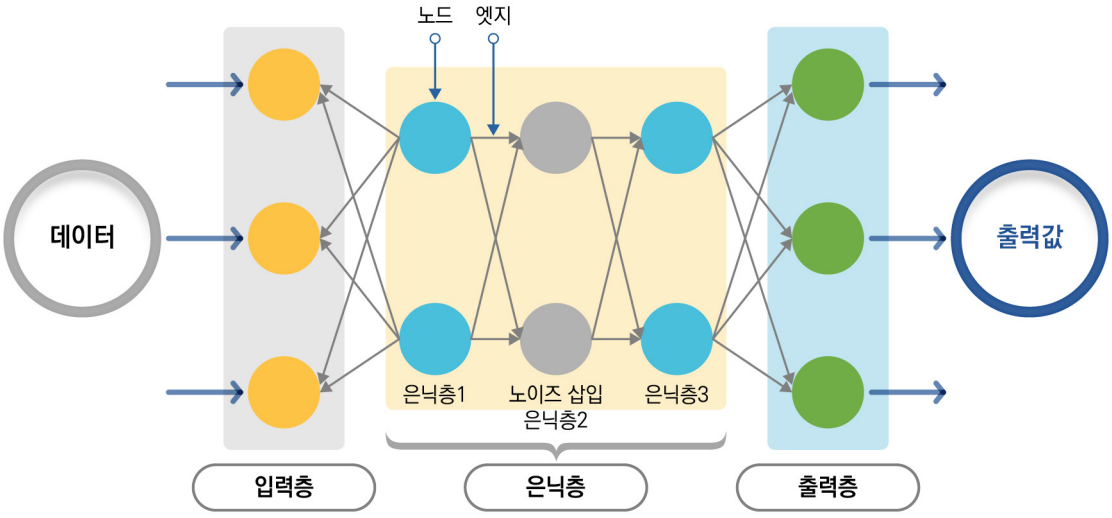
② 노드(Node)의 가중치(w)*에 노이즈를 삽입하여 적대적 예제로부터 완화

* 노드가 보유하고 있는 출력 강도를 수치화

〈 노이즈와 강건성의 상관관계 〉

- 구조가 단순하거나 알려진 AI 모델은 공격자가 모델의 구조를 유추 용이
- AI 모델 내부에 노이즈가 있을 시에는 공격자가 AI 모델을 쉽게 유추할 수 없어, 모델 인버전 공격, 회피 공격 등의 AI 모델 관련 공격에 안전
- 또한, 적대적 예제 등의 노이즈가 삽입된 샘플 등에 대해서 높은 성능을 보이기 때문에, 노이즈 삽입과 강건성 간에 상관관계가 있음
- 다만, 많은 노이즈 삽입은 모델의 성능 저하를 일으킬 우려가 있으며, 노이즈 삽입과 성능은 반비례적인 경향이 있음

〈 AI 모델 내부 구조 〉

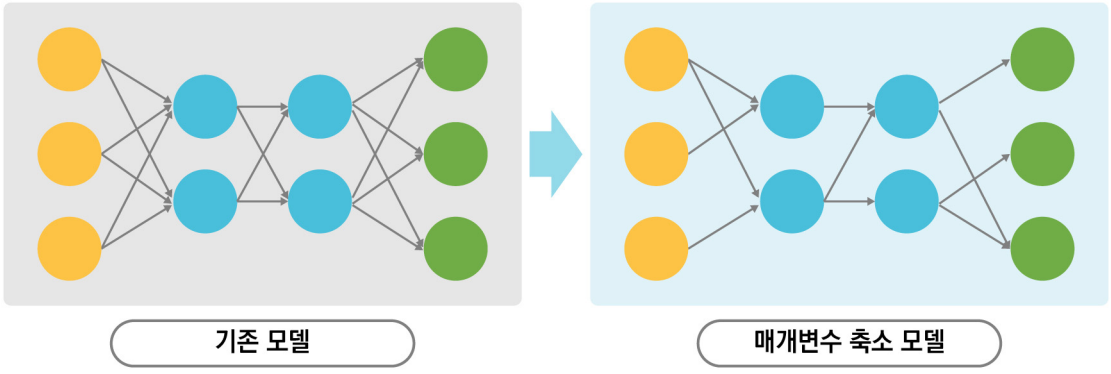


- AI 모델은 1개의 입력층(Input Layer)과 출력층(Output Layer), 1개 이상의 은닉층(Hidden Layer)으로 이루어져 있음
- AI 모델의 층은 노드로 이루어져 있으며, 각 노드는 가중치를 보유하며, 다른 노드와 엣지로 연결되어 전달받은 정보를 엣지를 통해 전달
- 위 AI 모델은 ①3개의 노드로 이루어진 입력층, ②2개의 노드를 보유한 3개의 은닉층, ③3개의 노드로 이루어진 출력층으로 구성되어 있음

✓ **(매개변수 축소)** AI 모델의 성능을 유지한 채, 구조를 변경하는 기법

- 해당기법을 통해서 공격자가 변형된 모델을 파악하기 어렵게 만들 수 있음

〈 파라미터 축소 기법 개요 〉



- 회피 공격은 AI 모델의 경사도와 경계선을 찾아, AI 모델의 결정을 회피하는 방식을 사용하기 때문에 경사도를 마스킹하여 노출을 방지하면 공격을 어렵게 할 수 있다.[7)]
- 또는 경사도 자체를 정규화하여 경사도에 대한 정보를 공격자가 유추하기 어렵게 한다.[8)]

7) Anish Athalye at al., “Obfuscated Gradients Give a False Sense of Security : Circumventing Defenses to Adversarial Examples”, 2018.
8) Papernot, Nicolas, et al., “Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks”, 2016.

3 **(신뢰성이 확보된 출처)** 사전학습 모델 활용 시에는 신뢰할 수 있는 출처로부터 받았는지 확인 후 활용한다.

- 방대한 데이터를 사전에 학습하여 생성하는 사전학습 모델은 알고리즘은 공개되어 있으나, 학습 데이터를 공개하는 경우는 거의 없다.
- 사전학습 모델이 학습한 데이터에 대한 신뢰성 보장이 어렵기 때문에 사용하기 전 사전학습 모델 출처의 신뢰성을 확인하는 것이 중요하다.

4 **(비잔티움 에러* 대처)** 연합학습 방식을 통해 학습을 진행할 경우는 비잔티움 에러를 방지할 방안을 마련한다.

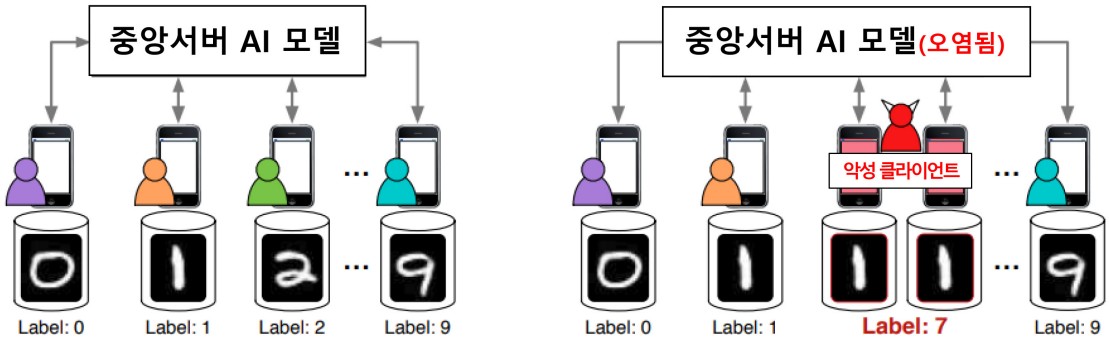
* 클라이언트에서 중앙서버로 메시지를 전달할 때 일부 클라이언트가 오염된 메시지를 전달하여 발생하는 오류

- 연합학습 방식의 AI 시스템을 운영하는 경우 악의적인 클라이언트, 또는 전송 과정 중 변조 등으로 인해 오염된 메시지를 전달받을 수 있다.
- 비잔티움 에러 발생 시 AI 모델이 오염될 수 있으므로 연합학습 방식 사용 시 악성 클라이언트와 오염된 메시지를 탐지·배제할 방안이 필요하다.

참고 *** 연합학습 방식에서의 악의적 행위**[9)]

✓ 악성 클라이언트에 의하여 중앙 서버의 AI 모델이 손상될 수 있음

〈 오염된 중앙서버의 AI 모델 비교 〉

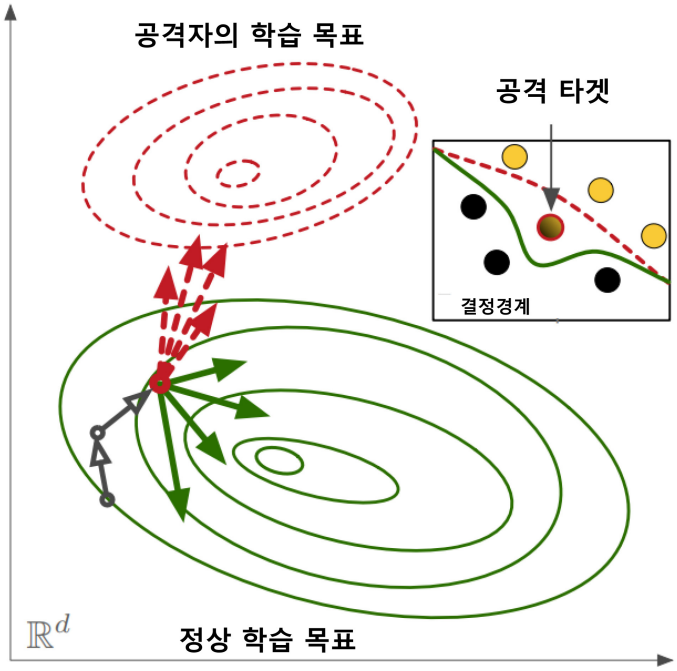


(1) 연합학습

(2) 오염된 악성 클라이언트로부터 공격받은 중앙서버 AI 모델

✓ 악성 클라이언트는 AI 모델을 비정상적인 방향으로 학습시키고자 하며, 이를 통해 회피 공격과 같은 적대적 공격, 교란 등의 행위 가능

〈 악성 클라이언트의 학습 방향 〉



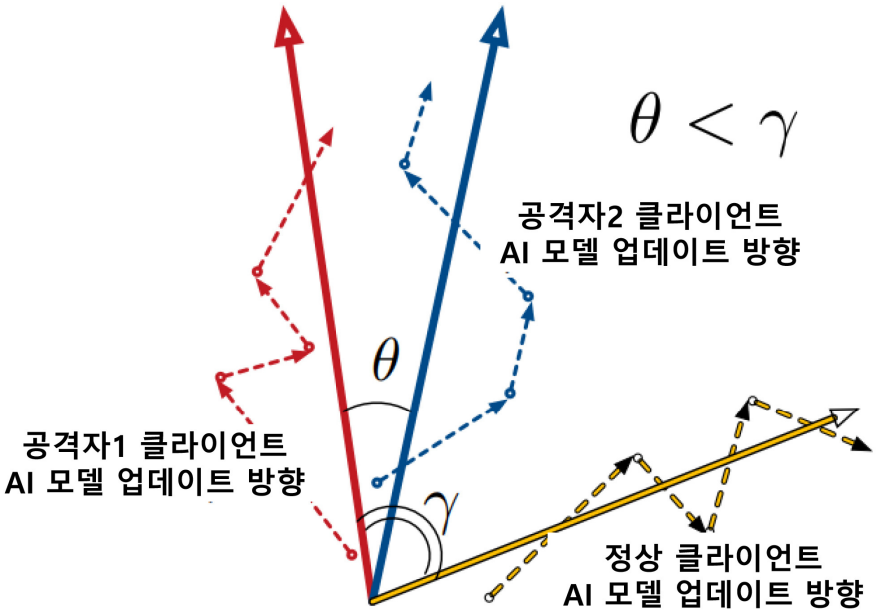
- 비잔티움 에러를 완화하기 위해서 클라이언트로부터 학습되는 과정을 확인하거나 통계적인 기법을 확인하는 방법이 연구되고 있다.
  - **(학습과정 확인을 통한 탐지[9])** 클라이언트로부터 중앙 서버의 AI 모델의 학습 정도가 변화되는 과정을 확인하여, 악성 클라이언트를 탐지한다.

참고 **FoolsGold 기법[9]**

✓ 클라이언트로부터 전달받은 매개변수를 적용한 학습 시에, 정상 클라이언트로부터 받은 매개변수와 악성 클라이언트로부터 받은 매개변수의 학습 방향이 다를 것이라 가정하고 이러한 특성을 통해 악성 클라이언트를 탐지

- 정상 클라이언트의 학습 방향과 공격자 간 벡터 각도는 γ로 표시할 수 있으며, 악성 클라이언트 간 각도는 θ로 표시
- 일정 θ와 γ를 기준으로 악성을 탐지

〈 FoolsGold 기법 개념 〉



9) Fung, Clement, Chris JM Yoon, and Ivan Beschastnikh., "Mitigating Sybils in Federated Learning Poisoning", 2018.

4 AI 모델 검증·평가

가. 개요

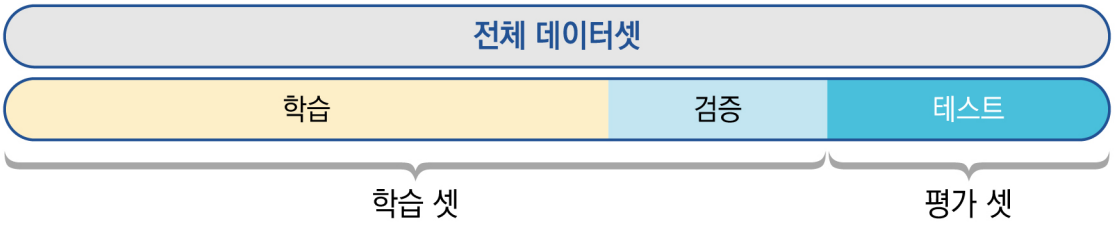
■ **학습 후 AI 모델의 성능을 확인하기 위해서 검증 및 평가(테스트)를 시행하는 단계이다.**

- **(모델 검증)** 학습이 완료된 모델을 검증하기 위한 단계로 학습 시에 이용된 데이터셋의 일부를 검증에 이용한다.

  ※ 홀드아웃 검증, K폴드 교체검증 등의 기법 활용 가능

- **(평가)** 학습과 검증 후에 모델의 최종 성능을 평가하기 위하여 테스트하는 단계로 학습·검증용 데이터셋과 별개의 평가 데이터셋을 준비해 평가에만 사용한다.

참고 **AI 서비스 개발 시의 데이터셋 구분**[10)]



나. 보안 고려사항

1 **(정보 관리)** 검증·테스트 시에 이용한 데이터셋, AI 모델 매개변수 등의 정보를 관리한다.

- AI 모델을 개발하는 데 있어, 투명성은 서비스의 품질 및 사고 대응에 도움이 되며, 보안 팀이나 개발자가 성능 저하 및 장애가 발생한 시기와 지점을 확인할 수 있다.

2 **(적대적 공격 테스트)** AI 모델을 대상으로 적대적 공격 등을 수행하여, 성능 수준을 평가한다.

- AI 모델 성능이 일정 수준 이하로 저하된다면, 적대적 예제를 다시 학습하거나 적대적 공격 탐지 기법 적용 등 강건성 확보 대책을 마련할 수 있다.

10) https://statology.org/validation-set-vs-test-set/

3 **(최종 출력값 확인)** 개인정보 등 민감한 정보가 이용자에게 출력되는지 확인한다.

- 민감한 정보가 출력되는 경우, 불특정 다수 또는 타인에게 노출이 되지 않도록 하는 방안을 마련한다.

  ※ AI 학습 데이터의 개인정보 활용과 관련된 상세한 사항은 개인정보보호법 및 인공지능(AI) 개인정보보호 자율점검표('21.5., 개인정보위) 참조

4 **(출력 횟수 제한)** AI 모델의 출력 횟수를 제한하여 모델의 정보 및 학습 데이터 유추를 어렵게 한다.

- 공격자는 AI 모델의 입·출력값을 기반으로 모델의 정보를 유추하고, 공격을 시도한다.

  ※ [참고3] 모델 추출 공격, [참고4] 모델 인버전 공격, [참고5] 회피 공격 참조

- 실제로 입·출력값 수집을 통해 아마존과 BigML의 유료 AI 모델을 99% 이상의 유사성으로 복제한 사례가 있다.[11)]

- AI 모델 정보의 유추와 모델 복제를 막기 위해서는 입·출력 횟수와 시간을 제한하는 방법 등을 적용할 수 있다.

- 출력 횟수를 제한하는 경우 AI 모델이 사용되는 업무나 서비스 특성을 고려*해 가급적 낮은 수준으로 정한다.

  * (예시) 챗봇의 경우 1분에 10회 미만 등

5 **(신뢰 점수 제한)** 정확도, 신뢰 점수* 등을 공개하지 않거나 범주화 등 비식별하여 제공함으로써 학습 데이터 및 모델 관련 정보 노출을 최소화한다.

* AI 모델이 데이터를 얼마나 믿을 수 있는지에 대한 점수 척도

- 공격자가 출력 결과에 대한 정확성, 신뢰 점수 등 모델에 대한 정보를 획득하면 모델 복제를 더 쉽게 할 수 있다.

  ※ [참고3] 모델 추출 공격, [참고4] 모델 인버전 공격, [참고5] 회피 공격 참조

- 따라서, 모델에 대한 정보 노출을 최소한으로 제한하여 공격을 어렵게 할 필요가 있다.

  ※ 다만, AI 의사결정 과정에 대한 설명의무가 있는 경우 설명에 불필요한 정보의 제공을 최소화한다.

11) Tramer et.al., "Stealing Machine Learning Models via Prediction APIs", 2016.

⑥ **(적대적 공격 탐지)** 적대적 공격을 탐지할 수 있는 기법을 적용한다.

- AI 모델은 적대적 공격에 취약할 수 있으며 공격이 발생하는 경우 차단 또는 재학습 등의 조치가 필요하므로 적대적 공격 탐지 방안을 고려해야 한다.
- 적대적 공격을 탐지하기 위한 기법으로는 노이즈 탐지[12], 입력데이터 인코딩[13], 중요도 맵 활용[14] 등이 있다.

[AI 모델 보안 고려사항]

AI 개발주기	점검항목
1. 데이터 수집	AI 학습 데이터에 대한 정보를 관리하고 있는가?
2. 데이터 전처리	학습 데이터 이상치(Outlier)를 식별·관리하였는가?
	학습 데이터의 변조 여부를 확인하였는가?
	예상할 수 있는 적대적 예제를 생성·학습하였는가?
3. 설계·학습	AI 모델 설계 시 알고리즘 선택 등에 있어 강건성을 고려하였는가?
	모델 튜닝 시 강건성을 확보 조치를 하였는가?
	(사전학습 활용시) 사전학습 모델을 신뢰할 수 있는 출처로부터 받았는가?
	(연합학습 활용시) 비잔티움 에러를 대처하는 방법을 적용하였는가?
4. 검증·테스트	검증·테스트 시에 이용한 데이터셋, AI 모델 매개변수 등의 정보를 관리하고 있는가?
	AI 모델을 대상으로 적대적 공격 등을 수행하여, 성능 수준을 확인하였는가?
	최종 출력값을 확인하였는가?
	모델의 출력 횟수를 제한하였는가?
	모델 정보 노출을 최소화하였는가?
	적대적 공격을 탐지할 수 있는 기법을 적용하였는가?

12) Kwon, Hyun, Hyunsoo Yoon, and Ki-Woong Park., "Acoustic-decoy:Detection of adversarial examples through audio modification on speech recognition system", 2020.
13) Weiling Xu, "Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks", 2017.
14) G.Ko and G.Lim, "Unsupervised Detection of Adversarial Examples with Model Explanations", 2011.

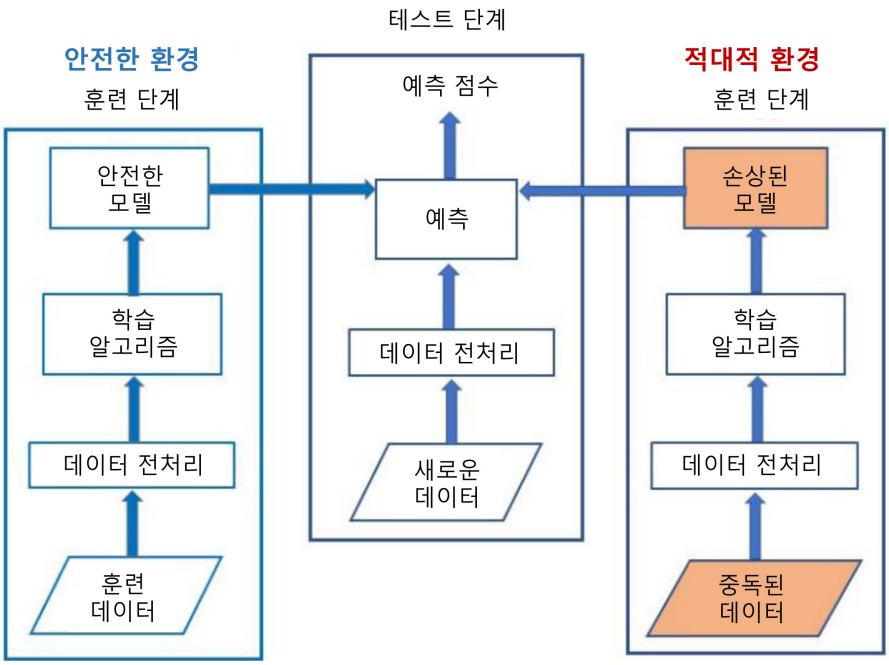
참고1 데이터 오염 공격

■ **오염된 데이터*가 AI 모델 학습 시에 활용되어, AI 시스템의 성능저하 및 오작동을 유발하게 되는 공격이다.**

* 적대적 예제 또는 악의적인 노이즈 등이 삽입된 데이터

- 데이터 오염 공격은 AI 모델 자체를 공격하는 공격으로서 이를 방지하기 위해서는 학습에 활용하는 데이터의 오염 여부를 학습 전에 확인하여야 한다.
- **(데이터 백도어 공격)** AI 알고리즘이 공격자의 트리거가 삽입된 악의적인 데이터를 학습하는 공격기법으로서 데이터 오염 공격의 일종이다.

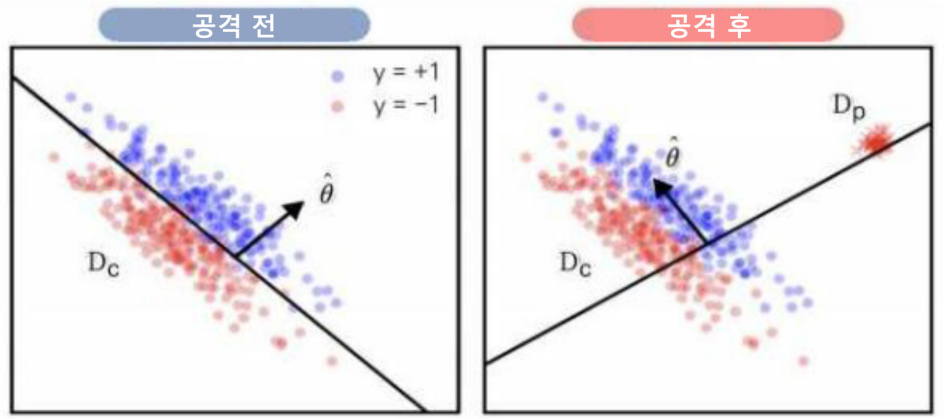
참고 **데이터 중독 프로세스**[15)]



15) Matthew Jagielski, et al., "Manipulating Machine Learning : Poisoning Attacks and Countermeasures for Regression Learning", 2018.

■ **AI 모델이 오염된 데이터를 학습하면 모델의 정확도가 저하된다는 것은 다수의 연구[16]를 통해 증명된 바 있다.[17]**

참고 **오염된 데이터 학습으로 인하여 경계면 변경 및 탐지 성능 저하[17]**



■ **데이터 오염 공격은 이상치 처리 등의 방법을 통해서 예방하거나 차단할 수 있다.**

- **(이상치 처리)** 임계치를 이상 점수(Outlier Score)로 정하고 이상 점수 밖에 위치하는 오염된 데이터를 찾거나[18], 클러스터링 기법을 활용하여 클러스터 밖에 위치하는 오염된 데이터를 찾는 방식을 활용할 수 있다.[19]

16) Shafahi Ali et al. “Poison frogs! targeted clean-label poisoning attacks on neural networks”, 2018.
17) https://towardsdatascience.com/poisoning-attacks-on-machine-learning-1ff247c254db
18) Paudice, Andrea, et al., “Detection of Adversarial Training Examples in Poisoning Attacks through Anomaly Detection.”, 2018.
19) Steinhardt, Jacob, Pang Wei W. Koh, and Percy S. Liang., “Certified Defenses for Data Poisoning Attacks.” , 2017.

참고2 모델 오염 공격

■ **모델 오염 공격은 연합학습*시 클라이언트에서 학습·생성된 악의적인 모델이 중앙서버의 AI 모델에 적용되어 성능저하 및 오작동을 유발하는 공격이다.**

- 서버가 각 클라이언트에서 전달되는 모델의 신뢰성 및 오염 여부를 확인하는 것이 어렵기 때문에 모델 오염 공격은 탐지하기 어렵다.

■ **FoolsGold 기법, 집계 알고리즘 적용 등을 통해서 모델 오염 공격을 예방할 수 있다.**

- **(집계 알고리즘 적용)** 알려진 집계 알고리즘(Krum[20], Bulyan[21], Trimmed Mean[22])을 사용하여 비잔티움 에러를 완화한다.
- **(FoolsGold 기법* 적용)** 연합학습 시 FoolsGold 기법을 통해서 악성 클라이언트를 판별하고 배제한다.[23]

  * 정상 클라이언트와 악성 클라이언트 간 경사도를 비교해서 서로 다른 업데이트 양상을 확인하여 악성을 판별하는 기법

- **(Feature Squeezing)** AI 모델의 예측 결과와 Squeezer 알고리즘*을 적용한 예측 결과를 비교함으로써, 적대적 예제를 탐지한다.[24]

  * 기존 입력값의 인코딩 단순화, 평활화 필터 등을 적용하여 특징 축소

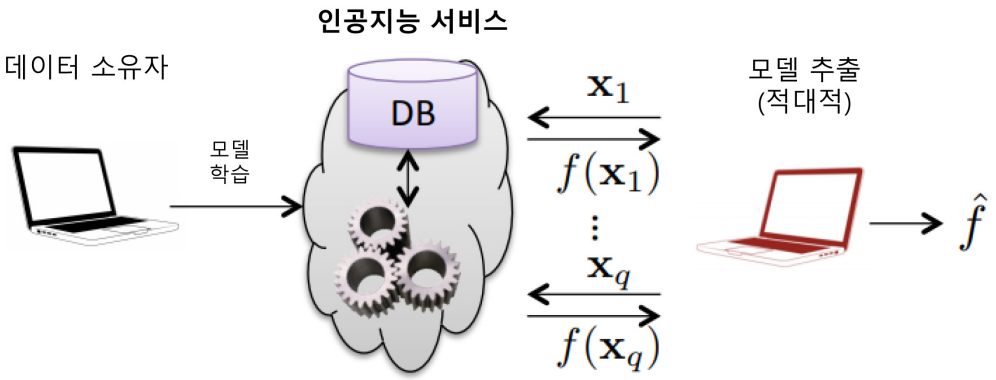
20) Peva Balachard, Rachid Guerraoui, Julien Stainer, et al. “Machine Learning with Adversaries : Byzantine Tolerant Gradient Descent”, 2017.
21) El Mahdi El Mhamdi, Rachid Guerraoui, and Sebastien Rouault, “The Hidden Vulnerability of Distributed Learning in Byzantium”, 2018.
22) Dong Yin, Yudong Chen, Kannan Ramchandran, and Peter Bartlett, “Byzantine-Robust Distributed Learning : Towards Optimal Statistical Rates”, 2018.
23) Fung, Clement, Chris JM Yoon, and Ivan Beschastnikh., “Mitigating Sybils in Federated Learning Poisoning”, 2018.
24) Weiling Xu, “Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks”, 2017.

참고3 모델 추출 공격

■ **모델 추출 공격은 원본 AI 모델로부터 유사한 모델을 추출하여 복제하는 공격이다.**

- 원본 AI 모델에 대한 대량의 쿼리를 통해 입·출력값을 수집하고 이를 학습하여 원본 AI 모델과 유사도가 높은 모델을 복제한다.

참고 **모델 추출 공격 프로세스**[25)]



■ **아마존(Amazon) 및 BigML이 MLaaS 형태로 제공하는 유료 AI 모델을 몇 분만에 99% 이상 유사도로 복제 가능하다는 것이 관련 연구[25)]로 증명된 바 있다.**

■ **다음과 같은 방법을 통해서 모델 추출 공격을 예방할 수 있다.**

- **(출력값 제한)** 출력 횟수와 시간을 제한하여 공격자의 입·출력 정보 수집을 어렵게 한다.
- **(예측 결과 변환)** 비식별 처리 등을 통해 출력값을 변환하면, 입·출력값으로부터 유사 모델을 생성하기 어려워져 모델 추출 공격을 예방할 수 있다.
- **(DP-SGD 기법*[26)])** 학습 과정에서 차등 프라이버시 기법을 적용함으로써, 경사도 등 모델과 관련된 정보 노출을 최소화하는 기법이다.

25) Tramer et.al., "Stealing Machine Learning Models via Prediction APIs", 2016.
26) Martin Abadi et al., "Deep Learning with Differential Privacy", 2016.

참고 *** DP-SGD(Differential Privacy - Stochastic Gradient Descent) 기법**

✓ SGD는 AI 모델을 학습* 방식의 대표적인 방법으로 입력 데이터를 작은 크기로 분할 집합(Mini Batch)하여 학습 진행

* AI 모델 학습은 새로운 데이터 입력으로 AI 모델 내의 노드 등의 수치가 최적화(Optimize)되는 과정으로 해석

✓ DP-SGD는 SGD방식에서 차등 프라이버시 기법을 적용하여 학습 진행

- 분할 집합마다 각 가중치를 구하고 최대 기울기 제한(Clip Gradient) 및 통계적 기반의 노이즈(Gaussian Noise)를 추가하여 학습 진행

✓ 해당 기법으로 공격자는 출력값을 기반으로 입력값을 유추, 모델 유추 등이 어려움

- **(목표 함수 노이즈[27])** 목표 함수(Object Function)*에 차등 프라이버시 기법 적용을 통해서, 입력값을 유추할 수 없도록 한다.

  * 학습 과정에서 입력 데이터에서의 실제 정답(레이블)과 AI 모델의 예측값 사이의 차이(손실값)를 판단하는 수식 함수

- **(데이터 분할 방식[28])** 민감 데이터를 분할하고 각 민감 데이터마다 각기 다른 분류기로 학습 및 각 분류기를 조합하여 학습하는 방식을 제안하였다.

  - 이 방식을 통해서, 출력값으로 원 입력값을 유추할 수 없도록 한다.

27) Chaudhuri, Kamalika, and Claire Monteleoni, “Privacy-preserving logistic regression”, 2009.
28) Papernot, Nicolas, et al., “Scalable private learning with pate”, 2018.

참고4 모델 인버전 공격

■ **모델 인버전 공격은 모델의 출력값으로부터 입력값을 유추하는 공격이다.**

- 공격자는 모델 종류, 파라미터 등 알려진 정보를 기반으로 대리모델*을 제작하여 공격에 활용할 수 있다.

  * 원본 모델의 예측을 정답으로 삼아 지도학습을 통해 학습한 유사모델

- 공격자는 쿼리를 통한 입·출력값, 신뢰 점수(Confidence Score)* 추측을 통하여 학습 데이터를 추측할 수 있다.

  * 신뢰구간에 모집단의 실제 평균값이 포함될 확률

■ **신뢰 점수를 기반으로 공격하여 입력값(원본 이미지)에 가까운 이미지를 생성한 사례도 존재한다.**[29)]

참고 **원본 이미지를 기반으로 복구한 이미지 예시**[29)]

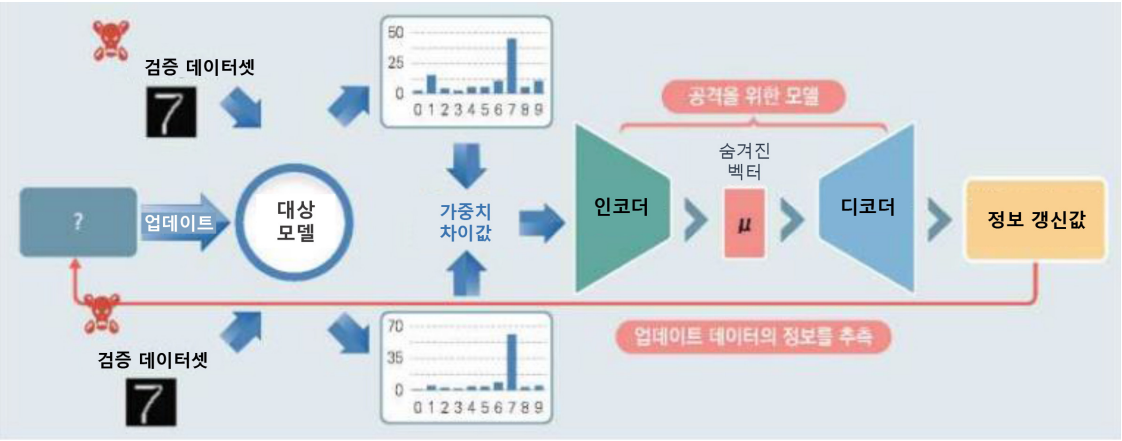
"원본 이미지"

"복구 이미지"

29) Fredrikson et.al,. "Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures.", 2015.

- 또한, 온라인 학습(Online Learning)*의 경우 AI 모델 재학습 전·후 동일한 입력값에 대한 출력값의 변화를 수집하여 재학습에 사용된 학습 데이터를 복구한 사례도 존재한다.[30)]

참고 **Updates-Leak 공격 모식도[30)]**



■ **모델 인버전 공격은 모델 추출 공격과 마찬가지로 입·출력값 보호를 통해 예방하고 차단할 수 있다.**

30) Ahmed Salem, "Update-Leak: Data Set Inference and Reconstruction Attacks in Online Learning", 2020.

참고5 회피 공격

■ **회피 공격[31]은 노이즈를 추가한 적대적 데이터를 입력하여, AI 모델이 잘못된 판단을 하도록 유도하는 공격이다.**

참고 **적대적 예제로 인한 오분류[31]**

"Panda" 노이즈 "Gibbon"

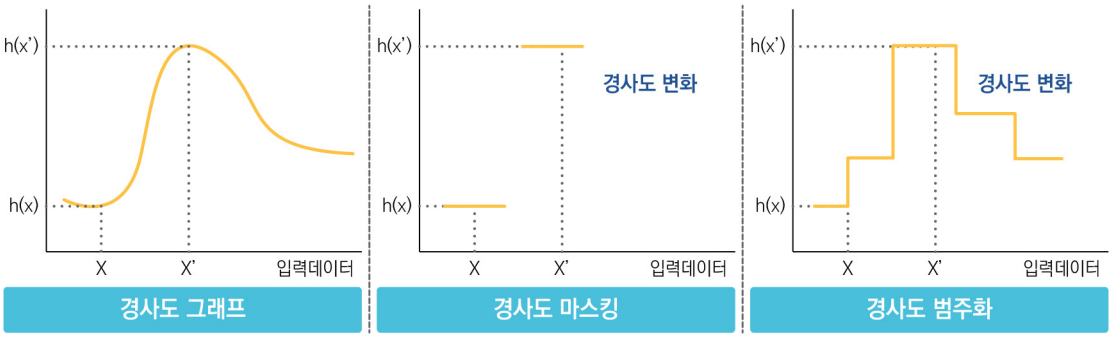
※ 회피 공격은 입력값에 사람의 눈으로는 구분하기 어려운 노이즈의 추가로 결괏값에 영향을 미치기 때문에, 데이터 관찰을 통한 탐지가 쉽지 않음

■ **적대적 예제 학습, 경사도 마스킹 등의 방법을 통해서 회피 공격을 예방하고 차단할 수 있다.**

- **(적대적 예제 학습[32])** 사전에 적대적 예제를 생성하여 학습함으로써, 적대적 예제를 통한 회피 공격을 방어할 수 있다.
- **(경사도(Gradient) 마스킹[33])** 공격자는 데이터를 입력하였을 시에 AI 모델의 학습 방향 등 변화되는 정도를 기반으로 적대적 예제를 생성할 수 있다.
  - AI 모델의 학습 방향은 데이터 입력 시에 AI 모델의 경사도를 통하여 확인할 수 있으며, 이를 공개하지 않거나 마스킹, 범주화 등의 기법을 활용하여 모델의 정보 노출을 최소화한다.

31) Ian J. Goodfellow et. al., "Explaining and Harnessing Adversarial Examples", 2015.
32) Zhang, Lemoine, and Mitchell, "Mitigating Unwanted Biases with Adversarial Learning", 2018.
33) Anish Athalye at al., "Obfuscated Gradients Give a False Sense of Security : Circumventing Defenses to Adversarial Examples", 2018.

참고 *** 경사도 마스킹(예시)**



- **(지식 증류**[34]**)** 모델의 경사도 자체를 정규화하여 적대적 공격의 학습 방향에 대한 노출을 최소화함으로써, 모델에 대한 정보를 보호하여 회피 공격을 어렵게 한다.
- **(입력값 변화**[35]**)** 입력값이 이미지인 경우 회전, 패딩 등 임의적인 변환을 통해서, 모델의 강건성을 향상할 수 있다.

34) Papernot, Nicolas, et al., "Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks", 2016.
35) Xie, Cihang, et al., "Mitigating Adversarial Effects through Randomization.", 2017.

별첨 AI챗봇서비스 보안성 체크리스트

분류	소분류	연번	보안성 체크리스트
공통	계정관리	1	챗봇 서비스 사용자별 역할(e.g., 관리자, 이용자)에 따라 계정을 생성하고 관리하는가?
	접근통제	2	챗봇 시스템에 대한 해킹 등을 방지하기 위해 정보보호시스템이 설치 및 운영되고 있는가?
		3	챗봇 시스템 관련 파일(e.g., AI 모델 파일, 소스코드, 학습데이터 등)에 대한 접근통제 정책 등록/변경/삭제 시 그 내역을 기록하고 정기적으로 적정성 및 이상 유무를 검토하는가?
	보안점검	4	챗봇 서비스 개발 시 보안 가이드라인 등을 참고하여 개발하고 오픈 전 시스템에 대한 보안점검을 실시하는가?
		5	챗봇 시스템에 대해 주기적으로 취약점을 점검하는가?
	입력 제한	6	개인정보를 처리하지 않는 챗봇의 경우 입력창에 개인정보를 입력하지 않도록 이용자에게 사전에 안내하는가?
		7	개인정보를 처리하지 않는 챗봇의 경우 챗봇 서비스 내 사용자 입력 시, 개인정보가 입력되거나 개인정보가 포함되는 파일이 업로드되는 것을 제한하는가?
	중요정보 보안	8	챗봇 서비스 관련 중요정보*에 대해 적절한 보호대책(계정/권한 관리, 접근통제, 암호화 등)이 적용되어 있는가? * 개인정보, AI 모델파일, 학습데이터(학습용, 검증용 등), 챗봇관리시스템, 챗봇인프라 관련 로그, 발화문 원문 저장 DB 등
		9	챗봇 서비스 내 중요정보* 암호화 시, 안전한 암호화 알고리즘을 사용하여 암호화되고 관리되는가? * 개인정보, AI 모델파일, 학습데이터(학습용, 검증용 등), 챗봇관리시스템, 챗봇인프라 관련 로그, 발화문 원문 저장 DB 등
		10	암호화에 사용되는 암호키에 대해 적절한 암호키 관리방안(e.g., 안전한 난수발생기를 통한 암호키 생성, 암호키 교체 주기 적용 등)을 수립하고 이행하는가?
	원격접속 금지	11	내부망 이외의 곳*에 위치한 챗봇서비스 관리자 시스템에 원격접속을 원칙적으로 금지하고 있는가? 원격으로 접속이 불가피한 경우,강화된 보안통제(e.g.,추가인증, 접속IP제한, 중요정보 노출 최소화 등)를 적용하는가? * 클라우드나 DMZ 구간 등

분류	소분류	연번	보안성 체크리스트
	악성코드 방지 대책	12	챗봇 시스템 내 악성코드에 대한 대책이 마련되어있는가?
	백업/복구 절차	13	챗봇 시스템에 대한 백업 및 복구절차가 수립/이행되고 있는가?
	패치 관리	14	챗봇 시스템 내 정보자산에 대한 패치 관리정책 및 절차가 적절히 수립(e.g., 인터넷 직접 접속을 통한 패치가 제한 등)되어 이행되는가?
	침해사고 대응	15	챗봇서비스 관련 침해사고 대응 절차를 수립하여 이행하는가?
	위험관리	16	AI를 활용한 챗봇서비스를 구축하는 경우, 이용자에게 잠재적으로 미칠 위험을 평가하고, 이를 관리하기 위한 위험관리 정책을 수립 이행하는가?
	외부자 계약	17	챗봇서비스 제공자와 같은 외부자와 계약 체결 시 정보보호 요구사항을 식별하고, 관련 내용을 계약서에 명시하여 그 이행 여부를 주기적으로 관리하는가?
	정보자산 식별	18	챗봇 서비스 관련 주요 정보자산(e.g., 관리 시스템, 인프라 등)을 식별하고 적절한 보호대책을 수립/이행하는가?
	학습데이터 관리	19	학습데이터 관리와 학습 통제 절차를 수립하고 이행하였는가?
		20	학습데이터 출처에 대한 신뢰성 평가 기준을 수립하고 이행하였는가?
		21	학습데이터 공격에 대한 보호대책이 마련되어있는가?
		22	학습데이터에 대한 업데이트 이력을 관리하고 있는가?
		23	AI 모델과 학습데이터 등의 주요파일 위변조 시 탐지방안 혹은 무결성 검증 방안이 있는가?
선택	개인정보 활용	24	(챗봇 서비스 내 개인정보 활용 시) 개인정보 및 민감정보의 활용 필요성을 사전에 검토하는가?
		25	(챗봇 서비스 내 개인정보 활용 시) 챗봇서비스에 개인정보 또는 민감정보를 활용하는 경우에는 안전성 확보에 필요한 조치를 이행하고 이를 정기적으로 점검하는가?
	클라우드 기반 챗봇 플랫폼 이용	26	(클라우드 기반 챗봇 플랫폼 이용 시) 외부기관과 동일 챗봇 플랫폼을 이용하는 경우, 해당 외부기관과의 네트워크 접근통제를 실시하는가?
		27	(클라우드 기반 챗봇 플랫폼 이용 시) 클라우드 기반 챗봇 서비스를 제공하는 정보처리 시스템의 중요도평가를 수행하는가?

분류	소분류	연번	보안성 체크리스트
		28	(클라우드 기반 챗봇 플랫폼 이용 시) 업무용 단말기 및 내부망 정보처리 시스템을 클라우드 서비스 제공자 구간에 위치한 내/외부망 정보처리 시스템에 연결하거나, 관리용 단말기를 클라우드 서비스에 연결해야 하는 경우, 적절한 보호대책이 수립되어 있는가?
		29	(클라우드 기반 챗봇 플랫폼 이용 시) 클라우드 접근에 필요한 키를 안전한 장소에 보관하고 접근권한을 최소화하는가?
	오픈소스 소프트웨어 사용	30	(오픈소스 소프트웨어 사용 시) 오픈소스 소프트웨어를 식별 및 관리하고, 승인완료한 오픈소스 소프트웨어만을 사용하는가?
	API 사용	31	(API 사용 시) API에 대한 적절한 보안대책(e.g.,사용하지 않는 API 식별 및 제거, API 관련 이상행위 모니터링 등)을 수립하여 적용하는가?
		32	(API 사용 시) API 변경 필요시, 변경 관리 절차에 따라 관리하고 변경내역을 문서화하여 관리하는가?
		33	(API 사용 시) API 호출에 필요한 키를 안전한 장소에 보관하고 접근권한을 최소화하는가?

금융분야 AI 보안 가이드라인

발행일 : 2023년 4월

발행인 : 김 철 웅

발행처 : 금융보안원
경기도 용인시 수지구 대지로 132

전 화 : (02) 3495-9000

〈비 매 품〉

금융분야
AI 보안 가이드라인

금융보안원
FINANCIAL SECURITY INSTITUTE