**인공지능기초 중간고사 대체과제**

**서술형 문제 보고서**

**202021497 생명공학과 이주철**

* **AI 보안 및 책임 있는 AI 서술형**

- 서술형 문제 보고서 템플릿 – AI 보안 및 책임 있는 AI -

**1. 서론**

- 현재, 이 시대에 많은 사람들은 AI 기술을 사용하고 있다. AI 기술의 빠른 확산 속에서 보안 리스크와 윤리적인 책임에 대한 문제는 피할 수 없다. AI 시스템이 의사결정 과정에 개입하고 인간의 삶에 커다란 영향을 미친다. 여러 영역에 확정되면서 도입 이전에 보안 취약점과 윤리적 책임에 대한 사전 검토는 필수적이어야 한다. 잘못된 판단, 시스템 장애, 정보 유출 나아가 사회적 혼란까지 초래할 수 있다. AI의 안정성과 윤리성을 책임지고 설계하는 것이 스타트업의 전략이 된다. 이를 바탕으로 AI 보안 리스크 관리와 책임 있는 AI 전략을 수립하는 것이 목적이다.

**2. AI 보안 리스크 식별 및 분석**

**1)** AI 시스템에서 발생할 보안 위협으로 첫 번째 Data Poisoning과 Infrastructure Attacks 이 두 가지에 관해 설명하려 한다.

Data Poisoning은 모델 학습에 사용되는 데이터를 조작하여 AI 모델의 성능을 악화시켜 사용자에게 영향을 끼친다.

Infrastructure Attacks는 물리적 및 가상 환경을 공격 대상으로 하며, AI가 운영되는 서버 환경에 해킹이나 Dos 공격을 가해 과부하를 일으켜 서버에 접근할 수 없게 만드는 보안 리스크이다.

**2)** Data poisoning은 의도적으로 데이터를 조작하여 모델이 편향적인 형태를 이루어 잘못된 출력을 내도록 유도하는 공격이다. 한 헬스케어 스타트업 에서는 환자의 유전자와 증상 데이터를 바탕으로 질병 예측 AI를 제공한다. 경쟁 회사는 몰래 환자들의 진료 데이터를 수집하는 API에 접근해서 의도적으로 의료 데이터를 조작한다. 예를 들면 알츠하이머가 없는 환자인데 알츠하이머가 있는 환자로 레이블링해서 추가한다. 그렇게 되면 학습된 모델은 알츠하이머 예측 정확도가 급격히 떨어지고, 알츠하이머가 아닌 사람에게 이 질병 가능성을 알려주는 False Positive 문제로 이어질 수 있다.

이러한 위협은 AI 모델이 훈련 데이터에 크게 의존한다는 원인에 의해 발생한다. 특히, 외부에서 데이터를 수집하고 제3자 API를 통해 정보를 수집한다면, 데이터 무결성 검증이 생략되는 구조적 허점이 존재한다. API 키나 인증 토큰 관리가 소홀하다면 보안 사고가 발생한다. 위 사례는 경쟁사가 공개 API를 통해 환자 데이터를 수집하는 경로를 악용했다.

Infrastructure Attacks는 AI가 운영되는 서버, 네트워크 등의 물리, 가상 서버에 대한 공격이며 모델 자체가 아닌 호스팅하는 환경을 노리는 방식이다. 한 자율주행 차량을 위한 서버를 구축하는 스타트업이 있다. 한 해커가 컨테이너 인증키를 탈취하여 내부에 백도어에 심었다. 그 후 서버가 조종당해 자율주행차들이 오작동하며 급정지, 급발진하는 사고가 발생한다. 기술적인 문제가 아닌 인프라 해킹이었다는 사실이 밝혀져 회사에 대한 보안 문제로 한바탕 소동이 일어났다.

위 사례에 대한 위협의 원인은 인증 시스템의 취약성과 보안 관제의 부재에 있다. 클라우드 기반의 컨테이너 환경을 사용하여 비용 효율을 높였다. 하지만 API 키나 인증 토큰 관리가 소홀해지면 보안 사고가 발생한다. 해커는 노출된 인증키를 통해 컨테이너 내부 접근 경로를 확보하고, 백도어를 심어 시스템을 장악했다. 이러한 키 관리나 접근 권한 최소화가 부족하여 생긴 위협 발생이다.

**3. 대응 방안 제시**

**1)** Data poisoning은 기술적으로 데이터 정합성 검증 시스템을 도입하여 이상치 탐지, 라벨 연관성 확인 등을 자동화하여 학습 데이터가 오염되는 것을 방지한다. 외부에서 수집된 데이터를 해시값이나 디지털 서명으로 검증하여 데이터 소스 인증 및 해시 검증으로 예방한다.

조직적으로 데이터 수집 프로토콜을 명확화하여 어떤 데이터 출처만 허용할지 또는 수집 방식에 대한 기준을 명시하는 것이다. 다른 방안으로는 교육 및 내부 감사 체계로 데이터 관리자에게 보안 리터러시 교육 제공 및 정기적으로 데이터 점검을 하게 한다. 해당 위협을 사전에 예방하고 대응하기 위해 이러한 방안들을 고안해 보았다.

위 방안들은 사후 예방이며, 사후 대응으로는 이상 예측 패턴 분석으로 특성 클래스에 이상하게 높은 확률을 주는 샘플을 탐지하여 이상 학습 여부를 판단한다. 오염된 데이터를 제거하고 롤백하여 기존 모델 버전으로 되돌린다.

한 시점에 무슨 데이터가 들어온 건지 로그 시스템을 유지한다.

2) Infrastructure Attacks는 API Key 및 Access Token을 주기적으로 회전하여 인증키가 유출되어도 일정 주기마다 바뀌기 때문에 공격이 불가하다. 또한, 방화벽 시스템을 설치하여 침입 탐지 및 차단 시스템으로 접근을 차단한다.

어느 서버에 무슨 권한이 있는지 관리자에게 제한하여 접근 권한을 최소화한다. 엔지니어 및 관리자를 대상으로 보안키 관리 같은 공격 대응 교육을 실시한다.

사후 대응은 공격 시도나 내부 데이터 유출 시도를 확인할 수 있게 하는 모니터링은 한다. 또는 로그 기반으로 서버 접근 기록, 컨테이너 동작 로그 등을 분석해 공격 경로를 추적한다. 조직적으로 보안 사고 대응 매뉴얼을 수립하여 어떻게 그 상황에 대응할지를 메뉴얼 화한다. 또는 외부 기관과 함께 보안 점검을 하여 취약점을 평가하고 수정한다.

**4. 책임 있는 AI 원칙 적용**

**1)** OECD가 제시한 원칙 중 3가지는 공정성, 설명 가능성, 책임성이다. 학습 데이터를 수집하는 단계에서 성별, 연령, 장애 유무, 나이, 지역 등 다양한 특성군을 균형 있게 설계한다. 편향성 탐지 알고리즘을 도입하여 특정 그룹에 불리한 예측이 반복되는지 확인한 후 공정성을 반영하여 모델 성능 평가 지표에 메트릭을 추가한다. 고령자나 청소년의 데이터가 부족할 수도 있기에 보완 데이터 확보 및 증강이 필요하다. 장애인 사용자에 대해 오작동이나 잘못된 결과 발생률을 따로 모니터링하고 개선할 부분을 개선하는 계획을 수립한다. 시력 장애, 청각 장애를 고려해 텍스트/음성을 기반 인터페이스를 멀티모달 대응 설계로 보완을 할 것이다.

설명 가능성은 사용자와 이해관계자가 AI 시스템의 작동 원리를 이해할 수 있어야 한다. 관리자나 고객이 AI 예측 결과에 이의를 제기할 수 있게 피드백 시스템을 내장한다. 설명 가능한 모델을 우선 사용하거나 이중 검증 체계를 도입하여 위험도가 높은 판단에 대해서는 이러한 방식으로 적용한다. 청소년 사용자는 직관적인 시각화와 설명 텍스트로 표현하여 쉽게 보여주고, 고령 사용자를 위해서는 음성 안내 혹은 문장을 요약하여 간단한 형태로 제공한다.

책임성은 AI 시스템의 오작동 혹은 피해가 발생했을 경우 인간이 최종 통제권을 가져야 한다. 모델 운영 중에 문제가 발생했을 경우 모델 로그를 기록, 버전관리하고 결과 추적 기능을 구현한다. 항상 사람의 검토를 거치게 하여 AI 의사결정 구조를 설계한다. 시스템이 잘못 판단할 경우에는 자동으로 중단하는 트리거를 탑재한다. 고령자는 디지털 접근성이 낮은 경우가 많아 인간 상담원이나 관리자와의 대화 즉, 소통 채널을 제공한다. 청소년과 유아를 위해서는 연령별 사용 제약 설정 기능을 도입하여 보호한다. 장애인을 대상으로는 자동화된 판별에 대해서 이의 제기 절차를 명확하게 시스템에 포함시켜 이해할 수 있도록 제공한다.

**5. AI 거버넌스 및 이해관계자 참여**

1) AI 시스템의 책임 있는 개발과 운영은 다양한 내부 및 외부 이해관계자가 적극적으로 참여해야 우리의 목표를 달성할 수 있다. 다양한 시각과 경험을 반영하면서 윤리적 오류와 보안 리스크를 사전에 방지할 수 있다.

AI 개발자는 기술 설계 및 구현을 하여 데이터 처리, 성능을 개선한다. 시스템 관리자 및 운영자는 운영 안정성 역할을 하며 AI가 배포된 서버 및 API의 보안, 로그 추적 등을 책임진다. 윤리 전문가와 법률 자문자는 윤리적 기준 제시 및 검토 역할을 하며 알고리즘 편향성, 인권 침해 가능성과 법적 리스크를 분석한다. 외부 감시자는 외부의 독립적 시각으로 시스템을 평가하고 감시한다. 이는 투명성 요구 및 제3자 검토 역할을 수행한다. 개발자는 윤리적 위험을 완전히 예측하기 어렵다. 윤리 전문가와 법적 자문자의 참여를 통해 편향, 차별, 오작동 리스크를 사전에 식별할 수 있다. 시스템 관리자는 보안상 이상 징후를 발견하여 실시간 로깅 체계를 제공한다. 이는 AI 운영 단계에서 책임성을 확보하는 데 기여한다. 물론 각 이해관계자가 1:1로 협업하는 것은 아니다. 개발에는 기술 전문 개발자뿐 아니라 윤리, 법, 책임 등 다양한 이해관계자의 참여가 필요하다.

윤리위원회는 AI 개발팀과 법률/윤리 전문가, 외부 자문위원으로 구성하며 데이터 수집의 정당성, 모델 편향성을 평가하고 알고리즘의 영향성을 분석할 것이다. 이를 바탕으로, 정기적으로 코드 리뷰와 모델 사전 검토 프로세스를 도입하고 정기적으로 회의를 진행한다.

AI 감사는 모델 로깅 및 추적 시스템으로 입출력 데이터와 의사결정 경로를 자동으로 기록하고 자동화된 이상 탐지를 통해 예측 패턴에서 이상값을 탐지하여 감사 체계를 도입한다. 내부 의사결정 책임을 구조화하기 위해 RACI 메트릭을 적용한다. 위험도가 높은 AI를 사용할 시 승인 체계를 도입하여 공공 서비스용 AI는 출시 전에 이사회 또는 윤리 위원회 승인을 필수적으로 받아야 하는 구조를 도입해야 한다.

**6. 결론**

본 보고서는 AI 시스템을 실서비스에 도입하기 전 고려해야 할 AI 보안 리스크 관리와 책임 있는 AI 전략의 수립에 대해 다양한 각도로 분석해 보았다. AI 시스템은 사회적 영향과 윤리적 영향을 동반하여 Data Poisoning, Infrastructure Attacks 같은 위협이 신뢰성과 안정성을 떨어뜨리는 요소가 된다. 이를 해결하기 위해서는 기술적 대응과 조직적 대응을 사전, 사후 대응으로 제시했다. OECD에서 제안하는 공정성, 설명 가능성 (투명성), 책임성을 원칙으로 실제 서비스에 반영하고, 특정 사회적 그룹에 대한 영향을 고려하여 설계 요소를 작성하였다. 이해관계자 간 협업의 중요성과 내부 거버넌스 체계를 어떻게 구현해야 할지도 역시 필수 요소로 다뤄 보았다. AI를 지속 가능한 서비스로 만들기 위해 사용자 신뢰와 사회적 수용성 확보에 있다. AI 모델 설계 초기 단계에서 보안과 윤리적 문제를 다루고 내제화한다. 내부의 다양한 부서와 이해관계자 간 협업을 강화하여 신뢰 가능한 사회적 파트너로 형성해야 한다.

**7. 참고자료**

**- Data Camp, AI SECURITY AND RISK MANAGEMENT 강의자료 참고**

**-** [**https://oecd.ai/en/ai-principles**](https://oecd.ai/en/ai-principles) **: OECD 가치 기반 원칙 5가지**

**-** [**https://www.ibm.com/kr-ko/think/topics/ai-governance**](https://www.ibm.com/kr-ko/think/topics/ai-governance) **: AI 거버넌스**

**- 맞춤법 검사와 문단 정리를 위해 ChatGPT를 사용했습니다.**