2024

신뢰할 수 있는 인공지능

개발 안내서

과학기술정보통신부
Ministry of Science and ICT

TTA 한국정보통신기술협회
Telecommunications Technology Association

일러두기

- 본 안내서는 과학기술정보통신부 「AI신뢰성 기반조성」 사업의 연구 결과로서 내용의 무단 전재를 금합니다.

- 아울러, 안내서의 내용을 가공·인용하는 경우 반드시 '과학기술정보통신부·한국정보통신기술협회 「2024 신뢰할 수 있는 인공지능 개발 안내서 – 스마트 치안 분야」'의 출처를 밝혀주시기 바랍니다.

- 본 안내서는 스마트 치안 분야 인공지능 서비스 및 제품을 개발하는 과정에서 참고 자료로 활용할 수 있도록 편찬되었습니다. 본 안내서는 기업의 업무 환경과 상황, 개발 목적 등을 고려하여 필요한 내용을 선택하여 활용하시기 바랍니다.

- 본 안내서의 치안분야·인공지능 동향 및 기술 정보는 2023년 12월 기준으로 서술되었습니다.

- 인공지능 신뢰성은 사회 구성원의 다양한 의견과 논의를 통해 합의와 공감대를 이루어야 하는 개념으로, 본 안내서가 이러한 담론의 수집과 논의의 장을 마련하는 촉매제가 되었으면 하는 바람입니다. 이를 위해 폭넓고 심도 있는 의견을 듣고 반영하고자 하오니, 많은 참여와 관심 부탁드립니다.

- 본 안내서는 한국정보통신기술협회가 운영하는 TrustOps 웹페이지(aitrustops.or.kr)에도 콘텐츠가 공개되어 있으므로 참고하시면 더 편리하게 활용하실 수 있습니다.

- 스마트 치안 외 분야는 「2024 신뢰할 수 있는 인공지능 개발 안내서 – 일반 분야」를 참고해주시기 바라며, 특화될 서비스 분야는 점차 확대해나갈 예정입니다.

CONTENTS

안내서 활용을 위한 체크리스트

안내서 활용을 위한 체크리스트

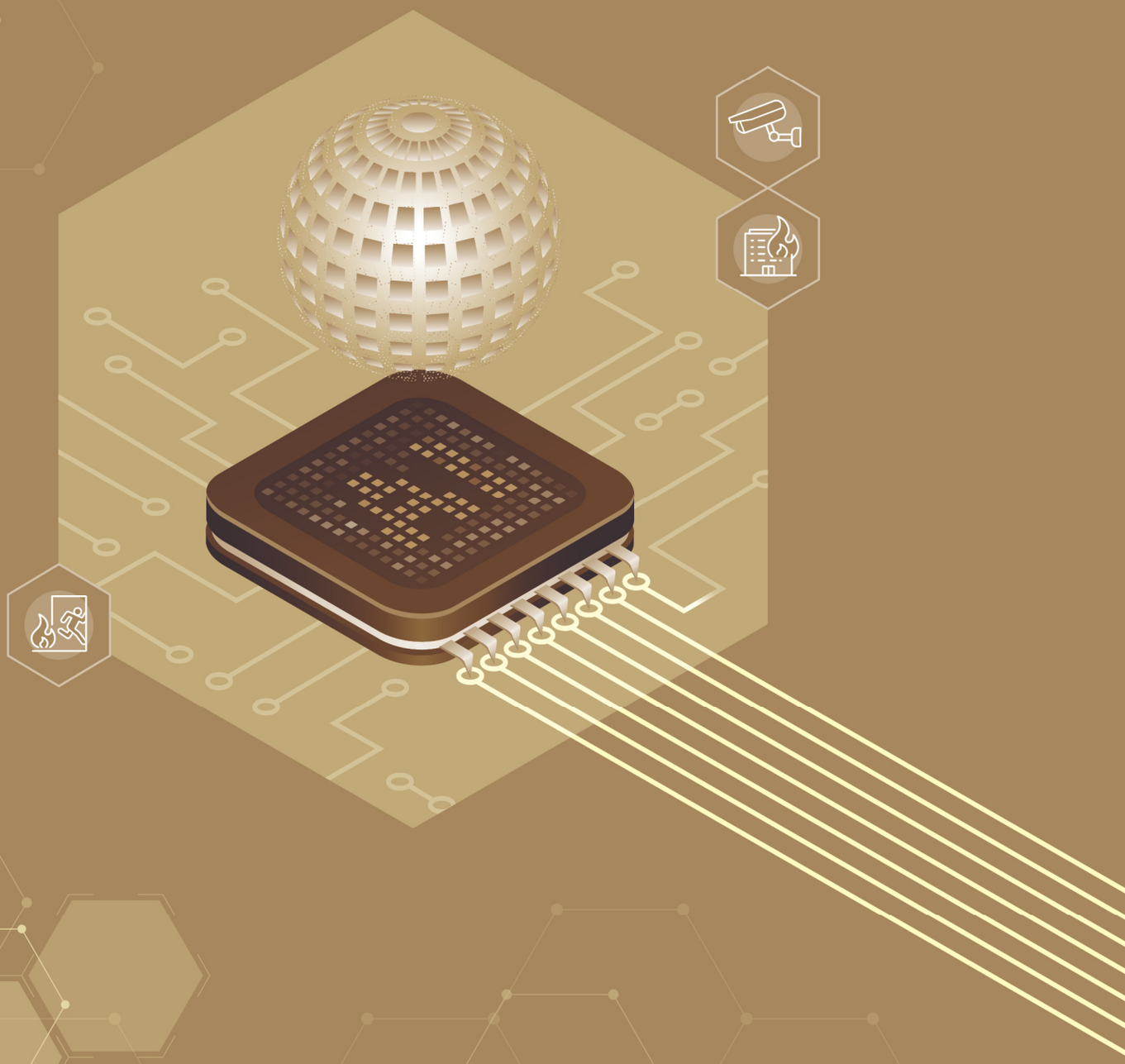
생명주기	요구사항 및 체크리스트		Yes	No	N/A
1 생명주기 관리	요구사항 01	인공지능 시스템의 위험 관리 계획 및 수행			
	01-1	인공지능 시스템 생명주기에 걸쳐 나타날 수 있는 위험 요소를 분석하였는가?	☐	☐	☐
	01-1a	인공지능 시스템의 위험 요소를 도출하고 이의 파급효과를 파악하였는가?	☐	☐	☐
	01-1b	인공지능 기술 적용을 어렵게 만드는 위험 요소가 있는지 확인하였는가?	☐	☐	☐
	01-2	위험 요소를 제거 및 방지하거나 영향을 완화하기 위한 방안을 마련하였는가?	☐	☐	☐
	01-2a	위험 요소별 완화 또는 제거 방안을 마련하였는가?	☐	☐	☐
	01-2b	위험 요소의 파급효과가 감소하였는지 확인하였는가?	☐	☐	☐
	요구사항 02	인공지능 거버넌스governance 체계 구성			
	02-1	인공지능 거버넌스에 대한 지침 및 규정을 수립하였는가?	☐	☐	☐
	02-1a	내부적으로 준수해야 할 인공지능 거버넌스에 대한 지침 및 규정을 마련하였는가?	☐	☐	☐
	02-2	인공지능 거버넌스를 위한 조직을 구성하고 인력 구성에 대해 검토하였는가?	☐	☐	☐
	02-2a	인공지능 거버넌스를 위한 조직을 구성하였는가?	☐	☐	☐
	02-2b	인공지능 거버넌스를 위한 조직은 전문성을 갖춘 인력으로 구성하였는가?	☐	☐	☐
	02-3	인공지능 거버넌스 체계가 올바로 이행되고 있는지 감독하고 있는가?	☐	☐	☐
	02-3a	인공지능 거버넌스에 대한 내부 지침 및 규정 준수 여부를 감독하고 있는가?	☐	☐	☐
	02-4	인공지능 거버넌스 조직이 신규 및 기존 시스템의 차이점을 분석하였는가?	☐	☐	☐
	02-4a	기존 동일 목적의 시스템과 비교하여, 신규 시스템이 개선할 수 있는 사항을 분석하였는가?	☐	☐	☐
	요구사항 03	인공지능 시스템의 신뢰성 테스트 계획 수립			
	03-1	인공지능 시스템의 특성을 고려한 테스트 환경을 설계하였는가?	☐	☐	☐
	03-1a	테스트 환경 결정 시 인공지능 시스템의 운영환경을 고려하였는가?	☐	☐	☐
	03-1b	가상테스트 환경이 필요한 인공지능 시스템의 경우, 시뮬레이터를 확보하였는가?	☐	☐	☐
	03-2	인공지능 시스템의 테스트 설계에 필요한 협의 체계를 구성하였는가?	☐	☐	☐
	03-2a	인공지능 시스템의 기대 출력을 결정하기 위한 협의 체계를 구성하였는가?	☐	☐	☐
	03-2b	설명가능성 및 해석가능성 확인을 위한 사용자 평가단을 구성하였는가?	☐	☐	☐
	요구사항 04	인공지능 시스템의 추적가능성 및 변경이력 확보			
	04-1	인공지능 시스템의 의사결정에 대한 추적 방안을 수립하였는가?	☐	☐	☐
	04-1a	인공지능 시스템의 의사결정에 대한 기여도 추적 방안은 확보하였는가?	☐	☐	☐
	04-1b	인공지능 시스템의 의사결정 추적을 위한 로그 수집 기능을 구현하였는가?	☐	☐	☐
	04-1c	지속적인 사용자 경험 모니터링을 위해 사용자 로그를 수집 및 관리하고 있는가?	☐	☐	☐

안내서 활용을 위한 체크리스트

생명주기	요구사항 및 체크리스트		Yes	No	N/A
1 생명주기 관리	04-2	학습 데이터의 변경 이력을 확보하고, 데이터 변경이 미치는 영향을 관리하였는가?	☐	☐	☐
	04-2a	데이터 흐름 및 계보lineage를 추적하기 위한 조치를 마련하였는가?	☐	☐	☐
	04-2b	데이터 소스 변경에 대한 모니터링 방안을 확보하였는가?	☐	☐	☐
	04-2c	데이터 변경 시, 버전관리를 수행하였는가?	☐	☐	☐
	04-2d	데이터 변경 시, 이해관계자를 위한 정보를 제공하는가?	☐	☐	☐
	04-2e	신규 데이터 확보 시, 인공지능 모델의 성능평가를 재수행하였는가?	☐	☐	☐
2 데이터 수집 및 처리	요구사항 05	**데이터 활용을 위한 상세 정보 제공**			
	05-1	데이터의 명확한 이해와 활용을 지원하는 상세한 정보를 제공하는가?	☐	☐	☐
	05-1a	정제 전과 후의 데이터 특성을 설명하였는가?	☐	☐	☐
	05-1b	학습 데이터와 메타데이터metadata를 구분하고 각 명세자료를 확보하였는가?	☐	☐	☐
	05-1c	보호변수protective attribute의 선정 이유 및 반영 여부를 설명하였는가?	☐	☐	☐
	05-1d	라벨링 작업자를 위해 교육을 시행하고 작업 가이드 문서를 마련하였는가?	☐	☐	☐
	05-2	데이터의 출처는 기록 및 관리되고 있는가?	☐	☐	☐
	05-2a	신뢰할 수 있는 출처로부터 제공되는 데이터셋을 사용하였는가?	☐	☐	☐
	05-2b	오픈소스 데이터셋을 활용하는 경우, 출처를 명시하였는가?	☐	☐	☐
	요구사항 06	**데이터 견고성 확보를 위한 이상abnormal 데이터 점검**			
	06-1	이상 데이터의 식별 및 정상 여부를 점검하였는가?	☐	☐	☐
	06-1a	전체 학습용 데이터 분포를 시각화하여 발생 가능한 오류들을 확인하였는가?	☐	☐	☐
	06-1b	학습 데이터 이상값 식별 기법을 적용하였는가?	☐	☐	☐
	06-2	데이터 공격에 대한 방어 수단을 강구하였는가?	☐	☐	☐
	06-2a	데이터 최적화를 통한 방어 대책을 마련하였는가?	☐	☐	☐
	요구사항 07	**수집 및 가공된 학습 데이터의 편향 제거**			
	07-1	데이터 수집 시, 인적·물리적 요인으로 인한 편향 완화 방안을 마련하였는가?	☐	☐	☐
	07-1a	인적 편향을 제거하기 위한 절차적, 기술적 수단을 적용하였는가?	☐	☐	☐
	07-1b	데이터의 다양성 확보를 위해 여러 수집 장치를 활용하였는가?	☐	☐	☐
	07-2	학습에 사용되는 특성feature을 분석하고 선정 기준을 마련하였는가?	☐	☐	☐
	07-2a	보호변수 선정 시 충분한 분석을 수행하였는가?	☐	☐	☐
	07-2b	편향을 발생시킬 수 있는 특성의 영향력을 완화하였는가?	☐	☐	☐
	07-2c	데이터 전처리 시 특성이 과도하게 제거되었는지 검토하였는가?	☐	☐	☐

생명주기	요구사항 및 체크리스트		Yes	No	N/A
2 데이터 수집 및 처리	07-3	데이터 라벨링 시, 발생 가능한 편향을 확인하고 방지하였는가?	☐	☐	☐
	07-3a	데이터 라벨링 기준을 명확히 수립하고 작업자에게 제공하였는가?	☐	☐	☐
	07-3b	다양한 데이터 라벨링 작업자를 섭외하기 위해 노력하였는가?	☐	☐	☐
	07-3c	다양한 데이터 라벨링 검수자를 확보하기 위해 노력하였는가?	☐	☐	☐
	07-4	데이터의 편향 방지를 위한 샘플링을 수행하였는가?	☐	☐	☐
	07-4a	편향 방지를 위한 샘플링 기법을 적용하였는가?	☐	☐	☐
3 인공지능 모델 개발	요구사항 08	**오픈소스 라이브러리의 보안성 및 호환성 점검**			
	08-1	오픈소스 라이브러리의 안정성을 확인하였는가?	☐	☐	☐
	08-1a	활성화된 오픈소스 라이브러리를 사용하였는가?	☐	☐	☐
	08-2	오픈소스 라이브러리의 위험 요소는 관리되고 있는가?	☐	☐	☐
	08-2a	사용 중인 오픈소스 라이브러리의 라이선스 준수사항을 이행하였는가?	☐	☐	☐
	08-2b	사용 중인 오픈소스 라이브러리의 호환성 및 보안취약점을 확인하였는가?	☐	☐	☐
	요구사항 09	**인공지능 모델의 편향 제거**			
	09-1	모델 편향을 제거하는 기법을 적용하였는가?	☐	☐	☐
	09-1a	개발하려는 모델에 맞게 편향제거 기법을 선택하였는가?	☐	☐	☐
	09-1b	편향성 평가 및 모니터링을 위한 정량적 지표를 선정하고 관리하는가?	☐	☐	☐
	요구사항 10	**인공지능 모델 공격에 대한 방어 대책 수립**			
	10-1	모델 공격이 가능한 상황을 파악하였는가?	☐	☐	☐
	10-1a	데이터 유형별 공격 가능한 적대적 사례를 확인하였는가?	☐	☐	☐
	10-2	모델 공격에 대한 방어 수단을 강구하였는가?	☐	☐	☐
	10-2a	모델 최적화를 통한 방어 대책을 마련하였는가?	☐	☐	☐
	요구사항 11	**인공지능 모델 명세 및 추론 결과에 대한 설명 제공**			
	11-1	인공지능 모델의 명세를 투명하게 제공하는가?	☐	☐	☐
	11-1a	시스템 개발 과정과 모델 작동 방식에 대한 세부 정보가 설명된 문서를 작성하였는가?	☐	☐	☐
	11-2	사용자가 모델 추론 결과의 도출 과정을 수용할 수 있도록 근거를 제공하는가?	☐	☐	☐
	11-2a	인공지능 모델에 적합한 XAI[eXplainable AI] 기술을 적용하였는가?	☐	☐	☐
	11-2b	XAI 기술 적용이 불가능한 경우, 기술 외 대안을 마련하였는가?	☐	☐	☐
	11-3	모델 추론 결과에 대해 사용자의 판단을 도울 수 있는 설명을 제공하는가?	☐	☐	☐
	11-3a	모델 추론 결과에 대한 설명이 필요한지 검토하였는가?	☐	☐	☐
	11-3b	사용자에게 인공지능 모델 추론 결과에 대한 설명을 제공하였는가?	☐	☐	☐

생명주기	요구사항 및 체크리스트	Yes	No	N/A
4 시스템 구현	**요구사항 12** 인공지능 시스템 구현 시 발생 가능한 편향 제거			
	12-1 소스 코드 및 사용자 인터페이스로 인한 편향을 제거하기 위해 노력하였는가?	☐	☐	☐
	12-1a 데이터 접근 방식 구현과정 등 소스 코드에서의 편향 발생 가능성을 확인하였는가?	☐	☐	☐
	12-1b 사용자 인터페이스 및 상호작용 방식으로 인한 편향을 확인하였는가?	☐	☐	☐
	요구사항 13 인공지능 시스템의 안전모드 구현 및 문제발생 알림 절차 수립			
	13-1 공격, 성능 저하 및 사회적 이슈 등의 문제 발생 시 대응 가능한 안전 모드를 적용하는가?	☐	☐	☐
	13-1a 문제 상황에 대한 예외 처리 정책이 마련되어 있는가?	☐	☐	☐
	13-1b 인공지능 데이터 및 모델 공격에 대해 시스템 측면의 방어 대책을 마련하였는가?	☐	☐	☐
	13-1c 인공지능 시스템의 의사결정으로 인한 파급효과가 크고 불확실성이 높은 경우, 사람의 개입을 고려하였는가?	☐	☐	☐
	13-1d 예상되는 사용자 오류에 대한 안내 및 대응을 제공하는가?	☐	☐	☐
	13-2 인공지능 시스템에서 문제가 발생할 경우, 시스템은 이를 운영자에게 전달하는 기능을 수행하는가?	☐	☐	☐
	13-2a 편견, 차별 등 윤리적 문제에 대한 알림 절차를 수립하였는가?	☐	☐	☐
	13-2b 시스템 성능 저하를 평가하기 위한 지표 및 절차를 설정하고 알림 절차를 수립하였는가?	☐	☐	☐
	요구사항 14 인공지능 시스템의 설명에 대한 사용자의 이해도 제고			
	14-1 인공지능 시스템 사용자의 특성(user characteristics)과 제약사항을 분석하였는가?	☐	☐	☐
	14-1a 사용자 특성에 따른 세부 고려사항을 분석하였는가?	☐	☐	☐
	14-2 사용자 특성에 따른 설명을 제공하는가?	☐	☐	☐
	14-2a 사용자 특성에 따른 설명 평가 기준을 수립하였는가?	☐	☐	☐
	14-2b 사용자가 이해하기 어려운 전문 용어 사용을 지양하였는가?	☐	☐	☐
	14-2c 사용자의 구체적인 행동과 이해를 이끌어낼 수 있도록 명확한 표현을 사용하였는가?	☐	☐	☐
	14-2d 설명이 필요한 위치와 타이밍은 적절한가?	☐	☐	☐
	14-2e 사용자 경험을 평가할 수 있는 다양한 사용자 조사 기법을 활용하였는가?	☐	☐	☐
5 운영 및 모니터링	**요구사항 15** 서비스 제공 범위 및 상호작용 대상에 대한 설명 제공			
	15-1 인공지능 서비스의 올바른 사용을 유도하기 위한 설명을 제공하는가?	☐	☐	☐
	15-1a 서비스의 목적과 목표에 대한 설명을 제공하는가?	☐	☐	☐
	15-1b 서비스의 한계와 범위에 대한 설명을 제공하는가?	☐	☐	☐
	15-2 사용자가 상호작용하고 있는 대상을 인지할 수 있도록 설명하는가?	☐	☐	☐
	15-2a 사용자와 인공지능이 상호작용하는 서비스 범위를 명시하였는가?	☐	☐	☐
	15-2b 서비스 내의 최종 의사결정을 수행하는 주체에 대해 사용자에게 설명하는가?	☐	☐	☐



PART 1

개요

01 안내서 발간 배경 및 목적

스마트 치안 영역에서 인공지능(AI) 기술은 다양한 측면에서 중추적인 역할을 한다. 기본적인 감시, 탐지 등 관련 목적뿐만 아니라 폭력 감지, 수사, 경고 시스템, 범죄 분석 등 전문 분야로까지 확대되고 있다. AI 기술의 적용 범위가 확대되면서, 법 집행 기관과 공공 안전 담당 조직에서도 기존의 업무 수행 방식을 AI 기술과 유기적으로 결합해 개선하려는 요구가 증가하고 있다.

이러한 배경에는 최근 지능형 범죄의 발생 빈도가 다른 범죄에 비해 나날이 증가해, 공공 안전 조직에서 인공지능 기술을 활용해 치안을 확보하려고 노력하기 때문이다.

스마트 치안 시스템은 법 집행 및 공공 안전 조치를 강화하고 공공 보안을 유지하고자 인공지능, 데이터 분석, 심지어 IoT(사물 인터넷) 애플리케이션을 사용하는 고급 기술이다. 이 시스템은 방대한 양의 데이터를 실시간으로 처리 및 분석하고, 범죄 활동을 예측하며, 리소스 할당을 최적화하고, 대응 시간을 개선할 수 있다는 특징이 있다. 이러한 시스템에는 감시 카메라, 센서 네트워크, 예측 분석, 스마트 커뮤니케이션 도구 등의 기능이 통합되어 법 집행 기관, 정부 기관에서 주로 사용되며, 범죄 예방 및 공공 안전을 관리하는 중요한 도구 역할을 한다.

하지만 이러한 기술적 변화에는 도전 과제가 뒤따른다. 데이터와 알고리즘의 편향성으로 인한 편견, 개인정보 침해, 개인의 권리 침해 가능성에 대한 우려가 제기되고 있다. 따라서 국제기구와 정부는 합법성, 윤리, 사회적 안전의 원칙을 지키고자 스마트 치안의 맥락에서 AI의 윤리적·합법적 사용을 면밀히 조사하고 있다. AI는 법 집행 발전의 핵심 원동력이 되었지만, 스마트 치안 분야에서 AI의 신뢰성을 실제로 구현하는 것은 지속적인 과제로 남아 있다. 법 집행 기관은 이러한 격차를 해소하여 AI 기반 치안이 높은 수준의 신뢰성을 유지함으로써 적법성과 윤리 기준을 준수하는 동시에 공공 안전을 강화하겠다는 약속을 이행하도록 보장해야 한다. 스마트 치안의 맥락에서 AI를 통합하는 것은 법 집행 과정에서 생성되는 방대한 양의 공공 데이터를 활용해야 할 필요성에서 비롯된다. 이러한 데이터를 활용하여 고품질의 효율적인 법 집행 서비스를 제공함으로써 새로운 가치를 창출하고 진화하는 대중의 요구를 충족할 수 있다.

또한, '스마트 치안 이니셔티브'(2021년 5월)[1], '법 집행에서의 인공지능'(2022년 1월)[2] 등 주요 법률의 시행으로 공공 안전 의사 결정 과정에서 인공지능의 참여와 역할이 크게 증가할 것으로 예상된다.

스마트 치안 기관/시설/정부/시의회/단체 등이 공공 안전 서비스 및 운영을 강화하고자 AI 도입을 적극적으로 추진하는 가운데, 데이터 및 알고리즘 편향으로 인한 편견, 프라이버시 침해, 개인의 권리 침해 가능성에 대한 우려도 제기되고 있다. 이러한 우려에 대응하여 유럽연합(EU) 등 국제기구는 'AI 기반 공공 안전 서비스 및 솔루션의 데이터 윤리 백서'(2020년 5월)에서 합법성, 윤리, 사회적 안전을 강조하며 공공 안전 서비스에서 AI의 윤리적·합법적 사용을 보장하는 필수 요소를 설명했다. 영국에서는 2019년 '공공 안전 분야에서의 인공지능 활용 가이드'라는 제목의 가이드[3]를 발간하여 법 집행 분야에서 인공지능의 활용과 확대를 촉진하고 있다. 한국에서는 2017년에 '지능정보사회를 위한 가이드라인'이 처음 발간되었다. 2019년 방송통신위원회는 인간 중심 서비스, 투명성, 설명 가능성, 책임성, 안전성, 비차별, 참여, 프라이버시, 데이터 거버넌스 등을 포괄하는 'AI 윤리 7대 원칙'을 발표했으며, 이는 스마트 치안 분야에도 융합적으로 적용될 수 있다. 또한, 최근 발표된 개인정보보호위원회의 '신뢰 기반 인공지능 데이터 규범'(2023년 8월)[4] 역시 데이터 규범을 확보하면서 신뢰 기반 인공지능 시스템 개발을 위한 원칙을 제시하고 있다.

02 스마트 치안 인공지능 신뢰성 동향

스마트 치안에서 신뢰할 수 있는 AI 개발을 보장하고자 유럽연합의 인공지능 고위급 전문가 그룹(AI HLEG)은 2019년 경제협력개발기구(OECD), 유럽연합 등의 기관에서 AI 신뢰성과 관련된 국제 표준 요구사항을 반영하여 '공공 안전에서 신뢰할 수 있는 AI를 구현하는 실무 가이드'를 발표했다.

그러나 국내외에서 발간된 AI 신뢰성 가이드는 주로 윤리적 관점에서 추상적인 항목을 제시하여 실제 법 집행 시나리오에서 실제 적용하는 데 한계가 있다. 따라서 AI 솔루션을 도입하려는 법 집행 기관은 검증 시스템을 구축하고 AI 신뢰성 특성을 충족하는 데 어려움을 겪고 있다.

본 개발 안내서는 스마트 치안 분야에서 AI 기술을 도입하는 데 실질적인 문제를 해결하고자 작성되었다. 치안 분야에 활용되는 AI 서비스의 신뢰성을 확보하도록 요구사항과 검증항목을 제시하고 있다. 본 개발 안내서가 스마트 치안 이니셔티브에 AI를 도입하고자 하는 관련 기업 및 개발자들에게 기초 자료로 활용되어 법적·윤리적 기준을 준수하면서 공공 안전 운영의 효과성과 효율성을 높일 수 있기를 기대한다.

02 스마트 치안 인공지능 신뢰성 동향

스마트 치안의 핵심 측면 중 하나는 개인과 공공의 안전을 보호하는 것이다. AI 기반 치안에 대한 신뢰를 구축하려면 커뮤니티의 적극적인 참여가 필요하다. 스마트 치안에서 AI의 신뢰성은 투명성, 공정성, 데이터 윤리, 책임성, 커뮤니티 참여와 관련된 다각적인 문제이다. AI 기반 치안에 대한 신뢰는 AI 개발자가 이러한 기술의 윤리적·합법적 안전한 사용을 우선시하고 우려 사항을 해결하고 대중의 신뢰를 구축하여 적극적으로 노력할 때만 달성할 수 있다. 본 개발 가이드를 통해 국내 스마트 치안 관련 기업 또는 기관들이 더욱 신뢰하는 스마트 치안 기술이나 시스템 및 서비스를 확보하는 데 기초 자료가 되기를 기대한다.

2.1. 스마트 치안 신뢰성 동향

최근 세계 주요 국가들은 표준 관련 기관 및 기술 단체와 함께 인공지능의 신뢰성을 제고하고자 상황별 맞춤형 대책을 적극적으로 추진하고 있다. 이 부분에서는 스마트 치안이 적용될 수 있는 분야를 살펴보고, 인공지능을 도입할 때 나타나는 도전 과제를 파악하며, 국내외에서 동시에 진행되는 정책 및 연구 동향을 살펴본다.

2.2. 스마트 치안 활용 영역

스마트 치안이란 인공지능(AI)을 비롯한 첨단 기술을 적용하여 치안 운영의 효율성, 효과성 및 전반적인 성과를 향상하는 것을 말한다. 이 접근 방식은 AI와 데이터 분석을 활용하여 의사 결정, 범죄 예방, 수사 및 지역 사회 안전을 지원한다.

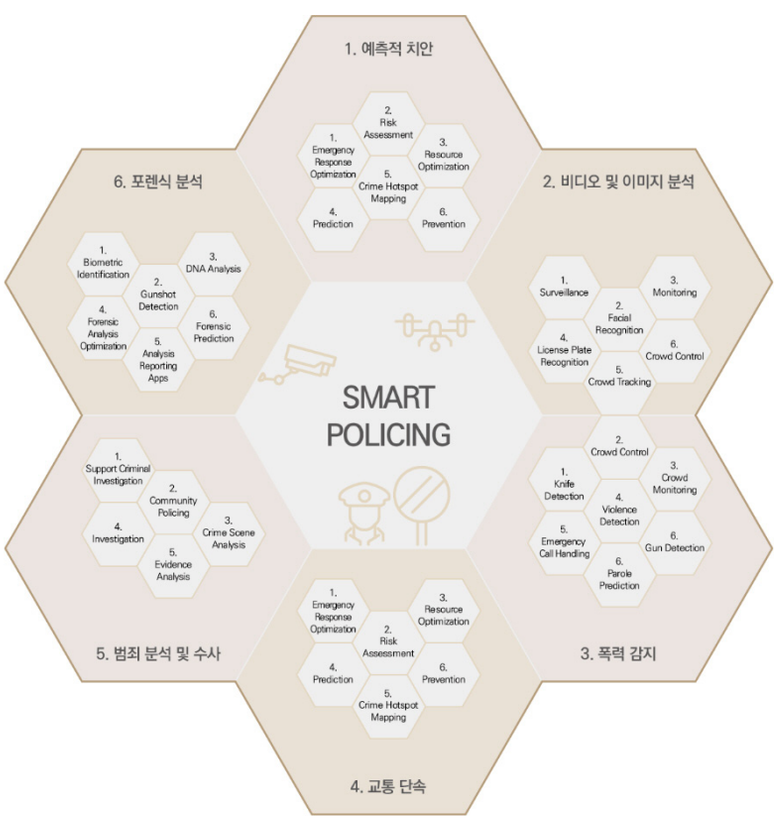
스마트 치안은 예측 치안, 감시, 증거 분석, 폭력 감지, 범죄 분석, 공공 안전 등 법 집행 기관의 다양한 영역에서 활용되고 있다. 또한 데이터 분석과 데이터 기반 의사 결정에 크게 의존한다.

이를 위해 법 집행 기관에서도 과거 범죄 데이터, 소셜 미디어 콘텐츠, 감시 영상 등을 포함한 방대한 양의 데이터를 수집해 분석한다. 이러한 데이터는 패턴, 추세 및 잠재적 우려 영역을 식별하는 데 사용되어 법 집행 전략에 정보를 제공한다. 개발된 시스템의 추론 결과가 개인의 공공 안전에 직접적인 영향을 미쳐 수집된 데이터의 편향성 완화 및 운영은 매우 중요하다.

치안 유지에 AI를 활용하는 주요 목적은 범죄율 감소, 공공 안전 강화, 법 집행 기관의 효율성 향상, 자원 배분 최적화이다. 또한 스마트 치안은 잠재적인 범죄 활동을 사전에 식별하여 과거 데이터, 범죄 데이터 등에 따라 사건을 예측함으로써 범죄가 발생하기 전에 법 집행 기관이 개입하도록 하는 것이 목표이다.

범죄 기록은 물론 소셜 미디어 활동까지 방대한 양의 데이터를 분석하여 패턴을 파악하고, 이상 징후를 감지하고, 법 집행 기관이 데이터 기반 의사 결정을 내리도록 지원하여 공공 안전을 지원하는 것이 치안 유지에 AI를 사용하는 진정한 목적이다. 스마트 치안 애플리케이션은 주요 기능과 목적에 따라 다음과 같이 몇 가지 주요 분야로 분류할 수 있다.

▼ 치안 분야에서 인공지능 활용 분야[5]



① **예측적 치안:** AI는 데이터를 분석하고 미래의 범죄 활동을 예측하는 데 사용되어 법 집행 기관이 효과적으로 자원을 할당하고 범죄가 발생하기 전에 예방하도록 한다. 그러나 AI 기반 예측 치안 시스템의 편견과 오용 가능성에 대한 우려가 있다. 예측 치안 시스템은 6가지 애플리케이션으로 나뉜다. 긴급 대응 최적화, 위험 평가, 자원 최적화, 예측, 범죄 핫 스팟 매핑, 예방이다.

② **비디오 및 이미지 분석:** AI 알고리즘은 감시 영상, 이미지 및 기타 시각적 데이터를 분석하여 관심 있는 사물, 사람, 이벤트를 식별할 수 있다. 이는 범죄 수사, 공공장소 모니터링, 법 집행 기관 직원 또는 시스템 사용자의 상황 인식 향상에 도움이 될 수 있다. 이러한 시스템은 사람이 접근하기에는 너무 위험하거나 접근하기 어려운 환경에서도 작동할 수 있어 운영의 전반적인 효율성과 안전성을 높일 수 있다. 비디오 및 이미지 분석은 6가지 애플리케이션으로 나뉜다. 감시, 얼굴 인식, 모니터링, 번호판 인식, 군중 추적, 군중 통제이다.

③ **폭력 감지:** 이것은 비디오 분석의 하위 영역으로 간주될 수 있지만, 행동 분석으로 인해 이 두 영역 사이에는 주요 차이점이 있다. AI는 감시, 수색 및 구조, 군중 통제 등 다양한 치안 업무를 하도록 드론 및 로봇 등 자율 시스템을 개발하고 배포하는 데 사용된다. 이를 통해 범죄나 위반 사항을 사전에 감지하고 사용자에게 사건에 대해 경고할 수 있다. 폭력 감지는 칼 감지, 군중 통제, 군중 모니터링, 폭력 감지, 긴급 전화 처리, 총기 감지, 가석방 예측 등 7가지 애플리케이션으로 나뉜다.

④ **교통 단속:** 교통 위반을 모니터링하고 교통법을 집행하고자 스마트 카메라와 번호판 인식 시스템에 AI가 사용된다. 교통 단속은 감시 및 추적, 신호등 제어, 드론 경찰 시스템, 번호판 인식, 교통 모니터링, 교통 관리 등 6가지 애플리케이션으로 나뉜다.

⑤ **범죄 분석 및 수사:** AI 도구는 대량의 데이터를 분석하고 패턴을 식별하며 범죄 수사를 지원할 인사이트를 제공하는 데 도움을 줄 수 있다. 범죄 분석 및 조사는 범죄 수사 지원, 지역 사회 치안, 범죄 현장 분석, 수사, 증거 분석 등 5가지 애플리케이션으로 나뉜다.

⑥ **포렌식 분석:** 지문 및 얼굴 인식과 같은 포렌식 분석을 강화하고 증거 처리의 정확성과 효율성을 개선하고자 AI 및 딥러닝 기술이 사용되고 있다. 포렌식 분석은 6가지 애플리케이션으로 나뉜다. 생체 인식, 총격 감지, DNA 분석, 포렌식 분석 최적화, 분석 기록 애플리케이션, 포렌식 예측이다.

이러한 범주는 스마트 치안 애플리케이션이 배포되는 주요 기능 및 분야를 나타낸다. 이러한 각 분야는 첨단 기술과 데이터 기반 접근 방식을 사용하여 법 집행 운영, 공공 안전 및 지역 사회 참여를 향상하는 것이 목표이다.

2.2. 스마트 치안 이슈 사례

스마트 치안 영역은 새로운 기술, 혁신적인 방법, 참신한 아이디어의 도입으로 인해 큰 변화를 보이고 있다. 이러한 법 집행의 진화는 데이터 분석, 인공지능, 통신 기술의 발전을 비롯한 여러 요인이 융합되어 촉진된다.

▼ 스마트 치안 사례

사례	설명
1	이 분야에서 주목할 만한 이니셔티브 중 하나는 범죄 감소 및 경찰 운영에 대한 혁신적인 접근 방식을 구현하는 프로젝트를 지원하는 2023 회계 연도 스마트 치안 이니셔티브 보조금 프로그램이다. 이 프로그램은 법 집행의 효율성을 높일 최첨단 솔루션을 채택하겠다는 의지를 보여 준다[6].
2	독립 자문 그룹이 수행한 치안 분야의 신흥 기술 검토는 법 집행에 새로운 기술과 신흥 기술을 통합하는 데 귀중한 통찰력을 제공했다. 이 보고서에는 기술을 활용하여 치안 관행을 개선할 연구 결과와 권장 사항이 요약되어 있다[7].
3	스마트 치안 관련 문헌에 대한 비판적 검토는 데이터 활용과 혁신적인 경찰 전략 간의 연관성을 보여 주며, 법 집행 분야의 핵심 발전 영역으로서 '스마트 치안'의 잠재력을 강조한다[8].
4	이러한 노력은 도시가 '스마트 시티'로 변모하는 법 집행의 광범위한 변화를 반영하며, 이러한 맥락에서 감시 기술과 치안의 의미에 대한 우려는 중요한 논의와 분석을 촉발한다[9].

기술이 계속해서 치안 환경을 재편함에 따라 이러한 발전이 사회적 가치와 기대에 부합하도록 윤리적·법적·사회적 차원을 고려하는 것이 필수적이다.

우리나라는 특히 사이버 범죄 치안과 관련하여 스마트 치안 문제를 적극적으로 해결해 왔다. 한국의 사이버 범죄 치안에 스마트 치안 기술을 도입하는 것에 대한 연구와 대중의 태도는 법 집행 기술의 최전선에 서려는 한국의 노력을 반영하는 연구 주제였다. 이 연구는 사이버 범죄와의 전쟁에서 스마트 치안 기술의 구현에 대한 한국 대중과 경찰의 다양한 태도를 분석하고자 했다[10]. 또한, 한국은 스마트 치안에 대한 다양한 측면이 논의되고 전시된 2023 대한민국 경찰 세계 엑스포 등의 행사[11]와 8월에 'ChatGPT'를 기반으로 한 범죄 데이터 학습 프로그램인 '폴리-엘렉트라'를 새롭게 선보인 것[12]에서 알 수 있듯이 스마트 치안에 대한 지속적인 관심을 보여 왔다. 이러한 이니셔티브는 스마트 기술을 통해 치안 역량을 강화하는 한국의 노력을 강조하고 있다.

2.3. 스마트 치안 정책 및 연구 동향

미국, 유럽, 캐나다, 한국을 포함한 주요 국가들은 스마트 치안 시스템을 발전시키는 데 인공지능(AI)의 신뢰성 확보가 중추적인 역할을 한다는 점을 인식하고 있다. 이들은 이러한 첨단 법 집행 기술의 지속적인 발전과 사회적·산업적 수용을 하도록 기본 요건을 마련했다. 해당 국가들은 치안 분야에서 AI의 신뢰성을 보장하는 정책을 적극적으로 옹호하고 시행하고 있다.

또한, 산업계와 학계 간의 시너지 효과는 스마트 치안 기술과 관련 구성 요소의 신뢰성을 강화하는 데 중요한 역할을 해 왔다. 이러한 협력 노력은 민간 부문으로 확대되어 AI 시스템의 신뢰성을 보장하는 지침을 수립할 중요한 연구 이니셔티브에 착수했다. 민간 부문의 주요 목표는 AI 신뢰성을 지속해서 평가하고 강화하는 자율적인 프레임워크를 만드는 것이다.

이와 관련하여 주요 국가들이 확고한 AI 신뢰성을 바탕으로 새로운 스마트 치안 시대를 열고자 힘을 모으는 것은 분명하다. 이들은 강력한 정책과 공동 연구를 통해 AI의 신뢰성을 일관되고 자율적으로 확인할 환경을 적극적으로 조성하고 있다.

▼ 주요국의 스마트 치안 신뢰성 관련 정책 동향

국가	주요 정책	특징
미국	심각한 범죄 문제를 해결할증거 기반 스마트 치안 이니셔티브(SPI)[1]	성과 측정 및 연구 파트너십
캐나다	토론토 경찰 서비스 위원회의 인공 지능(AI) 기술을 규율할 정책 초안 개발[13].	제안된 정책은 위험도에 따라 다섯 가지 범주를 설정: 극단적 위험, 고위험, 중간 위험, 저위험, 최저 위험. AI 기술의 다양한 적용에 따른 위험은 기술 개발 방식과 사용 방식에 따라 달라짐
	새로운 인공지능 기술 사용 정책 – 공개 자문[14]	스마트 치안 기술의 신뢰성을 높이고 효과성, 공정성, 정당성을 보장함과 동시에 안전 및 보안의 필요성과 프라이버시 및 인권의 균형을 맞추는 것을 목표로 함
유럽연합 (EU)	신뢰할 수 있는 인공지능을 위한 윤리 가이드라인(2019)	인공지능 사용으로 인해 발생하는 위험, 편견 및 기타 사회 윤리적 영향을 분석
	인공지능법[15]	인공지능 시스템에 다양한 버전의 투명성을 요구함. 고위험 AI의 의미를 정의하지 않고 충분한 투명성과 인간의 감독을 요구함. 인공지능법에서는 개인과 대중에게 특정 정보를 공개하도록 의무화. 또한 AI 시스템이 편견을 최소화하도록 설계되어야 하며 정기적으로 모니터링되어야 한다는 점을 강조.
	AI 사용을 위한 MEPs 확장 목록[16]	MEPs는 침투적이고 차별적인 AI 사용에 대한 금지를 포함하도록 목록 확장. 이 규칙은 위험 기반 접근 방식을 채택하여 AI 시스템 제공업체와 배포자에게 관련된 위험 수준에 따라 의무를 부과. 소셜 스코어링과 같이 허용할 수 없는 위험 금지. 실시간 및 사후 원격 생체 인식 시스템, 민감한 생체 인식 분류, 예측적 치안, 다양한 맥락에서 감정 인식, 공공 출처에서 얼굴 이미지의 비표적 스크래핑 등 침투적이고 차별적인 AI 사용을 포함하도록 금지 범위를 확대.
대한민국	사이버 범죄 치안에 스마트 치안 기술 도입[10]	혁신적인 방식으로 기술, 인텔리전스 및 데이터 사용
	범죄 위험도 예측 및 분석 시스템 소개: Pre-CAS (범죄 위험 예측 분석 시스템)	보안 및 공공 데이터의 통합 빅데이터를 활용한 최신 알고리즘 적용, AI를 활용한 범죄 위험도 분석 및 범죄 발생 건수를 예측

▼ 국내외 주요 산·학·연 스마트 치안 동향

국가	교육 기관 이름	활동 및 내용
한국	한국 인용 색인(KCI)	범죄 수사 및 개인정보 사용 위반을 조사 중인 연구 발표[17]
	경찰대학교	회귀 분석을 사용하여 범죄 발생과 지역의 공간적 영향을 조사[18]
	경기대학교	지역 사회의 전체 범죄와 관련하여 저지른 범죄 패턴과 낯선 사람의 거주지 또는 활동 분포를 조사[19]
	서울대학교	도시 범죄를 조사하고 서울, 인천, 경기 지역의 5대 범죄(살인, 강간/추행, 강도, 절도, 폭행)의 공간적 분포 특성을 분석한 연구[20]
전세계	스마트 치안 등 감시 목적의 AI 기술 사용 증가[21]	범죄 예방, 대응, 예측을 위한 알고리즘에 방대한 양의 데이터 제공
미국	뉴욕 경찰청	도메인 인식 시스템(DAS)은 뉴욕시 경찰국(NYPD)에서 개발한 소프트웨어로, 서로 다른 격리된 데이터 공간에 보관되는 중요한 정보를 효율적으로 중앙 집중화. DAS 영향 및 사용 정책은 NYPD와 파트너가 시스템을 책임감 있고 효과적으로 사용할 지침과 절차를 간략하게 설명

03 안내서 마련 과정

발간 배경에서 밝힌 바와 같이, 국내외로 스마트 치안에 대한 여러 가지 고려 사항 및 법적 제정이 이루어지며 많은 기관에서 관련 정보를 제시하나, 기술적 관점에서 상세한 구현 방법론을 정리한 사례는 없었다. 고려 사항

따라서 이 가이드는 데이터 과학자, 모델 개발자, 경찰, 공무원 등 이해관계자가 스마트 치안 시스템 개발 영역에서 실질적인 신뢰성을 확보하는 데 활용하는 포괄적인 프레임워크를 제공하고자 한다.

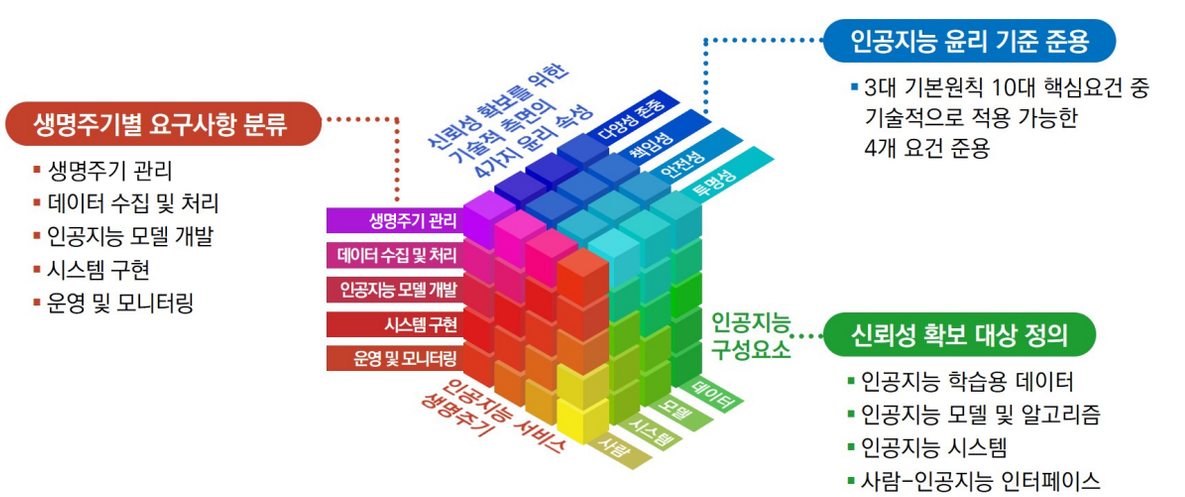
이러한 목표를 달성하고자 우리는 스마트 치안 분야의 특정 요구사항에 맞춘 전문 가이드를 개발했다. 치안을 위한 가이드 개발 과정에서 학계와 산업계의 전문가와 실무자들의 의견을 적극적으로 수렴했다.

또한 스마트 치안 및 스마트 치안 서비스를 제공하는 기업과의 협업을 촉진하는 것을 목표로 삼았다. 이들 기업과 공동 연구하여 사례 연구를 작성하고 피드백을 수렴하여 실제 적용 방안을 논의하고 컨설팅 활동을 진행함으로써 실질적인 활용도를 높이고자 노력했다.

3.1. 인공지능 신뢰성 프레임워크

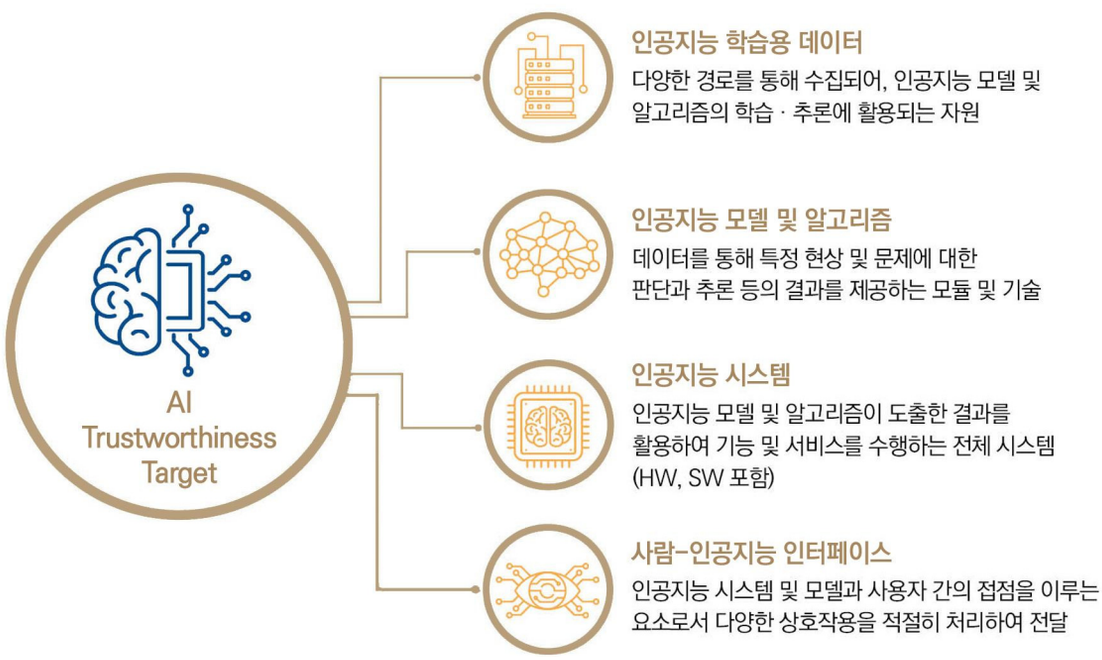
안내서 개발 과정 중 가장 우선해서 신뢰성 확보에 어떤 요소들이 실무적으로 고려되어야 하는지 탐색해 보았고, 그 결과 세 가지 설계 요소를 도출하여 안내서에 반영했다. 각 설계 요소는 요구사항과 검증항목 마련 시 모두 반영되었으며, 이러한 접근 방식을 아래 그림과 같이 매트릭스 형태로 체계화하여 '인공지능 신뢰성 프레임워크'로 정의했다. 이 프레임워크는 스마트 치안 분야뿐만 아니라 일반 분야 및 기타 산업에도 동일하게 적용된다.

▼ 인공지능 신뢰성 프레임워크



첫 번째는 인공지능 구성 요소이다. 인공지능을 구성하는 4가지 요소는 학습과 추론 기능을 수행하는 인공지능 모델 및 알고리즘, 인공지능 학습용 데이터, 실제 기능을 구현할 시스템, 사용자와 상호 작용하는 인터페이스가 있다. 각 구성 요소는 개별적으로 또는 통합적으로 인공지능 서비스의 생명주기에 따라 개별적으로 또는 통합적으로 개발, 검증 및 운영된다. 따라서 각 구성 요소의 신뢰성 확보 방안을 고민하고, 각 요소에 대한 요구사항과 검증항목을 제시하고자 했다. 각 요소에 대한 신뢰성 확보 방안은 다음과 같다.

▼ 인공지능 서비스 구성 요소



인공지능 서비스 구성 요소	신뢰성 확보 방안
인공지능 학습용 데이터	인공지능 학습 및 추론 과정에 활용하는 데이터를 대상으로 편향성 등이 배제되었는지 검증
인공지능 모델 및 알고리즘	인공지능이 모델 및 알고리즘에 따라 안전한 결과를 도출하며, 이에 대한 설명이 가능한지, 악의적인 공격에 강건한지 등을 검증
인공지능 시스템	인공지능 모델 및 알고리즘이 적용된 전체 시스템을 대상으로 인공지능이 추론한 대로 작동하는지, 인공지능이 잘못 추론한 경우의 대책이 존재하는지 등을 검증
사람-인공지능 인터페이스	인공지능 시스템 사용자·운영자 등이 인공지능 시스템의 동작을 쉽게 이해할 수 있으며, 인공지능 오작동 시 사람에게 알리거나 제어권을 이양하는지 등을 검증

두 번째, 인공지능 서비스 생명주기는 첫 번째에서 살펴본 인공지능 서비스 구성 요소들을 구현하고 운영하는 일련의 절차를 말한다. 기존 소프트웨어 시스템에서 다루는 엔지니어링 프로세스 및 생명주기와 유사하지만, 인공지능의 특성상 데이터 처리와 모델 개발 단계가 별도로 필요하고, 그 외 단계에서는 주요 활동의 정의가 조금씩 다르다. 현재 인공지능

또는 인공지능 서비스의 생명주기는 많은 문헌에서 6~8단계로 구분한다. 대표적으로 OECD와 ISO/IEC에서 제시한 생명주기가 있는데, 본 가이드는 두 기관에서 제시한 생명주기를 대표적인 사례로 참고하여, 실무자들이 쉽게 활용하도록 생명주기 단계의 성격과 활동을 왜곡하지 않는 선에서 아래와 같이 5단계로 정리하였다.

▼ 인공지능 서비스 생명주기별 주요 활동

생명주기 단계	주요 활동
1. 생명주기 관리	- 인공지능 시스템 관리 감독 조직 및 방안 마련 - 인공지능 시스템 위험 요소 분석 및 대응 방안 마련
2. 데이터 수집 및 처리	- 데이터 품질 확보, 데이터 사용자의 이해를 위한 정보 제공 방안 마련 - 데이터 라벨링 및 데이터셋 특성feature 문서화 - 인공지능 모델을 구축할 데이터셋 마련
3. 인공지능 모델 개발	- 비즈니스 목적에 따른 인공지능 모델 구현 - 구현된 인공지능 모델 확인 및 검증 - 인공지능 모델 튜닝, 데이터 분석, 추가로 필요한 데이터 수집 - 인공지능 모델에 대한 성능 평가
4. 시스템 구현	- 문제 발생 대비 안전 모드 구현 및 알림 절차 수립 - 인공지능 시스템 검증 및 사용자 설명에 대한 평가
5. 운영 및 모니터링	- 시스템 모니터링 및 인공지능 모델 재학습을 통한 성능 보장 - 모델 편향 탐지, 공정성, 설명 가능성 등 시스템 신뢰성 모니터링 - 치명적 문제 발생 시 해결 방안 마련

인공지능 서비스의 생명주기 단계는 반복적·순환적인 성격을 갖지만 반드시 순차적인 것은 아니다. 본 개발 안내서는 이해를 돕고자 1~5단계를 순차적으로 설명하였으나, 실제 데이터를 수집하고 가공하거나 모델을 개발, 운영하는 과정에서 순서는 달라질 수 있다.

세 번째, 인공지능 신뢰성에 필요한 요건을 정의하고자 '인공지능 윤리 기준'의 10대 핵심 요건을 준용하여 기술적 관점에서 필요한 요구사항과 검증항목으로 '다양성 존중', '책임성', '안전성', '투명성'을 도출하였다.

EC, OECD, IEEE, ISO/IEC 등 국제 기구에서는 인공지능 신뢰성의 하위 속성들을 세분화해서 제시한다. 특히 ISO/IEC 24028:2020 - Overview of trustworthiness in artificial intelligence는 신뢰성 확보에 필요한 고려 사항의 형태로 키워드를 제공한다. 여기에는 투명성, 통제 가능성, 견고성, 회복 탄력성, 공정성, 안전성, 개인 정보 보호, 보안성 등이 포함되나, 키워드 간 관계나 신뢰성과의 연관성은 정의되지 않았다. 이처럼 관점에 따라 유사해 보이지만 조금씩 다른 용어들이 여러 문헌에서 제각각 달리 정의되며, 아직 합의된 속성 분류나 정의가 없는 상태이다. 이에 앞서 언급한 EC, OECD, IEEE, ISO/IEC 등 다양한 기관에서 제시한 속성과 키워드를 종합적으로 분석하고, 국내 학계 전문가들의 의견을 수렴하여 합의점을 모색하였다. 이러한 폭넓은 의견 공유 과정을 통해 인공지능 신뢰성 속성을 도출한 후, 국가 인공지능 윤리 기준의 10가지 요구사항에 대응하는 기술적 요구사항을 최종 선정했다. 각 요구사항에 대한 정의는 다음과 같다.

▼ 인공지능 신뢰성 특성

신뢰성 특성	정의
다양성 존중	인공지능이 특정 개인이나 그룹에 대한 차별적이고 편향된 관행을 학습하거나 결과를 출력하지 않으며, 인종·성별·연령 등의 특성과 관계없이 모든 사람이 평등하게 인공지능 기술의 혜택을 받을 수 있는 것 – 관련 속성: 공정성·공평성fairness, 정당성justice – 관련 키워드: 편향bias, 차별discrimination, 편견prejudice, 다양성diversity, 평등equality – 국제 표준(ISO/IEC TR 24027:2021 – Bias in AI systems and AI aided decision making)에서는 공정성을 정의하지 않는다. 공정성은 복잡하고 문화·세대·지역 및 정치적 견해에 따라 다양하여 사회적으로나 윤리적으로 일관되게 정의하기 힘들기 때문이다.
책임성	인공지능이 생명주기 전반에 걸쳐 추론 결과에 대한 책임을 보장하는 메커니즘이 마련된 것 – 관련 속성: 책무성responsibility, 감사 가능성auditability, 답변 가능성answerability – 관련 키워드: 책임liability – 국제 표준(ISO/IEC TR 24028:2020) – Overview of Trustworthiness in artificial intelligence)에서의 정의: 엔터티의 작업이 해당 엔터티에 대해 고유하게 추적되도록 하는 속성
안전성	인공지능이 인간의 생명·건강·재산 또는 환경을 해치지 않으며, 공격 및 보안 위협 등 다양한 위험에 대한 관리 대책이 마련된 것 – 관련 속성: 보안성security, 견고성·강건성robustness, 성능 보장성reliability, 통제 가능성·제어 가능성controllability – 관련 키워드: 적대적 공격adversarial attack, 회복탄력성resilience, 프라이버시privacy – 국제 표준(ISO/IEC TR 24028:2020)에서의 정의: 용인할 수 없는 위험risk으로부터의 자유
투명성	인공지능이 추론한 결과를 인간이 이해하고 추적할 수 있으며, 인공지능이 추론한 결과임을 알 수 있는 것 – 관련 속성: 설명 가능성explainability, 이해 가능성understandability, 추적 가능성traceability, 해석 가능성interpretability – 관련 키워드: 설명 가능한 인공지능XAI, eXplainable AI, 이해도comprehensibility – 국제 표준(ISO/IEC TR 29119-11:2020 – Guidelines on the testing of AI-based systems)에서의 정의: 시스템에 대한 적절한 정보가 관련 이해관계자에게 제공되는 시스템의 속성

위와 같이 인공지능 신뢰성을 확보할 다양한 속성이 있으며, 각 신뢰성 속성에 대한 정의를 파악하는 것뿐만 아니라 신뢰성 속성 간 상호 의존 관계 역시 중요하게 고려되어야 한다. 예를 들어, 인공지능 서비스에 대한 과도한 투명성 요구는 프라이버시 관련 위험을 초래할 수 있다. 또한, 설명 가능성만으로는 투명성을 보장하기에 부족하지만, 설명 가능성은 투명성을 확보하는 데 중요한 요소 중 하나이다. 따라서, 인공지능 신뢰성 속성에 대한 충분한 이해를 바탕으로 인공지능 서비스를 제공하는 것이 중요하며, 해당 인공지능 서비스가 고려한 속성에 대해 적절하게 이행하는지 지속해서 검토해야 한다.

3.2. 스마트 치안 시스템의 주요 고려 사항 반영

본 안내서는 기술적 관점에서 상세한 방법론을 제시함으로써, 스마트 치안 시스템 및 서비스 개발 현장에서 실무자가 신뢰성 확보에 참고할 실무 지침서 성격의 자료를 지향한다. 따라서, 본 안내서는 일반 분야에서 다루는 구성 요소 및 생명주기를 바탕으로, 인공지능의 신뢰성을 확보하는 데 고려되어야 할 요소들을 스마트 치안 분야에 특화하여 정리했다.

첫 번째, 본 안내서에서 신뢰성 대상으로 다루는 스마트 치안의 범위는 스마트 치안 시스템에 사용될 모든 범위를 포함하지 않는다. 본 가이드는 탐지, 예측, 감시, 인식, 분석, 분류 활동에 직간접적으로 활용되는 인공지능을 대상으로 하며, 본문의 원활한 이해를 돕고자 필요한 경우 치안의 범위에 대한 몇 가지 예를 포함한다.

우리의 접근 방식은 첨단 기술과 공공 안전을 보장해야 하는 중대한 필요성의 교차점을 우선시한다. 이를 위해 포괄적인 시스템 관리, 위험 완화, AI 모델 개발, 배포 및 지속적인 모니터링을 안내하는 명확한 가이드라인 개발을 강조한다. 이러한 접근 방식은 지역 사회의 안녕이 걸린 스마트 치안 분야에서는 책임감 있고 윤리적인 AI 구현이 무엇보다 중요하다는 점을 인정한다. 우리는 혁신적인 기술과 스마트 치안 시스템에 내재한 윤리적 의무 및 사회적 책임을 융합하는 데 전념하고 있다.

▼ 스마트 치안 범위

대한민국에서는 경찰관 직무 집행법(이하 경찰법)을 통해 국민의 자유와 권리를 보호하고 사회의 공공질서를 유지하고자 경찰관이 직무를 수행하는 데 필요한 사항을 규정하고 있다. 경찰법에 따르면 경찰관은 다음 사항을 담당한다[22]:

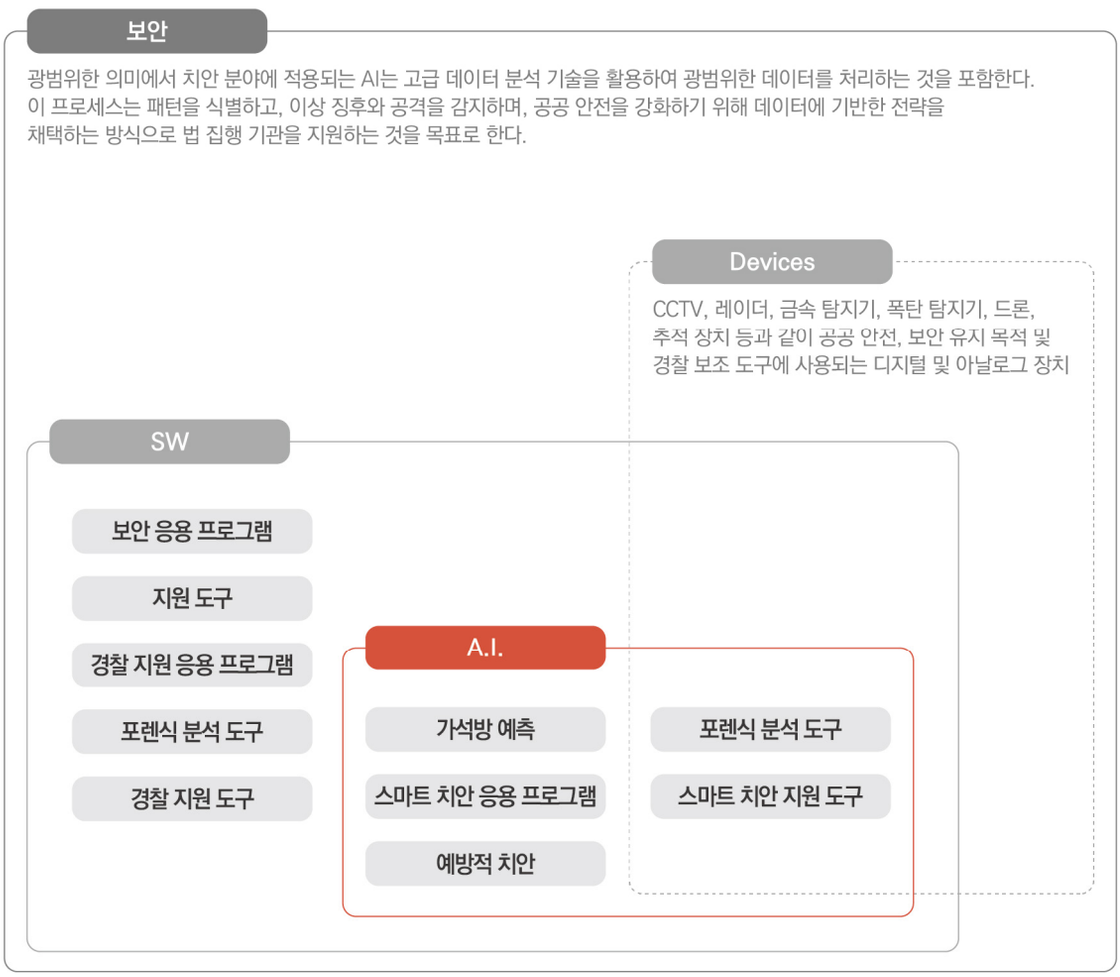
1. 국민의 생명, 신체 및 재산 보호
2. 범죄의 예방, 진압 및 수사

2.2. 범죄 피해자 보호

3. 주요 인사의 경호, 보안, 대간첩 및 대테러 작전 수행
4. 공공 안전에 대한 위험 예방 및 대응을 위한 정보의 수집, 생성 및 배포
5. 교통 통제 및 교통 위험 방지
6. 외국 정부 기관 및 국제기구와의 국제 협력
7. 기타 공공의 평화와 질서 유지

치안 분야에서 인공지능은 대부분 경찰관의 업무를 보조하고 공공의 안전을 유지하는 데 중요한 역할을 한다.

03 안내서 마련 과정



신뢰성 문제를 해결하고자 치안 분야의 AI는 엄격한 규정과 윤리 지침을 준수해야 한다. AI 시스템의 투명성과 책임성은 필수적이다. 치안 분야에 AI를 도입하면 안전 운영을 향상할 잠재력이 있지만, 신중하고 윤리적으로 그리고 기술이 신뢰할 수 있고 사회에 유익한지 확인하는 데 중점을 두고 수행해야 한다. 공공의 신뢰를 구축하고 법치를 유지하려면 치안 분야에서 AI의 이점과 개인 정보 보호, 편견, 윤리적 고려 사항의 균형을 맞추는 것이 필수적이다.

스마트 치안은 기술과 데이터를 활용하여 공공 안전을 개선한다. 그러나 이는 또한 신중하게 해결해야 하는 윤리 및 개인 정보 보호 문제를 야기한다. 스마트 치안이 법 집행 기관과 해당 기관이 서비스를 제공하는 지역 사회 모두에 혜택을 주려면 AI와 데이터를 책임감 있고 윤리적으로 사용하는 것이 중요하다.

두 번째, 스마트 치안 서비스를 구성하는 네 가지 필수 구성 요소를 다음 범위의 맥락에서 살펴보았다. 이 가이드는 주로 경찰, 공공 안전 및 보안 요원을 지원하는 데 사용되는 도구 및 애플리케이션과 관련 데이터 수집을 포함하는 스마트 치안 소프트웨어에 중점을 둔다. '인간-인공지능 인터페이스'에 관한 부분에서는 도메인 전문가와 공무원이라는 두 가지 주요 사용자 범주를 구분하여 방향성 있는 접근 방식을 제시한다.

▼ 스마트 치안 분야 인공지능의 서비스 구성 요소

구성요소	설명
인공지능 학습용 데이터	생체 정보 데이터, 이미지, 음성, 센서 데이터, 레이더 데이터, 범죄 기록, 사건 보고서, 비디오 영상 등을 통해 인공지능 학습 및 추론 과정에 활용하는 데이터에 편향성 및 공평성 등이 배제되었는지 검증
인공지능 모델 및 알고리즘	치안 분야의 인공지능 모델 및 알고리즘으로 안전한 결과를 도출하며 악의적인 공격에 강건한지 검증
인공지능 시스템	스마트 치안 인공지능이 판단 및 추론한 대로 작동하는지, 오류가 발생했을 때 대비 및 대책이 존재하는지 검증
사람-인공지능 인터페이스	스마트 치안 시스템이 경찰, 공공 기관, 법 집행관 또는 분석가, 운영자에게 치안과 관련된 정보를 이해하기 쉽게 접근하고 오작동을 방지하도록 인터페이스를 제공하는지 검증

세 번째, 치안 분야 인공지능 서비스의 생명주기는 스마트 치안 영역의 고유한 수요에 맞춘 다양한 활동을 포괄한다. 스마트 치안과 공공 안전 사이의 중요한 연결 고리를 인식하고 종합적인 시스템 관리와 사전 예방적 위험 대응 조치 수립에 우선순위를 두었다.

▼ 스마트 치안 서비스 생명주기별 주요 활동

생명주기 단계	설명
1. 생명주기 관리	비즈니스 이해, 표준 기반 관리 체계 구축 스마트 치안 시스템과 관련된 위험 요소 분석 및 대응 계획 수립 스마트 치안 시스템 개발에 필요한 각종 사전 승인 절차에 대한 확인 및 검증
2. 데이터 수집 및 처리	관련 데이터셋의 품질을 보장하고 데이터 사용자의 이해를 돕는 정보 제공 공공 보안과 관련된 공무원, 법 집행 담당자, 법률 고문, 범죄 분석가 등과 협력하여 치안 시스템을 위한 데이터 수집 및 처리
3. 인공지능 모델 개발	감시, 예측, 탐지, 분류, 예방, 리소스 할당 등 특정 애플리케이션에 맞는 인공지능 모델을 배포하고 검증 인공지능 모델의 성능을 평가하고 가상 테스트 시나리오를 포함한 테스트 계획을 수립 AI 모델의 편향성을 해결하고 감소/완화하는 전략을 수립
4. 시스템 구현	안전 모드를 구현하고 문제 발생 시 이해관계자에게 알리는 프로토콜을 수립 스마트 치안 시스템의 검증 프로세스에 대한 평가를 수행하고 시스템 사용자를 위한 사용자 친화적인 설명서를 개발
5. 운영 및 모니터링	스마트 치안 시스템 의 올바른 이해와 사용을 위한 교육 훈련 모델 편향성 감지, 공정성 보장, 설명 제공 등 시스템에 대한 신뢰성 모니터링 절차를 설정 문제 발생 시 이를 해결하는 대응 계획을 수립

3.3. 요구사항 및 검증항목 도출

다음 단계로 치안 분야 인공지능과 관련된 구체적인 요구사항과 검증항목을 도출했다. 우선 표준화 기구, 기술 단체, 국제기구, 주요국 에서 발표한 정책, 권고안, 표준 등을 기반으로 치안 분야 인공지능 신뢰성 확보에 필요한 기술 요구사항을 도출하고 구체화를했다. 이 과정에서 표준, 법안 및 규정을 신중하게 고려하여 신뢰와 효과를 최고 수준으로 보장하는 것이 가장 중요했다.

또한 ISO 13482:2014, 로봇 및 로봇 장치 – 개인 간호 로봇에 대한 안전 요구사항, ISO/IEC TR 24030, 정보 기술 – 인공지능(AI) – 사용 사례 및 ISO 22322:2022, 보안 및 회복 탄력성 – 비상 관리 – 공공 경보에 필요한 지침은 현장의 극단적인 요구사항을 강조하여 심사 대상에 포함하였다.

이와 함께 스마트 치안의 신뢰성을 높이고자 국내외에서 발표된 문서를 검토했다. 이 과정에서 중요한 정보를 개발 가이드에 통합하고 중복되는 내용은 제거하거나 압축했다. 참고 문헌은 다음과 같다.

▼ 인공지능 신뢰성 관련 주요 참고 문헌

기관명	발간 연원	권장 사항 및 표준 명칭
OECD	2019.05	인권과 민주적 가치를 존중하면서 혁신과 신뢰성을 촉진하는 인공지능(AI)에 대한 일련의 원칙
한국	2020	지능 정보화 기본법. 인공지능 영향 평가에 관한 법률
	2021.05	인권, 안전, 민주주의를 보장하는 AI 정책을 촉구하는 선언문
	2023.08	개인 정보 보호 위원회(PIPC)는 신뢰 기반 인공지능 데이터 규범에 대한 가이드라인을 발표[4]
미국(뉴욕 경찰청)	2021.04	Microsoft와 함께 개발한 네트워크형 도메인 인식 시스템(DAS)인 테러 공격 탐지 및 방지 도구의 영향 및 사용 정책에 대한 문서
영국	2020.10	Met 연구 윤리 위원회(MetREC)는 연구의 윤리적 고려 사항에 대해 메트로폴리탄 경찰에 독립적인 조언을 제공하며, 영국 최초의 치안 관련 연구 윤리 위원회
WEF	2021.06	조직을 위한 9가지 핵심 윤리적 AI 원칙
	2022.03	법 집행 기관에서 안면 인식 기술을 책임감 있게 사용하고자 9가지 원칙을 설명하는 백서를 발간. 이 백서는 유엔 지역 형사 사법 연구소(UNICRI), 인터폴, 네덜란드 경찰과 협력하여 개발
버지니아	2022.04	법 집행 기관의 기술 사용을 14가지 열거된 목적으로 제한하는 안면 인식 법안 제정
NIST	2023.01	AI 위험 관리 프레임워크(AI RMF)인 AI 거버넌스 솔루션
IEEE	2017.03	아동 및 학생 데이터 거버넌스를 위한 IEEE P7004 표준
	2019.03	IEEE P7001 자율 시스템의 투명성
	2019.03	IEEE P7002 데이터 개인 정보 보호 프로세스
국제표준화기구 (ISO/IEC)	2010.03	ISO 9241-210: 2010, 인간-시스템 상호 작용의 인체 공학 – 파트 210: 대화형 시스템을 위한 인간 중심 설계
	2014.02	ISO 13482:2014, 로봇 및 로봇 장치 – 개인 간호 로봇에 대한 안전 요구사항
	2018.02	ISO 31000:2018, 위험 관리 – 원칙 및 지침
	2020.05	ISO/IEC TR 24028:2020, 정보 기술 – 인공지능 – 인공지능의 신뢰성 개요
	2021.05	ISO/IEC TR 24030, 정보 기술 – 인공지능(AI) – 사용 사례
	2021.11	ISO/IEC TR 24027:2021, 정보 기술 – 인공지능(AI) – AI 시스템 및 AI 지원 의사 결정의 편향성
	2022.04	ISO/IEC 38507:2022, 정보 기술 – IT 거버넌스 – 조직의 인공지능 사용에 따른 거버넌스 영향
	2022.12	ISO 22322:2022, 보안 및 회복 탄력성 – 비상 관리 – 공개 경고 지침
	2023.02	ISO/IEC 23894:2023, 정보 기술 – 인공지능 – 위험 관리에 대한 지침

이를 통해 도출된 최종 요구사항은 아래 표와 같으며, 인공지능 윤리의 핵심 요구사항에 해당하는 결과도 표시되어 있다.

▼ 인공지능 신뢰성을 확보하는 기술적 요구사항과 윤리 요건 매칭 결과

요구사항	다양성 존중	책임성	안전성	투명성
요구사항 01 인공지능 시스템의 위험 관리 계획 및 수행		✓		✓
요구사항 02 인공지능 거버넌스 체계 구성	✓	✓	✓	✓
요구사항 03 인공지능 시스템의 신뢰성 테스트 계획 수립			✓	✓
요구사항 04 인공지능 시스템의 추적가능성 및 변경 이력 확보		✓		✓
요구사항 05 데이터 활용을 위한 상세 정보 제공		✓		✓
요구사항 06 데이터 견고성을 확보할 이상 데이터 점검			✓	
요구사항 07 수집 및 가공된 학습 데이터의 편향 제거	✓	✓		✓
요구사항 08 오픈 소스 라이브러리의 보안성 및 호환성 점검		✓	✓	
요구사항 09 인공지능 모델의 편향 제거	✓			
요구사항 10 인공지능 모델 공격에 대한 방어 대책 수립			✓	
요구사항 11 인공지능 모델 명세 및 추론 결과에 대한 설명 제공		✓		✓
요구사항 12 인공지능 시스템 구현 시 발생 가능한 편향 제거	✓			
요구사항 13 인공지능 시스템의 안전모드 구현 및 문제 발생 알림 절차 수립		✓	✓	✓
요구사항 14 인공지능 시스템의 설명에 대한 사용자의 이해도 제고				✓
요구사항 15 서비스 제공 범위 및 상호 작용 대상에 대한 설명 제공		✓		✓

3.4. 현장 적용 및 전문가 의견 수렴

신뢰성을 확보하고자 요구사항을 도출한 후에는 각 항목을 기술적 타당성, 효용성 및 포괄성 등의 관점에서 검토한 후 고도화했다. 각각의 세부 검증항목이 요구사항에 해당하는 내용이 맞는지(타당성), 개발 현장에서 실무적으로 활용 가능한 내용인지(효용성), 검증할 내용들이 과거부터 지금까지 연구 내용을 폭넓게 포함하는지(포괄성) 확인했다. 이를 위해 법 집행 공무원, 범죄 분석가, 기술 전문가, 스마트 치안 분야 산업계 전문가 등 다양한 이해관계자 그룹이 검토 및 자문 과정에 적극적으로 참여했다. 이들의 피드백과 인사이트는 내용을 구체화하는 데 큰 도움이 되었다.

또한 업계 및 학계 연구자, 기업 기획자, 개발 프로젝트 리더, 교수, 도메인 전문가와 협업해 관점을 더욱 풍부하게 했다. 이와 함께 스마트 치안 및 스마트 치안 서비스 전문 기업들과 협력하여 실제 적용할 연구 내용을 검토하고, 사례 연구를 준비하며 실제 사용성을 높이고자 소중한 피드백을 얻었다.

04 안내서 활용 대상

4.1. 활용 대상(대표 이해관계자·협력 대상) 정의 배경 및 기준

본 개발 안내서는 스마트 치안 분야 인공지능 제품 및 서비스의 개발 과정에 참여하는 다양한 조직과 개인이 활용할 수 있다. 특히, 기술적 관점에서 신뢰성에 중점을 두어야 하는 기획자, 아키텍트, 개발자, 품질관리자 등의 이해관계자들이 주요 대상이다. 이해관계자들은 제품·서비스의 신뢰성을 확보하기 위해 요구사항을 충족시키는 데 주력해야 하며, 이는 아래에 제시된 표를 통해 확인할 수 있다. 물론, 신뢰성과 연관된 문제가 발생했을 때 관련된 모든 책임을 이해관계자가 부담해야 한다는 의미는 아니다. 대표 이해관계자는 인공지능 생명주기 단계마다 요구사항을 만족 시키기 위한 대책을 수립하며, 자가 검증 시 각 검증항목의 만족 여부를 체크하는 주요 역할을 담당한다. 이 과정에서 효과적인 협력 체계의 필요성이 강조된다. 따라서, 대표 이해관계자는 한 명 이상의 협력 대상과 긴밀하게 협력하며, 이들 간의 협력 관계는 부록A에 기술되어 있다.

대표 이해관계자와 협력 대상은 한국SW산업협회KOSA가 국가직무능력표준NCS를 기반으로 개발한 IT분야역량체계 ITSQF에 근거해 정립되었다. 이를 통해, 국내 기업들이 본 개발 안내서를 활용하고자 할 때 참고할 수 있도록 하였다. 또한, 각 기업의 다양한 작무 체계에 맞게 적용하기 위해, 부록B에 제시된 각 직업·직무에 대한 정의를 참고하여 직무별 역할을 확인할 수 있다.

▼ 인공지능 생명주기 단계별 신뢰성 대표 이해관계자

생명주기 단계	대표 이해관계자(예)	관련 요구사항
1. 생명주기 관리	• 정보기술기획자 • IT감사자 • IT품질관리자	– 인공지능 시스템에 대한 위험관리 계획 및 수행 – 인공지능 거버넌스 체계 구성 – 인공지능 시스템의 신뢰성 테스트 계획 수립
2. 데이터 수집 및 처리	• 데이터아키텍트 • 데이터분석가	– 데이터의 활용을 위한 상세 정보 제공 – 데이터 견고성 확보를 위한 이상 데이터 점검 – 수집 및 가공된 학습 데이터의 편향 제거
3. 인공지능 모델 개발	• 인공지능SW개발자 • 인공지능아키텍트	– 오픈 소스 라이브러리의 보안성 및 호환성 확보 – 인공지능 모델의 편향 제거 – 인공지능 모델 공격에 대한 방어 대책 수립 – 인공지능 모델 명세 및 추론 결과에 대한 설명 제공
4. 시스템 구현	• 시스템SW개발자 • SW아키텍트 • UI/UX기획자	– 인공지능 시스템 구현 시 발생 가능한 편향 제거 – 인공지능 시스템의 안전 모드 구현 및 문제발생 알림 절차 수립 – 인공지능 시스템의 설명에 대한 사용자의 이해도 제고
5. 운영 및 모니터링	• 데이터베이스관리자 • 인공지능서비스기획자	– 인공지능 시스템의 추적가능성 및 변경이력 확보 – 서비스 제공 범위 및 상호작용 대상에 대한 설명 제공

4.2. 활용 기업 및 서비스 유형에 따른 적용 방안

본 개발 안내서는 다양한 규모와 형태의 기업과 기관에 적용될 수 있으며, 이에 따라 대표 이해관계자와 협력 대상의 직무 체계나 활동 범위가 달라질 수 있다. 특히, 스타트업과 같은 소규모 기업에서는 몇 명의 인력만으로 전체 활동을 수행할 수도 있을 것이다. 만약 대표 이해관계자의 직무를 수행하는 인력이 없다면 한 명 이상의 협력 대상이 그 역할을 맡을 수도 있다.

또한, 기업에는 제공하는 인공지능 서비스 유형에 따라 적용 방안이 달라질 수 있다. 다음 페이지에서 제시한 대표 이해관계자 및 협력 대상의 분류는 소비자 대상B2C, Business-to-Consumer 서비스 제공 기업에서 참고하기에 적합하다. 기업간B2B, Business-to-Business 서비스를 제공하는 경우에는 관련 표준('23년 12월 제정 예정인 TTA 단체 표준에 근거하며, 해당 표준에 대한 정보는 제정 완료 후 업데이트 예정)을 참고하는 것이 더 활용도가 높을 것이다. 표준에 근거한 요구사항별 이해관계자는 부록A를 참고하기 바란다.

이외에도 개발 안내서를 활용하는 환경에 따라 그 적용 방안은 다양해질 수 있다. 예를 들어, 개발하는 인공지능 제품·서비스의 산업 분야마다 해당 분야의 전문가 역시 적극 협업할 필요가 있다. 그리고 만약 대표 이해관계자 및 협력 대상이 모두 존재하지 않는 소규모 기업에서는 외부 전문가의 도움을 받을 수도 있다. 따라서, 다음 페이지와 부록A, B에 제시된 대표 이해관계자, 협력 대상, 직무별 역할 등의 내용은 참고 자료로 활용하는 것이 좋다.

05 안내서 활용 방법

본 안내서는 범용성을 갖추고자 인공지능 신뢰성 관점에서 기술적 고려가 필요한 요구사항 및 검증항목을 포괄적으로 수록하였다. 따라서, 기업 내부의 기술 역량, 제품 특성 등을 고려하여 적절한 요구사항과 검증항목을 선택하여 적용하고, 기업에서 제공 중인 서비스의 환경에 맞게 신뢰성을 확보하는 참고 자료로 활용하길 바란다. 더불어, 인공지능 신뢰성을 확보하려면 기술적 측면 외에도 윤리, 개인 정보 보호와 같은 법·제도적 측면도 함께 요구된다. 그러므로 본 안내서를 활용하기에 앞서 인공지능 윤리적 고려 사항을 점검하는 「인공지능 윤리기준 실천을 위한 자율점검표」와 개인정보보호의 준수 여부를 점검하는 「인공지능 윤리기준을 실천하는 자율점검표」와 개인정보보호의 준수 여부를 점검하는 「인공지능(AI) 개인정보보호 자율점검표」 등을 선행해서 검토할 것을 권고한다.

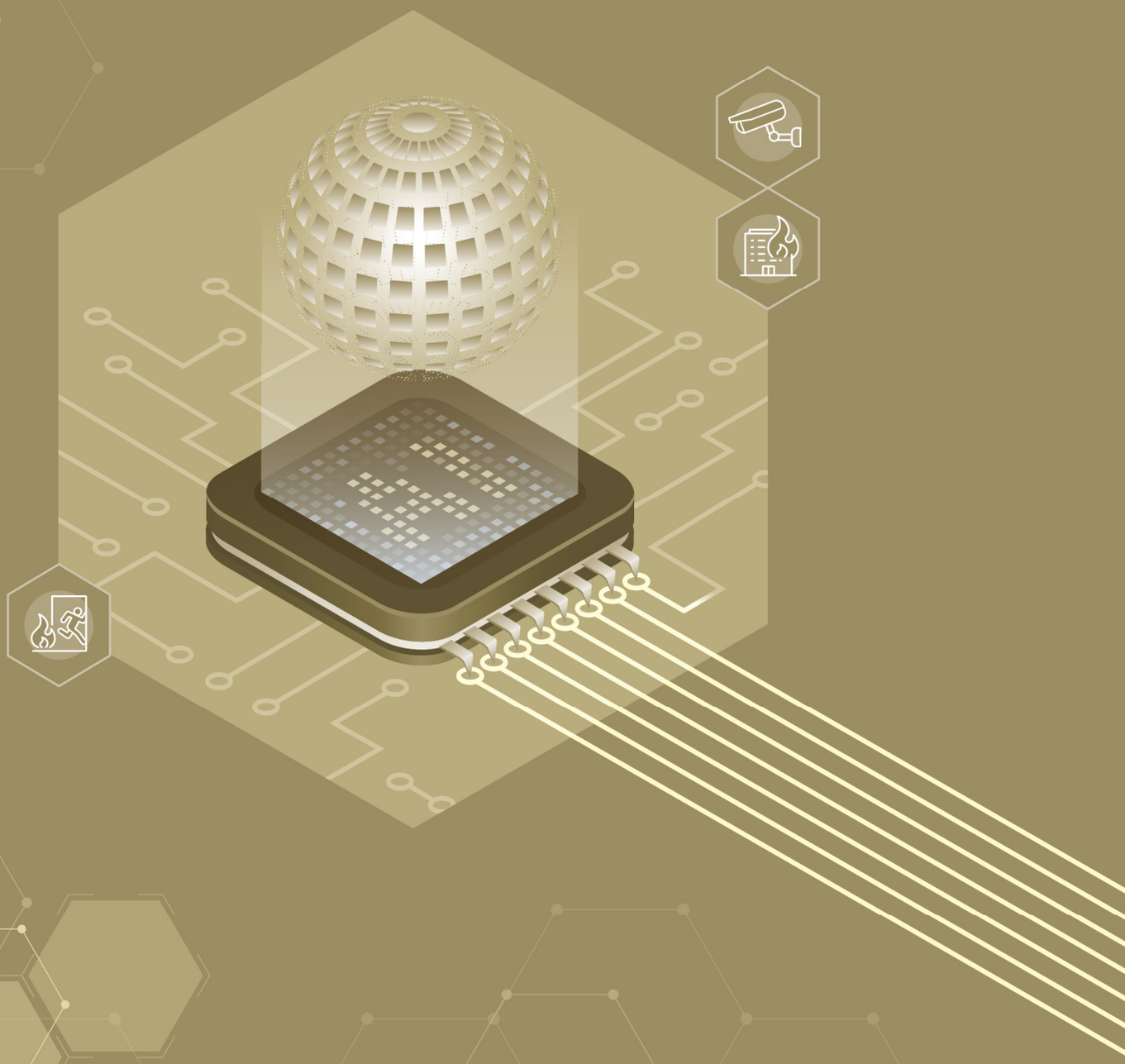
또한, 면밀한 검토가 필요한 것은 인공지능에 특화된 속성뿐만 아니라 시스템에 적용되는 기존의 속성도 마찬가지이다. 따라서 스마트 치안 시스템은 인공지능에 특화된 속성 외에도 성능 및 보안 등 전통적인 시스템 속성에 대한 검증을 거쳐야 한다. 이러한 다각적인 접근 방식을 통해 시스템이 실제 환경에서 효과적이고 안전하게 작동하는지 확인할 수 있다.

본 가이드는 다음과 같은 방법으로 효과적으로 활용할 수 있다.

① **위험 영향 분석**: 스마트 치안 시스템 도입에 따른 위험성을 평가하려면 도입 목적, 범위, 사고 위험성, 잠재적 사회적 결과 등을 필수로 분석해야 한다. 사소한 실수나 오류라도 안전이 보장되지 않으면 심각한 피해를 초래할 수 있다. 비즈니스 의사 결정자, 기획자, 개발자, 시스템 운영자 간 협업해 종합적인 영향 분석이 이루어져야 한다.

② **요구사항 선정**: ①의 결과를 바탕으로 개발 가이드에 설명된 요구사항과 세부 요구사항을 참조한다. 본 가이드는 스마트 치안 시스템의 신뢰성 확보에 필요한 요구사항을 선정하는 데 도움을 준다. 전문 이해관계자, 법률 고문, 개발자와 상의하여 가이드를 참고하는 것이 좋다. 요구사항이 불필요하다고 판단되는 경우 점검표에서 제외하여 'N/A[Not Applicable]'로 표시할 수 있다.

③ **자가 점검**: '②'에서 선정한 요구사항은 세부 요구사항 및 검증항목 본문을 참고하여 충족 여부를 점검한다. 이 과정에서 본 개발 안내서의 본문에 소개된 기술 및 기법 예시를 참고하여 요구사항을 충족하지 못할 경우 이를 해결할 만한 수단 또는 기술이 있는지 확인해 볼 것을 권고한다. 각 요구사항의 대표 행위자가 주도하여 협력 대상과 함께 검증항목의 충족 여부를 판단하는 데 필요한 절차서, 코드, 분석 자료 등의 관련 산출물을 확인하고, 테스트나 측정이 필요한 항목은 해당 활동을 수행한다. 검증항목에 따라 충족 여부를 정성적으로 평가할 수 있으나, 이는 '①'에서 분석한 서비스 영향 정도를 고려하여 대표 행위자와 협력 대상자가 협의하여 충족 여부를 판단할 수 있다.



PART 2

요구사항 및 검증항목

1. 생명주기 관리
2. 데이터 수집 및 처리
3. 인공지능 모델 개발
4. 시스템 구현
5. 운영 및 모니터링

목차

목차

목차

목차

01 생명주기 관리

책임성 투명성

요구사항 01 인공지능 시스템에 대한 위험 관리 계획 및 수행

- 스마트 치안 분야에서 인공지능 응용프로그램은 사용 목적의 중요성과 시스템의 추론 결과가 개인의 안전과 생명에 직접적인 영향을 미치므로 다른 분야보다 높은 확신과 정확성이 요구된다. 인공지능 응용프로그램은 독립형 시스템으로 작동하지만, 특히 기존 시스템과의 통합을 고려할 때 무결성과 연속성이 필요하다.
- 따라서 이러한 시스템이 사용자들에게 부정적인 영향을 미치지 않도록 그들의 안전을 보장하는 위험 관리가 필수적이다. 개인이 안전하려면 사전에 위험 요소를 확인하고, 각 위험의 심각도와 파급 효과를 분석하며, 그에 따른 대응책 마련이 필요하고 매우 중요하다.

01-1 인공지능 시스템 생명주기에 걸쳐 나타날 수 있는 위험 요소를 분석하였는가?

Yes ☐ No ☐ N/A ☐

- 위험 관리는 위험 인식identification, 위험 분석analysis, 위험 평가evaluation, 위험 대응treatment으로 구분한다. 이러한 네 가지 활동을 생명주기 단계별로 지속해서 반복하여 수행하여 신뢰성을 확보하고 위험을 제거 및 방지하여야 하고, 이에 대한 개념 및 정의는 ISO 31000:2018 – Risk Management에서 제공한다. 또한, 인공지능의 신뢰성 관점에서 살펴보아야 할 위험 요소를 인식, 분석 및 평가하는 방법론은 ISO/IEC 24028:2020[9]과 ISO/IEC 23894:2023 – Guidance on risk Management에서 제공한다[23].
- 이러한 표준 외에도 미국 국립표준기술연구소(NIST)에서 개발한 프레임워크도 있다. AI 위험 관리 프레임워크(AI RMF)로 개발자 및 사용자는 개발된 인공지능 시스템이 개인, 조직 및 사회에 미치는 위험을 관리한다[24].
- 또한 ISO 그룹은 로봇을 사용하는 ISO 13482:2014 표준도 발표하였다. 개발하는 시스템이 로봇 분야와 관련 있거나 교통 단속 및 제어하는 로봇 솔루션 시스템, 위험 조건에서 인간의 노동력을 대체하는 기술적 활용(예: 로봇 솔루션)이 포함될 때 이 표준의 섹션 4를 살펴보는 것도 고려해 볼 수 있다.
- 스마트 치안 시스템의 기술적·장비적 한계로 인한 불확실성, 물리적 한계 및 가변성, 오탐·미탐 등 탐지 관련 오류, 개인정보 유출, 외부 공격에 따른 보안 이슈 등으로 인한 문제도 고려하여야 한다.

01-1a 인공지능 시스템의 위험 요소를 도출하고 이의 파급효과를 파악하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 스마트 치안 시스템의 위험 요소는 소프트웨어 및 하드웨어 기반 시스템에서 발생하는 위험 요소와는 다르다. 소프트웨어/하드웨어의 결함이나 오류와 달리 스마트 치안 시스템의 추론 결과가 개인에게 직접적으로 영향을 미치므로 데이터 기반 분석의 특성으로 나타나는 편향성, 설명 가능성, 모델에 대한 공격 등의 위험 요소를 주의 깊게 도출하여야 한다. 이러한 요소의 분류와 주요 내용은 ISO/IEC 23894.2와 ISO/IEC 24028에 제시되어 있다. 개발된 스마트 치안 시스템이 공공 비상 경보 시스템과 관련될 때 ISO 22322, 로봇 시스템 사용과 관련될 때 ISO 13482:2014 표준 섹션 4에 제시된 잠재적 위험 요소 확보를 고려하여야 한다.
- 또한, 인공지능 알고리즘을 반영한 스마트 치안 시스템을 개발할 때 데이터 유출, 사이버 공격으로 인한 AI 의사 결정의 윤리적 영향 등 발생 가능한 모든 위험을 식별하고, 이로 인한 파급 효과(예: 인권 침해)를 파악하여야 한다.
- 개발된 시스템이 질문한 사용자를 자동으로 범죄자로 판단하거나 보행자가 폭력성 등을 보였을 때 사용자 승인 없이 자동으로 판단하는지 등 스마트 치안 시스템의 사용 정도
- CCTV 카메라, 센서 등 시스템 장비 분석
- 인공지능 시스템의 인지 성능 저하는 인권 침해(성별, 성별, 인종, 외모, 신체적 한계, 언어, 신념, 학력, 소속 단체 및 기타 편견과 관련된 인권 침해 등)로 이어진다. 이러한 성능 저하는 모델 추출 및 모델 회피 공격 등 다양한 모델 공격으로 인한 탐지 기능의 손실에 기인한다.
- 시스템 예상 사용자 분석
- 전기 및 전력 시스템(선로 전압, 누설 전류, 전기장 등), 열 에너지, 기계 에너지(중력, 진동 등)
- 위험 요소를 도출한 후에는 이를 야기하는 원인과 이에 따라 발생 가능한 결과를 분석하여야 한다. 발생 가능한 결과란 편향된 결과로 인해 차별이 발생하여 사회적으로 부정적인 영향을 미치는 현상 및 결과를 의미한다.
- 위험 요소의 발생에 따른 발생의 심각성, 빈도 등의 척도를 기준으로 위험의 크기 또는 수준을 평가한다. 이는 위험 요인의 파급 효과를 의미한다. 위험 요인 도출 후에는 다양한 환경이나 상황에 따른 관리 방안과 변동 효과를 분석하여야 하며, 분석 결과에 따라 인공지능 시스템의 수명 주기 동안 주기적인 추세 분석과 모니터링을 반복해서 수행하여야 한다.

참고 스마트 치안 시스템에서 가장 많이 관찰되는 인공지능(AI) 위험 요인들

위험 요소 예시[25]:

- 보안 취약점이 발생하기 쉬움: 개발된 시스템에 필요한 권한을 부여하면, 시스템 오작동 시, 시스템이 자체적으로 조명을 켜거나 문을 열거나 화재 경보 등을 자동으로 작동시킨다.
- 해커 공격의 취약함: 생년월일은 물론 신용카드 정보 등 개인정보를 많이 수집하는 모델 설계한 때, 개발된 시스템이 공격을 방어하기에 약한 때, 알 수 없는 공격자가 사용자의 보호 및 민감한 정보를 쉽게 도용한다. 또한, 극단적인 예로는 신원 도용으로 이어진다.
- 약한 암호로 인한 정보 유출: 개발자가 사용자 계정의 강력한 암호 생성을 지원하지 않는 모델 설계 시, 보안 취약점으로 이어지며, 도난당한 계정을 사용하여 보안 상실 등 극단적인 상황이 발생한다.
- 위치 추적: 개발자가 모델을 설계하여 사용자의 위치 정보를 기록하거나 스마트 디바이스를 통해 최종 사용자의 지리적 정보를 얻을 때, 스마트 디바이스 제공 업체에 의한 개인정보 유출로 이어진다.
- 식별되지 않은 보안 유출: CCTV 카메라, 센서 등 하드웨어가 포함된 시스템 개발 시, 이러한 장치 중 어느 하나에 보안 허점이 존재하므로 주의하여야 한다. 따라서 하이브리드 시스템을 다룰 때는 항상 장치의 허점과 취약점을 확인하여야 한다.
- 주택의 중요 기능을 제어하는 스마트 디바이스의 취약성: 이러한 장치가 공격자에 의해 공격당할 때, 주변의 화재 경보 등 보안 장비에 대한 액세스를 차단하는 방어 메커니즘을 설계하여야 한다.
- 데이터 조작: CCTV 영상이나 센서에서 수집된 데이터 등의 전송이 암호화되지 않은 상태로 서버로 이루어진다면, 공격자는 적대적인 공격 방법으로 해당 데이터에 쉽게 개입하고 조작한다. 따라서 개발자는 데이터 전송 과정의 트래픽을 확인하도록 고급 네트워크 모니터링 도구를 사용하여야 한다. 수집된 데이터의 보호 방법을 고려하지 않으면 데이터 유출, 신원 도용 또는 금융 사기 등의 문제가 발생한다. 또한 AI 모델, 특히 레이블이 지정된 데이터셋(지도 학습)를 사용하여 학습된 모델을 개발할 때는 학습에 사용된 데이터가 의도하지 않게 노출되거나 유출되는 것을 방지하도록 강력한 조치를 마련하는 것이 중요하다. 이러한 유출은 개인정보 보호, 보안 및 윤리적 문제 측면에서 심각한 결과를 초래하기 때문이다.

AI 모델에 대한 직접적인 공격: 개발된 모델의 오용은 위험 요소로 간주한다. 일부 예시는 다음과 같다:

- 개발된 모델이 인간의 감독 없이 자율 무기로 사용될 때
- 개인의 개인정보 보호나 시민의 자유에 대한 고려 없이 대규모 감시로 사용될 때
- 모델이 전이 학습을 공격하고자 개방될 때
- 모델이 딥 페이크 콘텐츠를 생성하고자 개방될 때
- 모델이 편향된 데이터로 훈련되거나 자동화 편향성에 노출될 때

01-1b 인공지능 기술 적용을 어렵게 만드는 위험 요소가 있는지 확인하였는가?

Yes ☐ No ☐ N/A ☐

- ISO/IEC 23894:2023에서는 위험 인식 단계에서 위험을 초래할 수 있는 위험 요소, 사건 또는 결과를 식별해야 한다고 말한다. 결과 식별은 조직, 개인, 커뮤니티, 집단, 사회에 대한 모든 결과를 대상으로 해야하며, 기술의 혜택을 경험하는 집단과 부정적인 결과를 경험하는 집단 간의 차이를 식별하는 데 특별한 주의를 기울여야 한다. 식별해야 할 결과의 예시는 다음과 같다.
  - ✓ 기회의 획득 또는 상실
  - ✓ 개인의 건강이나 안전에 대한 위협
  - ✓ 피해 복구를 위한 특정 기술에 대한 재정적 비용
- 만약 인공지능 기술이 극단적으로 부정적인 결과를 초래할 수 있다고 확인된 경우, 인공지능 기술 적용에 대해 재검토하여야 한다. UNESCO의 〈Recommendation on the Ethics of Artificial Intelligence〉와 같은 일부 문헌에서는 인공지능 기술을 적용하지 않아야 하는 특정 분야를 명시하고 있다.

참고 UNESCO, EU에서 언급한 인공지능 기술이 적용되지 말아야 할 분야의 예시

- Recommendation on the Ethics of Artificial Intelligence(UNESCO): Proportionality and Do No Harm
  - 인공지능 시스템은 소셜 스코어링social scoring이나 대규모 감시mass surveillance 목적으로 사용되어서는 안 된다.
- Artificial Intelligence Act(EU): Unacceptable risk
  - 허용할 수 없는 위험을 갖는 인공지능 시스템은 인간에게 위협이 되는 것으로 간주되어 금지되어야 할 시스템이다. 여기에는 다음이 포함된다:
    - · 사람이나 특정 취약 집단에 대한 인지 행동 조작(예: 어린이의 위험한 행동을 조장하는 음성 인식 장난감)
    - · 소셜 스코어링social scoring
    - · 안면 인식 등 실시간 원격 생체 인식 시스템

01-2 위험 요소를 제거 및 방지하거나 영향을 완화하기 위한 방안을 마련하였는가?

Yes ☐ No ☐ N/A ☐

- 01-1에서 분석된 각 위험 요소별로 대응 방안을 마련하여야 한다. 해당 대응 방안은 위험 요소의 원인을 제거하거나 잘못 계산된 결정의 파급 효과와 잠재적인 부정적 영향을 최소화함으로써 인권 침해와 개인의 피해를 방지하는 조치를 포함한다. 치안 시스템의 성격상 개인의 생명에 중요한 역할을 하므로 이러한 대응책을 신중히 고려하여야 한다.
- 스마트 치안 시스템은 편향으로 인해 인권을 침해하는 큰 잠재력이 있다. 따라서 대응방안 실행 및 운영 절차, 소프트웨어 및 하드웨어 기능, 모델 학습 기술 및 전략 등 기술적으로 적용하는 모든 방법을 포함한다.

- 이를 위해, ISO/IEC 24028:2020은 대응책의 분류를 제공한다. 또한, ISO/IEC TR 24030 '정보기술 – 인공지능 (AI) – Use cases' 표준은 치안 시스템에 대한 부분으로 7.18장에 해당하는 Use cases를 측정에 사용하고자 이해관계자들이 검토한다.
- 스마트 치안 시스템의 모든 이해관계자는 이를 고려하여야 하며, 위험 요인에 대한 대응 방안을 마련하고 위험이 제거 및 완화되었는지 확인하여야 한다.

01-2a 위험 요소별 완화 또는 제거 방안을 마련하였는가?

Yes ☐ No ☐ N/A ☐

- 위험 요소를 드러내거나 위험 요소를 식별하는 데 도움이 되는 적용 방법, 소프트웨어 및 하드웨어 기능, 모델 학습 기법 및 전략 등 기술적 방법론을 도출하거나 구현하여야 한다. 이러한 방법론에 대한 분류와 개략적인 내용은 ISO/IEC 24028 표준에 제시되어 있다.
- 01-1 에서 언급된 위험 요인에 대해서는 파급 효과가 가장 큰 위험 요소를 우선순위에 따라 대응 조치를 적용하여야 한다. 또한, 편향된 데이터 개입, 인공지능 시스템의 판단 결과에 인간 개입 등 위험 완화 조치를 고려하여야 한다. 이러한 시스템은 AI 모델과 함께 사용되는 시스템과 엣지 장치의 특성으로 인해 편향성이 높이므로 개발자들은 파급 효과를 식별할 때 주의하여야 한다.
- 인공지능의 규정 준수 및 평판적 위험은 전통적인 위험 관리 기능과는 다르다. 스마트 치안 시스템 수명주기의 각 구성 요소에 대한 가능한 위험 요인에 대한 대응 기술적 방법은 인공지능 시스템의 신뢰성 향상과 편향된 추론 결과의 발생 가능성을 감소하고자 신중히 분석되어야 한다. 대응이 적용된 후에는 위험이 실제로 제거, 방지 또는 완화되었는지 확인하고자 그 파급 효과를 재평가하여야 한다.

01-2b 위험 요소의 파급효과가 감소하였는지 확인하였는가?

Yes ☐ No ☐ N/A ☐

- 위험 요소를 발생시킬 수 있는 구현 및 운영 방식, 소프트웨어 및 하드웨어 기능, 모델 학습 기법 및 전략 등의 기술적인 방법론을 도출하여야 한다. 이러한 방법론에 대한 분류와 개략적인 내용은 ISO/IEC 24028:2020에 제시되어 있다.
- 앞서 위험 요소를 분석하는 과정에서 위험 요소의 파급효과를 평가하였는데, 파급효과가 가장 큰 위험 요소를 우선순위로 대응 방안을 적용해야 하며, 위험의 파급효과가 큰 경우 인공지능 시스템의 판단 결과에 대한 사람의 개입을 고려하는 등의 위험 완화 방안을 적용해야 한다.
- 대응 방안이 적용된 이후에는 파급효과를 재평가함으로써 위험 요소가 실제로 제거, 방지 혹은 이의 영향이 완화되었는지 확인하여야 한다.

안전성 | 다양성 존중 | 책임성 | 투명성

요구사항 02 인공지능 거버넌스governance 체계 구성

- 인공지능 기술을 치안 시스템에 사용하는 것은, 이러한 시스템이 개인의 생명, 안전 및 생활 안전에 직접적인 영향을 미쳐 개인 및 최종 사용자에게 심각한 피해를 초래할 위험에 대한 책임이 요구된다.
- 스마트 치안 시스템의 효과와 결과를 평가하고 준비하고자 조직 등은 인공지능 기술의 사용과 관련된 법률, 규정, 정책, 표준 및 지침을 마련한다. 신뢰성을 확보하고자 인공지능 시스템에 대한 신뢰를 확보하는 데 관련된 규정, 정책, 지침 등을 준수하는 것은 모델의 신뢰성을 확보하는 중요한 단계이다. 따라서, 내부적으로 시행할 규정, 정책 및 관련 표준/지침을 취득하고, 개발된 스마트 치안 시스템의 인공지능 거버넌스를 감독하고자 이러한 규정/지침 등을 관리하는 것이 필요하다.

02-1 인공지능 거버넌스에 대한 지침 및 규정을 수립하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능과 관련된 조직에서는 인공지능 시스템의 신뢰성을 확보하고자 거버넌스 체계를 구성할 필요가 있다. 인공지능 시스템은 학습이나 추론 과정에서 윤리 및 지식재산권IP, Intellectual Property 관련 문제나, 보안 및 개인정보 이슈가 발생하기 때문이다. 이러한 위험 요소에 대비하고자 내부적으로 인공지능 거버넌스에 대한 지침 및 규정을 수립하여야 한다[26].
- NIST의 AI RMFRisk Management Framework에서는 인공지능 시스템의 생명주기에 따라 내부 규정, 절차, 과정 및 실제 행위가 투명하고 효율적으로 이루어져야 한다고 언급한다. 즉, 인공지능과 관련된 법, 규제 관련 요구사항이 이해·관리되어 문서화하고, 위험 관리 절차와 산출물이 체계를 통해 투명하게 관리되어야 한다.
- 내부적으로 수립해야 할 규정은 활용 측면에 따라 크게 두 가지로 구분하여 마련한다.
  - ✓ 첫째, 인공지능 관련 법, 규제, 정책, 표준 및 지침을 채택·정리하여 내부적으로 이행해야 할 지침 및 규정을 수립하여야 한다.
  - ✓ 둘째, 인공지능 시스템 생명주기에 따른 조직의 역할과 책임을 명확하게 문서화하여야 한다.

02-1a 내부적으로 준수해야 할 인공지능 거버넌스에 대한 지침 및 규정을 마련하였는가?

Yes ☐ No ☐ N/A ☐

- 보통 전 세계적으로 인공지능 시스템을 규율하는 고전적인 접근 방식은 공법이다. 그러나 스마트 치안 체계에서 인공지능 거버넌스 체계의 기본 단계는 거버넌스 기본 원칙을 수립하는 것으로, 인공지능과 관련된 법, 규정, 정책 등을 이해한 후 내부적으로 윤리적 관점에서 이행해야 할 규정을 정의하여야 한다. 즉, 인공지능과 관련된 리스크를 인지하고 대비하려면 관련 표준 및 가이드라인을 도입하고 조직 내부 규정을 마련하여야 한다.
- 개발된 치안 시스템 체계에 대한 신뢰성을 확보하려면 인공지능 거버넌스 및 거버넌스 조직 전반의 업무, 역할, 의무, 책임 등을 명확히 하는 것이 필수적이다. 또한, 개발된 치안 시스템의 라이프 사이클 전반에 걸쳐 도출된 추론 결과를 관리하고자 가이드라인을 마련하고 효율적으로 관리하여야 한다.
- 치안 시스템의 신뢰성을 확보하는 윤리적·법적 의무, 개인정보 보호 관련 법률, 치안 시스템 모델 개발과 관련된 국제 및 국내 법률을 준수하고 거버넌스를 구축하여야 한다.
- 국제연합(UN), 세계경제포럼(WEF), 세계보건기구(WHO) 등의 기관은 인공지능에 대한 신뢰 구축이 목적으로 이러한 프레임워크 협약을 제정하는 권한을 부여받았다. 모델 인공지능 거버넌스 프레임워크는 WEF의 일부 프로젝트로 시작되었다[27].
- 이 모델 프레임워크는 내부 거버넌스 구조와 조치로 구성되며, 인공지능 증강 의사 결정, 운영 관리, 이해관계자 상호 작용 및 커뮤니케이션에 대한 인간의 참여 수준을 결정한다[28]. 이 프레임워크를 통해 다음을 할 수 있다.
- 조직이 구현된 관행을 평가하고 측정하도록 지원
- 조직의 구현 경험 및 사용 사례 향상
- 또한, 개인을 보호하고자 250개 이상의 국가가 세계인권선언(UDHR)으로 비준한 EU 원칙을 기반으로 개발된 피해 모델과 같은 모델을 프레임워크로 활용한다.

참고 2022년 개인정보 보호법 표준 해석 예시 사례[29]

특히 집행 권한이 없는 개발자/기관이 CCTV를 사용하여야 하는 감시 시스템을 구축하고자 할 때는, 이러한 데이터 설치, 활용 및 저장 사례를 확인하여 개인/공공 감시의 법률 제한 및 규제에 대한 인사이트를 얻는다.

단속 권한이 없는 개인 또는 공공 기관이 「개인정보 보호법」 제25조 제1항 제4호에 따른 '교통 정리를 학좌 필요할 때'에 CCTV를 설치·운영하는가: 이 조항은 일반적으로 불법 주정차, 신호 위반, 과속 등 교통법규 위반을 감시하는 목적과 관련 있다. 공공장소에서의 영상 감시에 대한 개인정보 보호법의 엄격한 개인정보 보호 제약을 고려할 때, 동법 제25조 제1항 제4호에 따른 설치 및 운영은 교통 감시 권한을 가진 기관에 한정된다고 해석하는 것이 타당하다(개인정보보호위원회 결정 제2020-03-045호 참조). 개인정보 보호법상 '개인정보처리자'로 분류되는 개인 또는 공공기관이 법 제2조 제7호에 따른 영상 정보 처리 기기를 설치·운영하려고 할 때, 별도의 표지에 목적을 명시하더라도 지방 자치 단체 또는 경찰의 교통 감시 권한 없이는 제25조 제1항 제4호에 따른 '필요한 교통 단속'을 하도록 영상 정보 처리 기기를 설치·운영할 수 없다.

기존에 방범용으로 사용하던 카메라에 교통 데이터 수집 기능을 통합하거나 불법 주정차 감시 등 다른 목적을 추가하여 기존 CCTV 시스템에 통합하는가: 개인정보 보호법 제25조 제1항은 공공장소에 영상 정보처리 기기를 설치·운영을 원칙적으로 금지하나, 제5호에서 교통 정보의 수집·분석 및 제공을 허용하는 예외 규정을 한다. 영상 정보 처리 기기를 설치-운영하려는 공공기관은 같은 법 제25조 제3항에서 대통령령으로 정하는 바에 따라 공청회, 전문가 의견 수렴 등 구체적인 절차를 거쳐야 한다. 동법 시행령 제23조 제1항에서는 이러한 장치를 설치하려는 공공기관의 장이 행정 예고, 전문가 자문 등의 절차를 규정한다. 또한, '개인정보 보호 표준 가이드라인' 제38조는 개인정보 보호법 시행령 제23조 제1항에 따라 영상 정보 처리 기기의 설치 목적을 변경할 때 특히 전문가 및 이해관계자의 의견을 구하는 것이 중요함을 강조한다. 지자체가 이러한 절차를 준수할 때 기존 CCTV 시스템을 방범과 교통 데이터 분석에 동시에 활용한다.

비공개 또는 출입이 제한된 장소에 CCTV를 설치할 때도 개인정보 보호법이 적용되는가: 개인정보 보호법 제25조 제1항은 예외적인 때를 제외하고는 공개된 장소의 영상 정보 처리 기기 설치 및 운영을 제한한다. 공공장소는 도로, 공원, 광장, 지하철역 등 접근에 제한이 없는 장소를 말한다. 출입이 통제되거나 특정 출입 요건이 있는 장소는 '공공장소'에 해당하지 않는다. 이러한 '공개된 장소가 아닌 곳'에 설치된 영상 정보 처리 기기는 개인정보 보호법 제25조의 적용을 받지 않는다. 다만, 사업자가 영상 정보와 관련된 개인정보 파일을 관리하는 등 사업 목적으로 해당 기기를 사용할 때는 '개인정보처리자'가 되어 법의 일반 원칙을 적용받는다. 표지판 설치 의무 등 공개된 장소에 있는 기기에 적용되는 의무가 공개된 장소가 아닌 곳에 있는 기기에까지 확대 적용되지는 않지만, 정보 주체가 직접 촬영되는 점을 고려하여 영상 정보 처리 기기를 보호하고자 개인정보 보호법에 준하는 조치를 하는 것이 바람직하다 (06-2b, 'Goalie II' 사례 참조).

분실물 회수 등 개인이 요청할 때 CCTV 영상에 대한 접근 권한을 부여해야 하는 법적 의무는 무엇인가: 개인정보 보호법 제35조는 '개인정보 열람'을 규정하며, 정보 주체는 자신이 처리하는 개인정보에 대한 열람을 요청한다. 열람 요구, 제한, 통지 방법 및 절차에 관한 구체적인 사항은 개인정보 보호법 시행령(제41조, 제42조, 제46조)에 명시한다. 정보 주체가 자신의 개인정보에 대한 열람을 요구할 때 개인정보처리자는 법 제35조 제4항에 따라 열람을 거절하거나 제한하는지를 판단한 후 그에 따라 처리하여야 한다. 또한 개인정보처리자는 제46조에 따라 요청자가 정보 주체 또는 정당한 대리인인지를 확인하여야 한다. 이러한 액세스 조항은 데이터 주체의 개인정보와 관련 있다는 점에 유의하는 것이 중요하다. 요청자의 데이터 이외의 개인정보에 대한 접근 권한을 부여하는 것은 타인의 프라이버시를 침해하므로 의무 사항이 아니다. '개인정보 보호 표준 가이드라인' 제46조에 명시된 것처럼 관련 없는 개인의 개인정보를 보호하는 조치가 권장된다.

02-2 인공지능 거버넌스를 위한 조직을 구성하고 인력 구성에 대해 검토하였는가?

Yes ☐ No ☐ N/A ☐

- 다양한 위험 요소를 인지하고 관련 규정을 마련하며, 이를 이행하도록 관리·감독하는 조직이 필요하다. 또한 예측하지 못한 편향성으로 인해 발생하는 피해에서 최종 사용자를 보호한다.
- 개인의 안전, 개인정보 보호 및 책임성, 추론 결과의 신뢰성, 가이드라인 준수 및 절차적 요건을 충족하는 규정 수립 등을 포함한 인공지능 거버넌스를 감독하여야 한다. 또한 이러한 조직은 각 담당자의 역할과 책임을 충분히 숙지하고 관련 역량을 갖춘 인력을 보유하여야 한다.
- 가능하면 인공지능 거버넌스 조직은 사용 사례에 대한 민첩하고 혁신적인 접근 방식을 유지하여야 하며, 개발된 시스템에 대한 거버넌스에 따라 조직 구성원을 교육하도록 지속해서 관리하여야 한다. 개발된 시스템에 거버넌스 프레임워크 솔루션을 도입할 때는, 조직 구성원에게 해당 프레임워크 교육도 제공하여야 한다.

02-2a 인공지능 거버넌스를 위한 조직을 구성하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 거버넌스 시스템을 구축하고자 조직을 형성하고 그들을 관리하는 것은 개발된 스마트 치안 시스템의 거버넌스를 유지하는 데 중요하다. 이에 내부 규정을 수립하고 이를 준수하도록 하여야 한다.
- 전담 조직을 통해 인공지능 거버넌스를 담당하는 책임자가 명확해지며, 스마트 치안 시스템에서 인공지능의 윤리적이고 책임 있는 사용을 보장하는 데 도움이 된다.
- 이 조직은 직접적인 이해관계자 외에도 인공지능의 미래에 관심 있는 정부 기관, 기업, 시민 사회 단체, 개인 등을 포함한다.

참고 스마트 치안 시스템의 인공지능 거버넌스 형성 조직의 역할

이 조직과 조직 구성원의 주요 기여와 책임은 개발된 AI 시스템의 거버넌스를 유지하는 것이다. 거버넌스 프로세스는 신뢰하고 신뢰성 있는 AI의 기본 요소이다. AI 시스템의 거버넌스란 사용자가 신뢰와 윤리적 원칙에 기반한 모델을 개발하도록 전체 AI 프로세스에서 '책임성'을 정의하고 확립하는 과정을 말한다. 따라서, 이 형성된 조직은 치안 시스템의 의사 결정 프로세스가 설명 가능하고 투명하며 공정하도록 보장하여야 한다. 따라서 조직은 신뢰하는 개발된 치안 시스템을 보장하고자 거버넌스 솔루션을 확보하거나 개발하여야 한다. 이에 조직은 또한 사용자에게 거버넌스에 대한 지침을 제시하는 거버넌스 모델 프레임워크를 확보하여야 한다[33]:

내부 거버넌스 구조 및 조치: IBM이 개발한 AI 거버넌스[30] 등 기존 또는 확립된 내부 거버넌스 구조를 채택하고 치안 시스템의 잠재적 사용자/대상과 관련된 가치, 위험 및 책임을 통합하는 조치를 한다. 또는 Microsoft의 AI 거버넌스 프레임워크의 구조를 채택한다[31]:

Microsoft에서 구성된 인공지능 거버넌스 시스템은 세 가지 구분된 역할을 요구한다.

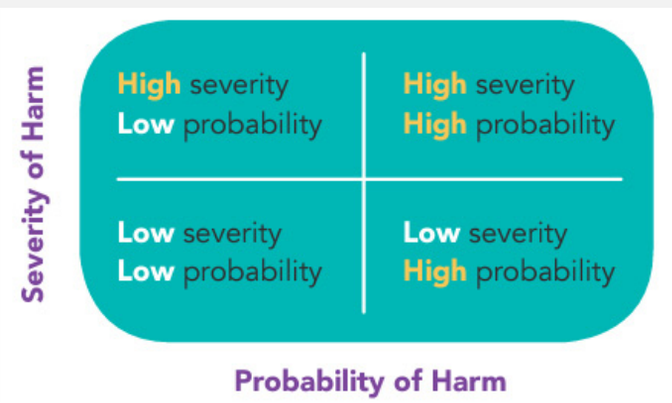
첫째, AI 윤리와 신뢰성에 대한 전문 지식을 기반으로 관련 규정을 수립하는 역할을 하여야 한다. 윤리에 관련된 법적 요구사항을 확인하고 지식재산권을 고려하여 위험에 대한 조치 방법을 규정하여야 한다.

둘째, AI 시스템 프로세스를 책임지고 관리하며 감독하는 역할이 필요하다. 담당자는 관련 규정에 따라 AI 시스템 프로세스가 올바르게 수행되는지 관리하고 감독하며, 회사 전체를 이끌고 책임져야 한다.

셋째, 규정의 구체적인 시행을 실제로 지원하는 역할이 필요하다. AI 거버넌스 시스템과 관련된 상세 규정이 정의되면, 각 부서와 분야에 따라 시행되도록 지원하여야 한다.

스마트 치안 시스템에 인간의 개입 수준 결정하기: 엔드 유저(조직/개인)에게 치안 시스템 사용 사례에 대한 위험 요소를 결정하도록 안내하는 방법을 수행한다. 스마트 치안 시스템에는 두 가지 가능한 개입 시나리오가 적용되어야 한다. 'Human-in-the-loop(결정 단계에서 완전한 개입을 제안하고, 엔드 유저가 시스템의 평가 및 추론 결과를 검토하고 관련 작업에 대한 최종 결정을 내릴 때)' 또는 'Human-over-the-loop(결정 단계에서 일부 개입을 제안하고, 엔드 유저가 시스템의 평가 및 추론 결과를 모니터링하거나 감독할 수 있으며 시스템의 결정에 개입할 수 있을 때)'이다. 어느 경우에도 시스템이나 개입된 사용자가 편향을 일으키면 개인들에게 대한 결과가 심각해질 수 있다.

의사 결정 과정에서 사람이 개입하는 정도* [33]



***'피해 가능성'은 대부분 개발된 보안 모델의 효율성과 정확성에 의존한다.**

운영 관리: 전체 치안 시스템 모델과 데이터 관리 프로세스를 개발하고 유지하는 동안 이 단계를 고려하여야 한다.

이해관계자 상호 작용 및 의사소통: 스마트 치안 시스템의 이해관계자와 솔루션 및 전략을 설계하고 관리 방법을 고려하여야 한다. 이에 효과적인 피드백 채널, 온라인/오프라인 상호 작용 및 '도움말 데스크' 솔루션 등을 확보한다.

이 외에도 ALTAI 조직은 AI 윤리와 관련된 문제에 대비하고자 AI 거버넌스 시스템 구축을 고려할 것을 권장한다.

02-2b 인공지능 거버넌스를 위한 조직은 전문성을 갖춘 인력으로 구성하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 거버넌스 담당 조직은 자신이 맡은 역할과 책임에 대해 충분히 인식한 인력으로 구성하여야 한다. 이들은 인공지능 생명주기에 걸친 모든 프로세스의 중심적인 역할로서, 담당자가 이를 충분히 인식한 후 책임지고 관리하여야 인공지능 시스템의 신뢰성을 확보하기 때문이다.
- 인공지능 거버넌스 담당 조직은 각기 다른 배경과 전문지식을 기반으로 충분히 숙련된 인력으로 구성해야 한다. 특히, 규정을 마련하는 역할을 맡은 담당자는 인공지능 윤리 및 신뢰성 분야의 원칙, 가이드라인, 표준 등에 대한 폭넓은 전문 지식이 필요하며, 이를 적절히 해석하여 조직 업무에 적용하는 기술력과 타 업무 담당자와 의사소통 역량이 필요하다. 또한, 정의된 규정을 실행하고 관리하고자 각 담당자에게 관련 교육을 제공하여 충분히 훈련하여야 한다.

02-3 인공지능 거버넌스 체계가 올바로 이행되고 있는지 감독하고 있는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 거버넌스 시스템을 운영하는 주체는 운영 결과에 대해 책임져야 하며, 이 책임은 위임할 수 없다. 따라서, 인공지능 거버넌스 운영 담당자는 조직이 내부 지침과 규정을 준수하는지 감독하여야 한다.
- ISO/IEC 38507:2022 – Governance implications of the use of artificial intelligence by Organizations에서 인공지능 거버넌스 체계는 인공지능 시스템에서 발생하는 위험에 따라 인공지능 시스템의 설계 및 사용에 대한 감독을 수행하여야 한다고 언급한다. 즉, 인공지능 거버넌스 체계를 통해 수립한 내부 규정을 조직이 적절히 이행하는지 감독하여야 한다.

02-3a 인공지능 거버넌스에 대한 내부 지침 및 규정 준수 여부를 감독하고 있는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 거버넌스 담당자는 인공지능 시스템 생명주기에 따라 조직이 내부 규정을 준수함을 확인 및 감독하여야 한다. 관련 이해관계자에게 신뢰성 있는 인공지능 시스템을 목표로 적절히 관리 및 통제됨을 입증하여야 한다.
- 특히, 인공지능 시스템의 위험 관리와 관련된 내부 규정을 이행하는지 감독함으로써 인공지능 시스템의 잠재적 위험에서 조직 및 이해관계자를 보호하고 조직의 역량이 향상된다.
- 따라서 인공지능 거버넌스 체계에서 감독을 담당하는 조직은 인공지능 시스템에 대한 이해를 바탕으로 역할에 대한 책임 및 권한을 명확히 인식하여 인공지능 시스템 생명주기에 걸쳐 모든 규정이 이행되는지 감독하여야 한다.

02-4 인공지능 거버넌스 조직이 신규 및 기존 시스템의 차이점을 분석하였는가?

Yes ☐ No ☐ N/A ☐

- 신규로 기획하는 스마트 치안 시스템은 사용 대상 및 역할 측면에서 기존 시스템과 무엇이 다른지 확인하고, 위험 항목을 점검한 결과를 바탕으로 시스템을 기획 및 설계하는 것이 필요하다. 이를 위해 보안 및 책임 프레임워크를 활용하여 개발한 스마트 치안 시스템이 안전하고 신뢰하며, 시중에 출시된 다른 시스템과 경쟁하는지 확인하여야 한다.
- 또한, 이전에 공개된 치안 시스템 리콜 사례를 비교 분석함으로써 잠재적인 위험을 분석하고 해결하여 신뢰성 있는 치안 시스템 모델을 개발하여야 한다. 스마트 치안 분야는 직접 현장 경험을 하여 AI 성능 평가가 가능한 점을 고려해 기존 시스템을 AI로 전환하거나 새로운 시스템을 시장에 도입하는 과정에서 발생하는 다양한 시나리오를 분석하여야 한다.

02-4a 기존 동일 목적의 시스템과 비교하여, 신규 시스템이 개선할 수 있는 사항을 분석하였는가?

Yes ☐ No ☐ N/A ☐

- 시장에서 이미 사용 중인 인공지능 시스템들이 존재하므로, 새로운 스마트 치안 시스템을 출시하기 전에, 개발된 시스템의 감지된 위험 수준과 기존 치안 시스템과 인지된 위험 수준을 비교하여 데이터 요구사항과 제어를 분석하여야 한다. 스마트 치안 시스템의 위험 수준은 주변 개인과 인간의 안전을 보장하고자 사용된다.
- 시스템의 복잡성이 증가함에 따라 산업 보안 법규에 더 구체적으로 다루어지기도 한다. 예를 들어, 스마트 홈 보안 생태계의 일부로서 응용되는 엣지 디바이스 사용 시, 개발자는 이러한 디바이스가 전체 개발된 시스템의 보안에 영향을 미칠 것을 예상하여야 한다. 새로 개발되는 치안 시스템은 시중에 나와 있는 기존 시스템과 동일한 수준의 안정성, 안전성, 유효성을 입증함과 동시에 객관적인 기준과 근거, 검증을 통해 안전성을 확보하여야 한다. 이러한 방향으로 비교 분석을 추구하여야 하며, 다양성 있는 개인들을 고려하여 시스템 신뢰성을 항상 확인하고 개발된 시스템에 대한 신뢰도를 관찰하고자 예측 가능한/예측 불가능한 편향을 완화하여야 한다.

참고 자동 출입국 심사 시스템의 정의 예시

비교 분석을 추진하고자, 개발된 시스템이 기본적인 요구사항을 충족하는지 확인하여야 한다. 이를 위해, 개발한 시스템 사용자들에게 제공할 서비스의 한계, 정의, 옵션 및 프레임워크를 명확하게 정의하여야 한다. 예를 들어, 국경에서 사용할 거짓 탐지 시스템 개발 시, 주어진 기본 기준을 준수하여야 한다. 예를 들어, 유럽 위원회(2021b)의 자동화된 국경 통제 시스템(ABC 시스템)의 정의는 다음과 같다:

"전자 출입 게이트 하드웨어, 문서 스캐닝 및 확인, 얼굴 인식 및 기타 생체 인증 등을 포함하여 다양한 인공지능 도구를 통합하여 국경 통행자의 신속한 처리를 용이하게 하고 보안을 강화하는 자동화된 이민 통제 시스템."

참고 보안 및 책임 프레임워크[32]

시장에 있는 다른 응용 프로그램/시스템들과 경쟁하려면, 개발한 스마트 치안 시스템이 안전하고 신뢰성 있게 사용되며 다른 시스템과 경쟁하는지 확인하여야 한다. 이를 위해 보안과 책임 프레임워크를 활용한다.

보안 및 책임 프레임워크의 전반적인 목적은 통합된 신생 기술을 포함한 모든 제품 및 서비스가 안전하고 신뢰성 있게 운영되며, 만약 피해 초래 시 효과적으로 그 피해를 해결하려고 함이다. 유럽 위원회의 정의를 고려할 때, 이 프레임워크를 사용하여 새로 개발된 솔루션/치안 시스템의 신뢰성을 유지하고 시장에 출시한다.

이러한 종류의 안전 및 책임 프레임워크는 개발된 치안 시스템뿐만 아니라 소비자/최종 사용자/대상 개인 등을 보호한다. 또한 시장에 존재하는 기존 시스템에 대해 새로 개발된 치안 시스템의 신뢰성을 제공한다.

유럽 연합(EU)의 안전 프레임워크: 시장에 있는 제품 사용 및 오용에 관여한다[33].

책임 프레임워크: 어떤 종류의 새로운 도전이나 문제가 다루어지는지 파악한다. 대부분 제품에 대한 이해를 파악하는 것이 목적이다[44].

안전성 투명성

요구사항 03 인공지능 시스템의 신뢰성 테스트 계획 수립

- 전통적인 소프트웨어와 달리, 인공지능은 추론 결과에 대한 불확실성uncertainty을 내포한다. 이러한 인공지능의 불확실성을 줄이는 것은 안전성과 같은 신뢰성 확보에 중요한 요소이다. 따라서 소프트웨어의 품질 확인을 위한 테스트 외에도 인공지능 시스템의 신뢰성 확인을 위한 테스트가 추가 요구된다. 테스트를 위해서는 인공지능 시스템의 복잡도complexity와 운영환경을 고려한 계획 수립이 필요하며, 계획에 따라 생명주기 전 단계에서 정기적·지속적 테스트를 수행한다.

  * 인공지능에 해당하는 속성뿐만 아니라 기존 소프트웨어 시스템에 적용되는 전통적 속성도 적용되었는지 확인이 필요하다. 따라서, 본 요구사항에 기술된 내용 외에도 시스템 성능, 보안 등 품질 관점의 검증 절차도 반드시 병행되어야 할 것이다.

03-1 인공지능 시스템의 특성을 고려한 테스트 환경을 설계하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템은 그 복잡도complexity나 위험도에 따라 가상 테스트 및 실환경 테스트를 모두 고려하여야 한다.
- 실제 테스트는 정확성을 제공하지만 시간과 비용 측면에서 타당성을 고려하여야 한다. 복잡한 작동 조건은 실제 테스트에 적합하지 않을 수 있다. 또한 인공지능 시스템이 개인과 물리적으로 상호작용하면 개인 인권의 위험을 초래한다. 국경 통제, 위협 검사, 범죄 예방, 감시, 교통 통제, 위험 통제 등을 하는 로봇 솔루션처럼 시나리오는 실제 테스트에 앞서 가상 시뮬레이션 테스트 환경을 구축하는 것이 좋다.
- 따라서 개발된 치안 시스템의 특성을 고려하여 적절한 테스트 환경을 결정하고 편향되지 않은 테스트 환경을 설계하는 것이 필요하다. 또한 테스트 환경에 필요한 주의 사항과 평가 항목은 다음과 같다.

테스트 환경 보안에 필요한 체크리스트

평가 항목	표머리
운영 환경	인공지능 시스템의 운영 환경이 복잡하고 다양한 이해관계자가 관련되는가?
개인정보 보호	인권 및 개인의 민감 정보 보호에 대한 잠재적 위협이 우려되는 시스템인가(국내는 제18조 및 제62조 참조)?
운영 비용	합리적인 시간과 비용으로 테스트를 수행하는가?
검증	임상 시험 전 검증하고자 특정 테스트가 필요한 항목은 무엇인가?
가상 환경	가상 환경이 실제 환경을 모방하는가, 시뮬레이션이 실제 환경을 제대로 반영하는가?
엣지 디바이스 점검	테스트 단계가 시작되기 전에 장비, 엣지 디바이스의 조정 및 보정이 잘 이루어졌는지 확인하는가?

03-1a 테스트 환경 결정 시 인공지능 시스템의 운영환경을 고려하였는가?

Yes ☐ No ☐ N/A ☐

- 스마트 치안 시스템은 개인의 생명과 안전에 직결되는 만큼 개발된 시스템의 인공지능 의사 결정에 따른 신뢰성 확보는 기본이자 필수 단계이다. 또한, 테스트 환경이 온프레미스on-premises나 가상 기반 등 실제 운영환경과 일치하는지에 대한 올바른 판단도 중요하다. 비디오 서비스는 트래픽이 집중되고 분석 프로세스에 대한 요구가 많아 시스템 과부하가 발생하므로 실제 환경을 모방한 테스트 환경을 만드는 것이 매우 중요하다. 또한 개발된 시스템의 안정성을 보장하고자 안정적인 인프라를 구축하는 가이드라인이 필요하다. 이는 실제 시나리오에서 인공지능 모델의 성능을 평가하는 데 도움이 된다.
- 인공지능에 대한 유네스코의 결정에 따르면 스마트 치안 시스템처럼 잠재적으로 인권을 위협한다고 확인된 인공지능 시스템은 출시 전에 이해관계자의 윤리적 영향을 평가하는 일환으로 광범위한 테스트(필요하면 모의 테스트도 포함)를 거쳐야 한다. 또한 ISO/IEC TR 24028 섹션 9.10에서는 표준 테스트 절차에 소프트웨어(모델) 유효성 검사 및 검증, 견고성 고려, 개인정보 관련 문제 고려, 시스템 예측 가능성 고려를 유지하여야 한다고 규정한다.
- 실제와 가상의 시뮬레이션 테스트 환경을 구축할 때, 실제 케이스를 모방한 현실적인 시나리오를 참조하여 현실적인 테스트 절차를 생성하고 가능하면 신뢰성을 보장하는 테스트 계획을 수립하는 것이 바람직하다. 테스트하고자 생성된 환경이 실제 세계에 가까울수록 테스트 결과에 매우 중요하며, 또한 개발된 시스템을 테스트하고자 사용자를 고용하기로 하면, 가짜 사용자/대상에게 실제 세계의 행동과 감정 환경을 모방하도록 경고하여야 한다.
- 테스트 환경에서는 장비(센서, 로봇 부품, 드론, CCTV 카메라 등)도 잘 고려하여야 하며, 이러한 장비 선택에 주의를 기울여야 한다. 테스트 장비를 선택할 때, 대부분의 사용자/조직/개인이 사용 환경에서 제공하는 평균 기술에 적합한 장비를 선택하여야 한다. 비디오 캡처 품질, 사운드 품질 등 외부 요인도 다수에게 적합하도록 조정이 필요하다.
- 또한, 시나리오를 구축하고 사용자를 직접 고용하는 대신 오픈 소스 데이터셋를 사용하거나 합성 데이터셋을 만들고 시스템을 테스트하는 증강 현실AR 및 가상 현실VR 테스트 환경 구축도 고려해 보아야 한다.

참고 모의 공격 데이터셋 활용 모의 테스트 시나리오 예시[34]

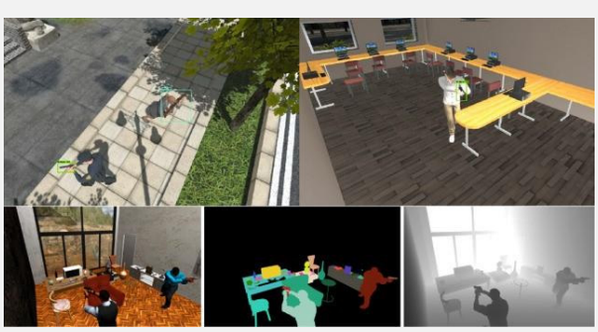
- 데이터셋을 활용한 테스트 환경은 Cam1, Cam7, Cam5라는 세 가지 다른 카메라를 사용하여 생성됨. 실험 설정은 모의 공격하여 수동으로 주석이 달린 데이터를 수집하는 과정에서 이루어졌으며, 연구자들은 당국의 허가를 받고 진행함
- 관련 법적 문서와 허가를 받는다면 적용 가능성이 상대적으로 높아 보임
  위의 실험에서, 연구자들은 대학 내에서 이루어지는 기회를 이용하였으나, 만약 다른 환경을 사용할 기회가 있고 법적인 허가를 받는다면, 이와 같은 테스트 환경/시나리오를 얻음

03-1b 가상 테스트 환경이 필요한 인공지능 시스템의 경우, 시뮬레이터를 확보하였는가?

Yes ☐ No ☐ N/A ☐

- 스마트 치안 시스템의 성격상 많은 경우 시스템을 테스트하고자 공개 데이터셋를 확보하는 경향이 있다. 그러나 여전히 개발 중인 치안 시스템의 종류에 따라 테스트 솔루션은 다양하게 접근한다. 만약 비용이 많이 들거나, 테스트 환경 구축이 어려운 시스템(예: 공공장소 감시 시스템을 테스트하는 환경, 감시 로봇 솔루션을 테스트하는 복잡한 교통 환경 등)을 개발한다면, 가상 시뮬레이션 테스트 환경을 고려하여야 한다.
- 가상 테스트 환경은 스마트 치안 시스템이 처리하도록 설계된 실제 세상의 시나리오와 조건을 정확하게 복제하여야 한다. 이는 다양한 환경 요소, 예를 들어 조명 조건, 날씨 조건, 물리적 장애물 등을 모방하여 시스템이 다양한 상황에서 성능을 평가하는 것을 포함한다. 또한 현실적이고 다양한 데이터를 생성하는 것은 스마트 치안 시스템의 능력을 평가하는 데 필수이다. 비디오 피드, 센서 입력, 네트워크 트래픽 등 시뮬레이션 된 데이터는 실제 세상의 이벤트를 모방하는 시나리오를 생성하는 데 사용되며, 이를 통해 시스템의 탐지, 인식, 반응 능력을 종합적으로 테스트한다.
- 가상 테스트 환경은 스마트 치안 시스템이 상호 작용하는 하드웨어 구성 요소와 네트워크 인프라를 에뮬레이션해야 한다. 이를 통해 다른 네트워크 구성과 배포 시나리오에서 시스템의 호환성, 성능, 확장성을 평가한다.
- 또한, 개발한 치안 시스템에 대한 공격 대응 능력 확인 필요 시에도 고려해 볼 수 있다. 이 목적을 이루고자, 가상 테스트 환경은 침입 시도, 데이터 유출, 네트워크 방해 등의 공격을 시뮬레이션하여 시스템이 위협을 효과적으로 탐지하고 완화하는지 평가한다. 이 외에, 데이터 개인정보 보호, 암호화, 접근 제어 또는 산업이나 지역에 특정한 다른 치안 관련 표준에 대한 시스템 능력도 평가 가능하다.

참고 Unity 합성 데이터셋 활용 모의 테스트 시나리오 예시[45]



- 데이터셋 환경은 Unity 게임 엔진에서 모델링하여 생성됨
  연구자들은 가상 테스트 환경에 여러 카메라를 설치하고 11개의 모델과 7개의 애니메이션을 수행함
- 이런 종류의 테스트 환경을 실제 세계에서 구축하려면 비용이 많이 들고, 정부나 지방 자치단체에 법적인 허가를 받기 어려운 단점이 있으므로, 개발 시 합성 데이터셋, Unity처럼 인공 환경 또는 시뮬레이션 테스트 환경을 사용한다면 이런 종류의 테스트 솔루션을 사용함

03-2 인공지능 시스템의 테스트 설계에 필요한 협의 체계를 구성하였는가?

Yes ☐ No ☐ N/A ☐

- 대부분의 인공지능 시스템은 복잡도가 높아 재현가능성reproducibility이 떨어져 투명성 확보에 어려움을 갖는다. 또한, 시스템의 복잡도는 기대 출력을 결정하는 테스트 오라클test oracle에 문제가 되기도 한다. 이에 따라 테스트가 통과 또는 실패했는지 그 여부를 판단하기 어렵다.
  - ✓ 인공지능 시스템의 테스트 오라클 문제를 다루기 위해. 기존 시스템을 부분적인 오라클로 사용할 수 있는 A/B 테스팅A/B testing, 입력값과 출력값 사이의 관계를 통해 시스템 동작을 확인하는 메타모픽 테스팅metamorphic testing 등의 테스팅 기법을 적용해볼 수 있다.
- 인공지능 시스템의 추론 결과에 대한 설명이 필요한 시스템이라면, 시스템 출력을 확인하는 대상 사용자에 따라 설명가능성*에 대한 평가 기준이 달라질 수 있다. 그리고 인공지능의 작동 방식을 이해하는 정도인 해석가능성interpretability의 평가 기준 역시 대상 사용자에 의존한다.

  * ISO/IEC TR 29119-11:2020에서는 설명가능성을 '인공지능 시스템이 주어진 결과를 어떻게 도출했는지 이해하는 정도'라고 정의하며, 해석가능성을 '인공지능 기술이 작동하는 방식에 대한 이해 정도'로 정의한다.
- 따라서 인공지능 시스템의 기대 출력에 대한 결정이나, 시스템 출력에 대한 설명가능성 및 해석가능성 평가 기준 수립에 필요한 협의 체계를 구축함으로써 협의체를 구성하고, 구성원 간 합의 도출을 통해 테스트를 설계하는 방식이 적절하다.

03-2a 인공지능 시스템의 기대 출력을 결정하기 위한 협의 체계를 구성하였는가?

Yes ☐ No ☐ N/A ☐

- 테스트 오라클의 문제 극복이 필요한 인공지능 시스템이라면, 시스템의 기대 출력을 결정하고자 해당 도메인의 내외부 전문가로 구성된 협의체를 구성하여야 한다. 이때 기대 출력을 결정하고자 여러 전문가가 동의하는 데 시간이 걸린다는 것을 인지하여야 한다.
- 협의체 전문가들은 단일 입력에 대해 다른 출력을 기대한다. 그러므로 협의체 운영 전에 전문가 합의에 대한 승인 기준을 미리 결정해 두어야 한다. 이때, 위반 예방이나 얼굴 탐지 작업 등 대부분에서 정확도, F1-Score, 재현율Recall, FPS를 사용하여 기대 출력의 수준을 합의한다.

참고 범죄 탐지 시스템에 대한 상담 시스템 대응 방안[35][36]

자문 과정에 참여할 대상을 결정하고, 이에는 법 집행 기관, 치안 전문가, 그 외 관련 당사자 등이 포함됨

스마트 치안 시스템의 구체적인 목표, 감지할 범죄 유형, 예상 결과 등, 자문 과정의 범위를 설정

해당 지역의 범죄 패턴, 가장 흔한 범죄 유형 그리고 현재 범죄를 탐지하고 예방하는 데 사용되는 방법에 관한 관련 정보를 수집

수집한 정보를 사용하여 스마트 치안 시스템의 잠재적 효과성을 분석하며, 이에는 범죄 탐지 및 예방 능력, 정확도 그리고 개인정보 보호와 시민 자유에 미칠 잠재적 영향을 포함

이해관계자들과 협력하여 스마트 치안 시스템의 예상 출력에 대한 피드백과 통찰력을 수집 및 분석
이는 워크숍, 설문 조사, 또는 다른 형태의 자문을 통해 이루어짐

받은 피드백을 바탕으로 스마트 치안 시스템의 예상 출력에 대한 권고 사항을 개발
이러한 권고 사항은 시스템의 목표, 잠재적인 위험과 이점 그리고 윤리적·법적 고려 사항을 고려하여야 함

스마트 치안 시스템의 예상 출력이 결정되면, 시스템을 구현하고 성능을 모니터링하여 원하는 결과를 달성하는지 확인

COMPAS 사례
2016년에 ProPublica에 의한 조사에서 오랫동안 미국 법원에서 사용된 COMPAS 알고리즘이 흑인 피고인에게 편향되어, 향후 범죄를 저지를 "고위험"으로 잘못 라벨링 하는 것이 백인 피고인보다 두 배나 많음을 발견함
Northpointe는 이러한 주장을 반박하며, 알고리즘이 의도한 대로 작동함을 주장함(자세한 내용은 04-2b 참조)
따라서 COMPAS 사례는 범죄 탐지 시스템 등스마트 치안 시스템의 윤리적·법적 영향 고려의 중요성을 강조함
이러한 시스템이 효과적이고, 윤리적이며, 이해관계자들의 필요와 기대에 부합하는지 보장이 중요함

03-2b 설명가능성 및 해석가능성을 확인을 위한 사용자 평가단을 구성하였는가?

Yes ☐ No ☐ N/A ☐

- 추론 결과에 대한 설명이 필요한 스마트 치안 시스템 개발 시, 시스템의 설명 가능성과 해석 가능성을 테스트할 때 치안 인공지능 시스템의 대상 사용자/개인이 시스템 출력과 작동 방식을 얼마나 쉽게 이해하는지 확인하여야 한다.
- 따라서, 사용자 평가단을 구성하여 설명을 어떤 난이도로 제공할지 결정하고, 이를 모델과 시스템 구현 시 반영하여야 한다. 이에, 생명주기 관리 단계에서 대상 사용자를 명확하게 정의한 후 사용자 평가단을 구성하여야 한다.
- 사용자 평가단의 평가 결과에 따라, 추론 결과가 편향되지 않는지 여부를 결정하는 기준 수립이 필요하다. 예를 들어, 평균 점수가 특정 점수 이상일 때 합격으로 판정하는 기준을 수립하거나, 그러한 값을 결정 내릴 때 가지치기 평균을 사용할지 여부 등 정성적 계산 기준을 수립하여야 한다.
- 또한, 정성적 계산 기준과 함께 인공지능 시스템이 훈련 데이터셋의 출처에 대해 충분한 설명을 제공하는지 그리고 훈련 데이터셋과 제공된 결과 사이의 상관관계를 자세히 설명하여야 한다. 이를 평가 기준으로 고려한다.

참고 '범죄 통제 및 예방하는 공공 감시 카메라'의 프로세스 평가에 대한 체크리스트 예시[37]

평가 그룹은 다음과 같은 사람들로 구성됨
계획자 및 도시 관계자, 범죄 분석가/기술자, 모니터링 요원, 조사자
사용자에게 시스템 설명 시 현재 단계 연구의 프로세스를 평가하고자 다음 연구 질문을 도출함
도시들이 왜 공공 감시 목적으로 공공 감시 기술에 투자를 결정하는가? 그들은 투자에서 무엇을 얻기를 희망하는가?
어떤 카메라를 구매하고 어떻게 배치 및 모니터링 결정에서 어떤 요인들이 어떤 역할을 하는가?
공공 감시 카메라의 투자 및 사용 결정에 공공이 어떻게 참여하는가?
카메라는 어떻게 실시간 체포를 지원하며, 어떻게 수사 목적에 사용되는가?
감찰 목적으로 공공 감시 카메라를 사용하는 장점과 한계는 무엇인가?

책임성 투명성

요구사항 04 인공지능 시스템의 추적가능성 및 변경 이력 확보

- 스마트 치안 인공지능 시스템은 사용자가 시스템에 대한 이해나 지식이 부족할 때 문제가 발생할 가능성이 높다. 따라서 성능, 사용 그래프, 사용 습관 등을 추적하고 정기적으로 분석하여 시스템 오류를 방지하여야 한다.
- 따라서 시스템 로그, 데이터 모니터링, 인공지능 모델과 사람 간 의사 결정 기여도 추적, 변경 이력 관리 등의 방법을 적용하여 문제 발생 원인을 추적하여야 한다. 또한 시스템 내 인공지능 모델의 성능 개선 시 변경된 데이터를 활용하여 모델 재학습 등을 수행하였을 때, 데이터의 변경 시점, 접근 사용자, 변경 내용 등을 모니터링하고 변경 이력을 관리하는 등 기술적 대응 방안을 확보한다.

04-1 인공지능 시스템의 의사결정에 대한 추적 방안을 수립하였는가?

Yes No N/A
☐ ☐ ☐

- 인공지능 시스템의 의사결정은 인공지능 모델이 자체 결정하거나 시스템 운영자 또는 사용자가 개입하여 내린다. 또한, 운영 중에도 학습이 이루어지도록 설계·개발된 인공지능 시스템이라면 학습 데이터와 모델에 대해 지속적인 모니터링이 필요하다.
- 인공지능 모델의 구축, 데이터셋, 시스템 자체 등 기능적 측면과 인공지능 시스템 운영자 및 사용자 등 인적 요인으로 인해 가능한 인공지능 시스템 추론 결과의 영향을 추적하고자 시스템 단계별로 로그 수집 대상 정보를 정의하고 모니터링을 지속하여야 한다.
- 또한 사용자 경험을 효과적으로 전달하고자, 의사 결정 결과에 대한 사용자 응답 시간 정보, 사용자 개입 요청, 시스템 상태 알림, 사용자 행동을 유발하는 시스템 지표 등을 수집, 분석, 관리하도록 한다.

04-1a 인공지능 시스템의 의사결정에 대한 기여도 추적 방안을 확보하였는가?

Yes No N/A
☐ ☐ ☐

- 인공지능 시스템의 결정에 대한 모델 기여도를 파악하려면 이전 모델의 추론 정보와 최종 결정에 대한 사람(예: 치안 담당자) 개입 여부 및 상호 작용 등의 정보가 추적되어야 하며, 시스템 결정에 대한 자세한 기여 기준을 내부적으로 수립하고 시스템 운영 과정에서 결정의 영향 정도를 추적한다(예: 로그 수집).
- 카메라, 엣지 디바이스, 레이더 등에서 얻은 다중 데이터의 평가 결과를 결합하여 의사 결정에 사용할 때, 각 출력의 기여 기준을 세분화하여 필요시 사용자에게 제공하도록 추적한다.

- 스마트 치안 시스템에서 의사 결정 과정의 기여도를 추적하여 시스템의 무결성, 보안 및 시스템의 신뢰성을 유지하고 관리 용이성 등을 달성한다. 이는 책임 추적, 규정 준수, 보안, 최적화 등 의사 결정에 필요한 데이터를 제공한다.

04-1b 인공지능 시스템의 의사결정 추적을 위한 로그 수집 기능을 구현하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템의 전 생명주기를 고려한 추적 가능성을 보장하려면 모델의 학습 과정, 운용 시 의사결정 결과, 사용자 입력 데이터 등의 정보에 대한 지속적인 수집이 필요하다. 따라서 시스템 프로세스별 로그를 수집할 정보를 선정하고, 정보 간 중요도를 정의한 뒤 로그 레코드 형식을 결정하여 로그를 수집하여야 한다.
- 스마트 치안 시스템의 로그 수집 기능은 카메라, 출입 통제 시스템 및 침입 탐지 시스템 등 다양한 소스에서 데이터를 캡처하고 기록 관리하며, 캡처한 데이터에는 이벤트, 경고, 사용자 활동 및 시스템 상태 업데이트가 포함된다.
- 특히 인공지능 시스템 운영 과정에서 오류 원인을 추적하려면 모델 구축 방법과 데이터셋 측면을 포함한 오류 원인의 분석이 필요하므로, 세 가지 측면을 고려하여 로그를 수집하여야 한다.

인공지능 시스템 운영에서 발생하는 오류의 원인 예시

오류 분류	오류의 원인 예시
모델 빌드 방식에 따른 오류	모델 및 데이터의 대상 선정, 수집, 정제, 라벨링에 대한 제어가 미흡하여 구축 절차, 구조, 학습 모델 측면에서 다양한 오류 데이터가 생성된다.
인공지능 구현상의 오류	개발된 인공지능 모델과 조직이 이미 보유한 기술을 연계하는 전략이 부족하다. 인공지능 구현 전략에서 개발된 모델과 조직이 기존에 사용하던 프로세스 및 기술의 호환성을 확인하지 않으면 오류가 발생한다.
데이터셋 측면의 오류	사용자들에게 시스템에 대한 경험에 대한 피드백을 제공받음 따라서 설문 조사, 피드백 양식 또는 사용자 콘텐츠의 감성 분석을 사용함

참고 스마트 치안 시스템에 대한 로그 수집의 직접적인 효과 조사[149]

오송 지하 차도 참사는 2023년 7월 16일 대한민국 충청북도 오송에서 발생한 침수 사고이다. 홍수로 인해 수많은 사상자가 발생했으며 당국의 재난 대응에 대한 비판이 이어졌다. 정부는 이 사건에 대한 조사에 착수했고, 재난 발생 후 경찰에 대한 검찰 수사가 요청되었다. 당국은 구조 활동이 지연되고 응급조치가 부적절했다는 보도와 함께 홍수에 대한 잘못된 대응으로 비판받았다.
감사 결과, 출동한 경찰관이 처음에는 현장에 출동했다고 보고했지만, 나중에 현장에 있었던 사실을 부인해 순찰차가 실제로는 A 지하도가 아닌 B 지하도로 간 것으로 밝혀진 이 사건은 잘못된 초기 정보로 인한 출동 오류를 감사에서 확인한 사례로 꼽힌다. 이 상황은 시스템에서 수집한 로그를 통해 밝혀졌다.

04-1c 지속적인 사용자 경험 모니터링을 위해 사용자 로그를 수집 및 관리하고 있는가?

Yes ☐ No ☐ N/A ☐

- 서비스 이용 로그 분석은 서비스 운영 상태에 관한 확인뿐만 아니라, 사용자가 겪는 문제가 무엇인지 확인하는 가장 기본적인 방법이다. 서비스 로그는 서비스가 운영되는 동안 지속해서 수집되며 서비스 고도화에 따라 다양한 형태로 누적된다.
- 사용자 로그는 사용자의 활동 및 시스템 이벤트를 기록하고 분석하고자 수집 및 저장하는 기록이다. 스마트 치안 시스템에서는 사용자의 로그인·로그아웃 시간, 오픈한 현관문·창문, 작동한 센서, 발생한 오류 및 시스템 문제 등의 정보가 포함된다. 사용자 로그를 수집하여 시간이 지남에 따라 사용자의 행동과 시스템 성능을 모니터링하고 분석 또한 가능하다.
- 서버 인프라에 대한 로그를 통해 서비스 운영 상태에 대한 모니터링을 수행하며, 사용자 상호 작용 로그는 사용자가 어떤 서비스를 많이 이용하고 어떤 서비스에서 오류를 겪는지 분석한다. 따라서 인프라 관점에서는 로그 분석 소프트웨어를 활용하며, 사용자 관점에서는 기업이 자체적으로 인터페이스 또는 상호 작용의 호출에 따른 로그를 수집하거나 로그 분석 도구를 활용한다.

스마트 치안 시스템의 사용자 로그 이벤트 수집 예시

치안 시스템 종류	로깅 항목	설명
감시	이벤트의 날짜 및 시간	활동 추적과 잠재적인 위협을 식별하는 데 활용
	카메라가 기록한 사건의 ID	사건의 위치를 식별하고, 카메라에서 영상을 검토하는 데 사용함
	감지된 사람 또는 물체의 ID	이벤트를 발생시킨 개인 또는 객체를 식별하는 데 사용함
	수행된 동작의 유형	특정 사건을 식별하는 데 사용함
범죄 예측	범죄의 날짜 및 시간	범죄 동향을 추적하고 범죄가 발생할 가능성이 높은 지역을 식별하는 데 활용함
	범죄의 종류	발생한 특정 범죄를 식별하고 투입할 자원 및 개입을 의사 결정하도록 사용함
	범죄 예측 발생 위치	범죄가 발생한 지역을 식별하고 이에 맞게 투입할 자원 및 개입을 의사 결정하도록 사용함
	범죄와 관련된 사회 및 경제적 요인	이는 범죄에 기여하는 요인을 파악하는 예측 모델 내에서 공정성을 추구하며, 실업률, 빈곤 수준 등 사회적·경제적 요인을 고려하는 것이 필수적이다.* 그러나 어떤 요인을 포함하고 얼마나 많은 가중치를 부여할지 결정하는 과정은 모델 편향을 방지하고 모델의 목표를 유지하고자 하는 가치와 일치하도록 안내하여야 한다. * 개인정보 보호법 제37조에 따라 정보 주체(개인)는 자신의 개인정보 처리 정지를 요구한다.
가정 치안 시스템	이벤트 날짜 및 시간	실내 장소의 활동과 잠재적인 위협을 추적하는 데 사용함
	작동된 센서 ID	이벤트의 위치를 식별하고 센서에서의 영상을 검토하는 데 사용함
	센서의 상태 (예: 작동됨, 작동되지 않음)	센서의 상태를 식별하고 이벤트가 가짜 경보인지 실제 위협인지를 판단하는 데 사용함
	발생한 이벤트의 유형 (예: 문 열림, 창문 깨짐)	해당 사건을 식별하고 적절히 조치하고자 사용함

04-2 학습 데이터의 변경 이력을 확보하고, 데이터 변경이 미치는 영향을 관리하였는가?

Yes No N/A
☐ ☐ ☐

- 인공지능 모델은 사용한 데이터에 따라 학습 모델도 함께 달라진다. 이로 인해 모델의 설계나 주요 파라미터들의 변경이 함께 이루어진다. 따라서 모델 개발 과정에서 학습 데이터가 변경될 때, 학습 데이터 버전 관리 및 변경이 발생한 원인을 추적하여야 한다.
- 또한, 신규 데이터(영상, 센서, 엣지 디바이스, 레이더, 합성 등)를 포함하여 인공지능 모델의 추가 학습이 필요할 때, 학습 데이터 변경으로 인한 모델의 성능 영향을 평가하고자 기존 학습 데이터에 추가된 신규 데이터 비율에 따른 모델 성능 변화 추적이 가능하도록 기록 및 관리하는 것이 바람직하다.
- 학습 데이터를 사용하거나 운영하는 이해관계자는 학습 데이터 변경 이력을 관리하고자 학습 데이터 버전 관리용 오픈 소스 도구를 사용하거나, 자체 시스템을 구축하는 방법 등을 고려하여야 한다. 이러한 학습 데이터 관리 시스템은 학습 데이터의 변경 원인, 변경된 학습 데이터의 구조, 학습 모델의 추론 결과, 모델 변경에 따른 성능 평가 결과 등에 대한 정보를 추적하고 관리하여야 한다.

참고 스마트 치안 시스템에서 데이터 흐름 및 계보 추적 시 도구 예시

스마트 치안 시스템에서 데이터 계보를 추적하고자 사용하는 다양한 도구와 기술들이 있다. 이러한 도구들은 데이터베이스, ETL 도구, BI 플랫폼 등 다양한 출처에서 메타데이터를 분석하여 데이터 계보 추적 과정을 자동화한다.

도구	설명
Spline[141]	Spline은 데이터베이스, 도구 등 다양한 소스에서 데이터를 분석하여 데이터 계보 추적 과정을 자동화하는 도구이며, 데이터 계보의 수집 및 쿼리 목적에 대한 API가 다.
OpenMetadata[142]	OpenMetadata는 오픈 소스 메타데이터 저장소로서, 데이터 카탈로그 작성, 발견 그리고 데이터 생태계 전반에 걸쳐 협업을 가능하게 한다. 이는 다양한 데이터 소스에 대한 데이터 계보를 관리하고 데이터 품질, 사용, 관리에 대한 인사이트를 제공한다.
Apache Atlas[193]	Apache Atlas는 데이터 관리와 메타데이터 프레임워크를 제공하는 오픈 소스로서, 데이터 계보 추적 기능을 제공한다. Hadoop 클러스터, 데이터베이스 그리고 기타 데이터 소스에 대한 데이터 계보를 추적한다.

04-2a 데이터 흐름 및 계보lineage 추적하기 위한 조치를 마련하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템은, 데이터의 변경으로 인해 모델의 확장이나 재설계 등의 시스템 변경이 발생한다. 따라서 시스템의 변경을 유도하는 데이터의 흐름 및 계보를 계속해서 추적하여야 한다.
- 데이터 흐름과 계보는 데이터 변경에 대해 역방향, 순방향 그리고 종단 간end-to-end 관점에서 추적하는 고려 사항은 다음과 같다.
  - ✓ 데이터 흐름과 계보 추적을 관리하고자 데이터 정책 팀을 설립하는 것이 유용한가?
  - ✓ 데이터 흐름과 계보를 추적하고자 메타데이터가 기록되고 유지되는가?
  - ✓ 데이터 흐름과 계보를 추적하고자 데이터 로딩, 매핑, 관리, 시각화 보고 기능을 구현하는 것이 유용한가?
  - ✓ 인공지능 개발 과정에서, 모델의 특성값을 저장하고 공유하는 특성 저장소 기능 구현이 유용한가?
  - ✓ 데이터를 원본까지 추적하는가?

04-2b 데이터 소스 변경에 대한 모니터링 방안을 확보하였는가?

Yes ☐ No ☐ N/A ☐

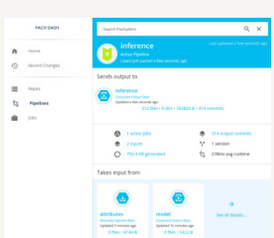
- 실시간으로 훈련 데이터를 수집하고 위협 가능성을 감지하거나 이상 예측을 하고자 인공지능 모델을 실시간으로 훈련하는 온라인 학습 방법을 적용하는 시스템은, 이 검증 항목이 고려 사항이 아닐 수 있다. 스마트 치안 시스템은, 훈련 데이터의 변경이 시스템 위협을 감지하고 대응하는 능력에 미치는 잠재적인 영향을 고려하는 것이 중요하므로, 데이터에 대한 모든 변경 사항은 데이터 개인정보 보호 및 치안과 관련된 관련 규정이나 표준을 준수하며 신중하게 테스트되고 검증되어 시스템 또는 사용자의 치안을 저해하지 않음을 보장하여야 한다.
- 스마트 치안 시스템 알고리즘을 개발하고자 오픈 소스 데이터셋을 사용할 때(04-1b 참조), 데이터셋의 변경 또는 업데이트가 빈번히 발생하므로 모델의 성능을 향상하고자 주기적으로 모니터링하여 최신 데이터셋을 반영하는 것이 필요하다.

04-2c 데이터 변경 시, 버전관리를 수행하였는가?

Yes No N/A
☐ ☐ ☐

- 인공지능 모델 개발 과정에서 학습 데이터의 업데이트, 오류로 인한 라벨링 재수행 등 데이터 변경이 이루어지면 학습 결과인 모델도 변경된다. 또한 이전에 학습에 사용한 데이터셋과 특성이 완전히 다르거나 데이터셋 전체를 교체할 때 성능이 크게 저하되며, 이때는 추가 학습이 필요하다.
- 따라서 학습 데이터의 변경이 수행될 때, 단순히 사용된 학습 데이터의 버전뿐만 아니라 해당 버전으로 학습한 인공지능 모델을 함께 관리하여야 한다. 특히, 신규 데이터의 추가로 인한 학습 데이터 변경이 필요할 때, 학습 혹은 테스트에 사용된 신규 데이터 비율을 기록하고, 그에 따른 모델의 성능 변화가 함께 추적 가능하여야 한다.
- 학습 데이터의 변경 사항을 추적하려면 버전 관리 시스템의 사용이 유용하다. Git, SVN과 같은 버전 관리 시스템을 사용하거나, DVC[Data Version Control], Pachyderm, Git LFS[Large File Storage], lakeFS, Delta Lake 등 전용 데이터 버전 관리 도구를 사용하여 데이터와 코드의 변경 사항을 추적하는 것을 포함한다. 또는 기계 학습 프로젝트의 자체적인 학습 데이터 버전 관리 시스템 구축도 가능하다.

참고 오픈 소스 기반 데이터 버전 관리 도구 예시

도구	설명
	DVC[Data Version Control] Tool[143] 오픈 소스 비주얼 스튜디오 코드 확장 프로그램 및 커맨드 라인 도구. Git 저장소 위에서 작동하며 Git과 유사한 커맨드 라인 인터페이스와 흐름을 가짐 DVC는 데이터와 기계 학습 테스트를 문서화 큰 파일, 데이터셋 디렉터리, 기계 학습 모델 등을 더 작은 메타 파일로 교체하여 관리(Git 처리가 쉬움). 즉, 소스 코드 관리와 분리된 원본 데이터를 가리킴. 데이터 저장: 프로젝트 내의 데이터를 코드 베이스와 별도로 사용하게 온프레미스 또는 클라우드 스토리지를 사용
	Pachyderm[144][145] 완전한 버전 관리 데이터 사이언스 플랫폼 커뮤니티 에디션 버전은 오픈 소스이며 어디에서나 배포 가능 기계 학습 수명 주기를 끝까지 관리하도록 도움 대표 기능: • 리포의 마스터 브랜치에서 데이터를 지속해서 업데이트 • 모든 유형, 크기, 파일 수를 지원 • 커밋은 중앙 집중화되고 트랜잭션 관리됨 • 팀이 서로 작업을 기반으로 작업을 공유, 변환, 업데이트하게 함

04-2d 데이터 변경 시, 이해관계자를 위한 정보를 제공하는가?

Yes ☐ No ☐ N/A ☐

- 다수의 이해관계자가 참여하는 인공지능 시스템의 개발 과정에서 데이터 변형으로 인한 인공지능 모델의 설계, 주요 초매개 변수의 변경 및 재학습 등의 조치를 이해하려면 이해관계자의 역할을 고려한 정보 제공이 필요하다.
- 데이터 변경에 따라 각 이해관계자에게 제공하는 정보는 다음과 같다.

데이터 변경 시 이해관계자 및 제공하는 정보 예시

이해관계자	제공 정보
비즈니스 결정권자	• 데이터 변경에 따른 모델의 세세한 변경점보다 기존 시스템의 목적, 서비스 의도 등의 변경점이나 시스템 전체의 방향성 등에 초점을 맞춘 정보
데이터 과학자	• 기존 데이터와 변경된 데이터의 특징, 포맷, 규모 등의 차이점 등의 정보
시스템 개발자	• 변경된 데이터 설명을 참고하여 기존 모델과의 호환성, 모델 구조 재설계, 모델 재학습 세부 전략(예: 목적함수, 학습 시간, 학습 알고리즘), 예상 출력 결과 변경점 등에 대한 정보
모델 검증자	• 변경된 테스트 데이터셋 구성, 재설계 및 재학습된 모델에 대한 주요 성능 평가 결과, 기존 모델과의 성능 비교 결과 등의 정보
모델 운영자	• 검증을 마친 변경 모델에 대한 운영 및 사용자 모니터링 결과 등을 수집 및 분석한 정보

04-2e 신규 데이터 확보 시, 인공지능 모델의 성능평가를 재수행하였는가?

Yes ☐ No ☐ N/A ☐

- 신규 데이터를 확보한 뒤, 인공지능 시스템에 사용하려면 기존 운영 중인 인공지능 모델과 성능 비교가 필요하다. 사람이 판단하기에 신규 데이터가 기존 학습 데이터와 유사하여도 학습된 인공지능 모델이 기존 학습 데이터에서 학습한 데이터 특성과 다르다.
- 스마트 치안 시스템에서는 인공지능 모델이 잠재적인 치안 위협을 감지하고 대응하는 데 자주 사용된다. 이러한 모델은 레이블이 지정된 예제 데이터셋으로 학습되고, 그 후에는 별도의 데이터셋에서 테스트를 진행한 후 성능을 평가한다. 그러나 새로운 데이터가 사용 가능해지면, 인공지능 모델을 재훈련하여 계속해서 위협을 감지하고 대응하도록 보장하는 것이 필요하다.
- 새로운 데이터 획득에 따른 성능을 평가하고자 다음과 같은 과정을 참고하며, 이 과정에서는 도메인 전문가와 협력을 고려한다.
  - ✓ 기존의 학습 모델과 관련된 대표적인 인공지능 모델을 보호하여 성능 평가와 비교 분석을 수행한다. 따라서 IBM QRadar Advisor with Watson 등 솔루션을 활용한다. 실제로는 인공지능을 활용한 사이버 보안 솔루션이나, 학습 루프 분석 기능을 통해 시스템을 평가하고 더 안정적인 업그레이드 프로세스를 생성한다[146].
  - ✓ 스마트 치안 시스템과 모델에 대한 적절한 성능 평가 지표를 선택한다.
  - ✓ 양적 및 질적 실험 방법의 선택, 실험 모델의 매개 변수 설정, 상세한 실험 계획 등 성능을 평가하는 시험을 설계한다.
  - ✓ 결과에 기반하여 새로운 데이터를 평가하거나 필요시 모델을 재설계, 확장, 재교육 결정 등 시험 진행하고 결과를 분석한다.
  - ✓ 치안 시스템과 관련된 새로운 데이터를 수집하고 적절한 클래스로 레이블을 지정한다. 데이터는 다양한 잠재적 위협과 시나리오를 커버할 정도로 다양성을 확보하여야 한다.

책임성 투명성

요구사항 05 데이터 활용 시 상세 정보 제공

- 폭력 감지, 얼굴 감지, 화재 감지, 거짓말 감지 등 프로젝트의 과제와 관련된 상황을 평가하고자 개발된 시스템의 훈련 및 테스트 단계에서 사용되는 데이터는 매우 중요하다. 그러나 획득한 데이터의 특성상 예측할 수 없는 부분이 존재하기에 정보의 사용에 관한 모든 법적 의무를 준수하여야 한다.
- 개인의 이름, 신원, 범죄 이력, 여권 정보 등 고도의 개인정보가 포함된 데이터셋 또는 개인 영상 및 감시 비디오 데이터를 획득하기로 한 때, 또는 개발한 지능형 치안 시스템의 맞춤형 데이터셋을 만들기로 한 때, 특히 원하는 데이터셋에 오픈 소스로 액세스할 수 없을 때 데이터 사용, 데이터 특성, 메타데이터 정보, 데이터 라벨링 절차, 라벨링 담당자에 대한 교육에 관한 포괄적인 세부 정보 제공이 필수적이다. 이러한 정보는 개발 상황에 맞게 조정하여 잠재적인 데이터 관련 문제를 사전에 해결하도록 하여야 한다.

05-1 데이터의 명확한 이해와 활용을 지원하는 상세한 정보를 제공하는가?

Yes ☐ No ☐ N/A ☐

- 메타데이터는 데이터를 설명하는 데이터로 정의하며, 원시 데이터의 특징들을 메타데이터에 기록하여 향후 데이터를 재활용하는 상황이나 동일한 형식의 데이터를 추가로 수집하여야 할 때 데이터에 대한 정보를 전달한다.
  - ✓ 메타데이터: 인종, 성별 주석, 거짓말 탐지에 사용되는 언어, 거짓말 탐지 결과에 영향을 미치는 트라우마 또는 심리적 장애의 유무, 상호 작용의 성격, 당사자 간 친밀도, 지역 교통 통계, 시설 및 국가별 위험 사고에 대한 통계적 인사이트 등의 특징적인 정보
- 메타데이터에 포함할 수 없을 때, 이러한 특성 정보를 별도로 제공한다. 또한 인종, 성별, 나이, 직업, 여권 정보, 방문 국가 정보, 공공장소에서의 행동, 숙련도, 특징적인 감정 상태 등의 보호변수는 신중하게 명시해야 한다.
- 또한, 학습 데이터 및 테스트 데이터의 상세 정보, 획득한 데이터의 일반적 상황 및 환경 요인, 시스템을 사용하거나 영향을 받는 대상자의 일반 정보, 메타데이터, 라벨링 운영 가이드 등을 확보하여 개발자뿐만 아니라 스마트 치안 시스템과 관련된 이해관계자가 수집된 데이터를 이해하고 편견이나 오류 가능성을 예방하도록 하여야 한다.
- 공공 감시 시스템, 국경용 거짓말 탐지기, 로봇을 이용한 교통 통제, 회의장 위협 검사 등 개발 시 법적 의무 이행을 입증하는 메타데이터에 관련 정보를 제공하거나 이 정보를 별도로 제공하여야 한다. 이해관계자에게 전달하여야 하는 정보의 예로는 데이터의 출처와 형식, 수집 방법, 정제 및 처리 기술, 데이터 라이선스 세부 정보, 잠재적 편견에 대한 보호 장치 역할을 하는 변수 등이 있다.

05-1a 정제 전과 후의 데이터 특성을 설명하였는가?

Yes No N/A
☐ ☐ ☐

- 데이터 정제 작업은 라벨링 작업 전 학습 데이터를 구축하는 데이터의 선별 및 처리 단계로서, 정제 과정을 거친 데이터만을 사용하는 사용자는 원시 데이터의 특성을 정확하게 파악할 수 없다. 따라서 향후 추가 데이터의 수집 가능성을 고려하여 정제하는 관련 정보와 정제 전과 후의 데이터 특성을 설명하여야 한다.
- 스마트 치안 분야에서 데이터를 저장하고 처리하는 비용이 증가함에 따라 데이터 정제 과정이 필수적이다. 예를 들어 국경 통제를 하는 거짓말 탐지기, 대규모 이벤트 시 위협 스크리닝, 가정 치안, 범죄/폭력 예방/탐지, 감시 시스템에서 비디오 데이터는 중요한 부분을 차지하므로, 데이터 정제에 더 큰 비용이 든다. 이러한 데이터의 특성상 대부분의 데이터는 깨끗하거나 신뢰성 있는 정보가 부족하다. 따라서 원시 데이터의 정제는 데이터를 준비하는 주요 과정 중 하나이다. 또한, 정제 전후의 데이터 특성에 대한 정보를 제공하여야 이해관계자들이 데이터를 적절하게 활용한다.
- 특히, 민감한 특성으로 인해 시스템이 편향된다. 따라서, 데이터 품질을 보장하고, 데이터 관리를 최적화하며, 데이터 구축 목적을 설정하고, 데이터 유형을 분석하며, 정제 과정의 표준 정보와 정제 도구를 제시하여야 한다.
- 다음은 데이터 유형별 데이터 정제 기준의 예이다.
- 이미지 데이터: 이미지 크기, 해상도, 비율, 이미지 품질, 촬영 장비, 개인정보, 저작권, 위치 정보(메타데이터에 의해 제공될 때) 등
- 비디오: 품질, 비디오 손실, 개인정보, 정치적 의견, 특정 인물 비하, 환경 조건 등
- 음성 데이터: 볼륨, 발음 정확도, 사투리, 잡음 및 간섭, 청취 불능(허용 범위에 따라), 개인정보, 저작권 등
- 텍스트 데이터: STT 처리의 정확도, 단어 의미론, 어휘 사용, 텍스트 길이, 텍스트의 문법 정확도, 텍스트 내용의 적절성, 주제와의 연관성 등(거짓말 탐지기를 STT 분석을 사용해 설계한 때)
- 비식별화 표준: 인간 얼굴, 표정에서 얻은 감정 상태, 눈동자 움직임, 몸짓, 미모 등 개인정보의 비식별화 (개인정보 보호법 고려*)

  * 비식별 처리로 익명화 기법을 사용하는 때, 익명화된 데이터는 일반적으로 비개인 데이터로 취급되며 개인 데이터와 동일한 엄격한 데이터 보호 규정이 적용되지 않는다. 반면, 비식별 처리로 가명 처리 기법을 사용 시 가명 처리된 데이터는 잠재적으로 개인과 연결되어 여전히 개인 데이터로 취급되므로 개인정보 보호법(제3조, 15조, 23조, 24조)에 따른 데이터 사용 및 수집 요건과 제한을 충족하는지 확인하여야 한다.

- 또한 영상정보처리기기(개인 사진 또는 사물 영상 등을 촬영하거나 네트워크를 통해 사진 또는 영상 등을 전송하고자 일정한 장소에 지속해서 설치되는 장치를 말하며, CCTV 카메라, 휴대 전화 카메라 등의 장치를 포함) 등 솔루션 사용 시 제약 사항에 유의하여야 하며, 개인정보 보호법에 따른 영상처리기기에 대한 정보를 포함하여야 한다(제25조, 제58조).
- 3D: 포인트 클라우드 획득, 메시 데이터 최적화, 표준 모델 생성 등

- 센서/ 로봇 장비: 단위, 누락 값, 센서의 기록 시간, 로봇 부품의 축 정보 등

※ 개인정보를 보호하고자 가명 처리 방식이 어려울 때, 도메인 인식 시스템(DAS)에서 사용하는 것과 같이 시스템 수준에서 개인의 개인정보 접근을 차단하는 극단적인 솔루션 사용을 고려하여야 한다(자세한 내용은 [38] 참고).

- 또한, 데이터셋을 구축하는 과정에서 일부 데이터는 데이터 품질을 향상하고자 다시 정제되며, 아래는 정제 후의 학습 데이터 특성을 설명하는 항목의 예이다.
- 데이터 속성 분석: 중복 방지, 이상 데이터 제거, 샘플링 등
- 통계적 설명 항목: 클래스별 훈련 데이터 수, 주제 수 등
- 환경적 설명 항목: 취득 지역(예: 지리적 위치, 날씨 조건, 실내 온도 등), 기록된 시간대, 기록된 장소(예: 실제 장소, 회사, 대학, 거리, 시나리오 환경 등), 개인정보, 영향을 받은 사람의 관련 정보, 공공 감시 시 기록의 거리/동네 정보 등

참고 감시 시스템의 데이터 확산 사용 사례[39]

연구원들이 자기 조직화 연산 신경망(Self-ONN)을 사용하는 드론용 군중 밀도 추정 모델인 DroneNet를 소개한다. DroneNet의 아키텍처 및 하이퍼파라미터 설정과 함께 DroneRGBT 데이터셋을 사용한 훈련 프로세스에 관해 설명한다.

DroneNet 모델에 대한 데이터 수집은 상하이테크 파트-B 및 CARPK처럼 벤치마크 데이터셋을 사용하여 수행되었다. 지상 실측 밀도 맵은 데이터셋마다 다른 시그마(σ) 값을 사용하여 생성되었다.

과적합을 피하고자 수평 뒤집기, 무작위 밝기 및 대비 등 데이터 증강 기법이 사용되었다.

데이터를 보강하고자 밀도 맵을 4x4의 그리드 크기로 분할하여 16개의 패치를 만들었다. 이를 통해 더 많은 훈련 샘플을 확보하고 데이터의 다양성을 높였다.

05-1b 학습 데이터와 메타데이터metadata를 구분하고 각 명세자료를 확보하였는가?

Yes No N/A
☐ ☐ ☐

- 스마트 치안 분야에서 사용하려는 데이터셋과 관련된 모든 정보는 가능하면 수집이 필요하다. 인공지능 훈련 데이터셋을 활용하려면 메타데이터Metadata라고 불리는 정보 파악 및 정리가 필요하다. 메타데이터는 JSON, XML 등의 형식으로 제공된다.

오픈 엑세스 데이터셋 아키텍처 분류

데이터 배포	데이터 이름	데이터 타입	인식 활용	판단 활용	제어 활용
AI Hub	대규모 한국어 DeepFake 탐지 데이터셋[40]	비디오	O (참가자의 원본 이미지 및 변조 이미지, 얼굴 스왑 인식)	X	X
	유동 인구를 분석하는 CCTV 영상 데이터[41]	비디오, 이미지	O (사람 감지, 까마귀 감지, 유동 인구 분석)	O	X
YouTube의 비디오 컬렉션	YouTube의 비디오 컬렉션[42]	RWF2000 시리즈	비디오	O YouTube의 원본 감시 동영상(폭력적, 비폭력적 행동으로 표시됨)	O
데이터 과학 및 전산 지능의 안달루시아 연구소	데이터 과학 및 전산 지능의 안달루시아 연구소[43]	무기 탐지	이미지	O (무기 유형 감지, 무기 인식)	O
음성 분석 및 통역 연구소 – University of Southern California	음성 분석 및 통역 연구소 – University of Southern California[44]	IEMOCAP 데이터베이스	Video, 오디오, 이미지, Text	O (얼굴 표정, 감정, 인간 이딕 상호 작용, 차원 특징(원자가, 활성화, 우성 등)	O
UCSD의 통계 비주얼 컴퓨팅 연구소 (SVCL)	UCSD의 통계 비주얼 컴퓨팅 연구소 (SVCL)[45]	UCSD 변칙 탐지 데이터셋	비디오	O (보행자 감시 영상 원시, 보행자 통로 내 이상 행동 감지)	O

- 메타데이터와 훈련 데이터는 분리되어야 하며, 각각에 대한 사양을 작성하여 개발자의 관점에서 인공지능 모델을 훈련할 때 쉽게 사용하도록 하여야 한다. 이러한 분리는 데이터에 특정 민족 그룹이나 한 지역만을 포함하는 등의 이유로 편향 문제를 일으키는 매우 개인적인 정보/특성/특징을 포함하여 중요하다.
- 예를 들어, AI Hub에서 제공하는 비디오 데이터셋 예시는 데이터의 유형(이미지, 비디오 등), 데이터 영역, 포맷, 유형, 출처, 라벨링 유형과 포맷 그리고 데이터 활용 서비스, 데이터 구축 연도 및 구축량에 대한 정보를 제공한다.

- 개발된 치안 시스템의 메타데이터에는 인종, 캐릭터 감정, 여권 정보, 교육/전문적 배경 정보, 공공장소에서의 행동 패턴, 방문한 나라, 이름 등 민감한 개인정보가 포함된다. 따라서 개인정보 보호법에 따라 이러한 데이터의 사용에 대한 가이드라인을 따라야 하며, 익명화 또는 비식별화하여야 한다. 또한, 항상 데이터셋 제공자에게 얻은 사용 허가 상태를 확인하여야 한다(14-1c 참조).

참고 훈련 데이터 및 메타데이터 명세

▪ AI Hub의 비디오 이미지 분야에서 '유동 인구를 분석하는 CCTV 비디오 데이터' 훈련 데이터 명세 예시 (2023년 3월 기준)

데이터 변경이력

버전	일자	변경내용	비고
1.3	2022-12-12	원천데이터, 라벨링데이터 수정	
1.2	2022-11-23	원천데이터, 라벨링데이터 수정	
1.1	2022-09-28	라벨링데이터 수정	Training, Validation > 라벨링데이터 _0928_add 개방
1.0	2022-07-14	데이터 최초 개방	

데이터 영역	영상이미지	**데이터 유형**	비디오 , 이미지
데이터 형식	jpg, mp4	**데이터 출처**	직접 수집
라벨링 유형	바운딩박스(이미지/동영상), 세그멘테이션	**라벨링 형식**	JSON
데이터 활용 서비스	특정 인상착의 인물 또는 그룹 탐색 서비스, 유동인구 분석을 통한 버스 노선 등의 도시 개혁 정책 개정에 활용 가능	**데이터 구축년도/ 데이터 구축량**	2021년/3분 클립 6,600개(330시간)

▪ AI Hub의 비디오 이미지 분야에서 '유동 인구를 분석하는 CCTV 비디오 데이터' 메타데이터 명세 예시 (2023년 3월 기준)

데이터 통계

구분	분류	항목	시간	비율	구분	분류	항목	시간	비율
1	장소	자갈치시장	49.5H	15%	1	시간대	새벽(0~9시)	65.1H	19.70%
2		서면영광도서 앞	33H	10%	2		오전(9~12시)	81.8H	24.80%
3		사직구장 주변 상권	66H	20%	3		오후(12~16시)	55.4H	16.80%
4		중구 중앙동 주변	66H	20%	4		저녁(16~20시)	65.2H	19.80%
5		동래역 근처	16.5H	5%	5		야간(20~24시)	62.5H	18.90%
6		사상 시외버스 터미널	66H	20%					
7		연산동 먹걸리 골목	33H	10%					

1. 데이터 포맷

과제명	주요 내용	수집 방법	데이터 구축량	데이터 형식
유동인구 분석을 위한 CCTV 영상 데이터	유동인구 분석을 위한 CCTV 영상 데이터 구축	직접 촬영	330시간 분량	영상/Json 맵핑

	사람	상점	자동차
예시			
속성값	성별, 연령대, 상·하의 유형 및 색상, 소지품, 애완동물 동행여부	공실, 주차장, 층수, 엘리베이터, 화장실 유무	이동방향, 상태
포맷	mp4, json	mp4, json	mp4, json

05-1c 보호변수protective attribute의 선정 이유 및 반영 여부를 설명하였는가?

Yes ☐ No ☐ N/A ☐

- 보호변수는 인종, 성별, 나이, 직업, 여권 정보, 방문한 국가 정보, 공공 행동, 능숙도, 특징적인 감정 상태, 관계 정보, 말하는 방식 등 일반적인 분야에서 사회적 문제를 일으키며, 추가 사용 시 프록시Proxy가 되므로 더욱 중요하다. 예를 들어, 높은 범죄 발생 지역인 불안 지역에 대한 감시 CCTV 시스템을 설계한다고 할 때, 주변 사람들의 공격적인 행동으로 시스템이 해당 지역을 향해 편향되고, 이 위치 변수는 해당 지역을 나타내는 프록시가 된다.
- 개발 단계에서 이러한 종류의 보호변수를 고려하지 않으면 예기치 않은 윤리적·문화적 편향 사례가 발생하므로, 모델의 설계 목적 외의 가능한 편향에 대비하고자, 훈련 데이터 수집 및 라벨링 단계에서 넓은 범위의 민족 다양성 등에 기반한 보호변수를 선택하고 반영하여야 한다.
- 또한, 이러한 정보들은 치안 시스템 모델에 대한 중요하고 보호되는 특성으로 사용되며, 보호변수가 신중하게 선택되지 않을 때, 한국은 개인정보 보호법 제15조, 제62조 및 개인정보 보호법 시행령 제18조에 따라 이는 개인권 침해로 이어진다. 이에, 수집된 데이터와 구축된 데이터의 미래 사용자를 고려한 인공지능 시스템의 개발 목적 그리고 데이터셋의 보호변수의 이유, 과정, 반영에 관해 설명한다.

참고 보호변수의 선정 및 반영이 필요한 시스템 예시

2016년 7월 7일 미국에서 Knightscope K5 자율 보안 로봇이 보안 검색 도중 16개월 된 아이가 위험하다고 인식하고 공격하여 부상을 입힘. 당국이 사건을 조사할 때, 이 로봇이 이전에 다른 아이를 보안 위협으로 간주하고 쫓은 것을 발견함. 이 사건은 나이 차이나 아이들에 대한 가치 차이가 시스템에 대한 프락시 발생으로 분석됨

출처: ABC7 News, “부모들이 Stanford 쇼핑센터의 보안 로봇이 아이를 다치게 한 것에 분노”, 2016-12.

보고서에 따르면 중국의 새로운 감시 시스템은 외국 기자들과 국제 학생들 그리고 신장의 무슬림 소수 민족을 추적하는 데 사용됨. 이 시스템은 이러한 목적으로 얼굴 인식과 생체 데이터 분석을 사용함

출처: 로이터, “독점: 중국성, 계획된 새로운 감시 시스템으로 기자들과 외국 학생들을 타깃으로 삼다”, 2021-11.

05-1d 라벨링 작업자를 위해 교육을 시행하고 작업 가이드 문서를 마련하였는가?

Yes ☐ No ☐ N/A ☐

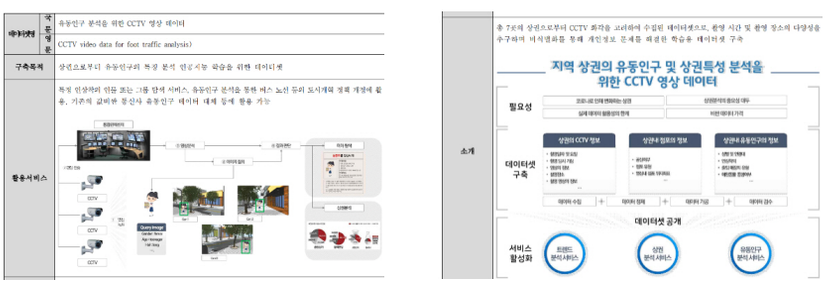
- 데이터 라벨링 작업은 인공지능 모델을 학습시키는 원시 데이터의 주석 Annotation, Labeling, 답변 작업에 해당한다. 특히 스마트 치안 시스템에 자체 데이터셋을 사용할 때를 포함하여, 인식하는 라벨링 작업의 평가는 개인의 생활과 안전에 영향을 미치므로 전문적이고 세부적으로 처리되어야 한다. 따라서, 전문가의 참여도 필수로 고려하여야 한다.
- 전문가의 분야는 프로젝트에 따라 다르다. 예를 들어 폭력 예방 시스템 개발 시 전문가의 분야는 범죄 분석가/프로파일러 등이 되어야 한다.

- 또한, 라벨링 작업의 품질은 다수의 전문가(범죄 전문가, 범죄 분석가, 법률 관계자 등) 및 작업자를 선택하고 합의하는 과정부터 확보되어야 하며, 운영자의 교육과 세부 작업 가이드에 대한 문서화가 중요하다.
- 음성, 텍스트, 시각, 센서, 엣지 장치, 개인정보, 생물학적 정보 등 매우 다양한 속성이 있는 데이터를 수집하여 라벨링할 때, 일반적으로 각 데이터를 동시에 재생, 제어, 라벨링하는 전문 도구가 제공된다. 이러한 전문 도구는 사용 방법, 라벨링 시 주의 사항 등을 손쉽게 학습하도록 가이드 문서를 갖추어야 한다.

참고 라벨링 작업자의 작업 가이드 문서 예시(출처: AI Hub)

데이터셋	작업 가이드 예시
유동 인구를 분석하는 CCTV 영상 데이터	그림 120 라벨링 작업 속성 예시 '멀티모달 비디오' 데이터셋 라벨링 가이드 문서 예시 어노테이션 수행 화면 예시

분류	속성명	속성 설명	데이터 타입	필수 여부	예시
info (CCTV)	year	촬영연도	int	O	2021
	version	버전정보	String	O	1.1
	date_created	촬영일자	String	O	2021/01/01 15:00:00
	day	촬영요일	String	O	fri
	weather	날씨	String	X	sunny
	username	사용자 ID	String	O	admin
video (CCTV)	file_name	파일명	String	O	sample.mp4
	resolution	비디오 해상도 및 컬러	Array[int]	O	[1920, 1080, 3]
	fps	촬영 프레임	int	O	3



단계	작업명	설명
1	저작도구 접속 및 로그인	• 크롬 브라우저를 통해 저작도구 홈페이지 접속 • 생성된 ID와 PWD 입력후 로그인
2	동영상 파일 업로드	• 프로젝트 생성 • 프로젝트 내 태스크 생성 및 파일 선택
3	라벨 추가 및 속성값 설정	• 작업화면 페이지로 이동 • 영상의 각 프레임 별로 식별가능한 객체에 대해서 라벨추가 버튼을 클릭하여 라벨을 추가하고, 해당 객체에 알맞은 속성값 설정
4	바운딩박스 검토	• 플레이 버튼을 클릭하여 재생한후, 바운딩박스가 식별되는 객체에 알맞게 음식이는지 검토
5	바운딩박스 수정	• 바운딩박스가 객체와 연결되지 않고 위치나 크기가 맞아닌 경우, 바운딩박스로 마우스를 올려 드래그 하거나, 기준에 맞지 않는 속성값에 대해서 수정
6	가공 데이터 저장	• 저장 버튼을 클릭하여 작업한 내용들을 저장

05-2 데이터의 출처는 기록 및 관리되고 있는가?

Yes No N/A
☐ ☐ ☐

- 학습 데이터의 품질은 인공지능 모델 성능에 영향을 미치는 핵심 요소 중 하나이다. 따라서 데이터를 수집하거나 생성하는 과정에서 품질을 확보하고자 노력하여야 한다. 스마트 치안 시스템 모델 학습은 데이터셋의 특성으로 인해 어려움이 있다. 개발된 인공지능 시스템의 정확도를 높이려면 신뢰하는 오픈 소스 데이터셋을 확보하거나 서로 다른 데이터셋을 안정적으로 결합하여야 한다.
- 오픈 소스 데이터셋 활용 시 다수의 사용자가 데이터셋 활용 과정에서 발견한 오류가 추후 발견되며, 이로 인한 데이터셋 수정, 재구축으로 인해 데이터 버전이 변경된다. 만약 데이터 버전이 변경되면 인공지능 모델의 동작에 영향을 주므로, 이러한 문제에 대응하고 상황을 추적하도록 오픈 소스 데이터셋의 명확한 출처, 빌드 시간, 오픈 소스 데이터셋 버전 등의 정보를 기록하고 관리하여야 한다.

05-2a 신뢰할 수 있는 출처로부터 제공되는 데이터셋을 사용하였는가?

Yes No N/A
☐ ☐ ☐

- 학습 데이터를 직접 생산한다면, 데이터 획득 시 수집 출처(예: 크라우드 워커, 아웃소싱 기관)의 객관성 확보가 필요하다. 또한 수집 대상인 데이터의 출처를 살펴 향후 지식재산권이나 개인정보 문제가 발생하는지 선제적으로 확인하여야 한다.
- 인공지능 기반 프로젝트에는 모델을 학습하는 데 많은 양의 데이터가 필요하여 오픈 소스 데이터셋이 선호되므로, 해당 데이터셋이 신뢰할 만한 수준의 품질인지 고려하여야 한다.
- TTA 정보통신 단체표준 TTAK.KO-10.1339 표준은 지도 학습 계열의 인공지능 기술에 활용되는 데이터를 획득할 때 출처의 신뢰성 확보 측면에서 고려하여야 할 내용을 정리하였다.

참고 지도 학습 시 데이터 품질 관리 요구사항(TTAK. KO-10.1339)

지도 학습 인공지능 모델에서 사용되는 데이터 획득 시 출처의 신뢰성을 확보하고자 고려해야 할 내용을 발표함
데이터 출처의 신뢰성을 위해 다음과 같은 요소를 고려함
제3자가 데이터 획득 시 개인정보보호, 지식재산권, 사전 승인/허가 등과 관련하여 정식으로 절차를 밟고 문제없이 획득하였는지 여부
제공하는 데이터셋의 규모가 충분히 커, 데이터 사용자가 원하는 학습용 데이터를 제공하는 데 문제가 없는지 여부
예) 규모가 충분히 않을 때, 데이터 획득을 재차 시도하고자 할 때 수급에 문제가 생기는지 여부
해당 데이터가 지속적인 업데이트 및 추가 제공 등이 이루어지는지 여부
데이터와 함께 설계서의 내용이 명확히 제공되는지 여부
해당 데이터의 활용 건수 및 인용 건수가 많아 범용성이 높은지 여부

05-2b 오픈 소스 데이터셋을 활용하는 경우, 출처를 명시하였는가?

Yes ☐ No ☐ N/A ☐

- 스마트 치안 시스템 모델 학습에 오픈 소스 데이터셋 사용 시 학습 시점에는 발견되지 않은 오류나 편향된 결과가 나올 수 있다.
- 또한, 편향된 결과나 오류는 사회 인식 변화에 따라 윤리적 문제로 이어진다. 이러한 상황은 개인이나 인권 단체에 의해 법적 소송으로 이어진다. 오픈 소스 데이터셋을 활용할 때 일부 제공업체는 데이터셋 특성에 대해 사용자에게 주의를 주므로 의도하지 않은 데이터 편향이 발생할 위험이 있다.[150][151]
- 오픈 소스 데이터셋을 활용하여 인공지능 모델 구축 시, 과거·현재·미래 시점에 발생하는 데이터 편향의 원인을 파악하고자 확보된 데이터의 명확한 출처와 관련 정보를 명시하여 관리하여야 한다.

안전성

요구사항 06 데이터 견고성 확보를 위한 이상abnormal 데이터 점검

- 스마트 치안 시스템에서 사용되는 데이터 특성의 분포를 시각화하여 라벨링 작업 오류를 확인하고, 메타데이터의 스키마 통계 분석 기법을 이용하여 데이터의 이상치를 식별·처리한다. CCTV, 스캐닝 장치, 카메라 녹화를 기반으로 한 시각 데이터는 알고리즘 평가 단계에서 개별 또는 정합 방식으로 라벨링 관련 오릇값을 식별·처리한다. 또한, 스마트 치안 모델이 고의적이거나 의도하지 않은 현장의 공격 상황을 이해하고, 공격이나 이상 데이터에 의해 발생하는 예측 불가한 편향을 방지하고자 모델 학습 전 데이터 준비·관리 단계에서 대응 방안을 마련하는 것이 필요하다.
- 통계적인 방법과 기법을 사용하여 이상치를 식별하고 처리할 때, 해당 작업에 대한 전문가들은 상호 검증하여 관련 데이터를 제외할지 반영할지 결정하고 문서화한다.

06-1 이상 데이터의 식별 및 정상 여부를 점검하였는가?

Yes ☐ No ☐ N/A ☐

- 이상 데이터란 학습용 데이터를 구성하는 데이터셋 수집 및 가공 과정에서 발생하는 다양한 오류error와 일반적인 데이터의 범위에서 크게 벗어난 데이터 이상값outlier을 포괄한다. 학습 데이터의 수집 및 가공 과정에서 발생하는 이상 데이터는 데이터상의 노이즈, 학습 데이터는 데이터상의 노이즈, 학습 데이터 내의 편향, 잘못된 라벨링, 라벨링 누락 등 다양한 원인에 의해 생기며 이를 해결하지 않으면 인공지능 모델의 성능 및 견고성을 확보하기 어렵다.
- 스마트 치안 분야에서 주민 신과 등의 비정형 데이터unstructured data를 학습에 활용 시, 데이터 전처리 과정에서 이상 데이터를 식별하는 별도의 기법을 마련하여야 한다.
- 인식 시스템의 인공지능 모델 학습 데이터 내 여러 센서나 엣지 디바이스에서 얻은 데이터를 시각화하여 라벨링 작업 결과에 오류가 없는지 확인한다. 센서 또는 엣지 디바이스 측정이 포함된 하이브리드 시스템 개발 시, 상태 모니터링과 시각화 솔루션 두 가지 모두를 사용하는 것을 고려하여야 한다. 또한, 메타데이터의 스키마schema를 분석하여 데이터의 이상값을 식별하고 예방 조치하고자 이상값 유무를 확인하여야 한다.

06-1a 전체 학습용 데이터 분포를 시각화하여 발생 가능한 오류들을 확인하였는가?

Yes ☐ No ☐ N/A ☐

- 데이터 전처리 단계에서 중요한 포인트 중 하나는 데이터셋의 분포를 파악하는 것이다. 데이터셋의 전체 분포를 활용함으로써 데이터를 시각화하여 추가적인 입력 오류를 식별한다. 데이터 분포 시각화는 사용자/개발자들이 가능한 오류일 때 이상치를 쉽게 검토하게 도와준다.
- 시각화 기법을 사용하면 인간의 실수로 인한 오류를 확인하고 인공지능 모델 훈련의 데이터 탐색과 이해에 매우 용이하다. 이를 위해 아래와 같은 항목을 단독 또는 복수 조합하고 각 정보를 시각화하여 분포를 확인한다.
  - ✓ 성별 조건의 예시: 여성, 남성
  - ✓ 시간 조건의 예시: 주간, 야간, ~분
  - ✓ 날씨 조건의 예시: 맑음, 눈, 비 등
  - ✓ 피부색 조건의 예시: 백인, 흑인 등
  - ✓ 교통경찰/인력 대신 로보틱 솔루션에 대한 개입 조건 예시: 완전 대체, 일부 대체, 이동/회전 대체
  - ✓ 거짓말 탐지기의 행동 상태 예시: 화남, 긴장, 행복, 중립, 협조, 외향성, 개방성, 성실성, 예민함 등
  - ✓ 거짓말 탐지기에 사용된 단어의 예시: 빈말, 욕설 등
- 데이터의 특성에 따라 데이터 분포를 시각화하는 다양한 기법이 있다.
  - ✓ 분포 그래프(전체 데이터의 평탄성, 평균, 분산, 편차 등을 사용하여 데이터 분포를 시각화)
  - ✓ 범주형 그래프(범주형 데이터를 시각화)
  - ✓ 행렬 그래프(이차원 행렬 데이터를 시각화)

데이터 분포 시각화 기법 예시

시각화 기법의 분류	설명
히스토그램 차트	변수의 히스토그램 형태로 데이터를 시각화한다.
커널 밀도 추정 그림	하나 또는 두 개의 변수에 대한 밀도 추정 그래프 형식으로 데이터를 시각화한다.
경험적 누적 분포 함수 그래프	전체 데이터의 누적 분포를 시각화한다.
루그 다이어그램	주변 분포도를 표시하고자 x/y축에 눈금을 표시하는 그래프를 작성하며, 이는 주로 다른 그래프를 보완하고자 함께 사용한다.
산점도[46]	각 데이터를 x/y축상의 점으로 나타내어 전체 데이터셋의 각 데이터를 시각화한다.
패시팅[59]	한 변수를 나타내는 열과 다른 변수를 나타내는 행을 사용하여 최대 두 가지 변수로 데이터를 표현한다.
시계열 그래프[59]	이 그래프 유형은 데이터셋의 시계열 표시에 초점을 맞춘다. 시계열 그래프는 x축에 시간을, y축에는 관련 특성의 결과 또는 측정값을 나타낸다.
박스 플롯과 릿지 플롯을 사용한 다중 모드 분포[59]	만약 데이터셋의 여러 모드/입력을 보여 주어야 한다면 이러한 플롯 유형이 선택된다. 주로 히스토그램 플롯과 함께 사용한다.

참고 데이터 분포를 확인하는 통계 분석 도구 [47]

일반적으로 분포하는 데이터[48]	정상적으로 분포하지 않은 데이터[49]	요구된 분포
T-test	Kruskal-Wallis test Mood's median test Kruskal-Walls test	Any
ANOVA	Mood's median test Kruskal-Walls test	Any
Paired t-test	One-sample sign test	Any
Bartlett's test Bartlett's test	Levene's test	Any
Individuals control chart	Run Chart	Any
Cp/Cpkanalysis	Cp/Cpkanalysis	Weibull, Log-normal, Largest extreme value, Poisson, Exponential, Binomial

데이터 분포를 확인하는 데이터 분석은 다음 도구들을 활용한다:[50]

Microsoft Power BI
SAP BusinessObjects
Sisense
TIBCO Spotfire
Thoughtspot
Qlik
SAS Business Intelligence
Tableau
Google Data Studio
Redash

06-1b 학습 데이터 이상값 식별 기법을 적용하였는가?

Yes ☐ No ☐ N/A ☐

- 고려 사항 데이터 이상값을 식별할 때는 데이터 전체에 통계적 기법을 적용하여 전체 데이터셋을 고려하였을 때 차별화되는 데이터 포인트를 찾아내는 방법이 주로 활용된다. 이와 관련된 대표적인 기법은 Z-점수, 사분위수 범위 등이 있는데 인식 작업을 하는 데이터셋에 적용할 때는 아래와 같은 특성을 정의하고 별도로 적용하여야 한다.
  - ✓ 얼굴 특성의 예: 너비, 높이, 형태 등
  - ✓ 얼굴 움직임의 예: 눈 깜빡임, 눈썹 올리기, 코주름 짓기, 입꼬리 올리기 등
  - ✓ 신체 부위의 예: 머리, 팔, 다리 등
  - ✓ 연령 차이의 예: 영아, 어린이, 성인
  - ✓ 성별의 예: 여성, 남성
  - ✓ 행동 분석의 예: 기쁨, 화남, 초조함 등
  - ✓ 물체의 예: 가방, 총, 나이프 등
  - ✓ 환경 조건의 예: 맑음, 비 오는 등
  - ✓ 레이더 유형의 예: 합성 개구리 레이더, 고주파 표면파 레이더, 클래식 레이더 등
- 일관된 분석 결과를 도출하려면 먼저 이상치 검색을 수행하여야 하며, 이상치가 포함된 데이터 분석은 모델 오류와 편향된 결과로 이어진다. 이상치는 합리적인 이상치와 비합리적인 이상치로 나뉜다. 합리적인 이상치는 정확하게 측정되지만 다른 데이터와 완전히 다른 추세나 특성을 보이는 이상치를 의미하며, 비합리적인 이상치는 입력 오류 등 데이터 오염에 의해 발생하는 이상치를 의미한다.
- 이상치 검색 시 마스킹 효과와 침몰 효과에 주의하여야 한다. 마스킹 효과는 일부 극단 값들에 의해 이상치로 분류되어야 할 측정값들이 정상 범주의 값으로 나타나는 현상이고, 침몰 효과는 정상 범주의 측정값들이 이상치에 가까운 동일한 이상치 값으로 나타나는 현상이다. 마스킹 효과와 침몰 효과를 해결하려면 강건한 중심값의 중심점과 이상치의 측정에 적게 영향을 받는 공분산 행렬을 사용하여야 한다.
- 또한, 메타데이터에도 이상치 식별 기법을 적용한다. 따라서 데이터의 스키마를 추론하고자 전체 메타데이터를 분석하고, 데이터셋의 통계와 비교하여 이상치를 식별한다[51].

접근 방법에 따른 이상치 탐지 방법의 분류

접근	이상치 탐지 방법의 분류
자료의 크기	작은 샘플, 대표 복사본
데이터의 차원	일차원, 이차원, 다차원
변수의 수	단변량, 이변량, 다변량
대상 변수의 존재 여부	가이드된 방법, 비지도 학습 방법
통계적 방법	인자 모형, 비인자 모형, 준인자 모형

데이터 구조에 따른 이상치 검색 방법의 분류

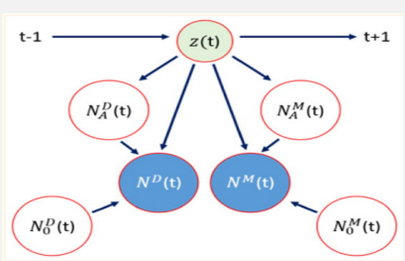
자료의 구조	이상치 검색 방법의 분류
단변량 데이터	시계열 데이터가 아니라 1의 변수들이 다수 있을 때
	• 표준화된 점수를 사용 • 수정된 표준화된 점수를 사용 • 통계적 가설 검정을 실시 • 사분위 범위를 계산 • 수정된 사분위 범위를 계산 • 준-사분위 범위를 사용
다변량 데이터	시계열 데이터가 아니며 변수의 수가 두 개 이상일 때
	• 회귀 진단에서 이상치를 탐색 • Mahalanobis Distance를 사용 • LOF(Local Outlier Factor)를 사용 • iForest(isolation Forest)를 사용
시계열 데이터	시계열 데이터일 때
	• Shewhart를 사용 • CUSUM을 사용 • 지수 가중 이동 평균을 사용 • Hidiroglou-Berthelot 방법을 사용 • 퓨전 기술을 사용

참고 스마트 홈 치안 시스템에서 이상값 탐지 검사 사용 사례[52]

이 논문에서 연구원들은 거주자에게 보안을 제공하는 데 스마트 홈의 역할을 탐구한다. 잠재적 위협을 식별하고 행동 이상을 탐지하는 이상 탐지 기술의 사용을 강조한다. 이 백서에서는 보안 위험으로 알려진 대상 거주자 상태 또는 주택 상태의 중요성에 관해 설명한다. 그런 다음 데이터에서 예상되는 행동에서 벗어난 패턴을 찾는 방법으로 이상 행위 탐지를 소개한다. 이 백서에서는 클러스터 센터까지의 거리를 기준으로 데이터 포인트를 클러스터링하거나 정규 분포 데이터의 z 점수를 계산하는 등 이상 징후 또는 이상값을 탐지하는 다양한 기법에 관해 설명한다.

가정 기반 보안의 이상값을 감지하여 리더는 몇 가지 핵심 사항을 식별할한다. 첫째, 보안 위험이 있다고 알려진 특정 거주자 상태 또는 주택 상태를 타게팅하는 것이 유용하다. 클러스터 센터까지의 거리를 기준으로 데이터 포인트를 클러스터링하거나 z 점수를 계산하는 등의 기법을 사용하면 이상값을 식별하는 데 도움이 된다. 패턴 데이터 내에서 이상값을 탐지하는 아이디어를 요약하면 다음과 같다:

- 위치 기반 이상 징후 탐지는 센서를 사용하여 집안에서 거주자의 위치를 모니터링하고 잠재적인 이상 징후 식별을 포함한다.
- 활동 기반 이상 징후 탐지는 계층적 클러스터링 또는 숨겨진 상태 조건부 랜덤 필드 등의 방법을 사용하여 개별 활동을 분석하고 각 활동 내에서 이상 징후를 탐지하는 데 중점을 둔다.



〈출처: Dahmen, Jessamyn, Diane J. Cook, Xiaobo Wang, Wang Honglei. “스마트 치안 홈: 보안 위협을 감지, 평가 및 대응하는 스마트 홈 기술 조사”, 2017[66]〉

06-2 데이터 공격에 대한 방어 수단을 강구하였는가?

Yes ☐ No ☐ N/A ☐

- 스마트 치안 알고리즘 중 인지 모델은 데이터 유형(예: 이미지 또는 센서 데이터 등)을 사용하는 공격에 취약하므로 대응 방안을 검토하고 적용하여야 한다.
- 스마트 치안 시스템은 특정 고유 정보(성별, 범죄 기록, 출신 지역, 인종, 국적, 외모 등) 사용 시 공격에 취약하므로 방어 방안을 마련하여야 한다.

참고 데이터 공격 및 방어 기술의 예시[152][153]

공격 기술 분류		공격 기법	대표적인 방어 기술
데이터 중독 공격	그라데이션 기반 GAN 기반 백도어 클린 레이블 포이즈닝	• AI 서비스는 일반적으로 모델 배포 후 수집된 새로운 데이터를 사용하여 입력 데이터 분포의 변화에 적응하도록 재학습된다(예: 침입 탐지 시스템). 이때 공격자는 신중하게 조작된 변조 데이터를 삽입하여 학습 데이터를 오염시켜 서비스의 정상적인 기능을 손상한다. ◦ 유형: FGSM, 홉스킵 점프 공격	• 데이터 살균 • 강력한 교육 • 인증된 방어 • 적대적 훈련 • 그러데이션 마스킹(증류) • 피처 스퀴징
회피 공격		• 공격자는 학습 모델이 입력을 올바르게 식별할 수 없도록 기존 입력 데이터에 미묘한 노이즈 차이를 추가하여 조작된 입력 데이터를 생성한다. 이러한 변화는 사람에게는 눈에 잘 띄지 않지만, 딥러닝 모델의 출력에 상당한 영향을 미친다. ◦ 유형: 의사결정 트리 만들기, 칼리니 & 와그너, 제로 차수 최적화	

* 언급된 기술은 05-2a에 자세히 설명되어 있다.

06-2a 데이터 최적화를 통한 방어 대책을 마련하였는가?

Yes No N/A
☐ ☐ ☐

- 대부분의 인공지능 모델은 여전히 보안 위협에 취약하다. 공격자는 모델 자체를 직접 공격할 수도 있지만, 모델 자체에 접근하기보다 데이터에 접근하기가 더 쉬워 훈련 데이터를 공격하기도 한다. 따라서, 적대적 공격을 방어하고 인공지능 서비스의 견고성을 높이는 다양한 방어 기법이 존재한다.
- 현재까지 완벽한 방어 기법은 없지만, 데이터 설계 및 모델 학습 단계에서의 회피 공격과 중독 공격에 방어하는 대표적 기법으로는 적대적 학습adversarial training, Gradient Masking 및 Feature Squeezing 등이 있다.

동영상에 대한 적대적 공격의 방어 기법 예시

방어 기술의 분류	방어 기술 내용
적대적 학습	• 위의 표에 언급된 이 방어 기법은 동영상에 대한 적대적인 공격에도 과도하게 사용된다.
적대적 예시 탐지	• 탐지 체계는 종종 탐지기를 로컬로 훈련하거나 적대적인 샘플의 특징 특성을 사용하여 입력이 손상하였는지 여부를 판단한다. 이 방어 기법은 프레임의 시간적 일관성을 기반으로 한다. 또 다른 접근 방식은 공간 정보를 사용하여 의미적 세분화에 대한 공격적 예시를 탐지하는 것이다. 이러한 모든 접근 방식은 공격 샘플을 추출하여 알고리즘 내의 도메인(또는 다른 도메인)에서 사용함을 기반으로 한다. 설계된 알고리즘은 원래 분류기의 결정 경계에 따라 입력 공간을 하위 공간으로 나눈 다음 하위 공간에서 모순되는 인스턴스를 분류하는 것이한다[53].
인증된 견고성	• 인증된 체계는 이론적 보장을 통해 규범에 구속된 적대적 섭동을 방어하고자 제안되었다. 이 접근 방식은 $\ell 2$ 규범 기반 접근 방식이며 모델의 인증된 $\ell 2$ 견고성을 기반으로 한다. 특히, 인증된 $\ell 2$ 견고성에 대한 경쟁 접근법이 적용되는 소규모 데이터셋에서 평활화는 더 높은 인증된 정확도를 제공한다[54].

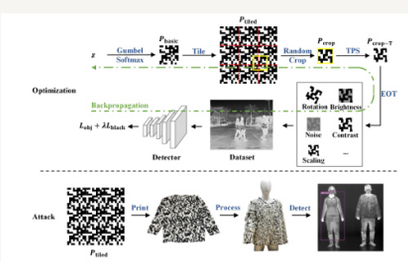
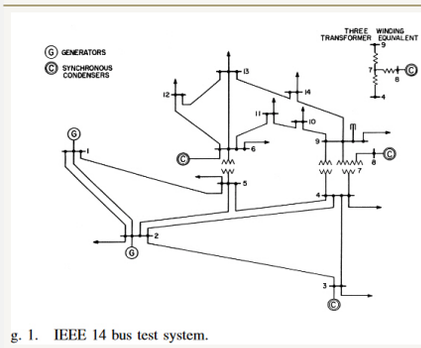
데이터 중독 공격에 대한 방어 기법 예시

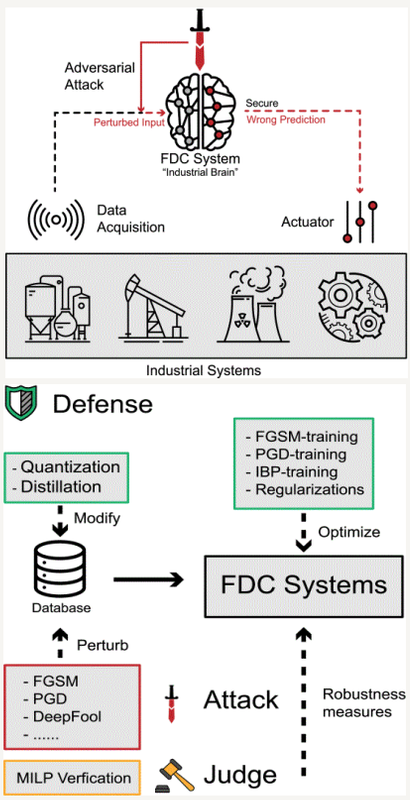
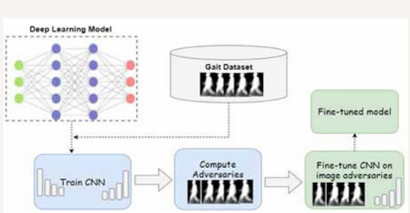
방어 기술의 분류	방어 기술 내용
데이터 살균	• 흔히 ROIN(부정적 영향에 대한 거부) 방법이라고도 한다. 특정 훈련 데이터가 모델에 부정적인 영향을 미칠 때 해당 부분을 훈련 세트에서 제거하는 기법이다. 과적합은 데이터셋가 특징 수보다 작을 때 발생하는 경향이 있다. 이 방법은 프로세스 후 모델을 재학습하고 탐지 처리 중에 그라디언트 및 H 행렬을 사용하여야 한다.
강력한 트레이닝	• 이 방법은 특징 가정에 제약을 가한다. 강력한 저차 행렬 접근 방식과 강력한 주성분 회귀를 개선하여 강력한 방어 성능이 목표인 방법이다. 선형 회귀 모델의 강력하게 학습하고자 각각의 반복마다 다른 잔차의 하위 집합을 계산하고자 가지치기 손실 함수를 사용하는 TRIM 방법은 잘 알려진 방어 기법이다.
인증된 방어 기능	• 이 기법에 대해 가장 잘 알려진 연구는 이 목적 때문에 만들어진 프레임워크[55]이다. 이 프레임워크는 특정 방어에 대한 전체 공격 영역을 검사하고자 설계되었으며, 이상 징후 제외 및 경험적 위험 최소화를 채택한다.

참고 적대적 공격에 대한 방어 연구 사례

스마트 치안 시스템은 일부 데이터 유형에 대한 공격이 집중적으로 이루어진다. 이러한 데이터는 비디오, 센서/엣지 디바이스 데이터이다. 스마트 치안 시스템은 의도된 용도로 사용되므로 지속해서 공격에 노출되어 사소한 보안 위반이 발생하고 그로 인한 오류는 치명적이다. 따라서 개발자는 방어 메커니즘을 선택할 때 모델에 대해 가능한 모든 공격을 고려하여야 한다.

동영상에 대한 적대적 공격과 관련한 방어 연구는 거의 없다. 다음은 동영상에 대한 적대적 공격과 관련한 연구를 소개한다.

연구	방어 기술 내용
 〈출처: 적외선 투명 의류: 실제 세계에서 다양한 각도에서 적외선 탐지기에서 숨기, 2022[57]〉	• 감시 시스템의 성공으로 인해 보안 위협을 매우 쉽게 탐지한다. 이러한 상황을 방지하고자 시스템에 대한 공격이 빈번하게 조직화한다. 범죄자들은 보안 감시를 회피하고자 물리적 공격 인스턴스를 사용하여 잠재적인 보안 위협을 일으킨다. • 공격자들은 이러한 시스템의 탐지 프로세스를 피하고자 패턴이 다양한 의복과 외부 신체적 특징에 집중한다. 옷의 패턴과 착용한 액세서리를 조작하여 시스템이 판단할 이미지를 조작한다. • 이 연구에서 연구진은 QR 코드를 패턴처럼 사용하여 적외선이 보이지 않는 옷을 개발하였다. 이 부착 기술로 공격자는 적외선 카메라와 센서를 사용하는 감시 시스템을 조작한다. • 방어 메커니즘으로 의미 없는 QR 코드 등 패턴을 학습 데이터셋에 추가하고 트랜스포머 기반 모델을 사용하여 개발된 모델에 대한 이러한 유형의 물리적 공격을 극복하는 것을 고려하여야 한다.
 g. 1. IEEE 14 bus test system. 〈출처: 스마트 그리드 상태 추정에 대한 악성 데이터 공격: 공격 전략 및 대응 방안, 2010[58]〉	• 이 연구에서는 스마트 그리드 상태 추정에 대한 화이트 박스 및 표적화되지 않은 악성 데이터 공격을 조사한다. 공격이 없는 상태와 공격이 탐지된 상태를 나타내는 두 가지 접근 방식을 기반으로 한 방어 설계를 제안한다. 먼저 시스템 입력 데이터를 획득한 후 통계 결과를 계산하고 임곗값과 비교하여 현재 입력이 적대적인 샘플인지 여부를 판단한다.

연구	방어 기술 내용
 〈출처: 공격과 방어: 데이터 기반 FDC 시스템의 적대적 보안, 2023년[59]〉	• 이 연구에서는 데이터 기반 결함 탐지 및 분류 시스템(FDC)의 보안에 초점을 맞춘다. 특히 어떤 유형의 적대적 공격과 방어 방법이 결함 분류기에 사용되었는지에 대해 연구한다. 또한, 산업적 공격 보안 벤치마크에 대해서도 다룬다. 연구진은 표형식 데이터에 대한 적대적 벤치마크를 제안하였다. 보안을 향상하려면 강력한 훈련, 기울기 마스킹, 입력 기울기에 대한 정규화를 통해 섭동의 안정성을 가져와야 한다. 이들은 주로 FDC와 데이터 기반 산업 시스템의 적대적 보안에 초점을 맞춘다.
 〈출처: 보편적인 적대적 공격에 대비한 보행 인식 기반 스마트 감시 시스템 보안, 2023 [60]〉	• 이 연구에서는 사람의 보행 동작에 초점을 맞춘 스마트 감시 시스템에 대한 실험을 수행하였다. • 저자는 GEI 이미지 데이터(CASIA 보행 데이터셋)를 사용하였다. • GEI 이미지에 대한 적대적 공격을 조사하고 부정확한 라벨링으로 인해 시스템 장애가 발생하는 것을 발견하였다. • 연구진은 GEI(감시 시스템) 이미지에 대한 이러한 적대적 공격에 대한 방어 메커니즘으로 모델의 적대적 재학습에 집중하였다.

다양성 존중 | 책임성 | 투명성

요구사항 07 수집 및 가공된 학습 데이터의 편향 제거

- 인간의 배경, 행동, 성별, 인종, 신체적 차이 등 데이터의 특성상 오픈 소스 공급 업체에서 편견 없는 데이터를 수집하거나 얻기 어렵다. 그중 인종적 편견은 폭력, 예방, 국경을 통제하는 거짓말 탐지기 등 스마트 치안 시스템에서 중요한 문제로 대두된다.
- 데이터 수집 및 처리 과정에서 발생하는 편향을 확인하여야 한다. 이는 데이터 수집 중에 편향 발생 가능성이 높고 모델의 추론 결과가 개인의 생명 및 안전에 직접적인 영향을 미치는 스마트 치안 시스템에서 특히 중요하다. 또한, 모델 학습 및 라벨링에 사용할 특성을 선택할 때 신중히 처리하여야 한다.
- 이미 편향성 검토가 완료된 데이터를 사용하거나, 사용자 프로필 특성상 다양성이 넓어 현실적으로 모든 사용자 그룹의 데이터를 검증하기 어려울 때는 샘플링 기법을 사용하여 데이터 검증을 고려하여야 한다.

07-1 데이터 수집 시, 인적·물리적 요인으로 인한 편향 완화 방안을 마련하였는가?

Yes ☐ No ☐ N/A ☐

- 데이터 편향은 데이터셋을 직접 수집할 때 인적·물리적 요인으로 인해 다양한 데이터가 수집되지 않아 데이터 편향이 발생한다. 데이터의 편향은 인공지능 알고리즘의 동작 성능에 문제를 일으키고, 시스템의 오동작으로 이어지므로 편향을 완화하는 노력이 필요하다.
- 구형 기술이나 레이더 장치 오작동 등 데이터 획득 시 사용된 장비나 적외선 카메라의 낮 기록 사용 등 환경적 요인으로 인해 물리적 편향이 발생한다. 따라서 특정 장비나 환경에서 수집한 데이터셋으로 학습시킨 알고리즘을 다른 장비나 환경에서 수집한 영상, 센서, 레이더, 합성 레이더/선박 데이터셋에 적용할 때 오류가 발생한다.
- 또한 수집 환경 및 제약 조건 등으로 인해 데이터 편향이 발생하므로 다양성을 확보하고 다양한 인적·물리적 편향 요소를 제거하고자 폭넓은 데이터를 사용하는 것이 바람직하다.

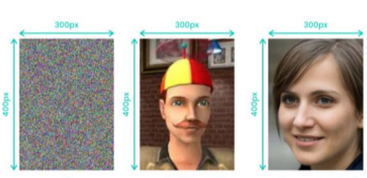
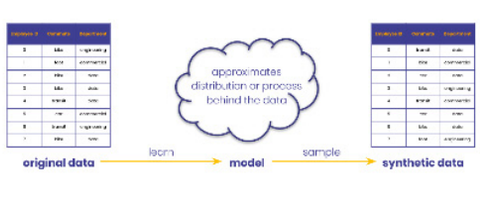
07-1a 인적 편향을 제거하기 위한 절차적, 기술적 수단을 적용하였는가?

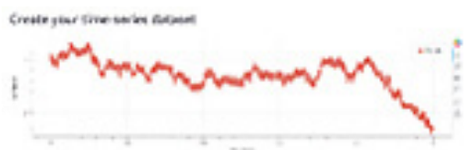
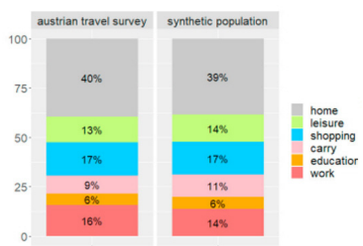
Yes No N/A
☐ ☐ ☐

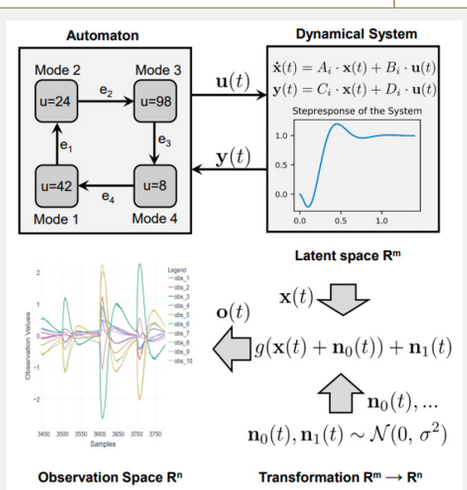
- 스마트 치안 분야 인공지능 시스템의 개발 및 고도화 시, 특정 예외를 제외하고 개인정보보호위원회 가이드라인에 따라 개인정보를 수집, 사용 또는 공유하기 전에 개인에게 통지하고 명시적인 동의를 얻어야 한다.
- GDPR 제5조와 제18조 및 제22조의 제한 사항(섹션 04-1c 및 [61] 참조)을 고려하면 이러한 상황은 적절하고 편향되지 않은 데이터셋을 찾는 것을 더욱 어렵게 한다. 이에 따라 수집된 데이터를 사용하여 목적에 적합한 데이터셋을 구축하여야 한다. 그러나 수집된 데이터 자체에 희귀 데이터나 특정 그룹, 지역, 인종, 성별 등에 편향된 데이터 불균형 등 인적 편향 문제가 발생한다.
- 인적 편향을 줄이고자 데이터 수집 작업의 가이드라인을 마련하고, 다양한 데이터 수집 작업자를 모집하여 특정 배경 및 성향 등을 배제하며, 수집 결과에 대한 검수자를 충분히 확보하여야 한다. 또한, 편향된 실제 데이터를 기반으로 합성 데이터를 생성하여 데이터의 다양성을 보장하고 편향성을 완화하여야 한다.

* 공공의 안전에 사용되는 인공지능과 같은 특정 상황은 예외이다. 개인정보 보호법 제15조에 명시된 것처럼, 주로 공공기관 등의 법적 의무 및 필수 업무와 관련될 때는 비동의 이용이 허용되는 것으로 간주한다.

참고 합성 데이터 생성 기법 사용 사례[63]

방법	방법 개요	
규칙 기반 및 지능형 규칙 기반	 〈출처: Manuel Pasieka, "합성 데이터 생성 방법과 합성 데이터 유형 비교", 2022[64]〉	• 필수 테스트 케이스에 대한 전체 커버리지 제공 • 프로덕션 데이터에 액세스하지 않고도 대량의 데이터를 빠르게 생성한다. • 상황에 따라 실제 데이터의 통계적 분포를 나타내지만, 항상 실제 데이터를 나타내는 것은 아니다.
통계 모델 기반	 〈출처: 크리스토프 웨마이어, "합성 데이터는 어떻게 생성하나요?", 2021[65]〉	• 실제 데이터가 충분하지 않을 때 무작위 샘플 분포 생성 • 실제 데이터를 대표함 • 하지만 실제 데이터 샘플의 종류와 크기에 따라 제한됨 • 테스트 목적에 필요한 데이터를 나타내지 않으며 프로덕션 데이터에 대한 액세스가 필요함

방법	방법 개요	
GAN	 〈Source: ML4ITS, “Synthetic Time-Series,”[88]〉	• 입력 및 출력 데이터가 모두 동일하게 유지되도록 보장한다. • 사실적인 합성 데이터 생성 • 학습에 대량의 실제 데이터가 필요하다. • 관계 무결성을 유지하기에는 너무 복잡하며 테스트를 완료하는 데 필요한 데이터를 제공하지 못한다.
Agent-based modeling	 〈Source: Felbermair et al., “Generating synthetic population with activity chains as agent-based model input using statistical raster census data” 2020[66]〉	• 실제 관찰된 행동을 기반으로 무작위 데이터 생성 • 실제 데이터를 샘플링하여 다양한 합성 데이터 생성 • 정확하게 표현하지 못함
Diffusion-based modeling	 〈Source: Voetman et al., “The Big Data Myth: Using Diffusion Models for Dataset Generation to Train Deep Detection Models” 2023[67]〉	• 사전 학습된 안정적인 확산 모델을 미세 조정하여 합성 데이터셋 생성 • 데이터셋을 생성하는 확산 모델을 사용하여 특히 스마트 치안 애플리케이션의 감시 및 탐지 목적으로 사용되는 심층 탐지 모델 훈련하기 • 그러나 관련 연구는 심층 객체 감지 모델에 필요한 방대한 양의 훈련 데이터에 대한 문제를 해결

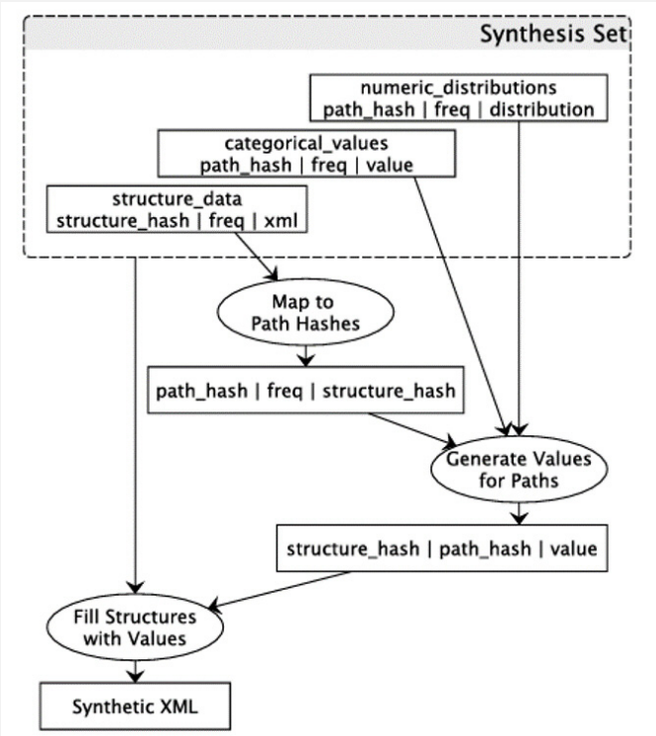


Exemplary comparison of MSE to MLE in the area of a change in the mode of the automation for a single observation variable.

	σ (MLE)	MSE
Mean value dataset	0.057	$3.88 \cdot 10^{-3}$
10 ms before mode change	0.099	$0.487 \cdot 10^{-3}$
5 ms after mode change	0.2	$1.2 \cdot 10^{-3}$

출처: Zimmering, B.; et al. Generating Artificial Sensor Data for the Comparison of Unsupervised Machine Learning Methods. Sensors 2021, 21, 2397. https://doi.org/10.3390/s21072397 [68]

Shuffle and output characteristics of the 2-phase MapReduce synthetic generator



Input		Phase 1		Phase 2	
Months	Documents	Shuffle (gzip)	Output	Shuffle (gzip)	Output
1	1.61 million	371 GB (73 GB)	198 GB	487 GB (69 GB)	365 GB
2	3.09 million	715 GB (141 GB)	394 GB	975 GB (143 GB)	733 GB
4	6.35 million	1,470 GB (290 GB)	810 GB	2,004 GB (294 GB)	1,506 GB
8	13.72 million	3,177 GB (627 GB)	1,750 GB	4,331 GB (635 GB)	3,255 GB

출처: Anderson, Jason W., et al. "Synthetic data generation for the internet of things." In 2014 IEEE International Conference on Big Data (Big Data), pp. 171–176. IEEE, 2014[69].

07-1b 데이터의 다양성 확보를 위해 여러 수집 장치를 활용하였는가?

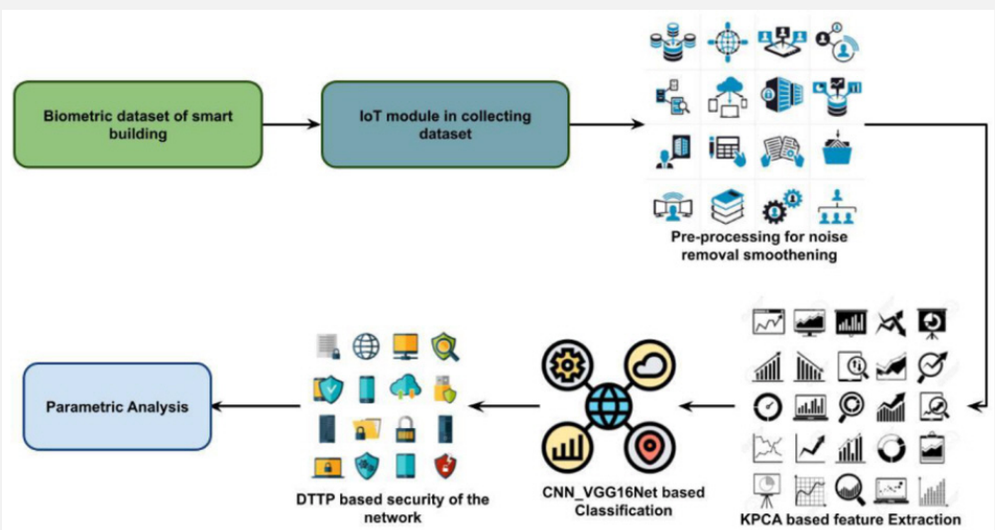
Yes ☐ No ☐ N/A ☐

- 스마트 치안 분야 시스템 개발 및 고도화하고자 데이터를 직접 수집 시, 수집 환경 및 제약 조건 등으로 인해 다양한 데이터를 확보하기 어려우므로 이기종 장치 데이터를 수집하여 다양성을 확보하여야 한다.
- 데이터의 수집 과정과 환경이 변화하므로 수집 후 데이터를 활용하려면 데이터 정합성을 유지하는 데이터 클렌징과 검수가 충분히 이루어져야 한다.

참고 더 안전한 인공지능 모델 구축 예시[70]

고도의 개인정보에 접근하거나 개인 또는 생체 데이터를 수집하는 시스템 개발 시 스마트 치안 시스템의 보안을 강화하고자 더욱 안전한 인공지능 모델을 구축하고, 추가적인 보안 조치를 하고, 이러한 데이터와 모델 출력/추론 결과 및 저장된 데이터를 보호하고자 특별히 다양한 보안 기법을 사용하는 것을 고려하여야 한다.

인공지능은 네트워크에서 이상 징후를 사전에 예측적으로 탐지하며, 안전한 데이터 전송을 하도록 노력하는 것도 현명한 결정이다. 예를 들어, 이 연구에서는 분류 및 특징 추출을 통한 안전한 데이터 전송과 생체 인증으로 침입자를 탐지하고자 개발된 시스템을 예로 든다. 여기서 침입자는 IoT 기반 스마트 빌딩 시스템의 생체 인식 데이터베이스를 수집하여 탐지한다. 개발된 모델의 출력과 시스템에서 처리되는 데이터 및 전체 네트워크는 결정적 신뢰 전송 프로토콜(DTTP)을 사용하여 보안을 유지한다. 제공되는 시스템의 아키텍처는 다음과 같다:



〈출처: Annadurai, C. et al. "Biometric Authentication-Based Intrusion Detection Using Artificial Intelligence Internet of Things in Smart City", 2022〉

07-2 학습에 사용되는 특성feature을 분석하고 선정 기준을 마련하였는가?

Yes No N/A
☐ ☐ ☐

- 보호변수 편향을 완화하려면 차별을 일으키는 민감한 특성들을 사전에 파악하는 것이 중요하며, 이에 데이터의 특성들을 분석하고, 해당 특성을 학습에 사용할지 그 선정 기준을 수립하는 것이 바람직하다. 아래와 같은 민감한 특성들은 사회적 물의를 일으키거나 인공지능 의사 결정에 차별을 일으킨다. 이와 같은 특성들은 데이터 학습 시 반영되지 않아야 하는 특성으로 선정하고, 이에 따라 발생하는 편향을 완화하여야 한다.

  ✓ 연령, 성별, 인종, 민족, 범죄 이력 및 사회적 출신, 언어, 신체적 한계 등

- 비디오 감시, 스마트 홈, 폭력 감지, 거짓말 탐지기 등 스마트 치안 시스템에는 개인정보 보호 문제가 수반된다. 영상 녹화의 잠재적인 사용 및 오용은 문제를 제기한다. 유럽의 일반 데이터 보호 규정(GDPR), 미국의 연방 정보 보안 관리법(FISMA), 한국의 개인정보 보호법 등 법적 안전장치는 데이터 사용에 대한 책임을 사용자에게 지게 하고, 특히 상업적 목적으로 수집된 데이터에 대한 책임을 보장한다[71].

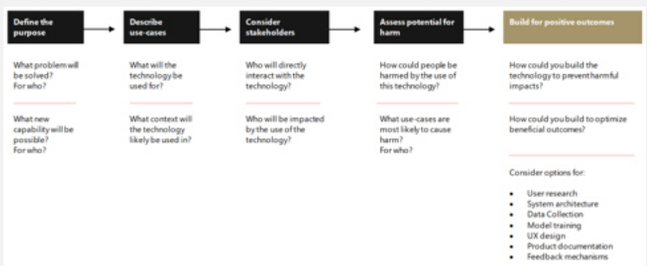
07-2a 보호변수 선정 시 충분한 분석을 수행하였는가?

Yes No N/A
☐ ☐ ☐

- 보호변수 선정 시 충분히 분석을 진행하지 않으면 모델 성능이 저하된다. 따라서 모델 추론 결과에 영향을 미치는 특성을 식별할 때 주어진 데이터셋에서 데이터의 일부분을 변경하거나 가중치를 재배치하면서 모델의 결과가 어떻게 변하는지 관찰하고 분석하여야 한다.
- 특히, 스마트 치안 분야 인공지능 알고리즘 개발 시 학습 데이터셋에 민감한 특성(예: 나이, 성별, 인종, 범죄 이력, 성격 특성, 생활 습관, 방문 국가, 보행 패턴, 종교, 방문 회의장, 집안 물품 등)이 기록된다면 보호변수를 설정하고, 보호변수가 편향된 결과에 얼마나 영향을 미치는지 분석하여 성능 결과가 어떻게 변하는지 비교 분석하여야 한다.
- 데이터셋의 편향성과 공정성을 식별하는 오픈 소스 도구(예: ML 공정성 짐, IBM의 AI 360 공정성, Aequitas, FairLearn 및 Google의 Facets, IBM AI 360 설명 가능성)와 데이터 변화에 따른 추론 결과의 추세를 시각화하는 도구(예: Google What If Tool)를 사용하여 설정한 보호 변수가 편향된 결과에 얼마나 영향을 미치고 성능이 어떻게 변하는지 알 수 있다.

참고 Microsoft 피해 모델링의 이해관계자/잠재적 사용자 프로세스 표[72]

개인정보 보호(보호변수) 및 보안 시 Microsoft는 식별 격차를 예측하는 데 도움이 되는 해로움 모델링이라는 디자인 모델을 설정한다. 또한 이 표를 평가하여 잠재적인 개인의 다양성을 고려하여 편견을 완화하는 차별 기준을 설정한다.



〈출처:https://learn.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/〉

07-2b 편향을 발생시킬 수 있는 특성의 영향력을 완화하였는가?

Yes ☐ No ☐ N/A ☐

- 편향을 완화하려면 데이터의 출처, 시스템 훈련에 사용된 방법, 잠재적 한계를 명확하게 문서화하여 투명성을 높이고 편향을 쉽게 식별하여야 한다. 또한, 다양한 관점을 고려하도록 개발 및 평가 프로세스에 이해관계자를 참여시켜야 한다.
- 교통 단속이나 위험한 환경에서 로봇 솔루션 사용 시, 환경적 요인, 직업적 위험, 인적 요인, 장비 요인 등의 영향을 받아 편향을 발생시킨다. 편향을 완화하는 간단한 접근법으로는 편향을 발생시키는 특성을 배제하는 특성 선택 기법feature selection을 적용하거나, 필터filter, 래퍼wrapper, 임베디드embedded 방법 등이 있다. 이러한 방법들은 데이터 내 특성들의 통계적 상관관계를 분석하여 높은 상관계수를 갖는 특성을 사용하거나, 특성 일부에 대해 좋은 성능을 갖는 부분 집합subset을 활용한다.
- 편향과 관련된 특성을 제거할 때, 다른 편향을 발생시키거나 강화하여 모든 경우에 효과적인 방법은 아니다. 따라서 편향을 완화하는 다양한 기법(예: 가중치 재지정, 라벨링 재지정, 변수 블라인딩, 샘플링)을 고려하여야 한다.

참고 스마트 치안 시스템의 편향성 해결 및 완화 사례 예시

얼굴 인식은, 훈련 데이터에 다양한 표현을 사용하면 피부색이 더 어둡거나 덜 일반적인 얼굴 특징을 가진 사람을 식별하는 편견을 해결하는 데 도움이 된다.

음성 인식 시스템은, 남성과 여성의 목소리가 동등하게 표현된 균형 잡힌 데이터셋을 통합하면 성별 편견이 완화된다.

정기적인 테스트와 평가를 통해 다양한 인구 통계 그룹에서 시스템 정확도의 불일치나 불균형을 파악할 수 있다. 편향이 발견되면 시스템을 재보정하고 공정성을 개선하는 적절한 조치를 할 수 있다.

특성 영향 완화에 사용하는 전략

데이터셋의 특성	완화 방법
데이터 수집	다양한 그룹의 사람들에게 데이터를 수집하여 시스템 학습에 사용되는 데이터셋이 모집단을 대표하도록 한다.
다양한 훈련 데이터	스마트 치안 시스템에서 편향성이 발생하는 주요 원인 중 하나는 편향된 학습 데이터이다. 이를 완화하려면 시스템을 개발하는 데 사용되는 학습 데이터가 다양하고 모집단을 대표하도록 하는 것이 중요하다. 여기에는 다양한 소스에서 데이터를 수집하고(이때 합성 데이터와 결합한 데이터셋) 성별, 인종, 연령 및 기타 관련 특성 측면에서 데이터의 균형을 맞추는 것이 포함된다.
정기 감사	스마트 치안 시스템은 편향되지 않도록 정기적으로 감사를 받아야 한다. 여기에는 시스템 훈련에 사용된 데이터를 분석하고 시스템이 내린 결정을 실제 결과와 비교하여 편향성 패턴을 파악하는 작업이 포함된다.

- 특성 선택 기법을 통해서 잠재된 편향을 완화하고 모델 성능을 향상하나, 지나치면 과적합overfitting 문제 혹은 오히려 편향의 원인이 된다.
- 특히, 모든 데이터에서 특성 선택 시행 시, 교차 검증에서 동일한 특성을 사용한 특성을 사용하므로 편향을 야기한다. 따라서 과도한 특성 선택 및 배제를 방지하는 점검이 필요하다.

과도한 특성 선택 및 배제 방지 점검표

체크리스트	조치
도메인 지식이 있는가?	있다면, 도메인 지식을 기반으로 임시 속성을 구성하는 것이 좋다.
쿼리티는 서로 관련이 있는가?	그렇지 않으면, 규모에 맞게 정규화하는 것이 좋다.
특성 간에 상호 의존성이 있는가?	그렇다면 관련 특성을 조합하여 특성 세트를 확장하는 것도 고려하라.
비용, 속도 등에 입력 변수를 제거하여야 하는가?	그렇지 않으면 특성을 분리하거나 특성 가중치의 합을 구성하는 것이 좋다.
모델에 대한 기능을 이해하거나 필터링하고자 기능을 개별적으로 평가하여야 하는가?	그렇다면 가변 순위 방법을 사용하는 것이 좋다.
예측기가 필요한가?	그렇지 않으면 특성을 선택할 필요가 없다.
데이터가 지저분한가?	그렇다면 최상위 변수를 사용하여 이상값을 제거하는 것이 좋다.
무엇을 먼저 해야 할지 아는가?	잘 모르겠다면 선형 예측자를 사용하고 포워드 선택 또는 제로 규범 임베디드 기법을 사용하라.
새로운 아이디어, 시간, 컴퓨팅 리소스, 충분한 데이터가 있는가?	있다면 다양한 방법을 시도해 보는 것이 좋다.
신뢰하는 솔루션을 원하는가?	그렇다면 여러 번 시도해 보고 부트스트랩을 사용하는 것이 좋다.

- 지도 학습 계열 인공지능 모델은 학습 데이터에 대한 라벨링이 요구된다. 그러나 라벨링 작업 시 작업자의 특정 의도가 반영되거나 실수로 인한 특성 정보 누락, 무의식적인 판단 등으로 편향이 발생한다.

- 이는 라벨링 작업자의 전문성 부족, 라벨링 도구 사용 미숙 및 작업 및 판단 기준의 일관성 결여 등이 원인이다. 라벨링 과정에서 발생하는 편향의 잠재적 원인을 사전에 파악하고, 라벨링 결과의 평가 및 작업 기준의 교육 등을 통해 편향 발생을 방지하여야 한다. 또한, 다양한 라벨링 작업자를 섭외하여 작업자별로 나타나는 편향을 최소화하거나, 검수자를 충분히 확보하여 편향 방지 작업을 수행하여야 바람직하다.

07-2c 데이터 전처리 시 특성이 과도하게 제거되었는지 검토하였는가?

Yes ☐ No ☐ N/A ☐

- 특성 선택 기법을 통해서 잠재된 편향을 완화하고 모델 성능을 향상시킬 수 있으나, 지나칠 경우 과적합 overfitting 문제 혹은 오히려 편향의 원인이 되기도 한다.
- 특히, 모든 데이터에서 특성 선택을 시행할 경우, 교차 검증에서 동일한 특성을 사용하게 되므로 편향을 야기할 수도 있다. 따라서 과도한 특성 선택 및 배제를 방지하기 위한 점검이 필요하다.

과도한 특성 선택 및 배제를 방지하기 위한 점검표

점검 항목	조치사항
도메인 지식을 가지고 있는가?	만약 가지고 있다면, 도메인 지식을 바탕으로 임시 특성들을 구성하는 것이 좋다.
특성들이 서로 연관 있는가?	만약 그렇지 않다면, 스케일을 맞추기 위해 정규화하는 것이 좋다.
특성들 사이에 상호 의존성이 있는가?	만약 그렇다면, 관련 있는 특성을 결합하여 특성 셋을 확장하는 것이 좋다.
입력 변수들을 비용·속도 등의 이유로 제거해야 할 필요가 있는가?	만약 그렇지 않다면, 특성들을 분리하거나, 특성의 가중치 합을 구성하는 것이 좋다.
모델에 대한 특성의 이해 혹은 필터링을 위해 특성들을 개별적으로 평가해야 하는가?	만약 그렇다면, Variable Ranking 방법을 사용하는 것이 좋다.
Predictor가 필요한가?	만약 그렇지 않다면, 특성 선택을 할 필요가 없다.
데이터가 지저분한가?	만약 그렇다면, Top Ranking Variable을 이용해 이상값을 제거하는 것이 좋다.
무엇을 먼저 해야 할지 아는가?	만약 모른다면, linear predictor를 사용하고, 전진 선택forward selection 기법이나 0-norm 임베디드 기법을 사용해보는 것이 좋다.
새로운 아이디어와 시간, 컴퓨팅 자원, 데이터가 충분한가?	만약 그렇다면, 다양한 방법을 시도하는 것이 좋다.
안정적인 솔루션을 원하는가?	만약 그렇다면, 여러 번 해보고 Bootstrap을 쓰는 것이 좋다.

07-3 데이터 라벨링 시, 발생 가능한 편향을 확인하고 방지하였는가?

Yes No N/A
☐ ☐ ☐

- 지도학습 계열 인공지능 모델은 학습 데이터에 대한 라벨링이 요구된다. 그러나 라벨링 작업 시에 작업자의 특정 의도가 반영되거나 실수로 인한 특성 정보 누락, 무의식적인 판단 등으로 편향이 발생할 수 있다.
- 이는 라벨링 작업자의 전문성 부족, 라벨링 도구 사용 미숙 및 작업 및 판단 기준의 일관성 결여 등이 원인이 될 수 있다. 라벨링 과정에서 발생할 수 있는 편향의 잠재적 원인을 사전에 파악하고, 라벨링 결과의 평가 및 작업 기준의 교육 등을 통해 편향 발생을 방지해야 한다. 또한, 다양한 라벨링 작업자를 섭외하여 작업자별로 나타날 수 있는 편향을 최소화하거나, 검수자를 충분히 확보하여 편향 방지 작업을 수행하는 것이 바람직하다.

07-3a 데이터 라벨링 기준을 명확히 수립하고 작업자에게 제공하였는가?

Yes No N/A
☐ ☐ ☐

- 스마트 치안 시스템 특성상 데이터는 고도로 전문화되며, 획득한 데이터에 대한 잘못된 해석, 라벨링 작업자의 특성 의도 반영 및 실수로 인한 특성 정보 누락 등으로 인해 편향이 발생한다.

참고 COMPAS 사례[73]

미국 법원에서 10년 넘게 인공지능 시스템은 재판 전 석방 및 특별 가석방 여부를 결정하는 주요 위협 평가에 사용되어 왔다.

COMPAS(대체 제재는 교정 범죄자 관리 프로파일링)는 미국 법원에서 피고가 재범자가 될 가능성을 평가하는 데 사용하는 Northpointe(현 Equivant)에서 개발한 사례 관리 및 의사 결정 지원 도구이다. 그러나 COMPAS의 정확성과 잠재적 편견에 대한 우려로 인해 논란이 되어 왔다.

이 시스템이 개발된 지 오래되지 않아 프로퍼블리카는 이 시스템이 인종 편향적인 결정을 내린다고 판단하였다.

프로퍼블리카는 “2년 동안 재범하지 않은 흑인 피고인이 백인 피고인에 비해 고위험군으로 잘못 분류될 가능성이 거의 2배(45% 대 23%)나 높다는 사실을 발견하였다.”고 밝혔다.

프로퍼블리카에 반대하는 이 연구[75]는 흑인과 백인의 재범률을 다르게 고려한 프로퍼블리카의 오류를 지적한다. 또한 프로퍼블리카가 고위험과 중간 수준을 ‘고위험’으로 분류한 것이 오탐률과 그에 따른 표적 집단 오류를 증가시켰다고 주장하였다.

COMPAS에 대한 또 다른 반대 의견으로, 이 연구에서는 피고의 인종과 성별에 따라 COMPAS 점수의 정확도가 달라진다는 사실을 발견하였다[76].

결론적으로 COMPAS 알고리즘은 인종적 편견, 투명성 부족, 정확성과 유용성의 한계, 형사 사법 시스템의 적법 절차 및 투명성 위반 가능성을 지속시킨다고 비판받아 왔다.

고려하여야 할 필수적인 측면은 알고리즘 설계에 내재한 공정성 개념이다. 예를 들어, COMPAS는 피고인의 재범 가능성을 정확하게 예측하는 데 중점을 두지만, ProPublica는 재범하지 않았지만 잘못 고위험군으로 분류된 피고인에 더 관심을 기울였다. 이러한 공정성 관점의 차이로 인해 통계적 불일치가 발생하였지만, 두 연구 모두 각자의 맥락에서 타당성이 있었다. 모델이 모든 공정성 기준을 충족하는 것은 불가능하므로 공정한 대우로 받아들이도록 정의하고 우선순위를 정하는 것이 중요하다[77].

〈출처: Aniket, AI for Crime Prevention and Detection – 5 Current Applications, 2019[74]〉

- 이러한 잠재적 편향은 다수의 라벨링 작업을 하는 가이드라인이 명확하지 않아 개인 판단에 의존하여 발생한다. 따라서 이를 파악하고 방지하려면 세부적인 라벨링 가이드라인을 마련함은 물론, 다양한 전문가와 협력하여 데이터 라벨링 프로세스에 주의를 기울여야 한다. 다음은 라벨링 가이드 교육 및 계획 수립 절차의 예시이다.
  - ✓ 의사 결정 프로세스 매핑: 전문가와 긴밀히 협력하여 데이터 라벨링 프로세스 수립을 포함한 세부적인 라벨링 표준화 지침을 수립한다.
  - ✓ 올바른 라벨링 도구 사용: 개인의 모든 종류의 개인 데이터를 포함하여 매우 민감한 데이터를 효과적으로 관리하고 라벨링의 주관성을 제거하고자 동일한 작업 환경을 제공한다.

참고 데이터 라벨링 접근법[78]

IBM의 연구에 따르면, 데이터의 라벨링 프로세스는 다음과 같은 경로를 따른다:

- 내부 라벨링: 기존 데이터 과학 전문가를 활용하여 추적을 간소화(정확도 향상 및 품질 향상)
- 합성 라벨링: 기존 데이터셋에서 새로운 프로젝트 데이터 생성(데이터 품질 및 시간 효율성 향상)
- 프로그래밍 방식의 라벨링: 스크립트를 사용하는 자동화된 데이터 라벨링 프로세스(시간 소비와 사람의 주석 작업 필요성 감소)
- 아웃 소싱: 사람을 고용하는 것은 높은 수준의 임시 프로젝트에 최적의 선택임(프리랜서 중심의 워크플로우를 개발하고 관리하는 데도 많은 시간이 소요됨).
- 크라우드소싱: 마이크로 태스킹 기능과 광범위한 배포로 인해 더 빠르고 비용 효율적임(작업자 품질, QA 및 프로젝트 관리는 크라우드소싱 플랫폼에 따라 다름).

07-3b 다양한 데이터 라벨링 작업자를 섭외하기 위해 노력하였는가?

Yes No N/A
☐ ☐ ☐

- 데이터 라벨링 단계에서 인적 편향을 줄이려면 다수의 데이터 라벨링 작업자 확보가 우선 요구된다. 또한, 라벨링 작업자들을 인구 통계학적 특성 및 배경지식 등이 다양하고 고르게 분포되도록 구성하는 것이 바람직하며, 주요 분포 고려 요소는 다음과 같다.
  - ✓ 인종, 종교, 성별, 민족, 장애 여부, 언어, 국적, 경제적 상황 등
- 작업자의 다양성을 검증하려면 크게 2가지를 확인하여야 한다. 첫째, 크라우드소싱crowdsourcing등의 방법을 도입하였는지 점검한다. 둘째, 데이터 라벨링 작업자의 인구 통계적 특성과 배경지식 등을 분석함으로써 실제로 다양하고 고르게 분포되도록 하여야 한다.
  - ✓ 크라우드소싱: 데이터 라벨링 과정에 라벨링 관련 교육을 받은 일반인이 참여하도록 외부 발주함을 의미하며, 이를 통해 기존 라벨링 작업자 집단보다 더욱 다양한 작업자를 확보함
- 다양한 라벨링 작업자를 채용하면 다음과 같이 여러 가지 면에서 큰 효과를 얻는다:
  - ✓ 정확성과 신뢰성 향상: 다양한 라벨링 작업자 그룹을 고용하면 데이터 라벨링 단계에서 인간의 편견을 줄이고, 이들은 다양한 관점, 경험, 배경으로 데이터 라벨링 작업에 참여한다. 서로 다른 관점의 작업자가 오류나 편견을 발견하므로 더욱 정확하고 신뢰하는 데이터 라벨링이 가능하다.
  - ✓ 편향성 감소: 다양한 그룹의 라벨링 작업자를 채용하면 라벨링 프로세스에서 편향성이 발생할 가능성이 줄어든다. 이는 실제 세계에 중대한 영향을 미치는 결정을 내리는 데 사용되는 스마트 치안 시스템의 맥락에서 특히 중요하다.
  - ✓ 효율성 향상: 라벨링 작업자 풀이 더 커지면 라벨링 프로세스가 더 빠르고 효율적으로 완료된다. 이는 스마트 치안 시스템의 개발 속도를 높이는 데 도움이 된다.
  - ✓ 사용자 요구사항에 대한 이해도 향상: 다양한 라벨 제작 작업자 그룹을 채용하면 다양한 사용자/개인 그룹의 요구사항을 더 잘 이해할 수 있다. 이는 다양한 배경과 경험을 한 다양한 사람이 사용하는 스마트 치안 시스템 개발에서 특히 중요하다.
  - ✓ 신뢰와 수용성 향상: 스마트 치안 시스템 개발에 다양한 라벨 제작자 그룹을 참여시키면 다양한 사용자 그룹 사이에서 이러한 시스템에 대한 신뢰와 수용성이 높아진다. 이는 이러한 시스템이 널리 채택되고 효과적으로 사용되도록 하는 데 중요하다.

07-3c 다양한 데이터 라벨링 검수자를 확보하기 위해 노력하였는가?

Yes ☐ No ☐ N/A ☐

- 다양한 데이터 라벨링 작업자를 확보하여도, 인적 편향은 발생한다. 따라서, 데이터 라벨링 검수자를 확보하고, 라벨링 결과가 데이터 수집 목적 및 데이터 스펙과 다른 부분은 없는지 등을 확인하며, 수정을 요청하는 등 작업하여야 한다.
- 데이터 라벨링 검수자 역시 데이터 라벨링 작업자와 마찬가지로 다양하고 고르게 분포하도록 구성하는 것이 바람직하다. 그러므로 크라우드소싱 등의 방법을 도입하였는지 그리고 검수자에 대한 조사와 분석을 통해 그 분포가 다양하고 고르게 형성되었는지 점검한다.
- 라벨링 검사 시에는 추후 데이터셋에 존재하는 사람의 특정 행동 패턴이나 시나리오를 분류하고, 분석 결과 검사 시 변호사, 전문가 등 많은 검수자를 확보하여야 한다.

07-4 데이터의 편향 방지를 위한 샘플링을 수행하였는가?

Yes ☐ No ☐ N/A ☐

- 샘플링은 모집단에서 일정한 기준으로 데이터를 추출하여 표본을 만드는 기법이다. 데이터 편향을 방지하기 위해 일정한 기준으로 데이터를 샘플링하여 다양한 시나리오와 상황을 학습시킨다. 이를 통해 시스템이 다양한 유형의 사건을 정확하고 효과적으로 식별하고 대응한다.
- 얼굴 및 신원 인식 등 데이터셋 내 클래스 불균형 문제는 부적절한 훈련으로 인해 의도하지 않은 편향을 유발하므로 샘플링 기법을 적용하여 방지한다.

07-4a 편향을 방지를 위한 샘플링 기법을 적용하였는가?

Yes ☐ No ☐ N/A ☐

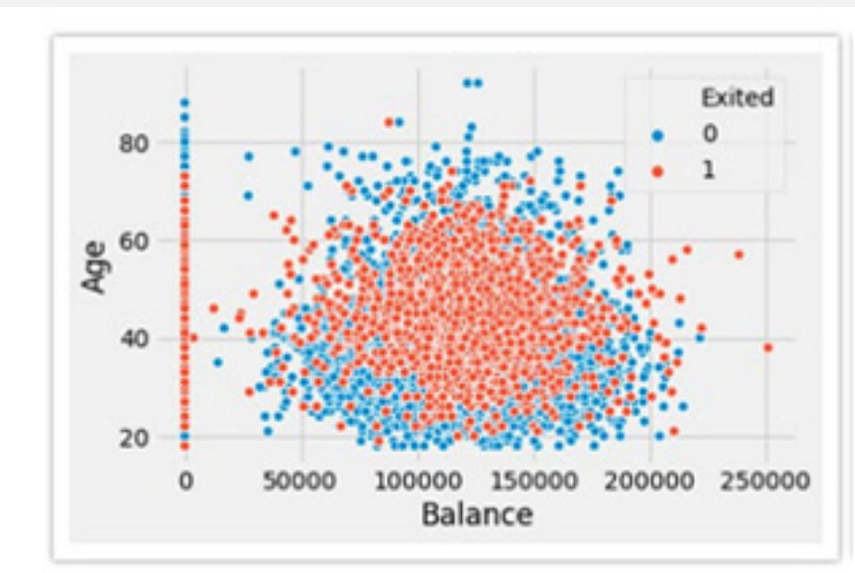
- 스마트 치안 분야 데이터셋에서 연령, 성별 또는 인종, 범죄 이력 등 가능성에 따라 사회적 편견 또는 차별을 일으키는 편향 요인이 있다(06-2a 참고). 이러한 다양한 차별 가능성에 따라 스마트 치안 서비스를 하는 인공지능 학습 데이터셋을 대상으로 할 때는 다음의 인구 통계학적 샘플링 기법을 적용한다.
  - ✓ 확률 샘플링: 단순 무작위 샘플링, 체계적 샘플링, 계층화 샘플링, 클러스터 샘플링
  - ✓ 비확률 샘플링: 편의 샘플링, 자발적 응답 샘플링, 의도적 샘플링, 스노볼 샘플링, 할당량 샘플링, 샘플링 할당량 샘플링
- 스마트 치안 시스템은 얼굴 및 신원 인식은 물론 적대감과 침착함 등의 클래스 구별 작업에서 자연스럽게 클래스 불균형 문제가 발생한다.

- 클래스 불균형 문제를 해결하고자 언더 샘플링, 오버 샘플링 기법 등을 적용한다. 객체 클래스의 불균형으로 편향이 예상되면 그로 인한 편향을 방지하는 샘플링 기법을 적용하고, 적용 과정에서 필요한 활동과 정보가 생성되었는지 확인한다.
  - ✓ 무작위 샘플링: 데이터를 샘플링하는 가장 간단한 방법이다. 이 방법에서는 전체 데이터 집합에서 데이터의 하위 집합을 무작위로 선택한다.
  - ✓ 오버 샘플링 및 언더 샘플링: 오버 샘플링은 해당 클래스의 데이터를 더 추가하여 소수 클래스의 대표성을 높이는 것을 포함한다. 반면에 언더 샘플링은 해당 클래스에서 일부 데이터를 제거하여 다수 클래스의 대표성을 줄이는 것이다. 이러한 방법을 사용하면 데이터 집합에서 서로 다른 클래스의 표현 균형을 맞출 수 있다.

참고 오버 샘플링 기법의 예

랜덤 오버 샘플링

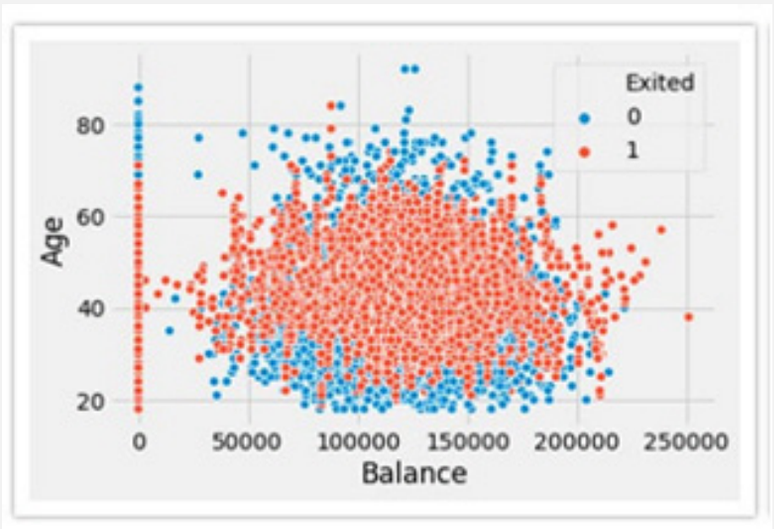
- 기존 소스 클래스를 단순히 복제하여 비율을 조정하는 방법
- 단순 복제로 인해 분포가 변하지 않고, 수가 증가하여 가중치를 더 많이 받는 원리
- 동일한 데이터가 확산하면 과적합의 위험이 있다.



〈출처: Satyam Kumar, 7 Over Sampling techniques to handle Imbalanced Data, 2020〉

SMOTE[79]

- 이 방법도 유용하며, 이 방법은 연속적인 기능만으로 설계되어 이 방법에 대해 언급하였다.
- 임의의 소수 클래스 데이터에서 가까운 소스 클래스 간에 새로운 데이터 생성하기
- 임의의 소스 클래스에 해당하는 관측치 X를 가져와서 X에 가장 가까운 가장 가까운 이웃 X(nn)을 찾은 다음, K X(nn)와 'X 사이에 새로운 데이터 X'를 대체한다.



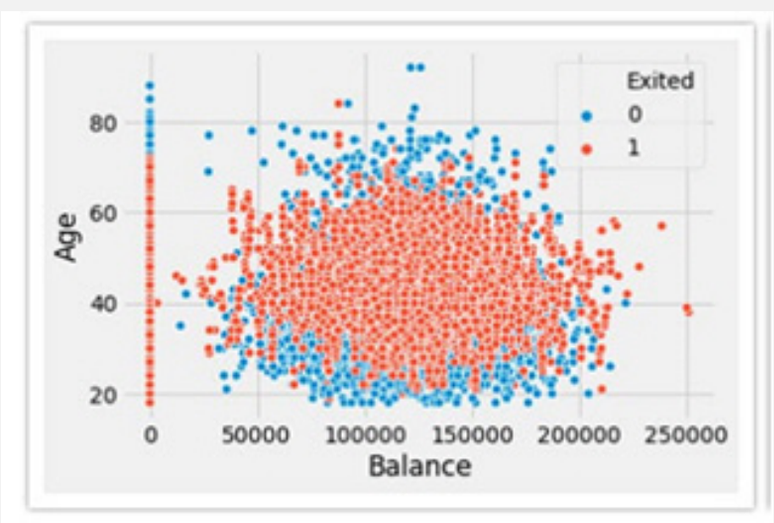
〈출처: Satyam Kumar, 7 Over Sampling techniques to handle Imbalanced Data, 2020〉

Borderline SMOTE[80]

- SMOTE에 약간의 변형이 있는 알고리즘
- 다수 집단과 소수 집단이 서로 인접한 경계선의 분포는 매우 중요하다.
- 경계선의 소수 클래스 데이터에 SMOTE 적용

ADASYN[Adaptive Synthetic Sampling][81]

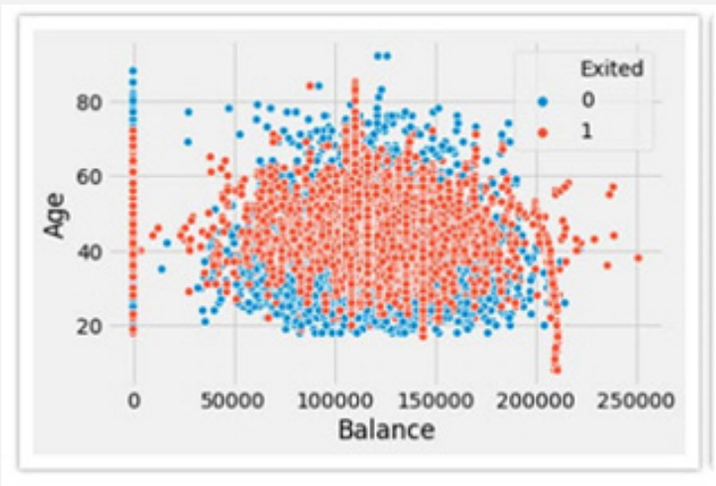
- 경계선 SMOTE에서 알고리즘이 약간 더 수정되었다.
- 경계선 근처에서 [위험, 안전, 소음]의 3가지로 판단하여 SMOTE로 진행한 부분에 가중치를 부여하여 적용한다.



〈출처: Satyam Kumar, 7 Over Sampling techniques to handle Imbalanced Data, 2020〉

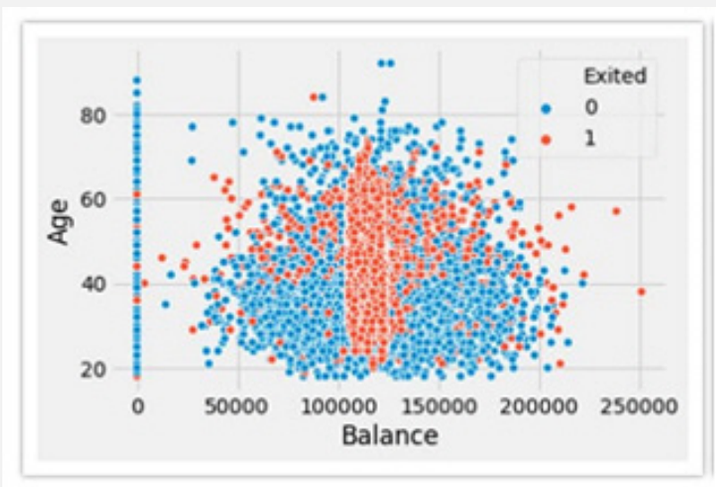
SVM SMOTE[82]

◦ 경계선 SMOTE의 또 다른 버전은 SVM의 오분류 지점을 식별하고자 설계된 SVM SMOTE이다.
◦ 이 방법에서 경계선 영역은 원래 훈련 데이터셋을 훈련한 후 SVM의 지원 벡터 포인트로 근사화된다.
◦ 그리고 각 소수 클래스 서포트 벡터의 경계선 영역에 합성 데이터를 생성한다.



K-Means SMOTE[83]

◦ 클래스 불균형 데이터 때문에 개발된 방법이다.
◦ 입력 공간의 소수 클래스 샘플을 생성한다.
◦ 고전적인 K 평균 클러스터링 알고리즘처럼 작동한다:
  - k-평균 방법으로 데이터 집합 클러스터링하기
  - 소수 클래스 샘플 수가 많은 것을 선택한다.
  - 소수 클래스 샘플에 더 많은 합성 데이터를 개별적으로 할당한다.



〈출처: Satyam Kumar, 7 Over Sampling techniques to handle Imbalanced Data, 2020〉

안전성 책임성

요구사항 08 오픈 소스 라이브러리의 보안성 및 호환성 점검

- 인공지능 모델 설계 및 개발 단계에서는 개발 기간을 단축하고 최신 기술 동향을 빠르고 유연하게 적용하고자 다양한 오픈 소스를 활용한다. 오픈 소스 라이브러리를 활용하기로 하였다면 사용할 라이브러리가 신뢰하는 수준인지, 안정적으로 업데이트 중인지, 주의하여야 할 라이선스 기준은 무엇인지 등 해당 오픈 소스의 버전을 지속해서 확인하여 운영 및 보안상의 위험 요소를 점검한다.

08-1 오픈 소스 라이브러리의 안정성을 확인하였는가?

Yes ☐ No ☐ N/A ☐

- 오픈 소스 라이브러리는 얼굴 인식은 물론 음성 인식 등 인공지능 분야의 연구와 다양한 분야의 협업을 향상하는 데 유용한 도구이며, 사용이 편리하여 대부분 개발자가 프로젝트에 사용하는 경향이 있다. 오픈 소스 라이브러리의 혁신성과 편의성 덕분에 거의 모든 수준의 연구/개발에 활용되며, 치안 시스템에도 활용된다.
- 오픈 소스 라이브러리는 특정 조직, 개인 또는 회사에서 관리한다. 오픈 소스를 운영하는 방식은 다양하므로 사전에 꼼꼼히 확인하여 향후 발생하는 리스크를 최소화하고 라이브러리가 얼마나 안정적인지 측정하여야 한다. 전문가들은 개발자가 오픈 소스 라이브러리를 프로젝트에 통합하기 전에 다음 사항을 확인하여야 한다고 경고한다[85].
  - ✓ 알려진 보안 취약점 확인
  - ✓ 라이선스 확인
  - ✓ 소프트웨어/라이브러리 버전을 확인하고 오픈 소스 라이브러리를 항상 최신 상태로 유지
  - ✓ 활발한 커뮤니티 확인
  - ✓ 문서 및 가이드라인 확인 (예: SBOM[Software Bill of Materials])
- 인공지능 모델을 개발하고자 프로젝트에 오픈 소스 라이브러리를 사용하기로 하였다면, 오픈 소스 라이브러리의 안정성을 확인하여야 한다. 오픈 소스 라이브러리의 안정성은 해당 오픈 소스 라이브러리의 사용자 수, 업데이트가 자주 이루어지는지, 이슈에 대한 대응이 신속하게 이루어지는지 등에 따라 달라진다.

08-1a 활성화된 오픈 소스 라이브러리를 사용하였는가?

Yes No N/A
☐ ☐ ☐

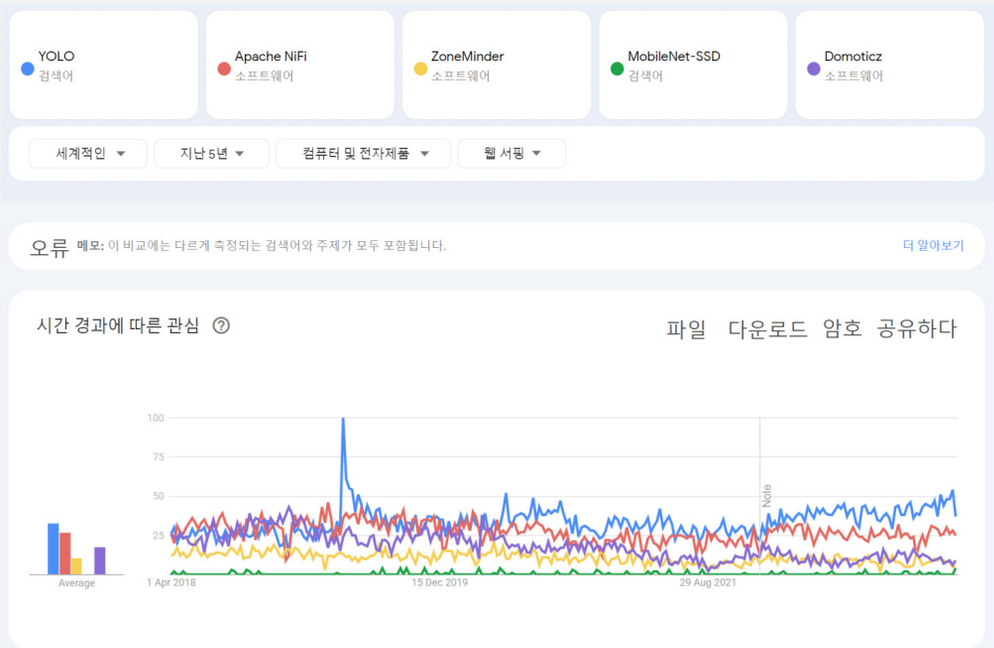
- 오픈 소스 라이브러리 제공자가 라이브러리 개선 및 업데이트를 중단하면, 개발 중인 스마트 치안 시스템에 위험을 초래하므로 오픈 소스 라이브러리 사용에서 안정성은 매우 중요한 요소이다[86].
- 인공지능 라이브러리의 안정성은 많은 개발자가 적극적으로 참여할 때 가능하다는 의견이다. 따라서, 사용하려는 오픈 소스 라이브러리의 개발 과정을 주의 깊게 살펴보아야 한다.
- '기업 공개소프트웨어 거버넌스 가이드 – 정보통신산업진흥원'에 따르면, 오픈 소스 프로젝트의 활성화 정도를 확인하는 것도 안정성을 확인하는 한 가지 방법이다. 해당 오픈 소스가 활발한 커뮤니티에서 논의되는지, 그 커뮤니티 내 구성원들이 적극적으로 협력하는지는 아주 중요한 선택의 표지석이다.
  - ✓ 오픈 소스 라이브러리를 GitHub에서 관리 중이라면, 오픈된 이슈 개수나 Pull Request 수, 마지막 커밋 일시 등을 통해 오픈 소스 개발이 얼마나 활발하게 이루어지고 지속해서 발전할 가능성이 어느 정도인지 파악하여야 한다.
  - ✓ 그 밖에도 해당 오픈 소스와 관련된 StackOverflow 질문 수, 오픈 소스 다운로드 수, Google 쿼리query 결과 수 등을 간단하게 측정하여 해당 라이브러리의 활성화 정도를 확인하여야 한다.
  - ✓ Redhat은, 오픈 소스 기반의 수익화 모델(호환성, 보안 강화, 기술 지원 등 제공)을 개발하고, 오픈 소스 라이브러리 업데이트 시 커뮤니티 내 구성원들이 제안한 개선 사항도 적용한다. 이처럼 수익화 모델 기반의 오픈 소스 라이브러리 역시 개인 및 기업의 참여가 활성화된 프로젝트로 판단하여야 한다.

참고 스마트 치안 분야 오픈 소스 라이브러리의 안정성 분석 예시

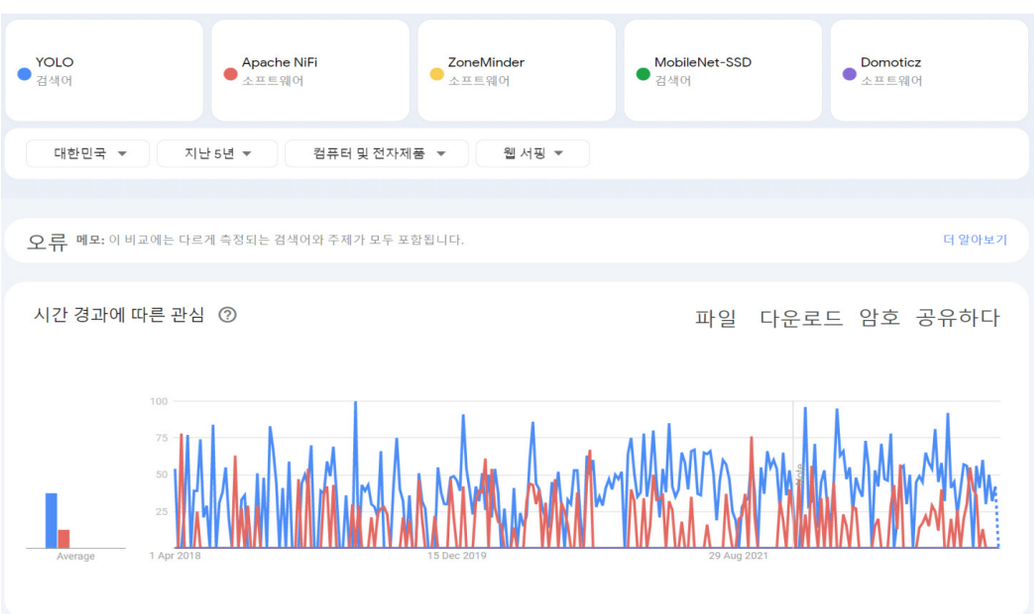
- GitHub 소스 분석 예시(2023.02.01. 기준)

항목 / 오픈 소스 라이브러리	오픈 이슈 개수	Pull Request 수	마지막 커밋 일시	Contributor 수	Used 수	Star 수	StackOverflow 질문 수
OpenCV	2,376	125	2023.03.30	1,448	–	67,634	70,936
TensorFlow	2,016	227	2023.03.31	3,345	–	172,609	81,483
YOLOv5	215	61	2023.03.30	309	–	36,858	592
YOLOv7	1,112	88	2023.03.04	29	–	9,500	1,757
YOLOv8	486	36	2023.03.30	62	428	5,560	–
Apache NiFi	–	26	2023.03.31	438	–	3,699	4,846
ZoneMinder	120	9	2023.03.29	252	–	4,092	19
Motion	3	–	2023.03.27	73	–	3,311	586
MobileNet-SSD	149	–	2021.06.28	7	–	1,893	425
Multiple Object Tracker	58	–	2023.03.08	5	–	1,913	10
Domoticz	92	4	2023.03.26	355	–	3,243	21
allennip	80	11	22. 11. 22	260	2,961	11,382	205

- Google Trends를 이용한 쿼리 분석 예시(2023.03.31. 기준): 전 세계적으로 지난 5년 동안, YOLO(You Only Look Once)와 Apache NiFi는 다른 라이브러리보다 더 높은 조회 수를 기록함. 트렌드 확인 시 주제를 컴퓨터 및 전자 제품으로 한정하여 확인



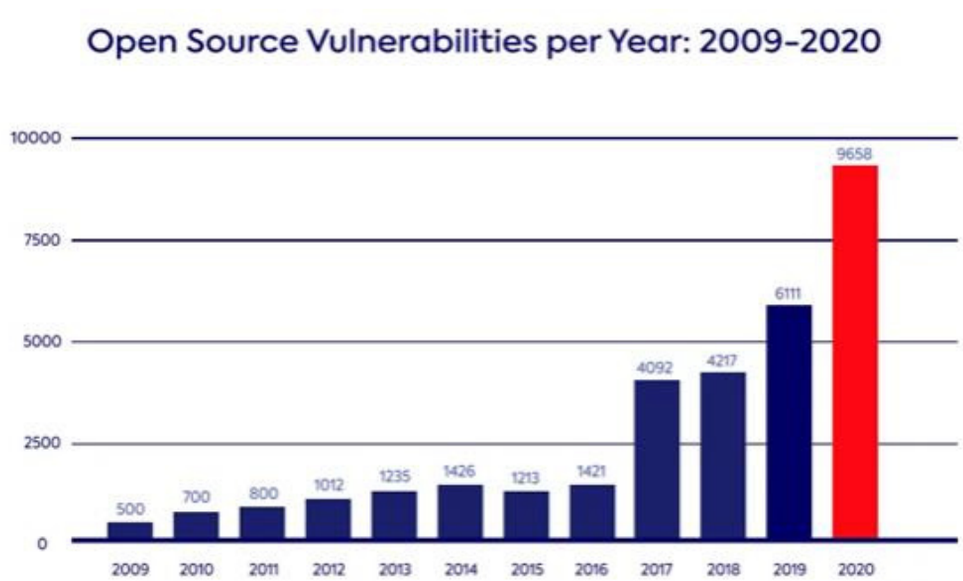
- 한국에서는 지난 5년 동안, YOLO와 Apache NiFi가 다른 라이브러리보다 많은 조회 수를 기록함
- Papers with Code를 이용한 쿼리 분석 예시(2023.03.31. 기준): Papers with Codes에서 'Crowed Counting' 쿼리에 대한 검색 결과 분석 – 논문과 함께 사용된 라이브러리 목록, 참조된 논문 수, 별의 수 및 논문 정보를 사용하여 언급된 라이브러리가 얼마나 신뢰할 수 있는지 확인함



08-2 오픈 소스 라이브러리의 위험 요소는 관리되고 있는가?

Yes No N/A
☐ ☐ ☐

- 2020년 시놉시스에서는 발간한 “2020 오픈 소스 보안 및 위험 분석(OSSRA) 보고서”에 따르면, 개발된 시스템의 50% 이상이 오픈 소스 라이브러리/컴포넌트를 사용하여 제작되는 것으로 확인되었다[84].
- 인공지능 모델 개발 시 오픈 소스 라이브러리 사용 및 위험 요소 관리는 일반적인 인공지능 개발과 비교하여 최소한 동일한 수준으로 관리하여야 한다. 개발 과정에서 사용되는 오픈 소스 라이브러리나 개발 환경 버전 변경에 따른 호환성을 고려하여 오픈 소스 라이브러리의 종류와 버전을 선택하여야 하며, 사용하는 오픈 소스 라이브러리에서 보안 취약점이 발견되므로 이러한 이슈를 점검하여 보안 위험도 관리가 필요하다.



〈출처: Kang Shik Shin, Evaluation Of Open Source Vulnerability Scanning Tools, KAIST CSRC Weblog, 2022〉

08-2a 사용 중인 오픈 소스 라이브러리의 라이선스 준수 사항을 이행하였는가?

Yes ☐ No ☐ N/A ☐

- 오픈 소스는 무료로 사용하지만, 라이선스별로 준수 사항은 별도로 규정된다. 그러므로 오픈 소스 라이브러리를 활용하여 인공지능 모델을 개발한다면, 사용할 오픈 소스 종류 및 라이선스 고지문을 확인하고, 허용 또는 의무 사항을 우선 숙지하여야 향후 발생하는 법률적 위험이 최소화된다. 스마트 치안 분야에서 오픈 소스 라이브러리의 라이선스 요구사항을 이행하는 것은 다음과 같은 이유로 중요하다.
  - ✓ 책임: 스마트 치안 시스템이 오픈 소스 소프트웨어 사용 시, 라이선스의 조건을 따르는 확보에 대한 책임이 있다. 이를 이행하지 않으면 법적인 문제와 잠재적 손해가 초래된다.
  - ✓ 보안: 오픈 소스 소프트웨어는 스마트 치안 시스템에서 기능과 보안을 강화하고자 자주 사용된다. 그러나 라이선스 요구사항을 준수하지 않고 오픈 소스 소프트웨어를 사용하면 시스템이 보안 취약점이나 익스플로잇에 취약해져 보안 위험이 발생한다.
  - ✓ 평판: 오픈 소스 라이선스를 준수하지 않으면 개발자 또는 공급 업체로서 평판과 신뢰도가 손상된다. 오픈 소스 커뮤니티의 부정적인 여론과 불신으로 이어져 고객이나 파트너를 유치하는 데 영향을 미친다(“Log4j 취약성 위기”[100]의 사례).
  - ✓ 협업: 오픈 소스 커뮤니티는 지식과 리소스의 협업과 공유를 통해 번창한다. 오픈 소스 라이선스를 준수하면 다른 사람의 지식재산을 존중할 뿐만 아니라 다른 사람이 사용하고 구축하도록 자신의 소프트웨어를 제공함으로써 커뮤니티에 기여한다.
- 다음은 OSI[Open Source Initiative][87]에 의해 제시된 오픈 소스 라이선스에 대한 요구사항이다.
  - ✓ 자유로운 재배포(Free Redistribution)
  - ✓ 소스 코드 공개(Source Code Open)
  - ✓ 2차 저작물 허용(Derived Works)
  - ✓ 저작자의 소스 코드 원형 유지(Integrity of The Author's Source Code)
  - ✓ 개인이나 단체에 대한 차별 금지(No Discrimination Against Persons or Groups)
  - ✓ 사용 분야에 대한 차별 금지(No Discrimination Against Fields of Endeavor)
  - ✓ 라이선스의 배포(Distribution of License)
  - ✓ 특정 제품에만 유용한 라이선스 금지(License Must not be specific to a product)
  - ✓ 다른 소프트웨어를 제한하는 라이선스 금지(License Must not restrict other software)
  - ✓ 기술 중립적인 라이선스 제공(License must be Technology-Neutral)

상위 활용 오픈 소스 라이브러리의 OSI 기준definition 분석

	Apache License 2.0	AGPL 3.0	Apache License 2.0	GPL 2.0	GPL 2.0	Apache License 2.0	GPL 3.0	GPL 3.0	Apache License 2.0
오픈 소스 라이브러리 OSI 기준	Open CV	YOLO v8	Apache NiFi	ZoneMinder	Motion	Mobile Net-SSD	Multiple object Tracker	Domoticz	allennlp
복제, 배포, 수정 권한 허용	○	○	○	○	○	○	○	○	○
배포 시 라이선스 사본 첨부	○	○	○	○	○	○	○	○	○
저작권 고지 사항 또는 Attribution 고지 사항 유지	○	○	○	○	○	○	○	○	○
배포 시 소스 코드 제공 의무와 범위	-	네트워크 서비스 포함 전체코드	-	전체코드	전체코드	-	전체코드	전체코드	-
조합 저작물 작성 및 타 라이선스 배포 허용	○	-	○	조건부 ○	조건부 ○	○	-	-	○
수정 내용 고지	-	○	-	-	-		○	○	
명시적 특허 라이선스의 허용	○	○	○	-	-	○	○	○	○
라이선시가 특허소송 제기 시 라이선스 종료	○	○	○	-	-	○	○	○	○
이름, 상표, 상호에 대한 사용 제한	○	○	○	-	-	○	-	-	○
보증의 부인	○	○	○	○	○	○	○	○	○
책임의 제한	○	○	○	○	○	○	○	○	○

08-2b 사용 중인 오픈 소스 라이브러리의 호환성 및 보안취약점을 확인하였는가?

Yes ☐ No ☐ N/A ☐

- 라이브러리의 버전 변경 과정에서 개발 환경, 언어, 도구 및 다른 라이브러리 버전과 호환되지 않는 호환성 문제가 초래된다. 따라서 오픈 소스 라이브러리 종류 및 버전 선택 시 라이브러리 간 의존성dependency을 파악하는 등 호환성을 고려하여야 한다.
- 사용 중인 오픈 소스 라이브러리에서 보안 취약점이 발견되기도 한다. 보안 취약점에 따른 영향을 최소화하고자 보안 취약점 및 버전 변경에 따른 릴리즈 노트release note를 확인하여 신속히 탐지 및 대응하여야 한다.
- 도구를 활용하여 오픈 소스의 보안 취약점을 확인하도록 한다. 다음은 무료로 이용하는 도구 중 일부이다.
  - ✓ Snyk(무료 및 유료 버전 제공)
  - ✓ WhiteSource Bolt(GitHub 및 Microsoft Azure 사용자 무료 사용 가능)
  - ✓ Labrador
  - ✓ Veracode SCA(무료 사용하는 커뮤니티 버전 제공)

- OpenVAS, OpenSCAP, OWASP, CVE[Common Vulnerabilities and Exposures details] 등 보안 취약점 기준 및 분석 도구를 통해 오픈 소스 라이브러리의 보안 취약점을 분석하여, 최근에 발견된 보안 위협의 내용과 라이브러리 개발진의 대응 정도를 이해할 수 있다.

참고 스마트 치안 분야 오픈 소스 라이브러리의 보안 취약성 분석 예시

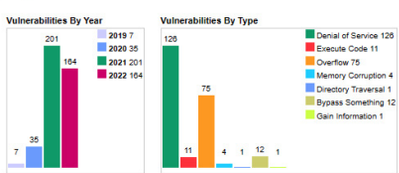
TensorFlow CVE[Common Vulnerabilities and Exposures] 예시(2023.02.01. 기준)

- 서비스 거부(DoS[Denial of Service]) 공격에 취약한 부분이 분석되었으며(31.0%), 오버플로 위협(18.4%)도 분석
- 2021년 201건에서 2022년 164건으로, 보고된 보안 취약점의 총수가 감소하여, 개발진이 보안 위협에 어느 정도 대응하는 것으로 이해됨

Vulnerability Trends Over Time

Year	# of Vulnerabilities	DoS	Code Execution	Overflow	Memory Corruption	Sql Injection	XSS	Directory Traversal	Http Response Splitting	Bypass something	Gain Information	Gain Privileges	CSRF	File Inclusion	# of exploits
2019	7	1	1	4											
2020	35	6	2	8	3										
2021	201	41	6	38	1			1		8	1				
2022	164	78	2	25						4					
Total	407	126	11	75	4			1		12	1				
% Of All		31.0	2.7	18.4	1.0	0.0	0.0	0.2	0.0	2.9	0.2	0.0	0.0	0.0	

Warning : Vulnerabilities with publish dates before 1999 are not included in this table and chart. (Because there are not many of them and they make the page look bad; and they may not be actually published in those years.)



2019~2022년 TensorFlow 라이브러리의 CVE 보안 취약점 분석 결과

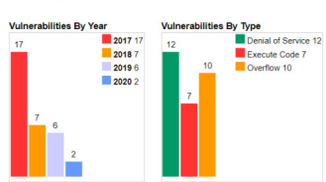
OpenCV CVE[Common Vulnerabilities and Exposures] 예시(2023.03.31. 기준)

- 서비스 거부 공격에 취약한 부분이 분석되었으며(37.5%), 코드 실행(21.9%)과 오버플로 위협(31%)도 분석됨
- 2017년 17건에서 2022년 2건으로 보고된 보안 취약점의 총수가 감소하여, 개발진이 보안 위협에 어느 정도 대응하는 것으로 이해됨

Vulnerability Trends Over Time

Year	# of Vulnerabilities	DoS	Code Execution	Overflow	Memory Corruption	Sql Injection	XSS	Directory Traversal	Http Response Splitting	Bypass something	Gain Information	Gain Privileges	CSRF	File Inclusion	# of exploits
2017	17	6	4	6											
2018	7	4	1	2											
2019	6	2													
2020	2		2	2											
Total	32	12	7	10											
% Of All		37.5	21.9	31.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

Warning : Vulnerabilities with publish dates before 1999 are not included in this table and chart. (Because there are not many of them and they make the page look bad; and they may not be actually published in those years.)



2017~2020년 OpenCV 라이브러리의 CVE 보안 취약점 분석 결과

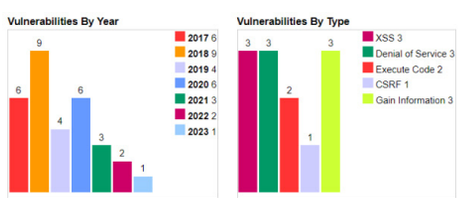
Apache NiFi CVECommon Vulnerabilities and Exposures 예시(2023.04.03. 기준)

- 서비스 거부 공격에 취약한 부분이 분석되었으며(9.7%), 코드 실행(6.5%), XSS (9.7%), 정보 획득(9.7%), CSRF 위협(3.2%)도 분석됨

Vulnerability Trends Over Time

Year	# of Vulnerabilities	DoS	Code Execution	Overflow	Memory Corruption	Sql Injection	XSS	Directory Traversal	Http Response Splitting	Bypass something	Gain Information	Gain Privileges	CSRF	File Inclusion	# of exploits
2017	6						2								
2018	9	2	2				1						1		
2019	4										1				
2020	6										1				
2021	3	1									1				
2022	2														
2023	1														
Total	31	3	2				3				3		1		
% Of All		9.7	6.5	0.0	0.0	0.0	9.7	0.0	0.0	0.0	9.7	0.0	3.2	0.0	

Warning : Vulnerabilities with publish dates before 1999 are not included in this table and chart. (Because there are not many of them and they make the page look bad; and they may not be actually published in those years.)



2017~2023년 Apache NiFi 라이브러리의 CVE 보안 취약점 분석 결과

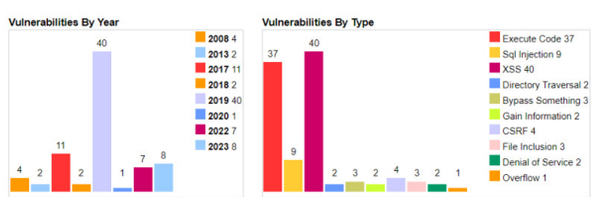
ZoneMinder CVECommon Vulnerabilities and Exposures 예시 (2023.04.03. 기준)

- 서비스 거부 공격에 취약한 부분이 분석되었으며(2.7%), 코드 실행(49.3%), 오버플로(1.3%), SQL 인젝션(12.0%), XSS(53.3%), 디렉터리 트래버설(2.7%), 우회(4.0%), 정보 획득(2.7%), CSRF(5.3%), 파일 포함 위협(4.0%) 등 다양한 부분에서 보안 취약점이 존재하는 것으로 분석됨
- 2019년 40건에서 2023년 8건으로 보고된 보안 취약점의 총수가 감소하여, 개발진이 보안 위협에 어느 정도 대응하는 것으로 이해됨

Vulnerability Trends Over Time

Year	# of Vulnerabilities	DoS	Code Execution	Overflow	Memory Corruption	Sql Injection	XSS	Directory Traversal	Http Response Splitting	Bypass something	Gain Information	Gain Privileges	CSRF	File Inclusion	# of exploits
2008	4		3			1	1								
2013	2		1					1							1
2017	11		2			1	5			1	2		2	1	
2018	2	2	2												
2019	40		23	1		4	30						1		
2020	1						1								
2022	7		3				2	1		1			1		
2023	8		3			3	1			1				2	
Total	75	2	37	1		9	40	2		3	2		4	3	1
% Of All		2.7	49.3	1.3	0.0	12.0	53.3	2.7	0.0	4.0	2.7	0.0	5.3	4.0	

Warning : Vulnerabilities with publish dates before 1999 are not included in this table and chart. (Because there are not many of them and they make the page look bad; and they may not be actually published in those years.)



2008~2023년 ZoneMinder 라이브러리의 CVE 보안 취약점 분석 결과

다양성 존중

요구사항 09 인공지능 모델의 편향 제거

- 인공지능 모델을 개발하는 과정에서 모델 종류나 시스템 목표에 따라 발생하는 편향을 제거하는 기법을 고려하여야 한다(자세한 내용은 08-1a에서 참조).

  * 요구사항 06-2에 언급된 것처럼 지역, 인종 차별, 방언 차이, 사회 경제적 또는 성차별 등 사회적·윤리적 문제가 있을 때만 해당한다.

  ※ 이때, 잠재적 사용자에게 결과의 편향성 이유에 대해 신뢰하고 수용 가능한 설명/정보를 제공하여야 한다. 또한 피해를 본 이용자에게서 제62조에 따라 고소당하는 것을 방지하고자 잠재적 이용자에게 확인서를 받는 것이 좋다(참고: 04-1c).

09-1 모델 편향을 제거하는 기법을 적용하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 모델은 데이터에 잠재된 편향을 학습하거나 심지어 편향을 더욱 증폭시키기도 한다. 따라서, 데이터 정제 단계에서 데이터에 잠재된 편향을 제거하는 방법뿐만 아니라, 모델 개발 과정에서도 모델 편향을 제거 또는 완화하는 기법을 적용하는 것이 바람직하다.
- 편향 완화 기법은 이를 적용하는 단계에 따라 3가지 방식으로 나뉜다. 모델 학습 전에 적용해야 할 편향 완화 기법$^{pre-processing}$, 모델 학습 중에 적용할 기법$^{in-processing}$, 모델 학습 이후 적용할 기법$^{post-processing}$이다. 구현하려는 인공지능 모델 및 목표 임무에 따라서 이 중 적절한 기법을 선택하여 적용하여야 한다.

09-1a 개발하려는 모델에 맞게 편향제거 기법을 선택하였는가?

Yes ☐ No ☐ N/A ☐

- 스마트 치안 분야에 활용되는 오픈 데이터셋은, 민감한 특성 정보를 포함하는 경우가 드물다. 따라서, 오픈 데이터셋을 활용하여 인지 알고리즘 개발 시, 사회·윤리적으로 문제가 되는 편향 제거 기법은 현시점에서 고려 대상은 아니다.
- 그러나, 추후 스마트 치안 시스템의 데이터셋에 민감한 특성(예: 인종, 성별, 나이 등)이 반영되거나, 시스템이 편향을 유발하면 다음과 같은 기법들을 고려하여 편향을 완화한다.
  - ✓ 알고리즘 공정성 기법: 알고리즘 공정성 기법은 스마트 치안 시스템에 적용되어 차별적 요인을 기반으로 결정하지 않다. 이러한 기법에는 편향을 제거하기 위해고자 의도적으로 조정된 데이터로 모델을 재학습시키는 등의 방법이 포함된다.

✓ 사람 감독: 스마트 치안 시스템은 사람의 감독을 포함하도록 설계되어야 한다. 이는 시스템이 경고를 표시할 때, 사람이 시스템을 모니터링하고 결정함을 포함하는 의미이다. 이는 시스템이 편향된 결정을 내리지 않도록 하며 시스템의 전체 정확도 또한 향상한다.

- 각 방식의 특성과 구현하려는 인공지능 모델 및 목표 임무에 맞게 적절한 기법을 선택하고 적용하여야 한다.

대표적 발생 가능 편향에 따른 적용 가능 기법

편향 유형	기법 (접근 방법)	기법 구분			설명
		Pre	In	Post	
인지 편향 congnitive bias	다양한 결정 계획 수립		✓		판단할 때 발생하는 경험 및 외형 인지 등에서 비롯된 편향으로, 다양한 팀 또는 전문가의 도움으로 인지 편향 완화 가능 (예: 개인이 가해자가 될 위험성을 추정하는 '가해 가능성 예측' 시 가해 가능성 수준 판단의 편향)
알고리즘 편향 algorithmic bias	가중치 재조정	✓	✓		데이터셋의 부분집합 불균형으로 인해 발생하는 편향으로 성별/인종 등에 따른 라벨의 조건부 확률에 기반하여 각 그룹-라벨 조합에 다른 가중치를 적용하여 완화 가능 (예: 밝은 피부색 관련 데이터 수가 높거나 우세하여 상대적으로 어두운 피부색의 인종이 치안 감시 대상으로 인식되는 편향 발생[88])
평가 편향 evaluation bias	임곗값		✓		통제 및 결정 단계에서 혼혈, 인종에 따라 평가/예측하는 등 윤리적 측면의 딜레마. 특정 조건에 반응하도록 훈련된 AI 모델에 대한 인간의 의사결정 편향[89] 등
자동화 편향 automation bias	자동화 시스템 감독			✓	스마트 치안 인공지능 모델의 자동 판단 선호도에 의한 편향으로 모델의 지속적인 감독을 통해 완화 가능 (예: 인공지능 모델이 얼굴 대칭이나 눈 모양 같은 일부 얼굴 특징을 읽는 능력을 향상하면서, 턱선, 뺨뼈, 눈 모양, 이마 등의 형태적 특징에서 학습한 인종, 출신 지역을 반영하여 무의식적으로 자동화를 예측함[90])

참고 AI 모델의 편향성 사례 예시

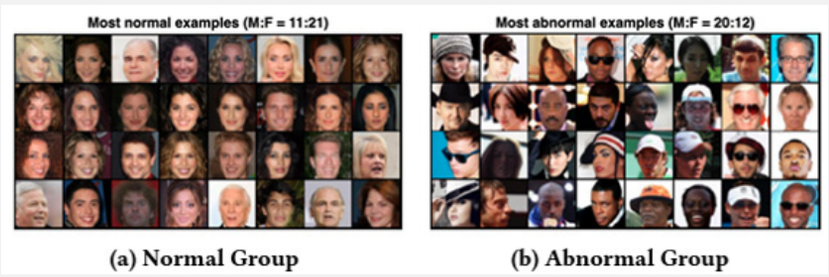
스마트 치안 시스템에 사용되는 인공지능 모델은 편향적이며, 특정 집단에 대한 불공정한 대우로 이어진다. 이를 파악하고자 주어진 스마트 치안 사례를 확인한다.

자동화된 치안 접근 방식은 종종 부정확하며, 인공지능 모델의 예측에 의문을 제기하지 않을 때가 많아 책임 공백으로 이어진다. 편향된 데이터를 사용하여 인공지능 모델을 학습시키면 특정 지역에서 과잉 치안이 이루어지고 위험한 피드백 루프가 발생한다[91].

법 집행 및 출입국 관리에 안면 인식 및 기타 알고리즘 기반 기술을 사용하면 인종 차별과 외국인 혐오가 심화할 위험이 있으며 인권 침해로 이어진다. 편향된 데이터는 AI 모델 학습에 사용되어 차별적인 결과를 초래한다[92].

경찰 관행의 과거 데이터는 알고리즘이 어떤 지역이 "나쁜" 지역이고 어떤 지역이 "좋은" 지역인지에 대한 태도를 반영하고 강화하는 결정을 내리는 피드백 루프를 만든다. 주로 체포 데이터에 기반한 인공지능은 실제 신고된 범죄와는 달리 편향 위험성이 더욱 높다[93].

또한 한 연구[94]에 따르면, 인공지능 시스템에 사용되는 이상 징후 감지 알고리즘은 아프리카계 미국인이나 피부색이 어두운 남성을 이상 징후로 예측할 가능성이 더 높다고 한다. 이는 스마트 치안 모델의 편향성이 특히 소수 인구와 여성에게 편향된 결과를 초래함을 나타낸다. 이 연구는 이러한 유형의 알고리즘이 유색 인종을 이상 징후로 식별할 가능성이 높으며, 이는 인종 프로파일링과 차별을 초래한다는 사실을 발견하였다.



〈출처: 공정한 심층 이상 징후 탐지를 향하여[125]〉

연구원들은 알고리즘 자체가 편향된 것은 아니지만, 알고리즘은 학습된 데이터를 통해 학습한다고 설명하였다. 데이터가 편향되면 알고리즘은 특정 특성을 정상과 연관시키는 방법을 학습하게 된다. 이 문제를 해결하고자 연구자들은 적대적 네트워크를 사용하여 민감한 속성과 학습된 표현 간 관계를 제거하는 공정한 이상 징후를 탐지하는 새로운 아키텍처를 제안하였다. 이 접근 방식은 공정하고 정확한 이상 징후 예측을 동시에 보장한다. 심층 이상 징후 탐지에 공정성을 추가하는 데 따르는 어려움은 알고리즘이 편향되지 않고 정확한 예측을 하도록 한다.
주어진 기사와 연구를 살펴보면 스마트 치안 모델에 사용되는 인공지능 모델이 편향되어 특정 그룹에 대한 불공정한 대우로 이어짐을 알 수 있다. 스마트 치안 시스템에 사용되는 인공지능 모델이 공정하고 편향되지 않도록 모니터링하고 감사하는 것이 중요하다.

09-1b 편향성 평가 및 모니터링을 위한 정량적 지표를 선정하고 관리하는가?

Yes ☐ No ☐ N/A ☐

- 편향성을 정량적으로 측정하는 지표는 아래 표와 같이 5가지 분류로 나누며, 개발하려는 모델과 임무 목표에 맞게 지표를 선정하고, 편향 완화 여부를 지속해서 측정 및 관리하는 것이 바람직하다.

편향을 정량적으로 측정하는 지표 분류

분류	지표
패리티parity 기반 지표	• 인구 통계학적statistical/demographic 형평성 지표, 차등적disparate 효과 지표
혼동 행렬confusion matrix 기반 지표	• 동등 기회equalized opportunity, Equalized Odds, 전체 정확도 형평성, 조건부 사용 정확도 형평성, 대응 형평성, 비보상 동등화
점수score 기반 지표	• 양성 및 음성 클래스 균형 지표
사후 가정counterfactual 기반 지표	• 사후 가정 공평성
개인individual 공평성 지표	• 일반화 엔트로피 지수, 셰일 지수

참고 편향성 평가 및 모니터링 지표

하나의 지표로는 이러한 시스템 편향성의 복잡성을 완벽하게 파악할 수 없으므로, 개발된 시스템의 복잡성에 따라 종합적인 평가를 하도록 다양한 지표를 조합하여야 한다.

- 거짓 양성률False Positive Rate, FPR: 모델이 위협이나 범죄자로 잘못 식별하는 비율. FPR이 높다는 것은 모델이 특정 그룹이나 개인에 대해 편향됨을 나타냄
- 거짓 음성률False Negative Rate, FNR: 모델이 실제 위협이나 범죄자를 식별하지 못하는 비율. 높은 FNR은 모델이 특정 그룹이나 개인에게서 위협을 탐지하는 데 효과적이지 않음을 나타냄
- 균등한 확률Equalized Odds: 모델의 예측이 다른 인구 통계학적 그룹에서도 동일하게 정확함을 보장하는 공정성의 척도. 균등한 확률 기준이 충족되지 않는다면, 이는 모델이 특정 그룹이나 지역의 개인이나 사건에 대해 편향됨을 나타냄
- 오류율 패리티Error-Rate Parity: 이 측정값은 두 그룹 간 오탐률과 오탐률이 비슷해야 함을 보장함
- 예측 균등성Predictive Parity: 모델의 예측이 인구의 다른 하위 그룹 간에 동일하게 정확함을 보장하는 공정성의 척도. 예측 균등성 기준이 충족되지 않는다면, 모델이 특정 하위 그룹에 대해 편향됨을 나타냄
- 보정Calibration: 시스템의 예측이 현실과 얼마나 잘 일치하는지(예측 동등성)를 측정하는 척도. 시스템이 예측에 대해 과신하거나 과소신뢰하면 편향성을 나타낸다. 특히 개발된 시스템에 센서, 레이더, 엣지 디바이스, 로봇 부품 등이 포함될 때 더욱 그러하다.
- 통계적 평등Statistical Parity: 이 기준은 위험도 점수의 분포가 두 그룹 간에 유사해야 함을 강조한다(결과적 공정성).
- 또한 클라인버그[95]에 따르면 오차율의 공정성 기준과 예측적 공정성 기준은 양립할 수 없다.
- 인구 통계학적 평등Demographic Parity: 이는 시스템의 결정이 인종, 성별 또는 연령등 인구 통계학적 요인에 영향받지 않도록 보장하는 또 다른 공정성 척도이다. 인구 통계학적 평등성 기준을 충족하지 못하면 시스템이 특정 그룹, 사회경제적 상황 또는 개인 또는 이벤트의 지역에 편향됨을 나타낸다.
- 정확도 불균등Accuracy Disparity: 다른 인구 통계학적 그룹/지역 등 간 정확도 차이를 측정하는 척도. 특정 그룹/지역이 모델의 예측에서 지속해서 정확도가 떨어진다면, 이는 편향을 나타냄

• 스마트 치안 시스템에서 인공지능 공정성 지표를 사용할 때, 여러 지표 간 내재한 비호환성 인식이 중요하다. 아래 연구들은 특정 인공지능이 다양한 공정성 기준을 동시에 만족시키기 어려움을 보여 주었다.

참고 사례 연구 – AI 시스템에서 차별을 방지하는 AI 공정성 평가[96]

이 연구는 한국과 미국을 중심으로 차별금지법의 중요성과 국가별 차별금지법의 차이에 대해 살펴본다. 한국의 차별금지법은 직접 차별과 간접 차별에 모두 대응하며, 차별적 대우에서 특정 보호변수를 배제함을 강조한다. 직접 차별에 대응하려면 입력 단계에서 해당 변수의 사용을 배제하여 문제의 근원을 해결하는 접근 방식이 사용된다. 간접 차별은 편향된 결과가 특정 그룹에 불균형적으로 영향을 미칠 때 발생한다. 간접 차별을 감지하는 기법에는 결과 동등성, 오류율 동등성, 예측 동등성 기준이 있다. 결과 동등성 기준은 결과를 실제 인구 비율과 일치시켜 간접 차별을 식별하는 데 중요한 역할을 한다.
또한 예측 동등성 기준은 AI 사용자 관점에서 신뢰도와 관련 있다. 반면, 오차율 동등성에는 오탐 등의 요소가 포함되며 공정성에 큰 영향을 미친다. 중요한 것은 이러한 기준을 차별 금지법에 통합하여 공정성 지표와 법적 의미 사이의 간극을 좁혀야 한다는 요구가 증가한다는 점이다.
따라서 직접 차별을 방지하는 원칙은 입력 단계에서 보호변수를 통제함을 포함하며, 간접 차별을 해결하는 것은 주로 결과의 평등성 달성에 중점을 둔다. 예측 동등성과 보정 기준은 신뢰성 확보의 핵심이다. 이러한 원칙은 의사 결정 과정의 공정성과 형평성을 보장하고자 스마트 치안 시스템을 비롯한 다양한 AI 애플리케이션에서 매우 중요하다.

안전성

요구사항 10 인공지능 모델 공격에 대한 방어 대책 수립

- 스마트 치안 알고리즘 중 위협 분석 시 인공지능 모델은 시각, 음성, 텍스트, 숫자, 위치, 센서, 엣지 디바이스, 생체 정보 데이터 등 중요한 정보가 포함되어 있기 때문에 적대적 의도가 있는 사용자에 의해 인공지능이 잘못된 의사결정을 하도록 유도하는 공격의 대상이 될 수 있으므로 이를 방지 또는 완화하기 위한 대책을 수립한다.

10-1 모델 공격이 가능한 상황을 파악하였는가?

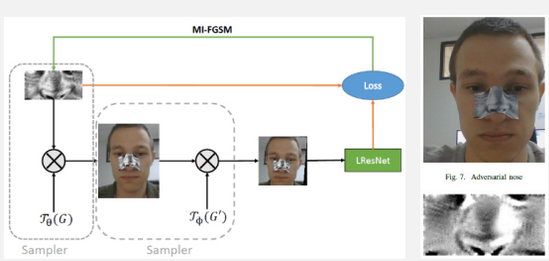
Yes ☐ No ☐ N/A ☐

- 적대적으로 생성된 입력과 같이 작은 변화에도 모델을 오동작하게 만드는 공격은 인공지능 시스템의 안전성을 위협할 수 있다. 따라서, 적대적 공격을 이해하고 적절한 대응 방안을 마련하여 인공지능 모델의 견고성을 향상시키는 것이 필요하다.
- 적대적 공격의 대표적 유형으로는 회피 공격evasion attack이 있다. 추론 중에 인공지능 모델을 속이기 위해 입력 데이터를 조작하는 것이다. 이러한 공격에 대응하는 방안을 수립하기 위해서는, 개발 중인 모델의 데이터 유형(예: 이미지, 텍스트, 오디오)별로 공격 가능한 적대적 사례를 파악하여야 한다.
- 스마트 치안 서비스에 적용 중인 모델을 대상으로 자행되는 추출 공격을 완화 또는 방어하고자 쿼리 횟수 제한, 난독화, 차등 프라이버시, 워터마킹, 능동적 방어 등의 방법을 적용한다.

참고 얼굴 인식 기반 감시 시스템에 대한 모델 회피 공격 예시 [108]

최근에는 특히 감시 시스템이 공격 이미지, 패치, 티셔츠 등을 사용하여 공격받는다. 공격자는 적대적 이미지를 통해 물리적으로 사기 사건을 일으켜 개발된 스마트 치안 시스템을 공격하려고 시도한다.

이 연구에서 연구자들은 실제 공격에 대한 얼굴 인식 시스템의 취약성에 대해 논의하고 이러한 시스템을 속이는 적대적인 패치를 만드는 방법을 제안한다. 연구자들은 얼굴의 여러 부위에 이러한 패치를 부착하여 그 효과를 연구하는 실험을 수행하였다. 사전 학습된 모델을 사용하여 다양한 사람의 임베딩을 수집한 결과 패치의 크기와 배치가 공격 성공에 영향을 미친다는 사실을 발견하였다. 실험 결과, 제안된 방법이 디지털 및 물리적으로 얼굴 인식 시스템을 성공적으로 속임을 보여 주었다.



〈출처: On Adversarial Patches: Real-World Attack on ArcFace-100 Face Recognition System, SIBIRCON, 2019[101]〉

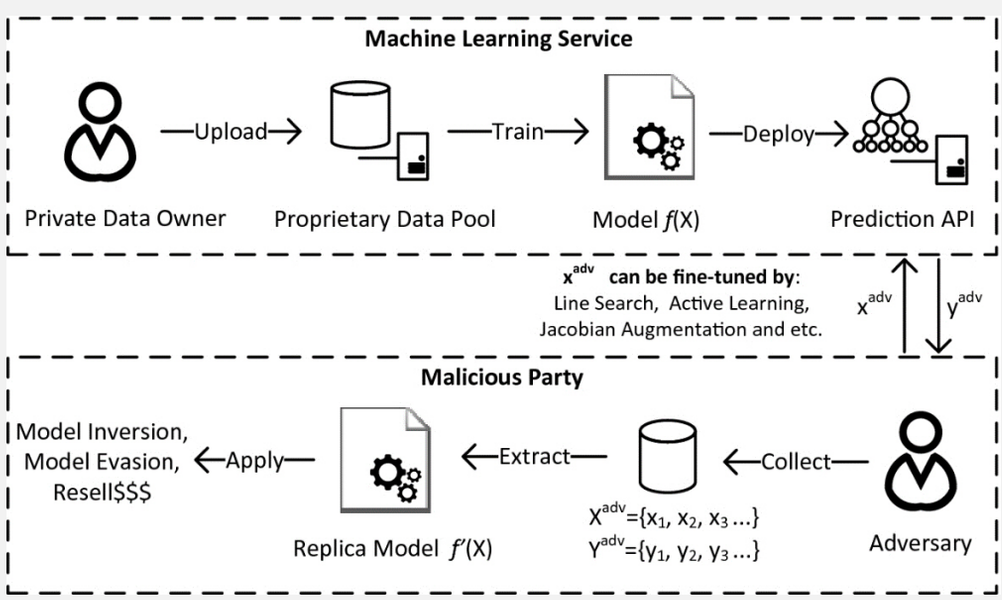
10-1a 데이터 유형별 공격 가능한 적대적 사례를 확인하였는가?

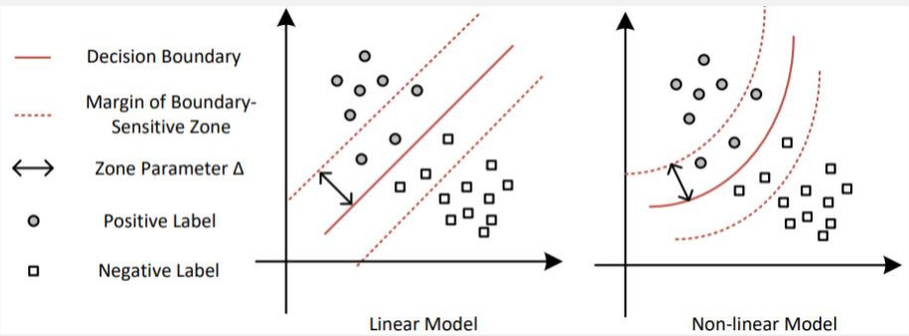
Yes No N/A
☐ ☐ ☐

- 적대적 공격에 관한 연구가 가장 활발히 이루어지고 있는 분야는 영상 분야로, 입력 이미지 공격이 주를 이룬다. 이미지는 텍스트나 오디오에 비해 픽셀값의 고차원 배열로 표현되는 복잡성으로 인해 적대적 사례를 생성하기가 비교적 쉽기 때문이다. 생성된 적대적 사례는 사람에게는 정상으로 보일 정도로 설계되지만, 모델의 예측을 변경시킬 수 있다. 이미지를 대상으로 한 적대적 사례의 예시는 다음과 같다.
  - ✓ CCTV 화면에 적대적 노이즈 현상을 추가하여 시스템이 차량의 번호판을 오인식하도록 유도
  - ✓ 의료 분야 영상에 적대적 노이즈를 추가하여 세그멘테이션 성능을 떨어뜨리도록 유도
- 텍스트 데이터 대상으로는, 문장에 대한 긍정 또는 부정에 대한 판별 모델에 대한 적대적 사례 연구가 진행되고 있다. 문장에서 중요 단어에 대한 후보군을 선정한 후 이를 대체하고 문법상 문제 여부, 유사도 등을 판단한 후에 오인식 확률이 높은 단어로 대체함으로써 공격이 가능하다.
- 오디오 데이터는 사람이 들을 수 없는 작은 노이즈를 입력에 추가하여 음성 인식 모델에 의해 잘못 인식되는 사례를 찾는 방법을 사용한다. 또한, 적대적 공격은 아니지만 오디오 데이터의 오인식 공격 방법으로써 특정 음으로 기계를 오작동시키거나, 사람이 들을 수 없는 영역대의 주파수를 이용하는 연구들도 진행된 바 있다.

참고 BDPL 경계 차등 비공개 레이어[100]

또한 모델 추출 공격에 관한 다른 연구로는 BDPL이 있다. BDPL 기법은 분류를 결정하는 기준과 그 주변 영역을 경계 민감 영역으로 지정하고 이 영역을 보호함으로써 외부의 입력이 민감 영역에 가까워질수록 결과에 노이즈가 섞여 모델 추출을 어렵게 한다.





〈출처: Zheng, H., Ye, Q., Hu, H., Fang, C., Shi, J. "BDPL: A Boundary Differentially Private Layer Against Machine Learning Model Extraction Attacks", 2019〉

10-2 모델 공격에 대한 방어 수단을 강구하였는가?

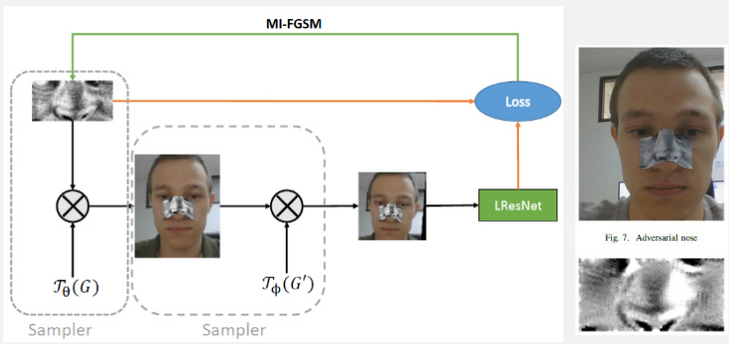
Yes No N/A
☐ ☐ ☐

- 06-1 에서 언급한 것처럼, 인공지능 서비스 운영 과정에서 입력 데이터에 최소한의 변조를 가해 예상과는 다른 결과를 출력하도록 하는 적대적 공격에 노출될 수 있다. 따라서, 인공지능 모델 개발 단계에서 대처 가능한 방안을 검토 및 적용하는 것이 바람직하다.
- 10-1 을 통해 현재 개발 중인 모델의 공격 가능한 상황을 파악하였다면, 모델 최적화model optimization를 통해 적대적 공격에 방어할 수 있다. 모델 최적화는 주로 성능 향상, 자원 효율성 향상, 학습 시간 단축, 모델 해석성 개선 등의 차원에서 활용되지만, 적대적 사례에 대한 효과적인 대응을 위해 활용되기도 한다. 모델 최적화를 통한 방어 대책을 통해 모델이 적대적 사례에 강건하게 동작하도록 한다.
- 완벽한 방어 방법은 없으므로 다양한 방어 기법과 지속적인 업데이트를 통합하는 계층화된 보안 전략이 핵심이다. 시스템 개발자는 보호하고자 엣지 디바이스, CCTV 카메라, 센서, 레이더에 방어 기술을 적용함을 고려하여야 한다.

참고 얼굴 인식 기반 감시 시스템에 대한 모델 회피 공격 예시 [108]

최근에는 특히 감시 시스템이 공격 이미지, 패치, 티셔츠 등을 사용하여 공격받는다. 공격자는 적대적 이미지를 통해 물리적으로 사기 사건을 일으켜 개발된 스마트 치안 시스템을 공격하려고 시도한다.

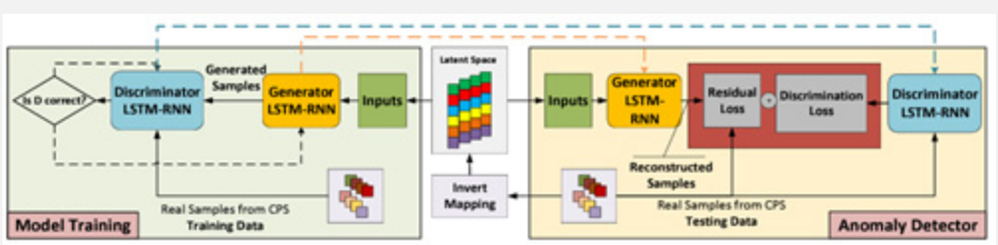
이 연구에서 연구자들은 실제 공격에 대한 얼굴 인식 시스템의 취약성에 대해 논의하고 이러한 시스템을 속이는 적대적인 패치를 만드는 방법을 제안한다. 연구자들은 얼굴의 여러 부위에 이러한 패치를 부착하여 그 효과를 연구하는 실험을 수행하였다. 사전 학습된 모델을 사용하여 다양한 사람의 임베딩을 수집한 결과 패치의 크기와 배치가 공격 성공에 영향을 미친다는 사실을 발견하였다. 실험 결과, 제안된 방법이 디지털 및 물리적으로 얼굴 인식 시스템을 성공적으로 속임을 보여 주었다.



〈출처: On Adversarial Patches: Real-World Attack on ArcFace-100 Face Recognition System, SIBIRCON, 2019[101]〉

참고 모델 추출 공격에 대한 방어 기술[99]

연구원들은 침입 이벤트(여러 시계열 데이터 및 데이터 스트림)를 지속해서 모니터링하는 데 사용하는 네트워크에 연결된 여러 센서 및 액추에이터의 이상 징후 탐지로 비지도 GAN 모델을 사용한다. 이 모델은 먼저 적대적 샘플을 생성하고 이를 판별 모델에 전달하여 실제 센서 데이터와 구별한다. 모델 추출의 구조는 다음과 같다.



〈출처: Dan Li et al. "Anomaly Detection with Generative Adversarial Networks for Multivariate Time Series", 2018〉

10-2a 모델 최적화를 통한 방어 대책을 마련하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 모델 개발 단계에서는 모델 최적화를 통해 인공지능 모델이 적대적 사례에 강건하게 대응할 수 있다. 대표적인 방어 대책으로는 Defensive Distillation, Gradient Regularization, Gradient Masking, Stochastic Network 등이 존재한다. 각 방안에 대한 설명 및 기법 예시를 다음 표에 정리하였다.
- 방어 대책을 선택할 때는 10-1 을 통해 파악한 데이터 유형별 적대적 사례를 먼저 확인하는 것이 좋다. 예를 들어 Defensive distillation의 경우, 텍스트 분류를 수행하는 신경망을 대상으로 한 적대적 사례에 대해 견고성을 크게 향상시키지 못하였다는 연구 결과가 존재한다. 따라서, 방어 대책을 적용할 때는 데이터 유형에 가장 적합한 방안을 선택하는 것이 필요하다.

AI 모델 회피 공격에 대한 방어 기법 예시

방어 기법 분류	방어 기법 내용
적대적 훈련 adversarial training	모델 훈련 시 훈련 데이터셋에 적대적 사례를 모방하는 적대적 샘플 데이터셋을 포함함 그러나, 모든 적대적 사례의 수를 고려하지 않는다면, 즉 충분한 수와 다양성이 있는 적대적 샘플 훈련 데이터셋이 보장되지 않는다면, 방어 기술로서의 성능이 보장되지 않음 • 적대적 학습 방법의 유형: 빠른 기울기 부호 방법fast gradient sign method, FGSM, 투영된 기울기 하강법projected gradient descent, PGD, 적대적 로짓 페어링adversarial logit pairing, ALP • 적대적 학습은 적대적 예제를 훈련 데이터에 포함하여 모델의 견고성을 향상하는 것을 의미하며, 적대적 샘플은 인공지능 모델을 속이고자 특별히 설계된 입력값임 이러한 데이터를 훈련 데이터에 포함하여 모델은 공격에 대해 더 저항력을 가지도록 학습시킴
입력 정제 input sanitization	모델에 입력 데이터를 제공하기 전에 조작이나 변경 여부를 확인함. 악의적인 입력 데이터에 의해 모델을 속이는 것을 방지함
입력 유효성 검사 input validation	모델에 입력 데이터를 제공하기 전에 조작이나 변경 여부를 확인함. 악의적인 입력 데이터에 의해 모델을 속이는 것을 방지함
랜덤화 randomization	입력 데이터에 무작위 노이즈나 변형을 도입하여 공격자의 회피 공격을 어렵게 함
앙상블 모델 model ensembling	동일한 두 모델(원본 모델과 적대적 공격을 판단하는 모델)의 추론 결과를 비교하여 두 결과 사이에 차이가 발생할 때 적대적 공격으로 판단함 또한, 특정 모델에 적용되는 적대적 공격을 불가능하게 만들고자 여러 학습 모델을 결합하여 최종 판단하는 방법을 적용함 • 모델 결합은 여러 모델의 출력을 결합하여 전체적인 정확도와 견고성을 향상함을 의미함. 이는 단일 모델이 적대적 공격에 속을 위험을 줄임
마그넷magnet 방법[102]	정상 데이터의 다양성manifold을 근사화하여 정상과 적대적인 예제를 구별함 적대적인 샘플을 다시 만들어, 작은 적대적인 경우를 올바르게 분류하는 왜곡에 효과적인 다양성 근처로 이동시킴
쿼리 수 제한	반복적인 쿼리를 시도하는 역전inversion 공격이나 모델 추출 공격을 방지하고자 반복적인 조회 수를 제한함
Stochastic Network	학습 모델의 불확실성을 다루기 위한 확률적인 요소를 도입하는 네트워크를 말한다. 이를 통해 모델의 결정을 불확실하게 만들어 적대적 사례에 대한 저항성을 높인다. (예: defensive dropout, Random Self-EnsembleRSE)
방어적 증류 defensive distilation[56]	DNN의 구현. 제안된 방법은 DNN 모델의 훈련 단계에 적용됨 제안된 방법은 훈련 모델의 소프트맥스 레이어에 사용되는 증류 온도distillation temperature라고 불리는 방법임
GAN 기반 방어	GAN은 회피 공격[62]에 대한 방어 메커니즘으로도 사용함 GAN 모델의 기본 아이디어는 판별자에 의해 감지되지 않는 새로운 이미지를 생성하고자 생성기를 사용할 때 판별자에게 잡히지 않도록 함을 목표로 함 APE-GAN 모델은 반대로, GAN 모델의 생성자에 적대적 이미지를 입력으로 제공하고, GAN 모델을 사용하여 이 공격받은 이미지를 공격받지 않은 형태로 변환함[103].
Gradient Regularization Gradient Masking	대부분의 적대적 공격은 모델 추론 과정에서의 경사gradient를 보고 공격이 이루어진다. 학습 모델의 경사가 출력으로 노출되는 것을 방지하는 것에 중점을 둔다. - Gradient Regularization: 모델의 경사를 일관된 형태로 유지(예: Bit Plane Feature Consistency (BPFC) regularizer, Second-Order Adversarial Regularizer (SOAR)) - Gradient masking: 출력에 노이즈를 추가하거나, 학습 중에 특정 부분을 제거함으로써 모델의 경사를 외부로부터 감춤(예: S2SNet)

책임성 투명성

요구사항 11 인공지능 모델 명세 및 추론 결과에 대한 설명 제공

- 인공지능 모델의 추론 결과만으로는 예측된 결과가 어떤 요소에 의해 도출되었는지 알기 어렵다. 또한, 시스템의 최종 결과를 얻고자 다수의 인공지능 모델이 사용된다. 이러한 과정에서 인공지능 모델의 예측 결과에 대한 사용자 신뢰를 확보하기 위해 사용된 모델 정보, 결과 도출 과정에 대한 설명 추론 결과에 대한 설명을 제공한다.

  * 사람이 인공지능 모델의 의사 결정 방식을 파악하도록 돕는 모델의 작동 방식에 대한 유용한 정보(예: 의사 결정 메커니즘, 의사 결정의 기초를 이루는 학습 데이터, 인공 신경망 내에서 사용된 변수와 가중치)

참고 AI 플랫폼 온라인 지원 사례

- 스마트 치안 시스템의 AI 모델과 관련된 몇 가지 주요 사양 및 개념:
  - 모델 아키텍처: AI 모델의 전반적인 설계 및 구조이다.
  - 학습 데이터: AI 모델은 추론 중에 접하게 될 데이터의 예가 포함된 대규모 데이터셋에서 학습하여야 한다.
  - 하이퍼파라미터: 학습 및 추론 중에 AI 모델의 동작을 제어하는 파라미터이다.
  - 추론 결과: 추론 중에 훈련된 AI 모델은 새로운 데이터를 가져와 학습된 지식을 기반으로 예측 또는 결정을 생성한다.
  - 정확도 및 성능 지표: AI 모델의 성능을 평가하는 데 사용된다.
  - 개인정보 보호 및 보안 고려 사항: 적절한 개인정보 보호 및 보안 조치가 마련되는지 확인함이 중요하다. 여기에는 데이터 암호화, 액세스 제어, 공격에 대한 예방 및 관련 규정 또는 표준 준수를 포함한다.

11-1 인공지능 모델의 명세를 투명하게 제공하는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템의 투명성을 확보하는 방안 중 하나는 인공지능 모델 또는 서비스의 개발, 테스트 및 배포 과정에서 발생한 다양한 결과를 문서로 작성하는 것이다.
- 스마트 치안 시스템은 이 문서가 사람들에게 결정에 대한 아이디어를 제공하고 시스템의 신뢰성을 확보한다. 획득된 모델을 설명하는 상세 문서를 작성하는 메커니즘을 구축하고, 모델의 목적, 입출력 정보, 성능, 편향성, 신뢰성 등의 결과를 사용자가 인공지능 모델과 관련된 정보를 요청할 때 투명하게 공개한다.
- 인공지능 모델의 주요 정보 및 구성 요소를 자세히 문서화함은 스마트 치안 시스템의 투명성 측면과 아울러 잠재적인 오류로 인해 영향을 받는 사용자의 이의 제기 시 추적성 측면에서도 중요하다.

11-1a 시스템 개발 과정과 모델 작동 방식에 대한 세부 정보가 설명된 문서를 작성하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템의 투명성을 높이고 시스템 사용자가 인공지능 기반 프로그램 구성 요소를 파악하는 정보 제공은 시스템 신뢰성을 높이는 데 중요한 요소이다. 이를 위해 인공지능 모델 개발 과정에서 모델의 명세를 작성한 모델 상세 문서 확보 시, 사용자에게 인공지능 시스템의 구성 요소를 파악하는 정보를 제공한다.
- 스마트 치안 시스템이 작동하는 방식을 설명하고자, 센서, 카메라 및 소프트웨어 알고리즘 등 주요 구성 요소에 대한 정보를 제공한다. 또한 악의적이거나 도망 중인 여행자, 무기, 보행자의 비정상적인 행동, 침입자, 화재 등 잠재적인 치안 위협을 감지하고 대응하는 방법, 다양한 환경(가정, 공공 또는 비즈니스)의 특정 요구사항을 충족하고자 사용자 정의 및 구성 방법 등을 설명한다.
- 이러한 내용을 포함하는 스마트 치안 시스템의 개발 프로세스와 모델 작동 방식을 자세히 설명하는 문서 내용을 다음과 같이 작성한다.

문서 섹션	문서 섹션 내용
개요	개발 중인 스마트 치안 시스템에 대한 개요와 그 필요성을 설명
요구사항	스마트 치안 시스템에 대한 특정 기능을 포함한 요구사항을 개요로 설명 예산이나 시간 제약과 같은 제약 사항도 고려 필요
설계	하드웨어 및 소프트웨어 구성 요소를 포함한 스마트 치안 시스템의 전반적인 설계 구조 설명 시스템의 아키텍처와 다양한 구성 요소(예: 센서, 레이더, 엣지 디바이스, CCTV 카메라 등)가 상호 작용하는 방식에 대해 함께 설명
구현	스마트 치안 시스템을 구현한 방법과 직면한 어려움 및 극복 과정을 설명 필요시, 구현을 설명하는 코드 조각이나 기술적인 다이어그램을 포함
테스팅	스마트 치안 시스템이 요구사항을 충족하는지 확인하고자 어떻게 테스트하였는지 설명 테스트 중에 발생한 문제와 그에 대한 대응 방안 설명 필요시, 해당 사례에 위치 정보나 날씨/예보 정보 등 추가
테스팅 결과	테스트 결과를 제시하고 스마트 치안 시스템의 결과를 평가 시스템의 향후 개선 사항이나 반복에 대한 권고 사항 제공
결론	문서의 핵심 요점을 요약하고 스마트 치안 시스템의 중요성 강조

- 위 내용 외에, 스마트 치안 시스템의 책임과 투명성을 유지하고자 IBM과 WEF[world economic forum]가 제안한 방법을 사용하거나 활용하여 AI 시스템의 투명성을 보장하는 문서를 작성한다[115][116].
  - ✓ 의사 결정 과정에 영향을 미치는 가중 요소(특징), 주요 가정, 사양 해석 시 주의 사항 및 이에 따른 치안 알고리즘의 위협 탐지 특성에 적합한 모델 사양 정보 등을 준비한다.

인공지능 모델 사양 정보 문서화 항목 예시[155]

모델 사양 구분	모델 사양 내용
모델 개요	모델 소유자, 성숙도 수준, 라이선스, 생성된 데이터 등
의도한 용도	기본 사용 사례, 보조 사용 사례, 사용자, 대응 지침, 윤리적 고려 사항 등
모델 상세	모델 디자인, 하이퍼파라미터 정보, 목적 함수, 공정성 제약 조건 등

모델 사양 구분	모델 사양 내용
활용한 데이터	데이터 소스, 인구 그룹, 변수, 전처리 방법 등
평가 정보	평가 데이터, 지표, 결과, 제한 사항 등
모니터링 정보	최근 평가 정보, 실패 정보, 버전 이름 등

문서화 작성 결과 검증 체크리스트 관점 예시[156]

문서 검증항목 구분	문서 검증항목 내용
완전성	정보가 누락되지 않음
간결성	사실 내용을 있는 그대로 효율적으로 전달함
관련성	모든 정보는 주제와 관련됨
뒷받침하는 증거	충분한 증거를 기반으로 주장을 지원함
어휘 선택	용어와 단어의 선택은 대상 독자에게 적합함
명확성	용어 및 기타 내용은 충분히 이해하며, 모호하거나 애매하거나 이해하기 어려운 부분은 없음
구조	논리적인 구조와 정보 흐름을 가짐
표현	표현 스타일(텍스트, 그래프, 표 등)은 콘텐츠에 적절함
탐색성	관심 있는 정보를 쉽게 찾도록 구성됨

11-2 사용자가 모델 추론 결과의 도출 과정을 수용할 수 있도록 근거를 제공하는가?

Yes ☐ No ☐ N/A ☐

- 사용자가 인공지능 모델의 추론 결과 및 인공지능 시스템의 동작을 신뢰하려면 시스템 사용자가 인공지능 모델이 제공하는 판단 혹은 추론 결과의 도출 과정을 이해할 수 있어야 하며, 사용자에게 이에 대한 설명 및 근거를 제시하는 것이 바람직하다.
- 스마트 치안 시스템에 대한 모델 추론 결과를 도출하는 프로세스에는 여러 단계가 포함되며 결과의 정확성과 신뢰성을 뒷받침하고자 섹션 10의 소개 부분에서 언급한 각 단계에서 증거를 수집하여야 한다.
  - ✓ 학습 데이터: 학습 데이터의 품질과 다양성은 정확성에서 중요하다. 데이터 적합성을 뒷받침하는 증거로서 수집 절차, 전처리 방법 및 소스 유효성 검증이 포함된다. 데이터의 구체적인 내용, 양, 다양성 및 품질에 대한 문서화는 적합성을 보장한다.
  - ✓ 모델 아키텍처와 하이퍼파라미터: 모델 구조와 하이퍼파라미터는 정확도에 영향을 미쳐, 선택된 내용, 설계 원칙, 구체적인 아키텍처 그리고 하이퍼파라미터 선택의 근거를 문서화함이 중요하다.
  - ✓ 모델 평가 지표: 정확도, FAR, 정밀도, 재현율, F1 점수 및 ROC 곡선 등 성능 지표는 정확성과 신뢰성을 검증한다. 평가 중에 이러한 지표를 문서화하여 그 효과를 입증한다.
  - ✓ 테스트와 검증: 획득한 데이터를 사용하여 모델 테스트 및 검증에서 얻은 절차, 테스트 데이터 및 결과를 증명한다. 이러한 교차 검증 프로세스는 정확도와 신뢰성을 강화한다.
  - ✓ 실제 배포: 스마트 치안 시스템에 모델을 배치한 결과, 실제 시나리오의 성능, 사례 연구, 사용자 피드백 및 보안 직원의 참여로 실용성과 효과성이 입증된다.

✓ 보안 조치: AI 모델과 스마트 치안 시스템에 구현된 보안 조치의 문서화는 신뢰성을 보장하며 잠재적인 위협이나 취약성을 보호한다.

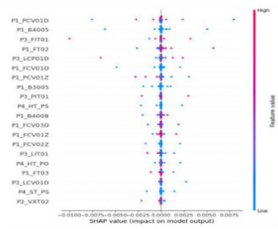
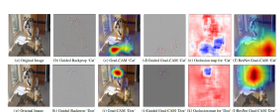
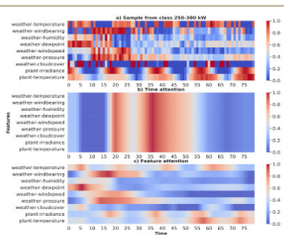
- 스마트 치안 시스템에 대한 모델 추론 결과를 도출하는 과정에서 투명성과 책임성을 보장하고자 증거를 철저히 문서화하고 검증함이 중요하다.
- 설명 방법의 하나로 인공지능의 행동, 추론 결과의 이면에 있는 논리를 사용자에게 설명하여 이해하기 쉽게 함[104]을 고려하는데, 이와 관련하여 섹션 10-1a에 제시된 내용을 참고하도록 한다.
- 그러나 개인이 이해하기 어려운 설명은 추론 결과에 부정적인 영향을 미치므로 기술 적용 시 전문가와 상담 및 충분한 논의가 필요하다.
- 또한, AI 모델의 추론 결과의 근거를 항상 설명하는 것은 아니므로, XAI 기술의 적용 이외의 대안을 사용하여 AI 시스템의 투명성을 확보하여야 한다. 따라서 XAI 기술의 적용 가능 여부를 고려하여 이 세부 요구사항의 검증항목을 선택적으로 적용하여야 한다.

11-2a 인공지능 모델에 적합한 XAI(eXplinable AI) 기술을 적용하였는가?

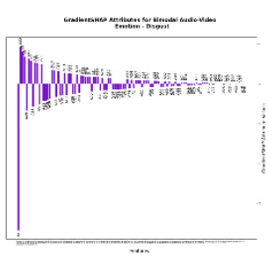
Yes ☐ No ☐ N/ ☐

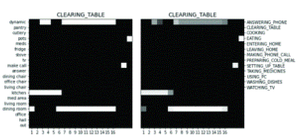
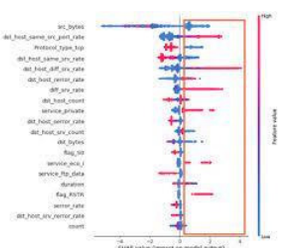
- 심층 학습 기술을 활용한 인공지능 시스템은 성능이 우수하지만 설명 가능성이 낮다. 이럴 때 모델 추론 결과에 대한 확신과 시스템 전체에 대한 신뢰도가 낮아지므로 사용자가 모델 추론 결과를 수용하는 근거를 확보하여야 한다.
- 대표적인 XAI 기술로 모델 비종속적인 설명 방법을 제공하는 LIME 지표나 모델 종속적인 설명 방법을 제공하는 LRP 지표 등을 이용한다. 다만, XAI 기술은 연구가 계속 진행 중이므로 기술 도입 전에 적합한 기법 선택이 중요하다.
- 스마트 치안 시스템의 투명성은, 텍스트 또는 시각화 등 다양한 접근 방식을 통한 XAI 기술을 사용하여 보장한다. 또한, 인공지능을 더 투명하고 신뢰하게 하려면 다음 사항을 고려한 XAI 기술 활용이 필요하다.
  - ✓ 스마트 치안 시스템은 이미지, 비디오, 생체 정보 등 다양한 데이터를 처리한다. XAI 기술을 도입하려면, 이런 다양한 데이터 유형에 맞게 적절한 기술을 찾아야 한다.
  - ✓ XAI 기술은 스마트 치안 시스템이 어떻게 결정을 내리는지 설명해 준다. 그러나 모든 치안 시스템에 동일한 방식으로 사용되는 표준 해결책은 없다. 어떤 기술을 사용할지는 시스템의 종류, 처리하는 데이터, 사용자의 요구사항 등에 따라 달라지며, 신중한 평가가 필요하다.
  - ✓ XAI 기술의 도입은 치안 시스템의 투명성과 신뢰성을 높이지만, 적용과 평가 과정에서 전문가와 상담 및 검토가 필요하다. 법적·윤리적, 개인정보 보호와 관련된 사항도 고려하여야 하므로 전략적인 접근이 필요하다.

참고 스마트 치안 분야 데이터 타입에 따른 XAI 적용 사례

데이터 종류	XAI 적용 사례
센서 데이터	 〈출처: E-SFD: Explainable Sensor Fault Detection in the ICS Anomaly Detection System, 2021〉	인공지능 기술의 의사 결정 과정은 아직 '블랙박스'로 남음 – 이 연구에서는 XAI 기술을 사용하여 센서 측정치와 데이터를 검토함 – 이 주석을 사용하여, 치안 전문가들은 이상을 감지한 후 응답 과정에서 이상을 일으킨 기능을 한 센서를 식별함 – 이런 방식으로, XAI 기술은 시스템 결과에 대한 선언적인 설명을 제공하는 데 유용함[105]
이미지 데이터	 〈출처: Grad-cam: Visual explanations from deep networks via gradient-based localization, 2017〉	연구에서는 적용된 XAI 방법들이 검출/분류 예측 결과에 어떤 영향을 미쳤는지 확인함 – 왼쪽 그림에서 보여 주듯이, 치안 시스템 모델이 예측하고자 사용한 영역을 시각화하여 설명을 제공받음 – 이는 사용자가 모델 예측의 근거를 이해하는 데 도움이 됨* [106] * 시각화된 예측의 투명도 값을 사용하거나 개인의 얼굴이나 신용카드 스냅샷 등의 개인 및 민감한 정보를 시각화할 때 개인정보 보호 문제를 반드시 고려하여야 함
시계열 데이터	 〈출처: Explainable deep neural networks for multivariate time series predictions, 2019〉	연구에서는 태양광 발전소에서 얻은 다변량 시계열 데이터의 특징을 시각화하고자 XAI 기술을 사용함[107] – 이 기술을 이용하여 연속/시계열 데이터로 얻은 데이터/특징의 예측 결과 시각화가 가능함

참고 스마트 치안 시스템에서 고려하는 XAI 기법 또는 알고리즘 예시

XAI 알고리즘	설명
Gradient SHAP[108]		PyTorch용 오픈 소스 XAI 라이브러리인 Captum의 제공 기능 – Gradient SHAP에서는 각 입력 샘플에 다양한 가우시안 노이즈를 추가한 다음, 기준선과 입력 사이에서 무작위 점을 선택하고, 선택한 무작위 점을 기준으로 그라디언트 값을 계산함 – 이 알고리즘에 따르면, 입력 특징은 독립적으로 간주하며, 모델 함수의 설명은 기준선과 입력 사이에서 선형임

XAI 알고리즘	설명
Grad CAM[109]	 시각화 목적으로 GradCAM 방법을 적용하면, 딥러닝 분류기에 의하여 X의 분류에 기여하는 각 픽셀을 보여주는 히트맵을 얻음 - 히트맵의 각 픽셀은 인터뷰 상황의 예측에 중요한 부분으로서 분류의 중요성(즉, 관련성)에 따라 [0], [1] 범위에서 값들을 가짐
LIME[110]	 LIME은 이미지뿐만 아니라 텍스트(자연어 생성(NLG)) 데이터에도 입력을 지원함 - LIME은 치안 시스템 모델을 포함한 복잡한 모델에 대한 로컬 설명을 생성하는 데 널리 사용되는 기술임 - 이 기술은 더욱 단순하고 해석 가능한 모델을 사용하여 특정 예측을 둘러싼 로컬 영역에서 보안 모델의 동작을 근사화하는 방식으로 작동함 - LIME은 치안 시스템 모델이 특정 결정을 내리는 데 영향을 미친 특징이나 속성에 대한 통찰을 제공함
SHAP[111]	 SHAP는 특정 변수를 위한 모든 변수 조합을 입력하고 결과를 비교하여 특정 변수가 예측에 얼마나 기여하는지 결정함 - 기여도가 높은 변수를 식별할 때는 LIME과 비슷하지만, 데이터에 가중치를 부여한다는 점에서 다름 - 좌측 그림에서, 파란색과 빨간색은 예측 결과에 대한 높음(파란색)과 낮음 기여(빨간색)를 나타내며, 막대의 크기는 영향의 정도를 나타냄 - SHAP는 개인 그룹 간에 값을 공정하게 분배하는 방법을 제공하는 게임 이론적 접근법이며, 치안 시스템 모델을 포함한 기계 학습 모델의 예측을 설명하는 데 적용됨 - SHAP 값은 특정 예측에 대한 각 특징 또는 속성의 기여도를 설명하는 데 사용하며, 모델의 전체적인 동작을 설명하는 데 사용됨[112]

11-2b XAI(eXplainable AI) 기술 적용이 불가능한 경우, 기술 외 대안을 마련하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 모델의 추론 결과 및 결정의 근거를 설명하는 것이 항상 가능한 것은 아니다. 특히 스마트 치안 분야에서는 개인정보 보호와 안전 문제도 함께 해결하여야 하므로, 설명 가능한 인공지능 기술에 설명이 항상 충분하지는 않을 수 있다[113].
- 투명성은 사용자와 개인이 시스템의 결정을 신뢰하게 하는 핵심 요소이다. 인공지능의 투명성은 보이는 부분만으로 제한되는데, XAI는 인공지능 시스템의 작동 방식을 쉽게 설명하려고 노력한다. 그러나 이것은 사용자의 개인 정보를 침해하는 설명을 제공하거나, 민감한 정보에 대한 이해가 어려운 복잡한 설명을 할 수도 있다[114].
- 설명을 더 의미 있게 만들려면 추가적인 도구와 방법, 예를 들어 규칙 기반 시스템, 의사 결정 트리, 베이지안 네트워크, 퍼지 논리 등이 필요하다. 복잡한 상황을 다루려면 전문 시스템도 고려하지만, 비용이 많이 들므로 예산을 신중히 고려하여야 한다.
- 스마트 치안 시스템에 어떤 XAI 기술을 선택할지는 시스템의 특별한 요구사항, 예를 들어 결정 과정의 복잡성, 사용 가능한 데이터의 양과 품질 그리고 비용과 자원 등에 따라 달라진다.

11-3 모델 추론 결과에 대해 사용자의 판단을 도울 수 있는 설명을 제공하는가?

Yes ☐ No ☐ N/A ☐

- 사용자에게 인공지능 모델의 추론 결과에 대한 설명을 제공하면, 사용자는 단순히 해당 인공지능 모델의 최종 결과뿐 아니라 그 결과가 도출된 수치적인 근거로 확률값, 불확실성(uncertainty) 등을 제공받는다. 이러한 정보는 사용자의 의사 결정에 도움이 되지만, 오히려 사용자의 혼란을 유발하므로, 정보 제공의 필요성을 사전 검토가 필요하다.

11-3a 모델 추론 결과에 대한 설명이 필요한지 검토하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템이 도출한 결과에 대한 설명 제공은, 사람들이 인공지능을 활용하여 의사 결정하는 데 도움이 되지만, 오히려 방해도 된다. 따라서 모든 경우에 추론 결과에 대한 설명을 제공하기보다는, 설명이 꼭 제공되어야 하는지를 확인하는 과정이 선행되어야 한다.
  - ✓ 투명성: 스마트 치안 시스템이 작동하는 방식과 결정의 근거에 대해 투명하여야 한다. 모델 추론 결과에 대한 설명을 제공함으로써 사용자/개인/시스템에 영향받는 사용자가 시스템 동작을 이해하고 신뢰하도록 도움을 줌

✓ 책임성: 스마트 치안 시스템이 잘못된 결정을 내렸을 때, 왜 그런 결정을 내렸는지 설명하는 것이 중요하다. 모델 추론 결과에 대한 설명을 제공함으로써 시스템 내의 오류나 편향을 식별하고 해결하는 데 도움됨

✓ 컴플라이언스: 개발된 시스템에 따라(어디에서 어떤 유형의 위협 감지 및 누구에게 영향을 미치는지) 인공지능 시스템이 내린 결정에 대한 설명을 제공하는 규제 요구사항이 있다. 이를 준수하지 않으면 법적 및 재정적인 처벌을 받음

✓ 사용자 신뢰: 스마트 치안 시스템의 모델 추론 결과에 대한 설명을 제공하여 사용자/개인과의 신뢰 구축에 도움이 된다. 사용자는 시스템이 어떻게 작동하고 왜 특정 결정을 내리는지 이해할 때 더 신뢰하는 경향임

✓ 시스템 개선: 모델 추론 결과에 대한 설명을 제공하여, 개발자들은 개선할 부분을 식별하고 시스템을 변경하여 성능을 향상하고 오류를 줄임

- 스마트 치안 시스템의 모델 추론 결과에 대한 설명을 제공하는 것은 투명성, 책임성, 준수, 사용자/개인 신뢰 및 시스템 개선을 보장하고자 중요하다. 올바른 판단을 내리도록, 인공지능 모델의 추론 결과에 대한 설명을 제공하지 않는 편이 더 나을 때 두 가지 예시는 다음과 같다.

  ✓ 첫째, 모델의 추론 결과에 대한 설명 제공 자체가 사용자의 의사 결정에 크게 영향을 미치지 않는다고 판단될 때다. 설명 제공으로 인해 미치는 영향을 명확하게 분석하지 않을 때, 자세한 설명을 제공하면 사용자의 의사 결정에 더 도움이 된다고 생각하겠지만, 예상과는 다르게 혼란을 초래한다. 예를 들어, 인공지능 시스템이 도출한 두 가지 결과가 있고, 각각의 예측 확률이 85.8%, 87.0%라면, 사용자는 어떤 결과를 활용하여 의사 결정을 할지 혼란스러워한다.

  ✓ 둘째, 예측 확률이 너무 높거나 낮을 때도 모델의 추론 결과에 대한 자세한 설명을 제공하지 않는 것이 낫다. 만약 시스템의 출력 결과에 대해 신뢰도가 100%라고 사용자에게 알릴 때, 사용자가 시스템의 출력 결과를 맹목적으로 수용하게 한다.

11-3b 사용자에게 인공지능 모델 추론 결과에 대한 설명을 제공하였는가?

Yes ☐ No ☐ N/A ☐

✓ 각 모델의 추론 결과가 참값과 일치할 확률을 계산하면, 이를 모델의 최종 의사 결정에 대한 설명으로 사용할 수 있다. 확률 변수의 분산 크기로, 인공지능 모델이 도출한 결과를 얼마나 확신하는지를 나타내는 불확실성을 모델 추론 결과에 대한 설명으로 고려해 볼 수 있다.

설명 제공 방안	설명 내용
텍스트 기반 설명	모델이 결정에 이르기까지 어떻게 도달하였는지 설명하는 텍스트 설명을 제공 예를 들어, 시스템의 센서가 감지한 특정 객체나 움직임 등 주요 요소를 설명하는 문장이나 단락을 제공함
시각화[149]	모델이 결정에 도달하는 과정을 사용자에게 시각화 기법을 사용하여 제시 예를 들어, 모델이 결정을 내리고자 주목한 입력 이미지의 영역을 강조하는 GradCAM 기법과 같은 히트맵을 표시하거나, 모델이 사용하는 다양한 특징 간 관계를 나타내는 그래프를 표시함

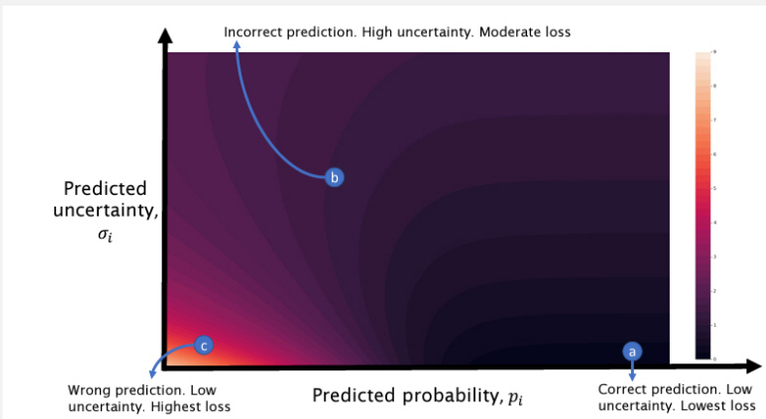
설명 제공 방안	설명 내용
상호 작용 기반 설명	상호 작용 설명을 통해 사용자들은 모델이 어떻게 결정에 이르렀는지 실시간으로 탐색함 예를 들어, 사용자들이 모델에 의해 사용되는 매개 변수를 조정하고 결정이 어떻게 변하는지 확인하도록 허용함
자연어 생성natural language generation, NLG[150]	자연어 생성 방안은 모델의 의사 결정 과정에 대한 사람이 읽는 설명을 자동으로 생성하는 기술(예: LIME 기술) 예를 들어, 시스템은 "시스템이 뒷마당에서 수상한 움직임을 감지하여 경보가 작동되었다."처럼 경보가 작동된 이유를 설명하는 문장을 생성함

- 그러나, 결정의 논리에 대한 타당한 증거를 제시하지 않고 모델 추론만 제시하는 것은 사용자, 개인, 비즈니스 제공자 등이 편향되었다고 비난받는다(06-3a의 COMPAS 사례). 이 문제를 해결하고자 인공지능 모델의 추론 결과에 대한 확률값을 측정하고, 추론 결과의 확신도를 나타내는 불확실성을 양적으로 측정하고 설명하는 것을 고려하여야 한다. 다음은 추론 결과의 확률과 불확실성에 따른 설명의 예시이다.

추론 확률(0~1)	불확실성(0~1)	설명 예시
0.98	0.01	98% 정확도 및 1% 불확실성으로, 거의 확실하게 침입을 탐지하였다.
0.98	0.90	98% 정확도로 침입을 탐지하였으나, 불확실성이 90%로, 사용자의 확인이 필요하다.

참고 사람 감지기의 불확실성 추정에 대한 연구 사례

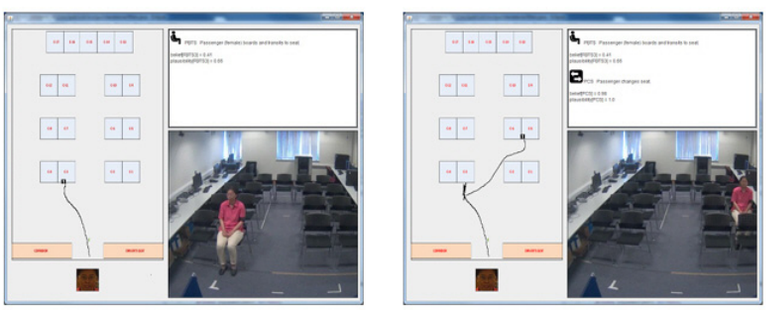
추론 결과의 불확실성을 설명하고자 다양한 연구가 진행된다.
다중 인스턴스 학습 프레임워크 내에서 프레임별 사람 감지 모델을 활용하여 시공간 행동 감지기를 훈련하는 지도 기반 학습 방법으로, 이 접근 방식은 불확실성 기반 손실 함수를 사용하여 레이블 노이즈와 불확실성 문제를 해결하고자 한다.
또한, 연구자들은 이 방법이 네트워크에 의해 만들어진 불확실성 예측을 통합하여 가방의 소음과 표준 다중 인스턴스 학습(MIL) 가정의 위반을 더 잘 다루도록 한다고 언급한다.



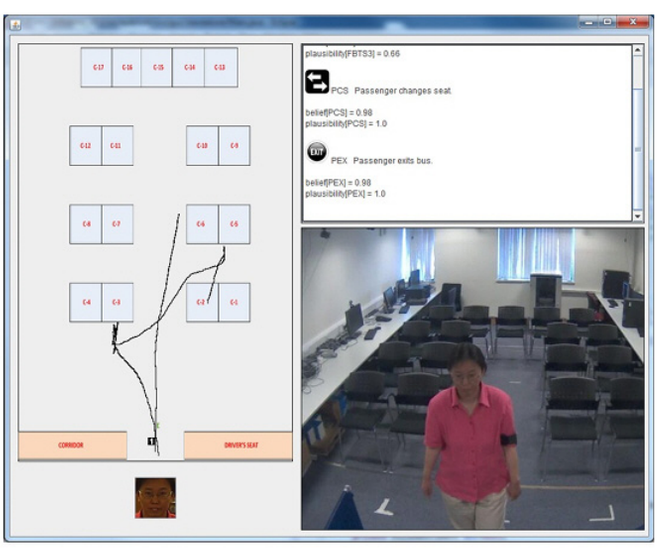
〈출처: Uncertainty-aware weakly supervised action detection from untrimmed videos, 2020[117] 〉

참고 비디오 감시 시스템의 불확실성 추정에 대한 연구 사례[118]

이 연구에서 연구자들은 Dempster-Shafer evidence 이론을 사용하여 스마트 교통 비디오를 감시하는 새로운 이벤트 인식 프레임워크를 조사한다.
추론 결과의 불확실성은 추론된 사건의 정확성과 신뢰성에 중대한 영향을 미치며, 이는 결국 의사 결정 과정에 영향을 미친다. 제안하는 프레임워크는 DS 이론을 사용하여 불확실성을 관리하고 신뢰도를 수치로 표현함으로써 이 문제를 해결하고자 하나, 추론 결과의 신뢰성은 증거를 수집하는 데 사용되는 감지 및 추적 시스템의 정확성과 신뢰성 그리고 시스템을 통해 증거를 전파하는 추론 과정에 따라 달라지는 것으로 확인된다.



(a) Instance 1 (b) Instance 2



(c) Instance 3

〈출처: Evidence reasoning for event inference in smart transport video surveillance, 2014[158]〉

04 시스템 구현

다양성 존중

요구사항 12 인공지능 시스템 구현 시 발생 가능한 편향 제거

- 인공지능 시스템 구현 단계에서 편향을 고려하지 않는다면, 시스템 설계자 또는 개발자의 배경지식이나 편견, 특정 인구 통계의 대표성 및 알고리즘 설계 선택 등으로 인공지능 시스템이 편향된다. 이러한 편향은 불공정한 결과와 시스템의 정확도 감소 등으로 이어진다. 따라서 발생 가능한 편향을 식별하고 이를 제거하는 방안을 고려하여 설계한다.

12-1 소스 코드 및 사용자 인터페이스로 인한 편향을 제거하기 위해 노력하였는가?

Yes ☐ No ☐ N/A ☐

- 스마트 치안 분야 인공지능 시스템은 복잡한 알고리즘과 다층 아키텍처로 인해 시스템 편향이 발생한다.[119].
- 인공지능 시스템과 인터페이스 및 상호 작용 측면에서 개발자 통찰력, 사용자 기대치, 영향을 받는 인구 통계, 경험 및 환경 콘텍스트 등의 관점에서 표현 편향presentation이나 순위 편향ranking bias 등이 발생하는지를 미리 확인하여 편향을 방지하도록 시스템을 설계하는 것이 바람직하다.
- 그 외에 인공지능 시스템의 구현 단계에서 편향을 방지하고자 작성된 코드를 주기적으로 검토하여 코드 구현 과정에서 특정 클래스 접근이 누락되지는 않았는지, 개발자의 편견이 코드에 반영되지는 않았는지 등을 확인하여야 한다.

참고 요구사항 프로세스에 사용되는 AI 시스템의 결과 편향으로 인한 실패/보고서

1. 2018년 중국 AI 교통 캠 케이스

• 중국의 얼굴 인식 시스템은 버스의 광고 포스터에서 유명한 사람의 얼굴을 무단 횡단으로 감지하였다.
  ◦ 얼굴 인식 시스템을 기반으로 한 응용프로그램은 규칙 위반 목록에 버스의 광고판에 사람을 게시하였다.
  ◦ 개발된 모델의 인지적 편향이 이 문제를 일으켰다.
  ◦ 경찰서는 모델이 업데이트되었다고 선언하였다.

2. 편견으로 인한 잘못된 얼굴 인식 2020

• 이 사건의 가능한 원인은 모델 학습 과정에서 불균형 데이터셋 사용 때문이다.
  ◦ 디트로이트 경찰서에서 사용하는 안면 인식 서비스는 사람들의 피부색을 차별한다.
  ◦ 강도 사건을 수사하는 동안 경찰서는 얼굴 인식 시스템을 사용한다.
  ◦ 편향된 결과가 발견된 후, TEANECK NJ- 타운십 의회는 편향 발생으로 인해 감시 비디오 영상에 얼굴 인식 기술 사용을 금지하였다[120].

12-1a 데이터 접근 방식 구현과정 등 소스 코드에서의 편향 발생 가능성을 확인하였는가?

Yes ☐ No ☐ N/A ☐

• 인공지능 시스템은 모델에서 활용한 데이터에 접근하는 방식이 코드상에 구현되는 과정에서 특정 클래스의 접근이 누락하는 등 다양한 형태의 편향이 발생한다.

• 인공지능 시스템이 의사 결정 과정에서 사용하는 정보(데이터, 클래스 등)를 사용하려고 하거나 모델 구축 시에 사전 경험, 규칙, 분석 등의 경험이 반영되어 의사 결정이 이루어지면 인지 편향과 확인 편향을 일으킨다. 이러한 편향을 줄이고자 다양한 배경과 경험을 지닌 전문가들의 검토를 통한 시스템 평가가 필요하다.

• 데이터 접근이나 모델의 편향을 파악하고 완화하는 작업은 시스템 구현 시 반영하여야 하며, 개발 중인 스마트 치안 시스템에 다음 접근 방법 중 적합한 방법을 선택할 수 있다.

✓ 스마트 치안 시스템에서 데이터 접근 방식의 편향을 식별하고자 사용된 데이터를 분석하여 특정 그룹의 과대 표현이나 과소 표현 등의 편향 원인을 찾고 해결하는 접근법이 필요하다(예: 11-1의 결함 있는 안면 인식 사례).

✓ 시스템의 광범위한 테스트와 검증 수행이 한 방법이다. 이것은 다양한 실제 상황에서 시스템을 테스트하고 결과에서 발생하는 편향 사례를 분석함을 포함한다. 시스템의 결과가 지속해서 편향된 결과를 산출한다면, 데이터 접근 방식의 구현에서 편향을 의미한다.

- 인공지능 시스템 설계 및 개발 단계에서 발생한 편향을 확인하고자 오픈 소스 도구(예: FairML, Google What-If Tool, ML Fairness Gym, IBM의 AI 360 Fairness, Aequitas, FairLearn)를 활용한다. 이러한 도구들은 주기적으로 출력 데이터의 통계를 분석하여 알려지지 않은 편향을 발견하거나, 미리 지정한 공평성 평가 지표에 따라 기능의 위험 여부를 알리는 등의 기능을 수행한다. 이 도구들을 활용함으로써 구현 과정에서 편향을 빨리 발견하고 대응하게 한다.

12-1b 사용자 인터페이스user interface 및 상호작용interaction 방식으로 인한 편향을 확인하였는가?

Yes ☐ No ☐ N/A ☐

- 시스템 설계 과정의 다양한 단계에서 편견이 도입됨을 이해하는 것이 중요하다. 이는 데이터 수집, 알고리즘 개발뿐 아니라, 사용자 인터페이스 디자인을 포함한 과정에서 발생한다. 스마트 치안 시스템에서 사용자 인터페이스와 상호 작용 방법으로 인한 편견을 식별하는 중요성은 다음과 같다.

  ✓ 사전 훈련된 인공지능 모델의 편향으로, 사용자 인터페이스나 상호 작용 방식에 편향성이 있다면, 시스템은 사용자 입력을 정확하게 해석하지 못하고, 편향을 증폭시켜 잘못된 결정과 효과적이지 않은 치안 조치로 이어짐

  ✓ 스마트 치안 시스템에서 사용자 인터페이스와 상호 작용 방법으로 인한 편견의 식별은 법적 및 윤리적 측면에서 중요함. 많은 국가에서는 다양한 특성을 기반으로 한 차별을 금지하는 법과 규정이 있으며, 편견이 있는 시스템은 법적 및 재정적인 조치를 받음. 또한, 윤리적 고려 사항에 따라 시스템은 배경이나 피부색과 관계없이 모든 사용자의 안녕과 안전을 보장하고자 설계되어야 함.

- 사용자의 상호 작용 편향을 방지하려면 사용자의 인터페이스 설계 및 구현 시에 편향 발생 가능성이 있는 요소(예: 표현 편향, 순위 편향)를 미리 인식하여 제거하여야 한다.

  ✓ 표현 편향: 정보가 제시되는 방식에 따라 발생하는 편향이다. 사용자들은 가장 두드러지는 내용에 주된 관심을 표현하는 경향이 있다. 예를 들어, 국경 치안관이 여행자/이민자의 프로필, 출신 지역, 인종, 범죄 기록, 사회 경제적 지위를 검토할 때, 인터페이스의 약간 불투명한 구성 요소의 정보를 간과하여 무의식적으로 국경 허가를 받은 여행자/이민자의 실제 허가 상황과는 관계없이 그들을 무시하기도 한다.

  ✓ 순위 편향: 정보가 노출되는 순서에 기반한다. 사용자들은 상위 결과가 가장 관련성이 높고 중요하다고 생각하여 하위 정보를 확인하지 않고 결정한다.

참고 사용자에게 알림을 보내는 치안 시스템 인터페이스 예시[121]

• 감시 시스템의 인터페이스
  ✓ 좌측의 얼굴 인식 결과를 표현할 때, 인종, 성별, 나이대 등과 무관하게 출력하여 시스템 사용자가 민감한 특성에 대해 편향된 의사 결정 또는 편견을 가지지 않도록 함

책임성 투명성

요구사항 13 인공지능 시스템의 안전 모드 구현 및 문제 발생 알림 절차 수립

- 스마트 치안 시스템에서는 비디오 영상, 접근 기록, 생체 인식 데이터 등 민감한 데이터를 수집하고 처리하므로 데이터 무단 엑세스, 도난 또는 오용 등의 문제가 빈번하게 발생한다. 따라서 스마트 치안 시스템에 안전 모드를 구현하고, 문제 발생 알림 절차를 수립하여 발생 가능한 문제 상황에 대비한다.

13-1 공격, 성능 저하 및 사회적 이슈 등의 문제 발생 시 대응 가능한 안전 모드를 적용하는가?

Yes ☐ No ☐ N/A ☐

- 많은 국가와 산업에는 기업이 개인 데이터를 보호하고 보안 위반을 방지하는 보안 조치를 구현하도록 요구하는 규정이 있다. 스마트 치안 시스템에서 안전 모드를 구현하려면 시스템 사용과 관련된 위험을 예방하거나 완화하는 오류 방지 메커니즘을 만드는 것을 포함하여야 한다. 안전 모드는 의도하지 않거나 악의적인 동작을 방지하고자 시스템이 제한된 용량으로 작동하거나 특정 기능에 대한 액세스를 제한하도록 설계된 상태이다.
- 스마트 치안 시스템에서도 외부의 공격, 인적 오류human error, 인공지능 모델의 성능 저하, 편향 발생으로 인한 사회적 물의, 사고 등이 예상될 때, 발생 원인을 사용자가 파악하고 해결하거나 사용자에게 정상적인 기능으로 복구하는 방법을 제시하여야 한다. 이러한 대처 방법이 작동하는 상태를 안전 모드라고 한다.
- 개인에게 영향을 미치는 시스템 오류는 의사 결정 시 해당 분야 전문가의 기술적 안전 조치가 동반되어야 한다. 자문 법률 전문가 또는 전문가는 제18조 및 제62조에 따라 고도로 개인정보 및 생물 정보 데이터를 보호하는 것이 좋다. 구현된 안전 모드 및 문제 통지 절차와 관련된 일부 일반 지침은 다음과 같다.
  - ✓ 안전 모드 정의: 여기에는 문제가 발생하거나 보안 위협을 감지할 때 등 특정 상황에서 시스템이 수행하여야 하거나 수행하지 말아야 하는 특정 동작 또는 작업을 정의하는 것이 포함된다. 안전 모드는 사용자나 치안 시스템 자체에 대한 잠재적인 피해를 보호하도록 설계되어야 한다.
  - ✓ 안전 모드 구현: 안전 모드가 정의되면 스마트 치안 시스템의 코드에 구현되어야 한다. 여기에는 안전 모드를 유발하는 특정 상황에 대한 특정 응답으로 시스템 알고리즘이나 프로그래밍을 수정하는 것이 포함된다. 안전 모드를 철저하게 테스트하여 의도한 대로 작동하는지 확인하는 것이 중요하다.
  - ✓ 모니터링 및 업데이트: 스마트 치안 시스템은 문제나 보안 위협의 징후가 있는지 정기적으로 모니터링하여야 한다. 문제 통지 절차는 효과적이고 효율적인지 확인하고자 정기적으로 검토하고 업데이트하여야 한다.

참고 ‘안전 모드’ 함수의 구현 예시

- 안전 모드 기준 정의: 먼저 개발된 치안 시스템에 대한 안전 모드 기준을 정의한다. 이를 위해 중요하고 핵심적인 기능 목록을 유지하여야 한다. 예를 들어 시스템을 활성화 및 비활성화하거나, 경보가 울릴 때나 카메라가 움직임을 감지하였을 때, 시스템이 비활성화되었을 때 등 다양한 이벤트에 대한 알림을 받는 기능 등이 있다. 이 예제에서는 치안 시스템 구성 요소 중 하나인 출입 통제 하위 시스템, 비디오 감시 시스템 또는 침입 탐지 시스템 등에서 하드웨어 또는 소프트웨어 오류를 감지할 때 안전 모드가 트리거되는 것으로 가정한다. 또한 안전 모드는 문제가 있는 하위 시스템을 비활성화하고 백업 시스템을 활성화하는 것으로 가정한다.
- 안전 모드 메커니즘 설계: 안전 모드 메커니즘은 구성 요소 오류 발생 시 시스템이 따라야 할 일련의 규칙과 절차로 구현된다. 예를 들어 출입 통제 하위 시스템이 오류를 발생시키면 시스템은 백업 출입 통제 시스템으로 전환하거나 수동 출입 통제 절차를 요구한다. 안전 모드 메커니즘은 또한 시스템 관리자나 유지 보수 팀에 오류를 처리하도록 통보를 트리거할 수 있다.
- 안전 모드 테스트: 안전 모드를 테스트하고자 시스템은 문제가 있는 구성 요소(예: 잠금이 열리지 않는 문 또는 녹화를 중단한 카메라)와 함께 시뮬레이션 된다. 안전 모드 메커니즘은 문제를 감지하고 백업 시스템으로 전환하거나 수동 절차로 전환하여야 한다. 시스템 관리자는 또한 오류와 안전 모드 활성화에 대한 통보를 받아야 한다.
- 예시
  - 불법 사용자가 스마트 치안 시스템이 설치된 건물에 접근하려고 할 때, 시스템의 얼굴 인식 알고리즘은 해당 개인이 인가되지 않았음을 감지하고 안전 모드를 트리거한다. 안전 모드 프로그램은 출입 통제 하위 시스템을 비활성화하고 해당 개인의 건물 입장을 방지한다. 동시에 시스템 관리자는 보안 위반과 안전 모드의 활성화에 대한 통보를 받는다.
  - 스마트 치안 시스템이 건물 내의 문 중 하나가 제대로 잠기지 않는 것을 감지할 때, 출입 통제 하위 시스템은 안전 모드를 트리거한다. 안전 모드 메커니즘은 문의 잠금을 실패할 때 해당 문 잠금을 비활성화하고 백업 잠금으로 전환하거나 수동 출입 통제 절차를 요구한다. 동시에 시스템 관리자는 문제와 안전 모드 활성화에 대한 통보를 받는다.

13-1a 문제 상황에 대한 예외 처리 정책이 마련되어 있는가?

Yes ☐ No ☐ N/A ☐

- 시스템에 문제가 발생하는 상황에서 기능 정지, 화면 전환 및 서비스 제공 초기 상태로 복구, 입력 거절, 의사 결정 회피 등의 예외 처리가 이루어지는지 확인하여야 한다.
- 이러한 예외 처리가 이루어질 때, 인공지능 시스템 사용자에게는 시스템 운영이 적절하지 않은 이유와 시스템의 대응에 대하여 설명을 제공하여야 한다.

참고 스마트 치안 시스템의 예외 처리 정책 예시[122][123]

조직 또는 시스템이 지원하도록 예외 처리 정책을 구현함으로써, 문제 상황에 빠르고 효과적으로 대응하며, 사고의 영향을 최소화하고 가능하면 빨리 시스템을 정상 작동 상태로 복구되도록 보장한다.

문제 상황 핵심 요소	예외 처리 정책
사건 보고 및 에스컬레이션escalation	사건이 어떻게 보고되고 에스컬레이션되어야 하는지를 정의한다. 누구에게 알려야 하는지, 어떤 정보를 제공해야 하는지 그리고 사건이 얼마나 빨리 해결되어야 하는지 등이 포함된다.
사건 분류 및 심각도	사건의 심각성과 시스템에 미치는 영향을 기반으로 사건을 분류하는 분류 체계를 정의한다. 이를 통해 사건 해결을 우선순위에 따라 진행하고 가장 중요한 문제를 먼저 해결하여야 한다.
안전사고 대응 절차	사건에 대응하는 절차를 정의한다. 사건의 영향을 완화하고자 취하여야 할 단계, 시스템을 정상 운영으로 복구하는 방법 그리고 앞으로 유사한 사건이 발생하지 않도록 예방하는 안내 방법을 포함한다.
의사소통 및 알림	사건 발생 시 의사소통의 처리 방안을 정의한다. 누구에게 알려야 적절한지, 어떤 정보를 제공해야 하는지 그리고 얼마나 자주 업데이트를 제공하여야 하는지 등이 포함된다.
사고 검토 및 분석	사건 검토 및 분석 절차를 정의하여 사건의 근본 원인을 파악하고 대응 효과를 평가하며 개선할 기회를 식별하여야 한다.

13-1b 인공지능 데이터 및 모델 공격에 대해 시스템 측면의 방어 대책을 마련하였는가?

Yes ☐ No ☐ N/A ☐

- 06-2 및 10-1 에서 언급한 적대적 공격 외에도, 인공지능 시스템은 데이터 및 모델을 대상으로 하는 다양한 공격에 노출될 수 있다. 따라서, 시스템 구현 단계에서 대처 가능한 방안을 검토 및 적용하는 것이 바람직하다.
- 시스템을 통한 데이터 및 모델 공격의 유형으로는 데이터 중독 공격data poisoning attack, 모델 추출 공격model extraction attack, 모델 전도 공격model inversion attack 등이 있다. 각 공격에 대한 설명은 다음 표에 정리하였다.

인공지능 데이터 및 모델 대상 공격 예시

공격 방법	설명
데이터 중독 공격	사용자의 입력을 통해 모델이 재학습되는 경우에, 인공지능 서비스 운영 과정에서 의도적으로 학습 데이터를 변질시켜 서비스의 정상적인 기능을 손상시키는 공격이다. 학습 데이터를 오염시킨다는 의미로, 데이터 오염 공격이라고도 한다.
모델 추출 공격	공격 대상 모델의 입력값과 결괏값을 분석하여 모델을 추출하는 공격이다. 모델에 쿼리query를 계속 던지면서 값을 분석하며, 반복적인 쿼리를 통해 모델을 유추하여 유사한 모델을 만들어 낼 수 있다. 추출 결과는 모델 전도 공격에 활용하기 위해 사용될 수 있다.
모델 전도 공격	모델에 수많은 쿼리를 던진 후 산출된 결괏값을 분석해 모델 학습에 사용된 데이터를 추출하는 공격이다. 모델을 학습시키는 데이터 안에 개인정보, 민감정보 등이 포함되어 있는 경우라면 전도 공격에 의해 중요 정보가 유출될 가능성이 있다.

- 위와 같은 공격에 대비하여, 시스템 구현 단계에서는 특정 기간 내에 수행할 수 있는 질의 수를 제한하여 모델 공격을 위한 반복적인 질의를 방어하거나, 기계학습을 기반으로 모델 공격에 대한 사전 탐지 및 경고 알림을 설정하는 등 능동적인 방어가 필요하다.

13-1c 인공지능 시스템의 의사결정으로 인한 파급효과가 크고 불확실성이 높은 경우, 사람의 개입을 고려하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템이 인공지능 모델의 판단 결과를 활용하여 시스템 동작을 제어하거나, 사람의 안전 및 환경에 영향을 주는 정보를 제공할 때, 사람의 개입이 필요할 때가 있다. 이는 인공지능 시스템의 동작 및 기능의 파급 효과가 크지만, 인공지능 모델이 도출한 판단 결과의 불확실성이 높을 때이다.
- 특히, 인공지능 모델이 치안 위협 감지 프로세스의 의사 결정에 사용될 때, 모델 훈련에 사용되는 민감한 데이터로 인해 편향을 일으킨다[128]. 따라서, 예외 처리 및 보안 기법 외에, 사람이 직접 혹은 부분적으로 개입하여 인공지능 모델의 불확실성을 해소하는 방안을 고려하여야 한다.

참고 인공지능의 의사결정에 대한 사람의 개입 정도

• ISO/IEC 24028:2020의 9.4 Controllability와 WEF(World Economic Forum) Companion to the Model AI Governance Framework에서는 도출된 위험의 심각도 및 발생빈도를 기반으로 인공지능의 의사결정에 대한 사람의 개입 정도를 아래와 같이 분류(Guiding questions 3.2)하였다.

구분	설명 및 정의
Human-in-the-loop	인공지능 시스템이 의사결정을 수행하지 않으며 사람이 수행하는 의사결정에 보조적인 용도로 사용된다. 예) 의료 진단/처방, 법 관련 집행
Human-out-of-the-loop	인공지능 시스템이 의사결정을 수행하며 사람이 개입하지 않는다. 예) 항공사 예비 부품 예측, 구매 상품 추천
Human-over-the-loop	인공지능 시스템이 의사결정을 수행하나 사람이 해당 결과를 모니터링하고 최종 결정에 개입한다. 예) 내비게이션

13-1d 예상되는 사용자 오류에 대한 안내 및 대응을 제공하는가?

Yes ☐ No ☐ N/A ☐

• 사용자 오류는 외적으로는 서비스 최종 결과물을 사용하는 사용자에게서, 내적으로는 서비스 결과를 생성하고자 내부 시스템을 사용하는 작업자에게서 비롯된다. 따라서 서비스 담당자는 다음과 같은 사용자 오류 유형을 이해하고 이와 관련되어 발생하는 오류를 사전에 정의하고 분석하여야 한다.

- ✓ 누락 오류: 수행하여야 할 작업을 누락하여 발생하는 오류
- ✓ 작위 오류: 수행하여야 할 작업을 부정확하게 수행하여 발생하는 오류
- ✓ 순서 오류: 수행하여야 할 작업 순서를 틀리게 수행하는 오류
- ✓ 시간 오류: 수행하여야 할 작업을 정해진 시간 내에 완수하지 못하여 발생하는 오류
- ✓ 불필요한 수행 오류: 작업 완수에 불필요한 작업을 수행할 때 발생하는 오류

• 이와 같은 사용자 오류에 대한 대응 계획을 다음과 같이 수립한다.

대응 계획	설명과 예시
제약 조건 설정	허용 가능한 옵션을 정의하여 표시하거나 사용자의 선택 사항을 어느 정도 제한하여 잘못된 사용자 입력을 방지한다.
시스템 제안 및 수정	자주 발생하는 사용자 실수를 수집하고, 실제 서비스 중 유사한 사용자 실수가 발생하면 시스템에서 수정을 유도하거나 올바른 입력을 제안한다. 예를 들어, 측정 실수가 잦은 데이터를 수집하면 이상값을 설정하여 재측정을 제안하거나, 시스템 작동 방식을 소개하는 데모 강사, 가이드라인, 백서, 동영상 등을 추가한다.
기본값 설정	먼저 서비스 제공 업체에서 결정한 기본값 또는 시스템에서 자주 사용하는 값을 기본값으로 제공하거나 관련 예제를 제공하면 사용자 오류가 줄어든다.

대응 계획	설명과 예시
재확인·결과 제공 및 취소	사용자에게 받은 입력을 다시 한번 확인하고 예상 결과를 미리 전달한다. 또한 잘못된 결과를 실행 취소하는 등의 기능을 통해 오류를 방지한다. 예를 들어, 표준 위협 탐지 프로세스는 시스템 기본값이 변경되어 사용자가 예상치 못한 새 값으로 프로세스를 강행하려고 할 때 변경 사항을 재확인하는 알림을 출력한다. 또한 이러한 변경 사항이 시스템 결과에 영향을 미칠 때 시스템 관리자 또는 유지 관리 팀에 관련 변경 사항을 처리하도록 알림을 제공하는 안전 모드 메커니즘을 트리거하는 기능도 설계하여야 한다.

13-2 인공지능 시스템에서 문제 발생할 경우, 시스템은 이를 운영자에게 전달하는 기능을 수행하는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템은 서비스 도중 외부의 공격, 민감한 정보의 오용 등 다양한 요인으로 편향이나 성능 저하 등이 발생하므로 시스템 운영자가 이를 파악하도록 시스템의 자체적인 점검 기능이나 사용자가 운영자에게 관련 의견을 전달하는 기능을 제공하여야 한다.
- 시스템의 자체 점검 기능은 서비스 성능 저하나 외부 공격에 대한 검사 등을 수행한 후 가능한 범위 내에서 이에 대응하고, 해당 사실을 시스템 운영자에게 전달하는 체계를 갖춰야 한다. 여기에는 사용자가 연락하는 헬프 데스크 또는 지원 센터를 설정하거나 실시간으로 문제를 감지하고 시스템 운영자에게 보고하는 자동화된 시스템 구현이 포함된다. 이 절차에는 보고된 문제에 대응하고 해결하는 단계도 포함되어야 한다.
- 스마트 치안 시스템의 성능은 다양한 원인으로 인해 저하되므로 지속적인 평가가 필요하며, 관리 지표 및 절차가 수립되었는지 확인하여야 한다

13-2a 편견, 차별 등 윤리적 문제에 대한 알림 절차를 수립하였는가?

Yes ☐ No ☐ N/A ☐

- 윤리적 문제 알림 프로세스에 대해서는, 먼저 인공지능 시스템 자체의 신뢰도를 평가하는 기준과 검사를 준비한다. 주요 검사 항목의 예시는 다음과 같다.
  - ✓ 인권, 사생활, 법률 및 환경을 보호하는 제한, 다양성 존중, 비침해, 공익*, 연대*, 개인 데이터 관리, 책임, 안전, 투명성, 라이선스 관리, 민감한 데이터의 사용 및 저장 등

    * 민간 기업 또는 이니셔티브는 해당 사항을 무시한다.
- 스마트 치안 시스템에서 편견과 차별에 해당하는 의사 결정일 때 시스템에서 이를 감지하는 방안을 마련하여야 하며 사용자가 발견하였다면 운영자에게 신고하는 기능도 개발되어야 한다. 이 외에도 편견과 차별 요소를 감지하면 대응 절차를 마련하고, 리포팅 시 행위자에게 관련된 모든 정보를 즉시 제공하여야 한다.

핵심 절차	절차 설명
명확한 의사소통	윤리적 문제를 신고하고자 명확한 의사소통 채널을 구축한다. 핫라인, 챗봇(인공지능 챗봇), 온라인 의사소통 채널, 양식 또는 익명으로 문제를 신고하는 이메일 주소 및 사용자 인터페이스를 통해 문제를 신고하는 절차를 포함한다.
에스컬레이션 및 리뷰 프로세스	윤리적 문제가 신속하게 검토되고 대응되도록 명확한 에스컬레이션 및 검토 절차를 수립한다. 신고된 문제를 조사하고자 전담 팀을 할당하고 필요하면 문제를 고위 경영진에게 에스컬레이션하는 절차를 포함한다.
조사 및 해결	윤리적 문제가 보고되면 가능하면 빠르게 철저히 조사되고 해결되어야 한다. 관련 데이터 수집, 인터뷰, 질문지, 관련 이해관계자와 Likert 척도 조사 등을 포함하며, 문제에 대한 대응 계획의 개발을 포함한다.
정기적 감사 및 검토	스마트 치안 시스템의 윤리적 문제를 사전에 식별하고 해결하는 데 도움이 된다. 사용자 피드백 및 입력의 정기적인 검토와 스마트 치안 시스템의 의사 결정 과정의 감사를 포함하며, 편견이나 차별 가능성이 있는 영역을 식별한다.
투명성 및 책임성 확보	스마트 치안 시스템은 투명하고 책임감 있게 설계되어야 한다. 윤리적 문제를 다루는 명확한 정책과 절차가 필요하며, 사용자는 윤리적 문제를 신고하는 절차에 대해 알아야 하고, 신고된 문제에 대한 스마트 치안 시스템의 대응은 명확하고 투명하게 사용자와 이해관계자에게 전달되어야 한다.

13-2b 시스템 성능 저하를 평가하기 위한 지표 및 절차를 설정하고 알림 절차를 수립하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템은, 서비스 배포 및 운영 단계에서 일반적인 소프트웨어와 달리 지속적인 데이터 축적, 서비스 기능 확장, 환경 변화 등의 이유로 성능 변화가 생긴다.
- 인공지능 시스템은 실제 서비스 운영 중 갑자기 성능이 저하되었을 때 원인을 바로 알기 어렵다. 특히, 스마트 치안 시스템은 신뢰성과 보안 유지가 중요하므로, 시스템 성능 저하를 지속해서 평가·관리하는 지표와 절차가 설정되었는지 점검이 필요하다.
- 스마트 치안 모델의 성능 모니터링 중 발견되는 데이터 및 모델 드리프트는 성능 저하의 주요 원인이며, 이러한 드리프트를 관리하고자 허용 범위를 명확히 정의하는 것이 중요하며 설정된 한계를 초과하면 모델 재학습 또는 업데이트가 필요하다[129].
  - ✓ 데이터 드리프트: 입력 데이터의 통계적 특성이 변화는 현상
  - ✓ 모델 드리프트: 입력 데이터와 예측 결과 사이의 관계가 변화는 현상
- 실제 스마트 치안 서비스 운영 중 인공지능 시스템의 갑작스러운 성능 저하의 원인을 실시간으로 식별하기는 어려워, 시스템의 성능 저하를 지속해서 평가하고 관리하는 지표와 절차를 시스템에 포함한다. 시스템 성능 점검 결과 성능 저하가 발견되면, 관련 정보에 대해 사용자와 시스템 운영자에게 알리는 절차 마련이 필요하다.
- 일반적으로 스마트 치안 시스템에 적용하는 대표적인 성능 지표로는 AUC, F1-score, 정밀도, 정확도, 재현율, 특이도, 위협 점수, 진양성, 진음성, 가양성, 가음성 등이 있다. 성능 저하가 확인되면 이를 시스템 운영자에게 보고하고 운영자는 성능 저하의 원인을 찾아 개선하는 절차를 마련하여야 한다.

• 스마트 치안 위협 탐지 모델에서 판단 오류, 추론 결과의 편향 또는 해석 불가능성 등 성능 저하가 발생하면, 해당 도메인 전문가는 데이터 재검토 및 성능 재평가를 진행한다.

참고 인공지능의 성능과 유용성을 검증하는 방법 예시[130]

〈개발된 모델의 성능에 대한 신뢰 검증 시 확인하여야 할 5가지 주요 질문〉

• 목표가 제대로 설정되었는가?

※ 실제 세계의 문제를 매핑하고, AI 시스템이 올바르게 문제를 해결하는지 여부를 물어보는 개념적 정확성에 관한 결정을 내리며, 보안 위협 탐지의 추론 결과가 소수 집단에 대한 차별을 초래하는지 등의 윤리적 문제를 포함

• AI 시스템은 소프트웨어 버그에서 자유로운가?
• AI 시스템은 적절하게 대표적인 데이터를 기반으로 하는가?
• AI 시스템은 이상치와 불가피한 데이터 오류를 처리하는가?
• AI 시스템의 정확도가 충분한가?

스마트 치안 시스템의 맥락에서 AI 시스템이 매우 정확한가에 대한 질문은 매우 중요하다. 이러한 시스템의 신뢰성 기준에는 예측 동등성 기준 및 집단 보정 기준 등의 측면이 포함된다. 스마트 치안에서 정확성을 보장하려면 다각적인 접근 방식이 필요하다. 입력 단계에서 보호변수의 구성은 직접적인 차별 사례 방지와 직접적으로 관련되어 매우 중요하다. 출력 단계로 넘어가면 간접 차별 방지와 직결되는 결과 동등성 기준을 만난다. 오차율 동등성 기준은 평가자의 관점에서 공정성을 보장하고 기회균등을 유지하는 역할을 한다. 마지막으로 예측도 동등성 기준은 평가자 입장에서 AI 시스템의 신뢰성을 확보하는 데 중요한 역할을 한다. 전반적으로 스마트 치안에서 AI 시스템의 정확성은 기술력뿐만 아니라 의사 결정 과정에서 공정성과 비차별을 유지하는 데 중요한 역할을 하는 윤리적 기반에 달려 있다.

※ 나머지 질문에 대한 답변은 수학적 연습이 필요함. 상세한 평가 방법들(08-1b, [131] 참조)은 다음과 같음

✓ ROC 곡선 분석
✓ 인구 통계적 균형
✓ 균등한 확률
✓ 예측 균형
✓ 보정 분석
✓ 정확도 차이

※ 개발된 치안 시스템의 성능을 검증하려면, 개발된 분야에 특정한 관련 항목의 성능 모니터링도 고려하여야 함. 예를 들어, 경찰 또는 정부 기관의 폭력/범죄 탐지 또는 감시 시스템을 개발할 때, 시스템 관련 항목에 대해 다음과 같은 접근법을 고려함

✓ 사용자 설문 조사 실시: 법 집행관이나 보안 직원 등 사용자들에게 피드백을 수집하여 스마트 치안 시스템의 성능에 대한 만족도를 평가. 설문 조사는 개발된 시스템의 정확성, 신뢰성, 사용성 그리고 보안 사건을 식별하고 대응하는 데 그 효과성에 대한 질문을 포함함[132]
✓ 운영 지표 평가: 얼굴 인식 알고리즘의 속도와 정확성, 거짓 긍정과 거짓 부정 비율 그리고 시스템의 응답 시간 등 스마트 치안 시스템의 운영 성능 지표를 평가. 이러한 지표들은 시스템의 성능과 원하는 목표를 달성하는 능력에 대한 통찰력을 제공함[178]
✓ 영향 평가: 경찰 활동에서 얼굴 인식 또는 기타 감시 도구 사용에 대한 영향을 이해하고자 영향 평가를 수행. 데이터 개인정보 보호, 편향, 시민 권리와 관련된 문제를 포함하여 커뮤니티에 대한 잠재적 이점, 위험, 영향을 평가. 이러한 평가는 스마트 치안 시스템의 전반적인 유용성과 윤리적 고려 사항을 평가하는 데 도움이 됨[133]
✓ 데이터 검증 고려: 다른 정보 또는 증거의 결과와 비교하여 스마트 치안 시스템의 성능을 검증. 예를 들어, 수사관은 시스템에서 얻은 데이터 포인트를 사용하여 사람의 알리바이를 검증하거나 범죄 현장에서 수집된 데이터와 시스템의 출력을 교차 참조함[134]
✓ 전문가와 협업: 인공지능, 기계 학습, 법 집행 분야의 전문가와 협력하여 스마트 치안 시스템의 성능과 유용성을 평가. 전문가들은 통찰력을 제공하고, 독립적으로 평가하며, 개선에 필요한 권고 사항을 제공함[135]

투명성

요구사항 14 인공지능 시스템의 설명에 대한 사용자의 이해도 제고

- 모델의 추론 결과에 대한 설명을 제공하는 기법을 적용하여도 사용자가 바로 이해하고 해석하기 어려울 때가 많다. 따라서 인공지능 시스템의 운영자 혹은 서비스 제공자는 사용자에게 제공되는 결과가 이해 가능한지understandable, 해석 가능한지interpretable, 설명 가능한지explainable를 평가한다.
- 감시, 폭력 감지, 국경 통과 또는 가석방 상담용 거짓말 탐지기 등의 스마트 치안 인공지능 시스템에서 사용자가 모델 결과를 이해하고 사용하는 것은 보안 인식을 향상하며, 스스로 위협에 대한 경계를 강화하여 더 안전한 환경을 조성한다. 따라서, 사용자의 이해를 향상하고자 사용자 특성 평가와 적절한 설명 제공이 중요하다.

14-1 인공지능 시스템 사용자의 특성user characteristics과 제약 사항을 분석하였는가?

Yes No N/A
☐ ☐ ☐

- 인공지능 시스템의 결과가 적절한지 평가하려면 먼저 해당 결과를 읽는 사용자를 고려하여야 한다. 사용자가 누군지에 따라 결과(설명)의 수준, 깊이, 맥락이 정해지는 만큼 사용자에 대한 자세한 분석이 수행되어야 한다.
- 만약, 사용자 특성을 고려하지 않으면 시스템의 부적절한 사용, 신뢰도 저하, 오경보, 보안 위반 누락 등으로 이어져 고객의 구매 결정에 영향을 미치며, 사용자의 다양성을 고려하지 않으면 차별이 발생한다.

14-1a 사용자 특성에 따른 세부 고려사항을 분석하였는가?

Yes ☐ No ☐ N/A ☐

- 스마트 치안 서비스 기획 단계에서 사용자의 선호도와 요구사항needs에 집중하였다면, 설명의 적절성을 평가하려면 각 사용자의 다양한 특성을 고려하여야 한다. 예를 들어, 서비스 사용자의 법 집행 절차에 대한 친숙도나 데이터 분석 경험에 따라 제공되는 설명 이해력 차이를 고려하여야 한다.
- 사용자 특성을 분석하고자 고려하여야 할 요소의 예시는 다음과 같다.

사용자 특성을 분석하고자 고려하여야 할 항목 예시

항목	세부 구분	고려해야 할 내용
나이	청소년, 성인, 노인 등	연장자 사용자와 젊은 사용자는 요구사항과 선호 사항이 다르다. 예를 들어, 큰 글꼴, 간단한 인터페이스, 음성 명령을 선호한다. 또한, 기술 수준의 차이로 인해 이 그룹은 젊은 세대에 비해 이해하는 용어나 어휘가 제한된다. 따라서, 노인 사용자는 글꼴 크기가 크고 탐색이 쉬운 간단한 인터페이스를 고려하여야 하며, 젊은 사용자는 기술적 전문성을 충족시키는 더 고급 기능을 고려하여야 한다.
성별	남성, 여성 등	치안 시스템 도메인에 따라 모델 및 인터페이스를 개발하는 동안 남성과 여성의 차이를 확인하여야 한다. 예를 들어, 화면 사용 시 행동 패턴의 주요 차이점으로 주로 사용하는 사용 습관을 기반으로 인터페이스를 설계하여야 한다.
인종 또는 언어	아시아인, 유럽인 등	다른 인종은 피부색이나 체형을 인식하는 평균 기준이 다르다. 사용자의 언어와 문화도 중요한 고려 사항이다. 스마트 치안 시스템은 사용자의 지역과 문화와 관련된 적절한 상징과 색상을 사용하여 다양한 언어와 문화를 수용하도록 설계되어야 한다.
신체적 능력	장애, 또는 일반인	신체적 제약이나 장애가 있는 사용자는 음성 명령, 조이스틱 컨트롤러 또는 터치패드 등 특수 기능이나 인터페이스를 사용하여야 시스템이 효과적으로 사용된다. 따라서, 이동 장애가 있는 사용자는 조이스틱 컨트롤러나 터치패드가 필요하고, 청각 장애가 있는 사용자는 시각적 알림이 필요하다.
기술 전문성	도메인 전문가, 컴퓨터 공학자 등	사용자의 기술 전문성은 또 다른 중요한 고려 사항이다. 기술적 전문성이 제한된 사용자는 시스템 설정 및 사용에 대해 더 많은 안내와 지원이 필요하며, 더 고급 기술 스킬을 가진 사용자는 더 복잡한 시스템과 더 많은 맞춤 설정 옵션을 선호한다.
지역	외각지, 교외, 도심, 바다, 땅 등.	사용자의 스마트 치안 시스템 활용 지역이 시스템을 설계할 때 고려하여야 할 요소이다. 예를 들어, 범죄율이 높은 지역의 사용자는 더 자주 또는 상세한 경고가 필요하며, 범죄율이 낮은 지역의 사용자는 적은 횟수의 경고가 필요하다.

14-2 사용자 특성에 따른 설명을 제공하는가?

Yes ☐ No ☐ N/A ☐

- 스마트 치안 서비스를 이용하는 사용자는 다양하여 인공지능 시스템의 결과가 서로 다른 입장에서 설명이 해석되고 오해가 생긴다. 따라서, 14-1에서 분석된 사용자 특성을 고려하여 설명을 평가하는 기준 항목을 수집하여야 하고, 설명 평가의 기준으로는 명확성, 구체성, 정확도 등을 고려하여야 한다.

14-2a 사용자 특성에 따른 설명 평가 기준을 수립하였는가?

Yes ☐ No ☐ N/A ☐

- 스마트 치안 분야에서는 다양한 사용자가 서비스를 이용하는 만큼 설명을 포괄적으로 평가하는 특성과 세부 항목을 정하는 단계가 필요하다. 설명의 평가 기준은 구체성, 명확성, 적절성 등의 항목이다. 세부 항목으로는 데이터 유형data type이나 모달리티modality에 따라 다르므로, 상세한 항목으로 고려하여야 한다.

사용자 특성에 따른 평가 항목 예시

구분	평가 항목
명확성	스마트 치안 시스템의 명확한 설명은 사용자가 시스템의 능력과 한계를 이해하는 데 필수이다. 평가 기준을 수립함으로써, 설명은 기술에 익숙하지 않은 사용자도 쉽게 이해하도록 만들어진다. • 사용자에게 다른 오해를 불러일으킬 만한 표현·단어·어휘는 없는가? • 불필요한 설명은 없는가? • 해당 설명을 통해 사용자가 기대하고 얻으려는 정보가 모두 들어 있는가? • 설명을 통해 인공지능 모델 결과의 이유를 이해하거나 받아들이기 쉬운가?
구체성	스마트 치안 시스템은 종종 특정 사용자의 요구를 충족시키도록 설계된다. 사용자 특성에 기반한 설명을 평가하는 기준을 설정함으로써, 시스템은 각 사용자의 특정 요구를 충족시키도록 맞춤화된다. • 사용자의 구체적 행동을 이끌도록 명확한 주어·목적어·동사를 활용하여 설명되는가? • 그래프/막대 등의 도구로 결과의 구체성을 지원하는가?
적절성	평가 기준의 설정은 스마트 치안 시스템에서 투명성을 증진한다. 사용자 특성과 일치하는 명확하고 상세한 설명을 제공함으로써, 시스템은 사용자와 신뢰를 구축하고 그들이 기술이 어떻게 작동하는지 이해하는 데 도움을 준다. • 주어진 설명이 사용자의 특정 지식수준을 요구하지는 않는가? • 배경지식 혹은 사전 경험이 필요하진 않은가? • 설명이 사용자에게 유용한가? • 독자를 고려한 전문 용어, 약어에 대한 설명을 제공하는가? • 설명이 제공되는 시점이 적절하였는가?
정확성	설명의 정확성은 사용자가 스마트 치안 시스템 사용에 대한 정보를 얻고 결정을 내리는 데 중요하다. 시스템은 설명이 정확하고 관련성이 있으며 최신 정보를 반영하도록 보장한다. • 설명과 함께 제공되는 자료의 그림과 설명이 모두 일치하는가? • 사전에 제공된 예상 결과의 설명과 실제 결과가 모두 일치하는가? • 내부 알고리즘과 정확히 일치하는 설명인가?

• 다음의 체크리스트에는 스마트 치안 시스템의 HMI에 대한 사용자 특성을 기반으로 한 디자인 권장 사항이 구성된다.

사용자 특성과 다양성을 기반으로 한 HMI 디자인 권장 사항

구분	권장 사항	명확성	구체성	적절성	정확성
인터페이스 디자인	사용자의 특성 또는 요구에 적합하게 인터페이스 디자인 • 예를 들어, 노인 사용자는 더 큰 글꼴과 더 간단한 레이아웃이 필요하고, 젊은 사용자는 더 고급이고 기능이 풍부한 인터페이스를 선호함 • 시스템이 대상 사용자가 쉽게 사용 가능하도록 설계하여, 명확하고 직관적이며 쉽게 탐색하는 방식으로 디자인	V	V	V	V
알람 설정	사용자의 특정 요구를 충족하게 사용자 정의 알람 설정 • 시스템의 민감도, 경보 빈도 및 수신되는 경보 유형을 조정 • 사용자가 시간대나 위치에 기반하여 경보 선호도 사용자 정의	V		V	
기기 호환성	필요시 스마트 치안 시스템 외 다양한 기기와의 호환성, 사용성을 고려 • 웹 브라우저나 모바일 앱을 통해 접근 및 시스템 원격 모니터링 등 지원			V	V
기술 지원	명확하고 포괄적인 문서를 제공, 이메일, 전화 또는 채팅을 통해 고객 지원팀에 문의하는 기능 등을 제공 • 사용자가 시스템을 효과적으로 사용하도록 교육 자료와 튜토리얼 제공	V	V	V	
물리/신체적 제약	신체적 제약이나 장애가 있는 사용자를 수용하도록 설계 • 음성으로 작동하는 명령이나 특수한 제어 장치(조이스틱 컨트롤러 또는 터치패드 등)의 사용 • 텍스트, 아이콘의 크기와 색상을 사용자가 조정하도록 제공하여 청각 장애가 있는 사용자에게 더욱 명확한 시각적 알림	V	V		
개인정보 고려	사용자의 개인정보를 보호하고 데이터 수집 및 저장에 대한 명확하고 투명한 정책을 제공 • 사용자가 수집되는 데이터와 그 사용 방법을 제어하는 기능을 제공(개인정보 보호법 제15조)			V	
현지화	사용자의 지역과 문화에 따라 다른 언어, 기호 및 색상을 사용하고, 다른 시간대와 지역적 관습을 수용하도록 설계	V	V	V	

14-2b 사용자가 이해하기 어려운 전문 용어 사용을 지양하였는가?

Yes No N/A
☐ ☐ ☐

- 스마트 치안 시스템은 제공하는 기능 및 정보에 따라 익숙하지 않은 용어나 전문 용어를 사용하면 의미를 이해하기 어려워 해석에 많은 시간이 소요되거나 의도한 내용을 전달하지 못한다.
- 따라서 텍스트로 설명한다면 다양한 독자를 배려하여 전문 용어를 최대한 지양하고, 필요시 용어에 대한 설명을 추가로 작성하여야 한다. 예를 들어, 자연어 처리 기술을 인터페이스에 적용하여 문장 내의 특정 단어나 복잡한 그래픽을 사용자의 수준에 맞는 적절한 단어 또는 설명으로 변환하여 제공한다.
- 반대로, 사용자가 도메인 전문가를 대상으로 할 때, 이해하는 시간을 단축하고자 전문가가 충분히 이해하는 수준의 전문 용어 사용을 권장한다.

홈 보안 용어, 전문 용어, 유행어 목록[136]

용어	정의
Biometrics	얼굴, 홍채 또는 지문 인식 등 인간의 고유한 외모 또는 행동 특성에 대한 측정 및 통계 분석
Bypass	액세스 권한을 얻고자 기존 시스템을 우회하는 행위
Cellular connectivity	셀룰러 네트워크를 통한 무선 통신
Hardwired	셀룰러 연결이 아닌 케이블 및 전선으로 연결되는 치안 시스템
Tamper	시스템이나 장치를 손상하거나 사용을 위태롭게 하는 방식으로 방해하는 행위
Prolific Offenders	다수의 범죄를 저지른 범죄자

14-2c 사용자의 구체적인 행동과 이해를 이끌어낼 수 있도록 명확한 표현을 사용하였는가?

Yes No N/A
☐ ☐ ☐

- 좋은 설명은 사용자의 구체적인 행동과 이해를 끌어내야 한다. 따라서 설명을 간결하고 명확하게 함으로써 모호하게 해석되지 않도록 작성하는 것이 중요하다.
- 시각적으로 성공·실패·경고·위험 등 결과에 따른 색상을 일관성 있게 유지해 줌으로써 사용자가 한눈에 시스템 결과를 이해하여야 한다. 그리고 텍스트, 그래픽, 음성으로 제공되는 설명에서는 지시 대명사를 사용하지 않고 대상을 명확하게 말해 주는 것을 예로 들 수 있다. 또한 비슷한 발음이 연이어질 때, 다른 단어로 대체하는 것이 바람직하다.

명확한 언어 표현 항목	설명
일반적 언어 사용	사용자들을 혼동시키는 기술 용어나 복잡한 용어를 사용하지 않고, 기술에 익숙하지 않은 사용자들에게도 이해하기 쉬운 일반적인 언어 사용
구체적 메시지 사용	시스템의 행동과 기능을 설명하고자 구체적인 언어 사용 • 예를 들어, “시스템이 경고를 보냅니다” 대신 “시스템이 휴대 전화로 알림을 보냅니다”와 같이 구체적인 메시지 제시

명확한 언어 표현 항목	설명
맥락 정보 제공	사용자가 특정 행동을 취하여야 하는 이유를 쉽게 이해하도록 맥락 정보를 함께 제공 • 예를 들어, “코드를 입력하십시오.” 대신 “로봇이 아님을 확인하고, 시스템을 활성화하고자 코드를 입력하십시오.” 등 맥락 정보 제공
능동 표현 사용	사용자가 적극적으로 반응하도록 능동적인 표현 사용 • 예를 들어, “알람이 활성화될 것이다.” 대신 “알람을 활성화하십시오.”등 능동적으로 요청함
피드백 제공	사용자에게 피드백을 제공하여 시스템이 성공적으로 완료 또는 실패하였음을 알림 • 예를 들어, 사용자가 치안 시스템을 활성화하고자 코드 입력 시 “인증 사용자이다. 시스템을 활성화합니다.” 등 시스템의 동작 성공에 대한 피드백을 제공함
테스팅 및 정제	사용자와 함께 시스템 표현을 테스트하고 피드백에 기반하여 개선함

14-2d 설명이 필요한 위치와 타이밍이 적절한가?

Yes No N/A
☐ ☐ ☐

- 잘 작성된 설명이 적절한 위치 및 타이밍에 나타나 이해를 돕는 것도 중요하다. 이를 위해 설명이 단발성이어야 하는지, 여러 번 반복하여 강조하여야 할지 숙고하고, 어느 위치에 놓여야 사용자가 잘 읽을지 고려하는 것이 필요하다.
- 특히, 불필요한 그래픽, 시각적 표현은 사용자에게 혼란을 야기하므로, 시각적 자료의 해상도, 위치 등의 조정도 필요하다. 예를 들어, 스마트 치안 시스템이 침입자나 화재 등 잠재적인 위협을 감지하였을 때, 관련된 시각화 지표나 도구와 함께 다음 단계에 대한 명확하고 간결한 안내를 제공한다면, 사용자가 위급 상황에서 안전하고 신속하게 대응할 것이다.

참고 스마트 치안 시스템의 알림 인터페이스 설계 고려 사항

알림 및 시스템 도구의 위치와 시간은 스마트 치안 시스템에서 매우 중요하다. 적절하게 설계된 알림은 사용자에게 잠재적 위협을 즉시 알려 주어 위험을 완화하고 자산을 효과적으로 보호하게 한다.

따라서 스마트 치안 시스템 인터페이스는 다음의 주요 디자인 사항을 고려하여 사용자에게 효율적이고 안전한 경험을 제공한다.

- 사용자 맞춤형 시스템 디자인 및 알림: 개인의 필요를 고려하여 사용자 중심의 접근 방식으로 설계[137]. 인공지능 모델이 실패할 때, 사람의 주의를 끄는 알림을 제공하는 등 사용자가 원하는 특정 유형의 경고와 알림을 받는 방법(예: 푸시 알림, 이메일, 앱 내 알림 등) 등을 설정[138]
- 그래픽 및 시각적 지표의 구성: 불필요한 그래픽 또는 시각적 표현은 보안 위협 결과의 혼란과 잘못 해석을 초래하여 중요한 상황에 대한 적시, 정확한 대응을 방해함. 적절한 균형과 시각적 표현의 명확성 확보가 시스템의 사용성과 보안 위협 감지 및 대응 능력을 향상함[139]
- 과도한 알림 회피: 과도한 알림으로 사용자를 피곤하게 하지 않아야 함. 논리 다이어그램을 사용하여 알람 상황의 중복을 제거하고, 관련성이 있고 필수적임을 보장하며 불필요한 경고와 사용자의 피로를 예방함[188]
- 간결한 데이터 표현: 사용자와 효과적으로 커뮤니케이션하고자, 알림 시스템은 데이터를 간결하고

이해하기 쉬운 방식으로 제시. 명확하고 직설적인 정보는 사용자가 상황을 빠르게 파악하고 적절하게 조치하게 하기 때문임[188]

- 적절한 승인 처리: 스마트 치안 시스템 인터페이스에는 사용자가 받은 알림을 승인하는 메커니즘이 필요함. 적절한 승인 처리는 사용자가 알림을 인지하고 확인하게 해 주어, 중요한 정보를 놓칠 위험을 줄여 줌[188]
- 시간 및 중요도: 알림은 이벤트의 시간과 중요도를 고려하여 전송되어야 함. 중요한 경고는 즉시 주의를 환기하여야 하나, 급하지 않은 경고는 사용자가 피곤하지 않도록 일괄 처리되거나 지연하여야 함[140]
- 알림 그룹화: 관련 알림을 그룹화하여 사용자가 불필요한 확인 및 중단을 회피하도록 함[140]

14-2e 사용자 경험을 평가할 수 있는 다양한 사용자 조사 기법을 활용하였는가?

Yes ☐ No ☐ N/A ☐

- 사용자 경험UX, User eXperience은 한 개인이 특정한 제품, 시스템 또는 서비스를 사용하며 느끼는 모든 것을 의미한다. 또한 그 개인이 인지하는 유용성, 사용 편의성, 효율성 등의 시스템 특성을 포함한다. 설명을 평가하고자 사용자 조사user research 기법을 활용한다.
- 사용자 조사 기법은 크게 접근 방식과 자료 획득 방식으로 구분한다. 우선, 사용자 조사 기법의 접근 방식에 따라 정량적(간접적) 조사와 정성적(직접적) 조사로 구분되며, 사용자 조사를 하고자 자료를 얻는 방식에 따라 사용자 행동을 통한 조사와 태도를 통한 조사로 구분한다. 접근 및 자료 획득 방식을 고려하여 적합한 사용자 조사 기법을 선정하고, 사용자 경험을 평가하는 것이 바람직하다.
- 접근 방식에 따른 사용자 조사 기법 구분 및 방법은 다음과 같다.
  - ✓ 정량적(간접적) 조사quantitative user research: 도구 등을 통하여 사용자의 행동이나 태도에 대한 데이터를 간접적으로 수집하는 방법
  - ✓ 정성적(직접적) 조사qualitative user research: 사용자의 행동이나 태도를 직접 관찰하는 방법
- 데이터 획득 방법에 따른 사용자 조사 기법 구분 및 방법은 다음과 같다.
  - ✓ 사용자 행동 기반 조사: 사용자가 무엇을 하는지 조사하는 방법(예: 분석, A/B 테스트, 눈동자 추적 등)
  - ✓ 사용자 태도 기반 조사: 사용자들이 무엇을 말하는지 조사하는 방법(예: 카드 정렬, 심층 인터뷰, 요구사항 연구 등)

사용자 조사	세부 구분	설명
정량적 (간접적) 조사	분석	스마트 치안 시스템과 사용자의 상호 작용에 대한 양적 데이터를 수집하는 데 사용함 • 사용자가 시스템을 어떻게 사용하는지, 어떤 기능이 가장 인기 있는지 그리고 시스템의 어떤 부분이 개선이 필요한지에 대한 통찰력을 제공함
	A/B 테스팅	스마트 치안 시스템의 다른 버전 또는 기능 등과 비교하여 사용자 경험 측면에서 어떤 버전이 더 효과적인지 평가하는 데 사용함 • A/B 테스트는 시스템의 두 개 이상의 버전을 만들고 사용자 경험 측면에서 어떤 것이 더 효과적인지 비교 분석하여 결정

사용자 조사	세부 구분	설명
	설문	사용자의 경험에 대한 양적 데이터를 수집하는 데 사용함 • 온라인 또는 직접 수행하며, 시스템의 기능에 대한 사용자의 인식에 대한 소중한 통찰력을 제공받음
정성적(직접적) 조사	인터뷰	사용자의 경험에 대한 질적 데이터를 수집하고자 인터뷰를 진행함 • 시스템과 관련된 사용자의 인식, 태도 및 행동에 대해 더 깊은 통찰력을 제공
	프로토타입 테스팅	스마트 치안 시스템의 사용 편의성을 평가하는 데 사용함 • 사용자들은 특정 작업을 수행하도록 요청받고, 관찰자는 시스템과 상호 작용을 관찰하고 기록함 • 시스템이 사용하기 어렵거나 사용자에게 혼란을 주는 영역에 대한 통찰력을 제공함
	포커스 그룹	스마트 치안 시스템에 대한 사용자의 경험에 대한 질적 데이터를 수집하는 데 사용됨 • 사용자들이 서로의 경험과 의견을 공유하게 하여 사용자들의 상호 작용에 대한 유용한 통찰력을 제공함

참고 스마트 치안 시스템에서 사용자 경험을 평가할 때 고려해야 할 체크리스트 예시

- 사용성: 사용자가 스마트 치안 시스템과 상호 작용이 얼마나 용이한지 파악한다. 사용성 평가에는 시스템의 인터페이스 디자인, 탐색의 용이성, 명확한 지시 사항 등의 요소가 포함된다.
- 효과성: 효과성은 스마트 치안 시스템이 사용자의 요구사항과 목표를 얼마나 잘 충족하는지 나타낸다. 이에는 시스템이 잠재적 위협을 감지하고 대응하는 능력, 시스템의 경고 및 알림의 정확성 그리고 시스템의 전반적인 성능 등의 요소가 포함된다.
- 효율성: 사용자가 스마트 치안 시스템을 사용하여 작업을 얼마나 빠르고 쉽게 수행하는지를 나타낸다. 효율성 평가에는 사용자의 동작에 대한 시스템의 반응 속도, 설정 및 선호도를 맞춤 설정하는 능력 그리고 관련 정보에 빠르게 접근하는 능력 등의 요소가 포함된다.
- 만족도: 사용자가 스마트 치안 시스템을 사용하는 전반적인 경험에 얼마나 만족하는지를 나타낸다. 이에는 시스템의 인지 가치, 사용의 용이성, 서비스의 품질 등의 요소가 포함된다.
- 접근성: 장애나 손상이 있는 사용자가 스마트 치안 시스템을 얼마나 쉽게 사용하는지를 나타낸다.

05 운영 및 모니터링

책임성 투명성

요구사항 15 서비스 제공 범위 및 상호 작용 대상에 대한 설명 제공

- 사용자가 인공지능 시스템이 제공하는 서비스를 올바르게 사용하고, 제공된 서비스를 오남용하지 않도록 서비스의 목적, 범위, 제한 사항, 면책 조항disclaimer, 상호 작용 대상을 포함한 내용을 설명한다.

15-1 인공지능 서비스의 올바른 사용을 유도하기 위한 설명을 제공하는가?

Yes ☐ No ☐ N/A ☐

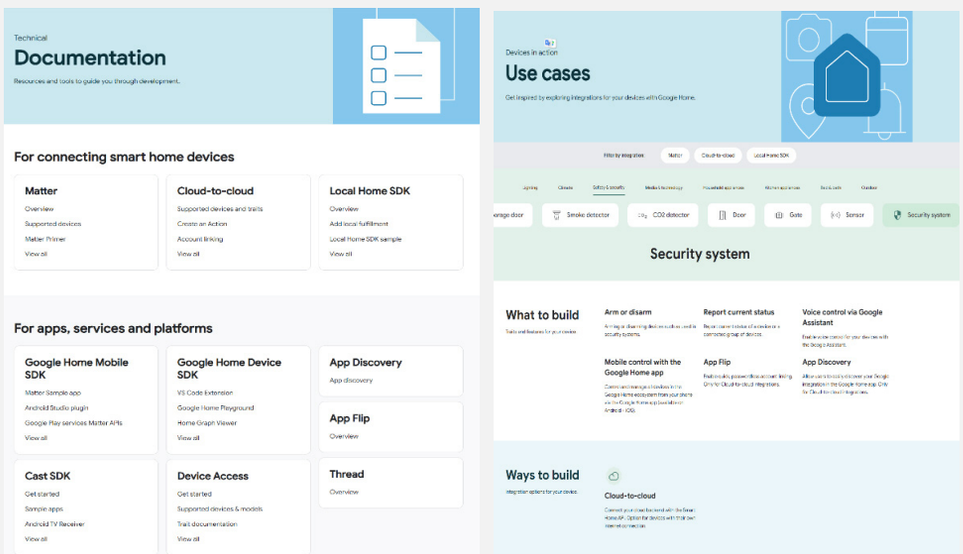
- 인공지능의 활용 범위가 넓어지면서 사용자가 서비스 기능을 실제 서비스 제공 범위보다 더 넓게 기대하여 오해할 때가 발생한다. 따라서 서비스 목적, 범위, 제한 사항, 면책 조항 등에 대한 설명을 제공함으로써 인공지능 기술의 오남용을 방지하고, 서비스에 대한 사용자의 기대치 조정이 중요하다.
- 또한 치안 서비스 분야의 인공지능 시스템은 다수 국민의 개인정보를 학습 데이터로 다루므로 사용자가 인공지능 서비스를 이용하는 과정에서 오남용하지 않고 올바르게 사용하도록 해당 정보 제공이 중요하다.
- 스마트 치안 인공지능 시스템의 사용자는 다양한 특성이 있으며, 특히 전문 지식에 대한 이해 수준이 초보자에서 전문가까지 다양하다. 따라서 서비스에 대해 제공되는 설명에 따라, 사용자의 해석 차이가 발생하고, 기대치가 잘못 설정된다. 스마트 치안 인공지능 시스템을 올바르게 사용하고자, 기술의 장점과 한계에 대해 교육하고, 효과적으로 사용하는 방법에 대한 지침을 제공하여야 한다.

스마트 치안 시스템의 설명 항목

설명 항목	세부 내용
스마트 치안 AI 시스템의 기능 이해	• 스마트 치안 AI 시스템은 기능과 성능이 매우 다양하므로, 사용자 설명서를 준비하고 시스템의 동작 방식에 대한 지침을 마련하여야 한다. • 이렇게 하면 사용자가 기능을 명확하게 이해하여 시스템을 효과적으로 활용하고, 일반적으로 발생하는 실수를 예방하여야 한다.
목적에 적합한 기능 사용	• 스마트 치안 AI 시스템과 함께 사용되는 장치는 각각 용도가 달라, 사용 설명서를 준비하고 시스템과 함께 사용하는 방법을 안내하여야 한다. • 특히, 특정 보안 기능에 활용되는 장치들은 각각의 사용법에 대한 정보를 제공하여야 한다.
경고 및 알림에 대한 정보 제공	• 대부분의 스마트 치안 인공지능 시스템은 사용자가 경고 및 알림을 설정하여, 잠재적인 보안 위협에 대한 정보를 실시간으로 확인할 수 있다. • 이처럼 보안 위협에 대해 사용자가 인지하는 방법 또는 기능에 대한 정보를 제공하여, 사용자가 보안 문제를 파악하도록 한다.
모니터링에 대한 정보 제공	• 일부 스마트 치안 인공지능 시스템은 전문가 모니터링 서비스를 제공하여, 전문가가 24시간 연중무휴로 사용자의 시스템을 모니터링하고, 긴급 상황 발생 시 사용자 또는 유관 기관에 알림을 보낸다. • 이처럼 시스템 외에 추가적인 서비스가 있다면, 사용자에게 관련 정보를 고지하고, 사용 및 운영 지침을 제공하도록 한다.

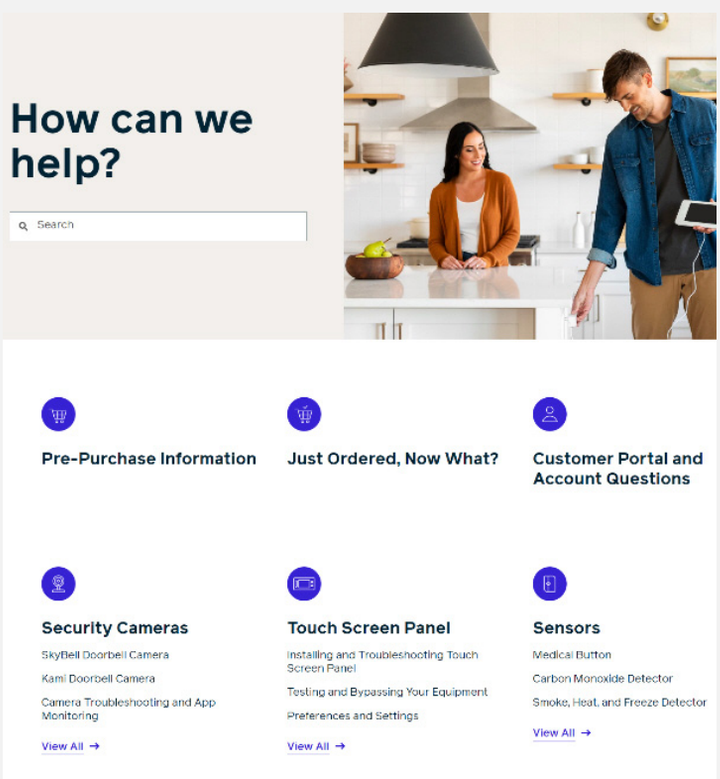
참고 AI 플랫폼 온라인 지원 사례

구글 홈은 구글에서 개발한 스마트홈 플랫폼으로, 사용자가 하나의 앱을 통해 스마트홈 및 치안 시스템을 생성, 제어, 관리하며, 서비스에 대한 지원 관련 문서를 제공한다.



〈Source: Google, Google Home Documentation[147]〉

보안 서비스를 제공하는 COVE사는 Help Center를 별도로 운영하며, 이 페이지를 통해 사용자에게 시스템에 대한 온라인 문서와 지침을 제공한다.

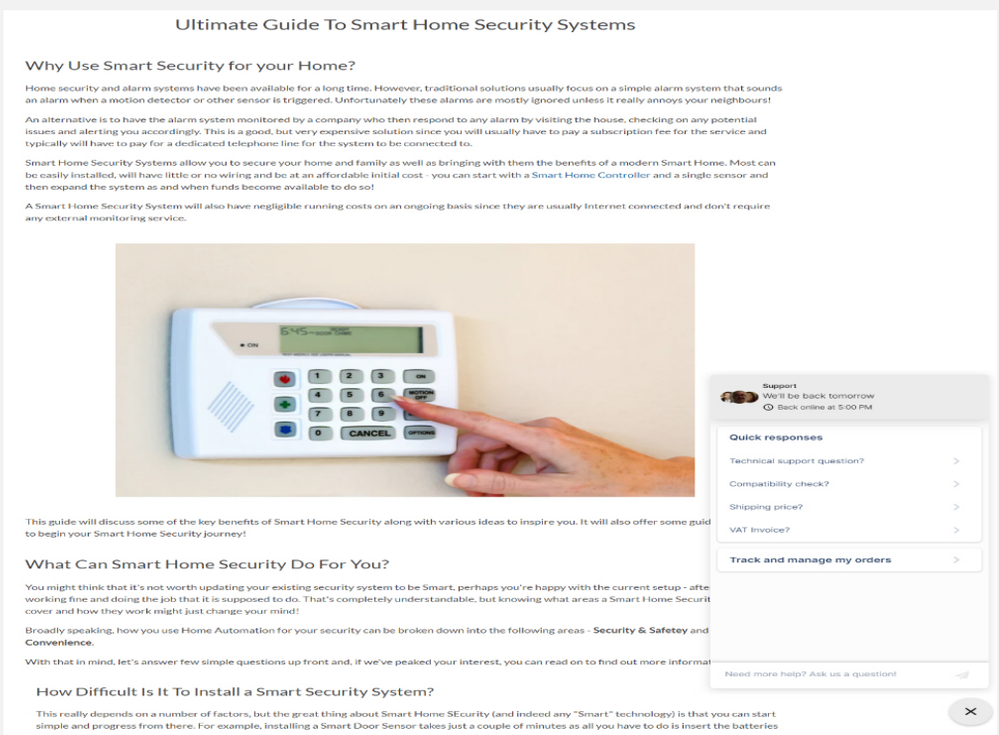


15-1a 서비스의 목적과 목표에 대한 설명을 제공하는가?

Yes No N/A
☐ ☐ ☐

- 서비스 목적goal은 서비스 제공사가 인공지능 시스템을 어떤 목적으로 제공하는지에 대한 방향성을 담은 것이며, 목표objective는 사용자가 해당 기능을 사용함으로써 무엇을 어떻게 구체적으로 얻는지를 의미한다. 사용자는 서비스 목적과 목표를 설명함으로써 사용 맥락에 맞는 적합한 기능을 선택하여 활용하여야 한다.
- 스마트 치안 시스템은 최종 사용자, 당국에 의해 직접 사용되므로 해당 사용자 그룹을 대상으로 서비스의 목적과 목표에 대한 설명은 비교적 쉽고 명확하다. 또한, 사용자가 일반 최종 사용자일 때, 서비스의 오용 또는 남용 가능성으로 전반적인 서비스에 대한 자세한 설명이, 미리 쉽게 이해하는 방식으로 제공되어야 한다. 또한, 시스템 완전 자동화human out of the loop, HOOTL 방식의 의사 결정 관점으로 수행되는 스마트 치안 시스템은, 서비스의 목적, 작동 방식 및 목표를 모든 사용자에게 미리 설명함으로써 사용자가 목적에 맞게 사용하도록 유도하여야 한다.

참고 스마트 홈 치안 시스템의 서비스 목적, 설명 사례[148]



서비스 목적: 집과 가족을 보호할 뿐만 아니라 현대적인 스마트 홈의 이점도 누린다. 또한 대부분 쉽게 설치하고 배선이 없으며 초기 비용이 저렴하다.
스마트폰으로 시스템을 제어하는지, 스마트 홈 시스템이 해킹되는지 등의 내용을 상세하게 설명해 준다.

15-1b 서비스의 한계와 범위에 대한 설명을 제공하는가?

Yes ☐ No ☐ N/A ☐

- 서비스 제공 범위와 한계를 설명함으로써 사용자가 기대치를 조정한다. 서비스 결과에 대한 품질은 사용자 그룹 특성, 사용 환경, 사용 데이터 등 다양한 요인에 영향받아 결과가 도출되므로 사용자에게 서비스 한계와 제공 범위에 대해 말하는 것이 중요하다.
- 스마트 치안 시스템은 시설물, 공공 도로, 교통, 가정, 사무실 및 기타 건물에 향상된 치안 및 감시 기능을 제공하고자 설계되며, 시스템이 제공하는 서비스 범위는 특정 제품이나 공급 업체에 따라 달라진다. 일부 시스템은 기본 모니터링 및 경보 기능을 제공하며, 다른 시스템은 얼굴 인식, 움직임 감지 및 자동 응답과 같은 고급 기능을 제공한다.

스마트 치안 서비스의 서비스 제한 및 서비스 제공 범위 설명 예시

서비스 제한 구분	설명
커버 영역 제한	스마트 치안 시스템은 일반적으로 커버하는 영역에 제한이 있다. 이는 특정 센서와 카메라에 의존하며, 핵심 영역에서 활동을 포착하고자 전략적으로 배치되어야 하기 때문이다. 사용자는 다양한 방법으로 이러한 제한 사항이나 조건 설정에 관해 확인하여야 한다.
오경보	스마트 치안 시스템의 한 가지 도전 과제는 잘못된 경보[false alarm]의 가능성이다. 이는 잘못된 설정이나 센서를 작동시키는 환경 조건을 포함한 다양한 요인에 의해 발생한다. 사용자는 어떤 조건에서 경보가 작동되는지 또는 다양한 방안으로 상세한 경보 정보를 알아야 한다.
기술적 이슈	스마트 치안 시스템은 기술에 의존하므로 소프트웨어 버그, 연결 문제 및 하드웨어 고장과 같은 기술적 문제에 영향을 받는다. 사용자는 보안 침해, 업데이트 또는 시스템 장애에 대한 자세한 보고서를 받는다.
개인정보 관리	스마트 치안 시스템이 더욱 발전함에 따라, 개인정보와 데이터 치안에 대한 우려가 증가한다. 일부 사용자는 자기 집과 개인 공간을 모니터링하는 카메라와 센서에 대한 아이디어에 불편함을 느낀다. 사용자들은 개인 및 민감한 데이터의 사용에 대해 알아야 하고, 개발자들은 개인 데이터 사용에 대해 투명하게 공개하여 사용자들에게 개인 데이터 사용에 대한 설명을 제공한다.

15-2 사용자가 상호작용하고 있는 대상을 인지할 수 있도록 설명하는가?

Yes ☐ No ☐ N/A ☐

- 최근 인공지능 시스템을 의인화함으로써 사용자가 친밀감을 향상하고 사용성을 높이려는 서비스가 많아진다. 그러나 인공지능 기술이 고도화되며 인간과 구분이 어려워져 사용자는 상호 작용의 대상이 사람인지, 시스템인지 혼란을 겪는다. 따라서 서비스 제공자는 사용자가 상호 작용하는 대상을 명확히 알림으로써 사용자가 겪는 혼란을 줄여야 한다.

15-2a 사용자와 인공지능이 상호작용하는 서비스 범위를 명시하였는가?

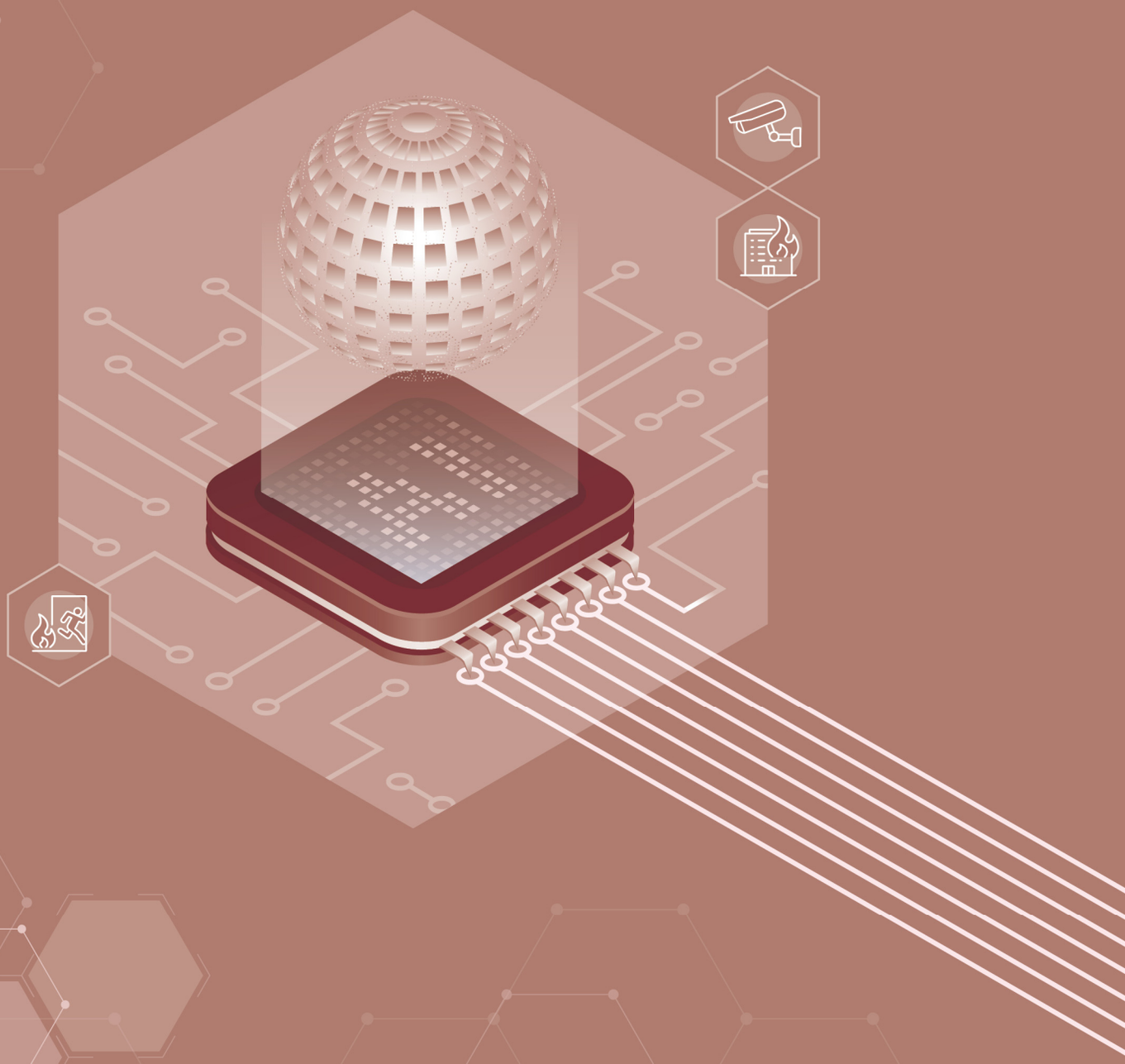
Yes ☐ No ☐ N/A ☐

- 서비스 내에서 사용자와 인공지능이 상호작용하는 범위를 명시해야 한다. 이로써 사용자는 어떤 작업이 자동화되고 어떤 작업을 직접 수행해야 하는지 이해할 수 있다. 또한, 인공지능과 상호작용하는 시점에는 사용자의 혼란을 방지하고 서비스에 대한 기대치를 조정할 수 있다.
- 특히 시스템 완전 자동화[HOOTL] 유형의 치안 시스템은, 인공지능이 서비스의 최종 의사 결정에 직접 또는 간접적으로 개입하는 방법을 명시한다.

15-2b 서비스 내의 최종 의사결정을 수행하는 주체에 대해 사용자에게 설명하는가?

Yes ☐ No ☐ N/A ☐

- 사용자에게 인공지능이 최종 의사결정을 내렸는지 또는 특정 결과에 기여했는지 등의 정보를 설명해야 한다. 예를 들어, 인공지능이 최종 의사결정을 내린 경우 사용자에게 해당 결정이 인공지능의 결과임을 명시적으로 사용자에게 전달해야 한다. 또한, 인공지능이 조언을 제시하고 최종 의사결정을 운영자가 내린 경우나, 사용자에게 최종 의사결정을 위임한 경우에도 관련 설명을 제공해야 한다.
- 미국 백악관에서 발표한 Blueprint for an AI Bill of Rights에서는 자동화 시스템이 채용이나 신용평가 등의 분야에서 사용될 경우 사람들의 삶에 깊은 영향을 미치기 때문에, 잠재적인 피해로부터 보호하기 위해 사용자에게 자동화 시스템의 활용 여부를 명시해야함을 언급하고 있다.
- 사용자가 시스템과 처음 상호 작용할 때, 시스템 능력과 제한을 설명하는 명확한 정보를 함께 제공한다. 사용자 매뉴얼 또는 가이드를 제공하거나, 시스템의 기본 기능을 설명하는 간단한 튜토리얼을 제공한다.



PART 3

부록

01 약어표

약어표

3D	Three-dimensional
ABC	Automated Border Control
ADASYN	Adaptive Synthetic Sampling
AI	Artificial intelligence
AIGA	Artificial Intelligence Governance and Auditing
ALP	Adversarial logit pairing
ALTAI	Assessment List for Trustworthy Artificial Intelligence
ANOVA	Analysis of variance
APE-GAN	Adversarial Perturbation Elimination with GAN
API	Application Programming Interface
AR	Augmented reality
ART	Adversarial Robustness Toolbox
AUC	Area under the curve
BDPL	Boundary Differentially Private Layer
C&W	Carlini-Wagner
CASIA	The Institute of Automation, Chinese Academy of Sciences
CCTV	Closed circuit television
COMPAS	Correctional Offender Management Profiling for Alternative Sanctions
CSRF	Cross-Site Request Forgery
CUSUM	Cumulative sum
CV	Cross validation
CVE	Common vulnerabilities and exposures
DACOBS	Davos Assessment of Cognitive Biases Scale
DAS	Domain Awareness System
DNA	Deoxyribonucleic acid
DNN	Deep neural network
DoD	United States Department of Defense
DoS	Denial of Service
DTTP	Deterministic trust transfer protocol
DVC	Data version control

01 약어표

EC	European Commission
ENISA	The European Union Agency for Cybersecurity
ETL	Extract, transform, and load
ETSI	European Telecommunications Standards Institute
EU	European Union
FAR	False acceptance rate
FDC	Fault detection and classification
FGSM	Fast Gradient Signed Method
FISMA	Federal Information Security Management Act
FNR	False negative rate
FPR	False positive rate
FPS	Frames per second
GAN	Generative adversarial network
GDPR	General Data Protection Regulation
GEI	Gait energy image
GradCAM	Gradient-weighted class activation mapping
GPL	General Public License
HMI	Human-machine interaction
HOOTL	Human-out-of-the-loop
HW	Hardware
IBM	International Business Machines
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
IEMOCAP	Interactive emotional dyadic motion capture database
IoT	Internet of Things
JSON	JavaScript Object Notation
ISO	International Organization for Standardization
KCI	Korea Citation Index
KS	Korean Standards
LED	Light-emitting diode
LIME	Local Interpretable Model-agnostic Explanations
MIL	Multiple instance learning
ML	Machine learning

ML4TIS	Machine Learning for Irregular Time Series Project
MLE	Maximum likelihood estimation
MSE	Mean squared error
N/A	Not applicable
NIST	National Institute of Standards and Technology
NLG	Natural Language Generation
NYPD	New York City Police Department
OECD	The Organization for Economic Cooperation and Development
OpenCV	Open Source Computer Vision Library
OSI	Opens Source Initiative
OSSRA	Open Source Security and Risk Analysis
OWASP	Open Worldwide Application Security Project
PC	Personal computer
PCA	Principal Component Analysis
PGD	Projected Gradient Descent
PIPA	Personal Information Protection Act
PRADA	Protecting against DNN Model Stealing Attacks
QA	Quality assurance
QR	Quick Response
RFE	Recursive Feature Elimination
RMF	Risk management framework
ROC	Receiver Operating Characteristic curve
ROC-AUC	Area under the Receiver Operating Characteristic curve
RONI	Reject on Negative Impact
RWF-2000	Real-world fighting video dataset with 2000 video clips
SAI	Securing Artificial Intelligence
SAP	Systems, Applications & Products in Data Processing
SAS	Statistical Analysis System
SCA	Software Composition Analysis
SHAP	Shapley Additive Explanations
SMOTE	Synthetic Minority Oversampling Technique
SPI	Security Parameters Index
SQL	Structured Query Language

01 약어표

STT	Speech-to-text
SVCL	Statistical visual computing laboratory
SVM	Support vector machine
SVN	Apache Subversion
SW	Software
TIBCO	The Information Bus Company
TTA	Telecommunications Technology Association
UCSD	University of California San Diego
UDHR	Universal Declaration of Human Rights
UN	United Nations
VAE	Variational autoencoder
VR	Virtual reality
WEF	World Economic Forum
WHO	World Health Organization
XAI	Explainable artificial intelligence
XML	Extensible Markup Language
XSS	Cross-site scripting

02 용어표

용어표

용어명	정의
보안 Security	보안은 기밀성, 무결성 및 가용성을 보장하고, 데이터, 시스템, 네트워크 및 기타 자산을 외부 위협에서 보호하는 일련의 조치와 원칙을 의미한다. 이는 정보 기술(IT) 분야뿐 아니라 물리적 환경과 사회적 측면에서 발생하는 위협에 대한 대비와 방어를 의미한다. 보안은 다양한 영역에서 중요한 역할을 한다. 아래는 몇 가지 중요한 보안 측면에 대한 개략적인 배경 정보이다: **기밀성(Confidentiality)**: 정보나 자산이 무단 액세스에서 보호되어야 한다. 이것은 민감한 데이터가 누출되지 않도록 하는 것을 의미한다. 기밀성을 유지하고자 데이터 암호화 및 엑세스 제어 등의 기술 및 절차를 사용한다. **무결성(Integrity)**: 정보나 자산이 무단 변경에서 보호되어야 한다. 이는 데이터가 정확하고 변조되지 않도록 함을 의미한다. 데이터 무결성을 보호하고자 체크섬 및 해시값을 사용하는 등의 방법이 사용된다. **가용성(Availability)**: 정보와 시스템은 필요할 때 항상 사용 가능하여야 한다. 이것은 서비스 중단을 방지하고 재해 복구 계획을 수립하여 시스템의 가용성을 보장함을 의미한다. **위협과 공격**: 악의적인 개체들은 다양한 방법으로 보안을 침해하려고 시도한다. 이러한 위협과 공격은 악성 코드, 해킹, 사회 공학, 물리적 침입 및 기타 다양한 형태로 나타난다. **보안 조치**: 보안을 강화하고자 다양한 조치가 취해진다. 이는 방화벽, 바이러스 백신, 액세스 제어, 교육 및 교육 등을 포함한다. **규정 및 규제**: 많은 산업과 국가에서는 보안을 강화하는 규정과 규제를 시행한다. GDPR(일반 데이터 보호 규정) 및 HIPAA(보건 정보 포트폴리오 보호법) 등 규정은 개인정보 보호와 의료 정보 보호를 강화하고자 도입되었다. **사용자 교육**: 보안은 기술적인 측면뿐 아니라 사용자 교육도 중요하다. 사용자는 안전한 비밀번호 사용, 피싱 공격에 대한 경각심 등을 갖고 행동하여야 한다.

02 용어표

용어명	정의
치안 Policing	치안은 특정 지역, 도시, 국가 또는 기타 장소에서 사람들과 재산에 대한 안전 보장과 관련된 개념이다. 치안은 범죄 예방, 법 집행, 비상 상황 대비 및 재난 관리 등 다양한 활동을 통해 유지된다. 치안은 사회적 안전과 개인 안전 모두를 포함하며, 공공장소 및 사회 전반에서 안전한 환경 유지가 목표이다. 치안은 사회적 안전과 개인 안전을 보호하고 강화함을 의미한다. 이는 정부, 경찰, 사회단체 및 개인 간 협력으로 달성된다. 치안은 다음과 같은 중요한 측면을 포함한다: **범죄 예방**: 범죄 예방은 범죄 발생률을 낮추는 데 중요한 역할을 한다.따라서 사회 프로그램, 교육, 환경 개선 및 경찰 활동을 통해 범죄를 방지하려고 한다. **법 집행**: 경찰 및 법 집행 당국은 법과 질서를 유지하고 범죄를 조사하며 범죄자를 처벌하는 역할을 한다. **비상 상황 대비**: 치안은 자연재해, 대규모 사고 또는 테러리즘 등 비상 상황 대비도 포함한다. 따라서 비상 계획 및 자원 마련과 국가 안전 유지가 중요하다. **재난 관리**: 치안은 홍수, 지진, 산불 및 기상 이변 등 재해에서 사람들을 보호하는 것을 포함한다. 따라서 비상 대응, 구조 작업 및 재해 관리 체계가 필요하다.
보안 취약점 Security Vulnerability	컴퓨터 시스템, 소프트웨어, 네트워크 또는 정보 시스템의 설계 또는 구현에서 발견된 보안 문제나 결함을 가리키는 용어이다. 이러한 취약점은 악의적인 공격자나 해커가 시스템에 접근하거나 악용하는 잠재적인 입구를 제공한다. 시스템의 정상적인 작동 중에 식별되거나 테스트 중에 드러나며, 이러한 취약점이 존재할 때, 시스템은 해킹, 데이터 유출, 무단 접근, 서비스 거부 공격 등 다양한 보안 위협에 노출된다. 보안 전문가들은 취약점을 식별하고 해결하여 시스템의 보안성을 강화하고 잠재적인 위협에서 보호한다.
CVE Common Vulnerabilities and Exposures	컴퓨터 시스템 및 소프트웨어에서 발견된 보안 취약점과 취약점에 대한 정보를 정리하고 표준화하는 데 사용되는 공개적인 식별 체계이다. 다양한 보안 전문가 및 조직 간 취약점 정보를 공유하고 협력하는 데 사용되며, 보안 취약점에 대한 효과적인 관리와 대응을 지원한다.
스마트 시티 Smart City	스마트 시티는 현대 도시 환경을 더 효율적으로, 지속 가능하게, 혁신적으로 관리하고 개선하는 개념이다. 스마트 시티는 다양한 정보 통신 기술과 인프라를 활용하여 도시 기능을 개선하며 시민 삶의 질을 향상하는 데 중점을 둔다. 이러한 개념은 기존의 도시 시설 및 서비스를 향상함과 동시에 환경, 경제, 사회 측면에서 지속 가능한 변화를 촉진한다. 스마트 시티의 주요 특징과 원칙은 다음과 같다: 정보 기술 활용, 지능적인 인프라, 지속 가능성, 시민 참여, 안전과 보안, 경제적 발전

02 용어표

용어명	정의
도메인 인식 시스템 Domain Awareness System	주로 보안 및 안전 관련 분야에서 사용되는 정보 기술 시스템이다. 특정 지역 또는 시스템 현황을 실시간으로 모니터링하고 분석하여 상황에 대한 포괄적인 이해와 관리를 제공한다. 도메인 인식 시스템은 다음과 같은 기능을 수행한다: 데이터 수집, 데이터 통합, 실시간 모니터링, 분석과 예측, 의사 결정 지원
공공 안전 영역 Public Safety Area	공공 안전은 도시, 지역, 국가 또는 다른 지역 사회에서 시민의 안전과 보호를 보장하고자 정부와 관련 당국, 기관 및 조직이 수행하는 활동 및 노력을 포함하는 분야를 가리킨다. 공공 안전은 다양한 측면과 상황에서 발생하는 위험, 재난, 범죄 및 응급 상황에 대응하며 시민의 생명, 재산 및 복지를 보호하고 증진하는 데 중점을 둔다. 공공 안전 영역은 다음과 같은 주요 구성 요소를 포함한다: 범죄 예방과 대응, 응급 서비스, 재난 관리, 교통안전, 보안 및 경찰 업무, 방재 및 위기 대응, 소방 서비스, 시민 교육과 인식
메타데이터 Metadata	데이터를 설명하는 정보의 집합을 가리키는 개념이다. 다시 말해, 메타데이터는 주로 다른 데이터를 분류, 관리, 검색, 이해 및 해석하는 데 사용되는 데이터의 특성과 속성을 설명하는 데이터이다. 메타데이터는 정보의 의미, 구조 및 내용을 더 잘 이해하고 조직화하는 데 도움을 준다. 메타데이터는 다양한 분야에서 중요한 역할을 한다. 예를 들어, 웹 검색 엔진은 웹 페이지의 제목, 메타 설명, 키워드 및 다른 서술 메타데이터를 분석하여 검색 결과를 제공하며, 라이브러리는 서술 메타데이터를 사용하여 도서 및 연구 자료를 분류하고 관리한다. 또한, 데이터베이스 시스템은 구조 메타데이터를 활용하여 데이터베이스의 스키마를 정의하고 쿼리를 실행한다.
보호변수 Protective attribute	데이터 분석 및 기계 학습에서 사용되는 중요한 개념 중 하나이다. 이 용어는 데이터셋에서 특정 개체 또는 그룹을 다른 개체 또는 그룹에서 보호하거나 구별하는 데 사용되는 특성 또는 속성을 가리킨다. 이러한 보호변수는 특정 인권 문제, 공정성 문제, 차별 문제 등을 다루는 데 중요한 역할을 한다. 주요 포인트: 보호변수의 목적, 데이터 프라이버시와 연결, 공정성과 균형, 보호변수 예시. 보호변수를 관리하고 활용함으로써, 데이터 분석 및 기계 학습 모델이 공정하며 차별 없는 예측과 결정을 내리도록 돕는 데 기여한다. 이는 공정한 사회와 기술 활용을 촉진하는 데 중요한 역할을 한다.

02 용어표

용어명	정의
오픈 소스 데이터 세스 Open Source Dataset	무료로 접근하고 사용하는 데이터의 모음을 나타낸다. 이러한 데이터셋는 주로 공개적으로 이용 가능하며, 보통 데이터의 복사, 수정, 공유 및 재배포가 허용된다. 오픈 소스 데이터셋는 연구, 교육, 소프트웨어 개발, 데이터 분석, 머신 러닝 및 다양한 다른 분야에서 활용된다. 오픈 소스 데이터셋는 다양한 분야에서 중요한 자원으로 활용되며, 오픈 소스 커뮤니티와 데이터 과학 커뮤니티에서 데이터 공유와 협업을 촉진하는 데 기여한다. 이러한 데이터셋는 공공 데이터, 연구 기관이나 기업에서 제공되거나 커뮤니티 기반으로 수집 및 유지 보수된다.
데이터 중독 Data Poisoning	일반적으로 데이터 보안 및 컴퓨터 과학 분야에서 사용되는 개념이다. 이 용어는 데이터의 무결성을 침해하거나 데이터를 손상하는 행위나 과정을 가리킨다. 데이터 중독은 데이터의 정확성, 신뢰성 및 무결성을 악화시키는 공격의 한 형태로 볼 수 있다. 주요 특징 및 예시는 데이터 조작, 악의적인 삽입, 데이터 유출 및 침투, 보안 취약점 이용, 암호화 해독, 악의적인 데이터 입력이다. 데이터 중독은 사이버 보안 위협 중 하나로, 정보 시스템 및 데이터의 안전을 위협한다. 따라서 데이터 중독을 방지하고 탐지하는 보안 조치와 솔루션을 적용하는 것이 중요하다. 이러한 조치에는 방화벽, 악성 코드 검출 및 침입 탐지 시스템(IDS), 취약점 관리 및 데이터 보안 솔루션 등이 포함된다.
데이터 회피 Data Evasion	일반적으로 치안 시스템, 침임 탐지 시스템(IDS), 침입 방지 시스템(IPS), 바이러스 스캐너 또는 다른 보안 메커니즘을 속이고자 시도하는 악의적인 활동을 가리킨다. 데이터 회피의 목표는 치안 시스템이나 소프트웨어가 악의적으로 활동하거나 차단하는 것을 우회하고 악의적인 동작을 감추는 것이다. 데이터 회피는 치안 시스템을 헷갈리게 하거나 악의적인 활동을 더 어렵게 탐지하도록 설계된다. 예를 들어, 다음과 같은 방식으로 데이터 회피를 수행한다: 암호화 회피, 시그니처 회피, 포맷 변환, 포함 및 숨김, 타이밍 공격
데이터 라벨링 Data Labeling	기계 학습 및 인공 지능 모델을 훈련하고자 데이터에 주석을 달거나 데이터 포인트를 범주 또는 클래스로 분류하는 과정을 가리킨다. 데이터 모델링은 모델의 학습을 지원하며 모델링 훈련 데이터로 패턴을 학습하고 예측을 수행하도록 한다. 데이터 라벨링은 다양한 분야에서 활용되며, 예를 들어, 음성 인식, 이미지 분류, 자연어 처리, 의료 진단, 자율 주행 자동차 및 다른 기계 학습 응용프로그램에서 필수적이다. 이 작업은 종종 인간 라벨러나 라벨링 플랫폼을 통해 수행되며, 큰 데이터셋를 처리할 때는 자동화된 라벨링 도구 및 기술도 사용된다.

02 용어표

용어명	정의
오픈 소스 라이브러리 Open Source Library	소프트웨어 개발자 및 컴퓨터 프로그래머가 무료로 사용하고 수정하는 소프트웨어 코드와 함수의 모음을 가리킨다. 이러한 라이브러리는 컴퓨터 프로그램을 개발하고 확장하는 데 도움을 주며, 일반적으로 특정 작업을 수행하는 함수, 클래스, 모듈 또는 코드 조각을 제공한다. 오픈 소스 라이브러리는 개발자 커뮤니티에 의해 지속해서 향상되며, 소프트웨어 개발 프로젝트의 생산성과 효율성을 향상하는 데 기여한다. 또한 오픈 소스 라이브러리를 사용함으로써, 많은 소프트웨어 프로젝트가 비용과 시간을 절약하면서도 고품질의 소프트웨어를 개발한다.
모델 추출 공격 Model extraction attack	기계 학습 모델 또는 머신 러닝 모델의 내부 구조와 동작을 이해하거나 복제하려는 악의적인 시도를 가리킨다. 이러한 공격은 주로 흑자 공격(Black-box Attack)과 백상 공격(White-box Attack)의 두 가지 방식으로 나뉜다. 모델 추출 공격은 주로 온라인 서비스, 클라우드 기반 머신 러닝 모델, 자연어 처리 모델, 이미지 분류 모델 및 기타 기계 학습 모델을 대상으로 한. 이러한 공격은 모델의 민감한 정보 유출, 보안 위험 및 데이터 개인정보 유출로 이어지므로 모델 보안 및 방어 메커니즘을 개발하는 연구가 중요하게 다뤄진다.
모델 회피 공격 Model evasion attack	모델 회피 공격(Model Evasion Attack)은 기계 학습 모델을 조작하거나 속이려는 악의적인 시도를 나타낸다. 이러한 공격은 모델이 주어진 입력 데이터를 잘못 분류하거나 잘못 예측하도록 하려는 목적으로 수행된다. 모델 회피 공격은 주로 분류 모델(classification models)을 대상으로 하며, 분류 모델은 입력 데이터를 여러 범주 중 하나로 분류하는 작업을 수행한다. 모델 회피 공격의 주요 특징과 방법: 입력 데이터 조작, 적대적 예제(Adversarial Examples), 모델 악용, 보안 검출 회피, 얕은 학습 모델 회피 공격은 보안과 프라이버시 문제를 일으키며, 모델의 견고성을 향상하고 이러한 공격을 방어하는 연구가 계속 진행 중이다.
XAI eXplainable Artificial Intelligence	해석 가능한 인공지능은 기계 학습 및 딥러닝 모델의 예측, 결정 또는 출력을 인간이 이해하고 설명하는 방식으로 제공하려는 노력을 가리킨다. XAI의 목표는 머신 러닝 및 인공지능 모델이 내부 작동 및 의사 결정 기준을 투명하게 밝히고, 사용자가 모델 동작을 이해하도록 하는 것이다. XAI는 다양한 분야에서 중요한 역할을 한다. 예를 들어, 의료 진단에서는 의사 및 환자가 모델의 의사 결정을 이해하고 신뢰하여야 한다. 금융 분야에서는 금융 모델이 왜 특정 대출을 승인 또는 거부하는지 설명하여야 한다. 자동차 자율 주행에서는 모델이 도로 상황을 어떻게 인식하고 판단하는지 이해하여야 한다. XAI 기술은 모델 해석성을 향상하는 다양한 방법을 개발하고, 이러한 기술은 모델의 믿음성을 높이고 모델의 사용자에게 더 나은 투명성을 제공하는 데 사용된다.

02 용어표

용어명	정의
사용자 인터페이스 User Interface	사용자와 컴퓨터 또는 다른 디지털 장치 간 상호 작용을 가능하게 하는 시스템 또는 소프트웨어의 일부분을 나타낸다. UI는 사용자가 디지털 시스템을 조작하고 컴퓨터 프로그램 또는 앱을 사용하는 데 필요한 그래픽, 텍스트, 음성 또는 기호적 요소를 포함하며 사용자와 시스템 간 정보 및 명령 흐름을 조절한다. 사용자 인터페이스 디자인은 사용자 경험(UX)과 밀접하게 관련되며, 사용자가 시스템을 쉽게 이해하고 조작하도록 하는 것이 중요하다. 사용자 인터페이스 디자이너는 사용자 중심의 디자인 원칙을 적용하여 직관적이고 효율적인 UI를 만들고, 사용자의 편의성과 만족도를 향상하는 데 기여한다.
안전 모드 Safe Mode	안전 모드는 주로 Microsoft Windows 운영 체제에서 널리 사용되며, 사용자가 문제 해결 및 시스템을 복구하고자 부트할 때 선택하는 옵션 중 하나이다. 안전 모드는 일반적으로 네트워크 기능이 비활성화되어 시스템 보안을 강화하고, 핵심 시스템 파일만 로드하여 안정성을 높이는 것을 특징으로 한다. 다른 운영 체제 및 소프트웨어에서도 비슷한 기능이 제공되며, 이는 시스템 문제 해결과 보안 검사를 하는 중요한 도구 중 하나로 사용된다.
사용자의 특성 User characteristics	컴퓨터 과학 및 인간-컴퓨터 상호 작용 분야에서 중요한 개념으로, 컴퓨터 시스템, 소프트웨어, 웹 사이트, 모바일 애플리케이션 등과 상호 작용하는 개별 사용자의 특징과 특성을 나타낸다. 이러한 특성은 사용자의 신체적·심리적·문화적·기술적·환경적 특징 및 선호도를 포함하며, 사용자 경험 디자인, 보안, 액세스 가능성, 개인화, 마케팅, 데이터 분석 등 다양한 분야에서 중요한 역할을 한다.
데이터 계보 Data Lineage	데이터 계보는 정보 기술 및 데이터 관리 분야에서 사용되는 개념으로, 데이터의 생성, 수정, 이동, 공유 및 삭제와 관련된 모든 활동을 기록하고 추적하는 프로세스를 가리킨다. 데이터 계보는 데이터의 변화를 체계적으로 관리하고 데이터의 신뢰성과 투명성을 유지하는 데 도움이 된다. 데이터 계보는 데이터 품질, 규정 준수, 보안, 오류 복구, 데이터 분석 및 보고 등과 관련된 다양한 목적으로 사용된다.
개인정보 보호 Metadata	개인정보 보호는 개인정보를 수집, 저장, 처리 및 공유하는 모든 조직 및 개인에 대한 의무이다. 이는 개인정보가 무단으로 액세스, 유출 또는 남용되지 않도록 보장하는 과정 및 원칙을 가리킨다. 개인의 기본적인 권리와 자유를 존중하도록 기업 및 정부 조직이 개인정보를 적절히 다루도록 하는 것이 중요하다.
퍼스널 케어 로봇 Personal Care Robot	퍼스널 케어 로봇은 인간의 건강, 편의 및 복지를 개선하고자 설계된 자동화된 기계 장치 또는 로봇이다. 이러한 로봇은 일상생활에서 다양한 기능을 수행하며, 환자, 노인, 또는 장애인 등 특별한 요구사항이 있는 개인들에게 특히 유용하다. 퍼스널 케어 로봇은 의료 관리, 건강 모니터링, 일상생활 지원, 심리적 지원 및 커뮤니케이션을 포함한 다양한 작업을 수행하며, 인공지능 및 센서 기술의 진보로 더욱 효과적으로 작동한다.

02 용어표

용어명	정의
딥 페이크 콘텐츠 Deepfake Content	인공지능(AI) 및 딥러닝 기술을 사용하여 생성된 가짜 비디오, 오디오, 이미지 또는 텍스트 콘텐츠를 지칭하는 용어이다. 이러한 콘텐츠는 기존 콘텐츠나 개인의 목소리, 외모, 언어 스타일을 모방하여 생성되며, 실제로는 해당 내용을 생성한 사람이나 캐릭터와 다르게 나타난다. 딥 페이크 콘텐츠는 주로 AI 모델, 특히 생성적 적대 신경망(GAN, Generative Adversarial Network)을 사용하여 만들어진다.
엣지 디바이스 Edge Device	엣지 디바이스는 컴퓨팅 자원과 데이터 처리 능력이 있는 장치로, 클라우드 컴퓨팅 리소스와 협력하여 데이터 처리 및 분석을 지역적으로 또는 분산된 환경에서 수행하는 장치를 가리킨다. 이러한 디바이스는 주로 현장에서 데이터를 생성하는 것과 관련 있으며, 데이터를 원활하게 수집, 처리, 저장 및 분석하는 데 사용된다. 엣지 디바이스는 네트워크 지연 시간을 줄이고 데이터 보안 및 개인정보 보호를 강화하는 데 도움이 된다.
크라우드소싱 Crowd Sourcing	일반 대중 또는 대중의 대규모 그룹, 일명 '크라우드(Crowd)'를 활용하여 정보, 아이디어, 자금, 노력 또는 기타 자원을 모으고 활용하는 비즈니스 또는 문제 해결 방법이다. 이러한 방식은 기업, 정부, 비영리 단체 등 다양한 조직이 해결하기 어려운 문제를 해결하거나 창의적인 아이디어를 얻는 데 사용된다. 크라우드소싱은 주로 온라인 플랫폼을 통해 이루어지며, 대부분은 보상 또는 인센티브를 통해 참여자들을 동원한다. 이 방법은 다양한 관심, 경험 및 전문 지식이 있는 다수의 사람을 결합함으로써 문제 해결과 혁신을 촉진한다.
위해 모델링 Harms Modeling	위해 모델링은 정보 시스템, 소프트웨어 애플리케이션, 물리적 시설 또는 기타 시스템의 보안을 강화하는 절차로, 잠재적 위험과 보안 취약점을 식별하고 이해하고자 사용되는 체계적인 방법론이다. 이 과정은 보안 공격에서 시스템을 보호하고 취약점을 사전에 식별하여 해결하는 방안을 도출하는 데 도움을 준다.
세계인권선언 UDHR(Universal Declaration of Human Rights)	세계인권선언은 1948년에 유엔 총회에서 채택된 중요한 국제 문서로, 모든 인간이 가진 기본적인 인권과 자유를 보호하고 존중하는 원칙을 명시한 문서이다. 이 선언은 모든 사람이 태어날 때 동등하고 자유로운 존재로서 존중받아야 한다는 원칙을 강조한다. 세계인권선언은 각국 정부와 국제 사회에 인권을 보장하고 증진하는 데 중요한 역할을 하며, 인권을 지키지 않는 행위에 대한 비판과 규탄을 담고 있다.
이니셔티브의 윤리 Ethics of Initiative	이니셔티브의 윤리는 조직이나 개인이 새로운 프로젝트, 아이디어 또는 활동을 개발하거나 시행할 때 미래의 결과 및 영향을 고려하고 윤리적인 원칙을 준수하는 데 중점을 둔 개념이다. 이니셔티브의 윤리는 행동의 출발점에 관한 윤리적 고려 사항을 다루며, 새로운 프로젝트나 변화를 추진할 때 윤리적인 판단과 책임을 강조한다. 이니셔티브의 윤리는 혁신과 개선을 추구하면서도 사회, 환경, 및 개인의 권리와 안전을 도모하며 균형을 유지하는 데 중요한 역할을 한다.

02 용어표

용어명	정의
치안 시스템 리콜 Security System Recall	치안 시스템 리콜은 기업 또는 조직이 치안 시스템 제품 또는 소프트웨어의 결함, 취약점 또는 기타 보안 위험으로 인해 해당 제품 또는 소프트웨어를 시장에서 회수하거나 수정하는 프로세스를 가리킨다. 이는 일반적으로 제품 또는 소프트웨어의 사용자에게 제품을 업그레이드하거나 보안 수정 사항을 적용하도록 하는 것을 포함한다. 치안 시스템 리콜은 보안 문제를 해결하고, 사용자의 데이터 및 시스템을 보호하고자 필요한 조치 중 하나로 이루어진다. 이 프로세스는 법적 규정 및 관련된 규정 및 보안 기준을 준수하여 진행된다.
STT 처리 Speech-to-Text Processing	STT 처리는 음성 신호를 텍스트로 변환하는 과정을 가리킨다. 이 기술은 음성 인식 및 음성 처리 분야에서 사용되며, 음성 입력을 컴퓨터가 이해하고 처리하는 텍스트로 변환하는 데 쓰인다. STT 처리는 음성 인식 시스템, 음성 검색, 자동 번역, 음성 명령 및 음성 기반 컴퓨터 인터페이스에서 중요한 역할을 한다. 일반적으로, 오디오 레코딩, 노이즈 제거, 음성 패턴 분석 및 음성을 텍스트로 변환하는 단계로 이루어진다. 이러한 기술은 음성 기반 응용 프로그램 및 시스템에서 사용자 경험을 향상하는 데 활용된다.
자기 조직화 연산 신경망 Self-Organizing Feature Map	자기 조직화 연산 신경망(SOFM)은 신경망의 한 종류로, 데이터의 비선형적인 특성을 탐색하고 데이터를 클러스터링하거나 분류하는 데 사용되는 비지도 학습 알고리즘이다. SOFM은 데이터의 내재한 구조를 발견하고 데이터의 특징을 추출하고자 주로 이용된다. 이 신경망은 지도 학습과는 달리 데이터에 대한 라벨이나 목표 출력이 없이 동작하며, 입력 데이터 간 유사성을 기반으로 데이터를 그룹화하는 데 도움을 준다.
하이퍼파라미터 Hyperparameter	기계 학습 및 딥러닝 모델의 학습 프로세스를 제어하고 조정하는 매개 변수이다. 이러한 파라미터는 모델 아키텍처나 학습 알고리즘과는 구별되며, 모델을 훈련하기 전에 사전에 설정되어야 한다. 하이퍼파라미터는 모델의 성능, 학습 속도 및 일반화 능력에 영향을 미친다.
AI 허브 AI Hub	AI 허브는 데이터셋, 모델, 도구, 교육 자료 등 다양한 인공지능(AI) 리소스에 대한 액세스를 제공하는 중앙 집중식 플랫폼 또는 리퍼지토리이다. AI 연구자, 개발자 및 애호가가 AI 프로젝트와 연구를 발전시키는 데 필요한 리소스를 검색, 공유 및 액세스하는 협업 공간 역할을 한다. AI 허브는 귀중한 AI 자산에 대한 단일 액세스 지점을 제공하여 AI 기술의 개발과 배포를 가속함으로써 AI 커뮤니티 내에서 혁신과 개발을 촉진하는 데 중요한 역할을 한다. 이러한 허브는 정부 기관, 교육 기관 또는 민간 조직에서 AI 관련 이니셔티브와 연구를 지원하고 촉진하고자 설립할 때가 많다.

용어명	정의
프락시 Proxy	컴퓨터 네트워크에서 중계 역할을 하는 서버 또는 소프트웨어를 가리킨다. 이 중계 서버 또는 소프트웨어는 클라이언트와 다른 서버 간 통신을 중계하고 제어하는 역할을 한다. 일반적으로 프락시는 클라이언트의 요청을 받아 중계 서버를 통해 인터넷에 있는 다른 서버로 전달한다. 프락시의 사용 목적은 다양하다. 주요 목적은 익명성 제공, 보안 강화, 캐시 서버를 통한 웹 페이지 로딩 최적화, 브랜드 위주의 인터넷 사용 제한 우회, 웹 트래픽 모니터링 그리고 접근 제어 등이다. 프락시 서버는 주로 웹 프락시, VPN(Virtual Private Network) 프락시 및 소켓 프락시 등 다양한 유형으로 사용된다.
프로파일러 Profiler	컴퓨터 프로그램, 애플리케이션 또는 시스템의 성능 및 동작을 분석하고 모니터링하고자 사용되는 소프트웨어 도구나 기술을 가리킨다. 프로파일러는 프로그램이 실행되는 동안 발생하는 다양한 활동을 추적하고 측정함으로써 성능 병목 현상을 식별하고 최적화하는 정보를 제공한다. 주로 프로그래머, 시스템 관리자 또는 성능 튜닝을 수행하는 전문가들에 의해 사용된다.
스마트 그리드 Smart Grid	스마트 그리드는 전기 공급 및 관리를 혁신적으로 개선하는 전력 시스템의 진화된 형태이다. 이러한 그리드는 고급 통신과 정보 기술을 통합하여 전력 생성, 분배, 및 소비를 효율적으로 조절하는 데 사용된다. 스마트 그리드는 전통적인 전력 인프라를 현대화하여 다음과 같은 목표를 달성한다: **효율성 향상**: 전력 네트워크를 더욱 효율적으로 운영하고 에너지 손실을 최소화한다. **신뢰성 증대**: 전력 공급의 안정성을 향상하여 정전 사고를 줄인다. **재생 에너지 통합**: 스마트 그리드는 재생 가능한 에너지원을 통합하여 친환경 전력 생산을 촉진한다. **고객 참여 증가**: 스마트 미터와 실시간 정보를 통해 고객이 에너지 사용을 모니터링하고 제어하는 기회를 제공한다. **데이터 분석 및 예측**: 스마트 그리드는 대량의 데이터를 수집하고 분석하여 전력 네트워크의 운영을 최적화하며 에너지 소비를 예측한다. 이러한 기능은 전력 공급자, 소비자 및 환경에 이익을 제공하며 전력 시스템의 혁신적인 발전을 나타낸다.

02 용어표

용어명	정의
하이브리드 치안 시스템 Hybrid Security System	하이브리드 치안 시스템은 물리적 보안 및 디지털 보안을 통합하여 조직 또는 개인의 보안을 강화하는 체계이다. 이 시스템은 전통적인 보안 요소와 현대적인 디지털 보안 요소를 결합하여 다양한 위협에서 보호한다. 이러한 시스템은 주로 실내 및 외부 감지 장치, CCTV 카메라, 액세스 제어, 바이오메트릭 인식, 네트워크 보안 및 암호화 기술을 통합한다. 하이브리드 치안 시스템은 보안 감시, 위험 관리, 사건 대응 및 보안 인프라의 통합적인 효율성을 향상하며 보안 환경의 다양한 측면을 다루는 데 사용된다. 이는 현대 사회에서 정보 보안과 물리적 안전의 결합으로 인해 중요성을 더하며, 기업, 정부 및 개인 보안 요구사항을 충족시키고자 채택된다.
레이더 데이터 Radar Data	레이더 데이터는 전파를 이용하여 물체의 위치, 거리, 속도, 크기, 형태 등 다양한 정보를 감지하고 측정하는 레이더 시스템에서 생성된 정보를 나타냄을 의미한다. 레이더는 주로 항공, 우주, 해양, 기상 및 군사 응용 분야에서 사용되며, 이 데이터는 보통 숫자 또는 이미지 형식으로 기록된다. 이러한 데이터는 항공기, 함선, 날씨 예측, 군사 작전, 항법 및 안전 시스템 등 다양한 분야에서 중요한 정보원으로 활용된다. 레이더 데이터는 레이더 시스템에 의해 수집된 실시간 정보로서, 이를 분석하여 다양한 목적에 활용한다.
합성 레이더 Synthetic Aperture Radar	레이더를 사용하여 지구 표면의 높은 해상도 영상을 생성하는 데 사용되는 센서 시스템이다. SAR 시스템은 특히 위성, 비행기 또는 드론에서 사용되며, 이동 중인 플랫폼에서 지상 또는 지하 대상을 모니터링하고 이미지로 표현하는 데 널리 활용된다. SAR은 레이더 파형을 반복하여 수집하고 처리하여 고해상도 합성 이미지를 생성한다. 이러한 이미지는 지형, 지질, 환경 모니터링, 자연재해 감지, 군사 정보 수집 및 다양한 다른 응용 분야에서 중요한 정보를 제공한다.
데이터 정리 Data Cleansing	데이터 스크러빙이라고도 하는 데이터 정리는 데이터셋의 오류나 불일치를 식별하고 수정하는 프로세스를 말한다. 이 중요한 데이터 관리 관행에는 누락된 정보, 중복, 서식 문제 및 기타 불일치 등 부정확성을 감지하고 수정하여 데이터가 정확하고 신뢰할 수 있으며 의도된 용도에 적합한지 확인하는 것이 포함된다. 데이터 정리는 데이터 품질을 개선하고, 오해의 소지가 있는 분석을 방지하며, 정보에 입각한 의사 결정 지원을 목표로 한다. 이 프로세스에는 일반적으로 데이터의 무결성과 유용성을 유지하는 데이터의 검증, 변환 및 보강이 포함된다.

용어명	정의
규제 샌드박스 제도 Regulatory Sandbox	정부 또는 규제 기관이 새로운 기술, 서비스 또는 비즈니스 모델을 실험하고 시험하도록 하는 혁신적인 규제 절차 또는 프로그램을 가리키는 용어이다. 이러한 제도는 기술 혁신과 창업을 촉진하며, 새로운 아이디어와 기술이 기존 규제에 어긋나지 않으면서 발전하도록 하고자 사용된다. 제품 또는 서비스의 개발자 및 제공 업체가 일정 기간 특정 규제 기준을 준수하지 않고 시장에서 실험하고 성장하도록 허용한다. 이를 통해 혁신을 촉진하고 새로운 비즈니스 모델이나 기술이 시장에서 성공할 기회를 제공한다. 규제 샌드박스 제도는 기존 규제 프레임워크와 혁신 사업의 상충을 최소화하면서 혁신을 지원하는 데 도움이 되며, 주로 금융 부문에서 사용될 때가 많지만 다른 산업 부문에서도 확대된다.
데이터 편향 Data Bias	데이터 편향은 데이터 집합 내 또는 데이터 수집 과정에서 발생하는 편향, 부정확성, 또는 비대표성을 나타내는 용어이다. 이는 주로 기계 학습 및 인공지능 모델에서 중요한 역할을 한다. 데이터 편향은 다양한 형태로 나타나며, 이는 데이터가 특정 그룹, 특정 위치, 특정 시간 범위에 집중되거나 특정 유형의 정보가 누락되어 발생한다. 데이터 편향은 모델의 결과나 예측에 왜곡을 일으키며, 이에 따라 모델이 부정확하거나 특정 그룹이나 사람에 대한 편견을 보인다. 데이터 편향을 감지하고 수정하는 것은 공정하고 신뢰성 있는 모델을 개발하고 유지하는 중요한 단계이다. 데이터 편향은 다양한 분야에서 발생하며, 사회, 의료, 금융, 범죄 예측 등 여러 분야에서 주목받는다. 이를 해결하고자 데이터 수집, 전처리, 모델 개발, 및 평가 단계에서 신중하고 투명한 접근 방식이 필요하다.
테일러링 Tailoring	테일러링은 주로 컴퓨터 과학 및 정보 보안 분야에서 사용되는 용어로, 특정한 요구사항에 따라 보안 정책, 절차 또는 시스템을 맞춤 설정하거나 조정하는 프로세스를 가리킨다. 이는 조직의 고유한 보안 요구사항을 충족하고자 일반적인 보안 솔루션 맞춤화를 의미하며, 공격자의 취약점 악용을 방지하거나 사전에 탐지하고자 사용된다. 테일러링은 일반적으로 보안 정책, 네트워크 구성, 암호화 및 액세스 제어와 관련된 보안 조치에 적용된다. 이를 통해 조직은 고유한 위협에 대비하고 기존의 일반적인 보안 조치로는 충분하지 않을 때를 대비한다.
깊은 통합 Deep Integration	정보 기술 및 소프트웨어 시스템에서 사용되는 용어로, 다양한 컴포넌트, 애플리케이션 또는 서비스를 원활하게 통합하고 상호 작용시키는 프로세스나 기술을 가리킨다. 이것은 종종 시스템, 서비스 또는 애플리케이션 간에 데이터 및 기능을 공유하고 상호 작용을 향상하려는 것이며, 다른 시스템 또는 서비스와의 연계성을 향상하는 데 중요하다.

02 용어표

용어명	정의
서비스 거부 공격 DoS[Denial of Service]	컴퓨터, 네트워크, 웹사이트 또는 온라인 서비스를 의도적으로 마비시키려는 공격 형태이다. 대상 시스템 또는 네트워크를 과도하게 부하를 주거나 리소스를 소모하여 서비스의 정상적인 기능을 방해하려는 것이 목표이다.
CSRF 위협 Cross-Site Request Forgery Threat	웹 애플리케이션 보안에서 발생하는 위협 중 하나로, 사용자가 자신의 의지와 동의 없이 웹 애플리케이션에서 비정상적인 요청을 실행하도록 유도하는 공격 기법을 가리킨다. 이러한 공격은 사용자의 웹 브라우저가 인증된 세션을 유지한 상태에서 악의적인 웹 페이지나 이메일 링크를 통해 실행된다. 공격자는 희생자의 권한을 도용하여 데이터 변경, 금전 이체, 비밀 정보 노출 등을 유발한다.
인공지능 모델 공격 AI Model Attack	인공지능 모델 공격은 악의적인 목적으로 혹은 시스템의 취약점을 이용하여 인공지능 모델을 속이거나 손상하는 과정을 가리킨다. 이러한 공격은 기계 학습 및 딥 러닝 알고리즘을 대상으로 하며, 그 목적은 모델의 성능을 왜곡하거나 모델의 예측 결과를 변조하여 악의적인 목표를 달성하고자 사용된다. 이러한 공격은 주로 모델의 입력 데이터를 조작하거나, 모델의 가중치를 조작함으로써 수행되며, 그 결과로 모델의 신뢰성과 안전성에 영향을 미친다.
에스컬레이션 Escalation	에스컬레이션은 어떤 문제나 상황이 더 심각해지거나 복잡해질 때, 일반적으로 상위 단계나 더 높은 권한이 있는 개인이 관여하도록 상승함을 의미한다. 이것은 조직, 프로젝트, 협력 관계 등 다양한 맥락에서 발생한다. 에스컬레이션은 주로 결정을 내리는 데 도움을 주거나 문제를 해결하고자 할 때 사용된다. 일반적으로 이것은 다음 단계로 관련 정보나 의사 결정을 요청함을 나타내며, 문제가 더 이상 현재 단계나 수준에서 처리하기 어려울 때 사용된다.
HML Hierarchical Multi-Level	계층적 다수준 데이터 모델이나 시스템을 가리키는 용어로, 데이터를 계층적으로 구성하거나 시스템을 다수준으로 분석 및 관리하는 과정에서 사용된다. 이 용어는 주로 정보 기술, 데이터베이스 설계, 그래픽 디자인, 조직 구조, 또는 교육 분야에서 활용된다. HML은 주로 데이터 또는 시스템을 위계적인 구조로 표현하고, 각 수준의 요소가 다음 수준의 요소를 포함하는 방식으로 조직된다. 이것은 복잡한 정보를 더욱 효과적으로 표현하고, 분석하고, 관리하고자 사용된다.
휴먼 아웃 오브 더 루프 Human Out of the Loop	휴먼 아웃 오브 더 루프는 기술 또는 자동화 시스템에서 인간의 개입을 제외함을 나타내는 용어이다. 이것은 주로 인공지능, 자동화 소프트웨어, 로봇 또는 다른 자동화된 시스템이 작업을 수행하거나 결정을 내릴 때 인간 개입이 필요하지 않을 때 사용된다. 휴먼 아웃 오브 더 루프는 작업을 효율적으로 자동화하고, 오류를 줄이며, 작업의 일관성을 유지하고자 사용된다. 이 용어는 기술 발전과 자동화의 증가와 함께 더 많이 사용되며, 특히 산업 자동화 및 인공지능 분야에서 주목받는다.

용어명	정의
휴먼 온 더 루프 Human on the Loop	휴먼 온 더 루프는 자동화된 시스템이나 인공지능(AI) 기술을 통해 처리되는 작업 중에서 인간의 개입, 감독 또는 결정이 필요한 프로세스를 가리킨다. 이 용어는 주로 자율 주행 차량, 산업 자동화, 치안 시스템 그리고 기타 자동화 기술 관련 분야에서 사용된다. 휴먼 온 더 루프는 기계 학습 및 인공 지능 시스템의 한계를 극복하고 안전성을 확보하고자 도입된다. 이러한 시스템은 대부분의 작업을 자동으로 처리하지만, 예외적인 상황 또는 복잡한 결정이 필요할 때 인간의 개입이 필요하다. 이는 안전 문제, 윤리적 고려 또는 정확한 결정을 보장하고자 중요하다.

03 요구사항별 이해관계자

요구사항별 이해관계자

관련 표준*에 근거한 요구사항별 이해관계자

* TTAK.KO-10.1497, 인공지능 시스템 신뢰성 제고를 위한 요구사항

요구사항 번호	IT분야역량체계[ITSQF] 기반 정의		관련 표준 기반 정의
	대표 이해관계자(예)	협력 대상(예)	이해관계자
요구사항 01	• 정보기술기획자	• 데이터분석가 • 인공지능아키텍트 • SW아키텍트	AI 제공자, AI 생산자, AI 고객, AI 파트너, AI 영향대상, 관계기관
요구사항 02	• IT감사자	• 정보기술기획자 • SW아키텍트 • 데이터분석가	AI 제공자, AI 생산자, AI 고객, AI 파트너, AI 영향대상, 관계기관
요구사항 03	• IT품질관리자	• 정보기술기획자 • 인공지능아키텍트	AI 제공자, AI 생산자, AI 고객, AI 파트너
요구사항 04	• 데이터분석가	• 데이터아키텍트 • 정보기술기획자	AI 생산자, AI 고객, AI 파트너, AI 영향대상
요구사항 05	• 데이터분석가	• 데이터아키텍트	AI 생산자, AI 파트너
요구사항 06	• 데이터아키텍트 • 데이터분석가	• IT품질관리자 • 인공지능아키텍트	AI 생산자, AI 파트너
요구사항 07	• 인공지능SW개발자	• SW아키텍트	AI 생산자, AI 파트너
요구사항 08	• 인공지능SW개발자	• 인공지능아키텍트 • IT품질관리자	AI 생산자, AI 파트너
요구사항 09	• 인공지능아키텍트	• 인공지능SW개발자 • 데이터분석가	AI 제공자, AI 생산자, AI 고객, AI 파트너, AI 영향대상
요구사항 10	• 인공지능SW개발자 • 인공지능아키텍트	• UI/UX기획자 • 시스템SW개발자	AI 생산자, AI 고객, AI 파트너
요구사항 11	• 시스템SW개발자	• IT품질관리자	AI 생산자, AI 고객, AI 파트너
요구사항 12	• SW아키텍트	• 보안사고대응전문가 • 정보기술기획자 • IT품질관리자	AI 생산자, AI 고객, AI 파트너
요구사항 13	• UI/UX기획자	• 인공지능서비스기획자 • UI/UX개발자	AI 제공자, AI 생산자, AI 고객, AI 파트너
요구사항 14	• 데이터베이스관리자	• 인공지능서비스관리자 • 인공지능아키텍트 • 데이터아키텍트	AI 제공자, AI 생산자, AI 고객, AI 파트너, AI 영향대상
요구사항 15	• 인공지능서비스기획자	• 인공지능서비스관리자	AI 제공자, AI 생산자, AI 고객, AI 파트너

04 이해관계자 정의

이해관계자 정의

IT분야역량체계I[TSQF]에서 제시한 대표 이해관계자·협력 대상의 직업·직무 정의

직업명	직무 정의
정보기술기획자	조직의 경영목표 달성하기 위하여 정보기술 전략을 기획하고, 거버넌스, 투자성과 분석, 운영 정책, 연구개발, 프로세스, 아키텍처 등 분야별 전략을 수립하는 일이다.
IT감사자	IT를 운영하는데 있어 거버넌스 차원의 관련법, 제도, 내부 정책, 역할, 가이드라인, 규범, 기술표준 등을 준수하도록 지속적인 통제관리를 수행하는 일이다.
IT품질관리자	IT품질목표를 달성하기 위하여 전사적인 품질정책 및 관리체계를 수립하고 품질향상을 위해 교육 및 관리활동 등을 수행하며, 프로젝트 차원에서의 품질보증 활동을 수행하는 일이다.
데이터분석가	다양한 형태의 데이터로부터 유용한 정보를 찾고 예측하기 위해, 목적에 적합한 분석 기법을 적용하여 전처리, 탐색적 분석, 분석 모델링, 시각화를 수행하는 일이다.
데이터아키텍트	전사아키텍처와 데이터품질관리에 대한 지식을 바탕으로 전사에서 보유한 정형데이터와 비정형 데이터를 체계적, 구조적으로 정의하고 검증, 관리하는 일이다.
인공지능SW개발자	인공지능서비스 기획 목적에 부합하는 서비스를 구축하기 위해 모델링 및 데이터 분석 결과를 인공지능 플랫폼 환경에서 기능, 인터페이스, 지식화를 구현하고, 검증하는 일이다.
인공지능아키텍트	인공지능서비스 목적을 달성하기 위하여 학습데이터 탐색 과정을 통해 적합한 인공지능 모델을 도출하고, 최적의 인공지능 플랫폼을 분석·설계하는 일이다.
인공지능SW개발자	인공지능서비스 기획 목적에 부합하는 서비스를 구축하기 위해 모델링 및 데이터 분석 결과를 인공지능 플랫폼 환경에서 기능, 인터페이스, 지식화를 구현하고, 검증하는 일이다.
시스템SW개발자	운영체제 환경에서 시스템 자원을 제어 및 관리하는 소프트웨어와 응용프로그램의 동작을 위한 시스템 플랫폼의 요구사항 분석 및 설계, 구현, 배포를 수행하는 일이다.
SW아키텍트	소프트웨어의 기능, 성능, 보안 등의 품질을 보장하고 소프트웨어를 구성하는 요소와 관계를 분석, 설계하여 전체적인 소프트웨어 구조를 체계화하는 일이다.
UI/UX기획자	서비스의 본질적 특성에 대한 이해를 기반으로 트렌드 분석, 사용자 이용 행태 분석 등을 통해 이해관계자 및 사용자의 요구를 발굴하고 사용성을 극대화할 수 있는 UI/UX를 설계 및 검증하여 서비스의 목적과 용도에 맞게 최적화된 UI를 제공하는 일이다.
데이터베이스관리자	데이터에 대한 요구사항으로부터 데이터베이스를 설계, 구축, 전환하고, 최적의 성능과 품질을 확보하도록 데이터베이스를 수정, 개선, 백업을 수행하는 일이다.
인공지능서비스기획자	인간의 지능으로 할 수 있는 일들을 시스템으로 구현하여 서비스로 제공하기 위한 인공지능 서비스의 목표를 설정하고 고객 요구사항 및 데이터 분석을 통해 인공지능 서비스 모델, 시나리오를 기획하여 실행계획을 수립하는 일이다.
UI/UX 개발자	사용자의 이용 행태와 트렌드, 기술 환경을 분석하고 새로운 사용자 경험(UX) 모델을 제시하여 이를 현실화시킬 수 있는 사용자 리서치, UI 아키텍처 설계, UII 구현 및 테스트, 디지털 컨텐츠 구현, 관련 가이드 제작 등을 수행하는 일이다.
인공지능서비스관리자	구축된 인공지능서비스를 체계적으로 운영하기 위하여 인공지능서비스 운영계획에 따라 품질을 유지하고 서비스를 개선하는 일이다.
보안사고대응전문가	보안사고의 위협정보를 탐지하고, 시스템 복구와 예방 전략을 수립하는 일과 서비스에 영향을 준 증거를 확보 후 분석하여 신속하게 대응하는 일이다.

* 출처: 정보기술산업 인적자원개발위원회, 한국소프트웨어산업협회, “2023 IT분야 역량체계 ITSQF 직무기술서”

05 참고문헌

참고문헌

[1] The Bureau of Justice Assistance (BJA), **Smart Policing Initiative (SPI),** [Online], Available: https://bja.ojp.gov/program/smart-policing-initiative-spi/overview

[2] The FBI, **Law Enforcement Act**, [Online], Available: https://www.govinfo.gov/content/pkg/COMPS-15718/pdf/COMPS-15718.pdf

[3] Central Digital and Data Office and Office for Artificial Intelligence, **A guide to using artificial intelligence in the public sector**, [Online], Available: https://www.gov.uk/government/collections/a-guide-to-using-artificial-intelligence-in-the-public-sector

[4] PIPC, **신뢰 기반 인공지능 데이터 규범, 첫 발 떼다**, [Online], Available: https://www.pipc.go.kr/np/cop/bbs/selectBoardArticle.do?bbsId=BS074&mCode=C020010000&nttId=9083

[5] Christopher Rigano, **Using Artificial Intelligence To Address Criminal Justice Needs**, [Online], Available: www.ojp.gov/pdffiles1/nij/252038.pdf

[6] The Bureau of Justice Assistance (BJA), **FY 2023 Smart Policing Initiative Grant Program**, [Online], Available: https://bja.ojp.gov/funding/webinar/fy23-spi-recording

[7] The Scottish Government's "Emerging Technologies in Policing", **Review of emerging technologies in policing: findings and recommendations**, [Online], Available: https://www.gov.scot/publications/review-emerging-technologies-policing-findings-recommendations/

[8] Afzal, Muhammad, and Panos Panagiotopoulos, "**Smart policing: A critical review of the literature,**" In Electronic Government: 19th IFIP WG 8.5 International Conference, EGOV 2020, Proceedings 19, pp. 59-70, 2020. 09.

[9] Joh, E. E., "**Policing the smart city**," International Journal of Law in Context, vol. 15 no. 2, pp. 177-182, 2019.

[10] Moon, HyungBin, Hyunhong Choi, Jongsu Lee, and Ki Soo Lee, "**Attitudes in Korea toward introducing smart policing technologies: Differences between the general public and police officers,**" Sustainability, vol. 9 no. 10, 2017. 10.

[11] **The Korea Police World Expo 2023**, [Online], Available: https://eng.police-expo.com/about/overview/

[12] 정민훈, **경찰, 오는 8월 '챗GPT' 기반 범죄 데이터 학습 'Poli-ELECTRA' 도입**, [Online], Available: https://www.asiatoday.co.kr/view.php?key=20230719010011230

[13] The Toronto Police Services Board, **Use Of New Artificial Intelligence Technology Policy – Public Consultation**, [Online], Available: https://tpsb.ca/ai

[14] Robertson, Kate, Cynthia Khoo, and Yolanda Song, **To surveil and predict: A human rights analysis of algorithmic policing in Canada,** [Online], Available: https://citizenlab.ca/wp-content/uplo

ads/2020/09/To-Surveil-and-Predict.pdf

[15] Fabio Cozzi et. al., **The EU Artificial Intelligence Act: What's the Impact?,** [Online], Available: https://www.mwe.com/insights/the-eu-artificial-intelligence-act-whats-the-impact/

[16] EU, **MEPs ready to negotiate first-ever rules for safe and transparent AI**, [Online], Available: https://www.europarl.europa.eu/news/en/press-room/20230609IPR96212/meps-ready-to-negotiate-first-ever-rules-for-safe-and-transparent-ai

[17] 홍지은 and 이동희, "**개인정보를 이용한 인공지능 알고리즘과 범죄 수사 – 'AI 수사관'에 대한 「개인정보보호법」 위반 결정을 중심으로 –**," 형사정책, vol. 34 no.2, pp. 43-73, 2022.

[18] 안재경, 우대식, 최이문, "**공간가중 포아송 회귀분석을 활용한 서울시 범죄신고발생의 공간이질성 분석**," 형사정책, vol. 33 no. 2, pp. 211-242, 2021.

[19] 장 현 석, "**외국인 비율이 동단위 범죄수준에 미치는 영향: GIS를 이용한 공간회귀분석**," 2022 치안정책연구, vol. 36 no. 3, 2022.

[20] 이 혜 인 and 김 경 민, "**수도권(서울◦경기◦인천)지역의 범죄발생 패턴: 공간자기상관성의 발견**," Korean Government Security Council, article 53, pp. 17-246, 2013. 12

[21] Feldstein, Steven, "**The global expansion of AI surveillance,**" Carnegie Endowment for International Peace, vol. 17, 2019. 09.

[22] Law Korea, **Police Officer Duties Execution Act**, [Online], Available: https://www.law.go.kr/%EB%B2%95%EB%A0%B9/%EA%B2%BD%EC%B0%B0%EA%B4%80%EC%A7%81%EB%AC%B4%EC%A7%91%ED%96%89%EB%B2%95

[23] White Paper Artificial Intelligence Technology, **Use Cases and Applications, Trustworthiness and Technical Standardization V. 1.1**, [Online], Available: https://portail-qualite.public.lu/dam-assets/publications/normalisation/2021/ilnas-white-paper-artificial-intelligence.pdf

[24] NIST, **AI Risk Management Framework**, [Online], Available: https://www.nist.gov/itl/ai-risk-management-framework

[25] Manasa Reddigari, **The 10 Biggest Security Risks in Today's Smart Home**, [Online], Available: https://www.bobvila.com/slideshow/the-10-biggest-security-risks-in-today-s-smart-home-53081

[26] W. G. Johnson and D. M. Bowman, "**A Survey of Instruments and Institutions Available for the Global Governance of Artificial Intelligence,**" in IEEE Technology and Society Magazine, vol. 40 no. 4, pp. 68-76, 2021. 12. doi: 10.1109/MTS.2021.3123745S.

[27] WEF, **Model Artificial Intelligence Governance Framework and Assessment Guide**, [Online], Available: https://www.weforum.org/projects/model-ai-governance-framework

[28] Singapore**, Artificial Intelligence Governance Framework,** [Online], Available: https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFrame

work2.pdf

[29] PIPC, **Guidelines**, [Online], Available: https://www.pipc.go.kr/np/cop/bbs/selectBoardArticle.do?bbsId=BS217&mCode=D010030000&nttId=8528#LINK

[30] IBM, **AI governance**, [Online], Available: https://www.ibm.com/products/cloud-pak-for-data/ai-governance

[31] Microsoft, **Microsoft Responsible AI Impact Assessment Guide**, [Online], Available: https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4ZzOI

[32] European Commission, "**Report from The Commission to The European Parliament, The Council and The European Economic and Social Committee Report on the safety and liability implications of Artificial Intelligence**," The Internet of Things and robotics, COM/2020/64 final, 2020. https://eur-lex.europa.eu/legal-content/en/TXT/?qid=1593079180383&uri=CELEX%3A52020DC0064

[33] European Commission, Communication from The Commission to The European Parliament, The European Council, The Council, **The European Economic and Social Committee and The Committee of The Regions Artificial Intelligence for Europe COM/2018/237 final**, [Online], Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN

[34] **Real-time gun detection in CCTV: An open problem**, [Online], Available:https://deepknowledge-us.github.io/US-Real-time-gun-detection-in-CCTV-An-open-problem-dataset/

[35] Chitnis, S., N. Deshpande, and A. Shaligram, "**An investigative study for smart home security: Issues, challenges and countermeasures,**" Wireless Sensor Network, vol. 8 no. 4, pp. 61-68, 2016. https://www.scirp.org/pdf/wsn_2016042516164326.pdf

[36] United Nations Office on Drugs and Crime, "**Criminal intelligence: manual for analysts. English, Publishing and Library Section**," United Nations Office, 2011. https://www.unodc.org/documents/organized-crime/Law-Enforcement/Criminal_Intelligence_for_Analysts.pdf

[37] La Vigne, Nancy G., Samantha S. Lowry, Joshua A. Markman, and Allison M. Dwyer, "**Evaluating the use of public surveillance cameras for crime control and prevention,**" Washington, DC: US Department of Justice, Office of Community Oriented Policing Services. Urban Institute, Justice Policy Center, pp. 1-152, 2011.

[38] New York Police Department, "**Domain Awareness System: Impact and Use Policy,**" 2021. 04. https://www.nyc.gov/assets/nypd/downloads/pdf/public_information/post-final/domain-awareness-system-das-nypd-impact-and-use-policy_4.9.21_final.pdf

[39] Khan, Muhammad Asif, Hamid Menouar, and Ridha Hamila, "**DroneNet: Crowd Density Estimation using Self-ONNs for Drones,**" In 2023 IEEE 20th Consumer Communications & Networking Conference (CCNC), pp. 455-460, 2023. https://ieeexplore.ieee.org/abstract/document/10059904

05 참고문헌

[40] AI Hub, **Dataset**, [Online], Available: https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=55

[41] AI Hub, **Dataset**, [Online], Available: https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=489

[42] Github, **RWF2000 – A Large Scale Video Database for Violence Detection**, [Online], Available: https://github.com/mchengny/RWF2000-Video-Database-for-Violence-Detection

[43] Andalusian Research Institute in Data Science and Computational Intelligence, **Weapon dataset**, [Online], Available: https://dasci.es/transferencia/open-data/24705/

[44] The Interactive Emotional Dyadic Motion Capture (IEMOCAP) Database, **IEMOCAP Dataset**, [Online], Available: https://sail.usc.edu/iemocap/

[45] Statistical visual computing laboratory (SVCL) at UCSD**, UCSD Anomaly Detection Dataset**, [Online], Available: http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm

[46] Irizarry, Rafael A, "**Introduction to data science: Data analysis and prediction algorithms with R,**" CRC Press, 2019. Available: http://rafalab.dfci.harvard.edu/dsbook/index.html

[47] Arne Buthmann, **Dealing with Non-normal Data: Strategies and Tools**, [Online], Available:https://www.isixsigma.com/normality/dealing-non-normal-data-strategies-and-tools/

[48] Rebecca Bevans, **Choosing the Right Statistical Test | Types & Examples,** [Online], Available: https://www.scribbr.com/statistics/statistical-tests/

[49] Rebecca Bevans, **Ordinal Data | Definition, Examples, Data Collection & Analysis,** [Online], Available: https://www.scribbr.com/statistics/ordinal-data/

[50] **Top 24 tools for data analysis and how to decide between them**, [Online], Available: https://www.stitchdata.com/resources/data-analysis-tools/

[51] TensorFlow, **Get started validating Tensorflow data**, [Online], Available: https://www.tensorflow.org/tfx/data_validation/get_started?hl=ko

[52] Dahmen, Jessamyn, Diane J. Cook, Xiaobo Wang, and Wang Honglei, "**Smart secure homes: a survey of smart home technologies that sense, assess, and respond to security threats,**" Journal of reliable intelligent environments, vol. 3, pp. 83–98, 2017.

[53] Yin, Xuwang, Soheil Kolouri, and Gustavo K. Rohde, "**Gat: Generative adversarial training for adversarial example detection and robust classification,**" arXiv preprint arXiv:1905.11475, 2019.

[54] Cohen, Jeremy, Elan Rosenfeld, and Zico Kolter, "**Certified adversarial robustness via randomized smoothing,**" In international conference on machine learning, pp. 1310–1320, 2019.

[55] Steinhardt, Jacob, Pang Wei W. Koh, and Percy S. Liang, "**Certified defenses for data poisoning attacks,**" Advances in neural information processing systems, vol. 30, 2017.

05 참고문헌

[56] Papernot, Nicolas, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami, "**Distillation as a defense to adversarial perturbations against deep neural networks,**" In 2016 IEEE symposium on security and privacy (SP), pp. 582-597, 2016.

[57] Zhu, Xiaopei, Zhanhao Hu, Siyuan Huang, Jianmin Li, and Xiaolin Hu, "**Infrared invisible clothing: Hiding from infrared detectors at multiple angles in real world,**" In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13317-13326, 2022.

[58] O. Kosut, L. Jia, R. J. Thomas, and L. Tong, "**Malicious data attacks on smart grid state estimation: Attack strategies and countermeasures,**"in Proc. IEEE SmartGridComm, pp. 220-225, 2010.

[59] Y. Zhuo, Z. Yin and Z. Ge, "**Attack and Defense: Adversarial Security of Data-Driven FDC Systems,**" in IEEE Transactions on Industrial Informatics, vol. 19 no. 1, pp. 5-19, 2023. 01. doi: 10.1109/TII.2022.3197190.

[60] Bukhari, Maryam, Sadaf Yasmin, Saira Gillani, Muazzam Maqsood, Seungmin Rho, and Sang Soo Yeo, "**Secure Gait Recognition-Based Smart Surveillance Systems Against Universal Adversarial Attacks,**" Journal of Database Management (JDM), vol. 34 no. 2, pp. 1-25, 2023.

[61] **South Korea - Data Protection Overview**, [Online], Available:https://www.dataguidance.com/notes/south-korea-data-protection-overview

[62] Randy Rieland, **Artificial Intelligence Is Now Used to Predict Crime. But Is It Biased?**, [Online], Available: https://www.smithsonianmag.com/innovation/artificial-intelligence-is-now-used-predict-crime-is-it-biased-180968337/

[63] **What is Synthetic Data Generation? The Complete Handbook**, [Online], Access: https://www.k2view.com/what-is-synthetic-data-generation

[64] Manuel Pasieka, **A comparison of synthetic data generation methods and synthetic data types,** [Online], Available: https://mostly.ai/blog/comparison-of-synthetic-data-types

[65] Christoph Wehmeyer, **How do you generate synthetic data?**, [Online], Available: https://www.statice.ai/post/how-generate-synthetic-data

[66] Felbermair, Samuel, Florian Lammer, Eva Trausinger-Binder, and Cornelia Hebenstreit, "**Generating synthetic population with activity chains as agent-based model input using statistical raster census data,**" Procedia Computer Science, vol. 170, pp. 273-280, 2020. https://www.sciencedirect.com/science/article/pii/S1877050920304695

[67] Voetman, Roy, Maya Aghaei, and Klaas Dijkstra, "**The Big Data Myth: Using Diffusion Models for Dataset Generation to Train Deep Detection Models,**" arXiv preprint arXiv:2306.09762, 2023.

[68] Zimmering, B.; et al. "**Generating Artificial Sensor Data for the Comparison of Unsupervised Machine Learning Methods,**" Sensors, vol. 21, 2397, 2021. https://doi.org/10.3390/s21072397

[69] Anderson, Jason W., et al., "**Synthetic data generation for the internet of things**," In 2014 IEEE

International Conference on Big Data (Big Data), pp. 171-176, 2014.

[70] Annadurai, C., I. Nelson, K. Nirmala Devi, R. Manikandan, N. Z. Jhanjhi, Mehedi Masud, and Abdullah Sheikh, "**Biometric Authentication-Based Intrusion Detection Using Artificial Intelligence Internet of Things in Smart City,**" Energies, vol. 15 no. 19, pp. 7430, 2022. https://doi.org/10.3390/en15197430

[71] Vijeikis, Romas, Vidas Raudonis, and Gintaras Dervinis, "**Efficient violence detection in surveillance,**" Sensors, vol. 22 no. 6, pp. 2216, 2022.

[72] Microsoft, **Harms Modeling**, [Online], Available: https://learn.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/

[73] Gurtowski, Maciej, and Jan Waszewski, "**Prejudices behind Algorithms: Automated Surveillance Systems as Tools of Segregation and Discrimination,**" Kultura i Edukacja, vol. 4 no. 122, pp. 94-109, 2018.

[74] Aniket, **AI for Crime Prevention and Detection -5 Current Applications**, [Online], Available: https://analyticsjobs.in/blog/artificial-intelligence-crime-used-5-ai-detect-crime/

[75] Deiterch, William; Mendoza, Christina; and Brennan, Tim, **COMPAS risk scales: Demonstrating Accuracy equity and predictive parity. [Report] Northpointe, Inc. Research Department**, [Online], Available: https://www.documentcloud.org/documents/2998391- ProPublicaCommentary-Final07061 6.html

[76] Brackey, Adrienne, "**Analysis of Racial Bias in Northpointe's COMPAS Algorithm**," PhD diss., Tulane University School of Science and Engineering, 2019. https://digitallibrary.tulane.edu/islandora/object/tulane%3A92018

[77] Mallika Chawla, COMPAS Case Study: Investigating Algorithmic Fairness of Predictive Policing, [Online], Available: https://mallika-chawla.medium.com/compas-case-study-investigating-algorithmic-fairness-of-predictive-policing-339fe6e5dd72

[78] IBM, **What is data labeling?**, [Online], Available: https://www.ibm.com/topics/data-labeling

[79] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, "**SMOTE: Synthetic Minority Over-sampling Technique,**" arXiv:1106.1813v1, 2011.

[80] H. Han, W. Y. Wang, B. H. Mao, "**Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning,**" ICIC 2005, Part I, LNCS 3644, pp. 878 - 887, 2005.

[81] H. He, Y. Bai, E. A. Garcia, S. Li, "**ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning,**" 2008 IEEE International Joint Conference on Neural Networks, 2008.

[82] Satyam Kumar, **7 Over Sampling techniques to handle Imbalanced Data**, [Online], Available: https://towardsdatascience.com/7-over-sampling-techniques-to-handle-imbalanced-data-ec51c8db349f

05 참고문헌

[83] Douzas, Georgios, Fernando Bacao, and Felix Last, "**Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE,**" Information Sciences, vol. 465, pp. 1-20, 2018.

[84] Kang Shik Shin, "**Evaluation Of Open Source Vulnerability Scanning Tools**," KAIST CSRC Weblog, 2022. https://csrc.kaist.ac.kr/blog/2022/03/11/evaluation-of-open-source-vulnerability-scanning-tools/

[85] Adam Murray, **7 Factors Developers Should Consider Before Choosing an Open Source Project**, [Online], Available: https://www.mend.io/resources/blog/7-factors-developers-should-consider-before-choosing-an-open-source-project/

[86] The World Economic Forum, **Model AI Governance Framework**, [Online], Available: https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf

[87] Open Source Initiative, **The Open Source Definition**, [Online], Available: https://opensource.org/osd

[88] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, **A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear**, [Online], Available: https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/

[89] Fu, Siyao, Haibo He, and Zeng-Guang Hou, "**Learning race from face: A survey,**" IEEE transactions on pattern analysis and machine intelligence, vol. 36 no. 12, pp. 2483-2509, 2014.

[90] M. Wang, **The Robots are Watching Us**, [Online], Available: https://www.hrw.org/news/2020/04/06/robots-are-watching-us#

[91] Hope Reese, **What Happens When Police Use AI to Predict and Prevent Crime?**, [Online], Available: https://daily.jstor.org/what-happens-when-police-use-ai-to-predict-and-prevent-crime/

[91] Nick Cumming-Bruce, **U.N. Panel: Technology in Policing Can Reinforce Racial Bias**, [Online], Available: https://www.nytimes.com/2020/11/26/us/un-panel-technology-in-policing-can-reinforce-racial-bias.html

[93] **Artificial Intelligence Is Now Used to Predict Crime. But Is It Biased?**, [Online], Available: https://www.smithsonianmag.com/innovation/artificial-intelligence-is-now-used-predict-crime-is-it-biased-180968337/

[94] Zhang, Hongjing, and Ian Davidson, "**Towards fair deep anomaly detection,**" In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pp. 138-148, 2021. https://arxiv.org/abs/2012.14961

[95] Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan, "**Inherent trade-offs in the fair det**

ermination of risk scores," arXiv preprint arXiv:1609.05807, 2016. https://arxiv.org/pdf/1609.05807

[96] 김병필, "**Assessing Fairness in Artificial Intelligence: Contextual Justification for Normative Criteria**," Doctor of Law dissertation, Graduate School of Seoul National University

Department of Law, 2023. 08.

[99] Li, Dan, Dacheng Chen, Jonathan Goh, and See-kiong Ng, "**Anomaly detection with generative adversarial networks for multivariate time series**," arXiv preprint arXiv:1809.04758, 2018. doi: 10.48550/arxiv.1809.04758

[100] Zheng, H., Ye, Q., Hu, H., Fang, C., Shi, J., "**BDPL: A Boundary Differentially Private Layer Against Machine Learning Model Extraction Attacks,**" In: Sako, K., Schneider, S., Ryan, P. (eds) Computer Security -ESORICS 2019. ESORICS 2019. Lecture Notes in Computer Science, vol 11735, 2019. https://doi.org/10.1007/978-3-030-29959-0_4

[101] Pautov, Mikhail, Grigorii Melnikov, Edgar Kaziakhmedov, Klim Kireev, and Aleksandr Petiushko, "**On adversarial patches: real-world attack on arcface-100 face recognition system,**" In 2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON), pp. 0391-0396, 2019. https://ieeexplore.ieee.org/abstract/document/8958134

[102] Dongyu Meng, Hao Chen, "**MagNet: a Two-Pronged Defense against Adversarial Examples,**" arXiv:1705.09064, 2017.

[103] Shen, Shiwei, Guoqing Jin, Ke Gao, and Yongdong Zhang, "**Ape-gan: Adversarial perturbation elimination with gan,**" arXiv preprint arXiv:1707.05474, 2017.

[104] Javed, Abdul Rehman, et. al., "**A Survey of Explainable Artificial Intelligence for Smart Cities**," Electronics, vol. 12 no. 4, pp. 1020, 2023. https://doi.org/10.3390/electronics12041020.

[105] Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra,"**Grad-cam: Visual explanations from deep networks via gradient-based localization**," In Proceedings of the IEEE international conference on computer vision, pp. 618-626. 2017.

[106] Assaf, Roy, and Anika Schumann, "**Explainable deep neural networks for multivariate time series predictions,**" In IJCAI, pp. 6488-6490, 2019.

[107] Talal Shaikh, Aaishwarya Khalane, Rikesh Makwana, et al. "**Evaluating Significant Features in Context-Aware Multimodal Emotion Recognition with XAI Methods,**" Authorea, 2023. 01.

[108] L. Arrotta, G. Civitarese, M. Fiori and C. Bettini, "**Explaining Human Activities Instances Using Deep Learning Classifiers,**" 2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA), pp. 1-10, 2022.doi: 10.1109/DSAA54385.2022.10032345.

[109] Adak, Anirban, Biswajeet Pradhan, Nagesh Shukla, and Abdullah Alamri, "**Unboxing Deep Lear**

ning Model of Food Delivery Service Reviews Using Explainable Artificial Intelligence (XAI) Technique," Foods, vol. 11 no. 14, 2019. https://doi.org/10.3390/foods11142019

[111] Molnar, C., **Interpretable Machine Learning (Second Edition): A guide for making black box models explainable,** [Online], Available: https://christophm.github.io/interpretable-ml-book/shap.html#fnref44

[112] Kleyton da Costa, "**SHAP Values: An Intersection Between Game Theory and Artificial Intelligence,**" [Online], Available: https://www.holisticai.com/blog/shap-values-game-theory-and-ai

[113] Jessica Newman, **Explainability won't save AI.", Brookings**, [Online], Available: https://www.brookings.edu/articles/explainability-wont-save-ai/.

[114] Zhao, Xuejun, Wencan Zhang, Xiaokui Xiao, and Brian Lim, "**Exploiting explanations for model inversion attacks,**" In Proceedings of the IEEE/CVF international conference on computer vision, pp. 682-692. 2021.

[115] **European Parliament Agrees on Position on the AI Act,** [Online],Available: https://www.huntonprivacyblog.com/2023/06/15/european-parliament-agrees-on-position-on-the-ai-act/.

[116] Piorkowski, David, John Richards, and Michael Hind, "**Evaluating a methodology for increasing AI transparency: A case study,**" arXiv preprint arXiv:2201.13224, 2022.

[117] A. Arnab, C. Sun, A. Nagrani, and C. Schmid, "**Uncertainty-aware weakly supervised action detection from untrimmed videos,**" in Proc. Eur. Conf. Comput. Vis. Cham, pp. 751-768, 2020. https://arxiv.org/pdf/2007.10703.pdf

[118] Hong, X., Ma, W., Huang, Y., Miller, P., Liu, W., and Zhou, H., "**Evidence reasoning for event inference in smart transport video surveillance,**" In ICDSC '14 Proceedings of the International Conference on Distributed Smart Cameras, 2014. https://doi.org/10.1145/2659021.2659040

[119] Schwartz, R., et. al., "**Towards a standard for identifying and managing bias in artificial intelligence,**" NIST Special Publication, vol. 1270, pp. 1-77, 2022.

[120] Isaiah McCall, **Teaneck just banned facial recognition technology for police. Here's why**, [Online], Available: https://www.northjersey.com/story/news/bergen/teaneck/2021/02/25/teaneck-nj-bans-facial-recognition-usage-police-citing-bias/6802839002/

[121] Tian, Ying-li, et. al. "**IBM smart surveillance system (S3): event based video surveillance system with an open and extensible framework,**" Machine Vision and Applications, vol. 19, pp. 315-327, 2008.

[122] National Institute of Standards and Technology (NIST), **Guidelines for Developing an Incident Response Plan**, [Online], Available: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-61r2.pdf

[123] ISO, **Information technology — Information security incident management**, [Online], Availabl

e: https://www.iso.org/standard/78974.html

[124] Diptiben Ghelni, "**Deep Learning and Artificial Intelligence Framework to Improve the Cyber Security,**" American Journal of Artificial Intelligence. Vol. x No. x, pp. x-x, 2022. https://d197for5662m48.cloudfront.net/documents/publicationstatus/90291/preprint_pdf/c12f4b6dfcb0ece3a42a357ad2203fac.pdf

[125] Akash Takyar, **AI Model Security: Concerns, Best Practices and Techniques**, [Online], Available: https://www.leewayhertz.com/ai-model-security/

[126] Veritis, **Anomaly Detection with MI & AI: An Introduction**, [Online], Available: https://www.veritis.com/blog/anomaly-detection-using-machine-learning/

[127] Chin K., **What is the Primary Method for Protecting Sensitive Data?**, [Online], Available: https://www.upguard.com/blog/protecting-sensitive-data

[128] van Bekkum, Marvin, and Frederik Zuiderveen Borgesius, "**Using sensitive data to prevent discrimination by artificial intelligence: Does the GDPR need a new exception?,**" Computer Law & Security Review, vol. 48, pp. 105770, 2023.

[129] The US Government Accountability Office (GAO), **An Accountability Framework for Federal Agencies and Other Entities**, [Online], Available: https://www.gao.gov/assets/gao-21-519sp.pdf

[130] Hand, David J., and Shakeel Khan, "**Validating and verifying AI systems,**" Patterns, vol. 1 no. 3, pp. 100037, 2020.

[131] the European Centre for Disease Prevention and Control (ECDC), **Data Quality Monitoring and Surveillance System Evaluation—A Handbook of Methods and Applications**, [Online], Available: https://www.ecdc.europa.eu/sites/default/files/media/en/publications/Publications/Data-quality-monitoring-surveillance-system-evaluation-Sept-2014.pdf

[132] Anton Maslov, "**Measuring the Performance of the Police: The Perspective of the Public,**"Public Safety Canada, [Online], Available: https://www.publicsafety.gc.ca/cnt/rsrcs/pblctns/2015-r034/index-en.aspx#perfmeas

[133] Nicol Turner Lee, Caitlin Chin, "**Police surveillance and facial recognition: Why data privacy is imperative for communities of color,**" Brookings, 2022, https://www.brookings.edu/articles/police-surveillance-and-facial-recognition-why-data-privacy-is-an-imperative-for-communities-of-color/

[134] Brian Ellis, "**3 Impacts of The Internet of Things (IoT) On Policing,**"California Peace Officers Association, 2017. https://cpoa.org/3-impacts-internet-things-iot-policing/

[135] Kevin Strom, "**Research on the Impact of Technology on Policing Strategy in the 21st Century, Final Report,**" National Criminal Justice Reference Service, 2017. https://www.ojp.gov/pdffiles1/nij/grants/251140.pdf

05 참고문헌

[136] **Home Security Glossary**, [Online], Available: https://www.homewatchgroup.com/home-security-glossary/

[137] Reisinger, Michaela R., Sebastian Prost, Johann Schrammel, and Peter F.,"**User requirements for the design of smart homes: dimensions and goals,**" Journal of Ambient Intelligence and Humanized Computing, pp. 1-20, 2022. https://link.springer.com/article/10.1007/s12652-021-03651-6

[138] Vivek Dev, "**Alarm Systems - Some key design considerations**,"Engineering Review, 2014. https://engmag.in/alarm-systems-some-key-design-considerations/

[139] **Notification UX: How to Design Notifications for A Better User Experience**, [Online], Available: https://userpilot.com/blog/notification-ux/

[140] Nick Babich, "**5 Types of UI Notifications And essential rules on when and how to use each type**,"UX Planet, 2022. https://uxplanet.org/5-types-of-ui-notifications-dbfbda284456

[141] Michael Segner, **Top 5 Open Source Data Lineage Tools (With User Reviews)**, [Online], Available: https://www.montecarlodata.com/blog-open-source-data-lineage-tools/

[142] **OpenMetadata**, [Online], Available: https://open-metadata.org/

[143] DVC, **Open-source Version Control System for Machine Learning Projects**, [Online], Available: https://dvc.org/

[144] Jakub Czakon, **Best 7 Data Version Control Tools That Improve Your Workflow With Machine Learning Projects**, [Online], Available: https://neptune.ai/blog/best-data-version-control-tools

[145] **Pachyderm**, [Online], Available: https://www.pachyderm.com/

[146] **9 Best AI Security Tools of 2023**, [Online], Available: https://renaissancerachel.com/best-ai-security-tools/

[147] Google, **Google Home Documentation**, [Online], Available:https://developers.home.google.com/docs

[148] Vesternet, **Ultimate Guide to Smart Home Security Systems**, [Online], Available:https://www.vesternet.com/en-global/pages/scenarios-smart-home-security-guide

[149] Yonhap News, 경찰, 오송 지하차도 참사 부실대응 논란에 항변…논란만 키워, [Online], Available:]https://www.yonhapnewstv.co.kr/news/MYH20230724000900641?input=1825m

[150] Osoba, Osonde A., William Welser IV, “An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence,” Santa Monica, CA: RAND Corporation, 2017. https://www.rand.org/content/dam/rand/pubs/research_reports/RR1700/RR1744/RAND_RR1744.pdf

[151] European Parliamentary Research Service, “The ethics of artificial intelligence: Issues and initiatives,” Scientific Foresight Unit (STOA) PE 634.452, 2020. 03. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf

[152] Koh, Pang Wei, Jacob Steinhardt, and Percy Liang., "Stronger data poisoning attacks break data sanitization defenses," Machine Learning, vol. 111, pp.1–47, 2022.

[153] Goldblum, Micah, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Mądry, Bo Li, and Tom Goldstein. "Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.45, no. 2, pp.1563–1580, 2022.

2024

신뢰할 수 있는 인공지능 개발 안내서 스마트 치안 분야

한국정보통신기술협회 신 준 호 단장
곽 준 호 팀장
김 송 이 책임
채 희 문 책임
조 경 우 책임
황 재 영 책임
신 예 진 책임
변 은 영 선임
오 상 훈 선임
강 상 연 전임

인 쇄 2024년 2월
발 행 2024년 2월
발 행 처 한국정보통신기술협회
발 행 인 손 승 현
편집·제작 ㈜디자인여백플러스
I S B N 979-11-89545-63-5