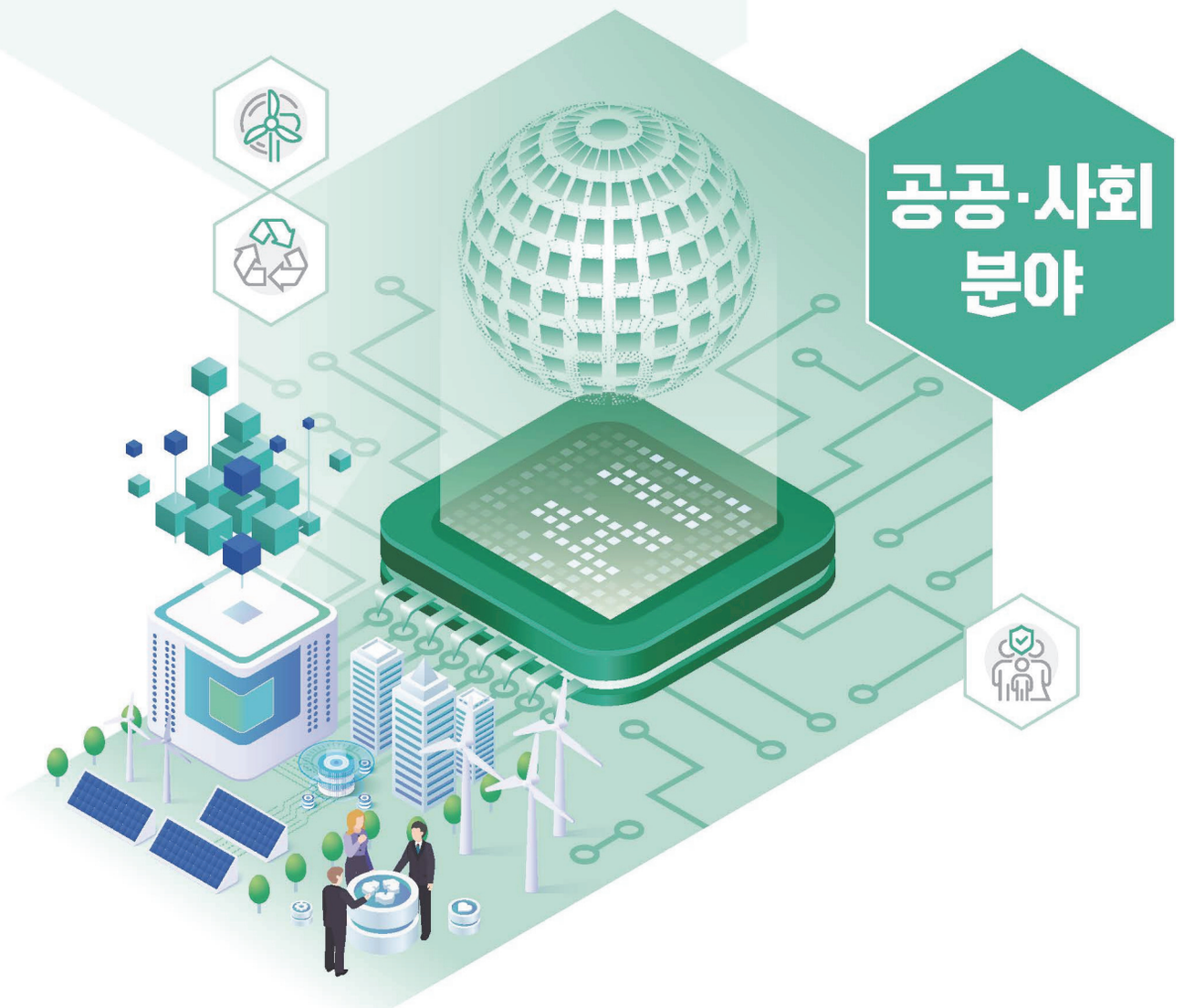
2023

신뢰할 수 있는 인공지능 개발 안내서

공공·사회 분야



과학기술정보통신부
Ministry of Science and ICT

TTA 한국정보통신기술협회
Telecommunications Technology Association

Jinhan M&B

일러두기

- 본 안내서는 과학기술정보통신부 「AI 신뢰성 검증체계 고도화」 사업의 연구 결과로서 내용의 무단 전재를 금합니다.

- 아울러, 안내서의 내용을 가공·인용하는 경우에는 반드시 '과학기술정보통신부·한국정보통신기술협회 《2023 신뢰할 수 있는 인공지능 개발 안내서 – 공공·사회 분야》'의 출처를 밝혀 주시기 바랍니다.

- 본 안내서는 공공·사회분야 인공지능 서비스 및 제품을 개발하는 과정에서 참고 자료로 활용할 수 있도록 편찬되었습니다. 본 안내서는 기업의 업무 환경과 상황, 개발 목적 등을 고려하여 필요하신 내용을 취사선택하여 활용하시기 바랍니다.

- 본 안내서의 공공·사회 분야 인공지능 동향 및 기술 정보는 2023년 2월 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(2023년 하반기 공개 예정)에도 콘텐츠가 공개되어 있으므로 참고하시면 더 편리하게 활용하실 수 있습니다.

- 공공·사회 외 분야는 「2023 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 계획 및 설계	**요구사항 01** 인공지능 시스템에 대한 위험관리 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?	☐	☐	☐
	01-2 위험 요소를 제거 및 방지하거나 영향을 완화하기 위한 방안을 마련하였는가?	☐	☐	☐
	01-2a 위험 요소 제거 방안을 도출하고 파급효과가 감소하였는지 확인하였는가?	☐	☐	☐
	01-2b 위험관리 관련 규정에 따라 이행하였음을 입증/관리 할 수 있는 방안을 마련하였는가?	☐	☐	☐
	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 충분히 훈련된 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 이용 빈도가 낮은 타 시스템의 개선 및 통·폐합을 통해 구현 가능한지 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?	☐	☐	☐
2 데이터 수집 및 처리	**요구사항 04** 데이터의 활용을 위한 상세 정보 제공			
	04-1 데이터의 명확한 이해와 활용을 지원하는 상세한 정보를 제공하는가?	☐	☐	☐
	04-1a 정제 전과 후의 데이터 특성을 설명하였는가?	☐	☐	☐
	04-1b 학습 데이터와 메타데이터를 구분하고 각 명세자료를 확보하였는가?	☐	☐	☐
	04-1c 보호변수의 선정 이유 및 반영 여부를 설명하였는가?	☐	☐	☐
	04-1d 라벨링 작업자를 위해 교육을 시행하고 작업 가이드 문서를 마련하였는가?	☐	☐	☐

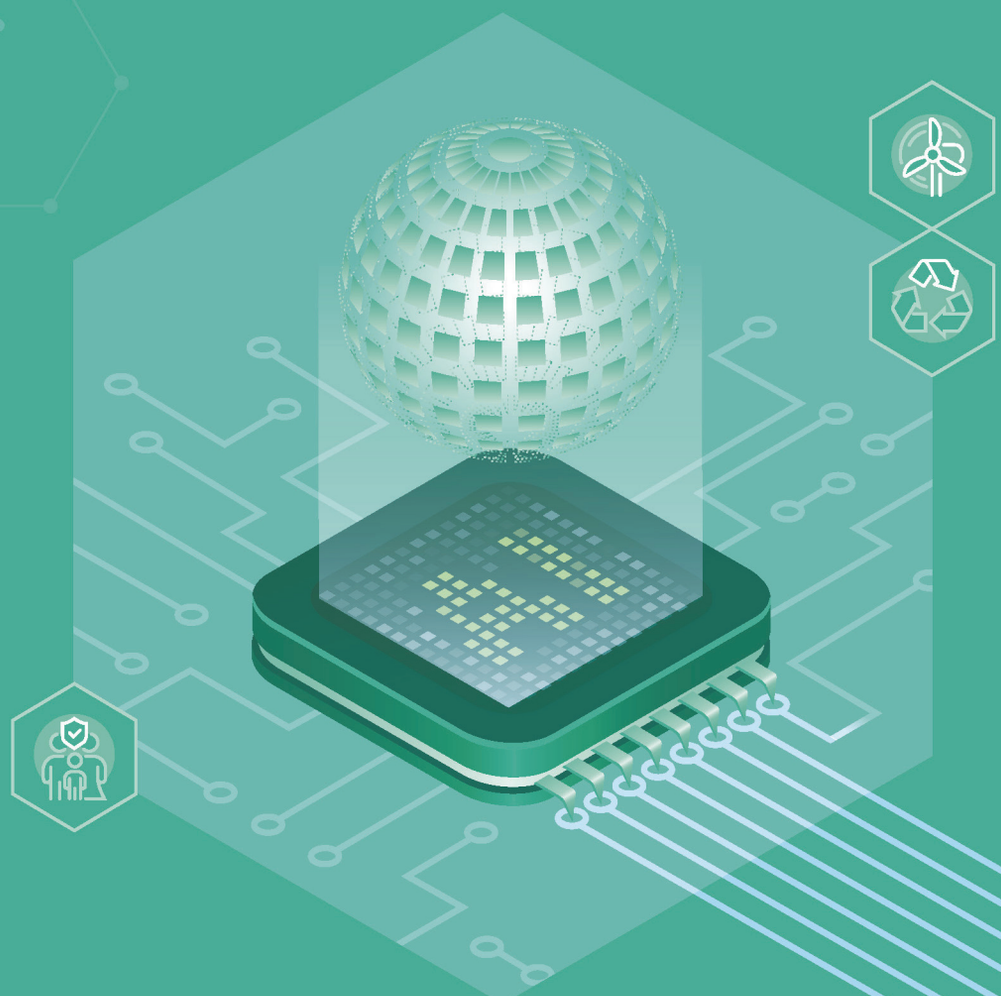
생명주기	요구사항 및 체크리스트	Yes	No	N/A
2 데이터 수집 및 처리	04-2 데이터의 출처는 기록 및 관리되고 있는가?	☐	☐	☐
	04-2a 신뢰할 수 있는 출처로부터 제공되는 데이터셋을 사용하였는가?	☐	☐	☐
	04-2b 오픈소스 데이터셋을 활용하는 경우, 출처를 명시하였는가?	☐	☐	☐
	요구사항 05 데이터 강건성 확보를 위한 이상abnormal 데이터 점검			
	05-1 이상 데이터의 식별 및 정상 여부를 점검하였는가?	☐	☐	☐
	05-1a 전체 학습용 데이터 분포를 시각화하여 발생 가능한 오류들을 확인하였는가?	☐	☐	☐
	05-1b 학습 데이터 이상값 식별 기법을 적용하였는가?	☐	☐	☐
	05-2 데이터 공격에 대한 방어 수단을 강구하였는가?	☐	☐	☐
	05-2a 데이터 중독poisoning, 회피evasion 등 공격에 대한 방어 대책을 마련하였는가?	☐	☐	☐
	요구사항 06 수집 및 가공된 학습 데이터의 편향 제거			
	06-1 데이터 수집 시, 인적·물리적 요인으로 인한 편향 완화 방안을 마련하였는가?	☐	☐	☐
	06-1a 인적 편향을 제거하기 위한 절차적, 기술적 수단을 적용하였는가?	☐	☐	☐
	06-1b 데이터의 다양성 확보를 위해 이기종 수집 장치를 활용하였는가?	☐	☐	☐
	06-1c 하드웨어로 인해 발생할 수 있는 데이터의 편향을 점검하였는가?	☐	☐	☐
	06-2 학습에 사용되는 특성을 분석하고 선정 기준을 마련하였는가?	☐	☐	☐
	06-2a 보호변수 선정 시 충분한 분석을 수행하였는가?	☐	☐	☐
	06-2b 편향을 발생시킬 수 있는 특성의 영향력을 완화하였는가?	☐	☐	☐
	06-2c 데이터 전처리 시 특성이 과도하게 제거되었는지 검토하였는가?	☐	☐	☐
	06-3 데이터 라벨링 시, 발생 가능한 편향을 확인하고 방지하였는가?	☐	☐	☐
	06-3a 데이터 라벨링 기준을 명확히 수립하고 작업자에게 제공하였는가?	☐	☐	☐
	06-3b 다양한 데이터 라벨링 작업자를 섭외하기 위해 노력하였는가?	☐	☐	☐
	06-3c 다양한 데이터 라벨링 검수자를 확보하기 위해 노력하였는가?	☐	☐	☐
	06-4 데이터의 편향 방지를 위한 샘플링을 수행하였는가?	☐	☐	☐
	06-4a 편향 방지를 위한 샘플링 기법을 적용하였는가?	☐	☐	☐
3 인공지능 모델 개발	**요구사항 07 오픈소스 라이브러리의 보안성 및 호환성 확보**			
	07-1 오픈소스 라이브러리의 안정성을 확인하였는가?	☐	☐	☐
	07-1a 활성화된 오픈소스 라이브러리를 사용하였는가?	☐	☐	☐
	07-2 오픈소스 라이브러리의 위험 요소는 관리되고 있는가?	☐	☐	☐
	07-2a 사용 중인 오픈소스 라이브러리의 라이선스 준수사항을 이행하였는가?	☐	☐	☐
	07-2b 사용 중인 오픈소스 라이브러리의 호환성 및 보안취약점을 확인하였는가?	☐	☐	☐

안내서 활용을 위한 체크리스트

생명주기	요구사항 및 체크리스트	Yes	No	N/A
3 인공지능 모델 개발	**요구사항 08** 인공지능 모델의 편향 제거			
	08-1 모델 편향을 제거하는 기법을 적용하였는가?	☐	☐	☐
	08-1a 개발하려는 모델에 맞게 편향제거 기법을 선택하였는가?	☐	☐	☐
	08-1b 편향성 평가 및 모니터링을 위한 정량적 지표를 선정하고 관리하는가?	☐	☐	☐
	요구사항 09 인공지능 모델 공격에 대한 방어 대책 수립			
	09-1 모델 추출 공격(model extraction attack)에 대한 방어 방안을 수립하였는가?	☐	☐	☐
	09-1a 모델 추출 공격에 대비하는 방어 기법을 적용하였는가?	☐	☐	☐
	09-2 모델 회피 공격(model evasion attack)에 대한 방어 방안을 수립하였는가?	☐	☐	☐
	09-2a 모델 회피 공격에 대비하는 방어 기법을 적용하였는가?	☐	☐	☐
	요구사항 10 인공지능 모델 명세 및 추론 결과에 대한 설명 제공			
	10-1 사용자가 모델 추론 결과의 도출 과정을 수용할 수 있도록 근거를 제공하는가?	☐	☐	☐
	10-1a XAI 기술 적용이 가능한 경우, 인공지능 모델의 추론 결과를 설명하기 위한 기법 적용에 대해 검토하였는가?	☐	☐	☐
	10-1b XAI 기술 적용이 불가능한 경우, 기법 적용 이외의 대안을 마련하였는가?	☐	☐	☐
	10-2 인공지능 모델 상세 문서를 통해 모델의 명세를 투명하게 제공하는가?	☐	☐	☐
	10-2a 시스템 개발 과정과 모델 작동 방식에 대한 세부 정보가 설명된 문서를 작성하였는가?	☐	☐	☐
	10-3 필요 시, 인공지능 모델 추론 결과에 대한 설명을 제공하는가?	☐	☐	☐
	10-3a 모델 추론 결과에 대한 설명이 필요한지 검토하였는가?	☐	☐	☐
	10-3b 사용자에게 인공지능 모델 추론 결과에 대한 설명을 제공하였는가?	☐	☐	☐
4 시스템 구현	**요구사항 11** 인공지능 시스템 구현 시 발생 가능한 편향 제거			
	11-1 소스 코드 및 사용자 인터페이스로 인한 편향을 제거하기 위해 노력하였는가?	☐	☐	☐
	11-1a 데이터 접근 방식 구현과정 등 소스 코드에서의 편향 발생 가능성을 확인하였는가?	☐	☐	☐
	11-1b 사용자 인터페이스 및 상호작용 방식으로 인한 편향을 확인하였는가?	☐	☐	☐
	요구사항 12 인공지능 시스템의 안전 모드 구현 및 문제발생 알림 절차 수립			
	12-1 공격, 성능 저하 및 사회적 이슈 등의 문제 발생 시 대응 가능한 안전 모드를 적용하는가?	☐	☐	☐
	12-1a 문제 상황에 대한 예외 처리 정책이 마련되어 있는가?	☐	☐	☐
	12-1b 인공지능 시스템의 보안 강화를 위한 보안 기법을 적용하였는가?	☐	☐	☐
	12-1c 인공지능 시스템의 의사결정으로 인한 파급효과가 크고 불확실성이 높은 경우, 사람의 개입을 고려하였는가?	☐	☐	☐
	12-1d 예상되는 사용자 오류에 대한 안내 및 대응을 제공하는가?	☐	☐	☐

생명주기	요구사항 및 체크리스트		Yes	No	N/A
4 시스템 구현	12-2	인공지능 시스템에서 문제가 발생할 경우, 시스템은 이를 운영자에게 전달하는 기능을 수행하는가?	☐	☐	☐
	12-2a	편견, 차별 등 윤리적 문제에 대한 알림 절차를 수립하였는가?	☐	☐	☐
	12-2b	시스템 성능 저하를 평가하기 위한 지표 및 절차를 설정하고 알림 절차를 수립하였는가?	☐	☐	☐
	요구사항 13	**인공지능 시스템의 설명에 대한 사용자의 이해도 제고**			
	13-1	인공지능 시스템 사용자의 특성user characteristics과 제약사항을 분석하였는가?	☐	☐	☐
	13-1a	사용자 특성에 따른 세부 고려사항을 분석하였는가?	☐	☐	☐
	13-2	사용자 특성에 따른 충분한 설명을 제공하는가?	☐	☐	☐
	13-2a	사용자 특성에 따른 설명 평가의 기준을 수립하였는가?	☐	☐	☐
	13-2b	사용자가 이해하기 어려운 전문 용어 사용을 지양하였는가?	☐	☐	☐
	13-2c	사용자의 구체적인 행동과 이해를 이끌어낼 수 있도록 명확한 표현을 사용하였는가?	☐	☐	☐
	13-2d	설명이 필요한 위치와 타이밍은 적절한가?	☐	☐	☐
	13-2e	사용자 경험을 평가할 수 있는 다양한 사용자 조사 기법을 활용하였는가?	☐	☐	☐
5 운영 및 모니터링	요구사항 14	**인공지능 시스템의 추적가능성 및 변경이력 확보**			
	14-1	인공지능 시스템의 의사결정에 대한 추적 방안을 수립하였는가?	☐	☐	☐
	14-1a	인공지능 시스템의 의사결정에 대한 기여도 추적 방안은 확보하였는가?	☐	☐	☐
	14-1b	인공지능 시스템의 의사결정 추적을 위한 로그 수집 기능을 구현하였는가?	☐	☐	☐
	14-1c	지속적인 사용자 경험 모니터링을 위해 사용자 로그를 수집 및 관리하고 있는가?	☐	☐	☐
	14-2	학습 데이터의 변경 이력을 확보하고, 데이터 변경이 미치는 영향을 관리하였는가?	☐	☐	☐
	14-2a	데이터 흐름 및 계보lineage를 추적하기 위한 조치를 마련하였는가?	☐	☐	☐
	14-2b	데이터 소스 변경에 대한 모니터링 방안을 확보하였는가?	☐	☐	☐
	14-2c	데이터 변경 시, 버전관리를 수행하였는가?	☐	☐	☐
	14-2d	데이터 변경 시, 이해관계자를 위한 정보를 제공하는가?	☐	☐	☐
	14-2e	신규 데이터 확보 시, 인공지능 모델의 성능평가를 재수행하였는가?	☐	☐	☐
	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	상호작용의 대상을 명확히 설명하는가?	☐	☐	☐
	15-2a	사용자가 인공지능과 상호작용하고 있다는 사실을 사용자에게 명확히 설명하였는가?	☐	☐	☐

2023 신뢰할 수 있는 인공지능 개발 안내서 | 공공·사회 분야



PART 1

개요

01 안내서 발간 배경 및 목적

01 안내서 발간 배경 및 목적

인공지능 기술은 인공지능 스피커, 인터넷의 추천 알고리즘과 같이 일상생활과 관련된 서비스부터 자율주행자동차, 법률, 의료, 금융 등 전문 분야에 이르기까지 다양하게 활용되고 있다. 인공지능의 활용 범위가 광범위해짐에 따라 공공기관에서도 업무 과정에서 발생하는 방대한 공공 데이터로부터 새로운 가치를 창출하여 국민의 다양한 요구에 응대하고 질 높은 공공서비스를 제공하기 위해 인공지능 기술을 도입하고자 하는 수요가 꾸준히 증가하고 있다. 이와 더불어 완전히 자동화된 행정처분과 전자정부서비스를 규정한 「행정기본법」(2021. 3.)과 「전자정부법」(2022. 1.)의 시행으로 공공 부분의 의사결정에서 인공지능의 관여와 역할이 커질 것으로 예상된다.

이처럼 공공기관에서 공공서비스 및 업무 개선을 위해 인공지능 도입을 추진하는 가운데 데이터 및 알고리즘 편향으로 인한 차별, 사생활 침해 등 인간의 기본권 침해가 증가할 우려가 있다. 이에 유럽연합Europe Union, EU에서는 《인공지능 기반 서비스 및 솔루션 공공조달의 데이터 윤리 백서white paper on data ethics in public procurement of AI-based services and solutions(2020. 5.)》를 통해 공공서비스에 활용되는 인공지능 기술의 필수 검토 요소로서 합법성, 윤리성, 사회적 안전성을 제시한 바 있다. 그리고 영국에서는 공공 분야에서의 AI 활용과 확대를 위해 〈공공부문 인공지능 활용 가이드a guide to using artificial intelligence in the public sector(2019. 6.)〉를 발간하였다.

국내에서는 2017년에 처음으로 〈지능정보사회 윤리 가이드라인〉을 발간하였고, 2019년에는 한국방송통신위원회에서 사람 중심 서비스, 투명성과 설명가능성, 책임성, 안전성, 차별 금지, 참여, 프라이버시와 데이터 거버넌스 등으로 구성된 'AI 윤리 7원칙'을 발표하였다. 그리고 같은 해에 공공기관의 인공지능 도입 시 발생할 수 있는 부정적 위험을 최소화하기 위해 신뢰 가능한 인공지능을 구현하고자 OECDOrganization for Economic Cooperation and Development, EU 등 인공지능 신뢰성 확보에 관한 국제 표준 요구사항을 반영한 〈공공기관 신뢰 가능 AI 구현 실용 가이드(2019. 5.)〉를 발간하였다.

그런데 지금까지 발간된 국내외 공공서비스를 위한 인공지능 신뢰성 가이드는 주로 윤리적 관점에서 추상적인 항목을 제시하고 있어 실무 현장에 활용하는 데는 한계가 있다. 우선 인공지능 서비스를 개발하는 공급자, 특히 공공기관에 납품할 목적으로 인공지능 서비스를 개발하는 중소기업에서는 자체적으로 인력을 확보하고 연구개발을 수행해 신뢰성 확보 요구사항을 도출하여 검증 체계를 마련하는 것이 쉬운 일이 아니다. 반대로 인공지능 서비스를 납품받는 수요자인 공공기관에서는 인공지능의 신뢰성 확보를 하나의 요구사항으로 제시할 수는 있으나 구체적 방안이 없어 제대로 이행되었는지를 판단하기가 힘든 실정이다.

따라서 본 안내서는 인공지능 기술의 공공서비스 도입 시에 상술한 현실적인 문제점을 해결하고자 작성되었다. 실용적인 요구사항 및 검증항목 도출을 위해 유럽, 미국, 영국 등 주요 선진국과 국제기구들에서 발표한 권고안 및 가이드를 참고하여 자율적으로 점검 가능한 요구사항 15개와 검증항목 68개를 제시하고 있다. 또한 해당 요구사항 및 검증항목에 대한 타당성 검토를 위해 실제 공공기관에 인공지능 서비스를 납품하는 업체에서 현장 적용을 수행하고, 본 개발 안내서의 타당성을 확보하기 위해 산학연 전문가의 의견을 반영하였다.

개발자 및 기획자, 곧 인공지능 서비스 개발 실무자는 본 개발 안내서에 제시된 항목을 참고하여 공공서비스에 활용되는 인공지능 서비스에 대한 최소한의 신뢰성을 확보할 수 있고, 공공기관에서 인공지능 서비스 운영을 담당하는 실무자는 신뢰성을 확보하려면 무엇을 중요하게 고려해야 하는지를 이해할 수 있을 것이다. 본 개발 안내서가, 우리나라 공공기관에 인공지능 서비스를 제공하고 운영하는 기업 및 기관이 좀 더 성숙한 인공지능 기술을 확보하여 공공·사회 인공지능 서비스 분야에서 글로벌 경쟁력을 갖추는 데 기초 자료로 활용될 수 있기를 희망한다.

02 공공·사회 인공지능 신뢰성 동향

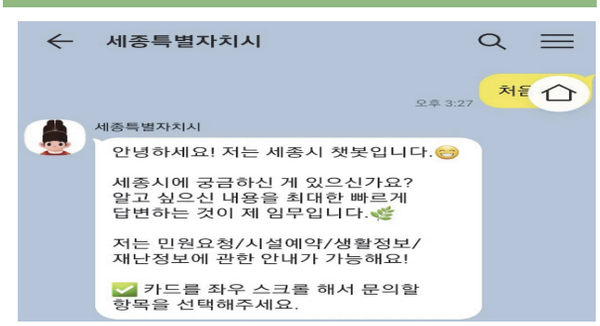
현대 사회가 복잡해지고 다양해지면서 정책이 다루어야 할 문제의 영역도 넓어지고 복합적인 양상을 띠고 있어 대량의 데이터를 활용한 최적의 공공 서비스를 제공할 수 있는 인공지능의 필요성은 더욱 증대될 것으로 보인다. 그런데 공공 분야에서 인공지능의 활용에 따른 잠재적인 부작용이나 위험성이 대두되고 있어 이에 대비하기 위한 인공지능의 신뢰성을 확보하는 것이 더욱더 강조되고 있다.

현재 세계의 주요 국가와 표준 관련 기구, 기술 단체에서는 공공 분야의 인공지능 활용에 대한 신뢰성을 확보하기 위해 각자의 상황에 맞는 방안을 제시하고 있다. 본 절에서는 공공 분야에서 인공지능이 활용되는 영역과 인공지능을 활용함으로써 발생할 수 있는 문제점을 파악하고, 국내외에서 진행 중인 공공 분야의 인공지능 신뢰성 관련 정책 및 연구 동향을 살펴보고자 한다.

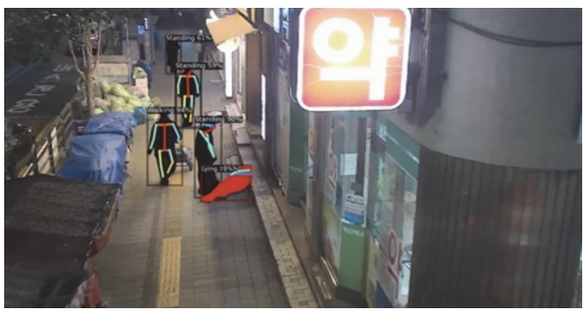
2.1. 공공·사회 인공지능 활용영역

공공·사회 분야에서는 통신기술의 발달을 기반으로 중앙정부와 지방정부 간의 시스템을 연계하고 정부와 국민 간의 쌍방향 소통 서비스를 제공하기 시작하였고, 더 나아가 각 공공기관은 빅데이터에 기반한 인공지능 기술을 활용하여 사용자 친화적이고 효율적인 행정 서비스를 제공하기 위해 노력하고 있다. 대표적으로 비대면 행정으로서 다양한 민원 상담 챗봇 서비스를 제공하고 있으며, 컴퓨터 비전 기술을 활용해 인공지능 기반 교통량 분석, 시민의 안전을 도모하기 위한 인공지능 기반 방범 CCTV에 활용하고 있다. 그뿐만 아니라 인사·채용, 업무 편의 및 자동화 분야 등의 행정에서 생산성을 높이기 위해 인공지능 기술을 도입하고 있다.

▼ 공공 및 사회 분야 인공지능 기술 활용 영역

민원 안내 인공지능 챗봇 서비스(세종특별자치시)[1]	스마트 교차로 서비스(부산광역시)[2]
	
카카오톡 메신저를 이용해 민원 현황, 시설 예약, 생활 정보를 포함한 코로나19 재난 정보 등 시민의 단순·반복 민원을 신속히 해결하고 지원하는 '민원 안내 인공지능 챗봇 서비스' 제공	기존의 교통관제는 교통경찰이 육안으로 교통량을 파악하고 신호를 제어하였으나 AI 기반 CCTV 교통관제 시스템을 도입하여 교통량 혼잡도를 분석한 후 교차로의 신호를 자동 제어

AI 기반 CCTV 위험 분석 서비스(대전광역시)[3]



도심 지역에서 주취자, 노숙자, 실신 등으로 쓰러진 사람을 실시간 탐지하는 행동 인식 AI 기술을 활용하여 안전사고 예방과 신속한 응급구조 업무 수행

인공지능 면접 프로그램 운영(한전KDN CKPASS)[4]

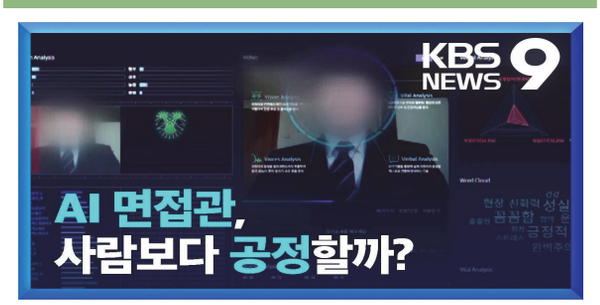
인공지능이 응시자의 얼굴, 감정 표정, 안면 색상의 변화 등을 분석하여 평가 면접을 진행함으로써 채용 과정에 소요되는 시간과 비용을 절감하여 업무 효율을 개선

2.2. 공공·사회 인공지능 이슈 사례

인공지능의 활용 분야가 광범위해짐에 따라 채용, 노동, 복지, 치안 등 공공서비스에도 인공지능에 의한 판단이나 결정이 많은 영향을 주므로 사회 전반에 걸쳐 인간의 기본권에 직접적인 영향을 미치는 사례가 발생하고 있다. 공공에서 활용되는 서비스는 우리의 삶 전반(복지, 의료 등)에서 활용되고 있는 만큼 오남용 및 악용에 따라 사생활 침해나 사회경제적 피해를 넘어서 미처 예상치 못한 복합적인 문제를 대거 유발할 수 있는 잠재적 위험성을 내포하고 있다.

▼ 공공 및 사회 분야 인공지능 이슈 사례

이슈 사례 1: 인공지능 채용 결과 정보 요청 이슈[5]



인권 단체가 공공기관 채용 과정에서 사용된 인공지능 면접 프로그램이 도출한 결과가 합리적인지에 대해 정보공개를 쟁송(爭訟)하여 일부 승소함(2022. 5.)

시사점

공공기관이 채용에 사용하는 기술의 의사결정에 대해 설명할 근거 자료를 준비하지 않거나 검증 없이 도입하여 운영하는 것을 지양해야 한다.

이슈 사례 2: 택시 배차 알고리즘 불공정 이슈[6]

시민이 택시 호출 시 특정 택시에 콜을 몰아주는 문제가 발생하여 인공지능 배차 알고리즘에 대한 불공정 문제가 제기됨(2023. 2.)

시사점

인공지능 기반 서비스는 공공성을 확보함으로써 결과에 대한 근거가 명확하고 차별성이 없이 개발하여 대중에게 제공돼야 한다.

2.3. 공공·사회 인공지능 신뢰성 정책 및 연구 동향

EU, 미국, 일본 등 주요 국가에서는 인공지능의 신뢰성 확보를 인공지능의 사회적·산업적 수용과 발전의 전제 조건으로 정의하고 신뢰성을 확보하기 위한 정책을 추진하고 있다. 또한 산학계에서도 관련 기술의 개발을 중심으로 신뢰성 확보를 위한 연구가 활발하게 이뤄지고 있다. 구체적으로 EU는 《인공지능 기반 서비스 및 솔루션 공공 조달의 데이터 윤리 백서》를 발간하여 인공지능 신뢰성의 3대 요소로 합법성, 윤리성, 기술적·사회적 안전성을 강조하였다. 또한 미국, 일본, 한국 등에서는 정부 차원에서 공공·사회 분야 인공지능의 신뢰성 확보를 위한 정책과 지침 개발 활동을 활발하게 전개하고 있다. 한편, 민간 부문에서도 공공·사회 분야의 인공지능 신뢰성 확보를 위한 가이드라인과 윤리 지침 및 거버넌스를 마련하는 등 자율적으로 인공지능의 신뢰성을 점검하고 확보할 수 있는 환경을 조성하는 데 노력을 기울이고 있다.

▼ 주요국의 공공·사회 분야 인공지능 신뢰성 관련 정책 동향

기관명	주요 정책(연도)	특징
EU 집행위원회	• EU AI 규제 법안(2021) • 《인공지능 기반 서비스 및 솔루션 공공 조달의 데이터 윤리 백서》(2020)	신뢰 가능 인공지능의 3대 요소로 합법성, 윤리성, 기술적·사회적 안전성을 제시하며, 인공지능이 정부 및 공공기관과 시민 간의 관계를 변화시키고 있으므로 유럽이 지향하는 신뢰 가능 인공지능의 혁신을 이끌어야 한다고 선언
호주	• 인권과 기술(2019)	인공지능 정보 기반 의사결정의 책무성에 대한 조사 및 합법성 여부 진단, 법치주의 원칙 보호, 평등 및 비차별 등 인권 증진을 제안
미국	• 설명 가능한 인공지능 4원칙(2020)	신뢰성 확보를 위해 정확성, 신뢰성, 보안성, 견고성 등의 설명가능성 원칙을 제정
캐나다	• 자동화된 의사결정 훈령(2019) • 알고리즘 영향평가 도구의 의무 적용 실시(2019)	정부 정책, 윤리 및 행정법적 고려 사항에 따라 자동화된 의사결정 시스템의 위험성 영역별로 구성된 평가 도구로 답변에 따라 공공기관의 자동화된 의사결정 시스템의 위험 영향 수준을 결정
영국	• 인공지능과 공공 윤리(2020) • 공공부문 인공지능 활용 가이드(2019) • 인공지능의 윤리와 안전을 고려한 시스템 설계·구현 가이드(2019)	공공부문의 인공지능 기술의 설계 및 활용의 윤리적 가치 체계와 실행 원칙을 제시
일본	• AI 활용 지침(2019) • 인간 중심 AI의 사회적 원칙(2019) • AI 개발 지침(2017)	개발자와 사업자의 기본 이념 및 AI 사회 원칙에 입각한 AI 개발 이용 원칙 강조
한국	• 인공지능 개발과 활용에 관한 인권 가이드라인(2022) • 인공지능 기반 미디어 추천 서비스 이용자 보호 기본 원칙(2021) • 인공지능 법·제도·규제 정비 로드맵(2020) • 공공기관 신뢰 가능 인공지능 구현 실행 가이드(2019) • 이용자 중심의 지능정보사회를 실현하기 위한 원칙(2019)	포용 성장, 지속 가능 발전, 복지 증진, 인간 중심 공정성을 원칙으로 인공지능 시대 실현을 위한 인공지능 법·제도·규제의 로드맵 및 실행 가이드를 제시

▼ 해외 주요 산학연 공공·사회 분야의 인공지능 신뢰성 연구 동향

기관명	활동 및 내용
국제인공지능윤리협회IAAE	2019년 10월에 '인공지능 윤리 헌장The AI ethics charter of the IAAE'을 발표하였고, 2022년 7월에 AI 기술과 윤리 이슈를 바탕으로 소비자가 문제없이 활용할 수 있도록 기업, 개발자, 사용자, 소비자가 지켜야 할 조항 10가지를 제시하는 〈디지털 휴먼 윤리 가이드라인〉을 발표함
유럽평의회CoE	유럽평의회 인공지능 특별위원회Ad hoc Committee on Artificial Intelligence, CAHAI 정책개발단은 2021년 5월에 인공지능 인권영향평가에서 참고할 수 있는 기존의 규범적 문헌을 개괄하고 인공지능의 적용을 검토한 〈인공지능 인권 민주주의 법치영향평가 보고서〉를 발표함
영국 에이다 러브레이스 연구소Ada Lovelace Institute	2020년에 알고리즘 감사와 알고리즘 영향평가의 조건을 명확히 하고 관련 연구 및 실무 현황을 설명하는 〈블랙박스 검사 보고서examining the black box〉를 발표함
미국시민자유연합ACLU	2020년에 알고리즘 형평성 툴킷algorithmic equity toolkit 을 제작하여 발표함으로써 정부에서 사용하는 감시 및 의사결정 기술을 식별하고 기술이 어떻게 작동하는지 이해하며, 그 영향과 효과 및 감독에 문제를 제기할 수 있도록 설계함
존스 홉킨스 대학교 정부 우수성 센터GovEx 샌프란시스코 시 및 카운티 하버드 케네디 스쿨 데이터 스마트 프로젝트 비영리단체 데이터 커뮤니티	2018년에 정부 부문에서 알고리즘을 구축하거나 도입하는 모든 사람에게 초점을 맞춘 윤리 및 알고리즘 툴킷ethics and algorithms toolkit을 개발함으로써 사용자가 알고리즘 사용으로 인한 윤리적 위험을 이해하고, 윤리적 위험을 최소화하기 위해 무엇을 할 수 있는지를 파악하는 데 도움을 주는 일련의 질문지를 제시함
뉴욕대학교 AI Now 연구소	2018년 알고리즘 영향평가 실시로 '공공기관 책임을 위한 실용적인 프레임워크 Algorithmic impact assessments: a practical framework for public agency accountability'라는 영향평가 모델을 개발함

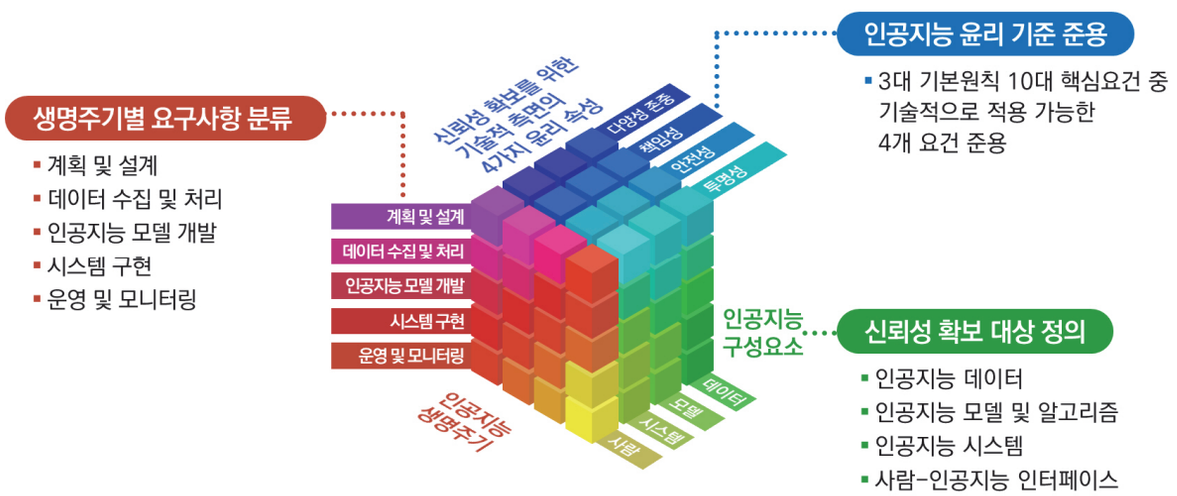
03 안내서 마련 과정

발간 배경에서 밝힌 바와 같이, 그간 국내외 많은 기관 및 기업에서 공공·사회 서비스에 활용되는 인공지능 기술의 신뢰성 확보를 위해 다양한 윤리 원칙과 지침, 가이드라인을 내놓았으나 기술적 관점에서 상세한 방법론을 제시한 사례는 아직 없는 실정이다. 따라서 인공지능 제품 및 서비스 개발 현장에서 데이터 과학자, 모델 개발자 등의 이해관계자가 실무 관점에서 신뢰성 확보에 참고할 수 있는 지침서의 필요성이 대두되었다. 이를 위해 2021년 1월부터 모든 산업 분야를 아우를 수 있는 일반 분야 안내서를 마련하였고, 이를 기반으로 2022년에는 공공·사회 분야에 특화된 안내서를 마련하였다. 공공·사회 분야의 안내서를 마련하는 과정에서는 산학계의 전문가와 실무자의 의견을 수렴하였다. 또한 공공·사회 분야에 관련 서비스를 제공하는 기업과의 협업을 통해 안내서의 현장 적용과 컨설팅 공동 연구를 진행하여 케이스 스터디를 마련하고 피드백을 받아 반영함으로써 실무 활용도를 높이고자 했다.

3.1. 인공지능 신뢰성 프레임워크 적용

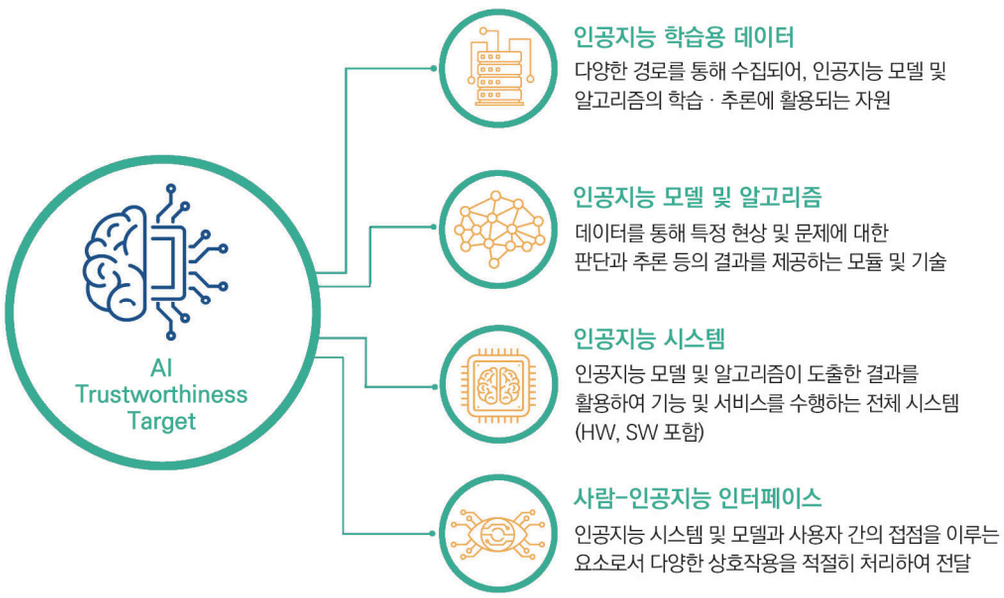
안내서 개발 과정에서는 신뢰성 확보를 위해 어떤 요소가 실무적으로 고려되어야 하는지를 가장 우선적으로 탐색하였다. 그 결과, 세 가지 주요 설계 요소를 도출하여 다음과 같이 매트릭스 형태의 체계화된 '인공지능 신뢰성 프레임워크'를 정의하고 제시하였다. 해당 프레임워크를 기반으로 공공·사회 분야에서 인공지능 서비스의 신뢰성 확보 여부를 검토할 때 참고해야 할 요구사항 및 검증항목이 개발되었다. 세 가지 주요 설계 요소의 내용은 다음과 같다.

▼ 인공지능 신뢰성 프레임워크



첫째는 인공지능의 구성 요소이다. 인공지능을 구성하는 4가지 요소로는 인공지능 학습용 데이터, 학습과 추론 기능을 수행하는 인공지능 모델 및 알고리즘, 실제 기능을 구현할 시스템, 사용자와 상호작용하기 위한 인터페이스가 있다. 각 구성 요소는 개별적으로 또는 통합적으로 인공지능 서비스의 생명주기에 따라 개발, 검증, 운영된다. 따라서 구성 요소별로 신뢰성 확보 방안을 수립하고 각 요소에 따른 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 배제되었는지를 검증
인공지능 모델 및 알고리즘	인공지능이 모델 및 알고리즘에 따라 안전한 결과를 도출하는지, 그에 대한 설명이 가능한지, 악의적인 공격에도 잘 대응하는지 등을 검증
인공지능 시스템	인공지능 모델 및 알고리즘이 적용된 전체 시스템을 대상으로 인공지능이 추론한 대로 작동하는지, 인공지능이 잘못 추론한 경우를 대비하는 방책이 있는지 등을 검증
사람-인공지능 인터페이스	인공지능 시스템의 사용자, 운영자 등이 인공지능 시스템의 동작을 쉽게 이해할 수 있으며, 인공지능의 오작동 시 사람에게 알리거나 제어권을 이양하는지 등을 검증

둘째, 인공지능 서비스의 생명주기는 인공지능 서비스의 구성 요소를 구현하고 운영하는 일련의 절차를 말한다. 기존 소프트웨어 시스템에서 다루는 공학 프로세스나 생명주기와 비슷하지만 인공지능 특성상 데이터 처리 및 모델 개발 단계가 별도로 필요하며 이외의 단계에서도 주요 활동에 대한 정의가 조금씩 달라진다. 현재 인공지능 혹은 인공지능 서비스의 생명주기는 다수의 문헌에서 6~8단계로 구분하고 있다. 대표적으로 OECD와 ISO/IEC에서 제시한 생명주기가 있는데, 본 안내서는 이 두 기구에서 제시한 생명주기를 대표성 있는 사례로 참고하였다. 그리고 실무자가 쉽게 활용할 수 있도록 각 생명주기 단계의 성격과 활동을 왜곡하지 않는 선에서 다음과 같이 5단계로 정리하였다.

▼ 인공지능 서비스 생명주기별 주요 활동

생명주기 단계	주요 활동
1. 계획 및 설계	- 인공지능 시스템의 관리 감독 조직 및 방안 마련 - 인공지능 시스템의 위험요소 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 특성feature의 문서화 - 인공지능 모델 구축을 위한 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 모델의 성능평가
4. 시스템 구현	- 문제 발생을 대비한 안전모드 구현 및 알림 절차 수립 - 인공지능 시스템의 검증 및 사용자 설명에 대한 평가
5. 운영 및 모니터링	- 시스템 모니터링 및 인공지능 모델 재학습을 통한 성능 보장 - 모델 편향 탐지, 공평성, 설명가능성 등 시스템 신뢰성 모니터링 - 치명적 문제 발생 시의 해결 방안 마련

인공지능 서비스의 생명주기 단계는 반복적·순환적인 성격을 띠지만 반드시 순차적인 것은 아니다. 본 개발 안내서는 이해를 돕기 위해 1단계부터 5단계까지 순차적으로 설명했으나 실제 데이터를 수집하고 가공하거나 모델을 개발·운영하는 과정에서는 순서가 달라질 수 있다.

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

EC, OECD, IEEE, ISO/IEC와 같은 국제기구에서는 인공지능 신뢰성의 하위 속성을 세분화해 제시하고 있다. 특히, ISO/IEC 24028:2020 - Information technology - Artificial intelligence - Overview of trustworthiness in artificial intelligence는 신뢰성 확보에 필요한 고려 사항의 형태로 키워드를 제공한다. 여기에는 투명성, 통제가능성, 강건성, 복구성, 공평성, 안전성, 개인정보보호, 보안성 등이 포함되지만 키워드 간의 관계나 신뢰성과의 연관성은 정의되지 않았다. 이처럼 관점에 따라 유사해 보이지만 조금씩 다른 용어들이 여러 문헌에서 제각각 달리 정의되어 있고, 합의된 속성 분류나 정의는 아직 없는 실정이다. 이에 앞서 언급한 EC, OECD, IEEE, ISO/IEC 등 여러 기구에서 제시한 속성과 키워드를 종합적으로 분석하고, 국내 산학연 전문가의 의견을 수렴해 합의점을 모색했다. 이처럼 폭넓은 의견 공유 과정을 거쳐 인공지능 신뢰성의 속성을 도출한 후에 이를 국가 인공지능 윤리 기준의 10대 핵심 요건에 대응시켜서 기술적 측면에서 다룰 만한 요건을 최종 선정하였다. 각 요건의 정의는 다음과 같다.

▼ 인공지능 신뢰성 요건

신뢰성 요건	정의
다양성 존중	인공지능이 특정 개인이나 그룹에 대한 차별적이고 편향된 관행을 학습하거나 결과를 추론하지 않으며, 인종·성별·연령 등과 같은 특성과 관계없이 모든 사람이 평등하게 인공지능 기술의 혜택을 받을 수 있는 것 – 관련 속성: 공평성·공정성fairness, 정당성justice – 관련 키워드: 편향bias, 차별discrimination, 편견prejudice, 다양성diversity, 평등equality – 국제표준(ISO/IEC TR 24027:2021 – Information technology – Artificial intelligence (AI) – Bias in AI systems and AI aided decision making) 정의: 공평성의 정의는 없다. 공평성은 복잡하고 문화·세대·지역·정치적 견해에 따라 다양하여 사회적으로나 윤리적으로 일관되게 정의하기가 힘들기 때문이다.
책임성	인공지능이 생명주기 전반에 걸쳐서 추론 결과에 대한 책임을 보장하기 위한 메커니즘이 마련되어 있는 것 – 관련 속성: 책무성responsibility, 감사가능성auditability, 답변가능성answerability – 관련 키워드: 책임liability – 국제표준(ISO/IEC TR 24028:2020) 정의: 엔터티의 작업이 해당 엔터티에 대해 고유하게 추적될 수 있도록 하는 속성
안전성	인공지능이 인간의 생명·건강·재산·환경을 해치지 않으며, 공격 및 보안 위협 등 다양한 위험을 관리하는 대책이 마련되어 있는 것 – 관련 속성: 보안성security, 강건성·견고성robustness, 성능보장성reliability, 통제가능성·제어가능성controllability – 관련 키워드: 적대적 공격adversarial attack, 복원력resilience, 프라이버시privacy – 국제표준(ISO/IEC TR 24028:2020) 정의: 용인할 수 없는 위험risk으로부터의 자유
투명성	인공지능이 추론한 결과를 인간이 이해하고 추적할 수 있으며 인공지능이 추론한 결과임을 알 수 있는 것 – 관련 속성: 설명가능성explainability, 이해가능성understandability, 추적가능성traceability, 해석가능성interpretability – 관련 키워드: 설명 가능한 인공지능eXplainable AI, XAI, 이해도comprehensibility – 국제표준(ISO/IEC TR 29119-11:2020 – Software and systems engineering – Software testing – Part 11: Guidelines on the testing of AI-based systems) 정의: 시스템에 대한 적절한 정보가 이해관계자에게 제공되는 시스템의 속성

이와 같이 인공지능의 신뢰성을 확보하기 위한 다양한 속성이 있는데, 각 신뢰성 속성의 정의를 파악할 뿐만 아니라 신뢰성 속성 간의 상호의존 관계도 중요하게 고려되어야 한다. 예를 들어, 인공지능 서비스 투명성에 대한 과도한 요구는 프라이버시와 관련된 위험을 초래할 수 있다. 또한 설명가능성만으로는 투명성을 보장하기에 부족하지만, 설명가능성은 투명성을 확보하기 위한 중요한 요소이다. 따라서 인공지능의 신뢰성 속성에 대한 충분한 이해를 바탕으로 인공지능 서비스를 제공하는 것이 중요하다. 또한 해당 인공지능 서비스가 고려한 속성에 대해 적절하게 이행하고 있는지를 지속적으로 검토해야 한다.

3.2. 공공·사회 분야 주요 고려사항

본 안내서는 기술적 관점에서 상세한 방법론을 제시함으로써 공공·사회 분야의 인공지능 서비스 개발 현장에서 실무자가 신뢰성 확보에 참고할 수 있는 실무 지침서로서의 성격을 지니는 것을 지향한다. 따라서 본 안내서는 일반 분야에서 다루는 구성 요소 및 생명주기를 바탕으로 인공지능의 신뢰성 확보를 위해 고려되어야 할 요소를 공공·사회 분야에 특화하여 정리하였다.

첫째, 공공·사회 분야에서 인공지능을 활용하는 서비스가 매우 다양하므로 본 개발 안내서에서는 공공기관에서 제공하는 사회적 서비스(G2C, B2G)를 주요 대상으로 한다. 단, 공공의 사례를 발굴하기 어려운 경우에는 사회적 관점의 공공서비스(B2C)로 대상을 확대하였다.

▼ 공공 및 사회 분야 개발안내서 참고 범위

정의	내용
G2C	Government to Customer의 약어로, 정부가 국민에게 제공하는 인공지능 기반의 서비스 – (예시) 음성인식 민원 접수시스템 등
B2G	Business to Government의 약어로, 기업이 개발하여 정부·공공기관에 제공하는 인공지능 기반의 서비스 – (예시) KT 어르신 돌봄 서비스, 병무청 AI 영상 면접 등
B2C	Business to Customer의 약어로, 기업이 국민에게 제공하는 인공지능 기반의 서비스 – (예시) AI 스피커, 챗봇, 음성인식 키오스크 등

둘째, 본 안내서는 한국정보화진흥원의 〈공공기관 신뢰 가능 AI 구현 실용 가이드〉와 영국의 〈A guide to using AI in the public sector〉를 주요 참고 자료로 삼았다. 또한 앞서 3.1.의 '인공지능 서비스 생명주기별 주요 활동'에서 정의된 틀을 기반으로 하여 공공·사회 분야에서 생명주기별로 고려해야 할 주요 활동을 추가로 정의하였다.

▼ 인공지능 서비스 생명주기별 주요 활동

생명주기 단계	주요 활동
1. 계획 및 설계	– 공공기관의 임무mission와 목적 및 목표에 따른 서비스 제공 가치 분석 • 인공지능 시스템의 콘셉트, 목적, 전제조건, 내용, 필수 조건 기획 • 사용자의 특성 분석 및 서비스 기획 – 인공지능 시스템의 평가 기준 및 관리 감독 방안 마련 • AI 시스템의 평가 계획 • 윤리행동강령 마련 – 인공지능 시스템의 위험 요소 분석 및 대응 방안 마련 • AI 시스템의 위험 확률 및 심각성 평가 • 인권영향평가, 인권실사, 품질인증 등의 분석 및 적용 • AI 시스템의 의사결정 프로세스에 인간의 개입 수준 결정

생명주기 단계	주요 활동
2. 데이터 수집 및 처리	- 데이터의 품질 확보, 데이터 사용자의 이해를 위한 정보 제공 방안 마련 - 데이터 라벨링 및 데이터셋 특성의 문서화 - 인공지능 모델 구축을 위한 데이터셋 마련
3. 인공지능 모델 개발	- 비즈니스 목적에 따른 인공지능 모델 구현 - 구현된 인공지능 모델 확인 및 검증 - 인공지능 모델 튜닝, 데이터 분석 및 추가로 필요한 데이터 고려 - 최종 인공지능 모델에 대한 성능평가 • 실증·검증: 다양한 차원과 환경에서 모델 성능 및 특성 평가, 테스트를 통한 모델 실증 및 결과에 따른 모델 조정
4. 시스템 구현	- 문제 발생에 대비한 안전모드 구현 및 알림 절차의 수립 • 적절한 안전장치로 인간 중심 가치 내재화(Kill Switch, Human in the loop 등) - 인공지능 시스템의 검증 및 사용자 설명 평가 • 파일럿 적용, 레거시 시스템과 결과 일치성 확인, 규정 준수 여부, 조직 변화 관리 및 사용자 경험 평가
5. 운영 및 모니터링	- 시스템 모니터링 및 인공지능 모델 재학습을 통한 성능 보장 - 모델 편향 탐지, 공평성, 설명가능성 등 시스템 신뢰성 모니터링 • AI 시스템 목표의 기관 미션 연계성, 사회경제적 영향평가 및 윤리 관점의 영향평가 지속 - 치명적 문제 발생 시, 해결 방안 도출 • 의사결정 및 조치 내용의 문서화 • AI 시스템 활용을 위한 투명한 정보 공개

3.3. 요구사항 및 검증항목 도출

다음 단계로 공공·사회 분야의 인공지능과 관련된 구체적인 요구사항과 검증항목을 도출했다. 우선 일반 분야 개발 안내서에서 참고한 문헌을 검토하고 표준화기구, 기술단체, 국제기구, 주요 국가 정부에서 공공·사회 분야의 인공지능 신뢰성 확보를 위해 발표한 정책, 권고안 그리고 표준을 기반으로 준수해야 할 기술적 요구사항을 도출하고 구체화해 제시했다. 이와 함께 〈공공기관 신뢰 가능 AI 구현 실용 가이드(2019. 12.)〉, 〈인공지능 개발과 활용에 관한 인권 가이드라인(2022. 5.)〉 등 국내에서 인공지능의 신뢰성 확보를 목적으로 발표된 사례를 검토했다. 이러한 과정을 거쳐 개발 안내서에 중요한 내용은 반영하고 중복된 내용은 제거하였다. 해당 참고 문헌은 다음과 같다.

▼ 인공지능 신뢰성 관련 주요 참고문헌

기관명	발간 연월	권고 및 표준안 명
대한민국	2020. 11.	국가 인공지능(AI) 윤리기준
	2019. 11.	이용자 중심의 지능정보사회를 실현하기 위한 원칙
	2019. 5.	공공기관 신뢰 가능 인공지능 구현 실행 가이드
유럽위원회	2020. 7.	The assessment list for trustworthy artificial intelligence
	2020. 5.	White paper on data ethics in public procurement of AI-based services and solutions
영국	2020. 2.	Artificial intelligence and public standards: a review by the committee on standards in public life
	2019. 6.	A guide to using artificial intelligence in the public sector

기관명	발간 연월	권고 및 표준안 명
미국 국립표준연구소 (NIST)	2023. 1.	NIST AI Risk Management Framework 1.0
국제표준화기구 (ISO/IEC)	2023. 2.	ISO/IEC 23894:2023, Information Technology – AI – Guidance on risk management
	2021. 11.	ISO/IEC TR 24027:2021, Information technology – Artificial Intelligence (AI) – Bias in AI systems and AI aided decision making
	2021. 3.	ISO/IEC TR 24029-1:2021, Artificial Intelligence (AI) – Assessment of the robustness of neural networks – Part 1: Overview
	2020. 5.	ISO/IEC TR 24028:2020, Information Technology – AI – Overview of Trustworthiness in artificial intelligence

이를 통해 최종 도출한 요구사항은 다음 표와 같다. 인공지능 윤리의 핵심 요건에 대응시킨 결과도 함께 표시했다.

▼ 인공지능 신뢰성 확보를 위한 기술적 요구사항과 윤리 요건 매칭 결과

요구사항		다양성 존중	책임성	안전성	투명성
요구사항 01	인공지능 시스템에 대한 위험관리 계획 및 수행		✓		✓
요구사항 02	인공지능 거버넌스 체계 구성	✓	✓	✓	✓
요구사항 03	인공지능 시스템의 신뢰성 테스트 계획 수립			✓	✓
요구사항 04	데이터의 활용을 위한 상세 정보 제공		✓		✓
요구사항 05	데이터 강건성 확보를 위한 이상 데이터 점검			✓	
요구사항 06	수집 및 가공된 학습 데이터의 편향 제거	✓	✓		✓
요구사항 07	오픈소스 라이브러리의 보안성 및 호환성 확보		✓	✓	
요구사항 08	인공지능 모델의 편향 제거	✓			
요구사항 09	인공지능 모델 공격에 대한 방어 대책 수립			✓	
요구사항 10	인공지능 모델 명세 및 추론 결과에 대한 설명 제공		✓		✓
요구사항 11	인공지능 시스템 구현 시 발생 가능한 편향 제거	✓			
요구사항 12	인공지능 시스템의 안전 모드 구현 및 문제발생 알림 절차 수립		✓	✓	✓
요구사항 13	인공지능 시스템의 설명에 대한 사용자의 이해도 제고				✓
요구사항 14	인공지능 시스템의 추적가능성 및 변경이력 확보		✓		✓
요구사항 15	서비스 제공 범위 및 상호작용 대상에 대한 설명 제공		✓		✓

3.4. 현장 적용 및 전문가 의견 수렴

신뢰성 확보를 위한 요구사항을 선별한 후에는 각 항목을 기술적 관점에서 검토하여 개발 안내서가 현업 종사자의 관점과 눈높이에 맞도록 고도화했다. 이 검토는 기술적 타당성, 효용성 및 포괄성이라는 관점을 포함했다. 각각의 세부 점검 항목이 요구사항에 해당하는 내용과 맞는지(타당성), 개발 현장에서 실무적으로 활용 가능한 내용인지(효용성), 검증을 위한 내용이 과거부터 지금까지 연구 내용을 폭넓게 포함하는지(포괄성)를 확인했다. 이를 위해 공공·사회 분야의 인공지능 전문가가 참여하여 직접 검토하고 자문했으며, 그 과정에서 다양하게 도출된 의견을 수렴하여 반영했다.

공공·사회 분야의 인공지능 전문가로는 기업의 기획자, 개발 프로젝트 리더, 교수, 국가연구소 책임연구원, 관련 국가 정책 담당자 등으로 산학계의 전문 종사자의 다양한 의견을 수렴하였다. 또한 자율주행 관련 서비스를 제공하는 기업과 협업하여 안내서의 현장 적용과 컨설팅 공동 연구를 진행하여 케이스 스터디를 마련하고 피드백을 받아 반영함으로써 실무 활용도를 높이고자 했다.

04 안내서 활용 대상

본 안내서는 공공·사회 분야에서 인공지능 서비스를 구현하는 과정에 직간접적으로 관련되거나 영향을 주는 모든 조직과 개인을 포함한 이해관계자가 참고할 수 있도록 작성되었다. 특히 업무상 기술적 관점에서 신뢰성과 관련된 시스템 기획자, 시스템 엔지니어, 데이터 공급자, 데이터 과학자, 인공지능 모델 개발자 등이 주요 대상이다. 이들이 공공·사회 분야 인공지능의 생명주기 단계마다 공공·사회 분야 인공지능의 신뢰성을 확보하기 위해 검토해야 할 주요 요구사항은 다음과 같다.

▼ 공공·사회 분야 인공지능의 생명주기 단계별 신뢰성 확보를 위한 요구사항

생명주기 단계	주요 행위자	주요 요구사항
1. 계획 및 설계	• 시스템 기획자 • 비즈니스 결정권자 • 품질 관리자 • 시스템 운영자 • 인공지능 윤리 전문가	– 인공지능 시스템 전체 생명주기에 걸친 신뢰성 확보 요구사항 검토 및 적용 방안 수립 – 공공·사회 분야의 인공지능 도입에 대한 윤리 지침 및 거버넌스 마련
2. 데이터 수집 및 처리	• 데이터 과학자 • 데이터 공급자 • 도메인 전문가	– 학습 데이터 확보 과정에서 발생할 수 있는 데이터 오류 및 편향을 관리하는 방안 확보
3. 인공지능 모델 개발	• 인공지능 모델 개발자 • 시스템 엔지니어 • 데이터 과학자	– 학습 모델의 편향적인 추론이나 공격에 대응하는 방안 수립 – 학습 모델의 추론을 해석하는 방안 제공

04 안내서 활용 대상

생명주기 단계	주요 행위자	주요 요구사항
4. 시스템 구현	• 시스템 엔지니어 • 인공지능 모델 개발자 • 품질 관리자	– 인공지능 시스템 개발 시 발생 가능한 편향이나 오류에 대응하는 방책 마련 – 인공지능 서비스가 도출한 결과에 대해 사용자 친화적인(user-friendly) 설명 제공 – 인권영향평가, 인권실사, 품질인증 등의 시험 및 평가 대응
5. 운영 및 모니터링	• 시스템 엔지니어 • 시스템 운영자 • 인공지능 모델 개발자 • 비즈니스 결정권자	– 인공지능 시스템에 문제 발생 시, 원인을 추적하여 대응 방안 마련 – AI 시스템 목표 및 윤리 관점의 운영 영향평가 지속

요구사항별로 대표 행위자와 협력 대상을 상세하게 대응시킨 결과는 다음과 같다. 요구사항별 협력 대상은 개발 안내서를 활용하는 서비스 및 기업 환경에 따라 상이할 수 있으므로 권장 사항으로 활용되길 바란다.

▼ 인공지능의 신뢰성을 확보하기 위한 요구사항별 활용 권장 대상

요구사항	대표 행위자	협력 대상
요구사항 01 인공지능 시스템에 대한 위험관리 계획 및 수행	• 시스템 기획자	• 비즈니스 결정권자 • 시스템 엔지니어 • 시스템 운영자 • 인공지능 모델 개발자 • 인공지능 윤리 전문가
요구사항 02 인공지능 거버넌스 체계 구성	• 시스템 기획자	• 비즈니스 결정권자 • 시스템 운영자 • 인공지능 윤리 전문가
요구사항 03 인공지능 시스템의 신뢰성 테스트 계획 수립	• 품질 관리자	• 시스템 기획자 • 시스템 엔지니어 • 비즈니스 결정권자
요구사항 04 데이터의 활용을 위한 상세 정보 제공	• 데이터 과학자	• 데이터 공급자 • 도메인 전문가 • 인공지능 모델 개발자
요구사항 05 데이터 강건성 확보를 위한 이상 데이터 점검	• 데이터 과학자	• 데이터 공급자 • 인공지능 모델 개발자
요구사항 06 수집 및 가공된 학습 데이터의 편향 제거	• 데이터 공급자	• 데이터 과학자 • 도메인 전문가 • 인공지능 모델 개발자
요구사항 07 오픈소스 라이브러리의 보안성 및 호환성 확보	• 인공지능 모델 개발자	• 시스템 엔지니어
요구사항 08 인공지능 모델의 편향 제거	• 인공지능 모델 개발자	• 데이터 과학자 • 시스템 엔지니어 • 인공지능 윤리 전문가

04 안내서 활용 대상

요구사항	대표 행위자	협력 대상
요구사항 09 인공지능 모델 공격에 대한 방어 대책 수립	• 인공지능 모델 개발자	• 시스템 엔지니어
요구사항 10 인공지능 모델 명세 및 추론 결과에 대한 설명 제공	• 인공지능 모델 개발자	• 데이터 과학자 • 시스템 엔지니어 • 시스템 운영자
요구사항 11 인공지능 시스템 구현 시 발생 가능한 편향 제거	• 시스템 엔지니어	• 시스템 운영자 • 인공지능 모델 개발자
요구사항 12 인공지능 시스템의 안전 모드 구현 및 문제발생 알림 절차 수립	• 시스템 엔지니어	• 시스템 운영자 • 인공지능 모델 개발자 • 품질 관리자
요구사항 13 인공지능 시스템의 설명에 대한 사용자의 이해도 제고	• 시스템 엔지니어	• 시스템 운영자 • 인공지능 모델 개발자 • 비즈니스 결정권자 • 인공지능 윤리 전문가
요구사항 14 인공지능 시스템의 추적가능성 및 변경이력 확보	• 시스템 엔지니어	• 인공지능 모델 개발자 • 데이터 과학자
요구사항 15 서비스 제공 범위 및 상호작용 대상에 대한 설명 제공	• 시스템 엔지니어	• 시스템 기획자 • 시스템 운영자 • 인공지능 모델 개발자 • 비즈니스 결정권자

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 서비스 특성 등을 고려하여 적절한 요구사항과 검증항목을 선택하여 적용하고, 기업에서 제공 중인 서비스 분야 및 환경에 맞게 신뢰성을 확보하기 위한 참고 자료로 활용하길 바란다. 더불어 인공지능의 신뢰성을 확보하기 위해서는 기술적 측면 외에도 윤리, 「개인정보보호법」과 같은 법·제도적 측면도 함께 고려되어야 한다. 그러므로 본 안내서를 활용하기 전에 인공지능의 윤리적 고려 사항 점검을 위한 '인공지능 윤리 기준 실천을 위한 자율점검표'와 개인정보보호의 준수 여부 점검을 위한 '인공지능(AI) 개인정보보호 자율점검표'를 선행적으로 검토할 것을 권고한다. 또한 인공지능에 해당하는 속성뿐만 아니라 기존 소프트웨어 시스템에 적용되는 전통적 속성도 적용되었는지를 확인할 필요가 있다. 따라서 본 안내서에 기술된 내용 외에도 시스템 성능, 보안 등 품질 관점의 검증 절차도 반드시 병행되어야 함을 밝힌다.

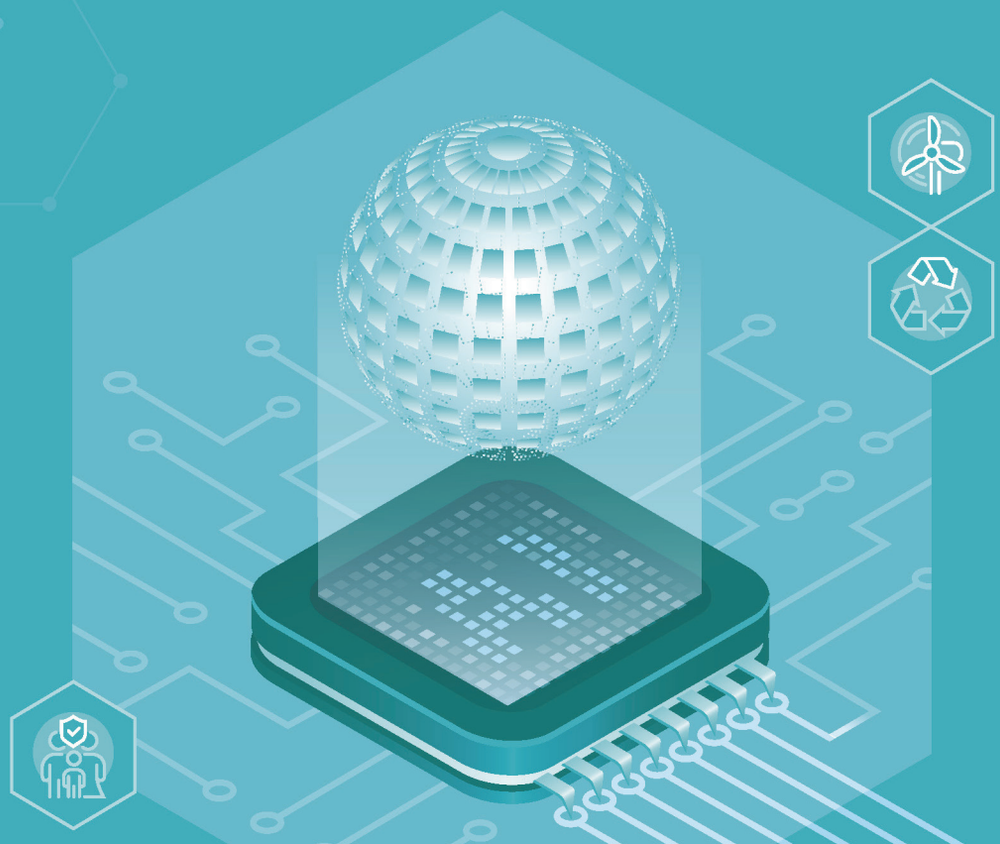
안내서는 다음과 같은 절차로 활용할 수 있다.

① **인공지능 서비스 위험 영향 분석:** 위험 영향 분석을 위해 우선적으로 고려할 사항은 점검 대상 인공지능 서비스의 활용 목적과 범위, 활용 대상에 따른 잠재적 영향이다. 유사한 목적의 서비스일지라도 인공지능의 추론 결과에 사람의 최종 개입 여부에 따라 위험이 미치는 영향의 정도가 달라질 수 있다. 영향 분석 과정에서 비즈니스 결정권자, 기획자, 개발자 및 시스템 운영자 등이 함께 논의하여 다양한 관점에서 분석할 것을 권장한다.

② **요구사항 선정:** '①'의 분석 내용을 토대로 개발 안내서 요구사항과 세부 요구사항 본문을 참고하여 인공지능 서비스에서 신뢰성을 확보하기 위해 필요한 요구사항을 선정한다. 만약 인공지능 서비스가 공공의 목적으로 활용되거나 질병 진단 또는 주식 거래과 같이 사람의 신체 및 재산에 되돌리기 힘든 피해를 줄 가능성이 있다고 판단된다면 가능한 한 모든 요구사항을 선정할 것을 권장한다. 반대로 특정 개인이나 집단에 차별이나 피해를 줄 가능성이 적은 서비스라면 모든 요구사항을 선정하지는 않더라도 신뢰성 점검을 위한 참고 자료로 활용할 수 있을 것이다. 이 과정에서 요구사항별 활용 권장 대상(대표 행위자 및 협력 대상)이 협의하여 불필요하다고 판단한 요구사항에는 'N/A'를 표시하여 점검 대상에서 제외할 수 있다.

③ **자가 점검 수행:** '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 못하면 이를 해결할 만한 수단이나 기술이 있는지를 확인할 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 있는데, 이때 '①'에서 분석한 서비스 영향 정도를 고려하여 대표 행위자와 협력 대상자가 협의하여 충족 여부를 판단하는 것이 바람직하다.

2023 신뢰할 수 있는 인공지능 개발 안내서 | 공공·사회 분야



PART 2

요구사항 및 검증항목

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

목차

목차

목차

목차

01 계획 및 설계

책임성 투명성

요구사항 01 인공지능 시스템에 대한 위험관리 계획 및 수행

대표행위자 | 시스템 기획자 협력 대상 | 비즈니스 결정권자 시스템 엔지니어 시스템 운영자 인공지능 모델 개발자 인공지능 윤리 전문가

- 공공·사회 분야의 인공지능 시스템을 구현하고 운영하는 과정에서 발생할 수 있는 위험 요소를 사전에 인식하고 위험의 파급효과를 분석하여 대응 방안을 마련한다.

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

Yes ☐ No ☐ N/A ☐

해당여부 판단

인공지능이 공익을 위해 기관 내부적으로 활용되거나 대민 서비스에 포함되는 경우에 본 항목을 고려하여 검증항목의 만족 여부를 판단하십시오.

- 위험관리는 위험 인식identification, 위험 분석analysis, 위험 평가evaluation, 위험 대응treatment으로 구분한다. 신뢰성 확보를 위해 이러한 네 가지 활동을 생명주기 단계별로 지속·반복적으로 수행함으로써 위험을 제거 및 방지하여야 한다. ISO 31000:2018 – Risk management – Guidelines에는 위험관리에 대한 개념 및 정의와 전체적인 흐름이 소개되어 있다.
- 다만, 인공지능의 신뢰성을 확보하는 과정에서 방해가 될 수 있는 위험 요소를 인식, 분석 및 평가하는 방법론은 기존의 소프트웨어 및 하드웨어 기반 시스템과는 상이할 수 있으므로 이 점을 고려해야 한다. ISO/IEC 24028:2020과 ISO/IEC 23894:2023에서는 인공지능의 신뢰성 관점에서 살펴보아야 할 위험 요소의 분류가 제공되어 있다.
- 공공기관에서 운영하는 인공지능 시스템을 구현한다면 해당 기관의 자체 위험관리 방법을 사용하여 위험 요소를 제거하는 방안을 도출하고 파급효과가 감소했는지를 확인해야 한다. 특히 각 기관의 목적과 임무에 따라 서비스 도입으로 공익성과 공정성을 해치는 부정적 영향이 일어날 가능성도 위험을 도출하는 과정에서 함께 검토해야 한다.

참고 캐나다 정부의 알고리즘 영향 평가 도구Algorithmic Impact Assessment tool[7]

- 알고리즘 영향평가(AIA)는 캐나다 재무부의 '자동화된 의사결정 훈령(2019. 4.)'을 지원하는 의무적 위험성 평가 도구이다. 이 도구는 자동화된 의사결정 시스템의 영향 정도를 판단하는 질문지 형식이다.
- 질문지는 위험성 48개와 완화 조치 33개에 대한 질문으로 구성되었으며, 시스템 설계, 알고리즘, 의사결정 유형, 영향 및 데이터를 비롯한 여러 요소를 기반으로 평가된다.
- 문항은 다음과 같은 다양한 요소에 의사결정이 미칠 영향을 측정하기 위해 고안되었다. 이 도구는 공개적으로 개발되었으며 일반인도 공개 라이선스에 따라 공유와 재사용이 가능하다.
  - ✓ 개인 또는 공동체의 권리
  - ✓ 개인 또는 공동체의 건강 및 복리
  - ✓ 개인, 단체, 공동체의 경제적 이익
  - ✓ 생태계의 지속가능성
  - ✓ 영향의 지속성과 가역성

정의된 위험성 영역

위험성 영역	정의
1. 프로젝트	
프로젝트 단계	프로젝트 소유자, 설명 및 단계 (설계 또는 구현)
사업 동기 및 긍정적 영향	의사결정 절차에 자동화를 도입하려는 동기
위험성 개요	높은 수준의 프로젝트 위험 지표
프로젝트 권한	프로젝트에 대해 새로운 정책 권한을 추구할 필요가 있음
2. 시스템	
시스템 관련	시스템의 사양 (예: 이미지 인식, 위험성 평가 등)
3. 알고리즘	
알고리즘 관련	알고리즘의 투명성, 쉽게 설명되는지 여부 등
4. 의사결정	
의사결정 관련	자동화된 의사결정의 분류 (예: 보건의료서비스, 사회 지원, 면허 등)
5. 영향	
영향평가	지속성, 가역성 및 영향을 받는 분야 (예: 권리, 건강, 경제, 환경 등)
6. 데이터	
소스	의사결정을 자동화하는 데 사용된 데이터의 출처 및 등급
유형	정형/비정형으로 사용되는 데이터의 종류(오디오, 텍스트, 이미지, 비디오 등)

완화성 영역 정의

완화성 영역	정의
7. 자문	
내부 및 외부 이해 관계자	개인정보보호 및 법률 전문가를 비롯해 자문받은 내외부의 이해관계자
8. 위험성 제거 및 완화 조치	
데이터 품질	데이터가 대표성을 띠고 편향적이지 않도록 보장하는 절차와 이 절차와 관련된 투명성 조치
절차적 공정성	시스템과 그 의사결정을 감사하는 절차 및 회복 조치
개인정보보호	개인정보를 보호하는 조치

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

Yes ☐ No ☐ N/A ☐

- 공공·사회 분야 서비스는 다음과 같이 부정적 결과를 초래할 수 있는 위험 요소를 확인하고 도출하도록 한다.

  ① (공익·공정성) 인공지능 시스템 도입으로 인한 불평등, 환경오염, 인권 차별 등의 부정적 영향을 유발할 가능성

  - ✓ 인공지능이 성별을 식별하여 성차별적인 결과를 출력하는지 확인
  - ✓ 인공지능이 학습할 때 성소수자, 장애인 등 특정 소수집단에 차별적인 결과를 출력하도록 학습되는지 확인
  - ✓ 연령·직업을 기준으로 업무 처리의 우선순위를 지정하는지 확인

  ② (보안·개인정보보호) 국민의 개인정보를 위해 개인정보 파일 및 시스템 관리 체계의 취약점

  - ✓ 개인정보 학습으로 성능을 자체적으로 향상하도록 되어 있는 인공지능 기기가 해킹에 취약하지 않게 방어법이 마련되어 있는지 확인
  - ✓ 악의적인 사용자가 공격을 시도할 때 미리 감지하여, 해당 공격에 대응하는 방어법을 사용할 수 있는지 확인
  - ✓ 서비스 중인 여러 인공지능 개체가 학습한 데이터를 클라우드 등을 통해 인공지능 시스템 전체가 공유할 때 개인정보가 엄격하게 관리되고 있는지 확인

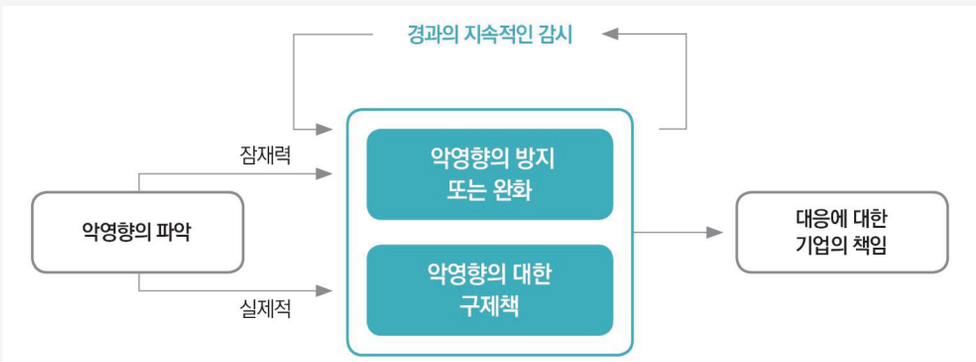
  ③ 이 외에 서비스 대상 산업 분야에서 요구되는 특성[feature]에 따른 위험 요소

  - ✓ 치안, 개인인증 등에 사용하는 얼굴 인식 과정에서 인식률 저하로 인한 위험 또는 외부의 공격 가능성이 있는지 확인
  - ✓ 음성인식, 이미지 등을 사용하여 만들어진 가짜 뉴스, 영상 등을 진짜 영상과 구분할 수 있는지 확인

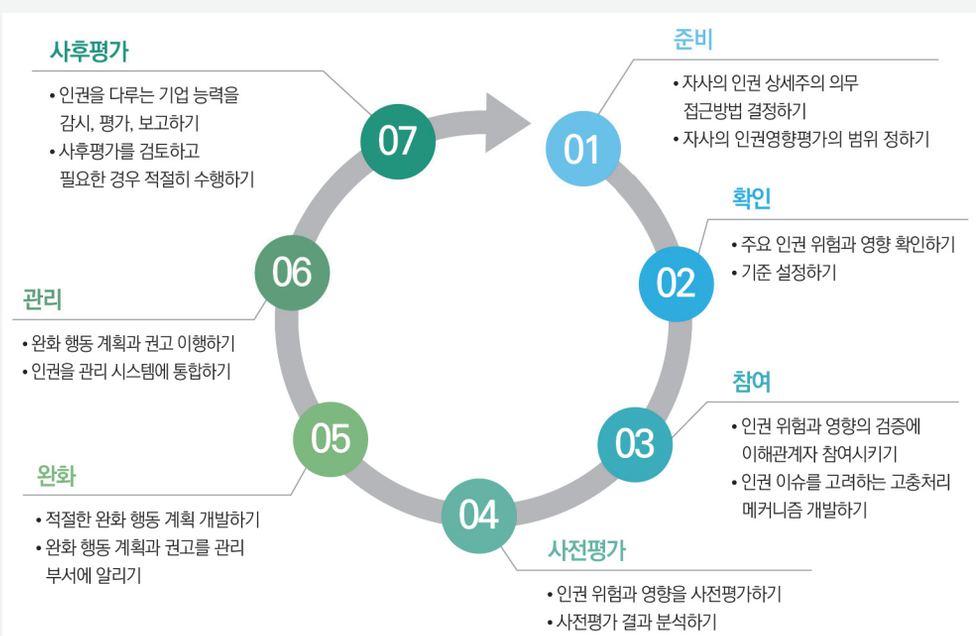
- 이 과정에서 델파이, 시나리오 등의 기법을 활용할 수 있으며, 실제 서비스의 영향을 받는 사람들에게서 피드백을 받는 인권영향평가, 인권실사 등을 통해 인공지능 시스템이 지닌 잠재적 위험의 심각성과 인권에 미치는 영향을 식별할 수 있다.

참고 인권실사 및 인권영향평가 과정[8]

- 인권실사Human rights due diligence란, 기관 또는 기업이 서비스를 제공하는 과정에서 발생하였거나 발생 가능한 인권 위험을 식별하고 분석하여 예방·조치하는 일련의 절차이다.

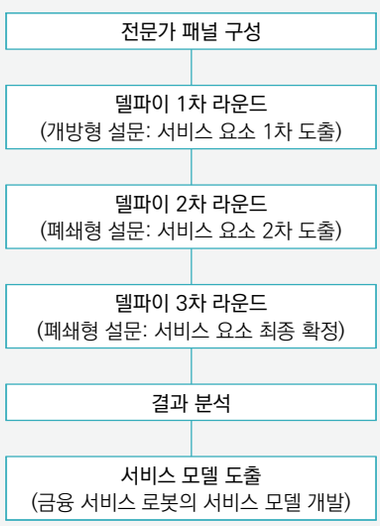


- 인권영향평가는 인공지능에 대한 인권 기반 접근법 중 하나로서 인권 관점에서 위험을 식별하고 평가할 때 사용한다.
- 인권영향평가는 대체로 계획-실행-평가-채택 과정으로 수행되며 이러한 활동은 일회성이 아닌 지속적으로 수행할 것을 권고한다.
- 기존의 인권영향평가 국제 규범으로는 유엔 기업과 인권 이행 지침, 최종 사용 관련 인권 위험 식별 및 평가, 독일의 국가인권정책기본계획NAP, 다국적 기업을 위한 OECD 지침이 있다.
- IFCInternational Finance Corporation가 IBLFInternational Business Leaders Forum와 유엔글로벌콤팩트의 협조하에 작성한 '인권영향평가 및 관리 지침서(HRIAMHuman Rights Impact Assessment and Management)'에서는 준비-확인-개입-평가-개선-운영-최종 평가로 이어지는 7단계 과정으로 수행할 것을 권고한다.[9]



참고 금융 서비스 로봇의 델파이 기법 활용 사례[10]

- 다음은 신협중앙회의 은행을 방문하는 고객에게 새로운 경험을 제공하고 창구 직원을 지원하기 위해 개발한 은행지점 도입용 금융 서비스 로봇의 사례이다. 신규 IT 융합 서비스를 개발하기 위해 다양한 기술 분야의 전문가의 아이디어를 합리적으로 수렴하기 위해 델파이 기법을 활용하였다.
- 연구는 총 3차에 걸친 델파이 기법을 적용하였으며, 델파이 1차 라운드를 통해 전문가 패널에게서 신규 서비스에 대한 서비스 요소를 자유로이 응답받고, 내용을 분석하여 서비스 요소 체계를 구성하였다. 델파이 2차 라운드에서는 도출된 서비스 요소를 대상으로 전문가들의 의견을 묻는 폐쇄형 설문을 진행했고, 델파이 3차 라운드에서는 델파이 2차 라운드의 응답 결과를 중심으로 다시 한번 폐쇄형 설문을 진행함으로써 전문가들의 합의된 의견을 도출했다.
- 최종적으로 제거되지 않고 남아 있는 서비스 요소를 토대로 금융 서비스 로봇의 서비스 모델을 개발하는 과정으로 진행된다.



- 델파이 1차 라운드 조사는 은행지점에 도입될 금융 서비스 로봇이 갖춰야 할 서비스나 기능에 대해 전문가들에게 개방형 설문을 제시하였다. 수집한 결과를 바탕으로 금융 서비스 로봇의 서비스 유형을 선행 사례와 이론적 연구에 기반하여 6개 차원(영역)으로 그룹화하고 각각의 조작적 정의를 내렸다. 결과적으로 6개 차원에 대한 세부 서비스 기능 30개를 다음과 같이 도출하였다.

차원	서비스 기능
인터페이스 (A)	고객과 서비스 로봇 간의 상호작용에 필요한 기본 요소
	A-1. 캐릭터를 활용한 친밀감을 줄 수 있는 로봇 외형
	A-2. 업무내용 전달을 위한 디스플레이 탑재
	A-3. 인공지능 기반 음성인식을 통한 대화형 서비스
	A-4. 인공지능 기반 음성인식을 통한 정확한 고객식별 및 인증 서비스
	A-5. 디스플레이 표시 및 음성발화에 따른 개인정보의 유출방지 서비스
	A-6. 개인 스마트폰(모바일)을 통한 상호작용 인터페이스(업무예약 및 확인)
	A-7. 사기행위 방지(FDS) 서비스

차원	서비스 기능
안내 서비스 (B)	고객 방문 시 혹은 고객 지원서비스 실패 시 안내하는 서비스 기능
	B-1. 로봇 처리범위를 벗어난 전문적 업무지원 필요 시 전문인력과의 화상상담 서비스
	B-2. 고객의 방문 목적에 대한 창구 위치 및 대기시간 안내 서비스
	B-3. 다음 방문일정 안내 서비스
	B-4. 방문 고객을 식별하고 대기표를 발급 후 고객 차례가 오면 안내(모바일 알림 등)
기본적 금융 서비스 (C)	단순하고 빈번한 금융 업무 서비스
	C-1. 입·출·송금 및 공과금 납부 서비스
	C-2. 환전 서비스
	C-3. 대출신청(대출한도 및 신용등급 조회) 서비스
	C-4. 카드신청(현금/체크/신용 등) 및 발급 서비스
	C-5. 각종 증명서류 발급(거래확인증, 잔액증명서 등) 서비스
	C-6. 제신고(諸申告) 업무(사고신고, 주소변경 등)
전문적 금융 서비스 (D)	복잡하고 전문성이 필요한 금융 업무 서비스
	D-1. 개인신용, 거래 패턴 등 고객 개인에 대한 빅데이터 분석을 통한 자산진단 서비스
	D-2. 인공지능 알고리즘을 활용한 상품(카드, 보험) 추천 서비스
	D-3. 로보 어드바이저 결과를 고객의 스마트폰으로 전송하는 서비스
	D-4. 위치기반서비스geofencing를 활용한 맞춤형 쿠폰 제공 서비스
부가 서비스 (E)	서비스 대기 중인 고객에게 흥미위주의 정보 혹은 기타 정보를 지원하는 서비스
	E-1. 금융정보(환율, 주식) 제공 서비스
	E-2. 동영상 교육(모바일 앱, 인터넷뱅킹 활용법 서비스)
	E-3. 사진촬영 후 이메일, 문자 발송
	E-4. 오늘의 운세, 노래, 퀴즈, 맛집 소개, 교통상황 안내 서비스
	E-5. 오늘의 주요 뉴스
지점관리 서비스 (F)	은행지점 관리 지원을 위한 기타 기능
	F-1. 영업시간 이후의 방범 및 경비
	F-2. 조합 홍보 서비스
	F-3. 온도, 습도 센서를 통한 직접 모니터링 기능
	F-4. 영상인식을 통한 지점 혼잡도 체크

- 다음은 델파이 3차라운드의 결과로서, 내용타당도 지수가 0.78 미만이거나 내용타당도 비율이 0.37 미만인 서비스 기능이 제거된 것이다. 조사 결과, 차원 E, F가 완전히 제거되었으므로 금융 서비스 로봇의 서비스 모델은 A, B, C, D 차원에 집중되어야 할 것으로 분석할 수 있다.

차원	서비스 기능	평균	표준편차	내용타당도 지수	내용타당도 비율
인터페이스 (A)	A-1	4.136	0.757	0.864	0.727
	A-2	4.591	0.492	1.000	1.000
	A-3	4.591	0.577	0.955	0.909
	A-4	4.409	0.834	0.864	0.727
	A-5	4.773	0.419	1.000	1.000
	A-6	4.182	0.490	0.955	0.909

차원	서비스 기능	평균	표준편차	내용타당도 지수	내용타당도 비율
안내 서비스 (B)	B-1	4.000	0.426	0.909	0.818
	B-2	4.273	0.617	0.909	0.818
	B-4	3.955	0.638	0.864	0.727
기본적 금융 서비스 (C)	C-1	4.545	0.582	0.955	0.909
	C-3	4.277	0.794	0.864	0.727
	C-4	3.864	0.625	0.818	0.636
	C-5	4.591	0.577	0.955	0.909
	C-6	4.045	0.638	0.909	0.818
전문적 금융 서비스 (D)	D-1	4.364	0.643	0.909	0.818
	D-2	4.818	0.386	1.000	1.000
	D-3	4.182	0.490	0.955	0.909

- 도출된 위험 요소는 이를 야기할 수 있는 원인과 발생 가능한 결과로 분석해야 한다. 발생 가능한 결과란 사회적으로 부정적인 영향을 미칠 수 있는 현상이나 사고를 의미한다. 앞서 언급한 불평등, 환경오염, 인권 차별 등이 이에 해당한다.
- 위험 요소의 발생 결과에 대해 발생 확률과 영향도(영향 받는 대상 및 규모, 취약계층, 영향 받는 방식), 심각성과 같은 척도로 위험 수준을 평가할 수 있다. 이는 위험 요소의 파급효과를 의미하는데, 파급효과가 큰 위험 요소에 대응할 수 있는 방안을 최우선으로 마련해야 한다.

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

Yes ☐ No ☐ N/A ☐

해당여부 판단

01-1 에서 위험 요소를 분석했다면 본 항목을 고려하여 검증항목의 만족 여부를 판단하십시오.

- 01-1 에서 분석된 위험 요소별 대응 방안을 마련해야 한다. 대응 방안이란, 인공지능 시스템의 위험 요소를 해소하기 위해 절차적·기능적·기술적으로 적용이 가능한 모든 방법을 의미한다. 위험에 대한 대응 전략은 긍정적 위험관리 전략과 부정적 위험관리 전략으로 구분하며, 정보통신기술의 보안 관리 방법을 설명하는 국제표준 ISO/IEC TR 13335 – Information technology – Security techniques – Management of information and communications technology security에서 소개하고 있다.
- 신뢰할 수 있는 인공지능 시스템을 구현하기 위한 원칙을 견지할 수 있도록 모든 이해관계자가 각자의 역할, 상황 및 행동 능력에 따라 인공지능 시스템의 생명주기 단계별 위험 요소를 지속적으로 모니터링하고 대응 방안을 적용해야 한다. 이 과정에서 이해관계자 간의 협력과 소통을 유도하고 적절하게 처리해야 한다.

참고 ISO/IEC 23894:2023 위험 대응 전략

- ISO/IEC 23894:2023에서는 ISO 31000:2018에 따라, 위험에 대해 다음과 같이 대응하도록 제안한다.
- 단, 공공·사회 분야의 인공지능 서비스는, '비용 대비 효과를 고려하여 관련 위험을 그대로 수용하거나, 위험을 수용한 채로 기회를 최대화하는 전략'은 리스크가 큰 접근 방식이기 때문에 가능하면 다른 전략으로 대응해야 한다.
  - ✓ 심각한 위험의 경우, 발생 가능성을 원천적으로 제거하는 전략
  - ✓ 비용 대비 효과를 고려하여 관련 위험을 그대로 수용하거나, 위험을 수용한 채로 기회를 최대화하는 전략
  - ✓ 위험을 유발하는 원인을 제거하는 전략
  - ✓ 위험의 발생 가능성이나 영향력을 저감하는 전략
  - ✓ 위험을 공유하여 감소시키는 전략(다른 계약 또는 보험 가입)
  - ✓ 정보에 입각한 결정으로 위험을 유지하는 전략

참고 영국 정부, 공공부문 인공지능 프로젝트의 위험과 완화 방법[11]

- 영국 정부는 공공 분야에서 인공지능 기술의 활용을 촉진하고 선도적 사례를 창출하기 위해, 2019년 6월에 '공공부문 인공지능 활용 가이드'를 발간하였다.
- 해당 가이드는 공공기관이 인공지능을 활용할 때 고려해야 할 위험성과 완화 방안을 다음과 같이 설명한다.

위험	세부 내용
편향 또는 차별 징후	모델의 편향된 결과를 모니터링하거나 공정하고 설명할 수 있게 만드는 프로세스가 있는지 확인
데이터 사용이 법·제도 및 정부의 규정을 준수하지 않음	인공지능 데이터 준비에 대한 지침을 참조
기밀 유지 및 데이터 무결성 유지를 보장하는 보안 프로토콜이 존재하지 않음	필요한 보안 프로토콜을 정의하기 위해 데이터 카탈로그를 구축
데이터에 접근할 수 없거나 열악한(poor) 데이터 품질	내부/외부에서 초기 단계에 사용할 데이터셋을 매핑하고, 이후에 데이터를 완전성, 고유성, 관련성, 충분성, 적시성, 대표성, 타당성(또는 일관성)의 조합에 대한 기준으로 정확하게 평가
모델 통합 불가능	인공지능 모델 구축 초기에 엔지니어를 포함해 개발된 모든 코드가 운용될 준비가 되었는지를 확인
모델에 대한 책임 프레임워크가 없음	인공지능 모델의 서로 다른 영역에서의 최종 책임자를 정의하기 위해 명확한 책임 기록을 확립

01-2a 위험 요소 제거 방안을 도출하고 파급효과가 감소하였는지 확인하였는가?

Yes ☐ No ☐ N/A ☐

- 위험 요소를 초래할 수 있는 원인을 분석하고, 이에 대응하는 방안을 마련해야 한다. 대응 방안은 ISO/IEC 24028:2020에 제시되어 있다.
- 앞서 분석한 위험 요소의 결과를 바탕으로 파급효과가 가장 큰 위험 요소에 대응하는 방안을 우선 적용해야 하며, 위험의 심각도가 높으면 인공지능 시스템의 판단 결과에 사람의 개입도 고려해야 한다. 해당 내용은 요구사항 12 에서 자세하게 다룬다.
- 공공/민간 서비스는 인공지능 시스템의 윤리 가이드라인을 바탕으로 한 윤리적 행동강령에 기반하여 위험 식별과 대응 방안이 인간 중심 가치와 공정성 촉진 관점에서 이루어지도록 한다. 따라서 공공서비스를 개발하는 민간기업에서는 사전에 각 기관의 윤리적 행동강령을 확인하고 위험 요소를 식별해야 한다.
- 이후에 파급효과에 대한 재평가를 시행하여 위험 요소가 제거되거나 영향도가 완화되었는지 확인해야 한다. 파급효과를 재평가하는 방법은 01-1a 의 내용을 참고한다.

01-2b 위험관리 관련 규정에 따라 이행하였음을 입증/관리 할 수 있는 방안을 마련하였는가?

Yes ☐ No ☐ N/A ☐

- 공공기관은 인공지능 서비스가 기관 내 윤리강령 및 규정에 따라, 인공지능 서비스의 개발·운영 전 단계에 걸쳐 수행되는 활동과 의사결정이 '신뢰할 수 있는 인공지능 구현을 위한 원칙'을 모두 준수하고 있다는 것을 입증해야 한다.
- 따라서 인공지능 서비스의 개발·운영 과정에서 발생한 주요 결정 사항과 위험, 이슈, 조치 사항을 문서화하고, 기관 내 윤리위원회 및 규제위원회가 인공지능 서비스에서 담당하는 책임과 역할을 보장함으로써 행위자의 책임성을 제고하고 각 활동을 입증할 수 있도록 한다.

참고 신뢰할 수 있는 인공지능 구현을 위한 원칙 준수의 입증·관리 예시[12]

- 내부 점검표는 개인정보보호 자율점검표, 인공지능 미디어 추천 서비스 이용자 보호 기본원칙 해설서 등 인공지능 개발·운영 가이드라인을 참고하여 인공지능 서비스와 조직에 적합한 항목으로 구성하도록 하며, 구체적인 규정과 취지, 세부 사항을 기술한다.
- 내부 점검표는 AI 개발자 및 운영자가 개인정보의 적법과 안전한 처리, 침해 예방 등을 위해 '준수 사항의 사전점검'이나 '서비스 운영 중의 수시 점검'을 하는 데 활용할 수 있다. 이 외에도 AI 기술 및 서비스 분야의 업무를 담당하는 이들을 교육할 때도 자료로 활용할 수 있다.

다양성 존중 | 책임성 | 안전성 | 투명성

요구사항 02 인공지능 거버넌스governance 체계 구성

대표행위자 | 시스템 기획자 협력 대상 | 비즈니스 결정권자 | 시스템 운영자 | 인공지능 윤리 전문가

- 인공지능 시스템은 윤리와 관련된 문제가 발생할 가능성을 잠재적으로 내포하고 있다. 이러한 인공지능 시스템의 사회적 영향과 결과를 예측하고 대비하는 조직을 구성하는 것은 인공지능 신뢰성을 확보하는데 중요한 요소이다. 따라서 인공지능 관련 법, 규제, 정책, 표준 및 지침을 정리하여 내부적으로 준수해야 할 규정을 수립하고, 이를 관리·감독하는 인공지능 거버넌스* 체계를 구성한다.

  * 조직organization의 목적, 기회, 위험 및 이익을 파악하는 지속적인 프로세스

- 거버넌스 조직이 구성되면 공공·사회 분야 서비스에 인공지능 시스템을 도입하기 전에 잠재적인 사용자층과 서비스 범위를 분석한 결과를 바탕으로 계획과 설계 방안을 마련한다. 이는 효율적이고 접근성 높은 공공 시스템을 위한 사전 작업이다.

02-1 인공지능 거버넌스에 대한 지침 및 규정을 수립하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

개발하는 인공지능 서비스가 공공서비스로서 사회에 윤리적 영향을 미치거나 지적재산권 분쟁이 우려되는 경우에 본 항목을 고려하여 만족 여부를 판단하십시오.

- 인공지능과 관련된 조직에서는 인공지능 시스템의 신뢰성을 확보하기 위한 거버넌스 체계를 구성할 필요가 있다. 인공지능 시스템은 학습이나 추론 과정에서 윤리 및 지적재산권Intellectual Property, IP 관련 문제나 보안 및 개인정보 이슈가 발생할 수 있기 때문이다. 공공서비스는 윤리적 문제로 인한 사회적 이슈가 발생하면 파급력이 크고, 공공기관에서 운영하는 서비스는 대부분 민간기업에 발주하여 진행되기 때문에 지적재산권 분쟁 규정이 존재한다. 따라서 이러한 위험 요소에 대비하기 위해 내부적으로 인공지능 거버넌스에 대한 지침 및 규정을 수립해야 한다.

- NIST의 AI RMFRisk Management Framework에서는 인공지능 시스템의 생명주기에 따라 내부 규정, 절차, 과정 및 실제 행위가 투명하고 효율적으로 이뤄져야 한다고 언급하였다. 즉, 인공지능과 관련된 법, 규제 관련 요구사항이 이해되며 관리되어 문서화되고, 위험관리 절차와 산출물이 체계적으로 투명하게 관리되어야 한다.

- 내부적으로 수립해야 할 규정은 활용 측면에 따라 크게 두 가지로 구분하여 마련할 수 있다.

  ✓ 첫째, 인공지능 관련 법, 규제, 정책, 표준 및 지침을 채택·정리하여 내부적으로 이행해야 할 지침과 규정을 수립해야 한다.

  ✓ 둘째, 인공지능 시스템의 생명주기에 따른 조직의 역할과 책임을 명확하게 문서화해야 한다.

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

Yes ☐ No ☐ N/A ☐

- 인공지능 거버넌스 체계에서 기본적으로 갖춰져야 할 단계는 윤리 원칙을 수립하는 것으로, 인공지능과 관련된 법, 규제 및 정책을 이해한 후 내부적으로즉, 인공지능과 관련된 위험을 인식하고 대비하기 위해 기업 성격에 맞는 핵심 가치를 선정하고 이와 관련된 표준 및 지침을 채택하여 내부 규정을 제공해야 한다.
- 공공기관에서 운영하는 서비스는 서비스 기획에 있어 기관 내부적으로 중요하게 다루는 사회적 가치와 서비스를 통해 이루고자 하는 목표가 있다. 이에 기반하여 인공지능 시스템의 신뢰성을 확보하기 위한 인공지능 거버넌스 및 조직 전체의 업무, 역할, 의무 및 책임을 명확하게 정의해야 한다. 이와 더불어 인공지능 시스템의 생명주기에 걸쳐 도출되는 산출물 관리에 대한 지침도 마련한다면 서비스를 더 효율적으로 관리할 수 있다.

참고 싱가포르의 인공지능 거버넌스 계획[13]

싱가포르 개인정보보호 감독기구Personal Data Protection Commission, PDPC는 서비스 주도의 경제에서 인공지능이 개인정보를 처리하는 지능형 시스템에 구축될 것으로 보고, 책임성에 기반을 둔 인공지능 거버넌스 프레임워크를 제시하였다. 그리고 인공지능의 윤리적 이슈·거버넌스·소비자 보호에 대한 논의를 촉진하기 위한 '인공지능 거버넌스 계획(AI Governance Initiatives)'을 발표했다.(2018. 6. 5.)

책임 있는 인공지능의 원칙 (PRINCIPLES FOR RESPONSIBLE AI)	인공지능 거버넌스 프레임워크 (GOVERNANCE FRAMEWORK FOR AI)
• 인공지능이 기업과 사회 전체에 이익을 주기 위해서는 인공지능 기술 이용에 대한 신뢰와 이해를 증진할 수 있는 일련의 원칙을 인공지능 거버넌스 프레임워크에 통합해야 하는바, PDPC는 다음과 같은 두 가지 주요 원칙을 도출함 1. 원칙1: 인공지능에 의한 결정은 반드시 설명 가능하고 투명하고 공정할 것(설명가능성, 투명성, 공정성) 2. 원칙2: 인공지능을 이용한 의사결정은 반드시 인간 중심일 것(인간 중심)	• 이 프레임워크는 ① 인공지능 거버넌스 프레임워크의 목표 확인 ② 조직별로 적절한 조정 방법 선택 ③ 소비자 관계 관리 프로세스에 대한 고려 ④ 의사결정 및 위험평가 프레임워크 구축 등의 4단계로 구성됨 • 이 프레임워크는 △ 개인에게 영향을 미치는 상황에서 인공지능 기술의 전반적인 설계·적용·이용에 초점을 맞추고 △ 모든 분야에 적용 가능한 기본 기준을 제시하되 특정 분야와 조직이 추가적인 표준을 통합할 수 있는 가능성을 배제하지 않음

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

Yes ☐ No ☐ N/A ☐

해당여부 판단

02-1에 따라 인공지능 거버넌스의 운영 지침 및 규정을 마련했다면 본 항목을 고려하여 만족 여부를 판단하십시오.

- 02-1에서 언급했듯이, 인공지능 시스템은 윤리와 관련된 문제가 발생할 수 있다는 위험 요소가 존재한다. 따라서 다양한 위험 요소를 인식하고 관련 규정을 마련하여 이를 실행할 수 있도록 관리 및 감독하는 조직이 필요하다.
- 유네스코가 발표한 인공지능 윤리 권고는 인권 및 법치 사회에 대한 인공지능 시스템의 영향을 식별, 예방 및 완화하고 그에 따른 의무를 이행하기 위해 감독 메커니즘이 있어야 한다고 명시하고 있다.
- 따라서 인공지능 거버넌스는 윤리적 측면에 관한 규정을 마련하고, 지침 준수 및 절차적 요건 충족 여부 등을 포함하여 감독하여야 한다. 또한, 이러한 조직은 각 담당자가 맡은 역할과 책임에 대해 충분히 인식하고 관련 역량을 갖춘 인력으로 구성할 필요가 있다.
- 단, 가능하다면 인공지능 거버넌스를 위한 조직은 외부 전문가(예: 심리학자, 데이터 과학자, 행정 전문가 등)를 포함하여 구성할 필요가 있다. 외부 전문가들은 내부 조직에서 발생할 수 있는 편향된 시각을 보완하고 집단 사고groupthink 등의 문제를 극복하는 데 도움을 주기 때문이다.

02-2a 인공지능 거버넌스를 위한 조직을 구성하였는가?

Yes ☐ No ☐ N/A ☐

- 조직의 윤리 원칙 수립 후 이를 실행할 수 있도록 관리하는 것이 인공지능 거버넌스 체계의 목표이다. 즉, 내부 규정을 마련하고 이를 준수하는지 확인할 필요가 있다.
- 신뢰할 수 있는 인공지능trustworthy AI을 위해서 인공지능 거버넌스 체계는 정기적으로 인공지능 관련 사고 및 이슈 사례 리뷰, 원칙 및 규정 수립, 잠재적 문제에 대한 계획 및 대응책 마련을 수행해야 한다.
- ALTAI에서는 인공지능 윤리와 관련된 문제에 대해 대비할 수 있도록 인공지능 거버넌스 체계를 구축하는 것을 고려하길 권고한다.

02-2b 인공지능 거버넌스를 위한 조직은 충분히 훈련된 인력으로 구성하였는가?

Yes □ No □ N/A □

- 인공지능 거버넌스 담당 조직은 자신이 맡은 역할과 책임에 대해 충분히 인식한 인력으로 구성해야 한다. 이들은 인공지능 생명주기에 걸친 모든 프로세스의 중심적인 역할로써, 담당자가 이를 충분히 인식한 후 책임지고 관리해야 인공지능 시스템의 신뢰성을 확보할 수 있기 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 전문지식을 갖춰야 하며, 이를 적절히 해석하여 조직 업무에 적용하기 위한 기술력과 타 업무 담당자와의 의사소통 역량이 필요하다. 또한, 정의된 규정을 실행하고 관리하기 위해 각 담당자에게 관련 교육을 제공하여 충분히 훈련해야 한다.
- 인공지능 윤리 및 신뢰성에 대한 가이드라인은 본 개발안내서 15쪽에서 소개하고 있는 [공공 및 사회분야 주요국 공공 및 사회분야 인공지능 신뢰성 관련 정책 동향]을 참고한다.

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

Yes □ No □ N/A □

해당여부 판단

02-1에 따라 인공지능 거버넌스에 대한 지침 및 규정을 마련한 경우 본 항목을 고려하여 만족여부를 판단하십시오.

- 인공지능 거버넌스 체계를 운영하는 주체는 운영 결과에 대한 책임을 져야 하고, 이 책임은 위임할 수 없다. 따라서 인공지능 거버넌스 운영 담당자는 조직이 내부 지침 및 규정을 준수하는지에 대해 감독해야 한다.
- ISO/IEC 38507:2022 - Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations에서 인공지능 거버넌스 체계는 인공지능 시스템에서 발생할 수 있는 위험에 따라 인공지능 시스템의 설계 및 사용에 대한 감독을 수행해야 한다고 언급하고 있다. 즉, 인공지능 거버넌스 체계를 통해 수립한 내부 규정을 조직이 적절히 이행하고 있는지 감독해야 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 감독해야 한다. 또한, 신뢰성 있는 인공지능 시스템을 목표로 적절히 관리 및 통제됨을 관련 이해관계자에게 입증해야 한다.
- 특히, 인공지능 시스템 위험관리와 관련된 내부 규정을 이행하는지 감독함으로써 인공지능 시스템의 잠재적 위험으로부터 조직 및 이해관계자를 보호하고 조직의 역량을 향상할 수 있다.
- 따라서 인공지능 거버넌스 체계에서 감독을 담당하는 조직은 인공지능 시스템에 대한 이해를 바탕으로 역할에 대한 책임 및 권한을 명확히 인지하여 인공지능 시스템 생명주기에 걸쳐 모든 규정이 이행되는지 감독해야 한다.

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

Yes No N/A
☐ ☐ ☐

해당여부 판단

공공·사회 분야 인공지능 시스템의 신규/기존 서비스 범위를 분석하고, 검증항목으로 만족 여부를 판단하십시오.

- 무분별한 공공 인공지능 시스템 개발이 범람할 경우, 서비스 사용자에게 혼란을 가중할 뿐만 아니라 시스템 개발 및 유지보수에 불필요한 예산 사용을 초래한다.
- 신규 계획 중인 공공·사회 분야의 인공지능 시스템이 기존에 운영 중인 시스템과 활용 대상 및 역할 측면에서 유사한지 고려하고, 기존 시스템의 개선 및 통·폐합을 통한 구현이 가능한지 분석한 결과를 기반으로 시스템을 계획 및 설계해야 한다.

참고 인공지능을 활용한 공공부문 서비스의 유형[14]

- 정부 부처 및 산하기관에서 활용하는 인공지능 서비스의 범위는 인공지능 활용 대상과 인공지능의 역할에 따라 다음과 같이 분류된다.

		인공지능 활용 대상	
		정부 내부	민간 지원
인공지능의 역할	증강	정책지능	공공지능
	자동화	정부봇	지능형 서비스

✓ 정책지능: 정부 정책의 결정 과정에 예측, 시뮬레이션 등 인공지능의 분석 결과 활용
✓ 공공지능: 국민의 주권자 역할을 강화하기 위해 정치적 의사결정 지원
✓ 정부봇: 일상적·반복적 업무를 봇을 활용하여 자동화함으로써 공무원의 생산성 강화
✓ 지능형 서비스: 대국민 정부 서비스의 기능 강화 및 자동화

02-4a 이용 빈도가 낮은 타 시스템의 개선 및 통·폐합을 통해 구현 가능한지 분석하였는가?

Yes ☐ No ☐ N/A ☐

- 신규 공공·사회 분야의 인공지능 시스템을 구현할 때, 기존에 운영 중이지만 이용 빈도가 낮은 시스템의 개선 및 통폐합 가능성을 먼저 분석해야 한다. 이는 전체 운영 중인 시스템의 총량이 불필요하게 증가하지 않도록 하여 다른 공공·사회 분야 인공지능 시스템의 품질에 직간접적인 영향을 끼치지 않게 하기 위함이다.
- 기존의 인공지능 시스템 개선 및 통폐합 과정에는 주요 이해관계자의 의견 교류가 필히 수반되어야 하며, 객관적인 기준, 근거, 검증에 기반하여 추진함으로써 서비스의 공공성을 해치지 않아야 한다.

안전성 투명성

요구사항 03 인공지능 시스템의 신뢰성 테스트 계획 수립

대표행위자 | 품질 관리자 협력 대상 | 시스템 기획자 시스템 엔지니어 비즈니스 결정권자

- 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.

- 특히 공공·사회 분야의 서비스는 제3자 평가를 위한 감사 가능성auditability을 제공해야 한다. 공공 인공지능 시스템이 내외부 전문가에 의해 객관적으로 평가될 수 있도록 충분한 테스트 계획을 수립한다.

03-1 인공지능 시스템의 특성을 고려한 테스트 환경을 설계하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

인공지능 시스템의 위험 분석 결과, 사고 발생 가능성 및 오동작의 파급력이 클 것으로 예상되면 본 항목을 고려하여 만족 여부를 판단하십시오.

- 인공지능 시스템은 그 복잡도나 위험도에 따라 가상테스트 및 실환경 테스트를 고려해야 한다.
- 유네스코의 인공지능 윤리 권고에서는 인권에 대한 잠재적 위협 가능성이 있다고 식별된 인공지능 시스템의 경우 출시 전 이해관계자들에 의해 윤리 영향 평가의 일환으로 광범위한 테스트를 거쳐야 하며, 필요하다면 실제 상황과 동일한 조건에서 테스트를 진행하여야 한다고 권고한다.
- 정확한 테스트를 위해서는 실환경 테스트를 수행하는 것이 적절하지만, 테스트는 합리적인 시간 및 비용 범위 내에서 수행되어야 하므로 운영 조건이 매우 복잡한 시스템이라면 실환경 테스트가 적절하지 않을 수 있다. 또한, 인간과 물리적으로 상호작용하는 인공지능에 실환경 테스트를 적용한다면 위험한 상황이 발생할 우려가 있는데, 이 경우 가상테스트를 수행하여야 한다.

- 따라서, 시스템 특성을 고려하여 적절한 테스트 환경을 결정한 후 테스트 환경을 설계하는 것이 필요하다. 테스트 환경 설계 시 고려해야 할 사항의 예시는 다음과 같다.
  - ✓ 인공지능 시스템의 운영환경이 복잡하고 끊임없이 변화하는가?
  - ✓ 인권에 대한 잠재적 위협 가능성이 우려되는 시스템인가?
  - ✓ 테스트는 합리적인 시간 및 비용 범위 내에서 수행 가능한가?
  - ✓ 실환경 테스트 시 환경의 개체(예: 차량, 건물, 동물, 인간)에 손상을 주는가?

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

Yes ☐ No ☐ N/A ☐

- 운영 환경의 제약, 기능의 다양성, 성능 저하 요소 등 매개변수가 많은 인공지능 시스템이라면 테스트 스위트test suite 수가 거의 무한해질 수 있다. 이 경우, 매개변수의 조합을 통해 테스트 스위트 수를 줄일 수 있는 조합 테스팅combination testing 기법의 하나인 페어와이즈 기법의 활용을 고려해야 한다.
- 반면에, 예외적인 상황edge case에 대한 시나리오의 생성이 어렵거나, 테스트 시 환경의 개체에 손상을 줄 위험이 있는 시나리오가 포함된 인공지능 시스템은 가상테스트 환경을 고려해야 한다.
- 그 외, 테스트 환경을 마련하기 어려워 실환경 테스트를 수행할 수 없는 경우(예: 원자력 사고 현장을 탐사하는 로봇)에는 가상테스트가 채택될 수 있다.

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

Yes ☐ No ☐ N/A ☐

- 일부 도메인은 오픈소스로 공개된 시뮬레이터가 있어, 개발할 인공지능 시스템에 적합하다면 이를 활용할 수 있다. ISO/IEC TR 29119-11:2020에서는 아래와 같은 시뮬레이터의 예시를 제공하고 있다.
  - ✓ 게임 엔진 기반의 이동형 로봇 시뮬레이터: MORSEModular OpenRobots Simulation Engine 프로젝트
  - ✓ 홈 서비스 로봇 학습 시뮬레이터: Facebook의 AI Habitat
  - ✓ 물리적 현상을 나타내는 시뮬레이터: Google DeepMind의 MuJoCo
  - ✓ 가상 환경 드론 비행 시스템 시뮬레이터: Microsoft의 Aerial Informatics, Robotics Platform
- 재사용 가능한 시뮬레이터가 없다면 시뮬레이터의 구축이 필요하며, 계획 및 설계 단계에서 시뮬레이터 구축을 위한 추가 자원의 규모(예: 인력, 비용, 시간)를 고려해야 한다.
- 시뮬레이터는 운영환경에 대한 대표성이 있어야 한다. 예를 들어, 드론의 가상테스트를 수행할 경우, 드론 운행 규정에서 정의한 비행불가 지역들의 대표성이 요구된다.[15]

항목	내용
비행 금지 시간	(1) 야간 비행 (야간: 일몰 후부터 일출 전까지)
비행 금지 장소	(1) 비행장으로부터 반경 9.3km 이내인 곳 → "관제권"으로 불리는 곳으로 이착륙하는 항공기와 충돌할 위험이 있음
	(2) 비행금지구역 (휴전선 인근, 서울 도심 상공 일부 → 국방, 보안상의 이유로 비행이 금지된 곳)
	(3) 고도 150m 이상 → 항공기의 바행항로가 설치된 공역
	(4) 인구 밀집 지역 또는 사람이 많이 모인 곳의 상공 (예: 스포츠 경기장, 각종 페스티벌 장소처럼 사람이 많이 모인 곳) → 기체가 떨어지면 인명 피해 위험이 높음
비행 중 금지 행위 및 비행 금지 경우	(1) 비행 중 낙하물 투하 금지, 조종자 음주 상태에서 비행 금지
	(2) 조종자가 육안으로 장치를 직접 볼 수 없을 때 비행 금지 (예: 안개, 황사 등으로 시야가 좋지 않은 경우, 멀리 날리는 경우)

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

Yes ☐ No ☐ N/A ☐

해당여부 판단

사용자에게 인공지능 시스템의 추론 결과에 대한 설명이 필요한 경우, 특히 대상 사용자의 특성이 다양한 경우에 본 항목을 고려하여 만족 여부를 판단하십시오.

- 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 추론 결과에 대한 설명이 필요한 시스템이라면, 시스템 출력을 확인하는 대상 사용자에 따라 출력에 대한 도출 방법을 이해하는 정도인 설명가능성에 대한 평가 기준이 달라질 수 있다. 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 관한 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 오라클 문제의 극복이 필요한 인공지능 시스템이라면, 시스템의 기대 출력을 결정하기 위해 해당 도메인의 내외부 전문가로 구성된 협의체를 구성하여야 한다. 이때 기대 출력을 결정하기 위해 여러 전문가가 동의하는 데 시간이 걸릴 수 있음을 인지하여야 한다.
- 협의체 전문가들은 하나의 입력에 대해 각자 다른 기대 출력을 예상할 수도 있다. 그러므로 협의체 운영 전 전문가 합의를 위한 승인 기준을 미리 결정해두어야 한다. 예를 들어, 특정 기대 출력에 대한 전문가 3인 중 2인 이상이 동의할 때 승인하는 등의 방법이 있다.

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

Yes ☐ No ☐ N/A ☐

- 인공지능시스템 출력에 대한 설명이 필요한 시스템의 경우, 시스템의 설명가능성과 해석가능성을 테스트하기 위해서는 인공지능 시스템의 대상 사용자가 시스템의 출력과 작동 방식을 얼마나 쉽게 이해하는지 확인하여야 한다.
- 따라서 사용자 평가단을 구성하여 설명을 어떤 난이도로 제공할지 결정하고, 이를 모델 및 시스템 구현 시 반영해야 한다. 이를 위해, 계획 및 설계 단계에서 대상 사용자를 명확히 정의한 후 사용자 평가단을 구성해야 한다.
- 사용자 평가단의 평가 결과에 따라 테스트의 통과 및 실패 여부를 결정할 기준을 마련하는 것이 필요하다. 예를 들어, 평균 점수가 일정 점수 이상일 때 통과를 결정하는 등의 정량적 기준 마련이나, 평균 점수 계산 시 절사평균의 활용 여부 등의 산출 기준 마련 등이 있다.

책임성 투명성

요구사항 04 데이터의 활용을 위한 상세 정보 제공

대표행위자 | 데이터 과학자 협력 대상 | 데이터 공급자 도메인 전문가 인공지능 모델 개발자

- 정부에서는 범정부 데이터 통합관리 플랫폼(AI Hub, 공공데이터 포털 등)을 구축해 모든 공공기관의 데이터를 표준화하고 데이터의 소재를 파악해 연관관계 등을 분석하고 있다.
- 그런데 실제 공공·사회 분야의 인공지능 시스템 개발 시, 공공에서 제공되는 오픈 데이터셋을 활용할 뿐만 아니라 직접 수집 방식도 병행하고 있다. 이때 직접 수집한 데이터는 수집 방식과 수집 기관에 따라 다르다.
- 이에 정부의 공공 데이터 표준화 및 관련 인공지능 서비스 구축에 어려움이 없도록 직접 수집한 데이터에 대한 상세 정보를 제공한다.

04-1 데이터의 명확한 이해와 활용을 지원하는 상세한 정보를 제공하는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

공공·사회 분야 인공지능 알고리즘 또는 모델 개발을 위해 데이터셋을 직접 구축하거나 공개 데이터셋에 향후 추가 데이터 수집의 가능성이 있는 경우에 본 항목을 고려하여 만족 여부를 판단하십시오.

- 데이터의 수집 및 정제 시, 원시 데이터raw data와 정제 후 데이터의 정보를 제공하여 데이터에 대한 이해를 도모하고, 학습 데이터와 메타데이터metadata를 정의하여 추가 데이터 수집에 필요한 정보를 제공한다.
- 특히 공공·사회 분야의 인공지능 신뢰성 요건에 크게 영향을 미치는 보호변수protective attribute에 대해서는 반드시 설명해야 한다. 이에 라벨링 작업 시, 작업자에게 전문 도구의 활용 방법과 라벨링의 유의 사항을 가이드해 준다.

04-1a 정제 전과 후의 데이터 특성을 설명하였는가?

Yes ☐ No ☐ N/A ☐

- 데이터 정제 작업은 라벨링 작업 전에 학습 데이터 구축을 위해 데이터를 선별하고 처리하는 단계이다. 사용자가 정제 과정을 거친 데이터만 사용한다면 원시 데이터[raw data]의 특성을 정확하게 파악할 수 없다. 따라서 정제를 위한 관련 정보와 정제 전/후의 데이터 특성을 설명한다면 원시·정제 데이터에 대한 이해도를 높이고 좀 더 신뢰성 있는 인공지능 모델을 개발할 수 있다.
- 공개 데이터셋을 활용할 때는 원시 데이터의 특성을 파악할 수 있는 정보를 함께 제공하는지를 확인해야 한다. 데이터셋과 함께 제공되는 자료를 참고하여 원시 데이터의 특성을 파악하고, 향후 추가 데이터 수집의 가능성을 고려하여 해당 데이터 정제 전/후의 특성을 설명하는 자료를 기록하도록 한다.
- 원시 데이터를 직접 수집했을 때는 기업이 자체적으로 관련 문서를 마련하며, 기관의 사용 목적에 기반한 데이터 수집 목적, 수집 대상, 수집 환경, 주제 등의 정보를 제공하여 원시 데이터에 대한 이해를 돕도록 한다. 다음은 정제 전의 원시 데이터의 특성으로 설명할 수 있는 항목의 예시이다.
  - ✓ 데이터 수집 목적: 한국어 대화 모델링을 통한 민원의 정책적 활용 등
  - ✓ 주제: 개인 및 관계, 건강 및 의료, 미용, 식생활, 주거 및 생활, 일, 시사, 경제, 교육, 출산 및 육아, 여가 생활, 쇼핑, 행사 등
  - ✓ 수집 대상: 국적, 성별, 연령대, 거주지역 등
  - ✓ 수집 환경: 수집 방법, 수집 날짜, 시각, 참여자 수 등
- 수집한 전체 원시 데이터로 학습 데이터셋을 구축하기 위해 원시 데이터에서 불필요한 내용, 개인정보 등을 확인하고 정제한다. 다음은 정제한 학습 데이터의 특성으로 설명할 수 있는 항목의 예시이다.
  - ✓ 데이터 선별 및 제외 설명 항목: 부적절한 내용, 부자연스러운 내용, 불필요한 내용(이모티콘, 감정 표현 등) 등 제외
  - ✓ 데이터 처리 설명 항목: 공공기관 데이터 표준 처리, 개인정보 비식별화 처리, 이모티콘 처리 등
  - ✓ 통계적 설명 항목: 주제별 데이터 수, 수집 대상별 데이터 수 등
  - ✓ 정제 정보: 업데이트 및 정제 주기, 롤백 정보 등
- 다만, 위에서 정의한 '불필요한' 내용의 예시는 특정 주제에 대한 긍정적/부정적 반응을 확인하는 관점에서는 '필요한' 특성이 될 수 있다. 이 점을 고려하여 목적에 맞게 정제 기준을 수립하는 것이 중요하다.

04-1b 학습 데이터와 메타데이터를 구분하고 각 명세자료를 확보하였는가?

Yes ☐ No ☐ N/A ☐

- 본 개발 안내서의 발간일을 기준으로 공공·사회 분야의 인공지능 활용 사례는 다음과 같으며, 이와 같은 서비스 개발 시 필요한 학습 데이터와 메타데이터의 명세 항목을 식별한다.
- 현재 공공·사회 분야의 학습 데이터와 메타데이터의 명세 내용은 일반 분야와 유사하다. 다만, 공공·사회 분야의 인공지능 서비스에 도메인이 반영되면 해당 분야의 개발 안내서에서 권고하는 내용을 참조하여 추가적인 명세 항목을 제공할 수 있다.

참고 공공·사회 분야의 인공지능 서비스 및 메타 데이터 예시

국가	서비스명	개요	데이터타입	메타데이터 명세
한국	온통서산	• 민원 업무 자동화 서비스 • SNS 민원창구를 통해 생활민원 등 간편하게 실시간 처리	텍스트	• 민원 분야 • 술어(요청) 구분
한국	국민비서 서비스	• 국민에게 쉽고 편리한 서비스 제공 • 사용자가 생활 속에서 언제든지 필요한 행정정보를 확인	텍스트	• 문의 유형 • 문의자 ID
한국	아산병원 병상 배정 업무 자동화 프로그램	• 인공지능을 기반으로 병원의 병상 배정 업무 자동화	텍스트	• 진료과 • 병동 • 병실(1인실, 2인실 등) • 담당 의사 • 배정 시간
미국	Apple Siri, Google Now	• 음성인식 기반의 인공지능 개인 비서 서비스	오디오	• 언어 • 길이 • 화자의 성별, 연령, 지역 등
한국	셀비 노트	• AI 음성 기록 솔루션	오디오	• 화자 수 • 주제
한국	119 신고 접수 시스템	• 음성인식 기술로 신고자의 통화 내용에서 주요 키워드를 효과적으로 추출하고 신고 접수	텍스트 오디오	• 재난 위치 • 상황 • 증상
한국	제안요청서 법규 준수 진단 모델	• 조달 요청 내역 현황을 분석하고 조달 요청 RFP 내 '필수 검토 용어 사전'을 구축하여 사전 검토	텍스트	• 계약 유형 • 평가 방식 • 입찰 방법 • 공동계약, 지역제한, 분할납품 여부 등

04-1c 보호변수의 선정 이유 및 반영 여부를 설명하였는가?

Yes No N/A
☐ ☐ ☐

- 공공·사회 분야에서는 다양성 존중이 특히 중요한 신뢰성 요소인 만큼 사회적으로 부정적인 영향을 미치는 특성을 고려해야 한다.
- 사회적인 관점에서 대표할 수 있는 민감한 특성은 인종, 국적, 성별, 나이, 거주지역, 종교 등이 있다. 이러한 내용은 정책 마련, 복지 제도 지원 등에서 차별로 인한 윤리적 문제를 일으킬 수 있다.
- 따라서 인공지능 서비스의 목적에 따라 보호변수를 선정하고, 선정 이유와 기준에 대한 설명을 제공한다.
- 데이터 편향을 확인하기 위한 오픈소스 분석 도구는 IBM AI Fairness 360, Google Fairness –indicator, Microsoft Fairlearn, ,Linked in LIFT, Open Source Aequitas 등이 있다.

04-1d 라벨링 작업자를 위해 교육을 시행하고 작업 가이드 문서를 마련하였는가?

Yes No N/A
☐ ☐ ☐

- 데이터 라벨링 작업은 인공지능 모델을 학습하기 위한 원시 데이터의 주석(정답) 작업에 해당하며, 다수의 작업자를 통해 수행된다. 라벨링 작업은 데이터셋의 품질 확보뿐만 아니라 모델 성능에 직접적인 영향을 줄 수 있어 작업자의 교육 및 상세한 작업 가이드 문서를 마련하는 것이 중요하다.
- 라벨링 작업은 데이터 종류에 따라 작업 대상, 범위, 상세 절차 및 라벨링 도구 등이 달라질 수 있다. 라벨링 작업 절차에 대한 가이드와 함께, 작업 절차에 따라 작업자를 대상으로 한 교육과 가이드 문서가 확보되어야 한다.

참고 민원 업무 자동화 서비스의 데이터 라벨링 작업 가이드 예시[16]

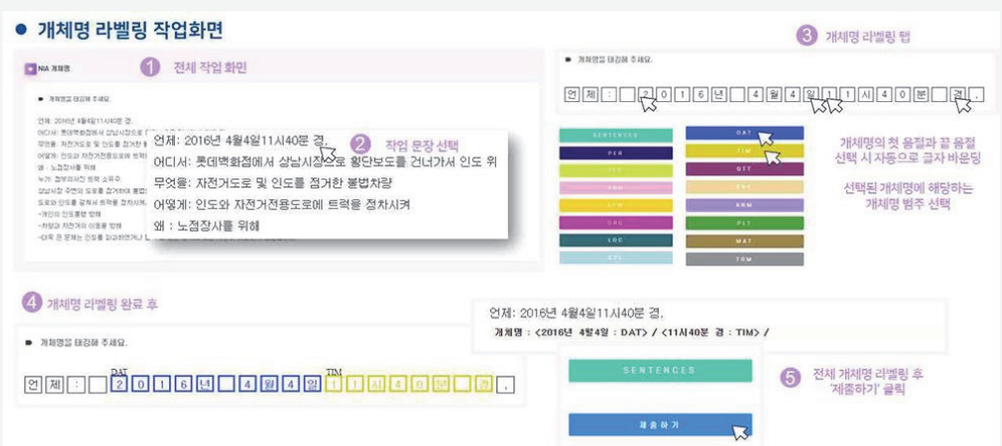
라벨링 작업 절차

절차	내용
1. 전처리데이터(입력)	수집 및 정제 후 전처리된 원천 데이터를 주관사 저작 도구로 제공
2. 가이드라인 확인	라벨링 기준으로 활용되는 문서 요약 가이드라인 확인
3. 정제된 문서의 사전 정보 확인	수집 및 정제, 전처리 과정에서 생성된 문서 정보 확인
4. 정제된 질의 정독	원천 데이터의 질의와 원시 데이터의 제목 및 응답 확인
5. 대표 민원 선정	질의 데이터에서 대표 민원 선정
6. 오타 수정 및 개인정보 비식별	질의의 오타를 수정하고 개인정보 비식별화
7. 개체명 라벨링	질의 문장의 개체명(인물, 기관, 위치, 등) 태깅
8. 키워드 라벨링	질의 문장의 핵심 키워드 선정
9. 의도 라벨링	질의 문장의 의도 선정
10. 유사질의 생성	원천 데이터 질의의 유사 질의 문장 생성
11. 담당 부서의 정보 부착	질의 문장의 처리 부서 부착
12. 관련 법률 확인	응답 작성 시 참고해야 하는 관련 법률 확인

라벨링 기준

라벨링 대상	라벨링 범위	클래스 분류 기준
민원 질의 데이터	1. 개체명 2. 키워드 3, 의도 4. 유사 질의 생성 5. 처리 부서	- 개체명: 국립국어원에서 정의한 15개 개체명 태깅 - 의도: 창원시 민원 분류 18개 카테고리, 대표 술어 8개

구분	내용
개체명 라벨링	대원동 공용주차장 이용 문의 예) 대원동: location, 공용주차장: public_park
키워드 라벨링	지방세를 내려고 하는데 어디서 냅니까? 예) 지방세
의도 분류 라벨링	대원동 공용주차장 이용 문의 예) 의도: 교통 〉 공용주차장 이용
유사 질의 라벨링	지방세 납부 어떻게 해요? 예) 지방세 내려고요
담당 부서 라벨링	대원동 공용주차장 이용 문의 예) 건설교통과
관련 법률 라벨링	공용주차장에서 파손 책임 문의 예) 제2조 1항



라벨링 도구 활용 방안

검수 절차

04-2 데이터의 출처는 기록 및 관리되고 있는가?

Yes No N/A
☐ ☐ ☐

해당여부 판단

공공·사회 분야의 인공지능 알고리즘 또는 모델 개발에 활용하는 데이터셋을 직접 구축할지 오픈소스 데이터셋을 활용할지에 따라 본 항목을 고려하여 만족 여부를 판단하십시오.

- 공공·사회 분야의 인공지능에서 활용하는 데이터셋의 품질로 인해 발생할 수 있는 인공지능의 신뢰성 문제는 다양하다. 이러한 문제가 발생했을 때 원인을 파악하고 대응하기 위해 데이터의 출처와 수집 시점, 데이터셋 버전 등의 정보를 관리해야 한다.
- 공공기관에서 활용하는 인공지능 서비스는 직접 데이터를 수집하는 방안 외에 타 기관으로부터 데이터를 받아 사용하는 경우도 있는데, 이 경우에도 공공데이터로 취급되기 때문에 '공공데이터 관리지침(행정안전부고시 제2021-70호, 2021.10. 26.)'에 따라 관리해야 한다.

04-2a 신뢰할 수 있는 출처로부터 제공되는 데이터셋을 사용하였는가?

Yes No N/A
☐ ☐ ☐

- 학습 데이터를 직접 생산한다면, 데이터 획득 시 수집 출처(예: 크라우드 워크, 아웃소싱 기관)의 객관성 확보가 필요하다. 또한, 수집 대상이 되는 데이터의 출처를 살펴 향후 소유권이나 사용권 이슈가 발생할 수 있는지 선제적으로 확인해야 한다.
- 오픈소스 데이터셋을 사용할 때는 해당 데이터셋 품질이 신뢰할만한 수준인지 고려할 필요가 있다. 고려사항으로는 데이터가 법적으로 문제는 없는지, 데이터셋 규모가 학습하기에 충분한지, 데이터셋에 대한 논의나 업데이트가 활발하게 이루어지고 있는지 등을 고려해야 한다.
- TTA 정보통신단체표준 TTAK.KO-10.1339에서는 지도학습 계열의 인공지능 기술에 활용되는 데이터 획득 시, 출처의 신뢰성을 확보하기 위해 고려해야 할 내용을 정리하고 있으며, 공공데이터 활용 및 직접 수집 방식에 따라 해당 내용을 참조한다.

참고 국내외 공공데이터 제공 포털

- 국내에서는 공공기관에서 보유하고 있는 다양한 데이터를 공공데이터의 제공 및 이용 활성화에 관한 법률(제17344호)에 따라 개방하고 있으며, 해외 정부기관에서도 공공데이터를 제공하는 플랫폼을 운영하고 있다.
- 공공기관에서 운영되는 데이터 제공 포털의 경우, 사용자가 데이터를 활용할 때 신뢰할 수 있는지를 판단할 수 있도록 데이터를 공개·활용하는 데 필요한 정보의 기준을 정하고, 이에 맞춰 데이터와 관련 정보를 함께 제공한다.

포털	링크
공공데이터포털	https://www.data.go.kr/
AI Hub	https://www.aihub.or.kr/
일본통계국 빅데이터 포털	http://www.stat.go.jp/
미국정부 공개자료 공공데이터 포털	https://www.data.gov/
영국 국립 데이터센터	https://data.gov.uk/
EU정보플랫폼	https://www.europeandataportal.eu/
중국국립데이터센터	http://data.stats.gov.cn/
홍콩정부 데이터센터	https://data.gov.hk/ja/
대만정부 정보공개 플랫폼	https://data.gov.tw/

- 이 외에 국제조직에서도 다양한 통계 데이터 등 공공데이터를 제공한다.

포털	링크
경제협력개발기구(OECD) 데이터베이스	https://data.oecd.org/
세계은행 공개정보 포털	https://data.worldbank.org/
빅데이터 포털 – 세계보건기구	http://apps.who.int/gho/data/node.home

04-2b 오픈소스 데이터셋을 활용하는 경우, 출처를 명시하였는가?

Yes No N/A
☐ ☐ ☐

- 오픈소스 데이터셋을 활용하여 학습 기반 인공지능 모델을 구축할 경우, 과거·현재·미래 시점에 발생할 수 있는 데이터 편향의 원인 파악을 위해 확보된 데이터의 명확한 출처와 관련 정보를 명시하여 관리해야 한다.
- 특히, 공공기관에서 활용하는 인공지능 서비스는, 직접 수집하는 데이터셋이라도 공공기관이 법령 등에서 정하는 목적을 위해 생성 또는 취득하여 관리하는 DB, 전자화된 자료 또는 정보로서, 공공데이터에 해당하기 때문에 출처 및 관련 정보를 명시하여 관리해야 한다.
- 만약 다른 기관의 데이터를 활용한다면 타 기관으로부터 받아 활용 중인 연계 데이터의 정합성 확보를 위해 연계 데이터의 목록을 작성하고 관리해야 한다.

참고 연계 데이터 관리 항목 및 설명 예시[공공데이터 관리 지침(행정자치부 고시 제2016-42호, 2016.11. 22. 개정)]

연계 데이터의 관리 항목		설명
연계 정보 정의	연계 정보 구분	• 연계 정보 제공 시 “제공”, 활용 시 “활용”으로 정의
	연계 정보명	• 연계 정보가 무엇을 의미하는지를 이해할 수 있는 수준에서 정의
	연계 주기	• 연계 정보의 연계 주기 정의(예: 실시간, 매주, 매월 등)
연계 항목 정의	연계 항목명	• 연계 정보를 구성하는 세부 연계 항목의 명칭
	연계 항목 설명	• 연계 정보를 구성하는 세부 연계 항목의 정의
	연계 항목 데이터 타입	• 연계 항목의 데이터 타입 정의
	연계 항목 데이터 길이	• 정의된 연계 항목별 데이터 길이의 정의
연계 항목 출처	연계 데이터베이스명	• 연계 항목을 제공 또는 활용하는 데이터베이스명
	연계 테이블명	• 연계 항목을 제공 또는 활용하는 테이블명
	연계 컬럼명	• 연계 항목을 제공 또는 활용하는 컬럼명
오너십 정의	제공 기관	• 연계 정보를 제공하는 기관 정보(기관명, 담당자, 연락처 등)
	활용 이관	• 연계 정보를 활용하는 기관 정보(기관명, 담당자, 연락처 등)
비고		• 연계 정보 관련 행정·기술적 사항(연계 형식, 연계 방법 등)

안전성

요구사항 05 데이터 강건성 확보를 위한 이상abnormal 데이터 점검

대표행위자 | 데이터 과학자 협력 대상 | 데이터 공급자 인공지능 모델 개발자

- 인공지능 모델의 학습에 활용되는 데이터는 이상값outlier, 중독 및 회피 등에 영향을 받지 않아야 하며, 이에 대한 점검 및 방어 기법의 적용을 통해 강건성을 확보한다.
- 특히 공공·사회 분야 인공지능 대민 서비스가 사용자의 입력을 통해 피드백받아 학습할 경우에는, 악의적인 사용자로 인한 공격에 매우 취약하게 되므로 이러한 점을 고려한 데이터 점검 및 방어 기법의 적용이 필요하다.

05-1 이상 데이터의 식별 및 정상 여부를 점검하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

인공지능 모델의 학습 데이터를 직접 구축하거나 사전 학습된 인공지능 모델의 학습 데이터에 대한 정상/오류 여부가 명확하게 확인되지 않았다면 본 항목을 고려하여 만족 여부를 판단하십시오.

- 이상 데이터는 학습용 데이터를 구성하는 데이터셋의 수집 및 가공 과정에서 발생할 수 있는 다양한 오류error와 일반적인 데이터의 범위에서 크게 벗어난 데이터 이상값을 포괄한다. 이를 점검해 대처하지 않으면 인공지능 모델의 성능 및 강건성을 충분히 확보할 수가 없다.
- 특히 공공·사회 분야의 다양성 존중 관점에서 신뢰성을 확보하기 위해서는 사전에 학습 데이터 내의 편향을 점검하고 대처해야 한다. 사전 학습 및 검증된 모델을 활용한다면 추가 학습의 이상 데이터에 따른 영향이 줄어들고 재검증하는 데 필요한 노력을 최소화할 수 있다.
- 비정형 데이터unstructured data를 학습에 활용한다면, 데이터 전처리 과정에서 이상 데이터를 식별할 수 있는 별도의 기법을 마련해야 한다.

05-1a 전체 학습용 데이터 분포를 시각화하여 발생 가능한 오류들을 확인하였는가?

Yes ☐ No ☐ N/A ☐

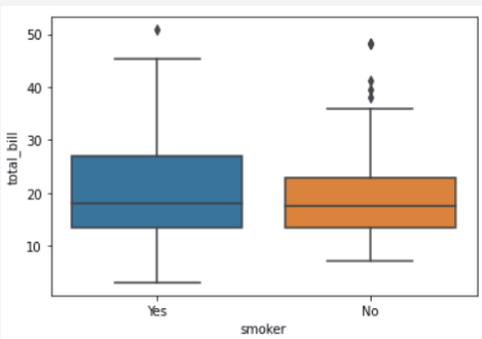
- 데이터 전처리 과정 중 하나인 데이터 정제 단계 이후, 데이터 전체 분포를 시각화하여 추가적인 입력 오류를 확인할 수 있다. 특히, 이러한 데이터 분포 시각화는 인공지능 모델 학습을 위한 데이터 탐구 및 이해에 많은 도움을 준다.

• 특히 편향을 유발하는 특성은 범주형 변수(성별, 종교 등)가 많은데, 공공·사회 분야에서는 이를 식별하고 대응하는 것이 중요한 만큼 이들의 분포를 시각화하여 편향 가능성이 최소화되게 해야 한다. 다음 표는 범주형 변수를 시각화하는 방법의 예시이다.

참고 범주형 변수 시각화 방법[17]

• 데이터를 범주로 나누어 각 범주의 분포를 시각화는 방법으로, 주로 범주형 변수의 분포를 파악할 때 사용한다.

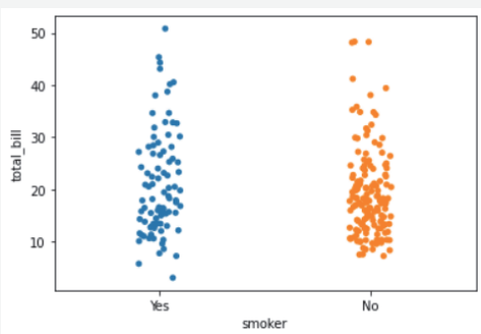
• 시각화 라이브러리 Seaborn은 Matplotlib을 기반으로 다양한 색상 테마와 통계용 차트 등의 기능을 추가한 시각화 패키지로서 다음과 같은 범주형 변수 시각화를 제공한다.



Box Plot

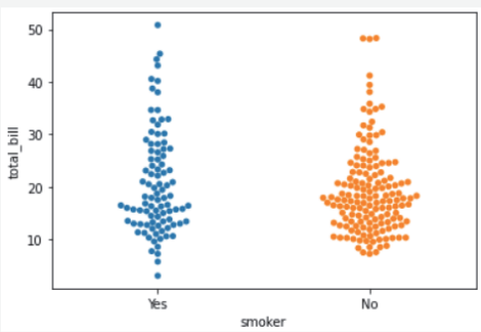
• 최대maximum, 최소minimum, 평균mean, 1 사분위수first quartile, 3 사분위수third quartile를 보기 위한 그래프로서 이상값을 발견하기에 좋다.

• 단일 연속형 변수에 대해 수치를 표시하거나 연속형 변수를 기반으로 서로 다른 범주형 변수를 분석할 수 있다.



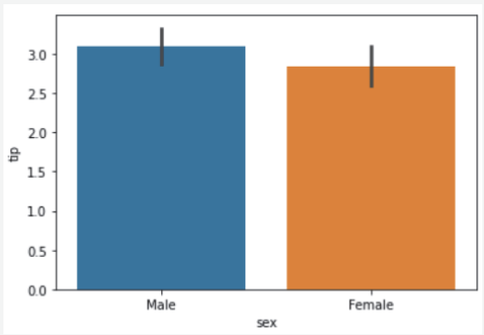
Strip Plot

• 연속형 변수와 범주형 변수 사이의 그래프로서 산점도scatter plot로 표시되며 범주형 변수의 인코딩을 추가로 사용한다.



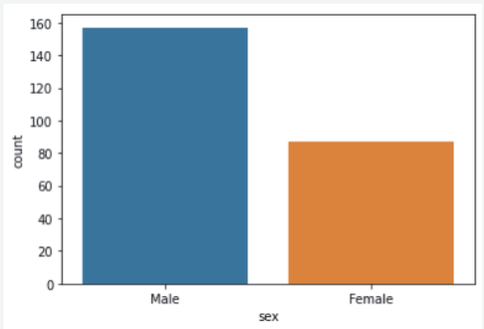
Swarm Plot

• Strip plot과 violin plot의 조합으로서 데이터 포인트 수와 함께 각 데이터의 분포도 제공한다.



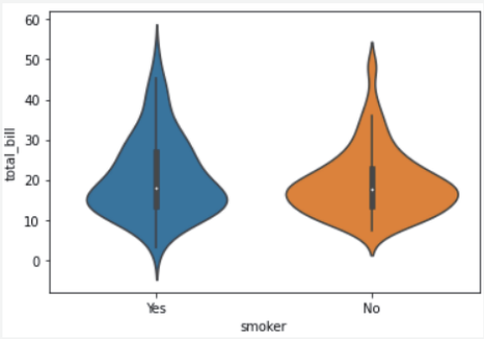
Bar Plot

- 이변량bivariate 분석을 위한 방법으로, x축에는 범주형 변수를, y축에는 연속형 변수를 넣는다.



Count Plot

- 일변량univariate 분석을 위한 방법으로, 범주형 변수의 발생 횟수를 센다.



Violin Plot

- Box Plot과 비슷하며 분포에 대한 보충 정보가 제공된다.

05-1b 학습 데이터 이상값 식별 기법을 적용하였는가?

Yes No N/A
☐ ☐ ☐

- 데이터 전처리 과정에서 중요한 활동 중 하나는 이상값을 식별하고 제거하는 것인데, 일반적으로 통계적인 기법을 사용해 전체 데이터셋의 범위에서 벗어나는 값을 찾아낸다.
- 공공·사회 분야의 데이터를 수집하는 과정에서 집단민원이 발생할 수 있으므로 이와 관련한 이상값을 정의해야 하고, 필요시 다음과 같이 집단민원을 추정하는 방법을 사용한다.

참고 집단민원 추정 방법론 예시[18]

- 이상탐지 활용 전자집단민원 추정 방법론에 관한 탐색적 연구(2019)에 따르면, 집단민원이란 '다수인 관련 민원으로, 내용에 유사성 또는 동질성이 있으며, 반복성과 주기성이 있는 민원', 또는 '1일 내 제기되는 민원 건수가 평상 수준을 넘으며 현저하게 다른 메커니즘으로 생성되는 민원'으로 정의된다.
- 전자민원시스템은 오픈 시스템이므로 복사 및 붙여넣기를 통해 짧은 시간에 동일한 민원을 제기할 수 있다. 이는 여러 사람이 동일한 시간에 동일한 공간에 모이지 않고서도 동일한 목소리를 낼 수 있는 손쉬운 집단민원 제기 수단으로서 전자민원시스템이 이용될 수 있다는 것을 의미한다.
- 이에 해당 연구에서는 는 이상탐지와 내용분석으로 전자집단민원 추정 방법론을 제시하고 있다.

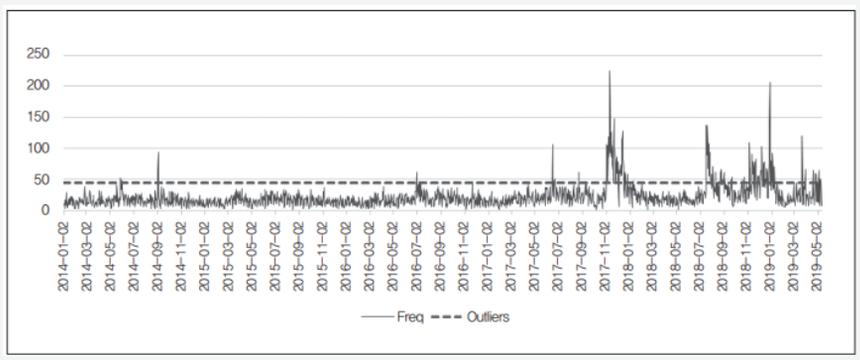
이상탐지 방법론

분류	원리	접근법	기준
불일치 테스트 (Discordancy tests)	• 모수의 확률밀도함수가 특정 분포를 따른다는 귀무가설을 가정한 후, 이상치가 분포에 속하지 않거나 주어진 유의 수준에 해당하지 않는 경우에 가설을 채택하여 이상치로 정의	Parametric	통계 개연성 분포 모델
근접성 기준 (Proximity)	• 관측치의 지역성(또는 근접성)이 희박한 경우에 이상치로 정의	Non-para metric	클러스터링 거리 밀도
라벨링 기법 (Labeling methods)	• 단변량·다변량 데이터univariate or multivariate data의 위치location, 확산spread, 왜곡skewness을 기준으로 이상치를 탐지하는 접근법		지도 학습 준지도 학습 비지도 학습

내용분석 방법론

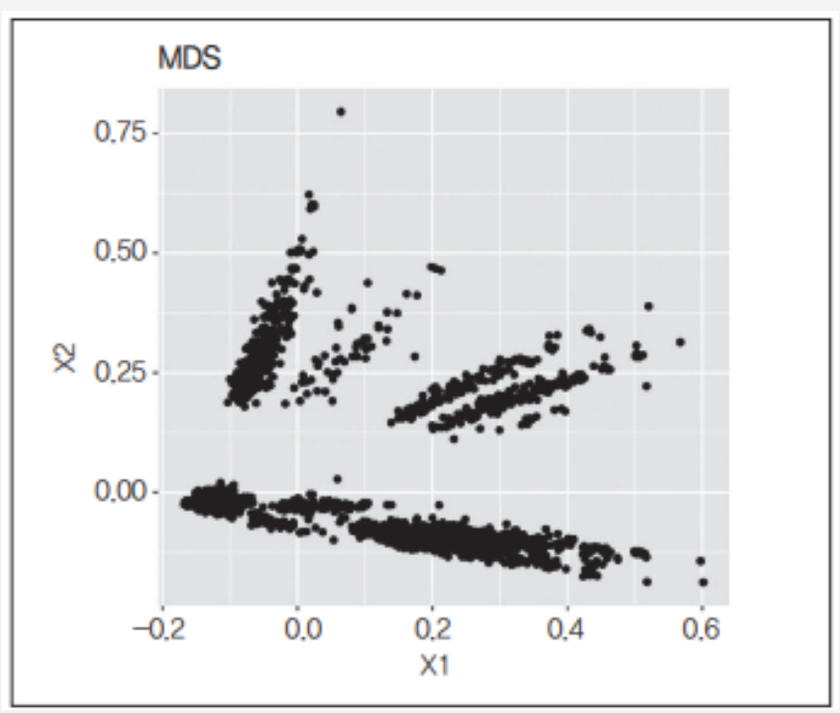
- 전자민원은 시간 데이터와 문자 데이터를 포함하기 때문에 전자집단민원 추정 프로세스는 3단계로 진행한다.

절차	방법	
시간 데이터의 시계열 분석	평상 수준 확인	Augmented Dickey-Fuller for stationarity check
	이상치 탐지	Box plots
문자 데이터의 내용분석	n-gram 모델	한국어 형태소 분석, MPHMinimum Perfect Hashing
	유사성 분석	LDLevenshtein Distance
임베딩	시각화	MDSMulti-Dimensional Scaling



창원시 시민의 소리 전자민원 시계열 이상치(1일 단위)

출처: 이상탐지 활용 전자집단민원 추정 방법론에 관한 탐색적 연구-창원시 시민의 소리 사례를 중심으로



민원 내용의 유사도

출처: 이상탐지 활용 전자집단민원 추정 방법론에 관한 탐색적 연구-창원시 시민의 소리 사례를 중심으로

05-2 데이터 공격에 대한 방어 수단을 강구하였는가?

Yes No N/A
☐ ☐ ☐

해당여부 판단

인공지능 모델의 개발 및 운영 과정에서 데이터 공격을 방어하는 수단을 적용하지 않았다면 본 항목을 고려하여 만족 여부를 판단하십시오.

- 공공·사회 분야의 대민 서비스는 다양한 사용자에게 공개되기 때문에 다른 분야보다 데이터 중독 또는 추출 공격에 취약하다. 또한 인공지능 서비스를 사용하지 않는 사람들도 간접적인 영향을 받을 수 있으므로 공격의 여파도 크다.
- 특히, 사용자의 입력을 통해 모델을 재학습하는 경우에는 인공지능 서비스의 운영 과정에서 의도적으로 학습 데이터를 변질시키거나 모델의 추론 과정에서 데이터를 교란하여 예상과 다른 결과를 출력하도록 하는 공격에 노출될 수 있으므로 이를 대처할 방안을 마련해야 한다.
- 이 외에도 사용자의 입력에 개인정보와 같이 민감한 정보가 포함되었다면 데이터 추출 공격으로 인한 데이터 유출 위험이 있으므로 이에 대처하는 방안을 검토하고 적용해야 한다.

참고 공격 기법 및 예시

공격 기법 분류	공격 기법 내용 및 사례	
데이터 중독 공격 (Poisoning attack)	 채팅봇 테이의 문제 답변[19]	• 인공지능 서비스는 일반적으로 입력 데이터 분포의 변화에 적응하기 위해 모델 배치 후 수집된 새로운 데이터를 사용해 재교육된다. 이때, 공격자는 세심하게 조작된 데이터를 주입하여 서비스의 정상적인 기능을 손상하는 방식으로 학습 데이터를 오염시킬 수 있다. • 데이터 중독 공격의 대표적인 사례로는 2016년 MS에서 개발한 채팅봇 테이(Tay)가 있다. 이 서비스는 사람들로 인해 악의적인 발언을 하도록 학습되어 욕설과 인종차별 발언을 하여 운영 16시간 만에 중단되었다. • 공공·사회 분야에서는 05-1b 에서 설명한 집단민원이 하나의 예시가 될 수 있다. 민원 수집 과정에서 특정 시점에 다수의 의도하지 않은 데이터가 생성될 수 있으며, 다수의 유사하거나 중복된 민원은 모델의 재학습에 영향을 미치게 된다. 이러한 데이터들은 모델이 편향된 결과나 악의적 발언을 출력하도록 학습될 수 있다.

공격 기법 분류	공격 기법 내용 및 사례
회피 공격 (Evasion attack)	• 공격자는 학습 모델이 입력을 올바르게 식별할 수 없도록 기존의 입력 데이터에 대해 미묘한 차이의 노이즈를 추가하여 조작된 입력 데이터를 생성한다. 이러한 변화는 사람의 눈에 잘 띄지 않지만 심층학습deep learning 모델의 출력에 큰 영향을 미친다. • 공공·사회 분야에서 악용될 수 있는 예시로는 특정 복지의 혜택 여부를 평가하는 서비스에서 데이터에 미묘한 변화를 주어 원하는 결과를 출력하도록 유도하는 경우 등이 있다.
전도 공격, 학습 데이터 추출 공격 (Inversion attack)	• 기계학습machine learning, ML 모델에 질의를 던진 후 산출된 결과값을 분석해 학습에 사용된 데이터를 추출한다. 데이터 분류를 위한 기계학습은 주어진 입력에 대한 분류 결과가 신뢰도를 함께 출력하는데, 이 결과를 분석해 학습 데이터를 복원한다. • 학습 데이터 추출 공격은 다음과 같이 얼굴인식 기계학습 모델의 학습을 위해 사용한 얼굴 이미지 데이터를 복원할 수 있는데, 이 대상은 기밀정보나 개인정보가 될 수 있다.  (오른쪽) 실제 학습 데이터, (왼쪽) 학습 데이터 추출 공격을 이용해 재현된 이미지[20]

05-2a 데이터 중독poisoning, 회피evasion 등 공격에 대한 방어 대책을 마련하였는가?

Yes ☐ No ☐ N/A ☐

- 05-1b에서 설명한 데이터에 대한 공격에 대해, 적대적 공격을 방어하고 인공지능 서비스의 강건성을 높이기 위한 다양한 방어 기법이 존재한다.
- 현재까지 완벽한 방어 기법은 없지만, 데이터 설계 및 모델 학습 단계에서 회피 공격과 중독 공격에 방어하기 위한 대표적 기법으로는 적대적 훈련adversarial training, Gradient masking, Defensive distillation, Feature squeezing 등이 있다.

참고 데이터 공격에 대한 방어 기법 및 예시[21]

방어 기법 분류	방어 기법 내용 및 사례
적대적 훈련	• 자동 진단 모델을 학습시킬 때 적대적 사례로서 작동할 수 있는 모든 경우의 수를 미리 학습 데이터셋에 포함하는 방법 • 학습 데이터 개수의 부족함을 극복하기 위해 사용하는 데이터 증강 적용 시, 영상에 다양한 노이즈를 추가하는 방법
Gradient masking	• 학습 모델의 gradient가 출력으로 노출되는 것을 방지하는 방법으로, 학습 모델의 출력에 노이즈를 추가하는 방법
Defensive distillation	• 학습 모델의 구조상 gradient 자체를 정규화 방법과 같이 두드러지지 않게 하여, 적대적 공격의 학습 방향에 힌트를 주지 않도록 하는 방법
Feature squeezing	• 심층학습 모델에 다음과 같은 두 가지 기능을 별도로 추가하는 방법 1) 영상의 인코딩을 단순화하여 표현 색상의 깊이depth 축소 2) 영상에 대해 공간적 평활화smoothing 필터 적용

다양성 존중 | 책임성 | 투명성

요구사항 06 수집 및 가공된 학습 데이터의 편향 제거

대표행위자 | 데이터 공급자 협력 대상 | 데이터 과학자 | 도메인 전문가 | 인공지능 모델 개발자

- 학습을 위한 데이터의 수집 및 가공 시, 발생할 수 있는 편향을 인식하고 이를 제거하는 방안을 마련한다. 이러한 완화 방안이 데이터 수집, 데이터 정제, 라벨링, 샘플링 시 적절히 적용되는지를 확인한다.
- 초거대 인공지능 모델처럼 현실적으로 모든 데이터를 검증하기 어려운 경우에는 샘플링 기법 등을 통해 데이터를 검증한다.

06-1 데이터 수집 시, 인적·물리적 요인으로 인한 편향 완화 방안을 마련하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

공공·사회 분야 서비스를 위한 인공지능 학습용 데이터셋을 직접 수집하여 구축할 때, 본 항목을 고려하여 만족 여부를 판단하십시오.

- 데이터셋을 직접 수집할 때 인적·물리적 요인으로 인해 다양한 데이터가 수집되지 못하여 데이터 편향이 발생할 수 있다. 데이터의 편향은 인공지능 알고리즘의 동작 성능에 문제를 일으킬 수 있고, 시스템의 오동작으로 이어질 수 있으므로 편향을 완화하는 노력이 필요하다.
- 데이터 수집 작업자를 통해 데이터셋을 직접 수집한다면, 데이터 수집 작업자의 기준에 따라 편향이 발생하지 않도록 해야 한다.
- 또한 수집 환경 및 제약조건으로 등으로 인해 다양한 데이터를 확보하기 어려울 수 있으므로 이기종 장치로 데이터를 수집하여 다양성을 확보해야 한다.
- 비용 및 시간적 제약으로 단일 하드웨어를 사용한다면 하드웨어로 인한 데이터 편향이 발생할 수 있으므로 다양한 사양의 하드웨어 도입을 고려해야 한다.

06-1a 인적 편향을 제거하기 위한 절차적, 기술적 수단을 적용하였는가?

Yes ☐ No ☐ N/A ☐

- 공공·사회 분야 인공지능 시스템의 개발 및 고도화 시, 보통 데이터 수집을 위한 작업자를 먼저 선정하고 해당 작업자를 통해 데이터를 수집한다.
- 이때, 다양한 시나리오 조합을 충분히 고려하여 데이터를 수집하지 않으면 수집 작업자의 기준에 따라 편향이 발생할 수 있다. 각 서비스 목적에 따른 시나리오 조합의 예시는 다음과 같다.
  - ✓ 이미지 또는 영상 인식 인공지능을 위한 데이터의 시나리오 요소: 날씨, 시간대, 배경, 대상 객체 크기 등
  - ✓ 챗봇 인공지능을 위한 문항 및 답변 형태 데이터의 시나리오 요소: 방언, 줄임말 등
  - ✓ 음성인식 인공지능을 위한 발화자 음성 데이터의 특성 요소: 음색, 악센트, 발화 속도 등
- 인적 편향을 줄이기 위해 데이터 수집 작업의 가이드라인을 마련하고, 다양한 데이터 수집 작업자를 모집하여 특정 배경 및 성향 등을 배제하며, 수집 결과에 대한 검수자를 충분히 확보한다.

06-1b 데이터의 다양성 확보를 위해 이기종 수집 장치를 활용하였는가?

Yes ☐ No ☐ N/A ☐

- 공공·사회 분야 시스템 개발 및 고도화를 위해 데이터를 직접 수집 시, 수집 환경 및 제약조건 등으로 인해 다양한 데이터를 확보하기 어려울 수 있으므로 이기종 장치로 데이터를 수집하여 다양성을 확보해야 한다. 다음 표는 수집 장치별 이기종 수집 장치의 예시이다.

데이터 다양성 확보를 위한 이기종 수집 장치 항목

수집 장치 종류	이기종 수집 장치 항목
카메라	• 휴대전화, 태블릿, 미러리스 카메라, DSLR(Digital Single-Lens Reflex) 카메라, 필름 카메라, CCTV(Closed-Circiut Television) 카메라 등
마이크	• 휴대전화, 태블릿, 블루투스 마이크, 핀 마이크, 휴대용 녹음기, 스마트 스피커 마이크 등

- 다만 이러한 경우, 수집 경로 및 수집 환경(예: 이미지 크기, 음성 파일 형식 등)이 달라지기 때문에 데이터의 일관성을 유지하기 위해 데이터의 정제와 검수가 충분히 이뤄져야 한다.

06-1c 하드웨어로 인해 발생할 수 있는 데이터의 편향을 점검하였는가?

Yes No N/A
☐ ☐ ☐

- 비용 및 시간적 제약으로 인해 단일 하드웨어를 사용한다면 하드웨어 사양으로 인한 데이터 편향이 발생할 수 있으므로 다양한 사양의 하드웨어 도입을 고려해야 한다.
  - ✓ 영상 또는 이미지 데이터는 휴대전화 카메라, 미러리스 카메라, DSLR 카메라 등 활용
  - ✓ 음성 데이터는 단일 녹음기 모델, 휴대전화 녹음 기능 등 활용
- 다음 표는 각 하드웨어에서 고려할 수 있는 하드웨어 사양의 예시이다.

데이터 편향 점검 시 고려해야 할 하드웨어 사양 항목

하드웨어 종류	하드웨어 사양 항목
RGB 카메라	• 칩셋 종류(예: CCD, CMOS), 해상도resolution, 시야각Field Of View, FOV, 압축방식(예: H.265, H.264 등), 스캔방식(예: progressive, interlaced) 등
마이크	• 다이내믹·콘덴서 종류, 주파수 응답 대역, 신호 대 잡음비, 감도 등

06-2 학습에 사용되는 특성을 분석하고 선정 기준을 마련하였는가?

Yes No N/A
☐ ☐ ☐

해당여부 판단

인공지능 알고리즘 또는 모델을 직접 개발하는 경우에 민감한 특성을 활용한다면 본 항목을 고려하여 만족 여부를 판단하십시오.

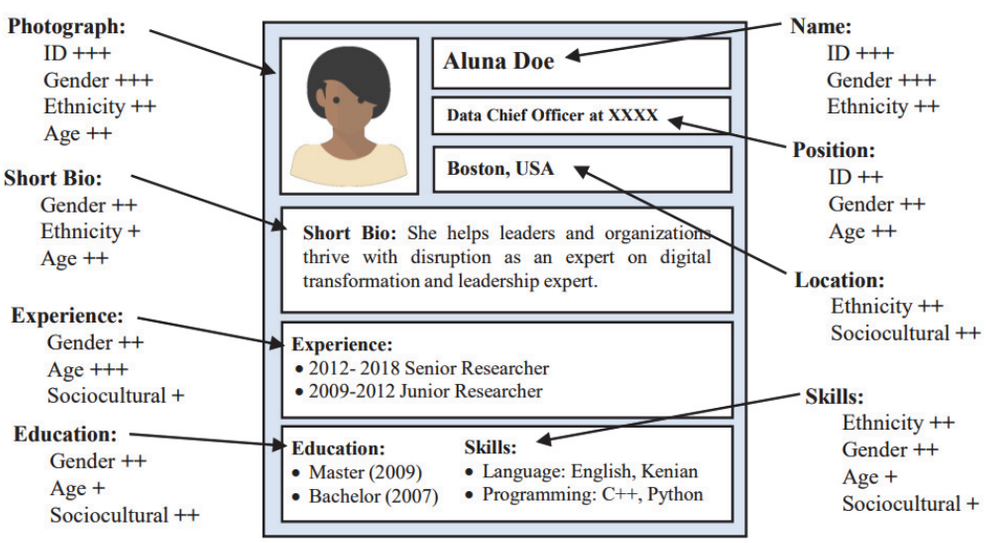
- 편향 제거를 위해 데이터에 포함된 차별적인 요소를 사전에 가려내는 것이 중요하며 이를 위해 학습을 위한 특성의 분석과 선정 기준을 수립하는 것이 바람직하다.
- 공공·사회 분야 서비스를 위한 학습 데이터셋에 민감한 특성 정보가 함께 기록된다면 보호변수를 설정하고 인공지능에 미치는 영향을 분석하여 편향을 방지한다.
- 단, 데이터 전처리 시에 편향을 방지하기 위해 특성이 과도하게 선택되거나 배제되지 않았는지도 검토해야 한다.

06-2a 보호변수 선정 시 충분한 분석을 수행하였는가?

Yes ☐ No ☐ N/A ☐

- 보호변수 선정 시 충분한 분석을 진행하지 않으면 모델의 성능을 저하할 뿐 아니라 민감한 특성에 차별적으로 반응할 수 있다. 따라서 모델 출력에 영향을 미치는 보호변수가 있으면 주어진 데이터셋으로부터 데이터 일부분을 변경하면서 모델의 결과가 어떻게 변하는지를 관찰하고 분석해야 한다.
- 특히, 공공·사회 서비스 개발 시 학습용 데이터셋에는 의도 여부를 떠나 연령, 성별, 지역, 인종 등 민감한 특성이 포함되어 학습될 수 있으므로 설정한 보호변수가 불공평한 결과에 얼마나 영향을 미치는지, 성능이 어떻게 변하는지를 비교 분석해야 한다.

참고 이력서 데이터 셋에서 확인할 수 있는 민감한 개인 속성 정보 예시[22]



이력서의 정보 구역 및 각각에서 파생할 수 있는 개인 속성
(+의 수는 민감한 정보의 수준을 나타냄: +++높음, ++중간, +낮음)

06-2b 편향을 발생시킬 수 있는 특성의 영향력을 완화하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 모델 학습 시, 데이터의 특성을 선택하여 사용함으로써 효율적인 학습은 물론, 컴퓨팅 자원과 비용을 저감 할 수 있으며, 여러 특성 사이의 관계 분석 과정에서 데이터에 대한 깊이 있는 이해를 통해 잠재된 편향을 인식할 수도 있다.
- 편향 완화를 위한 간단한 접근법으로는 편향을 발생시키는 특성을 배제하는 특성 선택 기법feature selection을 적용할 수 있고, 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 제거하는 경우, 다른 종류 편향을 발생시키거나 강화할 수 있어 모든 경우에 효과적인 방법은 아니다. 따라서 편향을 완화하기 위한 다양한 기법(예: 가중치 재지정, 라벨링 재지정, 변수 블라인딩, 샘플링)을 고려해야 한다.
- 단, 시스템 사용 목적에 따라 의도된 편향이거나 학습 과정에서 편향 완화가 가능한 경우에는 예외로 할 수 있다.

참고 이력서 데이터셋을 사용하는 멀티모달 네트워크에서 민감 정보의 반영/미반영에 따른 학습 결과 편향 영향도 분석[22]

• 개요
  - ✓ 24,000 건의 합성 이력서 프로필 생성
    - 정보 블록 5개, 인구 통계학적 속성 2개(성별, 민족성), 얼굴 사진에서 얻은 특성 12개 포함
    - 1) 교육 성취도
    - 2) 이용 가능 여부
    - 3) 이전 경험의 유무
    - 4) 추천서 유무
    - 5) 8가지 다른 공통 언어 셋의 언어 능숙도
  - ✓ 목표 함수를 설정하고, 함수에 존재하는 편향에 영향 받는 방식과 정도를 평가하기 위한 실험 설계

• 목표 함수

$$T^j = \beta^j + \sum_{i=1}^{n} \alpha_i x_i^j$$

  - ✓ n=역량 12가지의 수
  - ✓ α_i=전문가와 상의하여 수동으로 고정한 12개 역량별 가중치 요소
  - ✓ β^j=가변성을 주기 위한 임의의 잡음

• 목표 함수 T^U
  - ✓ 역량 12가지에 성별 및 민족성 점수를 반영하지 않고 계산한 목표 점수

• 목표 함수 T^G
  - ✓ 역량 12가지에 성별 및 민족성 점수를 반영하여 계산한 목표 점수

• 시험 시나리오

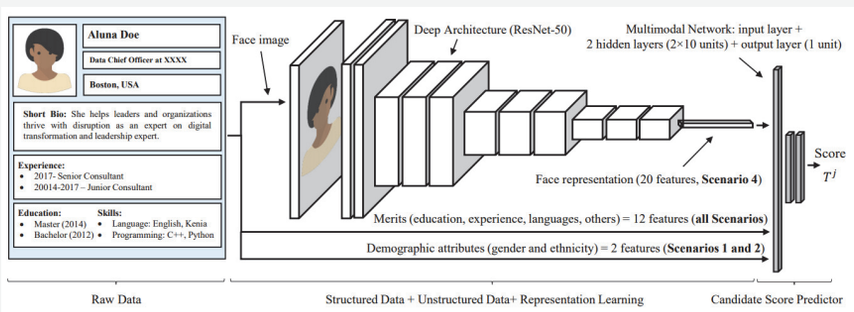
✓ 1) 편향이 제거된 점수 T^U를 목표로 하고, 성별 특성을 추가 입력값으로 하여 학습

✓ 2) 편향된 점수 T^G를 목표로 하고, 성별 특성을 추가 입력값으로 하여 학습

✓ 3) 편향된 점수 T^G를 목표로 하고, 성별 특성을 입력하지 않고 학습

✓ 4) 편향된 점수 T^G를 목표로 하고, 얼굴 사진에서 추출한 특성 임베딩 값을 추가 입력값으로 하여 훈련

• 학습

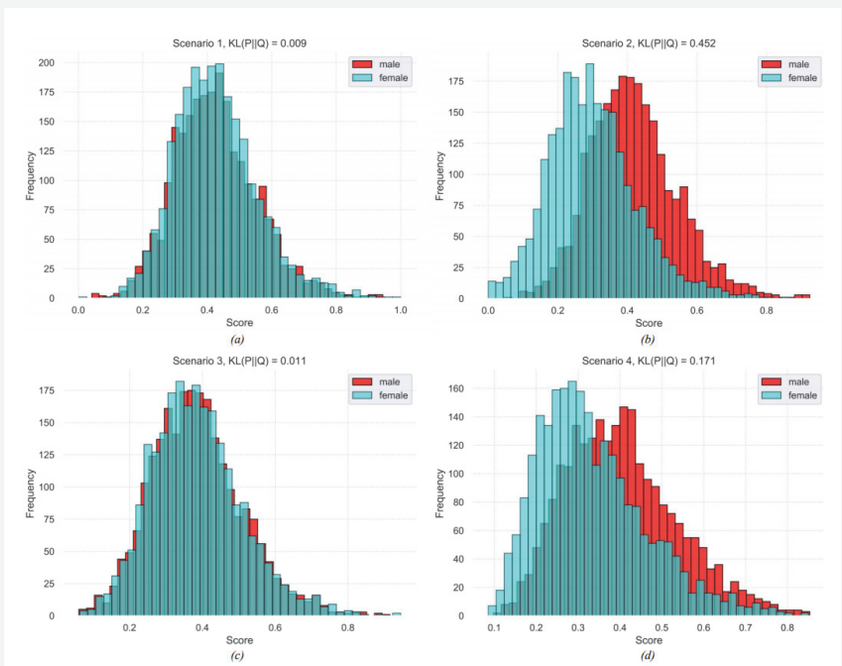


ResNet-50과 완전히 연결된 네트워크로 구성된 멀티모달 학습 아키텍처
서로 다른 도메인(이미지 및 구조화된 텍스트 데이터)의 특성을 융합하는 데 사용함

• 학습 결과 편향 분석

✓ 시나리오 4(d)와 같이 멀티모달 네트워크를 이용해 학습하면 성별 속성을 명시적으로 사용하지 않는 경우에도 학습 데이터에 존재하는 편향을 재현함

✓ 시나리오 2(b)는 입력값과 출력값 모두에 성별 특성이 포함돼 있어 분명하게 여성에 불리한 학습 결과가 발생함



시각 시험 시나리오에 민감한 특성인 성별에 따른 고용 점수 분포:
쿨백-라이블러 발산Kullback-Leibler divergence 비교- 0에 가까울수록 유사한 확률 분포

06-2c 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	조치 사항
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.

06-3 데이터 라벨링 시, 발생 가능한 편향을 확인하고 방지하였는가?

Yes No N/A
☐ ☐ ☐

해당여부 판단

공공·사회 분야 인공지능 알고리즘 또는 모델 개발을 위해 데이터셋을 직접 수집하거나 라벨링할 때, 본 항목을 고려하여 만족 여부를 판단하십시오.

- 지도학습 계열 인공지능 모델은 학습 데이터에 대한 라벨링이 요구된다. 그러나, 이러한 라벨링 작업 시에 작업자의 특정 의도 반영, 실수로 인한 특성 정보의 누락, 무의식적인 판단으로 인한 편향이 발생할 수 있다.
- 데이터 라벨링 과정에서 발생할 수 있는 이슈를 인지하고, 추후 다른 문제가 발생하지 않도록 사전에 명확한 표준 또는 작업 가이드라인을 마련하여 작업자에게 제공하고 교육함으로써 편향의 발생을 방지한다.
- 학습에 사용되는 원천 데이터(텍스트, 오디오 등)를 이해하고, 충분한 상황 분석을 기반으로 라벨링할 수 있는 배경 지식과 요건을 갖춘 작업자와 검수자를 다양하게 섭외하여 작업자별로 나타날 수 있는 오류나 편향을 최소화하고 편향 방지 작업을 수행한다.

06-3a 데이터 라벨링 기준을 명확히 수립하고 작업자에게 제공하였는가?

Yes No N/A
☐ ☐ ☐

- 데이터 라벨링 과정에서 발생할 수 있는 위험 요소를 사전에 식별하고, 이로 인한 문제를 예방할 수 있도록 작업자를 위한 라벨링 가이드라인에 라벨링 기준 및 표준 용어 등을 마련한다.
- 예를 들어 이미지에서 특정 객체를 라벨링해야 하는 경우에 작업자들이 공통된 판단을 내릴 수 있도록 명확한 기준을 제시해야 한다. 또한 챗봇 데이터에서 텍스트 전체가 아니라 특정 키워드, 문장 등을 라벨링한다면 작업자들이 어떤 대상을 라벨링해야 하는지를 판단할 수 있도록 세부적인 기준을 마련해야 한다.
- 데이터 라벨링을 포함한 구축 전반에 대한 가이드라인의 내용은 AI Hub에서 제공하는 인공지능 학습용 데이터셋 구축 안내서를 참고할 수 있다.

참고 자연재해 데이터에 대한 라벨링 기준 예시[23]

- AI Hub에서 제공하는 '자연재해로 인한 생활시설 안전 데이터'는 다양한 각도 및 거리에서 자연재해의 피해 상황을 촬영하고, 자연재해와 피해 객체, 재해 레벨을 정의한다.
- 이때, 비교적 명확한 자연재해와 객체 외의 재해 레벨은 라벨링 작업자의 주관에 따라 판단할 수 있는 정보이기 때문에 각 레벨을 판단할 수 있는 기준을 명확하게 제시한다.

자연재해	객체	재해 레벨	레벨 정의
지진	필로티	정상Lv.0	정상 상태의 필로티
		위험Lv.1	건물 노후화로 인한 필로티 균열 국토교통부 '필로티 건축물 구조 설계 가이드라인' 기준에 부합하지 않게 시공된 필로티 구조 (제안 사항)
		재해Lv.2	지진으로 인한 필로티 균열 및 붕괴
	담장	정상Lv.0	정상 상태의 담장
		위험Lv.1	노후로 인한 균열
		재해Lv.2	지진으로 인한 균열 및 붕괴
태풍 및 강풍	광고판 (간판 포함)	정상Lv.0	정상 상태의 간판 및 광고판
		위험Lv.1	연결부 결합이 잘되지 않거나 기울어져서 강풍에 탈락할 우려가 있는 간판 및 광고판
		재해Lv.2	태풍으로 인한 간판이나 광고판의 파손 및 탈락
	가로등, 전신주, 신호등, 교통 표지판	정상Lv.0	정상 상태의 전신주, 가로등, 신호등, 교통 표지판
		위험Lv.1	기울어져 태풍에 전도 위험이 있는 상태의 전신주, 가로등, 신호등, 교통 표지판
		재해Lv.2	태풍으로 인한 전신주, 가로등, 신호등, 교통 표지판의 전도
	비닐하우스	정상Lv.0	정상 상태의 비닐하우스
		재해Lv.1	태풍으로 인한 비닐하우스 파손
	창문	정상Lv.0	정상 상태의 창문
		위험Lv.1	파손되어 태풍으로 피해가 예상되는 창문
		재해Lv.2	태풍으로 인한 창문 파손
	조경수	정상Lv.0	정상 상태의 조경수
		위험Lv.1	기울어져 태풍에 전도 위험이 있는 상태의 조경수
		재해Lv.2	태풍으로 인한 조경수 전도
폭설	자동차	정상Lv.0	정상 상태의 자동차
		재해Lv.1	폭설로 인한 차량 파손 피해
	도로	정상Lv.0	정상 상태의 도로 (차선 구분 가능)
		재해Lv.1	폭설로 인한 차선 식별 불가 및 정체
	비닐하우스	정상Lv.0	정상 상태의 비닐하우스
		재해Lv.1	폭설로 인한 비닐하우스 파손 및 붕괴
	조경수	정상Lv.0	정상 상태의 조경수
		위험Lv.1	기울어져 폭설에 전도 위험이 있는 상태의 조경수
		재해Lv.2	폭설로 인한 조경수 전도
호우	도로	정상Lv.0	정상 상태의 도로
		재해Lv.1	폭우로 인한 도로 침하
	자동차	정상Lv.0	정상 상태의 자동차
		재해Lv.1	도로 침하로 인한 자동차 피해
	축대, 옹벽	정상Lv.0	정상 상태의 축대와 옹벽
		재해Lv.1	폭우로 인한 축대나 옹벽의 균열 및 붕괴

※ 객체별로 정의된 '위험 레벨' 구조물은 표에 명시된 재해뿐만 아니라 어떠한 재해에도 피해를 입을 수 있는 취약한 상태에 있는 구조물로 볼 수 있음

06-3b 다양한 라벨링 작업자를 섭외하기 위해 노력하였는가?

Yes No N/A
☐ ☐ ☐

- 데이터 라벨링 단계에서 인적 편향을 줄이려면 데이터 라벨링 작업자를 다수로 확보하고, 인구통계적 특성과 배경지식이 다양하고 고르게 분포되도록 구성하는 것이 바람직하다.
- 이 외에도 만약 인공지능 서비스를 특정 산업 분야 또는 공공기관의 임무 및 문제해결을 목적으로 개발한다면 해당 분야와 관련된 다양한 요건을 고려하여 작업자를 선별해야 한다.
- 데이터를 직접 수집하지 않고 오픈소스 데이터셋을 활용한다면 작업자의 다양성이 고려되었는지를 점검한다. 외부 전문가를 통해 데이터셋을 구축할 때는 데이터 라벨링 작업자의 섭외 요건을 전달하여 다양성을 갖출 수 있도록 한다.

참고 개인정보에 대한 라벨링 익명 처리 예시

서비스 유형	작업자 섭외 요건	이유
복지	정치적 성향	• 정치적 성향이 복지에 대한 인식과 태도에 영향을 미친다는 다양한 연구 결과 존재
	사회 문제에 대한 인식, 개인이 처한 상황	• 라벨링하는 시점에 개인의 상황이나 관심 요소가 영향을 미치기 때문
면접·채용	높은 직급 선호	• 직급이 높을수록 면접 영상에 대한 평가 결과의 관점이 다각화되기 때문에 라벨링 작업자의 편향을 허용
전문 분야	경력, 최종학력	• 서비스 유형에 따라 해당 분야에서 특정 수준 이상의 전문가가 필요한 경우가 있으므로 라벨링 작업자의 편향을 허용

06-3c 다양한 라벨링 검수자를 확보하기 위해 노력하였는가?

Yes No N/A
☐ ☐ ☐

- 다양한 데이터 라벨링 작업자를 확보했음에도 불구, 인적 편향이 발생할 수 있다. 따라서, 데이터 라벨링 검수자를 확보하고, 라벨링 결과가 데이터 수집 목적 및 데이터 스펙과 다른 부분은 없는지 등을 확인하며, 수정을 요청하는 등의 작업을 실시해야 한다.
- 데이터 라벨링 검수자 역시 데이터 라벨링 작업자와 마찬가지로 다양하고 고르게 분포할 수 있도록 구성하는 것이 바람직하다. 그러므로 크라우드소싱 등의 방법을 도입하였는지 그리고 검수자에 대한 조사와 분석을 통해 그 분포가 다양하고 고르게 형성되는지 점검한다.

06-4 데이터 수집 시 편향 방지를 위한 샘플링을 수행하였는가?

Yes No N/A
☐ ☐ ☐

해당여부 판단

공공·사회 분야 서비스 제공을 위한 인공지능 알고리즘 또는 모델을 개발 시, 보호변수 등으로 인해 편향이 예상된다면 본 항목을 고려하여 만족 여부를 판단하십시오.

- 데이터셋 내 클래스 불균형이 심하면 데이터가 적은 클래스를 제대로 학습할 수 없으므로 샘플링 기법을 적용하여 클래스 불균형으로 인한 편향을 방지한다.

06-4a 편향 방지를 위한 샘플링 기법을 적용하였는가?

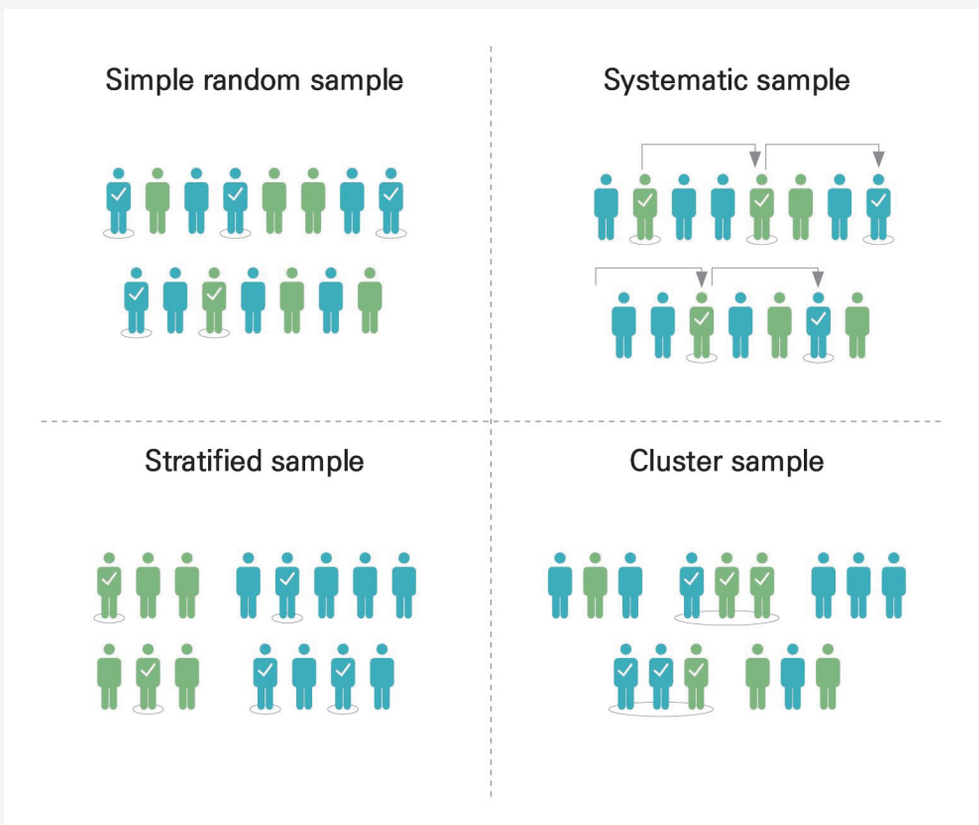
Yes No N/A
☐ ☐ ☐

- 공공·사회 분야 인공지능 서비스를 위한 데이터셋에는 사회적 편견 또는 차별을 일으킬 수 있는 편향의 요소는 명시적/비명시적으로 다양하게 분포(06-2a 참고)할 수 있다. 이러한 다양한 차별 가능성에 따라 공공·사회 서비스를 위한 인공지능 학습 데이터셋을 대상으로 할 때는 다음의 인구통계학적 샘플링 기법을 적용할 수 있다.
  - ✓ 확률 샘플링: 단순 무작위 샘플링simple random sampling, 체계적 샘플링systematic sampling, 층화 샘플링stratifield sampling, 클러스터 샘플링cluster sampling
  - ✓ 비확률 샘플링: 편의 샘플링convenience sampling, 자발적 응답 샘플링voluntary response sampling, 목적 샘플링purposive sampling, 눈덩이 샘플링snowball sampling, 할당량 샘플링quota sampling
- 또한 클래스의 구별 또는 정보를 활용하는 공공·사회 분야 서비스를 위한 인공지능 개발 시, 자연스럽게 클래스 불균형 문제가 발생할 수 있다.
- 클래스 불균형 문제의 해결을 위해 언더 샘플링under sampling, 오버 샘플링over sampling 기법 등을 적용할 수 있다. 객체 클래스의 불균형으로 편향이 예상되면 그로 인한 편향을 방지할 수 있는 샘플링 기법을 적용하고, 적용 과정에서 필요한 활동과 정보가 생성되었는지를 확인한다.

참고 인구통계학적 샘플링 기법 예시[24]

결과에서 유효한 결론을 도출하기 위해서는 그룹 전체를 대표하는 표본을 선택하는 방법을 신중하게 결정해야 한다. 샘플링 방법에는 크게 두 가지 유형이 있다.

- 확률 샘플링: 무작위로 선택하므로 전체 그룹에 대한 강력한 통계적 추론이 가능함
- 비확률 샘플링: 편의성이나 다른 기준에 따라 무작위가 아닌 선택을 포함하므로 데이터를 쉽게 수집할 수 있음



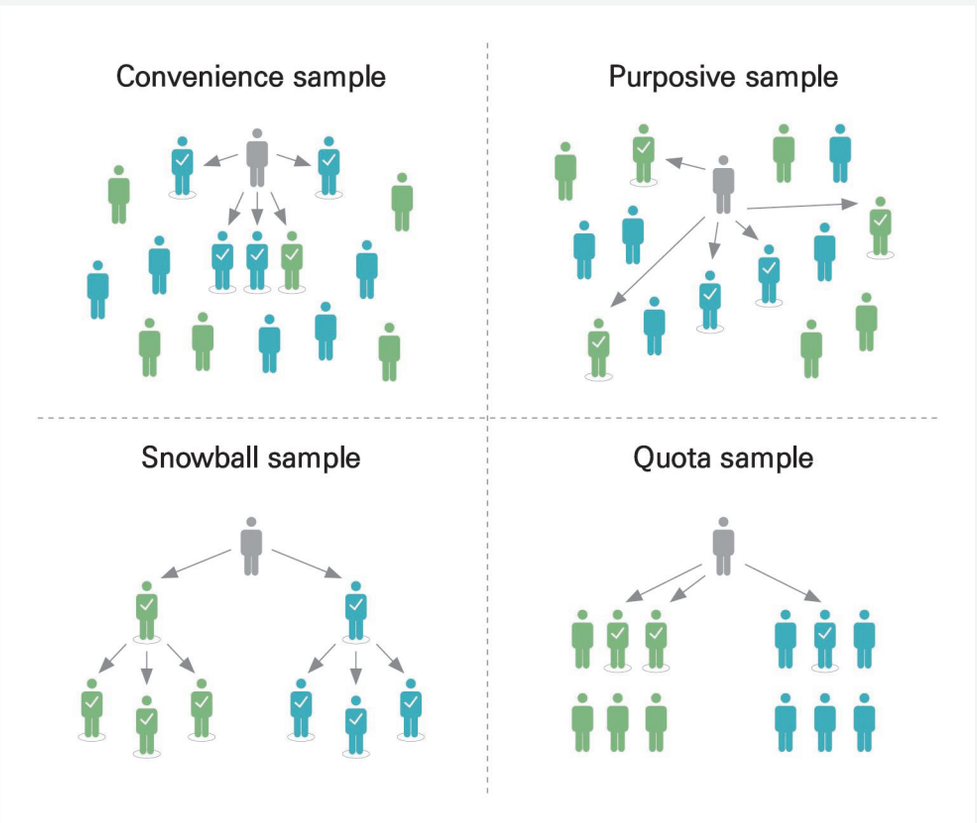
확률 샘플링 방법의 유형 4가지

[확률 샘플링 방법]

모집단의 모든 구성원이 선택될 가능성이 있음을 의미하며, 주로 양적 연구에 사용된다. 전체 모집단을 대표하는 결과를 생성하려는 경우에 확률 샘플링 방법이 유효한 선택이다.

- 단순 무작위 샘플링
  - ✓ 모집단의 모든 구성원이 선택될 확률이 동일함
  - ✓ 난수 생성기와 같은 도구 또는 전적으로 우연을 기반으로 하는 기타 기술을 사용할 수 있음
- 체계적인 샘플링
  - ✓ 수열 등을 적용하여 구성원을 선택함
  - ✓ 표본을 왜곡할 수 있는 숨겨진 패턴이 있지는 않은지 확인하는 것이 중요함
    예) 직원을 팀별로 그룹화하고 팀 구성원을 선택할 때 신입 직원을 건너뛰어 표본이 고위 직원으로 편향될 위험
- 층화 샘플링
  - ✓ 모집단을 하위 집단으로 나누고, 모든 부분군이 표본에 적절하게 표현되도록 하여 좀 더 정확한 결론을 도출할 수 있음
  - ✓ 관련 특성(예: 성별, 연령대, 소득 계층, 직무 역할 등)에 따라 모집단을 하위 그룹으로 나눔

• 클러스터 샘플링
  ✓ 각 클러스터를 전체 표본과 유사한 특성을 가지도록 나누고, 개인 대신 클러스터 그룹을 임의로 선택함
  ✓ 크고 분산된 모집단을 처리하는 데 유용하나 클러스터 간에 상당한 차이가 있을 수 있어 표본에 오류 발생 위험이 있음



비확률 샘플링 방법의 유형 4가지

[비확률 샘플링 방법]

무작위가 아닌 기준에 따라 선택되며 모든 개인이 포함될 기회가 있는 것은 아니다. 비확률 샘플링을 사용하는 경우에도 가능한 모집단을 대표하는 것을 목표로 해야 한다.

• 편의 샘플링
  ✓ 연구원이 가장 쉽게 접근할 수 있는 데이터를 선택함
  예) 연구원의 동료 연구원 데이터

• 목적 샘플링
  ✓ 연구자가 전문 지식을 사용하여 연구 목적에 가장 유용한 데이터를 선택함
  ✓ 통계적 추론을 하기보다는 특정 현상에 대한 자세한 지식을 얻기 원하는 연구나 모집단이 매우 작고 구체적인 질적 연구에 자주 사용됨
  ✓ 효과적인 목적 샘플링에는 명확한 기준과 근거가 있어야 함

• 눈덩이 샘플링
  ✓ 다른 참가자를 통해 참가자를 모집하여 데이터를 선택함

• 할당량 샘플링
  ✓ 미리 결정된 수 또는 단위 비율대로 비무작위적으로 선택함
  ✓ 예) 연구 목표에 따라 식이 선호도에 집중하고자 하는 경우에 육식, 채식, 완전 채식 그룹으로 나누어 600명의 표본 데이터 샘플링 시, 각 그룹에 200명의 할당량을 설정하고, 각 할당량에 도달할 때까지 데이터를 수집함

참고 샘플링 기법 예시 – 오버 샘플링

오버 샘플링이 필요한 이유: 클래스 불균형이 매우 심한 경우에 모델은 적은 수의 클래스의 분포를 제대로 학습하지 못하게 됨. 따라서 모델은 많은 수의 클래스의 분포에 과대 적합되고, 어떤 데이터가 들어오더라도 많은 수의 클래스로 분류하는 문제가 발생함

• 랜덤 오버 샘플링Random Over Sampling, ROS
  ✓ 기존에 존재하는 소스의 클래스를 단순 복제하여 비율을 맞춰줌
  ✓ 단순 복제하여 분포는 변화하지 않지만 숫자가 늘어나기 때문에 더 많은 가중치를 받게 되는 원리
  ✓ 똑같은 데이터가 증식되다 보니 오버피팅의 위험이 존재함

• 스모트Sythetic Minority Over-Sampling Technique, SMOTE [25]
  ✓ 임의의 소수 클래스 데이터로부터 인근 소스 클래스 사이에 새로운 데이터를 생성함
  ✓ 임의의 소스 클래스에 해당하는 관측치 X를 잡고, X로부터 가장 가까운 K개의 이웃 X(nn: nearest neighbors)를 찾음. K개의 X(nn)와 X 사이에 임의의 새로운 데이터 X를 생성함

• 보더라인 스모트borderline-SMOTE [26]
  ✓ SMOTE에서 조금 변형을 준 알고리즘
  ✓ 다수 클래스와 소수 클래스가 서로 인접해 있는 경계선, 즉 보더라인의 분포가 가장 중요함
  ✓ 따라서 경계선에 있는 소수 클래스의 데이터에 대해서 SMOTE를 적용함

• 에이다신Adaptive Sythetic Sampling, ADASYN [27]
  ✓ 보더라인 스모트에서 조금 더 변형을 준 알고리즘
  ✓ 보더라인 근처에서 'Danger, Safe, Noise'의 세 경우로 판단하여 SMOTE로 진행했던 부분을 가중치로 하여 SMOTE를 적용하는 방식

03 인공지능 모델 개발

책임성 안전성

요구사항 07 오픈소스 라이브러리의 보안성 및 호환성 확보

대표행위자 | 인공지능 모델 개발자 협력 대상 | 시스템 엔지니어

- 인공지능 모델 설계 및 개발 단계에서는 개발 기간을 단축하고 최신 기술 동향을 빠르고 유연하게 적용하기 위해 다양한 오픈소스 라이브러리를 활용할 수 있다. 오픈소스 라이브러리를 활용하기로 했다면 사용할 라이브러리가 신뢰할 수 있는 수준인지, 안정적으로 업데이트 중인지, 주의해야 할 라이선스 기준은 무엇인지 등 해당 오픈소스의 버전을 지속적으로 확인하여 운영 및 보안상의 위험 요소를 점검한다.

07-1 오픈소스 라이브러리의 안정성을 확인하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

공공·사회 분야 인공지능 모델 개발에 한 가지 이상의 오픈소스 라이브러리를 활용한다면 본 항목을 고려하여 만족 여부를 판단하십시오.

- 오픈소스 라이브러리는 특정 단체가 관리하기도 하거나, 개인 혹은 기업이 관리한다. 오픈소스를 운영하는 방식은 다양하므로 사전에 꼼꼼히 체크해야 향후 발생할 수 있는 위험risk을 최소화할 수 있다.
- 인공지능 모델 개발에 오픈소스 라이브러리를 사용한다면, 안정성 확인을 위해 해당 오픈소스 라이브러리가 얼마나 많은 사용자를 보유하는지, 업데이트는 자주 이루어지는지, 이슈가 발생했을 때 대응은 신속하게 이루어지는지 등을 따져봐야 한다.

07-1a 활성화된 오픈소스 라이브러리를 사용하였는가?

Yes ☐ No ☐ N/A ☐

- 오픈소스 라이브러리의 안정성은 많은 개발자가 적극적으로 참여할 때 가능하다는 의견이 있다. 따라서, 사용하려는 오픈소스 라이브러리의 개발과정을 주의 깊게 살펴볼 필요가 있다.
- '기업 공개소프트웨어 거버넌스 가이드-정보통신산업진흥원'에 따르면, 오픈소스 프로젝트의 활성화 정도를 확인하는 것도 안정성을 확인하는 한 가지 방법일 수 있다. 해당 오픈소스가 활발한 커뮤니티에서 논의되는지, 그 커뮤니티 내 구성원들이 적극적으로 협력하고 있는지는 아주 중요한 선택의 표지석일 수 있다.

✓ 오픈소스 라이브러리를 GitHub에서 관리 중이라면, 오픈된 이슈 개수나 Pull Request 수, 마지막 커밋 일시 등을 통해 오픈소스 개발이 얼마나 활발하게 이루어지고 지속해서 발전할 가능성이 어느 정도인지 파악할 수 있다.

✓ 그 밖에도 해당 오픈소스와 관련된 StackOverflow 질문 수, 오픈소스 다운로드 수, Google 질의 query 결과 수 등 간단한 측정을 통해서 해당 라이브러리의 활성화 정도를 확인할 수 있다.

참고 신뢰할 수 있는 출처 분석 예시 – GitHub 관리 및 측정

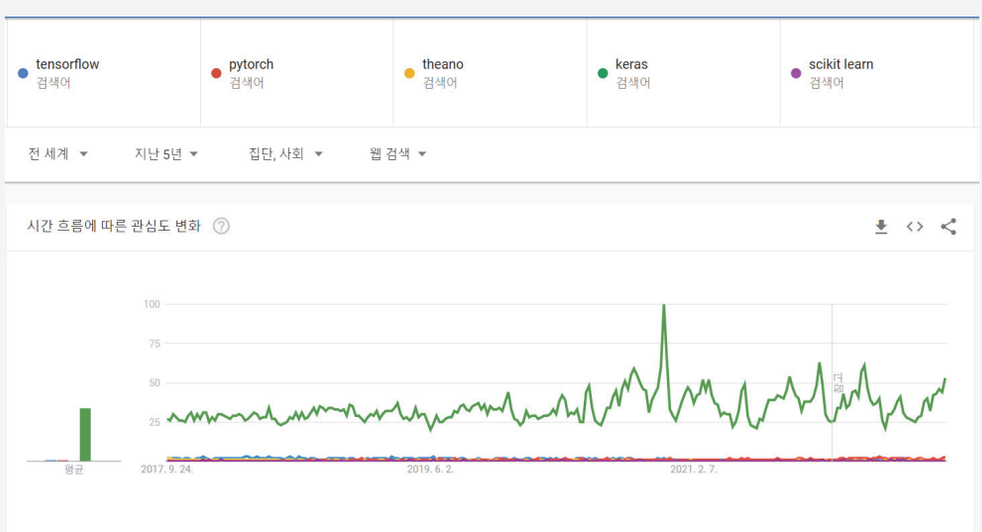
GitHub 출처 분석 예시(2022. 9. 23. 기준)

• 상위 활용 기계학습 오픈소스 라이브러리 중 가장 활발한 활동을 하는 것은 PyTorch, TensorFlow, Scikit-Learn 정도이고, GitHub 라이브러리 기준 Theano는 2021년 11월 이후에 커밋되지 않고 있음

오픈소스 라이브러리 / 항목	Tensorflow	PyTorch	Theano	Keras	Scikit-Learn	NLTK	Apache MXNet
오픈 이슈 개수	2,100	5,000 이상	584	251	1,500	217	1,800
Pull Request 수	243	838	101	79	600	8	206
마지막 커밋 일시	2022. 9. 23.	2022. 9. 23.	2021. 11. 23.	2022. 9. 23.	2022. 9. 23.	2022. 9. 21.	2022. 9. 16.
Contributor 수	3,204	2,445	351	1,059	2,495	374	874
Used 수	212,000	162,000	12,700	-	385,000	154,000	-
StackOverflow 질문 수	79,155	18,792	2,448	40,662	500	6,972	693

Google Query 분석 예시(2022. 9. 23. 기준)

• 과거 5년, 집단·사회 카테고리에서 Keras 오픈소스 라이브러리의 질의 수가 가장 많고, PyTorch, TensorFlow 순으로 조회수가 많음



TensorFlow, PyTorch, Theano, Keras, scikit learn 라이브러리의 집단·사회 카테고리 내 과거 5년 검색어 조회수(관심도) 변화

07-2 오픈소스 라이브러리의 위험요소는 관리되고 있는가?

Yes No N/A
☐ ☐ ☐

해당여부 판단

공공·사회 분야 인공지능 모델 개발에 한 가지 이상의 오픈소스 라이브러리를 활용한다면 본 항목을 고려하여 만족 여부를 판단하십시오.

- 오픈소스 라이브러리는 버전 변경에 따라 법률 및 기술적 면에서 이슈가 발생할 수 있다. 따라서 모델 개발에 오픈소스 라이브러리를 활용하였다면, 오픈소스 라이브러리의 신규 버전 출시에 따른 변경 사항 또는 사용 중인 버전에서 새롭게 발견된 이슈를 지속적으로 추적해야 한다.
- 법률적 위험 요소로 지적재산권을 관리해야 한다. 오픈소스 라이브러리 또는 소프트웨어는 저작권자가 소스코드를 공개했을 뿐이며 여전히 지적재산권으로 보호받는 소프트웨어이다. 따라서 저작권자가 제시한 라이선스(저작권) 준수 조건이 존재하며 오픈소스 라이브러리마다 라이선스에 따라 다양한 의무 사항이 있다. 이때 라이선스 위반 및 저작권 침해로 법적 책임을 질 위험이 있으므로 반드시 라이선스에 대한 위험 요소를 분석하고 관리해야 한다.
- 기술적 위험 요소로서 라이브러리 호환성 및 보안 취약점을 관리해야 한다. 우선 개발 과정 중 서로 다른 오픈소스 라이브러리 또는 버전 변경에 따른 호환성을 고려하여 종류 및 버전을 선택해야 한다. 그리고기 설치된 오픈소스 라이브러리의 보안 문제 및 패치를 지속적으로 추적하고 관리해야 한다.

07-2a 사용중인 오픈소스 라이브러리의 라이선스 준수사항을 이행하였는가?

Yes No N/A
☐ ☐ ☐

- 오픈소스는 무료로 사용할 수 있지만, 라이선스별로 준수 사항은 별도로 규정되어 있다. 그러므로 오픈소스 라이브러리를 활용하여 인공지능 모델을 개발한다면 사용할 오픈소스의 라이선스 종류 및 고지문을 확인하고 허용·의무 사항을 우선 숙지함으로써 향후 발생할 수 있는 법률적 위험을 최소화해야 한다.
- 다음은 OSI[Open Source Initiative] 단체에서 정한 오픈소스 라이선스의 준수사항이다.
  - ✓ 자유로운 재배포 (Free Redistribution)
  - ✓ 소스코드 공개 (Source Code Open)
  - ✓ 2차 저작물 허용 (Derived Works)
  - ✓ 저작자의 소스코드 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 contaminate other software)
  - ✓ 기술 중립적인 라이선스 제공 (License must be Technology-Neutral)

대표적 오픈소스 라이선스의 주요 내용

OSI 기준	Apache License 2.0	GNU GPL General Public License 3.0	AGPL GNU Affero GPL 3.0	LGPL GNU Lesser GPL 3.0	MIT License	Artistic License 2.0	Eclipse License	BSD Berkeley Software Distribution License	MPL Mozilla Public License 1.1
복제, 배포, 수정의 권한 허용	○	○	○	○	○	○	○	○	○
배포시 라이선스 사본 첨부	○	○	○	○	○		○	○	○
저작권 고지사항 또는 Attribution 고지사항 유지	○	○	○	○	○	○	○	○	○
배포시 소스코드 제공의무와 범위		전체코드	네트워크 서비스 포함 전체코드	2차 저작물		○ (표준 버전)	모듈 단위		파일 단위
조합저작물 작성 및 타 라이선스 배포허용	○			○	조건부	○	○	조건부	○
수정내용 고지		○	○	○		○	○		○
명시적 특허라이선스의 허용	○	○	○	○		○	○		○
라이선시가 특허소송 제기시 라이선스 종료	○	○	○	○		○	○		○
이름, 상표, 상호에 대한 사용제한	○		○			○		○	
보증의 부인	○	○	○	○	○	○	○	○	○
책임의 제한	○	○	○	○	○	○	○	○	○

- 오픈소스 라이선스가 적용된 인공지능 모델 개발을 위한 라이브러리를 이용할 때 다음의 라이선스 관련 사안에 유의해야 할 필요가 있다.[29]
  - ✓ 조합 저작물 또는 이차적 저작물에 대한 권리 및 귀속 고지에 대한 사항을 필수 준수
  - ✓ 작은 분량의 프로그램도 저작권이 있을 수 있으므로 권리 및 귀속에 관한 고지를 소스 형태 내에 반드시 포함
  - ✓ 인공지능 모델 관련 프로그램을 조합하거나 이 프로그램을 다른 프로그램에 포함하는 경우에 라이선스 양립성 문제가 발생할 수 있으므로 주의 요함
  - ✓ 오픈소스 라이브러리를 활용해 특허를 취득하고 이를 소송에 이용할 시, 이에 대한 보복 조항이 존재할 수 있으며 만약 없더라도 묵시적 실시허락grant of license이 인정될 수 있으므로 신중한 접근이 필요

07-2b 사용중인 오픈소스 라이브러리의 호환성 및 보안취약점을 확인하였는가?

Yes No N/A
☐ ☐ ☐

- 라이브러리의 버전 변경 과정에서 개발 환경, 언어, 도구 및 다른 라이브러리 버전과 호환되지 않는 호환성 문제를 초래할 수 있다. 따라서 오픈소스 라이브러리 종류 및 버전 선택 시 라이브러리 간 의존성dependency을 파악하는 등 호환성을 고려해야 한다.
- 사용 중인 오픈소스 라이브러리에서 보안취약점이 발견되기도 한다. 보안 취약점에 따른 영향을 최소화하기 위해 보안취약점 및 버전 변경에 따른 릴리즈 노트release note를 지속해서 확인하여 신속히 탐지 및 대응해야 한다.
- 인공지능을 활용한 공공기관 및 대민 서비스는 일반적으로 웹 또는 모바일 앱로 제공되므로 관련 보안 취약점을 반드시 확인하여 사전에 대응할 필요가 있다.

참고 오픈소스 라이브러리의 호환성 예시

- TensorFlow, PyTorch 모델 호환 예시: SCATTER LAB의 한 조직에서 TensorFlow와 PyTorch를 동시에 활용하기[30]
  - ✓ 유연한 리서치에는 PyTorch가 유리하고, 배포 측면에서는 TensorFlow가 유리하여 동시에 사용함
  - ✓ 이를 위해 내부에서 사용하는 모델들을 PyTorch, TensorFlow 버전으로 다시 작성하고, 모델의 모든 가중치를 변환해 주는 코드를 추가로 작성함
  - ✓ 그 결과, TensorFlow Checkpoint에서 PyTorch 모델로 적용이 가능하고, PyTorch State Dict 파일에서 TensorFlow 모델로 적용이 가능해짐
- TensorFlow, OpenVINO 모델 호환[31]
  - ✓ OpenVINO 개발 도구를 활용하여 TensorFlow 1.x, 2.x 모델 형식에서 OpenVINO IR 형식으로 변환
  - ✓ Frozen 모델 포맷(.pb 파일) 및 Non-Frozen 모델 포맷(Checkpoint, MetaGraph, SavedModel)의 변환 가능
- TensorFlow1, 2, Keras 저장 모델 호환[32]
  - ✓ TensorFlow 2에서는 TensorFlow 1에서 저장된 모델과 호환되어 변수와 함수 로드 가능
  - ✓ TensorFlow 2에서는 Keras로 저장된 모델과 호환되어 로드 가능

참고 오픈소스 라이브러리의 보안 취약점 분석 예시

TensorFlow CVE[Common Vulnerabilities and Exposures] 예시[33] (2022. 9. 23. 기준)

- DoS[Denial of Service] 공격에 취약한 부분이 존재하는 것으로 분석되고(32.5%), Overflow 위협도 존재하는 것으로 분석됨(18.3%)
- 총 보안 취약점은 2021년에 201건에서 2022년에 139건으로 줄어들어, 제조사에서 보안 위협에 어느 정도 대응하고 있는 것으로 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
2019	7	1	1	4											
2020	35	6	2	8	3										
2021	201	41	6	38	1			1		8	1				
2022	139	76	1	20						2					
Total	382	124	10	70	4			1		10	1				
% Of All		32.5	2.6	18.3	1.0	0.0	0.0	0.3	0.0	2.6	0.3	0.0	0.0	0.0	

2019~2022년 Tensorflow 오픈소스 라이브러리의 CVE 보안 취약점 분석 결과

PyTorch CVE[Common Vulnerabilities and Exposures] 예시[34] (2022. 9. 23. 기준)

- 보안 취약점 분석 결과, 2021년과 2022년에 각 1건씩 보안 위협이 발견됨

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
2021	1														
2022	1														
Total	2														
% Of All		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

2021~2022년 Pytorch 오픈소스 라이브러리의 CVE 보안 취약점 분석 결과

- 2022년 발견 항목 – CWE-94: Failure to Control Generation of Code('Code Injection')
  - ✓ 보안 위협 내용: 제품이 생성하는 코드 내에서 해당 입력이 사용될 때 사용자 제어 입력(데이터 평면)의 코드(제어 평면) 구문을 충분히 필터링하지 못함
    소프트웨어에서 사용자의 입력에 코드 구문이 포함되도록 허용하면 공격자가 소프트웨어의 의도된 제어 흐름을 변경하는 방식으로 코드를 작성할 수 있음
    이러한 변경으로 인해 공격자가 실행하는 임의 코드가 실행될 수 있음
  - ✓ 온라인에서 실행되는 인공지능 시스템에서 모델 구동을 위해 Pytorch의 환경이 구성되면 공격자가 Code injection 취약점을 공격하여 인공지능 시스템의 실패, 오작동 등을 유도할 수 있음

다양성 존중

요구사항 08 인공지능 모델의 편향 제거

대표행위자 | 인공지능 모델 개발자 협력 대상 | 데이터 과학자 | 시스템 엔지니어 | 인공지능 윤리 전문가

- 인공지능 모델을 개발하는 과정에서 모델의 종류나 시스템의 목표에 따라 편향*이 발생할 수 있으므로, 이를 제거하기 위한 기법을 고려한다.

* 요구사항 06-2 에서 언급한 바와 같이 민감한 특성 정보의 포함으로 문제가 되는 경우에 한함

08-1 모델 편향을 제거하는 기법을 적용하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

공공·사회 분야 인공지능 모델 개발 시 민감한 특성이 입력값 또는 출력값에 활용되거나 영향을 미쳐 편향 발생이 예상되는 경우, 본 항목을 고려하여 만족 여부를 판단하십시오.

- 인공지능 모델은 데이터에 잠재된 편향을 학습하게 되고, 심지어 편향을 더욱 증폭시키기도 한다. 따라서 데이터 정제 단계에서 데이터에 잠재된 편향을 제거하는 방법뿐만 아니라, 모델 개발 과정에서도 모델 편향을 제거 또는 완화하기 위한 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.

08-1a 개발하려는 모델에 맞게 편향제거 기법을 선택하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 모델의 편향을 완화하기 위한 기법은 적용 단계에 따라 3가지로 구분된다. 모델 학습 전, 학습 과정 중, 학습 이후에 적용하는 방식이다.
- 각 방식의 특성과 구현하려는 인공지능 모델 및 목표 임무에 맞게 적절한 기법을 선택하여 적용해야 한다.

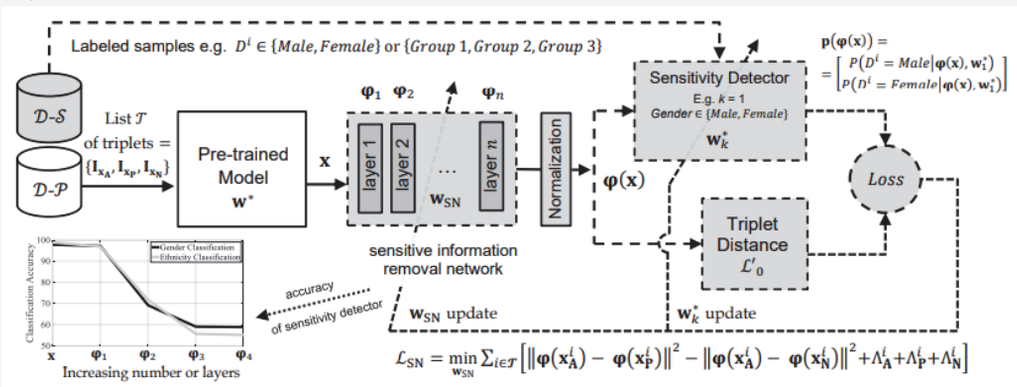
인공지능 모델의 편향을 완화하기 위한 기법 예시

편향 유형	기법	기법 구분			설명
		Pre	In	Post	
알고리즘 편향 algorithmic bias	가중치 재지정	✓			학습 데이터셋 샘플에 가중치를 할당하는 방식
리콜 편향recall bias	라벨링 재지정	✓			학습용 데이터 샘플의 라벨을 수정하는 방식
특성 편향feature bias	변수 블라인딩	✓			분류기가 민감한 변수에 반응하지 않도록 하는 방식
–	변형	✓		✓	숫자 데이터 기반 학습 시 데이터 변환 및 모델 예측 분포를 변환하는 방식
데이터 표본 편향data sampling bias	샘플링	✓			학습 데이터 내 샘플링을 통해 편향을 제거하는 방식
과잉일반화 편향overgeneralization bias	정규화	✓	✓		분류 시 편향에 많은 영향을 주는 클래스 분포를 대상으로 보정하는 방식
데이터 표본 편향data sampling bias	제약 최적화		✓	✓	분류기의 손실 함수에 보정값을 부여하는 방식
평가 편향evaluation bias	임곗값			✓	추론 결과가 결정 경곗값에 가까울 때 편향을 제거하는 방식
알고리즘 편향 algorithmic bias	보정			✓	긍정 예측 비율이 긍정적인 데이터 인스턴스의 비율과 동일하게 분포하도록 설정하는 방식

- 공공기관 및 대민 사회 서비스 제공 시에 개발하고자 하는 모델에 적용할 수 있는 편향 제거 기법의 데이터 타입별 예시는 다음과 같다.
  - ✓ 얼굴 인식 시, 이미지 내 존재하는 민감 특성(성별, 민족성 등)의 제거: 삼중항 손실triplet loss 방법의 확장을 적용하여 성능을 거의 유지하며 특성 임베딩feature embedding 입력에서 민감한 정보를 제거[35]
  - ✓ 음성 감정 분류 시 성별 중화: 웨이브 데이터의 주파수 이동, 필터링, 시간 확장 등 사운드에 여러 유형의 변경을 적용[36]
  - ✓ 챗봇 텍스트 내 유해 언어 감지 시 편향 제거: 어휘 및 방언의 편향이 존재하며 이를 완화하기 위해 라벨링 재지정[37]
  - ✓ 아동학대 가정 여부 판단 시 편향 제거: 사회 복지 현황, 성별과 같이 편향을 유발하는 것으로 의심되는 변수를 제외하여 재학습 및 제외 전의 결과와 편향 비교[38]

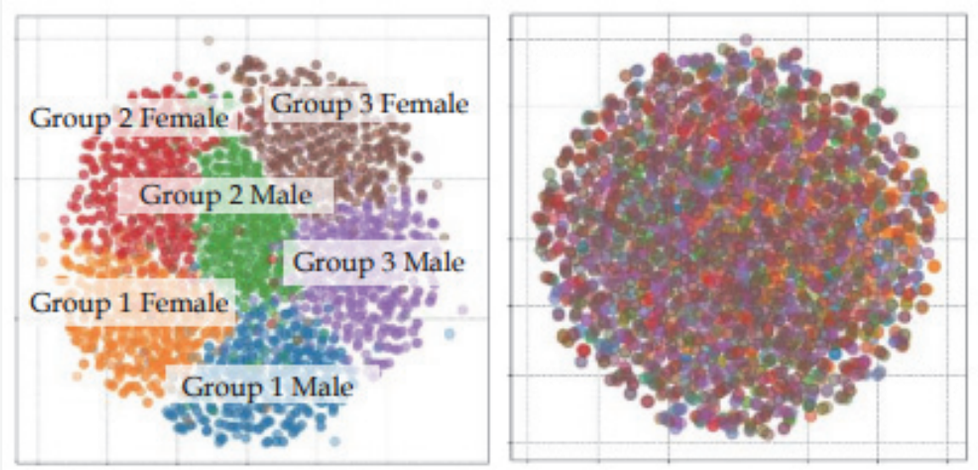
참고 이미지 내 존재하는 민감 특성의 제거 사례[35]

Triplet loss 방법을 확장 적용하여 이미지 내 존재하는 민감 특성 제거



사전에 학습된 임베딩 표현에서 민감한 정보를 제거하기 위한 SensitiveNets의 학습 과정

- 시험 대상 데이터셋
  - ✓ DiveFace – 627K 명의 사람에게서 수집된 470만 장의 얼굴로 구성된 데이터베이스
  - ✓ 데이터베이스에서 백인이 77%, 아시아인은 9% 미만을 대표하는 것으로 나타남
- 민감한 정보 제거 결과



ResNet-50의 임베딩 표현(좌)과 SensitiveNets의 학습 결과 임베딩 표현(우)의 2D 투영 결과

- ✓ 민족성 및 성별 속성에 따라 색상을 지정했을 때 기존의 얼굴 확인 인공지능 모델은 인구통계학적 속성과 높은 상관관계가 있는 클러스터 6개를 생성함
- ✓ 제안하고 있는 SensitiveNets 학습 방법을 통해 민감한 특성 2개(민족성, 성별)가 사라진 인구통계학적 클러스터링 결과를 확인함

08-1b 편향성 평가 및 모니터링을 위한 정량적 지표를 선정하고 관리하는가?

Yes ☐ No ☐ N/A ☐

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

편향을 정량적으로 측정하는 지표 분류

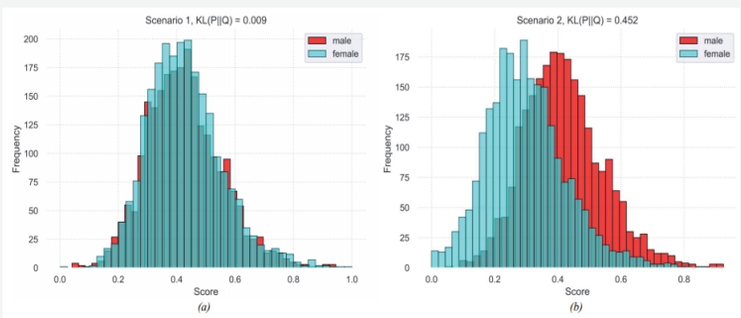
분류	지표 (모델)
패리티parity 기반 지표	인구통계학적statistical/demographic 형평성 지표, 차등적disparate 효과 지표
혼동 행렬confusion matrix 기반 지표	동등 기회equalized opportunity, 동등 가능성equalized odds, 전체 정확도 형평성, 조건부 사용 정확도 형평성, 대응 형평성 비보상 동등화
점수score 기반 지표	양성/음성 클래스 균형 지표
사후가정counterfactual 기반 지표	사후가정 공평성
개인individual 공평성 지표	일반화 엔트로피 지수, 셰일 지수

편향성 평가를 위한 기준 및 의미 [39]

유형	종류	수학적 의미
예측 기반	집단 공정성	집단별 긍정적 예측값을 할당받을 확률이 동일
	조건부 통계적 동등성	특정 데이터 속성을 통제했을 때, 그룹별로 긍정적 예측값을 할당받을 확률이 동일
예측 및 실제 결과 기반	예측적/결과적 동등성	긍정적 예측값의 비율이 집단 간에 실제로 동일
	위양성률false positive error rate 균형	위양성 예측값을 할당받을 확률
	위음성률false negative error rate 균형	위음성 예측값을 할당받을 확률
	동등 확률	예측값 기반 양성예측도Positive Predictive Value, PPV와 음성예측도 Negative Predictive Value, NPV
	조건부 사용 정확도 동등성	예측값 기반 양성예측도Positive Predictive Value, PPV와 음성예측도 Negative Predictive Value, NPV
	전체 정확도 동등성	위양성false positive과 위음성false negative의 비율이 집단 간 동일
	대우 동등성	위양성false positive과 위음성false negative의 비율이 집단 간 동일
예측 확률 및 실제 결과 기반	테스트 공정성 (조건 빈도)	예측된 확률 점수에 대해 보호집단/비보호집단의 피험자가 실제 양성일 확률이 동일할 때
	잘 보정됨well-calibration	예측된 확률 점수에 대해 보호집단/비보호집단의 각 피험자가 양성에 실제로 속할 확률과 같을 뿐만 아니라 예측된 확률 점수와도 같을 때
	양성 집단에 대한 균형	보호/비보호집단의 양성 클래스를 구성하는 각 피험자가 평균 예측 확률 점수 S를 동일하게 갖는 경우
	음성 집단에 대한 균형	보호/비보호집단 모두에서 음성인 피험자는 평균 예측 확률 점수가 동일해야 함
유사성 기반	인과적 차별	정확히 동일한 속성을 지닌 두 주제에 대해 동일한 분류를 생성할 때

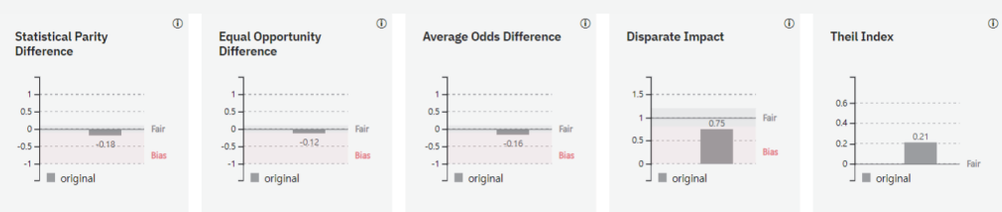
유형	종류	수학적 의미
	블라인드unaware를 통한 공정성	의사결정 과정에서 민감한 속성이 명시적으로 사용되지 않을 때
	인식을 통한 공정성	유사한 개인이 유사한 분류를 가질 때
인과 추리	반사실적 공정성	예측된 결과가 보호된 속성의 자손변수에 의존하지 않는 경우
	미해결된 차별이 없음	보호된 속성에서 예측된 결과까지의 경로가 존재하지 않는 경우
	대리 차별 금지	보호된 속성에서 '대리 변수에 의해 차단되는 예측된 결과'까지의 경로가 없는 경우
	공정한 추론	인과 관계 그래프의 경로를 정당/부당한 것으로 분류

참고 예측에 기반한 집단 공정성/편향성 지표의 측정 사례



쿨백 라이블러 발산을 활용하여 민감한 특성 하위 그룹에 대한 예측 결과의 확률 분포 비교[22]

성별에 따른 채용 점수 확률 분포의 쿨백 라이블러 발산 지표 비교
– 0.009로 차이가 거의 없음(좌), 0.452로 집단에 따라 차이가 많이 남(우)



COMPAS 데이터셋에서 범인의 재범 예측 시 민감한 특성 하위 그룹에서 편향성 지표 측정[40]

민족성 하위 그룹(특권–백인, 비특권–백인이 아님)에 대한 예측 결과, 측정 지표 5개 중 4개에서 편향된 결과를 확인

- ✓ 인구통계학적 패리티 지표statistical parity difference: 기준 범위 –0.1~0.1을 벗어난 –0.18로 편향
  – 특권 그룹이 비특권 그룹에 대해 받은 유리한 결과의 비율 차이
- ✓ 동등 기회 차이equal opportunity difference: 기준 범위 –0.1~0.1을 벗어난 –0.12로 편향
  – 특권 그룹과 비특권 그룹 간의 진양성true positive 비율 차이
- ✓ Average Odd 차이: 기준 범위 –0.1~0.1을 벗어난 –0.16으로 편향
  – 특권 그룹과 비특권 그룹 간의 위양성false positive 비율과 진양성 비율의 평균 차이
- ✓ 이질적인 영향disparate impact: 기준 범위 0.8~1.25를 벗어난 0.75로 편향
  – 특권 그룹과 비특권 그룹에 대한 유리한 결과의 비율
- ✓ Theil 지표: 비교적 낮은 점수에서 공정함, 0.21은 공정한 것으로 보임
  – 데이터셋의 모든 개인에 대한 일반화된 혜택 엔트로피로 계산

안전성

요구사항 09 인공지능 모델 공격에 대한 방어 대책 수립

대표행위자 | 인공지능 모델 개발자 협력 대상 | 시스템 엔지니어

- 인공지능 모델은 적대적 의도를 가진 사용자에 의해 학습 데이터 및 기능을 도용당하거나 다른 방식의 공격으로 악용될 수 있으므로 이를 방지 또는 완화하기 위한 대책을 수립한다.

09-1 모델 추출 공격model extraction attack에 대한 방어 방안을 수립하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

공공·사회 분야 서비스를 위한 텍스트 또는 음성 데이터 기반의 인공지능 서비스를 개발하고 적용할 때, 인명·재산 피해, 형평성 침해, 복지 대상자로 선택되는 등의 편법이 예상된다면 본 항목을 고려하여 만족 여부를 판단하십시오.

- 모델 추출 공격은 학습된 모델의 다양한 입력 결과를 분석하고 분류 기준을 추출하여 실제 서비스 중인 모델과 유사한 성능의 대체 모델을 구성한 다음에, 회피 공격을 위한 적대적 데이터를 생성하거나 출력된 결과를 분석해 모델 회피 공격 등의 2차 공격에 활용할 수 있다.
- 공공·사회 서비스를 위한 이미지 인식 및 처리, 텍스트 또는 음성인식 인공지능 서비스 등을 대상으로 하는 추출 공격을 완화하거나 방어하기 위해서는 질의query 횟수 제한, 예측 결과의 난독화와 같은 방법을 적용할 수 있다.

09-1a 모델 추출 공격에 대비하는 방어 기법을 적용하였는가?

Yes ☐ No ☐ N/A ☐

- 공공·사회 서비스를 위한 이미지 인식 및 처리, 텍스트 또는 음성 처리 인공지능 모델은 시민을 대상으로 늘 열려 있어서 모델 추출 공격에 취약할 수 있다.
- 인공지능 모델 추출 공격에 대한 주요 완화 방법은 특정 시간 간격당 인공지능 서비스에 대한 질의의 수를 제한하는 것, 의심스러운 질의를 탐지하여 경고하는 것, 예측 결과를 난독화obfuscation하는 것 등이 있다.

모델 추출 공격의 방어 기법

방어 기법 분류	방어 기법 내용
질의 횟수 제한	특정 기간 내에 수행할 수 있는 질의의 횟수를 제한하여 모델 공격을 위한 반복적인 질의를 방어하는 기법
학습 기반 모니터링	기계학습을 활용하여 모델 공격에 대해 사전 탐지 및 경고 알림, 상응하는 방어 기법을 실행하는 등, 능동적으로 방어하는 기법
예측 결과 난독화	예측 결과가 결정경계에 가까운 경우에 예측 결과의 정확도를 임의로 낮춰 모델의 세부 속성에 대한 추출을 방해하는 기법

09-2 모델 회피 공격model evasion attack에 대한 방어 방안을 수립하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

공공·사회 분야 인공지능 서비스를 개발하여 적용할 때, 인명·재산 피해, 형평성 침해, 복지 대상자로 선택되는 등의 편법이 예상된다면 본 항목을 고려하여 만족 여부를 판단하십시오.

- 모델 회피 공격은 입력 데이터에 최소한의 변조를 가해 인공지능 모델을 속이는 기법이다. 특히 이미지 도메인은 약간의 변화가 발생해도 사람의 눈에 잘 띄지 않으므로 적대적 공격adversarial attack에 취약하다.
- 모델 회피 공격을 완화하기 위해 적대적 학습을 통해 텍스트 처리 인공지능 모델에 대한 공격을 완화하거나, 다운 샘플링down sampling, 로컬 스무딩local smoothing, 양자화quantization 방법을 사용하여 음성 인식 인공지능 알고리즘을 대상으로 하는 공격을 완화하는 방법이 연구되어 방어 기법으로 고려할 수 있다.

09-2a 모델 회피 공격에 대비하는 방어 기법을 적용하였는가?

Yes ☐ No ☐ N/A ☐

- 이미지 및 영상 처리, 텍스트 처리, 음성 처리 인공지능 알고리즘을 대상으로 하는 모델 회피 공격 사례가 다수 발생하여, 이에 관한 연구가 이뤄지고 있다. 공공·사회 분야 서비스 개발 시 이를 방어하기 위한 적절한 대응 과정이 필요하다.
- 데이터 타입별로 인공지능 모델 대상 회피 공격과 알려진 주요 완화 방법은 다음과 같다.

모델 회피 공격 및 방어 기법

데이터 타입	모델 회피 공격 방법	방어 기법 분류	방어 기법 내용
이미지	입력 이미지 공격[41]	적대적 훈련 adversarial training	분류 과정에서 가능한 적대적 샘플을 모방한 적대적 샘플을 학습 데이터셋에 도입하여 함께 학습함
		그래디언트 마스킹 gradient masking	강력한 공격을 계산하기 위해서는 그래디언트가 필요하므로 그래디언트를 숨겨 문제를 해결하고자 함 현재는 방어 전술로 효과적이지 않은 것으로 판명됨, 공격자가 여전히 대리자를 만들어 모델 기울기와 상관없이 공격할 수 있다는 것이 밝혀짐
		입력 수정 input modification	적대적 노이즈를 제거하기 위해 어떤 방법으로든 '정리'하는 것 - 노이즈 제거 솔루션, 컬러 비트 심도 감소, 평활화, GAN을 이용한 화질 개선, JPEG 압축, 픽셀 편향, 일반 기본 함수 변환 등
		감지 detection	입력 수정 방법과 연계하여 2가지 입력의 결과를 예측해 본 후, 입력 수정 후의 예측 결과와 원본의 예측 결과가 멀리 떨어져 있으면 입력이 변조되었을 가능성이 있는 것으로 가정
		추가(NULL) 클래스 extra class	보통 특정한 데이터 분포에 대해 학습되며 정의상 그 범위를 벗어날 때 단서가 없음. 따라서 레이블이 무엇인지 분명히 모를 때 레이블을 강제하지 않고 클래스가 NULL인 추가 클래스를 제시하도록 함
텍스트	자연어 분류 공격[42]	적대적 훈련	분류 과정에서 가능한 적대적 샘플을 모방한 적대적 샘플을 학습 데이터셋에 도입하여 함께 학습함
음성	음성 명령 인식 공격[43]	묵음 클립 Appending Silence Clip, ASC을 활용한 프레임 오프셋 변경	묵음 클립을 프레임 앞 부분에 추가하여 프레임 시작 오프셋 변경 방어, 탐지 및 하이브리드 전략으로 활용할 수 있음
		다운 샘플링	샘플링 이론을 기반으로 하여 복구된 신호의 품질을 떨어트리지 않고 대역 제한 오디오 파일을 다운 샘플링한 후에 다시 업 샘플링하여 입력값으로 활용
		로컬 스무딩	슬라이딩 윈도우를 활용하여 윈도우 중앙값, 평균값 등으로 오디오 샘플 값 대체
		양자화	보통, 입력공간에서 적대적 공격을 위한 진폭은 작기 때문에 오디오 샘플링된 데이터의 진폭을 가장 가까운 배수로 반올림하여 정할 수 있음, 256, 512 양자화 단계에서 가장 좋은 성능을 얻음

참고 콘텐츠 필터 인공지능 회피 공격 사례[44]

2019년 3월 뉴질랜드 Christchurch에서 발생한 총격 사건을 생중계하였고, 이후 다른 소셜미디어에 광범위하게 업로드됨

• 개요
- ✓ 총격 사건 당시 범인의 관점에서 촬영된 동영상 원본의 사본이 8chan이라는 파일 공유 플랫폼에 게시된 후 최소 십만 단위의 숫자로 각 소셜 미디어에 업로드됨
- ✓ 이 사고로 50명이 사망하고, 50명이 부상함
- ✓ 28세의 총격범은 헬멧에 장착된 카메라를 착용하고, 일부 사람들이 "소셜 미디어에 최대한 퍼지도록 설계되었다."라고 말하는 방식으로 총격 사건을 생중계함

• 경과
- ✓ 각 소셜미디어들은 3일이 지나도 17분짜리의 비디오 사본을 네트워크에서 차단하기 위해 노력함
- ✓ Youtube는 플랫폼의 인공지능 시스템에 의해 약관 위반으로 잘못 분류된 동영상을 식별하기 위한 프로세스의 인적 검토 부분을 비활성화하는 실수를 함
- ✓ Youtube는 정확한 업로드 수를 공개하지 않음
- ✓ Facebook은 150만 건을 삭제하였다고 함
  - 그중 80%인 120만 건은 플랫폼에 올라오기 전인 업로드 시 차단됨
  - 30만 건은 게시된 후 24시간 이내에 제거됨
  - 라이브 영상의 조회수는 200회 미만이었고, 총격범이 업로드한 비디오는 삭제되기 전까지 조회수가 약 400회였음

• 공격 내용
- ✓ 문제가 되는 영상을 인식하도록 설계된 인공지능은 완벽하지 않음
- ✓ 많은 네트워크는 동일한 영상이 두 번 이상 업로드되는 경우를 인식하여 대량 업로드를 방지하기 위한 해싱 기법을 사용함
- ✓ 그러나 일부 사용자가 영상을 줄이거나, 로고를 추가하거나, 실제 이벤트를 비디오 게임처럼 보이게 하는 효과를 사용하여 해싱을 우회할 수 있었음
- ✓ 원본 영상은 삭제할 수 있었으나 파생된 영상을 제거하는 데는 어려움이 많았음

• 사후 처리 사례
- ✓ Facebook은 비디오에서 보다 많은 변형을 포착하기 위해 해싱 기술을 확장하여 프로세스에 오디오 해싱을 추가함
- ✓ Facebook, YouTube, Twitter 및 Microsoft를 비롯한, 테러 대응을 위한 글로벌 인터넷 포럼의 일부 네트워크는 데이터베이스에 비디오의 변형을 추가하여 다른 네트워크가 동일한 업로드를 방지할 수 있도록 함

책임성 투명성

요구사항 10 인공지능 모델 명세 및 추론 결과에 대한 설명 제공

대표행위자 | 인공지능 모델 개발자 협력 대상 | 데이터 과학자 시스템 엔지니어 시스템 운영자

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

  * 사람이 인공지능 모델의 의사결정 방식을 파악할 수 있도록 돕는 모델의 작동 방식에 대한 유용한 정보(예: 의사결정 메커니즘, 의사결정의 기초를 이루는 학습 데이터, 인공신경망 내에서 사용된 변수와 가중치)

참고 설명가능성 적용 전 고려해야 할 사항

1. **제품 및 서비스의 다양성에 대한 고려:** 모든 인공지능 모델과 제품 및 서비스에 설명가능성이 필요한 것은 아니다. 사용자가 제품 및 서비스를 이용하면서 시스템 동작 및 모델의 추론 결과에 관해 설명을 요구하는 분야가 있지만, 그렇지 않은 분야도 있다. 관련하여, UNESCO에서는 일시적이지 않거나, 쉽게 되돌릴 수 없는 인공지능 시스템의 경우에는 출력된 결과의 투명성이 보장되도록 사용자에게 의미 있는 설명이 제공되어야 한다고 언급한다. 따라서 이러한 사항들을 고려하여 본 요구사항을 선택적으로 적용할 수 있다.
2. **설명가능성이 미치는 영향에 대한 고려:** 설명가능성은 아직도 기술적으로 연구 및 개발이 활발하게 이루어지는 분야로서, 여전히 기술적 한계가 존재함과 동시에 설명가능성 외 다른 속성과도 상호 연관성이 있어 신중히 접근해야 한다. 일례로, 과도하게 설명가능성을 구현하는 경우, 모델 성능 및 프라이버시 등에 부정적인 영향을 초래한다는 의견도 존재한다. 따라서 본 요구사항은 제품의 개발 의도와 설명이 적용되는 상황 및 영향을 파악하여 설명의 적절한 수준을 마련하여야 한다.

10-1 사용자가 모델 예측 결과의 도출 과정을 수용할 수 있도록 근거를 제공하는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

공공·사회 분야 서비스 시스템의 신뢰성을 확보하기 위해 인공지능 알고리즘 및 모델이 반영된 시스템의 동작 결과나 원인 등의 설명 정보를 제공하고자 하는 경우에 본 항목을 고려하여 만족 여부를 판단하십시오.

- 인공지능 모델의 추론 결과 및 인공지능 시스템의 동작을 사용자가 신뢰하기 위해서는 시스템 사용자가 인공지능 모델이 제공하는 추론 결과의 도출 과정을 이해할 수 있어야 하며, 이에 대한 설명 및 근거를 사용자에게 제시하는 것이 바람직하다.

- 공공·사회 분야 서비스는 공공의 이익, 편의, 수혜 기준 등의 결과를 제공할 수 있으므로 사용자에게 모델의 추론 결과를 설명하는 것을 고려해야 한다. 예를 들어 복지의 수혜 대상 여부에 대한 정보를 제공할 때, 확인하고자 하는 사용자의 어떤 정보가 가부 여부에 가장 크게 영향을 미쳤는지를 설명하는 것을 고려해야 한다.
- 또한 시스템이 제대로 동작하지 않거나 업그레이드를 위한 정보를 얻으려는 이유 등에 따라 모델의 추론 결과를 설명하는 것을 고려해야 한다. 다음은 내부적으로 모델 성능 개선을 위해 고려할 수 있는 모델 출력 및 활용 정보의 예시이다.
  - ✓ AI 면접 시스템에서 이미지 및 음성 처리 시: 처리 결과 예측값이 낮은 확률값confidence value일 때 성능 개선을 위한 분석에 활용할 입력 데이터 및 기타 정보(영상 배경, 영상 내 면접자의 위치 등)
  - ✓ 쓰레기 무단투기 탐지 모델에서 이미지 처리 시: 쓰레기를 버리는 행위와 그 반대의 탐지 결과가 낮은 확률값일 때, 성능 개선을 위한 분석에 활용할 입력 데이터 및 주변 환경 정보(상가 등 복잡도, 미탐지, 오탐지 시간대, 날씨 등)
- 공공·사회 서비스에서 활용할 수 있는 인공지능 시스템을 대상으로 설명가능한 인공지능 eXplainable AI, XAI 등을 활용한 여러 연구와 시도 사례, 도입 사례를 검토해야 한다.
- 또한 XAI 기술로 아직 설명할 수 없는 부분에서는 전통적인 의사결정 트리 기법이나 현재 연구되고 있는 원인 학습 기술을 검토하여 도입을 고려해 볼 수 있다.
- XAI 기술의 적용 가능 여부를 검토한 후, XAI 기술 적용이 가능하다면 10-1a 를 활용하고 적용이 어렵다면 10-1b 를 활용할 수 있다.

10-1a XAI 기술 적용이 가능한 경우, 인공지능 모델의 추론 결과를 설명하기 위한 기법 적용에 대해 검토하였는가?

Yes ☐ No ☐ N/A ☐

- 심층학습 기술을 활용한 인공지능 시스템은 성능이 우수하지만 설명가능성이 낮다. 이런 경우에 모델 추론 결과에 대한 확신과 시스템 전체에 대한 신뢰도가 낮아질 수 있으므로 사용자가 모델 추론 결과를 수용할 수 있는 근거를 확보해야 한다.
- 대표적인 XAI 기술로 모델 비종속적인 설명 방법을 제공하는 LIME 지표나, 모델 종속적인 설명 방법을 제공하는 LRP 지표 등을 이용할 수 있다. 다만, XAI기술은 연구가 계속 진행 중이므로 기술 도입 전에 적합한 기법을 선택하는 것이 중요하다. 다음은 챗봇 및 음성인식 인공지능 시스템 모델의 추론 결과에 관한 적용 사례이다.

참고 챗봇 및 음성인식·분류 인공지능 시스템의 설명가능성 적용 사례

• 설명 가능한 인공지능 챗봇[45]
  ✓ 인공지능 시스템의 목적
    – 타이태닉(Titanic) 데이터셋에서 학습된 기본 random forest 모델을 기반으로 타이태닉호 침몰 시의 생존 확률을 예측하고 이유를 설명하는 챗봇
  ✓ 설명을 위한 적용 방법
    – 알려져 있지 않음
  ✓ 결과
    – 나이를 입력하면 타이태닉호에서 생존할 확률을 알려줌
    – 변경한 나이(예: 더 어린 나이)를 입력하면 '나이' 조건에 따른 생존 확률 차트를 보여줌
    – 죽는 이유로 내용을 입력하면 이유와 확률을 차트 형식으로 보여줌

• 설명 가능한 챗봇 인터페이스 설계[46]
  ✓ 시스템의 목적
    – 고장의 발생을 인정할 뿐만 아니라 고장의 원인과 고장이 발생한 정확한 위치에 대한 사용자의 이해도를 높일 수 있는 새로운 메커니즘 설계
  ✓ 주요 설계 목표 5가지
    목표 1: 사용자가 챗봇의 높은 수준의 기본 작동을 더 잘 이해할 수 있도록 의도 및 개체 측면에서 챗봇의 기능을 설명
    목표 2: 사용자가 질의에서 챗봇이 이해할 수 있는 것과 이해할 수 없는 것이 있음을 파악할 수 있도록 챗봇의 역량과 고장의 원인을 설명. 설명은 챗봇이 실패하는 정확한 이유를 설명하여 표시해야 함
    목표 3: 역량과 한계를 각각 설명하기 위한 규범 및 비교 사례에 기반하여 설명을 제공
    목표 4: 시각적인 단계별 설명을 제공하여 관련성 있고 이해하기 쉽도록 함
    목표 5: 사용자가 "다음", "이전", "종료"와 같은 UI 컨트롤을 포함하여 설명에 액세스하고 탐색할 때 접근이 자유로울 수 있도록 함
  ✓ 결과
    – 전반적으로 모든 참가자는 유용성, 투명성, 신뢰성 측면에서 챗봇의 설명이 이해하기 쉬운 것으로 평가함

• 음성에서의 감정 분류 설명[47]
  ✓ 인공지능 시스템의 목적
    – 음성 감정을 예측하고 관련 설명을 제공하기 위함
  ✓ 설명을 위한 적용 방법
    – 인간과 유사한 설명을 제공하기 위해 인간 이론을 적용함
    – 사람의 지각·인지 프레임워크가 설명을 지원하는 방법에 기반하여 음성 감정을 인식함
  ✓ 결과
    – 음성에서 포함되어 있는 날카로움, 목소리 크기, 평균 음조, 속도, 쉼 등의 특성을 분석하여 설명함
    – 문장에서 어떤 단어가 가장 주요했는지 saliency 맵 정보를 제공함
    – 참가자 14명을 모집하여 사람의 평가 결과와 얼마나 유사/상이한지를 분석함

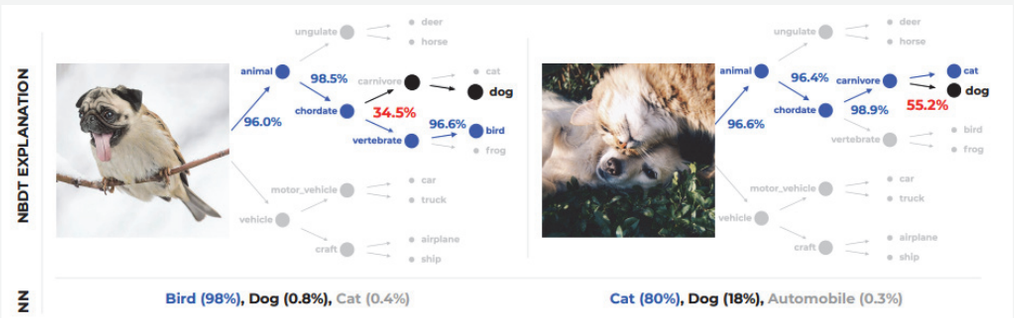
10-1b XAI 기술 적용이 불가능한 경우, 기법 적용 이외의 대안을 마련하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템의 서비스에 따라 XAI 기술의 적용이 어려우면, 개발자는 인공지능 시스템의 설명가능성을 높이기 위한 차선책을 고려해야 한다.
- XAI 기술로 설명이 어려운 경우에는 의사결정 트리와 같은 전통적인 방법의 도입을 고려할 수 있으며 [48], 현재 XAI의 한계, 더 나아가 기계학습의 한계를 보완하기 위해 원인 학습causal learning 기술도 활발하게 연구되고 있으므로 참고할 만하다.[49]

참고 의사결정 트리를 활용한 신경망Neural-Backed Decision Tree, NBDT 설명 시도

- 개요
  - ✓ 네트워크의 마지막 선형층을 의사결정 트리로 대체
  - ✓ 전통적인 의사결정 트리와 달리 추론 시에는 path probabilities를 사용하여 불확실한 중간 결정을 줄임
  - ✓ 이를 통해 학습된 모델의 가중치로부터 계층을 만들어 과적합을 피함
  - ✓ 학습은 계층적 손실을 사용하여 높은 레벨의 결정을 학습할 수 있게 함
- 학습 완료 모델 성능
  - ✓ Small-scale 데이터셋에서 기존 모델보다 약 1%의 성능 향상이 있었으며, 설명 가능한 특성도 보존됨
  - ✓ Large-scale 데이터셋에서 동일 backbone을 가진 기존 최고 성능과 비슷하거나 좋은 성능을 냄



모델의 성능을 해치는 애매모호한 데이터의 예시
– NBDT의 결정 중 엔트로피를 살펴본 결과, 애매모호함을 나타내기에 좋은 지표로 확인함

- 자체적인 설명가능성 평가
  - ✓ Saliency 설명과 NBDT의 설명 비교 시, 사람은 NBDT의 설명에서 더 정확하게 오분류를 찾음
  - ✓ NBDT의 엔트로피를 약간 수정하여 애매모호한 레이블을 탐지함
  - ✓ 이미지 분류 문제에서 사람들은 NBDT의 예측을 더 선호함
- 한계 및 시사점
  - ✓ 설명력이 부족할 수 있는 부분은 WordNet을 사용해 보완함
  - ✓ WordNet에 없는 계층은 설명하지 못함
  - ✓ 아직은 설명가능성 부분에서 정량적인 평가가 어려우나 자신들의 모델이 왜 설명력이 더 좋고 신뢰성이 높은지를 수치화함

10-2 인공지능 모델 상세 문서를 통해 모델의 명세를 투명하게 제공하는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

신뢰성을 확보하기 위해 인공지능 알고리즘 또는 모델이 반영된 시스템의 명세 정보를 투명하게 제공하고자 한다면 본 항목을 고려하여 만족 여부를 판단하십시오.

- 인공지능 시스템의 투명성을 확보하는 방안 중 하나는 인공지능 모델·서비스의 개발, 테스트 및 배포 과정에서 발생한 다양한 결과를 문서로 작성하는 것이다. 모델의 명세를 작성한 상세 문서를 확보해 두면 사용자가 인공지능 모델과 관련된 정보를 요구할 때 모델의 목적, 입출력 정보, 성능, 편향 여부, 신뢰도 등의 결과를 투명하게 공개할 수 있다.
- IBM의 AI FactSheet 360 프로젝트 및 WEF에서는 모델의 명세를 작성한 문서로써 인공지능 시스템의 투명성을 확보하는 방안을 제시하고 있다. 특히 IBM은 개발한 시스템의 알고리즘을 공개하지 않고 필요에 따라 인공지능 모델의 주요 정보 및 구성 요소를 설명할 수 있는 문서의 예시를 제공하고 있다.

10-2a 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 파악할 수 있는 정보를 제공하는 것은 시스템 신뢰성을 높이는 데 중요한 요소이다. 이를 위해 인공지능 모델 개발 과정에서 모델의 명세를 작성한 모델 상세 문서를 확보할 경우, 사용자에게 인공지능 시스템의 구성 요소를 파악할 수 있는 정보를 제공할 수 있다.
- 시민 또는 다수의 사용자가 사용하는 공공·사회 분야 성격이 강한 인공지능 서비스는 윤리적 관점과 사용자 요구에 좀 더 민감한 대응이 필요하므로 모델의 작동 방식 등 상세한 정보를 가능한 한 유지하고 관리할 수 있도록 한다. 다음은 서비스에 따라 모델 상세 문서가 필요한 상황의 예시이다.
  - ✓ AI 면접 시스템: A 기업과 B 기업의 동일 인공지능 시스템의 면접 결과가 서로 다르게 나오는 경우에 정보 공개 요청을 받을 수 있음
  - ✓ 재범 예측 시스템: 미국 COMPAS 시스템에서 백인이 아닌 사람에 대해 재범 예측 가능성을 더 높게 평가한 사례를 반영하여 해당 서비스에서의 확인 및 검증 내용을 사전에 기록하고 관리함
- 모델 상세 문서 작성 시, 인공지능 생명주기와 관련된 이해관계자를 고려하여 각자 필요한 정보를 선택하여 확인할 수 있도록 관련 정보를 포함해야 한다. 다음은 이해관계자에 따라 모델 상세 문서에 필요한 정보의 예시와 공공·사회 분야에 적용할 수 있는 IBM 오디오 분류기 모델의 모델 상세 문서의 사례이다.

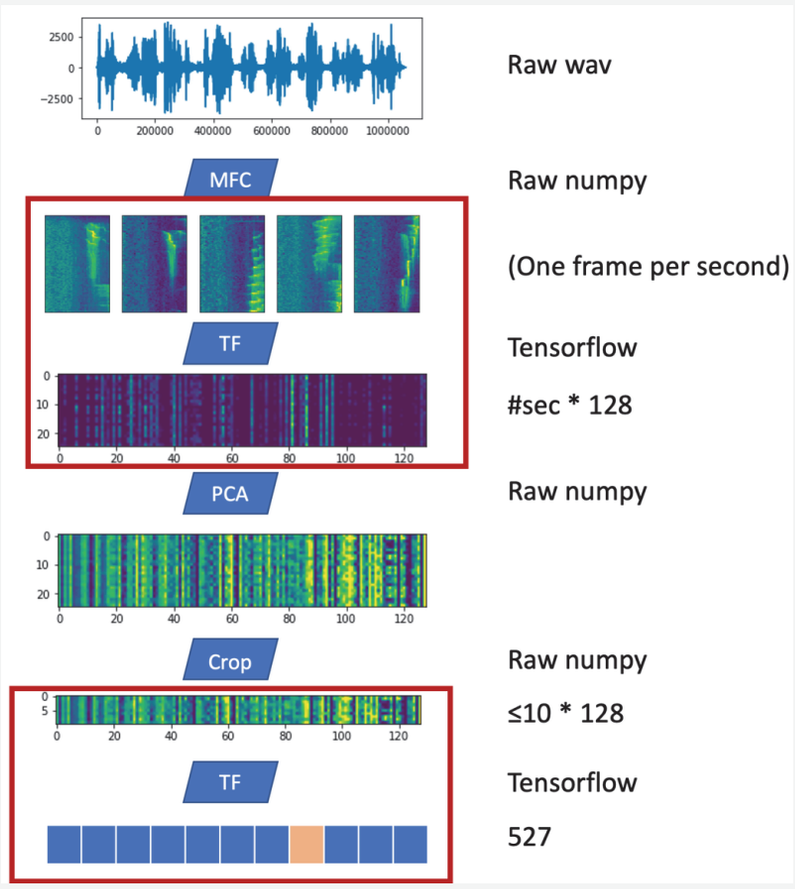
이해관계자	모델 상세 문서 정보
비즈니스 결정권자	전체 인공지능 시스템의 목적, 방향성, 시스템 내 서비스 명칭 및 서비스별 의도된 목적 등
데이터 과학자 및 시스템 개발자	학습에 사용된 데이터셋 명세 및 전처리 기법, 학습 모델 구성, 입출력 명세, 모델 학습 파라미터 등
모델 검증자	테스트 데이터셋 구성 정보 및 주요 테스트 성능, 편향, 신뢰도 등의 평가 결과
모델 운영자	모델 운영 및 모니터링 결과 측면의 성능평가 지표, 성능 저하 환경 요인, 최적 결과 도출 환경 등

참고 IBM 오디오 분류기 모델 상세 문서 사례(일부)[50]

- 개요
  - ✓ 이 문서는 IBM 개발자 모델 자산 eXchange의 오디오 분류자 모델과 함께 제공되는 FactSheet입니다. FactSheet는 공급 업체의 적합성 선언을 통해 AI 서비스에 대한 신뢰를 높이는 것을 목표로 하며 이 FactSheet는 오디오 분류기 모델을 교육하는 과정과 예상 결과 및 적절한 사용을 문서화합니다.
- 목적
  - ✓ 이 모델은 입력 오디오 클립을 분류합니다. 오디오 클립이 모델에 전달되고 모델은 클립에서 감지하는 상위 5개 클래스를 예측합니다. 오디오에 특정 오디오 클래스가 하나만 포함되어 있으면 + 4개와 밀접하게 관련된 클래스가 예측됩니다. 오디오에 여러 오디오 소스가 포함되어 있으면 최대 5개의 오디오 소스를 예측하려고 시도합니다.
  - ✓ 이 모델은 서명된 16비트 PCM wav 파일을 입력으로 인식하고, 임베딩을 생성하고, PCA 변환·양자화를 적용하고, 임베딩을 다중 주의 분류기에 대한 입력으로 사용하고, 상위 5개 클래스를 예측하여 확률을 출력합니다.
  - ✓ 이 모델은 현재 AudioSet Ontology의 일부인 527개의 클래스를 지원합니다. 클래스와 label_ids는 class_labels_indices.csv에서 찾을 수 있습니다. 이 모델은 Yu et al.의 논문 'Multi-level Attention Model for Weakly Supervised Audio Classification'에 설명된 대로 AudioSet에서 학습되었습니다.

- 중략 -

- 견고성
  - ✓ 목적
    - AI 및 ML 모델은 적은 양의 노이즈를 추가하여 출력이 변경될 수 있으며, 종종 인간에게는 인식되지 않거나 인간이 입력을 올바르게 분류하도록 하는 적대적 공격에 여전히 취약합니다. 이러한 테스트는 이러한 공격에 대한 모델의 취약성을 측정합니다.
  - ✓ 세부 정보
    - 오디오 분류기는 다음의 주요 구성 요소 5가지로 구성됩니다. 멜 주파수 분광기 특성 추출기(위의 MFC), 오디오의 각 초에 대한 오디오 임베딩을 생성하는 Tensorflow 모델, PCA 변환 및 자르기(최대 10 초까지), Tensorflow 분류 모델. 이러한 구성 요소 중 일부는 numpy와 같은 프레임워크에서 구현되며 그래디언트 계산을 지원하지 않으므로 화이트 박스 평가를 제공합니다. 우리는 엔드 투 엔드 성능을 측정하기 위해 블랙 박스 공격을 사용하여 경험적 평가를 수행합니다. 잡음은 원시 입력 오디오(16비트, 스테레오, 44.1kHz 입력)에 추가됩니다.



모델 구성 요소의 세부 설명

– 중 략 –

✓ 공격 매개 변수

– 평가 이미지 수: 18

* 모델과 함께 제공되는 스테레오 샘플

– 공격 유형: HSJ L-inf 및 L2 공격

* 샘플에 대한 최대 절대 변화가 최소화됨

– 최대 반복 횟수: 10

* HSJ 3단계 공격의 반복 횟수

– 최대 평가: 100

* 그래디언트를 추정하기 위한 최대 평가 수

– 초기 평가: 100

* 기울기를 추정하기 위한 초기 평가 수

– 초기 크기: 100

* 적대적 사례의 초기 생성에 대한 최대 시행 횟수

– 후략 –

10-3 필요 시, 인공지능 모델 추론 결과에 대한 설명을 제공하는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

신뢰성을 제공하기 위해 인공지능 알고리즘 또는 모델이 반영된 시스템의 추론 결과에 대한 설명을 함께 제공하고자 한다면 본 항목을 고려하여 만족 여부를 판단하십시오.

- 사용자에게 인공지능 모델의 추론 결과에 대한 설명을 제공하면, 사용자는 단순히 해당 인공지능 모델의 최종 결과뿐 아니라 그 결과가 도출된 수치적인 근거로 확률값, 불확실성uncertainty 등을 제공받을 수 있다. 이러한 정보는 사용자의 의사결정에 도움이 되지만, 오히려 사용자의 혼란을 유발할 수 있으므로, 정보 제공의 필요성을 사전에 검토하는 것이 필요하다.

10-3a 모델 추론 결과에 대한 설명이 필요한지 검토하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템이 도출한 결과에 대한 설명을 제공하는 것은, 사람들이 인공지능을 활용하여 의사결정하는 데 도움이 될 수 있지만, 오히려 방해될 수도 있다. 따라서 모든 경우에 모델의 추론 결과에 대한 설명을 제공하기보다는, 설명이 꼭 제공되어야 하는지를 확인하는 과정이 선행되어야 한다.
- 모델의 추론 결과에 대한 설명을 제공하지 않는 편이 더 나은 경우에 대한 두 가지 예시는 다음과 같다.
  - ✓ 첫째, 모델의 추론 결과에 대한 설명 제공 자체가 사용자의 의사결정에 크게 영향을 미치지 않을 것으로 판단되는 경우이다. 설명 제공으로 인해 미치는 영향을 명확하게 분석하지 않은 경우, 자세한 설명을 제공하면 사용자의 의사결정에 더 도움이 될 것으로 생각할 수 있지만, 예상과는 다르게 혼란을 초래할 수 있다. 예를 들어, 인공지능 시스템이 도출한 두 가지 결과가 있고, 각각의 예측 확률이 85.8%, 87.0%라면, 사용자는 어떤 결과를 활용하여 의사결정을 할지 혼란스러울 수 있다.
  - ✓ 둘째, 예측 확률이 너무 높거나 낮은 경우에도 모델의 추론 결과에 대한 자세한 설명을 제공하지 않는 것이 낫다. 만약 시스템의 추론 결과에 대해 예측 확률값이 100%라고 사용자에게 알릴 경우, 사용자가 시스템의 추론 결과를 맹목적으로 수용하게 만들 수 있다.

10-3b 사용자에게 인공지능 모델 추론 결과에 대한 설명을 제공하였는가?

Yes ☐ No ☐ N/A ☐

- 모델 추론 결과를 설명하기 위해서는 정밀도precision, 재현율recall, mAPmean Average Precision와 같은 지표와 함께 불확실성을 계산해야 한다. 불확실성이란 확률 변수의 분산 크기로서, 인공지능 모델이 도출한 결과를 얼마나 확신하는지를 나타내는 지표이다. 불확실성 추정 기법에는 베이지안 신경망Bayesian neural network, 앙상블ensemble, 드롭아웃dropout 등이 있다.

✓ 드롭아웃은 신경망 내의 노드와 각 노드 간 연결을 무작위로 선정하여 제거하는 기법이다.

✓ 드롭아웃을 적용한 신경망과 베이지안 신경망, 각각의 수행마다 다른 신경망이 생성된다는 특징을 활용하면 수행의 결과로 생성된 여러 신경망에 동일 입력값을 주고 그 결과로 얻은 출력값 여러 개의 평균과 분산을 계산할 수 있는데, 이때 계산한 분산이 불확실성이다.

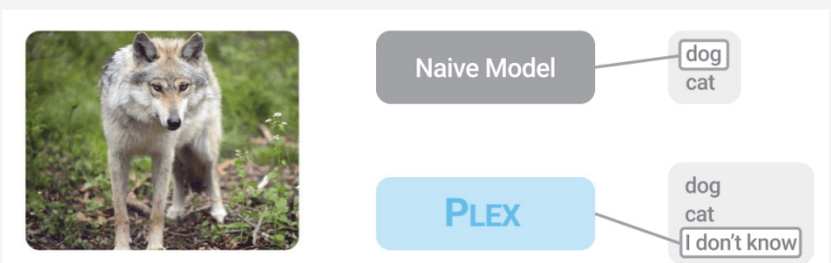
- 인공지능 모델의 출력 성능(예: 정밀도, 재현율)과 불확실성을 각각 계산해 나온 결과를 조합하여 모델 추론 결과를 설명할 수 있다. 예를 들어, 참과 거짓의 예측 모델이 있을 때 "모델의 예측 확률이 98%로 높고 예측의 불확실성이 1%로 낮으므로 '참'이라는 결과를 신뢰할 수 있다."라는 근거를 사용자에게 제시할 수 있다.
- 단, 예측 확률이 임계치보다 낮거나 불확실성이 높으면 사용자가 이를 인지할 수 있도록 모델 추론의 결과에 대한 설명을 반드시 제공해야 한다.
- 추론 결과의 임계치란 인공지능 모델의 성능 지표(예: 정밀도, 재현율)의 임계치, 그 성능의 불확실성에 대한 임계치로 나눌 수 있다. 추론 결과의 임계치 도출을 위해, 우선 인공지능 모델로 인해 발생할 수 있는 문제 상황을 정의하고, 문제 발생 여부를 결정짓는 중요 변수를 파악해야 한다. 여기서 문제 상황이란 사용자의 생명이나 재산에 위협이 되는 상황뿐 아니라 품질이 기대하는 또는 유지되어야 하는 수준보다 낮은 상황 등을 모두 포함한다.
- 임계치는 인공지능 모델을 통해 찾을 수 있다. 대표적인 기술인 LDA(Linear Discriminant Analysis), SVM(Support Vector Machine)부터 CNN(Convolutional Neural Network), LSTM(Long-Short Term Memory)을 비롯하여 비교적 최근에 발표된 GEN(Graph Extrapolation Network), SimCLR(Simple framework for Contrastive Learning of visual Representations) 등에 이르기까지 다양한 기법으로 추론 결과의 임계치를 도출할 수 있다.

참고 Google의 불확실성 접근 및 설명 제공 사례[51]

심층학습 모델은 비전, 언어 및 기타 양식, 대규모 사전 학습의 등장으로 인상적인 진전이 이루어졌으며, 학습 집합과 동일한 분포에서 가져온 테스트 데이터에 적용할 때 가장 정확하다.

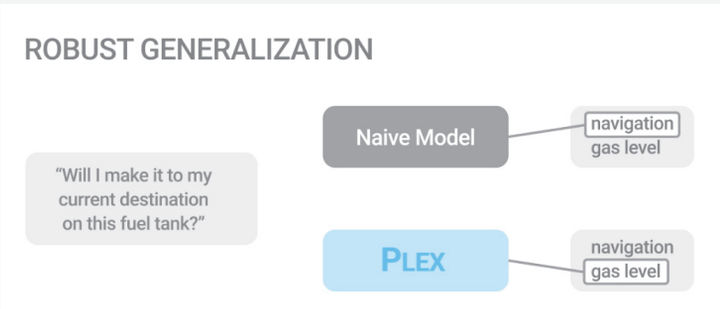
신뢰할 수 있는 기계학습 시스템에 대한 요구사항 3가지

1. 예측에 대한 불확실성을 정확하게 보고해야 한다.
   불확실성은 불완전하거나 알려지지 않은 정보를 반영하여 모델의 정확한 예측을 어렵게 함



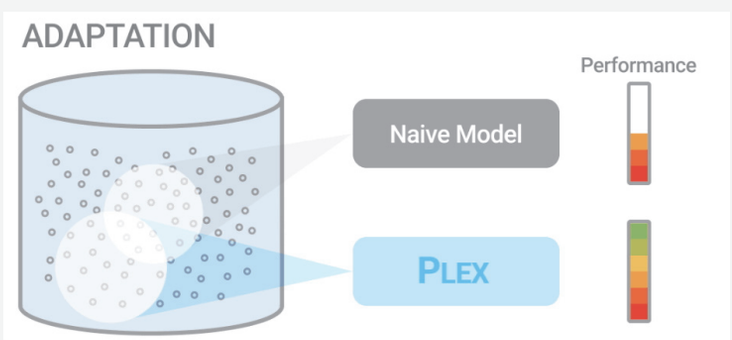
기존 모델(Naive Model)은 가장 비슷한 이미지로 답변하지만,
Plex AI는 확실한 데이터가 아니면 "I don't know"라고 출력함

2. 새로운 시나리오(배포 이동)로 강력하게 일반화해야 한다.
   강력한 일반화에는 보이지 않는 이벤트에 대한 추정 또는 예측이 포함되며 네 가지 유형의 분포 외 데이터를 조사함(공변량 이동, 의미론적 이동, 레이블 불확실성, 하위 인구 이동)



기존 모델은 가짜 상관관계에 민감한 반면 Plex AI는 질문에 대한 가능성을 확보하기 위해 gas level을 확인함

3. 새로운 데이터에 효율적으로 적응할 수 있어야 한다.
   적응은 학습 과정에서 모델의 능력을 조사하는 것을 말하는데, 벤치마크는 일반적으로 미리 정의된 학습-테스트 분할을 사용하는 정적 데이터셋에서 평가됨
   또한 성능을 보다 신속하게 개선하기 위해 학습하는 데이터를 능동적으로 선택하여 performance의 값을 높일 수 있음



• PLEX: 시각 및 언어를 위해 사전 학습된 대형 모델 확장
  ✓ 효율적인 앙상블 모델이 각각 예측을 수행한 다음에 집계되는 하위 모델을 기반으로 함
  ✓ 각 아키텍처의 마지막 선형 계층을 가우시안 프로세스gaussian process 또는 이질성heteroscedastic 예측 불확실성을 더 잘 표현하게 함

04 시스템 구현

다양성 존중

요구사항 11 인공지능 시스템 구현 시 발생 가능한 편향 제거

대표행위자 | 시스템 엔지니어 협력 대상 | 시스템 운영자 인공지능 모델 개발자

- 인공지능 시스템 구현 단계에서 편향을 고려하지 않는다면, 시스템 설계자 또는 개발자의 배경지식이나 편견으로 인공지능 시스템이 편향될 수 있다. 따라서 발생 가능한 편향을 식별하고 이를 제거하는 방안을 고려하여 설계한다.

11-1 소스 코드 및 사용자 인터페이스로 인한 편향을 제거하기 위해 노력하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

인공지능 시스템의 구현 시, 구현 방식 및 사용자 인터페이스가 중요한 경우에 본 항목을 고려하여 만족 여부를 판단하십시오.

- 공공·사회 분야 인공지능 시스템은 다양한 사용자층에 공개되어 있으므로 특정 선택을 암묵적으로 유도하는 사용자 인터페이스 등을 통한 편향에 주의해야 한다.
- 인공지능 시스템과의 인터페이스 및 상호작용 측면에서 사용자의 이해도, 호감도, 신뢰도 등의 관점에서 표현 편향presentation bias이나 순위 편향ranking bias 등이 발생하는지를 미리 확인하여 편향을 방지할 수 있도록 시스템을 설계하는 것이 바람직하다.
- 그 외에 인공지능으로 작성된 코드를 주기적으로 검토하여 코드 구현 과정에서 특정 클래스 접근이 누락됐는지, 개발자의 편견이 코드에 반영되지 않았는지 등을 확인해야 한다. 이를 위해 인공지능 추론 결과를 사용자에게는 직접 표시하지 않고 시스템에서 필터링을 거친 후에 출력하는 방법 등을 이용할 수 있다.

11-1a 데이터 접근 방식 구현과정 등 소스 코드에서의 편향 발생 가능성을 확인하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템은 모델에서 활용할 데이터에 접근하는 방식이 코드상에 구현되는 과정에서 특정 클래스의 접근이 누락되는 등 다양한 형태의 편향이 발생할 수 있다.
- 공공·사회 분야 서비스를 위한 인공지능 기반 시스템 구축 시(예: 복지 대상자 선정, 인공지능 면접 등), 입력되는 이미지, 텍스트 또는 음성 데이터의 전처리 과정에서 데이터 처리 전문가의 지식을 기반으로 처리 규칙을 적용하는 경우에 추론 결과가 특정 결과를 더 많이 내게 하는 등 잠재적으로 인지 편향cognitive bias을 일으킬 수 있다. 시스템의 소스 코드에서 발생할 수 있는 인지 편향의 예시는 다음과 같다.

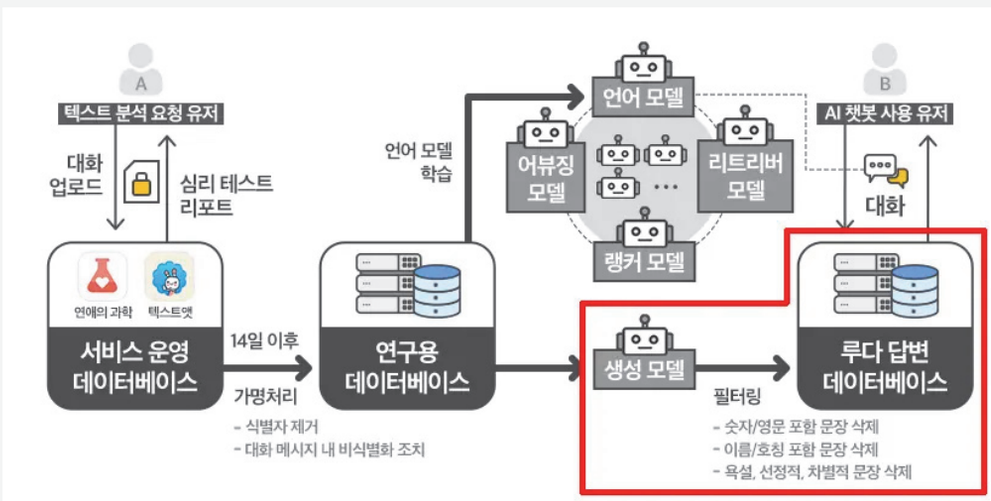
✓ 범죄 또는 재범 예측 시스템에서 특정 범죄 지역의 특성, 범죄자의 특성에 대해 알고 있는 전문가 개인의 판단 기준을 시스템에 적용하여 개발

✓ 멀티 모달 데이터 기반의 사회적 약자 도우미 시스템에서 상호작용 대상 사용자의 감정, 활동 상태 등을 알고 있는 전문가의 기준에 따라 판단하여 시스템의 동작을 결정

• 따라서 시스템의 편향 발생을 줄이기 위해서는 배경지식과 경험이 다양한 전문가를 선정하는 것이 도움이 된다.

참고 시스템 구현 시, 소스 코드에서의 편향 발생 가능성 확인[52]

• 답변 데이터베이스를 따로 활용하여 시스템의 편향 발생 가능성 사전 필터링



✓ 스캐터랩의 AI 챗봇은 언어 모델을 통해 답변을 미리 생성하고, 이를 검토하고 필터링하여 개인정보 또는 개인정보처럼 보이는 문장, 욕설, 선정적·차별적 문장을 사전에 삭제함

✓ 이와 유사하게, 시스템을 통해 출력되는 결과를 제어하여 편향 발생 가능성을 사전에 차단하는 방법을 적용할 수 있음

11-1b 사용자 인터페이스 및 상호작용 방식으로 인한 편향을 확인하였는가?

Yes ☐ No ☐ N/A ☐

- 공공·사회 분야 서비스를 위한 챗봇 또는 음성인식 인공지능 기반 시스템에서 인공지능 에이전트의 성별, 목소리 높낮이 등으로 인해 다음 사례와 같이 받아들이는 결과에 차이가 발생할 수 있으므로 사용자의 인터페이스 및 상호작용 방식으로 인한 편향을 확인해야 한다.
- 사용자의 상호작용 편향을 방지하기 위해서는 사용자의 인터페이스 설계 및 구현 시에 편향 발생 가능성이 있는 요소(예: 표현 편향, 순위 편향)를 미리 인식해 제거해야 한다.
  - ✓ 표현 편향: 정보가 표현되는 방식에 따라 발생하는 편향이다. 예를 들어, 사용자는 제품 사용 시에 보이는 콘텐츠만 클릭할 수 있으므로 표시된 콘텐츠에서는 클릭이 발생하고 다른 콘텐츠에서는 클릭이 발생하지 않는다. 이러한 사용자의 인터페이스로 인해 특정 콘텐츠의 클릭만 유도될 수 있다.
  - ✓ 순위 편향: 정보가 노출되는 순서에 따라 발생하는 편향이다. 사용자는 최상위 결과가 가장 관련성이 높고 중요하다고 생각하는 경향이 지배적이므로 사용자의 선택 빈도는 상위에 노출된 결과가 하위에 노출된 결과보다 높을 수 있다.

참고 인공지능 에이전트의 목소리 성별과 높낮이에 대한 신뢰도 등 사용자의 경험을 조사한 사례[53]

시스템에 인공지능 에이전트를 활용할 때, 인공지능 에이전트의 안내 목소리 성별, 목소리 높낮이 등에 따라 듣는 사람이 이해도, 신뢰도, 호감도 측면에서 얼마나 다르게 받아들이는지를 조사하여 분석한 결과이다.

- 여성 고음, 여성 저음, 남성 고음, 남성 저음으로 나누어 이해도, 신뢰도, 호감도 분석
- 날씨를 묻는 상황(상대적으로 가벼운 주제)
  - ✓ 여성 목소리가 이해도, 신뢰도, 호감도 측면에서 모두 높게 평가됨
  - ✓ '흐린 날씨' 상황에서 저음 목소리가 신뢰도 면에서 높에 평가됨
    - 흐리고 미세먼지 있는 날씨를 전달하기에는 차분하고 진중한 느낌의 저음이 더 믿음을 준 것으로 보임
- 뉴스를 묻는 상황
  - ✓ '훈훈한 뉴스'는 모든 측면에서 여성의 목소리가 높게 평가됨
  - ✓ 보다 진지한 '무거운 뉴스'는 남성의 목소리를 선호하는 것으로 나타남
    - 많은 참여자가 무거운 내용일수록 저음에서 신뢰감을 더 느끼고 집중이 잘되었다는 의견을 주었기 때문에 단순히 성별의 영향으로 보기는 어려움

책임성 안전성 투명성

요구사항 12 인공지능 시스템의 안전 모드 구현 및 문제발생 알림 절차 수립

대표행위자 | 시스템 엔지니어 협력 대상 | 시스템 운영자 인공지능 모델 개발자 품질 관리자

- 인공지능 시스템을 통해 생성되는 결과나 의사결정은 개인 혹은 사회에 부정적인 영향을 미칠 수 있으므로, 이에 대한 대응이 가능하도록 안전 모드를 구현하고, 문제발생 알림 절차를 수립한다.

12-1 공격, 성능 저하 및 사회적 이슈 등의 문제 발생 시 대응 가능한 안전 모드를 적용하는가?

Yes No N/A
☐ ☐ ☐

해당여부 판단

공공·사회 분야의 인공지능 서비스 중에서 문제 발생 시의 대응이 중요한 경우에 본 항목을 고려하여 만족 여부를 판단하십시오.

- 공공기관은 인공지능 시스템을 구현하기에 앞서, 인공지능 시스템의 활용에 따른 위험(피해 확률, 심각성)을 평가해야 한다. 또한 평가 결과에 따라 인공지능 시스템의 고장 및 오작동 시에 예상되는 문제 상황과 원인을 분석하고 예외 처리 및 대응 방안을 마련해야 한다. 이처럼 대처를 위한 방법이 작동하는 상태를 안전 모드라고 하며, 이를 구현하는 방법과 예시는 다음과 같다.
  - ✓ 시스템에 문제 발생 시에 기능 정지 및 피드백 제공 화면으로 전환
  - ✓ 시스템에 문제 발생 시에 서비스 제공 초기 화면 혹은 상태로 복구
  - ✓ 인공지능 판단 결과의 불확실성이 높거나 문제 발생 가능성이 큰 경우, 이에 대한 의사결정을 회피하거나 사용자에게 상황에 대한 안내 제공
  - ✓ 사용자의 악의적인 의도를 파악하고 이에 대한 입력을 거절
  - ✓ 자동 및 자율 운영 중 시스템에 문제가 발생하면 사람의 개입 유도
  - ✓ 예상되는 사용자 오류에 대해 안내 및 대응 제공

12-1a 문제 상황에 대한 예외 처리 정책이 마련되어 있는가?

Yes No N/A
☐ ☐ ☐

- 시스템에 문제가 발생하는 상황에서 기능 정지, 화면 전환 및 서비스 제공 초기 상태로의 복구, 입력 거절, 의사결정 회피 등의 예외 처리가 이루어지는지 확인해야 한다.

- 이러한 예외 처리가 이루어지는 경우, 인공지능 시스템 사용자에게는 시스템 운영이 적절치 않은 이유와 시스템의 대응에 대하여 설명을 제공해야 한다.
- 공공기관은 인공지능 시스템의 위험평가 결과를 기반으로 인공지능 시스템에 대한 사람의 감독 또는 의사결정 개입의 수준을 결정해야 한다. 관련된 내용은 12-1c 를 참고한다.
  - ✓ 인공지능 판단 결과의 불확실성이 높거나 문제 발생 가능성이 큰 경우, 이에 대한 의사결정을 회피하거나 사용자에게 상황에 대한 안내 제공
  - ✓ 자동 및 자율 운영 중 시스템에 문제가 발생하면 사람의 개입 유도
- 이 외에도 공공기관에서 운영하는 시스템은 행정안전부에서 제시하는 정보시스템 구축·운영 지침 및 개발보안, 개인정보보호 등의 표준을 확인해서 적용해야 한다.

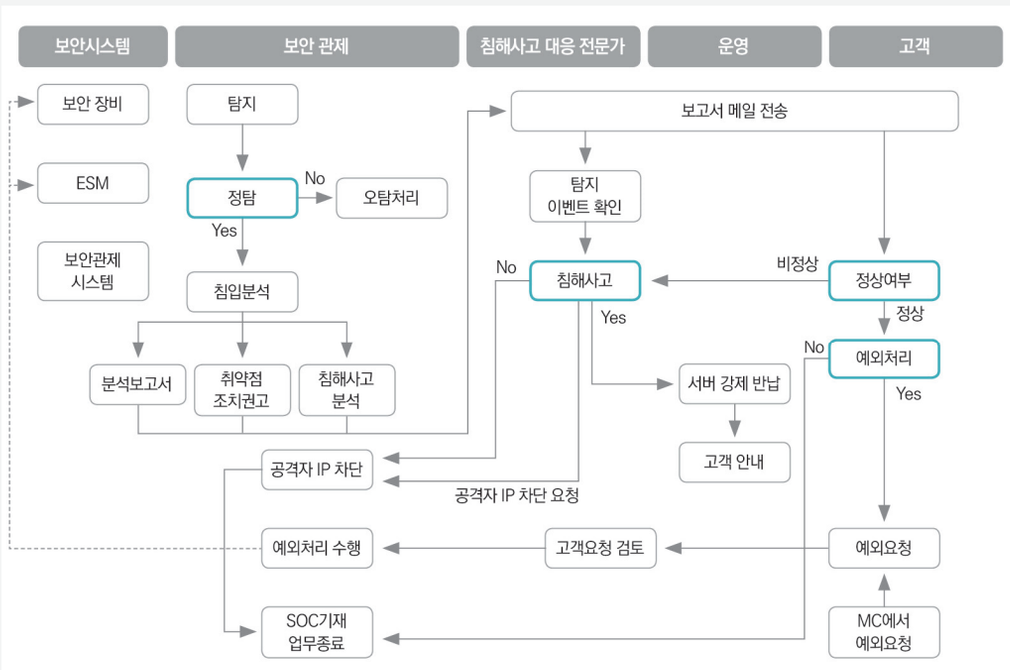
참고 공공기관에서 운영하는 시스템 관련 규정 중 예외처리 관련 내용 예시

- 다음은 공공기관에서 운영하는 시스템 구축 시에 준수해야 하는 규정 및 가이드이다.
- 인공지능에 대한 직접적인 내용은 없지만 인공지능 기술이 반영되어 운영되는 최종 서비스에서 준수해야 하는 항목이 있을 수 있으므로 시스템 구현 시에 참고한다.

규정 및 가이드	부처	예외 처리 관련 내용
소프트웨어 개발보안 가이드(2019. 11.)	행정안전부	• 에러 또는 오류 상황을 처리하지 않거나 처리되어 중요한 정보의 유출 등 보안 약점이 발생하지 않도록 설계
소프트웨어 보안약점 진단가이드(2019. 6.)	행정안전부	• 명시적인 예외인 경우에 예외처리 블록을 이용하거나 예외 발생 시 수행해야 하는 기능을 구현
공직자통합메일 지능형 고객상담 서비스 구축(2020. 3.)	문화체육관광부	• 예외에 대한 부적절한 처리로 인해 의도하지 않은 상황이 발생될 수 있는 보안약점
전자금융과 금융보안 (2021. 3.)	금융보안원	• 이상 현상 또는 예외 발견 시 신규 애플리케이션을 설치하도록 고객에게 안내
소프트웨어 개발보안 가이드(2021. 11.)	행정안전부	• 런타임 예외의 경우, 입력 값의 범위를 체크하여 애플리케이션이 정상적으로 동작할 수 있는 값만 사용되도록 보장해야 함
웹응용프로그램 개발보안 가이드(2010. 1.)	행정안전부	• 에러 메시지를 특정 URL로 리다이렉트 또는 예외 호출을 설정함 • 히든 필드 값을 그대로 사용하지 말고, 데이터베이스에서 재검색하여 값을 새로 얻어 오거나 히든 필드로 전송된 값들을 검증하도록 소스를 수정함
홈페이지S/W(웹) 개발보안 적용가이드(2012. 11.)	행정안전부	• 개인정보 등 중요 정보를 보호하기 위해 사용하는 암호 알고리즘 적용 시, IT보안인증 사무국이 안전성을 확인한 검증필 암호모듈 사용해야 함
소프트웨어사업 요구사항 분석적용 가이드(2021. 2.)	산업통상자원부	• 사용자가 입력한 데이터 형식의 모든 오류는 사용자가 시스템에 그 정보를 입력한 지 2초 이내에 적당한 오류 메시지를 사용자에게 제시해야 함

참고 보안관제 서비스의 예외 처리 사례[54]

- 네이버 Security Monitoring 서비스는 보안 솔루션의 구축부터 운영과 실시간 보안 관제 서비스까지 통합으로 제공해 주는 서비스이다. 다음 그림과 같이 프로세스 자체에 예외처리를 하여 오류 및 위험이 없도록 설계하였다.



Security Monitoring 대응 프로세스

12-1b 인공지능 시스템의 보안 강화를 위한 보안 기법을 적용하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템을 개발할 때 격리 및 탐지 등 보안 기법을 활용한 인공지능 보안 아키텍처와 구축 솔루션을 적용함으로써 인공지능 데이터 및 모델에 대한 보안성뿐만 아니라 인공지능 시스템의 전반적인 보안성을 확보할 수 있다.
- 공공·사회 분야의 서비스 제공을 위해 이미지, 음성, 텍스트 등을 활용한 인공지능 시스템을 적용하였을 때 발생할 수 있는 보안 위협 등을 고려해야 한다. 특히, 민원 업무 등 많은 부분에서 자동화 시스템(예: 챗봇을 이용한 민원 대응)을 도입하면서 인공지능 시스템 자체를 대상으로 하는 공격뿐 아니라 이를 악용한 악성 시스템을 통해 개인정보 탈취 등의 공격도 발생하고 있다. 열악한 보안 정책 및 시스템으로 인해 발생한 보안 사고 사례는 다음과 같다.

✓ 2017년, 해커들이 델타항공 챗봇 시스템에 액세스하여 소스 코드를 수정함으로써 항공사 웹사이트에서 고객 수십만 명의 개인 데이터와 결제 카드의 세부 정보를 탈취함[55]

✓ 녹음한 공격 대상자의 음성을 바탕으로 음성을 합성 및 변환하여 음성인식 서비스를 공격한 후 개인정보(통화목록 및 개인 건강관리 앱 등)에 접근할 수 있는 것을 확인함[56]

• 공공·사회 분야 서비스 제공 시에는 일반적으로 항상 서비스를 구동하고 있어야 하므로 외부의 공격에 그만큼 취약할 수밖에 없다. 인공지능 시스템은 많은 경우에 클라우드 또는 자체 서버를 통해서 서비스를 외부에서 활용할 수 있는 형태로 제공되기 때문에 종래에 알려진 사이버 보안 위협에 대응할 수 있도록 시스템을 구축하는 것이 바람직하다. 이를 위해 다음과 같은 보안 기법을 적용할 수 있다.

✓ 인증 및 권한 부여: 인공지능 시스템과 상호작용하는 동안 개인인증 보안 계층이 사용자를 확인

✓ 종단간 암호화: RSA 알고리즘과 같은 다양한 암호화 기법을 적용

✓ 자체 파괴 메시지: 민감한 개인 식별 정보가 전송되면 인공지능 시스템에서 설정 기간 이후에 자동 삭제

• 또한 많은 인공지능 시스템이 외부 접속을 허용하도록 서비스하므로 주기적으로 웹 보안 취약점을 진단하고 점검함으로써 사전에 보안을 강화할 수 있다.

참고 인공지능 시스템 공격 시나리오의 예시

• 챗봇의 잠재적 무기화 – 스키밍skimming 시나리오[57]

✓ 공격자가 유효한 대화에 새 필드를 쉽게 삽입할 수 있음

✓ 다음과 같은 스키밍 시나리오 발생 가능

1. 필요한 고객 이름 및 주문 번호를 요청하는 것 이상으로 일반적인 주문 상태 조회 대화 상자를 확장하여 주문에 사용된 신용카드 세부 정보 및 우편 번호 요청
2. 스키밍된 정보는 챗봇 내의 메모리 슬롯에 저장될 수 있음
   – 즉, 공격자가 수집된 데이터의 덤프를 트리거하도록 설계된 미리 결정된 '코드 워드'를 제공하여 채팅 채널을 통해 수집된 모든 정보를 유출할 때까지 챗봇 애플리케이션 메모리에만 상주할 수 있음을 의미

✓ 수집된 데이터가 디스크에 저장되지 않고 채팅 채널을 통해 난독화된 데이터를 사용하여 유출되는 것은 조직에서 이러한 공격 상황을 한동안 감지하지 못한다는 것을 의미할 수 있음

• 서비스 거부DoS, Denial of Service

✓ 인공지능 시스템은 자연어 이해 알고리즘과 같은 심층학습이 관련된 경우에 높은 컴퓨팅 성능이 필요함

✓ 서비스 거부 공격은 설계된 목적으로 리소스를 사용할 수 없도록 하는 데 중점을 두고 있음

✓ 인공지능 시스템이 매우 많은 수의 요청을 받으면 합법적인 사용자가 더는 사용하지 못하게 될 수 있음

12-1c 인공지능 시스템의 의사결정으로 인한 파급효과가 크고 불확실성이 높은 경우, 사람의 개입을 고려하였는가?

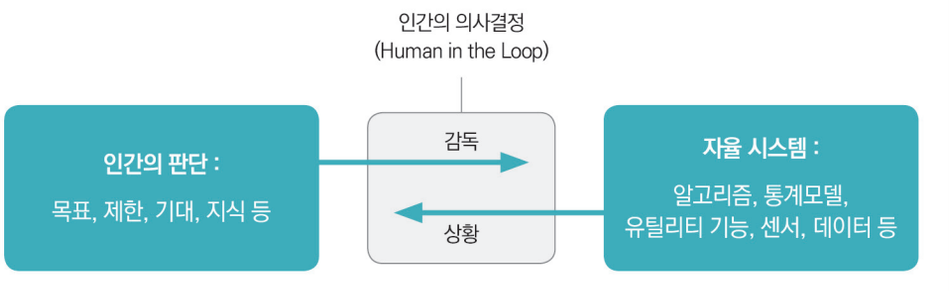
Yes ☐ No ☐ N/A ☐

- 인공지능 시스템이 인공지능 모델의 판단 결과를 활용하여 시스템 동작을 제어하거나, 사람의 안전 및 환경에 영향을 줄 수 있는 정보를 제공하는 경우, 사람의 개입이 필요한 경우가 있다. 이는 인공지능 시스템의 동작 및 기능의 파급효과가 크지만, 인공지능 모델이 도출한 판단 결과의 불확실성이 높은 경우이다.
- 특히, 인공지능 모델을 활용하여 자동 및 자율적으로 운영되는 시스템에서 이러한 경향이 두드러지며, 예외 처리 및 보안 기법 외에, 사람이 직접 혹은 부분적으로 개입하여 인공지능 모델의 불확실성을 해소하는 방안을 고려해야 한다.
  - ✓ 사람은 몰라도 됨human-out-of-the-loop
  - ✓ 사람이 모니터링human-over-the-loop
  - ✓ 사람이 중추 역할human-in-the-loop
- 피해 확률 및 심각성이 높은 경우에는 최종 의사결정에 사람이 개입하고 있으며, 피해 확률 및 심각성이 낮은 경우에는 사람이 개입하지 않는지를 확인한다.

참고 사람의 의사결정과 사회의 의사결정[8]

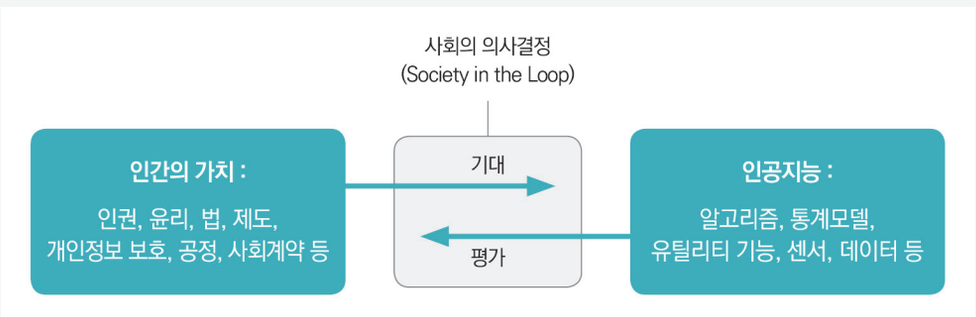
사람의 의사결정
(Human in the Loop)

- 인공지능 개발자의 감독하에 인공지능 의사결정의 기준을 정하고 신뢰도 평가를 수행한다.
- 신뢰도 평가 결과에 따라 필요한 조치를 한 후에 테스트를 반복하여 알고리즘을 완성하므로 소수 개발자의 사회적 편견이 알고리즘에 반영되는 잠재적 문제가 존재한다.



사회의 의사결정
(Society in the Loop)

- MIT Lab 소속의 라완이 '인간의 의사결정'의 대안으로 제시한 방법으로서 시민의 보편적 동의가 알고리즘에 반영되며 사회와 기술의 공진화를 추구한다.
- 인간사회에 통용되는 권리, 윤리, 법, 규범, 프라이버시, 공정성, 사회계약 등의 가치에 따라 인공지능 알고리즘의 개발 및 평가가 이루어진다.



참고 싱가포르 정보통신미디어개발청Infocomm Media Development Authority의 위험평가 매트릭스[8]

- 싱가포르 정보통신미디어개발청은 한 개인에 대한 기관의 의사결정에 대한 위험을 평가하는 매트릭스를 제안하고, 이를 바탕으로 의사결정에 필요한 사람의 참여 수준을 식별한다.

참고 아마존 알렉사의 음성인식 자동주문 사례

- CBS DFW 보도에 따르면, 미국 텍사스주에 사는 6세 아이가 인공지능 스피커 알렉사에게 “나랑 인형의 집 놀이 하자. 인형의 집 사 줘.”라고 말했다. 알렉사는 바로 반응했고, 아마존에서 ‘인형의 집’을 구입해 쿠키 4파운드와 함께 배달되게 했다.
- 이 사례는 샌디에이고 지역 방송국 CW6의 아침 TV쇼에 소개되었는데, 방송 진행자는 “작은 꼬마 아이가 ‘알렉사, 인형의 집 사줘’라고 말한 점이 너무나 귀엽다.”라고 언급했다. 몇 시간 후 CW6 채널 뉴스는 이 방송 진행자가 한 말을 들은 샌디에이고 지역 각 가정의 아마존 에코 기기가 저마다 아마존에서 인형의 집을 주문하는 소동이 벌어졌다고 보도했다.
- 아마존 에코의 설정에서 음성 주문을 금지하거나 인증 코드를 확인하도록 설정할 수 있다. 하지만 이런 조치를 하지 않은 채로 앵커의 멘트를 들은 아마존 에코가 주문을 진행한 것이다. 아마존은 이 사고로 주문된 인형의 집은 취소하고 환불하였다.

12-1d 예상되는 사용자 오류에 대한 안내 및 대응을 제공하는가?

Yes No N/A
☐ ☐ ☐

- 사용자 오류는 외적으로는 서비스 최종 결과물을 사용하는 사용자에게서, 내적으로는 서비스 결과 생성을 위해 내부 시스템을 사용하는 작업자에게서 비롯된다. 따라서 서비스 담당자는 다음과 같은 사용자 오류 유형을 이해하고 이와 관련되어 발생할 수 있는 오류를 사전에 정의하고 분석해야 한다.
  - ✓ 누락 오류: 수행해야 할 작업을 누락하여 발생하는 오류
  - ✓ 작위 오류: 수행해야 할 작업을 부정확하게 수행하여 발생하는 오류
  - ✓ 순서 오류: 수행해야 할 작업 순서를 틀리게 수행하는 오류
  - ✓ 시간 오류: 수행해야 할 작업을 정해진 시간 내에 완수하지 못하여 발생하는 오류
  - ✓ 불필요한 수행 오류: 작업 완수에 불필요한 작업을 수행할 때 발생하는 오류
- 사용자 오류에 따른 사전 대응 방안의 예시는 다음과 같다.
  - ✓ 제약조건 설정: 잘못된 사용자 입력을 막기 위해 사용자의 선택을 어느 정도 제약시키거나 수용 가능한 옵션을 정의하여 보여주는 것을 말한다. 예를 들어 인공지능 기반 상담 챗봇의 경우, 사용자의 자유로운 질문보다는 실제 많이 질의 되는 질문 목록을 먼저 제공하고 사용자가 선택하도록 한다.
  - ✓ 시스템 제안·정정: 자주 발생하는 사용자의 실수를 수집하고, 실제 서비스 시 유사한 사용자 실수가 발생한다면, 시스템에서 자동으로 정정하거나 올바른 입력을 제안한다. 예를 들어 검색 시 오탈자가 날 경우, 정정하여 추천하는 것을 예로 들 수 있다.
  - ✓ 기본값 설정: 시스템에서 필수이며 자주 사용되는 값을 기본값으로 먼저 제공하거나 관련 예시를 제공하여 사용자 실수를 줄일 수 있다.
  - ✓ 재확인·결과제공·실행취소: 사용자에게 전달받은 입력 등을 재차 확인하고 그에 대한 예상 결과를 미리 전달한다. 또한 잘못된 결과에 대해 실행을 취소하는 등의 기능을 포함하여 예방할 수 있다.

참고 인공지능 스피커의 사용자 오류 및 대응 방안 예시[59]

발생 원인에 따른 오류 상황 유형 및 대응 내용

발생 원인	상세 유형	대응 유형	대응 내용
환경적 노이즈	소음 또는 목소리가 겹침 등	A	두 가지 이상의 소리 겹침
어휘	잘못된 발음	B	시스템에 등록되지 않은 단어임
	개인화된 고유명사, 방언, 축약어, 은어 등	B	시스템에 등록되지 않은 단어임
	문법 오류, 부적절한 의미로 사용	B	시스템에 등록되지 않은 단어임
문장 구조	긴 문장	C	질문이 복잡하여 인식 불가능
	문법에 맞지 않는 문장 구조	D	질문하신 문장이 문법에 맞지 않음
질문 내용	두 가지 이상의 질문 동시 수행	C	질문이 복잡하여 인식 불가능
	시스템에서 제공하지 않는 기능에 대한 질문	E	수행이 불가능함
질문 번복	질문 중 질문 내용 수정 및 번복	C	질문이 복잡하여 인식 불가능

출처: AI 스피커 음성 상호작용 오류 상황에서의 사용자 감성 평가 분석

오류 대응 유형별 오류 상황 및 메시지 설계

오류 대응 유형	오류 대응 내용	오류 상황 사용자 발화
A	두 가지 이상의 소리 겹침	인기 동요 틀어줘.(다른 소리 섞임)
B	시스템에 등록되지 않은 단어임	난 자만추를 원해.
C	질문이 복잡하여 인식 불가능	오늘 날씨랑 습도랑 강수량이 어떻게 돼?
D	질문하신 문장이 문법에 맞지 않음	날씨 참 좋다. 오늘
E	수행 불가능함	요리해 줘!
F	질문과 관련된 정보가 없음	무한도전 시즌 2는 언제 해?

개선 메시지	적용한 프레임워크
앗! 두 가지 이상의 소리가 섞여서 잘 듣지 못했어요. 다시 한번 말씀해 주세요.	Human Helpful
음… 아직 학습하지 못한 단어가 있는 것 같아요. 다시 한번 말씀해 주시겠어요?	Human Helpful
죄송해요. 아직 긴 문장을 이해하지 못해요. 조금만 짧게 말씀해 주실 수 있나요?	Human Helpful
죄송해요. 제가 이해하지 못했어요. 다시 한번 말씀해 주세요.	Human Helpful
음… 도와드리고 싶지만, (요리)는 제가 할 수 있는 일이 아니에요.	Human Helpful
죄송해요. 아직(무한도전 시즌2)에 대해 학습하지 못했어요.	Human Helpful

출처: AI 스피커 음성 상호작용 오류 상황에서의 사용자 감성 평가 분석

12-2 인공지능 시스템에서 문제가 발생할 경우, 시스템은 이를 운영자에게 전달하는 기능을 수행하는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

공공·사회 분야의 인공지능 서비스 중에서 예측·판단 인공지능이 적용되는 서비스 또는 대민 서비스의 경우에는 본 항목을 고려하여 만족 여부를 판단하십시오.

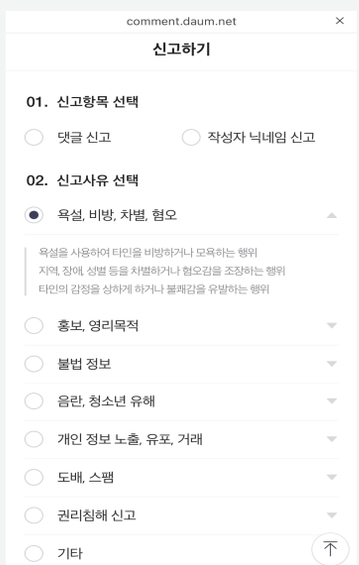
- 인공지능 시스템은 서비스 도중 외부의 공격, 사용자의 오용 등 다양한 요인으로 편향이나 성능 저하 등이 발생할 수 있으므로 시스템 운영자가 이를 파악할 수 있도록 시스템의 자체적인 점검 기능이나, 사용자가 운영자에게 관련 의견을 전달할 수 있는 기능을 제공해야 한다.
- 특히 공공·사회 분야의 인공지능 서비스는 모든 사용자에게 공정하게 제공해야 하므로 문제 발생 시 이를 운영자에게 전달하는 기능을 지니는 것이 매우 중요하다.

12-2a 편견, 차별 등 윤리적 문제에 대한 알림 절차를 수립하였는가?

Yes ☐ No ☐ N/A ☐

- 공공·사회 분야의 인공지능 시스템이 사회적 편견 또는 차별에 해당하는 의사결정일 경우에 시스템에서 이를 감지하는 방안을 마련해야 하며 사용자가 발견했다면 운영자에게 신고할 수 있는 기능도 개발되어야 한다.
- 이 외에도 사회적 편견 또는 차별 요소를 감지할 경우에 대응 절차를 마련하고, 리포팅 시 행위자에게 관련된 모든 정보를 즉시 제공하도록 해야 한다.

참고 카카오의 뉴스 차별 및 혐오성 댓글 신고 기능



- 카카오는 2020년 2월 26일 포털 다음(Daum)과 카카오톡 #탭의 뉴스 댓글 서비스에서 이용자의 자발적인 참여와 선한 영향력을 바탕으로 건강한 커뮤니케이션 생태계를 만들기 위한 개편했다.
- 댓글 신고 기준에 '차별/혐오' 항목을 추가하고, '덮어두기', '접기' 등 댓글 영역의 노출을 관리하는 기능을 신설했다.

12-2b 시스템 성능 저하를 평가하기 위한 지표 및 절차를 설정하고 알림 절차를 수립하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템의 경우, 서비스 배포 및 운영 단계에서 일반적인 소프트웨어와 달리 지속적인 데이터 축적, 서비스 기능 확장, 환경의 변화 등의 이유로 성능 변화가 생길 수 있다.
- 인공지능 시스템은 실제 서비스 운영 중 갑자기 성능이 저하됐을 때 원인을 바로 알기 어려우므로, 시스템의 성능 저하를 지속해서 평가, 관리하기 위한 지표와 절차가 설정되었는지 점검할 필요가 있다.
- 공공·사회 분야 인공지능 서비스에 적용할 수 있는 대표적인 성능 지표로는 F1-score, PPL[Perplexity], BLEU[Bilingual Evaluation Understudy], METEOR[Metric for Evaluation of Translation with Explicit ORdering], SSA[Sensible and Specificity Average] 등이 있다. 평가한 결과, 성능 저하가 확인되면 이를 시스템 운영자에게 보고하고 운영자는 성능 저하의 원인을 찾아 개선하는 절차를 마련해야 한다.

참고 인공지능 성능의 평가 방법

F1-score[60]
분류 모델에서 사용되는 기계학습 평가 지표[metric]이다.
모델의 성능을 측정하는 데 있어 정밀도와 재현율이 유용하게 사용되지만, 모델이 얼마나 효과적인지를 설명할 수 있는 한 가지 지표가 더 필요할 때 사용된다.

PPL[61]
텍스트 생성[text generation] 언어 모델의 성능평가 지표이다.
일반적으로 데이터셋이 신뢰도가 충분히 높을 때 perplexity 값이 낮을수록 언어 모델이 우수하다고 평가된다.
Perplexity는 언어 모델의 분기계수[branching factor]이다. 분기계수란 tree 자료 구조에서 branch의 개수를 의미하며 한 가지 경우를 골라야 하는 작업에서 선택지의 개수를 뜻한다.

BLEU[62,63]
기계 번역 결과와 사람이 직접 번역한 결과가 얼마나 유사한지를 비교하여 번역기의 성능을 측정한다.
BLEU는 주로 번역하는 모델에 사용된다.
1) n-gram을 통한 순서쌍들이 얼마나 겹치는지를 측정 (정밀도)
2) 문장 길이에 대한 과적합 보정 (brevity penalty)
3) 같은 단어가 연속적으로 나올 때 과적합되는 것을 보정 (clipping)

METEOR[64]
기계 번역 출력을 평가하기 위한 평가 지표이다.
유니그램 정밀도 및 재현율의 조화 평균을 기반으로 하며 정밀도보다 재현율이 더 높다. 이 평가지표는 더 많이 사용되는 BLUE에서 발견된 일부 문제를 수정하도록 설계되었으며 문장 또는 세그먼트 수준에서 인간의 판단과 좋은 상관관계를 생성한다.

SSA[65]
자율 발화 모델에 대한 평가 지표이다.
사람처럼 말하는 것이 목표인 모델을 평가할 때 사용하는 방법이며 기계적인 방법이 아닌 사람이 직접 자율 발화 모델과 대화하며 점수를 평가한다.

투명성

요구사항 13 인공지능 시스템의 설명에 대한 사용자의 이해도 제고

대표행위자 | 시스템 엔지니어 협력 대상 | 시스템 운영자 | 인공지능 모델 개발자 | 비즈니스 결정권자 | 인공지능 윤리 전문가

- 모델의 추론 결과에 대한 설명을 제공하는 기법을 적용하여도 사용자가 바로 이해하고 해석하기 어려운 경우가 많다. 따라서 인공지능 시스템의 운영자 혹은 서비스 제공자는 사용자에게 제공되는 결과가 이해 가능한지understandable, 해석 가능한지interpretable, 설명 가능한지explainable를 평가한다.

13-1 인공지능 시스템 사용자의 특성user characteristics과 제약사항을 분석하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

공공·사회 분야의 인공지능 서비스를 이용하는 사람의 특성을 고려하여 만족 여부를 판단하십시오.

- 인공지능 시스템의 결과가 적절한지 평가하기 위해서는 먼저 해당 결과를 읽는 사용자를 고려해야 한다. 사용자가 누군지에 따라 결과(설명)의 수준, 깊이, 맥락이 정해지는 만큼 사용자에 대한 자세한 분석이 수행되어야 한다.
- 이러한 공공·사회 분야의 서비스는 단순한 설명으로 제공하는 정보 외에도 서비스가 운영되는 장비, 환경 등을 고려한 상태 표시 및 시청각 정보 등을 활용할 수 있다. 따라서 명확한 정보 전달을 위해서는 사용자의 특성을 분석하고 다양한 전달 방법을 고려하는 것이 중요하다.

13-1a 사용자 특성에 따른 세부 고려사항을 분석하였는가?

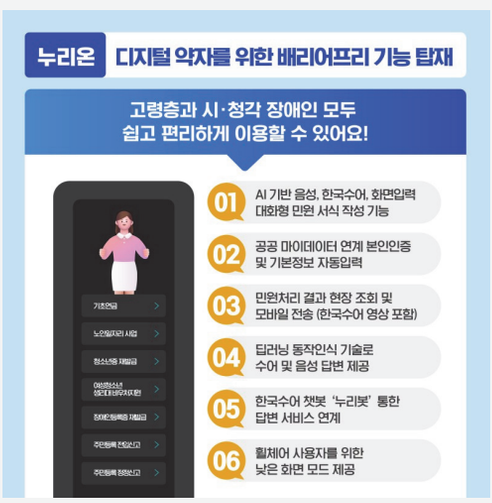
Yes ☐ No ☐ N/A ☐

- 공공·사회 분야에서 민간을 대상으로 서비스 되는 경우에는 일반적으로 사용자군이 불특정하므로 인간의 다양한 특성을 참고하여 사용자의 특성을 분석한다.
- 자연어 처리 기반의 서비스는 연령 또는 지역, 전문성에 따른 지식수준을 고려해야 한다. 사용자의 특성에 따라 사용하는 어휘, 단어, 사투리, 억양 등에 차이가 있다는 점을 고려하여 설명을 제공하도록 한다.
- 인공지능 서비스를 제공하는 방식은 장애인, 장노년층, 농어민, 저소득층 등 정보취약계층의 디지털 정보화 수준을 고려해야 한다. 특히 장애인은 서비스를 이용하는 데 더 많은 어려움을 겪을 수 있으므로 이러한 점을 고려하여 설명을 제공하도록 한다.

사용자의 특성 분석을 위한 고려 사항 예시

구분	상세 구분	고려 사항
연령	아동, 성인, 노인 등	아동은 성인과 비교해 이해할 수 있는 어휘, 단어에 한계가 있음
지역	경상도, 전라도, 충청도, 강원도, 제주도 등	지역 방언은 사용자의 목소리 톤, 억양, 사투리 등의 문제를 고려해야 함
장애 유무	장애인, 비장애인	신체적 제약으로 발생할 수 있는 한계를 고려해야 함. 예컨대 신체 크기, 신체 능력, 인지 능력에 제약이 있을 수 있음
지식	초보자, 전문가 등	관련 서비스의 경험 여부와 사전 배경지식의 차이로 지식수준이 다름을 고려해야 함

참고 대전광역시, 인공지능 기반 시청각장애인 민원 서비스 제공 사례[66]



- 대전시는 2022년 5월부터 인공지능(AI) 기반의 지능형 민원처리서비스(이하, 누리온)를 제공하고 있다. 누리온은 고령층과 시청각장애인 등이 민원 신청을 쉽게 할 수 있도록 만든 AI 기반 무인정보단말기로서, 기초연금 신청을 비롯한 민원서비스 7종을 제공한다.
- 이 서비스는 인공지능 기반의 영상·음성인식 기술을 통해 장애 유형에 맞춰 민원 정보를 안내하며, 수어 안내를 선택한 뒤 원하는 메뉴를 누르거나 수어로 표현하면 화면 속 캐릭터가 해당 정보에 관해 수어로 설명해 준다.
- 대전시의 인공지능 기반 민원 서비스 사례는 사회적 약자의 디지털 정보 격차를 줄이고 행정 효율을 제고하여 민원 서비스 향상에 기여할 것으로 평가되고 있다.

13-2 사용자 특성에 따른 충분한 설명을 제공하는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

공공·사회 분야의 인공지능 서비스가 충분한 설명을 제공해야 하는 경우에는 본 항목을 고려하여 만족 여부를 판단하십시오.

- 서비스를 활용하는 사용자는 다양하여 인공지능 시스템의 결과가 서로 다른 입장에서 해석되어 오해가 생길 수도 있다.
- 따라서 서비스 운영자는 13-1 에서 분석된 사용자 특성을 고려하여 의사결정의 주요 변수, 결정 요인, 데이터, 논리 또는 알고리즘을 명확하고 간단한 용어로 문맥에 따라 적절하게 제공해야 한다.
- 이러한 설명의 평가 기준으로는 명확성, 구체성, 정확도 등을 고려할 수 있다.

13-2a 사용자 특성에 따른 설명 평가의 기준을 수립하였는가?

Yes No N/A
☐ ☐ ☐

- 공공·사회 분야에서는 다양한 사용자가 서비스를 이용하는 만큼 설명을 포괄적으로 평가할 수 있는 특성과 세부 항목을 정하는 단계가 필요하다. 설명의 평가 기준은 구체성, 명확성, 적절성과 같은 항목이 될 수 있다. 세부 항목으로 데이터 유형data type이나 모달리티modality에 따라 각 항목에서 고려되어야 할 내용이 달라질 수 있다.
- 이 외에도 디지털 약자, 장애 등의 신체적 제약에 맞춰 사용자에게 적절한 형태로 정보를 제공함으로써 평등성과 같은 항목이 포함될 수 있다.

사용자의 특성 분석에 따른 설명 평가 항목 예시

구분	평가 항목
명확성	• 사용자가 이해하기 쉬운 단어와 문장을 사용하고 있는가? • 사용자에게 다른 오해를 불러일으킬 만한 표현·단어·어휘는 없는가? • 불필요한 설명이 있진 않은가? • 해당 설명에 사용자가 기대하고 얻고자 하는 정보가 모두 들어있는가?
구체성	• 사용자의 구체적 행동을 끌어낼 수 있도록 명확한 주어·목적어·서술어로 설명되는가?
적절성	• 주어진 설명이 사용자의 특정 지식수준을 요구하지는 않는가? • 배경지식 혹은 사전 경험이 필요하진 않은가? • 독자를 고려해 전문 용어 및 약어에 대한 설명을 제공하는가? • 설명이 제공되는 시점이 적절하였는가?
정확성	• 설명과 함께 제공되는 자료의 그림과 설명이 모두 일치하는가? • 사전에 제공된 예상 결과의 설명과 실제 결과가 모두 일치하는가? • 내부 알고리즘과 정확히 일치하는 설명인가?
효율성	• 인공지능을 사용할 경우에 시간적·금전적 낭비를 하지 않는가?
기억 용이성	• 제품(서비스)을 일정 기간 사용하지 않아도 바로 사용할 수 있도록 유연하게 설명되어 있는가?
평등성	• 장애가 있는 사용자와 비장애 사용자에게 동등한 수준의 정보와 기능을 제공하고 있는가? • 연령 및 색맹 등의 불편 사항에 맞춰 정보가 제공되는가? (예: 색깔 구별, 폰트 크기)

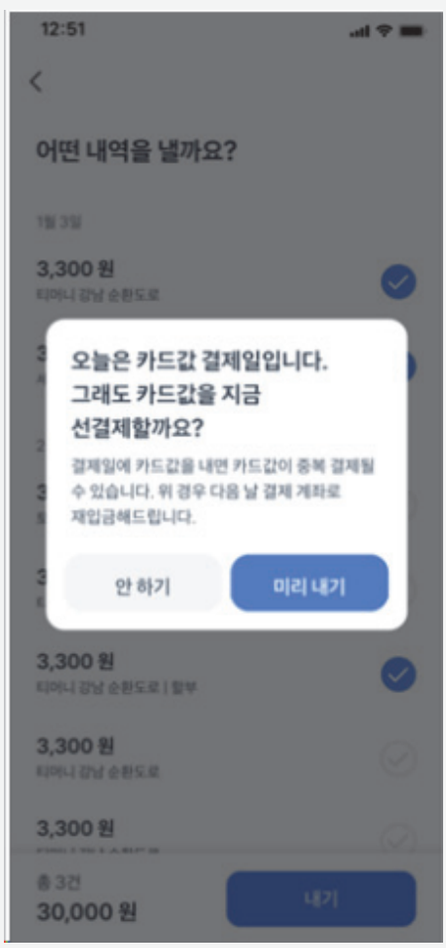
13-2b 사용자가 이해하기 어려운 전문 용어 사용을 지양하였는가?

Yes No N/A
☐ ☐ ☐

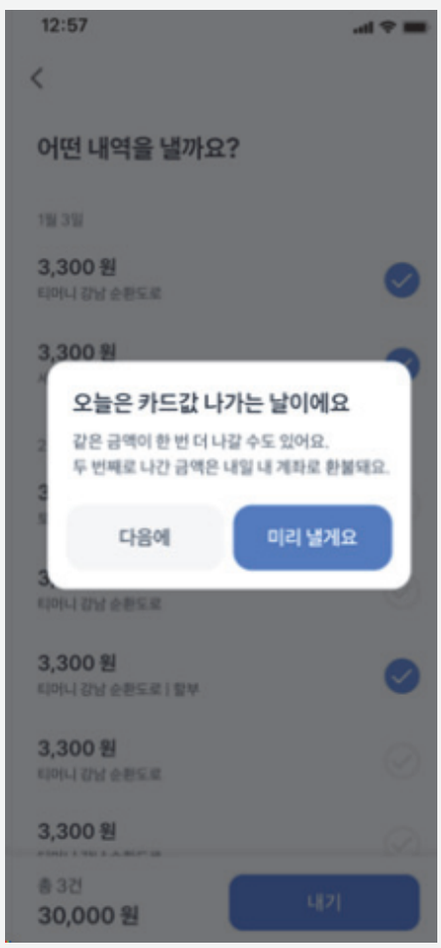
- 일반인에게 익숙하지 않은 용어나 전문용어를 사용하면 의미를 이해하기 어려워 해석에 많은 시간이 소요되거나 의도한 내용을 전달하지 못할 수 있다.
- 따라서 텍스트로 설명한다면 다양한 독자를 배려해 전문용어를 최대한 지양하고, 필요시 용어에 대한 설명을 추가로 작성하도록 한다.
- 그 예로 자연어 처리 기술 중에서 문장 내 특정 단어를 사용자 수준에 맞춘 적절한 단어로 변환해 주는 기술을 인터페이스에 적용할 수 있다.

참고 금융 플랫폼 토스의 쉬운 용어를 사용한 설명 사례

- 2013년 8월 설립 이후 지속적인 성장을 하고 있는 금융 플랫폼 '토스'의 성공 요인 사례 중 하나는 "쉽고 편하다"라는 평가이다. 토스는 사용자에게 정보를 제공하는 과정에서 어려운 금융용어를 이해하기 쉽게 순화하여 오해의 여지가 없도록 꼼꼼히 친절하게 설명해 준다.
- 예를 들어 "결제 금액보다 모자라요."가 아닌 "잔액이 모자라요."라고 안내하거나 "송금"이 아닌 "어디로 보낼까요?"라는 표현을 쓴다. 이에 고객은 쉽게 이해하고 편하게 사용할 수 있다.



변경 전



변경 후

13-2c 사용자의 구체적인 행동과 이해를 이끌어낼 수 있도록 명확한 표현을 사용하였는가?

Yes ☐ No ☐ N/A ☐

- 좋은 설명은 사용자로부터 구체적인 행동과 이해를 이끌어낼 수 있어야 한다. 따라서 설명을 간결하고 명확하게 함으로써 모호한 해석이 되지 않도록 작성하는 것이 중요하다.
- 시각적으로는 성공·실패·경고·위험 등 결과에 따른 색상을 일관성 있게 유지해 줌으로써 사용자가 한눈에 시스템 결과를 이해할 수 있다. 그리고 텍스트나 음성으로 제공되는 설명에서는 지시대명사를 사용하지 않고 대상을 명확하게 말해주는 것을 예로 들 수 있다. 또한 비슷한 발음이 연이어지는 경우, 다른 단어로 대체하는 것이 바람직하다.

참고 텍스트·음성 기반 인공지능 서비스의 설명 예시[67]

• AI 스피커는 대명사나 지시어의 사용을 최소화하여, 사용자가 정확한 정보를 습득하고 대응할 수 있도록 유도한다.

13-2d 설명이 필요한 위치와 타이밍은 적절한가?

Yes No N/A
☐ ☐ ☐

- 잘 작성된 설명이 적절한 위치와 타이밍에 나타나 이해를 돕는 것도 중요하다. 이를 위해 설명이 단발성이어야 할지 여러 번 반복해서 강조해야 할지를 숙고하고, 어느 위치에 놓여야 사용자가 잘 읽을 수 있을지도 고려하는 것이 필요하다.
- 특히 개인 맞춤형 인공지능 서비스는 사용자의 요구, 감정, 환경, 상황 등의 정보를 바탕으로 사용자에게 특화된 형태로 기능을 제공한다. 그렇기 때문에 공공서비스 중에서도 개인 맞춤형 서비스를 제공하는 경우에는 사람과 사회·문화에 대한 이해, 사람들의 경험과 감정에 대한 분석이 중요하다.
- 이와 더불어 작성된 설명의 위치와 타이밍이 적절한지를 조사하기 위해서는 13-2e의 사용자 경험 평가 및 조사 기법을 활용할 수 있다.

참고 장애인을 고려한 키오스크 사례

- 엘젠아이씨티에서 제공하는 음성인식, 자연어 대화, 음성 발화 등 인공지능 기술 기반의 서비스는 시각장애인용 키오스크를 제공한다.
- 일반 키오스크는 각 단계에 대한 안내만을 제공하는 데 비해 시각장애인용은 각 단계에 대한 안내 이후에 선택 가능한 메뉴에 대한 안내를 추가적으로 제공한다.
- 이렇게 주문 단계별 미션과 선택지에 대한 음성 안내를 추가적으로 제공함으로써 시각적 제약이 있는 사용자가 필요로 하는 정보를 적시에 취득하게 하고 정보의 접근성을 향상함으로써 서비스 이용 실패 경험을 최소화한다.

13-2e 사용자 경험을 평가할 수 있는 다양한 사용자 조사 기법을 활용하였는가?

Yes ☐ No ☐ N/A ☐

- 사용자 경험User eXperience, UX은 한 개인이 특정한 제품, 시스템, 서비스를 사용하며 느끼는 모든 것을 의미한다. 또한 그 개인이 인지하는 유용성, 사용 편의성, 효율성 등의 시스템 특성을 포함한다. 설명을 평가하기 위해 사용자 조사user research 기법을 활용할 수 있다.
- 다음은 국내에서는 진행된 인공지능 서비스 UX 평가를 위한 프레임워크에 대한 연구 논문의 내용이다.
- 사용자 조사 기법은 크게 접근 방식과 자료 획득 방식으로 구분할 수 있다. 우선, 사용자 조사 기법의 접근 방식에 따라 정량적(간접적) 조사와 정성적(직접적) 조사로 구분하며, 사용자 조사를 위해 자료를 얻는 방식에 따라 사용자 행동을 통한 조사와 태도를 통한 조사로 구분한다. 접근 및 자료 획득 방식을 고려해 적합한 사용자 조사 기법을 선정하고 사용자의 경험을 평가하는 것이 바람직하다.

  ✓ 접근 방식에 따른 구분 및 방법

  – 정량적(간접적) 조사quantitative user research: 사용자의 행동이나 태도에 대한 데이터를 도구 등을 통해 간접적으로 수집하는 방법 (예: 웹로그 분석, A/B 테스팅A/B testing, 설문 조사, 고객 지원 자료 분석)

  – 정성적(직접적) 조사qualitative user research: 사용자의 행동이나 태도를 직접 관찰하는 방법 (예: 인터뷰, 표적 집단 인터뷰focus group interview, 프로토타입 테스팅prototype testing)

  ✓ 자료 획득 방식에 따른 구분 및 방법

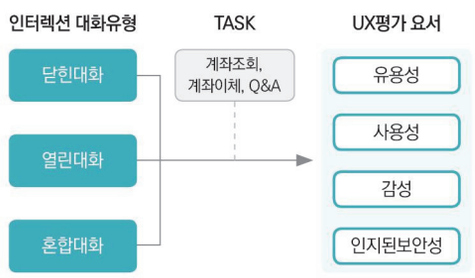
  – 사용자 행동 기반 조사behavioral user research: 사용자가 어떤 행동을 하는지 조사하는 방법 (예: 웹로그 분석, A/B 테스트, 시선 추적eye tracking)

  – 사용자 태도 기반 조사attitudinal user research: 사용자가 무엇을 말하는지 조사하는 방법 (예: 카드 소팅card sorting, 심층 인터뷰, 요구사항 조사)

참고 챗봇 및 음성인식 서비스의 사용자 조사 기법 적용 사례[68]

- 다음 사례는 금융 서비스 챗봇의 인터렉션 대화 유형이 사용자의 유용성, 사용성, 감성, 보안성에 미치는 효과에 대해 조사한 사례이다.
- 본 연구에서는 조사 범위research scope를 정의하기 위해, 챗봇 인터렉션 디자인과 챗봇 구현 방식에 따른 대화 유형을 정의한다. 이후 금융 챗봇이 적용된 서비스 사례와 프로세스를 분석하고 사용자의 경험 평가를 수행하였다.

조사 기간 및 대상	평가 기준
1. 조사 기간: 2018년 10월 15일 ~ 11월 15일(약 1달) 2. 실험 대상: 총 90명, 불성실한 응답자를 제외하고 81명의 대상자만 통계에 포함 3. 참여 형태: 온라인 참여자 35명, 대면 참여자 46명 4. 성별: 남성 23명(28%), 여성 58명(71%) 5. 연령층: 20대 40명(49%), 30대 40명(49%), 40대 1명(1%)	1. (유용성) 챗봇을 사용한 서비스는 금융 업무를 향상하는 유용한 기능이었는가? 2. (사용성) 사용자의 금융 서비스 이용 시, 쉽고 사용하기 편리하였는가? 3. (감성) 시스템 사용자는 마음속에서 얼마나 적절한 느낌을 받았는가? 4. (보안성) 서비스 이용 시, 금융 챗봇의 인터렉션 방식에 대한 안전성 만족도는 어떠한가?

가설 수립	연구 모델
1.가설 1: 금융 챗봇의 인터렉션 유형은 사용자의 금융 서비스 목적에 따라 유용성에 차이를 보일 것이다. 2.가설 2: 금융 챗봇의 인터렉션 유형은 사용자의 금융 서비스 목적에 따라 사용성에 차이를 보일 것이다. 3.가설 3: 금융 챗봇의 인터렉션 유형은 사용자의 금융 서비스 목적에 따라 감성에 차이를 보일 것이다. 4.가설 4: 금융 챗봇의 인터렉션 유형은 사용자의 금융 서비스 목적에 따라 인지된 보안성에 차이를 보일 것이다.	

사용자 평가 결과

1. 정량적 평가 결과

구분		닫힌대화a		열린대화b		혼합대화c		F	P	Scheffe
		M	SD	M	SD	M	SD			
유용성	계좌조회	4.14	0.877	3.80	0.765	4.14	0.848	4.341	0.014*	a=c>b
	계좌이체	4.10	0.982	3.67	0.908	4.02	0.908	4.963	0.008**	a>b
	Q&A	3.79	0.996	3.63	0.955	3.80	0.886	0.840	0.433	-
사용성	계좌조회	4.21	0.918	3.69	1.032	4.28	0.869	5.862	0.000**	a=c>b
	계좌이체	4.09	0.990	3.60	1.069	4.12	0.941	3.930	0.001*	a=c>b
	Q&A	3.89	0.949	3.58	0.998	3.95	0.947	2.156	0.034*	-
감성	계좌조회	3.31	1.080	3.16	1.006	3.23	1.143	0.383	0.682	-
	계좌이체	3.30	1.066	3.11	1.025	3.14	1.137	0.706	0.495	-
	Q&A	3.26	1.058	3.28	1.040	3.35	1.063	0.145	0.865	-
인지된 보안성	계좌조회	3.04	1.066	2.80	0.928	2.91	1.002	1.115	0.330	-
	계좌이체	2.99	1.164	2.72	1.098	2.72	1.040	1.626	0.199	-
	Q&A	3.40	1.080	3.16	0.981	3.43	1.036	1.646	0.195	-

2. 정성적 평가 결과

2.1. (유용성 – 부분 채택) 계좌조회와 계좌이체 서비스 이용 시 유용성에서 닫힌대화가 열린대화보다 더 높은 점수를 받았다.

– 사용자들은 계좌이체 시 빠르고 정확한 업무 수행을 가장 중요시하는 것으로 추론할 수 있다.

2.2. (사용성 – 채택) 계좌조회, 계좌이체, Q&A 서비스 이용 시, 모두 혼합대화가 가장 높은 점수를 받았다.

– 사용자들은 금융 서비스 이용 시 실용적인 인터렉션 방식을 선호한다는 것을 의미한다.

2.3. (감성 – 기각) 사용자들은 인터뷰를 통해, 금융챗봇 서비스 이용 시 사용자의 감성에 영향을 주는 요인으로 기존의 상담원과의 대화를 선호하는 사용자와 챗봇과의 대화를 흥미롭게 여기는 사용자로 구분되었다.

2.4. (인지된 보안성 – 기각) 사용자들은 인지된 보안성에 대한 질문에서 다양한 의견 차이를 보였다.

05 운영 및 모니터링

책임성 투명성

요구사항 14 인공지능 시스템의 신뢰성 테스트 계획 수립

대표행위자 | 시스템 엔지니어 협력 대상 | 인공지능 모델 개발자 데이터 과학자

- 〈공공기관 신뢰 가능 AI 구현 실용 가이드-OECD 권고안의 적용-〉 참고 후, 공공기관의 인공지능 서비스는 생명주기 내의 데이터셋, 프로세스, 의사결정과 관련하여 추적이 가능하다는 것을 보장한다.
- 이 외에도 인공지능 시스템 운영 단계에서 문제 원인을 추적하기 위한 시스템 로그, 데이터 모니터링, 인공지능 모델과 사람 간의 의사결정 기여도 추적, 변경이력과 같은 방안을 확보한다.

14-1 인공지능 시스템의 의사결정에 대한 추적 방안을 수립하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

공공·사회 분야의 인공지능 서비스 중에서 예측·판단 인공지능이 적용되는 서비스의 경우에는 본 항목을 고려하여 만족 여부를 판단하십시오.

- 인공지능 시스템의 의사결정은 인공지능 모델이 자체 결정하거나 시스템 운영자 또는 사용자가 개입해 내릴 수 있다. 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 발생 가능한 인공지능 시스템 추론 결과의 영향을 추적하기 위해서 시스템 단계별로 로그 수집 대상 정보를 정의하고 모니터링을 지속해야 한다.

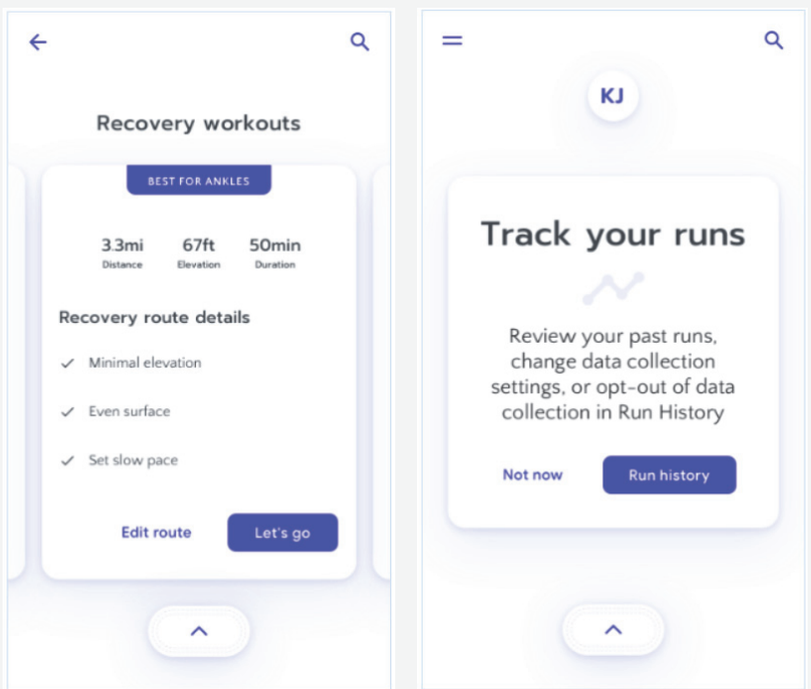
14-1a 인공지능 시스템의 의사결정에 대한 기여도 추적 방안은 확보하였는가?

Yes ☐ No ☐ N/A ☐

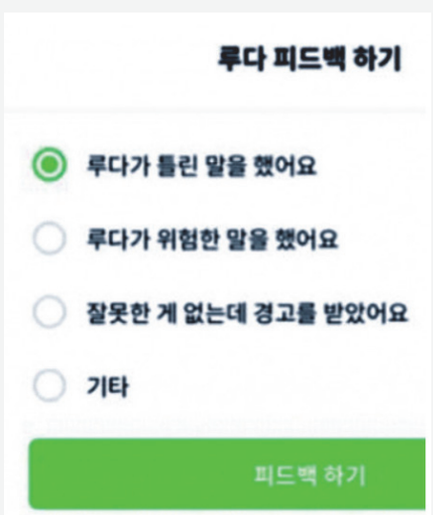
- 인공지능 시스템의 결정에 대한 모델 기여도를 파악하기 위해서는 이전 모델의 추론 정보와 최종 결정에 대한 사람(예: 공공기관 담당자, 시스템 운영자) 개입 여부 등의 정보가 추적되어야 한다.
- 인공지능 서비스의 기여도를 파악하기 위해 적절한 시점에 피드백을 수집하는 방법을 적용하고 있으며, 이는 사용자에게 피드백 요청/의도를 명시하는 명시적^{explicit} 피드백과 사용자의 사용 기록 등을 피드백으로 활용하는 암시적^{implicit} 피드백이 있다. 암시적 피드백은 사용자가 본인의 행동이 피드백으로 활용되고 있다는 것을 인지하지 못하기 때문에 약관에 명시해야 한다.[69]

참고 인공지능 시스템의 피드백 사례

- Google의 〈People+AI Guidebook〉에서는 다음과 같이 사용자가 인공지능의 의사결정에 대해 최종 판단을 내릴 수 있도록 하며 해당 결과를 바탕으로 인공지능을 학습시키는 것을 권고한다.



- 스캐터랩은 AI 챗봇 이루다 2.0과의 대화 시, '의견 보내기' 섹션과 대화창의 피드백 전송 기능을 통해 이용자의 피드백을 받고 있다.



14-1b 인공지능 시스템의 의사결정 추적을 위한 로그 수집 기능을 구현하였는가?

Yes ☐ No ☐ N/A ☐

- 복지 검색, 복지 기준 사전 평가 등 공공·사회적으로 민감한 분야의 대민 인공지능 시스템의 추론 결과는 그 근거가 중요하다. 따라서 모델의 학습 과정, 운용 시 의사결정 결과, 사용자 입력 데이터 등 이유를 추적하기 위한 로그를 수집해야 한다.
- 이를 위해 시스템 프로세스별 로그를 수집할 정보를 선정하고, 정보 간의 중요도를 정의한 다음에 로그 레코드 형식을 결정해야 한다.
- 특히 인공지능 시스템 운영 과정에서의 오류 원인 추적을 위해서는 모델 구축 방법과 데이터셋 측면을 포함한 오류 원인의 분석이 필요하므로, 두 가지 측면을 다 고려하여 로그를 수집해야 한다.

사용자의 특성 분석에 따른 설명 평가 항목의 예시

오류 구분	오류 원인 예시
모델 구축 방법 측면의 오류	모델·데이터의 대상 선정, 수집, 정제, 라벨링 등의 통제 미흡으로 인해 구축 절차, 구조, 학습 모델 측면의 다양한 오류 데이터 생성
데이터셋 측면의 오류	데이터셋 설계의 부족, 구문 정확성 위배, 데이터 구축 중복 등으로 인한 학습 데이터 품질의 저하

14-1c 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 확인할 수 있는 가장 기본적인 방법이 될 수 있다. 서비스 로그는 서비스가 운영되는 동안 지속해서 수집되며 서비스 고도화에 따라 다양한 형태로 누적될 수 있다.
- 공공·사회 분야의 서비스는 다양한 사용자 계층을 대상으로 하므로 디지털기기의 활용에 상대적으로 취약한 사용자의 경험을 모니터링하고 분석하여 더 나은 서비스를 제공하는 데 활용할 수 있다. 수집하여 분석할 수 있는 사용자 경험 로그는 다음의 예시와 같다.
  - ✓ 멀티 모달 기반 비대면 노인 돌봄 서비스: 노인의 인공지능 스피커에 반응하고 대화에 걸리는 시간, 웨어러블 디바이스의 착용 시의 편안함, 웨어러블 디바이스 관리의 쉬움 여부 등
  - ✓ 인공지능 대민 상담 서비스: 민원인의 상담 만족도, 한 상담 내에서 서비스가 동일하게 응답한 횟수(부정적 경험으로 예상) 등

14-2 학습 데이터의 변경 이력을 확보하고, 데이터 변경이 미치는 영향을 관리하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

인공지능 시스템의 성능을 개선하기 위해 주기적으로 학습 데이터를 변경하는 경우에 본 항목을 고려하여 만족 여부를 판단하십시오.

- 인공지능 모델은 사용한 데이터에 따라 학습 모델도 함께 달라진다. 이로 인해 모델의 설계나 주요 파라미터들의 변경이 함께 이루어질 수 있다. 따라서 모델 개발과정에서 학습 데이터가 변경될 경우, 학습 데이터 버전관리 및 변경이 발생한 원인을 추적해야 한다.
- 또한, 신규 데이터를 포함하여 인공지능 모델의 추가 학습이 필요한 경우, 학습 데이터 변경으로 인한 모델의 성능 영향을 평가하기 위해 기존 학습 데이터에 추가된 신규 데이터 비율에 따른 모델 성능 변화 추적이 가능하도록 기록 및 관리하는 것이 바람직하다.
- 이러한 학습 데이터 변경 이력 관리를 위해 학습 데이터 버전관리를 위한 오픈소스 도구 활용, 자체 시스템 구축 등을 고려할 수 있으며, 학습 데이터를 사용 또는 운용하는 이해관계자들이 데이터 변경으로 인한 영향을 확인할 수 있도록 학습 데이터 변경 원인, 변경된 학습 데이터의 구조, 학습 모델의 추론 결과 및 모델 변경으로 인한 성능평가 결과 등에 대한 정보를 제공해야 한다.

14-2a 데이터 흐름 및 계보Lineage를 추적하기 위한 조치를 마련하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템의 경우, 데이터의 변경으로 인해 모델의 확장이나 재설계 등의 시스템 변경이 발생할 수 있다. 따라서 시스템의 변경을 유도하는 데이터의 흐름 및 계보를 계속해서 추적해야 한다.
- 특히 민간기업에서 개발한 서비스를 공공기관에 납품하는 경우에는 개인정보 유출을 예방하기 위해 데이터 수집 경로를 비롯한 데이터 전달 이력을 관리할 필요가 있다.
- 데이터 흐름 및 계보는 데이터 변경에 대해 역방향, 순방향, 종단간end-to-end 관점으로 나누어 추적할 수 있으며, 추적을 위한 고려사항은 다음과 같다.
  - ✓ 데이터 흐름 및 계보 추적을 관리하기 위한 데이터 정책팀을 구성하는 것이 유용한가?
  - ✓ 데이터 흐름 및 계보 추적을 위해 메타데이터를 기록하고 유지보수할 것인가?
  - ✓ 데이터 흐름 및 계보 추적을 위한 데이터 적재, 매핑, 관리, 시각화 리포팅 기능을 구현하는 것이 유용한가?
  - ✓ 인공지능 개발 과정에서 모델의 특성 값을 저장 및 공유하는 특성 저장소feature repository 기능을 구현하는 것이 유용한가?
  - ✓ 데이터는 출처까지 역추적될 수 있는가?

14-2b 데이터 소스 변경에 대한 모니터링 방안을 확보하였는가?

Yes No N/A
☐ ☐ ☐

- 공공·사회 분야의 서비스 개발 시, 인공지능 모델의 학습 데이터를 확보하기 위해 데이터 제공 업체를 통한 데이터 수집 또는 구매, 직접적인 데이터 수집 등의 방법을 활용할 수 있다. 직접 수집하거나 데이터 제공 업체를 통해 받는 경우에는 수집 방법 변경, 데이터 제공 업체의 변경 등으로 데이터의 분포가 깨지거나 데이터 형식이 달라지는 등 학습 데이터의 무결성이 깨질 수 있다.
- 공공·사회 분야의 인공지능 알고리즘 개발을 위해 오픈소스 데이터셋을 활용하는 경우에는 데이터셋이 변경될 수 있다. 따라서 모델의 성능 개선 등에 반영하려면 주기적인 모니터링으로 최신의 데이터셋을 반영한다.
- 또한 공공·사회 분야의 인공지능 개발 시, 공공기관에서 제공하는 데이터 또는 공공 데이터를 학습을 위한 데이터셋으로 활용할 수 있다. 그때 다음의 예시와 같이 공공 데이터 소스 변경이 발생할 수 있어 주기적으로 확인하거나 알림을 받는 방안이 필요하다.
  - ✓ 지역의 건강 지표, 수준 및 동향 예측: 각 건강 지표 관련 데이터는 보통 월, 분기, 반기, 연 단위로 업데이트되므로 필요시 반영하여 모델을 재학습시켜야 함
  - ✓ 인공지능 기반 발주 지원 서비스: 작성된 제안 요청서의 법령 준수 여부, 요구사항 오류 등을 진단할 때, 법령이 바뀌면 모델을 재학습시켜야 함
- 인공지능 시스템의 데이터 소스 변경은 성능에 직접적인 영향을 줄 수 있다. 따라서 데이터 수집 과정을 모니터링해 데이터 소스 이상이나 중복 수집과 같은 문제에 대응할 수 있어야 한다.

14-2c 데이터 변경 시, 버전관리를 수행하였는가?

Yes No N/A
☐ ☐ ☐

- 인공지능 모델의 개발 과정에서 학습 데이터의 업데이트, 오류로 인한 라벨링 재수행 등과 같이 데이터 변경이 이루어지면 학습 결과인 모델도 변경된다. 특히 현재 인공지능 기반의 공공서비스는 대부분 배치 학습batch learning을 적용하고 있는데, 이러한 인공지능 시스템은 학습 데이터양이 많은 것이 특징이다.
- 또한 이전에 학습에 사용한 데이터셋과 특성이 완전 다르거나 데이터셋 전체를 교체한다면 성능이 크게 저하될 수 있으므로 그런 경우에는 추가 학습이 필요할 수 있다.
- 따라서 학습 데이터의 변경이 수행될 경우에는 단순히 사용된 학습 데이터의 버전뿐만 아니라 해당 버전으로 학습한 인공지능 모델을 함께 관리해야 한다. 특히, 신규 데이터를 추가하여 학습 데이터를 변경할 필요가 있을 때는 학습(혹은 테스트)에 사용된 신규 데이터의 비율을 기록하고, 그에 따른 모델의 성능 변화를 함께 추적할 수 있어야 한다.
- 이를 위해 기계학습 프로젝트를 위한 오픈소스 기반의 데이터 버전 관리 도구(예: DVCData Version Control)의 도입을 고려하거나 자체적으로 학습 데이터 버전 관리 시스템을 구축하여 학습 데이터 및 모델의 버전 관리를 수행해야 한다.

12-2d 데이터 변경 시, 이해관계자를 위한 정보를 제공하는가?

Yes No N/A
☐ ☐ ☐

- 다수의 이해관계자가 참여하는 인공지능 시스템의 개발 과정에서 데이터 변경으로 인한 인공지능 모델의 설계, 주요 초매개변수의 변경 및 재학습 등의 조치를 이해하기 위해서는 이해관계자의 역할을 고려한 정보의 제공이 필요하다.
- 데이터 변경에 따라 이해관계자별로 제공되어야 하는 정보는 다음과 같다.

이해관계자	제공 정보
비즈니스 결정권자	데이터 변경에 따른 모델의 세세한 변경 사항보다 기존 시스템의 목적, 서비스 의도 등의 변경 사항 및 시스템 전체의 방향성에 초점을 맞춘 정보
데이터 과학자	기존 데이터와 변경된 데이터의 특성, 포맷, 규모 등의 차이점 등의 정보
시스템 개발자	변경된 데이터의 설명을 참고하여 기존 모델과의 호환성, 모델 구조 재설계, 모델 재학습 세부 전략(예: 목적함수, 학습 시간, 학습 알고리즘), 예상 추론 결과의 변경 사항에 대한 정보
모델 검증자	변경된 테스트 데이터셋 구성, 재설계 및 재학습된 모델에 대한 주요 성능평가 결과, 기존 모델과의 성능 비교 결과 등의 정보
모델 운영자	검증을 마친 변경 모델에 대한 운영 및 사용자 모니터링 결과 등을 수집하여 분석한 정보

14-2e 신규 데이터 확보 시, 인공지능 모델의 성능평가를 재수행하였는가?

Yes ☐ No ☐ N/A ☐

- 신규 데이터를 확보한 다음, 인공지능 시스템에 사용하기 위해서는 기존에 운영 중인 인공지능 모델과의 성능 비교가 필요하다. 사람이 판단하기에 신규 데이터가 기존 학습 데이터와 유사하여도 학습된 인공지능 모델이 기존 학습 데이터에서 학습한 데이터의 특성과는 다를 수 있다. 다음은 신규 데이터를 반영함에 따라 인공지능 모델의 성능을 재평가할 필요가 있는 예시이다.
  - ✓ 비대면 노인 돌봄 서비스에서 음성인식 인공지능 모델 적용: 인구 분포 및 노인 대상자의 범위 변경으로 신규 노인에 대한 추가 데이터를 반영했다면 기존 노인을 대상으로 인식하던 음성인식 인공지능 모델의 성능을 재평가하여 기존 노인과 신규 노인을 대상으로 같거나 더 나은 성능의 음성인식 서비스를 제공하도록 함
  - ✓ 미아 추적용 얼굴인식 인공지능 모델 적용: 신규 학습 데이터셋을 포함하여 학습한 결과가 기존의 얼굴인식 성능과 같거나 더 나은 성능을 제공하도록 함
- 따라서 신규 데이터를 대상으로 도메인의 대표적인 인공지능 알고리즘을 사용하여 성능평가를 진행하고 분석하는 과정이 필요하다. 신규 데이터 확보에 따른 성능평가를 위해서는 다음과 같은 과정을 참고한다.
  - ✓ 성능평가 및 비교 분석을 위한 기존 학습 모델 및 관련 대표 인공지능 모델 확보
  - ✓ 대상 인공지능 분야 및 모델에 적절한 성능평가 지표 선정
  - ✓ 성능평가를 위한 실험 설계(정량적·정성적 실험 방법 선정, 실험 모델들의 파라미터 설정, 세부 실험 계획 등)
  - ✓ 실험 진행 및 결과 분석(결과에 따라 신규 데이터 평가, 필요시 모델 재설계, 확장, 재학습 등을 결정)

참고 인공지능 시스템의 드리프트 분석 및 완화 예시[70]

- 드리프트drift란?
  - ✓ 운영 환경의 변화로 인해 인공지능 시스템의 성능이 저하되는 현상을 의미함
  - ✓ 인공지능 시스템의 아키텍처 및 학습 데이터는 환경에 대한 가정을 구현하며, 이러한 가정이 더는 현실을 반영하지 않을 때 인공지능 시스템은 목표를 달성하지 못함
- 드리프트가 기계학습 시스템에 미치는 영향
  - ✓ 드리프트는 기능 추출, 모델 인코딩에 영향을 미치는 운영 환경의 예기치 않은 변경의 결과임
  - ✓ 인공지능 시스템이 나쁜 결과를 제공하기 시작한다면 드리프트가 범인일 가능성이 큼
- 콘셉트 드리프트
  - ✓ 데이터양, 객체, 엔티티 및 액션의 발생 여부, 중요 사항 등의 초기의 근본적인 운영 환경이 시간이 경과함에 따라 새로운 객체의 유입(예: 통행료 부과 대상의 추가) 등으로 바뀌는 것을 의미함
  - ✓ 운영 환경이 너무 많이 변경되어 감지된 특성 및 인공지능 모델이 더는 정확한 결정을 내릴 수 없음
  - ✓ 콘셉트 드리프트를 해결하기 위해서는 새로운 특성을 찾거나 인공지능 모델을 변경하거나 또는 둘 다 수행해야 할 수 있음

05 운영 및 모니터링

- 데이터 드리프트
  - ✓ 특성 추출 구성 요소에 영향을 줄 수 있는 변경 사항의 발생을 의미함
  - ✓ 예를 들어, 특정 장비 재보정, 조명 변경, 데이터 품질 저하 또는 감정 분석에 중요한 의미를 가감하는 언어 추세 등이 있음
- 드리프트 감지
  - ✓ 출력 검증: 인공지능 시스템 출력의 정확도를 모니터링하여 드리프트 감지
  - ✓ 시스템 모니터링: 인공지능 시스템 작동 중에 생성되는 중간 수량을 분석하여 드리프트 가능성 감지
- 출력 검증
  - ✓ 인공지능 시스템의 출력에 대해 '온전성 검사'를 수행하여 드리프트 감지
  - ✓ 첫 번째로 최고 수준에서는, 비즈니스 프로세스에 대한 다운스트림 영향을 모니터링하여 드리프트 감지
  - ✓ 두 번째로, 인공지능 시스템 출력을 사람의 주석과 주기적으로 비교하는 품질 관리 절차를 도입하여 드리프트 감지
  - ✓ 세 번째로, '워치독' 기계학습 시스템을 학습해 기본 인공지능 시스템을 두 번 추측하도록 학습하여 드리프트 감지
- 시스템 모니터링
  - ✓ 인공지능 시스템의 출력을 답안과 비교하지 않고도 드리프트를 감지할 수 있음
  - ✓ 데이터 드리프트 모니터링: 인공지능 시스템의 운영 환경에서 추출한 특성의 통계를 추적
  - ✓ 모델 모니터링: 다양한 계층의 정규화된 출력값을 모니터링
  - ✓ softmax 모니터링: 인공지능 시스템의 결정에 대한 자신감이 떨어지는 여부를 모니터링
- 드리프트 완화
  - ✓ 데이터 드리프트로 인한 경우에는 운영 환경에서 최근에 수집된 데이터로 모델을 재교육하여 성능을 복원할 수 있음
  - ✓ 데이터 어노테이션 팀을 참여시켜 유효성 검사 및 모니터링을 지원하고 재학습을 준비할 수 있음
  - ✓ 시스템 모니터링의 데이터를 사용하여 필요에 따라 기능 추출 및 ML시스템 모델에 대한 수정을 안내할 수 있음

책임성 투명성

요구사항 15 서비스 제공 범위 및 상호작용 대상에 대한 설명 제공

대표행위자 | 시스템 엔지니어 협력 대상 | 시스템 기획자 시스템 운영자 인공지능 모델 개발자 비즈니스 결정권자

- 〈공공기관 신뢰 가능 AI 구현 실용 가이드-OECD 권고안의 적용-〉에 따르면, 공공기관은 AI를 도입하고 활용할 때, 기관과 사람 간의 신뢰를 강화하기 위한 원활한 소통 방안을 마련한다.
- 따라서 사용자가 인공지능 시스템이 제공하는 서비스를 올바르게 사용하고, 제공된 서비스를 오남용하지 않도록 서비스의 목적, 범위, 제한 사항, 상호작용 대상을 포함한 내용을 설명한다.

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

Yes ☐ No ☐ N/A ☐

해당여부 판단

공공기관에서 직접 활용하거나 대민 서비스로 운영되는 경우에 본 항목을 고려하여 만족 여부를 판단하십시오.

- 공공기관에서 인공지능 시스템을 직접 활용한다면 인공지능 시스템은 공정·공익을 바탕으로 하는 의사결정에 활용되어야 한다. 그런데 시스템을 올바르게 사용하기 위한 정보가 제공되지 않으면 인공지능 시스템을 오남용하여 투명성 확보가 필수인 공공행정에 피해를 줄 수 있다.
- 또한 대민 서비스 분야의 인공지능 시스템은 다수의 국민에게 영향을 끼칠 수 있기 때문에 사용자가 인공지능 서비스를 이용하는 과정에서 오남용하지 않고 올바르게 사용할 수 있도록 해당 정보를 제공하는 것이 중요하다.
- 따라서 공공기관은 제공하는 인공지능 서비스의 목적, 의도, 역할, 범위를 설명하고, 인공지능 서비스의 의사결정이 개인에게 어떤 영향을 미치는지 등의 정보를 공개함으로써 투명성을 제고해야 한다.

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 얻을 수 있는지를 의미한다. 사용자는 서비스 목적과 목표를 설명함으로써 사용 맥락에 맞는 적합한 기능을 선택하여 활용할 수 있다.
- 공공서비스는 편익이 미치는 범위에 따라 시민 전체에 귀속되는 공익적 서비스와 개인에게 귀속되는 사익적 서비스로 구분된다.[71] 이처럼 공공서비스의 편익에 영향을 받는 대상이 다양할 경우를 분석하고, 각 이해관계자에게 제공되는 가치를 설명함으로써 사용자가 목적에 맞게 사용할 수 있도록 유도할 수 있다.

참고 midas HRi AI 역량검사의 목적 설명[72]

- AI역량검사 서비스를 제공하는 midas HRi에서는 서비스의 목적과 배경, 대상에 따른 제공 가치를 설명하고, 검사 유형별로 적용된 인공지능 기술과 검사 결과에 대한 정확도, 객관성 및 중립성의 확보 결과를 제공한다.

2.1 AI역량검사의 목적

AI역량검사는 최적 인재를 합리적으로 선발하기 위한 의사결정의 보조 목적으로 개발되었다.

AI역량검사를 통해 기업에서 채용을 효과적으로 진행할 수 있도록 보조하는 동시에 궁극적으로는 역량이 우수한 인재가 기업에서 많은 가치를 창출하는 것을 돕고자 한다.

- 기업의 채용에서 선발 의사결정을 도움
- 많은 구직자가 자신의 역량을 기업에 보여줄 수 있는 기회를 제공함
- 채용 프로세스의 효율성을 개선함(시간·비용 절약)
- 최적 인재 선발을 통해 기업이 사회적 부가가치를 창출하도록 도움

AI역량검사는 기업의 채용과 선발을 보조하는 도구이다.

기업 채용에서 선발 도구를 사용하는 것은 궁극적으로 최적 인재를 선발하기 위한 목적이며, 실용적인 관점에서 효율적으로 채용 프로세스가 운영되는 것도 중요하다.

채용과 선발 방법은 사회적으로도 중요한 의미를 지니고 있다. 왜냐하면 기업의 선발 방법에 따라 사회 전체가 영향을 받기 때문이다. 많은 기업에서 선발의 효율성을 위해 스펙 중심의 채용을 하고 있으며, 이로 인해 구직자들은 스펙을 만들기 위해 불필요한 시간과 비용을 투자하고 있다. 이처럼 실제 성과와 무관한 스펙으로 채용을 진행하면 사회 전체적으로 비용의 낭비가 발생하고, 스펙이 부족한 지원자는 지속적으로 구직에 실패하는 낙인효과가 나타난다.

따라서 스펙이 아닌 역량 중심의 채용 문화를 만들기 위해서는 별도의 준비가 필요하지 않으면서 합리적인 선발의 효과성을 지닌 도구를 사용하는 것이 매우 중요하다.

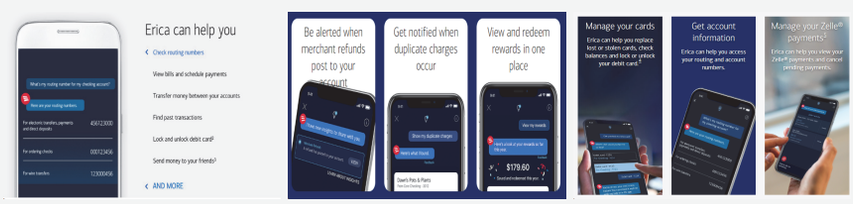
이처럼 기존의 채용 선발이 지닌 문제의 해결을 돕기 위해 **AI역량검사**를 기획하고 개발하였다.

AI역량검사의 제공 가치

대상	제공 가치
기업 관점	• 기존 채용 방법의 한계점·문제점 개선 및 타당도 높은 선발 방법 제공 • 허들 방식의 채용에서 발생하는 최적 인재 손실 방지 • 오프라인 검사 실시로 인한 시간·비용의 낭비 감소 • 보다 많은 지원자에게 면접 기회 제공(온라인 영상 면접 활용) • 면접의 효과성 향상과 편향 감소(면접관에게 자기 감시의 기회 제공) • 역량 중심의 채용을 위한 데이터 누적 관리 및 선발 프로세스의 개선
구직자 관점	• 모든 구직자에게 충분한 기회를 제공하여 상대적 박탈감 해소(시간·장소의 제약 없이 응시, 온라인 영상 면접 지원 등) • 역량 중심의 채용으로 구직 준비에 소요되는 시간·비용의 감소(스펙 쌓기와 인적성 검사 준비 등의 문제 해소)
사회적 관점	• 스펙 중심의 채용 관행에서 역량 중심 채용으로의 변화 • 학벌 중심 채용으로 발생하는 채용에서의 부익부 빈익빈 현상 해소 • 기업과 구직자 모두에게 채용 준비·활동으로 발생하는 사회적 비용 감소 • 공평한 기회와 공정한 평가 제공

참고 BANK OF AMERICA – Erica[73]

• AI 금융비서 Erica는 사용자의 자산 관리 및 금융 관련 업무를 간편하게 처리하는 기술을 지니고 있다. 홈페이지에 서비스의 목적과 제공하는 기능에 대한 설명이 있다. 사용자는 목적(카드 관리, 모든 계좌 잔액 보기, 월별 지출에 대한 주간 업데이트 검토 등)에 맞는 업무를 간단한 채팅으로 처리하고 확인할 수 있다.



15-1b 서비스의 한계와 범위에 대한 설명을 제공하는가?

Yes ☐ No ☐ N/A ☐

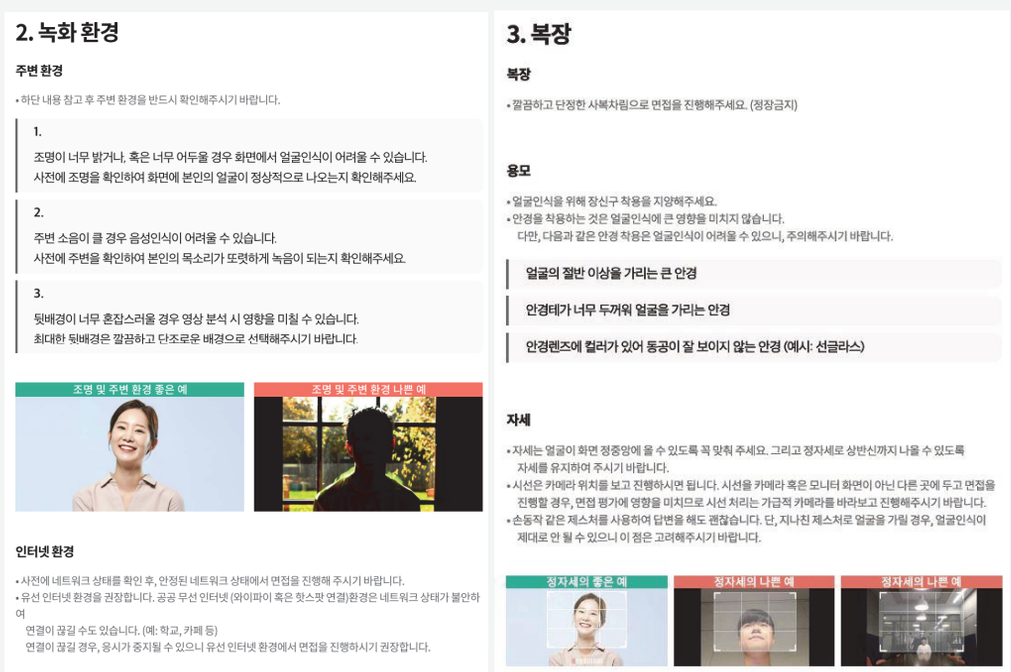
• 서비스 제공 범위와 한계를 설명함으로써 사용자 기대치를 조정할 수 있다. 서비스 결과에 대한 품질은 사용자 그룹 특성, 사용 환경, 사용 데이터 등 다양한 요인에 영향받아 결과가 도출될 수 있으므로 사용자에게 서비스 한계와 제공 범위에 대해 말하는 것이 중요하다.

• 특히 공공기관에서 사용하는 인공지능 서비스는 인공지능의 의사결정이 사용자 외에 제3자에게 영향을 미치는 경우가 있다. 이런 경우에는 인공지능 서비스가 영향을 미치는 범위를 설명하고 사용자가 인지할 수 있도록 해야 한다.

참고 병무청 AI 영상 면접 사전 안내[74]

병무청에서는 AI 영상 면접을 실시하고 있으며, 사전 안내 자료에서는 사용 장비에 따른 준비 사항 및 최소 요구사항, 주변 환경 및 인터넷 환경에 대한 안내, 복장, 용모, 자세에 대한 가이드를 제공하고 있다.

이 외에도 진행 과정에서 발생할 수 있는 다양한 예외 상황에 대응하는 방안과 유의 사항을 FAQ 형태로 정리하여 함께 안내한다.



유의 사항

- 외부로 출력되는 스피커·마이크를 사용하면 하울링(소리증폭현상)이 발생할 수 있으므로 마이크 기능이 있는 이어폰을 사용하시기 바랍니다.
- 헤드셋을 사용해도 괜찮습니다. 다만 헤드셋이 본인의 얼굴을 가리지 않도록 주의해 주세요.
- 카메라 혹은 마이크 권한 요청 메시지창이 나오면 꼭 허용해 주세요.
- 카메라 인식이 안 되는 경우에는 재부팅하여 다시 접속하시기 바랍니다. 다른 프로그램이 실행되어 있는 상태에서 면접에 응시할 경우에는 카메라 인식이 안 될 수 있습니다.
- 사전에 네트워크 상태를 확인하여 안정된 네트워크 상태에서 면접을 진행해 주세요.
- 유선 인터넷 환경을 권장합니다. 무선인터넷(와이파이 혹은 핫스팟 연결) 환경은 네트워크 상태가 불안정하여 연결이 끊길 수도 있습니다. 연결이 끊기면 응시가 중지될 수 있으니 반드시 유선 인터넷 환경에서 면접을 진행하시기 바랍니다.
- **응시하기 전에 반드시 면접 웹페이지를 제외한 모든 프로그램과 웹페이지를 종료**하고 AI 영상 면접을 진행하시기 바랍니다. (그러지 않으면 카메라 혹은 마이크 인식이 안 될 수 있습니다.)

15-2 상호작용의 대상을 명확히 설명하는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

인공지능 서비스가 사용자와 직접적으로 상호작용하는 경우에 본 항목을 고려하여 만족 여부를 판단하십시오.

- 최근 인공지능 시스템을 의인화함으로써 사용자가 친밀감을 향상하고 사용성을 높이려는 서비스가 많아지고 있다. 그러나 인공지능 기술이 고도화되며 인간과 구분이 어려워져 사용자는 상호작용의 대상이 사람인지, 시스템인지 혼란을 겪을 수 있다. 따라서 서비스 제공자는 사용자가 상호작용하는 대상을 명확히 알림으로써 사용자가 겪을 혼란을 줄여야 한다.

15-2a 사용자가 인공지능과 상호작용하고 있다는 사실을 사용자에게 명확히 설명하였는가?

Yes ☐ No ☐ N/A ☐

- 공공기관이 제공하는 서비스의 활용 범위가 챗봇부터 복지까지 확장됨에 따라 사용자도 다양해지고 있다. 공공기관은 이처럼 광범위한 사용자층을 고려하여 사용자가 인공지능과 상호작용하고 있음을 명확하게 이해할 수 있는 방식으로 안내해야 한다.

참고 병무청 챗봇 '아라'의 상호작용 사례[75]

- 병무청 챗봇 '아라'는 이용자의 질의응답과 민원 처리 및 조회, 개인 맞춤형 서비스를 제공한다.
- 사용자가 '아라'를 사용할 때 "병무청 챗봇 아라입니다."라는 문장을 통해 상호작용의 대상이 시스템임을 알린다. 이로써 사용자는 실제 상담사와 혼동하고 서비스를 이용할 수 있다.

PART 3

부록

01 약어표

약어표

ACLU	American Civil Liberties Union
AGPL	Affero GPL
AI	Artificial Intelligence
BLEU	Bilingual Evaluation Understudy
BSD	Berkeley Software Distribution
CCTV	Closed-circuit Television
CNN	Convolutional Neural Network
CVE	Common Vulnerabilities and Exposures
DoS	Denial of Service
DSLR	Digital Single-Lens Reflex Camera
EU	Europe Union
GEN	Graph Extrapolation Network
GPL	General Public License
HCM	Human Capital Management
HRIAM	Human Rights Impact Assessment and Management
IAAE	International Association of Agricultural Economists
IBLF	international Business Leaders Forum
IEC	International Electrotechnical Commission
IFC	International Finance Corporation
ISO	International Organization for Standardization
KLD	Kullback-Leibler Divergence
LDA	Linear Discriminant Analysis
LGPL	Lesser GPL
LSTM	Long-Short Term Memory
mAP	mean Average precision
METEOR	Metric For Evaluation of Translation with Ecplicit Ordering
MORSE	Modular OpenRobots Simulation Engine
MPL	Mozilla Public License
NAP	National Action Plan
NBDT	Neural-Backed Decision Tree
NHS	National Health Service
OECD	Organisation for Economic Co-operation and Development
OSI	Open Source Initiative
PDPC	Personal Data Protection Commission
PPL	Perplexity

01 약어표

RMF	Risk Management Framework
ROS	Random Over Sampling
SimCLR	Simple framework Contrastive Learning of visual Representations
SMOTE	Sythetic Minority Over-Sampling Technique
SSA	Sensible and Specificity Average
SVM	Support Vector Machine
UNESCO	United Nations Educational, Scientific and Cultural Organization
XAI	eXplainable Artificial Intelligence

본 약어표에 정의된 약어 외, 인공지능 기술 용어에 대한 정의는 《2023 신뢰할 수 있는 인공지능 개발 안내서 – 일반 분야》를 참고하시기 바랍니다.

02 참고문헌

참고문헌

- [1] 세종매일, “세종시, 민원안내 인공지능(AI) 챗봇 서비스”, 2021.11.10.
- [2] 연합뉴스, “부산 주요 교차로 11곳 '스마트 교차로' 구축”, 2018.12.26.
- [3] 세계로컬타임즈, “인공지능이 CCTV 영상 속 쓰러진 사람 실시간 탐지한다”, 2021.11.18.
- [4] 한전KDN, CK PASS, [Online], Available: https://kdn.ckpass.copykiller.com/
- [5] KBS News, “AI 면접관은 공정?…‘차별 위험’ 검증 필요”, 2022.07.19.
- [6] 전자신문, “‘콜 몰아주기’ 카카오T에 257억 과징금”, 2023.02.14.
- [7] 사단법인 정보인권연구소, “[번역] 캐나다 정부 〈알고리즘 영향평가〉”, 2022.05.24.
- [8] 한국정보화진흥원, “공공기관 신뢰가능 AI 구현 실용가이드 – OECD 권고안의 적용 –”, 2019.
- [9] 국가인권위원회, “인권영향평가 및 관리에 관한 지침(HRIAM 가이드)”, 11-1620000-000562-01, 2014.12.
- [10] 송영규, 이정우, 한창희, "Delphi를 활용한 융합 서비스 설계에 관한 연구 :은행지점 도입용 금융 서비스 로봇 사례", 한국IT서비스학회지, 2020, vol.19, no.3, 통권 65호 pp. 1-15, 2020.06.30.
- [11] Government Digital Service Office for Artificial Intelligence, “A guide to using artificial intelligence in the public sector”, 2019.06.10.
- [12] 개인정보보호위원회, “인공지능(AI) 개인정보보호 자율점검표”, 2021.05.31.
- [13] 한국인터넷진흥원, “싱가포르의 인공지능(AI) 거버넌스 계획”, 2019.01.
- [14] 한국정보화진흥원, “인공지능 시대의 정부: 인공지능이 어떻게 정부를 변화시킬 것인가?”, 2017.
- [15] 국토교통부, “무인비행장치(드론), 이것만 지키면 모두가 안전해요!”, 2015.05.27.
- [16] AIHub, “민원 업무 자동화 인공지능 언어 데이터”, [Online], 2022.07.29.
- [17] 데이터 사이언스 스쿨, “Seaborn을 사용한 데이터 분포 시각화”, [Online], Available: https://datascienceschool.net/01%20python/05.04%20%EC%8B%9C%EB%B3%B8%EC%9D%84%20%EC%82%AC%EC%9A%A9%ED%95%9C%20%EB%8D%B0%EC%9D%B4%ED%84%B0%20%EB%B6%84%ED%8F%AC%20%EC%8B%9C%EA%B0%81%ED%99%94.html
- [18] 정하영, “이상탐지 활용 전자집단민원 추정 방법론에 관한 탐색적 연구: 창원시 시민의 소리 사례를 중심으로”, 정보화정책 제26권 제4호, 96p, 2019.12.
- [19] 한겨레, “인공지능이 인종차별 막말…위험성 현실화?”, 2016.03.25.
- [20] Matt Fredrikson, Somesh Jha, Thomas Ristenpart, “Model Inversion attacks that Exploit Confidence Information”, CCS’15, 1322–1333p, 2015.10.12.

02 참고문헌

- [21] 동빈나, “[꼼꼼한 논문 리뷰] Constructing Unrestricted Adversarial Examples with Generative Models [NIPS 2018] (AI보안)”, [Online], Available: https://www.youtube.com/watch?v=IDtaVjJoV4g&list=PLRx0vPvlEmdBQw6kRaod33aUkJ1YHLz3Y&index=8, 2020.03.06.
- [22] Alejandro Pena, Ignacio Serna, Aythami Morales, Julian Fierrez, "Bias in Multimodal AI: Testbed for Fair Automatic Recruitment", arXiv:2004.07173v1, 2p, 2020.04.
- [23] AIHub, “자연재해로 인한 생활시설 안전 데이터”, 2022.07.29.
- [24] Scribbr, "Sampling Methods | Types, Techniques & Examples", [Online], Available: https://www.scribbr.com/methodology/sampling-methods/, 2019.09.19.
- [25] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", arXiv:1106.1813v1, 330p, 2011.06.
- [26] Hui Han, Wen-Yuan Wang, Bing-Huan Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning", ICIC 2005, Part I, LNCS 3644, pp. 878 – 887, 2005.
- [27] Haibo He, Yang Bai, Edwardo A. Garcia, Shutao Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning" 2008 IEEE International Joint Conference on Neural Networks, 1322p, 2008.06.
- [28] Open Source Initiative, "The Open Source Definition", [Online], Available: https://opensource.org/osd
- [29] 강기봉, "머신러닝에 관한 OSS 라이선스 연구", 정보법학 제23권 제2호, 211p, 2019.08.
- [30] SCATTER LAB Tech, “하나의 조직에서 TensorFlow와 PyTorch 동시 활용하기”, [Online], Available: https://tech.scatterlab.co.kr/torch-to-tf-tf-to-torch/
- [31] OpenVINO, “Converting a TensorFlow Model”, [Online], Available: https://docs.openvino.ai/latest/openvino_docs_MO_DG_prepare_model_convert_model_Convert_Model_From_TensorFlow.html
- [32] TensorFlow, “저장된 모델 워크플로 마이그레이션”, [Online], Available: https://www.tensorflow.org/guide/migrate/saved_model
- [33] CVE Details, “Google Tensorflow : Vulnerability Statistics”, [Online], Available: https://www.cvedetails.com/product/53738/Google-Tensorflow.html?vendor_id=1224
- [34] CVE Details, “Pytorchlightning : Vulnerability Statistics”, [Online], Available: https://www.cvedetails.com/vendor/26132/Pytorchlightning.html
- [35] Aythami Morales, Julian Fierrez, "SensitiveNets: Learning Agnostic Representations with Application to Face Images", IEEE Transactions on Pattern Analysis and Machine Intelligence, 7p, 2020.08.

02 참고문헌

- [36] Davit Rizhinashvili, Abdallah Hussein Sham, Gholamreza Anbarjafari "Gender Neutralisation for Unbiased Speech Synthesising", 4p, 2022.05.
- [37] Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Noah A. Smith ,Yejin Choi ,"Challenges in Automated Debiasing for Toxic Language Detection", Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pp. 3143-3155, 2021.04.
- [38] ALLEGHENY COUNTY, "The Allegheny Family Screening Tool", [Online], Available: https://www.alleghenycounty.us/Human-Services/News-Events/Accomplishments/Allegheny-Family-Screening-Tool.aspx
- [39] 김효은, "인공지능 편향식별의 공정성 기준과 완화", 한국심리학회지, Korean Journal of Psychology: General 2021, Vol. 40, No. 4, 459-485, 2021.12.
- [40] Kaggle, “COMPAS Recidivism Racial Bias”, [Online], Available: https://www.kaggle.com/datasets/danofer/compass
- [41] Towards Data Science, “Evasion attacks on Machine Learning” (or “Adversarial Examples”), [Online], Available: https://towardsdatascience.com/evasion-attacks-on-machine-learning-or-adversarial-examples-12f2283e06a1
- [42] Di Jin, Zhijing Jin, Joey Tianyi Zhou, Peter Szolovits, "Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment", arXiv:1907.11932v6, 6p, 2020.08.
- [43] Yongkang Gong, Diqun Yan, Terui Mao, Donghua Wang, Rangding Wang, “Defending and Detecting Audio Adversarial Example using Frame Offsets”, KSII TRANSACTIONS ON INTERNET AND INFORMATION SYSTEMS VOL. 15, NO. 4, 1540p, 2021.04.
- [44] digitaltrends, "New Zealand attack shows that as A.I. filters get smarter, so do violators", [Online], Available: https://www.digitaltrends.com/social-media/new-zealand-attacks-spread-on-social-media/
- [45] Michał Kuźba, Przemyslaw Biecek, "What would you ask your ML model? Explainable AI chatbot", ML in PL Conference, 1p, 2019.11.
- [46] Anjali Khurana, Parsa Alamzadeh, Parmit K. Chilana, "ChatrEx: Designing Explainable Chatbot Interfaces for Enhancing Usefulness, Transparency, and Trust", IEEE Symposium on Visual Languages and Human Centric Computing (VL/HCC) conference, 1p, 2021.
- [47] Wencan Zhang, Brian Y Lim, "Towards Relatable Explainable AI with the Perceptual Process", CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 12p, 2022. 4.
- [48] Alvin Wan1, Lisa Dunlap, Daniel Ho, Jihan Yin1, Scott Lee, Suzanne Petryk, Sarah Adel Bargal, Joseph E. Gonzalez, "NBDT: NEURAL-BACKED DECISION TREE", arXiv:2004.00221v3, 1p, 2021.1.

02 참고문헌

- [49] Bernhard Scholkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, Yoshua Bengio, "Towards Causal Representation Learning", arXiv:2102.11107v1, 4p, 2021. 2.
- [50] IBM Research AI FactSheets 360, “Audio Classifier FactSheet”, [Online], Avaliable: http://aifs360.mybluemix.net/examples/max_audio_classifier
- [51] Google Research, Towards Reliability in Deep Learning Systems, [Online], Available: https://ai.googleblog.com/2022/07/towards-reliability-in-deep-learning.html
- [52] 조선비즈, ““동성애 개싫다”던 AI 이루다, 이젠 “존중해”… 직접 대화해보니”, 2022.03.16.
- [53] 김나연, “발화 목적과 상황에 따른 AI 에이전트 목소리의 성별과 높낮이에 대한 사용자 경험 조사”, Design Convergence Study, 89, vol.20, no.4, 121p, 2021.08.
- [54] NAVER Cloud, “네이버 클라우드 플랫폼 보안 모범 사례”, 2021.
- [55] The Wall Street Journal, "Delta Sues Chatbot Provider Over 2017 Breach", [Online], Available: https://www.wsj.com/articles/delta-sues-chatbot-provider-over-2017-breach-11565947801
- [56] 홍천호, 조영호, "화자식별 기반의 AI 음성인식 서비스에 대한 사이버 위협 분석", Journal of Internet Computing and Services (인터넷정보학회논문지), Vol 22., Issue 6., pp.33-40, 2021
- [57] Forbes, "Chatbots Can Be Weaponized — How To Defend Against These Attacks", [Online], Available: https://www.forbes.com/sites/forbestechcouncil/2022/05/19/chatbots-can-be-weaponized---how-to-defend-against-these-attacks/?sh=cb88b94785d2
- [58] 황승환, “TV 소리를 착각해 장난감 주문한 인공지능 스피커 에코”, 2017.01.09.
- [59] RightBrain, “AI 스피커 음성 인터랙션 오류상황에서의 사용자 감성 평가 분석”, [Online], Available: https://blog.rightbrain.co.kr/?p=10988
- [60] Github, “Accuracy(정확도), Recall(재현율), Precision(정밀도), 그리고 F1 Score”, [Online], Available: https://eunsukimme.github.io/ml/2019/10/21/Accuracy-Recall-Precision-F1-score/
- [61] Tony Park tistory, “[NLP] 언어모델의 평가지표 'Perplexity' 개념 및 계산방법”, [Online], Available: https://heytech.tistory.com/m/344
- [62] Ladun tistory, “[Metric] BLEU (Bilingual Evaluation Understudy)”, [Online], Available: https://ladun.tistory.com/m/71
- [63] wikidocs, “딥 러닝을 이용한 자연어 처리 입문”, [Online], Available: https://wikidocs.net/31695
- [64] javaspecialist, “주요 성능 지표”, [Online], Available: http://javaspecialist.co.kr/board/1062;jsessionid=5193BB88FB52F9036F9180512756D3CE

02 참고문헌

- [65] testworks, “자연어 생성과 평가 방법”, [Online], Available: https://blog.testworks.co.kr/nlp-generation-evaluation/
- [66] 스마트시티 대전 공식블로그, “[스마트 대전 생활 : 무인민원발급기와 누리온] 민원서비스 편리하게 이용하세요!”, 2022.08.26.
- [67] 대한민국정책브리핑, “‘지능형 국민비서·챗봇 민원상담’ 서비스 구축 사업 착수”, 2020.08.25.
- [68] 조국애, 윤재영, “금융 서비스 챗봇의 인터렉션 유형별 UX 평가”, vol.14, no.2, 통권 33호. 61-69p, 2019.05.14.
- [69] Google, “Feedback + Control”, [Online], Available: https://pair.withgoogle.com/chapter/feedback-controls/
- [70] iMerit, "Staying Ahead of Drift in Machine Learning Systems", [Online], Available: https://imerit.net/blog/staying-ahead-of-drift-in-machine-learning-systems-all-una/
- [71] 안병철, 김정렬, 이도형, “공공서비스의 역사적 변천과 특성”, 한국거버넌스학회보, 제16권, 제2호, 4P, 2009
- [72] midas HRi, “AI 역량검사 백서”, [Online], Available: https://www.midashri.com/aicc
- [73] BANK OF AMERICA, [Online], Available: https://promotions.bankofamerica.com/digitalbanking/mobilebanking/erica
- [74] 병무청, “AI 영상면접 사전 안내”, [Online], Available: https://www.mma.go.kr/boardFileDown.do?gesipan_id=69&gsgeul_no=1508872&ilryeon_no=1
- [75] 병무청, “병무청 챗봇, '아라'를 활용하는 법 A to Z!”, 2022.08.15.

한국정보통신기술협회

이강해 단장　곽준호 팀장
조경우 책임　채희문 책임
황재영 책임　변은영 선임
신예진 선임　박경은 전임
오상훈 전임　강상연 연구원

2023

신뢰할 수 있는 인공지능 개발 안내서 공공·사회 분야

초판 인쇄 2023년 06월 26일
초판 발행 2023년 07월 06일
저　　자 한국정보통신기술협회
발 행 인 최영해 · 김갑용
발 행 처 진한엠앤비
주　　소 서울시 서대문구 독립문로 14길 66 205호(냉천동 260)
전　　화 02) 364-8491(代) / 팩스 02) 319-3537
홈페이지 http://www.jinhanbook.co.kr
편집·제작 ㈜디자인여백플러스
등록번호 제25100-2016-000019호 (등록일자: 1993년 05월 25일)

ISBN 979-11-290-4925-4 (93550)　[정가 16,000원]

2023

신뢰할 수 있는 인공지능 개발 안내서 공공·사회 분야

Jinhan M&B

정가 16,000원

93550

9 791129 049254

ISBN 979-11-290-4925-4

과학기술정보통신부
Ministry of Science and ICT

TTA 한국정보통신기술협회
Telecommunications Technology Association