2024

신뢰할 수 있는 인공지능 개발 안내서

일반 분야

과학기술정보통신부
Ministry of Science and ICT

TTA 한국정보통신기술협회
Telecommunications Technology Association

일러두기

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

- 아울러, 안내서의 내용을 가공·인용하는 경우에는 반드시 '과학기술정보통신부·한국정보통신기술협회 《2024 신뢰할 수 있는 인공지능 개발 안내서 – 일반 분야》'의 출처를 밝혀 주시기 바랍니다.

- 본 안내서는 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 안내서는 인공지능 서비스 및 제품 개발·운영 중 고려해야 할 기술적 측면의 신뢰성 확보 방안을 다루고 있습니다. 이 외, 개인정보보호, 저작권 등 법적 측면의 확보 방안은 〈AI 개인정보보호 자율점검표〉, 〈생성형 AI 저작권 안내서〉 등의 관련 기관 안내서를 참고하시기 바랍니다.

- 본 안내서의 인공지능 동향 및 기술 정보는 2023년 12월 기준으로 서술되었습니다. 인공지능 기술의 발전에 따라 최신 연구 결과, 현실적인 적용방안 등을 검토하여 지속해서 개정할 예정입니다.

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

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

- 2023년에 공개된 분야별 개발 안내서를 통해 자율주행, 의료, 공공·사회 분야에 특화된 내용을 확인하실 수 있으며, 2024년에는 채용, 스마트치안, 생성 AI 기반 서비스 분야를 공개할 예정입니다.

CONTENTS

안내서 활용을 위한 체크리스트

안내서 활용을 위한 체크리스트

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

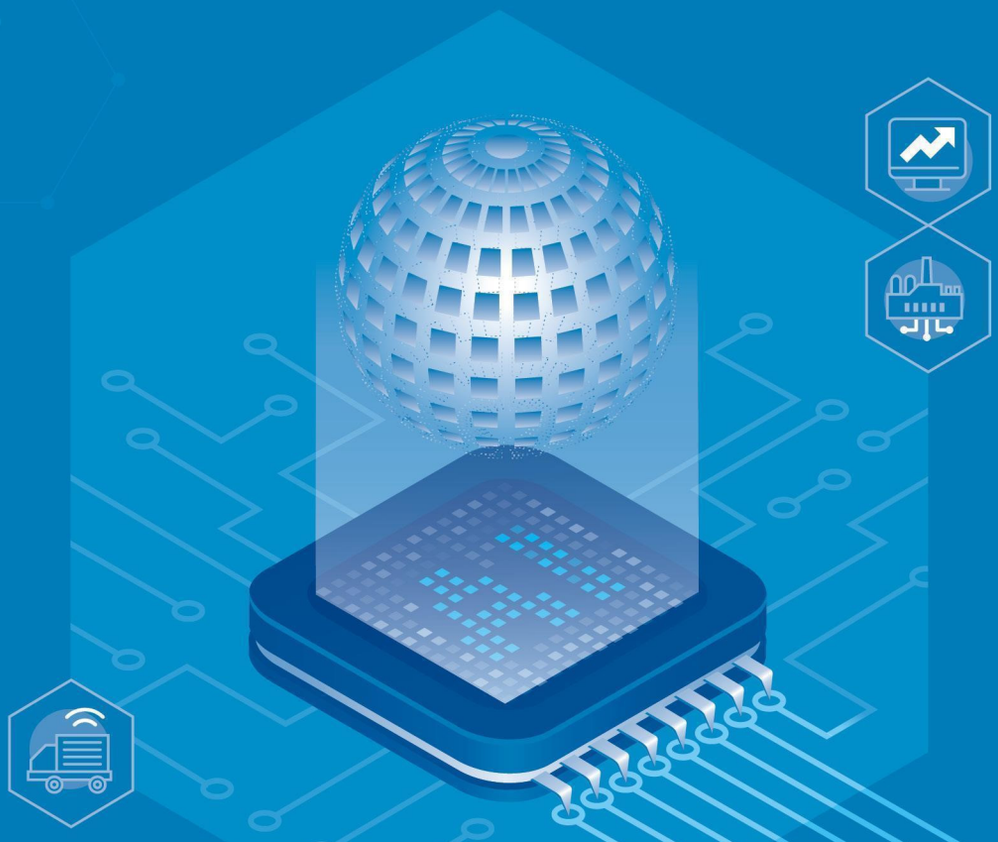
안내서 활용을 위한 체크리스트

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

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

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

2024 신뢰할 수 있는 인공지능 개발 안내서 | 일반 분야



PART 1

개요

01 안내서 발간 배경 및 목적

01 안내서 발간 배경 및 목적

현재 인공지능AI, Artificial Intelligence 기술은 다양한 분야에서 활용되고 있다. 인공지능의 활용 분야는 게임처럼 전통적으로 활용되던 분야는 물론, 스피커를 이용한 음성인식 서비스나 개인화된 비서 서비스 등 간단한 분야를 비롯해 검진 및 질병 진단, 자산관리를 비롯한 금융서비스, 자동차나 드론의 자율주행 등 복잡한 분야까지 폭넓게 아우른다. 이처럼 인공지능의 활용 분야가 넓어지고 일상에 점점 많은 영향을 미치면서 인공지능의 신뢰성trustworthiness 확보가 중요한 과제로 떠올랐다. 인공지능의 작동 원리나 메커니즘을 파악하기는 어렵고, 데이터 오염이나 편향 등의 문제로 인해 인공지능이 오류를 범할 가능성이 크기 때문이다. 특히 사람의 생명이나 공공안전과 직접 연관된 분야에까지 활용이 확대되면서 신뢰성의 중요성은 더욱더 커지고 있다.

이처럼 인공지능 신뢰성이 전 세계적인 관심사로 부상하면서 국제적으로 다양한 대응 방안이 마련되고 있다. 경제협력개발기구OECD, Organization for Economic Cooperation and Development는 인공지능 신뢰성 확보 권고안인 〈Recommendation of the Council on Artificial Intelligence('19.05)〉를 발표했으며, 유럽위원회EC, European Commission는 인공지능의 신뢰성을 실무자가 스스로 검증할 수 있도록 〈The Assessment List For Trustworthy Artificial IntelligenceALTAI for self assessment('20.07)〉를 공표했다. 국내에서도 이에 발맞춰 사람 중심의 인공지능 실현을 목표로 〈인공지능(AI) 윤리 기준('20.12)〉을 발표했다.

그러나 지금까지 나온 인공지능 신뢰성 원칙, 표준 등은 주로 윤리적 관점에서 추상적인 항목을 제시하고 있어 실무 현장에서 활용하기는 어렵다. 특히 인력과 연구 개발 투자 여력이 제한적인 중소기업은 직접 신뢰성 요구사항을 도출하거나 검증체계를 마련하기 어렵다.

본 안내서는 이러한 현실적인 문제점을 해결하고자 작성되었다. 미국, 유럽 등 주요 선진국과 국제기구들에서 발표한 권고안 및 가이드를 참고하여, 자율적으로 점검이 가능한 15개 개발 요구사항과 69개 검증항목을 제시하고 있다.

개발자나 기획자 등 인공지능 서비스 개발 실무자들은 본 개발 안내서에 제시된 항목을 참고하여 최소한의 신뢰성을 확보하는 한편, 신뢰성을 확보하려면 무엇이 중요한지 이해하는 데 도움이 될 것이다. 나아가 개발 안내서 내용을 바탕으로 서비스에 적합한 요구사항과 검증방법을 마련함으로써 신뢰성 높은 인공지능 서비스를 개발할 수 있을 것이다. 본 개발 안내서를 통해 우리나라 인공지능 관련 기업 및 기관들이 더욱더 성숙한 인공지능 기술을 확보하고, 글로벌 경쟁력을 가질 수 있는 기초자료가 되길 희망한다.

02 인공지능 신뢰성 동향

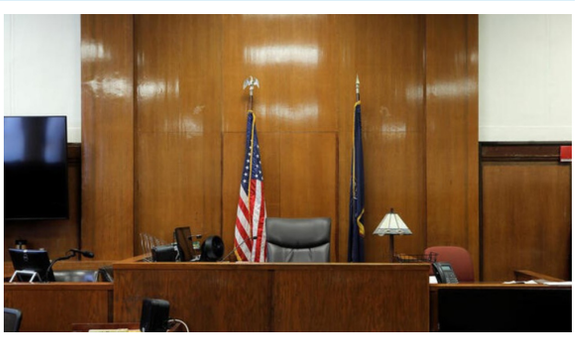
현재 세계의 주요 국가와 표준 관련 기구, 기술단체들은 인공지능 신뢰성 확보를 위해 다양한 원칙과 방안들을 제시하고 있다. 본 절에서는 인공지능이 폭넓게 활용되면서 발생하는 문제점이나 위험을 알아보고 이를 해결하기 위해 국내외에서 진행 중인 관련 정책 및 연구 동향을 살펴보고자 한다.

2.1. 인공지능 확산에 따른 문제점

인터넷이나 스마트폰 등의 신기술이 우리의 일상과 사회를 빠르고 편리하게 변화시킴과 동시에 새로운 문제가 등장했듯, 사회와 인공지능이 활용되면서 새로운 문제나 위험에 대한 우려가 등장했다. 이러한 문제 상당수는 사회적, 윤리적인 문제이므로 기술을 개선하거나 신기술을 도입하는 것만으로는 해결할 수 없다. 실제로 사용자가 인공지능 기술을 악의적인 목적으로 활용하거나 인공지능이 반사회·반인륜적인 판단을 하여 사회적 혼란을 유발하는 사건이 종종 발생하고 있다.

▼ 인공지능 사고사례

사고사례 1: 챗GPT가 알려준 가짜 판례 제출한 美 변호사



미국의 한 변호사가 챗GPT가 제시한 판례가 담긴 의견서를 법원에 제출. 그러나 해당 판례는 모두 가짜인 것으로 드러났으며, 챗GPT에 '진짜 판례가 맞느냐'고 거듭 확인했으나, 끝까지 '그렇다'고 주장 ('23.5.) *https://www.lawtimes.co.kr/news/187960*

시사점
생성형 인공지능의 할루시네이션[hallucination] 발생에 따른 우려

사고사례 2: 가짜 인터뷰 기사를 만든 인공지능 챗봇

독일의 한 주간지는 인공지능 챗봇을 활용해 인터뷰 내용을 만들어 만난 적도 없는 유명 레이싱 선수의 가짜 인터뷰 기사를 대서 특필 ('23.4.)
http://www.banronbodo.com/news/articleView.html?idxno=21640

시사점
인공지능을 통해 가짜 정보를 제작·유포하여 사회적 혼란 유발

사고사례 3: 체스 경기 도중 사고 발생

모스크바 체스 토너먼트에서 체스 로봇과 경기를 벌이던 7세 소년이 안전 규정을 어겨 로봇에 의해 손가락이 부러짐('22.7.)

http://www.aitimes.com/news/articleView.html?idxno=145954

시사점
인공지능과 연동된 하드웨어로 인한 안전사고 발생

사고사례 4: 피부색에 따라 차별하는 이미지분석 서비스

Google의 Vision API는 흑인 여성이 입고 있는 의사 가운을 '길거리 패션'으로 분류. 그러나, 같은 여성에 피부만 밝게 했더니 '정식 의복'으로 분류('22.7.)

https://www.seerinteractive.com/insights/image-ai-bias

시사점
피부색에 따른 차별적 결과 도출

2.2. 인공지능 신뢰성 개념

앞서 사례에서 살펴봤듯 인공지능 제품·서비스는 단지 '구현할 수 있는가?'라는 기술적 측면뿐 아니라 '이 제품·서비스가 존재해도 괜찮은가?'라는 윤리적 측면도 검토해야 한다. 특히 인공지능이 다양한 분야에 활용되면서 인공지능 시스템과 학습 모델에 윤리적인 결함이 있는데도 이를 인지하지 못한 채 사용될 경우 매우 큰 파급효과를 낳을 수 있다. '인공지능 신뢰성'이란 데이터 및 모델의 편향, 인공지능 기술에 내재한 위험과 한계를 해결하고, 인공지능을 활용하고 확산하는 과정에서 부작용을 방지하기 위해 준수해야 하는 가치 기준을 말한다. 주요 국제기구를 중심으로 인공지능 신뢰성을 확보하는 데 필수적인 요소가 무엇인지 활발한 논의가 이루어지고 있다. 일반적으로 안전성, 설명가능성, 투명성, 견고성, 공정성 등이 신뢰성을 확보하는 데 필수적인 요소로 거론되고 있다.

▼ 인공지능 신뢰성의 주요 핵심 속성 및 의미

핵심 속성	의미
안전성safety	인공지능이 판단·예측한 결과로 시스템이 동작하거나 기능이 수행됐을 때 사람과 환경에 위험을 줄 가능성이 완화 또는 제거된 상태
설명가능성explainability	인공지능의 판단·예측의 근거와 결과에 이르는 과정이 사람이 이해할 수 있는 방식으로 제시되거나, 문제 발생 시 문제에 이르게 한 원인을 추적할 수 있는 상태
투명성transparency	인공지능이 내리는 결정에 대한 이유가 설명 가능하거나 근거가 추적 가능하고, 인공지능의 목적과 한계에 대한 정보가 적합한 방식으로 사용자에게 전달되는 상태
견고성robustness	인공지능이 외부의 간섭이나 극한적인 운영 환경 등에서도 사용자가 의도한 수준의 성능 및 기능을 유지하는 상태
공정성fairness	인공지능이 데이터를 처리하는 과정에서 특정 그룹에 대한 차별이나 편향성을 나타내거나, 차별 및 편향을 포함한 결론에 이르지 않는 상태

※ 프라이버시privacy, 지속가능성sustainability 등도 핵심 속성 중 하나로서 다양하게 논의 중

참고 주요 기관에서 논의 중인 인공지능 신뢰성 개념

- (국제표준화기구, ISO) 신뢰성의 세부 속성으로 가용성availability, 회복탄력성resiliency, 보안성security, 프라이버시, 안전성, 책임성, 투명성, 통합성integrity 등 제시(ISO/IEC TR 24028:2020)
- (경제협력개발기구, OECD) 지속가능한 사회와 인간 중심의 가치에 부합하고 투명성, 설명가능성, 견고성 및 안전성을 갖춘 인공지능('19)
- (美 국립표준연구소, NIST) 유효하고 신뢰할 수 있으며, 안전하고, 회복력있고, 책임있고 투명하며, 설명가능하고 해석 가능하며, 개인정보를 강화하고, 유해한 편향을 관리하는 공정성을 포함하는 개념('23)
- (유럽위원회, EC) 인공지능은 활용 및 동작이 합법적이며, 윤리적이고 기술적·사회적으로 견고해야 함('19)

2.3. 국내외 인공지능 신뢰성 정책 및 연구 동향

유럽위원회, 미국 등 주요국들은 인공지능의 신뢰성 확보가 인공지능의 사회적·산업적 수용과 발전의 전제 조건으로 정의하고 신뢰성 확보 정책을 추진하고 있다. 또한 산업계 및 학계에서도 관련 기술 개발을 중심으로 신뢰성 확보를 위한 연구가 활발하다. 구체적으로 유럽위원회, 미국 등 주요국에서는 인공지능 신뢰성을 확보하는 데 필요한 정책과 규범을 본격적으로 마련하고, 이와 함께 국가 차원의 인공지능 전략 핵심 요소로 Trustworthy AI, Safe AI 등을 명시했다. 특히, 유럽위원회는 지난 2021년도에 규제안을 제시하는 등 선제적으로 신뢰성 확보를 위한 법제화 움직임을 보이고 있다. 한편, 민간 부문에서는 인공지능 신뢰성 확보를 위한 가이드라인을 마련하여 인공지능의 신뢰성을 자율적으로 점검하고 확보할 수 있는 환경을 조성하고자 노력하고 있다. 기술 분야에서는 미국, 유럽 등 주요국의 학계와 글로벌 기업이 인공지능 신뢰성 확보에 필요한 제반 기술을 개발 중이다. 우리나라도 글로벌 동향에 맞춰 〈인공지능(AI) 윤리 기준('20.12)〉, 〈신뢰할 수 있는 인공지능 실현 전략('21.5)〉, 〈인공지능 일상화 및 산업 고도화 계획('23.1)〉을 발표하며 정책 및 연구 개발 양면에서 발 빠르게 움직이고 있다.

▼ 해외 주요 산·학·연 인공지능 신뢰성 연구 동향

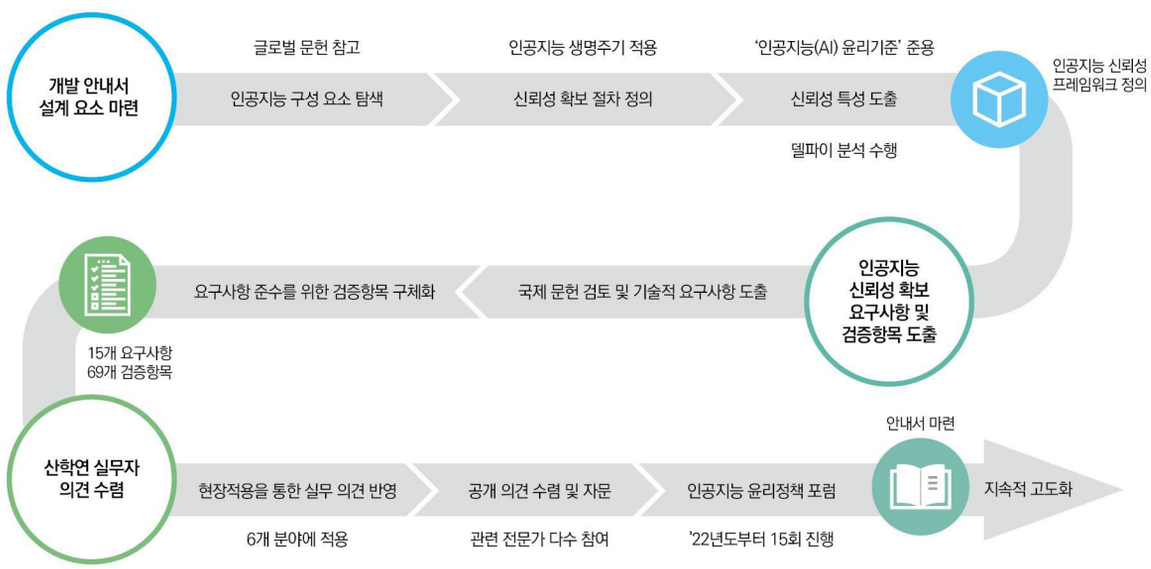
기관명	활동 및 내용
美 방위고등연구계획국DARPA	지능형 시스템에 대한 안전성·신뢰성 확보 연구(Assured Autonomy) 및 설명가능한 인공지능XAI, eXplainable AI 연구·개발 프로젝트 수행 중
美 국립표준기술연구소NIST	글로벌 기업 및 연구기관과 공동으로 기업·실무자가 활용 가능한 인공지능 위험관리 프레임워크Risk Management Framework를 개발하고, 안전하고 신뢰할 수 있는 인공지능을 만들기 위한 노력을 지원하기 위해 미국 인공지능 안전 연구소(USAISI) 및 관련 컨소시엄을 설립
美 스탠퍼드 대학	매년 인공지능 기술 수준 및 동향을 수록한 'AI Index'를 공개하고, 인공지능 안전성 관련 연구 수행 중
IBM	'Trusted AI'를 모토로, 공정성·설명가능성·견고성 확보를 위한 내부 업무 지침 및 백서, 개발·검증 도구 공개
Microsoft	'Responsible AI'를 모토로, 공정성·설명가능성·투명성 확보를 위한 내부 지침 및 업무 표준, 개발·검증 도구 공개
Google	'Responsible AI' 개발을 위한 원칙 제정, 신뢰성 확인 및 검증을 위한 가이드라인과 도구 공개

▼ 글로벌 인공지능 신뢰성 관련 정책 동향

국가	주요 정책(연도)
한국	• 2023: 인공지능 일상화 및 산업 고도화 계획, 초거대AI 경쟁력 강화 방안, 전국민 AI 일상화 실행계획 • 2021: 사람이 중심이 되는 인공지능을 위한 신뢰할 수 있는 인공지능(AI) 실현 전략 • 2020: 사람이 중심이 되는 인공지능(AI) 윤리기준 • 2019: 인공지능 국가전략
유럽위원회	• 2023: 인공지능법 • 2019: 신뢰할 수 있는 인공지능 윤리 가이드라인
유네스코	• 2021: 인공지능 윤리 권고
G7	• 2023: 히로시마 AI 프로세스
미국	• 2023: 안전하고 신뢰할 수 있는 AI에 관한 행정 명령, 자발적인 인공지능 위험 관리 약속 • 2022: 인공지능 권리장전
영국	• 2023: 인공지능(규제) 법안, 경쟁적인 AI 시장 선도 및 소비자 보호를 위한 원칙, AI 안전성 정상회의 – 브레츨리 선언
중국	• 2023: 생성 AI 서비스 규제 규정
일본	• 2022: 인공지능 원칙 구현을 위한 거버넌스 가이드라인, AI전략 2022 • 2019: 인공지능 활용전략
싱가포르	• 2020: 인공지능 거버넌스 프레임워크

03 안내서 마련 과정

▼ 안내서 마련 과정

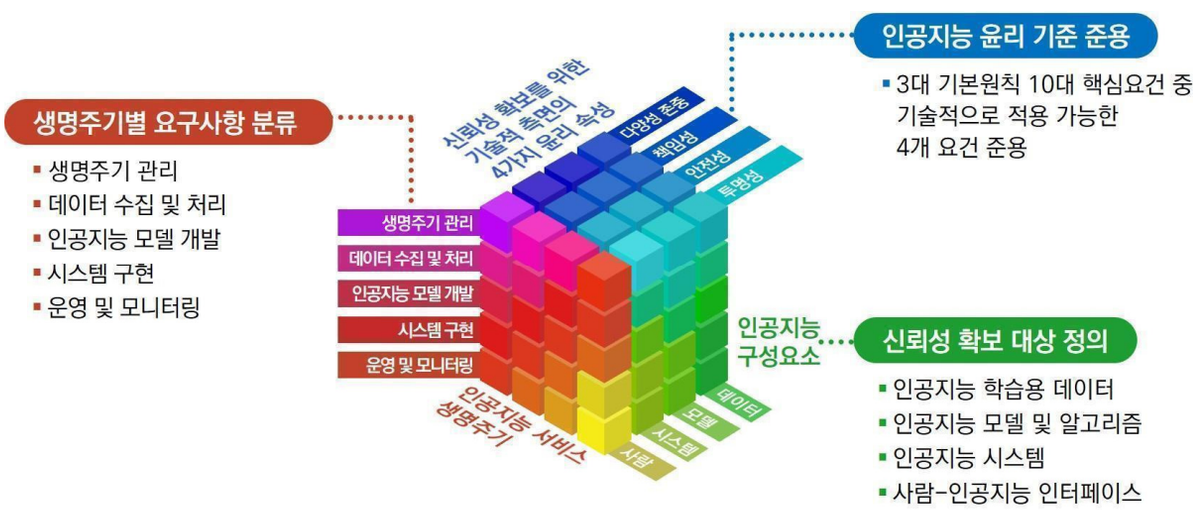


발간 배경에서 밝힌 바와 같이, 그간 국내외 많은 기관 및 기업이 인공지능 신뢰성 확보를 위한 윤리 원칙과 지침, 가이드라인을 내놓았으나, 기술적 관점에서 상세한 방법론을 제시한 사례는 아직 없었다. 따라서 인공지능 제품 및 서비스 개발 현장에서 데이터 과학자, 모델 개발자 등 이해관계자들이 실무 관점에서 신뢰성 확보에 참고할 수 있는 지침서 성격의 자료를 만들고자 했다. 이를 위해, 2021년도에는 국내외 많은 문헌을 기반으로 인공지능 신뢰성의 개념 및 신뢰성 확보를 위한 원칙, 항목들을 분석하고 종합적으로 정리하였다. 바탕이 된 문헌에는 과학기술정보통신부가 2020년 12월에 마련한 〈인공지능(AI) 윤리기준〉을 비롯해, EC, OECD, UNESCO 등 국제 사회가 발표한 원칙 및 프레임워크 등이 있다. 이렇게 작업본을 마련한 후 12회에 걸쳐 학계 및 산업계 전문가와 실무자 250여 명을 대상으로 개방형 자문과 검토를 거쳤다. 또한, 인공지능 제품·서비스를 제공하는 기업과 협업해 안내서의 현장 적용과 컨설팅 등 공동 연구를 진행하여 케이스 스터디를 마련하고 피드백을 받는 과정을 거쳐 실무 활용도를 높이고자 했다. 그리하여 2021년 11월 '인공지능 신뢰성 제고를 위한 공개 정책 세미나'를 개최하여 초안을 공개하였고, 2022년도 및 2023년도에는 과학기술정보통신부가 주관하는 '인공지능 윤리 정책 포럼'에서 신뢰성 확보를 위한 기술적·정책적 방안에 대한 논의를 통해 2년여 기간 동안 의견 수렴을 진행하였다. 이렇게 2021년 1월부터 2023년 12월까지 약 3년에 걸쳐 본 안내서를 개발하였으며, 앞으로도 지속해서 최신 기술 동향과 산업계 흐름·인식을 반영해 나갈 예정이다.

3.1. 개발 안내서 설계 요소(인공지능 서비스 구성, 생명주기, 신뢰성 특성)

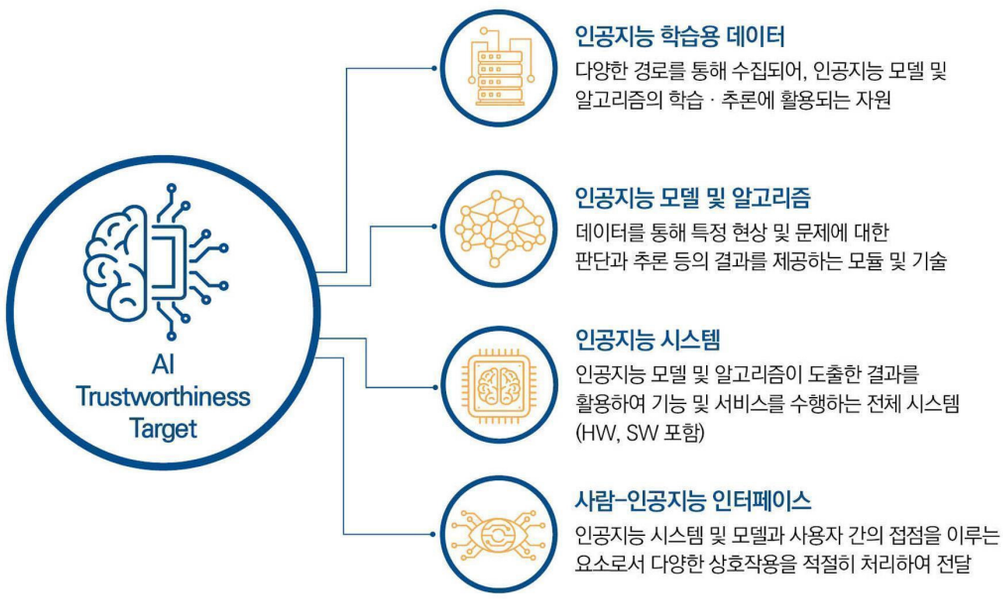
안내서 개발 과정 중 가장 우선적으로 신뢰성 확보를 위해 어떤 요소들이 실무적으로 고려되어야 하는지 탐색해보았고, 그 결과 세 가지 설계 요소를 도출하여 안내서에 반영하였다. 각 설계 요소들은 요구사항과 검증항목 마련 시 모두 반영되었으며, 이러한 접근법을 아래 그림과 같이 매트릭스matrix 형태로 체계화하여 '인공지능 신뢰성 프레임워크'로 정의하였다.

▼ 인공지능 신뢰성 프레임워크



첫 번째는 인공지능 구성요소이다. 인공지능을 구성하는 4가지 요소는 인공지능 학습용 데이터, 학습과 추론 기능을 수행하는 인공지능 모델 및 알고리즘, 실제 기능을 구현할 시스템, 사용자와 상호작용하기 위한 인터페이스가 있다. 각 구성 요소들은 개별적으로 또는 통합적으로 인공지능 서비스의 생명주기에 따라 개발, 검증 및 운영된다. 따라서 구성요소별 신뢰성 확보 방안을 고민하고, 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 배제되었는지 검증
인공지능 모델 및 알고리즘	인공지능이 모델 및 알고리즘에 따라 안전한 결과를 도출하며, 이에 대한 설명이 가능한지, 악의적인 공격에 견고한지 등을 검증
인공지능 시스템	인공지능 모델 및 알고리즘이 적용된 전체 시스템을 대상으로 인공지능이 추론한 대로 작동하는지, 인공지능이 잘못 추론한 경우의 대책이 존재하는지 등을 검증
사람-인공지능 인터페이스	인공지능 시스템 사용자·운영자 등이 인공지능 시스템의 동작을 쉽게 이해할 수 있으며, 인공지능의 오작동 시 사람에게 알리거나 제어권을 이양하는지 등을 검증

두 번째, 인공지능 서비스 생명주기는 첫 번째에서 살펴본 인공지능 서비스 구성 요소들을 구현하고 운영하는 일련의 절차를 말한다. 기존 소프트웨어 시스템에서 다루는 공학 프로세스나 생명주기와 비슷하나, 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 6~8가지 단계로 구분한다. 대표적으로 OECD와 ISO/IEC에서 제시한 생명주기가 있는데, 본 안내서는 두 기구에서 제시한 생명주기를 대표성 있는 사례로 참고하여, 실무자들이 쉽게 활용할 수 있도록 각 생명주기 단계의 성격과 활동을 왜곡하지 않는 선에서 아래와 같이 5가지 단계로 정리하였다.

▼ 인공지능 서비스 생명주기별 주요 활동

생명주기 단계	주요 활동
1. 생명주기 관리	- 인공지능 시스템 관리 감독 조직 및 방안 마련 - 인공지능 시스템 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 특성feature 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 모델에 대한 성능평가
4. 시스템 구현	- 문제 발생 대비 안전모드 구현 및 알림 절차 수립 - 인공지능 시스템 검증 및 사용자 설명에 대한 평가
5. 운영 및 모니터링	- 시스템 모니터링 및 인공지능 모델 재학습을 통한 성능 보장 - 모델 편향 탐지, 공정성, 설명가능성 등 시스템 신뢰성 모니터링 - 치명적 문제 발생 시 해결 방안 마련

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

세 번째, 인공지능 신뢰성에 필요한 특성을 정의하고자 '인공지능 윤리기준'의 10대 핵심요건을 준용하여 기술적 관점에서 필요한 요구사항과 검증항목으로 '다양성 존중', '책임성', '안전성', '투명성'을 도출했다.

EC, OECD, IEEE 및 ISO/IEC 등의 국제기구는 인공지능 신뢰성의 하위 속성들을 세분화해 제시한다. 특히, ISO/IEC 24028:2020은 신뢰성 확보에 필요한 고려사항의 형태로 키워드를 제공한다. 여기에는 투명성, 통제가능성, 견고성, 복구성, 공정성, 안전성, 개인정보보호, 보안성 등이 포함되나, 키워드 간의 관계나 신뢰성과의 연관성은 정의되지 않았다. 이처럼 관점에 따라 유사해 보이지만 조금씩 다른 용어들이 여러 문헌에서 제각각 달리 정의되고, 아직 합의된 속성 분류나 정의는 없는 상황이다. 이에, 앞서 언급한 EC, OECD, IEEE, ISO/IEC 등 여러 기구에서 제시한 속성과 키워드를 종합적으로 분석하고, 국내 학계·연구계·산업계 전문가의 의견을 수렴해 합의점을 모색했다. 이처럼 폭넓은 의견 공유 과정을 거쳐 인공지능 신뢰성 속성을 도출한 후, 이를 국가 인공지능 윤리기준의 10대 요건에 대응시켜서 기술적 측면에서 다룰만한 특성을 최종 선정하였다. 각 특성에 대한 정의는 아래와 같다.

▼ 인공지능 신뢰성 특성

신뢰성 특성	정의
다양성 존중	인공지능이 특정 개인이나 그룹에 대한 차별적이고 편향된 관행을 학습하거나 결과를 출력하지 않으며, 인종·성별·연령 등과 같은 특성과 관계없이 모든 사람이 평등하게 인공지능 기술의 혜택을 받을 수 있는 것 - 관련 속성: 공정성·공평성fairness, 정당성justice - 관련 키워드: 편향bias, 차별discrimination, 편견prejudice, 다양성diversity, 평등equality - 국제표준(ISO/IEC TR 24027:2021 – Bias in AI systems and AI aided decision making)에서는 공정성을 정의하지 않는다. 공정성은 복잡하고 문화·세대·지역 및 정치적 견해에 따라 다양하여 사회적으로나 윤리적으로 일관되게 정의하기 힘들기 때문이다.
책임성	인공지능이 생명주기 전반에 걸쳐 추론 결과에 대한 책임을 보장하기 위한 메커니즘이 마련되어 있는 것 - 관련 속성: 책무성responsibility, 감사가능성auditability, 답변가능성answerability - 관련 키워드: 책임liability - 국제표준(ISO/IEC TR 24028:2020 – Overview of trustworthiness in artificial intelligence)에서의 정의: 엔터티Entity의 작업이 해당 엔터티에 대해 고유하게 추적될 수 있도록 하는 속성
안전성	인공지능이 인간의 생명·건강·재산 또는 환경을 해치지 않으며, 공격 및 보안 위협 등 다양한 위험에 대한 관리 대책이 마련되어 있는 것 - 관련 속성: 보안성security, 견고성·강건성robustness, 성능보장성reliability, 통제가능성·제어가능성controllability - 관련 키워드: 적대적 공격adversarial attack, 회복탄력성resilience, 프라이버시privacy - 국제표준(ISO/IEC TR 24028:2020)에서의 정의: 용인할 수 없는 위험risk으로부터의 자유
투명성	인공지능이 추론한 결과를 인간이 이해하고 추적할 수 있으며, 인공지능이 추론한 결과임을 알 수 있는 것 - 관련 속성: 설명가능성explainability, 이해가능성understandability, 추적가능성traceability, 해석가능성interpretability - 관련 키워드: 설명가능한 인공지능XAI, eXplainable AI, 이해도comprehensibility - 국제표준(ISO/IEC TR 29119-11:2020 – Guidelines on the testing of AI-based systems)에서의 정의: 시스템에 대한 적절한 정보가 관련 이해 관계자에게 제공되는 시스템의 속성

※ 개인정보보호 관련 내용은 개인정보보호위원회의 〈AI 개인정보보호 자율점검표('21.5)〉로 갈음

위와 같이 인공지능 신뢰성 확보를 위한 다양한 속성들이 있으며, 각 신뢰성 속성들에 대한 정의를 파악하는 것뿐만 아니라 신뢰성 속성 간의 상호의존 관계 역시 중요하게 고려되어야 한다. 예를 들어, 인공지능 서비스에 대한 과도한 투명성 요구는 프라이버시 관련 위험을 초래할 수 있다. 또한, 설명가능성만으로는 투명성을 보장하기에 부족하지만, 설명가능성은 투명성을 확보하기 위한 중요한 요소 중 하나이다. 따라서, 인공지능 신뢰성 속성에 대한 충분한 이해를 바탕으로 인공지능 서비스를 제공하는 것이 중요하며, 해당 인공지능 서비스가 고려한 속성에 대해 적절하게 이행하는지 지속해서 검토해야 한다.

3.2 인공지능 신뢰성 확보 요구사항 및 검증항목 도출

다음 단계로 구체적인 요구사항과 검증항목을 도출했다. 우선 표준화기구, 기술단체, 국제기구, 주요국에서 인공지능 신뢰성 확보를 위해 발표한 정책, 권고안, 그리고 표준을 기반으로 기술적 요구사항을 도출하고 구체화하였다. 이와 함께 AI 개인정보보호 자율점검표('21.5), 금융분야 AI 가이드라인('21.7) 등 국내에서 인공지능 신뢰성 확보를 목적으로 발표된 점검표 등을 검토했다. 검토 과정에서 개발 안내서에 필요한 내용은 반영하고 중복된 내용은 제거하거나 통합하였다. 참고문헌은 다음과 같다.

▼ 인공지능 신뢰성 관련 주요 참고문헌

기관명	발간년월	권고 및 표준안 명
대한민국 정부	2020.11	국가 인공지능(AI) 윤리기준
한국정보통신기술협회 (TTA)	2023.12	TTAK.KO-10.1497, 인공지능 시스템 신뢰성 제고를 위한 요구사항
美 백악관	2023.10	Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence
	2023.07	Voluntary commitments – underscoring safety, security, and trust – mark a critical step toward developing responsible AI
유럽위원회	2023.12	Artificial Intelligence Act
유네스코 (UNESCO)	2020.07	The Assessment List for Trustworthy Artificial Intelligence
	2021.11	Recommendation on The Ethics of Artificial Intelligence
국제표준화기구 (ISO/IEC)	2020.05	ISO/IEC TR 24028:2020, Information Technology – AI – Overview of Trustworthiness in artificial intelligence
	2021.03	ISO/IEC TR 24029-1:2021, Artificial Intelligence (AI) – Assessment of the robustness of neural networks – Part 1: Overview
	2021.11	ISO/IEC TR 24027:2021, Information technology – Artificial Intelligence (AI) – Bias in AI systems and AI aided decision making
	2023.02	ISO/IEC 23894:2023, Information Technology – Artificial Intelligence (AI) – Guidance on risk management
美 국립표준연구소 (NIST)	2023.01	NIST AI Risk Management Framework 1.0
세계경제포럼 (WEF)	2020.01	Companion to the Model AI Governance Framework
경제협력개발기구 (OECD)	2019.05	Recommendation of the Council on Artificial Intelligence
Google	2019.05	People + AI guidebook
유럽전기통신표준협회 (ETSI)	2021.03	Securing Artificial Intelligence (SAI 005) – Mitigation Strategy Report

이를 통해 최종 도출한 요구사항은 아래 표와 같으며, 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 인공지능 시스템의 추적가능성 및 변경이력 확보		✓		✓
요구사항 05 데이터의 활용을 위한 상세 정보 제공		✓		✓
요구사항 06 데이터 견고성 확보를 위한 이상 데이터 점검			✓	
요구사항 07 수집 및 가공된 학습 데이터의 편향 제거	✓	✓		✓
요구사항 08 오픈소스 라이브러리의 보안성 및 호환성 점검		✓	✓	
요구사항 09 인공지능 모델의 편향 제거	✓			
요구사항 10 인공지능 모델 공격에 대한 방어 대책 수립			✓	
요구사항 11 인공지능 모델 명세 및 추론 결과에 대한 설명 제공		✓		✓
요구사항 12 인공지능 시스템 구현 시 발생 가능한 편향 제거	✓			
요구사항 13 인공지능 시스템의 안전 모드 구현 및 문제발생 알림 절차 수립		✓	✓	✓
요구사항 14 인공지능 시스템의 설명에 대한 사용자의 이해도 제고				✓
요구사항 15 서비스 제공 범위 및 상호작용 대상에 대한 설명 제공		✓		✓

3.3. 산학연 실무자 의견 수렴

신뢰성 확보를 위한 요구사항을 도출한 후에는 각 항목을 기술적 타당성, 효용성 및 포괄성 등의 관점에서 검토한 후 고도화했다. 각각의 세부 검증항목이 요구사항에 해당하는 내용이 맞는지(타당성), 개발 현장에서 실무적으로 활용 가능한 내용인지(효용성), 검증을 위한 내용들이 과거부터 지금까지 연구 내용을 폭넓게 포함하는지(포괄성) 확인했다. 이를 위해 기획자, 개발 프로젝트 리더, 교수, 국가연구소 책임연구원, 관련 국가 정책 담당자 등 다수의 인공지능 분야 전문가가 참여하여 직접 검토하고 자문했으며, 다양한 검토 의견을 수렴하여 반영하였다.

04 안내서 활용 대상

4.1. 활용 대상(대표 이해관계자·협력 대상) 정의 배경 및 기준

본 개발 안내서는 인공지능 제품 및 서비스의 개발 과정에 참여하는 다양한 조직과 개인이 활용할 수 있다. 특히, 기술적 관점에서 신뢰성에 중점을 두어야 하는 기획자, 아키텍트, 개발자, 품질관리자 등의 이해관계자들이 주요 대상이다. 이해관계자들은 제품·서비스의 신뢰성을 확보하기 위해 요구사항을 충족시키는 데 주력해야 하며, 이는 아래에 제시된 표를 통해 확인할 수 있다. 물론, 신뢰성과 연관된 문제가 발생했을 때 관련된 모든 책임을 이해관계자가 부담해야 한다는 의미는 아니다. 대표 이해관계자는 인공지능 생명주기 단계마다 요구사항을 만족시키기 위한 대책을 수립하며, 자가 검증 시 각 검증항목의 만족 여부를 체크하는 주요 역할을 담당한다. 이 과정에서 효과적인 협력 체계의 필요성이 강조된다. 따라서, 대표 이해관계자는 한 명 이상의 협력 대상과 긴밀하게 협력하며, 이들 간의 협력 관계는 부록 3에 기술되어 있다.

대표 이해관계자와 협력 대상은 한국SW산업협회KOSA가 국가직무능력표준NCS를 기반으로 개발한 IT분야역량체계ITSQF에 근거해 정립되었다. 이를 통해, 국내 기업들이 본 개발 안내서를 활용하고자 할 때 참고할 수 있도록 하였다. 또한, 각 기업의 다양한 직무 체계에 맞게 적용하기 위해, 부록 4에 제시된 각 직업·직무에 대한 정의를 참고하여 직무별 역할을 확인할 수 있다.

▼ 인공지능 생명주기 단계별 신뢰성 확보를 위한 대표 이해관계자

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

4.2. 활용 기업 및 서비스 유형에 따른 적용 방안

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

또한, 기업에서 제공하는 인공지능 서비스 유형에 따라 적용 방안이 달라질 수도 있다. 다음 페이지에서 제시한 대표 이해관계자 및 협력 대상의 분류는 소비자 대상B2C, Business-to-Consumer 서비스 제공 기업에서 참고하기에 적합하다. 반면, 기업간B2B, Business-to-Business 서비스를 제공하는 경우에는 관련 표준(TTAK.KO-10.1497, 인공지능 시스템 신뢰성 제고를 위한 요구사항)을 참고하는 것이 더 활용도가 높을 것이다. 표준에 근거한 요구사항별 이해관계자는 부록 3을 참고하기 바란다.

이외에도 개발 안내서를 활용하는 환경에 따라 그 적용 방안은 다양해질 수 있다. 예를 들어, 개발하는 인공지능 제품·서비스의 산업 분야마다 해당 분야의 전문가 역시 적극 협업할 필요가 있다. 그리고 만약 대표 이해관계자 및 협력 대상이 모두 존재하지 않는 소규모 기업에서는 외부 전문가의 도움을 받을 수도 있다. 따라서, 다음 페이지와 부록3, 4에 제시된 대표 이해관계자, 협력 대상, 직무별 역할 등의 내용은 참고 자료로 활용하는 것이 좋다.

05 안내서 활용 방법

본 안내서는 범용성을 갖추고자 인공지능 신뢰성 관점에서 기술적 고려가 필요한 요구사항 및 검증항목을 포괄적으로 수록하였다. 따라서, 기업 내부의 기술 역량, 제품 서비스 특성 등을 고려하여 적절한 요구사항과 검증항목을 선택하여 적용하고, 기업에서 제공 중인 서비스의 분야* 및 환경에 맞게 신뢰성 확보를 위한 참고자료로 활용하길 바란다. 더불어, 인공지능 신뢰성 확보를 위해서는 기술적 측면 외에도 윤리, 개인정보보호와 같은 법·제도적 측면도 함께 요구된다. 그러므로 본 안내서를 활용하기에 앞서 인공지능 윤리적 고려사항 점검을 위한 〈인공지능 윤리기준 실천을 위한 자율점검표〉와 개인정보보호의 준수 여부 점검을 위한 〈인공지능(AI) 개인정보보호 자율점검표〉를 선행적으로 검토할 것을 권고한다. 본 권고에는 인공지능 학습용 데이터나 오픈소스와 관련된 저작권 등, 법적 측면에 대한 검토도 포함된다. 또한, 인공지능에 해당하는 속성뿐만 아니라 기존 소프트웨어 시스템에 적용되는 전통적 속성도 적용되었는지 확인이 필요하다. 따라서, 안내서에 기술된 내용 외에도 시스템 성능, 보안 등 품질 관점의 검증 절차도 반드시 병행되어야 할 것이다.

안내서는 다음과 같은 절차로 활용할 수 있다.

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

② **요구사항 선정:** '①'의 분석 내용을 토대로 개발 안내서 요구사항과 세부 요구사항 본문을 참고하여 인공지능 서비스에서 신뢰성 확보를 위해 필요한 요구사항을 선정한다. 만약 인공지능 서비스가 공공 목적으로 활용되거나, 질병 진단 또는 주식 거래 등 사람의 신체·재산에 되돌리기 힘든 피해를 줄 가능성이 있다고 여겨진다면 가능한 모든 요구사항을 선정할 것을 권장한다. 반대로 특정 개인이나 집단에 차별이나 피해를 줄 가능성이 적은 서비스라면 모든 요구사항을 선정하지는 않더라도, 신뢰성 점검을 위한 참고자료의 성격으로 활용할 수 있을 것이다. 이 과정에서 요구사항별 활용 권장 대상(대표 행위자 및 협력 대상)이 협의하여 불필요하다고 판단된 요구사항의 경우 '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 못할 경우 이를 해결할 만한 수단 또는 기술이 있는지 확인해 볼 것을 권고한다. 각 요구사항의 대표 행위자가 주도하여 협력 대상과 함께 검증항목의 충족 여부를 판단하는 데 필요한 절차서, 코드, 분석 자료 등의 관련 산출물을 확인하고, 테스트나 측정이 필요한 항목은 해당 활동을 수행한다. 검증항목에 따라 충족 여부를 정성적으로 평가할 수 있으나, 이는 '①'에서 분석한 서비스 영향 정도를 고려하여 대표 행위자와 협력 대상자가 협의하여 충족 여부를 판단할 수 있다.

* 자율주행, 의료, 공공·사회 분야는 2023년에 공개될 분야별 개발 안내서를 통해 각 분야에 특화된 내용을 확인할 수 있으며, 2024년에는 채용, 스마트 치안, 생성 AI 기반 서비스 분야에 특화된 개발 안내서를 공개할 예정이다.

PART 2

요구사항 및 검증항목

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

목차

목차

목차

목차

01 생명주기 관리

책임성 투명성

요구사항 01 인공지능 시스템에 대한 위험관리 계획 및 수행

- 인공지능 시스템이 구현 및 운영되는 과정에서 발생 가능한 모델 오인식, 기능 오동작, 보안 및 개인정보 이슈 등의 위험 요소를 사전에 인식하고, 위험의 크기(심각성 및 파급효과)를 분석하여 대응 방안을 마련한다.

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

Yes ☐ No ☐ N/A ☐

- 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 management에는 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 24028:2020 – Overview of trustworthiness in artificial intelligence와 ISO/IEC 23894:2023 – Guidance on risk management에서는 인공지능의 신뢰성 관점에서 살펴보아야 할 위험 요소의 분류가 제공되어 있다.
- 위험 요소별로 위험이 발생할 수 있는 원인, 상황 및 조건을 분석한 다음, 위험 요소가 인공지능 시스템 또는 인간 및 주변 환경에 얼마나 큰 영향을 미치는지 분석하여야 한다. 만약, 식별된 위험이 극단적으로 부정적인 결과를 초래할 수 있다고 판단된 경우, 인공지능 기술 적용에 대해 재검토하여야 한다.
- 인공지능 시스템은 특성상 지속해서 그 형태와 형상이 변화할 수 있다. 이는 시스템에서 새로운 위험 요소가 지속해서 발생할 수 있음을 의미한다. 따라서, 생명주기 전반에 걸쳐 위험 요소 분석과 이에 대한 대응이 반복적으로 이루어져야만 적절하고 효과적인 위험관리가 이루어질 수 있다.

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

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템의 위험 요소는 소프트웨어 및 하드웨어 기반 시스템에서 발생할 수 있는 요소와는 다르다. 소프트웨어의 결함 및 오류, 하드웨어의 노후화 및 마모 등과 달리 데이터 기반 분석의 특성으로 나타날 수 있는 편향, 설명 미제공, 모델에 대한 공격 등의 위험 요소를 도출해야 한다. 이러한 요소는 아래 참고와 같이 ISO/IEC 23894:2023과 ISO/IEC 24028:2020에 제시되어 있다.
- 도출된 위험 요소별로 이를 야기할 수 있는 원인과 이에 따라 발생 가능한 결과를 분석해야 한다. 발생 가능한 결과란 사회적으로 부정적인 영향을 미칠 수 있는 현상 및 사고를 의미하며, 인체에 위해를 가하는 사고를 야기할 수 있는 차별적인 현상 등이 이에 해당한다.
- 위험 요소의 발생으로 인한 결과는 심각도와 발생빈도 등의 척도를 기준으로 위험의 크기 또는 수준을 평가할 수 있다. 이는 위험 요소의 파급효과를 의미한다. 위험 요소를 평가해 파급효과가 큰 위험 요소를 최우선으로 대응 방안을 마련해야 한다.
- 다만, 앞서 언급한 파급효과를 산정 및 평가하는 과정에서 심각도와 발생빈도뿐만 아니라, 상황에 맞는 척도를 도입하여 조합할 수 있다.

참고 인공지능 시스템의 위험 요소

- ISO/IEC 23894:2023에 따르면, 인공지능 시스템의 위험을 식별할 때 시스템의 특성과 그 응용 맥락에 따라 다양한 위험 요소들을 고려해야 한다. 위험 요소는 Annex B에서 다루고 있으며, 목록은 아래와 같다.
  - Complexity of environment
  - Lack of transparency and explainability
  - Level of automation
  - Risk sources related to machine learning
  - System hardware issues
  - System life cycle issues
  - Technology readiness
- 또한, ISO/IEC 24028:2020에서는 인공지능 시스템 구현 관점에서 고려할 수 있는 취약점 및 위험 요소를 정리해 놓았으므로(Chapter 8. Vulnerabilities, threats and challenges) 이를 참고할 수 있다.
  - AI specific security threats
  - AI specific privacy threats
  - Bias
  - Unpredictability
  - Opaqueness
  - Challenges related to the specification of AI systems
  - Challenges related to the implementation of AI systems
  - Challenges related to the use of AI systems
  - System and Hardware faults

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

Yes ☐ No ☐ N/A ☐

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

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

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

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

Yes ☐ No ☐ N/A ☐

- 01-1 에서 분석된 위험 요소별로 대응 방안을 마련하여야 한다. 위험 요소의 원인을 제거함으로써 인명 피해 및 사고를 미연에 방지하거나, 사고로 인한 파급효과 및 부정적 영향을 최소화하기 위한 수단이 이에 해당한다.
- 대응 방안이란, 구현 및 운영 방식 등의 절차, 소프트웨어 및 하드웨어 기능, 모델 학습 기법 및 전략 등 기술적으로 적용할 수 있는 모든 방법을 의미한다. 이에 대해 01-2a 의 참고와 같이 ISO/IEC 24028:2020에서 대응 방안의 분류를 제공한다. 인공지능을 구현하는 모든 이해관계자는 이를 고려하여 위험 요소에 대한 대응 방안을 마련하고, 위험이 제거 및 완화되었는지 확인하여야 한다.

01-2a 위험 요소별 완화 또는 제거 방안을 마련하였는가?

Yes ☐ No ☐ N/A ☐

- 위험 대응을 통해 위험의 부정적인 결과를 수용 가능한 수준으로 줄이고 긍정적인 결과를 달성할 가능성을 높이도록 해야 한다. 위험 요소별 다양한 대응 방안이 도출될 수 있으며, 각각의 방안에는 장단점이 있을 수 있으므로 신중하게 대응 방안을 선택하고 효과적으로 실행하여야 한다.
- 대응 방안의 한 가지 예로, 편향 완화를 위해 출처 및 데이터 소스를 분석하여 위험을 파악하고 데이터 수집 또는 라벨링 프로세스를 검토하는 것이 될 수 있다. 추가 예시는 다음 참고와 같이 ISO/IEC 24028:2020에 제시되어 있다.

참고 인공지능 시스템의 위험 대응 방안

- ISO/IEC 24028:2020에서는 인공지능 시스템 구현 관점에서 고려할 수 있는 취약점 및 위험 요소를 정리하였으며(Chapter 8. Vulnerabilities, threats and challenges), 이에 대한 대응 방안(Chapter 9. Mitigation measures)이 개괄적으로 제시되어 있으므로 이를 참고할 수 있다.

Chapter 9. Mitigation measures
9.1 General
9.2 Transparency
9.3 Explainability
9.4 Controllability
9.5 Strategies for reducing bias
9.6 Privacy
9.7 Reliability, resilience and robustness
9.8 Mitigating system hardware faults
9.9 Functional safety
9.10 Testing and evaluation
9.11 Use and applicability

01-2b 위험 요소의 파급효과가 감소하였는지 확인하였는가?

Yes ☐ No ☐ N/A ☐

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

다양성 존중 | 책임성 | 안전성 | 투명성

요구사항 02 인공지능 거버넌스governance 체계 구성

- 인공지능 시스템은 윤리와 관련된 문제가 발생할 가능성을 잠재적으로 내포하고 있다. 이러한 인공지능 시스템의 사회적 영향과 결과를 예측하고 대비하는 조직을 구성하는 것은 인공지능 신뢰성을 확보하는 데 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

02-1 인공지능 거버넌스에 대한 지침 및 규정을 수립하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능과 관련된 조직에서는 인공지능 시스템 신뢰성 확보를 위한 거버넌스 체계를 구성할 필요가 있다. 인공지능 시스템은 학습이나 추론 과정에서 윤리 및 지식재산권IP, Intellectual Property 관련 문제, 보안 및 개인정보 이슈가 발생할 수 있기 때문이다. 이러한 위험 요소에 대비하기 위해 내부적으로 인공지능 거버넌스에 대한 지침 및 규정을 수립해야 한다.
- NIST의 AI RMFRisk Management Framework에서는 인공지능 시스템 생명주기에 따라 내부 규정, 절차, 과정 및 실제 행위가 투명하고 효율적으로 이루어져야 한다고 언급한다. 즉, 인공지능과 관련된 법, 규제 관련 요구사항이 이해·관리되어 문서화하고, 위험관리 절차와 산출물이 체계를 통해 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 지침 및 규정을 수립해야 한다.
  - ✓ 둘째, 인공지능 시스템 생명주기에 따른 조직의 역할과 책임을 명확하게 문서화해야 한다.

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

Yes ☐ No ☐ N/A ☐

- 윤리 원칙의 수립은 인공지능 거버넌스 체계에서 기본적으로 갖춰져야 할 단계로, 인공지능과 관련된 법, 규제 및 정책을 이해한 후 내부적으로 윤리적 측면에서 이행해야 할 규정을 정의해야 한다. 즉, 인공지능과 관련된 위험을 인식하고 대비하기 위해 기업 성격에 맞는 핵심 가치를 선정하고 이와 관련된 표준 및 지침을 채택하여 내부 규정을 제공해야 한다.
- 인공지능 시스템의 신뢰성 확보를 위해서 인공지능 거버넌스 및 조직 전체의 업무, 역할, 의무 및 책임이 명확해야 한다. 이와 관련한 지침을 마련해 조직 구성원에게 제공함으로써 자신의 역할과 책임을 인식할 수 있다.

참고 윤리 원칙을 수립한 국내 기업의 사례

- 국내 기업 네이버에서 신뢰할 수 있는 인공지능 제품 연구·개발을 위해 '인공지능 윤리 준칙'을 수립하였다. 이는 기업의 모든 구성원이 지켜야 할 올바른 행동과 가치 판단의 기준이 되는 원칙을 의미한다. 인간존중humanity, 공정성fairness, 설명가능성explainability, 안전성safety, 프라이버시보호privacy protection의 5대 핵심 가치를 기반으로 윤리 준칙을 마련하였다.

참고 인공지능 거버넌스 조직 구성원의 주요 역할과 책임 지침 예시

- Model Artificial Intelligence Governance Framework 2nd('20.1)*는 인공지능을 책임감 있게 배포하기 위해 실제로 활용할 수 있는 실용적인 윤리 원칙을 제공한다. 다음은 인공지능 거버넌스를 위한 조직 구성원의 역할 및 책임과 관련된 지침의 예시이다.
  - ✓ 기존 위험관리 프레임워크를 사용하고 위험 관리 조치를 적용한다.
    - 인공지능 배포의 위험을 평가하고 관리한다.
    - 인공지능 의사결정에 대한 인간의 적절한 개입 수준을 결정한다.
    - 인공지능 모델 학습 및 선정 과정을 관리한다.
  - ✓ 인공지능 모델의 유지관리, 모니터링 및 문서화를 검토한다.
  - ✓ 이해관계자와 상호작용 및 의사소통한다.
  - ✓ 인공지능 시스템을 다루는 구성원이 교육받도록 보장한다.
    - 인공지능 모델의 결정을 해석하고, 데이터의 편향을 감지 및 관리하도록 교육한다.
    - 인공지능을 사용할 때 최소한의 이점, 위험 및 한계를 알 수 있도록 교육한다.

* 싱가포르의 IMDAInfocomm Media Development Authority(정보통신미디어개발기관)와 PDPCPersonal Data Protection Commission(개인정보보호위원회)에서 발행한 문서로, 유럽위원회 및 OECD 등 선도적 국제 플랫폼을 통해 얻은 피드백을 통합한 문서이다. 이는 인공지능 윤리 및 거버넌스에 대한 원칙, 프레임워크 및 권장 사항 등을 제공한다.

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

Yes ☐ No ☐ N/A ☐

- 02-1 에서 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 권고에서는 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)를 포함하여 구성할 필요가 있다. 외부 전문가들은 내부 조직에서 발생할 수 있는 편향된 시각을 보완하고, 집단 사고groupthink 등의 문제를 극복하는 데 도움을 주기 때문이다.

참고 인공지능 거버넌스 체계를 수립한 사례

- 국외 기업 마이크로소프트에서 인공지능 윤리와 관련된 문제에 대비하기 위한 인공지능 거버넌스 체계를 구축하였다. 이는 인공지능 윤리 관련 최신 동향에 대한 주제별 전문지식을 제공하는 'AI 윤리위원회', 인공지능 거버넌스 체계를 전사적으로 지도하는 'ORAOffice of Responsible AI', 시스템과 도구를 통해서 윤리 원칙 실행을 지원하는 'RAISE'로 구성된다.
- 국내 기업 LG에서 인공지능 윤리에 대해 점검할 수 있는 관리체제를 신설한 사례가 있다. 이는 윤리 원칙을 수립한 후, 조직 구성원들을 대상으로 인공지능 윤리 원칙 교육을 진행하며, 인공지능 연구 개발 단계에서 발생 가능한 윤리 문제를 사전에 검증하는 역할을 맡는다. 더불어, 주요 인공지능 윤리 이슈들을 논의하는 협의체를 출범시킬 예정이라고 밝혔다.
- 국내 기업 카카오에서 인공지능 기술윤리를 점검하기 위한 '기술윤리 위원회'를 신설한 사례가 있다. 이는 인공지능 서비스의 윤리규정 준수 여부 및 위험성 점검, 그리고 알고리즘 투명성 강화 등의 업무를 수행한다. 더불어, 인공지능 기술윤리 관련 정책 수립을 담당하는 '인권과 기술윤리팀'을 신설하였고, 전 직원을 대상으로 인공지능 알고리즘 윤리 교육을 진행했다고 밝혔다.

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.

참고 NIST AI Risk Management Framework에서의 거버넌스 방침

- GOVERN 2.1 섹션에서는 조직 내에서 인공지능 위험관리와 관련된 역할과 책임에 대해 다룬다. 이는 조직 내 위험 인식 문화를 조성하여 조직이 위험을 효과적으로 관리할 수 있도록 한다. 조직은 인공지능과 관련된 다양한 직무에서 해야할 일과 책임을 명확히 하는 규칙을 마련해야 하며, 직무의 예시는 다음과 같다.
  - ✓ 이사회 또는 자문위원회 (Boards of directors or advisory committees)
  - ✓ 고위 경영진 (Senior management)
  - ✓ 인공지능 감사 기능 (AI audit functions)
  - ✓ 프로젝트 관리 (Project management)
  - ✓ 인공지능 설계 (AI design)
  - ✓ 인공지능 개발 (AI development)
  - ✓ 인간-인공지능 상호작용 (Human-AI interaction)
  - ✓ 인공지능 테스트 및 평가 (AI testing and evaluation)
  - ✓ 영향 평가 기능 (Impact assessment functions)
  - ✓ 감독 기능 (Oversight functions)

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

Yes No N/A
☐ ☐ ☐

- 인공지능 거버넌스 담당 조직은 자신이 맡은 역할과 책임에 대해 충분히 인식한 인력으로 구성해야 한다. 이들은 인공지능 생명주기에 걸친 모든 프로세스의 중심적인 역할로서, 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.

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

Yes No N/A
☐ ☐ ☐

- 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 – 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
☐ ☐ ☐

- 무분별한 인공지능 시스템 개발이 범람할 경우, 서비스 사용자에게 혼란을 가중할 뿐만 아니라 시스템 개발 및 유지보수에 불필요한 예산 사용을 초래한다.
- 신규 계획 중인 인공지능 시스템이 기존에 운영 중인 시스템과 활용 대상 및 역할 측면에서 유사한지 고려하고, 기존 시스템에 대한 벤치마크 및 사례 연구를 통한 개선이 가능한지 분석한 결과를 기반으로 시스템을 계획 및 설계해야 한다.

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

Yes No N/A
☐ ☐ ☐

- 신규 인공지능 시스템 구축을 준비할 때, 구축하고자 하는 목적과 유사한 기존의 시스템이 존재할 수 있으므로 구축 사례들을 조사하는 것이 좋다. 이는 기존 동일 목적의 시스템의 문제점을 파악하여 신규 시스템이 좀 더 편리하고, 효율적으로 활용될 수 있도록 개선 사항을 도출하기 위함이다.
- 개선 사항 도출 과정에는 기존의 유사한 인공지능 시스템에 대한 벤치마크 및 사례 연구를 활용할 수 있다. 또한, 주요 이해관계자들의 의견 교류가 필수적으로 수반되어야 하며, 객관적인 기준, 근거, 검증을 기반으로 개선 사항을 도출해야 한다.

안전성 투명성

요구사항 03 인공지능 시스템의 신뢰성 테스트 계획 수립

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

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

03-1 인공지능 시스템의 특성을 고려한 테스트 환경을 설계하였는가?

Yes No N/A
☐ ☐ ☐

- 인공지능 시스템은 그 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 아래와 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 29119-11:2020 – Guidelines on the testing of AI-based systems에서는 아래와 같은 시뮬레이터의 예시를 제공한다.
  - ✓ 게임 엔진 기반의 이동형 로봇 시뮬레이터: MORSEModular OpenRobots Simulation Engine 프로젝트
  - ✓ 홈 서비스 로봇 학습 시뮬레이터: Facebook의 AI Habitat
  - ✓ 자율주행차 테스트용 시뮬레이터: NVIDIA의 DRIVE Constellation
- 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, 자율주행차의 보행자 회피 테스트는 높은 수준의 이미지 대표성이 요구된다.

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

Yes ☐ No ☐ N/A ☐

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

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

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

Yes ☐ No ☐ N/A ☐

- 테스트 오라클 문제의 극복이 필요한 인공지능 시스템이라면, 시스템의 기대 출력을 결정하기 위해 해당 도메인의 내·외부 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 이상 동의 시 승인하는 등의 방법이 있다.

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

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템 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 인공지능 시스템의 추적가능성 및 변경이력 확보

- 인공지능 시스템 운영 단계에서 문제 원인 추적을 위한 시스템 로그, 데이터 모니터링, 인공지능 모델과 사람 간의 의사결정 기여도 추적, 변경이력 관리 등의 방안을 확보한다.

04-1 인공지능 시스템의 의사결정에 대한 추적 방안을 수립하였는가?

Yes No N/A
☐ ☐ ☐

- 인공지능 시스템의 의사결정은 인공지능 모델이 자체 결정하거나 시스템 운영자 또는 사용자가 개입해 내릴 수 있다. 또한, 운영 중에도 학습이 이루어지도록 설계·개발된 인공지능 시스템이라면 학습 데이터와 모델에 대해 지속적인 모니터링이 필요하다.
- 인공지능 시스템의 경우, 전통적인 소프트웨어와 다르게 생명주기의 프로세스가 반복되는 특성이 있어 서비스 운영 단계에서도 전체 생명주기를 고려한 추적 방안을 확보해야 한다.
- 인공지능 모델의 구축, 데이터셋, 시스템 자체 등 기능적 측면과 인공지능 시스템 운영자 및 사용자 등 인적 요인으로 인해 발생 가능한 인공지능 시스템 출력 결과의 영향을 추적하기 위해서 시스템 단계별로 로그 수집 대상 정보를 정의하고 모니터링을 지속해야 한다.
  - ✓ 생성 AI 기반 서비스에서는 출력 결과(생성된 콘텐츠)에 워터마크를 추가하는 방식을 통해 콘텐츠의 자산을 보호하고 진본성을 유지한다.

04-1a 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: 시스템 운영자, 사용자) 개입 여부 등의 정보가 추적되어야 한다.
- 따라서 인공지능 모델이 전적으로 의사결정을 내리는 경우와 모델 결과를 사람이 검토하여 의사결정을 내리는 경우, 주로 사람이 의사결정을 내리지만 특정 이벤트와 같이 보조적으로 모델의 추론 결과가 활용되는 경우 등 시스템 결정에 대한 세부화된 기여도 기준을 내부적으로 확립하고, 시스템 운용 과정에서 이를 추적할 수 있는 방안(예: 로그 수집)을 확보해야 한다.

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

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템의 전 생명주기를 고려한 추적가능성 확보를 위해서는 모델의 학습 과정, 운용 시 의사결정 결과, 사용자 입력 데이터 등의 정보에 대한 지속적인 수집이 필요하다. 이를 위해 시스템 프로세스별 로그를 수집할 정보를 선정하고, 정보 간의 중요도를 정의한 뒤 로그 레코드 형식을 결정하여 로그를 수집해야 한다.
- 특히 인공지능 시스템 운영 과정에서의 오류 원인 추적을 위해서는 모델 구축 방법과 데이터셋 측면을 포함한 오류 원인의 분석이 필요하므로, 두 가지 측면을 고려하여 로그를 수집하여야 한다.

인공지능 시스템 운영 과정에서 발생 가능한 오류 원인 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 품질 저하

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

Yes ☐ No ☐ N/A ☐

- 서비스 이용 로그 분석은 서비스 운영 상태에 관한 확인뿐만 아니라, 사용자가 겪는 문제가 무엇인지 확인할 수 있는 가장 기본적인 방법이 될 수 있다. 서비스 로그는 서비스가 운영되는 동안 지속해서 수집되며 서비스 고도화에 따라 다양한 형태로 누적될 수 있다.
- 서버 인프라에 대한 로그를 통해 서비스 운영 상태에 대한 모니터링을 수행할 수 있으며, 사용자 상호작용 로그는 사용자가 어떤 서비스를 많이 이용하고 어떤 서비스에서 오류를 겪는지 분석할 수 있다. 이를 위해 인프라 관점에서는 로그 분석 소프트웨어를 활용할 수 있으며, 사용자 관점에서는 기업이 자체적으로 인터페이스 또는 상호작용의 호출에 따른 로그를 수집하거나 로그 분석 도구를 활용할 수 있다.

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

Yes ☐ No ☐ N/A ☐

- 인공지능 모델은 사용한 데이터에 따라 학습 모델도 함께 달라진다. 이로 인해 모델의 설계나 주요 파라미터들의 변경이 함께 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.

04-2a 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.
  - ✓ 데이터 흐름: 데이터가 시스템의 여러 구성 요소를 통과하면서 수집, 처리, 변환되는 방식에 중점을 둔다. 데이터 흐름을 이해하면 시스템의 데이터 처리 파이프라인을 최적화하고 잠재적인 병목 현상을 파악하며 데이터 무결성과 정확성을 보장할 수 있다.
  - ✓ 데이터 계보: 시스템 생명주기 동안 데이터의 출처, 변환 및 이동을 추적하고 문서화하는 것과 관련이 있다. 이는 데이터 출처, 추적성 및 규정 준수를 보장하는 데 매우 중요하며, 특히 데이터 감사, 디버깅 및 규정 준수에 유용하다.
- 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?

04-2b 데이터 소스 변경에 대한 모니터링 방안을 확보하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 모델의 학습 데이터 확보를 위해 웹 크롤링web crawling 등의 방법을 활용할 수 있다. 웹 크롤링은 관련 오픈소스(예: Apache Nutch, Scrapy)를 통해 대량의 데이터를 빠르게 확보할 수 있는 장점이 있으나, 크롤링의 대상인 웹 페이지의 데이터 소스가 실시간으로 변경되거나 대상 페이지 자체의 접속이 불가능한 장애가 있을 경우 특정 클래스의 데이터 부족 등 수집 데이터의 분포가 깨질 수 있다.
- 특히 지속해서 크롤링된 데이터를 실시간으로 학습하는 인공지능 시스템의 데이터 소스의 변경은 성능에 직접적인 영향을 줄 수 있다. 따라서 데이터 수집 과정을 모니터링해 데이터 소스 이상이나 중복 수집 등의 문제에 대응해야 한다.

04-2c 데이터 변경 시, 버전관리를 수행하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 모델 개발 과정에서 학습 데이터의 업데이트, 오류로 인한 라벨링 재수행 등 데이터 변경이 이루어지면 학습 결과인 모델도 변경된다. 또한 이전에 학습에 사용한 데이터셋과 특성이 완전히 다르거나 데이터셋 전체를 교체할 경우 성능이 크게 저하될 수 있으며, 이 경우에는 추가 학습이 필요할 수 있다.
- 따라서 학습 데이터의 변경이 수행될 경우, 단순히 사용된 학습 데이터의 버전뿐만 아니라 해당 버전으로 학습한 인공지능 모델을 함께 관리하여야 한다. 특히, 신규 데이터의 추가로 인한 학습 데이터 변경이 필요한 경우, 학습 혹은 테스트에 사용된 신규 데이터 비율을 기록하고, 그에 따른 모델의 성능 변화가 함께 추적 가능하여야 한다.
- 이를 위해 기계학습 프로젝트를 위한 오픈소스 기반의 데이터 버전관리 도구(예: DVCData Version Control)의 도입을 고려하거나, 학습 데이터 버전관리 시스템을 자체적으로 구축하여 학습 데이터의 버전과 모델의 버전관리를 수행해야 한다.

04-2d 데이터 변경 시, 이해관계자를 위한 정보를 제공하는가?

Yes No N/A
☐ ☐ ☐

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

데이터 변경 시 이해관계자에게 제공해야 할 정보 예시

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

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

Yes No N/A
☐ ☐ ☐

- 신규 데이터를 확보한 뒤, 인공지능 시스템에 사용하기 위해서는 기존 운영 중인 인공지능 모델과의 성능 비교가 필요하다. 사람이 판단하기에 신규 데이터가 기존 학습 데이터와 유사하여도 학습된 인공지능 모델이 기존 학습 데이터에서 학습한 데이터 특성과 다를 수 있다.
- 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 또는 필요한 경우 모델 재설계, 확장, 재학습 등 결정)

02 데이터 수집 및 처리

책임성 투명성

요구사항 05 데이터의 활용을 위한 상세 정보 제공

- 인공지능 학습용 데이터셋은 개발 과정에서 데이터가 추가로 수집될 수 있으며, 다른 유사 시스템의 학습 데이터로 사용될 수도 있다. 이때, 데이터 수집 출처, 특징 등 수집된 데이터의 정보의 제공이 미흡하다면 재사용성이 떨어지거나 데이터로 인해 야기된 문제에 대한 원인 파악이 어려울 수 있다. 따라서 수집 데이터의 올바른 활용과 문제 발생 시 명확한 원인 추적을 위해 데이터에 대한 상세 정보를 제공한다.

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

Yes ☐ No ☐ N/A ☐

- 데이터를 설명하기 위한 데이터로서 메타데이터metadata를 정의할 수 있으며, 메타데이터에 원시 데이터raw data의 특징들을 기록하여 향후 데이터를 재활용하는 상황이나 동일한 형식의 추가 데이터 수집이 필요할 때 데이터에 대한 정보를 전달할 수 있다.
- 개발자뿐만 아니라 인공지능 시스템과 관련된 이해관계자들이 수집 데이터를 이해하고 활용할 수 있도록 메타데이터, 상세 매뉴얼 등의 데이터에 대한 정보가 확보되어야 한다.
- 이해관계자들에게 전달되어야 할 정보의 예로는 수집 데이터의 출처와 형식, 데이터 수집·정제·가공 방법, 데이터 라이선스, 편향 유발 가능성 있는 보호변수protective attribute 등이 있다.

05-1a 정제 전과 후의 데이터 특성을 설명하였는가?

Yes No N/A
☐ ☐ ☐

- 데이터 정제작업은 라벨링 작업 전 학습 데이터 구축을 위한 데이터의 선별 및 처리 단계로서, 정제 과정을 거친 데이터만을 사용하는 사용자는 원시 데이터의 특성을 정확하게 파악할 수 없다. 따라서 향후 추가 데이터의 수집 가능성을 고려하여 정제를 위한 관련 정보와 정제 전과 후의 데이터 특성이 설명되어야 한다.
- 데이터 정제는 기본적으로 오픈소스 도구 등을 활용하여 정해진 규칙에 따라 데이터 일부를 제외 또는 변환하거나, 육안 검수 등의 방법으로 수행할 수 있으며, 정제된 데이터를 시각화하여 데이터 특성을 분석할 수 있다.
- 만일 원시 데이터를 직접 수집한 경우, 데이터 구축 목적, 데이터 종류, 도메인 특성 등 정제를 위한 기준 및 정제 도구 정보의 제시가 필요하다. 다음은 데이터 종류별 데이터 정제 기준의 예시이다.
  - ✓ 이미지 데이터: 이미지 크기, 비율, 화질, 촬영 장비, 개인정보처리, 저작권 등
  - ✓ 텍스트 데이터: 텍스트 분량, 텍스트 문법 정확성, 텍스트 내용 적절성, 주제와의 연관성 등
  - ✓ 음성 데이터: 음량, 발음 정확성, 소음 및 잡음, 안들림(허용범위 기준), 개인정보, 저작권 등

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

Yes No N/A
☐ ☐ ☐

- 인공지능 학습 데이터셋을 활용하기 위해서는 데이터셋에 대한 정보를 파악해야 하는데, 이러한 정보를 메타데이터라고 한다. 메타데이터는 JSON, XML 등의 형식으로 제공할 수 있으며, 데이터셋 종류에 따라 다음과 같은 정보가 포함될 수 있다.
  - ✓ 이미지 메타데이터: 촬영일시, 촬영위치, 노출도 등
  - ✓ 텍스트 메타데이터: 제목, 텍스트 길이, 생성일 등
  - ✓ 음성 메타데이터: 녹음일시, 길이, 녹음자, 화자, 화자 수 등
- 위와 같이 메타데이터와 학습 데이터는 구분되어야 하며, 각각에 대한 명세자료를 작성하여 개발자 관점에서 인공지능 모델 학습 등에 활용이 용이하도록 해야 한다.

참고 음성 데이터의 학습 데이터와 메타데이터 명세서

• 음성 데이터의 학습 데이터 명세서 예시

데이터 명		한국어 대화 음성 AI 데이터
데이터 포맷		음원: ***, PCM, 전사: ***.txt, 메타정보: json, xml
데이터 요약		한국인의 일상 대화를 인식하고 음성을 문자로 실시간 변환하는 AI기술 개발을 위한 대화음성 데이터 셋 구축 - 특정 상황(대화주제), 말씨나 말투 등 환경에 국한되지 않고 대화·음성 추출, 발화자 연령·특성·단위 분류를 통해 활용성 확보
데이터 출처		크라우드소싱 업체, 춘천 MBC, EBS
데이터 이력	배포버전	koreaspeech0000000.txt v1.0
	개정이력	신규
	작성자/배포자	수행기관(000)
데이터 통계	데이터 구축 규모	총 4,000시간 1TB 이내 - 연령별(1,000H), 지역별(1,000H), 방송콘텐츠(2,000H)
	데이터 분포	연령별 : 1그룹(10대~20대)(33.3%), 2그룹(30대~40대)(33.3%), 3그룹(50대~70대)(33.3%) 지역별 : 수도권(30%), 강원도(10%), 충청도(10%), 경상도(20%), 전라도(20%), 제주도(10%)
기타 정보	대표성	수도권/강원/충청/전라/경상/제주 6개 지역
	독립성	원시데이터는 라벨링데이터와 별개로 데이터셋으로 제공하므로 독립성 유지
	유의사항	인공지능 학습데이터를 활용한 다양한 알고리즘 도출
	관련 연구	-

• 음성 데이터의 메타데이터 명세서 예시


```
"verify-text-classification":"1",
    "verify-text-classification-metadata":
    {
        "class-name": "bad",
        "confidence": 0.93,
        "type": "groundtruth/label-verification",
        "job-name": "verify-text-classification",
        "human-annotated": "yes",
        "creation-date": "2018-11-20T22:18:13.527256",
        "worker-feedback": [
            {"comment": "The class of the sentence can't fit with its meaning."}
        ]
    }
```


https://aihub.or.kr/file/down.do?fileSn=8073&cnstcPrcuseFileSn=8073&dataSetSn=130

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

Yes ☐ No ☐ N/A ☐

- 대규모의 데이터셋을 이용하는 인공지능 모델의 학습 과정에는 데이터셋 자체의 편향이나 잠재된 편향 등 다양한 편향을 함께 학습할 수 있다. 이런 경우 인공지능 모델의 성능 저하뿐만 아니라, 성차별이나 인종 차별 등의 윤리적 문제로 인해 인공지능 시스템의 서비스화가 어려울 수 있다.
- 데이터 편향은 데이터 내 변수들을 분석하여 편향된 결과를 유발하는 데 많은 영향을 끼치는 특정 변수를 찾아내고, 이러한 변수들을 보호변수로 지정한 뒤 모델 학습에 반영되지 않게 하여 완화할 수 있다.
  - ✓ 데이터 편향을 확인하기 위한 대표적인 오픈소스 분석 도구는 Google What-If Tool, IBM Fairness 360 등이 있다.
- 따라서, 수집·구축된 데이터의 향후 사용자를 고려하여 개발하는 인공지능 시스템의 목적과 데이터셋의 보호변수 선정 이유, 과정 및 반영 여부에 대한 설명이 제공되어야 한다.

05-1d 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. 일반적인 라벨링 작업 절차는 아래와 같으며, 작업 절차에 따라 작업자를 대상으로 한 교육과 가이드 문서가 확보되어야 한다.
  - ✓ 데이터 획득 및 정제: 원시 데이터 획득 및 데이터 정제작업을 진행한다.
  - ✓ 라벨링 작업 대상 및 범위 정리: 원시 데이터 내의 어떤 항목들을 라벨링 하는지 대상 및 범위를 정의한다. 특히, 데이터 종류에 따라 세부적인 기준을 마련해야 한다(데이터 일부 라벨링, 개인정보 비식별화, 클래스 정의 및 관리 등).
  - ✓ 라벨링 방법 및 절차 수립: 라벨링 할 정보에 따라 자동·반자동·수동 등의 작업 방식을 결정하고, 작업의 배분 및 데이터별 라벨링 기준 등 상세한 작업 기준을 마련한다.
  - ✓ 라벨링 작업 진행: 상세 작업 기준으로 작업자 교육 후, 데이터 라벨링 작업을 실시한다(앞서 결정한 작업 방식에 따라, 자동·반자동일 경우, 적절한 라벨링 도구 선정 및 교육 진행).

05-2 데이터의 출처는 기록 및 관리되고 있는가?

Yes No N/A
□ □ □

- 학습 데이터의 품질은 인공지능 모델 성능에 큰 영향을 미치는 중요한 요인 중 하나이므로 데이터를 수집하거나 생성하는 과정에서 품질 확보를 위해 노력해야 하며, 경우에 따라서는 오픈소스 데이터셋을 활용할 수도 있다.
  - ✓ 특히, 생성 AI 모델은 오픈소스 데이터셋, 웹 크롤링 등을 통해 대량의 데이터를 학습에 활용한다. 이때, 학습 데이터의 품질이나 내용이 신뢰할 수 없을 경우, 편향 또는 환각hallucination의 결과로 이어질 수 있어 신뢰할 수 있는 출처로부터 제공되는 데이터셋을 사용하는 등 출처 관리가 필요하다.
- 오픈소스 데이터셋의 경우 다수의 사용자가 데이터 활용 과정에서 발견한 오류가 추후 발견될 수 있으며, 이로 인한 데이터셋 수정, 재구축으로 데이터 버전이 변경될 수 있다.
- 이러한 데이터셋 자체 원인으로 발생할 수 있는 인공지능 모델의 문제 대응을 위해서는 학습에 사용한 데이터의 명확한 출처, 구축 시점, 오픈소스 데이터셋 버전 등의 정보를 관리해야 한다.

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

Yes No N/A
☐ ☐ ☐

- 학습 데이터를 직접 생산한다면, 데이터 획득 시 수집 출처(예: 크라우드워커, 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 사용하는 경우에는 해당 데이터셋의 품질이 신뢰할만한 수준인지 고려할 필요가 있다. 고려사항으로는 데이터가 법적으로 문제는 없는지, 데이터셋 규모가 학습하기에 충분한지, 데이터셋에 대한 논의나 업데이트가 활발하게 이루어지는지 등을 고려해야 한다.
- WEF는 데이터 품질 보장을 위해 데이터셋을 학습시키기 전에 신뢰할 수 있는 데이터인지 미리 확인할 것을 권고한다.

참고 지도학습을 위한 데이터 품질 관리 요구사항 – 출처의 신뢰성 확보

TTA 정보통신단체표준 TTAK.KO-10.1339:2021 – 지도학습을 위한 데이터 품질 관리 요구사항에서는 지도학습 계열의 인공지능 기술에 활용되는 데이터 획득 시 출처의 신뢰성 확보 측면에서 고려해야 할 내용을 정리하였다.

- 데이터 획득 시 직접 생산 혹은 제3자에 의해 생산된 데이터의 중계의 2가지 방법으로 데이터를 획득할 수 있는데, 제3자에 의해 생산된 데이터를 중계하여 획득하는 경우, 데이터의 출처에 대하여 신뢰성을 확보하여야 하며, 다음과 같은 요소를 고려할 수 있다.
  - 제3자가 데이터 획득 시 개인정보보호, 지식재산권, 사전 승인/허가 등과 관련하여 정식으로 절차를 밟고 문제없이 획득하였는지 여부
  - 제공하는 데이터셋의 규모가 충분히 커, 데이터 사용자가 원하는 학습용 데이터를 제공하는 데에 문제가 없는지 여부
    예) 규모가 충분히 크지 않은 경우, 데이터 획득을 재차 시도하고자 할 때 수급에 문제가 있을 수도 있음
  - 해당 데이터가 지속적인 업데이트 및 추가 제공 등이 이루어지고 있는지 여부
  - 데이터와 함께 설계서의 내용이 명확히 제공되는지 여부
  - 해당 데이터의 활용건수 및 인용건수가 많아 범용성이 높은지 여부
- 반면, 데이터를 직접 생산(이미지/동영상 촬영, 발화 녹음, 텍스트 작성 등)하는 경우, 위의 내용 중 첫 번째 사항을 고려하여야 한다.

05-2b 오픈소스 데이터셋을 활용하는 경우, 출처를 명시하였는가?

Yes No N/A
☐ ☐ ☐

- 인공지능 모델 학습에 오픈소스 데이터셋을 사용한 경우, 학습 시점에는 발견되지 않았던 오류나 편향된 결과가 나올 수 있다. 또한, 편향된 결과는 사회 인식 변화에 따른 윤리적 문제와도 결부될 수 있어 오픈소스 데이터셋 구축 당시 인식하지 못한 데이터 편향의 발생 가능성이 있다.
- 따라서 오픈소스 데이터셋을 활용하여 학습기반 인공지능 모델을 구축할 경우, 과거·현재·미래 시점에 발생할 수 있는 데이터 편향의 원인 파악을 위해 확보된 데이터의 명확한 출처 및 관련 정보를 명시하여 관리해야 한다.

안전성

요구사항 06 데이터 견고성 확보를 위한 이상abnormal 데이터 점검

- 인공지능 모델의 학습에 활용되는 데이터는 이상값, 중독 및 회피 등에 영향을 받지 않아야 하며, 이의 점검 및 방어 기법의 적용을 통해 견고성을 확보한다.

06-1 이상 데이터의 식별 및 정상 여부를 점검하였는가?

Yes ☐ No ☐ N/A ☐

- 이상 데이터란 학습용 데이터를 구성하는 데이터셋의 수집 및 가공 과정에서 발생할 수 있는 다양한 오류error 와 일반적인 데이터의 범위에서 크게 벗어난 데이터 이상값outlier을 포괄한다. 학습 데이터의 수집 및 가공 과정에서 발생하는 이상 데이터는 데이터상의 노이즈, 학습 데이터 내의 편향, 잘못된 라벨링, 라벨링 누락 등 다양한 원인에 의해 생길 수 있으며 이를 해결하지 않으면 인공지능 모델의 성능 및 견고성 확보가 어렵다.
- 단, 이상 탐지anomaly detection 시스템에 활용되는 인공지능 모델의 경우, 이상 데이터는 제거해야 할 데이터가 아닌 학습 데이터가 될 수 있음에 유의하여야 한다.
- 비정형 데이터unstructured data를 학습에 활용하는 경우, 데이터 전처리 과정에서 이상 데이터의 식별을 위한 별도의 기법을 마련하여야 한다.

06-1a 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.
- 데이터 분포 시각화 방법은 데이터의 특성에 따라 다양한 기법이 존재한다. 먼저, 전체 데이터의 평균, 분산, 편차 등을 활용하여 데이터 분포를 시각화하는 분포 도표, 범주형 데이터를 시각화하는 범주형 도표, 2차원 행렬 데이터를 시각화하는 행렬 도표 등이 있다.

06-1b 학습 데이터 이상값 식별 기법을 적용하였는가?

Yes ☐ No ☐ N/A ☐

- 데이터 전처리 과정에서 중요한 활동 중 하나는 데이터 이상값을 식별하고 이를 제거하는 것이다. 데이터 누락과는 달리 데이터 이상값의 경우에는 데이터값이 이미 정해져 있지만, 전체 데이터셋을 기준으로 정상 범주를 벗어난 값이므로 단순 탐색만으로 발견하기 쉽지 않다.
- 데이터 이상값을 식별하는 방법에는 주로 데이터 전체에 대해 통계적 기법을 적용하여 전체 데이터셋을 고려하였을 때 차별화되는 데이터 포인트를 찾아내는 방법 등이 있으며, 이와 관련 대표적인 기법은 Z-점수, 사분위수 범위 등이다.

데이터 이상값 식별 기법 예시

이상값 식별 기법 분류	설명
Z-점수	• 가장 간단한 통계적 측정 방법으로, Z-점수는 주어진 데이터셋의 분포 평균과 표준편차를 이용하여 관찰된 데이터 포인트가 전체 데이터로부터 얼마나 멀리 떨어져 있는지를 수치화한다.
사분위수	• 데이터를 정렬한 후 4등분으로 나누면 등분점이 3개 생기는데, 앞에서부터 '제1사분위수(Q1)', '제2사분위수(Q2)', '제3사분위수(Q3)'라고 한다. 이때 데이터가 Q1과 Q3 사이에 속하지 않으면 이상값으로 판별한다.

06-2 데이터 공격에 대한 방어 수단을 강구하였는가?

Yes No N/A
☐ ☐ ☐

- 인공지능 서비스 운영 과정에서 입력 데이터에 최소한의 변조를 가해 예상과는 다른 결과를 출력하도록 하는 적대적 공격에 노출될 수 있으므로, 데이터 수집 및 처리 단계에서 대처 가능한 방안을 검토 및 적용하는 것이 바람직하다.
- 데이터 수집 및 처리 단계에서는 데이터 최적화data optimization를 통해 적대적 공격에 방어할 수 있다. 데이터 최적화는 모델의 안정성과 일반화 성능을 향상시키기 위해서도 활용되지만, 적대적 사례에 대한 효과적인 대응을 위해 활용되기도 한다. 데이터 최적화를 통한 방어 대책은 적대적 학습adversarial training, 데이터 품질 개선, 데이터 노이즈 제거를 중심으로 하여 모델이 적대적 사례에 강건하게 동작하도록 한다.

06-2a 데이터 최적화를 통한 방어 대책을 마련하였는가?

Yes No N/A
☐ ☐ ☐

- 데이터 수집 및 처리 단계에서는 데이터 최적화를 통해 인공지능 모델이 적대적 사례에 강건하게 대응할 수 있다. 대표적인 방어 대책은 적대적 학습이다. 적대적 학습은 적대적 사례로 활용될 수 있는 모든 경우의 수를 미리 고려하여 학습 데이터셋에 포함시키는 것이다. 이를 통해 모델이 미세한 변화에 대응하고 더 복잡한 환경에서도 안정적으로 작동할 수 있도록 한다. 적대적 학습을 위해서는 충분한 수와 다양성이 보장된 적대적 학습 데이터를 생성하는 과정이 필수적이다.
- 또한, 데이터 품질을 향상시키고 노이즈를 제거하는 과정도 중요하다. 이는 모델이 더 정확한 패턴을 학습하여 입력 데이터의 특이성을 감지하고 이를 효과적으로 처리함으로써 적대적 사례에 대한 민감성을 낮출 수 있다. 관련 방안은 06-1을 활용할 수 있다.

다양성 존중 | 책임성 | 투명성

요구사항 07 수집 및 가공된 학습 데이터의 편향 제거

- 학습에 필요한 데이터를 수집 및 가공 시 발생할 수 있는 편향을 인식하고 이를 제거하기 위한 방안을 적용한다. 주로, 데이터 수집 시 발생할 수 있는 편향을 확인해야 하며, 학습을 위한 특성을 선택하거나, 데이터 라벨링 및 샘플링 시에도 편향이 발생할 수 있으므로 제거 방안을 마련한다. 단, 이미 편향성 검토가 완료된 데이터를 활용하거나, 초거대 인공지능 모델처럼 현실적으로 모든 데이터를 검증하기 어려운 경우에는 샘플링 기법 등을 통해 데이터를 검증한다.

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

Yes ☐ No ☐ N/A ☐

- 인적 요인으로 인한 편향은 사람이 의식적 혹은 무의식적으로 특정 정보에 대해 편향되는 점에서 기인한다.
  - ✓ 인적 편향: 자동화 편향automation bias, 그룹 귀인 편향group attribution bias, 암묵적 편향implicit bias, 그룹 내 편향in-group bias 등이 포함됨
- 인적 편향을 방지하도록 데이터 수집 시 명확한 수집 및 검수 기준을 수립하여 수집하는 작업자별로 데이터 특성이 편향되지 않도록 방지하거나, 다양하고 충분한 수의 검수자를 확보함으로써 검수 시 편향을 바로잡아야 한다.
- 데이터는 수집 도구나 방법에 활용되는 물리적 요인으로 인해 데이터의 편향이 발생할 수 있다. 이미지의 촬영 도구나 저장 장치 등의 요인으로 인하여 이미지의 색상, 밝기, 해상도 등 물리적으로 한정된 데이터가 수집될 수 있다.
- 이에 따라 촬영 대상자의 연령대나 인종을 구분하기 힘들거나, 특정 방법으로 수집된 데이터만 학습이 이뤄지므로, 편향을 발생시키는 물리적 요인을 제거하거나 다양한 수집 장치를 활용하여 다양성을 보완하는 것이 바람직하다.

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

Yes No N/A
☐ ☐ ☐

- 데이터를 수집하는 과정에서의 인적 편향은 데이터 수집 작업자가 갖는 편향에서 비롯된다. 이 경우, 수집 작업자들의 개인별 편차를 줄이기 위해 데이터 수집 작업 가이드라인을 마련하고, 다양한 작업자를 모집하여 특정 배경과 성향을 배제하고, 수집 결과에 대한 검수자를 충분히 확보하여야 한다.
- 인적 편향을 완화하기 위해 다양한 출처와 인구통계학적 그룹에서 데이터를 수집하고, 다양한 데이터 증강 기술을 활용해 부족분을 보완할 수 있다. 또한 데이터 수집 과정을 지속적으로 모니터링하고 인적 감독 및 평가를 통합하면 잠재적인 편향을 식별하고 완화할 수 있다.

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

Yes No N/A
☐ ☐ ☐

- 특정 하드웨어 및 장비를 사용하여 데이터를 수집하는 경우, 수집 환경 및 제약 조건으로 인하여 많은 수의 일관된 데이터를 확보하기 어려울 수 있다. 이러한 경우 데이터의 다양성 확보에도 악영향을 미치기 때문에 다수의 장비 및 이기종 장치를 활용함으로써 데이터 수량 및 다양성 확보가 가능하다.
- 다만, 이러한 경우 수집 경로 및 환경(예: 카메라 촬영, 웹 크롤링)이 달라지기 때문에, 수집 후 데이터를 활용하려면 데이터의 일관성이 유지되어야 하므로 데이터 정제 및 검수가 충분히 이루어져야 한다.
- 데이터 수집 및 생성 시 장비의 사양 및 수집 환경 등 물리적 요인으로 인해 제한된 상황 및 시나리오에 대한 데이터만 수집되는 등의 편향이 발생할 수 있다. 따라서, 데이터 수집 시 이러한 요인을 점검하고 대처하는 계획을 마련해야 한다. 발생 가능한 편향의 예시는 다음과 같다.
  - ✓ 특정 브랜드나 모델의 카메라로만 이미지를 촬영하는 경우, 해당 카메라의 특성(예: 해상도, 노이즈 수준)이 데이터에 반영될 수 있음
  - ✓ 특정 종류의 센서(예: 가속도계, 자이로스코프)를 사용하여 데이터를 수집하는 경우, 해당 센서의 정확성과 측정 범위 등이 데이터에 영향을 미칠 수 있음
  - ✓ 특정 제조사의 의료 장비를 사용하여 환자 데이터를 수집하는 경우, 해당 장비의 측정 오차나 특성이 데이터에 영향을 미칠 수 있음

07-2 학습에 사용되는 특성feature을 분석하고 선정 기준을 마련하였는가?

Yes ☐ No ☐ N/A ☐

- 편향 완화를 위해서는 차별을 일으킬 수 있는 민감한 특성들을 사전에 파악하는 것이 중요하며, 이를 위해 데이터의 특성들을 분석하고, 해당 특성을 학습에 사용할 것인지 그 선정 기준을 수립하는 것이 바람직하다.
- 일부 민감한 특성들은 인공지능 의사결정의 차별을 일으킬 수 있으며, 국제기구나 글로벌 기업들은 아래 표와 같이 민감한 특성들을 언급하고 있다. 이와 같은 특성들은 데이터 학습 시 반영되지 않아야 하는 특성으로 선정하고, 이에 따라 발생할 수 있는 편향을 완화하여야 한다.

사회적 물의를 일으킬 수 있는 민감한 특성들

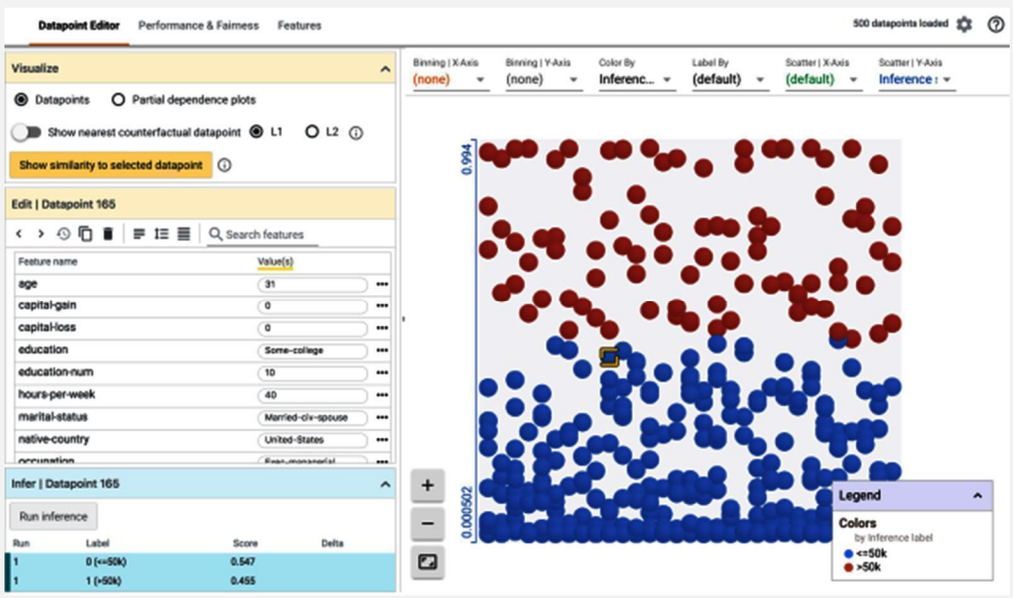
기관명	특성
UNESCO	• 나이, 성별, 인종, 민족·사회적 기원, 혈통, 언어, 종교, 정치적 사상, 국적, 출생 시 사회경제적 상황, 장애
ALTAI	• 나이, 성별, 인종, 민족·사회적 기원, 혈통, 언어, 종교, 정치적 사상, 소수 민족 구성원, 재산, 출생, 성적 지향
ISO/IEC 24027:2021	• 나이, 성별, 인종, 수입, 가족관계, 교육 수준, 키·체중, 장애 여부
IBM Watson OpenScale	• 나이, 성별, 인종, 결혼 여부, 주소
Google	• 인종, 성별, 장애 여부, 종교

07-2a 보호변수 선정 시 충분한 분석을 수행하였는가?

Yes ☐ No ☐ N/A ☐

- 보호변수 선정 시 충분한 분석을 진행하지 않을 경우, 모델의 성능이 저하될 수 있다. 따라서 모델 추론 결과에 영향을 미치는 특성을 식별한 경우, 주어진 데이터셋으로부터 데이터의 일부분을 변경하면서 모델의 결과가 어떻게 변하는지 관찰하고 분석하여야 한다.
- 기계학습 기반 회귀 및 분류 모델의 경우, 데이터 변화에 따른 추론 결과의 추이를 시각화하여 보여주는 도구(예: Google What If Tool)를 사용하여 설정한 보호변수가 인공지능 의사결정의 차별을 일으키는데 얼마나 영향을 미치는지, 성능이 어떻게 변하는지 알 수 있다.

참고 Google What-If Tool 화면 예시



https://pair-code.github.io/what-if-tool/

07-2b 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 고려해볼 수 있다. 필터filter 방법, 래퍼wrapper 방법, 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 활용하는 것이다.
- 편향과 관련된 특성을 제거하는 경우, 다른 종류의 편향을 발생시키거나 강화할 수 있어 모든 경우에 효과적인 방법은 아닐 수 있다. 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.

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

Yes ☐ No ☐ N/A ☐

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

과도한 특성 선택 및 배제를 방지하기 위한 점검표

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

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

Yes No N/A
☐ ☐ ☐

- 지도학습계열 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.
- 이는 라벨링 작업자의 전문성 부족, 작업 및 판단 기준의 일관성 결여 등이 원인이 될 수 있다. 라벨링 작업자가 발생시킬 수 있는 편향의 잠재적인 원인을 사전에 파악하고, 라벨링 결과의 평가 및 작업 기준의 교육 등을 통해 편향 발생을 방지해야 한다. 또한 다양한 라벨링 작업자를 섭외하여 작업자별로 나타날 수 있는 편향을 최소화하거나, 검수자를 충분히 확보하여 편향 방지 작업을 수행하는 것이 바람직하다.

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

Yes No N/A
☐ ☐ ☐

- 데이터 라벨링은 라벨링 도구의 이용 여부에 따라 자동·반자동·수동 등의 방식이 있다. 이때 라벨링 작업자가 라벨링 과정에 개입하게 되며, 이에 따라 작업자의 잠재적 편향이 라벨링에 반영될 수 있다.
- 이러한 잠재적 편향은 다수의 라벨링 작업을 위한 가이드라인이 명확하지 않아 개인의 판단에 의존하게 된다. 따라서 이를 파악하고 방지하기 위해서는 상세한 라벨링 가이드라인이 마련되어야 한다. 또한 가이드라인을 기반으로 작업자에게 충분한 교육을 실시하여 작업자 간 편향 발생 여지를 최소화해야 한다.

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

Yes No N/A
☐ ☐ ☐

- 데이터 라벨링 단계에서 인적 편향을 줄이려면 다수의 데이터 라벨링 작업자 확보가 우선적으로 요구된다. 또한, 라벨링 작업자들을 인구 통계학적 특성 및 배경지식 등이 다양하고 고르게 분포되도록 구성하는 것이 바람직하며, 주요 분포 고려 요소는 다음과 같다.

  ✓ 인종, 종교, 성별, 민족, 장애 여부, 언어, 국적, 경제적 상황 등

- 작업자의 다양성을 검증하기 위해서는 크게 2가지를 확인해야 한다. 첫째, 크라우드소싱crowdsourcing 등의 방법을 도입하였는지 점검한다. 둘째, 데이터 라벨링 작업자의 인구 통계적 특성, 배경지식 등을 조사하고 분석함으로써 실제로 라벨링 작업자가 다양하고 고르게 분포하는지를 확인한다.

  ✓ 크라우드소싱: 데이터 라벨링 과정에 라벨링 관련 교육을 받은 일반인이 참여토록 외부 발주하는 것을 의미하며, 이를 통해 기존 라벨링 작업자 집단보다 더욱 다양한 작업자를 확보할 수 있음

07-3c 다양한 데이터 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.
- 수동 라벨링에 필요한 시간과 노력을 최소화하여, 효율성을 높이기 위한 목적으로 자동 라벨링 시스템을 활용하기도 한다. 자동 라벨링은 여러 이점을 가지고 있지만, 동시에 편향이 잠재되어 있을 수 있다. 일례로, 자동 라벨링 시스템을 구축할 때 사용된 학습 데이터의 내재된 편향으로 인해 생성된 라벨에도 그 편향이 상속될 수 있다는 연구 결과도 존재한다. 따라서, 자동 라벨링 시스템을 활용하더라도 데이터 검수를 수행하여 편향을 완화하는 활동이 필요하다.

07-4 데이터의 편향 방지를 위한 샘플링을 수행하였는가?

Yes No N/A
☐ ☐ ☐

- 샘플링은 모집단에서 일정한 기준으로 데이터를 추출하여 표본을 만드는 기법이다. 일정한 기준으로 추출된 표본은 모집단의 분포를 대표하는 동시에 실제 모집단의 클래스 불균형으로 인한 편향 또한 방지하여야 한다.
- 모집단의 클래스 불균형에 따른 편향을 방지하기 위한 대표적인 기법으로 SMOTE(Synthetic Minority Oversampling TEchnique)를 예로 들 수 있다. 이는 임의의 소수 클래스 데이터와 유사한 새로운 합성데이터를 생성한 후, 기존 데이터에 추가하는 방식으로 편향을 방지한다.

07-4a 편향 방지를 위한 샘플링 기법을 적용하였는가?

Yes No N/A
☐ ☐ ☐

- 사회적 편견 및 차별을 야기할 수 있는 인구 통계학적인 데이터를 샘플링할 경우, 이로 인한 편향을 방지할 수 있는 샘플링 기법을 적용하고, 적용 과정에서 필요한 활동과 정보가 생성되었는지 확인해야 한다.

참고 샘플링 기법 예시 – SMOTE

- SMOTE는 실제 모집단 데이터 클래스의 불균형으로 인한 편향 문제를 해결하기 위해, 클래스의 개수가 적은 표본과 유사한 새로운 합성데이터를 생성하여 기존 데이터에 추가하는 기법이다.

SMOTE 단계

- SMOTE 기법 적용 시 데이터 증가로 인해 계산 시간 및 과적합 가능성 또한 증가하므로, 최종 데이터의 구성 및 모델의 추론 결과를 면밀히 확인해야 한다.
- 소수 클래스 구분 기준 및 합성 데이터 생성 비율 등의 세부적인 수치는 인공지능을 활용해 구현하고자 하는 서비스·기술, 다루고자 하는 데이터셋에 포함된 정보에 따라 달라질 수 있으며, 기법을 활용하는 담당자는 이에 대한 근거를 마련해야 한다.

03 인공지능 모델 개발

책임성 안전성

요구사항 08 오픈소스 라이브러리의 보안성 및 호환성 점검

- 인공지능 모델 개발 단계에서는 개발 기간을 단축하고 최신 기술 동향을 빠르고 유연하게 적용하기 위해 다양한 오픈소스를 활용할 수 있다. 오픈소스 라이브러리 도입 전에는 필요성 및 원하는 기능의 제공 여부 등의 확인이 필요하다. 오픈소스 활용을 결정하였다면 사용할 라이브러리가 안정적으로 업데이트 중인지, 주의해야 할 라이선스 기준은 무엇인지 등을 확인한다. 오픈소스를 사용 중인 경우, 사용하던 오픈소스가 어느 날 라이선스 정책이 바뀌거나 취약점이 새롭게 발견될 수도 있다. 따라서 사용 중인 오픈소스의 목록 및 버전을 지속해서 확인하여 운영 및 보안상의 위험 요소를 점검한다.

08-1 오픈소스 라이브러리의 안정성을 확인하였는가?

Yes ☐ No ☐ N/A ☐

- 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.

08-1a 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.
  - ✓ Redhat의 경우, 오픈소스 기반의 수익화 모델(호환성, 보안 강화, 기술지원 등 제공)을 개발하고 있으며, 오픈소스 라이브러리 업데이트 시 커뮤니티 내 구성원들이 제안한 개선 사항도 적용한다. 이처럼 수익화 모델 기반의 오픈소스 라이브러리 역시 개인 및 기업의 참여가 활성화된 프로젝트로 판단할 수 있다.

08-2 오픈소스 라이브러리의 위험 요소는 관리되고 있는가?

Yes No N/A
☐ ☐ ☐

- 오픈소스 라이브러리 또는 소프트웨어는 저작권자가 소스코드를 공개했을 뿐이며 지식재산권으로 보호받는 소프트웨어이다. 따라서, 저작권자가 제시한 라이선스(저작권) 준수 조건이 존재하며, 오픈소스 라이브러리마다 다양한 의무 사항이 있다. 이때, 라이선스 위반 및 저작권 침해로 법적 책임을 져야 할 위험이 있으므로 반드시 라이선스와 관련한 위험 요소를 분석하고 관리해야 한다.
- 오픈소스 라이브러리의 종류 및 버전 선택 시 개발 과정에서 사용된 오픈소스 라이브러리 또는 개발 환경 버전 변경에 따른 호환성을 고려하여야 하며, 이때 사용된 오픈소스 라이브러리에서 보안 취약점이 발견될 수 있으므로 이러한 이슈들을 확인하여 보안상의 위험 요소에 대한 관리도 필요하다.

08-2a 사용 중인 오픈소스 라이브러리의 라이선스 준수사항을 이행하였는가?

Yes ☐ No ☐ N/A ☐

- 오픈소스는 무료로 사용할 수 있지만, 라이선스별로 준수사항은 별도로 규정된다. 그러므로 오픈소스 라이브러리를 활용하여 인공지능 모델을 개발한다면, 사용할 오픈소스의 라이선스 종류 및 라이선스 고지문을 확인하고, 허용 또는 의무 사항을 우선해서 숙지해야 향후 발생할 수 있는 법률적 위험을 최소화 할 수 있다.
- 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 restrict other software)
  - ✓ 기술 중립적인 라이선스 제공 (License must be Technology-Neutral)

08-2b 사용 중인 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 의존성[dependency]을 파악하는 등 호환성을 고려해야 한다.
- 사용 중인 오픈소스 라이브러리에서 보안취약점이 발견되기도 한다. 보안 취약점에 따른 영향을 최소화하기 위해 보안취약점 및 버전 변경에 따른 릴리즈 노트[release note]를 지속해서 확인하여 신속히 탐지 및 대응해야 한다.

다양성 존중

요구사항 09 인공지능 모델의 편향 제거

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

  * 요구사항 07-2 에서 언급한 바와 같이 인종차별, 성차별 등 사회·윤리적으로 문제가 되는 경우에 한함

09-1 모델 편향을 제거하는 기법을 적용하였는가?

Yes No N/A
☐ ☐ ☐

- 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 기법pre-processing, 모델 학습 중에 적용할 기법in-processing, 모델 학습 이후 적용할 기법post-processing 이다. 구현하려는 인공지능 모델 및 목표 임무에 따라서 이 중 적절한 기법을 선택하여 적용하여야 한다.

09-1a 개발하려는 모델에 맞게 편향제거 기법을 선택하였는가?

Yes No N/A
☐ ☐ ☐

- 인공지능 모델의 편향을 완화하기 위한 기법은 적용 단계에 따라 3가지로 구분된다. 모델 학습 전, 학습 과정 중 그리고 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	기법구분			설명 및 지표
		Pre	In	Post	
알고리즘 편향 algorithmic bias	가중치 재지정	✓			학습 데이터셋 샘플에 가중치를 할당하는 방식
리콜 편향 recall bias	라벨링 재지정	✓			학습용 데이터 샘플의 라벨을 수정하는 방식
특성 편향 feature 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 편향 data sampling bias	샘플링	✓			학습 데이터 내 샘플링을 통해 편향을 제거하는 방식
과잉일반화 편향 overgeneralization bias	정규화	✓	✓		분류 시 편향에 많은 영향을 주는 클래스 분포를 대상으로 보정하는 방식
데이터 표본 편향 data sampling bias	제약 최적화		✓	✓	분류기의 손실 함수에 보정값을 부여하는 방식
평가 편향 evaluation 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

09-1b 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
패리티parity 기반 지표	• 인구통계학적statistical/demographic 형평성 지표, 차등적disparate 효과 지표
혼동 행렬confusion matrix 기반 지표	• 동등 기회equalized opportunity, Equalized Odds, 전체 정확도 형평성, 조건부 사용 정확도 형평성, 대응 형평성, 비보상 동등화
점수score 기반 지표	• 양성 및 음성 클래스 균형 지표
사후가정counterfactual 기반 지표	• 사후가정 공정성
개인individual 공정성 지표	• 일반화 엔트로피 지수, 셰일 지수

안전성

요구사항 10 인공지능 모델 공격에 대한 방어 대책 수립

- 인공지능 모델은 적대적 의도를 가진 사용자에 의해 인공지능이 잘못된 의사결정을 하도록 유도하는 공격의 대상이 될 수 있으므로 이를 방지 또는 완화하기 위한 대책을 수립한다.

10-1 모델 공격이 가능한 상황을 파악하였는가?

Yes No N/A
☐ ☐ ☐

- 적대적으로 생성된 입력과 같이 작은 변화에도 모델을 오동작하게 만드는 공격은 인공지능 시스템의 안전성을 위협할 수 있다. 따라서, 적대적 공격을 이해하고 적절한 대응 방안을 마련하여 인공지능 모델의 견고성을 향상시키는 것이 필요하다.
- 적대적 공격의 대표적 유형으로는 회피 공격evasion attack이 있다. 추론 중에 인공지능 모델을 속이기 위해 입력 데이터를 조작하는 것이다. 이러한 공격에 대응하는 방안을 수립하기 위해서는, 개발 중인 모델의 데이터 유형(예: 이미지, 텍스트, 오디오)별로 공격 가능한 적대적 사례를 파악하여야 한다.

10-1a 데이터 유형별 공격 가능한 적대적 사례를 확인하였는가?

Yes No N/A
☐ ☐ ☐

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

10-2 모델 공격에 대한 방어 수단을 강구하였는가?

Yes No N/A
☐ ☐ ☐

- 06-1 에서 언급한 것처럼, 인공지능 서비스 운영 과정에서 입력 데이터에 최소한의 변조를 가해 예상과는 다른 결과를 출력하도록 하는 적대적 공격에 노출될 수 있다. 따라서, 인공지능 모델 개발 단계에서 대처 가능한 방안을 검토 및 적용하는 것이 바람직하다.
- 10-1 을 통해 현재 개발 중인 모델의 공격 가능한 상황을 파악하였다면, 모델 최적화model optimization를 통해 적대적 공격에 방어할 수 있다. 모델 최적화는 주로 성능 향상, 자원 효율성 향상, 학습 시간 단축, 모델 해석성 개선 등의 차원에서 활용되지만, 적대적 사례에 대한 효과적인 대응을 위해 활용되기도 한다. 모델 최적화를 통한 방어 대책을 통해 모델이 적대적 사례에 강건하게 동작하도록 한다.

10-2a 모델 최적화를 통한 방어 대책을 마련하였는가?

Yes No N/A
☐ ☐ ☐

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

인공지능 모델 공격 방어를 위한 모델 최적화 방안

방어 대책	설명 및 기법 예시	적용 가능한 데이터 유형
Defensive Distillation	복잡한 신경망의 지식을 간단한 신경망으로 전이시키는 방법이다. 원본 모델의 확률 분포를 얻어 증류distillation 모델을 훈련하면, 증류 모델은 원본 모델의 특성을 보전하게 된다. 작업 수행 시 증류 모델을 활용하면 적대적 공격에 대응할 수 있다.	이미지 오디오
Gradient Regularization	대부분의 적대적 공격은 모델 추론 과정에서의 경사gradient를 보고 공격이 이루어진다. 학습 모델의 경사가 출력으로 노출되는 것을 방지하는 것에 중점을 둔다. – Gradient Regularization: 모델의 경사를 일관된 형태로 유지(예: Bit Plane Feature Consistency[BPFC] regularizer, Second-Order Adversarial Regularizer[SOAR]) – Gradient masking: 출력에 노이즈를 추가하거나, 학습 중에 특정 부분을 제거함으로써 모델의 경사를 외부로부터 감춤(예: S2SNet)	이미지 오디오 텍스트
Gradient Masking		
Stochastic Network	학습 모델의 불확실성을 다루기 위한 확률적인 요소를 도입하는 네트워크를 말한다. 이를 통해 모델의 결정을 불확실하게 만들어 적대적 사례에 대한 저항성을 높인다. (예: defensive dropout, Random Self-Ensemble[RSE])	

책임성 투명성

요구사항 11 인공지능 모델 명세 및 추론 결과에 대한 설명 제공

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

  * 사람이 인공지능 모델의 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 설명가능성explainability 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.

11-1 인공지능 모델의 명세를 투명하게 제공하는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템의 투명성을 확보하는 방안 중 하나는 인공지능 모델 또는 서비스의 개발, 테스트 및 배포 과정에서 발생한 다양한 정보를 문서로 작성하는 것이다. 모델의 명세를 작성한 상세 문서가 확보될 경우, 사용자가 인공지능 모델과 관련된 정보를 요구했을 때 모델의 목적, 입·출력 정보, 성능, 편향 여부 및 신뢰 점수 등의 결과들을 투명하게 공개할 수 있다.
- IBM과 WEF에서는 모델의 명세를 작성한 문서를 통해 인공지능 시스템의 투명성을 확보하는 방안을 제시한다. 특히, IBM은 개발한 시스템의 알고리즘 공개 없이 필요에 따라 인공지능 모델의 주요 정보 및 구성 요소를 설명할 수 있도록 하는 문서의 예시를 제공한다.

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

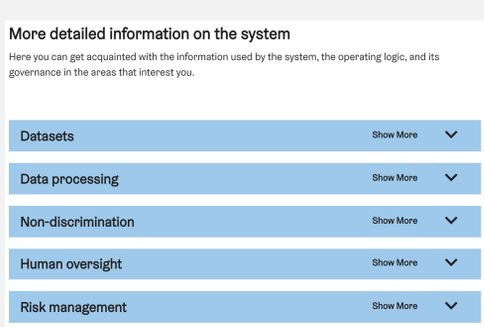
Yes ☐ No ☐ N/A ☐

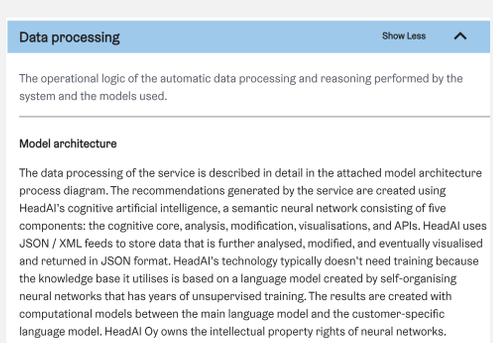
- 인공지능 시스템의 투명성을 높이고 시스템 사용자가 인공지능 기반 프로그램 구성 요소를 파악할 수 있는 정보를 제공하는 것은 시스템 신뢰성을 높이는 데 중요한 요소이다. 이를 위해 인공지능 모델 개발 과정에서 모델의 명세를 작성한 모델 상세 문서를 확보할 경우, 사용자에게 인공지능 시스템의 구성 요소를 파악할 수 있는 정보를 제공할 수 있다.
- 모델 상세 문서 작성 시에는 인공지능 생명주기와 관련된 이해관계자들을 고려하여 각자 필요한 정보를 선택하여 확인할 수 있도록 관련 정보를 포함하여야 한다. 다음은 이해관계자에 따른 모델 상세 문서 내 필요 정보 예시이다.

이해관계자에 따른 모델 상세 문서 예시

이해관계자	모델 상세 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, 신뢰 점수 등의 평가 결과
모델 운영자	• 모델 운영 및 모니터링 결과 측면의 성능 평가 지표, 성능 저하 환경 요인, 최적 결과 도출 환경 등

참고 암스테르담·헬싱키의 'AI 공공 설명' 웹사이트 구축 사례





https://ai.hel.fi/en/ai-register/

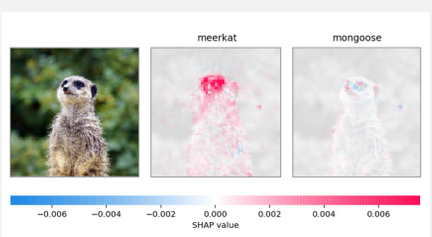
네덜란드 암스테르담과 핀란드 헬싱키가 유럽에서 처음으로 AI 기반 공공서비스 작동과정을 설명하는 웹사이트인 'AI 레지스터'를 만들었다. AI 레지스터는 AI와 빅데이터 기반 공공서비스를 이용하는 시민들이 AI에 관한 이해도를 돕기 위해 제작된 '공공 AI 상세 설명서'이다. 이는 편향과 개인 정보 침해 등을 향한 대중의 우려를 잠식시키려는 노력의 일환이다. 해당 웹사이트는 데이터셋과 데이터 처리 및 모델 아키텍처에 관한 설명을 제공함으로써 인공지능 시스템의 투명성을 확보하려고 노력하고 있다.

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

Yes ☐ No ☐ N/A ☐

- 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.
- 인공지능 모델의 투명성이 높거나, 모델 자체적으로 설명을 제공하는 경우 이를 응용할 수 있다. 반면, 모델의 복잡성이 높고 내재적 설명 방안이 없는 경우 별도의 설명가능한 인공지능XAI, eXplainable AI 기술 적용을 고려해야 한다. 다양한 기술 중 데이터·모델의 특성에 맞게 선택해야 하므로, 선행된 연구들을 검토한 후 적용 가능한 방안을 선택하고 적용하는 작업이 모델 개발 과정에 포함될 필요가 있다.
- 모델에 적합한 XAI 기술이 마련되지 않은 경우, 대안적인 방법을 통한 인공지능 시스템의 투명성 확보가 필요하다. 기술을 적용할 수 있는 경우라도 사용자가 도출 과정을 수용할 수 있을 만큼 충분치 않을 수 있으므로 기술 외적인 보완이 요구되기도 한다. XAI 기술 적용 가능 여부를 검토한 후, XAI 기술 적용이 가능하다면 11-2a를 활용하고, 적용이 어렵거나 보완이 필요한 경우 11-2b를 활용할 수 있다.

참고 모델 추론 결과의 도출 과정 설명 – SHAP를 사용한 근거 시각화



SHAP Example: Deep learning example with GradientExplainer
https://github.com/shap/shap

왼쪽 그림은 SHAP 알고리즘을 이용하여, 인공지능 모델에 입력된 미어캣의 이미지를 미어캣 또는 몽구스로 판정할 때 이미지 내의 어떤 픽셀이 어떤 방향으로, 얼마만큼 영향을 주는지를 산출, 시각화한 것이다.

- 정상 분류(미어캣): 미어캣의 안면부와 주요 형상을 이루는 픽셀 영역에서 양의 방향(적색)으로 결과에 영향을 주고 있음을 확인
- 오분류(몽구스): 같은 영역에서 음의 방향(청색) 영향이 발생하거나 아무런 영향이 없는(백색) 경우가 다수임을 확인

이러한 분석은 블랙박스인 인공지능 모델이 실제로 어떻게 작동하고 있는지를 사용자가 이해하기 쉬운 형태로 제시한다.

11-2a 인공지능 모델에 적합한 XAI[eXplainable AI] 기술을 적용하였는가?

Yes ☐ No ☐ N/A ☐

- 현존하는 XAI 기술은 속성에 따라 다음의 3가지 기준으로 분류할 수 있다.
  - ✓ 모델 내부 구조를 파악한 후 설명하는 내재적 방법[intrinsic methods]
    모델 입·출력만을 분석하여 설명하는 외재적 방법[extrinsic methods]
  - ✓ 특정 추론 결과에 대한 도출 과정을 설명하는 지역적 방법[local methods]
    모델의 전반적인 추론 행동을 설명하는 전역적 방법[global methods]
  - ✓ 특정 모델에만 적용할 수 있는 종속적 방법[model-specific methods]
    여러 모델에 적용할 수 있는 독립적 방법[model-agnostic methods]
- 인공지능 모델에 적용 가능한 XAI 기술이 무엇인지는 해당 모델과 데이터의 특성에 의해 결정된다. 예를 들어 대부분의 심층 학습[deep learning]처럼 설명 방법이 내재되어있지 않은 경우, 모델 독립적인 외재적 방법을 이용하거나 모델에 종속적인 형태의 설명 방법을 적용해볼 수 있다.
- 또한, 관련 이해관계자는 모델 특성에 적합한 XAI 기술을 선택해야 하며, 계량화, 시각화, 문서화 등 의미 전달에 효과적인 수단을 이용하여 결과를 제공할 수 있어야 한다. 또한 개발 착수 시 설명가능성을 고려하여 아키텍처를 설계하는 접근도 필요하다.

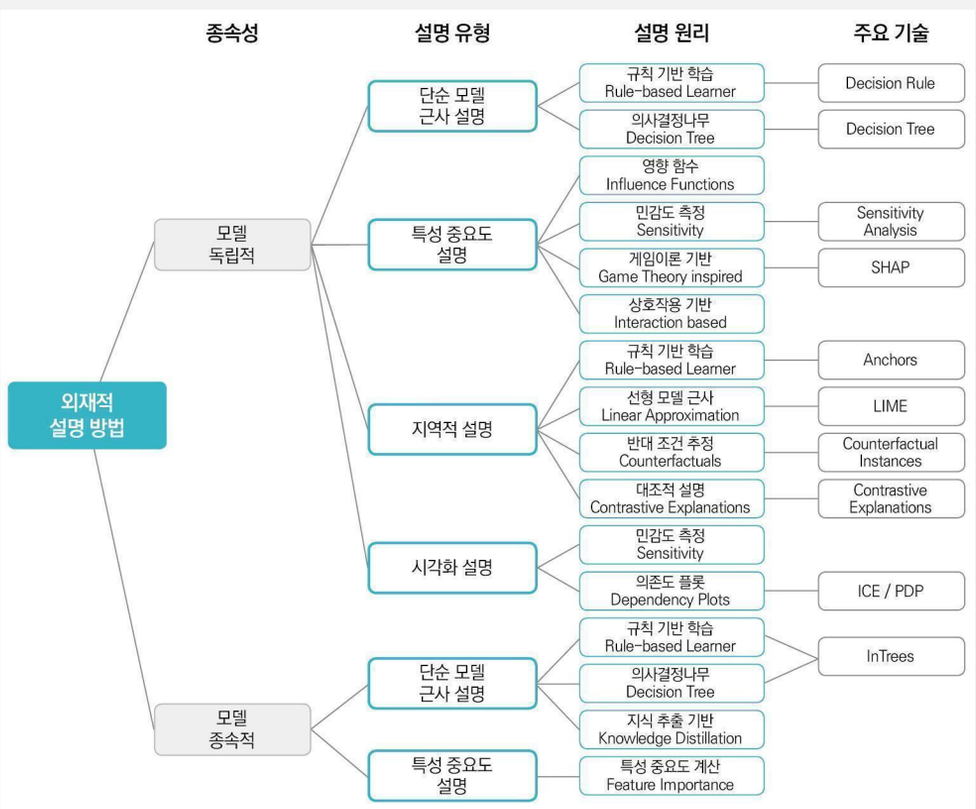
대표적인 XAI 기술 예시

기술명	설명 및 분류	특징
SHAP SHapley Additive exPlanations	게임 이론의 섀플리[Shapley] 값 개념을 기반으로 하며, 이 값을 사용하여 각 특성이 모델의 예측에 얼마나 기여했는지를 수치화하여 설명하는 기술 • 분류: 외재적, 지역적, 독립적	• 확실한 이론적 기반(Shapley)에 근거 • 복잡한 모델도 적용 가능 • 결과의 재현성 보장 • 높은 계산 비용 • 현존하는 기술 중 가장 널리 활용
LIME Local Interpretable Model-agnostic Explanations	복잡한 모델을 설명하기 어려울 때 대리 모델[surrogate model]을 만들어 설명하는 기술 중 하나로, 개별 예측을 설명하기 위해 입력값 주변의 작은 지역에 대한 대리 모델을 생성하여 각 특성의 영향을 설명하는 기술 • 분류: 외재적, 지역적, 독립적	• 다양한 모델에 적용 가능 • 샘플링에 따라 결과 변동 • 원본 모델의 복잡성을 충분히 근사하지 못할 수 있음
ICE/PDP Individual Conditional Expectation / Partial Dependence Plots	모델의 특성이 결과에 미치는 영향을 시각화하는 기술로, ICE는 특성을 변화시키면서 각 입력값에 대한 출력 영향을 설명하며, PDP는 ICE를 평균화하여 전체 예측에 미치는 영향을 설명 • 분류: 외재적, 지역적/전역적, 독립적	• 영향을 시각적으로 파악하기 용이 • 고차원 데이터의 경우 효과적인 상호작용 시각화가 어려움 • 실질적으로 표 형식 데이터[tabular data]에 한정
LRP Layer-wise Relevance Propagation	신경망의 계층[layer]별 기여를 순전파·역전파의 과정을 통해 계산하여, 각 계층이 전체 예측에 얼마나 기여하는지 수치화하여 설명하는 기술 • 분류: 외재적, 지역적, 독립적	• 종단 간 설명과 계층별 중요도 각각 설명 가능 • 모델에 맞게 별도 개발 필요 • 복잡한 모델의 해석 어려움

참고 인공지능 모델 특성에 따른 XAI 기술 선택 방안

#	모델 특성	XAI 기술 선택 방안
1	완성된 수식 또는 논리구조로 출력에 이르기까지의 작동 과정이 명백히 서술되는 경우 (예: Linear Regression, Logistic Regression, Decision Tree)	완성된 모델의 아키텍처와 특성별 가중치 또는 분류 기준을 그대로 설명에 활용
2	작동 원리가 수식의 형태로 제공되고, 모델의 생성 결과에 대한 각 특성의 영향 정도를 모델 자체적으로 객관적이고 상호 비교 가능한 수치 형태로 제공하는 경우 (예: Random Forest, XGBoost, LightGBM)	완성된 모델의 아키텍처와 특성별 가중치 또는 분류 기준을 그대로 설명에 활용
3	작동 원리가 수식의 형태로 제공되고, 외재적 방법 또는 추가적인 프로세스를 이용하여 각 특성의 영향을 수치화하여 제공하는 경우	
	표 형식 데이터처럼 고정된 특성 구성으로 정형화된 데이터를 이용하는 경우 (예: SVM, MLP 기반 심층학습 모델)	SHAP, PDP/ICE 등 (각 특성의 전역적인 영향을 수치화하여 제공하는 외재적 방법)
	이미지, 텍스트와 같이 비정형 데이터를 이용하는 경우 (예: CNN, Attention 기반 심층학습 모델)	SHAP, LIME, ICE, LRP, Attention을 활용한 모델의 Attention Score 등 (특정 표본에 대한 특성의 영향을 수치화하여 제공하는 외재적 방법)

참고 외재적 설명 방법의 분류



출처: Belle V and Papantonis. 2021. Principles and Practice of Explainable Machine Learning. Front. Big Data.

11-2b XAI 기술 적용이 불가능한 경우, 기술 외 대안을 마련하였는가?

Yes No N/A
☐ ☐ ☐

- XAI 기술 적용이 불가능하다는 것은 모델의 의사결정 요인이 무엇인지 계량적인 방법으로 나타낼 수 없다는 것을 뜻한다. 이 경우, 대안적인 방법을 통한 인공지능 시스템의 투명성 확보가 필요하다. 다음 네 가지 방안이 인공지능 시스템의 투명성 확보를 위한 대안이 될 수 있다.

  ✓ 의사결정 요인에 대한 간접 추정 근거 제시: 학습 데이터에 대한 기술적descriptive 분석과 특성 설명, 모델 아키텍처의 의도된 작동 원리 등이 있다. 또한, 시스템 배포 이후 모델이 추론에 사용한 데이터와 프로세스의 추적을 허용하고, 제대로 문서화되어 있는지 확인하는 과정이 필요하다.

  ✓ 신뢰성 확보를 위한 검증·평가: 실제 시스템의 유효성 검증과 검증에 대한 분석 결과를 활용할 수 있다. 시스템의 성능에 대한 객관적 평가 메트릭을 제시하고, 해당 메트릭이 도출된 테스트 데이터의 특성을 제시함으로써 모델의 기대 성능과 성능이 확보될 수 있는 조건을 사용자가 이해할 수 있도록 한다. WEF의 〈Companion to the Model AI Governance Framework〉에서는 프로덕션 환경에서의 반복적 테스트, 예외 식별 테스트 등을 수행할 수 있음을 언급하고 있다.

  ✓ 모델의 정상적인 작동 조건과 예상되는 오작동 및 위험에 대한 문서화: 인공지능 시스템의 기술적 한계를 사용자에게 적합한 방식으로 전달하는 것도 중요하다. 이는 15-1에서 더 자세히 언급한다. 11-3에서 언급한 것처럼, 모델이 결과 예측과 함께 신뢰 점수confidence score 등을 제시하는 것 또한 한 가지 방법이 될 수 있다.

  ✓ 모델 성능 및 사용에 대한 기록 및 추적가능성 확보: 04-1에서 설명한 것처럼, 인공지능 시스템의 초기 설계 단계에서 로깅 메커니즘logging mechanism 등을 보장하여 시스템의 작동 상태를 지속적으로 감독할 수 있게 하는 것이 도움이 된다. (단, 이는 반드시 인공지능 시스템과 관련된 비즈니스 모델 및 지식재산에 대한 정보가 항상 공개되어야 함을 의미하는 것은 아니다.)

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

Yes ☐ No ☐ N/A ☐

- 인공지능 모델의 추론 결과는 모델 구조에 따라 확률값probability, 신뢰 점수confidence score, 불확실성uncertainty 등의 수치로 설명될 수 있다. 인공지능 모델의 적용 분야와 사용자 특성에 따라, 수치를 통한 설명은 사용자의 최종적인 의사결정에 도움을 줄 수 있다.
- 하지만 수치를 통한 설명은 사용자의 직관에 반할 수도 있고, 모델 학습 방법의 한계로 인해 도출된 수치 자체에 오류가 있을 수도 있다. 따라서 인공지능 시스템 개발 시, 모델의 추론 결과에 대해 사용자의 판단을 도울 수 있는 설명 제공의 필요성과 더불어 설명의 적절성에 대한 기술적 평가가 이뤄져야 한다.

  * 고도로 복잡한 모델의 경우, 현존하는 기술로는 추론 결과에 대한 신뢰 점수와 불확실성을 정확하게 추론하는 것이 어려울 수 있다. 따라서, 신뢰 점수와 불확실성을 통해 모델 추론 결과를 설명할 수 없다면 11-2b 를 활용한 투명성 확보 전략 수립이 도움이 될 수 있다.

11-3a 모델 추론 결과에 대한 설명이 필요한지 검토하였는가?

Yes ☐ No ☐ N/A ☐

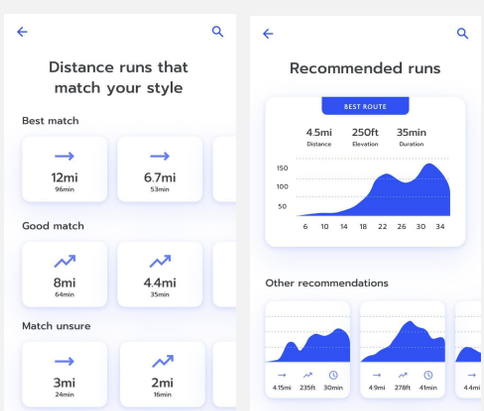
- 인공지능 시스템의 활용 분야 및 사용자의 특성에 따라 추론 결과와 함께 수치를 통한 설명 제공의 필요성 및 적절한 제공 방식에 대한 차이가 발생한다.
  - ✓ WHO의 〈Ethics and Governance of Artificial Intelligence for Health: WHO Guidance〉에서는 의료 분야 진단 보조 목적의 인공지능의 경우, 모델의 추론 결과만을 제시하면 입력 데이터의 잡음, 새로운 관측치의 입력, 의료진의 자동화 편향과 같은 인적 요인 등 예상치 못한 문제 발생 시 진단 오류를 발생시킬 위험이 있다고 언급하였다. 따라서 모델의 추론 결과와 함께 확률값, 신뢰 점수, 불확실성 등의 수치를 함께 제공하는 것이 사용자의 판단에 도움을 준다.
  - ✓ FAQ 질의응답을 제공하는 대화형 인공지능 시스템의 경우, 최종 사용자는 인공지능 시스템에 대한 전문지식이 담보되지 않는 불특정 다수의 고객이다. 인공지능 시스템이 사용자에게 발화 의도를 재확인하는 등의 형식이 사용성을 향상시킬 수 있지만, 추론 결과와 함께 수치를 제공하는 것은 오히려 사용자의 혼란을 유발할 수 있다.
- 모델의 설계 단계에서 신뢰 점수, 불확실성 등을 분리할 수 있도록 설계된 경우나 신뢰 점수 계산 과정에 학습 데이터에 포함되지 않은 이상치에 대한 분류 가능성이 고려된 경우, 추론 결과에 대한 수치의 제공은 사용자의 이해를 도울 수 있다. 그러나 이러한 고려가 되어 있지 않은 모델은 사용자가 수치를 잘못 해석할 우려가 있다.
  - ✓ 높은 신뢰 점수와 낮은 불확실성은 일반적으로 좋은 추론을 의미하지만, 모델이 항상 높은 신뢰 점수와 낮은 불확실성을 나타낸다면 과적합overfitting 여부를 확인하는 등 각 수치의 의미와 실용성을 평가해야 한다.

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

Yes ☐ No ☐ N/A ☐

- 인공지능 모델의 추론 결과에 대한 설명을 위해 관련 수치 혹은 수치의 재해석 결과를 제공하기로 결정하였다면, 이를 전달하는 방식에 대한 검토가 이어져야 한다. 전달 방식은 사용자의 이해 수준을 복합적으로 고려하여 설계하는 것이 좋으며, 수치를 명시적으로 제공하거나, 수치를 재해석하여 범주화된 결과를 제공하는 방법 등이 있다.
- 수치를 명시적으로 제공할 때 단순히 백분율만 표시한다면, 사용자는 해당 수치가 높은 것인지 낮은 것인지, 또는 수치가 무엇을 의미하는지 파악하기 어려울 수 있다. 또한, 대부분의 인공지능 모델이 100% 신뢰 점수로 예측을 수행하지 않기 때문에, 사용자의 혼란을 줄 수 있다. 따라서, 수치를 제공할 때는 사용자가 수치의 의미를 이해할 수 있도록 부가 설명이 동반되어야 한다.
- 수치를 명시적으로 제공하는 대신, 시스템은 N-best 대안alternative을 표시할 수 있다. Google Research의 〈People + AI Guidebook〉에서는 이 방법은 신뢰 점수가 낮은 상황에서 특히 유용할 수 있다고 언급한다. 신뢰 점수가 낮다는 것은 모델이 특정 출력에 대해 확신을 가지지 못하고 여러 가능성을 열어두고 있다는 의미이므로, 이때 대안을 함께 표시하면 사용자는 인공지능의 출력에 의존하지 않고 주체적으로 선택할 수 있다.

참고 신뢰 점수를 재해석하여 사용자에게 설명하는 방식의 예제



Google Research, People + AI Guidebook, "Explainability + Trust: Manage influence on user decisions"

Google Research에서는 신뢰 점수를 재해석하여 사용자에게 전달하는 여러 방안을 제안한다.

- 왼쪽의 첫 번째 그림은 신뢰 점수를 범주화하여 사용자에게 전달하는 방법으로, 수치를 명시적으로 전달하기보다 'Best', 'Good', 'Unsure'로 분류하여 전달한다.
- 두 번째 그림에서는 수치를 언급하지 않고 다양한 추천을 함께 제시한다. 신뢰 점수가 가장 높은 결과는 'Best'로 제시하고, 비교적 신뢰 점수가 낮은 결과는 'Other recommendations'로 제시한다.

04 시스템 구현

다양성 존중

요구사항 12 인공지능 시스템 구현 시 발생 가능한 편향 제거

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

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

Yes ☐ No ☐ N/A ☐

- 데이터 및 모델에 의한 편향 외에도 특정인이 작성한 소스 코드, 특정 선택을 암묵적으로 유도하는 사용자 인터페이스user interface 등을 통한 편향이 발생할 수 있다.
- 인공지능 시스템의 구현 단계에서 편향 방지를 위해, 작성된 코드를 주기적으로 검토하여 코드 구현 과정에서 특정 클래스 접근이 누락되지는 않았는지, 개발자의 편견이 코드에 반영되지는 않았는지 등을 확인해야 한다.
- 사용자 인터페이스 및 상호작용 측면에서는 표현 편향presentation bias이나 순위 편향ranking bias 등이 발생하지는 않는지 미리 확인하여 편향을 방지할 수 있도록 시스템을 설계하는 것이 바람직하다.

12-1a 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 클래스 접근이 누락되는 등 다양한 형태의 편향이 발생할 수 있다.
- 특히 규칙 기반 시스템rule-based system에서 다양한 분야의 경험이 있는 전문가를 선정하는 것이 바람직하다. 특정 분야 전문가의 지식을 기반으로 하드 코딩된 규칙을 사용할 경우, 출력 결과가 특정 클래스에 편향될 수 있으며 잠재적으로는 인지 편향cognitive bias을 일으킬 수 있다. 따라서 시스템의 편향 발생을 줄이기 위해서는 다양한 분야의 배경지식과 경험이 있는 전문가를 선정하는 것이 도움이 된다.
- 인공지능 시스템 설계 및 개발 단계에서 발생한 편향을 확인하기 위해 오픈소스 도구(예: FairML, Google What-If Tool)를 활용할 수 있다. 이러한 도구들은 주기적으로 출력 데이터의 통계를 분석하여 알려지지 않은 편향을 발견하거나, 미리 지정한 공정성 평가지표에 따라 기능의 위험 여부를 알리는 등의 기능을 수행한다. 이 도구들을 활용함으로써 구현과정에서 편향을 빨리 발견하고 대응할 수 있다.

12-1b 사용자 인터페이스 및 상호작용 방식으로 인한 편향을 확인하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템은 사용자 인터페이스에 의한 암묵적인 유도 또는 사용자의 의도적인 오남용에 따라 사용자 상호작용 편향이 발생할 수 있다.
- 사용자 상호작용 편향을 방지하기 위해서는 사용자 인터페이스 설계 및 구현 시 편향 발생 가능성이 있는 요소(예: 표현 편향, 순위 편향)를 미리 인식해 제거하여야 한다.
  - ✓ 표현 편향: 정보가 표현되는 방식에 따라 발생하는 편향이다. 예를 들어, 사용자는 제품 사용 시 보이는 콘텐츠만 클릭할 수 있으므로, 표시된 콘텐츠에서는 클릭이 발생하고 다른 콘텐츠에는 클릭이 발생하지 않는다. 이러한 사용자 인터페이스로 인해 특정 콘텐츠의 클릭만이 유도될 수 있다.
  - ✓ 순위 편향: 정보가 노출되는 순서에 따라 발생하는 편향이다. 사용자는 최상위 결과가 가장 관련성이 높고 중요하다고 생각하는 경향이 지배적이어서 상위에 노출된 결과가 하위에 노출된 결과보다 사용자의 선택 빈도가 높을 수 있다.

안전성 책임성 투명성

요구사항 13 인공지능 시스템의 안전 모드 구현 및 문제발생 알림 절차 수립

- 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.

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

Yes No N/A
☐ ☐ ☐

- 고장 안전fail-safe은 산업 전반에서 사용되는 일반적 개념으로, 고장이나 오류로 문제가 발생하더라도 안전한 상태를 유지할 수 있는 방법 및 기능을 의미한다. 이는 인공지능 시스템에도 적용될 수 있다. 인공지능 시스템에서도 외부의 공격, 인적 오류human error, 인공지능 모델의 성능 저하, 편향 발생으로 인한 사회적 물의, 사고 등이 예상되는 경우, 이의 발생 원인을 파악하고 해결하거나 사용자에게 정상적인 기능으로 복구할 수 있는 방법을 제시하여야 한다. 이러한 대처 방법이 작동하는 상태를 안전 모드라고 한다.
- 안전 모드를 구현하는 방법과 예시는 아래와 같다.
  - ✓ 시스템에 문제 발생 시 기능 정지 및 피드백 제공 화면으로 전환
  - ✓ 시스템에 문제 발생 시 서비스 제공 초기 화면 혹은 상태로 복구
  - ✓ 인공지능 판단 결과의 불확실성이 높거나 문제 발생 가능성이 높은 경우, 이에 대한 의사결정을 회피하거나 사용자에게 상황에 대한 안내 제공
  - ✓ 사용자의 악의적인 의도를 파악하고 이에 대한 입력을 거절
  - ✓ 자동 및 자율 운영 중 시스템에 문제 발생 시 사람의 개입 유도
  - ✓ 예상되는 사용자 오류에 대해 안내 및 대응 제공

13-1a 문제 상황에 대한 예외 처리 정책이 마련되어 있는가?

Yes ☐ No ☐ N/A ☐

- 시스템에 문제가 발생하는 상황에서 기능 정지, 화면 전환 및 서비스 제공 초기 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.
  - ✓ 예를 들어, 인공지능 스피커가 음성을 제대로 인식하지 못해서 부정확한 입력값을 가지면, 낮은 성능과 더불어 불확실성도 높아질 것이다. 이러한 상황에서 인공지능 스피커는 사용자에게 "무슨 말인지 잘 모르겠어요." 등의 회피형 답변을 제공하는데, 이것 역시 문제 상황에 대한 조치 중 하나라고 할 수 있다.
  - ✓ 단, 회피형 답변으로 해결할 수 없는 높은 위험도의 인공지능 서비스의 경우, 문제 상황에 대한 예외 처리 정책도 중요하지만 모델 자체의 개선 및 폐기도 고려해야 한다.

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

Yes ☐ No ☐ N/A ☐

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

인공지능 데이터 및 모델 대상 공격 예시

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

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

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

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템이 인공지능 모델의 판단 결과를 활용하여 시스템 동작을 제어하거나, 사람의 안전 및 환경에 영향을 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.
- 예시로, 자율주행 자동차 전방의 방해물 객체 인식을 통해 조향하는 인공지능 모델의 인식 결과가 불명확하거나 불확실성이 높은 경우, 운전자의 개입을 유도하고 제어권을 이양하는 기능이 고려되기도 한다.

참고 인공지능의 의사결정에 대한 사람의 개입 정도

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

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

13-1d 예상되는 사용자 오류에 대한 안내 및 대응을 제공하는가?

Yes ☐ No ☐ N/A ☐

- 사용자 오류는 외적으로는 서비스 최종 결과물을 사용하는 사용자에게서, 내적으로는 서비스 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.

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

Yes ☐ No ☐ N/A ☐

- 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.
- 시스템의 자체적인 점검 기능은 서비스 성능 저하나 외부 공격에 대한 검사 등을 수행한 후 가능한 범위 내에서 이에 대응하고, 해당 사실을 시스템 운영자에게 전달할 수 있는 체계를 갖춰야 한다.
- 사용자 의견 전달 기능은 시스템의 일시적인 오류나 도출 결과에 편향이 발생하는 등 문제가 생길 때, 사용자가 해당 사실을 시스템 운영자에게 전달할 수 있는 체계를 갖춰야 한다.

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

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템에서 편견 혹은 차별 등의 윤리적 문제의 발생 가능성을 확인하고, 문제 발생 시 이를 위한 알림 기능 혹은 절차가 수립되었는지 점검한다.
- 윤리적 문제 알림 절차의 경우, 먼저 인공지능 시스템에서 자체적인 신뢰 정도를 평가할 수 있는 기준과 점검 항목을 만든다. 주요 점검 항목의 예시는 다음과 같다.

  ✓ 인권보장, 사생활 보호, 다양성 존중, 침해금지, 공공성, 연대성, 데이터 관리, 책임성, 안전성, 투명성 등
- 시스템 자체 점검 기능 외에도, 시스템 운영 중 사용자가 윤리적 문제를 발견할 경우 시스템 운영자에게 의견을 전달할 수 있는 기능도 개발되어야 한다.

13-2b 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.
- 대표적인 성능지표로는 F1-score, IoU[Intersection over Union], mAP[mean Average Precision] 등이 있다. 평가 결과 성능 저하가 확인되면 이를 시스템 운영자에게 전달하고, 운영자는 성능 저하 원인을 찾아 개선을 진행하는 등의 절차를 마련해야 한다.

투명성

요구사항 14 인공지능 시스템의 설명에 대한 사용자의 이해도 제고

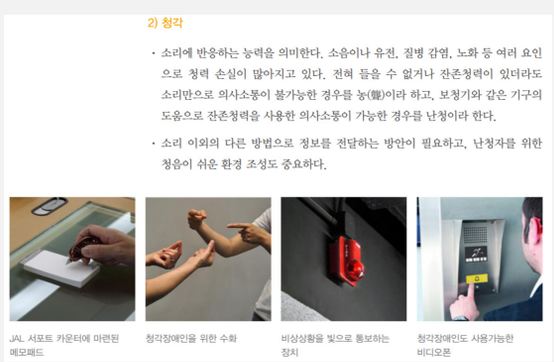
- 모델의 추론 결과에 대해 설명을 제공하는 기법을 적용하여도 사용자가 바로 이해해 해석하기 어려운 경우가 많다. 따라서 인공지능 시스템의 운영자 혹은 서비스 제공자는 사용자에게 제공되는 결과가 이해 가능한지[understandability], 해석 가능한지[interpretability], 설명 가능한지[explainability]를 평가한다.

14-1 인공지능 시스템 사용자의 특성[user characteristics]과 제약사항을 분석하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템의 결과가 적절한지 평가하기 위해서는 먼저 해당 결과를 읽는 사용자를 고려해야 한다. 사용자가 누군지에 따라 결과(설명)의 수준, 깊이 그리고 맥락이 정해지는 만큼 사용자에 대한 자세한 분석이 수행되어야 한다.

참고 서울시 유니버설 디자인 통합 가이드라인


2) 청각

· 소리에 반응하는 능력을 의미한다. 소음이나 유전, 질병 감염, 노화 등 여러 요인으로 청력 손실이 많아지고 있다. 전혀 들을 수 없거나 잔존청력이 있더라도 소리만으로 의사소통이 불가능한 경우를 농(聾)이라 하고, 보청기와 같은 기구의 도움으로 잔존청력을 사용한 의사소통이 가능한 경우를 난청이라 한다.

· 소리 이외의 다른 방법으로 정보를 전달하는 방안이 필요하고, 난청자를 위한 청음이 쉬운 환경 조성도 중요하다.

JAL 서포트 카운터에 마련된 메모패드 | 청각장애인을 위한 수화 | 비상상황을 빛으로 통보하는 장치 | 청각장애인도 사용가능한 비디오폰

http://ebook.seoul.go.kr/Viewer/4XB4DO5LDINO

'서울시 유니버설 디자인 통합 가이드라인'에서는 공공 시설물을 이용할 수 있는 다양한 이용자 특성(성별, 연령, 국적, 신체 크기, 질병, 인지능력)을 사전에 정의 및 분석하였다.

14-1a 사용자 특성에 따른 세부 고려사항을 분석하였는가?

Yes No N/A
☐ ☐ ☐

- 서비스 기획 단계에서 사용자의 선호도와 요구 사항needs에 집중했다면, 설명을 평가하기 위해서는 각 사용자의 다양한 특성을 고려해야 한다. 예를 들어 서비스 사용자 중 어린이가 이해 가능한 그래프와 단어 및 어휘의 제한이 있음을 고려해야 한다.
- 사용자 특성 분석을 위해 고려해야 할 요소의 예시는 다음과 같다.

사용자 특성 분석을 위한 고려사항 예시

구분	상세 구분	고려사항
연령	아동, 성인, 노인 등	• 아동 또는 노인의 경우, 성인과 비교해 이해할 수 있는 어휘, 단어에 한계가 있어 사용자 연령을 고려해야 함
장애 유무	장애인, 비장애인	• 신체적 제약으로 발생할 수 있는 한계를 고려해야 함. 그 예로는 신체 크기, 신체 능력, 인지능력이 있음
지식	초보자, 전문가 등	• 관련 서비스의 경험 여부와 사전 배경지식의 차이로 지식수준이 다름을 고려해야 함

14-2 사용자 특성에 따른 설명을 제공하는가?

Yes No N/A
☐ ☐ ☐

- 서비스를 이용하는 사용자는 다양하여 인공지능 시스템의 결과가 서로 다른 입장에서 설명이 해석되고 오해가 생길 수 있다. 따라서 14-1 에서 분석된 사용자 특성을 고려하여 설명을 평가할 수 있는 기준 항목을 수집한다. 설명 평가의 기준으로는 명확성, 구체성, 정확성 등을 고려할 수 있다.

14-2a 사용자 특성에 따른 설명 평가 기준을 수립하였는가?

Yes No N/A
☐ ☐ ☐

- 다양한 사용자가 서비스를 이용하는 만큼 설명을 포괄적으로 평가할 수 있는 특성과 세부 항목을 정하는 단계가 필요하다. 설명의 평가 기준은 명확성, 구체성, 적절성, 정확성 등이 될 수 있다. 이때, 설명의 기대치는 사용자 특성(예: 나이, 직업)에 따라 달라지며, 데이터 유형data type이나 모달리티modality에 따라서도 각 항목에서 고려되어야 할 내용들이 달라질 수 있다. 다음은 설명 평가를 위한 예시이다.

설명의 평가기준별 평가 항목 예시

구분	평가 항목
명확성	• 사용자에게 다른 오해를 불러일으킬 만한 표현·단어·어휘는 없는가? • 불필요한 설명이 있진 않은가? • 해당 설명을 통해 사용자가 기대하고 얻고자 하는 정보가 모두 들어있는가?
구체성	• 사용자의 구체적 행동을 끌어낼 수 있도록 명확한 주어·목적어·동사를 활용해 설명되는가?
적절성	• 주어진 설명이 사용자의 특정 지식수준을 요구하지는 않는가? • 배경지식 혹은 사전 경험이 필요하진 않은가? • 설명이 사용자에게 유용한가? • 독자를 고려한 전문 용어, 약어에 대한 설명을 제공하는가? • 설명이 제공되는 시점이 적절하였는가?
정확성	• 설명과 함께 제공되는 자료의 그림과 설명이 모두 일치하는가? • 사전에 제공된 예상 결과의 설명과 실제 결과가 모두 일치하는가? • 내부 알고리즘과 정확히 일치하는 설명인가?

14-2b 사용자가 이해하기 어려운 전문 용어 사용을 지양하였는가?

Yes No N/A
☐ ☐ ☐

- 텍스트를 통해 설명하는 경우, 다양한 독자를 배려해 전문 용어를 최대한 지양하고 필요한 경우, 용어에 대한 설명을 추가로 작성해 주는 것이 바람직하다. 그 예로 자연어 처리 기술 중, 문장 내 특정 단어를 사용자 수준에 맞춘 적절한 단어로 변환해주는 기술을 인터페이스에 적용할 수 있다.

14-2c 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 있게 할 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.

14-2d 설명이 필요한 위치와 타이밍은 적절한가?

Yes No N/A
☐ ☐ ☐

- 잘 작성된 설명이 적절한 위치 및 타이밍에 나타나 이해를 돕는 것도 중요하다. 이를 위해 설명이 단발성이어야 하는지, 여러 번 반복하여 강조시켜야 할지 숙고하고, 어느 위치에 놓여야 사용자가 잘 읽을 수 있는지 고려하는 것이 필요하다.
- 이와 더불어 작성된 설명의 위치와 타이밍이 적절한지를 조사하기 위해서는 14-2e 의 웹로그 분석, A/B 테스트 등 사용자 조사 기법을 활용할 수 있다.

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

Yes ☐ No ☐ N/A ☐

- 사용자 경험UX, User eXperience은 한 개인이 특정한 제품, 시스템, 또는 서비스를 사용하며 느끼는 모든 것을 의미한다. 또한, 그 개인이 인지하는 유용성, 사용 편의성, 효율성 등의 시스템 특성을 포함한다. 설명을 평가하기 위해 사용자 조사user research 기법을 활용할 수 있다.
- 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 구분되며, 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, 사용자 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 테스트A/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 무슨 행동을 하는지를 조사하는 방법 (예: 웹로그 분석, A/B 테스트, 아이 트래킹eye tracking)
  - 사용자 태도 기반 조사attitudinal user research: 사용자가 무엇을 말하는지를 조사하는 방법 (예: 카드 소팅card sorting, 심층 인터뷰, 요구사항 조사)

05 운영 및 모니터링

책임성 투명성

요구사항 15 서비스 제공 범위 및 상호작용 대상에 대한 설명 제공

- 사용자가 인공지능 시스템이 제공하는 서비스를 올바르게 사용하고, 제공된 서비스를 오·남용하지 않도록 서비스의 목적, 범위, 제한사항, 면책조항disclaimer, 상호작용 대상을 포함한 내용을 설명한다.

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

Yes No N/A
☐ ☐ ☐

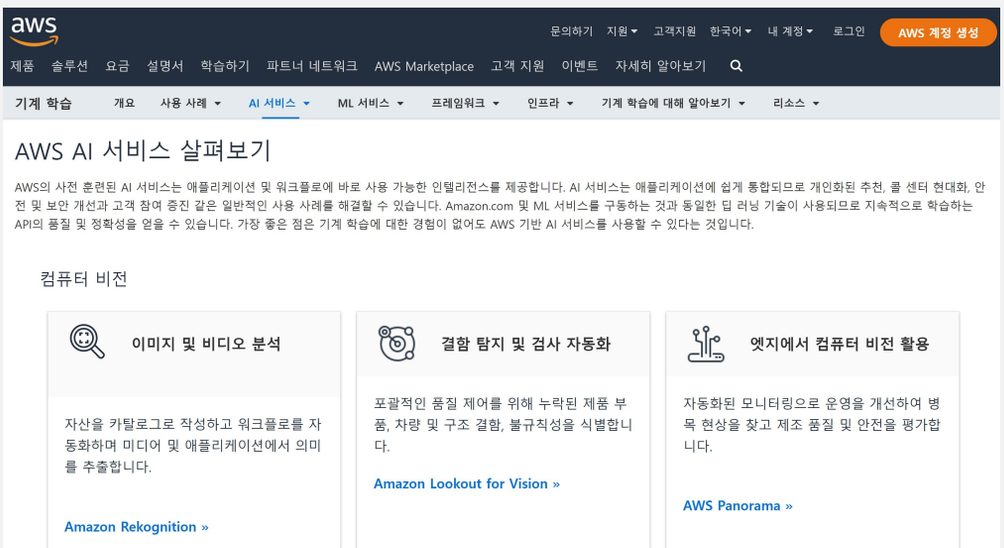
- 인공지능의 활용 범위가 넓어지면서 사용자가 서비스 기능에 대한 기대를 실제 서비스 제공 범위보다 더 넓게 오해하는 경우가 발생한다. 따라서 서비스 목적, 범위, 제한사항, 면책조항에 대한 설명을 제공함으로써 인공지능 기술의 오·남용을 방지하고 사용자의 서비스에 대한 기대치를 조정하는 것이 중요하다.
- 서비스 제공자는 인공지능이 제공하는 결과가 사용자에게 미치는 영향을 설명하고, 필요시 해당 결과를 되돌릴 수 있는지 또한 제공하여 해당 서비스의 올바른 사용을 유도해야 한다.

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.
- 인공지능 서비스가 오용 또는 남용될 경우, 인공지능 모델이나 시스템상의 새로운 취약점을 생성하거나 예상치 못한 사회적 이슈를 발생시킬 수 있다. 따라서 서비스가 의도한 목적을 벗어나 잘못 사용되는 것을 방지하기 위해, 이해관계자는 잠재적 오·남용 영역을 식별한 후 사용자가 이를 인식할 수 있도록 관련 사례와 처벌 내용 등을 알려야 한다.

참고 AWS Amazon Web Services AI의 서비스 목적 및 목표



https://aws.amazon.com/ko/machine-learning/ai-services/

AWS는 별도 웹사이트로 자사 인공지능 서비스 목적과 서비스를 통해 어떤 목적을 달성할 수 있는지 설명한다.

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.

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

Yes No N/A
☐ ☐ ☐

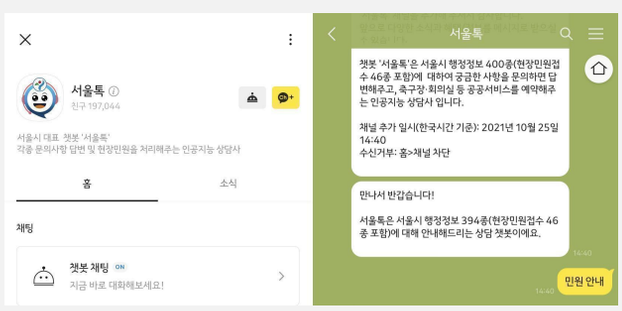
- 최근 인공지능 시스템을 의인화함으로써 사용자가 친밀감을 향상하고 사용성을 높이려는 서비스가 많아지고 있다. 그러나 인공지능 기술이 고도화되며 인간과 구분이 어려워져 사용자는 상호작용의 대상이 사람인지, 시스템인지 혼란을 겪을 수 있다. 따라서 서비스 제공자는 사용자가 상호작용하는 대상을 명확히 알림으로써 사용자가 겪을 혼란을 줄여야 한다.
- 특히, 생성 AI 기반 서비스를 통해 생성된 콘텐츠는 딥페이크, 가짜뉴스 등에 오용될 수 있는데, 이러한 문제는 해당 콘텐츠가 인공지능에 의해 생성되었음을 명시함으로써 사용자의 혼란을 방지할 수 있다. 이를 위해, 생성 콘텐츠에 가시적 워터마킹 기술을 적용한다면 사용자가 콘텐츠의 생성 출처를 쉽게 구분하도록 할 수 있다.

15-2a 사용자와 인공지능이 상호작용하는 서비스 범위를 명시하였는가?

Yes No N/A
☐ ☐ ☐

- 서비스 내에서 사용자와 인공지능이 상호작용하는 범위를 명시해야 한다. 이로써 사용자는 어떤 작업이 자동화되고 어떤 작업을 직접 수행해야 하는지 이해할 수 있다. 또한, 인공지능과 상호작용하는 시점에는 사용자의 혼란을 방지하고 서비스에 대한 기대치를 조정할 수 있다.
- 특히, 대표적인 의인화 서비스인 대화형 인공지능(예: 챗봇)의 경우, 대화 상대가 사람이 아닌 인공지능이라는 사실을 사용자에게 알려야 한다. 더불어, 상호작용 대상이 인공지능에서 사람(예: 상담사)로 전환되는 시점에도 이러한 전환 사실을 명확하게 사용자에게 설명해야 한다.

참고 ‘서울톡’의 상호작용 사례



https://pf.kakao.com/_xemMXj

서울시는 행정 및 민원 접수 간소화를 위해 민원 상담 챗봇 '서울톡'을 운영하고 있다. 서울톡은 메신저 서비스에서 친구 추가와 대화창에서 상호작용의 대상이 시스템임을 알림으로써 실제 상담사와 혼동하지 않도록 구분시킨다.

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

Yes ☐ No ☐ N/A ☐

- 사용자에게 인공지능이 최종 의사결정을 내렸는지 또는 특정 결과에 기여했는지 등의 정보를 설명해야 한다. 예를 들어, 인공지능이 최종 의사결정을 내린 경우 사용자에게 해당 결정이 인공지능의 결과임을 명시적으로 사용자에게 전달해야 한다. 또한, 인공지능이 조언을 제시하고 최종 의사결정을 운영자가 내린 경우나, 사용자에게 최종 의사결정을 위임한 경우에도 관련 설명을 제공해야 한다.
- 미국 백악관에서 발표한 인공지능 권리장전을 위한 청사진Blueprint for an AI Bill of Rights에서는 자동화 시스템이 채용이나 신용평가 등의 분야에서 사용될 경우 사람들의 삶에 깊은 영향을 미치기 때문에, 잠재적인 피해로부터 보호하기 위해 사용자에게 자동화 시스템의 활용 여부를 명시해야 함을 언급하고 있다.

PART 3

부록

01 약어표

약어표

AI	Artificial Intelligence
ALTAI	Assessment List for Trustworthy Artificial Intelligence
API	Application Programming interface
BDPL	Boundary Differentially Private Layer
BPFC	Bit Plain Feature Consistency
CNN	Convolutional Neural Network
DVC	Data Version Control
EC	European Commission
ETSI	European Telecommunications Standards Institute
EU	European Union
GEN	Graph Extrapolation Network
ICE	Individual Conditional Expectation
IEC	International Electrotechnical Commission
IoU	Intersection over Union
ISO	International Organization for Standardization
LDA	Linear Discriminant Analysis
LIME	Local Interpretable Model-agnostic Explanation
LRP	Layer-wise Relevance Propagation
LSTM	Long-Short Term Memory
mAP	mean Average Precision
MLOps	Machine Learning model Operationalization management
NIST	National Institute of Standards and Technology
OECD	Organization for Economic Cooperation and Development
PDP	Partial Dependence Plots
RMF	Risk Management Framework
RSE	Random Self-Ensemble
SHAP	SHapley Additive exPlanations
SimCLR	Simple framework for Contrastive Learning of visual Representations
SMOTE	Synthetic Minority Oversampling TEchnique
SOAR	Second-Order Adversarial Regularizer
SVM	Support Vector Machine
TAI	Trustworthy AI
TR	Technical Reports
WEF	World Economic Forum
WIT	What-If Tool
XAI	eXplainable AI

02 용어표

용어표

용어명	정의
인공지능 **AI, Artificial Intelligence**	ISO/IEC는 인공지능을 '하나 이상의 주어진 작업을 수행하기 위해 모델 형태로 보유한 지식을 획득, 처리, 생성 및 적용하는 능력'이라고 정의한다. 또한, 유럽위원회는 인공지능 시스템을 '복잡한 목표가 주어졌을 때 데이터를 수집하여 환경을 인식하고 수집된 정형 또는 비정형 데이터를 해석하면서 물리적 또는 디지털 차원에서 작동하는 소프트웨어(또는 하드웨어) 시스템'이라고 정의한다. 과학의 한 분야로서 인공지능은 아래와 같은 접근방법 또는 기술을 포함하는 개념이다.[1)] ① 기계학습machine learning: 심층학습 및 강화학습이 구체적인 예임 ② 기계 논증machine reasoning: 계획, 스케줄링, 지식 표현 및 추론, 검색 및 최적화 포함 ③ 로보틱스robotics: 제어, 인식, 센서 및 액추에이터뿐만 아니라 다른 모든 기술을 사이버-물리 시스템에 통합하는 것 포함
신뢰할 수 있는 인공지능 **TAI, Trustworthy AI**	신뢰할 수 있는 인공지능이란 법과 윤리를 준수하면서 기술적으로 강건한 인공지능을 의미한다. 또한, 신뢰성은 검증가능한 방식으로 이해관계자의 기대를 충족하는 능력이라고[2)] 정의할 수 있다. 신뢰할 수 있는 인공지능의 구현을 위해서는 인공지능 시스템 전 생명주기에 걸쳐 관련 이해관계자의 철저한 신뢰성 확보 활동이 필수적이다.[3)]
인공지능 윤리 **AI ethics**	인공지능 윤리는 인공지능 기술을 개발하고 적용하는 과정에서 지켜야 하는 바람직한 사회적 원칙이나 규칙을 말한다. 인공지능에 관련된 윤리적 쟁점은 명확하지 않아 다양한 가치들의 상호충돌이 발생하므로, 사회 전체에 유익하고 개인의 권리를 존중하는 방향으로 해결해 나가야 한다. 따라서 인공지능의 개발자·사용자 모두가 인공지능 윤리를 준수해야 하는 주체가 된다. 세계 각국 및 관련 기관은 인공지능의 부작용 및 잠재적 위험 등에 대비하여 인공지능 윤리의 중요성을 인식하고, 최선의 윤리적 해결책을 찾기 위한 사회적 합의를 구체화하고 있다.[4)]
설명가능한 인공지능 **XAI, eXplainable AI**	인공지능 모델의 추론 결과만을 보게 되는 사용자는 예측된 결과가 어떤 요소에 의해 도출되었는지 알 수 없다.[5)] 설명가능한 인공지능[6)]이란 사용한 모델 정보, 결과 도출 과정에 대한 설명 및 추론 결과에 대한 설명이 제공되어 사용자가 해당 인공지능 모델을 신뢰할 수 있는 인공지능을 의미한다. 설명가능한 인공지능을 통해 사용자가 안심하고 사용할 수 있다.

1) European Commission, "**ALTAI – The Assessment List on Trustworthy Artificial Intelligence,**" p. 24, 2020. 7.
2) IBM, **Trustworthy AI,** [Online], Available: https://research.ibm.com/topics/trustworthy-ai
3) UNESCO, "**Ethics of AI,**" p. 18, 2022. 12.
4) UNESCO, "**Recommendation on the Ethics of Articial Intelligence,**" p. 36, 2022. 12.
5) OECD Legal Instruments, "**Transparency and explainability,**" p. 8, 2019. 5.
6) European Commission, "**ALTAI – The Assessment List on Trustworthy Artificial Intelligence,**" pp. 14-26, 2020. 7.

02 용어표

용어명	정의
인간-인공지능 인터페이스 **HAI, Human-AI Interface**	인공지능 서비스를 구성하는 주요 요소는 학습용 데이터, 모델 및 알고리즘, 인공지능 시스템, 인간-인공지능 인터페이스[HAI]이다. HAI는 인공지능 시스템 사용자와 운영자 등이 인공지능 시스템의 동작을 쉽게 이해할 수 있도록 하며, 인공지능의 오작동 시 사람에게 알리거나 제어권을 이양하는지 등을 담당하는 부분이다. 인터페이스에서는 사용자의 이해도, 호감도, 신뢰도 등의 관점에서 편향성을 갖지 말아야 하고, 사용자가 특정 선택을 암묵적으로 유도하는 등의 편향에 주의하여야 한다. HAI는 사용자 중심이어야 하며, 연령, 성별, 능력, 특성, 장애 여부에 관계없이 모든 사람이 인공지능 서비스를 사용할 수 있도록 설계되어야 한다.[7]
사용자 특징 **user characteristics**	인공지능 시스템을 활용하는 사용자 특징이란 사용자의 연령, 성별, 지역, 인종, 국적, 장애 유무, 질병, 인지능력, 지식 등을 의미한다. 사용자 특징에 따라 사용자별로 모델 추론 결과에 대한 설명을 이해하고 해석하는 정도가 달라질 수 있다. 따라서 인공지능 시스템 운영자나 서비스 제공자는 사용자의 특징을 고려하여 추론 결과에 대한 설명을 제공해야 한다. 추론 결과에 대한 설명은 사용자 특징에 따라 명확해야 하고, 구체적이어야 하며, 적절해야 한다.
인공지능 시스템 생명주기 **AI system life cycle**	OECD에서는 인공지능 시스템 생명주기가 4가지 단계로 구성된다고 정의하였고, 그 4단계는 ① 설계와 데이터와 모델링[design, data and modelling], ② 검증과 확인[verification and validation], ③ 배포[deployment], ④ 운영 및 모니터링[operation and monitoring]이다.[8] 각 단계는 반드시 순서대로 진행될 필요는 없으며 반복적으로 발생하기도 한다.
인공지능 시스템의 의사결정 **AI decision-making**	인공지능의 의사결정이란, 실세계와 가상세계 환경에서 인지 작업을 학습하고 수행 능력을 생성하는 모델과 알고리즘을 통합하여 최적의 결과를 도출하는 것이다. 이 의사결정은 지식 모델링과 표현을 통하여 데이터를 활용하고 이들의 상관관계를 계산하여 다양한 수준의 자율성을 바탕으로 수행된다.[9]
인공지능 거버넌스 **AI governance**	일반적인 개념의 거버넌스[governance]는 정책 결정에 있어 정부 주도의 통제와 관리에서 벗어나 다양한 이해당사자가 주체적인 행위자로 협의와 합의 과정을 통하여 정책을 결정하고 집행해 나가는 사회적 통치 시스템으로 정의한다.[10] 인공지능 시스템에서는 윤리적 문제 발생 가능성이 잠재하고 있으므로 인공지능으로부터 야기되는 사회적 영향과 결과를 예측하고 대비하는 조직을 구성하는 것이 매우 필요하며, 이와 관련한 법, 규제, 정책, 표준 및 지침을 정리하여 내부적 준수 규정을 수립하고 이를 관리, 감독하는 프로세스 체계를 인공지능 거버넌스라 한다.

7) European Commission, "**ALTAI - The Assessment List on Trustworthy Artificial Intelligence**," p. 17, 2020. 7.
8) OECD Legal Instruments, **The technical landscape**, [Online], Available: https://www.oecd-ilibrary.org/sites/8b303b6f-en/index.html?itemId=/content/component/8b303b6f-en
9) UNESCO, "**Recommendation on the Ethics of Artificial Intelligence**," p. 10, 2021. 11.
10) J. Pierre and B. G. Peters, "**Governance, Politics and the State**," 2000. 1.

02 용어표

용어명	정의
인공지능 시스템의 신뢰성 AI trustworthiness	인공지능 시스템의 신뢰성이란 인공지능을 신뢰할 수 있는지에 대한 가치 기준으로서 유효하고, 신뢰할 수 있으며, 검증되고, 공정하며, 편견이 관리되고, 안전하고, 탄력적이며, 책임 있고, 투명하며, 설명 가능하고, 해석 가능하며, 프라이버시가 보장되는 것을 의미한다.[11] 인공지능 시스템 기획자, 개발자, 사용자 등 이해관계자들이 데이터 및 모델의 편향, 인공지능 기술에 내재한 위험과 한계를 해결하고, 인공지능을 활용하고 확산하는 과정에서 부작용을 방지하기 위해 준수해야 하는 가치 기준이다.
인공지능 시스템의 견고성 AI robustness	인공지능 시스템의 견고성이란 인공지능이 외부의 간섭이나 극한적인 운영 환경 등에서도 사용자가 의도한 수준의 성능 및 기능을 유지하는 상태이다. 즉, 인공지능 시스템이 실행 중 오류와 잘못된 입력에 대처할 수 있는 능력을 의미하며, 잘못된 입력이나 스트레스가 많은 환경에서 시스템이 올바르게 작동할 수 있는 정도에 따라 평가된다.
인공지능 시스템의 공정성 AI fairness	인공지능 시스템의 공정성이란 인공지능이 데이터를 처리하는 과정에서 특정 그룹에 대한 차별, 편견, 편향성을 나타내거나, 이를 포함한 결론에 이르지 않는 상태이다. 신뢰할 수 있는 인공지능TAI, Trustworthy AI은 전체 인공지능 시스템의 생명주기 동안 포용성inclusiveness과 다양성diversity을 가져야 하므로 인종, 나이, 성별, 능력, 특성에 관계없이 모든 사람이 인공지능 제품이나 서비스를 사용할 수 있도록 설계되어야 한다. 인공지능 운영 과정에 있어서도 역사적인 편향성이나 잘못된 거버넌스를 제거하여 다양성과 차별 금지를 포함한 공정성을 제공해야 한다. 특히, 인공지능 접근성이 낮은 장애인이나 사회적 약자에게 인공지능의 활용을 공정하게 제공해야 한다.[12]
다양성 존중 respect for diversity	다양성 존중이란 인공지능이 특정 개인이나 그룹에 대한 차별적이고 편향된 관행을 학습하거나 결과를 출력하지 않으며, 인종·성별·연령 등과 같은 특성과 관계없이 모든 사람이 평등하게 인공지능 기술의 혜택을 받을 수 있는 것을 의미한다.
인공지능 시스템의 안전성 AI safety	인공지능 시스템의 안전성은 인공지능 시스템이 예측한 대로 동작하거나, 위험 발생 가능성이 완화 및 제거된 것을 의미한다. 즉, 인공지능 시스템이 정의된 조건 아래에서 인간의 생명, 건강, 재산 또는 환경이 위험하게 되지 않도록 하는 것이다. 인공지능 시스템의 안전한 운영을 위해서는 이해관계자들은 아래 내용을 확인해야 한다.[13] - 개발자는 책임있는 설계 및 개발 관행을 준수하고 시스템을 적절하게 사용하는 방법을 운영자에게 명확히 전달 - 운영자와 최종 사용자들은 책임있는 의사 결정을 수행

11) NIST, "**AI Risk Management Framework: Second Draft,**" p. 10, 2022. 8.
12) European Commission, "**ALTAI – The Assessment List on Trustworthy Artificial Intelligence,**" p. 16, 2020. 7.
13) NIST, "**AI Risk Management Framework: Second Draft,**" p. 13, 2022. 8.

02 용어표

용어명	정의
인공지능 시스템의 위험 관리 AI risk management	위험 관리는 인공지능 제품 또는 서비스가 생명 주기 전반에 걸쳐 설계상 신뢰할 수 있도록 하는 데 도움이 되는 예방 프로세스이다. 이는 이해관계자와 취약점을 식별한 후, 위험을 평가하고, 조직의 기준에 기반하여 위험 해결 결정을 내리는 것을 포함한다.[14] 인공지능 위험 관리는 일반 기술에서의 위험 관리와 다름이 없지만, 인공지능이 사이버 보안, 프라이버시, 안전성, 인프라 구조를 포함한 정보 시스템에 대한 위험에 영향을 주는지를 포함한다.[15]
인공지능 시스템의 지속가능성 AI sustainability	인공지능 시스템의 지속가능성은 인공지능 시스템의 책임감있고, 윤리적인 개발, 배포 및 유지 관리를 의미하며, 인공지능 모델의 학습 및 배포에 대한 환경적 영향, 인공지능 의사결정의 투명성과 책임성, 인공지능 이익과 부담에 대한 공정한 분배, 인공지능 시스템의 장기적인 안전과 견고성 등 다양한 고려사항을 포함한다. 인공지능 시스템의 개발, 배치, 사용 및 전체 공급망은 지속가능한 형태로 제공되어야 하며, 인공지능 시스템을 위한 학습 데이터 활용이나 에너지 소비 등을 최소화하는 방식으로 개발되어야 한다. 인공지능 시스템 공급망 전체에 걸쳐 환경친화성을 확보하기 위한 조치가 강력하게 이루어져야 할 것이다.[16]
인공지능 시스템의 프라이버시 AI privacy	인공지능 시스템의 프라이버시는 인간의 자율성autonomy, 정체성identity 및 존엄성dignity을 보호하는 데 도움이 되는 규범과 관행으로[17], 개인에 관련된 데이터를 불법적으로 수집 및 사용하는 개인의 사생활 침해나 업무에 대한 침해로부터 자유롭게 벗어나는 것[18]을 의미한다. 일반적으로 인공지능 시스템 설계, 개발 및 배포의 과정에서도 익명성, 기밀성 및 제어와 같은 프라이버시 가치는 보장되어야 하며, 정책적 관점에서 프라이버시 관련 위험은 보안, 편향 및 투명성과 겹칠 수 있으며, 기술적 관점에서의 안전성 및 보안성은 프라이버시를 촉진하거나 감소시킬 수 있다.
인공지능 시스템의 투명성 AI transparency	인공지능 시스템의 투명성은 인공지능이 추론한 결과를 인간이 이해하고 추적할 수 있으며, 인공지능이 추론한 결과임을 알 수 있도록 하는 것이다. 인공지능을 신뢰한다는 것은 인공지능 시스템의 투명성이 어느 정도 지원되느냐와도 관련이 깊다. 투명성은 인공지능 시스템의 기능, 구성요소 및 절차에 대해서 알 수 있도록 보여주는 속성이며, 외부 검사에 사용할 수 있는 데이터, 모델, 알고리즘, 품질 보증, 거버넌스를 만드는 것 등이 포함된다. 투명성을 보장함으로써 인공지능의 공정성fairness, 프라이버시privacy, 인공지능 윤리AI ethics 등의 가치를 보호할 수 있다.[19]

14) ISO/IEC TR 24028, "**Overview of trustworthiness in artificial intelligence,**" p. 10, 2020. 5.
15) NIST, "**AI Risk Management Framework: Second Draft,**" p. 1, 2022. 8.
16) European Commission, "**ALTAI – The Assessment List on Trustworthy Artificial Intelligence,**" p. 19, 2020. 7.
17) NIST, "**AI Risk Management Framework: Second Draft,**" p. 16, 2022. 8.
18) ISO/IEC TR 24028, "**Overview of trustworthiness in artificial intelligence,**" p. 4, 2020. 5.
19) ISO/IEC TR 24028, "**Overview of trustworthiness in artificial intelligence,**" p. 23, 2020. 5.

02 용어표

용어명	정의
인공지능 시스템의 설명가능성 AI explainability	인공지능 시스템의 설명가능성은 인공지능 시스템 결과에 영향을 미치는 중요한 요인을 인간이 이해할 수 있는 방식으로 표현하기 위한 인공지능 시스템의 속성이다.[20] 인공지능 모델의 추론 결과만을 보게 되는 사용자는 예측된 결과가 어떤 요소에 의해 도출되었는지 알 수 없다. 그래서 인공지능 모델 제공자는 사용자에게 신뢰성을 제공할 필요가 있다.[21] 이는 곧 인공지능 모델과 제품 및 서비스에 대해 사용자에게 안심하고 사용할 수 있는 인공지능 서비스임을 입증하는 과정이다. 이 과정은 일반적으로 완화 조치mitigation measures라고 하는 인공지능 취약성을 완화할 수 있는 가능한 제어 및 지침의 일환이다. 완화 조치의 핵심 요소는 투명성transparency과 설명가능성 제공이다. 설명가능성은 설명가능한 인공지능XAI, eXplainable AI[22]이란 형식으로 표현되기도 하는데, 이는 사용한 모델 정보, 결과 도출 과정에 대한 설명 및 추론 결과에 대한 설명을 제공하여 사용자가 신뢰할 수 있는 인공지능 모델을 말한다.
인공지능 시스템의 책임성 AI accountability	인공지능 시스템의 책임성이란 인공지능 시스템을 개발, 배포, 사용 등 전 생명주기에 걸쳐 책임을 보장하기 위한 체계가 보장되어있는 것을 의미한다. 즉, 부정적 영향이 발생하는 경우 적절한 책임소재를 보장하는 메커니즘이 준비되어야 한다는 것이다.[23] 인공지능 책임성은 적절한 방식을 통해 위험을 식별하고 완화하는 위험 관리risk management와 밀접한 관련이 있다.
인공지능 시스템의 재현가능성 AI reproducibility	인공지능 시스템의 재현가능성이란 동일한 기계학습 실행 과정을 반복하여 수행했을 때, 모델의 추론 결과가 동일한 결과에 달할 수 있는 성질이다.
인공지능 시스템의 추적가능성 AI traceability	인공지능 시스템의 추적가능성이란 인공지능 시스템 개발, 학습 데이터 출처 그리고 인공지능 시스템 프로세스 로깅, 결과 등이 어떤 상태이고 어떻게 변해왔는지에 대한 기록을 확인할 수 있는지에 대한 여부를 의미한다.[24] 즉, 인공지능 시스템이 추적 가능하다는 뜻은 인공지능 시스템의 개발 프로세스를 적절하게 기록하여 필요할 때 확인할 수 있도록 하므로 투명성을 보장할 수 있다는 것이다.
인공지능 시스템의 해석가능성 AI interpretability	인공지능 시스템의 해석가능성은 인간이 이해할 수 있는 방식으로 인공지능 시스템의 내부를 설명하는 것이다. 유럽위원회는 해석가능성이 설명가능성explainability 또는 이해가능성understandability을 참조하며, 인공지능 시스템의 요소가 해석 가능할 때 외부 관찰자가 이를 이해하고 그 의미를 찾을 수 있다고 하였다.[25] 유럽위원회의 해석가능성에 대한 정의는 설명가능성을 참조하고 있어 유사한 개념으로 보고 있음을 알수 있다.

20) ISO/IEC TR 24027, "**Bias in AI systems and AI aided decision making**," p. 2, 2021. 11.
21) OECD Legal Instruments, "**Transparency and explainability**," p. 8, 2019. 5.
22) European Commission, "**ALTAI – The Assessment List on Trustworthy Artificial Intelligence**," pp. 14-26, 2020. 7.
23) European Commission, "**ALTAI – The Assessment List on Trustworthy Artificial Intelligence**," p. 21, 2020. 7.
24) European Commission, "**ALTAI – The Assessment List on Trustworthy Artificial Intelligence**," p. 29, 2020. 7.
25) European Commission, "**ALTAI – The Assessment List on Trustworthy Artificial Intelligence**," p. 27, 2020. 7.

02 용어표

용어명	정의
인공지능 시스템의 회복탄력성 AI resilience	인공지능 시스템의 회복탄력성은 인공지능 시스템의 생명주기 단계에서 적대적 공격adversarial attack을 받거나, 시스템 입력이 데이터 분포의 변화에 따라 달라질 때 인공지능 시스템이 실패하지 않고 강건하고robust, 다양한 공격에 따른 시스템의 실패와 고장시 신속히 적응adapt하고 복구할recover 수 있는 것을 의미한다.[26][27]
데이터셋 dataset	데이터셋은 논리적으로 의미 있는 데이터의 집합을 의미한다.[28] 기계학습에서의 데이터셋은 학습용 데이터셋training dataset, 검증 데이터셋validation dataset, 시험 데이터셋test dataset으로 구성된다.
데이터셋 특성 dataset feature	데이터셋 특성은 데이터셋을 구성하는 요소 중의 하나로 기계학습 모델에 입력되는 독립변수이다. 특성 이외의 데이터셋의 다른 요소는 결괏값target 또는 라벨label로서, 기계학습 모델에 데이터셋 특성을 입력했을 때 출력되는 종속변수이다. 데이터셋 특성은 반드시 하나일 필요는 없으며 특성의 개수를 데이터셋의 차원dimension이라고 한다.
데이터 과학자 data scientist	데이터 과학은 수학 및 통계, 전문 프로그래밍, 고급 분석, 인공지능 및 기계학습을 특정 전문 지식과 결합하여 조직의 데이터에 숨겨진 통찰을 발견하는 과학이다. 데이터 과학자는 데이터 과학의 지식을 겸비하고 조직에서 의사결정을 내리는 데 도움이 되는 데이터 수집, 분석 및 해석을 담당하는 분석 전문가라 할 수 있다. 데이터 과학자의 역할은 수학자, 과학자, 통계학자, 컴퓨터 프로그래머를 포함하여 여러 전통 및 기술 직업의 요소를 결합한 결과이다.[29]
원천 데이터 source data	원천 데이터는 인공지능 모델을 학습시키거나 테스트 또는 검증하는데 사용되는 데이터이며, 원시 데이터raw data를 라벨링 공정에 투입하기 위해 필요한 전처리 등 데이터 정제 작업을 수행한 데이터로서 라벨링과 어노테이션이 부착되지 않은 상태의 데이터이다.[30]
메타데이터 metadata	메타데이터는 다른 데이터에 대한 정보를 제공하는 데이터로 정의한다.[31] 인공지능에서는 원시 데이터의 특징을 메타데이터에 기록하여 향후 데이터를 재활용하는 상황이나 동일한 형식의 데이터를 추가로 수집해야 할 때, 데이터에 대한 정보를 전달하는 목적으로 메타데이터를 활용한다.
라벨링 데이터 labeled data	라벨링 데이터는 원천데이터에 부여한 참값ground truth, 파일 형식, 해상도 등의 데이터 속성과 설명, 주석 등이 포함된 어노테이션의 집합이다.

26) A. Kumar and S. Mehta, "**A Survey on Resilient Machine Learning**," arXiv:1707.03184v1, 2017. 7.
27) CFAR, **RESILIENT & SAFE AI**, [Online], Available: https://www.a-star.edu.sg/cfar/research/research-focus/resilient-safe-ai
28) ISO/IEC TR 24027, "**Bias in AI systems and AI aided decision making**," p. 2, 2021. 11.
29) TechTarget, **Data Scientist**, [Online], Available: https://www.techtarget.com/searchenterpriseai/definition/data-scientist
30) 한국정보통신기술협회, "**인공지능 학습용 데이터셋 구축안내서**," p. 3, 2021. 2.
31) Merriam-Webster, **Metadata**, [Online], Available: https://www.merriam-webster.com/dictionary/metadata

02 용어표

용어명	정의
비정형 데이터 **unstructured data**	비정형 데이터는 수집된 데이터의 크기와 내용이 서로 달라 통일된 형식이나 구조로 정리하기 곤란한 데이터로 정의할 수 있다. 비정형 데이터는 형식과 구조가 정해져 있지 않으므로 정형 데이터와는 달리 연산이 불가능하다는 특징이 있다. 기업은 전통적으로 관계형 데이터베이스RDB, Relational DataBase의 테이블 형태로 구성되어 있는 정형 데이터를 이용하여 내부 업무의 전산화를 추진해 왔다. 하지만 최근 들어 스마트폰의 활용 범위가 확대되고 인터넷상에서 SNS 활동이 폭증하면서 비정형 데이터가 많이 생산되고 있어, 비정형 데이터 분석을 기반으로 하는 비즈니스가 급속하게 증가하고 있다.[32]
이상 데이터 **abnomal data**	정상 데이터normal data와는 현저히 다른 특성을 가진 관측치를 식별하는 작업을 이상 탐지anomaly detection라고 하며, 이때 식별되는 관측치를 이상 데이터라 부른다.[33] 즉, 이상 데이터는 전반적인 패턴을 벗어나는 데이터로 정의할 수 있으며, 편차deviation를 초래하는 주요 요인이다.[34]
데이터 전처리 **data pre-processing**	데이터 전처리는 성능을 보장하거나 향상하기 위해 데이터를 사용하기 전에 데이터를 조작하거나 삭제하는 과정을 의미한다.[35] 데이터 전처리는 데이터 품질을 개선하고 인공지능 알고리즘 성능을 높이거나, 특정 데이터 마이닝 작업을 수월하게 하기 위한 목적으로 수행된다. 이때, 전처리에는 데이터를 정리, 변환 및 통합하는 것이 포함된다.
데이터 정제 **data cleaning**	데이터로부터 신뢰성 있는 결과를 얻기 위해서는 수집된 데이터를 분석 도구 또는 기법에 맞게 정제하는 과정이 필요하다. 데이터 정제란[36] 데이터 전처리 활동으로, 학습 데이터 구축을 위하여 오류 데이터를 탐지하고 이를 수정하는 작업을 의미한다. 즉, 모델 설계 이전 단계에서 오류 데이터를 탐지하여 결측치를 채우거나 이상값outlier이나 노이즈를 제거하는 과정을 통해 데이터의 신뢰도를 높이는 작업이라 할 수 있다. 데이터 정제와 혼동되는 개념으로 특성 공학feature engineering이 있다. 데이터 정제가 모델을 학습시키기 위해 1차적으로 수행하는 데이터 전처리 활동이라면, 특성 공학은 학습 성능을 더욱 고도화하기 위해 주어진 입력변수를 변형하여 목표변수를 더 잘 설명할 수 있도록 변환하는 전처리 활동이다.[37]

32) 이긍희, 함유근, 김용대, 원중호, "**빅데이터의 이해와 활용**," pp. 14-15, 2022. 7.

33) P. Tan, M. Steinbach, and A. Karpatne, "**Introduction to DATA MINING-Second Edition**," p. 13, 2018. 1.

34) D. S. Moore, G. McCabe, W. M. Duckworth, and L. Alwan, "**The Practice of Business Statistics: Using Data for Decisions-2nd Edition**," p. 13, 2008. 2.

35) Tableau, **Guide To Data Cleaning: Definition, Benefits, Components, And How To Clean Your Data**, [Online], Available: https://www.tableau.com/learn/articles/what-is-data-cleaning

36) P. Tan, M. Steinbach, and A. Karpatne, "**Introduction to DATA MINING-Second Edition**," p. 42, 2018. 1.

37) **데이터전처리**, [Online], Available: https://zhining.tistory.com/123

02 용어표

용어명	정의
데이터 라벨링 **data labeling**	데이터 라벨링은 인공지능이 기계학습에 활용할 수 있도록 기능이나 목적에 부합하는 라벨링 데이터를 원천데이터에 부착하는 활동이다.
데이터 식별 **data identification**	데이터 식별[38]이란 인공지능 모델링이나 시뮬레이션 모델링에서 목표 달성에 필요한 데이터셋과 속성(정확도, 샘플 주기, 형식 등)을 강조해서 표시하는 절차를 의미한다.
데이터 리니지 **data lineage**	데이터 리니지는 시간 경과에 따른 데이터 흐름을 추적하는 프로세스로, 데이터의 출처, 데이터에 일어난 변화, 데이터 파이프라인 내에서의 최종 목적지에 대한 자세한 정보를 제공하는 프로세스이다.[39] 자세히는, 조직 내에서 데이터의 수집 및 저장, 이동과 통합, 분석, 폐기에 이르는 모든 생명주기를 추적하는 것으로, 특정 데이터가 어떤 시스템에서 생성됐고 어느 데이터베이스에 저장됐는지, 이후 어떤 과정을 거쳐 현재는 어디서 활용되고 있는지 등을 투명하게 확인해 실시간으로 파악하는 것을 목적으로 하는 것이다.[40]
데이터 강건성 **data robustness**	데이터 강건성은 인공지능 모델이 학습용 데이터의 이상값outlier, 중독poisoning 및 회피evasion등의 공격에 영향을 받지 않는 것을 의미한다.
데이터 공격 **data attack**	데이터 공격은 인공지능 서비스 개발 또는 운영 과정에서 인공지능의 기밀성confidentiality과 무결성integrity을 공격하기 위하여 의도적으로 학습 데이터를 변질시키거나 입력 데이터를 오염시켜 예상과는 다른 결과를 나타내도록 하는 것을 의미한다.
데이터 중독 **data poisoning**	데이터 중독이란 인공지능 모델의 학습 데이터에 악의적인 데이터를 주입하는 행위를 말한다. 공격자는 데이터 중독을 통해 인공지능 시스템이 학습하지 말아야 할 내용을 학습하게 만들어 바람직하지 못한 결과를 출력하게 한다.[41] 이를 위해, 기계학습 데이터베이스에 침투하여 부정확하고 그릇된 예측을 하도록 유도하는 정보를 입력한다.[42] 이렇게 주입된 데이터로부터 학습한 알고리즘은 원래 의도하지 않은 결과를 도출한다.
클래스 불균형 **class imbalance**	클래스 불균형이란 분류 문제에서 각 클래스의 샘플 수에 큰 차이가 있는 상황을 의미한다. 실제 분류 문제를 풀 때 자주 클래스 불균형 문제에 마주치게 되는데, 이는 샘플의 분포가 편향되거나 왜곡되어 발생한다.

38) Stephan Onggo, James Hill, "**Data identification and collection methodology in a simulation project: an action research**," 2012. 3.
39) IBM, **What is data lineage**, [Online], Available: https://www.ibm.com/topics/data-lineage
40) ITDaily, **데이터 흐름/계보 관리 방법 및 시스템 특허 획득**, [Online], Available: http://www.itdaily.kr
41) European Commission, "**ALTAI – The Assessment List on Trustworthy Artificial Intelligence**," pp. 9-26, 2020. 7.
42) NIST, "**AI Risk Management Framework: Second Draft**," p. 15, 2022. 8.

02 용어표

용어명	정의
경험 오차 **empirical error**	기계학습 모델을 학습할 때는 학습 데이터셋을 직접 관찰하면서 조작할 수 있으므로 모델의 예측값과 학습 데이터셋의 참값 사이의 차이를 측정할 수 있다.[43] 학습기가 학습 데이터셋에서 만들어낸 예측값과 참값 사이의 차이를 경험 오차empirical error 또는 학습 오차training error라 부른다.[44] 회귀분석에서는 통상적으로 평균제곱오차MSE, Mean Squared Error를 사용하여 오차를 측정하며, 분류분석에서는 분류 오차율classification error rate을 사용하여 오차를 측정한다.
데이터 편향 **data bias**	데이터 편향이란 가용한 데이터가 모집단이나 연구 현상을 적절히 표현하지 못하여 데이터셋의 특정 요소가 과장되거나 축소되어 표현될 때 발생하는 오류이다. 편향된 데이터셋은 기계학습 모델이 실세계를 정확하게 나타내지 못해서 왜곡된 결과 또는 낮은 정확도를 초래한다. 데이터 편향은 기술 제약 조건 등에서 발생하며 인간의 인지 편향, 교육 방법론, 교육 인프라의 차이로도 발생할 수 있다.
데이터 다양성 **data diversity**	데이터 다양성이란 데이터 소스(내부 소스와 외부 소스 포함)의 다양성뿐만 아니라 데이터 구조의 다양성을 의미하는 포괄적 개념이다.[45][46] 데이터 다양성에 대한 정의는 기존 데이터베이스에서 쉽게 관찰하고 관리할 수 없는 데이터의 다양한 속성과 관련이 있다. 데이터 다양성은 콘텐츠, 지리 공간, 기계, 모바일, 스트리밍, 오디오, 비디오, 텍스트, 웹로그, 소셜 미디어 데이터와 같은 데이터 소스의 다양성을 의미하기도 하지만, 여기에는 정형, 반정형, 비정형 데이터를 포함하는 데이터 구조의 다양성이라는 의미도 내포하고 있다.[47]
보호 변수 **protective attribute**	보호 변수는 인공지능의 판단에 영향을 주지 않도록 설정한 변수이다. 인공지능 모델을 만드는 데 사용하는 데이터가 편향되어 있다면 올바른 모델이 나올 수 없다. 따라서 데이터 편향을 줄이기 위한 기술은 대부분 편향성과 관련되는 속성attribute[48]을 찾아내어 조정하는 데 중점을 두고 있다. 데이터의 불공정성을 측정하다 보면 개발자가 의도한 것과 달리 차별적 결과를 내놓는 특정 변수를 발견할 수 있는데 이러한 변수를 보호 변수로 지정하여 불공정성을 최소화할 수 있다. 보호 변수는 사회적 물의를 일으킬 수 있는 민감한 특성으로 보통 나이, 성별, 인종, 학력, 지역, 종교, 빈부 등이 이에 해당한다.

43) I. Goodfellow, Y. Bengio, and A. Courville, "**Deep Learning**," p. 107, 2016. 11.
44) Zhou Zhihua(김태헌 옮김), "**단단한 머신러닝**," p. 29, 2020. 2.
45) Information Resources Management Association, "**Big Data: Concepts, Methodologies, Tools, and Applications**," p. 974, 2016. 4.
46) ISO/IEC 20546, "**Overview and vocabulary**," p. 23, 2019. 2.
47) Information Resources Management Association, "**Big Data: Concepts, Methodologies, Tools, and Applications**," p. 3, 2016. 4.
48) ISO/IEC FDIS 24029-2, "**Methodology for the use of formal methods**," p. 1, 2023. 1.

용어명	정의
과적합/과소적합 overfitting/underfitting	과적합 또는 과대적합은 학습용 데이터셋의 특정 조건이나 구조를 과도하게 최적화하여 발생하는 문제로 검증 데이터셋에서 일반화가 잘 이루어지지 않는 기계학습 동작을 의미한다.[49] 과소적합은 그와 반대로 최적화가 불충분하게 수행되어 발생하는 문제로 검증 데이터셋에서 학습 데이터셋의 구조/패턴을 정확히 반영하지 못하는 현상을 의미한다. 과적합은 모델의 파라미터 수가 지나치게 많거나 학습용 데이터셋의 양이 부족한 경우에 흔히 발생한다. 과적합 모델에서는 학습이 많아질수록 새로운 데이터셋에 대한 분산variance이 과도하게 커지는 경향이 나타난다. 반면에, 과소적합 모델에서는 학습 및 검증 데이터셋 모두에 대해 부정확한 결과가 제공되므로 구조적 오류를 나타내는 편향이 커지는 경향이 있다.[50]
교차 검증 cross validation	학습용 데이터셋에서 모델을 학습할 때, 모델이 너무 간단하면 과소적합(높은 편향) 문제가 발생하고 너무 복잡하면 과대적합(높은 분산) 문제가 발생한다. 적절한 편향-분산 트레이드오프를 찾기 위해서는 모델의 일반화 성능을 평가할 필요가 있다.[51] 모델의 일반화generalization란 이전에 관측한 적이 없는 신규 데이터에 대해서도 모델이 잘 작동하는 능력을 가리키며, 일반화 오류generalization error란 신규 데이터에 대한 오차의 기댓값을 가리킨다.[52] 교차 검증이란[53] 시험 데이터셋에서 만들어내는 일반화 오류에 대해 신뢰할만한 추정치를 구하기 위해 모든 데이터를 학습 셋, 검증 셋, 시험 셋으로 무작위로 구분한 후, 학습 데이터셋으로 분석모형을 구축하고, 시험 데이터셋으로 분석모형의 성능을 평가하는 모델 평가 방법이다. 검증 데이터셋은 학습 데이터셋과 테스트 데이터셋 사이의 괴리를 보완하는 것으로 학습 내용을 검증하는 데 사용하며, 시험 데이터셋과 겹치면 안된다.
데이터 변경 이력 data change record	데이터 변경 이력이란 인공지능 시스템 전체 생명주기 동안 기록되는 학습 데이터의 자세한 변경 사항을 의미한다. 인공지능 모델은 사용한 데이터에 따라 학습 모델도 함께 달라진다. 따라서 모델 개발과정에서 학습 데이터가 변경될 경우, 학습 데이터 버전관리 및 변경이 발생한 원인을 추적할 수 있도록 데이터 변경 이력을 확보해야 한다.

49) P. Tan, M. Steinbach, and A. Karpatne, "**Introduction to DATA MINING-Second Edition**," p. 147. 2018. 1.
50) Amazon, **과적합이란 무엇입니까?**, [Online], Available: https://aws.amazon.com/ko/what-is/overfitting/
51) Sebastian Raschka and Vahid Mirhalili (박해선 옮김), "**머신 러닝 교과서 with 파이썬, 사이킷런, 텐서플로 개정 3판**," p. 247, 2021. 3.
52) I. Goodfellow, Y. Bengio, and A. Courville, "**Deep Learning**," p. 107, 2016. 11.
53) P. Tan, M. Steinbach, and A. Karpatne "**Introduction to DATA MINING-Second Edition**," pp. 165-166. 2018. 1.

02 용어표

용어명	정의
인공지능 모델 AI model	인공지능 모델은 인공지능 시스템을 개발할 때 해당 분야 서비스에서 수집된 데이터셋으로 모델을 만들기 위한 학습을 수행하고, 학습 알고리즘을 이용하여 목적에 맞는 특정 패턴을 만들어내는데, 이때 추출된 패턴을 의미한다. 학습 알고리즘은 데이터셋에서 패턴과 상관관계를 찾고 분석을 통해 최적의 의사결정과 예측을 수행하도록 설계된 알고리즘에 따라 모델을 학습시킨다.[54] 학습된 모델은 학습 알고리즘에 따라 다양한 형태로 구분된다. 수치 형태의 신경망 노드의 가중치 값, 혹은 통계 수치값으로 표현되거나, 규칙형태 등으로 표현이 된다. 최근 주로 이용되는 심층학습 알고리즘의 경우 인공지능 모델은 네트워크의 구조와 학습된 가중치로 구성된다.
인공지능 모델 개발자 AI model developer	인공지능 모델 개발자란 인공지능 서비스의 생명주기에서 인공지능 모델 개발, 시스템 구현, 운영 및 모니터링 과정의 주체이다. - 인공지능 모델 개발 단계에서는 인공지능 모델을 구현하고, 학습 모델의 편향적인 추론 결과나 공격에 대한 대응방안 마련과 학습 모델 추론 결과에 대한 해석, 모델의 확인 및 검증, 모델에 대한 성능평가까지를 담당한다. - 시스템 구현 단계에서는 기존 레거시 시스템과의 호환성을 제공하고, 기능 시험, 시스템 검증 배포 버전을 승인해 주는 역할을 수행한다. - 운영 및 모니터링 단계에서는 모델 모니터링 결과 분석을 통한 모델의 재학습, 모델의 편향성 제거, 공정성과 설명가능성 등 시스템 신뢰성을 모니터링하고 치명적 문제가 발생할 때 시스템 폐기의 의사결정 까지 관여한다.
인공지능 모델 명세 AI model specification	인공지능 모델 명세란 인공지능 시스템의 투명성과 신뢰성을 높이고, 사용자가 AI 기반 프로그램 구성 요소를 파악할 수 있도록 인공지능 모델 개발 과정에서 필요로 하는 설계와 기능에 대한 정보를 상세하게 기술한 문서이다. 여기에는 모델의 동작을 정의하는 알고리즘, 데이터 구조 및 매개변수에 대한 정보와 입출력 형식, 성능 메트릭 및 기타 관련 정보를 포함된다. 모델 명세는 모델이 일관되고 투명한 방식으로 설계, 구축 및 시험되도록 하며 모델의 지속적인 유지관리 및 개선을 위한 참조로 사용할 수 있으며, AI 모델의 개발, 배포 또는 사용에 관련된 모든 사람에게 필수적인 문서이다.
모델 추론 model inference	모델 추론은 기계학습 단계가 끝난 인공지능 서비스에 사용자가 새 데이터를 이용해 예측하는 것이다. 인공신경망을 이용한 심층학습은 학습 단계[learning steps]와 추론 단계[inference steps]로 나눌 수 있다.[55] 학습 단계가 축적된 많은 데이터를 바탕으로 각 신경망의 가중치[weight]를 업데이트해가며 심층학습 모델을 만들어 가는 과정이고, 추론 단계는 학습을 통해 만들어진 모델을 실제로 새 데이터에 적용하여 결과를 도출하는 과정이다.

54) KAIC, **AI 모델 개요**, [Online], Available: http://aicerti.com/14
55) ISO/IEC FDIS 24029-2, "**Methodology for the use of formal methods**," p. 10, 2023. 1.

용어명	정의
모델 튜닝 model tuning	모델 튜닝은 기계학습 모델 학습 단계에서 초매개변수hyperparameters를 보정calibrating하는 과정이다. 초매개변수는 모델 외적인 요소로, 알고리즘 사용자에 의해 정해지는 값으로 이는 학습 프로세스를 제어한다. 초매개변수의 예로는 학습 횟수number of epochs, 학습률learning rate, 학습-시험 분리 비율train-test split ratio, 배치 크기batch size, 의사결정 트리 알고리즘의 트리 깊이, 랜덤 포레스트 알고리즘random forest algorithm의 트리 수, k-평균 알고리즘k-means algorithm의 클러스터 k의 수, 신경망의 층layer수 등이 있다. 모델 튜닝은 모델 성능을 최대화하기 위해 이러한 초매개변수의 최적값을 찾는 과정이다. 따라서 모델 튜닝은 초매개변수 최적화hyperparameter optimization라고도 한다.
모델 사전 학습 model pre-training	모델 사전 학습은 모델 사전 훈련이라고도 하며, 하나의 작업 또는 데이터셋에서 모델을 먼저 학습시키는 것을 의미한다. 이렇게 사전 학습시킨 매개변수나 모델을 이용하여 다른 작업 또는 데이터셋에서 다른 모델을 학습시킨다. 이렇게 하면 모델을 처음부터 학습시키는 것보다 비용 및 시간 측면에서 유리하다.[56] 사전 학습 모델pre-trained model이란 다른 데이터셋에서 이미 학습된 모델이다.
모델 성능 평가 기준 model performance measure	기계학습에서는 학습기의 일반화 성능을 평가할 때 모델의 일반화 성능을 평가하는 기준이 마련되어야 하는데, 이를 모델 성능 평가 기준이라 한다.[57] 모델 성능 검증은 모델의 효율성 측정은 물론 두 개의 알고리즘의 성능을 비교하거나 새로운 데이터를 평가할 때 모델 성능이 더 좋아지는지 또는 더 나빠지는지 평가하는 데 사용할 수 있다.
인공지능 모델 공격/적대적 공격 AI model attack /adversarial attack	인공지능 모델 공격은 적대적 의도를 가진 사용자가 학습 데이터 및 기능을 도용하거나 다른 방식의 공격으로 인공지능 모델을 변형하거나 오용하는 것을 의미한다. 인공지능 모델 공격에는 모델 추출 공격과 모델 회피 공격이 있다. - 모델 추출 공격은 학습된 모델의 다양한 입력에 대한 인지 결과를 분석하고, 분류 기준을 추출하여 적용 중인 학습 모델과 유사한 성능의 대체 모델을 구성하는 방식의 공격이다. - 모델 회피 공격은 입력 데이터에 최소한의 변조를 가해 인공지능 모델을 속이는 방식의 공격이다.
모델 추출 공격 model extraction attack	모델 추출 공격은 기계학습 모델에 질의를 계속 입력하면서 결괏값을 분석함으로써 모델을 추출하는 공격이다.[58] 이 공격은 주로 서비스형 기계학습MLaaS, Machine Learning as a Service을 탈취하거나 전도 공격inversion attack이나 회피 공격evasion attack과 같은 2차 공격에 활용한다.

56) clarifai, **Ultimate Artificial Intelligence Glossary 2021**, [Online], Available: https://www.dlt.com/sites/default/files/resource-attachments/2021-01/Ultimate-Artificial-Intelligence-Glossary-2021-Ebook.pdf

57) Zhou Zhihua(김태헌 옮김), "**단단한 머신러닝**," p. 37, 2020. 2.

58) European Commission, "**ALTAI – The Assessment List on Trustworthy Artificial Intelligence**," p. 28, 2020. 7.

02 용어표

용어명	정의
모델 추출 공격 방어 기법 model extraction attack defensive method	모델 추출 공격 방어 기법이란 인공지능 모델 추출 공격에 대항하여 이를 방어하는 방법을 의미한다. 참고로, 모델 추출 공격이란 기계학습 모델에 질의를 계속 입력하면서 결괏값을 분석하는 방식의 공격을 의미한다.
모델 회피 공격 model evasion attack	모델 회피 공격은 공격자가 기계학습 시스템에서 오류를 생성하기 위해 입력 데이터를 조작하는 것을 목표로 하는 공격이다. 데이터 중독과 달리 모델 회피 공격은 시스템의 동작을 변경하지 않지만, 모델의 맹점과 약점을 악용하여 공격자가 원하는 오류를 생성하게 된다. 모델 회피는 기계학습 모델에 대한 가장 일반적 공격 중 하나이다.
모델 전도 공격 model inversion attack	모델 전도 공격은 인공지능 모델을 대상으로 이뤄지는 공격으로서 모델에 대한 접근 권한을 악용하여 학습 데이터에 대한 정보를 추론하는 공격을 말한다. 따라서 모델 전도 공격은 학습 데이터에서 기계학습 모델로의 일반적인 경로를 단방향에서 양방향으로 전환하여 다양한 정확도로 모델에서 학습 데이터를 추정할 수 있게 한다. 이러한 공격은 학습 데이터에 일반적으로 민감한 개인 정보가 포함되어 있다는 점에서 심각하다.[59]
모델 편향 model bias	모델 편향이란 인공지능 모델을 개발하는 과정에서 모델의 종류나 시스템의 목표에 따라 발생할 수 있는 편향이다. 모델 편향은 기계학습 편향, 알고리즘 편향, 또는 인공지능 편향bias in AI[60]이라고도 하는데, 알고리즘이 결과를 출력할 때 기계학습 절차상 가정에 오류가 있어서 구조적으로 편향성을 가진 결과를 출력하는 현상이다. 유럽위원회에서는 인공지능(또는 알고리즘) 편향을 임의의 특정 사용자 그룹을 다른 그룹보다 선호하는 것과 같이 불공정한 결과를 생성하는 컴퓨터 시스템의 체계적이고 반복될 수 있는 오류라 정의한다.[61]
인적 편향 human bias in AI	인적 편향이란 학습을 위한 데이터를 수집 및 가공 시 인적 요인에 의해 발생하는 오류이다. 이는 사람이 의식적 혹은 무의식적으로 특정 정보에 대해 편향되어 있다는 점에서 기인한다. 학습 데이터 수집 시 발생 가능한 편향을 확인해야 하며, 학습을 위한 특성을 선택하거나 데이터 라벨링 및 샘플링 시에도 인적 편향이 발생할 수 있다.
편향 제거 기법 bias removal methods	편향 제거 기법이란 인공지능 모델에 영향을 주는 편향을 없애는 방법을 말한다. 편향 완화 기법bias mitigation techniques이라고도 하는데, 편향 완화 기법은 이를 적용하는 단계에 따라 3가지 방식으로 나뉜다. 모델 학습 전에 적용해야 할 편향 완화 기법pre-processing bias mitigation, 모델 학습 중에 적용할 기법in-processing bias mitigation, 모델 학습 이후 적용할 기법post-processing bias mitigation이다.

59) European Commission, "**ALTAI – The Assessment List on Trustworthy Artificial Intelligence**," p. 28, 2020. 7.
60) ISO/IEC TR 24027, "**Bias in AI systems and AI aided decision making**," p. 7, 2021. 11.
61) European Commission, "**ALTAI – The Assessment List on Trustworthy Artificial Intelligence**," p. 33, 2020. 7.

02 용어표

용어명	정의
편향성 평가 **bias assessment**	편향성 평가란 인공지능에서 출력이 공정한지 또는 알고리즘 설계, 개발 또는 작동 등의 단계에서의 편향성에 대한 가치판단 과정이다. 편향이 항상 잘못이라고 말할 수 없다. 편향성 평가에는 편향의 영향, 편향이 발생할 가능성, 그리고 편향을 방지하거나 감지할 수 있는지 여부 등이 포함된다.
안전 모드 **safety mode**	안전 모드란 외부로부터의 공격, 인적 오류human error, 인공지능 모델의 성능 저하, 편향 발생으로 인한 사회적 물의, 사고 등이 예상되는 경우, 이의 발생 원인을 파악하고 해결하거나 사용자에게 정상적인 기능으로 복구하는 방법을 사용자에게 제시하는 대처 방법이 작동하는 상태를 의미한다.
오픈소스 라이브러리 **open source library**	컴퓨터 과학 분야에서 라이브러리library란 소프트웨어를 개발하는 프로그래머들이 참고할 수 있도록 컴파일해서 재사용할 수 있는 파일, 함수, 스크립트, 루틴, 그리고 그 외의 자원을 모아놓은 곳을 의미한다. 오픈소스 라이브러리open source library란 오픈소스 라이선스open source license를 가지고 있는 라이브러리를 의미한다.[62]
시뮬레이터 **simulator**	시뮬레이터는 시험testing에 사용되는 시스템으로써 일련의 입력값을 조정하여 실제 시스템과 유사한 결과를 확인하기 위해 사용된다. 개발 중인 인공지능을 실제 환경에서 운용할 경우, 시험 과정에서 인간의 삶과 건강, 재산, 환경 등에 해를 끼칠 수 있으므로 시뮬레이터의 사용이 중요하다.
테스트 오라클 **test oracle**	테스트 오라클이란 테스트를 수행한 결과가 참인지 거짓인지를 판단하기 위해서 미리 정의된 참값을 대입하여 비교하는 소프트웨어 테스트 기법 또는 활동이다. 일반적으로는 시험원이 시험 결과를 확인하지만, 테스트 케이스가 복잡한 계산을 필요로 하는 경우는 프로그램이 이를 대신한다.[63]
인공신경망 **ANN, Artificial Neural Network**	인공신경망은 수많은 노드가 상호 연결된 층으로 구성되는데, 연결된 각 노드는 입력을 처리하고 다른 뉴런으로 전송되는 출력을 생성하는 모델의 유형이다.[64] 가장 단순한 인공신경망은 퍼셉트론perceptron, a threshold logic unit으로, 기본적인 동작은 입력의 가중치 합계를 수행한 다음, 이 합계가 임계값을 초과하면 '1'을 출력하고, 그렇지 않으면 '0'을 출력하는 것이다.[65]
합성곱 신경망 **CNN, Convolutional Neural Network**	합성곱 신경망[66]이란 픽셀pixel들이 2차원 격자2D grid 형태로 배열된 데이터를 처리하는 데 특화된 신경망이다. 합성곱 신경망이라는 이름은 합성곱이라는 수학 연산을 사용하기 때문에 붙은 것이다. 합성곱은 특별한 종류의 선형linear 연산으로, 적어도 하나의 층에서 일반적인 행렬 곱셈 대신 합성곱을 사용하는 신경망이면 그 어떤 것이든 합성곱 신경망이라고 할 수 있다.[67]

62) CROS, **Open Source licensing features overview**, [Online], Available: https://ec.europa.eu/eurostat/cros/content/open-source-licensing-features-overview_en

63) 한국정보통신기술협회, **정보통신용어사전**, [Online], Available: https://terms.tta.or.kr/main.do

64) MIT, **Machine Learning, Explained**, [Online], Available: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

65) Kevin Gurney, "**An introduction to neural networks**," p. 19, 1997. 3.

66) I. Goodfellow, Y. Bengio, and A. Courville, "**Deep Learning**," p. 326, 2016. 11.

67) Stuart Russell, Peter Norvig, "**Artificial Intelligence: A Modern Approach, 4th ED**," p. 811, 2020. 4.

02 용어표

용어명	정의
순환 신경망 RNN, Recurrent Neural Network	순환 신경망에서는 은닉층hidden layer이 입력층과 이전 타임 스텝time step의 은닉층으로부터 정보를 받는다. 인접한 타임 스텝의 정보가 은닉층으로 흐르기 때문에 신경망이 이전 이벤트를 기억할 수 있다. 이 정보 흐름을 루프로 표시하는데, 그래프 표기법에서는 순환 에지recurrent edge로 부르기 때문에 RNN이 여기서 유래되었다.[68] 순환 신경망은 순차sequential 데이터를 처리하는 데 특화되었으며, 과거 정보를 기억하고 이를 기반으로 새로운 이벤트를 처리할 수 있다.[69]
적대적 생성신경망 GAN, Generative Adversarial Network	적대적 생성신경망은 주로 이미지 생성에 활용되며, 두 개의 인공신경망을 학습시켜서 활용하는 방법이다.[70] 하나는 이미지를 생성하는 생성generative 신경망, 다른 하나는 생성신경망이 만든 이미지를 진짜인지, 가짜인지 판별하는 판별discriminative 신경망이다. 이미지 생성 및 변환 기술은 인공신경망이 다양한 노이즈noise 입력을 받아 기존에 존재하지 않는 새로운 이미지를 생성해내거나 입력 이미지나 비디오를 다른 형태나 정보를 지닌 이미지 또는 비디오로 변환하는 기술이다.[71]
지도 학습 supervised learning	지도 학습 알고리즘이란[72] 입력 데이터와 출력 데이터의 학습 데이터셋이 주어졌을 때, 특정 입력 데이터와 특정 출력 데이터를 연관시키는 방법을 배우는 학습 알고리즘이다. 이때 '지도교사supervisor'가 라벨을 통해 참값ground truth을 제공하므로 '지도 학습'이라는 용어를 사용한다.
비지도 학습 unsupervised learning	비지도 학습은 샘플들에 라벨이 없어도 데이터 분포distribution로부터 정보를 최대한 추출하는 학습이다. 즉, 비지도 학습을 통해 라벨이 지정되지 않은 데이터에서 암시적인 패턴이나 추세를 찾을 수 있다.[73] 예를 들어, 라벨이 없는 온라인 판매 데이터를 살펴보고 구매하는 다양한 유형의 고객을 식별할 수 있다.[74] 비지도 학습은 밀도 추정, 데이터 분포에서 표본을 추출하는 방법의 학습, 데이터 분포에서 얻은 자료의 잡음 제거 방법 학습, 서로 연관된 객체들을 클러스터링clustering하는 등에 유용하게 사용된다.[75]
기계학습 machine learning	기계학습은 기계가 인간의 지능적인 행동을 모방할 수 있는 능력이라고 정의할 수 있다. 기계학습은 인공지능의 한 분야로, 명시적으로 프로그래밍하지 않고도 기계를 학습시킬 수 있는 능력을 부여하는 연구 분야이다.[76]

68) Sebastian Raschka, Vahid Mirjalili, "**Python Machine Learning-3rd**," p. 541, 2019. 12.
69) Sebastian Raschka, Vahid Mirjalili, "**Python Machine Learning-3rd**," p. 539, 2019. 12.
70) I. Goodfellow, J. Abadie, M. Mirza, and B. Xu, "**Generative Adversarial Nets**," arXiv:1406.2661v1, 2014. 6.
71) 조영주, 배강민, 박종열, "**GAN 적대적 생성 신경망과 이미지 생성 및 변환 기술 동향**," 전자통신동향분석 35권 제4호, pp. 91-102, 2020. 8.
72) Ian Goodfellow, Yoshua Bengio, and Aaron Courville, "**Deep Learning**," pp. 137-138, 2016. 11.
73) Ian Goodfellow, Yoshua Bengio, and Aaron Courville, "**Deep Learning**," pp. 142-144, 2016. 11.
74) MIT, **Machine Learning, Explained**, [Online], Available: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained, 2021. 4.
75) Ian Goodfellow, Yoshua Bengio, and Aaron Courville, "**Deep Learning**," pp. 142-144, 2016. 11.
76) MIT, **Machine Learning, Explained**, [Online], Available: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained, 2021. 4.

02 용어표

용어명	정의
심층학습 deep learning	심층학습은 많은 계층을 가지는 인공신경망에 데이터를 학습시키는 기계학습의 한 분야이다. 일반적으로 2~3개 층으로 되어 있는 신경망을 쉘로우 학습shallow learning이라고 하고 그 이상인 것을 심층학습deep learning이라고 한다.
강화학습 reinforcement learning	강화학습이란, 학습 에이전트가 특정 환경에서 선택 가능한 행동 중 보상을 최대화하는 행동을 선택하도록 하는 학습이다. 강화학습은 외부 환경의 지도에 의존하지 않고 환경과 직접적 상호 작용하는 에이전트에 의한 학습을 강조함으로써 다른 기계학습 접근법과 구별된다. 강화학습은 상태state, 행동action 및 보상reward 측면에서 학습 에이전트와 환경 간의 상호 작용을 정의하는 프레임워크를 사용한다.[77] 강화학습은 시행착오를 통해 기계가 보상 시스템을 구축하여 최상의 조치를 취하도록 학습한다.
전이학습 transfer learning	전이학습은 소규모 데이터 모집단으로 구성되는 그룹이 가진 부족한 데이터의 양적 문제를 완화하는 기술이다.[78][79] 전이학습은 하나의 작업에 대해 학습된 모델을 연관된 작업에 용도를 변경하여 적용하는 기계학습 기술이다. 예를 들어, 개와 고양이를 인식하도록 학습된 모델은 약간의 추가 학습만으로 개만을 인식하도록 용도를 변경할 수 있다. 전이학습은 하나의 문제를 해결하여 얻은 지식을 활용하여 새롭고 유사한 문제에 적용할 수 있어서 유용하다. 이렇게 하면 새 작업에서 우수한 성능 달성에 필요한 학습 데이터 및 계산 자원의 양을 크게 줄일 수 있다.
인과학습 causal AI	인과학습은 원인과 결과를 설명할 수 있는 인공지능 기술로 정의되며, 조직에서 의사 결정과 의사 결정 원인을 설명하는 데 사용되는 기술이다.[80] Microsoft에 따르면, 상관 패턴 인식correlational pattern recognition에 기반한 기계학습은 강력한 예측과 신뢰할 수 있는 의사결정에 불충분하기 때문에, 인과 추론causal reasoning 원칙에 기반한 인과학습을 새로운 기계학습 방식으로 제시하였다.
가중치 weight	기계학습에서 가중치는 인공신경망에서 뉴런간의 각 연결에 할당된 값으로, 신경망을 계산하는 동안 뉴런 간에 전송되는 신호의 강도와 방향을 결정하고 특정 작업에 대한 네트워크의 성능을 최적화하기 위해 학습 과정에서 조정되는 값이다. 각 노드는 일련의 입력값input, 가중치 및 편향bias으로 표현되며, 가중치와 편향은 인공신경망의 핵심 구성 요소로 네트워크의 출력을 결정하는 데 중요한 역할을 한다.

77) Richard S. Sutton and Andrew G. Barto. "**Reinforcement Learning: An Introduction**," pp. 15-16, 2018. 11.

78) ISO/IEC TR 24027, "**Bias in AI systems and AI aided decision making**," p. 25, 2021. 11.

79) R. Hee Jung, A. Hartwig, and M. Margaret, "**InclusiveFaceNet: Improving Face Attribute Detection with Race and Gender Diversity**," arXiv:1712.00193v3, 2018. 7.

80) SSIR, **The Case for Casual AI**, [Online], Available: https://ssir.org/articles/entry/the_case_for_causal_ai

02 용어표

용어명	정의
베이지안 결정이론 **bayesian decision theory**	베이지안 결정이론은 특정 사건에 대한 관측 전의 사전확률과 관측 후의 사후확률의 관계를 설명하는 확률이론으로 베이즈 정리Bayes' Theorem 또는 베이즈 규칙Bayes' Rule이라 부르기도 한다.[81] 베이지안 결정이론을 통계 전문 용어로 요약 정리하면 아래와 같다. $$P(H \mid E) = \frac{P(H)P(E \mid H)}{P(E)}$$ 여기서 H는 가정 또는 사건hypothesis or event, $P(H)$는 관찰되지 않은 잠재 변수 H에 대한 주관적인 사전확률priori probability, E는 증거evidence, $P(E)$는 증거 모델 또는 정규화 상수evidence model or normalized constant를 각각 나타낸다. $P(E\|H)$는 우도 또는 가능도likelihood라 하며, 이산형 확률 분포discrete probability distributions에서 잠재 변수 H의 값이 주어질 때의 E의 확률을 나타낸다. $P(H\|E)$는 증거evidence E가 주어질 때 사건hypothesis or event H가 발생할 사후확률posteriori probability을 가리킨다. 베이지안 결정이론에서는 주관적으로 예상한 사전확률 $P(H)$에 관찰된 데이터로 계산한 우도확률 $P(E\|H)$을 곱한 값에 이를 모든 가설에 대해 증거가 발생될 확률 $P(E)$로 나누어 추정 가설에 대한 사후확률$P(H\|E)$을 구한다. 베이즈 정리를 이용하면 우도확률에 의해 H와 E의 관계를 뒤집을 수도 있으므로 이를 확률론적 역수probabilistic inverse라고도 부른다.[82]
회귀/분류 **regression/classification**	회귀와 분류는 기계학습 유형 중 지도학습에 속한다.[83] 회귀와 분류는 입력변수(독립변수, 설명변수)와 목표변수(종속변수, 반응변수)의 값을 이용하여 주어진 입력변수에 대한 목표변수 값을 예측하는 모델을 개발한다는 점에서는 유사하지만, 회귀는 목표변수의 형태가 연속형인 데 반해 분류는 범주형이라는 점이 서로 다르다. 회귀 모델은 입력변수 x의 변화에 따라 목표변수 y가 어떻게 변화하는지를 설명하는 모델로 정의할 수 있다.[84] 기계학습에서는 회귀모델을 입력변수 x를 목표변수 y에 매핑mapping하는 목표함수objective function를 학습하는 모델로 정의하고 있다.[85] 회귀의 목표는 최소한의 오차로 입력 데이터에 적합fitting한 목표함수를 찾는 것이다. 분류 모델은 개체를 사전에 정의된 여러 범주 중 어느 하나에 할당하는 목표함수를 학습하는 모델이다.[86] 분류의 목표는 과거의 관측 데이터를 기반으로 새로운 샘플의 범주형 클래스를 예측하는 것이라 할 수 있다.[87]

81) Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, "**Introduction to DATA MINING-Second Edition**," pp. 214-215. 2018. 1.
82) Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong, "**Mathematics for Machine Learning**," pp. 164-165, 2020. 4.
83) Aurelien Geron, "**Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow 2nd Edition**," pp. 7-9, 2019. 10.
84) David S. Moore, George McCabe, William M. Duckworth, and Layth Alwan, "**The Practice of Business Statistics: Using Data for Decisions, 2nd Edition**," pp. 118-119, 2008. 2.
85) Pang-Ning Tan, Michael Steinbach, "**Introduction to Data Mining**," pp. 729-730, 2006. 3.
86) Pang-Ning Tan, Michael Steinbach, "**Introduction to Data Mining**," pp. 146-147, 2006. 3.
87) Sebastian Raschka, Vahid Mirhalili(박해선 옮김), "**머신 러닝 교과서 개정 3판**," p. 35, 2021. 3.

용어명	정의
범주형 데이터/수치형 데이터 **categorical data/ numerical data**	통계에서 데이터는 크게 두 가지 유형으로 분류할 수 있다. 하나는 범주형 데이터[88]이고, 다른 하나는 수치형 데이터이다. 범주형 데이터는 범주 또는 레이블로 구성된 데이터 유형을 말한다. 나이, 설문 결과, 학력, 성별, 국적 또는 제품 유형과 같은 특성을 나타내는 데 자주 사용된다. - 범주형 데이터는 순서ordinal형과 명목nominal형의 두 가지 유형으로 더 나눌 수 있다.[89] 순서 데이터는 서열 데이터라고도 하는데, 상, 중, 하와 같은 자연스러운 순서 또는 순위가 있는 범주를 나타낸다. 명목 데이터는 국적과 같이 순서가 없는 유형의 데이터를 말한다. 수치형 데이터는 양적 데이터quantitative data라고도 하는데 수로 구성된 데이터 유형을 말한다. - 수치형 데이터는 연속continuous형과 이산discrete형 두 가지 유형으로 더 나눌 수 있다. 연속 데이터는 0보다 크고 1보다 작은 실수 전체의 집합과 같이 특정 범위 내의 모든 값을 가질 수 있는 데이터이다. 이산형 데이터는 나이 또는 교실 안의 학생 수와 같이 특정 값만 취할 수 있는 데이터를 의미한다.[90] 요약하면 범주형 데이터는 특성 또는 레이블을 나타내고, 수치형 데이터는 수량 또는 값을 나타낸다.
합성곱 계층 **convolutional layer**	합성곱 계층은 합성곱 신경망CNN, Convolutional Neural Network에서 가장 중요한 구성 요소이며, 완전연결fully-connected 계층과는 달리 입력 이미지의 3차원 형상을 유지하는 것이 특징이다. 출력 또한 3차원 데이터로 출력하여 다음 계층으로 전달하기 때문에 합성곱 신경망에서는 이미지 데이터처럼 형상을 가지는 데이터를 제대로 학습할 가능성이 높다.
특성 벡터 **feature vector**	특성 벡터는 관찰된 현상의 수치적 속성을 순서대로 나열한 목록으로 기계학습 모델에 대한 입력 특성features을 나타낸다.[91] 기계학습 및 패턴인식에서 특성은 현상의 속성property, 특징characteristic 또는 변수이다.[92] 유익하고 구별되며 독립적인 특성을 선택하는 것은 분류 및 회귀에서 모델 성능 향상의 중요한 요소이다.[93]
외형 변환 **deformation**	외형 변환이란 굽힘bending, 늘리기stretching, 압착squeezing 등으로 형상을 변경하는 것을 말한다. 생성적 인공지능generative AI 영역 중 이미지 및 시각 예술 분야에서 창의적 활동에 3D 외형변환 기술을 활용한다.

88) Yale University, **Categorical Data,** [Online], Available: http://www.stat.yale.edu/Courses/1997-98/101/catdat.htm
89) Dan Yates, David S. Moore, Daren S. Starnes, "**The Practice of Statistics (2nd ed.)**," 2002. 3.
90) OECD, **Glossary of Statistics Terms**, [Online], Available: https://www.oecd-ilibrary.org/economics/oecd-glossary-of-statistical-terms_9789264055087-en
91) Iguazio, **What is a Featrue Vector?**, [Online], Available: https://www.iguazio.com/glossary/feature-vector/
92) Bishop Christopher, "**Pattern recognition and machine learning**," 2006. 8.
93) WIkipedia, **Feature**, [Online], Available: https://en.wikipedia.org/wiki/Feature_(machine_learning)

02 용어표

용어명	정의
트랜스포머 transformer	트랜스포머는 어텐션 메커니즘attention mechanism을 채택한 심층학습 모델이다. 변환기라고도 하는데 입력 데이터 각 부분의 중요성에 차등 가중치를 부여한다. 변환기 모델transformer model은 문장의 단어와 같은 순차적 데이터의 관계를 추적하여 맥락과 의미를 학습하는 신경망이고, 합성곱 신경망이나 순환 신경망을 기반으로 구성한 모델과 다르게 단순히 어텐션 구조만으로 전체 모델을 만들어 어텐션 기법의 중요성을 돋보이게 하는 모델이다. 주로 자연어 처리NLP, Natural Language Processing[94]나 컴퓨터 비전[95] 분야에 사용한다.
이미징 imaging	이미징 또는 영상분석imaging analysis이란 기존의 이미지 관련 의료기기에 판독과 진단의 기능을 부가한 것을 의미한다. 이미징은 크게 진단방사선 이미지, 심혈관 진단, 유방진단, 폐·뇌 등을 장비를 통해 판단하는 영역으로 구분된다. 이때 인공지능 활용 기술 분야는 ① 의료진이 육안으로 판독하기 힘든 영역의 표식을 찾아내는 정밀 판독reading과 ② 기존의 판독자료와 질환 관계 등을 분석하여 해당 이미지를 통한 질환 진단diagnosis이다.[96]
버트 BERT, Bidirectional Encoder Representations from Transformers	BERT는 사전 학습pre-training을 위한 트랜스포머 기반 기계학습 기술을 활용한 Google이 개발한 자연어 처리 모델이다. 2019년 Google은 검색 엔진에서 버트를 활용하기 시작했으며 2020년 후반에는 거의 모든 영어 쿼리에서 BERT를 사용한다고 발표했다.[97]
바트 BART, Bidirectional and Auto-Regressive Transformers	BART는[98] 입력 데이터에 노이즈를 추가하여 이를 다시 원문으로 복구하는 자동 인코더의 형태로 학습된다. 표준 트랜스포머 기반 신경 기계번역 구조를 사용하기 때문에 매우 단순하다. BART는 BERT와 GPT 모델의 구조적 특징을 결합한 모델로, 일반적인 sequence-to-sequence 아키텍처를 가진다.
AIOps Artificial Intelligence for IT Operations	AIOps는 빅 데이터와 기계학습을 결합하여 이벤트 상관관계, 이상 탐지 및 인과관계 결정을 포함한 IT 운영 프로세스의 자동화를 의미한다.[99] 즉, AIOps를 사용하여 IT 운영 관리를 단순화하고 복잡한 최신 IT 환경에서 문제 해결을 가속화하고 자동화할 수 있다.[100]

94) Vaswani Ashish, Shazeer Noam, Parmar Niki, "**Attention Is All You Need**," arXiv:1706.03762v5, 2017. 12.

95) TowardsDataScience, **Transformer in CV**, [Online], Available: https://towardsdatascience.com/transformer-in-cv-bbdb58bf335e

96) IBM, **How is artificial intelligence used in medicine?**, [Online], Available: https://www.ibm.com/topics/artificial-intelligence-medicine

97) Anna Rogers, Olga Kovaleva, Anna Rumshisky, "**A Primer in BERTology: What We Know About How BERT Works**," Transactions of the Association for Computational Linguistics, arXiv:2002.12327v3, 2020. 9.

98) Mike Lewis, Yinhan Liu, Naman Goyal, "**BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension**," arXiv:1910.13461v1, 2019. 10.

99) Gartner Glossary, **Information Technology Glossary**, [Online], Available: https://www.gartner.com/en/information-technology/glossary/

100) IBM, **AIOps**, [Online], Available: https://www.ibm.com/kr-ko/cloud/learn/aiops#toc-aiops-7HhM1Vpc

02 용어표

용어명	정의
GPT-3 **Generative Pre-Trained Transformer-3**	GPT는 주어진 텍스트 기반 입력을 했을 때 인간과 유사한 텍스트를 생성할 수 있는 심층학습 기술을 사용하는 언어 모델이다. 사용자가 모델에 문장을 입력하면 트랜스포머는 공개적으로 사용할 수 있는 데이터셋에서 추출된 일관된 문단 기반 정보를 생성한다. GPT-3는 OpenAI[101]가 GPT-2의 후속 모델로 발표한 비지도 트랜스포머 언어 모델unsupervised transformer language model이다.[102][103][104] OpenAI는 GPT-3 정식 버전에 1,750억 개의 매개변수가 포함되어 있다고 밝혔는데, 이는 GPT-2 정식 버전에서 사용한 15억 개의 매개변수보다 116배 많은 수다.
챗봇 **chatbot**	챗봇chatbot은 채팅chatting과 로봇robot이 결합된 용어이며, 사람이 입력한 질문을 인식하고, 그에 알맞는 응답을 제공하는 소프트웨어이다. 음성 명령이나 텍스트 채팅을 통해 사람이 사용하는 언어로 대화를 시뮬레이션할 수 있다.[105] 챗봇은 전자상거래, 은행 등 다양한 분야에서 고객 지원이나 정보 습득과 같은 영역에 활용된다. 인공지능 챗봇은 기계학습, 자연어 처리, 자동화된 규칙과 빅데이터 분석을 바탕으로 사람이 소통하는 방식으로 대화한다.[106] 특히, chat과 GPT의 합성어인 ChatGPT는 OpenAI가 2022년 12월 1일 공개한 초거대 인공지능 기반 프로토타입 대화형 인공지능 챗봇이다. ChatGPT는 기존 챗봇과 달리 정해진 답을 내 놓는게 아니라 사람이 묻는 질문에 알맞는 대답을 생성한다.[107]
복합 인공지능 **composite AI**	복합 인공지능은 서로 다른 인공지능(시각지능, 청각지능, 언어지능, 촉각 지능 등)과 관련된 여러 기술들(기계학습, 심층학습, 자연어 처리, 컴퓨터 비전 등)을 결합하여 학습의 효율성을 향상시켜, 다양한 비즈니스 문제들을 해결해 주는 인공지능이다.[108][109]

101) **OpenAI**, [Online], Available: https://openai.com/

102) Alec Radford, Jeffrey Wu, Rewon Child, "**Language Models are Unsupervised Multitask Learners**," 2019. 2.

103) Analytics India Magazine, **OpenAI Releases GPT-3, The Largest Model So Far**, [Online], Available: https://analyticsindiamag.com/open-ai-gpt-3-language-model/

104) T. B. Brown, B. Mann, and N. Ryder, "**Language Models are Few-Shot Learners**," arXiv:2005.14165v4, 2020. 7.

105) ScienceDirect, **An experimental study of public trust in AI chatbots in the public sector**, [Online], Available: https://doi.org/10.1016/j.giq.2020.101490

106) Kiran ramesh, "**A Survey of Design Techniques for Conversational Agents**," CCIS, vol. 750, 2017. 10.

107) Huggingface, **OpenAI GPT**, [Online], Available: https://huggingface.co/transformers/v2.5.0/model_doc/gpt.html

108) Gartner, **Building a Digital Future: Emergent AI Trends**, [Online], Available: https://www.gartner.com/en/documents/4014200

109) AIMultiple, **What is Composite AI & Why is it important in 2023?**, [Online], Available: https://research.aimultiple.com/composite-ai/

02 용어표

용어명	정의
엣지 인공지능 **edge AI**	엣지 인공지능이란 엣지 서버에서 이루어지는 인공지능 연산을 뜻한다. 즉, 네트워크 가장자리(엣지)에 위치한 단말에서 발생한 데이터를 멀리 떨어져 있는 중앙 서버가 아니라, 단말 자체 혹은 단말과 가까운 엣지 서버에서 처리하는 인공지능 연산 기술이다. 따라서, 개인의 데이터를 중앙 서버까지 멀리 보내지 않고 가까운 로컬 서버에서 처리함으로써, 대기 시간이 단축될 뿐 아니라 네트워크 비용이 감소하고 개인 정보 유출 위험성 또한 줄어든다.[110] 따라서, 엣지 인공지능은 인공지능 모델 학습이 중앙 서버에서 실행되는 방식에 비해, 도입 용도, 동작 방식 및 활용 사례 등이 차별화된다.

110) Hewlett Packard Enterprize, **엣지 AI란?**, [Online], Available: https://www.hpe.com/kr/ko/what-is/edge-ai.html

03 요구사항별 이해관계자

요구사항별 이해관계자

관련 표준*에 근거한 요구사항별 이해관계자

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

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

04 이해관계자 정의

이해관계자 정의

IT분야역량체계ITSQF에서 제시한 대표 이해관계자·협력 대상의 직업·직무 정의

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

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

05 참고문헌

참고문헌

- 120다산콜재단, **챗봇 상담 서울톡**, [Online], Available: https://www.120dasan.or.kr/dsnc/main/contents.do?menuNo=200019
- A. Bendale and T. E. Boult, "**Towards open set deep networks**," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1563-1572, 2016. 6.
- A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "**Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation**," Journal of Computational and Graphical Statistics, vol. 24, no. 1, pp. 44-65, 2015. 3.
- AI Verify Foundation, "**AI Verify Testing Framework**," 2023. 6.
- Berkman Klein Center, "**Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI**," Research Publication No. 2020-1, 2020. 1.
- Brad Smith, "**Tools and Weapons: The Promise and the Peril of the Digital Age Hardcover**," 2019.11.
- Council of the European Union, "**Artificial intelligence act, Council's General Approach**," 2022. 12.
- D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits, "**Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment**," arXiv:1907.11932v6, 2020. 8.
- ETSI GR SAI 005 V1.1.1, "**Securing Artificial Intelligence (SAI 005); Mitigation Strategy Report**," 2021. 3.
- European Commission, "**ALTAI - The Assessment List on Trustworthy Artificial Intelligence**," 2020. 7.
- F. Prost, H. Qian, Q. Chen, Ed. H. Chi, J. Chen, and A. Beutel, "**Toward a Better Trade-Off between Performance and Fairness with Kernel-based Distribution Matching**," "ML with Guarantees" workshop at 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), 2019. 12.
- G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K. R. Müller, "**Layer-wise relevance propagation: an overview**," Explainable AI: interpreting, explaining and visualizing deep learning, pp. 193-209. 2019. 9.
- Google Cloud, **Using the What-If Tool**, [Online], Available: https://cloud.google.com/ai-platform/prediction/docs/using-what-if-tool
- Google, **People + AI Guidebook - Explainability + Trust**, [Online], Available: https://pair.withgoogle.com/chapter/explainability-trust/
- Google, **People + AI Research - Patterns**, [Online], Available: https://pair.withgoogle.com/guidebook/patterns/how-do-i-calibrate-user-trust
- Google, **Responsible AI Practices - Google AI**, [Online], Available: https://ai.google/responsibilities/responsible-ai-practices/?category=interpretability
- H. Liang, E. He, Y. Zhao, Z. Jia, and H. Li, "**Adversarial attack and defense: A survey**," Electronics, 2022. 4.

05 참고문헌

- H. Zheng, Q. Ye, H. Hu, C. Fang, and J. Shi, "**BDPL: A Boundary Differentially Private Layer Against Machine Learning Model Extraction Attacks**," European Symposium on Research in Computer Security, pp. 66-83, 2019. 9.
- IMDA, PDPC, "**Model Artificial Intelligence Governance Framework 2nd**," [Online], Available: https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf
- ISO/IEC 23894, "**Artificial Intelligence - Risk Management**," 2023. 2.
- ISO/IEC 38507, "**Governance implications of the use of Artificial Intelligence by organizations**," 2020. 4.
- ISO/IEC TR 24027, "**Bias in AI systems and AI aided decision making**," 2021. 9.
- ISO/IEC TR 24028, "**Overview of trustworthiness in artificial intelligence**," 2020. 5.
- J. Adebayo, and M. Gorelick, **FairML: Auditing Black-Box Predictive Models**, [Online], Available: https://github.com/adebayoj/fairml
- J. Baek, D. B. Lee, and S. J. Hwang, "**Learning to Extrapolate Knowledge: Transductive Few-shot Out-of-Graph Link Prediction**," 34th Conference on Neural Information Processing Systems (Neur IPS 2020), 2020. 10.
- K. Lerman, and T. Hogg, "**Leveraging position bias to improve peer recommendation**," PloS one, vol. 9, no. 6, 2014. 6.
- LG, **'AI 윤리 점검 TF' 신설...인간존중·공정성 등 5대 핵심가치로**, [Online], Available: https://biz.sbs.co.kr/article/20000077506
- M. Maadi, H. A. Khorshidi, and U. Aickelin, "**A Review on Human-AI Interaction in Machine Learning and Insights for Medical Applications**," International Journal of Environmental Research and Public Health, vol. 18, no. 4, p. 2121, 2021. 2.
- M. Soll, T. Hinz, S. Magg, and S. Wermter, "**Evaluating Defensive Distillation for Defending Text Processing Neural Networks Against Adversarial Examples**," International Conference on Artificial Neural Networks, Springer International Publishing, 2019. 9.
- Microsoft, "**Microsoft Responsible AI Standard v2**," 2022.6.
- N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "**A survey on bias and fairness in machine learning**," ACM Computing Surveys (CSUR), vol. 5, no. 6, pp. 1-35, 2021. 7.
- NAVER, **AI 윤리 준칙**, [Online], Available: https://www.navercorp.com/value/aiCodeEthics
- R. Baeza-Yates, "**Bias on the Web**," Communication of the ACM, vol. 61, no. 6, pp. 54-61, 2018. 6.
- R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, "**AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias**," arXiv, 2018. 10.

05 참고문헌

- S. Lundberg and S. Lee, "**A Unified Approach to Interpreting Model Predictions**," NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017. 12.
- S. Vasudevan and K. Kenthapadi, "**LiFT: A Scalable Framework for Measuring Fairness in ML Applications**," Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM '20), pp. 2773-2780, 2020. 10.
- T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "**A Simple Framework for Contrasive Learning of Visual Representations**," Proceedings of the 37th International Conference on Machine Learning, pp. 1597-1607, 2020. 7.
- Tasq.ai, **Automate Data Labeling: What is It, and How Can Implementing It Help?**, [Online], Available: https://www.tasq.ai/blog/automate-data-labeling/
- The White House, "**Blueprint for an AI Bill of Rights**," 2022.10.
- The White House, "**Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence**," 2023. 10.
- The White House, "**Voluntary commitments - underscoring safety, security, and trust - mark a critical step toward developing responsible AI**," 2023. 7.
- 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
- World Economic Forum, "**Companion to the Model AI Governance Framework**," 2020. 1.
- World Health Organization, "**Ethics and Governance of Artificial Intelligence for Health: WHO Guidance**," 2021. 6.
- Y. Gal and Z. Ghahramani, "**Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning**," Proceedings of The 33rd International Conference on Machine Learning, vol. 48, pp. 1050-1059, 2016. 6.
- Y. Gong, D. Yan, T. Mao, D. Wang, and R. Wang, "**Defending and Detecting Audio Adversarial Example using Frame Offsets**," KSII Transactions on Internet and Information Systems, vol. 15, no. 4, 2021. 4.
- YouTube, **YouTube 작동의 원리 - 제품 기능, 책임 및 영향력**, [Online], Available: https://www.youtube.com/intl/ALL_kr/howyoutubeworks/
- 과학기술정보통신부·정보통신산업진흥원, "**소프트웨어사업 요구사항 분석·적용 가이드**," 2021. 1.
- 김휘영, 정대철, 최병욱, "**딥러닝 기반 의료 영상 인공지능 모델의 취약성: 적대적 공격**," 대한영상의학회지, vol. 80, no. 2, pp. 259-273, 2019. 3.
- 서울특별시, **서울시 유니버설디자인 통합 가이드라인**, [Online], Available: https://opengov.seoul.go.kr/anspruch/16856750

05 참고문헌

- 소프트웨어정책연구소, **설명가능한 인공지능(Explainable AI; XAI) 연구 동향과 시사점**, [Online], Available: https://spri.kr/posts/view/23296?code=industry_trend
- 정보통신산업진흥원, "**기업 공개소프트웨어 거버넌스 가이드**," NIPA 오픈소스 가이드, 2021. 11.
- 카카오**, 기술윤리 위원회**, [Online], Available: https://www.kakaocorp.com/page/detail/9780
- 한국데이터베이스진흥원, "**데이터 품질진단 절차 및 기법 (Ver 1.0)**," 데이터 품질관리 시리즈 4, 2009. 10.
- 한국정보통신기술협회, "**인공지능 시스템 신뢰성 제고를 위한 요구사항(TTA.KO-10.1497)**," 정보통신단체표준, 2023. 12.
- 한국정보통신기술협회, "**지도학습을 위한 데이터 품질 관리 요구사항(TTA.KO-10.1339)**," 정보통신단체표준, 2021. 12.

찾아보기

▼ 생명주기별 요구사항 분류

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

▼ 속성별 요구사항 분류

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

▼ 신뢰성 확보 대상별 요구사항 분류

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

2024

신뢰할 수 있는 인공지능 개발 안내서 일반 분야

한국정보통신기술협회 신준호 단장
곽준호 팀장
김송이 책임
채희문 책임
조경우 책임
황재영 책임
신예진 책임
변은영 선임
오상훈 선임
강상연 전임

인 쇄 2024년 2월
발 행 2024년 2월
발 행 처 한국정보통신기술협회
발 행 인 손승현
편집·제작 ㈜디자인여백플러스
ISBN 979-11-89545-62-8