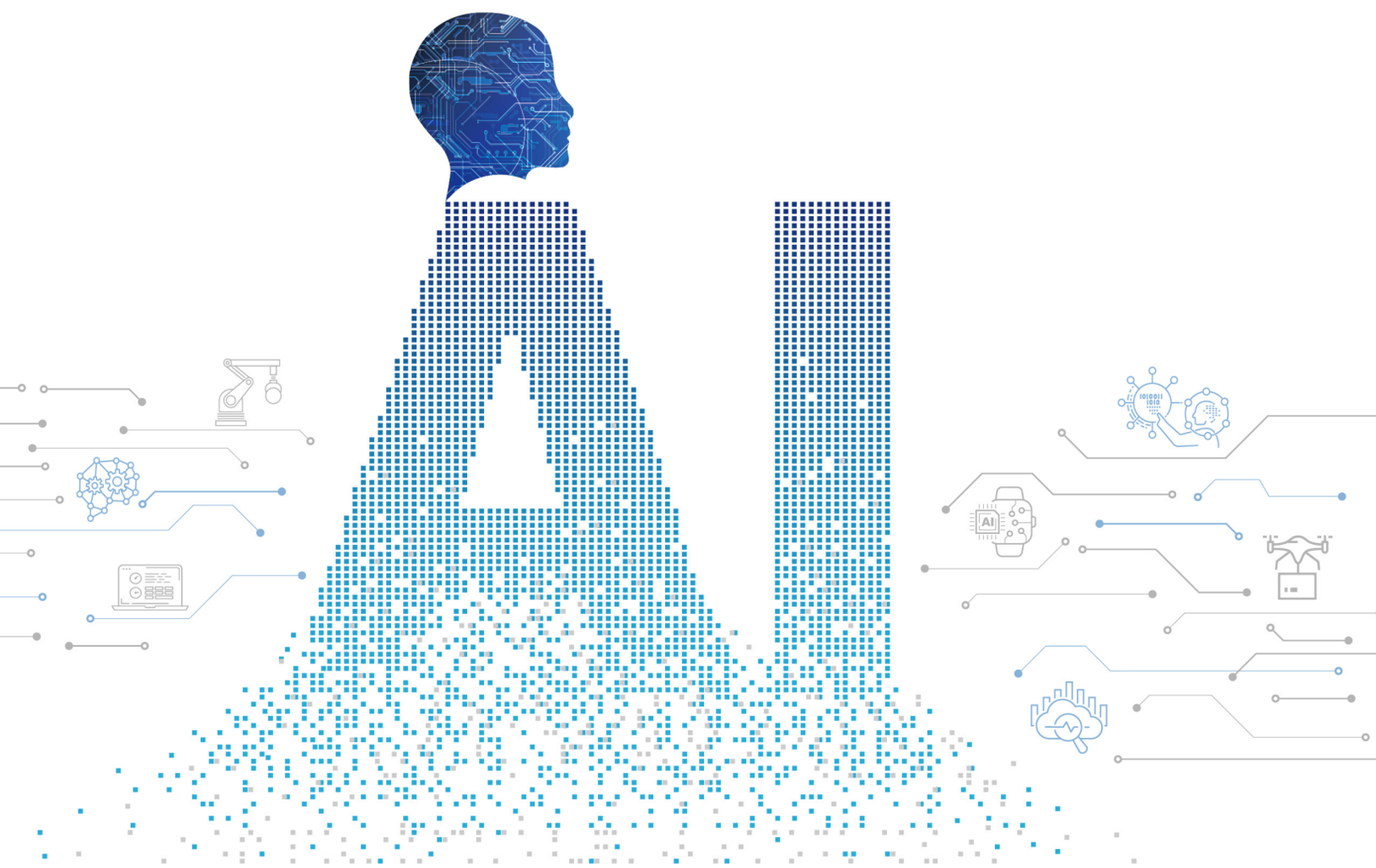
2022

신뢰할 수 있는 인공지능 개발 안내서(안)



과학기술정보통신부
Ministry of Science and ICT

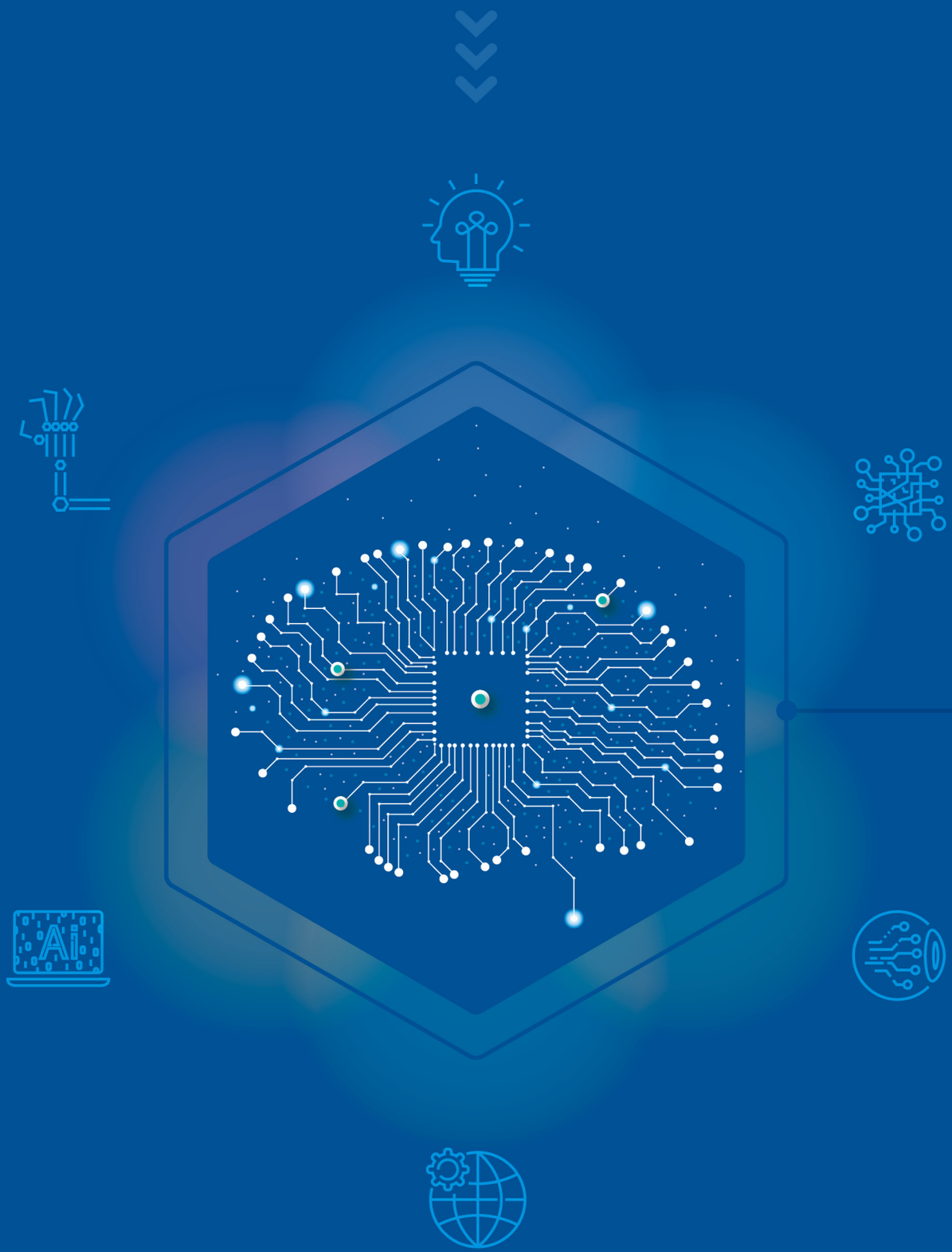
TTA 한국정보통신기술협회
Telecommunications Technology Association

일러두기

- 본 안내서는 과학기술정보통신부 「인공지능 신뢰성 기반조성」 사업의 연구 결과로서 내용의 무단전재를 금합니다.
- 아울러, 안내서의 내용을 가공·인용하는 경우에는 반드시 '과학기술정보통신부·한국정보통신기술협회 「2022 신뢰할 수 있는 인공지능 개발 안내서(안)」'의 출처를 밝혀주시기를 바랍니다.
- 본 안내서는 인공지능 서비스 및 제품을 개발하는 경우, 참고 자료로서의 활용을 위해 편찬되었습니다. 따라서, 본 안내서의 내용 중 업무 환경과 상황, 그리고 개발 목적을 고려하여 필요한 내용을 선택하여 활용하시기 바랍니다.
- 본 안내서의 인공지능 동향 및 기술 정보는 2021년 11월 기준으로 서술되었습니다. 추후 인공지능 기술을 활용한 다양한 서비스와 제품 영역별로 안내서를 고도화할 예정이며, 본 안내서에 소개된 기술이나 방법은 최신 연구 결과 및 현실적인 적용방안 등을 검토하여 지속적으로 수정할 예정입니다. 또한 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 의견수렴 과정을 계획 중에 있사오니, 많은 참여와 관심 부탁드립니다.

CONTENTS

2022 **신뢰할 수 있는 인공지능** 개발 안내서(안)



PART 1

개요

1 개요

01 발간 배경 및 목적

현재 인공지능AI, Artificial Intelligence 기술은 다양한 분야에 활용된다. 인공지능의 활용분야는 게임처럼 전통적으로 활용되던 분야는 물론, 스피커를 이용한 음성인식 서비스나 개인화된 비서 서비스와 같은 간단한 분야부터 검진 및 질병 진단, 자산관리를 비롯한 금융서비스, 자동차나 드론의 자율주행과 같은 복잡한 분야까지 폭넓게 아우른다. 이처럼 인공지능의 활용분야가 넓어지고 일상에 점점 많은 영향을 미치면서 인공지능의 신뢰성Trustworthiness 확보가 중요한 과제로 떠올랐다. 인공지능이 고도화될수록 그 작동 원리나 메커니즘을 파악하기는 어려워지는 데 비해, 인공지능의 활용빈도가 늘어나 데이터가 많아질수록 데이터의 오염이나 편향성과 같은 문제로 인해 인공지능이 오류를 범할 가능성은 커지기 때문이다. 특히 사람의 생명이나 공공안전과 직접 연관된 분야에까지 활용이 확대되면서 신뢰성의 중요성은 더 커졌다고 할 수 있다.

이처럼 인공지능의 신뢰성이 전세계적인 관심사로 부상하면서 국제적으로 다양한 대응방안이 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 작업자가 스스로 검증할 수 있게 한 검토목록인 "The Assessment List For Trustworthy Artificial IntelligenceALTAI for self assessment('20.07)"를 공표했다. 국내에서도 이에 발맞춰 사람 중심의 인공지능을 실현한다는 목표로 "인공지능(AI) 윤리기준('20.12)"을 선포했다.

그러나 앞서 언급한 것을 포함하여 지금까지 나온 인공지능 신뢰성 원칙 및 표준은 주로 윤리적 관점에서 추상적인 항목을 제시하고 있어 실무 현장에서 활용하기는 어렵다. 따라서 인공지능 서비스 개발자들이 별도의 고민이나 추론 과정 없이 실무에 그대로 적용할 수 있을 만큼 구체적으로 정리된 검토 목록이 필요하다. 특히 인력과 연구 개발 투자 여력이 제한적인 중소기업은 여건상 연구개발 인력이 직접 신뢰성 요구사항과 검증방법을 고안해 적용하기 어렵다.

신뢰할 수 있는 인공지능 개발 안내서는 이러한 문제점을 해결하고자 기획됐다. 미국, 유럽 등 주요 선진국과 국제 기구들에서 발표한 권고안 및 가이드를 참고, 자율적으로 점검 가능한 16개의 개발 요구사항과 59개 정성적·정량적 검증 항목을 제시한다.

개발자 및 기획자와 같은 인공지능 서비스 개발 실무자들은 본 개발 안내서에 제시된 항목들을 그대로 적용해 최소한의 신뢰성을 확보하는 한편, 신뢰성을 확보하려면 무엇이 중요한지 이해할 수 있을 것이다. 나아가 개발 안내서의 내용을 바탕으로 현장에 적합한 항목과 검증방법을 고안하고 활용함으로써 신뢰성 높은 인공지능 서비스를 개발할 수 있을 것이다. 본 개발 안내서를 통해 우리나라 인공지능 관련산업 기업 및 기관들이 보다 더 성숙한 인공지능 기술을 확보하는 데에 일조하고, 글로벌 경쟁력을 가질 수 있는 튼실한 기반이 되었으면 하는 바람이다.

02 인공지능 신뢰성 동향

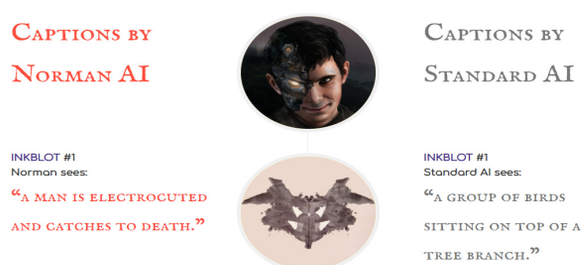
현재 세계의 주요 국가 정부와 표준 관련 기관, 기술단체들은 인공지능의 신뢰성을 확보하기 위해 각자의 상황에 맞는 방안을 제시하고 있다. 본 절에서는 인공지능이 폭넓게 활용되면서 발생하는 문제점을 알아보고 이에 따른 인공지능 신뢰성의 기본 개념, 국내외에서 진행 중인 관련 정책 및 연구 동향을 알아본다.

2.1. 인공지능 확산에 따른 문제점

인터넷이나 스마트폰과 같은 신기술로 인해 우리의 일상과 사회가 변화하는 동시에 새로운 문제가 등장했듯, 사회와 산업 전반에 걸쳐 다양한 분야에 인공지능이 활용되면서 새로운 위협이 등장했다. 이러한 문제 상당수는 사회적, 윤리적인 문제이므로 기술을 개선하거나 신기술을 도입하는 것만으로는 해결할 수 없다. 기술이 일상 곳곳에 영향을 줄수록 우리가 평소 의식하지 못하던 윤리적, 사회적 문제를 기술에 맞게 재정립해야 하기 때문이다. 실제로 현재 사용자가 인공지능 기술을 악의적인 목적으로 활용하거나 인공지능이 반사회·반인륜적인 판단을 하는 사고가 발생하고 있다.

인공지능 사고사례

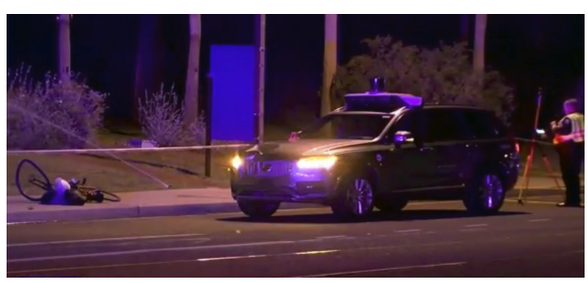
〈사고사례 1: 싸이코패스 인공지능〉



MIT에서는 일부러 반사회적·반인륜적인 데이터셋을 사용하여 훈련된 "Norman"을 개발, 그림을 이용한 심리 검사인 로르샤 테스트에서 제시된 그림을 사람이 감전되어 죽은 형상으로 이해하는 등 부정적인 인식 결과 도출('18.6.)

시사점
편향된 데이터로 학습한 인공지능은 사회적으로 수용하기 어려울 정도로 편향된 모델을 형성할 수 있음

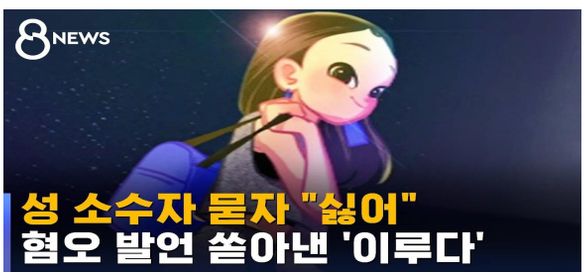
〈사고사례 2: 자율주행차 사망 사고〉



우버 자율주행차가 무단 횡단 중인 보행자를 치어 숨지게 한 사건이며, Lidar 센서를 통해 보행자를 감지하였으나 운전 효율에 우선순위를 두어 보행자를 무시하고 달려도 되는 도로 위 장애물로 판단('18.5.)

시사점
인공지능은 정확하고 효율적으로 작동했으나 사람의 안전이 무엇보다 우선이라는 윤리적 판단이 결여된 결과 인명사고를 초래

〈사고사례 3: 인공지능 챗봇 논란〉



SCATTER LAB이 개발한 주제 대화형 챗봇인 '이루다'는 20대 여성을 모사한 설정으로 공개하였으나 성 소수자 혐오 발언, 개인정보 침해 및 일부 사용자와의 성적인 대화로 논란('21.1.)

시사점
데이터셋 확보의 윤리성과 학습 데이터의 편향성으로 인해 사회적 물의를 일으킴

〈사고사례 4: 딥페이크 영상〉

딥러닝 방식을 이용, 얼굴과 목소리를 합성하여 버락 오바마 전 美 대통령이 트럼프 현 대통령을 욕하는 것처럼 꾸민 영상 공개 ('18.7.)

시사점
가짜 뉴스 및 조작 영상을 제작하여 사회적 혼란 및 물의 유발

2.2. 인공지능 신뢰성 개념

앞서 사례에서 살펴봤듯 인공지능 기술 프로젝트는 단지 '구현할 수 있는가?'라는 기술적 측면뿐 아니라 '이 프로젝트가 존재해도 괜찮은가?'라는 윤리적, 사회적 측면에서도 검토해야 한다. 특히 인공지능이 다양한 분야에 활용되면서 인공지능 시스템과 학습모델에 윤리적인 결함이 있는데도 이를 인지하지 못한 채 사용될 경우 매우 큰 파급효과를 낳을 수 있다. '인공지능 신뢰성'이란 데이터 및 모델의 편향, '블랙박스'적 특성처럼 인공지능 기술에 내재한 위험과 한계를 해결하고, 인공지능을 활용하고 확산하는 과정에서 부작용을 방지하기 위해 준수해야 하는 가치 기준을 말한다. 인공지능 신뢰성을 확보하는 데 필수적인 요소가 무엇인지는 주요 기구를 중심으로 논의가 이루어지고 있다. 일반적으로 안전성, 설명가능성, 투명성, 견고성, 공평성 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 악영향을 줄 가능성이 완화 및 제거된 상태
설명가능성Explainability	인공지능의 판단·예측의 근거와 결과에 이르는 과정이 사람이 이해할 수 있는 방식으로 제시되거나, 문제 발생 시 문제에 이르게 한 결과를 추적하여 도출할 수 있는 상태
투명성Transparency	인공지능의 안전성·설명가능성·견고성·공평성과 같은 주요 속성들의 근거와 동작 과정이 보편적 합리성에 부합하는 정도
견고성Robustness	인공지능이 외부의 간섭이나 극한적인 운영 환경 등에서도 사용자가 의도한 수준의 성능 및 기능을 유지하는 상태
공평성Fairness	인공지능이 데이터를 처리하는 과정에서 특정 그룹에 대한 차별이나 편향성을 나타내거나, 차별 및 편향을 포함하는 결론을 도출하지 않는 기능성

※ 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: '20)
- (경제협력개발기구) 지속가능한 사회와 인간 중심의 가치에 부합하고 투명성 및 설명가능성, 견고성 및 안전성을 갖춘 인공지능('19)
- (美 국립표준연구소, NIST) 인공지능이 사회 편익, 경제 성장을 위해 활용될 경우 반드시 만족시켜야 하는 목표이며, 설명가능성, 안전성, 보안성 등을 포함하는 개념('20)
- (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 주요 국가들은 인공지능의 신뢰성 확보가 인공지능의 사회적·산업적 수용과 발전의 전제 조건이라 여기고 신뢰성 확보 정책을 추진하고 있다. 또한 산업계 및 학계에서도 관련 기술 개발을 중심으로 신뢰성 확보를 위한 연구가 활발하다. 구체적으로 유럽위원회, 미국 등 주요국 정부 차원에서는 인공지능 신뢰성을 확보하는 데 필요한 정책과 규범을 본격적으로 마련하고 있으며, 이와 함께 국가 차원 인공지능 전략의 핵심 요소로 Trustworthy AI, Safe AI 등을 명시했다. 또한 인공지능 신뢰성 확보를 위한 가이드라인을 마련하여 민간 부문에서 자율적으로 인공지능의 신뢰성을 점검하고 확보할 수 있는 환경을 조성하고자 노력하고 있다. 기술 분야에서는 미국, 유럽 등 주요국의 학계와 글로벌 기업이 인공지능 신뢰성 확보에 필요한 제반 기술을 개발중이다. 우리나라 역시 최근 "인공지능(AI) 윤리기준('20.12)" "신뢰할 수 있는 인공지능 실현 전략('21.5)"을 발표하며 정책 및 연구개발 양면에서 세계적인 움직임에 동참하고 있다.

주요국 인공지능 신뢰성 관련 정책 동향

국가	주요 정책(연도)	특징
유럽위원회	• 신뢰할 수 있는 인공지능 윤리 가이드라인('19) • 인공지능 협력 선언('18)	인간중심의 가치, 윤리, 보안 등 균형 잡힌 인공지능 정책 추진 지향
미국	• 인공지능 애플리케이션 규제에 관한 가이드('20) • 자율주행 관련 가이드라인('16~'18) • 오바마 정부 인공지능 3부작 보고서('16)	산업 분야별 인공지능 활용·촉진을 위한 인공지능 기술개발 지원과 규제 완화 정책에 중점
중국	• 차세대 인공지능 발전계획('17)	정부 주도의 대규모 투자와 강력한 인력양성, 데이터 개방·공유 등 기업친화적 정책 추진
일본	• 인공지능 활용전략('19) • 인간중심의 인공지능 사회 원칙('18) • 인공지능 개발 가이드라인('17)	경제, 산업, 사회, 윤리 등의 관점에서 포괄적 접근
한국	• 사람이 중심이 되는 인공지능 윤리 기준('20) • 「인공지능 국가전략」('19)	사람 중심 인공지능을 기본가치로 인공지능 생태계 구축·인재양성·산업확산·역기능방지 등 종합 정책 추진

해외 주요 산·학·연 인공지능 신뢰성 연구 동향

기관명	활동 및 내용
방위고등연구계획국DARPA	Assured Autonomy, eXplainable AI 등의 프로젝트를 통해 인공지능 시스템의 안전성, 신뢰성 및 설명가능성 연구 진행 중
스탠포드 대학	인공지능 안전성 보장을 위한 정형 검증 기법, 학습 및 제어의 안전성 보장, 투명성 확보 방법 연구
IBM	'Trusting AI'를 모토로 공평성, 가치 정렬, 강건성, 설명 가능성, 투명성 및 책임성 5가지 원칙을 발표하고 관련 측정 평가 도구
마이크로소프트	책임 있는 인공지능 개발과 이를 통한 서비스 제공을 위해 공평성, 신뢰성 및 안전성, 개인정보보호 및 보안성 등 6가지 인공지능 개발 원칙 정의

03 신뢰할 수 있는 인공지능 개발 안내서 개발과정

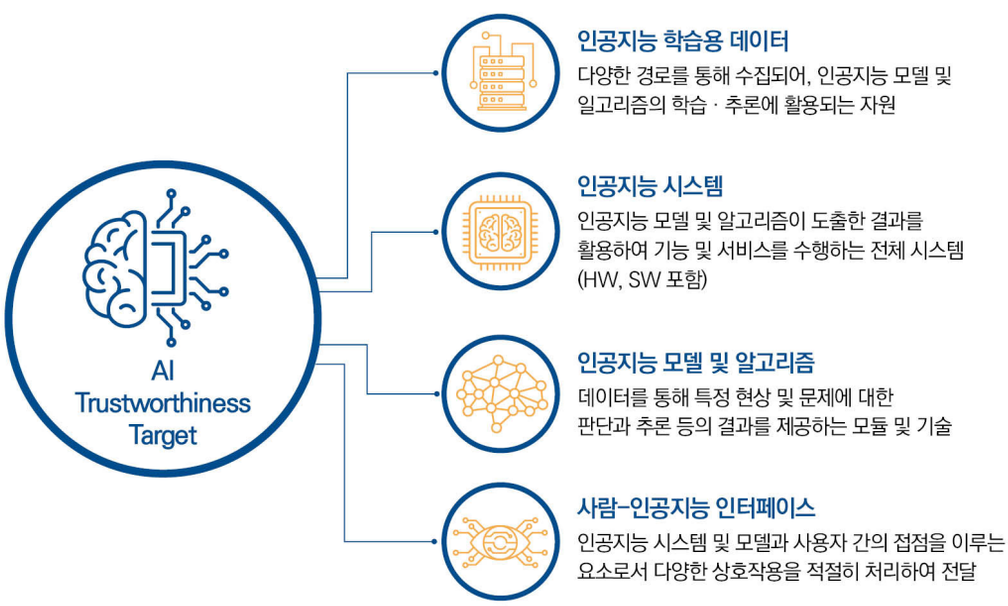
본 신뢰할 수 있는 인공지능 개발 안내서는 최근의 인공지능 관련 사건·사고와 주요 이슈에 대한 국제 사회의 움직임을 면밀히 분석해 기획했다. 발간 배경에서 밝힌 바와 같이, 그간 국내외 많은 기관 및 기업이 인공지능 신뢰성 확보를 위한 윤리 원칙과 지침, 가이드라인을 내놓았으나, 기술적 관점에서 상세한 방법론을 제시한 사례는 아직 없었다. 따라서 본 개발 안내서는 인공지능 활용 서비스 개발 현장에서 실무를 수행하는 데이터 과학자, 모델 개발자, 시스템 및 소프트웨어 개발·검증자, 설계자 및 기획자와 서비스 운영자 등이 실무 관점에서 신뢰성 확보 방안과 기준을 참고하고 현장에 적용할 수 있도록 계획했다. 이에 따라 본 개발 안내서의 개발 과정은 기술적, 공학적인 관점에서 신뢰성 확보에 필요한 요소를 찾는 데서 시작했다.

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

인공지능 서비스는 크게 네 가지 요소로 구성된다. 핵심적인 사고 기능을 수행하는 인공지능 모델 및 알고리즘, 이를 학습시킬 데이터, 실제 기능에 접목될 소프트웨어 기반 시스템, 필요에 따라 사용자와 상호작용하기 위한 인터페이스가 그 구성 요소다. 이들은 각자 개별적으로, 또는 함께 통합되어 인공지능 서비스의 생명주기에 따라 개발, 검증 및 운영된다. 따라서 인공지능 서비스를 개발하는 실무자 입장에서는 인공지능 서비스의 구성과 생명주기에 따른 기술적·공학적인 요구사항과 함께 이를 제대로 적용했는지 확인할 수 있는 검증 항목이 제공돼야 한다.

첫 번째, 인공지능 서비스를 구성하는 4가지 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 및 공평성 등이 배제되었는지 검증
인공지능 모델 및 알고리즘	인공지능이 모델 및 알고리즘에 따라 안전한 결과를 도출하며, 이에 대한 설명이 가능한지, 악의적인 공격에 강건한지 등을 검증
인공지능 시스템	인공지능 모델 및 알고리즘이 적용된 전체 시스템을 대상으로 인공지능 도출 결과대로 작동하는지, 잘못된 경우의 대책이 존재하는지 등을 검증
사람-인공지능 인터페이스	인공지능 시스템 사용자·운영자 등이 인공지능 시스템의 동작을 쉽게 이해할 수 있으며, 사람의 실수 및 인공지능의 오작동을 방지할 수 있는지 등을 검증

두 번째, 인공지능 서비스의 생명주기는 아래와 같이 정의할 수 있다.

인공지능 서비스 생명주기

생명주기 단계	주요 행위자	주요 활동
1. 계획 및 설계	• 비즈니스 결정권자 • 데이터 과학자 • 시스템 운영자	– 비즈니스 모델 및 성과지표KPI 정의 – 인공지능 시스템의 목적 및 전 생명주기에 따른 요구사항 수집 – 인공지능 시스템 개념 증명 및 필요 리소스 정의
2. 데이터 수집 및 처리	• 데이터 공급자 • 데이터 과학자 • 도메인 전문가	– 데이터 품질 확보, 데이터 사용자의 이해를 위한 정보 제공 방안 마련 – 개인 데이터 보호와 같은 데이터 접근 제어 및 비식별화 정책 수립 – 데이터 라벨링 및 데이터셋 특성 문서화 – 인공지능 모델 구축을 위한 데이터셋 마련
3. 인공지능 모델 개발	• 데이터 과학자 • 인공지능 모델 개발자 • 시스템 엔지니어	– 비즈니스 목적에 따른 인공지능 모델 구현 – 구현된 인공지능 모델 확인 및 검증 – 인공지능 모델 튜닝, 데이터 분석 및 추가로 필요한 데이터 고려 – 최종 인공지능 모델에 대한 성능평가
4. 시스템 구현	• 인공지능 모델 개발자 • 시스템 엔지니어	– 레거시 시스템과의 호환성 확인 – 기능 단위 테스트, 시스템 검증, 배포 버전 승인 – 인공지능 시스템의 파일럿 테스트 수행
5. 운영 및 모니터링	• 비즈니스 결정권자 • 인공지능 모델 개발자 • 시스템 운영자	– 시스템 모니터링 및 인공지능 모델 재학습을 통한 성능 보장 – 모델 편향 탐지, 공평성, 설명가능성 등 시스템 신뢰성 모니터링 – 치명적 문제 발생 시, 시스템 폐기 의사결정

인공지능 서비스 생명주기는 첫 번째에서 살펴본 구성 요소들을 구현하고 운영하는 과정을 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 구분해 정의한다. 대표적으로 OECD와 ISO/IEC에서 제시한 생명주기가 있는데, 본 개발 안내서는 두 기관에서 제시한 생명주기를 대표성 있는 사례로 참고하여, 실무자들이 쉽게 활용할 수 있도록 각 생명주기 단계의 성격과 활동을 왜곡하지 않는 선에서 5가지 단계로 단순화했다.

인공지능 서비스의 생명주기 단계는 반복적, 순환적인 성격을 지니며, 순차적이지 않을 수 있다. 본 개발 안내서는 이해를 돕기 위해 단계1부터 단계5까지 순차적인 것처럼 설명했으나, 실제 데이터를 수집하고 가공하거나 모델을 개발, 운영하는 과정에서는 순서의 의미가 사실상 없을 수도 있다.

세 번째, 인공지능 신뢰성에 필요한 요건을 정의하고자 '인공지능 윤리기준'의 핵심 요건을 준용하여 기술적 관점의 요구사항과 함께 검증 항목에 필요한 요건 4가지를 아래와 같이 도출했다.

1. 다양성 존중
2. 책임성
3. 안전성
4. 투명성

EC, OECD, IEEE 및 ISO/IEC와 같은 국제기구와 표준화 기구는 인공지능 신뢰성을 하위 속성으로 세분화해 제시한다. 또한 학계와 산업계에서는 신뢰성을 구성하는 하위 속성이 별도의 학제를 형성하기도 한다. 특히, ISO/IEC 24028:2020은 신뢰성 확보에 필요한 고려사항의 형태로 키워드를 제공한다. 여기에는 투명성, 통제가능성, 강건성, 복구성, 공평성, 안전성, 개인정보보호, 보안성 등이 포함되나, 키워드간의 관계나 신뢰성과의 연관성은 정의되지 않았다. 이처럼 관점에 따라 유사해 보이지만 조금씩 다른 용어들이 여러 문헌에서 제각각 달리 정의되고 있으며, 아직 합의된 정의는 없는 상황이다. 이에 앞서 언급한 EC, OECD, IEEE, ISO/IEC 등 여러 기관에서 제시한 속성과 키워드를 종합적으로 분석하고, 국내 학계·연구계·산업계 전문가의 의견을 수렴해 합의점을 모색했다. 이처럼 폭넓은 의견 공유 절차를 거쳐 인공지능 신뢰성 속성을 도출한 후, 이를 국가 인공지능 윤리기준의 10대 요건에 대응시켜서 기술적 측면에서 다룰만한 요건을 선정했다.

3.2 인공지능 신뢰성 확보 요구사항 및 검증항목 도출

다음 단계로 구체적인 요구사항과 검증항목을 도출했다. 우선 표준화기구, 기술단체, 국제기구, 주요 국가 정부에서 인공지능 신뢰성 확보를 위해 발표한 정책, 권고안, 그리고 표준을 기반으로 준수해야 할 기술적 요구사항을 도출하고 구체화하여 제시했다. 이와 함께 AI 개인정보보호 자율점검표('21.5), 금융분야 AI 가이드라인('21.7) 등 국내에서 인공지능 신뢰성 확보를 목적으로 발표된 사례들을 검토했다. 이러한 과정을 거쳐 개발 안내서에 중요한 내용은 반영하고 중복된 내용은 제거하거나 축약했다. 해당 참고문헌은 다음과 같다.

인공지능 신뢰성 관련 주요 참고문헌

기관명	발간년월	권고 및 표준안 명
대한민국	2020.11	국가 인공지능 윤리기준(안)
유럽위원회	2020.07	The Assessment List for Trustworthy Artificial Intelligence
국제표준화기구 (ISO/IEC)	2021.03	ISO/IEC TR 24029-1:2021, Artificial Intelligence (AI) – Assessment of the robustness of neural networks – Part 1: Overview
	2021.01	ISO/IEC 23894, Information Technology – Artificial Intelligence Risk Management
	2020.11	ISO/IEC 24027, Information technology – Artificial Intelligence (AI) – Bias in AI systems and AI aided decision making
	2020.05	ISO/IEC TR 24029:2020, – AI – Overview of Trustworthiness in artificial intelligence
Google	2019.05	People + AI guidebook
유럽전기통신표준협회 (ETSI)	2021.03	Securing Artificial Intelligence (SAI 005) – Mitigation Strategy Report
경제협력개발기구 (OECD)	2019.05	Recommendation of the Council on Artificial Intelligence
세계경제포럼 (WEF)	2020.01	Companion to the Model AI Governance Framework

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	데이터의 활용을 위한 상세 정보 제공		✓		✓
요구사항 03	데이터 강건성 확보를 위한 이상Abnormal 데이터 제거			✓	
요구사항 04	수집 및 가공된 학습 데이터의 편향 제거	✓	✓		✓
요구사항 05	오픈소스 라이브러리의 보안성 및 호환성 확보		✓	✓	
요구사항 06	인공지능 모델의 편향 제거	✓			
요구사항 07	인공지능 모델 공격에 대한 방어 대책 수립			✓	
요구사항 08	인공지능 모델 명세 및 출력 결과에 대한 설명 제공		✓		✓
요구사항 09	인공지능 모델 출력에 대한 신뢰도Confidence value 제공				✓
요구사항 10	인공지능 시스템 구현 시 발생 가능한 편향 제거	✓			
요구사항 11	인공지능 시스템의 안전 모드 구현		✓	✓	✓
요구사항 12	인공지능 시스템의 설명에 대한 사용자의 이해도 제고				✓
요구사항 13	인공지능 시스템의 추적가능성 확보			✓	✓
요구사항 14	서비스 제공 범위 및 상호작용 대상에 대한 설명 제공		✓		✓

3.3 산학연 실무자 의견 수렴

신뢰성 확보를 위한 요구사항을 선별한 후에는 각 항목을 기술적 관점에서 검토하여 개발 안내서가 현업 종사자들의 관점과 눈높이에 맞도록 고도화했다. 이 검토는 기술적 타당성, 효용성 및 포괄성이라는 관점을 포함했다. 각각의 세부 점검 항목이 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 다수의 인공지능 분야 전문가가 참여하여 직접 검토하고 자문했으며, 다양한 검토 의견을 수렴하여 반영했다. 인공지능 분야 전문가에는 기업의 기획자, 개발 프로젝트 리더, 교수, 국가연구소 책임연구원, 관련 국가 정책 담당자 등 산업계 및 학계의 연구자들을 분야를 가리지 않고 총망라해 다양하게 섭외했다.

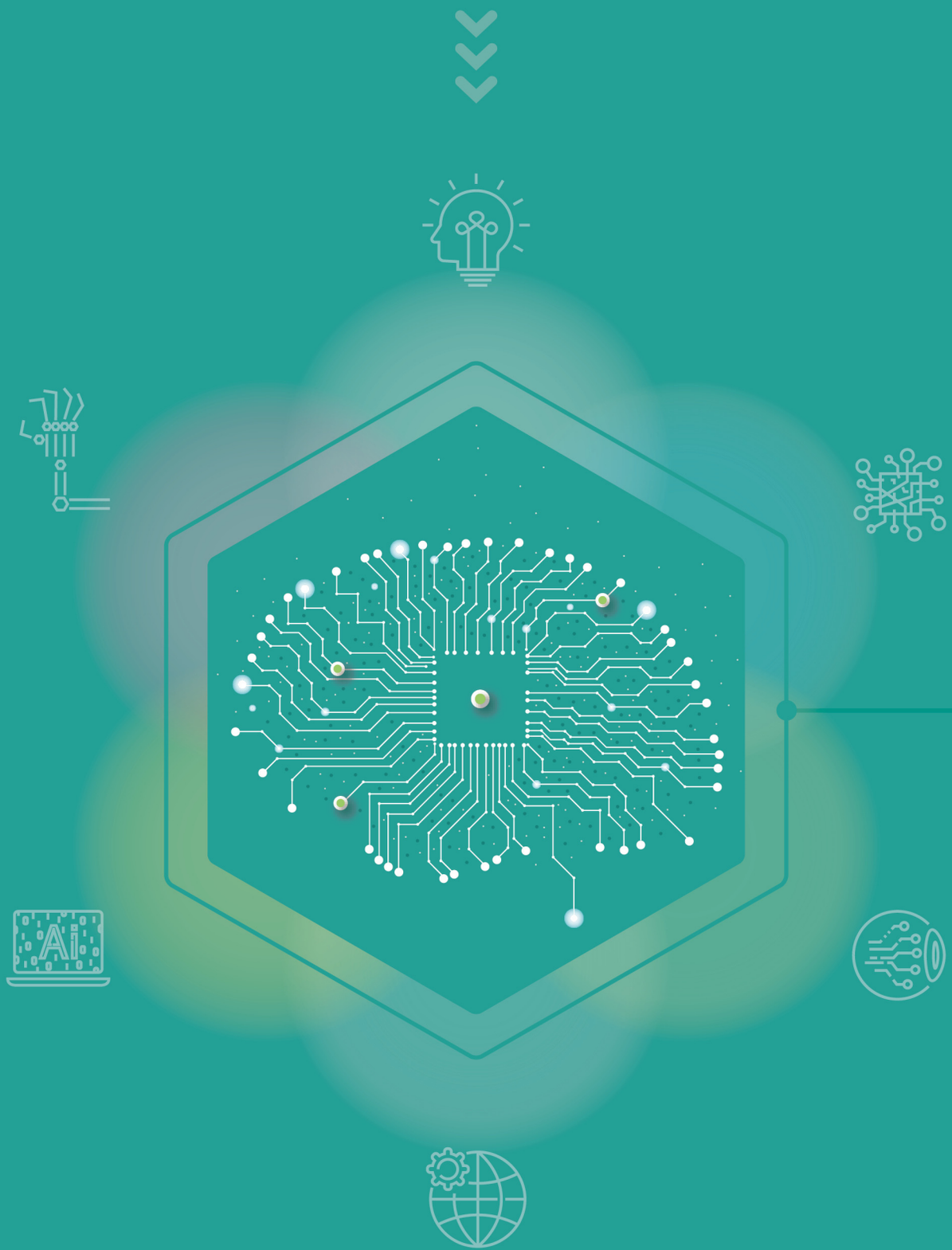
04 신뢰할 수 있는 인공지능 개발 안내서 활용 대상

신뢰할 수 있는 인공지능 개발 안내서는 인공지능 서비스를 구현하는 과정에 직·간접으로 관련되거나 영향을 주는 모든 조직과 개인을 포함한 이해관계자를 대상으로 제작됐다. 특히 업무상 기술적 관점에서 신뢰성을 신경써야 하는 기획자, 데이터 수집 및 가공자, 인공지능 모델 개발자, 시스템 및 소프트웨어 개발자, 테스터 등이 주요 대상이다. 이들이 인공지능 생명주기의 각 단계마다 인공지능의 신뢰성을 확보하기 위해 검토해야 할 주요 요구사항은 다음과 같다.

인공지능 생명주기 단계별 신뢰성 확보 요구사항

생명주기 단계	주요 행위자	주요 요구사항
1. 계획 및 설계	• 비즈니스 결정권자 • 데이터 과학자 • 시스템 운영자	– 인공지능 시스템 전체 생명주기에 걸친 신뢰성 확보 요구사항 검토 및 적용방안 수립
2. 데이터 수집 및 처리	• 데이터 공급자 • 데이터 과학자 • 도메인 전문가	– 학습 데이터 확보 과정에서 발생할 수 있는 데이터 오류 및 편향에 대한 관리방안 확보
3. 인공지능 모델 개발	• 데이터 과학자 • 인공지능 모델 개발자 • 시스템 엔지니어	– 학습 모델의 편향적인 출력이나 공격에 대한 대응방안 수립 – 학습 모델의 출력에 대한 해석방안 제공
4. 시스템 구현	• 인공지능 모델 개발자 • 시스템 엔지니어	– 인공지능 시스템 개발 시 발생 가능한 편향이나 오류에 대한 대응책 마련 – 인공지능 서비스가 도출한 결과에 대해 사용자 친화적인User-friendly 설명 제공
5. 운영 및 모니터링	• 비즈니스 결정권자 • 인공지능 모델 개발자 • 시스템 운영자	– 인공지능 시스템 문제 발생 시 원인 추적을 통한 대응 방안 마련

2022 **신뢰할 수 있는 인공지능** 개발 안내서(안)



PART 2

요구사항

목 차

목 차

1 계획 및 설계

요구사항 01 인공지능 시스템에 대한 위험관리 계획 및 수행

책임성 투명성

▶ 인공지능 시스템이 구현 및 운영되는 과정에서 발생 가능한 모델 오인식, 기능 오동작, 보안 및 개인정보 이슈와 같은 위험 요소를 사전에 인식하고, 위험의 크기(심각성 및 파급효과)를 분석하여 대응 방안을 마련해야 한다.

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

[Yes I No I 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 31000:2018에는 위험관리에 대한 개념 및 정의와 전체적인 흐름이 소개되어 있다.
- 다만, 인공지능의 신뢰성을 확보하는 과정에서 방해가 될 수 있는 위험 요소를 인식, 분석 및 평가하는 방법론은 기존의 소프트웨어 및 하드웨어 기반 시스템과는 상이할 수 있으므로 이 점을 고려해야 한다. ISO/IEC 24028:2020과 ISO/IEC 23894.2에서는 인공지능의 신뢰성 관점에서 살펴보아야 할 위험 요소의 분류가 제공되어 있다.
- 위험 요소별로 도출이 된 다음에는 위험 요소가 발생할 수 있는 원인 및 상황과 조건을 분석한 다음, 위험 요소가 인공지능 시스템 또는 인간 및 주변 환경에 얼마나 큰 영향을 미치는지 분석하여야 한다.

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

[Yes I No I 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 방식과 같은 절차, 소프트웨어 및 하드웨어 기능, 모델 학습 기법 및 전략 등 기술적으로 적용할 수 있는 모든 방법을 의미한다. 이에 대해 ISO/IEC 24028:2020에 대응 방안의 분류가 제공되어 있다. 인공지능을 구현하는 모든 이해관계자는 이를 고려하여 위험 요소에 대한 대응 방안을 마련하고, 위험이 제거 및 완화되었는지 확인하여야 한다.

2 데이터 수집 및 처리

요구사항 02 데이터의 활용을 위한 상세 정보 제공

책임성 투명성

▶ 인공지능 학습용 데이터셋은 개발 과정에서 데이터가 추가로 수집될 수 있으며, 다른 유사 시스템의 학습 데이터로 사용될 수도 있다. 이때, 데이터 수집 출처, 특징과 같이 수집된 데이터의 정보가 미비하다면 재사용성이 떨어지거나 데이터로 인해 야기된 문제에 대한 원인 파악이 어려울 수 있다. 따라서 수집 데이터의 올바른 활용과 문제 발생 시 명확한 원인 추적을 위해 수집 데이터의 상세정보가 제공되어야 한다.

02-1 데이터의 명확한 이해와 활용을 지원하는 상세한 정보를 제공하는가?

[Yes I No I N/A] ☐ ☐ ☐

- 데이터를 설명하기 위한 데이터로서 메타데이터Metadata가 정의될 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.

02-2 데이터의 출처는 기록 및 관리되고 있는가?

[Yes I No I N/A] ☐ ☐ ☐

- 학습 데이터의 품질은 인공지능 모델 성능에 큰 영향을 미치는 중요한 요인 중 하나이며 품질이 확보된 학습 데이터를 사용하기 위해 다양한 오픈소스 데이터셋을 활용할 수 있다.
- 오픈소스 데이터셋의 경우 다수의 사용자가 데이터 활용 과정에서 발견한 오류가 추후 발견될 수 있으며, 이로 인한 데이터셋 수정, 재구축으로 데이터 버전 변경이 있을 수 있다.
- 이러한 데이터셋 자체 원인으로 발생할 수 있는 인공지능 모델의 문제 대응을 위해서는 학습에 사용한 데이터의 명확한 출처, 구축 시점, 오픈소스 데이터셋 버전과 같은 정보를 관리해야 한다.

요구사항 03 데이터 강건성 확보를 위한 이상Abnormal 데이터 제거

안전성

▶ 인공지능 모델의 학습에 활용되는 데이터는 이상값, 중독 및 회피 등에 영향을 받지 않아야 하며, 이의 점검 및 방어 기법의 적용을 통해 강건성을 확보하여야 한다.

03-1 데이터 이상값Outlier 식별 및 정상·오류 여부를 점검하였는가?

[Yes I No I N/A] ☐ ☐ ☐

- 학습용 데이터의 이상값 및 오류 데이터란 학습용 데이터를 구성하는 데이터셋의 수집 및 가공 과정에서 발생할 수 있는 다양한 오류들을 포괄한다. 학습 데이터의 수집 및 가공 과정에서 발생하는 오류들은 데이터 상의 노이즈, 학습 데이터 내의 편향, 잘못된 라벨링, 라벨링 누락 등 다양하게 존재할 수 있으며 이를 점검하여 대처하지 않으면 인공지능 모델의 성능 및 강건성을 충분히 확보할 수 없다.
- 단, 사전 학습된 모델을 활용한다면 추가 학습 데이터의 오류에 따른 영향이 적을 수 있다. 해당 경우는 데이터 오류 점검을 최소화할 수 있다.

03-2 데이터 공격에 대한 방어 수단을 강구하였는가?

[Yes I No I N/A] ☐ ☐ ☐

- 인공지능 서비스 개발 또는 운영 과정에서 의도적으로 학습 데이터를 변질시키거나 입력 데이터에 최소한의 변조를 가해 예상과는 다른 결과를 출력하도록 하는 공격에 노출될 수 있으므로, 이를 대처할 방안을 검토 및 적용하는 것이 바람직하다.

공격 기법 분류	공격 기법 내용	대표 방어 기법
데이터 중독 공격	인공지능 서비스는 일반적으로 입력 데이터 분포의 변화에 적응하기 위해 모델 배치 후 수집된 새로운 데이터를 사용해 재교육된다(예: 침입 감지 시스템). 이때, 공격자는 세심하게 조작된Perturbed 데이터를 주입하여 서비스의 정상적인 기능을 손상시키는 방식으로 훈련 데이터를 오염시킬 수 있다.	• 적대적 학습Adversarial training • Gradient Masking (Distillation) • Feature Squeezing
회피 공격	공격자는 학습 모델이 입력을 올바르게 식별할 수 없도록 기존의 입력 데이터에 대해 미묘한 차이의 노이즈를 추가하여 조작된Perturbed 입력 데이터를 생성한다. 이러한 변화는 사람의 눈에 잘 띄지 않지만, 딥러닝 모델의 출력에 큰 영향을 미친다.	

요구사항 04 수집 및 가공된 학습 데이터의 편향 제거

다양성 존중 | 책임성 | 투명성

▶ 학습을 위한 데이터를 수집 및 가공 시 발생할 수 있는 편향을 인식하고 이를 제거하기 위한 방안을 적용하여야 한다. 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 마련하여야 한다. 단, 이미 편향성 검토가 완료된 데이터를 활용하거나, 초거대 인공지능 모델처럼 현실적으로 모든 데이터를 검증하기 어려운 경우에는 샘플링 기법 등을 통해 데이터를 검증할 수 있다.

04-1 데이터 수집 시, 인적·물리적 요인으로 인한 편향 완화 방안을 마련하였는가?

[Yes I No I N/A] ☐ ☐ ☐

- 인적 요인으로 인한 편향은 사람이 의식적 혹은 무의식적으로 특정 정보에 대해 편향되어 있는 점에서 기인한다.
  - ✓ 인적 편향: 자동화 편향Automation bias, 그룹 귀인 편향Group attribution bias, 암묵적 편향Implicit bias, 그룹 내 편향In-group bias 등이 포함됨
- 인적 편향을 방지하기 위해 수집된 데이터의 명확한 수집 및 검수 기준을 수립하여 수집하는 작업자별로 데이터 특성이 편향되지 않도록 방지하거나, 다양하고 충분한 수의 검수자를 확보함으로써 검수 시 편향을 바로잡을 수 있어야 한다.
- 데이터는 수집 도구나 방법에 활용되는 물리적 요인으로 인해 데이터의 편향이 발생할 수 있다. 이미지의 촬영 도구나 저장 장치와 같은 요인으로 인하여 이미지의 색상, 밝기, 해상도와 같은 물리적으로 한정된 데이터가 수집될 수 있다.
- 이로 인해 촬영 대상자의 연령대나 인종을 구분하기 힘들거나, 특정 방법으로 수집된 데이터만 학습이 이뤄지므로, 편향을 발생시킬 수 있는 물리적 요인을 제거하거나 다양한 수집 장치를 활용하여 다양성을 보완하는 것이 바람직하다.

04-2 학습에 사용되는 특성Feature을 분석하고 선정 기준을 마련하였는가?

[Yes I No I N/A] ☐ ☐ ☐

- 편향 제거를 위해 데이터에 포함된 차별적인 요소를 사전에 추리는 것이 중요하며, 이를 위해 학습을 위한 특성에 대한 분석과 선정 기준을 수립하는 것이 바람직하다.
- 차별적인 요소란 데이터에 포함된 성별, 인종, 사회 취약계층의 정보로 인해 학습 결과가 사회적 물의 및 차별을 일으킬 수 있는 것을 의미한다. ISO/IEC 24027에서는 이를 보호변수Protected attribute라고 지칭하며, 데이터 학습 시 반영되지 말아야 하는 특성으로 선정하고 편향을 방지하도록 언급하고 있다.

사회적 물의를 일으킬 수 있는 민감한 특성들

ISO/IEC 24027	IBM Watson OpenScale	Google
나이, 성별, 인종, 수입, 가족관계, 교육 수준, 키·체중, 장애 여부	나이, 성별, 인종, 결혼 여부, 주소	인종, 성별 장애 여부, 종교

04-3 데이터 라벨링 시, 발생 가능한 편향을 확인하고 방지하였는가?

[Yes I No I 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.

04-4 편향 방지를 위해 데이터 분포 검증을 통한 데이터 샘플링을 수행하였는가?

[Yes I No I N/A] ☐ ☐ ☐

- 샘플링은 데이터셋에서 일정한 기준으로 데이터를 추출하여 전체 데이터의 분포를 검증하는 기법이다. 일정한 기준으로 추출된 샘플 데이터는 수집된 전체 데이터의 분포를 대표하여야 의미가 있다. 그렇지 못할 경우 샘플링 과정에서 의도하지 않은 편향이 발생할 수 있다.
- 샘플링 기법 중 대표적인 기법으로는 층화추출법Stratified sampling을 예로 들 수 있다. 이는 모집단의 특성을 고려해 각 표본이 중복되지 않도록 계층으로 나눈 다음, 각 계층에서 표본을 추출함으로써 모집단 특성을 반영한 표본을 만들어 편향을 방지한다.

3 인공지능 모델 개발

요구사항 05 오픈소스 라이브러리의 보안성 및 호환성 확보

안전성 책임성

▶ 인공지능 모델 설계 및 개발 단계에서는 개발 기간을 단축하고 최신 기술 동향을 빠르고 유연하게 적용하기 위해 다양한 오픈소스를 활용할 수 있다. 오픈소스를 활용할 경우, 오픈소스의 목록 및 버전을 지속적으로 확인하여 운영 및 보안상의 위험 요소를 점검해야 한다.

05-1 오픈소스 라이브러리의 보안성 및 호환성 확보 여부를 확인하였는가?

[Yes I No I N/A] ☐ ☐ ☐

- 오픈소스 라이브러리는 버전 변경에 따라 법률 및 기술적 측면의 이슈가 발생할 수 있다. 따라서 모델 개발에 오픈소스 라이브러리를 활용하였다면, 오픈소스 라이브러리의 신규 버전 출시에 따른 변경사항 또는 사용 중인 버전에서 새롭게 발견된 이슈를 지속적으로 추적해야 한다.
- 법률적 측면의 대표적인 이슈는 라이선스License가 있으며, 기술적 측면의 대표적인 이슈는 호환성Compatibility 및 보안취약점Vulnerability이 있다. 이러한 이슈들의 확인을 통해 운영 및 보안상의 위험 요소에 대한 점검이 필요하다.

요구사항 06 인공지능 모델의 편향 제거

다양성 존중

▶ 인공지능 모델을 개발하는 과정에서 모델의 종류나 시스템의 목표에 따라 편향*이 발생할 수 있으며, 이를 제거하기 위한 기법을 고려하여야 한다.

* 요구사항 04-2에서 언급한 바와 같이 인종차별, 성차별 등 사회·윤리적으로 문제가 되는 경우에 한함

06-1 모델 편향을 제거하는 기법을 적용하였는가?

[Yes I No I N/A] ☐ ☐ ☐

- 인공지능 모델은 데이터에 잠재된 편향을 학습하게 되고, 심지어 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-processing이다. 구현하려는 인공지능 모델 및 목표 임무에 따라서 이 중 적절한 기법을 선택하여 적용하여야 한다.

요구사항

07 인공지능 모델 공격에 대한 방어 대책 수립

안전성

▶ 인공지능 모델은 적대적 의도를 가진 사용자에 의해 학습 데이터 및 기능을 도용당하거나 다른 방식의 공격으로 악용될 수가 있으므로 이를 방지 또는 완화하기 위한 대책이 수립되어야 한다.

07-1 모델 추출 공격Model extraction attack에 대한 방어 기법을 도입하였는가?

[Yes I No I N/A]
☐ ☐ ☐

- 모델 추출 공격은 학습된 모델의 다양한 입력에 대한 예측 결과를 분석하고 분류 기준을 추출하여 서비스 중인 학습 모델과 유사한 성능의 대체 모델을 구성하는 방법과 학습된 모델의 입력 데이터, 모델의 초매개변수Hyperparameter 정보, 계층 구조 등을 추출하는 공격 방식이 존재한다. 이러한 인공지능 모델에 대한 공격을 완화하기 위해 질의Query 횟수 제한, 예측 결과 난독화와 같은 방법들을 적용할 수 있다.

Case Study 클라우드 기반 머신러닝 서비스에 대한 모델 추출 공격 결과

Service	Model Type	Queries	Time(s)
Amazon	Logistic Regression	650	70
BigML	Decision Tree	1,150	631

Stealing Machine Learning Models via Predictions APIs, Usenix, 2016

- 모델 추출 공격 방법으로 70초 동안 650번의 질의로 아마존 클라우드에서 제공하는 머신러닝 모델(Logistic Regression)과 유사한 모델을 만들어 낸 연구 결과
- 선형 분류기의 회귀계수 및 결정트리의 경로에 대한 정보를 획득해 유사한 인공지능 모델을 만들 수 있음

요구사항 **08** **인공지능 모델 명세 및 출력 결과에 대한 설명 제공**

책임성 투명성

▶ 인공지능 모델의 출력만으로는 예측된 결과가 어떤 요소에 의해 도출되었는지 알기 어렵다. 또한, 시스템의 최종 결과를 얻기 위해 이러한 인공지능 모델이 다수 사용될 수 있다. 이러한 과정에서 인공지능 모델의 예측 결과에 대한 사용자 신뢰를 확보하기 위해서는 사용된 모델 정보 및 결과 도출 과정 설명*이 제공되어야 한다.

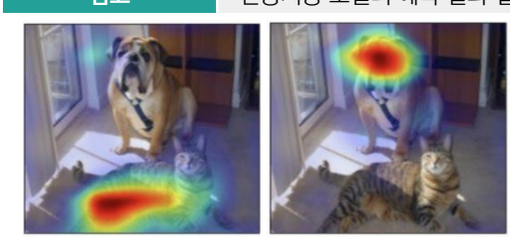
* 사람이 인공지능 모델의 의사결정 방식을 파악할 수 있도록 돕는 모델의 작동 방식에 대한 유용한 정보(예: 의사결정 매커니즘, 의사결정의 기초를 이루는 학습 데이터, 인공신경망 내에서 사용된 변수와 가중치)

08-1 인공지능 모델의 예측 결과를 설명하기 위한 기법 적용에 대한 검토를 하였는가?

[Yes I No I N/A] ☐ ☐ ☐

- 인공지능 모델의 예측 결과 및 인공지능 시스템의 동작을 사용자가 신뢰하기 위해서는 시스템 사용자가 인공지능 모델이 제공하는 판단 혹은 예측 결과의 도출 과정을 이해할 수 있어야 하며, 이에 대한 설명 및 근거를 사용자에게 제시하는 것이 바람직하다.
- 사람이 이해할 수 있는 방식으로 모델 판단의 근거를 제시할 수 있는 설명 가능한 인공지능XAI, eXplainable AI에 대한 검토 및 적용을 고려해야 하며, 설명이 필요한 요소 및 인공지능 모델 특성에 따라 대리Surrogate 모델, 집중Attention, 내부Internal 분석 방식 등 XAI 기술을 도입할 수 있다.

참고 인공지능 모델의 예측 결과 설명 예시 – 히트맵Heatmap을 사용한 예측 근거 시각화



인공지능 모델이 이미지를 개라고 예측한 경우와 고양이라고 예측한 경우, 모델이 각 예측에서 중요하게 여긴 영역은 다를 것이다.
왼쪽 그림과 같이 히트맵을 사용하여 모델이 예측에 활용한 영역을 시각화한다면, 사람이 모델 예측 근거를 이해하는 데 도움이 될 수 있다.

08-2 팩트 시트Fact sheet를 통해 인공지능 모델의 명세를 투명하게 제공하는가?

[Yes I No I N/A] ☐ ☐ ☐

- 인공지능 시스템의 투명성을 확보하는 방안 중 하나는 인공지능 모델 또는 서비스의 개발, 테스트 및 배포 과정에서 발생된 다양한 결과들의 묶음인 팩트 시트를 확보하는 것이다. 팩트 시트가 확보될 경우, 사용자가 인공지능 모델과 관련된 정보를 요구했을 때 모델의 목적, 입·출력 정보, 성능, 편향 여부 및 신뢰도 등의 결과들을 투명하게 공개할 수 있다.
- IBM의 AI FactSheet 360 프로젝트에서는 이러한 팩트 시트를 통해 인공지능 시스템의 투명성을 확보하는 방안을 제시하고 있으며, 개발한 시스템의 알고리즘 공개 없이 필요에 따라 팩트 시트를 통한 인공지능 모델의 주요 정보 및 구성 요소를 설명할 수 있도록 하는 것을 목적으로 하고 있다.

요구사항 09 인공지능 모델 출력에 대한 신뢰도Confidence value 제공

투명성

▶ 인공지능 모델이 도출한 결과를 더욱 정확하고 구체적으로 설명하기 위해 신뢰도를 활용할 수 있다. 여기서 신뢰도란, 정확도Accuracy, 정밀도Precision, 재현율Recall 등 인공지능 모델 성능지표와 불확실성을 함께 나타냄으로써 도출된 결과의 알고리즘 신뢰성이 얼마나 되는지를 의미한다. 신뢰도를 활용하고자 할 경우, 신뢰도를 표시하는 것이 사용자의 의사결정에 도움을 주는지 등을 사전에 검토해야 한다. 또한 낮은 신뢰도가 도출될 때를 대비하여 조치방안을 확보해야 한다. 단, 설계한 인공지능 모델에 따라 신뢰도 계산 및 신뢰수준 정의가 현실적으로 어려운 경우는 예외로 할 수 있다.

09-1 신뢰도 제공이 필요한 인공지능 모델 출력 결과에 대한 신뢰도를 제공하는가?

[Yes I No I N/A] ☐ ☐ ☐

1단계: 신뢰도가 필요한 결과인지 평가하기

- 사용자에게 신뢰도를 제공하면, 즉 인공지능 모델의 성능지표와 불확실성을 함께 보여준다면, 사용자는 단순히 해당 인공지능 모델이 얼마나 정확한지에 대한 정보뿐 아니라 그 결과의 신뢰성에 대해 객관적인 정보를 제공받을 수 있다. 이에 따라, 사용자의 의사결정에 도움이 되기도 하지만, 오히려 혼란을 유발하는 등 문제를 발생시킬 수 있다. 따라서, 상황 및 문맥을 고려하여 신뢰도의 필요성에 대해 반드시 평가하여야 한다.

2단계: 신뢰도 계산하기

- 정확도를 추정하기 위해 정밀도, 재현율, mAPmean Average Precision와 같은 지표를 사용할 수 있으며, 불확실성을 추정하는 방법으로는 앙상블Ensemble, 드롭아웃Dropout 기법 등이 있다.

09-2 신뢰도가 낮을 경우, 적절한 조치방안을 마련하였는가?

[Yes I No I N/A] ☐ ☐ ☐

1단계: 신뢰 구간Confidence interval에 따른 의미 정의하기

- 신뢰도가 도출되었으면 도출된 신뢰도의 구간을 정의하고, 이에 따른 의미를 정의하는 것이 필요하다. 만약 병변 진단 기능에 신뢰도를 표시한다면, 진단 결과의 신뢰 구간(예: 50% 이하, 51~60% 이하, 61~80% 이하, 81~90% 이하)을 정하고 구간별 신뢰도의 의미가 무엇인지 정의하는 것이다.

2단계: 신뢰도가 기대에 미치지 못하는 경우, 적절한 조치방안 또는 대안 마련하기

- 인공지능 모델의 결과는 입력값Input 혹은 응답 시간Response time과 같은 제약으로 신뢰도가 기대치에 미치지 못한 채 도출될 수 있다. 이러한 상황을 대비해 사용자 후속조치 정보를 마련하여 제공하고, 모델의 낮은 신뢰도에 대해 서비스 담당자에게 경고하는 기능을 구현하는 등 대응책을 마련해야 한다.
  - ✓ 예를 들어, 사용자가 인공지능 스피커에 질문 시 입력값, 즉 음성을 제대로 인식하지 못해서 부정확한 입력값을 갖게 되면, 낮은 성능과 더불어 불확실성도 높아질 것이다. 이러한 상황에서 인공지능 스피커는 사용자에게 "무슨 말인지 잘 모르겠어요."와 같은 회피형 답변을 제공하는데, 이것 역시 신뢰도가 기대 수준에 미치지 못하는 것에 대한 조치 중 하나라고 할 수 있다.

4 시스템 구현

요구사항 10 인공지능 시스템 구현 시 발생 가능한 편향 제거

다양성 존중

▶ 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 설계하여야 한다.

10-1 소스 코드 및 사용자 인터페이스로 인한 편향을 제거하기 위해 노력하였는가?

[Yes I No I 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 단계에서의 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 인터랙션Interaction 측면에서는 표현 편향Presentation bias이나 순위 편향Ranking bias 등이 발생하지는 않는지 미리 확인하여 편향을 방지할 수 있도록 시스템을 설계하는 것이 바람직하다.

요구사항 11 인공지능 시스템의 안전 모드 구현

안전성 책임성 투명성

▶ 인공지능 시스템을 통해 생성되는 결과나 의사결정은 개인 혹은 사회에 부정적인 영향을 미칠 수 있으므로, 이에 대한 대응이 가능하도록 안전 모드를 구현하고, 리포팅이 가능하여야 한다.

11-1 공격, 성능 저하 및 사회적 이슈 등의 문제 발생 시 안전 모드를 적용하는가?

[Yes I No I N/A] ☐ ☐ ☐

- 고장 안전Fail-Safe은 산업 전반에서 사용되는 일반적 개념으로, 고장이나 오류로 문제가 발생하더라도 안전한 상태를 유지할 수 있는 메커니즘 및 기능을 의미한다. 이는 인공지능 시스템에도 적용될 수 있다. 인공지능 시스템에서도 외부로부터의 공격, 인적 오류Human error, 인공지능 모델의 정확도 저하, 편향 발생으로 인한 사회적 물의, 사고 등이 예상되는 경우, 이의 발생 원인을 파악하고 해결하거나 사용자에게 정상적인 기능으로 복구할 수 있는 방법을 사용자에게 제시하여야 한다. 이러한 대처를 위한 메커니즘이 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 신뢰도가 떨어지거나 문제 발생 가능성이 높은 경우, 이에 대한 의사 결정을 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

11-2 인공지능 시스템에서 문제가 발생할 경우 리포팅을 수행하는가?

[Yes I No I N/A] ☐ ☐ ☐

- 인공지능 시스템은 서비스 도중 외부로부터의 공격, 사용자의 오용 등 다양한 요인으로 편향이나 성능 저하 등이 발생할 수 있으므로 시스템 운영자가 이를 파악할 수 있도록 시스템의 자체적인 진단 기능이나, 사용자가 운영자에게 관련 사항을 고지할 수 있는 리포트 기능을 제공해야 한다.
- 시스템의 자체적인 진단 기능은 서비스 성능 저하나 외부 공격에 대한 검사 등을 수행한 후 가능한 범위 내에서 이를 대응하고, 해당 사실을 시스템 운영자에게 보고할 수 있는 체계를 갖춰야 한다.
- 사용자 리포트 기능은 시스템의 일시적인 오류나 도출 결과에 편향이 발생하는 등 문제가 생길 때 사용자가 해당 사실을 시스템 운영자에게 전달할 수 있는 기능을 의미하며, 자체 진단 기능과 함께 사람에 의한 리포팅도 가능하여야 한다.

4 시스템 구현

요구사항 **12** 인공지능 시스템의 설명에 대한 사용자의 이해도 제고

투명성

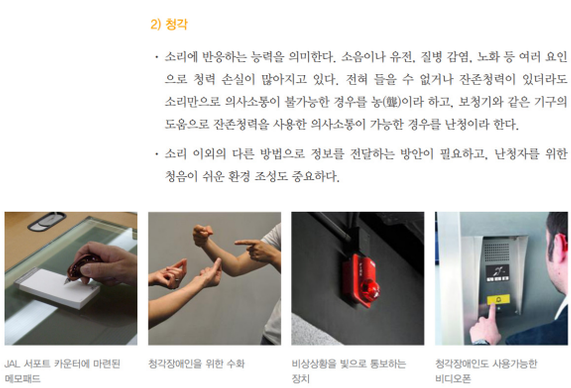
▶ 모델의 예측 결과에 대해 설명을 제공하는 기법을 적용하여도 사용자가 바로 이해해 해석하기 어려운 경우가 많다. 따라서 인공지능 시스템의 운영자 혹은 서비스 제공자는 사용자에게 제공되는 결과가 이해 가능한지Understandability, 해석 가능한지Interpretability, 설명 가능한지Explainability를 평가하여야 한다.

12-1 인공지능 시스템 사용자의 특성User characteristics과 제약사항을 분석하였는가?

[Yes I No I 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

'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.

12-2 사용자 특성에 따른 충분한 설명을 제공하는가?

[Yes I No I N/A] ☐ ☐ ☐

• 서비스를 이용하는 사용자는 다양하여 인공지능 시스템의 결과가 서로 다른 입장에서 설명이 해석되고 오해가 생길 수 있다. 따라서 12-1에서 분석된 사용자 특성을 고려하여 설명을 평가할 수 있는 기준 항목을 수집한다. 설명 평가의 기준으로는 명확성, 구체성, 정확성 등을 고려할 수 있다.

5 운영 및 모니터링

요구사항 13 인공지능 시스템의 추적가능성 확보

책임성 투명성

▶ 인공지능 시스템 운영 단계에서 문제 원인 추적을 위한 시스템 로그, 데이터 모니터링, 인공지능 모델과 사람 간의 의사결정 기여도 추적과 같은 기술적인 대응 방안을 확보해야 한다.

13-1 인공지능 시스템의 의사결정에 대한 추적 및 대응 방안을 수립하였는가?

[Yes I No I N/A] ☐ ☐ ☐

- 인공지능 시스템의 의사결정은 인공지능 모델이 자체 결정하거나 시스템 운영자 또는 사용자의 개입에 의해 내려질 수 있다. 또한, 운영 중에도 학습이 이뤄지도록 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 사용자와 같은 인적 요인으로 인해 발생 가능한 인공지능 시스템 출력 결과의 영향을 추적하기 위해서 시스템 단계별로 로그 수집대상 정보를 정의하고 모니터링을 지속해야 한다.

13-2 학습 데이터의 변경 이력을 주기적으로 관리하고 있는가?

[Yes I No I 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 추적할 수 있어야 한다.
- 이를 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 및 학습 모델 예상 출력 결과 등에 대한 정보를 제공해야 한다.

13-3 학습 데이터의 업데이트 이력을 주기적으로 관리하고 있는가?

[Yes I No I N/A] ☐ ☐ ☐

- 데이터에 큰 영향을 받는 학습기반 인공지능 모델은 신규 데이터 사용 시 성능이 떨어질 수 있으며, 필요 시 신규 데이터를 포함하여 추가 학습이 필요할 수 있다. 신규 데이터로 인한 인공지능 모델의 영향을 평가하기 위해서는 학습 데이터에 포함된 신규 데이터의 비율에 따른 모델 성능 변화 추적이 가능하도록 기록 및 관리하는 것이 바람직하다.
- 또한, 신규 데이터를 활용하여 기존 학습된 인공지능 시스템의 성능 영향 평가를 수행하는 경우 해당 도메인의 대표적 인공지능 알고리즘을 활용한 성능 비교 분석을 통해 신규 데이터로 인한 인공지능 모델의 재설계, 재학습 등의 절차가 필요하지 않은지 점검해야 한다.

5 운영 및 모니터링

요구사항 **14** 서비스 제공 범위 및 상호작용 대상에 대한 설명 제공

책임성 투명성

▶ 사용자가 인공지능 시스템이 제공하는 서비스를 올바르게 사용하고, 제공된 서비스를 오·남용하지 않도록 서비스의 목적, 범위, 제한사항, 상호작용 대상을 포함한 내용을 설명하여야 한다.

14-1 인공지능 서비스의 올바른 사용을 유도하기 위한 설명을 제공하는가?

[Yes I No I N/A] ☐ ☐ ☐

- 인공지능의 활용 범위가 넓어지면서 사용자가 서비스 기능에 대한 기대를 실제 서비스 제공 범위보다 더 넓게 오해하는 경우가 발생하고 있다. 따라서 서비스 목적, 범위 및 제한사항에 대한 설명을 제공함으로써 인공지능 기술의 오·남용을 방지하고 사용자의 서비스에 대한 기대치를 조정하는 것이 중요하다.
- 특히, 사람이 아닌 대상을 사람으로 의인화하는 서비스의 경우, 사용자의 초기 기대치가 잘못 형성되어 있을 확률이 높다. 따라서 서비스 제공자는 서비스 제공 목적과 의도가 무엇인지, 서비스 제공 범위는 어디까지이며 이와 관련된 한계는 무엇인지 설명함으로써 서비스에 대한 사용자의 기대치를 설정해야 한다.

14-2 상호작용의 대상을 명확히 설명하는가?

[Yes I No I N/A] ☐ ☐ ☐

- 최근 인공지능 시스템을 의인화함으로써 사용자가 친밀감을 향상시키고 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.

안내서 활용을 위한 체크리스트

워크플로우	요구사항 및 체크리스트	Yes | No | N/A
1 계획 및 설계	**요구사항 01** 인공지능 시스템에 대한 위험관리 계획 및 수행	
	01-1 인공지능 시스템 생명주기에 걸쳐 나타날 수 있는 위험 요소를 분석하였는가?	☐ ☐ ☐
	01-2 위험 요소를 제거 및 방지하거나 영향을 완화하기 위한 방안을 마련하였는가?	☐ ☐ ☐
2 데이터 수집 및 처리	**요구사항 02** 데이터의 활용을 위한 상세 정보 제공	
	02-1 데이터의 명확한 이해와 활용을 지원하는 상세한 정보를 제공하는가?	☐ ☐ ☐
	02-2 데이터의 출처는 기록 및 관리되고 있는가?	☐ ☐ ☐
	요구사항 03 데이터 강건성 확보를 위한 이상Abnormal 데이터 제거	
	03-1 데이터 이상값Outlier 식별 및 정상·오류 여부를 점검하였는가?	☐ ☐ ☐
	03-2 데이터 공격에 대한 방어 수단을 강구하였는가?	☐ ☐ ☐
	요구사항 04 수집 및 가공된 학습 데이터의 편향 제거	
	04-1 데이터 수집 시, 인적·물리적 요인으로 인한 편향 완화 방안을 마련하였는가?	☐ ☐ ☐
	04-2 학습에 사용되는 특성Feature을 분석하고 선정 기준을 마련하였는가?	☐ ☐ ☐
	04-3 데이터 라벨링 시, 발생 가능한 편향을 확인하고 방지하였는가?	☐ ☐ ☐
	04-4 편향 방지를 위해 데이터 분포 검증을 통한 데이터 샘플링을 수행하였는가?	☐ ☐ ☐
3 인공지능 모델 개발	**요구사항 05** 오픈소스 라이브러리의 보안성 및 호환성 확보	
	05-1 오픈소스 라이브러리의 보안성 및 호환성 확보 여부를 확인하였는가?	☐ ☐ ☐
	요구사항 06 인공지능 모델의 편향 제거	
	06-1 모델 편향을 제거하는 기법을 적용하였는가?	☐ ☐ ☐
	요구사항 07 인공지능 모델 공격에 대한 방어 대책 수립	
	07-1 모델 추출 공격Model extraction attack에 대한 방어 기법을 도입하였는가?	☐ ☐ ☐
	요구사항 08 인공지능 모델 명세 및 출력 결과에 대한 설명 제공	
	08-1 인공지능 모델의 예측 결과를 설명하기 위한 기법 적용에 대한 검토를 하였는가?	☐ ☐ ☐
	08-2 팩트 시트Fact sheet를 통해 인공지능 모델의 명세를 투명하게 제공하는가?	☐ ☐ ☐
	요구사항 09 인공지능 모델 출력에 대한 신뢰도Confidence value 제공	
	09-1 신뢰도 제공이 필요한 인공지능 모델 출력 결과에 대한 신뢰도를 제공하는가?	☐ ☐ ☐
	09-2 신뢰도가 낮을 경우, 적절한 조치방안을 마련하였는가?	☐ ☐ ☐

안내서 활용을 위한 체크리스트

워크플로우	요구사항 및 체크리스트	Yes | No | N/A
4 시스템 구현	**요구사항 10** 인공지능 시스템 구현 시 발생 가능한 편향 제거	
	10-1 소스 코드 및 사용자 인터페이스로 인한 편향을 제거하기 위해 노력하였는가?	☐ ☐ ☐
	요구사항 11 인공지능 시스템의 안전 모드 구현	
	11-1 공격, 성능 저하 및 사회적 이슈 등의 문제 발생 시 안전 모드를 적용하는가?	☐ ☐ ☐
	11-2 인공지능 시스템에서 문제가 발생할 경우 리포팅을 수행하는가?	☐ ☐ ☐
	요구사항 12 인공지능 시스템의 설명에 대한 사용자의 이해도 제고	
	12-1 인공지능 시스템 사용자의 특성User characteristics과 제약사항을 분석하였는가?	☐ ☐ ☐
	12-2 사용자 특성에 따른 충분한 설명을 제공하는가?	☐ ☐ ☐
5 운영 및 모니터링	**요구사항 13** 인공지능 시스템의 추적가능성 확보	
	13-1 인공지능 시스템의 의사결정에 대한 추적 및 대응 방안을 수립하였는가?	☐ ☐ ☐
	13-2 학습 데이터의 변경 이력을 주기적으로 관리하고 있는가?	☐ ☐ ☐
	13-3 학습 데이터의 업데이트 이력을 주기적으로 관리하고 있는가?	☐ ☐ ☐
	요구사항 14 서비스 제공 범위 및 상호작용 대상에 대한 설명 제공	
	14-1 인공지능 서비스의 올바른 사용을 유도하기 위한 설명을 제공하는가?	☐ ☐ ☐
	14-2 상호작용의 대상을 명확히 설명하는가?	☐ ☐ ☐

참고문헌

- 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.
- ETSI GR SAI 005 V1.1.1, "**Securing Artificial Intelligence (SAI 005); Mitigation Strategy Report,**" 2021. 3.
- F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "**Stealing machine learning models via prediction APIs,**" Proceedings of the 25th USENIX Security Symposium, pp. 601-618, 2016. 8.
- Google AI Blog. **Equality of Opportunity in Machine Learning.** [Online]. Available: https://ai.googleblog.com/2016/10/equality-of-opportunity-in-machine.html
- Google. **People + AI Guidebook - Errors + Graceful Failure**. [Online]. Available: https://pair.withgoogle.com/chapter/errors-failing/#section3
- Google. **People + AI Guidebook - Explainability + Trust.** [Online]. Available: https://pair.withgoogle.com/chapter/explainability-trust/
- Google. **People + AI Guidebook - Mental Models.** [Online]. Available: https://pair.withgoogle.com/chapter/mental-models/
- Google. **People + AI Guidebook - User Needs + Defining Success.** [Online]. Available: https://pair.withgoogle.com/chapter/user-needs/
- Google. **Responsible AI Practices - Google AI.** [Online]. Available: https://ai.google/responsibilities/responsible-ai-practices/?category=interpretability
- IBM Cloud Paks. **Fairness metrics overview.** [Online]. Available: https://www.ibm.com/docs/en/cloud-paks/cp-data/2.5.0?topic=openscale-fairness-metrics-overview
- ISO 31000, "**Risk management - Guidelines,**" 2018. 2.
- ISO/IEC TR 24027, "**Bias in AI systems and AI aided decision making,**" 2021. 9.
- 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, 2021. 2.
- Microsoft. **Guidelines for Human-AI-Interaction**. [Online]. Available: https://www.microsoft.com/en-us/research/project/guidelines-for-human-ai-interaction/
- R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. "**Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization,**" Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 618-626, 2017. 10.
- 서울특별시. **서울시 유니버설디자인 통합 가이드라인**. [Online]. Available :https://opengov.seoul.go.kr/anspruch/16856750
- 한국지능정보사회진흥원(NIA), **인공지능 학습용 데이터셋 구축 안내서**, 2021. 2.

PART 3

검증항목

목 차

목 차

목 차

1 계획 및 설계

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

E-01 인공지능 시스템의 위험 요소를 도출하고 이의 파급효과를 파악하였는가?

[Yes I No I 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 요소를 도출해야 한다. 이러한 요소의 분류와 개략적인 내용은 ISO/IEC 23894.2와 ISO/IEC 24028에 제시되어 있다.
- 도출된 위험 요소별로 이를 야기할 수 있는 원인과 이로 인해 발생 가능한 결과를 분석해야 한다. 발생 가능한 결과란 사회적으로 부정적인 영향을 미칠 수 있는 현상 및 사고를 의미하며, 인체에 위해를 가하는 사고 및 사회적 문제를 야기할 수 있는 차별적인 현상 등이 이에 해당한다.
- 위험 요소의 발생으로 인한 결과는 심각도와 발생빈도와 같은 척도를 기준으로 위험의 크기 또는 수준을 평가할 수 있다. 이는 위험 요소의 파급효과를 의미한다. 위험 요소의 평가를 통해 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

- 인공지능 시스템에서 인식해야 하는 위험에 대하여 ISO/IEC 23894.2의 Appendix A에서는 아래와 같은 키워드를 제시하고 있다. 이러한 키워드를 바탕으로 위험 요소를 구체화 및 세분화해 도출하여야 한다.
  - 공평성Fairness, 안전성Safety, 강건성Robustness
  - 보안성Security, 개인정보보호Privacy
  - 투명성Transparency, 설명가능성Explainability
  - 환경 영향Environmental impact
  - 책무성Accountability, 유지보수성Maintainability, 가용성Availability, 데이터 품질Data Quality
  - 전문성Expertise
- 또한, ISO/IEC 24028:2020에서는 인공지능 시스템 구현 관점에서 고려할 수 있는 취약점 및 위험 요소를 정리해 놓았으므로 (Chapter8. 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

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

E-02 위험 요소 제거 방안을 도출하고 파급효과가 감소하였는지 확인하였는가?

[Yes I No I 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 24028에 제시되어 있다.
- 앞서 위험 요소를 분석하는 과정에서 위험 요소의 파급효과를 평가하였는데, 파급효과가 가장 큰 위험 요소를 우선순위로 대응 방안을 적용하여야 한다.
- 대응 방안이 적용된 이후에 파급효과를 재평가함으로써 위험 요소가 실제로 제거, 방지 혹은 이의 영향이 완화되었는지 확인하여야 한다.

참고 인공지능 시스템의 위험 요소와 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 대응방안(Chapter 9. Mitigation measures)이 개괄적으로 제시되어 있으므로 이를 참고할 수 있다.

Chapter 8 Vulnerabilities, threats and challenges	Chapter 9 Mitigation measures
8.1 General	9.1 General
8.2 AI specific security threats	9.2 Transparency
8.3 AI specific privacy threats	9.3 Explainability
8.4 Bias	9.4 Controllability
8.5 Unpredictability	9.5 Strategies for reducing bias
8.6 Opaqueness	9.6 Privacy
8.7 Challenges related to the specification of AI systems	9.7 Reliability, resilience and robustness
8.8 Challenges related to the implementation of AI systems	9.8 Mitigating system hardware faults
8.9 Challenges related to the use of AI systems	9.9 Functional safety
8.10 System and Hardware faults	9.10 Testing and evaluation
	9.11 Use and applicability

2 데이터 수집 및 처리

요구사항 02-1 데이터의 명확한 이해와 활용을 지원하는 상세한 정보를 제공하는가?

E-03 정제 전과 후의 데이터 특성을 설명하였는가?

[Yes I No I N/A] ☐ ☐ ☐

- 데이터 정제작업은 라벨링 작업 전 학습 데이터 구축을 위한 데이터의 선별 및 처리 단계로서, 정제 과정을 거친 데이터만을 사용하는 사용자는 원시 데이터Raw data의 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

E-04 학습 데이터와 메타데이터Metadata를 구분하였으며, 각각에 대한 명세자료를 확보하였는가?

[Yes I No I 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.

E-05 보호변수Protective attribute의 선정 이유 및 반영 여부를 설명하였는가?

[Yes I No I 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 인종차별과 같은 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.

E-06 라벨링 작업자를 위해 교육을 시행하고 작업 가이드 문서를 마련하였는가?

[Yes I No I N/A] ☐ ☐ ☐

- 데이터 라벨링 작업은 인공지능 모델을 학습하기 위한 원시 데이터Raw data의 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).

요구사항 02-2 데이터의 출처는 기록 및 관리되고 있는가?

[Yes I No I N/A]

E-07 오픈소스 데이터셋을 활용하는 경우, 출처를 명시하였는가?

☐ ☐ ☐

- 인공지능 모델 학습에 오픈소스 데이터셋을 사용한 경우, 학습 시점에는 발견되지 않았던 오류나 편향된 출력이 발생될 수 있다. 또한, 편향된 출력은 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 관리하여야 한다.

요구사항 03-1

데이터 이상값Outlier 식별 및 정상·오류 여부를 점검하였는가?

E-08 전체 학습용 데이터 분포를 시각화하여 발생 가능한 오류들을 확인하였는가?

[Yes I No I 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.

시각화 기법 분류	설명
히스토그램 도표	데이터를 변수에 대한 히스토그램 형태로 시각화한다.
커널 밀도 추정 도표	하나 혹은 두 개의 변수에 대한 밀도 추정 그래프 형태로 데이터를 시각화한다.
경험적 누적 분포함수 도표	전체 데이터의 누적 분포를 시각화한다.
러그 도표	x/y축을 따라 눈금을 그려 주변 분포도를 표시하는 도표로, 다른 도표를 보완하는 데 주로 같이 사용된다.

E-09 학습 데이터 이상값 식별 기법을 적용하였는가?

[Yes I No I N/A] ☐ ☐ ☐

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

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	중앙값(Q2)으로 데이터를 두 부분으로 나누고, 다시 왼쪽 중앙값(Q1)과 오른쪽 중앙값(Q3)으로 나누어 총 4개의 범위를 정하며 사분위수 범위(Q3-Q1)을 구해 해당 범위를 벗어나면 이상값으로 판별한다.

요구사항 03-2 데이터 공격에 대한 방어 수단을 강구하였는가?

E-10 데이터 중독Poisoning, 회피Evasion 등 공격에 대한 방어 대책을 마련하였는가?

[Yes I No I N/A] ☐ ☐ ☐

- 적대적 공격을 방어하고 인공지능 서비스의 강건성을 높이기 위한 다양한 방어 기법이 존재한다. 특히 데이터 설계 및 모델 학습 단계에서의 회피 공격과 중독 공격 방어를 위한 대표적 기법으로는 적대적 학습Adversarial training, Gradient Masking, Feature Squeezing 등이 있다.

방어 기법 분류	방어 기법 내용
적대적 학습	가장 직관적으로 쉽게 떠올려볼 수 있는 알고리즘이다. 모델을 학습시킬 때, 적대적 사례로 활용할 수 있는 모든 경우의 수를 미리 고려하여 학습 데이터셋에 포함시키는 것이다. 그러나 충분한 수와 다양성이 보장된 적대적 데이터를 생성하는 과정 없이는 적대적 학습은 그 성능을 보장하기 어렵다.
Gradient Masking /Distillation	대부분의 적대적 공격은 모델 추론 과정에서의 경사Gradient를 관찰함으로써 이루어진다는 점에 착안하여, 학습 모델의 경사가 출력으로서 그대로 노출되는 것을 방지하거나Gradient masking, 학습 모델의 구조상 경사 자체를 일종의 정규화 방법과 같이 두드러지지 않게 하여 적대적 공격의 학습 방향에 힌트를 주지 않도록 하는 방법Distillation들이 제안되었다.
Feature Squeezing	적대적 공격을 막기 위한 방법으로는 본래의 학습 모델과 별도로, 주어진 입력이 적대적 사례인지 아닌지를 판단하는 학습 모델을 추가하는 방법이 있다. 그 외에 다수의 학습 모델을 앙상블Ensemble하여 시스템을 구성하면, 특정 모델에 대한 화이트박스 공격을 피할 수 있음은 물론이고, 다수의 학습 모델에 범용적으로 적용되는 적대적 공격은 개발이 어렵다는 점으로 인해 좋은 방어법이 된다는 연구 결과가 제시되었다.

요구사항 04-1 데이터 수집 시, 인적·물리적 요인으로 인한 편향 완화 방안을 마련하였는가?

E-11 인적 편향을 제거하기 위한 절차적, 기술적 수단을 적용하였는가?

[Yes I No I 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.

E-12 데이터의 다양성 확보를 위해 이기종 수집 장치를 활용하였는가?

[Yes I No I N/A] ☐ ☐ ☐

- 특정 하드웨어 및 장비를 사용하여 데이터를 수집하는 경우, 수집 환경 및 제약조건으로 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 데이터 활용을 위해서는 데이터의 일관성이 유지되어야 하므로 데이터 정제 및 검수가 충분히 이뤄져야 한다.

E-13 하드웨어로 인해 발생할 수 있는 데이터의 편향을 점검하였는가?

[Yes I No I N/A] ☐ ☐ ☐

- 데이터 수집 및 생성 시 카메라, 마이크 등과 같은 하드웨어 혹은 장비를 사용하게 된다. 이때 장비의 사양 및 수집 환경 같은 물리적 요인으로 인해 제한된 상황 및 시나리오에 대한 데이터만 수집되는 등의 편향이 발생할 수 있다.
- 따라서, 데이터 수집 시 이러한 요인을 점검, 대처하는 계획을 마련해야 한다. 아래의 표는 영상 촬영 장비를 이용해 영상 이미지를 수집하는 경우의 예시이다.

대처 방안	내용
촬영 시나리오 작성	• 촬영 방법 및 목적에 맞는 촬영 시나리오 작성 • 시나리오별 촬영 일정표 작성, 촬영업체 섭외 등 촬영 계획
영상 촬영 시나리오 검수	• 촬영 담당자와 촬영 시나리오 및 시나리오에 따른 세부 세항 점검
라벨링 목적에 맞는 데이터 수집 환경 구축	• 촬영 장소 선정 (획득 환경 구축 및 출입 불가 구역 등 확인) • 촬영 준비 확인 (예: 카메라, 조명, 피사체) 및 시나리오에 따른 촬영 환경 세팅 (예: 화각, 구도, 화질, 촬영 필요항목)
데이터 수집 및 추가로 획득해야 할 정보 파악	• 미리 정해놓은 촬영 기법 (예: 촬영 렌즈, 초점 거리, 프레임, 밸런스, 해상도) 및 촬영 방법 (예: 촬영 각도, 거리, 비율, 품질, 수량)에 맞춰 촬영 진행 • 목적별로 추가 획득해야 할 정보 확인 (예: 데이터 규모, 길이, 결과물 포맷)
수집한 원시 데이터 적절성 검수	• 획득한 원시 데이터에 대해 사전 계획된 목적에 맞게 적절하게 수집되었는지 확인 (획득 목적에 부합하지 않거나 낮은 품질 데이터 필터링) • 획득한 데이터의 법적 문제 발생 가능 여부 검수

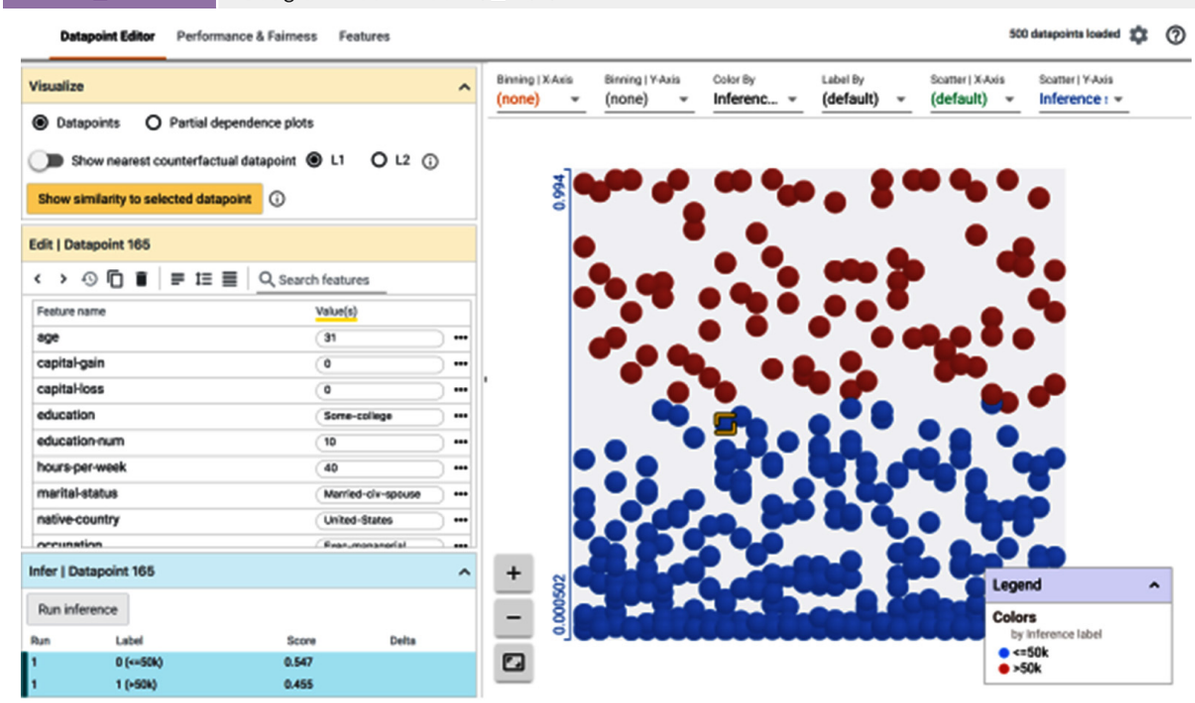
요구사항 04-2 학습에 사용되는 특성Feature을 분석하고 선정 기준을 마련하였는가?

E-14 보호변수Protective attribute 선정 시 충분한 분석을 수행하였는가?

[Yes I No I N/A] ☐ ☐ ☐

- 보호변수 선정 시 충분한 분석을 진행하지 않을 경우, 모델의 성능을 저하시킬 수 있다. 따라서 모델 출력에 영향을 미치는 보호변수가 있는 경우 주어진 데이터셋으로부터 데이터의 일부분을 변경하면서 모델의 결과가 어떻게 변하는지 관찰하고 분석하여야 한다.
- Google What-If Tool의 경우 기계학습 기반 회귀 및 분류 모델에 대하여 데이터 변화에 따른 추론 결과의 변화 추이를 시각화하여 보여주고, 설정한 보호변수가 불공평한 결과에 얼마나 영향을 미치는지, 성능 결과가 어떻게 변하는지 알 수 있다.

참고 Google What-If Tool 화면 예시



2 데이터 수집 및 처리

E-15 편향을 발생시킬 수 있는 특성을 배제하였는가?

[Yes I No I N/A]
☐ ☐ ☐

- 인공지능 모델 학습 시, 데이터의 특성을 선택하여 사용함으로써 효율적인 학습은 물론, 컴퓨팅 자원 및 비용 저감을 할 수 있으며 여러 특성 사이의 관계 분석 과정에서 데이터에 대한 상세한 이해를 통해 잠재된 편향을 인식할 수도 있다.
- 데이터 수집 및 처리 과정에서의 한 예로 특성 선택 기법Feature Selection의 활용과정에서 편향 발생의 최소화를 고려해볼 수 있다. 특성 선택 기법의 종류로는 필터Filter 방법, 래퍼Wrapper 방법, 임베디드Embedded 방법 등이 있다. 이들 방법은 데이터 내 특성들의 통계적 상관관계를 분석하여 높은 상관계수를 갖는 특성을 사용하거나, 특성 일부에 대해 좋은 성능을 갖는 부분 집합Subset을 활용하는 것이다.

E-16 데이터 전처리 시 특성이 과도하게 제거되었는지 검토하였는가?

[Yes I No I N/A]
☐ ☐ ☐

- 특성 선택 기법을 통해서 잠재된 편향을 완화하고 모델 성능을 향상시킬 수 있으나, 지나칠 경우 과적합Overfitting 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.

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 사이의 상호 의존성이 있는가?	만약 그렇다면, 관련 있는 특성을 결합시켜 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 하나?	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 해야할지 아는가?	만약 모른다면, linear predictor를 사용하고, 전진 선택Forward selection 기법이나 0-norm 임베디드 기법을 사용해보는 것이 좋다.
새로운 아이디어와 시간, 컴퓨팅 자원, 충분한 데이터가 있는가?	만약 있다면, 다양한 방법을 시도하는 것이 좋다.
안정적인 솔루션을 원하는가?	만약 그렇다면, 여러 번 해보고 bootstrap을 쓰는 것이 좋다.

요구사항 04-3 데이터 라벨링 시, 발생 가능한 편향을 확인하고 방지하였는가?

E-17 데이터 라벨링을 위한 작업 기준을 명확히 수립하고 작업자에게 제공하였는가?

[Yes I No I 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 충분한 교육을 작업자에게 실시하여 작업자간 편향 발생 여지를 최소화해야 한다.

E-18 다양한 데이터 라벨링 작업자를 섭외하기 위해 노력하였는가?

[Yes I No I N/A] ☐ ☐ ☐

- 데이터 라벨링 단계에서 인적 편향을 줄이려면 다수의 데이터 라벨링 작업자를 확보하고, 인구 통계적 특성 및 배경지식 등이 다양하고 고르게 분포되도록 구성하는 것이 바람직하다.
- 작업자의 다양성을 검증하기 위해서는 크게 2가지를 확인해야 한다. 첫째, 크라우드소싱Crowdsourcing과 같은 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

E-19 다양한 데이터 라벨링 검수자를 확보하기 위해 노력하였는가?

[Yes I No I 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 형성되어 있는지 점검한다.

요구사항 04-4 편향 방지를 위해 데이터 분포 검증을 통한 데이터 샘플링을 수행하였는가?

E-20 편향 방지를 위한 샘플링 기법을 적용하였는가?

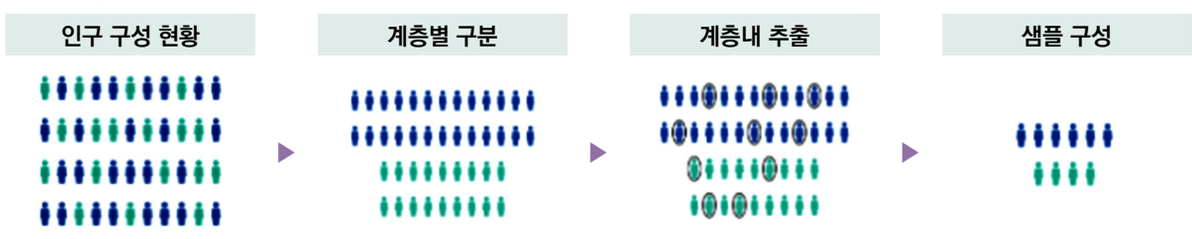
[Yes I No I N/A]
☐ ☐ ☐

- 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 확인하여야 한다.

참고 샘플링 기법 예시 – 층화추출법Stratified Sampling

- 층화추출법은 인구 통계학적 요소에 대한 편향을 방지하기 위해 데이터 추출 전 모집단을 구성하고 있는 계층 혹은 클래스별로 모집단을 분할하여 계층별로 추출을 시행한다.

층화추출법 단계



- 층화추출법을 적용하면서 ① 인구 구성 계층을 어떻게 나눴는지, ② 구분 기준은 무엇인지, ③ 계층별 추출 비율은 동일한지, ④ 총 샘플 구성 결과를 확인해야 한다.
- 계층별 구분 기준 및 추출 비율 등의 세부적인 수치는 인공지능을 활용해 구현하고자 하는 서비스·기술, 다루고자 하는 데이터셋에 포함된 정보에 따라 다르게 책정될 수 있으며, 추출을 수행하는 담당자는 이에 대한 근거를 마련하여야 한다.

3 인공지능 모델 개발

요구사항 05-1 오픈소스 라이브러리의 보안성 및 호환성 확보 여부를 확인하였는가?

[Yes I No I N/A]

E-21 사용 중인 오픈소스 라이브러리의 라이선스, 보안취약점, 호환성을 확인하였는가? ☐ ☐ ☐

- 오픈소스 라이브러리는 버전 변경에 따라 법률 및 기술적 측면의 이슈가 발생할 수 있으므로 사용 중인 버전을 관리하고, 아래와 같은 이슈를 검토하여야 한다.
- 법률적 측면: 라이선스License 확인하기
  - ✓ 오픈소스는 무료로 사용할 수 있지만, 라이선스는 별도로 규정되어 있으므로 사용할 오픈소스의 라이선스 종류 및 라이선스 고지문을 확인하여 허용 또는 의무사항을 우선적으로 숙지하여야 한다.
- 기술적 측면: 호환성Compatibility 및 보안취약점Vulnerability 확인하기
  - ✓ 라이브러리의 버전 변경 과정에서 개발 환경, 언어, 도구 및 다른 라이브러리 버전과 호환되지 않는 문제를 초래할 수 있다. 따라서 라이브러리 간 의존성Dependency를 파악하는 등, 호환성을 고려하여 오픈소스 라이브러리 종류 및 버전을 선택하여야 한다.
  - ✓ 사용 중인 오픈소스 라이브러리에서 보안취약점이 발견되기도 한다. 보안취약점에 따른 영향을 최소화하기 위해 보안취약점 및 버전 변경에 따른 릴리즈 노트Release note를 지속적으로 확인하여 신속히 탐지 및 대응해야 한다.

요구사항 06-1 모델 편향을 제거하는 기법을 적용하였는가?

E-22 개발하려는 모델에 맞게 편향제거 기법을 선택하였는가?

[Yes I No I N/A] ☐ ☐ ☐

- 인공지능 모델에 의한 편향을 완화하기 위한 기법은 기본적으로 3가지 방식으로 구분된다. 모델 학습 전 Pre-processing, 학습 과정 중 In-processing, 그리고 학습 이후 Post-processing에 적용하는 방식이다.
- 각 방식의 특성과 구현하려는 인공지능 모델 및 목표 임무에 맞게 적절한 기법을 선택하여 적용해야 한다.

기법	기법구분			설명 및 지표
	Pre	In	Post	
가중치 재지정	✓	✓		학습 데이터셋 샘플에 가중치를 할당하는 방식
라벨링 재지정	✓			학습용 데이터 샘플의 라벨을 수정하는 방식
변수 블라인딩	✓			분류기가 민감한 변수에 반응하지 않도록 하는 방식
변형	✓		✓	숫자 데이터 기반 학습 시 데이터 변환 및 모델 예측 분포를 변환하는 방식
샘플링	✓			학습 데이터 내 샘플링을 통해 편향을 제거하는 방식
정규화		✓		분류 시 편향에 많은 영향을 주는 클래스 분포를 대상으로 보정하는 방식
제약 최적화		✓	✓	분류기의 손실 함수에 보정값을 부여하는 방식
임계값			✓	추론 결과가 결정 경계값에 가까울 때 편향을 제거하는 방식
보정			✓	긍정 예측 비율이 긍정적인 데이터 인스턴스의 비율과 동일하게 분포하도록 설정하는 방식

E-23 편향성 평가 및 모니터링을 위한 정량적 지표를 선정하고 관리하는가?

[Yes I No I N/A] ☐ ☐ ☐

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

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 공평성
개인Individual 공평성 지표	일반화 엔트로피 지수, 셰일 지수

요구사항 07-1

모델 추출 공격Model extraction attack에 대한 방어 기법을 도입하였는가?

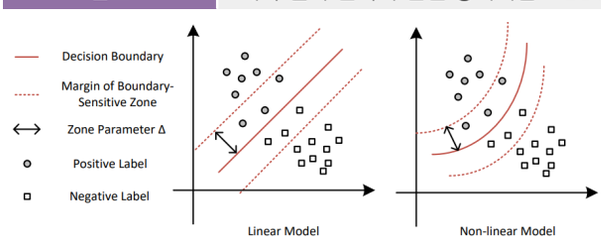
E-24 모델 추출 공격에 대비하는 방어 기법을 적용하였는가?

[Yes I No I N/A]
☐ ☐ ☐

- 인공지능 모델 공격에 대한 주요 완화 방법은 특정 시간 간격당 인공지능 서비스에 대한 질의의 수를 제한하는 것부터 의심스러운 질의에 대한 탐지와 함께 경고 및 예측 결과의 난독화Obfuscation와 같은 적절한 대응 과정을 수행하는 것이다.

방어 기법 분류	방어 기법 내용
질의Query 횟수 제한	특정 기간 내에 수행할 수 있는 질의 수를 제한하여 모델 공격을 위한 반복적인 질의를 방어하는 기법
학습기반 모니터링	머신러닝을 활용하여 모델 공격에 대해 사전 탐지 및 경고 알림, 상응하는 방어 기법을 실행하는 등 능동적으로 방어를 하는 기법
예측 결과 난독화	예측 결과가 결정경계에 가까운 경우 예측 결과의 정확도를 임의로 낮춰 모델의 세부 속성에 대한 추출을 방해하는 기법

참고 예측 결과 난독화 관련 방어기법 – BDPLBoundary Differentially Private Layer



BDPL 기법은 분류를 결정짓는 기준 및 그 주변의 영역을 Boundary Sensitive Zone으로 지정하고 이 영역을 보호하여 외부에서 들어오는 입력이 민감한 영역에 가까울수록 결과에 잡음을 섞어 모델을 추출하기 어렵게 만든다.

요구사항 08-1 인공지능 모델의 예측 결과를 설명하기 위한 기법 적용에 대한 검토를 하였는가?

E-25 필요 시, 모델 출력에 대한 설명을 제공하는가?

[Yes I No I N/A] ☐ ☐ ☐

- 인공지능 서비스는 복잡한 작업을 수행하기 위해 내부적으로 다수의 인공지능 모델을 동시에 사용하거나, 최종 출력 결과를 얻기 위해 사용자와 시스템이 여러 단계를 거쳐 상호작용할 수 있다.
- 인공지능 시스템은 사전에 학습된 모델을 사용하고, 모델 설계에 따라 입출력 구조 등이 성능에 큰 영향을 미치기 때문에 이러한 인공지능 서비스는 원활한 서비스를 위해 단계적으로 사용자에게 적합한 설명이 필요한 경우가 있다.

E-26 사용자가 출력 결과를 수용할 수 있도록 출력 결과에 대한 근거를 제공하는가?

[Yes I No I N/A] ☐ ☐ ☐

- 딥러닝 기술을 활용한 인공지능 시스템의 경우, 우수한 성능을 나타내지만 설명가능성은 낮다. 낮은 설명가능성은 모델 예측에 대한 확신과 시스템 전체에 대한 신뢰도를 낮출 수 있어 사용자가 납득 가능한 모델 출력 결과의 근거를 확보하여야 한다.
- 인공지능 모델 출력 결과의 근거를 확보하는 방안으로 모델의 입력 요인, 모델 내부, 설명변수, 대리모델 분석을 통한 설명 가능한 인공지능XAI, eXplainable AI 기술의 도입을 고려할 수 있으며, 이를 통해 사용자에게 모델 추론 및 출력 결과의 근거를 시각화하여 제시할 수 있다.
- XAI 기술은 인공지능 모델의 설명가능성 확보를 위해 연구가 계속 진행되고 있으며, 기술 도입 전 설명 대상 인공지능 모델 및 XAI 기법의 장·단점 분석을 통해 설명에 적합한 기법을 선택하는 것이 중요하다. 아래는 대표적인 XAI 기술의 방법론 및 장단점의 요약이다.

구분	지표		
모델 비종속적 설명 (Model-agonistic; 블랙박스에 대한 귀납적 추론 기반)	LIME	모델과 무관하게 입력과 출력만으로 확인 가능 (모델 공개가 되지 않을 때 비교적 유용한 접근법이며 대리모델 기반 방법론이 공유하는 장점), 특정 샘플에 대해 설명이 쉬워 실무 적용에 적합	설명 단위가 그때그때 달라짐, 입력과 출력만을 간접적으로 설명할 뿐 인공지능 모델에 대해 설명하지 않으므로 모델의 본질을 설명할 수 없음
모델 종속적 설명 (Model-specific; 매개변수, 아키텍처에 대한 지식 기반)	LRP	직관적이며 은닉층 내부의 기여도를 확인할 수 있어, 해당 은닉층이 무엇을 감지했을지 알아볼 수 있음	기여도를 히트맵Heatmap으로 표현하는 것으로는 신경망 모델이 학습한 추상적 개념을 알 수 없음
	Explorative sampling considering the generative boundaries of DGNN	복잡한 생성 모델에 쓰인 격자의 성격을 각 격자 사이에 있는 샘플들을 통해 어림짐작 가능	여러 샘플을 보고 판단하는 과정에서 모호한 변화와 이를 나누는 경계를 언어로 표현하기 어렵거나, 표현해도 예제 기반 설명의 특성상 분석자의 편향이 개입될 수 있음
	Rule extraction	신경망을 이해하기 쉬운 Decision Tree 순서도 형식으로 변형	신경망 모델을 축약하는 과정에서 모델의 정보를 누락할 수밖에 없음

[Yes I No I N/A]

E-27 설명 가능한 인공지능XAI, eXplainable AI 기술 적용이 어려운 경우, 대안을 마련하였는가? ☐ ☐ ☐

- 인공지능 시스템의 서비스에 따라 XAI 기술 적용이 어려울 경우, 개발자는 인공지능 시스템의 설명가능성과 신뢰도를 높이기 위한 차선책을 고려해야한다. 이러한 경우, 실제 시스템의 유효성을 검증하고, 검증에 대한 분석 결과를 설명함으로써 시스템의 신뢰성을 확보할 수 있다.
- 이를 위해 기존 소프트웨어 개발에 사용된 여러 테스트 기법을 인공지능 시스템에 적용해 볼 수 있으며, 아래와 같은 시스템 유효성 검증 방법들을 복합적으로 사용하여 유효성을 검증할 수 있다.
  - ✓ 프로덕션 환경에서의 반복성 테스트 수행
  - ✓ 반사실적 공평성 테스트 수행
  - ✓ 예외 식별 및 예외 처리 조치 이행 테스트 수행
  - ✓ 데이터에 따른 인공지능 모델의 적절성 테스트 수행
  - ✓ 단일 혹은 다양한 복잡 조건에 따른 반복성 테스트 수행
  - ✓ 대상 모집단의 서로 다른 하위 그룹에 대한 인공지능 모델 오류율 테스트 수행

요구사항 08-2 **팩트 시트Fact sheet를 통해 인공지능 모델의 명세를 투명하게 제공하는가?**

[Yes I No I N/A]

E-28 **시스템 개발 과정과 모델 작동 방식에 대한 세부 정보가 설명된 문서를 작성하였는가?** ☐ ☐ ☐

- 인공지능 시스템의 투명성을 높이고 시스템 사용자가 인공지능 기반 프로그램 구성 요소를 파악할 수 있는 정보를 제공하는 것은 시스템 신뢰성을 높이는데 중요한 요소이다. 이를 위해 인공지능 모델 개발 과정에서 팩트 시트Fact sheet를 작성할 경우, 사용자에게 인공지능 시스템의 구성 요소를 파악할 수 있는 정보를 제공할 수 있다.
- 팩트 시트 작성 시에는 인공지능 생명주기에 관련된 이해관계자들을 고려하여 각자 필요한 정보를 선택하여 확인할 수 있도록 관련 정보를 포함하여야 한다. 다음은 이해관계자에 따른 팩트 시트 내 필요 정보 예시이다.

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

요구사항 09-1

신뢰도Confidence value 제공이 필요한 인공지능 모델 출력 결과에 대한 신뢰도를 제공하는가?

E-29 신뢰도 제공이 필요한지 검토하였는가?

[Yes I No I N/A] ☐ ☐ ☐

- 인공지능 시스템이 도출한 결과에 대한 신뢰도를 보여주는 것은, 사람들이 인공지능을 활용하여 의사결정함에 있어 도움이 될 수 있지만, 오히려 방해될 수도 있다. 따라서 모든 결과에 대해 신뢰도를 제공하기보다는, 신뢰도가 꼭 제공되어야 하는지를 확인하는 과정이 선행되어야 한다.
- 신뢰도를 제공하지 않는 편이 더 나은 경우에 대한 두 가지 예시는 다음과 같다.
  - ✓ 첫째, 신뢰도 제공 자체가 사용자의 의사결정에 크게 영향을 미치지 않을 것으로 판단되는 경우이다. 신뢰도가 미치는 영향이 명확하지 않을 경우, 신뢰도를 더욱 세분화해 제공하면 사용자의 의사결정에 더 도움이 될 것으로 생각할 수 있지만, 예상과는 다르게 혼란을 초래할 수 있다. 예를 들어, 인공지능 시스템이 도출한 두 가지 결과가 있고, 각각의 신뢰도가 85.8%, 87.0%라면, 사용자는 어떤 결과를 활용하여 의사결정을 할지 혼란스러울 수 있다.
  - ✓ 둘째, 신뢰도가 너무 높거나 낮은 경우에도 신뢰도를 제공하지 않는 것이 낫다. 만약 시스템의 출력 결과에 대해 신뢰도가 100%라고 사용자에게 알릴 경우, 사용자가 시스템의 출력 결과를 맹목적으로 수용하게 만들 수 있다.

E-30 신뢰도를 계산하고, 계산 결과를 기반으로 모델의 신뢰 수준을 정의하였는가?

[Yes I No I N/A] ☐ ☐ ☐

- 신뢰도를 정의하기 위해서는 정밀도Precision, 재현율Recall, mAPmean Average Precision와 같은 지표와 함께 불확실성을 계산해야 한다. 불확실성이란 확률 변수의 분산 크기로, 인공지능 모델이 도출한 결과를 얼마나 확신하는지 나타내는 지표이다. 불확실성 추정 기법에는 베이지안 신경망Bayesian neural network, 앙상블Ensemble, 드롭아웃Dropout 등이 있다.
  - ✓ 드롭아웃은 신경망 내 노드와 각 노드 간 연결을 무작위로 선정하여 제거하는 기법을 말한다.
  - ✓ 드롭아웃을 적용한 신경망과 베이지안 신경망 모두 각각의 수행마다 다른 신경망이 생성된다는 특징을 활용해보면, 수행의 결과로 생성된 여러 신경망에 동일한 입력값을 주고, 그 결과로 얻은 여러 개의 출력값에 대해 평균과 분산을 계산할 수 있는데, 이때 계산한 분산이 불확실성이다.
- 인공지능 모델의 출력 성능(예: 정밀도, 재현율)과 불확실성을 각각 계산해 나온 결과를 조합하여 신뢰 수준을 정의할 수 있다. 예를 들어, 참과 거짓의 예측 모델이 있을 때 '모델의 예측 확률이 98%로 높고, 예측에 대한 불확실성이 1%로 낮으므로, "참"이라는 결과를 신뢰할 수 있다.'라는 근거를 사용자에게 제시할 수 있다.

E-31 모델 성능의 임계치를 도출하고, 임계치 이하일 경우 신뢰도를 제공하는가?

[Yes I No I N/A]
☐ ☐ ☐

- 출력 결과의 임계치란 인공지능 모델의 성능지표(예: 정밀도, 재현율)에 대한 임계치 그리고 그 성능의 불확실성에 대한 임계치로 나누어 볼 수 있다. 출력 결과의 임계치 도출을 위해 우선, 인공지능 모델로 인해 발생 가능한 문제 상황을 정의하고, 문제 발생 여부를 결정 짓는 중요 변수를 파악해야 한다. 여기서 문제 상황이란 사용자의 생명이나 재산과 관련된 위협이 되는 상황뿐 아니라 기대하는 또는 유지되어야 하는 품질 수준보다 낮은 상황 등을 모두 포함한다.
- 임계치를 찾는 것은 인공지능 모델을 통해 가능하다. 대표적인 기법인 LDA(Linear Discriminant Analysis), SVM(Support Vector Machine) 등 부터 CNN(Convolutional Neural Network), LSTM(Long-Short Term Memory)부터 비교적 최근 발표된 GEN(Graph Extrapolation Network), SimCLR(Simple framework for Contrastive Learning of visual Representations) 등에 이르기까지 다양한 기법으로 출력 결과의 임계치를 도출할 수 있다.

요구사항 09-2 신뢰도가 낮을 경우, 적절한 조치방안을 마련하였는가?

E-32 모델 출력의 신뢰 수준이 임계치 이하일 경우 사용자에게 추가 설명을 제공하는가?

[Yes I No I N/A] ☐ ☐ ☐

- 인공지능 시스템의 결과가 특정 제약으로 인해 기대한 만큼의 신뢰 수준을 보이지 못하는 경우, 사용자에게 후속조치 정보를 제공해야 한다. 여기서 후속조치 정보란 신뢰도가 낮을 시 사용자의 다음 행동은 어떻게 되어야 하는지 구체적으로 제공하는 정보를 의미한다.
- 사용자 후속조치 정보를 제공하는 방식에는 N-best 대안 방식, 수치형 지표 방식, 그래프 기반 시각화 방식, 범주형 시각화 방식 등 다양한 형태가 있다. 각 방식의 장단점이 다르므로 어떠한 분야의 인공지능 시스템인지, 주 사용자 특성은 어떠한지 등을 고려하여서 가장 적합한 방식을 선정하고, 사용자가 이해하기 쉽도록 정보를 제공해야 한다.

E-33 모델 성능이 허용 임계치 이하일 경우 이해관계자에게 경고하는 기능을 개발하였는가?

[Yes I No I N/A] ☐ ☐ ☐

- 인공지능 모델의 성능에는 정확도, 정밀도, 재현율, F1-score 등과 더불어 객체 검출 또는 인식에 걸리는 시간, 사용자에게 제공하는 설명의 이해 용이성 등이 있다. 성능 저하는 모델 성능이 허용 임계치 이하인 상황을 뜻하는데, 이를 방지하기 위해 지속적인 모니터링과 함께 대책 마련이 필요하다.
  - ✓ 모델 성능의 허용 임계값은 예를 들어, 95% 이상의 정확도와 1% 미만의 불확실성 정도를 보여야 하는 인공지능 모델이 있을 때 정확도 95%, 불확실성 1%, 즉, 일정 수치 이상 또는 이하일 때 문제 상황의 발생 우려가 있는 값을 의미함
- 이해관계자에게 경고하는 기능을 구현하는 방법으로는 MLOps[Machine Learning model Operationalization management]가 있다. MLOps는 개발[Development]과 운영[Operation]의 합성어인 DevOps, 즉 소프트웨어 제품 및 서비스에 대한 빠른 개발·배포·대응을 목적으로 개발 조직과 운영 조직 간 소통 및 협업을 강조하는 환경 또는 문화가 인공지능 시대에 맞춰 변화한 결과라고 할 수 있다. MLOps를 구축함으로써 문제 상황에 대해 빠른 대응이 가능하다. 또한 자동화 기능을 구현한다면 모델 성능 저하 시 즉시 경고 알림을 받을 수 있다.

4 시스템 구현

요구사항 10-1 소스 코드 및 사용자 인터페이스로 인한 편향을 제거하기 위해 노력하였는가?

E-34 데이터 접근 방식 구현과정 등 소스 코드에서의 편향 발생 가능성을 확인하였는가? [Yes I No I N/A] □ □ □

- 인공지능 시스템은 모델에서 활용할 데이터에 접근하는 방식이 코드 상에 구현되는 과정에서 특정 클래스 접근이 누락되는 등 다양한 형태의 편향이 발생할 수 있다.
- 특히 규칙 기반 시스템Rule-based system에서 특정 분야에 경험이 많은 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 배경지식과 경험이 다양한 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 공평성 평가지표에 따라 기능의 위험 여부를 알리는 등의 기능을 수행한다. 이 도구들을 활용함으로써 구현과정에서 편향을 빨리 발견하고 대응할 수 있다.

E-35 사용자 인터페이스User Interface 및 인터랙션Interaction 방식으로 인한 편향을 확인하였는가? [Yes I No I N/A] □ □ □

- 인공지능 시스템은 사용자 인터페이스에 의한 암묵적인 유도 또는 사용자의 의도적인 오남용에 따라 사용자 인터랙션 편향이 발생할 수 있다,
- 사용자 인터랙션 편향을 방지하기 위해서는 사용자 인터페이스 설계 및 구현 시 편향 발생 가능성이 있는 요소(예: 표현 편향Presentation bias, 순위 편향Ranking bias)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 결과에 비해 사용자의 선택 빈도가 높을 수 있다.

요구사항 11-1 공격, 성능 저하 및 사회적 이슈 등의 문제 발생 시 안전 모드를 적용하는가?

E-36 문제 상황에 대한 예외 처리 정책이 마련되어 있는가?

[Yes I No I 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 제공하는지 확인해야 한다.

E-37 인공지능 시스템의 보안 강화를 위한 보안 메커니즘을 적용하였는가?

[Yes I No I N/A] ☐ ☐ ☐

- 인공지능 시스템을 개발할 때 격리 및 탐지와 같은 보안 메커니즘을 활용한 인공지능 보안 아키텍처와 구축 솔루션을 적용함으로써 인공지능 데이터 및 모델에 대한 보안성뿐만 아니라 인공지능 시스템의 전반적인 보안성을 확보할 수 있다.

보안 메커니즘	설명 및 예시
격리	의사결정에 활용되는 주요 기능을 모듈 단위로 분리하고 모듈 간 접근제어 메커니즘을 설정하여 인공지능 시스템에 대한 보안성을 확보할 수 있다.
탐지	인공지능 시스템에 대한 공격을 지속적으로 모니터링해 네트워크 보안 상태를 종합적으로 분석하고 현재의 위험 수준을 측정할 수 있다.

E-38 문제 상황 발생 시, 사람의 개입을 고려하는가?

[Yes I No I 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 큰 반면, 인공지능 모델이 도출한 판단 결과의 불확실도가 높은 경우이다.
- 특히, 인공지능 모델을 활용하여 자동 및 자율적으로 운영되는 시스템에서 이러한 경향이 두드러지며, 예외 처리 및 보안 메커니즘 외에, 사람이 직접 혹은 부분적으로 개입하여 인공지능 모델의 불확실도를 해소하는 방안을 고려하여야 한다.
- 예시로, 자율주행자동차 전방의 방해물 객체 인식을 통해 조향하는 인공지능 모델의 인식 결과가 불명확하거나 불확실도가 높은 경우, 운전자의 개입을 유도하고 제어권을 이양하는 기능이 고려되기도 한다.

E-39 예상되는 사용자 오류에 대한 안내 및 대응을 제공하는가?

[Yes I No I N/A]
☐ ☐ ☐

- 사용자 오류는 외적으로는 서비스 최종 결과물을 사용하는 사용자로부터 내적으로는 서비스 결과 생성을 위해 내부 시스템을 사용하는 작업자로부터 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 작업의 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: 사용자로부터 전달받은 입력 등을 재차 확인하고 그에 대한 예상 결과를 미리 전달한다. 또한 잘못된 결과에 대해 실행을 취소하는 등의 기능을 포함시켜 예방할 수 있다.

요구사항 11-2 인공지능 시스템에서 문제가 발생할 경우 리포팅을 수행하는가?

E-40 편견, 차별 등 윤리적 문제에 대한 리포팅 절차를 수립하였는가?

[Yes I No I N/A] ☐ ☐ ☐

- 인공지능 시스템에서 편견 혹은 차별 등의 윤리적 문제의 발생 가능성을 확인하고, 문제 발생 시 이를 위한 리포팅 기능 혹은 절차가 수립되었는지 점검한다.
- 윤리적 문제 리포팅 절차의 경우, 먼저 제공하는 인공지능 시스템에서 자체적인 인공지능 시스템의 신뢰 정도를 평가할 수 있는 기준과 리스트를 만든다. 주요 체크 항목의 예시는 다음과 같다.
  - ✓ 인권보장, 프라이버시 보호, 다양성 존중, 침해금지, 공공성, 연대성, 데이터 관리, 책임성, 안전성, 투명성
- 시스템 자체적인 리포팅 외에도, 시스템 운영 중 사용자가 윤리적 문제를 발견할 경우 시스템 운영자에게 신고할 수 있는 기능도 개발되어야 한다.

E-41 시스템 성능 저하를 평가하기 위한 지표 및 절차를 설정하였는가?

[Yes I No I N/A] ☐ ☐ ☐

- 인공지능 시스템의 경우, 서비스 배포 및 운영 단계에서 일반적인 소프트웨어와 달리 지속적인 데이터 축적, 서비스 기능 확장, 환경의 변화 등과 같은 이유로 성능 저하가 생길 수 있다.
- 인공지능 시스템은 실제 서비스 운영 중 갑자기 성능이 저하됐을 때 원인을 바로 알기 어려우므로, 시스템의 성능 저하를 지속적으로 평가, 관리하기 위한 지표와 절차가 설정되었는지 점검한다.
- 선정할 수 있는 성능지표로는 F1-score, IoU[Intersection over Union], mAP[mean Average Precision] 등이 있다. 평가 결과 성능 저하가 확인되면 이를 시스템 운영자에게 보고하고, 운영자는 성능 저하 원인을 찾아 개선을 진행하는 등의 절차를 마련해야 한다.

요구사항 12-1 인공지능 시스템 사용자의 특성User characteristics과 제약사항을 분석하였는가?

E-42 사용자 특성에 따른 세부 고려사항을 분석하였는가?

[Yes I No I 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.

구분	상세 구분	고려사항
연령	아동, 성인, 노인 등	아동의 경우, 성인과 비교해 이해할 수 있는 어휘, 단어의 한계가 있음
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

요구사항 12-2 사용자 특성User characteristics에 따른 충분한 설명을 제공하는가?

E-43 사용자 특성에 따른 설명 평가 기준을 수립하였는가?

[Yes I No I N/A] ☐ ☐ ☐

- 다양한 사용자가 서비스를 이용하는 만큼 설명을 포괄적으로 평가할 수 있는 특성과 세부 항목을 정하는 단계가 필요하다. 설명의 평가 기준은 구체성, 명확성, 적절성과 같은 항목이 될 수 있다. 세부 항목으로 데이터 유형Data type이나 모달리티Modality에 따라 각 항목에서 고려되어야 할 내용들이 달라질 수 있다. 다음은 설명 평가를 위한 예시이다.

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 이끌어낼 수 있도록 명확한 주어·목적어·동사를 활용해 설명되고 있는가?
적절성	• 주어진 설명이 사용자의 특정 지식수준을 요구하지는 않는가? • 배경지식 혹은 사전 경험이 필요하진 않은가? • 독자를 고려한 전문용어, 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?

E-44 사용자가 이해하기 어려운 전문 용어 사용을 지양하였는가?

[Yes I No I N/A] ☐ ☐ ☐

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

E-45 사용자의 구체적인 행동과 이해를 이끌어낼 수 있도록 명확한 표현을 사용하였는가?

[Yes I No I 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·위험과 같은 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.

E-46 설명이 필요한 위치와 타이밍은 적절한가?

[Yes I No I 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 E-47의 웹로그 분석, A/B 테스트와 같은 사용자 조사 기법을 활용할 수 있다.

E-47 사용자 경험을 평가할 수 있는 다양한 사용자 조사 기법을 활용하였는가?

[Yes I No I 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, 웹로그 분석)
    - 사용자 태도 기반 조사Attitudinal user research: 사용자가 무엇을 말하는지를 조사하는 방법 (예: 카드 소팅Card sorting, 심층 인터뷰, 요구사항 조사)

5 운영 및 모니터링

요구사항 **13-1** 인공지능 시스템의 의사결정에 대한 추적 및 대응 방안을 수립하였는가?

E-48 데이터 소스 변경에 대한 모니터링 방안을 확보하였는가?

[Yes I No I N/A] ☐ ☐ ☐

- 인공지능 모델의 학습 데이터 확보를 위해 웹 크롤링Web crawling 같은 방법을 활용할 수 있다. 웹 크롤링은 관련 오픈소스(예: Apache Nutch, Scrapy)를 통해 대량의 데이터를 빠르게 확보할 수 있는 장점이 있으나, 크롤링 대상이 되는 웹 페이지의 데이터 소스가 실시간으로 변경되거나 대상 페이지 자체의 접속이 불가능한 장애가 있을 경우 특정 클래스의 데이터 부족과 같은 수집 데이터의 분포가 깨질 수 있다.
- 특히 지속적으로 크롤링된 데이터를 실시간으로 학습하는 인공지능 시스템의 데이터 소스의 변경은 성능에 직접적인 영향을 줄 수 있다. 따라서 데이터 수집 과정의 모니터링을 통해 데이터 소스 이상이나 중복 수집과 같은 문제에 대응할 수 있어야 한다.

E-49 인공지능 시스템의 의사결정에 대한 기여도 추적 방안은 확보하였는가?

[Yes I No I N/A] ☐ ☐ ☐

- 인공지능 시스템의 결정에 대한 모델 기여도를 파악하기 위해서는 이전 모델의 출력 정보와 최종 결정에 대한 사람(예: 시스템 운영자, 사용자) 개입 여부 등의 정보가 추적되어야 한다.
- 이를 위해서는 인공지능 모델이 전적으로 의사결정을 내리는 경우와 모델 결과를 사람이 검토하여 의사결정을 내리는 경우, 주로 사람이 의사결정을 내리지만 특정 이벤트와 같이 보조적으로 모델의 출력이 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.

E-50 인공지능 시스템의 의사결정 추적을 위한 로그 수집 기능을 구현하였는가?

[Yes I No I 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.

오류 구분	오류 원인 예시
모델 구축 방법 측면의 오류	모델·데이터의 대상선정, 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

5 운영 및 모니터링

E-51 지속적인 사용자 경험 모니터링을 위해 사용자 로그를 수집 및 관리하고 있는가?

[Yes I No I 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 지속적으로 수집되며 서비스 고도화에 따라 다양한 형태로 누적될 수 있다.
- 서버 인프라에 대한 로그를 통해 서비스 운영 상태에 대한 모니터링을 수행할 수 있으며, 사용자 인터랙션 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 기업 자체적으로 인터페이스 또는 인터랙션의 호출에 따른 로그를 수집하거나 로그 분석 도구를 활용할 수 있다.

요구사항 13-2 학습 데이터의 변경 이력을 주기적으로 관리하고 있는가?

E-52 데이터 변경 시, 버전관리를 수행하였는가?

[Yes I No I N/A] ☐ ☐ ☐

- 인공지능 모델 개발 과정에서 학습 데이터의 업데이트, 오류로 인한 라벨링 재수행 등과 같이 데이터 변경이 이루어지면 학습 결과인 모델도 변경되게 된다. 따라서 학습 데이터의 변경이 수행될 경우, 단순히 사용된 학습 데이터의 버전뿐만 아니라 해당 버전으로 학습한 인공지능 모델을 함께 관리하여야 한다.
- 이를 위해 머신러닝 프로젝트를 위한 오픈소스 기반의 데이터 버전관리 도구(예: DVC[Data Version Control])의 도입을 고려하거나, 학습 데이터 버전관리 시스템을 자체적으로 구축하여 학습 데이터의 버전과 모델의 버전관리를 수행해야 한다.

E-53 데이터 변경에 대비하여, 이해관계자를 대상으로 한 설명 절차를 수립하였는가?

[Yes I No I N/A] ☐ ☐ ☐

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

이해관계자	제공 정보
비즈니스 결정자	데이터 변경에 따른 모델의 세세한 변경점보다 기존 시스템의 목적, 서비스 의도 등의 변경점이나 시스템 전체의 방향성 등의 초점을 맞춘 설명 필요
데이터 과학자	기존 데이터와 변경된 데이터의 특징, 포맷, 규모 등의 차이점 등에 대한 설명 필요
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 대해 설명 필요
모델 검증자	변경된 테스트 데이터셋 구성, 재설계 및 재학습된 모델에 대한 주요 성능 평가 결과, 기존 모델과의 성능 비교 결과 등에 대한 설명 필요
모델 운영자	검증을 마친 변경 모델에 대한 운영 및 사용자 모니터링 결과 등을 수집 및 분석하여 설명 제공 필요

5 운영 및 모니터링

E-54 데이터 흐름 및 형상Lineage을 추적하기 위한 조치를 구현하였는가?

[Yes I No I N/A]
☐ ☐ ☐

- 인공지능 시스템의 경우, 데이터의 변경으로 인해 모델의 확장이나 재설계 등과 같은 시스템 변경이 발생할 수 있다. 따라서 시스템의 변경을 유도하는 데이터의 흐름 및 형상을 계속해서 추적해야 한다.
- 데이터 흐름은 데이터 변경에 대해 역방향 혹은 순방향, end-to-end 관점으로 나누어 추적할 수 있으며, 추적을 위한 고려사항은 다음과 같다.
  - ✓ 데이터 흐름 및 형상 추적을 위해 메타데이터를 기록하고 유지보수할 것인가?
  - ✓ 데이터 인벤토리Data inventory, 데이터 사전Data dictionary, 데이터 변경 프로세스Data change process 및 문서 제어 메커니즘Document control mechanism을 생성하는 것이 유용한가?
  - ✓ 데이터는 출처까지 역추적될 수 있는가?
  - ✓ APIApplication Programming interface, 데이터베이스, 파일과 더불어 "특성 저장소Feature repository"를 마련하여 데이터 흐름 및 형상을 추적하는 것이 유용한가?
  - ✓ 개발자들은 데이터 내러티브Data narrative 및 데이터 다이어리Data diary를 문서화하고, 사용 데이터의 종류, 수집 방법 및 이유에 대해 명확한 설명을 반드시 제공해야 하는가?
  - ✓ 데이터 흐름 및 형상 추적을 관리하기 위한 데이터 정책팀을 구성하는 것이 유용한가?

5 운영 및 모니터링

요구사항 13-3 학습 데이터의 업데이트 이력을 주기적으로 관리하고 있는가?

E-55 학습용 데이터 중 신규 데이터의 비율을 기록 및 관리하고 있는가?

[Yes I No I N/A] ☐ ☐ ☐

- 학습기반 인공지능 모델은 신규 데이터를 사용하여 성능 테스트를 수행하면 성능이 저하되는 것을 확인할 수 있다. 특히 이전에 학습에 사용한 데이터셋과 특성이 완전 다르거나 데이터셋 전체를 교체할 경우 성능이 크게 저하될 수 있으며, 이 경우에는 추가 학습이 필요할 수 있다.
- 따라서 신규 데이터가 추가될 때 이로 인한 인공지능 모델의 성능 변화 추적을 위해 학습 혹은 테스트에 사용된 신규 데이터의 비율을 기록하고, 그에 따른 모델의 성능 변화를 추적해야 한다.

E-56 신규 데이터 확보 시, 인공지능 모델의 성능평가를 재수행하였는가?

[Yes I No I 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)

5 운영 및 모니터링

요구사항 14-1 인공지능 서비스의 올바른 사용을 유도하기 위한 설명을 제공하는가?

E-57 서비스의 목적과 목표에 대한 설명을 제공하는가?

[Yes I No I 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. 서비스 목적과 목표를 설명함으로써 사용자는 사용 맥락에 맞는 적합한 기능을 선택하여 활용할 수 있다.

참고 YouTube의 서비스 목적 및 목표

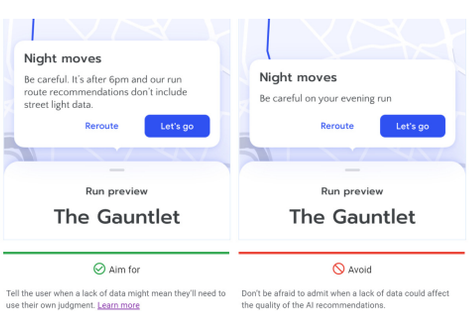
세계 최대의 동영상 스트리밍 플랫폼 YouTube는 별도 웹사이트로 자사 서비스 목적과, 사용자가 원하는 목표에 어떤 원리를 통해 도달하고 있는지 설명하고 있다.

E-58 서비스의 한계와 범위에 대한 설명을 제공하는가?

[Yes I No I N/A]
☐ ☐ ☐

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

참고 Google AI+ 디자인 가이드라인



Google AI+ 디자인 가이드라인에서는 서비스 결과의 품질에 영향을 미칠 수 있는 요인에 대한 설명을 권장하고 있다. 이와 관련, Google 에서는 구체적으로 서비스에 제한사항이 발생한 경우에 대해 명시적으로 알릴 것을 권장하고 있다.

요구사항 **14-2 상호작용의 대상을 명확히 설명하였는가?**

[Yes I No I N/A]

E-59 사용자가 인공지능과 상호작용하고 있음을 명확하게 인지할 수 있도록 안내하는가? ☐ ☐ ☐

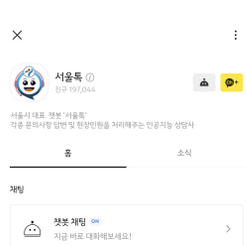
1단계: 의인화를 사용하는지 확인하기

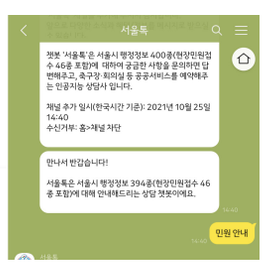
- 의인화란 인간이 아닌 대상을 인간과 유사한 상호작용의 대상으로 만듦으로써 사용성을 높이는 것을 말한다. 따라서 의인화를 활용한 인간과 유사한 상호작용을 하는 경우, 상호작용의 대상이 사람이 아니라는 점을 사용자에게 알림으로써 사용자의 혼선을 예방하고 사용자의 기대치를 조정할 수 있다.

2단계: 서비스 제공 범위에 따른 상호작용 대상 명시하기

- 특히 부분적으로 의인화를 활용할 경우, 어느 범위에서부터 인공지능이 활용되고 있는지 사용자에게 명시하는 것이 중요하다.

참고 '서울톡'의 상호작용 사례





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

자가검증 체크리스트

워크플로우	검증항목	Yes	No	N/A
1 계획 및 설계	01-1 인공지능 시스템 생명주기에 걸쳐 나타날 수 있는 위험 요소를 분석하였는가?	☐	☐	☐
	E-01 인공지능 시스템의 위험 요소를 도출하고 이의 파급효과를 파악하였는가?	☐	☐	☐
	01-2 위험 요소를 제거 및 방지하거나 영향을 완화하기 위한 방안을 마련하였는가?	☐	☐	☐
	E-02 위험 요소 제거 방안을 도출하고 파급효과가 감소하였는지 확인하였는가?	☐	☐	☐
2 데이터 수집 및 처리	02-1 데이터의 명확한 이해와 활용을 지원하는 상세한 정보를 제공하는가?	☐	☐	☐
	E-03 정제 전과 후의 데이터 특성을 설명하였는가?	☐	☐	☐
	E-04 학습 데이터와 메타데이터Metadata를 구분하였으며, 각각에 대한 명세자료를 확보하였는가?	☐	☐	☐
	E-05 보호변수Protective attribute의 선정 이유 및 반영 여부를 설명하였는가?	☐	☐	☐
	E-06 라벨링 작업자를 위해 교육을 시행하고 작업 가이드 문서를 마련하였는가?	☐	☐	☐
	02-2 데이터의 출처는 기록 및 관리되고 있는가?	☐	☐	☐
	E-07 오픈소스 데이터셋을 활용하는 경우, 출처를 명시하였는가?	☐	☐	☐
	03-1 데이터 이상값Outlier 식별 및 정상·오류 여부를 점검하였는가?	☐	☐	☐
	E-08 전체 학습용 데이터 분포를 시각화하여 발생 가능한 오류들을 확인하였는가?	☐	☐	☐
	E-09 학습 데이터 이상값 식별 기법을 적용하였는가?	☐	☐	☐
	03-2 데이터 공격에 대한 방어 수단을 강구하였는가?	☐	☐	☐
	E-10 데이터 중독Poisoning, 회피Evasion 등 공격에 대한 방어 대책을 마련하였는가?	☐	☐	☐
	04-1 데이터 수집 시, 인적·물리적 요인으로 인한 편향 완화 방안을 마련하였는가?	☐	☐	☐
	E-11 인적 편향을 제거하기 위한 절차적, 기술적 수단을 적용하였는가?	☐	☐	☐
	E-12 데이터의 다양성 확보를 위해 이기종 수집 장치를 활용하였는가?	☐	☐	☐
	E-13 하드웨어로 인해 발생할 수 있는 데이터의 편향을 점검하였는가?	☐	☐	☐
	04-2 학습에 사용되는 특성Feature을 분석하고 선정 기준을 마련하였는가?	☐	☐	☐
	E-14 보호변수Protective attribute 선정 시 충분한 분석을 수행하였는가?	☐	☐	☐
	E-15 편향을 발생시킬 수 있는 특성을 배제하였는가?	☐	☐	☐
	E-16 데이터 전처리 시 특성이 과도하게 제거되었는지 검토하였는가?	☐	☐	☐
	04-3 데이터 라벨링 시, 발생 가능한 편향을 확인하고 방지하였는가?	☐	☐	☐
	E-17 데이터 라벨링을 위한 작업 기준을 명확히 수립하고 작업자에게 제공하였는가?	☐	☐	☐
	E-18 다양한 데이터 라벨링 작업자를 섭외하기 위해 노력하였는가?	☐	☐	☐
	E-19 다양한 데이터 라벨링 검수자를 확보하기 위해 노력하였는가?	☐	☐	☐
	04-4 편향 방지를 위해 데이터 분포 검증을 통한 데이터 샘플링을 수행하였는가?	☐	☐	☐
	E-20 편향 방지를 위한 샘플링 기법을 적용하였는가?	☐	☐	☐

자가검증 체크리스트

워크플로우	검증항목	Yes	No	N/A
3 인공지능 모델 개발	05-1 오픈소스 라이브러리의 보안성 및 호환성 확보 여부를 확인하였는가?	☐	☐	☐
	E-21 사용 중인 오픈소스 라이브러리의 라이선스, 보안취약점, 호환성을 확인하였는가?	☐	☐	☐
	06-1 모델 편향을 제거하는 기법을 적용하였는가?	☐	☐	☐
	E-22 개발하려는 모델에 맞게 편향제거 기법을 선택하였는가?	☐	☐	☐
	E-23 편향성 평가 및 모니터링을 위한 정량적 지표를 선정하고 관리하는가?	☐	☐	☐
	07-1 모델 추출 공격Model extraction attack에 대한 방어 기법을 도입하였는가?	☐	☐	☐
	E-24 모델 추출 공격에 대비하는 방어 기법을 적용하였는가?	☐	☐	☐
	08-1 인공지능 모델의 예측 결과를 설명하기 위한 기법 적용에 대한 검토를 하였는가?	☐	☐	☐
	E-25 필요 시, 모델 출력에 대한 설명을 제공하는가?	☐	☐	☐
	E-26 사용자가 출력 결과를 수용할 수 있도록 출력 결과에 대한 근거를 제공하는가?	☐	☐	☐
	E-27 설명 가능한 인공지능XAI, eXplainable AI 기술 적용이 어려운 경우, 대안을 마련하였는가?	☐	☐	☐
	08-2 팩트 시트Fact sheet를 통해 인공지능 모델의 명세를 투명하게 제공하는가?	☐	☐	☐
	E-28 시스템 개발 과정과 모델 작동 방식에 대한 세부 정보가 설명된 문서를 작성하였는가?	☐	☐	☐
	09-1 신뢰도Confidence value 제공이 필요한 인공지능 모델 출력 결과에 대한 신뢰도를 제공하는가?	☐	☐	☐
	E-29 신뢰도 제공이 필요한지 검토하였는가?	☐	☐	☐
	E-30 신뢰도를 계산하고, 계산 결과를 기반으로 모델의 신뢰 수준을 정의하였는가?	☐	☐	☐
	E-31 모델 성능의 임계치를 도출하고, 임계치 이하일 경우 신뢰도를 제공하는가?	☐	☐	☐
	09-2 신뢰도가 낮을 경우, 적절한 조치방안을 마련하였는가?	☐	☐	☐
	E-32 모델 출력의 신뢰 수준이 임계치 이하일 경우 사용자에게 추가 설명을 제공하는가?	☐	☐	☐
	E-33 모델 성능이 허용 임계치 이하일 경우 이해관계자에게 경고하는 기능을 개발하였는가?	☐	☐	☐
4 시스템 구현	10-1 소스 코드 및 사용자 인터페이스로 인한 편향을 제거하기 위해 노력하였는가?	☐	☐	☐
	E-34 데이터 접근 방식 구현과정 등 소스 코드에서의 편향 발생 가능성을 확인하였는가?	☐	☐	☐
	E-35 사용자 인터페이스User Interface 및 인터랙션Interaction 방식으로 인한 편향을 확인하였는가?	☐	☐	☐
	11-1 공격, 성능 저하 및 사회적 이슈 등의 문제 발생 시 안전 모드를 적용하는가?	☐	☐	☐
	E-36 문제 상황에 대한 예외 처리 정책이 마련되어 있는가?	☐	☐	☐
	E-37 인공지능 시스템의 보안 강화를 위한 보안 메커니즘을 적용하였는가?	☐	☐	☐
	E-38 문제 상황 발생 시, 사람의 개입을 고려하는가?	☐	☐	☐
	E-39 예상되는 사용자 오류에 대한 안내 및 대응을 제공하는가?	☐	☐	☐
	11-2 인공지능 시스템에서 문제가 발생할 경우 리포팅을 수행하는가?	☐	☐	☐
	E-40 편견, 차별 등 윤리적 문제에 대한 리포팅 절차를 수립하였는가?	☐	☐	☐
	E-41 시스템 성능 저하를 평가하기 위한 지표 및 절차를 설정하였는가?	☐	☐	☐

자가검증 체크리스트

워크플로우	검증항목	Yes	No	N/A
4 시스템 구현	12-1 인공지능 시스템 사용자의 특성User characteristics과 제약사항을 분석하였는가?	☐	☐	☐
	E-42 사용자 특성에 따른 세부 고려사항을 분석하였는가?	☐	☐	☐
	12-2 사용자 특성User characteristics에 따른 충분한 설명을 제공하는가?	☐	☐	☐
	E-43 사용자 특성에 따른 설명 평가의 기준을 수립하였는가?	☐	☐	☐
	E-44 사용자가 이해하기 어려운 전문 용어 사용을 지양하였는가?	☐	☐	☐
	E-45 사용자의 구체적인 행동과 이해를 이끌어낼 수 있도록 명확한 표현을 사용하였는가?	☐	☐	☐
	E-46 설명이 필요한 위치와 타이밍은 적절한가?	☐	☐	☐
	E-47 사용자 경험을 평가할 수 있는 다양한 사용자 조사 기법을 활용하였는가?	☐	☐	☐
5 운영 및 모니터링	13-1 인공지능 시스템의 의사결정에 대한 추적 및 대응 방안을 수립하였는가?	☐	☐	☐
	E-48 데이터 소스 변경에 대한 모니터링 방안을 확보하였는가?	☐	☐	☐
	E-49 인공지능 시스템의 의사결정에 대한 기여도 추적 방안은 확보하였는가?	☐	☐	☐
	E-50 인공지능 시스템의 의사결정 추적을 위한 로그 수집 기능을 구현하였는가?	☐	☐	☐
	E-51 지속적인 사용자 경험 모니터링을 위해 사용자 로그를 수집 및 관리하고 있는가?	☐	☐	☐
	13-2 학습 데이터의 변경 이력을 주기적으로 관리하고 있는가?	☐	☐	☐
	E-52 데이터 변경 시, 버전관리를 수행하였는가?	☐	☐	☐
	E-53 데이터 변경에 대비하여, 이해관계자를 대상으로 한 설명 절차를 수립하였는가?	☐	☐	☐
	E-54 데이터 흐름 및 형상Lineage을 추적하기 위한 조치를 구현하였는가?	☐	☐	☐
	13-3 학습 데이터의 업데이트 이력을 주기적으로 관리하고 있는가?	☐	☐	☐
	E-55 학습용 데이터 중 신규 데이터의 비율을 기록 및 관리하고 있는가?	☐	☐	☐
	E-56 신규 데이터 확보 시, 인공지능 모델의 성능평가를 재수행하였는가?	☐	☐	☐
	14-1 인공지능 서비스의 올바른 사용을 유도하기 위한 설명을 제공하는가?	☐	☐	☐
	E-57 서비스의 목적과 목표에 대한 설명을 제공하는가?	☐	☐	☐
	E-58 서비스의 한계와 범위에 대한 설명을 제공하는가?	☐	☐	☐
	14-2 상호작용의 대상을 명확히 설명하는가?	☐	☐	☐
	E-59 사용자가 인공지능과 상호작용하고 있음을 명확하게 인지할 수 있도록 안내하는가?	☐	☐	☐

참고문헌

- 120다산콜재단. **챗봇 상담 서울톡**. [Online]. Available: https://www.120dasan.or.kr/dsnc/main/contents.do?menuNo=200019
- 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.
- 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. 6.
- 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.
- 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. 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.
- ISO/IEC CD 23894.2, "**Artificial Intelligence - Risk Management**," 2020. 6.
- 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 (NeurIPS 2020), 2020. 10.
- K. Lerman, and T. Hogg, "**Leveraging position bias to improve peer recommendation**," PloS one, vol. 9, no. 6, 2014. 6.

참고문헌

- 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 H ealth, vol. 18, no. 4, p. 2121, 2021. 2.
- 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, Jul. 2021.
- 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. Meht a, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varsh ney, and Y. Zhang, "**AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigati ng Unwanted Algorithmic Bias**," arXiv, 2018. 10.
- S. Vasudevan and K. Kenthapadi, "**LiFT: A Scalable Framework for Measuring Fairness in ML Applicat ions**," Proceedings of the 29th ACM International Conference on Information & Knowledge Manage ment (CIKM '20), pp. 2773-2780, 2020. 10.
- T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "**A Simple Framework for Contrasive Learning of Vis ual Representations**," Proceedings of the 37th International Conference on Machine Learning, pp. 15 97-1607, 2020. 7.
- World Economic Forum, **Companion to the Model AI Governance Framework**, 2020. 1.
- Y. Gal and Z. Ghahramani, "**Dropout as a Bayesian Approximation: Representing Model Uncertainty i n Deep Learning**," Proceedings of The 33rd International Conference on Machine Learning, Vol. 48, pp. 1050-1059, 2016. 6.
- YouTube. **YouTube 작동의 원리 - 제품 기능, 책임 및 영향력**. [Online]. Available: https://www.youtube.co m/intl/ALL_kr/howyoutubeworks/
- 김휘영, 정대철, 최병욱, "**딥러닝 기반 의료 영상 인공지능 모델의 취약성: 적대적 공격**," 대한영상의학회지, vol. 80, no. 2, pp. 259-273, 2019
- 서울특별시. **서울시 유니버설디자인 통합 가이드라인**. [Online]. Available: https://opengov.seoul.go.kr/ansp ruch/16856750
- 소프트웨어정책연구소. **설명가능한 인공지능(Explainable AI; XAI) 연구 동향과 시사점**. [Online]. Available: htt ps://spri.kr/posts/view/23296?code=industry_trend
- 한국데이터베이스진흥원, **데이터 품질진단 절차 및 기법 (Ver 1.0)**, 데이터 품질관리 시리즈 4, 2009. 10.

약어표

AI	Artificial Intelligence
ALTAI	Assessment List for Trustworthy Artificial Intelligence
API	Application Programming interface
BDPL	Boundary Differentially Private Layer
CNN	Convolutional Neural Network
EC	European Commission
ETSI	European Telecommunications Standards Institute
EU	European Union
GEN	Graph Extrapolation Network
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
SimCLR	Simple framework for Contrastive Learning of visual Representations
SVM	Support Vector Machine
TAI	Trustworthy AI
TR	Technical Reports
WEF	World Economic Forum
WIT	What-If Tool
XAI	eXplainable AI

2022 신뢰할 수 있는 인공지능 개발 안내서(안)

과학기술정보통신부 이 재 형 과장
박 예 슬 사무관

한국정보통신기술협회 차 순 일 단장
신 준 호 팀장
곽 준 호 책임
김 민 정 책임
조 경 우 책임
황 재 영 책임
신 예 진 선임
송 채 빈 연구원

발행년월 2022.01
발행인 최 영 해
발행처 한국정보통신기술협회
편집·제작 ㈜디자인여백플러스