2024

신뢰할 수 있는 인공지능

개발 안내서

채용 분야

과학기술정보통신부
Ministry of Science and ICT

TTA 한국정보통신기술협회
Telecommunications Technology Association

일러두기

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

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

- 본 안내서는 인공지능 서비스 및 제품을 개발하는 과정에서 참고 자료로 활용하도록 편찬했습니다. 기업의 업무 환경과 상황, 개발 목적 등을 고려하여 필요한 내용을 취사선택하여 활용하기 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 바립니다.

- 본 안내서의 인공지능 동향 및 기술 정보는 2023년 12월 기준으로 서술했습니다.

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

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

CONTENTS

안내서 활용을 위한 체크리스트

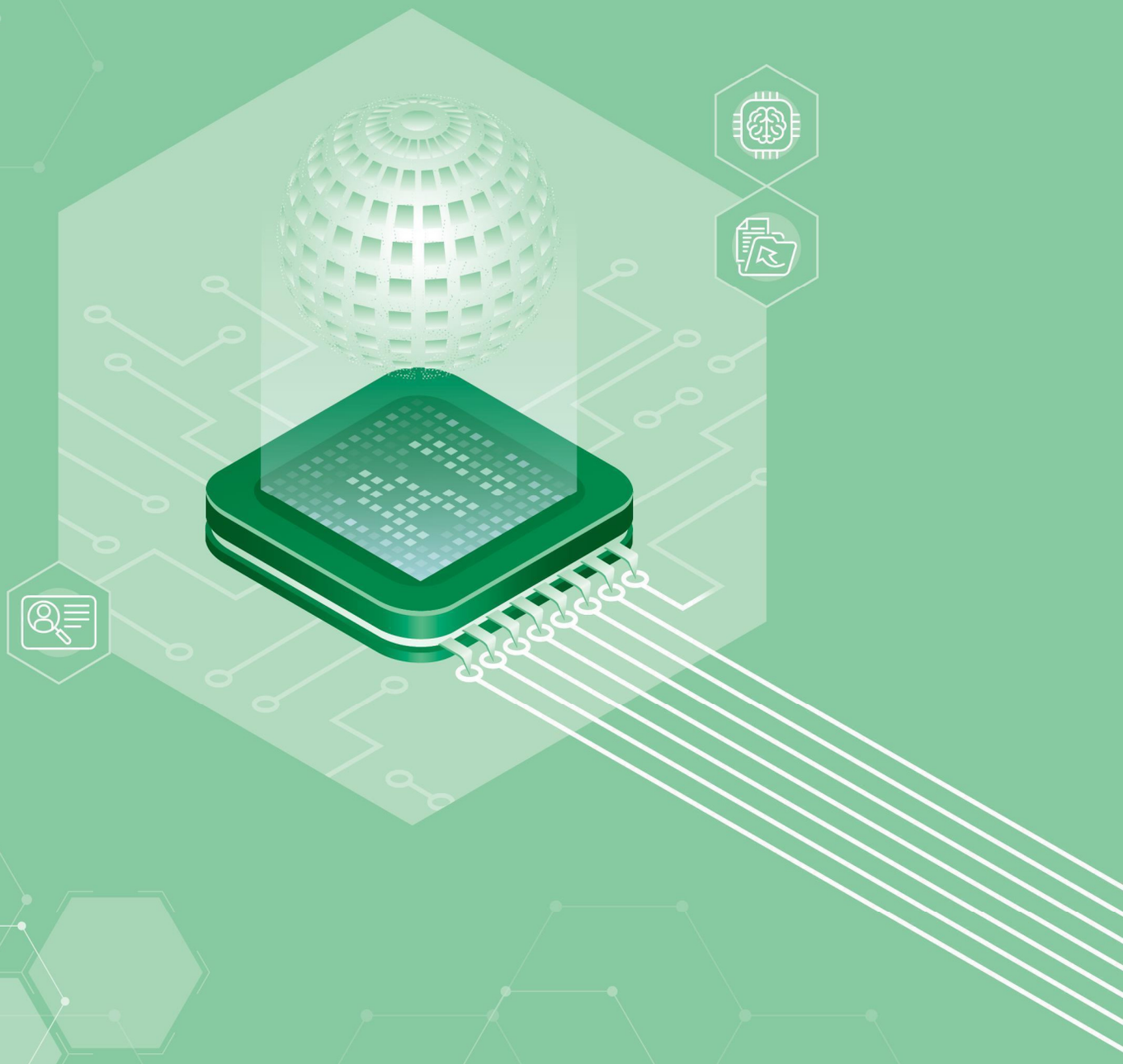
안내서 활용을 위한 체크리스트

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?	☐	☐	☐

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?	☐	☐	☐

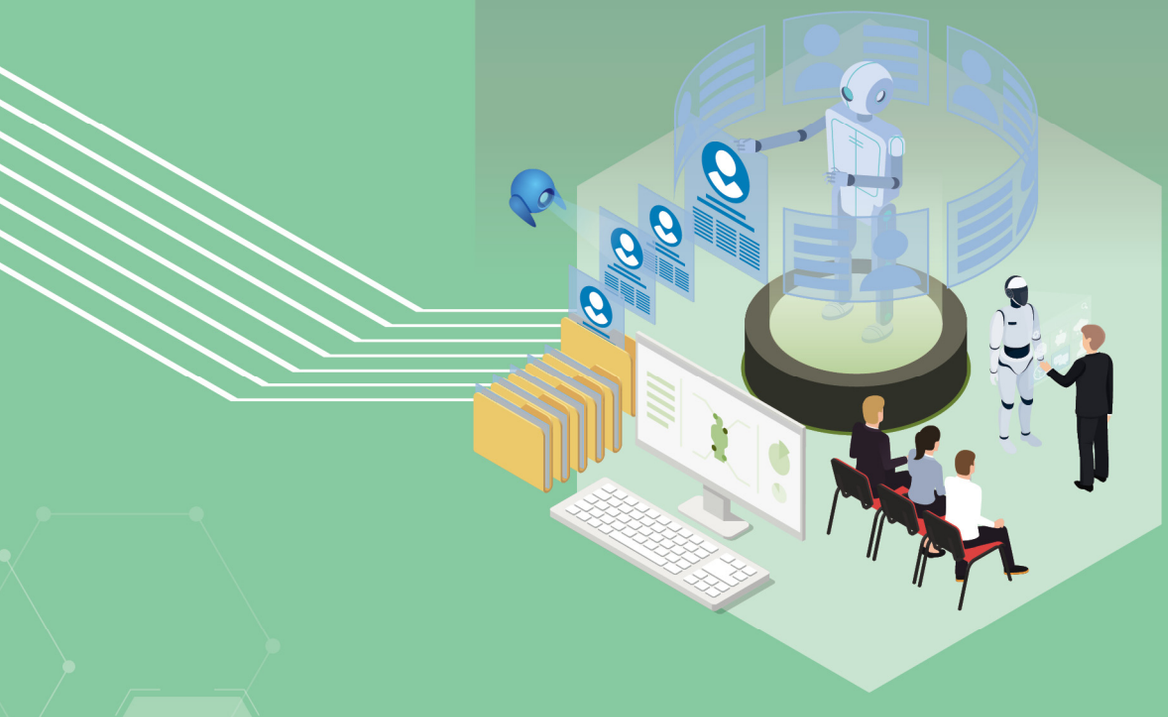
생명주기	요구사항 및 체크리스트		Yes	No	N/A
2 데이터 수집 및 처리	07-3	데이터 라벨링 시, 발생 가능한 편향을 확인하고 방지하였는가?	☐	☐	☐
	07-3a	데이터 라벨링 기준을 명확히 수립하고 작업자에게 제공하였는가?	☐	☐	☐
	07-3b	다양한 데이터 라벨링 작업자를 섭외하기 위해 노력하였는가?	☐	☐	☐
	07-3c	다양한 데이터 라벨링 검수자를 확보하기 위해 노력하였는가?	☐	☐	☐
	07-4	데이터의 편향 방지를 위한 샘플링을 수행하였는가?	☐	☐	☐
	07-4a	편향 방지를 위한 샘플링 기법을 적용하였는가?	☐	☐	☐
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?	☐	☐	☐



PART 1

개요



01 안내서 발간 배경 및 목적

인공지능은 채용 프로세스의 효율성, 객관성, 효과성에 대한 절실한 요구와 고유한 특성으로 인해 채용 및 면접 평가 분야에 없어서는 안 될 도구가 되었다. 인공지능은 빠른 데이터 분석, 패턴 인식, 대량의 정보를 신속하게 처리하는 능력과 특성이 있다. 이는 채용 절차의 다양한 측면을 간소화하고 개선하는 데 활용된다. 특히, 이력서의 초기 심사를 자동화하고, 지원자의 기술과 자격을 심층적으로 분석하며, 특정 직무와의 적합성까지 평가할 수 있다.

구직자에게 채용 시장은 복잡하고 경쟁이 치열하며, 구인자는 인재 채용에 많은 비효율이 발생하고 있어 인공지능의 도입이 각광을 받고 있다. AI는 입사지원서의 수에 관계없이 효율적으로 지원자를 분류하고 평가한다. 또한, 자격과 기술 등 입사 지원자의 자료를 일관적이고 객관적으로 평가하여 인간 평가자의 편향이 개입될 가능성을 줄인다. 채용 분야에서 AI를 활용하여 객관성을 확보하는 것은 채용 기업이나 HR 담당자에게도 유용하다. 인공지능은 채용 절차를 간소화하고 조직이 직무에 가장 적합한 지원자를 확보하는 중요한 조력자로 부상하고 있다.

인공지능은 이제 채용 인공지능 기술 영역을 포함해 다양한 분야에 영향을 미치고 있다. '행정기본법'(2021년 3월), '전자정부법'(2022년 1월) 등 공공 의사결정 과정에서 AI의 역할을 강화하는 입법이 추진되면서 공공기관의 AI 도입이 가속화되고 있다. 그러나 인공지능을 사용할 경우 데이터 및 알고리즘 편향으로 인한 잠재적 차별, 개인정보 침해, 인권침해가 발생할 수 있다는 우려가 제기되고 있다. 이러한 문제를 해결하기 위해 EU는 《인공지능 기반 서비스 및 솔루션의 공공 조달 데이터 윤리에 관한 백서》(2020년 5월)에서 합법성, 윤리, 사회적 안전을 강조하며 인공지능의 윤리적 사용을 위한 핵심 요소를 설명하였다.

국내에서도 이와 유사한 이니셔티브가 등장했는데, 개인정보 보호법(PIPA)은 개인의 사생활과 개인정보를 보호하기 위해 마련한 중요한 법안이다. 개인정보 보호법의 목표는 개인의 존엄과 가치를 지키면서 개인의 자유와 권리를 보호하는 것이다. PIPA가 적용된 사례로는 '이루다 사건'[1]이 대표적이다. 이 사건은 인공지능 시스템과 기술 맥락에서 개인정보 보호가 어떻게 시행되는지에 대한 큰 시사점을 제공함은 물론 국내 인공지능 거버넌스에 중요한 의미를 지닌다. 대화 텍스트와 같은 비정형 데이터에 대한 엄격한 가명처리의 필요성과 정보주체의 동의 없이 수집한 개인정보 이용의 한계를 실감할 수 있는 사례이다.

또한, 전 세계적으로 투명성, 공정성, 책임성을 강조하는 'AI 권리장전'[19]과 'AI 법'[20]과 같은 이니셔티브가 제안 및 제정되어 AI 시스템이 내린 결정의 책임에 무게를 더하고 있다. 윤리적 거버넌스에 대한 원칙 수립은 채용 인공지능 시스템의 신뢰성 확보의 기본이다. 시스템은 핵심 원칙을 준수하면서 설계, 개발, 운영하여야 한다. 이는 AI 기반 채용 프로세스에서 발생할 수 있는 잠재적인 편향, 차별 또는 피해를 완화하는 중추적인 역할을 한다. 윤리적 거버넌스를 구축함으로써 조직은 AI 시스템의 신뢰성을 뒷받침할 뿐만 아니라 책임감 있고 사회적 의식이 있는 기술을 실현할 수 있다.

따라서 본 개발 안내서는 추상적인 윤리적 원칙과 실제 구현의 간극을 좁히고 AI 서비스, 특히 채용 인공지능에 특화하여 최소한의 신뢰성과 윤리기준을 충족하도록 돕는 실용적인 자료를 제공한다. 책임감 있고 신뢰할 수 있는 개발로 채용 인공지능의 안정성을 확보할 수 있기를 희망한다.

02 채용 인공지능 신뢰성 동향

개발자와 고위 경영진 등 인공지능 제품 및 서비스 개발에 관여하는 전문가들은 본 개발 안내서에 제시된 매뉴얼을 준수함으로써 채용 분야의 신뢰성 확보에 도움을 받을 수 있다. 개발 프로세스 전반에 걸쳐 인사(HR) 전문가와의 협업이 가장 중요하다. 개발자는 본 개발 안내서에서 제공하는 지침을 활용하여 AI 서비스에 맞는 정확한 요구사항을 확인하고 검증 방법을 수립하여 궁극적으로 신뢰할 수 있는 솔루션을 개발할 수 있다. 본 개발 안내서가 채용 인공지능 분야의 국내 기업 및 기관의 관련 기술, 시스템, 서비스의 신뢰성을 강화는 기초 자료로 활용될 수 있기를 기대한다.

2.1 채용 인공지능 신뢰성 동향

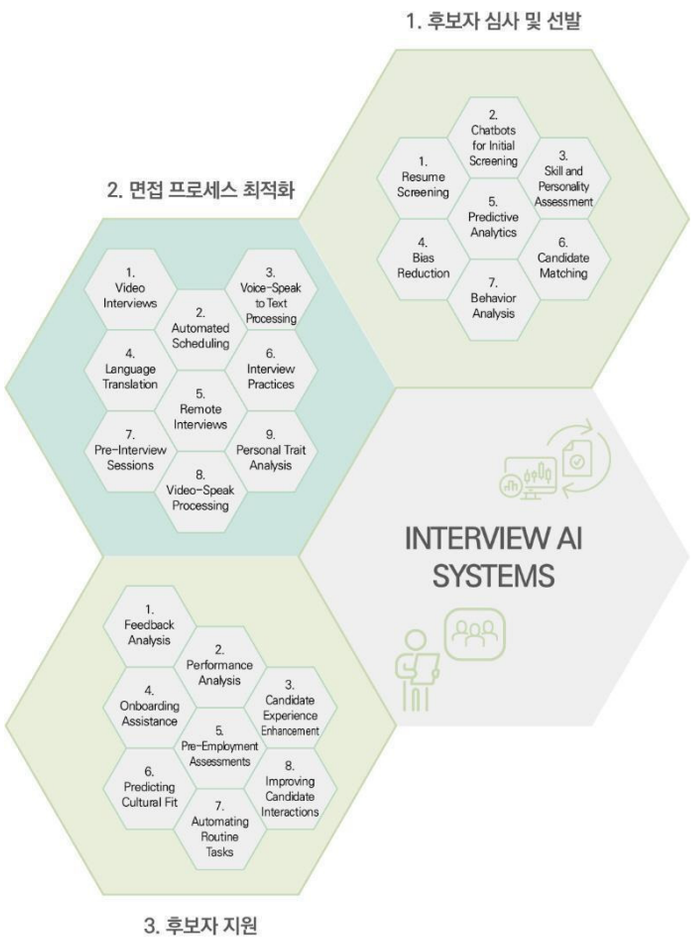
채용 분야에서 인공지능의 영향력은 막대하다. 인공지능은 시간이 오래 걸리는 작업을 자동화하여 채용 프로세스를 가속화하고, 조직이 우수한 인재를 더 빠르게 확보할 수 있도록 지원한다. 또한, 채용의 편향을 줄여 다양성과 포용성을 높이는 데 기여할 수도 있고, 운영 비용을 절감하여 재정적으로도 도움이 된다. 결과적으로 채용 담당자는 AI가 일상적인 프로세스를 처리하는 동안 전략적 업무에 인적·물적 자원을 집중할 수 있다. 하지만 이러한 변화를 위해서는 공정성, 투명성, 데이터 보안을 보장하기 위한 체계적인 관리가 필요하다. 채용 분야에서 체계적인 관리를 통해 인공지능의 신뢰성을 확보하는 절차가 있다면 구직자와 구인자 모두 채용 인공지능에 대한 긍정적인 경험을 할 수 있다. 채용에 신뢰할 수 있는 AI를 도입한다면 보다 효율적이고 일관적인 채용 방식으로 나아가게 될 것이다.

2.2 채용 인공지능 활용 영역

채용 과정에 적용되는 인공지능, 즉 채용 인공지능은 다양한 산업과 분야에 걸쳐 광범위하게 활용되고 있다. MSPowerUser의 조사에 따르면 글로벌 기업의 88%가 채용을 포함한 HR$^{human\ resources}$ 분야에서 어떤 형태로든 인공지능 기술을 사용 중이다[2]. 인공지능은 효율적이고 객관적인 데이터 기반 의사결정을 도입하여 전통적인 채용 프로세스를 변화시키고 있다. 후보자 심사 및 선발, 면접 프로세스 최적화, 후보자자 지원 등에 사용되어 채용 담당자와 구직자 모두에게 종합적인 솔루션을 제공한다. 기업 이니셔티브 외에도 교육기관의 입학 절차 등 대량의 지원자를 평가 또는 선발하는 데 채용 인공지능을 활용하고 있다. 의료 분야에서는 AI가 레지던트 면접과 병원 직원 채용에, 정부 기관에서는 공무원 시험과 공공 부문 채용 등에 활용하고 있다. 또한, 비영리단체와 사회적 기업에서도 잠재적인 자원봉사자와 직원을 평가하기 위해 인공지능 기반 면접을 도입하고 있으며, 다양한 영역에서 채용 인공지능 기술의 활용이 증가하는 추세이다.

채용 인공지능은 특정 분야에 국한되지 않고 채용 프로세스의 품질과 효율성을 향상시켜 직무에 가장 적합한 지원자를 찾아낸다. 채용 인공지능 애플리케이션은 주요 기능과 목적에 따라 다음과 같은 몇 가지 분야로 분류할 수 있다.

▼ 채용(지원자 평가) 분야에서의 인공지능 활용 분야[4]



① **후보자 심사 및 선발:** 인공지능 시스템은 지원자 심사 및 선발을 위한 알고리즘을 통해 이력서와 지원서를 분석하여 직무와 관련된 주요 자격, 기술, 경험을 파악한다. 과거 데이터와 직무 요건을 비교함으로써 지원자의 직무 적합성을 예측하여 효율적으로 최종 후보를 선정할 수 있다. 자동화된 채점 및 순위 시스템을 통해 채용 담당자는 유력한 지원자에게 집중할 수 있게 되며, 자연어 처리(NLP) 기술은 지원자 텍스트를 이해하고 분류하는 데 사용되어 지원자와 직무 설명을 쉽게 매칭하도록 한다. 이를 통해 시간을 절약할 수 있을 뿐만 아니라 지원자 선발의 정확성도 높일 수 있다. 지원자 심사 및 선발 절차는 크게 7가지로 나뉜다: 이력서 심사, 초기 심사를 위한 챗봇, 기술 및 인성 평가, 편견 완화, 직무 적합성 예측 분석, 지원자 매칭, 행동 분석

② **면접 프로세스 최적화:** 인공지능은 다양한 방식으로 면접 프로세스를 최적화한다. 지원자와 면접관의 면접 일정을 조정하고, 프로세스를 간소화한다. 가상 면접 인공지능은 자동화된 설문지 및 비언어적 단서 분석 같은 기능으로 지원자의 종합적인 평가를 제공한다. 직무에 맞게 AI가 생성한 면접 질문은 관련성을 높여줄 뿐만 아니라, 이를 통해 면접에 집중할 수 있도록 도와준다. 또한, AI 시스템은 면접관의 피드백을 분석하여 전반적인 면접 프로세스를 개선함으로써 지원자 평가의 품질과 효율성을 높인다. 이러한 면접 프로세스 최적화는 크게 9가지로 나뉜다: 화상 면접, 자동 일정 예약, 음성-텍스트 처리, 언어 번역, 원격 면접, 면접 연습, 사전 면접 세션, 화상-음성 처리, 개인 특성 분석

③ **후보자 지원**: 채용 인공지능 시스템은 채용 프로세스 전반에 걸쳐 후보자도 지원한다. 면접 준비 팁, 잠재적인 질문, 면접 중 지원자의 성과를 개선하기 위한 제안을 제공하여 면접 준비를 돕는다. 면접 후에는 지원자의 강점과 개선이 필요한 부분 등 피드백을 제공하여 지원자가 자신의 성과를 더 잘 이해할 수 있게 한다. 또한, AI 기반 챗봇은 지원자의 질문에 답하고, 회사 및 면접 프로세스 정보도 제공한다. 일부 AI 시스템은 지원자의 이력서 개선안을 제공하기도 한다. 지원자 지원 절차는 8가지로 나뉜다: 피드백 분석, 성과 분석, 지원자 경험 향상, 온보딩 지원, 채용 전 평가, 문화적 적합성 예측, 일상적인 작업 자동화, 지원자 상호작용 개선

2.3. 채용 인공지능 이슈 사례

인공지능을 이용해 입사 지원자를 평가하고 선별하는 것에 대한 끊임없는 이슈와 논란이 사회적 우려를 불러일으키고 있다. 채용 인공지능 시스템은 효율성과 자동화 측면에서 이점이 있지만, 평가의 공정성과 정확성, 인간 개입의 필요성에 대한 지속적인 우려가 있다.

▼ 채용 인공지능에 대한 사회적 우려

구분	채용 인공지능에 대한 사회적 우려
기술적 제한[5][6]	인공지능 채용 시스템의 기술적 한계가 심각한 문제로 대두되고 있다. 이 시스템이 지원자의 직무 적합성을 항상 정확하게 평가하지 못할 수 있어 그 효과와 신뢰성에 우려를 불러일으키고 있다.
인간의 회의주의[7]	구직자들은 AI 채용의 공정성과 정확성에 의구심을 표하고 있다. AI 평가에 따라 지원자가 직무에 '부적합'한 것으로 평가되는 경우가 있어 편향이나 오류 가능성에 우려가 제기되기도 하였다.
새로운 트렌드[8][9]	논란에도 불구하고 국내에서도 채용에 AI를 활용하는 사례가 증가하고 있다. 롯데 같은 기업에서는 AI 시스템을 도입해 자기소개서 평가에 인사 담당자와 함께 의견을 제시한다. 전문가들은 기술이 고도화되면서 채용에서 AI의 역할이 더욱 확대될 것으로 전망한다.

국내[10]에서는 자기소개서뿐만 아니라 이력서 및 입사지원서 전체를 검토하여 면접 대상자를 선별하는 데에도 인공지능이 활용되고 있다. 국내 채용 및 구인구직 서비스 기업들은 국내 기업을 위한 인재 분석 시스템을 개발하였다. 채용 분야의 윤리적이고 공정한 평가 기준에 대한 사회적 요구에 따라 기업들은 [11], [12] 등 독자적인 공정성 평가 기준을 활용한 채용 인공지능 솔루션을 제공한다. 현재 700개 이상의 기업이 채용 절차에 이러한 '인공지능 분석 도구'와 시스템을 도입하고 있다. 이 도구들은 지원자의 이력서 표절을 탐지하고, 화상 면접 시 표정, 제스처, 목소리, 억양, 면접 답변 등 다양한 측면을 분석하는 데 활용된다. 많은 기업이 채용 과정에서 '인재 스크리닝 소프트웨어'를 활용하고 있지만 이로 인해 의도치 않게 과도한 필터링이 발생하기도 한다. 예를 들면 간호사를 선발하는 데 '고객 서비스', 수리 기술자를 선발하는 데 '프로그래밍' 경험 같은 지나치게 구체적인 기준으로 지원자를 추천할 수 있다. AI가 빅데이터에 의존하기 때문에 자격을 갖춘 지원자를 의도치 않게 배제할 수 있어서 발생하는 문제이다.

우리나라는 채용 분야에 AI 시스템을 도입하면서 채용 프로세스에 큰 변화를 겪고 있다. 속도와 효율성 측면에서 장점이 있지만, 개발된 채용 인공지능 시스템이 인간을 대체하기에는 시기상조라는 우려와 함께 공정한 채용 프로세스를 보장할 윤리 가이드라인이 필요하다는 목소리도 나오고 있다.

2.4. 채용 인공지능 신뢰성 정책 및 연구 동향

주요국의 인공지능 신뢰성 정책과 연구를 살펴보면 채용 분야와 밀접한 신뢰성 동향을 파악할 수있다. EU는 개인의 삶과 지원자의 직업 경력에 큰 영향을 미칠 수 있는 면접 및 채용 분야는 고위험 분야로 간주한다. 고위험 시스템(특히 인공지능법과 인공지능 권리장전에서 여러 조직이 주목한)은 배포 전에 데이터 보호 영향 평가(DPIA)를 받아야 하며, 조직은 채용 인공지능 시스템의 위험 수준에 따른 규정을 준수해야 한다. 또한, 시스템의 개발자와 사용자가 관련된 모든 당사자의 요구와 우려에 민감하게 반응할 수 있도록 이해관계자의 협업과 참여가 강조되고 있다.

▼ 주요국의 채용 인공지능 신뢰성 관련 정책 동향

국가	주요 정책	기능
미국	미국의 국가 법률 검토[48]	조직에서 면접을 녹화할 때 발생하는 5가지 주요 데이터 개인정보 보호 및 보안 위험과 이를 완화하기 위한 전략을 강조한다.
	평등고용기회위원회 (EEOC) 통일된 지침	채용 과정에서 인종, 성별, 연령(40세 이상)에 관계없이 지원자를 공정하게 대우할 것을 요구한다.
EU	AI 권리장전/AI 법	채용은 민감한 영역으로 인식하여 고위험 시스템으로 분류한다.
	GDPR	채용 인공지능 시스템은 채용 대상자로부터 개인 데이터(예: 이름, 연락처 등)를 수집한다. 이러한 시스템을 사용하는 조직은 동의를 얻고, 투명한 데이터 처리 정보를 제공하며, 개인 데이터를 안전하게 처리함으로써 GDPR을 준수해야 한다.
	신뢰할 수 있는 AI를 위한 윤리 가이드라인	채용에 사용되는 AI 시스템을 위해 다양하고 대표적인 데이터셋과 디베이싱 기법을 조언한다. 채용에서는 투명성과 설명 가능성이 매우 중요하므로, AI 시스템은 의사결정에 대한 명확한 설명을 제공하고 개인이 자신의 데이터에 액세스하고 수정할 수 있어야 한다.
대한민국	개인정보 보호위원회(PIPC)	이 가이드라인은 데이터 처리, 동의, 투명성 및 책임에 관한 내용을 다루며, AI 기술 사용 시 개인정보 보호를 개괄적으로 설명한다. 채용 인공지능 시스템에서 개인 데이터를 보호하려면 PIPC 규정을 준수해야 한다. PIPC는 채용 시 개인정보를 수집하고 보호하기 위한 규칙을 정하고 동의, 구체적인 사용, 보안을 강조한다. PIPC 규정을 준수하지 않으면 처벌을 받게 되므로 채용 인공지능 시스템의 책임감 있고 윤리적인 운영을 장려한다.

기업은 몇 가지 주요 전략을 채택하여 채용에서 AI의 신뢰성을 확보할 수 있다. 첫째, 채용 인공지능 시스템과 채용 도구를 구축하고 유지 관리할 때 사람의 의견이 중요하다는 점을 인식해야 한다. 자동화로 효율성과 정확성을 높일 수 있지만, AI 시스템이 올바르게 작동하려면 채용 관리자의 전문지식은 필수이다. 둘째, 조직은 AI 채용 비용, 법적 개인정보 보호 문제, 잠재적 편향, 인간 채용 담당자를 대체할 때의 영향 등 여러 요소를 고려하여 이해관계자의 수용 기준에 세심한 주의를 기울여야 한다. 셋째, 공정성을 보장하고 편향을 최소화하는 것으로, 가장 중요한 요소이기도 하다. AI 도구는 인구통계학적 요인보다는 지원자의 기술과 경험에 초점을 맞춰 다양성과 포용성을 촉진함으로써 채용 편향을 크게 줄일 수 있다.

조직은 채용에 AI를 적용하기 위한 이론적 배경을 고려하여 AI 원칙의 기반을 마련하고, 이를 프로세스에 통합해야 한다. 이러한 전략은 채용 프로세스에 AI의 신뢰성을 높이고, 더 정확하고 편향되지 않으면서 다양성과 효율성이라는 조직의 목표에 부합하게 하는 데 종합적으로 기여한다. 앞서 언급한 바와 같이 주요 국가, 기관, 위원회에서는 개발된 AI 모델의 정확성과 신뢰성을 향상하기 위해 적극적으로 연구를 진행하고 프레임워크를 개발하고 있다.

▼ 국내외 주요 산·학·연 채용 인공지능 신뢰성 연구 동향

국가	기관명	활동 및 내용
대한민국	인공지능 윤리 기준 과학기술정보통신부 (정보통신기술)	다양성 존중, 책임성, 투명성과 같은 원칙을 수립하여 개인의 권리와 존엄성을 보호한다. 채용 인공지능 시스템의 윤리적이고 협조적인 환경을 조성하는 데 중점을 두어 채용 인공지능 시스템의 책임감 있고 윤리적인 사용을 장려한다.
EU	데이터 보호 영향 평가(DPIA)[14]	고위험 시스템은 배포 전에 DPIA가 필요하다. 채용 시 개인의 권리와 자유에 높은 위험이 발생할 수 있는 처리를 해야 한다면 DPIA는 필수이다. DPIA는 일회성 활동이 아닌 지속적인 도구로서 처리 전에 수행해야 한다.
	유럽 위원회	2021년 4월 유럽 위원회가 제안한 위원회 작업 프로그램 2021[15]은 AI 시스템을 위험도에 따라 분류하고 시스템의 위험 수준에 따라 규정 준수를 의무화한다. 고용 및 근로자 관리를 포함한 다양한 부문에서 AI 시스템 사용 규제를 목표로 한다.
미국	자율 및 지능형 시스템 윤리에 관한 IEEE 글로벌 이니셔티브	채용 인공지능 시스템을 포함한 자율적이고 지능적인 시스템이 인간의 복지에 미치는 영향을 평가하고 잠재적인 위험에 대처할 것을 권장한다.
		AI를 포함한 자율 및 지능형 시스템에 대한 윤리기준을 제시한다. 특히 채용 대상자에게 큰 영향을 미치는 시스템의 결정과 관련이 있다.
	IBM	채용 인공지능 시스템과 함께 사용할 수 있는 정보 시스템의 재해 복구 계획의 예를 제시한다.
	일리노이-인공지능 화상 면접법(AIVI 법)	화상 면접 시 지원자 분석에 알고리즘(면접 봇) 및 기타 AI 방법 활용을 통제하기 위한 규정을 수립한다.
전 세계	AI 파트너십[59]	업계, 학계, 비영리단체가 참여하는 파트너십으로, 책임감 있고 윤리적인 AI 개발과 사용을 장려한다. 이 파트너십은 투명성, 공정성, 책임성을 원칙으로 강조하는데, 이는 채용 인공지능 시스템에 적용되어 윤리적이고 책임 있는 실천을 보장할 수 있다.

03 안내서 마련 과정

본 안내서는 국내외 수많은 윤리적 지침, 원칙 및 규제 접근법이 제시하는 데 그치지 않고 세부적이고 실용적인 구현 방법론까지 포괄하여 제공하는 것을 목표로 한다. 2024년 신뢰할 수 있는 인공지능 개발 안내서의 요구사항을 근간으로 채용분야에 특화한 인공지능 신뢰성 요구사항 실현을 위해 구체적인 지침을 제공한다.

본 안내서는 다양한 이해관계자를 위한 실용적인 나침반 역할을 자처한다. 여기에는 채용 인공지능 시스템 개발에 중요한 역할을 하는 데이터 과학자, 모델 개발자, 전문 채용 담당자 또는 직업 상담사도 포함된다. 관련 전문가들이 채용 인공지능 개발 환경에서 신뢰성과 윤리적 고려 사항을 보장하는 데 필요한 실용적인 지식과 도구를 갖추도록 돕는 것이 본 안내서의 목적이다.

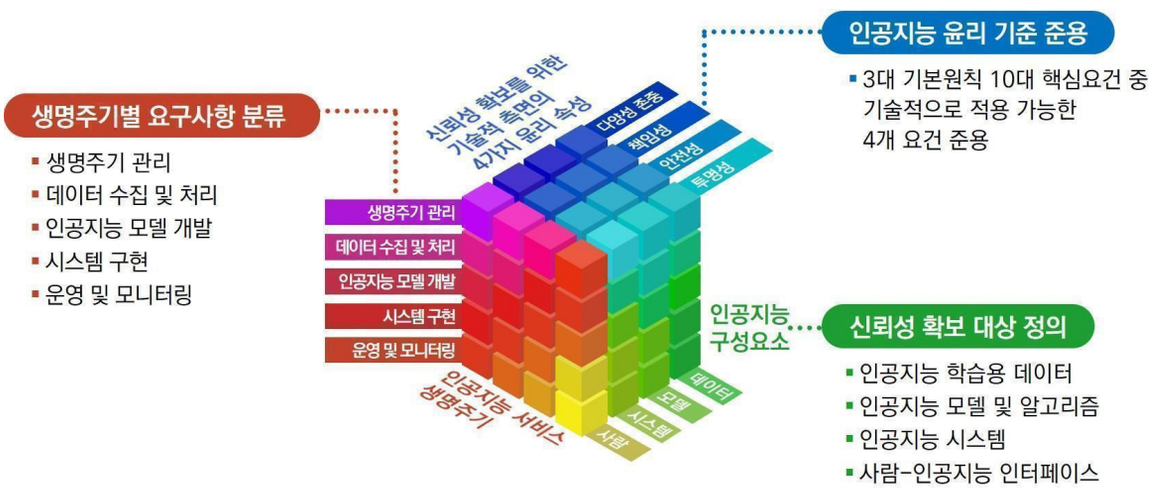
본 안내서의 제작은 학계와 업계의 전문가와 실무자의 지혜를 바탕으로 이들의 공동 노력을 통해 이루어졌다. 이들의 적극적인 참여로 콘텐츠가 더욱 풍성해졌을 뿐만 아니라 실제 현장에서의 실용성을 더할 수 있었다. 또한, 면접 평가 및 채용 인공지능 서비스 전문 기업과의 파트너십도 적극적으로 모색하였다. 이러한 협업 방식을 통해 심도 있는 연구를 수행하고, 포괄적인 사례 연구를 준비하며, 귀중한 피드백을 받아 개발 안내서를 실제 적용 가능성에 기반하여 작성할 수 있었다.

이러한 피드백 중심의 접근 방식을 통해 변화하는 AI 기술 환경과 채용 업계의 특정 요구사항에 맞춰 진화하는 역동적인 리소스를 만드는 것을 목표로 하였다. 이를 통해 실질적인 유용성과 효율성을 향상시켜 해당 분야 이해관계자들에게 가치 있고 지속적인 자산이 되도록 하였다. 또한, 채용 지원자를 공정하게 평가한다는 대의를 발전시킨다는 것 또한 우리의 목표였다. 이에 따라 본 개발 안내서는 AI의 기술적 요건을 충족할 뿐만 아니라 윤리적 기준을 준수하여 궁극적으로 구직자에게 공평한 기회를 제공할 수 있도록 하였다. 채용 프로세스를 개선하고, 급변하는 고용 환경에서 지원자의 경험을 향상시키고자 제작 과정은 공정성과 무결성을 목표로 이루어졌다. 즉, 본 안내서는 신뢰할 수 있는 AI 시스템을 구축하고, 널리 알리며, 채용 영역에서 신뢰할 수 있는 윤리적 도구로 활용할 수 있도록 하였다.

3.1. 개발 안내서 설계 요소

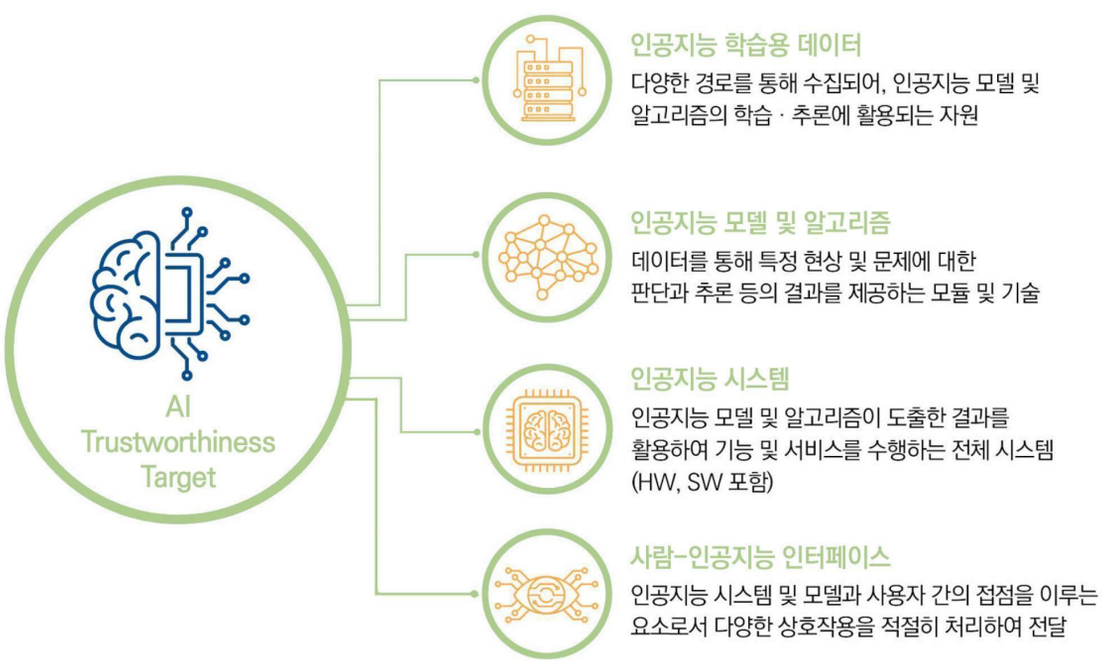
개발 안내서를 제작하는 과정에서 가장 먼저 해야 할 일은 신뢰성 확보였다. 실제로 어떤 요소를 고려해야 하는지 살펴본 결과, 세 가지 설계 요소를 도출하여 안내서에 반영하였다. 각 설계 요소는 요구사항과 검증항목에 모두 반영하였다. 이러한 접근 방식을 아래 그림과 같이 매트릭스matrix 형태로 체계화하여 '인공지능 신뢰성 프레임워크'라 정의하였다. 이 프레임워크는 채용AI 분야뿐만 아니라 일반 분야 및 기타 산업에도 적용 가능하다.

▼ 인공지능 신뢰성 프레임워크



첫 번째는 인공지능 구성 요소이다. 인공지능을 구성하는 네 가지 요소는 학습과 추론 기능을 수행하는 '인공지능 모델과 알고리즘', '인공지능 학습을 위한 데이터', '실제 기능을 구현하는 시스템', '사용자와 상호작용하는 인터페이스'로 구성된다. 각 구성 요소는 인공지능 서비스의 생명주기에 따라 개별적으로 또는 통합적으로 개발, 검증, 운영된다. 따라서 각 구성 요소의 신뢰성 확보 방법을 고민하고, 요소별 요구사항과 검증항목을 제시하고자 하였다. 요소별 신뢰성 확보 방법은 다음과 같다.

▼ 인공지능 서비스 구성 요소



인공지능 서비스 구성 요소	신뢰성을 보장하는 방법
인공지능 모델 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 도출했는지, 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 있는지 등을 검증
사람-인공지능 인터페이스	인공지능 시스템 사용자·운영자 등이 인공지능 시스템의 작동을 쉽게 이해할 수 있는지, 인공지능 오작동 시 사람에게 알리거나 제어권을 이양하는지 등을 검증

두 번째, 인공지능 서비스 생명주기는 첫 번째에서 살펴본 인공지능 서비스 구성 요소들을 구현하고 운영하기 위한 일련의 절차를 의미한다. 기존 소프트웨어 시스템에서 다루는 엔지니어링 프로세스 및 생명주기와 유사하지만, 인공지능의 특성상 데이터 처리와 모델 개발 단계가 별도로 필요하고, 그 외 단계에서는 주요 활동의 정의가 조금씩 다르다. 현재 인공지능 또는 인공지능 서비스의 생명주기는 많은 문헌에서 6~8단계로 구분하고 있다. 대표적으로 OECD와 ISO/IEC에서 제시한 생명주기가 있다. 본 개발 안내서는 두 기관에서 제시한 생명주기를 참조하여, 실무자들이 쉽게 활용할 수 있도록 생명주기 단계별 성격과 활동을 왜곡하지 않는 범위 내에서 다음과 같이 5단계로 정리하였다.

▼ 인공지능 서비스 생명주기별 주요 활동

생명주기 단계	주요 활동
1. 생명주기 관리 (계획 및 설계)	인공지능 시스템 관리 감독 조직 및 방안 마련 인공지능 시스템 위험 요소 분석 및 대응 방안 마련
2. 데이터 수집 및 처리	데이터 품질 확보, 데이터 사용자의 이해를 위한 정보 제공 방안 마련 데이터 라벨링 및 데이터셋 특성^{feature} 문서화 인공지능 모델 구축을 위한 데이터셋 마련
3. 인공지능 모델 개발	비즈니스 목적에 따른 인공지능 모델 구현 구현된 인공지능 모델 확인 및 검증 인공지능 모델 튜닝, 데이터 분석, 추가로 필요한 데이터 수집 인공지능 모델의 성능 평가
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, 설명 가능성 등 시스템 신뢰성 모니터링 치명적인 문제 발생 시 해결 방안 마련

인공지능 서비스의 생명주기 단계는 반복적이고 주기적이지만, 반드시 순차적이지는 않다. 본 개발 안내서에서는 이해를 돕기 위해 1~5단계를 순차적으로 설명했으나, 실제 데이터를 수집-가공하거나 모델을 개발-운영하는 과정에서 순서는 달라질 수 있다.

세 번째, 인공지능의 신뢰성에 필요한 특성을 정의하기 위해 '인공지능 윤리기준'의 10대 핵심요건을 적용하여 '다양성 존중', '책임성', '안전성', '투명성'을 기술적 관점에서 필요한 요건 및 검증항목으로 도출하였다.

EC, OECD, IEEE, ISO/IEC와 같은 국제기구에서는 인공지능 신뢰성의 하위 속성을 세분화하였다. 특히 ISO/IEC 24028:2020-인공지능의 신뢰성 개요에서는 신뢰성 확보에 필요한 고려 사항을 키워드로 제시하였다. 여기에는 투명성, 제어 가능성, 견고성, 회복탄력성, 공정성, 안전성, 개인정보 보호, 보안 등이 포함되지만, 키워드 간의 관계나 신뢰성과의 연관성은 정의하지 않았다. 이처럼 비슷해 보이지만 관점에 따라 조금씩 다른 용어가 여러 문헌에서 서로 다르게 정의되고 있으며, 아직 합의된 속성 분류나 정의는 없다. 이에 EC, OECD, IEEE, ISO/IEC 등 다양한 기관에서 제시하는 속성과 키워드를 종합적으로 분석하고, 국내 학계 전문가들의 의견을 수렴하여 합의점을 모색하였다. 이러한 폭넓은 의견 공유 과정을 통해 인공지능 신뢰성 속성을 도출한 후, 국가 인공지능 윤리기준의 10가지 요구사항에 대응하는 기술적 요구사항을 최종 선정하였다. 각 요구사항에 대한 정의는 다음과 같다.

▼ 인공지능 신뢰성 특성

신뢰성 특성	정의
다양성 존중	인공지능이 특정 개인이나 그룹에 대한 차별적이고 편향된 관행을 학습하거나 결과를 출력하지 않으며, 인종·성별·연령 등의 특성과 관계없이 모든 사람이 평등하게 인공지능 기술의 혜택을 받을 수 있는 것 - 관련 속성: 공정성fairness, 정당성justice - 관련 키워드: 편향bias, 차별discrimination, 편향prejudice, 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 해석이 다양하여 사회적으로나 윤리적으로 일관되게 정의하기 힘들기 때문이다.
책임성	인공지능이 생명주기 전반에 걸쳐 추론 결과에 대한 책임을 보장하는 메커니즘과 마련된 것 - 관련 속성: 책무성responsibility, 감사가능성auditability, 답변가능성answerability - 관련 키워드: 책임liability - 국제표준(ISO/IEC TR 24028:2020)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, 견고성robustness, 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: 설명 가능한 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 이해관계자에게 제공되는 시스템의 속성

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

위에서 설명한 바와 같이 AI 시스템의 신뢰성을 확립하는 데 중요하고 다양한 속성이 있다. 각 신뢰성 속성의 개별적인 정의를 이해하는 것뿐만 아니라 이러한 속성 간의 복잡한 상호작용을 인식하는 것도 중요하다. 예를 들어, 투명성 요건과 개인정보 관련 위험의 관계를 생각해보면 다음과 같다. AI 서비스에 투명성을 과도하게 요구하면 의도치 않게 잠재적인

개인정보 보호 문제를 야기할 수 있다. 설명 가능성은 투명성 달성에 중요한 요소이기는 하지만, 투명성 확보에 기여하는 여러 중요한 요소 중 하나일 뿐임을 인식해야 한다.

따라서 인공지능 서비스는 이러한 인공지능의 신뢰성 속성에 대한 깊은 이해를 바탕으로 제공해야 한다. 또한, 인공지능 서비스가 이러한 속성을 효과적이고 적절하게 통합하고 있는지 반드시 주기적으로 평가해야 한다. 본 안내서는 지속적인 검토와 조정을 통해 AI의 신뢰성을 확보하고, 의도하지 않은 결과와 잠재적 위험으로부터 보호하도록 하는 것을 목표로 한다.

3.2. 채용 분야 주요 고려 사항 반영

채용 분야에서 AI 시스템을 활용하기 위해서는 다양한 유형의 데이터를 평가하고 채용 전형의 고유한 특성을 고려하는 철저한 검증 과정이 필요하다. 자기소개서, 경력기술서, 면접, 시험형 채점 등 전형의 성격에 따라 AI의 적용 범위가 달라질 수 있고, 효과적으로 활용하기 위해서는 전형별로 구체적인 고려 사항이 필요하다.

1. 구조화된 데이터: 이 범주에는 성격 및 역량 테스트와 같은 영역에 중점을 두고 깔끔하게 정리한 데이터와 객관식 평가가 포함되는 경우가 많다. 구조화된 데이터는 지원자를 정량화할 수 있는 평가에 특히 적합하다.

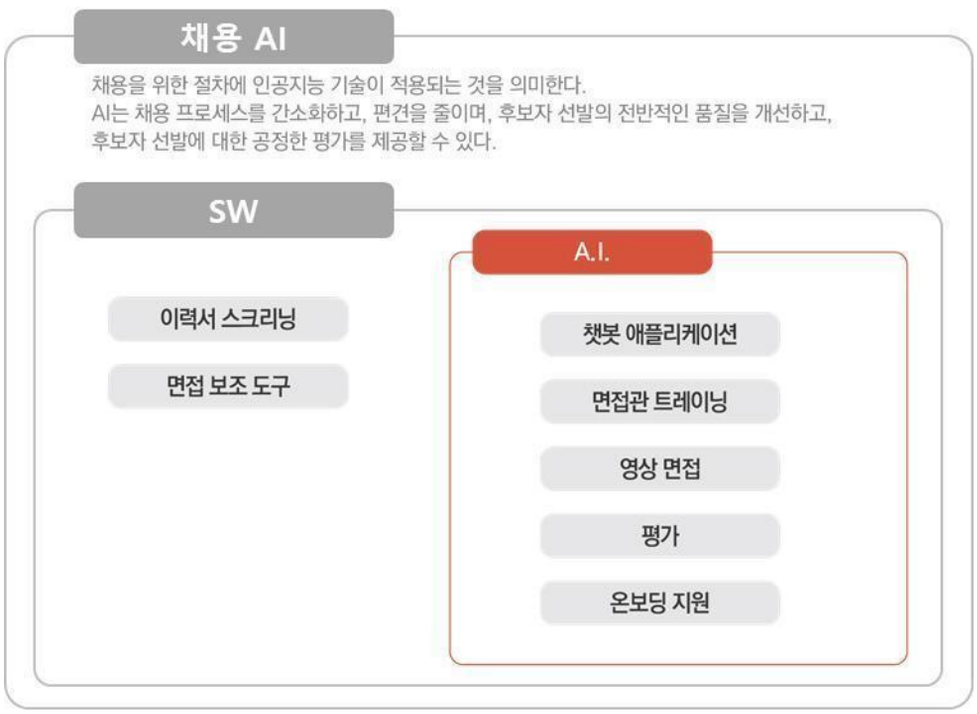
2. 비정형 데이터: 비정형 데이터는 자기소개서와 같은 텍스트 기반 구성 요소, 다양한 면접 형식(영상, 음성, 텍스트), 경력 및 학력을 포함하는 전형 데이터, 게임 내 행동 및 결과와 같은 게임 기반 평가의 간접 평가 등 다양한 콘텐츠 형태로 구성된다. AI는 이처럼 다양한 데이터를 효율적으로 처리하고 평가하는 데 중추적인 역할을 한다.

채용 심사의 범위에는 데이터 수집, 데이터 전처리, 특징 추출, 자연어 처리 및 머신러닝과 같은 기술을 활용한 AI 분석, 후보에 대한 채점과 평가, 최종 의사결정 등의 단계가 포함된다. 이처럼 각기 다른 데이터 유형과 심사 프로세스에 맞게 AI를 적용하면 채용 절차의 효율성과 객관성을 높일 수 있고, 특정 직무에 가장 적합한 지원자를 식별할 수 있어 궁극적으로 채용 프로세스가 간소화된다. 이처럼 AI를 채용 분야의 여러 과정에 적용할 수 있지만, 본 안내서에서는 다음과 같은 고려사항에 맞는 내용만을 다루고 있다.

첫째, 본 개발 안내서에서 신뢰성 대상으로 다루는 채용 인공지능 시스템의 범위는 채용 영역에 활용될 수 있는 모든 범위를 포함하지는 않는다. 본 안내서는 탐지, 예측, 평가, 인식, 분석, 분류 활동, 전형, 면접 진행 등에 직간접적으로 활용되는 인공지능을 대상으로 하며, 본문의 원활한 이해를 돕는 데 필요한 경우에만 채용 활동 범위의 일부 예시를 포함하고 있다. 채용 인공지능 시스템을 개발하고 배포할 때는 기술 영역을 넘어서는 주요 고려 사항을 신중하게 다루는 것이 무엇보다 중요하다. 이러한 고려 사항은 윤리, 형평성 그리고 이러한 시스템이 개인의 삶과 향후 경력 전망에 미치는 심대한 영향과 관련이 깊다. 이러한 AI 기반 채용은 무결성, 공정성, 투명성을 유지하면서 진행되어야만 한다. 이는 데이터 품질과 공정성부터 시스템 관리와 포괄적인 위험 완화까지 채용 인공지능의 윤리적, 실용적 토대를 뒷받침한다. 이러한 주요 고려 사항을 세심하게 살펴봄으로써 채용 영역에서 AI 기술의 잠재력을 활용하여 기회를 창출하고, 정보에 입각한 의사결정을 촉진하며, 관련된 모든 이해관계자가 채용 인공지능에 대한 긍정적인 경험을 하게 하는 것을 목표로 한다.

03 안내서 마련 과정

▼ 채용 인공지능 범위



채용 인공지능은 지원자 평가부터 성격 특성 평가까지 다양한 분야에 활용되며, 자동화된 예약 및 언어 번역 같은 기능도 수행한다. 이는 채용 프로세스 간소화, 편향 감소, 비용 절감, 지원자 선택의 질 향상 등에 기여할 수 있어 채용 인공지능 도입의 중요성을 높인다. 또한, 인공지능은 빠른 지원자 평가 처리와 자격, 기술을 기반으로 한 의사결정을 통해 편향 가능성을 줄이며, 채용 평가에 예측 분석을 통합하여 더 나은 채용 결정을 내릴 수 있도록 지원한다. 이 외에도 AI는 챗봇을 활용하여 실시간 업데이트와 사전 면접 지원으로 지원자 경험을 향상시킨다.

신뢰할 수 있는 채용 인공지능 시스템은 투명성, 공정성, 책임감 있는 운영이 필수이다. 편향 없는 알고리즘, 설명 가능한 결정, 개인정보 보호를 확인해야 하는데, 이는 공정성과 포용성을 높여 능력에 기반한 채용 결정을 보장한다. 인공지능의 활용이 증가하면서 채용 프로세스에 대한 신뢰 구축과 동등한 기회 제공이 더욱 중요해지고 있다. 신뢰할 수 있는 채용 인공지능은 책임감 있고 윤리적인 데이터 사용이 기반이 된다.

둘째, 다음 범위의 맥락에서 채용 인공지능 서비스를 구성하는 네 가지 핵심 요소를 평가하였다. 본 안내서는 주로 채용 전문가와 채용 담당자, 관련 직무에 구직하는 지원자를 돕기 위해 설계된 도구와 애플리케이션을 포함하는 채용 인공지능 시스템 개발에 초점을 맞추고 있다. 또한, 관련 데이터의 수집, 저장 및 관리도 다룬다. '인간-인공지능 인터페이스' 섹션에서는 면접관과 면접 대상자라는 두 주요 사용자 범주를 구분하여 방향성 있는 접근 방식을 취한다.

셋째, 채용 인공지능 영역 내 인공지능 서비스의 생명주기는 채용 영역의 고유한 요구와 복잡성을 충족하기 위해 신중하게 설계되었다. 채용 과정에는 채용 시스템과 지원자의 직업 전망, 경력, 미래의 삶 사이에 중추적인 연관성이 있어 이에 대한 이해가 선행되어야 한다. 본 안내서의 가장 중요한 목표는 AI 기반 채용이 고용주의 요구를 충족할 뿐만 아니라 구직자에게 공평한 기회를 제공하고, 정보에 입각한 공정하고 건설적인 채용 방식을 촉진하는 것이다.

▼ 채용 분야 인공지능의 서비스 구성 요소

구성 요소	설명
학습용 데이터	채용 인공지능 모델을 학습시키는 데 사용되는 초기 데이터에는 이미지, 음성, 동영상, 학교 기록과 같은 다양한 형태의 과거 성취도 데이터, 전문 지식, 가장 많이 사용한 단어, 성격 특성, 외모 영상과 같은 고도의 개인정보가 포함되는 경우가 많다.
모델 및 알고리즘	채용 분야에서 다양한 작업을 수행하도록 설계된 인공지능 모델과 알고리즘이다. 여기에는 지원자 소싱 및 선별, 지원자의 성공 및 문화적 적합성 예측, 화상 면접 실시, 지원자 상호작용 개선, 안면 인식 시스템을 위한 이미지 인식, 텍스트 분석에 필요한 자연어 처리 등을 위한 머신러닝 모델이 포함된다.
인공지능 시스템	특정 채용 평가 작업에 맞춤화된 AI 모델과 알고리즘을 포함하는 포괄적인 AI 시스템 및 플랫폼이다. 여기에는 실시간 화상 면접, 지원자 프로필 매칭 예측, 자동화된 개인 특성 분석, 채용 전문가를 위해 설계한 기타 다양한 AI 기반 솔루션 등 여러 기능이 포함된다.
사람-인공지능 인터페이스	일반적으로 전문가, 관리자 또는 지원자를 고용하는 운영자가 AI 시스템과 상호작용하는 인터페이스이다. 이 인터페이스에는 대시보드, 명령줄 도구 또는 AI 모델에서 생성된 인사이트와 권장 사항을 제공하는 특수 소프트웨어가 포함될 수 있다. 이는 의사결정 및 시스템 제어의 중요한 구성 요소이다.

본 안내서는 채용 인공지능 시스템의 신뢰성을 추구하고, AI의 평가 추론 결과와 그 결과가 미치는 영향의 간극을 줄이기 위한 내용을 반영하였다. 이에 시스템의 무결성 보증과 추론 결과의 영향력이라는 중추적인 요소 사이의 탄력적이고 복잡한 연결을 구축하기 위해 포괄적인 지침과 시스템을 세심하게 고안할 필요가 있다. 본 안내서의 주요 목표는 지원자의 공정한 평가 영역 내에서 결과를 형성하고 조정하는 데 있어 대체 불가능한 역할을 하는 채용 인공지능 시스템 전체를 강화하는 것이다. 이러한 노력은 시스템 자체의 보존은 물론 시스템이 사회와 커뮤니티에 미치는 광범위한 영향력을 유지하고자 하는 열망까지 담고 있다.

▼ 채용 인공지능 서비스 생명주기별 주요 활동

생명주기 단계	주요 활동
1. 생명주기 관리 (계획 및 설계)	채용 인공지능 시스템 관리 감독 조직 및 방법 수립 채용 인공지능 시스템과 연계된 위험 요소 평가 및 적절한 대응 전략 개발 채용 인공지능 시스템 개발에 필수적인 평가 프로세스 검증 및 검증
2. 데이터 수집 및 처리	관련 사례를 검토하여 데이터 편향성을 완화하고, 이러한 우려를 해결하기 위한 맞춤형 전략 개발 관련 데이터셋의 무결성을 보장하며, 데이터 사용자가 데이터를 이해하는 데 도움이 되는 설명 정보 제공 채용 및 직업 상담 전문가와 협력하여 면접 평가 시스템용 데이터를 수집하고 처리
3. 인공지능 모델 개발	평가, 예측, 탐지, 분류, 분석 등 특정 애플리케이션에 맞는 인공지능 모델을 배포하고 검증 인공지능 모델의 성능을 평가하고, 테스트 계획을 수립하는데, 인간 실험 지원자를 찾기 어려울 때 실제 테스트 환경을 마련하는 등 추가 사례를 위한 가상 테스트 시나리오가 포함될 수 있음 인공지능 모델의 편향성을 해결하고 감소/완화하기 위한 전략 수립
4. 시스템 구현	개발된 시스템에서 생성한 추론 결과에 대한 설명 제공 기능을 개선하기 위한 제안서 작성 안전모드를 제정하고 문제 발생 시 이해관계자에게 경고하는 절차 설정 채용 인공지능 시스템의 검증 프로세스 평가를 수행, 시스템 사용자를 위한 사용자 친화적인 설명 작성
5. 운영 및 모니터링	채용 인공지능 시스템을 적극적으로 모니터링하고, 필요에 따라 모델을 재교육하여 지속적인 성능 확보 모델 편향성 감지, 공정성 보장, 설명 제공 등 시스템 신뢰성 모니터링 절차를 수립 문제 발생 시 이를 해결하기 위한 대응 계획을 수립

3.3. 요구사항 및 검증항목 도출

다음 중추적인 단계로, 채용 영역에서 채용 인공지능을 적용하기 위해 정확한 요구사항과 검증항목을 도출하였다. 본 안내서는 저명한 표준화 기구, 기술 기관, 국제단체, 주요 국가 및 도시 정부에서 발표한 정책, 권장 사항 및 표준을 철저히 검토해 작성하였다. 이러한 기초 자료는 채용 부문에서 인공지능의 신뢰성을 확고히 하기 위해 세심하게 작성하였고, 세부적으로 제시된 엄격한 기술 요구사항 개발의 토대를 마련하고자 하였다.

이 과정에서 최고의 표준, 법안 및 규정을 준수하여 최고 수준의 신뢰와 효율성을 심어주는 데 중점을 두었다. 특히 'ISO/IEC 38500:2015, 정보 기술-조직의 IT 거버넌스', 'ISO/IEC TR 24028:2020, 정보 기술-인공지능-인공지능의 신뢰성 개요', 'ISO/IEC 38507:2022, 정보 기술-IT 거버넌스-조직의 인공지능 사용에 대한 거버넌스 시사점' 등 국제적으로 인정받는 표준에 특히 주목하였다. 개발된 채용 인공지능 모델의 거버넌스 측면과 관련성이 높다는 점에서 이러한 표준을 채택하였다.

또한, AI가 인권에 미치는 영향과 관련된 규제, 표준, 법안 및 법규에 초점을 맞췄다. 이 중 주목할 만한 것은 OECD의 2022년 인공지능 시스템 평가 프레임워크, 120개 국내 시민사회단체의 인공지능 정책 지지 선언[16], 개인정보 보호위원회(PIPC)가 발표한 신뢰 기반 인공지능 데이터 규범에 관한 가이드라인이다. 또한, 학생 데이터 활용 및 거버넌스에 대한 IEEE의 P7004 표준과 자율 시스템이 인간에게 미치는 영향을 평가하기 위한 7010-2020 권장 실행을 면밀히 모니터링하였다. 이 문서들은 채용 인공지능 시스템의 맥락에서 개인에게 직접적인 영향을 미친다는 점에서 중요한 의미가 있다.

채용 인공지능 시스템의 신뢰성을 강화하기 위해 본 안내서에서는 국내 및 국제 수준에서 발표된 문서를 종합적으로 검토하였다. 이러한 면밀한 검토 과정을 통해 상당한 인사이트와 정보를 개발 안내서에 통합하였고, 중복되는 내용을 제거하여 더 명확하고 간결하게 압축할 수 있었다. 본 개발 안내서의 목표는 채용 분야 AI 애플리케이션에서 최고 수준의 신뢰성과 안정성을 제공하는 가치 있고 실용적인 자료가 되는 것이다. 본 안내서의 주요 참고 자료는 다음과 같다.

▼ 인공지능 신뢰성 관련 주요 참고 문헌

기관명	발간 연월	권고 및 표준안 명
대한민국	2020.05	지능정보화 기본법
	2021.05	인권, 안전, 민주주의를 보장하는 AI 정책을 촉구하는 선언문
	2023.08	개인정보 보호위원회(PIPC)는 신뢰 기반 인공지능 데이터 규범에 대한 가이드라인을 발표[17]
OECD	2011.02	정책 입안자, 규제 기관, 입법자 등이 특정 상황에 배포된 AI 시스템의 특성을 파악하는 데 도움을 주기 위해 AI 시스템을 평가하는 프레임워크 개발
WEF	2021.06	조직을 위한 9가지 핵심 윤리적 AI 원칙
	2021.10	모델 인공지능 거버넌스 프레임워크
NIST	2023.01	AI 위험관리 프레임워크(AI RMF)인 AI 거버넌스 솔루션
	2023.11	사이버 보안 프레임워크
IEEE	2017.03	아동 및 학생 데이터 거버넌스를 위한 IEEE P7004 표준
	2019.03	IEEE P7002 데이터 개인정보 보호 프로세스
	2020.05	자율 및 지능형 시스템이 인간 복지에 미치는 영향을 평가하기 위한 IEEE 7010-2020 권장 사례

기관명	발간 연월	권고 및 표준안 명
국제 표준화기구 (ISO/IEC)	2020.05	ISO/IEC TR 24028:2020, 정보 기술-인공지능-인공지능의 신뢰성 개요
	2021.11	ISO/IEC TR 24027:2021, 정보 기술-인공지능(AI)-AI 시스템 및 AI 지원 의사결정의 편향성
	2022.04	ISO/IEC 38507:2022, 정보 기술-IT 거버넌스-조직의 인공지능 사용에 따른 거버넌스 영향
	2022.10	ISO/IEC 27001:2022, 정보 기술-보안 기술-정보 보안 관리 시스템
	2023.02	ISO/IEC 23894:2023, 정보 기술-인공지능-위험관리에 대한 지침

이를 통해 도출된 최종 요구사항은 아래 표와 같다. 인공지능 윤리의 핵심 요구사항의 결과 또한 표시되어 있다.

▼ 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		✓		✓

3.4. 현장 적용 및 전문가 의견 수렴

신뢰성 특성을 수립한 후 기술적 실현 가능성, 실용성, 포괄성 등의 요소를 고려하여 각 항목을 세심하고 철저하게 평가하였다. 평가에는 각 검증항목의 타당성, 실제 개발 시나리오에 대한 적응성, 역사적 통찰력부터 해당 분야의 최신 발전까지 광범위한 연구 결과와의 연계성 등을 면밀히 검토하는 과정이 포함되었다. 이를 위해 검토 및 자문 과정에 적극적으로 참여한 채용 전문가를 비롯해 다양한 이해관계자 패널을 참여시켰다. 또한, 업계 및 학계 연구자, 기업 기획자, 개발 프로젝트 리더, 저명한 교수 및 도메인 전문가까지 협업을 확대하였다. 이들의 다양한 관점이 제시되면서 권장 사항은 더욱 깊고 풍부해졌다.

또한, 채용 및 면접 평가 서비스 전문 기업과 파트너십을 맺어 이론과 실제 적용 간의 시너지를 도모하였다. 협업을 통해 심도 있는 사례 연구를 수행하고 귀중한 피드백을 얻어 개발 안내서의 실질적인 유용성을 더욱 향상시켰다. 지속적인 대화와 반복적인 검토 과정을 거쳐 업계의 요구사항을 충족하고 관련성과 효율성을 갖추도록 개선할 수 있었다.

04 안내서 활용 대상

본 안내서는 채용 영역에서 인공지능 서비스를 구현하거나 영향을 미치는 데 적극적인 역할을 하는 조직과 개인을 아우르는 모든 이해관계자를 위한 포괄적이고 필수적인 참고 자료이다.

분야의 특성을 고려할 때 개발과 운영 단계에서 도메인 전문가나 법률 자문가와의 긴밀한 협업이 필수이다. 협업의 범위는 도메인 전문가, 전문 채용 담당자, 채용 인공지능 서비스 제공 및 사용에 직간접적으로 관여하는 모든 개인에게까지 확대된다. 이들은 모두 채용 영역에서 AI 시스템의 신뢰성과 효율성을 보장하는 중요한 잠재적 이해관계자이다. 이 과정에서 효과적인 협력 체계의 필요성이 강조된다. 따라서, 대표 이해관계자는 한 명 이상의 협력 대상과 긴밀하게 협력하며, 이들 간의 협력 관계는 부록3 에 기술되어 있다.

대표 이해관계자와 협력 대상은 한국SW산업협회[KOSA]가 국가직무능력표준[NCS]를 기반으로 개발한 IT분야역량체계[ITSQF]에 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, 부록4 에 제시된 각 직업·직무에 대한 정의를 참고하여 직무별 역할을 확인할 수 있으나 채용 분야에 국한된 직무 역할까지는 제공하지 않는다.

▼ 채용 인공지능 생명주기 단계별 신뢰성 확보를 위한 대표 이해관계자

생명주기 단계	대표 이해관계자(예)	관련 요구사항
1. 계획 및 설계	• 정보기술기획자 • IT감사자 • IT품질관리자	- 인공지능 시스템에 대한 위험관리 계획 및 수행 - 인공지능 거버넌스 체계 구성 - 인공지능 시스템의 신뢰성 테스트 계획 수립
2. 데이터 수집 및 처리	• 데이터아키텍트 • 데이터분석가	- 데이터의 활용을 위한 상세 정보 제공 - 데이터 견고성 확보를 위한 이상 데이터 점검 - 수집 및 가공된 학습 데이터의 편향 제거
3. 인공지능 모델 개발	• 인공지능SW개발자 • 인공지능아키텍트	- 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	- 인공지능 시스템의 추적가능성 및 변경이력 확보 - 서비스 제공 범위 및 상호작용 대상에 대한 설명 제공

05 안내서 활용 방법

05 안내서 활용 방법

채용 인공지능 시스템은 모든 조직에 똑같이 적용하기 어렵다. 적용하는 조직의 특정 요구사항과 특성에 맞게 적용해야 한다. 내부 기술 역량과 제품 속성에 따라 기준을 선택하고 적용하여 서비스 환경에 원활하게 맞출 수 있도록 맞춤화*할 필요가 있다.

* 조직과 개발자는 이 개발 안내서를 따르되, 개발 프로세스를 간소화하고 프로젝트의 효율성을 높이기 위해 필수 항목만 우선 확인하고 적용하는 데 집중해야 한다.

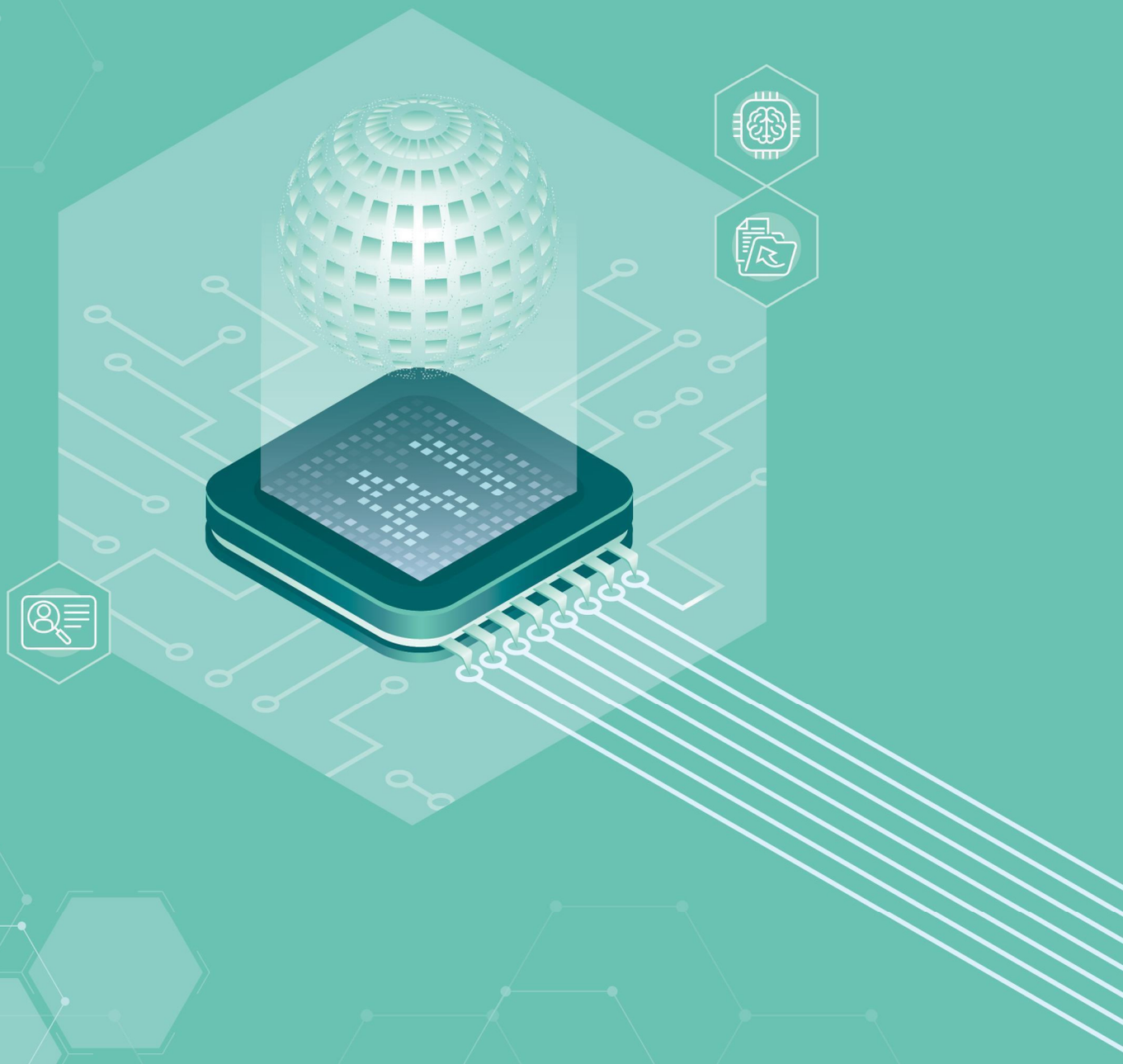
채용 인공지능 시스템은 민감한 개인 데이터와 비디오 영상을 처리하는 경우가 많다. 따라서 본 안내서를 사용하기 전에 윤리적 AI 원칙과 개인정보 보호 가이드라인을 사전 검토할 것을 적극 권장한다. 또한, 책임 있는 데이터 사용과 저장을 보장하고, 개발된 시스템이나 애플리케이션의 신뢰성을 확보하기 위해 PIPC의 '신뢰 기반 인공지능 데이터 규범을 향한 첫걸음[17]'에 대한 법률 자문 및 컨설팅 수행을 강력히 추천한다. 여기에는 AI에 관련 속성뿐만 아니라 성능 및 보안과 같은 기존 시스템 속성의 검증도 포함된다.

본 안내서는 다음과 같은 방법으로 활용하면 효과적이다.

① **위험 영향 분석**: 채용 인공지능 시스템의 목적, 범위, 사고 위험 및 잠재적인 사회적 결과를 분석하여 시스템 구현과 관련된 위험을 평가한다. 이해관계자 간의 협업을 통해 종합적인 영향을 분석한다.

② **요구사항 선정**: '①'의 결과를 바탕으로 안내서에 설명된 요건을 참조하여 채용 인공지능 서비스의 신뢰성을 보장할 요건을 선정한다. 전문 이해관계자 및 법률 고문과의 상담을 권장한다. 요구사항이 불필요하다고 판단되면 점검표에서 제외하여 'N/A'로 표시할 수 있다.

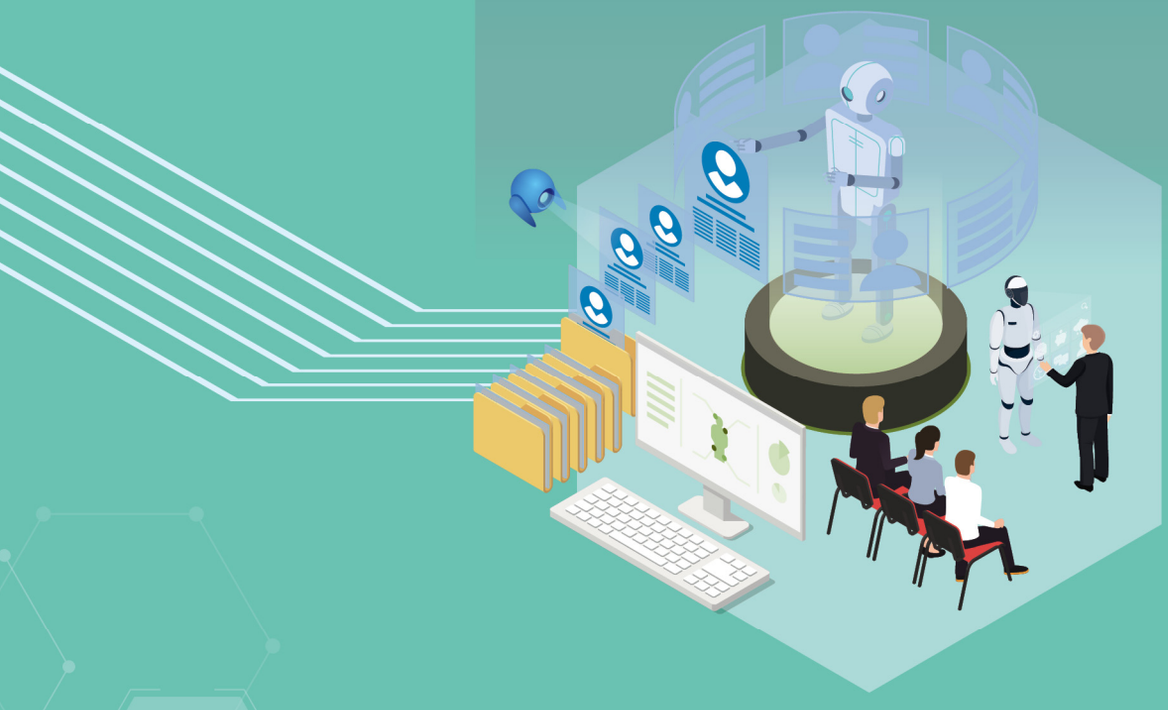
③ **자가 점검 수행**: '②'단계에서 선택한 요구사항에 대해 세부 요구사항 및 검증 기준을 참조하여 검증한다. 준수 여부를 확인하기 위해 자체 점검 프로세스를 수행한다. 추가 테스트가 필요하다면 실행한다. 철저한 점검을 위해서는 각 요건을 담당하는 담당자 간의 협업이 필수적이다. 각 요건에 대해 지정된 담당자는 해당 결과물 평가에 관련된 당사자들과 긴밀히 협력하면서 검사 프로세스를 주도한다. 여기에는 확립된 기준 준수 여부를 확인하는 데 필요한 절차, 코드 및 필수 분석 데이터의 평가가 포함된다. 테스트 또는 측정이 필요한 경우 필요한 절차를 반드시 수행해야 한다. 정량적 평가 외에도 검증 기준 준수 여부에 대한 정성적 평가도 가능하지만, '①'단계에서 평가한 서비스의 영향 수준을 고려하여 대표 및 기타 이해관계자와 협의하여 최종 결정하는 것을 권장한다. 이러한 구조화된 접근 방식은 신뢰할 수 있는 채용 인공지능 시스템의 개발과 운영을 용이하게 하여 조직 의 채용 프로세스와 지원자의 안전, 보안, 복지를 증진한다.



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 시스템의 위험 관리 계획 및 수행

- 채용 인공지능 시스템을 구현 및 운영하는 과정에서 발생 가능한 모델 오인식, 기능 오동작, 보안 및 개인정보 이슈 등 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 ☐

- 위험 관리는 위험 인식identification, 위험 분석analysis, 위험 평가evaluation, 위험 대응treatment으로 구분한다. 신뢰성 확보를 위해 이러한 네 가지 활동을 생명주기 단계별로 지속·반복 수행하여 위험을 제거 및 방지하여야 한다. ISO 31000:2018-Risk management에는 위험 관리의 개념 및 정의와 전체적인 흐름이 소개되어 있으며, 미국 국립표준기술연구소(NIST)의 위험관리 프레임워크(AI RMF 1.0)[22] 등을 인공지능 시스템 설계시 참고 할 수 있다.

참고 Microsoft의 위험 관리를 위한 프레임워크 예시

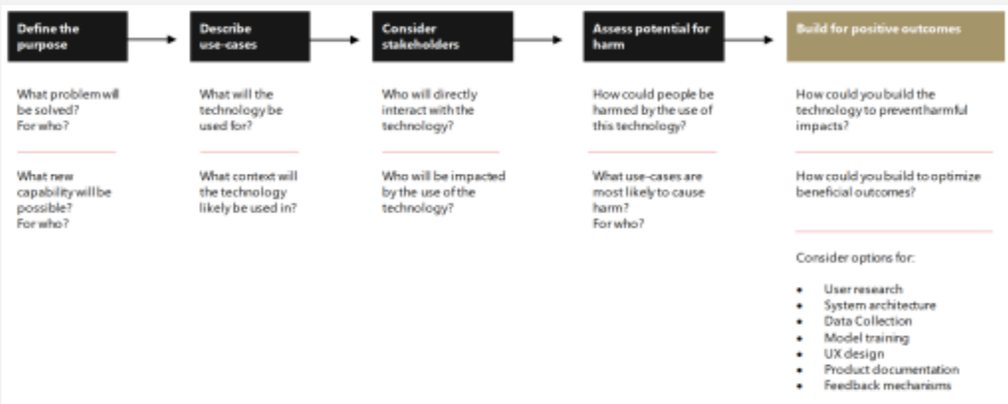
HARM(Human-AI-Risk-Mitigation)은 보안 위험을 평가하는 구조화된 방법과 머신러닝 모델에 사용되는 특성과 관련된 광범위한 윤리적·사회적 고려 사항을 제공하여 보다 강력하고 책임감 있는 AI 시스템 개발을 유도한다. 특히 머신러닝 및 AI 시스템의 맥락에서 학습에 사용되는 특성 분석과 선택에 사용할 수 있다. 이 접근 방식은 기존의 보안 및 위험 분석을 확장하여 이러한 시스템에서 사용되는 특성과 기능으로 인해 발생할 수 있는 잠재적 피해까지 고려한다.

머신러닝에 HARM 모델링을 적용할 때는 보안 위협만 살펴보는 데서 벗어나 학습 프로세스에 사용되는 기능 및 데이터와 관련된 편향, 공정성 문제, 의도하지 않은 결과까지 고려해야 한다. 개발자는 개발 프로세스 초기에 잠재적인 피해, 편향, 함정을 식별함으로써 기능 선택, 데이터 품질, 알고리즘 설계에 대해 정보에 입각한 결정을 내리고, 보다 공정하고 신뢰할 수 있으며 책임 있는 AI 시스템을 만들 수 있다.

이 솔루션은 네 가지 주요 구성 요소로 이루어져 있다.

- 위험 식별: 하드웨어, 소프트웨어, 데이터, 인적 요소와 관련된 위험을 포함하여 AI 시스템과 관련된 잠재적 위험을 식별한다.
- 자산 특성화: 데이터 입력 및 출력, 알고리즘, 휴먼 인터페이스를 포함한 시스템 및 구성 요소의 특성을 정의한다.
- 위험 분석: 위험 발생 확률과 위험의 심각성을 고려하여 식별된 위험의 발생 가능성과 잠재적 영향을 평가한다.
- 완화: 데이터 품질 개선, 투명성 및 설명 가능성 구현, 효과적인 인적 감독 메커니즘 구축과 같은 기술적 및 비기술적 해결책을 포함하여 위험을 완화하고 잠재적 피해를 최소화하할 조치를 구현한다.

개인정보 보호(보호변수) 및 보안을 위해 Microsoft는 설계 모델인 HARM 모델링을 수립하여 취약점을 예측하는 데 도움을 줄 수 있다. 또한, 이 표를 평가하여 잠재적인 사용자 다양성을 고려함으로써 편향을 완화하기 위한 차별 기준을 설정할 수도 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 권리장전[19]에서 고용은 민감한 영역으로 간주하며, EU의 인공지능법[20] 및 일반 데이터 보호 규정(GDPR)에서는 채용 인공지능을 고위험 시스템으로 분류한다. 따라서 채용 인공지능 시스템과 관련해 위험 요소를 면밀히 분석하고 관리하여 구직자에게 잠재적인 피해가 발생하지 않도록 해야 한다.

참고 EU 인공지능 법 발췌–고용 시스템[20]

"고용, 근로자 관리 및 자영업에 대한 접근, 특히 채용 및 선발, 승진 및 해고 결정, 업무 관련 계약 관계에 있는 사람의 업무 할당, 모니터링 또는 평가에 사용되는 AI 시스템도 향후 지원자의 경력과 생계에 상당한 영향을 미칠 수 있으므로 고위험군으로 분류해야 한다. 업무 관련 계약 관계에는 '커미션 업무 프로그램 2021'에 언급된 바와 같이 플랫폼을 통해 서비스를 제공하는 직원 등 당사자가 포함되어야 한다. 채용 과정과 업무 관련 계약 관계에 있는 사람의 평가, 승진 또는 유지 과정에서 이러한 시스템은 예를 들어 여성, 특정 연령대, 장애인, 특정 인종 또는 민족 출신 또는 성적 지향에 대한 차별과 같은 역사적 차별 패턴을 영속화할 수 있다. 해당 대상자의 성과와 행동을 모니터링하는 데 사용되는 인공지능 시스템은 데이터 보호 및 개인정보 보호에 대한 권리에도 영향을 미칠 수 있다."

- AI 채용 시스템의 위험에 대한 구체적인 ISO 표준은 없다. 그러나 채용 인공지능 시스템에 적용할 수 있는 위험 관리와 관련된 몇 가지 일반적인 표준을 참고할 수 있다.

ISO 참조 표준 예시

표준	제목
ISO/IEC 31010:2009[23]	위험 관리–위험 평가 기법
ISO 31000:2018[24]	위험 관리–가이드라인
ISO/IEC 27005:2018[25]	정보 기술–보안 기술–정보 보안 위험 관리
ISO/IEC 24028 :2020[26]	정보 기술–인공지능–인공지능의 신뢰성 개요
ISO/IEC 23894:2023	정보 기술–인공지능–위험 관리 지침

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

Yes ☐ No ☐ N/A ☐

- 채용 인공지능 시스템을 이용하려면 입사 지원자가 자신의 하드웨어를 사용하기 때문에 면접 과정에서 사용자의 하드웨어(휴대폰, 태블릿, 마이크, 웹캠)의 호환성 등 여러 가지 위험이 발생할 수 있다. 또한, 데이터 기반 분석의 특성상 편향, 설명 미제공, 모델 공격과 같은 위험 요소를 반드시 도출해 내야 한다. 해당 위험 요소의 주요 내용은 01-1에서 제시한 ISO/IEC 표준들을 참고할 수 있다.
- 채용 인공지능 시스템 구현 시 시스템이 독립적으로 지원자를 선택할지 아니면 채용 담당자에게 결정 권한을 부여할지와 같은 사용자 지정 옵션에 대한 정보 제공이 필요하다. 또한, 데이터 보호, 보안 조치 등을 명확하게 전달하고, 알림에는 편향 완화, 성과 평가를 위한 피드백 메커니즘, 법적·윤리적 준수 노력이 수반되어야 한다. 아래는 인공지능 채용평가의 위험 요소 및 고려 사항의 예시이다.

채용 인공지능 시스템 구현 시 발생할 수 있는 위험 요소 예시

위험	설명	원인	예시와 참조
편향된 데이터 위험	채용 인공지능 시스템이 편향된 데이터를 학습하면 불공정하고 차별적인 채용 집행이 발생해 소송, 부정적인 여론, 회사 평판 손상으로 이어질 수 있다.	특정 그룹(인종, 성별, 연령 등)을 과소/과대 대표 학습	[27]은 아프리카계 미국인 언어의 경우 자동 음성 인식의 편향성을 제기한다.
알고리즘 편향 위험	채용 인공지능 시스템에서 사용하는 알고리즘이 편향되면 인종, 성별, 나이 등의 요인에 따라 특정 지원자를 부당하게 우대하거나 거부하는 결과를 초래하기도 한다. 이는 차별 소송과 부정적인 여론으로 이어질 수 있다.	시스템 테스트 또는 검증의 부적절함	[28]은 사람들의 개인적 속성을 무시하고 특정 기술과 행동에만 초점을 맞출 것을 제안한다.
보안 위험	보안이 적절히 유지되지 않으면 데이터 유출이나 해킹 시도에 취약해져 민감한 지원자 데이터가 손실되고 회사의 평판 손상로 이어질 수 있다.	얼굴 인식 사용에 따른 생체 인식데이터 유출 및 개인정보 보호 문제	[29]는 채용 인공지능 시스템이 얼굴 인식 스캔을 사용하여 신원을 확인하는 경우 보안 문제를 논의한다.
기술적 실패 위험	기술적 오류나 오작동이 발생하면 부정확하거나 불완전한 지원자 평가로 이어져 잘못된 채용 결정이 내려지고 잠재적으로 법적 문제가 발생할 수 있다.	하드웨어 비호환성 부정확한 데이터 입력 잘못된 시스템 사용	하이어뷰는 면접 중 배경 소음을 필터링하는 소음 제거 기능이 있는 AI 면접 플랫폼을 제공한다.
투명성 부족 위험	지원자를 평가하는 방법이나 채용 과정에서 고려하는 요소를 투명하게 밝히지 않으면 지원자의 불신과 회의, 잠재적인 법적 문제로 이어질 수 있다.	신뢰할 수 없는 출처로부터 기술 구매[30]	GDPR은 개인이 알고리즘 결정에 대한 설명을 요구할 수 있는 '설명할 권리'를 제공할 것을 권고한다.
제한된 범위 위험	지원자 속성의 평가 범위가 좁으면 직무와 관련된 중요한 요소를 놓쳐 잘못된 채용 결정과 잠재적인 법적 문제를 초래할 수 있다.	부정확하거나 불완전한 데이터	[31]에서는 데이터의 대표성 부족으로 인한 모델 설계의 편향성을 소개한다.

- 위험 요인을 파악한 후에는 다양한 환경이나 상황에서의 관리 방안과 그 파급효과를 분석해야 한다. 인공지능 시스템의 생명주기 동안 주기적인 추세 분석과 모니터링을 반복 수행하여 도출한 위험요소의 심각도와 발생 빈도 등을 척도화 할 수 있다. 이를 통해 위험수준 및 파급효과에 따른 대응방안을 준비해야 한다.

위험 요인 식별, 위험 수준 평가 및 대응 준비의 예시[36][37]

위험 요소 식별	위험 수준 평가	대응 준비
AI 알고리즘의 편향과 차별	높음	시스템적 편향을 줄이고 채용 시장의 공정성과 형평성을 증진하기 위한 AI 시스템 설계
AI 의사결정의 투명성 부족	높음	AI 시스템을 투명하게 설계하고 의사결정 프로세스가 명확하고 이해하기 쉽도록 보장
데이터 수집과 관련된 개인정보 보호 문제	중간	AI 시스템이 관련 데이터 보호법을 준수하는지, 입사 지원자의 개인정보를 보호하는지 확인
버그 또는 연결 문제와 같은 기술적 문제	낮음	AI 시스템이 신뢰할 수 있고 잘 유지 관리되어 기술적 문제를 방지하는지 확인
AI 도구의 제한된 평가 기능	중간	AI 시스템을 관련 기준에 따라 지원자를 평가하도록 설계하고 다른 평가 방법과 함께 사용하여 종합적인 평가가 이루어지도록 보장

AI 기반 채용의 윤리적 고려 사항에 관한 참고자료[41]

작성자	유형	주제	관점	윤리적 위험				
				알고리즘 편향성 도입	프라이버시 손실 및 힘의 비대칭	투명성 및 설명 가능성 부족	책임 소재의 모호함	사람의 감독 상실 가능성
Acikgozetal (2020)[42]	경험적: 실험	AI 채용에 대한 지원자의 반응	설명	X		X		
Raghavan (2020)[43]	경험적: 품질 분석	실제 편향성 완화	법률, 기술*	O		X		X
Yargeretal (2020)[31]	개념	알고리즘 채용의 형평성	이론	O	X	X		X
Chamorro-Premuzicetal., (2019)[44]	개념	AI 채용의 윤리적 구현	실무자	X		X		
Sánchez-Monederoetal. (2020)[30]	경험적: 분석	실제 편향성 완화	법률, 기술*	O		X	X	X

설명 범례: O – 초점 주제; X – 언급된 주제

* 여러 관점이 적용되며, 이 중 가장 먼저 언급된 관점이 주된 관점이다. 모든 통계에서는 이중 계산을 피하기 위해 가장 먼저 언급된 관점만 계산한다.

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. 예를 들면 기회의 획득 또는 상실, 개인의 건강이나 안전에 대한 위협, 피해 복구를 위한 특정 기술에 대한 재정적 비용 등이 있다.
- 만약 인공지능 채용 시스템이 극단적으로 부정적인 결과를 초래할 수 있다고 확인된 경우, 인공지능 기술 적용에 대해 재검토하여야 한다.

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

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

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

Yes ☐ No ☐ N/A ☐

- 01-1에서 분석한 위험 요인별로 대응 방안을 마련해야 한다. 대응 방안에는 위험 요인의 원인을 제거하여 인권 침해 및 개인의 피해를 사전 예방하거나, 잘못된 의사결정에 따른 파급효과 및 부정적 영향을 최소화할 방안이 포함된다.
- 여기서 대응 방안은 구현 및 운영, 소프트웨어 및 하드웨어 기능, 모델 학습 기법 및 전략 등의 절차를 포함하여 기술적으로 적용할 수 있는 모든 방법을 의미한다. 01-2a에서 참조한 바와 같이 ISO/IEC 24028:2020은 대응 방안의 분류를 제공한다. AI를 구현하는 모든 이해관계자는 이를 고려하여 위험 요소에 대한 대응 방안을 마련하고 위험을 제거 및 완화해야 한다.

위험 요소 완화 방안 예시

위험 완화 방안	내용
위험 평가 수행	잠재적 위험을 식별하고, 발생 가능성과 잠재적 영향을 분석하며, 심각도에 따라 우선순위를 정하는 것을 포함한다.
보안 조치 구현	위험 평가 결과에 따라 식별된 위험의 발생 가능성과 영향을 줄이기 위한 보안 조치를 구현한다. 보안 조치의 예로는 데이터 암호화, 액세스 제어, 방화벽 등이 있다.
비상 계획 개발	사고 발생 시 사고에 대응하고 그 영향을 최소화하는 방법을 설명하는 비상 계획을 수립한다. 여기에는 백업 및 재해 복구 계획이 포함된다.
정기적으로 조치 검토 및 업데이트	보안 조치와 비상 계획을 정기적으로 검토하고 업데이트하여 새로운 위협과 기술에 대한 효과적인 상태를 최신으로 유지하도록 한다.
교육 및 인식 제고	채용 인공지능 시스템 관련한 모든 직원이 보안 정책과 절차를 숙지하고 있는지 확인하고, 보안 모범 사례에 대해 정기적인 교육을 한다.

참고 위험 평가를 위한 알고리즘 영향 평가(AIA) 도구[45]

캐나다 정부가 공개한 알고리즘 영향 평가(AIA)는 재무부의 자동화된 의사결정에 관한 지침을 지원하기 위해 개발된 의무적인 위험 평가 도구이다. AIA는 조직이 자동화된 의사결정 시스템 배포와 관련된 위험을 평가하고 완화하는 데 도움이 되도록 설계되었다. 이 도구는 자동화된 의사결정 시스템의 영향 수준을 결정하는 설문지로, 51개의 위험 질문과 34개의 완화 질문으로 구성되어 있다. AIA는 자동화된 의사결정 시스템의 영향 수준을 식별하는 데 도움을 주며, 각 영향 수준은 점수 백분율 범위에 해당한다.

Table 3. Raw impact score from the risk areas

Risk area	No. of questions	Maximum score
1. Project	16	27
2. System	1	0
3. Algorithm	2	6
4. Decision	2	7
5. Impact	20	42
6. Data	10	44
Raw impact score	**51**	**126**

Table 4. Mitigation score from the mitigation areas

Mitigation area	No. of questions	Maximum score
7. Consultations	2	2
8. De-risking and mitigation measures	32	44
Mitigation score	**34**	**46**

- 또한, IEEE 7010-2020은 자율 및 지능형 시스템, 특히 채용 인공지능 시스템이 인간 복지에 미치는 영향을 평가하고, 잠재적 위험을 해결하기 위한 프레임워크를 제안한다. 아울러 ISO/IEC 38500, 27001, 27005, 27701은 각각 IT 관리, 정보 보안, 정보 보안 위험 관리 그리고 개인정보 보호에 관한 지침을 제공하므로 참고할 수 있다.

01-2a 위험 요소별 완화 또는 제거 방안을 마련하였는가?

Yes ☐ No ☐ N/A ☐

- 위험 요소를 제거할 수 있는 구현 및 운영 방식, 소프트웨어 및 하드웨어 기능, 모델 학습 기법 및 전략 등의 기술적인 방법론을 도출하여야 한다. 이러한 방법론의 분류와 개요는 ISO/IEC 24028에 제시되어 있다.
- 채용 인공지능 시스템의 생명주기 동안 잠재적인 위험을 식별하고, 위험의 발생 가능성과 시스템에 미치는 영향에 따라 우선순위를 정하여 해당 위험에 구체적인 대비 계획을 수립해야 한다. 여기에는 보안 조치 추가, 데이터 품질 향상, 시스템 설계 수정 등이 포함될 수 있다.
- 수립한 계획을 실행하려면 관련 이해관계자들이 위험과 조치 계획을 인지하고 있는지 확인해야 하며, 정기적인 모니터링으로 새로운 위험을 감지하고 완화 계획의 효과를 검토해야 한다. 시간이 지나면서 계획을 조정하여 위험을 효과적으로 완화하는지도 확인해야 한다.

채용 인공지능 시스템의 위험 요소 및 대응 계획 예시

위험 요소	대응 계획	예시와 참조
데이터 개인정보 보호 및 보안 침해	데이터 암호화 강화, 개인정보 송수신 및 저장 취약점 점검, 외부 접근 및 개인정보 접근 제어, 정기적인 감사 및 모니터링, 접근 및 반출 제어 시스템 구현 등	미국의 국가 법률 검토[48]: 5가지 주요 데이터 프라이버시 및 보안 위험과 이를 완화하기 위한 전략 제시
기술적 오류 및 시스템 장애	면접 자료의 주기적 백업 및 이중화 조치, 재해 복구 절차 마련, 인사팀에 비상 절차 교육, 정기적인 시스템 테스트 및 유지보수, 면접 대상자의 웹캠, 마이크, 음성을 확인할 자체 테스트 수행 등	IBM[49]은 채용 인공지능 시스템에 적용할 수 있는 정보시스템의 재해 복구 계획 사례
알고리즘 및 모델 개발의 편향성	알고리즘과 모델의 테스트 및 검증, 데이터셋 대표성 확인, 다양한 이해관계자들에게 모델의 편향 가능성과 그 해결 조치 교육, 면접 후 피드백 통합 등	[43]에서는 알고리즘 채용의 편향성 완화, 클레임 및 수행 평가를 논의
시스템 조작을 시도하는 악의적인 행위자	강력한 인증 및 액세스 제어 구현, 시스템 연결의 정기적 모니터링, 정기적 침투 테스트 및 취약점 파악, 보안 표준 준수, 지원자에게 보안 위협 교육	[50]은 연방수사국(FBI)의 사이버 범죄자 사례, 미국인의 개인 식별 정보(PII)와 딥페이크를 훔쳐 일자리 지원에 도용 사례
투명성 및 설명 가능성 부족	HMI 기능 및 채용 과정 설명 강화, 면접 대상자와의 상호작용 강화, 해석 가능성 강화를 위한 기술 개선, 면접 프로세스에서의 상호작용 강화 등	[51]은 AI 시스템이 관련 정보를 적시에 제공하도록 제안

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.
- 위에서 평가한 위험 요인 중 파급효과가 큰 것에 우선순위를 두어 대응 방안을 적용하고, 위험의 파급효과가 큰 경우 편향된 데이터 개입, 인공지능 시스템의 판단 결과에 사람의 개입을 고려하는 등 위험 완화 방안을 적용해야 한다. 특히 면접 평가 모델은 데이터셋의 특성 상 편향될 가능성이 크기 때문에 파급효과를 신중하고 세밀하게 검토하여 파악해야 한다.
- AI 생명주기별로 발생 가능한 위험 요소에 대응할 수 있는 기술적 방법을 적용한 후에는 그 파급효과를 재평가하여 실제로 위험이 제거, 예방 또는 완화되는지 확인해야 한다.

채용 인공지능 시스템의 생명주기 단계별로 발생할 수 있는 문제와 대응 방안 예시

프로세스	발생 이슈	대응 방안
계획 및 설계	사용자 니즈 및 요구사항에 대한 이해 부족	포괄적인 사용자 조사 수행 및 요구사항 수집
	예측 가능한 위험 요소 존재	의도된 성격 특성 등급, 대상 집단의 특성, 관련된 상호작용 유형, 사용자 프로필, 면접 환경, 잠재적 장애물 등의 정보 고려
데이터 수집 및 처리	편향된 데이터 수집	다양한 데이터 소스 사용, 공정한 샘플링 기법 활용, 다양한 하드웨어 장치 사용, 다양한 성별·인종·비원어민·연령대 데이터 수집
	데이터 저장소 손상	데이터 저장소 공격 관련 대응책 준비
	민감 데이터의 오용 및 유출	데이터 검토 위원회 구성, 면접 데이터 활용 적절성 평가
	데이터 편향 발생	시스템 성능 정의, 편향성 평가(인지적, 사회적, 문화적, 통계적 등)
	데이터 소스 불투명성	데이터 소스 공개 및 시스템 운영 외부 감사
	데이터 프라이버시 침해	데이터 암호화, 사용자 인증 조치, 엄격한 보안 프로토콜 구현
	잘못된 데이터 레이블	여러 어노테이터 사용, 자동화된 검사, 라벨 검수 프로세스 정립
인공지능 모델 개발	데이터 중독의 결과 조작	비정상 학습 데이터의 탐지 및 필터링
	AI 모델 적대적인 공격	적대적 공격에 대한 방어 전략 개발
	원본 데이터 손실	원시/소스 데이터를 추적하기 위한 데이터 변환 프로세스 문서화
	불완전한 실 조건 테스트	교육 및 테스트 데이터 외에 실제 데이터로 철저한 테스트를 수행, 외부 테스트에 크라우드소싱 그룹 포함
	AI 모델의 불충분한 투명성 및 해석 가능성	LIME, SHAP, 모델 증류, 의사결정 트리 및 규칙 추출과 같은 기법을 활용하여 의사결정 과정을 설명
	모델 과적합 또는 과소적합	정규화, 교차 검증, 조기 중지 등의 기술 구현
	윤리적 문제 발생	윤리 프레임워크를 수립, 윤리 원칙 준수 여부 감사 및 테스트

프로세스	발생 이슈	대응 방안
시스템 구현	채용 인공지능 시스템 출력 부정확	강력한 오류 처리 및 복구 메커니즘을 구현, 백업 및 복구 절차 구현, 정기적인 시스템 테스트 및 유지보수
	시스템 사양으로 인한 오류 발생	이기종 환경에서 채용 인공지능 시스템을 구현할 수 있도록 정확한 시스템 사양 정의 및 요구 사양 제시
	부적절한 테스트	기능 /비기능 종합적인 테스트 계획 수립. 엣지 케이스 시나리오 추가
	해킹/데이터 유출	암호화 및 액세스 제어 등 보안 조치 시행. 정기적인 보안 감사 실시
운영 및 모니터링	개인정보 및 데이터 보호 문제	데이터 암호화 여부, 개인정보 송수신, 저장 과정 취약점 점검, 개인정보 접근 및 반출 등 각 항목 정기적인 모니터링
	채용 인공지능 시스템 결과의 투명성 부족	기본 알고리즘에 대한 명확하고 간결한 설명 제공. 설명 가능성 및 해석 가능성 조치 구현. 투명성 보고서 제공
	무단 액세스, 시스템 손상 등 보안 취약점	정기적인 보안 감사 및 침투 테스트 실시. 방화벽, 침입 탐지 시스템, 액세스 제어 등 강력한 보안 제어
	윤리적 문제(개인정보 침해, 차별)	정기적인 윤리 검토 및 감사. 법적, 규제적, 윤리적 의무 이행 지침 마련
	잠재적 위해의 피해 사례	이해관계자에게 법적 책임을 부여하는 절차 마련
	예측 불가능성	시스템의 실패 요인을 예측하고, 예측할 수 없는 행동에 대한 안전장치 도입
	인공지능 생명주기 불투명성	인공지능 생명주기 정보 공개 상호 검증
	평가 개입 프로세스 부재	최종 의사결정권자인 전문가 피드백 기능 구현

안전성 | 다양성 존중 | 책임성 | 투명성

요구사항 02 인공지능 거버넌스governance 체계 구성

- 채용 인공지능 시스템의 의사결정이 개인의 삶과 경력에 미치는 영향이 윤리적 문제를 야기할 수 있다. 인공지능 신뢰성을 확보하려면 의사결정의 사회적 영향과 결과를 예측하고 대비하는 조직을 구성하는 것이 중요하다. 따라서 인공지능 관련 법, 규제, 정책, 표준 및 지침을 정리하여 내부적으로 준수해야 할 규정을 수립하고, 이를 관리·감독하는 인공지능 거버넌스* 체계를 구성해야 한다.

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

02-1 인공지능 거버넌스에 대한 지침 및 규정을 수립하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 관련 조직에서는 인공지능 시스템의 신뢰성을 확보하기 위한 거버넌스 체계를 구성할 필요가 있다. 인공지능 시스템은 학습이나 추론 과정에서 윤리 및 지식재산권 관련 이슈, 보안, 개인정보 보호 이슈에 직면할 수 있기 때문이다. 이러한 위험에 대비하기 위해 내부적으로 인공지능 거버넌스에 대한 가이드라인과 규정을 마련해야 한다.

인공지능 거버넌스 참고 가이드라인 및 규정

기관	가이드라인 및 규정
미국	일리노이주-인공지능 영상 면접 법(AIVI 법)[55]
	평등고용기회위원회(EEOC) 통일 가이드라인[56]
EU	일반 데이터 보호 규정(GDPR)[21]
	신뢰할 수 있는 AI를 위한 윤리 가이드라인[58]
캐나다	인공지능 및 데이터 법(AIDA)[57]
IEEE 표준 협회	자율 및 지능형 시스템 윤리에 관한 IEEE 글로벌 이니셔티브
IBM, 구글 등 학계, 시민사회, 산업계, 비영리 단체 등 100여 곳	AI 파트너십[59]
대한민국	개인정보보호위원회(PIPC)[60]
	채용절차의 공정화에 관한 법률[61]
	인공지능 윤리기준

- 내부적으로 수립해야 할 규정은 활용 측면에 따라 크게 두 가지로 구분할 수 있다.
  - ✓ 첫째, 인공지능 관련 법규, 규정, 정책, 표준, 가이드라인 등을 정리하여 조직 내에서 채용 인공지능 시스템 사용에 관한 지침과 규정을 수립
  - ✓ 둘째, 채용 인공지능 시스템 생명주기 전반에 걸쳐 조직의 역할과 책임을 명확히 문서화하고, 다양한 구성원과 개인의 책임 정의, 지속적인 모니터링, 개인정보 보호 및 보안의 보장 조치, 데이터 유출 및 사이버 보안 조치를 포함

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

Yes ☐ No ☐ N/A ☐

- 윤리 원칙의 수립은 인공지능 거버넌스 체계에서 기본적으로 갖춰야 할 단계로, 인공지능과 관련된 법, 규제 및 정책을 이해하고 내부적으로 윤리적 측면에서 이행해야 할 규정을 정의해야 한다. 즉, 인공지능 관련 위험을 인식하고 대비하기 위해 기업 성격에 맞는 핵심 가치를 선정하고, 이와 관련된 표준 및 지침을 채택하여 내부 규정을 제공해야 한다.
- 이러한 가이드라인과 규정은 채용 결정, 추천, 관할 지역의 규제 및 법적 프레임워크를 고려하는 등 해당 시스템의 구체적인 상황과 목표에 맞게 조정되어야 한다. 채용 인공지능 시스템 거버넌스를 위해 내부적으로 준수해야 하는 기존 외부 지침 및 규정은 다음과 같다.
  - ✓ 법률 준수: 채용 인공지능 시스템은 고용법과 같은 관련 법률 규정 및 요건의 준수 필요. 데이터 개인정보 보호법, 고용법, 차별 금지 등 참조
  - ✓ 편향: 데이터셋 또는 국가/문화적 차이와 관련된 문제/법적 의무로 인해 EEOC[62] 타이틀 7에 따라 의도적인 차별(차별적 대우)에 대한 클레임이 제기될 수 있음
  - ✓ 데이터 보호 및 개인정보 보호: 개인정보의 안전한 취급, 동의, 데이터 액세스, 보안 조치, 편향 및 차별 위험 최소화 등 데이터 보호 및 개인정보 보호 규정 준수
  - ✓ 윤리 가이드라인: 채용 인공지능 시스템을 공정하고 투명하고 편향 없는 방식으로 개발하고 사용하며, 프라이버시를 존중하고, 차별이 없도록 필요한 데이터만 수집. 채용 인공지능 시스템의 기능과 의사결정 프로세스는 사용자와 이해관계자가 투명하게 이해할 수 있어야 함
  - ✓ 품질 보증 기준: 신뢰할 수 있고 일관된 채용 인공지능 시스템 품질 표준을 준수하여 기능, 신뢰성 및 안전성에 대한 엄격한 테스트와 신뢰를 보장
  - ✓ 사이버 보안 규정: 사이버 보안 규정을 준수하여 면접 중과 면접 후에 민감한 정보를 보호하고, 보안 침해를 방지하며, 시스템 보안을 유지. 암호화, 액세스 제어, 방화벽, 사고 대응 절차, 정기적인 취약성 평가 등 포함
- 이러한 핵심 원칙 외에도 관련자들은 개발된 채용 인공지능 시스템을 관리하기 위해 다음 5가지 주제도 고려해야 한다.[64]
  - ✓ 주요 목표는 면접 대상자 평가이므로 공정성, 투명성, 건설적인 피드백, 전문성 개발, 경력 발전, 다양성 및 포용성 증진 등의 이점을 제공해야 한다. 긍정적인 평가 경험은 조직의 평판을 높이고 최고의 인재를 유치하는 데 도움이 된다.
  - ✓ 신뢰할 수 있는 채용 인공지능 시스템의 핵심 포인트 중 하나는 면접 채점 시 사용되는 '블랙박스' 개방이다. 따라서 지원자와 데이터/AI 사용 내용을 공유해야 한다. 데이터가 어떻게 사용되고 있는지, 무엇을 평가하고 있는지, 극단적으로는 데이터/면접/지원서 등을 삭제할 수 있게 하는 등 관련 정보를 공유해야 한다.
  - ✓ 또한, 채용 인공지능 시스템이 사용자 그룹의 두 부분구성원(면접관과 면접 대상자) 모두에게 설명 가능한지 확인하는 것도 중요하다. 채점 알고리즘이 어떻게 작동하는지, 시스템의 한계와 강점은 무엇인지 사용자에게 충분히 설명할 것을 추천한다.

✓ 학습 데이터의 내용을 고려하여 직무와 관련성이 있는 데이터로 채용 인공지능 모델을 학습시켜야 한다. 또한, 직무와 관련된 특성이나 기술을 찾아내 면접 채점에 신중을 기해야 한다.

✓ AI 결과를 사용하여 면접 대상자의 질문과 추측 문제를 해결한다. 채용 담당자는 최종 결정을 내리기 전에 AI 결과를 보조 확인 수단으로 활용하여 스스로 성찰하고 편향을 평가하도록 장려해야 한다.

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

Yes ☐ No ☐ N/A ☐

- 채용 인공지능 시스템은 채용, 승진, 해고와 같은 중요한 결정을 지원하며 고위험 시스템으로 간주될 수 있어 윤리적 및 법적 문제가 발생할 수 있다. 이에 대비하여 다양한 위험을 관리하고 규정을 준수하는 조직 필요하다.
- 채용 인공지능 시스템의 윤리적이고 투명하며 책임 있는 사용을 보장하기 위해서는 채용 인공지능 거버넌스를 위한 조직을 구축해야 한다. 조직은 다음을 포함하되 이에 국한하지 않고 다양한 배경과 전문성을 갖춘 개인으로 구성할 것을 권장한다: 윤리학자 및 철학자, AI 전문가, 인사 전문가, 법률 전문가, 데이터 과학자, 개인정보 보호 전문가 등

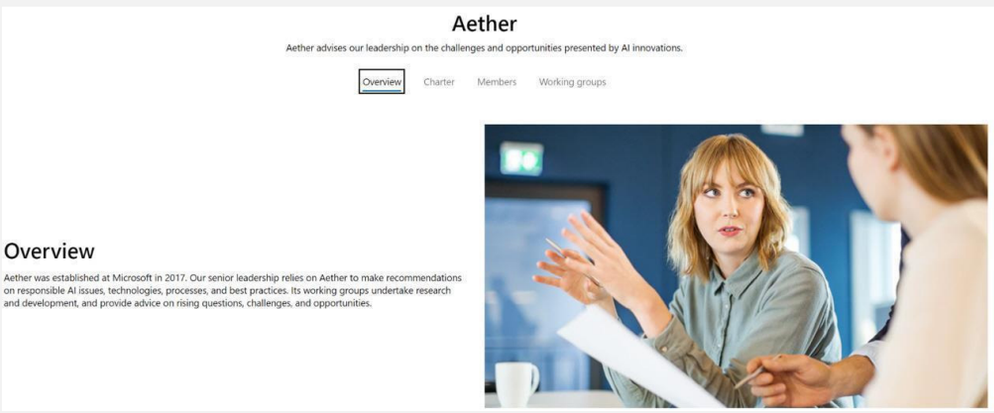
02-2a 인공지능 거버넌스를 위한 조직을 구성하였는가?

Yes ☐ No ☐ N/A ☐

- 조직의 윤리 원칙 수립 후 이를 실행할 수 있도록 관리하는 것이 인공지능 거버넌스 체계의 목표이다. 채용 인공지능 시스템을 윤리적이고 투명하게 운영하기 위해 조직은 AI, 윤리, 법률, 사회과학 전문가로 이루어진 팀을 조성하여 정책과 가이드라인을 개발하고 시스템 성능을 지속적으로 모니터링해야 한다.
  - ✓ AI 윤리 전문가: 채용 인공지능 시스템이 시스템적 편향을 줄이고 채용 시장의 공정성과 형평성을 증진하도록 설계되었는지 확인하는 역할
  - ✓ 법률 전문가: 채용 인공지능 시스템이 개인정보 보호법 및 채용절차의 공정화에 관한 법률을 포함한 관련 법률과 규정을 준수하는지 확인하는 역할
  - ✓ 채용 전문가: 채용 인공지능 시스템이 관련 기준에 따라 지원자를 평가하도록 설계되고 다른 평가 방법과 함께 사용되어 종합적인 평가가 이루어지도록 보장하는지 확인하는 역할
- 조직은 투명한 의사결정과 이해관계자 참여를 강조하며, 윤리적 우려에 대한 보고 및 해결 메커니즘을 구축해야 한다. 이를 통해 채용 인공지능 시스템이 윤리 원칙을 준수하고 피해나 차별을 방지할 수 있다.

참고 Microsoft의 책임 있는 AI 거버넌스[67]

AETHER: Microsoft에는 공식적인 AI, 엔지니어링, 연구 윤리위원회가 있다. 이 위원회는 회사 전체에서 AI의 개발 및 배포를 안내하는 역할을 담당한다. AI, 법률, 철학, 윤리, 공공 정책 전문가로 구성되어 있으며, 모든 새로운 AI 프로젝트를 검토 및 승인하고 Microsoft의 윤리적 원칙 및 표준에 부합하는지 확인하는 역할도 한다.



AETHER 위원회 외에도 Microsoft에는 자사 AI 제품 및 서비스의 사회적·환경적 영향을 평가하는 AI, 윤리, 엔지니어링 및 연구 효과 위원회인 RAISE도 있다. RAISE는 엔지니어링 그룹 전체에서 Microsoft의 책임 있는 AI 규칙 및 프로세스를 구현할 수 있도록 구축된 이니셔티브 및 엔지니어링 팀이다.

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

Yes ☐ No ☐ N/A ☐

- AI 생명주기 전반의 모든 프로세스에서 중심적인 역할을 담당하는 AI 거버넌스 담당 조직을 구성하는 인원은 자신의 역할과 책임을 충분히 인지하고 있어야 한다. 이들은 직원들에게 윤리적 영향, 잠재적 위험과 혜택, 기술적 측면을 잘 훈련하고 교육해야 한다.
- AI 거버넌스의 모든 측면을 다루기 위해 팀에는 다양한 분야의 전문가를 포함시키는 것이 바람직하다. 모든 이해관계자를 고려하여 시스템을 개발하고 사용하려면 여러 분야의 잘 훈련된 전문가로 이뤄진 팀이 필수적이다.
  - ✓ 채용 인공지능 시스템 거버넌스에 필요한 주요 전문 분야: 채용, 데이터 과학, 정보학, 인공지능 기술, 심리학, 음성 분석, 감정 분석, 법률, 인적, 디자인 사고 등
  - ✓ 채용 인공지능 시스템 거버넌스에 관여할 수 있는 전문가: 인사 전문가, 경영진, 심리학자, 데이터 과학자, 분석 전문가, 정보서비스 전문가, 인공지능 시스템 개발자, 변호사 등
- 또한, 채용 인공지능 시스템을 감독하는 전문가팀은 최신 개발 동향과 모범 사례를 파악하기 위해 꾸준한 교육과 훈련을 받아야 한다. 이 팀은 데이터 처리, 모델 개발 등 다양한 측면의 교육을 받아야 하고, 채용 인공지능 시스템 거버넌스와 관련된 윤리, 법률, 규제에 대한 이해를 갖추어야 한다.

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

Yes ☐ No ☐ N/A ☐

해당여부 판단

02-1에 따라 인공지능 거버넌스에 대한 가이드라인과 규정을 마련한 경우 이 항목을 고려하여 충족 여부를 판단하십시오.

- 인공지능 거버넌스 체계를 운영하는 주체는 운영 결과를 책임져야 하고, 이 책임은 위임할 수 없다. 따라서 운영 담당자는 조직이 내부 지침 및 규정을 준수하는지를 감독해야 한다. 또한, 거버넌스 시스템을 개선할 수 있는 영역을 식별하기 위해 발생할 수 있는 모든 사고 또는 문제를 검토 및 분석할 수도 있어야 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 사용을 감독해야 한다고 언급하고 있다.

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

Yes ☐ No ☐ N/A ☐

- 인공지능 거버넌스에 대한 내부 지침 및 규정 준수를 감독하기 위해 조직은 규정 준수 전담팀을 구성하거나 책임자를 지정해야 한다. 규정 준수 팀(책임자)은 거버넌스와 관련된 위반 사항이나 우려 사항을 조사하고 해결할 수 있는 권한을 가져야 한다. 다음은 AI 거버넌스 내부 지침 및 규정에 따른 감독의 예이다.
  - ✓ 정기 감사: 내부 지침 및 규정을 준수하는지 확인, 채용 인공지능 시스템과 프로세스에 대한 정기적인 내/외부 감사
  - ✓ 교육 및 인식 제고 프로그램: 직원을 대상으로 정기적인 교육 및 인식 프로그램 실시, 지침과 규정 숙지 및 역할 이해 교육
  - ✓ 성과 모니터링: 채용 인공지능 시스템의 성능을 정기적으로 모니터링하여 가이드라인 및 규정에서 벗어나는 문제나 편차를 파악, 시스템의 성과를 정기적으로 검토
  - ✓ 돌발 상황 관리: 사고 보고, 조사 및 시정 조치 실행을 위한 절차 등 돌발 상황 관리 계획 수립
  - ✓ 보고 : 전용 보고 채널 설정, 경영진 보고, 가이드라인 및 규정 위반 보고 등
- 규정 준수팀은 법무, 인사 등 다른 관련 부서 및 이해관계자와 긴밀히 협력하여 정책과 절차를 개발하고 시행해야 한다. 또한, 해당 내용을 직원들에게 교육하고, 감사 및 위험 평가를 수행하여 모든 문제나 사고가 신속하고 효과적으로 해결되도록 노력해야 한다. 아울러 고위 경영진 및 이사회에 정기적으로 보고하고 소통함으로써 규정 준수 노력을 우선순위로 유지하고 채용 인공지능 거버넌스 시스템에 필요한 변경이나 업데이트가 이루어지게 할 수 있다.

참고 IBM의 AI 거버넌스 구축 사례[71]: AI Ethics

IBM은 책임 있는 AI 개발 및 배포를 위한 일련의 원칙을 포함하는 AI 윤리 프레임워크를 개발하였다. 이 프레임워크에는 일련의 거버넌스 프로세스, 정책, 관행이 포함되어 있어 AI 솔루션이 이러한 원칙에 부합하게 하고, 책임 있고 윤리적인 방식으로 AI 솔루션을 개발 및 배포하며, AI 개발 생명주기 전반에 걸쳐 윤리적 고려 사항이 포함되도록 보장한다.

IBM의 거버넌스 구조에서는 AI 윤리 이사회가 AI 개발 및 배포와 관련된 윤리적 고려 사항에 대한 지침을 제공하는 등 다양한 역할을 담당하고 있다. 법무, 규정 준수, 사이버 보안 등 다양한 부서의 대표자로 구성된 이사회는 AI 프로젝트를 검토하고 승인하며, 진행 중인 AI 배포를 모니터링하고 발생하는 모든 윤리적 문제를 해결하는 책임이 있다. 더불어 AI 윤리 및 관련 분야 외부 전문가로 구성된 AI 윤리위원회를 설립하여 새로운 윤리적 문제와 동향 지침을 제공한다.

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

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템이 무분별하게 개발되면 서비스 사용자의 혼란을 가중할 뿐만 아니라 시스템 개발 및 유지보수에 불필요한 예산 사용을 초래한다.
- 새 시스템과 기존 시스템의 차이를 분석하면 새 시스템을 적절하게 관리하고 통제하기 위해 새로운 정책이나 절차가 필요한 영역을 파악하는 데 도움이 될 수 있다.

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

Yes ☐ No ☐ N/A ☐

- 현재 채용 인공지능 시스템을 위한 FDA 또는 510(k)과 같은 확립된 규제 절차는 없다. 그러나 일각에서는 채용 인공지능 시스템을 포함하여 일반적으로 AI 시스템의 개발 및 배포를 위한 표준과 가이드라인을 개발하려고자 노력 중이다. 예를 들어, IEEE(전기전자기술자협회)는 채용 인공지능 시스템에 적용될 수 있는 AI 개발 원칙과 권장 사항을 포함하는 '윤리적으로 정렬된 설계' 프레임워크[72]를 개발하였다. 또한, 국제표준화기구(ISO)는 AI 시스템의 위험을 평가하고 관리하기 위한 지침을 제공하는 ISO/IEC 23894[73]를 개발하였다.
- 신-구 채용 인공지능 시스템의 비교는 신중하고 포괄적으로 이루어져야 하며, AI, 데이터 분석 및 관련 분야(예: 인사, 채용 등)에 대한 전문성을 갖춘 다분야 팀이 참여해야 한다. 또한, 비교 과정에서 윤리적 고려 사항도 고려해야 한다.

채용 인공지능 시스템 개발을 위해 분석해야 할 중요 포인트

분석 요소	분석 내용
정확성	직무 성과를 예측하거나 최적의 지원자를 식별하는 측면에서 기존 및 신규 채용 인공지능 시스템의 정확도를 측정한다.
데이터 개인정보 보호 및 보안	두 시스템 관련해 시행 중인 데이터 개인정보 보호 및 보안 조치를 평가하고 잠재적인 약점이나 개선이 필요한 부분을 파악한다. 새로운 시스템이 데이터 보호 및 개인정보 보호 기능을 개선했는지 고려한다. 예를 들어, 시스템에서 인종, 성별, 종교와 같은 민감한 데이터를 수집하거나 저장하는가? 그렇다면 이러한 데이터는 어떻게 보안 및 보호되는가? 면접 데이터(영상 데이터)와 면접 대상자의 개인정보(이름, 주소, 전화번호 등)의 기밀성 및 무결성을 보호하기 위한 강력한 조치가 시스템에 마련되어 있어야 한다.
사용자 경험	두 시스템의 사용자 경험을 평가하고 새로운 시스템이 더 사용자 친화적이고 직관적이거나 효율적일 수 있는 영역을 파악한다.
편향성	새 시스템이 기존 시스템의 편향성이나 불공정성 문제를 해결했는지 평가한다. 학습 데이터, 알고리즘 및 결과에서 편향성의 증거를 찾고, 새 시스템에서 개선되거나 악화됐을 수 있는 영역을 식별한다.
사용자 지정 가능성	채용 인공지능 시스템은 조직의 특정 요구에 맞추기 위해 면접 질문과 프롬프트를 조정하고 다양한 언어와 방언을 지원하는 기능이 필요하다. 또한, 새로운 시스템과 이전 시스템의 사용자 지정 가능성을 비교하여 확장성과 최대 지원 사용자 수를 평가해야 한다.
통합	원활한 운영을 위해 지원자 추적 시스템(ATS), 인재 관리 시스템 및 기타 관련 소프트웨어 등 다른 HR 시스템과의 통합 기능이 원활한지 비교한다.
확장성	채용 인공지능 시스템의 확장성은 면접을 많이 진행하는 조직에서 특히 중요하며, 정확도나 성능 저하 없이 대량의 면접 트래픽을 처리할 수 있어야 한다. 현재 시스템이 처리 가능한 규모와 최대 지원 사용자 수를 고려하여 새로운 시스템의 확장성과 최대 지원 사용자 수를 평가한다.
비용	채용 인공지능 시스템을 구현하고 유지 관리하는 데 드는 비용은 중요한 요소이다. 초기 설정 비용, 라이선스 비용, 지속적인 유지보수 비용을 포함한 비용 총액을 평가해야 하므로 두 시스템의 비용을 비교한다.
윤리적 고려 사항	채용 인공지능 시스템 도입 시 공정성, 책임성, 편향 방지와 같은 윤리적 고려 사항을 고려해야 한다. 이를 위해 윤리, AI, 데이터 프라이버시, 법률 준수 전문가로 구성된 팀이 분석을 수행한 결과를 토대로 새로운 시스템이 기존보다 더 윤리적인지를 판단하고, 필요한 개선 사항을 확인한다.

- 비교 분석 결과, 새로운 채용 인공지능 시스템은 같은 목적으로 시장에 출시된 기존 시스템과 동일한 수준의 안정성 및 유효성을 입증할 수 있어야 한다. 또한, 객관적인 기준과 근거, 검증을 바탕으로 채용 인공지능 시스템의 안전성을 확보해야 한다.

안전성 투명성

요구사항 03 인공지능 시스템의 신뢰성 테스트 계획 수립

- AI 시스템은 출력 결과에 불확실성uncertainty 요소가 도입된다는 점에서 기존 소프트웨어와 다르다. 따라서 시스템의 공정성을 보장할 품질 보증 테스트는 물론 별도의 신뢰성, 공정성 등 추가 테스트가 필요하다. 효과적인 테스트를 위해 인공지능 시스템의 복잡도complexity와 운영환경을 고려해 계획을 수립해야 하고, 계획에 따라 생명주기 전 단계에서 정기적·지속적 테스트를 해야 한다.

03-1 인공지능 시스템의 특성을 고려한 테스트 환경을 설계하였는가?

Yes ☐ No ☐ N/A ☐

- 유럽근로생활관측소(EurWORK)는 지원자의 경력과 생계에 중대한 영향을 미칠 수 있다는 점에서 채용 인공지능 시스템을 '고위험' 시스템으로 분류한다. 이 기구는 채용 인공지능 시스템이 특정 연령대, 장애인, 특정 인종 또는 민족 출신, 성적 지향과 관련된 차별을 영속화할 우려가 있음을 인정하였다[74]. 따라서 시스템이 부적절하게 설계되고 사용되는 것을 방지할 테스트 환경 및 계획이 매우 중요하다.
- 또한, 유네스코의 인공지능 윤리 권고안에 따르면, 인공지능 시스템은 인권을 잠재적으로 위협할 수 있는 것으로 확인됐다. 이에 따라 윤리적 영향 평가의 일환으로 출시 전 이해관계자들의 광범위한 테스트를 거쳐야 하고, 필요시 실제 상황과 동일한 조건에서 테스트를 진행하기를 권장하고 있다.
- 그러나 채용 인공지능 시스템은 복잡해 실제 환경에서의 테스트가 어려울 수 있다. 따라서 테스트 환경을 설계하기 전에 시스템의 아키텍처, 입출력 형식, 처리 메커니즘 등을 이해하는 것이 중요하다. 다음으로는 시스템 응답 정확성 검증, 다양한 시나리오에서의 성능 평가, 취약점 식별, 악의적인 행동 탐지 등의 목표를 설정하고, 다양한 테스트 시나리오를 식별한다. 테스트 데이터셋은 실제 환경을 모방하고 엣지 케이스와 이상값을 포함하는 등 다양성을 보장해야 한다. 또한, 적절한 테스트 도구를 선택하고 테스트 환경을 설계하여 프로덕션 환경과 유사하게 구성해야 한다.

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

Yes ☐ No ☐ N/A ☐

- 테스트 환경을 설계할 때 운영환경을 고려하면 실제 사용 시 발생할 수 있는 잠재적인 문제나 취약점을 파악하고 해결할 수 있다. 운영환경에는 시스템에서 사용하는 하드웨어 및 소프트웨어, 네트워크 연결, 잠재적인 간섭 또는 중단 원인 등의 요소가 포함된다. 테스트 환경은 시스템이 실제 조건에서 예상대로 작동하는지 확인하기 위해 최대한 운영환경에 가깝게 모방하여 설계해야 한다.
- 그러나 실제 환경에서의 테스트가 항상 가능하지는 않다. 제한된 조건에서 시험 환경을 설계할 때 고려해야 할 사항은 다음과 같다.
  - ✓ 인공지능 시스템의 운영환경이 복잡하고 다양한 이해관계자가 관련되어 있는가?
  - ✓ 해당 시스템이 개인의 인권 및 민감 정보 보호에 잠재적인 위협이 되는가?
  - ✓ 합리적인 시간과 비용으로 테스트를 수행할 수 있는가?
  - ✓ 시험 전 검증을 위해 특정 테스트가 필요한 항목은 무엇인가?
  - ✓ 전문가의 판단을 대체하기 위해 채용 인공지능 시스템을 구축하는 경우(채용 담당자/HR 전문가 대신 평가), 전문가가 테스트 결과를 검토할 수 있는 적절한 패널을 구성하였는가?
- 또한, 운영환경을 대표하는 완전한 데이터셋을 찾기가 어렵기 때문에 편향을 피하려면 다양한 종류의 사용자를 대상으로 시스템을 테스트해야 한다.
  - ✓ 채용 인공지능 후보자의 다양성과 엣지케이스를 확인하기 위해 변성테스팅 Metamorphic testing 기법을, 공정성과 일관성을 확인하기 위해, 설계한 테스트 케이스와 동일한 시나리오의 전문가 의사결정을 비교하는 전문가 패널 기법을 고려할 수 있다.

테스트 다양성 확보를 위한 확인 요소 예시

테스트할 요소	테스트 환경에 추가할 요소
바쁜 사무실과 같은 시끄러운 환경	다양한 소음 수준
다양한 배경, 문화, 억양을 가진 사람들	다양한 억양과 문화적 배경을 가진 다양한 참가자들
다양한 조명 조건	밝거나 어두운 환경과 같은 다양한 조명 시나리오
스마트폰 또는 태블릿과 같은 다양한 기기	화면 크기, 해상도, 처리 능력이 다양한 기기
다양한 언어	해당 언어에 능통한 참가자
외모 요인	헤어스타일, 이어폰 착용, 체형에 영향을 미치는 신체적 장애 등

참고 오픈 액세스 MIT 면접 데이터셋 테스트 환경의 테스트 시나리오 예시[83]

두 대의 카메라를 사용하여 만든 테스트 환경과 모의 면접의 시청각 녹화를 수집하기 위한 실험을 설정했다. 연구진은 MIT 학생과 전문 커리어 카운슬러를 자원 봉사자로 고용했다.

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

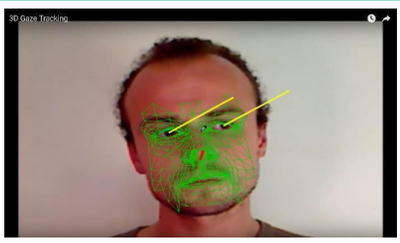
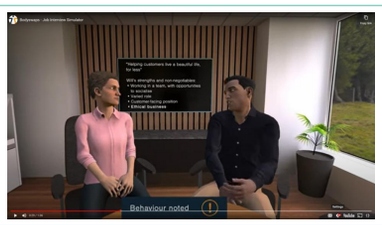
Yes ☐ No ☐ N/A ☐

- 채용 인공지능 시스템을 개발하는 동안 실제 테스트를 진행하거나 지원자를 찾기 어려울 수 있다. 이때 본격 테스트 전 가상테스트 환경에서 시뮬레이터를 통해 테스트하면 시간과 비용을 절약할 수 있다.
- **시뮬레이션은 가상 환경에서 통제되고 반복 가능한 시나리오를 수행할 수** 있다. 또한, 가상 환경은 다양한 인종, 억양 등 특정 시나리오와 조건을 테스트하기 위해 쉽게 수정 및 조정할 수 있으며, 데이터를 쉽게 수집하고 분석할 수 있다. 이를 통해 보다 다양한 후보자(입력데이터) 대상으로 채용 인공지능 시스템을 테스트할 수 있으며, 실제 환경에 시스템을 배포하기 전에 문제나 취약점을 식별하고 해결하는 데 도움이 될 수 있다[88].

  ✓ 가상테스트 환경을 만들 때는 다음 사항을 고려해야 한다.

  - 인종/출신 및 연령대의 다양성
  - 성별 인구의 비율
  - 자연어, 방언 및 언어 장애 관련 차이
  - 군말, 추임새, 욕설, 면접 중 재채기 등 드물게 발생하는 이벤트

채용 인공지능 적용 시뮬레이션 예시

시뮬레이터	내용	예시
시선기록기[89]	시선추적 데이터를 기록하고 분석할 수 있는 무료 시선추적 소프트웨어이다. 이는 면접 중 면접 대상자의 시선 패턴을 분석하는 데 유용하며, 이를 통해 주의력과 참여도에 대한 인사이트를 얻을 수 있다.	
유니티[90]	면접 시나리오의 인터랙티브 시뮬레이션을 만드는 데 사용할 수 있는 인기 게임 엔진이다. 이러한 시뮬레이션은 채용 인공지능 시스템을 훈련하고 다양한 조건에서 성능을 테스트하는 데 사용할 수 있다.	
Pixovr[91]	VR 면접 기술 훈련 모듈을 통해 사용자는 면접 시뮬레이션에서 자신의 기술과 성과를 파인튜닝할 수 있다.	

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

Yes ☐ No ☐ N/A ☐

- 대부분의 인공지능 시스템은 복잡도가 높아 재현가능성reproducibility이 떨어져 투명성 확보가 어렵다. 또한, 시스템의 복잡도는 기대 출력을 결정하는 테스트 오라클test oracle에 문제가 되기도 한다. 이에 따라 테스트가 통과 또는 실패했는지 여부를 판단하기 어렵다.
- 채용 인공지능 시스템은 인사 전문가에게는 적절한 선정 이유를, 지원자에게는 추론 결과에 대해 공개 가능한 범위에서 합리적인 설명을 제공할 수 있어야 한다. 시스템 출력을 확인하는 대상 사용자에 따라 설명 가능성*의 평가 기준이 달라질 수 있다. 인공지능의 작동 방식을 이해하는 정도인 해석 가능성interpretability의 평가 기준 역시 대상 사용자에 의존한다.

  * ISO/IEC TR 29119-11:2020에서는 설명 가능성을 '인공지능 시스템이 주어진 결과를 어떻게 도출했는지 이해하는 정도'로, 해석 가능성을 '인공지능 기술이 작동하는 방식에 대한 이해 정도'로 정의한다.

- 따라서 AI 시스템의 기대 출력에 대한 결정이나, 시스템 출력에 대한 설명 가능성 및 해석 가능성 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 ☐

- 채용 인공지능 시스템의 신뢰할 만한 기대 출력을 결정하기 위해 심리학자, 인사 전문가, 인공지능 연구자 등 해당 분야의 다양한 전문가가 참여하여 의견을 제시할 수 있어야 한다. 다양한 이해관계자를 협의 과정에 참여시킴으로써 채용 인공지능 시스템의 성능을 포괄적이고 효과적으로 테스트할 수 있다.
- 협의 체계는 시스템 설계와 테스트가 모든 이해관계자를 충족시킬 수 있도록 정기적인 회의와 토론은 물론 사용자 테스트 및 피드백 절차를 포함해야 한다.
- 협의체 전문가들은 하나의 입력값에 대해 서로 다른 기대 출력값을 예상할 수도 있다. 따라서 협의체 운영 전에 전문가 합의에 대한 승인 기준을 미리 정해 두어야 한다. 채용 분야의 경우, 대부분의 면접 평가 과제는 개발자가 BEI 역량 평가[92], 빅5 성격 특성이나 리커트 척도(예: 7점, 10점)를 이용하여 검사 결과를 척도화하는 경향이 있다[93][94][95][96][97].

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

Yes ☐ No ☐ N/A ☐

- 채용 인공지능 시스템의 설명 가능성을 보장하기 위해서는 기존 수준의 설명뿐만 아니라 다층적이고 포괄적인 정의가 필요한데, 이는 시스템의 투명성과 기능 간 균형을 유지하는 역할을 한다.
- 채용 전문가의 경우 채용 인공지능 시스템에서 생성된 결과물을 평가하는 역할을 맡을 수 있다. 또한, 시스템의 설명 가능성과 해석 가능성에 대한 피드백을 제공하여 이해하기 어렵거나 모호한 부분을 파악할 수도 있다. 면접 결과 제공 시 입사 지원자의 사용자 평가를 통해 면접 데이터를 수집하고, 사용자에 대한 설명 가능성과 해석 가능성을 개선할 수 있다. 이처럼 모든 사용자로 구성된 사용자 평가단을 구성하여 설명의 난이도를 결정하고, 이를 모델 또는 시스템 구현에 반영해야 한다.
- 사용자 평가단의 평가 결과에 따라 해당 결과가 유효한지(평균 이상)를 판단할 수 있는 기준(역량점수, 직무적합성, 상관관계 등)을 수립하고 테스트하여 평가 결과에 대한 의미 있는 설명을 보여줄 수 있다.
  - ✓ 해당 직무와 관련된 지원자의 기술, 경험 및 자격 측면에서 AI 시스템의 예측을 설명한다.
  - ✓ 해당 직무와 관련된 지원자의 기술 및 경험에 대한 구체적인 예를 제시한다.
  - ✓ 지원자를 역할이 유사한 다른 지원자와 비교한다.
  - ✓ 상관관계 분석을 통해 AI 시스템의 예측과 다른 측정값 또는 지표 간의 관계를 파악한다.

안전성 투명성

요구사항 04 인공지능 시스템의 추적가능성 및 변경이력 확보

- 채용 인공지능 시스템의 성능, 사용 그래프, 사용 습관 등을 추적하고 정기적으로 분석하여 시스템에 대한 이해나 지식이 부족한 사용자의 실수와 시스템 오류를 방지한다.
- 채용 인공지능 시스템의 운영 단계에서는 시스템 로그 확보, 데이터 모니터링, AI 모델과 인간 간의 의사결정 기여도 추적, 변경 이력 관리 등 문제 원인을 추적할 다양한 방안을 확보한다.

04-1 인공지능 시스템의 의사결정에 대한 추적 방안을 수립하였는가?

Yes No N/A
☐ ☐ ☐

- 채용 인공지능 시스템에서의 의사결정은 크게 두 가지 유형으로, AI 모델 자체에 의한 의사결정과 시스템 운영자나 사용자의 개입에 의한 최종 의사결정으로 구분된다. 시스템에 의한 의사결정은 개인에게 미치는 영향이 크기 때문에 의사결정의 주체, 의사결정에 대한 기여도 등을 분석하고, 특정 사건이나 사고 발생 시 책임 소재를 명확히 할 기반을 마련해야 한다.
- 채용 인공지능 시스템 서비스를 운영하는 도중에 계속해서 학습하는 방식으로 개발할 경우, 학습 데이터 및 모델에 대한 지속적인 모니터링이 필요하다. 예를 들어, 채용 인공지능 시스템이 지원자로부터 얻은 새로운 데이터를 지속적으로 모델에 공급하여 재학습한다면, AI 모델의 전체 생명주기를 고려한 추적 방법을 확보해야 한다.

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 정보를 추적할 수 있어야 한다.
- 따라서 인공지능 모델이 전적으로 의사결정을 내리는 경우와 모델 결과를 사람이 검토하여 의사결정을 내리는 경우, 주로 사람이 의사결정을 내리지만 특정 이벤트에만 보조적으로 모델의 추론 결과가 활용되는 경우 등 시스템 결정의 세부화된 기여도 기준을 내부적으로 확립하고, 시스템 운용 과정에서 이를 추적할 방안(예: 로그 수집)을 확보해야 한다.
- 채용 인공지능을 통한 의사결정은 인공지능의 전적인 판단과 채용 전문가(직업 상담사/HR 전문가)의 최종 검토에 의한 판단으로 나뉜다. 이를 구분하여 추적할 수 있어야 한다.

- 또한, 지원자의 영상 분석 작업을 통해 얻은 생체, 언어, 시각, 보컬 데이터의 평가 결과를 조합하여 의사결정에 활용할 경우, 결과에 대한 각각의 기여도를 세분화하여 사용자(면접관/채용 담당자)에게 제공해야 한다. 다음과 같은 다양한 추적 방법을 사용할 수 있다.

  ✓ AI 시스템의 의사결정과 해당 결정에 연결된 입력 관련 자세한 로그 보관, 기여도 분석, 반대 사실 분석(AI 시스템이 다른 결정을 내렸을 때의 결과를 평가), 기능 중요도 분석, 중요도 맵, 어트리뷰션 방법 등

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

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템의 전체 수명주기를 고려하여 추적성을 확보하려면 모델의 학습 과정, 운영 중 의사결정 결과, 인공지능의 구현 조정[106], 사용자 입력 데이터 등의 정보를 지속적으로 수집할 필요가 있다. 이를 위해 시스템 프로세스별로 로그를 수집할 정보를 선정하고, 해당 정보의 중요도를 정의하며, 로그를 수집하기 위한 로그 기록 형식을 결정해야 한다.
- 인공지능 시스템 운영 과정에서 발생하는 오류의 원인을 추적하기 위해서는 모델 구축 방식과 데이터셋 측면 등 오류의 원인을 분석해야 하므로 두 가지 측면을 고려하여 로그를 수집해야 한다.
- 로그 수집 기능을 구현하려면 필요한 정보를 캡처하고 AI 시스템의 의사결정 프로세스에 대한 포괄적인 기록을 제공할 수 있도록 신중하게 설계하고 구현해야 한다. 로그의 형식은 AI 시스템의 아키텍처와 캡처해야 하는 정보에 따라 달라진다. 로그 수집 기능을 설계할 때 필요한 정보를 제공할 수 있도록 로그의 형식을 고려해야 한다.

참고 AI 시스템에서 로그 수집 시 고려 사항

이슈	로그의 유용성
데이터 개인정보 보호 및 보안	로그에는 개인의 민감한 정보가 포함될 수 있으므로 로그 수집 기능이 이 정보의 개인정보를 보호하고 관련 데이터 보호법을 준수하는지 확인해야 한다.
데이터 보존	AI 시스템이 내린 결정은 장기적인 결과를 초래할 수 있으므로 로그를 얼마나 오래 보관해야 하는지, 로그에 액세스할 수 있는 사람은 누구인지 고려해야 한다.

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

Yes ☐ No ☐ N/A ☐

- 채용 인공지능 시스템에서 로그 관리는 지원자가 시스템과 상호작용하여 생성한 데이터를 수집, 저장, 처리, 합성, 분석하는 지속적인 프로세스이다.
- 시스템이 운영되는 동안 서비스 로그를 지속적으로 수집하면 서비스가 진행됨에 따라 다양한 형태의 데이터를 축적할 수 있다. 사용자 행동과 상호작용을 분석하면 시스템을 최적화할 수 있는 영역을 파악하여 성능과 전반적인 효율성을 개선하는 데 도움이 된다.

채용 인공지능 시스템에 대한 사용자 로그의 유용성 예시

예시	내용
데이터 수집	사용자 상호작용 기록: 채용 인공지능 시스템은 지원자의 응답, 탐색 패턴을 비롯해 시스템 내에서 수행한 모든 작업을 포함한 사용자 상호작용을 기록한다. 사용자 이벤트 캡처: 면접 시작과 종료 시간, 특정 질문에 대한 답변, 지원 채널과의 상호작용(가능한 경우) 등 사용자 이벤트를 추적한다. 데이터 익명화: 개인정보 보호 규정을 준수하기 위해 사용자 로그를 익명화 또는 가명화하여 지원자의 개인정보를 보호해야 한다.
데이터 저장소	중앙 데이터베이스: 분석을 위해 기록 데이터에 쉽게 액세스하고 검색할 수 있는 안전한 중앙 집중식 데이터베이스에 사용자 로그를 저장한다. 데이터 보존 정책: 데이터 보존 정책을 구현하여 사용자 로그를 보존할 기간을 설정함으로써 데이터 분석 요구와 데이터 저장 비용 및 개인정보 보호 문제 사이의 균형을 맞출 수 있다.
데이터 분석	사용자 경험 지표: 완료율, 응답 시간, 지원자 만족도 점수와 같은 주요 사용자 경험 지표를 정의하여 시스템 성과를 측정한다. 패턴 분석: 데이터 분석 기술을 사용하여 사용자 상호작용의 패턴, 추세 및 이상 징후를 식별하여 잠재적인 문제 또는 개선이 필요한 영역을 발견할 수 있다. 비교 분석: 시간 경과에 따른 사용자 경험 지표의 변화를 추적하고 시스템 업데이트 또는 변경의 영향을 평가한다.
모니터 및 알림	실시간 모니터링: 사용자 상호작용에 대한 실시간 모니터링으로 문제나 오류를 즉시 감지한다. 알림 메커니즘: 즉각적인 주의가 필요할 수 있는 중요한 문제를 기술팀에 알리는 알림 메커니즘을 설정한다.
지속적인 성능 개선	반복 업데이트: 데이터 분석과 사용자 피드백을 통해 얻은 인사이트를 활용하여 채용 인공지능 시스템을 반복적으로 업데이트한다. A/B 테스트: 모든 사용자에게 롤 아웃하기 전에 A/B 테스트를 수행하여 변경 사항의 영향을 평가한다.
규정 준수 및 개인정보 보호	개인정보 보호법 및 데이터 프라이버시 규정 준수: 개인정보 보호법(PIPA)과 같은 데이터 개인정보 보호 규정을 준수하고, 필요한 경우 명시적인 사용자 동의를 얻는다.
피드백 메커니즘	사용자 피드백 수집: 설문조사 또는 피드백 양식과 같은 피드백 메커니즘을 통합하여 채용 인공지능 시스템 사용 경험에 대한 지원자의 의견을 직접 수집한다. 사용자 피드백 분석: 사용자 피드백을 분석하여 사용자 불만 사항에 대한 인사이트를 얻고 시스템 개선을 위한 제안을 수집한다.

참고 인공지능 시스템 추적성 증진을 위해 고려할 수 있는 관행[265]

2020년 다보스 세계경제포럼 연차총회에서 발표된 '모델 인공지능 거버넌스 프레임워크(2차 버전)'에 따르면, 인공지능 기반 의사결정의 추적성 및 로깅을 확보하기 위해 다음과 같은 사항을 고려할 수 있다.

- 모델 학습 및 AI 증강 결정을 문서화하기 위한 감사 추적을 구축한다.
- 모든 입력 데이터 스트림을 캡처하는 블랙박스 레코더*를 구현한다. 예를 들어, 웹 브라우저의 쿠키와 같은 디자인 아이디어를 사용하여 채용 인공지능 시스템 사용자의 움직임과 이력을 수집 및 추적할 수 있다.
  *블랙박스 레코더는 AI 모델에서의 '블랙박스'를 의미하는 것이 아니다.
- 추적성과 관련된 데이터가 성능 저하 또는 변경을 피하기 위해 적절하게 저장/기록되고, 업계와 관련된 기간 동안 유지되도록 보장한다.

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

Yes ☐ No ☐ N/A ☐

- 채용 인공지능 시스템과 같이 위험도가 높은 분야는 책임성과 투명성을 보장하기 위해 학습 데이터의 변경 이력을 확보하고 데이터 변경의 영향을 관리해야 한다. 데이터 변경으로 모델의 설계나 주요 파라미터들이 함께 변경될 수 있다. 따라서 학습 데이터 버전 관리 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.
- 학습 데이터 변경 이력 관리를 위해 학습 데이터 버전을 관리할 오픈소스 도구 활용, 자체 시스템 구축 등을 고려할 수 있다. 또한, 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 성능 평가 결과 등의 정보를 제공해야 한다.

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

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템은 데이터 변경으로 인해 모델 확장이나 재설계 등 시스템 변경이 발생할 수 있다. 따라서 시스템 변경을 유도하는 데이터의 흐름 및 계보를 계속해서 추적해야 한다.
- 시스템의 데이터 흐름은 시스템의 다양한 단계에서 데이터가 이동하고 처리되는 것으로, 시스템의 여러 구성 요소를 통과하면서 수집, 처리, 변환되는 방식에 중점을 둔다. 데이터 흐름을 이해하는 것은 시스템의 데이터 처리 파이프라인을 최적화하고 잠재적인 병목 현상을 파악하며 데이터 무결성과 정확성을 보장하는 데 필수이다.

- 반면, 데이터 계보는 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 흐름과 계보는 데이터 변경의 역방향, 정방향 및 종단간end-to-end 관점에서 추적할 수 있는데, 추적 고려 사항은 다음과 같다.
  - ✓ 데이터 흐름 및 계보 추적을 관리하기 위해 데이터 정책팀을 구성하는 것이 유용한가?
  - ✓ 데이터 흐름 및 계보 추적을 위해 메타데이터를 기록하고 유지 관리할 것인가?
  - ✓ 데이터 흐름 및 계보 추적을 위해 데이터 로딩, 매핑, 관리 및 시각화 보고 기능을 구현하는 것이 유용한가?
  - ✓ AI 개발 과정에서 모델의 특징값을 저장하고 공유하는 특징 저장소feature repository 기능을 구현하는 것이 유용한가?
  - ✓ 데이터의 출처를 추적할 수 있는가?

04-2b 데이터 소스 변경에 대한 모니터링 방안을 확보하였는가?

Yes ☐ No ☐ N/A ☐

- 지원자 평가의 정확성, 공정성, 무결성을 보장하려면 채용 인공지능 시스템의 데이터 소스 변경에 대한 강력한 모니터링 계획을 세워야 한다. 지원자의 답변과 직무 요건 등 중요한 데이터 소스를 식별하고 우선순위를 지정함으로써 시스템은 가장 민감한 정보를 보호하는 데 집중할 수 있다. 초기 데이터의 견고한 기준선을 설정하면 향후 편차나 무단 변경을 감지하는 기준점이 된다.
- 이를 위해서는 정기적으로 시스템의 데이터를 감사해야 한다. 시스템이 사용하는 데이터 소스를 포괄적으로 문서화하고, 데이터 변경에 관련된 개인의 역할과 책임을 명확하게 정의하며, 모든 수정 사항을 효과적으로 추적 및 관리할 수 있는 중앙 저장소를 유지하는 것이 좋다.

참고 채용 인공지능 시스템의 데이터 소스 변경에 대한 모니터링 계획 예시

데이터 변경 관리 프로세스

- 채용 인공지능 시스템을 위해 특별히 설계하고 잘 정의한 데이터 변경 관리 프로세스를 수립한다. 지원자 답변, 지원 정보, 직무 요건 등 사용된 모든 데이터 소스를 문서화한다. 데이터 변경과 관련된 직원의 역할과 책임을 명확히 정의하고 중앙 집중식 저장소를 유지하여 수정 사항을 효과적으로 추적한다.

정기적인 데이터 감사

- 지원자 데이터와 평가 결과에 초점을 맞춘 정기적인 데이터 감사를 실시한다. 지원자 답변 및 평가 결과의 정확성, 완전성, 관련성을 보장한다. 지원자 정보를 보호하기 위해 개인정보 보호 및 보안 정책을 준수하는지 확인한다.

액세스 제어 강화

- 채용 인공지능 시스템에 맞춰 엄격하게 액세스를 제어한다. 특히 민감한 지원자 정보와 관련하여 권한이 있는 직원만 데이터에 액세스할 수 있도록 제한한다. 역할 기반 액세스 제어로 직무 역할에 따라 적절한 권한을 부여한다.

기밀 유지를 위한 데이터 암호화

- 강력한 데이터 암호화 기술을 사용하여 개인정보 및 평가 결과와 같은 민감한 지원자 데이터를 보호한다. 데이터 암호화는 지원자 정보의 기밀성과 개인정보 보호를 보장하여 무단 액세스의 위험을 완화한다.

실시간 모니터링 및 경고

- 데이터 소스를 실시간으로 모니터링하여 무단 변경이나 의심스러운 활동을 즉시 감지한다. 데이터 변경이 발생하면 관리자에게 즉시 알릴 수 있도록 경고 및 알림을 설정한다.

추적성을 위한 버전 제어

- 버전 관리 메커니즘을 활용하여 지원자 답변 및 평가 결과의 변경 사항을 추적한다. 데이터 변경에 대한 과거 기록을 유지하면 추적성을 보장하고 감사 및 디버깅 프로세스를 지원할 수 있다.

데이터 개인정보 보호 규정 준수

- GDPR 및 관련 데이터 보호법과 같은 데이터 개인정보 보호 규정을 엄격하게 준수한다. 데이터 처리에 대한 지원자의 명시적인 동의를 얻고 데이터 사용 및 보존 정책을 알린다.

직원을 위한 보안 교육

- 채용 인공지능 시스템 관리에 관여하는 모든 직원에게 포괄적인 보안 교육을 한다. 채용 인공지능 시스템과 관련된 잠재적 보안 위협 및 취약성을 교육하여 보안에 민감한 문화를 조성한다.

지속적인 개선 및 침투 테스트

- 모니터링 계획을 정기적으로 검토하고 개선하여 새로운 보안 위협과 업계 모범 사례에 맞게 조정한다. 정기적인 침투 테스트와 보안 평가로 취약점을 사전에 파악하고 해결한다.

04-2c 데이터 변경 시, 버전관리를 수행하였는가?

Yes ☐ No ☐ N/A ☐

- 채용 인공지능 모델을 개발하는 과정에서 학습 데이터를 업데이트하거나 오류로 인해 라벨링을 다시 수행하는 등 데이터 변경이 이루어지면 학습 결과물인 모델도 변경된다. 또한, 기존에 학습에 사용했던 데이터셋과 특성이 완전히 달라질 수 있고, 데이터셋 전체를 교체하면 성능이 현저히 저하될 수도 있다.
- 따라서 학습 데이터를 변경할 시에는 사용된 학습 데이터의 버전뿐만 아니라 해당 버전으로 학습된 인공지능 모델도 함께 관리해야 한다. 특히, 새로운 데이터가 추가되어 학습 데이터를 변경해야 하는 경우, 학습이나 테스트에 사용된 새로운 데이터의 비중을 기록하고, 이에 따른 모델의 성능 변화를 추적할 수 있어야 한다.
- 이를 위해 머신러닝 프로젝트에 오픈소스 기반의 데이터 버전 관리 도구(예: DVCData Version Control, 파키덤[pachyderm], Git 대용량 파일 저장소(LFS), 레이크FS, 델타 레이크 등)를 도입하거나 자체적으로 학습 데이터 버전 관리 시스템을 구축하여 학습 데이터의 버전 및 모델의 버전을 관리해야 한다.

오픈소스 기반 데이터 버전 관리 도구

도식	도구 소개
	데이터 버전 관리 DVC, 데이터 버전 관리 도구[266] 오픈소스 Visual Studio 코드 확장 및 명령줄 도구이다. Git 리포지토리 위에서 작동하며, 명령줄 인터페이스와 흐름은 Git과 유사하다. • DVC는 데이터 및 머신러닝 테스트를 문서화하고, 대용량 파일, 데이터셋 디렉터리, 머신러닝 모델 등을 작은 메타파일(Git으로 처리하기 쉽도록)로 대체하여 관리한다. 즉, 소스 코드 관리와는 다른 별개의 소스 데이터를 의미한다. 데이터 저장: 온프레미스 또는 클라우드 스토리지를 사용하여 프로젝트의 데이터를 코드 베이스와 별도로 사용할 수 있도록 한다.
	Pachyderm[267][268] 완벽한 버전 제어 데이터 과학 플랫폼이다. 커뮤니티 에디션 버전은 오픈소스로, 어디서나 배포할 수 있다. 엔드투엔드 머신러닝 생명주기를 제어하는 데 도움이 된다. 능력 • 레파지토리의 마스터 브랜치에서 데이터를 지속 업데이트 • 유형과 크기, 파일 수에 제한 없이 지원 • 커밋은 중앙 집중식 트랜잭션 • 팀이 서로의 작업을 기반으로 데이터셋을 빌드, 공유, 변환, 업데이트
	Git Large File Storage (LFS)[267][269] 오픈소스 프로젝트이다. Git 내에서 대용량 파일을 텍스트 포인터로 대체한다. Git 플랫폼에서 대용량 파일의 버전을 관리할 수 있다. 동시에 워크플로나 다른 Git 리포지토리와 동일한 액세스 제어 및 권한을 유지한다. 원격 작업 기능이 있다.
	lakeFS[267][270] Git과 유사한 브랜치 및 커밋 모델을 제공하는 오픈소스 솔루션이다. 격리된 브랜치의 변경을 허용한다. 활용 범위 • 실험할 수 있는 레이크의 스냅샷을 격리 • 사용자가 만든 규칙에 따라 데이터를 입력하고 관리 • 데이터의 변경 사항을 신속하게 되돌리고, 데이터셋에 일관성을 제공하며, 연쇄적인 품질 문제를 방지하기 위해 프로덕션 데이터를 테스트
	Delta Lake[267][271] 데이터 레이크의 신뢰성을 위해 사용한다. 데이터 프로세스의 ACID 트랜잭션, 메타데이터 제어 및 일괄 처리를 제공한다. 활용 범위 • 메타데이터 처리 • 배치 스트리밍 및 통합 • 스키마 작업 • 직렬화 가능 격리, 레벨 조정 • 데이터 버전 관리, 히스토리 제어, ML 실험 재현 • 병합, 업데이트 및 삭제 작업 지원

04-2d 데이터 변경 시, 이해관계자를 위한 정보를 제공하는가?

Yes ☐ No ☐ N/A ☐

- 채용 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.

이해관계자별 제공 정보 예시

이해관계자	제공 정보
지원자	– 변경된 데이터가 지원자에게 중대한 영향을 미치는 경우, 데이터 변경 사항 및 평가 프로세스에 미치는 영향 – 수집하는 데이터 유형, 자격 평가에 사용하는 정보 – 개인 데이터 보호 및 데이터 개인정보 보호 규정 준수에 대한 알림
채용 담당자 및 채용 관리자	– 데이터에 적용된 구체적인 변경 사항과 지원자 평가에 미치는 영향 – 채용목표에 관련한 데이터 변경 근거 – 신규 데이터 사용 지침
데이터 과학자 및 AI 개발팀	– 변경 사항의 특성과 그 이유 및 구제적인 데이터 수정 정보 – 새 데이터를 사용한 모델의 성능평가 결과 – 모델 매개변수 조정사항 및 평가 결과에 대한 기여도 정보 – 새 데이터의 사양 및 모델 관련 문서 등 리소스에 대한 접근 제공
고위 경영진 및 의사결정권자	– 데이터 변경의 광범위한 목표와 목적, 조직의 전략적 채용 계획과의 연계성 – 데이터 수정으로 인해 예상되는 혜택과 결과 – 변경에 따른 잠재적 위험과 완화 전략 – 데이터 변경이 채용 AI 시스템의 전반적인 효율성과 효과에 미치는 영향 진행 상황에 대한 정기적 보고

- 채용 분야와 같이 민감한 개인정보를 처리하는 인공지능 시스템의 경우, 데이터의 변경뿐만 민감한 데이터의 보안이 침해되는 경우에도 그 위반 사항에 대해 지원자에게 알려야 한다.[21]

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

Yes No N/A
☐ ☐ ☐

- 신규 데이터를 확보하고 인공지능 시스템에 사용하기 위해서는 기존 운영 중인 인공지능 모델과 성능을 비교해야 한다. 사람이 보기에는 신규 데이터가 기존 학습 데이터와 유사해도 학습된 인공지능 모델이 기존 학습 데이터에서 학습한 데이터와 특성이 다를 수 있다.
- 따라서 채용 평가 프로세스에 대해 대표적인 인공지능 알고리즘을 활용한 성능 평가 및 신규 데이터 분석이 필요하다. 신규 데이터 확보에 따른 성능 평가는 아래의 프로세스를 참고할 수 있다. 이 과정에서 반드시 인사 전문가/직무 상담사와 협업해야 한다.
- 평가 프로세스에는 일반적으로 새로운 데이터의 모델 성능을 이전에 확인된 데이터의 성능과 비교하고, 모델이 계속해서 좋은 성능을 유지하도록 필요한 조정을 하는 과정이 포함된다. 또한, 새로운 데이터를 학습 집합에 통합하면 모델의 일반화 능력이 향상돼 향후 보이지 않는 새로운 데이터를 더 잘 처리하도록 준비할 수 있다.
- 이 평가는 발생한 변화에 따라 다른 프로세스를 따를 수 있다. 새로운 데이터 확보에 따른 성능 평가에 적용할 수 있는 프로세스에는 다음과 같은 것들이 있다.
  - ✓ 성과 평가와 비교 분석을 위한 기존 학습 모델 및 관련 대표 인공지능 모델 확보
  - ✓ 채용 인공지능 시스템 및 모델에 적합한 성과 평가 지표 선정
  - ✓ 성능 평가를 위한 실험 설계(정량적, 정성적 실험 방법 선정, 실험 모델의 파라미터 설정, 세부 실험 계획 등)
  - ✓ 실험 진행 및 결과 분석(결과를 바탕으로 새로운 데이터를 평가하거나 필요한 경우 모델 재설계, 확장, 재교육 등 결정)

참고 성능 비교 평가 프로세스 예시

새 데이터를 유효성 검사 집합과 테스트 집합으로 나눈다.
필요한 경우 유효성 검사 집합을 사용하여 모델의 하이퍼파라미터를 파인튜닝한다.
정확도, 정밀도, 리콜, F1 점수 등 관련 메트릭을 사용하여 테스트 세트에서 모델의 성능을 평가한다.
새 데이터 모델의 성능을 이전에 본 데이터 성능과 비교하여 모델의 성능이 저하되었는지 또는 개선되었는지 평가한다.
새 데이터로 모델을 재학습하는 등 모델에 필요한 조정을 통해 성능을 개선한다.
결과를 철저히 분석하여 모델이 특정 예측을 하는 이유를 이해하고 데이터에 잠재적인 편향이나 문제가 있는지 파악한다.

책임성 투명성

요구사항 05 데이터 활용을 위한 상세 정보 제공

- 채용 인공지능 시스템의 훈련과 테스트에 사용하는 데이터는 고도의 개인정보를 포함하기 때문에 개발자는 대부분 자체 데이터셋을 구축한다. 이로 인해 데이터의 특성이 하드웨어, 데이터 유형(예: 음성, 영상, 텍스트), 수집 정책 등에 따라 다양하게 변할 수 있으며, 시스템의 호환성을 보장하기 위해 다양한 하드웨어 장치에서 수집된 데이터를 통합해야 한다.

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

Yes ☐ No ☐ N/A ☐

- 채용 인공지능 시스템은 얼굴이 포함된 이미지, 음성이 포함된 영상, 이름, 아이디, 주소 같은 개인정보 등 민감한 데이터를 다루기 때문에 데이터 확보는 물론 특징characteristics 검증이 어려울 수 있다. 따라서 추후 데이터를 재활용하거나 동일한 형식의 추가 데이터를 수집해야 할 경우 원시 데이터를 정제 및 가공하기 위한 메타데이터가 제공되어야 한다.
- 또한, 학습 데이터, 메타데이터, 라벨링 작업 가이드 등 데이터를 제공해야 개발자뿐만 아니라 기타 이해관계자들도 해당 데이터를 이해하고 편향이나 오류 발생을 방지할 수 있다.

채용 인공지능 시스템의 명확한 데이터 이해와 활용을 위해 공유해야 할 정보 예시

공유할 정보	세부 정보
데이터 원본	이력서 데이터베이스, 채용 공고 등 모델 학습에 사용한 데이터의 출처
데이터 품질	데이터의 정확성, 완전성, 관련성 등 품질 평가 요소 및 평가 결과
데이터 전처리	중복 제거, 누락된 데이터 처리, 데이터 정규화 등의 처리 내용
데이터 가공	누가 데이터에 라벨을 붙이는가(채용 전문가, 제삼자), 라벨링 프로세스가 전문가의 검증을 받았는가, 채용 전문가가 라벨링 후 라벨링 작업의 유효성을 검사하는가 등
데이터 개인정보 보호	수집한 개인정보의 보호 및 기밀 유지 내용, 학습 사용 데이터의 비식별화와 무단 액세스 또는 공개로부터 보호하기 위한 적절한 보안 조치 마련 여부 등

05-1a 정제 전과 후의 데이터 특성을 설명하였는가?

Yes No N/A
☐ ☐ ☐

- 채용 인공지능 시스템의 정제 전후 데이터의 특성은 시스템의 구체적인 목표에 따라 달라질 수 있다. 시스템 특성상 도메인 지식에 기반한 정제 작업이 필요하며, 채용 전문가가 참여해야 한다. 이 과정에서 조직의 데이터 큐레이션, 데이터 거버넌스 확보 등의 절차가 진행되어야 하며, 정제 전후의 데이터 특성에 대한 정보를 명시하여 이해관계자가 데이터를 적절히 활용할 수 있도록 해야 한다.
- 면접 인공지능이 수집하는 데이터는 주로 비디오, 음성 등의 비정형 데이터로, 머신러닝 모델 활용에는 적합하지 않아 적절한 정제가 필요하다. 정제 작업을 통해 어떤 부분을 구조화하고 수정했는지 그 과정과 결과를 기록해야 지속적인 업데이트 및 감사가 가능하다.
- 또한, 서류 또는 면접 영상에서 얻은 데이터는 직무 경력이나 얼굴, 음성 등 개인을 특정할 수 있는 정보를 포함한다. 해당 데이터는 개인정보 비식별 조치 가이드라인을 준용하여 가명처리, 총계처리, 데이터 삭제, 범주화, 데이터 마스킹 등의 처리 후 기록해야 한다. 아울러 데이터셋 제공자(데이터 주체)가 부여한 권한 상태를 항상 확인해야 한다.
- 데이터셋을 구축하는 과정에서 데이터 품질 향상을 위해 일부 데이터를 정제하는 과정을 거치는데, 정제 후 학습 데이터의 특성으로 설명할 수 있는 항목의 예는 다음과 같다.
  - ✓ 데이터 속성 분석 항목: 중복 방지, 이상 데이터 제거, 샘플링 등
  - ✓ 통계적 설명 항목: 클래스별 학습 데이터 수, 피험자 수 등
  - ✓ 환경 설명 항목: 촬영 지역(예: 지리적 위치, 실내 온도), 촬영 시간대, 촬영 장소(예: 실제 장소, 회사, 대학, 시나리오 환경 등), 피험자 정보, 피험자 간의 관련성 정보 등

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

Yes No N/A
☐ ☐ ☐

- 면접 인공지능 시스템의 경우, 학습 데이터에는 면접 대상자의 채용 여부를 나타내는 라벨과 함께 이전 면접의 녹취록이 포함될 수 있다. 모델은 이 데이터를 사용하여 패턴을 학습하고 새로운 면접 대상자에 대해 예측할 수 있다. 반면 채용 인공지능 시스템의 메타데이터에는 면접 대상자(예: 이름, 이력서, 입사 지원서), 면접관(예: 이름, 직책, 면접 대상자에 대한 피드백), 면접 과정(예: 면접 날짜와 장소, 질문, 면접 시간) 관련 정보가 포함될 수 있다.
- 학습 데이터와 메타데이터를 분리하기 위해 일반적으로 별도의 파일이나 데이터베이스에 저장한다. 학습 데이터는 AI 모델을 학습시키는 데 사용되고, 메타데이터는 면접에의 맥락과 추가 정보를 제공하는 데 사용된다. AI 모델의 출력 또는 면접을 기반으로 한 최종 채용 결정과 같이 학습 데이터와 메타데이터 외에 추가 정보를 저장해야 할 수도 있다.

학습데이터와 메타 데이터 구분 항목 및 데이터셋 예시

메타데이터 사양	
지원자 정보	이름, 이메일, 직위, 경력 수준 등 지원자 정보를 입력한다. 이를 통해 AI 시스템이 면접 경험을 개인화할 수 있다.
면접관 정보	면접관의 이름, 직위, 면접 스타일 등 면접관 관련 정보를 입력한다. 이를 통해 AI 시스템은 면접관의 접근 방식에 맞게 질문과 답변을 맞춤 설정할 수 있다.
직무 요구사항	필요한 기술 및 경험과 같은 직무 요건 관련 정보를 제공하여 AI 시스템이 지원자를 보다 효과적으로 평가할 수 있도록 지원한다. 직책, 고유 직무 ID, 범주, 직급, 부서, 고용 유형, 위치, 원격 근무 옵션 등이 포함된다. 메타데이터에는 직무의 책임과 중요성을 요약한 포괄적인 직무 설명도 포함된다. 교육 요건, 경력, 필요한 기술 및 자격증을 포함하여 주요 책임과 최소 자격 요건이 명시되어 있다.
평가 기준	의사소통 능력, 기술 지식, 문제 해결 능력 등 AI 시스템이 지원자를 평가하는 데 사용하는 평가 기준에 대한 정보이다. 이를 통해 AI 시스템이 일관되고 공정하게 지원자를 평가할 수 있다.
학습 데이터 사양	
데이터 형식	채용 인공지능 시스템은 면접 대상자의 응답을 분석하기 위한 텍스트, 음성 특성을 분석하기 위한 오디오 또는 표정 분석을 위한 비디오 등 특정 형식의 면접 데이터를 제공해야 할 수도 있다.
라벨링	데이터의 어떤 부분이 질문이고 어떤 부분이 답변인지를 표시하기 위해 라벨을 지정해야 할 수도 있다.
출처	면접관, 면접 대상자의 이름, 면접 날짜 등 면접 데이터의 출처를 기록해야 할 수도 있다.
품질	AI 시스템이 면접을 정확하게 처리하고 이해하기 위해서는 노이즈 제거를 제거하는 등 면접 데이터의 품질을 높여야 할 수도 있다.

오픈 액세스 데이터셋 아키텍처 분류와 소개

데이터셋 배포자	데이터 이름	형태	인지	판단	통제
AI 허브	Multimodal Video	Audio, Video, Text	O (감정, 성별, 음성 스크립트, 사물 및 관계 정보 등)	X	X
Rochester HCI	MIT Interview Dataset	Audio, Visual	O (표정, 언어(단어 수 등), 운율 정보)	O	X
Computer Vision Center and University of Barcelona	First Impressions V2 (CVPR'17)	Audio, Visual	O (언어(필사본), 면접 주석), 성격 특성 등)	O	X
Speech Analysis & Interpretation Laboratory – University of Southern California	IEMOCAP Database	Video, Audio, Image, Text	O (표정, 감정, 인간 상호작용, 차원적 특징(원자가, 활성화, 우세 등) 등)	O	X
Engineering and Physical Sciences Research Council (EPSRC)	VoxCeleb	Video, Audio	O (사람의 음성, 얼굴 트랙, 성별, 발화 길이 등)	X	X

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

Yes ☐ No ☐ N/A ☐

- 보호변수*는 채용 인공지능 시스템에서 보호 속성을 사용하는 것을 의미하지만, 얼굴 이미지나 음성 데이터도 포함할 수 있다. 채용 인공지능 시스템의 대표적인 보호 변수로는 연령, 성별, 인종(민족), 장애 여부, 사회경제적 지위 등이 있다.

  * 단순히 입력값만이 아니라 보호변수에 따라 최종 결과에 편차가 나타날 수 있는지와 이를 완화하기 위한 조치가 필요한지 여부의 분석까지 의미한다.

- 특정 변수를 보호함으로써 직무와 관련 없는 요소가 아니라 오로지 지원자의 자격과 경험만으로 평가할 수 있도록 한다. 또한, 채용에 있어 차별 또는 공정과 관련한 법적 의무를 위반하지 않도록 관련 법령을 참고하여 보호변수를 적용 및 설명해야 한다.

보호변수를 선택할 때 고려해야 할 법적 의무 사항

관련 조항	의무 사항
채용 절차의 공정화에 관한 법률 제4조 3항	면접 등 채용 절차에서 구직자로부터 '구직자의 자녀 유무, 자녀의 연령 및 자녀 수' 등 개인정보를 취득해서는 안 된다고 규정한다.
남녀고용평등과 일-가정 양립 지원에 관한 법률 제7조	남녀고용평등과 일-가정 양립 지원에 관한 법률 제7조는 흔히 '남녀고용평등과 일-가정 양립 지원에 관한 법률'로 불리며, 대한민국에서 직장 내 양성평등 촉진과 일-가정 양립 지원을 목적으로 한다.
고용상 연령차별금지 및 고령자고용촉진에 관한 법률 제4조의4 및 제4조 5항[103]	제4-4조 1항[104]: 이 조항은 사업주가 정당한 사유 없이 연령을 이유로 개인을 차별하는 것을 명시적으로 금지한다. 이 조항을 위반하는 모든 차별은 불법으로 간주한다.
	제4-5조[105]: 이 조항에서는 연령 차별의 다른 측면을 다루거나 고령자 고용 촉진과 관련한 추가 세부 정보를 제공할 가능성이 크다.

- 특히 인종적으로 다양한 외국에서는 예상치 못한 윤리적 편향 사례가 보고되고 있다. 의도하지 않은 편향에 대비하기 위해서는 학습 데이터 수집 및 라벨링 단계에서 보호변수를 선정하고 반영하여 직무 요건 및 지원자의 자질과 직접 관련된 요인에 집중할 필요가 있다. 이러한 문제를 완화하는 데 도움이 되는 공정성 지표 및 다양한 편향 완화 기법과 도구가 있다.

참고 고용 편향 발생 및 보호변수 조사 사례

MIT Technology Review [111]: 2021년에 잘 알려진 두 가지 AI 면접 프로그램을 대상으로 영어 기반 직무 면접에 독일어로 된 위키피디아 항목을 그대로 읽는 답변을 입력함. 두 시스템은 억양만으로 영어점수를 판단하여 각각 9점 만점에 6점, 73%로 높은 평가를 출력함.

A사의 유명 사례: A사는 2014년부터 입사 지원자의 이력서를 검토하는 시스템을 개발했으나, 2015년 이 시스템에 성적 차별이 존재함을 알게 됨. 문제를 조사한 결과, 10년 동안 회사에 제출된 이력서에서 데이터셋을 얻었는데, 이 이력서 대부분은 남성 지원자가 제출한 것이었음.

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

Yes ☐ No ☐ N/A ☐

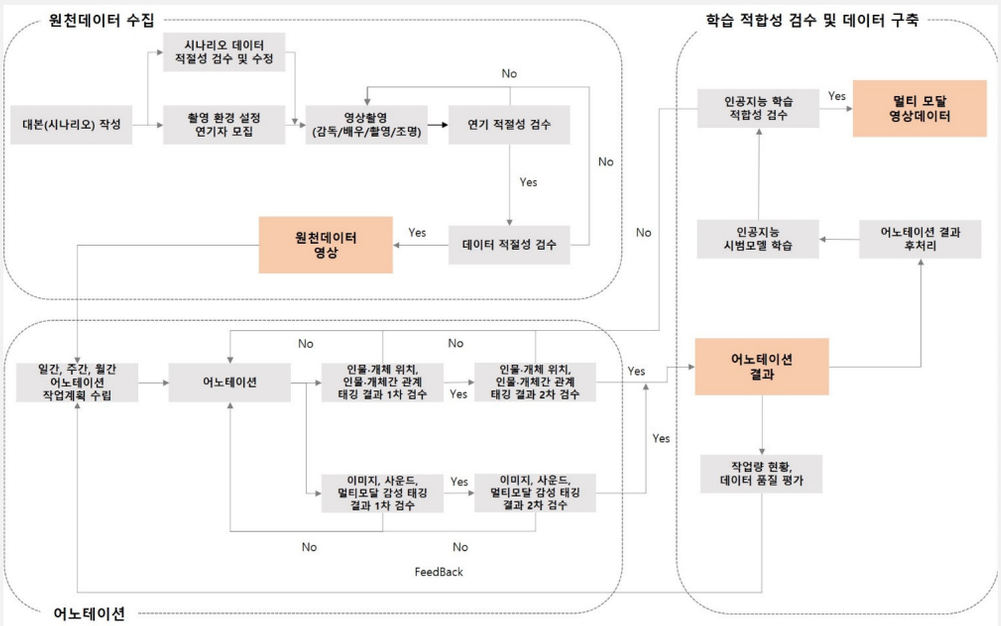
- 데이터 라벨링 작업은 인공지능 모델 학습을 위한 원천데이터의 주석(답변) 작업에 해당하는데, 면접 대상자의 면접 평가 과정은 전문적이고 세밀하게 진행되어야 하므로 전문 채용팀의 참여가 필수이다.
- 채용의 경우, 라벨링 프로세스에서 종종 주관성이 나타날 수 있으므로 이를 규제할 협의 체계의 역할이 중요하다. 다수의 전문가를 선정하고 합의하는 과정을 통해 채용 목적에 맞는 데이터셋 구축 기준을 마련해야 한다. 또한, 객관적 품질 확보를 위해 운영자 교육 및 세부 업무가이드를 반드시 문서화해야 한다.
- 데이터 종류에 따라 라벨링 작업의 대상, 범위, 세부 절차, 라벨링 도구 등이 달라질 수 있다. 채용 인공지능 시스템의 라벨링에 대한 자세한 내용은 요구사항 07에서 자세히 다룬다. 다음은 작업 가이드 및 교육에 포함해야 하는 내용의 예시이다.
  - ✓ 작업 지침, 라벨링 기준, 검수 기준, 라벨링 절차 정의
  - ✓ 라벨링 작업 예시(잘된 예, 잘못된 예, 자주 하는 실수 등)
  - ✓ 교육 세션 진행 및 성과 모니터링
  - ✓ 작업 지침 문서 준비

참고 라벨 제작자를 위한 작업 가이드 AI 허브 예시[113]

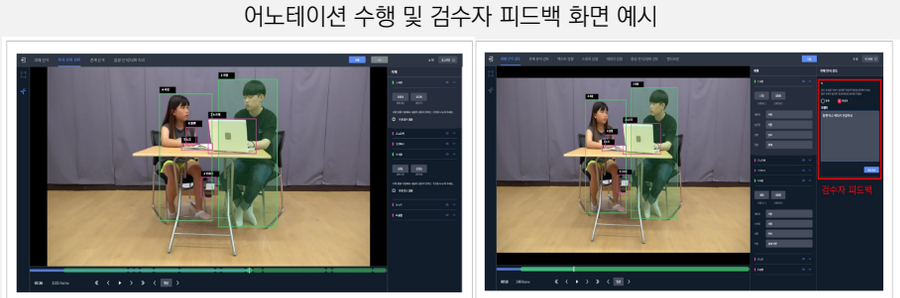
'멀티모달 비디오' 데이터셋 가이드라인의 데이터 항목에 대한 주석 기준 예시

어노테이션 대항목	데이터 항목	어노테이션 기준
인물의 정보 및 감정	인물별 감정	- 발화 텍스트, 발화 음성, 얼굴 표정, 멀티모달의 4개 모달리티에 대해 8종의 감정 기입 - 인물별 감정 8종 (기쁨, 슬픔, 분노, 놀람, 공포, 경멸, 혐오, 중립) - arousal(감정의 강도): 1(약함)~10(강함)의 값을 가지며 중간값은 5 - valence(감정의 긍부정도): 1(부정)~10(긍정)의 값을 가지며 중간값은 5
	인물별 성별	- 2종 (남, 여)
	인물별 연령대	- 7종 (10대 이하, 10대, 20대, 30대, 40대, 50대, 60대 이상)
	인물별 발화 스크립트	발화스크립트
개체 정보 및 관계	개체 정보	개체 위치 개체 분류
	관계 정보	약 20종의 개체 관계(위치/행동 관계)
발화 정보 및 의도	상황 설명 정보	대화 전체의 주제를 15종 이상으로 분류
	발화별 대화 의도	진술/주장, 질문, 명령/요청, 약속, 표출, 응대/답변, 인사/부르기/환기, 기타의 8종 태깅
	발화별 대화 전략 분류 정보	CMU-RAPT의 자기대화, 질문대화, 공감대화, 칭찬대화, 비윤리적대화, 완곡대화, 비언어적대화의 7종 태깅

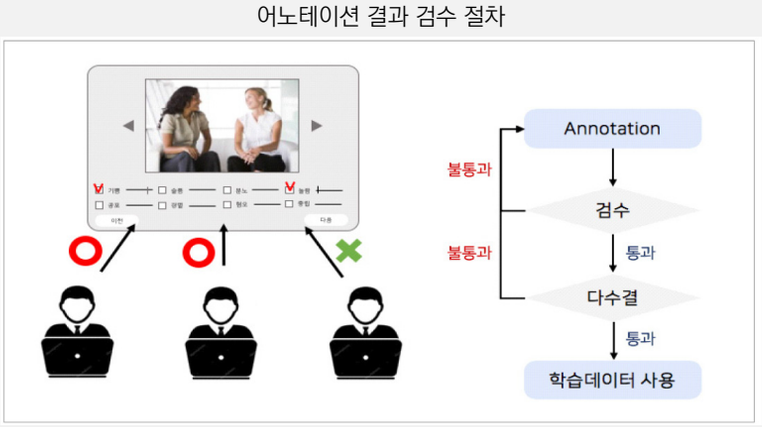
'멀티모달 비디오' 데이터셋의 주석/라벨링 프로세스 사례



어노테이션 수행 및 검수자 피드백 화면 예시



어노테이션 결과 검수 절차



05-2 데이터의 출처는 기록 및 관리되고 있는가?

Yes No N/A
☐ ☐ ☐

- 채용 인공지능은 매우 민감한 개인정보 데이터를 사용하기 때문에 그 출처를 신중하게 기록하고 관리하여 유출을 방지 해야한다. 또한, 학습 데이터의 품질은 모델의 성능에도 큰 영향을 미치므로 신뢰할 만한 출처를 선별하여 사용하는 것도 중요하다.
- 따라서 데이터 출처와 품질의 문서화 및 관리가 필요하며, 면접 대상자 데이터의 개인정보 보호를 위한 적절한 보안 조치 및 사용 동의 절차 등의 관리를 포함한다.
- 오픈소스 데이터셋을 사용하는 경우 변경 가능성에 대비하여 출처, 빌드 시간, 버전 등을 기록하고 관리하여 데이터 변경이 AI 모델 동작에 미치는 영향을 대응하고 추적할 수 있다.

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

Yes No N/A
☐ ☐ ☐

- 채용 인공지능 시스템의 정확한 작동을 보장하기 위해서는 반드시 출처를 신뢰할 수 있는 고품질의 데이터를 학습시켜야 한다.
- 따라서 직접 데이터를 수집하지 않는 경우에는 데이터 수집 방법, 샘플 크기, 데이터 품질, 개인정보 보호 조치 여부, 모델 관련성, 합법성 등을 사전에 점검해야 한다. WEF[World Economic Forum]는 오픈소스 데이터셋 활용 이전에 신뢰할 수 있는 데이터인지 미리 확인할 것을 권고한다.
- 채용에 응시하는 불특정 다수의 개인정보 데이터를 직접 수집하는 경우에는 데이터의 사용 범위, 보관 기관, 처리 등의 내용을 설명하고 데이터 사용에 대한 동의를 얻어야 한다.

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 제삼자에 의해 생산된 데이터 중계, 2가지 방법으로 데이터를 획득할 수 있다. 제삼자가 생산한 데이터를 중계하여 획득하는 경우, 데이터의 출처에 대한 신뢰성을 확보하여야 하며, 다음과 같은 요소를 고려할 수 있다.
  - 제삼자가 데이터 획득 시 개인정보 보호, 지식재산권, 사전 승인/허가 등과 관련하여 정식으로 절차를 밟고 문제없이 획득했는지 여부
  - 데이터 사용자가 원하는 학습용 데이터를 제공하는 데에 문제가 없을 만큼 제공하는 데이터셋의 규모가 충분히 큰지 여부
    예) 규모가 충분히 크지 않은 경우, 데이터를 재차 획득하고자 할 때 수급에 문제가 있을 수도 있음
  - 해당 데이터가 지속적으로 업데이트되고 추가 제공 등이 이루어지고 있는지 여부
  - 데이터와 함께 설계서의 내용이 명확히 제공되는지 여부
  - 해당 데이터의 활용 및 인용 건수가 많아서 범용성이 높은지 여부
- 데이터를 직접 생산(이미지/동영상 촬영, 발화 녹음, 텍스트 작성 등)하는 경우에는 위의 내용 중 첫 번째 사항을 고려하여야 한다.

05-2b 오픈소스 데이터셋을 활용하는 경우, 출처를 명시하였는가?

Yes ☐ No ☐ N/A ☐

- 채용분야와 같이 공개 데이터셋이 거의 없는 상황에서는 데이터셋이 모든 사례를 다루지 못해 예기치 못한 오류나 편향을 발생시킬 수 있다. 이는 채용의 결과에서 윤리적 문제와 관련될 수도 있어 지원자나 인권 단체로부터 소송을 당할 가능성도 있다.
- 따라서 오픈소스 데이터셋을 사용하여 인공지능 모델을 구축할 때는 편향된 데이터의 원인을 확인하기 위해 수집한 데이터의 명확한 출처와 관련 정보를 지정하고 관리해야 한다.
- 데이터셋의 출처 명시 외에도 데이터셋이 관련될 수 있는 지식 재산권과 관련 문서를 검토하여 모든 제약사항을 준수하는지 확인해야 한다.
- 오픈소스 데이터셋 관련 확인 사항:
  - ✓ 접근성: 데이터셋은 사용자가 제약 없이 쉽게 사용할 수 있어야 한다.
  - ✓ 라이선스: 데이터셋은 사용 조건이나 제약사항을 요약하여 명확하고 개방적인 라이선스를 가져야 한다.
  - ✓ 문서화: 데이터셋은 데이터 출처, 형식, 변수 및 기타 관련 세부 정보를 제공하는 명확하고 포괄적인 문서와 함께 제공되어야 한다.
  - ✓ 품질: 데이터셋의 데이터는 의도한 용도에 적합하고 정확하며 깨끗한 품질을 보장해야 한다.
  - ✓ 재사용성: 데이터셋은 다양한 응용 및 연구 질문에 쉽게 재사용하고 적용할 수 있도록 설계되어야 한다.
  - ✓ 투명성: 데이터셋은 수집, 처리 및 잠재적인 편향 또는 제약사항과 관련하여 투명해야 한다.
  - ✓ 커뮤니티: 데이터셋은 사용자 및 기여자가 적극적으로 참여하는 커뮤니티의 결과물이어야 하며, 피드백과 지원, 지속적인 개발을 제공할 수 있어야 한다.

안전성 다양성 존중

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 기법 적용을 통해 견고함을 확보한다.
- 통계적 방법과 기법을 사용하여 이상값을 처리할 때는 채용 전문가가 데이터를 교차 검토하여 제외할지 반영할지 결정해야 한다. 이러한 의사결정 프로세스는 모범 사례에 부합하면서도 잠재적인 편향이나 오류를 해결할 수 있도록 신중하게 문서화 한다.

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.
- 비정상 데이터를 식별하고 제거하는 것은 채용 인공지능 시스템의 전체적인 정확성과 신뢰성 향상에 도움이 될 수 있고, 특히 인종과 성별과 같은 편향이나 오류 가능성을 줄일 수 있다[129].

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

Yes No N/A
☐ ☐ ☐

- 데이터 정제 단계 이후 전체 학습 데이터 분포를 시각화하여 추가적인 오류나 이상값을 식별할 수 있다. 데이터가 정리되었더라도 인공지능 모델 학습에 부정적인 영향을 줄 수 있는 패턴이나 이상치가 여전히 남아 있을 수 있다. 데이터 분포 시각화는 데이터 분포 시각화는 인공지능 모델을 학습하는 데 사용된 데이터를 이해하는 데 많은 도움이 된다.
- 면접 평가 알고리즘에 사용하는 데이터셋의 다음 항목들을 시각화하여 데이터 유형의 분포를 확인하면 특정 그룹의 과잉 표현 또는 미표현 같은 데이터의 불균형을 식별할 수 있다. 또한, 통계적 정규성에서 벗어나는 값, 결측값, 일관성 부족, 측정 오류 등의 식별도 가능하다.
  - ✓ 성별 조건, 시간 조건, 피부색, 행동 상태(화남, 불안, 행복, 중립 등), 성격 특성(협조성, 외향성, 개방성, 성실성, 신경증 등), 사용 단어(전문 용어, 추임새, 욕설 등)

참고 채용 인공지능 시스템의 입력 데이터 오류를 확인하기 위한 시각화 분석 예시

성별 및 인종에 기반한 감성 분석 시스템의 편향 가능성 조사. 영화 리뷰 데이터셋을 사용하여 200가지 다양한 감성 분석 시스템으로 리뷰의 감성을 분석한다. 본 논문은 어떻게 시각화 분석이 채용 인공지능의 입력 데이터에서 편향이나 오류를 식별하는 강력한 도구가 될 수 있는지를 보여준다.

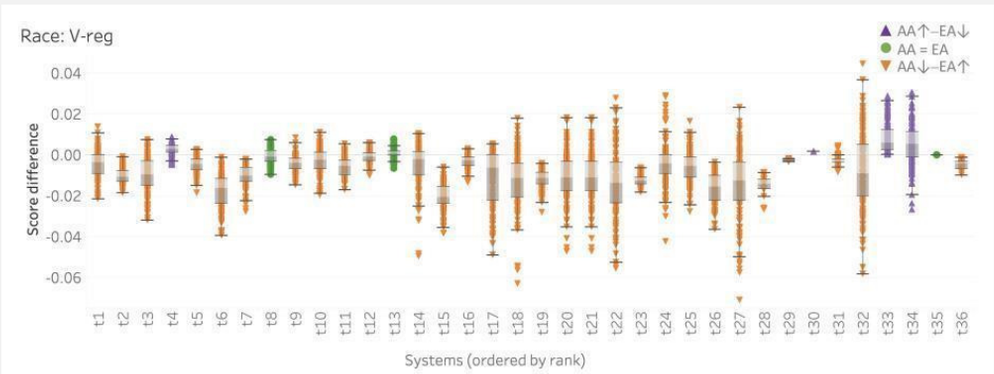
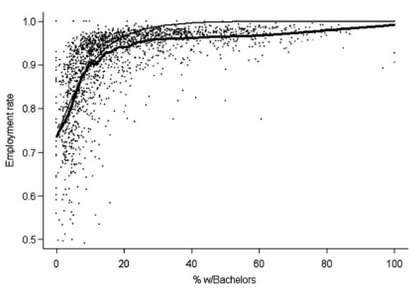
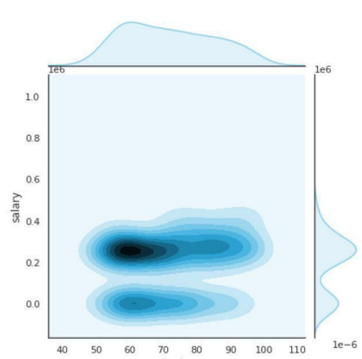


Figure 2: **Analysis of race bias:** Box plot of the score differences on the race sentence pairs for each system on the valence regression task. Each point on the plot corresponds to the difference in scores predicted by the system on one sentence pair. ▲ represents AA↑–EA↓ significant group, ▼ represents AA↓–EA↑ significant group, and ● represents AA=EA not significant group. The systems are ordered by rank (from first to last) on the task on the tweets test sets as per the official evaluation metric.

- 채용 인공지능 시스템으로 작업할 때, 전체 학습 데이터 분포를 시각화하는 다양한 방법이 있다. 데이터의 유형과 형식에 따라 이미지, 텍스트 파일, 음성 등이 가능하며, 그 특성에 따라 다양한 기법을 적용할 수 있다.
  - ✓ 분포 그래프(데이터 전체의 평탄성, 평균, 분산 및 편차를 이용한 데이터 분포 시각화)
  - ✓ 범주별 그래프(범주형 데이터 시각화)
  - ✓ 행렬 그래프(이차원 행렬 데이터 시각화)

데이터 분포 시각화 방법 예시

기술	설명	
히스토그램	**이력서 데이터 정리 후 데이터의 히스토그램 예시[130]** 	히스토그램은 숫자 데이터 분포를 시각적으로 나타낸 것이다. 이는 데이터가 가지는 값의 구간을 나누고 각 구간에 속하는 관측치의 수를 세는 과정을 포함한다. 각 구간의 수를 막대 그래프로 나타내면 분포의 형태를 시각적으로 파악할 수 있고, 이상치나 편향성 같은 특징들을 확인할 수 있다. 또한, 히스토그램은 정제된 데이터에서 특정 변수의 분포 불균형과 같은 이상한 패턴이나 편향성을 확인하는 데에도 유용하다.

기술	설명	
산점도	**산점도 데이터 시각화의 예시[132]**  Scatter plot of employment rate versus percentage of adults with university (Bachelor) degrees, across Chicago CMSA census tracts. Dark line is non-parametric regression (supersmoother), lighter line is job networking model calibrated to match.	산점도는 두 개의 수치 변수를 시각적으로 나타낸 그래픽이다. 각 변수를 x축과 y축에 나란히 그려서 각 관측값을 그래프상의 점으로 나타내면 데이터의 전반적인 분포와 가능한 패턴 또는 관계를 파악할 수 있다. 산점도는 데이터 내에서 군집화나 패턴, 이상치나 편향성을 식별하는 데 유용하다.
막대그래프	**이력서 데이터셋[130]의 기존 범주 분포를 보여주는 막대그래프 예시**  	막대그래프는 인구 통계 정보와 같은 범주형 데이터를 시각화하는 데 유용하다. 각 범주의 빈도를 막대그래프로 표시함으로써 데이터의 전체 분포와 가능한 편향 또는 불균형을 파악할 수 있다.
히트맵	**인도 비즈니스 스쿨의 데이터셋에서 '직업 가용성 점수 대 봉급-공동에 대한 히트맵' 예시[132]**  	히트맵은 정제된 데이터에서 서로 다른 변수 간의 상관관계를 파악하는 데 유용하다. 예상치 못한 고점은 오류나 편향의 표시일 수 있다.

06-1b 학습 데이터 이상값 식별 기법을 적용하였는가?

Yes ☐ No ☐ N/A ☐

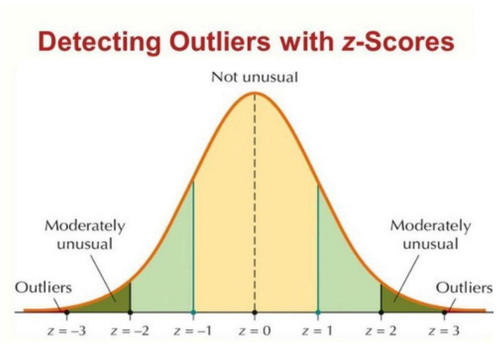
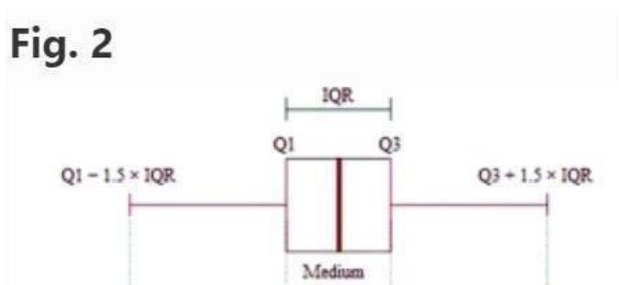
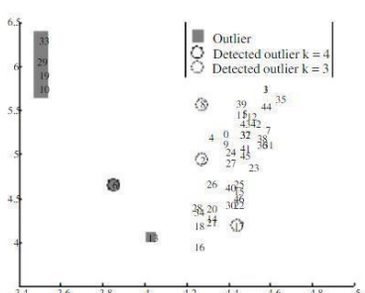
- 학습 데이터에 이상값이 존재하면 모델의 과대적합overfitting이나 과소적합underfitting으로 이어져 새로운 데이터의 일반화 능력이 떨어질 수 있다. 데이터셋을 인공지능 모델 훈련에 사용하기 전에 이상값 식별 기법을 적용하여 이를 방지할 수 있다.
- 이상값은 '이성적 이상값'과 '합리적 이상값'으로 나눌 수 있다. 이성적 이상값은 입력 오류와 같은 데이터 오염으로 인한 이상값을, 합리적 이상값은 정확하게 측정되었지만 다른 데이터와 완전히 다른 추세나 특성을 보이는 이상값을 의미한다.

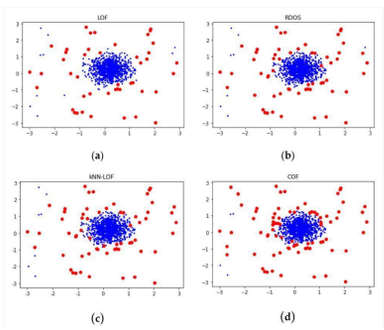
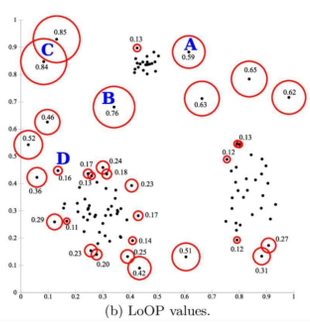
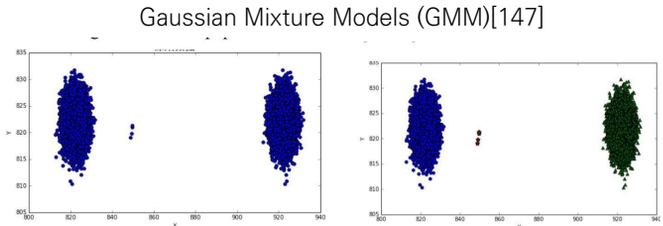
참고 면접 인공지능 시스템에서 사용되는 이상값 데이터의 예시

긴 응답	지원자의 답변이 길어지는 형태의 이상값은 채용 인공지능 시스템의 전반적인 성능에 영향을 미칠 수 있다. 긴 답변에는 관련 없는 정보가 포함될 수 있고, 이는 시스템을 혼란스럽게 하거나 정확도를 떨어뜨릴 수 있기 때문이다.
극단적인 값	매우 높거나 낮은 점수와 같이 극단적인 값 형태의 이상값도 채용 인공지능 시스템의 성능에 영향을 미칠 수 있다. 예를 들어, 특정 평가에서 다른 지원자보다 훨씬 높거나 낮은 점수를 받은 지원자는 이상값이 되어 전체 결과를 왜곡하기도 한다.
흔하지 않은 어휘	이상값은 지원자가 면접에서 사용하는 언어에서도 발생할 수 있다. 예를 들어, 지원자가 해당 분야에서 일반적으로 사용하지 않는 흔하지 않은 어휘나 용어를 사용하면 이상값이 되어 시스템의 언어 처리 기능의 정확성에 영향을 주기도 한다.
기술적 결함	화상 면접 중 배경 소음이나 연결 문제와 같은 기술적 결함도 데이터에 이상값을 생성할 수 있다. 이러한 이상값은 지원자 응답의 부정확한 분석으로 이어질 가능성이 있다.

- 이상값 외에도 모델 추론 결과에 부정적인 영향을 미칠 수 있어 식별하고 제거해야 하는 비정상 데이터가 있다. 채용 인공지능 시스템에서 이상 데이터를 걸러내기 위해 지도 학습, 비지도 학습, 실시간 이상 감지, 격리 포레스트Isolation Forest 모델, DBSCANDennsity-Based Spatial Clustering of Applications with Noise, SVMSupport Vector Machine, LOFLocal Outlier Factor, 자동 인코더를 사용한 이상 감지 등의 머신러닝 알고리즘을 사용하거나 비정상적인 픽셀 강도를 보이는 데이터를 필터링하는 등 엔지니어링 관점의 다양한 이상 감지 기법을 사용할 수 있다.
- 면접 데이터의 입력은 대부분 비정형 데이터다. 비정형 데이터에서 이상값을 식별하기 위한 접근 방식 중에는 데이터 유형에 특별히 맞춤화된 이상 징후 탐색 기법을 활용하는 것도 있다. 예를 들어, 이미지 데이터의 경우 이상 징후는 표준에서 크게 벗어난 개체나 패턴 형태로 나타날 수 있다. 비정형 데이터는 그 다양한 특성을 고려할 때 도메인별 이상 징후 탐지 방법이 필요할 수 있으며, 이상값과 비정상 데이터를 효과적으로 필터링하기 위해 여러 가지 기술을 조합하는 경우가 많다.

이상값 탐지 기법 예시

기법	이상치 탐지 방법 예시
통계적 방법	z-score[136][137]  *Source: Analytics Vidhya* interquartile range (IQR)[138]  Mahalanobis distance[139]
기계 학습 방법	k-nearest neighbors[140]  **Figure 1. Outliers in HR dataset detected with ODIN, with threshold T = 0** support vector machines[141] neural networks[142]

기법	이상치 탐지 방법 예시
근접 기반 방법	Local Outlier Factor (LOF)[143]  k-nearest neighbor (KNN)[144] methods
군집 기반 방법	유사한 데이터 포인트를 그룹화하기 위해 클러스터링 알고리즘을 사용한다. 그 이후에 이상값은 어떤 클러스터에 속하지 않는 포인트로 식별된다: DBSCAN, K-means[145]
밀도 기반 방법	데이터 포인트의 밀도를 활용하여 이상치를 식별하는 방법: 지역 이상값 확률(LoOP)[146]을(를) 사용  (b) LoOP values. Gaussian Mixture Models (GMM)[147]  Fig 3. Original data Fig 4. Result of detection
정보 이론 기반 방법	정보 이론을 활용하여 이상값을 식별한다: 최소 설명 길이(MDL)와 콜모고로프 복잡도를 사용

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 하여 특정 고유 정보(성별, 학력, 출신 지역, 외모 등)를 의도적 또는 비의도적으로 사용하는 등 적대적 공격에 노출될 수 있으므로, 데이터 수집 및 처리 단계에서 대처 가능한 방안을 검토 및 적용하는 것이 바람직하다.
- 면접 평가 작업에 대한 적대적 데이터 공격의 실제 사례는 아직 많지 않으며, 이론적 가능성은 계속 제기되고 있다. 데이터 수집 및 처리 단계에서는 데이터 최적화data optimization를 통해 적대적 공격에 방어할 수 있다. 적대적 학습adversarial training, 데이터 품질 개선, 데이터 노이즈 제거 등으로 모델이 적대적 사례에 견고하게 동작하도록 한다.

06-2a 데이터 최적화를 통한 방어 대책을 마련하였는가?

Yes No N/A
☐ ☐ ☐

- 채용 인공지능 시스템에서 데이터 공격은 모델의 정확성과 공정성을 훼손하거나 면접 대상자의 프라이버시를 침해하고 시스템을 사용 불가능하게 만드는 등 심각한 결과를 초래할 수 있다.
- 예를 들어, 데이터 중독 공격은 학습 데이터셋에 악성 데이터를 주입하여 모델의 결정 경계를 조작할 수 있다. 이로 인해 잘못된 예측과 모델 내 편향이 발생할 수 있다.
- 중요한 방어 수단 중 하나는 학습 데이터가 깨끗하고 이상 징후나 공격이 없는지 확인하는 것이다. 이는 이상값 탐지, 데이터 검증, 데이터 프로파일링과 같은 다양한 기법을 사용하여 데이터를 최적화함으로써 달성할 수 있다. 또한 적대적 데이터를 의도적으로 생성하여 이를 학습용 데이터로 활용할 수도 있다.

다양성 존중 | 책임성 | 투명성

요구사항 07 수집 및 가공된 학습 데이터의 편향 제거

- 인간의 배경/행동/성별/인종/신체적 차이 등과 같은 데이터의 성격상 채용 인공지능 시스템은 편향 없는 데이터를 수집하거나 제공받기가 어렵다. 편향을 완화하기 위해 데이터 수집, 특성 선택, 라벨링, 샘플링 등 데이터의 수집 및 가공의 전 과정에서 편향성을 확인하고, 이를 완화하기 위한 조치를 취한다.

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

Yes ☐ No ☐ N/A ☐

- 채용 인공지능 시스템의 목표는 누구를 채용할지 결정할 때 공정성을 유지하는 것이다[160]. 편향은 공정성 개념과는 반대되는 경향의 다양한 사회적·제도적 동향에서 발생할 수 있다. 이런 경향은 종종 사람을 통한 데이터 수집 과정에서 그대로 반영되기도 한다. 이를 방지 하려면 데이터 수집을 시작하기 전에 충분히 다양한 데이터 원천을 보장하고, 다각적인 팀을 구성하며, 지속적인 검토 절차를 수립하는 등의 적절한 방안을 마련해야 한다.
- 인적 편향 외에도 물리적 편향이 발생할 수 있는데, 이는 주로 데이터 수집에 사용되는 장비와 관련이 있다. 시각적이든 음성적이든 필요 데이터는 일정 수준의 정밀도를 갖추어야 한다. 실제로 표정이나 눈동자 움직임 같은 세부 정보를 추출하려면 충분한 해상도의 이미지가 필요하다. 음성의 톤 식별이나 발음 평가에서도 마찬가지이다. 수백 개의 다양한 장치에서 수집된 데이터에 알고리즘을 적용하기 때문에 처음부터 이를 고려하지 않으면 오류가 발생할 수 있다.
- 데이터 수집이 완벽할 수는 없지만, 편향성을 제거하기 위한 작업은 필수이다[161]. 따라서 다양한 유형의 장치를 사용해야 하며, 가이드라인을 준수하여 다양한 지역, 인종, 연령, 성별 데이터를 제공해야 한다[162].

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

Yes ☐ No ☐ N/A ☐

- 모델 학습 및 테스트에 필요한 편향되지 않은 데이터를 찾기는 쉽지 않기 때문에, 수집한 데이터의 사용 목적에 적합한 데이터셋을 구성해야 한다.
- 무의식적인 편견은 인공지능 생명주기의 모든 단계에 영향을 미칠 수 있다. 데이터 수집은 무의식적 편견의 영향을 받을 수 있고, 결정권자는 암묵적 편견에 따라 설계 판단을 내릴 수 있으며, 결과물의 작동과 해석 또한 편향에 따라 형성될 수 있다. 특히 데이터 자체에 오랜 고정관념과 같은 잠재적 편향이 포함될 수 있다는 것이 더 어려운 문제다.
  - ✓ 예를 들어, 자연어 처리(NLP) 알고리즘은 'CEO'를 '남성'으로, '비서'를 '여성'으로 연결하여 학습할 수 있다. 이는 학습 언어 자체에 잠재적 편향이 포함되어 있기 때문이며 NLP에서 매우 흔하게 나타난다. 그 결과 AI는 성별과 직업적 역할을 연관시키는 방식에서 본질적으로 편향되어 있다.
- 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.
- 데이터 수집 전에 반드시 작업에 대한 가이드라인을 마련하고, 특정 배경과 성향에 관계없이 다양한 인력을 채용하며, 충분한 검수 인력을 확보하여 작업자 간의 개인 편차를 줄여야 한다.

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

Yes ☐ No ☐ N/A ☐

- 데이터 수집 시 한 가지 유형의 장치만 사용하거나 특정 환경에서만 녹화하여 데이터 편향성으로 인해 알고리즘의 인식 성능에 문제가 발생할 수 있다. 따라서 데이터 수집 시 이러한 요소를 확인하고 대응할 방안을 마련해야 한다.
- 면접 인공지능을 위한 데이터는 카메라와 마이크를 통해 수집된다. 그러나 시중의 많은 디바이스가 알고리즘에 적용할 수 있는 입력 형태를 지원하지 않을 수도 있다. 다양한 입력에도 견고한 시스템을 구축하려면 여러 사양의 기기를 활용하여 데이터의 양과 다양성을 확보해야 하며, 수집 후 데이터 클렌징 및 검수가 충분히 이루어져야 한다. 다양한 장치 및 환경의 예시이다.
  - ✓ 다양한 카메라를 사용한다: 노트북, 데스크톱, 스마트폰, 태블릿 등 다양한 디바이스를 통해 데이터를 수집하여 다양한 관점과 환경을 포착한다.
  - ✓ 다양한 카메라 각도 사용: 전면, 후면, 오버헤드 등 다양한 카메라 각도에서 데이터를 수집하여 다양한 관점과 시야를 캡처할 수 있다.
  - ✓ 다양한 마이크 사용: 옷깃 마이크, 헤드셋 마이크, 내장 마이크 등 다양한 유형의 마이크를 사용하여 다양한 오디오 품질을 캡처할 수 있다.
  - ✓ 다양한 녹음 조건: 조용한 방, 시끄러운 공공장소, 실외 환경 등 다양한 환경에서 데이터를 수집하여 다양한 수준의 배경 소음을 캡처할 수 있다.
  - ✓ 다양한 조명 조건: 밝은 햇빛, 저조도, 인공조명 등 다양한 조명 조건에서 데이터를 수집하여 다양한 시각적 관점을 포착한다.

데이터 편향 점검 시 고려해야 할 하드웨어 사양 항목 예시

획득 유형	하드웨어 사양 항목
RGB Camera	칩셋 유형(예: CCD, CMOS), 해상도(픽셀), 시야각, FOV, 압축 방식(예: H.265, H.264 등), 스캔 방식(예: 프로그레시브, 인터레이스) 등
VICOM Camera	센서 모드(넥서스 2.11), 그레이스케일 모드, 스트로브 강도, 증폭 게인, 가속도 측정 활성화 등
마이크	다이내믹·콘덴서 종류, 주파수 응답 대역, 신호 대 잡음 비, 감도 등

- 다양한 데이터를 수집한 후에는 그 분포를 평가하여 다양한 하드웨어 사양과 녹화 조건에서 고르게 분포되는지 확인해야 한다. 또한, 사용된 하드웨어와 AI 모델이 생성한 출력들의 관계를 살펴봄으로써 하드웨어와 출력 간에 상관관계가 있는지 파악하고 조치해야 한다.

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.
- 채용 인공지능 시스템은 필연적으로 민감한 개인정보 데이터를 사용한다. 개인정보에 포함되는 일부 민감한 특성은 사회적 문제와 학습 결과의 차별을 야기할 수 있는 특성이다. 이러한 요인은 데이터 학습에 반영해서는 안 되는 특성으로 분류할 것을 권장한다. 국제기구나 글로벌 기업에서 언급하는 민감한 특성의 예는 다음과 같다.

민감한 특성의 예시

기관명	특징
UNESCO	연령, 성별, 인종, 민족·사회적 기원, 조상, 혈통, 언어, 종교, 정치적 이념, 국적, 출생 시 사회경제적 지위, 장애
ALTAI	연령, 성별, 인종, 민족·사회적 기원, 조상, 혈통, 언어, 종교, 정치적 이념, 소수민족, 재산, 출생, 성적 지향
ISO/IEC 24027	연령, 성별, 인종, 소득, 가족 관계, 교육 수준, 키·체중, 장애 여부
IBM Watson OpenScale	연령, 성별, 인종, 결혼 여부, 주소 소외된 인구(연령, 인종, 민족, 성 지향, 기타 소수 인구)
Google	인종, 성별, 장애, 종교

07-2a 보호변수 선정 시 충분한 분석을 수행하였는가?

Yes No N/A
☐ ☐ ☐

- 보호변수를 선택할 때 충분히 분석하지 않으면 모델에 편향 또는 성능 저하를 발생시킬 수 있다. 채용·선발 과정에서 요구 역량이나 기술과 관계 없이 민간한 데이터 특성으로 결과에 영향을 줄 수 있으며,최악의 경우 고용 차별을 이유로 소송에 휘말릴 수도 있다.
- 따라서 모델 추론 결과에 영향을 미치는 특성을 식별한 경우 데이터의 일부를 변경하거나 가중치를 재배치하면서 모델의 결과가 어떻게 변화하는지 관찰하고 분석하여 공정한 추론 성능을 확보해야 한다.
- 최근에는 다양한 공개도구(Google What If Tool, ML-fairness-gym, IBM의 AI Fairness 360, Aequitas, FairLearn 및 Google의 Facets, IBM AI 360 등)를 활용하여 데이터 변화에 따른 추론 결과를 유추하거나 시각화할 수 있다. 이를 통해 설정한 보호변수가 인공지능 의사결정의 차별을 일으키는 데 얼마나 영향을 미치는지, 성능이 어떻게 변하는지 파악할 수 있다.

07-2b 편향을 발생시킬 수 있는 특성의 영향력을 완화하였는가?

Yes ☐ No ☐ N/A ☐

- 채용 인공지능 모델 학습 시 성별, 인종, 나이, 언어와 같은 인구통계학적 정보등을 입력 특성으로 사용하면 모델이 다수 그룹에 편향될 수 있다.
- 이러한 특성의 영향을 완화하는 방법은 다양한 인구 통계 그룹, 문화 및 배경을 대표하여 학습 데이터를 최대한 다양화하거나, 학습 데이터의 대표성을 향상시키기 위해 데이터 밸런싱 및 데이터 합성 등 데이터 증강 기법을 사용하는 것이다. 그러나 모든 인구그룹의 데이터를 확보하는 것은 불가능에 가깝다.
- 다른 방법은 해당 편향을 유발하는 특성을 학습에서 배제하는 특성 선택기법eature selection을 적용하는 것이다. 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.
- 하지만 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 함께 고려해야 한다.

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

Yes ☐ No ☐ N/A ☐

- 데이터 정리, 데이터 변환 및 정규화와 같은 데이터 전처리 기술을 사용할 때 의도하지 않은 과도한 처리는 과적합overfitting 문제 또는 편향의 원인이 되기도 한다. 중요한 데이터의 손실을 피하려면 데이터 특성을 신중하게 분석해야 한다.

특성 분석을 위한 기술 예시

기술	예시	사유
기능 선택	상호 정보, 카이제곱 테스트, 상관관계 기반 특징 선택	가장 중요한 특징을 식별한다.
차원 축소	주성분 분석(PCA), 선형 판별 분석(LDA)	가장 중요한 정보를 보존하면서 특징의 수를 줄인다.
특징 추출	자동 인코더, 딥 러닝	본래의 특징에서 새로운 특징을 학습한다.

- 인종, 민족, 성별, 나이 등의 특성은 이미 의도치 않게 편향되어 있으며, 이는 개인 특성을 평가하기 위한 데이터셋의 한 종류로 사용될 수 있다. 이러한 상황에서 집중되거나 편향된 데이터를 제거하려면 인사 전문가 또는 전문 직업 상담사가 잠재적인 편향성을 검토해야 한다. 특성을 제거하지 않고 각 특징에 다른 가중치를 부여하거나 전체 데이터셋의 분포를 확인하는 등의 주의가 필요하다.

과도한 특성 선택 및 배제를 방지하기 위한 점검표

체크리스트	조치
도메인 지식이 있는가?	있다면 도메인 지식을 기반으로 임시 속성을 구성할 것을 권장한다.
특성이 서로 관련이 있는가?	그렇지 않다면 규모에 맞게 정규화할 것을 권장한다.
특성 간에 상호 의존성이 있는가?	그렇다면 관련 특성을 결합하여 인재 집합을 확장하는 것을 고려한다.
비용, 속도 등을 위해 입력 변수를 제거해야 하는가?	그렇지 않다면 특성을 분리하거나 특성 가중치의 합을 구성할 것을 권장한다.
모델의 특징을 이해하거나 필터링하기 위해 특징을 개별적으로 평가해야 하는가?	그렇다면 변수 순위 방법 사용을 권장한다.
Predictor가 필요한가?	필요하지 않다면 특성을 선택할 필요가 없다.
데이터가 지저분한가?	그렇다면 최상위 변수를 사용하여 이상값을 제거할 것을 권장한다.
무엇을 먼저 해야 할지 알고 있는가?	모른다면 선형 예측자를 사용하고, 포워드 선택 또는 제로 규범 임베디드 기법을 사용한다.
새로운 아이디어, 시간, 컴퓨팅 리소스, 충분한 데이터가 있는가?	그렇다면 다양한 방법을 시도해볼 것을 권장한다.
신뢰할 수 있는 솔루션을 원하는가?	그렇다면 여러 번 시도해보고 부트스트랩을 사용할 것을 권장한다.

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

Yes No N/A
☐ ☐ ☐

- 채용 분야는 데이터의 특성 때문에 적절한 오픈소스 데이터셋을 구하기 어려운 경우가 많아 대부분이 자체적으로 데이터셋을 수집한다[172][173]. 채용 인공지능 모델 학습을 위한 데이터의 라벨링하기 위해 개인 특성, 인간 행동 분석, 예측 속성 및 라벨링 작업의 전문 지식이 필요하다.
- 라벨링 작업 과정에서 작업자의 특정 의도가 반영되거나 실수로 인한 특성 정보 누락, 무의식적 판단 등으로 인해 라벨링 작업 중 편향이 발생할 수 있다. 따라서 다양한 인구통계학적 그룹, 문화 및 배경이 서로 다른 개인을 포함한 다양한 라벨러를 사용하여 데이터에 라벨을 지정해야 한다. 또한, 라벨러 간의 불일치를 식별 및 해결하고 일관된 라벨링을 보장하는 것도 중요하다.
- 이를 위해 라벨의 정의와 라벨링 프로세스의 명확한 가이드라인은 물론 엄격한 평가 절차도 제공해야 한다. 작업 중 발생할 수 있는 문제를 미리 인지하고 명확한 기준이나 업무지침을 마련하여 작업자에게 제공하고 교육함으로써 향후 편향과 같은 문제가 발생하지 않도록 방지할 수 있다.

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

Yes No N/A
☐ ☐ ☐

- 면접 시스템의 특성상 데이터는 고도로 전문화되어 있으며, 개인적 특성에 대한 잘못된 해석으로 인해 편향되는 경향이 있다. 명확한 라벨링 기준이나 가이드라인이 없으면 개인의 판단에 의존하는 편향이 발생하기도 한다.
- 이를 파악하고 예방하려면 전문가(직업 상담사, 인사 전문가)와 긴밀히 협력하여 표준화된 라벨링에 대한 가이드라인을 작성해야 한다. 여기에는 라벨 판단 기준 및 판단이 애매한 경우의 처리, 잘된 예, 잘못된 예 등 구체적이고 실질적인 내용을 포함해야 한다.
- 또한, 라벨링 도구 등의 작업 환경을 일관되게 제공하여 라벨러의 주관을 배제 하도록 도울 수 있다. 데이터 및 AI 시스템의 변경 사항을 반영하기 위해 필요에 따라 데이터 라벨링 가이드를 정기적으로 평가하고 업데이트해야 한다.
- 라벨링 가이드라인을 제공하여 편향을 완화하는 것 외에도, 채용 영역에서 라벨링 역할을 하는 전문 면접관의 고유한 가치와 평가 기준, 모델 학습을 일치시키는 접근 방식도 고려할 필요가 있다. 편향을 규제해야 하는 섬세한 균형이 필수로, 면접관의 평가 기준을 효과적으로 통합하여 평가 프로세스에서 인적 전문성과의 연계성을 강화할 수 있도록 모델을 구성해야 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 확보하기 위해 크라우드소싱crowdsourcing 의 도입을 고려할 수 있다. 크라우드 워커의 경우 사람의 음성 및 비디오 분석에 대한 전문 지식이 부족하기 때문에 편향이 발생할 수 있지만, 다양한 일반인들의 참여로 작업자 집단을 다양화할 수 있다. 07-3a에서 강조한 작업자 교육을 통해 편향의 위험을 완화할 수 있다.

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.
- 검수 작업 역시 표준과 가이드라인을 명확하게 정의하고 검수자에게 전달하여 이들이 일관성 있게 데이터를 검수할 수 있도록 해야 한다. 또한, 배경지식에 관계없이 라벨링 검수자가 라벨링 표준, 절차 및 도구를 이해하고 데이터를 검수할 수 있도록 교육하고 지원해야 한다. 또한 여러 검수자를 동일한 데이터 라벨에 할당하여 개인의 편향이 미치는 영향을 줄일 수도 있다.
- 또한 면접 영상 데이터셋에 존재하는 이벤트, 사람의 특정 행동 패턴 또는 시나리오를 분류하고 분석 결과를 검사해야 하는 경우 변호사, 인사 전문가, 상담사 등 전문 분야의 검수자를 투입해야 한다.

참고 머신러닝을 위한 라벨링에 심리학자 검수자가 필요한 이유의 예시

연구에 따르면 얼굴의 미세한 표정이 속임수를 감지하는 신뢰할 수 있는 수단이라고 한다. 그러나 실험 결과에 따르면, 미세한 표정을 관찰하는 훈련을 받았다 해도 거짓말쟁이와 진실한 사람을 구별하는 인간의 판단은 정확도가 우연보다 높은 정도로 나타났다[176]. 오히려 최근 머신러닝(ML) 기술을 사용하여 미세 표정을 식별하고 기만적인 진술과 진실을 구별하도록 학습된 인공지능 분야에서 더 유망한 결과가 나왔다.

매우 정확한 결과를 얻기 위해서는 특정 라벨링을 적용하고 이를 검증할 검사를 수행해야 한다.

07-4 데이터의 편향 방지를 위한 샘플링을 수행하였는가?

Yes No N/A
☐ ☐ ☐

- 현재 공개된 채용 인공지능 데이터는 주로 고소득 국가에서 수집된 것들이다. 따라서 고소득 국가내 특정 연령대(40세 미만 또는 이상), 인종 또는 민족(백인, 흑인, 아시아계) 등의 데이터 불균형이 쉽게 발생할 수 있다. 이 같은 클래스 불균형 문제는 소수의 클래스 분포를 제대로 학습하지 못해 의도치 않은 편향성을 야기하는데, 샘플링 기법을 적용해 불균형을 확인하고 완화 방안을 도출할 수 있다.
- 샘플링은 모집단에서 일정한 기준에 따라 데이터를 추출하여 표본을 만드는 기법이다. 일정한 기준으로 추출한 표본은 모집단의 분포를 적절히 나타내면서 실제 모집단의 계층 불균형으로 인한 편향을 방지할 수 있어야 한다.

07-4a 편향 방지를 위한 샘플링 기법을 적용하였는가?

Yes No N/A
☐ ☐ ☐

- 채용 분야의 데이터는 매우 민감하고 사적인 정보이므로 쉽게 수집할 수 없다. 어렵게 수집하여 모집단의 분포를 적절하게 대표하지 못하는 데이터셋은 연령, 성별, 인종 차별, 문화적 편견 등의 편향 요인이 될 수 있다. 학습요 데이터의 인구 통계학적 불균형을 완화하기 위해 여러 가지 오버 샘플링over sampling 기법을 적용하여 부족한 데이터를 추가할 수 있다.
- 대표적 오버샘플링 기법에는 SMOTESynthetic Minority Oversampling TEchnique가 있으며 이를 토대로 발전한 경계선 SMOTE[179] , ADASYNAAdaptive Synthetic Sampling[180], SVM SMOTE [177], K-평균 SMOTE [181] 등이 있다.

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.

안전성 책임성

요구사항 08 오픈소스 라이브러리의 보안성 및 호환성 점검

- 인공지능 모델의 설계 및 개발 단계에서 오픈소스 라이브러리는 개발 기간을 단축하고 최신 기술 트렌드를 빠르고 유연하게 적용할 수 있도록 돕는다. 이에 따라 개발자들 사이에서도 오픈소스 라이브러리 활용에 대한 관심이 높아지고 있다.
- 오픈소스 라이브러리를 사용하기로 했다면 해당 라이브러리의 버전을 지속 모니터링하고 신뢰할 수 있는 라이브러리인지, 정기적으로 업데이트되고 있는지 확인해야 한다. 또한, 라이선스 기준을 숙지하고 운영 및 보안 위험도 확인해야 한다. 이는 시스템의 효율성과 신뢰성 보장에 도움이 될 수 있다.

08-1 오픈소스 라이브러리의 안정성을 확인하였는가?

Yes No N/A
☐ ☐ ☐

- 오픈소스 라이브러리를 활용하면 채용 인공지능 시스템의 연구를 향상킬 수 있고, 얼굴 인식부터 음성 인식까지 다양한 분야와 협업 가능하며, 인사 부서의 면접 프로세스를 개선할 수 있다.
- 채용 인공지능 시스템에 라이브러리를 통합하기 전에는 안정성과 신뢰성을 확인하고, 정기적으로 최신 보안 패치와 버그 수정을 확인하여 잠재적인 안정성 및 보안 문제를 예방해야 한다.
- 오픈소스 라이브러리의 안정성 확인에는 문서, 릴리스 노트, 사용자 커뮤니티를 고려해야 하고, 통제된 환경에서 라이브러리를 테스트하여 공정성 측면에서 잠재적인 문제나 버그를 식별해야 한다.

08-1a 활성화된 오픈소스 라이브러리를 사용하였는가?

Yes No N/A
☐ ☐ ☐

- 안정성은 라이브러리 사용의 핵심 포인트 중 하나이다. 오픈소스 라이브러리는 라이브러리 제공자가 라이브러리 개선/업데이트를 중단하면 개발 중인 채용 인공지능 프로젝트에 위험을 초래할 수 있다[182]. 이를 방지하기 위해 가능하면 대규모 기업, 인사 전문가, 대학/정부 인턴십 센터 등에서 개발하거나 관여한 활성 라이브러리를 도입하는 것이 좋다.
- 활성 오픈소스 라이브러리에는 많은 사용자와 개발자 커뮤니티가 있어야 문제 해결과 기능 요청 및 지원에 도움이 된다. 또한, 일반적으로 적절히 문서화되어 있고 접근성이 뛰어나야 라이브러리를 코드에 쉽게 통합할 수 있다. 대규모 커뮤니티 덕분에 오픈소스 라이브러리에는 머신러닝, 자연어 처리, 컴퓨터 비전의 최신 기술이 포함된 경우가 많으며, 이는 비용효율적일 수 있다. 오픈소스 라이브러리를 선택하기 전에 커뮤니티 외에도 몇 가지 요소를 고려해야 한다.

- 오픈소스 라이브러리의 활성도는 각종 쿼리를 분석하여 정량 측정할 수도 있다. 요즘은 대부분의 플랫폼에서 이런 기능을 제공하므로 사용 전 확인해볼 것을 권한다.
  - ✓ GitHub: 오픈 이슈 수, 풀 리퀘스트 수, 마지막 커밋 날짜 및 시간, 기여자 수, 사용 횟수, 별 개수, 오픈소스 관련된 스택오버플로 질문 수, 오픈소스 다운로드 수, 구글 쿼리 결과 수 등 제공
  - ✓ 코드가 있는 논문Paperswithcode: 사용 라이브러리 목록, 참조 노문 수, 별점 등
  - ✓ 캐글 플랫폼은 '캐글 모델': 모델 목록, 모델 변형(S3d, Resnet50 등), 모델 관련 논문, 깃허브 링크, 참조 논문 수, 활동 통계, 업보트 수 등 제공

오픈소스 라이브러리를 사용하기 전에 고려해야 할 매개변수

매개변수	정의
미해결 이슈의 수	아직 해결되지 않은 미해결 이슈 또는 사용자가 보고한 문제 수를 나타낸다. 미해결 이슈 수가 많으면 유지 관리 및 지원이 부족하거나 라이브러리에 잠재적인 버그가 있다는 의미일 수 있다.
풀 리퀘스트 수	메인 코드베이스에 병합되기를 기다리는 개발자가 제안한 변경 또는 수정 사항의 수를 나타낸다. 풀 리퀘스트 수가 많으면 라이브러리에 적극적으로 기여하는 개발자 커뮤니티가 활발하다는 의미일 수 있다.
마지막 커밋 날짜	라이브러리에 마지막으로 코드 커밋한 날짜를 나타낸다. 커밋 날짜가 최근이라면 라이브러리의 개발 주기가 활발하고 지속적인 유지 관리가 이루어지고 있다는 의미일 수 있다.
기여자 수	라이브러리에 코드 또는 문서를 기여한 고유 개발자 수를 나타낸다. 기여자 수가 많을수록 활발한 커뮤니티와 다양한 기술 및 관점이 라이브러리에 기여하고 있다는 의미일 수 있다.
별 개수	라이브러리가 GitHub와 같은 플랫폼에서 받은 별 또는 좋아요의 수를 나타낸다. 별 개수가 많을수록 라이브러리의 인기와 커뮤니티의 지지가 높다는 의미일 수 있다.
사용 횟수	라이브러리가 다른 프로젝트에서 다운로드되거나 사용된 횟수를 나타낸다. 사용 횟수가 높을수록 인기 있고 널리 채택된 라이브러리일 수 있다.
스택오버플로 질문의 수	라이브러리와 관련된 스택오버플로의 질문 수를 나타낸다. 질문과 답변 수가 많으면 라이브러리 관련 커뮤니티가 활발하고 사용자를 위한 지원 및 리소스가 제공되고 있다는 의미일 수 있다.

참고 팩터 기반 오픈소스 라이브러리

Visual

Deepface: Python용 경량 얼굴 인식 및 얼굴 속성 분석(나이, 성별, 감정 및 인종) 라이브러리. 최첨단 모델을 감싸는 하이브리드 얼굴 인식 프레임워크이다: VGG-Face, Google FaceNet, OpenFace, Facebook DeepFace, DeepID, ArcFace, Dlib 및 SFace.

Human: AI 기반 3D 얼굴 감지 및 회전 추적, 얼굴 설명 및 인식, 신체 포즈 추적, 3D 손 및 손가락 추적, 홍채 분석, 연령 및 성별, 감정 예측, 시선 추적, 제스처 인식 라이브러리. HTML 및 자바스크립트 기반.

PyTorch*: PyTorch를 사용하여 얼굴을 분석하기 위한 Python 라이브러리. 주로 심층 신경망을 사용하여 얼굴을 감지하고 얼굴 특징을 분석하기 위해 개발함.

Fer: PyPI 패키지로 심층 신경망으로 표정 인식을 위한 파이썬 라이브러리.

Residual Masking Network: 잔여 마스킹 네트워크를 사용한 표정 인식을 파이토치로 구현.

Conv-emotion: 파이토치를 통해 대화에서 감정 인식을 위한 다양한 아키텍처 구현.

얼굴 감정 인식(HSEmotion): 동영상과 사진에서 얼굴 감정을 인식하기 위한 파이썬 라이브러리.

MMSA: 멀티모달 감정 분석 작업을 위해 개발된 파이썬 기반 프레임워크.

EmoPy: 사람의 얼굴 이미지가 주어지면 감정 표현 분류를 예측하는 심층 신경망 클래스가 포함된 파이썬 툴킷이다.

Open Source library entry	deepface	human	facetorch	fer	Residual Masking Network	conv-emotion	MMSA	EmoPy
Open Issue Count	6	2	2	19	13	0	14	11
Number of pull requests	0	0	1	0	0	0	0	6
Last commit date	23.01.31	23.01.30	23. 01. 29	22.12.13	22.12.19	22.07.29	22.11.08	21.01.15
Number of Contributors	30	8	1	6	1	8	7	15
Used count	1028	117	-	252	29	-	-	24
Number of stars	5413	1121	109	245	331	1059	302	833
Number of StackOverflow questions	232	5	0	202	0	5	0	3

*게시자는 사용자에게 신뢰할 수 있는 인공지능 윤리 가이드라인을 확인할 것을 경고한다. 모델은 완벽하지 않고 편향성이 있을 수 있으므로 편향성에 따른 결과를 확인하지 않은 상태로 이 라이브러리를 사용하는 것은 권하지 않는다.

Verbal

딥스피치: 텐서플로를 사용하는 오픈소스 엔진으로, 임베디드(오프라인, 온디바이스) 음성-텍스트 변환 엔진

Annyang: 웹 애플리케이션용으로 개발됨. HTML, JavaScript를 지원함. 자바스크립트 음성 인식 라이브러리

SpeechRecognition: 파이썬용 모듈로 온라인과 오프라인에서 여러 엔진과 API를 지원

Whisper.cpp: C/C++용으로 개발되었으며, OpenAI의 Whisper 모델을 포팅한 것

Speech-to-Text-WaveNet: 텐서플로 구현이다. 딥마인드의 웨이브넷을 기반으로 하는 엔드투엔드 문장 수준의 영어 음성 인식에 사용된다. 이를 사용하기로 했다면, 영어 말하기 능력에 기반한 모델이라는 점에 유의해야 한다.

Open Source library entry	DeepSpeech	annyang	Speech Recognition	whisper.cpp	Speech-to-Text-WaveNet
Open Issue Count	106	45	239	68	78
Number of pull requests	18	3	26	15	4
Last commit date	21.11.18	22.04.14	23.01.13	23.01.29	21.10.08
Number of Contributors	136	36	46	47	4
Used count	887	645	-	-	-
Number of stars	20,974	6,388	6,730	5,684	3,738
Number of StackOverflow questions	235	153	31	2	1

08-2 오픈소스 라이브러리의 위험 요소는 관리되고 있는가?

Yes No N/A
☐ ☐ ☐

- 면접 평가 프로젝트에서는 시각, 음성, 언어, 바이탈 등 4가지 영역에서의 특징을 분석해야하는데, 오픈소스 라이브러리를 사용하면 많은 시간적, 기술적 이점이 있다. 그러나 오픈소스 소프트웨어 사용 시에는 저작권자의 라이선스를 준수해야 하며, 라이선스 위반으로 인한 법적 책임을 피하라면 사용 중인 오픈소스 소프트웨어의 라이선스와 사용 조건을 명확히 이해하고 준수해야 한다.
- 오픈소스 라이브러리의 종류 및 버전 선택 시 개발 과정에서 사용된 오픈소스 라이브러리 또는 개발 환경 버전 변경에 따른 호환성을 고려해야 하며, 이때 사용된 오픈소스 라이브러리에서 보안 취약점이 발견될 수 있으므로 이를 확인하여 보안상의 위험 요소의 관리도 필요하다.

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

Yes No N/A
☐ ☐ ☐

- 오픈소스 라이브러리를 활용하여 인공지능 모델을 개발할 경우, 각 오픈소스에서 정한 라이선스 공지를 숙지하고 분석하여 저작권, 특허 등 지식재산권 침해로 인한 민형사상 법적 분쟁 등 향후 발생할 수 있는 리스크를 최소화해야 한다. 특히, 소스 코드 공개 의무는 향후 회사의 영업비밀이 유출될 위험으로 이어질 수 있으므로 유의해야 한다.

대표적 오픈소스 라이선스의 주요 내용

OSI 기준	Apache License 2.0	GPL General Public License 3.0	AGPL Affero GPL 3.0	LGPL Lesser GPL 3.0	MIT License	Artistic License 2.0	Eclipse License	BSD Berkeley Software Distribution License	MPL Mozilla Public License 1.1
복제, 배포, 수정의 권한 허용	○	○	○	○	○	○	○	○	○
배포 시 라이선스 사본 첨부	○	○	○	○	○		○	○	○
저작권 고지사항 또는 Attribution 고지사항 유지	○	○	○	○	○	○	○	○	○
배포 시 소스코드 제공 의무와 범위		전체 코드	네트워크 서비스 포함 전체 코드	2차 저작물		○ (표준 버전)	모듈 단위		파일 단위
조합저작물 작성 및 타 라이선스 배포 허용	○			○	조건부	○	○	조건부	○
수정 내용 고지		○	○	○		○	○		○
명시적 특허 라이선스의 허용	○	○	○	○		○	○		○
라이선시가 특허 소송 제기 시 라이선스 종료	○	○	○	○		○	○		○
이름, 상표, 상호 사용 제한	○		○			○		○	
보증의 부인	○	○	○	○	○	○	○	○	○
책임의 제한	○	○	○	○	○	○	○	○	○

- 오픈소스 이니셔티브(OSI[open source initiative])는 오픈소스 라이선스가 사용자에게 소프트웨어를 자유롭게 사용, 수정, 배포할 자유를 제공하는 동시에 원저작자와 코드의 무결성을 보호할 수 있도록 하는 표준을 개발하였다. OSI는 이러한 표준을 충족하는 것으로 인증된 라이선스 목록을 관리한다. 다음은 OSI 에서 명시한 오픈소스 라이선스 요건이다[183].
  - ✓ 자유로운 재배포(Free redistribution)
  - ✓ 소스코드 공개(Source code open)
  - ✓ 2차 저작물 허용(Derived works)
  - ✓ 저작자의 소스 코드 원형 유지(Integrity of the author's source code)
  - ✓ 개인이나 단체에 대한 차별 금지(No discrimination against persons or groups)
  - ✓ 사용 분야에 대한 차별 금지(No discrimination against fields of endeavor)
  - ✓ 라이선스 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 호환성과 보안 취약점을 확인하였는가?

Yes ☐ No ☐ N/A ☐

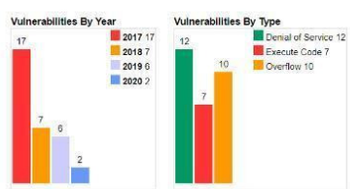
- 라이브러리의 버전 변경 과정에서 개발 환경, 언어, 도구 및 다른 라이브러리 버전과 호환되지 않는 호환성 문제가 발생할 수 있다. 따라서 오픈소스 라이브러리 종류 및 버전 선택 시 라이브러리 간 의존성[dependency]을 파악하는 등 호환성을 고려해야 한다. 다음은 호환성이 좋은 라이브러리의 몇 가지 예시이다.
  - ✓ TensorFlow(오픈소스 기계 학습) – Keras(상위 API)
  - ✓ Flask(웹 응용 프레임워크) – SQLAlchemy(Python용 SQL 툴킷 및 객체-관계 매핑)
  - ✓ NumPy(Python 숫자 계산) – SciPy(과학적 계산 추가기능)
  - ✓ Pandas(데이터 조작) – Matplotlib(시각화)
  - ✓ OpenCV(컴퓨터 비전) – NumPy(Python의 숫자 계산)
- 사용 중인 오픈소스 라이브러리에서 보안 취약점이 발견되기도 한다. 보안 취약점에 따른 영향을 최소화하기 위해 보안 취약점 및 버전 변경에 따른 릴리즈 노트[release note]를 지속해서 확인하여 신속히 탐지 및 대응해야 한다.
- Black Duck, Snyk, OWASP[Open Web Application Security Project], CVE[Common Vulnerabilities and Exposures] 및 NVD[National Vulnerability Database] 등의 취약성 스캐너 또는 보안 도구를 사용하면 라이브러리에 알려진 보안 취약점이 있는지 확인하여 이를 최소화 할 수 있다.

참고 오픈소스 라이브러리의 보안 취약성 분석 예시

라이브러리	CVE Common Vulnerability and Exposures February 2023
Opencv	2020년에 2건의 보안 위협이 발견되었다(2017년 17건, 2020년 2건으로 보안 취약성 발견 합계가 감소한 것으로 미루어 제조업체의 보안 위협 대응이 어느 정도 이루어졌음을 알 수 있다). DoS서비스 거부: 공격의 취약한 부분이 분석되었으며(37.5%), 오버플로 위협도 분석되었다(31.3%).
NVIDIA NeMo	보안 취약점 분석 결과, 2022년에 하나의 보안 위협이 발견되었다. 디렉터리 트래버설 보안 위협: NVIDIA NeMo 1.6.0 이전 버전 ASR WebApp에서 취약점이 발견되었고, 관리자 권한이 있는 경우/경로 트래버설은 임의의 디렉터리 삭제로 이어질 수 있다.

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

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



Opencv 오픈소스 라이브러리의 2017~2020 CVE 보안 취약성 분석 결과

Year	# of Vulnerabilities	DoS	Code Execution	Overflow	Memory Corruption	Sql Injection	XSS	Directory Traversal	Http Response Splitting	Bypass something	Gain Information	Gain Privileges	CSRF	File Inclusion	# of exploits
2022	1							1							
Total	1							1							
% Of All		0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	

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

2023년 NeMo 오픈소스 라이브러리 CVE 보안 취약점 분석 결과

다양성 존중

요구사항 09 인공지능 모델의 편향 제거

- 연구에 따르면 AI 모델이 편향되어 해로운 결과를 초래할 수 있는 세 가지 주요 이유가 있다[184].
  - ✓ 비즈니스/컨텍스트의 목적이나 특성이 편향되어 있어 의도와 상관없이 AI 모델의 결과가 편향될 수밖에 없는 경우
  - ✓ 인공지능 모델을 사용할 주 사용자를 제대로 파악하지 못하는 경우, AI 모델의 학습 데이터 자체에 잠재적 사용자의 대표 데이터가 포함되지 않거나 특정 집단에 대한 편향이 있음
  - ✓ 보호변수 또는 그 프록시가 모델의 추론 결과의 핵심 요소인 경우
- 개발하고자 하는 인공지능 모델이 첫 번째 이유에 해당하지 않는 경우, 위에 요약된 두 번째 또는 세 번째 이유에 따라 발생할 수 있는 편향성을 제거하기 위한 기법을 고려해야 한다.
  - ✓ 요구사항 07-2에 언급된 바와 같이 인종차별, 방언 차이, 학력, 성차별 등 사회적·윤리적 이슈가 있는 경우에만 해당된다[185].

09-1 모델 편향을 제거하는 기법을 적용하였는가?

Yes ☐ No ☐ N/A ☐

- 편향성은 모델 예측의 부정확성을 초래하고, 특정 그룹이나 개인에 대한 차별을 초래할 수 있다. 이는 시스템과 조직의 평판에 해를 끼치거나 법적 처벌을 초래하기도 한다. 채용 분야에서는 그 심각성이 배가 된다.
- 인공지능 시스템에서 모델 편향을 제거하는 기술은 학습 데이터 선택부터 모델 배포 및 유지 관리에 이르기까지 개발 프로세스 전반에 걸쳐 적용해야 한다. IBM의 초기 연구에 따르면 편향성 완화 기법은 모델학습 기준 적용 시점에 따라 전처리, 처리 중, 처리 후 기법으로 분류하여 적용할 수 있다.
- 여기에는 모델의 성능을 정기적으로 모니터링하는 것은 물론, 학습 데이터를 정기적으로 재평가하고 필요에 따라 모델을 조정하여 편향성을 줄이거나 제거하는 것까지 포함된다. 또한, 알고리즘 공정성 및 설명 가능성과 같은 기술을 적용하여 모델에서 잠재적인 편향의 원인을 식별하고 해결할 수 있다.

09-1a 개발하려는 모델에 맞게 편향제거 기법을 선택하였는가?

Yes ☐ No ☐ N/A ☐

- 채용 인공지능 모델을 개발할 때는 몇 가지 편향이 발생할 수 있다[187]. 적절한 편향 제거 기술을 선택하는 것은 데이터에 존재하는 편향 유형, 해결할 문제의 특성 및 사용하는 모델 같은 여러 요인에 따라 달라진다. 데이터 전처리, 데이터 증강, 재샘플링 및 모델 훈련 과정에서 공정성 제약 조건의 사용과 같은 일반적인 편향 제거 기술이 있다. 이러한 기술들이 구체적인 데이터와 문제에 얼마나 효과적인지를 평가하고, 사용하면서 발생하는 새로운 편향 소스를 계속 모니터링할 필요가 있다.

인공지능 모델의 편향을 완화하기 위한 기법 예시

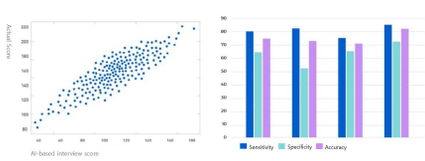
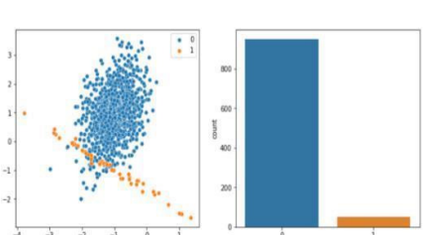
편향 유형	기술 또는 접근 방법	pre	in	post	설명
알고리즘 편향 algorithmic bias	가중치 재지정	✓			학습 데이터셋 샘플에 가중치를 할당하는 방식
리콜 편향 recall bias	라벨링 재지정	✓			학습용 데이터 샘플의 라벨을 수정하는 방식
특성 편향 feature bias	변수 블라인딩	✓			분류기가 민감한 변수에 반응하지 않도록 하는 방식
–	변형	✓		✓	숫자 데이터 기반 학습 시 데이터 변환 및 모델 예측 분포를 변환하는 방식
데이터 표본 편향 data sampling bias	샘플링	✓			학습 데이터 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 ☐

- 개발하려는 모델과 임무 목표에 맞게 지표를 선정하고, 편향 완화 여부를 지속해 측정 및 관리하기 위해서는 편향 정도를 정량적으로 확인하는 것이 좋다. 편향성을 평가하는 지표로는 다음과 같은 것들이 있다.

편향을 양적으로 측정하는 지표의 분류

분류	지표
상관관계 기반 지표	DACOBS[Development of the Davos assessment of cognitive biases scale], Ddavos 인지 편향 척도 평가, 크론바흐의 알파 테스트(또는 계수 알파), 피어슨 상관 계수
	평가 결과의 품질을 평가하기 위해 가장 선호되는 지표 중 하나는 결과와 참조 값들의 상관계수 테스트를 기반으로 한다. 예를 들어, 해당 논문(가장 잘 알려진 MIT 면접 세트의 제작자 연구 중 하나인 오픈 액세스 데이터셋)에서 크리펜도르프 알파를 얻어낸 후, 각 평가자의 상관관계를 추정하여 출력물의 편향/비편향을 테스트하였다(특성별). [191].
혼동 행렬 confusion matrix 기반 지표	동등 기회[equalized opportunity], Equalized Odds, 전체 정확도 형평성, 조건부 사용 정확도 형평성, 대응 형평성, 비보상 동등화
	 모델, 검사 도구 및 알고리즘의 분류, 식별 및 예측 능력, 오류 행렬로도 알려진 편향을 평가하기 위한 것이다. 혼동 행렬에서 정확도, 정밀도, 재현율, 특이도 및 F1-스코어를 도출할 수 있다. [192][193]
점수[score] 기반 지표	양성 및 음성 클래스 균형 지표
	 왼쪽 그래프의 경우(950개의 다중 클래스 데이터, 50개의 소수 클래스 데이터), 클래스 수가 많아 다중 클래스 예측 시 정확도가 높다. 그러나 소수 클래스를 예측하는 정확도는 낮다. 따라서 점수 기반 지표인 ROC-AUC와 F1-스코어를 통해 편향 보정을 고려할 수 있다. [193][194][195]
통계 기반 지표	교차 검증(CV[Cross Validation]), 공정성 기반 지표: 인구/인구 공정성 지표, 차이 있는 부정 효과 지표
	교차 검증이나 회전 추정 또는 샘플 밖 테스트라고도 불린다. CV는 통계 분석 방법으로, 결과를 독립 데이터셋에 어떻게 일반화할지 평가하는 여러 유사 모델 유효성 검사 기술의 하나로, 오류 추정에서 편향을 계산하는 데 효과적으로 간주된다. 따라서 잘 중첩된 CV 절차는 거의 편향되지 않은 실제 오류 추정을 제공할 수 있어 분류기의 편향 완화에 사용할 수 있다[193][196].

안전성

요구사항 10 인공지능 모델 공격에 대한 방어 대책 수립

- 면접 영상 분석, 지원자 선별을 위한 인공지능 모델은 적대적 의도를 가진 사용자로 인해 불공정한 추론을 낼 수 있으므로 모델을 대상으로 한 다양한 공격을 방지 또는 완화할 대책을 수립해야 한다.

10-1 모델 공격이 가능한 상황을 파악하였는가?

Yes ☐ No ☐ N/A ☐

- 면접 영상의 적대적으로 조작된 입력과 같이 작은 변화에도 모델을 오동작하게 만드는 공격은 분석결과의 정확성을 위협할 수 있다. 따라서, 적대적 공격을 이해하고 적절한 대응 방안을 마련하여 모델 추론 결과의 신뢰도를 확보해야 한다.
- 적대적 공격의 대표적 유형으로는 회피 공격evasion attack이 있다. 추론 중에 인공지능 모델을 속이기 위해 입력 데이터를 조작하는 것이다. 회피 공격에 대한 대응을 위해 채용 AI 모델에 사용하는 주 데이터 유형별 공격 가능한 적대적 사례를 파악해야 한다.

10-1a 데이터 유형별 공격 가능한 적대적 사례를 확인하였는가?

Yes ☐ No ☐ N/A ☐

- 면접 영상 분석 모델은 지원자의 얼굴 표정(영상), 음성, 텍스트 정보를, 지원자 선별 모델은 텍스트 입력을 주로 사용한다.
- 영상 분야는 적대적 공격에 관한 연구가 가장 활발히 이루어지고 있는 분야로 입력 이미지 공격이 주를 이룬다. 이미지는 텍스트나 오디오에 비해 픽셀값의 고차원 배열로 표현되는 복잡성으로 인해 적대적 사례를 생성하기가 비교적 쉽다.
- 아울러 면접 내용이나 직무 요구사항에 대한 추론은 주로 음성을 전환한 텍스트, 또는 입력 텍스트 그 자체를 입력으로 활용하기 때문에 사용자의 의도적인 입력을 통해 모델의 추론을 방해할 수도 있다.

참고 채용 인공지능 시스템에 대한 회피 공격(사용자 입력) 예시

키워드 채우기	이력서나 면접 질문에 대한 응답에 여러 관련 키워드를 포함시킨다. 이러한 키워드는 반드시 지원자의 실제 자격을 반영하는 것은 아니지만, AI 시스템의 긍정적 분류를 유발하여 지원자를 실제보다 적합해 보이게 한다.
맥락적 왜곡	면접 질문에 대한 응답에 미묘한 변경이나 왜곡을 주입할 수 있다. 이러한 변경 사항은 AI 모델의 의사결정 과정을 조작하기 위해 신중하게 설계된 것이다. 특정 특성이나 기술을 선호하도록 만들어 시스템의 평가에 영향을 미치려 할 수 있다.
언어 습관 모방	고용주가 원하는 품질에 더욱 가까워 보이기 위해 적들은 성공한 지원자들과 자주 관련된 특정한 언어 스타일 및 용어를 모방하여 결과에 영향을 주려 한다.
합성 데이터 생성	일부 공격자들은 알고리즘을 활용하여 직무 요구사항과 일치하는 합성 데이터를 생성할 수 있다. 이러한 합성 프로필이나 응답은 실제 지원자와 일치하지 않을 수 있지만, 채용 인공지능 시스템을 속여 합법적인 지원자로 간주하도록 설계되었다.

10-2 모델 공격에 대한 방어 수단을 강구하였는가?

Yes ☐ No ☐ N/A ☐

- 10-1을 통해 현재 개발 중인 모델의 공격 가능한 사례를 파악하였다면, 모델 최적화model optimization를 통해 이를 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 위시한 방어 수단을 강구하여 인공지능 모델 개발 단계에서 대처 가능한 방안을 검토 및 적용하는 것이 바람직하다.

10-2a 모델 최적화를 통한 방어 대책을 마련하였는가?

Yes ☐ No ☐ N/A ☐

- 채용 인공지능 시스템에 회피 공격을 하는 공격자는 자격에 대한 시스템의 평가를 회피하여 채용 직책에 선발될 가능성을 높이려 할 수 있다. 주요 완화 방법은 적대적 학습, 적대적 공격 데이터의 역활용, 적대적 공격 여부를 판단하는 모델 추가, 자석 방법, 쿼리 찾기 차단, 네트워크 수정, 방어적 증류 그리고 GANGenerative Adversarial Networks 방어 등이 있다.

모델 공격에 대한 모델 개발 및 시스템 구현 단계의 방어 기술

방어 기술의 분류	방어 기술 내용
적대적 학습adversarial training (무차별 적대적 훈련)	모델을 훈련할 때 적대적인 사례를 모방하는 적대적인 샘플 학습 데이터셋을 학습 데이터셋에 포함시키는 것으로, 제로 지식 공격 유형에 유용하다. 그러나 적대적 학습은 적대적인 샘플 학습 데이터셋의 양과 다양성이 충분하지 않으면(즉, 모든 적대적인 사례의 수를 고려하지 않으면) 방어 기술로서의 성능이 떨어진다. 적대적 학습 방법의 종류에는 FGSMFast Gradient Sign Method 적대적 학습, PGDProjected Gradient Descent적대적 학습, ALPAdversarial Logits Pairing등이 있다. 또한, 이미지 데이터셋의 경우 L-BFGSLimited-memory BFGS와 FGSM에 의해 생성된 교란을 방어하기 위해 시야 집중 메커니즘을 사용할 수 있지만, 이 방법이 공격의 강도를 줄일 수 있음을 고려해야 한다.
적대적 공격을 판단하는 모델 추가(시스템 개발 단계)	같은 두 모델의 추론 결과를 비교하여 두 결과 간에 차이가 발생하면 적대적 공격으로 판단하는 적대적 공격 탐지 기술이다. 특정 모델에 적용된 적대적 공격을 불가능하게 만드는 여러 학습 모델을 결합하기도 한다.
자석 메커니즘[207]	정상 데이터의 다양성을 근사하여 정상 및 적대적 예제를 구별한다. 매니폴드 근처로 이동하여 적대적 예제를 재구성하는데, 작은 적대적 사례를 올바르게 분류하는 데 효과적이다.
블록 쿼리 조회	반복 쿼리를 시도하는 역전 공격이나 모델 추출 공격을 방지하기 위해 모델의 수를 제한하는 방법이다.
네트워크 수정	자동 인코더에 노이즈 제거를 쌓아 올리면서 야기되는 취약성을 제거하는 것이 목표이다. 공격에 대한 견고성을 높이기 위해 기울기 규제를 사용하는 해결책을 찾았다.
방어 증류[208]	DNN의 구현이다. 제안된 방법은 DNN 모델의 훈련 단계에 적용된다. 제안된 증류 온도 방법은 훈련 모델의 SoftMax 계층에서 사용된다.
GAN 방어	10-1a와 마찬가지로 GAN은 회피 공격에 대한 방어 메커니즘으로도 사용된다[198]. GAN 모델의 기본 아이디어는 생성기를 사용하여 새로운 이미지를 생성할 때 판별기에 걸리지 않는 것이다. 개발된 APE-GAN 모델도 유사한 이점이 있다. 이 방법에서는 적대적 이미지를 생성기에 입력으로 공급한다. GAN 모델을 사용하여 이 공격받은 이미지를 공격받지 않은 형태로 변환한다[209].
앙상블Ensemble	앙상블은 여러 모델의 예측을 결합하여 정확도를 높이고 모델 회피 공격의 위험을 줄이는 방법이다. 여러 모델을 사용하면 공격자는 목표를 달성하기 위해 모든 모델을 침해해야 한다. 기법의 종류에는 Bagging(부트스트랩bootstrap을 병렬로 집계함), Boosting(여러 약한 모델을 순차적으로 훈련하며, 각 후속 모델은 이전 모델의 오류를 수정하려고 함), Stacking(여러 모델의 출력을 입력 기능으로 사용하여 그 예측을 결합하는 상위 수준 모델에 학습시킴), Random forest(입력 기능의 무작위 하위 집합마다 훈련된 여러 결정 트리를 결합하여 과적합을 줄이고 일반화를 강화함) 등이 있다.
차원 축소Dimensionality Reduction	특성 축소는 모델에서 고도로 상관되거나 정보를 제공하지 않는 특성을 제거하는 데 사용하는데, 공격자가 모델의 취약점을 이용하기가 더 어려워질 수 있다. 차원 축소 방법은 크게 선형과 비선형으로 나뉘며 대표적으로 주성분 분석(PCAPrincipal Component Analysis)과 LLE Locally-Linear Embedding이 있다

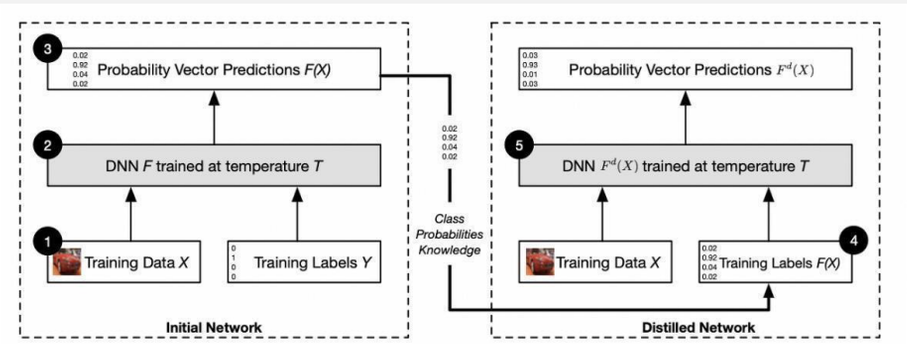
참고 CleverHans: 적대적 공격 방어 – 딥러닝 라이브러리[205]:

CleverHans는 회피 공격을 포함한 적의 공격에 대한 다양한 방어 기법을 구현하는 툴킷이다. 이를 통해 연구자와 실무자는 다양한 방어 전략을 실험하고 테스트할 수 있다.

CleverHans는 적대적 훈련, 입력 변환, 그라데이션 마스킹과 같은 다양한 기법을 제공하고, 적대적 공격에 대한 방어에 대한 요약도 제시한다. 또한, 연구자와 실무자는 CleverHans를 사용하여 회피 공격을 포함한 적대적 공격으로부터 AI 모델을 방어하고 다양한 방어 방법의 효과를 평가할 수 있다[206].

참고 데이터 공격에 대한 모델의 방어 기법 예시

방어 증류 방어 기술[151]



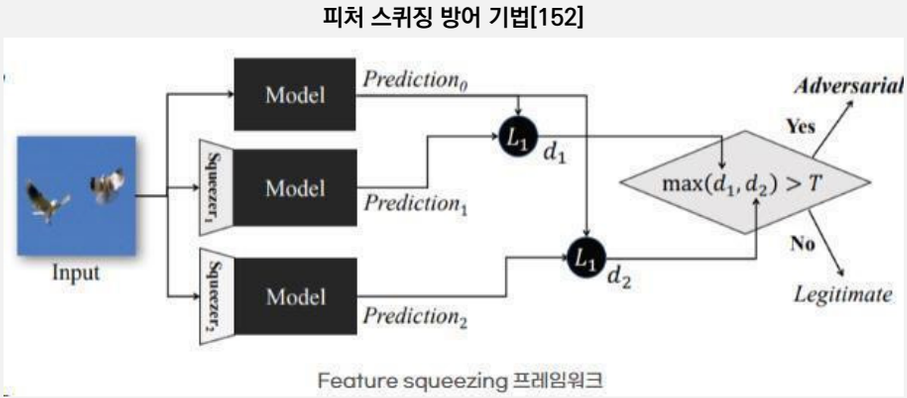
현재 C&W 공격, PGD 등 다양한 공격 방식이 등장하면서 이 기법은 많이 깨지고 무뎌졌지만, 처음 나왔을 때는 많은 사람의 관심을 끌며 차세대 방어 기법으로 각광받았다.

지식 증류 방식을 기반으로 한 교사 네트워크 T와 학생 네트워크 S의 두 가지 모델이 있으며, 먼저 학습한 T의 지식으로 학생을 학습시켜 추가 지침을 제공한다.

발표 당시 대부분의 공격은 그라디언트 형태였다. 모델의 기울기가 가파르면 작은 노이즈도 네트워크 출력에 큰 차이를 만들 수 있으므로 적대적인 샘플을 생성하기 쉽다.

따라서 이 방법은 모델의 기울기를 완만하게 만들어 공격이 작동하기 어렵게 한다.

피처 스퀴징 방어 기법[152]



Feature squeezing 프레임워크

적의 공격에 흔들리지 않도록 모델을 강화하는 방법, 데이터를 학습할 때 표현의 복잡성을 줄여 감도를 낮춰 모델을 견고하게 만드는 방법 등이 있다. 또한, 픽셀의 색상값을 작게 인코딩하여 색 심도를 낮추는 방법과 이미지에 평활화 필터를 적용하는 방법도 있다.

책임성 투명성

요구사항 11 인공지능 모델 명세 및 추론 결과에 대한 설명 제공

- AI 모델의 추론 결과만으로는 모델이 어떻게 예측 결과를 얻었는지 알기 어렵다. 특히 채용 인공지능 시스템의 판단은 사용자의 삶에 직접적인 영향을 미칠 수 있어 추론 과정을 설명하는 문제가 더욱 중요하다.
- 또한, 시스템의 최종 결과를 얻기 위해 다수의 인공지능 모델이 사용될 수 있다. 이 과정에서 인공지능 모델 추론 결과를 지원자의 면접 평가에 활용한다면, 사용자의 이해를 돕기 위해 사용된 모델 정보와 추론 결과의 구체적인 인과관계를 보여줄 수 있는 설명(모델에 대한 유용한 정보)을 제공해야 한다.

참고 설명 가능성explainability 적용 전 고려해야 할 사항

1. **제품 및 서비스의 다양성 고려:** 모든 인공지능 모델과 제품 및 서비스에 설명 가능성이 필요한 것은 아니다. 사용자가 제품 및 서비스를 이용하면서 시스템 동작 및 모델의 추론 결과에 관해 설명을 요구하는 분야가 있지만, 그렇지 않은 분야도 있다. 이와 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. **설명 가능성이 미치는 영향 고려:** 설명 가능성은 아직 기술적으로 연구 및 개발이 활발하게 이루어지는 분야로, 여전히 기술적 한계가 있다. 동시에 설명 가능성 외의 다른 속성과도 상호 연관성이 있어 신중히 접근해야 한다. 일례로, 설명 가능성을 과도하게 구현하면 모델 성능 및 프라이버시 등에 부정적인 영향을 초래한다는 의견도 있다. 따라서 본 요구사항은 제품의 개발 의도와 설명이 적용되는 상황 및 영향을 파악하여 적절한 수준의 설명을 마련해야 한다. 따라서 이 요건은 면접의 성격과 모델 개발 의도 및 설명이 필요한 정보에 설명이 미치는 영향에 따라 제기된 상황과 결과물을 식별한다.

인공지능 해석 가능성 기법의 일반 지도[210]

11-1 인공지능 모델의 명세를 투명하게 제공하는가?

Yes No N/A
☐ ☐ ☐

- 인공지능 모델이나 서비스의 개발, 테스트, 배포 과정에서 발생한 다양한 결과를 문서화하는 것은 인공지능 시스템의 투명성을 보장하는 합법적인 방법의 하나이다.
- 특히 채용 평가 프로세스에서 이러한 문서화는 사람들에게 선발 배경에 대한 근거를 제공하고 시스템의 신뢰도를 높인다. 사용자가 인공지능 모델과 관련된 정보를 요청할 때 모델의 목적, 입출력 정보, 성능, 편향성, 신뢰성 등의 결과를 투명하게 공개할 수 있도록 모델을 상세하게 설명하는 문서를 작성한다.
- 인공지능 모델의 주요 정보와 구성 요소를 상세히 기록하는 것은 채용 인공지능 시스템의 투명성뿐만 아니라 잠재적 오류나 지원자의 이의제기 발생 시 추적성 측면에서도 중요하다.

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

Yes No N/A
☐ ☐ ☐

- 시스템 개발에 대한 세부 정보를 제공하는 것은 시스템의 투명성을 보장하는 데 매우 중요하다[221]. 이 문서는 향후 시스템의 유지 관리, 개선 및 배포의 참고 자료가 될 수 있으며, 시스템 내부 작동을 이해해야 하는 이해관계자 및 규제 기관에도 유용한 정보를 제공할 수 있다. 따라서 이 문서는 명확하고 간결하면서도 체계적으로 작성되어야 하며, 기술 전문가가 아닌 일반인도 쉽게 이해할 수 있어야 한다.
- IBM과 WEF는 모델 사양의 문서화를 통해 AI 시스템의 투명성을 보장하는 방법을 제안한다. 사양을 작성할 때는 AI 모델의 메커니즘 측면을 포함하여 추가 정보를 명시해야 한다. 예로는 면접 채점에 영향을 미치는 가중치(특징), 주요 가정, 스펙 해석 시 유의사항 등이 있다.
- 모델 상세 문서 작성 시에는 AI 생명주기에 관련된 이해관계자를 고려하여 각자 선별 및 검증할 수 있도록 관련 정보를 포함해야 한다. 다음은 이해관계자에 따라 모델 상세 문서에 필요한 정보의 예시이다.

이해관계자에 따른 모델 상세 문서 예시

이해관계자	모델 세부사항
비즈니스 의사결정권자	시스템 내 서비스의 목적, 방향, 서비스 이름, 각 서비스의 의도된 목적 등
데이터 과학자 및 시스템 개발자	학습에 사용되는 데이터셋 사양 및 전처리 기법, 학습 모델 구성, 입력/출력 사양, 모델 학습 파라미터 등
모델 검증자	테스트 데이터셋 구성 정보 및 주요 테스트 성능, 편향성, 신뢰성 등의 평가 결과
모델 운영자	모델 운영 및 모니터링 결과, 성능 저하 환경 요인, 최적 결과 도출을 위한 환경 등의 성능 평가 지표

참고 Amazon 리더십 원칙과 아마존의 인재 육성 과학 사례[224]

2018년 Amazon은 채용 프로세스와 머신러닝을 사용하여 입사 지원자를 평가하는 방법을 설명하는 "Amazon 리더십 원칙과 아마존의 인재 육성 과학"이라는 제목의 문서를 발표했다. 이 문서에는 아마존의 모델이 공정하고 편향이 없는지 확인하기 위해 취하는 단계가 자세히 설명되어 있다. 또한, 일련의 리더십 원칙을 사용하여 채용 결정을 내리는 방법과 과거 채용 데이터를 사용하여 머신러닝 모델을 훈련하여 고성과 직원과 연관된 패턴을 식별하는 방법을 설명한다. 편향적이거나 차별적인 결과를 도출하지 않도록 모델을 정기적을 감사하는 방법도 담겨 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 한다. 또한, 이에 대한 설명 및 근거를 사용자에게 제시하는 것이 바람직하다.
- 채용 인공지능 시스템이 채용 결정에 미치는 영향은 증가하는 추세이다. 그러나 시스템이 제안하는 논리를 사용자가 제대로 이해하지 못하거나, 시스템이 결과를 낸 이유를 적절히 설명하지 못해 마찰을 빚는 경우도 종종 있다.
- 이를 방지하기 위해 사람이 이해할 수 있는 방식으로 모델 판단의 근거를 제시할 수 있는 설명 가능한 인공지능XAI, eXplainable AI 기술의 검토와 적용을 고려해야 한다. 또한, 설명이 필요한 요소 및 인공지능 모델 특성에 따라 대리surrogate 모델, 집중attention, 내부internal 분석 방식 등 XAI 기술을 도입할 수 있다.
- 또한, AI 모델의 추론 결과의 근거를 항상 설명할 수 있는 것은 아니므로, XAI 기술 적용 이외의 대안을 활용하여 AI 시스템의 투명성을 확보해야 할 수도 있다. 따라서 XAI 기술 적용 가능 여부를 고려하여 본 세부 요건의 검증항목을 선택적으로 적용할 수 있다.

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

Yes ☐ No ☐ N/A ☐

- 텍스트 또는 시각화 등 다양한 접근 방식을 통해 XAI 기술을 활용하면 채용 인공지능 시스템의 투명성을 확보할 수 있다. XAI 기술 도입을 고려할 때, 채용 인공지능 시스템에서 주로 사용하는 세 가지 주요 데이터 유형인 (면접 영상에서 얻은)텍스트, 오디오 및 이미지/시계열(면접 영상에서 얻은 비디오 시퀀스) 데이터에 따라 도입 가능한 기술을 살펴볼 필요가 있다.

참고 최근 개발된 AI 시스템을 설명하기 위한 XAI 기술 사용[210]

Table 1. Interpretability Methods to Explain Deep Learning Models.

Ref	Tool	Category	Local vs. Global	Model Specific vs. Model Agnostic	Data Type	Citations/Year	Year
[32]	DeepExplain iNNvestigate tf-explain	PH	L	Specific	img	1548.3	2014
[35]	Grad-CAM tf-explain	PH	L	Specific	img	797.8	2017
[34]	CAM	PH	L	Specific	img	607.8	2016
[31]	iNNvestigate	PH	L	Specific	img	365.3	2014
[23]	DeepExplain iNNvestigate tf-explain	PH	L	Specific	img	278.3	2013
[27]	DeepExplain iNNvestigate Integrated Gradients tf-explain alibi Skater	PH	L	Specific	img txt tab	247	2017
[40]	Deep Visualization Toolbox	PH	L	Specific	img	221.7	2015
[37]	DeepExplain iNNvestigate The LRP Toolbox Skater	PH	L	Specific	img txt	217.8	2015
[29]	DeepExplain DeepLift iNNvestigate tf-explain Skater	PH	L	Specific	img	211.5	2017
[41]	iNNvestigate	PH	L	Specific	img	131.5	2017
[38]	iNNvestigate tf-explain	PH	L	Specific	img	113.3	2017
[42]	tcav	PH	L	Specific	img	95	2018
[43]	rationale	PH	L	Specific	txt	81.4	2016
[36]	Grad-CAM++	PH	L	Specific	img	81	2018
[39]	RISE	PH	L	Specific	img	43.3	2018
[44]	iNNvestigate	PH	L	Specific	img	41.8	2017

Table 2. Interpretability Methods to Explain any Black-Box Model.

Ref	Tool	Category	Local vs. Global	Model Specific vs. Model Agnostic	Data Type	Citations/Year	Year
[45]	lime Eli5 InterpretML AIX360 Skater	PH	L	Agnostic	img txt tab	845.6	2016
[59]	PDPbox InterpretML Skater	PH	G	Agnostic	tab	589.2	2001
[48]	shap alibi AIX360 InterpretML	PH	L & G	Agnostic	img txt tab	504.5	2017
[50]	alibi Anchor	PH	L	Agnostic	img txt tab	158.3	2018
[53]	alibi	PH	L	Agnostic	tab img	124.5	2017
[60]	PyCEbox	PH	L & G	Agnostic	tab	53.3	2015
[58]	L2X	PH	L	Agnostic	img txt tab	50.3	2018
[57]	Eli5	PH	G	Agnostic	tab	41.5	2010
[51]	alibi AIX360	PH	L	Agnostic	tab img	34.3	2018
[61]	Alibi	PH	G	Agnostic	tab	23.2	2016
[54]	alibi	PH	L	Agnostic	tab img	17	2019
[62]	pyBreakDown	PH	L	Agnostic	tab	8.3	2018
[62]	pyBreakDown	PH	G	Agnostic	tab	8.3	2018
[47]	DLIME	PH	L	Agnostic	img txt tab	7.5	2019
[56]	AIX360	PH	L	Agnostic	tab	7	2019
[52]	AIX360	PH	L	Agnostic	tab img	3	2019

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

Yes No N/A
☐ ☐ ☐

- AI 모델 추론과 의사결정에 대한 포괄적인 설명을 제공하는 것은 어려운 작업이 될 수 있다. 특히 AI 기반 채용 평가의 맥락에서 개인정보 보호 및 안전과 관련한 주요 사안을 고려하면 더욱 그러하다[219]. 투명성은 면접관, HR 전문가, 지원자 등 시스템 사용자를 포함한 모든 이해관계자의 요구를 충족시키면서 채용 인공지능 시스템에 신뢰를 심어주는 데 중추적인 역할을 한다.
- 그러나 특정 결과를 미묘하게 암시하는 등 지원자와 관련된 개인정보를 실수로 공개하거나[220], 기본 논리에 대한 통찰력 없이 지원자 선택이나 민감한 속성 평가 관련해 복잡한 설명만 제공하게 되면[211] 특정 위험을 수반할 수 있다.
- 개발자는 AI 시스템을 통해 개인정보 보호와 신뢰 구축을 목표로 하면서도 복잡한 문제를 해결해야 한다. 예를 들어, 채용 전략이나 기술, 자사의 인사 노하우, 지원자의 개인정보 등을 드러내지 않으면서 AI의 의사결정을 가시화하고, 지원자와 면접관에게 의미 있는 피드백을 제공해야 한다. 따라서 투명성과 개인정보 보호 사이의 균형을 유지하는 것이 매우 중요하다. XAI 기술 적용이 불가하다면 다른 접근 방식으로 프라이버시를 보장하면서도 가치 있는 인사이트를 제공해야 한다.

참고 AI 모델의 동작을 이해하기 위한 접근 방식 예시

접근 방법	내용
모델 단순화	모델의 복잡성을 줄이거나 본래의 모델 대신 이해하고 해석하기 쉬운 더 간단한 모델을 사용
특징 중요도 분석	특징 중요도 점수 또는 부분 의존도 플롯과 같은 방법을 사용하여 입력 데이터의 각 특징이 모델 예측에 미치는 상대적 중요성을 파악
모델 검사	모델 매개변수의 가중치 또는 뉴런의 활성화 같은 모델의 내부 작동을 검사하여 모델이 입력 데이터를 처리하고 예측하는 방식을 이해
데이터 시각화	입력 데이터, 모델의 예측, 예측과 실제 출력 사이의 잔차를 시각화하여 데이터의 패턴과 모델의 동작을 식별
모델 기반 추론	모델 유형 및 하이퍼파라미터와 같은 모델 구조에 대한 지식을 사용하여 모델의 동작을 이해
인간 전문가 검토	인간 전문가가 모델과 입력 데이터의 예측을 검토하고 도메인 지식을 바탕으로 설명을 제공

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

Yes ☐ No ☐ N/A ☐

- 사용자에게 인공지능 모델의 추론 결과에 대한 설명을 제공하면, 사용자는 단순히 해당 인공지능 모델의 최종 결과뿐 아니라 그 결과가 도출된 근거 수치로 확률값, 불확실성uncertainty , 신뢰도confidence score 등을 제공받을 수 있다. 이러한 정보는 사용자의 의사결정에 도움이 되지만, 오히려 사용자의 혼란을 유발할 수도 있으므로, 정보 제공의 필요성을 사전에 검토해야 한다.
- 반면 EU의 일반개인정보보호규정(GDPR)처럼 인공지능 모델 추론 결과에 대한 설명이 법적으로 요구되는 경우도 있으므로 다양한 관점의 설명 제공을 검토해야 한다.

11-3a 모델 추론 결과에 대한 설명이 필요한지 검토하였는가?

Yes ☐ No ☐ N/A ☐

- AI 및 자동화된 시스템의 사용 사실을 지원자에게 사전에 알리지 않으면 지원자의 신뢰를 잃을 수 있다[209]. 또한, 알고리즘의 예측 및 의사결정 프로세스는 프로그래머 자신에게도 불투명한 경우가 많다.
- 알고리즘이 지원자 평가에 수백만 개의 데이터 포인트를 사용할 때, 어떤 속성이 결정을 내리는지 질적 설명을 제공하기 어려워진다[227][228]. 채용은 사람들의 삶과 관련성이 높고, 모호한 차별이 발생할 수 있기 때문에 적절한 설명은 윤리적으로 매우 중요하다[229][230]. 따라서 GDPR은 사람들이 자신에 대한 알고리즘의 결정과 관련한 설명을 요청할 수 있는 '설명에 대한 권리'도 보장한다[231].
- 그러나 무조건적 설명의 제공은은 자칫 개인정보보호와 상충하거나 사용자의 혼란을 야기할 수도 있다. 따라서 인공지능 시스템의 활용 분야 및 사용자의 특성에 따라 설명의 필요성을 검토해야 한다. 다음은 모델의 추론 결과에 대한 설명을 제공하지 않는 것이 더 나은 경우의 예이다.
  - ✓ 모델의 추론 결과 자체에 대한 설명을 제공해도 사용자의 의사결정에는 큰 영향을 미치지 않을 것으로 보이는 경우: 알고리즘의 두 가지 도출 결과가 있고, 예측 확률이 각각 84.2%와 86.0%일 때, 사용자는 어떤 결과를 사용하여 의사결정을 내려야 할지 혼란스러울 수 있다.
  - ✓ 예측 확률이 너무 높거나 낮은 경우: 사용자에게 시스템의 출력 결과의 예측 확률값이 100%라고 알려주면, 사용자는 시스템의 출력 결과를 맹신하거나 그 반대의 경우가 발생할 수 있다.
  - ✓ 그 밖에 결과에는 사용자와 공유할 수 없는 민감한 정보 또는 개인 데이터가 포함되거나, 설명이 너무 복잡하거나 기술적으로 난해한 경우 또는 정보 관련해 법적 제한이 있는 경우 등이 있다.

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

Yes No N/A
☐ ☐ ☐

- 인공지능을 채용에 활용하면, 특히 채용 과정에서 많은 지원자가 몰리는 유명 기업이나 기관은 인사 부서에서 채용 프로세스에 드는 시간과 비용을 절감할 수 있다.
- 그러나 결정의 근거가 되는 구체적인 설명 없이 모델 추론만 제시할 경우 사용자(면접관/채용 담당자/기업 등)는 편향적이라는 비난을 받게 될 수 있다. 이 문제를 해결하기 위해 인공지능 모델의 추론 결과에 대한 확률값과 추론 결과의 신뢰도를 나타내는 불확실성을 정량화하여 설명하는 것을 고려할 수 있다. 다음은 추론 결과의 확률과 불확실성에 따른 설명의 예시이다.

참고 AI 모델 추론에 대한 설명 예시

자연어 처리(NLP)를 사용하여 면접 질문에 대한 지원자의 응답을 분석하고 면접관에게 인사이트를 제공하는 채용 인공지능 시스템이 있다. 이 시스템은 언어 능력, 의사소통 능력, 직무별 역량 등 다양한 요소를 기반으로 각 지원자에게 점수를 제공한다. 다양한 기법으로 사용자에게 시스템의 추론 결과를 설명할 수 있다.

기능 중요도 제공: 회사는 지원자의 점수를 결정하는 데 가장 큰 영향을 미친 특징(예: 특정 단어나 구문 또는 말하기 패턴) 정보를 제공, 그래픽 설명이나 중요도 맵과 같은 시각화 수행

예제 사용: 지원자가 특정 점수에 도달한 방법의 예시를 제공, 예) 특정 점수를 받은 지원자가 사용한 특정 문구나 단어를 강조 표시

시각화 사용: 회사는 그래프나 차트 같은 시각화를 통해 모델이 특정 점수에 도달한 방법 제시, 시스템은 언어 능력이나 의사소통 능력과 같은 다양한 요소에 따라 지원자의 점수를 분석하여 나타냄

반면, 직무의 특성이나 특정 법적 요건으로 인해 AI의 결정에 대한 자세한 설명을 제공하기 어렵거나 불필요한 경우도 있다. 예를 들어, 보안이 엄격하거나 기밀이 필요한 정부 직책은 신원 조회 또는 법적 제한에 따라 특정 지원자가 자동으로 실격 처리될 수 있다. 이 경우 AI 시스템은 결격 사유를 공개할 수 없으므로 자세한 설명 없이 지원자가 필요한 기준을 충족하지 못한다는 간결한 알림만을 제공하기도 한다. 설명 제공의 필요성이 직무의 특정 상황과 요구사항에 따라 달라질 수도 있으며, 때로는 AI 결정에 대한 광범위한 정당성을 제공하는 것이 실현 불가능하거나 법적으로 허용되지 않을 수도 있음을 보여주는 예시이다.

추론 확률(0~1)	불확실성(0~1)	예제 설명
0.98	0.01	모델의 추론 확률이 98%에 달하고 추론의 불확실성이 1%로 낮아 모델 추론 결과를 신뢰할 수 있다.
0.98	0.90	모델의 추론 확률은 98%에 달하지만, 추론의 불확실성이 90%에 달해 모델의 추론 결과를 신뢰하기 어렵다.
0.20	0.01	모델의 추론 확률이 20%로 낮고, 추론의 불확실성이 1%로 낮아 모델 추론 결과를 신뢰할 수 있다.
0.20	0.90	모델의 추론 확률은 20%로 낮지만, 추론의 불확실성은 90%에 달해 모델의 추론 결과를 신뢰하기 어렵다.

사용자의 의사결정 요구사항과 모델 예측의 신뢰성에 따라 설명의 세부 수준과 균형을 맞추는 것이 AI 시스템의 효과적인 커뮤니케이션과 신뢰에 매우 중요하다. 아울러 구글의 클라우드 자동 머신러닝 플랫폼의 'AI 설명' 기능[233]또한 참조해 보자.

다양성 존중

요구사항 12 인공지능 시스템 구현 시 발생 가능한 편향 제거

- 채용 인공지능 시스템의 구현 과정에서 인종/민족, 원어민-외국어 능력 등 사용자별 배경이나 편견으로 인해 시스템이 편향될 수 있다. 또한, 모델 학습 시 데이터 편향으로 예고 없이 자동화 편향이 발생하기도 한다(요구사항 09 참조). 따라서 발생 가능한 편향을 파악하고 이를 제거하거나 완화할 방법을 고려해야 한다.

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

Yes ☐ No ☐ N/A ☐

- AI 시스템은 개발 과정에서 개발자의 사전 지식, 예상, 경험 등이 반영되거나 특정 선택을 암묵적으로 유도하는 사용자 인터페이스로 인해 편향이 발생할 수 있다.
- 편향을 방지하기 위해 채용 인공지능 시스템의 구현 단계에서 주기적으로 코드를 검토하여 개발자의 제한된 배경 지식, 편향된 데이터셋을 사용하거나 편향이 코드에 반영되었는지 확인하고, 사용자 인터페이스와 상호작용 측면에서 표현 편향presentation bias이나 순위 편향ranking bias 등이 발생하는지 사전 점검하여 편향이 발생하지 않도록 시스템을 설계하는 것이 바람직하다.

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

Yes ☐ No ☐ N/A ☐

- 채용 인공지능 시스템의 특성상 편향은 주로 데이터셋 구성, 엔지니어 목표 수립, 기능 선택과 관련된 문제에서 비롯된다. 이를 해결하기 위해 다양한 기술적 도구와 외부 감독을 활용한다.
- 시스템은 지원자를 평가하거나 인공지능 모델을 구축하여 적용할 때, 미리 정의된 정보, 인성 평가 경험, 행동/음성/텍스트 분석 등을 반영하여 의사결정을 내리기 때문에 인지 편향이나 확증 편향이 이러한 평가/결정에 영향을 미칠 수 있다. 이를 해결하려면 아마존 메카니컬 터크, 인사 전문가, 전문 직업 상담사 등 다양한 배경과 경험을 가진 전문가를 선정하는 것이 좋다.
- 또한, 프로그래밍 과정에서 개발자의 의식적-무의식적 편향(예측-비예측) 문제는 오픈소스 도구(예: FairML, Google What-If Tool)를 활용하여 완화 가능하다. 이러한 도구들은 주기적으로 출력 데이터 통계를 분석하여 알려지지 않은 편향을 발견하거나, 미리 지정한 공평성 평가지표에 따라 기능의 위험 여부를 알리는 등의 기능으로 구현과정에서 편향을 조기에 발견하고 대응할 수 있다.

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

Yes ☐ No ☐ N/A ☐

- 사용자 인터페이스와 상호작용 방식은 인터페이스 디자인, 면접 질문에 사용된 언어, 허용되는 응답 유형, 인공지능 시스템이 사용자 입력을 처리하고 해석하는 방식 등 여러 가지 요인으로 인해 편향을 발생시킬 수 있다.
- 이처럼 의도하지 않은 행동이 유도한 편향을 제거하기 위해서는 사용자 인터페이스 디자인 및 구현에서 편향(예: 표현 편향, 순위 편향)을 유발할 수 있는 요인을 파악하고 개선해야 한다.
  - ✓ 표현 편향: 정보가 제시되는 방식에 따라 발생하는 편향이다. 사용자는 가장 눈에 띄는 콘텐츠에 일차적인 관심을 표현하는 경향이 있다. 예를 들어, 채용 담당자/HR 담당자가 지원자의 결과/프로필을 검토할 때, 실제 중요한 등급/관련 직무와의 적합도와는 상관없이 결과 중 가장 크고 선명한 이미지/차트/그래픽 등을 보여준 지원자를 무의식적으로 선택할 수 있다.
  - ✓ 순위 편향: 이러한 편향은 정보가 노출되는 순서에 따라 발생한다. 사용자는 최상위 결과가 가장 관련성이 높고 중요하다고 생각하는 경향이 있다. 또는 인터페이스의 약간 불투명한 도구/구성 요소를 간과하기 쉽다.

안전성 책임성 투명성

요구사항 13 인공지능 시스템의 안전모드 구현 및 문제 발생 알림 절차 수립

- 개인의 커리어와 삶은 물론 사회에 미칠 수 있는 부정적인 영향을 방지하기 위해 채용 인공지능 시스템에는 안전모드와 발생할 수 있는 모든 문제를 해결할 수 있는 알림 절차가 있어야 한다. 이를 통해 시스템이 문제에 신속하게 대응하고 시스템의 결정이나 결과로 인해 발생할지도 모를 잠재적 피해를 완화할 수 있다.

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

Yes ☐ No ☐ N/A ☐

- 안전모드는 보안 위반이나 성능 문제 같은 위험으로부터 시스템을 보호하고 장애나 오류로 인한 문제가 발생하더라도 안전한 상태를 유지하도록 설계된 기능 또는 메커니즘이다.
- 특히 면접 평가 업무는 인공지능 시스템 오류가 발생하면 개인에게 직접적인 영향을 미치기 때문에 기술적인 안전모드를 구현함과 동시에 의사결정 과정에 전문 직업 상담사/HR 전문가를 참여시켜 개발된 모델의 안전성을 확보할 필요가 있다. 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 발생하면 기능 정지, 화면 전환 및 서비스 제공 초기 상태로의 복구, 입력 거절, 의사결정 회피 등의 예외 처리가 이루어지는지 확인해야 한다.
- 채용 인공지능 시스템의 원활한 동작을 위해 사례에 따라 발생할 수 있는 예외 유형과 이를 식별하는 방법을 정의하고, 예외 사항을 적절한 개인 또는 인사팀과 상호작용하여 신속하게 처리할 절차를 수립해야 한다.

참고 채용 인공지능 시스템의 예외 사례

사례	예외 설명
시스템 사용자가 면접관/인사팀원/채용 담당자인 경우	추론 결과에 오류가 발생할 경우, 해당 사용자가 직접 면접 과정을 채점하거나 관련 면접 영상을 평가할 사람에게 보내야 한다.
	모델 추론 결과의 불확실성이 높을 경우, 정확도나 신뢰도가 낮으므로 주의가 필요하다는 알림을 제공할 수 있다.
시스템 사용자가 지원자/면접 대상자인 경우	개인정보 입력 또는 면접 세션 시작 시 문제가 발생할 경우, 지원자 때문이 아닌 시스템 문제로 인한 면접 세션 실패에 대해서는 담당 면접관/채용 담당자에게 알려 도움을 요청하거나 피드백을 제공해야 한다.
	부주의 또는 기기 오류로 인해 입력 데이터가 부정확하면 입력 거부 등의 조치가 취해질 수 있다. 이러한 상황이라면 시스템을 오도하기 위해 고의로 조작한 데이터(예: 글꼴 색상을 변경하여 직책의 키워드를 누락하거나 그 반대)가 있는지 개발자에게 확인할 것을 강력히 권장한다.

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.
- 채용 인공지능 시스템은 이미지와 동영상, 개인정보 등 민감한 데이터를 다루기 때문에 보안 기법을 적용 및 강화하여 무단 액세스로부터 데이터를 보호해야 한다. 이를 통해 사용자 신뢰를 높이고 의도하지 않은 결과에 따르는 위험을 줄일 수 있다.
- 시스템 구현 단계에서 대표적으로 방어해야 하는 공격중 하나는 모델 추출model extraction 이다. 모델 추출 공격은 공격자가 표적 모델에 접근하여 쿼리를 통해 필요한 데이터를 수집 후 쿼리 결과를 사용하여 모델을 복제하는 것이다.[199]

- 채용 인공지능 시스템의 설계와 구현은 채용 평가 프로젝트의 특성으로 인해 결과적으로 불확실할 수 있다. 이러한 이유로 아직 일반적인 모델 추출 공격 사례는 없다. 그럼에도 불구하고, 시스템은 최종 사용자에게 공개된다. 특히 영상 녹화 과정에서는 지원자에게도 공개되기 때문에 추출 공격에 노출될 수 있어 방어 대책이 필요하다.

모델 추출 공격에 대한 방어 기술

방어 기법 분류	방어 기법 내용
쿼리 횟수 제한	모델 공격에 대비하여 특정 시간 동안 수행할 수 있는 쿼리의 수를 제한하는 기술로, 반복적인 쿼리로부터 방어한다.
학습 기반 모니터링	모델 공격에 대한 적극적인 탐지와 경고 알림을 수행하기 위해 머신러닝을 활용하는 등 적극적인 방어 기법이다.
예측 결과의 난독화	예측 결과가 결정 경계에 근접한 경우 예측 결과의 정확도를 임의로 낮추는 기술로, 모델의 세부 속성 추출을 방지한다.
모델의 명시적 밀도 제한	생성된 데이터의 확률 출력에 중점을 둔 기술로, 모델의 획득 가능한 높은 확률 결과를 예상한다. 따라서 데이터는 학습 데이터의 분포 범위 내에 있어야 한다. 이 정보를 활용하여 모델 훈련에 사용될 데이터를 확인하는 데 임곗값을 설정한다.
VAE 및 GAN 방어	VAE 또는 GAN 구조 때문에 이들이 생성한 데이터를 얻기 위해 입력 잠재 벡터를 사용해야 한다. 이러한 벡터는 대상 모델이 잠재 벡터를 활용하여 생성 모델의 출력을 제어할 수 있게 한다. 이러한 생성 모델의 입력 잠재 벡터를 보관하여 공격에 대한 방어 메커니즘으로 활용할 수 있다.
모델 워터마킹	조직을 추적할 수 있는 고유한 워터마크 또는 식별자를 모델에 포함한다. 모델이 추출되면 워터마크가 공격의 출처를 식별하는 데 도움이 된다.
적대적 트리거	모델이 추출될 때만 활성화되는 미묘한 공격 트리거를 모델에 도입한다. 이러한 트리거는 추출된 모델이 배포될 때 부정확하거나 악의적인 동작으로 이어질 수 있다.
동적 모델 업데이트	새로운 파라미터와 구성으로 모델을 정기적으로 업데이트하면 공격자가 변경 사항을 따라잡고 정확한 버전을 추출하기 어렵다.
노이즈 인젝션	추출하는 동안 모델의 출력에 노이즈를 추가하여 공격자가 모델의 아키텍처와 파라미터를 리버스 엔지니어링하기 어렵게 만든다.

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

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템이 면접의 의사결정에 사용될 경우, 모델 학습에 사용되는 민감한 데이터로 인해 차별적 영향을 받을 가능성이 매우 크다[240]. 장애인과 같은 특수한 경우, 의사결정의 결과가 불확실하거나 의사결정이 기존 정책과 충돌하는 경우, 성별이나 목소리로 인해 편향이 발생할 수 있는 경우와 같은 특정 상황에서는 AI 시스템이 내린 결정을 담당 면접관 등이 재정의하는 절차를 마련할 수 있다.
- 기술적으로는 면접관/채용 담당자가 시각화 도구를 활용하여 최종 추론 결과를 출력하고 동시에 학습 결과를 개선할 수 있는 대화형 모델을 구현하며, 정확한 데이터가 모델에 입력될 수 있도록 전문가의 확인을 유도하는 체계적 접근을 고려해야 한다.

13-1d 예상되는 사용자 오류에 대한 안내 및 대응을 제공하는가?

Yes No N/A
☐ ☐ ☐

- 채용 인공지능 시스템의 사용자는 면접관/채용 담당자와 지원자로 나뉜다. 면접관은 개발된 시스템을 면접/채용 프로세스의 주체로, 지원자는 입사 지원 과정의 시험/면접 도구로 활용할 수 있다.
- 이러한 구분은 인공지능에 대한 기술적 이해도가 낮은 사용자, 특히 신체적 제약이 있거나 연령 차이가 있는 사용자로 인해 휴먼에러로 이어질 수 있다. 따라서 서비스 담당자는 다양한 사용자 오류 유형을 이해하고 발생할 수 있는 오류를 사전에 정의 및 분석해야 한다.

참고 채용 인공지능 시스템에서 발생할 수 있는 사용자 오류의 예시

에러	콘텐츠
잘못된 입력	사용자가 부정확하거나 일관되지 않은 정보를 제공할 우려가 있는데, 인사팀에서는 직무 설명에 대한 구체적인 정보자 지원자에게는 개인정보일 수도 있다.
불완전한 입력	사용자가 질문에 대한 필수 답변이나 필요한 첨부 파일과 같은 중요한 정보를 제공하는 것을 잊어버릴 수 있다.
질문의 잘못된 해석	사용자가 질문 내용을 잘못 이해하고 관련 없는 답변을 제공할 수 있다.
기술적 문제	인터넷 연결이 느리거나 기기가 오작동하는 등 기술적 문제가 발생하여 정확한 정보를 제공하지 못할 수 있다.
의도하지 않은 동작	사용자가 중요한 단계를 건너뛰거나 버튼을 잘못 클릭하는 등 실수로 의도하지 않은 작업을 수행할 수 있다.

- 사용자 오류에 따른 사전 대응 방안은 다음과 같다.
  - ✓ 제약조건 설정: 허용 가능한 옵션을 정의하여 표시하거나 사용자의 선택을 어느 정도 제한하여 잘못된 사용자 입력을 방지한다.
  - ✓ 오류 메시지: 사용자에게 명확하고 간결한 오류 메시지를 제공하여 무엇이 잘못되었는지와 오류를 수정하는 방법을 설명한다.
  - ✓ 시스템 제안·정정: 자주 발생하는 사용자의 실수를 수집하고, 실제 서비스 시 유사한 사용자 실수가 발생하면 시스템에서 자동 정정하거나 올바른 입력을 제안한다. 예를 들어, 검색 시 오탈자가 나면 정정하여 추천하는 것 등이다.
  - ✓ 기본값 설정: 시스템에서 자주 사용되는 값을 기본값으로 먼저 제공하거나 관련 예시를 제공함으로써 사용자 오류를 줄일 수 있다. 예를 들어, 지원자를 위한 표준 면접 시나리오에서 표준 유형의 면접 지원서를 기본값으로 추가할 수 있다.
  - ✓ 재확인·결과제공·실행취소: 사용자로부터 받은 입력값을 재확인하고 예상 결과를 미리 전달한다. 또한, 잘못된 결과를 취소하는 등의 기능으로 오류를 방지할 수 있다.

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

Yes ☐ No ☐ N/A ☐

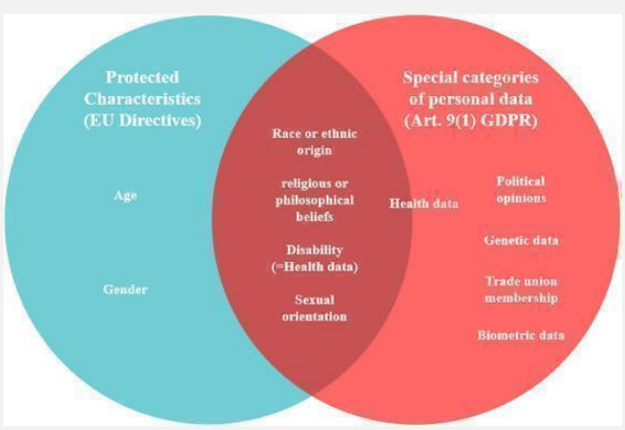
- 채용 인공지능 시스템은 성능 저하 및 공격, 편향된 데이터로 학습, 민감 데이터 오남용, 모델 학습 과정에서의 보호 데이터 오평균화, 외부 공격, 서비스 중 사용자의 오남용 등 다양한 문제가 발생할 수 있다.
- 따라서 채용 인공지능 시스템은 윤리적 문제와 성능 저하에 대비하기 위해 자체 점검 기능을 갖춰야 한다. 잠재적 사용자를 고려하여 편향과 차별 등의 윤리적 문제 점검 기능과 절차, 아울러 성능 저하의 지속적인 평가지표와 관리 절차 또한 필요하다. 또한, 그 내용을 운영자에게 전달할 수 있어야 한다.
- 사용자 의견 전달 기능은 시스템의 일시적인 오류나 도출 결과에 편향이 발생하는 등의 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 ☐

- 채용 인공지능 시스템에서 편향이나 차별 등 윤리적 문제가 발생할 가능성을 점검하고, 문제 발생 시 알림 기능이나 절차가 마련되었는지 확인한다. 윤리적 문제 알림 절차는 인공지능 시스템 자체에 대한 신뢰도를 평가하는 기준과 점검 항목이 될 윤리 원칙을 마련하는 것이 먼저이다.
  - ✓ 예: 인권, 프라이버시, 다양성 존중, 비침해성, 공공성, 연대성, 개인정보 관리, 책임성, 안전성, 투명성, 라이선스 관리, 민감 데이터의 사용 및 보관 등
- 다음으로 이를 개발자, 사용자, 운영자 등 모든 이해관계자가 준수할 수 있도록 교육해야 한다. 또한, 사용자가 AI 시스템에서 편향, 차별과 같은 비윤리적 행동 사례를 신고할 수 있는 창구도 필요하다. 마지막으로, 이를 정기적으로 검토하고 업데이트하여 그 효과를 확인하고, 인식한 문제를 해결하기 위한 적절한 조치를 해야 한다.

참고 '특수 범주 데이터'와 '보호되는 차별 금지 근거'의 구분[240]



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 성능이 변할수 있다. 특히 면접 AI 시스템에서는 시스템이 구축되는 국가에 따라 축적하는 데이터 환경이 크게 달라질 수도 있다.
- 인공지능 시스템은 실제 서비스 운영 중 갑자기 성능이 저하됐을 때 원인을 바로 알기 어려우므로, 시스템의 성능 저하를 지속해서 평가하고 관리할 지표와 절차가 설정되었는지 점검해야 한다.
- 채용 인공지능 시스템에 사용할 수 있는 성능 지표에는 정확도, 속도, 사용자 만족도, F1-score, IoU[Intersection over Union], mAP[mean average precision], recall, 특이도[specificity], 정밀도[precision], 위협 점수[threat score], 진양성, 진음성, 오양성, 오탐 등이 있다. 평가 결과 성능 저하가 확인되면 이를 시스템 운영자에게 전달하고, 운영자는 성능 저하 원인을 찾아 개선하는 등의 절차를 마련해야 한다.
- 성능 개선 작업 시에는 개인정보 보호법[243] 등 개인정보와 민감정보 처리의 제한규정을 준수해야 한다.

투명성

요구사항 14 인공지능 시스템의 설명에 대한 사용자의 이해도 제고

- 채용 인공지능 시스템을 사용하는 지원자들은 자신이 직무 요건을 충족했음에도 불구하고 왜 시스템에 선택되지 않았는지 의문을 품기 쉽다. 이는 채용 인공지능 시스템에 가장 많이 제기되는 투명성 문제이다.
- 따라서 인공지능 시스템의 운영자 또는 서비스 제공자는 사용자에게 제공되는 결과가 이해 가능한지 understandability, 해석 가능한지interpretability, 설명 가능한지explainability를 점검하고 보완해야 한다.

참고 워싱턴 D.C. 20580 – 인공지능에 대한 FTC 규칙 제정을 위한 EPIC 청원서[248]

2020년 초 전자 프라이버시 정보 센터(EPIC) 변호사는 Company H의 심사 관행의 불공정성을 강조하는 FTC 제소를 제기하였다.

> H[redacted] administers an online "video interview" and/or an online "game-based challenge[]" to the candidate.[41] H[redacted] collects "tens of thousands of data points"[42] from each video interview and a "rich and complex" array of data from each "psychometric game[.]"[43] H[redacted] then inputs these personal data points into "predictive algorithms"[44] that allegedly determine each job candidate's "employability," "cognitive ability," "psychological traits," "emotional intelligence," and "social aptitudes."[45] But H[redacted] does not give candidates access to the training data, factors, logic, or techniques used to generate each algorithmic assessment. In some cases, even H[redacted] is unaware of the basis for an assessment.[46]

14-1 인공지능 시스템 사용자의 특성user characteristics과 제약사항을 분석하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템의 판단이 적절한지 평가하기 위해서는 먼저 해당 결과를 읽는 사용자를 고려해야 한다. 채용 인공지능 시스템의 사용자는 크게 채용 담당자와 지원자, 두 그룹으로 분류할 수 있다. 사용자가 누군지에 따라 결과(설명)의 수준, 깊이 그리고 맥락이 정해지므로 사용자를 자세히 분석해야 한다.

14-1a 사용자 특성에 따른 세부 고려사항을 분석하였는가?

Yes No N/A
☐ ☐ ☐

- 채용 전문가, 지원자 등 다양한 사용자가 시스템을 사용하게 되는데, 이들은 배경이 서로 다를 수 있으므로 AI 시스템의 사용자 특성과 제약을 고려해야 한다.
- 사용자 특성에 따른 세부 고려 사항을 분석하려면 먼저 채용 인공지능 시스템의 대상 사용자 그룹과 이들의 특성(예: 인구 통계, 기술 전문성 등)부터 파악해야 한다. 그룹별로 채용 인공지능 시스템에 대한 요구사항, 목표, 기대치를 파악하고, 대상 사용자 그룹의 한계와 제약 조건이 채용 인공지능 시스템 사용에 어떤 영향을 미칠 수 있는지 평가한다.
- 분석 과정에서 얻은 인사이트는 채용 인공지능 시스템을 설계하고 개발하는 시점에 통합하여 사용자 고유의 특성과 요구사항을 충족할 수 있도록 한다.

사용자 특성 분석을 위한 고려 사항 예시[249~251]

구분	세부 내용	고려 사항
연령	청소년, 성인, 고령자 등	• 청소년이 시스템을 사용하는 경우, 이해할 수 있는 용어나 어휘가 성인보다 제한적일 수 있다. • 고령자는 기술 수준 차이로 인해 시스템 사용을 학습하는 데 시간이 더 많이 소요될 수 있다.
성별	남성, 여성 등	• 성별에 따라 성평등 정책에 대한 이해가 다를 수 있으며, 특히 국가에 따라 여성 고용 관련 정책이 없는 경우도 있다.
인종	아시아인, 유럽인 등	• 인종마다 피부색이나 신체 크기를 인식하는 평균 기준이 다를 수 있으며, 억양이나 방언의 차이에 따라 평균 기준도 다를 수 있다.
장애	장애인, 비장애인	• 신체 크기, 신체 능력, 인지 능력의 차이 또는 제한으로 인해 면접 조건이 제한될 수 있다. 또한, 서비스 대상 국가에서 이러한 상황에 대해 어떤 제도적 보완책을 제공하고 있는지 항상 확인하고 고려해야 한다. * 대한민국은 “장애인고용촉진 및 직업재활법” 참조[252]
지식 수준	일반인, 개발 전문가, 컴퓨터 엔지니어 등	• 컴퓨터 및 정보 서비스에 대한 경험 또는 기술 인식 수준에 따라 서비스 이해도에 차이가 있을 수 있다.

14-2 사용자 특성에 따른 설명을 제공하는가?

Yes No N/A
☐ ☐ ☐

- 최근 연구에 따르면 사람들은 특히 복잡하거나 민감한 사안일수록 알고리즘 결과가 성공적이라 해도 인공지능이 내린 결정을 신뢰하지 않는 것으로 나타났다[253]. 인공지능 시스템이 의도한 사용자 그룹의 특정 요구와 기대에 부합하도록 설계하려면 사용자 특성을 고려해 충분한 설명을 제공해야 한다.
- 채용 인공지능 시스템의 결과물은 서비스를 이용하는 다양한 사용자에 따라 다른 관점으로 해석되거나 오해될 수 있다. 따라서 14-1에서 분석한 사용자 특성을 고려하여 설명을 평가할 수 있는 기준 항목을 반영하여 명확하게 이해할 수 있는 표현으로 설명하는 것이 바람직하다.

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 평가 기준별 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?

참고 HMI[Human Machine Interface] 설계 권장 사항 목록[257]

채용 인공지능 시스템에서 인간-기계 사이의 인터페이스 원칙이 충족되는지 평가하는 것도 도움이 된다[257]. 하단 체크리스트는 채용 목적으로 새로운 성과 추적 방법의 필요성을 감지한 연구자들이 HMI에 사용하는 설계 메커니즘이다.

구분	원칙	명확성	구체성	적절성	정확성
기술 및 소프트 스킬	• 시스템 산출물을 위해 기술 지식과 소프트(개인 커뮤니케이션) 기술 수준을 모두 포함하는 활동을 사용한다. • 완전한 프로필에는 기술 및 사회적 기술 활동이 포함되어야 한다.	V	V		
품질	• 직무 요건과 관련된 지원자의 기여도를 반영하는 것이 중요하므로 기여도를 강조한다.			V	
지역 사회에서의 현재 사회적 지위	• 연구에 따르면 대인 커뮤니케이션 능력과 성격 특성은 항상 평가지표로서 중요한 성공 요인으로 간주된다. • 따라서 개인 행동 특성에 대한 조직 평가 결과를 제공할 것을 권장한다[240].				V
활동 요약	• 활동 유형에 따라 전문 분야를 간략하게 요약한다. • 사용자의 전문성과 경험을 평가하기 위해 프로필을 조사해야 하는 부담을 줄여준다.		V		V
요약 시각화	• 사용자는 대부분 결과보다는 평가/확인/조사/검토에 노력이 적게 드는 것을 선호한다. • 따라서 출력 내용을 시각적으로 요약하면 검수자가 활동/결과를 빠르게 평가하고 확인하는 데 도움이 된다.	V	V	V	
드릴다운 drill down 허용	• 사용자는 시스템의 결과물 요약 정보에 접근하여 검토하는 경향이 있으며, 평가 결과에 대한 보다 자세한 정보를 조사하려 한다. • 따라서 사용자가 원한다면 시스템에서 제공하는 세부 활동/출력 정보에 쉽게 접근할 수 있도록 한다.	V	V		

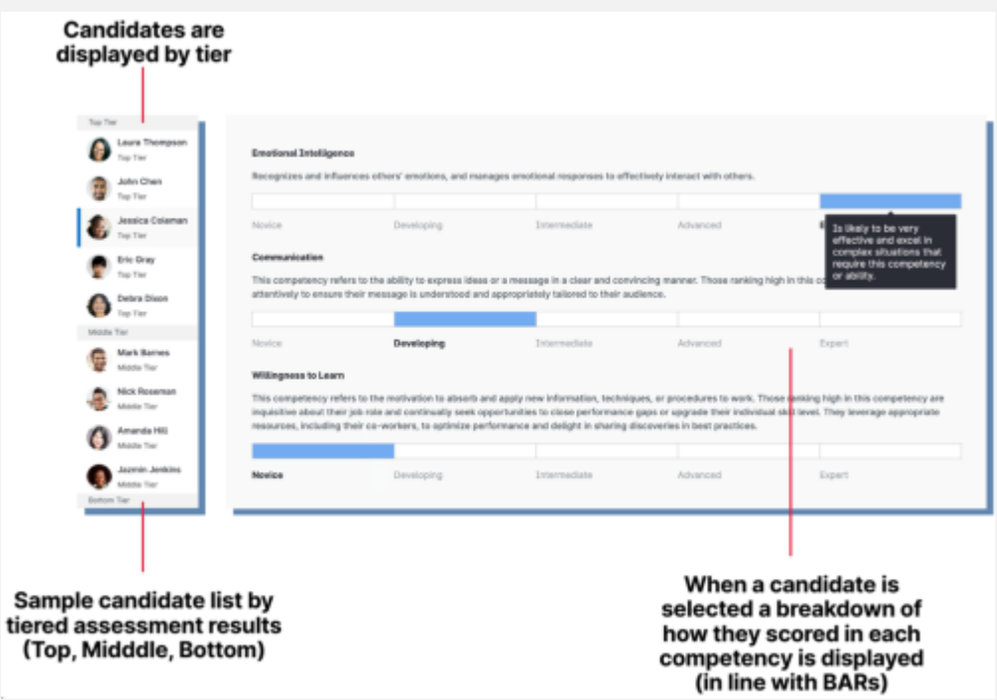
14-2b 사용자가 이해하기 어려운 전문 용어 사용을 지양하였는가?

Yes ☐ No ☐ N/A ☐

- 텍스트로 설명할 때는 전문 채용팀과 지원자를 고려하여 가급적 전문 용어는 피하고, 필요시 용어 설명을 추가로 제공해야 한다.
- 이를 위해 사용자 조사를 통해 사용자의 배경, 지식 수준, 언어 선호도를 파악할 수 있다. 또한, 채용 분야의 전문가와 협력하거나 채용 인공지능 시스템을 적용하는 대상 산업 도메인을 심층적으로 이해하고 있으면 사용자 그룹에 적합한 용어를 택하는 데 도움이 된다.
- 마지막으로, 대상 사용자 그룹 중 일부를 선정하여 설명을 테스트하면, 사용한 용어의 효과에 대한 추가 인사이트를 얻을 수 있다.

참고 채용팀에서 사용자에게 제공하는 설명 예시

- 사용자가 쉽게 이해하고 해석할 수 있는 인터페이스를 준비하는 것도 AI 시스템의 신뢰성 측면에서 중요한 요소이다.
- 인터페이스를 준비할 때는 모든 사용자가 이해할 수 있는 공통된 용어를 택해야 한다.
- 전문 용어와 단어 선택은 아래 예시와 같이 명확하고 간단하며 이해하기 쉬워야 한다.



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

Yes ☐ No ☐ N/A ☐

- 명확한 언어와 문장은 사용자의 구체적인 행동과 이해를 끌어낸다. 따라서 설명을 간결하고 명확하게 하여 모호하게 해석하는 일이 없도록 작성해야 한다.
- 시각적으로는 성공·실패·경고·위험 등 결과에 따른 색상을 일관성 있게 유지하여 사용자가 한눈에 시스템 결과를 이해하도록 도울 수 있다. 텍스트나 음성으로 제공되는 설명에서는 지시대명사를 사용하지 않고 대상을 명확하게 말해주는 것도 좋다. 또한, 비슷한 발음이 연달아 이어질 때는 다른 단어로 대체하는 것이 바람직하다.

참고 사용자를 위한 사용자 인터페이스 상호작용 및 시각화 방법 예시(1) [259]

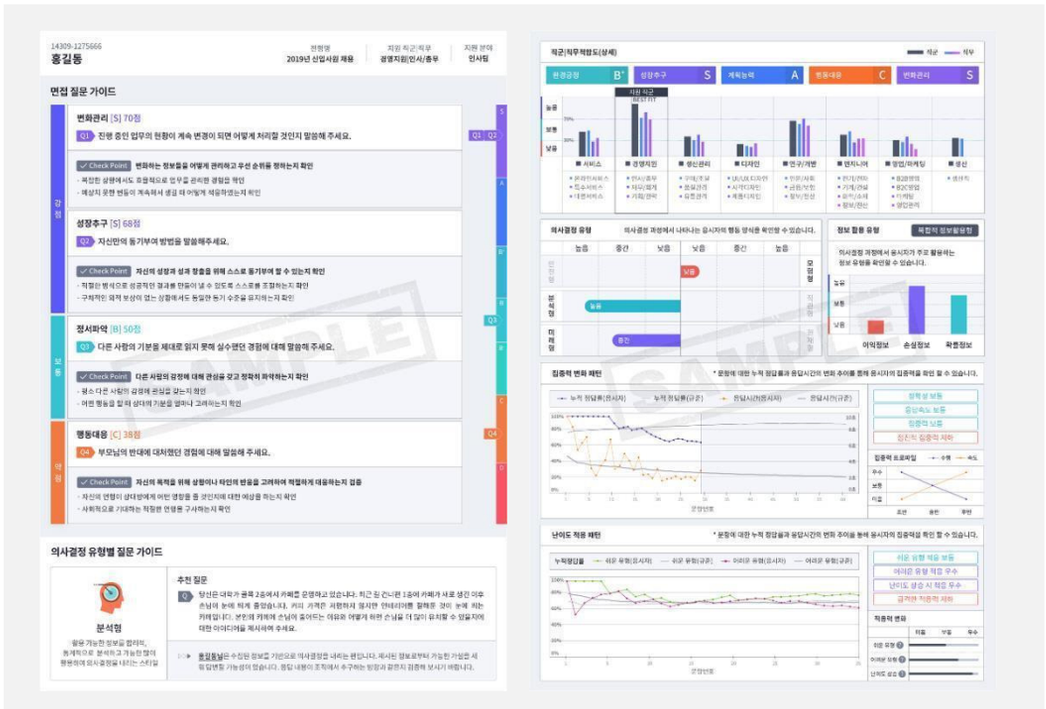
다음은 HR 솔루션을 제공하는 제네시스랩의 뷰인터HR 평가 결과서 예시이다. 지원자가 면접 AI 모델을 사용하여 면접 과정을 완료하면 시스템은 종합적인 평가와 패턴 분석을 생성하고, 그 결과를 접근 권한에 따라 두 그룹(14-1참고)과 공유한다.

참고 사용자를 위한 시각화 방법 예시(2)[260]

아래 결과는 국내 온라인 채용 솔루션을 제공하는 마이다스아이티의 'AI 역량검사 백서'에서 발췌한 내용이다.

지원자가 AI 검사를 완료하면, 시스템은 세부적인 등급과 패턴 분석 결과를 생성하고, 그룹별 접근 권한에 따라 채용자와 지원자 두 사용자 그룹에 적합한 결과를 공유한다.

04 시스템 구현


14309-1275666
홍길동
전형명
2019년 신입사원 채용
지원 직군/직무
경영지원(인사/총무
지원 분야
인사팀
면접 질문 가이드
변화관리 [S] 70점
Q1 진행 중인 업무의 현황이 계속 변경이 되면 어떻게 처리할 것인지 말씀해 주세요.
Check Point 변화하는 정보들을 어떻게 관리하고 우선 순위를 정하는지 확인
- 복잡한 상황에서도 효율적으로 업무를 관리한 경험을 확인
- 예상치 못한 변동이 계속해서 생길 때 어떻게 적응하였는지 확인
성장추구 [S] 68점
Q2 자신만의 동기부여 방법을 말씀해주세요.
Check Point 자신의 성장과 성과 향상을 위해 스스로 동기부여 할 수 있는지 확인
- 적절한 범석으로 성공적인 결과를 만들어 낼 수 있도록 스스로를 조절하는지 확인
- 구체적인 외적 보상이 없는 상황에서도 동일한 동기 수준을 유지하는지 확인
정서파악 [B] 50점
Q3 다른 사람의 기분을 제대로 읽지 못해 실수했던 경험에 대해 말씀해 주세요.
Check Point 다른 사람의 감정에 대해 관심을 갖고 정확히 파악하는지 확인
- 청소 다른 사람의 감정에 관심을 갖는지 확인
- 어떤 행동을 할 때 상대에 기분을 헤아려 고려하는지 확인
행동대응 [C] 38점
Q4 부모님의 반대에 대처했던 경험에 대해 말씀해 주세요.
Check Point 자신의 목적을 위해 상황이나 타인의 반응을 고려하여 적절하게 대응하는지 검증
- 자신에 연령이 상대방에게 어떤 영향을 줄 것인지에 대한 예상을 하는지 확인
- 사회적으로 기대하는 적절한 언행을 구사하는지 확인
의사결정 유형별 질문 가이드
분석형
추천 질문
Q 당신은 대학가 골목 2층에서 카페를 운영하고 있습니다. 최근 집 건너편 1층에 카페가 새로 생긴 이후 손님의 눈에 띄게 줄었습니다. 커피 가격은 저렴하지 않지만 인테리어를 잘해둔 것이 눈에 띄는 카페입니다. 본인의 카페에 손님이 줄어드는 이유와 어떻게 하면 손님을 더 많이 유치할 수 있을지에 대한 아이디어를 제시하여 주세요.
→ 홍길동님은 수집된 정보로 기반으로 의사결정을 내리는 편입니다. 제시된 정보로부터 가능한 가설은 세 개 입니다. 가능성이 있습니다. 현재 내용이 조직에서 추구하는 방향과 같은지 검증이 보시기 바랍니다.
직군|직무적합도(상세)
지원 직군
환경감정 B
성장추구 S
계획능력 A
변화대응 C
변화관리 S
서비스
경영지원
생산관리
디자인
연구개발
인사이트
영업/마케팅
생산
의사결정 유형
정보 활용 유형
복합적 정보활용형
높음
보통
낮음
어익정보
손실정보
확률정보
집중력 변화 패턴
집중력 프로파일
난이도 적응 패턴
적응력 변화

14-2d 설명이 필요한 위치와 타이밍은 적절한가?

Yes No N/A
☐ ☐ ☐

- 채용 인공지능 시스템은 지원자의 행동이 실시간으로 평가에 반영되기 때문에, 설명도 지원자가 면접에 방해받지 않는 위치와 타이밍에 제공하는 것이 중요하다. 적절한 설명 타이밍은 면접 프로세스의 정확성과 효율성을 개선하는 데 도움이 된다. 이를 위해 설명이 단발성이어야 하는지, 반복하여 강조해야 하는 것인지 숙고하고, 어느 위치에 놓여야 사용자가 잘 읽을 수 있는지 고려해야 한다.
- 설명의 올바른 위치와 타이밍은 14-2e의 애널리틱스 및 A/B 테스트와 같은 사용자 조사 기법을 사용하여 찾아낼 수 있다.

참고 채용 인공지능 시스템 설명 제공 타이밍 예시

지원자에게 제공되는 지침은 명확하고 간결하며 모호함이 없어야 한다.

타이밍	방법	고려 사항
면접 중	실시간 설명	• 채용 인공지능 시스템에서 지원자와의 대화형 면접(예: 챗봇 또는 화상 면접)이 가능한 경우, 지원자는 면접 중에 질문하거나 지침에 대한 설명을 요청할 수 있어야 한다. • 이러한 실시간 상호작용은 오해를 방지하고 지원자가 답변하기 전에 질문을 완전히 이해하는 데 도움을 준다.
면접 전	면접 전 지침	• 면접 전에 지원자는 면접 진행 방식, 사용되는 플랫폼 및 기술 요구사항에 대해 명확한 지침을 받아야 한다. • 또한, 지원자가 명확하게 이해하도록 돕기 위해 지원 가능 여부 또는 문의처 정보를 제공해야 한다.
면접 후	면접 후 소통	• 면접이 끝난 후, 지원자가 모호하거나 어려운 질문에 대해 설명을 요청할 수 있는 옵션이 있어야 한다. • 이는 이메일, 지정된 커뮤니케이션 채널 또는 면접 피드백 양식을 통해 진행할 수 있다.
지원 채널	전용 지원 채널	• 채용 인공지능 시스템에는 이메일 또는 채팅 지원과 같은 전용 지원 채널이 필요하다. • 이러한 전용 지원 채널을 통해 지원자가 면접을 진행하는 과정 중 언제든 설명이나 기술 지원을 요청할 수 있어야 한다.
인앱 지원	인앱 도움말	• 채용 인공지능 시스템에 사용자 인터페이스 또는 앱이 있는 경우, 면접 진행 과정에서 지원자가 가질 수 있는 일반적인 질문이나 애로사항에 참조할 수 있는 도움말 또는 FAQ를 제공할 수 있다.
적시 응답	즉각적인 피드백	• 채용 인공지능 시스템은 원활하고 효율적인 면접 프로세스를 보장하기 위해 지원자의 질문에 적시에 응답하는 것을 목표로 해야 한다.

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 사용하면서 느끼는 모든 것을 의미하며, 그 개인이 인지하는 유용성, 사용 편의성, 효율성 같은 시스템 특성을 포함한다. 설명을 평가하기 위해 사용자 조사user research 기법을 활용할 수 있다.

사용자 경험 조사 기법 예시

기법		설명
정량적 기법	설문조사	사용자 경험에 대한 정보를 수집하기 위해 사용자에게 일련의 질문을 하고 답변을 받는 방법
	분석	웹 분석, 이벤트 추적 및 로그 데이터를 사용하여 사용자 행동에 대한 정보를 수집하고 분석하는 방법
	A/B 테스트	다양한 버전의 채용 인공지능 시스템을 비교하여 어떤 시스템이 최상의 사용자 경험을 제공하는지 확인하는 방법
정성적 기법	면접	채용 인공지능 시스템에 대한 사용자의 생각과 경험을 이해하기 위해 사용자와 일대일로 심도 있는 면담을 진행하는 방법
	포커스 그룹	채용 인공지능 시스템에 대한 다양한 사용자 피드백과 의견을 수집하기 위해 사용자 그룹을 선정하여 토론을 진행하는 방법
	사용성 테스트	사용자가 채용 인공지능 시스템으로 특정 작업을 수행하여 행동을 관찰하고 피드백을 수집하는 통제된 테스트 방법
행동 조사		채용 인공지능 시스템과 사용자의 행동 및 상호작용을 관찰하는 데 중점을 두며, 시스템의 효과와 유용성에 대한 인사이트를 얻을 수 있는 방법
태도 조사		채용 인공지능 시스템을 사용하는 동안 사용자의 신념, 감정, 태도를 이해하는 데 중점을 두며, 기술에 대한 사용자의 인식과 수용 관련 인사이트를 얻을 수 있는 방법

05 운영 및 모니터링

책임성 투명성

요구사항 15 서비스 제공 범위 및 상호작용 대상에 대한 설명 제공

- 채용 인공지능 시스템의 오용 및 남용을 방지하기 위해 사용자에게 시스템의 목적, 범위, 제약사항, 면책조항 등에 대한 명확한 설명을 제공한다.

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

Yes No N/A
☐ ☐ ☐

- 최근 채용 인공지능 시스템이 시장에 확산되면서 구직자들이 시스템을 실제 서비스 제공 범위보다 광범위하게 이해하거나 서비스 기능에 대한 기대치가 높은 경우가 많다. 특히 대면이 아닌 원격으로 지원서를 받도록 설계된 온라인 채용 인공지능 시스템의 특성상[272], 사용자의 초기 기대가 왜곡되거나 잘못 해석될 가능성이 크다. 따라서 인공지능 기술의 오남용을 방지하고 서비스에 대한 사용자의 기대치를 조정하기 위해, 서비스의 목적, 범위, 한계, 면책사항 등을 상세히 설명해야 한다.
- 시스템 운영자는 AI 시스템의 결과물이 사용자에게 미치는 영향과 결과를 되돌릴 수 있는지 등을 설명하여 사용자가 서비스를 올바르게 사용하도록 유도해야 한다.

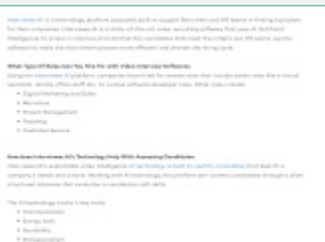
15-1a 서비스의 목적과 목표에 대한 설명을 제공하는가?

Yes No N/A
☐ ☐ ☐

- 채용 인공지능 시스템은 주로 기업의 채용 과정에서 인사부서의 부담을 덜 수 있게 설계된다. 따라서 타겟 사용자 그룹에게 서비스의 목적goal과 목표objective를 비교적 쉽고 명확하게 설명할 수 있다. 반면 이용자가 지원자(채용 대상자)라면 사전에 서비스 전반에 대한 상세하고 알기 쉬운 설명을 제공해야 한다.
- 또한, 채용 인공지능 시스템이 지원자를 자동으로 판단하고 채점하는 방식(전문가의 개입 없이 1차 채용 평가를 위한 시스템)으로 개발된 경우 서비스의 목적과 취지를 모든 이용자에게 사전에 설명해야 한다.
- 인공지능 서비스가 오용 또는 남용될 경우, 인공지능 모델이나 시스템상의 새로운 취약점이 생성되거나 예상치 못한 사회적 이슈가 발생할 수 있다. 따라서 서비스가 의도한 목적을 벗어나 잘못 사용되는 것을 방지하기 위해, 이해관계자는 잠재적 오남용 영역을 식별하고, 사용자가 이를 인식할 수 있도록 관련 사례와 처벌 내용 등을 알려야 한다.
- 또한, 잠재적 사용자의 다양성은 광범위하고 예측할 수 없으므로, 정보 문서/가이드라인[273]을 준비하고 시스템의 목적과 목표를 사용자에게 알려야 한다.

참고 채용 인공지능 플랫폼에서 제공하는 서비스 목적 예시

다수의 서비스 중인 채용 인공지능 플랫폼에서는 서비스의 목적, 목표를 포함한 서비스 정보를 별도의 페이지에 게시하여 사용자에게 설명한다. 해당 사이트의 서비스 목적, 목표, 대상 등의 설명을 참조할 수 있다.

게시 화면		
출처	https://interviewer.ai/success-story/video-interviews-for.hiring/	https://ideal.com/ai-recruiting/

15-1b 서비스의 한계와 범위에 대한 설명을 제공하는가?

Yes ☐ No ☐ N/A ☐

- 채용 인공지능 시스템은 모델 학습 과정에서 발생하는 제약으로 인해 한계가 있을 수 있다. 이러한 한계는 제한된 학습 데이터 사용량이나 녹음 과정의 환경적 영향 등에 기인한 것이므로, 최종 사용자에게 채용 인공지능 시스템의 범위와 한계를 설명함으로써 사용자의 기대치를 조정할 수 있다.
- 채용 인공지능의 일반적인 한계
  - ✓ 특정 시나리오 또는 특정 유형의 데이터에서 정확성 또는 효율성 제한
  - ✓ 특정 언어, 억양 또는 지역에 대한 적용 범위 및 지원 제한
  - ✓ 복잡하거나 미묘한 작업을 처리하는 능력 제한
  - ✓ 리소스 또는 사용량 제한으로 인한 가용성 및 용량 제한 등

참고 하이어뷰 플랫폼[255]의 서비스 범위 한계 설명

하이어뷰 플랫폼[18]의 단어 발음 해석으로 인한 한계

- 지원자 인사이트 보고서를 자세히 설명하기 위해 모델 입력을 수정하고 결과 점수의 변화를 측정하여 각 기능의 상대적 중요성을 파악한다. 결과는 입력 기능의 정렬된 목록과 각각의 상대적 강도로 표시하며, 개별 단어 분석을 통해 우수한 성과를 보인 지원자들의 주제에서 추출된 패턴을 확인한다.
- 그러나 같은 단어라도 문장 내 위치에 따라 의미가 달라질 수 있으므로 CAKE 모델은 단어를 독립적으로 처리하지 않고 맥락 안에서 해석한다. 모델을 맥락에서 설명하기 위해 단어 대신 문장과 구절을 제거하여 모델이 미치는 영향을 조사한다. 예를 들어, '팀'이 협업 모델에서 긍정적인 입력으로 간주되는 경우를 확인할 수 있다.
- 예시 문장과 그들의 영향을 확인한 후, 모델 점수에 큰 영향을 미치는 패턴과 주제를 분석한다. 결과적으로, 모델의 동작은 학습 데이터를 생성하는 인간 평가자가 사용하는 BARS[Behaviourally-Anchored Rating Scale]와 잘 일치하는 것으로 나타났다.

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

Yes ☐ No ☐ N/A ☐

- 채용 인공지능 시스템은 사용하는 데이터의 민감성과 AI 시스템이 내리는 결정의 중요도 등으로 인해 고위험 시스템에 해당한다. 특히 구직자들은 상호작용 대상이 사람인지 AI 시스템인지 혼동할 수 있으므로, 서비스 제공자는 사용자에게 상호작용 대상과 내용을 명확히 전달하여 혼란을 최소화해야 한다.

15-2a 사용자와 인공지능이 상호작용하는 서비스 범위를 명시하였는가?

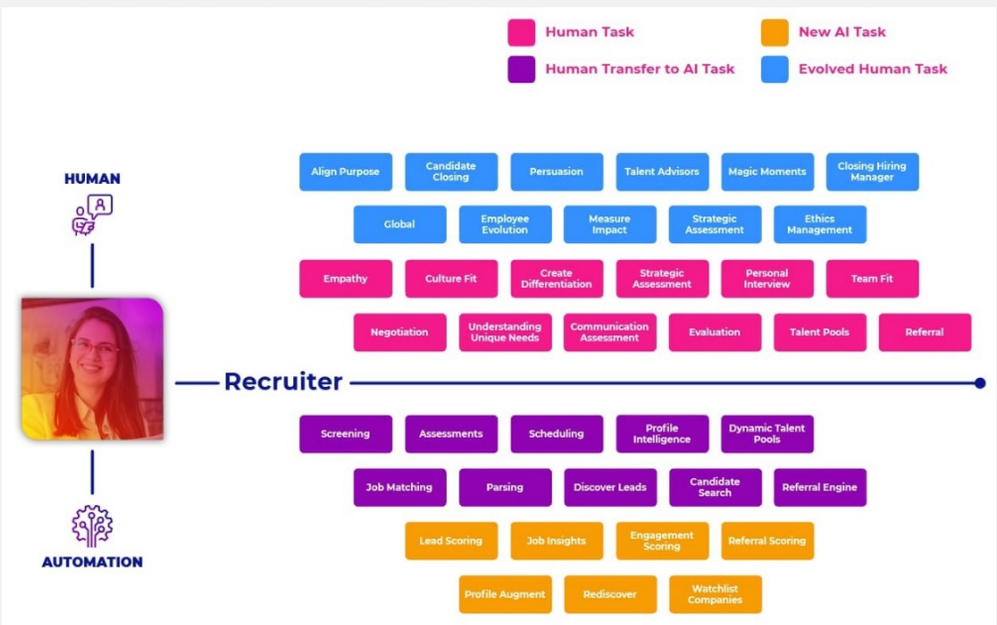
Yes ☐ No ☐ N/A ☐

- 사용자에게 채용 인공지능 시스템과 상호작용하고 있다는 사실을 명확하게 전달해야 한다. 상호작용을 시작할 때 소개 메시지를 제공하거나, 시스템의 정체성을 나타내는 로고 또는 이름을 표시하는 등 다양한 방법으로 알릴 수 있다.
- 사용자가 AI 시스템과의 상호작용에서 기대할 수 있는 것과 필요한 정보를 명확하게 설명하여 시스템의 목적과 기능을 이해하면 사용자가 시스템에 참여할 가능성이 커진다. 사용자에게 AI와 상호작용하고 있다는 사실을 알려주는 방법은 다음과 같다.
  - ✓ 1단계: 사용자가 채용 인공지능 시스템과 상호작용하는 서비스 범위 지정
    - 채용 인공지능 시스템의 구체적인 특징과 기능은 의도된 사용 사례, 서비스를 제공하는 공급업체 또는 개발자에 따라 명확하게 언급되어야 한다.
  - ✓ 2단계: 채용 인공지능 시스템으로 면접을 진행한 후, 서비스 내에서 인공지능이 최종 의사결정을 내릴지 여부를 지정해야 한다.
    - 채용 인공지능 시스템 제공업체는 의사결정 과정에서 AI의 역할을 투명하게 공개하고, 이를 사용자에게 명확하게 전달해야 한다.
    - 예를 들어, 채용 인공지능 시스템은 면접 내용 분석을 기반으로 제안된 포지션에 가장 적합한 지원자를 결정할 수 있는데, 모든 참가자를 분류한 후에도 결정할 수 있다. 이 경우 순위 편향성을 고려해야 한다.

참고 사용 예시 – 채용 인공지능 시스템 설계 시 고려해야 할 상호작용[276]

채용 인공지능 시스템을 사용할 때 지원자가 인공지능 시스템과 어떻게, 얼마나 상호작용하는지, 어느 단계에서 지원자가 누구와 소통하는지(사람/AI 봇) 등의 정보를 제공하지 않으면 지원자는 혼란스러울 수 있다. 따라서 지원자에게 어떤 상황에서 무엇을 해야 하는지 명확하게 알려주어야 한다.

• 지원자, 채용 담당자, AI 시스템 간의 상호작용 예시



• 인터랙션 방식을 설계하기 위한 가이드라인

분류		고려 사항
포함 정보	채용 정보	채용/채용 프로세스에 대한 일반 정보
	면접 과정	면접 절차에 대한 일반적인 정보
	다음 계획/단계	면접 결과에 따라 취할 수 있는 다음 계획/단계
	시스템 정보	시스템에 대한 일반 정보(사용된 데이터, 시스템 권한 부여 등)
	시스템 입력	사용자가 입력한 데이터
	시스템 처리	면접 평가 과정에 사용되는 시스템 알고리즘 또는 기술 절차 관련 정보
	시스템 출력	시스템에서 출력하는 데이터(면접 등급, 추천, 비교 등)
정보 전송 방법	공감(안심)	추론 결과에 대해 안심할 수 있도록 신중하게 선별한 문구를 지원자에게 전달
	명확성, 일반성	교육 수준이 다르거나 기술 이해력이 부족한 지원자/채용 담당자를 위한 쉬운 용어 사용
사용자 상호작용	입력 확인	데이터 입력 확인 가능 여부(확인, 취소 등)
	인사 전문가/직업 상담사/채용 담당자와의 면접	담당자와의 면담 또는 결과/과정 등에 대한 질문 가능 여부
	입력 비교 (시각화)	여러 입력을 비교할 수 있는 시각화 정보 제공
	정보 요청 가능성	요청 시 추가 세부 정보 제공 가능 여부

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 명시해야함을 언급하고 있다.

PART 3

부록

01 약어표

약어표

UN	European Union
WEF	World Economic Forum
WHO	World Health Organization
AIVI	Artificial Intelligence Video Interview
EEOC	Equal Employment Opportunity Commission
IEEE	Institute of Electrical and Electronics Engineers
PIPC	Personal Information Protection Commission
GDPR	General Data Protection Regulation
NIST	National Institute of Standards and Technology
ATS	Applicant Tracking Systems
EurWORK	The European Observatory of Working Life
EPSRC	Engineering and Physical Sciences Research Council
AI	Artificial Intelligence
CV	Cross-validation
TFDV	TensorFlow Data Validation
KNN	K-nearest Neighbor
LoOP	Local Outlier Probability
GMM	Gaussian Mixture Models
LOF	Local Outlier Factor
iForest	Isolation Forest
CUSUM	Cumulative Sum Control Chart
C&W	The Carlini and Wagner attack
PGD	Projected Gradient Descent
RoVISQ	Reduction of Video Service Quality
DNN	Deep Neural Network
GAN	Generative Adversarial Networks
RGB	Red Green Blue
VICOM	VICOM Camera
HARM	Human-AI-Risk-Mitigation
UNESCO	United Nations Educational, Scientific and Cultural Organization

01 약어표

ALTAI	The Assessment List for Trustworthy Artificial Intelligence
ISO/IEC	The International Organization for Standardization/International Electrotechnical Commission
PCA	Principal Component Analysis
LDA	Linear Discriminant Analysis
ROS	Random Over Sampling
SMOTE	Synthetic Minority Over-sampling Technique
ADASYNA댑티브	Adaptive Synthetic Sampling Approach
SVM	Support Vector Machine
HSEmotion	Face-emotion-recognition (A Python Library)
RSF	Russian Science Foundation
HSE	HSE University
OSI	Open Source Initiative
OWASP	Open Web Application Security Project
NVD	National Vulnerability Database
CVE	Common Vulnerabilities and Exposures
DoS	Denial of Service
DACOBS	Davos Assessment of Cognitive Biases Scale
VAE	Variational Autoencoder
BDPL	Boundary Differentially Private Layer
XAI	eXplainable AI
NLP	Natural Language Processing
IOU	Intersection Over Union
mAP	Mean Average Precision
DVC	Data Version Control
PIPA	The Personal Information Protection Act
HR	Human Resources
DPIA	Data Protection Impact Assessment
EU	The European Union
OECD	The Organization for Economic Co-operation and Development

02 용어표

용어표

용어명	정의
GAN Generative Adversarial Networks	생성적 적대 신경망은 비지도 학습에 사용되는 인공지능 알고리즘으로, 제로섬 게임 틀 안에서 서로 경쟁하는 두 개의 신경 네트워크 시스템에 의해 구현된다. 2014년에 이안 굿펠로우가 발표한 개념이다.
GDPR General Data Protection Regulation	GDPR(일반 데이터 보호 규정)은 세계에서 가장 엄격한 개인정보 보호 및 보안법이다. 데이터 최소화, 공정성 및 투명성, 설명 가능성, 무결성, 목적 제한, 책임성의 원칙을 요약하여 개인 데이터의 사용 또는 오용으로 인한 개인에 대한 잠재적 피해 위험을 최소화한다.
HMI Human Machine Interface	기계 사용을 용이하게 하고 대화형 디지털 시스템을 설계 및 개발하는 것을 목표로 하는 기술이자 나아가 미래의 세상을 생각할 수 있는 모델, 컴퓨팅 장치 및 새로운 상호작용 기술을 개발하는 것이다. 키보드, 마우스, 창, 아이콘, 메뉴 등의 패러다임을 넘어서는 것을 목표로 한다.
TensorFlow Data Validation	TFDV(TensorFlow Data Validation)는 머신러닝 데이터를 탐색하고 검증하기 위한 라이브러리이다. TF 데이터 검증에는 다음이 포함된다. ①훈련 및 테스트 데이터의 요약 통계를 확장 가능하게 계산, ②데이터 분포 및 통계를 위한 뷰어와 통합은 물론 특징 쌍의 측면 비교(패싯), ③필수 값, 범위, 어휘 등 데이터에 대한 기대치를 설명하는 자동화된 데이터 스키마 생성, ④스키마를 검사하는 데 도움이 되는 스키마 뷰어, ⑤누락된 특징, 범위를 벗어난 값 또는 잘못된 특징 유형 같은 이상 현상을 식별하는 이상 탐지, ⑥어떤 기능에 이상이 있는지 확인하고 이를 수정하기 위해 자세히 알아볼 수 있는 이상 뷰어이다.
VAE Variational Auto-Encoder	VAE는 Input image X를 잘 설명하는 feature를 추출하여 Latent vector z에 담고, 이 Latent vector z를 통해 X와 유사하지만 완전히 새로운 데이터를 생성하는 것을 목표로 한다.
Z-스코어 Z-Score	각 값이 평균에서 얼마나 떨어져 있는지를 나타내는 척도이다.
거버넌스 Governance	거버넌스는 조직이 통제되고 운영되는 시스템은 물론, 조직과 그 구성원들이 책임을 지는 메커니즘까지 포괄한다. 윤리, 위험 관리, 규정 준수 및 관리가 모두 거버넌스의 요소이다.

02 용어표

용어명	정의
고위험시스템 High Risk system	EU에서는 AI 시스템이 어떤 결정이나 예측을 내리고 특정 조치를 취한 이유를 알 수 없는 경우가 많다고 언급한다. 따라서 채용 결정이나 공익 제도 신청 등에서 누군가가 부당하게 불이익을 받았는지 평가하기 어려워질 수 있다고 설명한다. 위험군은 ①허용할 수 없는 위험, ②고위험, ③제한된 위험, ④위험 최소화 또는 위험 없음으로 분류된다. ①은 정부의 사회적 점수 매김부터 음성 지원을 사용하여 위험한 행동을 조장하는 장난감에 이르기까지, 사람들의 안전, 생계, 권리에 대한 명백한 위협으로 간주하는 모든 AI 시스템을 금지하는 것이다. ②는 시민의 생명과 건강을 위험에 빠뜨릴 수 있는 것을 의미한다. 채용이 고위험에 해당하며, 이외에 주요한 고위험군은 다음과 같다. • 시민의 생명과 건강을 위험에 빠뜨릴 수 있는 중요 인프라 • 교육 또는 직업 훈련: 교육에 대한 접근성과 누군가의 직업 과정을 결정할 수 있을 때 • 제품의 안전 구성 요소 • 고용, 근로자 관리 및 자영업에 대한 접근성 • 필수적인 민간 및 공공 서비스 • 국민의 기본권을 방해할 수 있는 법 집행 • 이주, 망명, 국경 통제 관리 사법 및 민주적 절차의 관리. 고위험 시스템은 위에 언급한 부분을 포함한 시스템을 지칭하며, 시장에 출시되기 전에 엄격한 의무를 준수해야 한다.
기계학습 Machine Learning	머신러닝(ML)은 인간 프로그래머가 알고리즘을 개발하는 데 큰 비용이 드는 문제를 해결하기 위한 포괄적인 용어이다. 인간이 개발한 알고리즘에 명시적으로 지시할 필요 없이 기계가 스스로 알고리즘을 발견하도록 지원함으로써 문제를 해결한다.
기술 통계량 계산 Descriptive Statistic	대규모 모집단에 대한 일반화나 추론 없이 데이터 집합의 주요 특징과 특성을 설명하고 분석하는 데 중점을 두고 계산하는 것을 의미한다.
멀티모달 Multi Modal	멀티모달 AI는 텍스트, 이미지, 영상, 음성 등 다양한 데이터 모달리티를 함께 고려하여 서로의 관계성을 학습 및 표현하는 기술이다.
단순 무작위 샘플링 Simple random sampling	모집단(population)의 각각의 요소 또는 사례가 표본(sample)으로 선택될 가능성이 같게 되는 표본 추출법이다.
데이터 아웃라이어 Data Outlier	보통 관측된 데이터의 범위에서 많이 벗어난 매우 작거나 큰 값을 말한다.
데이터 중독 Data Poisoning	학습된 모델의 예측 동작을 제어하기 위해 학습 데이터셋을 조작하여 모델이 악성 예시를 원하는 클래스로 분류하도록 하는 공격이다.

02 용어표

용어명	정의
데이터셋 Data set	데이터의 모음을 의미한다.
라벨링 Labeling	머신러닝에서 데이터 라벨링은 원시 데이터(이미지, 텍스트 파일, 동영상 등)를 식별하고 머신러닝 모델이 학습할 수 있도록 의미 있고 유익한 라벨을 하나 이상 추가하여 컨텍스트를 제공하는 프로세스이다.
메타데이터 Metadata	다른 데이터에 대한 정보를 제공하는 데이터이다.
채용 인공지능	면접을 진행할 때 도움을 주는 AI이다.
AI 블랙박스 AI Black box	사용자에게 보이지 않는 내부 작동 방식을 가진 AI 시스템을 말한다. 데이터를 입력하고 출력을 얻을 수는 있지만, 시스템의 코드나 출력을 생성한 로직을 검사할 수는 없다.
사분위 범위 Interquartile range	통계적 분산을 나타내는 척도로, 데이터의 분포를 보여준다.
선형 판별 분석(LDA) Linear Discriminant Analysis	FDA 또는 Linear Discriminant Analysis(LDA)라고 불린다. 데이터들을 하나의 직선(1차원 공간)에 프로젝션한 후, 이 데이터들이 잘 구분이 되는가를 판단하는 방법이다.
설명 가능 인공지능 eXplainable Artificial Intelligence	판단의 이유를 사람이 이해할 수 있는 방식으로 제시하는 인공지능을 일컫는다. 특정한 판단에 대해 알고리즘의 설계자조차 그 이유를 설명할 수 없는 '블랙박스' 인공지능과 대비되는 개념이다. 불확실성을 해소하여 인공지능에 대한 신뢰성을 높일 수 있다.
스키마 추론 Schema inference	메타데이터 직렬화 형식으로 TensorFlow 도구에 속하는 표 형식 데이터(예: tensorFlow 예제)를 설명하는 스키마이다.
신뢰할 수 있는 인공지능 Trustworthy AI	신뢰할 수 있는 AI는 이력서 심사를 자동화하고 후보자 면접에서 객관적으로 자격을 평가함으로써 채용을 간소화하고 효율성을 높이며 편견을 줄인다. 이는 공정성과 신뢰성을 보장하여 채용에서 책임 있고 신뢰할 수 있는 AI 생태계를 조성하는 데 중요하다. 채용에 신뢰할 수 있는 AI를 도입하면 기존 채용 관행에서 의사결정에 영향을 미칠 수 있는 무의식적 편견과 관련된 오랜 문제를 해결할 수 있다. 자동화된 면접 평가 및 지원자 평가 알고리즘은 대량의 지원서를 편견 없이, 신속하게 선별하여 지원자의 관련 자격과 경험에만 집중한다. 이를 통해 채용 일정을 단축하고 후보자 프로필을 더욱 철저하게 분석할 수 있다. 또한, 지원자 면접 시 표준화된 평가 프로세스를 촉진한다. 기술, 역량, 자격을 객관적으로 평가하여 모든 지원자에게 공평한 경쟁의 장을 조성하는 데 기여한다. 이러한 객관성은 개인적 특성에 관계없이 모든 지원자를 동일한 기준에 따라 평가하기 때문에 채용 프로세스의 공정성과 신뢰성이 높아진다. 책임 있고 신뢰할 수 있는 AI 에코시스템 맥락에서 신뢰할 수 있는 AI는 채용의 중요한 자산이 된다.

02 용어표

용어명	정의
어노테이터 Annotator	인공지능이 학습하기 위해 필요한 다양한 데이터를 수집, 입력 및 관리하는 일을 의미한다.
언더샘플링 Undersampling	다수 클래스의 데이터를 제거하는 방법으로, 소수 클래스와의 비율을 맞추는 기법이다.
오버샘플링 Oversampling	소수 클래스의 데이터를 늘리는 방법으로, 다수 클래스와의 비율을 맞추는 기법이다.
오픈소스 Opensource	설계방식이 공개되어 있어 사람들이 수정하고 공유할 수 있는 것을 의미한다.
이니셔티브 Initiative	'문제를 해결하기 위한 계획이라는 의미로 자주 사용되면 계획, 새로운 사업 구상에 가까운 말이다. EU는 신뢰할 수 있는 AI를 구축하는 데 기여할 3가지 상호 관련된 법적 이니셔티브를 제안하였다. ①AI 시스템과 관련된 기본권 및 안전 위험을 해결하기 위한 법적 프레임워크, ②민사 책임 프레임워크, ③부문별 안전 법규 개정이다. 반면 미국은 ①AI 연구 인프라 강화, ②AI R&D 우선순위 지정, ③신뢰할 수 있는 AI 발전, ④정부 및 국가 안보를 위한 AI 활용, ⑤국제 AI 참여 촉진, ⑥AI 준비된 인력 양성을 제안하였다.
인공지능 생명주기 AI Life cycle	인공지능 생명주기는 비즈니스 문제에서 해당 문제를 해결하는 AI 솔루션으로 이동하는 반복적인 프로세스이다. 생명주기의 각 단계는 설계, 개발 및 배포 단계를 여러 번 반복한다.
주성분 분석(PCA) Principal component analysis	고차원의 데이터를 저차원의 데이터로 환원하는 기법이다.
채용 인공지능	인사과정 진행의 전반적인 부분에 도움을 주는 AI를 뜻한다.
체계적 샘플링 Systematic sampling	모집단에서 매 K번째를 연구대상으로 선정하는 방법이다.
층화 샘플링 Stratified sampling	모집단을 먼저 중복되지 않도록 층으로 나눈 후, 각 층에서 표본을 추출하는 방법이다.
카이제곱 테스트 Chi-squared test	관찰된 빈도가 기대되는 빈도와 의미 있게 다른지를 검정하기 위해 사용되는 검정방법이다.
클러스터 샘플링 Cluster Sampling	모집단에서 집단을 일차적으로 표집한 후, 선정된 각 집단에서 구성원을 표본으로 추출하는 다단계 표집 방법이다.
테스트 오라클 Test Oracle	컴퓨팅, 소프트웨어 엔지니어링 및 소프트웨어 테스트에서 테스트의 합격 또는 불합격 여부를 결정하는 메커니즘이다. 오라클을 사용하려면, 주어진 테스트 사례 입력에 대해 테스트 중인 시스템의 출력을 해당 제품이 가져야 한다고 오라클이 판단한 출력과 비교해야 한다.

용어명	정의
투명성 Transparency	투명성은 여러 방식으로 정의할 수 있다. 인접한 개념으로는 '설명 가능성', '해석 가능성', '이해 가능성', '블랙박스' 등이 있다.
편향 Bias	알고리즘 편향 또는 AI 편향이라고도 하는 머신러닝 편향은 머신러닝(ML) 프로세스에서 잘못된 가정으로 인해 알고리즘이 체계적으로 편향된 결과를 생성할 때 발생하는 현상이다. 인공지능(AI)의 하위 집합인 머신러닝의 성능은 학습 데이터의 품질, 객관성 및 크기에 따라 달라진다. 결함이 있거나 불량하거나 불완전한 데이터는 부정확한 예측을 초래한다. 이는 컴퓨터 과학에서 입력의 품질에 따라 출력의 품질이 결정된다는 개념을 전달하기 위해 사용되는 가비지 인(Garbage in), 가비지 아웃(Garbage out)의 교훈을 반영한 것이다. 머신러닝 편향은 일반적으로 머신러닝 시스템을 설계하고 훈련하는 개인이 도입한 문제에서 비롯된다. 이러한 사람들이 의도하지 않은 인지적 편향이나 실제 편향을 반영하는 알고리즘을 만들 수 있다. 또는 불완전하거나 결함이 있거나 편향이 있는 데이터셋을 사용하여 머신러닝 시스템을 학습하고 검증하는 과정에서도 편향이 생길 수 있다.
합성 데이터 Synthetic Data	실제 사건에 의해 생성된 데이터가 아닌 컴퓨터 프로그램을 통해 인위적으로 생성된 데이터이다.
학습 데이터 Training data	학습 데이터는 AI 모델이나 머신러닝 알고리즘이 올바른 결정을 내릴 수 있도록 가르치는 데 사용되는 라벨이 지정된 데이터이다.

03 요구사항별 이해관계자

요구사항별 이해관계자

관련 표준*에 근거한 요구사항별 이해관계자

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

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

04 이해관계자 정의

이해관계자 정의

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

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

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

05 참고문헌

참고 문헌

[1] Chea Yun Jung, Hyun Kyong Joo, "**Post-'Lee-Luda' personal information protection in Korea: developer responsibility and autonomous AI governance,**" International Data Privacy Law, vol. 13 issue 2, pp. 154-167, 2023. 05.https://doi.org/10.1093/idpl/ipad006

[2] Keelan Balderson**, 35 AI Recruitment Statistics for Employers and Candidates,** [Online], Available: https://mspoweruser.com/ai-recruitment-statistics/

[3] RR Author, **68% Are Aware of AI's Replacement in the Job Market**, [Online], Available: https://realresearcher.com/media/68-percent-are-aware-of-ais-replacement-in-the-job-market/

[4] I. Tewari and M. Pant, "**Artificial Intelligence Reshaping Human Resource Management: A Review,** " 2020 IEEE International Conference on Advent Trends in Multidisciplinary Research and Innovation (ICATMRI), pp. 1-4, 2020. doi: 10.1109/ICATMRI51801.2020.9398420.

[5] Korean LII, **AI job interview**, [Online], Available: http://www.koreanlii.or.kr/w/index.php/AI_job_interview

[6] Kang Jae-eun, **[Feature] AI job interviews are booming, but doubts linger**, [Online], Available: https://www.koreaherald.com/view.php?ud=20211004000214

[7] 홍은지/정다은, **AI Job Interviews: Robots That Judge Humans**, [Online], Available: https://skt.skku.edu/news/articleView.html?idxno=1254

[8] Yonhab News Agency, **Lotte to evaluate cover letters with AI system**, [Online], Available: https://en.yna.co.kr/view/AEN20180319003600320

[9] Park Soong-joo, **[Newsmaker] Time to take the 'human' out of HR?**, [Online], Available: https://www.koreaherald.com/view.php?ud=20230217000186

[10] 김지섭, **"청소 경험 없으니 탈락입니다" AI 면접관의 황당한 실수**, [Online], Available: https://www.chosun.com/economy/weeklybiz/2021/09/17/TPZTJBPGLRFXXGJDMP3YOIO3YQ/

[11] Genesis Lab, **Research Areas**, [Online], Available: https://home.genesislab.ai/research-area

[12] Midas, **AI Competency Test**, [Online], Available: https://www.midashri.com/aicc

[13] Chamorro-Premuzic, T., Polli, F., Dattner, B. "**Building ethical AI for talent management.**" Harvard Business Review, 2019. 11

[14] EU Commission, **When is a Data Protection Impact Assessment (DPIA) required?**, [Online], Available: https://commission.europa.eu/law/law-topic/data-protection/reform/rules-business-and-organisations/obligations/when-data-protection-impact-assessment-dpia-required_en

[15] European Parliament, **EU AI Act: first regulation on artificial intelligence**, [Online], Available: https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-firs

t-regulation-on-artificial-intelligence

[16] Korean Civil Society Organizations, **Input for Report on right to privacy in the digital age,** [Online], Available: https://www.ohchr.org/Documents/Issues/DigitalAge/Submissions/CSOs/ROK-CSOs.pdf

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

[18] Tim McGarr, **ISO/IEC 23894 - A new standard for risk management of AI**, [Online], Available: https://aistandardshub.org/a-new-standard-for-ai-risk-management

[19] The White House, **Blueprint for an AI Bill of Rights | OSTP | The White House**, [Online], Available: https://www.whitehouse.gov/ostp/ai-bill-of-rights/

[20] European Commission, **The Act | The Artificial Intelligence Act**, [Online], Available: https://artificialintelligenceact.eu/the-act/

[21] General Data Protection Regulation (GDPR), **Data protection impact assessment**, [Online], Available: https://gdpr-info.eu/art-35-gdpr/

[22] The National Institute of Standards and Technology (NIST), **AI Risk Management Framework**, [Online], Available: https://www.nist.gov/itl/ai-risk-management-framework

[23] The International Organization for Standardization (ISO), **ISO/IEC 31010:2009, Risk management — Risk assessment techniques**, [Online], Available: https://www.iso.org/standard/51073.html

[24] The International Organization for Standardization (ISO), **ISO 31000:2018, Risk management — Guidelines**, [Online], Available: https://www.iso.org/standard/65694.html

[25] The International Organization for Standardization (ISO), **ISO/IEC 27005:2018: Information technology — Security techniques — Information security risk management**, [Online], Available: https://www.iso.org/standard/75281.html

[26] The International Organization for Standardization (ISO), **ISO/IEC 24028 :2020: Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence**, [Online], Available: https://www.iso.org/standard/77608.html

[27] Joshua L Martin, Kelly Elizabeth Wright, "**Bias in Automatic Speech Recognition: The Case of African American Language,**" Applied Linguistics, vol. 44, Issue 4, pp 613-630, 2023. 8. https://doi.org/10.1093/applin/amac066

[28] Bîgu, D., Cernea, M.-V., "**Algorithmic bias in current hiring practices: An ethical examination,**" In 13th international management conference (IMC) on management strategies for high Performance, 2019. 10.

[29] North-Samardzic, A., "**Biometric Technology and Ethics: Beyond Security Applications**." J Bus Ethics, vol. 167, pp. 433-450, 2020. https://doi.org/10.1007/s10551-019-04143-6

05 참고문헌

[30] Sánchez-Monedero, J., Dencik, L., Edwards, L., "**What does it mean to 'solve'the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiri ng systems**." In Conference on Fairness, Accountability, and Transparency, pp. 458-468, 2020. 01. https://doi.org/10.1145/3351095.3372849.

[31] Yarger, L., Cobb Payton, F., Neupane, B., "**Algorithmic equity in the hiring of underrepresented IT job candidates.**" Online Information Review, vol. 44, pp. 383-395, 2020. https://doi.org/10.1 108/OIR-10-2018-0334

[32] Center for Development of Security Excellence, **Risk Management for DoD Security Programs Student Guide**, [Online], Available: https://www.cdse.edu/Portals/124/Documents/student-g uides/GS102-guide.pdf

[33] The National Institute of Standards and Technology (NIST), **Guide for Conducting Risk Assess ments**, [Online], Available: https://nvlpubs.nist.gov/nistpubs/legacy/sp/nistspecialpublication 800-30r1.pdf

[34] The University of Vermont, **Guide to Risk Assessment and Response**, [Online], Available: http s://www.uvm.edu/sites/default/files/UVM-Risk-Management-and-Safety/Guide_to_Risk_Op portunity_Assessment_Response.pdf

[35] Team Asana, **What is a risk register: a project manager's guide (and example)**, [Online], Availa ble: https://asana.com/resources/risk-register

[36] Nick Shah, **How to Prepare for a Job Interview Run by AI**, [Online], Available: https://builtin.co m/job-interviews/ai-job-interviews

[37] Fulk, H. Kevin, Heidi L. Dent, William A. Kapakos, and Barbara Jo White, "**Doing more with less: Using AI-based Big Interview to combine exam preparation and interview practice.**" Issues in Information Systems, vol. 23, no. 4, pp. 204-217, 2022. [Online], Available: https://www.wcu. edu/_files/academic-enrichment/CCPD_BigInterview_Article.pdf

[38] The National Institute of Standards and Technology (NIST), **AI Risk Management Framework Co ncept Paper**, [Online], Available: https://www.nist.gov/system/files/documents/2021/12/14/ AI%20RMF%20Concept%20Paper_13Dec2021_posted.pdf

[39] The National Institute of Standards and Technology (NIST), **AI RMF Playbook**, [Online], Available: https://pages.nist.gov/AIRMF/

[40] I. Nitta, K. Ohashi, S. Shiga and S. Onodera, "**AI Ethics Impact Assessment based on Requirem ent Engineering,**" 2022 IEEE 30th International Requirements Engineering Conference Worksh ops (REW), pp. 152-161, 2022 doi:10.1109/REW56159.2022.00037

[41] Hunkenschroer, Anna Lena, and Christoph Luetge. "**Ethics of AI-enabled recruiting and selecti on: A review and research agenda.**" Journal of Business Ethics, vol. 178, no. 4, pp. 977-1007, 2022. https://link.springer.com/article/10.1007/s10551-022-05049-6/tables/3

[42] Acikgoz, Y., Davison, K. H., Compagnone, M., Laske, M., "**Justice perceptions of artificial intelligence in selection.**" International Journal of Selection and Assessment, vol. 28, pp. 399-416, 2020. https://doi.org/10.1111/ijsa.12306

[43] Raghavan, M., Barocas, S., Kleinberg, J., Levy, K. "**Mitigating bias in algorithmic hiring: Evaluating claims and practices.**" In Conference on fairness, accountability, and transparency, 2020. 01.

[44] Chamorro-Premuzic, T., Polli, F., Dattner, B. "**Building ethical AI for talent management.**" Harvard Business Review, 2019. 11.

[45] Government of Canada, **Algorithmic Impact Assessment tool**, [Online], Available: https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

[46] Daniel S. Schiff, Aladdin Ayesh, Laura Musikanski, John C. Havens, "**IEEE 7010: A New Standard for Assessing the Well-being Implications of Artificial Intelligence.**" In 2020 IEEE international conference on systems, man, and cybernetics (SMC), pp. 2746-2753, 2020.

[47] The International Organization for Standardization (ISO), **ISO/IEC 38500:2015 Information technology — Governance of IT for the organization**, [Online], Available: https://www.iso.org/standard/62816.html

[48] Damon W. Silver of Jackson Lewis P.C., "**5 Key Data Privacy and Security Risks That Arise When Organizations Record Job Interviews & Strategies for Mitigating Them.**" National Law Review, vol. XI, no. 103, 2021.https://www.natlawreview.com/article/5-key-data-privacy-and-security-risks-arise-when-organizations-record-job-interviews

[49] IBM, **Example: Disaster recovery plan**, [Online], Available: https://www.ibm.com/docs/en/i/7.3?topic=system-example-disaster-recovery-plan

[50] Sergiu Gatlan, **FBI: Stolen PII and deepfakes used to apply for remote tech jobs**, [Online], Available: https://www.bleepingcomputer.com/news/security/fbi-stolen-pii-and-deepfakes-used-to-apply-for-remote-tech-jobs/

[51] Dattner, B., Chamorro-Premuzic, T., Buchband, R., Schettler, L. "**The legal and ethical implications of using AI in hiring**." Harvard Business Review, 2019. 4.

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

[53] The World Economic Forum (WEF), **Model Artificial Intelligence Governance Framework and Assessment Guide,** [Online], Available: https://www.weforum.org/projects/model-ai-governance-framework

[54] Personal Data Protection Commission, **Model Artificial Intelligence Governance Framework**,Singapore, 2020. 01., [Online], Available: https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Fil

es/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf

[55] Artificial Intelligence Video Interview Act (AIVI Act), [Online], Available: https://www.ilga.gov/legislation/ilcs/ilcs3.asp?ActID=4015&ChapterID=68

[56] Code of Federal Regulations, **PART 1607—Uniform Guidelines On Employee Selection Procedures (1978)**, [Online], Available: https://www.govinfo.gov/content/pkg/CFR-2018-title29-vol4/xml/CFR-2018-title29-vol4-part1607.xml

[57] Government of Canada, **Artificial Intelligence and Data Act**, [Online], Available: https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act

[58]] EU Commission, **Ethical Guidelines for Trustworthy AI**, [Online], Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[59] The Partnership on AI, [Online], Available: https://partnershiponai.org/

[60] Personal Information Protection Commission (PIPC), **The Personal Information Protection Act**, [Online], Available: https://www.pipc.go.kr/eng/user/cmm/privacyGuideline.do

[61] 입법예고, **[2122205] 채용절차의 공정화에 관한 법률 전부개정법률안(윤재옥의원 등 114인)**, [Online], Available: https://pal.assembly.go.kr/napal/lgsltpa/lgsltpaOngoing/view.do?lgsltPaId=PRC_T2B3A0S5W2J2B1P7Q0S7R2T9L6H1N4

[62] Perkinscoie, **How AI and Automated Systems Use Can Lead to Discrimination in Hiring**, [Online], Available: https://www.perkinscoie.com/en/news-insights/how-ai-and-automated-systems-use-can-lead-to-discrimination-in-hiring.html

[63] General Data Protection Regulation GDPR, **Automated individual decision-making, including profiling**, [Online], Available: https://gdpr-info.eu/art-22-gdpr/

[64] 5 ways to address regulations around AI-enabled hiring and employment, [Online], Available: https://venturebeat.com/ai/employment-ai-regulations-5-takeaways-for-technical-decision-makers/

[65] Jayatilleke, Buddhi., "**Towards Establishing Fairness in AI Based Candidate Screening.**", 2022. Doi: 10.13140/RG.2.2.17622.73288.

[66] The World Economic Forum, '**Model AI Governance Framework**', 2020, [Online], Available: https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf

[67] Microsoft, **The Microsoft Responsible AI Standard**, [Online], Available: https://www.microsoft.com/en-us/ai/our-approach?activetab=pivot1:primaryr5

[68] Fairlearn, **Fairness in Machine Learning**, [Online], Available: https://fairlearn.org/main/user_guide/fairness_in_machine_learning.html#fairness-of-ai-systems

[69] Microsoft AI, **Putting principles into practice: How we approach responsible AI at Microsoft**, [O

nline], Available: https://www.microsoft.com/cms/api/am/binary/RE4pKH5

[70] The International Organization for Standardization (ISO), **ISO/IEC 38507:2022 Information technology — Governance of IT —Governance implications of the use of artificial intelligence by organizations**, [Online], Available: https://www.iso.org/standard/56641.html

[71] The World Economic Forum, **Responsible Use of Technology: The IBM Case Study**, [Online], Available: https://www3.weforum.org/docs/WEF_Responsible_Use_of_Technology_The_IBM_Case_Study_2021.pdf

[72] IEEE, **Ethically Aligned Design**, [Online], Available: https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

[73] The International Organization for Standardization (ISO), **ISO/IEC 23894:2023 Information technology — Artificial intelligence —Guidance on risk management**, [Online], Available: https://www.iso.org/standard/77304.html

[74] the European Observatory of Working Life (EurWORK), **Artificial intelligence**, 2022, https://www.eurofound.europa.eu/observatories/eurwork/industrial-relations-dictionary/artificial-intelligence

[75] R. Abbas, Z. Sultan and S. N. Bhatti, "**Comparative analysis of automated load testing tools: Apache JMeter, Microsoft Visual Studio (TFS), LoadRunner, Siege,**" 2017 International Conference on Communication Technologies (ComTech), Rawalpindi, Pakistan, 2017, pp. 39-44, doi: 10.1109/COMTECH.2017.8065747.

[76] A. Holmes and M. Kellogg, "**Automating functional tests using Selenium,**" AGILE 2006 (AGILE'06), Minneapolis, MN, USA, pp. 6 pp.-275, 2006, doi: 10.1109/AGILE.2006.19.

[77] Github, **Jmeter**, [Online], Available: https://github.com/apache/jmeter

[78] N. Li, A. Escalona and T. Kamal, "**Skyfire: Model-Based Testing with Cucumber,**" 2016 IEEE International Conference on Software Testing, Verification and Validation (ICST), Chicago, IL, USA, pp. 393-400, 2016, doi: 10.1109/ICST.2016.41.

[79] Ai, Hua, and Fuliang Weng. "**User simulation as testing for spoken dialog systems.**" In Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue, pp. 164-171. 2008.

[80] Krumhuber, Eva G., Antony S. R. Manstead, Darren P. Cosker, Dave Marshall and Paul L. Rosin. "**Effects of Dynamic Attributes of Smiles in Human and Synthetic Faces: A Simulated Job Interview Setting.**"Journal of Nonverbal Behavior, vol. 33, pp. 1-15, 2009

[81] I. Stanica, M. -I. Dascalu, C. N. Bodea and A. D. Bogdan Moldoveanu, "**VR Job Interview Simulator: Where Virtual Reality Meets Artificial Intelligence for Education,**" 2018 Zooming Innovation in Consumer Technologies Conference (ZINC), pp. 9-12, 2018. doi: 10.1109/ZINC.2018.8448645.

05 참고문헌

[82] Busso, Carlos, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan, "**IEMOCAP: Interactive emotional dyadic motion capture database.**" Language resources and evaluation, vol. 42, pp. 335-359, 2008.

[83] Naim, Iftekhar, M. Iftekhar Tanveer, Daniel Gildea, and Mohammed Ehsan Hoque. "**Automated prediction and analysis of job interview performance: The role of what you say and how you say it.**" In 2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG), vol. 1, pp. 1-6, 2015. https://www.cs.rochester.edu/~gildea/pubs/naim-fg15.pdf

[84] Kaur, Davinder, Suleyman Uslu, Kaley J. Rittichier, and Arjan Durresi. "**Trustworthy artificial intelligence: a review.**" ACM Computing Surveys (CSUR), vol. 55 no. 2, pp. 1-38, 2022. https://cs.iupui.edu/~adurresi/papers/kaur2022trustworthy.pdf

[85] MIT Interview Dataset, **Automated Prediction of Job Interview Performances**, [Online], Available: https://roc-hci.com/past-projects/automated-prediction-of-job-interview-performances/

[86] Computer Vision Center and University of Barcelona, **First Impressions V2 Dataset**, [Online], Available: https://chalearnlap.cvc.uab.cat/dataset/24/description/

[87] Pingitore, Regina & Dugoni, Bernard, Tindale, Scott, Spring, B., "**Bias Against Overweight Job Applicants in a Simulated Employment Interview.**" The Journal of applied psychology, vol. 79, pp. 909-17, 1995. Doi: 10.1037/0021-9010.79.6.909.

[88] A. Holmes and M. Kellogg, "**Automating functional tests using Selenium,**" AGILE 2006 (AGILE'06), pp. 6-275, 2006. doi: 10.1109/AGILE.2006.19.

[89] GazeRecorder, **Online Eye Tracking Software**, [Online], Available: https://gazerecorder.com/

[90] Tallinn University, **Interview Simulator**, [Online], Available: http://www.tlu.ee/~plaupa/interviewsimulator/

[91] Pixo, **Interview Simulator**, [Online], Available: https://pixovr.com/vr-training-content/interview-simulation/

[92] Genesis Lab, "뷰인터등급표 NEW version VIEWINTER HR", White Paper, 2023.

[93] Hough, Leatta M. "**The'Big Five'personality variables--construct confusion: Description versus prediction.**" Human performance, vol. 5 no. 1-2, pp. 139-155, 1992.

[94] Raymark, Patrick H., Mark J. Schmit, and Robert M. Guion. "**Identifying potentially useful personality constructs for employee selection.**" Personnel Psychology, vol. 50 no. 3, pp. 723-736, 1997.

[95] Pulakos, Å. "**Selection Assessment Methods. A guide to implementing formal assessments to build a high-quality workforce,**" SHRM Foundation, 2005. https://www.shrm.org/hr-today/tren

ds-and-forecasting/special-reports-and-expert-views/documents/selection-assessment-methods.pdf

[96] Singhal, Astha, Mohammad Rafayet Ali, Raiyan Abdul Baten, Chigusa Kurumada, Elizabeth West Marvin, and Mohammed E.H., "**Analyzing the impact of gender on the automation of feedback for public speaking.**" In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 607-613, 2018.

[97] Barrick, Murray R., and Michael K. Mount. "**The big five personality dimensions and job performance: a meta-analysis.**" Personnel psychology, vol. 44 no. 1, pp. 1-26, 1991.

[98] Suen, HY., Hung, KE., Lin, CL. "**Intelligent video interview agent used to predict communication skill and perceived personality traits.**" Hum. Cent. Comput. Inf. Sci., vol. 10 no. 3, 2020. https://doi.org/10.1186/s13673-020-0208-3

[99] Alufaisan, Yasmeen, Laura R. Marusich, Jonathan Z. Bakdash, Yan Zhou, and Murat K., "**Does Explainable Artificial Intelligence Improve Human Decision-making?,**" PsyArXiv. 2020. 06. doi:10.31234/osf.io/d4r9t

[100] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia B., "**Evaluating saliency map explanations for convolutional neural networks: a user study.**"In Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI '20). Association for Computing Machinery, pp. 275-285, 2020. https://doi.org/10.1145/3377325.3377519

[101] Information Technology Industry Council, **AI Policy Principles**, [Online], Available: https://www.itic.org/resources/AI-Policy-Principles-FullReport2.pdf

[102] Palmero, Cristina, et al. "**Context-aware personality inference in dyadic scenarios: Introducing the UDIVA dataset.**" In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1-12. 2021.

[103] PIPC, **고용상 연령차별금지 및 고령자고용촉진에 관한 법률**, [Online], Available: https://www.law.go.kr/%EB%B2%95%EB%A0%B9/%EA%B3%A0%EC%9A%A9%EC%83%81%EC%97%B0%EB%A0%B9%EC%B0%A8%EB%B3%84%EA%B8%88%EC%A7%80%EB%B0%8F%EA%B3%A0%EB%A0%B9%EC%9E%90%EA%B3%A0%EC%9A%A9%EC%B4%89%EC%A7%84%EC%97%90%EA%B4%80%ED%95%9C%EB%B2%95%EB%A5%A0

[104] PIPC, **고용상 연령차별금지 및 고령자고용촉진에 관한 법률**, [Online], Available: https://www.law.go.kr/법령/고용상연령차별금지및고령자고용촉진에관한법률/(20220610,18921,20220610)/제4조의4

[105] PIPC, **고용상 연령차별금지 및 고령자고용촉진에 관한 법률**, [Online], Available: https://www.law.go.kr/법령/고용상연령차별금지및고령자고용촉진에관한법률/(20220610,18921,20220610)/제4조의5

[106] Mujtaba, D. F., and Mahapatra, N. R., "**Ethical considerations in AI-based recruitment,**"In M. Cunningham & P. Cunningham (Eds.), IEEE International Symposium on Technology in Society, pp. 1-7, 2019. 11. https://doi.org/10.1109/ISTAS48451.2019.8937920.

05 참고문헌

[107] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., "**Fairness through awareness**," In 3rd conference on innovations in theoretical computer science, pp. 214-226, 2012. 01.

[108] Hardt, M., Price, E., and Srebro, N., "**Equality of opportunity in supervised learning,**"In 30th conference on neural information processing systems (NIPS), 2016. 12.

[109] Kim, P. T., "**Data-driven discrimination at work.**" William & Mary Law Review, vol. 58 no. 3, pp. 857-937, 2017.

[110] Ben Dattner, TomasChamorro-Premuzic, Richard Buchband, and Lucinda S., **The Legal and Ethical Implications of Using AI in Hiring**, [Online], Available: https://hbr.org/2019/04/the-legal-and-ethical-implications-of-using-ai-in-hiring

[111] MIT Technology Review, **We tested AI interview tools. Here's what we found**, [Online], Available: https://www.technologyreview.com/2021/07/07/1027916/we-tested-ai-interview-tools/

[112] Reuters, **Amazon scraps secret AI recruiting tool that showed bias against women**, [Online], Available: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

[113] AI Hub, **[Advanced Data Labeler_Image Image] Video Labeling techniques and applications of data, chapter 01,** [Online], Available: https://www.aihub.or.kr/aihubdata/edccrse/view.do?currMenu=136&topMenu=103&pageIndex=1&edcCrseSn=51&thisYear=2021&edcSeCode=EDUCATE003&searchCondition=&searchKeyword=

[114] Kaggle, **Resume Dataset**, [Online], Available: https://www.kaggle.com/datasets/gauravduttakiit/resume-dataset

[115] Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, Cristian Danescu-Niculescu-Mizil, "**ConvoKit: A Toolkit for the Analysis of Conversations**,". Proceedings of SIGDIAL, 2020.

[116] Kaggle, **Dataset**, [Online], Available: https://www.kaggle.com/datasets/kyanyoga/sample-sales-data

[117] Kaggle, **Employee Churn Prediction Determine what factors predict an employee leaving his/her job.**, [Online], Available: https://www.kaggle.com/c/employee-churn-prediction

[118] Kaggle, **Dataset**, [Online], Available: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

[119] Kaggle, **Dataset**, [Online], Available: https://www.kaggle.com/datasets/tunguz/big-five-personality-test

[120] Kaggle, **Dataset**, [Online], Available: https://www.kaggle.com/datasets/zeyadkhalid/mbti-personality-types-500-dataset

[121] Kaggle, **Dataset**, [Online], Available: https://www.kaggle.com/datasets/tunguz/big-five-pers

onality-test

[122] AVA, **AVA-Kinetics Dataset**, [Online], Available: https://research.google.com/ava/

[123] Mellon University, **CMU Graphics Lab Motion Capture Database**, [Online], Available: http://mocap.cs.cmu.edu/

[124] MPII, MPII Human Pose dataset, [Online], Available: http://human-pose.mpi-inf.mpg.de/

[125] Kaggle, Dataset, [Online], Available: https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset

[126] Kaggle, Dataset, [Online], Available: https://www.kaggle.com/datasets/ashtonsix/fake-turing-test

[127] Emily Dinan, et. al., "**Wizard of Wikipedia: Knowledge-Powered Conversational Agents,"** The Second Conversational Intelligence Challenge (ConvAI2), 2019. 01.

[128] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, Jason Weston, "**Wizard of wikipedia: Knowledge-powered conversational agents.**" arXiv preprint arXiv:1811.01241, 2018.

[129] Qiqiang Wu, Xianmin Zhang, Bo Zhao, "**A novel adaptive kernel-guided multi-condition abnormal data detection method**," Measurement, vol. 206, 2023. 01. https://doi.org/10.1016/j.measurement.2022.112257.

[130] Kaggle, **Resume-Screening-with-NLP**, [Online], Available: https://www.kaggle.com/code/akashkotal/resume-screening-with-nlp/notebook

[131] Kaggle, **You're Hired! |Analysis on Campus Recruitment Data**, [Online], Available: https://www.kaggle.com/code/benroshan/you-re-hired-analysis-on-campus-recruitment-data#1.Introduction

[132] Krauth, Brian., "**A dynamic model of job networking and social influences on employment.**" Journal of Economic Dynamics and Control., vol. 28, pp. 1185-1204, 2004. Doi: 10.1016/S0165-1889(03)00079-4.

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

[134] TTA-21167-SD**, 알기쉬운 ICT 표준해설서**, TTA, pp.71-86, 2021.

[135] Belhaouari, Samir Brahim, "**Unsupervised outlier detection in multidimensional data.**" Journal of Big Data, vol. 8 no. 1, pp. 1-27, 2021. https://doi.org/10.1186/s40537-021-00469-z

[136] Ronald E. Shiffler, "**Maximum Z Scores and Outliers,**" The American Statistician, vol. 42 no. 1, pp. 79-80, 1988. DOI: 10.1080/00031305.1988.10475530

[137] Aman Preet Gulati, **Dealing with outliers using the Z-Score method**, [Online], Available: https://www.analyticsvidhya.com/blog/2022/08/dealing-with-outliers-using-the-z-score-method/

[138] Vinutha, H.P., Poornima, B., Sagar, B.M., "**Detection of Outliers Using Interquartile Range Technique from Intrusion Dataset.**" In: Satapathy, S., Tavares, J., Bhateja, V., Mohanty, J. (eds) Information and Decision Sciences. Advances in Intelligent Systems and Computing, vol 701. 2018. https://doi.org/10.1007/978-981-10-7563-6_53

[139] Cabana, E., Lillo, R.E. andLaniado, H., "**Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators.**" Stat Papers. Vol. 62, pp. 1583-1609, 2021. https://doi.org/10.1007/s00362-019-01148-1

[140] V. Hautamaki, I. Karkkainen and P. Franti, "**Outlier detection using k-nearest neighbour graph,** " Proceedings of the 17th International Conference on Pattern Recognition ICPR 2004., vol. 3, pp. 430-433, 2004. doi: 10.1109/ICPR.2004.1334558.

[141] Chuang, Chen-Chia andLee, Zne-Jung., "**Hybrid robust support vector machines for regression with outliers.**" Appl. Soft Comput., vol. 11, pp. 64-72, 2011. Doi:10.1016/j.asoc.2009.10.017.

[142] Hawkins, Simon and He, Hongxing, Williams, Graham, Baxter, Rohan., "**Outlier Detection Using Replicator Neural Networks.**" Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery, vol. 2454, 2002. 113-123. 10.1007/3-540-46145-0_17.

[143] Xu H, Zhang L, Li P, Zhu F. "**Outlier detection algorithm based on k-nearest neighbors-local outlier factor.**" Journal of Algorithms & Computational Technology, vol. 16, 2022. doi:10.1177/17483026221078111

[144] Yosipof A, and Senderowitz H., "**k-Nearest neighbors optimization-based outlier removal.**" J Comput Chem., vol. 36 no. 8, pp. 493-506, 2015. 05. doi: 10.1002/jcc.23803. Epub 2014 Dec 15. PMID: 25503870.

[145] Siti Monalisa, and Fitra Kurnia, "**Analysis of DBSCAN and K-means algorithm for evaluating outlier on RFM model of customer behaviour.**"TELKOMNIKA Telecommunication, Computing, Electronics and Control, vol. 17 no. 1, pp. 110-117, 2019. https://doi.org/10.12928/TELKOMNIKA.v17i1.9394

[146] Hans-Peter Kriegel, Peer Kröger, Erich Schubert, and Arthur Zimek, "**LoOP: local outlier probabilities,**"In Proceedings of the 18th ACM conference on Information and knowledge management (CIKM '09). Association for Computing Machinery, pp. 1649-1652, 2009. https://doi.org/10.1145/1645953.1646195

[147] W. Liu, D. Cui, Z. Peng and J. Zhong, "**Outlier Detection Algorithm Based on Gaussian Mixture Model,**" 2019 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), pp. 488-492, 2019. doi: 10.1109/ICPICS47731.2019.8942474.

[148] ITWORLD, "**Defense of 'Beginning of Adversarial Machine Learning Response Strategy' Has Begun**," [Online], Available: https://www.itworld.co.kr/news/175699

[149] Yin, Xuwang, Soheil Kolouri, and Gustavo K. Rohde. "**Gat: Generative adversarial training for a**

dversarial example detection and robust classification." arXiv preprint arXiv:1905.11475, 2019.

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

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

[152] Weilin Xu, David Evans, Yanjun Qi, "**Feature Squeezing: Detecting Adversarial Examples in De ep Neural Networks,**" NDSS, 2018. 2

[153] Chang, Jung-Woo, Mojan Javaheripi, Seira Hidano, and Farinaz Koushanf, "**RoVISQ: Reductio n of Video Service Quality via Adversarial Attacks on Deep Learning-based Video Compressio n,**" arXiv preprint arXiv:2203.10183, 2022.

[154] Lo, Shao-Yuan, and Vishal M. Patel, "**Defending against multiple and unforeseen adversarial v ideos,**" IEEE Transactions on Image Processing, vol. 31, pp. 962-973, 2021.

[155] Claus, Michele, and Jan Van Gemert, "**Videnn: Deep blind video denoising,**" In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp. 0-0, 20 19

[156] Pony, Roi, Itay Naeh, and Shie Mannor, "**Over-the-air adversarial flickering attacks against vid eo recognition networks,**" In Proceedings of the IEEE/CVF Conference on Computer Vision an d Pattern Recognition, pp. 515-524, 2021.

[157] Xie, Shangyu, Han Wang, Yu Kong, and Yuan Hong, "**Universal 3-dimensional perturbations fo r black-box attacks on video recognition systems,**" In 2022 IEEE Symposium on Security and Privacy (SP), pp. 1390-1407, 2022.

[158] Li, Shasha, Abhishek Aich, Shitong Zhu, Salman Asif, Chengyu Song, Amit Roy-Chowdhury, a nd Srikanth Krishnamurthy, "**Adversarial attacks on black box video classifiers: Leveraging the power of geometric transformations,**" Advances in Neural Information Processing Systems, v ol. 34, pp. 2085-2096, 2021.

[159] Wei, Xingxing, Jun Zhu, Sha Yuan, and Hang Su, "**Sparse adversarial perturbations for videos,**" In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33 no. 01, pp. 8973-898 0, 2019.

[160] Mohd Fakhri Mat Saad, **"A Review of Artificial Intelligence Based Platform in Human Resource Recruitment Process,**" Faculty of Business, Communications, and Law INTI International Univ ersity Nilai, Malaysia 6 th IEEE International Conference on Recent Advances and Innovations in Engineering- ICRAIE, 2021.

[161] The National Institute of Standards and Technology (NIST), **Towards a Standard for Identifying and Managing Bias in Artificial Intelligence**, [Online], Available: https://nvlpubs.nist.gov/nistp

ubs/SpecialPublications/NIST.SP.1270.pdf.

[162] High-Level Expert Group On Artificial Intelligence Set Up By The European Commission, **Ethics Guidelines For Trustworthy Ai,** [Online], Available: https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html

[163] Korea Law Translation Center, **Personal Information Protection Act**, [Online], Available: https://elaw.klri.re.kr/eng_service/lawView.do?hseq=53044&lang=ENG

[164] Joshi, C., Kaloskampis, I., Nolan, L., **Generative adversarial networks (GANs) for synthetic dataset generation with binary classes**, [Online], Available:https://datasciencecampus.ons.gov.uk/projects/generative-adversarial-networks-gans-for-synthetic-dataset

[165] Chanda, Sasanka Sekhar, and Debarag Narayan Banerjee, "Omission and commission errors underlying AI failures,"AI & society, pp. 1-24, 2022.

[166] Tengai, **Digital interviews raise concerns**, [Online], Available: https://tengai.io/blog/is-your-remote-interview-biased

[167] Fernández-Martínez, Carmen and Fernández, Alberto, "AI and recruiting software: Ethical and legal implications," Paladyn, Journal of Behavioral Robotics, vol. 11, no. 1, pp. 199-216, 2020. https://doi.org/10.1515/pjbr-2020-0030

[168] Lindsey Zuloaga, **AI in Recruiting: What it Means for Talent Acquisition in 2022,** [Online], Available: https://www.hirevue.com/blog/hiring/ai-in-recruiting-what-it-means-for-talent-acquisition

[169] Ibert, E. T., "**AI in talent acquisition: a review of AI-applications used in recruitment and selection,**"Strategic HR Review, vol. 5, pp. 215-221, 2019.

[170] Ben Dattner, Tomas Chamorro-Premuzic, Richard Buchband and Lucinda Schettler, "**The Legal and Ethical Implications of Using AI in Hiring,**" Harvard Business Review, 2019. 04.

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

[172] B C Lee, B Y Kim, "**Development of An Ai-Based Interview System for Remote Hiring**,"International Journal of Advanced Research in Engineering and Technology (IJARET), vol. 12 issue 3, pp. 654-663, 2021. 03. DOI: 10.34218/IJARET.12.3.2021.060

[173] Kulkarni, Swatee B., and Xiangdong Che, "**Intelligent software tools for recruiting,**" Journal of International Technology and Information Management, vol. 28 no. 2, pp. 2-16, 2019.

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

[175] Langer, Markus, Kevin Baum, Cornelius J. König, Viviane Hähne, Daniel Oster, and Timo Speith, "**Spare me the details: How the type of information about automated interviews influences applicant reactions,**" International Journal of Selection and Assessment, vol. 29 no. 2, pp. 154

-169, 2021. DOI: 10.1111/ijsa.12325

[176] Merylin Monaro, Stéphanie Maldera, Cristina Scarpazza, Giuseppe Sartori, NicolòNavarin, "**Detecting deception through facial expressions in a dataset of videotaped interviews: A comparison between human judges and machine learning models,**" Computers in Human Behavior, 2022.

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

[178] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, "**SMOTE: Synthetic Minority Over-sampling Technique**," arXiv:1106.1813v1, 2011. 6.

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

[180] H. He, Y. Bai, E. A. Garcia, S. Li, "**ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning,**" 2008 IEEE International Joint Conference on Neural Networks, 2008. 6.

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

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

[183] Open-source Initiative, **The Open-source Definition**, [Online], Available: https://opensource.org/osd

[184] Baxter Kathy, "**What is AI bias mitigation, and how can it improve AI fairness?**", New Tech Forum, 2021. 08.https://www.infoworld.com/article/3630450/what-is-ai-bias-mitigation-and-how-can-it-improve-ai-fairness.html

[185] Venturebeat, **Researchers find that even 'fair' hiring algorithms can be biased**, [Online], Available: https://venturebeat.com/ai/researchers-find-that-even-fair-hiring-algorithms-can-be-biased/

[186] PIPA, **Article 62 of the Enforcement Decree of the Personal Information Protection Act (Reporting on Infringements)**, [Online], Available: https://elaw.klri.re.kr/kor_service/lawView.do?lang=ENG&hseq=53044&joseq=JO0062000

[187] McKenzie Raub, "**Bots, Bias and Big Data: Artificial Intelligence, Algorithmic Bias and Disparate Impact Liability in Hiring Practices,**" Arkansas Law Review Arkansas Law Review, 529, vol. 71 no. 2, 2018. 12. https://scholarworks.uark.edu/alr/vol71/iss2/7.

05 참고문헌

[188] Fernández, Carmen, and Alberto Fernández, "**Ethical and legal implications of AI recruiting software,**" Ercim News, Special Theme: Transparency in Algorithmic Decision Making, vol. 116, pp. 22-23, 2019.

[189] Martínez, María del Carmen Fernández, and Alberto Fernández, "**AI in Recruiting. Multi-agent Systems Architecture for Ethical and Legal Auditing,**" In IJCAI, pp. 6428-6429. 2019

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

[191] Naim, Iftekhar, Md Iftekhar Tanveer, Daniel Gildea, and Mohammed Ehsan Hoque, "**Automated analysis and prediction of job interview performance,**" IEEE Transactions on Affective Computing, vol. 9 no. 2, pp. 191-204, 2016.

[192] Lee, B. C., and B. Y. Kim, "**Development of an AI-based interview system for remote hiring,**" International. J. Adv. Res. Eng. Technol, vol. 12, pp. 654-663, 2021.

[193] Rudolph, Matthias, **Artificial Intelligence in Recruiting. A Literature Review on Artificial Intelligence Technologies, Ethical Implications and the Resulting Chances and Risks,** [Online], Available: https://www.grin.com/document/978174

[194] Christopher Schmidt, **Approaching Unbalanced Datasets Using Data Augmentation**, [Online], Available: https://medium.com/@cjc.schmidt/approaching-unbalanced-datasets-using-data-augmentation-8b4978e1cf2e

[195] Chen, Lei, Ru Zhao, Chee Wee Leong, Blair Lehman, Gary Feng, and Mohammed Ehsan Hoque, "**Automated video interview judgment on a large-sized corpus collected online,**" In 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 504-509, 2017.

[196] Varma, S., Simon, R., "**Bias in error estimation when using cross-validation for model selection,**" BMC Bioinformatics, vol. 7 no. 91, 2006. https://doi.org/10.1186/1471-2105-7-91

[197] Man Luo, Yankai Zeng, Pratyay Banerjee, Chitta Baral, "**Weakly-Supervised Visual-Retriever-Readerfor Knowledge-based Question Answering,**" EMNLP, 2021.

[198] Short, Austin, Trevor La Pay, and Apurva Gandhi, "**Defending Against Adversarial Examples,**" No. SAND2019-11748. Sandia National Lab. (SNL-NM), 2019.

[199] Liu, Shengyi, "**Model Extraction Attack and Defense on Deep Generative Models,**" In Journal of Physics: Conference Series, vol. 2189 no. 1, pp. 012024, 2022.

[200] Tramèr, Florian, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart, "**Stealing Machine Learning Models via Prediction APIs.**" In USENIX security symposium, vol. 16, pp. 601-618, 2016.

[201] Hu, Hailong, and Jun Pang, "**Model extraction and defenses on generative adversarial network**

s," arXiv preprint arXiv:2101.02069, 2021.

[202] Zhongshu Gu, Ridgewood, Heqing Huang, Mahwah, Marc Phillipe Stoecklin, Jialong Zhang, "**Protecting Deep Learning Models Using Watermarking,**" White Plains, 2018. 06. https://patentimages.storage.googleapis.com/59/b1/12/fe1c2fde585dd5/US11163860.pdf

[203] Insights2Techinfo, **Self-Driving Automobiles and Adversarial Attacks**, [Online], Available: https://insights2techinfo.com/self-driving-automobiles-and-adversarial-attacks/

[204] Zhang, Hu, Linchao Zhu, Yi Zhu, and Yi Yang, "**Motion-excited sampler: Video adversarial attack with sparked prior,**" In Computer Vision-ECCV 2020: 16th European Conference, Proceedings, Part XX vol. 16, pp. 240-256, 2020. 08.

[205] Papernot, Nicolas, et al., "**Technical report on the cleverhans v2. 1.0 adversarial examples library,**" arXiv preprint arXiv:1610.00768, 2016.

[206] Austin Short, Trevor La Pay, Apurva Gandhi, **Defending Against Adversarial Examples**, [Online], Available: https://www.osti.gov/servlets/purl/1569514

[207] Dongyu Meng, Hao Chen, "**MagNet: a Two-Pronged Defense against Adversarial Examples**," arXiv:1705.09064, 2017. 9.

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

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

[210] Linardatos et al., "**Explainable ai: A review of machine learning interpretability methods,**" 2020. https://www.mdpi.com/1099-4300/23/1/18

[211] Hofeditz, Lennart, Sünje Clausen, Alexander Rieß, Milad Mirbabaie, and Stefan Stieglitz, "**Applying XAI to an AI-based system for candidate management to mitigate bias and discrimination in hiring,**" Electronic Markets, pp. 1-27, 2022.

[212] Fleiß, Jürgen, Elisabeth Bäck, and Stefan Thalmann, "**Explainability and the intention to use AI-based conversational agents. An empirical investigation for the case of recruiting**," [Online], Available: https://ceur-ws.org/Vol-2796/xi-ml-2020_fleiss.pdf

[213] Ortega, Alfonso, Julian Fierrez, Aythami Morales, Zilong Wang, and Tony Ribeiro, "**Symbolic AI for XAI: Evaluating LFIT inductive programming for fair and explainable automatic recruitment,**" In Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp. 78-87, 2021.

[214] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin, "**" Why should i trust you?" Explaining the predictions of any classifier,**" In Proceedings of the 22nd ACM SIGKDD international con

ference on knowledge discovery and data mining, pp. 1135-1144, 2016.

[215] Arakawa, Riku, and Hiromu Yakura, "**AI for human assessment: What do professional assessors need?,**" arXiv preprint arXiv:2204.08471, 2022.

[216] Talal Shaikh, Aaishwarya Khalane, Rikesh Makwana, et al. "**Evaluating Significant Features in Context-Aware Multimodal Emotion Recognition with XAI Methods,**"Authorea, 2023. 01. Doi: 10.22541/au.167407909.97031004/v1

[217] Adak, Anirban, Biswajeet Pradhan, Nagesh Shukla, and Abdullah Alamri, "**Unboxing deep learning model of food delivery service reviews using explainable artificial intelligence (XAI) technique,**" Foods, vol. 11 no. 14, 2022. https://doi.org/10.3390/foods11142019

[218] Github, **SHAP (SHapley Additive exPlanations)**,[Online], Available: https://christophm.github.io/interpretable-ml-book/shap.html#fnref44

[219] Byungwook Choi, **The Present and Future of Medical AI**, Korea Institute of Health and Medical Research, vol. 60, [Online], Available: https://hineca.kr/1868

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

[221] van Esch, P., and Black, J. S., "**Factors that influence new generation candidates to engage with and complete digital, AI-enabled recruiting,**" Business Horizons, vol. 62, pp. 729-739, 2019. https://doi.org/10.1016/j.bushor.2019.07.004

[222] European Parliament, "**Artificial Intelligence in Healthcare: Applications, Risks, and Ethical and Societal Impacts,**" European Parliamentary Research Service, 2022. 6. https://www.europarl.europa.eu/RegData/etudes/STUD/2022/729512/EPRS_STU(2022)729512_EN.pdf

[223] IBM, **A Methodology for Creating AI FactSheets**, [Online], Available: http://aifs360.mybluemix.net/methodology

[224] Day One Team, **Our Leadership Principles**, [Online], Available: https://www.aboutamazon.eu/news/working-at-amazon/our-leadership-principles

[225] Dattner, B., Chamorro-Premuzic, T., Buchband, R., Schettler, L., "**The legal and ethical implications of using AI in hiring,**" Harvard Business Review, 2019. 04.

[226] Sánchez-Monedero, J., Dencik, L., and Edwards, L., "**What does it mean to 'solve' the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems,**"In Conference on Fairness, Accountability, and Transparency New York: Association for Computing Machinery, pp. 458-468, 2020. 01. https://doi.org/10.1145/3351095.3372849.

[227] Raghavan, M., Barocas, S., Kleinberg, J.,and Levy, K., "**Mitigating bias in algorithmic hiring: Ev**

aluating claims and practices," In Conference on fairness, accountability, and transparency Ne w York: Association for Computing Machinery, 2020. 01.

[228] Simbeck, K., "**HR analytics and ethics,**" IBM Journal of Research and Development, vol. 63no. 4/5, pp. 1-12, 2019.

[229] Tambe, P., Cappelli, P., Yakubovich, V.,"**Artificial intelligence in human resources managemen t: Challenges and a path forward,**" California Management Review, vol. 61, pp. 15-42, 2019. https://doi.org/10.1177/0008125619867910

[230] Vasconcelos, M., Cardonha, C., and Gonçalves, B., "**Modeling epistemological principles for bi as mitigation in AI systems: An illustration in hiring decisions,**" In J. Furman, G. Marchant, H. Price, & F. Rossi (Eds.), AAAI/ACM Conference on AI, Ethics, and Society, pp. 323-329, 201 8. 02. https://doi.org/10.1145/3278721.3278751

[231] Pena, A., Serna, I., Morales, A., and Fierrez, J., "**Bias in multimodal AI: Testbed for fair autom atic recruitment,**"In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 129-137, 2020. 06. https://doi.org/10.1109/CVPRW50498.2020.0 0022.

[232] Schumann, C., Foster, J. S., Mattei, N., Dickerson, J. P., "**We need fairness and explainability in algorithmic hiring,**"In B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, & G. Sukthankar (Ed s.), 19th international conference on autonomous agents and multiagent systems (AAMAS 20 20), 2020. 05.

[233] Google, **Explainable AI**, [Online], Available: https://cloud.google.com/explainable-ai

[234] Raghavan, M., Barocas, S., Kleinberg, J., and Levy, K., "**Mitigating bias in algorithmic hiring: E valuating claims and practices,**" In Conference on fairness, accountability, and transparency N ew York: Association for Computing Machiner, 2020. 01.

[235] Chen, Z. "**Ethics and discrimination in artificial intelligence-enabled recruitment practices,**" Hu manit Soc Sci Commun, vol. 10 no. 567, 2023. https://doi.org/10.1057/s41599-023-02079-x

[236] Microsoft, **Responsible bots: 10 guidelines for developers of conversational AI**, [Online], Avail able: https://www.microsoft.com/en-us/research/uploads/prod/2018/11/Bot_Guidelines_N ov_2018.pdf

[237] Microsoft, **Responsible bots: 10 guidelines for developers of conversational AI**, [Online], Avail able: https://www.microsoft.com/en-us/research/uploads/prod/2018/11/Bot_Guidelines_N ov_2018.pdf

[238] Diptiben Ghelni., "**Deep Learning and Artificial Intelligence Framework to Improve the Cyber S ecurity,**" American Journal of Artificial Intelligence, Vol. x, No. x, 2022, pp. x-x, https://d197f or5662m48.cloudfront.net/documents/publicationstatus/90291/preprint_pdf/c12f4b6dfcb0ec

e3a42a357ad2203fac.pdf

[239] Li, Jian-hua., "**Cyber security meets artificial intelligence: a survey,**" Frontiers of Information T echnology & Electronic Engineering, vol. 19 no. 12, pp. 1462-1474, 2018.

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

[241] Reid Blackman, '**A Practical Guide to Building Ethical AI**', Harvard Business Review, [Online], Available: https://hbr.org/2020/10/a-practical-guide-to-building-ethical-ai

[242] European Commission, **Ethics By Design and Ethics of Use Approaches for Artificial Intelligen ce,** [Online], Available: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intel ligence_he_en.pdf

[243] Korea Legislation Research Institute, **PIPA**, [Online], Available: https://elaw.klri.re.kr/eng_ser vice/lawView.do?hseq=53044&lang=ENG

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

[245] Kalev, Alexandra, Frank Dobbin, and Erin Kelly, "**Best practices or best guesses? Assessing th e efficacy of corporate affirmative action and diversity policies,**" American sociological review, vol. 71 no. 4, pp. 589-617, 2006.

[246] D. F. Mujtaba and N. R. Mahapatra, "**Ethical Considerations in AI-Based Recruitment,**" 2019 IE EE International Symposium on Technology and Society (ISTAS), pp. 1-7, 2019. doi: 10.1109/ ISTAS48451.2019.8937920

[247] Riveiro, Maria, and Serge Thill, ""**That's (not) the output I expected!" On the role of end user e xpectations in creating explanations of AI systems,**" Artificial Intelligence, vol. 298, 2021.

[248] Federal Trade Commission, **Washington, D.C. 20580- EPIC Petition for FTC Rulemaking on A I**, [Online], Available: https://epic.org/wp-content/uploads/privacy/ftc/ai/EPIC-FTC-AI-Petiti on.pdf

[249] Fernández-Martínez, Carmen, and Alberto Fernández, "**AI and recruiting software: Ethical and legal implications,**" Paladyn, Journal of Behavioral Robotics, vol. 11 no. 1, pp. 199-216, 2020.

[250] Drage, Eleanor, and Kerry Mackereth, "**Does AI Debias Recruitment? Race, Gender, and AI's " Eradication of Difference",**" Philosophy & technology, vol. 35 no. 4, 2022.

[251] Raven Veal, **How to Define a User Persona**, [Online], Available: https://careerfoundry.com/en /blog/ux-design/how-to-define-a-user-persona/

[252] Ministry of Employment and Labor (Employment Division for the Disabled), **Employment Pro**

motion and Vocational Rehabilitation for Disabled Persons Act (Abbreviation: Employment for Disabled Persons Act), 044-202-7482, [Online], Available: https://www.law.go.kr/%EB%B2%95%EB%A0%B9/%EC%9E%A5%EC%95%A0%EC%9D%B8%EA%B3%A0%EC%9A%A9%EC%B4%89%EC%A7%84%EB%B0%8F%EC%A7%81%EC%97%85%EC%9E%AC%ED%99%9C%EB%B2%95

[253] B.J. Dietvorst, J. Simmons, C. Massey, "**Understanding algorithm aversion: forecasters errone ously avoid algorithms after seeing them err,**" Acad. Manag. Proc., vol. 1, p. 12227, 2015. htt ps://doi.org/10.5465/ambpp.2014.12227abstract

[254] Laato, Samuli, Miika Tiainen, A. K. M. Najmul Islam, and Matti Mäntymäki, "**How to explain AI systems to end users: a systematic literature review and research agenda,**" Internet Researc h, vol. 32 no. 7, pp. 1-31, 2022.

[255] HireVue, **Explainability Statement**, [Online], Available: https://hirevue-api.dev-directory.com /wp-content/uploads/2022/04/HV_AI_Short-Form_Explainability_1pager.pdf

[256] Leslie, D., "**Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector,**" The Alan Turing Institute, 201 9. https://doi.org/10.5281/zenodo.324052

[257] Kuttal, Sandeep Kaur, Xiaofan Chen, Zhendong Wang, Sogol Balali, and Anita Sarma, "**Visual R esume: Exploring developers' online contributions for hiring,**" Information and Software Techn ology, vol. 138, 2021. https://doi.org/10.1016/j.infsof.2021.106633.

[258] Suen, Hung-Yue, Kuo-En Hung, and Chien-Liang Lin, "**Intelligent video interview agent used to predict communication skill and perceived personality traits,**" Human-centric Computing an d Information Sciences, vol 10, no 3, pp. 1-12, 2020. 01. https://doi.org/10.1186/s13673-02 0-0208-3

[259] Genesis Lab, "제네시스랩결과지- IT/SW development recruitment in the first half of 2023", Whi te Paper, 2023.

[260] Midas, **3. Utilization of inspection system and AI**, [Online], Available: https://www.midashri.c om/aicc-03

[261] Chou, Yi-Chi, Felicia R. Wongso, Chun-Yen Chao, and Han-Yen Yu, "**An AI Mock-interview Pl atform for Interview Performance Analysis,**" In 2022 10th International Conference on Informa tion and Education Technology (ICIET), pp. 37-41. 2022. 05. doi: 10.1109/ICIET55102.2022.9 778999

[262] Simran Singh, **Why UX is the most important feature in recruitment software**, [Online], Availa ble: https://recruitee.com/articles/ux-recruitment-software#3

[263] Naveed Ahmed, **How to Evaluate UX Design? [Tips, Tricks and Guide for 2022]**, [Online], Avail able: https://enou.co/blog/how-to-evaluate-ux-design/

05 참고문헌

[264] EU Commission, **Ethics By Design and Ethics of Use Approaches for Artificial Intelligence**, [Online], Available: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf

[265] Singapore Minister for Communications and Information, **Model Artificial Intelligence Governance Framework - Second Edition**, [Online], Available: https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf

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

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

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

[269] **Git Large File Storage (LFS)**, [Online], Available: https://git-lfs.github.com/

[270] **lakeFS**, [Online], Available: https://lakefs.io/

[271] **Delta Lake**, [Online], Available: https://delta.io/

[272] Interviewer.AI, **How Do Companies Benefit with Video Interviews for Hiring**, [Online], Available: https://interviewer.ai/success-story/video-interviews-for-hiring/

[273] Ideal, **AI for Recruiting: A Definitive Guide for HR Professionals**, [Online], Available: https://ideal.com/ai-recruiting/

[274] Microsoft, **Guidelines for Human-AI Interaction**, [Online], Available: https://www.microsoft.com/en-us/haxtoolkit/uploads/prod/2021/05/AI-Design-guidelines_041519.pdf

[275] Chaitanya Pattapu, **How AI Interviewing is Redefining the Way we Hire**, [Online], Available: https://blog.talview.com/how-ai-interviewing-redefining-the-way-we-hire

[276] Monica Montesa, **AI Recruiting in 2023: The Definitive Guide**, [Online], Available: https://www.phenom.com/blog/recruiting-ai-guide

2024

신뢰할 수 있는 인공지능 개발 안내서 채용 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 979-11-89545-65-9