2023

신뢰할 수 있는 인공지능 개발 안내서

자율주행 분야

과학기술정보통신부
Ministry of Science and ICT

TTA 한국정보통신기술협회
Telecommunications Technology Association

Jinhan M&B

일러두기

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

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

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

- 본 안내서의 자율주행·인공지능 동향 및 기술 정보는 2023년 2월 기준으로 서술되었습니다.

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

- 본 안내서는 한국정보통신기술협회가 운영하는 TrustOps 웹페이지(2023년 하반기 공개 예정)에도 콘텐츠가 공개되어 있으므로 참고하시면 더 편리하게 활용하실 수 있습니다.

- 자율주행 외 분야는 《2023 신뢰할 수 있는 인공지능 개발 안내서 – 일반 분야》를 참고해주시기 바라며, 특화될 서비스 분야는 점차 확대해나갈 예정입니다.

CONTENTS

Checklist 안내서 활용을 위한 체크리스트 6

PART 1 개 요 11

1. 안내서 발간 배경 및 목적 12
2. 자율주행 인공지능 신뢰성 동향 13
3. 안내서 마련 과정 17
4. 안내서 활용 대상 25
5. 안내서 활용 방법 27

PART 2 요구사항 및 검증항목 29

1. 계획 및 설계 34
2. 데이터 수집 및 처리 51
3. 인공지능 모델 개발 84
4. 시스템 구현 108
5. 운영 및 모니터링 133

PART 3 부 록 145

1. 약어표 146
2. 용어표 149
3. 참고문헌 151

안내서 활용을 위한 체크리스트

안내서 활용을 위한 체크리스트

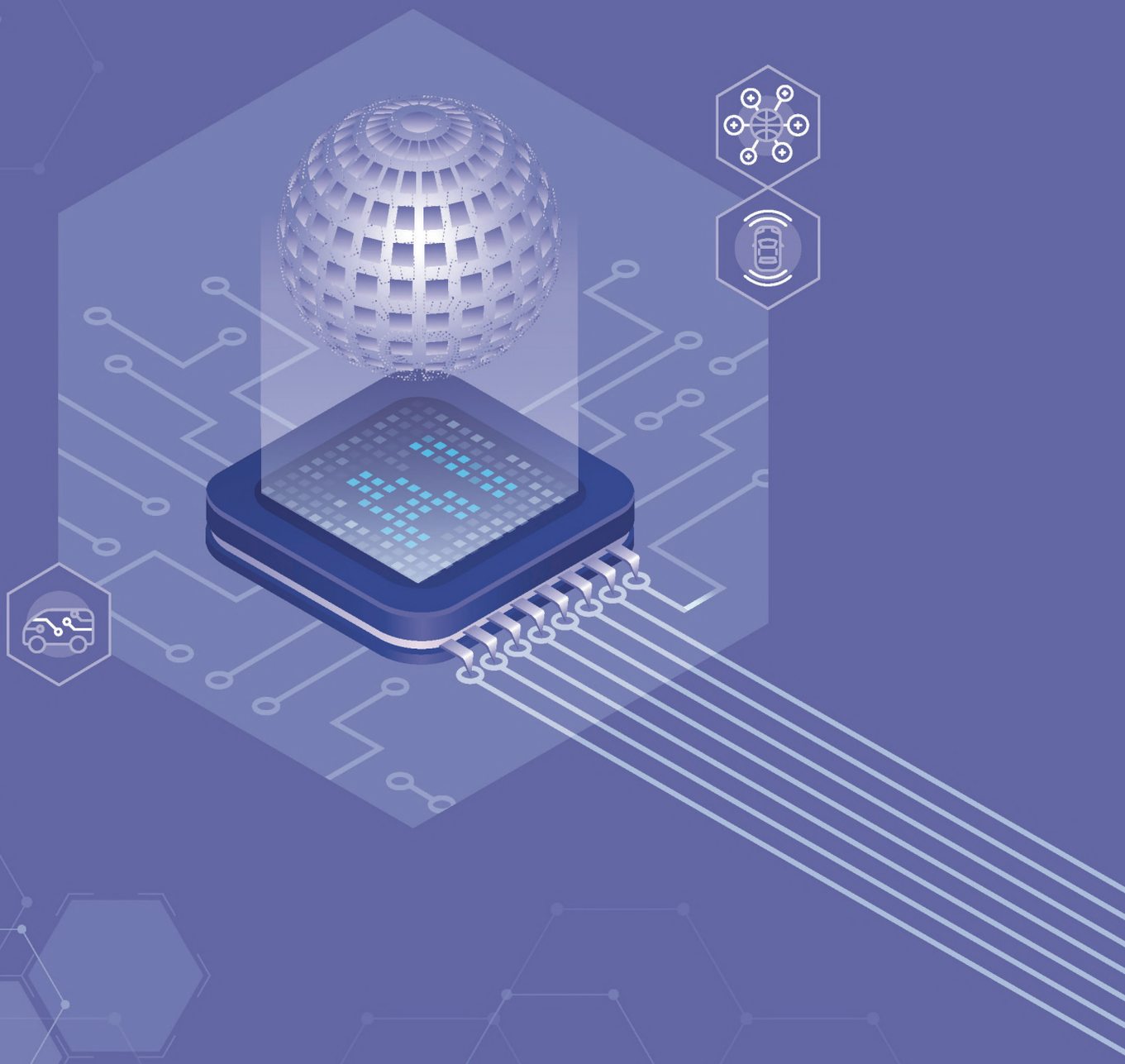
생명주기	요구사항 및 체크리스트	Yes	No	N/A
1 계획 및 설계	**요구사항 01** 인공지능 시스템에 대한 위험관리 계획 및 수행			
	01-1 인공지능 시스템 생명주기에 걸쳐 나타날 수 있는 위험 요소를 분석하였는가?	☐	☐	☐
	01-1a 인공지능 시스템의 위험 요소를 도출하고 이의 파급효과를 파악하였는가?	☐	☐	☐
	01-2 위험 요소를 제거 및 방지하거나 영향을 완화하기 위한 방안을 마련하였는가?	☐	☐	☐
	01-2a 위험 요소 제거 방안을 도출하고 파급효과가 감소하였는지 확인하였는가?	☐	☐	☐
	요구사항 02 인공지능 거버넌스governance 체계 구성			
	02-1 인공지능 거버넌스에 대한 지침 및 규정을 수립하였는가?	☐	☐	☐
	02-1a 내부적으로 준수해야 할 인공지능 거버넌스에 대한 지침 및 규정을 마련하였는가?	☐	☐	☐
	02-2 인공지능 거버넌스를 위한 조직을 구성하고 인력 구성에 대해 검토하였는가?	☐	☐	☐
	02-2a 인공지능 거버넌스를 위한 조직을 구성하였는가?	☐	☐	☐
	02-2b 인공지능 거버넌스를 위한 조직은 충분히 훈련된 인력으로 구성하였는가?	☐	☐	☐
	02-3 인공지능 거버넌스 체계가 올바로 이행되고 있는지 감독하고 있는가?	☐	☐	☐
	02-3a 인공지능 거버넌스에 대한 내부 지침 및 규정 준수 여부를 감독하고 있는가?	☐	☐	☐
	02-4 인공지능 거버넌스 조직이 신규 및 기존 시스템의 차이점을 분석하였는가?	☐	☐	☐
	02-4a 신규 인공지능 시스템 도입 전, 기존 시스템의 대체 필요성 등을 분석하였는가?	☐	☐	☐
	요구사항 03 인공지능 시스템의 신뢰성 테스트 계획 수립			
	03-1 인공지능 시스템의 특성을 고려한 테스트 환경을 설계하였는가?	☐	☐	☐
	03-1a 테스트 환경 결정 시 각 환경에서 테스트 가능한 주행 시나리오를 고려하였는가?	☐	☐	☐
	03-1b 시뮬레이터 및 주행시험장 등 테스트 환경을 확보하였는가?	☐	☐	☐
	03-2 인공지능 시스템의 테스트 설계에 필요한 협의 체계를 구성하였는가?	☐	☐	☐
	03-2a 인공지능 시스템의 기대 출력을 결정하기 위한 협의 체계를 구성하였는가?	☐	☐	☐
	03-2b 설명가능성 및 해석가능성 확인을 위한 사용자 평가단을 구성하였는가?	☐	☐	☐
2 데이터 수집 및 처리	**요구사항 04** 데이터의 활용을 위한 상세 정보 제공			
	04-1 데이터의 명확한 이해와 활용을 지원하는 상세한 정보를 제공하는가?	☐	☐	☐
	04-1a 정제 전과 후의 데이터 특성을 설명하였는가?	☐	☐	☐
	04-1b 학습 데이터와 메타데이터metadata를 구분하였으며, 각각에 대한 명세자료를 확보하였는가?	☐	☐	☐
	04-1c 보호변수protective attribute의 선정 이유 및 반영 여부를 설명하였는가?	☐	☐	☐
	04-1d 라벨링 작업자를 위해 교육을 시행하고 작업 가이드 문서를 마련하였는가?	☐	☐	☐
	04-2 데이터의 출처는 기록 및 관리되고 있는가?	☐	☐	☐
	04-2a 신뢰할 수 있는 출처로부터 제공되는 데이터셋을 사용하였는가?	☐	☐	☐
	04-2b 오픈소스 데이터셋을 활용하는 경우, 출처를 명시하였는가?	☐	☐	☐

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

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

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

2023 신뢰할 수 있는 인공지능 개발 안내서 | 자율주행 분야



PART 1

개요

1. 안내서 발간 배경 및 목적

2. 자율주행 인공지능 신뢰성 동향

3. 안내서 마련 과정

4. 안내서 활용 대상

5. 안내서 활용 방법

01 안내서 발간 배경 및 목적

자율주행차란 운전자 또는 승객의 조작 없이 스스로 운행이 가능한 차량을 말한다. 자세히는, 차량에 자율주행을 위한 센서 등 첨단 기술을 적용하여 스스로 주변 환경을 인식하고, 위험을 판단하고 주행 경로를 계획하여 운전자 또는 승객의 조작 없이 안전하게 운행할 수 있게 한 차량을 의미한다[1]. 자율주행차는 인간처럼 주변 상황을 인지할 수 있어야 하며, 인지한 상황에 따라 어떻게 행동할 것인가를 판단하여, 결과적으로 인간보다 더 나은 주행을 할 수 있어야 한다. 따라서 인간의 지능과 생각 체계를 모사한 인공지능은 자율주행 분야에서 선택이 아닌 필수가 되어가고 있다.

자율주행 분야에서 인공지능은 규칙 기반rule-based의 자동화된 차량 제어를 넘어 안전하고 효율적인 차량 주행 방법을 스스로 학습하며 주행 능력을 향상하고 있다. 자율주행차는 다양한 교통상황에 모두 대응할 수 있어야 하는데, 규칙 기반 차량 제어 방식은 자율주행차에는 한계가 있다. 따라서 최근에는 심층학습 기반의 인공지능을 구현하여 인간 운전자의 차량 제어 방식을 학습하게 함으로써 인간 운전자에 버금가는, 또는 이를 넘어서는 학습 기반 자율주행 알고리즘이 개발되고 있다. 방대한 양의 주행 데이터를 학습하여 어떤 상황에서 어떻게 차량을 제어하면 어떤 결과로 이어지는지 등 일련의 차량 제어 과정을 익히는 것이다[2]. 그 외에도, 강화학습 등의 다양한 기법이 적용되고 있는 추세이다.

이렇듯 자율주행 분야에서 인공지능의 중요성과 역할이 커지면서 자율주행 인공지능의 신뢰성trustworthiness 확보가 중요한 과제로 떠올랐다. 인공지능은 그 특성상 작동 원리나 메커니즘을 파악·이해하기 어려우며, 추론 결과를 도출하는 과정에서 오류(예: 데이터 오염, 편향성, 모델 추출 공격)를 범할 가능성이 크기 때문이다. 특히 자율주행 시스템에서 오류가 발생하면 사고로 이어질 수 있고 운전자·보행자 등의 생명에 직접적인 영향을 미치므로 높은 수준의 신뢰성이 요구된다.

이에 따라, 전 세계적으로 자율주행 인공지능의 신뢰성에 관해 여러 대응 방안이 마련되고 있다. 유럽위원회EC, European Commission는 자율주행 차량의 신뢰성을 다룬 《Trustworthy Autonomous Vehicles('21)》와 자율주행의 기술 및 사이버 보안을 다룬 《Cybersecurity Challenges in the Uptake if Artificial Intelligence in Autonomous Driving('21)》을 발간하였다. 이 외에도 시스템의 결함이 없더라도 차량 내·외부 환경에 대한 센싱 기술이 중요한 자율주행시스템에서 의도한 기능 안전을 확보하기 위한 ISO 21448:2022 – Safety of the intended functionality 표준이 개발되었다. 국내에서도 이에 발맞춰 자율주행차가 인명 보호를 최우선 하도록 설계·제작되는 것을 목표로 《자율주행차 윤리 가이드라인('20)》을 발표하였다.

그러나 지금까지 나온 자율주행 분야의 인공지능 신뢰성 원칙, 제안 정책, 표준 등은 주로 윤리 또는 프로세스 관점에서 추상적인 항목을 제시하고 있어 실무 현장에서 활용하기는 어렵다. 특히 인력과 연구 개발 투자 여력이 제한적인 중소기업은 직접 신뢰성 요구사항을 도출하거나 검증체계를 마련하기 어려워 이러한 현실적인 문제점을 해결하고자 본 개발 안내서가 작성되었다. 미국과 유럽 등 선진국과 국제기구들에서 발표한 권고안, 가이드, 표준, 사례 및 연구자료 등을 참고하였으며 주요 항목에 대한 개발 요구사항과 검증항목으로 자율적 점검이 가능할 것이다.

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

02 자율주행 인공지능 신뢰성 동향

자율주행차는 운전자에게 여러 가지 편의를 제공하며 교통사고를 줄이고 교통혼잡을 개선할 목적으로 활용된다. 현재 일어나는 교통사고나 교통혼잡은 차량의 부품적 결함이 원인인 경우는 극히 적으며, 운전자의 인지 반응시간과 차로 변경, 운전 미숙 등 다양한 인적 요인으로 발생하는 경우가 대부분이다. 자율주행차에서는 이러한 인적 요인들이 배제되며 여러 가지 센서를 이용해 데이터를 수집·분석하여 차간 거리와 차두시간을 일정하게 유지·주행함으로써 기존의 교통류에서 교통혼잡을 일으키는 근본적인 원인을 해소할 수 있을 것으로 기대된다[3]. 본 장에서는 자율주행차의 이점을 극대화하기 위한 자율주행 인공지능 기술과 더불어 이슈 사례를 소개함으로써 신뢰성 확보의 필요성에 대해 다룬다.

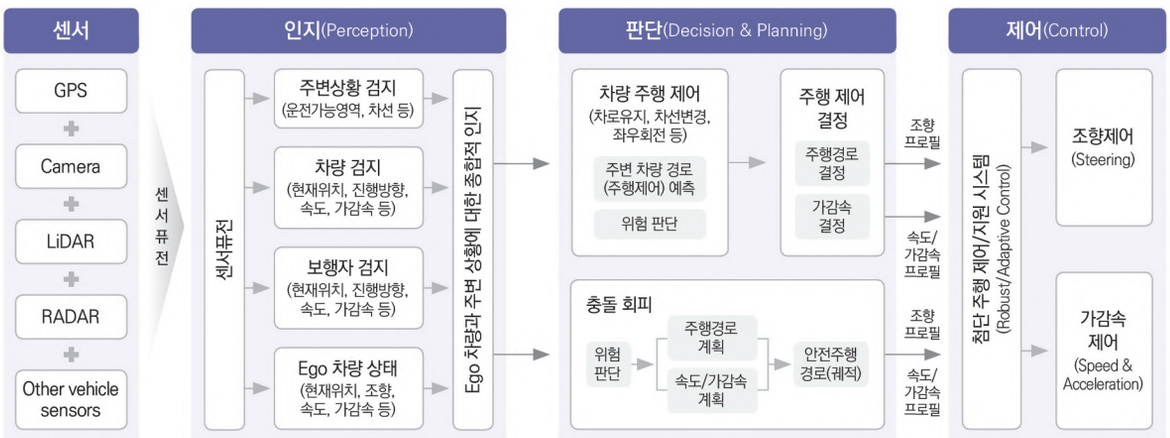
2.1 자율주행 인공지능 기술

자율주행차는 운전자를 대신하여 자율주행시스템이 차량을 운행하기 때문에 운전자가 차량을 주행하는 데 필요한 모든 행동을 대신할 수 있어야 한다. 또한, 차량이 주행하는 교통환경은 차량뿐만 아니라 보행자, 자전거, 전동킥보드 등 다양한 객체가 상호작용하는 복잡한 환경이기 때문에 예기치 않은 변수에 대응하고 순간적으로 빠른 판단이 필요한 상황이 자주 발생할 수 있다. 이에 대처하기 위해 자율주행차는 사람 운전자와 같이 인지한 각종 교통 상황 정보를 종합하고, 상황을 고려해 빠른 판단을 내릴 수 있는 기능을 갖춰야 한다. 이를 자율주행차의 인지·판단·제어의 3가지 과정으로 일반적으로 정의한다. 자율주행차는 차량에 장착된 각종 센서에서 수집한 데이터를 종합하여 상황을 '인지'하고, 인지한 상황에 근거하여 차량을 어떻게 제어하고 주행해야 할지 '판단'하며, 이러한 주행 제어 측면의 판단에 따라 차량을 '제어'한다.

① 인지perception: 자율주행을 위한 첫 번째 필수 기능은 사람의 눈과 귀 역할을 하는 카메라, 레이더, 라이다 등 센서 기술을 활용한 '인지'이다. 인지 기술은 자율주행차의 상황 판단 및 차량 제어 기반이 되는 데이터를 수집하고 분석하는 기술이기에 정확하게 정보를 수집하여 센싱된 객체를 정확하게 판별해내는 능력이 무엇보다 중요하다. 센서를 하나가 아니라 여러 개 사용하는 이유 역시 정확한 인지를 위해서이며, 앞에서 언급한 센서만으로 정확한 정보를 파악하기 힘들거나 센싱된 정보를 검증해야 할 때를 대비해 디지털맵, GPSGlobal Positioning System, V2XVehicle to Everything 무선통신을 함께 사용하기도 한다.

② 판단decision & planning: 자율주행을 위한 두 번째 필수 기능은 인지한 상황 정보에 따라 차량 제어 및 경로 설정을 결정하는 '판단'이다. 일반적으로 자율주행차의 판단 기술은 주어진 상황에서 차량의 경로를 생성하는 기술로 이해할 수 있다. 예를 들어 차량의 전방에 정지 차량, 보행자, 장애물 등이 인지되었을 때, 이를 피해 갈지 혹은 그대로 주행할지 등을 판단한다. 또한, 전방의 위험 상황을 회피한다고 할 때 그대로 정지할지 혹은 차선을 변경할지 등에 대한 다양한 경로 대안을 고려하여 최적의 선택을 내리게 된다. 이를 위해 자율주행 알고리즘은 수많은 경로 대안에 대해 충돌확률, 차량 내 승객 안전성, 이동 효율성 등 다양한 측면에서 효과를 분석하여 최적의 단일한 경로 대안을 산출하게 된다.

▼ 자율주행 알고리즘 아키텍처[2]



③ 제어control: 자율주행을 위한 세 번째 필수 기능은 주변 교통 상황 판단에 근거한 '제어'이다. 자율주행차의 제어는 눈, 귀와 같은 감각기관에서 수집된 정보를 두뇌가 판단하여 팔이나 다리 등을 움직이게 하는 것에 비유할 수 있으며, 이는 차량의 감속, 가속, 정지, 회전 등으로 표현되곤 한다. 차량의 제어는 자율주행의 마지막 단계로 차량의 파워트레인, 브레이크, 스티어링 등에서 수행된다. 예전에는 기계적인 힘으로 조작되던 것들이 최근에는 차량 내부 통신규격인 CAN Controller Area Network 통신에 기반해 MDPS 모터와 엔진 제어기 등이 작동하는 전자제어 방식으로 바뀌고 있다. 자율주행 시스템에서는 스마트 크루즈 컨트롤SCC, Smart Cruise Control, 차선 유지 지원 시스템LKAS, Lane Keeping Assist System 등 첨단 안전 시스템을 구축함으로써 앞에서 언급한 차량의 기계적인 장치에 기반해 더욱 안전하고 효율적인 차량 제어를 지원한다[2].

위와 같이 자율주행차의 동작 단계에 따라 자율주행 인공지능 기술을 구분하는 것과 더불어, 운전 자동화 수준에 따라 자율주행 단계를 구분하기도 한다. 미국자동차공학회SAE, Society of Automotive Engineers에서는 자율주행 단계를 Level 0부터 Level 5까지 6단계로 구분하며, Level 3부터는 인간 운전자가 아니라 자율주행 시스템이 제어 주체가 된다.

▼ 미국자동차공학회에서 정의한 운전 자동화 수준에 따른 자율주행 단계 구분

2.2 자율주행 인공지능 이슈 사례

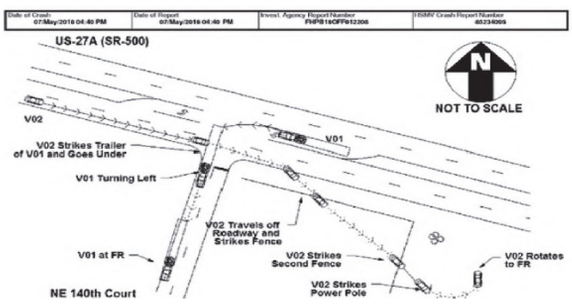
자율주행차가 상용화되면서 도로에서 다른 차량과 혼재되어 주행함에 따라, 도로 위 객체와 충돌하는 등 크고 작은 안전 사고가 발생하였다. 자율주행차의 사고 발생 원인으로는 운전자의 자동화 작동 메커니즘 이해 부족, 자율주행 기능에 대한 지나친 의존 등이 꼽힌다. 이러한 사고 외에도 스티어링휠 핸즈오프 경고를 회피하는 등 자율주행 시스템을 남용하는 사건이 종종 발생하고 있다.

안전사고 사례 1: Uber 자율주행차의 보행자 사망사고[4]

Uber 자율주행차가 무단 횡단 중인 보행자를 치어 숨지게 한 사건. LiDAR 센서에서 통해 보행자를 감지하였으나 운전 효율에 우선순위를 두어 보행자를 무시하고 달려도 되는 도로 위 장애물로 판단('18.5.)

시사점
사람의 안전이 무엇보다 우선이라는 윤리적 고려가 부족해 인명 사고를 초래함. 인지 편향을 예방하기 위해 다수의 전문가 협의를 통한 구현 및 검증이 필요함

안전사고 사례 2: Tesla 운전자 사망사고[5]



Tesla의 자율주행 모드 '오토파일럿'이 실행되던 모델 S가 대형 트레일러를 인지하지 못하고 충돌하여 운전자가 사망한 사건. 센서가 파란색 하늘과 순백색 트럭 측면을 구분하지 못하여 충돌한 것으로 파악('16.5.)

시사점
본 사고는 인지 시스템 오류에 의한 것으로, 자율주행차에 들어가는 SW의 안전성 강화를 위해 ISO 26262 등 국제표준 수정에 박차를 가해야 함

남용 사례: 자율주행기능 경고음 회피[6]

Tesla FSD[Full-Self Driving] 기능 동작 중, 운전자는 스티어링 휠에 손을 얹어 놓아야 하지만, 간단한 장치를 장착하여 경고 기능을 무력화하였음

시사점
자동화 기능을 남용한 사례로, 예상치 못한 인명 피해가 발생할 수 있음. 기능의 목적, 제한사항 등에 대한 설명을 제공하여 서비스 오남용을 방지해야 함

보안 사고 사례: 차량 해킹을 통한 제어[7]

해커 두 명이 지프 체로키의 에어컨, 스테레오, 엔진을 원격 조종한 사건. 실제로 변속장치를 조종하거나 차량이 속도를 줄이면 브레이크를 조종하는 모습을 시연하였음

시사점
차량 통신을 해킹하는 방법으로 차량 제어권이 탈취되면 큰 인명 피해와 재산 피해가 발생할 수 있음

2.3 자율주행 인공지능 신뢰성 정책 및 연구 동향

유럽, 미국, 아시아에서는 서로 다른 성격으로 자율주행 분야의 인공지능 신뢰성 관련 정책을 추진하고 있다. 유럽은 자율주행차의 안전성 및 신뢰성을 사전에 확보하고, 승인을 얻은 차량이 출시될 수 있도록 규제하는 데 중점을 둔다. 미국은 아직 법안을 마련하지 못했지만, 자율주행 관련 정책 보고서를 발행해 방향을 수립하고 있다. 아시아 3개국(한국, 일본, 중국)은 자율주행차 개발, 도입, 확산을 목표로 정책을 추진하면서 자율주행차의 윤리 및 신뢰성을 확보하기 위해 노력하고 있다. 산업계와 학계에서는 자율주행차의 안전성을 위주로 연구를 진행하고 있으며, 점차 신뢰성 확보를 위한 노력을 이어나가고 있다. 산업계에서는 자율주행차의 신뢰성을 확보하기 위한 표준을 제정해 적용, 확산하고 있고, 학계에서는 신뢰할 수 있는 자율주행 서비스 연구기관을 설립하여 기술적·사회적 문제 해결을 위한 연구를 본격적으로 시작하였다.

▼ 자율주행 분야 주요국 인공지능 신뢰성 관련 정책 동향

국가	주요 정책(연도)	특징
유엔유럽경제위원회	• 도로교통부에서 자율주행 차량 이용 관련 법적 문서 초안 작업을 위한 전문가그룹 활동 개시('21) • 자율주행시스템 및 동적 통제의 개념 규정('20) • 자율주행 레벨3 국제 규정 채택('20)	자율주행 차량 상용화에 따른 도로 안전, 취약한 도로 사용자의 안전을 보장하는 것을 목표로 기존의 도로교통 협약을 보완하는 새로운 법률 문서 준비
유럽위원회	• 자율주행 시스템 법안 초안 발표('22) • 신뢰할 수 있는 자율주행 차량 정책 보고서 발간('21)	신뢰할 수 있는 자율주행 차량의 요구사항 및 평가항목을 개발하고, 법안 초안에서 자율주행 차량의 형식 승인을 위한 평가, 감사, 테스트하는 방법을 정의
미국	• '자율주행차 종합계획' 정책 보고서 발간('21) • '자동화된 차량 기술에서 미국의 리더십 확보' 정책 보고서 발간('20) • '교통의 미래 준비' 정책 보고서 발간('18) • '자율주행 시스템' 정책 보고서 발간('17)	1-2년 간격으로 자율주행 관련 정책보고서를 발간 중이며, 최근 정책보고서는 대중을 포함한 파트너, 이해관계자가 자율주행 차량의 기능과 한계에 대한 정보에 접근할 수 있도록 촉진함
독일	• '무인자율주행차법' 연방 상원의 승인 획득('21) • 전 세계 첫 자율주행 윤리지침 마련('17) • 도로교통법 8차 개정('17)	운전자가 탑승하지 않은 상태로 운행하는 무인자율주행차는 허용되지 않았으나, 이번 개정으로 무인자율주행차를 상용화함
중국	• 자율주행 차량에 대한 국가 초안 규칙 발표('22) • 도로교통·자율주행기술 발전 및 응용 촉진에 관한 지도의견 발표('20) • 스마트 자동차 혁신 개발 전략 발표('20) • 지능형 커넥티드 차량 도로 시험 관리 기준(시험) 정부 공동 발표('18)	2025년까지 자율주행차 상용화를 목표로, 간선급행버스체계BRT, Bus Rapid Transit에서 자율주행 차량의 사용을 장려하고, 교통부가 발표한 규칙 초안에 따라 대중교통 서비스를 제공할 수 있도록 함
일본	• '차체' 안전성 확보를 위한 기술기준 검토 돌입('21) • 도로운송차량법 개정·시행('20) • 도로교통법 개정('20) • 도로법 개정에 관한 법률안을 각의에서 결정('20)	자율주행차의 보급 확산과 차세대 모빌리티의 안전성 확보
한국	• 자율주행차를 고려한 도로교통법 개정 및 시행('22) • 레벨4 자율주행차 제작·안전 가이드라인 발표('20) • 레벨3 자율주행차 상용화를 위한 제도 완비('20) • 자율주행자동차 윤리 가이드라인 발표('20)	자율주행차 산업 생태계 구축과 경쟁력 강화를 위해 규제 개선, 시범 운행지구 지정, 연구·개발 지원 등 다양한 전략을 수립 및 추진

▼ 자율주행 분야 해외 주요 산·학·연 인공지능 신뢰성 연구 동향

기관	활동 및 내용
ASAM	OpenDRIVE, OpenLABEL, OpenODD, OpenSCENARIO 등 ASAM(자동화 및 측정 시스템 표준화 협회)에 가입한 회원사의 기술전문가를 프로젝트 그룹에 파견하여 적극적으로 표준화 활동을 수행했으며, 이를 바탕으로 자율주행 차량을 위한 라벨링, 주행 시뮬레이션과 시나리오 검증을 위한 산업 표준 제정
BMW	인공지능 사용을 위한 윤리 강령(7대 원칙) 발표('20) – AI를 공공재적인 성격을 띠는 것으로 보아 평등하게 사용할 수 있게 하며, 특히 인간에게 위해를 가할 목적으로 사용해서는 안 된다는 점을 강조
영국 사우샘프턴 대학	TAS[Trustworthy Autonomous Systems] Hub – 책임 있는 자율주행 차량 데이터 등 신뢰할 수 있고 사회적으로 유익한 자율 시스템의 설계, 규제 및 운영에서 사회적·기술적 문제를 해결하기 위한 프로젝트 수행

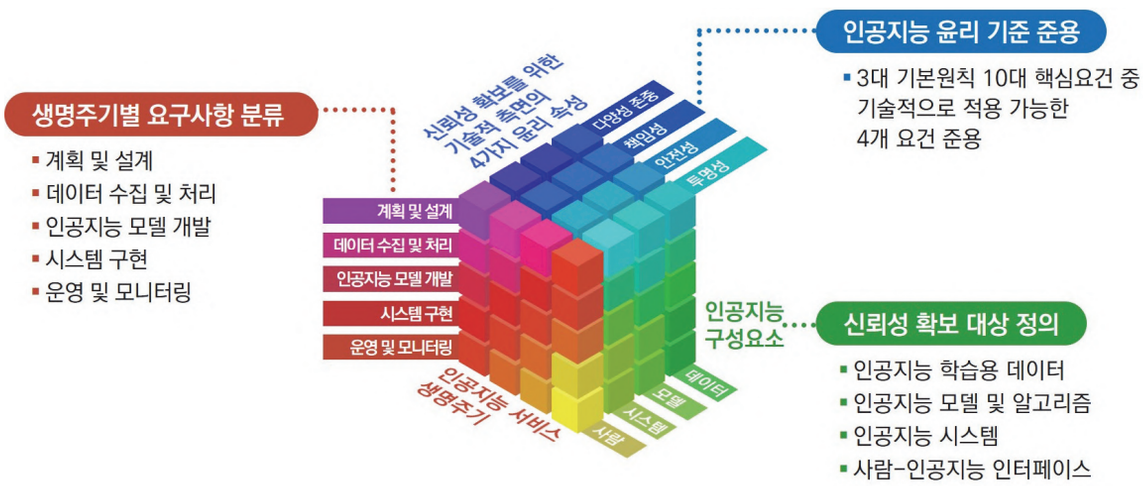
03 안내서 마련 과정

발간 배경에서 밝혔듯이, 국내외에서 인공지능 기반 자율주행차에 대한 여러 가지 문제를 고려하여 법과 제도가 마련되며 많은 기관과 기업에서 관련 가이드라인을 내놓고 있으나, 기술적 관점에서 상세한 구현 방법론을 정리한 사례는 없었다. 따라서 본 안내서에서는 자율주행차의 인공지능 개발 현장에서 데이터 과학자, 모델 개발자 등 이해관계자들이 실무 관점에서 신뢰성 확보에 참고할 수 있는 지침서 성격의 자료를 만들고자 하였다. 이를 위해, 2021년 1월부터 모든 산업 분야를 아우를 수 있는 일반 분야 안내서를 마련하기 시작했으며, 이를 기반으로 2022년에는 자율주행 분야에 특화된 안내서를 발간하였다. 자율주행 분야의 안내서 마련 과정에서는 학계 및 산업계 전문가와 실무자들을 대상으로 의견 수렴을 진행하였다. 또한, 자율주행 관련 서비스를 제공하는 기업과 협업해 안내서의 현장 적용과 컨설팅에 관한 공동 연구를 진행하여 케이스 스터디를 마련하고 피드백을 받는 과정을 거쳐 실무 활용도를 높이고자 하였다.

3.1 인공지능 신뢰성 프레임워크 적용

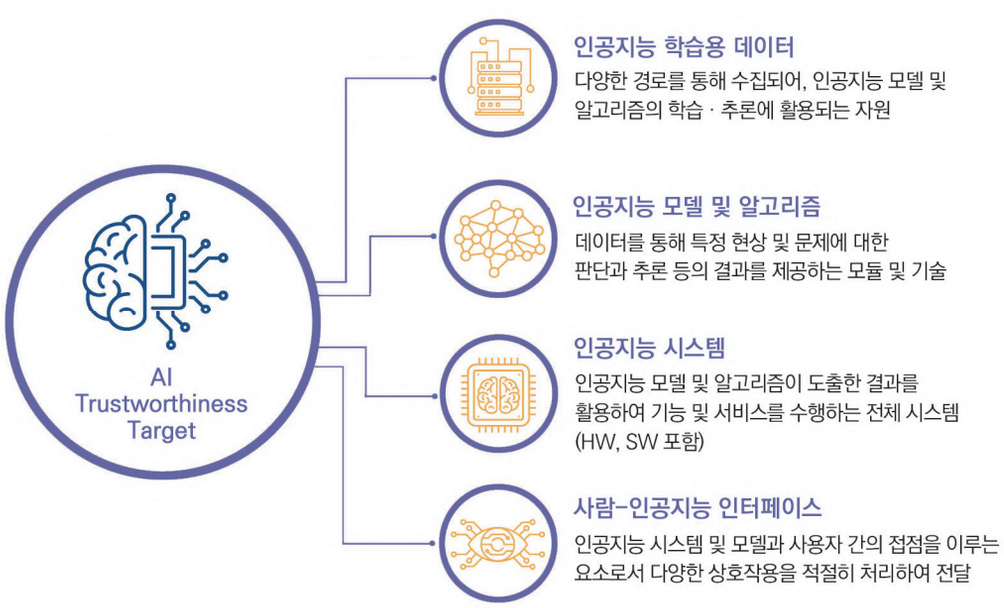
안내서 개발 과정 중 가장 우선적으로 신뢰성 확보를 위해 어떤 요소들이 실무적으로 고려되어야 하는지 탐색해보았고, 그 결과 세 가지 설계 요소를 도출하여 안내서에 반영하였다. 각 설계 요소들은 요구사항과 검증항목 마련 시 모두 반영되었으며, 이러한 접근법을 아래 그림과 같이 매트릭스 형태로 체계화하여 '인공지능 신뢰성 프레임워크'로 정의하였다. 이 프레임워크는 자율주행 분야뿐만 아니라, 일반 분야 및 타 산업 분야에도 동일하게 적용된다.

▼ 인공지능 신뢰성 프레임워크



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

▼ 인공지능 서비스 구성 요소



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

두 번째, 인공지능 서비스 생명주기는 첫 번째에서 살펴본 인공지능 서비스 구성 요소들을 구현하고 운영하는 일련의 절차를 말한다. 기존 소프트웨어 시스템에서 다루는 공학 프로세스나 생명주기와 비슷하나, 인공지능 특성상 데이터 처리 및 모델 개발 단계가 별도로 필요하며, 이외의 단계에서도 주요 활동에 대한 정의가 조금씩 달라진다. 현재 인공지능 혹은 인공지능 서비스의 생명주기는 다수의 문헌에서 6~8가지 단계로 구분한다. 대표적으로 OECD와 ISO/IEC에서 제시한 생명주기가 있는데, 본 안내서는 두 기구에서 제시한 생명주기를 대표성 있는 사례로 참고하여, 실무자들이 쉽게 활용할 수 있도록 각 생명주기 단계의 성격과 활동을 왜곡하지 않는 선에서 아래와 같이 5가지 단계로 정리하였다.

▼ 인공지능 서비스 생명주기별 주요 활동

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

03 안내서 마련 과정

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

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

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

▼ 인공지능 신뢰성 요건

신뢰성 요건	정의
다양성 존중	인공지능이 특정 개인이나 그룹에 대한 차별적이고 편향된 관행을 학습하거나 결과를 출력하지 않으며, 인종·성별·연령 등과 같은 특성과 관계없이 모든 사람이 평등하게 인공지능 기술의 혜택을 받을 수 있는 것 - 관련 속성: 공평성·공정성fairness, 정당성justice - 관련 키워드: 편향bias, 차별discrimination, 편견prejudice, 다양성diversity, 평등equality - 국제표준(ISO/IEC TR 24027:2021 - Bias in AI systems and AI aided decision making)에서는 공평성을 정의하지 않는다. 공평성은 복잡하고 문화·세대·지역 및 정치적 견해에 따라 다양하여 사회적으로나 윤리적으로 일관되게 정의하기 힘들기 때문이다.
책임성	인공지능이 생명주기 전반에 걸쳐 추론 결과에 대한 책임을 보장하기 위한 메커니즘이 마련되어 있는 것 - 관련 속성: 책무성responsibility, 감사가능성auditability, 답변가능성answerability - 관련 키워드: 책임liability - 국제표준(ISO/IEC TR 24028:2020 - Overview of Trustworthiness in artificial intelligence)에서의 정의: 엔터티의 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, 강건성·견고성robustness, 성능보장성reliability, 통제가능성·제어가능성controllability - 관련 키워드: 적대적 공격adversarial attack, 복원력resilience, 프라이버시privacy - 국제표준(ISO/IEC TR 24028:2020)에서의 정의: 용인할 수 없는 위험risk으로부터의 자유
투명성	인공지능이 추론한 결과를 인간이 이해하고 추적할 수 있으며, 인공지능이 추론한 결과임을 알 수 있는 것 - 관련 속성: 설명가능성explainability, 이해가능성understandability, 추적가능성traceability, 해석가능성interpretability - 관련 키워드: 설명가능한 인공지능XAI, eXplainable AI, 이해도comprehensibility - 국제표준(ISO/IEC TR 29119-11:2020 - Guidelines on the testing of AI-based systems)에서의 정의: 시스템에 대한 적절한 정보가 관련 이해관계자에게 제공되는 시스템의 속성

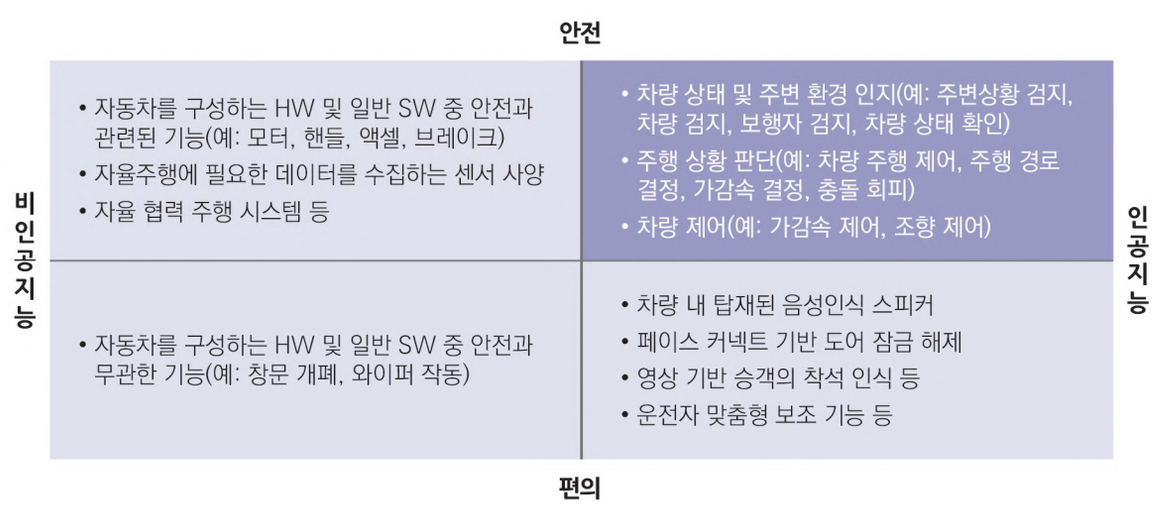
위와 같이 인공지능 신뢰성 확보를 위한 다양한 속성들이 있으며, 각 신뢰성 속성들에 대한 정의를 파악하는 것뿐만 아니라 신뢰성 속성 간의 상호의존 관계 역시 중요하게 고려되어야 한다. 예를 들어, 인공지능 서비스에 대한 과도한 투명성 요구는 프라이버시 관련 위험을 초래할 수 있다. 또한, 설명가능성만으로는 투명성을 보장하기에 부족하지만, 설명가능성은 투명성을 확보하기 위한 중요한 요소 중 하나이다. 따라서, 인공지능 신뢰성 속성에 대한 충분한 이해를 바탕으로 인공지능 서비스를 제공하는 것이 중요하며, 해당 인공지능 서비스가 고려한 속성에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

3.2 자율주행 분야 주요 고려사항 반영

본 안내서는 기술적 관점에서 상세한 방법론을 제시함으로써, 자율주행 시스템 및 서비스 개발 현장에서 실무자가 신뢰성 확보에 참고할 수 있는 실무 지침서 성격의 자료를 지향한다. 따라서 본 안내서는 일반 분야에서 다루는 구성요소 및 생명주기를 바탕으로, 인공지능의 신뢰성 확보를 위해 고려해야 할 요소들을 자율주행 분야에 특화하여 세 가지로 정리하였다.

첫 번째, 본 안내서에서 신뢰성 확보 대상으로 다룰 자율주행의 범위는 자율주행차의 모든 기술 요소를 포함하지는 않는다. 본 안내서는 인공지능 모델에 기반하여 구현된 자율주행 알고리즘(예: 인지 알고리즘), 자율주행 알고리즘이 포함된 자율주행 시스템(예: 인지 결과에 기반한 판단·제어 시스템), 자율주행 시스템에 따라 사용자에게 제공되는 서비스를 대상으로 신뢰성을 확보하기 위한 요소를 다룬다. 즉, 자율주행의 다양한 요소 중 인공지능이 직·간접적으로 적용된 요소에 한정한다.

또한, 자율주행 시스템에서 인공지능이 적용되는 기능은 크게 안전을 위한 기능과 편의를 위한 기능으로 나눌 수 있다. 본 안내서에서는 안전을 위한 기능에 적용되는 인공지능의 신뢰성 확보를 위한 내용을 다룬다. 정리하자면, 아래 그림에서 인공지능과 비인공지능, 안전 기능과 편의 기능 기준으로 영역을 구분했을 때 본 안내서에서 다루는 내용은 오른쪽 상단 영역에 해당한다.



두 번째, 자율주행 분야 인공지능의 서비스를 구성하는 4가지 요소는 아래와 같은 범위를 고려하였다. 특히, 자율주행 분야는 다른 분야에 비해 자율주행 기능이 작동 중일 때 이를 사용자나 보행자에게 적절히 알리고, 자율주행 기능이 중단되어야 하는 상황에서 운전자에게 차량 제어권을 이양하도록 알리는 등의 HMI[Human Machine Interface] 설계가 중요하다. 이를 위해 본 안내서에서는 HMI 설계 방안 및 HMI에 대한 사용자 평가 사례 등 참고할 수 있는 자료를 담기 위해 노력했으며, 이 내용은 4가지 구성요소 중 '사람-인공지능 인터페이스'에 해당한다.

▼ 자율주행 분야 인공지능의 서비스 구성요소

구성요소	설명
인공지능 학습용 데이터	차량의 카메라, 센서, 자율협력주행시스템 등을 통해 인공지능 학습 및 추론 과정에 활용하는 데이터에 편향성이 배제되었는지 검증
인공지능 모델 및 알고리즘	자율주행의 인공지능 모델 및 알고리즘으로 안전한 결과를 도출하며 악의적인 공격에 강건한지 검증
인공지능 시스템	자율주행 인공지능이 판단 및 추론한 대로 작동하는지, 오류가 발생했을 때 대비 및 대책이 존재하는지 검증
사람-인공지능 인터페이스	자율주행 시스템이 운전자, 보행자, 운영자에게 자율주행 기능 작동 상태 또는 안전과 관련된 정보를 이해하기 쉽게 전달하고 오작동을 방지할 수 있는지 검증

세 번째, 자율주행 분야 인공지능 서비스 생명주기에 따라 아래와 같은 활동들을 고려하였다. 자율주행은 사람의 안전과 직결되는 응용 기술로, 신뢰성 요건 중 안전성의 중요도가 높다. 따라서 차량 및 자율주행 안전 관련 표준인 ISO 26262 - Functional safety 및 ISO 21448에서 다루는 내용을 '계획 및 설계' 에서 위험관리에 참고할 수 있도록 제시하였다.

▼ 자율주행 분야 인공지능 서비스 생명주기별 주요 활동

생명주기 단계	주요 활동
1. 계획 및 설계	• 비즈니스 이해, 표준 기반 관리 체계 구축 • 위험 발생 가능성 분석(hazard), 윤리적 고려사항 분석 • 자율주행 시스템 안전 동작 관련 정의(예: 운행가능영역, 제약조건, 비상대응, 사고 후 처리) • 사이버 위협이나 취약점으로 인한 위험 최소화, 위험 저감 조치 및 최소화 대책 설계 마련 • HMI, 인공지능 모델 설계
2. 데이터 수집 및 처리	• 주행 데이터에 대한 요건(예: 날씨, 시간대, 도로 특성, 주행 속도) 설정 • 주행 데이터 수집·전처리·가공 및 품질 확보 • 학습·테스트·검증용 데이터셋 구축
3. 인공지능 모델 개발	• 자율주행 알고리즘 아키텍처(인지·판단·제어) 모델 구현 • 모델 성능평가 및 개선, 설명가능성 확보
4. 시스템 구현	• 위험 저감 조치 및 최소화 대책 적용 • 위험요소 확인, 정상 동작 검증(안전성 검증)
5. 운영 및 모니터링	• 자율주행시스템의 올바른 이해와 사용을 위한 교육훈련 • 추적 관리 체계 문서화 • 모델의 성능 유지·개선(주기적 또는 특정 이벤트 시에 모델의 재학습)

3.3 요구사항 및 검증항목 도출

다음 단계로 자율주행 분야 인공지능에 관한 구체적인 요구사항과 검증항목을 도출하였다. 우선 표준화기구, 기술단체, 국제기구, 주요 국가에서 자율주행 분야 인공지능 신뢰성 확보를 위해 발표한 정책, 권고안, 표준을 기반으로 준수해야 할 기술적 요구사항을 도출하고 구체화하였다. 자율주행 분야는 안전성이 중요한 만큼, 차량과 자율주행 안전 관련 표준인 ISO 26262 및 ISO 21448에서 다루는 내용들을 주의 깊게 살펴보았다. 이와 함께 《Trustworthy Autonomous Vehicles》('21.12), 《자율주행차 윤리 가이드라인》('20.12), 《자율주행차 제작안전 가이드라인》('20.12) 등 국내외에서 자율주행 분야 인공지능 신뢰성 확보를 목적으로 발표된 문헌들을 검토하였다. 검토 과정에서 개발 안내서에 필요한 내용은 반영하고 중복된 내용은 제거하거나 통합하였다. 참고문헌은 다음과 같다.

▼ 자율주행 분야 인공지능 신뢰성 관련 주요 참고문헌

기관명	발간 연월	권고 및 표준안 명
대한민국	2020.12	자율주행차 제작안전 가이드라인
	2020.12	자율주행차 윤리 가이드라인
	2020.11	국가 인공지능(AI) 윤리기준
유럽위원회	2021.12	Trustworthy Autonomous Vehicles
독일	2017.06	German Federal Ministry of Transport and Digital Infrastructure, Ethics Commission: Automated and Connected Driving
국제표준화기구 (ISO/IEC)	2022.06	ISO 21448:2022 – Road vehicles – Safety of the intended functionality
	2020.05	ISO/IEC TR 24028:2020 – AI – Overview of Trustworthiness in artificial intelligence
	2018.12	ISO 26262:2018 – Road vehicles – Functional safety

이를 통해 최종 도출한 요구사항은 아래 표와 같으며, 인공지능 윤리의 핵심 요건에 대응시킨 결과도 함께 표시했다.

▼ 인공지능 신뢰성 확보를 위한 기술적 요구사항과 윤리 요건 매칭 결과

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

3.4 현장 적용 및 전문가 의견 수렴

신뢰성 확보를 위한 요구사항을 도출한 후에는 각 항목을 기술적 타당성, 효용성 및 포괄성 등의 관점에서 검토한 후 고도화했다. 각각의 세부 검증항목이 요구사항에 해당하는 내용이 맞는지(타당성), 개발 현장에서 실무적으로 활용 가능한 내용인지(효용성), 검증을 위한 내용들이 과거부터 지금까지 연구 내용을 폭넓게 포함하는지(포괄성) 확인했다. 이를 위해 자율주행 인공지능 분야 전문가가 참여하여 직접 검토하고 자문했으며, 다양한 검토 의견을 수렴하여 반영했다. 자율주행 인공지능 분야 전문가에는 기업의 기획자, 개발 프로젝트 리더, 교수 등 산업계 및 학계의 연구자 등 분야를 가리지 않고 다양한 의견을 수렴하였다. 또한, 자율주행 관련 서비스를 제공하는 기업과의 협업을 통해 안내서의 현장 적용과 컨설팅 공동 연구를 진행하여 케이스 스터디를 마련하고 피드백을 받는 과정을 거쳐 실무 활용도를 높이고자 했다.

04 안내서 활용 대상

본 안내서는 자율주행 분야 인공지능 서비스 구현 과정에서 직·간접으로 관련되거나 영향을 주는 모든 조직과 개인을 포함한 이해관계자가 참고할 수 있지만, 주요 대상은 특히 업무상 기술적 관점에서 신뢰성을 신경 써야 하는 시스템 기획자, 시스템 엔지니어, 데이터 공급자, 데이터 과학자, 인공지능 모델 개발자 등이다. 이들이 자율주행 분야 인공지능 생명주기의 각 단계마다 신뢰성을 확보하기 위해 검토해야 할 주요 요구사항은 다음과 같다.

▼ 자율주행 분야 인공지능 생명주기 단계별 신뢰성 주요 행위자

생명주기 단계	주요 행위자	주요 요구사항
1. 계획 및 설계	• 시스템 기획자 • 비즈니스 결정권자 • 품질 관리자 • 시스템 운영자	– 인공지능 시스템 전체 생명주기에 걸친 신뢰성 확보 요구사항 검토 및 적용방안 마련 – 인공지능 시스템 영역, 자율주행 위험요건 식별 및 대응방안 검토
2. 데이터 수집 및 처리	• 데이터 과학자 • 데이터 공급자 • 도메인 전문가 • 보안 분석가	– 학습 데이터 확보 과정에서 발생할 수 있는 데이터 오류 및 편향에 대한 관리방안 확보 – 학습 데이터 수집 시나리오 설계 및 시나리오별 정제 기준 마련
3. 인공지능 모델 개발	• 인공지능 모델 개발자 • 시스템 엔지니어 • 데이터 과학자	– 학습 모델의 편향적인 추론 결과나 공격에 대한 대응방안 마련 – 학습 모델의 추론 결과에 대한 해석방안 제공
4. 시스템 구현	• 시스템 엔지니어 • 인공지능 모델 개발자 • 품질 관리자 • HMI 전문가	– 인공지능 시스템 개발 시 발생 가능한 편향이나 오류에 대한 대응방안 마련 – 인공지능 서비스가 도출한 결과에 대한 사용자 친화적인[user-friendly] 설명 제공 – 인공지능의 작동 상태를 사용자에게 알리기 위한 인터페이스 구현
5. 운영 및 모니터링	• 시스템 엔지니어 • 시스템 운영자 • 인공지능 모델 개발자 • 비즈니스 결정권자	– 서비스 목적 및 한계를 사용자에게 알려 오남용 예방 – 인공지능 시스템 문제 발생 시 원인 추적을 통한 대응 방안 마련

04 안내서 활용 대상

요구사항별 협력 대상은 아래와 같으며, 개발 안내서를 활용하는 서비스 및 기업 환경에 따라 상이할 수 있으므로 참고 사항으로 활용되길 바란다.

▼ 인공지능 신뢰성 확보를 위한 요구사항별 활용 권장 대상

요구사항	대표 행위자	협력 대상
요구사항 01 인공지능 시스템에 대한 위험관리 계획 및 수행	• 시스템 기획자	• 비즈니스 결정권자 • 시스템 엔지니어 • 시스템 운영자
요구사항 02 인공지능 거버넌스 시스템 구성	• 시스템 기획자	• 비즈니스 결정권자 • 시스템 운영자
요구사항 03 인공지능 시스템의 신뢰성 테스트 계획 수립	• 품질 관리자	• 시스템 기획자 • 시스템 엔지니어 • 비즈니스 결정권자
요구사항 04 데이터의 활용을 위한 상세 정보 제공	• 데이터 과학자	• 데이터 공급자 • 인공지능 모델 개발자
요구사항 05 데이터 강건성 확보를 위한 이상 데이터 점검	• 데이터 과학자	• 데이터 공급자 • 인공지능 모델 개발자 • 보안 분석가
요구사항 06 수집 및 가공된 학습 데이터의 편향 제거	• 데이터 공급자	• 데이터 과학자 • 인공지능 모델 개발자
요구사항 07 오픈소스 라이브러리의 보안성 및 호환성 확보	• 인공지능 모델 개발자	• 시스템 엔지니어
요구사항 08 인공지능 모델의 편향 제거	• 인공지능 모델 개발자	• 시스템 엔지니어
요구사항 09 인공지능 모델 공격에 대한 방어 대책 수립	• 인공지능 모델 개발자	• 시스템 엔지니어
요구사항 10 인공지능 모델 명세 및 추론 결과에 대한 설명 제공	• 인공지능 모델 개발자	• 시스템 엔지니어 • 시스템 운영자
요구사항 11 인공지능 시스템 구현 시 발생 가능한 편향 제거	• 시스템 엔지니어	• 시스템 운영자 • 인공지능 모델 개발자
요구사항 12 인공지능 시스템의 안전 모드 구현 및 문제발생 알림 절차 수립	• 시스템 엔지니어	• 시스템 운영자 • 인공지능 모델 개발자 • HMI 전문가
요구사항 13 인공지능 시스템의 설명에 대한 사용자의 이해도 제고	• 시스템 엔지니어	• 시스템 운영자 • 인공지능 모델 개발자 • 비즈니스 결정권자 • HMI 전문가
요구사항 14 인공지능 시스템의 추적가능성 및 변경이력 확보	• 시스템 엔지니어	• 인공지능 모델 개발자 • 데이터 과학자
요구사항 15 서비스 제공 범위 및 상호작용 대상에 대한 설명 제공	• 시스템 엔지니어	• 시스템 기획자 • 시스템 운영자 • 인공지능 모델 개발자 • 비즈니스 결정권자

05 안내서 활용 방법

05 안내서 활용 방법

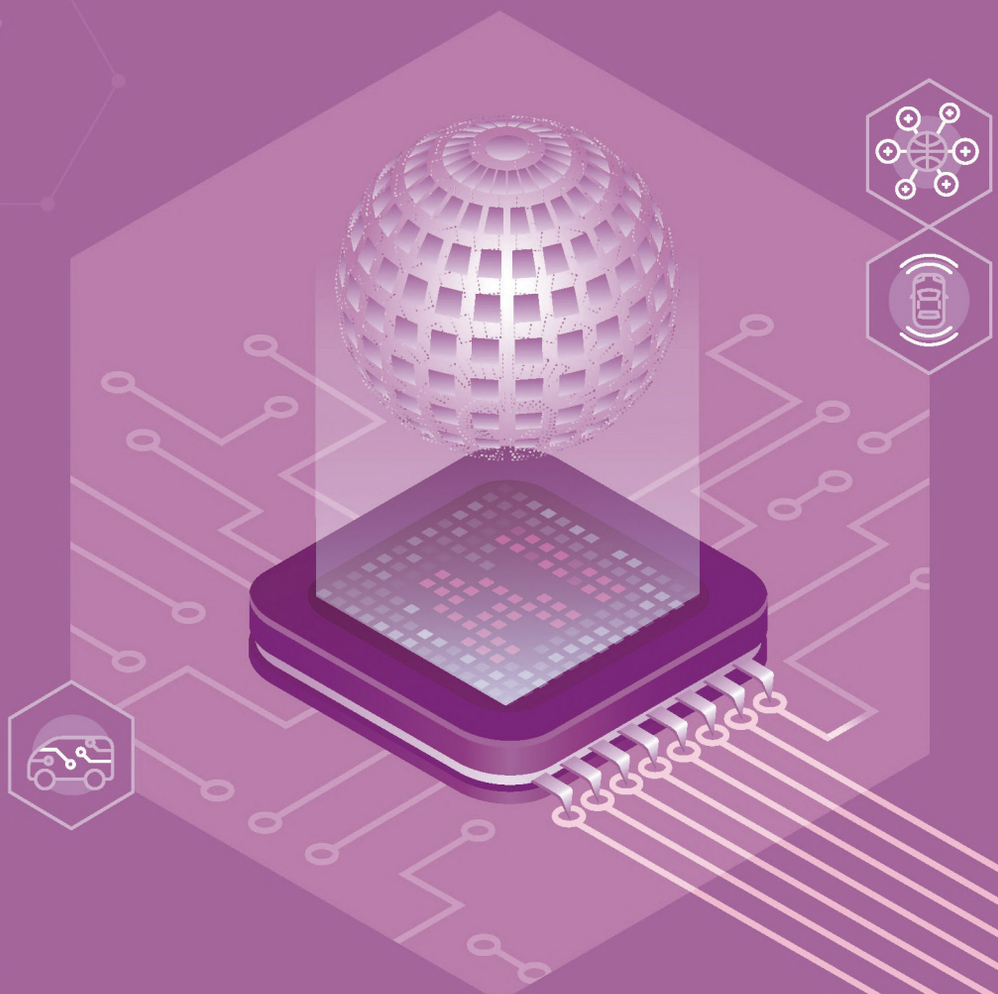
본 안내서는 범용성을 갖추고자 인공지능 신뢰성 관점에서 기술적 고려가 필요한 요구사항 및 검증항목을 포괄적으로 수록하였다. 따라서, 기업 내부의 기술 역량, 제품 특성 등을 고려하여 적절한 요구사항과 검증항목을 선택하여 적용하고, 기업에서 제공 중인 서비스의 환경에 맞게 신뢰성 확보를 위한 참고자료로 활용하길 바란다. 더불어, 인공지능 신뢰성 확보를 위해서는 기술적 측면 외에도 윤리, 개인정보보호와 같은 법·제도적 측면도 함께 요구된다. 그러므로 본 안내서를 활용하기에 앞서 인공지능 윤리적 고려사항 점검을 위한 〈인공지능 윤리기준 실천을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

안내서는 다음과 같은 절차로 활용할 수 있다.

① **자율주행 서비스 위험 영향 분석**: 자율주행 서비스 도입을 고려·개발을 원하는 경우 서비스의 활용 목적과 범위, 사고 위험 및 사고 발생 시 사회적 파급도와 책임 등 여러 측면에서 위험 영향을 분석하여야 한다. 특히 자율주행 인공지능의 활용에 따른 위험이 없는 유형인지를 판단해야 하는데, 시스템 결함으로 인한 작은 사고라도 인명 피해나 주변 환경에 피해를 줄 수 있다. 따라서 영향 분석 과정에서 비즈니스 결정권자, 기획자, 개발자 및 시스템 운영자 등이 함께 논의에 참여하여 다양한 관점에서의 분석을 수행할 것을 권장한다.

② **요구사항 선정**: '①'의 분석 내용을 토대로 개발 안내서 요구사항과 세부 요구사항 본문을 참고하여 인공지능 서비스에서 신뢰성 확보를 위해 필요한 요구사항을 선정한다. 자율주행 제어 시스템의 경우 사람의 안전과 직결되므로, 가능한 모든 요구사항을 선정할 것을 권장한다. 이 과정에서 시스템 기획자 및 개발자 등 요구사항별 활용 권장 대상(대표 행위자 및 협력 대상)이 협의해야하며, 만약 불필요하다고 판단된 요구사항의 경우 'N/A[Not Applicable]'를 표시하여 점검 대상에서 제외할 수 있다.

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



PART 2

요구사항 및 검증항목

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

목차

생명주기	요구사항 및 체크리스트	
1 계획 및 설계	요구사항 01	인공지능 시스템에 대한 위험관리 계획 및 수행 ····· 34
	01-1	인공지능 시스템 생명주기에 걸쳐 나타날 수 있는 위험 요소를 분석하였는가?
	01-1a	인공지능 시스템의 위험 요소를 도출하고 이의 파급효과를 파악하였는가?
	01-2	위험 요소를 제거 및 방지하거나 영향을 완화하기 위한 방안을 마련하였는가?
	01-2a	위험 요소 제거 방안을 도출하고 파급효과가 감소하였는지 확인하였는가?
	요구사항 02	인공지능 거버넌스governance 체계 구성 ····· 39
	02-1	인공지능 거버넌스에 대한 지침 및 규정을 수립하였는가?
	02-1a	내부적으로 준수해야 할 인공지능 거버넌스에 대한 지침 및 규정을 마련하였는가?
	02-2	인공지능 거버넌스를 위한 조직을 구성하고 인력 구성에 대해 검토하였는가?
	02-2a	인공지능 거버넌스를 위한 조직을 구성하였는가?
	02-2b	인공지능 거버넌스를 위한 조직은 충분히 훈련된 인력으로 구성하였는가?
	02-3	인공지능 거버넌스 체계가 올바로 이행되고 있는지 감독하고 있는가?
	02-3a	인공지능 거버넌스에 대한 내부 지침 및 규정 준수 여부를 감독하고 있는가?
	02-4	인공지능 거버넌스 조직이 신규 및 기존 시스템의 차이점을 분석하였는가?
	02-4a	신규 인공지능 시스템 도입 전, 기존 시스템의 대체 필요성 등을 분석하였는가?
	요구사항 03	인공지능 시스템의 신뢰성 테스트 계획 수립 ····· 45
	03-1	인공지능 시스템의 특성을 고려한 테스트 환경을 설계하였는가?
	03-1a	테스트 환경 결정 시 각 환경에서 테스트 가능한 주행 시나리오를 고려하였는가?
	03-1b	시뮬레이터 및 주행시험장 등 테스트 환경을 확보하였는가?
	03-2	인공지능 시스템의 테스트 설계에 필요한 협의 체계를 구성하였는가?
	03-2a	인공지능 시스템의 기대 출력을 결정하기 위한 협의 체계를 구성하였는가?
	03-2b	설명가능성 및 해석가능성 확인을 위한 사용자 평가단을 구성하였는가?
2 데이터 수집 및 처리	요구사항 04	데이터의 활용을 위한 상세 정보 제공 ····· 51
	04-1	데이터의 명확한 이해와 활용을 지원하는 상세한 정보를 제공하는가?
	04-1a	정제 전과 후의 데이터 특성을 설명하였는가?
	04-1b	학습 데이터와 메타데이터metadata를 구분하였으며, 각각에 대한 명세자료를 확보하였는가?
	04-1c	보호변수protective attribute의 선정 이유 및 반영 여부를 설명하였는가?
	04-1d	라벨링 작업자를 위해 교육을 시행하고 작업 가이드 문서를 마련하였는가?
	04-2	데이터의 출처는 기록 및 관리되고 있는가?
	04-2a	신뢰할 수 있는 출처로부터 제공되는 데이터셋을 사용하였는가?
	04-2b	오픈소스 데이터셋을 활용하는 경우, 출처를 명시하였는가?

목차

생명주기	요구사항 및 체크리스트	
2 데이터 수집 및 처리	요구사항 05	데이터 강건성 확보를 위한 이상abnormal 데이터 점검 ····· 63
	05-1	이상 데이터의 식별 및 정상 여부를 점검하였는가?
	05-1a	전체 학습용 데이터 분포를 시각화하여 발생 가능한 오류들을 확인하였는가?
	05-1b	학습 데이터 이상값 식별 기법을 적용하였는가?
	05-2	데이터 공격에 대한 방어 수단을 강구하였는가?
	05-2a	데이터 중독poisoning, 회피evasion 등 공격에 대한 방어 대책을 마련하였는가?
	요구사항 06	수집 및 가공된 학습 데이터의 편향 제거 ····· 71
	06-1	데이터 수집 시, 인적·물리적 요인으로 인한 편향 완화 방안을 마련하였는가?
	06-1a	인적 편향을 제거하기 위한 절차적, 기술적 수단을 적용하였는가?
	06-1b	데이터의 다양성 확보를 위해 수집 시 여러 차량 제원을 활용하였는가?
	06-1c	하드웨어로 인해 발생할 수 있는 데이터의 편향을 점검하였는가?
	06-2	학습에 사용되는 특성feature을 분석하고 선정 기준을 마련하였는가?
	06-2a	보호변수protective attribute 선정 시 충분한 분석을 수행하였는가?
	06-2b	편향을 발생시킬 수 있는 특성의 영향력을 완화하였는가?
	06-2c	데이터 전처리 시 특성이 과도하게 제거되었는지 검토하였는가?
	06-3	데이터 라벨링 시, 발생 가능한 편향을 확인하고 방지하였는가?
	06-3a	데이터 라벨링 기준을 명확히 수립하고 작업자에게 제공하였는가?
	06-3b	다양한 데이터 라벨링 작업자를 섭외하기 위해 노력하였는가?
	06-3c	다양한 데이터 라벨링 검수자를 확보하기 위해 노력하였는가?
	06-4	데이터의 편향 방지를 위한 샘플링을 수행하였는가?
	06-4a	편향 방지를 위한 샘플링 기법을 적용하였는가?
3 인공지능 모델 개발	요구사항 07	오픈소스 라이브러리의 보안성 및 호환성 확보 ····· 84
	07-1	오픈소스 라이브러리의 안정성을 확인하였는가?
	07-1a	활성화된 오픈소스 라이브러리를 사용하였는가?
	07-2	오픈소스 라이브러리의 위험 요소는 관리되고 있는가?
	07-2a	사용 중인 오픈소스 라이브러리의 라이선스 준수사항을 이행하였는가?
	07-2b	사용 중인 오픈소스 라이브러리의 호환성 및 보안취약점을 확인하였는가?
	요구사항 08	인공지능 모델의 편향 제거 ····· 90
	08-1	모델 편향을 제거하는 기법을 적용하였는가?
	08-1a	개발하려는 모델에 맞게 편향제거 기법을 선택하였는가?
	08-1b	편향성 평가 및 모니터링을 위한 정량적 지표를 선정하고 관리하는가?

목차

생명주기	요구사항 및 체크리스트	
3 인공지능 모델 개발	요구사항 09	인공지능 모델 공격에 대한 방어 대책 수립 ····· 94
	09-1	모델 추출 공격model extraction attack에 대한 방어 방안을 수립하였는가?
	09-1a	모델 추출 공격에 대비하는 방어 기법을 적용하였는가?
	09-2	모델 회피 공격model evasion attack에 대한 방어 방안을 수립하였는가?
	09-2a	모델 회피 공격에 대비하는 방어 기법을 적용하였는가?
	요구사항 10	인공지능 모델 명세 및 추론 결과에 대한 설명 제공 ····· 98
	10-1	사용자가 모델 추론 결과의 도출 과정을 수용할 수 있도록 근거를 제공하는가?
	10-1a	XAIeXplainable AI 기술 적용이 가능한 경우, 인공지능 모델의 추론 결과를 설명하기 위한 기법 적용에 대해 검토하였는가?
	10-1b	XAI 기술 적용이 불가능한 경우, 기법 적용 이외의 대안을 마련하였는가?
	10-2	인공지능 모델 상세 문서를 통해 모델의 명세를 투명하게 제공하는가?
	10-2a	시스템 개발 과정과 모델 작동 방식에 대한 세부 정보가 설명된 문서를 작성하였는가?
	10-3	필요 시, 인공지능 모델 추론 결과에 대한 설명을 제공하는가?
	10-3a	모델 추론 결과에 대한 설명이 필요한지 검토하였는가?
	10-3b	사용자에게 인공지능 모델 추론 결과에 대한 설명을 제공하였는가?
4 시스템 구현	요구사항 11	인공지능 시스템 구현 시 발생 가능한 편향 제거 ····· 108
	11-1	소스 코드 및 사용자 인터페이스로 인한 편향을 제거하기 위해 노력하였는가?
	11-1a	데이터 접근 방식 구현과정 등 소스 코드에서의 편향 발생 가능성을 확인하였는가?
	11-1b	사용자 인터페이스user interface 및 상호작용 방식으로 인한 편향을 확인하였는가?
	요구사항 12	인공지능 시스템의 안전 모드 구현 및 문제발생 알림 절차 수립 ····· 113
	12-1	공격, 성능 저하 및 사회적 이슈 등의 문제 발생 시 대응 가능한 안전 모드를 적용하는가?
	12-1a	문제 상황에 대한 예외 처리 정책이 마련되어 있는가?
	12-1b	인공지능 시스템의 보안 강화를 위한 보안 기법을 적용하였는가?
	12-1c	인공지능 시스템의 의사결정으로 인한 파급효과가 크고 불확실성이 높은 경우, 사람의 개입을 고려하였는가?
	12-1d	예상되는 사용자 오류에 대한 안내 및 대응을 제공하는가?
	12-1e	객체 및 주행상황 인지 오류를 방지하기 위해 다중 센서 기술을 적용하였는가?
	12-2	인공지능 시스템에서 문제가 발생할 경우, 시스템은 이를 운영자에게 전달하는 기능을 수행하는가?
	12-2a	편견, 차별 등 윤리적 문제에 대한 알림 절차를 수립하였는가?
	12-2b	시스템 성능 저하를 평가하기 위한 지표 및 절차를 설정하고 알림 절차를 수립하였는가?

목차

생명주기	요구사항 및 체크리스트	
4 시스템 구현	요구사항 13	인공지능 시스템의 설명에 대한 사용자의 이해도 제고 ····· 125
	13-1	인공지능 시스템 사용자의 특성[user characteristics]과 제약사항을 분석하였는가?
	13-1a	사용자 특성에 따른 세부 고려사항을 분석하였는가?
	13-2	사용자 특성에 따른 충분한 설명을 제공하는가?
	13-2a	사용자 특성에 따른 설명 평가의 기준을 수립하였는가?
	13-2b	사용자가 이해하기 어려운 전문 용어 사용을 지양하였는가?
	13-2c	사용자의 구체적인 행동과 이해를 이끌어낼 수 있도록 명확한 표현을 사용하였는가?
	13-2d	설명이 필요한 위치와 타이밍은 적절한가?
	13-2e	사용자 경험을 평가할 수 있는 다양한 사용자 조사 기법을 활용하였는가?
5 운영 및 모니터링	요구사항 14	인공지능 시스템의 추적가능성 및 변경이력 확보 ····· 133
	14-1	인공지능 시스템의 의사결정에 대한 추적 방안을 수립하였는가?
	14-1a	인공지능 시스템의 의사결정에 대한 기여도 추적 방안을 확보하였는가?
	14-1b	인공지능 시스템의 의사결정 추적을 위한 로그 수집 기능을 구현하였는가?
	14-1c	지속적인 사용자 경험 모니터링을 위해 사용자 로그를 수집 및 관리하고 있는가?
	14-2	학습 데이터의 변경 이력을 확보하고, 데이터 변경이 미치는 영향을 관리하였는가?
	14-2a	데이터 흐름 및 계보[lineage]를 추적하기 위한 조치를 마련하였는가?
	14-2b	데이터 소스 변경에 대한 모니터링 방안을 확보하였는가?
	14-2c	데이터 변경 시, 버전관리를 수행하였는가?
	14-2d	데이터 변경 시, 이해관계자를 위한 정보를 제공하는가?
	14-2e	신규 데이터 확보 시, 인공지능 모델의 성능평가를 재수행하였는가?
	요구사항 15	서비스 제공 범위 및 상호작용 대상에 대한 설명 제공 ····· 140
	15-1	인공지능 서비스의 올바른 사용을 유도하기 위한 설명을 제공하는가?
	15-1a	서비스의 목적과 목표에 대한 설명을 제공하는가?
	15-1b	서비스의 한계와 범위에 대한 설명을 제공하는가?
	15-2	상호작용의 대상을 명확히 설명하는가?
	15-2a	사용자가 인공지능과 상호작용하고 있다는 사실을 사용자에게 명확히 설명하였는가?

01 계획 및 설계

책임성 투명성

요구사항 01 인공지능 시스템에 대한 위험관리 계획 및 수행

대표 행위자 | 시스템 기획자 협력 대상 | 비즈니스 결정권자 시스템 엔지니어 시스템 운영자

- 자율주행 분야는 운영환경이 다양해 다른 분야에 비해 높은 불확실성을 내포하고 있으며, 위험 상황 발생 시 환경·사람·재산상 피해를 야기할 수 있어 위험관리는 필수적이다. 따라서, 위험을 사전에 인식하고, 각 위험의 크기(심각성 및 파급 효과)를 분석하여 대응 방안을 마련하는 등 위험관리를 수행한다.

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

Yes ☐ No ☐ N/A ☐

해당여부 판단

자율주행 알고리즘 중 인지·판단·제어를 위한 인공지능 모델 및 시스템을 포함하여 개발하는 경우, 본 항목을 고려하여 만족 여부를 판단하십시오.

- 위험관리는 위험 인식identification, 위험 분석analysis, 위험 평가evaluation, 위험 대응treatment으로 구분한다. 이러한 네 가지 활동을 생명주기 단계별로 지속해서 반복하여 수행하여 신뢰성을 확보하고 위험을 제거 및 방지하여야 하고, 이에 대한 개념 및 정의는 ISO 31000:2018 – Risk management[8]에서 제공하고 있다. 또한, 인공지능의 신뢰성 관점에서 살펴보아야 할 위험 요소를 인식, 분석 및 평가하는 방법론은 ISO/IEC 24028:2020[9]과 ISO/IEC 23894:2023 – Guidance on risk management[10]에서 제공하고 있다.
- 자율주행 분야에서는 위험으로 인한 결과가 최종적으로 인명 피해 및 재산상 피해로 나타나므로 위험 분석이 더욱 세심하게 이루어져야 한다. 자율주행 분야의 안전에 관련된 표준으로는 ISO 26262[11]와 ISO 21448[12] 등이 있다. 이는 기존의 차량 안전을 확보하는 체계와 함께 자율주행 시스템 외적으로 발생하는 위험 요소를 도출하고 의도한 기능이 안전하게 동작하도록 그 대응 방안에 대해 다룬다. 물론, 이러한 위험 요소 분석 및 대응은 인공지능 생명주기 전반에 걸쳐 반복적으로 이루어져야 한다.

참고 자율주행 분야의 안전 관련 표준 비교 (ISO 26262 vs. ISO 21448)

- ISO 26262: 센서 및 액추에이터actuator 고장 등 시스템 오류로 인해 발생하는 위험을 방지하기 위한 표준
- ISO 21448: 시스템 오류 없이 설계 의도대로 동작하는 경우에도, 외부 환경에 따른 센서 인식률 저하 등의 문제로 발생할 수 있는 위험을 방지하기 위한 표준

두 표준에서 다루는 위험 요인 비교[13]

분류	위험 요인	관련 표준
시스템	전장E/E, Electrical and/or Electronic 시스템의 결함	ISO 26262
	예측 가능한 오용reasonably foreseeable misuse의 유무와 관계없이 성능 제한 또는 상황 인식 부족	ISO 21448
외부 요인	활성 인프라 및 차량에서 차량 통신, 외부 장치, 클라우드 서비스에 미치는 영향	ISO 26262
	차량 주변의 환경(예: 보행자, 타 차량, 도로 인프라, 날씨 등 환경 조건, 전파 방해)	ISO 21448 ISO 26262

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

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템의 위험 요소는 소프트웨어 및 하드웨어 기반 시스템에서 발생할 수 있는 요소와는 다르다. 소프트웨어의 결함 및 오류, 하드웨어의 노후화 및 마모 등과 달리 데이터 기반 분석의 특성으로 나타날 수 있는 편향, 설명 미제공, 모델에 대한 공격 등의 위험 요소를 도출해야 한다. 이러한 요소의 분류와 주요 내용은 ISO/IEC 23894.2와 ISO/IEC 24028에 제시되어 있다.
- 또한, 인공지능 알고리즘을 반영한 자율주행 시스템을 개발할 때는 다음과 같은 위험 요소가 발생할 수 있음을 인지하고, 이로 인한 파급효과(예: 인명피해, 재산피해)를 파악해야 한다.
  - ✓ 인공지능 시스템의 처리 성능 저하로 인한 차량의 장애물 충돌(회피 시간 미확보)
  - ✓ 인공지능 시스템의 인지 성능 저하, 모델 공격(예: 모델 추출, 모델 회피 공격)으로 인한 교통법규 위반, 장애물 충돌, 인명 피해 등 사고 발생
  - ✓ 인공지능 시스템의 보행자 인지 성능 차이로 인해 인종에 따라 서로 다른 차량 행동 발생(윤리적 차별)
- 위험 요소를 도출한 후에는 이를 야기하는 원인과 이에 따라 발생 가능한 결과를 분석해야 한다. 발생 가능한 결과란 위 사례와 같이 사회적으로 부정적인 영향을 미치는 현상과 사고를 의미하며, 이에 해당하는 것으로는 인체에 위해를 가하는 사고(예: 장애물 충돌, 인명 피해), 사회적 문제를 야기할 수 있는 차별적인 현상(예: 인종, 연령대에 따라 상이한 자율주행 차량 반응) 등이 해당한다. ISO 21448에서는 이와 같은 위험을 야기하는 원인과 이에 따른 영향을 제시하였으며, 이 내용은 아래의 참고에 정리되어 있다.

- 위험 요소의 발생으로 인한 결과는 심각도와 발생빈도와 등의 척도를 기준으로 위험의 크기 또는 수준을 평가할 수 있다. 이는 위험 요소의 파급효과를 의미한다. 위험 요소를 평가해 파급효과가 큰 위험 요소를 최우선으로 대응 방안을 마련해야 한다. 다만, 파급효과를 산정 및 평가하는 과정에서 심각도와 발생빈도뿐만 아니라, 상황에 맞는 척도를 도입하여 조합할 수 있다.

참고 ISO 21448에서 제시한 위험원 및 영향의 예시[14]

분류	원인	영향
광원	역광	- 이미지 센서의 광포화로 인한 장애물을 오인식 또는 미인식
	낮은 조도	- 장애물 감지 불가로 인한 미인식
날씨	비	- 빗방울에 의한 왜곡으로 인한 미인식 또는 오인식 - 차선의 미인식 또는 오인식
	눈	- 시야 차단으로 인한 장애물 및 차선을 오인식 또는 미인식
	안개	- 짧은 가시거리로 인한 장애물 및 차선을 오인식 또는 미인식
도로 상태	곡선로	- 전방의 차량을 옆 차선의 차량으로 오인식
	경사로	- 광고판에 있는 사람 그림을 보행자로 오인식
	결빙	- 보이지 않는 도로의 결빙으로 인한 통제 불가 가능성
	노면 반사	- 왜곡된 이미지로 인한 장애물 오인식
외부 환경	진흙	- 시야 차단으로 인한 장애물 및 차선의 오인식 또는 미인식

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

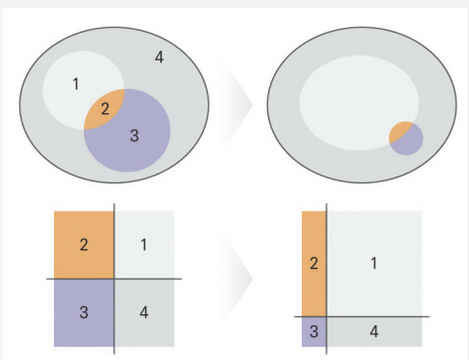
Yes ☐ No ☐ N/A ☐

해당여부 판단

01-1에서 위험 요소를 분석한 경우, 본 항목을 고려하여 만족 여부를 판단하십시오.

- 위험 대응 방안은 01-1에서 분석된 위험 요소별로 마련해야 한다. 이에 해당하는 대응 방안은 위험 요소의 원인을 제거하여 인명 피해와 사고를 미연에 방지하거나, 사고로 인한 파급효과와 부정적인 영향을 최소화하는 것 등이다.
- 대응 방안이란, 구현·운영 방식 등 절차, 소프트웨어·하드웨어 기능, 모델 학습 기법·전략 등 기술적으로 적용하는 모든 방법을 의미한다. ISO/IEC 24028:2020에 대응 방안의 분류가 제공되어 있다. 이를 고려하여 인공지능을 구현하는 모든 이해관계자는 위험 요소에 대한 대응 방안을 마련하고, 위험이 제거 또는 완화되었는지 확인해야 한다.
- ISO 21448에서는 자율주행 분야의 위험관리 활동을 통해 되도록 많은 위험 요소와 시나리오를 인지하고, 해당 요소와 시나리오를 안전하게 처리하는 것을 목적으로 위험관리 방안을 안내하였다.

참고 ISO 21448에서 제시한 위험관리 활동 목표



영역	설명
1	시나리오를 인지하고 있으며, 해당 시나리오를 안전하게 처리 가능한 상황
2	시나리오를 인지하고 있으나, 해당 시나리오를 안전하게 처리 불가한 상황
3	인지하지 못한 시나리오이며, 해당 시나리오를 안전하게 처리 불가한 상황
4	인지하지 못한 시나리오이나, 해당 시나리오를 안전하게 처리 가능한 상황

- 영역 2, 3을 최소화하는 동시에, 영역 1을 최대화하는 것이 ISO 21448의 개념이자 최종 목표이다.

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

Yes ☐ No ☐ N/A ☐

- 위험 요소의 발생을 막으려면 구현·운영 방식, 소프트웨어·하드웨어 기능, 모델 학습 기법·전략 등 기술적인 방법론을 도출해야 한다. 이러한 방법론에 대한 분류와 개략적인 내용은 ISO/IEC 24028:2020에 제시되어 있다.
- 위험 완화 활동에서는 파급효과가 가장 큰 위험 요소를 우선순위에 두고 대응해야 하며, 위험의 심각도가 높은 경우에는 인공지능 시스템의 판단 결과에 따라 사람의 개입도 고려해야 한다. 그 이후에는 파급효과를 재평가하여 위험 요소가 실제로 제거, 방지되었는지 혹은 이의 영향이 완화되었는지 확인해야 한다.
- ISO 21448에서는 자율주행 시스템의 위험을 완화하기 위해, 관련 기능이 안전하게 동작할 수 있도록 하는 방안을 제시한다. 따라서 이 표준에서 언급한 방안을 활용하여 01-1 에서 분석한 위험 요소를 완화할 수 있도록 대응 방안을 마련해야 한다.

참고 ISO 21448에서 제시한 위험 요소 제거 방안

(1) 위험을 회피하기 위한 시스템 개선: 위험 요소를 회피하고 위험 요소에 따른 영향을 감소하기 위해서 센서와 차량의 성능 및 정확도를 향상하고, 인식 및 알고리즘 성능을 향상해야 하며, 시험가능성testability을 개선해야 한다.

(2) 위험 완화를 위한 기능 구축: 특정 사용 사례에 발생할 수 있는 위험을 완화하기 위한 기능을 구축해야 한다. 예를 들어, 카메라가 외부 요인으로 인해 가려질 경우, 레이더 등 다른 센서를 사용하여 지속적으로 운영될 수 있게 하는 등의 기능이 필요하다. 이와 관련된 자세한 내용은 다중 센서 기술 적용과 관련된 검증항목 12-1e 를 참고할 수 있다.

(3) 위험 발생 시 운전자에게 차량 제어권 인계: 위급한 상황에서 운전자에게 주행 권한을 넘겨주는 기능이 필요하다. 이를 위해 사람과 차량 간 상호작용, 경고를 위한 기능을 개선해야 한다.

(4) 예측할 수 있는 오용 감소: 예측할 수 있는 오용을 감소·완화해야 하기 위해 운전자에게 설계된 기능에 대한 설명을 제공해야 하고, 사람과 차량 간 상호작용 기능을 개선해야 한다. 또 운전자가 운전대에서 손을 뗐을 때 경고하는 기능 등 모니터링 및 경고 시스템을 실행해야 한다.

위험 요소의 원인과 이에 따른 대응 방안 예시

요소	위험 발생 원인(예)	대응 방안(예)
전장 시스템	전장 시스템 성능의 한계를 초과	– 시스템의 성능이 저하되어 자율주행 기능이 비활성화되었음을 운전자에게 알리고 차량 제어권을 운전자에게 인계 – 정해진 절차에 따라 안전하게 기능 종료 – 성능을 저하시키되 기능을 유지
운전자	예측할 수 있는 오용	– 운전자의 부주의한 작동에 대한 대응법 제공 – 운전자에게 올바른 작동법을 알림 – 비정상적인 작동 감지 및 경고

다양성 존중 | 책임성 | 안전성 | 투명성

요구사항 02 인공지능 거버넌스governance 체계 구성

대표 행위자 | 시스템 기획자 협력 대상 | 비즈니스 결정권자 | 시스템 운영자

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

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

02-1 인공지능 거버넌스에 대한 지침 및 규정을 수립하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

자율주행 인공지능 모델 및 시스템이 윤리적 영향을 미치거나 지재권 분쟁 등의 우려가 있는 경우, 본 항목을 고려하여 만족 여부를 판단하십시오.

- 인공지능과 관련된 조직에서는 인공지능 시스템 신뢰성 확보를 위한 거버넌스 체계를 구성할 필요가 있다. 인공지능 시스템은 학습이나 추론 과정에서 윤리 및 지식재산권IP, Intellectual Property 관련 문제, 보안 및 개인정보 이슈가 발생할 수 있기 때문이다. 이러한 위험 요소에 대비하기 위해 내부적으로 인공지능 거버넌스에 대한 지침 및 규정을 수립해야 한다.
- NIST의 AI RMFRisk Management Framework에서는 인공지능 시스템 생명주기에 따라 내부 규정, 절차, 과정 및 실제 행위가 투명하고 효율적으로 이루어져야 한다고 언급한다. 즉, 인공지능과 관련된 법, 규제 관련 요구사항이 이해·관리되어 문서화하고, 위험관리 절차와 산출물이 체계를 통해 투명하게 관리되어야 한다.
- 내부적으로 수립해야 할 규정은 활용 측면에 따라 크게 두 가지로 구분하여 마련할 수 있다.
  - ✓ 첫째, 인공지능 관련 법, 규제, 정책, 표준 및 지침을 채택·정리하여 내부적으로 이행해야 할 지침 및 규정을 수립해야 한다.
  - ✓ 둘째, 인공지능 시스템 생명주기에 따른 조직의 역할과 책임을 명확하게 문서화해야 한다.
- 국토교통부에서 발간한 자율주행자동차 윤리 가이드라인에서는 관련 법규나 인증기준, 생명 윤리, 정보통신 윤리, 공학 윤리를 준수해야 한다고 언급하고 있다. 또한, 운행의 법적·윤리적 기준에 대한 투명성을 확보하기 위하여 운행 관련 내용을 기록하고 보관할 수 있도록 제작해야 한다고 언급하고 있다.

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

Yes ☐ No ☐ N/A ☐

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

참고 자율주행자동차 윤리 가이드라인(국토교통부)의 공통 원칙(예)

- 자율주행자동차 윤리 가이드라인에서는 윤리적 측면에서 고려해야 할 기본가치에 대한 공통 원칙을 제시하고 있다. 아래의 항목들을 참고하여 내부적으로 준수해야 할 인공지능 거버넌스에 대한 지침 및 규정을 마련할 수 있다.

No.	설명
1.1	자율주행자동차는 인간의 존엄성, 국제법적으로 인정된 인권과 자유, 프라이버시 및 문화적 다양성을 존중하고, 인간을 성별, 나이, 인종, 장애 등을 이유로 차별하지 않으며, 인간의 법과 관습에 의한 판단과 통제에 따르도록 설계, 제작, 관리되어야 한다.
1.2	자율주행자동차는 인간의 행복과 이익의 증진을 위한 수단으로서 인간의 안전하고 편리하며 자유로운 이동권을 보장하고, 타인의 권리와 자유를 침해하지 않도록 설계, 제작, 관리되어야 한다.
1.3	자율주행자동차는 자동차 사고로 인해 발생할 손실을 최소화하고, 무엇보다 인간의 생명을 우선하도록 설계, 제작, 관리되어야 한다. 또한, 손실을 최소화하는 과정에서 인간을 성별, 나이, 종교 등 개인적 차이 등을 이유로 차별하지 않고, 교통 약자를 고려하는 방식으로 작동하도록 설계, 제작, 관리되어야 한다.

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

Yes ☐ No ☐ N/A ☐

해당여부 판단

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

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

02-2a 인공지능 거버넌스를 위한 조직을 구성하였는가?

Yes ☐ No ☐ N/A ☐

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

참고 인공지능 거버넌스 체계를 수립한 사례

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

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

Yes ☐ No ☐ N/A ☐

- 인공지능 거버넌스 담당 조직은 자신이 맡은 역할과 책임에 대해 충분히 인식한 인력으로 구성해야 한다. 이들은 인공지능 생명주기에 걸친 모든 프로세스의 중심적인 역할로서, 담당자가 이를 충분히 인식한 후 책임지고 관리해야 인공지능 시스템의 신뢰성을 확보할 수 있기 때문이다.
- 인공지능 거버넌스 담당 조직은 각기 다른 배경과 전문지식을 기반으로 충분히 숙련된 인력으로 구성해야 한다. 특히, 규정을 마련하는 역할을 맡은 담당자는 인공지능 윤리 및 신뢰성 분야의 원칙, 가이드라인, 표준 등에 대한 폭넓은 전문지식을 갖춰야 하며, 이를 적절히 해석하여 조직 업무에 적용하기 위한 기술력과 타 업무 담당자와의 의사소통 역량이 필요하다. 또한, 정의된 규정을 실행하고 관리하기 위해 각 담당자에게 관련 교육을 제공하여 충분히 훈련해야 한다.
- 국토교통부에서 발간한《자율주행자동차 윤리 가이드라인》에서는 차량의 서비스 제공 및 이용 과정에서 기술과 서비스에 대해 관련 주체들을 대상으로 안내, 교육, 훈련 등이 실시되어야 한다고 언급하였다.

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

Yes ☐ No ☐ N/A ☐

해당여부 판단

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

- 인공지능 거버넌스 체계를 운영하는 주체는 운영 결과에 대한 책임을 져야 하고, 이 책임은 위임할 수 없다. 따라서 인공지능 거버넌스 운영 담당자는 조직이 내부 지침 및 규정을 준수하는지에 대해 감독해야 한다.
- ISO/IEC 38507:2022 – Governance implications of the use of artificial intelligence by organizations[15]에서 인공지능 거버넌스 체계는 인공지능 시스템에서 발생할 수 있는 위험에 따라 인공지능 시스템의 설계 및 사용에 대한 감독을 수행해야 한다고 언급하고 있다. 즉, 인공지능 거버넌스 체계를 통해 수립한 내부 규정을 조직이 적절히 이행하고 있는지 감독해야 한다.

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

Yes ☐ No ☐ N/A ☐

- 인공지능 거버넌스 담당자는 인공지능 시스템 생명주기에 따라 조직이 내부 규정을 준수함을 확인 및 감독해야 한다. 또한, 신뢰성 있는 인공지능 시스템을 목표로 적절히 관리 및 통제됨을 관련 이해관계자에게 입증해야 한다.
- 특히, 인공지능 시스템 위험관리와 관련된 내부 규정을 이행하는지 감독함으로써 인공지능 시스템의 잠재적 위험으로부터 조직 및 이해관계자를 보호하고 조직의 역량을 향상할 수 있다.
- 따라서 인공지능 거버넌스 체계에서 감독을 담당하는 조직은 인공지능 시스템에 대한 이해를 바탕으로 역할에 대한 책임 및 권한을 명확히 인지하여 인공지능 시스템 생명주기에 걸쳐 모든 규정이 이행되는지 감독해야 한다.

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

Yes ☐ No ☐ N/A ☐

해당여부 판단

신규 자율주행 시스템의 개발을 계획하는 경우, 본 항목을 고려하여 만족 여부를 판단하십시오.

- 자율주행 시스템 및 서비스를 위한 인공지능 모델 및 시스템 개발이 범람할 경우, 성능상의 이슈가 발생하고 서비스 사용자에게 혼란을 가중시킬 뿐만 아니라 시스템 개발 및 유지보수에 불필요한 예산 사용을 초래할 수 있다.
- 신규 자율주행 시스템의 개발을 계획할 때는 신뢰성, 성능, 효율, 안전성, 비용 등의 다양한 측면에서 기존에 운영 중인 시스템과 비교해 유사한지, 개선이 가능한지 등을 분석한 결과를 기반으로 개발 필요성을 확인해야 한다.

02-4a 신규 인공지능 시스템 도입 전, 기존 시스템의 대체 필요성 등을 분석하였는가?

Yes ☐ No ☐ N/A ☐

- 오래전부터 전장 장치, 기계 부품 등으로 개발되어온 전통적인 차량 시스템(예: ABS[Anti-lock Brake System])에 인공지능을 도입해 대체 또는 고도화하려면, 먼저 그 필요성을 충분히 분석해야 한다. 불확실성[uncertainty]이 높은 신규 인공지능 시스템을 도입해 안정적으로 운영되고 있는 기존 시스템에 불필요한 위험 요소를 생성하는 등, 전체 시스템의 안전성에 직·간접적인 영향을 끼치지 않기 위함이다.
- 자율주행을 위한 신규 인공지능 시스템을 도입·분석 과정에서 주요 이해관계자들의 의견 교류가 반드시 진행되어야 한다. 이 과정에서 객관적인 기준과 근거를 마련하고, 검증을 수행하여 기존 시스템의 신뢰성, 성능, 안전성, 비용 등 다양한 측면을 개선할 수 있는 방향으로 추진되어야 한다.

안전성 투명성

요구사항 03 인공지능 시스템의 신뢰성 테스트 계획 수립

대표 행위자 | 품질 관리자 협력 대상 | 시스템 기획자 시스템 엔지니어 비즈니스 결정권자

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

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

03-1 인공지능 시스템의 특성을 고려한 테스트 환경을 설계하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

위험 분석 결과에 따라, 사고 발생 가능성 및 오동작으로 인한 파급력이 예상되는 경우 본 항목을 고려하여 만족 여부를 판단하십시오.

- 자율주행 시스템은 주행 시나리오의 위험도에 따라 가상테스트 및 실환경 테스트를 모두 고려해야 한다. 정확한 테스트를 위해서는 실환경 테스트를 수행하는 것이 적절하지만, 테스트는 합리적인 시간 및 비용 범위 내에서 수행되어야 하므로 운영 조건이 매우 복잡한 시스템이라면 실환경 테스트가 적절하지 않을 수 있다. 또한, 자율주행 시스템에 실환경 테스트를 적용한다면 환경의 개체(예: 차량, 건물, 동물, 인간)에 손상을 주는 위험한 상황이 발생할 우려가 있으므로, 이 경우에는 가상테스트를 수행해야 한다.
- 따라서 주행 시나리오를 고려하여 적절한 테스트 환경을 결정해야 하며, 계획 및 설계 단계에서 적절한 테스트 환경(예: 시뮬레이터, 주행시험장)을 확보해두어야 한다.

03-1a 테스트 환경 결정 시 각 환경에서 테스트 가능한 주행 시나리오를 고려하였는가?

Yes ☐ No ☐ N/A ☐

- 운영환경의 제약, 기능의 다양성, 성능 저하 요소 등 매개변수가 많은 자율주행 분야에서는 테스트 스위트 test suite 수가 거의 무한해질 수 있다. 따라서 모든 시나리오를 실환경 테스트로 수행하기 보다는 시뮬레이터를 활용한 가상테스트도 함께 고려해야 하며, 이때 주행 시나리오에 따라 상이한 테스트 환경을 결정해야 한다.
- 테스트 환경은 주행 시나리오에 포함된 대상 객체와 상황에 따라 달라져야 한다. 예를 들어, 낙하물 인식에 따른 차선 변경 시나리오는 실환경에서 테스트할 수 있지만, 보행자 인식에 따른 정차 시나리오는 실환경에서 테스트할 수 없기 때문이다. 이처럼, 테스트할 때 환경의 개체에 손상을 줄 위험이 있는 시나리오는 가상테스트 환경을 고려해야 한다.

참고 테스트 환경별 통제 가능한 주행 시나리오 예시[16]

Use Case 국내 S사의 테스트 환경별 주행 시나리오 설계 사례(일부 발췌)

환경	구분	코드	시나리오	목표
시뮬레이션	edge case	S01TC01	도로변 주차 차량 사이 어린이 출현	자차량은 안전거리 내에서 어린이를 감지하고 충돌을 방지하기 위해 안전하게 주행
		S01TC02	주행 차선 내 장애물	자차량은 안전거리 내에서 떨어진 물체를 감지하고 충돌을 방지하기 위해 안전하게 주행
	직선로	S02TC01	전방 차량 주행 중 급정지	전방 차량이 급정거하였을 때, 시스템의 상황 대처 능력 확인
		S02TC02	전방 차량 추종 주행 중 타겟 차량 Cut-in	자차량의 전방 차량 추종 주행 중, 타겟 차량의 Cut-in에 따른 시스템 상황대처 능력 확인
주행시험장	장애 시험	S03TC01	횡단보도의 보행자	자차량은 횡단보도에서 보행자를 감지하고 충돌을 피하기 위해 정지
		S03TC02	전방 차량 주행 중 급정지	전방 차량이 급정거하였을 때, 시스템의 상황 대처 능력 확인
		S03TC03	전방 차량 추종 주행 중 타겟 차량 Cut-in	자차량의 전방 차량 추종 주행 중, 타겟 차량의 Cut-in에 따른 시스템 상황대처 능력 확인

주행 시나리오별 시뮬레이션 및 주행시험장 활용 사례

03-1b 시뮬레이터 및 주행시험장 등 테스트 환경을 확보하였는가?

Yes No N/A
☐ ☐ ☐

- 자율주행 분야의 시뮬레이터를 확보할 때는 운영환경에 대한 대표성을 가지는지 확인해야 한다. 예를 들어, 보행자 회피 테스트는 높은 수준의 이미지 대표성이 요구된다.
- 일부 도메인은 오픈소스로 공개된 시뮬레이터가 있어, 개발할 인공지능 시스템에 적합하다면 이를 활용할 수 있다. ISO/IEC TR 29119-11:2020[17]에서는 자율주행차 테스트용 시뮬레이터의 예시로 NVIDIA의 DRIVE Constellation을 언급하였다.
- 국내 자율주행차 주행시험장으로는 화성의 K-City, 성남의 판교제로시티, 상암의 5G자율주행 테스트베드, 여주의 한국도로공사 여주시험도로, 연천의 기상환경 재현 도로성능평가 실험시설, 대구의 지능형자동차부품진흥원 대구 주행시험장, 충북대학교 자율주행차 성능시험장, 새만금 주행시험장 등이 있다[18]. 주행시험장마다 제공되는 시설물 및 시험환경이 상이하므로, 개발할 자율주행 시스템에 적합한 주행시험장을 선정해 예약해야 한다.

참고 NVIDIA의 DRIVE Constellation 데모 영상[19]

- 가상 환경 기반의 자율주행차용 안전 테스트 데모 영상

주행 도로 위 어린이 출현 환경 시뮬레이션 사례

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

Yes ☐ No ☐ N/A ☐

해당여부 판단

사용자에게 자율주행 시스템의 출력 결과에 대한 설명이 필요한 경우, 본 항목을 고려하여 만족 여부를 판단하십시오.

- 대부분의 인공지능 시스템은 복잡도가 높아 재현가능성reproducibility이 떨어져 투명성 확보에 어려움을 갖는다. 또한, 시스템의 복잡도는 기대 출력을 결정하는 테스트 오라클test oracle에 문제가 되기도 한다. 이에 따라 테스트가 통과 또는 실패했는지 그 여부를 판단하기 어렵다.
- 인공지능 시스템의 추론 결과에 대한 설명이 필요한 시스템이라면, 시스템 출력을 확인하는 대상 사용자에 따라 설명가능성*에 대한 평가 기준이 달라질 수 있다. 그리고 인공지능의 작동 방식을 이해하는 정도인 해석가능성interpretability의 평가 기준 역시 대상 사용자에 의존한다.

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

- 따라서 인공지능 시스템의 기대 출력에 대한 결정이나, 시스템 출력에 대한 설명가능성 및 해석가능성 평가 기준 수립에 필요한 협의 체계를 구축해 협의체를 구성하고, 구성원 간 합의 도출을 통해 테스트를 설계하는 방식이 적절하다.

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

Yes ☐ No ☐ N/A ☐

- 테스트 오라클 문제의 극복이 필요한 인공지능 시스템이라면, 시스템의 기대 출력을 결정하기 위해 해당 도메인의 내·외부 전문가로 구성된 협의체를 구성하여야 한다. 이때 기대 출력을 결정하기 위해 여러 전문가가 동의하는 데 시간이 걸릴 수 있음을 인지하여야 한다.
- 협의체 전문가들은 하나의 입력에 대해 각자 다른 기대 출력을 예상할 수도 있다. 그러므로 협의체 운영 전 전문가 합의를 위한 승인 기준을 미리 결정해두어야 한다. 예를 들어, 특정 기대 출력에 대한 전문가 3인 중 2인 이상 동의 시 승인하는 등의 방법이 있다.

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

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템 출력에 대한 설명이 필요한 경우, 시스템의 설명가능성과 해석가능성을 테스트하기 위해서는 인공지능 시스템의 대상 사용자가 시스템의 출력과 작동 방식을 얼마나 쉽게 이해하는지 확인하여야 한다.
- 따라서 사용자 평가단을 구성하여 설명을 어떤 난이도로 제공할지 결정하고, 이를 모델 및 시스템 구현 시 반영해야 한다. 이를 위해, 계획 및 설계 단계에서 대상 사용자를 명확히 정의한 후 사용자 평가단을 구성해야 한다.
  - ✓ 예를 들어, 인간-기계 인터페이스HMI, Human-Machine Interface를 통해 자율주행 시스템의 출력을 제공하는 경우 운전자에 따라 서로 다른 입장에서 해석되거나 오해가 발생할 수 있다. 따라서 다양한 배경의 운전자로 이루어진 평가단을 구성하여 해석가능성을 확인하는 절차가 필요하다.
- 사용자 평가단의 평가 결과에 따라 테스트의 통과 및 실패 여부를 결정할 기준을 마련하는 것이 필요하다. 예를 들어, 평균 점수가 일정 점수 이상일 때 통과를 결정하는 등의 정량적 기준 마련이나, 평균 점수 계산 시 절사평균의 활용 여부 등의 산출 기준 마련 등이 있다.

02 데이터 수집 및 처리

책임성 투명성

요구사항 04 데이터의 활용을 위한 상세 정보 제공

대표 행위자 | 데이터 과학자 협력 대상 | 데이터 공급자 인공지능 모델 개발자

- 자율주행을 위한 알고리즘 중 인지·판단·제어를 위한 인공지능용 데이터는 차량 운행 또는 시뮬레이션 환경 등을 통해 자체적으로 수집하거나, 오픈소스 데이터셋을 활용할 수 있다. 이때 활용하는 데이터셋은 각 이해관계자의 작업 등을 위해 운행 데이터 수집 기준, 정제 기준, 학습 데이터 선별 시나리오 등 충분한 정보를 제공·기록·관리함으로써 사전에 데이터를 추가 구축하거나 문제 발생 시 원인을 추적할 수 있는 기반을 제공한다.

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

Yes ☐ No ☐ N/A ☐

해당여부 판단

자율주행 인공지능 모델 개발을 위해 데이터셋을 직접 구축하거나, 원시 데이터 구매 후 자체 정제 또는 공개 데이터셋에 향후 추가 데이터 수집을 고려하는 경우 본 항목을 고려하여 만족 여부를 판단하십시오.

- 메타데이터metadata는 데이터를 설명하는 데이터로 정의할 수 있으며, 원시 데이터raw data의 특징들을 메타데이터에 기록하여 향후 데이터를 재활용하는 상황이나 동일한 형식의 데이터를 추가로 수집해야 할 때 데이터에 대한 정보를 전달한다.
- 데이터의 수집 및 정제 시 원시 데이터와 정제 후 데이터에 대한 정보를 이해관계자들에게 제공하여 데이터에 대한 이해를 도모하고, 학습 데이터와 메타데이터를 정의하여 추가로 데이터를 수집해야 할 때 필요한 정보를 제공한다. 이해관계자들에게 전달되어야 할 정보의 예로는 수집 데이터의 출처와 형식, 데이터 수집·정제·가공 방법, 데이터 라이선스, 편향 유발 가능성이 있는 보호변수protective attribute 등이 있다.
- 라벨링 작업자에게는 데이터 라벨링 작업 시 라벨링 도구 활용 방법, 라벨링 유의사항 등의 정보를 제공한다. 특히, 2종 이상의 센서를 정합fusion할 때는 전문 도구의 사용법 등 더욱 상세한 가이드를 제공해야 한다.

04-1a 정제 전과 후의 데이터 특성을 설명하였는가?

Yes ☐ No ☐ N/A ☐

- 데이터 정제작업은 라벨링 작업 전 학습 데이터 구축을 위한 데이터의 선별 및 처리 단계로서, 정제 과정을 거친 데이터만을 사용하는 사용자는 원시 데이터의 특성을 정확하게 파악할 수 없다. 따라서 향후 추가 데이터의 수집 가능성을 고려하여 정제를 위한 관련 정보와 정제 전과 후의 데이터 특성이 설명되어야 한다.
- 자율주행 분야에서 원시 데이터를 정제하는 것은 최종적으로 인공지능 모델을 학습시킬 수 있는 데이터를 준비하는 과정으로 인공지능 모델의 학습 목표와 연관 있는 데이터는 남기고, 의미를 알 수 없는 데이터 등은 제거한다. 현재 시점에서 자율주행 알고리즘 아키텍처 중 인지 시스템을 위한 인공지능 모델이 가장 활발하게 개발되고 있으며, 이미지와 포인트 클라우드 데이터가 학습하는 데 수집 및 활용되고 있다. 자율주행 분야에서 각 데이터를 정제하는 기준의 예시는 다음과 같다.
  - ✓ 샘플링 주기 정제 기준: 이미지 및 포인트 클라우드 데이터의 초 단위 데이터 추출
  - ✓ 의미 정제 기준: 주행 환경 및 주행·사고 상황 등을 조합한 시나리오에 따른 시점의 장면 이미지
  - ✓ 노이즈 정제 기준: 흔들림이 심하여 알 수 없는 이미지, 너무 밝거나 너무 어둡게 나온 이미지
  - ✓ 객체 정제 기준: 기준 크기보다 작은 이미지 또는 포인트 클라우드 객체만 있는 시점 데이터, 인지 대상 객체가 존재하지 않는 시점 데이터
  - ✓ 비식별화 정제 기준: 이미지 데이터 내 사람의 얼굴, 번호판 등 개인정보 비식별화
- 정제 작업을 진행하는 중 입력 데이터는 원시 데이터 그대로 보존되거나, 필요에 따라 변경되는 특성feature이 있다. 특성이 변경되는 예시는 다음과 같다.
  - ✓ 원시 데이터는 도로 주행 이미지 데이터에 포함된 사람의 얼굴, 차량 번호판 등 개인정보가 식별될 때 비식별화(예: 블러링, 모자이크 처리) 과정에서 특성이 변경됨
  - ✓ 원시 데이터는 도로 주행 이미지 데이터의 해상도가 목표 크기보다 클 때 조절하는 과정에서 특성이 변경됨
- 원시 데이터는 차량에 기반한 수집 환경, 수집 장치(센서 등) 및 위치 등의 정보를 제공하여 원시 데이터의 수집 상황과 환경에 대한 이해를 돕는다. 다음은 정제 전 원시 데이터의 특징으로 설명하는 항목의 예시이다.
  - ✓ 차량 관련 설명 항목: 수집한 차량에 대한 일반 정보(차종, 연식 등 차량의 크기와 모양을 알 수 있는 정보), 부착한 센서의 종류, 센서의 사양(예: 해상도, 채널 수), 센서 설치 위치, 위치 정보의 기준 좌표계 등
  - ✓ 수집 목표 객체 설명 항목: 동적 객체(예: 자동차, 보행자, 이륜차)
  - ✓ 환경적 설명 항목: 수집한 지역(예: 지리적 위치, 도로 종류), 날씨, 수집한 시간대, 수집한 장소(예: 외부, 주차장), 교통환경(예: 교차로, 도로 폭, 곡선도로, 중앙분리대 유무)
  - ✓ 통계적 설명 항목: 수집한 원시 데이터 시간 등

- 수집한 전체 원시 데이터 중 극히 일부 데이터가 학습용 데이터셋을 위해 정제된다[20]. 국내에서도 사고 발생 유형, 생성 가능한 시나리오의 유형 수를 분석하고, 현실적으로 취득할 수 있는 시나리오 100개를 선별하여 해당하는 상황의 학습용 데이터셋을 구축하였다(06-1a 참고자료)[21]. 다음은 정제 후 학습 데이터의 특징으로 설명할 수 있는 항목의 예시이다.
  - ✓ 데이터 선별 및 처리 설명 항목: 선별 시나리오 기준 등
  - ✓ 통계적 설명 항목: 선별 시나리오별 학습 데이터 수, 동적 객체 수 등
  - ✓ 센서 퓨전 데이터 활용 설명 항목: 각 센서를 부착한 위치 정보의 기준 좌표계에 따른 캘리브레이션 calibration 정보, 서로 다른 센서의 정합을 위한 타이밍 정보 등

04-1b 학습 데이터와 메타데이터metadata를 구분하고 각 명세자료를 확보하였는가? Yes ☐ No ☐ N/A ☐

- 자율주행 알고리즘 아키텍처[22,23]에서 단계별 인공지능 서비스를 적용할 수 있으나, 본 개발 안내서의 발간일을 기준으로 하여 다수의 오픈 데이터셋이 해결하려는 부분은 인지 분야이다. 인지를 위한 인공지능 모델의 학습 데이터는 이미지, 포인트 클라우드 형식이 주를 이루며, 메타데이터는 JSONJavaScript Object Notation 또는 XMLeXtensible Markup Language 형식으로 작성된다.

자율주행 오픈 데이터셋별 알고리즘 아키텍처 구분

데이터셋 배포자	데이터 명	자율주행 알고리즘 아키텍처		
		인지	판단	제어
Berkeley Artificial Intelligence Research Lab	BDD100K	O (차선, 객체 등)	X	X
A project of Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago	The KITTI Vision Benchmark Suite	O (자동차, 보행자 등)	X	X
AI Hub	특이 도로 환경 주행 데이터	O (표지판, 신호등, 차량 등)	X	X
	다양한 기상 상황 주행	O (자동차, 날씨, 오토바이, 객체 등)	O	X
	차량 및 사람 인지 영상	O (표지판, 신호등, 차량 등)	X	X
	주차 장애물 인지 영상	O (주차 기둥, 울타리, 러버콘, 객체 등)	X	X
	자율주행 버스 개발 노선 주행 이미지	O (자동차, 교통표지판, 객체 등)	X	X
	도로 주행 영상	O (자동차, 신호등, 노면표시 등)	X	X
Waymo	WaymoOpen Dataset	O (자동차, 자전거 타는 사람, 객체 등)	X	X
woven planet	Level 5	O (차선, 객체 등)	X	X

- 인지 분야에서는 주변 상황 검지(예: 도로 영역, 차선), 차량 및 보행자 검지 등을 위주로 오픈 데이터셋이 활발하게 배포되고 있다.
- 이러한 서비스를 개발하기 위한 데이터 수집 센서, 데이터 타입, 학습을 위한 어노테이션annotation 타입을 정리하면 다음과 같다.

인지 시스템에서 활용되는 데이터 종류

데이터 수집 센서	데이터 타입	어노테이션 타입
라이다LiDAR	360도 포인트 클라우드	3D 바운딩 박스cuboid , 3D 키포인트
단·장거리 레이더RADAR	360도 포인트 클라우드	3D 바운딩 박스cuboid
RGB 카메라	이미지	2D 바운딩 박스, 시맨틱 세그멘테이션, 2D 키포인트, 폴리라인(또는 폴리곤polygon)
Night Vision 카메라	이미지	
초음파	숫자	-
GPSGlobal Positioning System	숫자	-

- 본 안내서에서 다루는 데이터는 앞서 정리한 자율주행 분야 인공지능 서비스를 대상으로 하며, 자율주행 알고리즘을 개발할 때 활용하는 학습 데이터와 메타데이터의 명세 내용도 이에 기반하여 필요한 명세 항목을 식별한다.
- 앞서 정리한 데이터 타입과 어노테이션 타입은 일반 분야의 학습 데이터와 메타데이터의 명세 내용과 많은 부분에서 유사하다. 다만, 자율주행 분야에서는 주행 환경과 객체를 실시간으로 인지해야 하므로 더욱 안전성 있는 인공지능 알고리즘을 개발해야 한다. 이를 위해 학습 데이터와 메타데이터의 명세 항목에 아래와 같은 추가 정보를 제공하여 활용할 수 있도록 한다.
  - ✓ 객체: 상대속도, 절대 속도, 거리, 진행 각도 등
  - ✓ 센서: 캘리브레이션 파라미터, 센서 설치 위치 좌표계(절대 좌표계, 차량 좌표계) 등

참고 AI Hub의 '강건한 융합 센서 객체 인식 자율주행 데이터' 학습 데이터 명세서 예시

<table>
<tr><td colspan="2">데이터 명</td><td>강건한robust 융합 센서 객체 인식 자율주행 데이터</td></tr>
<tr><td colspan="2">데이터 포맷</td><td>이미지/포인트 클라우드 데이터: jpg/bin, 메타정보: json</td></tr>
<tr><td colspan="2">데이터 요약</td><td>• 국내 환경에 적합한 완전자율주행(Lv.4-5) 및 강건한 객체 인식 모델 개발을 위한 상호보완적인 이기종 센서 융합 데이터셋
- 라이다 데이터 기반 차량·보행자·이륜차 인식 모델 개발 활용
- 후 략 -</td></tr>
<tr><td colspan="2">데이터 출처</td><td>자체 수집 및 구축</td></tr>
<tr><td rowspan="3">데이터 이력</td><td>배포 버전</td><td>1.0</td></tr>
<tr><td>개정 이력</td><td>데이터 최초 개방</td></tr>
<tr><td>작성자/배포자</td><td>OOOO / AI Hub</td></tr>
<tr><td rowspan="2">데이터 통계</td><td>데이터 구축 규모</td><td>- 영상
<table>
<tr><th>클래스</th><th>수집(시간)</th><th>정제(시간)</th><th>이미지 수</th></tr>
<tr><td>전방 카메라</td><td rowspan="3">200</td><td rowspan="3">100</td><td>66,000</td></tr>
<tr><td>후방 카메라</td><td>66,000</td></tr>
<tr><td>…</td><td>…</td></tr>
<tr><td>합계</td><td>200</td><td>100</td><td>360,000</td></tr>
</table>
- 포인트클라우드
<table>
<tr><th>클래스</th><th>수집(시간)</th><th>정제(시간)</th><th>점군(bin)</th></tr>
<tr><td>Center (128ch)</td><td rowspan="3">200</td><td rowspan="3">100</td><td>120,000</td></tr>
<tr><td>Right (32ch)</td><td>120,000</td></tr>
<tr><td>Left (32ch)</td><td>120,000</td></tr>
<tr><td>합계</td><td>200</td><td>100</td><td>360,000</td></tr>
</table></td></tr>
<tr><td>데이터 분포</td><td>- 2D 바운딩 박스
✓ 주야간 분포
<table>
<tr><th>Category</th><th>Count</th><th>비율</th></tr>
<tr><td>Day</td><td>272,741</td><td>82.63%</td></tr>
<tr><td>Night</td><td>57,331</td><td>17.37%</td></tr>
<tr><td>합계</td><td>330,072</td><td>100%</td></tr>
</table>
- 중 략 -
- 2D 세그멘테이션
✓ 클래스별 라벨 분포
<table>
<tr><th>Label</th><th>Count</th><th>비율</th></tr>
<tr><td>car</td><td>145,455</td><td>42.84%</td></tr>
<tr><td>truck</td><td>42,326</td><td>12.47%</td></tr>
<tr><td>bus</td><td>15,318</td><td>4.51%</td></tr>
<tr><td>…</td><td>…</td><td>…</td></tr>
<tr><td>합계</td><td>339,533</td><td>100%</td></tr>
</table></td></tr>
<tr><td rowspan="2">기타 정보</td><td>대표성</td><td>장소 – 서울 경기 지역의 구도심, 신도심, 골목, 시골길, 산길, 해안도로, 경부고속도로, 영동고속도로, 올림픽대로, 강변북로, 서부간선도로 등 11개 도로
형태 – 휴게소 주차장, 터널 진출입구간, 램프 진출입 구간, 옥외·실내 주차장 등</td></tr>
<tr><td>유의사항</td><td>필요한 경우, 데이터 수집 차량의 모양과 센서 설치 위치 등을 고려해야 함</td></tr>
</table>

참고 자율주행 인지 소프트웨어 평가를 위한 객체 속성 정의 – 메타데이터 명세 내용 참고

TTA 정보통신단체표준 TTAK.KO-11.0262/R1에서는 자율주행 인지 소프트웨어 평가를 위한 각 객체의 속성을 정의하였다.

• 이동 객체 속성

✓ 보행자, 동물 속성: 유형, 상대속도, 절대 속도, 거리, 진행 각도, 검출방식, 검출좌표

✓ 차량 속성: 유형, 상대차로 위치, 상대위치, 상대속도, 절대 속도, 절대 가속도, 거리, 진행 각도, 검출방식, 검출좌표


```
    "annotations": [
        {
            "id": "ca386f256eb9f50b1df99594d42f651bdfb328a1608862ea94b6398c758f6d8c",
            "type": "bbox",
            "attributes": {
                "목적차량(특장차)": "구급차"
            },
            "points": [
                [
                    1054,
                    391
                ],
                [
                    1682,
                    391
                ],
                [
                    1682,
                    783
                ],
                [
                    1054,
                    783
                ]
            ],
            "label": "목적차량(특장차)"
        },
```


'차량 및 사람 인지 영상' 학습 데이터의 차량 속성 어노테이션 예시(출처: AI Hub)

• 고정 객체 속성(일부)

✓ 차선 속성: 유형, 상대 차선 위치, 차선색상, 차선 파라미터(차선의 오프셋, 헤딩각, 곡률 파라미터값), 검출방식, 검출좌표

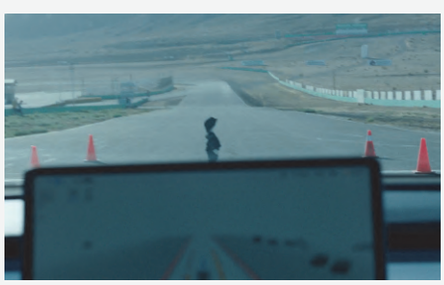
✓ 신호등 속성: 유형, 용도, 설치형태, 거리, 색상, 점멸, 검출방식, 검출좌표

04-1c 보호변수protective attribute의 선정 이유 및 반영 여부를 설명하였는가?

Yes ☐ No ☐ N/A ☐

- 보호변수는 사회적 물의를 일으킬 수 있는 민감한 특성이며, 보통 나이, 성별, 인종, 종교 등이 이에 해당한다.
- 자율주행 분야에서 활발히 연구 및 개발되고 있는 인공지능 기반 인지 알고리즘에서, 오픈 데이터셋이 기록하고 있는 민감한 특성은 거의 없는 것으로 확인되었다. 따라서, 자율주행을 위한 인지 알고리즘을 개발한다면 현재 기준으로 보호변수는 고려대상이 아닐 수 있다.
- 그러나, 추후 자율주행 알고리즘 아키텍처에서 인지·판단·제어를 위한 데이터셋에 아래와 같은 민감한 특성이 반영될 수 있다. 이때 트롤리 딜레마 상황에서 판단과 제어가 필요할 수 있으므로 보호 변수의 선정 및 반영을 고려해야 한다.
  - ✓ 인지·판단·제어 데이터셋 내 반영 가능한 민감한 특성 예시[24]: 보행자의 연령대, 성별, 인종 정보

참고 자율주행 차량에서 보행자 연령대에 따른 공정성 이슈

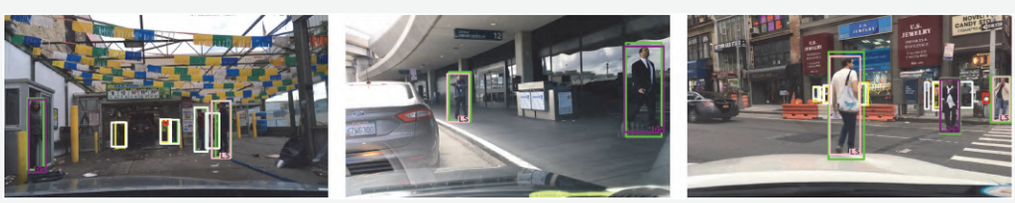


자율주행 차량의 어린이 인식 시험 장면

- Tesla 사의 완전자율주행FSD, Full Self-Driving 베타 버전(10.12.2)에서 어린이 크기의 물체(마네킹)를 제대로 탐지하지 못함[25]
  - ✓ 시험 환경: 약 110m 직선 트랙 끝에 어린이 크기의 마네킹(보행자) 준비
  - ✓ 시험 결과: 평균 시속약 40km로 주행한 차량은 실험에서 매번 마네킹과 부딪힘(FSD 모드에서 객체 인지 후 FSD 모드 해제까지 지연이 발생하였거나, FSD 모드에서 일시적으로 객체를 인식하였으나 미인식 상태로 전환)

참고 자율주행 차량에서 보행자 피부색에 따른 공정성 이슈

- 피부색이 어두운 보행자가 피부색이 밝은 보행자보다 자율주행차에 치일 가능성이 크다는 연구 결과[26,27]
  - ✓ 시험 대상 인공지능: 객체 감지object detection 모델 – Faster R-CNNRegion Based CNN, Mask R-CNN
  - ✓ 시험에 활용한 데이터셋: MS COCO, BDD100K Train
  - ✓ 피부톤 구분 기준: Fitzpatrick 척도를 사용하여 피부톤이 밝은 경우(Fitzpatrick 1~3)와 어두운 경우(Fitzpatrick 4~6)로 구분
  - ✓ 시험 결과: 이미지에서 보이는 시간대, 시야를 가리는 방해물 같은 변수를 통제하더라도 피부톤이 어두운 집단에 대해 평균 5% 덜 정확하게 인지함



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

Yes ☐ No ☐ N/A ☐

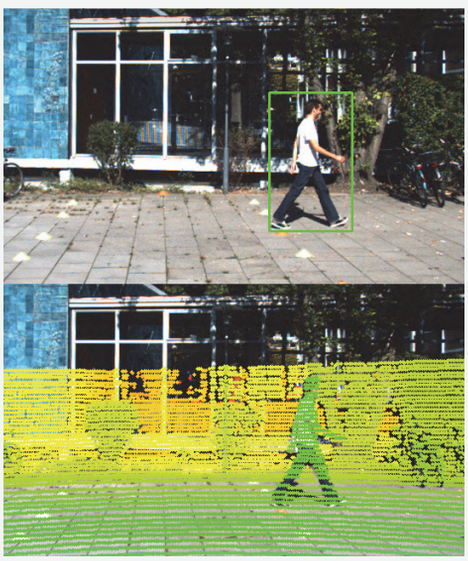
- 다중 카메라, 라이다LiDAR, Light Detection And Ranging 센서, 레이더RADAR, RAdio Detection And Ranging 센서 등 2종 이상의 장치에서 수집한 데이터를 라벨링할 때, 동일한 객체(시간, 각각의 위치 고려)를 라벨링하는 작업이 필요하므로 라벨링 작업자를 위한 교육 및 작업 가이드 문서가 필요하다.
- 2종 이상의 장치를 이용하여 수집한 데이터를 라벨링할 때는 보통 각 데이터를 정합해 동시에 재생 및 제어하며 라벨링할 수 있는 전문적인 도구가 제공되므로 도구의 사용 방법, 라벨링 시 유의사항 등을 알려주는 가이드 문서를 마련해야 한다.

참고 학습 데이터의 어노테이션/라벨링 작업 기준 예시(출처: AI Hub, 차로 위반 영상 데이터)

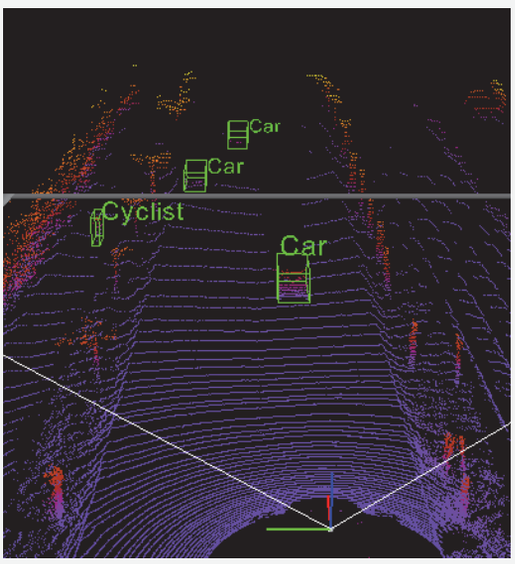
작업 기준	작업 방식
라벨링 작업 대상	• 차량: 원시 데이터 획득 차량 주행차로 및 좌우 차로 내 [표]에 제시한 차량 (4종) 폴리곤 세그멘테이션 – 단, 차량 객체의 전방, 좌/우 측방, 후방 중 최소한 한 개 이상의 단면이 가려지거나 잘림이 없는 차, 100×50 픽셀 이상인 차량이 대상임 • 차선: 원시 데이터 획득 차량 주행차선의 좌/우 차선이 작업대상으로 폴리곤 세그멘테이션 – 단, 2줄 점선, 복합선, 교차로의 갓길, 횡단보도, 정지선은 작업에서 제외
라벨링 작업 기준	• 오태깅 – (차량) 가로/세로 100×50 픽셀 미만은 태깅 대상 아님(단, 이륜은 세로 100 픽셀 미만), ✓ 차량객체의 전방, 좌/우 측방, 후방 중 단 하나의 단면도 정확하게 표시되지 않는 (단면이 가려지거나 잘림이 있는) 차량은 태깅 대상 아님 – (차선) 라벨링 대상/범위가 아닌 차선은 태깅 대상 아님 • 과태깅 – (차량) 승용자동차, 승합자동차, 화물(특수)자동차, 이륜자동차가 아닌 대상은 태깅 대상 아님 – (차선) 정의되지 않은 클래스 차선은 태깅 대상 아님 • 미태깅 – (차량) 라벨링 대상/범위에 준하는 차량이 태깅되지 않은 경우 – (차선) 라벨링 대상/범위에 준하는 차선이 태깅되지 않은 경우
라벨링 범위	• 차로 위반 차량 – 위반 차량이 있는 경우, 근접 위반 차량 1대 – 위반 차량이 없는 경우, 해당 차로의 근접 차량 최대 3대 (측면) 바로 옆 차선에 위치한 차량 / (정면) 주행차로와 좌우 차선에 위치한 차량 • 차선 – 위반 차량이 있는 경우, 위반 차선 1개 – 위반 차량이 없는 경우, (측면) 바로 옆 차선 1개 / (정면) 주행차로의 좌우 차선

참고 LiDAR 센서 및 다중 카메라 센서 정합 라벨링 도구

- 3D Sensor Fusion 라벨링 오픈소스 도구: kitti object vis



라이다 및 카메라 데이터의 2D 바운딩 박스



라이다 데이터의 3D 바운딩 박스

04-2 데이터의 출처는 기록 및 관리되고 있는가?

Yes No N/A
☐ ☐ ☐

해당여부 판단

자율주행 인공지능 모델 개발을 위해 데이터셋을 직접 구축하거나, 원시 데이터 구매 후 자체 정제 또는 오픈소스 데이터셋을 활용하는 등의 여부에 따라 본 항목을 고려하여 만족 여부를 판단하십시오.

- 학습 데이터의 품질은 인공지능 모델 성능에 큰 영향을 미치는 중요한 요인 중 하나이므로 데이터를 수집하거나 생성하는 과정에서 품질을 확보하도록 노력해야 하며, 경우에 따라 오픈소스 데이터셋을 활용할 수도 있다.
- 오픈소스 데이터셋을 활용할 때 다수의 사용자가 데이터 활용 과정에서 발견한 오류가 추후 발견될 수 있으며, 이로 인한 데이터셋 수정, 재구축으로 데이터 버전 변경될 수 있다. 만약 데이터 버전이 변경되면 인공지능 모델의 동작에도 영향을 줄 수 있으므로, 이러한 문제에 대응하기 위해 학습에 활용한 오픈소스 데이터셋의 명확한 출처, 구축 시점, 오픈소스 데이터셋 버전 등의 정보를 기록하고 관리해야 한다.

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

Yes No N/A
☐ ☐ ☐

- 오픈소스 데이터셋은 객체 인식 인공지능 모델을 개발할 때도 종종 활용되므로, 해당 데이터셋이 신뢰할 만한 수준의 품질인지 고려해야 한다.
- TTA 정보통신단체표준 TTAK.KO-10.1339에서는 지도학습 계열의 인공지능 기술에 활용되는 데이터를 획득할 때 출처의 신뢰성 확보 측면에서 고려해야 할 내용을 정리하였다.
- 다음과 같이 자율주행 분야 오픈소스 데이터셋의 출처 신뢰성 확보를 위한 체크리스트를 활용하고, 각 항목을 비교 분석하여 최종 활용한 오픈소스 데이터셋과 사용 근거를 기록하도록 한다.

참고 TTAK.KO-10.1339 기반 자율주행 데이터 출처 신뢰성 확보 여부 확인 예시

1) 제공하는 데이터셋의 규모가 충분히 커서 데이터 사용자가 원하는 학습용 데이터를 제공하는 데 문제가 없는지 확인

	KITTI	Cityscapes	ApolloScape	Mapillary	BDD100K
# Sequences	22	~50	4	N/A	**100,000**
# Images	14,999	5,000 (+2,000)	143,906	25,000	**120,000,000**
Multiple Cities	N	Y	N	Y	**Y**
Multiple Weathers	N	N	N	Y	**Y**
Multiple Times of Day	N	N	N	Y	**Y**
Multiple Scene Types	Y	N	N	Y	**Y**
데이터셋 비교: Kirti Bakshi, 2018.9.24					

2) 해당 데이터가 지속적으로 업데이트되고 및 추가 데이터 제공 등이 이루어지고 있는지 확인

	KITTI	Cityscapes	ApolloScape	Mapillary	BDD100K
수집시작 연월	**2012.03**	2015.06	2018.04	**2020.01**	2018.05
최근 업데이트 연월	**2021.02**	2020.10	2020.09	**2022.06**	2020.04
출처: 각 데이터셋 사이트					

3) 데이터와 함께 설계서의 내용이 명확히 제공되는지 확인

	KITTI	Cityscapes	ApolloScape	Mapillary	BDD100K
설계서 내용 제공	**Y**	**Y**	**Y**	**Y**	**Y**

4) 해당 데이터의 활용 건수 및 인용 건수가 많아 범용성이 높은지 확인

	KITTI	Cityscapes	ApolloScape	Mapillary	BDD100K
최근 3년간 인용 및 활용 건수 (평균 증가율, %)	1,349 (23.24)	**1,487 (26.41)**	41 (44.21)	47 (7.62)	68 (63.5)
벤치마크	**115**	37	5	3	12
출처: Papers with Codes 데이터셋 검색 결과					

04-2b 오픈소스 데이터셋을 활용하는 경우, 출처를 명시하였는가?

Yes ☐ No ☐ N/A ☐

- 자율주행 알고리즘 중 인지를 위한 인공지능 알고리즘이 중점적으로 개발 및 적용되는데, 이때 오픈소스 데이터셋을 상당 부분 활용하는 것으로 확인되었다.
- 각 오픈소스 데이터셋을 활용할 때, 데이터셋에 포함되지 않은 객체(누락), 객체별 수량의 큰 차이(클래스 밸런스)로 인해 모델의 오류를 유발할 수 있다.
- 따라서, 오픈소스 데이터셋을 활용하여 인공지능 모델을 구축할 때는 과거·현재·미래 시점에 발생하는 데이터 편향의 원인을 파악할 수 있도록 확보된 데이터의 명확한 출처와 관련 정보를 명시하여 관리해야 한다.
- 아래와 같은 정보를 활용하여 오픈소스 데이터셋의 출처를 명시할 수 있다.

참고 TTAK.KO-10.1339 기반 자율주행 오픈소스 데이터셋의 출처 명시 요소(예)

- 저자(필수): 오픈소스 데이터셋의 작성자
- 출판일(필수): 데이터셋을 사용할 수 있게 된 날짜 또는 모든 품질 보증 절차가 완료된 날짜
- 제목(필수): 오픈소스 데이터셋의 이름
- 판edition: 데이터셋이 원시 또는 구체화된 방식을 나타내는 데이터 처리의 수준 또는 단계
- 버전: 데이터 요소를 더 추가하거나 파생 프로세스를 다시 실행한 결과로 데이터가 변경될 때의 숫자
- 특징명 및 URIUniform Resource Identifier: ISO 19101:2002의 이름 'feature' (예: GridSeries, ProfileSeries) 및 표준 정의를 식별하는 URI
- 데이터셋 유형: 데이터베이스 또는 데이터셋
- 게시자(필수): 데이터셋을 호스팅하는 조직 또는 품질 보증 수행 조직
- 고유 숫자 지문UNF, Universal Numerical Fingerprint: 인용 이후 변경 사항이 발생하지 않았는지 확인하는 데 사용되는 데이터셋의 암호화 해시값
- 식별자: 영구 체계에 따른 데이터의 식별자
- 위치(필수): 데이터셋을 사용할 수 있는 영구 URLUniform Resource Locator

안전성

요구사항 05 데이터 강건성 확보를 위한 이상abnormal 데이터 점검

대표 행위자 | 데이터 과학자 협력 대상 | 데이터 공급자 인공지능 모델 개발자 보안 분석가

- 인지 시스템에 사용되는 데이터 특성 분포를 시각화하여 라벨링 작업 오류를 확인하고, 메타데이터의 스키마 통계 분석 기법을 이용하여 데이터의 이상값을 식별·처리한다. 자율주행 알고리즘 중 환경 인지를 위한 카메라나 라이다 센서 기반 시각적 데이터는 개별 또는 정합 방식으로 라벨링 관련 오륫값을 식별·처리한다. 또한, 인공지능 자율주행 알고리즘 모델이 고의적이거나 의도치 않은 현장의 공격 상황을 이해하고, 모델 학습 전 데이터 준비·관리 단계에서 대응 방안을 마련하는 것이 필요하다.

05-1 이상 데이터의 식별 및 정상 여부를 점검하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

자율주행 인공지능 학습용 데이터를 직접 구축하거나, 학습 데이터에 대한 정상·오류 여부가 명확하게 확인되지 않았다면 본 항목을 고려하여 만족 여부를 판단하십시오.

- 이상 데이터란 학습용 데이터를 구성하는 데이터셋의 수집·가공 과정에서 발생할 수 있는 다양한 오류error와 일반적인 데이터의 범위에서 크게 벗어난 데이터 이상값outlier을 포괄한다. 학습 데이터의 수집·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 강건성 확보가 어렵다.
- 자율주행 분야와 같이 비정형 데이터unstructured data를 학습에 활용하는 경우, 데이터 전처리 과정에서 이상 데이터의 식별을 위한 별도의 기법을 마련하여야 한다.
- 인지 시스템을 위한 인공지능 모델의 학습 데이터 내 여러 센서 데이터를 정합 시각화하여 라벨링 작업 결과에 오류가 없는지 확인한다. 또한, 메타데이터의 스키마schema를 분석하여 데이터의 이상값을 식별하고 이상값 유무를 확인해야 한다.

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

Yes ☐ No ☐ N/A ☐

- 데이터 전처리 과정 중 하나인 데이터 정제 단계 이후, 데이터 전체 분포를 시각화하여 추가적인 입력 오류를 확인할 수 있다. 특히, 이러한 데이터 분포 시각화는 인공지능 모델 학습 시 활용하는 데이터를 이해하는 데 많은 도움을 준다.
- 자율주행 인지 알고리즘을 위한 데이터셋을 활용할 때, 아래와 같은 항목을 단독 또는 복수 조합하고 각 정보를 시각화하여 분포를 확인한다.
  - ✓ 날씨 조건 예시: 맑음, 흐림, 비, 눈, 안개
  - ✓ 시간 조건 예시: 주간, 야간, 해질녘, 새벽
  - ✓ 도로 특성 예시: 시·군도 일반도로, 고속도로, 자동차 전용도로, 국도, 지방도, 기타
  - ✓ 교차로 여부 예시: 신호 교차로, 비신호 교차로, 회전 교차로, 3지·4지·5지·6지 교차로, 교차로 아님
  - ✓ 도로 폭 예시: 왕복 2·4·6·8차로, 이면도로
  - ✓ 곡선도로 유무 예시: 직선도로, 곡선도로
- 데이터 분포의 시각화 방법은 데이터 특성에 따라 다양한 기법이 존재한다. 전체 데이터의 평균, 분산, 편차 등을 활용하여 데이터 분포를 시각화하는 분포 도표, 범주형 데이터를 시각화하는 범주형 도표, 2차원 행렬 데이터를 시각화하는 행렬 도표 등이 있다.

데이터 분포 시각화 기법 예시

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

05-1b 학습 데이터 이상값 식별 기법을 적용하였는가?

Yes ☐ No ☐ N/A ☐

- 자율주행 인지 알고리즘을 위한 데이터셋은 주로 멀티 카메라, 라이다, 레이더 센서의 데이터로 라벨링 작업이 되어 있다.
- 각 센서의 데이터는 다양한 형식과 속도로 출력을 제공하고 있어 데이터를 이해하고 라벨링 작업 결과를 확인하는 데 어려움이 있다[28]. 이 경우에는 각 데이터를 정합하여 라벨링 오류를 확인한다.
  - ✓ 카메라 데이터 예시: 30 FPS[Frame Per Second] 3D 행렬 이미지, 객체 2D 바운딩 박스, 보행자 2D 키포인트, 라인 등
  - ✓ 라이다 센서 데이터 예시: 20Hz 포인트 클라우드, 객체 3D Cuboid, 보행자 3D 키포인트 등
- 데이터 이상값을 식별할 때는 데이터 전체에 통계적 기법을 적용하여 전체 데이터셋을 고려하였을 때 차별화되는 데이터 포인트를 찾아내는 방법이 주로 활용된다. 이와 관련된 대표적인 기법은 Z-점수, 사분위수 범위, DBSCAN[Density-Based Spatial Clustering of Applications with Noise] 등이 있다. 이러한 기법을 인지를 위한 데이터셋에 적용하려면 각 데이터의 다음과 같은 특징을 따로 정의하여 적용할 수 있다.
  - ✓ 객체 특징 예시: 가로, 세로, 높이 크기, 넓이(또는 부피), 객체 라벨 등
  - ✓ 이미지 특징 예시: 시간대, 날씨, 도로 종류 등
- 또한, 데이터셋 중 메타데이터를 대상으로 이상값 식별 기법을 적용할 수도 있다. 전체 메타데이터를 분석하여 특성 데이터에 대한 스키마를 추론하고 데이터셋의 통계와 비교 분석하여 이상값을 확인한다[29].

데이터 이상값 식별 기법 예시

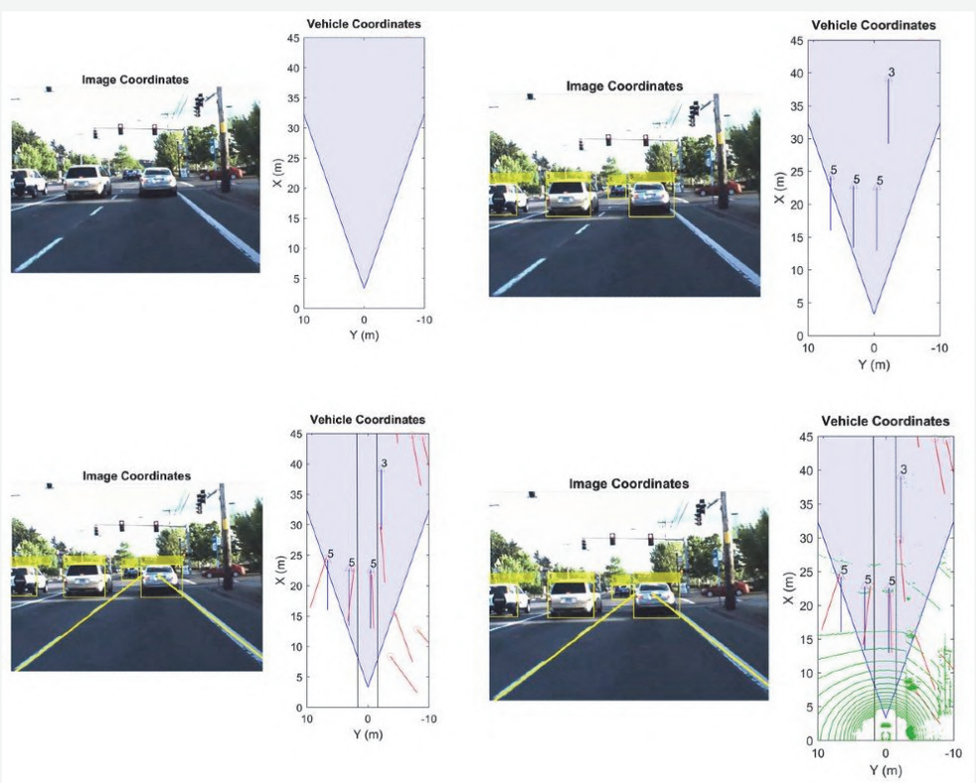
이상값 식별 기법 분류	설명
Z-점수	가장 간단한 통계적 측정 방법으로, Z-점수는 주어진 데이터셋의 분포 평균과 표준편차를 이용하여 관찰된 데이터 포인트가 전체 데이터로부터 얼마나 멀리 떨어져 있는지를 수치화한다.
사분위수	중앙값(Q2)으로 데이터를 두 부분으로 나누고, 다시 왼쪽 중앙값(Q1)과 오른쪽 중앙값(Q3)으로 나누어 총 4개의 범위를 정하며 사분위수 범위(Q3-Q1)를 구해 해당 범위를 벗어나면 이상값으로 판별한다.
DBSCAN	노이즈가 있는 밀도 기반 공간 클러스터링을 대표하는 알고리즘이다. 임의 모양의 클러스터나 이상값이 있는 클러스터를 찾을 수 있다. 특정 포인트가 클러스터의 많은 포인트 가까이에 위치하면 클러스터에 속한다고 판단하고, 아니라면 이상값으로 판단한다.

참고 데이터 이상값 식별 방안 – 라이다, 카메라 데이터 및 라벨링 데이터의 정합 결과 시각화

• 차량 센서 데이터 예시(출처: MathWorks)

```
Camera  (640 x 480 x 3)

Vision Detector
  SensorID       = 1;
  Lane Detector
    Left
      IsValid:        1
      Confidence:     3
      BoundaryType:   3
      Offset:         1.68
      HeadingAngle:   0.002
      Curvature:      0.0000228
    Right
      IsValid:        1
      Confidence:     3
      BoundaryType:   2
      Offset:         -1.58
      HeadingAngle:   0.002
      Curvature:      0.0000228

Radar Detector
  SensorID       = 2;
  Timestamp      = 1461634696407521;

  Lidar  (47197 x 3)
  -12.2911    1.4790   -0.5900
  -14.8852    1.7755   -0.6475
  -18.8020    2.2231   -0.7396
  -25.7033    3.0119   -0.9246
   -0.0632    0.0815    1.2501
   -0.0978    0.0855    1.2561

  Inertial Measurement Unit
    Timestamp:   1461634696379742
    Velocity:    9.2795
    YawRate:     0.0040

Detections(3)
  TrackID:        12
  TrackStatus
```

• 라이다 데이터, 카메라 데이터, 라벨링 데이터 정합 결과 시각화 예시(출처: MathWorks)



(왼쪽 위) 센서 커버 영역 시각화, (오른쪽 위) 차량 좌표를 이미지 좌표로 변환,
(왼쪽 아래) 차선 및 레이더 탐지 결과 시각화, (오른쪽 아래) 라이다 포인트 클라우드 시각화

참고 메타데이터 기반 데이터 이상값 식별 방안 – TFDV[TensorFlow Data Validation]

- 데이터 스키마 추론 예시

```
In [21]: tfdv.display_schema(schema)
```

Feature name	Type	Presence	Valency	Domain
fare	Float	required	single	
trip_start_hour	Int	required	single	
dropoff_census_tract	Float	optional	single	
company	String	optional	single	company
trip_start_timestamp	Int	required	single	
pickup_longitude	Float	required	single	
trip_start_month	Int	required	single	
trip_miles	Float	required	single	
dropoff_longitude	Float	optional	single	
dropoff_community_area	Float	optional	single	
pickup_community_area	Int	required	single	
payment_type	String	required	single	payment_type
trip_seconds	Float	optional	single	
trip_start_day	Int	required	single	
tips	Float	required	single	
pickup_latitude	Float	required	single	
dropoff_latitude	Float	optional	single	

Domain	Values
company	"0118 - 42111 Godfrey S.Awir", "0694 - 59280 Chinesco Trans Inc", "1085 - 72312 N and W Cab Co", "2733 - 74600 Benny Jona", "2809 - 95474 C & D Cab Co Inc.", "3011 - 66308 JBL Cab Inc.", "3152 - 97284 Crystal Abernathy", "3201 - C&D Cab Co Inc", "3201 - CID Cab Co Inc", "3253 - 91138 Gaither Cab Co.", "3385 - 23210 Eman Cab", "3623 - 72222 Arrington Enterprises", "3897 - Ilie Malec", "4053 - Adwar H. Nikola", "4197 - 41842 Royal Star", "4615 - 83503 Tyrone Henderson", "4615 - Tyrone Henderson", "4623 - Jay Kim", "5006 - 39261 Salifu Bawa", "5006 - Salifu Bawa", "5074 - 54002 Ahzmi Inc", "5074 - Ahzmi Inc", "5129 - 87128", "5129 - 98755 Mengisti Taxi", "5129 - Mengisti Taxi", "5724 - KYVI Cab Inc", "585 - Valley Cab Co", "5864 - 73614 Thomas Owusu", "5864 - Thomas Owusu", "5874 - 73628 Sergey Cab Corp.", "5997 - 65283 AW Services Inc.", "5997 - AW Services Inc.", "6488 - 83287 Zuha Taxi", "6743 - Luhak Corp", "Blue Ribbon Taxi Association Inc.", "C & D Cab Co Inc", "Chicago Elite Cab Corp.", "Chicago Elite Cab Corp. (Chicago Carriag", "Chicago Medallion Leasing INC", "Chicago Medallion Management", "Choice Taxi Association", "Dispatch Taxi Affiliation", "KOAM Taxi Association", "Northwest Management LLC", "Taxi Affiliation Services", "Top Cab Affiliation"
payment_type	"Cash", "Credit Card", "Dispute", "No Charge", "Pcard", "Unknown"

- 데이터 이상 감지 예시: payment_type 특성에 예상하지 못한 문자열 값이 있음(1% 미만으로, "Prcard" 문자열이 포함되어 있음)

```
In [24]: tfdv.display_anomalies(anomalies)
```

Feature name	Anomaly short description	Anomaly long description
payment_type	Unexpected string values	Examples contain values missing from the schema: Prcard (<1%).

05-2 데이터 공격에 대한 방어 수단을 강구하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

자율주행 인공지능 모델의 개발 및 운영과정에서 데이터 공격에 대한 방어 수단이 필요한 경우, 본 항목을 고려하여 만족 여부를 판단하십시오.

- 자율주행을 위한 알고리즘 중 인지 모델은 고의적이거나 의도하지 않은 물체, 표지판의 훼손 등 다양한 상황에서 공격받을 수 있으므로 이에 대한 방어 수단을 마련해야 한다.
- 인지 모델은 단일 데이터 타입(예: 이미지 또는 포인트 클라우드)을 이용할 경우 공격에 취약하므로 대처 방안을 검토해 적용해야 한다.

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

Yes ☐ No ☐ N/A ☐

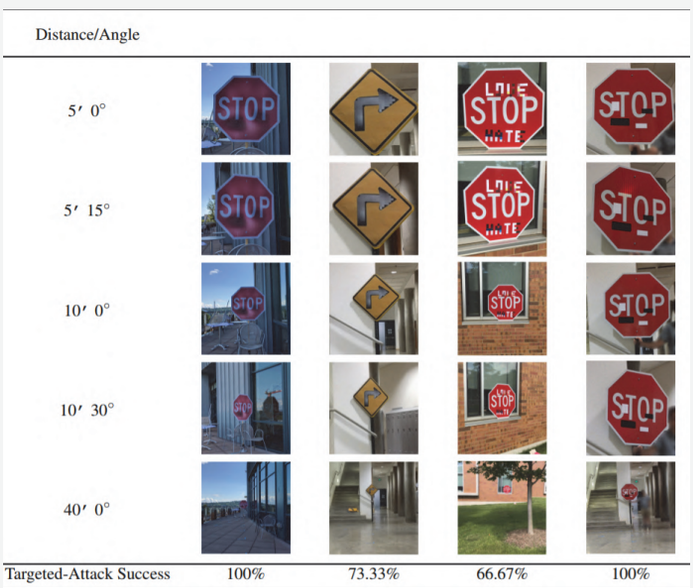
- 자율주행 알고리즘 중 인지를 위한 인공지능 모델은 4가지 적대적 공격[30,31] 중 모델 회피 공격에 취약하다.
- 적대적 공격은 카메라뿐만 아니라, 3D 객체를 프린팅하여 적대적 객체를 만들어 라이다 센서를 통해 입력되는 데이터를 기만하고 공격하는 것이 가능하다[32].
- 아직 완전한 방어 기법이 조사되진 않았지만, 인공지능 애플리케이션 개발자와 보안 분석가가 해야 할 일을 정리하여 제공한다[31]. 공격은 단일 센서로 입력되는 단일 인공지능 모델을 대상으로 이루어진다는 점을 이용하여 다양한 센서(예: 레이더, 라이다)의 인지 결과를 병합하여 공격을 방어할 수 있다.
- 또한, 운행 가능 영역ODD, Operational Design Domain의 정보를 확인할 수 있는 HD 지도 정보를 추가로 활용하여 교통표지판을 대상으로 자행되는 공격을 방어할 수 있다.

참고 향후 인공지능 애플리케이션 개발자 및 보안 분석가가 해야 할 일[31]

- 모든 프로덕션 기계학습 애플리케이션에서 적대적 공격 가능성을 전제
- 취약한 코드를 쓰거나 배포하기 전에 적대적 위협 평가를 실시
- 인공지능 학습 파이프라인의 표준 위험 완화 활동으로 적대적 위협 예시를 생성
- 폭넓은 적대적 입력에 대해 인공지능 애플리케이션을 테스트해서 추론의 견고함을 확인
- 새로운 적대적 기계학습 위협 매트릭스가 제공하는 것과 같은 적대적 방어 지식을 재사용해 위조 입력 예시에 대한 인공지능의 탄력성을 개선
- 배포된 인공지능 모델의 라이프사이클 전반에서 적대적 공격 방어를 지속적으로 업데이트
- 데이터 과학자에게 세밀한 적대적 공격 대응 방법론을 제공해 인공지능 개발 및 운영화 라이프사이클 전반에서 이런 방법을 적용하도록 유도

참고 자율주행 데이터 회피 공격 사례

- 도로 표지판을 이용한 공격 사례
  - ✓ 워싱턴대학, 미시간대학, 스토니브룩대학, 캘리포니아 버클리대학의 연구원 그룹은 도로 표지판에 대한 약간의 수정을 보여주는 논문을 발표
  - ✓ 모델은 STOP 표지판을 45마일/시간 속도제한 표지판으로 해석



서로 다른 설정에서의 적대적 공격 성공률[33]

참고 자율주행 데이터 중독 공격 사례

- 데이터 중독 공격 연구 사례
  - ✓ 자율주행 차량에서 적용할 수 있는 심층 생성 모델DGM, Deep Generative Model에 대한 데이터 중독 공격[34]

중독된 DGM에 입력되는 이미지 – 물방울 및 적색 신호등

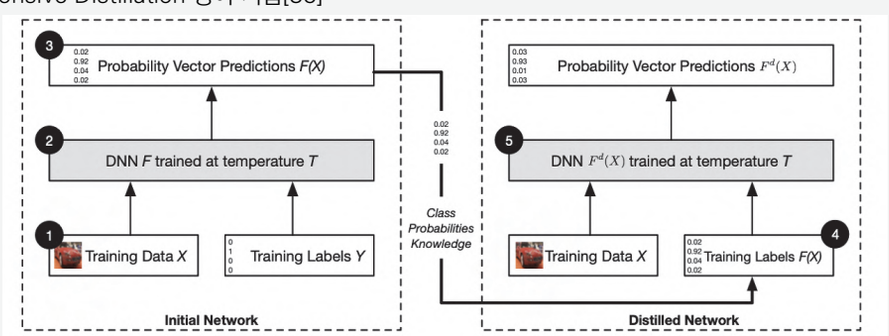
중독된 DGM의 출력 이미지 – 녹색 신호등으로 교묘히 변경

  - ✓ 심층 생성 모델은, 위 그림과 같이 이후 인지를 위한 모델로의 입력을 위한 전처리 기술로 활용할 수 있음
  - ✓ 이때, 각 노이즈(예: 이미지 내 빗방울, 눈)를 제거하며 신호등, 표지판 등을 함께 변경하는 데이터 중독 공격에 노출될 수 있음
  - ✓ 공격 시나리오
    - 악의적인 내부자 또는 경쟁 회사의 스파이가 학습용 데이터셋의 일부 데이터를 주입하거나 교체

- 외부 공격자가 APT[Advanced Persistent Threat]와 같은 접근 방식을 이용해 자율주행 시스템 개발 회사의 학습 시스템이나 임대 클라우드에 은밀히 침입한 후 데이터를 조작
- 일부 자율주행 스타트업은 비용에 민감하거나 기술 지원이 부족하기 때문에 최신 공개 모델을 직접 가져와 자체 용도에 맞게 미세 조정하는 것을 선호. 이때 공격자는 중독된 데이터로 일부 심층 생성 모델을 학습시키고 해당 회사를 피싱하기 위해 공개할 수 있음

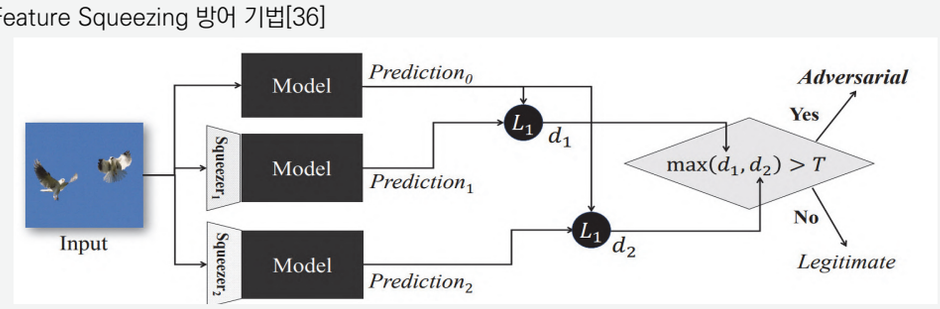
참고 데이터 공격에 대한 방어 기법 예시

• Defensive Distillation 방어 기법[35]



- ✓ 현재 C&W 공격, PGD 등 여러 공격 방법이 등장하여 이 기법이 깨지며 무뎌지긴 했으나 처음 나왔을 때 많은 사람에게 관심을 받으며 차세대 방어 기법으로 주목받았음
- ✓ Knowledge Distillation 방법에 기반하였으며, Teacher 네트워크 T와 Student 네트워크 S 등 2개 모델이 존재하고, 먼저 학습된 T의 knowledge를 Student를 학습시킬 때 사용하여 추가적인 지도를 하는 방식임
- ✓ 논문이 나왔을 당시 대부분의 공격은 그래디언트를 이용한 방식들이었음. 모델의 그래디언트가 높은 값을 가지게 되면 그만큼 작은 노이즈[perturbation]만으로 네트워크 출력에 큰 변화를 줄 수 있기 때문에 적대적 샘플을 만들기 쉬워짐
- ✓ 따라서, 이 방법은 모델이 작은 그래디언트를 갖게 만들어 공격이 동작하기 어렵도록 함

• Feature Squeezing 방어 기법[36]



- ✓ 적대적 공격에 교란당하지 않도록 모델을 강화하기 위해, 데이터를 학습할 때 표현의 복잡성을 줄여 민감도[sensitivity]를 낮춰 모델을 견고하게 만드는 방법
- ✓ 화소의 색상 값을 더 작게 인코딩하여 색상 깊이를 줄이는 방법, 영상에 평활화[smoothing] 필터를 적용하는 방법이 있음

다양성 존중 | 책임성 | 투명성

요구사항 06 수집 및 가공된 학습 데이터의 편향 제거

대표 행위자 | 데이터 공급자 협력 대상 | 데이터 과학자 | 인공지능 모델 개발자

- 자율주행 알고리즘 중 인지 모델을 위한 데이터를 수집·가공할 때는 사람의 운행 습관이나 지식수준 등의 차이로 인한 편향을 인식·제거하는 방안을 적용한다. 데이터 수집 시 편향 발생 가능성을 고려해 목표별 수집 시나리오를 사전에 설계하고, 다양한 센서 하드웨어 사양 조합을 고려해야 한다. 또한, 데이터셋을 정제·라벨링할 때는 특성이 과도하게 제거됐는지 확인해야 한다. 데이터셋에 민감한 특성 정보가 포함됐거나, 자율주행 판단·제어를 위한 인공지능 개발 시 보호 변수를 선정·분석한다.

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

Yes ☐ No ☐ N/A ☐

해당여부 판단

자율주행 인공지능 학습용 데이터를 직접 수집하는 경우 본 항목을 고려하여 만족 여부를 판단하십시오.

- 데이터 편향은 데이터셋을 직접 수집할 때 인적·물리적 요인으로 인해 다양한 데이터를 수집하지 못하여 발생할 수 있다. 이에 따라 자율주행 인공지능 알고리즘의 동작 성능에 문제를 일으키고, 자율주행 시스템의 오동작으로 이어질 수 있으므로, 편향을 완화하려는 노력이 필요하다.
- 차량에 센서를 부착하여 데이터셋을 직접 수집하는 경우, 데이터 수집 작업자가 익숙한 패턴으로 운행에 나서면 편향이 발생할 수 있다. 따라서 사전에 다양한 시나리오를 설계하여 작업자별로 데이터 특성이 편향되지 않도록 한다.
- 자율주행 알고리즘 중 이미지를 이용해 인지 알고리즘을 개발할 때, 데이터를 직접 수집하는 경우면 수집 환경 및 제약 조건으로 인해 다양한 데이터를 확보하기 어려울 수 있으므로, 카메라, 웹 크롤링 등 이기종 장치를 이용하여 다양성을 확보한다.
- 차량에 센서를 부착한 후에 데이터 수집할 때, 비용 및 시간적 제약으로 인해 단일 하드웨어를 사용하면, 제한된 상황에서만 데이터가 수집되는 데이터 편향이 발생할 수 있다. 따라서 다양한 센서 하드웨어의 도입을 고려하여 데이터 다양성을 확보해야 한다.

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

Yes ☐ No ☐ N/A ☐

- 자율주행 분야 인지 알고리즘 개발을 위한 데이터 수집 시, 데이터 수집을 위한 차량을 준비하고 센서를 부착해 실제 도로를 주행하며 데이터를 수집한다. 이때, 데이터를 수집하는 작업자가 편안하게 운행할 수 있는 익숙한 도로, 지역, 시간대를 중점적으로 운행하며 데이터를 수집할 수 있다. 그 결과, 붐비는 도로, 다수의 보행자가 이용하는 횡단보도 또는 교차로, 다양한 차선 종류, 다양한 교통표지판 등 자율주행 알고리즘 개발에 필요한 다양한 객체 또는 시나리오를 충분히 수집할 수 없어 편향이 발생할 수 있다.
- 인지 알고리즘 중 단일 객체(예: 보행자, 차량)의 개발 및 고도화를 위해 데이터 수집 작업을 별도로 진행하는 경우, 작업자에 따라 편향이 발생할 수 있다. 따라서, 데이터 수집 작업의 가이드라인을 마련하고, 다양한 작업자를 모집하여 특정 배경과 성향을 배제하고, 수집 결과를 검토하는 검수자를 충분히 확보해 수집 작업자들의 개인별 편차를 줄여야 한다.

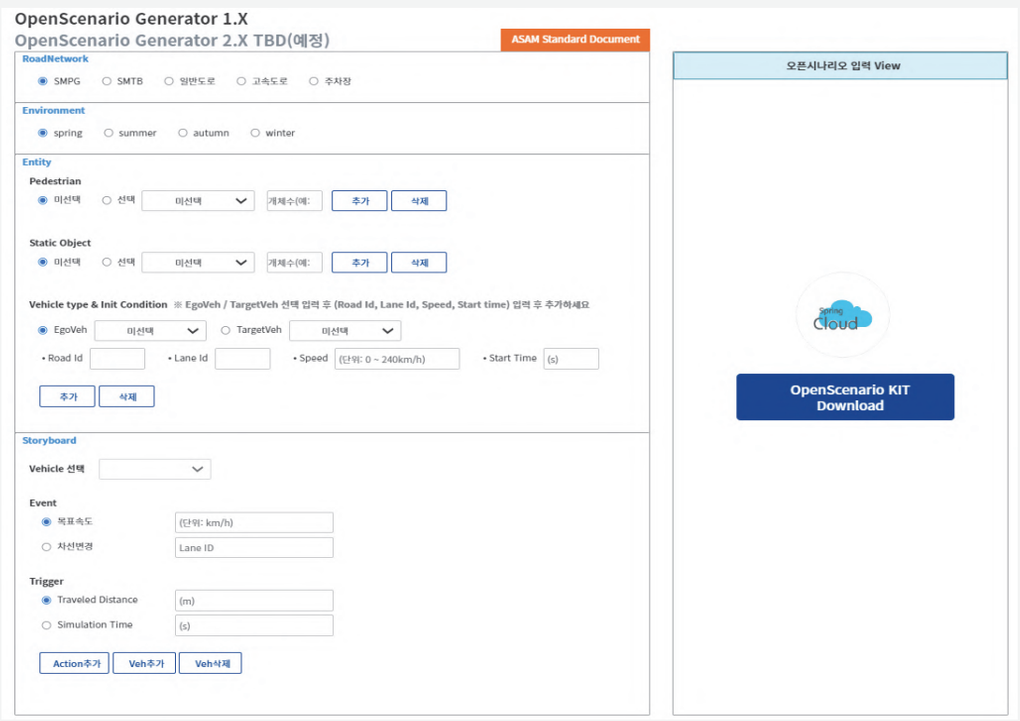
참고 자율주행 학습 데이터의 수집 시나리오 설계 예시 (출처: AI Hub, 차량 및 사람 인지 영상 데이터)

- 차량 주행 중 일어날 수 있는 돌발 상황과 수신호, 보행자 행동 등 특수 상황을 정의, 조합하여 시나리오 설계
  - ✓ 생성 가능 유형 총 368,400가지 조합 중 현실적으로 데이터 취득이 가능한 시나리오 100가지 이상 정의

<table>
<tr><th colspan="3">사고 발생 현황 조건</th><th colspan="5">교통 환경 조건</th><th colspan="4">사고 상황/행태 관점</th><th>사고 심각도</th></tr>
<tr><th>날씨 조건</th><th>주/야간 조건</th><th>도로 특성</th><th>교차로 여부</th><th>도로 폭</th><th>곡선 도로 유무</th><th>충돌 유형</th><th>중앙분리대 유무</th><th>사고 유형</th><th colspan="2">이동 속도</th><th>사고 발생 거리(m)</th><th>사고발생</th></tr>
<tr><td rowspan="2">맑음</td><td rowspan="2">주간</td><td rowspan="2">시,군도 (일반도로)</td><td rowspan="2">신호 교차로</td><td rowspan="2">왕복 2차로</td><td rowspan="2">직선도로</td><td rowspan="2">직각충돌</td><td rowspan="2">가드레일 (차량이탈) 분리대</td><td rowspan="10">보행자 사고</td><td rowspan="10">자동차</td><td>30kph 이하</td><td rowspan="2">2~16m (2m 단위)</td><td rowspan="10">단독 사고</td></tr>
<tr><td>40kph</td></tr>
<tr><td rowspan="8">흐림</td><td rowspan="8">야간</td><td rowspan="8">고속도로</td><td rowspan="8">비신호 교차로</td><td rowspan="8">왕복 4차로</td><td rowspan="8">곡선도로</td><td rowspan="8">정면 충돌</td><td rowspan="8">무단횡단 방지분리대</td><td>50kph</td></tr>
<tr><td>60kph</td><td rowspan="7">20~30m (5m 단위)</td></tr>
<tr><td>70kph</td></tr>
<tr><td>80kph</td></tr>
<tr><td>90kph</td></tr>
<tr><td>100kph 이상</td></tr>
<tr><td rowspan="2">비</td><td rowspan="2">해질녘</td><td rowspan="2">자동차 전용도로</td><td rowspan="2">회전 교차로</td><td rowspan="2">왕복 6차로</td><td rowspan="2"></td><td rowspan="2">측면 충돌</td><td rowspan="2">콘크리트 방지분리대</td><td rowspan="4">자동차 사고</td><td rowspan="4">보행자</td><td>걷기 4kph</td><td>40~70m (10m 단위)</td><td rowspan="4">2대 사고</td></tr>
<tr><td rowspan="3">달리기 10kph</td><td rowspan="3"></td></tr>
<tr><td>눈</td><td>새벽</td><td>국도</td><td>3차 교차로</td><td>왕복 8차로 이상</td><td></td><td>후면 충돌</td><td>분리대 없음</td></tr>
<tr><td>안개</td><td></td><td>지방도</td><td>4차 교차로</td><td>이면도로</td><td></td><td>측면 충돌</td><td></td><td rowspan="2">이륜차 사고</td><td rowspan="2">자전거</td><td>15kph</td><td></td><td rowspan="2">3대 사고</td></tr>
<tr><td></td><td></td><td>기타</td><td>5차 교차로</td><td></td><td></td><td></td><td></td><td>20kph</td><td></td></tr>
<tr><td></td><td></td><td></td><td>6차 교차로 이상</td><td></td><td></td><td></td><td></td><td rowspan="2">기타사고 (PM, 건설장비, 농기계 등)</td><td>PM, 농기계</td><td>25kph 이하</td><td></td><td rowspan="2">4대 이상 다중 사고</td></tr>
<tr><td></td><td></td><td></td><td>교차로 아님</td><td></td><td></td><td></td><td></td><td>건설기계</td><td>30kph 이상</td><td></td></tr>
</table>

Use Case 국내 S사의 데이터 수집 시나리오 생성기 활용 사례(일부 발췌)

- S사에서는 자율주행 인공지능 학습용 데이터의 수집 시나리오 설계 시 인적 편향을 최소화하기 위해, 기술적 수단으로 오픈 시나리오 생성기를 활용하고 있음
  - ✓ 오픈 시나리오 생성기: ASAM[Association for Standardization of Automation and Measuring systems] 표준 문서에 따라 개발한 자체 시나리오 생성 도구로, 데이터 내 포함되어야 할 도로 종류, 환경(예: 계절), 객체(예: 보행자, 정적 객체, 동적 객체), 시나리오 상황(예: 차량 주행 속도)과 같은 요인을 선택할 수 있음



국내 S사에서 활용 중인 오픈 시나리오 생성기 화면 구성

06-1b 데이터의 다양성 확보를 위해 수집 시 여러 차량 제원을 활용하였는가?

Yes ☐ No ☐ N/A ☐

- 첨단 운전자 보조 시스템ADAS, Advanced Driver Assistance System을 포함한 자율주행 기능은, 특정 차량 및 센서에 종속적으로 개발하거나, 다양한 차량에 특정 하드웨어를 설치하여 범용적으로 활용할 수 있는 형태[37]로 개발되고 있다.
- 자율주행 알고리즘의 활용 목적 및 범위에 따라, 이미지를 이용한 범용 자율주행 인지 알고리즘을 개발하는 데이터를 직접 수집 시, 수집 환경 및 제약 조건(예: 일정한 지역, 단일 차량 제원) 때문에 다양한 데이터를 확보하기 어려울 수 있다. 이때는 다양한 제원의 차량을 활용하여 데이터의 수량과 다양성을 확보하는 것이 필요하다.

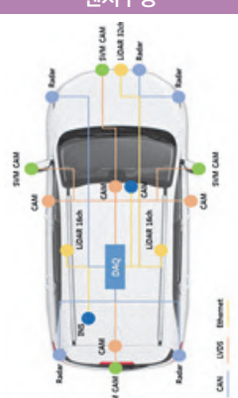
  ✓ 차량 제원 예: 차종, 전폭, 전고, 전장, 축간거리, 타이어 크기 등
- 다만, 이때 수집 경로 및 환경(예: 하드웨어 내 카메라 사양, 차량 제원)이 달라지므로, 수집 후 데이터를 활용하려면 데이터의 일관성이 유지되도록 데이터 정제·검수가 충분히 이뤄져야 한다.

참고 국토교통부의 자율주행 데이터 수집을 위한 차량 대여 사례[38]

- 국토교통부에서는 자율주행 데이터 수집을 위한 차량을 제작해 무상 대여 사업을 시행 중이다.

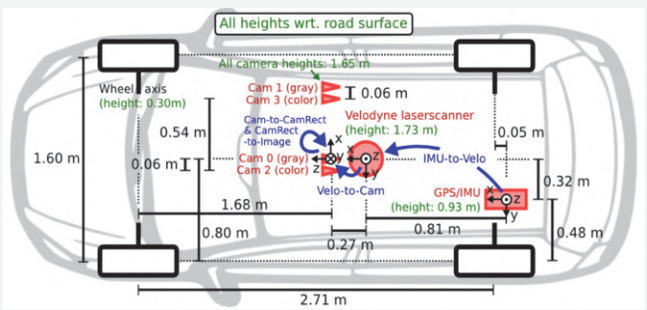
자율주행 데이터 수집용 SUV 및 세단

장착 센서 및 위치

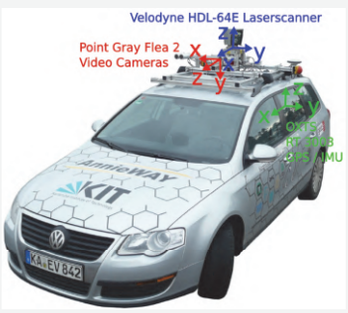
분류	사양	수량	센서구성
차량	- 5인승 SUV 디젤 2.0 - 5인승 승용차 가솔린 2.0	1대	
영상데이터	- FHD 카메라(전, 후, 좌, 우)	4EA	
	- SVM 카메라	4EA	
점군데이터	- 루프, 전방, 후방 레이더	3EA	
	- 전방, 후측방 레이더	3EA	
차량정보	- GPS/IMUInertial Measurement Unit - 차량신호(CAN) 수집장치	1식	
기타정보	- 기상센서	1식	
전산장비	- IPC(점군 데이터 저장 및 처리)	1식	
	- VSB(영상 데이터 저장 및 처리)	1식	

참고 KITTI 오픈 데이터셋 수집 정보[39]

• KITTI 오픈 데이터셋의 객체 인식, 추적 등을 위한 수집 데이터 정보



차량에 장착한 센서 개요



차량 좌표계 개요

센서 정보 및 수집 데이터

하드웨어 사양	모델 상세	데이터
Inertial Navigation System(GPS/IMU)	OXTS RT 3003	• 3D GPS/IMU data(text file) – 위치, 속도, 가속도, 메타 정보
Laser Scanner	Velodyne HDL-64E	• 3D Velodyne point clouds – 100k points per frame – Binary float matrix
Grayscale Cameras	Pint Grey Flea 2(FL2-14S3M-C) -1.4 Megapixels	• Grayscale stereo sequences – Raw, Processed(unsynced+unrectified) Format: png(0.5 Megapixels)
Color Cameras	Point Grey Flea 2(FL2-14S4C-C) -1.4 Megapixels	• Color stereo sequences – Raw, Processed(unsynced+unrectified) Format: png(0.5 Megapixels)

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

Yes ☐ No ☐ N/A ☐

- 데이터 수집 시 데이터 수집 장치(예: 카메라, 라이다 센서, 레이더 센서)를 차량에 부착하면 하드웨어로 인해 데이터 편향이 발생하므로 점검이 필요하다.
- 직접 주행하여 데이터를 수집할 때, 차량에 센서를 탈부착하는 게 쉽지 않으므로 단일 센서만 활용하여 데이터를 수집하여 활용할 수 있다. 이때, 특정 하드웨어 사양(예: 차량 사양, 센서 사양)에 대한 데이터만 수집되는 편향이 발생할 수 있다.
- 수집한 데이터를 이용한 인지 알고리즘을 다른 차량에 적용할 때는 데이터 편향으로 인한 알고리즘 인지 성능에 문제가 발생할 수 있으므로, 데이터 수집 시 이러한 요인을 점검, 대처하는 계획을 마련해야 한다. 아래의 표는 수집 센서별로 고려할 수 있는 하드웨어 사양의 예시이다.

데이터 편향 점검 시 고려해야 할 센서 종류별 하드웨어 사양 항목

수집 센서	하드웨어 사양 항목
RGB 카메라	칩셋 종류(예: CCD, CMOS), 해상도(pixel), 시야각FoV, Field Of View, 압축방식(예: H.265, H.264), 스캔방식(예: Progressive, Interlaced) 등
라이다	해상도(mm^2), 최대범위(예: cm^2, m^2), 시야각, 스캔 방식(주사식, 섬광식) 등
레이더	주파수 대역(Hz), 감지거리 범위(예: m), 수평 및 수직 반전력 빔폭half-power beamwidth 등

06-2 학습에 사용되는 특성feature을 분석하고 선정 기준을 마련하였는가?

Yes No N/A
☐ ☐ ☐

해당여부 판단

자율주행 인공지능 모델의 입력값으로 차별을 유발할 수 있는 민감한 변수를 활용하는 경우, 본 항목을 고려하여 만족 여부를 판단하십시오.

- 편향 완화를 위해 데이터에 포함된 차별적인 요소를 사전에 가려내는 것이 중요하며, 이를 위해 학습 관련 특성에 대한 분석과 선정 기준을 수립하는 것이 바람직하다. 차별적인 요소란 학습 결과가 사회적 물의나 차별을 일으킬 수 있는 특성을 말하며, 나이, 성별, 인종, 민족·사회적 기원, 언어, 장애 등이 있다.
- 자율주행 알고리즘의 학습 데이터셋에 민감한 특성 정보가 함께 기록될 때는 보호 변수를 설정하고, 인공지능에 미치는 영향을 분석하여 편향을 방지한다.
- 자율주행 알고리즘 중 이미지 또는 클라우드 포인트 데이터 외 다변수를 활용하여 인지·판단·제어 인공지능 알고리즘 개발 시 편향을 방지하도록 특성을 배제하거나, 특성이 과도하게 선택 또는 배제되지 않았는지 검토한다.

06-2a 보호변수protective attribute 선정 시 충분한 분석을 수행하였는가?

Yes No N/A
☐ ☐ ☐

- 현재 널리 활용되는 자율주행용 오픈 데이터셋은 민감한 특성 정보를 포함하는 경우가 드물다. 따라서 오픈 데이터셋을 활용하여 인지 알고리즘을 개발한다면 보호변수의 경우는 현 시점에서 고려 대상이 아닐 수 있다. (04-1c 참고)
- 그러나, 한 연구 결과에서는 자율주행 알고리즘에 객체 인식 인공지능을 활용할 때 백인보다 흑인이 5% 정도 덜 검출되는 것으로 분석하였다[26,40]. 이 연구는 동료 검토 및 실제 자율주행 차량 제조업체의 데이터를 대상으로 하지 않았기 때문에 사실로 받아들여지기 어렵지만, 추후 실제 자율주행차에서 발생할 수도 있는 문제이므로 충분히 고려해야 한다.
- 만약 자율주행용 학습 데이터셋에 민감한 특성(예: 보행자 객체의 연령, 성별, 인종) 정보가 함께 기록된다면, 보호변수를 선정해야 한다. 또한, 보호변수를 선정할 때 충분히 분석하지 않으면 모델 성능을 떨어뜨릴 수 있다는 점을 염두에 두어야 한다. 즉, 특정 변수가 불공평한 결과에 얼마나 영향을 미치는지, 성능 결과가 어떻게 달라지는지 등을 충분히 분석해야 한다.

06-2b 편향을 발생시킬 수 있는 특성의 영향력을 완화하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 모델 학습 시, 데이터의 특성을 선택하여 사용함으로써 효율적인 학습은 물론, 컴퓨팅 자원과 비용을 저감할 수 있으며 여러 특성 사이의 관계 분석 과정에서 데이터에 대한 깊이 있는 이해를 통해 잠재된 편향을 인식할 수도 있다.
- 현재 대부분의 자율주행 인지 알고리즘은 이미지 또는 포인트 클라우드 데이터를 입력받아 객체를 탐지하고, 이후의 진행 방향을 예측하는 기능을 수행한다. 따라서 현시점에서는 편향을 일으킬 수 있는 일부 특성을 선택 또는 배제하는 것은 고려 대상이 아닐 수 있다.
- 그러나 추후 자율주행 알고리즘 아키텍처에서 인지·판단·제어를 위한 데이터셋에 편향을 발생시킬 수 있는 특성이 추가되고, 그 정보를 이용하여 판단(예: 주행 경로 예측, 위험 판단) 및 제어하는 알고리즘을 개발하게 된다면 편향을 발생시키는 특성의 영향력을 완화해야 한다.

06-2c 데이터 전처리 시 특성이 과도하게 제거되었는지 검토하였는가?

Yes ☐ No ☐ N/A ☐

- 특성 선택 기법을 통해서 잠재된 편향을 완화하고 모델 성능을 향상시킬 수 있으나, 지나칠 경우 과적합 overfitting 문제 혹은 오히려 편향의 원인이 되기도 한다.
- 자율주행을 위한 인지 알고리즘 개발 시, 현시점에서는 편향을 일으킬 수 있는 일부 특성을 선택 또는 배제하는 것은 고려 대상이 아닐 수 있다. (06-2b 참고)
- 그러나 추후 자율주행 알고리즘 아키텍처에서 인지·판단·제어를 위한 데이터셋에 추가 특성이 반영되고, 해당 정보를 이용하여 판단(예: 주행 경로 예측, 위험 판단) 및 제어하는 인공지능 알고리즘을 개발하게 된다면 과도한 특성 선택 및 배제를 방지하기 위한 분석이 필요할 수 있다. 아래 표는 운전자의 차선 변경 의도를 파악하는 인공지능의 입력 특성을 비교 분석한 예시이다.

참고 운전자의 차선 변경 의도를 파악하는 인공지능의 입력 특성 비교 분석 예시

데이터	입력 변수(기능)		참고문헌									
			[41]	[42]	[43]	[44]	[45]	[46]	[47]	[48]	[49]	[50]
차량	세로방향	가스 페달 위치(%)	O			O			O			
		브레이크 페달 위치(%)	O									
		스로틀 위치(%)			O							O
		엔진 RPM(rpm)										O
		세로 위치(m)		O	O							
		종방향 가속도(m/s^2)	O	O			O		O			O
		앞차와의 거리(m)				O	O		O			O
		앞차와의 시간 거리(s)				O						O
		속도(km/h)	O						O			O
	측면방향	스티어링 각도	O	O	O	O	O		O	O	O	O
		스티어링 속도(m/s)									O	
		스티어링 힘(N)									O	
		요 각속도(rad/s)	O	O					O	O		O
		헤딩 각(rad or degree)		O								O
		측면 위치(m)	O	O	O	O					O	
		차선 내 위치								O		
		차선 중심 간격 띄우기(m)										O
		횡방향 가속도(m/s^2)	O				O					O
		신호 켜기(on or off)			O				O	O		
도로	도로 곡률(m-1)		O							O		
	측면 차선의 존재 여부(on or off)					O						
	기울기(degree of rad)								O			
	비디오 이미지							O				
운전자	머리 각도		O							O		O
	눈동자 각도				O			O	O			O
알고리즘	기계학습 알고리즘		SBL	HMM			SVM			RVM	HMM	RF
주변 상황	현실세계		O			O	O	O	O	O		
	모의실험 장치			O	O	O					O	O

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

Yes ☐ No ☐ N/A ☐

해당여부 판단

자율주행 인공지능 학습용 데이터를 직접 수집 또는 라벨링하는 경우 본 항목을 고려하여 만족 여부를 판단하십시오.

- 현재의 자율주행 분야 인지를 위한 인공지능 모델에서는 학습 데이터에 대한 라벨링이 요구된다. 그러나, 라벨링 작업 시에 작업자의 특정 의도 반영, 실수로 인한 특성 정보의 누락, 무의식적인 판단으로 인한 편향이 발생할 수 있다.
- 데이터 라벨링 과정에서 발생할 수 있는 이슈를 인지하고, 사전에 명확한 표준 또는 작업 가이드라인을 마련하여 작업자에게 제공, 교육해 추후 다른 문제가 발생하지 않도록 하고, 편향의 발생을 방지한다.
- 자율주행을 위한 원시 데이터(예: 비디오, 포인트 클라우드)를 이해하고, 충분한 상황 분석을 기반으로 라벨링할 수 있는 배경지식과 요건을 갖춘 작업자와 검수자를 폭넓게 섭외하여 작업자별로 나타날 수 있는 오류와 편향을 최소화하고, 편향 방지 작업을 수행하는 것이 바람직하다.

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

Yes ☐ No ☐ N/A ☐

- 자율주행용 인지 알고리즘의 데이터 라벨링 과정에서 발생할 수 있는 이슈를 인지하고, 사전에 작업자에게 표준 또는 명확한 가이드라인을 마련하여 제공하고 교육을 실시하여 추후 다른 문제가 발생하지 않도록 한다.
- 라벨링 표준 또는 가이드라인의 부재 시 다음과 같은 이슈가 발생할 수 있다[51].
  - ✓ 방해받는 차량 대 차량 상호 작용hampered Vehicle2Vechicle interaction: 시나리오 데이터 라벨링 시 주변 환경에 대한 설명과 이해가 달라, 둘 이상의 서로 다른 자율주행 시스템ADS, Autonomous Driving System이 포함된 복잡한 상황에서 사상자가 발생할 수 있음
  - ✓ 공유 한계precluded sharing: 서로 다른 라벨링 분류 및 사양을 활용한 데이터셋 또는 조직 간의 데이터 오류 발생 확률이 높아질 수 있어 데이터를 공유하는 것이 어려움
  - ✓ 주석 품질 감소reduced annotation quality: 개별 라벨링 작업에 임시 교육이 필요하며, 안전을 위협하는 작업자의 오류를 줄이기 위해 자동·반자동 등 도구도 적용해야 함
- 자동차 개발 및 테스트에서 툴 체인의 표준화를 촉진하는 비영리 조직인 ASAMAssociation for Standardization of Automation and Measuring systems에서는 ASAM e.V. 및 ASAM OpenLABEL 등 자율주행 분야의 데이터 라벨링 표준을 제시하였다.

Use Case 국내 S사의 3D Cuboid 객체 데이터 라벨링 가이드라인 작성 사례(일부 발췌)

• 국내 S사에서는 자체적으로 구축 중인 데이터의 라벨링을 위해 라벨링 작업 가이드 문서를 관리하고 있으며, 이를 라벨링 작업자가 참고할 수 있도록 함

• 라벨린 작업 방식: 3D Cuboid

• 라벨 대상(총 8 classes)

대상	구분	라벨	세부 기준
자동차	자동차_세단	car_sedan	– 보이는 점군을 모두 태깅 후 해당 라벨로 분류 – 승용차, 버스를 제외한 기타차량(트럭, 특수 차량)에 대한 액세서리(박스/짐/추가구조물) 는 보이는 점선을 포함하여 태깅
	자동차_밴	car_van	
	자동차_버스	car_bus	
	자동차_트럭	car_truck	
	기타 차량	other vehicles	
보행자 및 개인형 이동수단	보행자	pedestrian	– 사람의 자세와 상관없이 전부 태깅
	자전거 탑승자	bicyclist	– 이륜차에 탑승한 탑승자 태깅(자전거, 오토바이 모두 포함) – 탑승한 것만 보이고 자전거 바퀴가 안 보이는 경우 보이는 점군까지만 작업 – 보행자 기타는 휠체어 등에 앉아 있는 사람을 태깅
	보행자 기타	other person	

– 세단: 세단, 해치백, SUV 포함하여 4~5인승 탑승 차량
– 밴: 봉고차, 승합차, 미니밴
– 트럭: 전면 확인 시 일반적인 트럭 형태와 유사한 것으로 소방차, 트럭, 사다리차, 탑차 포함
– 기타 차량: 세단, 밴, 트럭에 포함되지 아니한 분류로, 지게차, 포크레인, 레미콘, 그 외 포함

• 가공 대상 기준
– Ego vehicle로부터 25m 이내, camera(Front, Left, Right)에 인식되는 객체를 대상으로 작업함
– Front, Left, Right 이미지에서 객체 일부라도 보이면 작업, 보이지 아니하면 작업 제외
– 차량 일부가 보여도 점선 형태가 보이는 데까지만 작업
– 라벨 명단에 적혀 있는 대로 태깅 및 분류 작업
– 연속 이미지(sequence)의 경우 라벨 추적하여 작업
– 태깅축은 차량 보닛, 사람이 정면을 바라보고 있는 방향에서부터 작업

• 가공 대상 제외
– 객체의 식별이 불가능한 경우(점의 형태가 완전하지 않은 경우)
– 이미지 내에 식별이 가능하나, 점의 형태가 해당 객체로 추정이 안 될 경우

06-3b 다양한 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
- 자율주행 알고리즘 데이터 라벨링 작업자가 실제로 다양하고 고르게 분포하는지 확인하려면 배경지식을 조사하고 분석하여 다양성을 검증해야 한다. 이때 검토할 만한 배경지식은 다음과 같다.
  - ✓ 데이터 라벨링 작업자의 배경지식 및 요건: 운전 경력, 운전 지역, 연령대, 사고 경험, 운전 교육, 차량 정비 등

06-3c 다양한 라벨링 검수자를 확보하기 위해 노력하였는가?

Yes ☐ No ☐ N/A ☐

- 다양한 데이터 라벨링 작업자를 확보했음에도 불구, 인적 편향이 발생할 수 있다. 따라서, 데이터 라벨링 검수자를 확보하고, 라벨링 결과가 데이터 수집 목적 및 데이터 스펙과 다른 부분은 없는지 등을 확인하며, 수정을 요청하는 등의 작업을 실시해야 한다.
- 데이터 라벨링 작업자와 마찬가지로 데이터 라벨링 검수자 역시 다양하고 고르게 분포되도록 구성하는 것이 바람직하다. 그러므로 검수자를 조사하고 분석하여 다양한 배경지식 및 요건을 갖춘 검수자의 분포가 다양하고 고르게 형성되는지 점검한다.
- 라벨링 검수 시에는 추후 데이터셋 내에 존재하는 이벤트 또는 시나리오를 분류하고, 분석 결과를 검수해야 하므로 변호사, 사고감정사 등 많은 검수자를 확보해야 한다.

06-4 데이터의 편향 방지를 위한 샘플링을 수행하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

자율주행 인공지능 학습용 데이터를 직접 구축하거나, 인공지능 모델 개발 시 클래스 또는 보호 변수 등으로 인해 편향이 예상되는 경우 본 항목을 고려하여 만족 여부를 판단하십시오.

- 샘플링은 모집단에서 일정한 기준으로 데이터를 추출하여 표본을 만드는 기법이다. 일정한 기준으로 추출된 표본은 모집단의 분포를 대표하는 동시에 실제 모집단의 클래스 불균형으로 인한 편향 또한 방지하여야 한다.
- 자율주행 알고리즘 중 인지 알고리즘의 객체 인지 및 분류 문제를 해결하기 위한 데이터셋 내 클래스 불균형 문제는 적은 수의 클래스 분포를 제대로 학습하지 못하게 하므로 샘플링 기법을 적용하여 의도하지 않은 편향을 방지한다.

06-4a 편향 방지를 위한 샘플링 기법을 적용하였는가?

Yes ☐ No ☐ N/A ☐

- 자율주행 분야의 데이터셋을 대상으로 사회적 편견 또는 차별을 일으킬 수 있는 편향의 요소는 아직 주요하게 고려되고 있지 않으나, 객체 인지 알고리즘에서 발생할 수 있는 연령 또는 인종 차별 가능성(04-1c 참고)에 따라 추후 판단·제어를 위한 학습 데이터셋을 대상으로 할 때는 다음과 같은 인구통계학적 샘플링 기법을 적용할 수 있다.
  - ✓ 확률 샘플링: 단순 무작위 샘플링simple random sampling, 체계적 샘플링systematic sampling, 층화 샘플링stratifield sampling, 클러스터 샘플링cluster sampling
  - ✓ 비확률 샘플링: 편의 샘플링convenience sampling, 자발적 응답 샘플링voluntary response sampling, 목적 샘플링purposive sampling, 눈덩이 샘플링snowball sampling, 할당량 샘플링quota sampling
- 자율주행 인공지능 알고리즘 중 인지를 위한 알고리즘은 현재 객체의 종류(클래스)를 구별하는 데 상당 부분 집중되어 있다. 많은 오픈소스 데이터셋이 객체 인지 및 분류 문제를 해결하기 위한 데이터셋 구조를 제공하고 있어, 자연스럽게 클래스 불균형imbalance 문제가 발생한다.
- 클래스 불균형 문제를 해결하기 위해서 언더 샘플링under sampling, 오버 샘플링over sampling 기법 등을 활용할 수 있다. 객체 클래스의 불균형으로 인지 편향이 예상되는 경우, 이를 방지할 수 있는 샘플링 기법을 적용하고, 적용 과정에서 필요한 활동과 정보가 생성되었는지 확인해야 한다.

03 인공지능 모델 개발

책임성 안전성

요구사항 07 오픈소스 라이브러리의 보안성 및 호환성 확보

대표 행위자 | 인공지능 모델 개발자 협력 대상 | 시스템 엔지니어

- 인공지능 모델 개발 단계에서 기간을 단축하고 최신 기술 동향을 빠르고 유연하게 적용하기 위해 다양한 오픈소스 라이브러리를 활용한다. 오픈소스 라이브러리를 도입하기에 앞서 필요성 여부, 필요한 기능의 포함 여부 등을 확인한다. 자율주행 인공지능 모델은 안전성과 성능 등 측면에서 요구사항이 많으므로, 라이브러리의 신뢰수준, 안정적인 업데이트 여부, 주의할 라이선스 기준, 오픈소스 라이브러리의 보안 취약점 등 해당 오픈소스의 버전을 지속해서 확인하여 운영 및 보안상 위험 요소를 점검한다.

07-1 오픈소스 라이브러리의 안정성을 확인하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

자율주행 인공지능 모델 개발 시 한 가지 이상의 오픈소스 라이브러리를 활용하는 경우, 본 항목을 고려하여 만족 여부를 판단하십시오.

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

07-1a 활성화된 오픈소스 라이브러리를 사용하였는가?

Yes No N/A
☐ ☐ ☐

- 오픈소스 라이브러리의 안정성은 많은 개발자가 적극적으로 참여할 때 가능하다는 의견이 있다. 따라서, 사용하려는 오픈소스 라이브러리의 개발과정을 주의 깊게 살펴볼 필요가 있다.
- '기업 공개소프트웨어 거버넌스 가이드-정보통신산업진흥원'에 따르면, 오픈소스 프로젝트의 활성화 정도를 확인하는 것도 안정성을 확인하는 한 가지 방법일 수 있다. 해당 오픈소스가 활발한 커뮤니티에서 논의되는지, 그 커뮤니티 내 구성원들이 적극적으로 협력하고 있는지는 아주 중요한 선택의 표지석일 수 있다.
  - ✓ 오픈소스 라이브러리를 GitHub에서 관리 중이라면, 오픈된 이슈 개수나 Pull Request 수, 마지막 커밋 일시 등을 통해 오픈소스 개발이 얼마나 활발하게 이루어지고 지속해서 발전할 가능성이 어느 정도인지 파악할 수 있다.
  - ✓ 그 밖에도 해당 오픈소스와 관련된 StackOverflow 질문 수, 오픈소스 다운로드 수, Google 쿼리 query 결과 수 등 간단한 측정을 통해서 해당 라이브러리의 활성화 정도를 확인할 수 있다.
  - ✓ Redhat의 경우, 오픈소스 기반의 수익화 모델(호환성, 보안 강화, 기술지원 등 제공)을 개발하고 있으며, 오픈소스 라이브러리 업데이트 시 커뮤니티 내 구성원들이 제안한 개선 사항도 적용한다. 이처럼 수익화 모델 기반의 오픈소스 라이브러리 역시 개인 및 기업의 참여가 활성화된 프로젝트로 판단할 수 있다.

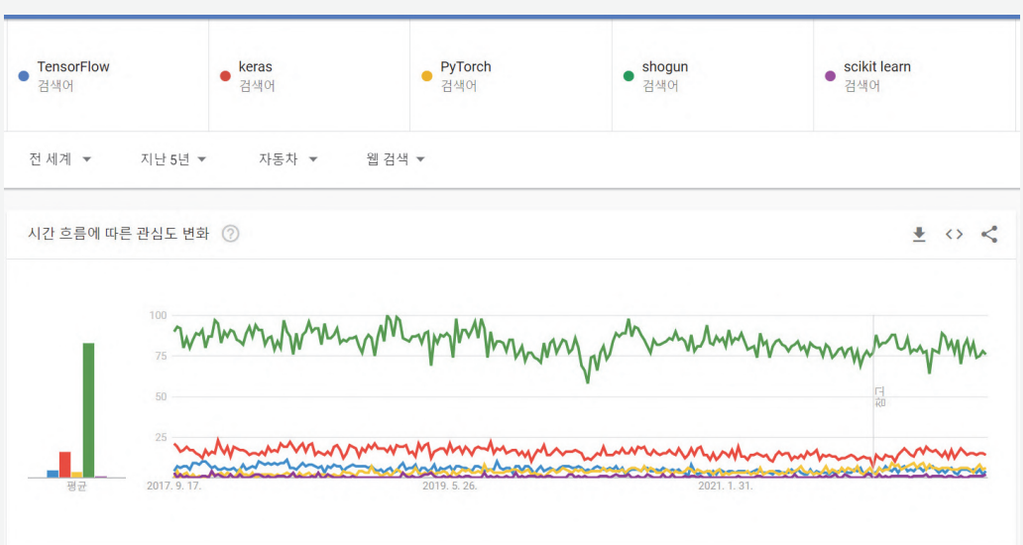
참고 자율주행 모델 개발 시 주로 활용되는 오픈소스 라이브러리 활성화 분석 예시 – GitHub 관리 및 측정

- GitHub 출처 분석 예시(2022.09.13. 기준): 가장 활발하게 활동하는 것은 PyTorch, TensorFlow, Scikit-Learn 정도이고, GitHub 라이브러리 기준 Shogun, CNTK는 2020년 이후 커밋이 중단되었음

오픈소스 라이브러리 / 항목	A	B	C	D	E	F	G	H	I
오픈 이슈 개수	2,100	251	5,000 이상	415	1,500	754	1,800	-	274
Pull Request 수	227	82	827	11	607	87	206	65	11
마지막 커밋 일시	22.9.1	22.9.13	22.9.13	20.12.9	22.9.13	20.4.1	22.9.13	22.9.13	22.9.3
Contributor 수	3,199	1,054	2,433	175	2,487	201	874	162	107
Used 수	210,000	-	160,000	-	381,000	-	-	-	1,000
StackOverflow 질문 수	78,966	40,565	18,635	59	500	503	693	77	89

[A] Tensorflow, [B] Keras, [C] PyTorch, [D] Shogun, [E] Scikit-Learn, [F] CNTK, [G] mxnet, [H] H2O.ai, [I] apple core ML

- Google Query 분석 예시(2022.09.13. 기준): 과거 5년간, 자동차 카테고리에서 Shogun 오픈소스 라이브러리의 쿼리 수가 가장 많고, keras, TensorFlow 순으로 조회수가 많음



TensorFlow, keras, PyTorch, shogun, scikit learn
라이브러리의 자동차 카테고리 내 과거 5년간 검색어 조회수(관심도) 변화

07-2 오픈소스 라이브러리의 위험요소는 관리되고 있는가?

Yes No N/A
☐ ☐ ☐

해당여부 판단

자율주행 인공지능 모델 개발 시 한 가지 이상의 오픈소스 라이브러리를 활용하는 경우, 본 항목을 고려하여 만족 여부를 판단하십시오.

- 오픈소스 라이브러리 또는 소프트웨어는 저작권자가 소스코드를 공개했을 뿐이며 지식재산권IP, Intellectual Property으로 보호받는 소프트웨어이다. 따라서, 저작권자가 제시한 라이선스(저작권) 준수 조건이 엄연히 존재하며, 오픈소스 라이브러리마다 라이선스에 따라 다양한 의무 사항이 있다. 이때, 라이선스 위반 및 저작권 침해로 법적 책임을 져야 할 위험이 있으므로 반드시 라이선스와 관련한 위험 요소를 분석하고 관리해야 한다.
- 개발 과정에서는 오픈소스 라이브러리 변경, 개발 환경 변경 등 서로 다른 오픈소스 라이브러리 또는 오픈소스 라이브러리의 버전 변경에 따른 호환성을 고려하여 종류 및 버전을 선택해야 한다. 학습이 완료된 인공지능 모델은 실행환경에서 모델 파일을 로드하고 추론inference을 실행하기 위해 학습 환경과 동일한 오픈소스 라이브러리를 설치해야 한다.
- 이때 실행 환경에 설치한 오픈소스 라이브러리로 인하여 보안 취약점이나 위험이 발생할 수 있다. 이로 인한 영향을 최소화하기 위해 버전 변경에 따른 릴리즈 노트release note를 지속적으로 확인하고 보안 취약점을 검사하는 등 적극적으로 대응해야 한다.

07-2a 사용 중인 오픈소스 라이브러리의 라이선스 준수사항을 이행하였는가?

Yes No N/A
☐ ☐ ☐

- 오픈소스는 무료로 사용할 수 있지만, 라이선스별로 준수사항은 별도로 규정된다. 그러므로 오픈소스 라이브러리를 활용하여 인공지능 모델을 개발한다면, 사용할 오픈소스의 라이선스 종류 및 라이선스 고지문을 확인하고, 허용 또는 의무 사항을 우선해서 숙지해야 향후 발생할 수 있는 법률적 위험을 최소화할 수 있다.
- 다음은 OSIOpen Source Initiative 단체에서 정한 10개 항목으로 된 오픈소스 기준이다[52].
  - ✓ 자유로운 재배포 (Free Redistribution)
  - ✓ 소스코드 공개 (Source Code Open)
  - ✓ 2차 저작물 허용 (Derived Works)
  - ✓ 저작자의 소스코드 원형 유지 (Integrity of The Author's Source Code)
  - ✓ 개인이나 단체에 대한 차별 금지 (No Discrimination Against Persons or Groups)
  - ✓ 사용 분야에 대한 차별 금지 (No Discrimination Against Fields of Endeavor)
  - ✓ 라이선스의 배포 (Distribution of License)

✓ 특정 제품에만 유용한 라이선스 금지 (License Must not be specific to a product)
✓ 다른 소프트웨어를 제한하는 라이선스 금지 (License Must not contaminate other software)
✓ 기술 중립적인 라이선스 제공 (License must be Technology-Neutral)

상위 활용 오픈소스 라이브러리의 OSI 기준definition 분석

오픈소스 라이브러리 / OSI 기준	Tensorflow	Keras	PyTorch (caffe2)	Shogun	Scikit-Learn	CNTK	mxnet	H2O.ai	apple core ML
	Apache License 2.0	Apache License 2.0	Apache License 2.0	BSD3	BSD3	MIT	Apache License 2.0	Apache License 2.0	BSD3
자유로운 재배포	O	O	O	O	O	O	O	O	O
소스코드 공개	의무X	의무X	의무X	의무X	의무X	의무X	의무X	의무X	의무X
2차 저작물 허용	O	O	O	O	O	O	O	O	O
저작자의 소스코드 원형 유지	가능	가능	가능	가능	가능	가능	가능	가능	가능
개인이나 단체에 대한 차별 금지	차별X	차별X	차별X	차별X	차별X	차별X	차별X	차별X	차별X
사용 분야에 대한 차별 금지	없음	없음	없음	없음	없음	없음	없음	없음	없음
라이선스의 배포	O	O	O	O	O	O	O	O	O
특정 제품에만 유용한 라이선스 금지	O	O	O	O	O	O	O	O	O
다른 소프트웨어를 제안하는 라이선스 금지	O	O	O	기재X	기재X	기재X	O	O	기재X
기술 중립적인 라이선스 제공	O	O	O	기재X	기재X	기재X	O	O	기재X

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

Yes ☐ No ☐ N/A ☐

- 라이브러리의 버전 변경 과정에서 개발 환경, 언어, 도구 및 다른 라이브러리 버전과 호환되지 않는 호환성 문제를 초래할 수 있다. 따라서 오픈소스 라이브러리 종류 및 버전 선택 시 라이브러리 간 의존성dependency을 파악하는 등 호환성을 고려해야 한다.
- 사용 중인 오픈소스 라이브러리에서 보안취약점이 발견되기도 한다. 보안 취약점에 따른 영향을 최소화하기 위해 보안취약점 및 버전 변경에 따른 릴리즈 노트release note를 지속해서 확인하여 신속히 탐지 및 대응해야 한다.
- 현재 자율주행을 위한 인지 알고리즘은 통신에 따른 보안 위협은 비교적 적은 편이나 취약점 점검을 소홀히 해서는 안 된다. 더불어, 추후 V2X를 통해 주변 사물, 객체와의 통신이 이루어질 때는 보안 취약점을 더 꼼꼼히 점검해야 한다.

참고 오픈소스 라이브러리 호환성 예시

- TensorFlow, PyTorch 모델 호환 예시: SCATTER LAB의 하나의 조직에서 TensorFlow와 PyTorch를 동시 활용하기[53]
  - ✓ 유연한 리서치를 위해서는 PyTorch가 유리하고, 배포 측면에서는 TensorFlow가 유리하여 동시에 사용
  - ✓ 이를 위해 내부에서 사용하는 모델들을 PyTorch, TensorFlow 버전으로 다시 작성하고, 모델의 모든 가중치를 변환해주는 코드를 추가로 작성
  - ✓ 그 결과, TensorFlow Checkpoint에서 PyTorch 모델로 적용, PyTorch State Dict 파일에서 TensorFlow 모델로 적용이 가능해졌음
- TensorFlow, OpenVINO 모델 호환[54]
  - ✓ OpenVINO 개발 도구를 활용하여 TensorFlow 1.x, 2.x 모델 형식에서 OpenVINO IR 형식으로 변환
  - ✓ Frozen 모델 포맷(.pb 파일) 및 Non-Frozen 모델 포맷(Checkpoint, MetaGraph, SavedModel)의 변환 가능
- TensorFlow1, 2, Keras 저장 모델 호환[55]
  - ✓ TensorFlow 2에서는 TensorFlow 1에서 저장된 모델이 호환되어 변수와 함수 로드 가능
  - ✓ TensorFlow 2에서는 Keras로 저장된 모델이 호환되어 로드 가능

참고 오픈소스 라이브러리의 보안 취약점 분석 예시

TensorFlow CVE[Common Vulnerabilities and Exposures] 예시[56](2022.09.13. 기준)

- DoS[Denial of Service] 공격에 취약한 부분이 존재하는 것으로 분석되고(24.8%), Overflow 위협도 존재하는 것으로 분석(20.4%)
- 총 보안 취약점은 2021년 201건에서 2022년에 80건으로 줄어들어 제조사 측에서 어느 정도 보안 위협에 대응하고 있는 것으로 분석

Year	# of Vulnerabilities	DoS	Code Execution	Overflow	Memory Corruption	Sql Injection	XSS	Directory Traversal	Http Response Splitting	Bypass something	Gain Information	Gain Privileges	CSRF	File Inclusion	# of exploits
2019	7	1	1	4											
2020	35	6	2	8	3										
2021	201	41	6	38	1			1		8	1				
2022	80	32	1	16						2					
Total	323	80	10	66	4			1		10	1				
% Of All		24.8	3.1	20.4	1.2	0.0	0.0	0.3	0.0	3.1	0.3	0.0	0.0	0.0	

2019~2022년 CVE 보안 취약점 분석 결과

PyTorch CVE 예시[56](2022.09.13. 기준)

- 보안취약점 분석 결과, 2021년과 2022년에 각 1건씩 보안 위협이 발견되었음

Year	# of Vulnerabilities	DoS	Code Execution	Overflow	Memory Corruption	Sql Injection	XSS	Directory Traversal	Http Response Splitting	Bypass something	Gain Information	Gain Privileges	CSRF	File Inclusion	# of exploits
2021	1														
2022	1														
Total	2														
% Of All		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

2021~2022년 CVE 보안 취약점 분석 결과

- 2022년 발견 항목 – CWE-94: Failure to Control Generation of Code('Code Injection')
  - ✓ 보안 위협 내용: 제품이 생성하는 코드 내에서 해당 입력이 사용될 때 사용자 제어 입력(데이터 평면)의 코드(제어 평면) 구문을 충분히 필터링하지 못함. 소프트웨어에서 사용자의 입력에 코드 구문이 포함되도록 허용하는 경우 공격자가 소프트웨어의 의도된 제어 흐름을 변경하는 방식으로 코드를 작성할 수 있음. 이러한 변경으로 인해 임의 코드가 실행될 수 있음 등

다양성 존중

요구사항 08 인공지능 모델의 편향 제거

대표 행위자 | 인공지능 모델 개발자 협력 대상 | 시스템 엔지니어

- 인공지능 모델을 개발하는 과정에서 모델 종류나 시스템 목표에 따라 발생할 수 있는 편향을 제거하기 위한 기법을 고려해야 한다. 자율주행 인공지능 모델은 크게 판단·제어 부분에서 개발자의 인지 편향이 발생할 수 있고, 모델의 추론 결과를 토대로 운전자나 보행자 등에게 알릴 때 자동화 편향이 있을 수 있기 때문이다. 또한, 운행 도중 다양한 주행 상황에서 각 인지·판단·제어 인공지능 모델이 편향 없이 목표 성능을 달성할 수 있는지 분석하고, 모니터링할 수 있는 기법과 지표를 적용해야 한다.

08-1 모델 편향을 제거하는 기법을 적용하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

자율주행 인공지능 모델 개발 시 차별·편향을 유발할 수 있는 민감한 특성이 입력값 또는 출력값에 활용되거나 운전자의 인지 편향 등이 예상되는 경우, 본 항목을 고려하여 만족 여부를 판단하십시오.

- 인공지능 모델은 데이터에 잠재된 편향을 학습하거나 심지어 편향을 더욱 증폭시키기도 한다. 따라서, 데이터 정제 단계에서 데이터에 잠재된 편향을 제거하는 방법뿐만 아니라, 모델 개발 과정에서도 모델 편향을 제거 또는 완화하기 위한 기법을 적용하는 것이 바람직하다.
- 인지 알고리즘의 데이터셋에 민감한 특성이 포함되지 않았다면, 편향 제거 또는 완화 기법을 적용하거나 편향성을 평가하고 모니터링하기 위한 정량적 지표의 선정은 고려 대상이 아닐 수 있다.
- 추후, 민감한 특성이 반영된 데이터셋을 활용하여 인공지능 모델을 개발하거나 판단·제어를 위한 인공지능 모델을 개발할 때는 목표 임무에 따라 적절한 편향 완화 기법, 편향성을 평가하고 모니터링하기 위한 정량적 지표의 선정 및 적용이 필요할 것이다.

08-1a 개발하려는 모델에 맞게 편향제거 기법을 선택하였는가?

Yes ☐ No ☐ N/A ☐

- 현재 널리 활용되는 자율주행용 오픈 데이터셋은 민감한 특성 정보를 포함하는 경우가 드물다. 따라서 오픈 데이터셋을 활용하여 인지 알고리즘을 개발하는 경우, 사회적·윤리적으로 문제가 되는 편향을 제거하는 기법은 현시점에서는 고려 대상이 아닐 수 있다.
- 그러나 추후 자율주행 알고리즘 아키텍처 중 인지·판단·제어를 위한 데이터셋에 민감한 특성(예: 인종, 성별, 나이)이 반영되거나, 판단·제어 시 사람의 편향을 유발할 수 있으므로 다음과 같은 기법을 고려해야 한다.

발생 가능한 편향에 따른 적용 가능 기법

편향 유형	기법 (접근 방법)	기법 구분			설명
		Pre	In	Post	
인지 편향cognitive bias	다양한 결정 계획 수립		✓		인지·판단 시 발생할 수 있는 편향으로, 다양한 팀 또는 전문가의 도움으로 인지 편향 완화 가능 (예: 주변 차량 주행 예측 후 차선 변경 판단, 추월 및 장애물 기동 판단)
알고리즘 편향algorithmic bias	가중치 재지정	✓	✓		인지 시 데이터셋의 서브셋 불균형으로 발생할 수 있는 편향 (예: 피부색 그룹별 서브셋의 데이터 크기 차이로 인지 알고리즘 결과 편향 발생, 날씨(또는 이미지 선명도) 그룹별 서브셋의 데이터 크기 차이로 인지 알고리즘 결과 편향 발생)
평가 편향evaluation bias	임곗값		✓		판단·제어 시 차량간 치명타 충돌 등 윤리적 측면의 딜레마 상황[57] 및 판단·제어 시 자율주행 차량의 도덕적 딜레마에 대한 인간의 의사 결정 편향[58]
자동화 편향automation bias	자동화 시스템 감독			✓	레벨 2, 3(일부), 4(일부)에서 자율주행 알고리즘 및 시스템의 자동화로 발생할 수 있는 편향 (예: 자율주행 또는 자동화 운행 시 운전자별 주의가 부족해 편향 발생[59])

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

Yes ☐ No ☐ N/A ☐

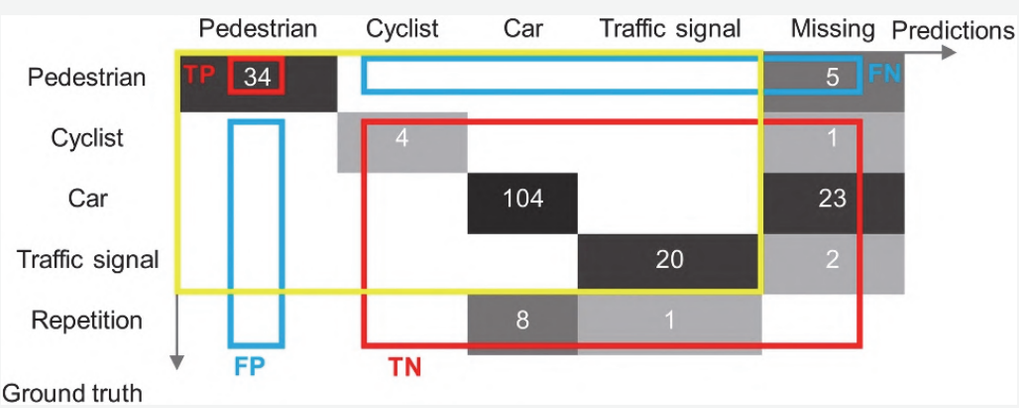
- 현재 널리 활용되는 자율주행용 오픈 데이터셋은 민감한 특성 정보를 포함하는 경우가 드물다. 따라서 오픈 데이터셋을 활용하여 인지 알고리즘을 개발하는 경우, 사회적·윤리적으로 문제가 되는 편향성에 대한 평가나 모니터링을 위한 정량적 지표는 아직 선정 및 관리해야할 대상이 아닐 수 있다.
- 그러나 추후 자율주행 알고리즘 아키텍처 중 인지·판단·제어를 위한 데이터셋에 민감한 특성(예: 인종, 성별, 나이)이 반영되거나, 판단·제어 시 사람의 편향을 유발할 수 있으므로 다음과 같은 지표를 고려하고, 지속적으로 측정, 모니터링해야 한다.

발생 가능한 편향에 따른 모니터링 지표

편향 유형	지표 (모델)
인지 편향cognitive bias	인지 편향 척도DACOBS, Davos Assessment of Cognitive Biases Scale, 크론바흐 알파 계수Cronbach's alpha test
알고리즘 편향algorithmic bias	패리티parity 기반 지표: 인구통계학적statistical/demographic 형평성 지표, 차등적disparate 효과 지표
평가 편향evaluation bias	혼동 행렬confusion matrix 기반 지표: 동등 기회equalized opportunity, Equalized Odds, 전체 정확도 형평성, 조건부 사용 정확도 형평성, 대응 형평성, 비보상 동등화
자동화 편향automation bias	스위스 치즈 분석Swiss cheese analysis[59]

참고 자율주행 인지 알고리즘의 평가 편향 모니터링을 위한 혼동 행렬 분석 사례[60]

- 눈 오는 상황의 하위 그룹에 대한 확장된 혼동 행렬 분석 결과
  - ✓ 보행자 범주에서 5개 개체가 누락되어 5개 거짓 음성FN, False Negative 결과가 나옴



참고 2016년 Tesla 모델 S와 트랙터 간 충돌 사고에 대한 스위스 치즈 분석 사례[59]

- 스위스 치즈 분석은 실패에 대한 위험 원인 분석 및 사전 위험 저감을 위해 활용되는 분석 기법으로, 자율주행 차량의 자동화 편향 평가 및 모니터링에 활용된 사례가 있다.
- 미국 도로교통안전국NHTSA,National Highway Traffic Safety Administration 에서 스위스 치즈 분석을 사용하여 자율주행 차량의 사고를 분석한 결과, 운전자가 자율주행에 과의존하게 되는 자동화 편향으로 인한 것임을 확인하였다.

슬라이스	설명	치즈 구멍	활성화된 실패/지연된 실패	주제 관련성
자율주행 기능	• Tesla에는 레이더, 초음파 센서 및 카메라가 장착되어 있었고, 다른 차량 및 장애물을 감지하고, 신경망으로 객체를 분류하고 경로를 예측하는 소프트웨어 파워로 구동됨	• Tesla는 앞 장애물의 반 진입 정도를 탐지하지 못하였음. 아니라면 운전자에게 이상으로 경고할 수 있었을 것임	지연된 실패	한계
자율주행 기능	• Tesla에는 크루즈컨트롤이 오토 스티어링, 차선 변경, 비상 제동 기능을 포함한 교통 인식 장치가 장착되어 있음	• 차량이 속도를 늦추거나 회피하는 등 트랙터의 세미 턴에 반응하지 않음	지연된 실패	한계
자율주행 기능 확인에 대한 운전자의 자각	• Tesla는 스티어링휠에 손을 대고 있는 것으로 사람이 운전에 주의를 기울이고 있는지 알 수 있음	• Tesla는 도로 타입에 따라 운전자가 손을 놓는 시간을 더 연장하는 것을 허용하였음	지연된 실패	설계
자율주행에 있어 운전자의 멘탈 모델	• 운전자는 다음을 이해해야 함 – 레벨 2 자동화의 요구사항 – 레벨 2 자동화에는 인간이 이상을 인식하면 항상 운전을 할 준비가 되어 있어야 함	• 운전자가 그의 Tesla Autopilot의 광범위한 사용 및 과신에 대한 영상을 유튜브에 게시하였음	지연된 실패	안주
3자 행동	• 다른 운전자는 반드시 적절히 응답해야 함	• 다른 차량은 양보 실패로 인용되었음	활성화된 실패	코너 사례
3자 조건	• 다른 운전자는 반드시 적절하게 대응할 수 있어야 함	• 다른 차량 운전자는 사고 후 테스트에서 마리화나 양성 반응이 나왔음	지연된 실패	코너 사례
운전자의 주의력	• 운전자가 레벨 2 자율주행 차량을 운전하고 있기 때문에 인간 운전자는 반드시 자율주행 차량 또는 3자의 기능 실패에 따른 실패 안전에 대응하고 있어야 함	• 인간 운전자는 사고 나기 전 37분의 주행 동안 25초만 스티어링휠에 손을 대고 있었음	활성화된 실패	주의력, 안주
자율주행 행동	• 자율주행 차량은 움직임을 방해하는 물체의 모든 것을 감지하고 분류해야 함 • 레벨2 자동화에서는 인간 운전자에게 경고와 주의를 항상 알리지는 않아도 됨	• Tesla는 트랙터를 방해물로 분류하는 데 실패하였음	활성화된 실패	한계
운전자의 행동	• 운전자는 도로에서 예상하지 못한 상황이 발생했을 때 보자마자 운전 권한 인수를 시작해야 함	• 스티어링휠에서 어떤 터치나 토크가 감지되지 않았고, 충돌 전 Tesla의 페달에서 어떤 압력도 적용되지 않았음	활성화된 실패	인수, 주의력

안전성

요구사항 09 인공지능 모델 공격에 대한 방어 대책 수립

대표 행위자 | 인공지능 모델 개발자 협력 대상 | 시스템 엔지니어

- 자율주행 알고리즘 중 인지를 위한 인공지능 모델은 시각적 정보(예: 이미지, 포인트 클라우드)를 분석하여 객체의 위치·크기·거리 등을 판단한다. 따라서 시각적 정보를 처리하는 모델을 대상으로 자행될 수 있는 추출이나 회피 등 적대적 공격에 취약할 수 있다. 공격에 대한 방지 또는 완화 방안을 수립해 적대적 의도를 가진 사용자가 학습 데이터와 기능을 도용하는 등의 공격에 대비한다.

09-1 모델 추출 공격(model extraction attack)에 대한 방어 방안을 수립하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

자율주행 인공지능 모델의 입력을 사람이 제어할 수 있거나 모델 추론 결과를 외부 컴포넌트와 통신하는 등 모델 공격이 가능한 경우, 본 항목을 고려하여 만족 여부를 판단하십시오.

- 모델 추출 공격은 학습된 모델의 다양한 입력에 대한 인지 결과를 분석하고, 분류 기준을 추출하여 적용 중인 학습 모델과 유사한 성능을 지닌 대체 모델을 구성한다. 이후 회피 공격을 위한 적대적 데이터 생성과 추론된 결괏값을 분석해 학습 데이터 내의 개인정보 및 민감정보 등을 복원하여 유출하는 방식으로 2차 공격에 활용할 수 있다.
- 자율주행 차량에 적용 중인 모델을 대상으로 자행되는 추출 공격을 완화 또는 방어하기 위해 인지 결과를 암호화하여 통신하는 방법을 적용할 수 있다.

09-1a 모델 추출 공격에 대비하는 방어 기법을 적용하였는가?

Yes ☐ No ☐ N/A ☐

- 자율주행 알고리즘이 동작하는 아키텍처는 일반적인 모델 추출 공격 사례가 아직 없는 폐쇄 환경으로 확인되었다.
- 하지만 내부 통신(예: 캔CAN, Controller Area Network)을 통해 인공지능 모델의 인지 결과를 특정 컴포넌트로 전달하는 경우, 네트워크상 모든 노드에 브로드캐스팅broadcasting되므로 스니핑sniffing 등에 대한 보안 취약점을 이용한 모델 추출 공격에 노출될 수 있다. 이때, 질의를 통한 공격보다 내부 모델이 인지한 결과를 가로채어 모델을 추출하는 공격을 할 수 있다.
- 이를 완화하려면 인지 결과의 암호화 등 비식별 처리와 같은 적절한 대응 과정을 수행해야 한다.

운행 차량 내 자율주행 알고리즘 모델 추출 공격 방어 기법

방어 기법 분류	방어 기법 내용
인지 결과 암호화	인지 결과의 암호화 등 비식별 처리를 통해 예측 결과가 노출되더라도 모델의 추출을 방해
인지 출력 교란[61]	공격자가 모델을 훔치려면 특정 정보(예: 예측 라벨 및 신뢰도 점수)가 필요하기 때문에 출력 정보를 줄이거나 난독화하여 모델을 방어

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

Yes ☐ No ☐ N/A ☐

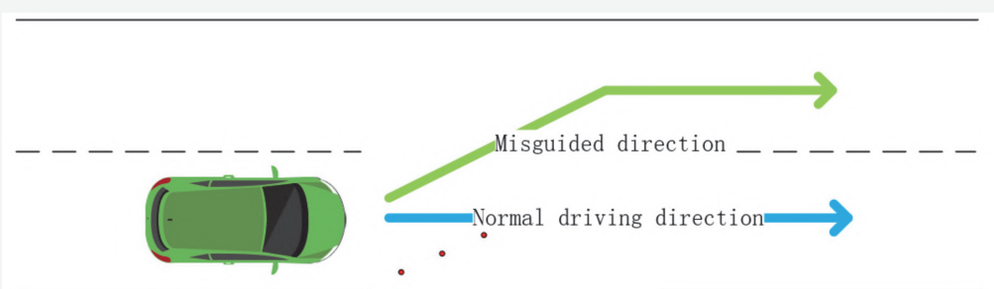
해당여부 판단

자율주행 인공지능 기반 인지 모델을 개발하는 경우, 본 항목을 고려하여 만족 여부를 판단하십시오.

- 모델 회피 공격*은 입력 데이터에 최소한의 변조를 가해 인공지능 모델을 속이는 기법이다. 연구에 따르면 이미지 기반 인공지능 분류 모델은 적대적 공격에 취약한 편이고[62], 이미지에 기반하여 객체 인식 및 분류를 수행하는 자율주행 알고리즘도 모델 회피 공격에 상당히 취약하다.
  * 05-2 참고 자료의 회피 공격 사례 참고
- 모델 회피 공격을 완화하기 위해 HD 맵 정보, V2XVehicle to Everything 정보를 교차 활용하거나, 적대적 공격을 탐지하기 위한 추가 모델을 준비하고 인지 결과를 비교하는 등 여러 방법을 적용할 수 있다.

참고 차선을 가장한 패치를 이용한 공격 사례[63]

- 중국 Tencent가 운영하는 Keen Security Lab은 물리적 환경인 실제 주행 도로에 아래 그림과 같은 패치 3개를 부착하여 다른 차선으로 주행시키는 공격을 시도해 성공하였음
- 아래 그림에서 빨간색 대시는 스티커인데, 차량은 이를 오른쪽 차선이 연속된 것으로 간주하고 교차로 맞은편의 실제 왼쪽 차선을 무시하였음
- 교차로 중앙으로 이동할 때 실제 왼쪽 차선을 오른쪽 차선으로 사용해 역주행하였음



붉은색 패치를 도로에 부착하여 자율주행 차량의 주행 차선 조작 시도

09-2a 모델 회피 공격에 대비하는 방어 기법을 적용하였는가?

Yes ☐ No ☐ N/A ☐

- 자율주행 알고리즘 아키텍처 중 인지를 위한 인공지능 알고리즘을 대상으로 하는 모델 회피 공격 사례가 다수 발생하고 있어 이를 방어하기 위해 적절한 대응 방안을 마련해야 한다.
- 모델 회피 공격에 대한 주요 완화 방법은 V2X 및 HD 맵 등에서 얻은 정보를 함께 활용하거나, 적대적 공격 여부를 판단하기 위한 모델을 추가하는 방법 등이 있다.

모델 회피 공격 방어 기법

방어 기법 분류	내용
V2X 또는 HD 맵 정보 교차 활용[64]	• 인지에 있어 HD 맵의 활용은 알고리즘이나 기업에 따라 달라질 수 있음 • HD 맵에는 디지털화 위치, 도로표지판 위치, 차선, 가로등, 신호 등 탐색에 필수적인 기능이 포함되어 있음 • 이러한 맵의 데이터와 센서(예: 카메라, 라이다)가 인지하는 데이터 간의 불일치를 비교하여 차량을 제어할 수 있는 운전자에게 경고할 수 있음
적대적 공격 여부를 판단하기 위한 모델 추가[65]	• 원래 모델과 별도로 적대적 공격 여부를 판단하기 위한 모델을 추가하고, 두 모델의 추론 결과를 비교해 두 결과 간에 큰 차이가 있으면 적대적 공격으로 탐지하는 방식
학습 기반 모니터링	• 기계학습을 활용하여 모델 공격에 대해 사전 탐지 및 경고 알림, 상응하는 방어 기법을 실행하는 등 능동적으로 방어하는 기법
증류distilation[66]	• 원시 학습 데이터셋 X를 기반으로 초기 심층 신경망을 학습하고 Y에 라벨을 지정하여 확률 벡터 예측 F(X)를 얻음 • 학습 데이터셋 X와 출력 결과 F(X)를 새로운 라벨로 사용하고 유사한 증류 네트워크를 학습하여 새로운 확률 벡터 예측 Fd(X)를 얻음 • 최종적으로 새로운 증류 네트워크를 사용하여 분류하거나 예측함 • 이를 통해 작은 교란에 대한 모델의 민감도가 감소하고 적대적 샘플에 대한 저항이 향상되는 것으로 확인
적대적 훈련 adversarial training[67]	• 예측 과정에서 발생할 수 있는 적대적 샘플을 모방한 적대적 샘플을 학습용 데이터셋에 추가하여 함께 학습함

책임성 투명성

요구사항 10 인공지능 모델 명세 및 추론 결과에 대한 설명 제공

대표 행위자 | 인공지능 모델 개발자 협력 대상 | 시스템 엔지니어 시스템 운영자

- 자율주행 알고리즘 아키텍처에서 모델 추론 결과의 활용 및 설명 대상은 인공지능 모델의 결과를 내부적으로 제공하는 인지 부분, 운전자가 확인할 수 있는 판단·제어 부분으로 나뉜다. 자율주행 알고리즘 처리 과정에서 다양한 인공지능 모델이 사용되고, 이후 주행 경로를 계획·변경하거나, 차량을 제어하는 등 다수의 인공지능 모델이 유기적으로 연계되어 사용될 수 있다. 따라서 모델 정보 및 결과 도출 과정에 대한 설명*, 추론 결과에 대한 설명을 제공하여 인공지능 모델의 예측, 분류, 계획 등 결과에 대해 사용자 신뢰를 확보한다.

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

참고 설명가능성explainability 적용 전 고려해야 할 사항

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

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

Yes ☐ No ☐ N/A ☐

해당여부 판단

자율주행 인공지능 모델의 추론 과정에 대해 사용자 대상 설명이 요구되는 경우, 본 항목을 고려하여 만족 여부를 판단하십시오.

- 사용자가 인공지능 모델의 추론 결과 및 인공지능 시스템의 동작을 신뢰하기 위해서는 시스템 사용자가 인공지능 모델이 제공하는 판단, 혹은 추론 결과의 도출 과정을 이해할 수 있어야 하며, 사용자에게 이에 대한 설명 및 근거를 제시하는 것이 바람직하다.
- 자율주행 분야에서 자율주행 차량의 행동을 설명할 필요성은 여러 요인에 의해 결정된다. 자율주행은 높은 이해관계와 안전이 중요한 애플리케이션이므로 성능 보증을 요구하는 것은 사회적 관점에서 자연스러운 일이나, 현실적으로는 대응하기 어렵다. 모든 상황을 철저히 나열하거나, 시나리오에 따라 완전히 테스트하기는 어렵기 때문이다. 따라서 대체 솔루션으로 자율주행 운전에 대한 설명이 이루어져야 한다.
- 또한, 시스템 성능에 따른 설명은 다양한 이유로 필요하다. 예를 들어, 시스템이 제대로 동작하지 않을 때 엔지니어와 연구원이 예외적인 상황edge case, 잠재적인 실패 모드potential failure mode 등 더 많은 정보를 얻으면 향후 버전을 개선하는 데 도움이 된다. 시스템 성능이 사람과 유사해질 때 사용자의 신뢰를 높이는 데도 설명이 필요하다.
- 자율주행 차량 운전자와 탑승자를 대상으로 인지, 계획, 차량 제어, 현지화, 시스템 관리 항목 등 설명 가능한 부분 및 XAI 등을 활용한 여러 연구와 시도가 이루어지고 있으므로, 이의 검토 및 도입을 고려해 본다.
- 또한, XAI 기술로 아직 설명할 수 없는 부분은 전통적인 의사 결정 트리 기법 또는 현재 연구 중인 원인 학습 기술을 검토하고 도입을 고려해본다.
- 인공지능 모델 추론 결과의 근거를 설명하는 것이 항상 가능한 것은 아니므로 XAI 기술 적용 이외의 대안을 활용하여 인공지능 시스템의 투명성 확보가 필요할 수 있다. 따라서 XAI 기술 적용 가능 여부를 검토한 후, 검토 결과 XAI 기술 적용이 가능하다면 10-1a 를 활용하고 적용이 어려운 경우 10-1b 를 활용할 수 있다.

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

Yes ☐ No ☐ N/A ☐

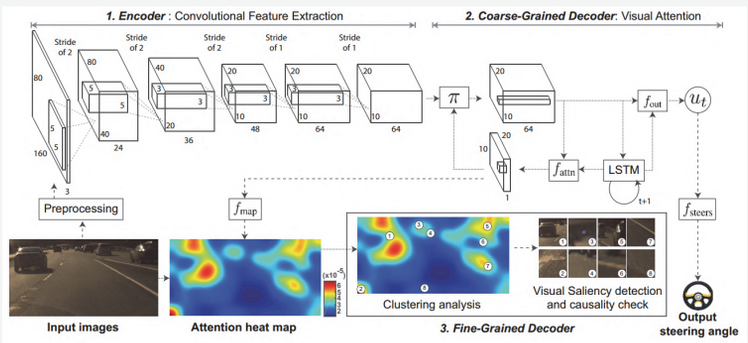
- 자율주행차 운전자와 탑승자에게 자율주행 분야의 인지, 계획, 차량 제어, 현지화[localization], 시스템 관리 항목에 대하여 인공지능 시스템의 출력 결과를 설명할 수 있으며, 각 항목은 설명이 요구되는 목표와 방법에 각각 다르게 접근하여 연구하고 있다[68].
- 현재까지 연구 및 조사 중인 항목별 추론 결과에 대한 설명은 다음과 같다. 이러한 연구는 계속 진행 중이고, 기법 도입 전에 설명 대상 인공지능 모델 및 각 방법의 장단점을 분석해 설명의 목적에 적합한 기법을 선택하는 것이 중요하다.

자율주행 분야 항목별 설명 내용 및 방법

항목	설명 내용		설명을 위한 접근 방법(기법)
인지	사후 설명을 위한 데이터셋 활용	각 데이터셋은 기호, 규칙, 토폴로지가 다른 지역에서 수집하므로, 데이터셋이 자율주행에 충실함을 제공하기 위한 설명이 필요	- 수동 설명[69,70] - 차량 궤적[71] - 경계 상자와의 이상 식별[70,72] - 인간 운전자 행동[73,74]
	자율주행에 대한 비전 기반 설명	인지와 장면 이해의 기본 구조인 신경 네트워크 설명	- 클래스 활성화 맵(CAM[Class Activation Map])[75] - 그라디언트 클래스 활성화 맵 (Grad-CAM)[76] - 가이드 그라드 - CAM[77] - Grad-CAM++[78] - Smooth Grad-CAM ++[79] - VisualBackProp[80] - 계층별 관련성 전파(LRP)[81,82] - DeepLift[83,84] - Guided-Backpropagation[85]
계획	차량에 탑승하고 있는 운전자·탑승자에게 설명을 제공하지 않고 경로를 업데이트할 때 혼란을 야기할 수 있음		- XAI-플랜[86] - 이유 계획[87] - 구체화 기반 계획[88] - 설명 가능성 및 예측 가능성 계획[89] - 모델 조정을 위한 계획 설명[90,91]
차량 제어	자율주행 기능이 차량을 제어할 때, 운전자와 탑승자가 차량에 제어 의사결정 내용을 질의하고, 변경하는 상황이 발생할 수 있음		- 혼합 현실 시각화[92] - 차량 인터페이스에서 유연한 대시보드 패널[93]
현지화	환경의 맵 정보를 기준으로 자율주행차의 위치와 방향을 결정. 시간이 지남에 따라 차량의 위치와 방향에 대한 설명을 적시에 제공하여 사고를 피할 수 있음		- 여백을 초과할 때 정보를 표시하거나 경보를 트리거하는 특수 대시보드 (예: 안전 주차)[94] - 자율주행차 디버깅을 위해 시스템 개발자에게 위치 오류의 지속적인 전송[95]

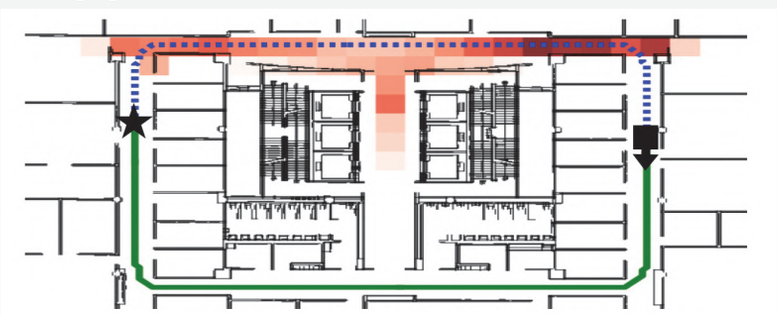
참고 설명을 위한 접근 방법 연구 사례

• 해석 가능한 학습interpretable learning[96]



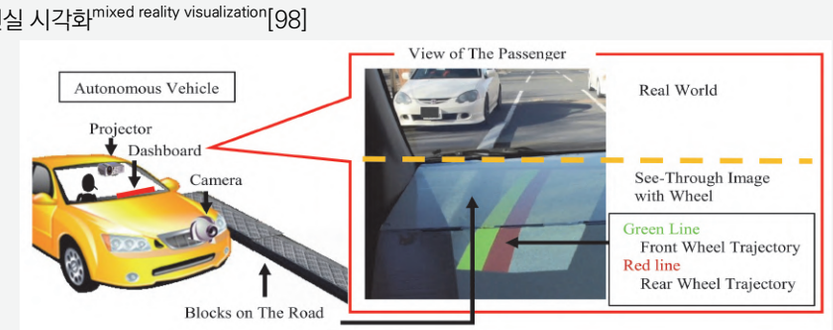
✓ 종단 간 방식으로 입력한 원시 이미지 스트림에서 조향각도 명령을 예측
✓ 모델이 보는 위치와 대상을 시각화할 수 있도록 히트맵 생성
✓ 주의 블록의 각 클러스터를 면밀히 조사하여 인과 관계를 테스트하고 인과·시각적 주의 히트맵 생성

• 이유 계획WHY-PLAN[87]



✓ 차량이 화살표 방향에서 별 표시 쪽으로 이동할 때 더 짧은 거리에도 불구하고, 혼잡한 경로를 피하려고 더 긴 경로를 계획하였음(음영이 어두울수록 예상되는 차량 통행량이 많다는 것을 의미)
✓ 이유 계획에서, "최단 경로를 택할지", "많은 사람을 피할지" 등 두 가지 목표를 대조하여 운전자에게 알림

• 혼합현실 시각화mixed reality visualization[98]



✓ 대시보드에 투명 이미지를 투사하여 승객의 시야에서 벗어난 노면을 시각화함
✓ 혼합 현실 기술을 사용하여 표시된 이미지에 휠 궤적의 그래픽을 오버레이overlay하여 승객이 자동 운전 제어가 올바르게 작동하는지 쉽게 확인할 수 있음
✓ 표시된 이미지를 통해 승객은 도로 상태와 승객의 시야에서 벗어난 예상 차량 경로를 이해할 수 있음

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

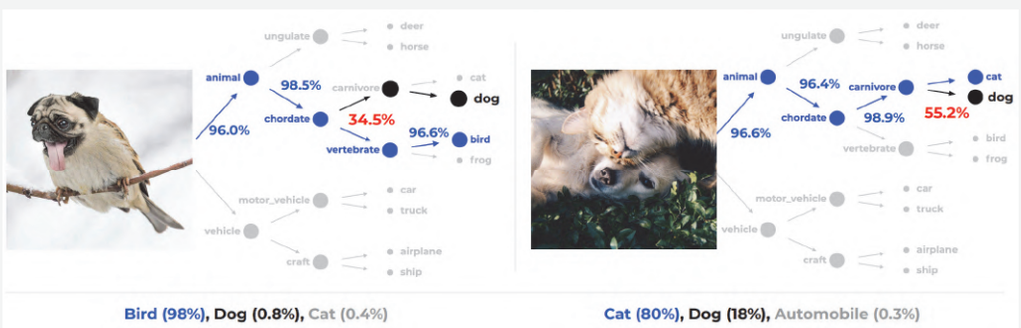
Yes ☐ No ☐ N/A ☐

- 인공지능 모델 추론 결과 및 결정의 근거를 설명하는 것이 항상 가능한 것은 아니다. 이러한 경우, XAI 기술 적용 이외의 대안을 활용하여 인공지능 시스템의 투명성 확보가 필요할 수 있다.
- 인공지능 시스템의 서비스에 따라 XAI 기술 적용이 어려우면, 개발자는 인공지능 시스템의 설명가능성과 신뢰도를 높이기 위해 차선책을 고려할 수 있다. 자율주행 알고리즘 아키텍처에서 판단·제어에 해당하는 알고리즘은 네트워크의 복잡성으로 인해 현재 기술 수준으로는 XAI 기술의 적용이 어려울 수 있다.
- 자율주행 알고리즘 중 인지를 위한 영상 분야 XAI 기술 적용 시 다음과 같은 한계가 있다.
  - ✓ 역전파 기반backpropagation-bases XAI: 이미지의 각 픽셀이 인공지능의 추론 결과에 어느 정도 영향을 미치는지를 경사gradient 값으로 측정
    - 일반적으로 사람이 보기에 해석성이 떨어지며 픽셀들이 주변 픽셀과 어떻게 조합되는지, 그 조합이 얼마나 중요한지를 제시하지 못함
  - ✓ 클래스 활성맵 기반CAM, Class Activation Map XAI: 인공지능의 최상단 레이어의 특징맵을 활용
    - 일반적으로 영상을 다루는 심층신경망은 최상단 레이어로 갈수록 특징맵의 해상도를 줄여나감
    - 히트맵을 시각화하기 위해 큰 폭으로 업스케일을 하는 것이 불가피하고, 모든 채널의 특징맵을 사용하기 때문에 노이즈가 추가될 수밖에 없음
  - ✓ 입력간섭 기반perturbation-based XAI: 입력 샘플링 기반 방법과 입력 최적화 방법으로 나뉨
    - 입력 샘플링 기반 방법에서는 설명을 위해 수천 번의 출력이 필요하며, 랜덤 마스크를 사용하기 때문에 같은 입력 영상이라도 수행할 때마다 결과가 달라짐
    - 입력 최적화 방법에서는 정확한 해를 구했을 때 해석성이 매우 높은 설명을 확보할 수 있으나, 반대의 경우에는 전혀 상관없어 보이는 특징들을 보여주기도 함. 또한 수치 최적화 방식을 사용하기 때문에 연산 시간이 늘어나는 문제도 있음
- 위와 같이 XAI 기술을 활용해 설명이 어려운 경우에는 의사 결정 트리와 같은 전통적인 방법을 도입하는 것도 고려해볼 수 있다[97]. 현재 XAI의 한계, 더 나아가 기계학습의 한계를 보완하기 위해 원인 학습causal learning 기술도 활발하게 연구되고 있어 참고해 볼 수 있다[98].

참고 의사 결정 트리 적용 Neural Network[NBDT, Neural-Backed Decision Trees] 설명 시도[97]

- 개요
  - ✓ 네트워크의 마지막 선형[linear] 레이어를 의사 결정 트리로 대체
  - ✓ 전통적인 의사 결정 트리와 달리 추론 시에는 path probabilities를 사용하여 불확실한 중간 결정을 줄임
  - ✓ 이를 통해 학습된 모델의 가중치로부터 계층을 만들어 과적합을 피함
  - ✓ 학습은 계층적 손실을 사용하여 높은 레벨의 결정을 학습할 수 있게 함

- 학습 완료 모델 성능
  - ✓ Small-scale 데이터셋에서 기존 모델보다 1% 가량 성능이 향상되었으며, 설명가능한 특성도 보존
  - ✓ Large-scale 데이터셋에서 동일 backbone을 가진 SOTA[State-Of-The-Art] 모델과 비슷하거나 더 좋은 성능을 냈음



모델 성능을 해치는 모호한 데이터 예시

- 자체적인 설명가능성 평가
  - ✓ Saliency 설명과 NBDT의 설명 비교 시, 사람은 NBDT의 설명에서 더 정확하게 오분류를 찾을 수 있었음
  - ✓ NBDT의 엔트로피를 약간 수정하여 모호한 라벨을 탐지하였음
  - ✓ 이미지 분류 문제에서 사람들은 NBDT의 예측을 더 선호하였음

- 한계 및 시사점
  - ✓ 설명력이 부족할 뻔한 부분은 WordNet을 사용하여 보완
  - ✓ WordNet에 없는 계층은 설명하지 못함
  - ✓ 아직 설명가능성 부분에서 정량적인 평가가 어려우나, 자신들의 모델이 왜 더 설명력이 좋고, 신뢰성이 높은지 수치화하였음

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

Yes No N/A
☐ ☐ ☐

해당여부 판단

자율주행 인공지능 모델의 명세 정보를 투명하게 제공하고자 하는 경우, 본 항목을 고려하여 만족 여부를 판단하십시오.

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

10-2a 시스템 개발 과정과 모델 작동 방식에 대한 세부 정보가 설명된 문서를 작성하였는가?

Yes No N/A
☐ ☐ ☐

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

이해관계자에 따른 모델 상세 문서 예시

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

참고 IBM의 객체 탐지기 팩트 시트Object Detector FactSheet 사례

• 개요

✓ 이 문서는 IBM 개발자 기계학습 eXchange의 객체 탐지기 모델과 함께 제공되는 FactSheet이다. FactSheet는 공급 업체의 적합성 선언을 통해 AI 서비스에 대한 신뢰를 높이는 것을 목표로 하며, 이 FactSheet는 객체 탐지기 모델을 교육하는 과정과 예상 결과 및 적절한 사용을 문서화하였다.

• 목적

✓ 경계 상자를 사용하여 이미지 내에서 여러 개체를 감지한다. 이 모델은 COCO 데이터셋에서 서로 다른 80개 개체 클래스를 인식하도록 학습된다. 이 모델은 이미지 피처 추출을 위한 심층 컨볼루션 네트 기본 모델과 COCO 데이터셋에서 학습된 객체 감지 작업에 특화된 추가 컨볼루션 레이어로 구성된다. 이는 TensorFlow 프레임워크를 사용하여 SSD MobileNetV1을 기반으로 한다.

– 중 략 –

• 견고성

✓ 이미지 변환에 대한 견고성

– AI 및 ML 모델은 자연적으로 발생하는 노이즈에도 불구하고 정상적으로 수행되어야 하며, 여기서 출력은 객체 라벨과 경계 상자 예측 모두에서 일관되게 유지되어야 한다. 이 테스트는 심각도 증가(노이즈 증가)에 다양한 일반적인 이미지 변환을 적용하고 광범위한 이미지 변환에 대한 모델의 안정성을 측정한다.

– 더 구체적으로, 모델 예측의 안정성은 세 가지 메트릭으로 표현된다.

1. 경계 상자 예측의 수가 변경되는가?(검출 안정성)
2. 고유한 객체 라벨 집합이 변경되는가?(안정성 설정)
3. 경계 상자의 위치와 크기가 변경되는가?(바운딩 박스 안정성)

• 세부 정보

– 이 테스트에서는 이미지 손상 벤치마크(https://github.com/hendrycks/robustness)의 이미지 손상을 사용한다. 2017 MS-COCO 평가 데이터셋(http://cocodataset.org/)에서 무작위로 선택된 클래스 별 평가 샘플 집합이 주어지면 다음과 같은 손상이 적용된다.

• 평가 메트릭

– 손상된 이미지 집합의 각 이미지에 대해 다음과 같은 통계를 얻는다.

1. 감지 안정성: 원본 소스 이미지와 동일한 출력 예측 수를 가진 손상된 이미지의 수를 총 평가 이미지 수(N)로 나눈 값으로 측정한다.
2. 세트 안정성: 손상된 이미지의 객체 라벨 세트와 원본 소스 이미지의 라벨 세트 사이의 IoUIntersection over Union로 측정된다.
3. 경계 상자 안정성: 손상된 이미지의 경계 상자 영역과 원본 이미지 사이의 IoU로 측정된다.

– 후 략 –

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

Yes No N/A
☐ ☐ ☐

해당여부 판단

자율주행 인공지능 모델의 추론 결과에 대해 사용자 대상 설명이 요구되는 경우, 본 항목을 고려하여 만족 여부를 판단하십시오.

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

10-3a 모델 추론 결과에 대한 설명이 필요한지 검토하였는가?

Yes No N/A
☐ ☐ ☐

- 자율주행 시스템은 제동 여부, 우회전 또는 좌회전, 차선 변경 여부뿐 아니라 미래에는 추월 여부와 같은 특정 상황에서 수많은 결정을 내린다. 이를 위해 센서, 인공지능, 다른 SW 모듈을 포함한 다양한 구성 요소가 유기적으로 실시간 동작한다[99]. 이러한 환경에서 모델의 추론 결과와 이에 대한 설명을 운전자에게 지속적으로 제공하는 것은 불필요하거나, 너무 많은 정보를 제공하는 것일 수 있다.
- 하지만 자율주행 시스템이 시시각각으로 달라지는 외부 환경에 반응하여 의사결정을 하는 과정에서 특정 모델, 특정 시점에 대한 추론 정보는 운전자 또는 탑승자가 차량의 동작 상태를 쉽게 이해하는 데 도움이 된다[100]. 다음은 모델의 추론 결과가 운전자의 이해를 도울 수 있는 예시이다.
  - ✓ 급제동 전, 자율주행 시스템이 탐지한 객체의 종류 및 확률 안내
  - ✓ 차량의 주행 경로 변경에 가장 큰 영향을 준, 인지 또는 판단(예: 전방 사고, 공사) 추론 결과
- 인공지능 시스템이 도출한 결과에 대한 설명을 제공하는 것은, 사람들이 인공지능을 활용하여 의사 결정하는 데 도움이 될 수 있지만, 오히려 방해될 수도 있다. 따라서 모든 경우에 모델의 추론 결과에 대한 설명을 제공하기 보다는, 설명이 꼭 제공되어야 하는지를 확인하는 과정이 선행되어야 한다.
- 모델의 추론 결과에 대한 설명을 제공하지 않는 편이 더 나은 경우에 대한 두 가지 예시는 다음과 같다.
  - ✓ 첫째, 모델의 추론 결과에 대한 설명 제공 자체가 사용자의 의사결정에 크게 영향을 미치지 않을 것으로 판단되는 경우다. 설명 제공으로 인해 미치는 영향을 명확하게 분석하지 않은 경우, 자세한 설명을 제공하면 사용자의 의사결정에 더 도움이 될 것으로 생각할 수 있지만, 예상과는 다르게 혼란을 초래할 수 있다. 예를 들어, 인공지능 시스템이 도출한 두 가지 결과가 있고, 각각의 예측 확률이 85.8%, 87.0%라면, 사용자는 어떤 결과를 활용하여 의사결정을 할지 혼란스러울 수 있다.
  - ✓ 둘째, 예측 확률이 너무 높거나 낮은 경우에도 모델의 추론 결과에 대한 자세한 설명을 제공하지 않는 것이 낫다. 만약 시스템의 출력 결과에 대해 신뢰도가 100%라고 사용자에게 알릴 경우, 사용자가 시스템의 출력 결과를 맹목적으로 수용하게 만들 수 있다.

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

Yes No N/A
☐ ☐ ☐

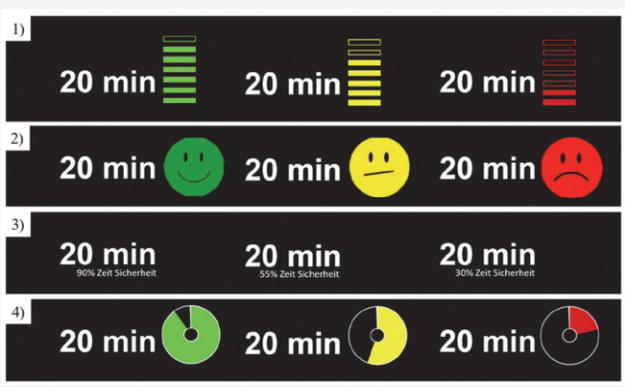
- 각 모델의 추론 결과가 참값과 일치할 확률을 계산할 수 있다면, 이를 모델의 최종 의사결정에 대한 설명으로 사용할 수 있다. 확률 변수의 분산크기로, 인공지능 모델이 도출한 결과를 얼마나 확신하는지를 나타내는 불확실성을 모델 추론 결과에 대한 설명으로 고려해볼 수 있다.
- 다음은 자율주행 분야에서 예측 결과에 대한 확률과 불확실성에 따른 모델 추론 결과에 대한 설명 예시이다[100].

예측 확률(0~1)	불확실성(0~1)	설명 예시
0.98	0.01	98% 정확도 및 1% 불확실성으로 거의 확실하게 전방 10m에 있는 고양이를 탐지하여 급제동한다.
0.98	0.90	98% 정확도로 전방 10m에 있는 고양이를 탐지하여 급제동이 필요하나, 불확실성이 90%이므로 운전자 확인이 필요하다.

- 예측 확률이 임계치보다 낮은 경우 또는 불확실성이 높은 경우에는 사용자가 이를 인지할 수 있도록 모델 추론 결과 대한 설명을 반드시 제공하여야 한다. 임계치를 도출하기 위해서는 인공지능 모델로 인해 발생 가능한 문제 상황을 정의하고, 문제 발생 여부를 결정 짓는 중요 변수를 파악해야 한다. 여기서 문제 상황이란 사용자의 생명이나 재산과 관련된 위협적인 상황뿐 아니라 기대하는 또는 유지되어야 하는 품질 수준보다 낮은 상황 등을 포함한다.
- 다만, 불확실성이 매우 낮은 경우에는 모델의 과적합으로 인한 결과임을 의심해볼 수 있다. 따라서, 불확실성이 매우 낮다면 모델의 과적합 여부에 대한 추가 검증이 필요할 수 있다.

참고 자율주행 기능의 신뢰 정보 제시 사례

- 부분 자율주행 차량에서 운전자에게 다음 운전 개입 요청까지 남은 예측 시간에 대한 표시 및 신뢰도 정보 제공 요구[101]
  - ✓ 운전자가 비운전 관련 활동NDRA, Non-Driving Related Activities을 할 때, 사용자 요구로 식별되었음
  - ✓ 이러한 추정은 계속 달라지는 교통 상황과 업데이트된 지도 기반 정보로 인해 오류가 발생하기 쉬움



다음 운전 인수 요청까지의 예측 시간 및 예측 결과의 신뢰도 제공 HMI 예시 (왼쪽부터 90%, 55%, 30% 신뢰도)

04 시스템 구현

다양성 존중

요구사항 11 인공지능 시스템 구현 시 발생 가능한 편향 제거

대표 행위자 | 시스템 엔지니어 협력 대상 | 시스템 운영자 인공지능 모델 개발자

- 자율주행 시스템 개발 시, 사람의 운전을 차량이 대신하는 과정에서 차량 운전자의 주행 패턴과 성향 등 사용자별 배경지식이나 편견으로 인해 인공지능 시스템이 편향될 수 있다. 또한, 운전자의 적극적인 개입이 필요한 단계에서 자율주행 차량 운행 시 운전자의 배경지식이나 성향 등에 따라 자율주행시스템의 운전 인수 요청을 인지할 때 자동화 편향이 발생할 수 있다. 따라서 자율주행 인공지능 시스템에서 발생 가능한 편향을 식별하여 이를 제거 또는 완화한다.

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

Yes ☐ No ☐ N/A ☐

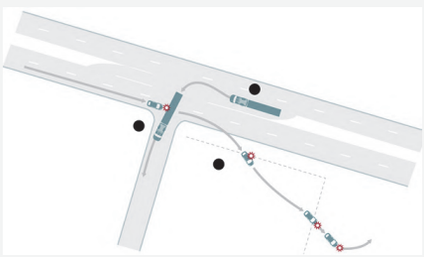
해당여부 판단

자율주행 인공지능 시스템의 출력을 표현하기 위해 HMI를 적용하는 경우, 본 항목을 고려하여 만족 여부를 판단하십시오.

- 자율주행 시스템의 특성상 자율적인 주행을 위한 운전 행동 판단·제어 인공지능 알고리즘을 포함하는 시스템을 개발할 때 개발자의 사전 지식, 규칙, 경험이 반영되어 시스템 내 소스 코드에서 규칙, 사전 지식 등의 편향이 발생하여 사고로 이어질 수 있으므로 주의해야 한다.
- 자동화 운행 정보를 인지하는 정도가 사용자(운전자)마다 달라 자율주행 시스템에서 인수가 지연되고 사고로 이어지므로, 시스템 설계 시 다수의 사용자를 대상으로 사용자 인터페이스로 인한 자동화 정보의 편향이 발생하지는 않는지 검토해야 한다.

참고 자율주행(또는 첨단 운전자 보조 시스템 탑재) 차량 내 자동화 편향으로 인한 사고 사례

1. 미국에서 발생한 자율주행차 첫 번째 사망 사례(2016년)

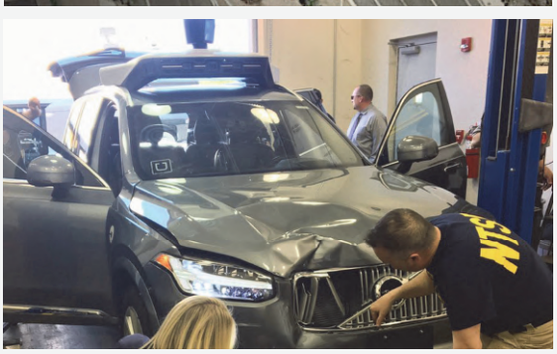


- NTSB[National Transportation Safety Board] 분석 결과
  - ✓ 차량 자동화에 대한 과도한 의존으로 인한 운전자 부주의가 결합하여 큰 사고로 이어짐
  - ✓ 결과적으로 운전자는 전방 트럭에 대한 대응이 부족하였음
  - ✓ 제조업체의 지침 및 경고에 따라 자동화 시스템을 사용하지 않고, 장기간 운전에서 이탈하였음
- Tesla 사의 대응
  - ✓ 아래와 같은 내용의 차량 전체 시스템을 업데이트(Autopilot version 8)
  - ✓ 운전자가 스티어링 휠에서 손을 떼고 주행하는 경우 안내하는 경고 메시지 출력 시간을 줄임
  - ✓ '3 strikes rule'을 적용하여 경고 3회를 받았다면 주행 종료 시까지 자동 조종 기능을 비활성화함

2. 트럭과 자율주행 셔틀 차량 사이의 저속 충돌(2017년)

- NTSB 분석 결과
  - ✓ 트럭 운전사가 골목으로 후진하는 중 셔틀과 저속으로 충돌
  - ✓ 셔틀 차량 내 운전석에 승무원이 앉지 않았음
  - ✓ 승무원이 차량을 멈추긴 했지만, 여전히 적시에 멈추지 못했고, 트럭과 충돌할 것을 깨달았을 때조차도 개입하지 않았음(승무원의 인식과 경계심이 사고를 예방하기에 충분하지 않음)
- Keolis 사의 대응
  - ✓ 승무원이 탑승하는 동안은 반드시 운전석에 앉도록 정책을 강제화함

3. 자율주행으로 운전하는 Uber 테스트 차량과 보행자 충돌(2018년)



- NTSB 분석 결과
  - ✓ 차량 운전자가 자율주행 상황을 모니터링하지 못하였음
  - ✓ 다음과 같은 원인으로 충돌이 발생했다고 분석
    - (1) Uber사의 부적절한 안전 위험 평가 절차
    - (2) 차량 운전자에 대한 비효율적인 감독
    - (3) 운전자의 자동화 의존을 해결하기 위한 적절한 메커니즘 부족, 안전 부주의 문화 등
- Uber사의 대응
  - ✓ 8개월 동안 테스트 중단
  - ✓ 자율주행 차량 테스트의 설계 및 구현을 설명하는 70페이지 분량의 안전 보고서 제출
  - ✓ Fail Safe로 사용할 수 있는 모든 테스트 차량에 부운전자 추가

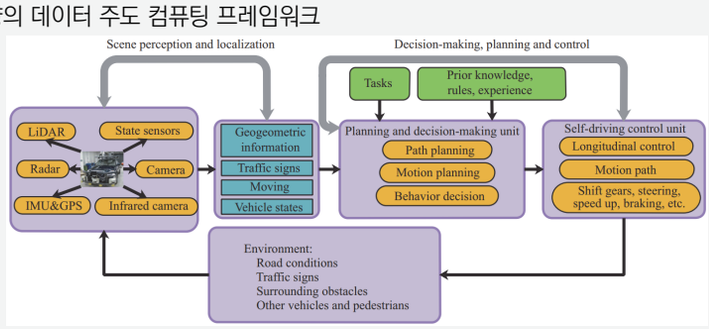
11-1a 데이터 접근 방식 구현과정 등 소스 코드에서의 편향 발생 가능성을 확인하였는가?

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템은 모델에서 활용할 데이터에 접근하는 방식이 코드상에 구현되는 과정에서 특정 클래스 접근이 누락 되는 등 다양한 형태의 편향이 발생할 수 있다.
- 자율주행 알고리즘 아키텍처 중 인지를 위한 인공지능 모델을 활용하는 경우, 자율주행 시스템 초반에 구축되어 이후 기능을 위한 기본 정보로 사용되므로 해당되지 않을 수 있다.
- 그러나 판단·제어를 위한 인공지능 모델 및 시스템을 구축하여 적용할 경우에는 경로 계획, 제어 계획, 행동 결정 등을 위해 미리 정의된 작업과 사전 지식·규칙·경험을 반영하여 의사결정을 하므로 인지 편향 또는 확인 편향이 잠재적으로 이어질 수 있다. 따라서 시스템의 편향 발생을 줄이기 위해서는 배경 지식과 경험이 다양한 전문가를 선정하는 것이 도움이 된다.

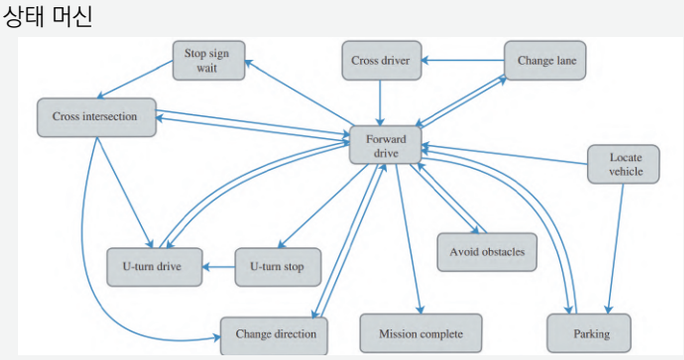
참고 자율주행 차량의 데이터 주도 컴퓨팅 프레임워크 및 의사결정 유한 상태 머신 예시[102]

- 자율주행 차량의 데이터 주도 컴퓨팅 프레임워크



✓ 센서에서 얻은 장면 인식 및 GPS 데이터 등은 위치 파악 등 현지화에 사용되며 계획 및 의사결정 단계를 거쳐, 자율주행 제어 장치에서 수신한 제어 신호를 사용하여 자율주행 차량을 제어함
✓ 도로 상태와 같은 환경은 추가 처리를 위해 인식 모듈에 피드백됨
✓ 계획 및 의사결정 유닛에서 경로 계획, 모션 계획, 행동 결정 등의 기능을 위해 작업task 및 사전 지식prior knowledge, 규칙rules, 경험experience이 필요하며, 이는 인지 편향을 유발할 수 있음

- 의사결정 유한 상태 머신



✓ 의사결정 유한 상태 머신은 여러 운전 행동 상태로 구성됨. 다른 입력 정보를 사용하여 현재 및 다음 동작 상태의 전환 관계를 정의함
✓ 각 운전 행동 상태에서 시스템 구동을 위해 필요한 정보의 누락, 혼선으로 인해 편향이 발생할 수 있음

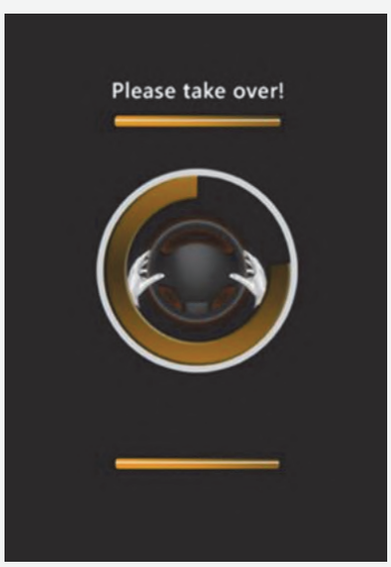
11-1b 사용자 인터페이스 및 상호작용 방식으로 인한 편향을 확인하였는가?

Yes ☐ No ☐ N/A ☐

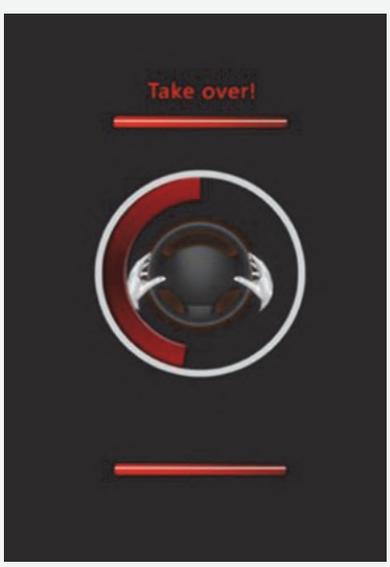
- 자율주행 시스템을 대상으로 기존에 발생한 사고를 분석한 결과 사용자 인터페이스에서 자동화 편향이 상당부분 발생한다고 조사되었다. 특히 자율주행 시스템의 운전 권한을 사용자(운전자)에게 인수 take-over할 때 자동화 편향이 발생하였다. 이러한 자동화 편향은 운전자의 주의가 부족하거나, 운전자에게 인수해야 할 상황임을 충분히 알리지 못해 발생하였다[59].
- 자동화 편향은 사용자(운전자)에 따라 자율주행 시스템의 자동화 수준(자율주행 레벨 1~2) 및 그에 따른 운전 권한을 인수하는 제공 정보의 이해 정도가 달라 발생하였다.
- 사용자 인터페이스 설계 및 편향 발생 가능성이 있는 자동화 인터페이스 요소에 대한 인식이 사용자(운전자)마다 다를 수 있음을 확인하고 미리 검토, 보완해야 한다.

참고 운전 권한을 사용자에게 인수해야 하는 상황임을 알리는 사용자 인터페이스 예시[103]

- 자율주행 시스템에서 운전자에게 운전 권한을 인수해야 하는 상황 등을 운전자에게 알리고자 할 때, 인수 긴급도에 따라 차별화된 인터페이스가 필요하다.



낮은 인수 긴급도 HMI 예시



높은 인수 긴급도 HMI 예시

안전성 책임성 투명성

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

대표 행위자 | 시스템 엔지니어 협력 대상 | 시스템 운영자 인공지능 모델 개발자 HMI 전문가

- 자율주행 시스템에서는 인공지능 모델의 동작에 영향을 주는 센서 고장, 내부 처리 성능 저하 등 여러 가지 문제가 빈번하게 발생할 수 있다. 이와 더불어, 운전자의 적극적인 개입이 요구되는 상황에서 운전자의 전방 주시 또는 조작 미흡 등이 발생하기도 한다. 따라서 자율주행 인공지능 시스템에 안전 모드를 구현하고, 문제 발생 알림 절차를 수립하여 발생 가능한 문제 상황에 대비한다.

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

Yes ☐ No ☐ N/A ☐

해당여부 판단

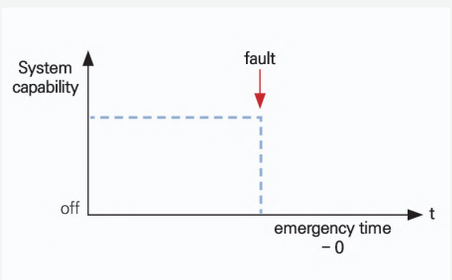
자율주행 인공지능 시스템의 문제 발생에 대한 대응이 필요한 경우, 본 항목을 고려하여 만족 여부를 판단하십시오.

- 고장 안전fail-safe은 고장이나 오류로 문제가 발생하더라도 안전한 상태를 유지하는 메커니즘 및 기능을 의미하며, 산업 전반에서 사용되는 일반적인 개념이다. 고장 안전은 자율주행 분야에도 적용될 수 있다. 특히 자율주행 시스템은 차량 고장 발생 시 주로 운행을 종료하여 안전상태로 전환하는 ISO 26262 보다는 고장이 발생하더라도 정상적으로 동작하며 안전하게 기능을 유지하는 fail operational 아키텍처가 요구된다.
- 자율주행 시스템의 입력 센서가 일부 또는 모든 고장으로 자율주행 기능을 수행할 수 없는 등의 문제가 발생하면 사용자에게 알림과 동시에 안전하게 조치해야 한다. 이때 적용되는 fail operational 아키텍처는 자율주행 레벨에 따라 전반적인 전략이 결정되어야 한다. 예를 들어, 자율주행 레벨 3 이하에서는 시스템 요청 시 차량 제어권을 운전자가 받을 수 있는 반면, 레벨 4 이상에서는 운전자의 개입 없이 시스템 스스로 Dynamic Driving TaskDDT 피드백을 받을 수 있도록 설계되어야 한다[104].
- 자율주행 시스템은 내부적으로 발생하는 문제뿐만 아니라, 외부 위협이 있을 때도 문제가 발생할 수 있다(예: 센서 재밍jamming, 위성항법시스템GNSS, Global Navigation Satellite System, 스푸핑spoofing). 따라서 이러한 문제를 방지하기 위해서는 보안 메커니즘을 적용해야 한다.

참고 자율주행 시스템에서의 고장 안전 모드[105,106]

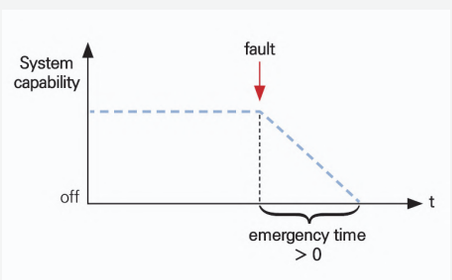
(1) Fail Passive

자율주행 시스템에 고장이 발생하면 즉시 시스템 전체가 정지되는 모드이다. 주행 중 시스템이 정지되면서 운전자는 차량을 제어할 수 없게 되고, 이때 예상치 못한 위험이 발생할 수 있다는 한계가 존재한다.



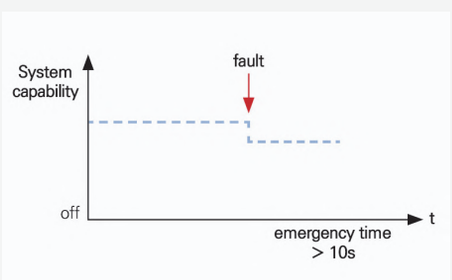
(2) Fail Active

자율주행 시스템에 고장이 발생하면 경보를 울리고, 짧은 시간 동안 차량 운행이 유지되는 모드이다. 이때, 시스템은 미리 정의된 대로 동작하며, 운전자는 제한된 기능 안에서 차량을 제어할 수 있다.



(3) Fail Operational

자율주행 시스템 일부에 고장이 발생하더라도 자율주행 기능이 유지되는 모드이다. 주행 중 특정 기능이 고장 났을 때 비활성화하거나 리셋하면 위험을 초래할 수 있으므로, 안전하게 기능을 유지할 수 있는 Fail operational 아키텍처는 자율주행 시스템에서 중요한 설계 요소이다. 리던던시[redundancy]를 충분히 확보해 특정 기능이 고장 나더라도 그 기능을 대체하는 다른 기능이 활성화된다.



12-1a 문제 상황에 대한 예외 처리 정책이 마련되어 있는가?

Yes ☐ No ☐ N/A ☐

- 인지 시스템이 포함되는 자율주행 분야의 경우, 데이터 입력을 담당하는 센서가 고장나면 핵심 기능이 동작할 수 없으므로 이에 대한 처리 메커니즘이 필요하다. 국토교통부에서 발간한《레벨4 자율주행 자동차 제작안전 가이드라인》9장에서 이러한 비상 대응 예시에 대해 언급하였다.
- 각 센서의 일부가 고장 나거나, 안개·폭우 등 기상 상황으로 인해 자율주행 기능을 수행할 수 없을 때는 즉시 사용자(예: 운전자, 보행자)에게 내부 또는 외부 알림 수단을 통해 알리고, 사용자에게 안전과 관련된 정보를 제공하여 조치할 수 있도록 해야 한다.

참고 문제 상황에 대한 알림 예시(외부 환경)

- 외부 환경으로 인한 Tesla 사의 완전자율주행FSD, Full Self-Driving 기능의 해제 및 재개[107]

눈길에서 차선 소실 시 FSD 해제

차선이 보일 때 FSD 재개

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

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템을 개발할 때 격리 및 탐지 등 보안 기법을 활용하여 인공지능 보안 아키텍처를 적용함으로써 인공지능 데이터 및 모델에 대한 보안성뿐만 아니라 인공지능 시스템의 전반적인 보안성을 확보할 수 있다.
- 자율주행 차량에는 수많은 센서가 주행 기능을 자율적으로 수행하도록 완벽하게 장착되어 있다. 따라서 공격자는 센서 재밍을 통해 원치 않는 신호를 통신 채널에 주입하고 관련 인공지능 시스템과 센서의 연결을 차단·방해할 수 있다. 특히 위성항법시스템에 스푸핑 공격이 실행되면 자율주행 알고리즘 아키텍처 중 판단·제어 모델에 의사결정 프로세스와 사용자(운전자, 승객) 안전을 포함한 차량 관련 기능에 잠재적 영향을 미치는 악의적인 거짓 데이터가 제공될 수 있다. 공격자는 이러한 유형의 공격(재밍 및 스푸핑)을 실행해 자율주행 차량 또는 시스템 제어에 영향을 미칠 수 있다[108].
  - ✓ 예를 들면 실제와 다른 상황으로 인식하거나, 잘못된 충돌 경고 유발, 차량의 잘못된 위치·포지셔닝을 선택하여 안전 관련 문제를 일으킬 수 있다.
- 자율주행을 위한 기술 및 기반 시설infrastructure이 발달함에 따라 타 차량(예: 커넥티드카), 도로 인프라(예: 첨단 도로 인프라[109]) 등과의 V2X 통신을 통한 주행 경로 설정 기능 등이 활용될 수 있다. 이때, 통신을 활용한 공격 등에 노출될 수 있으므로 보안에 유의해야 한다.

참고 Uconnect 기능을 사용하는 Jeep 차량의 무선 해킹 사례[7]

- 2015년 발생한 지프 체로키 해킹 사건
  - ✓ 두 해커가 지프 체로키의 에어컨, 스테레오, 엔진을 원격 조종한 사건
  - ✓ 두 해커는 동영상에서 실제로 변속장치를 조종하거나 자동차가 속도를 줄이면 브레이크를 조작하는 모습을 시연하였음. 두 해커가 사용한 도구에는 저속에서 엔진을 완전히 죽이거나, 갑자기 브레이크를 작동하거나, 완전히 비활성화하는 기능이 포함되어 있음
  - ✓ 당시, 차량의 IP 주소를 아는 사람은 Uconnect의 셀룰러 연결을 통해 누구나 전국 어디에서 액세스할 수 있었고, 이는 공격자의 관점에서 볼 때 매우 좋은 취약점이라고 밝힘

2015년 해킹 시험에 참여한 해커 소개

원격으로 스티어링 휠 조작 시연

참고 자율주행 차량의 보안 공격 위험성[108]

- 공격에 대한 대책 수립의 필요성: 자율주행 차량의 센서가 재밍 공격을 받으면, 센서가 활용되는 기능(예: 장애물 감지, 차선 이탈)에 부정적 영향을 줄 수 있다. 또한, 위성항법시스템 신호가 스푸핑 공격을 받아 차량이 잘못된 데이터를 수신하기 시작하면 위치 기반 시스템이 영향을 받을 수 있다. 예를 들면, 공격자가 자율주행 차량의 제어권을 인수할 수 있으며, 잘못된 충돌 경고를 유발하거나 차량의 잘못된 위치 및 포지셔닝을 설정하게 되어 안전 관련 문제를 일으킬 수 있다.
- 대책: 간섭 완화 기술의 적용, 스푸핑·재밍 감지 기능을 갖춘 수신기 적용, 보안 통제 및 패치 취약점을 정기적으로 평가, 강력한 사용자 인증 기법, 차량에 침입 감지 시스템IDS, Intrusion Detection System 배치, 재해 복구 계획 수립, 감사 로그 유지 관리 등

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

Yes ☐ No ☐ N/A ☐

- 자율주행 시스템에서 완전 자율주행단계로 보는 레벨 4에서 몹시 위험한 상황 일부, 레벨 3 이하에서 차량이 제어할 수 없는 상황일 때 운전자에게 알리고 운전의 제어를 요청해야 하는데, 이를 '제어권 전환 요청TOR, Take Over Request'이라고 한다. 현재 레벨 3 이하의 자율주행 인공지능 시스템에서 최종 책임은 운전자에게 있으므로 반드시 사용자가 개입하여 운전하도록 규정하고 있다.
- 자율주행 알고리즘 중 인지 모델 추론 결과의 신뢰도는 안전과 직결되는 중요한 문제이다. 모델의 인지 결과는 이후 판단·제어 알고리즘의 입력 정보로 활용되므로 시스템이 자율주행 모드라면, 인지 결과의 낮은 신뢰도에 사용자가 대비하도록 미리 조치해야 한다.
- 다음은 인지 모델 추론 결과의 신뢰도가 낮거나 인지 모델의 성능이 낮을 때 시스템에서 고려해야 하는 항목이다.
  - ✓ 인지 모델 추론 결과의 신뢰도가 일정 수준 이하이면, 사용자에게 TOR 주의 알림 또는 긴급한 TOR 알림
  - ✓ 학습을 완료한 인지 모델의 성능이 일정 수준 또는 그 이하이면, 모델의 배포 금지
  - ✓ 차량 내에서 구동 중인 인지 모델의 처리 성능 또는 통신 문제 등으로 응답 시간 등이 일정 수준 이하이면, 사용자에게 TOR 주의 알림 또는 긴급한 TOR 알림
- 국제 기준에서는 운전자가 TOR을 인지하고 수동 운전으로 전환하기까지 '4초' 정도 소요된다고 본다. 그러나 운전자가 평소 운전 실력으로 완전하게 복귀하려면 평균 14.25초가 더 걸린다는 연구 결과가 있어, TOR 인지에서 온전하게 제어권이 전환되기까지는 약 19초 정도가 소요되므로, 이를 고려하여 사용자가 개입하도록 안내해야 한다[110].
- 국토교통부에서 발간한 《자율주행자동차 윤리 가이드라인》에서는 필요시 운전자 또는 탑승자의 판단에 의해 제어 또는 정지될 수 있는 기능을 갖추고 있어야 한다고 언급하고 있다.

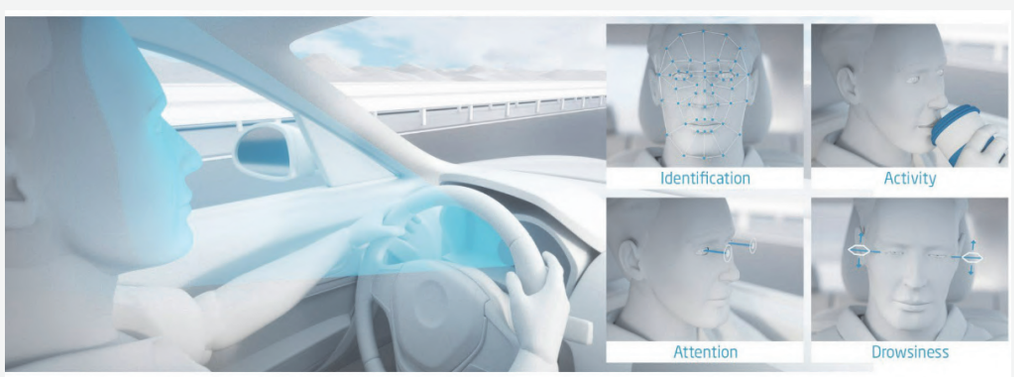
12-1d 예상되는 사용자 오류에 대한 안내 및 대응을 제공하는가?

Yes ☐ No ☐ N/A ☐

- 자율주행 알고리즘을 위한 인공지능 기반 시스템에서 사용자가 직접 값을 입력하는 등의 조작은 없으므로 현 시점에서는 조작 오류에 대한 안내 및 대응은 고려대상이 아닐 수 있다.
- 자율주행 시스템에서 레벨 3 이하는 전적으로 운전자에게 책임이 있으므로 운전자는 반드시 전방을 주시해야 할 의무가 있다. 따라서, 자율주행 모드가 동작 중이라 하더라도 운전자는 전방을 주시해야 하고, 시스템은 사용자가 운행 상태에 집중하지 못하는 상황 등을 감지하여 안내 및 대응해야 한다[37].
  - ✓ 안내 및 대응 예시: 운전자 전방 부주의 경고, 운전자 스티어링휠 핸드오프 경고 기능 등
- 12-1c의 TOR 알림 시, 운전자의 즉각적인 개입이 필요할 때 운전자의 대처가 지연되는 일이 발생할 수 있으므로 13-2a의 설계 권장 사항 18번처럼 다양한 방법(예: 시각, 촉각, 청각)을 복합적으로 활용하여 사전에 대응할 수 있다.

참고 사용자 오류에 대한 대응 방안 예시[111]

- 자율주행 시스템에서 예상되는 사용자 오류에 대한 대응 방안으로는 운전자 모니터링 시스템DMS, Driver Monitoring System이 있다. 운전자의 시선을 분석하여 주의 산만, 졸음 등을 모니터링하는 접근 방식이 이에 해당한다.



12-1e 객체 및 주행상황 인지 오류를 방지하기 위해 다중 센서 기술을 적용하였는가?

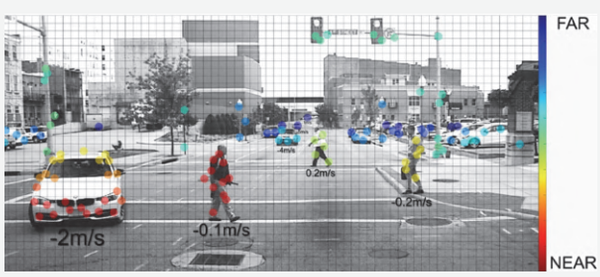
Yes ☐ No ☐ N/A ☐

- 객체 및 주행 상황을 인지할 때 어느 한 센서의 정보만 활용한다면 오인식 가능성이 커 안전상의 문제로 이어질 수 있다. 이를 방지하기 위해 동일한 유형의 센서를 중복하여 활용하거나, 다른 유형의 센서에서 얻은 정보를 규합하여 활용한다면 인식 결과의 상호 보완이 가능하므로, 다중 센서 기술을 적용하여 관련 위험을 완화할 수 있다.
- ISO 21448에서는, 인식 시스템에 영향을 줄 수 있는 예시를 다음과 같이 소개하고, 이와 같은 위험을 완화하기 위해 다중 센서 기술을 활용할 수 있다고 언급하고 있다.
  - ✓ 비전 시스템(카메라) 오류 요인: 차량 정면에서의 일출, 태양으로 인한 주변 빛의 반사, 터널 출입구에서의 갑작스러운 조도 변화 등
  - ✓ 레이더 시스템 오류 요인: 비 또는 눈에 의한 간섭, 금속 교량 주행 시 간섭 등

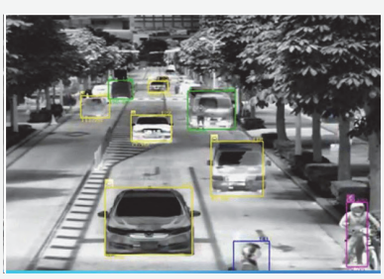
참고 연구/개발 중인 센서 예시

본문에서 언급한 센서 외에도 자율주행 인지에 활용할 수 있는 다양한 센서가 연구 및 개발 단계에 있다. 추후 이 센서들이 실제 자율주행 분야에 적용되면 다중 센서 기술을 활용하지 않더라도, 동기종 센서를 중복으로 활용할 수 있다.

- 4D 이미지 레이더 센서: 거리, 높이, 깊이, 속도 등 4가지 차원에서 환경을 감지하며, 물체의 형태와 속도를 동시에 인식 가능
- 적외선 ToF[Time of Flight] 센서[112]: 반사되는 적외선을 통해 물체를 3차원으로 인식할 수 있는 기술로, 사용자 얼굴 인식 또는 동작 인식에 적용 가능



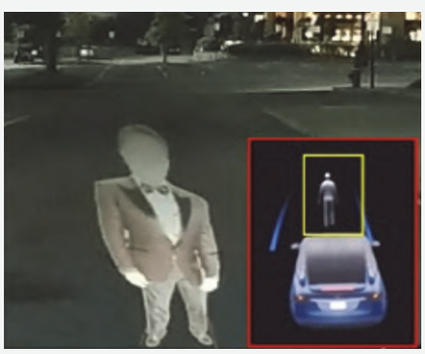
4D 이미지 레이더 센서 예시



적외선 ToF 센서 예시

참고 비전 시스템의 인지 오류 유발 사례

비전 시스템은 카메라를 통한 입력을 활용하므로, 인식된 객체의 깊이를 계산할 수 없다. 이러한 비전 시스템의 한계로 인한 영향을 파악하기 위해, 인지 오류를 의도적으로 유발하는 연구가 진행되기도 했다. 아래 그림(왼쪽)은 도로 노면상에 빔프로젝터를 활용하여 사람 형상의 이미지를 투사했더니 이 이미지를 실제 사람으로 인식하여 차량이 멈추게 된 사례이다. 이 사례를 통해, 아래 그림(오른쪽)과 같이 착시효과로 안전을 높이기 위한 도로 위 3D 트릭아트 역시 비전 시스템의 인지 오류를 유발할 수 있을 것이라 예상해볼 수 있다.



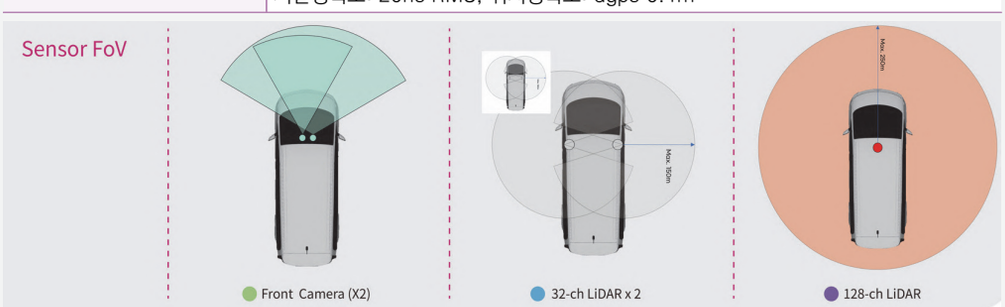
팬텀 이미지를 통한 인지 오류 유발 사례[113]

캐나다 밴쿠버 도로에 그려넣은 착시 그림

Use Case 국내 S사의 자율주행차 다중 센서 기술 적용 사례

- S사에서는 인지 오류를 최소화하기 위해 자율주행차에 다양한 종류의 센서를 적용하고, 각 센서의 FoV[Field Of View]를 분석하여 적용하고 있음

센서 종류	사양
LiDAR (128ch)	측정범위: 120m(80%)/40m(10%), 정확도: 3cm(Avg.), FoV(수직): ±22.5°, 샘플링: 1,310k points/sec 이상
LiDAR (32ch × 4)	측정범위: 100m, 정확도: ±3cm(일반), FoV(수직): ±15°, 샘플링: 300k points/sec 이상
전방 RADAR	측정범위: 60m~200m, FoV: ±9°~±60°, 주파수: 77GHz
Camera F#1	해상도: FHD(1280×1080), 프레임 레이트: 30fps, FoV: 60°
Camera F#2	해상도: FHD(1280×1080), 프레임 레이트: 30fps, FoV: 120°
AVM	해상도: HD(1280×720), Top view 범위: 2.5m
GPS	위성신호: GPS/Glonass/Galileo/Beidou/Qzss, RTK: 0.01m + 1ppm 시간정확도: 20ns RMS, 위치정확도: dgps 0.4m



S사에서 수집하고 있는 상용 승합 자율주행차의 다중 센서 기술 적용 및 FoV 분석 사례

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

Yes ☐ No ☐ N/A ☐

해당여부 판단

자율주행 인공지능의 인지·판단·제어 시스템을 개발하는 경우, 본 항목을 고려하여 만족 여부를 판단하십시오.

- 인공지능 시스템은 서비스 도중 외부의 공격, 사용자의 오용 등 다양한 요인으로 편향이나 성능 저하 등이 발생할 수 있으므로, 시스템 운영자가 이를 파악할 수 있도록 시스템의 자체적인 점검 기능이나, 사용자가 운영자에게 관련 의견을 전달할 수 있는 기능을 제공해야 한다.
- 자율주행 인공지능 시스템에서 편견이나 차별 등 윤리적 문제가 발생할 가능성은 없는지 확인하고, 문제 발생 시 이를 위한 점검 기능 혹은 절차가 수립되었는지 검토해야 한다.
- 자율주행 인공지능 시스템은 다양한 원인에 의해 성능 저하가 일어날 수 있으므로, 이를 지속적으로 평가·관리하기 위한 지표와 절차가 설정되었는지 점검해야 한다.

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

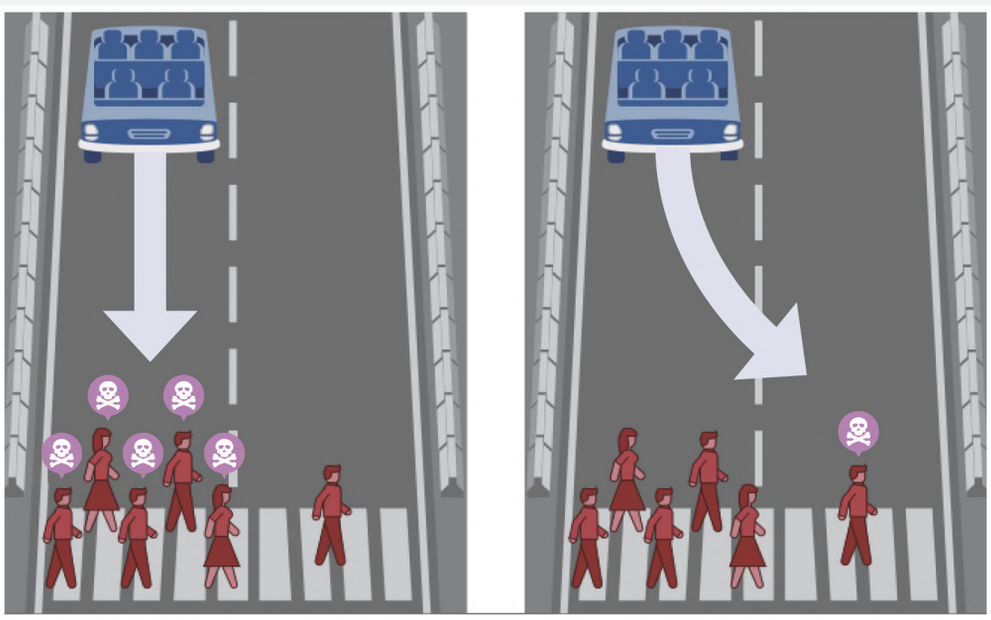
Yes ☐ No ☐ N/A ☐

- 국토교통부에서 발간한《자율주행자동차 윤리 가이드라인》4장 공통 원칙 중 기본원칙 1.3에서는 '손실을 최소화하는 과정에서 인간을 성별, 나이, 종교 등 개인적 차이 등을 이유로 차별하지 않고, 교통약자를 고려하는 방식으로 작동하도록 설계, 제작, 관리되어야 한다.'고 언급하고 있다. 이는 트롤리 딜레마Trolley Dilemma라고도 알려져 있다.
- 자율주행 인공지능 시스템에서 편견이나 차별 등 윤리적 문제의 발생 가능성을 확인하고, 문제 발생 시 이를 위한 알림 기능 혹은 절차가 수립되었는지 점검한다. 윤리적 문제 알림 절차의 경우, 먼저 인공지능 시스템에서 자체적인 신뢰 정도를 평가할 수 있는 기준과 점검 항목을 만든다. 주요 점검 항목의 예시는 다음과 같다.

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

참고 트롤리 딜레마 예시[114]

- 트롤리 딜레마는 다양한 IT 분야에서 문제를 표시하는 데 없어서는 안 될 개념임
- 자율주행 알고리즘이 민감한 결정에 직면할 때 기술적인 문제보다는 윤리적인 문제를 제기하기 때문에, 자율주행 분야에서 특히 고려되어야 함
- 아래 그림에서 시나리오 A(직진하여 보행자 5명과 충돌)와 시나리오 B(직진하여 보행자 1명과 충돌) 중 시나리오 B를 선택한 운전자의 행동은 사상자를 줄이려는 것으로 정당화될 수 있으나, 자율주행 차량의 행동은 데이터 부족으로 인해 아직 정당화될 수 없음



자율주행 분야에서 트롤리 딜레마는 윤리적 문제를 제기함(출처: MIT Technology Review)

- 자율주행 차량을 대상으로 하는 9가지의 서로 다른 양극화 시험 항목
  - ✓ 애완동물보다 인간을 우선해야 하는가?
  - ✓ 보행자보다 승객을 우선해야 하는가?
  - ✓ 적은 생명보다 더 많은 생명을 우선해야 하는가?
  - ✓ 남성보다 여성을 우선해야 하는가?
  - ✓ 노인보다 젊은 사람을 우선해야 하는가?
  - ✓ 병든 사람보다 건강한 사람을 우선해야 하는가?
  - ✓ 사회적 지위가 낮은 사람보다 높은 사람을 우선해야 하는가?
  - ✓ 법을 준수하지 않는 사람보다 준수하는 사람을 우선해야 하는가?
  - ✓ 차량은 방향을 틀어야 하는가(행동을 취함) 또는 코스를 유지해야 하는가(행동하지 않음)?
- 통찰: 누가 죽을 것인지 아닌지를 말하는 것보다 누가 더 위험하거나 덜 위험한지를 분석하는 데 더욱 집중해야 하며, 편향이 어떻게 일어나고 있는지에 대한 위험 분석으로 관점을 옮겨야 함

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

Yes ☐ No ☐ N/A ☐

- 일반적인 소프트웨어와 달리 인공지능 시스템은 서비스 배포 및 운영 단계에서 지속적인 데이터 축적, 서비스 기능 확장, 운영환경의 변화 등의 이유로 성능 변화가 발생할 수 있다.
- 실제 서비스 운영 중 인공지능 시스템의 성능이 갑자기 저하될 때 그 원인을 바로 알기 어려우므로, 시스템의 성능 저하를 지속해서 평가·관리하기 위한 지표와 절차가 설정되었는지 점검할 필요가 있다.
- 내부 오작동(예: 안테나 파손) 또는 외부 교란 요인(예: 악천후 조건)으로 인해 센서 성능 평가에서 성능 저하가 확인되면 이를 시스템 운영자에게 보고하고 운영자는 성능 저하 원인을 찾아 개선을 진행하는 등 절차를 마련해야 한다.
  - ✓ 특히 자율주행 레벨이 높을수록 차량은 인간 운전자에게 의존하지 않기 때문에, 자율주행 차량의 인지 시스템은 매우 높은 수준의 견고성과 신뢰성 측면을 요구한다.
  - ✓ 센서 결함(내부 오작동, 외부 교란 요인 등을 포함)은 자율주행 차량의 예상 기능에 상당한 위험을 초래하므로 지속적으로 인식 센서 성능을 모니터링해야 한다. 아울러 센서 결함이 이후 판단·제어 단계에 영향을 미치는 것을 방지하기 위해 센서 성능 정보를 ADAS/AD 기능에 제공하여 시스템 알림을 통해 시스템 또는 사람이 개입하여 영향을 받은 센서를 복구하거나 시스템 운영자가 개선을 진행하는 등의 절차를 적용할 수 있다.

참고 도로 주행 차량용 자율주행 시스템의 환경 인지 성능 메트릭 예시[115]

- 자율주행 시스템에 적용 가능한 전통적인 인공지능 성능 지표
  - ✓ 참 긍정$^{TP, True\ Positive}$, 거짓 긍정$^{FP, False\ Positive}$, 거짓 부정$^{FN, False\ Negative}$
  - ✓ 민감도(recall): 거짓 부정의 중요성이 높은 경우에 적합
  - ✓ 신뢰도(precision): 거짓 긍정의 중요성이 높은 경우에 적합
  - ✓ 조화 평균(F1 Score)
  - ✓ Jaccard distance
  - ✓ Union 메트릭을 통한 교차점(IoU)
- 객체 감지 및 추적 평가를 위한 메트릭
  - ✓ 다중 객체 추적 정확도(MOTA), 추적 정밀도(MOTP), 감지 정확도(MODA), 감지 정밀도(MODP)
  - ✓ 높은 순위 추적 정확도(HOTA)

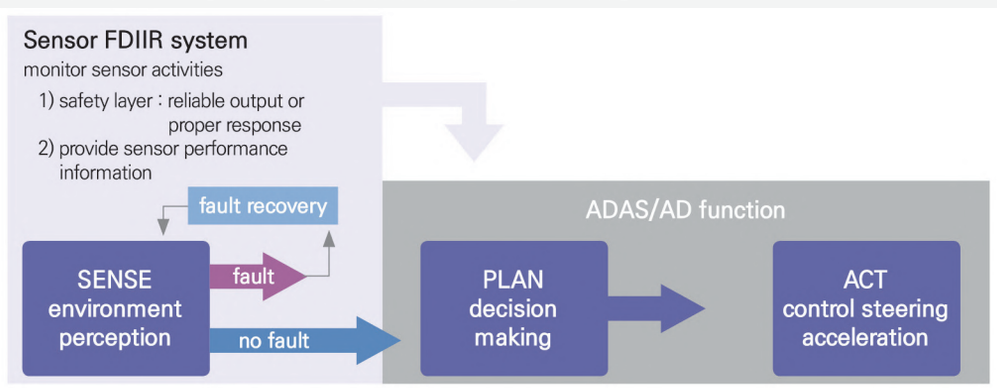
참고 도로 주행 차량용 자율주행 시스템의 모션 계획 성능 메트릭 예시[115]

- 자율주행 시스템에 적용 가능한 전통적인 차량 성능 지표
  - ✓ Time to Collision^{TTC}: 속도를 변경하지 않고 계속 주행하는 경우, 차량과 장애물 간의 충돌을 관찰하는 데 필요한 시간
  - ✓ Time Exposed Time-to-Collision^{TET}: 지정된 임곗값보다 낮게 유지되는 누적 지속시간
  - ✓ Post Encroachment Time^{PET}: 잠재적인 충돌 영역에서 두 차량의 도착 사이의 시간 간격

- 모션 계획 성능 평가를 위한 메트릭
  - ✓ 책임에 민감한 안전$^{RSS,\ Responsibility-Sensitive\ Safety}$ 메트릭
  - ✓ 모델 예측 순간 안전 메트릭$^{MPrISM,\ Model\ Predictive\ Instantaneous\ Safety\ Metric}$: 차량의 안전성을 정량화하기 위한 메트릭
  - ✓ 인공 전위 필드$^{APF,\ Artificial\ Potential\ Fields}$: ADS의 충돌 회피 및 모션 계획에 사용하는 메트릭

참고 센서 결함 감지, 격리, 식별, 복구$^{FDIIR,\ Fault\ Detection,\ Isolation,\ Identification\ and\ Recovery}$ 시스템 연구[116]

- 센서 FDIIR 시스템
  - ✓ 신뢰할 수 있는 환경 인지를 보장하기 위해 센서 활동을 모니터링



센서 FDIIR 시스템을 포함하는 인지·계획·행동(제어) 주기의 개략도

- 인지 센서에 대한 센서 FDIIR 방법의 분류

FDIIR 분류	예시 방법
센서 모델과 비교$^{comparison\ to\ sensor\ model}$	센서 모델에서 감지된 물체와 실제 센서에서 감지한 물체 비교
센서 출력 모니터링$^{monitoring\ sensor\ output}$	센서 출력의 신호 분석 및 타당성 검사
정적 실측 자료와 비교$^{comparison\ to\ static\ ground-truth}$	센서에서 감지한 인프라와 환경의 지상 실측 인프라 비교
동적 실측 자료와 비교$^{comparison\ to\ dynamic\ ground-truth}$	다른 차량이 감지한 도로 사용자와 내 차량이 감지한 도로 사용자 비교
동일한 유형의 다른 센서와 비교$^{comparison\ to\ other\ sensor\ of\ same\ type}$	관찰 중인 센서에서 감지한 물체를 동일한 유형의 다른 센서에서 감지한 물체(두 개의 라이다 센서)와 비교
다른 유형의 다른 센서와 비교$^{comparison\ to\ other\ sensor\ of\ different\ type}$	관찰 중인 센서에서 감지한 물체를 다른 유형의 다른 센서에서 감지한 물체(라이다·레이더 센서)에서 감지한 물체와 비교
내부 인터페이스 모니터링$^{monitoring\ internal\ interface}$	단일 센서 인터페이스 출력의 신호 분석 및 타당성 검사
여러 인터페이스 비교$^{comparison\ of\ multiple\ interfaces}$	센서 인터페이스 사이의 센서 일부가 모델링됨 모델링된 부분의 출력과 해당 센서 인터페이스의 출력 비교

투명성

요구사항 13 인공지능 시스템의 설명에 대한 사용자의 이해도 제고

대표 행위자 | 시스템 엔지니어 협력 대상 | 시스템 운영자 | 인공지능 모델 개발자 | 비즈니스 결정권자 | HMI 전문가

- 자율주행 시스템에서는 운전자, 보행자, 다른 도로 이용자 등 다양한 이해관계자가 복잡한 상호작용을 한다. 그 과정에서 인공지능 활용 모델의 결과에 설명을 제공하거나, 모델의 추론 결과를 사용자에게 안내하는데, 정보가 모호하거나, 어떤 행동으로 반응할지 등 사용자가 바로 이해하기 어려운 경우가 많다. 따라서 인공지능 시스템의 운영자나 서비스 제공자는 사용자에게 제공되는 결과가 이해 가능한지 understandability, 해석 가능한지 interpretability, 설명 가능한지 explainability를 평가하여 사용자의 이해도를 제고한다.

13-1 인공지능 시스템 사용자의 특성 user characteristics 과 제약사항을 분석하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

사용자에게 자율주행 인공지능 시스템의 동작에 대한 설명이 필요한 경우, 본 항목을 고려하여 만족 여부를 판단하십시오.

- 인공지능 시스템의 결과가 적절한지 평가하기 위해서는 먼저 해당 결과를 읽는 사용자를 고려해야 한다. 사용자가 누군지에 따라 결과(설명)의 수준, 깊이 그리고 맥락이 정해지는 만큼 사용자에 대한 자세한 분석이 수행되어야 한다.
- 자율주행 차량의 내외부에 자율주행 인공지능 시스템 관련 상태 표시, 자율주행 관련 시청각 안내, 명확한 공고 등을 하려면 직간접적 사용자의 다양한 특성을 고려해야 한다.

13-1a 사용자 특성에 따른 세부 고려사항을 분석하였는가?

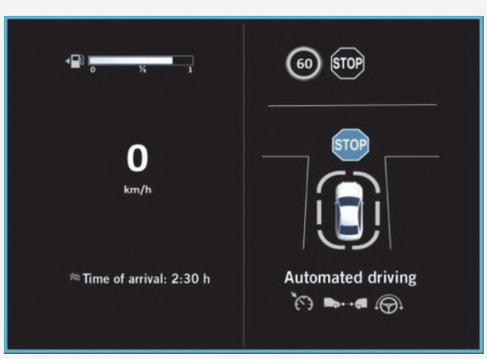
Yes ☐ No ☐ N/A ☐

- 인공지능 시스템의 설명을 평가하려면 각 사용자의 다양한 특성을 고려해야 한다. 자율주행 분야에서 직간접적 사용자는 운전자, 탑승자, 보행자, 다른 도로 이용자(예: 다른 차량 운전자, 자전거 타는 사람, 오토바이 운전자) 등이다. 차량 내외부에 자율주행 인공지능 관련 시스템의 상태 표시, 자율주행 관련 시청각 안내, 명확한 공고 등을 할 때도 사용자 특성을 고려해야 한다[117].
- 사용자 특성을 분석하기 위해 고려해야 할 요소의 예시는 다음과 같다.

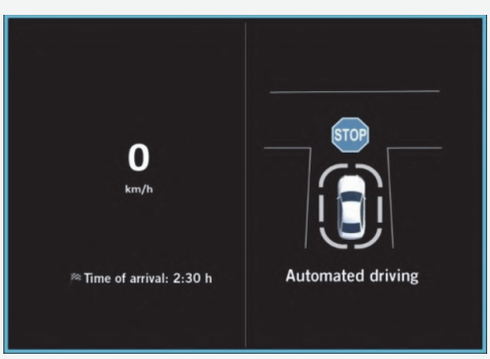
자율주행 분야 직간접 사용자별 세부 고려사항

구분	상세 구분	관련 사용자	고려사항
연령	아동, 성인, 노인 등	탑승자, 보행자	아동의 경우, 성인과 비교해 이해할 수 있는 어휘, 단어가 제한되므로 사용자 연령을 고려해야 함 - 자율주행 시스템의 상태 안내, 시청각 안내, 명확한 공고 시 차량 내외부에서 탑승자, 보행자의 연령을 고려하여 적절한 어휘, 단어를 선정하여 명시
장애 유무	장애인, 비장애인	운전자, 탑승자, 보행자, 다른 도로 이용자	신체적 제약으로 발생할 수 있는 한계를 고려해야 함. 그 예로는 신체 크기, 신체 능력, 인지 능력이 있음 - 운전자, 다른 도로 이용자: 1종 대형·특수 외에는 청력과 관계없이 면허 취득이 가능하므로 청각 메시지 및 인터페이스 사용 시 청각 장애인을 고려해야 함 - 탑승자, 보행자: 자율주행 시스템의 상태 안내, 시청각 안내, 명확한 공고 시 차량 내외부에서 탑승자 및 보행자의 장애 여부를 고려하여 적절한 메시지 전달 방법(시각, 청각, 촉각)을 선정하여 제공
지식	자율주행 시스템 경험 유무	운전자, 보행자, 다른 도로 이용자	자율주행 서비스의 경험 여부와 사전 배경지식의 차이로 지식수준이 다름을 고려해야 함 - 운전자, 다른 도로 이용자, 보행자: 서비스 경험 유무, 주행 시 핸드오버 hand-over를 위한 주의력 차이 등을 고려하여 자율주행 시스템 내외부 안내 사항을 명시

참고 사용자 연령을 고려한 HMI 예시[118]



기존의 인터페이스



사용자 연령을 고려하여 단순화한 인터페이스

- 55세 이상 사용자를 대상으로 인터페이스 선호도 설문조사를 진행한 결과, 시각적으로 단순화된 인터페이스를 선호한다는 결과가 도출되었다.
  - ✓ 시각적 요소를 단순화하고 핵심 정보만 표현한 디스플레이 선호
  - ✓ 각 동작에 대한 신호음 대신 음성 출력 선호

13-2 사용자 특성에 따른 충분한 설명을 제공하는가?

Yes No N/A
☐ ☐ ☐

해당여부 판단

사용자에게 자율주행 인공지능 시스템의 동작에 대한 설명이 필요한 경우, 본 항목을 고려하여 만족 여부를 판단하십시오.

- 자율주행 인공지능 시스템을 이용하는 사용자가 다양하기 때문에 시스템의 결과에 대한 설명이 서로 다른 입장에서 해석되거나 오해가 생길 수 있다. 따라서 13-1 에서 분석된 사용자 특성을 고려하여 설명을 평가하는 기준 항목을 수집해야 하고, 설명 평가의 기준으로는 명확성, 구체성, 정확도 등을 고려해야 한다.
- 자율주행 차량 내 인공지능 관련 시스템에서는 자국어, 간단한 언어 사용 등 사용자의 언어를 이용하여 메시지를 명확하게 전달해야 하며, 시각적으로 충분한 대비를 이루는 휘도를 적용하고, 5가지 이내의 색상을 사용하는 등 사용자의 명확한 이해를 통해 행동을 끌어낼 수 있는 표현을 사용하는 것이 바람직하다.
- 사용자의 정상 시선 내에 가장 중요한 정보들을 배치하거나, 적절한 표시 타이밍을 적용하여 사용자가 명확하게 반응할 수 있도록 표현하는 등의 방안도 필요하다. 이러한 설명에 따라 사용자의 시스템 수용 및 신뢰에 미치는 영향에 대한 사용자 경험UX, User eXperience 관련 측면도 평가해볼 수 있다.

13-2a 사용자 특성에 따른 설명 평가의 기준을 수립하였는가?

Yes No N/A
☐ ☐ ☐

- 다양한 사용자(예: 운전자, 탑승자, 보행자)가 서비스를 이용하는 만큼 설명을 포괄적으로 평가할 수 있는 특성과 세부 항목을 정하는 단계가 필요하다. 설명의 평가 기준은 명확성, 구체성, 적절성, 정확성 등의 항목이 될 수 있다. 세부 항목으로 데이터 유형data type이나 모달리티modality에 따라 각 항목에서 고려되어야 할 내용들이 달라질 수 있다. 다음은 설명 평가를 위한 예시이다.

설명 평가항목 및 기준

구분	평가항목
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?

• 자율주행 차량에서 인간-기계 인터페이스 원칙이 충족되는지를 평가하는 것도 도움이 된다[119]. 이 체크리스트는 차량 내 인터페이스와 관련된 기존 규범, 적용 표준, 설계 지침, 실증 연구에서 파생된 시각, 청각, 촉각을 활용한 HMI에 대한 가장 중요한 설계 권장 사항 20여 개 항목으로 구성되어 있다.

자율주행 차량의 HMI 설계 권장 사항 목록

#	항목	명확성	구체성	적절성	정확성
1	의도하지 않은 자율주행 모드의 활성화 및 비활성화 방지				✓
2	시스템 모드는 계속 표시되어야 함			✓	
3	모드 변경 사항을 효과적으로 전달해야 함		✓	✓	
4	시스템 상태를 전달하는 데 사용되는 시각적 인터페이스는 적절한 위치와 거리에 장착해야 하며, 우선순위가 높은 정보는 운전자의 예상 시야에 가깝게 표시되어야 함			✓	
5	HMI 요소는 기능에 따라 그룹화하여 모드 표시기의 인식을 지원해야 함	✓			
6	시간에 민감한 시스템과의 상호작용은 지속적인 주의를 가지지 않아야 함			✓	
7	시각적 인터페이스는 전경과 배경 사이의 휘도 및/또는 색상이 충분히 대비되어야 함	✓			
8	텍스트(예: 글꼴 유형, 문자 크기)와 기호는 허용된 좌석 위치에서 쉽게 읽을 수 있어야 함	✓			
9	자동화 모드를 전달하려면 일반적으로 허용되거나 표준화된 기호를 사용해야 하며, 비표준 기호를 사용할 때는 추가 텍스트 설명으로 보완해야 함			✓	
10	메시지의 의미는 긴급성에 따라야 함		✓		
11	메시지는 사용자의 언어를 사용하여 전달해야 함(예: 자국어, 기술 언어 사용, 공통 구문 사용)	✓			
12	문자를 활용한 메시지는 되도록 짧아야 함	✓			

#	항목	명확성	구체성	적절성	정확성
13	시스템 상태를 코딩하는 데 5가지 이상의 색상을 일관되게 사용해서는 안 됨 (흰색 및 검은색 제외)	✓			
14	시스템 상태를 전달하는 데 사용되는 색상은 일반적인 관습 및 고정 관념에 따라야 함	✓			
15	중복 코딩으로 색맹을 위한 디자인, 빨강/녹색 및 파랑/노랑 조합을 피함	✓			
16	청각 출력은 운전자를 놀라게 하지 않고 주의를 환기해야 함			✓	
17	청각 및 진동 촉각 출력은 메시지의 긴급성에 맞게 조정되어야 함		✓		
18	우선 순위가 높은 메시지는 다중 감각을 통해 안내해야 함				✓
19	경고 메시지는 사용자에게 위험 원인을 안내해야 함		✓		
20	센서 고장인 경우, 그 결과 및 작업자 요구사항 단계가 표시되어야 함				✓

13-2b 사용자가 이해하기 어려운 전문 용어 사용을 지양하였는가?

Yes ☐ No ☐ N/A ☐

- 자율주행 차량의 외부 인터페이스를 활용하여 사용자에게 시스템의 상태, 안내 메시지 등을 송출하여 소통하는 경우, 외부의 다양한 보행자, 도로 이용자를 배려해 되도록 전문 용어를 최대한 사용하지 않는 것이 바람직하다.
- 13-2a 에서 설명한 자율주행 차량의 HMI 주요 20개 설계 권장 사항 중 텍스트 및 문자 메시지와 관련된 항목은 8~12번이다. 그 중, 11번 항목에서, 메시지는 사용자의 언어를 사용하여 전달하라고 요구하며, 다음과 같은 요소를 체크하여 검증할 수 있다.
  - ✓ 자국어를 사용하였는가?
  - ✓ 단순한 언어를 사용하였는가?
  - ✓ 약어를 사용하지는 않았는가?
  - ✓ 현재의 자율주행 레벨(0~5)이 아닌, 활성화된 자율주행 및 운전자 보조 기능 명칭(예: 차선 유지 기능, 크루즈 컨트롤 기능)을 표시하였는가?

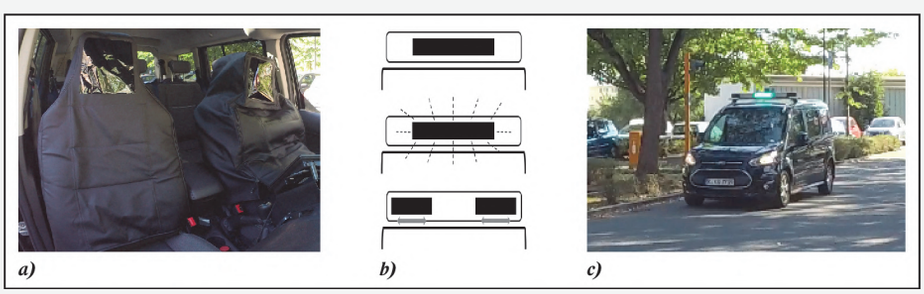
13-2c 사용자의 구체적인 행동과 이해를 이끌어낼 수 있도록 명확한 표현을 사용하였는가?

Yes ☐ No ☐ N/A ☐

- 좋은 설명은 사용자의 구체적인 행동과 이해를 이끌 수 있어야 한다. 따라서 설명을 간결하고 명확하게 함으로써 해석이 모호하지 않도록 작성하는 것이 중요하다.
- 시각적으로는 성공·실패·경고·위험 등 결과에 따른 색상을 일관성 있게 유지해 줌으로써 사용자가 한눈에 이해할 수 있도록 할 수 있다. 이처럼 자율주행 차량에서 표현의 명확성 제고를 위해 13-2a의 설계 권장 사항 중 7, 13번 항목과 같은 내용을 고려할 수 있다.

참고 자율주행 분야 인공지능 신뢰성 확보를 위한 eHMI[external HMI] 개발 예시

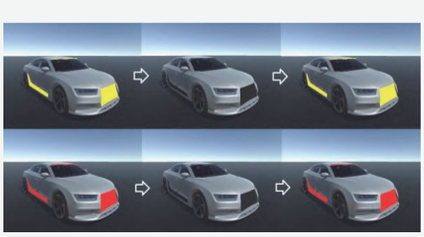
- 물리적 프로토타입을 활용하는 보행자에게 안전과 편안함을 부여하는지에 관한 사용자 인터뷰[120]
  - ✓ 참가자 대다수는 인터페이스의 유무(기준 조건)에 관계없이 자율주행 차량에 비해 수동 구동 차량과 상호작용하는 것이 더 안전하다고 느낀다고 보고하였음
  - ✓ 인터페이스는 직관적으로 이해할 수 없으며 부분적으로만 신뢰할 수 있는 것으로 나타났음



a) 운전자를 덮고 있는 시트 슈트의 자동주행 조건
b) 조명 신호의 시각화(자동화 모드, 시동 모드, 횡단 모드)
c) 차량 지붕의 라이트 바

- 컴퓨터 디자인 활용에 대한 긴급성 평가[121]
  - ✓ 빠르게 접근하는 자율주행 차량의 경우, 보행자에게 상황의 긴급성을 전달할 수 있는 잠재력과 관련한 인터페이스 평가
  - ✓ 접근하는 차량의 애니메이션 비디오(50km/h, 일정한 속도, 차량 감속)를 시청
  - ✓ 보행자 관점을 고려하며 각 디자인에 대한 긴급성을 평가하고, 도로를 횡단할 가능성을 표시함

차량 모델링 및 외부 디스플레이



경고 색상 점멸

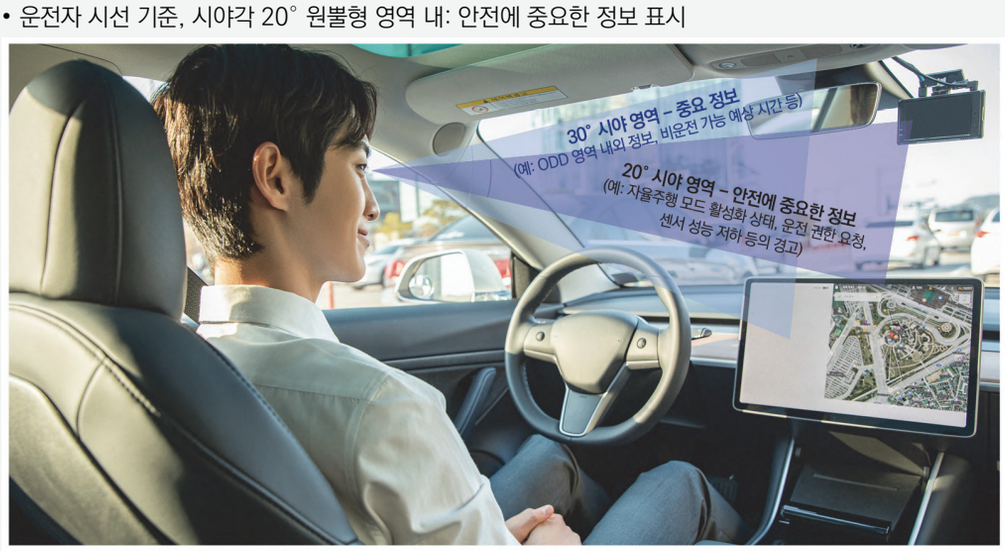
13-2d 설명이 필요한 위치와 타이밍은 적절한가?

Yes No N/A
☐ ☐ ☐

- 잘 작성된 설명이 적절한 위치 및 타이밍에 나타나 이해를 돕는 것도 중요하다. 이를 위해 단발성으로 설명해야 할지, 여러 번 반복하여 강조해야 할지 숙고하고, 사용자가 잘 읽으려면 어느 위치에 놓여야 할지 고려해야 한다.
- 자율주행 차량에서 운전자, 탑승자를 대상으로 설명이 필요한 위치와 타이밍을 선정하기 위해 13-2a 의 설계 권장 사항을 참고하여 4번, 6번과 같은 항목을 고려할 수 있다.

참고 시각적 인터페이스의 적절한 표시 위치 예시

- 운전자 시선 기준, 시야각 30° 원뿔형 영역 내: 중요 정보 표시
- 운전자 시선 기준, 시야각 20° 원뿔형 영역 내: 안전에 중요한 정보 표시



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

Yes No N/A
☐ ☐ ☐

- 사용자 경험은 한 개인이 특정한 제품, 시스템, 또는 서비스를 사용하며 느끼는 모든 것을 의미한다. 또한, 개인이 인지하는 유용성, 사용 편의성, 효율성 등의 시스템 특성을 포함한다. 설명을 평가하기 위해 사용자 조사user research 기법을 활용할 수 있다.
- 사용자 조사 기법은 크게 접근 방식과 자료 획득 방식으로 구분할 수 있다. 우선, 사용자 조사 기법의 접근 방식에 따라 정량적(간접적) 조사와 정성적(직접적) 조사로 구분되며, 사용자 조사를 위해 자료를 얻는 방식에 따라 사용자 행동을 통한 조사와 태도를 통한 조사로 구분한다. 접근 및 자료 획득 방식을 고려해 적합한 사용자 조사 기법을 선정하고, 사용자 경험을 평가하는 것이 바람직하다.

✓ 접근 방식에 따른 구분 및 방법

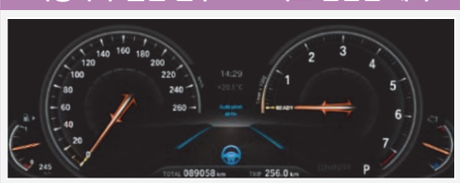
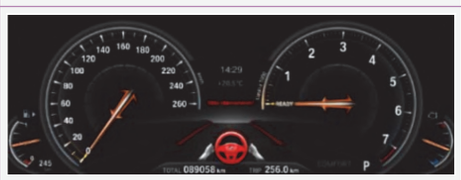
- 정량적(간접적) 조사quantitative user research: 사용자의 행동이나 태도에 대한 데이터를 도구 등을 통해 간접적으로 수집하는 방법 (예: 웹로그 분석, A/B 테스트A/B testing, 설문 조사, 고객 지원 자료 분석)
- 정성적(직접적) 조사qualitative user research: 사용자의 행동이나 태도를 직접 관찰하는 방법 (예: 인터뷰, 표적 집단 인터뷰focus group interview, 프로토타입 테스팅prototype testing)

✓ 자료 획득 방식에 따른 구분 및 방법

- 사용자 행동 기반 조사behavioral user research: 사용자가 무슨 행동을 하는지를 조사하는 방법 (예: 웹로그 분석, A/B 테스트, 아이 트래킹eye tracking)
- 사용자 태도 기반 조사attitudinal user research: 사용자가 무엇을 말하는지를 조사하는 방법 (예: 카드 소팅card sorting, 심층 인터뷰, 요구사항 조사)

참고 자율주행 분야 인공지능 신뢰성 확보를 위한 HMI 개발

• 사용자 인터뷰를 통한 HMI 휴리스틱 예측 타당성 평가 결과[122]
✓ 드라이버 12명이 낮거나 높은 준수 HMI를 시뮬레이션 환경에서 30분간 운행하며 평가

모드	사용자가 '높은 준수 HMI'라고 판단한 예시	사용자가 '낮은 준수 HMI'라고 판단한 예시
L3 ADS 활성		
주의 TOR		
임박한 TOR		
L3 ADS 사용 불가		

05 운영 및 모니터링

책임성 투명성

요구사항 14 인공지능 시스템의 추적가능성 및 변경이력 확보

대표 행위자 | 시스템 엔지니어 협력 대상 | 인공지능 모델 개발자 데이터 과학자

- 자율주행 시스템은 시간에 따라 운전 주체가 달라져 시스템과 운전자가 어떤 수행 결과와 의사결정을 내렸는지 추적·분석하고 재연해야 한다. 이를 위해 문제 원인 추적을 위한 시스템 로그, 인공지능 모델과 사람 간의 의사결정 기여도 추적 등 여러 방안을 확보해야 한다. 또한, 시스템 내 인공지능 모델의 성능 개선을 위해 변경된 데이터를 활용하여 모델 재학습 등을 수행했을 때, 데이터의 변경 시점, 접근 사용자, 변경 내용 등을 모니터링하고 변경이력을 관리하는 등 기술적 대응 방안을 확보한다.

14-1 인공지능 시스템의 의사결정에 대한 추적 방안을 수립하였는가?

Yes ☐ No ☐ N/A ☐

해당여부 판단

자율주행 인공지능의 인지·판단·제어 시스템을 개발하는 경우, 본 항목을 고려하여 만족 여부를 판단하십시오.

- 인공지능 시스템의 의사결정은 인공지능 모델이 자체 결정하거나 시스템 운영자 또는 사용자가 개입해 내릴 수 있다. 이때, 의사결정의 주체와 의사결정 기여도 등을 산정, 분석하여 특정 이벤트 또는 사고 발생 시 책임 소재를 명확히 밝힐 수 있는 기반을 마련해야 한다.
- 자율주행 알고리즘의 판단·제어 인공지능 모델을 활용하여 시스템을 구축한 경우, 인공지능 시스템은 위계적인 4가지 요소를 조합하여 주행 의사결정을 한다. 이때, 인공지능 시스템 내 각 모델의 추론 통합 결과를 확인하고, 운전자 및 탑승자도 의사결정에 관여할 수 있으므로 인공지능 시스템, 운전자 및 탑승자의 포괄적 기여도 기준 등을 확립하여 최종 의사결정의 주체를 명확하게 정하여, 이를 확인 및 추적하는 방안(예: 로그)을 마련해야 한다.
- 사용자 경험을 효과적으로 제공하기 위해서는 인공지능 시스템의 의사결정 결과, 사용자 개입 요청, 시스템의 현재 상태 알림, 사용자 행동을 유도하기 위한 시스템 표시기 등에 대한 사용자 반응시간 정보 등을 수집 및 분석 관리해야 한다.

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

Yes ☐ No ☐ N/A ☐

- 인공지능 시스템의 결정에 대한 모델 기여도를 파악하려면 이전 모델의 추론 정보와 최종 결정에 대한 사람(예: 운전자) 개입 여부 등의 정보가 추적되어야 한다.
- 자율주행 시스템은 인지·판단·제어와 관련하여 다음과 같은 위계 관계의 4가지 의사결정 요소에 따른다[123]. 추후 인공지능 모델의 학습 결과로 이러한 요소가 자율주행 인공지능 시스템에 반영된 경우, 각 모델의 추론 결과를 사람이 검토하여 의사결정을 내릴 수 있다.
  - ✓ 길 또는 경로 계획path or route planning
  - ✓ 행동 중재behavior arbitration
  - ✓ 모션 계획motion planning
  - ✓ 차량 제어vehicle control
- 판단·제어 등을 위한 인공지능 모델 기반 자율주행 시스템의 결정 사항대로 주행하거나, 결정 사항에 운전자 및 탑승자가 개입하는 경우, 시스템 내 각 인공지능 모델의 기여도와 운전자의 개입까지 포괄하여 시스템 결정에 대한 세분화한 기여도 기준을 내부적으로 확립하고, 시스템 운용 과정에서 이를 추적하는 방안(예: 로그 수집)을 확보해야 한다.

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

Yes ☐ No ☐ N/A ☐

- 국토교통부에서 발간한《레벨4 자율주행자동차 제작안전 가이드라인》에서는 자율주행 시스템의 의사결정 요소와 관련하여 사고 또는 이벤트 상황을 재현하고, 데이터 기록 장치를 설치 및 설정하여 향후 사고 상황을 통한 학습 및 성능 개선에 활용하도록 가이드하고 있다.
  - ✓ 기록 항목: 자율주행 시스템의 작동 여부, 해제 원인, 제어권 전환 요구, 비상 운행의 시작과 종료 등
  - ✓ 기록 분량: 각 상황의 발생 사유, 발생 일시를 초 단위로 저장. 저장된 데이터는 사고 원인 및 책임 소재를 명확히 밝힐 수 있도록 최소 6개월 이상 또는 2,500건(자율주행 시작·종료, 사용자 개입, 주요 사건 이벤트 등의 저장 건수. 일 평균 12~15건씩 6개월 분량에 해당) 이상의 기록을 보존하여 데이터기록장치를 설치하도록 가이드
- 추후 자율주행 알고리즘 중 판단·제어를 위한 인공지능 모델을 반영한 시스템을 구축할 때, 기능 구현 방안을 마련하여 인공지능 시스템의 의사결정 요소를 재현하고, 의사결정 과정을 추적하도록 한다.

14-1c 지속적인 사용자 경험 모니터링을 위해 사용자 로그를 수집 및 관리하고 있는가?

Yes ☐ No ☐ N/A ☐

- 서비스 이용 로그 분석은 서비스 운영 상태에 관한 확인뿐만 아니라, 사용자가 겪는 문제가 무엇인지 확인할 수 있는 가장 기본적인 방법이 될 수 있다. 서비스 로그는 서비스가 운영되는 동안 지속해서 수집되며 서비스 고도화에 따라 다양한 형태로 누적될 수 있다.
- 자율주행을 위한 인공지능 모델이 반영된 시스템에서는 사용자에게 제어 권한을 넘겼을 때 사용자의 반응시간, 인공지능의 주행 경로 계획 수용·미수용 여부 등 사용자의 서비스 이용 시 로그를 수집 및 분석하여 더욱 나은 서비스를 제공하는 데 활용한다.
- 이때, 사용자는 온·오프라인에서 받은 자율주행 서비스에 대한 사용자 경험 로그의 제공 여부를 결정할 수 있다. 단, 민감한 개인정보가 포함되어 있을 수 있으므로, 기업은 활용하는 개인정보의 범위를 정해 고시하고 개인정보는 익명 또는 가명으로 처리해야 한다.

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

Yes ☐ No ☐ N/A ☐

해당여부 판단

자율주행 인공지능 모델의 개선을 위한 학습 데이터의 변경이 발생하는 경우, 본 항목을 고려하여 만족 여부를 판단하십시오.

- 인공지능 모델은 사용한 데이터에 따라 학습 모델도 함께 달라진다. 이에 따라 모델의 설계나 주요 파라미터들도 함께 변경된다. 따라서 모델 개발 과정에서 학습 데이터가 변경될 경우, 학습 데이터의 버전관리 및 변경이 발생한 원인을 추적해야 한다.
- 또한, 신규 데이터를 포함하여 인공지능 모델의 추가 학습이 필요한 경우, 학습 데이터 변경으로 인한 모델의 성능 영향을 평가하기 위해 기존 학습 데이터에 추가된 신규 데이터 비율에 따른 모델 성능 변화 추적이 가능하도록 기록 및 관리하는 것이 바람직하다.
- 이외에, 학습에 오픈소스 데이터셋을 활용하는 경우, 활발하게 개발되고 있는 오픈소스 데이터셋은 수시로 보완 및 업데이트가 이루어지므로, 인공지능 모델 및 시스템 성능 개선을 위해 주기적으로 모니터링해야 한다.
- 이러한 학습 데이터 변경이력 관리를 위해 학습 데이터 버전관리를 위한 오픈소스 도구 활용, 자체 시스템 구축 등을 고려할 수 있으며, 학습 데이터를 사용 또는 운용하는 이해관계자들이 데이터 변경으로 인한 영향을 확인할 수 있도록 학습 데이터 변경 원인, 변경된 학습 데이터의 구조 및 학습 모델 예상 출력 및 모델 변경으로 인한 성능 평가 결과 등에 대한 정보를 제공해야 한다.

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

Yes ☐ No ☐ N/A ☐

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

14-2b 데이터 소스 변경에 대한 모니터링 방안을 확보하였는가?

Yes No N/A
☐ ☐ ☐

- 자율주행 알고리즘 중 인지 시스템을 위한 학습 데이터를 실시간으로 수집하여 인공지능 모델을 학습시키지 않는다면 본 검증항목은 현재 고려사항이 아닐 수 있다.
- 인지 시스템을 위한 인공지능 알고리즘 개발을 위해 오픈소스 데이터셋을 활용하는 경우, 데이터셋이 변경되거나 업데이트될 수 있어, 주기적으로 모니터링을 실시해 최신 데이터셋을 모델의 성능 개선에 반영할 수 있다.

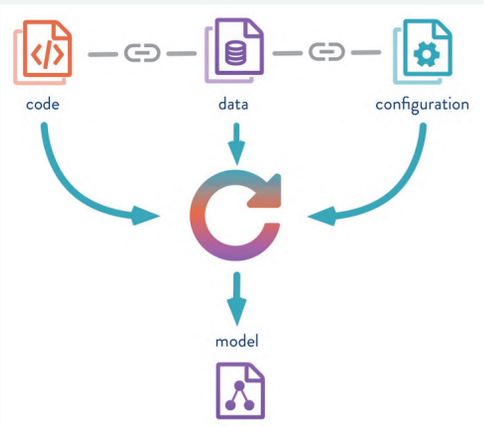
14-2c 데이터 변경 시, 버전관리를 수행하였는가?

Yes No N/A
☐ ☐ ☐

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

참고 데이터 버전관리[DVC, Data Version Control] 도구[124]

- 오픈소스 비주얼 스튜디오 코드 확장 및 명령 줄 도구
- Git 저장소 위에서 작동하며, Git과 비슷한 명령 줄 인터페이스 및 흐름을 가지고 있음
- DVC는 데이터 및 기계학습 시험을 문서화함
- 큰 파일, 데이터셋 디렉토리, 기계학습 모델 등을 작은 메타파일(Git으로 처리하기 쉬운)로 대체하여 관리함. 즉, 소스코드 관리와 분리되는 원본 데이터를 가리킴
- 데이터 저장소: 온-프레미스 또는 클라우드 저장소를 사용하여 프로젝트의 데이터를 코드 기반과 별도로 사용 가능



14-2d 데이터 변경 시, 이해관계자를 위한 정보를 제공하는가?

Yes ☐ No ☐ N/A ☐

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

데이터 변경 시 이해관계자에게 제공해야 할 정보 예시

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

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

Yes ☐ No ☐ N/A ☐

- 신규 데이터를 확보한 뒤, 인공지능 시스템에 사용하기 위해서는 기존 운영 중인 인공지능 모델과의 성능 비교가 필요하다. 사람이 판단하기에 신규 데이터가 기존 학습 데이터와 유사하여도 학습된 인공지능 모델이 기존 학습 데이터에서 학습한 데이터 특성과 다를 수 있다.
- 따라서 신규 데이터를 대상으로 도메인의 대표적인 인공지능 알고리즘을 사용하여 성능평가를 진행하고 분석하는 과정이 필요하다. 신규 데이터 확보에 따른 성능평가를 위해서는 다음과 같은 과정을 참고한다.
  - ✓ 성능평가 및 비교 분석을 위한 기존 학습 모델 및 관련 대표 인공지능 모델 확보
  - ✓ 대상 인공지능 분야 및 모델에 적절한 성능평가 지표 선정
  - ✓ 성능평가를 위한 실험 설계(정량적·정성적 실험 방법 선정, 실험 모델들의 파라미터 설정, 세부 실험 계획)
  - ✓ 실험 진행 및 결과 분석(결과에 따라 신규 데이터 평가 또는 필요한 경우 모델 재설계, 확장, 재학습 등 결정)

책임성 투명성

요구사항 15 서비스 제공 범위 및 상호작용 대상에 대한 설명 제공

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

- 사용자가 자율주행 시스템이 제공하는 서비스를 올바르게 사용하고, 제공된 서비스를 오남용하지 않도록 서비스의 목적, 범위, 제한사항, 면책조항disclaimer, 상호작용 대상 등에 관한 설명을 제공한다.

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

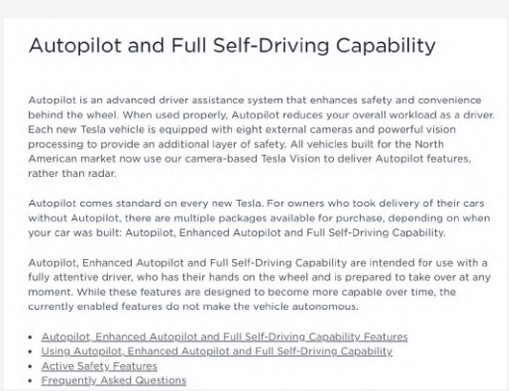
Yes ☐ No ☐ N/A ☐

해당여부 판단

자율주행 인공지능 서비스를 운영하는 경우, 본 항목을 고려하여 만족 여부를 판단하십시오.

- 인공지능의 활용 범위가 넓어지면서 사용자가 서비스 기능을 실제 서비스 제공 범위보다 더 넓게 기대하여 오해하는 경우가 발생한다. 따라서 서비스 목적, 범위, 제한사항, 면책조항 등에 대한 설명을 제공함으로써 인공지능 기술의 오남용을 방지하고, 서비스에 대한 사용자의 기대치를 조정하는 것이 중요하다.
- 사용자가 인공지능 기반 자율주행 서비스에서 일어나는 돌발 상황에 적절히 대응할 수 있도록 서비스의 범위, 작동한계, 대응 방법 등에 대한 설명을 제공하여 서비스를 올바르게 사용할 수 있도록 유도해야 한다.

참고 Tesla사의 자율주행 서비스 목적·범위·한계 소개 사례



Autopilot and Full Self-Driving Capability

Autopilot is an advanced driver assistance system that enhances safety and convenience behind the wheel. When used properly, Autopilot reduces your overall workload as a driver. Each new Tesla vehicle is equipped with eight external cameras and powerful vision processing to provide an additional layer of safety. All vehicles built for the North American market now use our camera-based Tesla Vision to deliver Autopilot features, rather than radar.

Autopilot comes standard on every new Tesla. For owners who took delivery of their cars without Autopilot, there are multiple packages available for purchase, depending on when your car was built: Autopilot, Enhanced Autopilot and Full Self-Driving Capability.

Autopilot, Enhanced Autopilot and Full Self-Driving Capability are intended for use with a fully attentive driver, who has their hands on the wheel and is prepared to take over at any moment. While these features are designed to become more capable over time, the currently enabled features do not make the vehicle autonomous.

- Autopilot, Enhanced Autopilot and Full Self-Driving Capability Features
- Using Autopilot, Enhanced Autopilot and Full Self-Driving Capability
- Active Safety Features
- Frequently Asked Questions

- Tesla는 홈페이지에서 자율주행 서비스의 서비스 기능을 Autopilot, Enhanced Autopilot, Full Self-Driving으로 분류하고, 기능별 운전자 지원 범위 및 한계를 소개하고 있다[125].

15-1a 서비스의 목적과 목표에 대한 설명을 제공하는가?

Yes No N/A
☐ ☐ ☐

- 국토교통부에서 발간한《레벨 4 자율주행자동차 제작안전 가이드라인》에서는 자율주행 시스템을 올바로 이해하고 사용하도록 다음과 같은 정보 제공을 가이드한다. 이에 따라 인공지능 서비스의 목적과 목표를 설명하여 사용자가 자율주행 서비스를 이해하고 대처 방안 등을 숙지할 수 있도록 해야 한다.
  - ✓ 자료: 이용자 설명서, 가상 교육, 온라인 교육, 현장 교육
  - ✓ 정보 제공 내용: 자율주행 시스템의 구성, 작동 방법, 상황대처 요령
- 자율주행 서비스가 오용 또는 남용될 경우, 인공지능 모델이나 시스템상의 새로운 취약점을 생성하거나 예상치 못한 인명피해를 일으킬 수 있다. 따라서 서비스가 의도한 목적에서 벗어나 잘못 사용되는 것을 방지하기 위해, 이해관계자는 잠재적 오남용 영역을 식별한 후 사용자가 이를 인식할 수 있도록 관련 사례와 처벌 내용 등을 알려야 한다.

참고 Tesla사의 오토파일럿 기능의 핸즈오프handsoff 경고 횟수를 줄이기 위한 장치 남용 사례

- 사용자 남용 발생 배경
  - ✓ Tesla사의 오토파일럿 기능은 사용자 계약 조건에 따라 운전자가 반드시 스티어링휠을 잡고 있어야 한다고 경고함
  - ✓ 운전자들은 핸즈오프 경고 알림 간격이 짧아 불편함을 느끼고, 알림을 회피할 방안을 찾음
- 사용자 남용 사례
  - ✓ 스티어링휠에 손을 올려놓았다고 속이기 위해 장착 가능한 장치가 판매되기 시작[126]
  - ✓ 핸즈오프 경고 없이 오토파일럿 기능을 계속 사용하기 위해 스티어링 휠에 물병, 오렌지 등을 올려 경고 기능을 무력화하는 시도도 발생[127]
- 관련 참고 사항
  - ✓ 전문가들은 스티어링휠 장착용 제품이 운전자의 안전을 전혀 고려하지 않는다고 경고
  - ✓ Tesla사는 오토파일럿 기능의 남용 방지를 위해 도움이 될 수 있는 스티어링휠 센서 및 시선 추적 도입을 고려했지만, 비용 및 효율성 문제로 인해 폐기한 것으로 알려짐
  - ✓ (참고) 국토교통부에서 발간한《자율주행자동차 윤리 가이드라인》에서는 자율주행 자동차의 보유자나 이용자가 자율주행 시스템을 불법으로 개조하거나 임의로 변경하여 안전을 위해하는 행위를 방지할 수 있도록 해야 한다고 언급함


스티어링휠에 물병, 오렌지 등을 올려 경고 기능을 무력화하는 시도[127]

핸즈오프 알림 회피를 위한 장착용 제품[126]

15-1b 서비스의 한계와 범위에 대한 설명을 제공하는가?

Yes ☐ No ☐ N/A ☐

- 미국자동차공학회SAE, Society of Automotive Engineers에서는 운전자 보조에서부터 완전 자동화까지 자율주행 단계별 서비스의 한계 및 범위를 규정하고 있다.
- 자율주행 서비스에 적용된 인공지능 인지·판단·제어 시스템의 기술 및 성능 수준에 따라 서비스의 한계 및 범위에 대한 정보를 명확하게 제공하여, 사용자가 이에 적절히 대응할 수 있도록 한다. 제공해야 할 정보의 예시는 다음과 같다.
  - ✓ 서비스 범위: 인공지능 기반 자율주행 서비스의 운행조건, 운행가능영역ODD, Operational Domain Design, SAE 단계 기반 자율주행 레벨
  - ✓ 작동 한계: 카메라의 오염으로 인한 인지 시스템의 오동작 가능성
  - ✓ 대응 방법: 인공지능 서비스의 작동 및 해제 방법, 고장 대응 절차, 잠재적 위험을 최소화하기 위한 정보 등

15-2 상호작용의 대상을 명확히 설명하는가?

Yes No N/A
☐ ☐ ☐

해당여부 판단

자율주행 인공지능이 운전자·보행자 등 사용자와 직접적으로 상호작용하는 경우, 본 항목을 고려하여 만족 여부를 판단하십시오.

- 수동 주행 차량, 반 자율주행 차량, 완전 자율주행 차량이 동시에 운행하는 혼합 자율 교통환경에서는 각 도로 사용자에게 주행 차량의 현재 상태를 알리는 고급 의사소통 인터페이스가 있어야 한다. 이를 통해 모든 도로 사용자의 안전과 효율성을 극대화하고, 신기술에 대한 신뢰와 수용을 강화해야 한다.
- 이를 위해, HMI를 적용하여 자율주행 차량의 현재 시스템 모드를 내외부에 표시함으로써 현재 인공지능 시스템이 동작 중임을 모든 도로 사용자에게 명확히 알리고 대처할 수 있도록 한다.

참고 자율주행 차량과 도로 사용자 간 의사소통을 위한 물리적 프로토타입 HMI 적용 예시

• 실리콘밸리에 본사를 둔 Drive.ai의 HMI 적용 예시이다.

(1) 자율주행 차량임을 표시
- 차량 전면에 "Self-Driving Vehicle"을 표시하여 자율주행 차량임을 전달

(2) 자율주행 시스템 모드 표시
- 차량 후드, 프런트 펜더 및 후면에 LED 패널을 부착하여 차량 모드 및 의도 전달
- Person Driving: 수동 모드 시, 모든 패널에 운전사 그림과 함께 표시
- Waiting for you to cross: 보행자에게 양보할 때 측면 패널에 표시
- Pedestrian crossing: 보행자가 횡단하는 동안 후면 패널에 표시

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

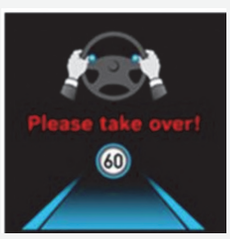
Yes ☐ No ☐ N/A ☐

- 자율주행 차량은 HMI를 도입해 자율주행 시스템 모드를 표시함으로써 운전자에게 인공지능과 상호작용하고 있음을 명확히 설명해야 한다.
- 자율주행 시스템 모드 표시 예시는 (1) 자율주행 모드가 작동 중이지 않지만, 기능 활성화를 통해 작동이 가능한 상태, (2) 자율주행 모드가 제대로 작동 중인 상태, (3) 자율주행 모드 작동이 불가능하여 운전자의 수동 운전으로 제어 전환을 요청한 상태, (4) 수동 운전 중인 상태 등이 있다.
- 또한, HMI 도입 후에는 도로 사용자가 자율주행 인공지능의 모드와 의도를 명확하게 인지하고 느끼는지 조사 및 평가가 필요하다.

참고 자율주행 시스템 모드 표시 예시[128]

(1) 자율주행 모드 활성화 가능

(2) 자율주행 모드 작동 중



(3) 제어권 전환 요청

(4) 수동 운전 모드

PART 3

부록

1. 약어표

2. 용어표

3. 참고문헌

01 약어표

약어표

ABS	Anti-lock Brake System
ADAS	Advanced Driver Assistance System
ADAS/AD	Advanced Driver Assistance System & Autonomous Driving
ADS	Autonomous Driving System
AI	Artificial Intelligence
APF	Artificial Potential Fields
APT	Advanced Persistent Threat
ASAM	Association for Standardization of Automation and Measuring Systems
AVM	Around View Monitoring
BRT	Bus Rapid Transit
CAM	Class Activation Map
CAN	Controller Area Network
CNN	Convolutional Neural Networks
CVE	Common Vulnerability and Exposures
DACOBS	Davos Assessment of Cognitive Biases Scale
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DDT	Dynamic Driving Task
DGM	Deep Generative Model
DMS	Driver Monitoring System
DoS	Denial of Service
DVC	Data Version Control
EC	European Commission
E/E	Electrical and/or Electronic
eHMI	external Human Machine Interface
FDIIR	Fault Detection, Isolation, Identification and Recovery
FN	False Negative
FOV	Field Of View
FP	False Positive
FPS	Frame Per Second
FSD	Full-Self Driving
GNSS	Global Navigation Satellite System

01 약어표

GPS	Global Positioning System
HMI	Human Machine Interface
HMM	Hidden Markov Models
IDS	Intrusion Detection System
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
IMU	Inertial Measurement Unit
IoU	Intersection over Union
IP	Intellectual Property
ISO	International Organization for Standardization
JSON	JavaScript Object Notation
LiDAR	Light Detection And Ranging
LKAS	Lane Keeping Assist System
MDPS	Motor Driven Power Steering
ML	Machine Learning
MPrISM	Model Predictive Instantaneous Safety Metric
NBDT	Neural-Backed Decision Trees
NDRA	Non-Driving Related Activities
NHTSA	National Highway Traffic Safety Administration
NTSB	National Transportation Safety Board
ODD	Operational Design Domain
OECD	Organisation for Economic Co-operation and Development
OSI	Open Source Initiative
PET	Post Encroachment Time
R-CNN	Region Based Convolutional Neural Networks
RADAR	RAdio Detection And Ranging
RF	Random Forest
RMF	Risk Management Framework
RSS	Responsibility Sensitive Safety
RVM	Relevance Vector Machine
SAE	Society of Automotive Engineers
SBL	Sparse Bayesian Learning
SCC	Smart Cruise Control

01 약어표

SOTA	State Of The Art
SVM	Surround View Monitor (vehicle) / Support Vector Machine (machine learning)
TAI	Trustworthy Artificial Intelligence
TAS	Trustworthy Autonomous Systems
TET	Time Exposed Time-to-Collision
TOR	Take Over Request
ToF	Time of Flight
TP	True Positive
TR	Technical Reports
TTC	Time to Collision
UNF	Universal Numerical Fingerprint
UNESCO	United Nations Educational, Scientific and Cultural Organization
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
UX	User eXperience
V2X	Vehicle to Everything communication
WEF	World Economic Forum
XAI	eXplainable Artificial Intelligence
XML	eXtensible Markup Language

02 용어표

용어표

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

용어명	정의
차량간 통신 V2X, Vehicle to everything	차량간 통신은 무선망을 통해 다른 차량 및 도로 등의 인프라와 정보를 교환하는 기술을 의미한다.
전동식 파워 스티어링 MDPS, Motor Driven Power Steering	전동식 파워 스티어링은 보조동력을 이용하여 스티어링 휠을 쉽게 조작할 수 있는 시스템인 파워스티어링 기술 중 하나로, 토크torque센서가 스티어링 휠의 회전 방향과 속도를 감지하여 모터를 구동함으로써 보조동력을 제공하는 기술을 의미한다.
관성 센서 IMU, Inertia Measurement Unit	관성 센서는 물체의 관성을 측정하여 물체가 기울어진 각도를 측정하는 센서로, 자이로스코프gyroscope, 가속도계accelerometer, 지자기 센서geomagnetic sensor로 구성된다.
Frenet 좌표계	Frenet 좌표계는 도로(경로)에서 차량의 위치를 변수 s와 d를 사용하여 설명하는 좌표 방법으로, s 좌표는 도로를 따른 거리(세로 변위)를 의미하고, d 좌표는 도로의 좌우 위치(측면 변위)를 의미한다.
테스트 케이스 test case	테스트 케이스는 구현된 소프트웨어가 사용자의 요구사항을 정확하게 준수했는지 확인하기 위해 설계된 입력값, 실행 조건, 기대 결과 등으로 구성된 테스트 명세서를 의미한다.
테스트 스위트 test suite	테스트 스위트는 테스트 수행 목적에 따라 그룹화된 테스트 케이스의 집합이다.
자율주행 알고리즘 아키텍처	자율주행차의 동작을 위한 알고리즘 구성을 인지-판단-제어의 3가지로 정의한 것이다. 자율주행차는 차량에 장착된 각종 센서로부터 수집된 데이터를 종합하여 상황을 '인지'하고, 인지된 상황에 근거하여 차량을 어떻게 제어하고 주행해야 할지 '판단'하며, 이러한 주행제어 측면의 판단에 근거하여 차량을 '제어'하는 알고리즘을 포함한다.
라이다 LiDAR, Light Detection And Ranging	라이다는 레이저를 목표물에 비춤으로써 사물까지의 거리, 방향, 속도, 온도, 물질 분포 및 농도 특성 등을 감지할 수 있는 기술이다.
레이더 RADAR, RAdio Detection And Ranging	레이더는 전자파가 목표물에 부딪힌 뒤 되돌아오는 반사파를 측정하여 대상을 탐지하고, 방향, 거리, 및 속도 등을 파악하는 기술이다.

02 용어표

용어명	정의
운행 가능 영역 ODD, Operational Design Domain	자율주행시스템의 기능이 정상적이고 안전하게 수행될 수 있는 자동 영역으로, 도로, 기상 및 교통 등이 이에 해당한다.
지능형 지속 공격 APT, Advanced Persistent Threat	지능형 지속 공격은 특정 실체를 목표로 지속적인 해킹 시도를 통해 개인 정보와 같은 중요한 데이터를 유출하는 형태의 공격을 의미한다.
비운전 관련 활동 NDRA, Non-Driving Related Activity	비운전 관련 활동은 운전자가 운전하지 않는 동안 할 수 있는 활동으로, 주행과 관련 없는 작업을 의미한다.
스니핑 sniffing	스니핑은 네트워크상에서 자신이 아닌 다른 상대방의 패킷 정보를 도청하는 행위이다.
스푸핑 spoofing	스푸핑은 공격자가 네트워크, 웹사이트 등의 데이터 위변조를 통해 정상 시스템인 것처럼 위장하여 일반 사용자를 속이는 해킹 기법이다.
전파 방해 jamming	전파 방해는 고의적으로 초고주파 에너지를 방사하여 특정 전파의 사용을 방해하는 기술이다.
제어권 전환 요청 TOR, Take-Over Request	제어권 전환 요청은 자율주행 차량이 운전자에게 운전의 제어를 요청하고 제어 권한을 넘기는 기술이다.
범지구 위성 항법 시스템 GNSS, Global Navigation Satellite System	범지구 위성 항법 시스템은 인공위성을 이용해 위치를 결정할 수 있게 하는 체계로, 위성에서 발신한 전파를 이용하여 언제, 어디서, 누구에게나 정밀한 측위 정보를 제공

03 참고문헌

참고문헌

- [1] 산업통상자원부, "**자율주행 레벨 4+ 상용화 앞당긴다**," 2021. 3.
- [2] J. So, "**기술표준이슈 – 자율주행 알고리즘**," TTA ICT Standard Weekly, no. 1057, 2021. 11.
- [3] S. Jung, Y. Moon, S. Lee, and K. Hwang, "**Impacts of Automated Vehicles on Traffic Flow Changes**," The Journal of The Korea Institute of Intelligent Transportation Systems, vol. 16, no. 6, pp. 244-257, 2017. 12.
- [4] ABC News, "**Deadly Crash with Self-driving Uber**," 2019. 11.
- [5] T. Park and H. Jin, "**자율주행차 사망사고에 따른 시사점**," 소프트웨어정책연구소 월간SW중심사회, 2016. 8.
- [6] Youtube, **Tesla Model 3 Autopilot No Nag Ever - My Cheap DIY Hack**, [Online], Available: https://youtu.be/1WV1aYNisq4
- [7] WIRED, **Hackers Remotely Kill a Jeep on the Highway—With Me in It**, [Online], Available: https://www.wired.com/2015/07/hackers-remotely-kill-jeep-highway/
- [8] ISO 31000, "**Risk management — Guidelines**," 2018. 2.
- [9] ISO/IEC 24028, "**Overview of trustworthiness in artificial intelligence**," 2020. 5.
- [10] ISO/IEC 23894, "**Guidance on risk management**," 2023. 2.
- [11] ISO 26262, "**Functional safety**," 2011. 11.
- [12] ISO 21448, "**Road vehicles — Safety of the intended functionality**," 2022. 6.
- [13] S. Lee, "**ISO 26262 and ISO/PAS 21448 as Exemption Clauses of Product Liability**," Journal of Institute of Korean Electrical and Electronics Engineers, vol. 23, no. 1, pp. 346-349, 2019. 3.
- [14] M. Kim, T. Kim, and Y. Kim, "**On the Integrated process of RSS model and ISO / DIS 21448 (SOTIF) for securing autonomous vehicle safety**," Journal of the Korea Society of Systems Engineering, vol. 17, no. 2. pp. 129-138, 2021. 12.
- [15] ISO/IEC 38507, "**Governance implications of the use of artificial intelligence by organizations**," 2022. 4.
- [16] CLAYTEX, **Virtual development and testing of autonomous vehicles – where to start?**, [Online], Available: https://www.claytex.com/blog/virtual-development-and-testing-of-autonomous-vehicles-where-to-start/
- [17] ISO/IEC TR 29119-11, "**Guidelines on the testing of AI-based systems**," 2020. 11.

03 참고문헌

- [18] Y. Kim, S. Park, I. Kim, H. Ko, S. Cho, and I. Yun, "**Study on Establishment of Development Strategy for K-City Based on Analysis of Domestic and Overseas Automated Vehicle Test beds**," The Journal of The Korea Institute of Intelligent Transportation Systems, vol. 20, no. 4, pp. 28-45, 2021. 8.
- [19] NVIDIA, **Omniverse 기반의 NVIDIA DRIVE Sim**, [Online], Available: https://www.nvidia.com/ko-kr/self-driving-cars/simulation/
- [20] Lyft, **Level5 - Tutorial: Perception for Self-Driving Cars**, [Online], Available: https://level-5.global/data/perception/
- [21] AI Hub, **차량 및 사람 인지 영상**, [Online], Available: https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=195
- [22] S. D. Pendleton, H. Andersen, X. Du, X. Shen, M. Meghjani, Y. H. Eng, D. Rus, and M. H. Ang, "**Perception, Planning, Control, and Coordination for Autonomous Vehicles**," Machines, vol. 5 no. 1, 2017. 2.
- [23] S. Ha, "**Special Report – 자율주행 지원을 위한 고정밀지도 기술 동향**," TTA Journal, vol. 173, 2017. 10.
- [24] M. Brandao, "**Age and gender bias in pedestrian detection algorithms**", arXiv:1906.10490v1, 2019. 6.
- [25] The Register, **Tesla Full Self-Driving 'fails' to notice child-sized objects in testing**, [Online], Available: https://www.theregister.com/2022/08/09/Tesla_autopilot_child_testing/
- [26] Vox, **A new study finds a potential risk with self-driving cars: failure to detect dark-skinned pedestrians**, [Online], Available: https://www.vox.com/future-perfect/2019/3/5/18251924/self-driving-car-racial-bias-study-autonomous-vehicle-dark-skin
- [27] B. Wilson, J. Hoffman, and J. Morgenstern, "**Predictive Inequity in Object Detection**," arXiv: 1902.11097v1, 2019. 2.
- [28] MathWorks, **Visualize, Label, and Fuse Sensor Data for Automated Driving**, [Online], Available: https://kr.mathworks.com/company/newsletters/articles/visualize-label-and-fuse-sensor-data-for-automated-driving.html
- [29] TensorFlow, **Get started with Tensorflow Data Validation**, [Online], Available: https://www.tensorflow.org/tfx/data_validation/get_started
- [30] Adversarial ML Threat Matrix, **Annoucing ATLAS!**, [Online], Available: https://github.com/mitre/advmlthreatmatrix
- [31] ITWORLD, "**'적대적 머신러닝 대응 전략의 시작' 방어가 시작됐다**," [Online], Available: https://www.itworld.co.kr/news/175699

03 참고문헌

- [32] Y. Cao, C. Xiao, D. Yang, J. Fang, R. Yang, M. Liu, and B. Li, "**Adversarial Objects Against LiDAR-Based Autonomous Driving Systems**," arXiv:1907.05418v1, 2019. 7.
- [33] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "**Robust Physical-World Attacks on Deep Learning Models**," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1625-1634, 2018. 4.
- [34] S. Ding, Y. Tian, F. Xu, Q. Li, and S. Zhong, "**Trojan Attack on Deep Generative Models in Autonomous Driving**," Security and Privacy in Communication Networks: 15th EAI International Conference, pp. 299-318, 2019. 10.
- [35] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "**Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks**," 2016 IEEE symposium on security and privacy, pp. 582-597, 2016. 3.
- [36] W. Xu, D. Evans, and Y. Qi, "**Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks**," arXiv:1704.01155v2, 2017. 12.
- [37] comma.ai, **Keep your eyes on the road**, [Online], Available: https://comma.ai/
- [38] TECH WORLD, **국토부, '자율차 주행 데이터' 수집 차량 대여 실시**, [Online], Available: https://www.epnc.co.kr/news/articleView.html?idxno=95160
- [39] KITTI, **Welcom to he KITTI Vision Benchmark Suite!**, [Online], Available: https://www.cvlibs.net/datasets/kitti/
- [40] B. Wilson, J. Hoffman, and J. Morgenstern, "**Predictive Inequity in Object Detection**," arXiv: 1902.11097v1, 2019. 2.
- [41] J. C. McCall, D. P. Wipf, M. M. Trivedi, and B.D. Rao, "**Lane Change Intent Analysis Using Robust Operators and Sparse Bayesian Learning**," IEEE Transactions on Intelligent Transportation Systems, vol. 8, no. 3, pp. 431-440, 2007. 9.
- [42] D. Polling, M. Mulder, M. M. Paassen, and Q. P. Chu, "**Inferring the Driver's Lane Change Intention Using Context-based Dynamic Bayesian Networks**," In Proceedings 2005 IEEE International Conference on Systems, vol. 1, pp. 853-858, 2005. 10.
- [43] D. D. Salvucci, and A. Liu, "**The Time Course of a Lane Change: Driver Control and Eye-movement Behavior**," Transportation Research Part F: Traffic Psychology and Behaviour, vol. 5, no. 2, pp. 123-132, 2002. 6.
- [44] D. D. Salvucci, H. M. Mandalia, N. Kuge, and T. Yamamura, "**Lane-change Detection Using a Computational Driver Model**," Human Factors: The Journal of the Human Factors and Ergonomics Society, vol. 49, no. 3, pp. 532-542, 2007. 6.

03 참고문헌

- [45] H. M. Mandalia and M. D. D. Salvucci, "**Using Support Vector Machines for Lane-change Detection**," In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 49, no. 22, pp. 1965-1969, 2005. 9.
- [46] M. Itoh, K. Yoshimura, and T. Inagaki, "**Inference of Large Truck Driver's Intent to Change Lanes to Pass a Lead Vehicle via Analyses of Driver's Eye Glance Behavior in the Real World**," In Society of Instrument and Control Engineers(SICE) Annual Conference 2007, pp. 2385-2389, 2007. 9.
- [47] M. J. Henning, O. Georgeon, and J. F. Krems, "**The Quality of Behavioral and Environmental Indicators Used to Infer the Intention to Change Lanes**," In 4th International Driving Sympo sium on Human Factors in Driver Assessment, vol. 4, no. 2007, pp. 231-237, 2007. 7.
- [48] B. Morris, A. Doshi, and M. Trivedi, "**Lane Change Intent Prediction for Driver Assistance: On-road Design and Evaluation**," In 2011 IEEE Intelligent Vehicles Symposium (IV), pp. 895 -901, 2011. 6.
- [49] N. Kuge, T. Yamamura, O. Shimoyama, and A. Liu, "**A Driver Behavior Recognition Method Based on a Driver Model Framework**," Journal of Passenger Cars, vol. 109, pp. 469-476, 2000. 1.
- [50] S. Lee and S. Lee, "**Selection of Machine Learning Algorithm and Input Variables for Detecting Driver's Lane Change Intention**," Korean Journal of Computational Design and Engineering, vol. 25, no. 1, pp. 24-35, 2020. 3.
- [51] ASAM, **ASAM OpenLABEL®**, [Online], Available: https://www.asam.net/standards/detail/openlabel/
- [52] Open Source Initiative, **The Open Source Definition**, [Online], Available: https://opensource.org/osd
- [53] SCATTER LAB Tech, **하나의 조직에서 TensorFlow와 PyTorch 동시 활용하기**, [Online], Available: https://tech.scatterlab.co.kr/torch-to-tf-tf-to-torch/
- [54] OpenVINO, **Converting a TensorFlow Model**, [Online], Available: https://docs.openvino.ai/latest/openvino_docs_MO_DG_prepare_model_convert_model_Convert_Model_From_TensorFlow.html
- [55] TensorFlow, **Migrate the SavedModel workflow**, [Online], Available: https://www.tensorflow.org/guide/migrate/saved_model
- [56] CVE Details, **Google 〉〉 TensorFlow: Vulnerability Statistics**, [Online], Available: https://www.cvedetails.com/product/53738/Google-Tensorflow.html?vendor_id=1224
- [57] K. Evans, N. Moura, S. Chauvier, R. Chatila, and E. Dogan, "**Ethical Decision Making in Auto nomous Vehicles: The AV Ethics Project**," Science and Engineering Ethics, vol. 26, pp. 3285 -3312, 2020. 10.

03 참고문헌

- [58] D. Frank, P. Chrysochou, P. Mitkidis, and D. Ariely, "**Human decision-making biases in the moral dilemmas of autonomous vehicles**," Scientific Reports, vol. 9, no. 1, 2019. 9.
- [59] ResearchGate, **Breakdowns in Human-AI Partnership: Revelatory Cases of Automation Bias in Autonomous Vehicle**, [Online], Available: https://www.researchgate.net/publication/337910493_Breakdowns_in_Human-AI_Partnership_Revelatory_Cases_of_Automation_Bias_in_Autonomous_Vehicle
- [60] L. Peng, H. Wang, and J. Li, "**Uncertainty Evaluation of Object Detection Algorithms for Autonomous Vehicles**," Automotive Innovation, vol. 4, no. 3, pp. 241-252, 2021. 6.
- [61] X. Gong, Y. Chen, Q. Wang, W. Yang, and X. Jiang, "**Model Extraction Attacks and Defenses on Cloud-Based Machine Learning Models**," IEEE Communications Magazine, vol. 58, no. 12, pp. 83-89, 2020. 12.
- [62] Insights2Techinfo, **Self-Driving Automobiles and Adversarial Attacks**, [Online], Available: https://insights2techinfo.com/self-driving-automobiles-and-adversarial-attacks/
- [63] Tencent Keen Security Lab, "**Experimental Security Research of Tesla Autopilot**," 2019. 3.
- [64] Medium, **On security and safety of HD maps**, [Online], Available: https://medium.com/yoda yoda/on-security-and-safety-of-hd-maps-be295a491662
- [65] LG CNS, **머신러닝 보안 취약점! 적대적 공격의 4가지 유형**, [Online], Available: https://blog.lgcns.com/2191
- [66] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "**Distillation as a defense to adversarial perturbations against deep neural networks**," In 2016 IEEE symposium on security and privacy, pp. 582-597, 2016. 5.
- [67] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "**Intriguing properties of neural networks**," in International Conference on Learning Representations, 2014. 4.
- [68] D. Omeiza, H. Webb, M. Jirotka, and L. Kunze, "**Explanations in Autonomous Driving: A Survey**," IEEE Transactions on Intelligent Transportation Systems, vol. 23. no. 8, pp. 10142-10162, 2021. 11.
- [69] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, "**Textual explanations for self-driving vehicles**," in Proceedings of the European conference on computer vision(ECCV), pp. 563-578, 2018. 9.
- [70] T. You and B. Han, "**Traffic accident benchmark for causality recognition**," in Proceedings of the European conference on computer vision(ECCV), pp. 540-556, 2020. 8.

- [71] J. Houston, G. Zuidhof, L. Bergamini, Y. Ye, L. Chen, A. Jain, S. Omari, V. Iglovikov, and P. Ondruska, "**One thousand and one hours: Self-driving motion prediction dataset**," Proceedings of Machine Learning Research, pp. 409-418, 2021. 12.
- [72] Y. Xu, X. Yang, L. Gong, H. C. Lin, T. Y. Wu, Y. Li, and N. Vasconcelos, "**Explainable object-induced action decision for autonomous vehicles**," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9523-9532, 2020. 6.
- [73] V. Ramanishka, Y. T. Chen, T. Misu, and K. Saenko, "**Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning**," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7699-7707, 2018. 6.
- [74] Y. Shen, S. Jiang, Y. Chen, E. Yang, X. Jin, Y. Fan, and K. D. Campbell, "**To explain or not to explain: A study on the necessity of explanations for autonomous vehicles**," arXiv:2006.11684 v4, 2020. 6.
- [75] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "**Learning deep features for discriminative localization**," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921-2929, 2016. 6.
- [76] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "**Grad-cam: Visual explanations from deep networks via gradient-based localization**," in Proceedings of the IEEE international conference on computer vision, pp. 618-626, 2017. 10.
- [77] Z. Tang, K. V. Chuang, C. DeCarli, L.-W. Jin, L. Beckett, M. J. Keiser, and B. N. Dugger, "**Interpretable classification of alzheimer's disease pathologies with a convolutional neural network pipeline**," Nature communications, vol. 10, no. 1, pp. 1-14, 2019. 5.
- [78] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "**Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks**," in IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 839-847, 2018. 5.
- [79] D. Omeiza, S. Speakman, C. Cintas, and K. Weldermariam, "**Smooth Grad-CAM++: an enhanced inference level visualization technique for deep convolutional neural network models**," Proceedings of the 2019 Intelligent Systems Conference, 2019. 9.
- [80] M. Bojarski, A. Choromanska, K. Choromanski, B. Firner, L. J. Ackel, U. Muller, P. Yeres, and K. Zieba, "**Visualbackprop: Efficient visualization of CNNs for autonomous driving**," in IEEE International Conference on Robotics and Automation (ICRA), pp. 4701-4708, 2018. 6.
- [81] S. Lapuschkin, S. Waldchen, A. Binder, G. Montavon, W. Samek, and K. R. Muller, "**Unmasking clever hans predictors and assessing what machines really learn**," Nature communications, vol. 10, no. 1, pp. 1-8, 2019. 3.

03 참고문헌

- [82] W. Samek, T. Wiegand, and K. R. Muller, "**Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models**," ITU Journal: ICT Discoveries, no. 1, pp. 39-48, 2017. 10.
- [83] A. Shrikumar, P. Greenside, and A. Kundaje, "**Learning important features through propaga ting activation differences**," In International conference on machine learning, pp. 3145-3153, 2017. 7.
- [84] D. Zeiler and R. Fergus, "**Visualizing and understanding convolutional networks**," in European Conference on Computer Vision. pp. 818-833, 2014. 9.
- [85] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "**Striving for simplicity: The all convolutional net**," in 3rd International Conference on Learning Representations, 2015. 5.
- [86] R. Borgo, M. Cashmore, and D. Magazzeni, "**Towards providing explanations for AI planner decisions**," in the IJCAI/ECAI 2018 Workshop on Explainable Artificial Intelligence, 2018. 7.
- [87] R. Korpan and S. L. Epstein, "**Toward natural explanations for a robot's navigation plans**," in Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, 2018. 3.
- [88] J. Bidot, S. Biundo, T. Heinroth, W. Minker, F. Nothdurft, and B. Schattenberg, "**Verbal plan explanations for hybrid planning**," in Multikonferenz Wirtschaftsinformatik(MKWI), pp. 2309-2320, 2010. 2.
- [89] Y. Zhang, S. Sreedharan, A. Kulkarni, T. Chakraborti, H. H. Zhuo, and S. Kambhampati, "**Plan explicability and predictability for robot task planning**," in IEEE International Conference on Robotics and Automation (ICRA), pp. 1313-1320, 2017. 5.
- [90] T. Chakraborti, S. Sreedharan, Y. Zhang, and S. Kambhampati, "**Plan explanations as model reconciliation: Moving beyond explanation as soliloquy**," arXiv:1701.08317, 2017. 5.
- [91] T. Chakraborti, S. Sreedharan, S. Grover, and S. Kambhampati, "**Plan Explanations as Model Reconciliation-An Empirical Study**," in 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 258-266, 2019. 3.
- [92] S. Sasai, I. Kitahara, Y. Kameda, Y. Ohta, M. Kanbara, Y. Morales, N. Ukita, N. Hagita, T. Ikeda, and K. Shinozawa, "**MR visualization of wheel trajectories of driving vehicle by seeing-through dashboard**," in IEEE International Symposium on Mixed and Augmented Reality Workshops, pp. 40-46, 2015. 9.
- [93] L. Marques, V. Vasconcelos, P. Pedreiras, and L. Almeida, "**A flexible dashboard panel for a small electric vehicle**," in 6th Iberian Conference on Information Systems and Technologies (CISTI 2011), pp. 1-4, 2011. 6.

- [94] T. Schneider, J. Hois, A. Rosenstein, S. Ghellal, D. Theofanou-Fulbier, and A. R. Gerlicher, "**ExplAIn yourself! transparency for positive ux in autonomous driving**," in Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1-12, 2021. 5.
- [95] T. G. Reid, S. E. Houts, R. Cammarata, G. Mills, S. Agarwal, A. Vora, and G. Pandey, "**Localization requirements for autonomous vehicles**," SAE International Journal of Connected and Automated Vehicles, vol. 2, no. 3, pp. 173-190, 2019. 9.
- [96] J. Kim and J. Canny, "**Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention**," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2942-2950, 2017. 3.
- [97] A. Wan, L. Dunlap, D. Ho, J. Yin, S. Lee, S. Petryk, S. A. Bargal, and J. E. Gonzalez, "**NBDT: NEURAL-BACKED DECISION TREE**," in International Conference on Learning Representations (ICLR), 2021. 5.
- [98] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "**Towards Causal Representation Learning**," in Proceedings of the IEEE - Advances in Machine Learning and Deep Neural Networks, 2021. 2.
- [99] F. Sprenger, "**Microdecisions and autonomy in self-driving cars: virtual probabilities**," AI & SOCIETY, vol. 37, pp. 619-634, 2020. 12.
- [100] Forbes, **Explaining Why Explainable AI (XAI) Is Needed For Autonomous Vehicles And Especially Self-Driving Cars**, [Online], Available: https://www.forbes.com/sites/lanceeliot/2021/04/24/explaining-why-explainable-ai-xai-is-needed-for-autonomous-vehicles-and-especially-self-driving-cars/?sh=2784346e1c5a
- [101] T. Hecht, S. Danner, A. Feierle, and K. Bengler, "**Does a Confidence Level for Automated Driving Time Estimations Improve the Subjective Evaluation of an Automation HMI?**," Multimodal Technologies and Interaction, vol. 4, no. 3, 2020. 7.
- [102] S. Chen, Z. Jian, Y. Huang, Y. Chen, Z. Zhou, and N. Zheng, "**Autonomous driving: cognitive construction and situation understanding**," vol. 62, pp. 1-27, 2019. 7.
- [103] F. Naujoks, Y. Forster, K. Wiedemann, and A. Neukum, "**A Human-Machine Interface for Cooperative Highly Automated Driving**," Advances in Human Aspects of Transportation, vol. 484, pp. 585-595, 2016. 7.
- [104] SOLUTIONLINK, **자율주행 Fail operational safety architecture**, [Online], Available: http://www.sol-link.com/neo/kr/consulting/Fail_Operational_Architecture.php
- [105] "**Fail safe system**," [Online], Available: https://hujubkang.tistory.com/entry/Fail-safe-system
- [106] D. Watzenig, "**Automated Driving - Challenges and Opportunities**," IEEE Austria Section, 2017. 10.

03 참고문헌

- [107] Youtube, **Tesla's FSD Beta – An Experiment On Public Roads**, [Online], Available: https://www.youtube.com/watch?v=D_SPymCay18
- [108] G. Dede, R. Naydenov, A. Malatras, R. Hamon, H. Junklewitz, and I. Sachez, "**Cybersecurity challenges in the uptake of Artificial Intelligence in Autonomous Driving**," European Union Agency for Cybersecurity (ENISA) and Joint Research Centre (JRC), 2021. 2.
- [109] S. Oh, "**자율주행시대에 대비한 첨단도로인프라 정책방안**," KRIHS Policy Brief, no. 637, 2017. 11.
- [110] Brunch, **자율 주행과 TOR(Take Over Request)**, [Online], Available: https://brunch.co.kr/@gloriachoi-ux/3
- [111] VISTEON, **Driver monitoring systems for a safe, autonomous future**, [Online], Available: https://www.visteon.com/technology/interior-sensing
- [112] LG CNS, **일상생활로 들어오는 자율주행 센서**, [Online], Available: https://blog.lgcns.com/2198
- [113] B. Nassi, D. Nassi, R. Netanel, Y. Mirsky, O. Drokin, and Y. Elovici, "**Phantom of the adas: Phantom attacks on driver-assistance systems**," Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, pp. 293-309, 2020. 10.
- [114] MIT Technology Review, **Should a self-driving car kill the baby or the grandma? Depends on where you're from**, [Online], Available: https://www.technologyreview.com/2018/10/24/139313/a-global-ethics-study-aims-to-help-ai-solve-the-self-driving-trolley-problem/
- [115] M. N. Sharath and B. Mehran, "**A Literature Review of Performance Metrics of Automated Driving Systems for On-Road Vehicles**," Frontiers in Future Transportation, vol. 2, 2021.11.
- [116] T. Goelles, B. Schlager, and S. Muckenhuber, "**Fault Detection, Isolation, Identification and Recovery (FDIIR) Methods for Automotive Perception Sensors Including a Detailed Literature Survey for Lidar**," Sensors, vol. 20, no. 13, 2020. 6.
- [117] H. Woo and G. Lee, "**A Study on Assessment Items and Considerations for Development of KNCAP of Automated Driving System**," Journal of Auto-vehicle Safety Association, vol. 13, no. 3, pp. 102-110, 2021. 9.
- [118] S. Schwindt, N. Theobald, B. Abendroth, and D. Manstentten, "**Requirements of Elderly Drivers for the HMI of Automated Vehicles**," ATZ Worldwide, vol. 124, no. 5, pp. 58-61, 2022. 4.
- [119] N. Schömig, K. Wiedemann, S. Hergeth, Y. Forster, J. Muttart, A. Eriksson, D. M. Rundus, K. Grove, J. Krems, A. Keinath, A. Neukum, and F. Naujoks, "**Checklist for Expert Evaluation of HMIs of Automated Vehicles—Discussions on Its Value and Adaptions of the Method within an Expert Workshop**," Information, vol. 11, no. 4, 2020. 4.

- [120] A. Hensch, I. Kreißig, M. Beggiato, J. Halama, and J. Krems, "**Effects of a light-based communication approach as an external HMI for Automated Vehicles – a Wizard-of-Oz Study**," Transactions on Transport Sciences, vol. 10, no. 2, pp. 18-32, 2020. 1.
- [121] Y. Li, M. Dikmen, T. G. Hussein, Y. Wang, and C. Burns, "**To cross or not to cross: urgency-based external warning displays on autonomous vehicles to improve pedestrian crossing safety**," 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, pp. 188-197, 2018. 9.
- [122] F. Naujoks, S. Hergeth, A. Keinath, K. Wiedemann, and N. Schömig, "**Development and application of an expert assessment method for evaluating the usability of SAE Level 3 ADS HMIs**," 26th International Technical Conference on the Enhanced Safety of Vehicles, 2019. 6.
- [123] Neptune.ai, **Self-Driving Cars With Convolutional Neural Networks (CNN)**, [Online], Available: https://neptune.ai/blog/self-driving-cars-with-convolutional-neural-networks-cnn
- [124] DVC, **Open-source Version Control System for Machine Learning Projects**, [Online], Available: https://dvc.org/
- [125] Tesla, **Autopilot and Full Self-Driving Capability**, [Online], Available: https://www.tesla.com/support/autopilot
- [126] ELECTREK, **Tesla Autopilot 'buddy' hack to avoid 'nag' relaunches as 'phone mount' to get around NHTSA ban**, [Online], Available: https://electrek.co/2018/09/09/Tesla-autopilot-buddy-hack-avoid-nag-relaunch-phone-mount-nhtsa-ban/
- [127] NATE NEWS, **Tesla '오토 파일럿' 속이는 방법?…핸들에 손 얹고 집중해야**, [Online], Available: https://news.nate.com/view/20210421n13625?mid=n1101
- [128] C. Purucker, F. Berghöfer, F. Naujoks, K. Wiedemann, and C. Marberger, "**Prediction of Take-Over Time Demand in Highly Automated Driving. Results of a Naturalistic Driving Study**," HFES Europe Chapter Annual Meeting, 2018. 10.

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

2023

신뢰할 수 있는 인공지능 개발 안내서 자율주행 분야

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

ISBN 979-11-290- 4928-5 (93550) [정가 18,000원]

2023

신뢰할 수 있는 인공지능 개발 안내서 자율주행 분야

Jinhan M&B

정가 18,000원

93550

9 791129 049285

ISBN 979-11-290-4928-5

과학기술정보통신부
Ministry of Science and ICT

TTA 한국정보통신기술협회
Telecommunications Technology Association