Transactions of KSAE, Vol. 34, No. 4, pp.497-506 (April, 2026)

pISSN 1225-6382 / eISSN 2234-0149
DOI http://dx.doi.org/10.7467/KSAE.2026.34.4.497

End-to-End 자율주행 AI의 안전성 확보를 위한 기술 동향과 국제 안전 표준 관점의 분석

손 준 우[*1,2] · 박 명 옥[2] · 정 석 찬[3] · 김 성 희[4]

(주)소네트 자율주행연구팀[1] · 대구경북과학기술원 ABB연구부[2] · 동의대학교 인공지능학과[3] · 동의대학교 산업ICT기술공학과[4]

Analysis of Technology Trends and International Safety Standards for Ensuring the Safety of End-to-End Autonomous Driving AI

Joonwoo Son[*1,2] · Myoungouk Park[2] · Seok Chan Jeong[3] · Sung-Hee Kim[4]

[1]Autonomous Driving R&D Team, SonnetAI, 243 Techno Jungang-daero, Dalseong-gun, Daegu 43017, Korea
[2]Division of ABB, Daegu Gyeongbuk Institute of Science and Technology, Daegu 42988, Korea
[3]AI Grand ICT Research Center, Graduate School of Artificial Intelligence, Dong-eui University, Busan 47340, Korea
[4]Department of Industrial ICT Engineering, Graduate School of Artificial Intelligence, Dong-eui University, Busan 47340, Korea
(Received 27 February 2026 / Revised 2 March 2026 / Accepted 2 March 2026)

Abstract : End-to-end (E2E) learning-based autonomous driving has emerged as a promising paradigm that directly maps sensor inputs to vehicle control commands using data-driven models. Compared to conventional modular architectures, E2E approaches offer advantages in terms of architectural simplification and holistic learning of complex driving contexts from large-scale data. However, since E2E systems rely on probabilistic decision-making, exhibit limited explainability, and remain vulnerable to distribution shifts, edge cases, and long-tail scenarios, those benefits actually introduce fundamental challenges in safety assurance. This paper reviews recent technological developments in E2E autonomous driving and systematically analyzes key AI safety issues from a system-level perspective. Core challenges—including rare operational scenarios, uncertainty under out-of-distribution conditions, and limitations in traceability and accountability—are examined in relation to existing automotive safety frameworks. In particular, the paper investigates how ISO 21448 (Safety of the Intended Functionality, SOTIF) and UL 4600 can be reinterpreted and applied to learning-based autonomous driving systems to complement traditional failure-based functional safety standards. To address the structural mismatch between E2E architectures and existing safety standards, this paper discusses rule-based safety shields and scenario-based validation as practical and standard-compatible mechanisms for mitigating non-failure-based risks and constructing evidence-driven safety arguments. The analysis demonstrates that instead of E2E autonomous driving invalidating existing safety frameworks, it actually necessitates their complementary and systematic integration to achieve robust safety assurance in learning-based autonomous driving systems.

Key words : End-to-end autonomous driving(End-to-end 자율주행), AI safety(인공지능 안전성), Safety of the Intended Functionality (SOTIF)(의도된 기능 안전(SOTIF)), Safety shield(안전 보호 계층), Scenario-based validation(시나리오 기반 검증)

1. 서 론

자율주행 기술은 지난 10여 년간 인공지능, 센서(카메라, 라이다, 레이다), 정밀지도 및 고정밀 측위, 고성능 컴퓨팅 기술의 발전과 함께 빠르게 고도화되어 왔다. 산업적으로는 첨단 운전자 보조 시스템(Advanced Driver Assistance Systems, ADAS)에서 출발하여 조건부 자동화 및 고도 자동화 단계로 기술적 확장이 지속되고 있으며, 학계와 산업계 전반에서도 실제 도로 환경에서의 주행

*Corresponding author, E-mail: json@dgist.ac.kr

신뢰성과 안전성 확보를 핵심 경쟁력으로 인식하고 있다. 최근에는 단순히 자율주행 기능의 성능을 향상시키는 것을 넘어, 해당 기능이 실제 서비스 환경에서 안전하게 운영될 수 있음을 객관적이고 설득력 있게 입증하는 문제가 점차 중요한 과제로 부상하고 있다. 이는 자율주행 기술의 주요 논점이 더 이상 기능 성능 지표의 향상에 국한되지 않고, 시스템 차원의 '안전성 보증(Safety assurance)' 확보 및 검증체계 구축으로 확장되고 있음을 시사한다.[1)]

전통적으로 자율주행 시스템은 인지(Perception), 판단(Planning), 제어(Control)로 구성된 모듈형 아키텍처를 기반으로 발전해 왔다. 이러한 구조는 각 기능별 역할과 책임이 비교적 명확하다는 장점을 가지며, 고장 모드 및 영향 분석(Failure Mode and Effects Analysis, FMEA)과 같은 공학적 안전 분석 기법을 적용하기에 용이하다. 실제로 ISO 26262와 같은 기존 자동차 기능 안전 표준은 이러한 모듈형 구조를 전제로 고장 기반 위험 분석과 안전 메커니즘 설계 기준을 제시해 왔다.[2)] 그러나 자율주행 기능이 고도화되고 주행 환경의 복잡성이 증가함에 따라, 모듈 간 인터페이스의 복잡성 증가와 시스템 전체 관점에서의 최적화 한계가 지속적으로 지적되고 있다.

이러한 한계를 극복하기 위한 대안으로, 최근에는 센서 입력으로부터 차량 제어 명령을 직접 출력하는 End-to-End(E2E) 자율주행 구조가 주목받고 있다. E2E 자율주행은 초기의 단순한 조향각 예측 모델에서 출발하여[3)] 조건 기반 모방학습, 시계열 모델링, Transformer 기반 구조, 대규모 데이터 학습 등으로 빠르게 발전해 왔다.[4-7)] 이러한 접근은 주행 상황의 맥락 정보와 복합적인 교통 상호작용을 통합적으로 학습할 수 있다는 점에서 잠재적인 장점을 가진다. 그러나 동시에 E2E 자율주행은 학습 데이터 분포에 대한 높은 의존성, 예외 상황(Edge case)에 대한 취약성, 분포 이동(Distribution shift) 문제, 그리고 제한된 설명 가능성 등 해결해야 할 안전성 이슈를 내재하고 있다.[8-14)]

최근 AI Safety 관점에서 이러한 문제를 완화하기 위한 다양한 연구가 제안되고 있다. 대표적으로 불확실성(Uncertainty) 추정 및 OOD(Out-Of-Distribution) 감지, 시나리오 기반 검증, 시뮬레이션 기반 반례(Counterexample) 탐색, 규칙 기반 제약과의 결합, 그리고 학습 기반 정책과 독립적인 안전 보호 계층을 결합한 하이브리드 구조 등이 논의되고 있다.[15-20)] 그럼에도 불구하고, 다수의 선행 연구는 개별 기술 요소의 성능 또는 적용 가능성 검토에 초점을 맞추는 경향이 있으며, 해당 기술들이 국제 안전 표준이 요구하는 안전성 검증 및 인증 체계와 어떠한 방식으로 연계될 수 있는지에 대한 체계적인 분석은 상대적으로 부족한 실정이다.

특히 ISO 21448(Safety Of The Intended Functionality, SOTIF)[13)]과 UL 4600[21)]은 각각 고장이 없는 상태에서 발생하는 위험과 자율 시스템의 안전성 논증(Safety Case) 체계를 다루는 핵심 표준임에도 불구하고, E2E 자율주행 관련 AI 안전 기술과 해당 표준들을 체계적으로 매핑하고, 이를 안전성 관점에서 종합적으로 분석한 리뷰 연구는 부족한 실정이다. 이에 본 논문은 E2E 자율주행 기술에 대한 AI 안전성 이슈를 국제 안전 표준의 관점에서 재해석함으로써, 표준-기술 간 연결 구조를 명확히 제시하는 것을 목표로 한다. 이를 통해 자율주행 AI 안전성 논의를 성능 중심 접근에서 안전성 보증 중심 접근으로 확장하는데 기여하고자 한다.

2. 자율주행 시스템 구조

2.1 전통적 모듈형 자율주행 구조

전통적인 자율주행 시스템은 Fig. 1과 같이 인지(Perception), 판단(Planning), 제어(Control)로 구성된 모듈형(Modular) 구조를 기반으로 설계되어 왔다. 인지 모듈은 카메라, 라이다(LiDAR), 레이다(Radar) 등 다양한 센서로부터 수집된 데이터를 처리하여 주변 환경을 인식하며, 객체 검출, 차선 인식, 주행 가능 영역 추정, 상태 추정 등의 기능을 수행한다. 판단 모듈은 인지 결과를 기반으로 현재 주행 상황을 해석하고, 주행 목표에 부합하는 행동 전략을 결정한 뒤 경로(Path) 또는 궤적(Trajectory)을 계획한다. 제어 모듈은 계획된 궤적을 추종하기 위해 조향, 가속, 제동과 같은 저수준 제어 명령을 생성하며, 이를 차량의 물리적 동작으로 변환한다.

이와 같은 모듈형 구조는 기능 안전(Functional safety) 관점에서 비교적 명확한 이점을 제공한다. 각 모듈은 비교적 명확한 입력과 출력을 가지므로, 고장 모드 및 영향 분석(Failure Mode and Effects Analysis, FMEA), 고장 트리 분석(Fault Tree Analysis, FTA)과 같은 전통적인 안전 분석 기법의 적용이 용이하다. 또한 안전 요구사항을 모듈 단위로 할당하고 검증할 수 있어, 시스템의 안전성을 단계적으로 논증하는 접근이 가능하다. 실제로 ISO 26262는 이러한 모듈형 개발 체계를 전제로 고장 기반

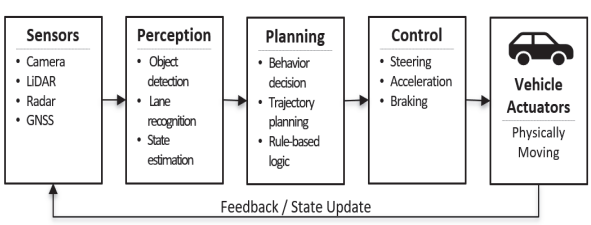

Fig. 1 Conventional modular autonomous driving architecture

위험 분석과 안전 메커니즘 설계를 규정하고 있으며, 오랜 기간 자동차 전장 시스템과 ADAS 개발에서 핵심적인 안전 프레임워크로 활용되어 왔다.[2)]

그러나 자율주행 기능이 고도화되고 주행 환경이 복잡해짐에 따라, 모듈형 구조의 한계도 점차 부각되고 있다. 모듈 간 인터페이스가 다층화됨에 따라, 인지 단계에서 발생한 경미한 오류 또는 불확실성이 판단 및 제어 단계로 전파되면서 시스템 수준에서 증폭될 가능성이 존재한다. 또한 실제 주행 데이터에는 인지・판단・제어 전 과정에 걸친 복합적인 상관관계가 내재되어 있으나, 모듈형 구조에서는 이를 단계별 처리과정에서 분리함으로써 충분히 활용하지 못할 가능성이 있다. 결과적으로 이러한 제약은 주행 상황의 맥락과 상호작용을 통합적으로 고려할 수 있는 새로운 자율주행 구조로의 전환 요구로 이어지고 있다.

2.2 End-to-End 자율주행 구조

End-to-End(E2E) 자율주행은 센서 입력으로부터 차량의 제어 명령(조향, 가감속 등)을 직접 출력하는 학습 기반 구조를 채택한다. 전통적인 인지-판단-제어의 단계를 명시적으로 분리하지 않고, 전체 주행 과정을 단일 정책(Policy)으로 모델링한다는 점에서 기존 모듈형 구조와 근본적으로 구별되는 접근이다. 이러한 개념은 초기에는 전방 카메라 영상으로부터 조향각을 직접 예측하는 비교적 단순한 형태로 제안되었으며,[3)] 자율주행을 지도 학습(Supervised learning) 기반의 정책 학습 문제로 정식화할 수 있음을 보여주었다는 점에서 중요한 출발점으로 평가된다.

이후 E2E 자율주행 연구는 단순한 반사적(Reflexive) 제어를 넘어, 보다 복잡한 주행 맥락을 반영하기 위한 방향으로 빠르게 확장되었다. 조건(Condition) 기반 모방학습은 주행 명령(예: 좌회전, 직진 등)을 입력으로 제공함으로써 동일한 환경에서도 주행 의도에 따른 상이한 행동을 학습할 수 있음을 보였으며,[4)] 이는 E2E 구조가 고수준 의사결정 요소를 내재화할 수 있음을 시사하였다. 또한 시계열 모델링, 순환 신경망(Recurrent neural networks), Attention 메커니즘의 도입은 과거 상태와 주변 교통 흐름을 고려한 연속적인 주행 결정을 가능하게 하였다.[5)]

최근에는 트랜스포머(Transformer) 기반 구조와 대규모 데이터 학습을 결합한 E2E 자율주행 접근이 주요 연구 흐름으로 부상하고 있다.[6,7)] 해당 방식은 센서 데이터를 토큰(Token) 형태로 임베딩한 뒤, 공간적・시간적 Self-attention을 통해 주행 장면을 통합적으로 표현한다. 이를 통해 다중 교통 주체 간의 상호작용, 장기적인 주행 맥락, 희소하지만 중요한 이벤트를 단일 모델 내에서 처리할 수 있는 가능성이 제시되고 있다. 특히 대규모 주행 로그와 결합될 경우, 다양한 환경 조건과 도로 유형에 대한 일반화 성능을 향상시킬 수 있다는 점에서 학계와 산업계 모두에서 높은 관심을 받고 있다.

E2E 자율주행의 핵심적인 기술적 장점은 주행 전반을 단일 최적화 문제로 다룰 수 있다는 점이다. 명시적인 규칙 기반 설계나 모듈 간 인터페이스 정의가 없어도 인간 운전자의 암묵적 판단과 다양한 변수 간 복합적 상관관계를 데이터로부터 직접 학습할 수 있다. 이는 시스템 구조의 단순화 뿐만 아니라, 수작업 규칙 설계 및 반복적 튜닝에 소요되는 비용을 절감할 수 있는 잠재력을 가진다. 또한 충분한 데이터와 연산 자원이 확보된 환경에서는 다양한 주행 시나리오에 대한 성능 개선이 비교적 신속하게 이루어질 수 있다는 점에서 E2E 접근의 중요한 장점으로 평가된다.

이러한 이유로 E2E 자율주행은 높은 기술적 완성도와 확장 가능성을 인정받으며, 차세대 자율주행 아키텍처의 유력한 후보로 자리 잡고 있다. 그러나 동시에, E2E 자율주행은 학습 기반 의사결정 특성으로 인해 기존 기능 안전 접근과는 상이한 위험 양상을 보일 수 있다. 따라서 E2E 자율주행을 실제 서비스 환경에 확대 적용하기 위해서는 이러한 기술적 장점을 유지하면서도 안전성을 체계적으로 확보할 수 있는 AI Safety 관점의 분석과 표준 기반 검증 전략의 수립이 필수적으로 요구된다.

3. End-to-End 자율주행의 주요 AI 안전성 이슈

3.1 Edge Case 및 Long-tail 문제

실제 도로 환경에서는 발생 빈도는 낮으나 사고로 직결될 수 있는 다양한 예외 상황(Edge case)이 존재한다. 예를 들어 보행자의 돌발 행동, 임시 공사 구간에서의 비정형적 주행 조건, 복수 교통 참여자 간의 복합 상호작용 등이 이에 해당한다. 이러한 상황은 통계적으로 희소하게 발생하기 때문에 학습 데이터에 충분히 포함되기 어렵고, 결과적으로 데이터 기반 학습에 의존하는 E2E 자율주행 시스템에서 취약 요인으로 작용한다.

E2E 자율주행 시스템은 학습 데이터 분포에 의존성이 높아 희소 상황(Long-tail)에서 성능 저하가 발생할 가능성이 높다. 특히 모델이 높은 신뢰도로 잘못된 행동을 출력하는 과잉 확신 오류(Overconfidence error) 현상은 단순한 성능 저하를 넘어 심각한 안전 문제로 이어질 수 있다. 기존의 주행 거리 기반 검증 방식만으로는 이러한 희소 이벤트의 발생 가능성과 위험성을 충분히 다루기 어렵다는 한계점도 지적되고 있다.[22)]

이러한 배경에서 최근에는 예외 상황을 체계적으로

식별하고 위험도가 높은 경계 조건을 중심으로 안전성을 평가하는 시나리오 기반 검증 접근의 필요성이 강조되고 있다. 즉, 평균적인 주행 성능 지표를 개선하는 것뿐만 아니라, 잠재적 위험이 내재된 경계 조건을 체계적으로 검증하는 프로세스가 E2E 자율주행 안전성 확보의 핵심 요소로 자리 잡고 있다.

3.2 분포 이동과 불확실성

E2E 자율주행 모델은 학습 데이터 분포와 실제 운행 환경 간의 일치성을 전제로 성능을 발휘한다. 그러나 국가별 교통 문화와 규범, 도로 구조의 차이, 기상 및 조명 조건의 변화 등은 불가피하게 분포 이동(Distribution shift)을 초래한다. 이는 모델의 인지 및 판단의 성능 저하로 이어질 수 있으며, 궁극적으로 실제 서비스 환경에서 예기치 못한 위험을 발생시킬 가능성을 내포한다.

특히 모델이 미학습 시나리오에 대해 과잉 확신을 기반으로 제어 명령을 출력하는 경우, 이는 단순한 성능 저하를 넘어 치명적인 사고로 직결될 수 있다. 따라서 E2E 자율주행에서는 모델 출력의 정확도뿐 아니라, 해당 출력 결과에 내재된 불확실성을 함께 평가하고 이를 보수적인 제어 행동 또는 회피 전략으로 연결하는 메커니즘이 필수적으로 요구된다.

이러한 요구를 충족하기 위한 핵심 기술 요소로는 불확실성 추정과 OOD(Out-Of-Distribution) 감지가 제시된다. 해당 기술들은 안전 보호 계층(Safety layer) 또는 fallback 전략과 결합됨으로써 분포 이동 상황에서 시스템 수준의 안전성을 강화하는 역할을 수행한다.[23)]

3.3 설명 가능성과 책임성 문제

E2E 자율주행은 내부 의사결정 과정이 명시적으로 분해되지 않는 기술적 특성을 가지며, 이러한 블랙박스적 요소는 사고 원인 분석 시 판단 근거를 추적하는 데 제약을 초래한다. 결과적으로 E2E 자율주행 시스템의 제한된 설명 가능성은 신뢰성 확보 측면에서 중요한 한계로 지적된다.[22)]

설명 가능성(Explainability)의 부족은 단순한 기술적 문제를 넘어, 책임성(Accountability) 및 제도적 수용성 측면에서 실질적인 장애 요인으로 작용한다. 실제 도로 환경에서 자율주행 서비스가 운영될 경우, 사고 발생 이후 시스템의 판단 근거와 위험요인을 규제 기관, 보험사, 서비스 운영자 및 사용자에게 설명하는 것은 필수적인 요구사항이다. 그러나 E2E 자율주행과 같이 학습 기반 정책이 직접 제어 명령을 생성하는 구조에서는, 특정 사고의 원인이 인지 오류, 판단 오류, 데이터 편향 또는 학습 과정에서의 한계 중 어떤 요인에 기인하는지 명확히 분리·구분하기가 어렵다. 이러한 특성은 사고 책임의 귀속과 법적 및 제도적 판단을 복잡하게 만들며, 사회적 수용성 측면에서도 중요한 도전 과제로 작용한다.[20,23)]

또한 설명 가능성의 문제는 안전성 검증 및 인증 절차에도 직접적인 영향을 미친다. 기존의 안전성 검증 체계는 시스템의 동작 원리와 위험 완화 메커니즘을 명시적으로 기술하고, 이에 기반한 검증 증거를 단계적으로 제시하는 방식을 전제로 한다. 그러나 E2E 자율주행 시스템은 내부 상태와 의사결정 경로가 명시적으로 드러나지 않기 때문에, 전통적인 개발 프로세스에서 요구되는 요구사항 추적성(Traceability) 확보에 구조적 제약이 따른다. 이로 인해 E2E 자율주행 시스템의 안전성 입증은 단일 기술 요소의 성능 검증만으로는 충분하지 않으며, 시스템 차원의 보완 전략과 함께 안전성 논증(Safety Case) 구조를 구성할 필요성이 제기된다.

이러한 배경에서 최근 연구들은 E2E 자율주행 모델의 의사결정 과정을 해석하기 위한 다양한 접근을 시도하고 있다. 예를 들어, 주의(Attention) 가중치 시각화, 중간 표현(Feature) 분석, 정책(Policy) 출력에 대한 민감도 분석 등은 학습 기반 모델의 의사결정을 부분적으로 해석하려는 시도로 제안되고 있다.[24)] 그러나 이러한 기법들은 주로 사후 분석(Post-hoc analysis)에 해당하며, 실제 안전성 검증이나 인증에 요구되는 수준의 결정론적 근거를 충족하기 어렵다는 한계가 존재한다.

결과적으로 E2E 자율주행 시스템의 설명 가능성 문제는 개별 모델의 기술적 해석 가능성 확보만으로 해결되기 어렵고, 시스템 수준에서의 안전성 논증 프레임워크와 결합되어야 한다. 이는 SOTIF 및 Safety Case 접근과 연계하여 해석될 수 있다.[13,21)] 즉, E2E 자율주행의 설명 가능성 한계는 개별 모델 투명성의 향상이라는 문제를 넘어, 표준 기반 안전성 논증 구조 내에서 어떻게 관리되고 보완될 수 있는지를 함께 고려해야 할 핵심 이슈로 볼 수 있다. 이와 관련하여 4장에서 표준-기술 매핑 및 Safety Case 구성 방안을 논의하고자 한다.

4. 기존 안전 표준과 E2E 자율주행 기술 간의 구조적 부적합성 분석

4.1 ISO 26262 적용의 한계

ISO 26262는 하드웨어 및 소프트웨어 고장을 중심으로 위험을 분석하고 이를 체계적으로 완화하기 위한 자동차 기능 안전(Functional safety) 표준이다.[2)] 이는 시스템 구성 요소의 고장(Failure)을 주요 위험 원인으로 가정하며, 고장 발생 가능성과 그 영향을 기반으로 안전 요구사항과 안전 메커니즘을 정의한다. 이러한 고장 기반 접근은 전통적인 모듈형 자율주행 시스템 또는 ADAS와

같이 기능이 명확히 분리된 시스템에 대해 효과적인 안전 분석 및 설계 프레임워크로 활용되어 왔다.

그러나 E2E 자율주행 시스템에서 나타나는 위험 특성은 ISO 26262의 기능 안전 가정과 구조적으로 상이할 수 있다. 즉, E2E 자율주행에서 발생하는 주요 위험 요인은 하드웨어 고장이나 명시적인 소프트웨어 결함보다는 학습 데이터의 한계, 분포 이동(Distribution shift), 희소 상황(Long-tail)에 대한 일반화 실패 등 학습 기반 의사결정의 불완전성에서 기인하는 경우가 많다. 이러한 위험은 개별 구성 요소가 정상적으로 동작하는 상태에서도 발생할 수 있으며, 전통적 '고장' 개념으로 환원하기 어렵다.

또한, E2E 자율주행은 센서 입력부터 제어 출력까지를 단일 정책(Policy)으로 통합하여 학습하는 구조적 특성으로 인해, 기능 단위의 명확한 구분 및 요구사항 추적성(Traceability) 확보가 제한될 수 있다. 이로 인해 특정 주행 오류를 단일 구성 요소의 고장으로 귀속시키거나, 고장 원인과 결과를 명시적으로 연결하는 ISO 26262 기반의 안전 분석을 시스템 전체에 적용하는 데에는 근본적인 한계가 존재한다. 다만 명시적 고장 가능성이 존재하는 구성 요소에 대해서는 ISO 26262의 적용이 유효하다.

이와 같이 ISO 26262는 구성 요소 수준의 기능 안전 확보에는 유효하나, 학습 기반 의사결정에서 발생하는 비고장(Non-failure) 기반 위험을 충분히 포괄하지 못하는 한계가 있다. 따라서, E2E 자율주행의 안전성 확보를 위해서는 고장 기반 기능 안전 접근을 보완할 수 있는 확장된 안전성 프레임워크가 요구되며, 이는 고장이 없는 상태에서의 위험을 다루는 ISO 21448(SOTIF)과, 시스템 차원의 안전성을 논증 구조로 제시하는 UL 4600을 통해 실현 될 수 있다.[13,21]

4.2 ISO 21448(SOTIF) 관점에서의 해석

ISO 21448(SOTIF)는 시스템 구성 요소의 고장이 존재하지 않더라도 발생할 수 있는 위험을 다루는 안전 표준으로, 기능 불충분(Functional insufficiency)과 이를 촉발하는 조건(Triggering condition)을 핵심 개념으로 정의한다.[13] 이러한 관점은 E2E 자율주행 시스템에서 나타나는 위험 특성을 해석하는데 특히 적합하다. E2E의 위험은 전통적인 고장보다는 특정 조건하의 성능 불충분에 기인하며, 이는 SOTIF의 기능 불충분 개념으로 적절히 해석될 수 있다.

Table 1은 E2E 자율주행의 대표적인 안전 기술 요소들을 ISO 21448 및 UL 4600의 주요 개념과 대응시키고, 각 기술의 안전적 함의를 정리한 것이다. Table 1에서 나타난 바와 같이, 학습 기반 E2E 정책은 평균적인 주행 상황에서는 높은 성능을 보일 수 있으나, 예외 상황(Edge case)나 희소 상황(Long-tail)에서는 주행 맥락에 대한 해석과 대응 능력이 결여될 가능성이 존재한다. 이러한 상황들은 SOTIF 관점에서 기능 불충분을 촉발하는 조건으로 해석될 수 있으며, 결과적으로 안전성 검증은 평균적인 성능 지표 중심의 검증을 넘어 경계 조건(Boundary condition) 및 희소 상황 중심의 검증 전략으로 확장될 필요가 있음을 의미한다.

Table 1 Mapping of key E2E safety technologies to ISO 21448 and UL 4600

Technology / Concept	ISO 21448 (SOTIF)	UL 4600	Key implication
End-to-end policy	Functional insufficiency (FI) (Clause 6: Hazard identification)	Safety-critical element (Section 8: Autonomous system self-awareness)	Learning-based decision-making is a primary source of hazards without component failure
Edge case / Long-tail	Triggering condition (TC) (Clause 7: Identification of triggering conditions)	Hazardous scenario (Section 6: Risk assessment)	Verification & Validation (V&V) must focus on boundary and rare operational conditions
Distribution shift	FI-inducing condition (Clause 7: Analysis of functional insufficiencies)	ODD boundary risk (Section 8.2: ODD monitoring)	Explicit definition and management of the ODD are essential
Rule-based safety shield	Risk reduction measures (Clause 9: SOTIF strategy & mitigation)	Independent safety mechanism (Section 8.1: Functional safety)	Deterministic safety layers complement probabilistic policies
Uncertainty gating	FI awareness (Clause 7 & 10: Insufficiency identification/Verification)	Risk mitigation (Section 8: Self-awareness & monitoring)	Awareness of model uncertainty is a safety-relevant function
MRM / Fallback	Risk reduction (Clause 12: Validation of SOTIF)	Fallback strategy (Section 9: Operational states & MRM)	Capability for safe failure is a key safety requirement

또한 분포 이동(Distribution shift)은 E2E 정책의 성능을 저하시켜 기능 불충분을 유발하는 대표적 조건으로 간주될 수 있다. 학습 환경과 실제 운행 환경 간의 차이는 의도된 기능의 안전한 수행을 저해할 수 있으며, 이는 SOTIF에서 정의하는 위험 발생 메커니즘과 직접적으로 연결된다. 이러한 관점에서 운행설계영역(ODD)의 명확한 정의와 관리, 그리고 ODD 경계에서의 위험 식별 및 평가는 E2E 자율주행 안전성 확보에 있어 필수적인 요소로 부각된다.

더 나아가 불확실성 인식(Uncertainty awareness) 및 OOD(Out-Of-Distribution) 감지는 기능 불충분 상태를 사전에 감지하고 위험을 완화하기 위한 핵심 기술 요소로 해석될 수 있다. 즉, 모델 출력의 신뢰도를 정량적으로 평가하고, 불확실성이 높은 구간에서는 보수적인 행동 정책으로 전환하거나 안전 보호 계층(Safety layer)을 활성화하는 메커니즘은 SOTIF 관점에서 기능 불충분 위험을 감소시키는 주요 수단으로 기능한다.

따라서 ISO 21448(SOTIF)은 E2E 자율주행 시스템에서 발생하는 예외 상황, 분포 이동, 불확실성 문제를 고장 기반이 아닌 기능적 관점에서 구조화하여 해석할 수 있는 체계적인 틀을 제공한다. 이는 E2E 자율주행 안전성 논의를 단순한 성능 개선 차원을 넘어, 의도된 기능의 안전한 수행을 보증하기 위한 검증 및 완화 전략으로 확장하는데 핵심적인 역할을 한다.

4.3 UL 4600과 Safety Case 접근

UL 4600은 특정 기술 또는 아키텍처를 강제하지 않는 기술 중립적(Technology-neutral) 안전 표준으로, 자율 시스템이 주어진 운행 설계 영역(ODD) 내에서 안전하다는 주장을 논리적 구조와 검증 가능한 증거를 통해 입증할 것을 요구한다.[21] 이러한 Safety Case 기반 접근은 내부 동작이 명시적으로 구분되지 않는 E2E 자율주행 시스템의 특성에 부합하는 효과적인 프레임워크로 평가된다.

UL 4600 관점에서 학습 기반 E2E 정책은 안전에 중대한 영향을 미치는 핵심 요소(Safety-critical element)로 간주되며, 해당 정책이 어떤 가정 하에서 안전하게 동작하는지에 대한 명확한 주장과 이를 뒷받침하는 근거의 제시가 요구된다. 또한 Table 1에서 정리한 위험 시나리오(Hazardous scenario) 개념은 E2E 자율주행에서 문제로 지적되는 예외 상황(Edge case) 및 희소 상황(Long-tail)을 구조적으로 다루기 위한 분석 도구로 활용될 수 있다. 특히 시나리오 기반 검증 결과는 Safety Case를 구성하는 핵심적인 증거로 포함될 수 있으며, 정책의 성능 한계와 위험 노출 조건을 체계적으로 제시하는 역할을 수행한다.

또한 UL 4600은 학습 기반 정책의 불확실성을 보완하기 위한 독립적인 안전 메커니즘(Independent safety mechanism)의 중요성을 강조한다. 예를 들어 규칙기반 안전 보호 계층(Rule-based safety shield)과 같은 결정론적(Deterministic) 보호 계층은 확률적 특성을 갖는 학습 기반 정책의 잠재적 실패를 구조적으로 완화할 수 있으며, 이는 UL 4600에서 요구하는 위험 경감(Risk mitigation) 전략으로 해석될 수 있다. 더 나아가 불확실성 기반 제어 제한(Uncertainty gating)은 위험 수준에 따른 시스템 동작을 조절할 수 있는 수단을 제공하며, 최소 위험 상태(Minimum Risk Maneuver, MRM) 또는 대응(Fallback) 전략은 안전한 실패(Safe failure)를 보장하기 위한 핵심 요소로 Safety Case에 포함될 수 있다.

결과적으로 UL 4600은 E2E 자율주행의 안전성을 개별 기술 요소의 완전성으로 입증하기보다 정책의 한계, 위험 시나리오, 위험 완화 전략, 그리고 검증 결과를 유기적으로 연결하여 안전성을 논증할 수 있는 구조를 제공한다. 이는 ISO 26262와 ISO 21448(SOTIF)이 단독으로는 충분히 포괄하기 어려운 시스템 수준의 안전성 보증을 보완하며, E2E 자율주행을 실제 서비스 환경에 적용하기 위한 실질적인 논증 프레임워크로 기능한다.

이상을 종합하면, ISO 26262, ISO 21448(SOTIF), UL 4600은 자율주행 시스템–특히 End-to-End(E2E) 자율주행–의 안전성을 다층적으로 설명하고 확보하기 위해 상호보완적으로 활용되어야 하는 프레임워크로 이해될 수 있다. ISO 26262는 센서, 액추에이터, 통신 인터페이스 등 명시적인 고장이 발생 가능한 구성 요소 수준에서 기능 안전을 확보하는 데 핵심적인 역할을 수행한다. 반면, 학습 기반 의사결정의 불완전성, 분포 이동(Distribution shift), 희소 상황(Long-tail)과 같이 고장이 없는 상태에서도 발생 가능한 위험은 ISO 21448에서 정의하는 기능 불충분(Functional insufficiency)과 촉발 조건(Triggering condition) 개념을 통해 적절히 구조화될 수 있다.

한편 UL 4600은 특정 위험 유형이나 기술 요소를 개별적으로 규정하기보다는, 시스템이 주어진 운행 설계 영역(ODD) 내에서 안전하다는 주장을 어떠한 논리 구조로 구성하고, 어떤 검증 가능한 증거를 통해 이를 입증할 것인지에 초점을 두는 Safety Case 중심 접근을 제공한다. 이러한 접근은 내부 구조가 명시적으로 구분되지 않는 E2E 자율주행 시스템의 특성과 부합하며, SOTIF 관점에서 식별된 위험 요소와 ISO 26262 기반의 기능 안전 활동, 그리고 다양한 기술적 위험 완화 수단을 하나의 논증 구조로 통합하는 역할을 수행한다. 결과적으로, Fig. 2와 같이, E2E 자율주행의 안전성 확보는 단일 표준의 적용만으로 해결될 수 있는 문제가 아니라, 고장 기반 기능 안전, 비고장 기반 위험 해석, 그리고 시스템 수준 안전

Fig. 2 E2E autonomous driving safety assurance cycle

성 논증을 결합한 상호 보완적 프레임워크(Complementary framework) 을 통해 실현될 수 있을 것이다.

5. 기존 안전 표준과 E2E 자율주행 기술의 보완적 대응 방안

4장의 구조적 부적합성 분석을 토대로, 본 장에서는 학습 기반 정책의 장점은 유지하되 독립적인 안전 기능을 보완하는 다층적 안전 아키텍처를 설계하는 실무적 대안을 제시하고자 한다. 특히 명시적 기능 분해가 어려운 E2E 구조에서는 위험을 특정 구성 요소의 고장으로 귀속시키거나 고장 원인과 결과를 단계적으로 연결하는 전통적 안전 분석 방식이 제한될 수 있다. 또한 평균적 주행 성능과 별개로 희소 상황이나 분포 이동 환경에서 예측 가능성이 저하될 수 있으며, 이는 시스템 차원의 안전성 확보에 있어 핵심 위험 요인으로 작용한다. 이러한 특성은 E2E 자율주행의 안전성 확보가 단일 표준이나 개별 기술 요소로만 해결되기 어렵다는 점을 시사한다.

이러한 구조적 제약을 완화하기 위해 본 장에서는 학습 기반 정책의 효율성과 규칙 기반 메커니즘의 견고함을 결합한 다층적 안전 아키텍처와 시나리오 기반 검증 방안을 제안한다. 구체적으로 규칙 기반 안전 보호 계층(Rule-based safety shield)과 시나리오 기반 검증(Scenario-based validation)을 중심으로, E2E 자율주행 기술이 기존 안전 표준과 정합성을 확보할 수 있는 보완적 경로를 제시하고자 한다.

5.1 구조적 안전성 보완을 위한 규칙 기반 안전 보호 계층

규칙 기반(Rule-based) 안전 보호 계층은 학습 기반 E2E 자율주행 정책을 대체하는 것이 아니라, 해당 정책의 출력을 감시하고 제약하는 독립적인 안전 메커니즘으로 설계된다. 일반적으로 안전 보호 계층은 E2E 정책이 생성한 행동을 실시간으로 평가하여, 물리적 한계 초과, 교통 법규 위반, 충돌 위험 증가 등 명확히 정의 가능한 안전 제약 조건을 초과하는 경우 해당 행동을 차단하거나 수정한다. 이러한 개념은 강화학습 분야에서 제안된 Shielding 접근법에서도 확인되며, 학습 정책의 탐색 자유도를 유지하면서도 안전 사양 위반을 방지하는 구조로 활용되어 왔다.[25)]

이러한 규칙 기반 안전 보호 계층의 핵심 장점은 학습 기반 정책의 불확실성과 확률적 실패 가능성을 결정론적(Deterministic) 규칙으로 보완할 수 있다는 점이다. 예를 들어, 차량 동역학적 한계, 최소 안전 거리, 교통 법규 등과 같이 명확히 정의 가능한 제약 조건은 학습 모델의 예측 오류와 무관하게 결정론적으로 적용될 수 있으며, 시스템 수준에서의 안전 제약을 일관되게 유지하는 기반이 된다. 또한 불확실성 추정 결과를 기반으로 보수적인 행동을 강제하거나, 최소 위험 상태(MRM)로 전환하는 로직을 포함함으로써, E2E 정책의 신뢰성이 낮은 상황에서도 시스템 수준의 안전성을 유지할 수 있다.

표준 관점에서 볼 때, 규칙 기반 안전 보호 계층은 ISO 21448(SOTIF)에서 요구하는 기능 불충분(Functional insufficiency) 및 촉발 조건(Triggering condition) 완화 수단으로 해석될 수 있으며, UL 4600에서는 학습 기반 정책과 독립적인 위험 완화 메커니즘으로서 중요한 Safety Case를 구성하는 핵심적 논증 근거로 활용될 수 있다. 즉, 안전 보호 계층은 E2E 자율주행이 기존 안전 표준과 양립할 수 있도록 하는 구조적 연결 고리 역할을 수행하며, 학습 기반 시스템은 본질적으로 안전하지 않다는 비판에 대응할 수 있는 실효성 있는 방안을 제시한다.

5.2 증거 기반 안전성 보증을 위한 시나리오 기반 검증

규칙 기반 안전 보호 계층이 위험을 구조적으로 완화하는 수단이라면, 시나리오 기반 검증은 이러한 완화 전략이 실제로 유효함을 입증하기 위한 핵심 방법론이다. E2E 자율주행의 안전성은 평균적인 주행 성능지표나 누적 주행 거리만으로 설득력 있게 입증되기 어렵다. 특히 사고로 직결될 수 있는 위험은 대부분 경계 조건이나 희소 상황에서 발생하기 때문에, 안전성 검증 역시 이러한 상황을 중심으로 설계되어야 한다.

PEGASUS 프로젝트를 비롯한 시나리오 기반 검증 연

구는 실제 도로 환경에서 발생 가능한 위험 상황을 체계적으로 식별하고 분류한 뒤, 이를 기반으로 검증 시나리오를 구성하는 접근법을 제시하였다.[26,27] 해당 방식은 무작위 주행 거리 기반 검증의 비효율성을 보완하고, 제한된 시험 자원으로도 위험 밀집도(Risk density)가 높은 조건에서의 안전성을 평가할 수 있도록 한다. 특히 E2E 자율주행에서는 예외 상황 및 희소 상황 시나리오가 주요 위험 원인으로 작용할 수 있으므로, 시뮬레이션 기반 시나리오 검증은 필수불가결한 검증 전략으로 간주 될 수 있다.[28]

UL 4600의 Safety Case 관점에서 시나리오 기반 검증은 단순한 시험 결과를 넘어, 시스템이 어떠한 가정과 조건 하에서 안전한지를 명확히 설명하는 구조화된 증거로 활용될 수 있다. 예를 들어, 특정 ODD 경계 조건에서 E2E 정책의 출력이 어떻게 제한되며, 안전 보호 계층이 어떤 방식으로 개입하는지를 시나리오 단위로 제시함으로써, 안전성 논증의 객관성과 설득력을 크게 향상시킬 수 있다. 이는 규제 기관 및 보험사 등 다양한 이해관계자와의 신뢰 구축을 위한 핵심적 역할을 수행하며, 결과적으로 E2E 자율주행 시스템의 실질적인 상용화 가능성을 높이는 토대가 된다.

6. 결 론

본 논문은 E2E 자율주행 기술 동향을 AI Safety 관점에서 분석하고, 기존 자동차 안전 표준과의 구조적 부적합성을 해소하기 위한 다층적 보완 방안을 제시하였다. 자율주행 기술의 패러다임이 성능 중심에서 서비스 차원의 안전성 보증(Safety assurance)으로 전환되고 있다는 점에 주목하여, E2E 구조의 특성상 발생하는 안전 이슈를 체계적으로 분석하고, 이를 기존 안전 표준과 결합하는 다층적 대응 방안을 제안하였다.

이에 본 논문은 ISO 21448(SOTIF)의 기능 불충분(Functional insufficiency) 및 촉발 조건(Triggering condition) 개념을 통해 E2E 위험을 기능적 관점에서 구조화할 수 있음을 보였으며, UL 4600의 Safety Case 접근이 시스템 수준의 안전성 주장을 논증과 증거 기반으로 구성하는 데 효과적임을 확인하였다. 또한 규칙 기반 안전 보호 계층, 불확실성 기반 제어 제한, MRM/Fallback 전략, 시나리오 기반 검증 등 E2E 안전 기술 요소를 표준의 핵심 개념 및 요구사항에 매핑함으로써 표준-기술 간 연결 구조를 명확히 하였다. 마지막으로, 5장에서 규칙 기반 안전 보호 계층을 통한 결정론적 안전 보완과 경계 조건 중심 시나리오 기반 검증을 통한 증거 기반 안전성 논증을 E2E 자율주행의 현실적인 보완 전략으로 제안하였다.

종합적으로 본 논문은 E2E 자율주행을 기존 표준과 대립시키기보다는 기능 안전, 비고장 기반 위험 관리(SOTIF), Safety Case 기반 안전성 논증을 결합하는 다층적 접근을 통해 안전성 확보가 필요함을 제시하였다.

6.1 한계점 및 향후 연구

본 연구는 E2E 자율주행의 안전성 이슈를 AI safety 관점에서 정리하고, 국제 안전 표준 기반의 개념적 프레임워크를 제시하는 리뷰 연구에 초점을 두었다는 점에서 몇 가지 한계를 가진다. 첫째, 표준 조항과 기술 요소 간 매핑은 주로 논리적 정합성과 개념적 대응 관계를 중심으로 제시되었으며, 실제 개발 프로젝트에서의 적용 사례나 정량적 효과(예: 잔여 위험 수준, 안전성 향상 폭 등)를 실증적으로 제시하는 데에는 한계가 있었다. 둘째, 설명 가능성과 책임성의 중요성을 논의하였으나, 인증 수준에서 요구되는 추적성(Traceability) 확보와 설명 증거(Evidence)의 구성 방식을 표준 친화적으로 정립하기 위한 구체적 절차와 방법론 제안은 향후 과제로 남겨 두었다. 셋째, 시나리오 기반 검증의 필요성을 강조하였으나, 시나리오 커버리지의 정량화, 시뮬레이션-시험장-실도로 간 상호 검증, 그리고 ODD 경계 관리의 운영 절차에 대한 논의는 후속 연구를 통해 보완될 필요가 있다.

향후 연구에서는 실제 운행 데이터와 시험 결과를 기반으로 SOTIF 위험 분석과 UL 4600의 Safety Case를 통합하는 실증 연구가 필요하며, 규칙 기반 안전 보호 계층(Rule-based safety shield)과 불확실성 기반 제어 제한(Uncertainty gating)을 포함한 안전 아키텍처의 설계·검증 방법론을 보다 구체화함으로써 학습 기반 정책의 한계를 체계적으로 보완할 수 있는 실효성 있는 안전 가이드라인을 정립할 필요가 있다. 더 나아가 다양한 지역 및 환경 변화에 따른 분포 이동을 고려한 ODD 정의 및 갱신 전략, 장기 운영 관점에서의 안전성 증거 축적과 갱신 프로세스를 체계화함으로써, E2E 자율주행의 안전성 보증 체계를 보다 실천적이고 지속 가능한 방향으로 발전시킬 필요가 있다.

후 기

본 연구는 산업통산자원부 한국산업기술기획평가원 자율주행기술개발혁신사업(과제명: 지정구역기반 Point to point 이동 Lv.4 승합차급 자율주행 차량플랫폼 기술개발, 과제번호: 20014361)의 연구비 지원 및 과학기술정보통신부 대구경북과학기술원 기관고유사업(과제번호: 26-IT-03)의 연구비지원에 의해 수행되었습니다.

References

1) P. Koopman and M. Wagner, “Toward a Framework for Highly Automated Vehicle Safety Validation,” SAE 2018-01-1071, 2018.
2) International Organization for Standardization, Road Vehicles — Functional Safety, ISO 26262, 2018.
3) M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, K. J. Zhao and K. Zieba, “End to End Learning for Self-Driving Cars,” arXiv preprint arXiv:1604.07316, 2016.
4) F. Codevilla, M. Müller, A. López, V. Koltun and A. Dosovitskiy, “End-to-End Driving via Conditional Imitation Learning,” IEEE International Conference on Robotics and Automation (ICRA), pp.4693-4700, 2018.
5) K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz and A. Geiger, “TransFuser: Imitation with Transformer-Based Sensor Fusion for Autonomous Driving,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.45, No.11, pp.12878-12895, 2023.
6) J. Hwang, R. Xu, H. Lin, W. C. Hung, J. Ji, K. Choi, D. Huang, T. He, P. Covington, B. Sapp, Y. Zhou, J. Guo, D. Anguelov and M. Tan, “Emma: End-to-End Multimodal Model for Autonomous Driving,” arXiv preprint arXiv:2410.23262, 2024.
7) Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, L. Lu, X. Jia, Q. Liu, J. Dai, Y. Qiao and H. Li, “Planning-Oriented Autonomous Driving,” Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pp.17853-17862, 2023.
8) K. Chitta, A. Prakash and A. Geiger, “Neat: Neural Attention Fields for End-to-End Autonomous Driving,” IEEE/CVF International Conference on Computer Vision (ICCV), pp.15793-15803, 2021.
9) L. Chen, X. Zhu, J. Dai, Y. Qiao and H. Li, “End-to-End Autonomous Driving: Challenges and Frontiers,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.
10) M. Borg, C. Englund, K. Wnuk, B. Duran, C. Levandowski, S. Gao, Y. Tan, H. Kaijser, H. Lönn and J. Tornqvist, “Safely Entering the Deep: A Review of Verification and Validation for Machine Learning and a Challenge Elicitation in the Automotive Industry,” Journal of Automotive Software Engineering, Vol.1, No.1, pp.1-19, 2020.
11) F. Codevilla, E. Santana, A. M. López and A. Gaidon, “Exploring the Limitations of Behavior Cloning for Autonomous Driving,” IEEE/CVF International Conference on Computer Vision (ICCV), pp.9329-9338, 2019.
12) M. Arjovsky, L. Bottou, I. Gulrajani and D. Lopez-Paz, “Invariant Risk Minimization,” arXiv preprint arXiv:1907.02893, 2019.
13) International Organization for Standardization, Road Vehicles — Safety of the Intended Functionality, ISO 21448, 2022.
14) É. Zablocki, H. Ben-Younes, P. Pérez and M. Cord, “Explainability of Deep Vision-Based Autonomous Driving Systems: Review and Challenges,” International Journal of Computer Vision, Vol.130, No.10, pp.2425-2452, 2022.
15) D. Bogdoll, M. Nitsche and J. M. Zöllner, “Anomaly Detection in Autonomous Driving: A Survey,” IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.4488-4499, 2022.
16) B. Jiang, S. Chen, Q. Xu, B. Liao, J. Chen, H. Zhou, Q. Zhang, W. Liu, C. Huang and X. Wang, “VAD: Vectorized Scene Representation for Efficient Autonomous Driving,” IEEE/CVF International Conference on Computer Vision (ICCV), pp.8340-8350, 2023.
17) N. Hanselmann, K. Renz, K. Chitta, A. Bhattacharyya and A. Geiger, “KING: Generating Safety-Critical Driving Scenarios for Robust Imitation via Kinematics Gradients,” European Conference on Computer Vision (ECCV), pp.335-352, 2022.
18) S. Shalev-Shwartz, S. Shammah and A. Shashua, “On a Formal Model of Safe and Scalable Self-Driving Cars,” arXiv preprint arXiv:1708.06374, 2017.
19) N. Webb, D. Smith, C. Ludwick, T. Victor, Q. Hommes, F. Favaro, G. Ivanov and T. Daniel, “Waymo’s Safety Methodologies and Safety Readiness Determinations,” arXiv preprint arXiv:2011.00054, 2020.
20) W. Zeng, W. Luo, S. Suo, A. Sadat, B. Yang, S. Casas and R. Urtasun, “End-to-End Interpretable Neural Motion Planner,” IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.8660-8669, 2019.
21) UL Standards and Engagement, ANSI/UL 4600: Standard for Safety for the Evaluation of Autonomous Products, 2023.
22) N. Kalra and S. M. Paddock, “Driving to Safety:

How Many Miles of Driving Would It Take to Demonstrate Autonomous Vehicle Reliability?" Transportation Research Part A: Policy and Practice, Vol.94, pp.182-193, 2016.
23) R. Michelmore, M. Kwiatkowska and Y. Gal, "Evaluating Uncertainty Quantification in End-to-End Autonomous Driving Control," arXiv preprint arXiv:1811.06817, 2018.
24) S. Jain and B. C. Wallace, "Attention Is Not Explanation," arXiv preprint arXiv:1902.10186, 2019.
25) M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Niekum and U. Topcu, "Safe Reinforcement Learning via Shielding," Proceedings of the AAAI Conference on Artificial Intelligence, Vol.32, No.1, pp.2669-2678, 2018.
26) PEGASUS Project, "Scenario-Based Safety Validation of Automated Driving," German Federal Ministry for Economic Affairs and Energy, Project Report, 2019.
27) M. Park and J. Son, "Reference Test Scenarios for Assessing the Safety of Take-Over in a Conditionally Autonomous Vehicle," Transactions of KSAE, Vol.27, No.4, pp.309-317, 2019.
28) H. Kim, G. Jo and J. Son, "Implementation and Verification of Virtual Environment for Autonomous Driving System Development," Transactions of KSAE, Vol.29, No.4, pp.331-336, 2021.