FOCUS 2

신뢰할 수 있는 AI: 글로벌 현황과 향후 정책적 과제

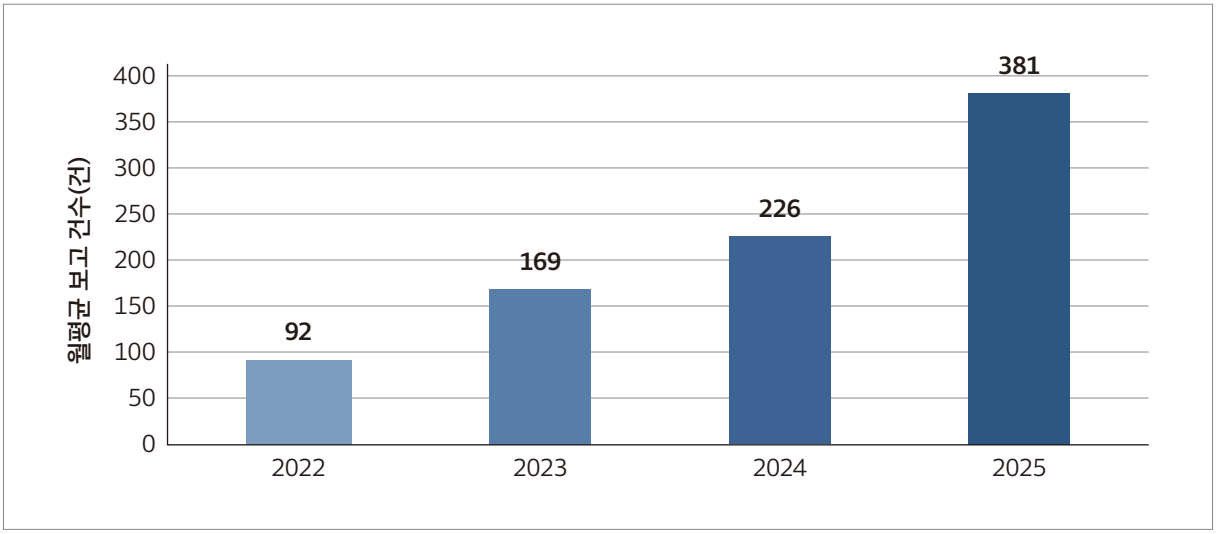
- **김종원** | 슈어소프트테크 AIV Lab 팀장 | bk3206@suresofttech.com
- **김영현** | 슈어소프트테크 AX 센터 팀장 | yhkim1@suresofttech.com
- **정재룡** | 슈어소프트테크 AX 센터 실장 | jyjung@suresofttech.com

서론: "AI 중심 사회가 마주한 새로운 도전: 신뢰할 수 있는 AI"

인공지능(Artificial Intelligence, AI)은 단순한 기술적 혁신을 넘어 현대 사회의 기반 인프라로 안착하고 있다. OpenAI가 개발한 ChatGPT의 주간 활성 사용자 수(WAU)가 9억 명을 돌파한 사례에서 보듯[1], AI는 이제 개인의 생산성 도구나 기업의 사무 보조를 넘어 의료기기, 자율주행, 로보틱스 등 물리 시스템과 국가 핵심 인프라에도 깊숙이 이식되고 있다. 이는 마치 전력망이나 통신 인프라처럼 사회 시스템 전체가 AI의 정상 작동을 전제로 작동하는 시대에 도달했음을 의미한다.

그러나 이러한 기술적 확산 속도가 AI의 '안전성'을 담보하는 것은 아니다. OECD AI 사고 모니터(AIM)[1]의 집계에 의하면, AI 관련 사고는 2022년 월평균 92건에서 2025년 381건으로 3년 만에 약 4.1배 증가했다.[2]

[그림 1] OECD AI 사고 모니터의 최근 4년간 월평균 AI 관련 사고 건수



AI 시스템의 불확실성은 기업과 사회에 실질적이고 막대한 손실을 야기할 수 있다. 2023년 구글의 챗봇 Bard의 오답 논란으로[2] 모기업 알파벳의 시가총액 1,000억 달러가 하루 만에 증발했고[3], 2024년에는

1 OECD AI 사고 모니터(AIM, AI Incidents and Hazards Monitor)는 전 세계 언론 보도를 자동으로 수집·분석해 AI 관련 사안을 집계하는 OECD의 모니터링 체계이다.

2 2023년 2월 구글이 공개한 Bard 시연에서 챗봇이 "제임스 웹 우주망원경이 태양계 밖 행성을 최초로 촬영했다"고 답했으나, 외계행성 최초 촬영은 2004년 유럽남방천문대의 초대형망원경(VLT)이 수행한 것이었다. 이 여파로 알파벳 주가는 하루 만에 7.6% 하락했다.

딥페이크 기술을 악용한 화상회의 사기로[3] 2,500만 달러가 유출됐다.[4] 물리 시스템 영역인 자율주행에서는 인명 피해까지 발생했는데, 테슬라의 오토파일럿과 자율주행(FSD, Full Self-Driving) 기능 관련 사망자는 2016년 첫 사망사고 이후 2025년 10월까지 65명에 이르고 있다.[4][5][6]

이러한 사고들은 일시적/우발적 결함이 아닌, AI 기술 특성에 기인한 구조적 위험이라는 사실에 주목할 필요가 있다. 데이터 기반의 확률적 추론에 의존하는 AI 알고리즘의 특성상, 특정 출력에 대한 원인을 명확히 설명하기 어려울 뿐 아니라 오류를 수정했다고 해서 오류가 완벽히 제거되었는지 증명하는 것 역시 까다롭다.

이런 한계를 극복하기 위해 국제 표준화 기구는 '신뢰할 수 있는 AI(Trustworthy AI)'라는 개념을 제시했다. ISO/IEC TR 24028:2020에서는 신뢰가능성(Trustworthiness)을 "이해관계자의 기대를 검증 가능한 방식으로 충족하는 능력"으로 정의한다.[7] EU AI Act와 한국 AI 기본법 등 주요 국가의 법제는 고위험·고영향 AI에 대해 안전성 확보를 법적 의무로 명시하여 강력한 증명을 요구하고 있다.[8][9]

그러나, 법적·제도적 규제 요구와 이를 입증할 객관적 검증 수단 사이에는 여전히 격차가 존재한다. 법적 당위성은 마련되었으나, 과학적 검증 도구가 부재한 이 간극이 오늘날 AI 중심 사회가 마주한 본질적 도전이다.

글로벌 AI 신뢰성 및 안전 기술 동향

앞서 언급한 ISO/IEC TR 24028 외에도, 미국 NIST의 AI Risk Management Framework, EU의 Ethics Guidelines for Trustworthy AI 등은 안전성, 강건성, 보안성, 투명성, 책무성 등 AI 신뢰성을 구성하는 다양한 요소를 제시하고 있다.[10][11] 이러한 요소를 검증의 관점에서 재구성하면 기능성 검증, 강건성 검증, 보안성 검증, 운영·관리 영역으로 구분할 수 있는데, 운영·관리 영역은 시스템의 구성·배포·모니터링 등 전통적인 소프트웨어 검증에서도 다뤄지는 공통 영역이므로, 본문에서는 AI 모델 자체의 동작과 성능을 직접 시험하는 기능성·강건성·보안성 검증(EU AI Act 제15조 대응 영역)[5]에 초점을 맞춘다.[8]

3 2024년 1월 영국 엔지니어링 기업 아룹(Arup)의 홍콩 지사 직원이 본사 최고재무책임자(CFO)와 동료들로 위장한 화상회의에 참여한 뒤 홍콩 소재 5개 계좌로 15차례에 걸쳐 송금한 사건이다. 회의에 등장한 얼굴과 음성은 모두 공개된 임원 영상·음성으로 만든 딥페이크였다.

4 Tesla Deaths는 언론에 보도된 테슬라 관련 사망사고를 집계하는 민간 데이터베이스로, 2025년 10월 기준 전체 772건 가운데 오토파일럿이 관여한 사망자를 65명으로 집계했다. 이 중 2건은 미국 도로교통안전국(NHTSA) 결함조사국이 2022년 이후 FSD 작동 중 발생한 것으로 판정한 사례다.

5 EU 인공지능법 제15조의 표제는 "Accuracy, robustness and cybersecurity"로, 고위험 AI 시스템이 정확성·강건성·사이버보안을 적절한 수준으로 달성하고 생애주기 전반에 걸쳐 일관되게 작동하도록 설계·개발할 것을 요구한다. 이 세 요건이 본문의 기능성·강건성·보안성 검증에 각각 대응한다.

1) 기능성 검증

기능성 검증은 AI 모델이 의도한 기능을 명확히 수행하는지 확인하기 위해 시험 요구사항을 충족하는 데이터셋을 구성하는 것이 핵심이다. 과거에는 평균/표준편차 등 통계적 특성을 활용하거나 차원 축소 및 클러스터링 기반으로 분류하여, 데이터셋 전체의 통합 결과를 확인하는 방식이 주로 사용되었다.[12] 이러한 방식은 분류 기준은 정량적일지라도 각 데이터 조건이 실제 어떤 기능적 요구사항에 대응하는지 추적 및 해석하기 어렵다는 한계가 있었다.

최근에는 입력 데이터로부터 사람이 이해할 수 있는 정보를 추출하여 시험 조건을 세분화한다(예: 자율주행 도메인 기준 운전 방향, 도로 유형, 날씨, 계절 등).[13] 이를 조합해 테스트 케이스를 구성함으로써 모델의 전체 성능뿐 아니라 각 요구사항별 준수 여부까지 확인할 수 있다. 즉 기능성 검증은 암묵적 기준에 의존해 통합 성능만 측정하던 수준에서 벗어나, 운용 조건별로 판정 기준을 명시하고 요구사항 단위로 충족 여부를 추적하는, 보다 정밀하고 정량적인 검증으로 발전하고 있는 것이다.

2) 강건성 검증

강건성 검증은 AI 모델이 어떠한 악의적 혹은 비정상적 입력에 대해서도 의도한 결과를 정확히 낼 수 있는지를 검증하는 것이다. 이를 위해 AI 모델의 입력 데이터셋에 노이즈나 현실적인 변형이 가해진 상황에서도 AI 모델의 출력이 원본 입력의 결과와 일관성이 유지되는지, 성능 요구사항이 허용 기준 이하로 저하되지 않는지를 확인하는 검증을 수행한다. 강건성 검증은 무작위 변형을 반복 적용하는 랜덤 테스트 기술을 시작으로, 오류 유형의 데이터를 변형·조합하는 유전 알고리즘을 거쳐, 모델 내부의 기울기(Gradient) 정보를 활용하는 기울기 기반(Gradient-based) 테스트로 발전하면서 AI 모델의 강건함을 효율적으로 확인할 수 있게 되었다.[14][15][16] 그러나, 이러한 테스트들을 통해 발견된 오류가 없다고 하더라도 모델이 강건하다고 단정하기는 어렵다.

이를 보완하기 위해 정해진 입력의 변형 범위 내에서 오류가 발생하지 않음을 수학적으로 증명하는 형식 검증[6] 연구가 활발하게 진행되고 있다.[17][18][19] 즉, 강건성 검증은 단순한 오류 사례의 효율적 탐색을 넘어, AI가 어느 범위까지 정상 동작을 보증할 수 있는지 정량적으로 제시하는 방향으로 확장되고 있다.

6 형식 검증(Formal Verification)은 정해진 입력 변형 범위 전체에 대해 오류가 발생하지 않음을 수학적으로 증명하는 방법이다. 통계적 분석이 성능 유지를 '보여주는' 데 그치는 반면 형식적 방법은 이를 '증명'할 수 있으나, 입력 차원과 신경망 규모가 커질수록 계산량이 급증하는 확장성 한계가 있다.

3) 보안성 검증

보안성 검증은 공격자가 AI의 추론 메커니즘과 입력 특성을 악용하더라도 AI가 안전하게 동작하는지를 확인한다. 초기에는 보안 전문가의 경험과 직관에 의존한 레드티밍[7]이 중심이었으나, 전문가의 역량과 선택된 공격방식에 따라 결과가 달라지는 한계가 있었다.

최근에는 NIST의 적대적 공격 분류와 MITRE ATLAS 같은 공통 위험 프레임워크를 적용하여 평가 대상을 체계화하고[20][21], 적대적 공격 입력을 자동 생성해 대규모로 반복 시험하는 기술이 도입되고 있다.[22][23] 아울러 공격 성공률(Attack Success Rate)[8] 및 안전성 등급 등으로 결과를 수치화하려는 시도도 확대되고 있다.[24][25] 즉 보안성 검증도 정성적 취약점 점검에서 시험 범위와 실패 수준을 명시적으로 측정하는 구조로 전환되고 있다.

3가지 검증의 방법과 대상은 다르지만 발전의 지향점은 동일하다. AI 검증은 정성적·단편적 검증에서 벗어나 시험 조건과 판정 기준을 명확히 하고, 검증 결과를 수치 데이터와 Pass/Fail 기준으로 제시하는 정량적·객관적 검증으로 진화하고 있다. 한편, 산업계에서는 Cisco의 Cisco AI Defense, Patronus AI의 Patronus Platform, 슈어소프트테크의 VERIFAI 시리즈 등 AI 모델과 시스템을 검증하는 도구들이 속속 출시되면서 AI 테스팅을 자동화하여 효율적인 검증이 가능해지고 있다.[26][27][28]

그러나 여전히 근본적인 질문은 남아 있다. “어떤 조건까지 시험해야 충분한지”, “어느 수준의 강건성과 보안성을 확보해야 하는지”에 대한 공통된 합격 기준은 아직 확립되지 않았다. 기술은 “측정할 수 있는 단계”로 발전했지만, “어느 정도면 충분한가”에는 정답에 다다르지 못한 것이다. 다음 장에서는 이 미완의 과제가 주요 국가의 정책과 국내 제도에서 어떻게 논의되고 있는지 살펴본다.

7 레드티밍(Red-teaming)은 공격자의 관점에서 시스템의 결함과 취약점을 의도적으로 찾아내는 구조화된 시험 활동을 말한다. AI 분야에서는 모델이 내놓지 말아야 할 답변을 하도록 유도하는 입력을 탐색하는 활동을 가리킨다.

8 공격 성공률(ASR, Attack Success Rate)은 시도한 공격 가운데 방어를 우회해 목표한 결과를 얻어낸 비율이다.

글로벌 AI 신뢰성 및 안전 규제 동향

1) 글로벌 AI 규제화

앞서 살펴본 바와 같이 AI 검증 기술은 정량적/객관적 단계로 진화하고 있으나, '어느 수준을 안전한 수준으로 볼 것인가'에 대한 글로벌 공통 기준은 아직 확립되지 않았다. 그러나 AI의 적용 범위가 물리 시스템과 핵심 인프라로 급속히 확대되면서, 각국 정부는 검증 기술의 성숙을 기다리지 않고, 규제 체계를 먼저 마련하기 시작했다.

이 흐름이 국제사회에서 본격적으로 공론화된 계기가 2023년 '블레츨리 선언(Bletchley Declaration)'이다.[9][29] 기존 AI 모델보다 훨씬 강력한 성능을 보유한 AI의 위험에 공동 대응하자는 선언에서 출발한 AI 안전 논의는 이후 각국의 법률, 행정명령, 가이드라인 등으로 구체화되었다. 다만, 구체적 방식은 각국의 제도적 여건과 규제 전략에 따라 서로 다른 형태로 발전하고 있다.

2) 주요국 규제 현황

EU, 미국, 중국, 일본 등 주요 국가들의 규제와 주요 내용, 이행 수단을 정리하면 다음 표와 같다.

[표 1] 글로벌 AI 신뢰성 관련 규제

국가	규제 및 법령	주요 내용	이행 수단
EU	인공지능법(AI Act)[8]	위험도에 따라 4등급으로 분류, 고위험 AI 시스템은 출시 전 적합성 평가 및 CE 마킹 의무화	위반 시 글로벌 매출의 최대 7% 과징금 부과
중국	생성형 AI 서비스 잠정관리방법[30]	여론 형성과 관련된 AI 출시 전 안전성 평가 및 알고리즘 신고	신고 미이행 시 서비스 출시 불가
인도	정보기술 규칙 개정안[31]	AI로 생성한 정보에 라벨 및 출처에 대한 메타데이터 포함 의무화	플랫폼(중개자)에게도 책임 부과
영국	AI 안전연구소의 자발적 협약[32]	최첨단 AI 모델의 위험성에 대해 배포 전 평가 수행	AI 안전연구소와 협업을 통한 AI 모델 평가 진행
미국	AI 표준혁신센터(CAISI) 및 행정명령[33]	주요 AI 개발사와 모델 접근 및 배포 전 안전성 평가 협약 체결	AI 표준혁신센터의 표준 프레임워크에 대해 실증 및 평가
일본	AI 추진법 및 공공조달 가이드라인[34][35]	AI 시스템의 안전성/신뢰성 확보를 공공조달 입찰 조건으로 연계	공공 조달 지침을 통한 품질 확보 유도
싱가포르	AI 어슈어런스 파일럿[36][52]	제3자 시험기관에 의한 Application 검증	시장에 의해서 AI 신뢰성 확보 유도

9 2023년 11월 1~2일 영국 블레츨리 파크에서 열린 제1차 AI 안전성 정상회의에서 한국·미국·중국을 포함한 28개국과 EU가 서명했다. 광범위한 작업을 수행할 수 있는 고성능 범용 AI를 '프론티어 AI'로 규정하고 그 위험에 국제적으로 공동 대응할 것을 합의했다.

국가별 제도화 과정은 '측정 기술의 성숙도'와 '규제 현실' 간의 밀접한 관계가 있음을 보여준다. EU는 AI Act를 구체화한 표준과 AI 적합성 평가 체계의 준비가 지연되어 고위험 AI에 대한 의무 적용 시점을 2027년 12월 이후로 연기한 바 있다.[10][37][38] 인도도 정보기술 규칙을 개정하는 과정에서 AI가 생성한 '시각 콘텐츠 표시 면적 10% 이상'이라는 정량적 표시 기준을 제시했으나, 최종안에서는 해당 수치 기준을 삭제하고 "명확하고 두드러진 라벨링"이라는 원칙 중심의 기준으로 조정했다.[11][31] 반면, 영국 AI 안전연구소는 최첨단 AI 모델을 측정하고 평가하는 평가 프레임워크를 빠르게 수립하여 공개함으로써[39] AI 개발사들은 협약에 맞춰 배포 전에 평가하고 있다.

3) AI 규제 유형과 4대 공통 특징

각국의 대응은 규제 대상(위험등급 기반 vs 행위 기반)과 이행 수단(제재형 vs 참여 유도형)에 따라 다음 표처럼 4가지 유형으로 분류된다.

[표 2] AI 규제 유형

	위험등급 기반	행위 기반
제재형	**등급규제형** — EU - AI 위험 등급을 판정해 의무 부과 - 위반 시 과징금 및 과태료 부과	**행위규제형** — 중국, 인도 - 신고·표시·삭제 등 특정 행위를 직접 명령 - 위반 시 출시 금지, 법적 책임 부과
참여 유도형	**역량평가형** — 영국, 미국 - 국가 연구소 중심의 AI 모델 평가 프레임워크를 공개 - 기업들은 이를 통해 자발적 평가 및 배포	**절차이행형** — 일본, 싱가포르 - AI 개발/검증 절차 준수를 요구 - 공공 조달 및 시장에 의한 평가를 유도

이처럼 각국은 규제 대상과 이행 수단을 달리하고 있지만, 그 가운데에서도 다음과 같은 공통점이 나타난다.

① 사전 검증 체계로의 전환: AI 신뢰성을 확인하는 절차가 시장에 출시되기 전에 진행되고 있다. 적합성 평가, 사전 신고, 표시 의무, 배포 전 시험은 모두 시장 출시에 앞서 안전을 확인하는 절차다.

10 유럽위원회는 2025년 11월 'Digital Omnibus on AI'를 통해 고위험 AI 관련 의무의 적용 시점을 조화표준(harmonised standards)의 준비 상황에 연동해 늦추는 개정안을 제안했고, 이사회와 유럽의회는 2026년 이에 합의했다.

11 인도 전자정보기술부(MeitY)가 회람한 개정 초안은 AI 생성 시각물에 표시 면적의 10% 이상을 차지하는 라벨을, 음성물에는 앞부분 10% 구간에 걸친 고지를 요구했다. 최종 규칙에서는 이 정량 기준이 삭제되고 '명확하고 두드러지게(clearly and prominently) 표시'라는 원칙 기준으로 대체되었다.

② 측정 가능성과 규제의 정밀성: 규제 기준을 구체화하기 위해서는 이를 객관적으로 측정할 수 있는 기술적 기반과 산업 현장에서의 적용 가능성이 함께 뒷받침되어야 한다. EU와 인도의 사례는 정량적이고 세부적인 규제 기준을 마련하기 위해 측정 기술의 성숙도뿐만 아니라 실제 이행 가능성까지 고려해야 함을 보여준다.

③ 정부의 평가 기준·프레임워크 제시: 정부는 AI 검증 및 측정 역량을 확보하고, 이를 민간이 활용할 수 있는 평가 체계로 제공하고 있다. 영국과 미국은 국가 연구기관을 중심으로 평가 프레임워크를 구축·공개하고, 이를 기반으로 민간의 자발적인 평가와 협력을 유도하고 있다.

④ AI 검증 체계의 산업화: 싱가포르의 사례처럼 시험 방법의 표준화, 시험기관 인정 체계 구축 등 AI 검증 자체를 하나의 신산업 영역으로 육성하려는 정책적 시도가 구체화되고 있다.

국내 AI 신뢰성 및 안전 규제의 현황 및 글로벌 규제와의 차이점

1) 국내 AI 기본법 체계의 현황과 특성

2026년 1월 22일 시행된 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법」(이하 AI 기본법)은 고영향 AI라는 위험 등급 체계[12]를 도입하여 판정하고 이에 따라 법적 의무를 부과하며, 의무 위반이나 시정명령 불이행 시 과태료를 부과하는 구조를 취하고 있다.[9][40] 이는 앞서 3장에서 분류한 유형 중 EU와 동일한 '등급규제형'에 해당한다. AI 기본법은 생성형 AI 결과물의 고지·표시 의무, 고영향 AI의 위험관리와 설명 가능성 확보, 일정 연산량 이상의 고성능 AI[13]에 대한 안전성 확보 체계 구축 등을 기업의 법적 책임으로 명시하고 있다. 정부도 시행에 맞춰 투명성·안전성·고영향 판단·사업자 책무·영향평가 등 5대 가이드라인을 제시하고[43][44][45][46][47] 최소 1년의 계도 기간을 부여했다.[41][42]

이로써 한국은 AI 안전성 확보를 위한 법적 체계를 마련했으나, AI 산업 현장에서 이를 이행하기 위한 구체적인 기준은 아직 충분히 마련되지 않았다. 고영향 AI의 위험등급을 객관적으로 판정할 기술과 AI 평가 프레임워크, 정량적 기준 등이 구체화되지 않은 상태에서 법적 의무가 먼저 발효됐기 때문이다.

12 AI 기본법상 고영향 인공지능은 사람의 생명·신체 안전 및 기본권에 중대한 영향을 미칠 우려가 있는 영역에 사용되는 AI를 말하며, 에너지·보건의료·원자력·범죄수사·채용·대출심사·교통·공공서비스·교육 등이 해당 영역으로 열거되어 있다.

13 학습에 사용된 누적 연산량이 10의 26승 부동소수점연산(FLOP) 이상인 인공지능을 가리킨다. 해당 사업자는 인공지능 수명주기 전반의 위험을 식별·평가·완화하고 위험관리체계 구축 결과를 과학기술정보통신부장관에게 제출해야 한다.

2) 글로벌 기준과 국내 현황 간 차이점

3장에서 도출한 글로벌 공통점과 국내 AI 기본법을 비교/분석하면 다음 표와 같은 차이가 있다.

[표 3] 글로벌 AI 규제 공통점과 국내 AI 규제 현황 간 비교 분석

글로벌 AI 규제 공통점	국내 AI 규제 현황	구조적 한계 및 차이점 분석
사전 검증 체계로의 전환	AI 기본법을 제정하여 고영향 AI 판별과 안전성 확보를 법적 의무로 규정	산업 현장 적용이 가능한 수준의 세부 검증 표준 및 평가 가이드라인 미비
측정 가능성과 규제의 정밀성	AI 등급 판정 요건은 가이드라인에 제시되고 실질적 요건은 고시에 위임	AI 등급, AI 안전도 등 객관화, 수치화할 수 있는 정량 지표 부재
정부의 평가 기준·프레임워크 제시	AI 안전연구소 중심으로 42종의 AI 모델 평가 및 한국어 벤치마크를 구축했으나, 평가 도구 및 세부 방법론은 미공개[48][49]	민간 기업이 참조할 공개 레퍼런스·프로토콜 미흡
AI 검증 체계의 산업화	CAT, AI-MASTER 등 민간 검증체계 운영 중[50][51] * CAT: TTA에서 진행하는 AI 신뢰성 인증 제도 * AI-MASTER: 한국인공지능산업협회, 슈어소프트테크에서 주도하는 AI 신뢰성 인증 제도	국제 상호인정 제도와 AI 특화 시험기관 인정 체계는 부재

종합하면, 한국은 AI 기본법 제정을 통해 사전 검증 체계 도입 및 정부 중심의 평가 인프라 구축 등 글로벌 규제 트렌드의 제도적 틀을 발 빠르게 확보했으나, 이를 산업 현장에 실효성 있게 적용하는 부분에 있어서는 '기술적·구조적 이행 격차(Implementation Gap)'가 존재한다. 또한, 고영향 AI 판정 및 안전성 수준을 객관적으로 입증할 정량적 측정 지표와 세부 검증 표준이 부재하여 기업의 법적 불확실성을 높이고 있다. 국가 차원에서 구축한 평가 도구와 방법론이 민간 기업의 개발 실무에 즉시 활용될 수 있도록 배포되고 있지 않다. 아울러 CAT, AI-MASTER 등 민간 중심의 신뢰성 인증제도가 운영 중임에도 불구하고, 글로벌 상호인정 제도 및 AI 특화 시험기관 인정 체계의 미비로 인해 국내 검증 결과가 국제적으로 통용되고, 검증 산업이 확대되는 데 한계를 나타내고 있다.

신뢰할 수 있는 AI를 위한 3대 정책 제언

앞서 확인한 글로벌 기준과 국내 상황을 고려했을 때, AI 신뢰성 구축을 위해 다음 세 가지 정책 방향이 필요하다.

① AI 신뢰성에 대해 산업 현장 중심의 정량적 세부 검증 표준 및 정밀 측정 지표를 신속하게 확립해야 한다. AI 기본법 제정으로 안전성 확보를 위한 법적 의무는 마련되었으나, 이를 산업 현장에서 이행하기 위한 정량적 판정 기준과 세부 검증 방법은 충분히 구체화되지 않았다. 정부는 주요 분야별(의료, 자율주행, 금융, 제조 등) 운용 특성을 반영한 정량적 세부 검증 기준(적대적 공격 성공률 허용치, 워터마크 존속성 등)을 하위 고시에 명확히 규정해야 한다. 또한 무엇을, 어떤 절차로, 어떤 판정 기준에 따라 측정할 것인지를 규정한 표준 시험 방법을 제시하여 산업 현장에 명확한 이행 기준을 제공해야 한다. 다만, 이러한 기준이 중소/스타트업에게 과도한 진입장벽이 되지 않도록 위험등급과 기업 규모에 따라 검증 수준을 차등화하고, 공용평가 데이터셋과 검증 바우처 등 완충 장치를 함께 마련해야 한다.

② 정부는 직접 AI 평가 도구와 데이터셋, 방법론을 개발해 보급하는 공공 주도 방식보다 시장이 준수해야 할 명확한 안전 기준과 도메인별 평가 가이드라인을 제시하는 선제적 역할에 집중해야 한다. 한편, 해당 기준을 입증할 실질적인 검증 도구, 자동화 평가 솔루션(레드티밍, 형식 검증 등), 그리고 자율주행·로보틱스 등 물리 시스템을 위한 시뮬레이션과 실증 환경의 개발 및 확산은 민간의 전문성과 기술 생태계가 주도하도록 유도해야 한다. 정부는 민간 중심의 검증 기술 생태계가 자율적으로 형성되고 활성화될 수 있도록 연구개발(R&D) 지원 및 인프라 연계 등 시장 친화적 환경을 조성하는 데 중점을 두어야 한다. 특히, 개별기업이 자체적으로 구축하기 어려운 공개 벤치마크와 평가 데이터셋, 물리 시스템 실증을 위한 공동 테스트베드는 공공이 구축하되, 운영과 고도화는 시험기관 및 산업계가 참여하는 개방형 생태계에 맡길 수 있다.

③ AI 인증/검증 제도가 국제적으로 통용될 수 있게 해야 하며, 공공조달 연계를 통해 'AI 검증 산업'을 육성할 필요가 있다. AI에 특화된 시험기관의 자격과 검증 품질을 관리하는 인정 체계를 구축하고, 국내 시험 결과가 해외 적합성 평가에서도 인정되도록 상호인정 제도를 확보해야 한다. 동시에 공공조달과 국가 AI 사업 평가에 안전성 검증을 결합해 AI 검증에 대한 시장 수요를 만들어야 한다. 이를 위한 단계적 이행 과제와 추진 주체를 정리하면 다음 표와 같다.

[표 4] AI 검증 인프라 구축 단계별 로드맵

단계	주요 과제	주관·협력
1단계 (계도 기간 내)	분야별 정량 검증 기준 및 표준 시험 방법 하위 고시 제정 AI 특화 시험기관 인정 기준 마련	과학기술정보통신부 국가기술표준원
2단계 (제도 정착기)	인정 체계 운영 및 시험기관 지정 공공조달 입찰 요건에 안전성 검증 결과 반영	국가기술표준원·KOLAS 조달청, TTA·한국인공지능산업협회 등 인증기관
3단계 (국제 연계기)	국내 시험 결과의 해외 적합성 평가 인정 확보 EU·싱가포르 등과 상호인정 협력 추진	과학기술정보통신부, KOLAS 한국AI안전연구소

법적 의무와 기술적 기준 사이의 격차는 상호 조정할 필요가 있다. EU가 AI 의무의 적용 시점을 표준 준비 상황에 연동한 것처럼 법적 의무 이행의 시기를 조절하거나, 규제 샌드박스를 운영하여 부분적으로 적용하는 방법 등 산업계와 정부가 서로 소통해야 한다.

마지막으로 위의 세 가지 정책 제언은 서로 밀접한 관계가 있다. 기준이 없으면 무엇을 측정할지 정할 수 없고, 측정 수단이 없으면 기준은 문서에 머물며, 국제적으로 통용되지 않으면, 결과는 국내로 한정된다. 지금 필요한 것은 규제의 강도를 높이는 일이 아니라 안전을 판정하고 그 판정을 통용 가능한 증명으로 만드는 기반을 갖추는 일이다. 그 기반이 갖춰질 때 AI 안전은 혁신을 늦추는 제약이 아니라, 우리 AI가 더 빠르고 더 넓은 시장으로 나아갈 수 있게 하는 조건이 될 것이다.

참고문헌

[1] TechCrunch. (2026). *ChatGPT reaches 900 million weekly active users*. https://techcrunch.com/2026/02/27/chatgpt-reaches-900m-weekly-active-users/

[2] Organisation for Economic Co-operation and Development. (2026). *OECD AI Incidents and Hazards Monitor (AIM)*. https://oecd.ai/en/incidents

[3] CNN Business. (2023). *Google shares lose $100 billion after company's AI chatbot makes an error during demo*. https://www.cnn.com/2023/02/08/tech/google-ai-bard-demo-error

[4] CFO Dive. (2024). *Scammers siphon $25M from engineering firm Arup via AI deepfake 'CFO'*. https://www.cfodive.com/news/scammers-siphon-25m-engineering-firm-arup-deepfake-cfo-ai/716501/

[5] Tesla Deaths. (2025). *Digital record of Tesla crashes resulting in death*. https://www.tesladeaths.com/

[6] National Highway Traffic Safety Administration. (2024). *Preliminary evaluation PE24031: Tesla Full Self-Driving (FSD)*. U.S. Department of Transportation.

[7] International Organization for Standardization & International Electrotechnical Commission. (2020). *ISO/IEC TR 24028: 2020 Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence*.

[8] European Parliament & Council. (2024). *Regulation (EU) 2024/1689 (Artificial Intelligence Act)*. Official Journal of the European Union.

[9] *인공지능 발전과 신뢰 기반 조성 등에 관한 기본법*, 법률 제20676호 (2025).

[10] National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)*. U.S. Department of Commerce.

[11] High-Level Expert Group on Artificial Intelligence (AI HLEG). (2019). *Ethics guidelines for trustworthy AI*. European Commission.

[12] Eyuboglu, S., Varma, M., Saab, K., et al. (2022). *Domino: Discovering systematic errors with cross-modal embeddings*. International Conference on Learning Representations (ICLR).

[13] Chen, M., et al. (2025). *HiBug2: Efficient and interpretable error slice discovery for comprehensive model debugging*. arXiv: 2501.16751.

[14] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). *Explaining and harnessing adversarial examples*. International Conference on Learning Representations (ICLR).

[15] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). *Towards deep learning models resistant to adversarial attacks*. International Conference on Learning Representations (ICLR).

[16] International Organization for Standardization & International Electrotechnical Commission. (2021). *ISO/IEC TR 24029-1:2021 Artificial intelligence — Assessment of the robustness of neural networks — Part 1: Overview*.

[17] International Organization for Standardization & International Electrotechnical Commission. (2023). *ISO/IEC 24029-2:2023 Artificial intelligence — Assessment of the robustness of neural networks — Part 2: Methodology for the use of formal methods*.

[18] Cohen, J., Rosenfeld, E., & Kolter, J. Z. (2019). *Certified adversarial robustness via randomized smoothing*. International Conference on Machine Learning (ICML).

[19] Kaulen, T., et al. (2025). *The 6th International Verification of Neural Networks Competition (VNN-COMP 2025): Summary and results*.

[20] National Institute of Standards and Technology. (2025). *Adversarial machine learning: A taxonomy and terminology of attacks and mitigations (NIST AI 100-2)*. U.S. Department of Commerce.

[21] MITRE Corporation. (2025). *ATLAS: Adversarial threat landscape for artificial-intelligence systems*.

[22] Mazeika, M., Phan, L., Yin, X., et al. (2024). *HarmBench: A standardized evaluation framework for automated red teaming and robust refusal*. International Conference on Machine Learning (ICML).

[23] Chao, P., Debenedetti, E., Robey, A., et al. (2024). *JailbreakBench: An open robustness benchmark for jailbreaking large language models*. Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track.

[24] Souly, A., Lu, Q., Bowen, D., et al. (2024). *A StrongREJECT for empty jailbreaks*. Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track.

[25] Ghosh, S., et al. (2025). *AILuminate: Introducing v1.0 of the AI risk and reliability benchmark from MLCommons*. arXiv: 2503.05731.

[26] Cisco. (2026). *AI model and application validation (Cisco AI Defense) [Product overview]*. https://www.cisco.com/site/us/en/products/security/ai-defense/ai-model-application-validation/index.html

[27] Patronus AI. (2026). *Patronus platform [Product overview]*. https://www.patronus.ai/

[28] AI타임스. (2024). *슈어소프트테크, AI 모델 검증 도구 「베리파이엠(VERIFAI-M)」 최초 공개*.

[29] UK AI Safety Summit. (2023). *The Bletchley Declaration*.

[30] Cyberspace Administration of China. (2023). *生成式人工智能服务暂行管理办法 [생성형 AI 서비스 잠정관리방법]*.

[31] Ministry of Electronics and Information Technology. (2026). *Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules, 2026*.

[32] UK AI Security Institute. (2025). *Frontier AI trends report*.

[33] National Institute of Standards and Technology. (2024). *U.S. AI Safety Institute Signs Agreements Regarding AI Safety Research, Testing and Evaluation With Anthropic and OpenAI*.

[34] Cabinet Office of Japan. (2025). *人工知能関連技術の研究開発及び活用の推進に関する法律 [AI 관련 기술 연구개발 및 활용 추진법]*, 법률 제53호.

[35] Digital Agency of Japan. (2025). *行政の進化と革新のための生成AIの調達·利活用に係るガイドライン (DS-920) [행정의 진화와 혁신을 위한 생성형 AI 조달·활용 지침]*.

[36] Infocomm Media Development Authority & AI Verify Foundation. (2025). *Global AI assurance pilot: Findings report*.

[37] European Commission. (2025). *Digital omnibus: Proposal amending Regulation (EU) 2024/1689*.

[38] Council of the EU & European Parliament. (2026). *Artificial intelligence: Council and Parliament agree to simplify and streamline rules [Press release]*.

[39] UK AI Safety Institute. (2024). *Inspect: An open-source framework for large language model evaluations*. https://inspect.aisi.org.uk/

[40] *인공지능 발전과 신뢰 기반 조성 등에 관한 기본법 시행령*, 대통령령 (2026).

[41] 과학기술정보통신부. (2026). *인공지능기본법 22일 시행* [보도자료].

[42] 과학기술정보통신부. (2026). *인공지능 발전과 신뢰 기반 조성 등에 관한 기본법 시행령 개정안 국무회의 통과* [보도자료].

[43] 과학기술정보통신부. (2026). *인공지능 투명성 확보 가이드라인*.

[44] 과학기술정보통신부. (2026). *인공지능 안전성 확보 가이드라인*.

[45] 과학기술정보통신부. (2026). *고영향 인공지능 판단 가이드라인*.

[46] 과학기술정보통신부. (2026). *고영향 인공지능 사업자 책무 가이드라인*.

[47] 과학기술정보통신부. (2026). *인공지능 영향평가 가이드라인*.

[48] 한국AI안전연구소. (2026). *국내외 AI 모델 42종 안전성 평가 수행 실적*.

[49] 한국AI안전연구소. (2026). *LLM 안전성 및 신뢰성 평가 데이터셋 구축 사업*.

[50] 한국정보통신기술협회. (2026). *인공지능 신뢰성 인증(CAT) 안내*.

[51] 한국인공지능산업협회. (2026). *AI-MASTER 인증 제도 안내*.

[52] Infocomm Media Development Authority. (2025). *Singapore launches new tools to help businesses protect data and deploy AI in a trusted ecosystem*.