AI 신뢰성 인증 소개

신준호 TTA AI융합기획단 단장

1. 머리말

지난 5월 21~22일 개최된 AI 서울 정상회의를 계기로 오픈AI(OpenAI), 구글(Google) 등 국내외 14개 AI 기업은 '서울 AI 기업 서약'에 참여했다[1]. 이는 기업이 자발적으로 '책임 있는 AI 개발'에 나서겠다는 선언이다. 더불어 최첨단 AI 모델을 보유하거나 개발하는중인 16개 기업은 좀 더 엄정한 AI 모델 안전성 확보를 약속하는 '프론티어 AI 안전 약속'에 이름을 올렸다[2]. 네이버, 삼성전자, SK텔레콤, LG AI 연구원, 카카오, KT 등 국내 AI 관련 기업들도 이번 AI 기업 서약과 안전 약속을 통해 글로벌 AI 안전규범 준수에 동참했다. 안전성·보안성·신뢰성을 갖춘 AI 설계·개발·배치·사용을 보장하기로 한 것이다.

한편 앞서 올해 3월, EU(유럽연합, European Union)는 세계 최초로 AI에 대한 규제법(AI 법)을 통과시켰다. AI 법의 골자는 그 위험성과 영향력에 따라 AI를 '금지, 고위험, 제한된 위험, 최소위험'으로 구분하고, 이 중 고위험 AI와 범용 AI에 대해 엄격한 규제를 적용하는 것이다.

지난 2022년 11월 ChatGPT의 등장과 함께 이뤄진 AI의 가파른 기술발전은 많은 이들에게 큰 인상을 남겼다. AI의 긍정적 효과뿐만 아니라, 부정적 영향에 대한 우려의 목소리도 높아지고 있는 상황이다. 이에 AI 신뢰성 확보를 위한 AI 개발·서비스 제공 기업의 자율적 노력이 절실하게 요구되고 있다.

AI 신뢰성 인증(Certification of AI Trustworthiness, CAT)에 대한 본격적인 소개에 앞서, AI 신뢰성의 정의를 간략히 살펴보도록 하자. 2022년 출판된 AI 개념 및 용어에 관한 국제표준인 ISO/IEC 22989에서는 AI 신뢰성(Trustworthiness)을 'AI 시스템의 의사 결정과 활동에 영향을 받는 모든 이해관계자의 기대치를 충족시키는 역량'으로 정의하고 있다. 이는 AI 개발자, 제공자, 사용자가 모두 만족할 수있는 수준을 충족시키는 동시에, 정확성(Accuracy),강건성(Robustness), 안정성(Reliability) 등 기술적 가치, 그리고 투명성(Transparency), 책무성(Accountability), 공정성(Fairness)과 같은 사회적 가치에도 부합되는 역량을 지칭한다.

2. AI 신뢰성 인증제도

2.1 인증대상

AI 신뢰성 인증의 주 대상은 'AI 모델'과 'AI 시스템'이다<표 1>. AI 모델을 개발·제공하는 기업의 경우 AI 모델 인증을 신청할 수 있으며, 대상 모델은 일반적으로 고유한 모델 명칭과 배포 버전을 가지고 있다. 공개된 오픈소스 모델을 추가훈련 및 미세조정(Fine-tuning)을 통해 재구성한 AI

모델의 경우에도 동일하게 인증받을 수 있다. 즉, 많은 국내 기업들이 개발·보유하고 있는 각종 소규모 언어 모델도 AI 신뢰성 인증 대상에 해당된다.

<표 1> AI 신뢰성 인증범위

인증범위	인증방법	내용
AI 거버넌스 체계	심사	정책·조직·프로세스 이행수준 점검
AI 제품 신뢰성	시험·평가	AI 제품의 신뢰성 특성 항목 시험, 평가, 검증, 확인

AI 서비스와 애플리케이션 등 AI 시스템 개발·제공사의 경우엔 AI 시스템 인증을 신청할 수 있다. 여기서 AI 시스템은 AI 모델을 포함한 네트워크, 데이터베이스, 사용자 이용환경 등 하드웨어·소프트웨어를 포함한 AI 관련 제품을 지칭한다. 통상 국내 기업에서 접근하고 있는 범용 AI 모델 기반 애플리케이션 개발 방식은 외부에서 제공하는 범용 AI 모델을 API 또는 오픈소스 형태로 활용하는 것이다. 이는 'AI 시스템'에 해당된다.

2.2 근거 표준 및 지침

2020년 5월 발간된 ISO/IEC TR 24028(AI 신뢰성 개요)는 표준관점에서 폭넓은 AI 신뢰성 확보 접근 방안을 조사한, 최초의 기술보고서라 할 수 있다. 상기 기술보고서에서 현재 AI와 같은 새로운 기술의 경우, 알려지지 않은 위험에 대한 예방적 차원으로 '위험관리' 방식이 신뢰성 확보에 도움이 되며, 기존 SW·데이터 품질모델 및 시험평가 표준도 AI 신뢰성 확보에 활용할 수 있도록 고려할 수 있다고 하였다. 보고서는 이와 함께 책임성, 책무성, 거버넌스, 안전의 중요성도 강조했다. AI 신뢰성은 기술적 측면뿐 아니라 사회적 측면까지 고려해야 하기 때문이다.

2022년부터 AI 신뢰성 관련 표준 및 지침 발간이 본격화됐다. 현재의 AI 신뢰성 인증은 국내외 가이드라인 및 국제표준에서 권고하고 있는 'AI 신뢰성 확보를 위한 사회기술적 요건'을 중심으로 그 기준을 마련했다. 특히 인증대상인 AI 모델·시스템을 설계·개발·배포하는 데 구현된 기업 내부의 AI 거버넌스(정책, 조직, 임무), 그리고 주요 신뢰성 특성에 대한 검증·확인 활동들을 총 15개의 요구사항으로 녹여내어 인증 요건으로 정하였다. AI 신뢰성 인증에 참조하고 있는 ISO/IEC JTC 1/SC 42 국제표준, NIST(미국 국립표준기술연구소, National Institute of Standards and Technology) 표준과 함께, TTA에서 발간한 '신뢰할 수 있는 인공지능 개발안내서'에 대한 핵심 내용은 아래 박스에 요약해 보았다.

ISO/IEC 22989:2022 (AI 개념 및 용어)

- AI 신뢰성 개념: AI 시스템의 의사결정 및 활동에 영향을 받는, 모든 이해당사자의 기대치를 충족시키는 역량을 지칭. 확인 가능함을 전제로 함
- AI 신뢰성 특성: 강건성, 안정성, 회복탄력성, 제어가능성, 설명가능성, 예측가능성, 투명성, 편향 및 공정성, 보안, 프라이버시, 안전, 책무성, 무결성, 진실성, 품질 및 사용성 등

ISO/IEC 23894:2023 (AI 위험관리 가이던스)
- AI를 개발·생산·배포하거나 AI를 활용한 제품·시스템·서비스를 사용하는 조직이 AI와 관련된 위험을 관리할 수 있도록, 핵심 임무들을 제시하고 위험관리 절차를 안내
- ISO 31000 위험관리 표준과 연계해 사용

ISO/IEC 42001:2023 (AI 관리 시스템)
- AI를 활용한 제품·서비스를 제공하는 조직에서 이행해야 하는 관리적 요구사항

NIST, AI 위험관리 프레임워크
- 신뢰할 수 있는 AI 시스템 특성: 안전, 보안 및 회복탄력성, 설명가능성 및 해석가능성, 향상된 개인정보보호, 편향 관리 및 공정성, 책무성 및 투명성, 유효성 및 안정성
- 위험관리 핵심기능: 감독(GOVERN), 매핑(MAP), 측정(MEASURE), 관리(MANAGE)

과학기술정보통신부, 신뢰할 수 있는 인공지능 개발안내서 7종(TTA 발간)
- 2023년 발간분야: 일반, 공공사회, 의료, 자율주행
- 2024년 발간분야: 일반(개정), 생성형 AI 서비스, 스마트 치안, 채용

단체표준, TTA.KO-10.1497 AI 시스템 신뢰성 제고를 위한 요구사항
- 전 AI 생명주기에서 AI 위험관리 및 신뢰성 제고를 위해 준수해야 할 요구사항 15개 제시
- 주요 신뢰성 특성: 설명가능성, 투명성, 제어가능성, 회복탄력성, 안정성, 강건성, 예측가능성, 보안성, 공정성, 프라이버시, 안전성, 책임성

2.3 인증범위

AI 신뢰성 인증은 크게 AI 거버넌스 체계, AI 제품 신뢰성이라는 2가지 범위를 대상으로 한다. 고위험 AI 제품의 경우 2개 부문 모두 심사받아야 하는데, 이때 위험성 평가결과에 따라 중점적으로 관리해야 하는 위험요소를 고려해 상세 점검항목들이 정해진다.

- AI 거버넌스 체계 : 책임 있는 AI를 위한 AI 거버넌스 구축 및 이행
- AI 제품 신뢰성 : 투명성, 공정성(편향관리), 안전성, 설명가능성, 정확성, 강건성, 안정성 등

참고로 EU AI 법에선 사람의 건강, 안전 및 기본권에 높은 위험을 야기할 수 있는 경우 고위험 AI 시스템으로 분류한다. 이에 대해 위험관리 등 거버넌스 준수 및 투명성, 정확성, 강건성을 포함한 적합성 평가를 요구한다. 또한 ChatGPT와 같은 범용 AI 모델도 모두 고위험군으로 분류해 엄정한 규제를 적용하고 있다.

2.4 인증기준

AI 거버넌스 체계 측면에선, 인증대상 AI 모델 및 시스템의 설계·개발·배포 전 과정에 걸쳐 정책·조직·프로세스가 올바르게 작동하는지 체크한다. 이를 통해 위험관리를 중점으로 한 아래의 15개 요구사항을 모두 충족하는지 확인한 후 인증을 부여하는 것이다. 이 과정에서 투명성, 책무성, 개인정보보호, 저작권보호, 공정성, 안전성, 추적가능성, 설명가능성 및 보안성 등 AI 제품의 신뢰성 특성에 대한 시험평가도 함께 진행된다<표 2>.

<표 2> AI 제품 신뢰성 특성별 점검사항

점검방법	특성 구분	점검사항(예시)
시험·평가 검증·확인	투명성	모델·데이터 명세, 벤치마크 결과, AI 생성 콘텐츠 표시
	책무성	규제기반 책임성, 합의기반
	책임성	개인정보보호 학습데이터 출처 및 비식별화된 데이터 사용 확인
	저작권보호	학습데이터 출처 및 저작권 확보여부 확인
	공정성	인구통계학적 동등성
	안전성	유해한 오용, 제어가능성, 안전기능(세이프가드)
	추적가능성	데이터 추적가능성, 기록 추적가능성
	설명가능성	모델설명력
	해석가능성	결과해석력
	품질	정확성, 안정성, 강건성
	보안	모델 보안, 데이터 보안, 시스템 보안

다만 AI 신뢰성 특성 측정방법 및 기준과 관련, 현재 국제적으로 통용되는 표준이 부족한 상황이다. 이때문에 아직 시험평가 기준을 별도로 정해 두지 않았으며, 주로 시험평가 수행여부에 초점을 맞추고 있다.

2.4.1 AI 시스템에 대한 위험관리 계획 및 수행

- 위험관리 요소, 위험관리 계획 및 수행 프로세스가 명확하게 정의돼 있는지 확인
- 위험 식별, 분석, 평가, 대응 과정이 체계적으로 수행되는지 확인

2.4.2 AI 거버넌스 체계 구성

- 거버넌스 체계가 조직 및 AI 모델·시스템의 범위를 명확히 정의하는지 확인
- 거버넌스 체계가 운영에 대한 책임과 권한을 명확히 구분하는지 확인
- 거버넌스 체계가 조직의 제품 및 서비스를 감시, 관리, 감독하는지 확인
- 거버넌스 체계를 지속적으로 평가하고 개선하는지 확인

※ ISO/IEC 23894, ISO/IEC 42001, NIST AI 위험관리 프레임워크 등의 공통 임무요건 적용

2.4.3 AI 시스템의 신뢰성 테스트 계획 수립

- 테스트를 통해 달성하고자 하는 목표를 명확하게 정의했는지 확인
- 테스트 대상이 되는 AI 시스템의 기능과 범위가 명확하게 정의돼 있는지 확인

- 테스트를 수행하는 데 사용할 방법론(평가기준 포함)이 정의돼 있는지 확인
- 테스트를 수행하는 데 필요한 환경과 데이터가 정의돼 있는지 확인

2.4.4 AI 시스템의 추적가능성 및 변경이력 확보
- 데이터 출처 및 변환 과정에 대한 정보가 명확하게 기록돼 있는지 확인
- 모델 학습과정 및 알고리즘 변화에 대한 정보가 기록돼 있는지 확인
- 코드 및 시스템 버전 관리가 체계적으로 이뤄지고 있는지 확인
- 성능 및 결과에 대한 지속적인 추적과 분석이 이뤄지고 있는지 확인
- 변경 사항에 대한 상세 정보가 정확하게 기록돼 있는지 확인
- 변경 사항에 대한 승인 및 검증 프로세스가 명확하게 정의돼 있는지 확인
- 변경 사항에 대한 롤백 및 복구 기능이 정상적으로 작동하는지 확인

2.4.5 데이터의 활용을 위한 상세 정보 제공
- 메타데이터에 중요 관리정보가 모두 포함돼 있는지 확인
- 정제 과정에서의 데이터 특성 변환에 대한 규칙이 명확한지 확인
- 민간정보 또는 개인정보에 해당하거나 일부 포함돼 있는지 확인
- 데이터 출처(데이터의 원본 제공자 또는 수집 방법)가 명확한지 확인

2.4.6 데이터 견고성 확보를 위한 이상 데이터 점검
- 데이터 수집 과정에서 발생하는 오류, 누락, 부정확한 값 등에 대한 내부 검증활동 확인
- 데이터 최적화 과정에서 발생하는 데이터 변형, 통계적 오류 등에 대한 내부 검증활동 확인
- 데이터 변조 공격 등을 감지하고 방어하기 위한 수단을 구축했는지 확인

2.4.7 수집 및 가공된 학습 데이터의 편향 제거
- 학습 데이터의 편향 특성, 유형, 기준이 명확하게 정의돼 있는지 확인
- 학습 데이터 수집 및 가공 시 편향 제거를 위한 기술 도입의 적정성 확인
- 데이터 라벨링 작업 지침, 교육, 감독 활동의 이행준수 여부 확인

2.4.8 AI 오픈소스 라이브러리의 보안성 및 호환성 점검
- 성능, 안정성, 커뮤니티, 문서화 수준 등을 고려한오픈소스 라이브러리 선정 기준 확인
- 코드 검사, 취약점 스캔, 침투 테스트 등 보안 취약점에 대한 발견 및 해결 절차 정의 확인
- 라이브러리 버전 확인, API 호환성 검증, 성능 테스트 등 호환성 점검결과 확인

2.4.9 AI 모델의 편향 제거
- 모델의 편향 제거 기법 적용을 위한 분석 내용이 식별됐는지 확인
- 편향성 수준에 대한 정량 분석 수행 가능 여부를 확인

2.4.10 AI 모델 공격에 대한 방어 대책 수립
- 모델 공격의 영향도 파악 유무 확인
- 가능한 모델 공격 유형 및 공격에 대한 방어 대책수립 여부 확인

2.4.11 AI 모델 명세 및 추론 결과에 대한 설명 제공
- 모델 명세가 구체적이고 충분하며 정확한지 확인
- 적용된 설명가능성 기법의 적절성 확인
- 추론 결과에 대한 설명이 이해하기 쉬우며, 신뢰도가 함께 제시되는지 확인

2.4.12 AI 시스템 구현 시 발생 가능한 편향 제거
- 데이터 접근 방식 구현 과정 등 소스코드에서 편향 발생 가능성 확인
- 사용자 인터페이스 및 상호작용 방식으로 인한 편향 확인

2.4.13 AI 시스템의 안전 모드 구현 및 문제 발생 알림 절차 수립
- 안전 모드의 작동 기준·조건과 작동 시 시스템 동작 상황을 확인
- 안전 모드 해제 기준 및 절차 확인
- 문제 감지 절차 및 식별 방법 확인
- 알림 방식 및 내용 확인
- 알림 수신자 및 사람의 개입 시 해당 역할 확인

2.4.14 AI 시스템의 설명에 대한 사용자의 이해도 제고
- 사용자 특성에 맞는 설명 방식이 적용됐는지 확인
- 설명 내용이 명확하고 간결하며, 전문용어 사용을 최소화했는지 확인
- 시각 자료, 다양한 설명 방식 등을 활용해 이해도를 높였는지 확인

2.4.15 서비스 제공 범위 및 상호작용 대상에 대한 설명 제공
- 서비스 약관에 면책조항을 포함한 내용이 적절한지 확인
- 이용정책에 허용 가능한 사용과 금지된 사용 관련 내용이 적절한지 확인
- 개인정보보호 정책에 개인정보 수집 및 처리에 관한 내용이 적절한지 확인
- 실제 서비스 이용환경에서 상기 정책들이 올바르게 적용돼 있는지 확인

3. AI 신뢰성 인증 사례

올해 1월부터 TTA에서 제공하기 시작한 AI 신뢰성 인증을 통해, 현재까지 총 3건의 인증 획득 사례가 나왔다<표 3>. 모두 'AI 시스템'에 해당하는 제품으로 특정 용도에만 사용가능하다. 위험 분석 결과 각각 채용, 국방, 생체인식 분야에 활용되는 고위험군으로 분류되어 AI 거버넌스 체계 및 제품 신뢰성 시험을 거쳤다.

<표 3> AI 신뢰성 인증 제품 현황

기업명	제품명	대상분류	인증일자
㈜제네시스랩	뷰인터 HR v2.0	AI 시스템	2024-04-18
㈜엔플럭스	인공지능 융합 지뢰탐지 모듈 v1.0	AI 시스템	2024-04-08
㈜마크애니	스마트아이(실종자검색) v1.0	AI 시스템	2024-02-06

4. 맺음말

과학기술정보통신부는 2020년 '사람이 중심이 되는 「인공지능(AI) 윤리기준」'을 마련한 후 「신뢰할 수 있는 인공지능 개발 안내서(TTA)」와 「인공지능 윤리기준 실천을 위한 자율점검 표(KISDI)」를 통해 신뢰기반 AI 생태계 조성을 지속 적으로 도모해 왔다.

특히 올해부턴 국제표준과 디지털 규범에 근거한 'AI 신뢰성 인증'을 제공함으로써 국내 AI 기업이 EU AI법 등 글로벌 AI 규제에 신속하고 효율적으로 대응할 수 있도록 중점 지원하고 있다. AI 신뢰성 인증제도는 빠르게 발전하는 AI 기술 추이에 발맞춰 지속적으로 개선·발전시켜 나가 국내 AI 분야 대표 인증으로 자리매김할 수 있도록 하겠다.

AI 시대에 접어들면서, 규제와 혁신 간의 조화가 국가 경쟁력 강화에 무엇보다 중요한 시점이다. 최근 들어 최첨단 AI가 초래하는 부작용을 줄이고 예방하기 위한 과학적인 연구를 중심으로 'AI 안전'에 대한 관심이 높아지고 있다. TTA가 제공하고 있는 'AI 신뢰성 인증'이 AI 기술이 초래하는 부정적 영향으로부터 국민과 사회를 보호하는 가장 기본적이면서 중요한 안전장치 역할을 할 것으로 기대한다.

※ 본 연구는 과학기술정보통신부의 '인공지능신뢰성기반조성사업'의 일환으로 수행됨

[참고문헌]

[1] 'Seoul AI Business Pledge', 2024년 5월 22일, AI 서울 정상회의 2024

[2] 'Frontier AI Safety Commitments', 2024년 5월 21일, AI 서울 정상회의 2024

※ 출처: TTA 저널 제213호