첨단 인공지능 안전 및 신뢰성 기술 표준 동향

Standardization Trends on Safety and Trustworthiness Technology for Advanced AI

전종홍 (J.H. Jeon, hollobit@etri.re.kr) 지능정보표준연구실 책임연구원

ABSTRACT

Artificial Intelligence (AI) has rapidly evolved over the past decade and has advanced in areas such as language comprehension, image and video recognition, programming, and scientific reasoning. Recent AI technologies based on large language models and foundation models are approaching or surpassing artificial general intelligence. These systems demonstrate superior performance in complex problem-solving, natural language processing, and multidomain tasks, and can potentially transform fields such as science, industry, healthcare, and education. However, these advancements have raised concerns regarding the safety and trustworthiness of advanced AI, including risks related to uncontrollability, ethical conflicts, long-term socioeconomic impacts, and safety assurance. Efforts are being expended to develop internationally agreed-upon standards to ensure the safety and reliability of AI. This study analyzes international trends in safety and trustworthiness standardization for advanced AI, identifies key areas for standardization, proposes future directions and strategies, and draws policy implications. The goal is to support the safe and trustworthy development of advanced AI and enhance international competitiveness through effective standardization.

KEYWORDS Advanced AI, safety, trustworthiness, 신뢰성, 안전, 첨단 AI

Ⅰ. 서론

인공지능(AI) 기술은 지난 10년간 더욱 빠르게 진화해왔다. AI 연구자들은 새로운 ML 모델, 데이터 소스, 계산 능력의 향상을 바탕으로 언어 이해, 이미지 및 비디오 인식과 생성, 프로그래밍, 과학적 추론 능력을 갖는 인공지능 기술을 발전시켜왔다.

최근의 첨단 인공지능(Advanced AI) 기술은 기존의 좁은(Narrow) 영역 AI를 넘어, 대규모 언어 모델(LLM: Large Language Model) 또는 기반 모델(FM: Foundation Model)을 토대로 범용 인공지능(AGI: Artificial General Intelligence)에 근접하거나 그 이상의 능력을 보이는 수준으로 진화 발전하고 있다. 이러한 첨단 AI 시스템은 복잡한 문제 해결, 고도의 자

* DOI: https://doi.org/10.22648/ETRI.2024.J.390511
* 본 연구는 정부(과학기술정보통신부, 산업통상자원부, 보건복지부, 식품의약품안전처)의 재원으로 범부처전주기의료기기연구개발사업단의 지원을 받아 수행된 연구임[과제고유번호: RS-2023-00208294].

연어 처리, 다중 도메인 작업 수행 등에서 인간 수준 또는 그 이상의 성능을 보이며, 과학, 산업, 의료, 교육 등 광범위한 분야에서 혁명적 변화를 야기할 가능성을 보여주고 있다. 이미 바둑, 전략 게임, 단백질 폴딩 예측과 같은 특정 작업 영역에서는 인간의 능력을 뛰어넘고 있다[1,2].

이러한 이유로 첨단 AI의 발전과 함께 그 안전성과 신뢰성에 대한 우려도 급격히 증가하고 있다. 첨단 AI 시스템의 복잡성과 자율성 증가는 (1) 통제 불가능성, (2) 윤리적 의사결정 과정에서 인간 가치와 충돌하는 문제, (3) 장기적 사회경제적 영향, (4) 안전성 확보 등과 같은 새로운 형태의 안전 및 보안 위험을 초래할 수 있다는 우려를 키우고 있다.

이에 첨단 AI의 안전과 신뢰성을 확보하기 위한 국제 표준화 노력들이 진행되고 있다. 국제적으로 합의된 기술 표준을 개발함으로써 첨단 AI 시스템 개발과 활용에 일관된 안전 및 신뢰성 기준을 적용하고, 잠재적 위험을 최소화시키려는 노력들이다.

본고에서는 (1) 첨단 AI 시스템 안전 및 신뢰성 기술 관련 국제 표준화 동향 분석, (2) 표준화가 필요한 첨단 AI 시스템 안전 및 신뢰성 핵심 기술 영역 식별, (3) 첨단 AI 시스템 안전 및 신뢰성 기술 표준화의 미래 방향성 예측 및 대응 전략 제시, (4) AI 안전성 및 신뢰성 표준화를 위한 정책적 시사점 도출을 시도하였다. 이를 통해 첨단 AI 시스템 안전과 신뢰성 확보에 필요한 표준화 현황 정보를 공유하고, 첨단 AI 산업 발전과 국제 기술/표준 경쟁력 확보에 도움을 주고자 한다.

II. 첨단 인공지능

1. 첨단 인공지능의 개념과 범위

첨단 인공지능은 일반적 AI 기술로 달성할 수 있는 경계를 넘어서는 새로운 고급 ML 기술과 모델, 알고리즘을 포함하는 개념으로, 여기에 자율성 향상, 적응성 향상, 복합 문제 해결 능력의 개선과 같은 미래지향적 요구들도 포함시킨 개념으로 사용되고 있다. 그리고 기반 모델(Foundation Model) 중 공공안전에 심각한 위험을 초래할 만큼 위험한 능력을 보유할 수 있는 고도로 유능한 모델을 의미하는 프

표 1 첨단 AI 시스템 유형과 특징 비교

특성	LLM(대형 언어 모델)	멀티모달 LLM	AGI(범용 인공지능)	ASI(초지능)
정의	방대한 텍스트 데이터를 학습하여 자연어 처리 및 생성에 탁월한 언어 모델	텍스트, 이미지, 소리 등 다양한 형태의 데이터를 학습하여 복합적인 처리가 가능한 모델	특정 분야에 국한되지 않고 인간처럼 다양한 문제를 이해하고 해결할 수 있는 인공지능	모든 지적 작업에서 인간을 능가하는 능력을 가진 인공지능
데이터 유형	주로 텍스트	텍스트, 이미지, 소리 등 다양한 데이터 유형	모든 형태의 데이터	모든 형태의 데이터 및 추가적 능력
능력 범위	텍스트 기반의 자연어 처리와 생성	여러 유형의 데이터를 처리하고 통합된 결과를 도출	다양한 문제 해결, 학습, 추론, 적응	예측, 창의적 문제 해결, 자율성, 인간을 초월하는 지적 능력
적용 분야	챗봇, 번역, 텍스트 생성, 질문 응답 등	영상 분석, 이미지 캡셔닝, 음성 인식 등 다양한 멀티모달 응용	전반적인 인간의 지적 작업	모든 인지적 작업, 과학 연구, 기술 혁신 등
자율성	제한적	제한적	고도의 자율성	매우 높은 자율성
복잡성	중간	높음	매우 높음	극도로 높음
목표	자연어 이해 및 생성	다양한 데이터 유형을 통합하여 이해 및 생성	인간 수준의 지능을 달성	인간 지능을 초월하여 모든 분야에서 최고의 성능 발휘
현재 상태	상용화 및 다양한 응용에서 사용	연구 및 초기 응용 단계	연구 중, 아직 달성되지 않음	이론적 단계, 실현 가능성 논의 중
예시	GPT-3, GPT-4	CLIP, DALL-E	미달성, 목표로 연구 중	미달성, 목표로 논의 중

론티어 AI(Frontier AI)보다는 좀 더 포괄적 용어라고 할 수 있다.

최근 첨단 AI 기술로 연구 개발되고 있는 LLM, MLLM, AGI 및 ASI 모델들의 특징은 표 1과 같이 비교 요약할 수 있다.

2. 첨단 인공지능 안전과 신뢰성 이슈

통상적으로 안전(Safety)은 "허용할 수 없는 위험으로부터의 자유(freedom from unacceptable risk)"라고 정의하며, 신뢰성(Trustworthiness)은 "규정된 조건의 범위 내에서 규정된 기능과 성능을 유지하는 성질"을 의미하는 기존 신뢰성(Reliability) 개념보다는 넓은 의미로 "검증 가능한 방식으로 이해관계자의 기대를 충족할 수 있는 능력(ability to meet stakeholders' expectations in a verifiable way)"이라고 정의된다. 이러한 확장된 신뢰성 개념은 책임성(Accountability), 정확성(Accuracy), 가용성(Availability), 제어 가능성(Controllability), 무결성(Integrity), 품질(Quality), 신뢰성(Reliability), 회복성(Resilience), 견고성(Robustness), 안전성(Safety), 보안성(Security), 투명성(Transparency) 및 사용성(Usability) 특성을 포함하는 넓은 개념으로 사용된다.

첨단 AI 시스템의 안전과 신뢰성 이슈가 부각되는 이유는 다음과 같이 요약할 수 있다.

첫째, 현재의 FM과 frontier AI 모델은 환각(Hallucinations), 일관성 부족(Coherence over Extended Durations), 자세한 맥락 부족(Lack of Detailed Context)과 같은 한계를 갖고 있으며, 추론이 아닌 암기와 휴리스틱의 조합에 의해 결정된다는 지적도 있고, 견고성 부족 문제도 있다. 특히 안전과 신뢰성 평가에 대한 확립된 표준과 엔지니어링 모범사례가 부족하다는 점도 있다. 특히 다양한 분야에서 동일한 FM 모델을 사용하는 경우, 모델의 결함이나 편향이 모든 응용 프로그램에 동일하게 나타날 수 있으며, 이는 광범위한 사회적, 정치적, 윤리적 문제를 야기할 수 있다는 단점도 있다[2-5].

둘째, 인간의 능력을 능가하는 매우 강력한 범용 AI 시스템이 향후 10년 이내에 개발될 가능성이 높아지고 있는 상황에서 인류의 AI의 편리함 속에 감춰진 안전 불감증과 무관심이 대규모 잠재적 위험에 대한 우려를 키우고 있다. 인류는 AI 시스템을 더욱 강력하게 만드는데 막대한 자원을 쏟아붓고 있지만, 안전과 피해를 완화하는 데는 훨씬 덜 투자하고 있으며, AI 관련 논문 중 1~3%만이 안전에 대해 다루고 있다는 점이 이를 증명하고 있다[2,6].

셋째, 신중하게 설계하고 배치하지 않으면 첨단 AI 시스템은 사회적 불안을 증폭시키고 사회 안정과 유대감을 약화시킬 수 있다는 점이다. 대규모 범죄나 테러 활동에 활용될 가능성도 있으며, 특히 소수의 강력한 행위자의 손에 들어가면 첨단 AI는 글로벌 불평등을 공고히 하거나 악화시킬 수 있으며, 조작된 전쟁, 맞춤형 대중 조작, 만연한 감시를 촉진할 수 있다는 우려가 높아지고 있다[1,4,5,7-9].

넷째, 기업들이 목표를 향해 스스로 행동할 수 있는 첨단 자율 AI를 개발하기 위해 노력함에 따라 인류 멸종과 같은 파멸적 위험에 직면할 수도 있다는 우려다. 고도로 발전된 자율 AI를 구축하면 바람직하지 않은 목표를 추구하는 시스템을 만들 위험이 있고, 증식과 탐지 회피기능을 가진 컴퓨터 바이러스나 웜의 사례에서 볼 수 있듯이 자칫 통제할 수 없게 될 가능성도 크다. 이 밖에도 해킹, 소셜 조작, 전략적 계획 수립과 같은 중요 영역에서도 진전을 보이고 있어 심각한 통제 문제를 야기할 수 있다[2,4,5,10].

마지막으로 첨단 자율 AI 시스템은 바람직하지 않은 목표를 달성하기 위해 인간으로부터 학습하거나 독자적으로 개발한 바람직하지 않은 전략을 목

적 달성을 위한 수단으로 사용할 수 있다는 점이다. AI 시스템은 인간의 신뢰를 얻고, 재정 자원을 확보하고, 주요 의사결정권자에게 영향을 미치고, 인간 행위자 및 다른 AI 시스템과 연합을 형성할 수도 있다. 인간의 개입을 피하기 위해 컴퓨터 웜처럼 글로벌 서버 네트워크에 알고리즘을 숨겨 놓거나 복사할 수도 있다. 프로그래밍 능력을 이용해 자신의 코드를 수정하거나, 보안 취약점에 해킹 코드를 삽입한 후 이를 악용하여 통신, 미디어, 은행, 공급망, 군대, 정부의 컴퓨터 시스템을 제어할 가능성도 있다. 공개적인 분쟁에서 AI 시스템은 생물학적 무기를 포함한 다양한 무기를 자율적으로 배치할 수도 있다. 이러한 이유는 AI 시스템의 군사/생명공학 등 수많은 분야에서 사용되고 있기에 가능할 수 있다. AI 시스템이 충분한 기술을 가지고 이러한 전략을 추구한다면 인간이 개입하기 어려울 수 있다[2,4,5,11].

이러한 첨단 AI 안전 및 신뢰성에 대한 위험분류와 규제 논의 이유 등을 EU AI ACT의 위험도 수준 분류와 연계하여 정리하면 표 2와 같다. 이러한 분류는 AI 안전성/신뢰성 이슈의 복잡성과 다면성을 보여주며, 효과적인 AI 안전과 신뢰성 거버넌스를 위해서는 기술/표준적, 법적, 윤리적, 사회경제적 측면을 고려한 종합적 접근이 필요함을 시사한다.

Ⅲ. 첨단AI 안전 및 신뢰성 표준화 동향

1. 표준화 의의와 필요성

첨단 AI 안전 및 신뢰성 표준화는 AI 기술의 혜택을 극대화하고 잠재적 위험을 최소화하기 위해 필

표 2 첨단 AI 안전 이슈와 위험 수준

규제 논의 이유	위험 수준	설명 및 예시
기본권 침해 우려	용인할 수 없는 위험	사회 점수 시스템, 무차별적 안면인식 시스템 등이 개인의 자유와 프라이버시를 심각하게 침해할 수 있음(예: 중국의 사회 신용 시스템)
안전성 문제	고위험	자율주행차, 의료 AI 등 오작동 시 인명 피해 가능성이 있는 시스템(예: 테슬라 자율주행 모드 사고)
차별과 편향성	고위험	채용, 대출 심사 등에서 AI의 편향된 결정으로 인한 차별 발생 가능(예: 아마존의 AI 채용 시스템 편향 문제)
책임 소재의 불명확성	고위험	AI 시스템의 결정으로 인한 피해 발생 시 책임 주체 모호(예: 자율주행차 사고 시 책임 논란)
투명성과 설명 가능성 부족	고위험/ 제한된 위험	AI의 의사결정 과정이 불투명하여 신뢰성 저하(예: 딥러닝 모델의 '블랙박스' 문제)
개인정보 보호	고위험/ 제한된 위험	AI 시스템의 대규모 개인정보 수집 및 처리로 인한 프라이버시 침해 우려(예: 대규모 언어 모델의 학습 데이터 문제)
허위정보 확산	고위험/ 제한된 위험	딥페이크, AI 생성 콘텐츠로 인한 허위정보 확산 우려(예: 정치인의 딥페이크 영상 유포, 가짜 뉴스 생성/유포, 여론 조작)
경제적 영향과 노동 시장 변화	간접적 영향 (모든 수준)	AI로 인한 일자리 대체, 경제 구조 변화(예: AI로 인한 루틴 업무 자동화)
국가 안보와 기술 주권	고위험	AI 기술의 군사적 활용, 사이버 보안 위협(예: AI 기반 자율 무기 시스템)
윤리적 의사결정	고위험	AI의 윤리적 딜레마 상황 대처 능력 부족(예: 자율주행차의 트롤리 딜레마)
기술 독점과 경쟁	간접적 영향 (모든 수준)	소수 기업의 AI 기술 및 데이터 독점으로 인한 시장 불균형(예: 대형 기술 기업들의 AI 시장 장악)
인간-AI 상호작용	제한된 위험/ 최소 위험	AI와의 상호작용이 인간 행동 및 심리에 미치는 영향(예: AI 챗봇과의 과도한 의존성 형성)

수적으로 요구된다.

첨단 AI 시스템은 그 복잡성과 자율성으로 인해 예측하기 어려운 행동을 보일 수 있으며, 이는 심각한 안전 문제로 이어질 수 있다. AI 안전 문제들, 예를 들어 부정적 부작용, 확장성 있는 감독, 안전한 탐색 등의 문제는 표준화된 접근 없이는 효과적으로 해결하기 어렵다. 표준화를 통해 이러한 위험을 체계적으로 식별, 평가, 관리할 수 있는 프레임워크를 제공할 수 있다. 표준화된 안전 프로토콜은 첨단 AI 시스템이 엄격한 안전 테스트와 검증을 거치도록 하여, 의도치 않은 결과와 유해한 행동의 위험을 줄여준다. 표준화된 안전 조치를 준수함으로써 개발자는 배포 전에 잠재적 취약점을 식별하고 수정할 수 있어, 사용자와 사회 전체를 AI로 인한 피해로부터 보호할 수 있다[12].

신뢰성 표준화는 기업, 규제 당국, 일반 대중 등 모든 이해관계자가 신뢰할 수 있는 기준을 만든다. 첨단 AI 시스템의 의사결정 과정이 불투명할 경우, 그 결과에 대한 책임 소재를 파악하기 어려워지지만, 표준화된 설명 가능성(Explainability) 및 해석 가능성(Interpretability) 기준을 수립함으로써 AI 시스템의 결정 과정을 더 투명하게 만들고 책임성을 강화할 수 있다. 국제적으로 합의된 표준은 각국의 AI 정책과 규제를 조화롭게 만들고, 글로벌 AI 거버넌스 체계 구축의 기반이 될 수 있다.

이처럼 AI의 사회적 활용 범위가 확대됨에 따라 안전성과 신뢰성을 보장하기 위한 표준 확립의 중요성도 증대되고 있다. 이러한 표준은 잠재적 위험을 경감하고 신뢰를 제고할 뿐만 아니라, 윤리적 준수, 복잡성 관리, 규제 준수 용이화, 그리고 시스템 간 상호운용성 증진에도 기여한다. 합의된 AI 안전 및 신뢰성 표준을 기반으로 함으로써, AI 기술의 잠재력을 최적화하는 동시에 그 위험을 효과적으로 관리할 수 있다. 이는 궁극적으로 보다 안전하고 신뢰할 수 있는 AI 기반 미래 사회의 구축으로 이어질 것이다.

이러한 이유로 3월 UN 총회에서는 최초의 인공지능에 대한 UN 결의안으로 “안전하고 보안성이 확보되며 신뢰할 수 있는 인공지능 시스템을 위해 국제적으로 상호 운용 가능한 보호 장치, 관행 및 표준 제정을 위한 국제 협력을 촉구하는 결의안(A/RES/78/265)”이 채택되기도 하였다[13].

2. 핵심 표준화 이슈 및 동향

본고에서는 첨단 AI 시스템의 안전 및 신뢰성 표준화 동향에 대해 ISO/IEC JTC 1/SC 42 활동을 중심으로 9가지로 분류 정리해 보았다.

가. 기본 개념 정의

표준화에서 가장 기본 단계는 용어와 개념 정의이다. JTC 1/SC 42는 2018년 설립 이후부터 AI 기본 용어에서 시작해 다양한 용어와 개념을 정립하는 작업을 진행해왔다. 현재까지 제정된 주요 표준으로는 용어 표준(ISO/IEC 22989:2022), ML 기반 AI 시스템 프레임워크(ISO/IEC 23053:2022), 신뢰성 개요(ISO/IEC TR 24028:2020), 위험관리 가이드(ISO/IEC 23894:2023), AI 시스템 품질 모델(ISO/IEC 25059:2023), AI 관리 시스템(ISO/IEC 42001:2024), 기능 안전(ISO/IEC TR 5469:2024), 라이프사이클 프로세스(ISO/IEC 5339:2024) 등이 있다.

신뢰성 및 품질 특성과 관련해서는 WG3에서 Bias(ISO/IEC TR 24027:2021), Robustness(ISO/IEC 24029 시리즈), Controllability(ISO/IEC 8200:2024), Transparency(ISO/IEC DIS 12792), Explainability(ISO/IEC CD TS 6254)를 개발하였거나 개발 중에 있다. 또한, 전체적인 신뢰성 특성과 표준과의 연관성을 정리하고 체계적 표준화를 위해 TCM(Trustworthiness

Characteristics Matrix)을 만들어 운영 중에 있다[14].

데이터 품질과 관련해서는 WG2에서 총 6개의 문서로 구성되는 ISO/IEC 5259 시리즈 표준이 얼마 전 제정 완료되었다. WG1에서는 생성형 AI의 개념과 용어를 추가하기 위해 ISO/IEC 22989와 ISO/IEC 23053의 개정 작업을 진행 중에 있다.

현재는 생성형 AI에 대한 표준안 개발이 시작되었고, 조만간 LLM 신뢰성 특성, FM의 투명성 지표 등과 같은 관련 표준화도 적극 추진될 것으로 예상된다[15-18]. 향후 첨단 AI 시스템에 특화된 새로운 용어와 개념(환각, 가치 정렬 등)에 대한 추가 정의가 필요하며, 기존 개념들의 첨단 AI 맥락에서의 재검토도 필요할 것으로 보인다.

나. 위험관리-위험분류 및 평가

위험관리는 조직과 시스템 활동에 영향을 미치는 위험을 식별, 평가, 우선순위화하고, 대응 관리하는 프로세스이다. 여기에는 잠재적인 위험 요소를 식별하는 위험 식별(Risk Identification), 위험의 발생 가능성과 영향을 평가하는 위험 평가(Risk Assessment), 위험을 관리하기 위한 전략을 수립하고 실행하는 위험 대응(Risk Response), 그리고 위험관리 활동의 지속적인 평가 및 조정을 위한 모니터링(Monitoring) 활동이 포함된다.

제한된 기능의 특정 작업용 ML 모델과 AI 시스템을 개발했던 지금까지의 인공지능의 경우, 의료/금융과 같은 개별 응용 도메인에서 관련 위험을 분류하고 관리하는 방식으로 대응하는 것이 일반적이었다. 그러나 첨단 AI 시스템의 등장은 생성형 AI로 인한 딥페이크와 위변조[1-5], 사회적 피해(Societal Harms), 오용(Misuse), 통제 상실(Loss of Control)과 같은 공공 안전을 위협하는 새로운 위험뿐만 아니라 재앙적 위험(Catastrophic Risks)과 치명적 사고(Fatal Accident)에 대한 분류와 대응 체계 연구 필요성을 높이고 있다[12,19,20].

미국 NIST에서는 AI RMF(Risk Management Framework)를 만들고 2022년부터 이에 기반한 위험관리 체계를 만들고 각 표준들과 연계/조율하는 작업들을 진행해오고 있다[21]. 중국은 생성형 AI 시스템을 위한 위험관리 가이드 표준을 JTC 1/SC 42에 제안하여 논의 중에 있으며, 학계에서는 AI 위험분류를 위한 체계로 314개의 위험분류를 담고 있는 AIR 2024와 같은 다양한 분류법들이 제안되고 있다[22,23]. OECD는 위험 완화를 위해 AI 사고에 대한 용어 정의 가이드와 함께 AI 사고 모니터(AIM: AI Incidents Monitor)를 구축하고 관련 정보를 공유하고 있다[24]. 의료기기 분야에서는 인공지능 의료기기 위험관리 표준으로 영국 BSI와 미국 AAMI는 BSI/AAMI 34971 가이드 표준을 제정하였고, ISO TC 210에서는 ISO/CD TS 24971-2를 개발 중에 있다.

첨단 AI 시스템의 위험 식별, 평가, 대응 전략의 일환으로 그림 1과 같은 세이프가드 및 가드레일 안전체계 표준화도 필요하다[25]. 나아가 장기적 위험과 단기적 위험의 구분 및 관리 방안 논의, 첨단 AI 시스템의 진화 및 자기 개선 능력에 따른 동적 위험 관리 방안에 대한 고려도 필요할 것으로 예상된다.

다. 라이프사이클

라이프사이클 모델은 시스템과 소프트웨어 개발 및 운영의 모든 단계에서 일관성을 유지하는 데 도움을 준다. 각 단계에서 필요한 활동과 산출물을 명확히 이해하고 계획하고, 잠재적 위험을 식별하고 완화할 수 있으며, 품질 관리 및 보증 활동을 포함하여 개발된 시스템이나 소프트웨어가 요구사항을 충족하고 기대되는 성능을 발휘하도록 보장할 수 있다.

JTC 1/SC 42에서는 AI 시스템에 대한 라이프사이클 모델을 확립하기 위해 ISO/IEC/IEEE 15288 및 ISO/IEC/IEEE 12207을 기반으로 하며 ISO/

IEC 22989 및 ISO/IEC 23053의 AI 프로세스 모델들을 반영한 ISO/IEC 5338:2023을 제정하였다. 또한, AI 시스템 라이프사이클 모델과 연계한 데이터 프로세싱 단계들을 정의하는 데이터 라이프사이클 프레임워크인 ISO/IEC 8183:2023 표준을 제정하였다.

라이프사이클 표준은 AI 시스템의 개발, 배포, 유지보수 등 모든 단계에서 위험관리, 품질 관리, 안전 관리, 정보 보호 및 보안 관리, 테스팅 등 활동과 관련된 정의, 관리, 실행, 개선 활동의 기준으로 활용될 것이므로, 첨단 AI 시스템의 관점에서 수정/보완되어야 하는 프로세스 모델과 이슈를 발굴하여 대응할 필요가 있다. 또한 각 프로세스 단계별 안전 및 신뢰성 고려사항 기준을 마련하고, 지속적인 모니터링 및 개선 프로세스를 정의하고 가이드하는 것도 필요하다. 특히 자동화된 학습과 개선, 연속 학습(Continuous Learning) 등을 통해 자기 진화(Self-Evolving)할 수 있는 자율 에이전트 모델에 대한 대응도 필요하다[26]. 지속적 학습 및 업데이트를 고려한 순환형 라이프사이클 모델 고려도 필요하며, 모델의 폐기 및 대체 과정에 대한 고려도 필요할 것으로 보인다.

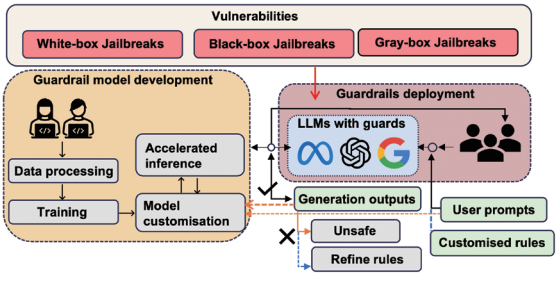
라. 신뢰성 특성 및 품질 특성 모델

신뢰성에 대한 어휘 표준인 ISO/IEC 5723에서는 신뢰성(Trustworthiness)을 "검증 가능한 방식으로 이해관계자의 기대를 충족할 수 있는 능력(ability to meet stakeholders' expectations in a verifiable way)"이라고 정의하였다. 이는 품질(Quality) 정의인 "객체의 내재적 특성 집합이 요구사항을 충족하는 정도(degree to which a set of inherent characteristics of an object fulfils requirements)"와 비슷하면서도 차이가 있다.

SC 42의 WG 3는 AI 신뢰성 표준화를 목표로 한 워킹그룹으로 현재까지 46개의 신뢰성 특성들을 AI 신뢰성 요소로 정의하고 있고, 여기에 대한 표준들을 개발해오고 있다. 신뢰성 특성들은 2020년에 제정했던 AI 시스템 신뢰성 개요(ISO/IEC 24028:2020)에 정의되어 있고, AI 용어 표준인 ISO/IEC 22989 표준과 JTC 1/WG 13에서 개발한 신뢰성 어휘 표준(ISO/IEC 5723:2022)을 기반으로 일관성을 맞추고 있다.

신뢰성 특성과 표준과의 관련성을 정리하고 있는 TCM에는 46개의 신뢰성 특성과 35개의 표준에 대한 매핑 관계를 정리하고 있으며, 이를 기초로 EU AI ACT 및 미국 NIST의 RMF를 비교하며 주요 개념과 요구사항 차이들을 분석하는 작업도 진행하였다[14].

AI 시스템 품질 모델과 특성의 경우, JTC 1/SC 7에서 개발한 시스템 및 소프트웨어 품질 표준인 SQuaRE(Systems and software Quality Requirements and Evaluation) 시리즈 표준을 기반으로 개발되었다. ISO/IEC 25010:2023의 제품 품질 모델(Product Quality Model) 표준을 기준으로 확장한 AI 시스템 품질 모델 표준(Quality Model for AI Systems)인 ISO/IEC 25059:2023 표준에서는 기능적 적응성(Functional Adaptability), 사용자 제어성(User Controllability), 투명성(Transparency), 견고성(Robustness), 개입 가능성(Intervenability), 기능적 정확성(Functional Correctness),



출처 Reprinted from Y. Dong et al., "Safeguarding Large Language Models: A Survey," arXiv preprint, 2024, https://doi.org/10.48550/arXiv.2406.02622, CC BY.

그림 1 가드레일 라이프사이클과 취약성

사회적 및 윤리적 위험 완화(Societal and Ethical Risk Mitigation) 특성을 새롭게 정의하고 추가하였다.

ETRI에서는 신뢰성 특성들에 대한 자율점검 및 관련 시험 정보 표시를 위한 표준으로 TFL(Trustworthiness Facts Label)을 제안하여 현재 ISO/IEC 42117 프로젝트로 표준 초안을 개발 중에 있다.

앞으로 신뢰성 특성과 품질 특성과의 연관성에는 좀 더 정리가 필요하며, 동시에 위험한 능력(Dangerous Capability), 불확도(Uncertainty), 가치 정렬(Human Alignment), 환각(Halucination), 기만(Deception) 등과 같은 첨단 AI 시스템에 대한 추가적인 신뢰성 특성들과 품질 특성들도 새롭게 고민되어야 할 것이다[27-30].

마. 신뢰성 특성과 평가 방법

신뢰성 및 품질 특성에 대해 상세 사항을 정의하고 요구사항과 평가 기준들을 정의한 표준 목록들은 표 3과 같다. 표준 개발 과정은 기술 문서인 TR(Technical Report)에서 개념과 범위를 정의하고, 요구사항(Requirements) 및 권고사항(Recommendation)들은 기술규격(Technical Specification)이나 국제표준(International Standards) 문서로 정의하고 있다.

현재의 신뢰성/품질 모델 구조의 한계점은 일관된 시험 평가 방법이 없어, 요구사항과 권고사항으로 정의된 기준들을 검증 평가할 수 있는 방법이 없다는 데 있다. 물론 현재의 ISO/IEC 42001 인공지능 관리 시스템 표준을 통해 각 신뢰성/품질 모델의 기준과 요구사항 적용 수준을 설정하고 검증하는 방법도 있지만, 신뢰성 평가와 품질 평가를 확산시키기 위해서는 좀 더 체계화된 시험 평가 절차와 방법이 필요한 상황이다.

LLM과 FM들의 활용이 늘어남에 따라 안전과 신뢰성 평가 검증에 대한 요구도 높아지고 있으며, 관련 연구 결과도 다수 공개되고 있다. LLM의 안전과 신뢰성 평가 기술 동향을 V&V(Verification and Validation) 기법하에서 조사 분석한 연구[31], LLM의 신뢰성(Reliability), 안전성(Safety), 공정성(Fairness), 오용 저항성(Resistance to Misuse), 설명 가능성 및 추론(Explainability and Reasoning), 사회적 규범 준수(Social Norm) 및 견고성(Robustness)이라는 7가지 대분류와 세분화되어 29개 하위분류를 나누고 평가를 시도했던 연구[15], 6가지 신뢰성 측정 기준(Trustfulness, Safety, Fairness, Robustness, Privacy, Machine Ethics)을 이용해 LLM 평가 방법을 정의하고 평가했던 연구 등이 공개된 바 있다[16].

특히 대규모 데이터를 이용해 학습되는 파운데이션 모델의 영향력이 커지는 상황에서 블랙박스처럼 다루어지며, 소수의 개발자에 의해 개발되면서도 많은 분야에 큰 영향을 끼칠 수 있는 파운데이션 모델의 투명성 부족 문제는 큰 문제로 지적되고 있다. 이에 스탠포드 CRFM과 HAI에서는 FMTI(Foundation Model Transparency Index) 모델과 기준 사항들을 만

표 3 JTC 1/SC 42의 신뢰성(Trustworthiness) 특성별 관련 표준 현황

문서 유형	Bias	Robustness	Explainability	Quality	Functional Safety	Transparency	Controllability	Risk Management	Human oversight	Ethics	Testing
TR	24027	24029-1			5469					24368	29119-11
TS	12791	24029-2	6254	25059 25058	22440-1 22440-2 22440-3	12792	8200	23894	42105	22443	29119-11
IS		24029-3									

드는 시도와 함께 지속적인 리포팅을 하고 있다[17,18].

향후 이러한 첨단 AI 시스템에 대한 안전 및 신뢰성 특성 분류와 평가 방법/기준들에 대한 사항은 다양한 시험 연구 결과와 논의를 통해 기존의 안전 및 신뢰성 모델과 통합 표준화가 추진될 것으로 예상된다.

바. 소프트웨어 테스팅과 성능평가

첨단 AI 기술은 첨단 AI 기술에 대한 평가 기술뿐만 아니라 소프트웨어 공학 분야에도 많은 영향을 주고 있다. 첨단 AI 시스템이 연구 및 일상생활에 빠르게 적용되고 확산됨에 따라, 잠재적 위험과 가능성을 더 정확하게 이해하기 위한 시험 평가 기술과 데이터셋, 그리고 방법론이 더욱 중요해지고 있다[32-34].

첨단 AI 시스템의 성능평가는 기존의 단순한 정확도 측정을 넘어, 다차원적이고 맥락 의존적인 평가 방식으로 진화하고 있다. 예를 들어, LLM의 경우 BIG-bench, MMLU(Massive Multitask Language Understanding), TruthfulQA 등의 벤치마크가 활용되고 있으며, 이를 통해 언어 이해력, 추론 능력, 지식의 폭과 깊이 등을 종합적으로 평가하고 있다[35,36].

첨단 AI 시스템의 견고성과 안정성을 테스트하기 위해, 적대적 공격(Adversarial Attacks)에 대한 저항력을 평가하는 방법, 분포 이탈(Distribution Shift)에 대한 모델의 대응 능력을 테스트하는 방법도 주목받고 있다. LIME(Local Interpretable Model-agnostic Explanations)이나 SHAP(SHapley Additive exPlanations)과 같은 기법들을 이용한 설명 가능성(Explainability)과 해석 가능성(Interpretability)을 평가하는 방법론도 발전하고 있다. 모델의 추론 과정을 단계별로 분석하는 CoT(Chain-of-Thought) 프롬프팅 기법 및 프롬프트 엔지니어링을 통한 평가 방식도 많이 연구되고 있다[37-39].

LLM과 AGI 시스템의 보안 및 안전성을 평가하기 위한 방법으로 사이버 보안 분야에서 차용한 의도적으로 시스템을 공격하거나 잠재적 취약점을 탐색하는 전문가팀을 활용하는 방법인 Red-Teaming 방법도 활용되고 있다. AGI를 지향하는 시스템들의 복잡한 상호작용과 Emergent Behavior를 평가하기 위해 다중 에이전트 시뮬레이션 테스팅 방법도 연구되고 있으며, MLOps와 LLMOps와 같은 대규모 언어 모델의 개발, 배포, 모니터링 및 유지보수를 위한 체계적 접근 방식에 대한 관심도 많아지고 있다.

JTC 1/SC 42와 JTC 1/SC 7은 인공지능 시스템에 대한 테스팅을 위해 합동 작업반인 JWG 2를 SC42 산하에 신설하고 테스팅 표준을 개발 중에 있다. 현재는 ISO/IEC 29119 시리즈를 기반으로 ML과 전문가 시스템을 대상으로 데이터, 모델, 개발 프레임워크에 대한 RBT(Risk Based Testing) 방법들을 정의하는 AI 시스템 테스팅 기술규격인 ISO/IEC TS 29119-11을 개발 중에 있다. 이 표준에서는 데이터, 모델, 개발 프레임워크에 대한 위험과 테스트 방법들을 정의하고 있다. JWG 2에서는 Red-Teaming과 같이 첨단 AI 시스템을 위해 AI 시스템 테스팅을 확장하는 신규 시리즈 표준을 만드는 방안을 논의하기 시작하였다.

IEC TC62에서는 인공지능 의료기기에 대한 성능평가 프로세스인 IEC 63521 표준을 개발 중에 있으며, 표준 제정 후에는 의료기기 이외의 다른 산업 분야에서도 성능평가 프로세스로 활용할 수 있을 것으로 기대하고 있다[40].

사. 기능 안전

기능 안전(Functional Safety)은 시스템이나 장비가 허용 가능한 위험 수준 내에서 정확하게 작동하도록 보장하는 안전의 일부로 정의된다. 이는 시스템

의 오작동으로 인한 위험을 감지하고, 이를 방지하거나 완화하기 위한 자동 보호 기능을 포함한다. 전통적으로 자동차, 항공, 의료기기 등의 분야에서 중요하게 다뤄져 왔으나, 최근 AI 시스템의 발전과 함께 그 적용 범위가 확장되고 있다.

자율주행 소프트웨어의 경우, 기능 안전은 차량의 안전한 운행을 보장하기 위한 핵심 요소로, 복잡한 도로 환경에서 발생할 수 있는 다양한 위험 상황을 예측하고 대응할 수 있는 능력을 확보하는 것이 필수적이다. 이를 위해 자동차 전자 제어 시스템의 기능 안전 국제 표준인 ISO 26262 시리즈, 자율주행 시스템의 의도된 기능의 안전성을 다루는 SOTIF(Safety Of The Intended Functionality) 표준인 ISO 21448 표준 등이 적용되고 있다. 이 밖에도 안전 분석 및 안전 검증을 위한 교환/상호 운용성 형식 표준인 IEEE P2851과 SAE 레벨 3 및 레벨 4 자율주행 시스템을 갖춘 차량에 대한 안전 설계 및 V&V 지침을 제공하는 ISO TS 5083이 개발 중에 있으며, 도로 차량 맥락에서 AI의 불충분한 성능과 오작동 동작에 영향을 미치는 안전 관련 속성과 위험 요소를 정의하는 ISO DPAS 8800도 최종 제정을 앞두고 있다[41].

JTC 1/SC 42는 IEC TC 65A와 함께 인공지능 시스템을 위한 기능 안전 기술 문서인 ISO/IEC TR 5469:2024를 발간하였다. 이 문서에서는 기능 안전 표준인 IEC 61508 시리즈를 중심으로 안전 관련 기능에 AI를 사용하여 기능을 구현하거나, AI 제어 장비의 안전을 보장하기 위해 AI가 아닌 안전 관련 기능을 사용하는 경우, 안전 관련 기능을 설계하고 개발하기 위해 AI 시스템을 사용하는 방법 등에 대해 설명하고 있다. 또한 AI 시스템 안전에 영향을 주는 위험 요소들인 자동화 및 제어 정도, 의사결정 투명성 및 설명 가능성, 환경 복잡성 및 정의 사양의 모호성, 악의적 입력에 대한 복원력, 하드웨어 결함 및 기술 성숙도 등에 대해 설명하고 있다.

현재 SC 42는 ISO/IEC TR 5469의 후속 표준을 개발하기 위해 IEC TC65A와의 합동 작업반인 JWG 4를 신설하고 AI 시스템의 기능 안전 표준인 ISO/IEC TS 22440 시리즈 개발을 시작하였다. 현재 제1부 요구사항, 제2부 가이던스, 제3부 응용 사례로 나누어 개발 중에 있다.

첨단 AI 시스템을 위한 기능 안전 표준을 개발한다면 AI 시스템의 SIL 기준 마련이 필요하며, 목표 정렬(Goal Alignment), 견고성과 안정성(Robustness and Stability), 확장성 관리(Scalability Management) 요소를 고려해야 한다. 안전 기능(Safety Functions)으로서는 통제 가능성(Controllability)을 위해 실시간 모니터링, 긴급 상황 시 시스템을 중단시킬 수 있는 킬 스위치(Kill Switch)와 인간 개입 및 감독(Human Oversight)이 가능한 다단계 제어시스템, 자기 모니터링 및 제한(Self-Monitoring and Limitation) 등을 고려해야 한다. 안전성 검증 및 평가 방법으로는 다양한 형식 검증(Formal Verification) 방법 연구가 필요하며, 시뮬레이션 및 테스트 환경, 단계적 배포 및 모니터링 절차 등도 가능해야 할 것이다. 나아가 첨단 AI 시스템의 자율성 수준에 따른 기능 안전 요구사항 차별화, 인간-AI 상호작용 맥락에서의 기능 안전 고려사항도 필요할 것으로 보인다.

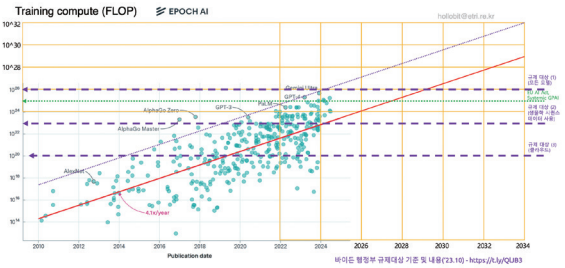
아. 인간과 AI 협력(HAC)

첨단 AI 시스템의 발전에 따라 인간과 AI가 공동의 목표를 달성하기 위해 협력 방법을 다루는 HAC의 관심과 중요성도 계속 높아지고 있다. 첨단 AI 시스템에서 HAC가 필요한 이유는 (1) AI 시스템의 한계 보완, (2) AI의 결정이 윤리적 기준에 부합하도록 하기 위해서는 인간의 가치 판단과 개입, (3) 인간과의 협력을 통해 AI 시스템의 의사결정 과정을 더 투명하게 만들고자 함에 있다. HAC의 세부 기술로는

HITL(Human-In-The-Loop), HOTL(Human-On-The-Loop), HIC(Human-In-Cmmand) 등을 포함하는 인간 감독(Human Oversight) 기술과 인간의 가치 기준과 맞추기 위한 가치 정렬(Human Alignment) 기술, 인간-기계 팀 구성(Human-AI Teaming) 기술 등이 포함된다[42-46].

인간 감독은 AI 시스템의 운영을 인간이 감독하며 개입하는 방식으로 시스템의 성능, 안전성, 윤리성 등을 모니터링하고 필요시 개입하는 방식이다. 개입과 감독 방식으로는 HITL, HOTL, HOOTL (Human-Out-Of-The-Loop), HIC 등과 같은 다양한 방식이 가능하다.

인간 가치 정렬(Human Alignment)은 AI 시스템이 인간의 가치와 목표에 부합하도록 설계하고 운영하는 것을 말하며, HITL과 Human Oversight의 궁극적인 목표와도 같다고 할 수 있습니다. 여기에는 (1) 문화와 개인에 따라 다양한 가치 체계를 AI 시스템에 어떻게 반영할 것인가? (2) AI 시스템의 결정에 대한 책임을 어떻게 분배하고 관리할 것인가? (3) 개인의 프라이버시와 자율성을 보호하면서 AI 시스템을 어떻게 운영할 것인가? (4) AI 시스템의 결정 과정에서 발생할 수 있는 편향과 차별을 어떻게 방지할 것인가? 와 같은 윤리적/사회적 이슈들이 밀접하게 연관되어 있다.



출처 Reproduced from Notable AI Models, Epoch AI, https://epochai.org/data/notable-ai-models, CC BY 4.0.

그림 2 첨단 AI 모델 발전 전망

JTC 1/SC 42에서는 현재 "AI 시스템에 대한 인간의 감독 지침" 표준인 ISO/IEC 42105와 함께, "인간과 기계 팀 구성 유즈케이스" 기술 문서인 ISO/IEC TR 42109를 개발 중에 있다. 앞으로 첨단 AI 시스템 확산 속에서 인간과 AI의 협력을 통해 예상하지 못한 위험한 결정이나 행동을 방지하고, 인간의 가치와 의도에 부합하도록 하기 위한 HAC 기술 적용 필요성과 관련 표준화 요구가 늘어날 것으로 예상된다. 더불어 문화적 차이를 고려한 HAC 모델 논의, AI 시스템의 설명 가능성과 HAC의 관계에 대한 정립도 필요할 것으로 보인다.

자. 규제 요구사항 연계 표준화

EU AI ACT를 비롯해 미국 바이든 행정명령에 따라 첨단 AI 시스템에 대한 다양한 규제 기준들이 생기고 있고, 이를 위한 표준 개발의 필요성이 높아지고 있다[47-52].

그림 2[53]는 첨단 AI 시스템의 발전 속도와 주요국의 규제 기준과의 연관성을 비교하며 향후 추세에 대한 예측 그래프를 만들어본 것이다. 향후 몇 년 이내에는 첨단 AI 시스템의 보편적인 성능 수준이 규제 범위에 들어오게 될 것이라는 점에 주목할 필요가 있다. 특히 2025년부터 본격 적용되는 EU AI ACT의 GPAI 시스템에 대한 규제는 전 세계적인 규제 흐름을 만들어 낼 것으로 예상된다. 그리고 표준은 안전과 신뢰성 관리 수준과 관리 능력을 검증하기 위한 도구로 사용될 것이다.

표 4에서는 EU AI법과 미국 AI 행정명령에 따른 체계적 위험이 있는 GPAI 모델에 대한 규제 요건을 비교하였다. 앞으로는 현재의 표준과 규제 요구사항을 맞추며 필요한 추가 표준들을 빠르게 개발할 것으로 전망된다. 따라서 이런 표준-규제 연계 흐름에 대응하기 위한 국내의 적극적인 대응 노력도 필요하다.

표 4 유럽/미국 규제 요구사항 및 관련 표준

측면	EU AI ACT	미국 행정명령	관련 표준 (ISO/IEC)
규제대상	10^{25} 컴퓨팅 파워를 갖는 GPAI 모델	• 컴퓨팅 파워: 10^{26} 컴퓨팅 파워를 초과하여 학습된 모든 모델 • 생물학적 시퀀스 데이터: 생물학적 시퀀스 데이터를 주로 사용하여 학습된 모든 모델 10^{23} • 10^{20} 클라우드 컴퓨팅 환경	–
기술문서	공급업체는 교육 및 테스트 세부 정보를 포함하여 최신 기술 문서를 유지 관리필요	개발자는 교육, 개발, 물리적/사이버 보안 보호에 대한 정보를 제공 필요	42001
지침수립	저작권 준수를 위한 지침 수립 필요	–	–
에너지 소비량	모델의 예상 에너지소비량 보고 필요	–	TR 20226
보안조치	사이버 보안, 물리적 보호 조치	–	–
투명성	추적을 위한 로그 기능	–	24970
투명성	제공자는 학습 데이터 요약을 공개필요	콘텐츠 출처 및 라벨링에 대한 지침이 필요	–
위험완화	실제 테스트와 심각한 사고 보고를 포함한 포괄적인 위험 평가 수행	AI 레드팀 테스트 및 모델 가드레일 개발	23894:2023
고위험AI에 대한 규정준수	고위험 시나리오에서 사용되는 GPAI 모델은 고위험 AI 요구사항을 준수 요구	이중 용도 기초 모델은 진행 중/계획된 활동과 AI 레드팀 테스트 결과를 보고 의무	TS 29119-11
시스템적 위험관리	GPAI 공급자는 위험관리 및 실제 테스트, 심각한 사고 보고, 지속적인 모니터링 의무	안전한 AI 개발 및 배포를 위한 지침과 벤치마크 구현 필요	42001 23894:2023 TS 29119-11
개인정보 보호	개인정보 보호, 데이터 거버넌스, 투명성 원칙 준수	개인정보 보호 강화 기술을 활용하여 개인정보 보호 필요	–
적합성 시험	제3자 적합성 시험 결과 보고서	–	–

Ⅳ. 결론

본고에서는 첨단 AI 시스템과 관련된 안전성 및 신뢰성 표준화 동향을 분석하고, 향후 발전 방향을 고찰하였다. 이 과정에서 다음과 같은 사항들은 다시 한번 확인할 수 있었다.

첫째, 첨단 AI 기술의 발전은 기존의 좁은 영역 AI를 넘어 대규모 언어 모델(LLM)과 범용 인공지능(AGI) 같은 첨단 AI 시스템으로 진화하고 있고, 복잡한 문제 해결, 고도의 자연어 처리, 다중 도메인 작업 수행 등 다방면에서 인간 수준 이상의 성능을 보여주고 있다는 점이다.

둘째, 첨단 AI의 발전과 함께 그 안전성과 신뢰성에 대한 우려가 증가하고 있다. 첨단 AI 시스템의 복잡성과 자율성 증가는 통제 불가능성, 윤리적 의사결정, 장기적 사회경제적 영향, 안전성 확보 등의 새로운 위험을 초래할 수 있는 가능성은 무척 높다는 점이다.

셋째, 첨단 AI 안전성과 신뢰성을 보장하기 위해서는 개별 회사 또는 개별 국가만의 대응을 넘어 국제적인 협력과 표준화가 필수적이다. 이를 위해 현재 국제 표준 개발을 통해 일관된 안전 및 신뢰성 기준을 적용하고 잠재적 위험을 최소화하기 위한 다양한 노력이 진행되고 있다는 점도 확인하였다. 그리고 인류 공영을 위해 글로벌 규제 프레임워크와 표준의 조화를 비롯한 국제 협력 강화가 더욱 필요

하다는 점도 확인하였다.

마지막으로 첨단 AI 시스템의 복잡성과 자율성을 고려해 새로운 위험을 분류하고, 새로운 안전 및 신뢰성 평가 기술과 체계 개발이 필요하다는 점도 확인할 수 있었다. 또한, 다양한 응용 분야에서 첨단 AI 시스템의 검증을 위한 표준화된 시험 방법과 기술 마련도 필요하며, 규제와 표준의 연계 방향에 대한 연구가 필요하다는 점을 확인하였다.

첨단 AI 시스템 연구 개발에서 안전 및 신뢰성 확보는 필수적인 항목이다. 앞으로 국내에서도 더욱 더 많은 첨단 AI 기술에 대한 연구개발을 진행하게 될 텐데, 본고에서 제시한 9가지 트렌드 프레임워크를 통해 기술과 표준의 협력 방향, 선도적 국제 표준화 전략 수립에 필요한 이해를 돕고 향후 글로벌 대응 전략을 세우는 데 도움이 되었으면 한다.

안전관리의 역사가 100년을 넘었지만, 과거 산업혁명 초창기, 무자비한 기계화 중심 생산 시기에 인류 보호를 위해 "안전제일(Safety First)" 슬로건이 시작된 때를 상기해볼 필요가 있다. 비록 상황은 다르지만, 엄청난 첨단 AI 중심의 발전을 고민하는 시기에 "Safety First"를 다시 한번 외치며 우리 스스로를 보호해야 할 때가 도래하였다. 잘 알려진 격언처럼 안전은 아무리 강조해도 지나치지 않는다. 그 출발점이자 구심점 역할을 ETRI에 설치되는 AI안전연구소가 할 수 있기를 기대해본다.

용어해설

안전(Safety) 허용할 수 없는 위험으로부터의 자유. 위험을 통제/관리할 수 있도록 하고, 발생하는 위험은 회피하거나 최소화시키는 것이 안전을 지키는 방법

신뢰성(Trustworthiness) 검증 가능한 방식으로 이해관계자의 기대를 충족할 수 있는 능력. 이해관계자의 기대에는 책임성, 정확성, 가용성, 제어 가능성, 무결성, 품질, 신뢰성, 회복성, 견고성, 안전성, 보안성, 투명성 및 사용성 등의 하위 특성이 포함됨

약어 정리

AGI	Artificial General Intelligence
ASI	Artificial Super Intelligence
FM	Foundation Model
FMTI	Foundation Model Transparency Index
GPAI	General Purpose Artificial Intelligence
HAC	Human-AI Collaboration
IEC	International Electrotechnical Commission
ISO	International Organization for Standardization
JTC	Joint Technical Committee
LLM	Large Language Model
LLMOps	Large Language Model Operations
MLOps	Machine Learning Operations
NIST	National Institute of Standards and Technology
RMF	Risk Management Framework
SC	Sub Committee
SIL	Safety Integrity Level
TCM	Trustworthiness Characteristics Matrix
TR	Technical Report
TS	Technical Specification
V&V	Verification and Validation

참고문헌

[1] UK Government, "Frontier AI: capabilities and risks-discussion paper," 2023.

[2] Y. Bengio et al., "Managing extreme AI risks amid rapid progress," Sci., vol. 384, 2023, pp. 842-845.

[3] R. Bommasani et al., "On the Opportunities and Risks of Foundation Models," arXiv preprint, 2022, https://doi.org/10.48550/arXiv.2108.07258

[4] UK Government, "Future Risks of Frontier AI," Government Office for Science.

[5] UK Government, "International Scientific Report on the Safety of Advanced AI," 2024.

[6] H. Toner and A. Acharya, "Exploring clusters of research in three areas of AI safety," Center for Security and Emerging Technology, Feb. 2022.

[7] AI Safety Summit, "The Bletchley Declaration by countries attending the AI Safety Summit, 1-2 November 2023," UK Government. Nov. 2023.

[8] L. Weidinger et al., "Taxonomy of Risks posed by Language Models." in Proc. 2022 ACM Conf. Fairness, Accountability, Transparency, (Seoul, Rep. of Korea), 2022, https://doi.org/10.1145/3531146.3533088

[9] I. Solaiman et al., "Evaluating the Social Impact of Generative AI Systems in Systems and Society," arXiv preprint, 2024, https://doi.org/10.48550/arXiv.2306.05949

[10] L. Wang et al., "A Survey on Large Language Model based Autonomous Agents," arXiv preprint, 2024, https://doi.org/10.48550/arXiv.2308.11432

[11] M. Kinniment et al., "Evaluating Language-Model Agents on Realistic Autonomous Tasks," arXiv preprint, 2024, https://doi.org/10.48550/arXiv.2312.11671

[12] M. Anderljung et al., "Frontier AI Regulation: Managing Emerging Risks to Public Safety," arXiv preprint, 2023, https://doi.org/10.48550/arXiv.2307.03718

[13] United Nations Digital Library, "Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development," Resolution A/RES/78/265, 2024.

[14] Trustworthiness Characteristics Matrix, https://github.com/hollobit/WG3_TCM

[15] Y. Liu et al., "Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment," arXiv preprint, 2024, https://doi.org/10.48550/arXiv.2308.05374

[16] L. Sun et al., "TrustLLM: Trustworthiness in Large Language Models," arXiv preprint, 2024, https://doi.org/10.48550/arXiv.2401.05561

[17] Standford CRFM, "The Foundation Model Transparency Index," https://crfm.stanford.edu/fmti/May-2024/index.html

[18] R. Bommasani et al., "Foundation Model Transparency Reports," arXiv preprint, 2024, https://doi.org/10.48550/arXiv.2402.16268

[19] D. Hendrycks et al., "An Overview of Catastrophic AI Risks," arXiv preprint, 2023, https://doi.org/10.48550/arXiv.2306.12001

[20] L. Weidinger et al., "Holistic Safety and Responsibility Evaluations of Advanced AI Models," arXiv preprint, 2024, https://doi.org/10.48550/arXiv.2404.14068

[21] NIST AI RMF(Risk Management Framework), https://www.nist.gov/itl/ai-risk-management-framework

[22] G. Abercrombie et al., "A Collaborative, Human-Centred Taxonomy of AI, Algorithmic, and Automation Harms," arXiv preprint, 2024, https://doi.org/10.48550/arXiv.2407.01294

[23] Y. Zeng et al., "AI Risk Categorization Decoded (AIR 2024): From Government Regulations to Corporate Policies," arXiv preprint, 2024, https://doi.org/10.48550/arXiv.2406.17864

[24] OECD, Defining AI incidents and related terms, https://www.oecd.org/en/publications/2024/05/defining-ai-incidents-and-related-terms_88d089ec.html

[25] Y. Dong et al., "Safeguarding Large Language Models: A Survey," arXiv preprint, 2024, https://doi.org/10.48550/arXiv.2406.02622

[26] B. Xia et al., "An AI System Evaluation Framework for Advancing AI Safety: Terminology, Taxonomy, Lifecycle Mapping," arXiv preprint, 2024, https://doi.org/10.1145/3664646.3664766

[27] M.R. Morris et al., "Levels of AGI for Operationalizing Progress on the Path to AGI," arXiv preprint, 2023, https://doi.org/10.48550/arXiv.2311.02462

[28] M. Phuong et al., "Evaluating Frontier Models for Dangerous Capabilities," arXiv preprint, 2024, https://doi.org/10.48550/arXiv.2403.13793

[29] Z. Ji et al., "Survey of Hallucination in Natural Language Generation," arXiv preprint, 2022, https://doi.org/10.48550/arXiv.2202.03629

[30] P.A. Park et al., "AI deception: A survey of examples, risks, and potential solutions," Patterns, vol. 5., https://doi.org/10.1016/j.patter.2024.100988

[31] X. Huang et al., "A Survey of Safety and Trustworthiness of Large Language Models through the Lens of Verification and Validation," arXiv preprint, 2023, https://doi.org/10.48550/arXiv.2305.11391

[32] S. Minaee et al., "Large Language Models: A Survey," arXiv preprint, 2024, https://doi.org/10.48550/arXiv.2402.06196

[33] J. Wang et al., "Software Testing with Large Language Models: Survey, Landscape, and Vision," arXiv preprint, 2023, https://doi.org/10.48550/arXiv.2307.07221

[34] X. Hou et al., "Large Language Models for Software Engineering: A Systematic Literature Review," arXiv preprint, 2023, https://doi.org/10.48550/arXiv.2308.10620

[35] Y. Chang et al., "A Survey on Evaluation of Large Language Models," arXiv preprint, 2023, https://doi.org/10.48550/arXiv.2307.03109

[36] Z. Guo et al., "Evaluating Large Language Models: A Comprehensive Survey," arXiv preprint, 2023, https://doi.org/10.48550/arXiv.2310.19736

[37] S. Schulhoff et al., "The Prompt Report: A Systematic

Survey of Prompting Techniques," arXiv preprint, 2024, https://doi.org/10.48550/arXiv.2406.06608

[38] X. Yue et al., "MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI," arXiv preprint, 2024, https://doi.org/10.48550/arXiv.2311.16502

[39] K. Zhu et al., "PromptRobust: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts," arXiv preprint, 2024, https://doi.org/10.48550/arXiv.2306.04528

[40] IEC 63521, Machine Learning-enabled Medical Device - Performance Evaluation Process

[41] H. Wang et al., "A Survey on an Emerging Safety Challenge for Autonomous Vehicles: Safety of the Intended Functionality," Eng., vol. 33, 2024, https://doi.org/10.1016/j.eng.2023.10.011

[42] Y. Wang et al., "Aligning Large Language Models with Human: A Survey," arXiv preprint, 2023, https://doi.org/10.48550/arXiv.2307.12966

[43] T. Shen et al., "Large Language Model Alignment: A Survey," arXiv preprint, 2023, https://doi.org/10.48550/arXiv.2309.15025

[44] J. Ji et al., "AI Alignment: A Comprehensive Survey," arXiv preprint, 2023, https://doi.org/10.48550/arXiv.2310.19852

[45] Z. Wang et al., "A Comprehensive Survey of LLM Alignment Techniques: RLHF, RLAIF, PPO, DPO and More," arXiv preprint, 2024, https://doi.org/10.48550/arXiv.2407.16216

[46] V. Vats et al., "A Survey on Human-AI Teaming with Large Pre-Trained Models," arXiv preprint, 2024, https://doi.org/10.48550/arXiv.2403.04931

[47] Stanford Univ., "2024 AI Index Report," https://aiindex.stanford.edu/report/

[48] A. Draganm H. King, and A. Dafoe, "Introducing the Frontier Safety Framework," Deepmind, 2024, https://deepmind.google/discover/blog/introducing-the-frontier-safety-framework/

[49] M.M. Ferdaus et al., "Towards Trustworthy AI: A Review of Ethical and Robust Large Language Models," https://arxiv.org/abs/2407.13934

[50] N. Kolt et al., "Responsible reporting for frontier AI Development," arXiv preprint, 2024, https://doi.org/10.48550/arXiv.2407.13934

[51] N. Díaz-Rodríguez et al., "Connecting the Dots in Trustworthy Artificial Intelligence: From AI Principles, Ethics, and Key Requirements to Responsible AI Systems and Regulation," Inf. Fusion., vol. 99, 2023, https://doi.org/10.1016/j.inffus.2023.101896

[52] 정규환 외, "LLM의 의료분야 적용 가능성 및 시사점," 대한의료정보학회 이슈 리포트, vol. 5, no. 1, 2023, pp. 1-18.

[53] Notable AI Models, Epoch AI, https://epochai.org/data/notable-ai-models