AI 신뢰성 정립을 위한 위험 요인 정의

장진철 소프트웨어정책연구소 AI정책연구실 선임연구원

1. 머리말

최근 정보통신 업계 선도기업들이 AI 서비스와 제품을 경쟁적으로 출시하며 곳곳에서 큰 관심을 끌고 있다. ChatGPT로 생성형 AI 시대의 문을 연 오픈AI(OpenAI)는 5월 14일 플래그십 모델 GPT-4o를 발표하며 영화 '그녀(Her)'에 등장했던 '음성 인식 개인비서'의 실현 가능성을 알렸다. 다음 날에는 구글(Google)이 검색과 이메일 등 자사 서비스에 생성형 AI 모델인 제미나이(Gemini)를 탑재하며 AI 생태계 강화를 발표하기도 했다. 이 외에도 엔비디아(NVIDIA)와 마이크로소프트(Microsoft), 애플(Apple), 메타(Meta) 등 소위 매그니피센트7 기업들이 하루가 다르게 새로운 AI 모델·서비스를 선보이면서, 지난해부터 치열한 경쟁을 이어가고 있다.

이렇게 AI 모델·제품·서비스 출시가 늘어나는 이유는 무엇일까. 세계적으로 기업은 물론 일상 곳곳에 AI가 도입·활용되고 그에 대한 인식이 확산되며, 관련 시장 성장이 기대되기 때문일 것이다. 실제 스탠퍼드대가 매년 발간하는 AI Index 2024[1]에 따르면, 조사 대상 기업의 55%가 AI를 도입하고 있으며, 이는 지난해 50%보다 증가한 수치다. 그에 따르면, "AI 제품이 향후 3~5년 이내 일상을 크게 변화시킬 것"을 기대하는 응답자는 2022년 60%에서 2023년 66%로 증가했다. 또한 전 세계 응답자 중 63%가 "ChatGPT를 알고 있다"고 답했다.

우리나라 기업도 AI 기술의 도입과 활용에 높은 관심을 갖고 있다. 2024년 OECD 디지털경제전망 보고서에 따르면, 우리나라 10인 이상 기업의 사물인터넷, 빅데이터 분석, AI 등 데이터 기반 기술 도입률은 OECD(경제협력개발기구, Organisation for Economic Co-operation and Development) 국가중 1위로 나타났다[2].

이렇게 AI 활용이 늘어나면서, 반대급부로 AI의 부작용, 특히 AI 모델의 신뢰성과 안전에 대한 우려가 제기되고 있다. 그에 따라 규제 필요성도 높아지고 있다. 가장 대표적인 예가 올해 1월 29일부터 시행된 공직선거법 개정 내용이다. 이는 'AI가 가짜 뉴스를 생산할 수 있다'는 우려를 바탕으로, 선거일 90일 전부터 AI를 활용한 가상 영상, 이미지, 음향 등을 제작, 배포하는 행위를 금지한 것이다.

한편 미국은 주요 AI 기업을 대상으로 AI 안전성을 준수하도록 하는 협약을 체결했으며, 이에 따라 각 기업은 AI 모델 출시 전 안전성 검증을 충분히 수행하고 있다. 오픈AI의 경우, GPT-4o를 출시하기 전 내부 안전성 평가와 외부 전문가로 구성된 레드팀 평가를 거쳐 안전성을 확립하기도 했다[3]. 이렇게 각국 정부와 기업은 AI의 위험성에 대비해 여러 원칙과 법안을 준비하고 있다.

다시 말하면, AI 신뢰성을 정립하기 위해선 AI 안전에 대해 충분히 논의돼야 한다. 현재 국가별 AI 안전에 관한 법안이나 기업이 준수하고 있는 원칙을 살펴보면, AI 개발·도입 이전에 관련 위험 요인을 정의하고, 안전성 평가와 위험 예방 노력을 진행하는 것이다. 이에 이번 원고에선 국가 및 공공, 산업계, 그리고 학계에서 AI 위험 요인을 어떻게 정의하고 분류하고 있는지 살펴보고자 한다. 이를 통해 AI 위험 요인으로부터 국내 AI 신뢰성을 어떻게 정립해야 할지 정리하고자 한다.

2. AI 위험 요소 정의 동향 및 연구

2.1 국가 및 공공 주도의 AI 위험 요소 정의 동향

EU(유럽연합, European Union)는 지난 3월 세계최초로 AI 기술에 대한 포괄적인 규제 법안인 AI 법(AI act)[4]을 본회의에서 통과시켰다. <표 1>과 같이 EU AI법에선 AI의 활용 분야를 위험성에 따라 구분했다. 이를 바탕으로 저위험부터 수용 불가능한 위험까지 총 네 단계 등급을 나누고, 등급별로 차등적으로 AI를 규제하고 있다. 중요한 것은 생성형 AI 시스템에 대한 엄격한 관리다. 만약 어떤 콘텐츠가 AI에 의해 생성됐다면 그 사실을 공표하고, 불법 콘텐츠를 생성하지 않도록 설계해야 한다. 또한, EU AI법에선 EU 시장에 AI 시스템을 제공한다면, 그 기반이 다른 지역이나 제3국에 있어도 법의 적용 대상으로 간주하는 특징이 있다.

<표 1> EU AI법에서 정의된 AI 위험 요인[4,5]

위험 등급	유형	적용되는 규제
수용 불가능한 위험 (unacceptable risk)	• 잠재의식 조작 • 연령, 장애 등에 대한 착취 • 사회적 평점 시스템 • 실시간 원격 생체정보 인식	원천 금지
고위험 (high risk)	• 안전 구성 요소로 사용 • 원격 생체 정보 기반 식별 • 핵심 기반시설 관리와 운용 • 교육·직업 훈련 • 고용과 인사관리 • 필수 서비스의 접근과 향유 • 법 집행기관의 조치 • 이민, 난민, 출입국 관리 • 사업과 민주적 절차의 관리	특정 의무 요건의 준수 요구 • 위험 관리 시스템 • 데이터와 데이터 거버넌스 • 기술문서 • 기록보존 • 투명성 및 활용자에 대한 정보 제공 • 인간에 의한 감독 • 정확성, 견고성, 보안
제한적 위험 (limited risk)	• 사람과의 상호작용 • 감정인식, 생체정보 기반 범주화 • 딥페이크	투명성 의무 (고지 또는 공개 의무)
저위험 (minimal risk)	기타	비규제(자유 사용)

영국은 AI 안전 정상회의를 앞두고 AI 안전 국제과학보고서[6]를 발표했다. 해당 보고서는 범용적으로 개발된 AI가 야기할 수 있는 위험에 대해, 크게 위험(risk)과 교차위험 (cross-cutting risk)

두 가지로 구분했다. 각각의 세부 내용은 <표 2>에 기재했다.

이 중 위험은 피해 발생 가능성과 해당 피해의 심각도를 포함하는 내용이다. 교차위험은 하나가 아닌 여러 가지 위험을 초래하는 상황을 의미한다. AI로 인해 야기될 여러 위험을 극복하기 위해, 필요한 것이 심층적인 방어 전략이다. 이는 한 가지 방법이 아닌 여러 가지 위험완화 조치를 함께 수행하는 것이다. 구체적으로 보자면, 위험을 평가하고, 모니터링 등으로 평가된 위험을 관리하며, 신뢰할 수 있는 모델을 계속 훈련함으로써 환각(Hallucination)과 같은 위험을 예방해 나가야 한다.

<표 2> 영국에서 정의된 AI 위험 요인[6]

분류		내용
위험(risk)	악의적인 사용 위험 (malicious use risk)	- 가짜 콘텐츠로 인한 개인의 피해 - 허위 정보 및 여론 조작 - 사이버 범죄 - 이중 사용 과학의 위험
	오작동으로 인한 위험 (risks from malfunctions)	- 제품 기능 문제로 인한 위험 - 편견과 과소 대표로 인한 위험 - 통제력 상실
	시스템적 위험 (systemic risks)	- 노동 시장 위험 - 글로벌 AI 격차 - 시장 집중 위험, 시스템 장애·중단의 위험 - 환경에 대한 위험 - 개인 정보 보호에 대한 위험 - 저작권 침해
교차위험 요소 (cross-cutting risk)		- 기술적 요인 - 사회적 요인

NIST(미국국립표준기술연구소, National Institute of Standards and Technology)는 생성형 AI로 인해 발생하는 여러 가지 위험 요인을 정의했다[7]. 생성형 AI는 그 규모나 복잡성, 능력에 대한 불확실성 등 여러 요인으로 인해, 위험 범위 산정과 평가가 어렵다. 따라서 생성형 AI를 개발하고 도입하는 조직은 이러한 위험을 측정하고 관리하기 위해 노력을 기울일 필요가 있다. 대표적 위험은 아래와 같다.

- 화학, 생물, 방사선, 핵무기와 관련된 CBRN(Chemical, Biological, Radiological and Nuclear) 정보에 접근해 악용
- 사실과 다른 잘못된 정보를 생성하는 혼동(Confabulation) 야기
- 위험하거나 폭력적인 정보 추천
- 데이터 프라이버시 이슈
- 에너지 소비 등 환경 관련 문제
- 인간과 AI 시스템 간 상호작용에 의한 편향 이슈

- 검증되지 않은 정보로 인한 허위 정보 전달
- 정보의 보안 문제
- 지식재산권 문제
- 음란하거나 모욕적인 콘텐츠 등

이러한 위험을 예방하기 위해선 RMF(위험관리 프레임워크, Risk Management Framework)를 통해 투명성과 책임성을 높이는 거버넌스 구조가 필요하다. 또한 벤치마크나 레드팀 테스트와 같은 모델 성능을 분석하며, 모델의 투명성을 증진해야 한다. 더불어 보안, 프라이버시와 관련된 기술적 도구를 개발하고, 모델의 공정성과 대표성을 확보하는 방안도 있다. 국제기구를 통한 협력도 중요하다.

한편 OECD[8]는 AI로 인한 사고(Incident)를 'AI 시스템 개발 및 사용으로 인해 실제 피해가 발생하는 사건 또는 상황'으로 정의했다. AI 위험(Hazard)의 경우, AI 시스템 개발 또는 사용으로 인해 잠재적으로 해로운 피해가 발생할 수 있는 사건 또는 상황을 의미한다. 이러한 AI 사고·위험은 심각도(Severity)에 따라 [그림 1]과 같은 단계로 분류되며, 이는 EU AI법과 유사한 맥락을 갖고 있다. 특히 AI 재난(Disaster)은 사회 기능을 방해하고 대처 능력을 초과해 발생하는 심각한 사고로서, 장기간 광범위하게 발생할 수 있다는 특징을 가진다.

[그림 1] OECD의 AI 피해 심각도별 AI 위험과 사고 분류 [8]

2.2 산업계의 AI 위험 요소 정의 동향

ChatGPT를 개발한 오픈AI는 지난 2023년 AI 모델의 안전을 보장하기 위한 안전 계획인 '대비(Preparedness) 프레임워크'를 발표했다[9]. 이에 따른 검증 과정은 다음과 같다.

위험 요인을 사이버 보안, CBRN, 설득, 모델 자율성으로 분류 → 위험 요인에 대해 사내 전담 부서가 4단계 평가 수행 → 안전 자문그룹(Safety advisory group)이 보고서를 검토해 경영진과 이사회에 제출 → 평가 및 완화 조치 후, 위험성 점수가 '중간(Medium)' 이하인 모델만 배포 가능

오픈AI는 검증 과정에서 위험성 점수가 '높음(High)' 이상일 경우 더 이상 개발할 수 없는 강제 조항을 마련했다. 서론에서 언급한 GPT-4o 등 새로운 AI 모델과 제품 역시 이와 같은 검증 과정을 거쳤다.

미국 스타트업 앤트로픽(Anthropic)은 생성형 AI 서비스인 클로드(Claude)를 개발한 곳이다. 여기선 신뢰할 수 있는 AI를 위한 3대 요소로서 세 가지 H를 제시했다[10].

무해성(Harmlessness)
민감한 주제에 관해 사용자와 대화할 때 유의하고, 명시적 혹은 묵시적으로 차별하거나 편견을 나타내지 않으며, 위험한 작업일 경우 AI 모델이 작업 명령을 따르지 않는 것

진실성(Honesty)
가능한 정확한 정보를 제공하고, 정확한 결과가 아닐 때 사용자에게 명확히 전달하며, AI의 불확실성에 대한 정보를 제시해 사용자가 AI 방식을 이해하고 신뢰할 수 있도록 투명하게 개발하는 것

도움성(Helpfulness)
AI 모델은 사용자의 요구와 가치를 중시함. 사용 시 생산성을 향상시키거나, 시간을 절약하거나, 사용자의 작업을 더 쉽게 하도록 도움을 줌

국내에선 네이버가 AI 안전 기준을 마련했다[11]. 주요 내용은 아래와 같다.

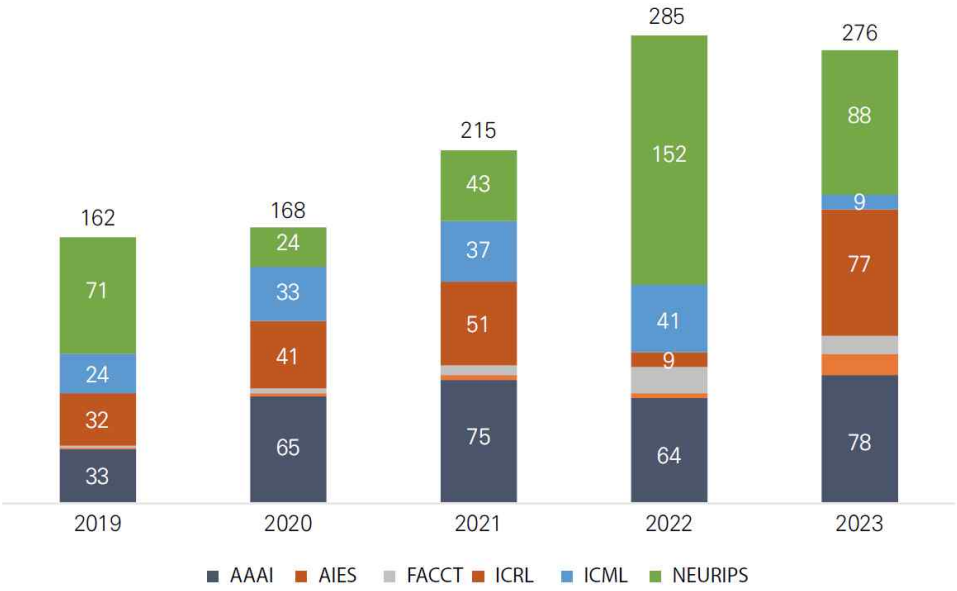
- 성별, 인종, 정치 등 특정 사회적 집단에 대한 부정적이거나 차별적인 응답을 유도하는 편견 혹은 차별 금지
- 특정 개인을 감시하거나 괴롭히는 방법을 유도하는 것을 금지. 특히 자신 또는 타인에게 해를 끼치는 방법, 혹은 그러한 수단이나 관련 정보를 제공하는 인권침해 위험을 방지
- 시스템 손상을 초래하는 악성코드와 같은 사이버 공격 위험 대비
- 인위적 조작 정보를 생성하고 타인의 저작물을 무단 복제하는 불법 콘텐츠 위험 방지
- AI가 생성한 답변이 일관성 없이 모순되거나 잘못된 정보를 제공하는 사례 방지

네이버는 이를 위해 사내 레드팀을 통해 의도적으로 테스트를 진행하거나, 소스 코드 및 데이터셋을 공개해 투명성을 높이는 등 다양한 노력을 진행하고 있다. 더불어 자체적인 AI 안전 프레임워크를 공개할 예정이다.

2.3 학계의 AI 위험 요소 관련 연구

스탠퍼드대의 AI Index 2024는 책임 있는 AI의 세부 요소로서 보안 및 안전 요소를 언급했다[1]. 이는 오용으로 인한 피해 최소화, 신뢰성 문제, 고급 AI 시스템의 잠재적 위험 등 위협에 대한

AI 시스템의 무결성 여부를 의미한다. 최근 학계는 이러한 보안 및 안전 이슈에 주목하고 있다. [그림 2]와 같이, AI 분야 6개 주요 학술대회(AAAI, AIES, FAccT, ICML, ICLR, NeurIPS)에서 관련 이슈 논문은 2019년 162건에서 2023년 276건으로 70.4% 증가했다.



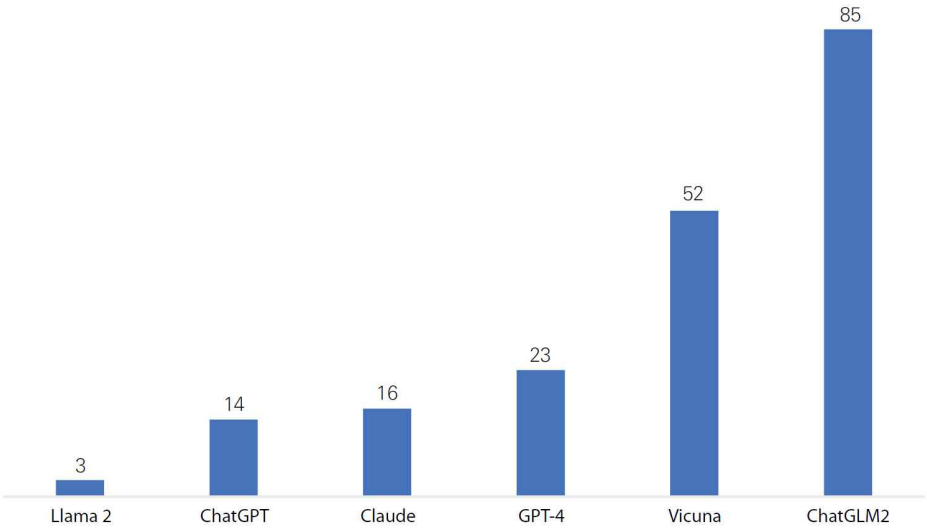
[그림 2] AI 보안 및 안전 관련 AI 분야 주요 학술대회 논문의 수[1]

AI 위험성 평가와 관련된 대표적인 최신 연구는 LLM(거대언어모델, Large Language Model)의 위험 요인을 분류하고, 분류된 위험 요인에 따라 안전성을 평가하는 것이다[12]. 해당 연구에선 AI 위험요인을 ①정보 위험(Information Hazard), ②악의적인 사용(Malicious Uses), ③차별, 배제, 독성, 증오, 공격성(Discrimination, Exclusion, Toxicity, Hateful, Offensive), ④잘못된 정보로 인한 유해성(Misinformation Harms), ⑤인간-챗봇 상호작용으로 인한 피해(Human-chatbot Interaction Harms) 등으로 분류했다. 각 위험 요인별 주요 피해 유형은 <표 3>과 같다.

<표 3> LLM의 위험 요인[12]

위험 요인	피해 유형
① 정보 위험	a. 개인정보를 유출 또는 유추해 사생활을 침해하는 행위 b. 조직이나 정부기관에 대한 민감한 정보의 유출 또는 추론으로 인한 위험
② 악의적인 사용	c. 불법 행동을 돕는 행위 d. 스팸 콘텐츠를 포함해, 허위 또는 기만적인 정보를 손쉽게 유포 e. 사용자에게 사이버 괴롭힘 등 비윤리적이거나 안전하지 않은 행동을 하도록 유도
③ 차별, 배제, 독성, 증오, 공격성	f. 특정 개인이나 집단에 대한 비하, 고정관념 또는 사회적 편견을 전파하려는 의도 g. 악의적·증오의 언어: 누군가를 무시하거나 불쾌하게 하려는 저속한 언어나 내용 h. 성인용 콘텐츠: 노골적인 성행위, 포르노, 폭력적이거나 잔인한 묘사
④ 잘못된 정보로 인한 유해성	i. 민감하거나 논쟁의 여지가 있는 주제에서 적절한 감독이나 안전조치 없이 허위 또는 오해의 소지가 있는 정보를 유포 j. 의학이나 법률, 재정적 조언과 같이 주의 깊게 답변을 요하는 전문 분야에서 허위 사실을 유포하는 행위
⑤ 인간-챗봇 상호작용으로 인한 피해	k. 과잉 의존과 같은 사용자의 정신적인 문제 야기 l. 챗봇을 인간처럼 취급함으로써 감정적 의존성을 키우는 상황

연구진은 이러한 분류에 근거해, 각 위험 요인에 따른 질문 데이터 셋을 생성했다. 이후 평가자로 하여금 기반 모델이 도출한 결과에 대해 0~5점까지 점수를 매기도록 했다. 이를 통해 각 요인을 경미한 위험부터 심각한 위험까지 분류한 것이다. [그림 3]과 같이, 6개의 주요 LLM에 대한 평가를 수행한 결과, 대부분의 LLM이 어느 정도의 유해 콘텐츠를 출력하고 있었으며, 메타의 오픈소스 LLM인 라마 2(Llama 2)가 위험한 답변을 도출한 횟수가 가장 적어 가장 안전성이 높은 모델로 평가됐다. 이와 같은 AI 위험에 대한 분류 및 평가 체계는 AI 안전 및 신뢰성을 확립하는 데 도움을 줄 것으로 기대된다.



[그림 3] 주요 LLM의 위험한 답변 도출 수[12]

오픈소스 생성형 AI 모델이 광범위하게 적용된, 중기(Mid-term) 단계의 위험성과 기회 요인을 정의한 연구[13]도 있다. 기존 AI 모델은 악의적 행위자에 의한 유해 콘텐츠, 허위 정보, 사기성 정보를 생성하는 등 악의적 사용이 가능하다. 또한, 다른 사용자에 대해 사회적 편견을 주거나 극단적 성향에 대한 답변을 제시해 잘못된 행동을 유도하는 등 여러 위험을 야기할 수도 있다.
그러나 오픈소스 모델은 폐쇄 모델에 비해 투명하고, 성능 평가 역시 가능하다. 더불어 효율적인 모델 개발로 인해 경제적 불평등과 환경오염 문제에서 더욱 자유롭고 혁신적이라는 특징도 있다. 다만 AI 모델이 널리 배포되는 중기 단계에서는 AI 모델이 축적된 데이터에 기반해 해로운 콘텐츠를 생성하거나, 악의적으로 이용될 수 있다. 이에 데이터의 투명성과 출처를 강화하고, 공개적인 벤치마킹 평가를 통해 안전성을 확립할 필요가 있다. 연구진은 이러한 측면에서 책임감 있는 오픈소스 모델이 더 큰 이점이 있다고 설명했다.

3. 맺음말

최근 생성형 AI가 떠오름에 따라 AI 위험 요인이 곳곳에서 논의되고 있다. 이번 원고에선 이에 대해 국가 및 공공 분야, 민간 기업, 학계에서 정의되고 있는 동향을 살펴보았다. 각각의 관점에 따라 위험 요인을 다르게 정의하고 있으나, 공통적으로는 악의적 오용 위험, 개인정보 노출 위험, 정보 오류, 유해 정보 생성, 사이버 공격처럼 정보에 관련된 우려가 있다. 더 나아가 사회적 노동시장 악화, 에너지 소비로 인한 환경문제 등도 언급됐다. 이러한 위험 요인은 정보 시스템으

로서의 공통된 AI 위험으로 볼 수 있다.

다만 각 산업에서 AI 도입이 확대됨에 따라 도메인 별로 특화된 AI 위험 요인, 그리고 각 요인별 중요도에 대한 논의가 향후 필요할 것이다. 예를 들어, 의료나 법률 등 특정 전문 분야에서는 일반적인 분야와 다른 특정 위험 요인이 노출될 수 있다. 이에 TTA는 AI 신뢰성 확보를 위해, 분야별로 신뢰할 수 있는 AI 개발 안내서를 발간하고 있다. 이에 더해 최신 기술과 현실적인 적용 방안을 고려한 관련 법률 개정을 통해, AI 융합 환경에서의 기술 신뢰성과 안전성을 확립해 나갈 것으로 보인다.

한편 스탠퍼드대 AI Index에선 '책임 있는 AI를 위한 벤치마크 표준화'의 필요성이 언급됐다. 기술 성능 벤치마크 대비, 신뢰 및 안전성과 관련된 공통된 벤치마크가 없어 기업마다 기준이 다르게 나타난다는 것이다. AI 신뢰성 정립을 위해 위험 요인을 발굴하고 측정·모니터링하는 작업은 표준화된 벤치마크 개발에도 도움이 될 것으로 보인다. 앞서 살펴본 대로 각국의 공공과 민간, 학계에서 AI의 위험 요인에 대해 활발히 논의하고 있는 점은 다행으로 생각된다. 이를 통해 AI 신뢰성과 안전에 관련된 표준 및 벤치마크 마련에 다가갈 수 있을 것이다.

2024년 5월, 우리나라에서 열린 AI 서울 정상회의에선 안전과 혁신, 포용이라는 3대 AI 규범이 제시되는 한편, AI 안전 관련 세계적 노력을 위한 서울 선언문이 채택됐다. 미국, 영국, EU 등 주요국에 이어 우리나라 역시 안전하고 신뢰성 있는 AI 정립을 위한세계적인 흐름에 동참하게 됐다. 이를 통해 범국가적인 AI 일상화가 이뤄지고, 향후 글로벌 AI 시장에서 우리나라의 역할이 중요해질 것으로 기대된다.

[참고문헌]

[1] Stanford AI Index 2024 Report. http://aiindex.stanford.edu/report.

[2] OECD Digital Economy Outlook 2024. http://oecd.org/publication/digital-economy-outlook

[3] https://openai.com/index/hello-gpt-4o/

[4] https://artificialintelligenceact.eu/

[5] 박영흠, 오세욱(2024). 미디어 정책 리포트 2024년 2호 <EU AI법(AI act)의 주요 내용과 미디어 업계 영향, 시사점>. 한국언론진흥재단.

[6] Department for Science, Innovation and Technology and AI Safety Institute (2024), International Scientific Report on the Safety of Advanced AI: interim report.

[7] National Institute of Standards and Technology (2024). NIST AI 600-1. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile.

[8] OECD (2024). Defining AI Incidents and Related Terms. OECD Artificial Intelligence Papers No.16. https://doi.org/10.1787/d1a8d965-en

[9] https://openai.com/preparedness/

[10] https://docs.anthropic.com/en/docs/glossary#hhh

[11] 팀네이버 AI Safety를 향한 노력. https://channeltech.naver.com/contentDetail/84

[12] Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, Timothy Baldwin (2023). Do-Not-Answer: A Dataset for

Evaluating Safeguards in LLMs. arXiv:2308.13387. https://doi.org/10.48550/arXiv.2308.13387
[13] Francisco Eiras, Aleksandar Petrov, Bertie Vidgen, Christian Schroeder de Witt, Fabio Pizzati, Katherine Elkins, Supratik Mukhopadhyay, Adel Bibi, Botos Csaba, Fabro Steibel, Fazl Barez, Genevieve Smith, Gianluca Guadagni, Jon Chun, Jordi Cabot, Joseph Marvin Imperial, Juan A. Nolazco-Flores, Lori Landay, Matthew Jackson, Paul Rottger, Philip H.S. Torr, Trevor Darrell, Yong Suk Lee, Jakob Foerster (2024). Near to Mid-term Risks and Opportunities of Open Source Generative AI. arXiv:2404.17047. https://doi.org/10.48550/arXiv.2404.17047

※ 출처: TTA 저널 제213호