AI 신뢰성 확보를 위한 우리나라 정보통신 기업의 현황과 시사점

장진철 소프트웨어정책연구소 AI정책연구실 선임연구원

1. 머리말

2022년 오픈AI(OpenAI)의 챗GPT(ChatGPT) 출시 이후, 생성형 AI를 중심으로 한 AI 기술은 검색, 커머스 등 전통적인 정보통신 산업은 물론 금융, 의료, 교육과 같은 다양한 분야 제품·서비스에 빠르게 내재화되며, 전 산업을 빠르게 변화시키고 있다. 맥킨지앤컴퍼니(McKinsey&Company) 조사[1]에 따르면, 2025년 전 세계 기업 조직 88%가 AI를 도입하고 있으며 이는 2024년 대비 10%p, 2017년 대비 68%p 증가한 수치다. 특히, 생성형 AI를 도입한 기업 조직의 비중은 2025년 79%로, 2023년 33% 대비 46%p 증가하며 가파른 성장세를 보이고 있다. 정보통신, 지식관리 산업 중 심층 연구, 서비스 관리 등의 영역에서 AI 에이전트 활용이 증가하면서, AI 활용 폭이 점차 넓어지고 있다.

특히, 우리나라 주요 정보통신 기업들은 초거대 언어모델 및 멀티모달 모델을 자사 플랫폼 및 네트워크 인프라에 결합해 신규 서비스와 비즈니스 모델을 창출하고 있다. 한편으론, AI 시스템 성능뿐 아니라 책임있는 AI 구현을 위한 안전성, 투명성 등 AI 신뢰성 확보가 핵심 과제로 부상하고 있다. 이용자와 사회가 신뢰하지 못하는 AI는 단기간에 빠르게 확산될 수는 있으나, 중장기적으로는 규제 리스크와 평판 리스크를 초래해 산업 경쟁력을 저해할 수 있다. 이에, AI 신뢰성은 기술적 선택이 아니라 필수적인 사안으로 인식되고 있다. 실제로, 2025년 공개된 SK텔레콤 유심 정보 유출 사고를 비롯한 국내 통신사의 보안 사고는 가입자 이탈 등 기업 평판에 영향을 끼쳤으며, 법적 조정 및 소송이 이어지고 있다.

국제적으로도 AI 신뢰성을 제도적으로 담보하려는 움직임이 본격화되고 있다. EU(유럽연합, European Union)는 세계 최초의 포괄적 AI 규제법인 EU AI법(AI Act)을 제정해, 생체인식, 신용평가, 직원 채용 등 사회 경제적 영향이 큰 영역의 AI를 고위험(High-Risk) 시스템으로 분류하고, 이들에 대해 데이터 관리, 거버넌스, 설명가능성, 기본권 영향평가, 등록 의무 등 엄격한 요구사항을 부과하고 있다[2]. 미국은 규제법 대신 AI 위험관리 프레임워크(AI RMF(Risk Management Framework))를 중심으로, 기업이 자율적으로 AI 위험을 식별, 평가, 완화할 수 있도록 지원하는 위험 기반 접근을 제시하며 신뢰할 수 있는 AI(trustworthy AI) 구현을 촉구하고 있다[3]. 이러한 규범과 표준은 글로벌 정보통신 기업들의 AI 개발과 운영 방식에 직접적인 영향을 끼치고 있으며, 국내 기업의 움직임과 관련 법률 제정에도 기여하고 있다.

우리나라도 정부와 민간을 중심으로 AI 신뢰성 확보를 위한 정책·제도 정비와 실무 지침 수립을 병행해 왔다. 특히 지난 2024년 말 국회 본회의를 통과한 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법(인공지능 기본법)」은 AI 기술·서비스 전반에 대한 기본 원칙, 책임 구조, 위험기반 접근,

신뢰성 확보 의무, 데이터 활용·보호체계 등을 국가 차원에서 종합적으로 규율한 법적 기반을 마련했다. 인공지능 기본법은 고영향 AI 정의 및 관리 기준, AI 신뢰성 확보 의무, 이용자 보호 조치, AI 사고 대응 및 책임, AI 규제 샌드박스 등 정부·기업·이용자 간 역할을 보다 명확히 설정함으로써, 정보통신 기업들이 준수해야 할 신뢰성 관리 기준의 방향성을 제시하고 있다. 이는 특히 정보통신 인프라, 데이터 처리 능력을 핵심 자산으로 하는 국내 기업들이 AI 신뢰성 확보를 중요한 과제로 인식하는 계기가 됐다.

특히 통신사, 인터넷 플랫폼 기업, 클라우드, 데이터센터 사업자 등 정보통신 기업들은 AI가 자사 서비스의 기반 인프라와 밀접히 결합돼 있다는 특성을 가지고 있기에, AI로 인한 오류·편향 등의 문제가 대규모 이용자와 사회 전반에 동시에 전파될 수 있다. 이는 단일 서비스 차원을 넘어 디지털 신뢰 인프라에 대한 국민적 신뢰를 훼손하고, 나아가 국가 차원의 디지털 경쟁력에도 부정적 영향을 줄 수 있다. 이런 점에서, 정보통신 기업의 AI 신뢰성 확보는 공공재적 성격을 가지는 중요한 과제로 볼 수 있다. 실제로, 최근 통신망 운용과 같은 여러 분야에서 자율적으로 AI 기술 역량을 활용함에 따라, AI 모델 신뢰성 확보가 통신업계에서는 안전과 기술 주권에 필수적 요소가 되고 있다.

이와 같은 국내외 동향을 종합해 보면 우리나라 정보통신 기업들에게 있어, AI 신뢰성 확보는 규제 준수 차원을 넘어 글로벌 시장 경쟁우위 확보와 사회적 수용성 제고를 위한 전략적 과제임을 알 수 있다. 이번 원고에선 세계 통신 분야 주요 AI 사고를 OECD AI 사고 데이터베이스(OECD AI Incident Database)를 토대로 살펴본다. 또한 우리나라 정보통신 기업의 AI 신뢰성 확보 노력에 대해선 국내 국가승인통계인 '2024년 인공지능산업실태조사'를 통해 알아본다. 결론으로는 이러한 국내외 동향을 바탕으로 향후 정보통신 업계에서 AI 신뢰성 확보를 위한 시사점을 도출하고자 한다.

2. 세계의 통신 분야 주요 AI 사고

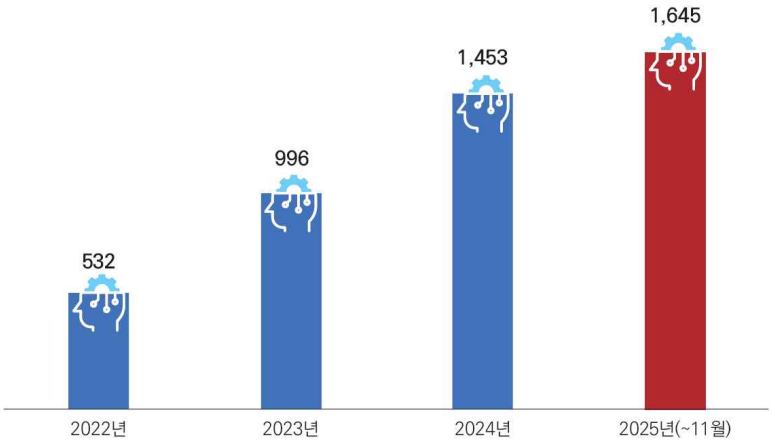
OECD(경제협력개발기구, Organisation for Economic Co-operation and Development)는 2019년 AI 원칙(AI Principles)을 채택[4]한 이후, AI 위험에 관한 국제적 정보 공유와 정책 기반 강화를 위해 다양한 글로벌 협력 인프라를 구축해 왔다. 그 중 하나인 AIM(AI Incidents Monitor)[5]은 전 세계에서 발생한 AI 관련 사고(Incident) 및 안전상 문제 사례를 수집·정리해 연구자, 기업, 정책 입안자가 참고할 수 있도록 하는 공개 데이터베이스다.

해당 데이터베이스는 뉴스 기사, 기업 보고서, 학술문헌, 시민 단체 보고서 등 공개적으로 확인 가능한 자료를 기반으로 구축된다. 수집된 사건은 표준화된 템플릿에 따라 정형화돼 기록되며, AI 모델 개발자와 사용자, 정책 실무자, 언론 등 다양한 이해관계자가 참고할 수 있도록 분류체계가 정비돼 있다. AIM은 처음엔 비영리 단체인 PAI(Partnership on AI) 주도로 구축됐으나, 이후 OECD AI Policy Observatory가 이를 통합해 국제적 표준 정보 플랫폼으로 발전시켰다. 현재는 AI 시스템이 사회에 미친 부정적 영향이나 위험 요소를 체계적으로 기록·분석함으로써 AI 신뢰성 정책을 수립하는 데 중요한 근거 자료로 활용되고 있다[6].

OECD AI 사고 데이터베이스는 AI 사고에 대해서, AI 시스템 개발·사용 또는 오작동으로 인해 직간접적으로 피해가 발생하는 사건·상황 또는 일련의 사건으로 규정했다. 여기엔 △ 개인 또는 집단의 부상 또는 건강 피해, △ 중요 인프라의 관리 및 운영 중단, △ 인권 침해 또는 기본권, 노동권, 지적 재산권을 보호하기 위한 관련 법률에 따른 의무 위반, △ 재산, 지역 사회 또는 환경에 대한 피

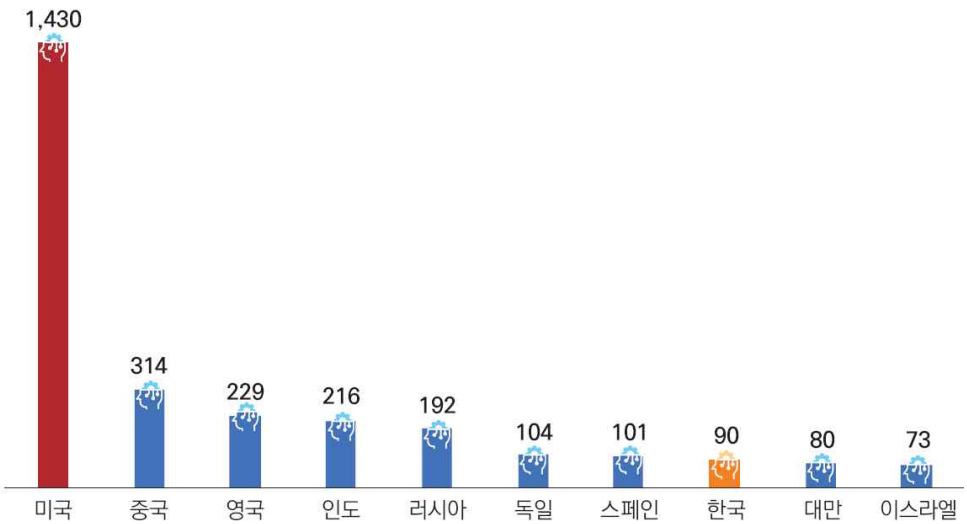
해가 포함된다. 즉, OECD AI 사고 데이터베이스는 기술적 오류뿐 아니라 사회적, 조직적 요인도 다루며, 개발자는 물론 운영자, 플랫폼 사업자 등 다양한 주체의 책임을 모두 고려해 거버넌스 체계를 정립하고자 한다.

필자는 통 신 산업 A I 사고 데이터 분 석을 위해 2022년 1월부터 2025년 11월까지의 사고 데이터를 추출했다(2025.12.1. 기준). 그중 통신 분야 추출을 위해 산업(Industry)군 중 IT 인프라스트럭처 및 호스팅(IT Infrastructure and Hosting), 디지털 보안(Digital Security) 두 가지를 선정했다.



[그림 1] 연도별 통신 산업의 AI 사고 건수(~2025.11.)

[그림 1]은 연도별 통신 산업에서의 AI 사고 수다. 기간 내 총 사고 건수는 4,626건으로 나타났다. 같은 기간 동안 AI 소프트웨어, 자율주행, 로봇 등을 포함한 전체 AI 사고 건수는 9,926건이며, 통신 분야 AI 사고 건수는 전체 절반 수준인 46.6%을 차지하고 있다. 통신 산업 AI 사고 중 67%가 2024년 이후 최근 2년 내에 발생하고 있으며, 2022년 532건에서 2025년(1~11월) 1,645건으로 약 3배 이상 증가했다. 통신 산업의 국가별 AI 사고 건수는 다음 [그림 2]와 같다. 미국이 1,430건으로 전체 사고의 31%를 차지했으며, 중국, 영국, 인도 등의 순으로 나타났다.



[그림 2] 국가별 통신 산업의 AI 사고 건수(~2025.11.)

통신 산업에서의 주요 AI 사고 사례는 다음과 같다. 사이버 범죄자들이 WormGPT, KawaiiGPT 등 AI 챗봇에 대한 접근 권한을 통해 피싱 이메일, 랜섬웨어, 악성코드 생성을 유도하고, 이로 인해 사이버 범죄 증가, 재산 및 지역 사회 피해가 유발된 사례[7]가 있다. 이란에선 AI가 사이버 공갈 협박에 악용되면서 범죄자들이 피해자들의 소셜 미디어 사진을 기반으로 가짜 이미지와 영상을 제작한 후, 현금을 요구한 사례[8]가 있다. 한편, 북미 지역에선 AI 데이터센터의 과도한 전기 소비가 겨울철 정전 위험을 증가시키지 않을까 우려[9]한다. AI 학습 및 추론량 급증은 데이터센터 전력 수요를 늘리고, 이는 극한의 추위 속에서 전력망 용량을 초과해 중요 인프라에 지장을 초래할 수 있다.

3. 우리나라 기업의 AI 신뢰성 노력

국내 기업의 AI 신뢰성과 관련한 노력을 살펴보는 정량적 방법으로는 인공지능산업실태조사를 활용하는 방법이 있다. 인공지능산업실태조사[10]는 우리나라 AI 공급 산업에 대한 정책 수립을 뒷받침하기 위한 통계다. 2019년 승인돼 현재까지 매년 AI 공급 기업을 대상으로 조사가 진행됐다. 해당 조사는 기본적인 기업 현황인 매출액, 인력, 수출 통계는 물론, 학습용 데이터 확보 현황과 같이 전반적인 AI 정책 수립을 지원하기 위한 문항으로 구성돼 있다. AI 기업의 신뢰성 노력 현황 관련 문항은 AI 안전 글로벌 이슈가 부상하기 시작한 2024년 처음으로 추가됐다. 관련 질문은 <표 1>과 같다.

<표 1> AI 신뢰성 및 안전 관련 2024년 인공지능산업실태조사 조사 문항

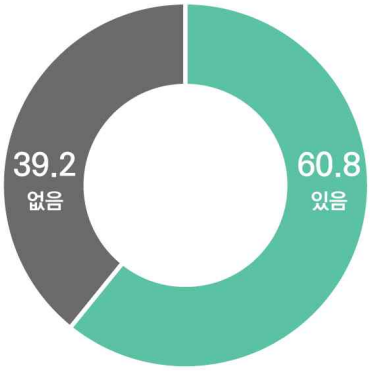
Q. 귀사는 인공지능 신뢰성 및 안전성 확보를 위한 활동과 관련된 전담조직(또는 인력)을 보유하고 있습니까?

① 예 ② 아니오

Q. 귀사에서는 다음의 인공지능 신뢰성 및 안전성 확보를 위한 활동(노력)을 수행하고 있습니까?

구분	설명 및 예시	선택
① 제3자 검증 (외부전문기관)	KTR(한국과학융합시험원구원), 와이즈스톤 ICT시험인증연구소 등	□ 예 □ 아니오
② 연구개발(R&D)	XAI 등 신뢰성 기술 연구	□ 예 □ 아니오
③ AI 모델 검증 및 테스트	자체 개발 또는 응용하고 있는 AI 모델의 정확성·신뢰성 검증을 위한 테스트	□ 예 □ 아니오
④ 데이터 품질 관리	자체 개발 또는 응용하고 있는 AI 모델 학습에 사용되는 데이터의 품질관리(데이터 편향성/공정성 검토)	□ 예 □ 아니오
⑤ 보안 관리	자체 개발 또는 응용하고 있는 AI 시스템의 보안 강화를 위한 활동(사이버 공격 및 방어 메커니즘 도입)	□ 예 □ 아니오
⑥ 투명성/설명 가능성 제고	AI 모델의 의사결정 과정을 설명할 수 있는 메커니즘 마련, 고객 또는 사용자에게 AI 의사결정 과정에 대한 정보 제공	□ 예 □ 아니오
⑦ 모니터링 및 리스크 관리	운영 중인 AI 시스템에 대한 모니터링 및 안전성 평가, 사용자 피드백 수집, AI 시스템 오류에 대비한 대응 계획 사전 마련 등	□ 예 □ 아니오
⑧ 교육 및 훈련	AI 신뢰성 및 안전성에 대한 내부 직원 교육, 훈련 프로그램 운영 등	□ 예 □ 아니오
⑨ 없음		

[그림 3]은 <표 1> 문항에 대한 응답 결과다. AI 신뢰성 및 안전성 확보를 위한 활동과 관련된 전담조직(또는 인력) 보유 여부는 '있음' 60.8%, '없음' 39.2%로 나타났다. <표 2>와 같이, 기업 규모별로는 종사자 규모가 클수록, 매출액 규모가 클수록 신뢰성 및 안전성과 관련한 전담조직을 보유하고 있는 비율이 높았다.

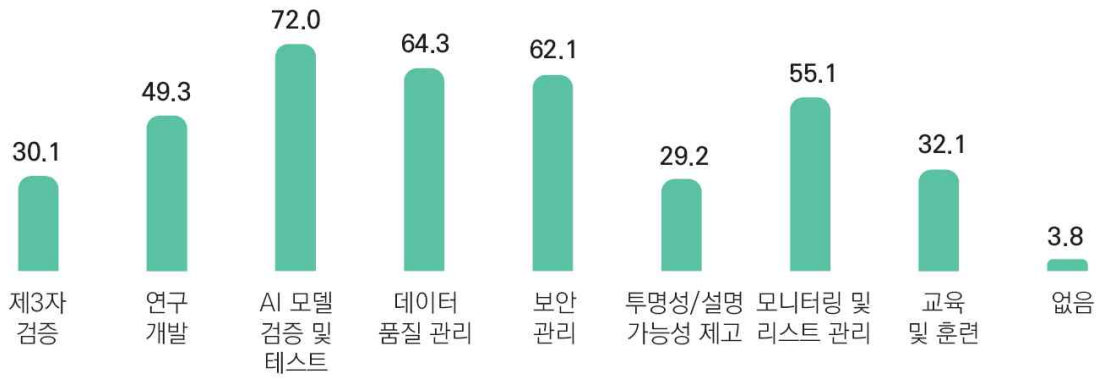


[그림 3] AI 신뢰성 및 안전성 확보를 위한 전담조직(또는 인력) 보유 여부

<표 2> AI 신뢰성 및 안전성 확보를 위한 전담조직(또는 인력) 보유 여부 상세결과

구분		사례 수	있음	없음
전체		2,517	60.8	39.2
종사자 규모	1,000인 이상	31	86.2	13.8
	100~1,000인 미만	239	64.5	35.5
	10~100인 미만	1,204	69.0	31.0
	10인 미만	1,043	49.7	50.3
매출액 규모	1,000억 이상	64	77.6	22.4
	100억~1,000억 미만	304	64.1	35.9
	10억 이상~100억 미만	904	66.5	33.5
	1억 이상~10억 미만	755	58.2	41.8
	1억 미만	490	50.0	50.0

[그림 4] 및 <표 3>과 같이, 기업의 AI 신뢰성 및 안전성 확보를 위한 활동(노력)으로는 'AI 모델 검증 및 테스트'가 72.0%로 가장 높았으며, '데이터 품질 관리(64.3%)', '보안 관리(62.1%)', '모니터링 및 리스크 관리(55.1%)'가 뒤를 이었다.



[그림 4] 기업의 AI 신뢰성 및 안전성 확보를 위한 활동(노력)

<표 3> 기업의 AI 신뢰성 및 안전성 확보를 위한 활동(노력) 상세결과

구분		사례 수	① 제3자 검증 (외부전문 기관)	② 연구 개발 (R&D)	③ AI 모델 검증 및 테스트	④ 데이터 품질 관리	⑤ 보안 관리	⑥ 투명성/ 설명 가능 성 제고	⑦ 모니 터링 및 리스크 관리	⑧ 교육 및 훈련	⑨ 없음
전체		2,517	30.1	49.3	72.0	64.3	62.1	29.2	55.1	32.1	3.8
종사자 규모	1,000인 이상	31	27.3	68.6	82.4	65.1	75.4	47.8	79.2	79.2	14.1
	100~1,000인 미만	239	29.2	59.5	80.3	55.4	70.6	35.1	48.4	44.8	4.5
	10~100인 미만	1,204	31.5	52.1	76.8	70.3	64.3	31.8	60.1	36.4	2.6
	10인 미만	1,043	28.8	43.2	64.3	59.5	57.1	24.3	50.3	22.8	4.8
매출액 규모	1,000억 이상	64	16.8	56.7	73.9	43.9	70.8	32.2	63.4	63.6	19.0
	100억~1,000억 미만	304	37.0	59.4	81.8	63.1	74.0	38.3	51.3	44.0	1.5
	10억 이상~100억 미만	904	28.0	52.2	74.6	70.4	64.4	31.1	58.4	36.2	3.3
	1억 이상~10억 미만	755	26.7	46.6	68.5	66.1	56.0	27.3	56.7	27.7	4.8
	1억 미만	490	36.7	40.9	66.3	53.7	58.6	22.7	48.1	19.5	2.8

우리나라 주요 통신 기업의 AI 신뢰성 확보를 위한 노력 사례는 다음과 같다[6]. 먼저 SK텔레콤은 2024년 AI 거버넌스 원칙인 'T.H.E AI'를 공개하고 안전 관련 전담 조직을 신설했다[11]. 'T.H.E. AI'는 '통신기술 기반(by Telco)', '사람을 향한, 사람을 위한(for Humanity)', '윤리적 가치 중심(with Ethics)'이라는 원칙을 따르는 AI 거버넌스를 상징한다. 이는 SK텔레콤 고객에 대한 신뢰와 안전을 기반으로, 잠재적 AI 위험으로부터 고객을 보호하고, 다양성·평등·공정 가치·인류 복지를 증진하면서, AI 의사결정의 투명성·개인 정보보호·윤리적 책임을 약속하기 위한 체계다. 이어, SK텔레콤은 AI 거버넌스 기본 원칙을 구체화한 AI 행동규범을 수립하고 사내 전 구성원이 실천 서약에 참여한 바 있다.

KT는 2024년 인간 존엄성과 공공성 증진이라는 기본 가치를 기반으로 AI 윤리원칙을 제정[12]했다. KT의 5가지 핵심 윤리 원칙(ASTRI)은 북극성 길잡이별(Astri)처럼 KT의 모든 Responsible AI 논의의 방향을 가리키는 이정표 역할을 수행하며, 이 원칙들은 글로벌 Responsible AI 흐름뿐 아니라 우리나라에서 통용되는 사회·문화적 가치들을 반영했다. 신뢰할 수 있는 AI 추진을 위한 체계로써, KT는 사내에 RAIC(책임 있는 인공지능 센터, Responsible AI Center)를 설립하고, 윤리성·신뢰성을 갖춘 AI를 제공할 수 있도록 책임 있는 AI 프레임워크를 연구개발·수립하는 조직을 구성했다. KT RAIC는 AI 기술이 사용자에게 유익한 가치를 제공할 수 있도록 관련된 위험을 최소화하기 위한 연구를 수행하며, AI 시스템의 취약점을 분석해 위험 수준에 대한 관리체계를 구축하기 위해 노력하고 있다. KT는 AI 제품 서비스를 기획·검증하는 단계에서 자사 모든 AI 제품 및 서비스가 Responsible AI 윤리원칙을 준수할 수 있도록 내부평가 절차를 수립하고 있다. 이를 통해 KT는 각 AI 개발 단계에서 발생할 수 있는 리스크를 정의하고, 이를 완화하기 위한 다양한 활동을 시행하고 있다. KT는 AI 리스크 분석 및 완화를 위해 리스크 정의, 평가 및 완화, 배포 및 모니터링 절차를 수행하고 있다.

4. 맺음말

이번 원고에선 OECD AI 사고 데이터베이스를 토대로 세계의 통신 분야 주요 AI 사고 살펴보고, 이어 우리나라 정보통신 기업의 AI 신뢰성 확보 노력에 대해 2024년 인공지능산업실태조사 조사 결과를 바탕으로 알아보았다. AI 사고에서 통신 기업이 차지하는 비중은 절반 가량으로, 그 중요도는

매우 높다고 할 수 있다. 또한 AI 기술 발전에 따라 발생한 글로벌 AI 사고 현황을 정리한 OECD 데이터베이스를 살펴보면서, AI 사고의 연도별·국가별 주요 통계를 통해 AI 사고의 지속적 증가를 확인했다. 이에, 정보통신 기업에서 AI 신뢰성 확보를 위한 노력이 중요하다고 할 수 있으며, 국내 기업들 역시 AI 사고에 대비하기 위한 전담조직 및 프레임워크를 운영하고 있다.

2024년 인공지능산업실태조사[10] 결과에서, 기업은 안전과 신뢰성을 위한 활동으로 모델 검증 및 테스트, 데이터 품질 관리에 주력하고 있음을 알 수 있다. 네트워크 운영의 공정성과 신뢰성을 확보하는 것이 무엇보다 중요하며, 이를 위해선 AI 학습 데이터의 품질 및 편향성이 중요하다고 할 수 있다. 한편으론 최근 잦아진 민간 기업·정부 시스템 해킹 및 재난 피해 사례에서와 같이, 보안 관리를 비롯한 재난 대응에도 노력이 필요한 상황이다. 이에, AI 기반 통신 시스템의 신뢰성을 보증하기 위한 종합적인 노력이 필요할 것으로 보인다.

[참고문헌]

[1] McKinsey (2025), The state of AI in 2025: Agents, innovation, and transformation. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

[2] 박강민, 장진철, 안성원 (2024), 유럽연합 인공지능법(EU AI Act)의 주요내용 및 시사점. 소프트웨어정책연구소 이슈리포트. https://spri.kr/posts/view/23764

[3] 이해수, 유재흥 (2025), 미국의 AI안전·신뢰성 정책 추진 현황과 시사점. 소프트웨어정책연구소 이슈리포트. https://spri.kr/posts/view/23815

[4] OECD(2019), AI Principles. https://www.oecd.org/en/topics/ai-principles.html

[5] AIM: AI Incidents and Hazards Monitor. https://oecd.ai/en/incidents

[6] 장진철, 안성원, 김예진(2025), AI 신뢰성 및 윤리 제도 연구. 소프트웨어정책연구소 연구보고서. https://spri.kr/posts/view/23864

[7] Malicious AI Chatbots Facilitate Cybercrime via Dark Web. https://oecd.ai/en/incidents/2025-11-25-92a6

[8] AI-Generated Deepfakes Used for Cyber Extortion in Ardabil. https://oecd.ai/en/incidents/2025-11-26-3c97

[9] AI Data Centers Raise Risk of Winter Power Outages in North America. https://oecd.ai/en/incidents/2025-11-23-5549

[10] 과학기술정보통신부(2025). 2024년 인공지능산업실태조사. https://spri.kr/posts/view/23852

[11] SK Telecom(2025). AI 거버넌스. https://www.sktelecom.com/view/esg/aigovernance.do

[12] KT(2024), Responsible AI Report. https://corp.kt.com/KT_RAI_Report.pdf

※ 출처: TTA 저널 제222호