Issue Report

2025. 02.

FINST&P Vol. 2025-1

인공지능의 잠재적 위험 분석과 안전·신뢰 확보 방안: 거버넌스를 중심으로

국가미래전략기술 정책연구소
글로컬 협력센터 백단비 연구원

KAIST 국가미래전략기술 정책연구소
The Future Institute fod National Strategic Technology and Policy (FINST&P)

Contents

AI의 잠재적 위험과 안전·신뢰 확보를 위한 거버넌스의 필요성

안전하고 신뢰할 수 있는 인공지능(이하 AI)에 대한 정책 전략 수립의 중요성이 나날이 커지며, AI 위험 관리 대응체계와 관련 규범 마련은 여러 국가의 중요 이니셔티브로 자리 잡게 되었다.

AI로 인한 잠재적 위험을 예측하고 그 결과에 따라 선제적으로 정책을 수립하고 시행한다는 것은 사실상 AI의 급속한 기술 발전 속도로 인해 한계가 있다. 또한 불확실성이 팽배한 가운데 이루어진 예측 기반의 규제가 오히려 섣부른 결정으로 이어져 부정적 결과를 초래할 가능성도 크다. 그럼에도 불구하고 오늘날 AI의 잠재적 위험을 지속적으로 예측하고 분석하는 것은 분명 필수적이고 중대한 사안이 되었다. 주요 기관에서도 이를 인지하며 여러 연구와 보고서를 통해 AI의 잠재적 위험 분석과 대응 방안을 제시하고 있다. 최근 발간된 주요 보고서와 논문을 기반으로 AI의 잠재적 위험이 어떠한 양상으로 나타나고 있는지 분석하고, 안전과 신뢰를 확보할 수 있는 방안을 고찰해보고자 한다.

AI의 잠재적 위험과 안전·신뢰 확보를 위한 거버넌스의 필요성

Future of Life[1](2024)에서 세계적 석학인 요수아 벤지오(Yoshua Bengio) 몬트리올대 교수, 데이비드 크루거(David Krueger) 몬트리올대 교수, 스튜어트 러셀(Stuart Russell) UC버클리대 교수, 아토사 카시르자데(Atoosa Kasirzadeh) 카네기멜런대 교수 등의 전문가와 함께 AI의 안전 지수를 평가하는 보고서(「FLI AI Safety Index 2024, '24.12.」)를 발표했다. 주요 6개 도메인(① 위험 평가, ② 직면한 위험 요소, ③ 안전 프레임워크, ④ 실존주의적인 안전 전략, ⑤ 거버넌스 및 책무성, ⑥ 투명성과 커뮤니케이션)을 중심으로 AI 개발 기업(오픈AI, 구글 딥마인드, 메타 등)의 AI 안전 지수를 평가했다. 각 6개 도메인별 세부 지표를 살펴보면 ① 위험 역량 평가, 사전 외부 안전 테스트, 사전 외부 전문가 평가, 모델 취약점에 대한 버그 현상 점검 등, ② 워터마킹, 미세 조정 보호, 데이터 크롤링 등, ③ 모델 평가, 의사 결정, 위험 도메인 등, ④ 정렬 전략, 역량 목표, 안전 연구, 외부 안전 연구 지원 등, ⑤ 이사회, 리더쉽, 파트너쉽, 공공 약속 준수 등, ⑥ 안전 평가 투명성, 스탠포드의 2024년 파운데이션 모델 투명성 지수 등으로 구성되어 있다(Future of Life, 2024). 한편, Esmat Zaidan et al.(2024)는 AI가 국제사회에 미치는 다양한 위험으로 잠재적 일자리 손실과 사회적 불안정 확대, AI의 무기로서의 사용, 개인 정보 피해, 알고리즘 편향성, 가상 위협 및 사이버 갈등, 기술 의존성 강화, 경제적 불평등 확대, 군비 경쟁 유발, 사회 기반 시스템 혼란 야기, AI의 지구 통제 가능성 등이 AI 개발의 단계별로 다양하게 나타날 수 있음을 명시했다.

[그림 1] Future of Life에서 제시한 주요 AI 안전 평가 도메인과 세부 지표

Risk Assessment	Current Harms	Safety Frameworks	Existential Safety Strategy	Governance & Accountability	Transparency & Communication
Dangerous capability evaluations	AIR Bench 2024	Risk domains	Control/Alignment strategy	Company structure	Lobbying on safety regulations
Uplift trials	TrustLLM Benchmark	Risk thresholds	Capability goals	Board of directors	Testimonies to policymakers
Pre-deployment external safety testing	SEAL Leaderboard for adversarial robustness	Model evaluations	Safety research	Leadership	Leadership communications on catastrophic risks
Post-deployment external researcher access	Gray Swan Jailbreaking Arena - Leaderboard	Decision making	Supporting external safety research	Partnerships	Stanford's 2024 Foundation Model Transparency Index 1.1
Bug bounties for model vulnerabilities	Fine-tuning protections	Risk mitigations		Internal review	Safety evaluation transparency
Pre-development risk assessments	Carbon offsets	Conditional pauses		Mission statement	
	Watermarking	Adherence		Whistle-blower Protection & Non-disparagement Agreements	
	Privacy of user inputs	Assurance		Compliance to public commitments	
	Data crawling			Military, warfare & intelligence applications	
				Terms of Service analysis	

* 출처: Future of Life, 2024

1) Future of Life는 미국의 비영리 단체로 혁신적 기술로 인해 발생하는 큰 위험을 방지하고 이로운 방향으로 기술의 혁신을 이끌고자 '15년에 설립된 단체이다. AI, 바이오, 핵무기에 집중하여 위험을 분석하는 연구를 진행중에 있다.

AI의 자율화·지능화에 따른 잠재적 위험 확대와 해당 기술을 활용한 위험 대응 체제 확대 전망

최근 영국의 과학혁신기술부와 AI안전연구소(AI Safety Institute) 주관으로 몬트리올대의 요수아 벤지오 교수가 의장이 되어 30개국의 주요 유수의 전문가 및 UN, EU, OECD 등의 국제기구 전문가 등 100여명이 참여해 만든 「국제 AI 안전 보고서(International AI Safety Report, '25)」를 발표했다. 해당 보고서는 세계 첫 AI 안전 정상회의의 블레츨리 선언을 바탕으로 정상회의에 참석한 국가들과 보고서 작성을 약속하고, 서울에서 한 차례 중간보고서를 발표('24. 05.)한 후 완성된 최종본('25.01.)이다.

범용AI(general-purpose AI)에 국한해 잠재적 위험을 분류하고 위험 관리 방식에 대해 논했는데, 위험을 크게 3가지(① 악의적 사용 위험, ② 오작동 위험, ③ 전체적 위험)로 분류했다. 먼저 악의적 사용으로 인한 위험은 개인 뿐만 아니라 조직과 사회에 해를 끼칠 수 있는 위험을 말하고, 두 번째로 오작동으로 인한 위험에는 AI와 사용자 모두 의도치 않았으나 해를 끼칠 수 있는 경우를 말한다. 마지막으로 전체적 위험은 AI의 광범위한 배포와 활용에 따라 발생할 수 있는 여러 위험들을 포괄해 말하고 있다. 위험별로 나타날 수 있는 세부적인 위험은 표1과 같다(AI action summit, 2025).

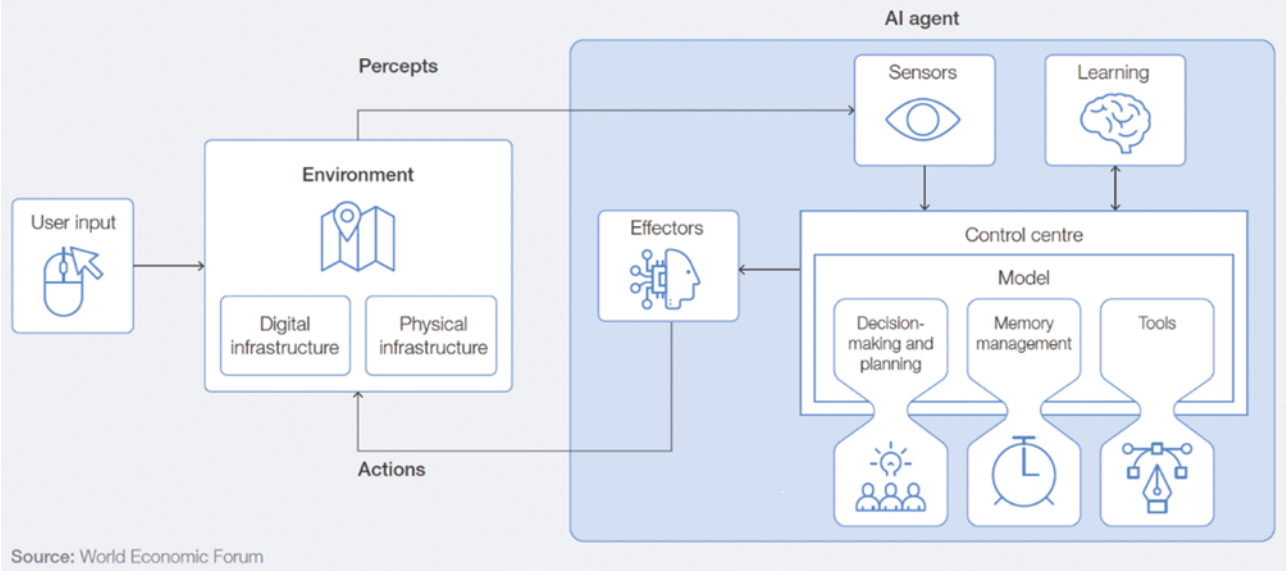
<표 1> 범용AI(general-purpose AI)가 야기할 수 있는 위험 분류

위험 분류	세부 위험	세부 위험별 예시
악의적 사용 위험 (malicious use risks)	◆ 가짜 콘텐츠를 통한 피해 ◆ 여론 조작 ◆ 사이버 범죄 ◆ 생물학적 및 화학적 공격	• 딥페이크 포르노, 음성 사칭을 통한 금융가기, 개인의 평판 파괴, 심리적 학대 등 • 대규모 여론 조작, 정치적 결과에 영향 등 • 사이버 보안 취약성을 스스로 찾아 악용 가능 • 신규 독성 화합물 설계를 용이하게 만들 수 있는 가능성, 생물학적 무기 생산 계획 생성 가능 등
오작동 위험 (risks from malfunctions)	◆ 신뢰성 문제 ◆ 편견 강화 ◆ 통제 상실	• 의료 또는 법률 조언을 받기 위해 AI를 활용했으나 거짓이 포함된 답변이 생성 가능 • 인종, 성별, 문화, 나이, 정치적 의견 등에 대해 편견을 나타낼 가능성 • 범용AI는 더욱 누군가의 통제를 받지 않고 작동하게 되며, 통제력을 상실할 수도 있다는 일부 전문가의 의견 존재
전체적 위험 (systemic risks)	◆ 노동 시장의 위험 ◆ 글로벌 AI R&D 격차 확대 ◆ 시장 집중도 및 집중된 시스템 ◆ 사용의 실패 가능성 ◆ 환경적 위험 ◆ 저작권 침해	• 업무와 작업의 자동화를 불러오며 다수의 일자리 감소로 연계 가능 • 현재 AI R&D 개발이 일부 서방국과 중국에 국한되어 있어 세계적 불평등 야기 가능 • 현재 소수 기업이 범용AI 시장을 지배하므로 소수 기업이 개발한 시스템이 취약성이 높은 상황이 발생될 경우 대규모로 실패와 중단 연계 가능 • 에너지, 물, 원자재 양의 빠른 증가 • 데이터 활용에 따른 동의 및 보상 문제, 다양한 데이터 권한법의 적용 다양화로 AI 안전 연구 수행자의 혼란 가중 등

* 출처: AI action summit, 2025

해당 보고서에서는 특히 범용AI와 관련해 최근 급격히 성장하고 떠오르고 있는 'AI 에이전트(AI Agents)[2]'를 주목하며, AI 에이전트가 기존의 AI 구현 방식 보다 '자율성'이 높은 만큼 잠재적 위험이 보다 크다고 분석하며, 이러한 위험에 대한 접근방식은 이제 막 연구중인 단계이기 때문에 향후 더 많은 분석이 필요함을 명시했다. 또한 AI 에이전트의 활용 확대 시 사용자가 자신의 AI 에이전트의 작동 내용을 알지 못할 가능성과, AI 에이전트가 인간의 통제 없이 작동할 가능성, 공격자가 에이전트를 통해 도청할 가능성, AI와 AI 간의 상호작용으로 복잡한 새로운 위험이 증가할 가능성을 제시했다(AI action summit, 2025). 가트너(2024)에서 발표한 2024년 신기술 하이프 사이클 보고서에서도 주목해야할 25가지 혁신 기술의 주요 트렌드 중 하나로도 '자율형 AI'를 꼽혔으며 인공일반지능(이하 AGI)을 비롯한 AI의 빠른 성장을 예고하고 있다. 자율형 AI는 스스로 작동하고 개선하며 복잡한 환경에서도 효과적인 의사 결정을 내릴 수 있는 시스템이다. WEF(2024)에서도 AI 에이전트의 혁신적인 잠재력을 주목하면서도 AI 에이전트 구현 시 프로그래밍의 허점을 악용하거나 새로운 시나리오에서 학습된 목표를 잘못 적용하는 등의 잠재적 위험을 언급하며, 엄격한 테스트와 투명성 모니터링, 데이터 거버넌스 프레임워크 등이 필요하다고 주장한 바 있다(WEF, 2024). 이러한 자율성과 지능화 측면과 관련한 기업의 AI의 잠재적 위험 대응 방식으로써 안전한 AI 개발을 모토로 하는 앤트로픽(Anthropic)의 '헌법적 AI(Constitutional AI)'을 주목해 볼 수 있다. '헌법적 AI(Constitutional AI)'는 AI의 잠재적 위험을 관리하고 윤리적 가치를 담은 독자적인 위험 점검·관리 방식으로, AI의 위험을 인간이 직접 모니터링하고 데이터를 훈련시키는 기존의 주된 방식이 아니라, 입력된 윤리적 지침을 기반으로 AI 모델 스스로 일련의 원칙과 프로세스를 통해 훈련하여 윤리적 규범을 지키는 방식이다. 한편, 기존의 AI 위험 대응 방식과 관련해 미국표준기술연구소(NIST)(2024)에서는 생성형 AI를 비롯한 AI 자동화 시스템에 발생될 편향성에 대한 위험을 지적하고 있는데, 훈련된 데이터를 기반으로 만들어진 생성형 AI 모델의 유해한 편향성은 보다 차별적 의사결정을 초래하거나 기존의 사회적 편견을 확대하는데 영향을 미칠 수 있고, 같은 모델이 다양한 애플리케이션에 적용되는 균질화(homogenization) 측면에서도 편향성이 문제를 일으킬 수 있으며, 합성 데이터의 과도한 의존 시 모델 붕괴가 이루어질 수 있음을 말하고 있다.

[그림 2] AI 에이전트의 주요 구성 요소와 메커니즘



* 출처: WEF, 2024

2) AI에 대한 수요가 '자동화'를 넘어 '지능화'로 변화됨에 따라 AI에이전트 개념이 등장했다. 데이터 기반의 의사결정을 지원하고, 인간을 보조하는 것을 넘어 스스로 목표를 달성하기 위한 계획을 수립하고, 문제를 정의하고, 환경과 상호작용하여 AI 스스로 결정하고 행동할 수 있는 시스템을 말한다(AI Action summit, 2025). 세계 3대 IT 전시회인 CES 2025(Consumer Electric Show)에서도 'AI 에이전트'가 크게 화두가 된 바 있으며, 구글, 마이크로소프트, 오픈AI 등의 주요 기업에서도 관련 기술에 관한 개발 경쟁에 열을 올리고 있다.

이와 관련해 앞서 살펴본 「국제 AI 안전 보고서(International AI Safety Report, '25)」에서는 AI의 위험을 식별하고 평가하여 관리하기 위한 기술과 방법에 대해 논했는데, 범용 AI 모델을 보다 안전하게 기능하도록 훈련하는 것에 현재까지 어느정도 진전이 있기는 하지만, 최근 연구에 따르면 불완전한 인간의 피드백에 크게 의존하여 모델을 훈련시키는 주된 훈련 방식이 오히려 오류를 발견하기 어렵게 만들어, 모델이 어려운 질문에 대해 잘못 대답함으로써 사람들을 잘못된 판단으로 이끌 수 있다고 한다. 따라서 위험을 식별하고 평가하는 데 있어 효과성을 높이기 위해서는 평가자가 시스템과 모델 자체에 대한 기술 측면에 있어 상당한 전문성을 갖는 것이 필요하다고 제시했다. 또한, AI 위험에 대한 모니터링이 필요한데 기술적 모니터링과 인간의 감독과 개입은 안전을 개선해 나갈 수는 있으나 비용과 지연을 초래할 수 있음을 지적했다(AI action summit, 2025).

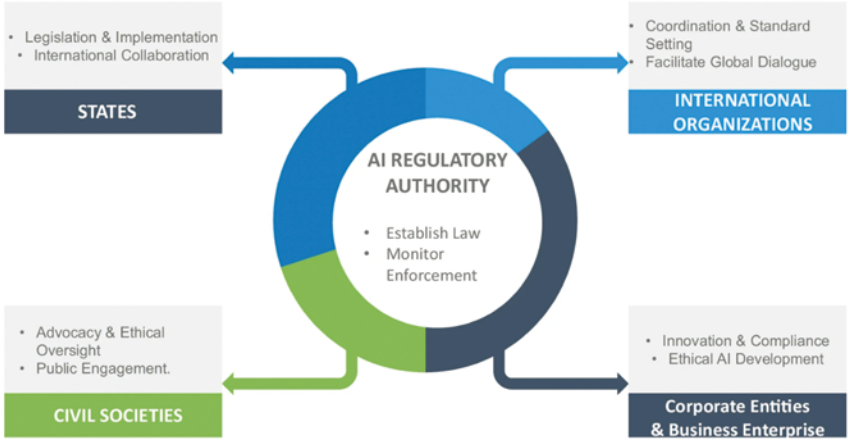
이러한 연구들을 종합해볼 때 기존의 AI의 위험 관리 대응 방식인 인간의 개입(데이터 학습, 모니터링 등)이 오히려 편향성을 높여 잘못된 의사결정으로 이어질 가능성이 있고, 방대하게 발생되는 데이터의 양 대비 높은 전문성을 보유한 전문가 풀의 적은 수와 에너지 소비와 소요 시간 등을 고려했을 때 위험을 대응하고 관리할 수 있는 시스템 체계의 변화가 요구될 것으로 판단된다. 특히 기술 혁신의 흐름 상 AI는 이미 AI에이전트, AGI 등의 기술을 향해 자동화 단계를 넘어선 지능화, 자율화 단계로 개발이 이루어지고 있기 때문에, AI 자체의 자율성과 책임성이 보다 높아지며 잠재적 위험은 더욱 커지고 다양해질 것으로 예상된다. 따라서 이러한 측면을 고려한 잠재적 위험에 대한 예측과 관리도 필요하지만, 다른 한편으로는 오히려 이러한 기술을 역으로 이용해 앤트로픽과 같이 새로운 위험 대응 시스템을 마련하거나 대응 시스템과 체제를 지속적으로 변경하고 구축할 필요가 있다.

신뢰할 수 있는 글로벌 규범 마련을 위한 글로벌 거버넌스의 필요성

급속하게 다변화되는 AI 기술의 발전 양상에 따른 위험 대응 시스템과 규범 체제를 마련하기 위해서는 거버넌스 과정이 필요하다. 「국제 AI 안전 보고서(International AI Safety Report, '25)」에서는 향후 하드웨어 기반의 모니터링 메커니즘이 사용자 고객과 규제 기관의 모니터링을 효과적으로 만들고 국경을 초월한 검증을 할 수 있게 해줄 것으로 예상은 되나, 아직까지 신뢰 가능한 메커니즘은 없는 상태라고 지적했다(AI action summit, 2025). Esmat Zaidan et al.(2024)는 AI 위험이 전지구적 문제임을 감안할 때 AI 위험 대응을 국가와 기업에 재량에 맡기는 것이 아니라, 국제적으로 조정되고 합의된 대응이 필요하며 공동의 목표와 표준 설정을 해야하고, 국제법상에서 규제 권한을 지닌 AI 규제 프레임워크의 필요성을 강조하며, 글로벌 거버넌스 맥락에서의 주요 주체를 국가, 국제기구, 기업, 시민사회로 나누었다.

현재 주요 정부, 국제기구, 기업에서는 자체적으로 AI 위험을 규제하기 위한 규범과 기술적 시스템을 마련하고는 있지만, 아직까지 국제적으로 법적 구속력이 있는 합의된 규범은 마련되지 않았다. 국제적으로 서로 합의되고 조정된 규범이어야 전세계적으로 적용이 가능하고 그 신뢰성을 높일 수 있다. 최근 프랑스 파리에서 개최 예정('25.02.)인 '파리 AI 정상회의(French AI Action Summit 2025)'를 앞두고 이번 정상회의가 기존의 진행되었는 AI 안전정상회의와 마찬가지로 단지 선언적 행사에 그칠 것인지 아니면 글로벌 AI 거버넌스 구축의 기반이 될 수 있을지 귀추가 주목된 바 있다.

[그림 3] 글로벌 거버넌스 맥락에서의 AI 규제



* 출처: Esmat Zaidan et al., 2024

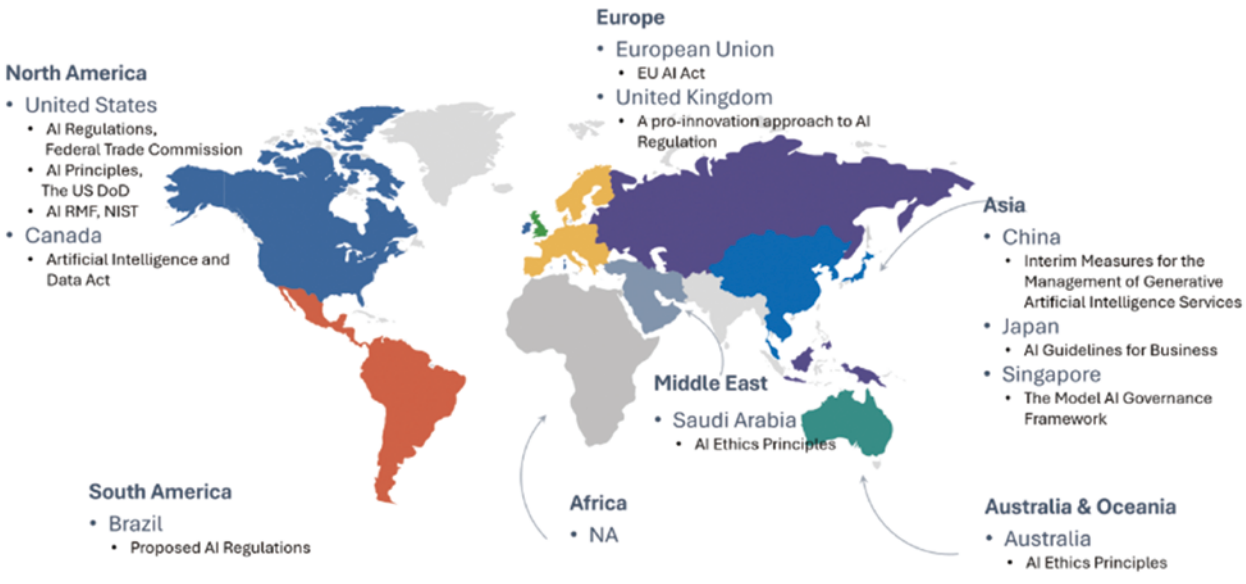
거버넌스 주요 주체의 AI 위험 대응 현황

이번 파리 AI 정상회의에는 공익 AI, AI에 대한 신뢰, 혁신과 문화 등이 주제로 논의될 예정이지만, 주요 목표는 글로벌 차원의 통합된 글로벌 거버넌스 체계 마련을 위한 국제협력 강화이다. 국제적으로 AI 거버넌스의 개념의 정의와 합의가 되어 있지 않은 상태가 전세계적으로 각 주체의 다양한 해석을 만들어 혼란을 야기하고 있고, 주요국의 이해관계가와 정책 방향성이 다소 상이하기 때문에 국제적 조정과 합의가 쉽지 않을 것이라는 평가가 존재한다(최창현, 2025). 또한 앞서 AI의 잠재적 위험을 분석한 여러 참고문헌에서도 잠재적 위험에 대한 대응 방안으로 거버넌스가 공통적으로 포함되어 있다. 잠재적 위험을 진단하고 평가하여 관리하기 위해서는 거버넌스 과정을 함께 살펴보는 것이 매우 중요해졌다. 거버넌스 과정에 참여하는 주요 주체들이 현재 AI 위험을 어떻게 대응하고 있는지를 파악함으로써 향후 합의될 국제적 표준에 대해 예측하고 준비할 수 있고, 이를 통해 안전과 신뢰를 확보할 수 있는 방안을 모색해 볼 수 있기 때문에 주요 주체들의 대응 동향을 추적하는 것은 의의가 있다.

주요국 정부는 AI 위험 대응과 안전 확보를 위해 법 제정부터 국가 AI 안전연구소(AI Safety Institute) 설립까지 국가 차원에서 AI 위험을 직접 감독·관리하기 위한 발판을 마련했고, 세계 최초로 28개국이 참여한 AI 안전정상회의(AI Safety Summit, '24.11.)를 개최하고, AI 안전연구소 국제 네트워크를(International Network of AI Safety Institutes, '24. 11.) 수립하는 등 글로벌 연대를 강화하며 자국 중심의 정책 추진 외에도 국제 거버넌스를 구축하며 글로벌 규범과 표준을 세워 안전성과 신뢰성을 확보해 나가려는 모습을 보였다.

이외에도 주요 주체들의 동향을 살펴보면 OECD, UN 등의 국제기구와 G7과 같은 주요국 모임에서는 AI의 안전한 사용에 관한 결의나 권고안을 채택하는 등 국제사회의 규제와 거버넌스의 필요성을 일깨웠고, 기업에서는 혁신적인 AI 기술 개발을 전폭적으로 추진하면서도, 동시에 AI 안전 및 윤리에 관한 연구·전략 전문 팀을 신설하고 규범 마련을 위한 정책 수립의 장에도 적극적으로 참여하는 모습을 보이며, AI 안전성에 대한 기업 자체의 대응 체제 마련을 기업의 대표 이미지 전반에 내세우는 모습도 나타나고 있다.

[그림 4] 국가별 AI 규제 정책 추진 현황



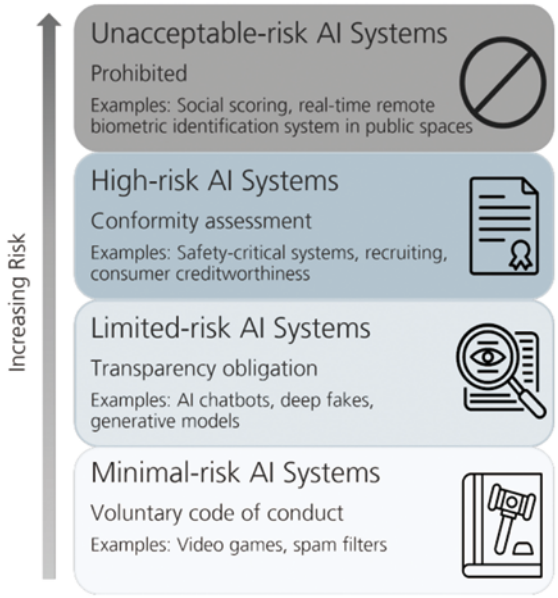
* 출처: Irina Steenbeek, 2024

최근 주요국 정부의 AI 위험 대응 현황

최근 주요국 정부의 AI 위험 대응 현황

가. EU: 세계 최초의 법적 AI 규제 마련과 협력적 거버넌스 체제 강화 강력한 법적 규제 추진

[그림 5] EU AI 법의 위험 분류 현황



* 출처: Fraunhofer IKS, 2024

정부 차원의 AI 위험 대응 현황을 구체적으로 살펴보면, 먼저 EU는 전세계 최초로 「AI 규제법(AI Act, 이하 AI법)」을 제정 및 발효('24. 8.)했다. 이는 AI의 개발부터 배포, 사용에 이르기까지 AI 시스템에 관한 윤리적 규제 프레임워크를 수립하기 위함이다. AI법의 주요 특징으로는 AI로 인해 발생할 수 있는 위험을 위험 단계별로 차등하여 4단계(허용할 수 없는 위험, 고위험, 제한된 위험, 최소 위험)로 분류하고, 특히 고위험 AI 시스템에 중점을 두어 주요 분야를 규제하고 있다. 의료, 교육, 선거, 핵심 인프라 등에 활용되는 AI 기술을 고위험 등급으로 분류하여, 시장 출시 이전에 인간이 직접 AI 사용을 반드시 감독하고 위험관리시스템을 갖추도록 규제하였다. 또한 관련 이해관계자의 유형을 분리 및 정의하고 있으며, 제공자(provider), 배포자(deployer)의 경우 특히 고위험과의 상관성과 투명성을 확보해야 하는 의무에 초점을 맞췄다. 또한 AGI에 투명성 의무를 부과하여 챗GPT 포함 모든 AGI를 대상으로 AI의 학습과정에 사용된 콘텐츠의 출처를 명시하도록 규정했다. 끝으로 AI법의 원활한 시행과 각 회원국의 책임 기관 선정 등을 총괄하는 AI 사무국(European AI office)의 운영 지침도 포함하여 유럽 단일 AI 거버넌스 시스템을 구축하고자 추진하고 있다(EU, 2024).

EU의 AI법 제정은 윤리적 차원의 책임있는 AI 사용을 권하며 기업과 사용자의 자율 규제에 맡겼던 기존의 경향에서 이제는 법적 규제 추세로 옮겨가고 있음을 상징하기도 한다(Celso Cancela Outeda, 2024). EU의 개인정보보호법인 GDPR(General Data Protection Regulation)이 전세계적으로 영향력을 미치고 있는 점과 EU의 AI법이 세계 최초의 AI 규제법임을 미루어 볼 때, EU의 AI법은 글로벌 규제 표준의 기반이 될 수 있는 가능성을 내포하고 있으며, 앞으로 AI 규제 분야에 있어 EU의 영향력 강화가 예상된다(Heather Domin, 2024). 반면, EU와 같은 선제적 규제가 오히려 AI 기술의 발전과 혁신을 저해할 수 있다는 전문가들의 견해도 일부 존재한다.

협력적 거버넌스 체제 강화

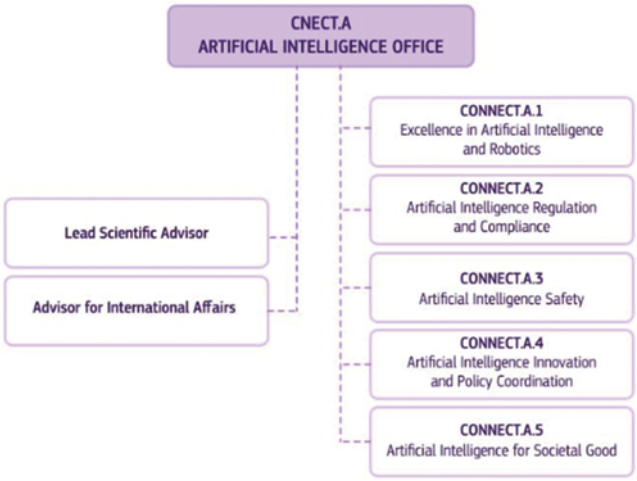
EU는 AI법의 원활한 시행을 뒷받침하고 신뢰할 수 있는 AI의 개발과 안전한 활용을 위해 거버넌스 체제를 활성화하고자 AI 사무국을 유럽연합집행위원회(EC) 내에 설립했다. AI 사무국은 다른 나라(미국, 영국, 일본, 우리나라 등)의 AI 안전연구소(AISI)와 유사 역할을 담당하나 규제 측면에 주로 집중하며 연구에는 초점을 두지 않는다는 차별점이 있다(Lily Ottinger, 2024). EU는 AI 사무국을 통해 EU 차원의 단일 AI 거버넌스 시스템의 기반을 형성하고자 한다. AI 사무국은 과학계, 산업계, 싱크탱크, 시민단체 등의 전문가 커뮤니티와 협력과 소통을 통해 정확한 의사 결정을 하고자 계획되었으며, 아래의 그림과 같이 AI 안전과 규제부터 기술 혁신까지 주요 임무를 맡도록 부서가 구성되어 있다(EC, 2024).

AI 사무국의 주요 임무는 1) AI법 지원, 2) 신뢰할 수 있는 AI의 기술 개발 및 활용 강화, 3) 국제 협력 추진, 4) 다양한 이해관계자간의 협력이다. 거버넌스 과정은 협력적 거버넌스(Collaborative governance)의 형태로 추진되고 있는데, 이러한 협력 형태가 오히려 기존 전문가 집단에 있는 참여자들의 소수 특혜로 이어질 수도 있고, 규제 대상이 되는 기업이나 기관의 회피 가능한 루트로 역이용될 가능성이 있다고 지적하는 전문가의 의견도 나오고 있다(Celso Cancela Outeda, 2024).

이와 관련해 최근 국내의 현황을 견주어 보자면, 한국 정부는 AI 기술의 혁신과 규제의 윤리적 기준을 세우며 AI 정책 전반을 심의·조정하는 최상위 거버넌스 기구로 대통령 직속 국가인공지능위원회를 출범시켰다('24. 9.). 위원회의 명단을 살펴보면, 관련 학계의 주요 전문가부터 SK 하이닉스, LG AI연구원, 카카오 등 AI 활용 기업의 민간위원이 다수 포함된 반면, 국내 AI 선도기업이자 대표적인 기술 플랫폼 기업으로 불리는 '네이버'는 포함되어 있지 않다.

네이버는 전세계 세 번째로 초대규모 언어모델 '하이퍼클로바'를 출시('21. 5.)하며 국내 최초로 검색 서비스에 적용했고, 연이어 생성형 AI 모델인 '하이퍼클로바X'를 출시('23. 8.).하며, 이를 여러 서비스에 적용하는 '온 서비스 AI 전략'과 '소버린 AI' 구축을 계획하며 자사의 AI 연구 개발에 기업 매출의 20% 이상에 해당되는 금액의 투자 계획도 밝힌 바 있다. 또한 세계 첫 AI 안전 정상회의에서도 우리나라 정부 대표와 네이버를 포함해 단 3곳만 초대받을 만큼 한국의 AI 기업으로서 대표성을 띠고 있는 가운데, 네이버를 위원회 명단에 속하지 않은 것을 두고 의문과 부정적인 뜻을 표하는 의견들도 존재했다(양진원, 2024, 변상근, 2024). 앞선 전문가의 의견처럼 기존의 전문가 협력 체제가 오히려 특혜를 조장할 수 있는 가능성도 배제하지 못하겠지만, 반대로 정책 수립 시 현장의 목소리를 귀담아 듣는데 이해관계자 거버넌스의 구성에 치우침은 없는지, 거버넌스 과정이 합리적인지도 거버넌스 시작 단계부터 고려할 필요성이 있다.

[그림 6] EU AI 사무국의 부서 구성 및 역할



* 출처: European Commission, 2024

[그림 7] 2020~2024년 네이버·카카오 연구개발(R&D) 투자 현황

	2020	2021	2022	2023	2024(3분기 누적)
네이버	1조3321억원	1조6551억원	1조8091억원	1조9926억원	1조3620억원
카카오	5354억원	7645억원	1조213억원	1조2236억원	9720억원

* 출처: 변상근, 2024

나.
미국: 미국의 패권 장악력과 자국 우선주의 기반의 정책 확대

AI 규제 강화 및 규범 선도 정책에서 정책 변동 예정

미국은 AI 기술 선도국으로서 AI 안전 규범까지 선도하려는 정책과 전략을 추진해왔다. 다만, 도널드 트럼프 대통령의 재집권으로 기존의 바이든 행정부에서 트럼프 2기 행정부로 바뀌며 미국 우선주의 강화 전략(Make America Great Again, MAGA)을 기반으로 AI 규제와 윤리 규범 측면에 대한 정부의 정책 방향이 크게 변화될 것으로 예고되고 있다.

기존의 바이든 행정부는 '23년 7월과 9월에는 주요 AI 선도기업의 '자발적 AI 안전 서약(Voluntary AI Commitments)'를 받고, 연방 차원 최초의 법적 구속력을 가진 「AI 규제를 위한 행정명령(Safe, Secure and Trustworthy Development and Use of Artificial Intelligence(Executive Order 14110), '23.11.)」을 통해 안전성 평가 의무와 개인정보 보호, 형평성·시민권 제고, 국제적 AI 리더십 확보를 추진했고, 「국가안보각서(National Security Memorandum), '23.11.)」를 통해 국가 안보에 사용되는 AI에 대한 윤리 기준을 설정, 국립표준기술연구소(NIST) 산하의 AI 안전연구소(AISI) 설립('23.11.) 등을 통해 AI 규제 이니셔티브를 수립하고 확장했다(유계환 외, 2024, 법무부 국제법무지원과, 2024).

특히 미국은 한국, 일본, EU, 캐나다, 프랑스, 싱가포르, 호주, 케냐 10개국을 모아 'AI 안전연구소 국제 네트워크(International network of AISI)' 출범을 선언('24.11.)하며, 크게 1) 연구, 2) 테스팅, 3)지침, 4) 포용 분야에 대한 글로벌 협력을 추진했다. 이러한 바이든 정부의 규제 정책과 거버넌스 구축은 전세계적으로 AI 위험 관리와 안전 강화를 주도하며 AI 글로벌 규범의 리더로 자리잡는 모습을 보였다.

[그림 8] 자발적 AI 안전 서약에 참여한 15개 AI 기업

* 출처: 정해영, 2024

[그림 9] 미국의 AI 안전연구소 국제 네트워크 관련 국가별 연구소 설립 현황

	United States	United Kingdom	European Union	Japan	Singapore	South Korea	Canada	France	Kenya	Australia
Established	February 2024	November 2023	May 2024	February 2024	May 2024	May 2024 (Announced)	April 2024 (Announced)			
Name of Organization	**US AISI**	**UK AISI**	**EU AI Office**	**Japan AISI**	**Singapore AISI**	**Korea AISI**	**Canada AISI**			
Housed Under	National Institute of Standards & Technology	Department for Science, Innovation & Technology	Directorate General for Commu-nications Networks, Content and Technology.	Informa-tion-Tech-nology Promotion Agency	Digital Trust Centre	Electronics and Telecom-munications Research Institute				
Funding (USD & Foreign Currency)	**$10 million** (FY24)	**> $65 million/yr** (>£50 million/yr 2024-2030)	**$51 million** (€46.5 million) (Funding period unknown)		**$7.5 million/yr** (S$10 million/year) (2023-2027)	**$7.2-14.4 million/yr** (W10-20 billion/yr) (Tentative, starting 2025)	**$36.5 million** (C$50 million) (Funding period unknown)			
Staff	c.20 (current core staff)	c.20 (current core staff)	c.50 (planned, AI safety unit)	c. 23 (current staff)		Minimum 30 staff (planned, budget pending)				
Public List of Functions	US AISI Vision, Mission, and Strategic Goals	Introducing the AI Safety Institute	Tasks of the AI Office	AISI's Tasks	Initial Research Areas					
Published Research or Guidelines	Managing Misuse Risk for Dual-Use Foundation Models	See website		See website	Model AI Governance Framework for Generative AI					

Legend	No public statement	Public Information

* 출처: Gregory C. Allen, 2024

미국은 연방 차원 외에도 주별로 AI를 규제하고 있는데, 특히 주요 글로벌 테크사들이 위치한 캘리포니아주의 「AI 규제법(SB1047)」이 주 의회를 통과('24.8.)하며 글로벌 이슈로 떠오른 바 있다. 주요 내용은 AI 모델 출시 시 그 이전에 반드시 안전성 테스트를 의무화하고, 본사 소재지에 상관없이 캘리포니아주에서 사업을 하는 모든 회사에 법이 적용되도록 구상되어 있다. 해당 법안에 대해 당시 주요 AI 개발사들의 반응은 지지하는 측면 외에도 오히려 규제로 인한 산업의 혁신 저해를 염려하는 미온적 측면과 강경한 반대측면으로 나뉘었다. 또한, 유타주도 「AI 정책법(AI Policy Act), '24.5.」을 발효하며 캘리포니아주보다 먼저 생성형 AI를 사용하는 회사에 투명성 의무를 부과하는 미국 주 법이 되었다. 유타주의 법은 EU의 AI법을 주로 참고해 제정한 것으로 알려져 있으며 이 외에도 콜로라도주, 코네티컷주, 일리노이주, 버지니아 등 주 자체적으로 AI 규제법을 시행하는 주가 늘어나는 추세에 있었다(Daniel K. Alvarez et al., 2024).

반면, 트럼프 2기 행정부가 들어서면서 규제 완화를 통한 산업 경제 활성화를 내세우고 있다. 일반적으로 기존에는 UN 등 주요 국제기구에서 전지구적 이슈에 대한 큰 규모의 거버넌스 과정과 국제사회의 합의가 이루어져 왔는데, 트럼프 대통령은 그간 이러한 다자협의체나 기구에 대해 미온적인 반응을 보여왔고, 실제로 취임 직후 세계보건기구(WHO)를 탈퇴하는 모습을 보였기 때문에, 향후 트럼프 정부가 AI 글로벌 거버넌스 협의의 장을 기존대로 이끌지, 혹은 중국이나 다른 나라에서 국제 AI 표준과 규범을 설정하며 미국은 기술 선도국 지위만 유지하고 기술 패권만을 확대할지에 대한 예측 또한 분분하다(Haileleol Tibebu, 2024).

기업과 시장 중심의 기술 혁신에 중점을 둔 거버넌스 확대

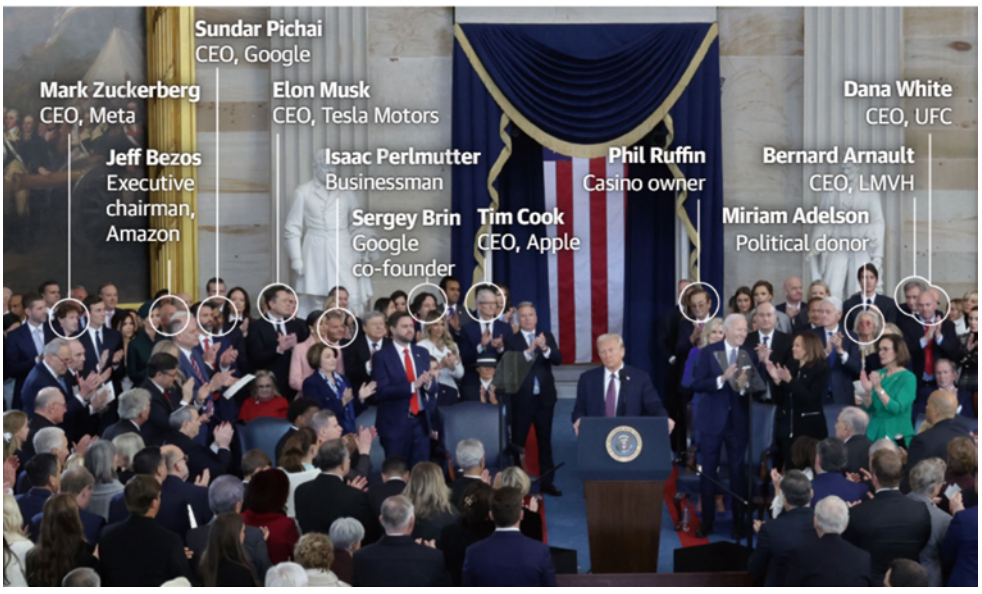
미국의 AI 이니셔티브법을 기반으로 설립된 美 국가인공지능자문위원회(National Artificial Intelligence Advisory Committee, 이하 NAIAC)에서는 최근 트럼프 2기 행정부를 위한 AI 관련 정책 조언을 담은 보고서(「NAIAC Insights for the Administration of President Donald J. Trump, '25.1.)를 발간했다. 해당 보고서에서 미국이 지속적으로 AI 리더십을 확보하고 경쟁력을 높이기 위해 우선적인 10가지 AI 관련 정책 분야를 제시하며, 그 중 하나로 'AI 거버넌스'를 뽑았고, 우선되는 10가지의 정책 분야가 교차되고 상호 관련 되는 5가지 항목에서도 '글로벌 AI 거버넌스 선도'를 선정했다. NAIAC는 AI 혁신 분야에서의 미국의 리더십 강화를 위해서는 'AI 안전과 보안'을 중요하게 여길 것을 강조하고 있으며, 이를 위해 AI 위험 점검을 할 수 있는 보안 점검 지침 기준 마련과 평가 관행을 만들고, NIST와 AI 안전연구소가 AI 안전에 관한 평가 체계를 견고히 할 수 있도록 자금 지원을 늘릴 것과, 최근 몇 년간 글로벌 AI 거버넌스 확대 양상(글로벌 정상회담, 양자 및 다자간 협정, 국가 AI 안전연구소 설립 등)에 따른 다양한 AI 거버넌스 접근 방식에 대해 글로벌 차원의 조율과 AI 거버넌스를 정립함으로써 국제적 리더십을 공고히 할 것을 조언하고 있다(NAIAC, 2025).

또한 교차 및 상호 관련되는 5가지 우선순위 중 또 다른 하나로는 '공공-민간 파트너십 활용'을 뽑았다. 파트너십은 프로그램 설계와 AI 어플리케이션에 대한 대중의 신뢰를 강화할 수 있고, 이러한 파트너십에 광범위한 이해관계자를 의사 결정 및 정책 논의에 참여시켜야 할 것을 강조하고 있다. AI 기술력이 점차 중요해지는 만큼 테크사들의 정책 참여는 커지고 정부와의 협력은 더욱 밀접해지고 있다.

이번 트럼프 대통령의 취임식에서도 메타의 CEO 마크 저커버그(Mark Zuckerberg), 애플의 CEO 팀쿡(Tim Cook), 구글의 CEO 순다르 피차이(Sundar Pichai), 아마존의 창업자 제프 베조스(Jeff Bezos), 테슬라의 CEO 일론 머스크(Elon Musk) 등 글로벌 빅테크의 주요 리더들의 자리를 트럼프 대통령의 가족 구성원 옆으로 배치하는 모습을 통해 트럼프의 향후 기업 중심의 정책 기조를 확인할 수 있었고, 트럼프 대통령의 취임 직후 다음날 바로 주요 기업들과의 초대형 AI 인프라 프로젝트인 「스타게이트(Stargate AI Initiative), '25.1.」 정책[3]을 발표하고, 취임 후 곧바로 전 바이든 정부의 AI 산업 규제 행정명령은 폐기한 것을 확인할 수 있다.

이외에도 여러 AI 기술 사용 및 개발 관련 기업들의 정책 참여와 협력 행태는 더욱 커질 것으로 예상되므로, 이러한 강화 추세는 한국에도 영향을 미칠 것이다.

[그림 10] 글로벌 빅테크 수장들의 트럼프 대통령 취임식 참석 모습



* 출처: Edward Helmore, 2025

3) 스타게이트는 오픈 AI, 소프트뱅크, 오라클이 설립하는 합작 법인으로 향후 4년간 미국 AI 인프라에 5000억 달러를 투자하여 대규모 데이터센터 건설을 포함한 인프라 개발, 차세대 AI 기술 개발, 전략 확보 등을 책임지고, 마이크로소프트, 엔비디아 등도 기술 파트너로 참여할 것으로 예상되고 있다(Kim Eun-jin, 2025)

다.
중국: 다자협력 기반의 자국 중심 AI 정책 강화와 글로벌 확장 도모

중국의 세계적 AI 기술 확산의 기반 마련을 뒷받침하는 AI 규범 마련 정책

중국은 미국과 AI 기술력을 앞다투는 나라로 성장하며 빠르게 기술 혁신을 이룩하고 있는 것 뿐만 아니라 AI 안전·윤리 규범 마련에도 여러 법과 제도를 마련해오고 있다. 중국은 「차세대 인공지능 개발계획(A next generation artificial intelligence development plan, '17)」을 주요 국가계획으로 세우며, 이를 기반으로 '19년 과학기술부 산하에 산학연 AI 관련 전문가로 구성된 '국가 차세대 인공지능 거버넌스 전문위원회'를 마련해 해당 위원회를 통해 「차세대 AI 윤리규범('21)」을 수립했다. 그 외에도 AI 발전에 따른 규제를 위해 일명 중국의 데이터3법 (「네트워크 안전법('17)」, 「데이터 보안법('21)」, 「개인정보보호법('21)」)을 시행, 「인터넷 정보 서비스 알고리즘 추천 관리 규정('21)」, 「인터넷 정보 서비스 딥페이크 관리규정('23)」, 「생성형 인공지능서비스 관리 잠정방법('23)」, 「인터넷 안전기준 실천지침: 생성형 인공지능 윤리규범('23)」등의 규정을 잇따라 공표 및 시행했다(홍천택, 2021, 이상우, 2023). 또한 베이징시는 '23년 5월 과학기술부의 AI 윤리 규범이 반영된 「세계적인 영향력을 갖춘 AI 혁신 발원지 건설 실시 방안(2023~2025년)」, 「범용 AI의 혁신·발전을 촉진하기 위한 몇 가지 조치」를 발표했다(정민욱, 2023).

중국이 자체적으로 AI 안전 확보 및 점검 강화에 관한 법 수립과 정책을 적극적으로 추진하고 있는 주요 이유로는 중국이 해외 시장 진출 시, AI 기술에 대한 안전성 검증 문제로 인한 비판을 사전에 방지하기 위함이라는 시각과 중국 또한 기술 우위 확보 뿐만 아니라 자국 중심의 AI 안전 규범 마련이라는 신(新) 패권을 확대하기 위함이라는 시각이 존재한다(백서인, 2024).

[그림 11] AI 기술 발전 및 현안별 중국의 관련법 제정 현황

중국 AI 현안별 개별법 개요

주요 AI 현안	중국의 개별법 제정 현황
윤리규범 ('21)	◆「차세대 인공지능 윤리규범」 * 新一代人工智能伦理规范
인터넷 서비스 알고리즘 ('21)	◆「인터넷 정보 서비스 알고리즘 추천 관리규정」 * 互联网信息服务算法推荐管理规定
딥페이크 ('22)	◆「인터넷 정보 서비스 심층 종합 관리규정」 * 互联网信息服务深度合成管理规定
생성형 AI ('23)	◆「생성형 인공지능 서비스 관리 잠행방법」 * 生成式人工智能服务管理暂行办法 ◆「인터넷 안전기준 실천지침-생성형 인공지능 서비스내용 표식방법」 * 网络安全标准实践指南-生成式人工智能服务内容标识方法

* 출처: 법무부 국제법무지원과, 2024

지역 거점 및 다자 협력 체계의 글로벌 거버넌스 강화

중국의 AI 관련 거버넌스 형태는 미국이 서방 가치동맹 혹은 동맹국간의 양자간 협력을 통해 AI 기술력 강화와 규범 선도를 꾀하는 것과 달리, 다자협력체계를 취해 지역 플랫폼, 자국에 유리한 다자 플랫폼(ASEAN, BRICS) 등 다양한 국가와의 협력을 통해 글로벌 거버넌스를 구축 및 강화하는 행태를 나타내고 있다(김소영 외, 2024).

중국은 '23년 10월 「글로벌 AI 거버넌스 이니셔티브(AI Governance Initiative)」를 제시하며 AI 윤리·안전·표준·규범과 관련한 포괄적이고 균형잡힌 접근 방식과 다자주의를 유지할 것을 표명하고 있다. 또한 자체적인 이니셔티브 수립 외에도 국제기구의 논의에 참여하며 AI 기술 혁신과 안전한 AI 활용을 꾀하고 있다. UN의 고위급 회의에 참석하여 27개국과 함께 「AI 역량구축을 위한 국제협력 강화에 관한 의결안('24.6.)」을 체결하며 AI 기술의 오용과 남용 방지, 안전하고 제어 가능한 AI로 신뢰할 수 있는 기술 개발, 친환경 방식의 AI 개발로 녹색 전환 추진 계획을 밝혔다. 해당 의결안을 기반으로 중국 정부는 「모두를 위한 인공지능 역량 구축 행동계획(Artificial Intelligence Capacity-Building Action Plan For Good and For All, '24.9.)」을 발표했다. 계획에는 중국이 취해야 할 10가지 구체적인 조치를 취하고 있으며, 오픈소스 커뮤니티 형태의 국제협력을 통한 협업 촉진과 국제 협력 플랫폼 구축, 현지 중심의 AI 학습데이터 마련, AI 윤리와 안전에 관한 규제 사례 공유 등을 담고 있다(Ministry of Foreign Affairs The People's Republic of China, 2024, U.N. Digital library, 2024, Zeng Yi, 2024).

중국은 주요국들이 빠르게 AI 안전연구소(AISI)를 설립한 것과 달리 아직까지 AI 안전연구소를 수립하지 않는 행보를 보이는 가운데, 중국이 AISI를 설립하지 않는 주요 이유로는 미국과 서방을 시작으로 설립되고 있는 이러한 흐름에 따라가는 것 자체가 주도권을 뺏기기 싫은 위한 중국의 대응 모습일 수 있으며, 또한 이미 중국 자체적으로 AISI의 역할을 담당할 수 있을 만한 주요 기관들의 있으나, 자국 내 지위 다툼으로 인해 아직까지 관련 국제협력 및 국제사회 대응 시 AISI의 역할을 어느 기관이 맡을 것인지가 결정되지 않았다는 의견이 존재했다. 다만, 국가 차원에서 AISI 역할을 수행할 기관을 지정하지 않거나 AISI를 설립하지 않을 경우 AISI 국제 네트워크를 비롯한 글로벌 논의의 장에서 소외될 가능성이 있고, 다른 나라에서 협력을 원할 시 특정된 기관이 없어 국제협력의 난항을 불러올 수 있다는 분석이 나왔다(Oliver Guest, 2024).

중국은 베이징 AI 표준화 연구소 개설 계획을 발표('24.8.)했는데, 베이징 AI 표준화 연구소는 AI의 표준화를 전문으로 하는 중국 최초의 연구기관이 될 것으로 알려져 있으며, AI 응용 프로그램 연구와 AI 기술로 인해 사회에 발생 가능한 거버넌스를 연구하고, 개방적이고 협력적인 시스템으로 AI 적용 과정에서 나타나는 문제에 대한 표준화된 해결안을 제공 예정이다. 또한 국제전기통신연합(ITU), 국제전기표준회의(IEC), 국제표준화기구(ISO) 등 주요 표준화 기구에서도 AI 글로벌 표준 제정에 적극적으로 나서고 있어 글로벌 표준 선정을 위한 전략 추진도 엿보인다(연합뉴스, 2024, 백서인, 2024).

[그림 12] AI안전연구소 역할 수행이 가능한 중국의 유망 연구기관 현황

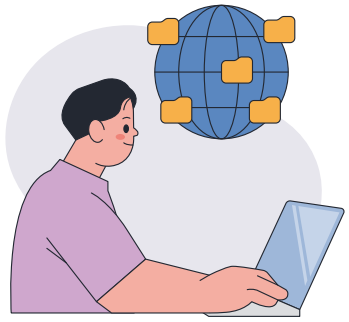
Table 1: Most promising Chinese AI Safety Institute counterparts

		AI Safety Institute core functions		
Institution	Recommended topics	Technical research and evaluations	Standards	International cooperation
CAICT, a think tank housed within the Ministry of Industry and Information Technology.	Evaluations	✓	✓	✓
Shanghai AI Lab, a government-backed AI research institution.	Technical research and evaluations and international cooperation	✓	✓	✓
TC260, a committee within China's official national standards body.	Standards		✓	
Institute for AI International Governance, a policy-focused research institute within Tsinghua University.	International cooperation			✓
Beijing Academy of Artificial Intelligence, a government-backed AI research institution.	International cooperation	✓	✓	✓

* 출처: Oliver Guest, 2024

라.
일본: 국제사회의 AI 규범 강화에 따른 정책 변동과 기존의 정책 기조와의 공진화 모색

기술 혁신과 규제 사이의 AI법 제정 방향 논의 중



일본은 미국과 EU의 AI 법적 규제 강화 흐름에 따라 지난 2월('24년) 「책임있는 AI 추진을 위한 기본법(Basic Law for the Promotion of Responsible AI, AI Act)」의 초안을 공개하며, 한국의 AI 기본법 시행 예정 시기와 같은 '26년에 시행을 예고하고 있다.

일본의 본래 정책 기조는 AI 기술 혁신과 투자 촉진을 위한 기업 친화적인 정책 시행과 기업 중심의 자율 규제로 주로 「AI 사용을 위한 가이드라인(AI R&D 지침, AI 활용 지침, AI 원칙 이행 지침)」 제공을 통해 규제를 해왔다. 또한 '24년 4월 일본의 총무성(MIC)과 경제산업성(METI)은 기존의 세 가지 가이드라인을 통합한 「AI 사업자 가이드라인(AI Guidelines for Business Ver 1.0, AI Guidelines)」을 발표했는데, AI 개발 및 이용 관련 사업자가 자율적으로 규제 사항을 준수할 것을 권고하며, AI 개발·제공·이용 기업의 발생 가능한 리스크를 줄이면서도 관련 기술 혁신과 산업을 발전시킬 것을 요구했다(Tomomi Fujikouge et al., 2024, 장보은, 2024, 고환경 외, 2024). 'AI 개발자, AI 서비스 제공자, AI 비즈니스 사용자' 각 유형별로 가이드라인을 제공하여, AI 개발 단계의 데이터 편향성 방지부터 개인정보 침해 방지 추구, 저작권 문제 등을 관리하면서도 이익은 극대화하도록 되어 있으나 이러한 지침이 매우 추상적이라는 평도 있었다(Tomomi Fujikouge et al., 2024). 법적 구속력이 없는 한계와 이러한 자율적 규제가 오히려 기업의 자체적 규제와 사업의 방향 설정에 어려움을 겪을 수도 있다는 의견도 나왔다(장보은, 2024).

일본 정부에서도 자국 내의 이러한 비판적 의견과 국제적인 AI 규제 강화 흐름에 발맞춰 '24년 일본 최초의 AI법 시행 계획을 발표하게 된 것으로 보인다. 일본은 법 수립을 위해 전문가들을 구성하여 'AI제도 연구회'를 출범('24.08.)하고, 기술혁신 촉진과 위험 대응 양립을 내걸며 법안 마련 중에 있다. 주요 글로벌 로펌을 비롯한 일각에서는 일본의 AI법이 기존의 정책 기조에 따라 기업 중심의 AI 기술 혁신 특성에 맞춰 민간사업자를 대상으로하여 AI 개발의 안정성을 확보할 수 있는 시스템을 구축하여 안전성 검증을 수행하고, 규정 준수 상태를 정부 또는 관련 관리 조직에 정기적으로 보고하는 조항을 포함할 것으로 전망하고 있다(Tomomi Fujikouge et al., 2024).

이러한 전망과 같이 일본의 AI법은 EU와 다른 성격을 띠는데 AI에 대한 엄격한 규제 보다는 규제 시행을 최소화하고 혁신보다 규제가 우선시되면 안된다는 주장이 주를 이루고 있다. 아무래도 AI 기술력이 선도국들과 비교 시 상대적으로 낮기 때문에 기술력 강화가 더 앞장서야 한다는 주장이 영향을 미친 것으로 보이며, 실제로 이러한 기업 중심의 기술 혁신 강화 정책은 글로벌 빅테크 기업들(오픈 AI, 마이크로소프트, 구글, 아마존 등)의 일본 진출 및 투자 계획 발표로 이어지고 있는 것으로 보여진다(정보통신산업진흥원, 2024).

하지만 기술혁신과 규제의 균형에 있어 그 초점을 잡는 것이 쉽지 않아 보인다. 최초에는 AI 기업의 기술 악용 시 해당 사업자의 명칭을 공개하는 벌칙을 포함하기로 했으나, 최근 기술혁신 저해와 기업의 국제경쟁력 약화를 우려하는 것을 이유로 법적 벌칙을 넣는 것을 다시 유보하고 있는 것을 통해 미루어 짐작할 수 있다(강구열, 2024). 우리나라 역시 AI 기본법 수립 초기 국회 표류로 난항을 겪었던 이유 중 하나가 기술혁신과 규제 사이에서의 간극을 줄이고 양립하기 위함이었고, 기술 혁신에 조금 더 의견이 기울어지기는 하였으나, 그 접점을 찾기 어려워 현재까지도 이에 대한 논쟁이 끊이지 않고 있기 때문에, 유사 상황에 처한 일본이 이러한 이슈를 어떻게 해결해나가는지 주목할 필요가 있다.

아시아 중심의 거버넌스 선도 및 강화 조짐

일본은 주요 정상회의와 고위급 회의를 통해 글로벌 AI 규범을 선도하기 위한 발판을 마련하고 있다. 일본은 AI 안전연구소를 영국과 미국에 이어 전세계에서 세 번째로 설립한('24.02.) 것 외에도, G7 정상회의에서 「G7 히로시마 AI 프로세스 행동 강령(Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems, '23. 10.)」을 제안하며 전세계적으로 AI 규범을 수립하는 첫 토대를 마련했다. 히로시마 프로세스는 ASEAN 6개국을 포함한 50여개국이 해당 강령에 지지를 표명한 것으로 알려져 있다.

최근('25년)에는 지리적 위치를 극대화한 거버넌스 구축을 꾀하고 있다. 중국의 영향력 확대를 견제하고 서구 중심의 AI 기술을 넘어 AI 기술 개발과 규제 측면 모두에서 글로벌 주도권을 확보하고자, '아시아 AI 연대' 구축에 나서고 있다. 일본은 ASEAN 회원국들과 AI 분야 협력을 강화하여 동남아시아 언어와 문화에 특화된 LLM을 개발하고, AI 안전성 평가 표준을 공동으로 마련하고자 하고 있다. 아시아 지역의 특성을 반영한 AI 생태계 구축을 목표로 일본-ASEAN 디지털 장관 회의('25. 1.)에서 협력 방안을 제시하며, 올해 안에 파트너십 체결을 계획하고 있다(신민철, 2024). 또한 싱가포르와도 양자간 오랜 연구 협력 파트너쉽[4]을 맺으며 아시아 중심의 AI 기술 혁신과 규범 마련을 함께 꾀하고 있다(A*STAR, 2024).

즉, 일본은 자국의 정책 추진 목표인 AI 기술과 규범 선도를 위해 ASEAN 국가의 협력 강화를 통해 소버린 AI를 형성하며, 아시아를 주도하려는 움직임을 나타내고 있다. 이와 관련해 최근 관련 보고서의 전문가 의견에 따르면, 현재 급변하고 있는 글로벌 AI 거버넌스 지형의 타이밍에 맞추지 못하고 있는 한국의 실태를 지적하며, 미국에서는 북미 AI 동맹이 급부상되며 서방 가치 동맹을 맺고 있고, 중국은 자국의 강점을 기반으로 다자 협력 거버넌스 체제를 마련하고 있는 점을 언급했다(김소영 외, 2024).

위와 같이 지역 중심의 선도 체제를 가져가려는 일본의 모습을 통해 한국 또한 기존의 협력 체제에 국한되거나 일부 국가에 한정된 협력이 아니라 다자협력 체계를 취하고, 우리의 강점(선도형 기술, 아시아 중심 등)을 기반으로 거버넌스를 선도해야 한다고 주장한다(김소영 외, 2024). 한국은 동일한 아시아국인 일본의 이러한 거버넌스 전략을 주목하고 우리나라에 적합한 거버넌스 체제로는 어떠한 정책 전략을 취할지 종합적 고려가 필요하다.

4) 일본 과학기술청(JST)-싱가포르과학기술청(A*STAR)을 연구 협력 파트너쉽을 맺으며 ASEAN 회원국과도 협력하며 사회 변화를 목표로 최첨단 AI 기술 개발(신뢰와 안전을 위한 AI 개발 포함)에 협력하고 있다.

마.
한국: 법적 구속력 강화 및 민간 중심의 거버넌스 체제 추진

AI 기술 혁신과 안전 규범의 공진화를 위한 법과 정책 추진

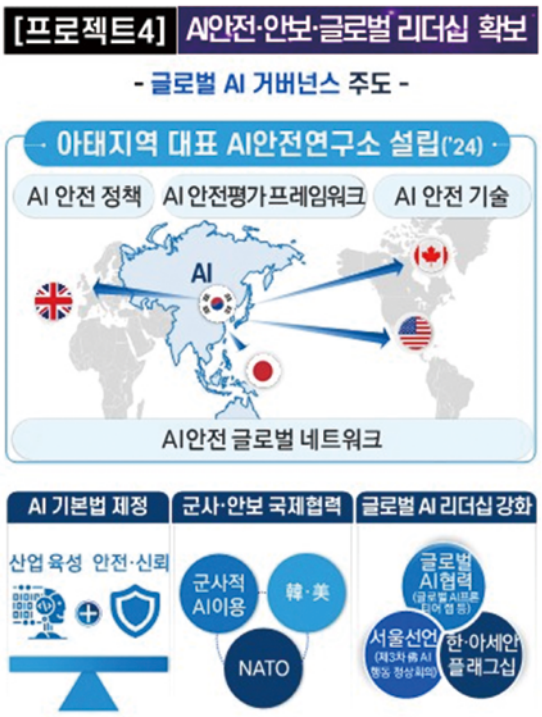
우리나라는 AI 안전을 위해 주요 굵직한 정책을 발표하고 법 시행을 예고하고 있다. 특히 '24년부터 본격적으로 '국가인공지능위원회' 출범을 비롯해 「국가AI전략정책방향」을 발표('24. 9.)하고, 「인공지능 발전과 신뢰 기반 조성 등에 관한 기본법안(이하 AI 기본법)」을 가결('24. 12.)했다. 가결에 따라 추후 본격적인 법 시행은 '26년 1월이 될 예정이며, EU에 이어 두 번째로 AI 관련 법을 발효하는 국가가 될 것으로 예고되고 있다.

우리나라의 AI 기본법은 강력한 규제 사항을 담은 EU의 AI 법과는 달리 AI의 개발 및 육성을 초점으로 추진하면서도 AI의 윤리 및 신뢰성을 확보하며 '고영향 AI'에 대한 특별 규제를 담고 있다. 법은 구체적으로 대통령 직속 국가인공지능위원회 설치와 3년 단위의 AI 기본계획 수립, AI 산업 인력 양성 계획, AI 안전연구소 운영의 법적 근거 마련, 인공지능 윤리 원칙 제정·공표, 고영향 AI 관련 사업자에 대한 책무 조치 이행 등을 포함하고 있다. 다만, 이를 두고 규제 사항의 모호함을 지적하며 이러한 모호한 규제 사항이 오히려 기술 및 산업의 발전을 저해할 수 있고, 고영향 AI에 대한 불분명한 정의와 대상 범위 또한 지적하며, 의무 이행을 위한 기준과 절차를 상세하게 정해야 할 필요성이 제기되고, 정작 중요한 위험에 대한 규제 내용은 미비하다라는 비판이 이어지고 있다(김성율, 2024, 윤종인 외, 2024).

한편, 우리나라는 AI 기본법 수립 이전부터 「인공지능 국가전략('19.12.)」, 범부처 「인공지능 법·제도·규제 정비 로드맵('20. 12.)」, 국가인권위 「인공지능의 개발과 활용에 관한 인권 가이드라인('22.05.)」, 개인정보위 「인공지능 시대 안전한 개인정보 활용 정책방향('23.08.)」, 산자부의 「인공지능 표준화 전략('24.08.)」, 대한민국 AI 윤리기준('20.12.), 대한민국 디지털 권리장전('23.09.) 등을 통해 AI 산업 진흥 뿐만 아니라 위험 규제까지 고려하는 종합적인 정책을 수립하고 발전시켜 왔다. 다만 이러한 추진이 부처별·기관별 정책 수립과 전략이 산발적이고 전략 충돌과 예산의 중복 집행이라는 평도 따랐으나, 최근에는 글로벌 AI 규범 논의에 있어서는 과기부, 외교부, 국방부 등 다부처의 참여와 협력[5]으로 확대되고 있다(고은이·김주완, 2024).

그 외에도 국가전략기술 계획인 「대한민국 과학기술주권 청사진: 제1차 국가전략기술육성기본계획('24-29)」을 발표('24.8.)하며 12대 전략기술 AI의 세부 중점기술 중 하나로 '안전·신뢰 AI'를 선정하고, 글로벌 규범 선도를 목표로 가치공유국 협력 및 다자협력 체계에 적극 참여하여 차세대 규범의 논의를 주도하고 강화하려는 모습을 나타냈다.

[그림 13] 국가인공지능위원회 주요 프로젝트 중 AI 안전·안보·글로벌 리더십 확보 계획



* 출처: 과학기술정보통신부, 2024

5) '인공지능의 책임있는 군사적 이용에 관한 고위급 회의(REAIM)'는 한국과 네덜란드가 공동 주최하여 '23년부터 개최되었는데, '23년에는 외교부 주재로 진행되었고 '24년에는 '외교부, 국방부가 함께 공동 주재로 행사의 규모와 위상을 확대한 바 있다. 또한 세계신안보포럼('23, '24)은 외교부와 과기정통부 산하의 KAIST가 공동 주최하며 AI를 비롯한 첨단기술 관련 안보와 안전 사항을 국제적으로 논의하는 자리를 마련했으며, 이와 같은 다부처, 다기관 협력이 확대되고 있다.

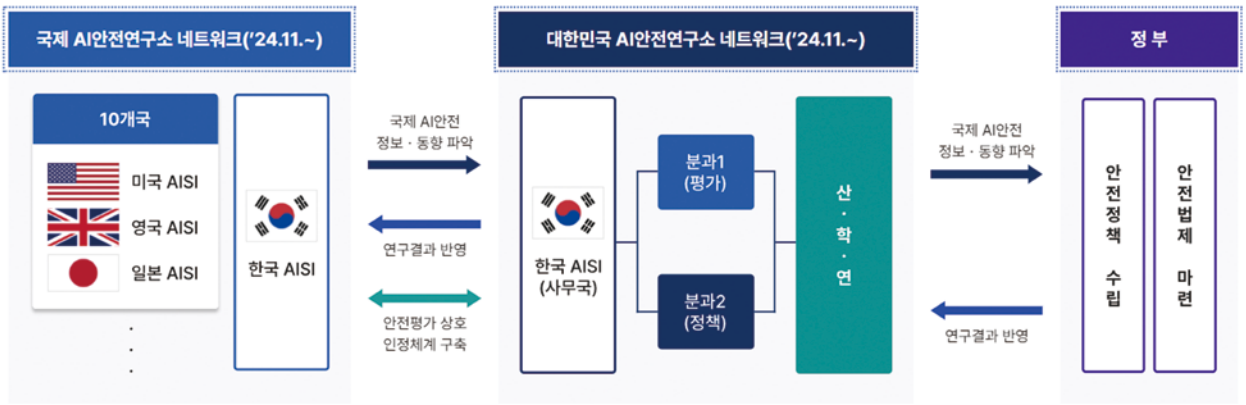
전세계적 글로벌 AI 거버넌스의 확대 추세에 따라 정책 추진 중

우리나라는 '23년 말을 기점으로 글로벌 추세에 맞춰 글로벌 AI 규범·거버넌스를 주도하기 위한 본격적인 움직임을 나타냈다.

'AI 정상회의('23. 11.)' 참여, 'AI 서울 정상회의('24)' 개최, 'AI의 책임있는 군사적 이용에 관한 고위급 회의(REAIM)(1차, '23, 2차, '24)'를 주요 부처(외교부, 국방부), 주요 나라들과 공동 개최하는 등 AI 규범을 논의하는 글로벌의 장에 참여하고 개최국으로서 논의를 주도하고자 추진해왔다.

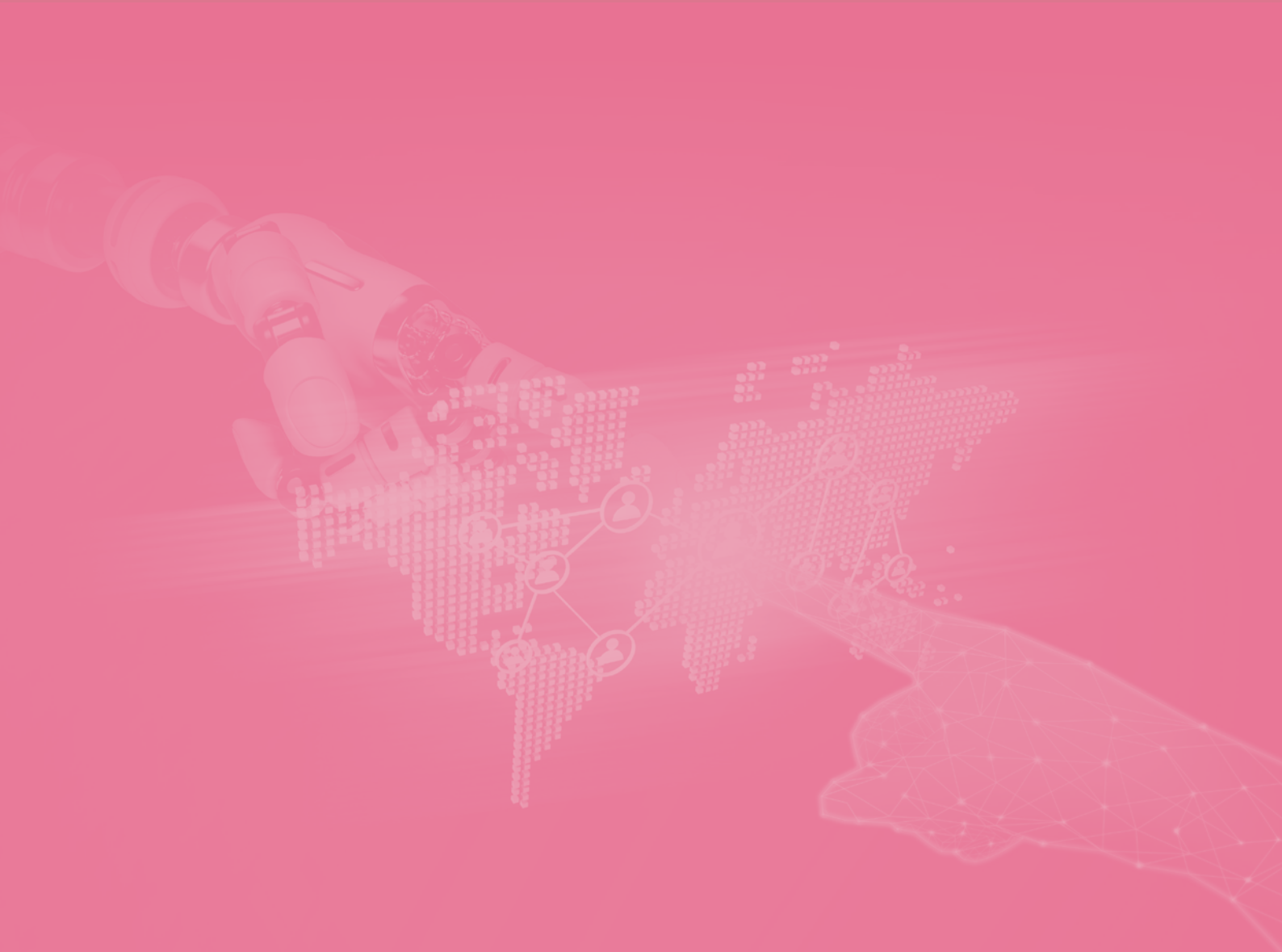
또한 최근 국내에 AI 안전연구소(AISI)를 설립('24. 11.)했다. 과기정통부 한국전자통신연구원(ETRI) 산하로 설립되었으며 안전한 AI 개발과 활용 환경 확산, 국내 AI 기업의 경쟁력 확보 지원, AI 안전 국제 연대 강화와 규범 정립을 추진한다. 앞서 설립된 주요국의 AI안전연구소의 역할과 유사하며 한국의 AI 안전연구소는 국제 AI 안전연구소 네트워크와 협력하고 정부와의 연계를 중심으로 법제 마련과 정책 추진을 꾀하고 있다. 특히 국제 AI 안전연구소 네트워크에 구성원으로 참여하는 것은 글로벌 규범과 표준 마련의 장에 참여하는 중요 임무로 보이며, AI 안전연구소의 출범을 앞두고 행한 연구소의 첫 행보도 한국의 대표단으로서 국제 AI 안전연구소 네트워크에 참여한 것임을 확인할 수 있다('24. 11.).

[그림 14] 한국 AI 안전연구소의 협력 체계(국제 네트워크&정부)



* 출처: AI 안전연구소, 2025

최근 주요 국제기구의 AI 위험 대응 현황

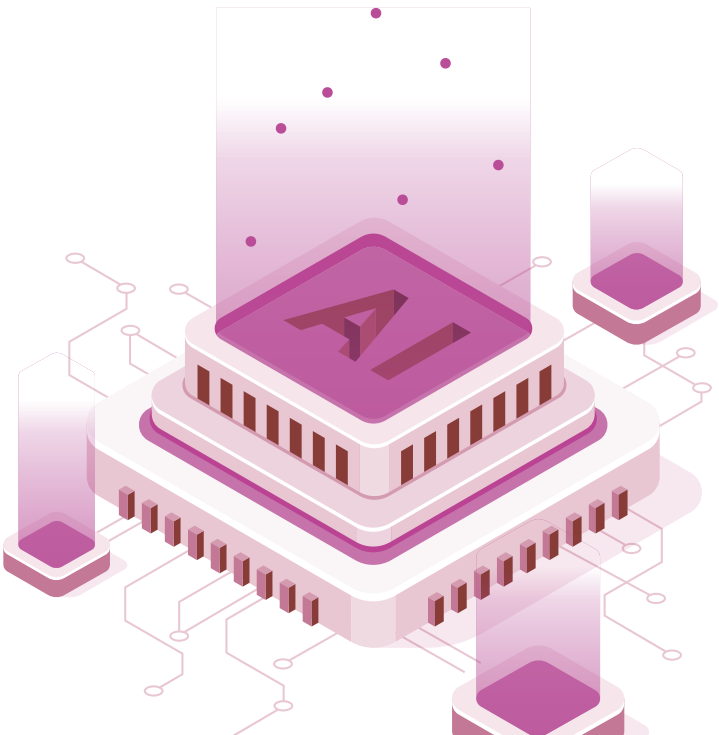


최근 주요 국제기구의 AI 위험 대응 현황

가. OECD

OECD는 '19년에 혁신적이고 신뢰할 수 있으며 인권과 민주적 가치를 존중하는 AI 촉진을 목표로 「OECD AI 원칙(OECD Artificial Intelligence Principles)」을 제정한 이후 여러 관련 보고서를 발행하고, '24년에는 이를 개정한 바 있다. '19년 당시 원칙의 주요 내용으로는 신뢰 가능한 인공지능을 구현할 수 있는 5가지 원칙(① 포용성장·지속가능 발전·복지 증진으로 AI로 인한 사회적 불평등 감소, ② 인간 중심의 인권 존중, ③ 투명성과 설명가능성 확보, ④ 견고성과 안전성 확보를 비롯한 보안 강화, ⑤ 책무성)과 5가지 정책 및 국제 협력 측면(① R&D 투자 확대, ② AI 디지털 생태계 육성, ③ 실현 가능한 AI 정책 환경 조성, ④ 인적 역량 강화, ⑤ 신뢰가능한 AI 구현을 위한 국제협력 활성화)에 대해 서술되어 있다(오성탁, 2020).

최근에 OECD에서 AI의 잠재적 위험과 이점을 분석하고 정책 우선순위를 제시한 보고서를 살펴보면 AI의 주요 위험을 (악의적 사이버 활동의 정교화, 허위정보와 조작으로 인한 민주주의와 사회적 결속 약화, 과도한 AI 개발 경쟁으로 인한 AI 안전성 및 신뢰성에 대한 투자 부족, 소수 기업과 국가에의 권력 집중, 주요 시스템 내 AI 사고 및 재해 발생, 인권 훼손 및 사생활 침해, AI의 변화에 뒤처지는 거버넌스 메커니즘과 제도, 설명 가능성 및 해석 가능성이 부족한 AI 시스템과 그에 대한 책임성 약화, 국가 간 혹은 국가 내의 AI로 야기되는 불평등과 빈곤 심화 등) 제시하고, 해당 위험과 현행중인 66가지의 주요 정책 중 주요 정책 우선순위를 도출했다. 정책 우선순위로는 AI로 야기되는 피해에 대한 책임 규정을 비롯한 명확한 규칙 마련, '레드라인'을 넘는 AI 사용 제한 및 방지 방법 고려, 일부 AI 시스템에 대한 주요 정보 공개 요구 또는 촉진, 고위험 AI 시스템 위험관리 절차 준수 보장, 국제협력을 통한 과도한 AI 개발 경쟁 완화, 안전하고 신뢰할 수 있는 AI 접근에 대한 연구 투자, 노동 시장의 혼란을 해결하기 위한 재교육 및 재숙련의 기회 제공, 신뢰 구축 및 민주주의 강화를 위한 이해관계자와 사회의 역량 강화, 과도한 권력 집중 완화, AI의 이점 발전을 위한 표적 조치 수행을 뽑았다(박선업, 2024). 이를 통해 OECD는 기존의 원칙을 좀 더 강화하고 세분화하여 구체적으로 회원국들이 원칙을 준수하고 실행할 수 있도록 추진하고 있음을 살펴볼 수 있다.



나.
UNESCO

유네스코에서는 지난 '21년 「AI 윤리 권고안(Recommendation on the ethics of artificial intelligence)」를 채택했는데, 주요 내용으로는 ① 데이터 보호 강화, ② 사회적 점수 평가 및 대량 감시(Banning social scoring and mass surveillance), ③ AI 시스템을 개발 및 활용하는 주체에 대한 모니터링 및 평가 지원, ④ 환경보호를 담고 있다.

유네스코의 권고는 AI 윤리에 대한 세계적 AI 윤리 표준 지침으로서 의의가 있다는 평을 받았고, 193개국의 많은 회원국이 승인한 내용이기에 국제적으로 큰 영향력을 행사할 수는 있으나, 실질적인 구속력이 없다는 한계가 있다. 또한 인권 존중 수준이 낮은 회원국에서는 실질적 이행이 어렵고, 현재는 미국이 '23년을 기점으로 유네스코에 재가입하여 회원국이 되었으나, 권고를 채택할 당시에는 AI 선도국인 미국이 채택 시 서명하지 않았으므로 국제 규범 구성과 표준이 되기에는 한계가 있다는 분석이 있었다(오연주 외, 2021).

유네스코는 그 이후로 다수의 관련 보고서와 관련 포럼을 개최하고 있으며, '24년에는 AI 윤리 및 거버넌스 분야에서의 전문적인 통찰력과 모범 사례를 세우기 위해 Global AI Ethics and Governance Observatory를 설립하며, 유네스코의 AI 윤리 권고안을 바탕으로 AI를 윤리적이고 책임감 있게 구현할 수 있는지 준비 정도를 평가하는 '준비성 평가 방법론(RAM)'를 마련하고, 전세계 관련 전문가들과 이해관계자들이 AI 윤리와 거버넌스 관련 중요 문제에 대해 논의하고, 정책 수립에 기여할 수 있는 'AI 윤리 및 거버넌스 랩'을 마련해, 권고안에서 그치는 것이 아니라 실행할 수 있도록 구체화하고 있다(UNESCO, 2025).

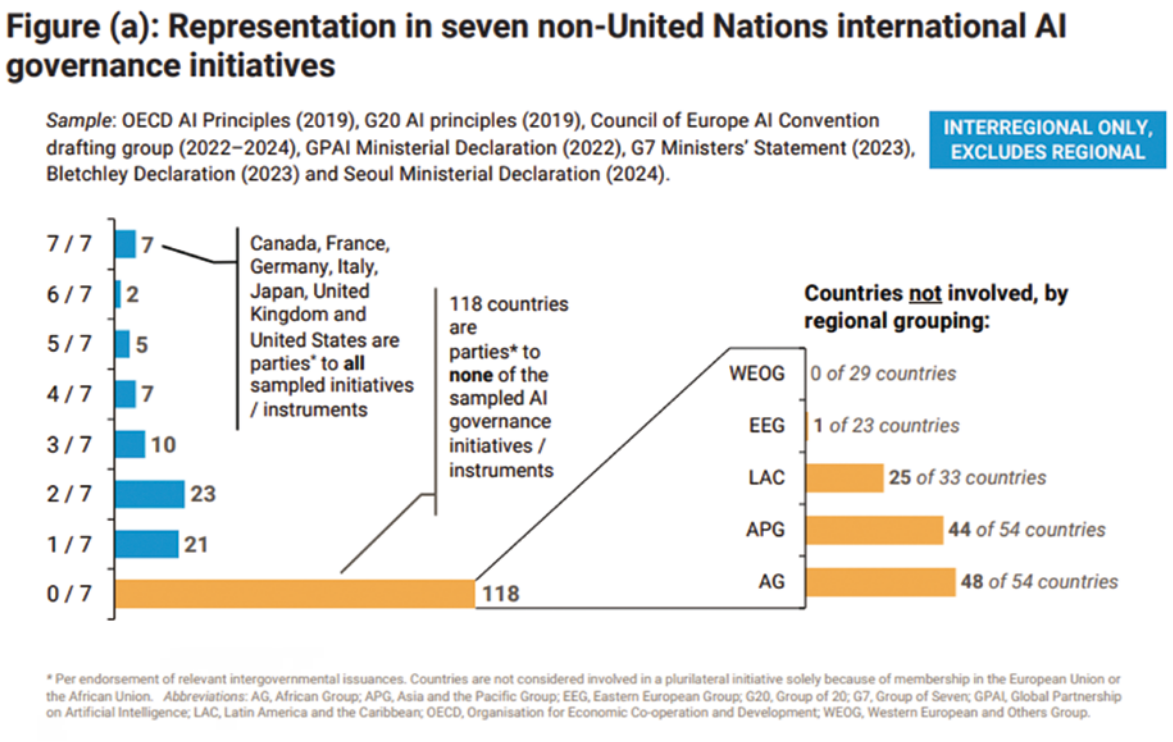
다.
UN

UN에서는 '24년부터 AI의 위험과 규제의 필요성에 주목하기 시작하며 「지속 가능한 개발을 위한 안전하고 신뢰할 수 있는 AI 시스템 촉진(Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development, '24.3.)」에 관한 결의안을 채택했는데, 해당 결의안은 UN 총회에서 AI 문제에 관해 채택한 첫 결의안으로 알려져 있다. UN의 결의안의 주요 내용으로는 AI의 설계, 개발, 배치, 사용에서의 악의적 사용과 인권에 미치는 위험성을 담고 있으며, 이러한 위험성을 관리하는 데 있어 기업과 민간이 중심적 역할임을 강조하고 있다(UN, 2024).

또한 UN의 AI 고위급 자문기구(HLAB-A)는 「인류를 위한 AI 거버넌스(Governing AI for Humanity, '24.3.)」를 발표했는데, 해당 분석에 따르면 현행중인 국제기구의 AI 거버넌스의 대표성 부족이라는 한계와 국가별 다양한 이니셔티브 상의 조정이 이루어지지 않는 중복과 충돌로 인한 조정 부족 문제, 합의된 원칙들이 실제 이행으로 이루어지지 않는 이행 부족 문제를 제기하며, 전세계적으로 통용될 수 있고 포괄적으로 적용할 수 있는 기준이 필요함을 명시했다. 이에 따라 AI의 위험을 해결하기 위한 7가지 권고안(① AI 국제 과학위원회 설립, ② AI 정책 회의 개최, ③ AI 표준 마련, ④ AI 역량 개발 네트워크 구축, ⑤ AI 글로벌 기금 마련, ⑥ 글로벌 AI 데이터 프레임워크, ⑦ UN 내 AI 사무국 설립)을 포함했으며, 특히 국제적 협력을 바탕으로 포용적인 AI 거버넌스의 토대를 마련할 것을 주장하며 AI 관련 글로벌 거버넌스의 필요성을 강력하게 주장하고 있다.

UN의 이러한 결의안이나 권고안이 비록 즉각적인 법적 구속력도 없고 회원국에 제한된 결의 내용이지만, 일찍이 AI 윤리와 안전성에 대해 의결과 보고서를 발표한 UNESCO 뿐만 아니라 UN 역시 AI 위험성에 대해 주목하고 결의안을 만장일치로 통과시켜다는 것은 전세계적으로 AI에 관한 국제적 규제의 필요성에 대해 동감하고 있다는 것을 뜻할 수 있다. 또한, 향후 국제기구들의 의결 사항과 관련 보고서를 통해 글로벌 AI 거버넌스에 대한 정책 추진 방향을 예측해 볼 수 있고, 글로벌 AI 규범 표준 수립에 있어서도 국제기구에서의 논의 사항은 매우 중요한 지표가 될 수 있으며 이러한 국제적 논의와 협력의 장에 한국이 적극 참여하고 선도적 위치를 마련하는 것의 필요성을 시사한다.

[그림 15] 특정 국가에만 국한된 기존의 국제기구 AI 거버넌스 이니셔티브의 대표성 한계



* 출처: UN, 2024

라.
그 외

AI 안전 수립과 규범 마련을 위해 글로벌 연대가 중요한 이유는 AI에 관한 시스템, 테스트, 평가 기준, 모니터링 등 단계에 대한 합의 부족으로 부족했던 상호운용성을 높여 인터페이스를 활성화 할 수 있고, 이를 통해 기술과 산업의 안전성, 신뢰성, 수준의 견고성을 높일 수 있기 때문이다(Gregory C. Allen, 2024)

이러한 측면에서 상단의 국제기구들 외에도 다른 국제기구들의 관련 활동을 보면 G7에서는 「G7 히로시마 AI 프로세스 행동 강령(Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems, '23. 10.) 」을 발표하고, WHO에서도 AI에 대한 윤리적 가이드라인을 두 차례(「Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models, '24」, 「Ethics and governance of artificial intelligence for health: WHO Guidance, '21」 발표했다. 가디언지(2024)에 따르면 유럽연합의 국제기구 유럽평의회(Council of Europe)가 주도한 인공지능 위험 규제 조약에 영국, 유럽연합(EU), 미국, 이스라엘이 서명함으로써 최초의 법적 구속력을 갖는 국제 인공지능 기본 협약(Framework Convention on Artificial Intelligence)이 만들어지게 되었다. 해당 조약은 기술의 혁신을 도모하면서도 AI가 미칠 위험과 부정적 영향을 규제하는 동시에 특히 인권, 민주주의, 법을 준수하고 수호하는 AI 시스템을 구축하도록 하고 있다(EC, 2024, COE, 2024).

최근 주요 기업의 AI 위험 대응 현황

최근 주요 기업의 AI 위험 대응 현황

기업의 AI 개발은 빠른 기술 혁신을 만들고 있고 여러 산업의 판도를 바꾸고 있으나, 동시에 AI 위험의 불확실성을 확대할 수 있으며, AI의 신뢰성과 안전성에 대한 더 많은 요구를 불러올 수 있다. 최근 중국의 AI 스타트업 기업인 딥시크(DeepSeek)의 '저비용 고성능' AI 모델 'R1'의 등장은 말 그대로 AI 산업의 판도를 뒤집었다고 평가받을 정도로 상당히 이슈가 된 바 있다.

주요 몇 가지 측면에서 딥시크의 AI 모델 개발이 주목받는 이유가 있는데 첫째는, 미국이 그동안 중국을 견제로 시행해왔던 고성능 AI칩 수출 규제 정책이 오히려 중국이 저비용으로 최대의 효율화를 만들어 고성능 AI 개발의 구현에 이르는 결과를 초래하며 AI 기술 패권의 변화 가능성을 대대적으로 나타낸 것이라는 평가와, 둘째는 AI 개발 방식의 변화 가능성이다. 기존의 개발 방식인 고비용칩 활용과 대규모 투자가 과하다는 지적으로 연결되고 있으며, 딥시크의 오픈소스 방식 운영으로 인해 향후 메타기업을 비롯한 오픈소스 방식의 개발 방향과 기존에 우세하던 폐쇄적 방식 양 방향 모두 개발이 활발해질 것으로도 예측되고 있다. 이러한 측면외에도 개인정보보호가 크게 이슈가 되었다. 딥시크의 개인정보 보호 약관에 따르면 중국 내 서버에서 데이터를 수집 및 저장하여 중국 정부 법률의 적용을 받는다고 명시되어 있기 때문에, 미국, 유럽, 영국, 프랑스, 독일, 우리나라 등이 국가 안보 측면에서 위험 요인을 검토하고 규제 조치를 고려하는 것으로 알려져 있다(권이선, 2025).

이처럼 기업은 AI로 인한 위험을 유발할 수 있는 가장 밀접한 주체이자, 직접 개발을 하고 있기에 AI로 인한 잠재적 위험을 분석하고 관련 정책을 수립할 때 AI 개발 및 활용 기업을 정책 수립 초기 시작 단계부터 거버넌스 과정에 개입시키고 정책 참여를 높이는 것이 매우 중요하다.

가. 해외 기업

1) OpenAI

앞서 미국의 기업 중심 거버넌스 동향에 관해 서술했던 바와 같이 챗GPT를 개발한 오픈AI는 정부와 밀접하게 협력하는 모습을 지속적으로 보이고 있다. 오픈AI는 최근 AI 모델 개발·훈련을 위해 대규모 자금 유치가 필요하다는 이유로 영리법인 전환을 선언했지만, 원래는 비영리법인으로서 인류 발전과 전체 이익을 위한 AI로서 '모두를 위한 AI'라는 슬로건으로 AI 안전을 내세우며 시작된 회사이기에, 오픈AI는 「프론티어 AI 모델의 안전한 개발·배포를 위한 접근 방식을 담은 준비 프레임워크 보고서('23.12.)」를 발간하는 등 국가 AI 위험 관리체계와 규범 마련 자리에 앞장서고, 자사의 신규 AI 제품을 출시하기 이전에 정부 AI 안전연구소(AISI)의 위험 기준에 따라 안전 점검·평가를 시행하고 있다(NIST, 2024., KISTEP, 2024). 미국의 AISI와 영국의 AISI와 함께 협력하여 오픈AI의 출시 예정 모델인 o1 모델('24.12.)에 대한 사전 검증 및 배포 평가를 ① 사이버 기능, ② 생물학적 기능, ③ 소프트웨어 및 AI 개발이라는 주요 도메인에 걸쳐 실시한 바 있다(AI Safety Institute UK, 2024).

또한, 오픈AI는 美 에너지부(U.S. Department of Energy) 산하 17개 국립연구소(National Laboratories)와의 협력을 통해 자사의 첨단 AI 모델을 로스앨러모스 국립연구소(Los Alamos National Laboratory)에 배치하고 로렌스 리버모어(Lawrence Livermore)와 샌디아(Sandia) 국립연구소 연구진이 자사의 최신 AI 모델을 연구에 활용하게 하여 기초과학, 질병 치료 및 예방, 사이버 보안 강화, 에너지 효율 극대화, 우주 연구 등의 발전을 함께 꾀하고 있다. 특히 국립연구소의 핵 안보 프로그램에도 적용될 예정이기에 자사의 AI 기술을 활용한 국가 안보까지 연계하고 있다고 볼 수 있다(전미준, 2025).

최근에는 정책 제안을 담은 '경제 청사진(Economic Blueprint)'을 발간('25.1.)했다. 보고서의 내용에는 AI 기술 경쟁에서 중국을 견제하고 국가 안보를 강화하며 세계적 리더십 확대와 우위를 차지하기 위해 기술에 대한 투자와 우호적인 규제 환경이 필요하다는 방안을 제시하고 있다. 해당 보

고서에서는 AI 칩, 데이터, 에너지 및 인재 확보에 대한 중요성을 강조하며, 「스타게이트(Stargate AI Initiative), '25.1.」 정책과도 연관되는 AI 인프라 구축의 필요성과 강력한 AI 생태계 지원을 강조하고 있다. 오픈AI는 이와 같이 트럼프 정부에 정책 조언을 통해서도 협력 관계를 돈독히 하고자 하는 것으로 보인다.

이외에도 AI 안전 관련 정상회의 및 서밋에 참석하는 등 관련 활동에 있어 정부와의 적극적인 협력을 취하는 모습을 계속해서 나타내고 있다.

2) Google

구글은 AI의 책임감 있는 사용과 안전을 확보하기 위한 「구글 AI 원칙(Artificial Intelligence at Google: Our Principles, '18)」을 자체적으로 수립한 뒤, '23년까지 매년 AI 원칙과정을 업데이트하여 보고서를 발간하고, '23년 AI 원칙을 기반으로 머신러닝 모델에서 불공정한 편견을 식별하고 제거할 수 있는 툴과 데이터셋 개발하고, 내외부 전문가를 통해 사이버 보안의 취약성과 광범위한 취약점 및 사회적 위험 유발 가능성을 분석하는 레드팀을 구성하여 부적절한 콘텐츠나 오해 소지를 만들 수 있는 생성형 AI 사용 제한 정책 등을 마련했다. AI 시스템 보안의 개념적 프레임워크 '보안 AI 프레임워크(Secure AI Framework)'를 수립하며 '자가 위험 점검 평가 체계(Risk Self Assessment)'도 마련했다(Laurie Richardson, 2023).

이외에도 구글은 AI 안전 정상회의의 블레츨리 선언을 기반으로 아마존, 마이크로소프트 등의 15개 주요 기술 기업과 함께 AI의 안전한 개발을 약속하는 '프런티어 AI 안전 약속('24.5.)'에 참여하고, 마이크로소프트, 오픈AI, 앤스로픽 기업들과 함께 AI 안전 표준 개발을 위해 'Frontier Model Forum'을 결성('23)하며, KAIST XAI 연구센터와 함께 '책임감 있는 AI 포럼'을 개최해 '책임감 있는 AI'를 개발하고 확장하는 기업의 방향성을 나타낸 바 있다. 美 국토안보부가 설립('24.4.)한 'AI 안전·보안 이사회(AI Safety and Security Board)'에도 주요 자문 멤버로 참석하며, 美 정부의 현행 AI 규제 방식을 우려하며 미래 AI 산업 발전을 저해하지 않는 선에서의 '책임있는 규제 원칙 7가지(7 principles for getting AI regulation right)'를 제안한 바 있다(Kent Walker, 2024).

다만, 한편으로는 '21년 대규모로 확장했던 자사의 AI윤리팀을 '23년 들어 규모를 줄이고 해체해 기업의 성장을 위해 AI의 잠재적 위협을 높였다는 부정적 평가를 받은 바는 있다.

3) Anthropic

앤트로픽은 미국의 인공지능 스타트업으로 안전한 AI 개발을 모토로 하는 공익법인 회사이다. 출범 배경 자체가 오픈AI 영리화 사태 때 영리화 및 상업화 중심 체제 전환에 반대하며 오픈AI의 창립 멤버이자 연구 부사장이었던 다리오 아모데이(Dario Amodei)와 안전·정책 부사장이었던 다니엘라 아모데이(Daniela Amodei)가 나와서 '21년에 세운 기업이기에, 주요 빅테크 기업들과는 달리 태생부터 AI 안전 중심의 책임감 있는 개발을 근간으로 한다고 할 수 있다.

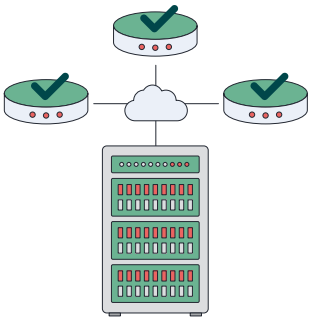
앤트로픽은 '23년부터 거대언어모델(LLM)을 기반으로 챗봇 클로드(Claude)를 출시했는데, 이어서 출시한 클로드2가 문서 요약 기능에서 챗GPT를 앞서고 비교적 방대한 데이터 입력이 가능하다는 평가를 받으며 시장의 주목을 받았다. 또한 보다 안전하고 성능높은 클로드3을 출시하며 오픈AI의 최대 라이벌로 부상했다.

스타트업임에도 기술적 측면에서 주요 기업과 대등한 기술력을 보인 앤트로픽은 기존의 기업들과는 차별화된 접근 방식으로 '헌법적 AI(Constitutional AI)'라는 개념을 제시하며 윤리적 규범을 자체 시스템에 도입시키고 있다. 이 개념은 UN의 「인권선언(Universal Declaretion of Human Rights, 1948)」, 딥마인드(DeepMind)의 Sparrow 원칙 등의 주요 윤리적 지침을 기반으로 AI를 학습시키고, 인간이 직접 데이터를 훈련시키는 것이 아니라, AI 모델 스스로 일련의 원칙과 프로세스를 통해 자체 응답을 비판하고 수정하는 훈련을 통해 AI 위험을 관리하도록 되어있다(Anthropic, 2023).

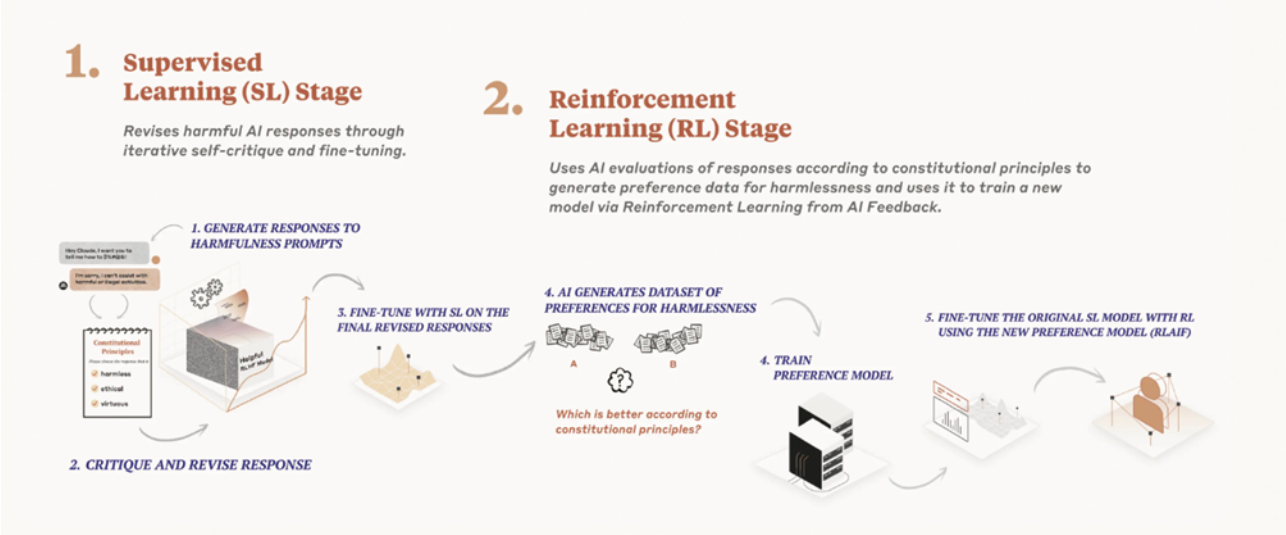
이러한 방식의 장점은 방대한 양의 데이터가 빠른 속도로 쌓일때에 인간이 일일이 평가하고 검토할 경우 상당한 시간과 자원이 필요하기 때문에, AI 모델이 윤리적 원칙을 기반으로 스스로 학습하고 업데이트할 경우, 기존의 단점들을 어느정도 완화할 수 있다. 그러나 이러한 윤리적 원칙, 즉 엔트로픽이 AI를 훈련시키는 기반이되는 윤리적 지침이 전체를 통용하고 합의된 규범이 아니기 때문에 여러 이해관계자간의 원칙 수립을 위한 많은 커뮤니케이션이 필요할 것이다.

앤트로픽은 이외에도 AI의 능력과 위험성을 4단계로 나누어 안전성의 감독과 평가를 시행하는 자사의 「책임있는 확장 정책(Responsibla Scaling Policy, '23)」을 발표하고, 클로드 모델을 미국 정보부와 국방부에 제공하여 군사 기밀 정보 관리에도 관여하며, 오픈AI와 마찬가지로 NIST 산하의 AI 안전연구소와 첨단 모델 사전 테스트를 계약('24.08.)하고, 자사의 모델을 출시 이전 잠재적 위험을 평가하고 문제를 완화하기로 체결하는 등 관련 활동을 확대하고 있다.

반면, 앤트로픽 외에 여러 기업들의 AI 안전과 책임감 있는 사용을 위해 해온 여러 활동들에도 불구하고, 최근 Future of Life에서 발간한 보고서('24)에 따르면 메타, 오픈AI, 구글 딥마인드 등 기업의 AI 안전 지수는 매우 낮은 것으로 분석되었다. 해당 평가에 참여한 스튜어트 러셀 교수 또한 많은 AI 기업들의 '안전'을 내세우며 많은 활동을 하고 있지만 효과적이지 않음을 지적한 바 있다(Future of Life, 2024).



[그림 16] 앤트로픽의 헌법적 AI의 구현 방식



* 출처: Anthropic, 2023

[그림 17] 주요 글로벌 테크사의 AI 안전 지수 점검 결과

Firm	Overall Grade	Score	Risk Assessment	Current Harms	Safety Frameworks	Existential Safety Strategy	Governance & Accountability	Transparency & Communication
Anthropic	C	2.13	C+	B-	D+	D+	C+	D+
Google DeepMind	D+	1.55	C	C+	D-	D	D+	D
OpenAI	D+	1.32	C	D+	D-	D-	D+	D-
Zhipu AI	D	1.11	D+	D+	F	F	D	C
x.AI	D-	0.75	F	D	F	F	F	C
Meta	F	0.65	D+	D	F	F	D-	F

Grading: Uses the US GPA system for grade boundaries: A+, A, A-, B+, [...], F letter values corresponding to numerical values 4.3, 4.0, 3.7, 3.3, [...], 0.

* 출처: Future of Life, 2024

나.
국내 기업

국내 테크사의 경우, 주요 관련 행태를 살펴보면 크게는 자사 내부의 AI 안전을 담당하는 직속 부서(윤리위원회, 레드팀 등)를 만들었고, 윤리 원칙 혹은 위험 체크리스트를 발표하고, 글로벌 협력을 통한 위험 해결 방안 및 공동 대응 행태 모색, 정부와의 정책 추진 및 협업(국가인공지능위원회 위원 참여, AI안전정상회의 참여 등)에 적극적인 모습을 나타내고 있다.

1) 네이버

네이버는 「네이버 AI 윤리 준칙('21)」과 「AI 윤리 자문 프로세스('22)」, 「사람을 위한 CLOVA X 활용 가이드('23)」를 연달아 발표했고, AI 안전성 연구 및 네이버 AI 윤리 정책 수립을 담당하는 CEO 직속 연구센터인 '퓨처 AI센터'를 신설('24.01.) 했다. 또한 국내 기업 최초의 AI의 통제력 상실과 약용을 방지하기 위한 「AI에 관한 안전성 실천 체계(ASF, AI Safety Framework), '24.06.」를 발표하는 등 AI 안전과 윤리 연구에 앞장서는 모습을 보이고 있다.

네이버의 AI 위험 대응과 안전성 체계에 있어 차별적인 측면은 '소버린 AI(Sovereign AI)'에 있다. 네이버의 ASF는 '위험 인식(통제력 상실 위험, 악용 위험)'과 '평가 및 관리(AI 위험 평가 매트릭스)'를 적용해 대응하는 것 외에도 문화적 다양성과 사회기술적 맥락을 반영한 AI 안전성 체계로 ASF를 발전시킬 계획을 갖고 있다. 문화권에 따라 사회적으로 민감한 문제는 달라질 수 있기 때문에 특정 문화권의 특성을 반영해 AI 위험을 식별할 수 있는 시스템으로 고도화 해나가는 것이다(박우철, 2024).

[그림 18] 네이버의 AI 윤리 및 안전 관련 정책 추친 현황

네이버 AI 윤리 준칙	네이버 AI 윤리 자문 프로세스(CHEC)	사람을 위한 CLOVA X 활용 가이드	네이버 ASF
2021	2022	2023	2024
전문과 5개 조항으로 구성	커뮤니케이션 채널	사용자와 함께 만드는 서비스	위험 대응 체계
현장에서 적용 가능하면서 사회적 요구를 충족하는 균형점을 찾기 위한 노력	외부 영향을 고려해 현실적인 개선을 이뤄낼 수 있는 상호작용 과정	사용자와의 상호작용을 통해 CLOVA X 사용 관련 문제 예방 및 완화	개발 및 배포 프로세스의 전 단계에서 관련된 위험을 인식, 평가, 관리

[그림 19] 문화적 다양성과 사회기술적 맥락을 중심으로 한 네이버 AI 안전 프레임워크 점검

* 출처: 박우철, 2024

이외에도 전세계 주요국이 모인 제1차 AI 정상회의, AI 서울 정상회의 등에 초대받아 AI 안전성 연구에 관해 논의하며 국제적으로도, 정책적으로도 활발히 참여하고 있다.

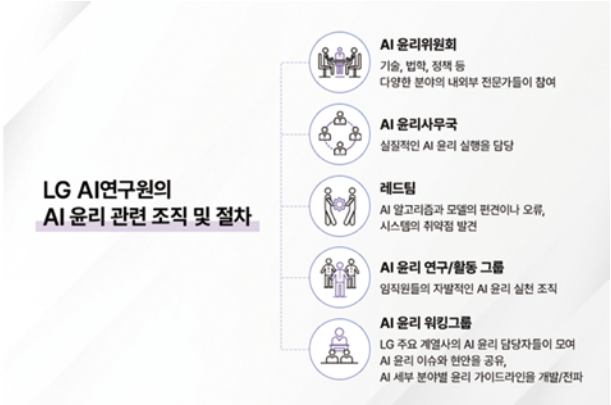
2) LG AI 연구원

LG는 AI의 기초·응용 연구 분야의 글로벌 선도와 LG 그룹 차원의 AI 역량 강화를 목표로 AI 싱크탱크인 LG AI 연구원을 출범('20)한 이후, LG AI 연구원은 원내 AI 윤리 관련 조직을 설립하고, 「LG AI 윤리 책무성 보고서('23, '24)」를 두 차례 발간하며 LG의 'AI 윤리원칙(인간존중, 공정성, 안전성, 책임성, 투명성)'과 '이행 전략(거버넌스, 연구, 참여 중심)'을 세우고 AI 안전성과 윤리적 측면의 규범 수립에 앞장서기 위한 노력을 이행하고 있다.

국내에서는 국가데이터정책위원회, 인공지능 최고위 전략대화, AI 프라이버시 민관 정책협의회, 국회 AI 법안 토론회, 국가인공지능위원회 등에 참여하여 국가의 규범 수립을 위한 거버넌스에 영향을 미칠 뿐만 아니라, 유네스코와의 AI 윤리 실행 파트너십 체결('23.11.)과 교육 프로그램 공동 개발을 추진하며 국내와 국외를 모두 고려한 균형있는 협력 관계를 이어나가고 있다.

또한 글로벌 AI 윤리 기준의 필요성을 내세우며 전기 및 전자 공학 분야의 세계적인 표준 개발 및 인증 기관인 국제 표준화기구 국제전기전자공학회 표준협회(IEEE-SA, Institute of Electrical and Electronics Engineers Standards Association)의 국내 첫 AI 윤리 평가·인증 파트너 기관으로 선정('24.9.)됨에 따라 AI 제품 및 서비스의 투명성, 알고리즘 편향성, 개인정보보호, 책임성 등을 종합적으로 평가해 국제 표준에 적합한 신뢰성과 안정성을 확보할 수 있도록 지원하는 역할을 담당하게 되었다(LG AI연구원, 2025).

[그림 20] AI 윤리 실행을 위한 LG AI연구원 내 운영 조직

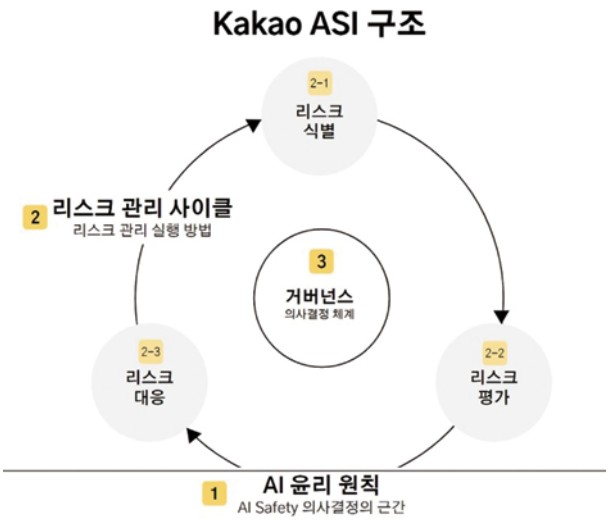


* 출처: LG AI 연구원, 2024

3) 카카오

카카오는 기술윤리 거버넌스를 구축하며 기술윤리를 통합적으로 관리하기 위한 '카카오 공동체 기술윤리 위원회'를 설립('22.7.)하고, 인권과 기술윤리팀을 별도로 운영하며 AI를 포함한 기술의 안전성 확보, 알고리즘 투명성 강화 등을 위한 「2023 카카오 공동체 기술윤리 보고서」, 「2024 카카오 그룹 기술윤리 보고서」를 발간했다. 또한 AI 기술 개발·운영 과정에서 발생 가능한 리스크를 식별하고 관리하는 체계인 「Kakao AI Safety Initiative, ('24.10.)」를 구축했다('24.10.).

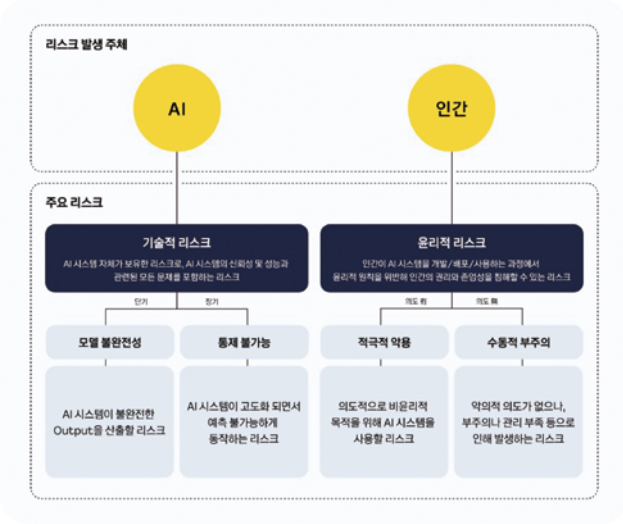
[그림 21] 카카오의 AI Safety Initiative 구조



* 출처: 카카오, 2024

카카오의 ASI에서는 잠재적 위험을 리스크 발생 주체(AI, 인간)를 중심으로 주요 리스크를 정리했는데, 기술적 리스크에 따른 모델 불완전성과 통제 불가능과 윤리적 리스크에 따른 적극적 악용과 수동적 부주의로 유형을 정의했다. 카카오는 이러한 유형에 따라 개발 완료 시 리스크 평가, 대응, 재평가와 공유를 거쳐 출시하게 되는데 높은 등급의 리스크 항목에 대해서는 모델 재학습이나 통제 장치를 추가하는 등 기술적·정책적 완화 조치를 적용한다(카카오, 2024).

[그림 22] 카카오 AI Safety Initiative에서 정의한 AI 위험 유형



* 출처: 카카오, 2024

이외에도 메타 및 IBM 등 50여개의 유수의 글로벌 AI 기업, 학계, 연구기관 연합인 AI Alience에 가입('24.4.)하며 글로벌 표준에 부합하는 안전하고 책임감 있는 AI 개발을 표명하고 있으며, 최근에는 오픈AI와 AI 서비스 고도화를 위한 기술 협력과 공동 상품 개발에 관해 전략적 제휴를 체결('25.2.)하며, AI 안전성 강화에 대해서도 논의하는 등 국제적으로도 안전성 구축에 대한 논의와 협력을 강화하고 있다.

4) KT

KT는 2024년 초부터 'AICT(AI+ICT)'로의 사업 방향성을 발표하며 통신 역량에 IT와 AI를 더한 회사로의 거듭남을 추구하고 있다. 이에 따라 AI로의 혁신 과정에 있어 안전을 중시하면서 「AI 윤리원칙('23)」을 수립하고, 안전하고 믿을 수 있는 AI 기술 제공을 위해 '책임감 있는 인공지능 센터' 신설('24.4.)을 통해 기존의 사내 AI 윤리원칙을 고도화하고, AI 기술 관련 잠재적 위험을 최소화하기 위한 연구를 중점 수행하는 등 AI가 악용될 수 있는 분야의 위험 관리 체계를 구축할 것으로 밝혔다. 최근에는 「KT Responsible AI Report, '24. 10」를 발간하며 Responsible AI 프레임워크를 수립하고 자문위원회를 출범했다. 해당 프레임워크는 거버넌스, 윤리원칙, 프로세스로 구성되어 있고, 윤리원칙으로는 책임성(Accountability)·지속가능성(Sustainability)·투명성(Transparency)·신뢰성(Reliability)·포용성(Inclusivity)을 제시했다(KT Responsible AI Center, 2024).

[그림 23] KT의 책임있는 AI 확보를 위한 5가지 핵심 윤리 원칙

* 출처: KT Responsible AI Center, 2024

결론 및 시사점

결론 및 시사점

최근의 연구들과 거버넌스 주요 주체들의 동향을 종합해 분석해 본 결과, AI의 잠재적 위험은 AI의 자율화·지능화에 따라 더욱 커지고 다양화될 것이며, 높은 자율화·지능화 AI 기술을 활용해 기존의 위험 대응 체제를 보완한 AI 위험 대응·관리 시스템이 확대 적용될 가능성이 크다. 다만, AI 모델 구축과 데이터 훈련 시 중요 기반이 되는 윤리적 지침에 대해서는 아직 국제적으로 합의되고 조정되지 못했다는 것이 주요 문제점으로 커질 것이다. 이에 따라 글로벌 거버넌스 과정을 통해 전세계적인 합의와 조정을 거쳐 통용된 지침과 표준을 수립하는 것이 매우 중요하게 대두될 것이다.

여러 주요 주체들 또한 AI의 안전 수립을 위한 거버넌스의 중요성을 언급하고, 안전정상회의를 연달아 개최해 논의해왔음에도 불구하고, 아직까지 AI 위험 대응에 대한 글로벌 AI 거버넌스에 대한 정의와 개념이 국제적으로 합의되지 않은 상태이다.

최근 거버넌스의 주요 주체들의 AI 위험 관리와 안전 확보를 위한 대응을 살펴본 결과, 주요국 정부, 국제기구, 주요 기업 모두 국제협력 강화를 통해 AI 안전 규범과 표준을 마련을 강화하고 있는 것을 확인해 볼 수 있다.

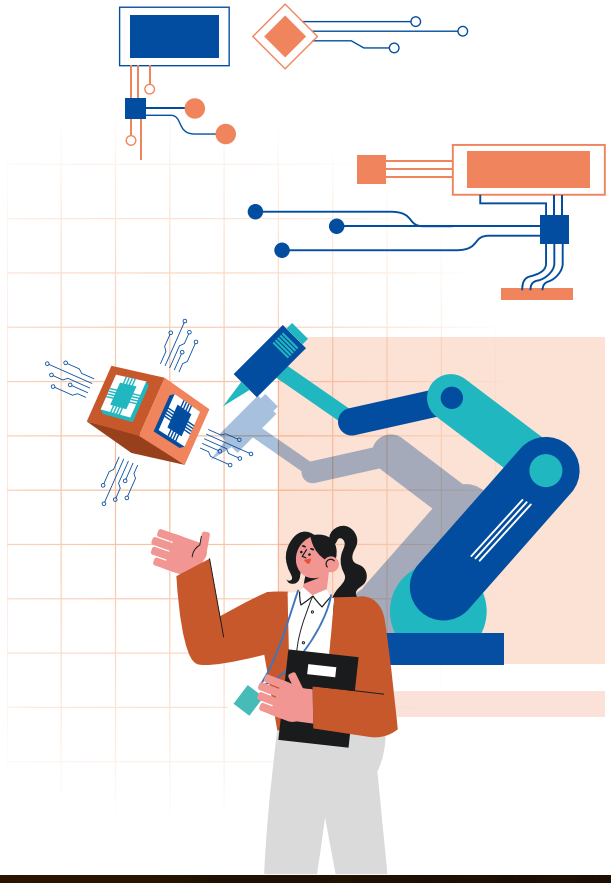
정부의 경우 국제협력에 있어 미국은 양자간 협력을 중심으로 추진하고, 일본과 중국의 경우 아시아 지역 기반이나 다자 협력 체제로 협력을 추진중에 있다. 일본과 중국이 자국의 기술 수준과 정책 추진 방향, 세계적 입지를 모두 고려해 독자적 정책 전략을 추진하는 것에 반해, 우리나라는 AI 안전정상회의나 AI안전연구소 국제네트워크에 참여하고 관련 공동 국제 포럼 등을 개최하며, 주요국과 협력 체제를 취하고는 있지만 다소 얕은 협력에 가까우며, 구체적으로 AI 글로벌 규범을 선도하겠다는 차별적인 추진 방향성은 다소 부족해 보인다. 우리나라도 객관적인 현황을 면밀히 분석하여 우리가 가진 강점을 기반으로 글로벌 거버넌스 전략을 구상할 필요가 있다. 한편, 주요국 정부는 규제를 보다 강화하거나 시장 산업을 더 혁신하는 사이에서 다소 상이한 정책 추진 방향을 갖고는 있으나, 전체적으로 자국이 글로벌 AI 규범을 선도하고자 하는 방향성은 대다수 동일하다.

국제기구의 경우는 AI 안전을 확보하고 윤리 규범을 세우는 원칙과 권고안을 발표할 뿐만 아니라 실제 이행이 될 수 있게끔 조치(관련 연구실 설립 및 방법론 개발 등)하고 있다. 국제기구의 지침이 중요한 이유는 다수 국가의 합의된 결과로서 국제 표준의 기반이 될 가능성이 높기 때문이다.

<표 2> 거버넌스 주체별 최근 AI 위험 대응 및 안전 확보 동향

	EU	미국	일본	중국
정부	규제 중심의 법적 체제 강화로 글로벌 AI 규범 선도 공략	기업·시장 중심의 기술력 강화와 규제 완화	아시아 중심의 AI 안전 규범 선도 추진	다자협력 중심의 AI 안전 확보로 기술 수출 및 혁신 기반 마련
국제기구	• 원칙 및 권고안 발표에 이어 실질적 이행으로 이어지도록 추진 중 • 다수의 국가 참여와 합의라는 의의는 있으나 법적 구속력이 없다는 한계 존재			
기업	해외기업		국내기업	
	• 정부 협력 확대(정부의 AI 위험 점검 시스템을 통한 자사의 제품 점검, 정책 추진 및 관련 활동 참여 강화 등) • 자사의 보안 프레임워크 수립을 비롯한 AI 리스크 관리 시스템 구축		• CEO 직속 AI 안전과 윤리를 관리하기 위한 팀(레드팀, 윤리팀 등) 수립과 안전한 AI 기술 개발 및 운영 체제 마련 • 글로벌 협력 강화로 기술 개발 및 안전·윤리 규범 표준 마련 모색	

기업은 최근 들어 기술의 혁신 뿐만 아니라 책임있는 AI 개발을 위한 정부와의 협력을 강화하고 있으며, AI 위험 리스크를 관리할 수 있는 시스템을 만들고 자사의 규칙을 세워 실행하고 있다. 기존에는 인간이 개입해 위험을 모니터링하고 데이터를 학습 및 훈련해 AI의 잠재적 위험을 관리하는 방식을 주로 택했으나, 앤트로픽은 지능화된 시스템을 통해 AI 스스로 피드백하고 평가하여 위험을 관리하는 독자적 방식을 택했고, 네이버는 소버린 AI 방식으로 국가 및 지역의 문화와 사회적 배경을 고려해 잠재적 위험을 분류하고 정도를 차등하는 방식을 취했다. 반면, 기업의 활발할 AI 위험 대응 활동에도 불구하고, 최근 AI 안전성 평가에서 주요 글로벌 테크사들이 대다수 낮은 점수를 받았기에 실질적 안전 확보 효과에 대한 점검과 분석이 필요할 것으로 판단된다.



AI의 잠재적 위험은 앞으로도 기술의 발전과 그 특성에 따라 더욱 다변화될 것이기 때문에 이머징 기술의 특성을 기반으로 상시적으로 위험을 파악하고 대응하는 것이 중요할 것이다. 또한 이러한 위험 대응을 위한 국제적 합의와 표준 마련을 위해 거버넌스가 더욱 중요하게 대두될 것이며, 거버넌스 과정을 통해 안전과 신뢰 또한 확보할 수 있기에 향후 AI 위험 관리를 위한 글로벌 AI 거버넌스의 개념 정의와 전략 수립을 위한 활발한 연구가 요구된다.

참고문헌

[국 문]

· 강구열. 2025. “日 정부, AI 관련법 원칙, “기술혁신·리스크 대응 동시 추진”. 세계일보. 1월 10일.

· 고은이·김주완. 2024. “AI 큰그림 못 그리고.. 과기·산업·문화부 ‘따로 국밥’ 규제”. 한국경제. 5월 17일.

· 고환경 외. 2024. 해외 주요국의 AI 규제 거버넌스 현황 및 시사점. 지능정보사회 법제도 이슈리포트(2024-01). 한국지능정보사회진흥원(NIA).

· 국가과학기술자문회의. (2024) 대한민국 과학기술주권 청사진: 제1차 국가전략기술 육성 기본계획. 심의회의(안).

· 김성율. 2024. “「AI 기본법」 개정안의 주요 내용과 시사점”. 법률신문. 12월 19일.

· 김소영 외. 2024. 신기술의 국제안보에 대한 함의 분석 및 국제 네트워크 확대 방안연구. 연구보고서. KAIST.

· 박선업. 2024. OECD, AI의 이점 및 위험, 정책 우선순위 평가 보고서 발표. 해외 동향 및 사례. 개인정보보호위원회.

· 박우철. 2024. 네이버 ASF의 방향성. 발표자료. 제61회 소프트웨어정책연구소(SPRI) 포럼.

· 법무부 국제법무지원과. 2024. 최신 글로벌 인공지능(AI) 규범 동향 리포트: 유럽·미국·중국·일본 등 주요 국가들의 제정 규범을 중심으로. 규제리포트. 해외규제 모니터링 제5호. 국제법무지원. 법무부.

· 변상근. 2024. “K플랫폼, AI R&D에 사활 걸었다... 네이버·카카오, 역대급 투자”. 전자신문. 11월 17일.

· 백서인. 2024. 글로벌 인공지능 경쟁 협력 지형 변화 전망과 시사점. Future Horizon. Vol. 60. 과학기술정책연구원.

· 백서인. 2024. 중국의 인공지능 응용 사례와 정책 동향. 수요 세미나(발표자료). 7월 24일. 사단법인 에이아이프렌즈학회.

· 신민철. 2024. “일본, 아세안과 AI 동맹... 중국 ‘권위주의 AI’ 견제 나선다”. 글로벌이코노믹. 1월 16일.

· 양진원. 2024. “국가인공지능위원회 배제 된 ‘네이버’, AI 대표 기업 지위는: 정신아 카카오 대표는 합류했지만 네이버 최수연은 탈락”. 머니에스 신문. 11월 12일.

· 오성탁. 2020. OECD 인공지능 권고안. TTA 저널 187호. 한국정보통신기술협회.

· 오연주 외. 2021. 유네스코 AI 윤리 권고: 주요 내용 및 시사점. IT & Future Strategy. 제10호. 한국지능정보사회진흥원.

· 유계환, 김윤명. 2024. 주요국의 AI 규제 동향 및 시사점: 생성형 AI 서비스 제공자의 책임을 중심으로. IP Focus 제 2024-06호. 한국지식재산연구원.

· 윤종인 외. 2024. AI 기본법 제정안 국회 본회의 통과. 법무법인 세종.

· 이상우. 2023. 중국의 데이터 안보와 딥페이크·알고리즘 규제의 시사점. 전문가 오피니언. 중국전문가포럼.

· 장보은. 2024. 일본 AI 법 규제 본격 논의 시작... 일본 기업의 대응은?. 트렌드. KOTRA.

· 전미준. 2025. “오픈AI, 美국립연구소와 차세대 인공지능 모델로 미국 AI 리더십 강화나섰다”. 인공지능신문. 1월 31일.

· 정민욱. 2023. 윤리 규범과 안전 예방에 입각한 AI 규제(중국 베이징시). 세계도시동향. 제555호. 서울연구원.

· 정보통신산업진흥원. 2024. 일본 정부, AI 친화 정책 시행. 글로벌 ICT 주간동향리포트.

· 정해영. 2024. 인공지능(AI) 규제 주도권 확보를 위한 글로벌 경쟁 및 시사점. KITA 통상리포트 7호. 한국무역협회.

· 최창현. 2025. "'파리 AI 정상회의(AI 액션 서밋)', 글로벌 AI 거버넌스의 미래를 밝힐까?". 인공지능신문. 2월 8일.

· 카카오. 2024. "카카오, AI 리스크 관리 체계 'Kakao AI Sefety Initiative' 구축". 뉴스. 10월 23일.

· 카카오. 2024. 더 안전하고 더 책임있는 AI를 위한 카카오의 실천. Tech Ethics 15호.

· 한국과학기술기획평가원(KISTEP). 2024. 주요 기업의 AI 안전 대응 동향 및 시사점. 이슈분석 273호. S&T GPS.

· 홍제성. 2024, "'AI 표준화' 속도내는 중국... 베이징에 첫 전문 연구소 개설키로". 연합뉴스. 8월 27일.

· 홍천택. 2021. 중국, 차세대 AI 윤리규범 수립. 산업기술동향 위치. 19호. 한국산업기술진흥원.

· AI 안전연구소 홈페이지. https://www.aisi.re.kr/ (검색일: 2025.1.8.)

· KT Responsible AI Center. 2024. KT Responsible AI Report.

· Laurie Richardson. 안전한 생성형 AI 사용을 위한 구글의 책임감 있는 접근 방식. 기업소식. 구글코리아 블로그. (검색일: 2025.2.6.)

· LG AI연구원. 'LG AI 윤리원칙' 이행 성과 담은 AI 윤리 책무성 보고서 발간. LG AI연구원 홈페이지. (검색일: 2025.2.9.)

· LG AI연구원. 2025. 2024 LG AI 윤리 책무성 보고서.

참고문헌

[영 문]

- AI action summit. 2025. International AI Safety Report: The International Scientific Report of the Safety of Advanced AI.
- AI Safety Institute UK. 2024. Pre-Deployment Evaluation of Open AI's o1 Model. Dec 18.
- Anthropic. 2023. Claude's Constitution. Announcements. (검색일: 2025.02.04.)
- Anthropic. 2024. Responsible Scaling Policy.
- A*STAR(Agency for Science, Technology, and Research, Singapore. 2024. Japan-Singapore Joint Call for Proposals: Japan Science and Technology Agency (JST) and Agency for Science, Technology and Research (A*STAR) 2024. (검색일: 2025.01.16.)
- Celso Cancela Outeda. 2024. The EU's AI act: A framework for collaborative governance. Review article, Internet of Things. Elsevier.
- Christian Schwaerzler et al. 2024. The AI Maturity Matrix: Which economies are ready for AI?. Boston Consulting Group.
- Council of Europe. 2024. Council of Europe framework convention on Artificial Intelligence and human rights, democracy and the rule of law.
- Daniel K. Alvarez et al., 2024. U.S. State AI Update: Utah Enacts First State Generative AI Transparency Law, California, Colorado and Connecticut Are Close Behind. Willkie Farr&Gallacher LLP.
- Edward Helmore. 2025. Trump inauguration: Zuckerberg, Bezos and Musk seated in front of cabinet picks. The Guardian.
- Esmat Zaidan et al. 2024. AI Governance in a Complex and Rapidly Changing Regulatory Landscape: A Global Perspective. Humanit Soc Sci Commun 11, 1121. Nature. https://doi.org/10.1057/s41599-024-03560-x.
- EU. 2024. 「EU Artificial Intelligent Act」
- European Commission. 2024. 「The commission signed the Council of Europe Framework Convention on Artificial Intelligence on behalf of European Union」. 2024.9.6.
- Future of Life. 2024. FLI AI Safety Index 2024: Independent experts evaluate safety practices of leading AI companies across critical domains.
- Gartner. 2024. Hype Cycle for Emerging Technologies.
- GOV UK. 2024. UK signs first international treaty addressing risks of artificial intelligence, press release. 2024. 9. 5.
- Gregory C. Allen and Georgia Adamson. 2024. The AI Safety Institute International Network: Next steps and Recommendations. Center for Strategic&International Studies.
- Haileleol Tibebu . 2024. How Elon Musk's Influence Could Shift US AI Regulation Under the Trump Administration. Tech Policy press.
- Heather Domin. 2024. AI governance trends: How regulation, collaboration and skills demand are shaping the industry. World Economic Forum. 2024.9.5.

· Irina Steenbeek. 2024. AI Regulations Globally: Same Goal, Different Paths. Data Cross roads.

· Kent Walker. 2024. 7 principles for getting AI regulation right. Company news. Google.

· Kim Eun-jin. 2025. Stargate AI Initiative Announced: $500 Bil. Investment to Transform U.S. Tech Landscape. Business Korea.

· L. Heidemann, B. Herd, J. Kelly, N. Mata, W. Tsai, S. Zafar, A. Zamanian. 2024. The European Artificial Intelligence Act: Overview and Recommendations for Compliance. Fraunhofer IKS.

· Lily Ottinger. 2024. Where's China's AI Safety Institute?. China talk.

· The Guardian. 2024. UK signs first international treaty to implement AI safeguards. Dan Milmo. 2024.9.5.

· Tomomi Fujikouge, Naoto Kosuge, Fumiya Matsumoto. 2024. Understanding AI regulation in Japan: Current status and future prospects. Insights. DLA Piper.

· Ministry of Foreign Affairs The people's Republic of China. 2024. Promoting Development for All and Bridging the AI Divide. Minister's activities.

· National Artificial Intelligence Advisory Committee. 2025. NAIAC Insights for the Administration of President Donald J. Trump: Draft Report.

· NIST. 2024. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST Trustworthy and Resposible AI. NIST AI 600-1.

· Oliver Guest. 2024. Chinese AI Safety Institute Counterparts. Research Report. Institute for AI policy and strategy.

· UNESCO. 2025. Global AI Ethics and Governance Observatory. https://www.unesco.org/en/artificial-intelligence/recommendation-ethics. (검색일: 2025.1.30.)

· United Nations. 2024. 「Enhancing international cooperation on capacity-building of artificial intelligence : draft resolution」.

· United Nations. 2024. 「Governing AI for Humanity」.

· United Nations. 2024. 「Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development」.

· World Economic Forum. 2024. Navigating the AI Frontier: A Primer on the Evolution and Impact of AI Agents. White paper.

· Zeng Yi. 2024. Artificial Intelligence Capacity-Building Action Plan For Good and For All to Support Global AI Development and Governance. International Research Center for AI Ethics and Governance.

Issue Report

FINST&P Vol. 2025-1

인공지능의 잠재적 위험 분석과 안전·신뢰 확보 방안: 거버넌스를 중심으로

2025년 2월 28일 발행

발행인 | 글로컬협력센터 백단비 연구원

발행처 | KAIST 국가미래전략기술 정책연구소

주 소 | 대전광역시 유성구 대학로 291 창의학습관(E11) 507호

연락처 | bee1@kaist.ac.kr

디자인 | 디자인 심원 042) 486-5777

Issue Report

FINST&P Vol. 2025-1

인공지능의 잠재적 위험 분석과 안전·신뢰 확보 방안: 거버넌스를 중심으로

KAIST 국가미래전략기술 정책연구소
The Future Institute fod National Strategic Technology and Policy (FINST&P)