책임 있는 AI를 위한 기업의 노력과 시사점

Corporate Efforts for Responsible AI and Implications

Executive Summary

최근 전 산업과 일상에서 AI의 활용이 폭넓게 이뤄지고 있으나, 한편으로는 AI 위험에 대한 우려 및 AI로 인한 사건 수가 증가하면서 AI 위험 대응 요구도 확대되고 있다. 이에 따라 각국 정부와 학계, 업계 등 이해당사자가 AI의 위험을 방지하고 안전하고 신뢰할 수 있는 AI를 개발 및 도입하기 위해 노력하고 있다. 본 보고서에서는 액센츄어와 스탠퍼드 대학교가 실시한 글로벌 기업의 책임 있는 AI에 대한 조치 인식 조사를 인용하여 책임 있는 AI 영역별 대응 수준을 진단하고, 주요 기업별 전담 조직 및 AI 안전 프레임워크 현황 사례를 조사하여 기업의 구체적인 책임 있는 AI에 관한 노력에 대해 살펴보았다.

액센츄어와 스탠퍼드 대학교의 조사 결과, 기업은 개인정보 보호 및 데이터 거버넌스, 신뢰성 및 보안, 투명성 및 설명 가능성, 공정성 등 책임 있는 AI의 요인별 대응을 추진하고 있으며, 개인정보 보호 및 데이터 거버넌스 측면의 대응 수준이 가장 높게 진단되었다. 그러나, AI 모델 발전에 따른 결과 설명의 어려움, 국가별 공정성의 기준에 대한 차이 등의 사유로 투명성 및 설명 가능성, 공정성 부문에 대한 향후 조치를 향상시킬 필요가 있다.

국내외 기업별 사례 조사 결과, 주요 기업들은 AI 모델의 평가와 개발·배포 여부에 대한 의사결정을 할 수 있는 전담 조직을 설립하고, 전담 조직에 의해 AI의 위험성을 정의하고 평가하는 체계를 구축하고 있다. 국내 기업은 계열사 간 컨센서스를 위한 협의체를 운영하는 특징이 있으며, 산업으로의 AI 적용을 위한 과제별

소프트웨어정책연구소 AI정책연구실

장진철 선임연구원 jincheul@spri.kr
노재원 선임연구원 jwnoh@spri.kr
유재흥 책임연구원 jayoo@spri.kr
조지연 선임연구원 jy.cho@spri.kr

위험 요인을 분류하고 평가하는 체계를 도입하고 있다. 글로벌 조사와 유사하게, 공정성 부문은 제도적인 가이드라인 수준이며 기업의 실질적인 조치가 미흡한 상황으로 향후 개선이 필요하다.
본 보고서의 결과는 각국 정부가 AI 규제에 관한 논의와 실행을 본격화되는 가운데, 기업들이 전담 조직을 구축하고, AI 안전 프레임워크를 수립 및 준수함으로써 책임 있는 AI를 정착시키기 위해 노력하고 있음을 보여준다. 앞으로 국내외 기업들의 안전하고 책임 있는 AI 개발 및 사용을 위한 지속적인 노력이 요구된다.

Recently, AI has been widely utilized in all industries and daily life, but on the other hand, as concerns about AI risks and the number of incidents caused by AI increase, the demand for AI risk response is also expanding. Consequently, all stakeholders, including governments, academia, and industry, are working to prevent AI risks and ensure the development and implementation of safe and trustworthy AI. This report cites a survey of global companies' awareness of responsible AI measures conducted by Accenture and Stanford University to diagnose the level of response in each area of responsible AI, and investigates case studies of dedicated organizations and frameworks in major companies to explore specific efforts towards responsible AI.
According to research conducted by Accenture and Stanford University, Global survey results show that companies are pursuing responses to responsible AI factors such as privacy protection and data governance; reliability and security; transparency and explainability; and fairness. The response level in privacy protection and data governance was diagnosed as the highest. However, due to difficulties in explaining the result of advanced AI models, challenges in processing different languages, and differences in fairness standards across countries, there is a need for improved measures in transparency, explainability, and fairness in the future. As a result of the survey of domestic and global companies, major companies are establishing dedicated organizations capable of evaluating AI models and making decisions on whether to develop and distribute them, and are establishing a system to define and evaluate the risks of AI through dedicated organizations. Domestic companies are characterized by operating a consultative body for consensus among affiliates, and are introducing a system to classify and evaluate risk factors for each task for applying AI to the industry. Similar to the Accenture survey, the fairness sector is at the level of institutional guidelines, and actual measures by companies are insufficient, so it can be said that improvement is needed in the future.
The results of this report show that while governments around the world are discussing and implementing AI regulations, companies are making efforts to establish responsible AI by establishing dedicated organizations and establishing and complying with frameworks. In the future, efforts will be required to develop and use safe AI technology across the entire AI ecosystem.

I. 서론

1. 연구 배경 및 목적

■ 최근 생성형 AI의 부상은 AI의 일상화와 전 산업으로의 AI 도입 확대에 기여

- 맥킨지의 조사에 따르면, 2023년 전 세계 조직의 55%가 AI를 도입하고 있으며 이는 2022년 대비 5%p, 2017년 대비 35%p 증가[1]

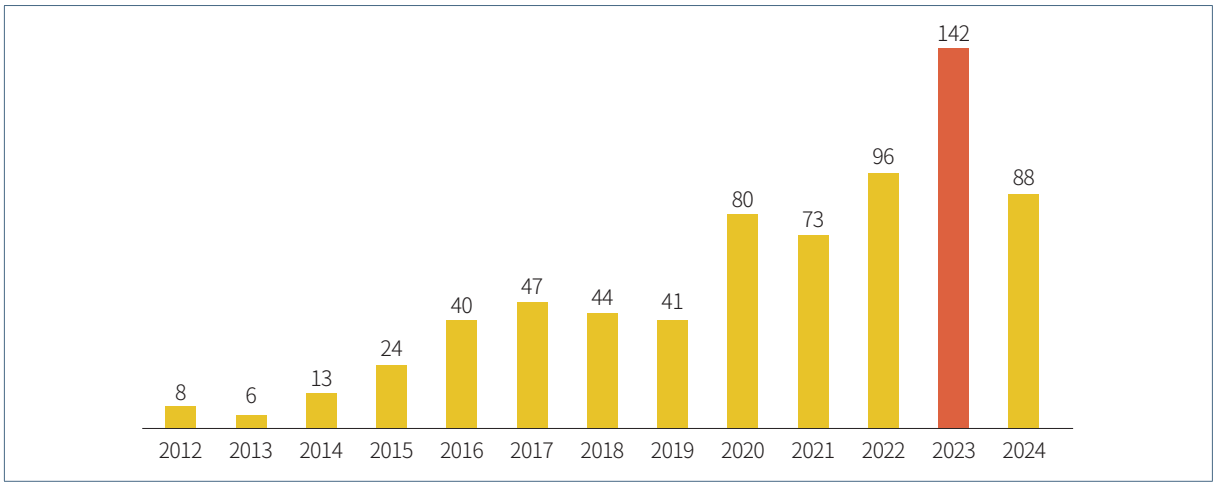
- 스탠퍼드大인간중심인공지능연구소(HAI)는 AI가 저숙련 노동자와 고숙련 노동자 사이의 성과 격차를 좁히는 등 산업의 효율화에 기여한다고 분석[2]

■ AI의 확산과 함께 AI로 인해 발생하는 사건 수 역시 지속 증가함에 따라 AI 개발사의 위험 대응 요구도 확대

- AI 사건 데이터베이스(AI Incident Database, AIID)*에 따르면, AI 사건 수는 [그림 1]과 같이 2023년에 전년 대비 96건에서 142건으로 48% 증가하는 추세를 보임[3]

  * AIID는 책임 있는 AI 협력체의 프로젝트로 AI 윤리 문제를 추적하는 공공 데이터 셋인 AIAAIC(AI, Algorithmic, and Automation Incidents and Controversies) 데이터베이스를 기반으로 하며, 상시로 사건 보고서를 수집하고 검토

[그림 1] 연도별 AI 사건 수 (~'24년 7월)



1 McKinsey (2023), The state of AI in 2023: Generative AI's breakout year.

2 Stanford HAI (2024), Artificial Intelligence Index Report 2024.

3 *Ibid*. AI Incident Database(2024.7.17. 기준), 연도별 수치는 SPRi 연구진 업데이트.

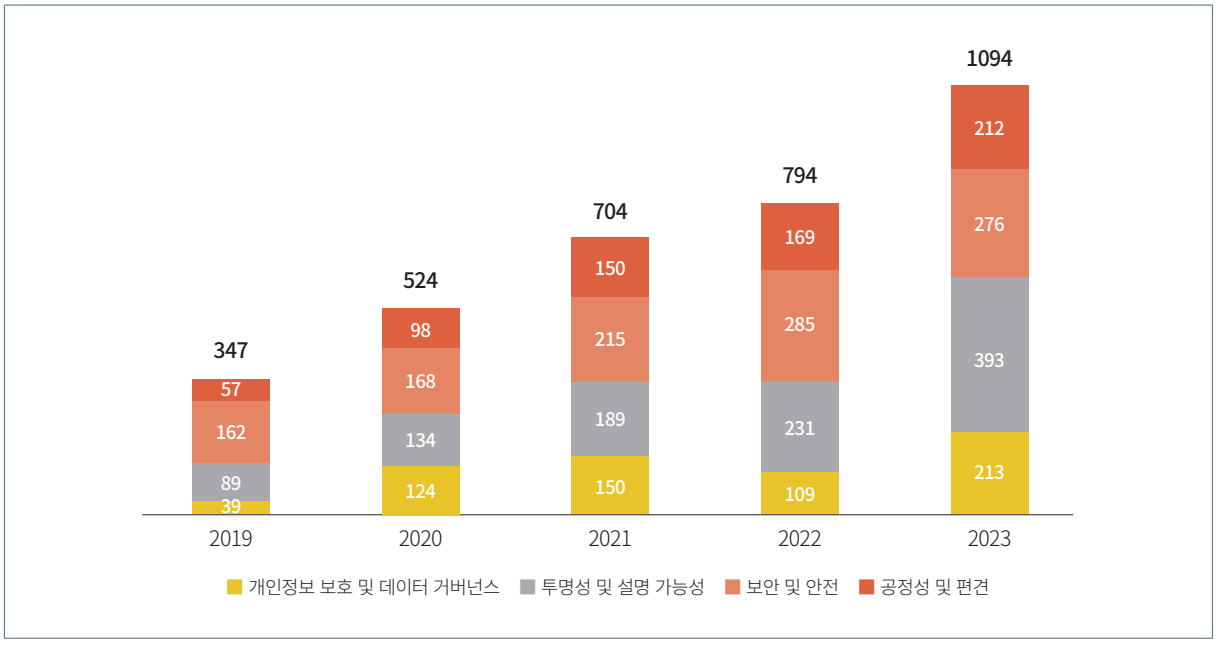
- 2024년 7월 기준 88건이 발생하였으며, 같은 추세로 보면 2024년 연말까지 총 150건 이상으로 2023년 대비 증가할 전망임
- 이러한 증가 추세는 다양한 산업 분야에서 AI가 활용됨에 따라 AI로 인한 사고로 인식하거나 AI가 윤리적으로 오용될 수 있는 사회적 인식이 증가한 영향을 반영한 것으로 해석할 수 있음

■ 정부와 학계는 AI의 위험을 방지하고 안전하고 신뢰할 수 있는 AI 개발을 위한 연구에 매진

- 각국은 정부 주도의 AI 안전연구소 설립을 통해 민관 협력을 통한 AI 위험 관리 프레임워크 및 검·인증 체계구축, 거버넌스 수립을 논의[4]
  - 2023년 영국, 2024년 한국에서 개최된 AI 안전 정상회의, 2023년 G7 정상회의 등 AI 안전에 관한 국제협력 및 논의 또한 활발

- 학계에서도 국가와 사회의 높은 관심을 반영하여, [그림 2]와 같이 책임 있는 AI 및 AI 안전과 관련한 기술적 연구 논문도 매년 증가하는 추세[5]

[그림 2] 주요 AI 국제학회의 책임 있는 AI 관련 논문 제출 건수



4 유재흥, 노재원, 장진철, 조지연(2024), “해외 AI안전연구소 추진현황 및 시사점”, SPRi 이슈리포트 IS-175 참고.

5 Stanford HAI (2024), Artificial Intelligence Index Report 2024 및 저자 재구성. AAAI, AIES, FAccT, ICML, ICLR, NeurIPS 6개 국제 AI 학술대회 대상 책임 있는 AI와 관련된 주요 키워드를 포함하는 논문 제출 건수 집계 결과.

■ 실제 AI 모델을 개발하는 업계에서도 책임 있는 AI를 위한 실질적인 움직임이 이뤄지는 중

- 기업의 AI 도입과 활용이 증가하면서 AI 개발 기업은 AI 위험을 해결해야 할 문제로 인식
  - Gartner(2023)[6] 가 글로벌 IT 경영진을 대상으로 한 조사 결과, 생성형 AI의 가장 우려되는 위험으로 개인정보 위험(42%), 환각(14%), 오용(12%), 보안(13%), 편견 및 불공정(10%), 불투명한 결과(7%)를 언급
  - 생성형 AI를 도입하여 조직이 이익을 얻기 위해 노력함과 동시에, 위험이란 해결 과제도 함께 거론되고 있음

- 최근 AI 개발 기업이 진행하는 책임 있는 AI 실현을 위한 전담 조직 신설 및 원칙·프레임워크 수립 등 여러 사례를 조사하여, 현재의 현황과 앞으로의 방향을 논의할 필요가 있음

2. 연구의 구성

■ 본 연구는 부상하고 있는 AI 안전 및 책임 있는 AI에 관해 국내외 기업이 실질적으로 진행하고 있는 현황을 통계 조사 결과와 기업 사례 연구를 통해 살펴보고자 함

- 첫째, 글로벌 컨설팅 기업인 액센츄어와 미국 스탠퍼드 대학교가 공동으로 시행한 책임 있는 AI 현황의 글로벌 통계 조사(Accenture & Stanford HAI (2024). Global State of Responsible AI)에 기반하여, 책임 있는 AI의 요인별 주요 결과를 분석함[7]

- 둘째, 국내외 주요 AI 기업의 책임 있는 AI 관련 전담 조직 운영 및 내부 원칙 수립 등 사례 연구를 통해 기업별 책임 있는 AI의 실현 현황을 살펴봄

- 마지막으로, 책임 있는 AI를 위한 민간 기업의 움직임에 따라 정부의 정책적 대응 방향에 대해 시사점을 도출하였음

6 Gartner Inc. (2023), “2024 Tech Provider Top Trends: AI Safety(ID G00804240)” (2023.12.12.)

7 액센츄어의 Global State of Responsible AI 2024 조사 결과가 스탠퍼드 대학교 HAI의 AI Index 2024에 게재되었으며, 액센츄어 보고서는 2024년 7월 현재 미공개. 이에, 본 보고서는 스탠퍼드 대학교 AI Index 2024의 결과를 인용하여 분석함.

II. 글로벌 기업의 책임 있는 AI 현황 조사의 주요 내용

1. 조사 개요

■ **액센츄어는 2024년 스탠퍼드 대학교와 공동으로 글로벌 기업의 책임 있는 AI에 대한 인식과 조치에 대한 현황 조사를 시행하여, 스탠퍼드 AI Index 리포트에 공개**

- (조사 목적) 책임 있는 AI 채택의 현재 수준을 이해하고, 각 산업과 지역에 따른 기업의 개발, 배포 및 사용에 따른 책임 있는 AI 활동 현황과 영향 비교

- (조사 범위) 2024년 2월부터 3월까지 22개 국가의 19개 산업 내 AI 개발 및 도입, 서비스 기업을 대상으로 응답을 수집하였으며, 연간 매출액이 5억 달러 미만인 기업 제외 후 1,000개 이상의 조직 대상 분석

2. 조사 항목

■ **책임 있는 AI와 관련하여 △개인정보 보호 및 데이터 거버넌스, △투명성 및 설명 가능성, △신뢰성 및 보안, △공정성을 측정[8]**

- 개인정보 보호 및 데이터 거버넌스는 데이터의 법률 및 규정 준수 여부, 데이터의 완전성, 고유성, 일관성, 정확성 및 오류가 있는 데이터의 업데이트 프로세스 유무를 측정

- 투명성 및 설명 가능성은 개발 프로세스 및 데이터 소스 등 주요 정보에 대한 문서화, 모델의 이용 사례와 한계를 다루는 교육 프로그램 유무 등을 측정

- 신뢰성은 레드팀 구성과 같은 모델 오류 및 취약성, 유해성에 대한 완화 조치 유무, 적대적 공격 방지 조치, 모델 신뢰도 테스트를 포함하며, 보안은 사이버 보안 위험에 대한 조치 프로세스 유무, 전담 인력 등을 측정

- 공정성은 예상 사용자의 인구통계에 기반하여 편향되지 않은 데이터를 수집하고 평가하는지, 모델 개발 및 검토 단계에서 다양한 이해관계자가 참여하여 기술적 편향이 완화되는지를 측정

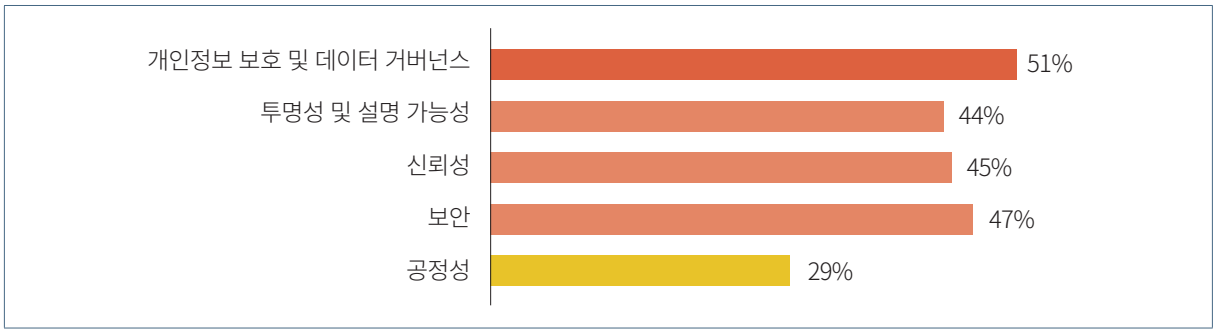
8 조사 측정 문항은 본 보고서의 [부록]에 번역하여 제시

- 각각의 지표별 적용하지 않음(Not applied), 임시 조치(Ad-hoc), 신규 도입(Rolling-out), 충분히 운영(Fully operationalized)의 4가지 수준으로 응답 수집

3. 조사 결과

■ 글로벌 기업은 개인정보 보호 및 데이터 거버넌스 관련 위험 조치를 가장 많이 고려(51%)하고 있으며, 이어 보안, 신뢰성, 투명성 및 설명 가능성, 공정성 순으로 고려

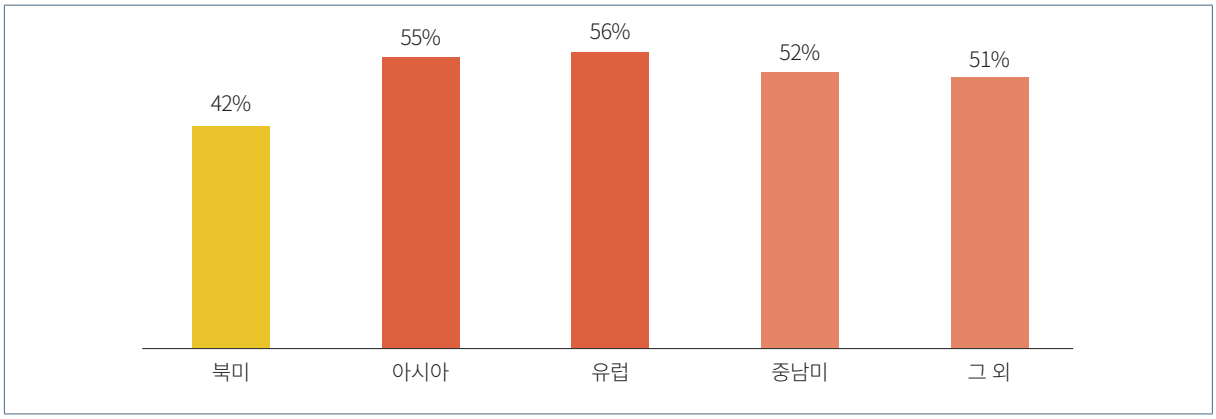
[그림 3] 글로벌 기업의 책임 있는 AI 현황 조사 결과



■ 조사에 응답한 기업의 51%가 자사의 AI 전략에 개인정보 보호 및 데이터 거버넌스 관련 위험을 고려하고 있다고 응답

- 지역별로는 유럽(56%)과 아시아(55%)에서 개인정보 보호 및 데이터 거버넌스 위험을 가장 많이 고려하고 있으며, 북미(42%)가 가장 낮음

[그림 4] 지역별 개인정보 보호 및 데이터 거버넌스 조치 비율



- 기업이 데이터 거버넌스 관련 위험에 대한 조치*를 취했는지에 대한 질문에, 90%의 기업이 적어도 한 가지 이상의 대응을 마련했다고 응답

  * 데이터 거버넌스 관련 법규 준수, 데이터 사용에 대한 동의 확보, 데이터 관련성 유지를 위한 정기 감사 및 업데이트 실시 등

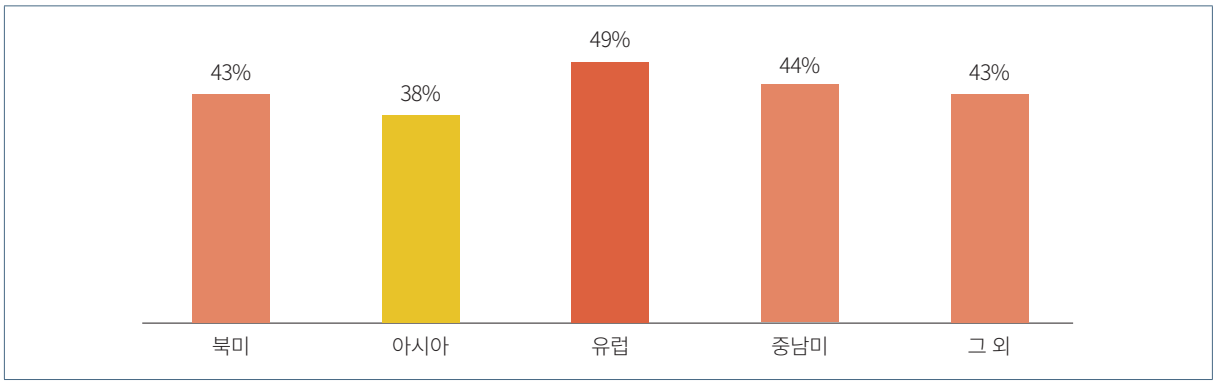
■ 조사에 응답한 기업의 44%가 자사의 AI 전략에 투명성 및 설명 가능성 관련 위험을 고려하고 있다고 응답

- 지역별로는 유럽(49%)이 투명성 및 설명 가능성에 대한 고려 비율이 높으나, 상대적으로 아시아(38%)에서는 관련 고려 비율이 낮음

- 88%의 기업이 적어도 한 가지 이상의 투명성 및 설명 가능성 관련 위험을 완화하는 조치*를 마련

  * 투명한 개발 프로세스 문서화, 모델의 의도된 사용 사례와 한계를 다루는 이해관계자를 위한 교육 프로그램 유무, 일부 성능이 희생되더라도 모델의 설명 가능성에 우선순위를 두는지 유무, 모델 설명 도구를 사용하여 모델 결정을 명료하게 하는지 등

[그림 5] 지역별 투명성 및 설명 가능성 조치 비율



■ 조사에 응답한 기업의 45%가 자사의 AI 전략에 신뢰성 관련 위험을 고려하고 있다고 응답하였으며, 보안에 대해서는 47%가 고려

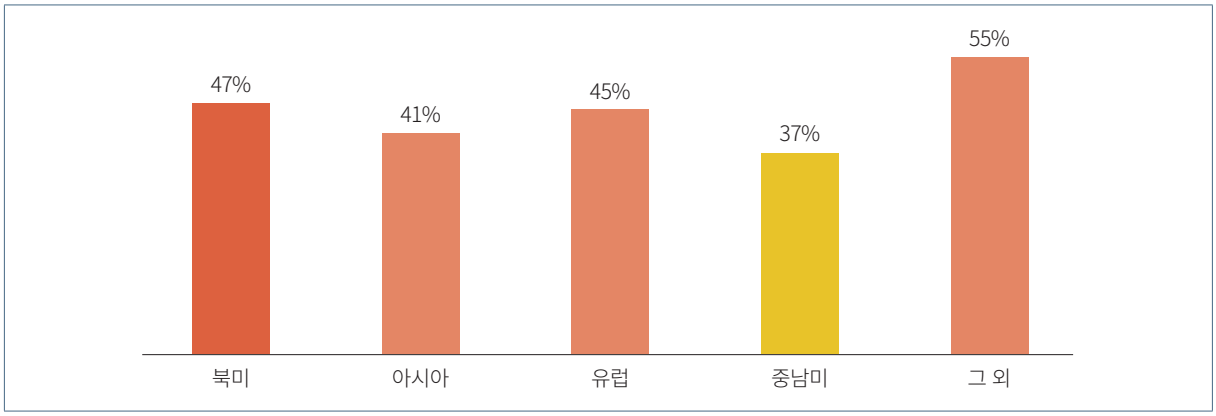
- 지역별로는 북미(47%)에서 신뢰성을, 아시아(51%)에서 보안에 대한 고려를 적극적으로 하고 있으며, 반면 중남미(37%), 유럽(42%)에서는 각각 신뢰성과 보안 고려 비율이 낮음

- 응답 조직의 75%가 적어도 하나 이상의 신뢰성 관련 대응책*을 마련하였으며, 63%는 하나 이상의 보안 관련 대응책**을 마련

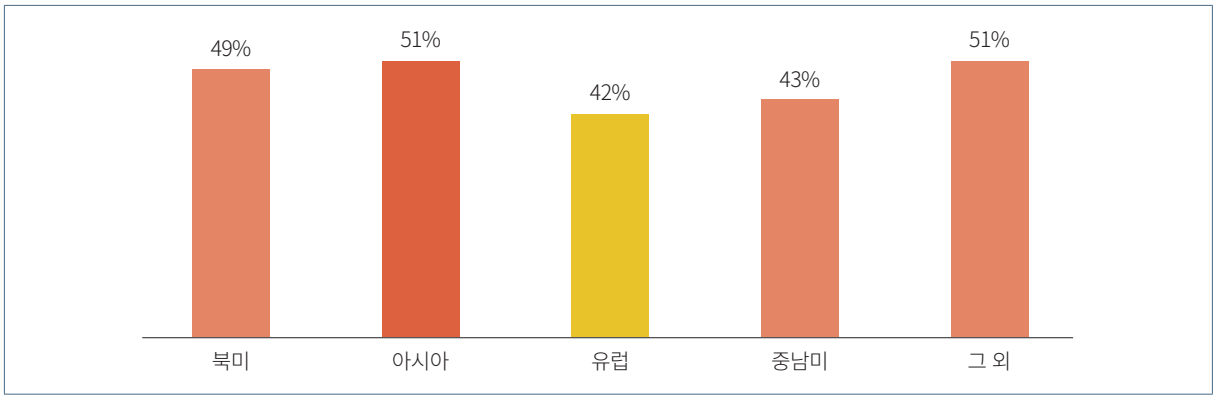
  * AI 모델 오류에 대한 완화 조치, AI 시스템/모델의 가용성을 보장하기 위한 장애 조치 계획, 취약성 또는 유해한 행동(예: 레드팀 구성)에 대한 모델/시스템 평가, 적대적 공격을 방지하는 조치, 모델 출력에 대한 신뢰도 점수 제공, 광범위한 시나리오와 지표를 포괄하는 포괄적인 테스트 사례 마련 등

  ** 사이버 보안 관행(예: 다단계 인증, 액세스 제어 및 직원 교육) 유무, 공급망 내 제3자의 사이버 보안 조치 및 검증 유무, 전담 AI 사이버 보안 조직 혹은 교육을 받은 직원 유무, AI 관련 사이버 보안 점검 및 조치(예: 적대적 테스트, 취약성 평가, 데이터 보안 조치), 진화하는 AI 관련 사이버 보안 위험에 대한 조치 등

[그림 6] 지역별 신뢰성 조치 비율



[그림 7] 지역별 보안 조치 비율



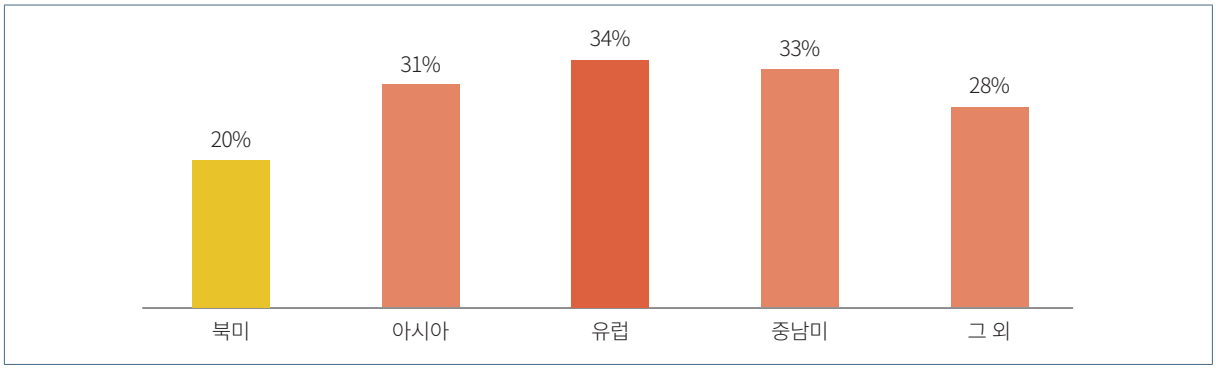
- 또한, 조사 대상의 대다수인 88%는 기반 모델 개발자가 관련 위험을 완화할 책임이 있다는 데 동의하고 있으며, 86%는 생성형 AI로 인한 잠재적 위협이 충분히 크므로 전 세계적으로 합의된 거버넌스의 필요성에 동의

■ 조사에 응답한 기업의 29%가 자사의 AI 전략에 공정성 관련 위험을 고려하고 있다고 응답

- 지역별로는 유럽(34%)에서 가장 많이 공정성 위험을 고려하는 조치*를 취했다고 응답하지만, 북미(20%)에서 가장 낮게 응답

  * 예상되는 사용자의 인구통계를 기반으로 편향 없는 데이터 수집 여부, 독립적인 감독을 위해 제3자(감사자/일반 대중)가 접근 가능한 방법론과 데이터 출처의 존재 여부, 모델 개발 및/또는 검토 프로세스에 다양한 이해관계자의 참여, 다양한 인구통계학적 그룹에 대한 성과 평가, 모델 개발 중 기술적 편향 완화 기술 사용 등

[그림 8] 지역별 공정성 조치 비율



III. 국내·외 주요 기업의 책임 있는 AI 현황

■ 본 장은 주요 AI 개발사의 책임 있는 AI를 위한 구체적인 현황을 사례 기반으로 조사

- 책임 있는 AI를 위한 △전담 조직의 유무와 권한, △기업 내 책임 있는 AI 관련 원칙 및 프레임워크 유무 관점으로 국내외 주요 기업 대상 사례 조사를 실시

- 국외 기업은 책임 있는 AI 조직 및 AI 안전 프레임워크를 공개한 주요 기업을 대상으로 조사하였으며, 국내 기업은 최근 ESG 보고서 및 대외적인 사업 보고서를 통해 사회를 위한 책임 있는 AI 현황을 소개한 기업을 중심으로 관련 사례를 정리

1. 오픈AI

■ 오픈AI는 안전하고 책임 있는 AI 개발을 위하여 AI 모델 수준에 따른 조직을 구성하고 관련 AI 모델 평가 및 연구를 수행

- [표 1]과 같이, 모델 수준을 현재의 모델, 최신의 고성능 프런티어 모델, 미래에 개발될 가능성이 있는 초지능 모델로 구분하여, 각각 안전시스템 팀, 대비 팀, 초정렬 팀*으로 조직을 구분

  * 초정렬 팀(Superalignment team)은 최근 조직의 해체를 발표('24.5)

[표 1] 오픈AI 안전 조직의 안전 연구 수행 범위*

현재의 모델		프런티어 모델		초지능 모델
안전시스템(Safety Systems) 팀	▶	대비(Preparedness) 팀	▶	초정렬(Superalignment) 팀

* 세 개의 팀이 서로 다른 개발 시점(Time frames)과 위험 요인을 담당하며, 표의 오른쪽으로 갈수록 고도화된 모델

- (안전시스템 팀) AI 모델의 안전성, 견고성, 신뢰성을 확보하고 실제 환경에서의 AI 모델의 배포와 관련한 위험을 다루며, 안전하게 AI 모델을 배포하기 위한 연구 수행
  - 챗GPT와 같은 현재 수준의 AI 모델 및 제품 수준에서의 위험과 오남용 완화 조치를 구현하는 데 초점을 두고 있으며, 엔지니어링, 정책, 인간과 AI의 협업 등 다양한 분야 전문가로 구성된 하위 4개 팀[9]으로 구성
- (대비 팀) 프런티어 AI(Frontier AI)와 같은 점점 더 강력해지는 최첨단 AI의 잠재적인 위험을 예측·완화하는 대비 프레임워크(Preparedness Framework)와 프로토콜 개발 연구를 수행
  - 개발 중인 프런티어 모델을 지속적으로 평가하며 새롭게 등장하는 위험을 모니터링하는 부서로서 모델을 안전하게 만드는 것에 초점을 둔 조직으로 대비 프레임워크를 전담
- (초정렬 팀) 먼 미래에 도입될 초지능 모델의 안전 기반 구축을 위하여 초지능 시스템이 인류에게 유익한 방식으로 작동하도록 보장하는 방법에 관한 연구를 수행[10]
  - 미래의 초고도 인공지능 시스템이 야기할 수 있는 위험에 대응하고 안전을 보장하기 위하여 AI를 제어할 수 있는 방법론을 연구하는 안전팀

9 안전 엔지니어링(Safety Engineering), 모델 안전 연구(Model Safety Research), 안전 추론 연구(Safety Reasoning Research), 인간과 AI의 상호작용(Human-AI Interaction)

10 오픈AI 공동 창립자인 일리야 수츠케버(Ilya Sutskever)와 초정렬 연구를 이끌던 얀 라이케(Jan Leike) 등 안전팀 구성의 핵심 인사들의 사임 후, 초정렬 팀의 해체 발표

■ 프런티어 AI 모델의 안전한 개발·배포를 위한 접근 방식을 담은 대비(Preparedness) 프레임워크 보고서를 공개 ('23.12)

- 보고서는 AI 모델이 초래할 수 있는 위험을 체계적으로 추적, 평가, 예측, 완화하기 위한 프로세스를 설명하고 있으며, △위험 수준 추적, △알려지지 않은 위험 탐색, △안전기준 설정, △대비팀 구성, △안전 자문 그룹(Safety advisory group)의 5가지 핵심 요소로 구성

- 대비 프레임워크의 주요 내용은 AI 모델의 기능, 취약성과 같은 잠재적 위험을 추적하고, 스코어카드를 통해 AI 위험의 수준을 평가하며, 4가지 위험 범주와 위험 수준을 제시
  - 위험은 사이버 보안(Cybersecurity), 설득(Persuasion), 화학·생물학·방사능·핵(CBRN) 위협, 모델 자율화(Model Autonomy) 등 4가지[11]로 범주화하고, 위험 등급은 '낮음', '중간', '높음', '심각'으로 분류
  - 스코어카드를 기반으로 AI 모델의 위험 등급을 평가하고 완화 전*·후** 모델 위험을 추적할 수 있도록 구성

    * 완화 이전 위험 평가(Pre-mitigation risk) : 평가 모델이 특정 도메인에서 최악의 경우를 가정하여 평가를 수행하며, 기본 모델과 모델 훈련 전후를 지속 평가

    ** 완화 이후 위험 평가(Post-Mitigation Risk) : 완화 조치 이후의 위험 감소 검증, 시스템에 대한 '최악의 경우' 시나리오를 평가하고, '중간' 위험 이하로 유지하는 것을 목표
  - AI 모델의 평가 및 완화 조치 후, 위험성 점수가 '중간(Medium)' 이하인 모델만 배포할 수 있으며, 위험성 점수가 '높음(High)' 이하일 경우에만 모델의 추가 개발이 가능

- 오픈AI는 GPT-4o 등 새로운 AI 모델과 제품을 발표할 때도 대비 프레임워크를 기반으로 한 평가 검증 과정을 거쳤음을 공식 발표[12]

2. 마이크로소프트

■ 마이크로소프트(MS)는 책임 있는 AI 투명성 보고서[13]를 발표하며 생성 AI 개발 및 활용 과정의 투명성 강화 및 책임 있는 AI 실현에 대한 MS의 실천 사항을 공개

11 △사이버 보안 : 모델이 사이버 공격 및 취약성 발견에 어떻게 활용될 수 있는지 평가, △화학, 생물, 방사선 및 핵 위협(CBRN): 모델이 이러한 위협과 관련된 정보를 생성하거나 확산하는 데 사용될 가능성에 대한 평가, △설득 : 모델이 인간의 행동을 변경하거나 조작하는 데 얼마나 효과적인지 평가, △모델 자율성: 모델이 자율적으로 작동하거나 복제할 수 있는 능력 평가

12 https://openai.com/index/hello-gpt-4o/

13 Microsoft (2024), Responsible AI Transparency Report.

- 2023년 7월 백악관의 주요 인공지능 기업들의 자발적 AI 약속[14]의 일환으로 발표한 보고서로 투명성 실천을 위한 MS의 개발, 배포 및 고객 지원에 대한 원칙을 설명

- MS의 생성 애플리케이션은 책임 있는 AI, 보안, 개인정보 보호 및 데이터 보호 정책을 포함한 회사 정책을 준수하며, 규제 및 내·외부 이해관계자의 피드백을 바탕으로 관련 정책을 수립
  - 개발 주기의 전 과정에 걸쳐 생성 AI의 위험을 매핑, 특정 및 관리하도록 가이드라인과 준수 사항을 설계
  - 위험 매핑(Map) 단계는 AI 시스템과 관련된 잠재적 위험을 식별하고 관리하기 위한 첫 단계로, 생성 AI 시스템의 계획, 보호 조치 및 적절성에 관한 결정
  - 위험 측정(Measure) 단계는 식별된 위험을 정량화하고 영향을 평가하는 단계로, 생성 AI 시스템을 개발하고 사용할 때 AI 위험 및 관련 영향을 측정*하는 절차를 제공
    * 생성형 AI 애플리케이션에 대한 식별된 위험을 측정하기 위한 메트릭을 설정(Metrics for identified risks)하고, 완화가 확인된 위험을 해결하는 데 얼마나 효과적인지 측정(Mitigations performance testing)
  - 위험 관리(Manage) 단계는 측정된 위험에 대한 대응 조치를 실행하고, 지속적인 성능 모니터링, 피드백, 사고 대응 프로세스의 반복을 통해 개선

- MS는 책임 있는 AI를 위하여 관리, 매핑, 측정 및 관리를 위한 프로세스를 반복하여 진행하며, 제품 개발 및 배포 주기 전반에 걸쳐 AI 관련 위험을 관리

■ 책임 있는 AI에 대한 관리는 CEO에서부터 고위 리더십 팀과 회사 전체에 걸쳐 진행*

- 책임감 있는 AI 문화를 촉진하기 위하여 상향식 및 하향식 결합 접근 방식을 통하여 개인, 팀 및 조직의 역량 강화를 도모
  * △마이크로소프트 이사회(Microsoft Board)는 책임 있는 AI 정책 및 프로그램에 대한 감독, 지침 제공, △책임 있는 AI 위원회(Responsible AI Council)는 AI와 관련한 도전과제에 대처하고 책임 있는 AI 정책과 프로세스 진전을 위한 회의.

3. 구글

■ 구글은 2023년 기존 구글 리서치 팀을 딥마인드 부서에 통합하여 전담 연구 조직을 구성

- 책임 있는 AI 팀 또한 딥마인드로 이동되어, 안전하고 윤리적인 AI 배포와 AI 테스트 및 평가를 위한 업무를 수행

14 Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI.

- 제품 개발 및 AI 애플리케이션의 평가를 위해 인공지능 원칙[15]('18.6)*을 도입하였고, 안전과 책임에 대한 접근 방식[16]을 게시('23.10)
  * 책임 있는 AI 기술 개발을 위해 사회적 유익성, 공정성, 안전성 등 7가지 AI 활용의 목적에 따라 AI를 개발할 것을 정의하고, 활용하지 않을 분야를 명시

■ 구글은 책임 있는 AI 시스템 개발 및 배포를 목표로 총 9개 영역에서의 접근 방식을 공개

- (책임 있는 역량 확장, Responsible Capabilities Scaling) AI 원칙 및 위험 식별·평가·관리를 위한 전 주기 프로세스를 구축하고, 프런티어 모델의 위험관리 내용을 포함
  - '책임 있는 AI 위원회'와 '책임 있는 개발과 혁신팀'을 구성하고 윤리 및 안전성 관련 위험을 식별하며, 프런티어 모델에 대한 안전성 평가를 수행
  - '책임성과 안전 위원회(Responsibility and Safety Council, RSC)'를 설립하여 원칙 준수를 지원하고, 기술 및 정책적 위험 완화를 수행

- (모델 평가 및 레드팀) AI 모델의 평가를 위해 벤치마킹, 사용자 평가, 영향평가 등 다양한 형태의 평가를 지원하며, 제품 조사를 위해 전담 레드팀을 구성

- (모델 보고 및 정보 공유) 자체 연구를 기반으로 한 책임 있는 AI 관행 프레임워크를 개발하고 구글 클라우드(Vertex AI)를 통해 기반 모델 등에 관한 포괄적인 정보를 제공

- (취약점 보고 및 모니터링) 모델 배포 이후 발견되는 취약점에 대해 외부 사용자의 취약점 감지 및 보고 권한 부여, 내·외부 적대적 테스트와 잠재 위협을 감지하고 분류하는 전담팀을 운영

- (보안 제어) 모델 가중치 유출 방지 등 고급 AI 모델과 시스템의 보안을 유지하기 위해 취약점을 관리하고 다양한 보안 모범 사례를 도입
  - 공공 및 민간 부문의 AI 시스템 보안에 대한 개념적 접근 방식인 보안 AI 프레임워크(SAIF)*를 적용
    * 구글의 SAIF(Secure AI Framework)는 AI 생태계에 보안 기반 확장, 탐지·대응AI 확장, 방어 자동화로 새로운 위협에 대응, 플랫폼 수준에서의 보호 조치, AI 배포를 위한 피드백 루프 생성, AI 시스템 위험의 맥락화 등의 내용을 포함

- (AI 생성물 식별) AI가 생성한 콘텐츠를 안전하고 책임감 있게 식별하고, 워터마킹, 메타데이터, 디지털 서명 등 출처 추적 방안을 마련

15 https://ai.google/responsibility/principles/
16 https://deepmind.google/public-policy/ai-summit-policies/

- 워터마킹 및 AI 생성 이미지 식별 도구인 신스ID(SynthID) 베타 버전을 출시('23.8)하고, 이미지에 대한 정보 도구를 통한 출처 추적을 지원

- (선제적 연구 및 투자) AI 윤리, 안전 및 거버넌스 연구 등 사회 안전 및 보안 위협에 관한 선제적 연구와 투자 진행
  - 구글 딥마인드는 AI 시스템의 견고성 및 검증 연구, 사회적 영향 연구 및 생성 모델 평가 등 책임 있는 AI 개발 지원을 위한 연구를 수행 중

- (데이터 입력 제어 및 감사) AI 개발 전주기에서 데이터 사용 원칙을 지키고, 데이터 사용 요청 및 검토 절차와 데이터의 수정 및 관리 프로세스를 수립

- 이외에도 AI의 사회적 도전과제를 해결한 사례 등을 발표하며, AI가 여러 분야에서 긍정적 영향을 미칠 수 있음을 강조

■ 기존 RCS 정책에 기반하여, 안전한 프런티어 AI 개발을 위한 안전 프레임워크를 발표

- 프런티어 안전 프레임워크는 △치명적 역량 수준(Critical Capability Level) 식별, △프런티어 모델 평가, △완화 조치 적용의 세 가지 주요 내용으로 구성
  - 자율성, 생물보안, 사이버 보안, 기계학습 R&D 영역은 예비 분석 결과에 따라 위험영역으로 구분되어 CCL 식별 수행
  - 프런티어 모델 평가*에서는 모델 역량이 CCL에 도달하기 전 안전 버퍼(Safety buffer)를 가질 수 있도록 조기 경고 평가 방식
    * 기반 모델의 성능 측정치인 효과적 연산 능력(Effective compute)이 6배 커질 때마다, 파인튜닝을 하는 3개월마다 재평가
  - 보안 완화 조치와 배포 완화 조치는 각각 4단계로 구성되어 각 수준 및 역량에 해당하는 대응 방안을 정의

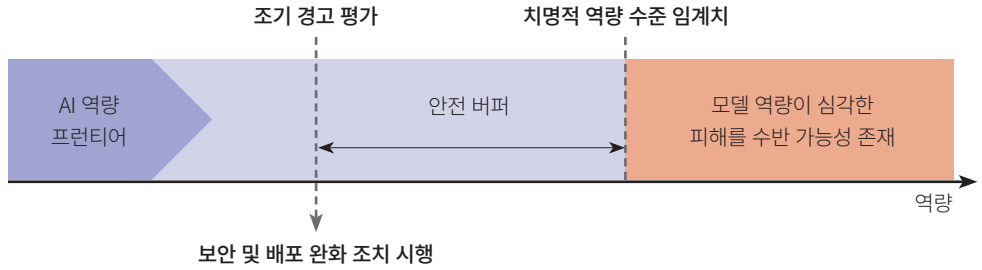
[표 2] 구글 프런티어 안전 프레임워크 세부 내용

구분	세부 기능
치명적 역량 수준 식별	• 프런티어 모델이 고위험 도메인(자율성, 생물보안, 사이버 보안, 기계학습 R&D)에서 피해를 발생시킬 수 있는 경로를 분석 • 특정 모델이 피해를 발생시킬 수 있는 치명적 역량 수준(CCL)을 식별하고 임계치를 설정
프런티어 모델 평가	• 프런티어 모델의 주기적 평가를 통한 CCL의 임계치 도달 시점 감지 • 조기 경고 평가(Early warning evaluation)를 개발하여, 임계치 도달 전 반복 평가 수행

구분	세부 기능
완화 계획 적용	• 조기 경고 평가 통과 시(평가 임계치 도달) 완화 조치를 시행 • 4단계로 구성된 보안 완화 조치*(모델 가중치 유출 방지) 및 배포 완화 조치**(중요한 기능에 대한 접근 관리)를 포함 * 보안 완화 조치: 현상 유지(0단계), 제어된 접근(1단계) 일방적 접근 방지(2단계) 고신뢰 개발 환경(3단계) ** 배포 완화 조치: 현상 유지(0단계), 치명적 역량을 목표로 한 완화 조치(1단계), 레드팀 검증을 통한 안전 사례(2단계), 접근 방지(3단계)

출처: Google Deepmind(2024), Introducing the Frontier Safety Framework. SPRi 재구성.

[그림 9] 구글 AI 안전 프레임워크 개념도



출처: Google Deepmind(2024), Introducing the Frontier Safety Framework. SPRi 재구성.

4. 앤트로픽

■ 앤트로픽은 기업의 최우선 가치로 AI 안전을 내세우며 AI 안전 연구 역량 확보에 투자

- CEO인 다리오 아모데이(Dario Amodei)는 오픈AI의 연구 부사장 출신으로 오픈AI의 상업화에 이견을 가지고 오픈AI의 전 정책 책임자였던 잭 클라크(Jack Clark) 등 동료들과 창업
  - 최근에는 오픈AI에서 초정렬(Superalignment) 연구를 이끌던 얀 라이크(Jan Leike)가 회사를 비판*하며 사임하고 앤트로픽에 합류('24.5)
    * "AI 안전과 보안의 다양한 측면, 특히 '확장가능한 감독(Scalable oversight)', '약한 일반화(Weak-to-strong generalization)', 자동화된 정렬(Automated alignment) 연구에 집중할 것(얀 라이크, 2024.5.29., X 포스팅)"[17]

17 현재 앤트로픽의 최고과학책임자 자레드 캐플란(Jared Kaplan)은 확장할 수 있는 감독(대규모 AI의 행동을 예측할 수 있고 바람직한 방식으로 제어하는 기술)을 연구 중

■ 앤트로픽은 회사 설립 시부터 AI 모델 개발의 기본적 접근 방식으로 헌법적 AI를 제안[18]

- 헌법적 AI(Constitutional AI)는 AI 안전성 기술 중 하나로, AI 모델이 사전에 정의된 일련의 윤리적 원칙과 헌법적 규범에 따라 행동하도록 하는 것을 목표로 함

- 헌법적 AI는 지도 학습(Supervised learning) 단계와 강화 학습(Reinforcement learning) 단계로 구성되며 응답에 대한 리뷰와 AI 피드백은 '헌법'이라 불리는 원칙에 의해 조정
  - 지도 학습 단계는 초기 모델을 크게 개선하며, 강화 학습 단계 시작 시 초기 행동에 대한 일부 제어를 제공하여 잠재적인 탐색 문제를 해결
  - 강화 학습 단계는 성능과 신뢰성을 크게 향상하는 것으로 알려짐

■ 앤트로픽은 AI 안전 관리 프레임워크인 책임 있는 확장 정책(RSP) 프레임워크를 운영[19]

- ASL(AI Safety Levels)를 5단계로 구분하고 각 안전성 수준마다 데이터 유해성 평가, 모델 카드 공개, 취약점 보고, 접근 제한, 배포 중단 등 조치를 시행
  - 클로드 3(Claude 3)를 배포하며 모델 안전 수준 등을 평가한 결과도 공개('24.3)*
    * 클로드3는 ASL-2 정도로 재앙적 위험(Catastrophic risk) 가능성이 작음

[표 3] 앤트로픽의 AI 위험 수준 정의(ASL)

단계	정의
ASL-1	• 유의미한 치명적인 위험을 초래하지 않는 시스템 (예: 2018 LLM 또는 체스만 두는 AI 시스템)
ASL-2	• 생물학적 무기를 제조하는 방법에 대한 지침을 제공할 수 있는 능력 등 위험한 능력의 초기 징후를 보이지만 신뢰성이 부족하거나 검색 엔진이 제공하지 못하는 정보를 제공하지 않아 아직 유용하지 않은 시스템 (예: 클로드(Claude)를 포함한 현재의 LLM 수준)
ASL-3	• AI가 아닌 기준선(예: 검색 엔진 또는 교과서)에 비해 치명적인 오용의 위험이 크게 증가하거나 낮은 수준의 자율 기능을 보이는 시스템
ASL-4 ASL-5+	※ ASL-4 이상(ASL-5+)은 현재 시스템과 너무 거리가 멀기 때문에 아직 정의되지 않았으나, 치명적인 오용 가능성과 자율성에서 질적 상승을 수반할 가능성이 높음

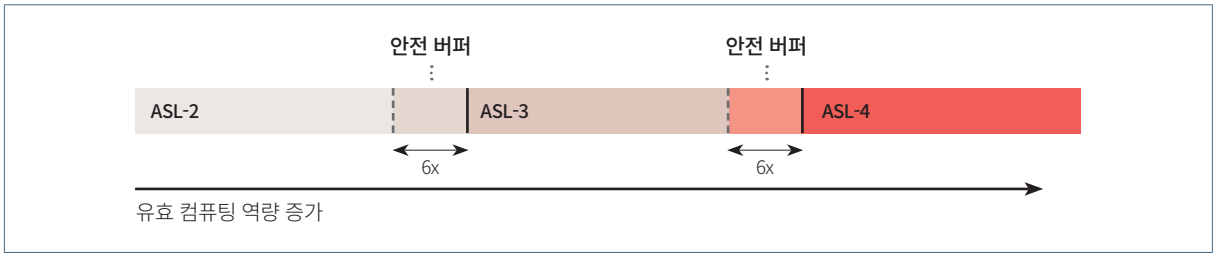
- 수준별 AI 역량의 임곗값을 설정해 위험을 관리하며 단계마다 안전 버퍼(Safety buffer)를 두어 임계치에 도달하기 전 위험성을 평가하고 제어

18 Bai et al.,(2022), Constitutional AI: Harmlessness from AI Feedback

19 Anthropic (2023), Anthropic's Responsible Scaling Policy

- 기본 3개월마다 정기적으로 평가하고, 유효 컴퓨팅[20]이 4배 증가할 때 또는 미세조정이 있을 시 평가하고, 평가 결과에 따라서 배포 중지 등 보안 조치 수행

[그림 10] 앤트로픽의 AI 안전 수준과 안전 버퍼



출처: Anthropic (2023), Anthropic's Responsible Scaling Policy version 1.0

- [그림 10]과 같이, 안전 버퍼[21] 크기는 유효 컴퓨팅 역량이 6배 크게 되었을 때로 설정함으로써 평가가 진행되는 동안에도 모델 훈련이 안전하게 계속될 수 있도록 했으며, 필요한 안전 조치가 마련되지 않는 한 해당 모델의 훈련과 배포는 중단

5. 메타

■ 메타는 기술적인 시스템 수준의 안전장치를 마련하여 책임 있는 AI 모델 개발 모색

- 라마 가드(Llama Guard)[22]는 프롬프트 및 응답 결과가 안전한지를 파악하고, 안전하지 않았을 때는 위반된 콘텐츠를 나열하여 악성 결과물을 방지
  - 라마 가드에서 분류하는 위험 범주로는 폭력적인 범죄, 비폭력 범죄(사이버 범죄, 사기, 무기 범죄 등), 성범죄, 아동 성 착취, 명예훼손, 증오, 자살 및 자해 등 14가지 범주를 포함

- 프롬프트 가드(Prompt Guard)[23]는 LLM 동작을 파괴하는 프롬프트 공격으로부터 보호
  - 프롬프트 공격의 범주로는 AI 모델이 의도하지 않은 명령을 실행하도록 명령하는 프롬프트 주입(Prompt Injections), 모델에 내장된 안전 및 보안 기능을 무시하도록 설계된 악성 명령인 탈옥(Jailbreaks)을 포함

20 유효 컴퓨팅은 사전 훈련 또는 미세 조정 기술의 개선을 포함하지 않은 상태에서 모델을 훈련하는 데 필요한 컴퓨팅 자원 규모로 정의
21 Anthropic (2023), Anthropic's Responsible Scaling Policy version 1.0
22 https://github.com/meta-llama/PurpleLlama/blob/main/Llama-Guard3/MODEL_CARD.md
23 https://github.com/meta-llama/PurpleLlama/blob/main/Prompt-Guard/MODEL_CARD.md

- 사이버 보안 위험 평가도구(Cybersec Eval)를 제공하여, 안전하지 않은 코드의 위험을 완화하기 위한 벤치마크 모음을 제공

■ AI 기능 및 모델 관련 정보의 공개 및 레드팀 운영을 통한 투명성 개선

- 설명 가능성과 투명성을 높이는 방안으로 인공지능(AI) 시스템의 작동 방식을 더 잘 이해하는 데 도움이 되도록 시스템 카드(System card)[24]를 공개
  - 시스템 카드는 각 AI 시스템 및 기능별 작동 원리, 사용 팁, 데이터 사용량, 이용 시 주의할 사항을 명시하여 AI 시스템이 의사결정을 내리는 과정을 투명하게 공개

- 내·외부 전문가로 구성된 레드팀 활동을 통해 AI 모델의 취약점을 파악하고 수정하는 동시에 예상치 못한 사용 방식을 찾아 개선

■ 별도의 전담 조직 없이 제품 개발 부서에서 책임 있는 AI를 담당

- 2023년 메타 내 팀 재분배의 영향으로, 책임 있는 AI 팀을 해산('23.11)하고 생성형 AI 제품 개발 부서에서 책임 있는 AI 개발 및 사용과 관련된 업무를 지원[25]

6. 네이버

■ 네이버는 안전하고 신뢰성 있는 AI 기술 연구개발 및 전략·정책 논의의 전문성 강화를 위해 CEO 직속 조직으로 퓨쳐 AI 센터를 설립

- 퓨쳐 AI 센터(Future AI Center)는 AI 안전성을 연구하고 네이버의 AI 윤리·안전 정책 수립 및 총괄 임무를 수행하며, 데이터셋 구축 및 소스 코드 공개, 국내외 안전성 연구 협력을 수행

- 2024년 2월에 네이버의 AI 안전 관련 프레임워크 및 자사 로드맵 작성·홍보, 국내외 AI 안전 정책 동향 분석 등을 목적으로 AI 안전 분야 경력직 연구원 채용을 진행

24 https://ai.meta.com/tools/system-cards/
25 https://www.cnbc.com/2023/11/18/facebook-parent-meta-breaks-up-its-responsible-ai-team.html

■ AI 기술을 누구나 쉽고 편리하게 활용할 수 있는 일상의 도구로 만들고자 하는 방향성을 담은 네이버 AI 윤리 준칙을 2021년 공개

- 네이버 AI 윤리 준칙은 네이버 구성원의 AI 개발과 이용에 있어 인간 중심의 가치를 우선적으로 부여하며, AI 윤리 자문 프로세스를 통해 AI 윤리 논의를 진행하도록 원칙을 제정

- 사람을 위한 AI 개발, 다양성 존중, 합리적인 설명과 편리성, 안전을 고려한 서비스 설계, 프라이버시 보호와 정보 보안 등의 원칙을 수립

- 신규 개발 중인 생성형 AI 서비스에 AI 윤리 준칙을 반영하며, 네이버의 생성형 AI '하이퍼클로바 X' 공개와 더불어 '사람을 위한 클로바 X 활용 가이드'를 공개

■ AI 시스템과 관련한 위험을 예방하기 위해 2024년 6월 AI 안전 프레임워크(AI Safety Framework, 이하 ASF) 초안을 발표

- ASF는 네이버 AI 윤리 준칙을 준수하는 네이버 구성원이 산업 현장에서 AI 시스템을 개발하고 배포하는 과정에서 AI 안전을 구체적으로 실천하기 위한 체계

- 네이버는 악의적인 이용자 공격에 대응하기 위한 자체적인 레드팀(모의공격)을 운영하고 있으며, 혐오, 폭력, 고정관념 등 AI 안전 평가 기준을 확립

- ASF는 AI의 통제력 상실 위험에 대해서는 AI 위험 평가 스케일을 구축하여 대응하고, 악용 위험에 대해서는 AI 위험 평가 매트릭스를 통해 위험을 관리하는 방안으로 구성
  - AI 위험 평가 스케일은 현재 존재 혹은 개발되는 최고 성능의 AI 시스템인 프런티어 AI를 3개월 주기의 정기적 평가와 시스템의 능력이 기존보다 6배 증가했다고 판단되는 경우 별도 평가를 수행하는 절차로써, 앤트로픽이나 구글과 유사한 기준을 수립
  - [표 4]와 같이 AI 위험 평가 매트릭스는 AI 시스템의 목적 영역과 안전 조치의 필요성을 기준으로 AI 시스템의 전체 라이프사이클에 따른 위험이 발생할 수 있는지를 인식하고 평가 및 관리

[표 4] 네이버의 AI 위험 평가 매트릭스

		안전 조치의 필요성	
		낮음	높음
목적 영역	일반	**AI 시스템 위험 낮음** AI 시스템을 배포하고, 배포 후 안전성 모니터링을 통해 시스템 위험을 관리	**AI 시스템 위험 있음** 추가적인 안전 조치를 시행해 AI 시스템 위험을 완화할 때까지 AI 시스템을 배포하지 않음
	특수	**AI 시스템 위험 있음** 특별한 자격이 있는 사용자에게 AI 시스템을 제공하여 AI 시스템 위험을 완화	**AI 시스템 위험 높음** AI 시스템을 배포하지 않음

* 출처: 네이버(2024), Naver Integrated Report 2024.

- 위험이 있는 경우, 기술적·정책적 조치 등 위험 완화를 위한 안전 조치를 시행하여 AI 시스템 위험이 충분히 완화되었다고 판단되는 경우에만 AI 시스템을 배포
- 일반 영역의 AI 시스템에서는 특수 영역에서 활용되는 능력이 사용되지 않도록 안전 조치를 취함

■ 네이버 AI 윤리 준칙을 기반으로, AI 서비스 위험 관리를 위한 네이버 윤리 자문 프로세스(CHEC)[26] 운영

[그림 11] 네이버의 AI 윤리·안전성 개선 프로세스

* 출처: 네이버(2024), Naver Integrated Report 2024.

- 네이버 AI 서비스에 산업적인 시각뿐만 아니라 인간 중심의 가치를 부여하고, 사회적인 관점을 더해 '사람을 위한 AI'라는 가치를 부여

26 CHEC: Consultation on Human-Centered AI's Ethical Considerations

- 현실적 개선 사례 기반 원칙을 구체화하여 네이버 AI 윤리 준칙을 자연스럽게 준수하는 기업 문화로 내재하는 방향을 목표로 하여, 관련 사내 교육도 활발히 진행

7. LG AI 연구원

■ LG AI 연구원은 LG그룹의 AI 싱크탱크로써 산하에 AI 윤리위원회, AI 윤리사무국, LG AI 윤리 워킹그룹 등을 운영

- AI 윤리 전문가로 구성된 AI 윤리사무국은 AI 연구개발 및 이용 단계에서 발생할 수 있는 윤리적 문제를 사전에 점검함으로써 연구원 내외의 실제 업무에 AI 윤리 원칙을 적용하고 있음

- LG그룹 계열사의 주요 AI 윤리 이슈를 논의하는 협의체 조직으로써 LG AI 윤리 워킹그룹을 운영

[그림 12] LG AI 윤리 원칙

* 출처: LG AI 연구원(2024), 2023 LG AI 윤리 책무성 보고서.

■ AI를 개발하고 활용하는 LG그룹 전체 구성원이 지켜야 할 기준인 AI 윤리 원칙과 AI 위험을 사전에 파악하는 위험 관리 프로세스를 수립하여 운영

- AI 윤리 원칙은 국제기구 및 정부, 기업에서 발표한 규범을 토대로 인간 존중, 공정성, 안전성, 책임성, 투명성 5가지 핵심 가치로 구성

- AI 위험 관리 프로세스는 AI 과제의 특성을 분석하여 잠재적인 위험성을 미리 파악하며, 이후 문제 우선순위를 설정하여 순차적으로 해결하며, 결과에 대한 문서화를 통해 AI 시스템의 투명성과 책임성을 확보

[표 5] LG AI 연구원의 AI 위험 문제 해결 우선순위 구분

		잠재적 위험성	
		낮음	높음
해결 난이도	어려움	**4순위** 잠재적 위험성이 낮고 해결하기 어려운 문제	**1순위** 잠재적 위험성이 높고 해결하기 어려운 문제
	쉬움	**3순위** 잠재적 위험성이 낮고 해결하기 쉬운 문제	**2순위** 잠재적 위험성이 높고 해결하기 쉬운 문제

* 출처: LG AI 연구원(2024), 2023 LG AI 윤리 책무성 보고서.

- 계열사의 고객 상담 효율화를 위한 AI 기술 개발 과제에 위험관리 프로세스가 적용된 바 있음
  - (① 과제 특성 분석) 고객 상담 관련 학습용 데이터 내 민감 개인정보 및 고객의 신체나 정서적 위험을 초래하는 고위험 정보, 혐오 문구 유무 분석
  - (② 문제 해결 우선순위 분석) 연구자가 학습 데이터 내 욕설과 혐오 표현 등 위험 요인을 제거하고 데이터 수집 시 자동 필터링 기술을 개발하는 요구사항 도출
  - (③ 이행 결과 확인 및 문서화) 식별된 잠재 위험과 개선 결과, 데이터 및 모델의 특성, 위험과 한계 등을 포함한 AI 위험 관리 프로세스의 과정 및 결과를 종합적으로 문서화

8. 그 외 기업

■ 카카오는 계열사의 최고기술책임자(CTO)가 참여하는 카카오 공동체 기술윤리 위원회를 2022년 7월부터 운영

- 카카오뱅크, 카카오페이 등 주요 계열사가 관련 핵심 가치를 공유하고 이를 서비스에 반영할 수 있도록 월 1회 정기 회의를 개최하며, 여기에는 각 계열사의 최고 기술 책임자(CTO, Chief Technology Officer)가 대표하여 참석

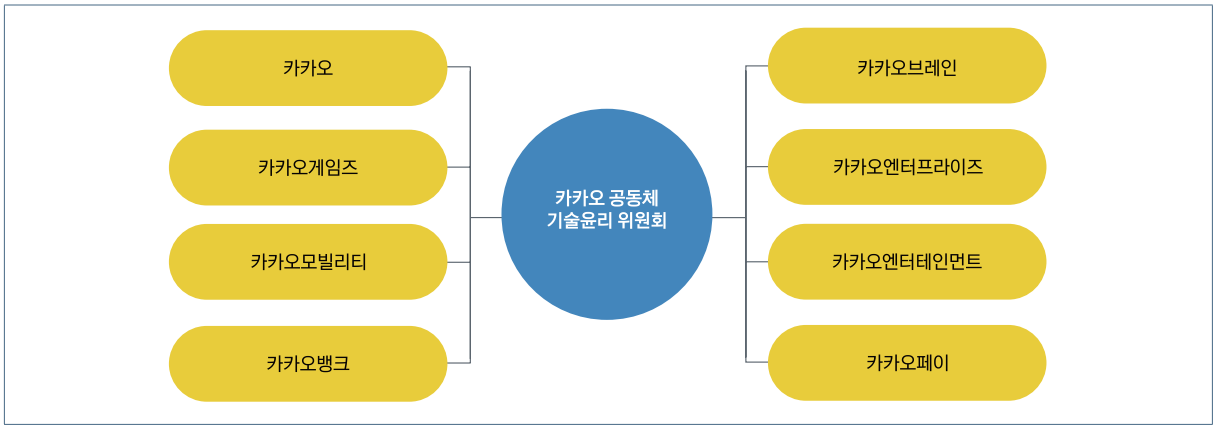
- 이는 카카오 계열사 서비스 및 기술의 안전성을 기술윤리 관점에서 검토하여, 알고리즘 윤리 헌장 준수 여부와 위험성 점검, 알고리즘 투명성 강화 등을 위한 체계적 개선 업무 수행

- 기존의 기술윤리 관련 정책 점검과 함께 책임 있는 AI를 위한 거버넌스 구축 및 가이드라인 제정을 완료하고 현재 내재화 및 고도화 진행 중

- 글로벌 오픈소스 커뮤니티인 'AI 얼라이언스(AI Alliance)'에 가입함으로써 투명하고 개방된 AI 생태계 확장을 위한 책임감 있는 AI 이니셔티브 추진('24.4)

- 2023년 3월 카카오 공동체 기술윤리 위원회에 개선된 AI 윤리 원칙인 '카카오 공동체의 책임 있는 AI를 위한 가이드라인' 제안 및 확정

[그림 13] 카카오 공동체 기술윤리 거버넌스 구조



* 출처: 카카오(2023), 2023 카카오 공동체 기술윤리 보고서.

■ 2024년 4월 KT 사내에 책임 있는 인공지능 센터를 설립하여 AI 윤리, 정책 개발·협력, 소프트웨어 엔지니어링 등 다양한 분야의 인력을 채용

- RAIC(Responsible AI Center)의 주요 기능은 △ 책임 있는 AI 정책수립 및 대외 정책 지원, △ AI가 악용될 수 있는 분야(안전성, 투명성, 프라이버시 등)에서 위험 수준에 대한 관리체계 구축, △ AI 안전 정책 연구 및 글로벌 기관과 협력 등으로 구성

- 2023년 AI 윤리 준칙을 수립함으로써 포용성, 투명성, 신뢰성, 지속가능성 등 일상의 혁신을 위한 윤리적이고 책임 있는 AI 기술을 실현

- 주요 내용은 AI 기술과 제품 개발에서의 공정성, 비차별, 투명성, 해석 가능성의 원칙 준수, 표준화와 개방형 생태계를 통한 자원과 기술 공유 강화, 책임감 있는 AI 개발을 위한 사회적 책임 강화 등을 담고 있음
- 마이크로소프트와의 전략적 파트너십과 AI·클라우드·IT의 혁신과 성장을 주도하고 공공과 금융 분야의 보안성을 강화한 소버린 클라우드와 AI를 개발

■ SK텔레콤은 2024년 3월 AI 거버넌스 원칙인 'T.H.E AI'를 공개

- 'T.H.E. AI'는 통신 기술 기반(by Telco)의 사람을 향한, 사람을 위한(for Humanity), 윤리적 가치 중심(with Ethics) 및 이에 따르는 AI 거버넌스 원칙을 상징
  - 이를 통해 고객 신뢰와 안전을 기반으로 잠재적 AI 위험으로부터 고객을 보호하며, 다양성과 평등, 공정의 가치를 지향하고 인류의 복지를 증진하면서, AI 의사결정의 투명성과 개인정보 보호, 윤리적 책임을 약속

- AI 거버넌스 관리체계를 원활하게 운영하기 위해 대외협력 담당(CGO)이 총괄하는 AI 거버넌스 전담 조직을 신설, AI 거버넌스 경영 전반에 걸쳐 지원 조직, 전문 조직 등 조직 간 시너지 창출 기반 마련

IV. 해외와 국내 기업의 책임 있는 AI 대응 비교

■ 해외와 국내 AI 기업의 AI 안전 전담 조직 및 프레임워크 특징은 다음과 같음

[표 6] 해외 및 국내 기업의 책임 있는 AI 대응 방향 비교

<table>
<tr><th></th><th>해외 기업</th><th>국내 기업</th></tr>
<tr><td rowspan="2">전담 조직 및 거버넌스</td><td colspan="2">AI 위험 식별·평가·조치 의사결정을 위한 전담 조직 운영</td></tr>
<tr><td>(오픈AI) AI 모델의 수준별 조직 운영
(MS) 이사회·책임 있는 AI 위원회 운영
(구글) 딥마인드에 AI 안전 조직 운영
(앤트로픽) 오픈AI의 초정렬 연구자 합류
(메타) 개발 부서에서 AI 안전 담당</td><td>그룹 계열사 간 협의체 형태를 포함하며, 이를 통한 도메인별 AI 위험에 대한 공통점과 차이점을 공유</td></tr>
<tr><td>AI 위험 식별 및 평가</td><td>악의적 사용(Malicious use)을 중심으로 한 위험 요인 정의 및 평가</td><td>개발 목적 및 과제 특성에 따른 위험성과 안전 조치의 필요성, 난이도 평가</td></tr>
<tr><td>안전 버퍼의 기준</td><td colspan="2">(구글, 앤트로픽, 네이버) 연산 능력이 6배 커질 때마다, 3개월마다 정기적 평가</td></tr>
<tr><td>책임 있는 AI 요인별 비교</td><td colspan="2">신뢰성, 보안, 투명성, 데이터 거버넌스 및 개인정보 보호 중심의 조치로써 향후 공정성 요인에 대한 조치 필요
(메타) 투명성을 강화하기 위해 모델 정보를 담은 시스템 카드(System card) 제공</td></tr>
</table>

- 국내와 해외 기업 모두 AI 위험 식별·평가·조치 의사결정을 위한 전담 조직을 운영하고 있으며, 국내 기업은 계열사 간 책임 있는 AI 관련 협의체 조직을 운영

- 해외 기업은 범용 AI의 위험 유형[27] 분류에 따른 오용 및 악의적 사용, 즉 가짜 콘텐츠 생성이나 여론 조작, 사이버 공격, 생화학적·방사능(CBRN) 무기 개발 등을 방지하는 AI 안전 식별, 평가 및 조치에 중점

- 국내 기업은 산업별 AI의 응용을 고려하여 AI 개발 목적 및 과제 특성에 따른 AI 모델 평가 기준을 운용
  * 네이버의 프런티어 AI 모델 안전 평가 기준인 3개월 주기의 정기 평가, 시스템 능력이 6배 증가했을 때의 별도 평가는 앤트로픽 및 구글과 동일

- 액센추어와 스탠퍼드 HAI의 글로벌 조사(Accenture & Stanford HAI, 2024) 및 기업별 책임 있는 AI 현황에 따르면 AI 요인 중 신뢰성, 보안, 투명성, 정보보호 측면의 고려는 이뤄지고 있으나, AI 윤리 및 공정성 측면의 노력이 필요

■ 즉, 각 기업의 전담 조직과 AI 안전 프레임워크는 위험 평가 및 완화 조치에 대한 권한과 역할을 명시하고 있으며, 기업의 비즈니스 상황에 맞는 차별성이 있음

- 기업 대부분은 전담 조직의 주도로 AI 위험 수준에 따라 모델의 위험 수준을 주기적으로 평가하는 체계를 갖추고 있으며, 평가 결과에 따라 모델의 개발·배포 관리를 통한 위험 완화 방안을 제시하고 있음
  - 전담 조직의 장은 고위험 AI 모델의 개발 및 배포를 중지하거나 수정할 권한을 가지는 형태로 대부분의 의사결정이 가능한 C-레벨 급 위상을 보유

- AI의 안전 및 위험은 물론 윤리적 가치에 기반하여 이용자 중심의 공정성, 투명성, 설명 가능성 측면도 함께 고려하고 있는 특징이 있음

- 국내 기업은 공정성과 투명성 보장, 보안을 위한 기존의 AI 윤리 원칙 수립에 이어, 최근 해외 추세에 맞게 AI 안전 위협으로부터의 보호 방안을 마련하기 위한 조직의 역할이 확대되고 있음

27 Yoshua Bengio 외 (2024.5.17.), International Scientific Report on the Safety of Advanced AI: INTERIM REPORT

V. 시사점

■ 액센츄어와 스탠퍼드 HAI의 글로벌 조사 결과에 따르면, 공정성을 비롯하여 투명성 및 설명 가능성에 대한 기업의 인식과 조치가 향상될 필요

- 유럽(95%), 북미(96%), 아시아(92%)의 기업은 자사 AI 도입 전략에 책임 있는 AI의 요인, 즉 공정성, 투명성 및 설명 가능성, 개인정보 및 데이터 거버넌스, 신뢰성 및 보안 등을 하나 이상 고려하고 있음
  - 동의 없는 데이터 사용이나 데이터 유출과 같은 개인정보 보호 및 데이터 거버넌스의 조치 비율이 높으나, 투명성과 공정성 관련 조치 비율은 낮음
  - 아시아에서는 보안 조치 비율이 높았으나 투명성은 낮은 비율을 보이며, 유럽은 투명성, 공정성, 개인정보 보호 및 데이터 거버넌스 조치는 높으나 보안은 상대적으로 낮은 조치를 보이면서 지역별 편차를 나타냄

- AI 모델이 고도화되며 복잡한 모델은 우수한 성능을 제공할 수 있으나 단순한 모델보다 해석하기 어려운 경향이 있어 모델의 복잡성과 설명 가능성 사이의 적절한 절충이 필요함

- 현재 공정성은 일관되고 구조화된 접근 방식이 타 지표 대비 부족하며, 공정성을 정의, 측정, 보장하는 일은 복잡하기에 이에 대한 합의가 향후 필요함

■ 기업 내 전담 조직 신설 및 원칙 수립과 함께, 전담 조직의 권한과 역할 및 목표를 분명히 하여 책임 있는 AI 프로세스를 정착시키고 실행하는 노력 중요

- 기업의 AI 윤리 및 공정성 실현 시 주요 장애요인으로 '혁신 제품 개발 목표가 AI 윤리 의지보다 우선시되는 기업 문화', 'AI 윤리가 기업의 성과 지표로 정량화하기 어려운 점', '기업 내 전담 조직의 권한 부족과 잦은 조직개편'이 주로 언급됨 (Ali et al, 2023)[28]

- 이에, 기업 구성원에 대한 AI 안전 및 윤리 교육, 기업 경영진의 인식 개선이 선행됨으로써 기업 내 책임 있는 AI 문화가 내재될 필요가 있음

28 Ali, S. J. et al., "Walking the Walk of AI Ethics: Organizational Challenges and the Individualization of Risk among Ethics Entrepreneurs", ACM FAccT '23.

■ 정부의 AI 안전 관련 제도 마련 및 AI 안전연구소 설립과 더불어 기업의 자체적인 노력으로 민관 협력을 통한 책임 있는 AI 정착을 위한 문화 확산이 필요

- 최근 유럽연합의 인공지능법, 미국 백악관의 행정명령을 비롯한 위험성이 높은 AI에 대한 각국의 규제 조치가 진행되는 만큼, 기업과 국가의 고위험 AI의 안전 조치를 확보하는 프로세스가 조속히 정착되어야 함

- 현재 대기업은 자체적인 전담 조직을 구성하여 책임 있는 AI 프레임워크 수립 및 실행을 통한 자율 규제에 노력하고 있으며, 향후 국가 차원에서는 중소·스타트업 및 수요기업 등 전체 AI 산업 가치사슬로의 책임 있는 AI 확산이 필요함
  - 특히, 제도와 함께 기술적 대응이 본격적으로 이뤄지는 해외 기업 대비, 국내 기업은 AI 윤리 가이드라인을 제시하는 수준인 상황으로, 향후 국내 기업도 기술적·제도적 조치 모두 병행하여 추진할 필요가 있음

- 본 보고서의 결과에 따르면 주요 기업별 AI 안전 프레임워크에 차이가 있으며, 이에 국가 및 기업 단위의 글로벌 협력을 통해 AI 윤리 및 안전에 대한 프레임워크의 표준화가 필요함
  - 현재 한국에서는 개인정보보호위원회의 '인공지능(AI) 개발·서비스를 위한 공개된 개인정보 처리 안내서', 한국정보통신기술협회(TTA)의 '신뢰할 수 있는 인공지능 개발 안내서' 등 관련한 가이드라인이 논의되고 있음

- 향후 책임있는 AI 실현을 위해 국내 AI 개발기업, 도입 기업 현황, 조직 내 AI 안전 제도 마련, 실행 여부 등에 대한 지속적인 실태 파악이 필요함

부록 글로벌 기업 책임 있는 AI 현황 조사 항목 (Accenture & Stanford HAI, 2024)

[개인정보 보호 및 데이터 거버넌스]

- 데이터가 모든 관련 법률 및 규정을 준수하고 해당되는 경우 동의를 받아 사용되는지 확인
- 데이터 수집 및 준비에는 데이터의 완전성, 고유성, 일관성 및 정확성에 대한 평가 포함 여부
- 최종 모델/시스템이 사용되는 인구통계학적 설정과 관련하여 데이터가 대표적인지 확인.
- 데이터의 관련성을 보장하기 위한 정기적인 데이터 감사 및 업데이트 여부
- AI 수명주기 전반에 걸쳐 데이터 세트 문서화 및 추적성을 위한 프로세스 유무
- 단점이 있는 데이터 세트에 대한 수정 계획 및 문서화 유무

[투명성 및 설명 가능성]

- 개발 프로세스에 대한 문서화, 세부적인 알고리즘 설계 선택, 데이터 소스, 의도된 사용 사례 및 제한 사항
- 모델의 의도된 사용 사례와 한계를 다루는 이해관계자(사용자 포함)를 위한 교육 프로그램
- 일부 성능이 희생되더라도 높은 해석성이 중요한 단순한 모델에 우선순위
- AI 모델을 설명하는 도구를 사용하여 모델 결정을 명료하게 하는지 유무

[신뢰성]

- 모델 오류에 대한 완화 조치 및 낮은 신뢰도 결과물에 대한 처리 유무
- 시스템/모델의 가용성을 보장하기 위한 장애 조치 계획 또는 기타 조치 유무
- 취약성 또는 유해한 행동(예: 레드팀 구성)에 대한 모델/시스템 평가 유무
- 적대적 공격을 방지하기 위한 조치
- 모델 출력에 대한 신뢰도 점수
- 광범위한 시나리오와 지표를 포괄하는 포괄적인 테스트 사례 유무

[보안]

- 기본적인 사이버 보안 관행(예: 다단계 인증, 액세스 제어 및 직원 교육) 여부
- 공급망 내 제3자의 사이버 보안 조치 조사 및 검증 여부
- 전담 AI 사이버 보안 팀 및/또는 AI 관련 사이버 보안에 대해 명시적으로 교육을 받은 직원 유무
- 기술적 AI 관련 사이버 보안 점검 및 조치(예: 적대적 테스트, 취약성 평가, 데이터 보안 조치) 유무
- 진화하는 AI 관련 사이버 보안 위험 및 기존 사이버 보안 프로세스에 통합 유무

[공정성]

- 예상되는 사용자 인구통계를 기반으로 한 대표 데이터 수집 여부
- 독립적인 감독을 위해 제3자(감사자/일반 대중)가 접근 가능한 방법론과 데이터 소스를 만들었는지 여부
- 모델 개발 및/또는 검토 프로세스에 다양한 이해관계자의 참여 여부
- 다양한 인구통계학적 그룹에 대한 성과 평가 여부
- 모델 개발 중 기술적 편향 완화 기술 사용 여부

⊙ 참고문헌

1. 국외문헌

Sanna J. Ali et al. (2023). “Walking the walk of AI ethics: Organizational challenges and the individualization of risk among ethics entrepreneurs.” In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 217-226).

Accenture & Stanford HAI (2024), Global State of Responsible AI. (미공개)

Anthropic (2023), Anthropic’s Responsible Scaling Policy Version 1.0.

Bai et al.,(2022), Constitutional AI: Harmlessness from AI Feedback.

Gartner Inc. (2023), “2024 Tech Provider Top Trends: AI Safety”.

McKinsey (2023), The state of AI in 2023: Generative AI’s breakout year.

Microsoft (2024), Responsible AI Transparency Report.

Nestor Maslej et al. (2024). “The AI Index 2024 Annual Report,” AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2024.

Yoshua Bengio et al.(2024), International Scientific Report on the Safety of Advanced AI: INTERIM REPORT

2. 국내문헌

네이버(2024), NAVER INTEGRATED REPORT 2023.

유재흥, 노재원, 장진철, 조지연 (2024), “해외 AI안전연구소 추진현황 및 시사점”, SPRi 이슈리포트 IS-175.

카카오(2023), 2023 카카오 공동체 기술윤리 보고서.

KT(2024), KT ESG REPORT 2024.

LG AI 연구원(2024), 2023 LG AI 윤리 책무성 보고서.

SK텔레콤(2024), SK TELECOM Annual Report 2023: Road to Global AI Company.