기술주권

안전·신뢰 AI

한국기술교육대학교 구본진
(前 KISTEP 전략기술정책단)

국가전략기술 기술주권 브리프:
안전·신뢰 AI

구본진

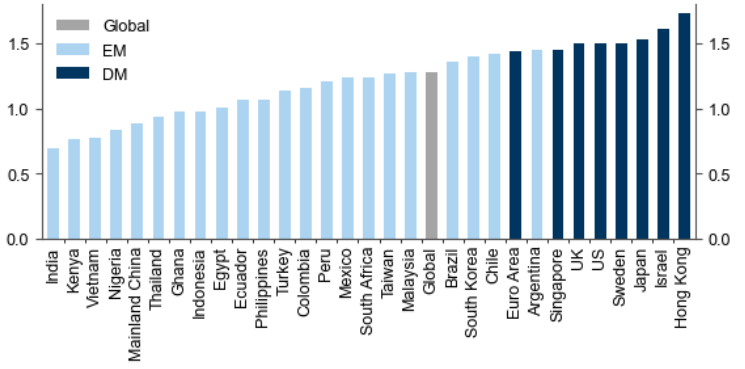
KISTEP 한국과학기술기획평가원
Korea Institute of S&T Evaluation and Planning

Ⅰ 작성배경

AI는 다양한 산업과의 접목으로 경제 성장에 기여하고 있고, 시장 및 투자 규모도 폭발적으로 증가

- AI 조기 채택 기업은 연간 2~3pp의 노동생산성 증가를 보이고*, 국가의 AI 채택은 향후 10년간 전 세계 연간 생산성 성장을 1.4pp 증가시킬 수 있을 것으로 추정(연간 글로벌 GDP 7% 증가에 기여)

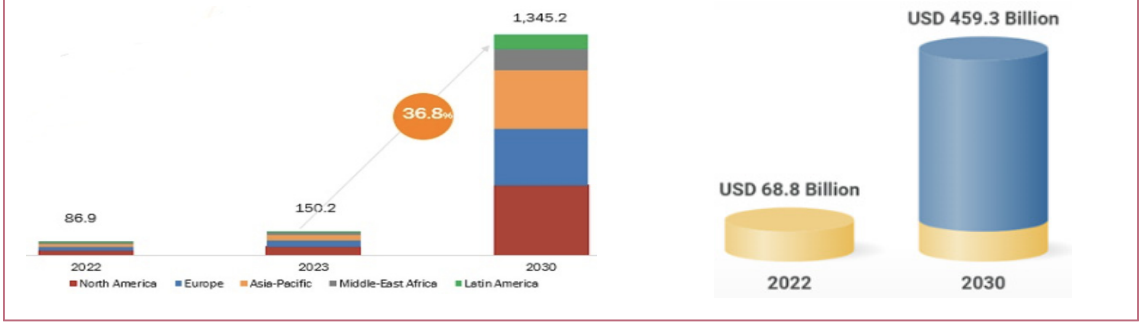
  * (연간 노동생산성 증가율, 단위 pp) Alederucci et al., (2022): 6.8, Czarnitzki et al. (2023): 2.7, Behrens & Trunschke (2020): 2.6, Acemoglu et al., (2022): 1.9, Bessen & Righi (2019): 1.7



[그림 1] AI 채택을 통한 국가 및 글로벌 연간 생산성 성장 추정치

※ 자료출처: Goldman Sachs Global Investment Research (2023)

- 글로벌 AI 시장 규모는 '30년까지 1조 3,452억 달러로 확대될 것으로 전망되며 全 분야의 주요 기업들이 AI를 중심으로 전면적인 체질 전환을 추진 중



[그림 2] 글로벌 AI 시장 전망

※ 자료출처: Markets and Markets (2022), Research and Markets (2023)

AI 기술이 핵심 전략기술로 부상하면서 주요국 간 AI 기술패권 경쟁 격화

- 美-中 두 축을 중심으로 AI 생태계 블록화가 진행 중이며 양국 간 경쟁 강도는 지속적으로 강화되고 있는 상황
  - 미국은 군사적 사용 가능성을 근거로 AI, 반도체, 양자컴퓨터 3개 분야 중국 기업에 대한 투자를 금지하였고, 최근 저사양 AI 첨단 반도체에 대한 중국 수출 금지 조항을 추가(수출 제재 우회 통제 포함)
  - 중국은 美 Micron 반도체의 중국 내 판매금지와 갈륨, 희토류 수출을 통제, AI 기술을 수출 제한 목록에 추가하고 반외국제재법을 시행

- 미국과 중국 정부는 우월한 기술력을 갖춘 AI 기업 확보를 국가 이익의 최우선으로 인식하여 관련 빅테크 기업 보호·육성 정책을 적극 추진
  - 미국은 중국 AI 기업 견제를 위하여 상원·하원의 빅테크 반독점 주요 법안들*을 대부분 폐기하였고, 정책 방향을 자국 빅테크 규제에서 육성·보호로 선회
    * Ending Platform Monopolies Act, American Innovation and Choice Online Act, Open App Markets Act 등
  - 중국은 마윈 사태 이후 규제 중심의 자국 빅테크 기업 정책 방향이 기술 경쟁력 확보를 위한 자국 빅테크 지원·육성으로 변경
    ※ ('23.7.) 중국 Li Quang 총리는 국가 경제에 있어 중국 플랫폼 기업의 중요성 강조, 중국 최고 경제 규제 당국은 플랫폼 기업에 대한 지원을 약속하는 등 규제 완화 기조 본격화

AI 기술의 확산 및 응용 분야의 빠른 확장으로 예상치 못한 AI 역기능 문제도 급속히 증가

- 현재 빠르게 확산 중인 주력 AI 시스템은 높은 비선형성과 블랙박스적 특성이 있고, 이로 인해 불확실성을 내포
  - AI 시스템의 비선형성 및 블랙박스적 특성은 복잡한 패턴과 상호작용을 효과적으로 학습할 수 있고, 시스템의 복잡성을 감소시킬 수 있는 장점이 있으나 모델의 내부 동작을 해석하고 이해하기 어렵게 만들어 신뢰성, 설명 가능성, 투명성을 낮추는 단점도 보유

- AI의 윤리적 오용 관련 사건을 추적하는 AIAAIC DB에 따르면 AI 사건과 논란 건수는 '12년 이후 26배 증가
  ※ ('12) 10건 → ('21) 260건 (출처: Stanford Univ. HAI AI Index Report, 2023)

AI 역기능에 대한 우려로 윤리적 AI 기술에 대한 기술적·사회적 요구가 증대되고 있고, 한국을 포함한 주요국들은 AI 안전성·신뢰성 확보를 위한 다양한 정책을 추진

- 현재까지는 AI 기술발전을 통한 시스템 성능 향상에 집중하였다면 이제는 AI 윤리에 대한 기술적 요구가 높아질 것으로 예상

  ※ Gartner는 responsible AI 기술이 5~10년 내 주류가 되어 기술의 정점에 진입할 것으로 예상

Benefit ↓	Years to Mainstream Adoption Less Than 2 Years ↓	2 - 5 Years ↓	5 - 10 Years ↓	More Than 10 Years ↓
Transformational	Computer Vision	Composite AI Decision Intelligence Deep Learning Edge AI Generative AI Intelligent Applications Physics-Informed AI	Foundation Models Natural Language Processing Neuromorphic Computing Responsible AI	Artificial General Intelligence Autonomous Vehicles
High		AI Cloud Services AI Maker and Teaching Kits AI TRiSM Data-Centric AI Data Labeling and Annotation Digital Ethics Synthetic Data	AI Engineering Causal AI Knowledge Graphs ModelOps Operational AI Systems Smart Robots	
Moderate				
Low				

[그림 3] AI에 대한 우선순위 매트릭스

※ 자료출처: Gartner Hype Cycle for AI (2022)

- 주요국 정부는 안전·신뢰 AI 생태계 조성을 위한 다양한 정책을 추진 중

  ※ AI 윤리기준 및 가이드라인 수립, 안전성 강화를 위한 기술개발 장려, 규제 및 표준화 추진, 다자간 협력 강화 등

본 고에서는 안전·신뢰 AI 경쟁에 효과적으로 대응할 수 있도록 관련 글로벌 동향 및 주요 이슈를 분석하고, 향후 기술주권 확보를 위한 정책 수립 방향을 제언

- 한국을 포함한 주요국의 안전·신뢰 AI 기술·산업·정책 동향 분석을 통한 주요 이슈 및 시사점을 도출
- AI 주요 학회 프로시딩(proceeding) 논문 분석을 통해 우리의 경쟁력을 진단
- 급변하는 안전·신뢰 AI 시장에서의 기술주권 확보를 위한 정책과 전략을 제시

Ⅱ 글로벌 기술/산업/정책 동향

1. 기술동향

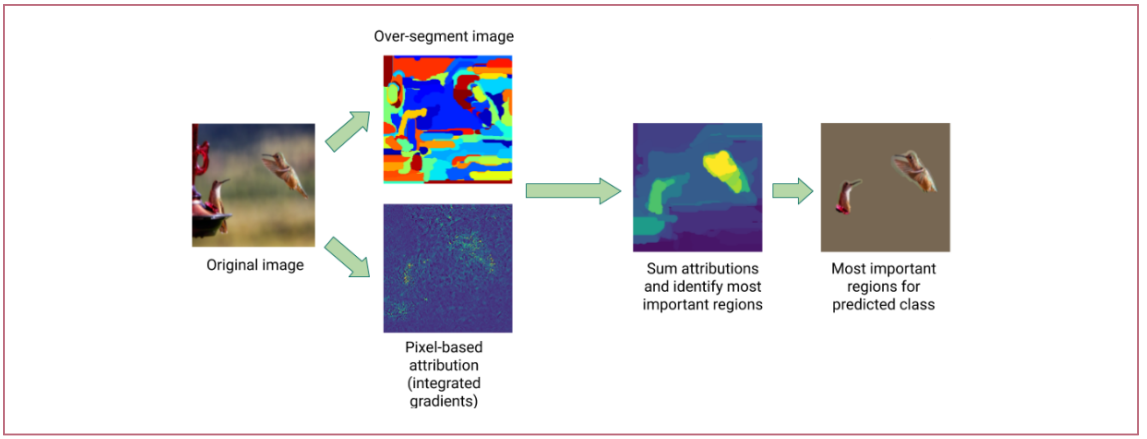
■ AI 안전성/신뢰성에 대한 상세 규정은 부재한 상황이나 글로벌 빅테크 기업들은 자체적인 연구를 통해 안전하고, 신뢰할 수 있는 AI 시스템 구축을 위한 Responsible AI Toolkit*들을 개발·공개 중

* AI 시스템을 안전하고, 신뢰할 수 있으며 윤리적 방식으로 평가·개발·배포하는 접근방식인 responsible AI를 실제로 운영함에 도움이 되는 통합 도구 및 기능 모음

● Amazon은 기초 모델 평가, 생성형 AI 안전 장치, 편향 감지 도구, 모델 동작 원리 이해 도구, 모니터링 도구, 거버넌스 개선 도구 등의 책임있는 AI 구축을 위한 다양한 리소스를 제공

- (Model Evaluation on Amazon Bedrock) 사용자가 정확성, 안전성 등의 지표를 기반으로 특정 사용 사례에 가장 적합한 기초모델을 평가, 비교, 선택할 수 있도록 지원
- (Bedrock Guardrails) 사용자가 생성 AI 어플리케이션에서 피해야 할 주제를 지정하여 제한된 카테고리에 속하는 쿼리 및 응답을 자동으로 감지하고 방지하도록 지원
- (Amazon SageMaker Clarify) 데이터 준비 단계, 모델 훈련 이후 단계, 배포된 모델에서 특정 속성을 검사하여 잠재적 편향을 감지함으로써 사용자가 편향을 완화하는 데 도움을 주고, 모델 동작에 대한 가시성을 향상하여 이해관계자에게 투명성 및 설명가능성 제공
- (Amazon SageMaker Model Monitor) 배포된 모델에서 부정확한 예측을 자동으로 감지하여 사용자에게 경고
- (Amazon Titan Image Generator) 생성형 이미지를 제공하는 프로그램으로 부적절한 사용자 입력·모델 출력을 감지 및 제거하는 필터링 기능 구축, 모델 출력의 인구통계학적 다양성 개선 등 모델 개발 프로세스의 각 단계에서 책임있는 AI를 고려하여 AI 생성 이미지를 식별하는 메커니즘을 제공
- (ML Governance from Amazon SageMaker) 기계학습(ML) 모델에 대한 보다 높은 수준의 제어 및 가시성을 제공하여 ML 프로젝트의 거버넌스 개선 지원

- Google은 Explainable AI, Model Cards, TensorFlow Open-source Toolkit 등을 개발 및 제공하여 구조화된 방식으로 모델 투명성 향상을 지원
  - (Explainable AI) ML 모델의 결과를 파악하고 해석할 수 있는 도구로 해석 가능하고 포괄적인 AI 설계, 최종 사용자의 신뢰도와 투명성 제고, 모델 거버넌스 간소화를 지원
  - (Model Cards) ML 모델의 투명성과 책임성 향상을 위한 도구로 모델의 목적, 설계, 훈련 데이터, 성능, 편향 등에 대한 정보를 제공하여 사용자가 모델의 잠재적 한계를 파악하여 이를 책임감 있게 사용하도록 지원
    ※ 모델의 편향에 대한 정보, 모델의 제한 사항 설명 등을 제공
  - (TensorFlow Open-source Toolkit) Explainable AI, Model Cards 도구를 포함하여 데이터의 공정성 및 편향 문제 완화, 공정한 결과를 얻기 위한 ML 모델 학습, 개인 정보 보호를 준수하는 ML 모델 학습, 공정성 지표 등을 통한 ML 모델 평가, 해석 가능하고 포용적인 ML 모델 개발 등을 통합적으로 지원
  - (Vertex Explainable AI) 개발자와 사용자가 ML 모델의 예측을 이해하고 해석할 수 있도록 도와주는 Google 클라우드 AI 플랫폼(Vertex AI) 내 도구 및 프레임워크 도구로 두 가지 주요 유형의 설명을 제공
    ※ (기능 기반 설명) Sharpley값 및 통합 그래디언트, XRAI 등의 기술을 통해 특정 예제에 대한 모델의 예측에 영향을 미치는 각 입력 기능의 상대적 중요성을 식별
    ※ (예시 기반 설명) 입력 기능의 특정 변화가 특정 모델의 예측에 미치는 영향을 보여줌으로써 인터랙티브하게 사용자의 모델 동작 탐색을 지원
    ※ (기존 Explainable AI와의 차별성) 다양한 설명 가능 방법 및 도구를 단일 통합 플랫폼으로 지원, 다양한 모델 유형 작업 지원, 고급 설명 가능성 기능 지원 등



[그림 4] XRAI 기술 개념도

※ 자료출처: Google Cloud Vertex AI https://cloud.google.com/vertex-ai/docs/explainable-ai/overview

- Microsoft는 데이터 과학자가 AI를 공정성, 투명성, 책임성 원칙에 따라 책임 있게 운영할 수 있도록 모델 디버깅 및 책임 있는 의사 결정 기능을 지원하는 Responsible AI Dashboard를 개발·배포
  - (데이터 분석) 데이터 세트 분산 및 통계를 이해하고 탐색
  - (Fairlearn 패키지) 모델 성능 평가 및 모델의 그룹 공정성 문제*를 평가
    ※ 모델의 예측이 다양한 사람들/집단에 미치는 영향
  - (Erroranalysis 패키지) 잘못된 데이터 코호트를 신속하게 식별하여 ML 실무자에게 모델 실패 분포에 대한 더 상세한 정보를 제공
  - (InterpretML 패키지) 글로벌 설명, 로컬 설명 기능 등 ML 모델의 예측에 대한 인간이 이해할 수 있는 설명을 생성하여 모델의 심층 진단을 지원
    ※ (해석 기술) SHAP text, SHAP vision, Guided Backprop, Guided gradCAM, Integrated Gradients, XRAI, D-RISE 등
  - (Counterfactual Analysis and What-If) 작업 입력 변경 시 모델이 무엇을 예측할지 설명함으로써 사용자가 ML 모델이 입력 변경에 어떻게 반응하는지 이해하고, 디버깅할 수 있도록 지원
    ※ (반사실 예제 생성방법) Randomized search, Genetic search, KD tree search 등
  - (EconML 패키지) 모델의 처리 기능이 실제 결과에 미치는 인과 영향을 확인

- IBM은 watsonx.governance 및 개별 도구들을 통해 AI 모델 책임성, 투명성, 설명 가능성 지원
  - (watsonx.governanace) 자동화, 모니터링, 캡처, 협업 기능을 지원하여 AI 책임성 향상
    ※ (자동화) 운영 위험, 정책, 규정 준수, 재무 관리, IT 거버넌스 및 내/외부 감사를 아우르는 자동화된 확장형 거버넌스, 위험 및 규정 준수(GRC) 도구로 모델 위험을 사전에 감지 및 완화하고 AI 규정을 시행 가능한 정책으로 변환하여 자동으로 적용
    ※ (모니터링) AI 수명주기 전반에 걸쳐 모델을 모니터링, 분류 및 제어하여 편향성, 드리프트 및 모델 재교육의 필요성을 선제적으로 식별하고 완화하여 예측 정확도를 향상
    ※ (캡처) 모델 검증자와 승인자가 팩트 시트에 액세스하여 모델 라이프사이클 세부 정보를 항상 최신 상태로 확인할 수 있도록 지원하여 설명 가능성을 높이고 감사, 이해관계자, 주주 또는 고객 요청에 대한 지원을 제공
    ※ (협업) 협업 도구와 사용자 기반 동적 대시보드, 차트 및 차원별 보고를 지원하여 프로세스에 대한 가시성 및 설명 가능한 AI 결과 향상
  - (AI Privacy 360) AI 기반 솔루션의 개인 정보 보호 위험 평가를 지원하고 관련 개인 정보 보호 요구사항을 준수하는데 도움이 되는 도구

- (AI Adversarial Robustness 360) 개발자와 연구자가 회피, 중독, 추출 및 추론과 같은 적대적 위협으로부터 ML 모델과 애플리케이션을 평가·방어할 수 있는 도구
- (AI Fairness 360) 데이터 세트와 ML 모델에서 원치 않는 편향을 규명하고 이러한 편향을 완화하는 최첨단 알고리즘을 확인하는 포괄적인 오픈소스 측정항목 도구
- (AI Explainability 360) ML 모델이 AI 애플리케이션 수명 주기 전반에 걸쳐 다양한 수단을 통해 레이블을 예측하는 방법을 이해하는 것을 지원
- (Uncertainty Quantification 360) AI가 불확실성을 표현할 수 있는 기능을 제공하여 AI의 안전한 배포 및 사용에 중요한 투명성을 강화
- (AI FactSheets 360) AI 모델이나 서비스의 생성·배포에 관한 관련 정보*의 모음을 제공하는 도구로 모델의 투명성을 강화

  * 모델의 목적과 중요도, 데이터 세트, 모델 또는 서비스의 측정된 특성, 모델이나 서비스의 생성 및 배포 프로세스 중에 취한 조치 등

● META는 사진 및 동영상 속 사물과 사람을 분류하고 감지함에 있어 AI 모델의 공정성을 평가하는 FACET* 벤치마크 도구와 잠재적인 통계 편향에 대한 측정값을 조기에 체계적으로 표면화하는 Fairness Flow 도구를 개발

  * Fairness in Computer Vision EvaluaTion

- (FACET) 데이터 세트는 인간이 라벨링한 50,000명의 이미지 32,000개로 구성되며 AI 모델의 잠재적 편견을 심층적으로 조사하기 위해 다양한 인구통계학적 속성, 직업 및 활동을 다루며 개발자가 AI 모델의 편견을 이해·완화하는 작업을 지원
- (Fairness Flow) 그룹별 라벨링 및 모델 성능에 대한 높은 수준의 통계적 이해를 제공하여 분석할 수 있는 모델 유형과 관련된 시스템, 프로세스 및 정책에 대한 심층적인 조사를 지원

2. 산업 동향

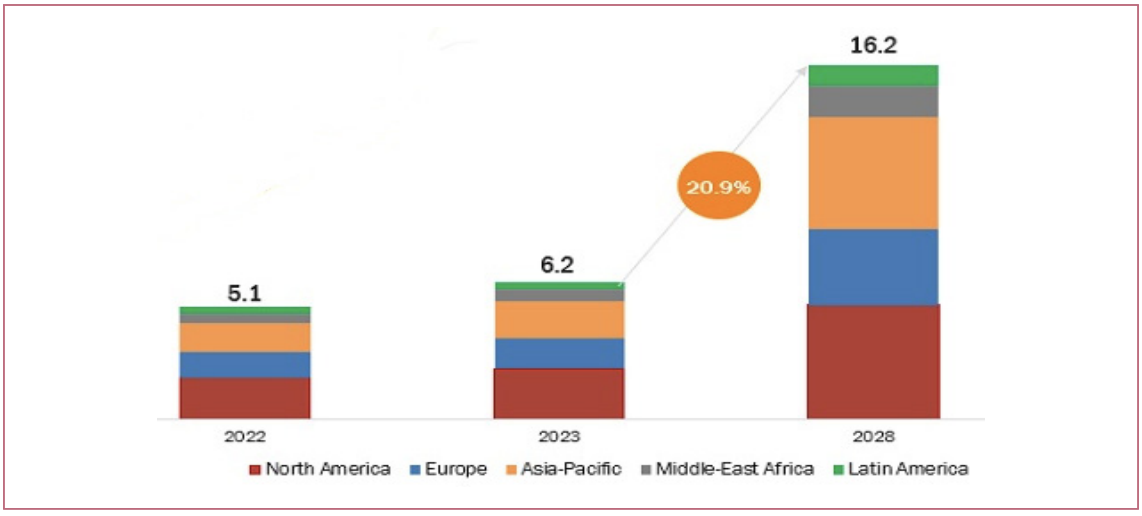
AI 안전성·신뢰성 확보를 위한 AI 거버넌스 시장과 이를 구현함에 필요한 설명가능한 AI(XAI), 데이터 프라이버시 보호 기술 시장규모도 지속적으로 성장 중

- AI 거버넌스는 AI 시스템의 책임 있는 개발, 배포 및 운영을 보장하는 것을 목표로 하는 도구, 프레임워크 및 프로세스 등을 의미하며 해당 시장 규모는 '22년 기준 약 1.2억 달러로 추정되었고, 향후 CAGR 48%로 성장하여 '32년 약 54억 달러 규모로 확대 예상

  ※ (출처) Global Market Insights, https://www.gminsights.com/industry-analysis/ai-governance-market

- 설명가능한 AI(XAI) 글로벌 시장 규모는 '23년 기준 약 62억 달러로 추정되고 있고, 향후 CAGR 20.9%로 성장하여 '28년 약 162억 달러 규모로 확대 예상

  - XAI의 글로벌 시장은 미국, 캐나다, 영국, 프랑스 등 북미와 유럽의 ICT분야 선도국들이 주도하고 있으며, 성장세는 중국, 인도와 같은 개발 도상국이 더 높을 것으로 예상



[그림 5] XAI 글로벌 시장 규모 예측('22~'28)

※ 자료출처: Markets and Markets, https://www.marketsandmarkets.com/Market-Reports/explainable-ai-market-47650132.html

- 데이터 프라이버시 SW 시장 규모는 '21년 기준 약 16.5억 달러 규모로 추정되었고, 향후 CAGR 40%로 성장하여 '29년 약 240억 달러 규모로 확대 예상

  ※ (출처) Maximize, https://www.maximizemarketresearch.com/market-report/data-privacy-software-market/145730/

글로벌 빅테크 기업들은 AI 안전성·신뢰성 확보를 위하여 윤리원칙 확립, AI 거버넌스 구축, 관련 R&D 수행, 이해관계자 협의 등 다양한 노력을 추구

- (Amazon) AI의 책임감 있고 안전한 사용 촉진을 위해 내부 AI 정책을 수립하고, 백악관, 정책 입안자, 기술 산업, 연구원 및 AI 커뮤니티와 지속적으로 협력, 관련 R&D* 등을 추진

  * (예) 개방형 언어 생성 편향 측정, 얼굴 인식 편향 벤치마킹 개발, 공정한 대표 학습 방법 개발 등

- (Google) AI 개발 원칙 수립 및 지속적 업데이트, 설계·배포하지 않을 AI 애플리케이션 설정, AI 거버넌스 구축 및 운영, 관련 국제 표준 기구 및 이해관계자와의 협력, 내부 교육·R&D* 추진 등

  * (예) ML 공정성을 위한 MinDiff, XAI, 적대적 견고성 평가, ML 과정 피부색 평가 개선 등

- (Microsoft) 백악관 및 이해관계자와의 협력, NIST AI 위험 관리 프레임워크 구현, 인간 중심 AI를 위한 연구, AI와 인류 번영의 미래에 대한 연구 등 추진, AI 안전성 확보를 위한 AI Red Team 운영 등

- (Meta) 책임 있는 AI 핵심 요소 강조, AI 공정성과 투명성 확보를 위한 데이터 세트 및 도구 구축, 인종 전반에 걸쳐 플랫폼에서 AI 모델의 공정성을 의미 있게 측정할 수 있는 데이터 접근 방법 개발, 광고 게재의 공정성 향상을 위한 연구 등

- (OpenAI) 안전한 AI 개발을 위한 헌장 제정, AI 문제 해결을 위한 안전팀 신설 및 운영, AI 신뢰성 확보를 위한 업계 및 정책 입안자와의 협력, 책임감 있는 사용을 위한 모범 사례 개발 및 플랫폼 오용 모니터링 등

- 생성형 AI 관련 기업인 Amazon, Google, Meta, Microsoft, OpenAI, Anthropic, Inflection AI의 임원들은 Biden 대통령과 만나 백악관이 개발한 안전하고 보안이 유지되며 신뢰할 수 있는 AI 관련 약속에 자발적으로 동의('23.7.)

  - (안전) 제품 출시 전 AI 시스템에 대한 독립적인 전문가의 내부 및 외부 보안 테스트를 통해 제품의 안전성을 보장하고 전 세계의 AI 위험 관리 노력에서 얻은 인사이트 공유
  - (보안) 사이버 보안 및 내부 위협 보호 장치에 투자하고 제3자의 취약점 발견 및 보고를 촉진하여 보안을 최우선하는 시스템 구축

    ※ 보안 이슈 발생 시 특정 입력이 출력에 미치는 영향 등 알고리즘 편향의 중요한 결정 요인을 결정하는 AI 알고리즘의 "공개되지 않은 모델 가중치" 공개 등
  - (신뢰) 워터마킹 시스템과 같이 콘텐츠가 AI로 생성된 시기를 사용자가 알 수 있도록 하는 '강력한 기술 메커니즘'을 개발하여 대중의 신뢰 확보

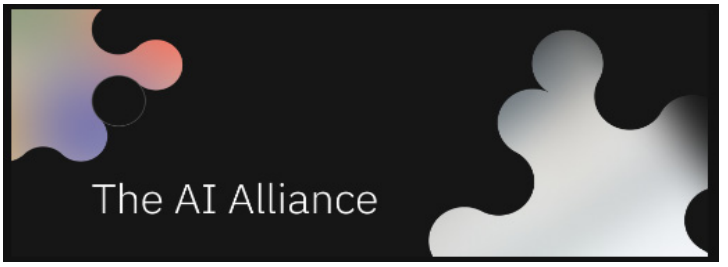
또한 글로벌 빅테크 기업들은 안전·신뢰 AI 분야 주도권 확보를 위한 연합 결성

- 폐쇄형 AI 진영인 OpenAI, Google, Microsoft는 대규모 ML 모델의 안전하고 책임감 있는 개발을 보장할 기관인 Frontier Model Forum의 결성을 발표('23.7)
  - AI 안전 연구 발전, 모범 사례 규명, 다방면에 걸친 협업, AI가 사회의 가장 큰 과제를 해결할 수 있도록 지원
    - ※ 프론티어 모델의 책임감 있는 개발을 촉진하고, 위험을 최소화하며 역량과 안전성에 대한 독립적이고 표준화된 평가를 가능하게 함에 일조
    - ※ 프론티어 모델의 책임감 있는 개발 및 배포를 위한 모범 사례 규명
    - ※ 정책 입안자, 학계, 시민 사회, 기업이 함께 협력하고 신뢰와 안전 위험에 대한 지식 공유
    - ※ 기후 변화 완화 및 적응, 암 조기 발견 및 예방, 사이버 위협 퇴치 등의 문제를 해결하기 위한 애플리케이션 개발 지원
  - 안전·신뢰 AI 분야 연구를 촉진하기 위한 1,000만 달러 이상의 AI 안전기금 조성('23.10.)

[그림 6] Frontier Model forum

※ 자료출처: Frontier Model Forum

- 최근 개방형 AI 진영인 Meta와 IBM도 안전·신뢰 AI 발전을 위한 AI Alliance를 출범('23.12.)
  - Meta와 IBM은 개방적이고 안전하며 책임감 있는 AI를 발전시키기는 것을 목표로 한 50개 이상의 조직으로 구성된 국제 연합 AI Alliance를 출범
  - 이 동맹에는 Advanced Micro Devices, Dell Technologies, Intel, NASA, Linux Foundation 등의 기술 기업, 선도적인 연구 대학, 과학 기관이 회원으로 참여
  - AI 시스템의 책임감 있는 개발·사용을 가능하게 하는 벤치마크 및 평가 리소스 개발·배포, 활발한 AI HW 생태계 육성, 글로벌 AI 기술 구축 및 탐색적 연구 지원, AI 규제에 대한 대중 담론과 정책 입안자에게 정보를 제공하는 교육 리소스 개발 등



[그림 7] AI Alliance

※ 자료출처: IBM Newsroom

국내 기업들도 대기업을 중심으로 책임 있는 AI 구현을 위해 윤리지침 및 가이드라인 수립, 관련 위원회 운영, 이해관계자와의 협력 등을 지속

- (네이버) '22년 3월 "네이버 AI 윤리원칙" 발표 및 AI 윤리위원회를 설립·운영하며 인공지능과 관련된 윤리적인 문제에 대한 조언 제공, 투명성과 사용자 프라이버시에 대한 고려 강조, AI 윤리 교육 프로그램 운영 및 AI 윤리 연구* 지원

  * ACL 2023에서 초거대 AI 윤리 관련 논문 7개 채택('23.7.)

- (삼성전자) '19년 공정성, 투명성, 책임성 등 3대 원칙을 중심으로 한 "삼성 AI 윤리원칙" 발표, AI 윤리기준 정립 및 공동연구를 수행하는 국제 컨소시엄인 'Partnership on AI' 가입, 교육 프로그램 운영을 통한 직원 인식 제고 등을 추진

- (카카오) '18년 국내 기업 최초의 알고리즘 윤리규범 마련, 투명성, 공정성, 프라이버시 보호 등에 중점을 둔 '카카오 AI 윤리 플랫폼'을 구축하여 인공지능 기술과 서비스의 윤리적인 측면을 강화하고, '22년 국내 기업 최초의 전사적 AI 윤리 논의 기구인 '공동체 기술윤리위원회' 신설, AI 윤리 교육 프로그램 운영 및 AI 윤리 연구 지원 등 추진

- (LG 전자) '22년 'AI 윤리원칙' 발표 및 'AI 윤리 점검 TF'와 LG 계열사들이 참여하는 'AI 윤리 워킹 그룹' 협의체 운영, '23년 11월 UNESCO와 AI 윤리 실행·확산을 위한 협력 의향서 체결하여 AI 윤리 영향 평가, 데이터 프라이버시, 보안을 보장하는 거버넌스 모델 추구, AI 윤리 관련 공개 온라인 강좌 운영 등 추진

- (SK 텔레콤) '21년 AI 윤리 가치를 사규에 반영하며 AI 추구가치로 사람 중심의 AI, 사회적 가치, 무해성, 기술 안전성, 공정성, 투명성 등을 강조, AI 윤리를 시스템적으로 정착시키기 위해 사규를 포함한 사내 프로세스와 고객 소통 채널을 구축하고 외부 전문가 자문단 운영 예정

최근 국내 대기업들도 거대언어모델(LLM)의 안전성·신뢰성 확보를 위한 컨소시엄 구축('23.10.)

- 네이버, 카카오, SKT, KT, LG AI 연구원 등은 컨소시엄을 구축하고, 한국지능정보사회진흥원의 'AI 학습용 데이터 구축사업'에 지원하여 LLM의 신뢰성 및 윤리성 평가에 사용될 기준 데이터셋 구축 예정
- 해당 컨소시엄은 안전·신뢰 AI 논의를 통하여 과기부 AI 규제 설정에 일조할 것으로 전망

3. 정책 동향

한국을 포함한 주요국은 AI 기술의 긍정적 측면을 극대화하기 위한 AI 촉진 정책과 부정적 측면을 완화/예방하기 위한 안전·신뢰 AI 정책을 동시에 추진 중

〈표 1〉 주요국 AI 촉진 정책 및 안전·신뢰 AI 정책

국가명	주요 AI 촉진 정책	주요 안전·신뢰 AI 정책
한국	• AI 국가전략 ('19.12) • K-클라우드 프로젝트 ('23.6) • 초거대 AI 경쟁력 강화방안 ('23.7) • 대한민국 초거대 AI 도약방안 ('23.9)	• AI 윤리기준 ('20.12) • 신뢰할 수 있는 인공지능 실현전략 ('21.5) • AI-저작권법 제도개선 TF ('23.2) • 안전한 개인정보 활용 정책 ('23.8) • 디지털 권리장전 ('23.9) • AI 신뢰성 검인증 ('23.12)
미국	• 국가 AI 이니셔티브 ('21.1) • 국가 AI R&D 전략계획 ('23.5)	• AI 권리장전 ('22.10) • AI 위험관리 프레임워크 ('23.1) • 백악관-생성 AI 기업 간담회 ('23.7) • 행정명령 E14410 ('23.10)
중국	• 차세대 AI 발전규획 ('17.7) • AI 활용 고도화 지도의견 ('22.7)	• 인터넷 정보서비스 알고리즘 추천 관리 규정 ('22.3) • 인터넷 정보서비스 심층합성 관리규정 ('23.1) • 생성 AI 서비스 관리 임시 시행 방법 ('23.7) • 글로벌 인공지능 거버넌스 이니셔티브('23.11)
일본	• AI 전략 2019 ('19.6) • AI 전략회의 ('23.5)	• G7 히로시마 정상회의 ('23.5) • 지적재산 추진계획 2023 ('23.6)
영국	• 국가 AI 전략 ('21.10) • 국방 AI 전략 ('22.6) • 과학 및 기술 프레임워크 ('23.3)	• 프론티어 AI 태스크포스 ('23.4) • 기반 모델 검토 보고서 ('23.9) • 프런티어 AI 안전을 위한 기업정책 지침 ('23.10) • AI 안전연구소 설립 제안 ('23.11) • 안전한 AI 개발 지침 ('23.11)
EU	• EU를 위한 AI ('18.4) • ELLIS 연구소 ('18.12) • Euro HPC 프로그램 ('23.9)	• 신뢰할 수 있는 AI 가이드라인 ('19.4) • AI Act ('21.4)

● 한국의 AI 촉진 정책

- (AI 국가전략, ’19.12) 과기부는 국가차원에서 AI 산업을 지원·육성하기 위한 R&D 지원, 교육 및 인력 양성 등 3대 분야 9대 전략, 100대 실행과제를 제시
- (K-클라우드 프로젝트, ’23.6) 과기부는 세계 최고 수준의 초고속·저전력 국산 AI 반도체를 개발해 데이터센터에 적용하여 국내 클라우드 경쟁력을 강화

  * 3단계에 걸쳐 추진되며 2023년 착수한 1단계 사업에서는 현재 상용화 초기 단계의 국산 NPU를 데이터센터에 적용하고 클라우드 기반 AI 서비스까지 제공하는 실증 사업 진행

- (초거대 AI 경쟁력 강화방안, ’23.7) 과기부는 국내 초거대 AI 생태계를 활성화하고 중소·벤처기업 및 공공 부문에 초거대 AI를 선도적으로 도입하기 위해 ‘민간의 첨단 초거대 인공지능 활용지원 사업’을 2023년부터 추진
- (대한민국 초거대 AI 도약방안, ’23.9) 과기부는 AI 국제협력 확대, 전 국민 AI 일상화 추진, 디지털 권리장전 수립, AI 윤리·신뢰성 확보 등 디지털 모범국가를 향한 ‘대한민국 인공지능 도약방안’을 발표

● 한국의 안전·신뢰 AI 정책

- (AI 윤리기준, ’20.12) 인간성(Humanity)을 위한 AI 3대 원칙과 10대 요건을 담은 국가 인공지능 윤리기준 수립 및 공개

  * (3대 기본원칙) 인간의 존엄성 원칙, 사회의 공공선 원칙, 기술의 합목적성 원칙 (10대 핵심요건) 3대 기본원칙을 실천하고 이행할 수 있도록 인공지능 개발~활용 전 과정에서 ① 인권 보장, ② 프라이버시 보호, ③ 다양성 존중, ④ 침해금지, ⑤ 공공성, ⑥ 연대성, ⑦ 데이터 관리, ⑧ 책임성, ⑨ 안전성, ⑩ 투명성의 요건 충족

- (신뢰할 수 있는 인공지능 실현전략, ’21.5) “누구나 신뢰할 수 있는 인공지능, 모두가 누릴 수 있는 인공지능 구현”을 위한 3대 전략 10대 실천과제 제시

  * (3대전략) 신뢰 가능한 인공지능 구현 환경 조성, 안전한 인공지능 활용을 위한 기반 마련, 사회 전반의 건전한 인공지능 의식 확산

- (AI-저작권법 제도개선 TF, ’23.2) 문체부는 AI 법률, 저작권, 콘텐츠, 인공지능 전문가들로 구성된 AI-저작권법 제도개선 워킹그룹을 발족하고 및 2023년 연내 가이드 수립

  * (가칭) ‘저작권 관점에서의 AI 산출물 활용 가이드(안)’도 마련할 예정

- (안전한 개인정보 활용 정책, ’23.8) 개인정보보호위원회는 ‘인공지능 시대 안전한 개인정보 활용 정책방향’을 통해 AI 개발·서비스 기획-데이터 수집-AI 학습-서비스 제공 등 단계별로 개인정보를 어떠한 원칙과 기준에 입각하여 처리할 수 있는지 가이드라인 제시

* 개보위는 'AI 프라이버시 전담팀'을 '23년 10월 신설하고 AI 모델·서비스를 개발·제공하는 사업자와 소통창구를 마련하여 사안별로 개인정보 처리의 적법성, 안전성 등에 대한 법령해석을 지원하거나 규제 샌드박스 적용을 검토하는 등 적극적인 컨설팅 수행

- (디지털 권리장전, '23.9) 과기부는 디지털 심화 시대에 맞는 국가적 차원의 기준과 원칙을 제시하기 위해 디지털 질서 규범의 기본 방향을 담은 '디지털 권리장전' 공개

  * (5대 기본원칙) '자유와 권리의 보장', '공정한 접근과 기회의 균등', '안전과 신뢰의 확보', '디지털 혁신의 촉진', '인류 후생의 증진'

- (AI 신뢰성 검인증, '23.12) 과기정통부는 한국정보통신기술협회(TTA)와 함께 인공지능 기반 제품·서비스를 대상으로 수행

  * 인공지능 윤리기준 10대 원칙 중 기술적으로 구현 및 검증 가능한 △다양성 존중 △책임성 △투명성 △안전성을 중심으로 15개의 신뢰성 요구사항과 67개의 검증항목을 통해 제품·서비스의 신뢰성 확보 여부를 시험하고 인증

● 미국의 AI 촉진 정책

- (국가 AI 이니셔티브, '21.1) 미국은 '20년 국가 AI 이니셔티브법을 제정하고 '21년 미국 최초의 국가 AI 전략으로서 '국가 AI 이니셔티브'를 추진

  * 법에 의거해 국가 AI 이니셔티브실(National Artificial Intelligence Initiative Office)을 신설하고 연방 정부의 AI 관련 국가 이니셔티브 및 프로젝트 업무 조정

- (국가 AI R&D 전략계획, '23.5) 미국의 AI 기술 리더십 확보를 위한 국가 AI 연구개발 전략서로 '16년 최초 수립, '19년 개정된 후 '23년 5월 2차 개정안 발간

  * 9개 전략 목표: ① 근본적이고 책임 있는 AI 연구에 장기적으로 투자, ② 인간-AI 협업의 효과적인 방법 개발, ③ AI의 윤리적, 법적, 사회적 영향을 이해하고 해결, ④ AI 시스템의 안전과 보안을 보장, ⑤ AI 학습 및 테스트를 위해 공유할 수 있는 공개 데이터 세트 및 환경 개발, ⑥ 표준 및 벤치마크를 통해 AI 시스템 측정 및 평가, ⑦ 국가 AI R&D 인력 수요에 대한 이해 향상, ⑧ AI 발전을 가속화하기 위해 공공-민간 파트너십 확대, ⑨ AI 연구의 국제협력에 대한 원칙적이고 조정된 접근방식을 확립[2023년 개정판에 추가]

● 미국의 안전·신뢰 AI 정책

- (AI 권리장전, '22.10) 백악관 내 과학기술정책국(OSTP)는 AI 시스템 설계, 이용, 배포 과정에서 시민 권리 보호를 위한 지침 서로 AI 권리장전 마련

  * 5대 원칙 제시: 안전하고 효과적인 시스템, 알고리즘 차별 금지, 개인정보보호, 고지 및 설명, 인간 대안/검토 사항/비상 대책 검토

- (AI 위험 관리 프레임워크, '23.1) 상무부 산하 국립표준기술연구소(NIST)는 공공 및 민간에서 활용하도록 AI 위험 관리 프레임워크(AI Risk Management Framework1.0)를 마련

- (백악관-생성 AI 기업 간담회, ’23.7) 백악관에서 대표적 생성 AI기업* 대표를 소집하여 안전하고 보안이 유지되며 투명한 AI 개발에 대한 백악관-기업 간 공동 성명 발표

  * 7개 참여 기업: 아마존, 앤스로픽, 구글, 인플렉션, 메타, 마이크로소프트, 오픈AI

- (행정명령 E14410, ’23.10) 안전하고 신뢰할 수 있는 AI에 관한 행정명령 공표

  * 행정명령에는 △새로운 안전 및 보안 표준 제정, △소비자 사생활 보호, △형평성과 시민 권리 증진, △소비자들에게 이익 제공, △ 노동자 보호, △혁신과 경쟁 촉진, △국제적 파트너들과 협력해 미국의 리더십 증진, △책임있고, 효과적인 정부 활용을 보장한다는 내용 포함

● 중국의 AI 촉진 정책

- (차세대 AI 발전규획, ’17.7) 2030년까지 인공지능 세계 1위 도약을 목표로 연구 개발, 산업 육성, 인재 양성 등 시장 활성화 정책 제시

  * ’17년 12월에는 차세대 AI 산업발전 3개년 실행계획(2018~2020)을 수립하고 추진

- (AI 활용 고도화 지도의견, ’22.7) AI 육성을 위해 AI 주요 활용사례 구축, AI 활용 혁신 능력제고, AI 활용사례 개방 가속화, AI 혁신 강화 등 5개 분야 15대 중점 과제 규정

● 중국의 안전·신뢰 AI 정책

- (인터넷 정보서비스 알고리즘 추천 관리 규정, ’22.3) 사용자의 사용 기록과 개인 정보를 바탕으로 콘텐츠를 배치하거나 추천하는 플랫폼을 사업자를 대상으로 불량 정보가 확산되지 않도록 공급자 의무를 부과

- (인터넷 정보서비스 심층합성 관리규정, ’23.1) 딥러닝 합성 서비스 제공 업체에 딥러닝 합성 서비스 사용자의 신원 인증과 콘텐츠 관리, 안전 관리 책임 등을 부과

- (생성 AI 서비스 관리 임시 시행 방법, ’23.7) 국가인터넷정보판공실은 국가발전개혁위원회, 교육부, 과기부, 공업정보화부, 공안부, 광전총국과 함께 중국 대중에게 글, 그림, 기타 콘텐츠를 생성하는 AI 서비스에 대한 관리 지침을 발표하고 8월 15일부터 시행

  * 공급자 의무: 국가 규정에 따른 안전 평가, 데이터훈련 및 라벨링, 콘텐츠 관리, 콘텐츠 표기, 개인정보보호, 운영상 조치 의무 부여

- (글로벌 인공지능 거버넌스 이니셔티브 ’23.11) 개인정보의 보호와 허위정보 유포 제약과 함께, AI 개발에서 국가간 차별없는 동등한 기회 부여 및 권리 보장 제안

● 일본의 AI 촉진 정책

- (AI전략 2019, ’19.6) 아베 정부는 ’18년 9월 설치한 통합혁신전략추진회의에서 일본의 인공지능 종합 전략인 ‘AI 전략 2019’를 수립하고 발표

  * 일본의 AI 전략은 환경 변화를 반영하여 세부 목표 및 과업을 조정하여 ’21년, ’22년 개정됨

- (AI 전략회의, '23.5) '23년 5월 생성 AI를 포함한 국가 전략을 수립하고 정책 방향성을 제시하는 AI 정책 총괄 자문기구인 'AI전략회의'를 신설하고 AI 리스크 대응, AI 이용 촉진, AI 개발력 강화를 위한 정책* 수립

  * AI 전략회의, 'AI에 관한 잠정적 논점 정리('23.5)'라는 보고서 공개

● 일본의 안전·신뢰 AI 정책

- (G7 히로시마 정상회의, '23.5) 히로시마에서 열린 G7 정상회의에서 신뢰할 수 있는 AI를 위한 국제 지침을 연말까지 수립한다는 '히로시마 AI 프로세스' 제안

  * 2023년 10월에 AI 개발자를 위한 국제 지침과 지켜야 할 국제 행동 규범(11개 항목)에 대해 합의

- (지적재산 추진계획 2023, '23.6) 생성 AI와 저작권 관계를 구체적인 사례에 따라 정리하고, 필요한 방안을 검토하는 시책 방향성을 제시하고 추후 구체적 사례 분석 및 관련 법률 정비를 지속적으로 진행 예정

● 영국의 AI 촉진 정책

- (국가 AI 전략, '21.10) 향후 10년간 AI 분야에서 초강대국 리더십 확보를 국가의 AI 비전으로 삼아 전략적 투자, 포괄적 혜택, 거버넌스 3개 영역에서 단기(향후 3개월), 중기(6~12개월), 장기(12개월 이상) 정책 과제 제시

  * 2022년 7월에는 국가 AI 전략의 추진 상황을 점검하고, 향후 구체적인 추진방안을 설정한 국가 AI 전략-실행계획(National AI Strategy-AI Action Plan) 발표

- (국방 AI 전략, '22.6) 4대 목표로 국방을 'AI 지원' 조직으로 변혁, 국방 이점을 위해 속도와 규모에 맞게 AI를 채택하고 활용, 영국의 국방·안보 AI 생태계 강화, 보안·안정성 및 민주적 가치를 촉진하기 위해 글로벌 AI 개발 구체화로 설정

- (과학 및 기술 프레임워크, '23.3) 영국 과학혁신기술부는 AI와 슈퍼컴퓨터, 양자 등의 분야에서 영국이 글로벌 과학기술 강국이 되기 위한 인프라, 투자, 기술 강화에 2030년까지 3억 7천만 파운드의 투자 계획을 담은 'UK 과학 및 기술 프레임워크' 발표

● 영국의 안전·신뢰 AI 정책

- (프론티어 AI 태스크포스, '23.4) 태스크포스는 정부 내 스타트업으로 ARC Evals, RAND, Trail of Bits를 비롯한 주요 기술 기관과 협력하며 AI의 최전선에서 위험을 평가할 수 있는 AI 연구팀을 구축하는 임무를 수행

  * '23년 7월 시작된 Foundation Model Taskforce가 2023년 9월 이름을 Frontier AI TF로 이름을 변경하였고 2023년 11월 발표한 AI안전연구소(AI Safety Institute)로 기능 대체 예정

- (기반 모델 검토 보고서, '23.9) 영국 경쟁시장청(CMA)은 AI 기반 모델의 잠재력과 위험성을 동시에 평가하면서 기반 모델이 시장을 독점함으로써 경쟁을 저해하고, 이용자 피해 유발 가능성을 최소화하기 위한 기업의 7가지 원칙* 제시

  * 책임성, 접근성, 다양성, 선택가능성, 유연성, 공정성, 투명성 확보

- (프런티어 AI 안전을 위한 기업정책 지침, '23.10) 영국 과학혁신기술부는 프런티어 AI 개발 기업들을 위한 지침서를 개발해 보급

  * 지침에는 책임 있는 역량 확장, 모델 평가와 레드팀 구성, 모델 신고와 정보공유, 보안 통제, 취약점 신고 절차 수립, AI 생성 결과물의 출처 확인, AI 위험 연구, 모델 오용 방지와 감시, 데이터 입력 제어와 감사 방법 포함

- (AI 안전연구소 설립 제안, '23.11) 영국 총리는 '23년 11월 영국 블레츨리 파크에서 열린 'AI Safety Summit'에서 AI 안전·신뢰성 확보를 위해 AI 안전연구소 설립 제안하며 글로벌 AI 안전 연구의 허브 표방

  * 영국 블레츨리 파크에서 열린 AI안전정상회의(AI Safety Summit)에서 미국, 중국, 한국, EU를 포함 28개 국가가 AI위험성 규제에 합의한 '블레츨리 선언' 발표

- (안전한 AI 개발지침, '23.11) 영국 국가사이버보안센터(NCSC)와 미국 사이버안보 및 인프라안보국(CISA) 공동 개발, 우리나라(NIS) 포함 21개국 협력 기관으로 참여해 'Guidelines for secure AI system development' 개발

● EU의 AI 촉진 정책

- (EU를 위한 AI, '18.4) EU는 '18년 4월 EU 차원의 AI 전략인 'EU를 위한 AI'를 발표하였으며 경제 전반에 걸쳐 기술적 산업적 역량 및 AI 활용 증진, 사회·경제적 변화 준비, 윤리적·법적 프레임워크 확보를 3대 목표로 제시

- (ELLIS 연구소, '18.12) 유럽 학습 및 지능형 시스템 연구소(ELLIS*)는 유럽의 인공지능 기술주권 확보를 위한 '18년 설립된 범유럽 AI 연구 네트워크로 14개 연구프로그램, 박사 및 박사후 과정 인력 배출, AI 연구거점 구축(16개국 41개)을 진행 중

  * European Laboratory for Learning and Intelligent Systems

- (Euro HPC 프로그램, '23.9) '23년 9월 유럽집행위원회는 중소기업, 스타트업, 및 공공기관에 유럽 보유 슈퍼컴퓨팅 인프라* 활용을 지원하는 정책 발표

  * EU는 엑사스케일 전 단계의 슈퍼컴퓨터 '루미(핀란드)', '마레노스트럼 5(스페인)', '레오나르도(이탈리아)'를 보유하고 있으며 독일과 프랑스에서는 엑사스케일 슈퍼컴퓨터를 가동할 예정

● EU의 안전·신뢰 AI 정책

- (신뢰할 수 있는 AI 가이드라인, '19.4) 유럽집행위원회는 적법하고, 윤리적이며, 견실한 AI 시스템의 설계, 개발, 이용을 위한 원칙, 요구사항 및 자율 점검을 위한 평가 목록 개발을 목적으로 신뢰할 수 있는 인공지능 가이드라인을 마련하고 배포

  * 7대 요구사항: 인간의 감독권 허용, 기술적 견고성과 안전성 확보, 프라이버시 보호와 데이터 접근 규정 마련, 투명성 확보, 다양성·차별금지·공정성, 사회적·환경적 웰빙 지향, 책임성

- (AI Act, '21.4) 유럽의회는 '21년 4월 EU 내에서 출시 또는 서비스되는 인공지능에 대한 규제법으로 'AI Act'를 마련하고 '23년 12월 유럽의회, 집행위원회, 유럽연합 이사회 3자 합의를 마쳤으며 발효를 앞둔 상태

  * 법안에서는 AI에 제한 사항과 공급자의 안전 조치 등을 포함하고 있으며 위반 시 최대 3,500만달러 또는 매출의 7%의 과징금 처벌 규정을 포함

4. 요약 및 시사점

(이슈 종합 1) 최근 AI 안전성·신뢰성 확보에 대한 논의가 더욱 활발해지고 있으며 주요국 정부도 AI 안전 규제 설정 움직임 강화

● AI 기술 발전을 강조하던 대표적인 AI 학계 거장들도 이제는 AI 윤리 및 안전의 필요성을 주장하며 관련 문제 해결에 집중

- AI 선구자인 Geoffrey Hinton 교수는 AI 윤리·안전 문제 해결을 위하여 Google을 사임하였고('23.5.), 편견, 투명성, 책임, 개인 정보 보호 및 윤리원칙 준수 문제를 다루는 프레임워크와 지침을 개발하는데 기여할 계획

  ※ Hinton 교수는 New York Times와의 인터뷰에서 AI 분야 개척에 대한 후회를 말하며 AI가 허위 정보를 조장하고 일자리를 없애는 것에 대한 우려를 표명

- 딥러닝 모델 창시자인 Yoshua Bengio 교수도 AI 기술의 급속한 발전에 따른 예상치 못한 문제 발생으로 평생의 업적에 대해 '상실감'을 느꼈다고 언급하였고('23.6.), 글로벌 AI 윤리 논쟁을 주도

  ※ Bengio 교수는 최근 BBC와의 인터뷰에서 "AI가 얼마나 빨리 발전할지 알았다면 안전을 최우선으로 생각했을 것"이라고 발언

● EU는 세계 첫 AI 규제법인 'AI Act' 합의로 생성형 AI 등을 포함한 AI 규제 환경 구축 의지 공고화

- AI Act는 AI를 활용한 제품이나 서비스에 대해 '위험 기반 접근방식'을 취하고, 기술보다는 AI 사용 규제에 초점을 맞춘 세계 최초 AI 규제 법안
- 이는 약 4억 5천만 명의 EU 거주자에게 적용될 예정이나 이는 전 세계적으로 영향을 주어 글로벌 AI 규제 설정을 주도할 것으로 예상

※ "AI Act는 유럽의 판도를 바꿀 뿐만 아니라 관할권 전반에 걸쳐 AI를 규제하는 글로벌 모멘텀을 크게 강화할 세계 최초의 포괄적·수평적·구속력 있는 AI 규제" (Anu Bradford 교수, Columbia Law School)

● 자율규제 형식을 통해 AI 기술패권 강화 및 산업 육성을 도모하던 美, 中, 日 등도 최근 AI의 안전성·신뢰성 확보를 위한 정부 중심 규제로의 정책 기조 변화를 시사

- (美) '22년 하반기부터 정부 중심의 AI 규제 관련 지침/가이드라인을 제정·발표하고 있고, '23년부터 상무부가 AI Risk Management Framework를 마련하고, AI Safety Institute 설립을 추진함으로써 향후 AI 안전 규제가 공공·민간에 공통으로 적용될 것을 시사
- (中) 미국과의 AI 기술패권 경쟁을 위해 가장 느슨한 AI 규제 기조를 유지해오다 '23년 이후 딥러닝·생성형 AI 서비스를 대상으로한 세계 최고 수준의 강력한 규제를 창설

※ 사업자와 사용자, 데이터 및 알고리즘을 동시에 규제

- (日) 자율규제 형식을 유지하다 '23년 이후 AI 안전 관련 정상회의, 생성형 AI 지재권 관련 규제를 추진하고 있고, 최근 기시다 총리는 AI 기술을 개발하거나 사용하는 모든 기업을 위한 정부 지침 최종안을 채택하며 '24년 1월 AI Safety Institute 설립 계획을 발표

(이슈 종합 2) 주요국들은 안전·신뢰 AI 규제 표준 설정 목적의 규범 클러스터를 형성하여 서로 협력하되 자국에 유리한 룰(rule) 확보를 위해 경쟁할 것으로 예상

● (EU 진영) 'AI Act' 합의로 글로벌 AI 규제 선도와 자국 기업·산업 보호 기반 마련

- 세계 최초의 AI 규제 법안을 통해 글로벌 AI 규제 선두주자로서 EU의 입지를 강화
- 동시에 해당 법안에는 상대적으로 열위인 자국 내 AI 기업 및 산업 보호와 육성을 위하여 글로벌 AI 빅테크 기업들을 규제하기 위한 조항* 및 징벌적 벌금 부과 조항**을 포함

* (예) OpenAI, Google, Meta 등의 생성형 AI, LLM 및 범용 AI 시스템을 대상으로 모델 출시 전 투명성 의무 준수 요구(기술 문서 작성, EU 저작권법 준수, 학습 사용 데이터 공개 등 포함)

** 위반 기업들에게는 750만 유로(약 107억원) 또는 매출액의 1.5%에서 최대 3,500만 유로(약 497억원) 또는 전세계 매출액의 7%까지 벌금 부과

- (미국 및 우방국 진영) 미국과 영국은 AI Safety 파트너십을 준비 중이며 미국 사이버보안·인프라 보안국과 영국 국립사이버보안센터는 전 세계적으로 합의된 최초의 지침인 AI 안전·신뢰 지침인 'Guidelines for Secure AI System Development'을 공동으로 개발
- (중국) 자체적인 세계 최고 수준의 AI 안전 규제로 높은 진입장벽을 구축하여 중국 내수 AI 시장 보호를 도모

시사점

- 안전·신뢰 AI 환경 구축을 위한 적극적인 대응 부재 시 향후 AI 핵심 분야(AI 모델링, 컴퓨팅 등)과 같이 경쟁우위를 확보하지 못할 위험이 존재
- AI 거버넌스 및 관련 기술 시장의 급성장은 관련 솔루션을 개발하고 구현하는 기업에 성장 기회를 제공
- AI 정책 방향 설정 시 기술의 기회-위험 간 균형 유지 필요

Ⅲ 경쟁력 분석

분석 개요

3장에서는 AI 주요 학회들의 프로시딩(proceeding) 논문을 분석하여 국가별 안전·신뢰 AI 경쟁력 분석

※ AI 분야는 특허출원 비중이 낮고, 최신 동향을 분석하고자 프로시딩 논문 중심으로 분석

(데이터 수집) 안전·신뢰 AI 관련 키워드 도출 및 해당 키워드를 통한 AI 주요 학회별·연도별 안전·신뢰 AI 프로시딩 수집

- 산·학·연 전문가 인터뷰를 통한 안전·신뢰 AI 관련 키워드를 도출하였고, 이를 바탕으로 데이터 추출 논리 및 키워드 쿼리를 〈표 2〉와 같이 설정

〈표 2〉 안전·신뢰 AI 논문 추출을 위한 데이터 추출 논리·키워드 쿼리

변수명	데이터 추출 논리	데이터 추출 키워드 쿼리
AI 신뢰성	AI and 신뢰성	(“artificial intelligence” or “machine learning” or “deep learning”) and (“privacy protection” or “personal information protection” or “personal data protection” or “rigorous” or “safe” or “robust” or “secure” or “explainable” or “interpretable” or “transparent” or “fair” or “unbiased” or “impartial” or “data provenance” or “predictable” or “nonharmful” or “comprehensive” or “traceable” or “explicable” or “nondiscriminatory” or “inclusive”)
AI 개인정보보호	AI and 개인정보보호	(“artificial intelligence” or “machine learning” or “deep learning”) and (“privacy protection” or “personal information protection” or “personal data protection” or “data provenance”)
AI 견고성	AI and 견고성	(“artificial intelligence” or “machine learning” or “deep learning”) and (“rigorous” or “safe” or “robust” or “secure” or “predictable” or “nonharmful”)
AI 설명가능성	AI and 설명가능성	(“artificial intelligence” or “machine learning” or “deep learning”) and (“explainable” or “interpretable” or “transparent” or “comprehensive” or “traceable” or “explicable”)
AI 공정성	AI and 공정성	(“artificial intelligence” or “machine learning” or “deep learning”) and (“fair” or “unbiased” or “impartial” or “nondiscriminatory” or “inclusive”)

- 〈표 2〉를 바탕으로 AI 주요 학회별·연도별 안전·신뢰 AI 프로시딩을 수집
  - AI 주요 학회들 중 논문 DB, DB access 등의 상황을 고려하여 ICML(International Conference on Machine Learning, ICLR(International Conference on Learning Representations), NeurIPS(Neural Information Processing Systems) 학회의 최근 3개년 안전·신뢰 AI 프로시딩을 수집

※ 논문 저자 소속기관(affiliation)의 국가명이 명시되지 않은 경우 분석 대상에서 제외

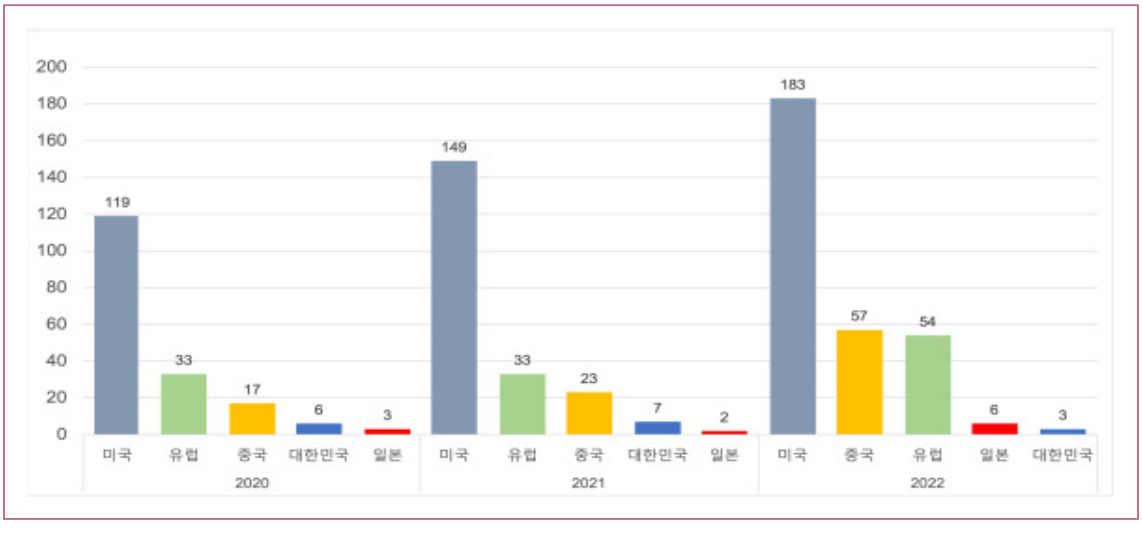
〈표 3〉 AI 주요 학회별 안전·신뢰 AI 프로시딩 수집 결과

	ICML			ICLR			NeurIPS		
연도	안전신뢰 논문 수	학회논문 전체 수	비중	안전신뢰 논문 수	학회논문 전체 수	비중	안전신뢰 논문 수	학회논문 전체 수	비중
2020	65	1084	6.00%	22	686	3.21%	87	1898	4.58%
2021	48	1183	4.06%	38	860	4.42%	118	2333	5.06%
2022	61	1233	4.95%	63	1191	5.29%	154	2830	5.44%
합계	174	3500	4.97%	123	2737	4.49%	359	7061	5.08%

(데이터 분석 결과) 수집한 결과를 국가별·연도별로 분석

※ 국가 분류는 학회 프로시딩의 저자의 소속기관의 국가를 기준으로 분류

※ 만약 저자별 소속 기관이 상이할 시 제1저자 소속기관의 국가를 기준으로 국가를 판정



[그림 8] 국가별·연도별 게재된 AI 주요 학회 안전·신뢰 AI 프로시딩 수

(시사점) ’22년도까지 AI 주요 학회의 안전·신뢰 AI 관련 프로시딩 비중은 4~5% 내외이고, 해당 데이터 기준 미국이 다른 주요국 대비 월등히 많은 수의 프로시딩을 게재 중

- AI 분야 권위자들을 중심으로 안전·신뢰 AI 기술의 중요도가 최근에 강화되고 있기에 ’22년까지는 비교적 낮은 비중으로 안전·신뢰 AI 프로시딩이 발간되었을 것으로 추정
- 프로시딩 수 분석 기준 한국과 일본은 주요국 대비 안전·신뢰 AI 프로시딩 수가 현저히 부족한 측면
- 안전·신뢰 AI가 기술과 제도를 중심으로 클러스터화 되고 있는 동향을 고려할 때 한국은 향후 관련 기술패권 경쟁에서 뒤처질 위험이 존재

(한계점) 분야 특성 및 데이터 한계로 인한 경쟁력 분석의 한계가 존재

- 타 국가전략기술 분야 대비 AI 분야는 기술의 빠른 변동성으로 논문/특허가 큰 의미를 갖지 못하는 측면이 존재

  ※ 논문 게재/특허 출원 기간 내 기술표준/유효성이 저하될 수 있어 연구결과를 바로 공개하는 경우가 다수

- 이에 그나마 정량분석이 가능한 AI 주요 학회 프로시딩을 중심으로 국가 경쟁력을 간접적으로 측정·분석

  ※ AI 주요 학회들은 학술지 대비 프로세스가 짧기 때문에 프록시 데이터로 적합한 측면

- 따라서 타 분야의 학술지 논문 분석과 해당 결과를 동일시하기에는 한계가 존재

Ⅳ 정책 제언

(대응 방향) AI 규제강도에 따른 정책수단의 장단점과 글로벌 추세를 고려하여 기민하게 대응할 필요

- 현재 주요국들의 AI 안전성 및 신뢰성 확보를 위한 정책 방향과 규제 강도는 확정되지 않았으나, 최근 정부 중심 규제로의 정책 기조 변화가 감지되고 있어 여러 가능성에 대비하여 시나리오를 마련하여 대비 필요

〈표 4〉 규제 강도에 따른 안전·신뢰 AI 정책 수단 및 수단별 장·단점

규제 강도	안전·신뢰 AI 정책 수단	내용	수단별 장·단점
가장 낮음	전문가 중심의 자율 규제 권고	관련 기술 개발자/서비스 제공자가 전문가 집단의 참여/협력을 통해 자발적으로 통제를 추구하는 방식	• (장점) 자율기반의 관련 기술/생태계 발전 촉진 • (단점) 전문가 위원회 구성에 대한 객관성 결여 위험, 예측 못한 문제 발생 시 법적 책임 모호
다소 낮음	인증체계 구축·운영	정부가 이해관계자에게 자발적인 인증 획득을 유도하여 관련 분야 위험 대응 체계를 구축할 수 있도록 유인하는 방식	• (장점) 물리적 대응 체계 구축·운영 가능 • (단점) 인증 주체 지정, 인증 수준 및 절차에 대한 이해관계자 합의에 오랜 시간 소요
다소 높음	개인적 권리 설정	정부가 이용자들에게 알고리즘 구조 및 결과에 대한 설명을 기술개발자/서비스 제공자들에게 요구할 수 있는 권한을 부여하는 방식	• (장점) 인공지능 기술/제품/서비스 개발자가 딥러닝의 인과관계 모호성을 최소화하여 제품/서비스를 개발할 수 있도록 유도 • (단점) 기업 부담 가중, 설명 수준에 대한 규정 등을 설정하는데 상당한 시간 소요
매우 높음	직접적 행정 규제 설정	정부가 기술개발/서비스 제공 가능 영역을 설정하고, 이외 영역에 대해서는 법적 규제를 가하는 방식	• (장점) 역기능 피해 및 분쟁 최소화 • (단점) 진화하고 있는 인공지능 기술을 법으로 명확하게 규제하기 어려운 측면

(국제협력) 글로벌 안전·신뢰 AI 환경 조성에 주도적 역할을 수행하되 일정 수준의 국내 산업·기업 보호 조치 마련 필요

- 글로벌 안전·신뢰 AI 논의 및 관련 규제/기술 표준 설정에 적극적으로 참여

- 정부 간 협력, 국제기구, 민간 협력 등 다양한 방식을 활용하여 AI 시스템의 안전성, 공정성, 투명성, 설명 가능성, 인간 통제 등 다양한 분야에 대한 국제적 기준*을 마련하고, 이를 준수하도록 유도

  * (예) ISO/IEC JTC 1/SC 42: WG3 분과에서 AI 신뢰성(trustworthiness) 관련 표준 제정

- 정보 공유 및 협력을 위한 플랫폼 구축을 통해 AI 안전성·신뢰성 관련 모범 사례 공유, 연구 결과 및 규제 접근 방식 교환 등을 추진

  ※ 한국 정부는 미국 및 우방국 진영에서 글로벌 AI 규제 형성 방안을 논의하고 있고('23년 11월 AI Safety Summit 참여, '24년 상반기 AI의 안전한 활용을 위한 미니 정상회의를 영국과 공동 주최 예정), 향후에도 해당 진영을 유지할 것으로 예상

- 편향 탐지/완화 기술, AI 모델의 설명 가능성 및 조절가능성 기술, 프라이버시 보호 기술 등 중요한 AI 안전 문제 해결을 위한 기술을 연구하는 공동연구 이니셔티브 개발

● 안전·신뢰 AI 관련 국제 협력을 추진함과 동시에 국내 AI 산업·기업 보호 조치를 마련하는 투트랙(two-track) 전략 검토 필요

- 향후 확정될 글로벌 스탠다드에 맞추어 안전·신뢰 AI 정책을 추진하면서도 국내 기업에 지나치게 무겁지 않은 부담을 주지 않도록 균형을 유지할 필요

- 나아가 법적 규제를 마련하여 AI 시스템의 안전성, 공정성, 투명성, 설명 가능성, 인간 통제 등 다양한 분야에 대한 국내 기준을 마련하고, 이를 준수하도록 의무화하여 내수 시장을 보호하고, 국내 AI 산업과 기업의 경쟁력을 강화할 필요

  ※ (예) 클라우드 산업 분야의 클라우드보안인증제도(CSAP)와 같이 내수 시장에 별도 기준을 설정·적용하여 국내 AI 산업 및 기업 보호를 도모하는 방안도 검토 필요

(정부) AI 안전성·신뢰성 확보를 위한 제도적 안전망 구축과 관련 핵심기술 R&D 지원 추진

● 글로벌 AI 규제 동향, 국내 AI 기술·산업 특수성을 고려하여 혁신과 안전 사이의 균형을 맞추는 한국형 AI 규제 프레임워크 설계

● AI에 의한 피해 발생시 이에 대한 사업자 책임 부여 및 이용자 구제 방안 마련

※ AI 위험성 평가, AI 보험제도, AI 피해로부터 이용자 사전 보호(사전고시) 및 사후 보상 제도 등 수립

● Responsible AI 분야 핵심기술 확보를 지향하는 R&D 지원

- 설명 가능한 AI 기술, 편향성 완화 기술, 개인정보보호 기술, AI 오용 탐지·방지 기술, AI 동작/구조 안전성 검증 기술 등의 개발을 지원하여 향후 EU GDPR, AI Act 등 세계 최고 수준의 AI 신뢰성 체계와 호환 가능한 안전하고 신뢰성 있는 AI 모델 및 서비스 제공 기술 확보

- AI 연구자, 개발자, 이용자, 윤리, 사회 과학, 법률 전문가 간 협력을 장려하여 AI가 사회적으로 미치는 영향을 파악하고, 관련 이해관계자 간 협업을 지원

(기업) 적극적인 AI 신뢰성 확보 기술 투자, 포괄적 위험 평가, 전략적 파트너십 형성 등을 통해 안전·신뢰 AI 분야 경쟁력 확보

- 한국어 데이터 및 사용 환경에 특화된 Responsible AI Toolkit의 개발·배포에 적극적으로 투자
  - AI 시스템의 편견, 편향성, 투명성 및 데이터 개인정보보호를 해결하기 위한 기술, 안전 프로토콜 및 메커니즘을 구현하는 기술 확보 및 배포
- 외부 전문가 중심의 안전·신뢰 AI 위원회를 구축하여 AI 시스템/서비스 배포 전 관련 잠재적 위험을 식별하고 완화하는 포괄적 위험 평가 프로세스를 구축
- 업계 내 전략적 파트너십을 형성하여 AI 안전성·신뢰성 관련 모범 사례를 공유하고, AI 기술에 대한 안전 및 신뢰성 평가 기술을 공동으로 개발
- 정부 기관과 적극적으로 협력하여 관련 사례·경험·전문 지식을 공유하며 AI 안전 및 신뢰성 관련 문제 해결 정책 수립 과정에 참여

(대학) 안전·신뢰 AI 관련 전문 교육 프로그램 개발, 기존 커리큘럼 확대 등을 통해 관련 인재 양성

- 기업 및 정부와 협력하여 AI 안전성과 신뢰성에 초점을 맞춘 전문 프로그램 개발
  - AI 개발 및 배포, 관련 규제 분야 전문가 대상의 AI 안전성·신뢰성 제고를 위한 맞춤형 교육 프로그램 개발
  - AI로부터 파생되는 윤리적 문제를 식별하고, 이를 예방/방지/수습할 수 있는 안전 조치를 구현하는 교육 프로그램 개발
- AI 윤리, 안전, 거버넌스를 컴퓨터 과학 및 엔지니어링 필수 커리큘럼에 반영하여 AI 안전·신뢰 의식을 갖춘 전문 인재 양성
- 연구 파트너십, 인턴십, 지식 교환 프로그램을 지원하여 학계와 업계 간 협력 촉진

참 고 문 헌

- 고학수 외 (2021). 유럽 연합 인공지능법안의 개요 및 대응 방안.
- 구본진 (2022). 디지털 전환의 미래사회 위험이슈 및 정책적 대응 방향: 인공지능을 중심으로. 기술혁신연구, 30(1), 1-20.
- 이중희 (2023). 중국의〈생성형 인공지능 서비스 관리 잠정 방법〉에 대한 분석: 배경과 쟁점
- 소프트웨어정책연구소 (2019). 일본의 인공지능 전략 동향 : AI 전략 2019.
- 소프트웨어정책연구소 (2021). 유럽(EU)의 인공지능 윤리 정책 현황과 시사점 : 원칙에서 실천으로.
- 소프트웨어정책연구소 (2022). 인공지능 신뢰체계 정립방안 연구.
- 소프트웨어정책연구소 (2023). 글로벌 AI 신뢰성 정책 동향 연구.
- 한국무역협회 (2021). 중국 인공지능 산업 동향과 시사점: 중국의 AI 굴기와 성공전략.
- 한국지능정보사회진흥원 (2022). 최신 AI 불확실성 정량화 동향 및 시사점.
- 한국지능정보사회진흥원(2023). 중국 인터넷 정보 서비스의 심층합성 관리 규정.
- 한국지식재산연구원 (2022). 일본 지적재산 추진계획 2022의 주요내용과 시사점.
- Acemoglu, D., Autor, D., Hazell, J., & Restrepo, P. (2022). Artificial intelligence and jobs: Evidence from online vacancies. Journal of Labor Economics, 40(S1), S293-S340.
- Behrens, V., & Trunschke, M. (2020). Industry 4.0 related innovation and firm growth. ZEW-Centre for European Economic Research Discussion Paper, (20-070).
- Bessen, J., & Righi, C. (2019). Shocking technology: what happens when firms make large IT investments?.
- Gartner (2022). What's New in Artificial Intelligence from the 2022 Gartner Hype Cycle.
- Hatzius, J. (2023). The Potentially Large Effects of Artificial Intelligence on Economic Growth (Briggs/Kodnani). Goldman Sachs.
- Markets and Markets (2022). AI Market by Offering Technology, Business Functio, Vertical, and Region - Global Forecast to 2030.
- Research and Markets (2023). Artificial Intelligence (AI) - Global Strategic Business Report.
- Stanford University (2023). HAI AI Index Report 2023.

[웹페이지]

- https://aws.amazon.com/ko/machine-learning/responsible-ai/resources/
- https://cloud.google.com/responsible-ai?hl=ko
- https://responsibleaitoolbox.ai/introducing-responsible-ai-dashboard/#dashboard-components
- https://www.ibm.com/kr-ko/products/watsonx-governance
- https://ai.meta.com/responsible-ai/
- https://www.gminsights.com/industry-analysis/ai-governance-market
- https://www.marketsandmarkets.com/Market-Reports/explainable-ai-market-47650132.html
- https://www.maximizemarketresearch.com/market-report/data-privacy-software-market/145730/

필자 소개

❑ **구본진**

- 한국기술교육대학교 산업경영학부 조교수
- 前 한국과학기술기획평가원 전략기술정책단 부연구위원
- 041-560-1608 / bonkoo@koreatech.ac.kr

[KISTEP 브리프 발간 현황]

발간호 (발행일)	제목	저자 및 소속	비고
112 (24.01.08.)	무기발광 디스플레이	진영현·오세미 (KISTEP)	기술주권
113 (24.01.12.)	2022년 우리나라와 주요국의 연구개발투자 현황	이새롬·한웅용 (KISTEP)	통계분석
114 (24.01.12.)	2022년 우리나라와 주요국의 연구개발인력 현황	이새롬·한웅용 (KISTEP)	통계분석
– (24.01.22.)	KISTEP Think 2024, 10대 과학기술혁신정책 아젠다	강현규·이민정 (KISTEP)	이슈페이퍼 (제357호)
– (24.01.25.)	국가연구개발 성과분석 프레임워크 개발 및 적용	박재민·문해주·김수민·박서현 (건국대학교) 이호규(고려대학교) 강승규(한국조달연구원)	이슈페이퍼 (제358호)
115 (24.01.25.)	세계경제포럼(WEF) Global Risks 2024 주요 내용 및 시사점	이미화 (KISTEP)	혁신정책
116 (24.01.25.)	기후변화와 기후 지구공학	정의진·임현 (KISTEP)	미래예측
117 (24.01.26.)	단백질 구조예측 및 디자인	전수진·한민규 (KISTEP)	기술동향
– (24.01.29.)	신약개발 분야 정부 R&D 현황과 효율성 제고 방안	송창현·엄익천(KISTEP) 김순남(국가신약개발사업단) 이원희(유한양행)	이슈페이퍼 (제359호)
– (24.01.31.)	반도체 분야 정부연구개발투자의 효과성 분석과 개선방안	김준희·엄익천(KISTEP) 오승환(경상국립대학교) 전주경(한국특허기술진흥원)	이슈페이퍼 (제360호)
118 (24.02.01.)	인공지능이 변화시킬 미래 연구수행 모습	이상남 (KISTEP)	미래예측
119 (24.02.13.)	EU 인공지능(AI) 규제 현황과 시사점	강진원·김혜나 (KISTEP)	혁신정책
– (24.02.15.)	'생성형 인공지능' 시대의 10대 미래유망기술	박창현 (KISTEP)	이슈페이퍼 (제361호)

발간호 (발행일)	제목	저자 및 소속	비고
– (24.02.29.)	과학기술 전공자 취업 현황 분석 및 시사점	이정재·박수빈·이원홍 (KISTEP)	이슈페이퍼 (제362호)
120 (24.03.07.)	국가R&D 국외수혜정보 보고 제도 주요 내용 및 시사점	황인영·정정규 (KISTEP)	혁신정책
121 (24.03.19.)	2022년 한국의 과학기술논문 발표 및 피인용 현황	김용희 (KISTEP)	통계분석
122 (24.03.20.)	브렉시트(Brexit) 이후 영국의 과학기술 동향	임현지·이가원·홍미영 (KISTEP)	기술동향
123 (24.03.27.)	'과학기술협력에 관한 격년 보고서(2022년 NSTC ISTC)'의 이행사항 점검 결과와 시사점	도계훈·강진원·김혜나 (KISTEP)	혁신정책
124 (24.04.01.)	호라이즌 유럽(Horizon Europe)의 연구데이터 정책과 시사점	이민정·송창현 (KISTEP)	혁신정책
125 (24.04.01.)	안전·신뢰 AI	구본진 (한국기술교육대학교)	기술주권