GLOBAL DATA MARKET TRENDS

Monthly Trends Report

데이터 중심 AI, 글로벌 현황

2025년 7월

K data
한국데이터산업진흥원

일러두기

◆ 해당 자료는 한국데이터산업진흥원에서 발간한 “월간 동향 리포트” 보고서입니다

◆ 급변하는 데이터 산업의 흐름을 쉽게 파악할 수 있도록 시의성 및 시사성이 높은 해외 데이터 산업 관련 뉴스 정보, 주요국별 데이터 산업의 현황 자료 및 ‘Data-Centric AI: A Systematic Review of Methods’, ‘Challenges and Future Directions, Data-Driven Breakthroughs and Future Directions in AI Infrastructure: A Comprehensive Review’, ‘2025 Technology and Innovation Report Inclusive Artificial Intelligence for Development’ 보고서의 요약 정리를 제공합니다.

◆ 주요 국가별 데이터 산업 뉴스 및 전문지를 일정기간 모니터링한 후, 월별로 글로벌 주요 국가의 핵심 정책 이슈와 이달의 비즈니스, 기술 이슈 등을 요약하여 제공합니다.

- ◯ 정책/전략: 글로벌 데이터 산업 육성 및 규제 정책 동향
- ◯ 비즈니스: 글로벌 데이터 비즈니스 이슈
- ◯ 기 술: 글로벌 데이터 산업의 기술 부문 이슈
- ◯ 사 례: 글로벌 데이터 기업 및 기관 사례 분석
- ◯ 보 고 서: 글로벌 데이터 산업 주요 이슈 관련 보고서 요약 정리

◆ 본 자료는 진흥원 홈페이지(http://www.kdata.or.kr)를 통해 매월 발간되며, 최신 데이터 산업 정보를 소개합니다.

Contents

Contents

1 데이터 중심 AI, 글로벌 현황

요약

▷ 국가, 산업 분야별 현황 및 기업/기관 사례

- 미국은 AI 연구 인프라와 공공 데이터 개방을 통한 민간 혁신을 촉진하고 있으며, EU는 데이터 규제와 공유 플랫폼을 기반으로 신뢰성 높은 AI 생태계를 조성
- 데이터 중심 AI 비즈니스 모델은 데이터 수집·가공 및 제공 서비스, 데이터 마켓플레이스, AI 통합 개발 도구·플랫폼, 내부 데이터와 생성형 AI를 연계하는 모델 등이 존재하며, 산업별 특성과 데이터를 반영해 설계된 산업 특화 AI(Vertical AI)가 주목받고 있음
- AI 성능은 데이터 수집과 전처리에 좌우되며 AI가 학습 가능한 형태로 가공하는 기술이 중요한 만큼 데이터를 어노테이션하거나 시뮬레이션 기반으로 생성하는 기술이 발전하는 추세
- 헬스케어 산업에서는 의료 영상 데이터에 기반한 AI 진단 모델, 제조 산업에서는 고장 예측 및 유지보수 최적화를 위한 산업 장비 데이터 분석 AI를 활용

▷ 심층 분석 – 데이터 중심 AI 현황

○ 데이터 중심 AI는 모델 개선이 아닌 데이터 품질 향상 중심의 접근 필요

- 데이터 중심 AI는 모델 아키텍처나 알고리즘 개선보다 학습 데이터의 품질과 구조적 설계에 중점을 두는 방식으로 접근
- 노이즈 제거, 정밀한 라벨링, 데이터 증강 등의 전처리 과정이 핵심 기술로 작용하고 있으며 대표성 있는 샘플 확보와 구조화된 데이터셋 구축이 중요

○ 기존 AI 접근은 모델 구조 개선에 과도하게 집중되어 있는 상황

- 아키텍처나 알고리즘 고도화에 편중된 발전 경로를 따라왔으며, 이로 인해 성능 향상의 한계가 반복적으로 노출

○ 데이터 중심 AI를 갖추기 위해 데이터 수집, 라벨링, 정리, 증강, 유지 관리, 자동화 등 프로세스별 주요 방법론 및 주요 기술(합성 데이터, 연합 학습, 데이터 사이트 패러다임 등) 확보 필요

○ 또한, 국가별로 책임 있는 AI를 위한 데이터 구축 정책을 마련

- 다수의 국가가 AI 등장 이전에 데이터 정책을 수립했으나, 현 시점에는 데이터 정책의 업데이트가 필요한 상황이며, AI 생태계와 데이터 기반 혁신이 필요함

○ 향후 과제로는 데이터 중심 접근의 체계화 및 국제적 협력 기반 마련

- 선진국과 개발도상국 간의 AI 역량 격차를 해소하기 위해서는 글로벌 데이터 공유 기준 마련, AI 훈련용 공공 데이터의 국제 공동 활용 등 필요

Contents

Ⅱ 데이터산업 분야별 현황

1 정책/전략

1) 미국

개요

▷ **AI 연구 인프라와 공공 데이터 개방을 통한 민간 혁신 촉진**

○ 미국은 AI 분야에서 민간 주도의 기술 발전을 강점으로 하며, 이를 뒷받침하기 위해 연방정부가 데이터 개방, 연구 인프라 제공, 규범 개발에 집중하고 있음

○ 'National AI Initiative Act(2020)[1)]' 이후 AI 전략은 데이터 접근성과 활용성 강화에 초점을 두고 전개되고 있으며, 특히 AI의 발전은 양질의 데이터에 달려 있다는 인식 하에 공공 데이터 인프라 확대와 AI 학습 자원 제공을 중점적으로 추진 중임

주요 내용

정책/제도	내용
National AI Initiative Act(2020)[1]	– 2021년 「National AI Initiative Act」에 따라 설계된 연구 인프라로, AI 연구자에게 연산 자원, 공공·민간 데이터, 툴킷, 교육 자원 등을 통합 제공 – GPU 기반 슈퍼컴퓨터와 고품질 정부 데이터를 연구자에게 제공함으로써, 대기업 중심의 연산 자원 독점을 완화하고 학계·스타트업의 AI 경쟁력 확보를 지원
Executive Order 14158 – Establishing and Implementing the President's Department of Government Efficiency[2]	– Executive Order 14158은 연방 정부 내 중복된 기능을 통합하고, 데이터 기반 정책 결정을 확대하기 위해 부처 간 데이터 공유 및 통합 체계를 강화하도록 규정함 – 특히 데이터 무결성과 접근성을 제고하고, 분석 역량 강화를 위한 공공 데이터의 적극적 활용을 핵심 과제로 제시함
Executive Order 14179 - Removing Barriers to American Leadership in Artificial Intelligence[3]	– Executive Order 14179는 AI 인프라 구축을 저해하는 규제를 완화하고, 고성능 컴퓨팅 및 데이터 자원 접근성을 높이기 위한 국가 차원의 액션 플랜 수립을 지시함 – Stargate 프로젝트는 이에 발맞춰 민간 주도로 초대형 AI 데이터센터와 GPU 인프라를 미국 전역에 구축하는 이니셔티브로 추진됨
Federal Data Strategy (FDS)[4]	– Federal Data Strategy(FDS)는 2024년까지 바이든 행정부가 연방 정부의 데이터 거버넌스 강화와 공공 데이터 활용 확대를 위해 추진한 전략임 – FDS는 모든 연방 기관에 데이터 책임자를 지정하고, 데이터의 수집·관리·활용에 관한 10대 원칙과 40개의 실행과제를 제시함

1) 공식 명칭은 'National Artificial Intelligence Initiative Act of 2020'이며, 미국 연방정부 차원의 AI 전략법으로 AI 연구개발, 교육, 데이터 및 연산 자원 접근성 확대 등을 종합적으로 추진하기 위한 법률

2) EU

개요

▷ **데이터 규제와 공유 플랫폼을 기반으로 신뢰성 높은 AI 생태계 조성**

○ EU는 AI 개발과 활용에 있어 투명성, 신뢰성을 핵심 가치로 설정하고, 이를 실현하기 위한 데이터 기반 법·제도 및 인프라 구축에 집중하고 있음

○ 특히 데이터 중심 AI 구현을 위해 Data Governance Act, AI Act 등의 법제와 함께 산업별 데이터 스페이스(Data Spaces)를 병행하여 추진 중이며, 이는 EU가 디지털 주권을 확보하고, 미국·중국과의 기술 격차를 줄이기 위한 전략의 일환임

주요 내용

정책/제도	내용
Data Governance Act (DGA, 2023)[5]	– 공공·민간·자발적 데이터 공유를 촉진하기 위한 법적 틀로, 데이터 중개기관(Data Intermediary)을 도입 – 의료, 농업, 제조 분야 등 민감한 공공 데이터를 신뢰 기반으로 공유할 수 있도록 관리 체계를 법제화 – DGA는 비개인정보뿐 아니라 익명화된 개인정보도 포함해, AI 학습데이터로 활용 가능한 범위를 확대함
European Data Spaces (EDS)[6]	– 2025년까지 총 14개 분야에 걸쳐 공공 및 민간 데이터를 연계해 산업별 공동 데이터 플랫폼을 구축 – 각 스페이스는 데이터 상호운용성, 메타데이터 표준, 접근 권한 설정 등을 포함하며, AI 활용을 위한 고품질 데이터셋 제공에 중점 – AI 개발자는 특정 산업용으로 인증된 신뢰할 수 있는 데이터셋을 활용 가능함
Digital Europe Programme (2021~2027)[7]	– 데이터, 인공지능, 사이버보안, 고성능 컴퓨팅(HPC) 등 핵심 디지털 역량 확보를 위해 총 1.76억 유로 규모로 편성된 EU의 전략적 재정 지원 프로그램 – 데이터 중심 AI 강화를 위해 공공·민간 데이터 인프라 구축, 데이터 공간(Data Spaces), 클라우드 플랫폼 등을 우선 지원 – AI 모델 학습 및 실증을 위한 테스트베드, 신뢰 가능한 AI 도입 촉진 사업, 중소기업 대상 AI·데이터 기술 역량 강화 프로젝트 등도 포함되어 있음
AI Act[8]	– AI 시스템을 위험 기반으로 분류(4단계: 금지, 고위험, 일반, 최소위험)하고, 고위험 AI는 데이터의 품질, 정확성, 편향 여부에 대한 엄격한 요구 부과 – 데이터 측면에서는 '학습·검증 데이터의 문서화', '편향 점검 결과 보고', '데이터 출처 명시' 등이 필수 – 공공안전, 금융, 의료 등 분야의 AI는 학습데이터 기준에 따라 사용 허용 여부가 달라짐

3) 한국

개요

▷ **AI 산업 고도화와 일상화를 위한 데이터 기반 정책 강화**

○ 한국은 데이터 중심 AI 생태계 조성을 위해 정책적 기반을 다각도로 구축하고 있음

○ 범정부 차원의 국가 전략 수립과 산업 적용 확대는 물론, AI 신뢰성 확보를 위한 윤리·보안 체계 정비도 병행되고 있으며, 고품질 데이터의 구축·개방 확대, AI 학습 데이터의 활용성 제고, 민감정보에 대한 특례 제도 도입 등은 데이터 기반 AI 기술의 실질적 확산을 유도하고 있음

주요 내용

정책/제도	내용
신뢰할 수 있는 인공지능 실현전략[9]	– '사람 중심' 신뢰 가능한 AI 실현을 위한 데이터 기반 정책 방향을 제시 – AI 학습용 데이터의 품질·신뢰성 확보, 편향 제거, 윤리적 설계 가이드라인 수립 등 데이터 활용의 책임성과 안전성을 강조함 – AI 윤리 기준 정립 및 데이터 가공·활용 전 주기의 검증 체계 도입도 주요 과제로 포함됨
AI 일상화 및 산업 고도화 계획[10]	– AI 학습용 데이터를 민간이 쉽게 활용할 수 있도록 품질 인증 및 데이터 가공 지원 체계를 강화함 – 산업별 초거대 AI 개발을 위한 데이터 패키지 제공, 클라우드 인프라 연계 등을 통해 AI 생태계 고도화를 추진함 – 특히 의료, 금융, 제조 등 중점 분야를 중심으로 AI 실증·상용화를 유도하고, 데이터 중심 R&D 연계도 강조됨
국가 AI 전략 정책방향[11]	– 2024년 9월 발표된 대통령 주재 국가AI위원회 전략으로, 데이터 중심 AI 생태계 구축을 핵심 방향으로 설정함 – AI 학습용 고품질 데이터 클러스터 조성, 데이터 개방 및 거래 활성화, 산업별 데이터 연계 체계 강화 등이 핵심임 – 민간 데이터 활용성 제고와 동시에 AI 신뢰성 확보를 위한 데이터 품질 기준 마련도 병행 추진됨
원본 데이터 활용 허용 AI 특례(예정)[12]	– 개인정보보호위원회가 2025년 업무계획에서 제시한 제도 기반으로, AI 개발을 위한 민감정보 활용을 허용하는 특례 규정을 포함함 – 의료, 에너지 등 고위험 분야에서 원본 데이터의 가명처리 및 재사용을 가능하게 하여 AI 개발과 데이터 혁신의 균형을 모색함 – 동시에 활용 조건으로 안전성, 비식별성, 설명가능성 기준도 강화되어 데이터 중심 AI의 윤리적 기반 마련에 기여함

2 비즈니스

1) 데이터 중심 AI 비즈니스 모델 예시

주요 내용

▷ **데이터 수집·가공 및 제공 서비스**[13]

○ 수요기업이 AI 학습용으로 활용할 수 있도록 고품질 데이터를 수집·정제하여 제공하는 비즈니스 모델

- 주요 수집 데이터는 비정형 데이터(텍스트, 음성, 이미지 등)로, 대규모 모델 학습에 적합한 구조로 가공함
- Scale AI은 이 분야의 대표 기업으로, 다양한 산업 분야에서 기업의 AI 모델 개발, 개선 및 배포를 지원하는 종합적인 데이터 플랫폼으로 발전하고 있음

▷ **데이터 마켓플레이스**[14]

○ 데이터 공유 및 협업을 지원하는 온라인 상점으로, 데이터 공급자와 수요자를 연결하여 생태계 참여자에게 안전한 환경에서 데이터 및 관련 서비스를 사고 팔 수 있는 기회를 제공함

- 시장 및 비즈니스 조사, 인구 통계 데이터, 마케팅 및 광고 데이터, 과학 데이터 등 다양한 정보를 얻을 수 있으며, 기업들은 마켓플레이스를 활용하여 제품 및 서비스를 제공하고 수익을 창출할 수 있음
- AWS Data Exchange, Snowflake Marketplace, Oracle Data Marketplace가 대표적 사례

▷ **AI 통합 개발 도구·플랫폼**[15]

○ 기업이 수집한 데이터를 기반으로 맞춤형 AI 모델 및 서비스를 구축할 수 있는 개발 도구를 제공

- AI 서비스를 개발 및 배포·관리하고 업무에 적용할 수 있는 End-to-End AI 통합 개발 플랫폼으로, 모델 구축, 배포, 확장 등 포괄적인 환경을 원하는 개발자와 데이터 사이언티스트가 활용할 수 있는 플랫폼임
- 대표 기업 및 서비스로는 Google Vertex AI, IBM Watsonx Orchestrate 등이 있음

▷ **내부 데이터와 생성형 AI 연계**[16]

○ 기업 내부 데이터와 생성형 AI를 결합해 데이터 분석, 업무 자동화, 의사 결정 등을 지원하는 비즈니스 모델

- 문장 생성 시 데이터베이스에서 검증된 정보를 검색하고 반영하는 RAG 기술이 적용되었으며, Chat GPT Enterprise, 솔트룩스 LUXIA 등이 대표 사례

▷ **또한, 산업별 특성과 데이터를 반영해 설계된 Vertical AI(산업 특화 AI)가 주목받고 있으며, 수요기업은 이를 통해 각 분야의 고유한 요구사항을 효과적으로 해결하고 업무 효율성을 극대화하고 있음**[17]

2) 데이터 유통 및 거래 플랫폼 동향

주요 내용

▷ 데이터 마켓플레이스의 부상[18]

○ 데이터 자체를 상품처럼 사고팔 수 있는 데이터 마켓플레이스(Data Marketplace)가 본격적으로 시장에 자리 잡고 있음

- AWS Data Exchange, Snowflake Marketplace, SAP Data Marketplace 등 주요 글로벌 플랫폼은 민간·공공 데이터를 등록하고 이용자가 조건에 따라 구입할 수 있도록 함
- 제공되는 데이터 유형은 실시간 거래 데이터, IoT 센서 데이터, 소비자 행태 데이터, 위성영상 등으로 다양함
- 특히 머신러닝 학습에 특화된 정형·비정형 데이터셋의 거래가 활발하며, 고품질 메타데이터와 API 제공이 핵심 경쟁력임

▷ 유통 구조 및 수익 모델의 다변화

○ 대부분의 데이터 거래 플랫폼은 구독형 모델(Subscription), 사용량 기반 요금제(Pay-per-use), 수익 공유 모델(Revenue Sharing) 기반의 복합 수익모델을 채택함

- 일부 플랫폼은 거래 전 데이터 미리보기, 품질 검증 기능, 신뢰 등급을 도입해 구매 신뢰성을 제고하고 있음

구분	내용
구독형 모델 (subscription)	- 일정 기간동안 동일한 유형의 데이터를 지속적으로 사용하는 기업에게 유리하며, Snowflake는 실제 사용량을 기준으로 비용을 정산함
사용량 기반 요금제 (pay-per-use)	- 초기 진입장벽을 낮춰 다양한 산업군의 중소기업이 접근 가능하게 함
수익 공유 모델 (revenue sharing)	- AWS Data Exchange는 데이터 제공자와 수익을 분배하는 구조를 운영

▷ 국내 정부·민간 주도 플랫폼 확산[19]

○ 정부는 데이터 기반 인공지능 생태계 조성을 위해 공공·민간 데이터의 유통 기반 강화와 데이터 거래 활성화, AI 학습 목적의 특화 거래 지원 등을 정책 방향으로 제시하고 있음

- '18년 정부 주도로 빅데이터 플랫폼 사업이 시작된 이후, AI 데이터 구축 사업과 데이터거래소가 출범하면서 데이터 경제 활성화의 기반이 마련됐으며, 산림·환경·제조 등 다양한 산업 분야에서 데이터 수집, 유통, 활용이 확대되고 있음
- 민간·학계 또한 KT의 주도하에 고려대, 두산디지털이노베이션, 중앙일보 등과 함께 'K-데이터 얼라이언스'를 결성하여 한국어 데이터를 중심으로 한 AI 학습 생태계 조성을 추진하고 있음

3) 시장 트렌드 및 투자 흐름

주요 내용

▷ 데이터 중심 AI 시장의 고속 성장

○ 초거대 AI, 생성형 AI의 확산은 데이터 수요를 급증시켜, 데이터 기반 서비스 시장을 AI 가치사슬의 핵심 영역으로 부상시킴

- 특히 RAG(Retrieval-Augmented Generation) 기반 생성형 AI는 기업 내부 데이터와 외부 지식 데이터의 결합을 요구해, 데이터 확보가 경쟁력의 핵심이 되고 있음
- 이에 따라 데이터 품질관리, 분산데이터 연계, 도메인 특화 데이터 수집 등 전방위적인 데이터 서비스 수요가 증가 중임

▷ 투자자 관심의 전환: 데이터와 인프라 중심[20]

○ 글로벌 투자자들은 기존 AI 모델 개발보다는 데이터 수집·정제·유통 인프라와 데이터 품질관리 솔루션에 집중하고 있음

- 2024년 전 세계 민간 AI 투자 규모는 약 1,004억 달러로, 이 중 20~30%가량이 데이터센터, 유통 플랫폼, 보안·통합 솔루션 등 데이터 인프라 분야에 집중된 것으로 추정됨[2)]
- 미국의 Boldstart Ventures는 데이터 보안과 문서 기반 RAG 기술 기업에 2.5억 달러 규모의 펀드를 조성했으며, Kleiner Perkins 등도 의료·법률 데이터 기반 생성형 AI 스타트업에 집중 투자 중임

▷ 산업별 특화 데이터 수요 증가

○ 헬스케어, 금융, 제조, 교육 등 분야별로 도메인 특화 AI에 맞춤형 데이터셋을 구축하려는 움직임이 활발함

- 루닛, 뷰노 등은 고해상도 의료영상 데이터를 수집·분류하여 AI 진단 정확도를 획기적으로 높였고, 이는 투자 유치에도 큰 역할을 함
- 또한 금융권에서는 이상 거래 탐지(FDS), 고객 행동 예측을 위한 실시간 로그 데이터 기반 분석 서비스에 대한 수요가 증가하고 있음

▷ 공공·민간 협력 통한 데이터 자산화 확대

○ 데이터 유통 시장의 신뢰도 확보를 위해 정부-민간 협력 기반의 품질 인증, 거래 중개 플랫폼, 데이터 신뢰 프레임워크 마련이 필요함

- EU는 공공과 민간 데이터를 분리·연계하는 European Data Spaces를 중심으로 산업별 데이터 기반 AI 활성화 전략을 강화하고 있음

2) CB Insights 기반 Crowdfund Insider 보도와 Newmark의 2024년 데이터센터 투자 보고서에 따르면 2024년 전 세계 민간 AI 기업의 자금 조달은 1,004억 달러에 달했으며, 미국 내 데이터센터 건설에만 315억 달러가 투자되었음

3 기술

1) 데이터 수집 및 가공 기술

개요

▷ **AI 성능을 좌우하는 데이터 수집과 전처리의 핵심 기반**

○ 데이터 중심 AI의 출발점은 다양한 원천으로부터 수집된 신뢰할 수 있는 데이터이며, 이를 AI가 학습 가능한 형태로 가공하는 기술이 중요함

○ 센서, 로그, IoT, 웹 등 다양한 소스에서 데이터를 확보하고, 이를 자동화된 방식으로 어노테이션하거나 시뮬레이션 기반으로 생성하는 기술이 발전 중임

주요 내용

▷ **실시간 데이터 수집 기술**

○ AI 학습을 위한 데이터는 IoT, 센서, 웹, 로그 등 다양한 경로에서 실시간으로 수집되고 있음

- 제조 현장에서는 설비 로그 데이터, 헬스케어 분야에서는 생체 센서 데이터, 물류에서는 위치·경로 정보가 실시간 수집되며, 대규모 데이터를 신속히 처리하기 위해 Kafka, Flink 등 스트리밍 처리 기술이 병행 활용됨

▷ **자동 어노테이션 및 라벨링 도구**[21]

○ 수작업에 의존하던 데이터 라벨링 과정을 자동화하거나 반자동화하는 기술이 급속히 확산되고 있음

- Hugging Face의 'AutoTrain', 라벨링툴 'CVAT', 'Label Studio' 등이 주요 사례이며, 음성·이미지·텍스트 전 영역에 적용 가능함

▷ **합성데이터(Synthetic Data) 생성 기술**[22]

○ 실제 데이터가 부족하거나 개인정보 문제로 활용이 어려운 경우, 시뮬레이션이나 알고리즘을 통해 유사 데이터를 생성함

- 자율주행 분야는 가상환경 기반의 주행 시나리오 생성 및 시뮬레이션, 의료 분야는 MRI·CT 등 영상 생성, 금융은 가상 거래 시뮬레이션에 활용되며, GAN, Variational Autoencoder, 물리기반 렌더링(PBR) 방식 등이 있음

▷ **센서 퓨전(Sensor Fusion) 기반 데이터 통합 기술**[23]

○ 서로 다른 종류의 센서(LiDAR, 카메라, 레이더, IMU 등)로부터 수집된 데이터를 통합하여 더 정확하고 신뢰도 높은 정보를 생성하는 기술

- 각 센서는 고유의 장점과 한계를 지니며, 이를 상호 보완함으로써 개별 센서보다 더 나은 인식 결과를 도출하고 데이터를 종합적으로 해석할 수 있음

2) 데이터 품질 관리 기술

개요

▷ **AI 성능 향상을 위한 데이터 품질 기술의 고도화**

○ AI의 정확성과 신뢰성 확보를 위해 데이터 품질을 진단·정제·통합·관리하는 기술이 핵심 기반으로 부상하고 있으며, 정량화된 품질 지표와 자동화 도구를 통해 오류와 불균형을 사전에 식별하고 개선하여 산업별 표준을 준수하도록 데이터셋을 최적화함

주요 내용

▷ **데이터 품질 측정을 위한 정량화 지표**[24]

○ 정확성(Accuracy), 완전성(Completeness), 일관성(Consistency), 최신성(Timeliness)을 기본 4대 지표로 통용

- 공공기관은 데이터 품질관리지표(DQI)를 활용하고, 기업은 Snowflake, Ataccama, Talend 등 툴을 도입해 품질을 자동 진단함

▷ **학습용 데이터의 품질 고도화 기준**[25]

○ AI 학습용 데이터에는 클래스 비율, 라벨 신뢰도, 이상치 유무, 중복률 등 특수 품질 기준이 적용됨

- 학습 데이터의 불균형(Class Imbalance)을 개선하기 위해 오버샘플링, 언더샘플링, 클래스 가중치 조정 등의 기법이 사용되며, 라벨링 품질은 교차검증, 노이즈 제거 알고리즘, 다중 주석자 합의(Consensus)[3] 방식으로 관리됨

▷ **데이터셋 정제 및 통합 기술**

○ 학습에 활용되는 대규모 데이터셋의 오류 제거와 통합을 위한 기술이 활용되고 있음

- 이상치 탐지, 중복 제거, 누락값 보완 등을 통해 원천 데이터의 품질을 정제함
- 스키마 정렬, 속성 일치 등 통합 기술을 통해 출처가 다양한 데이터의 정합성과 일관성을 확보함

▷ **데이터 품질 진단 및 자동화 도구**[26]

○ 데이터의 품질을 실시간 진단하고 리포트를 제공하는 자동화 도구들이 상용화되고 있음

- Great Expectations, Soda.io, Deequ, OpenDQ 등은 품질 규칙 정의, 검증, 경고 등의 기능을 제공함
- 이를 통해 데이터 파이프라인 내 품질 검증을 자동화하고, 이상 발생 시 조기 대응이 가능함

▷ **산업별 품질 표준 준수 지원 기술**

○ 금융, 의료, 제조 등 산업별 품질 기준을 자동 점검하는 기술이 확산되고 있음

- 각 산업에 요구되는 정확도, 형식, 단위, 데이터 항목의 존재 여부 등을 규칙 기반 또는 ML 기반으로 검사함
- 이를 통해 AI 학습용 데이터가 실사용 요구조건을 만족하도록 관리할 수 있음

3) 데이터 중심 학습 기술

개요

▷ 모델 중심에서 데이터 중심으로 전환되는 AI 학습 전략

○ 데이터 중심 AI는 모델 구조보다 학습에 투입되는 데이터의 품질과 구성 방식이 성능의 핵심 변수임

○ 이러한 접근법은 데이터를 선별하거나 구조화하는 방식으로 학습 효율과 결과의 정밀도를 높이고 있으며, 특히 RAG 기반 생성형 AI처럼 데이터 검색과 연결이 핵심인 구조에서는 데이터 중심 학습 전략이 필수적으로 적용됨

주요 내용

▷ 커리큘럼 학습(Curriculum Learning)[27]

○ AI 모델이 학습할 데이터를 난이도나 품질에 따라 점진적으로 투입하여 학습 효율을 높이는 전략임

- 기초적·일반적인 샘플을 먼저 학습한 후, 복잡하거나 예외적인 샘플을 나중에 학습시키는 방식으로 구성됨
- 의료, 법률, 과학문서 등 고신뢰 분야에서 특히 성능 향상 효과가 큼

▷ 능동 학습(Active Learning)[28]

○ 모델이 학습에 가장 도움이 되는 데이터를 스스로 선택하여 학습하는 방식으로, 데이터 비용과 시간 낭비를 최소화할 수 있음

- 모델이 혼동하거나 확신이 낮은 데이터를 선별해 반복 학습하거나 라벨링을 요청함
- 특히 대규모 비정형 데이터셋에서 주목받으며, 실시간 인간-모델 상호작용 기반 반응형 학습 구조로 발전 중임

▷ RAG(Retrieval-Augmented Generation) 기반 학습[29]

○ 생성형 AI가 단순한 사전학습 지식이 아닌, 실제 데이터베이스·문서·사내 지식 등을 실시간으로 불러와 응답을 생성하는 구조임

- 문서를 벡터화하고 검색시스템과 연동하여 사용자 질의와 가장 관련 높은 데이터를 추출하고, 이를 기반으로 LLM이 응답을 생성하며, 고객상담, 보험약관, 사내 매뉴얼, 기술문서 등 텍스트 기반의 정답형 AI 구현에 활용됨

▷ 데이터 중심 MLOps 기술[30]

○ AI 모델 성능 개선을 데이터 품질 및 흐름 중심으로 접근하기 위한 자동화된 학습 운영 체계

- 기존 MLOps가 모델의 배포와 반복 학습 자동화에 집중했다면, 데이터 중심 MLOps는 데이터 버전, 변경 이력, 품질 메타데이터 등을 통합 관리하여 데이터 기반 학습 최적화를 가능하게 함

3) 여러 annotator의 라벨 중 다수결 또는 평균을 통해 최종 라벨을 결정하는 방식으로, 라벨의 신뢰도와 품질을 높이기 위한 기법

Contents

기관/기업 사례

III

1 데이터 중심 AI 업종별 활용 사례

1) 헬스케어 산업[31]

개요

▷ **루닛(Lunit): 의료영상 데이터 기반 AI 진단**

○ 의료 영상은 고해상도 이미지, 정형화된 진단 정보, 민감한 개인정보가 결합된 고부가가치 데이터로서, 정확한 라벨과 임상지식 기반의 학습이 필요한 대표적인 데이터 중심 AI 분야임

○ 영상 판독 AI는 의사의 진단 보조 역할을 하며, 빠르고 정확한 이상징후 감지를 통해 진단 효율성과 의료 접근성을 동시에 개선함

주요 내용

▷ **활용 사례**

○ 루닛은 흉부 엑스레이 및 유방촬영 영상 데이터를 수집·정제하여 폐암, 유방암, 결핵 등 질환을 조기 진단하는 AI 솔루션 'Lunit INSIGHT'를 개발함

○ 약 300만 건 이상의 의료영상 데이터를 확보하기 위해 국내외 병원, 공공의료기관과 협업 및 공동 연구를 진행하였으며, 방사선 전문의가 참여한 고정밀 어노테이션을 통해 학습 데이터를 고도화함

○ 학습 시 데이터 불균형 해소를 위해 정상 사례(음성)의 비중을 조절하고, 소수 사례 증강을 통해 모델 성능을 향상시킴

○ 의학적으로 중요한 병변 부위를 강조하기 위해 Grad-CAM 기반 시각화 기술을 적용하여 의료진이 직접 검토 가능한 형태로 결과를 제공함

▷ **성과**

○ 루닛 AI 솔루션은 2023년 기준 전 세계 40여 개국, 600개 병원에 도입되었으며, 실제 임상에서 민감도 95% 이상, 판독 속도 30% 단축 효과가 보고됨

○ 유럽 CE 인증, 미국 FDA 승인, 국내 신의료기술 인증 등 다국적 인증을 획득하며 의료기기로 상용화 및 글로벌 시장 진출 기반을 마련함

○ AI 판독 신뢰성 향상을 통해 의료 사각지대에서의 조기 진단률 증가에 기여함

2) 제조 산업[32]

개요

▷ **Siemens: 고장 예측을 위한 산업 장비 데이터 기반 AI 시스템**

○ Siemens는 대형 산업 설비에서 발생하는 시계열 센서 데이터를 활용해 설비 고장 예측과 유지보수 최적화를 위한 AI 기반 시스템을 구축함

○ 복잡한 산업 기기와 분산된 공정 데이터를 통합 분석함으로써, 데이터 기반의 설비 의사결정 자동화를 목표로 함

주요 내용

▷ **활용 사례**

○ Siemens는 공장 현장에 설치된 장비에서 온도, 진동, 전류, 압력, 회전속도 등 다양한 센서 데이터를 초 단위로 수집하여, 장비 상태의 정상/이상 패턴을 사전 정의함

○ 데이터를 수집하는 'Industrial Edge' 장치는 공정 현장에서 전처리를 수행하고, 사전에 필터링된 데이터를 Siemens Cloud Platform으로 전송함

○ 수집된 데이터는 시계열 이상 탐지 알고리즘, 결함 분류 모델, 고장 가능성 예측 회귀모델 등을 통해 분석되며, 이를 통해 실시간 상태 모니터링, 이상 경고, 유지보수 일정 제안이 이루어짐

○ 고장이 드물게 발생하는 특성상, 합성 고장 데이터(Synthetic Failure Logs)를 시뮬레이션을 통해 생성하여 모델 성능을 보완하고 있음

○ 또한, AI 모델은 공정별 상황에 맞춰 커스터마이징되며, 특정 공정에 최적화된 학습 셋으로 지속적인 재학습이 진행됨

▷ **성과**

○ 일부 고객 공정에서는 설비 고장 전 탐지 정확도 92% 이상, 예측 정비 시스템 도입으로 장비 다운타임 40% 이상 감소

○ 연간 유지보수 비용 수백만 유로 절감, 설비 교체 주기 최적화, 예비 부품 재고 최소화 등의 효과가 확인됨

○ 현재 Siemens MindSphere는 미국 GE, 독일 BASF, 프랑스 Renault 등 주요 글로벌 제조업체의 공정에 적용되어, 유럽 스마트팩토리 전략의 핵심 기반으로 자리잡고 있음

○ 데이터 중심 AI 기술을 통해 '사후 정비 → 사전 예방'으로 전환한 대표적 성공 사례로 평가받음

3) 금융 산업[33]

개요

▷ **JPMorgan Chase: 트랜잭션 로그 기반 AI 사기 거래 탐지 시스템**

○ JPMorgan Chase는 세계 최대 금융기관 중 하나로, 매일 수억 건의 실시간 트랜잭션 데이터를 처리하고 방대한 로그 데이터를 활용하여 정교한 AI 기반 이상거래 탐지 시스템(FDS)을 구축함

○ 기존의 규칙 기반 시스템의 한계를 넘어, 고객별 거래 패턴의 미묘한 변화도 실시간으로 포착할 수 있는 머신러닝 기반 탐지 체계로 전환하고 있음

주요 내용

▷ **활용 사례**

○ JPMorgan은 자체 개발한 AI 시스템 'Hawk AI'를 통해 카드 결제, ATM 출금, 송금, 로그인 등 다양한 트랜잭션 로그를 통합 분석함

○ 트랜잭션별 시계열 벡터를 생성하고, 정상 행동의 패턴을 모델링한 후, 새로운 거래가 기존 패턴과 어느 정도 괴리가 있는지를 이상 점수로 환산함

○ 지도학습 기반 분류모델과 더불어, 실제 사기 거래 데이터가 부족한 특성을 고려하여 반지도 학습(semi-supervised learning), AutoEncoder 기반 이상치 탐지 등 다양한 방식이 혼합되어 사용됨

○ 탐지 결과에 대한 신뢰 확보를 위해 SHAP, LIME 기반의 설명가능 AI 기법(XAI)을 적용, 금융 규제기관 및 내부 감사 부서에 대한 설명 책임을 충족함

○ 또한 다크웹에서 수집된 사기 유형 데이터, 글로벌 사기 IP 목록, 디바이스 ID 등 외부 위협 인텔리전스 데이터를 내부 트랜잭션 로그와 융합하여 탐지 정밀도를 높임

▷ **성과**

○ 탐지 정확도는 기존의 규칙 기반 시스템 대비 25% 향상, 탐지 속도는 평균 수분에서 30초 이내로 단축

○ 연간 수천 건의 고위험 사기 거래를 사전에 차단하며, 수백억원 이상의 손실 방지 효과 달성

○ 2024년 기준, Hawk AI 시스템은 미국 금융감독기관이 제시한 AI 리스크 관리 프레임워크에 부합하는 모범 사례로 인증받았으며, 내부적으로는 고객 CS 응대 부담이 감소하고, AI가 사전 대응한 탐지 성공률이 87% 이상으로 나타나는 등 실질적인 성과를 입증함

Contents

심층분석 – 데이터 중심 AI 현황

Ⅳ

1 데이터 중심 AI 현황

▷ 데이터 중심 AI는 모델 개선이 아닌 데이터 품질 향상 중심의 접근 필요[34]

○ 데이터 중심 AI는 모델 아키텍처나 알고리즘 개선보다 학습 데이터의 품질과 구조적 설계에 중점을 두는 방식으로 접근

- AI 성능은 데이터의 정확도, 다양성, 대표성 등에 따라 민감하게 반응하며, 이는 단순한 모델 구조보다 실질적 성능 향상에 직접적인 기여를 하는 항목
- 고도화된 모델이라 하더라도 부정확하거나 편향된 데이터를 학습할 경우, 오히려 왜곡된 결과를 초래할 수 있어 데이터의 중요성이 더욱 부각되고 있음

○ 노이즈 제거, 정밀한 라벨링, 데이터 증강 등의 전처리 과정이 핵심 기술로 작용

- 데이터 정제 기술은 오탐률·노이즈를 제거하여 학습 효율을 높이며, 모델의 학습 수렴 속도를 향상시킴
- 라벨링 자동화 기술과 증강 기법은 부족하거나 편향된 데이터셋을 보완하여, 다양한 상황에 대한 학습과 대응력을 강화하는 핵심 기술임

○ 대표성 있는 샘플 확보와 구조화된 데이터셋 구축이 중요

- 현실 세계의 특성을 반영한 균형 잡힌 데이터셋은 AI의 일반화 능력과 신뢰도를 높이는 핵심 자산
- 잘 설계된 데이터 구조는 모델 학습의 방향성과 성능 안정성 확보에 결정적 역할을 수행

▷ 주요 AI 성과는 데이터 접근성과 활용에 기반[35]

○ GPT 시리즈, ImageNet 등은 대규모 고품질 데이터셋 접근이 발전을 주도한 사례

- GPT 시리즈는 획기적인 아키텍처보다도 광범위하고 정제된 텍스트 데이터 학습을 통해 성능을 비약적으로 향상
- ImageNet의 대규모 라벨링 데이터는 시각 인식 분야에서 딥러닝의 전환점을 만들며 데이터의 역할을 증명

○ 고성능 컴퓨팅과 알고리즘은 데이터 중심 학습을 보조

- 컴퓨팅 자원과 알고리즘은 데이터의 대량 처리와 고속 학습을 가능하게 하지만, 본질적인 성능 향상의 중심은 아닌 것으로 나타났으며 이는, 데이터 중심 전략이 알고리즘 중심 접근보다 더 안정적이고 재현 가능한 혁신을 이끌 수 있음을 시사하고 있음

▷ AI 성능 향상의 관건은 고품질 데이터 확보 전략[36]

○ 알고리즘 구조나 연산 성능의 한계로 인해 AI에서 데이터의 중요성이 더욱 부각되고 있음

- 고도화된 알고리즘도 낮은 품질의 데이터에서는 한계를 드러내며, 데이터가 AI 발전을 저해하는 주요 요인이 되는 상황

- 무어의 법칙[4] 폐기와 GPU 처리 성능의 포화는 연산 자원의 획기적 발전이 점점 어려워지고 있음을 시사

○ 차세대 AI 혁신은 데이터 접근성과 활용 방식에 좌우

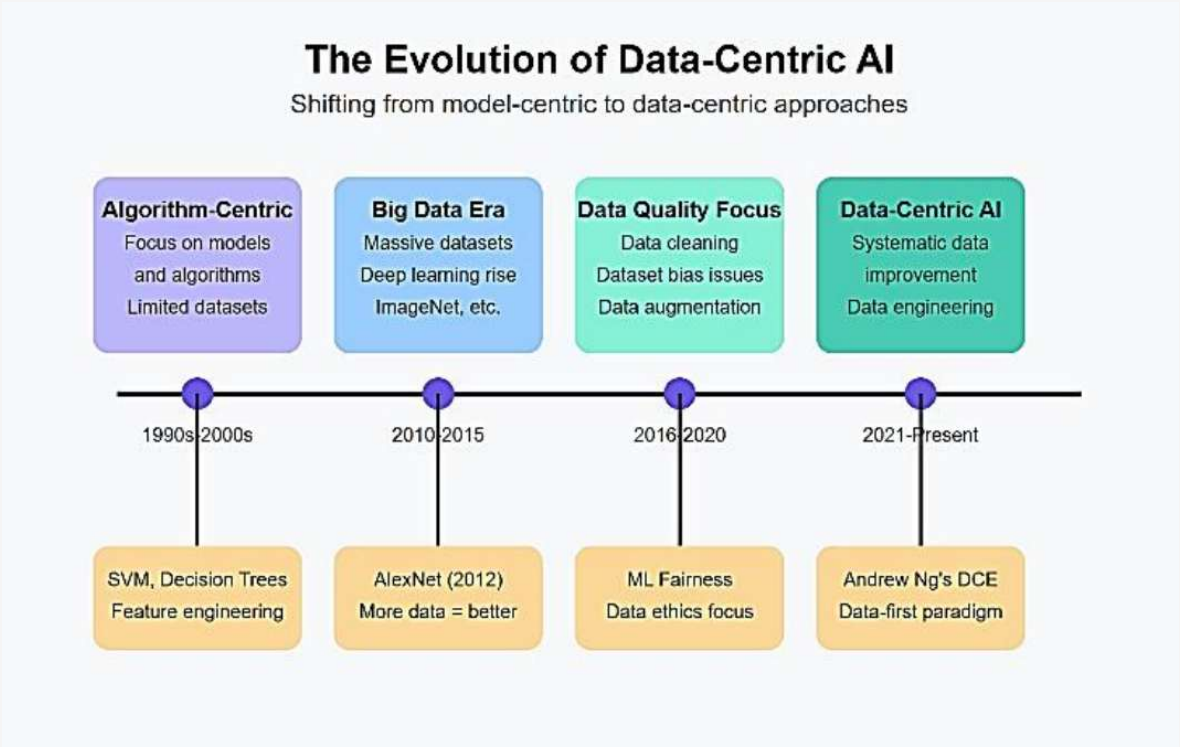
- 향후 AI의 발전은 더 크고 정제된 데이터셋의 확보, 공유, 윤리적 활용 기반 위에서 가능할 것으로 예상
- 데이터 중심 접근은 신뢰성 있는 AI 시스템 구축을 위한 핵심 전략으로 자리 잡고 있으며, 이를 위한 정책적·기술적 기반 마련이 필요

▷ 기존 AI 접근은 모델 구조 개선에 과도하게 집중[37]

○ 기존 AI 연구는 모델 아키텍처나 알고리즘 고도화에 편중된 발전 경로를 따라왔으며, 이로 인해 성능 향상의 한계가 반복적으로 노출

- 최신 LLM의 발전 역시 근본적으로는 데이터의 확장과 품질 확보에 의존하는 경향을 보이며, 알고리즘 혁신만으로는 성능 개선이 제한적
- 고도화된 모델 구조가 복잡성과 연산 자원 부담을 동반하면서 실용성과 효율성의 균형 문제를 야기

< 데이터 중심 AI의 발전 - 모델 중심에서 데이터 중심 접근 방식으로의 전환 >



* 출처 : Data-Centric AI: A Systematic Review of Methods, Challenges and Future Directions

4) 마이크로칩 기술의 발전 속도에 관한 일종의 법칙으로 마이크로칩에 저장할 수 있는 데이터 분량이 18개월마다 두 배씩 증가한다는 법칙. 인텔의 창립자 고든 무어가 1965년에 발견한 관찰 결과로 캘리포니아 공과대학교의 교수 카버 미드가 발견자 고든 무어의 이름을 따 명명했다. 무어의 법칙은 시간이 지남에 따라 수정 및 보완되고 있었으나, 2016년 2월에 반도체 업계가 경제성을 이유로 포기를 선언하면서 법칙이 폐기됨

▷ 데이터 중심 AI 발전 현황

No.	시기	주요 내용	기타 주요 사항
1	모델 중심 (1990년대-2000년대)	알고리즘과 모델에 집중, 제한된 데이터셋 사용	SVM, 결정 트리, 특성 공학
2	빅 데이터 시대 (2010-2015)	대규모 데이터셋 사용, 딥러닝의 부상, ImageNet 등	AlexNet (2012), 더 많은 데이터 = 더 나은 결과
3	데이터 품질 중심 (2016-2020)	데이터 정제, 데이터셋 편향 문제 해결, 데이터 증강	ML 공정성, 데이터 윤리 중심
4	데이터 중심 AI (2021-현재)	체계적인 데이터 개선, 데이터 엔지니어링	Andrew Ng의 DCE (Data-Centric Engineering), 데이터 우선 패러다임

▷ 국가 간 AI 격차는 데이터, 인프라, 인재의 복합적 불균형에서 발생하며 모델 중심 접근은 한계가 존재[38]

○ 국가 간 혹은 기관 간 AI 역량 격차는 알고리즘 자체보다 고품질 데이터 접근성, 컴퓨팅 인프라, 전문 인력 확보 여부에 따라 심화되는 구조

- 고성능 GPU와 대규모 데이터셋을 보유한 일부 국가 및 기관이 AI 기술을 독점하거나 선점하는 경향
- 알고리즘은 공개되더라도 실제 적용 및 활용을 위한 기반 여건이 갖춰지지 않으면 AI 활용도는 제한적

○ 알고리즘 소스는 공개되는 경우가 많아 확산이 빠른 반면, 양질의 데이터셋 구축은 국가마다 여건 차이가 커 격차를 심화시키는 요소로 작용

- 데이터 인프라, 수집 체계, 가공 역량 등이 AI 경쟁력의 결정적 차별점으로 부각
- 데이터 접근성과 품질 격차가 인재 및 인프라 격차와 함께 복합적으로 작용하여 국가 간 AI 역량 불균형을 초래

▷ 데이터 접근성 및 품질은 AI 성능의 핵심 격차 요인이며, 인프라 및 인재를 통한 AI 개발·실험은 지속 확산의 조건

○ 학습 데이터의 양과 질은 AI 성능의 가장 직접적인 기반이며, 다양한 언어, 지역, 산업 데이터를 보유한 국가가 우위를 점함

- 공공 데이터의 개방 수준과 표준화 여부가 민간 활용 가능성을 결정
- 저개발국가나 중소기관은 적절한 데이터셋을 구축·관리할 자원이 부족하여 개발 역량 확보에 제약을 받음

○ 클라우드 기반 연산 환경과 병렬처리 기술 등 고성능 컴퓨팅 자원의 접근성이 AI 개발과 실험의 필수 요건

- 국가나 기관의 재정 능력에 따라 AI 학습 및 추론에 필요한 자원 비용의 투자 여력이 결정됨
- 이와 함께 데이터 엔지니어, 모델 설계자, 정책 전문가 등 다양한 인재군 확보가 기술 내재화의 관건

▷ 데이터 중심 AI의 필요성과 대응 방향 제시

○ 데이터 중심 AI 심층분석 파트는 모델 중심의 기존 접근 방식에서 벗어나, 데이터 품질·활용·인프라에 초점을 맞춘 데이터 중심 AI의 필요성을 조명하고 정책·기술 측면에서의 전환 방향을 제시함

- AI 성능의 한계를 극복하고, 공정하고 신뢰 가능한 AI 구현을 위해 데이터 기반 접근의 확대 필요
- 알고리즘의 고도화만으로 해결되지 않는 AI 편향성, 신뢰성, 일반화 문제를 해결할 수 있는 대안 탐색

▷ 기술·정책·국가간 협력 관점에서의 다양한 요소를 제시

○ 본 보고서는 데이터 중심 AI의 기술적 핵심 요소, 주요 정책 과제, 국가 간 격차를 해소하기 위한 협력 방향을 총체적으로 파악

- 데이터 품질 개선, 학습용 데이터셋 최적화, 모델 효율화 전략 등 기술 접근 정리
- 데이터 인프라 확대, 공공 데이터 개방 및 표준화, 국제적 데이터 공유 및 격차 해소 전략 등 정책 과제를 제시

2 데이터 중심 AI 방법론[39]

▷ A. 데이터 수집

○ AI 시스템의 성공 여부는 데이터가 얼마나 좋은지에 달려 있음. 데이터 중심 AI(DCAI[5])에서 데이터 수집은 다양하고 대표적이며 편향성이 제거된 대규모 데이터 세트를 만드는 것에서 시작함

- 데이터 수집은 과거에는 조직이 명확한 기준 없이 임의로 데이터를 수집하는 것을 의미했으나, DCAI 패러다임에서는 인구 통계, 환경, 엣지 케이스 등 모델이 특정 집단에 치우치지 않도록 편향을 방지하기 위해 특정 그룹을 샘플링하는 것을 뜻함
- 능동적 학습 접근법은 예측의 불확실성이 크거나 고유한 특성을 지닌 데이터를 전략적으로 선별하는 방식으로 활용되며, 이를 통해 데이터 활용의 효율성을 높일 수 있음

○ DCAI에서 데이터 수집은 단순 집계, 기회주의적 데이터 집계가 아닌 체계적인 설계 프로세스임

- 적대적 예제(Adversarial Example)를 샘플링하는 기법은 표준 샘플링에서 놓치기 쉽고 분류하기 어려운 데이터 세트의 품질을 높일 수 있으며, 주로 작업 환경의 실제 복잡성을 대표하는 데이터 세트를 만듦
- 잘못된 샘플링은 의료, 금융, 자율 시스템과 같은 고위험 영역에서 상당한 위험을 초래할 수 있으므로 조직은 데이터 세트의 특성과 수집 절차가 포함된 공식적인 데이터 보고서를 작성하고, 데이터 세트의 사용 목적과 방식을 문서화함

○ 데이터 중심 AI에서 수집된 데이터의 품질은 학습 데이터의 정합성, 신뢰성, 다양성에 직접적으로 영향을 미치며, 이는 곧 모델 성능을 좌우하는 핵심 요소라는 인식이 강함

- 부정확하거나 편향된 데이터는 고도화된 모델의 성능을 저하시킬 수 있음
- 특히 대규모 언어 모델(LLM)의 경우, 데이터 품질이 모델의 일반화 능력과 윤리성에 직접적 영향을 미침

○ 또한, 개인정보 보호, 분산된 데이터 소유 구조로 인해 발생하는 데이터 수집의 어려움을 해결하기 위한 방안으로 연합 학습(Federated Learning) 기술이 주목받고 있음

▷ B. 데이터 라벨링

○ 데이터를 수집한 후, 일관되고 정확한 라벨을 붙이는 것이 중요하며, 라벨 노이즈는 모델 성능 저하를 유발하는 주요 오류 요인임

- 라벨 오류는 모델의 결정 경계(Decision boundaries)와 일반화(Generalization)에 큰 영향을 미치며, 예시 중 일부만 잘못된 라벨을 붙여도 성능에 큰 차이를 일으킬 수 있음
- 이상탐지 기반 정제, 논리적 일관성 검증, 자동화된 전처리 기술 등이 활용되며, 정제 수준에 따라 학습 성능과 학습 시간, 추론 결과의 정확도가 유의미하게 변화

5) Data-Centric AI: 모델 중심 AI에서 벗어나 데이터 품질과 데이터의 설계에 초점을 맞춘 AI 개발 접근 방식

○ 수동 라벨링은 일반적으로 많이 사용하는 방법이나, 비용이 많이 들고 주관적인 해석, 주석자의 피로, 불명확한 라벨링 지침으로 인해 신뢰할 수 없는 경우가 많음

- 최신 DCAI에서는 다양한 품질 관리 방법을 사용하여 라벨의 정확성을 보장하며 명확한 라벨링 방향, 중복 라벨링에 대한 다수결 투표, 어려운 사례에 대한 전문가 평가 등을 통해 품질을 향상시킴

○ 휴리스틱 라벨링[6] 함수와 지식 기반을 인코딩하여 대규모로 라벨링하고 활용함에 따라 약한 감독(weak supervision)과 같은 프로그램적 라벨링 접근법이 증가하고 있음

- 통계적 방법(예: 스노클)을 활용하여 라벨 오류를 추정하고 영향을 줄이는 방식이 사용됨

▷ C. 데이터 정제

○ 우리가 사용하는 데이터 세트에는 라벨이 잘못 지정된 예시, 손상된 파일, 중복, 이상값 등 실제 사용에서 발생하는 노이즈가 포함될 가능성이 높음

- 데이터 정제에서는 노이즈 변수를 체계적으로 처리하여 학습에 영향을 줄 수 있는 인위적인 훈련 사례에 의존하지 않도록 해야 함

○ 또한, 대규모 언어 모델의 효율성은 양질의 사전학습 데이터와 설계된 입력 프롬프트에 의해 좌우됨

- 데이터 정제 과정에서 오류, 편향, 중복 등을 사전에 제거해 불필요한 계산 최소화
- 프롬프트 설계는 모델의 응답 일관성과 정확도에 직접 영향을 미치므로, 문맥 유도와 목적 적합성을 고려한 구조가 요구됨

○ 최근에는 오류 수정보다 오류 탐지가 강조되고 있음

- CleanLab과 같은 도구는 교차 검증 기법을 사용하여 라벨이 불확실한 부분을 식별
- 통계적 포일, 클러스터링 불일치 감지, 유사성 기반 중복 제거는 데이터 세트의 품질을 더욱 향상시킴

○ DCAI에서 데이터 정제는 일회성 작업이 아닌 지속적이고 주기적인 프로세스로 보는 시각이 존재

- 모델은 최신 모델을 사용하여 시간이 지남에 따라 개선될 것이며, 새로운 오류 사례를 발견할 경우 오래된 데이터 세트를 검토하고 정리해야 함. 이에, 효과적인 정리는 단순히 범주를 수정하는 것 이상을 수행
- 혼란을 야기하는 미묘한 실수를 제거하고 평가 데이터 세트가 배포의 기본 컨텍스트와 관련이 있는지 확인

○ 데이터 정제가 충분하지 않으면 모델이 편향된 학습 경로 또는, 비정상적인 패턴에 의존해 학습할 위험이 존재하며, 이는 실제 환경에서 치명적인 오류로 이어질 수 있음

▷ D. 데이터 증강

○ 데이터 증강은 데이터 세트가 작거나 불균형하거나 다양할 때 필요하며, 모델이나 데이터 세트 구조를 변경하지 않고 원본 데이터에 변형을 가하여 데이터를 보강할 수 있음

6) 명확한 규칙이나 경험적 지식에 기반하여 데이터를 라벨링하는 방식. 예를 들어, 특정 조건을 만족하는 데이터를 긍정적으로, 다른 조건을 만족하는 데이터를 부정적으로 라벨링하는 방식

- 일반적인 컴퓨터 비전에서 데이터 증강은 회전, 스케일링, 뒤집기, 색상 지터링(color jittering) 또는 자르기를 통해 추가 데이터를 합성하는 방식으로 이루어짐
- 자동 증강, 랜드 증강 또는 적대적 증강과 같은 최근 연구에서는 증강 전략을 학습하고 검증 성능을 극대화하기 위해 최상의 변환 집합을 선택함
- 텍스트의 경우 단어 교체, 문장 순서 재배열, 요약/확장 등을 통해 문장 다양성을 확보하며, 이미지·음성·IoT 센서 데이터에서도 회전, 필터 적용, 노이즈 삽입 등의 방식으로 활용됨

○ NLP에서 데이터 증강은 원래 문장의 의미를 유지하는 것이 어려운 작업임

- 효과가 입증된 데이터 증강 전략은 역번역, 동의어 대체, 의역, 트랜스포머 기반 생성, 적대적 교란 텍스트 등

○ 합성 데이터 생성의 경우 생성 모델(GAN, VAE, 확산 모델) 기반 완전히 새로운 예시 생성 가능

- 생성된 데이터는 실제 데이터의 대체물로 사용될 수 있으며, 특히 개인 정보 보호가 중요한 분야(예: 금융, 의료)에서 유용함
- 생성된 데이터를 통해 실제 데이터에 대한 접근이 제한되는 상황에서도 학습을 지원할 수 있음

▷ E. 데이터 유지 관리

○ 데이터 유지 관리는 중요하지만 종종 과소평가되는 부분

- 실제 환경은 분포의 변화, 인구의 변화, 센서 드리프트, 적의 변화 등 동적인 특성을 지니고 있기 때문에 정적인 데이터 세트만으로는 모델 성능의 안전성을 보장할 수 없음
- 프로덕션 시스템은 원격 측정 데이터를 분석하여 학습 데이터에 잘 나타나지 않거나 존재하지 않은 새로운 패턴을 발견하고, 이를 재학습/개발 데이터 수집에 반영해야 함

○ 데이터 세트 버전 관리는 중요하나, 기존에는 모델에 대한 버전 관리의 견고성이 부족

- 데이터 세트 변경 및 버전 관리를 체계적으로 추적할 수 있는 최신 시스템이 필요하며, 재현성, 가독성, ML 데이터 파이프라인 관리가 가능하도록 설계 필요
- 데이터 품질을 유지하는 것은 단기적인 작업이 아니라 지속적이고 주기적인 활동이어야 하며, 모델이 환경 변화에 계속 적응할 수 있도록 해야 함

▷ F. 데이터 중심 AI의 자동화 전략

○ AI의 자동화는 DCAI 방법의 규모를 확장할 수 있는 원동력이 됨. 초기 단계의 수동 데이터 정리, 라벨링, 증강 방식은 더 많은 사전 작업과 수동 데이터 배포를 의미했음

- 데이터 규모의 급증과 반복 작업의 필요성이 증가함에 따라 데이터 자동화 시스템이 지속적으로 개발되고 있음
- 기초 모델이나 대규모 언어 모델(LLM)을 사용하는 자동 라벨링 방법은 라벨이 부족한 상황에서도 빠르게 초기 데이터셋을 구축할 수 있게 함
- 라벨 오류 감지를 위한 자동화 알고리즘은 비지도 또는 반지도 방법을 사용해 오류나 불일치를 찾을 수 있음
- 강화 학습 에이전트는 특징을 최적으로 선택 및 라벨링하고, 합성 특징을 생성하기 위한 파이프라인을 실행할 수 있음

○ 한편, 데이터의 양보다 질을 우선하는 최적화 기법이 새롭게 부상함에 따라 단순히 데이터의 양을 늘리는 것이 아닌, 학습 성능 향상에 기여하는 데이터를 선별하고 효율적으로 구성하는 방식이 주목받고 있음

- 대규모 데이터셋의 무차별적 학습은 연산 비용과 시간 소모를 증가시킴
- 모델 성능에 영향력이 큰 데이터 중심으로 구성하면 계산 자원을 절감하면서도 성능 유지 가능

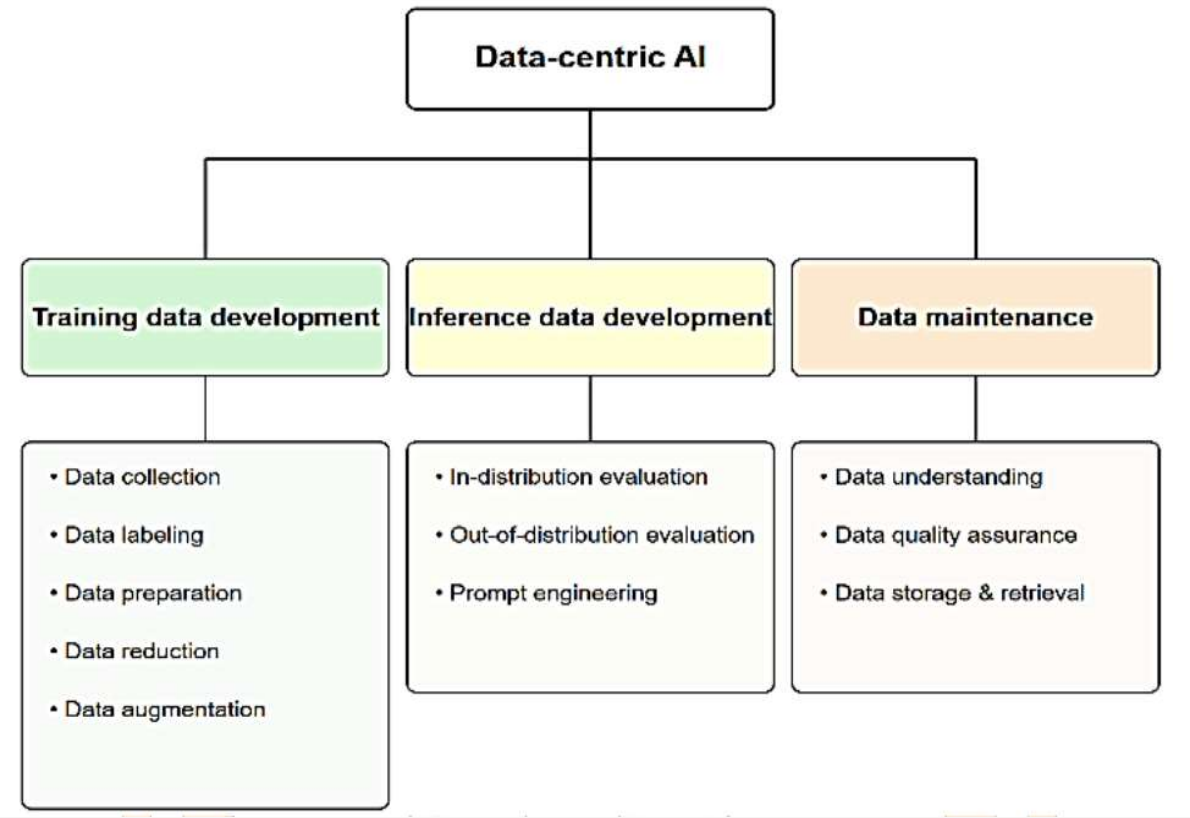
○ 학습 성능에 기여도가 높은 데이터를 선택적으로 사용하는 핵심 샘플 선별 기술의 적용이 증가함

- 커버리지 기반, 다양성 기반, 정보 이득 기반 알고리즘 등 다양한 기준으로 샘플 선택

○ 자동화에는 새로운 위험이 수반됨

- 자동화 시스템의 결과물에 맹목적으로 의존하여 품질 점검이 생략될 경우, 오류가 대규모로 전파될 수 있음
- 최신 DCAI 방법에서는 사람이 자동화 방법의 잘못된 판단을 감독하고 검증하며 수정할 수 있도록 하여 신뢰할 수 있고 가치 있는 자동화를 유지하는 휴먼 인 더 루프 자동화 방법을 고려하고 있음
- 향후, 자동화 범위는 증가하나, 인간의 지속적인 개선 및 의사결정 판단을 내릴 수 있도록 하는 방식으로 발전 예상

< 학습 데이터 개발, 추론 데이터 개발, 데이터 유지 관리 개요 >



* 출처 : Data-Centric AI: A Systematic Review of Methods, Challenges and Future Directions

○ 훈련데이터 개발 항목

- 데이터 수집: 다양한 출처에서 편향을 최소화하고 대표성 있는 데이터를 수집하는 과정
- 데이터 라벨링: 정확한 라벨링을 통해 모델 학습에 필요한 정보를 제공하는 작업
- 데이터 전처리: 데이터를 모델 학습에 적합한 형식으로 변환하고 오류를 제거하는 과정
- 데이터 축소: 데이터 세트의 크기를 줄여 연산 자원을 절감하고 학습 효율성을 높이는 작업
- 데이터 증강: 데이터 변형을 통해 모델의 일반화 능력을 향상시키는 기법

○ 추론 데이터 개발 항목

- 분포 내 평가: 훈련 데이터와 동일한 분포에서 모델 성능을 평가하는 기법
- 분포 외 평가: 훈련 데이터와 다른 분포에서 모델 성능을 평가하여 일반화 능력을 확인하는 방식
- 프롬프트 엔지니어링: 입력 텍스트를 설계하여 모델의 응답을 최적화하는 기법

○ 데이터 유지 관리 항목

- 데이터 이해: 수집된 데이터의 특성과 패턴을 분석하여 인사이트를 도출하는 과정
- 데이터 품질 보증: 데이터의 정확성, 일관성, 신뢰성을 보장하여 품질을 유지하는 작업
- 데이터 저장 및 검색: 데이터를 효율적으로 저장하고 필요한 정보를 빠르게 검색할 수 있도록 하는 기술

3 데이터 중심 AI를 위한 기술 현황[40]

▷ 효과적인 데이터 사용을 위해 대체 데이터 사용 기술이 중요

○ AI는 고품질, 관련성, 상호 운용성이 확보된 데이터 세트에 의존하나, 이러한 데이터가 부족하고 접근 비용이 높아 대체 데이터 기술이 중요해지고 있음

▷ 학습 효과 기반의 동적 데이터셋 구성

○ 모델 학습 진행에 따라 데이터셋을 유동적으로 변경해 최적의 학습 흐름 구성

- Curriculum Learning, Self-Paced Learning 등 학습 단계에 따라 난이도 조절
- 특히 LLM과 같은 대규모 모델에서 적절한 데이터 흐름 제어가 성능에 큰 영향을 미침

▷ 정보 밀도 기반 학습으로 연산 효율성 확보

○ 전체 데이터셋이 아닌 정보 밀도가 높은 핵심 데이터 중심 학습을 통해 연산 자원 절감 및 효율 극대화

- 중복 데이터나 정보량이 낮은 샘플은 제거하고, 의미적 다양성과 중요도를 반영한 데이터만 선택
- 대표적인 방법으로는 중요 문장 요약, 대표 샘플 클러스터링, 핵심 개체 기반 정제 등 활용

▷ 모델 복잡도와 성능 간 비효율 극복을 위한 대안

○ 모델 중심 접근은 연산 자원과 학습 시간의 기하급수적 증가를 동반하며, 일정 수준 이상의 성능 개선에는 한계가 존재

- 반면, 데이터 중심 접근은 상대적으로 적은 연산 자원으로도 효율적 학습이 가능
- 특히 데이터셋의 중복 제거, 다양성 확보, 고품질 데이터 확대 등을 통해 모델의 일반화 성능 향상

▷ 데이터 사이트 패러다임(Data Site Paradigm)[7)]

○ 데이터 사이트 패러다임은 연구자가 원시 데이터를 직접 다운로드하는 대신 코드를 데이터 사이트로 전송하여 데이터 분석을 수행하는 방식

- 데이터 소유자는 코드를 검토하고 승인한 후, 로컬에서 모델을 실행하여 결과만 연구자에게 반환
- 이 접근 방식은 데이터 로컬리티를 유지하면서도 안전하고 책임감 있는 분석을 가능하게 함

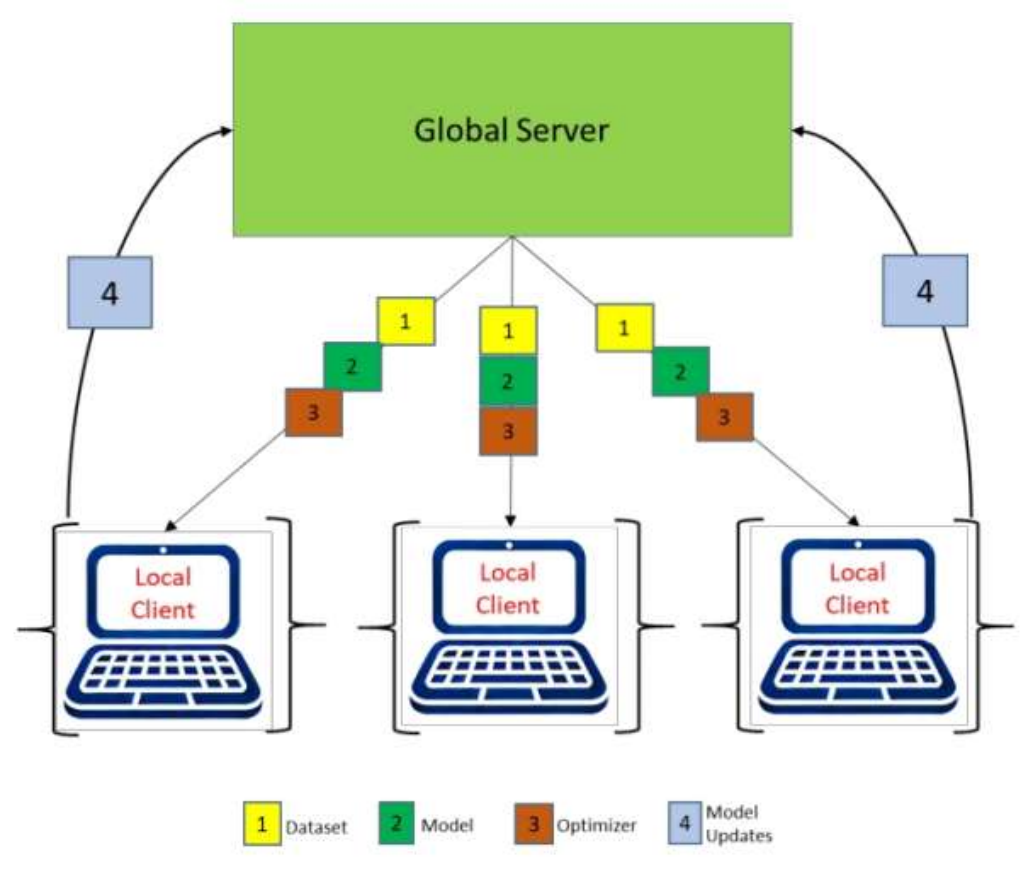
7) 원시 데이터 세트의 직접 다운로드를 방지하기 위해 개발된 기술로, 민감한 데이터 세트를 전송하는 대신 데이터가 있는 곳으로 모델이나 쿼리를 보내 데이터의 로컬리티를 유지하면서 안전하고 책임감 있는 분석이 가능함. 원격 실행, 감사 로깅, 프로젝트 거버넌스, 자동화된 정책 시행 등의 기능을 PET와 통합하여 워크플로우를 지원하며 자동화를 통해 보안과 확장성이 향상될 수 있음

▷ 연합 학습(Federated Learning)

○ 연합 학습은 원시 데이터를 중앙 서버로 전송하지 않고도 모델 학습을 가능하게 하는 기술

- 데이터 보유자가 로컬에서 모델을 업데이트하고, 업데이트된 모델 매개변수만 공유하여 중앙 코디네이터가 중앙에서 이를 집계 후 글로벌 모델을 생성
- 이 방식은 모바일 디바이스, IoT 시스템, 병원 네트워크와 같은 개인정보 보호가 중요한 환경에서 특히 효과적
- 또한, 민감 데이터가 원본을 벗어나지 않도록 보장하여 사용자 개인정보를 보호하고 법적 프레임워크를 준수할 수 있도록 설계됨

< 연합 학습 모델 학습 프로세스 >



* 출처 : Data-Driven Breakthroughs and Future Directions in AI Infrastructure: A Comprehensive Review

▷ 개인정보 보호 강화 기술(Privacy-Enhancing Technologies)

○ PET는 데이터 처리 파이프라인 전반에 걸쳐 개인정보 보호를 보장하는 일련의 기술을 포함

- 동형 암호화, 차별 프라이버시, 보안 다자간 계산(SMPC) 등의 방법이 포함
- PET는 의료, 국방, 금융 등 엄격한 규정 준수가 요구되는 분야에서 이미 실행 가능한 솔루션으로 배포되고 있음

▷ 합성 데이터 및 모의 데이터(Synthetic and Simulated Data)

○ 모의 데이터는 무작위로 생성되거나 익명화된 값을 포함하면서 실제 데이터의 구조적 특성을 복제하는 인공 데이터 세트임

- 시스템 아키텍처 테스트, API 검증, 사용자 인터페이스 검증 등 초기 개발 단계에서 주로 사용
- 합성 데이터는 실제 데이터의 통계적 속성을 모델링하여 생성된 데이터로, 학습, 검증 및 성능 벤치마킹에 효과적으로 사용
- 두 데이터 모두 개인정보 보호 문제를 최소화하면서 다양한 데이터 세트를 생성할 수 있음
  - 예: 병원 데이터의 통계적 분포와 일치하는 합성 데이터를 만들어 개인정보를 침해하지 않고도 모델을 개발함
- 그러나 생성 과정이 계산 집약적이고 복잡할 수 있으며, 실제 이벤트에서 발견되는 모든 범위의 분산을 포착하지 못할 가능성이 존재함

4 데이터 중심 AI 정책

▷ 책임 있는 AI를 위한 데이터 구축[41]

○ 데이터는 지식 경제의 핵심 생산요소이므로 데이터 정책 업데이트 필요

- 많은 국가들이 AI 등장 이전에 이미 데이터 정책을 수립했지만, 이를 업데이트해야 하며, 일부 국가들은 여전히 국가 데이터 프레임워크 부족
- 데이터 정책은 데이터베이스가 경제 전반에서 상호 운용성과 접근성을 갖출 수 있도록 보장해야 하며, 데이터의 입력부터 출력까지 전 과정에서 개인정보 보호를 위해 이용자의 동의를 기반으로 해야 함
- 편향성을 고려한 데이터 처리 방법과 지식 재산권 및 공정성 문제를 해결하는 정책이 필요함
- AI 개발을 지원하기 위해 지식 재산권 규정을 재검토하고, 공공-민간 협력을 촉진하는 메커니즘을 마련하는 것이 중요함

▷ AI와 데이터 보호 및 개인정보 보호[42]

○ 데이터 보호와 개인정보 보호는 데이터 중심 AI 정책의 핵심 요소임

- AI 시스템은 민감한 개인정보를 다루기 때문에, 데이터 보호와 개인정보 보호에 대한 공공의 우려를 해소해야 함
- 이를 통해 신뢰를 구축하고 AI 채택을 촉진할 수 있으므로, 법적 틀을 마련하여 AI 모델이 민감한 정보를 처리할 때의 위험을 최소화하는 규제와 기준이 필요함

▷ 디지털 격차 해소 및 데이터 인프라 구축[43]

○ 전 세계 약 3분의 1은 인터넷에 접근할 수 없어 디지털 격차가 발생하고 있음

- 이는 AI 활용 및 개발에 대한 참여를 방해하며, 디지털 리터러시 확산을 지연시킴
- 국가들은 디지털 인프라를 강화하고, AI 기술을 국가 개발 목표와 일치시키기 위한 정책을 수립해야 함
- 특히 개발도상국은 AI를 국가적 우선순위로 설정하고, 선제적 AI 정책을 수립해야 함

▷ AI 생태계와 데이터 기반 혁신[44]

○ AI 생태계는 데이터의 활용도에 따라 발전함

- 데이터를 중심으로 한 혁신 생태계를 구축하려면, 데이터와 AI 기술의 연구 및 개발을 촉진하는 정책이 필요함
- AI 연구 개발, 기업 지원, 공공부문 AI 활용 확대 등을 통한 AI 산업 생태계를 지원해야 함
- 데이터 기반 기업 지원 정책을 마련하고, 데이터 공유와 협력을 통해 산업 전반의 혁신을 촉진해야 함

▷ AI 정책 수립의 글로벌 불확실성 대응[45]

○ AI 기술의 발전은 불확실성과 위험을 동반함

- 이를 대응하지 않으면 더 큰 비용이 초래될 수 있음
- 각국은 AI의 빠른 발전에 발맞춰 정책을 유연하게 대응하고, 국제적으로 협력해야 함
- 개발도상국은 글로벌 협력 네트워크에 참여하여 자원 공유와 지식 교류를 촉진해야 하며, AI 관련 정책을 점진적으로 확산시켜야 함

▷ AI 정책의 국제적 차원에서의 대응[46]

○ AI는 국제적 특성을 가지므로, 국경을 초월한 데이터 정책을 수립해야 함

- 글로벌 시장에서의 AI 확산에 대응하는 정책적 대응이 필요함
- 국가들은 클라우드 컴퓨팅과 데이터 의존도를 관리하며, 국내 서비스 시장의 발전이 억제되지 않도록 주의 필요
- AI 전용 프로그램이나 오픈 데이터 이니셔티브를 통해 오픈 데이터를 수집하고 제공함으로써 데이터 통합과 협업을 효율화해야 함
- 미국과 EU는 다양한 규제 및 윤리 기준을 통해 데이터 정책을 선도하고 있으며, 아래 표에서는 기타 국가들의 데이터 구축 정책 사례를 소개함

< 데이터 프레임워크 구축이 필요한 국가별 정책 현황 >

Chile	Germany	India	Colombia	Singapore
Data Observatory	Mobility Data Space	Ethical Guidelines for AI in Biomedical Research and Healthcare	Sandbox on privacy by design and by default in AI projects	Computational Data Analysis Provision
Facilitate AI adoption by supporting data availability	Apply AI systems to specific industries through sectoral data marketplace	Ensure privacy, safety and security in data and algorithmic decisions	Support AI solutions that respect personal information and rights	Revise copyright law to support AI development with data accessibility and security
Key Actions	*Key Actions*	*Key Actions*	*Key Actions*	*Key Actions*
• Open data platforms leveraging public-private-academia collaborations • Provide data-based services and analyses across fields	• Launch a market-based platform to exchange data for the mobility sector • Incentivize participation with financial remuneration	• Prioritize human data privacy and security • Set processes to ensure representativeness and accountability in development and deployment of AI in health	• Create a secure environment for the experimentation of AI • Promote public-private collaboration to foster mutual learning	• Introduce exceptions and favor computational data analysis and machine learning • Implement safeguards to protect the commercial interests of copyright owners

*출처 : UNCTAD, 2025 Technology and Innovation Report Inclusive Artificial Intelligence for Development

5 결론

▷ 모델 중심에서 데이터 중심으로의 패러다임 전환

○ 기존 AI 개발은 고도화된 알고리즘 설계와 복잡한 모델 구조 개선에 초점을 맞추는 '모델 중심' 접근이 주를 이뤘으나, 최근에는 데이터 품질과 구성의 중요성을 강조하는 '데이터 중심' 접근으로 전환

- LLM을 포함한 최신 모델들이 동일한 구조를 유지하면서도 더 큰 데이터셋으로 성능을 향상시킨 사례 다수
- 학습 알고리즘이나 아키텍처 변경 없이, 데이터 품질과 커버리지 개선만으로 성능 향상 가능

▷ 실제 사례에서 입증된 데이터 중심 접근의 효율성

○ ChatGPT 등 대표적인 대규모 모델들도 구조 혁신보다는 대규모·고품질 데이터의 확보와 활용을 통해 성능 도약을 이룸

- ImageNet 사례처럼, 대량의 라벨링된 이미지 데이터가 컴퓨터 비전 분야의 성능을 비약적으로 향상
- 알고리즘보다 데이터에서 AI 발전의 핵심 동력이 비롯된다는 분석이 지속적으로 제기됨

▷ 연합 학습, PET, 데이터 사이트 패러다임 등 개인정보 보호 기술의 고도화

○ 개인정보 보호와 데이터 활용을 동시에 달성하기 위한 기술이 빠르게 진화하고 있음

- 연합 학습은 데이터를 중앙 서버로 전송하지 않고 로컬에서 모델을 학습시켜 개인정보 노출을 방지함
- PET는 동형 암호화, 차별 프라이버시, 보안 다자간 계산 등을 통해 데이터 처리 과정의 프라이버시를 보호함
- 데이터 사이트 패러다임은 원시 데이터를 다운로드하지 않고 코드만 전송하여 분석을 수행함으로써 데이터 로컬리티를 유지하며 보안성을 향상함
- 합성 데이터는 실제 데이터의 통계적 특성을 반영해 생성되며, 개발·검증·학습에 활용 가능하면서도 민감정보 노출 위험을 줄일 수 있음
- 이러한 기술들은 의료, 금융, IoT 등 민감한 분야에서 안전하고 책임감 있는 데이터 활용 기반을 마련하고 있음

▷ 데이터 중심 AI 실현을 위한 정책적 대응 필요

○ AI는 국가 간 경계를 넘나드는 기술인 만큼, 데이터 수집·활용·보호 전반에 걸친 국제적 협력이 필수적

- 각국은 AI의 급속한 발전 속도에 대응해 유연한 정책 수립이 필요하며, 데이터 기반 AI 생태계를 뒷받침할 수 있는 제도적 기반을 강화해야 함
- 특히 개발도상국은 글로벌 협력 네트워크에 참여해 오픈 데이터, 기술 자원, 전문 지식 등을 공유하고, 점진적으로 AI 정책을 확산시켜야 함
- 오픈 데이터 이니셔티브나 AI 전용 프로그램을 통해 공공 데이터 통합 및 활용 기반을 조성함으로써, 데이터 중심 AI의 실현 가능성을 높일 수 있음
- 이러한 정책적 대응은 데이터 격차로 인한 기술 불균형을 완화하고, AI의 공정하고 지속가능한 발전을 뒷받침하는 핵심 요소로 작용함

- 데이터 인증제도, 품질 평가 기준, 활용 이력 기반 검증 체계 등이 도입되어야 AI 학습 데이터 신뢰성이 제고됨
- 민간 데이터의 거래 활성화를 위한 계약 표준화, 익명화된 데이터의 안전한 공유 기반 마련도 중요함
- 또한 법적 불확실성을 줄이기 위해 '데이터 책임성', '데이터 기반 서비스 안전성' 등 데이터 중심 AI 시대에 적합한 규범 정비가 병행되어야 함

▷ 향후 과제는 데이터 중심 접근의 체계화와 국제적 협력 기반 마련

○ 선진국과 개발도상국 간의 AI 역량 격차를 해소하기 위해서는 글로벌 데이터 공유 기준 마련, AI 훈련용 공공 데이터의 국제 공동 활용 등이 필요함

- 기술적 격차뿐 아니라 데이터 인프라와 제도적 격차도 국제 협력을 통해 단계적으로 해소되어야 함

Contents

참고 문헌 V

1 참고 문헌

[1] CONGRESS,GOV, H.R.6216 - National Artificial Intelligence Initiative Act of 2020

[2] The American Presidency Project. Executive Order 14158—Establishing and Implementing the President's "Department of Government Efficiency", 2025

[3] COVINGTON, January 2025 AI Developments – Transitioning to the Trump Administration, 2025

[4] Federal Data Strategy 2021 Action Plan

[5] European Commission, European Data Governance Act

[6] PromethEUs, Common European Data Spaces and the EU Vision of Data Markets

[7] European Commission, New Digital Europe Programme invests over €176 million in European digital capacities and tech

[8] Centuro Global, Data Governance, The EU AI Act and the Future of Global Mobility, 2024

[9] 과학기술정보통신부, 신뢰할 수 있는 인공지능 구현 전략 발표, 2021

[10] 인공지능 윤리 소통채널, [정책] 인공지능 일상화 및 산업 고도화 계획, 2023

[11] 과학기술정보통신부, 국가인공지능 전략 정책방향, 2024

[12] 개인정보보호위원회, 안전한 개인정보, 신뢰받는 인공지능 시대 -「2025년 개인정보보호위원회 주요 정책 추진계획」발표 -, 2025

[13] LeadrPro, What is Scale AI? The Ultimate Guide to the Data Engine Powering Modern AI, 2024

[14] databricks, 데이터 마켓플레이스 또는 데이터 마켓이란?

[15] CLOVA Studio, KT AI Studio / dasarpAI, Google AI Studio vs Vertex AI, 2024

[16] HelloT, 루시아 2.5'의 솔트룩스, 성능·비용 면에서 딥시크 넘었다., 2025 / cody, OpenAI의 ChatGPT 엔터프라이즈: 가격, 혜택 및 보안

[17] ALCHERA, 버티컬 AI: 산업별 혁신의 열쇠, 2025

[18] Medium, Unlocking the Power of Data with AWS Data Exchange: Your Guide to Seamless Data Access and Monetization, 2024

[19] 아이뉴스24, KT, 'K 데이터 얼라이언스' 출범…한국적 AI 생태계 본격 시동, 2025 / ZDNET Korea, "데이터는 국부 원천...데이터 융합으로 새 시장 창출해야", 2021

[20] Crowdfund Insider, Private AI Firms Raised a Record $100.4B in 2024 – Report, 2025/DCD, Capital deployment in US data construction reached $31.5bn in 2024 – Newmark, 2025

[21] CVAT, CVAT vs. Label Studio: Which One to Choose?, 2024

[22] V7, What is Synthetic Data in Machine Learning and How to Generate It, 2022

[23] Medium, Sensor Fusion With Kalman Filter, 2023

[24] SUPERWISE, A gentle introduction to ML fairness metrics, Common fairness metrics, 2022

[25] SpringerOpen, Resampling approaches to handle class imbalance: a review from a data perspective, 2025

[26] TELMAI, Open source data quality tools you can't ignore: Great Expectations vs. Soda vs. Deequ, 2023

[27] Number Analytics, Curriculum Learning: The Future of AI Training, 2025

[28] neptune.ai, Active Learning: Strategies, Tools, and Real-World Use Cases, 2023

[29] Google Cloud, 검색 증강 생성(RAG)이란 무엇인가요?

[30] DVC, Automate Your ML Pipeline: Combining Airflow, DVC, and CML for a Seamless Batch Scoring Experience, 2023

[31] EMPR, FDA Clears AI Software Lunit INSIGHT MMG to Detect Breast Cancer, 2021

[32] SIEMENS, Machine health check: automating maintenance, 2020

[33] amity solutions, AI in Banking: How JPMorgan Uses AI to Detect Fraud, 2025

[34] 2025 Technology and Innovation Report Inclusive Artificial Intelligence for Development 28p, Data-Driven Breakthroughs and Future Directions in AI Infrastructure: A Comprehensive Review 1~2p, Data-Centric AI: A Systematic Review of Methods, Challenges and Future Directions 3p

[35] Data-Driven Breakthroughs and Future Directions in AI Infrastructure: A Comprehensive Review 4p

[36] Data-Driven Breakthroughs and Future Directions in AI Infrastructure: A Comprehensive Review 5p

[37] Data-Driven Breakthroughs and Future Directions in AI Infrastructure: A Comprehensive Review 5p

[38] 2025 Technology and Innovation Report Inclusive Artificial Intelligence for Development 136p, 159p

[39] Data-Centric AI: A Systematic Review of Methods, Challenges and Future Directions 5~8p

[40] Data-Driven Breakthroughs and Future Directions in AI Infrastructure: A Comprehensive Review 6~8p

[41] 2025 Technology and Innovation Report Inclusive Artificial Intelligence for Development 130p

[42] 2025 Technology and Innovation Report Inclusive Artificial Intelligence for Development 128p

[43] 2025 Technology and Innovation Report Inclusive Artificial Intelligence for Development 128p

[44] 2025 Technology and Innovation Report Inclusive Artificial Intelligence for Development 135~139p, 145~148p

[45] 2025 Technology and Innovation Report Inclusive Artificial Intelligence for Development 135~139p, 145~148p

[46] 2025 Technology and Innovation Report Inclusive Artificial Intelligence for Development 145~148p

2 URL

- https://www.presidency.ucsb.edu/documents/executive-order-14158-establishing-and-implementing-the-presidents-department-government
- https://www.insidegovernmentcontracts.com/2025/02/january-2025-ai-developments-transitioning-to-the-trump-administration/
- https://strategy.data.gov/assets/docs/draft-2021-federal-data-strategy-action-plan.pdf
- https://www.congress.gov/bill/116th-congress/house-bill/6216
- https://digital-strategy.ec.europa.eu/en/policies/data-governance-act
- https://www.prometheusnetwork.eu/blog/common-european-data-spaces-and-the-eu-vision-of-data-markets/
- https://digital-strategy.ec.europa.eu/en/node/12522/printable/pdf
- https://www.centuroglobal.com/article/data-governance-eu-ai-act/
- https://ai.kisdi.re.kr/aieth/bbs/B0000085/view.do?nttId=349&menuNo=400014&pageIndex=1
- https://www.msit.go.kr/bbs/view.do?bbsSeqNo=94&mId=113&mPid=238&nttSeqNo=3184952&pageIndex=3&ref=ai-ethics.kr&sCode=user
- https://www.msit.go.kr/bbs/view.do?sCode=user&mId=113&mPid=238&bbsSeqNo=94&nttSeqNo=3180239
- https://www.pipc.go.kr/np/cop/bbs/selectBoardArticle.do?bbsId=BS074&mCode=C020010000&nttId=10928
- https://www.leadrpro.com/blog/what-is-scale-ai-a-comprehensive-overview
- https://www.databricks.com/kr/glossary/data-marketplace
- https://clova.ai/en/clova-studio
- https://www.ncloud.com/product/aiService/clovaStudio#overview
- https://main—dasarpai.netlify.app/dsblog/google-ai-studio-vs-vertexai/
- https://tech.ktcloud.com/entry/KT-AI-Studio-%EC%84%9C%EB%B9%84%EC%8A%A4-%EC%86%8C%EA%B0%9C
- https://www.hellot.net/news/article.html?no=97927
- https://meetcody.ai/ko/blog/openai%EC%9D%98-chatgpt-%EC%97%94%ED%84%B0%ED%94%84%EB%9D%BC%EC%9D%B4%EC%A6%88-%EB%B9%84%EC%9A%A9-%EC%9D%B4%EC%A0%90-%EB%B0%8F-%EB%B3%B4%EC%95%88/
- https://www.alchera.ai/resource/blog/vertical-AI
- https://mihirpopat.medium.com/unlocking-the-power-of-data-with-aws-data-exchange-your-guide-to-seamless-data-access-and-76063e04de85
- https://www.inews24.com/view/1865977
- https://zdnet.co.kr/view/?no=20210222220603
- https://www.crowdfundinsider.com/2025/02/235834-private-ai-firms-raised-a-record-100-4b-in-2024-report/
- https://www.datacenterdynamics.com/en/news/capital-deployment-in-us-data-construction-reached-315bn-in-2024-newmark/
- https://www.cvat.ai/resources/blog/cvat-or-label-studio-which-one-to-choose
- https://www.v7labs.com/blog/synthetic-data-guide
- https://medium.com/%40satya15july_11937/sensor-fusion-with-kalman-filter-c648d6ec2ec2
- https://superwise.ai/blog/gentle-introduction-ml-fairness-metrics/
- https://fairlearn.org/main/user_guide/assessment/common_fairness_metrics.html
- https://journalofbigdata.springeropen.com/articles/10.1186/s40537-025-01119-4
- https://www.telm.ai/blog/open-source-data-quality-tools/
- https://www.numberanalytics.com/blog/curriculum-learning-future-of-ai-training
- https://neptune.ai/blog/active-learning-strategies-tools-use-cases
- https://cloud.google.com/use-cases/retrieval-augmented-generation
- https://dvc.org/blog/automate-your-ml-pipeline-combining-airflow-dvc-and-cml-for-a-seamless-batch-scoring-experience
- https://www.empr.com/home/news/fda-clears-ai-software-lunit-insight-mmg-to-detect-breast-cancer/
- https://www.plm.automation.siemens.com/media/global/ru/Siemens%20SW%20Machine%20health%20check%20automating%20maintenance%20White%20Paper_tcm52-85813.pdf
- https://www.amitysolutions.com/blog/ai-banking-jpmorgan-fraud-detection
- https://www.researchgate.net/publication/391989884_Data-Centric_AI_A_Systematic_Review_of_Methods_Challenges_and_Future_Directions
- https://arxiv.org/abs/2505.16771
- https://unctad.org/publication/technology-and-innovation-report-2025

GLOBAL DATA MARKET TRENDS

Monthly Trends Report

2025년 7월

발행일 2025.07.31.
발행처 한국데이터산업진흥원 서울시 중구 세종대로 9길 42, 부영빌딩 8층
기 획 데이터정책실 정책기획팀

K data 한국데이터산업진흥원