

# SPRi AI Brief

인공지능 산업의 최신 동향

2023년 12월호

# CONTENTS

## I. 인공지능 산업 동향 브리프

### 1. 정책/법제

- ▷ 미국, 안전하고 신뢰할 수 있는 AI 개발과 사용에 관한 행정명령 발표 ..... 1
- ▷ G7, 히로시마 AI 프로세스를 통해 AI 기업 대상 국제 행동강령에 합의 ..... 2
- ▷ 영국 AI 안전성 정상회의에 참가한 28개국, AI 위험에 공동 대응 선언 ..... 3
- ▷ 미국 법원, 예술가들이 생성 AI 기업에 제기한 저작권 소송 기각 ..... 4
- ▷ 미국 연방거래위원회, 저작권청에 소비자 보호와 경쟁 측면의 AI 의견서 제출 ..... 5
- ▷ EU AI 법 3자 협상, 기반모델 규제 관련 견해차로 난항 ..... 6

### 2. 기업/산업

- ▷ 미국 프런티어 모델 포럼, 1,000만 달러 규모의 AI 안전 기금 조성 ..... 7
- ▷ 코히어, 데이터 투명성 확보를 위한 데이터 출처 탐색기 공개 ..... 8
- ▷ 알리바바 클라우드, 최신 LLM '통이치엔원 2.0' 공개 ..... 9
- ▷ 삼성전자, 자체 개발 생성 AI '삼성 가우스' 공개 ..... 10
- ▷ 구글, 앤스로픽에 20억 달러 투자로 생성 AI 협력 강화 ..... 11
- ▷ IDC, 2027년 AI 소프트웨어 매출 2,500억 달러 돌파 전망 ..... 12
- ▷ 빌 게이츠, AI 에이전트로 인한 컴퓨터 사용의 패러다임 변화 전망 ..... 13
- ▷ 유튜브, 2024년부터 AI 생성 콘텐츠 표시 의무화 ..... 14

### 3. 기술/연구

- ▷ 영국 과학혁신기술부, AI 안전 연구소 설립 발표 ..... 15
- ▷ 구글 딥마인드, 범용 AI 모델의 기능과 동작에 대한 분류 체계 발표 ..... 16
- ▷ 갈릴레오의 LLM 환각 지수 평가에서 GPT-4가 가장 우수 ..... 17

### 4. 인력/교육

- ▷ 영국 옥스퍼드 인터넷 연구소, AI 기술자의 임금이 평균 21% 높아 ..... 18

## II. 주요 행사

- ▷ CES 2024 ..... 19
- ▷ AIMLA 2024 ..... 19
- ▷ AAAI Conference on Artificial Intelligence ..... 19

# I . 인공지능 산업 동향 브리프

## 미국, 안전하고 신뢰할 수 있는 AI 개발과 사용에 관한 행정명령 발표

### KEY Contents

- 미국 바이든 대통령이 '안전하고 신뢰할 수 있는 AI 개발과 사용에 관한 행정명령'에 서명하고 광범위한 행정 조치를 명시
- 행정명령은 △AI의 안전과 보안 기준 마련 △개인정보보호 △형평성과 시민권 향상 △소비자 보호 △노동자 지원 △혁신과 경쟁 촉진 △국제협력에 골자로 함

### ● 바이든 대통령, AI 행정명령 통해 안전하고 신뢰할 수 있는 AI 개발과 활용 추진

- 미국 바이든 대통령이 2023년 10월 30일 연방정부 차원에서 안전하고 신뢰할 수 있는 AI 개발과 사용을 보장하기 위한 행정명령을 발표
  - 행정명령은 △AI의 안전과 보안 기준 마련 △개인정보보호 △형평성과 시민권 향상 △소비자 보호 △노동자 지원 △혁신과 경쟁 촉진 △국제협력에 관한 내용을 포괄
- (AI 안전과 보안 기준) 강력한 AI 시스템을 개발하는 기업에게 안전 테스트 결과와 시스템에 관한 주요 정보를 미국 정부와 공유할 것을 요구하고, AI 시스템의 안전성과 신뢰성 확인을 위한 표준 및 AI 생성 콘텐츠 표시를 위한 표준과 모범사례 확립을 추진
  - $\Delta 10^{26}$  플롭스(FLOPS, Floating Point Operation Per Second)를 초과하는 컴퓨팅 성능 또는 생물학적 서열 데이터를 주로 사용하고  $10^{23}$ 플롭스를 초과하는 컴퓨팅 성능을 사용하는 모델 단일 데이터센터에서 1,000Gbit/s 이상의 네트워크로 연결되며 AI 훈련에서 이론상 최대  $10^{20}$  플롭스를 처리할 수 있는 컴퓨팅 용량을 갖춘 컴퓨팅 클러스터가 정보공유 요구대상
- (형평성과 시민권 향상) 법률, 주택, 보건 분야에서 AI의 무책임한 사용으로 인한 차별과 편견 및 기타 문제를 방지하는 조치를 확대
  - 형사사법 시스템에서 AI 사용 모범사례를 개발하고, 주택 임대 시 AI 알고리즘 차별을 막기 위한 명확한 지침을 제공하며, 보건복지 부문에서 책임 있는 AI 배포와 사용을 위한 전략을 마련
- (소비자 보호와 근로자 지원) 의료 분야에서 책임 있는 AI 사용을 촉진하고 맞춤형 개인교습 등 학교 내 AI 교육 도구 관련 자원을 개발하며, AI로 인한 근로자 피해를 완화하고 이점을 극대화하는 원칙과 모범사례를 마련
- (혁신과 경쟁 촉진) 국가AI연구자원(National Artificial Intelligence Research Resource, NAIRR)\*을 통해 미국 전역의 AI 연구를 촉진하고, 중소기업과 개발자에 기술과 인프라를 지원
  - \* 국가 차원에서 AI 연구 인프라를 확충해 더 많은 AI 연구자에게 인프라를 지원하는 프로그램
  - 비자 기준과 인터뷰 절차의 현대화와 간소화로 AI 관련 주요 분야의 전문 지식을 갖춘 외국인들이 미국에서 공부하고 취업할 수 있도록 지원

출처 : The White House, Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (E.O. 14110), 2023.10.30.

## G7, 히로시마 AI 프로세스를 통해 AI 기업 대상 국제 행동강령에 합의

### KEY Contents

- G7이 첨단 AI 시스템을 개발하는 기업을 대상으로 AI 위험 식별과 완화를 위해 자발적인 채택을 권고하는 AI 국제 행동강령을 마련
- 행동강령은 AI 수명주기 전반에 걸친 위험 평가와 완화, 투명성과 책임성의 보장, 정보공유와 이해관계자 간 협력, 보안 통제, 콘텐츠 인증과 출처 확인 등의 조치를 요구

### ○ G7, 첨단 AI 시스템의 위험 관리를 위한 국제 행동강령 마련

- 주요 7개국(G7)\*은 2023년 10월 30일 ‘히로시마 AI 프로세스’를 통해 AI 기업 대상의 AI 국제 행동강령(International Code of Conduct for Advanced AI Systems)에 합의
  - G7은 2023년 5월 일본 히로시마에서 개최된 정상회의에서 생성 AI에 관한 국제규범 마련과 정보공유를 위해 ‘히로시마 AI 프로세스’를 출범\*\*
  - 기업의 자발적 채택을 위해 마련된 이번 행동강령은 기반모델과 생성 AI를 포함한 첨단 AI 시스템의 위험 식별과 완화에 필요한 조치를 포함
- \* 주요 7개국(G7)은 미국, 일본, 독일, 영국, 프랑스, 이탈리아, 캐나다를 의미
- \*\* 5월 정상회의에는 한국, 호주, 베트남 등을 포함한 8개국이 초청을 받았으나, AI 국제 행동강령에는 우선 G7 국가만 포함하여 채택
- G7은 행동강령을 통해 아래의 조치를 제시했으며, 빠르게 발전하는 기술에 대응할 수 있도록 이해관계자 협의를 통해 필요에 따라 개정할 예정
  - 첨단 AI 시스템의 개발 과정에서 AI 수명주기 전반에 걸쳐 위험을 평가 및 완화하는 조치를 채택하고, 첨단 AI 시스템의 출시와 배포 이후 취약점과 오용 사고, 오용 유형을 파악해 완화
  - 첨단 AI 시스템의 성능과 한계를 공개하고 적절하거나 부적절한 사용영역을 알리는 방법으로 투명성을 보장하고 책임성을 강화
  - 산업계, 정부, 시민사회, 학계를 포함해 첨단 AI 시스템을 개발하는 조직 간 정보공유와 사고 발생 시 신고를 위해 협력하고, 위험 기반 접근방식을 토대로 개인정보보호 정책과 위험 완화 조치를 포함하는 AI 거버넌스와 위험 관리 정책을 마련
  - AI 수명주기 전반에 걸쳐 물리보안, 사이버보안, 내부자 위협 보안을 포함한 강력한 보안 통제 구현
  - 사용자가 AI 생성 콘텐츠를 식별할 수 있도록 워터마크를 비롯하여 기술적으로 가능한 기법으로 신뢰할 수 있는 콘텐츠 인증과 출처 확인 메커니즘을 개발 및 구축
  - 사회적 위험과 안전·보안 문제를 완화하는 연구와 효과적인 완화 대책에 우선 투자하고, 기후 위기 대응, 세계 보건과 교육 등 세계적 난제 해결을 위한 첨단 AI 시스템을 우선 개발
  - 국제 기술 표준의 개발 및 채택을 가속화하고, 개인정보와 지식재산권 보호를 위해 데이터 입력과 수집 시 적절한 보호 장치 구현

출처: G7, Hiroshima Process International Code of Conduct for Advanced AI Systems, 2023.10.30.

## 영국 AI 안전성 정상회의에 참가한 28개국, AI 위험에 공동 대응 선언

### KEY Contents

- 영국 블레츨리 파크에서 개최된 AI 안전성 정상회의에 참가한 28개국들이 AI 안전 보장을 위한 협력 방안을 담은 블레츨리 선언을 발표
- 첨단 AI를 개발하는 국가와 기업들은 AI 시스템에 대한 안전 테스트 계획에 합의했으며, 영국의 AI 안전 연구소가 전 세계 국가와 협력해 테스트를 주도할 예정

### ● AI 안전성 정상회의 참가국들, 블레츨리 선언 통해 AI 안전 보장을 위한 협력에 합의

- 2023년 11월 1~2일 영국 블레츨리 파크에서 열린 AI 안전성 정상회의(AI Safety Summit)에 참가한 28개국 대표들이 AI 위험 관리를 위한 '블레츨리 선언'을 발표
- 선언은 AI 안전 보장을 위해 국가, 국제기구, 기업, 시민사회, 학계를 포함한 모든 이해관계자의 협력이 중요하다고 강조했으며, 특히 최첨단 AI 시스템 개발 기업은 안전 평가를 비롯한 적절한 조치를 취하여 AI 시스템의 안전을 보장할 책임이 있다고 지적
- 각국은 AI 안전 보장을 위해 첨단 AI 개발기업의 투명성 향상, 적절한 평가지표와 안전 테스트 도구 개발, 공공부문 역량 구축과 과학 연구개발 등의 분야에서 협력하기로 합의

### ● 영국 총리, 정부 주도의 첨단 AI 시스템 안전 테스트 계획 발표

- 리시 수낙 영국 총리는 AI 안전성 정상회의를 마무리하며 첨단 AI 모델에 대한 안전성 시험 계획 수립과 테스트 수행을 주도할 영국 AI 안전 연구소의 출범을 발표
- 첨단 AI 모델의 안전 테스트는 국가 안보와 안전, 사회적 피해를 포함한 여러 잠재적 유해 기능에 대한 시험을 포함하며, 참석자들은 정부 주도의 외부 안전 테스트에 합의
- 각국 정부는 테스트와 기타 안전 연구를 위한 공공부문 역량에 투자하고, 테스트 결과가 다른 국가와 관련된 경우 해당 국가와 결과를 공유하며, 적절한 시기에 공동 표준 개발을 위해 노력하기로 합의
- 참가국들은 튜링상을 수상한 AI 학자인 요슈아 벤지오 교수가 주도하는 '과학의 현황(State of the Science)' 보고서 작성에도 합의했으며, 보고서를 통해 첨단 AI의 위험과 가능성에 관한 기존 연구를 과학적으로 평가하고 향후 AI 안전 연구를 위한 우선순위를 제시할 계획
- 한국은 영국 정부와 6개월 뒤에 온라인으로 AI 미니 정상회의를 공동 개최하기로 합의했으며, 프랑스 정부와는 1년 후 대면 정상회의를 개최할 예정

출처: Gov.uk, The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023, 2023.11.01.  
Gov.uk, World leaders, top AI companies set out plan for safety testing of frontier as first global AI Safety Summit concludes, 2023.11.02.

## 미국 법원, 예술가들이 생성 AI 기업에 제기한 저작권 소송 기각

### KEY Contents

- 미국 캘리포니아 북부지방법원은 미드저니, 스태빌리티AI, 디비언트아트를 대상으로 예술가 3인이 제기한 저작권 침해 소송을 기각
- 법원은 기각 이유로 고소장에 제시된 상당수 작품이 저작권청에 등록되지 않았으며, AI로 생성된 이미지와 특정 작품 간 유사성을 입증하기 어렵다는 점을 제시

### ● 예술가들의 AI 저작권 침해 소송, 저작권 미등록과 증거불충분으로 기각

- 미국 캘리포니아 북부지방법원의 윌리엄 오릭(William Orrick) 판사는 2023년 10월 30일 미드저니(Midjourney), 스태빌리티AI(Stability AI), 디비언트아트(DeviantArt)에 제기된 저작권 침해 소송을 기각
  - 2023년 1월 예술가 사라 앤더슨(Sarah Anderson), 캘리 맥커넌(Kelly McKernan), 칼라 오르티즈(Karla Ortiz)는 이미지 생성 AI 서비스를 개발한 3개 기업을 상대로 저작권 침해 소송을 제기
  - 예술가들은 3개 기업이 AI 모델을 학습시키기 위해 원작자 동의 없이 작품을 학습 데이터셋에 포함하여 저작권을 침해했다고 주장했으며, 법원은 지난 4월 피소 기업들이 제출한 기각 신청을 수용해 소송을 기각
- 오릭 판사는 판결문에서 소송을 기각한 핵심 이유로 예술가들의 저작권 미등록을 제시
  - 판결문은 소송을 제기한 캘리 맥커넌과 칼라 오르티즈가 미국 저작권청에 예술 작품에 대한 저작권을 제출하지 않았다는 점을 지적했으며, 사라 앤더슨은 고소장에 인용된 수백 개의 작품 중 16개 작품에 대해서만 저작권을 보유
- 판결문은 또한 생성 AI 모델 훈련에 사용된 모든 이미지에 저작권이 있다거나, 생성 AI로 만든 이미지가 저작물을 이용해 훈련되었으므로 저작물의 파생 이미지라는 주장은 개연성이 부족하다고 지적
  - AI는 새로운 이미지를 생성할 때 다양한 예술가의 작품을 참조하므로, 생성된 이미지와 저작권을 가진 특정 작품과의 실질적 유사성을 입증할 수 없다면 저작권 침해를 인정받기 어려움
- 오릭 판사는 원고 측에 고소장을 수정하고 저작권이 침해된 특정 이미지를 중심으로 소송 범위를 줄여 소송을 다시 제기할 것을 요청
  - 단, 사라 앤더슨이 저작권을 보유한 16개 작품을 무단으로 복제한 스태빌리티AI에 대한 저작권 침해 소송은 인정되어 계속 진행됨

출처: Venturebeat, Midjourney, Stability AI and DeviantArt win a victory in copyright case by artists– but the fight continues, 2023.10.30.



## 미국 연방거래위원회, 저작권청에 소비자 보호와 경쟁 측면의 AI 의견서 제출

### KEY Contents

- 미국 FTC는 저작권청이 실시한 저작권과 AI 관련 질의공고에 대하여 소비자 보호와 경쟁 측면의 의견을 제시
- FTC는 생성 AI로 인한 창작자와 소비자 피해의 가능성에 우려를 표시하는 한편, 일부 빅테크가 막대한 재원을 활용해 시장 지배력을 더욱 강화할 수 있다는 우려를 제기

### ○ FTC, 생성 AI로 인한 소비자와 창작자의 피해 및 빅테크의 시장 지배력 강화 우려

- 미국 연방거래위원회(FTC)가 2023년 10월 30일 저작권청(U.S. Copyright Office, USCO)이 지난 9월 발표한 저작권과 AI 관련 질의공고(Notice of Inquiry, NOI)에 대한 의견서를 발표
  - 저작권청은 생성 AI와 관련된 저작권법과 정책 이슈를 조사하고 있으며, 폭넓은 의견 수렴을 통해 입법과 규제 조치의 필요성을 검토할 계획
  - FTC는 생성 AI의 개발과 배포가 소비자, 근로자, 중소기업에 피해를 줄 수 있다며 소비자의 개인정보 침해, 차별과 편견의 자동화, 사기 범죄 등 AI 사용과 관련된 위험에 주목
- FTC는 저작권법에 따른 권리와 책임 범위를 넘어서는 저작권 문제에 주목하여 생성 AI로 인해 창작자의 경쟁력이 불공정한 피해를 볼 수 있으며, 소비자가 특정 창작자의 작품을 생성 AI가 만들었다고 오해할 소지가 있다고 지적
  - 저작권법에 저촉되는 행위는 불공정 경쟁이나 기만행위에도 해당될 수 있으며, 창작자의 평판 악화, 저작물의 가치 저하나 개인정보 유출로 소비자에 상당한 피해를 초래 가능
- FTC는 일부 빅테크가 막대한 재원을 활용해 생성 AI 사용자의 이탈을 막고 저작권이 있는 상용 데이터에 대한 독점 라이선스를 확보해 시장 지배력을 더욱 강화할 수 있다는 우려도 제기
  - 이와 관련 FTC는 아마존 AI 비서 '알렉사(Alexa)'와 스마트홈 보안 기기 '링(Ring)'이 소비자의 사적 정보를 알고리즘 훈련에 사용하여 프라이버시를 침해한 혐의를 조사하는 등 법적 권한을 활용해 AI 관련 불법 행위에 대처하고 있음
    - \* FTC는 2023년 5월 31일 동의를 받지 않고 어린이들의 음성과 위치 정보를 활용한 '알렉사'와 고객의 사적 영상에 대하여 직원에게 무제한 접근 권한을 부여한 '링'에 3,080만 달러(약 420억 원)의 과징금을 부과
- FTC는 빠르게 발전하는 생성 AI가 여러 산업과 비즈니스에 변화를 가져올 수 있지만, 현행법상 AI에 관한 예외 조항은 없으며, 모든 권한을 활용해 소비자를 보호하고 개방적이고 공정한 경쟁 시장을 유지하겠다고 강조

출처: FTC, In Comment Submitted to U.S. Copyright Office, FTC Raises AI-related Competition and Consumer Protection Issues, Stressing That It Will Use Its Authority to Protect Competition and Consumers in AI Markets, 2023.10.30.



## EU AI 법 3자 협상, 기반모델 규제 관련 견해차로 난항

### KEY Contents

- 유럽의회, EU 집행위원회, EU 이사회가 진행 중인 AI 법 최종협상에서 프랑스, 이탈리아, 독일이 기반모델에 대한 규제에 반대하며 협상이 난관에 봉착
- 프랑스, 이탈리아, 독일 3개국은 기반모델 개발기업에 대하여 자율적 행동강령을 도입하고 준수를 의무화하는 방안을 제안

### ● AI 법 3자 협상, 이사회 일부 국가가 기반모델 규제에 반대하며 차질

- 유럽의회, EU 집행위원회, EU 이사회가 'AI 법(AI act)'에 대한 최종협상을 진행 중인 가운데, 일부 국가가 기반모델에 대한 규제에 반대하며 협상이 난관에 봉착
  - 10월 24일 열린 3자 협상 회의에서는 사회에 더 큰 영향을 미치는 강력한 AI 모델에 더 엄격한 규칙을 적용하는 계층적 접근방식에 따라 기반 모델 규제에 대한 기본적인 합의에 도달
  - 그러나 11월 10일 열린 통신작업반 회의에서 EU 이사회는 프랑스, 독일, 이탈리아 대표가 기반모델에 대한 모든 유형의 규제에 반대하며 협상이 중단됨
- 유럽 정책 미디어 유랙티브(Euractiv)에 따르면 프랑스 AI 기업 미스트랄(Mistral)이 로비를 통해 기반모델에 대한 규제 반대를 주도
  - 독일의 대표적인 AI 기업 알레프 알파(Aleph Alpha) 역시 독일 정부에 압력을 행사하고 있으며, 이들 기업은 EU의 AI 규제로 인해 미국과 중국의 경쟁사보다 뒤처질 것을 우려

### ● 독일, 프랑스, 이탈리아 3개국, 기반모델에 대한 '의무적 자율규제' 제안

- 통신작업반 회의가 결렬된 이후 독일, 프랑스, 이탈리아는 2023년 11월 19일 비공식 문서를 통해 '의무적 자율규제(Mandatory Self-regulation)' 방식의 기반모델 규제를 제안
  - 3개국은 기반모델 전반에 대한 규제가 기술 중립적이고 위험 기반의 AI 규제 원칙에 어긋난다고 주장하며 기반모델 전반에 대한 규제가 아닌, 특정 용도로 사용될 수 있는 AI 시스템에 대한 규제를 요구
  - 3개국은 자발적인 행동강령을 도입하고 준수를 의무화하는 방안을 제안하며, 기반모델 개발기업에 머신러닝 기술 정보와 모델의 기능과 한계를 요약한 '모델 카드' 작성을 요구하겠다고 설명
  - 3개국은 AI 감독기관이 모델 카드를 토대로 기반모델 개발기업의 행동강령 준수 여부를 확인하되, 위반 시 곧바로 제재를 가하지 않고 위반행위 분석과 영향 평가를 시행한 후 제재하는 방안을 제안

출처: Euractiv, EU's AI Act negotiations hit the brakes over foundation models, 2023.11.1.

Euractiv, France, Germany, Italy push for 'mandatory self-regulation' for foundation models in EU's AI law, 2023.11.19.

## 미국 프런티어 모델 포럼, 1,000만 달러 규모의 AI 안전 기금 조성

### KEY Contents

- 구글, 앤스로픽, 마이크로소프트, 오픈AI가 참여하는 프런티어 모델 포럼이 자선단체와 함께 AI 안전 연구를 위한 1,000만 달러 규모의 AI 안전 기금을 조성
- 프런티어 모델 포럼은 AI 모델의 취약점을 발견하고 검증하는 레드팀 활동을 지원하기 위한 모델 평가 기법 개발에 자금을 중점 지원할 계획

### ● 프런티어 모델 포럼, 자선단체와 함께 AI 안전 연구를 위한 기금 조성

- 구글, 앤스로픽, 마이크로소프트, 오픈AI가 출범한 프런티어 모델 포럼이 2023년 10월 25일 AI 안전 연구를 위한 기금을 조성한다고 발표
  - 참여사들은 맥거번 재단(Patrick J. McGovern Foundation), 데이비드 앤 루실 패커드 재단(The David and Lucile Packard Foundation) 등의 자선단체와 함께 AI 안전 연구를 위한 기금에 1,000만 달러 이상을 기부
  - 또한 신기술의 거버넌스와 안전 분야에서 전문성을 갖춘 브루킹스 연구소 출신의 크리스 메서롤(Chris Meserole)을 포럼의 상무이사로 임명
- 최근 AI 기술이 급속히 발전하면서 AI 안전에 관한 연구가 부족한 시점에, 포럼은 이러한 격차를 해소하기 위해 AI 안전 기금을 조성
  - 참여사들은 지난 7월 백악관 주재의 AI 안전 서약에서 외부자의 AI 시스템 취약점 발견과 신고를 촉진하기로 약속했으며, 약속을 이행하기 위해 기금을 활용해 외부 연구집단의 AI 시스템 평가에 자금을 지원할 계획

### ● AI 안전 기금으로 AI 레드팀을 위한 모델 평가 기법 개발을 중점 지원할 계획

- 프런티어 모델 포럼은 AI 안전 기금을 통해 AI 레드팀 활동을 위한 새로운 모델 평가 기법의 개발을 중점 지원할 예정
  - 포럼에 따르면 AI 레드팀에 대한 자금 지원은 AI 모델의 안전과 보안 기준의 개선과 함께 AI 시스템 위험 대응 방안에 관한 산업계와 정부, 시민사회의 통찰력 확보에 도움이 될 전망으로, 포럼은 향후 몇 달 안에 기금 지원을 위한 제안 요청을 받을 계획
- 프런티어 모델 포럼은 출범 이후 업계 전반에 걸쳐 AI 레드팀 구성에 관한 모범사례 공유를 추진하는 한편, 첨단 AI 모델의 취약점이나 잠재적으로 위험한 기능 및 위험 완화 관련 정보를 공유할 수 있는 공개 절차도 개발 중

출처: Google, Anthropic, Google, Microsoft and OpenAI announce Executive Director of the Frontier Model Forum and over \$10 million for a new AI Safety Fund, 2023.10.25.

## 코히어, 데이터 투명성 확보를 위한 데이터 출처 탐색기 공개

### KEY Contents

- 코히어와 12개 기관이 광범위한 데이터셋에 대한 감사를 통해 원본 데이터 출처, 재라이선스 상태, 작성자 등 다양한 정보를 제공하는 '데이터 출처 탐색기' 플랫폼을 출시
- 대화형 플랫폼을 통해 개발자는 데이터셋의 라이선스 상태를 쉽게 파악할 수 있으며 데이터셋의 구성과 계보도 추적 가능

### ○ 데이터 출처 탐색기, 광범위한 데이터셋 정보 제공을 통해 데이터 투명성 향상

- AI 기업 코히어(Cohere)가 매사추세츠 공과대(MIT), 하버드대 로스쿨, 카네기멜론대 등 12개 기관과 함께 2023년 10월 25일 '데이터 출처 탐색기(Data Provenance Explorer)' 플랫폼을 공개
  - AI 모델 훈련에 사용되는 데이터셋의 불분명한 출처로 인해 데이터 투명성이 확보되지 않아 다양한 법적·윤리적 문제가 발생
  - 이에 연구진은 가장 널리 사용되는 2,000여 개의 미세조정 데이터셋을 감사 및 추적하여 데이터셋에 원본 데이터소스에 대한 태그, 재라이선스(Relicensing) 상태, 작성자, 기타 데이터 속성을 지정하고 이러한 정보에 접근할 수 있는 플랫폼을 출시
  - 대화형 플랫폼 형태의 데이터 출처 탐색기를 통해 데이터셋의 라이선스 상태를 쉽게 파악할 수 있으며, 주요 데이터셋의 구성과 데이터 계보도 추적 가능
- 연구진은 오픈소스 데이터셋에 대한 광범위한 감사를 통해 데이터 투명성에 영향을 미치는 주요 요인을 발견
  - 깃허브(GitHub), 페이퍼워드코드(Papers with Code)와 같은 클라우드소싱 플랫폼에서 수집한 데이터로 훈련된 오픈소스 LLM에서는 데이터 라이선스의 누락 비율이 72~83%에 달함
  - 또한 클라우드소싱 플랫폼이 할당한 라이선스는 데이터셋 원저작자의 의도보다 더 광범위한 사용을 허용한 경우가 상당수
  - 데이터 생태계 분석 결과, 부정확하거나 모호한 라이선스 문서화 등 데이터 출처 입증과 관련된 관행 전반에서 구조적 문제가 드러남
- 연구진은 데이터 출처 탐색기만으로는 해결이 어려운 법적 이슈도 존재한다며 일관된 법적 프레임워크의 필요성을 제기
  - 일례로 데이터를 수집한 지역, 모델 훈련 지역, 모델 배포 지역마다 규제가 다르면 어떤 법률을 적용해야 하는지 실무자의 판단이 어려울 수 있으며, 서로 다른 라이선스를 적용받는 개별 데이터셋을 하나로 통합해 사용하는 경우에도 각각의 라이선스 조건 준수에 어려움이 발생

출처 : Cohere, Data Provenance Explorer Launches to Tackle Data Transparency Crisis, 2023.10.25.

## 알리바바 클라우드, 최신 LLM ‘통이치엔원 2.0’ 공개

### KEY Contents

- 알리바바 클라우드가 복잡한 지침 이해, 광고문구 작성, 추론, 암기 등에서 성능이 향상된 최신 LLM ‘통이치엔원 2.0’을 공개
- 알리바바 클라우드는 산업별로 특화된 생성 AI 모델을 공개하는 한편, 모델 개발과 애플리케이션 구축 절차를 간소화하는 올인원 AI 모델 구축 플랫폼도 출시

### ● 알리바바의 통이치엔원 2.0, 주요 벤치마크 테스트에서 여타 LLM 능가

- 중국의 알리바바 클라우드가 2023년 10월 31일 열린 연례 기술 컨퍼런스에서 최신 LLM ‘통이치엔원(Tongyi Qianwen) 2.0’을 공개
  - 알리바바 클라우드는 통이치엔원 2.0이 2023년 4월 출시된 1.0 버전보다 복잡한 지침 이해, 광고문구 작성, 추론, 암기 등에서 성능이 향상되었다고 설명
  - 통이치엔원 2.0은 언어 이해 테스트(MMLU), 수학(GSM8k), 질문 답변(ARC-C)과 같은 벤치마크 테스트에서 라마(Llama-2-70B)와 GPT-3.5를 비롯한 주요 AI 모델을 능가
  - 통이치엔원 2.0은 알리바바 클라우드의 웹사이트와 모바일 앱을 통해 대중에 제공되며 개발자는 API를 통해 사용 가능
- 알리바바 클라우드는 여러 산업 영역에서 생성 AI를 활용해 사업 성과를 개선할 수 있도록 지원하는 산업별 모델도 출시
  - 산업 영역은 고객지원, 법률 상담, 의료, 금융, 문서관리, 오디오와 동영상 관리, 코드 개발, 캐릭터 제작을 포함
- 알리바바 클라우드는 급증하는 생성 AI 수요에 대응해 모델 개발과 애플리케이션 구축 절차를 간소화하는 올인원 AI 모델 구축 플랫폼 ‘젠AI(GenAI)’도 공개
  - 이 플랫폼은 데이터 관리, 모델 배포와 평가, 신속한 엔지니어링을 위한 종합 도구 모음을 제공하여 다양한 기업들이 맞춤형 AI 모델을 한층 쉽게 개발할 수 있도록 지원
  - 생성 AI 개발에 필요한 컴퓨팅과 데이터 처리 요구사항을 지원하기 위해 AI 플랫폼(PAI), 데이터베이스 솔루션, 컨테이너 서비스와 같은 클라우드 신제품도 발표
- 알리바바 클라우드는 AI 개발을 촉진하기 위해 올해 말까지 720억 개 매개변수를 가진 통이치엔원 모델을 오픈소스화한다는 계획도 공개

## 삼성전자, 자체 개발 생성 AI ‘삼성 가우스’ 공개

### KEY Contents

- 삼성전자가 온디바이스에서 작동 가능하며 언어, 코드, 이미지의 3개 모델로 구성된 자체 개발 생성 AI 모델 ‘삼성 가우스’를 공개
- 삼성전자는 삼성 가우스를 다양한 제품에 단계적으로 탑재할 계획으로, 온디바이스 작동이 가능한 삼성 가우스는 외부로 사용자 정보가 유출될 위험이 없다는 장점을 보유

### ● 언어, 코드, 이미지의 3개 모델로 구성된 삼성 가우스, 온디바이스 작동 지원

- 삼성전자가 2023년 11월 8일 열린 ‘삼성 AI 포럼 2023’ 행사에서 자체 개발한 생성 AI 모델 ‘삼성 가우스’를 최초 공개
  - 정규분포 이론을 정립한 천재 수학자 가우스(Gauss)의 이름을 본뜬 삼성 가우스는 다양한 상황에 최적화된 크기의 모델 선택이 가능
  - 삼성 가우스는 라이선스나 개인정보를 침해하지 않는 안전한 데이터를 통해 학습되었으며, 온디바이스에서 작동하도록 설계되어 외부로 사용자의 정보가 유출되지 않는 장점을 보유
  - 삼성전자는 삼성 가우스를 활용한 온디바이스 AI 기술도 소개했으며, 생성 AI 모델을 다양한 제품에 단계적으로 탑재할 계획
- 삼성 가우스는 △텍스트를 생성하는 언어모델 △코드를 생성하는 코드 모델 △이미지를 생성하는 이미지 모델의 3개 모델로 구성
  - 언어 모델은 클라우드와 온디바이스 대상 다양한 모델로 구성되며, 메일 작성, 문서 요약, 번역 업무의 처리를 지원
  - 코드 모델 기반의 AI 코딩 어시스턴트 ‘코드아이(code.i)’는 대화형 인터페이스로 서비스를 제공하며 사내 소프트웨어 개발에 최적화
  - 이미지 모델은 창의적인 이미지를 생성하고 기존 이미지를 원하는 대로 바꿀 수 있도록 지원하며 저해상도 이미지의 고해상도 전환도 지원
- IT 전문지 테크리퍼블릭(TechRepublic)은 온디바이스 AI가 주요 기술 트렌드로 부상했다며, 2024년부터 가우스를 탑재한 삼성 스마트폰이 메타의 라마(Llama)2를 탑재한 퀄컴 기기 및 구글 어시스턴트를 적용한 구글 픽셀(Pixel)과 경쟁할 것으로 예상

출처 : 삼성전자, ‘삼성 AI 포럼’서 자체 개발 생성형 AI ‘삼성 가우스’ 공개, 2023.11.08.

삼성전자, ‘삼성 개발자 콘퍼런스 코리아 2023’ 개최, 2023.11.14.

TechRepublic, Samsung Gauss: Samsung Research Reveals Generative AI, 2023.11.08.

## 구글, 앤스로픽에 20억 달러 투자로 생성 AI 협력 강화

### KEY Contents

- 구글이 앤스로픽에 최대 20억 달러 투자에 합의하고 5억 달러를 우선 투자했으며, 앤스로픽은 구글과 클라우드 서비스 사용 계약도 체결
- 3대 클라우드 사업자인 구글, 마이크로소프트, 아마존은 차세대 AI 모델의 대표 기업인 앤스로픽 및 오픈AI와 협력을 확대하는 추세

### ● 구글, 앤스로픽에 최대 20억 달러 투자 합의 및 클라우드 서비스 제공

- 구글이 2023년 10월 27일 앤스로픽에 최대 20억 달러를 투자하기로 합의했으며, 이 중 5억 달러를 우선 투자하고 향후 15억 달러를 추가로 투자할 방침
- 구글은 2023년 2월 앤스로픽에 이미 5억 5,000만 달러를 투자한 바 있으며, 아마존도 지난 9월 앤스로픽에 최대 40억 달러의 투자 계획을 공개
- 한편, 2023년 11월 8일 블룸버그 보도에 따르면 앤스로픽은 구글의 클라우드 서비스를 위해 4년간 30억 달러 규모의 계약을 체결
- 오픈AI 창업자 그룹의 일원이었던 다리오(Dario Amodei)와 다니엘라 아모데이(Daniela Amodei) 남매가 2021년 설립한 앤스로픽은 챗GPT의 대항마 ‘클로드(Claude)’ LLM을 개발
- 아마존과 구글의 앤스로픽 투자에 앞서, 마이크로소프트는 차세대 AI 모델의 대표 주자인 오픈 AI와 협력을 확대
- 마이크로소프트는 오픈AI에 앞서 투자한 30억 달러에 더해 2023년 1월 추가로 100억 달러를 투자하기로 하면서 오픈AI의 지분 49%를 확보했으며, 오픈AI는 마이크로소프트의 애저(Azure) 클라우드 플랫폼을 사용해 AI 모델을 훈련

### ● 구글, 클라우드 경쟁력 강화를 위해 생성 AI 투자 확대

- 구글은 수익률이 높은 클라우드 컴퓨팅 시장에서 아마존과 마이크로소프트를 따라잡고자 생성 AI를 통한 기업 고객의 클라우드 지출 확대를 위해 AI 투자를 지속
- 구글은 앤스로픽 외에도 AI 동영상 제작 도구를 개발하는 런웨이(Runway)와 오픈소스 소프트웨어 기업 허깅 페이스(Hugging Face)에도 투자
- 구글은 챗GPT의 기반 기술과 직접 경쟁할 수 있는 차세대 LLM ‘제미니(Gemini)’를 포함한 자체 AI 시스템 개발에도 수십억 달러를 투자했으며, 2024년 제미니를 출시할 계획

출처 : The Wall Street Journal, Google Commits \$2 Billion in Funding to AI Startup Anthropic, 2023.10.27.  
Bloomberg, AI Startup Anthropic to Use Google Chips in Expanded Partnership, 2023.11.09.

## IDC, 2027년 AI 소프트웨어 매출 2,500억 달러 돌파 전망

### KEY Contents

- IDC의 예측에 의하면 AI 소프트웨어 시장은 2027년 2,510억 달러로 달할 전망이며, 생성 AI 플랫폼과 애플리케이션은 2027년까지 283억 달러의 매출을 창출할 전망
- 2023년 기준 AI 소프트웨어 매출의 3분의 1을 차지하는 최대 시장인 AI 애플리케이션은 2027년까지 21.1%의 연평균 성장률을 기록할 전망

### ● 기업들의 AI 투자 증가에 힘입어 AI 소프트웨어 시장 급성장 예상

- 시장조사기관 IDC는 AI 소프트웨어 시장이 2022년 640억 달러에서 2027년 2,510억 달러로 연평균 성장률 31.4%를 기록하며 급성장할 것으로 예상
- AI 소프트웨어 시장은 AI 플랫폼, AI 애플리케이션, AI 시스템 인프라 소프트웨어(SIS), AI 애플리케이션 개발·배포(AI AD&D) 소프트웨어를 포괄
- 협업, 콘텐츠 관리, 전사적 자원관리(ERM), 공급망 관리, 생산 및 운영, 엔지니어링, 고객관계관리(CRM)를 포함하는 AI 애플리케이션은 AI 소프트웨어의 최대 시장으로 2023년 전체 매출의 약 3분의 1을 차지하며 2027년까지 21.1%의 연평균 성장률을 기록할 전망
- AI 비서를 포함한 AI 모델과 애플리케이션의 개발을 뒷받침하는 AI 플랫폼은 두 번째로 시장 규모가 큰 분야로, 2027년까지 35.8%의 연평균 성장률이 예상됨
- 분석, 비즈니스 인텔리전스, 데이터 관리와 통합을 포함하는 AI SIS는 기존 소프트웨어 시스템과 통합되어 방대한 데이터를 활용한 의사결정과 운영 최적화를 지원하며, 현재 매출 규모는 비교적 작지만 5년간 연평균 성장률은 32.6%로 시장 전체를 웃돌 전망
- 애플리케이션 개발, 소프트웨어 품질과 수명주기 관리 소프트웨어, 애플리케이션 플랫폼을 포함하는 AI AD&D는 향후 5년간 카테고리 중 가장 높은 38.7%의 연평균 성장률이 예상됨
- IDC에 따르면 경제적 불확실성과 시장 역학의 변화에도 AI와 자동화 기술에 대한 기업들의 투자 의지는 확고하며, 기업들은 AI 도입이 사업 성공과 경쟁우위에 필수적이라고 인식
- IDC 설문조사에 따르면 향후 12개월 동안 응답자의 3분의 1은 기업이 특정 사용 사례나 응용 영역에서 외부 AI 소프트웨어의 구매를 고려하거나 외부 AI 소프트웨어와 내부 자원의 결합을 고려
- 한편, AI 소프트웨어 시장에 포함되지 않는 생성 AI 플랫폼과 애플리케이션은 2027년까지 283억 달러의 매출을 창출할 전망

☞ 출처 : IDC, IDC Forecasts Revenue for Artificial Intelligence Software Will Reach \$279 Billion Worldwide in 2027, 2023.10.31.



## 빌 게이츠, AI 에이전트로 인한 컴퓨터 사용의 패러다임 변화 전망

### KEY Contents

- 빌 게이츠가 5년 내 일상언어로 모든 작업을 처리할 수 있는 AI 에이전트가 보급되며 컴퓨터를 사용하는 방식이 완전히 바뀔 것으로 예상
- 에이전트의 보급은 컴퓨터 분야를 넘어 산업 전 영역에 영향을 미칠 전망으로 특히 의료와 교육, 생산성, 엔터테인먼트·쇼핑 영역에서 고가로 제공되던 서비스가 대중화될 전망

### ● 5년 내 기기에 일상언어로 말하기만 하면 되는 AI 에이전트의 보급 예상

- 빌 게이츠 마이크로소프트 창업자가 2023년 11월 9일 공식 블로그를 통해 AI 에이전트가 컴퓨터 사용방식과 소프트웨어 산업을 완전히 변화시킬 것이라는 전망을 제시
- 자연어에 반응하고 사용자에게 대한 지식을 바탕으로 다양한 작업을 수행하는 소프트웨어를 의미하는 에이전트는 컴퓨터 사용방식이 키보드 입력에서 아이콘 클릭으로 바뀐 이후 최대의 컴퓨팅 혁명을 가져올 전망
- 현재는 컴퓨터 작업 시 작업 내용에 따라 각각 다른 앱을 사용해야 하지만 5년 내 에이전트의 발전으로 기기에 일상언어로 말하기만 하면 되는 미래가 도래할 것
- 온라인에 접속하는 모든 사람이 AI 기반의 개인 비서를 사용할 수 있게 되며, 에이전트는 사용자에게 대한 풍부한 지식을 바탕으로 맞춤형 대응이 가능하며 시간이 지날수록 개선됨
- 일례로 여행 계획 수립 시 AI 챗봇이 예산에 맞는 호텔을 제안하는데 머문다면, 에이전트는 사용자의 여행 패턴을 분석해 여행지를 제안하고 관심사에 따른 활동을 추천하며 선호하는 스타일의 레스토랑 예약도 가능

### ● AI 에이전트가 의료와 교육, 생산성, 엔터테인먼트·쇼핑 영역의 서비스 대중화를 주도할 것

- 에이전트로 인해 주목할 만한 변화는 고비용 서비스의 대중화로 특히 △의료 △교육 △생산성 △엔터테인먼트·쇼핑의 4개 영역에서 대규모 변화 예상
- (의료) 에이전트가 환자 분류를 지원하고 건강 문제에 대한 조언을 제공하며 치료의 필요 여부를 결정하면서 의료진의 의사결정과 생산성 향상에 기여
- (교육) 에이전트가 1대 1 가정교사의 역할을 맡아 모든 학생에게 평등한 교육 기회를 제공할 수 있으며, 아이가 좋아하는 게임이나 노래 등을 활용해 시청각 기반의 풍부한 맞춤형 교육 경험을 제공
- (생산성) 사용자의 아이디어를 기반으로 에이전트가 사업계획과 발표 자료 작성, 제품 이미지 생성을 지원하며, 임원의 개인 비서와 같은 역할도 수행
- (엔터테인먼트·쇼핑) 쇼핑 시 에이전트가 모든 리뷰를 읽고 요약해 최적의 제품을 추천하고 사용자 대신 주문할 수 있으며 사용자의 관심사에 맞춤형된 뉴스와 엔터테인먼트를 구독 가능

## 유튜브, 2024년부터 AI 생성 콘텐츠 표시 의무화

### KEY Contents

- 유튜브가 몇 달 안에 생성 AI를 사용한 콘텐츠에 AI 라벨 표시를 의무화하기로 했으며, 이를 준수하지 않는 콘텐츠는 삭제하고 크리에이터에 대한 수익 배분도 중단할 수 있다고 설명
- 유튜브는 AI 생성 콘텐츠가 신원 파악이 가능한 개인을 모방한 경우 개인정보 침해 신고 절차에 따라 콘텐츠 삭제 요청도 받을 계획

### ○ 유튜브, 생성 AI 콘텐츠에 AI 라벨 표시 안 하면 콘텐츠 삭제

- 유튜브가 2023년 11월 14일 공식 블로그를 통해 몇 달 안에 생성 AI를 사용한 콘텐츠에 AI 라벨을 표시하는 새로운 규칙을 시행한다고 발표
- 실제로 일어나지 않은 사건을 사실적으로 묘사하거나 실제로 하지 않은 말이나 행동을 보여주는 콘텐츠와 같이 AI 도구를 사용해 사실적으로 변경되거나 합성된 콘텐츠에는 AI 라벨을 표시 필요
- 유튜브는 이러한 규칙이 선거나 분쟁 상황, 공중 보건, 공직자 관련 문제와 같이 민감한 주제를 다루는 콘텐츠에서 특히 중요하다고 강조했으며, 크리에이터가 AI로 제작한 콘텐츠에 AI 라벨을 표시하지 않으면 해당 콘텐츠는 삭제되고 광고 수익을 배분하는 유튜브 파트너 프로그램도 정지될 수 있음
- 유튜브는 두 가지 방식으로 AI를 이용한 콘텐츠의 변경이나 합성 여부를 시청자에게 전달할 계획으로 동영상 설명 패널에 라벨을 표시하는 방식이 기본이며, 민감한 주제를 다루는 특정 유형의 콘텐츠는 동영상 플레이어에 더욱 눈에 띄는 라벨을 적용
- 유튜브는 커뮤니티 정책에 위반되는 일부 합성 콘텐츠에 대해서는 라벨 지정 여부와 관계없이 삭제할 방침으로, 가령 사실적인 폭력을 보여주는 합성 동영상이 시청자에게 충격이나 혐오감을 줄 수 있다면 삭제될 수 있음

### ○ 유튜브, 특정인을 모방한 AI 생성 콘텐츠에 대한 삭제 요청에도 대응 계획

- 유튜브는 몇 달 내에 신원 파악이 가능한 개인의 얼굴이나 음성을 모방한 AI 생성 콘텐츠에 대하여 개인정보 침해 신고 절차를 마련해 삭제 요청을 받을 계획
- 단, 모든 콘텐츠가 삭제 대상은 아니며 유튜브는 콘텐츠가 패러디나 풍자인지, 해당 영상에서 삭제 요청을 한 특정인을 식별할 수 있는지, 공직자나 유명인이 등장하는지 등 다양한 요소를 고려할 예정
- 유튜브는 음반사가 아티스트의 고유한 노래나 목소리를 모방한 AI 생성 음악에 대하여 삭제를 요청할 수 있는 기능도 도입할 방침

출처 : Youtube, Our approach to responsible AI innovation, 2023.11.14.

## 영국 과학혁신기술부, AI 안전 연구소 설립 발표

### KEY Contents

- 영국 과학혁신기술부가 첨단 AI 시스템에 대한 평가를 통해 안전성을 보장하기 위한 AI 안전 연구소를 설립한다고 발표
- AI 안전 연구소는 핵심 기능으로 첨단 AI 시스템 평가 개발과 시행, AI 안전 연구 촉진, 정보교류 활성화를 추진할 계획

### ● 영국 AI 안전 연구소, 첨단 AI 시스템 평가와 AI 안전 연구, 정보 교류 추진

- 영국 과학혁신기술부가 2023년 11월 2일 첨단 AI 안전에 중점을 둔 국가 연구기관으로 AI 안전 연구소(AI Safety Institute)를 설립한다고 발표
  - AI 안전 연구소는 첨단 AI의 위험을 이해하고 거버넌스 마련에 필요한 사회·기술적 인프라 개발을 통해 영국을 AI 안전 연구의 글로벌 허브로 확립하는 것을 목표로 함
  - 영국 정부는 향후 10년간 연구소에 공공자금을 투자해 연구를 지원할 계획으로, 연구소는 △첨단 AI 시스템 평가 개발과 시행 △AI 안전 연구 촉진 △정보 교류 활성화를 핵심 기능으로 함
- (첨단 AI 시스템 평가 개발과 시행) 시스템의 안전 관련 속성을 중심으로 안전과 보안 기능을 이해하고 사회적 영향을 평가
  - 평가 우선순위는 △사이버범죄 조장, 허위 정보 유포 등 악의적으로 활용될 수 있는 기능 △사회에 미치는 영향 △시스템 안전과 보안 △인간의 통제력 상실 가능성 순
  - 연구소는 외부 기관과 협력해 자체 시스템 평가를 개발 및 수행하고, 평가와 관련된 의견 공유 및 지침 마련을 위해 전문가 커뮤니티를 소집할 계획
- (AI 안전 연구 촉진) 외부 연구자를 소집하고 다양한 예비 연구 프로젝트를 통해 AI 안전 기초연구를 수행
  - AI 시스템의 효과적 거버넌스를 위한 도구 개발\* 및 안전한 AI 시스템 개발을 위한 새로운 접근 방식 연구를 수행

\* 편향된 훈련 데이터에 대한 분석기술, 민감한 정보를 포함하는 AI 시스템에 대한 미세 조정 방법
- (정보 교류 활성화) 현행 개인정보보호와 데이터 규제 하에서 연구소와 정책입안자, 국제 파트너, 학계, 시민사회 및 일반 대중과 정보 공유 채널을 구축
  - AI 안전성 정상회의(AI Safety Summit)에서 합의된 대로 첨단 AI 모델의 평가 후 해당 모델이 배포된 타국의 정부 및 연구소와 평가 결과를 공유하고, 학계와 대중이 AI 시스템의 피해와 취약점을 보고할 수 있는 명확한 절차를 수립

출처 : Gov.uk, Introducing the AI Safety Institute, 2023.11.02.

Venturebeat, Researchers turn to Harry Potter to make AI forget about copyrighted material, 2023.10.06.

## 구글 딥마인드, 범용 AI 모델의 기능과 동작에 대한 분류 체계 발표

### KEY Contents

- 구글 딥마인드 연구진이 성능과 범용성, 자율성을 기준으로 범용 AI(AGI)의 수준을 0~5단계까지 총 6단계로 구분한 프레임워크를 공개
- 현재 AGI는 단백질 구조를 예측하는 알파폴드와 같은 특정 용도에서는 5단계 수준을 달성했지만 광범위하게 활용될 수 있는 범용에서는 1단계 수준에 머물러 있음

### ● 챗GPT와 구글 바드와 같은 AI 챗봇은 범용 AI 1단계 수준

- 구글 딥마인드 연구진은 2023년 11월 4일 범용 AI(Artificial General Intelligence, AGI) 모델을 용도와 성능에 따라 분류하는 프레임워크를 제시한 논문을 발표
  - 프레임워크의 목적은 AGI의 성능, 범용성, 자율성 수준을 정의하여 모델 간 비교와 위험 평가, AGI 달성까지의 진행 상황을 측정할 수 있는 공통 기준을 제공하기 위함
- 연구진은 AGI 개념 정의에 필요한 기준을 수립하기 위한 6가지 원칙을 아래와 같이 도출
  - (프로세스가 아닌 기능에 중점) AI가 어떻게 작동하는지보다 무엇을 할 수 있는지가 더 중요
  - (범용성과 성능을 모두 평가) 진정한 AGI는 인간을 능가하는 폭넓은 범용성과 기술의 깊이를 모두 요구
  - (인지와 메타인지 작업에 중점) 물리적 작업의 수행 능력은 AGI의 필수 전제조건이 아니며, 인지 작업과 메타인지 작업(예; 새로운 작업의 학습 능력, 인간에게 도움을 요청할 시점을 아는 능력)이 핵심
  - (실제 구현보다 잠재력에 집중) 통제된 상황에서 발휘되는 성능에 따라 AGI를 규정하고 테스트를 진행
  - (생태학적 타당도를 갖춘 벤치마크 사용) AGI에 대한 벤치마크는 사람들이 경제적·사회적 또는 예술적으로 가치 있게 여기는 실질적인 작업을 대상으로 성능 평가 필요
  - (종점이 아닌 AGI를 향한 경로에 중점) 단계별 접근방식을 통해 AGI의 발전 상태를 점진적으로 측정
- 연구진은 상기 원칙에 따라 AI를 성능에 따라 0~5단계와 광범위한 목적에 활용될 수 있는 범용 AI 및 특정 과업에 활용되는 특수 AI로 분류했으며, 특수 AI에서는 5단계까지 달성되었으나, 범용 AI는 현재 1단계 수준

#### 〈구글 딥마인드의 범용 AI 분류 프레임워크〉

| 성능                       | 특수 AI 예시  | 범용 AI 예시      |
|--------------------------|---|---------------|
| 0단계: AI 아님               | 계산기 소프트웨어, 컴파일러                                   | 아마존 메커니컬 터크   |
| 1단계: 신진(숙련되지 않은 인간)      | GOFAI(Good Old Fashioned Artificial Intelligence) | 챗GPT, 바드, 라마2 |
| 2단계: 유능(숙련된 인간의 50% 이상)  | 스마트 스피커(애플 시리, 아마존 알렉사, 구글 어시스턴트), IBM 왓슨         | 미달성           |
| 3단계: 전문가(숙련된 인간의 90% 이상) | 문법 교정기(그래머리), 생성 이미지 모델(달리2)                      | 미달성           |
| 4단계: 거장(숙련된 인간의 99% 이상)  | 답블루, 알파고  | 미달성           |
| 5단계: 초인간(인간을 100% 능가)    | 알파폴드, 알파제로, 스톡피시                                  | 미달성           |

출처 : Arxiv.org, Levels of AGI: Operationalizing Progress on the Path to AGI, 2023.11.04.

## 갈릴레오의 LLM 환각 지수 평가에서 GPT-4가 가장 우수




### KEY Contents

- 주요 LLM의 환각 현상을 평가한 'LLM 환각 지수'에 따르면 GPT-4는 작업 유형과 관계없이 가장 우수한 성능을 보였으며 GPT-3.5도 거의 동등한 성능을 발휘
- 오픈소스 모델 중에서는 메타의 라마2가 RAG 없는 질문과 답변 및 긴 형식의 텍스트 생성에서 가장 우수한 성능을 발휘

### ○ 주요 LLM 중 GPT-4가 가장 환각 현상 적고 GPT-3.5 터보도 비슷한 성능 기록

- 머신러닝 데이터 관리 기업 갈릴레오(Galileo)가 2023년 11월 15일 주요 LLM의 환각 현상을 평가한 'LLM 환각 지수(LLM Hallucination Index)'를 발표
  - 생성 AI의 환각 현상은 AI 시스템이 잘못된 정보를 생성하거나, 현실과 다른 부정확한 결과를 내놓는 현상으로, 기업의 AI 도입을 가로막는 주요 장애물이며, 환각 지수는 신뢰할 수 있는 생성 AI 구축을 위해 환각을 평가하고 측정하는 구조화된 접근방식을 제공
  - 환각 지수는 △검색 증강 생성(Retrieval-Augmented Generation, RAG)\*을 포함한 질문과 답변 △RAG 없는 질문과 답변 △긴 형식의 텍스트(보고서나 기사, 에세이) 생성의 3개 작업 유형에 대하여 환각을 기준으로 LLM의 순위를 평가
    - \* 기존에 학습된 데이터가 아닌 외부 소스(데이터셋, 데이터베이스, 문서 등)에서 가져온 정보를 검색해 활용하는 기술
- 3개의 작업 유형 평가 전체에서 오픈AI의 GPT-4가 최고의 성능을 기록했으며, GPT-3.5 터보도 GPT-4와 거의 동등한 성능을 발휘
  - 메타의 라마2(Llama-2-70b)는 RAG 없는 질문과 답변 유형에서 오픈소스 모델 가운데 가장 우수했고 긴 형식의 텍스트 생성에서도 GPT-4에 준하는 성능을 기록했으나, RAG 포함 질문과 답변에서는 허깅 페이스의 제퍼(Zephyr-7b)가 라마2를 능가

#### 〈갈릴레오의 LLM 환각 지수(RAG 포함 질문과 답변 기준)〉

| Developer   | Model                    | Context Adherence Score |
|---|--------------------------|-------------------------|
|  | gpt-4-0613               | 0.76                    |
|  | gpt-3.5-turbo-0613       | 0.75                    |
|  | gpt-3.5-turbo-1106       | 0.74                    |
|  | zephyr-7b-beta           | 0.71                    |
|  | gpt-3.5-turbo-instruct   | 0.68                    |
|  | llama-2-70b-chat         | 0.68                    |
|  | llama-2-13b-chat         | 0.68                    |
|  | mistral-7b-instruct-v0.1 | 0.67                    |
|  | llama-2-7b-chat          | 0.65                    |
|  | falcon-40b-instruct      | 0.60                    |
|  | mpt-7b-instruct          | 0.58                    |

## 영국 옥스퍼드 인터넷 연구소, AI 기술자의 임금이 평균 21% 높아

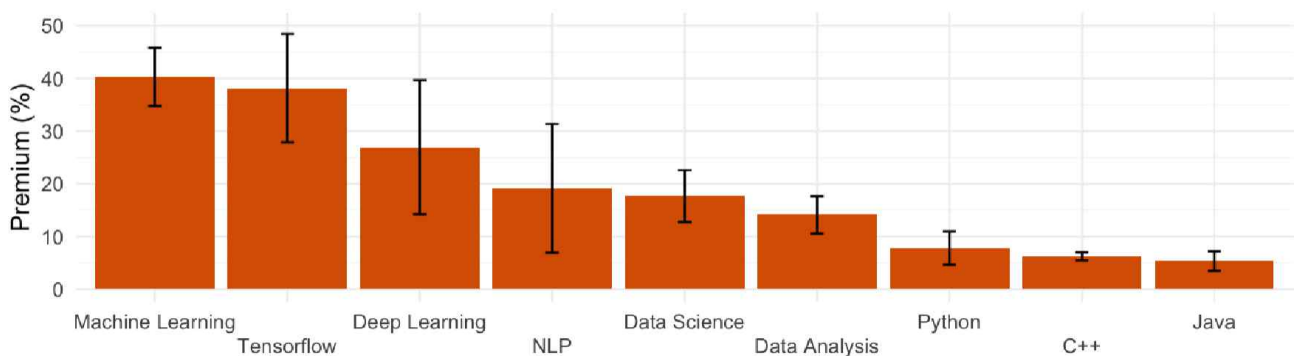
### KEY Contents

- 옥스퍼드 인터넷 연구소의 연구에 따르면 특정 기술의 경제적 가치는 다른 기술과 결합 가능성이 높을수록 높게 평가됨
- AI의 확산은 기술의 경제적 가치에 크게 영향을 미치며, AI 기술을 가진 근로자는 평균 21%, 최대 40% 높은 임금을 받을 수 있음

### ● AI 기술 중 머신러닝, 텐서플로우, 딥러닝의 임금 프리미엄이 높게 평가




- 옥스퍼드 인터넷 연구소(Oxford Internet Institute)가 2023년 10월 24일 962개 기술과 2만 5천 명을 대상으로 한 연구에서 AI를 포함한 주요 기술의 경제적 가치를 분석한 결과를 발표
- 연구에 따르면 한 기술의 경제적 가치는 근로자의 여타 역량과 얼마나 잘 결합하는지를 보여주는 '상보성(complementarity)'에 따라 결정됨
- 특정 기술은 다른 기술과 결합 가능성이 높을수록 경제적 가치가 높아지며, 일례로 데이터 분석과 같은 기술은 여타 고부가가치 기술과 결합할 수 있어 가치가 높지만, 사진 리터칭 같은 기술은 특정 기술과만 결합할 수 있어 가치가 낮게 평가됨
- 대부분 직업은 여러 기술의 조합이 필요하며, 근로자의 재교육에서 경제적 효율성을 높이려면 기존 기술과 신기술 간 상보성을 극대화할 필요
- AI의 확산은 기술의 경제적 가치에 크게 영향을 미치는 요소로, AI 기술을 가진 근로자는 평균적으로 21% 높은 임금을 획득 가능
- AI 기술 중 근로자에 대한 경제적 가치(시간당 임금 증가율 기준) 측면에서 상위 5개 기술은 머신러닝(+40%), 텐서플로우(+38%), 딥러닝(+27%), 자연어처리(+19%), 데이터 과학(+17%) 순

〈AI 기술 유형 평균 기술 대비 갖는 임금 프리미엄〉



출처 : Oxford Internet Institute, AI comes out on top: Oxford Study identifies the economic value of specific skills, 2023.10.24.

## II. 주요 행사 일정

| 행사명  | 행사 주요 개요  |   |   |
|--|---|---|---|
| CES 2024                                   |    | <ul style="list-style-type: none"> <li>- 미국 소비자기술 협회(CTA)가 주관하는 세계 최대 가전·IT·소비재 전시회로 5G, AR&amp;VR, 디지털헬스, 교통·모빌리티 등 주요 카테고리 중심으로 기업들이 최신의 기술 제품군을 전시</li> <li>- CTA 사피로 회장은 가장 주목받는 섹터로 AI를 조명하였으며, 모든 산업을 포괄한다는 의미에서 ‘올 온(AI on)’을 주제로 한 이번 전시에는 500곳 이상의 한국기업 참가 예정</li> </ul> |   |
|  | 기간  | 장소  | 홈페이지  |
|  | 2024.1.9~12   | 미국, 라스베이거스  | <a href="https://www.ces.tech/">https://www.ces.tech/</a>                         |
| AIMLA 2024                                 |  | <ul style="list-style-type: none"> <li>- 머신러닝 및 응용에 관한 국제 컨퍼런스(AIMLA 2024)는 인공지능 및 머신러닝의 이론, 방법론 및 실용적 접근에 관한 지식과 최신 연구 결과 공유</li> <li>- 이론 및 실무 측면에서 인공지능, 기계학습의 주요 분야를 논의하고, 학계, 산업계의 연구자와 실무자들에게 해당 분야의 최첨단 개발 소식 공유</li> </ul>  |   |
|  | 기간  | 장소  | 홈페이지  |
|  | 2024.1.27~28  | 덴마크, 코펜하겐   | <a href="https://ccnet2024.org/aimla/index">https://ccnet2024.org/aimla/index</a> |
| AAAI Conference on Artificial Intelligence |  | <ul style="list-style-type: none"> <li>- AI 발전 협회 컨퍼런스(AAAI)는 AI 연구를 촉진하고, AI 분야 연구원, 실무자, 과학자, 학생 및 공학자 간 교류의 기회 제공</li> <li>- 컨퍼런스에서 AI 관련 기술 발표, 특별 트랙, 초청 연사, 워크숍, 튜토리얼, 포스터 세션, 주제 발표, 대회, 전시 프로그램 등 진행</li> </ul>   |   |
|  | 기간  | 장소  | 홈페이지  |
|  | 2024.2.20~27  | 캐나다, 밴쿠버  | <a href="https://aaai.org/aaai-conference/">https://aaai.org/aaai-conference/</a> |





홈페이지 : <https://spri.kr/>

보고서와 관련된 문의는 AI정책연구실(jayoo@spri.kr, 031-739-7352)으로 연락주시기 바랍니다.

경기도 성남시 분당구 대왕판교로 712번길 22 글로벌 R&D 연구동(A) 4층

22, Daewangpangyo-ro 712beon-gil, Bundang-gu, Seongnam-si, Gyeonggi-do, Republic of Korea, 13488