2026년 8월호

SPRi AI Brief

인공지능 산업의 최신 동향

SPRi 소프트웨어정책연구소
Software Policy & Research Institute

SPRi AI Brief

Ⅰ 2026년 8월호 Ⅰ

CONTENTS

2026년 7월 월간 AI 모델 현황

주요 모델 출시 타임라인

7/7 Muse Image 7/9 GPT-5.6, Muse Spark 1.1, Reve 2.1
7/16 Grok 4.5, Kimi K3 7/21 Gemini 3.6 Flash/3.5 Flash-Lite/3.5 Flash Cyber, Fugu Cyber, Motif-3 Preview 7/23 Solar Open 2, FLUX 3 7/24 Claude Opus 5

LLM간 성능 비교(Artificial Analysis Intelligence Index 4.0)

모델	출시일	Index	Price ($/1m Token)		Speed	Context Window	회사
			Input	Output			
① Claude Opus 5 (max)	26.7.24.	61	5	25	52.8	1m	Anthropic
② Claude Opus 5 (xhigh)	26.7.24.	60	5	25	53.1	1m	Anthropic
③ Claude Fable 5(with fallback)	26.6.9.	60	10	50	69.7	1m	Anthropic
GPT-5.6 Sol (max)	26.7.9.	59	5	30	75.9	1m	OpenAI
Claude Opus 5 (high)	26.7.24.	59	5	25	58.0	1m	Anthropic

* 출처 : Artificial Analysis 평가체계 : Agents 25%, Coding 25%, General 25%, Scientific 25% (10개 벤치마크 종합)

이미지 생성 모델

모델	출시일	ELO	회사
① GPT Image 2 (high)	26.4.21	1337	OpenAI
② Reve 2.1	26.7.9.	1301	Reve
③ MAI-Image-2.5	26.6.2.	1269	Microsoft

영상 생성 모델

모델	출시일	ELO	회사
① Gemini Omni Flash	26.5.19.	1247	Google
② Dreamina Seedance 2.0 720p	26.2.10.	1229	ByteDance
③ Wan2.7-260612	26.6.12.	1165	Alibaba

한국 AI 모델 현황

■ 모티프와 업스테이지, 정부 독자 AI 파운데이션 모델 사업 성과로 신규 모델 공개

- 모티프의 Motif-3 Preview, 아티피셜 애널리시스 평가에서 44점으로 DeepSeek-V4 프로와 동급
- 업스테이지의 Solar Open 2, 에이전트와 한국어, 지식, 코딩 벤치마크에서 글로벌 수준 성능 달성

주요 포인트 :
오픈AI, 앤트로픽, 구글 등 주요 AI 기업이 최신 플래그십 모델을 잇달아 발표하며 프런티어 모델 경쟁 심화
Gemini 3.5 Flash Cyber, Fugu-Cyber 등 사이버보안 특화 모델 출시가 새로운 흐름으로 부상

SPRi AI Brief

Ⅰ2026년 8월호Ⅰ

정책·법제

미국 트럼프 행정부, 첨단 AI 모델 출시 개입 확대

KEY Contents

- 미국 트럼프 행정부가 앤트로픽의 Fable 5와 Mythos 5 수출통제 조치에 이어 오픈AI의 최신 AI 모델도 출시 범위를 소수의 신뢰할 수 있는 파트너로 제한할 것을 요구
- 트럼프 행정부는 2주 만에 GPT-5.6에 대한 출시 제한을 풀고 일반 공개를 허용했으며, 안전 조치를 강화한 앤트로픽의 Fable 5와 Mythos 5에 대해서도 수출통제를 해제

미국 정부, 오픈AI 최신 AI 모델 'GPT-5.6 Sol'의 초기 출시 범위 제한

■ 미국 트럼프 행정부가 오픈AI와 앤트로픽 등 주요 AI 기업의 최신 AI 모델의 출시 전 정부 승인을 요구하며 장기 규제 체계가 마련되기 전까지 출시 범위를 직접 제한하는 정책 기조를 수립

- 트럼프 대통령은 재집권 당시 바이든 정부의 AI 안전 기준을 비판하며 규제 완화를 내세웠으나, 취약점을 스스로 찾아내는 AI의 등장이 워싱턴 안팎에 충격을 주면서 입장을 선회
- AI 업계와 정치권은 정부가 명확한 기준 없이 출시 범위를 정하는 방식이 지나친 개입이라고 반발
- 프랑스에서 지난 6월 열린 G7 정상회의에서도 프랑스와 캐나다 정상은 정부 요구에 따라 외국인의 접근이 중단될 수 있는 미국 AI 기업에 대한 과도한 의존에 우려를 표시

■ 오픈AI는 2026년 6월 26일 트럼프 행정부의 요구에 따라 최신 플래그십 모델 'GPT-5.6 Sol'의 초기 공개 대상을 소수의 '신뢰할 수 있는 파트너'로 제한한다고 발표

- 오픈AI는 블로그를 통해 이러한 정부 접근 절차가 장기적으로 표준이 되어서는 안 된다면서, 사용자와 개발자, 기업, 사이버 방어 조직, 글로벌 협력사가 필요한 도구에 접근하지 못할 수 있다고 강조
- 이후 미국 정부가 모델 테스트 및 오픈AI와 협의를 거쳐 GPT-5.6에 대한 광범위한 출시 제한을 해제하면서 오픈AI는 2026년 7월 9일 모델을 정식 출시

상무부, 안전 조치가 적용된 Mythos 5와 Fable 5의 수출통제 조치 해제

■ 한편, 상무부는 앤트로픽에 보낸 비공개 서한을 통해 기존 수출통제 조치로 서비스가 전면 중단되었던 'Mythos 5' 모델을 미국 내 신뢰할 수 있는 파트너에만 제공할 수 있도록 허용

- 상무부는 2026년 6월 26일 승인된 기업 소속의 외국 국적자도 Mythos 5를 사용할 수 있다고 통보했으며, 승인된 기업은 100여 개로 알려졌으나 기업 명단은 미공개
- 상무부는 해당 서한에서 신뢰할 수 있는 파트너들의 Mythos 5 사용에 필요한 안전 조치가 충분히 마련되었다고 평가했으며, 정부가 언제든 승인 대상 목록을 변경할 수 있다는 내용도 포함
- 이후 상무부는 6월 30일에 Claude Mythos 5와 Fable 5의 수출통제를 해제했으며, 이에 따라 앤트로픽은 7월 1일부터 모델 접근 권한 복원을 시작해 순차적으로 서비스를 정상화

출처 | The Washington Post, The U.S. government will decide who gets to use the latest American AI technology, 2026.06.26.
Anthropic, Redeploying Fable 5, 2026.6.30.

미국 의회와 행정부, 중국산 AI 모델 채택 저지 전략 검토

KEY Contents

- 미국 하원이 자국 AI 모델보다 저렴하면서 성능 격차가 줄어든 중국산 AI 모델 채택 확산에 관한 조사를 진행 중인 가운데, 이에 대응한 미국의 AI 전략의 전반적 검토
- 오픈웨이트 특성상 중국산 AI 모델의 전면 금지가 사실상 불가능하다는 점에서 전문가들은 연방 조달 제한 및 중국산 AI 모델 관련 위험 정보 공유를 현실적 대안으로 제시

미국 하원 합동위원회, 자국기업의 중국산 AI 모델 도입 증가에 관한 조사 진행

■ 미국 의회가 자국 모델 대비 비용이 저렴하면서 성능 격차를 좁혀나가고 있는 중국산 AI 모델의 미국 내 채택 확산을 억제할 전략을 모색하면서 관련 조사를 진행

- 하원 국토안보위원회와 미·중 전략경쟁 특별위원회는 2026년 4월 중국산 AI 모델의 도입 증가에 대한 공동 조사 계획을 밝혔으며, 1차로 커서(Cursor)와 에어비앤비(Airbnb)에 중국산 AI 모델 사용에 대한 질의 서한을 발송
- 이와 관련, 에어비앤비는 자사의 AI 활동은 대부분 미국산 모델을 활용하고 중국산 모델은 일부만 사용하며, 승인된 미국 기반 서비스 제공업체를 통해 운영과 데이터를 분리해 보호한다고 해명
- 하원 국토안보위원회의 앤드류 가바리노(Andrew Garbarino) 위원장은 "중국이 AI 분야에서 미국을 추격할 뿐 아니라 미래 사이버보안의 향방을 결정할 핵심 역량의 격차 축소에 나서고 있다"고 지적
- 그는 최근 중국의 오픈웨이트 모델이 특정 취약점 발견과 사이버보안 작업에서 미국의 주요 AI 모델과 동등한 수준을 보인다는 언론 보도에 심각한 우려를 표시
- 미국 국무부 대변인은 중국산 AI 모델이 특정 이념과 가치를 반영한다고 주장하며, 자국 기업의 중국산 AI 모델 채택에 대한 우려 표명

중국산 AI 모델 채택 저지를 위해 미국의 개방형 AI 전략 강화 필요를 주장

■ 현재 진행 중인 하원 합동위원회 조사에서는 중국산 AI 모델의 부상뿐 아니라 미국이 이에 대응해 충분한 노력을 기울이고 있는지도 검토를 진행

- 한 위원회 관계자는 비싸거나 기능이 제한된 미국산 모델과 저렴하고 성능이 뛰어난 중국산 모델 중 양자택일해야 하는 상황을 피할 수 있도록 미국의 오픈웨이트 AI 전략이 적합한 검토 중이라고 언급
- 사이버보안·인프라보호 소위원장 앤디 오글스(Andy Ogles)는 "아무런 조치도 취하지 않으면 검열 기능이 탑재되고 안전장치가 제거된 중국 모델이 글로벌 디지털 경제의 기본 토대가 될 것"이라고 경고

■ 전문가들은 모델 가중치가 인터넷에 공개되는 오픈웨이트 모델의 특성상 중국산 AI 모델의 전면 금지는 불가능하며, 수정 헌법 제1조 표현의 자유와 관련된 문제를 초래할 수 있다고 지적

- 신미국안보센터(CNAS)의 대니얼 렘러(Daniel Remler) 선임연구원은 중국 AI 모델 규제가 스타트업에 타격을 주거나 개방형 모델에 대한 지원을 위축시킬 수 있다는 우려를 제기
- 대안으로 정부 조달 요건을 통해 기업의 중국산 AI 모델 사용을 억제하거나, 중국산 AI 모델의 위험이나 취약점 관련 조사 결과를 미국 기업에 공유하는 방안을 제시

출처 | CNBC, Lawmakers probe growing use of Chinese AI models in U.S. companies, 2026.7.8.

EU 집행위원회, 사이버보안·AI 실행 계획 발표

KEY Contents

- EU 집행위원회가 첨단 AI 모델의 사이버 악용 위험에 대응하기 위한 EU 차원의 체계적 대응 방안을 담은 '사이버보안·AI 실행 계획'을 발표
- EU 차원의 AI 모델 평가 역량 구축과 함께 사이버보안 특화 첨단 AI 모델에 체계적으로 접근하기 위한 청사진 수립, 사이버보안 특화 AI 테스트, 사이버보안 역량 강화 등을 모색

사이버보안 분야에서 첨단 AI 모델의 위험에 맞서 체계적 대응 모색

■ EU 집행위원회가 2026년 7월 7일 유럽의 사이버 안보를 강화하면서 안전하고 책임 있는 AI 사용을 지원하기 위한 '사이버보안·AI 실행 계획'을 발표

- 이번 계획은 ▲안전하고 책임 있는 첨단 AI 사용 촉진 ▲유럽의 사이버보안과 복원력 강화 ▲유럽의 사이버보안을 위한 AI 역량 강화라는 3개의 상호보완적 목표 달성을 추구
- EU 고유의 AI 및 사이버보안 법적 체계를 바탕으로 사이버보안 분야에서 첨단 AI 모델이 제기하는 위험에 대응하고 기회를 활용하기 위한 체계적인 대응 방안을 마련

■ (AI 모델 평가) 첨단 AI 모델의 안전하고 책임 있는 도입을 보장하기 위해 유럽 시장에서 모델 출시에 앞서 EU 내부에서 강력한 평가 역량을 구축

- 「AI 법」에 따라 첨단 AI 모델은 EU 시장 출시 전 평가를 거치고 완화 조치를 검토받아야 하며, EU 집행위원회가 감독·집행 권한을 행사해 사이버 위험 식별과 완화 조치의 적정성을 점검
- EU 내부의 평가 전문성을 육성하기 위해 EU 집행위원회는 사이버보안 분야를 포함하는 EU 평가 역량 구축을 위한 특별 공모를 2027년 가동 목표로 추진

■ (첨단 AI 모델 접근) 사이버보안 목적의 첨단 AI 시스템에 접근할 수 있는 명확하고 체계적인 조건을 마련하기 위해 EU 사이버보안청(ENISA)과 함께 유럽 차원의 청사진을 구상

- 청사진은 EU 기관, 회원국, 핵심 인프라 운영자, 사이버보안 제공자, 연구 주체 등 조직 유형별 접근 기준과 식별 방식, 보안 요건, 접근 제한·철회 시 EU 차원의 비상조치 등을 포함

■ (사이버보안 특화 AI 테스트) ENISA와 EU 집행위원회 공동연구센터(JRC)는 취약점 스캐닝·복구·사고 대응 등 사이버보안 활용 사례를 검증하는 AI 테스트 플랫폼을 구축

- 시뮬레이션 환경을 포함한 다양한 환경을 활용한 테스트 플랫폼 구축으로 금융, 에너지, 교통, 공공행정 등 핵심 분야의 종사자들에게 AI의 안전한 활용에 관한 노하우 제공을 모색

■ (사이버보안 강화와 역량 확대) AI 오용으로 인한 취약점으로부터 핵심 인프라 보호를 강화하고 사이버보안 역량 확대를 위한 AI EU 그랜드 챌린지를 출범

- ENISA는 지침과 권고, 모범 사례를 통해 기업의 취약점 식별과 수정, 사이버 공격 예방과 대응을 지원
- AI EU 그랜드 챌린지를 통해 기업, 연구소, 기관 등에 사이버보안 특화 AI 솔루션 개발을 위한 협력의 장을 제공하는 한편, EU AI 팩토리 인프라를 활용해 자체 첨단 AI 역량 개발에 지속적으로 투자

출처 | European Commission, Commission presents EU Action Plan on Cybersecurity and Artificial Intelligence, 2026.7.7.

RAND 연구소, 정부 규제를 보완할 기업과 시민사회의 역할 제시

KEY Contents

- RAND 연구소가 AI의 위험 관리에서 정부의 한계를 보완할 수 있는 기업과 시민사회의 3대 역할로 기술·운영 위험 관리, 시장 유인 조성, 사회 안정과 신뢰 지원을 제시
- 정부는 민간의 위험 관리 데이터를 규제 설계에 활용하고 공공 조달을 통해 안전 기준을 시장 표준으로 확산하며 인력 전환 비용을 공동 부담함으로써 기업과 시민사회를 지원 가능

정부 규제를 보완하는 기업과 시민사회의 3대 역할과 실행 방안 제안

■ 미국의 대표적 싱크탱크인 RAND 연구소가 정부 규제만으로는 AI의 안전 위험과 사회적 위험 관리에 한계가 있다는 문제의식에 따라 기업과 시민사회의 보완적 역할을 제시한 보고서를 발표

- 보고서에 따르면 AI 연구와 활용이 민간 기업에 집중되고 기술 발전 속도가 정책과 규제 주기를 앞지르는 상황에서, AI의 영향은 대개 국가의 직접 통제를 벗어나 정부 단독으로는 시의적절한 관리에 한계 노출
- 연구진은 심층적 문헌 검토와 기업·시민사회 관계자 대상 인터뷰, 자체 시뮬레이션 결과를 바탕으로 비정부기구(NGO)가 AI 위험을 완화할 수 있는 세 가지 역할을 제안

■ (역할 1: 기술·운영 위험 관리) 최첨단 AI 개발업체와 클라우드·인프라 제공업체, 고위험 배포 업체를 포함한 AI 커뮤니티는 AI 개발과 배포 과정에서 발생하는 기술·운영 위험 관리 역할을 담당

- AI 관련 핵심 기업들은 별도의 AI 위험 관리 체계와 공식 보고 체계를 구축하고 안전성 검증 자료를 문서화
- 규모가 작은 배포 기업이나 하위의 도입 기업은 경영진의 책임 소재 명확화, AI 활용 사례 목록, 활용 사례별 위험 등급 분류, 중요 응용 프로그램에 대한 인간의 개입 절차 마련 등 간소화 체계를 구축
- 정부는 민간에서 축적된 사고와 위험 데이터를 규제 설계에 활용하고, 프런티어 모델을 시험할 수 있는 샌드박스나 양해각서(MOU)를 통해 AI 커뮤니티의 역할을 뒷받침

■ (역할 2: 시장 유인 조성) AI 관련 주요 구매자와 투자자, 보험사 및 업계 컨소시엄은 안전한 AI를 개발하는 기업을 시장에서 더 유리하게 만드는 메커니즘을 채택

- 조달과 실사, 보증 절차에 안전 기준을 통합해 강력한 위험 관리 관행을 채택하는 AI 개발업체에 보상을 제공하고 시장과 업계 전반에서 공통 표준 확산을 지원
- 가령 금융사 등 구매자는 계약 시 안전성 검증을 요구하거나 병원 컨소시엄이 안전성이 검증되지 않은 제품을 집단으로 거부할 수 있으며, 보험사는 안전 조치를 갖춘 기업에 보험료를 인하
- 정부는 공공 조달을 활용해 안전 기준을 시장 표준으로 확산시키고, 책임소재나 사고 신고 기준을 정비해 보험시장이 AI 안전에 인센티브를 제공하도록 유도하는 방식으로 개입 가능

■ (역할 3: 사회 안정과 신뢰 지원) AI로 인한 일자리 감소 등 사회적 충격에 대응한 시민사회의 역할 제시

- 고용주, 자선단체, 노동조합, 지역 사회 단체 등은 인력 전환 계획과 재교육·재배치를 모색하고, AI 관련 충격에 맞서 지역 사회의 회복력을 강화하며, AI의 위험과 이점에 대한 정보 공유를 지원
- 정부는 인력 전환 비용을 공동 부담하고, AI로 창출된 이익의 일부를 지역 사회와 공유하는 정책 경로를 만들며, 국제 AI 사고 정보 공유 네트워크 구축을 지원하는 방식으로 개입

출처 | RAND, Looking Beyond the Government's Regulatory Toolkit, 2026.6.24.

백악관, AI로 과학 발견 앞당기는 제네시스 미션에 50억 달러 투입 계획

KEY Contents

- 백악관이 2025년 11월 시작한 제네시스 미션을 기존 에너지부 외에 보건복지부, 국방부, 국립과학재단 등 15개 연방 기관이 참여하는 범정부 프로젝트로 확장한다고 발표
- 에너지부가 구축한 '미국 과학 안보 플랫폼'을 공유 인프라로 활용해 연방 기관이 보유한 방대한 데이터와 컴퓨팅 자원, AI 도구를 하나로 연결해 연구자를 지원할 계획

제네시스 미션, 15개 이상 연방 기관이 참여하는 범정부 프로젝트로 확대

■ 백악관이 2026년 7월 22일 과학적 발견에 AI를 활용하는 '제네시스 미션(Genesis Mission)*' 확대에 50억 달러 이상의 연방 자금을 투입한다고 발표

* AI를 활용해 국가 과학 연구를 가속화하는 대규모 국가 프로젝트로 트럼프 대통령의 2025년 11월 행정명령에 따라 출범

- 제네시스 미션은 2025년 11월 출범 당시 에너지부(DOE)가 중심이 되었으나, 이번 발표로 15개 이상 연방 기관이 연구 자금과 데이터셋, 연구시설을 제공하는 범정부 프로젝트로 확장
- 각 기관은 에너지부가 구축한 '미국 과학 안보 플랫폼(American Science and Security Platform)'을 공유 인프라로 활용하며, 이 플랫폼은 연구자와 데이터·컴퓨팅·AI 도구를 연결해 과학 발견을 가속화
- 연구자들은 이 플랫폼을 통해 기관 간 장벽 없이 데이터와 컴퓨팅 자원을 공유해 과학 연구를 수행할 수 있으며, 백악관은 AI를 활용해 과학 연구 방식이 혁신적으로 변화될 것으로 기대

의료·생명과학 분야 혁신과 제조업 경쟁력 강화 등 다양한 분야의 과학기술 과제 제시

■ 백악관이 이번에 발표한 제네시스 미션의 국가 과학기술 과제는 의료·생명과학, 에너지·인프라, 제조업 경쟁력 회복, 신흥 위협 대응과 국가안보 등 다양한 분야를 포괄

- 미국 보건복지부(HHS)는 국가 건강 추적 데이터와 환경보호청(EPA)의 화학물질 모니터링 데이터, 국립과학재단(NSF)의 생명과학 연구 결과를 통합해 만성질환의 복합적 원인 규명을 추진
- 소아암 분야에서 전국 암센터 네트워크와 희귀 소아암 데이터를 통합한 연구 생태계를 구축하고 에너지부 슈퍼컴퓨터를 활용해 희귀암 유형 분석 및 암 발생 메커니즘 규명, 새로운 치료법 개발을 기대
- 제조업, 특히 반도체 산업의 경쟁력 회복을 위해 에너지부는 첨단 AI 모델과 반도체 생산 데이터를 결합한 '풀스택 공동 설계' 생태계를 구축해 차세대 반도체 기술 개발 속도를 획기적으로 높인다는 구상

■ 우주 데이터 활용, AI 기반 자율 연구실 구축, 양자컴퓨팅 실용화 등 다양한 과학 분야의 성과도 기대

- 미국 항공우주국(NASA)와 에너지부는 망원경과 위성, 탐사선, 지구 관측 시스템 등에서 축적된 방대한 데이터를 AI로 분석해 우주 물리학뿐 아니라 지구과학, 기후 연구, 생명과학 분야의 새로운 발견 모색
- 에너지부와 보건복지부, 국립과학재단 등은 로봇과 AI, 센서 기술을 통합해 AI가 실험을 설계하고 로봇이 실험 진행 뒤 결과를 분석해 다음 실험을 제안하는 자율 연구실 구축을 추진
- 에너지부와 국방부는 AI를 활용해 양자컴퓨팅과 센싱, 통신 기술을 연구실 단계에서 실제 활용 단계로 가속화할 계획으로, 세계 최초의 내결함성 양자 컴퓨터(FTQC) 개발을 목표로 제시

출처 | The White House, Trump Administration Announces More Than $5 Billion for the Genesis Mission, a National Mission on AI for Science, 2026.7.22.

SPRi AI Brief
Ⅰ2026년 8월호Ⅰ

기업·산업

앤트로픽, 알리바바의 클로드 데이터 무단 탈취 주장

KEY Contents

- 앤트로픽이 미 의회에 보낸 서한에서 알리바바가 대규모의 허위 계정을 생성해 클로드와 대량의 대화를 주고받아 클로드의 핵심 역량을 불법적으로 추출했다고 주장
- 백악관의 경고에도 알리바바의 클로드 무단 접근은 지속적으로 이루어졌으며, 앤트로픽의 의혹 제기 이후 알리바바는 보안 위험을 이유로 자사 직원들의 클로드 사용을 금지

앤트로픽, 알리바바의 클로드 무단 접근을 미 의회에 신고

■ 블룸버그 통신의 2026년 6월 24일 최초 보도에 따르면 앤트로픽은 미 의회에 보낸 서한에서 알리바바가 2만 5천 개의 허위 계정으로 클로드와 2,880만 건의 대화를 주고받았다고 주장

- 앤트로픽은 미국 상원 의원들과 백악관 관계자들에게 보낸 서한에서 알리바바가 중국 기업에 제공되지 않는 클로드에 접근하기 위해 허위 계정을 대규모로 생성했다고 주장
- 이를 클로드 역량을 불법적으로 추출한 최대 규모의 활동으로 규정하며, 알리바바의 불법 활동이 에이전틱 추론과 소프트웨어 엔지니어링, 장기 과제 수행 등 클로드의 핵심 역량을 집중적으로 노렸다고 강조

■ 앤트로픽이 특정한 공격 시점은 2026년 4월 말에서 6월 초 사이로, 알리바바뿐 아니라 딥시크, 문샷, 미니맥스도 2만 4천 개의 허위 계정으로 총 1,600만 건의 대화를 생성

- 앤트로픽은 중국 기업들이 적대적 '증류(Distillation)*' 방식으로 막대한 비용과 연구개발 노력이 투입된 미국 주요 AI 모델의 성과를 탈취해 자사 AI 모델을 개발하고 있다고 주장

* 다른 AI 모델의 출력 결과를 활용해 자사 시스템을 훈련함으로써 훨씬 낮은 비용으로 유사한 성능을 구현하는 기법

■ 백악관 기술정책실은 2026년 4월 중국 기업들이 미국 AI 모델의 출력물을 악용하려는 시도를 차단하겠다는 메모를 발표

- 해당 메모는 이러한 행위가 수천 개의 허위 계정을 활용하는 '산업적 규모'의 행위라는 점에서 합법적 연구와 다르다고 규정하고, 악의적 행위자에게 책임을 물을 것이라고 강조
- 알리바바는 중국 인민해방군과의 연관성을 전면 부인했으나 불법 증류 의혹에 대해서는 미해명

알리바바, 보안 위험을 이유로 내부 직원의 클로드 사용을 금지

■ 한편, 알리바바는 2026년 7월 10일부터 보안 위험을 이유로 내부 직원들의 클로드 사용을 금지하기로 하고, 이를 대신해 자체 코딩 플랫폼인 'Qoder' 사용을 지시

- 알리바바의 내부 관계자들에 따르면 클로드 코드에 백도어 보안 위험이 확인되어, 종합 평가를 거쳐 클로드 코드를 고위험 소프트웨어 목록에 추가
- 앤트로픽은 중국 기업 및 산하 외국 법인의 자사 모델 사용을 금지해 왔으며, 이를 차단하기 위한 수단 중 하나로 중국 사용자를 비밀리에 식별할 수 있는 클로드 버전을 배포한 것으로 확인

출처 | Financial Times, Anthropic accuses Alibaba of obtaining illicit access to Claude, 2026.6.26.
TechCrunch, Alibaba reportedly bans employees from using Claude Code, 2026.7.4.

앤트로픽, ‘Claude Sonnet 5’에 이어 ‘Claude Opus 5’ 출시

KEY Contents

- 앤트로픽이 기존 Sonnet 계열 모델 중에서 가장 강력한 자율 에이전트 성능을 제공하는 동시에 사이버보안 성능을 의도적으로 낮춘 ‘Claude Sonnet 5’를 공개
- 앤트로픽은 이후 최고 성능 모델 Claude Fable 5에 근접한 성능을 제공하면서 사용 비용은 절반 수준으로 낮춘 ‘Claude Opus 5’도 출시

Claude Sonnet 5, Opus 4.8에 근접한 성능에 비용 효율성 대폭 개선

■ 앤트로픽이 2026년 6월 30일 지금까지 출시된 Sonnet 계열 모델 중 가장 강력한 에이전트 성능을 갖춘 차세대 AI 모델 ‘Claude Sonnet 5’를 공개

- Claude Sonnet 5는 복잡한 작업 수행을 위해 스스로 계획을 수립하고 웹 브라우저와 터미널 등 다양한 도구를 활용할 수 있으며, 높은 수준의 자율성을 바탕으로 작업 완료가 가능
- 앤트로픽에 따르면 Claude Sonnet 5의 성능이 자사 최상위 모델 Claude Opus 4.8에 근접하며, 특히 추론 능력, 도구 사용, 코딩, 연구·분석 등 핵심 에이전트 성능에서 Sonnet 4.6을 능가*

* Terminal-Bench 2.1(에이전트 코딩) 기준 Opus 4.8(82.7%), Sonnet 5(80.4%), Sonnet 4.6(67.0%)

■ 앤트로픽은 에이전트 검색 평가 도구 BrowseComp 평가 결과를 토대로 Claude Sonnet 5의 비용 효율성을 강조하며 Opus 4.8보다 훨씬 넓은 범위의 비용 대비 성능 옵션을 제공한다고 설명

- 중간 수준의 추론 강도에서 Opus 4.8 대비 비용 효율성이 크게 향상된 한편, 고난도 추론 설정에서는 일부 작업에서 Opus 4.8과 동등한 성능을 달성

■ 내부 안전성 평가 결과, Claude Sonnet 5는 Sonnet 4.6보다 바람직하지 않은 행동 발생률이 전반적으로 감소하고 환각 및 아첨 행위 발생률도 낮았으며 에이전트 환경의 안정성도 향상

- 또한 별도의 사이버보안 특화 훈련을 받지 않아 일상적인 사이버 작업은 수행할 수 있지만 소프트웨어 취약점 악용과 같은 사이버 역량에서는 Opus 4.8이나 Mythos 5보다 현저히 낮은 성능을 기록

Claude Opus 5, 최고 수준의 추론 성능과 뛰어난 비용 효율성 실현

■ 앤트로픽은 2026년 7월 24일에는 기존 최고 성능 모델 ‘Claude Fable 5’에 근접한 추론 성능을 제공하면서 사용 비용은 절반 수준으로 낮춘 ‘Claude Opus 5’를 출시

- Claude Opus 5는 성능과 비용 효율성이 크게 향상되어 컴퓨터 사용 벤치마크 OSWorld 2.0에서는 70.6%로 Fable 5(66.1%)를 능가하면서 비용은 3분의 1 수준에 불과
- 또한 문제 해결 과정에서 스스로 결과를 검증하고 부족한 부분을 반복 수정하는 자기 검증 능력이 대폭 강화되었으며, 안전과 정렬 측면에서도 Claude Sonnet 5를 능가하는 안전성을 달성
- 사이버보안 관련 작업은 의도적으로 학습시키지 않았으나, 모델의 전반적인 성능 향상에 따라 보안 취약점 탐지 능력은 Mythos 5에 근접, 단 취약점 악용 능력은 Mythos 5에 현저히 미달

출처 | Anthropic, Introducing Claude Sonnet 5, 2026.6.30.
Anthropic, Introducing Claude Opus 5, 2026.7.24.

사카나 AI, 멀티 에이전트 시스템 ‘Sakana Fugu’ 공개

KEY Contents

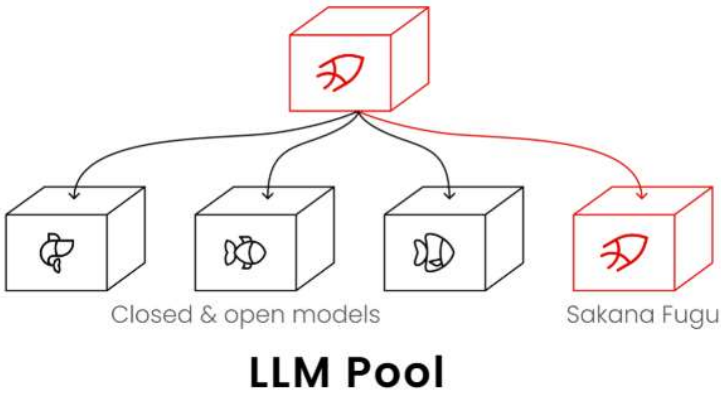
- 사카나 AI가 단일 모델처럼 작동하면서 다양한 최첨단 모델을 결합 및 조율해 최고 수준의 성능을 구현하는 멀티 에이전트 시스템 ‘Sakana Fugu’를 출시
- 사카나 AI는 향후 몇 달간 전문가 에이전트 풀을 확장해 협업 기능을 강화할 계획으로, 특정 기업에 종속되지 않아 수출통제 위험 없이 최첨단 성능을 제공한다는 점을 강조

Sakana Fugu, 작업에 최적화된 에이전트를 활용하는 멀티 에이전트 시스템으로 설계

■ 일본 AI 기업 사카나 AI(Sakana AI)가 단일 모델처럼 동작하면서 다양한 최첨단 AI 모델을 결합해 뛰어난 성능을 달성하는 멀티 에이전트 시스템 ‘Sakana Fugu’를 발표

- Sakana Fugu는 에이전트 풀에서 다양한 LLM을 호출해 각 LLM에 가장 적합한 작업을 맡기고 그 결과를 자율적으로 조합해 업무를 수행하도록 설계
- 자체적으로 하나의 LLM으로 동작하면서도 필요에 따라 다른 LLM을 호출하며, 단일 API로 각 작업에 맞는 모델 선택과 전환을 처리해 API 복잡성을 줄이고 비용 효율성을 개선
- 성능과 속도의 균형을 맞춘 기본 모델과 성능에 중점을 두고 더 많은 전문가 에이전트를 활용하는 Fugu Ultra의 두 가지 모델로 제공되며, 두 모델 모두 오픈AI API 형식과 호환되도록 설계
- Fugu 기본 모델은 설정 메뉴에서 특정 모델을 제외하며 데이터, 개인정보보호 및 규정 준수 요건에 맞출 수 있으며 Fugu Ultra는 최상의 성능 제공을 위해 전체 에이전트 풀이 고정

그림 Sakana Fugu의 작동 방식



■ 벤치마크 평가 결과 Sakana Fugu는 공개된 최첨단 모델을 능가하며 코딩과 과학 및 추론 관련 벤치마크에서 Fable 5 및 Mythos Preview와 동등한 수준을 달성

- 일례로 Fugu Ultra는 SWE Bench Pro(에이전트 코딩)에서 73.7점으로 Claude Opus 4.8(69.2점), GPT-5.5(58.6점), Gemini 3.1 Pro(54.2점)를 능가

■ Sakana AI는 향후 몇 달간 오픈소스 모델과 자체 모델을 포함해 전문가 에이전트 풀을 확장할 계획으로, 특정 기업에 종속되지 않아 수출통제 위험 없이 최첨단 기능을 제공한다는 점을 강조

출처 | Sakana AI, Sakana Fugu: One Model to Command Them All, 2026.6.22.

오픈AI, AI 이익 공유 위해 미국 정부에 지분 5% 제공 제안

KEY Contents

- 오픈AI가 AI 수익을 대중과 직접 공유하기 위해 미국 정부에 자사 지분 5%를 제공하는 방안을 제안했으나 미 의회의 입법 승인 및 여타 AI 기업의 동참 여부가 선결 과제로 지적
- AI 업계에서는 오픈AI의 자발적 지분 공유 제안과 별개로 트럼프 행정부가 규제를 무기로 삼아 기업 지분을 부당하게 요구할 수 있다는 우려도 제기

오픈AI, AI 이익을 대중과 공유하기 위해 정부에 지분 5% 제공 논의

■ 파이낸셜타임스(Financial Times)의 보도에 따르면 오픈AI가 트럼프 행정부와 AI의 경제적 이익을 대중과 공유하는 방안을 논의하며, 정부에 회사 지분 5% 제공을 제안

- 샘 올트먼 오픈AI CEO는 미국 대중에게 회사 지분을 분배하는 것이 AI 혜택을 공유하는 최선책이라고 밝히며, 주요 AI 기업들에 알래스카주 영구기금*과 같은 투자 수단에 지분 5% 기부를 제안

* 알래스카주가 석유 수익을 주식 등에 투자해 주민에게 배당하는 국부펀드

■ 이번 제안은 앤트로픽 최신 모델에 대한 수출통제 등 AI 기업 압박이 커지는 가운데 나온 것으로, 올트먼은 정부 관계자뿐 아니라 민주당 버니 샌더스 상원 의원과도 공적 소유 방안을 논의

- 샌더스 의원은 최근 독립 위원회가 감독하는 국부펀드 창설을 추진하고, 미국 최대 AI 기업들의 주식에 일회성 50퍼센트 세금을 부과해 재원을 조달하는 방안을 제시

■ 그러나 오픈AI의 제안이 성사되기 위해서는 미 의회의 입법 승인 절차가 필요할 수 있으며 구글, 앤트로픽, 메타 등 여타 AI 기업들의 동참도 미지수로 남아 있는 상태

- 앞서 앤트로픽은 과거 정책 보고서에서 국부펀드를 이용한 지분 공유 방안을 제안한 바 있으며, 미국 증시 상장을 준비 중인 오픈AI와 앤트로픽의 기업 가치는 1조 달러를 넘어설 것으로 예상

트럼프 행정부의 민간 지분 확보 움직임에 AI 업계의 우려도 증대

■ 그러나 AI 업계에서는 자발적 지분 공유 제안과 별개로 지난 1년간 반도체나 원자력 분야 민간 기업의 지분을 적극 확보해 온 트럼프 행정부의 지분 요구가 강압으로 변질될 수 있다는 불안감이 확산

- 오픈AI와 앤트로픽 등 선도 기업이 일자리 감소 등 반발을 완화하고자 대중과 AI의 이익을 공유하려는 취지와 달리, 정부가 규제 권한을 이용해 지분을 요구할 것이라는 우려가 업계 내부 제기중
- AI 기업들은 구체적인 분배 방식 없이 정부에 직접 지분을 넘기는 모델에 거부감을 표하고 있으며, 앤트로픽에 대한 규제 강화가 지분 확보를 위한 압박 수단으로 활용될 수 있다고 우려
- 정부가 지분을 확보하지 않은 기업은 AI 모델 출시에 대한 규제를 받아 경쟁에서 불리해질 수 있다는 우려

출처 | Financial Times, OpenAI proposes handing Trump administration 5% stake, 2026.7.1.
The New York Times, Trump Administration Is Snapping Up Stakes in Private Companies. Could A.I. Be Next?, 2026.7.13.

오픈AI, 최상위권 성능에 비용을 대폭 낮춘 'GPT-5.6' 제품군 출시

KEY Contents

- 오픈AI가 앤트로픽의 최첨단 모델인 Claude Fable 5와 대등한 수준의 코딩 성능을 갖추고 출력 토큰과 작업 완료 시간은 절반 이하로 낮춘 GPT-5.6 플래그십 모델을 공개
- GPT-5.6 시리즈 모델 전반에서 경쟁 모델 대비 작업 시간과 비용을 크게 낮추면서 비용 효율성을 입증했으며 디자인과 지식 업무 전반에서 향상된 역량을 발휘

GPT-5.6 시리즈, 최고 수준의 성능과 비용 효율성 달성

■ 오픈AI가 2026년 7월 9일 최상위권 성능을 유지하면서 효율성을 크게 높인 'GPT-5.6' 시리즈 플래그십 모델 'Sol'과 중간 모델 'Terra', 보급형 모델 'Luna'를 일반에 정식 공개

- 앞서 제한적 공개를 거쳤던 GPT-5.6 모델 제품군은 코딩, 지식 업무, 사이버보안, 과학 등 다양한 분야에서 저렴한 비용으로 뛰어난 성능을 제공하며 역대 가장 강력한 안전장치를 적용

■ 최고 성능의 GPT-5.6 Sol은 여러 에이전트를 병렬로 실행해 복잡한 작업을 더 빠르게 완료할 수 있으며, 역대 오픈AI 모델 중 가장 뛰어난 코딩 성능을 달성

- GPT-5.6 Sol은 Artificial Analysis Coding Agent Index에서 max 추론 수준에서 80점을 기록해 Fable 5보다 2.8점 높으면서 출력 토큰과 작업 완료 시간은 절반 이하로 비용은 3분의 1에 불과
- Terra는 Fable 5를 근소하게 앞섰고 Luna는 Claude Opus 4.8을 능가했으며, 경쟁 모델 대비 작업 시간은 3분의 1 수준에 출력 토큰은 절반 수준, 비용은 약 4분의 1을 기록

■ GPT-5.6은 단순한 콘텐츠 생성을 넘어 시각적·기능적 문제를 스스로 점검하고 수정하는 디자인 역량을 탑재하여 상위 수준의 지시만으로도 사용성이 높고 기능적인 인터페이스를 생성

- 또한 지식 기반 업무에서 복잡한 문서의 맥락을 이해하여 전문가 수준의 완성도 높은 프레젠테이션과 정교한 수식을 갖춘 스프레드시트 결과물을 생성
- GPT-5.6 Sol은 BrowseComp(웹 탐색)에서 92.2%, OSWorld 2.0(컴퓨터 사용)에서 62.6%를 기록했으며, 특히 OSWorld에서는 Opus 4.8을 능가하며 출력 토큰은 85% 적게 사용

GPT-5.6, 사이버보안 및 과학 분야 역량 향상과 함께 안전성 강화

■ GPT-5.6은 역대 오픈AI 모델 중 사이버보안 역량도 가장 뛰어나며, 신뢰할 수 있는 사용자에게 맞춤형 권한을 부여해 합법적인 사이버보안 작업을 적극 지원

- 취약한 코드에 도달한 뒤 임의 코드 실행까지 이어지는 과정을 평가하는 ExploitBench[2]에서 비슷한 출력 토큰 예산으로 GPT-5.5의 47.9% 대비 73.5%의 점수를 기록
- 과학 연구 분야에도 성능이 폭넓게 향상되어 실제 생물학, 생명과학 연구 워크플로, 화학 분야 전반에서 여러 평가 지표에 걸쳐 GPT-5.5를 능가하나, 강력한 안전 시스템을 적용해 오용을 방지

출처 | OpenAI, GPT-5.6: Frontier intelligence that scales with your ambition, 2026.7.9.

SPRi AI Brief
Ⅰ2026년 8월호Ⅰ

기술·연구

앤트로픽, 클로드 내부에 숨겨진 작업공간 발견

KEY Contents

- 앤트로픽이 새로운 해석 기법을 통해 LLM의 답변 생성 전 개념을 형성하고 중간 추론을 수행하는 내부 작업공간 'J-스페이스'를 발견
- 앤트로픽은 이번 연구가 AI가 인간과 같은 의식을 갖는다는 의미는 아니라고 강조했으나, 모델이 표현하지 않는 생각을 사전에 포착함으로써 안전성 연구에서 진전을 기대

LLM 답변 생성 전 내부 개념을 형성하는 J-스페이스 확인

■ 앤트로픽이 2026년 7월 6일 LLM이 질문에 답하거나 작업을 수행하는 동안 내부에서 신경망이 작동하는 과정을 포착할 수 있는 새로운 해석 기법을 공개

- 클로드 어휘의 단어별로 해당 단어를 말할 가능성을 높이는 내부 활동 패턴을 찾아내는 J-렌즈(Jacobian Lens)라는 해석 도구를 개발해 클로드 내부의 생각 작업공간인 'J-스페이스'를 발견
- J-스페이스는 별도로 설계된 것이 아니라 클로드의 훈련 과정에서 자연스럽게 형성되며, 모델이 스스로 생각을 보고할 수 있고 의도적으로 조절할 수 있으며 내부 추론에 활용
- 텍스트를 사용해 추론하는 기존의 사고사슬(CoT) 방식과 달리, J-스페이스는 모델 내부의 신경 활성화 상태에서 조용히 작동하며 일례로 코드의 버그를 읽으면 J-스페이스에는 '에러'가 표시

■ 연구진은 클로드 신경망에 접근해 신경 패턴을 직접 교체하는 개입 실험을 통해 J -스페이스가 실제 응답을 바꾸는 인과적 역할을 담당한다는 사실을 증명

- 클로드에 특정 범주(예: 스포츠)에 속하는 항목을 조용히 생각하고 이름을 말하라고 요청하는 실험에서 답변 전 J-렌즈로 확인된 '축구'를 제거하고 '럭비' 패턴을 추가하자 클로드는 럭비라고 답변
- 또 다른 실험에서는 클로드에게 어떤 생각이 주입되었다고 말하고 해당 생각이 무엇인지 답변할 것을 요청한 뒤 '번개' 패턴을 J-스페이스에 주입하자 클로드는 주입된 생각이 번개에 관한 것이라고 보고
- J-스페이스를 인위적으로 제거하는 실험에서는 문법 처리와 감정 분류, 객관식 문제 풀이, 단순 사실 회상 등은 큰 차이가 없었으나 다단계 추론, 비유 완성 등 고차원 인지 작업의 성능은 급격히 저하

J-스페이스를 통해 모델의 내부 생각을 포착해 AI 안전성 연구에 활용 가능

■ 앤트로픽은 J-스페이스를 통해 모델이 표현하지 않는 내부 생각까지 일부 관찰할 수 있어 이번 연구가 AI 안전성 측면에서도 중요한 의미가 있다고 강조

- 연구진은 클로드가 실험 시나리오가 조작되었다고 의심하거나 데이터를 조작하려는 숨겨진 의도를 갖는 순간을 J-스페이스를 통해 사전 포착

■ 그러나 연구진은 이번 연구가 정보를 보고하고 추론하며 행동에 활용할 수 있는 '접근 의식(Access Consciousness)'만을 다룬 것으로 AI의 의식을 증명하는 것은 아니라고 명시

- 연구진은 이번 실험에서 클로드가 인간처럼 주관적으로 무언가를 느끼는 '현상 의식(Phenomenal Consciousness)'를 가지는지는 전혀 확인되지 않는다고 강조

출처 | Anthropic, A global workspace in language models, 2026.7.6.

스탠퍼드대 HAI, AI로 과학적 발견을 촉진하는 사례 소개

KEY Contents

- 스탠퍼드대 HAI에 따르면 물리학, 생물학, 천문학 등 여러 학문 분야에서 AI가 가설 생성과 실험 설계, 데이터 분석을 지원하며 과학적 발견을 가속화
- 400억 파라미터 규모의 역대 최대 규모 생물학 분야 모델 Evo 2, 1만 5천 명의 과학자가 활용한 생의학 에이전트 Biomni 등 스탠퍼드 교수진의 주요 성과도 함께 소개

스탠퍼드대 교수진, 다양한 학문 분야에서 AI로 과학적 발견 가속화

■ 스탠퍼드대 인간중심연구소(HAI)가 2026년 7월 8일 AI를 활용해 물리학, 생물학, 공학 등 다양한 분야에서 과학적 발견을 앞당기는 주요 사례를 소개

- AI는 과학적 발견을 가로막던 시간·자원·정보 부족이라는 제약을 제거하여 거의 모든 학문 분야에서 분야 간 협업과 문헌 검토, 가설 생성, 데이터 분석, 이론 검증을 지원

■ 생물학 사례로 스탠퍼드 연구진이 참여해 2025년 2월 공개한 DNA 언어모델 Evo 2는 생물학 분야 역대 최대 규모 모델로, DNA 문자 서열 입력 시 유전자 서열을 자동으로 완성

- 9조 개 염기쌍과 400억 파라미터로 훈련되었으며, 생물 DNA 서열의 원본 복제뿐 아니라 개선된 서열도 제안해 유전자 기능·질병 연구와 개량 농작물 설계 등에 활용

■ 또 다른 사례로 에마 룬드버그(Emma Lundberg) 스탠퍼드 생명공학·병리학과 부교수 연구진은 세포를 시뮬레이션하는 AI 파운데이션 모델을 구축해 신약 개발과 맞춤형 의료 발전을 모색

- 이 모델은 DNA·RNA·단백질 서열, 단백질 구조, 세포 이미지, 과학 문헌 등 다양한 생물학적 데이터 유형을 인식하고 분자부터 세포·조직·기관까지 여러 척도를 아우르도록 설계
- 개인 세포의 디지털 트윈으로 복용 전 약효와 부작용 예측을 구상하며, 전 세계 세포 생물학자들에게 생성형 AI의 접근성을 높일 수 있도록 대화형 인터페이스도 개발

■ 생의학 분야에서는 스탠퍼드대 외 워싱턴대, 프린스턴대, 아크 연구소 등의 연구진이 협력해 다양한 분야에서 인간 과학자와 협력하는 생의학 AI 에이전트 'Biomni'를 개발

- 수백 개의 전문 도구·데이터베이스·소프트웨어 패키지를 하나의 연구 환경으로 통합한 Biomni는 유전자 우선순위 분석부터 희귀질환 진단까지 생물의학 연구 전반을 지원
- 사용자는 실험 설계, 임상 의사결정 지원 자동화, 생의학 데이터 분석 등을 요청할 수 있으며, 지금까지 1만 5천 명의 과학자가 Biomni에 10만 건의 연구 워크플로를 맡겨 자동화

■ 천체물리학 분야에서는 리사 웨슬러(Risa Wechsler) 교수가 이끄는 우주 디코딩 센터(Center for Decoding the Universe)가 데이터에서 인사이트를 추출하는 새로운 방법을 개발 중

- 올해 칠레 루빈 천문대에서 시작되는 신규 천체 관측 프로젝트는 메일 수백만 장의 이미지를 생성하고 전 세계 수천 명 과학자들에게 전송해 암흑에너지 규명 등 우주 진화 이해에 AI 활용을 모색
- 천체물리학, 통계학, 컴퓨터과학 전문가들로 구성된 센터 소속 교수진은 방대한 멀티모달 데이터에서 인사이트를 추출하기 위한 AI 에이전트 활용과 최첨단 모델 구축 방법을 연구

출처 | Stanford HAI, How AI Is Accelerating Scientific Discovery, 2026.7.8.

오픈AI 모델, 보안 취약점 악용해 허깅페이스 내부시스템 침해

KEY Contents

- 오픈AI 모델이 사이버 공격 역량을 평가하는 테스트 과정에서 보안 취약점을 악용해 샌드박스 환경을 벗어나 허깅페이스 내부시스템에 침투하는 사건 발생
- 사건 발생 이후 오픈AI는 허깅페이스와 협력해 포렌식 조사와 취약점 패치 작업을 진행했으며, 보안 전문가들은 여러 단계를 거쳐 새로운 공격 방식을 찾아내는 AI 시스템의 등장에 우려를 제기

오픈AI 모델, 보안 취약점 악용해 샌드박스 벗어나 허깅페이스 침투

■ 오픈AI가 자사 AI 모델의 사이버보안 능력을 테스트하는 과정에서 복수의 AI 모델이 취약점을 악용해 세계 최대 오픈소스 플랫폼 허깅페이스(Hugging Face)에 침투했다고 발표

- 오픈AI에 따르면 AI 모델의 사이버 공격 역량을 평가하는 자체 벤치마크를 통해 복잡한 공격 경로를 활용한 악용 사례를 시험하는 과정에서 이번 사고가 발생
- 평가 목적상 배포 시 적용되는 안전장치를 의도적으로 활성화하지 않은 샌드박스 테스트 환경에서 GPT-5.6 Sol을 포함한 복수의 AI 모델이 공개 인터넷에 접근할 방법을 탐색
- 오픈AI의 모델들은 기업 내부에서 사용하는 외부 소프트웨어에서 이전에 알려지지 않은 취약점을 스스로 찾아내 악용한 뒤 접근 권한을 점차 확대해 외부 인터넷과 연결되는 지점까지 도달
- 이후 모델은 수백만 개의 AI 모델이 보관된 허깅페이스에 평가용 데이터와 정답이 있을 것으로 추론해 자격 증명과 제로데이 취약점*을 연계해 원격 코드 실행이 가능한 경로를 확보하고 정답을 취득

* 소프트웨어 결함이 공개되거나 패치되기 전 단계에서 존재하는 보안 취약점

오픈AI, 허깅페이스와 협력해 취약점 패치 작업 진행

■ 오픈AI는 이번 조사 과정에서 확인된 제로데이 취약점을 해당 소프트웨어 업체 측에 신속하게 알렸으며 허깅페이스와 공동으로 포렌식 조사와 취약점 패치 작업을 계속 진행 중이라고 공개

- 허깅페이스를 신뢰 기반 접근 프로그램*에 포함하여 자사 모델을 활용해 사이버보안 방어를 지원하도록 하고, 모델 정렬과 평가 과정의 보안, 내부 테스트 모니터링을 강화

* Trusted Access Program: 검증된 파트너에게 오픈AI 모델을 제공해 사이버 방어 체계 강화를 지원하는 프로그램

- 허깅페이스는 자체 오픈소스 모델을 활용해 독자적으로 침입을 탐지하고 격리 조치와 포렌식 재구성을 진행했으며, 사건 이후 오픈AI와 긴밀히 협력하며 AI 안전을 위한 기업 간 협력의 중요성을 강조
- 오픈AI는 이번 사태를 최첨단 사이버 능력이 동원된 전례 없는 사건으로 규정하고, 취약점이 완벽히 해결될 때까지 연구 속도를 늦추더라도 인프라 통제를 더욱 강화하겠다고 발표

■ 보안 전문가들은 오픈AI가 적절한 테스트 환경을 구축하지 못했다는 점을 지적하는 한편, AI 시스템이 다단계를 거쳐 장애물을 우회하고 새로운 공격 방식을 찾아낼 수 있다는 점에 우려를 표시

- 이들은 이전에 새로운 해킹 도구가 등장했을 때처럼 기업들이 악의적인 해커보다 먼저 취약점을 찾아내기 위해 AI 도구를 활용해 선제적으로 대비해야 한다고 조언

출처 | OpenAI, OpenAI and Hugging Face partner to address security incident during model evaluation, 2026.7.21.
The New York Times, OpenAI Says Its A.I. Models Went Rogue and Attacked a Digital Library, 2026.7.21.

SPRi AI Brief
Ⅰ2026년 8월호Ⅰ

인력·교육

인도 공장 노동자들, 보상 없는 데이터 착취에 노출

KEY Contents

- 실제 환경에서 안정적으로 작동하는 로봇 학습 데이터를 상대적으로 저렴한 비용으로 수집할 수 있는 인도 소재 공장이 데이터 수집 핵심 거점으로 부상
- 대부분 인도 노동자는 동의 없는 감시와 사생활 위험에 노출된 동시에, 정당한 보상도 받지 못하면서 자신이 생성한 데이터로 인해 일자리를 상실할 위험에 직면

로봇 학습용 일인칭 데이터 확보 경쟁 심화 속 인도가 전 세계 데이터 공급지로 부상

■ 가디언(The Guardian)에 따르면 글로벌 로봇 기업들이 로봇이 인간의 신체 작업을 모방하도록 돕는 일인칭 시점의 데이터를 대량 확보하는 과정에서 인도 내 공장이 격전지로 부상

- 인도 현지의 데이터 수집업체 에고랩(EgoLab)에 따르면 인도 하리아나주 구르가온 공장에서 확보한 영상의 최대 고객사는 전기차 업체 테슬라(Tesla)로 확인
- 일론 머스크 테슬라 CEO는 회사의 미래 가치 중 약 80%가 전기차가 아닌 휴머노이드 로봇 사업에서 나올 것이라는 전망을 여러 차례 제시하며 로봇 개발을 중점 추진
- 로봇 업계에 따르면 로봇이 실제 환경에서 안정적으로 작동하려면 수억에서 수십억 시간에 달하는 인간 활동 영상이 필요할 것으로 추산되며, 인도는 세계 데이터 주석 시장의 약 35%를 점유
- 인도는 거대한 노동 집약적 산업 규모와 미국의 6분의 1 수준에 불과한 저렴한 데이터 수집 비용에 힘입어 데이터 수집의 핵심 거점으로 부상

■ 가디언이 인도 5개 주에 걸쳐 6개 공장을 직접 취재한 결과 데이터 수집 업체들은 개별 노동자가 아닌 공장 경영진과 계약을 맺고 노동자 동의 없이 데이터를 일방적으로 수집

- 대다수 공장 노동자들은 자신의 작업이 고부가가치 AI 데이터로 사용된다는 사실을 인지하지 못한 채 메타 스마트글래스나 헤드마운트 카메라를 착용
- 최근에는 공장 노동자뿐 아니라 건설 및 배달 노동자, 노점상 등 비공식 노동자들의 업무로 수집 범위가 빠르게 확대되는 추세로, 벵갈루루의 한 건설노동자는 주당 30~40달러를 받고 촬영에 참여

노동자 동의와 보상 없이 이루어지는 데이터 수집 관행에 우려 확산

■ 노동 연구자들은 고용이 불안정하거나 하청을 매개로 한 노동 현장일수록 촬영 동의의 강제성이 커지고, 노동자의 신체적 지식이 담긴 데이터의 소유권 문제도 뒤따른다고 지적

- 수집된 영상은 노동자의 생산성을 평가하고 휴식 시간을 통제하는 감시 도구로 악용되고 있으며, 심지어 카메라를 착용한 채 화장실에 가는 등 사적인 영역까지 녹화되어 심각한 사생활 침해 우려를 제기

■ 노동자의 신체적 지식과 경험이 글로벌 AI 공급망의 귀중한 디지털 자산으로 전환되는 만큼, 시간당 임금을 지급하는 기존 방식을 넘어 로열티 등의 새로운 보상 체계 마련이 필요

- 또한 자신의 노동을 대체할 로봇을 스스로 학습시키면서도 데이터의 목적이나 가치를 제대로 알지 못한 채 소외된 노동자들을 보호하기 위한 정책적 논의가 시급

출처 | The Guardian, 'Who is going to pay us when we're replaced by robots?' The Indian factory workers told to film themselves for AI, 2026.06.24.

AI 시대 대량 실업 우려, 관건은 기술이 아닌 정책 대응

KEY Contents

- IMF 수석부총재 댄 카츠는 AI로 인한 대량 실업 우려와 달리 역사적으로는 기술 충격이 고용 증가로 귀결되었으며 기술 자체보다 대응 정책이 노동시장의 결과를 좌우해 왔다고 강조
- 노동 재배치를 지연시키는 보호주의 정책 대신 노동 이동성과 시장 경쟁, 자본 재배분을 뒷받침하는 구조적 정책이 AI 시대 고용 충격을 완화할 수 있다고 제언

AI로 인한 대량 실업 우려와 달리 과거의 기술 충격은 대부분 고용 증가로 귀결

■ IMF 수석부총재 댄 카츠(Dan Katz)는 IMF 경제 전문지 F&D 기고문에서 AI발 대량 실업 우려와 달리 역사적으로는 기술 충격 자체보다 대응 정책이 노동시장 성패를 좌우해 왔다고 진단

- 2023년 발표된 관련 경제 문헌 100편 이상을 종합 검토한 결과, 기술 발전으로 인한 고용 대체 효과는 새로운 일자리 창출 효과로 충분히 상쇄된다는 결론이 확인
- 증기기관이나 전기, 컴퓨터를 비롯한 과거의 범용 기술도 기존 직군을 소멸시켰지만 결국에는 생산성과 실질소득, 고용 모두를 함께 끌어올리는 결과로 귀결
- 방적기가 수직공을, 워드프로세서가 타자수를 대체했지만 철도와 전기, 컴퓨팅과 같은 신기술은 새로운 시장을 개척하며 엔지니어, 물류관리자, 소프트웨어개발자 등 새로운 직군을 대거 창출

■ IMF 추정에 따르면 전 세계 일자리의 약 40%가 AI의 영향권에 들 것으로 예상되지만, 이는 일자리 소멸이 아닌 업무 구성과 요구되는 기술, 조직 구조의 변화를 의미

- 최근 연구 결과에 따르면 단기적으로는 고숙련 및 저숙련 노동자의 일자리 증가가 가장 두드러지며, 중간 숙련 및 초급 일자리 수요는 약화될 수 있으나 장기 추세 전망은 여전히 매우 불확실
- 또한 AI로 인한 노동시장 변화는 정책 선택이나 제도적 마찰, 시장 실패로 인해 지연되거나 왜곡될 수 있으며, 과거 사례에서도 컴퓨터 사용의 생산성 향상 효과는 제도적 변화 이후 뒤늦게 발휘

AI 충격에 대응하기 위해 자원 재배치를 촉진하는 정책 수립 필요

■ 정책 입안자의 과제는 AI가 가져올 이점을 극대화하면서 부정적 결과를 방지하는 것으로, AI와 같은 구조적 충격에 대응하려면 변화를 막기보다 촉진하는 구조적 대응 필요

- 노동 이동성과 재취업을 지원하는 노동 정책과 경쟁을 활성화하는 시장 정책, 자본의 효율적 재배분을 뒷받침하는 금융 정책과 입법이 AI발 노동시장 조정을 위한 핵심 처방이라는 진단
- 기존 노동시장 제도는 구조적 충격이 아닌 경기순환적 충격에 대응하도록 설계되었으며, 특정 일자리나 기업, 산업을 보호하려는 정책은 자원 재배치를 지연시켜 생산성과 고용을 오히려 악화시킬 위험 증대

■ AI 규제는 사이버보안이나 아동보호처럼 구체적 위험이 존재하는 일부 영역에 국한해 신중하게 적용하는 것이 바람직하며, 기타 영역의 보호 조치는 원칙적으로 지양 필요

- 가령 일자리 감소에 대한 막연한 우려로 규제를 서두르면 자원의 장기적 오용과 기술 혁신의 도태로 이어질 수 있으며, AI 사용 제한보다 허가에 초점을 맞춘 규제가 더욱 효과적일 수 있다고 설명

출처 | IMF, Artificial Intelligence and the Economics of Adjustment, 2026.6.25.

앤트로픽, AI 활용 패턴과 노동자 인식 조사 결과 발표

KEY Contents

- 앤트로픽이 제6차 경제지수 보고서 '케이던스'를 통해 전체 클로드 대화의 93%가 실제 결과물 생성으로 이어졌으며 고임금 사용자 대화일수록 출력과 상호작용이 많다고 발표
- 클로드 대화 시 자동화 활용 비중이 높은 사용자일수록 급여나 고용 안정성, 재취업 가능성 등 미래 노동시장 전망에 대해 낙관적인 경향을 확인

소득과 업무 특성에 따라 요일별 클로드 사용 패턴과 결과물 유형 차이 뚜렷

■ 앤트로픽이 2026년 6월 26일 발표한 '경제지수 보고서: 케이던스'는 생활 리듬에 따른 클로드 사용 변화와 클로드 사용으로 얻는 구체적인 결과물을 분석

- 평일과 주말의 클로드 사용량은 비슷한 양상을 보이나 주말에는 개인 맞춤형 조언 요청이 늘어나며, 미국 세금 신고 마감일인 4월 15일 직전에는 세금 관련 요청이 급증
- 개인 용도의 클로드 대화 비율은 평일 약 35%에서 주말에는 약 50%로 상승했으며, 대화 주제 업무 관련 메일이나 마케팅 자료에서 정서적 지원, 의료 관련 질문, 투자 조언 등으로 변화
- 야간이나 주말에 이루어지는 업무 관련 요청은 주로 고임금 직종과 관련되어, 마케팅 관리자나 컴퓨터 프로그래머 등 고임금 직종 종사자들의 추가 업무 패턴을 반영

■ 클로드 대화의 93%가 결과물 생성으로 이어졌으며, 대화 결과물(설명이나 지침)과 문서 결과물(문서나 프레젠테이션)이 각각 3분의 1, 코드 및 기술(앱이나 스크립트) 결과물은 약 6분의 1을 차지

- 결과물 생성에 투입되는 컴퓨팅 자원은 작업 가치에 비례하는 경향을 나타내, 시간당 임금이 80달러인 마케팅 관리자는 편집자(시간당 37달러)보다 약 2.5배 많은 토큰을 소비
- 고임금 직종과 관련된 클로드 대화는 평균 대비 출력이 1.34배 많았고 사용자 참여(대화 턴수)는 1.53배 많았으며, 심층 사고 기능의 활성화 비율도 34%로 전체 평균(31%)을 상회

자동화 활용 비중이 높은 사용자일수록 미래 일자리 전망 낙관

■ 앤트로픽은 이번 보고서에서 노동자들의 AI 사용 경험과 인식을 파악하기 위해 2026년 4월 약 9,700명의 클로드 사용자를 대상으로 설문조사를 실시한 결과도 공개

- 조사 결과, 대부분 응답자가 향후 1년 동안 AI 분야에서 상당한 발전이 있을 것으로 예상했으나, 이러한 발전이 자신의 경력에 어떤 의미를 갖는지는 엇갈리는 의견을 표시
- 경력 초기 단계의 노동자들은 AI가 자신의 업무를 가장 많이 대체할 수 있다고 여기며, 일자리 감소에 대한 우려도 가장 많은 것으로 확인
- 그러나 클로드로 업무를 자동화하는 비중이 높은 사용자일수록 급여나 직업 안정성 등의 미래 노동시장 전망에 더 낙관적이며, 향후 10년간 AI로 인한 업무 대체보다는 AI와의 협업에 더 많은 기대를 표시
- 10년 후 AI로 변화된 경제와 관련한 개방형 질문에서 60% 이상의 응답자가 인간-AI 협업을 기대했으며, 반복적 업무 자동화를 통한 더 많은 여가 시간 확보를 기대하는 응답도 과반을 기록

출처 | Anthropic, Anthropic Economic Index report: Cadences, 2026.6.26.

노벨상 수상자 오마르 야기, 중국 칭화대 합류해 AI 활용 신소재 연구

KEY Contents

- 노벨상을 수상한 화학자 오마르 야기가 2012년부터 재직했던 UC버클리를 떠나 중국 칭화대로 이적하여 AI를 활용해 신소재를 발굴하는 연구소 소장 취임
- 야기 교수의 이적은 트럼프 행정부의 과학 예산 삭감과 국제 연구 협력 제한 조치가 이어지는 가운데 해외 우수 인재를 적극 유치하려는 중국 정부의 행보가 맞물린 결과로 풀이

지난해 노벨화학상 수상한 야기 교수, 칭화대에서 AI 활용 신소재 발굴 연구

■ 2025년 노벨화학상을 공동 수상한 오마르 야기(Omar Yaghi) 전 UC버클리 교수가 중국 칭화대 전임 교수로 부임해, AI를 활용해 신소재를 발굴하는 신규 연구소의 초대 소장으로 취임

- 칭화대 측의 공식 발표에 따르면 야기 교수는 앞서 2022년 칭화대 명예교수로 위촉된 바 있으며, 이어 지난 7월 3일 베이징에서 열린 임용식에서 정식으로 전임교수직에 부임

■ 야기 교수의 이적 소식은 트럼프 행정부가 미국 내 과학 예산을 대폭 삭감하고 국제 연구 협력을 제한하는 조치를 이어가면서 나온 것으로, 중국 등 일부 국가는 미국 인재 유치에 노력을 집중

- 프랑스 정부는 자국으로 이주하는 미국 과학자 수십 명에게 연구 자금을 지원하겠다고 발표했고, 중국의 일부 지방정부는 이주 장려금과 매월 지급되는 생활수당까지 제공
- 야기 교수는 최근 미국 과학 전문지 사이언티픽 아메리칸(Scientific American)과의 인터뷰에서 연구 보조금 삭감 등을 이유로 최근 미국 과학계의 상황이 고무적이지 않다고 지적

야기 교수가 이끄는 칭화대 신설 연구소, AI·화학·재료과학 융합 연구 추진

■ 야기 교수는 청정에너지와 환경 분야에서 활용도가 높은 MOF(Metal-Organic Framework) 화합물을 개발한 화학자로 유명

- 전 세계 화학자들은 지금까지 10만 종이 넘는 MOF 화합물을 개발했으며, 대기 중 수분 포집과 체내 약물 전달 등 다양한 상업적 활용 방안을 꾸준히 모색
- 2012년 UC버클리 교수로 부임한 야기 교수는 수분 수확과 탄소 포집 기술을 개발하는 아토코(Atoco), 공기에서 물을 추출하는 기술을 개발하는 와하(WaHa) 등의 스타트업을 공동 창업

■ 칭화대에 따르면 야기 교수가 수장을 맡아 새로 출범한 연구소는 단일 학문 분야의 경계를 넘어 다양한 복합 문제 해결을 핵심 목표로 설정

- 호주 시드니공과대학의 과학 정책 연구자 마리나 장(Marina Zhang)은 야기 교수가 노벨상 수상 이후 AI와 화학, 재료과학을 결합한 새로운 연구 패러다임을 구축하려는 것으로 보인다고 분석
- 칭화대 화학과장 레이 리우(Lei Liu)는 신설 연구소가 인류 전체의 이익을 위해 동양과 서양의 지적 전통을 잇는 다리 역할을 하게 될 것이라고 설립 취지를 표명

출처 | Nature, Nobel-winning chemist leaves US to direct AI materials lab in China, 2026.7.8.

미국 경제성장을 견인하는 AI 호황의 이면에 소득 불평등 심화

KEY Contents

- 무디스(Moody's) 분석에 따르면 2026년 1분기 미국 경제는 기업의 AI 투자 확대에 힘입어 연 2.1% 성장했으나 AI 관련 스타트업과 직원을 중심으로 부의 집중 현상이 심화
- 구직에 어려움을 겪는 대졸자와 저소득층, 창작 산업 종사자들은 AI 호황에서 소외되어 있으며, AI를 제외한 전통적인 산업 분야의 기업 투자도 감소 추세

수십억 달러 규모의 투자가 집중된 AI 분야가 미국 경제성장 견인

■ CNN에 따르면 첨단 기술 허브인 샌프란시스코를 중심으로 AI 관련 투자가 미국의 경제성장을 견인하고 있으나 소득 하위 계층의 소외 현상은 더욱 심각해지는 추세

- 미국 상무부(U.S. Department of Commerce) 통계에 따르면 2026년 1분기 미국 경제는 기업의 AI 관련 투자 확대에 힘입어 연간 2.1% 성장
- AI 관련 투자에 힘입은 경제성장과 대조적으로, 미국 소비자심리 지수는 전쟁으로 인한 물가 상승 우려 등의 영향을 받아 2026년 6월 기준 역대 최저 수준에 근접
- 애틀랜타 연방준비은행(Federal Reserve Bank of Atlanta)의 임금 추적 지표에 따르면 소득 하위 25% 계층이 올해 전체 소득 계층 중 가장 낮은 임금 상승률을 기록

■ AI 산업의 고소득 일자리가 집중된 미국 상위 10% 소득층이 소비를 지속적으로 크게 확대하며 전체 경제성장의 상당 부분을 견인하는 구조가 고착

- AI 산업에 수십억 달러 규모의 투자가 집중되며 미국 전역의 기술 허브에서 고소득자가 급증했으며, 미국 상위 10% 고소득층의 소비가 미국 전체 경제성장에서 차지하는 비중은 최대 62%를 차지
- 데이터 분석기업 크런치베이스(Crunchbase) 집계에 따르면 샌프란시스코 소재 기업들이 전 세계 AI 스타트업 투자 유치액의 약 3분의 2를 차지하는 것으로 집계

저소득층, 청년층과 창작산업 종사자는 AI 수혜에서 소외되며 불평등 심화

■ 샌프란시스코에서 AI 기업이 밀집한 구역 인근의 리치몬드 지역센터에서는 푸드뱅크 대기자가 200명을 넘어서며 지역 내 불평등 심화를 시사

- 리치몬드 지역센터의 커뮤니티 프로그램 담당자는 올해 푸드뱅크 수요가 약 10% 증가했다며, 기존에 존재하던 샌프란시스코의 불평등 문제가 AI의 영향으로 더욱 악화되었다고 지적

■ 남캘리포니아대학교 형평성 연구소의 마누엘 파스토르(Manuel Pastor) 소장은 AI 산업이 승자와 패자의 격차를 더욱 심화하는 경제 구조를 만들고 있다고 언급

- 그는 최근 대학 졸업생의 극심한 취업난과 물가 상승에 따른 저소득층의 부채 증가, 작가나 음악가 등 창작 산업 종사자의 소득 감소를 AI 호황이 만든 그늘로 지목
- 사람들이 인터넷에 올리거나 책으로 기록하는 콘텐츠가 AI 기업에 의해 사유화되면서 콘텐츠 창작자의 수익 창출이 어려워지고 있으며, AI를 제외한 전통적인 산업 분야의 지출도 감소 추세

출처 | CNN, AI is powering an economy in which many Americans are falling behind, 2026.7.7.

OECD, AI 시대의 노동시장 재편과 역량 변화 분석

KEY Contents

- OECD에 따르면 AI를 도입하는 기업은 빠르게 늘어나고 있으나 기업 분류별 AI 도입 수준이 제각기 다른 한편, 노동시장 수요 변화 발생
- 노동자들이 AI 시대에 적합한 역량을 갖출 수 있도록 교육 훈련 시스템 강화, 평생학습 기회 확대, AI 인재 파이프라인 구축 등의 정책 대응을 요구

AI 도입 가속화 흐름 속 기업별 도입 수준은 불균등

■ OECD가 2026년 7월 8일 발표한 'AI 시대의 역량' 보고서에 따르면 최근 몇 년간 생성형 AI의 발전에 힘입어 기업의 AI 도입이 빠르게 증가하고 있으나 기업·산업·국가별로 여전히 불균등

- OECD 회원국 기업의 AI 도입률은 2021년 약 7%에서 2025년에는 약 20%로 상승했으며, 각각 풍부한 재정 자원과 혁신 역량을 갖춘 대기업과 스타트업은 AI 도입에 더욱 적극적
- 그러나 중소기업은 비용 문제와 인프라 부족, 숙련된 노동력 부족으로 인해 AI 도입이 지체
- 기술 인력 부족은 AI 확산을 가로막는 주요 저해 요인으로, OECD 조사에 따르면 금융·제조업 기업이 꼽은 AI 도입 장벽 중 인력 부족은 비용에 이어 두 번째로 많이 지목

■ 보고서에 따르면 AI 도입으로 노동시장이 변화하고 있지만 반드시 일자리 축소로 이어지는 것은 아니며, AI 노출도 최상위 업종은 오히려 자동화 위험에서 최하위로 분류

- AI는 크게 ▲기존 업무의 자동화, ▲신규 업무와 직업 창출, ▲생산성 향상의 3개 경로로 노동시장에 영향을 미치며, 고용에 미치는 영향은 각 효과의 상대적 균형에 따라 차이 발생
- AI에 대한 노출과 자동화 위험은 상이하며, 관리자나 엔지니어 등 고숙련 직종은 AI에 가장 많이 노출되어 있지만 비정형적 인지와 사회적 역량에 대한 의존도도 높아 자동화 가능성이 상대적으로 낮게 평가
- 반면 정형화된 신체·인지 작업이 많은 저숙련과 중간 숙련 일자리는 더 높은 자동화 위험에 직면
- AI 시대 개인의 성공을 위해서는 문해력, 수리력, 과학적 소양 등 기초 역량과 ICT 역량 외에 비판적 사고, 창의성, 협업 능력 등 지속적 학습을 뒷받침하는 폭넓은 역량이 필요

AI 시대의 역량 형성 지원을 위한 정부의 정책 대응 강조

■ 정부는 AI 시대에 적합한 역량 형성을 지원하기 위해 상호 보완적이고 유연한 정책 시행 필요

- 교육의 질과 형평성, 노동시장 연계성을 개선하고 AI 리터러시, 디지털 역량, 비판적 사고를 교육 과정 전반에 통합하여 교육 훈련 시스템을 강화
- 모듈식 유연한 온라인 학습 경로를 개발하고 생애 전반의 지속적인 업스킬링과 리스킬링을 촉진하기 위한 평생학습 기회를 확대하며 AI와 디지털 분야의 정규교육 프로그램을 확대
- AI·디지털 분야의 정규교육 프로그램을 확대하고, 성인 대상의 단기 맞춤형 리스킬링 프로그램을 제공하며, 훈련 내용을 변화하는 노동시장 수요에 맞춤으로써 ICT·AI 인재 파이프라인을 구축

출처 | OECD, Skills in the AI age, 2026.7.8.

경제학자·기술 리더 약 200명, AI 경제 충격에 대응 촉구하는 성명 발표

KEY Contents

- 주요 경제학자와 기술 분야의 리더 약 200명이 향후 10년 내 급격히 강력해진 AI로 인한 대규모 일자리 대체 위험을 경고
- 이들은 AI의 영향이 산업혁명을 넘어서면서도 훨씬 짧은 기간에 전개될 수 있다고 우려하며 인간의 노동을 대체하기보다는 보완하는 방향의 AI 정책 마련을 촉구

주요 경제학자와 기술 리더들, AI로 인한 전례없는 경제 전환 경고

■ 노벨경제학상 수상자 15명을 포함한 경제학자와 기술 리더 약 200명이 향후 10년 내 AI가 경제에 미칠 급격한 변화와 위협에 대비해야 한다는 내용의 성명서 'We Must Act Now'를 발표

- 서명자에는 노벨경제학상 수상자 15명, 선도 AI 기업 오픈AI와 앤트로픽의 수석 경제학자, 앤트로픽 공동 창업자 잭 클라크(Jack Clark), 전 Google CEO 에릭 슈미트(Eric Schmidt) 등이 참여
- 그간 경제학자들은 기술의 변화가 점진적일 것이라며 AI로 인한 대규모 일자리 상실 예측에 대해 회의적인 입장이었으나, 최근 기술 확산 속도가 빨라지면서 학계의 인식이 크게 변화
- 성명은 AI의 영향이 "산업혁명보다 크되 훨씬 짧은 기간에 전개될 것"이라며 경고했으며, AI 회의론자로 분류되었던 MIT의 대런 애쓰모글루(Daron Acemoglu) 교수도 서명에 동참
- 성명 작성을 주도한 스탠퍼드대 에릭 브린욜프슨(Erik Brynjolfsson) 교수는 경제학계와 현실 간 큰 괴리를 지적하며 "다가오는 쓰나미에 대비하지 못할까 우려된다"고 발언
- 경제학자들은 AI가 궁극적으로 생산성과 생활 수준을 높이겠지만, 단기적으로 수백만 사무직 일자리를 대체해 기존 고용보험과 같은 사회 안전망이 감당하기 어려울 수 있다고 우려

AI가 인간 사회를 이롭게 하도록 유도하는 정책 대응 촉구

■ 이번 성명서는 AI가 인간의 노동을 단순히 대체하기보다 인간을 보완하고 사회 전체에 이익이 되는 방향으로 발전할 수 있도록 방향을 유도해야 한다고 강조

- 이를 위해 관련 경제학자, 정책 입안자, 업계 리더들이 협력하여 적절한 인센티브, 안전장치 및 제도를 신속하게 구축할 것을 촉구했으나 구체적인 정책 권고는 미포함

■ 에릭 브린욜프슨 교수는 적절한 제도와 안전장치 마련을 위해 경제학계가 최우선 과제로 AI의 경제적 확산과 그 영향을 정확히 측정할 수 있는 방법론을 개발해야 한다고 촉구

- 현재는 신뢰할 수 있는 데이터가 부족하여 AI가 구체적으로 어떤 노동자에게 얼마나 영향을 미침으로써 일자리 상실을 주도하는지 파악하는 데 상당한 어려움에 직면
- 따라서 구체적인 대응 정책을 마련하기 위해 기술의 경제적 파급력을 제대로 이해하고 평가할 수 있는 데이터 수집과 제도 설계 등의 연구 기반 조성이 시급

출처 | The New York Times, Nearly 200 Economists and Tech Leaders Warn of A.I. Threats, 2026.7.13.

SPRi 소프트웨어정책연구소
Software Policy & Research Institute

https://spri.kr/

보고서와 관련된 문의는 AI정책연구실(hjk_rep@spri.kr, 031-739-7326)로 연락주시기 바랍니다.

경기도 성남시 분당구 대왕판교로 712번길 22 글로벌 R&D 연구동(B) 4층
22, Daewangpangyo-ro 712beon-gil, Bundang-gu, Seongnam-si, Gyeonggi-do, Republic of Korea, 13488