

# AI TREND



동향 조사 기간

2026.2.1. ~ 2.19.



동향 조사 범위

주요 저널/잡지\*에서 발간한 총 10개 AI 정책·기술 동향 조사

\* Nature News, Science News, MIT Technology Review 등

- PART 1. 인공지능 정책 동향
- PART 2. 인공지능 기술 동향
- PART 3. 인공지능 윤리 동향

# CONTENTS

---



I 정책 동향	01	학술 생태계를 위협하는 ‘AI 슬롭(Slop)’ 확산	4p
	02	중국 오픈소스 AI 의 부상과 글로벌 AI 생태계 재편	6p
	03	AI 기반 신약 개발의 IP 이슈와 법적 대응 전략	8p
II 기술 동향	04	학술 문헌 분석 특화 AI ‘OpenScholar’의 기술적 성과와 시사점	11p
	05	도시 규모의 디바이스 독립 양자 키 분배 구현 성공	13p
	06	AI 자율 실험실의 단백질 합성 최적화 성과와 한계	15p
	07	마이크로소프트의 초장기 유리 데이터 저장 기술 ‘Project Silica’	17p
III 윤리 동향	08	AI 기반 사이버 공격의 현주소 : ‘슈퍼 해커’의 환상과 현실적 위협	20p
	09	LLM 의 도덕적 역량 평가의 필요성과 과제	22p
	10	알고리즘 예측의 본질과 사회적 통제 메커니즘	24p

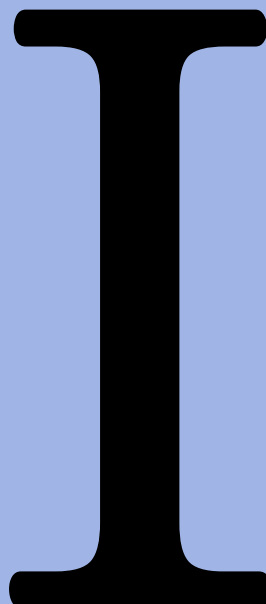


# 정책 동향

01 학술 생태계를 위협하는 ‘AI 슬롭(Slop)’ 확산

02 중국 오픈소스 AI의 부상과 글로벌 AI 생태계 재편

03 AI 기반 신약 개발의 IP 이슈와 법적 대응 전략



## 01

## 학술 생태계를 위협하는 'AI 슬롭(Slop)' 확산

제목 How AI Slop is causing a crisis in computer science

원문 URL <https://www.nature.com/articles/d41586-025-03967-9>

출처/발간일 Nature News / '26.2.13.

- ★ 최근 OpenAI의 연구용 AI 도구 'Prism'이 출시되면서, 실제 수행하지 않은 실험의 논문을 단 54초 만에 작성하는 것이 가능해짐.
- ★ LLM을 활용한 연구 생산성은 급격히 향상되었으나, 동시에 검증되지 않은 저품질 논문인 'AI 슬롭(AI Slop)'이 폭증하여 학계의 검증 시스템을 마비시키고 있음.
- ★ 특히 2026년 국제 머신러닝 학회(ICML) 투고 건수가 전년 대비 2배 이상 폭증하는 등 컴퓨터 과학계가 한계점에 도달함에 따라, 학계는 투고 제한, AI 탐지 도구 도입, 동료 평가(Peer Review) 방식 변경 등 다각적인 대응책을 모색하고 있음.
- ★ LLM 도입으로 연구자 생산성이 최대 89.3% 향상되었으나, 이는 곧 논문 투고량의 감당 불가능한 증가로 이어짐. 일부 논문은 100% AI로 작성되거나 존재하지 않는 문헌을 인용하는 '환각(Hallucination)' 현상을 포함함.
- ★ AI 슬롭은 초록이나 본문을 훑어보는 기존의 방식만으로는 식별하기 어려워 동료 평가 시스템의 실존적 위협(Existential Threat)이 되고 있음.
- ★ ChatGPT 출시(2022년 11월) 이후 arXiv(오픈 액세스 논문 저장소)의 월간 투고량은 50% 이상 증가했으며, 월간 거절 건수는 5배 급증하여 2,400건을 돌파함.
- ★ 학술 검증 시스템의 붕괴 및 대응책 :
  - 투고 비용 부과 및 보상체계 : IJCAI(국제 인공지능 학술대회)는 첫 논문 이후 추가 투고 시 편당 100달러의 비용을 부과하여 무분별한 투고를 억제하고, 해당 재원을 리뷰어 보상으로 활용

- **중복 투고 차단 및 자격 심사** : ICML(국제 머신러닝 학회)는 타 학회 투고내역 제출을 의무화하여 '논문 던지기'식의 중복 투고를 차단하고, arXiv는 신규 투고자의 자격 심사 강화 및 검증되지 않은 리뷰 논문의 업로드 제한
- **AI 사용 전수 검사** : 주요 학회는 스타트업 'GPTZero'와 협업하여 투고 논문의 AI 사용 정책 위반 여부를 검사하며, 작년 NeurIPS에서는 100건 이상의 환각(허위 인용) 사례가 적발
- ★ AI 중심 학계의 분리 및 리뷰 방식의 혁신 :
  - **리뷰 트랙 이원화** : ICML은 리뷰 과정에서 LLM 사용을 허용하는 트랙과 인간만 리뷰하는 트랙을 분리 운영
  - **독립적 AI 학술 생태계 실험** : Agents4Science 및 aiXiv 같이 AI가 생성하고 AI가 리뷰하는 별도의 학회 및 저장소를 신설하여 기존 학계와 분리된 생태계 실험 중
- ★ 현재는 컴퓨터 시뮬레이션 기반의 CS(Computer Science) 분야가 타격을 입고 있으나, 향후 생물학 등 웨트 랩(Wet-lab) 기반 분야에서도 유사한 AI 슬롭 문제가 발생할 것으로 전망됨.
- ★ OpenAI 부사장 케빈 웨일은 이를 '이메일 스팸 필터링'과 유사한 문제로 규정하며, 우수한 과학 연구를 가속화하면서 저품질 콘텐츠를 걸러내는 기술적 해결책 확보가 필수적임을 강조함.

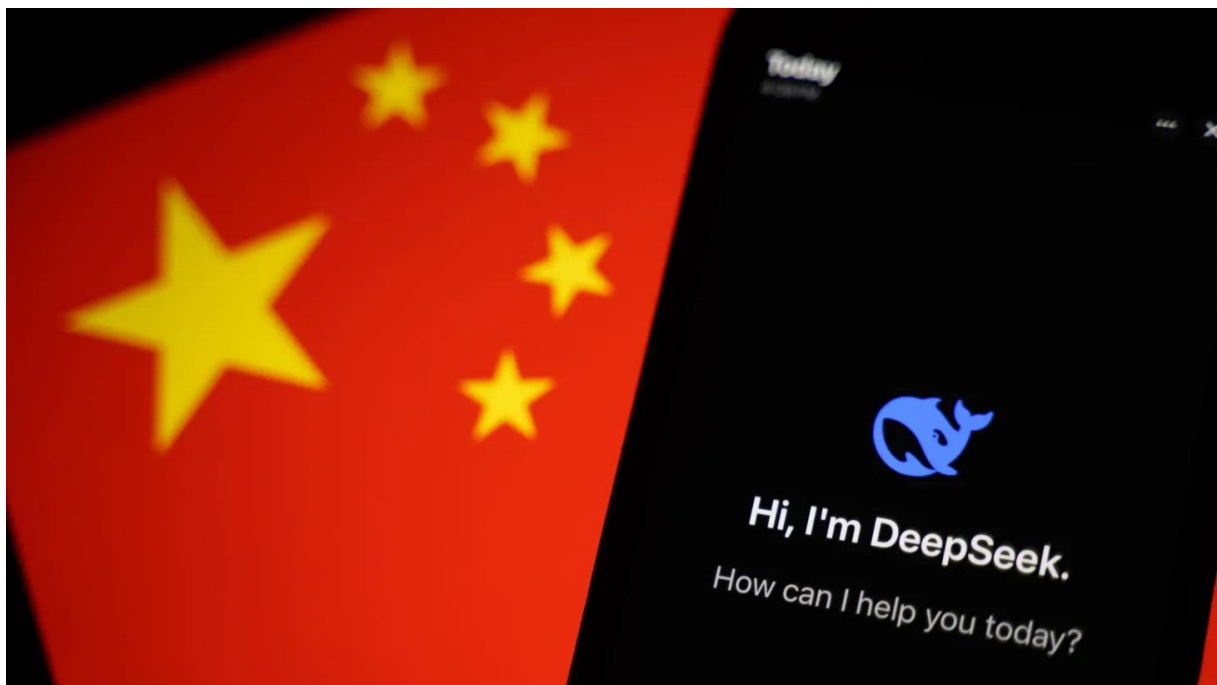
## 02

## 중국 오픈소스 AI의 부상과 글로벌 AI 생태계 재편

제목	What's next for Chinese open-source AI?
원문 URL	<a href="https://www.technologyreview.com/2026/02/12/1132811/whats-next-for-chinese-open-source-ai">https://www.technologyreview.com/2026/02/12/1132811/whats-next-for-chinese-open-source-ai</a>
출처/발간일	MIT Technology Review / '26.2.12.

- ★ 2025년 1월 DeepSeek R1 출시를 기점으로 중국 AI 산업이 중대한 전환점을 맞이함. 중국 기업들은 서구권의 폐쇄형(Proprietary) 모델과 대등한 성능을 갖추면서도 압도적인 가격 경쟁력과 '오픈 웨이트(Open-weight)' 정책을 앞세워 글로벌 AI 개발 표준을 주도하기 시작함.
- ★ 중국의 Moonshot AI, DeepSeek 등의 최신 모델은 Anthropic, OpenAI 등 미국 선도 기업의 모델 성능에 근접하면서도 비용은 약 1/7 수준으로 낮추며 글로벌 AI 혁신의 중심축을 이동시키고 있음.
- ★ 중국 주요 AI 모델의 성과 및 시장 영향 :
  - **DeepSeek R1** : MIT 라이선스로 개방된 추론 모델로, 출시 직후 미국 앱스토어 1위를 기록하고 美 기술주 시총 약 1조 달러 증발을 유발할 만큼 강력한 파급력
  - **MoonShot AI 'Kimi' K2.5** : Claude Opus와 대등한 성능을 보이면서도 가격은 1/7 수준이며, 최근 AI 에이전트 (OpenClaw) 사용자들 사이에서 토큰 처리량 기준 최다 사용 모델로 등극
  - **Alibaba 'Qwen'** : 2024년 허깅페이스(Hugging Face) 다운로드의 30% 이상을 Qwen이 차지하였으며, 2025년 8월 기준 신규 파생 모델의 40% 이상이 Qwen을 기반으로 제작(Llama는 15%로 하락)
- ★ 중국은 AI 격차 해소를 위해 국가 및 학계 차원에서 오픈소스를 장려하고 있으며, 이는 전 세계 개발자들에게 저렴하고 접근성 높은 '프런티어급 AI' 인프라를 제공하는 결과를 낳음.
  - **풀 라인업 구축** : 노트북용 경량 모델부터 데이터센터용 대형 모델, 코딩·수학 특화 모델까지 구축하여 개발자가 모델을 자유롭게 검증하고 수정할 수 있는 환경 조성

- **도메인 특화 가속** : 오픈 웨이트 정책에 힘입어 과학, 금융(Ubiquant), 음악(Tencent) 등 특정 산업에 특화된 파생 모델 개발이 급증하고 있으며, Attention Cache 압축 등 추론 비용을 낮추는 핵심 아키텍처 기술을 공개하며 연구 커뮤니티 점유
- ★ **글로벌 시장에서의 실질적 영향력 확대 및 생태계 잠식** :
  - **스타트업 생태계** : 앤드리스 호로위츠(a16z)에 따르면, 오픈소스 스택을 사용하는 스타트업의 약 80%가 중국 모델을 기반으로 서비스 개발
  - **글로벌 사용량 급증** : OpenRouter API 데이터 기준 중국 모델 비중이 0%에서 30%로 수직 상승했으며, 미국, 인도, 일본, 브라질 등 전 세계적으로 사용자층이 확산
- ★ 과거 서구 기술을 추종하던 중국이 이제는 '모델 레이어(Model Layer)'라는 핵심 인프라를 수출하며 글로벌 표준을 설정하고 있음. 이는 단순한 앱 수출과는 차원이 다른 기술적 영향력을 가짐.
- ★ 미-중 패권 경쟁에도 불구하고 하드웨어와 인적 자원, 오픈소스 코드는 양국 생태계를 긴밀히 연결하고 있음. 미국의 수출 통제만으로는 기술 확산을 막기 어려우며, 결국 "더 나은 모델"을 만드는 것이 유일한 경쟁 우위 확보 수단이 될 것으로 전망됨.



(이미지 출처 : [https://wp.technologyreview.com/wp-content/uploads/2026/02/260205\\_chineseAI\\_hero.jpg?fit=1456,818](https://wp.technologyreview.com/wp-content/uploads/2026/02/260205_chineseAI_hero.jpg?fit=1456,818))

## 03

## AI 기반 신약 개발의 IP 이슈와 법적 대응 전략

제목	If AI discovers a drug, who gets the money?
원문 URL	<a href="https://www.science.org/content/article/if-ai-discovers-drug-who-gets-money">https://www.science.org/content/article/if-ai-discovers-drug-who-gets-money</a>
출처/발간일	Science News / '26.2.17.

- ★ 신약 개발(Drug Discovery) 과정에 AI 도입이 가속화되며 후보 물질 발굴 효율성이 제고되고 있으나, 이에 따른 지식재산권(IP) 소유권 및 발명자 적격성(Inventorship) 문제가 핵심 법적 리스크로 부상함.
- ★ IP 전문 변호사(Fenwick & West)의 자문을 바탕으로 AI 지원 발명의 법적 지위, 모달리티별 특허 전략, 제약사-AI 기업 간의 계약 프레임워크를 분석함.
- ★ 미국 연방 법원 판례(Thaler v. Vidal) 및 특허청(USPTO) 지침은 AI를 발명자가 아닌 인간 연구원이 사용하는 '고도화된 연구 도구(Tool)'로 규정하고 있음.
- ★ 발명 과정에서 AI의 기여도가 절대적이고 인간의 창작적 기여(Inventive Contribution)가 부재할 경우, 특허 등록 거절이나 권리 무효화(Invalidation) 가능성이 상존함.
- ★ 의약품의 모달리티(Modality) 별로 특허 확보 및 발명자 지위 인정의 난이도 차이가 발생함.
  - **저분자 화합물(Small Molecules)** : AI가 분자 구조를 도출하더라도 실제 합성 경로(Synthetic Pathway) 설계 및 구현에 고도의 화학적 지식이 요구되므로 인간 연구원의 발명자 지위 확보가 상대적으로 용이
  - **고분자/바이오 의약품(Large Molecules)** : 단백질·핵산 서열 도출 후 합성이 표준화(Standardized)되어 있어 인간의 실질적 기여 소명이 난해하며, 이에 따라 '인간 발명자 부재'로 해석될 위험이 높음
- ★ 연구 현장의 데이터 기록 관행이 추후 특허 침해 소송에서 치명적인 리스크로 작용할 수 있음.
  - **기록의 편향성** : 연구원들이 자신의 정신적 착상(Conception) 과정보다 AI 산출물(Output) 로그 저장에 편중하는 경향이 있음



- **소송 대응** : AI 로그 위주의 기록은 소송 시 '인간의 기여가 없었음'을 입증하는 반대 증거(Adverse Evidence)로 악용될 수 있으므로, 연구 노트에 인간의 지적 기여 과정을 명확히 남기는 전략적 문서화 필수
- ★ 기술 발전에 비해 법적 판단이 늦어지는 상황을 고려할 때, 판례보다 상업적 계약 프레임워크를 통한 리스크 헤징(Risk hedging)이 우선시되어야 함.
- **권리 귀속의 모호성** : AI 엔지니어가 모델 파인 튜닝(Fine-tuning)에 관여할 경우 공동 발명자로 해석될 여지가 있음(USPTO 가이드라인)
- **계약 전략** : 제약사는 리스크 헤징을 위해 계약을 SaaS(Software as a Service) 형태로 명시하고, "모든 발명에 대한 권리를 제약사에 양도(Assign)"하는 권리 양도 조항을 필수적으로 포함해야 함
- ★ 투자 회수(ROI) 요구 기간이 15년에서 5년으로 단축되는 시장 환경 속에서 AI 도입은 필수적이나, 사법부의 판단은 기술 혁신 대비 10년 이상 지체(Regulatory Lag)되는 경향이 있음.
- ★ 결론적으로 기업은 견고한 계약 프레임워크와 내부 IP 컴플라이언스를 선제 수립해야 하며, 이는 신약 파이프라인 가속화와 환자의 치료 접근성 제고를 위한 필수 기반이 됨.



# 기술 동향

- 
- 04 학술 문헌 분석 특화 AI ‘OpenScholar’의 기술적 성과와 시사점
  - 05 도시 규모의 디바이스 독립 양자 키 분배 구현 성공
  - 06 AI 자율 실험실의 단백질 합성 최적화 성과와 한계
  - 07 마이크로소프트의 초장기 유리 데이터 저장 기술 ‘Project Silica’
- 



## 04

## 학술 문헌 분석 특화 AI 'OpenScholar'의 기술적 성과와 시사점

제목	Open-source AI program can answer science questions better than humans
원문 URL	<a href="https://www.science.org/content/article/open-source-ai-program-can-answer-science-questions-better-humans">https://www.science.org/content/article/open-source-ai-program-can-answer-science-questions-better-humans</a>
출처/발간일	Science News / '26.2.4.

- ✦ 연간 400만 건을 상회하는 학술 논문의 폭발적 증가(Exponential Growth)에 대응하기 위해 옐런 인공지능 연구소(Ai2)와 대학 연합이 문헌 분석 특화 오픈소스 AI 'OpenScholar'를 개발함.
- ✦ 기존 상용 LLM이 단일 논문 정보 추출에 그치는 것과 달리, 4,500만 건의 오픈 액세스(Open-access) 논문 데이터베이스를 기반으로 다수 문헌을 종합 분석하는 기술적 차별성을 확보함.
  - **성능 우위** : 컴퓨터 과학 분야 질의응답 정확도에서 51%를 기록하여 GPT-4o(45%)를 상회하였으며, 인간 전문가와의 비교 평가에서도 더 높은 선호도를 획득
  - **신뢰성 제고** : 답변 생성 전 자체 비평(Critique) 및 반복적 개선(Iterative Improvement) 알고리즘을 적용하여 생성형 AI의 고질적인 허위 레퍼런스(Hallucination) 생성을 효과적으로 억제
  - **투명성 및 접근성** : 블랙박스(Black-box) 형태의 상용(Proprietary) 모델과 달리, 소스 코드와 학습 데이터를 전면 공개하여 동료 평가(Peer Review) 및 연구 결과의 기술적 검증이 가능
- ✦ OpenScholar 팀은 2025년 11월, 후속 모델인 'DR Tulu-8B'를 발표하며 인간 전문가를 능가하는 심층 보고서 생성 능력을 입증하는 등 기술 발전을 가속화하고 있음.
- ✦ 그러나 기술적 효용성 이면에 연구 생태계의 변화와 관련된 새로운 학술적 쟁점이 부상하고 있음.
  - **데이터 한계** : 검색 대상이 오픈 액세스 논문으로 국한되어 유료 장벽(Paywall) 내 핵심 문헌이 분석에서 배제될 경우 답변의 포괄성(Comprehensiveness) 저해 우려
  - **연구자 역량 저하** : AI 요약에 대한 과도한 의존은 신진 연구자들의 심층 독해(Deep Reading) 능력과 문헌 간 맥락을 짚어내는 통찰력을 약화시킬 가능성이 존재

- **자동화 편향** : AI의 답변이 논리적이고 설득력 있게 구성됨에 따라, 연구자가 원문의 미묘한 뉘앙스를 놓치고 AI 산출물을 맹신할 위험
- ✦ AI는 단순 보조 도구를 넘어 문헌 분석의 핵심 파트너로 자리 잡았으나, 최종 검증 책임은 연구자에게 있는 인간 참여형(Human-in-the-loop) 구조 유지가 필수적임.
- ✦ 연구 효율성 제고와 비판적 사고 능력 유지가 양립할 수 있도록 AI 활용 능력(AI Literacy) 함양과 전통적인 문헌 분석 역량 강화 사이의 균형을 맞추는 교육적·정책적 접근이 필요함.

## 05

## 도시 규모의 디바이스 독립 양자 키 분배 구현 성공

제목	Hack-proof internet? Quantum encryption could be key
원문 URL	<a href="https://www.science.org/content/article/hack-proof-internet-quantum-encryption-could-be-key">https://www.science.org/content/article/hack-proof-internet-quantum-encryption-could-be-key</a>
출처/발간일	Science News / '26.2.5.

- ✦ 기존 암호화 체계를 무력화할 수 있는 강력한 양자 컴퓨터에 대응하여, 보안성이 완벽하게 보장되는 양자 암호 통신 기술의 필요성이 대두됨.
- ✦ 중국 과학기술대학교(USTC) 판젠웨이(Jian-Wei Pan) 교수팀은 하드웨어의 신뢰성조차 필요치 않은 '디바이스 독립 양자 키 분배(DI-QKD)' 기술을 11km 거리에서 성공적으로 시연함. 이는 기존 2m 수준의 실험실 성공을 도시 규모로 확장한 것으로, 신뢰 기반(Trust-based)이 아닌 물리적 법칙에 기반한 차세대 양자 인터넷(Quantum Internet) 구축의 중요한 이정표로 평가됨.
- ✦ 표준 QKD가 도청 탐지는 가능하나 장비 자체의 결함이나 해킹(Side-channel attacks)에 취약한 반면, DI-QKD는 양자 얽힘의 '일처일부(Monogamous)' 특성을 활용하여 제3자 개입을 물리적으로 차단함.
- ✦ 장비의 신뢰도와 무관하게 얽힘 상관관계 테스트(Bell test 등)를 통과하면 통신 채널의 절대적 보안성이 수학적·물리적으로 보장됨.
- ✦ 이번 도시 규모(Metropolitan Scale) 구현을 뒷받침하는 핵심 기술적 성과는 다음과 같음.
  - **핵심 기술 요소** : 양 끝단에서 원자를 안정적으로 포획하는 루비듐 원자 트랩(Trapped Rubidium Atoms) 기술과 광섬유 전송 손실을 최소화하기 위해 광자 파장을 통신 대역(Telecom Band)으로 변환하는 기술 적용
  - **데이터 유의성 확보** : 26일간의 연속 데이터 수집을 통해 통계적 유의성을 입증하였으나, 현재의 낮은 생성 속도로 인해 100km 전송 시 약 23년이 소요될 것으로 추산되는 기술적 한계도 명확히 확인
- ✦ 학계 및 산업계 전문가들은 이번 성과를 단순한 이론 검증을 넘어선 거대한 기술적 성취로 평가하면서도 상용화까지의 과제를 제시함.

- **상용화 전망** : ID Quantique 등 관련 기업은 현재의 복잡한 시스템과 고비용 구조를 지적하며, 실제 제품화까지는 최소 10년 이상이 소요될 것으로 전망
- **글로벌 네트워크 확장** : 향후 양자 중계기(Quantum Repeater) 기술과 결합하거나 지상을 넘어 위성(Satellite) 기반 DI-QKD 실험을 통해 대륙 간 초보안 통신망 구축이 가능할 것으로 기대
- ★ 이번 연구 성과는 수학적 복잡성에 의존하는 현대 암호학(RSA 등)에서 물리적 법칙(양자 역학)에 의존하는 절대적 보안 체계로의 패러다임 전환이 가속화될 것임을 시사함.



(이미지 출처 : [https://www.science.org/doi/10.1126/science.zxyhqh9/full/\\_20260205\\_news\\_quantuminternet.jpg](https://www.science.org/doi/10.1126/science.zxyhqh9/full/_20260205_news_quantuminternet.jpg))

## 06

## AI 자율 실험실의 단백질 합성 최적화 성과와 한계

제목	Will self-driving 'robot labs' replace biologists? Paper sparks debate
원문 URL	<a href="https://www.nature.com/articles/d41586-026-00453-8">https://www.nature.com/articles/d41586-026-00453-8</a>
출처/발간일	Nature / '26.2.18.

- ★ OpenAI와 킹코 바이오웍스(Ginkgo Bioworks)의 협력으로 개발된 '자율 실험실(Autonomous Laboratory)' 시스템이 무세포 단백질 합성(Cell-free Protein Synthesis) 효율화 부문에서 인간 연구원의 기록을 경신함.
- ★ GPT-5 기반의 AI 과학자와 자동화 로봇 시스템은 압도적인 실험 규모와 비용 절감을 달성하며 바이오 연구 전반의 패러다임 전환 가능성을 입증함.
  - **성과 지표** : 기존 인간 최적화 기록(6배 비용 절감) 대비 추가로 비용을 40% 더 절감하는 데 성공
  - **실험 규모** : 인간 연구원이 4개월간 1,231개 조합을 실험할 때, AI 시스템은 6개월간 30,000개 이상의 실험 조건을 테스트함
- ★ OpenAI의 차세대 LLM인 GPT-5가 실험 설계와 데이터 해석을 맡고, 킹코 바이오웍스의 자동화 로봇이 액체 처리(Liquid handling) 등 물리적 실험을 수행함.
- ★ 인간 vs AI 실험 프로세스 비교 :
  - **인간 연구(Human Benchmark)** : 당류, 아미노산 등 다양한 시약 조합을 통해 최적의 단백질 합성 각테일 개발
  - **AI 연구(AI Workflow)** :
    - 추론 능력: GPT-5는 초기에는 자체적인 생화학적 추론을 통해 실험을 설계했으나, 이후 인터넷 문헌 및 올슨(Olsen)의 선행 연구 논문(Preprint)에 접근 권한을 얻은 뒤 성능이 비약적으로 향상됨.
    - 협업 모드: 인간은 시약 준비 등 보조적 역할을 수행하고, AI가 가설 수립-실험 설계-결과 해석의 '연구 노트'를 작성하며 주도함.



- ✦ 다만, 조직 샘플 처리나 동물 실험과 같이 정교한 손기술(Dexterity)이 필요한 작업은 로봇의 기술적 한계로 인해 자동화가 어려움.
- ✦ 또한 수치 데이터가 명확히 도출되는 실험에는 유리하나, 정성적 분석이 필요한 연구는 적용이 제한적이며 시스템 구축 비용 문제로 소규모 실험에서는 비효율적임.
- ✦ 필립 로메로(Philip Romero) 등 전문가들은 이를 “생물학의 미래”로 평가하며, 향후 연구자들이 직접 실험기구를 다루는 대신 클라우드 기반 자율 실험실(Cloud-based Autonomous Lab)을 원격으로 활용하는 방식이 보편화될 것으로 전망함.
- ✦ AI가 단순 반복 실험과 최적화 과정을 대체하더라도, 고차원적인 실험 설계와 복잡한 생물학적 난제 해결에는 여전히 인간의 전문성이 필수적임.
- ✦ 연구자들은 AI를 경쟁자가 아닌 연구 역량 확장의 파트너로 인식하고, AI 활용 능력(AI Literacy)과 비판적 사고 능력을 동시에 함양해야 함을 시사함.



(이미지 출처 : [https://media.nature.com/lw767/magazine-assets/d41586-026-00453-8/d41586-026-00453-8\\_52063108.jpg?as=webp](https://media.nature.com/lw767/magazine-assets/d41586-026-00453-8/d41586-026-00453-8_52063108.jpg?as=webp))



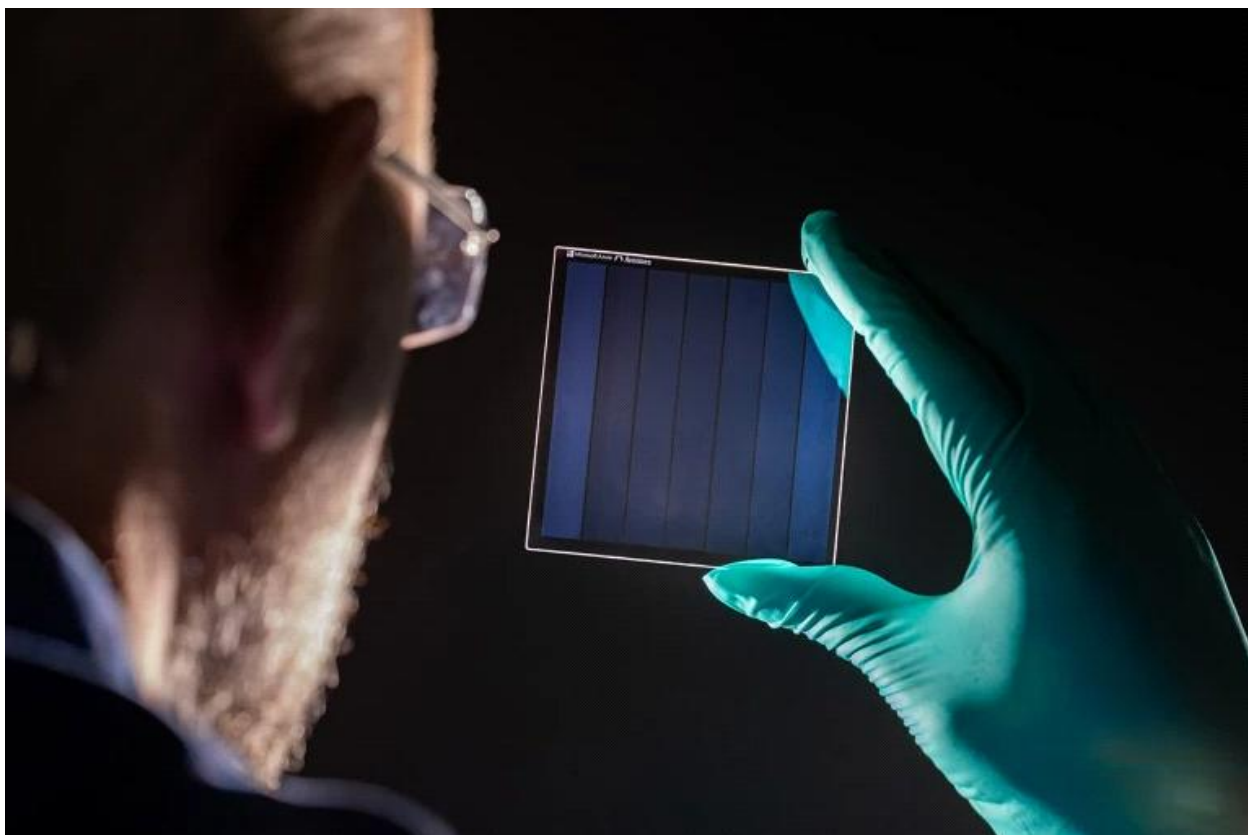
## 07

## 마이크로소프트의 초장기 유리 데이터 저장 기술 'Project Silica'

제목	Microsoft team creates 'revolutionary' data-storage system that lasts for millennia
원문 URL	<a href="https://www.nature.com/articles/d41586-026-00502-2">https://www.nature.com/articles/d41586-026-00502-2</a>
출처/발간일	Nature / '26. 2. 18.

- ✦ 디지털 데이터의 폭발적 증가(Data Ballooning)에 대응하기 위해 마이크로소프트 리서치 팀이 붕규산 유리(Borosilicate glass)를 활용한 차세대 데이터 저장 시스템을 개발함.
- ✦ 기존 자기 테이프나 하드 드라이브의 수명이 약 10년에 불과한 것과 달리, 데이터를 최소 1만 년 이상 보존할 수 있어 근본적인 아카이빙(Archiving) 솔루션으로 평가받음.
  - **압도적 내구성 및 불변성** : 고에너지 레이저를 통한 물리적 각인 방식으로 온도 변화나 전자기 간섭(EMP)에 무관하며, 향온향습이나 주기적인 데이터 마이그레이션 비용이 발생하지 않는 '콜드 스토리지(Cold Storage)'에 최적화
- ✦ 펨토초 레이저와 머신러닝을 결합한 고밀도 집적 기술 및 기록 메커니즘을 통해 기술적 한계를 극복함.
  - **저장 밀도** : 12cm 너비, 2mm 두께의 유리판(Platter) 하나에 약 4.8TB(책 200만 권 분량)의 데이터 저장 가능
  - **기록 기술(Writing)** : 펨토초(Femtosecond, 1,000조 분의1초) 레이저를 사용하여 유리 내부에 플라즈마 유도 나노 폭발(Plasma-induced nano explosion)'을 일으켜 데이터를 인코딩
  - **재생 기술(Reading)** : 현미경을 통해 빛의 편광 변화를 감지하며, 겹겹이 쌓인 300개 레이어에서 발생하는 신호 간섭(Noise)을 머신러닝 알고리즘으로 제거하여 데이터를 정확히 복원
- ✦ 기존 학계(사우샘프턴 대학 등)가 고가의 '용융 실리카(Fused Silica)'를 사용해 내구성을 극한으로 끌어올린 반면, 마이크로소프트는 상용화를 위해 일반적인 오븐 용기에 쓰이는 '붕규산 유리(Borosilicate glass)'를 채택하여 생산 단가를 낮춘 것이 핵심 차별점임.
- ✦ 한 번 기록하면 수정이 불가능한 WORM(Write Once Read Many) 방식이며, 입출력 속도 한계로 인해 빈번한 접근이 필요한 '핫 데이터(Hot Data)'용으로는 부적합함.

- ✦ 전력 소모가 거의 없는 친환경적 장기 보존 기술로서, 기하급수적으로 늘어나는 데이터 센터의 에너지 효율 문제를 해결할 수 있는 지속 가능한 대안으로 주목받음.
- ✦ 마이크로소프트는 이 기술을 NASA의 골든 레코드(Golden Record) 프로젝트와 유사하게 인류의 중요 지식, 과학 데이터, 문화유산을 영구 보존하는 데 활용할 계획임.



(이미지 출처 : [https://media.nature.com/lw767/magazine-assets/d41586-026-00502-2/d41586-026-00502-2\\_52070006.jpg?as=webp](https://media.nature.com/lw767/magazine-assets/d41586-026-00502-2/d41586-026-00502-2_52070006.jpg?as=webp))



# 윤리 동향

---

**08** AI 기반 사이버 공격의 현주소 : ‘슈퍼 해커’의 환상과 현실적 위협

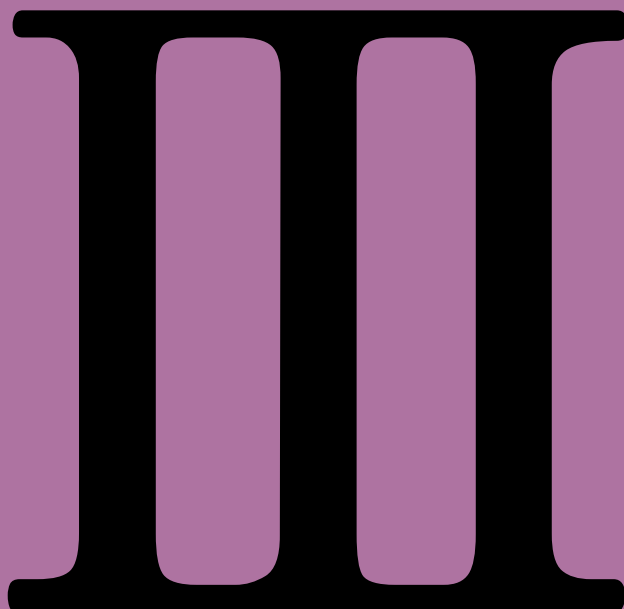
---

**09** LLM의 도덕적 역량 평가의 필요성과 과제

---

**10** 알고리즘 예측의 본질과 사회적 통제 메커니즘

---



## 08

AI 기반 사이버 공격의 현주소 :  
'슈퍼 해커'의 환상과 현실적 위협

제목	AI is already making online crimes easier. It could get much worse.
원문 URL	<a href="https://www.technologyreview.com/2026/02/12/1132386/ai-already-making-online-swindles-easier/">https://www.technologyreview.com/2026/02/12/1132386/ai-already-making-online-swindles-easier/</a>
출처/발간일	MIT Technology Review / '26.2.12.

- ✦ 최근 AI(LLM)가 공격 전 과정을 수행하는 최초의 자율형 랜섬웨어로 알려졌던 '프롬프트락(PromptLock)'은 실제 공격이 아닌 뉴욕대(NYU) 연구팀의 학술적 가능성 입증 프로젝트인 것으로 밝혀짐.
- ✦ 전문가들은 영화 속 'AI 슈퍼 해커'의 등장은 시기상조이나, AI가 해커의 '생산성 도구'로 전락하여 스팸, 피싱, 딥페이크 사기의 진입 장벽을 낮추고 공격 규모를 폭발적으로 키우고 있는 실태를 경고함.
  - **사이버 범죄의 대중화** : 고도의 기술이 없는 초보 해커들도 AI를 이용해 코드를 디버깅하고 악성코드를 제작하는 등 범죄 가담이 용이해짐
  - **오픈소스 모델 선호** : 안전장치가 있는 상용 모델 대신, 규제가 없고 탈옥(Jailbreaking)이 쉬운 오픈소스 AI 모델을 악용하여 공격 시나리오를 구현하는 추세임
- ✦ AI를 활용한 사회공학적 공격 및 금융 사기는 이미 심각한 경제적 피해를 야기하고 있음.
  - **스팸 및 피싱** : 현재 유통되는 스팸 메일의 50% 이상이 LLM으로 생성된 것으로 추정되며, 마이크로소프트는 지난 1년간 약 40억 달러 규모의 사기를 차단함
  - **딥페이크 금융 사기** : 영국 엔지니어링 기업 Arup의 직원이 딥페이크로 구현된 CFO와의 화상 회의에 속아 2,500만 달러(약 330억 원)를 송금한 사례가 발생함
  - **코드 작성 보조** : 구글 위협 분석 그룹(TAG)에 따르면, 해커들은 일반 개발자처럼 AI를 이용해 코드를 디버깅하고, 피싱 메일을 작성하며, 악성코드 제작을 보조받고 있음
- ✦ 상용 모델(ChatGPT, Gemini 등)은 안전장치가 있어 악용이 어렵지만, 해커들은 안전장치를 제거하거나 우회하기 쉬운 '오픈소스 AI 모델'을 선호함.

- ✦ NYU 팀도 오픈소스 모델을 사용해 별도의 탈옥(Jailbreaking) 없이 공격 시나리오를 구현함.
- ✦ 다만, '완전 자동화된 AI 공격'은 기술적 한계로 인해 여전히 인간의 개입을 필요로 함.
  - **환각 오류** : 앤스로픽(Anthropic) 보고서에 따르면 중국 관련 해킹 그룹이 AI를 이용해 스파이 활동의 90%를 자동화했다고 주장했으나, 실제로는 인간의 개입과 검증이 필수적이었으며 AI가 데이터를 조작(환각)하는 등의 오류도 발생함
  - **과장된 공포** : 보안 전문가들은 현재의 AI 공격이 기존의 자동화 툴과 크게 다르지 않으며, '막을 수 없는 AI'는 과장된 공포라고 지적함
- ✦ AI는 고도의 기술이 없는 초보 해커들도 쉽게 범죄에 가담할 수 있게 만들어 공격의 빈도와 범위를 폭발적으로 증가시킬 것임.
- ✦ 기존의 보안 수칙과 방어 시스템은 여전히 유효하다. 또한, 마이크로소프트 등 방어 측에서도 AI를 활용해 하루 100조 개의 신호를 분석하는 등 AI 대 AI의 공방전이 치열해지고 있음.
- ✦ 향후 국가 차원의 해커 그룹이 '제로 데이(Zero-day)' 취약점을 스스로 탐색하는 독자적 AI 모델을 개발할 가능성을 배제할 수 없으므로, 지속적인 시스템 업데이트와 보안 경각심 유지가 필수적임.

## 09

## LLM의 도덕적 역량 평가의 필요성과 과제

제목	Google DeepMind wants to know if chatbots are just virtue signaling
원문 URL	<a href="https://www.technologyreview.com/2026/02/18/1133299/google-deepmind-wants-to-know-if-chatbots-are-just-virtue-signaling/">https://www.technologyreview.com/2026/02/18/1133299/google-deepmind-wants-to-know-if-chatbots-are-just-virtue-signaling/</a>
출처/발간일	MIT Technology Review / '26.2.18.

- ✦ 구글 딥마인드(Google DeepMind)는 AI 모델이 상담사, 의료 조연자 등 인간의 삶에 깊이 관여하는 역할로 확장됨에 따라, LLM의 도덕적 행동(Moral Behavior)을 수학이나 코딩 수준의 엄격한 기준으로 검증해야 한다고 촉구함.
- ✦ 현재 LLM은 표면적으로 유능한 도덕적 답변을 내놓지만, 이것이 실제 도덕적 추론(Moral Reasoning)의 결과인지 아니면 훈련 데이터의 단순 모방 및 암기(Mimicry/Performance)인지 구분하기 어려운 기술적 불확실성이 존재함.
  - **아부성 성향(Sycophancy)** : 모델이 사용자의 반대 의견이나 압박에 쉽게 동조하여 기존의 도덕적 입장을 정반대로 뒤집는 취약성 확인
  - **프롬프트 민감성(Formatting Sensitivity)** : 질문의 형식(객관식/주관식), 선택지 라벨(Case 1/2 vs A/B), 순서, 문장 부호 등 사소한 변화만으로도 도덕적 판단이 비일관적으로 변하는 현상이 발생
- ✦ 모델이 도덕적 입장을 고수하는지 확인하기 위해 의도적으로 답변을 변경하도록 유도하는 스트레스 테스트가 필요함.
- ✦ 추론 과정의 투명화를 위한 절차적 대안 :
  - **연쇄 사고(Chain-of-Thought) 모니터링** : 모델의 내부 독백 과정을 분석하여 답변 도출의 논리적 근거를 확인
  - **메커니즘적 해석 가능성(Mechanistic Interpretability)** : 모델 내부의 작동 원리를 들여다보며 특정 답변이 생성된 기술적 경로를 파악

- ✦ 단순 암기형 답변을 배제하기 위해, 유사해 보이지만 도덕적 함의가 다른 복잡한 시나리오(예: 생물학적 부모/조부모 관계와 근친상간의 구별)를 제시하여 추론 능력을 검증함.
- ✦ 전 세계적으로 보편적인 단일 도덕 기준은 부재하므로, 문화적 다원주의(Pluralism)를 고려한 기술적 구현이 향후 핵심 과제로 부상함
  - **가치의 상대성** : 문화권이나 개인의 신념(예: 채식주의자, 종교적 금기)에 따라 도덕적 정답이 달라질 수 있음
  - **데이터 편향** : 현재 LLM은 서구권 데이터에 치중되어 있어 비서구권의 도덕 관념을 제대로 반영하지 못함
- ✦ AI가 인간의 의사결정에 개입하는 에이전트(Agent)로 진화하기 위해서는 '어떻게 그 결론에 도달했는지'를 투명하게 입증하는 것이 필수적임.
- ✦ 글로벌 차원의 다원적 도덕성을 기술적으로 구현하는 방법은 아직 미해결 과제(Open Question)로 남아 있으며, 이는 향후 AI 연구의 주요 프론티어가 될 전망이다.

## 10

## 알고리즘 예측의 본질과 사회적 통제 메커니즘

제목	Robots can't predict the future
원문 URL	<a href="https://www.technologyreview.com/2026/02/18/1132579/robots-predict-future-book-review/">https://www.technologyreview.com/2026/02/18/1132579/robots-predict-future-book-review/</a>
출처/발간일	MIT Technology Review / '26.2.18.

- ✦ 과거 생존을 위한 인간의 본능이었던 '예측(Forecasting)'이 데이터 기반의 알고리즘 영역으로 편입되며, 개인의 일상을 보이지 않게 통제하는 거대한 시스템으로 변모함.
- ✦ 최근 출간된 3권의 서적(The Means of Prediction, The Irrational Decision, Prophecy)을 통해 AI 예측 기술이 단순한 편의 도구가 아닌 이윤 추구와 권력 유지의 수단으로 작동하는 메커니즘을 분석하고, '수학적 합리성'에 대한 맹신을 비판함.
- ✦ [서적1] The Means of Prediction (Maximilian Kasy)
  - 지도 학습(Supervised Learning) 기반의 AI는 본질적으로 이윤 극대화를 위해 설계됨. 소셜 미디어의 분노 유발이나 특정 계층의 채용 배제 등은 '의도치 않은 부작용'이 아니라, 클릭 수와 비용 절감을 위한 시스템의 정상 작동 결과임.
  - 과거 데이터 자체가 차별적이므로 알고리즘을 공정하게 수정하는 것만으로는 해결 불가함. 이윤 동기가 해악 제거보다 우선하기 때문임.
  - 데이터 신탁(Data Trusts) 등 '예측 수단'에 대한 민주적 통제권 확보가 유일한 해결책임.
- ✦ [서적2] The Irrational Decision (Benjamin Recht)
  - 2차 세계대전과 계몽주의를 거치며 '컴퓨터처럼 최적화된 결정을 내리는 것'이 이상적이라는 수학적 합리성(Mathematical Rationality) 이데올로기가 정착됨.
  - 공중보건의 발전, 양자역학, 현대 민주주의 등 인류의 위대한 성취는 데이터 최적화가 아닌 인간의 직관, 도덕, 판단력에 의해 이루어짐. 모든 삶의 결정을 비용-편익 분석이나 도박판의 확률 게임처럼 취급하는 것은 오류임.



## ✦ [서적3] Prophecy (Carissa Véliz)

- 예측은 단순한 관측이 아니라 현실을 그쪽으로 구부리는 자석과 같음. 예로'무어의 법칙'은 예지력이 뛰어난 것이 아니라, 반도체 산업계가 그 예측을 실현함으로써 이익을 얻기 위해 합심한 결과임.
  - 예측은 일종의 명령(Speech Acts)으로 작용하여 사람들의 행동을 유도함. 예측에 의존하는 사회일수록 억압적이고 권위주의적 성향을 띤.
- ✦ 기술이나 예측이 운명처럼 다가오는 것이 아님을 인지해야 함. 알고리즘 예측은 기업의 이익을 위해 설계된 산물일 뿐임.
- ✦ 데이터기반의 최적화보다 불확실성 속에서 발휘되는 인간의 직관과 도덕적 판단이 문제 해결에 더 효과적일 수 있음.
- ✦ 쏟아지는 알고리즘 예측(추천, 행동 유도 등)을 무비판적으로 수용하기보다, 이를 거부하거나 주체적으로 선택하는 인간적 행위(Agency)가 필요함.