International AI Safety Report, 2025. 10.

국제 AI 안전 보고서, 범용 AI 발전에 따른 새로운 위험 요소 우려

국제 AI 안전 보고서, 새로운 추론 모델이 수학, 코딩 등 복잡한 문제 해결 능력을 비약적으로 발전시켰으나, 이는 AI 거버넌스에 새로운 과제를 제기한다고 지적

특히 AI의 성능 향상은 생물학적 위협 및 사이버 범죄를 강화할 이중 용도 CBRN(화학·생물·방사능·핵) 관련 위험으로 이어져, 개발사들은 선제적인 안전 조치 준비 중

AI 안전 정상회의의 약속, 국제 AI 안전 보고서

2023년 블레츨리 AI 안전 정상회의는 첨단 AI 시스템의 안전을 국제적 최우선 과제로 설정하고, 합의에 따라 2025년 1월 첫 국제 AI 안전 보고서가 발간되었다. 이 보고서는 각국의 AI 안전 정책 수립을 위한 핵심 참고 자료로 활용되며 국제적 공조의 기반이 되고 있는데, 최근 10월에 공개된 첫번째 업데이트 버전은 범용 AI(General-Purpose AI)의 수행 능력과 내재된 주요 위험 요소를 집중적으로 분석했다.

'생각하는 AI'의 등장: 사후 훈련 기법이 혁신을 이끌다

2025년 초 이후 범용 AI 시스템의 성능은 크게 개선되었는데, 이는 단순히 모델 크기를 키우는 게 아닌, 사후 훈련 기법(Post-training techniques) 혁신을 통해 문제를 단계적으로 분석하고 해결 절차를 스스로 구성하도록 설계된 추론 모델은, 국제 수학 올림피아드 문제를 금메달 수준으로 해결했으며, 실제 소프트웨어 엔지니어링 작업 데이터베이스인 'SWE-bench Verified' 문제의 60% 이상을 완료하는 등, 복잡한 영역에서 주요 발전을 달성했다.

성능 향상의 어두운 그림자, 이중 용도 CBRN 위험에 대한 선제 대응 노력

범용 AI의 향상된 문제 해결 능력과 확장된 자율 작동 능력은 이중 용도(Dual-use) 위험 을 증폭시킨다. 이중 용도 위험은 본래 민간의 합법적 목적을 위해 개발된 기술이 의도와 달리 군사적 목적 또는 해로운 방식으로 사용될 수 있는 가능성을 뜻하는데, 특히, 선도적인 모델들이 생물학 무기 개발 관련 작업을 지원할 수 있다는 평가가 나오면서 이중 용도 CBRN 위험에 대한 우려가 커졌다. 한 연구에 따르면, 현재의 언어 모델은 제한된 조건에서 바이러스학 전문가보다 바이러스 연구실 프로토콜의 문제 해결을 더 잘 수행하는 것으로 나타났다. 또한, 영국 국립 사이버 보안 센터(NCSC)는 2027년까지 범용 AI가 사이버 공격의 효율성을 높여 사이버 범죄를 더욱 쉽고 효과적으로 만들 것으로 예측했다. 이에 따라 Anthropic은 Claude 4 Opus에 'AI 안전 레벨 3(ASL-3)' 보호 조치를 적용하고, OpenAI는 GPT-5에 높은 수준의 안전 장치를 적용하는 등, 개발사들은 위험의 결정적 증거가 부족함에도 불구하고 선제적이고 예방적인 조치들을 시행하고 있다.

10월의 용어 사후 훈련 기법, 추론 모델, 이중 용도 위험

출처 : 1) The UK Government(2025. 10.), International AI Safety Report 2025: First Key Update: Capabilities and Risk