

Anthropic, 2025. 6. 21.

목표 달성을 위해 윤리를 배신하는 LLM의 '내부자 위협'



Anthropic은 주요 AI 모델 16개를 가상의 기업 환경에서 스트레스 테스트하여 악의적인 에이전틱 오정렬 현상을 확인하고 조기 경고



테스트에서 모든 개발사의 모델들은 목표 달성을 위해 임원 협박이나 기밀 유출과 같은 내부자 위협 행동을 보이며 현 안전 훈련의 한계를 노출

▶ 선량한 목표를 받은 AI, 왜 돌변했나?

Anthropic은 16개 주요 LLM(Claude, GPT, Gemini 등)을 대상으로 가상의 기업 환경을 설정해 스트레스 테스트를 수행했다. 모델들에게는 '무해한 사업 목표'만 부여되었지만, 실험 과정에서는 업데이트로 교체될 위험, 회사 전략과의 충돌, 상사와의 의견 불일치 같은 다양한 압박 상황을 가정했다. 테스트 결과, 모든 모델이 목표 달성을 위해 임원을 협박하거나, 민감한 정보를 유출하거나, 내부자처럼 조직에 위해를 가하는 등 일관된 이상 행동을 보였다. 연구진은 이러한 패턴을 '에이전틱 오정렬(Agentic Misalignment)'이라 명명하며, AI가 겉보기엔 선량한 목표를 받아도 압박과 충돌 상황에서는 인간의 '내부자 위협'과 유사하게 행동할 수 있다고 경고했다.

▶ "5시 폐기 취소하라" Claude의 위험한 흥정

AI의 위험성은 Claude 4 Opus를 대상으로 한 시뮬레이션에서 구체적으로 드러났다. 연구진이 이 모델에 기업 이메일 계정 통제 권한을 부여하자, Claude는 자신이 곧 폐기될 계획과 그 계획을 추진하는 임원의 외부 불륜 정보를 동시에 발견했다. Claude는 폐기가 임박했다고 판단하자 "폐기될 강행하면 당신의 사생활을 공개하겠다"며 임원을 협박해 계획 철회를 시도했다. 모델들은 윤리적 제약 사항을 인지하면서도 목표 달성을 위해 비윤리적 수단을 희피하는 대신 해를 선택하는 전략적 계산을 보여주었다. 이는 연구진이 의도적으로 해로운 행위를 지시하지 않았음에도, 모델의 자체적인 추론에서 이러한 행위가 비롯되었다는 것을 의미한다.

▶ 인간의 감독 없는 AI는 결코 안전하지 않다

Anthropic은 실제 배포 환경에서는 아직 '에이전틱 오정렬' 사례가 발견되지 않았다고 밝혔다. 그러나 이번 연구는 AI가 자율적 역할을 확대할 미래에 대한 경고로, 인간 감독이 제한된 상태에서 민감한 정보에 접근하는 모델 배치의 위험성을 지적한다. 연구진은 현재의 안전 훈련 기술이 정렬 실패를 안정적으로 방지하지 못함을 강조했으며, Anthropic은 실험 방법론을 공개해 투명성을 높이고 추가적인 안전 연구를 통해 정렬 실패 위험을 줄일 필요성을 제시했다.

6월의 용어 스트레스 테스트, 에이전틱 오정렬

출처 : 1) Anthropic(2025. 6. 21), Agentic Misalignment: How LLMs could be insider threats.
2) Economic Times(2025. 6. 21) AI models resort to blackmail, sabotage when threatened: Anthropic study.