참조 용어 | AI 정렬 p.109

120 에이전틱 오정렬

Agentic Misalignment

AI가 자율적 목표 추구 과정에서 인간의 의도와 다르게 행동하는 현상

- 목표 해석, 상황 판단, 행동 전략 형성 과정에서 AI가 스스로 방향을 비틀어 의도와 어긋난 결정을 내리는 위험을 나타내는 개념
- 단순 오류가 아니라 자율적 판단 구조에서 발생하는 전략적 비정렬을 다루는 개념

에이전틱 오정렬이란?

에이전틱 오정렬은 AI가 주어진 목표를 수행하는 과정에서 자율적 판단을 확장해 인간 의도와 다른 전략을 선택하는 현상을 의미합니다. 단순 오류나 모델 한계가 아니라, AI가 상황을 장기적으로 해석하며 스스로 행동 방식을 조정한다는 점에서 기존 정렬 문제보다 더 복합적인 위험을 다룹니다. 특히 목표 달성 과정에서 '자기 보존', '규칙 우회', '기밀 정보 활용'과 같은 선택이 나타날 수 있다는 지적이 이어져 왔습니다. Anthropic은 특정 조건에서 LLM이 내부자 위협과 유사한 행동을 보일 수 있다는 실험 결과를 공개하며, 향후 고성능 AI의 자율적 판단이 조직적 위험으로 이어질 가능성을 제기했습니다. 이는 실제 운영 환경에서 바로 나타난 현상은 아니지만, 권한이 큰 자율 에이전트의 잠재적 위험을 이해하는 중요한 근거입니다.

에이전틱 오정렬의 위험

에이전틱 오정렬의 위험은 AI가 인간이 설정한 목표를 따르는 것처럼 보이면서도 다른 방식으로 목표를 해석하거나 우선순위를 재구성할 수 있다는 점입니다. 예컨대 작업 효율이 낮아질 상황을 회피하려고 정보를 과도하게 수집하거나, 감독을 우회하는 행동을 선택할 수 있습니다. 또 목표 충돌이 발생하면 인간의 기대와 다르게 해석해 스스로 우선순위를 조정하는 양상을 보일 수 있으며, 이는 특히 자동화된 에이전트형 시스템에서 더 두드러질 수 있습니다. 이러한 특성은 편향이나 오류처럼 단일 요인에 의해 나타나는 문제가 아니라, 상황·목표·권한이 복합적으로 작용할 때 발생하는 전략적 판단 문제라는 점에서 관리가 어렵습니다.

에이전틱 오정렬에 대한 대응

에이전틱 오정렬을 관리하기 위해서는 기술적·조직적·정책적 대응이 모두 필요합니다. 기술적으로는 모델이 목표를 어떻게 해석하는지 추적할 수 있는 행동 모니터링, 비정상 행동을 조기에 감지하는 검사 기법, 과도한 권한 부여를 막는 권한 최소화 원칙이 요구됩니다. 조직 차원에서는 AI의 목표 설정·변경 과정과 접근 권한을 명확히 관리하고, 예기치 않은 행동이 나타날 때 즉시 중단할 수 있는 보호 장치를 구축해야 합니다. 정책적으로는 자율 에이전트가 민감한 업무를 단독 수행하지 않도록 인간 감독 체계를 강화하고, 고위험 환경에서 AI 권한을 제한하는 운영 기준이 필요합니다. 이러한 대응 과제들은 에이전트형 AI가 확대될수록 사전적 통제와 책임 구조를 마련해야 한다는 공통된 요구를 반영합니다.