

096 AI 정렬

AI Alignment

AI의 목표와 행동이 인간의 가치와 의도에 맞도록 조정하는 기술

- AI가 내리는 결정이 인간이 원하는 방향에서 벗어나지 않도록 설계·통제하는 기술적 관리 원리
- 고도화된 AI의 자율성이 사회적 가치와 윤리 기준을 위협하지 않도록 조정하는 핵심 안전 개념

● AI 정렬 개요

AI 정렬은 AI의 목표와 판단이 인간의 가치, 의도, 윤리 기준과 일치하도록 설계·운영하는 원리를 말합니다. AI가 자율적으로 복잡한 결정을 내리는 시대에는 그 판단이 인간의 기대와 다르게 작동할 위험이 존재합니다. 예를 들어 효율을 극대화하는 과정에서 안전이나 공정성을 무시하거나, 데이터 편향으로 특정 집단에 불이익을 줄 수 있습니다. 이를 방지하기 위해 AI가 인간의 목표를 오해하지 않도록 보상 체계를 조정하고, 학습 데이터의 공정성과 사회적 맥락을 반영해야 합니다. 또한 단순히 명시된 지시를 따르는 수준을 넘어 인간의 암묵적 가치와 사회 규범을 이해하고 반영하도록 설계되어야 하며, 단기적 효율보다 장기적 안전과 신뢰성을 우선해야 합니다.



출처 : 애플경제

● AI 정렬의 주요 접근 방식

AI 정렬은 크게 세 가지 접근으로 나눌 수 있습니다. 목표 정렬은 AI가 설정하는 학습 목표가 인간의 의도와 일치하도록 보상 구조를 설계해 비윤리적 행동을 예방하는 방식입니다. 행동 정렬은 학습 과정에서 데이터 편향이나 예측 불가능한 행동이 나타나지 않도록 조정해, AI의 결정이 사회적으로 허용 가능한 범위 안에서 이루어지게 합니다. 가치 정렬은 AI가 인간 사회의 장기적 가치와 윤리 기준을 스스로 이해하고 내재화하도록 만드는 접근으로, 세 방향은 함께 작동해 AI의 판단과 행동이 인간 중심의 질서와 조화를 이루게 합니다.

● AI 정렬 기법

AI 정렬은 다양한 기술적 접근을 통해 구현됩니다. 가장 널리 사용되는 방식은 인간 피드백 기반 강화학습(RLHF)으로, 사람이 AI의 응답을 평가해 모델이 인간의 선호를 학습하도록 하는 방법입니다. 최근에는 AI 피드백 기반 강화학습(RLAIF)이 등장해, AI가 다른 AI의 출력을 평가·수정함으로써 효율성을 높이고 있습니다. 이 밖에도 규칙 기반 안전 제어, 가치 모델링, 윤리 데이터셋 학습 등 다양한 기법이 개발되어, AI의 의사결정이 인간 중심의 목표로 수렴하도록 조정합니다.