089 AI 안전

AI Safety

AI의 예측 불가능한 위험으로부터 인간과 사회를 보호하도록 설계된 체계

- AI의 오작동·악용·편향 등 잠재적 위험을 사전에 식별·통제해 안정성을 확보하는 관리 체계
- AI 신뢰성과 윤리의 기반이 되는 기술적 전제이자 사회적 안전장치

AI 안전의 개념

AI 안전은 인공지능이 인간의 의도와 일관되게 작동하면서 사회적 위해를 최소화하도록 설계·관리하는 체계를 말합니다. 초기에는 기술적 오작동이나 오류 방지를 의미했지만, 현재는 AI의 자율성과 범용성이 커지면서 예측 불가능한 판단·악용·편향 등 사회적 위험까지 포괄하게 되었습니다. 이는 설계–개발–배포–운영–폐기의 AI 생명주기 전반에 걸친 위험 식별, 평가, 대응을 포함합니다. 결과적으로 AI 안전은 AI 신뢰성의 하위 요소이자, AI 윤리를 현실화하는 실천적 기반으로 작동합니다.

AI 안전의 구성 요소

AI 안전의 핵심 구성 요소는 예측 가능성, 견고성, 인간 통제 가능성, 검증 가능성, 책임성 등으로 요약됩니다.

- 예측 가능성: AI가 의도된 목적 내에서 일관되게 작동하도록 보장하는 능력
- 견고성: 데이터 오류·적대적 공격·환경 변화에도 안정적 성능을 유지
- 인간 통제 가능성: 시스템이 자율적으로 판단하더라도 인간이 개입·중단할 수 있는 가능성
- 검증 가능성: 모델의 의사결정 과정과 결과를 외부에서 평가·검증할 수 있도록 기록·투명화
- 책임성: 사고 시 명확한 책임 주체와 대응 절차를 확보

AI 안전에 관한 국제적 정책 동향

AI 안전은 각국의 AI 정책에서 중요한 요소로 부상하고 있습니다. 영국에서 개최된 AI Safety Summit을 계기로 국제 AI 안전연구소(AISI) 네트워크가 출범했으며, 참석국들은 국제 AI 안전보고서를 발간하기로 약속했습니다. AISI 네트워크에는 영국, 미국, 캐나다, 일본, 싱가포르, 한국 등 주요국이 참여하고 있으며, 초거대 모델의 위험 평가, 공동 테스트 데이터셋 구축, 검증 기준의 국제 정합성 확보 등 여러 다양한 AI 안전 이슈를 다룹니다. 이외에도 EU는 AI Act를 통해 위험 기반 접근을 제도화하고 고위험 시스템에 대한 사전 평가와 인증을 요구하고 있습니다. 우리나라 역시 AI 신뢰성 검증 가이드라인과 「AI 기본법」 추진과 함께 AISI 참여를 통해 국제 협력 기반을 강화하고 있습니다.