참조 용어 | AI 안전 p.102, AI 윤리 p.107

088 AI 신뢰성

AI Trustworthiness

안전하고 공정하며 투명하게 작동하는 신뢰할 수 있는 AI 체계

- AI가 사회적 가치와 법적 기준을 충족하면서 예측 가능하게 작동하도록 설계된 시스템을 의미
- 인간 중심의 윤리 원칙을 바탕으로 안전성·공정성·설명 가능성을 확보한 AI를 지칭

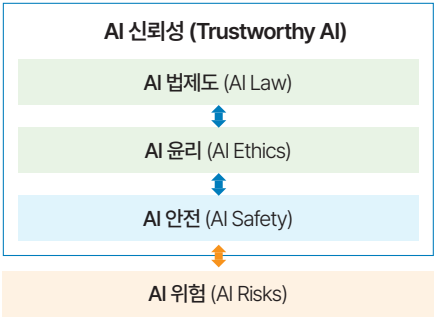
AI 신뢰성의 개념

AI 신뢰성은 인공지능이 사회와 개인에게 안전하고 책임 있게 작동하는지를 판단하는 핵심 기준입니다. 단순히 성능이 뛰어난 AI가 아니라, 사용자가 AI의 결과를 이해하고 신뢰할 수 있도록 안전성과 투명성을 확보한 상태를 의미합니다. 이는 AI 안전과 AI 윤리를 포괄하는 상위 개념으로, AI가 정확하게 작동하고 오류나 편향을 최소화한 '안전성'을 갖추는 것은 신뢰의 기술적 전제이며, 인권과 공정성, 투명성 같은 '윤리 원칙'은 그 신뢰를 사회적 차원으로 확장시키는 기반이 됩니다.

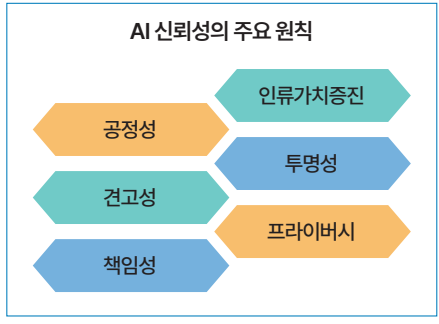
AI 신뢰성의 주요 원칙

AI 신뢰성의 개념 정의는 기관, 학자에 따라 조금씩 다르지만, 대체로 공정성, 견고성, 책임성, 인류가치증진, 투명성, 프라이버시 보호 등의 원칙으로 구성되며 국제표준 및 국가별 평가·인증의 기초가 되고 있습니다.

- 공정성: 알고리즘 편향을 방지하고 특정 집단의 불이익을 차단
- 견고성: 외부 공격·데이터 변동에도 안정적으로 작동
- 책임성: AI 결과에 대해 명확한 책임 주체를 설정
- 인류 가치 증진: 기술 발전이 인간의 존엄성·공공선과 조화되도록 함
- 투명성: 의사결정 과정을 이해 가능한 형태로 공개
- 프라이버시 보호: AI 전 과정에서 개인정보 오용을 최소화하고 데이터 주권을 보장



출처 : AI 안전의 개념과 범위 (SPRi)



출처 : AI 신뢰성 및 윤리 제도 연구 (SPRi)