참조 용어 | AI 신뢰성 p.101

044 설명가능한 AI / XAI

Explainable AI

AI의 판단 근거를 사람이 이해할 수 있도록 설명하는 기술

- AI의 의사결정 과정을 투명하게 드러내어 사람이 이해·검증할 수 있게 하는 기술
- 복잡한 블랙박스형 모델의 한계를 보완해 신뢰성과 책임성을 높이는 것이 목표

XAI의 개념

설명가능한 AI(XAI)는 인공지능이 내린 결과의 근거와 과정을 사람이 이해할 수 있는 형태로 제시하는 기술입니다. 딥러닝 모델은 수많은 매개변수와 연산 과정을 거쳐 결과를 도출하지만, 내부 작동 원리를 직접 해석하기 어렵습니다. 이러한 '블랙박스 문제'를 해결하기 위해 XAI는 AI의 판단 근거를 시각화하거나 규칙화하여 투명하게 제시합니다. 즉, 단순히 결과를 제시하는 수준을 넘어, 그 결론에 이르는 논리적 과정을 드러내어 사용자와 개발자가 신뢰할 수 있는 판단 구조를 제공합니다.

XAI의 접근 방식

XAI의 접근 방식은 크게 두 가지입니다. 하나는 내재적 설명성으로, 애초에 구조가 단순하고 논리적인 모델을 설계해 결과를 사람이 직접 이해할 수 있게 만드는 방식입니다. 예를 들어 의사결정나무나 선형 회귀 모델처럼 계산 과정이 명확한 경우가 이에 해당합니다. 다른 하나는 사후적 설명으로, 복잡한 신경망 모델이 이미 도출한 결과를 시각화하거나 규칙으로 해석하는 방법입니다. 이러한 기법은 모델의 복잡한 내부 연산을 사람이 이해할 수 있는 형태로 번역하여, AI의 결정 과정을 간접적으로 해석하게 합니다.

XAI의 필요성

AI가 의료 진단, 금융 심사, 채용 평가 등 사회 전반의 의사결정에 활용되면서 판단의 투명성과 설명 가능성은 필수 요건이 되었습니다. 설명가능한 AI는 결과의 근거를 명확히 제시함으로써 사용자 신뢰를 높이고, 오류나 편향이 발생했을 때 원인을 추적할 수 있도록 합니다. 또한 법적·윤리적 책임을 명확히 하여, 신뢰할 수 있는 AI(Trustworthy AI) 구축의 기반이 됩니다. 특히 공공 정책과 산업 규제 분야에서는 알고리즘의 공정성과 책임성을 확보하는 수단으로 중요하게 작용하며, 사회적 수용성을 높이는 역할을 하고 있습니다.

XAI의 과제

XAI는 투명성과 신뢰성을 강화하지만 완전한 해법은 아닙니다. 설명력을 높이기 위해 모델을 단순화하면 성능이 저하될 수 있고, 복잡한 모델은 여전히 해석이 어렵습니다. 또한 사용자가 이해하기 쉬운 설명이 반드시 사실적이거나 정확한 설명을 의미하지는 않습니다. 설명력과 성능 간의 균형, 그리고 다양한 문화·언어·데이터 환경을 고려한 설명 기준 마련이 앞으로의 핵심 과제입니다. 그럼에도 XAI는 AI 윤리와 책임성을 실현하는 핵심 기술로, 인간 중심의 신뢰 가능한 AI 발전 방향을 제시하는 기반으로 평가됩니다.