스터디 12.17.(화) 21:30

12.17. 금융권 DS 현직자 강의

→ 서류(성장과정, 직무역량) : 어려움, 목표, 노력, 성과

AI 윤리와 설명가능성에 대해서

개요

현대 사회에서 AI의 급속한 발전은 우리의 삶을 근본적으로 변화시키고 있습니다.

특히 딥러닝을 비롯한 고도화된 AI 기술이 의료 진단, 금융 의사결정, 자율주행 등 중요한 판단이 요구되는 영역에 도입되면서, AI 시스템의 윤리적 설계와 의사결정 과정의 투명성 확보는 그 어느 때보다 중요한 과제로 대두되고 있습니다.

AI 윤리와 설명가능성에 대한 논의가 활발해진 배경에는 여러 사회적 사건들이 있습니다.

예를 들어, 2016년 ProPublica의 조사에서 드러난 형사사법 시스템의 AI 편향성 문제는 큰 파장을 일으켰습니다. 해당 시스템은 흑인 범죄자의 재범 위험도를 백인에 비해 체계적으로 높게 평가했으며, 이는 AI 시스템의 편향성과 공정성 문제를 사회적 담론의 중심으로 끌어올렸습니다.

또한 2018년 발생한 우버의 자율주행차 사고는 AI 시스템의 의사결정 과정과 책임소재의 투명성 문제를 부각시켰습니다.

이러한 사회적 맥락 속에서, AI 윤리와 설명가능성이라는 두 가지 핵심 주제를 깊이 있게 다루고자 합니다. 특히 AI 시스템이 인간의 기본권과 존엄성을 존중하면서도 혁신적 가치를 창출할 수 있는 균형점을 찾는 것에 주목합니다.

AI 윤리의 핵심 가치와 원칙

AI 윤리는 단순한 도덕적 지침이나 규범적 원칙의 집합을 넘어, AI 기술의 개발과 적용 과정 전반에 걸쳐 고려되어야 할 포괄적인 가치 체계입니다. 이는 기술 발전이 인간의 기본권과

사회적 가치를 침해하지 않으면서도 혁신적 발전을 이룰 수 있도록 하는 균형점을 제시합니다.

특히 최근 들어 AI 기술이 점차 고도화되고 그 영향력이 확대되면서, AI 윤리의 중요성은 더욱 부각되고 있습니다.

인간 중심의 AI 개발

AI 기술은 궁극적으로 인간의 삶의 질을 향상시키고 사회 발전에 기여하는 도구로서 개발되어야 한다는 것이 AI 윤리의 가장 근본적인 원칙입니다. 이는 단순한 선언적 가치를 넘어, 구체적인 기술 개발과 적용 과정에서 실천되어야 할 지침들을 포함합니다.

먼저, 인간의 자율성과 존엄성 존중이라는 측면에서 AI 시스템은 인간의 자기결정권을 침해해서는 안 됩니다. 예를 들어, 의료 AI 시스템이 진단과 치료 방침을 제시할 때, 이는 의료진과 환자의 최종 판단을 보조하는 역할에 국한되어야 하며, 그들의 의사결정 권한을 대체해서는 안 됩니다. 최근 개발되고 있는 의료 AI 시스템들은 이러한 원칙을 반영하여, 진단 결과와 함께 그 근거가 되는 데이터와 분석 과정을 함께 제시하는 방식을 채택하고 있습니다.

프라이버시와 개인정보 보호는 또 다른 중요한 축을 형성합니다. AI 시스템이 학습하고 처리하는 데이터에는 개인의 민감한 정보가 포함될 수 있으며, 이는 엄격한 보호와 관리가 필요합니다. 특히 의료 데이터나 금융 정보와 같은 민감 정보를 다루는 AI 시스템의 경우, 데이터 수집부터 저장, 처리, 폐기에 이르는 전 과정에서 강력한 보안 조치와 함께 정보주체의 권리를 보장하는 체계가 구축되어야 합니다.

사회적 공공선의 추구 역시 간과할 수 없는 가치입니다. AI 기술은 단순히 경제적 효율성이나 기술적 성능만을 목표로 해서는 안 되며, 사회 전체의 복지 향상에 기여해야 합니다. 예를 들어, AI 기술을 활용한 교육 시스템은 교육 격차를 해소하고 개인별 맞춤형 학습을 지원함으로써 교육의 형평성과 질적 향상에 기여할 수 있습니다.

투명성과 설명가능성

AI 시스템의 투명성과 설명가능성은 AI에 대한 사회적 신뢰를 구축하는 핵심 요소입니다. 특히 AI가 중요한 의사결정에 관여하는 경우, 그 결정 과정과 근거를 이해관계자들이 이해할

수 있는 방식으로 설명할 수 있어야 합니다. 이는 단순히 기술적인 투명성을 넘어, AI 시스템이 사회적으로 수용되고 책임 있게 운영되기 위한 필수 조건입니다.

알고리즘의 작동 원리 공개는 매우 신중하게 접근해야 할 문제입니다. 기업의 영업 비밀과 지적 재산권을 보호하면서도, 필요한 수준의 투명성을 확보하는 것이 중요합니다. 예를 들어, AI 시스템이 사용하는 주요 변수들과 의사결정의 기본 원칙은 공개하되, 구체적인 알고리즘 구현 방식은 보호하는 방식을 채택할 수 있습니다.

의사결정 과정의 추적 가능성은 특히 금융, 의료, 법률과 같은 중요한 의사결정 분야에서 핵심적입니다. AI 시스템이 특정한 결정을 내린 과정을 사후에 검토하고 검증할 수 있어야 합니다. 이는 문제가 발생했을 때 원인을 파악하고 해결하는 것은 물론, 시스템의 지속적인 개선을 위해서도 필수적입니다.

설명가능한 AI(XAI)의 기술적 구현

설명가능한 AI(eXplainable AI, XAI)는 AI 시스템의 의사결정 과정과 결과를 인간이 이해할 수 있는 방식으로 설명하는 기술을 의미합니다. 이는 AI 시스템의 블랙박스 문제를 해결하고 신뢰성을 확보하기 위한 핵심 요소로 자리잡고 있습니다. 특히 최근에는 딥러닝 모델의 복잡성이 증가하면서 설명가능성의 중요성이 더욱 부각되고 있습니다.

주요 기술적 접근 방법

설명가능한 AI를 구현하기 위한 기술적 접근은 크게 두 가지 방향으로 나눌 수 있습니다. 첫째는 모델 자체를 해석 가능하게 설계하는 방식이고, 둘째는 복잡한 모델의 판단 과정을 사후에 설명하는 방식입니다.

LIME(Local Interpretable Model-agnostic Explanations)은 후자의 대표적인 예시입니다. LIME은 복잡한 모델의 특정 예측 결과에 대해, 그 주변에서 단순한 해석 가능한 모델을 만들어 설명을 제공합니다. 예를 들어, 의료 영상 진단에서 AI가 특정 부위를 암으로 진단했다면, LIME은 그 판단의 근거가 된 영상의 특정 영역을 하이라이트하여 보여줄 수 있습니다.

SHAP(SHapley Additive exPlanations)은 게임 이론의 샤플리 값을 기반으로 각 특성의 예측 기여도를 계산합니다. 이는 특히 금융 분야에서 널리 활용되고 있습니다. 예를 들어, 대

출 심사 과정에서 AI가 신용 평가를 할 때, SHAP 값을 통해 각각의 요소(소득, 직업, 거래 이력 등)가 최종 결정에 얼마나 영향을 미쳤는지 정량적으로 설명할 수 있습니다.

프로토타입 네트워크(Prototype Networks)는 좀 더 직관적인 접근 방식을 취합니다. 이는 데이터의 대표적인 예시들을 기반으로 판단을 내리는 방식으로, 인간의 사례 기반 추론 과정과 유사합니다. 의료 진단 시스템에서 특정 질병을 진단할 때, 유사한 과거 사례들을 함께 제시함으로써 의료진이 판단의 근거를 직관적으로 이해할 수 있게 합니다.

SENN(Self-Explaining Neural Networks)은 신경망 자체를 설명 가능한 형태로 설계하는 접근 방식입니다. 이는 모델의 구조 자체에 해석 가능성을 내장함으로써, 별도의 설명 기법 없이도 판단 과정을 이해할 수 있게 합니다. SENN은 특히 고위험 분야에서 활용되는 AI 시스템에 적합한데, 실시간으로 설명이 필요한 자율주행차량이나 의료기기 등에 적용될 수 있습니다.

산업별 구체적 적용 사례

의료 분야에서는 설명가능한 AI의 적용이 특히 중요합니다. 진단과 치료 방침 결정에 AI를 활용할 때, 의료진과 환자 모두가 그 판단의 근거를 명확히 이해할 필요가 있기 때문입니다. 예를 들어, IBM의 Watson for Oncology는 암 치료 방침을 추천할 때, 관련 의학 문헌과 임상 데이터를 함께 제시하여 의료진이 추천의 근거를 검토할 수 있게 합니다.