Special Topic **05**

장동익 팀장
한국교통연구원
모빌리티데이터융합연구팀

AI 신뢰성 확보를 위한 필수 조건, 설명 가능성(XAI)

XAI, 인간과 AI 간 신뢰 보정을 수행하는 핵심 상호작용 메커니즘 될 수 있어

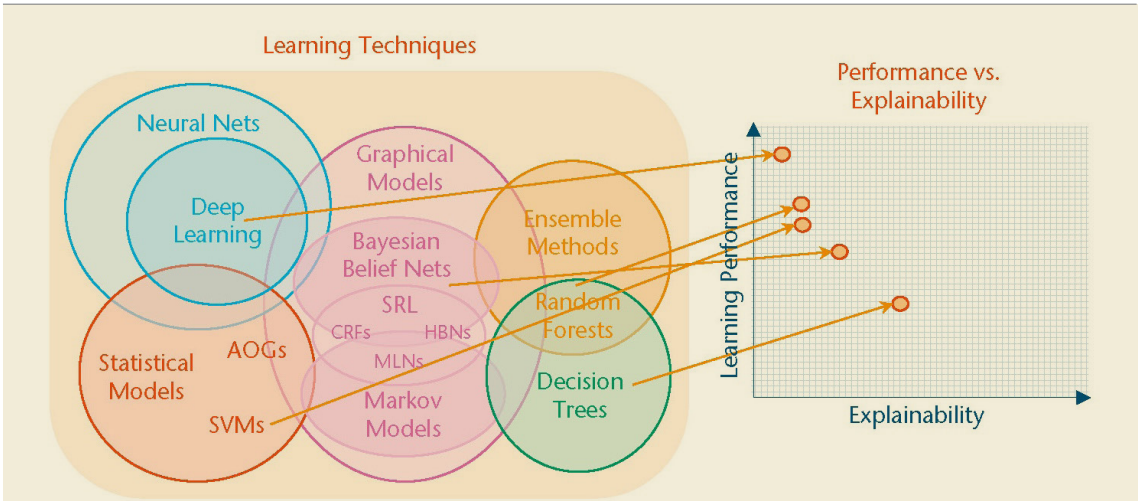
설명 가능한 AI

AI(Artificial Intelligence, 인공지능)의 발전 역사는 데이터에서 규칙을 찾아내고 이를 사람이 이해할 수 있도록 하는 여정이었다. 1970~80년대 AI의 초기 단계에서는 미리 정의된 규칙에 따라 작동하는 규칙 기반(rule-based) 전문가 시스템(Expert System)이 주요 접근법이었기 때문에, 특정 결과가 도출된 이유와 근거를 인간이 명확하게 이해할 수 있었다. 선형 회귀(Linear Regression)나 의사결정나무(Decision Tree) 모델은 구조 자체가 투명해 예측 과정을 사람이 직관적으로 파악할 수 있는 대표적인 예였다.

그러나 1990년대 이후 서포트 벡터 머신(SVM), 랜덤 포레스트와 그레이디언트 부스팅 같은 앙상블 트리(Ensemble Tree) 등 새로운 머신러닝 기법이 등장했고, 2010년대에는 딥러닝(Deep Learning)과 같은 복잡한 모델들이 압도적인 성능을 보이며 AI 기술의 패러다임을 근본적으로 변화시켰다. 이 과정에서 모델의 구조는 기하급수적으로 복잡해졌고, 내부 의사결정 과정을 직관적으로 이해하기 어려운 '블랙박스' 문제가 더욱 심각해졌다.

2016년 알파고(AlphaGo)의 등장은 이러한 문제에 대한 대중적 인식을 크게 확산시켰다. 알파고는 당대 최고 수준의 인간 기사를 압도적으로 능가하는 기량을 선보였지만, 프로 기사들조차 알파고가 특정 수를 둔 이유를 명확히 설명하지 못했다. 이는 "AI는 탁월하게 잘 두지만, 그 이유는 알 수 없다"는 인식을 넓게 퍼지게 만들었고, AI가 인간의 이해와 신뢰를 얻기 위해서는 '왜'라는 근본적인 질문에 답할 수 있어야 한다는 사회적 요구를 촉발했다.

이러한 요구에 대응하여 2016년 미국 국방고등연구계획국(DARPA)은 설명 가능한 인공지능(eXplain



자료: Gunning and Aha(2019).

그림1 ML·AI 모형들의 예측력(성능) vs. 설명 가능성

able Artificial Intelligence, XAI) 프로그램을 출범시켰다. 이 프로그램의 목표는 군사 및 안보 환경에서 AI가 높은 성능을 유지하면서도 인간이 이해할 수 있는 설명을 제공하여 효과적으로 협력할 수 있도록 하는 것이었다. DARPA는 이를 위해 설명 가능성을 내재화한 모델 설계, 사용자 친화적 인터페이스 개발, 실제 임무 환경에서의 설명 효과 검증이라는 세 가지 접근 방향을 제시했다.

또한 이러한 흐름은 AI의 높은 예측력이 반드시 모델의 근본적인 타당성을 보장하지 않는다는 '클레버 한스(Clever Hans[1]) 현상'에 대한 경각심을 불러일으켰다. AI에서도 이와 유사하게, 모델이 실제 문제 해결에 필요한 본질적 특징이 아니라 훈련 데이터에 내재된 무관한 패턴이나 편향(bias)에 의존해 높은 예측력을 보이는 경우가 있을 수 있기 때문이다. 예를 들어, 폐암 진단 AI가 폐암의 실제 병리적 특징이 아니라 X-ray 이미지의 특정 주석(annotation)에 기반해 진단을 내린다면, 이는 '블랙박스' 내부의 잘못된 추론을 의미한다. 이러한 오류와 잠재적 편향은 XAI를 통해 조기에 발견하고 수정할 수 있으며, 따라서 XAI는 단순한 예측 결과 해석을 넘어 모델의 숨겨진 결함을 진단하고 보완하는 전략적 디버깅 도구로서 중요한 가치를 지닌 것으로 볼 수 있다.

1) 20세기 초 독일의 말. 클레버 한스는 복잡한 산수 문제를 풀 수 있는 것처럼 보였으나, 실제로는 조련사와 관객의 미세하고 무의식적인 신호에 반응했던 것에 불과했다.

XAI 기술 프레임워크

최근 AI가 다루는 데이터는 대규모화, 고차원화되고 있으며, 모델 구조는 점점 더 복잡해지고 있다. 이에 따라 딥러닝과 강화학습에 특화된 XAI 기법들이 계속해서 등장하고 있다. 강화학습 XAI(XRL, eXplainable Reinforcement Learning)에서는 순차적 의사결정 과정에서 AI 에이전트가 왜 특정 상태에서 특정 행동을 선택했는지, 그리고 어떤 보상 요인이 그 결정에 영향을 미쳤는지를 심층적으로 분석하는 것이 핵심이다. 이를 위해 정책 시각화, 상태 기여도 분석, Q-value 분해 등의 첨단 기법들이 활용되며, 특히 신호 제어 AI와 자율주행 전략 해석에 뛰어난 효과를 보여준다. XRL은 단순히 최종 결과를 해석하는 것을 넘어, 에이전트가 특정 상황에서 특정 행동을 선택한 심층적 메커니즘과 그 결정을 유도한 보상 요인을 정밀하게 분석할 수 있게 하므로 자율주행, 교통 신호 제어 등 복잡한 순차적 의사결정 문제에서 AI 에이전트의 행동 원리를 밝혀내는 데 필수적인 요소가 되고 있다.

현재 주로 활용되는 XAI 기술은 크게 두 가지 접근법으로 나눌 수 있다. 첫째는 모델 자체의 구조가 투명하고 이해하기 쉬운 내재적 해석(White-box) 방식이고, 둘째는 복잡한 모델의 예측 결과를 사후에 설명하는 사후 설명 방식이다. 내재적 해석 모델은 선형 회귀나 의사결정나무처럼 모델의 작동 원리가 명확하여 그 자체로 해석이 가능한 모델을 의미한다. 이러한 모델은 규제 준수와 디버깅에 유리하지만, 실제 복잡하고 비선형적인 문제에서는 성능 면에서 한계가 있다. 사후 설명 방식은 고성능 블랙박스 모델을 그대로 유지하면서, 예측 결과가 나온 후 그 이유를 해석하는 기법이다. 이 방식은 모델의 구조와 무관하게 적용 가능한 모델 불가지론적(Model-agnostic) 기법과 특정 모델 구조를 전제로 하는 모델 특정적(Model-specific) 기법으로 나뉜다.

> “
> **강화학습 XAI**에서는
> 순차적 의사결정 과정에서
> AI 에이전트가 왜 특정 상태에서
> 특정 행동을 선택했는지,
> 그리고 어떤 보상 요인이
> 그 결정에 영향을 미쳤는지를
> **심층적으로 분석**하는 것이 **핵심**이다.
> ”

모델 불가지론적 기법

LIME(Local Interpretable Model-agnostic Explanations)

특정 예측 주변의 국소 영역에서 복잡한 블랙박스

모델을 간단한 선형 회귀나 의사결정나무와 같은 해석 가능한 모델로 근사하여, 예측에 영향을 미친 핵심 특징을 도출한다. 예를 들어, 이미지의 일부 영역을 흐리게(perturbation) 처리한 후 예측 변화를 관찰해 중요한 슈퍼픽셀(super-pixel)을 찾아내는 방식이다.

SHAP(SHapley Additive exPlanations)

협력 게임 이론의 Shapley 값을 활용해 각 특징의 기여도를 공정하고 일관되게 분배하는 기법이다. 모든 특징 조합을 고려해 높은 정확도로 기여도를 계산할 수 있지만, 계산 비용이 크다는 단점이 있어 실제로는 TreeSHAP 등 근사 알고리즘이 주로 사용된다.

모델 특정적 기법

Grad-CAM

합성곱 신경망(CNN)의 그래디언트를 활용해 예측에 기여한 이미지 영역을 히트맵으로 시각화하고, Score-CAM은 그래디언트 없이 활성값 기반의 안정적인 시각화를 제공한다.

LRP(Layer-wise Relevance Propagation)

출력값을 입력 특징에 단계적으로 분배하여 기여도를 계산한다.

Integrated Gradients

기준 입력에서 실제 입력까지의 경로를 따라 기울기를 적분해 특징 기여도를 정량화한다.

원본 이미지 | 고양이 식별 Grad-CAM | 강아지 식별 Grad-CAM

자료: Selvaraju et al.(2020).

그림1 Grad-CAM을 이용하여 고양이와 강아지를 식별 모형에서 예측에 가장 크게 기여하는 부분을 보여주는 예시

ProtoPNet

설명형(self-explainable) 신경망으로 학습 과정에서 확보한 대표 사례(prototype)를 예측 근거와 함께 제시함으로써, 전문가가 AI의 결정을 직관적으로 이해하고 신뢰를 형성하도록 돕는다.

사후 설명 방식은 설명 모델이 원본 모델의 근사치이므로 실제 작동과 설명 내용 간 불일치 문제(정합성)가 발생할 수 있다. 이를 극복하고 성능과 해석 가능성을 동시에 확보하기 위해 블랙박스와 화이트박스 모델을 상황에 따라 결합하는 하이브리드(Hybrid) 접근법이 제안되고 있다. 예를 들어, 자율주행 분야에서는 LIME의 연산 효율성과 SHAP의 설명 정밀도를 통합한 프레임워크를 통해 실시간 해석과 설명 품질의 균형을 맞추려는 시도가 진행되고 있다(Tahir, 2024).

최근 대규모 언어모델(LLM)을 포함한 생성형 AI의 급속한 발전은 XAI의 활용 방식과 범위를 근본적으로 변화시키고 있다. 전통적인 XAI 기법이 생성하는 수치, 그래프, 히트맵 등의 기술적 결과물은 여전히 비전문가에게는 이해하기 어려울 수 있다. 그러나 LLM은 이러한 XAI의 복잡한 결과물을 사용자 친화적인 자연어 보고서나 스토리텔링 형식으로 설명해주며, 단순한 정보 번역을 넘어, 사용자의 배경지식과 상황 맥락에 완벽하게 맞춘 최적화된 설명을 제공하고 있다. 프롬프트 엔지니어링은 이러한 맞춤형 설명을 구현하는 핵심 기술로, XAI가 산출한 데이터를 기반으로 초심자용 쉬운 설명부터 전문가용 상세 보고서까지 다양한 수준의 설명을 정교하게 제어할 수 있게 해준다. 더 나아가 생성형 AI는 정적인 보고서 제공을 넘어, 실시간 상호작용을 통해 설명을 지속적으로 보완하고 갱신하는 '설명 대화' 방식을 구현할 수 있다. 사용자의 질문에 XAI 분석 결과를 바탕으로 즉각적이고 상세한 대화형 응답을 제공함으로써, XAI를 단순한 해석 도구에서 인간과 AI가 능동적으로 협력하는 지능형 인터페이스로 혁신적으로 발전시키고 있다. 그러나 이러한 잠재력을 온전히 실현하기 위해서는 LLM이 생성하는 설명의 정합성, 완전성, 일관성 등 다차원적 품질을 어떻게 평가하고 보장할 것인지에 대한 새로운 기술적, 윤리적 도전 과제들이 여전히 남아 있다.

XAI 거버넌스

AI 신뢰성에 관한 논의는 이제 국제적인 규범과 표준을 통해 구체적인 모습을 갖추고 있다. 이는 XAI가 단순한 선택적 기술 요소가 아니라, 이제는 법적 의무이자 제도적 준수 사항으로 자리 잡았음을 명확히 보여주고 있다.

GDPR(General Data Protection Regulation)

유럽연합의 GDPR은 자동화된 의사결정과 관련하여, 당사자가 해당 결정의 논리와 주요 요인을 인간이 이해할 수 있는 방식으로 설명받을 권리(Right to Explanation)를 공식적으로 명시했다. 이를 통해 금융 분야에서는 대출 거절 이유를, 채용 과정에서는 탈락 근거를 요구할 수 있는 법적 기

> “
> **AI 신뢰성**에 관한 논의는
> **국제적인 규범과 표준**을 통해
> 구체적인 모습을 갖추고 있다.
> 이는 **XAI**가 단순한 선택적 기술 요소가
> 아니라, 이제는 **법적 의무**이자
> **제도적 준수 사항**으로 자리 잡았음을
> 명확히 보여주고 있는 것이다.
> ”

반이 마련되고 있다. 다만, 이 권리가 본문 조항(Article 22)이 아닌 부속 설명문인 Recital 71에 언급되어 있어, 법적 구속력의 범위와 해석을 둘러싼 논쟁이 계속되고 있으며, 이에 따라 EU AI Act와 같은 후속 법안에서 더욱 구체적이고 강제력 있는 규정이 보완되고 있다.

EU AI Act

최초로 위험 기반 접근을 도입한 AI 전용 법안으로, AI 시스템을 위험 수준별로 분류하여 규제한다. 허용할 수 없는 위험을 초래하는 AI 시스템의 사용은 전면 금지하며, 의료, 교통, 금융 등 생명과 안전에 직접적인 영향을 미치는 고위험 AI 시스템(High-Risk AI Systems, HRAIS)에 대해서는 엄격한 투명성과 설명 가능성 의무를 부과한다. 이는 추상적 권리를 모든 분야에 일괄적으로 적용하는 대신, 고위험 영역에 집중해 실질적인 책임과 규제를 강화하려는 정책적 방향을 명확히 보여주고 있다.

OECD AI 권고안

포용적 성장, 인간 중심 가치, 투명성과 설명 가능성, 견고성과 안전, 책임성 등 다섯 가지 가치 기반 원칙과 함께, 신뢰할 수 있는 AI R&D 투자, 디지털 생태계 조성, 위험 기반 규제 환경 구축, 인력 역량 강화, 국제 협력 등 다섯 가지 정책 권고를 제시했다. 이는 EU AI Act, GDPR과 같은 구체적 법안의 철학적·정책적 기반이 되었으며, XAI를 단순한 기술적 과제를 넘어 사회적 가치와 공정성을 구현하는 핵심 도구로 자리 잡게 하는 데 크게 기여하였다.

규범이 AI의 책임 있는 개발과 활용에 대한 당위성과 원칙을 제시한다면, 국제 표준은 이를 실무적으로 구현하기 위한 구체적인 방법론을 제공한다. ISO/IEC JTC 1/SC 42는 AI 신뢰성을 체계적으로 관리하기 위한 표준을 제정하는 핵심 기구로, 다음과 같은 주요 표준을 제시했다.

ISO/IEC 42001(AI 관리 시스템, AIMS)

AI 관리 시스템(AIMS)의 요구사항을 규정하여, 윤리적이고 안전하며 투명한 AI 개발과 운영을 위한 조직 차원의 거버넌스 체계를 제시한다.

ISO/IEC 23894(리스크 관리)

AI 리스크 관리 표준으로, 위험의 식별, 분석, 평가, 대응, 모니터링에 관한 구조화된 절차를 제공한다.

ISO/IEC TR 24027(편향 완화)

권고적 기술보고서로 인간 인지 편향, 데이터 편향, 공학적 편향 등 AI 시스템의 다양한 편향 유형을 식별하고 완화하는 프레임워크를 제시한다.

ISO/IEC 25059(품질 평가)

AI 시스템의 학습성, 불완전 데이터 처리, 확률적 추론, 설명 가능성, 공정성 등 고유한 특성을 반영한 품질 모델을 제공한다.

또한, 미국 국립표준기술연구소(NIST)의 AI RMF (AI Risk Management Framework)는 AI 전 생애주기에 걸친 리스크 관리를 위해 Govern, Map, Measure, Manage라는 네 가지 핵심 기능을 제시하며, 조직이 실무적으로 적용 가능한 가이드라인을 제공하고 있다. 이처럼 국제 규범과 표준은 '설명 가능성'이라는 추상적 가치를 실현 가능한 기술적, 관리적 지침으로 구체화하고, 이를 통해 AI 관리 원칙과 절차를 완성하는 핵심 토대를 형성한다.

교통 분야 XAI 도입 전략

교통 분야는 AI 기술의 도입이 빠르게 확산되고 있지만, 동시에 시민의 생명과 직결된 고위험 의사결정 영역이기 때문에 XAI 도입의 중요성이 더욱 부각되고 있다. 교통 수요 예측, 신호 제어, 자율주행과 같은 분야는 투명성과 안전성, 책임성이 담보되지 않는다면 사회적 수용성을 이끌어내기 어렵다. 특히 자율주행에서 피할 수 없는 사고와 같은 윤리적 판단이 필요한 상황에서는 AI의 의사결정 근거와 과정에 대한 명확한 설명이 반드시 요구되고 있다.

최근 XAI 도입의 핵심 패러다임은 인간 중심 AI(Human-Centered AI, HCAI)로 전환되고 있다. HCAI는 AI가 인간의 역할을 대체하는 것이 아니라, 인간의 능력과 가치를 존중하고 강화하는 방향으로 설계되어야 한다는 접근법이다. 기존의 AI가 효율성과 자동화에 집중했다면, HCAI는 공정성과 투명성, 접근성을 보장하면서 사용자 경험을 중심에 둔다. 이를 위해 개발 초기부터 심리학자, 윤리학자, 도메인 전문가 등 다양한 분야의 협업을 포함하고, 최종 사용자의 참여를 통한 공동 설계 원칙을 강조하고 있다.

HCAI 원칙을 교통 분야 XAI에 적용하기 위해서는 사용자 맞춤형 설명 전략이 필수적이다. 인간-컴퓨터 상호작용 연구에 따르면, 과도한 설명 정보는 사용자의 인지 부하를 증가시켜 오히려 이해를 방해할 수 있으므로, 설명은 정보량과 표현 방식을 사용자의 역할과 상황에 최적화되어야 한다. 예를 들어, 자율주행 시스템에 XAI를 적용하기 위해서는 각 부문별로 다음과 같은 전략을 반영해야 할 것이다.

> 최근 **XAI 도입**의 핵심 패러다임은
> **인간 중심 AI**(Human-Centered AI, HCAI)로
> 전환되고 있다. HCAI는 AI가
> 인간의 역할을 대체하는 것이 아니라,
> **인간의 능력과 가치**를 **존중**하고 **강화**하는
> 방향으로 설계되어야 한다는 접근법이다.

운전자·탑승자

운전자·탑승자에게는 자율주행 시스템의 결정 근거가 즉각적이고 직관적으로 전달되어야 한다. 복잡한 텍스트 대신 색상 강조, 아이콘, 음성 경고 등 시각·청각 요소를 활용해 인지 부하를 줄이고 신속한 판단을 지원하여야 한다.

관제사·운영자

관제사·운영자에게는 교통 관제 시스템의 의사결정 이유를 대화형 대시보드로 시각화하여 제공되어야 한다. 이를 통해 관제사는 AI의 의사결정 요인을 명확히 이해하고, 필요한 경우 제안을 검증하고 수정할 수 있어야 한다.

정책 결정자·개발자

정책 결정자·개발자에게는 장기 전략 수립을 위해 변수별 영향도 분석, 모델 편향 및 위험 분석 결과, 재현 가능한 설명 보고서 등을 제공하여야 한다. 이를 통해 시나리오별 비교 분석을 통해 정책 수립과 규제 준수를 동시에 지원하여야 한다.

결론 및 정책 제언: XAI 거버넌스 구축을 위한 로드맵

XAI는 AI 시스템의 신뢰성을 확보하는 핵심 기술이지만, 실제 도입 과정에서는 여전히 사회·제도적, 기술적 난제들이 존재한다. 먼저 사회·제도적 측면의 난제들은 다음과 같다.

설명 책임의 경계

AI 생애주기에 관여하는 다양한 이해관계자(모델 개발자, 데이터 제공자, 시스템 운영자, 최종 사용자 등) 중 누가 설명의 품질과 신뢰성에 대한 최종적인 책임을 질 것인지에 대한 명확한 합의가 필요하다.

설명 의무 범위

의료, 교통 등 고위험 분야에만 설명 의무를 제한할 것인지, 아니면 모든 AI 시스템에 동일하게 적용할 것인지에 대한 정책적 결정이 요구된다.

설명 남용(Explanation Abuse) 위험

지나치게 단순화되거나 수사적으로 포장된 설명은 사용자의 판단을 왜곡하고 AI에 대한 맹목적

신뢰를 야기할 수 있다. 예를 들어, 자율주행차가 갑자기 정차한 이유를 '전방의 장애물 때문'이라고만 설명하고, 실제로는 단순한 그림자에 반응했음에도 불구하고 사용자가 그 설명을 맹목적으로 신뢰하는 상황이 발생할 수 있다. 이를 방지하기 위해 설명 품질에 대한 최소 기준과 검증 절차가 마련되어야 한다.

기술적 측면의 난제들은 다음과 같다.

설명 품질 평가 지표의 표준화

정확성, 완전성, 일관성 등을 평가하는 다양한 지표(Feature Attribution Precision: FAP, Feature Attribution Recall: FAR, Feature Attribution F1-score: FAF1 등)가 제안되고 있으나, 국제적으로 통용될 수 있는 통합 체계는 아직 부재하다.

인지 부하 관리

설명은 양보다는 적절성이 중요하며, 사용자의 역할과 상황에 맞춘 맞춤형 설명 가이드라인이 필요하다.

이러한 과제를 해결하기 위해 XAI 거버넌스 로드맵을 다음과 같이 제시할 수 있다.

기반 구축 단계

- AI 시스템의 위험 수준을 평가하고, ISO/IEC 42001 기반 AI 관리 시스템을 도입한다.
- ISO/IEC TR 24027에 따라 데이터 편향을 정기적으로 분석·완화하는 기술적·절차적 기반을 마련한다.

> "XAI는 **인간의 판단을 돕고 상호 보완적인 관계를 구축**하는 데 핵심적인 역할로 자리매김해야 한다. 그러므로 설명 책임의 경계, 설명 남용 방지, 품질 지표 표준화 등은 **공통의 규범**과 **가이드라인**을 **수립**해야 해결할 수 있는 복합적 과제이다."

기술 내재화 단계

- 교통 수요 예측, 신호 제어, 자율주행 등 핵심 AI 시스템에 LIME, SHAP, Grad-CAM과 같은 사후 설명 기법을 통합한다.
- 모델 예측 근거를 체계적으로 기록·관리하는 내부 프로세스를 구축한다.

사용자 확장 단계

- LLM과 프롬프트 엔지니어링을 활용해 이해관계자별 맞춤형 설명 인터페이스를 개발한다.

- 운전자, 관제사, 정책 결정자 등 각 사용자군의 인지 특성을 반영한 HCAI 원칙을 적용해 설명의 효과를 극대화한다.

거버넌스 강화 단계

- 설명의 정확성, 완전성, 일관성을 평가하는 품질 지표를 마련하고 주기적으로 측정한다.
- 의사결정 로그와 설명 보고서를 보관하여 책임성과 재현성을 보장하는 감사 체계를 확립한다.

이 로드맵을 실행하면 XAI는 단순한 의사결정 과정의 해석 도구를 넘어 인간과 AI 간 신뢰 보정을 수행하는 핵심 상호작용 메커니즘이 될 수 있다. 과도하게 단순한 설명은 맹목적 신뢰를, 지나치게 복잡한 설명은 불필요한 인지 부하를 초래하므로, 결국 설명의 양과 질을 사용자의 상황과 역할에 맞춰 최적화하는 것이 성공적인 XAI 거버넌스의 핵심이기 때문이다.

XAI는 단순히 문제를 해결하는 기술을 넘어, 인간의 판단을 돕고 상호 보완적인 관계를 구축하는 데 핵심적인 역할로 자리매김해야 한다. 그러므로 설명 책임의 경계, 설명 남용 방지, 품질 지표 표준화 등은 AI 기술, 법률, 사회과학, HCI 전문가들이 협력하여 공통의 규범과 가이드라인을 수립해야 해결할 수 있는 복합적 과제이다.

AI의 설명 가능성은 단순한 기술적 과제가 아니라, 사회적 신뢰와 법적 책임을 담보하기 위한 핵심 요건으로 자리 잡고 있다. 유럽의 GDPR과 AI Act가 이를 제도적으로 강화해온 것처럼, 우리나라 역시 2026년부터 시행될 인공지능 기본법에서 고영향 인공지능에 대한 투명성과 책임성 확보를 명시하고 있다. 이는 곧 설명 가능한 AI가 선택이 아닌 필수가 될 것임을 보여주며, 앞으로 기술 개발과 정책 수립 모두에서 그 중요성은 더욱 커질 것이다.

참고문헌(References & Further Readings)

1. Gunning, D. & Aha, D. W.(2019). DARPA's Explainable Artificial Intelligence (XAI) Program, AI Magazine, 40, 44-58.
2. DARPA. (2016). Explainable Artificial Intelligence (XAI) program overview. Defense Advanced Research Projects Agency.
3. Gunning, D., Vorm, E., Wang, J., & Turek, M. (2021). DARPA's Explainable AI (XAI) Program: A Retrospective.
4. ISO/IEC. (2019-2024). JTC 1/SC 42 Artificial Intelligence Standards. International Organization for Standardization / International Electrotechnical Commission.
5. OECD. (2019). Recommendation of the Council on Artificial Intelligence.
6. European Parliament and Council of the European Union. (2016). General Data Protection Regulation (GDPR), Regulation (EU) 2016/679. Official Journal of the European Union.
7. European Commission. (2024). Artificial Intelligence Act. Official Journal of the European Union.
8. U.S. National Institute of Standards and Technology. (2023). AI Risk Management Framework (AI RMF 1.0). NIST Special Publication 1270.
9. Samek, W & Müller, K.-R.(2019). Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Towards Explainable Artificial Intelligence
10. Selvaraju, R.R., Cogswell, M., Das, A. et al.Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Int J Comput Vis128, 336-359
11. Tahir, H. A., Alayed, W., Hassan, W. U., & Haider, A. (2024). A Novel Hybrid XAI Solution for Autonomous Vehicles: Real-Time Interpretability Through LIME-SHAP Integration. Sensors, 24(21), 6776.