**Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)**

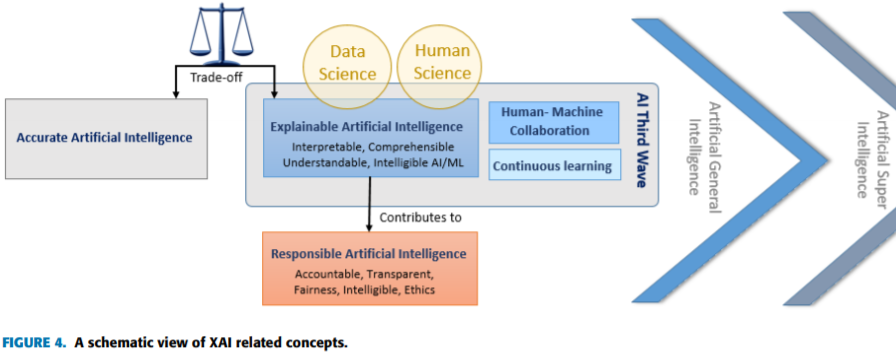
[**https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8466590**](https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8466590)

**[00. Key Terms]**

|  |  |
| --- | --- |
| Interpretable Machine Learning | **입력 데이터가 출력 데이터에 어떻게 수학적으로 매핑**되는지 사람이 연구하고 이해할 수 있는 머신러닝 모델 |
| Black-box problem | **블랙박스:** 내부 디자인, 구조, 구현을 전혀 노출시키지 않는다.  **화이트박스:** 내부 디자인 등을 전적으로 노출시킨다. |
| Responsible AI | AI가 도덕적이고 윤리적인 고려를 하는 것   * **Accountability:** 결정과 행동을 사용자에게 설명할 수 있는 것 * **Responsibility:** 사용자들의 역할, AI 시스템의 결정과 행동에 대한 능력, 오류 또는 예상치 못한 결과에 대응하는 것 |
| Accurate AI | 정확한 예측의 비율 |
| Data Science | 인공지능 모델은 보통 training set과 test set을 필요로 한다. 데이터 사이언스는 **데이터 분석, 머신러닝 등과 관련된 분야**이다. |
| Social Science | 설명은 사회적 상호 작용의 한 가지이고 이것을 위한 일반적인 원칙은 **사회 내부의 각 개인과 관련되어 있다**. |
| 3rd wave AI | **Artificial Intelligence 3.0 (AI 3.0):** AI 시스템이 설명 모델을 구축하고, 새로운 상황에 직면했을 때 그것을 학습하고 추론한다. (continuous learning) 또한 **Interactive AI**는 인간과 자연스러운 소통이 가능하다. |
| Artificial General Intelligence (AGI) | 사람이 할 수 있는 어떤 행동이든 학습하여 수행할 수 있는 인공지능으로 **인공지능 분야의 궁극적 목표**이다. |

**[01. XAI의 필요성]**

|  |  |
| --- | --- |
| 옳음을 증명 | 인공지능/딥러닝 시스템이 편향되거나 차별적인 결과를 도출하는 문제점 등으로 인해 **논란이 발생하고, 따라서 그 판단이 옳음을 증명**해야 한다. |
| 제어를 위한 설명 | 상황이 잘못되는 것을 예방할 수 있다. |
| 성능 향상을 위한 설명 | **설명 및 이해 가능한 모델은 보다 쉽게 성능이 향상**될 수 있다. (사용자가 왜 시스템이 특정 값을 출력했는지 알 수 있기 때문) |
| 발견을 위한 설명 | 설명을 통해 새로운 사실이나 정보를 알 수 있다. (AlphaGo Zero 등) |



**[02. XAI의 응용 분야]**

|  |  |
| --- | --- |
| 운송 | 자율주행 자동차는 교통 사고를 줄이고 향상된 이동성을 발휘하지만 설명 가능한 AI 문제에 직면한다. (잘못 분류하면 사고가 발생하기 쉽기 때문) |
| 헬스케어 | 인공지능 모델에 의해 환자를 다루는 문제 등 |
| 법률 | 모델이 정직하고 차별적이지 않아야 한다. |
| 금융 | 데이터 보안, 공정성의 문제 등에 의해 XAI를 필요로 한다. |
| 군사 | DAPRA 프로젝트 때문에 XAI가 중요하게 다루어지고 있다. |

**[03. XAI의 설명 가능 전략]**

|  |
| --- |
| Complexity Related Methods |
| 머신러닝 모델의 복잡도는 해석 가능성과 직결된다. 즉 더 복잡한 모델일수록 해석하기 어려우므로, **모델을 간단히 만드는 것이 XAI에 도움**이 된다.   * **Bayesian Rule Lists (BRL):** Decision Tree에 기반한 모델 * Generalized Additive Model에 기반한 학습 방법을 pneumonia 문제에 응용 * 이미지의 내용을 자동으로 학습하는 attention-based 모델 * **SLIM (Sparse Linear Model)** 을 이용한 data-driven scoring system   + 사용자를 충분히 이해시킬 수 있는, 제안된 시스템의 해석 가능성 |

|  |
| --- |
| Scoop Related Methods |
| 모델의 해석 가능성은 자동화된 모델을 이해할 수 있는 정도를 의미하는데, 이것은 **모델의 전반적인 행동 양상과 특정 데이터에 대한 예측이라는 두 영역**으로 나눌 수 있다.   * **Global interpretability:** 모델의 전반적인 행동 양상 관련 * **Local interpretability:** 특정 데이터에 대한 예측 관련   **Global Interpretability:** **전체 로직에 대한 이해를 가능**하게 하고 **모든 서로 다른 가능한 경우에 대해 추론**할 수 있게 한다.   * 전반적인 수준의 결정에 대한 정보를 제공하는 머신러닝 모델에 적합하다. * **예시:** GIRP (recursive partitioning을 이용한 global model interpretation)   **Local Interpretability:** 모델의 **특정한 결정 또는 예측에 대한 이유**를 설명한다.   * 모델이 어떤 데이터에 대해 특정한 결정을 왜 했는지를 설명할 수 있게 한다. * **예시:** LIME for Local Interpretable Model-Agnostic Explanation |
| (NEXT SLIDE) |

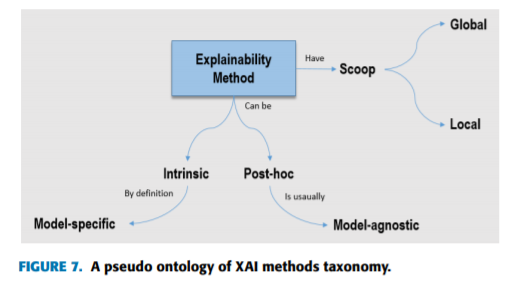
**[03. XAI의 설명 가능 전략]**

|  |
| --- |
| (PREVIOUS SLIDE) |
| Model Related Methods |
| Model agnostic은 XAI 모델이 **어떤 종류의 머신러닝 알고리즘 또는 모델에도** 적용할 수 있는지, 아니면 **특정한 알고리즘/모델에만** 적용할 수 있는지를 나타낸다.  **Model-Specific Interpretability:** 특정한 종류의 모델에만 국한되며, 특정한 종류의 해석이 필요하면 **그 해석을 제공하는 모델에서 제한적인 정보만 제공한다**는 문제점이 있다. 따라서 최근에는 model-free한 해석 방법에 대한 요구가 있다.  **Model-Agnostic Interpretability:** 특정 머신러닝 모델에 국한되지 않고 다양한 모델에 적용할 수 있다. 즉, **예측을 설명과 분리**하는 모델이다.   * 보통 **post-hoc** 모델이며, 일반적으로 ANN을 해석하는 데 사용되며 local 또는 global 해석 모델이 될 수 있다.   **<Visualization>**  머신러닝 모델을 이해하는 아이디어는 **패턴을 탐색하기 위하여 그 모델을 시각화**하는 것이다.   |  |  | | --- | --- | | Surrogate Models | 복잡한 모델을 설명하기 위한 간단한 모델이다.   * Decision Tree와 같은 **해석 가능한 모델**이며, **원래 블랙박스 모델에서의 예측을 기반으로 학습**하여 복잡한 모델을 해석한다. | | Partial Dependence Plot (PDP) | 1개 이상의 입력 변수와 블랙박스 모델의 예측 간의 **평균적인 partial relationship을 시각화**하는 데 사용된다. | | Individual Conditional Expectation (ICE) | PD가 모델의 작업에 대한 coarse한 view를 제공한다면, ICE는 **PDP의 출력을 disaggregate하여 상호 작용과 개별적인 차이를 제공**하려고 한다. |   **<Knowledge Extraction>**  머신러닝 모델이 어떻게 작동하는지 설명하기 어렵지만 **알고리즘이 hidden layer의 cell을 수정하므로 모델의 내부를 표현할 수 있을 것**으로 보인다.   |  |  | | --- | --- | | Rule Extraction | 매우 복잡한 모델로부터 insight를 얻는 방법으로는 **Rule Extraction**이 있으며, 그 효용을 증명하는 작업에서는 **symbolic하고 comprehensive한 접근을 제안**한다. | | Model Distillation | 모델을 압축하여 **Deep Neural Network로부터 정보(dark knowledge)를 전달**하는 방법이다.   * 처음에 계산 복잡도를 줄이기 위하여 제안되었으나, 나중에 모델 해석을 위하여 사용되기 시작했다. | |
| (NEXT SLIDE) |

**[03. XAI의 설명 가능 전략]**

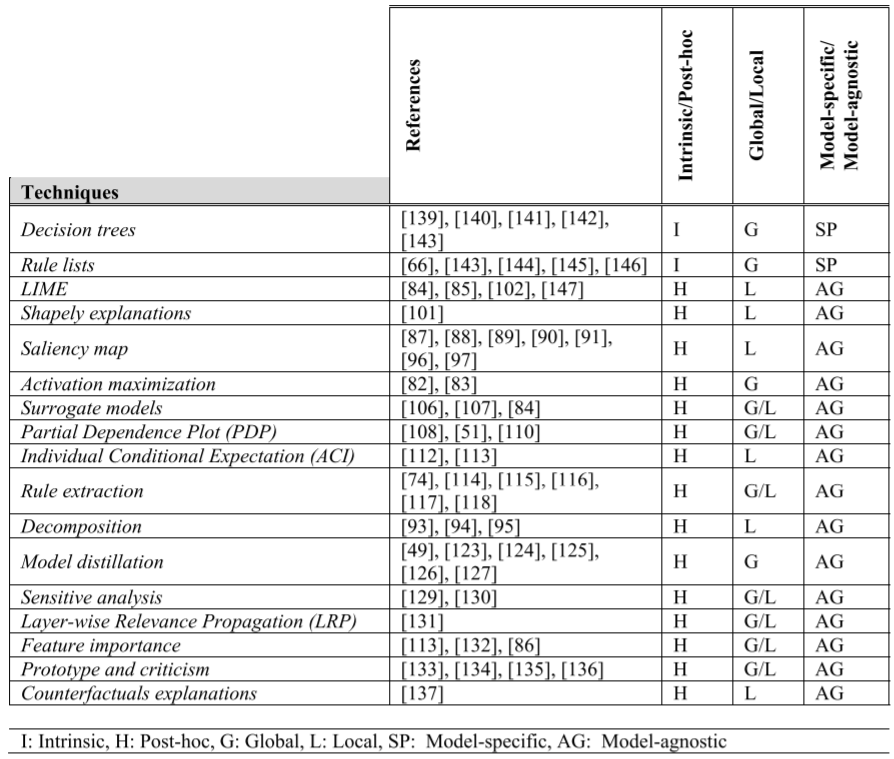
|  |
| --- |
| (PREVIOUS SLIDE) |
| Model Related Methods |
| **<Influence Methods>**  Input의 값을 바꾸는 등의 방법으로 **각 feature의 중요도를 평가**한다.   |  |  | | --- | --- | | Sensitivity Analysis  (SA) | Sensitive는 신경망에서 **특정 입력값 또는 가중치의 조정에 의해 출력값이 얼마나 영향을 받는지**를 의미한다. Sensitivity Analysis는 agnostic explanation 기술로 여겨진다. | | Layer-wise Relevance Propagation (LRP) | 예측 함수를 output layer에서 시작하여 input layer까지 backward propagation한다. (**relevance conservation**) | | Feature Importance | 모델의 예측 오류는 **각 feature의 중요도를 측정**하기 위하여 각 feature를 조합한 후에 측정되는데, **중요하지 않은 feature에 대한 값들의 조합은 모델에 의해 무시**된다. |   **<Example-Based Explanation>**  **dataset에서 특정 인스턴스를 선택**하여 머신러닝 모델의 행동을 설명한다.   |  |  | | --- | --- | | Prototypes and Criticisms | Prototype은 데이터에서 추출한 대표적인 instance이며, 따라서 **그 아이템 구성이 과도한 일반화를 유발하는 Prototype과의 유사도에 의해 결정**된다.   * 이를 방지하기 위하여 **예외적인 경우가 필요**하며, 이를 criticism이라고 한다. | | Counterfactuals Explanations | Unconditional Counterfactual Explanation이란 자동화된 결정에 대한 새로운 종류의 설명이다.   * 알고리즘을 전부 설명할 필요 없이 **대안적인 결정으로 유도되는 최소 조건**을 설명한다. | |

지금까지 정리한 것을 요약하면 다음 그림과 같다.



**[04. XAI 측정: 설명 평가]**

|  |
| --- |
| **XAI의 설명 측정에 대한 3가지 주요 기준:**   * **Application-grounded:** 어플리케이션이 설명을 제공하고 **최종 사용자가 테스트**할 수 있게 한다. * **Human-grounded:** 비교적 간단한 application-grounded evaluation이라고 할 수 있으며, **실험은 도메인의 전문가보다는 일반인을 대상**으로 진행한다. * **Functionality-grounded:** **사람을 필요로 하지 않는다**. 이미 어떤 클래스의 모델이나 인증된 regularizer를 가지고 있을 때 가장 적합하다. |



**[05. Human in the Loop]**

|  |  |
| --- | --- |
| Human-like explanations | * **Contrastive:** 사람은 왜 사건 E가 발생했는지 묻지 안호, 그 대신 왜 **다른 사건 F 대신 E가 발생했는지** 묻는다. * 모든 원인 대신 **1~2가지의 가능한 원인**에 대하여 설명한다. * 설명은 사회적 소통, 상호작용이므로 **설명자는 피설명자의 mental model에 영향**을 주어야 한다. |
| Human-friendly explanations | **XAID (eXplainable AI for Designers):** 게임 디자이너가 XAID를 통해 **AI/머신러닝 모델과 함께 생산**할 수 있도록 하는 인간 중심의 접근을 제안하였다. |