NISTIR 8312

설명가능한 인공지능의 4가지 원칙

미국 상무부 산하 국립표준기술연구소(NIST)는 2021년 9월 <설명가능한 인공지능의 4가지 원칙>에 대한 보고서를 펴냈습니다.
NIST는 이 보고서에서 설명이 요구되는 분야에 도입되는 인공지능이 보장해야 하는 설명의 요건으로 설명성, 의미성, 설명의 정확성, 지식의 한계성이라는 4가지 원칙을 제시하였습니다.

NIST는 이 보고서 외에도 AI 위험관리 프레임워크(AI RMF) 등 인공지능 기술 및 시스템에 대한 신뢰를 구축하기 위한 여러 자료를 발간하였으며, 인공지능 기술에 대한 국제 표준을 주도하고 있습니다.

번역: 정보인권연구소(초벌번역은 기계번역의 도움을 받았습니다. 각주 제외)

이 간행물은 다음 링크에서 무료로 다운로드할 수 있습니다.
https://doi.org/10.6028/NIST.IR.8312

National Institute of
Standards and Technology
U.S. Department of Commerce

NISTIR 8312

설명가능한 인공지능의 4가지 원칙

P. 조나단 필립스
카리나 A. 한
피터 C. 폰타나
에이미 N. 예이츠
크리스틴 그린
정보접근부 정보기술연구실

데이비드 A. 브로니아토프스키
정보기술연구소

마크 A. 프르지보키
정보접근부 정보기술연구실

이 간행물은 다음 링크에서 무료로 다운로드할 수 있습니다:
https://doi.org/10.6028/NIST.IR.8312

2021년 9월

미국 상무부 장관 지나 *M.* 레이몬도

국립표준기술연구소
제임스 K. 올호프, 미국 상무부 표준기술차관의
통상업무 수행, 국립표준기술연구소 소장

이 문서는 실험 절차나 개념을 적절하게 설명하기 위해
특정 상업 조직, 장비 또는 재료를 언급할 수 있습니다.
이러한 표시는 미국 국립표준기술연구소의 추천이나 보증을 의미하지 않으며,
해당 조직, 재료 또는 장비가 반드시 목적에 가장 적합하다고 암시하는 것도 아닙니다.

국립표준기술연구소
기관 간 또는 내부 보고서 8312
Natl. Inst. Technol. Interag. Intern. 8312, 43 페이지 (2021년 9월)

이 간행물은 다음 링크에서 무료로 다운로드할 수 있습니다:
https://doi.org/10.6028/NIST.IR.8312

초록

설명가능한 인공지능(AI) 시스템의 기본 속성을 구성하는 설명가능 AI의 4가지 원칙을 소개합니다. 우리가 제안하는 설명가능 AI시스템이란, 출력과 프로세스에 증거 또는 이유가 수반되고, 개별 사용자가 이해할 수 있는 설명을 제공하며, 시스템의 출력 생성 과정을 정확하게 반영하는 설명을 제공하고, 시스템이 설계된 조건 하에서 그 출력이 충분한 신뢰 수준에 도달한 경우에만 작동하는 것입니다. 이 4가지 원칙을 각각 설명성, 의미성, 설명의 정확성, 지식의 한계성이라고 부릅니다. 이 4가지 원칙은 컴퓨터 과학, 공학, 심리학 분야를 비롯하여 설명가능 AI의 다학제적 특성을 포괄하기 위하여 여러 이해관계자의 참여 속에 개발되었습니다.
만능 설명은 존재하지 않기 때문에 사용자마다 다른 유형의 설명이 필요할 것입니다. 그래서 우리는 설명의 다섯 가지 범주를 제시하고 설명가능 AI의 이론을 요약합니다. 설명가능 알고리즘의 주요 클래스에 해당하는 영역에서 알고리즘에 대하여 개괄합니다. 기본적인 비교 연구로 사람들이 제공하는 설명이 4가지 원칙을 얼마나 잘 따르는지를 평가합니다. 이러한 평가들은 설명가능 AI시스템을 설계할 때 제기되는 문제들에 대한 통찰력을 제공할 것입니다.

키워드

인공지능(AI), 설명가능 AI, 설명가능성, 신뢰할 수 있는 AI.

핵심 요약

AI 분야는 방대하고 복잡하며 지속적으로 진화하고 있습니다. 컴퓨팅 성능이 발전하고 데이터 세트가 점점 커지면서 다양한 애플리케이션 영역에서 AI 알고리즘의 사용 가능성을 탐구하거나 개발하고 있는데 이러한 상황은 사용자의 다양성 및 그와 관련된 위험의 가능성을 안고 있습니다. AI 커뮤니티는 신뢰할 수 있는 AI시스템을 위한 바람직한 특성 중 하나로 설명가능성을 추진해 왔습니다. NIST는 AI 커뮤니티와 협력하여 AI에 대한 신뢰를 구축하는 데 필요한 기술 특성을 추가적으로 확인했습니다. 설명가능성 및 해석가능성 외에도 시스템의 신뢰성을 뒷받침하기 위한 AI시스템의 다양한 특성으로 정확성, 개인정보 보호, 신뢰도, 견고성, 안전성, 보안성(복원력), 유해한 편향의 완화, 투명성, 공정성, 책임성 등이 제시됩니다. AI시스템의 설명가능성 및 여타의 특성은 AI 수명주기의 다양한 단계에서 상호적으로 작용합니다. 이들은 모두 매우 중요하지만, 이 연구에서는 설명가능 AI시스템의 원칙에 초점을 맞춥니다.

이 보고서는 설명가능 AI시스템의 기본적인 전제 조건으로 생각되는 4가지 원칙을 소개합니다. 설명가능 AI에 대한 이 원칙들은 NIST의 공개 워크샵과 공개 의견 수렴에 광범위한 AI 커뮤니티가 참여하여 만들어졌습니다. 모든 AI시스템에 설명이 필요하지는 않다는 점을 잘 알고 있습니다. 하지만 설명할 수 있도록 고안되었거나 설명이 필요한 AI시스템의 경우 다음 4가지 원칙을 준수할 것을 제안합니다:

설명성: 시스템이 출력 및 프로세스에 대한 증거 또는 이유를 제시하거나 포함합니다.

의미성: 시스템이 대상 소비자가 이해할 수 있는 설명을 제공합니다.

설명의 정확성: 설명이 출력을 생성한 이유를 올바르게 반영하거나 시스템의 프로세스를 정확하게 반영합니다.

지식의 한계성: 시스템이 설계된 조건 하에서만 작동하고 그 출력이 충분한 신뢰 수준에 도달하였을 때 작동합니다.

이 작업을 하면서 우리는 프로세스 기반 설명과 출력 기반 설명의 중요성뿐 아니라 설명의 목적과 양식의 중요성을 인식하였습니다. 예를 들어, AI 개발자와 디자이너가 설명에 대해 요구하는 바는 정책 입안자나 최종 사용자와 매우 다를 수 있습니다. 따라서 설명이 요청되는 이유와 설명이 전달되는 방식은 AI 사용자별로 다를 수 있습니다. 4가지 원칙은 AI시스템이 정보 수신자인 인간과 상호작용하는 측면을 많이 고려하였습니다. 주어진 상황의 요구사항, 당면한 업무, 소비자 등은 상황에 적합한 것으로 간주되는 설명 유형에 모두 영향을 미칩니다. 이러한 상황에는 규제기관 및 법률 상의 요구사항, AI시스템의 품질 관리, 고객 관계 등이 포함될 수 있지만 이에 국한되지 않습니다. 설명가능 AI시스템의 4가지 원칙은 광범위한 동기, 사유 및 관점을 포괄하고자 했습니다. 이 원칙을 통해 설명에서 고려해야 할 맥락적 요소를 정의하고 설명 품질을 측정하는 기반을 마련할 수 있을 것입니다.

AI 분야의 복잡성을 고려할 때 시간이 흐를수록 원칙들이 더욱 정교해지고 커뮤니티의 의견을 반영하며 나아질 것이라고 생각합니다. 설명가능성 외에도 AI의 신뢰성에 영향을 미치는 여러 다른 사회 기술적 요소가 있다는 사실을 충분히 이해하고 있습니다. 설명가능한 AI시스템의 원칙에 대한 이번 작업은, 훨씬 더 큰 단위로 진행 중인 NIST AI 포트폴리오 사업의 일부로서, 이 사업은 신뢰할 수 있는 AI의 데이터, 표준, 평가, 검증 및 인증 등 AI 측정에 필요한 모든 것을 다룹니다. NIST가 계측 기관이므로 설명가능 AI시스템의 초기 원칙을 정의하는 것은 향후 측정과 평가 활동에 대한 로드맵 역할을 합니다. AI에 대한 우리 기관의 목표와 활동은 법적 의무, 대통령실 지침, 미국 산업계, 다른 연방 기관 및 글로벌 AI 연구 커뮤니티의 요구사항에 따라 우선순위를 정하고 정보를 얻습니다. 현재의 작업은 훨씬 더 큰 영역의 한 단계에 불과하며, 더 큰 AI 분야와 마찬가지로 이 작업 역시 시간이 흐르면서 계속 진화하고 발전할 것으로 예상됩니다.

목차

그림 목록

1. 소개

저자 중 한 명은 아버지가 암 진단을 받았을 때 종양 전문의와 상담했습니다. 종양 전문의는 암의 상태에 대하여 설명하고 치료 전략과 옵션에 대해 알렸습니다. 또 그는 아버지의 질문에 답하고 치료에서 자신이 맡는 역할에 대하여 설명했습니다. 아버지는 자신이 이 과정의 파트너이며 어느 정도 통제권을 가지고 있다고 느꼈습니다. 아버지는 치료 과정에 대하여 의미 있고 이해하기 쉬운 설명을 들었기 때문에 치료에 대해 신뢰할 수 있었습니다. 의사의 병상 매너도 아버지의 마음에 들었습니다. 의료 기법이 변화하면서 좋은 병상 매너를 갖추는 것이 이제 당연해졌습니다. 인공지능(AI) 시스템이 진단에 관여할 경우 의사에게 권장 사항을 설명함으로써 좋은 병상 매너에 기여할 수 있습니다.

의료 진단은 AI시스템이 사람의 삶에 영향을 미치는 결정에 관여하는 한 가지일 뿐입니다. 다른 예시로는 대출 신청을 평가하고 형벌 양형을 제시하는 시스템이 있습니다. 이 결정들의 특성으로 인해 AI시스템의 출력에 설명을 수반하기 위한 알고리즘, 방법, 기법을 개발하려는 움직임이 활발해졌습니다. 어느 정도 이런 움직임이 촉발된 것은 자동화된 시스템을 비롯한 모든 국가의 의사결정이 그 결정의 근거에 대한 정보를 제공해야 한다는 법률과 규정으로 인한 것입니다. 신뢰할 수 있는 AI를 만들고자 하는 열망이 동기를 부여하기도 하였습니다.

설명가능 AI는 AI시스템의 신뢰를 특징짓는 여러 속성 중 하나입니다. 다른 속성으로는 복원력, 신뢰도, 편향성, 책임성 등이 있습니다. 일반적으로 이러한 용어는 단독으로 정의되지 않고 원칙의 일부 또는 집합으로 정의됩니다. 그 정의는 저자에 따라 다르며, 사회가 AI시스템에 기대하는 규범에 초점을 맞추고 있습니다. 설명가능한 시스템에 대한 요구에 따르면, 답변의 근거를 명확히 제시하지 못하였을 경우 사용자가 해당 시스템에 부여하는 신뢰 수준에 영향을 미칠 것으로 보입니다. 시스템이 편향적이거나 불공정할 것이라는 의심은 개인과 사회에 피해를 가져올지 모른다는 우려로 이어질 수 있습니다. 이는 이 기술의 사회적 수용과 채택을 늦출 수 있습니다.

설명에 대한 요구가 증가함에 따라 이 분야는 AI시스템에서 좋은 설명을 규정하는 원칙적 방법론이 필요해 졌습니다. 첫째, 설명을 사람이 받기 때문에 설명의 특성은 인간 중심적이어야 합니다. 둘째, 사람이 이해할 수 있어야 합니다. 셋째, 설명에는 출력물을 생성하는 시스템의 프로세스가 정확하게 반영되어야 합니다. 설명에 대한 신뢰를 높이기 위해서는 시스템이 설계된 조건을 벗어나 작동하는 경우를 표시해야 합니다. 좋은 설명에 대한 이 핵심 개념은 설명가능 AI에 대한 4가지 원칙의 기초가 됩니다.

이 원칙들은 설명가능 AI의 목표를 달성하기 위해 알고리즘이 작동하는 방식에 영향을 미칠 수 있습니다. 그러나 이 개념의 초점은 알고리즘 방식이나 컴퓨팅 자체에 있지 않습니다. 또 이 원칙은 시스템이 배치되어 사용되는 용도와도 관련이 없습니다. 우리는 그보다 설명을 듣는 인간을 중심으로 4가지 원칙을 구성하여 제시합니다. 이 원칙들은 설명의 품질, 우수성, 정확성, 한계 등 설명의 구성요소를 측정하기 위한 구조를 제시합니다. 구조화된 방식으로 설명을 측정하는 것은 설명 품질을 측정할 수 있는 구체적인 정의를 발전

시키는 데 필수적입니다. 이 원칙들은 이 분야의 향후 연구 방향에 대한 지침이 될 수 있습니다. 4가지 원칙은 설명가능 AI의 측정, 정책적 고려, 안전, 사회적 수용 등 AI 기술의 여타 측면에 대해서도 토대가 될 수 있습니다.

2장에서는 이 원칙들을 제시하고 논의합니다. 3장에서는 설명에 대하여 폭넓은 관점을 채택하여 이를 개념화합니다. 설명가능 AI가 초래하는 위험, 특히 원칙이 충족되지 않을 때 발생할 수 있는 위험에 대하여 간략히 설명합니다(4장). 현재의 작업을 상황에 맞게 반영하기 위하여, 설명가능 AI의 방법론과 평가 지표, 그리고 설명가능 AI에 대한 여타의 기성 원칙들을 검토합니다(5, 6, 7장). 마지막으로, 기존 문헌을 검토하여 인간의 경우 AI에 도입한 원칙을 어느 정도 충족하는지 평가합니다(8장). 인간과 기계에 대한 성능 기대치는 다를 수 있습니다. 상황에 따라서 이러한 기대치가 적절할 수도 있고 그렇지 않을 수도 있지만, 이를 비교할 수 있는 기준선이 필요합니다.

2. 설명가능 AI의 4가지 원칙

설명가능 AI시스템에 대한 4가지 기본 원칙을 제시합니다. 이 원칙들은 AI시스템이 정보 수신자인 인간과 상호작용하는 측면을 많이 고려하였습니다. 주어진 상황의 요구 사항, 당면한 업무, 소비자 모두 상황에 적합한 것으로 간주되는 설명 유형에 영향을 미칩니다. 이러한 상황에는 규제기관 및 법률 상의 요구사항, AI시스템의 품질 관리, 고객 관계 등이 포함될 수 있지만 이에 국한되지 않습니다. 4가지 원칙은 광범위한 동기, 사유 및 관점을 포괄하고자 했습니다. 이 원칙들은 설명을 생성하는 시스템에 적용되며, 머신러닝 기술뿐 아니라 모든 범위의 AI 기술을 지원합니다.

원칙을 자세히 살펴보기 전에 이 문서에서 사용되는 세 가지 핵심 용어인 설명, 출력, 프로세스를 조작적으로 정의합니다. **설명**이란 시스템의 출력 또는 프로세스와 관련된 증거, 토대 또는 추론입니다. 시스템의 **출력**이란 i) 시스템의 출력 또는 ii) 작업을 수행하는 기계 또는 시스템이 취한 조치로 정의합니다. 시스템의 출력은 작업마다 다릅니다. 대출 신청의 경우 출력은 승인 또는 거부라는 결정입니다. 추천 시스템의 경우 추천 영화 목록이 출력될 수 있습니다. 문법 검사 시스템의 경우 출력은 문법 오류 및 수정 권장 사항이 될 것입니다. 분류 시스템의 경우 객체 식별자 또는 스팸 탐지기가 될 수 있습니다. 자동 운전의 경우 내비게이션 자체가 될 수 있습니다. **프로세스**는 시스템의 기반이 되는 절차, 설계 및 시스템의 작업흐름(workflow)을 의미합니다. 여기에는 시스템에 대한 문서, 시스템 개발에 사용된 데이터 또는 저장된 데이터에 대한 정보, 시스템에 대한 관련 지식이 포함됩니다.

설명가능 AI의 4가지 원칙은 다음과 같이 요약할 수 있습니다.

설명성: 시스템이 출력 및 프로세스에 대한 증거 또는 이유를 제시하거나 포함합니다.

의미성: 시스템이 대상 소비자가 이해할 수 있는 설명을 제공합니다.

설명의 정확성: 설명이 출력을 생성한 이유를 올바르게 반영하거나 시스템의 프로세스를 정확하게 반영합니다.

지식의 한계성: 시스템이 설계된 조건 하에서만 작동하고 그 출력이 충분한 신뢰 수준에 도달하였을 때 작동합니다.

이러한 정의는 아래에서 보다 자세한 맥락에 적용됩니다. 그림 1은 원칙을 보여 주며 설명가능한 시스템으로 간주되려면 먼저 설명이 있거나 접근 가능한 증거가 포함되어 있어야 함을 나타냅니다.

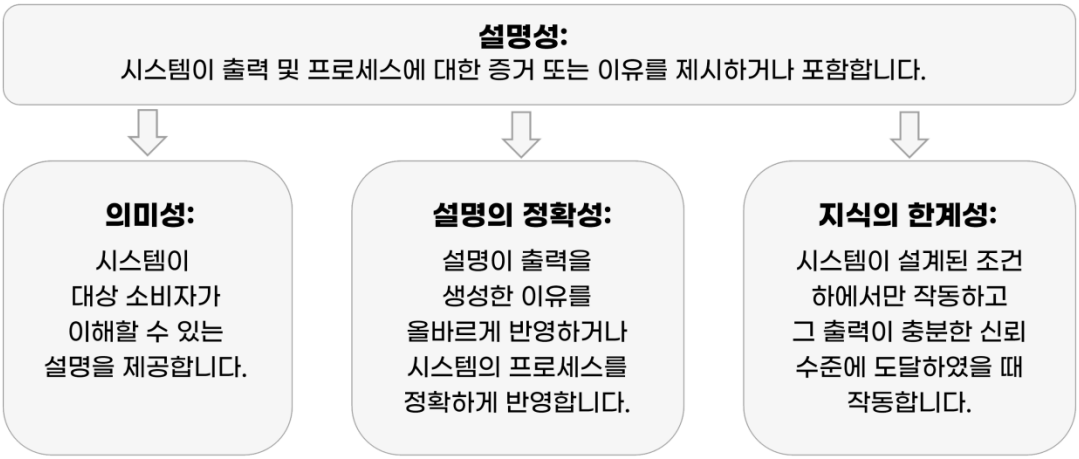

그림 1. 설명가능한 인공지능의 4가지 원칙에 대한 그림. 화살표는 시스템이 설명가능하려면 설명을 제공해야 함을 나타냄. 나머지 세 가지 원칙은 이러한 설명의 기본 속성임.

2.1. 설명성

설명성 원칙이란, 시스템이 설명가능한 것으로 간주되기 위해서는 AI시스템의 결과 또는 프로세스와 관련된 증거, 지원기능 또는 추론을 제공해야 한다는 것입니다. 설명 원칙 그 자체는 설명이 올바른지, 유익한지, 이해하기 쉬운지 여부와 무관합니다. 이 원칙은 설명에 어떤 품질 지표도 부과하지 않습니다. 이 요소들은 설명의 의미성과 설명의 정확성 원칙의 구성요소입니다. 실제로 설명은 주어진 시스템과 시나리오에 따라 달라질 수 있으며, 그래야 합니다. 즉, 설명이 실행되거나 시스템에 포함될 수 있는 방법은 매우 다양할 것입니다. 광범위한 애플리케이션에 적용하기 위해 우리는 의도적으로 설명에 대해 광범위한 정의를 채택하였습니다.

2.2. 의미성

의도된 수신자가 시스템의 설명을 이해하면 시스템이 의미성 원칙을 충족하는 것입니다. 설명에는 설명을 의미있게 만드는 여러가지 공통점이 있습니다. 예를 들어, 시스템이 특정 방식으로 작동한 이유를

설명하는 것이 특정 방식으로 작동하지 않은 이유를 설명하는 것보다 더 이해되기 쉬울 수 있습니다. 각각의 사람이 "좋은" 설명이라고 생각하게 하는 바는 여러 요소에 달렸습니다. 따라서 개발자는 의도된 목표 청중을 고려해야 합니다. 사람들이 정보를 중요하거나 관련성이 있거나 유용하다고 생각하는 데에는 여러 가지 요소가 영향을 미칩니다. 여기에는 개인의 사전적인 지식과 경험, 사람 간의 전반적인 심리적 차이 등이 포함됩니다. 사람들이 의미 있다고 생각하는 내용도 업무나 시스템에 대한 경험이 쌓이면서 시간이 지나면 달라질 것입니다. 또한 사람들은 집단별로 시스템의 설명에서 다른 것을 원하게 됩니다. 집단은 시스템에 대한 역할이나 관계에 따라 광범위하게 정의할 수 있습니다. 예를 들어, 시스템 개발자는 설명에서 원하는 바가 최종 사용자와 다를 가능성이 높습니다.

무엇이 의미 있는 것으로 여겨지는지는 청중에 따라서 뿐 아니라 설명의 목적별로도 달라질 수 있습니다. 주어진 상황에서 여러 시나리오와 요구사항에 따라 무엇이 중요하고 유용한지가 결정 됩니다. 청중의 요구, 전문성의 수준, 당면한 문제 또는 질의와의 관련성을 이해하였을 때 의미성의 원칙을 충족할 수 있습니다. 목적에 대한 자세한 설명은 3절에서 확인할 수 있습니다.

의미성 원칙을 측정하는 일은 계속 이루어져야 할 작업 분야입니다(7.1절). 다양한 청중에게 적합한 측정 프로토콜을 개발하는 것이 과제입니다. 우리는 이를 부담으로 생각하기보다 설명의 맥락에 대한 인식과 이해가 모두 AI 설명의 품질을 측정하는 역량을 보탤 것이라고 생각합니다. 따라서 이러한 요소의 범위를 제한하면 목표 지향적이고 의미 있는 방식으로 설명을 실행할 수 있는 가능성을 구속할 수 있습니다.

2.3. 설명의 정확성

설명성과 의미성 원칙은 모두 목표 청중이 이해할 수 있는 설명을 생성할 것을 시스템에 요구할 뿐입니다. 이 두 가지 원칙은 설명이 시스템의 출력 생성 프로세스를 정확하게 반영할 것을 요구하지 않습니다. 설명의 정확성 원칙은 시스템의 설명에 정확성을 부여합니다.

설명의 정확성은 결정의 정확성과 별개의 개념입니다. 결정의 정확성은 시스템의 판단이 옳은지 옳지 않은지를 나타냅니다. 시스템의 결정 정확성과 무관하게, 이에 수반된 설명은 시스템이 어떻게 결론이나 조치에 도달했는지 정확하게 설명할 수도 있고 그렇지 않을 수도 있습니다. AI 연구자들은 알고리즘 및 시스템 정확도에 대한 표준 측정 기준을 개발했습니다. 이렇게 수립된 결정의 정확성 지표가 존재하지만, 설명의 정확성에 대한 성능 지표에 대해서는 개발 중에 있습니다. 7.2절에서는 이 주제에 대해 현재 진행 중인 연구를 검토합니다.

또한 설명의 정확성은 설명의 세부 수준을 고려해야 합니다. 일부 청중이나 목적에 따라서 간단한 설명으로 충분할 수 있습니다. 추론을 할 때 중요한 요점에 간단명료하게 초점을 맞추거나 방대한 세부 사항 없이 높은 수준의 추론을 제공할 수 있습니다. 간단한 설명에는 알고리즘의 출력 생성 프로세스를 완전하게 규정할 수 있는 뉘앙스가 부족할 수 있습니다. 그러나 이러한 뉘앙스는 시스템 전문가처럼 특정한 청중에게만 의미 있을 수 있습니다. 이는 인간이 복잡한 주제를 설명하는 방식과 유사합니다. 신경과학

교수가 동료에게 새로운 연구 결과를 설명할 때 방대하고 기술적인 세부 사항을 들 수 있습니다. 같은 연구 결과를 학부생에게 설명할 때는 관련된 더 중요한 세부 사항을 설명하기 위해 내용을 압축하고 변경할 수 있습니다. 같은 교수라도 교육을 받지 않은 친구나 부모에게는 그 결과를 매우 다르게 설명할 수 있습니다.

여기서 중요한 것은 설명의 정확성과 의미성이 상호작용한다는 점입니다. 자세한 설명은 시스템의 처리 프로세스를 정확하게 반영할 수 있지만 특정 청중에게 유용하고 쉬운 접근 방식을 희생시킬 수 있습니다. 마찬가지로 짧고 단순한 설명은 잘 이해될 수 있지만 시스템의 특성을 충분히 반영하지 못할 수 있습니다. 이러한 고려 사항들을 감안하면 이 원칙은 설명의 정확성 측정 기준 면에서 유연성을 허용합니다.

2.4. 지식의 한계성

앞의 원칙들은 시스템이 설계되었거나 가지고 있는 지식 범위 내에서 작동한다고 암묵적으로 가정합니다. 지식의 한계성 원칙은 시스템이 작동하도록 설계되거나 승인되지 않은 경우 또는 답변이 신뢰할 수 없는 경우를 식별하는 것을 말합니다. 지식의 한계성을 식별하고 선언함으로써 부적절한 판단을 내릴 수 있는 때 판단하지 않도록 응답을 보장하는 것입니다. 이 원칙은 오인될 소지가 있거나 위험하거나 부당한 결과를 방지하여 시스템에 대한 신뢰를 높일 수 있습니 다.

시스템이 지식의 한계에 도달하거나 이를 초과할 수 있는 방식에는 두 가지가 있습니다. 한 가지 방식은 시스템에 대한 작업이나 쿼리가 해당 영역을 벗어나는 경우입니다. 예를 들어 새의 종류를 분류하기 위해 구축된 시스템에 사용자가 사과의 이미지를 입력할 수 있습니다. 시스템은 입력된 이미지에서 새를 찾을 수 없으므로 응답을 제공할 수 없다는 답변을 내놓을 수 있습니다. 이는 답변인 동시에 설명이기도 합니다. 두 번째 방식은, 내부 신뢰도 임계값에 따랐을 때 가장 가능성이 높은 답변의 신뢰도가 너무 낮은 경우입니다. 새 분류 시스템의 예시를 다시 살펴보자면 새의 입력 이미지가 너무 흐릿하여 새의 종류를 판단하기 어려울 수 있습니다. 이 경우 시스템은 이미지가 새라고 인식하지만 이미지의 품질이 낮다고 확인할 수 있습니다. 예시적으로는 "이미지에서 새를 찾았지만 이미지 품질이 너무 낮아 새를 식별할 수 없습니다."와 같이 출력할 수 있습니다.

2.5. 요약

설명가능 AI시스템에 대한 광범위한 요구와 적용 분야를 감안해 보았을 때, 한 시스템이 두 가지 유형 이상의 설명을 생성할 수 있을 때 보다 설명가능하고 원칙을 더 잘 충족한 것으로 볼 수 있습니다. 또한 설명의 정확성을 평가하는 데 사용되는 지표가 보편적이거나 절대적이지 않을 수 있습니다. 현재 설명가능 AI의 방법론을 개발하고 검증하기 위한 작업이 계속 진행 중입니다. 이러한 노력에 대한 개요는 6절과 7절에 나와 있습니다. 여기서 4가지 원칙은 설명 자체가 사용자의 요구를 충족하는지 여부를 고려하는 방법론적 지침으로 사용됩니다.

설명가능 AI는 활발히 연구되고 있는 분야입니다. 이 분야가 새로운 지식과 데이터로 성장함에 따라 이 시스템에 대한 우리의 이해와 활용도도 달라질 것입니다. 따라서 원칙은 시스템의 필요에 대해 생각하는

방법을 안내합니다. 이러한 원칙들은 새로운 과제와 질문에 접근하는 바탕이 됩니다.

3. 설명의 목적과 양식

설명의 광범위한 범주를 다루기 위해서 우리는 설명을 목적과 양식이라는 두 가지 속성으로 규정합니다. 목적은 사람이 누군가 설명을 요청하는 이유가 무엇인지 또는 설명을 통해 어떤 질문에 답하고자 하는지에 대한 것입니다. 양식은 설명이 전달되는 방식을 말합니다.

청중의 경우 설명의 목적과 설명이 제공하는 정보에 큰 영향을 미칩니다. 이 정보는 다양한 집단의 사람들, 그리고 이들이 시스템에서 맡은 역할에 따라 달라집니다. 시스템의 구축자는 AI 모델의 디버깅 또는 학습 데이터의 평가에 대한 설명을 요청할 수 있습니다. 규제 당국은 시스템이 명시된 규제 요건을 충족하는지에 대하여 문의할 수 있습니다.

설명의 목적은 결국 양식에 영향을 미칩니다. 그림 2에서 우리는 양식의 세 가지 요소인 세부 수준, 인간과 기계의 상호작용 정도, 그리고 형식을 시각화했습니다. 이 속성들이 전부는 아니고 설명은 다양한 형태를 취할 수 있습니다. 그래도 우리는 이 요소들이 4가지 원칙의 충족과 밀접한 관련이 있다고 강조합니다. 따라서 이들 요소를 고려하는 것이 설명 생성의 토대가 됩니다. 아래에서 이에 대해 더 자세히 설명합니다.

세부 수준은 범위로 표시되는데 희박함부터 광범위함까지 있습니다. 희박함은 제공되는 정보의 양이 간략하고 제한적이며 높은 수준의 세부 사항이 부족하다는 의미입니다. 희박한 설명의 예시로는 알람 시스템에서 내리는 결정에 대한 설명을 들 수 있습니다(예: "과열로 인해 시스템의 프로세스가 느려졌습니다."). 광범위한 설명은 시스템에 대해 자세한 정보를 포함하거나 많은 양의 정보를 제공할 수 있습니다(예: 관련 시스템 정보가 포함되어 그 과정을 이해할 수 있는 보고서)

우리는 인간과 기계의 상호작용 정도를 선언적 설명, 일방향 상호작용, 쌍방향 상호작용의 세 가지 범주로 분류합니다. 선언적 설명의 경우에는 시스템이 설명을 제공하고 그 이상의 상호작용이 없습니다. 이와 같은 설명 방식은 현재 설명가능 AI의 방법으로 가장 많이 사용되고 있습니다(6절). 예를 들어, 대출 신청 시스템은 항상 승인 또는 거부만을 출력할 수 있습니다. 객체 분류기는 중요도 맵을 출력할 수 있습니다. 모델 카드는 시스템에 대하여 미리 판단한 정보를 포함할 수 있습니다. 선언적 설명은 "객체 분류기가 왜 이런 판단을 내렸습니까?"와 같은 기본 쿼리에 기반한 것입니다. 사람이 질의 내용을 변경할 수는 없습니다(시스템 자체를 변경하여 다른 결과를 산출하는 경우는 제외).

더 높은 수준의 상호작용은 일방향 상호작용입니다. 이 경우 시스템에 입력된 쿼리 또는 질문에 기반하여 설명이 결정됩니다. 예를 들어, 사용자가 시각화하고자 하는 요소에 따라 그래픽으로 출력될 수도 있습니다. 이를 통해 설명의 소비자는 추가적으로 탐색하거나 다른 쿼리를 제출할 수 있습니다.

상호작용 수준이 가장 높은 범주를 쌍방향 상호작용이라고 정의합니다. 이 경우는 사람들 간 대화를 모델링합니다. 사람이 더 깊이 탐색할 수도 있고, 기계가 다시 조사하거나 질문을 명확히 하거나 새로운

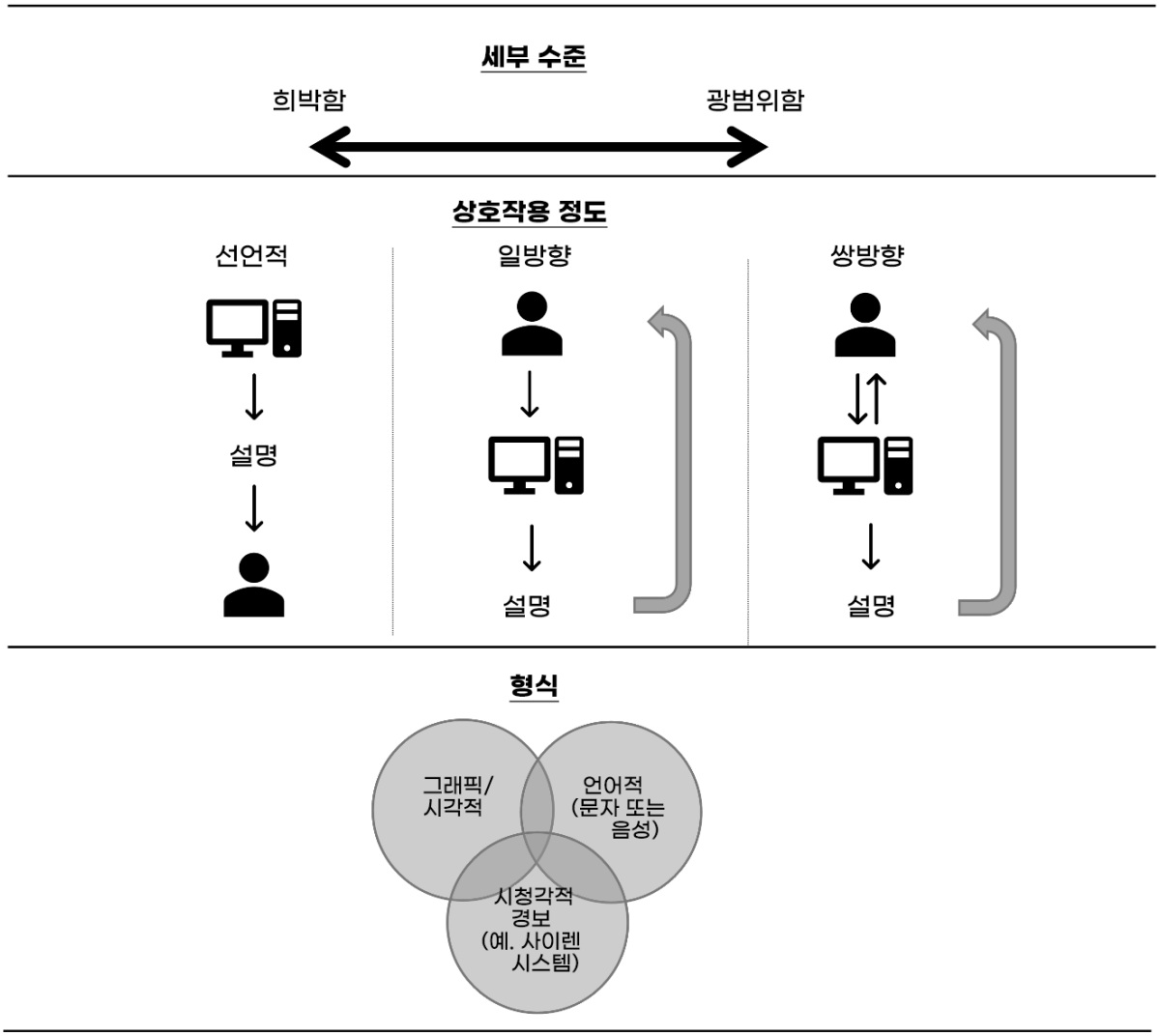

그림 2. 저자들의 설명 양식에 대한 도해

탐색 경로를 제시할 수도 있습니다. 예를 들어, 시스템이 사용자에게 추가 세부 정보를 묻거나 다른 질문을 제안할 수 있습니 다. 저자들이 알기로 쌍방향 상호작용은 아직 존재하지 않습니다. 이를 개발하는 것이 향후 연구 방향입니다.

설명의 형식에는 그래픽/시각적, 언어적, 시청각적 경보가 있습니다. 그래픽 형식의 예시로는 데이터 분석의 결과물이나 중요도 맵 등이 있습니다. 언어 형식에는 문자로 된 출력물이나 보고서뿐 아니라 음성 같은 청각적 출력물이 포함될 수 있습니다. 이러한 시각적 및 언어적 형식은 청중이 설명을 기대하고 주의를 기울이고 있다고 가정하고 있습니다. 또다른 형식의 설명은 주의하고 있지 않은 청중을 집중시킵니다. 사이렌이나 조명 시스템은 다양한 경보음, 조명 점멸 패턴 및 밝은 색을 사용한 설명으로 청중의 주의를 끌 있습니다. 예를 들어, 특정 사이렌 음조나 패턴으로 주의가 필요한 시스템의 상태를 나타낼 수 있습니다.

목적에 맞는 설명을 생성하고 4가지 원칙을 충족하기 위해서는 각각의 양식 요소를 검토해야 합니다. 어떤 경우에는 단순하고 선언적인 설명이 의미를 최적화하는 데 가장 적합한 양식일 수 있습니다. 토네이도가 발생했을 때처럼 기상 비상 사태인 경우가 이에 해당합니다. 미국 기상청의 현재 날씨 경보는 "토네이도 주의보: 조치를 취하세요!"라는 경보와 간단한 설명으로 작동합니다. 이 경보는 "토네이도 주의보" 라는 간단한 설명으로 "조치를 취하"도록 합니다. 측정 기준에 따르면 이런 설명이 아주 정확한 수준이라고 간주되지는 않을 것입니다. 토네이도 경보가 선포된 이유를 설명에 포함하지 않고 최소한의 세부 정보만 제공하였기 때문입니다. 그러나 이 예시에서는 간결하게 설명하는 것이 해당 상황에 적절한데, 이는 많은 사람이 이해할 수 있고 신속한 조치를 취할 수 있기 때문입니다. 최소한의 정보이기는 했지만 경보와 함께 추가 정보를 제공하면 기상 경보에 대한 규정 준수(예: '울음소리 효과') 문제를 해결하는 데 도움이 될 수 있습니다. 시스템 디버깅과 같은 다른 사례에서는 시스템 내부 단계에 대한 정보를 설명에 포함할 수 있습니다. 이때는 설명이 길어질 수 있고 분야별 전문 용어가 포함될 수 있습니다. 청중은 설명을 검토하고 다음 조치를 결정하는 데 시간과 노력을 더 들여야 할 수 있습니다. 이 경우 더 자세한 정보를 사용자가 선호하는 형식으로 제공하는 것이 도움될 수 있으며, 쌍방향 상호작용이 중요해질 수 있습니다.

이처럼 다양한 목적, 양식 및 고려 사항들이 설명의 여러 범주와 유형을 나타냅니다. 이는 설명이 필요한 시스템의 범주를 다룰 때 유연할 필요가 있음을 시사합니다. 설명이 제공되는 상황은 다양하므로 4가지 원칙은 다양한 양식을 적절히 적용할 것을 권장합니다. 어떤 설명은 다른 설명보다 더 쉽게 이루어질 수 있으며, 설계자는 서로 다른 목표를 추구할 때의 장단점을 고려해야 합니다.

4. 설명가능 AI의 위험 관리

위험은 "목표의 불확실성으로 인한 영향"으로 정의되며, 부정적인 결과(위협)와 긍정적인 결과(기회)를 모두 포함합니다. 위험 관리는 위험을 정의, 평가, 완화하는 데 사용할 수 있는 프로세스입니다. 설명가능 AI는 모델 또는 시스템의 위험을 평가, 측정 또는 예측하여 인공지능의 위험을 완화할 수 있습니다.

설명은 취약점을 테스트하는 데 사용할 수 있습니다. 반대로 6.3절에서 설명한 것처럼 설명가능 AI가 적대적 공격 등 그 자체로 위험을 초래할 수 있습니다. 이 절에서는 후자, 즉 설명가능 AI로 인해 발생할 수 있는 잠재적 위험을 관리하는 문제에 초점을 맞춥니다.

모든 설명가능 AI시스템에는 위협과 기회라는 잠재적 위험이 존재합니다. 이해관계자가 목표와 일반적인 위험의 트레이드오프를 받아들일 준비가 되어 있는 정도를 이해관계자의 위험 선호도라고 합니다. 많은 위험 관리 전략은 식별, 분석, 대응, 모니터링 및 검토 라는 구성요소를 공통적으로 가지고 있습니다. 설명가능 AI를 위해서는 위험 관리 전략에서 4가지 원칙을 고려해야 합니다.

첫 번째 원칙인 설명성은 설명가능 AI를 위해 필요하지만, 설명은 그 자체로 긍정적이든 부정적이이든 위험을 초래합니다. 설명으로 인해 초래될 수 있는 부정적인 결과로는 독점적인 세부 정보가 노출되는 것입니다. 단순한 설명은 시스템의 내부 작동을 노출하지 않을 수 있습니 다. 그러나 독립적 쿼리 여러 개 또는 쌍방향 상호작용을 통해 여러 건의 설명이 서로 연결될 경우 지적 재산이 노출될 수 있습니다. 최종 사용자가 시스템을 이해하려면 얼마나 많은 설명에 접근할 수 있어야 할까요? 이 수는 설명의 범위에 따라 달라질 수 있 습니다. 여기에는 시스템의 지식의 한계성에 대한 설명이 포함됩니다.

그러나 설명은 긍정적인 결과도 가져올 수 있습니다. 사용자가 시스템을 더 잘 이해할 수 있습니다. 또 이는 시스템에 대한 신뢰도 향상과 같은 개선으로 이어질 수 있습니다. 공정신용보고법(FCRA) 및 유럽연합 일반 개인정보보호 규정(GDPR) 제13조와 같은 규정을 준수하기 위해 설명이 필요할 수도 있습니다.

설명은 청중에게 의미 있는 것이어야 합니다. 이는 두 번째 원칙인데 자체적인 위험을 수반합니다. 의미 있는 설명은 시스템에 대해 더 깊은 통찰력을 제공할 수 있지만 내부 작동을 노출시켜 지적 재산이나 시스템 취약성을 노출시킬 수 있습니다. 반면 의미 없는 설명은 무시되거나 설명으로 인정받지 못할 위험에 처할 수 있습니다.

유용한 설명이 되려면 설명이 의미 있어야 할 뿐 아니라 정확해야 한다는 것이 세 번째 원칙입니다. 이와 관련된 잠재적 위험은 일반적으로 모델 위험으로 알려져 있으며, 이는 유효하지 않거나 잘못 적용된 모델에서 부정적 결과가 파생될 수 있다는 의미입니다. Federal Reserve System(2011)이 언급한 바와 같이, 이 유형 위험의 두 가지 주요 원인 중 하나는 잘못된 결과를 초래하는 모델의 근본적인 오류를 듭니다. 부정확한 설명은 시스템의 작동 방식이나 결과에 대하여 오인이나 오해를 초래할 수 있습니다. 이는 최종 사용자에게도 부정적인 위험이지만, AI시스템이 더 큰 시스템의 일부로 사용되었을 때 AI가 인간에게 편향을 낳을 수도 있습니다.

얼굴 인식에서 인간 얼굴 검사자가 AI 알고리즘에서 얼굴의 어느 부분이 유용한지 정보를 받을 수 있었다고 봅시다. 정확한 설명은 심사관이 한 쌍의 얼굴을 더 정확하게 평가하는 데 도움이 될 수 있지만 부정확한 설명은 잘못된 결정으로 이어질 수 있습니다. 사법 체계에서는 피고인을 다시 구금할지 여부에 대한 결정 같은 경우에 AI 알고리즘이 사용되고 있습니다. 알고리즘이 어떻게 결과에 도달하는지에 대한 부정확한 설명은 사법적인 오심을 초래할 수 있습니다. 정확한 설명은 보다 정의로운 사회를 만드는 데

도움이 될 수 있습니다.

모델 위험의 또 다른 주요 원인은 모델을 잘못 사용하거나 지식적인 차원의 한계를 넘어서는 것 입니다. 네 번째 원칙인 시스템의 지식의 한계성을 설명하는 설명은 모델이 범위를 벗어나 작동하지 않는다는 확신을 주고 신뢰를 키울 수 있습니다. 시스템의 한계를 설명하는 것은 시스템의 내부 작동을 잠재적으로 노출할 수 있는데, 다른 설명에서 수집한 정보와 결합할 경우 더욱 그럴 수 있습니다.

소프트웨어 노출의 잠재적 위험의 경우 상황, 범주 및 위험 수준 별로 다양합니다. 최종 사용자가 누구인가요? 최종 사용자가 개발자 등 조직 내부에만 있는 경우 이에 대한 관리 전략은 외부 고객을 포함하는 경우와 다를 수 있습니다.

설명가능 AI는 시스템에 새로운 위협을 가져옵니다. 하지만 새로운 기회도 제공합니다. 결과가 위협이 될지 기회가 될지는 청중에 따라 달라질 수 있습니다. 위험 관리는 이러한 요소와 다른 요소의 장단점과 가능성을 고려합니다. 위험을 평가할 때 흔히 평가하는 두 가지 요소는 위험의 발생가능성과 결과의 영향력입니다.

AI의 경우 일반적으로 위험 관리 프레임워크를 개발할 필요가 있습니다. 관련하여 NIST가 2021. 7. 29. 에 정보를 요청한 바 있습니다. 위험 관리에 대한 자세한 내용은 참고문헌 22, 37, 128을 참조하세요.

5. 문헌에 나타난 원리의 개요

설명가능 AI의 이론과 성질은 다양한 관점에서 논의되어 왔으며, 그 관점들에는 공통점과 차이점이 있습니다.

Lipton(2018)은 설명가능한 기법을 투명성과 사후 해석가능성이라는 두 가지 큰 범주 로 구분합니다. Lipton은 투명한 설명을 시스템이 어떻게 산출물에 도달했는지를 어느 정도 반영하는 것으로 정의합니다. 하위 클래스는 시뮬레이션 가능성으로, 사람이 전체 모델을 파악할 수 있어야 한다고 봅니다. 이는 설명이 시스템의 내부 작동을 반영한다는 것을 의미합니다. 사후 설명은 "종종 모델의 작동 방식을 정확하게 설명하지는 못하지만, 그럼에도 불구하고 머신러닝 실무자와 최종 사용자에게 유용한 정보를 제공할 수 있습니다." 예를 들어, 어떤 새는 해당 학습 세트에서 홍관조와 유사하기 때문에 홍관조입니다.

Rudin(2019)은 최첨단 정확도를 위해 해석가능성을 희생해야 한다고 가정해서는 안 된다고 주장합니다. 그들은 중요한 의사결정의 경우, 동일한 수준의 정확도를 가진 해석가능한 모델이 존재하지 않는다는 사실을 증명하지 못하는 한 블랙박스 모델을 피해야 한다고 권고합니다. 이 문서의 나머지 부분에서는 블랙박스를 폐쇄박스(closed-box)라고 지칭할 것입니 다. Rudin 등(2021)은 이전 연구를 기반으로 해석가능한 기계 학습의 5가지 원칙과 10가지 과제를 제시하였습니다.

Mueller 등(2021)은 사용자 중심 설명가능 AI시스템의 몇 가지 기본 개념을 검토합니다. 이들은 이 개념을 바탕으로 자기 설명 점수표(Self-Explanation Scorecard)를 고안하였고 사용자 중심의 설계 원칙을 제시하였습니다.

Broniatowski(2021)는 심리학의 관점에서 해석가능성과 설명가능성이 머신러닝 시스템에 대한 분명한 요구사항이라는 주장을 펼쳤습니다. 결과 분석은 시스템의 출력이 다양한 유형의 사용자에 맞게 조정되어야 함을 의미합니다.

Wachter 등(2017)은 설명이 설명 정확도 속성을 충족해야 한다고 주장합니다. 그들은 반사실적(counterfactual) 설명으로도 충분하다고 주장합니다. "예측에 대한 반사실적 설명은 해당 예측을 사전 정의된 출력으로 변경하는 특징값의 가장 작은 변화를 나타냅니다." 예를 들어, 플랫폼에 15분 일찍 도착했다면 기차를 탈 수 있었을 것이라는 식입니다. 반사실적 설명은 시스템의 내부 작동을 드러낼 필요가 없습니다. 이런 속성으로 반사실적 설명은 지적 재산을 보호하는 것으로 인정됩니다.

Gilpin 등(2018)은 설명가능 AI에 대한 일련의 개념을 정의합니다. 지금의 이 연구에서 의미성 및 설명의 정확성 원칙과 유사하게, Gilpin 등은 설명은 해석가능성과 완전성 사이에서 절충이 이루어져야 한다고 제언합니다. 또한, 이들은 이러한 절충이 시스템의 중요 한계점을 모호하게 만들어서는 안 된다고 말합니다.

Doshi-Velez and Kim(2017)은 설명이 사용자나 소비자에게 의미 있는지를 측정하는 중요한 문제를 다룹니다. 이들은 설명의 효율성을 측정하는 과학적 프레임워크를 제시합니다. 논문에서는 설명가능한 시스템의 해석가능성에 대한 테스트를 시작할 때 필요한 요소에 대해 논의합니다. 이 논문은 개념으로서 이러한 원칙을 다루는 것과 측정 기준 및 평가 방법을 만드는 것 사이에 차이가 있다고 강조합니다.

영국 ICO[정보공개와 개인정보보호를 소관하는 영국의 국가기관 - 역주]와 앨런 튜링 연구소(2020)는 설명가능 AI를 위해서 따라야 할 원칙을 제시한 바 있습니다. 이 원칙은 투명할 것, 책임성이 있을 것, 운영 상황을 고려할 것, AI시스템이 영향을 받는 개인과 광범위한 사회에 미치는 영향을 고려할 것 등입니다. 이 기관들은 원칙에 대한 논의 외에도, 프로세스 기반 설명과 결과 기반 설명의 문제, 근거의 문제, 누가 어떤 결정을 내렸는지에 대한 책임 문제, 데이터에 대한 설명 문제, 시스템 사용의 공정성, 안전성, 영향을 극대화하는 설계 조치 등 설명에 포함되는 다양한 요소에 대하여 논의합니다.

Barredo Arrieta 등(2020)은 설명가능성 또는 해석가능성을 설명하기 위하여 여러 문헌에서 사용되는 다양한 용어인 이해 가능성(understandability), 파악 가능성(comprehensibility), 해석가능성(interpretability), 설명가능성(explainability) 및 투명성(transparency)에 대해 논의합니다. 이들은 이러한 용어가 모두 어떻게 다르면서도 서로 연관되어 있는지에 대해 논의합니다.

Weller(2019)는 투명성의 유형과 설명의 다양한 사용자 또는 소비자 계층을 다루는 방법에 대해 논의합니다. 설명의 정확성 원칙과 유사하게 이 논문에서는 설명의 충실성을 다음과 같이 소개합니다.

> *... 중요한 세부 사항을 숨기지 않고 진정한 이해를 정확하게 나타낸다는 의미에서 주어진 설명이 충실하다면 사회에 대체로 유익합니다. 충실하다는 개념은 정확하게 정의하기 어려울 수 있습니다. 이는 법정에서 가끔 "진실, 모든 진실, 오로지 진실만을 말하라"는 지침의 정신과 비슷한 맥락입니다.*

이 관점들에는 공통점과 차이점이 모두 존재합니다. 4가지 원칙과 유사하게, 이들의 공통점은 설명의 존재 여부, 설명이 얼마나 의미 있는지 여부, 설명이 얼마나 정확하거나 완전한지를 구분하는 개념을 포함한다는 것입니다. 차이점이 있긴 하겠지만 이 관점은 설명가능한 시스템을 개발하는 데 지침이 됩니다. 이론들 간 중요한 의견 불일치가 있는 부분은 설명의 의미성과 정확성에 대한 상대적 중요성 문제입니다. 이 의견 불일치는 여러 원칙의 균형을 동시에 충족하기가 어렵다는 사실을 나타냅니다. 애플리케이션의 상황, 커뮤니티 및 사용자의 요구사항, 특정한 작업에 따라 각 원칙의 중요성이 결정될 것입니다.

6. 설명가능 AI 알고리즘의 개요

연구자들은 AI시스템을 설명하기 위해 다양한 알고리즘을 개발해 왔습니다. 다른 문헌을 참고하여 우리는 자체 해석가능 모델과 사후 설명의 두 가지 범주로 설명을 정리했습니다. 자체 해석가능 모델은 사람이 직접 읽고 해석할 수 있는 알고리즘 모델(또는 알고리즘 그 자체에 대한 표현)입니다. 이 경우 모델 자체가 설명입니다. 사후 설명은 알고리즘의 작동 방식에 대한 아이디어를 제공하기 위해 알고리즘을 묘사, 설명 또는 모델링하는 다른 소프트웨어 도구에 의해 생성되는 설명입니다. 사후 설명은 알고리즘의 작동 방식에 대한 내부 지식이 없이도 선택한 입력에 대한 출력을 조회할 수 있다면 알고리즘에 사용할 수 있는 경우가 많습니다.

우리는 다양한 설명의 하위 유형과 사용 가능한 모든 설명을 모두 언급하기보다 널리 사용되는 몇 가지 예시와 분류에 우선 주목하고, 이후 설명가능 AI에 대한 다양한 설문조사를 독자에게 소개합니다.

6.1. 자체 해석가능 모델

자체 해석가능 모델은 그 자체로 설명이 되는 모델입니다. 자기 해석가능 모델은 모델 전체를 전역적으로 설명할 뿐만 아니라, 입력 시뮬레이션이 각각의 모델 입력을 살펴보는 과정을 통해 결정 각각에 대한 지역적 설명을 제공할 수 있습니다.

가장 일반적인 자체 해석가능 모델로는 의사결정 트리와 회귀 모델(로지스틱 회귀 등)이 있습니다. 기본 의사결정 트리와 기본 회귀 모델보다 정확도가 향상된 다양한 해석가능 모델을 만드는 작업이 진행 중입니다. 이러한 모델에는 의사결정 목록, 의사결정 집합, 프로토타입(각 클래스의 대표 표본. Kim 등, 2014), 입력 집합을 완전히 분류하는 특징 조합 규칙(Kuhn 등, 2020), 베이지안 규칙 목록, 가산 의사결정 트리 및 의사결정 트리의 향상된 변형 등이 있습니다.

일부 문헌에서는 자체 해석가능 모델의 정확성과 해석가능성의 트레이드오프를 주장합니다. 즉, 자체 해석가능 모델은 사후 모델보다 정확도가 떨어지며, 이는 모델을 더 정확하게 만드는 것과 인간에게 더 의미 있게 만드는 것 사이에 상충 관계가 있기 때문입니다. 그러나 Rudin(2019), Rudin과 Radin(2019)은 이에 동의하지 않으며, 정확성-해석가능성 사이에 트레이드오프가 반드시 존재하는 것은 아니고 많은 경우 결정 정확도의 손실 없이도 해석가능 모델을 사용할 수 있다고 주장합니다.

6.2. 사후 설명

사후 설명은 지역적(local) 설명과 전역적(global) 설명의 두 가지 종류로 분류됩니다. 지역적 설명은 의사결정의 하위 집합을 설명하거나 의사결정별로 설명합니다. 전역적 설명은 해석할 수 없는 모델을 근사화하는 모델을 생성합니다. 경우에 따라 전역적 설명은 특정 입력에 대해 시뮬레이션하여 해당 입력 개별에 대한 지역 설명을 제공하는 방식으로 지역적 설명을 제공할 수도 있습니다. 간단한 예시로, 로지스틱 회귀(자체 해석가능 모델 또는 불투명 모델에 대한 사후 근사치일 수 있음)를 생각해 보겠습니다. 회귀계수는 모든 입력을 설명하는 전역적 설명을 제공합니다. 그러나 입력을 가중치와 함께 대입한 다음 해당 가중치를 사용하여 알고리즘의 출력을 설명할 수 있습니다.

다음 하위 절에서 이러한 설명 각각에 대해 논합니다. 지역적 설명은 6.2.1절에서, 전역적 설명은 6.2.2절에서 설명합니다.

6.2.1. 지역적 설명

지역적 설명은 입력의 하위 집합을 설명합니다. 가장 일반적인 유형의 지역적 설명은 결정별 또는 단일 결정에 대한 설명으로, 단일 입력 지점에 대한 알고리즘의 출력 또는 결정에 대하여 설명합니다.

일반적으로 사용되는 지역적 설명 알고리즘 중 하나는 LIME(Local Interpretable Model- Agnostic Explainer)입니다. LIME은 결정을 내리고 주변 지점을 쿼리하여 지역적 결정을 나타내는 해석가능 모델을 구축한 다음, 이 모델을 사용하여 기능별 설명을 제공합니다. 기본으로 선택되는 모델은 로지스틱 회귀 분석입니다. 이미지의 경우, LIME은 각 이미지를 슈퍼픽셀로 나눈 다음 모델의 임의 검색 공간에서 쿼리를 수행합니다. 여기서 어떤 슈퍼픽셀을 생략하고 모두 검정색(또는 사용자가 선택한 색상)으로 대체할지 다양하게 변경해 봅니다.

또 다르게 일반적으로 사용되는 지역적 설명 알고리즘은 SHAP(SHapley Additive exPlanations) 입니다. SHAP은 게임 이론에 착안하여 시나리오를 연합 게임으로 변환한 다음 해당 게임에서 샤플리 값을 산출함으로써, 회귀 문제에 대한 입력에 있어 기능별로 중요도를 제시합니다. SHAP은 특징을 플레이어로, 특징 값과 기본값을 전략으로, 시스템 출력을 보상으로 취급하여 입력에서 연합 게임을 형성합니다. 샤플리 값과 연합 게임에 대한 자세한 내용은 Ferguson(2014)을 참조하세요.

또 다른 일반적인 지역적 설명은 반사실입니다. 반사실은 "만약 입력이 이번 신규 입력대로였다면 시스템은 다른 결정을 내렸을 것"라고 말하는 설명입니다. 이러한 설명 방식에서는 서로 많이 다른 사례가 여러 가지 제시되곤 하지만, 반사실적 설명은 하나의 사례를 제시합니다. 시스템이 다른 결정을 내린다는 점을 제외하고는 가능한 한 입력과 유사한 사례를 제공하는 것이 바람직합니다. 그러나 일부 시스템에서 는 여러 개의 반사실 사례를 하나의 설명으로 제시할 수 있습니다. Ustun 등(2019)은 로지스틱(또는 선형) 회귀 모델에 대한 반사실 설명을 개발했습니다. 반사실은 변경될 특정 기능의 양으로 표현됩니다.

이미지 데이터의 문제에 대한 것으로 인기 있는 또다른 지역적 설명의 유형은 중요도 픽셀입니다.

중요도 픽셀(saliency pixel)은 각 픽셀이 분류 결정에 기여하는 정도에 따라 해당 픽셀에 색상을 지정합니다. 최초의 중요도 알고리즘 중 하나는 클래스 활성화 맵(Class Activation Maps, CAM) 입니다. CAM을 개선한 인기 있는 픽셀 알고리즘은 GRAD-CAM입니다. GRAD-CAM은 모든 컨볼루션 네트워크를 설명할 수 있도록 CAM을 일반화했습니다.

또한 Koh와 Liang(2017)의 지역적 설명은, 결정을 내린후 각 학습 데이터 포인트가 특정 결정에 미친 영향에 대한 추정치를 생성합니다. 또다른 지역적 설명으로는 ICE(Individual Conditional Expectation)가 있습니다. ICE 곡선은 데이터 인스턴스에 대한 어떤 특징의 변화로 인한 한계 효과를 보여줍니다.

6.2.2. 전역적 설명

전역적 설명은 전체 알고리즘에 대한 사후 설명을 생성합니다. 여기에서 알고리즘이나 시스템에 대하여 전역적 모델을 생성하는 것을 포함하기도 합니다.

전역적 설명 중 하나는 부분의존성 플롯(PDPs)입니다. PDPs는 특징(특정 데이터 열 또는 구성 요소의 값)이 변할 때 예측된 응답의 한계 변화를 보여줍니다. PDPs는 특징과 응답 사이의 관계가 선형적인지 아니면 더 복잡한지 판단하는 데 유용합니다.

심층 신경망에서 이러한 전역적 알고리즘 중 하나는 TCAV(Testing with Concept Activation Vectors) 입니다. TCAV는 신경망 상태를 CAV(Concept Activation Vectors)라는 인간 친화적 개념의 선형 가중치로 표현하여 보다 사용자 친화적인 방식으로 신경망을 설명하고자 합니다. 색상이 이미지 분류기의 결정에 어떤 영향을 미치는지 알아보기 위하여, TCAV가 색상을 포함하는 CAV를 학습하여 이미지 분류 알고리즘을 설명하는 데 사용되었습니다.

전역적인 설명을 제공하는 데 사용되는 두 가지 시각화 방식으로는 PDPs와 ICE가 있습니다. PDPs는 특징(특정 데이터열 또는 구성요소의 값)이 변할 때 예측된 응답의 한계 변화를 보여줍니다. PDPs는 특징과 응답 사이의 관계가 선형적인지 아니면 더 복잡한지 판단하는 데 유용합니다. ICE 곡선은 더 세분화되어 있으며 데이터의 각 인스턴스에서 어떤 특징의 변화로 인한 한계 효과를 보여줍니다. ICE 곡선은 PCP에서 시각화된 관계가 모든 ICE 곡선에서 동일한지 확인하는 데 유용하며, 잠재적인 상호작용을 식별하는 데 도움이 될 수 있습니다.

프로토타입은 각 클래스의 대표 샘플이며, 때로 6.1절에서 언급한 것처럼 자체 해석이 가능한 모델일 뿐만 아니라 신경망에 대한 전역적인 설명으로 사용되기도 합니다.

전역적 설명을 생성하는 또 다른 방법은 다양한 입력에 대해 취해진 지역적 설명을 요약하는 것입니다. LIME의 변형인 SP-LIME은 하위 모듈식 선택을 사용하여 가장 관련성이 높은 지역 LIME의 설명을 요약하여 설명합니다. 또 다른 방법은 시스템에서 전역적 모델을 학습하여 사후 모델을 근사화하는 것으로, 의사결정 집합(decision set) 또는 반사실 규칙의 요약과 같은 것이 있습니다.

6.3. 설명가능성에 대한 적대적 공격

설명의 정확성(원칙 3)은 설명의 중요한 구성요소입니다. 어떤 경우 설명의 정확도가 100%가 아니면, 공격자가 이를 악용하여 작은 입력 섭동에 대한 분류기의 출력을 조작하고 시스템의 편향을 숨길 수 있습니다. 첫째, Lakkaraju and Bastani(2020)는 설명이 폐쇄박스의 예측을 모방할 수 있다고 하더라도 설명의 정확성 면에서 불충분하며 이러한 시스템은 사용자를 오도하는 설명을 생성할 수 있다고 말했습니다. Slack 등(2020)은 오인의 소지가 있는 설명을 생성하는 접근 방식에 대하여 설명했습니다. 이들이 이를 수행한 방식은 모든 입력 데이터 인스턴스에 대하여 분류를 일치시키지만 입력 지점의 작은 섭동에 대해서 출력을 변경하는 스캐폴딩[리버스 엔지니어링과 같은 의미 - 역주]을 특정 분류기 주변에 생성하여, 지역적으로만 쿼리했을 때 전역적 시스템의 동작을 모호하게 만드는 것입니다. 즉, 시스템이 학습 세트 인스턴스와 유사한 인스턴스를 입력으로 제공하는 LIME과 같은 도구에 의해 설명될 것으로 예상되는 경우, 시스템은 대체 프로토콜을 개발하여 학습 및 테스트 세트에서 시도한 분류 방식과 다른 인스턴스를 결정에 사용하려 합니다. 이는 시스템이 분류를 요청받는 시도가 무엇인지 예상하게 하여 설명자를 오도할 수 있습니다. 이와 유사한 또다른 접근 방식을 Aivodji 등(2019)이 입증한 바 있습니다. 이들은 폐쇄박스 모델을 사용하여 모델을 페어워시(fairwash)해서 원래 모델과 비슷하지만 훨씬더 공정하게 해석이 가능한 모델의 앙상블을 생성하여 원래 모델의 불공정성을 숨깁니다. Dimanov 등(2020)은 설명을 조작하기 위해 모델을 약간 교란하는 또 다른 예시를 보여줍니다. Hall 등(2019)이 논의한 설명가능 AI의 취약점 중 하나는 개발자가 폐쇄박스 모델에서 불공정성을 은폐할 가능성입니다. Kindermans 등(2019)의 연구에 따르면 많은 중요도 픽셀 설명에 입력 불변성이 결여되어 있으며, 이는 입력의 작은 변화가 출력이나 관련 픽셀의 속성을 크게 바꿀 수 있음을 의미합니다.

7. 설명가능 AI 알고리즘에 대한 평가

이 절에서는 설명가능 AI 알고리즘을 평가하는 최신 기술을 요약합니다. 이 보고서에서는 설명가능 AI 알고리즘에 대한 평가를 그 평가 원칙에 따라 구분하였습니다. 이 절의 설명성 원칙(원칙 1)은 설명가능 AI 알고리즘의 개요(제6절)에서 다루었고 해당 절에서 최신 설명 기법을 검토했습니다. 이 절에서는 설명의 의미성(원칙 2)과 설명의 정확성(원칙 3)을 측정하는 최신의 기법에 대하여 검토합니다. 우리가 아는 한 알고리즘에 대한 지식의 한계성(원칙 4)을 개발하고 평가하는 작업은 제한적으로 이루어지고 있습니다. 따라서 이 절에서는 지식의 한계성에 대한 평가를 논하지 않습니다.

7.1. 의미성 평가

설명의 의미성을 측정하는 한 가지 방법은 인간의 시뮬레이션 가능성을 측정하는 것입니다. 이는 기본적으로 사람이 기계 학습 모델을 이해하여 모델과 동일한 입력 데이터를 가지고 합리적인 시간 내에 모델 자체에서 예측을 생성할 수 있을 정도로 모델의 매개 변수를 이해할 수 있는 능력입니다.

모델 자체를 시뮬레이션할 수 있는 능력은 높은 수준의 이해도를 반영합니다. 이는 일반적으로 자체 해석가능 모델에서 모델의 복잡성을 측정하기 위한 방법으로 사용됩니다.

여러 연구에서 인간의 시뮬레이션 가능성을 시험했습니다. Lage 등(2019) 및 Lage 등(2019)은 인간의 결과 정확도, 응답 시간을 측정하고 모델 시뮬레이션의 주관적 난이도에 대한 인적 설문조사를 실시했습니다. Hase와 Bansal(2020)은 인간이 주어진 입력에 대해 시스템의 출력을 예측하는 순방향 시뮬레이션과 인간에게 입력과 출력이 주어지는 반사실 시뮬레이션이라는 두 가지 종류의 인간 시뮬레이션 가능성에 대해 설명합니다. 이들은 입력이 특정 방식으로 변경될 경우 시스템이 어떤 출력을 낼지 예측해야 합니다. 설명을 평가할 때는 다양한 설명에 대한 사용자 정확도의 변화를 측정하여 순방향 시뮬레이션과 반사실 시뮬레이션을 평가했습니다. Poursabzi-Sangdeh 등(2019)은 주택 가격에 대한 다양한 로지스틱 회귀 모델을 시뮬레이션하는 사람의 정확도를 측정했습니다. Slack 등(2019)은 '만약의 경우(what-if)' 시뮬레이션 가능성 평가를 수행하였는데 이 경우 사용자가 설명과 함께 입력값을 받습니다. 그 다음 사용자는 주어진 입력에서 약간 교란된 새로운 입력(새 입력은 만약의 경우 또는 반사실에 대한 입력)으로 모델을 시뮬레이션하도록 요청받습니다.

유의미성을 평가하는 또 다른 전략은 인간에게 제공된 시스템의 출력을 입력으로 사용하여 작업을 완료하도록 요청한 다음, 작업에 소요된 시간과 결정의 정확도를 측정하는 것입니다. Poursabzi-Sangdeh 등(2019)이 이를 수행한 방식은 인간에게 주택 가격이 얼마일 것으로 생각하는지 예측하도록 하였을 뿐 아니라 모델이 주택 가격을 어떻게 예측할지도 묻는 것이었습니다(이 단계에서 인간은 모델에 동의하지 않을 수 있습니다). Kim 등(2014)은 사례의 힘을 활용했습니다. 이들의 모델인 베이지안 사례 모델(BCM)은 다양한 요리 레시피의 프로토타입 을 학습했습니다. 인간에게는 프로토타입의 재료만 제공되었고, 각 레시피를 얼마나 잘 분류할 수 있는지 측정했습니다. Lai와 Tan(2018) 은 속임수 탐지 작업에서 의미성을 테스트했습니다. 이 작업은 호텔 리뷰가 진짜인지 거짓인지 판단하는 것이었습니다. 리뷰 자체만 제공되었을 때와 기계의 설명과 함께 제공되었을 때 인간의 속임수 탐지 정확도를 비교했습니다. 이 비교를 통해 기계의 도움/설명을 받았을 때와 받지 않았을 때 인간의 판단 정확도를 비교할 수 있습니다. Lakkaraju 등(2019) 은 인간에게 설명을 보고 결정을 내리도록 하고 그 정확도와 응답 시간을 측정함으로써 의사결정 세트의 다양한 복잡성의 해석가능성을 평가했습니다. Mac Aodha 등(2018)은 인간이 설명을 제공하는 시스템으로 교육받은 경우와 설명을 제공하지 않는 시스템으로 교육받은 경우의 정확도를 비교하는 방식으로 설명을 평가했습니다. Schmidt와 Biessmann (2019) 은 시스템의 설명이 있든 없든 주어진 과제를 완료하도록 사용자를 모집하고 각 사용자의 총 소요 시간과 의사결정 정확도를 측정했습니다. Anderson 등(2020)은 AI 교육을 받지 않은 사람들에게 강화 학습 에이전트의 동작을 설명하는 두 가지 기법을 연구했습니다. 이들은 설명이 없는 경우, 두 가지 설명을 각각 따로 설명하는 경우, 두 가지 설명을 모두 설명하는 경우 등 다양한 설명 조건을 테스트했습니다. 전반적으로 사람이 두 가지 기법, 즉 중요도 맵과 보상-분해 막대(reward-decomposition bars)를 결합했을 때 가장 정확했습니다.

의미는 주관적인 평가로도 측정되었습니다 Hoffman 등(2019)은 좋은 설명에 대한 다양한 기준을

논하고 설명 만족도(Explanation Satisfaction Scale)를 제시했습니다. Holzinger 등(2020)은 설명을 비교하기 위해 시스템 인과성 척도(SCS)를 개발 했습니다. Lage 등(2019)은 인간의 시뮬레이션 가능성을 평가하기 위해 인간에게 모델 시뮬레이션의 난이도를 주관적으로 평가할 것을 요청하기도 했습니다. Rajagopal 등(2021)은 사용자에게 설명의 다양한 속성을 평가하게 요청하는 실험을 실시했습니다.

모델의 크기나 복잡성에 대한 지표를 모델의 해석가능성을 측정하는 척도로 사용하기도 합니다. Lakkaraju 등(2016)은 사용자에게 제공된 정보가 결론을 내리기에 충분한지 물어봄으로써 모델의 해석가능성을 측정했습니다. Poursabzi-Sangdeh 등(2019)은 두 가지 유형의 모델에서 참가자들이 모델의 실제 예측을 더 가깝게 시뮬레이션할 수 있는지를 테스트하여 비교했습니 다. 그들은 '정보 과부하'로 인해 더 투명한 모델보다 더 적은 정보(덜 투명한 모델)가 더 나은 결과를 가져올 수 있다는 것을 발견했습니다. Lage 등(2019) 은 복잡성이 인간의 시뮬레이션 능력에 미치는 영향을 측정했습니다. 이 구상은 복잡성의 수준과 유형이 다른 유형보다 투명성에 어느 정도 영향을 미칠 수 있다고 봅니다. Lakkaraju 등(2019)은 인간에게 의사결정을 요청하고 도움 설명을 제공한 후 이들이 얼마나 빨리, 얼마나 정확하게 의사결정을 내리는지 측정했습니 다. Narayanan 등(2018)은 다양한 유형의 출력 복잡성이 인간의 수행에 어떤 영향을 미치는지 비교했습니다. Bhatt 등(2020)은 '기능 중요도' 설명을 정량화하는 복잡성 지표를 설계했습니다.

7.2. 설명의 정확성 평가

설명의 정확성은 '충실도' 작업과 밀접한 관련이 있습니다. 여러 연구에서 설명의 충실도를 평가했습니다. 이를 테스트하는 한 가지 방법은 시스템 출력을 실측 자료(ground truth)로 삼고 그 사후 설명을 머신러닝 지표를 사용해 평가하는 방식으로 모델을 시뮬레이션하는 것입니다. Lakkaraju 등(2019)은 이 전략을 따르면서도 각 인스턴스에 최대 하나의 설명이 있고 모든 인스턴스가 사후 설명 모델에 의해 설명되는지 검토했습니다. Mohseni 등(2020)이 제안한 두 번째 방법은 인간이 설명을 평가하고 설명의 정확성을 평가하기 위해 "온전성 검사(sanity checks) "을 실시하는 것입니다. 세 번째 방법은 시스템에 다양한 입력을 설명하도록 요청하는 것입니다. 대부분의 경우 입력은 적응형입니다. 새로운 입력은 제공된 설명에 따라 이전 입력의 약간 변경된 버전입니다. 그런 다음 실험으로 입력의 변화에 따른 출력의 변화와 설명이 변경된 기능의 중요도를 측정합니다. Samek 등(2017)은 중요도 픽셀을 사용하여 설명의 정확도를 평가했습니다. 이들은 가장 중요한 픽셀을 점차적으로 삭제하고 분류 점수가 얼마나 변화하는지 측정했습니다. 이 구상은 중요한 픽셀이 결정 정확도에 더 많은 영향을 미친다면, 픽셀이 삭제될수록 시스템이 이미지를 원래 클래스로 분류할 가능성이 낮아진다는 것입니다. Hooker 등(2019) 은 중요한 특징이 제거되었을 때 시스템의 성능이 저하되는지 테스트했습니다. 이들은 중요한 픽셀을 제거한 다음 시스템을 재학습시키고 재학습된 시스템의 결정 정확도를 측정하는 전략을 적용했습니다. Yeh 등(2019)은 '비충실성 측정법'을 개발하여 설명의 정확성을 평가했습니다. Alvarez Melis와 Jaakkola(2018)는 모델의 고순위 특징을 제거하고 분류 확률의 하락을 측정하여 설명의 정확성, 즉 충실도를 평가했습니다. 또한 입력에 화이트 노이즈를 추가하고 설명이 얼마나 변화하는지 측정하는 방식으로 설명의 정확성을 측정했습니다.

Adebayo 등(2018)은 학습된 모델의 변화에 따른 설명의 변화량을 측정하여 심층 신경망에 대한 중요도 픽셀 설명의 정확성을 평가했습니다. Sixt 등(2019)은 중간 컨볼루션 레이어를 무작위로 추출하고 중요도 픽셀을 비교하여 중요도 픽셀의 품질을 평가했습니다. 또한 레이블이 실제 레이블일 때와 무작위 레이블일 때의 중요도 픽셀을 비교했습니다. Qi 등(2020)은 설명과 관련이 있다고 판단되 는 이미지 픽셀을 추가하거나 삭제하는 방식으로 설명의 정확성을 평가했습니다. 그런 다음 새로운 이미지에 대한 시스템의 점수를 비교했습니다. Bhatt 등(2020)은 유사한 입력은 유사한 특징 중요도를 설명한다는 의미의 민감도와 설명의 변화는 입력의 변화와 상관관계가 있어야 한다는 의미의 충실도를 모두 확인하는 방식으로 '특징 중요도' 설명의 정확성을 평가했습니다.

반사실 설명의 품질은 Wachter 등(2017)이 검토하였습니다. 반사실 설명은 "시스템이 해당 입력에 따른 결정을 변경하기 위해 [기존] 입력에서 변경되어야 하는 최소량은 얼마인가?"라는 질문에 답해야 합니다. 따라서 그들은 반사실이 원래 데이터 포인트에서 얼마나 멀리 떨어져 있는지를 검토했습니다.

8. 설명가능 AI에 대한 비교 집단으로서 인간

인간과 AI시스템의 성능을 고려할 때 성능 기대치에 대한 의견 간에는 상당히 큰 차이가 있습니다. 어떤 사람들은 기계에 인간보다 훨씬 더 높은 기준을 적용해야 한다고 주장하는 반면, 다른 사람들은 기계는 단지 인간만큼만 잘하는 것으로 충분하다고 생각합니다. 기계가 인간보다 얼마나 더 뛰어나야 하는가라는 질문처럼 철학적으로 중요한 차이에서 유래한 흥미롭고 어려운 질문들이 쏟아집니다. 어떤 면에서 더 나아야 할까요? "인간만큼"을 어떻게 측정할 수 있을까요? 이러한 철학적 논쟁에서 어느 쪽에 속하든, 인간의 능력을 기준으로 삼는 일이 도움이 됩니다. 이 절에서는 인간의 의사결정이 4가지 원칙에 어느 정도 부합하는지에 대해 설명합니다.

인공지능과 별개로, 혼자서 일하는 인간도 설명할 수 있기를 기대하며 중대한 결정을 내립니다. 예를 들어 의사, 판사, 변호사, 법의학자는 종종 자신의 판단에 대한 근거를 제시해야 하는 경우가 많습니다. 이렇게 제시된 설명은 어떻게 4가지 원칙에 부합할까요? 우리는 외부 사건(예: "왜 하늘이 파란색인가?" 또는 "왜 어떤 사건이 발생했나?")이 아닌, 인간 자신의 판단과 결정(예: "왜 이런 결론이나 선택에 도달했나?")에 대한 설명에 엄격하게 초점을 맞췄습니다. 설명이 수반되는 외부 사건은 인간의 추론과 예측을 공식화하는 데 도움이 될 수 있습니다. 이는 설명가능 AI에 대한 열망과 일치합니다. 그러나 다음에 설명하는 바와 같이 인간이 스스로 판단, 결정, 결론을 내리기 위해 만든 설명은 대체로 신뢰할 수 없습니다. 설명가능 AI의 비교 집단으로서 인간은 설명가능 AI시스템의 벤치마크 지표 개발에 정보를 제공하고, 인간과 기계의 협업 동학에 대한 이해를 높일 수 있습니다.

8.1. 설명성

이 원칙은 어떤 시스템이 설명가능한 것으로 간주되려면 그 시스템이 설명을 제공해야 한다고 말합니다. 이 절에서는 인간이 자신의 판단과 결정에 대해 설명을 생성하는지 여부와 그렇게 하는 것이 의사결정자에게 유익한지 여부에 초점을 맞출 것입니다. 8.2절에서는 인간의 설명이 의미가 있는지에 대해, 8.3절에서는 이러한 설명의 정확성에 대해 논의할 것입니다.

인간은 다양한 유형의 설명을 생성할 수 있습니다. 그러나 언어적 설명은 의사결정과 추론 과정을 방해할 수 있습니다. 전문 지식이 쌓일수록 기본 과정은 의식적으로 인식할 수 없는 자동화된 과정이 되므로 구두로 설명하기가 더 어려워지는 것으로 생각됩니다. 이는 인공지능 자체에도 비슷한 긴장을 낳는데, 높은 정확성에 대한 욕구가 설명가능성의 감소를 동반하곤 하는 것으로 생각됩니다.

보다 일반적으로 말하자면, 의식적 인식이 제한적인 상태에서 발생하는 과정들은 결정 자체를 명시적으로 표현하도록 요구하였을 때 해를 입을 수 있습니다. 그 예시로 거짓말 탐지를 들 수 있습니다. 사람이 진실을 말하는지 거짓을 말하는지를 명시적으로 판단하는 거짓말 탐지는 일반적으로 부정확합니다. 그러나 거짓말 탐지 정확도는 명시적 판단을 우회하여 암묵적 범주화 작업을 통해 판단을 내렸을 때 향상될 수 있었습니다. 이는 거짓말 탐지가 무의식적인 과정일 수 있으며, 의식적 판단을 강요할 때 방해받을 수 있음을 시사합니다.

이러한 연구 결과들을 종합하면, 인간의 일부 평가는 명시적인 판단이나 설명이 필요한 경우보다 자동적이고 암묵적으로 남겨두는 것이 더 효과적일 수 있음을 시사합니다. 인간의 판단과 의사결정은 종종 폐쇄박스처럼 작동하며, 이 폐쇄박스 과정에 간섭하면 결정의 정확성에 해를 끼칠 수 있습니다.

8.2. 의미성

시스템이 이 원칙을 충족하기 위해서는 의도된 목표 청중이 이해할 수 있고 이해하기 쉬운 설명을 제공해야 합니다. 이를 위해 우리는 다른 사람이 어떻게 결론에 도달했는지를 추론하는 인간의 능력에 초점을 맞췄습니다. 이러한 고려는 여기서 다음과 같은 의미입니다: 1) 청중이 설명을 제공하는 사람이 의도한 것과 동일한 결론에 도달하는지 여부, 2) 설명을 기반했을 때 청중이 결론에 대해 서로 동의하는지 여부입니다.

설명가능 AI에서 인간과 인간의 상호작용과 유사한 사례로는 법의학자가 일반인(예: 배심원단)에게 법의학 증거를 설명하는 경우를 들 수 있습니다. 현재 법의학자가 결과를 보고하는 방식과 일반인이 그 결과를 이해하는 방식 사이에 차이가 있습니다. 잭슨 등(2015)은 배심원에게 제시되는 증거의 유형과 배심원이 그 증거를 이해하는 능력에 대해 광범위하게 연구했습니다. 그들은 법의학자들의 설명 대부분이 오해의 소지가 있거나 혼동을 일으키기 쉽다는 사실을 발견했습니다. 따라서 이러한 설명은 "의미성"이라는 내부 기준을 충족하지 못합니다. 이 분야의 과제는 설명의 개선 방법을 배우고, 제안된 해결책이 항상

일관된 결과를 가져오지 않는다는 것을 이해하는 것입니다.

사람들이 당면한 질문에 따라 서로 다른 설명 유형을 기대하거나, 의견 형성을 주도하는 맥락이 무엇인지, 만족스러운 설명으로 간주되는 바의 개인차 등, 다른 사람에게 의미 있는 설명을 제공하는 것은 복잡한 문제를 포함합니다. 따라서 무엇이 의미 있는 것으로 간주되는지는 상황과 사람에 따라 다릅니다.

8.3. 설명의 정확성

이 원칙은 시스템의 설명이 특정 결과물을 생성하는 이유나 그 과정을 정확하게 반영하는 것을 말합니다. 인간의 경우 이는 의사결정 과정에 대한 설명이 그 결정 이면에 있는 정신적 과정을 진정으로 반영하는 것과 유사합니다. 이 절에서는 이러한 측면에만 초점을 맞추었습니다. 설명의 품질이나 일관성에 대한 평가는 이 원칙의 범위를 벗어납니다.

설명의 정확성과 관련된 내적 성찰의 유형에 있어서, 사람들이 의사결정에 대한 자신의 이유를 제시하더라도 이것이 정확하거나 의미 있는 내성을 반영하지는 않는다는 사실이 잘 알려져 있습니다. 이를 '내성적 착각'이라고 부르는데, 이 용어는 자신의 내면을 들여다봄으로써 얻은 정보가 가치가 있다는 잘못된 생각에 근거하였을 경우를 나타냅니다. 사람들은 개인적 의견과 같이 불변하는 결정의 이유조차 도 조작합니다. 사실, 말로 표현할 수 있는 사람들의 의식적인 추론이 항상 의사결정 표현 이전에 일어나는 것은 아닙니다. 그보다는 사람들이 결정을 내린 다음 사후에 그 결정에 대한 이유를 적용한다는 증거가 있습니다. 신경과학적 관점에서 볼 때, 결정의 신경적 표식은 사람이 의식적으로 인식하기 최대 10초 전에 발생할 수 있습니다. 이 발견은 의사결정 과정이 우리가 의식적으로 인식하기 훨씬 전에 시작된다는 사실을 시사합니다.

사람들은 대부분 자신이 정확하게 성찰할 수 없다는 것을 인식하지 못합니다. 이는 사람들이 자신의 이전 결정을 정확하게 기억하지 못하는 '선택적 실명'에 대한 연구를 통해 입증되었습니다. 이러한 부정확한 기억에도 불구하고 참가자들은 실제로 한 적이 없는 선택의 이유를 제시합니다. 장기기억을 사용하지 않는 연구의 경우, 참가자들은 지각적 판단을 평가하는 방식에 대해서도 인식하지 못하는 것으로 나타났습니다. 예를 들어, 사람들은 누군가의 신원을 파악하기 위해 어떤 얼굴 특징을 사용했는지 보고할 때 부정확하게 말합니다.

설명 정확도에 대한 정의에 따랐을 때, 위와 같은 결과는 인간이 이 기준을 안정적으로 충족한다는 생각을 뒷받침하지 않습니다. 알고리즘의 경우와 마찬가지로 인간의 의사결정 정확도와 설명 정확도는 별개입니다. 수많은 작업에서 인간은 매우 정확할 수 있지만 의사결정 과정을 말로 표현할 수는 없습니다.

8.4. 지식의 한계성

이 원칙은 시스템이 1) 설계된 조건 하에서 2) 출력이나 조치가 충분한 신뢰 수준에 도달했을 때만 작동한다는 것을 말합니다. 이 원칙에 있어 우리는 메타인지, 즉 자신의 생각에 대한 사고라는 광범위한 영역을 좁혔습니다.

여기서는 인간이 자신의 능력과 정확성을 올바르게 평가하는지, 답을 모른다고 언제 보고해야 할지 아는지에 대해 초점을 맞췄습니다. 사람들이 자신의 지식 한계를 평가할 수 있는지 테스트하는 방법에는 여러 가지가 있습니다. 한 가지 방법은 참가자에게 다른 사람과 비교하여 자신이 과제를 얼마나 잘 수행했거나 수행할 것이라고 생각하는지 예측하도록 요청하는 것입니다(예: 다른 과제 수행자와 비교하여 점수가 몇 백분위수에 속할지). 지식 한계 인식을 테스트하는 또 다른 방법은 응답 신뢰도를 측정하는 것으로, 신뢰도가 높을수록 자신이 맞을 가능성이 높다고 믿는다는 것을 나타냅니다.

잘 알려진 더닝-크루거 효과에서 알 수 있듯이, 대부분의 사람들은 다른 사람에 대한 자신의 능력을 부정확하게 평가합니다. 전문가를 포함한 사람들은 일반적으로 자신의 능력을 명시적으로 추정하라는 요청을 받았을 때 자신의 정확도와 능력을 잘 예측하지 못한다는 것도 비슷한 연구결과입니다. 그러나 최근 얼굴 인식에 대한 더닝-크루거 효과를 재현한 연구에 따르면 사람들은 자신의 정확도를 정확하게 예측하지 못했지만, 능력 추정치는 과제 난이도에 따라 달라지는 것으로 나타났습니다. 이는 정확한 값(예: 다른 사람에 대한 수행 예측 백분위수 또는 예측 정답률)은 틀릴 수 있지만, 사람들은 예측 수행의 방향을 적절히 조절할 수 있음을 시사합니다(예: 과제 난이도가 자신에게 어느 정도 어려운지 아는 등).

다양한 판단과 결정의 경우에서, 사람들은 피드백이 없는 경우에도 자신이 언제 오류를 범했는지 알고 있는 경우가 많습니다. 목격자의 증언을 예로 들면, 확신과 정확성 사이에 약한 관련성이 있는 것으로 여러 차례 밝혀졌지만, 심문 과정과 사건-기억 시점 간 길어지는 시간 안에 '오염'이 없는 경우 사람의 확신은 정확성을 예측할 수 있습니다. 따라서 지식의 한계를 평가하는 데 있어 인간의 단점은 설명 자체를 생산할 때와 유사합니다. 명시적으로 설명을 요구할 경우, 이러한 설명은 전문 지식을 통해 얻은 자동화된 과정을 방해할 수 있으며, 실제 인지 과정을 정확하게 반영하지 못하는 경우가 많습니다. 마찬가지로, 이 절에서 설명한 것처럼 사람들이 다른 사람과 비교하여 자신의 능력 수준을 명시적으로 예측하거나 추정하도록 요청받으면 부정확한 경우가 많습니다. 그러나 이러한 명시적 판단에 비해 주어진 결정에 대한 확신을 평가하도록 요청하면 사람들은 우연 이상의 수준으로 자신의 정확도를 측정할 수 있습니다. 이는 사람들이 자신의 지식 한계에 대한 통찰력을 가지고 있음을 시사하지만, 이러한 통찰력은 경우에 따라 제한적이거나 약할 수 있습니다.

9. 토론 및 결론

우리는 설명가능 AI시스템의 기본 요소를 포착하기 위해 4가지 원칙을 도입했습니다. 이 원칙은 설명가능한 시스템의 다양한 구성요소를 다룰 수 있는 프레임워크를 제공합니다. 이 4가지 원칙은 시스템이 설명을 생성하고, 그 설명이 인간에게 의미가 있으며, 설명이 시스템의 프로세스를 정확하게 반영하고, 시스템은 그 지식의 한계를 표현해야 한다는 것입니다. 이 원칙은 시스템이 4가지를 모두 따를 때 힘을 발휘합니다. 설명을 제공하지만 이해할 수 없거나 정확하지 않거나 지식의 한계를 벗어나는 시스템은 가치가 떨어집니다. 시스템 결과에 대한 사용자의 수용도에도 사실상 영향을 미칠 수 있습니다.

설명가능 AI를 개발하고 평가하는 데는 다양한 접근 방식과 철학이 있습니다. 컴퓨터 과학적 접근법은 다양한 계산 방식 및 그래픽 기술과 관점에서 설명가능 AI의 문제를 다루며, 이는 유망한 돌파구로 이어질 수 있습니다. AI 설명의 효과와 그 효과의 배후에 있는 인간적 요소를 고려하면서 인간을 최전선에 두는 분야가 꽃을 피우고 있습니다. 저희의 4가지 원칙은 이러한 인간과 기계의 상호작용을 탐구할 수 있는 다학제적 프레임워크를 제공합니다. 시스템의 실질적인 요구가 이러한 원칙을 어떻게 다루거나 무시할 것인지에 영향을 미칠 것입니다. AI 커뮤니티는 이러한 요구들을 염두에 두고 4가지 원칙을 조정하고 적용하면서 궁극적으로 광범위한 적용 범위를 포착해낼 것입니다.

설명가능 AI의 초점은 양질의 설명을 제공할 수 있는 시스템의 역량을 향상시키는 것이었습니다. 여기서는 인간이 AI를 위해 설정한 것과 동일한 원칙을 충족할 수 있는지에 대해 다루었습니다. 우리는 인간이 여기 제시된 원칙을 충족하는 데 있어 제한적인 역량만 발휘한다는 사실을 보여주었습니다. 이는 AI시스템을 비교할 수 있는 벤치마크를 제시합니다. 최근의 규제는 사회적 기대를 반영하여 설명가능 AI에 대한 요건을 부과하여 AI시스템에 어느 정도의 책임성을 부과하고 있습니다. 설명가능 AI가 발전함에 따라 AI시스템의 특정 부분이 인간에 비해 사회적 기대와 목표를 더 잘 충족하는 것을 볼 수도 있습니다. 인간과 기계의 협업에서 AI시스템과 인간 모두의 설명가능성을 이해함으로써 각각의 강점을 통합한 구현을 추구할 수 있는 길이 열리고, 인간이나 AI시스템 중 한쪽만의 능력을 넘어서는 설명가능성을 잠재적으로 향상시킬 수 있습니다.

이 보고서에서는 설명가능한 의사결정과 관련된 인간적 요소에 제한적으로만 초점을 맞추었습니다. 인간과 설명가능한 기계 간의 상호작용에 관해서는 많은 것을 더 배우고 연구해야 합니다. 이 보고서의 범위를 벗어나지만, AI와 인간 사이의 인터페이스를 고려할 때 인간의 인식과 의사결정을 이끄는 일반적인 원리를 이해하는 것은 설명가능 AI 분야에 매우 유익한 정보가 될 수 있습니다. 인간의 경우 일반적으로 더 간단하고 일반적인 설명을 선호하는 경향이 있습니다. 그러나 앞서 설명한 것처럼 어떤 설명이 높은 품질로 간주되는지는 개인차가 있습니다. 의사결정의 맥락과 의사결정의 유형도 이에 영향을 미칠 수 있습니다. 인간은 다른 요소와 분리하여 결정을 내리지 않습니다. 의식적으로 인지하지 않더라도 사람들은 첫인상, 성격적 특성 판단, 배심원단 결정 등 의사결정에 관련 없는 다양한 정보를 통합합니다. 동일한 정보가 제공되더라도 맥락, 개인의 편견, 정보가 제시되는 방식이 의사결정에 영향을 미칩니다. 설명가능 AI의 맥락에서 이러한 인간적 요소를 고려하는 일은 이제 막 시작되었습니다.

설명가능 AI가 성공하기 위해서는 AI 커뮤니티가 인간과 인공지능 시스템 간의 인터페이스를 연구해야 합니다. 인간과 기계의 협업은 정확도 측면에서 매우 효과적인 것으로 나타났습니다. AI 설명가능성에 대해서도 인간과 기계의 협업에 유사한 돌파구가 있을 수 있습니다. 여기에 정의된 원칙은 시스템 출력에 대하여 사용자가 더 깊이 이해할 수 있도록 함으로써 설명가능 AI가 더 안전한 세상을 이끌 수 있는 지침과

철학을 제공합니다. 의미 있고 정확한 설명은 사용자가 이 정보를 적용하여 자신의 행동을 조정하거나 결정에 이의를 제기할 수 있는 역량을 강화하게끔 합니다. 개발자와 감사 기관은 설명을 통해 시스템을 적절하게 개선, 유지 관리 및 배치할 수 있는 능력을 갖추게 됩니다. 설명가능 AI는 복잡한 AI시스템의 여러 측면을 안전하게 작동하고 신뢰하게 하는 데 기여합니다. 4가지 원칙이라는 공통 프레임워크와 정의는 복잡한 현실의 시스템에 필요한 설명가능 AI 방법론의 진화를 촉진합니다.

[참고문헌 생략]