TechDispatch

설명가능한 인공지능

설명가능한 인공지능이란 무엇이고 어떻게 구현할 수 있을까요? 2023 년 11 월 EDPS(유럽 개인정보보호 감독관)는 설명가능한 인공지능에 대하여 이해하기 쉽게 설명하는 자료를 펴냈습니다.
인공지능이 개인정보보호법에서 요구하는 투명성을 준수하기 위해서는 설명가능성이 구현될 필요가 있습니다. 이 자료는 인공지능의 투명성, 해석가능성, 설명가능성의 개념을 설명하고 현재의 기술 수준에서 설명이 가능한 방식(화이트박스)과 불가능한 방식(블랙박스)을 구분하여 소개한 후 인공지능 의사결정에서 여전히 인간의 개입이 중요하다는 사실을 강조하였습니다.

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

EDPS
European Data Protection Supervisor

HTML	ISBN 978-92-9242-715-3	ISSN 2599-932X	doi: 10.2804/132319	QT-AD-23-002-EN-Q
PDF	ISBN 978-92-9242-716-0	ISSN 2599-932X	doi: 10.2804/802043	QT-AD-23-002-EN-N

"블랙박스" 효과

보건의료, 금융, 교통, 제조업, 엔터테인먼트 등의 분야에서 인공지능(AI)의 도입이 빠르게 증가하고 있습니다.
최근 몇 년간 AI 의 인기가 높아진 것은 대량의 정보 처리나 패턴 식별 등에서 작업을 자동화할 수 있다는 점, 그리고 대중적으로 널리 사용할 수 있다는 점 덕분입니다.[1]
최근 몇 년간 큰 인기를 끈 AI 의 두 가지 사례로는 *ChatGPT 와* 같은 대규모 언어 모델[2](LLM)이나 Stable *Diffusion 과* 같은 텍스트-이미지 모델[3]이 있습니다.[4]

그러나 AI 의 사용이 증가하고 있음에도 불구하고 이러한 시스템 중 상당수는 AI 시스템을 제공하는 자('제공자'), AI 시스템을 배치하는 자('배치자'), 그리고 AI 시스템의 사용에서 영향을 받는 사람 모두에게 불투명한 방식으로 운영되고 있습니다. 복잡한 AI 시스템 분야에서 이들 시스템의 제공자조차도 자신이 구축한 시스템의 결정과 결과를 설명할 수 없는 경우가 많습니다.

이 현상을 일반적으로 **"블랙박스" 효과** 라고 칭합니다.

1. 불투명한 AI 시스템의 위험

머신러닝(ML) 또는 딥러닝(DL)과 같은 AI 시스템은 인간의 명확한 프로그래밍이 아니라 자체적인 학습 과정 [5]을 통해 학습한 알고리즘을 사용합니다.

AI 모델은 학습 과정을 통하여 특정 입력 요소(예: 임상 증상) 사이에서 새로운 상관관계를 발견할 수 있으며, 수많은(수백만 개) 매개변수가 상호 작용하며 관여하는 매우 복잡한 모델에 기반해서 의사 결정 또는 예측(예: 의료 진단)을 내릴 수 있습니다. 이는 AI 전문가조차도 그 결과물이 어떻게 생성되는지 이해하기 어렵게 만듭니다(Peters, 2023).

이러한 상황에서 시스템이 어떠한 결정을 내린 이유는 시스템 사용자와 시스템의 영향을 받는 사람 모두에게 불분명할 수 있습니다. 그 결과 '블랙박스' 효과는 AI 시스템에 대해 잘못된 신뢰 또는 과도한 의존을 초래할 수 있으며, 두 경우 모두 개인에게 부정적인 결과를 초래할 수 있습니다.

오늘날의 사회에서는 사용자가 특정 기술을 사용하기 위해 해당 기술의 작동 원리를 이해할 필요가 없으며, 특정 기술의 작동 방식을 완전히 이해하지 못하는 경우가 많다고 주장할 수도 있습니다. 예를 들어, 자동 변속기의 작동 원리를 실제로 설명할 수 있는 운전자가 얼마나 될까요? 이러한 상황에서 AI 기술이 공공 기관을 비롯하여 자동화된 의사 결정(또는 의사 결정 지원)을 위해 구현되는 경우가 많다는 사실을 염두에 두어야 합니다. 이 경우 대부분 투명성과 책무성이 필수적인 법적 요구사항입니다.

따라서 여기에 AI 결정의 기반이 되는 로직을 감추는 '블랙박스' 효과는 허용되지 않습니다.

한편 기술적 관점에서 자동차 자동 변속기의 경우를 생각해 보았을 때 또 다른 차이점이 있습니다. 자동차가 작동하는 경우와 달리 AI 엔지니어는 표면 아래에서 무슨 일이 일어나고 있는지 완전히 이해하지 못할 수도 있다는 점입니다.

이러한 불투명성은 AI 의 결정을 이해하기 어렵게 만들 뿐 아니라 편향성,[6] 부정확성 또는 소위 '환각' 등 AI 시스템의 결함을 숨길 수 있기 때문에 개인에게 직접적인 영향을 미칠 수도 있습니다.[7]

불충분하게 설계, 개발 또는 테스트된 알고리즘은 잠재적으로 차별적이거나 개인에게 해로운 결과를 초래할 수 있습니다.

예를 들어 AI 를 사용하여 입사 지원자를 선발하는 경우, 편향된 학습 데이터로 인하여 시스템이 의도치 않게 특정 인구 집단이나 배경을 가진 지원자를 선호할 수 있습니다. 시스템이 '블랙박스'인 경우 특정 지원자가 탈락하거나 선발된 이유를 파악하기 어려워 편향을 식별하고 해결하는 것이 더 어려워질 수 있습니다.[8]

또다른 예로 의료 진단 AI 모델을 사용하는 경우, 편향된 학습 데이터로 인하여 특정 인구 집단의 특정 상태를 편향적으로 놓치거나 오진할 수 있습니다. 모델이 '블랙박스'인 경우 의료 전문가가 의사 결정의 근거를 이해하기 어려워져서 편향 우려를 해결할 수 있는 가능성마저 저해합니다.[9]

차별적인 결과만 '블랙박스'의 문제인 것이 아닙니다. 투명성 자체가 부족하면 자동화된 결정의 영향을 받는 사람들이 기본적인 로직과 그 잠재적 영향력을 이해하는 것 또한 방해받을 수 있습니다. 예를 들어 신용 평가에 사용되는 AI 모델의 경우, 은행 고객이 자신의 금융 생활에 영향을 미치는 자동화된 결정에 대하여 아무런 지식을 얻지 못할 수 있습니다.[10]

더 중요한 점은 정부가 사용하는 자동화된 의사결정 시스템으로 인하여 개인이 영향을 받을 수 있는데, 그 운영이나 기능이 기존 법률에 완전히 명확하게 잘 정의되어 있지 않을 수 있다는 것입니다.[11]

2. 설명가능한 인공지능이란 무엇인가요?

설명가능한 인공지능(XAI)은 AI 시스템의 조치와 결정에 대하여 명확하고 이해할 수 있는 설명을 제공하는 AI 시스템의 기능입니다. 이 기술의 핵심 목표는 의사 결정 프로세스의 기반이 되는 메커니즘을 설명함으로써 인간이 이러한 시스템의 동작을 이해할 수 있도록 하는 것입니다.

그러나 설명가능성을 개선하려는 많은 노력은 목표 사용자의 요구를 효과적으로 해결하기보다는 주로

AI 연구자 자신에게 맞춤한 설명으로 이어지는 경우가 많습니다. 이는 복잡한 의사 결정 모델에 대하여 만족스러운 설명을 정의하는 책임을 이러한 모델을 자세히 이해하고 있는 AI 전문가에게 맡기는 것입니다(Miller T. H., 2017).

이상적으로는 XAI 가 시스템의 역량과 이해한 바를 설명할 수 있어야 하고, 과거 조치, 진행 중인 프로세스 및 향후 단계를 설명할 수 있어야 하며, 조치의 기반이 되는 관련 정보를 공개할 수 있는 기능을 포함해야 합니다(Gunning, 2019).

투명성, 해석가능성 및 설명가능성

AI 의 맥락에서 투명성, 해석가능성 및 설명가능성의 개념은 공식적인 정의를 가지고 있지 않으며 때때로 같은 의미로 사용되기도 합니다.
본 문서에서는 이를 서로 다른 개념으로 다음과 같이 해석합니다.

- **투명성이란** 특정 모델을 이해할 수 있는 능력을 말합니다. 가장 엄격한 의미에서는 한 **사람이 전체 모델을 한 번에 파악할 수 있을 때 그 모델이 투명하다고 할 수 있습니다**. 투명성은 전체 모델의 수준, 개별적 구성 요소(예: 매개변수)의 수준, 특정 학습 알고리즘의 수준에서 검토될 수 있습니다. 두 번째, 덜 엄격한 투명성의 개념으로는 모델의 각 부분(예: 각 입력, 매개변수, 계산)이 직관적인 설명을 허용하는 것을 의미할 수 있습니다(Lepri, 2018).

투명한 AI 시스템은 이해관계자가 의사결정 프로세스를 검증 및 감사할 수 있도록 허용함으로써 책무성을 확보할 수 있게끔 하고, 편향이나 불공정성을 감지하며, 시스템이 윤리 기준 및 법적 요구사항에 맞추어 운영될 수 있도록 보장합니다.

- **해석가능성은 주어진 "블랙박스" 모델 또는 결정에 대한 인간의 이해 정도를** 말합니다(Lisboa, 2013)(Miller T. H., 2017). 해석가능성이 불충분한 모델은 "결과 결정이 제시되었을 때 입력된 정보로부터 특정 분류에 도달한 방법이나 이유를 구체적으로 알 수 없다는 의미에서 불투명"합니다(Burrell, 2016).

해석가능한 AI 모델을 통해 인간은 입력이 주어졌을 때 모델이 무엇을 예측할지 예상하고 모델이 언제 실수하였는지 이해할 수 있습니다.

- AI 의 **설명가능성은 특정 모델 예측이나 결정에 대하여 명확하고 일관된 설명을** 제공하는 데 중점을 둡니다. 특정 결과에 대하여 인간이 이해할 수 있는 정당성이나 이유를 제공함으로써 "AI 시스템이 왜 이러한 특정한 예측을 하였는가?"와 같은 질문에 답하는 것을 목표로 합니다. 설명가능성은 해석가능성을 기반으로 하지만 인간과 컴퓨터의 상호 작용, 법률, 윤리 등 다른 분야 및 영역도 고려하는 것입니다(Thampi, 2002).

설명가능성은 사용자, 규제 기관 및 이해관계자가 AI 가 생성한 결과의 근거를 이해하는 데 도움이 되므로 사람의 생명이나 민감한 정보가 걸려 있는 주요 애플리케이션에서 특히 중요합니다.

설명가능성은 AI 시스템에 대한 신뢰를 구축하는 데 중요합니다. 그러나 시스템이 충분히 해석가능한 경우에는 설명이 필요하지 않을 수도 있습니다. 이는 (덜 복잡한) 특정 유형의 AI 에서 더 쉽게 달성될 수 있습니다.

예를 들어 규칙 기반 시스템 또는 전문가 시스템은 규칙과 특정한 전문 지식을 사용하여 조언이나 진단을 제공하는 AI 의 하위 집합으로, 일반적으로 의료, 물류 및 금융 분야에서 사용됩니다. 이러한 시스템이 충분한 투명성을 제공하고 적절한 전문 지식을 갖춘 전문가가 이를 해석해 주는 경우 사용자를 위한 설명가능성 메커니즘을 [별도로] 구현할 필요성이 없는 경우도 있습니다.

그러나 해당 분야의 전문가가 아닌 사람이 시스템을 완전히 이해하려면 XAI 가 필요할 수 있습니다.

3. 설명가능한 AI 에 도달할 수 있는 방식

AI 설명가능성에 도달할 수 있는 방식은 두 가지 범주로 나눌 수 있습니다. 한 가지는 해석가능성이 시스템 설계에 내장되어 있어 자체적으로 해석가능한 모델이고, 다른 한 가지는 시스템의 동작을 먼저 관찰한 후에 설명하는 사후 설명 모델입니다.

자체적으로 해석가능한(또는 "화이트박스") 모델은 데이터 입력이 출력 또는 대상 변수에 어떤 영향을 미치는지 이해하기 쉽게 보여주는 알고리즘을 특징으로 합니다. 반면에 "블랙박스" 모델은 그 자체로는 설명할 수 없습니다. 이러한 설명가능성 부족은 시스템 설계자가 의도적으로 난독화하였거나(Xu, 2018) 모델이 복잡하기 때문에 발생할 수 있습니다.

"화이트 박스" 접근 방식: 자체적으로 해석가능한 모델

'화이트박스' 모델에서는 사용된 알고리즘을 이해하기 쉽고, 입력된 특징이 출력 또는 목표 변수로 변환되는 방식을 해석할 수 있습니다. 목표 변수를 예측하는 데 가장 중요한 특징을 파악할 수 있으며, 이들 특징을 이해할 수 있습니다(Thampi, 2002).
해석가능성은 전체 모델, 개별적인 구성 요소(예: 입력 매개변수) 또는 특정한 학습 알고리즘의 수준 등 다양한 수준에서 제공될 수 있습니다.

'화이트 박스' 모델의 두 가지 예시로는 의사결정 트리와 선형 회귀를 들 수 있습니다.[12]

의사결정 트리 모델의 예로는 수신 이메일이 스팸인지 아닌지를 자동으로 판단하는 이메일 분류 시스템을 들 수 있습니다. 이 모델은 먼저 '스팸' 또는 '스팸 아님'으로 분류된 이메일 데이터 세트에 대해 학습하고, 데이터를 그 특징에 따라 재귀적으로 분할하여 마치 나무와 같은 구조를 만듭니다. 각 노드에서 트리는 이메일 분류 측면에서 가장 많은 정보를 얻는 기능을 선택합니다. 그 결과 의사결정 트리는 순서도와 같은 구조로 시각화할 수 있습니다. 각 노드는 조건(예: "이메일에 '무료'라는 단어가 포함되어 있는가?")을 나타내고 각 가지는 해당 조건에 따라 가능한 결과를 나타냅니다. 트리의 잎은 최종 분류("스팸" 또는 "스팸 아님")를 나타냅니다.

그러나 특정 유형의 AI 는 내재적인 복잡성과 해석가능성 부족으로 인해 일정한 어려움을 겪습니다. 보다 복잡한 아키텍처의 예시로는 여러 레이어가 상호 연결된 인공 신경으로 구성된 신경망 아키텍처를 들 수 있습니다. 여기서 각 레이어는 연산을 수행하고 다음 레이어로 신호를 전달합니다. 복잡한 아키텍처의

또다른 예시로는 3 개 이상의 레이어로 구성된 신경망인 딥러닝 알고리즘이 있습니다. 이 경우 대부분 모델 내부를 설명하는 표현이 모델 자체에 대해서 만큼이나 이해하기 어려울 수 있습니다(Lipton, 2018).

이러한 사실은 모델이 항상 자체적으로 해석할 수 있기를 기대하는 것이 비현실적이라는 점을 시사합니다. 따라서 복잡한 시스템에는 사후 접근 방식이 더 적합해 보입니다.

"블랙박스" 접근 방식: 사후 설명

사후 접근 방식에서는 모델의 결정이 내려진 후 설명이 생성되며, 이는 전역적(global) 또는 지역적(local)인 것으로 구분할 수 있습니다.

전역적 설명은 AI 모델의 동작과 의사 결정 프로세스에 대한 전반적인 지식을 제공하며, 모델의 동작에 광범위하게 적용되는 패턴, 일반 추세 및 통찰성을 포착하는 것을 목표로 합니다(예: 시스템이 채용 공고에 가장 적합한 후보자를 선택하는 방법은 무엇인가요?).

전역적 설명 기법의 예시로는 '특징 중요도'(Breiman, 2001)가 있습니다. 이는 모델의 의사 결정 프로세스에서 가장 영향력 있는 특징 또는 변수를 식별하여, 모델의 예측 또는 분류에 가장 큰 영향을 미치는 입력 요소를 이해하는 데 도움을 줍니다. 예를 들어 음악 추천 시스템에서는 사용자의 청취 기록, 장르 선호도, 노래의 메타데이터와 같은 요소가 중요한 특징이 될 수 있습니다.

또 다른 전역적 설명 기법으로는 '규칙 추출'(Craven, 1996)이 있습니다. 이는 복잡한 모델의 동작을 모방하여 사람이 읽을 수 있는 규칙 또는 의사결정 트리를 생성합니다. 이러한 규칙은 의사 결정 프로세스에 대하여 전역적인 지식을 제공하고 해석가능성을 부여합니다. 예를 들어 의료 진단 모델에서는 증상, 검사 결과 및 환자 특성에 대한 특정한 조합으로 특정 진단을 나타낸 규칙을 추출할 수 있습니다(예: "환자의 나이가 50 세 이상이고 혈압이 높으면 고혈압으로 진단").

반면에 지역적 설명은 특정 결과(예: "내 입사 지원서가 거부된 이유")에 대한 AI 모델의 의사 결정 프로세스에 초점을 맞춥니다. 지역적 설명은 전체 모델에 적용되는 전역적 설명 대신, 특정 사례에 대한 모델의 동작을 명확히 하고 특정 예측이나 결정이 내려진 이유에 대하여 이해하는 것을 목표로 합니다.

지역적 설명에 대한 두 가지 기법의 예시로는 LIME 과 SHAP 이 있습니다.

LIME 은 Local Interpretable Model-agnostic Explanations 의 약자인데(Ribeiro, 2016), 이는

입력 데이터에 섭동을 생성(조작)하여, 원래 속성의 일부 값만 변경하는 일련의 인공 데이터를 생성하고, 모델의 출력을 관찰하는 기법입니다. LIME 은 이러한 관찰을 통하여 해석가능한 "대리" 모델을 생성하고 이를 설명할 수 있습니다. 대리 모델은 더 단순하고 해석이 가능하므로 입력된 특징이 모델의 결정에 어떻게 기여하는지 사용자가 이해할 수 있습니다.

예를 들어, LIME 은 소득, 신용 점수, 고용 이력 등 다양한 특성을 바탕으로 신청자의 대출을 승인할지 여부를 판단하는 데 사용될 수 있습니다. 이러한 시나리오에서 LIME 은 신청자의 높은 신용 점수와 안정적인 고용 이력이 의사 결정에 가장 긍정적인 영향을 미쳐서 대출을 승인했음을 보여줄 수 있습니다. LIME 은 입력과 출력 내용을 검토해서 어떤 특징이 평가에서 큰 비중을 차지하였는지 설명하는 더 단순한 (대리) 모델을 생성할 수도 있습니다.

Shapley Additive Explanations, 즉 SHAP(Lundberg, 2017)은 모델의 각 특징에 값을 할당하는 협동 게임 이론 [13]에 기반한 방법입니다. 이 기법은 가능한 모든 특징 조합을 고려하여 특정 사례 예측에 미치는 각 특징의 기여도를 계산합니다. 이 기법은 특징의 중요도에 대하여 통합된 척도를 제공하며 지역적 수준에서 모델의 결정을 설명하는 데 도움이 됩니다.

예를 들어, 평방 피트, 침실 수, 도심과의 거리 등의 특징을 기반으로 주택 가격을 예측하는 머신러닝 모델이 있는데, 특정 주택이 특정 가격으로 예측된 이유를 이해해야 할 필요가 있다고 가정해 보겠습니다. SHAP 를 주택 특성에 적용하면 각 특징이 이 주택에 대한 모델의 예측과 전체 주택 평균에 대한 모델의 예측 간 차이에 얼마나 기여했는지 파악하는 데 도움이 됩니다. 이러한 통찰성은 예측을 유도하는 요인과 이러한 요인이 서로 어떻게 상호 작용하는지 이해하는 데 도움이 될 수 있습니다.

그러나 LIME, SHAP 및 기타 섭동 기반 사후 설명 방법의 잠재적 약점을 보여주는 연구 결과도 나와 있습니다(Slack, 2020) (Lakkaraju, 2020). 이러한 방법으로 주입된 섭동은 정상적인 입력 데이터와 구별될 수 있기 때문에 모델은 이를 다른 것에서 분리하도록 지시할 수 있습니다. 또한 악의적인 개발자의 경우 섭동 기반 입력을 탐지하였을 때 겉으로는 "편향되지 않은" 출력을 제공하도록 하면서 매우 편향적이고 차별적인 모델을 만들 수도 있습니다.

실제로 여러 연구에서 사후 설명 방법을 신뢰할 수 있는 것으로 간주해서는 안 된다고 제언합니다. (Vale, 2022)는 "사후 설명 방법의 사용은 많은 경우 유용하지만 이 방법에는 한계가 있기 때문에 고위험 의사 결정에서 모델 결과의 공정성을 보장하는 유일한 메커니즘으로 의존하는 것을 금지한다 "고 합니다. 다른 연구(Bordt, 2022)에서는 "기술적, 철학적 관점에서 이러한 설명 방식은 알고리즘이 특정 결정에 도달한 '고유하고 진정한 사유'를 결코 밝혀낼 수 없다"고 말합니다. 이 연구는 "최악의 경우, 이러한 설명은

그렇지 않은 경우에도 '정당하거나' '객관적인' 결정이 내려졌다고 우리를 호도할 수 있다"고 결론내렸습니다.

따라서 모델의 공정성을 평가할 때는 '블랙박스' 접근 방식의 한계를 고려해야 합니다.

4. XAI 와 개인정보 보호

AI 결정에 대하여 투명한 통찰성을 제공하는 XAI 의 능력은 투명성, 책무성, 공정성 등 개인정보 보호의 여러 원칙을 준수하는 데 기여할 수 있습니다.

투명성

개인정보 처리의 투명성은 개인정보 보호의 핵심 원칙입니다. 개인정보는 적법하고 공정하며 정보주체와 관련하여 **투명한 방식으로** 처리되어야 합니다. 또한 컨트롤러[한국 개인정보보호법의 개인정보처리자에 해당함 – 역주]는 **명확하고 평이한 언어를 사용하여** 정보주체의 개인정보 처리와 관련된 모든 정보를 간결하고 **투명하며** 이해하기 쉽고 접근하기 쉬운 형태로 제공하기 위해 적절한 조치를 취해야 합니다.

또한 컨트롤러는 프로파일링을 비롯한 자동화된 결정의 존재에 대한 정보, 그리고 적용 로직에 대한 의미 있는 정보를 개인정보 수집 시 정보주체에게 제시해야 합니다.[14]

설명가능한 AI 는 AI 시스템이 어떻게 개인정보를 처리하고 결론에 도달하였는지에 대한 통찰성을 제공하여 결론 또는 결정을 이끌어낸 '추론'을 이해할 수 있도록 합니다. 또한 이러한 시스템의 영향을 받는 배치자와 개인이 의사 결정 프로세스를 이해하고 상호 작용할 수 있는 기회를 확대함으로써 이들의 역량을 강화할 수 있습니다.

일반적으로 말해 투명성은 AI 시스템의 사용에 대한 신뢰와 믿음을 조성해야 합니다. 나아가 투명성은 특히 공공 행정기관의 의사 결정을 지원하는 등 몇몇 경우에서 법적인 요구사항이 되며, 이 경우 해당 결정을 정당화해야 할 의무를 법적으로 부과받습니다.[15]

개인정보 컨트롤러의 책무성

조직은 개인정보 처리가 적법하고 투명한 방식으로 수행되도록 보장할 책임이 있습니다. 이러한 책임에는 개인정보 보호 원칙을 준수하는 메커니즘뿐 아니라 프로세스를 효과적으로 감독하고 감사할 수 있도록

지원하는 메커니즘을 구현해야 할 필요성도 포함됩니다.

시스템에 대한 책무성과 이해도가 높아지면 개인정보 컨트롤러가 수행해야 하는 위험에 대한 평가(예: 개인정보 보호 영향 평가를 수행하는 경우)도 더 잘 이루어질 수 있습니다.

XAI 를 적절하게 구현하면, 감사를 용이하게 만들 뿐 아니라, 조직으로 하여금 AI 기반 의사결정에 대한 책무성을 갖추고 책임 있는 AI 개발을 촉진하고 이러한 기술에 대한 대중적 신뢰를 조성하며 해당되는 규제 기준에 따라 AI 를 사용하는 데 중요한 역할을 할 수 있습니다.[16]

데이터 최소화

개인정보 보호 중심 설계 및 기본설정의 원칙은 데이터 최소화와 같은 개인정보 보호 원칙을 구현하기 위해 기술적 및 관리적 조치를 적용할 필요가 있다고 강조합니다. AI 의사 결정 프로세스에서 가장 영향력 있는 요소와 특징을 밝혀내는 XAI 의 기능은 개인정보의 수집, 보관 및 처리를 감소시키는 데 직접적으로 기여할 수 있습니다.

XAI 는 의사 결정에 중요한 데이터 포인트를 식별하여 조직이 개인정보 보호 규정을 준수할 수 있도록 도와줍니다. XAI 가 제공하는 통찰성은 보다 집중적이고 표적화한 데이터를 수집하려는 노력으로 이어져 개인의 프라이버시 침해를 최소화하는 동시에 정확하고 효과적인 AI 기반 결과를 얻을 수 있습니다.

특수 범주 개인정보

AI 학습에는 특수 범주 개인정보[민감정보에 해당함 - 역주]가 사용될 수 있는데, 이 정보는 잘못 취급하거나 오용될 경우 개인정보에 높은 위험을 초래할 수 있습니다. 예를 들어, 학습 데이터에서 종교나 성적 지향과 같은 특수 범주 개인정보를 유추할 수 있는 경우 AI 알고리즘의 불투명성은 특수 범주 개인정보 처리와 그 결과에 미치는 영향에 대하여 우려를 낳을 수 있습니다.

머신러닝 모델과 같은 AI 시스템은 특정 속성과 정보주체와 관련된 정보 간의 상관관계를 식별할 수 있으며, 이를 프록시[대리변수] 속성이라고 합니다. 특정 상황에서 프록시 속성은 개인에 대한 특수 범주 개인정보를 추론하는 데 사용될 수 있습니다.

예를 들어, 일부 도시에서는 우편 번호와 인종 집단 간에 강한 상관관계가 있을 수 있으며, 이는 우편번호 속성

을 인종에 대한 프록시로 만들 수 있습니다. AI 시스템은 학습 중에 이러한 상관관계를 파악하고, 예를 들어 신용 평가 결정을 내릴 때 이 프록시 속성에 기반한 의사 결정을 내릴 수 있습니다.

그러나 개인에 대한 이러한 추론은 완전히 잘못될 위험이 있습니다. XAI 는 개발자와 사용자가 특수 범주 개인정보와 의사 결정을 연결시킬 수 있는 프록시 속성을 식별하는 데 도움을 줄 수 있습니다.

XAI 를 구현한다고 해서 **자동으로 개인정보 보호 규정을 준수하게 되는 것이 아니라**는 점을 강조하는 것이 중요하겠습니다.
그럼에도 XAI 는 컨트롤러가 개인정보 처리 업무를 개인정보 보호 원칙에 따라 수행하였으며, 그 목적, 성격, 맥락, 범위는 물론 자연인의 자유와 권리에 미치는 심각한 위험의 발생가능성을 고려하였음을 입증할 때 유용한 기술적 조치가 될 수 있습니다.

5. XAI 구현과 관련된 위험

설명가능성은 AI 시스템의 투명성과 신뢰성을 증진할 수 있는 잠재력을 가지고 있지만, 이를 도입하는 것이 컨트롤러, 개발자, 엔지니어 및 정보주체에게 위험을 초래할 수도 있습니다. 설명가능성을 구현할 때는 다음과 같은 위험을 완화하기 위한 예방 조치를 취해야 합니다.

잘못된 해석(misinterpretation)

XAI 는 구현 방식에 따라 청중이 이해하기에 너무 복잡하거나 너무 기술적인 설명을 낳거나, AI 모델의 전체적인 복잡성을 파악하지 못하는 방식으로 지나치게 단순화한 설명을 유발할 수 있습니다(**'블랙박스 접근 방식: 사후 설명'** 장 참조). 두 경우 모두 개인의 오해를 초래할 수 있습니다.

처리와 관련된 정보는 명확하고 평이한 언어를 사용하여 간결하고 투명하며 이해하기 쉽고 접근하기 쉬운 방식으로 정보주체에게 제공되어야 합니다. 따라서 설명은 전문 용어와 기술적 복잡성을 피하고 이해하기 쉬운 방식으로 이루어져야 합니다.

잘못된 해석의 위험을 줄이기 위해서, 조직은 먼저 설명의 대상이 되는 다양한 **이해관계자를 파악해야 합니다.** 그다음 각 청중에 따른 설명의 세부 수준을 조정해야 합니다. 설명은 명확하고 쉬운 언어를 사용하여 문제의 복잡성과 개인의 이해 수준 사이의 간극을 좁혀야 합니다. 그래픽 표현을 비롯하여 사용자 친화적인 인터페이스를 사용하면 설명 과정을 용이하게 만들 수 있지만, 그렇다고 해서 시스템을 지나치게 단순화해서는 안 됩니다. 설명이 AI 시스템의 동작을 정확하게 반영하고 사용자가 불완전하거나 부정확한 설명으로 오도되지 않으려면 XAI 방법에 대하여 신중하게 검증하고 테스트할 필요가 있습니다.

또한 조직은 XAI 설명이 명확할 뿐 아니라 **중립적**이어서 기존의 편향을 강화하지 않도록 보장해야 합니다.

시스템의 악용 가능성

조직은 개인정보 처리로 인해 기밀성, 무결성 및 가용성을 비롯하여 개인에게 미치는 위험이 발생하였을 때 적합한 수준의 보안을 보장하는 기술적 및 관리적 조치를 적절하게 구현해야 합니다.

XAI 의 맥락에서 이는 AI 시스템을 악용하여 잠재적으로 개인에게 영향을 미칠 수 있는 개인정보나 세부 정보를 노출할 위험을 방지하는 것을 의미합니다.

이 논문(Kuppa, 2021)에서는 XAI 가 제공하는 정보를 사용하여 안티바이러스 시스템을 대상으로 멤버십 추론 공격과 적대적 공격 등 여러 유형으로 공격하는 경우를 소개합니다. 저자에 따르면 *"반사실적 설명방법(counterfactual explanation)은 공격자가 입력 공간에서 수렴하기 어려운 '블랙박스' 최적화의 문제를 해결하는 대신 적대적인 샘플을 더 빨리 찾는 방법으로 사용될 수 있습니다. 공격자는 반사실적 설명을 사용하여 공격 경로를 최적화할 수 있습니다."*

이를 위해서는 시스템의 투명성과 민감한 구성 요소의 보호 사이에 신중한 균형을 이룰 필요가 있습니다.

영업 비밀 공개

마찬가지로, XAI 는 재산적 정보 또는 민감한 영업 전략을 노출시켜 AI 시스템의 제공자(또는 배치자)의 사업상 경쟁력에 손실을 가져오는 위험도 야기합니다.

책무성의 원칙, 그리고 개인정보 중심 설계 및 기본설정 원칙이 이 문제와 관련이 있습니다.

조직은 처리 수단이 결정되는 순간부터 특히 이러한 시스템의 사용으로 인해 영향을 받을 수 있는 개인에게 유익한 정보를 설명이 제공할 수 있도록 XAI 를 구현하는 메커니즘을 구축해야 합니다. 이는 독점 알고리즘, 영업 비밀 또는 기타 상업적으로 민감한 세부 정보를 과도하게 노출시키지 않고도 수행할 수 있습니다.

배치자의 AI 시스템 과의존

설명은 인간이 AI 의 추천을 정확성에 관계없이 '맹목적으로' 받아들일 가능성(자동화 편향)을 높일 수 있습니다. 특히 의료 분야와 같이 오류로 인한 비용이 높은 분야에서 성공적인 '인간-AI 상호작용'을 위해서는 인간의 비판적 참여가 필수적인 요소입니다(Gajos, 2022).

조직은 최소한 개인이 컨트롤러에게 인적 개입을 요구하고, 자신의 관점을 표현하며, 해당 결정에 이의를 제기할 수 있는 메커니즘이 XAI 구현에 포함되어 있는지 확인해야 합니다. 이는 프로파일링을 비롯한 자동화된 처리에만 기반하여 법적 효력을 가지거나 자신에게 중대한 영향을 미치는 결정에 종속되지 않을 개인의 권리에 부합하는 조치입니다. 보다 일반적으로 XAI 는 결정을 내리는 책임자가 결정을 통제하고 균형 잡힌 시각을 유지하며 AI 시스템에 지나치게 의존하지 않도록 보장하는 것을 목표로 합니다.

AI 시스템에 과도하게 의존하는 위험 문제를 해결하기 위해 조직은 중대한 결과, 특히 신체적 또는 경제적 피해에 대한 위험이나 개인과 집단의 권리와 자유에 대한 위험을 초래하는 결정에 인간이 참여하고 인간이 감독하는 것을 적극적으로 장려해야 합니다.

기술 발전이 책임감 있고 사회적으로 수용 가능한 결정으로 이어지려면 AI 의 한계에 대하여 명확한 의사소통이 필요합니다. 이러한 시스템의 사용으로 인해 영향을 받는 사람들은 필요한 경우 인간의 개입을 요청하도록 (그리고 쉽고도 적시에 인간의 지원을 받을 수 있도록) 장려되어야 합니다.

본질적으로 XAI 는 상당한 잠재력을 보여주지만, AI 의 신뢰성을 향상시키는 데 있어 그 중요성과 한계에 대한 포괄적인 이해가 수반되어야 합니다. 이를 위해서는 포괄적이고 심층적인 위험 평가, AI 시스템의 기능에 대한 지속적인 모니터링, 개인정보 보호 당국과 관할 부문별 감독 기관(예: 근로 감독 기관, 보건의료 감독 기관, 금융 감독 기관 등) 간의 협력 노력을 통하여, 책임감 있고 안전한 구현을 보장해야 합니다.

5. 인적 요소의 중요성

AI 시스템을 설명하는 접근 방식이 무엇이든, 설명은 궁극적으로 사람들에게 관련 있고 의미가 있는 것이어야 하므로 인간적인 측면을 고려하는 것이 필수적입니다. 인간은 대조적인 설명에 대한 선호, 선택성, 설명에 대한 신뢰, 설명의 맥락화 능력 등 여러 가지 요인에 따라 서로 다른 방식으로 정보를 인식하고 처리합니다.

설명가능성을 제공할 때 고려해야 할 사항

인간은 대조적인 설명을 선호합니다

사람들은 "왜?"에 대해 알고 싶어하는 것을 넘어 "왜 Q 대신 P 라는 사건이 일어났는가?"라고 질문하는 경향이 있습니다. 대조적인 설명은 선택지 간의 주된 차이점을 강조함으로써 복잡한 의사결정 과정을 단순화하고 개인이 과거의 선택으로부터 배우고 자신의 의사결정 전략을 개선할 수 있는 바탕이 됩니다. (Miller T., 2019)

인간은 선택적입니다

복잡한 설명에 직면했을 때, 사람들은 가장 핵심적이거나 관련성이 높은 측면에 선택적으로 집중하고 덜 중요하다고 생각되는 세부 사항을 걸러낼 수 있습니다. 또한 기존에 알고 있는 지식과 일치하는 설명에 끌리는 경향이 있을 수 있습니다. (미텔슈타트, 2019)

인간이 설명을 신뢰해야 합니다.

신뢰성은 시스템의 정확성과 신뢰도에 대한 문제로 생각될 수 있지만, 개인이 주어진 설명을 얼마나 신뢰하는지에 대한 문제로 생각될 수도 있습니다. 너무 복잡하거나 불완전하거나 부정확한 설명은 전체 시스템에 대한 불신을 유발할 수 있습니다. (리베이로, 2016)

설명은 맥락에 맞아야 합니다

XAI 시스템이 자신의 기능과 이해를 설명할 수 있어야 하지만 모든 설명은 AI 시스템 사용자의 작업, 능력, 기대치에 따라 달라지는 맥락 속에서 이루어지는 것입니다. (Gunning, 2019)

설명은 사회적입니다

설명은 대화나 상호작용의 일부로 제시되는 지식의 전달이므로, 피설명자의 생각에 대한 설명자의 생각에 상대적인 방식으로 이루어집니다. 개인 대 집단의 행동, 규범 및 도덕 등의 영향을 받습니다. (Miller T., 2019)

또한 설명을 제공할 때는 '목표 청중'을 고려해야 합니다.

예를 들어, 감독 당국이 시스템을 조사할 때는 AI 가 서비스하는 활동과 관련된 법률(예: AI 노무 관리 시스템의 경우 작업장의 보건 및 안전 조건)을 준수하는지 확인하기 위해 더 자세한 설명이 필요할 수 있습니다.

결론

신뢰할 수 있는 AI 가 되려면 무엇보다도 투명하고, 책무성을 갖추며, 윤리적이어야 하고 [17], 설명가능한 AI 는 이러한 특정 요구사항을 충족하는 데 중요한 역할을 할 수 있습니다.

EDPS 의 관점에서 볼 때, XAI 의 개념은 '인간 중심' AI 를 개발하겠다는 약속을 구현합니다. AI 가 내린 결정의 '이유'를 설명함으로써 AI 기반 결정으로 인해 권리가 침해된 개인이 진정 제기 등 의미 있는 방식으로 디지털 환경에 참여할 수 있도록 합니다.

XAI 는 AI 결정의 근거를 공개함으로써 개인이 자신의 개인정보가 어떻게 취급되고 있는지 이해할 수 있는 통찰력을 제공합니다. XAI 가 제공하는 투명성은 조직과 사용자 간의 신뢰를 강화할 뿐만 아니라 핵심적인 개인정보 보호 원칙에도 부합합니다.

또한, XAI 는 정보주체가 부당한 결정을 식별하고 이의를 제기할 수 있도록 하여 의사 결정의 공정성을 높이고 평등한 대우와 차별 금지 원칙을 촉진할 수 있습니다. 그러나 XAI 는 단순히 투명성을 향한 하나의 걸음에 그치지 않습니다. 이는 기계가 주도하는 프로세스와 인간이 추구하는 정당성, 신뢰, 공정성 사이의 간극을 메우기 위한 도약입니다. XAI 의 역할은 단순한 설명을 넘어 인간의 이해, 윤리적 판단, 공감적 배려가 AI 기술 사용에 있어 필수적인 면모라는 인식을 내포합니다. 인간적 차원은 여전히 필수적입니다. 인공지능의 혁신적인 잠재력을 받아들일 때, 인공지능 시스템의 사용이 개인은 물론 사회 전체에 중대한 영향을 미칠 수 있다는 점을 기억하는 것이 중요합니다. 이러한 맥락에서 인공지능은 기술을 넘어 인간의 이해, 책임성, 윤리적 감독에 대한 필요성을 인공지능의 잠재력과 결합하는 문제가 됩니다.

XAI 의 도입은 AI 의 장래 기술적인 기능 정의에 기여할 뿐 아니라 인권, 윤리, 책무성을 지키기 위한 인류 공동의 책임 또한 규정할 것입니다.

그러나 XAI 의 등장으로 인해 여러 잠재적 문제들도 발생하고 있어 세심한 주의가 필요합니다.

위에서 강조한 바와 같이, 복잡하거나 지나치게 단순화한 설명은 오해를 불러일으킬 수 있고, XAI 가 잘못 사용되거나 악의적으로 사용될 경우 시스템 작동을 정당화하기 위한 '설득 연습'으로 변질될 수 있습니다. 또한 영업 비밀에 대한 과도한 보호는 투명성을 저해할 수 있고, 나아가 (Adadi, 2018)에서 강조한 바와 같이 XAI 의 재정적 비용도 고려해야 합니다. *"AI 시스템을 설명가능하게 만드는 데에는 의심할 여지없이 비용이 많이 듭니다. AI 시스템의 개발과 실제 질의 방식 모두에서 상당한 자원이 요구됩니다."*

AI 시스템의 기능(및 대중적 요구)이 커져감에 따라 AI 개발자가 새로운 혁신을 추구하기 위해 윤리적 고려 사항을 무시할 위험도 커지고 있습니다.

하나의 사회 구성원으로서 우리는 이러한 발전이 기본권, 특히 프라이버시 및 개인정보 보호에 대한 기본권을 보호하는 방식으로 이루어지도록 요구하고 보장해야 할 책임이 있습니다.[18]

EU 기관, 기구, 사무소 및 청을 감독하는 개인정보 보호 당국으로서, 우리 기관은 AI 시스템의 사용이 개인정보 보호 원칙에 부합하고 해당 법률에 명시된 규칙을 준수하도록 보장할 책임이 있습니다.

이 간행물은 유럽 개인정보 보호 감독관(EDPS)의 기술 및 개인정보 보호 부서에서 작성한 간략한 보고서입니다. 이 보고서는 새로운 기술에 대한 사실적인 설명을 제공하고 사생활 및 개인정보 보호에 미칠 수 있는 영향을 논의하는 것을 목표로 합니다. 이 간행물의 내용은 EDPS 의 정책적 입장을 의미하지 않습니다.

이슈 작성자: 비토르 베르나르도

편집자: 편집자: 마시모 아토레시, 자비에 라레오, 루이스 벨라스코

연락처: techdispatch@edps.europa.eu

테크디스패치 간행물을 구독하거나 구독 취소하려면 다음 주소로 메일을 보내주십시오.

techdispatch@edps.europa.eu. 개인정보 보호 고지는 EDPS 웹사이트에서 온라인으로 확인할 수 있습니다.

edps.europa.eu


@EU_EDPS


EDPS


European Data Protection Supervisor


@EDPS@social.network.europa.eu


@EDPS@tube.network.europa.eu