01_Recent Trends in Explainable AI

💡 KAIST 김재철AI대학원 XAI Tutorial 2024 강의를 통해 안전하고 신뢰할 수 있는 AI 시스템을 구현하는 데에 필수 요소인 설명가능한 인공 지능의 원리와 최근 동향을 살펴보자.

설명가능 인공지능의 현재와 미래

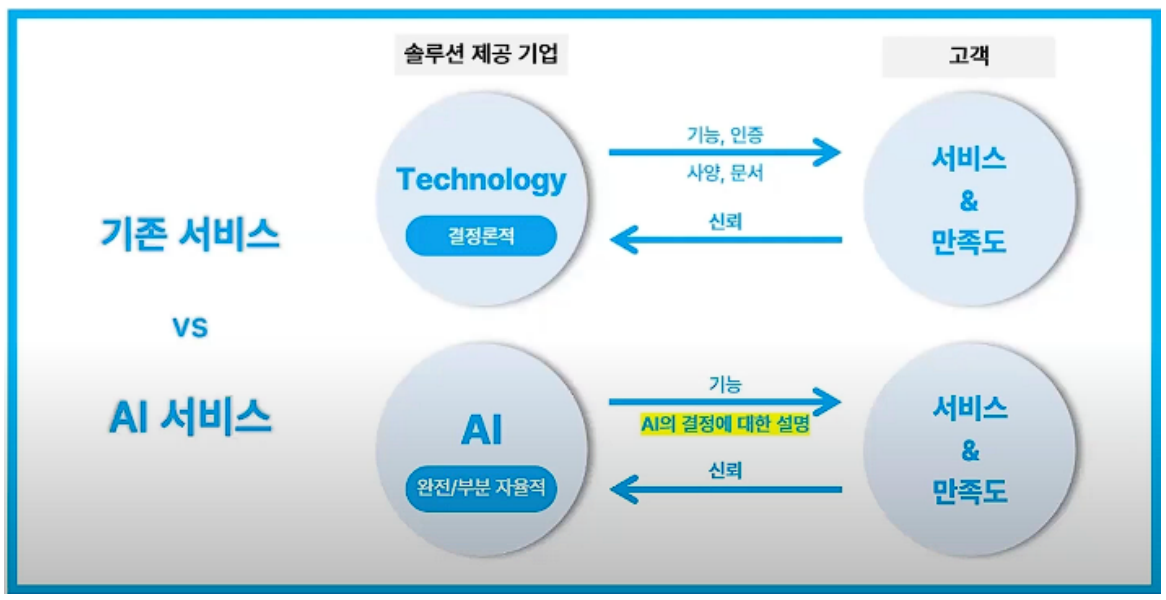
1. 인공지능의 현재

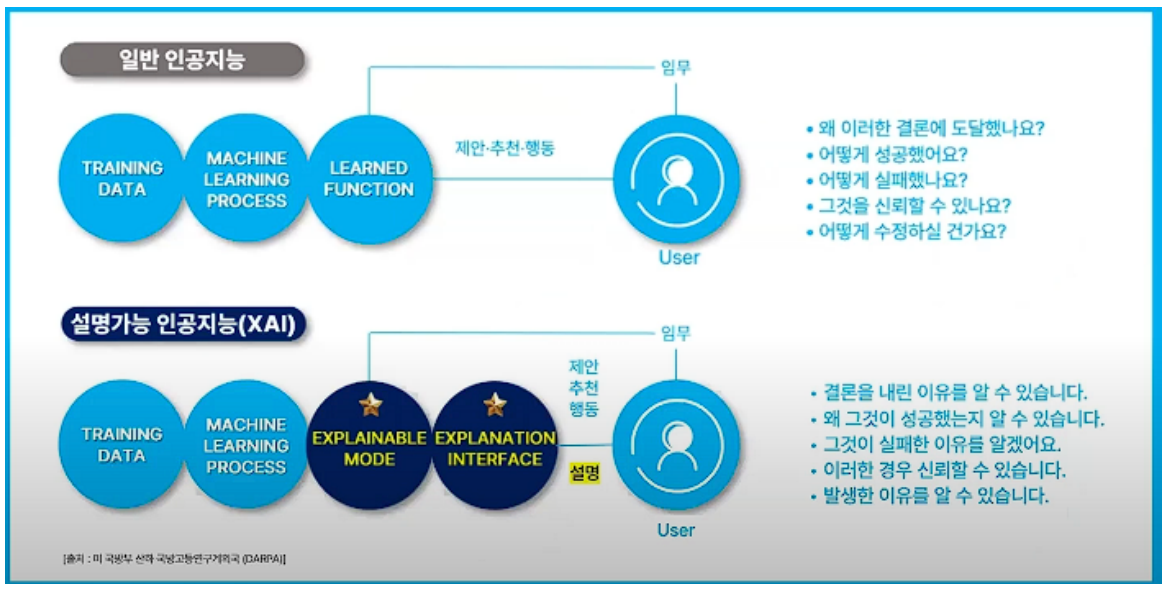
- 2016년 : 여러 이미지 인식 성능의 급격한 향상(SegNet→PSPNet)
- 2023년 : 생성형 AI의 언어능력 향상(ChatGPT)
- 2017 ~ 2030까지 AI로 인한 GDP변화가 전 세계 GDP변화의 15% 비중을 차지할 것으로 전망

---

2. 인공지능 미래

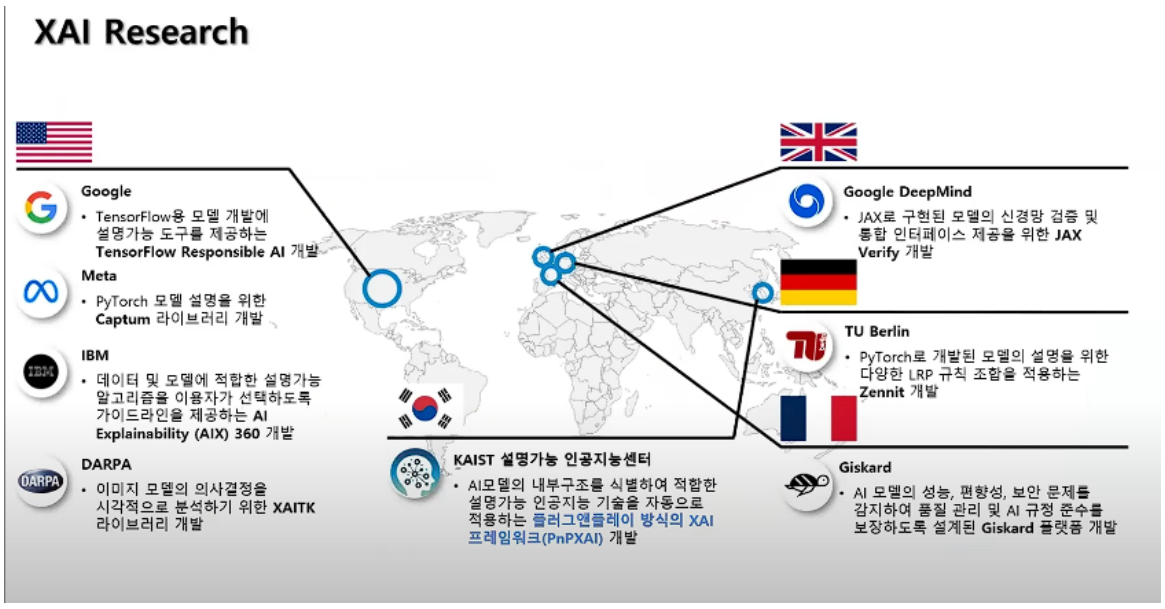
- 기존의 서비스들은 우리가 예측한대로 행동하지 않으면 너무 불편하게 느꼈다.
  - Ex) 리모콘으로 TV를 켜면 켜져야된다. 안켜지면 답답함
- But, AI 서비스는 예측가능하지 않은 결과를 허용한다. 더 나은 정보가 있다면 정제된 상태로 얻기를 원한다.





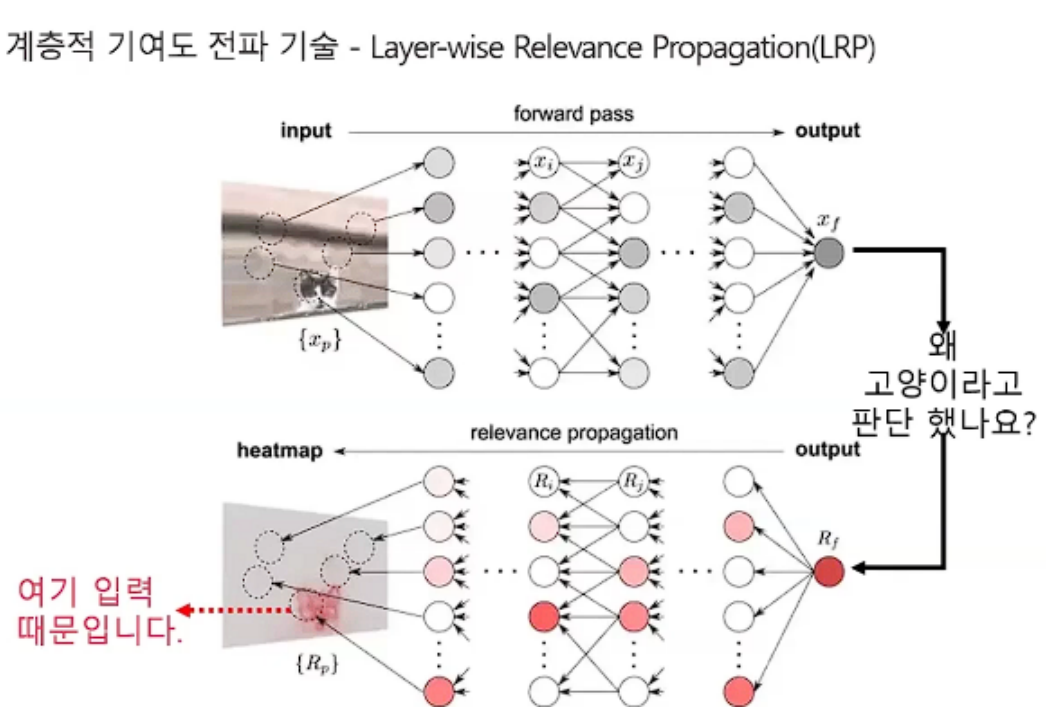
3. 인공지능 신뢰성

- AI 시스템을 만드는 사람 입장에서는 AI의 공정성, 안정성, 신뢰성은 아주 중요하다.
- 어떤 문제가 발생했을 때, 데이터의 문제인지, 알고리즘의 문제인지, 모델의 문제인지, 사용자의 문제인지 정확하게 파악할 수 있어야 한다.



NVIDIA의 PilotNet(자율주행 딥러닝)을 설명하는 AI(예시)

- LRP의 목적은 입력이 있을 때 '왜 고양이라고 판단 했는가'라는 질문에 어디를 보고 고양이라고 했는지 설명해주는 것이다.
- 고양이의 부분적 특징(고양이 귀)만을 파악해서 고양이라고 했는지, 아니면 고양이를 완벽히 이해해서 고양이라고 했는지를 알 수가 있다.
- 설명 가능한 인공지능의 알고리즘도 bias가 존재한다.
  - 알고리즘을 적용했을 때 실제 큰 틀에서는 비슷하게 결과를 주지만, 강론에서 보면 어떤 면에서는 러프하거나 정교한 다양한 알고리즘들이 존재한다.
  - 그러나 이들은 bias가 섞여서 나오면 한 두개의 알고리즘으로 어떤 것이 문제인지 파악하기 어렵다.
  - 결과적으로 여러 개의 알고리즘을 겹쳐서 보는 방법이 중요하다.
- 설명에서 끝이 아니고 '모델에 혹시 문제가 있지 않은가', '이 케이스에 대해서 우리가 일반화 성능을 잘 이끌어낼 수 있도록 데이터가 괜찮은가'등도 중요하다.



XAI 요약

- AI에 대해서 설명하는 새로운 방법을 만드는 것도 중요하지만, 기존의 기법들을 잘 만들고 표준화의 입장에서 정의를 정확히 하는 것도 중요하다.

- 설명성에 대한 evaluation을 어떻게 할 것인지, 각 domain마다 설명을 잘 맞추어서 하는지, multi demensional하게 설명하는지, social한 impact에 대해서 잘 고려하는지.. 등등 XAI에 있어서 중요한 점들이다.