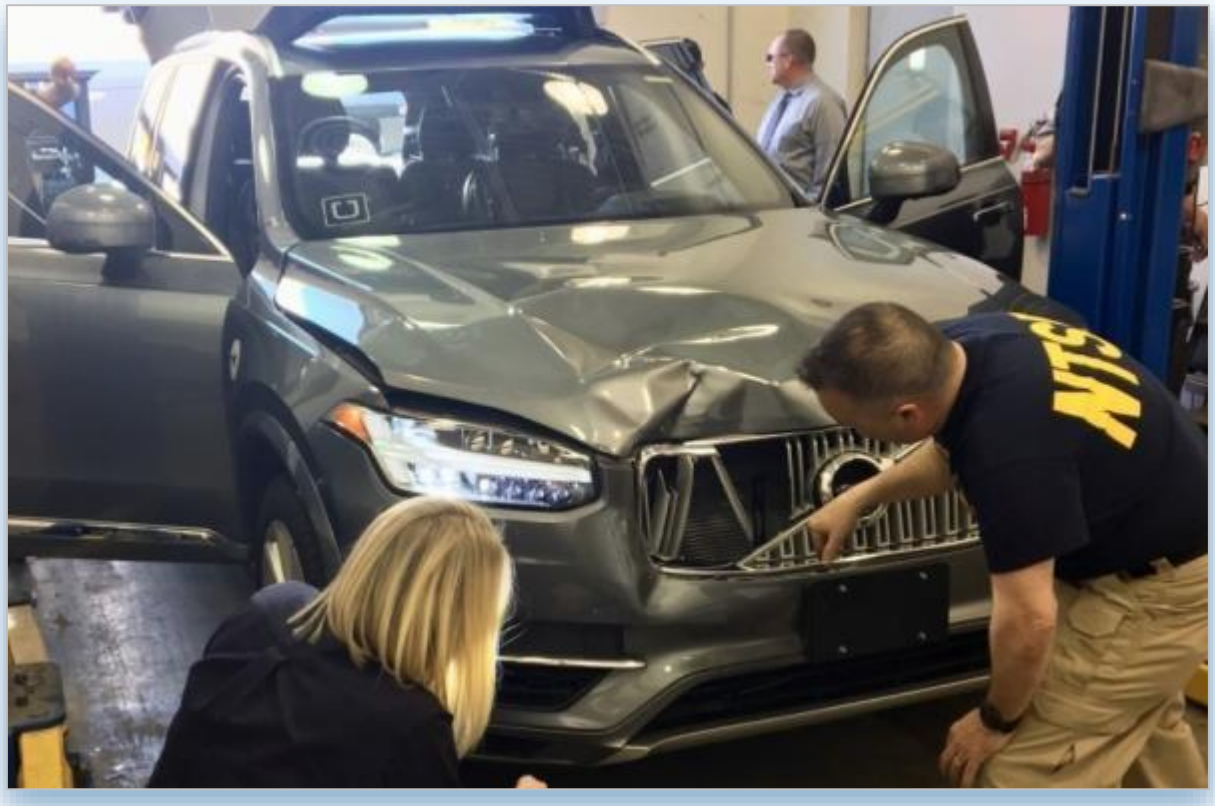
INEEJI EMPOWERING EXPERTISE

Explainable Artificial Intelligence to Connect AI and Human

KAIST Kim Jaechul Graduate School of AI
INEEJI Corporation

Jaesik Choi

Do We Understand the Internal Mechanism of AI Models?



Uber's self-driving car killed a pedestrian (Marc 18th, 2018)

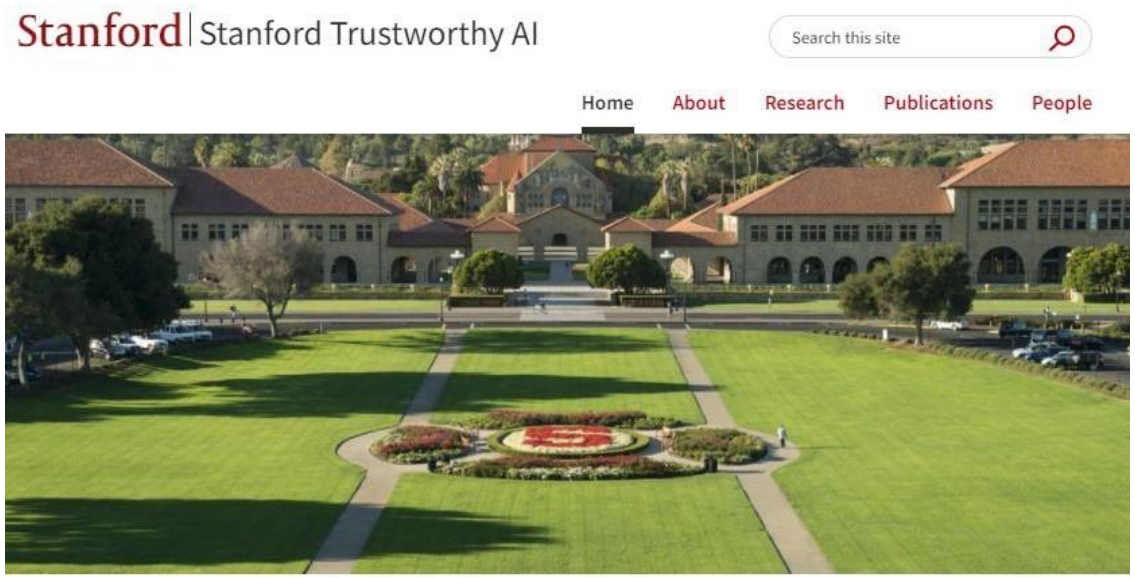
The ‘safety driver’ was watching a TV show (June 22th, 2018)

EU General Data Protection Regulations (GDPR)

항목	내용
Right to be forgotten	17 : When customers do not want, the personal contents should be elemen **eliminated**
Llimit of AI decision	22 : Customers have the right not to be handled by **AI algorithm**
Rright to explanation	13-14 : Customers have the right to receive **proper explanations on the decisions made by AI algorithms**
Fines	Up to **4% of total global revenue**
Enact	**2018/05/28**

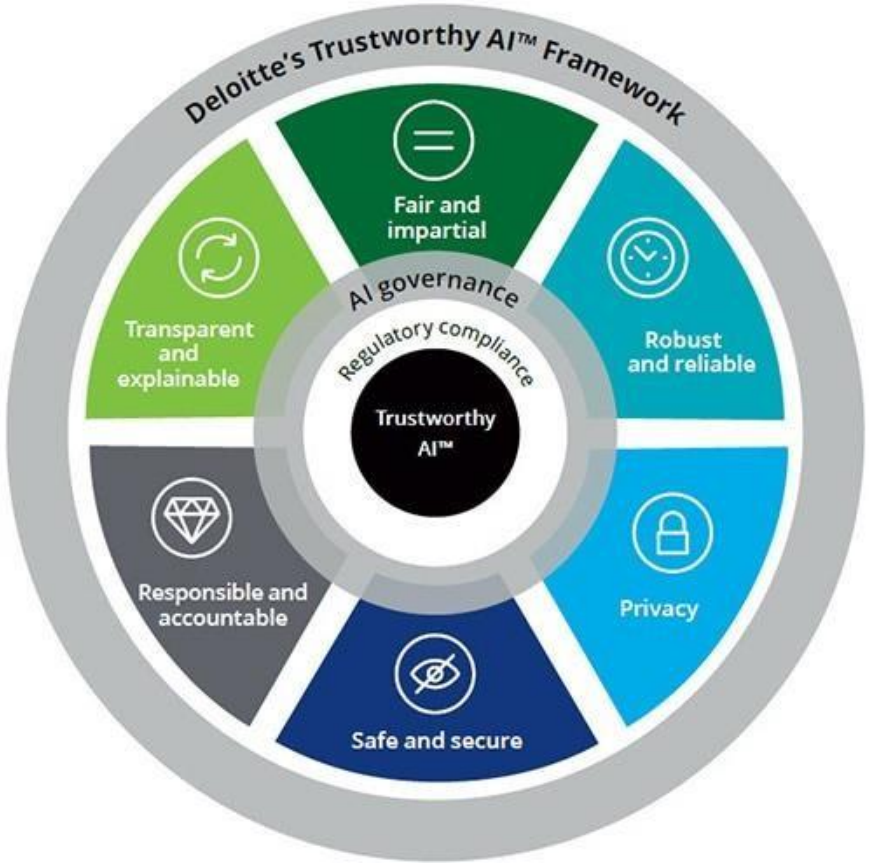
In the area of high risk AI, the fine will be up to 6% of total global revenue

Trustworthy AI



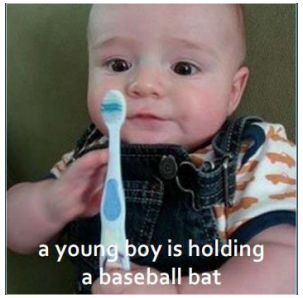
Our Mission

Stanford Trustworthy AI aims to **supercharge** innovations in **artificial intelligence** with **human understanding**. We engage in translational research across fairness, explainability, privacy, and robustness, guided by ethics.

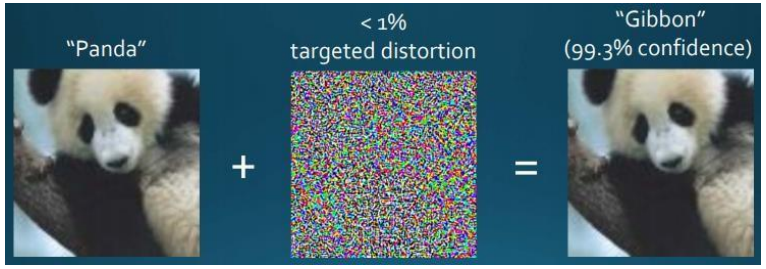


A DARPA Perspective on AI – Three Waves of AI

DESCRIBE	CATEGORIZE	EXPLAIN
Handcrafted Knowledge	*Statistical Learning*	*Contextual Adaptation*



Statistically impressive, but individually unreliable

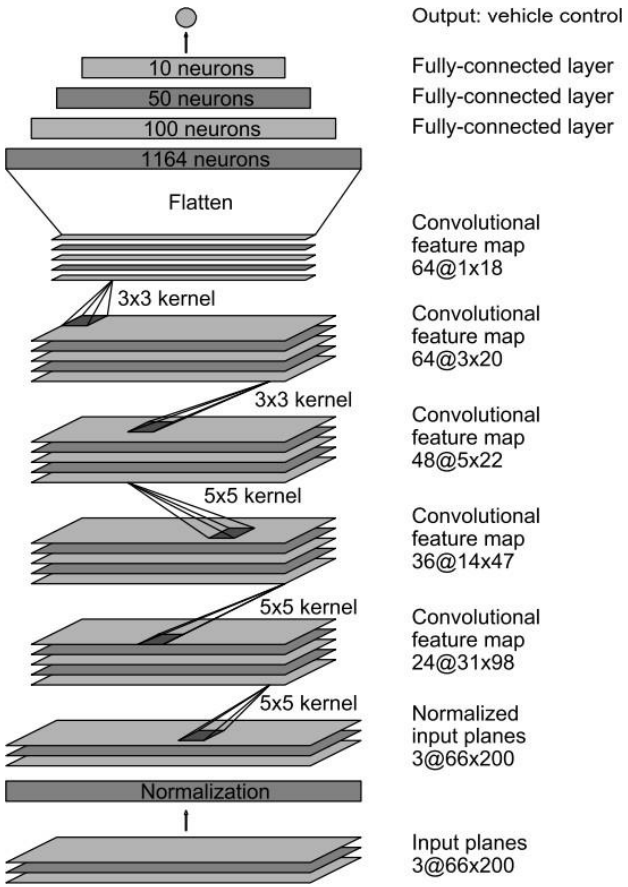


Inherent flaws can be exploited

Skewed training data creates Maladaptation

(Some) Initiatives: XAI in EU

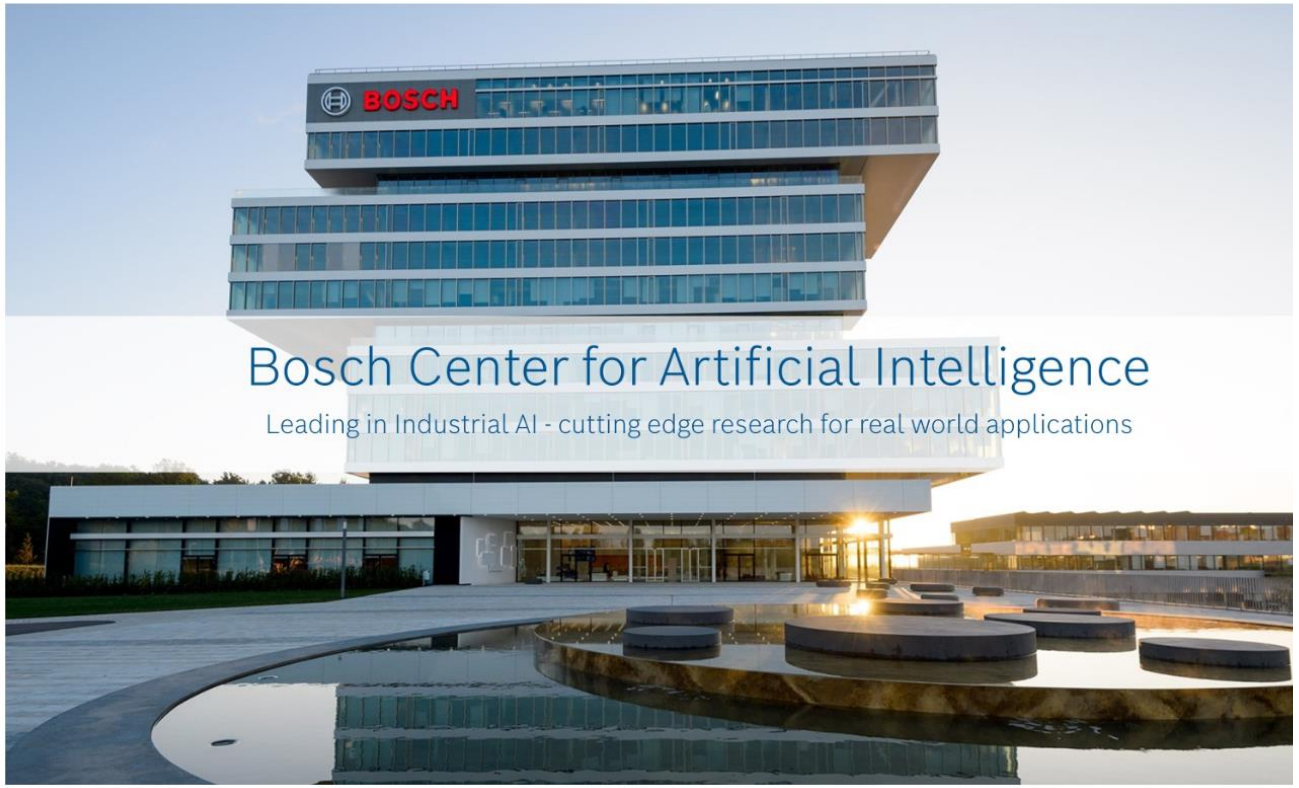
Explainable AI for PilotNet of NVIDIA



PilotNet 구조도
NVIDIA/Google 연구팀, 2017년

Explaining the decisions of PilotNet

Bosch Center for Artificial Intelligence



Introduction

The Bosch Center for Artificial Intelligence (BCAI) was created in 2017 out of existing competence centers to deploy cutting-edge AI technologies across Bosch products and services creating solutions that are "Invented for life". We employ roughly 270 associates worldwide, dedicated to over 185 projects within seven locations: Germany, India, United States, Israel and China.

Using data from Bosch's various business divisions, we conduct cutting-edge research that focuses on safe, secure, robust, and explainable AI. We design and implement AI for smart, connected, and autonomous technologies across Bosch business sectors.

By collaborating with international partners, we are committed to fostering scientific exchange. We actively look for opportunities to expand our research network further and to collaborate with industry thought leaders.

Venture Capital's Investment on XAI

1	미국	Kyndi	KYNDI	$28.5M	9	일본	Skydisc	SKYDISC	¥1.7B
2		Fiddler	Fiddler	$45.2M	10	영국	Intellibonds	intellibonds	500K
3		EquBot	EQUBOT	$2.1M	11		Factmata	FACTMATA	$3.6M
4		Convoy	CONVOY	$675,5M	12		Genie AI	GENIE AI	£2M
5		Zest	ZEST AI	$250M	13		Imandra	IMANDRA REASONING AS A SERVICE	$12.6M
6		Truera	truera	$17.3M	14	프랑스	DreamQuark	DreamQuark	$19.1M
7		Digite	ANTHROP\C	$124M	15	캐나다	FAIRLY	FAIRLY	-
8	일본	Hacarus	HACARUS	$24.3M	16		DarwinAI	DARWIN AI	$11.9M

The team analyzes the internal mechanism
of Deep Neural Network

Explainable AI Program in Korea – Part I

Goal → Human-level Learning and Inference to overcome the limitations of Deep Neural Networks

AS-IS (Deep Learning)

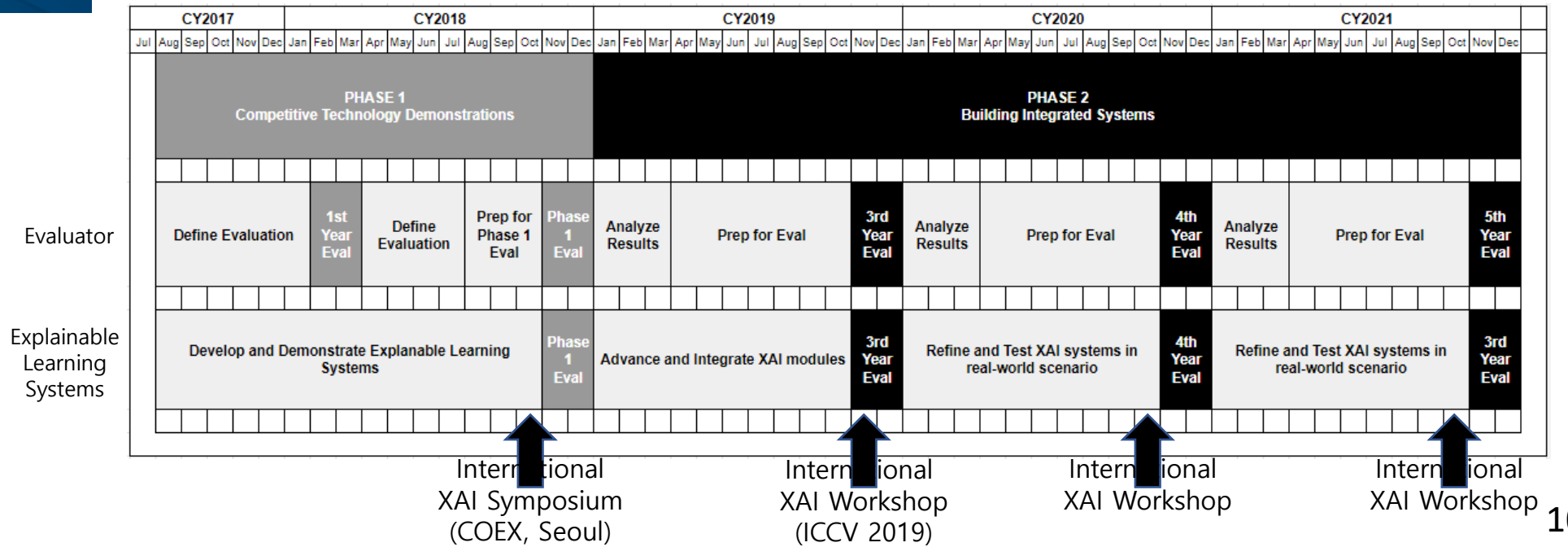
- It is hard to know the decision, so called **Blackbox** model
- It does not work well when we **do not have enough training data**

TO-BE (Human-level Learning/Inference)

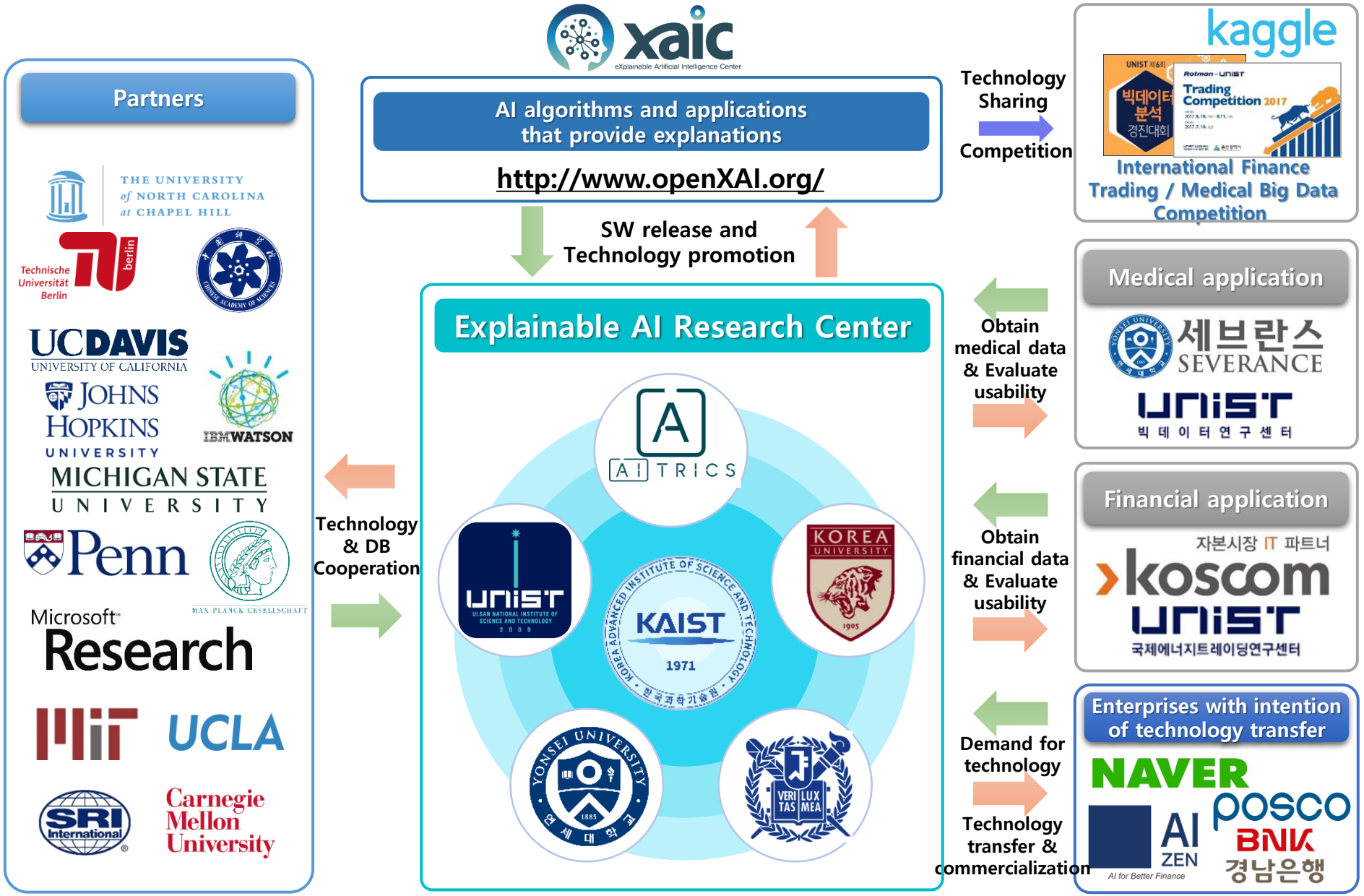
- **Explainable learners** which can provide the reasons of decisions
- Learning explainable models even with **data deficient environment**

Fund → Institute of Information & Communication Technology Promotion (IITP) under Ministry of Science and ICT (MSICT) as part of Innovative Growth Engine Project

Period → July 2017 ~ December 2021 (54 months)



Explainable AI Program in Korea – Part I



Explainable AI Program in Korea – Part I

2018 International XAI Symposium

ICCV 2019
Seoul, Korea

2019 ICCV Workshop on

Interpreting and Explaining Visual Artificial Intelligence Models

Saturday, November 2nd, 2019

@ COEX 318AB, Seoul, Korea

2019 International XAI Workshop

http://xai.kaist.ac.kr/workshop/2019/

KDD2020 Tutorial on

Interpreting and Explaining Deep Neural Networks: A Perspective on Time Series Data

2020 KDD Tutorial

http://xai.kaist.ac.kr/Tutorial/2020/

Trevor Darrell
Professor
UC Berkeley

Wojciech Samek
Head of Machine Learning Group
Fraunhofer Heinrich Hertz Institute

David Bau
PhD student
MIT

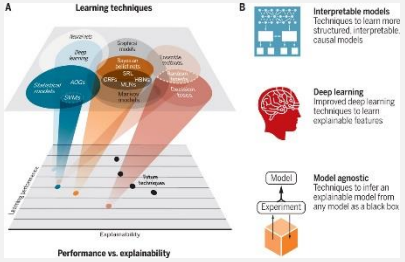
Ludwig Schubert
Software Engineer
OpenAI

Explainable AI Program in Korea – Part I

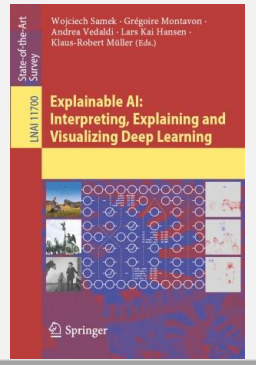
Research Results

AI Top Conference Papers (ICML, NeurIPS, AAAI, …) 87

Top Journal Papers (Science Robotics, …) 45



Book Edited (Explainable AI: Springer)



Technology Transfer

Patents (Registration) 37(2)

Industrial Projects 11

Manufacturing	Healthcare	Finance	Mobile
Steel, Semiconductor			
Process Explain	ICU monitoring	Credit Rating	Robust Generation

Open Source/Meetings

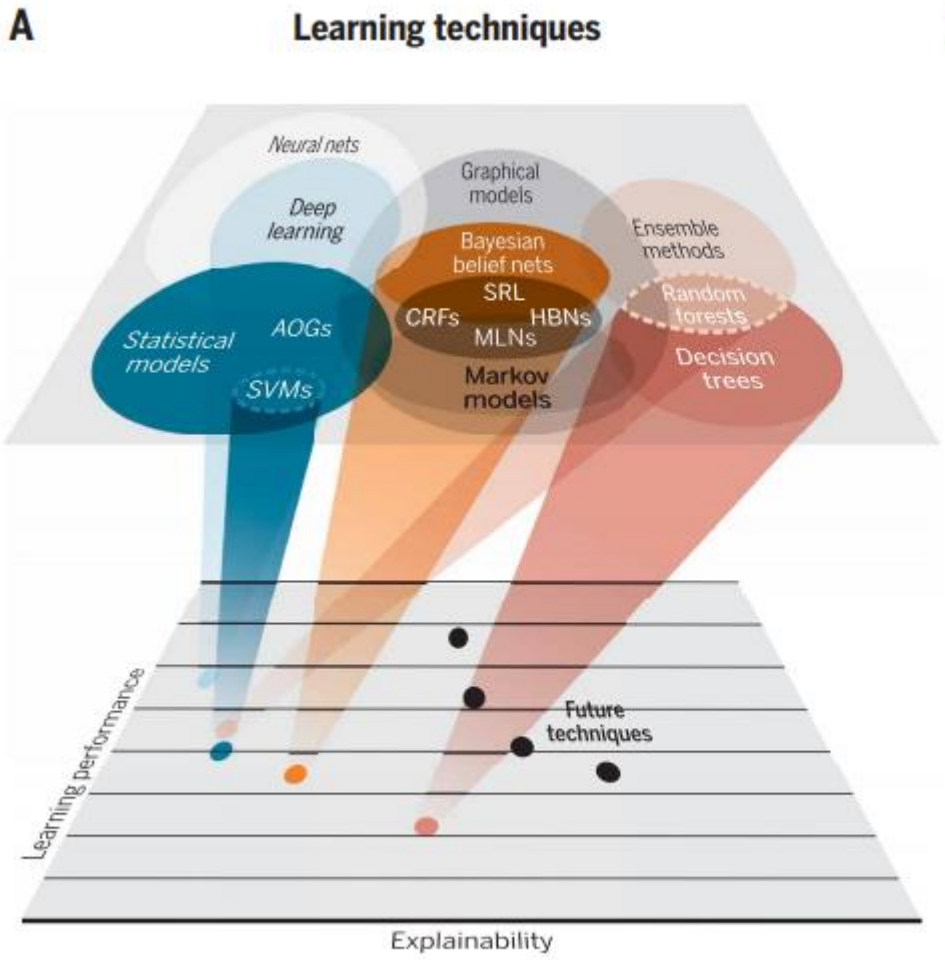
Open Source Projects 44
github.com/OpenXAIProject

Online Tutorial 31

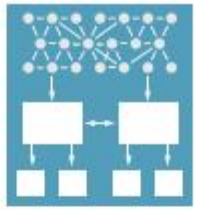
Open Workshop 10

International Gathering 3

Methodology in Explainable Artificiail Intelligence

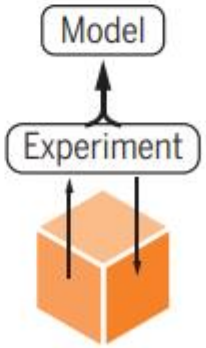


Performance vs. explainability



Interpretable models
Techniques to learn more structured, interpretable, causal models

Deep learning
Improved deep learning techniques to learn explainable features



Model agnostic
Techniques to infer an explainable model from any model as a black box

D. Gunning et al., Science Robotics, 2019

Attribution Preservation in Network Compression for Reliable Network Interpretation

Maintaining explainability while compressing neural networks

Prior art [UMD/NEC Labs, ICLR 2017] → **SOTA [KAIST, NeurIPS 2020]**

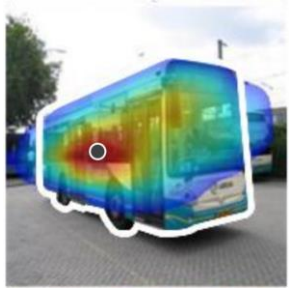
Before

Explanation
Loss

After

Before

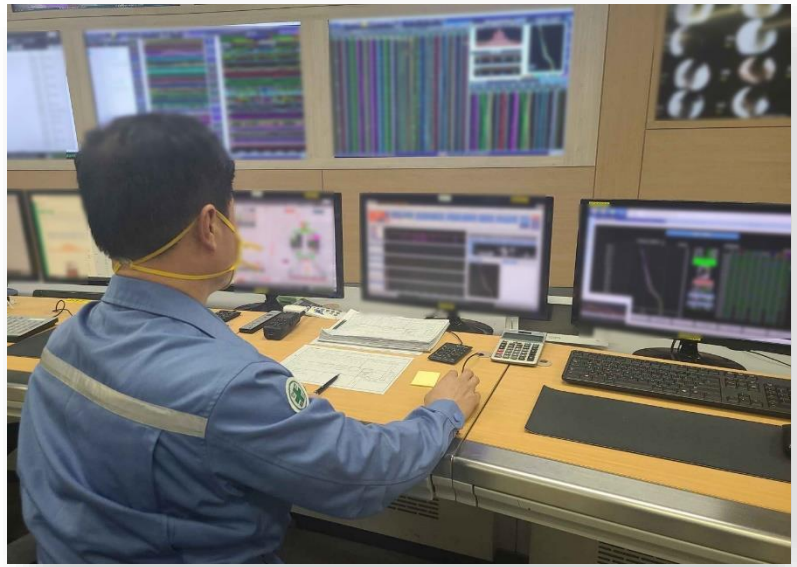
Explanation
Preserved



After

[1] H. Li et al., "Pruning filters for efficient convnets." *ICLR,* 2017.
[2] G. Park et al., "Attribution Preservation in Network Compression for Reliable Network Interpretation." *NeurIPS*, 2020.

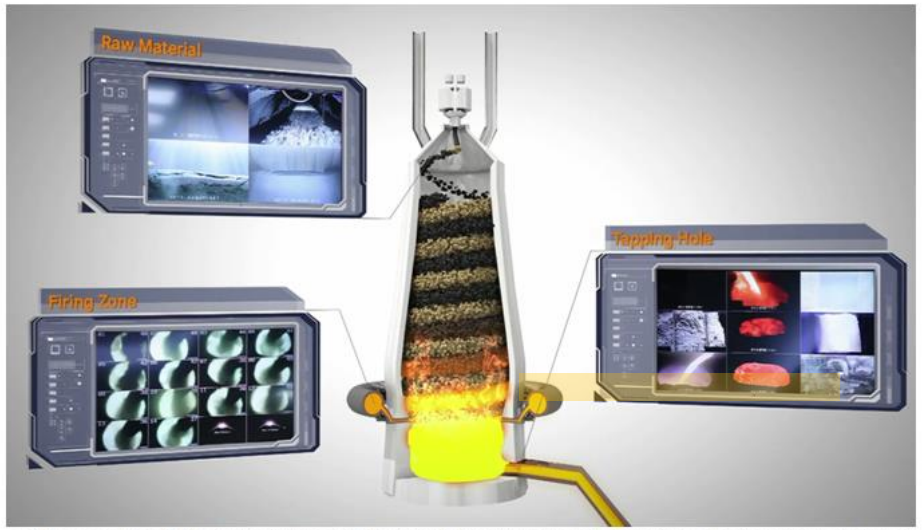
POSCO Smart Blast Furnace



포스코 포항제철소 고로에 '설명 가능 인공지능' 기술 적용한다

2고로와 3고로에 XAI 확대 적용해 품질, 생산성 향상 기대

김재광 기자(=경북) stmkjki@polinews.co.kr 등록 2020.02.10 16:16:16



▲ 스마트 고로 내부 일러스트. 이미지 빅데이터 기반 인공지능으로 고로 상부에서는 통기성, 중부에서는 연소성, 하부에서는 용선온도를 스스로 제어해 쇳물을 생산한다. <사진제공=포스코>

포스코가 '인공지능 용광로'로 불리는 포항제철소 2고로와 3고로에 '설명 가능 인공지능'(XAI·Explainable AI) 기술을 적용한다.

XAI는 인공지능이 의사결정을 내린 이유를 설명해주는 시스템으로 인공지능의 활용성을 높일 수 있는 차세대 AI기술이다.

기존 인공지능 시스템은 주어진 자료를 정확히 분석하고 예측할 수 있으나 그 결과에 대한 원인을 알기 어려운 단점이 있었다. 그러나 XAI는 결과에 대한 핵심 원인을 파악 할 수 있어 인공지능에 대한 신뢰성을 획기적으로 높일 수 있다.

The First International Standard on XAI initiated from Korea

ISO/IEC JTC 1/SC 42 **N 782**

ISO/IEC JTC 1/SC 42 "Artificial intelligence"
Secretariat: **ANSI**
Committee Manager: **Benko Heather Ms.**

Official Form 4 - NP - Information technology -- Artificial intelligence -- Objectives and methods for explainability of ML models and AI systems

Document type	Related content	Document date	Expected action
Ballot / Reference document	Project: ISO/IEC NP TS 6254 Ballot: ISO/IEC NP TS 6254 (restricted access)	2020-11-16	**VOTE** by 2021-02-09

Description

SC 42 N 782 is a NP for ballot to approve the proposal "Information technology -- Artificial intelligence -- Objectives and methods for explainability of ML models and AI systems" and has also been issued via the electronic balloting procedure with the ballot opening on 17 November 2020. SC 42 N 711 is the Draft Document related to the Form 4 contained in SC 42 N 782. Votes should be submitted by 9 February 2021. Any comments submitted with votes should be provided in the standard format.

韓, 인공지능 국제표준화 주도권 강화한다

조명의 기자 | 승인 2020.11.04 17:45

[테크월드=조명의 기자]

인공지능(AI) 국제 표준화 회의에서 우리나라는 한국판 뉴딜 정책의 핵심인 인공지능의 표준화를 위해 인공지능 데이터의 프레임워크와 서비스 생태계, 머신러닝 데이터 품질, 인공지능 신뢰성 등에서 국제표준 논의를 주도했다.

국립전파연구원
National Radio Research Agency

과학기술정보통신부 국립전파연구원과 산업통상자원부 국가기술표준원은 10월 20일부터 30일까지 온라인으로 개최된 '제6차 인공지능 국제 표준화 회의(ISO/IEC JTC1/SC42)'에 산·학·연·관 전문가 33명이 우리나라 대표단으로 참가한 결과, ▲인공지능 서비스 생태계 표준화를 위한 신규 특별작업반 설립, ▲설명가능한 인공지능(XAI)의 신규 국제표준 제안(NP), ▲신러닝 데이터 품질 신규 국제표준안 작업 지속 등의 성과를 거뒀다고 밝혔다.

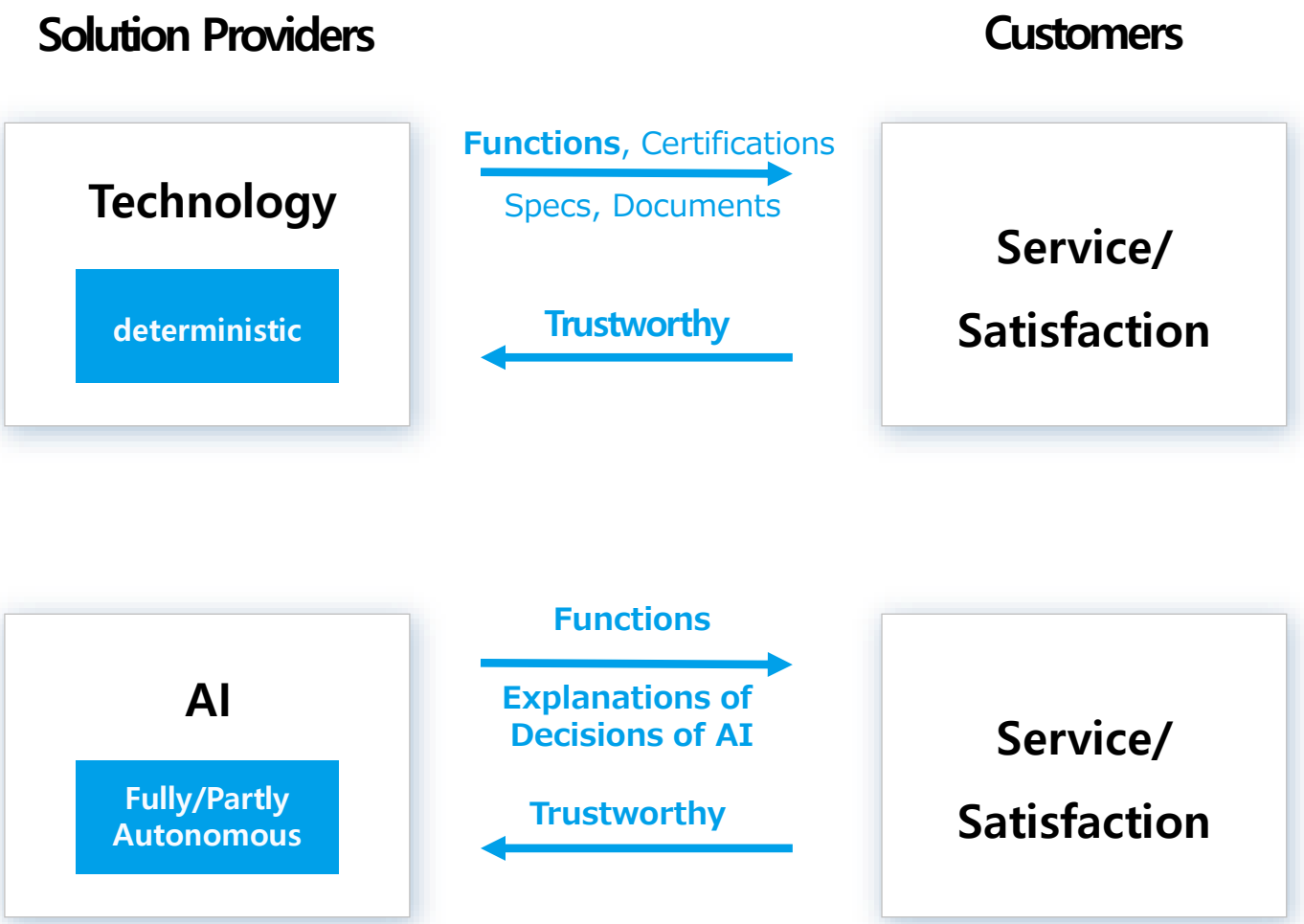
아울러 우리 대표단은 지난 1년 동안 우리나라가 주도한 인공지능 데이터 특별작업반 운영 결과를 공유했으며, '인공지능 데이터 프레임워크'에 관한 신규 국제표준안 제안을 위해 논의를 이어갈 계획이다.

이외에도 '설명가능한 인공지능 시스템 개발 지침'을 신규 표준화 과제로 제안(서울시립대 이재호 교수)했으며, 2021년 초에 신규 과제로 최종 채택될 예정이다.

설명가능한 인공지능은 예를 들면, 인공지능을 이용한 금융대출 심사결과에 대해 그 결정 과정과 이유 등을 소비자에게 설명해 주는 등 인공지능의 신뢰성을 높이는 기술이다.

Existing Services vs AI Services

기존 제품 vs AI 제품



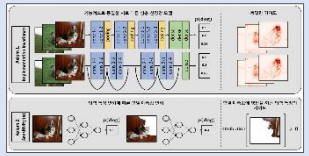
One of the Most Accurate XAI Algorithms – INEEJI/KAIST

April 2022 ~ December 2026 (57 months)

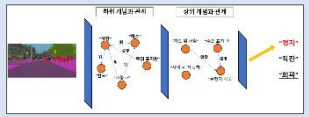
[1] Plug and Play XAI

AI Model Explanation

Input Explanation

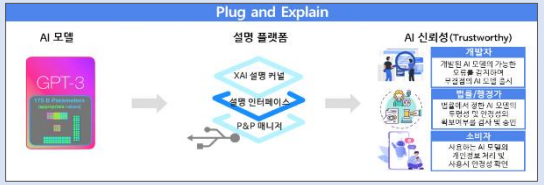


Internal Explanation



<플러그앤플레이 설명가능 인공지능 개발을 위한 기관별 역할>

AI Model Plug and Play XAI



General Method to Explain AI Models

Test Applications

Tech Collaborations

Explaining Internal Decision of AI

Collaobrators

Industries

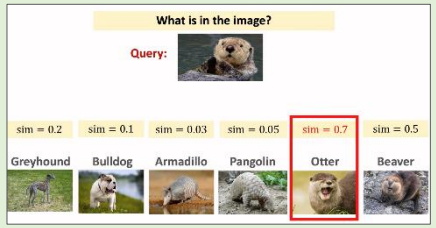
posco, NAVER, linewalks, SAMSUNG, neurophet, SK telecom, HYUNDAI, SAMSUNG, 삼성서울병원, INEEJI, 하나금융TI, kakaobrain

International Collaboration

EPFL, 上海科技大学 ShanghaiTech University, UCL, PHBS 北京大学汇丰商学院, TUM, VinAI RESEARCH, BIFOLD, Fraunhofer HHI, IBM, UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN, UMKC, University of Nottingham UK | CHINA | MALAYSIA

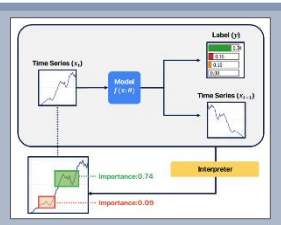
[2] Causal Inference Explanation

Causality Discovery



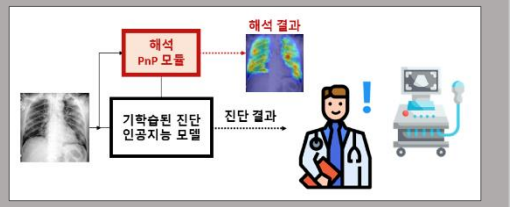
Few Shot based XAI LEarning

[3] XAI for General Audience



Explanations with Commonsense Knowledge for General Public

[4] XAI for Experts



Explanations for Expert to Improve the Work Performance

Learn, Practice and Generate Knowledge to Solve Some of the World's Greatest Problems in AI.

jaesik.choi@kaist.ac.kr

http://xai.kaist.ac.kr/