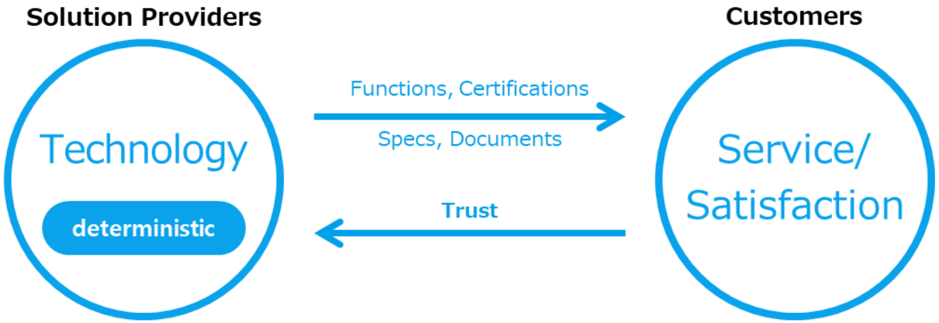
Principles and Applications of Explainable Artificial Intelligence

Jaesik Choi

Director, Explainable Artificial Intelligence Center, KAIST
CEO, INEEJI Corp.

Why Explainable AI(XAI)?

Existing Services

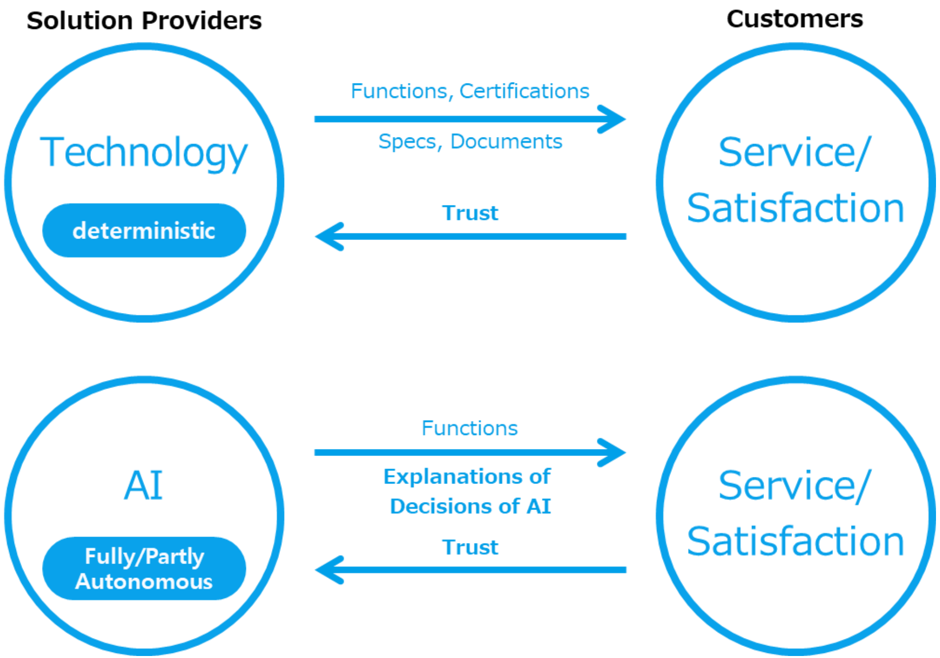


Why Explainable AI(XAI)?

Existing Services

VS

AI Services

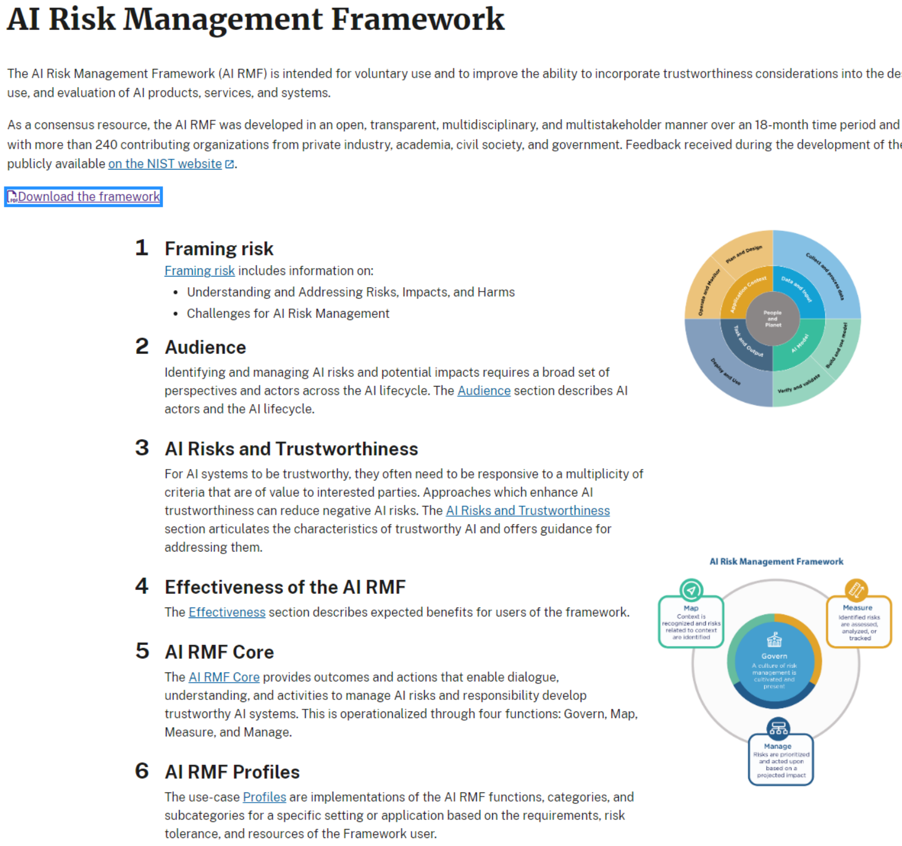


EU: General Data Protection Regulations (GDPR)

Items	Contents
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

US: NIST AI Risk Management Framework



AI Risk Management Framework

The AI Risk Management Framework (AI RMF) is intended for voluntary use and to improve the ability to incorporate trustworthiness considerations into the design, development, use, and evaluation of AI products, services, and systems.

As a consensus resource, the AI RMF was developed in an open, transparent, multidisciplinary, and multistakeholder manner over an 18-month time period and in collaboration with more than 240 contributing organizations from private industry, academia, civil society, and government. Feedback received during the development of the AI RMF is publicly available on the NIST website.

Download the framework

1 **Framing risk**

Framing risk includes information on:

- Understanding and Addressing Risks, Impacts, and Harms
- Challenges for AI Risk Management

2 **Audience**

Identifying and managing AI risks and potential impacts requires a broad set of perspectives and actors across the AI lifecycle. The Audience section describes AI actors and the AI lifecycle.

3 **AI Risks and Trustworthiness**

For AI systems to be trustworthy, they often need to be responsive to a multiplicity of criteria that are of value to interested parties. Approaches which enhance AI trustworthiness can reduce negative AI risks. The AI Risks and Trustworthiness section articulates the characteristics of trustworthy AI and offers guidance for addressing them.

4 **Effectiveness of the AI RMF**

The Effectiveness section describes expected benefits for users of the framework.

5 **AI RMF Core**

The AI RMF Core provides outcomes and actions that enable dialogue, understanding, and activities to manage AI risks and responsibility develop trustworthy AI systems. This is operationalized through four functions: Govern, Map, Measure, and Manage.

6 **AI RMF Profiles**

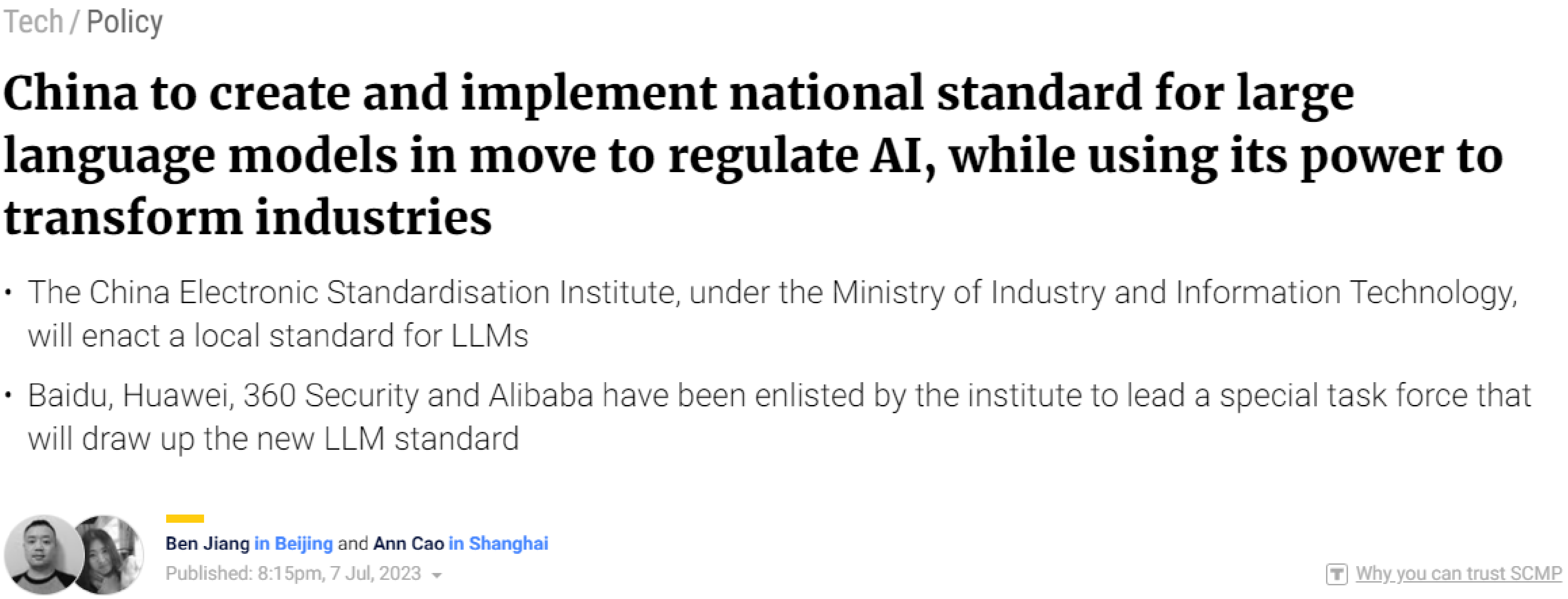
The use-case Profiles are implementations of the AI RMF functions, categories, and subcategories for a specific setting or application based on the requirements, risk tolerance, and resources of the Framework user.

Plan and Design
Collect and process data
Operate and Monitor
Application Context
Data and Input
People and Planet
Task and Output
AI Model
Build and use model
Deploy and Use
Verify and validate
AI Risk Management Framework
Map
Context is recognized and risks related to context are identified
Measure
Identified risks are assessed, analyzed, or tracked
Govern
A culture of risk management is cultivated and present
Manage
Risks are prioritized and acted upon based on a projected impact

- US NIST Established the procedure of Trustworthy and Responsible AI Resource Center(AIRC, https://airc.nist.gov/Home) to support organizations/institutes to build responsible AI (March 2023).
- **AI Risk Management Framework (AI RMF) asses the trustworthiness of AI product, service, design, and implementation.**
- 240 organizations (companies, academia, organizations and government) agree the process.

https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF

China: forming a national standards for testing large language models



Tech / Policy

China to create and implement national standard for large language models in move to regulate AI, while using its power to transform industries

- The China Electronic Standardisation Institute, under the Ministry of Industry and Information Technology, will enact a local standard for LLMs
- Baidu, Huawei, 360 Security and Alibaba have been enlisted by the institute to lead a special task force that will draw up the new LLM standard

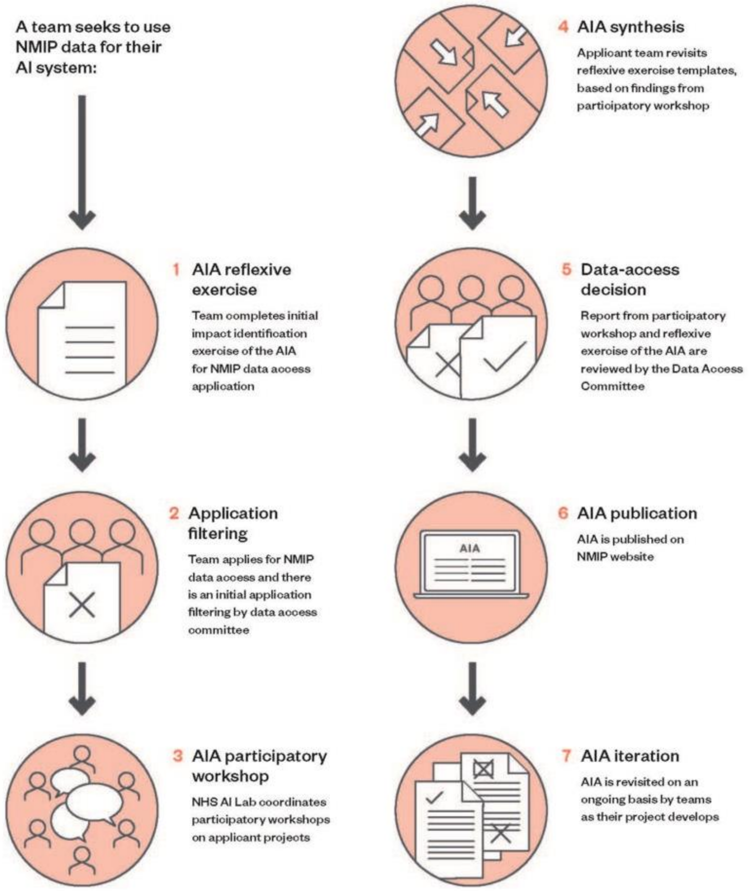
Ben Jiang in Beijing and Ann Cao in Shanghai

Published: 8:15pm, 7 Jul, 2023

Why you can trust SCMP

https://www.scmp.com/tech/policy/article/3226942/china-create-and-implement-national-standard-large-language-models-move-regulate-ai-while-using-its

UK: Impact Assessment of AI Health Care



A team seeks to use NMIP data for their AI system:

1 AIA reflexive exercise

Team completes initial impact identification exercise of the AIA for NMIP data access application

2 Application filtering

Team applies for NMIP data access and there is an initial application filtering by data access committee

3 AIA participatory workshop

NHS AI Lab coordinates participatory workshops on applicant projects

4 AIA synthesis

Applicant team revisits reflexive exercise templates, based on findings from participatory workshop

5 Data-access decision

Report from participatory workshop and reflexive exercise of the AIA are reviewed by the Data Access Committee

6 AIA publication

AIA is published on NMIP website

7 AIA iteration

AIA is revisited on an ongoing basis by teams as their project develops

- **Ada Lovelace Institute conducts the Impact assessment of AI algorithm** (Algorithmic Impact Assesment, AIA) on National Health Service (NHS) UK
- All companies needs approval from the AIA process when they wish to build AI models with NIMP data of NIS. NHS finally select the companies and organization who can access the NIMP data.
- **NHS NMIP AIA process emphasize the accountability and the transparency of AI models.**

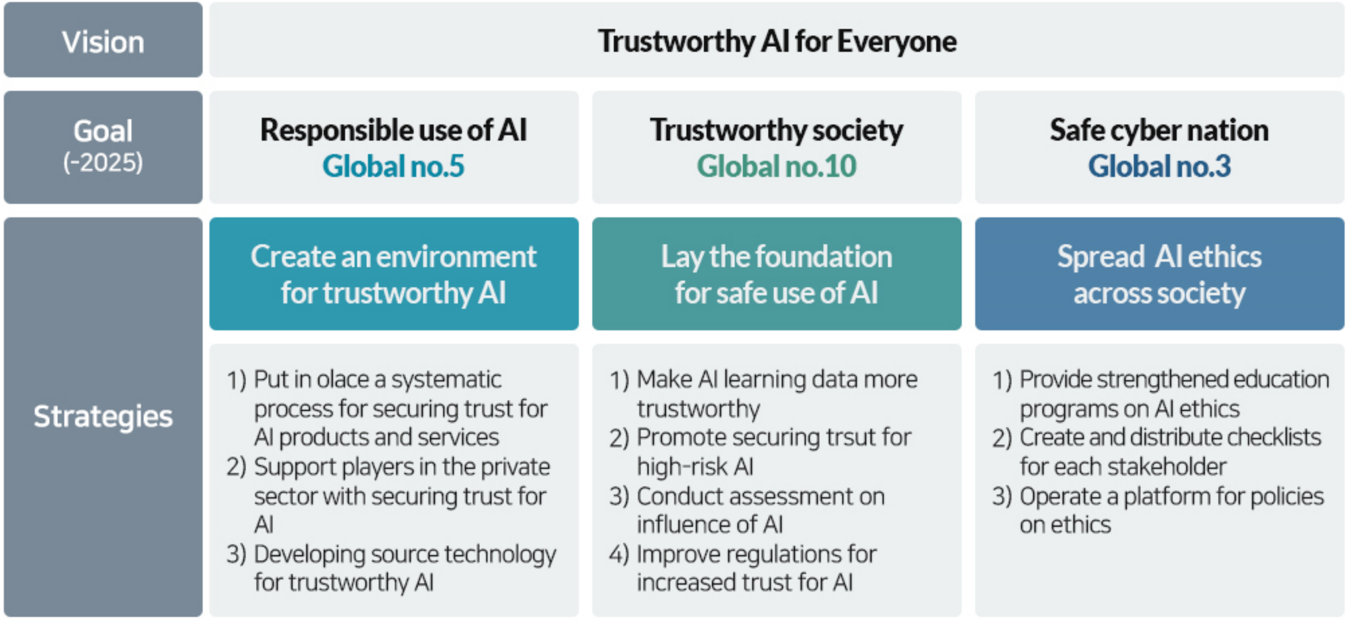
https://www.koraia.org/chtml/board.php?template=base&com_board_basic=file_download&com_board_id=12&com_board_idx=363&com_board_file_seq=0

Strategy to Realize AI Trustworthy in Korea

June 2021

Vision, goals, detailed strategies of trustworthy artificial intelligence

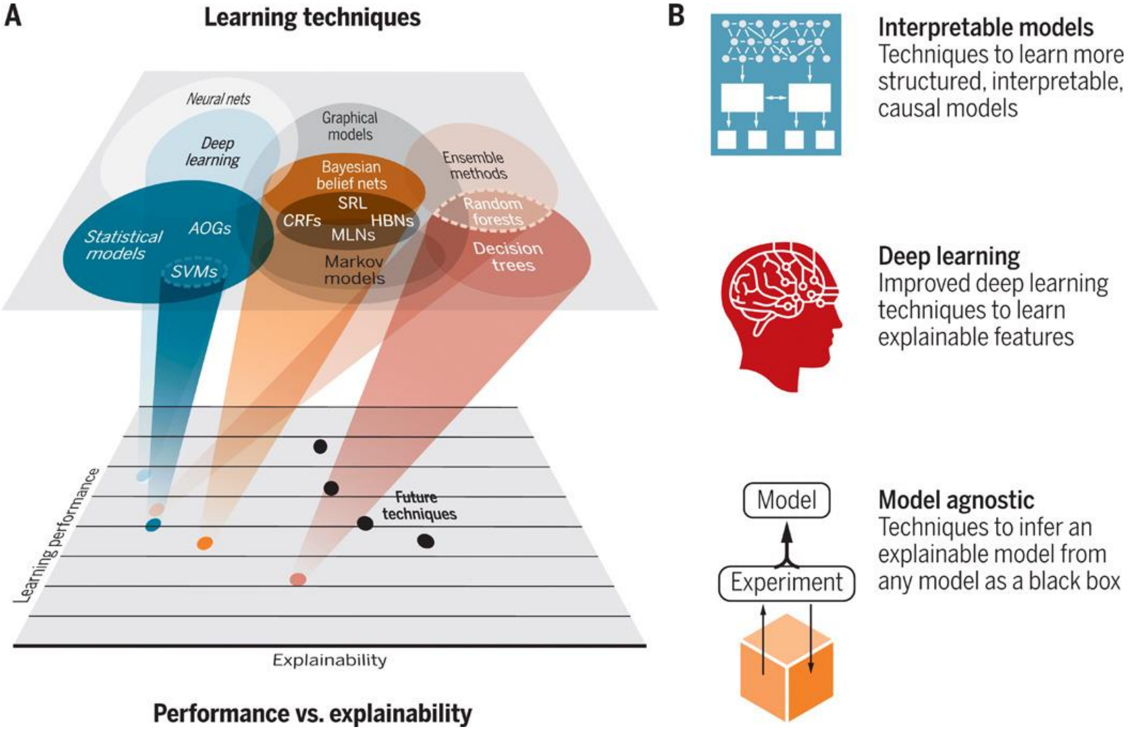
The strategy has the vision of “realize trustworthy artificial intelligence for everyone” and will be implemented step by step until 2025, based on the three pillars of ‘technology, system, ethics’ and 10 action plans.



Vision	Trustworthy AI for Everyone		
Goal (-2025)	Responsible use of AI Global no.5	Trustworthy society Global no.10	Safe cyber nation Global no.3
Strategies	Create an environment for trustworthy AI	Lay the foundation for safe use of AI	Spread AI ethics across society
	1) Put in olace a systematic process for securing trust for AI products and services 2) Support players in the private sector with securing trust for AI 3) Developing source technology for trustworthy AI	1) Make AI learning data more trustworthy 2) Promote securing trsut for high-risk AI 3) Conduct assessment on influence of AI 4) Improve regulations for increased trust for AI	1) Provide strengthened education programs on AI ethics 2) Create and distribute checklists for each stakeholder 3) Operate a platform for policies on ethics

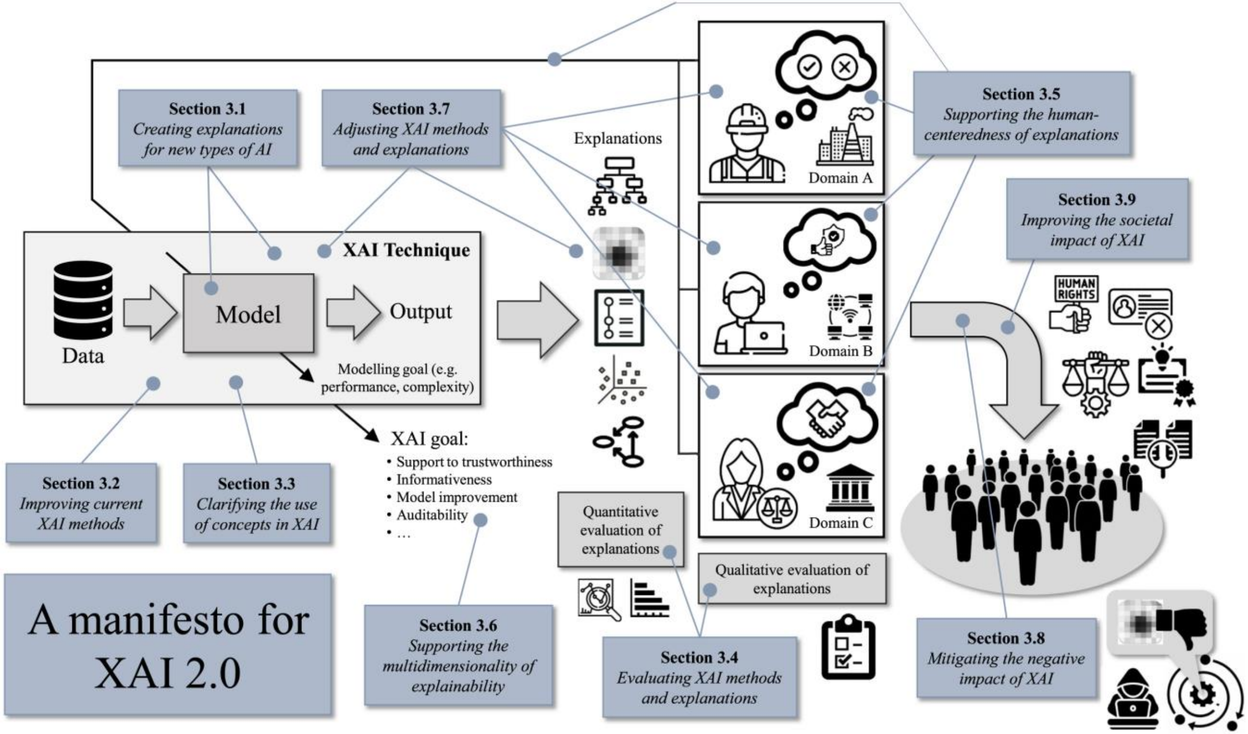
UK	▸ Established 5 codes of ethics (2018 Apr), a guide to using AI in the public sector (2019 June.), a guideline for explainable AI (2020 May)

XAI – Explainable Artificial Intelligence



- David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf and Guang-Zhong Yang, *XAI—Explainable artificial intelligence*, **Science Robotics**, 4(37), 2019.

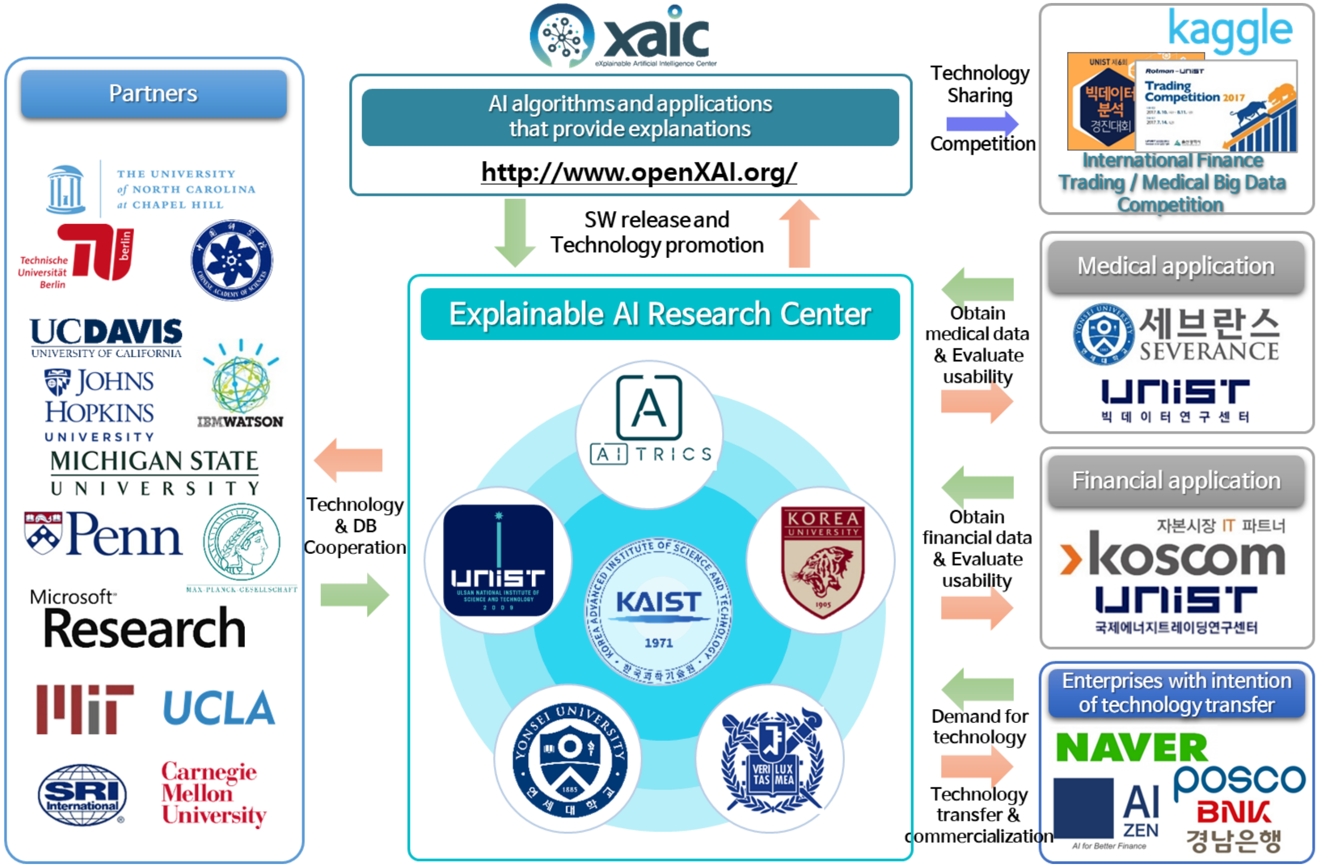
Explainable Artificial Intelligence (XAI) 2.0:
A manifesto of open challenges and interdisciplinary research directions


Section 3.1
Creating explanations for new types of AI
Section 3.7
Adjusting XAI methods and explanations
Explanations
Domain A
Section 3.5
Supporting the human-centeredness of explanations
Section 3.9
Improving the societal impact of XAI
XAI Technique
Data
Model
Output
Modelling goal (e.g. performance, complexity)
Domain B
HUMAN RIGHTS
XAI goal:
• Support to trustworthiness
• Informativeness
• Model improvement
• Auditability
• …
Section 3.2
Improving current XAI methods
Section 3.3
Clarifying the use of concepts in XAI
Quantitative evaluation of explanations
Domain C
Qualitative evaluation of explanations
A manifesto for XAI 2.0
Section 3.6
Supporting the multidimensionality of explainability
Section 3.4
Evaluating XAI methods and explanations
Section 3.8
Mitigating the negative impact of XAI

- Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, Richard Jiang, Hassan Khosravi, Freddy Lecue, Gianclaudio Malgieri, Andrés Páez, Wojciech Samek, Johannes Schneider, Timo Speith and Simone Stumpf, ***Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions***, *Information Fusion*, 2024.

Explainable AI Program in Korea

July 2017
- Present



xaic
eXplainable Artificial Intelligence Center

AI algorithms and applications that provide explanations

http://www.openXAI.org/

SW release and Technology promotion

Explainable AI Research Center

Partners

Technology & DB Cooperation

Technology Sharing

Competition

International Finance Trading / Medical Big Data Competition

Medical application

Obtain medical data & Evaluate usability

Financial application

Obtain financial data & Evaluate usability

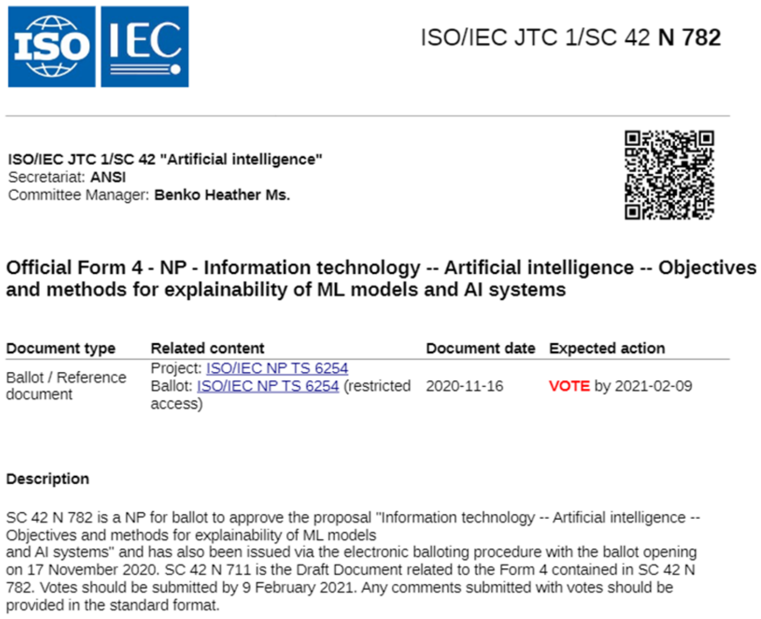
Enterprises with intention of technology transfer

Demand for technology

Technology transfer & commercialization

International Standard of XAI

Participant	Title	Organization	Stage	Number	Date	Country
Jaeho Lee	Objectives and methods for explainability of ML models and AI systems	ISO/IEC JTC 1/SC 42	NP	ISO/IEC NP TS 6254	2020-11-16	Switzerland



ISO IEC

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.

국립전파연구원
National Radio Research Agency

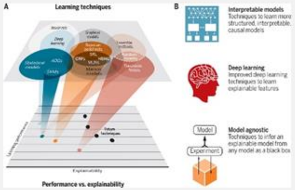
- **The First International Standard on XAI Initiated by Korea**

Explainable AI Program in Korea

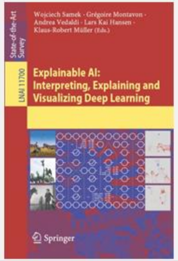
Research Results

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

Top Journal Papers (Science Robotics, …) 45


A
Learning techniques
Deep learning
Graphical models
Ensemble methods
Statistical models
AOGs
SRL
CRFs
HBNs
MLNs
Markov models
SVMs
Future techniques
Explainability
Performance vs. explainability
B
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
Model
Experiment

Book Edited (Explainable AI: Springer)


Wojciech Samek · Grégoire Montavon · Andrea Vedaldi · Lars Kai Hansen · Klaus-Robert Müller (Eds.)

State-of-the-Art Survey

LNAI 11700

Explainable AI: Interpreting, Explaining and Visualizing Deep Learning

Springer

Technology Transfer

Patents (Registration)	37(2)
Industrial Projects	11

Manufacturing	Healthcare	Finance	Mobile
posco Steel SAMSUNG 삼성전자 Semiconductor	세브란스병원 SEVERANCE HOSPITAL 강남세브란스병원 GANGNAM SEVERANCE HOSPITAL 국민건강보험 일산병원 National Health Insurance Service Ilsan Hospital SNUH 분당서울대학교병원	IBK 기업은행 신한은행 KEB 하나은행	NAVER SAMSUNG 삼성전자
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

KDD2020 Tutorial on

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

One of the Most Accuracte XAI Technique

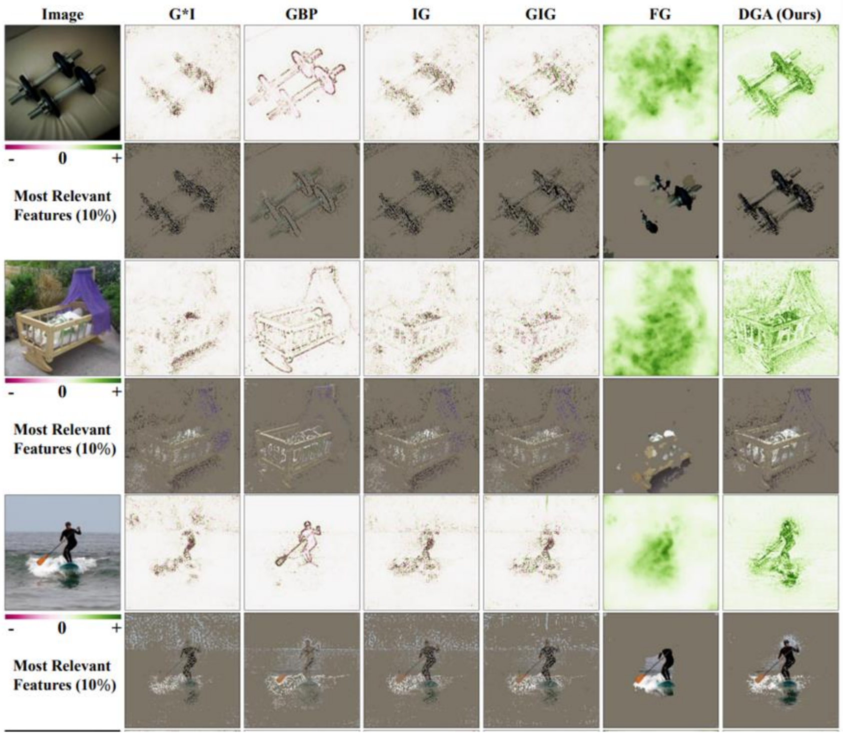


Image
G*I
GBP
IG
GIG
FG
DGA (Ours)
- 0 +
Most Relevant Features (10%)
- 0 +
Most Relevant Features (10%)
- 0 +
Most Relevant Features (10%)

Google TensorFlow
Explainable AI Toolkit

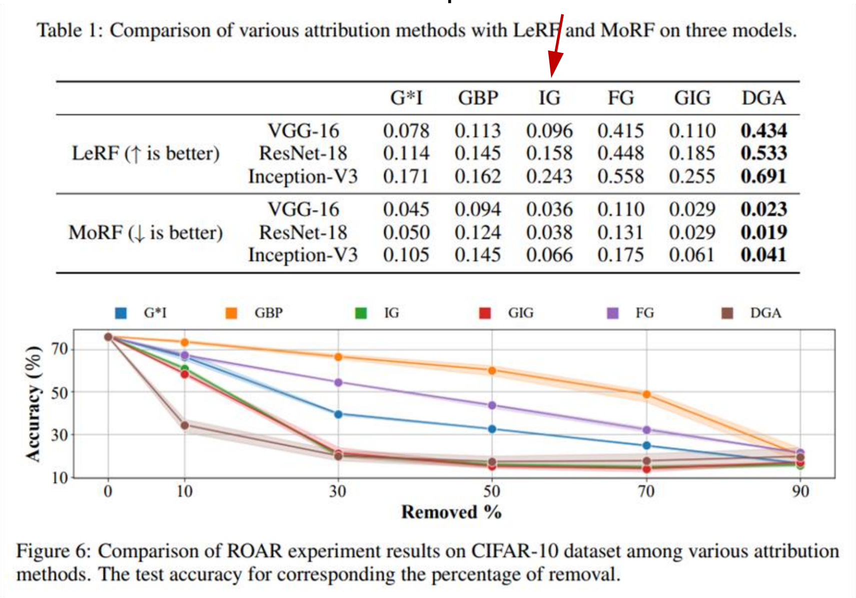


Table 1: Comparison of various attribution methods with LeRF and MoRF on three models.

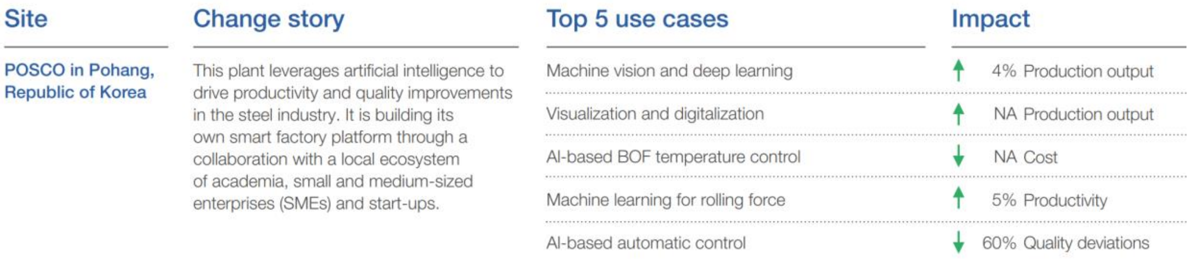
		G*I	GBP	IG	FG	GIG	DGA
LeRF (↑ is better)	VGG-16	0.078	0.113	0.096	0.415	0.110	**0.434**
	ResNet-18	0.114	0.145	0.158	0.448	0.185	**0.533**
	Inception-V3	0.171	0.162	0.243	0.558	0.255	**0.691**
MoRF (↓ is better)	VGG-16	0.045	0.094	0.036	0.110	0.029	**0.023**
	ResNet-18	0.050	0.124	0.038	0.131	0.029	**0.019**
	Inception-V3	0.105	0.145	0.066	0.175	0.061	**0.041**

Figure 6: Comparison of ROAR experiment results on CIFAR-10 dataset among various attribution methods. The test accuracy for corresponding the percentage of removal.

Giyoung Jeon (INEEJI), Haedong Jeong, Jaesik Choi,
Distilled Gradient Aggregation: Purify Features for Input
Attribution in the Deep Neural Network, NeurIPS, 2022

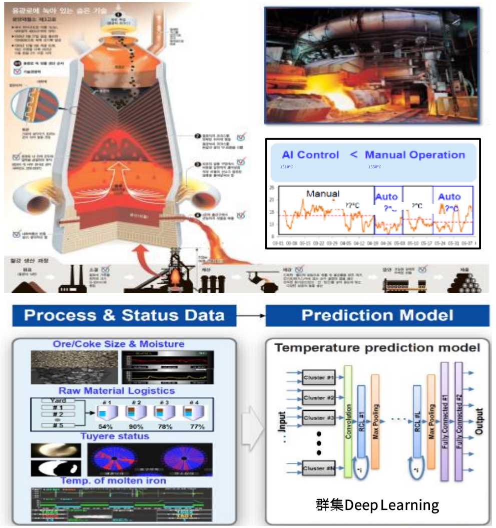
Success stories of INEEJI (Start-up Team) clients

Partner	Contents	Media	Year
POSCO	Using deep learning technology, artificial intelligence has created a "Smart Blast Furnace" that learns, predicts, and manages data. In the past, people used to check the temperature with pictures taken every two hours, but now they can predict and automatically control even the heat after an hour using an algorithm called deep learning. The Pohang 2nd furnace, where this technology is applied, has increased the amount of iron produced per day by 240 tons. It can produce 85,000 medium-sized cars annually. In fact, the average annual production of Pohang 2nd furnace improved by 5% compared to the previous one, and fuel costs were reduced by 1%.	ChosunIlbo	2021



Site	Change story	Top 5 use cases	Impact
POSCO in Pohang, Republic of Korea	This plant leverages artificial intelligence to drive productivity and quality improvements in the steel industry. It is building its own smart factory platform through a collaboration with a local ecosystem of academia, small and medium-sized enterprises (SMEs) and start-ups.	Machine vision and deep learning	↑ 4% Production output
		Visualization and digitalization	↑ NA Production output
		AI-based BOF temperature control	↓ NA Cost
		Machine learning for rolling force	↑ 5% Productivity
		AI-based automatic control	↓ 60% Quality deviations

https://www3.weforum.org/docs/WEF_Global_Lighthouse_Network.pdf



AI prediction model for Acute Kidney Injury

SNUH 서울대학교병원

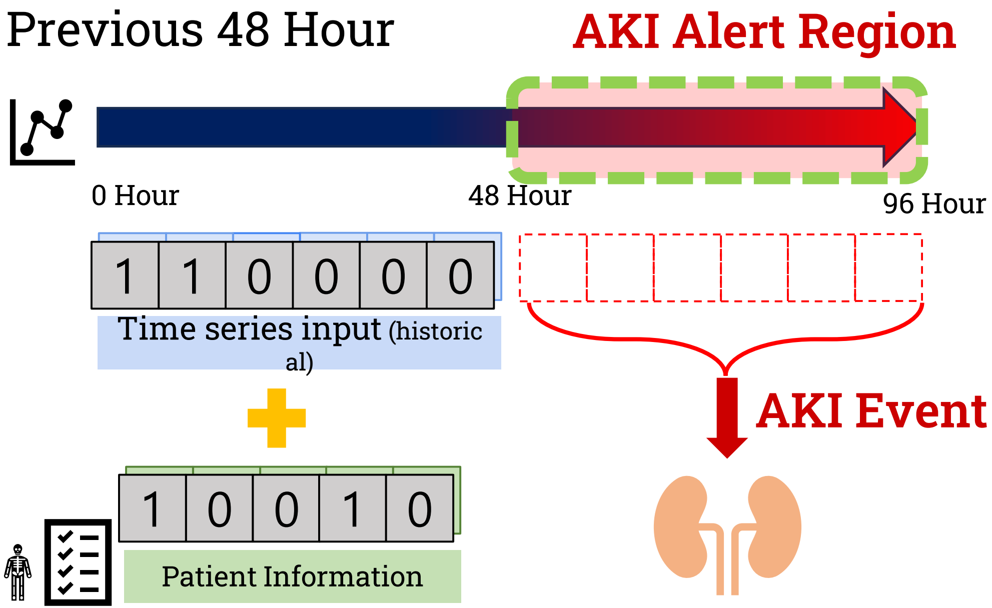
- Developing model
- Internal validation

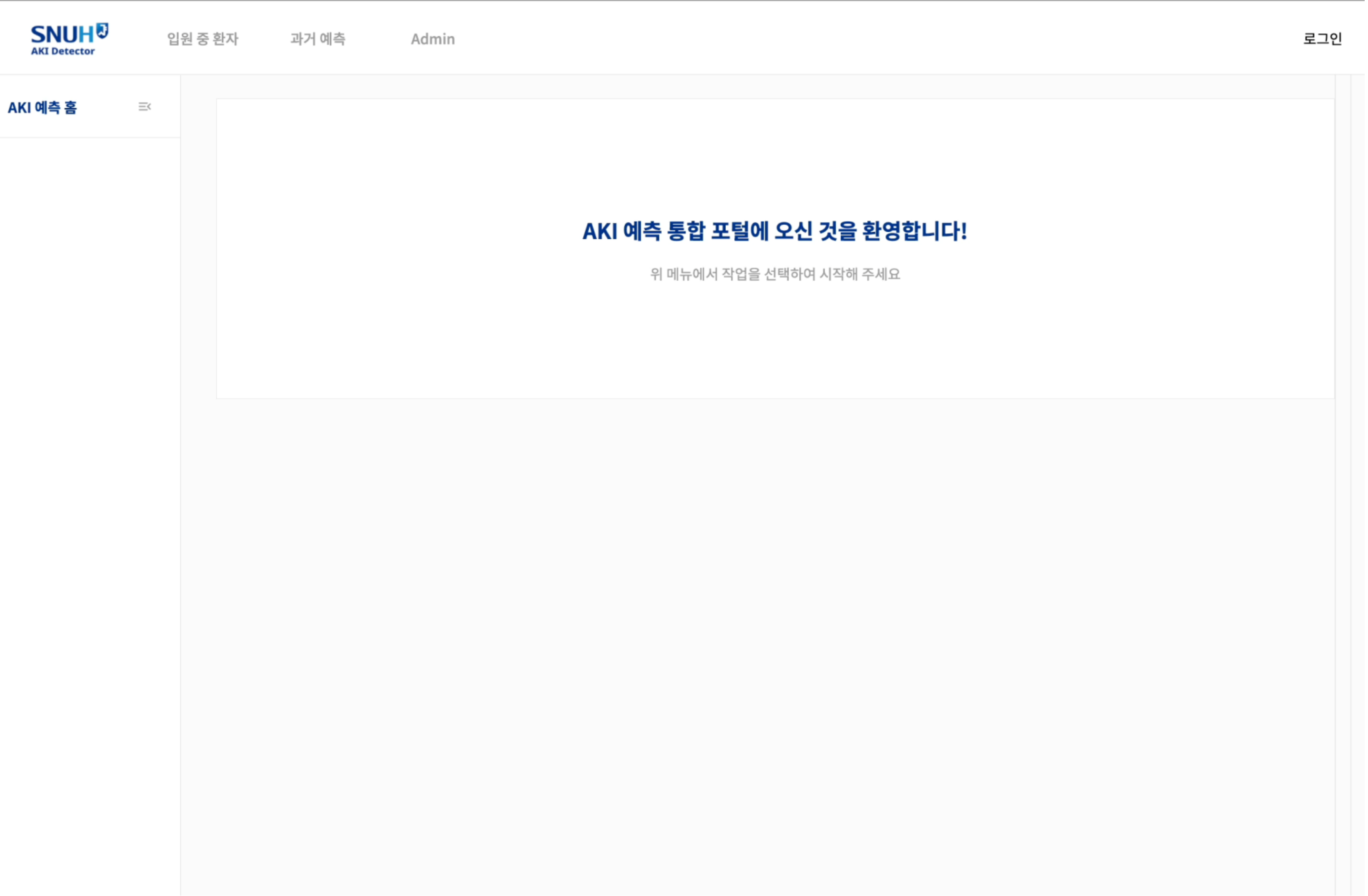
SNUH 분당서울대학교병원

- External validation
- Clinical Test

Current day	Patient info	Dynamic TS	Target (AKI)	Stay length
2	P_x	day(1,2)	day(3,4)	5
3		day(2,3)	day(4,5)	
4		day(3,4)	day(5,6)	

🟦 Used on prediction





AKI 예측 통합 포털에 오신 것을 환영합니다!

위 메뉴에서 작업을 선택하여 시작해 주세요

AI prediction model for Acute Kidney Injury

The training set comprised data from 183,221 patients at Seoul National University Hospital (2013-2017).

At Seoul National University Bundang Hospital (2020-2021),
we randomly selected 74 patients from departments with high AKI rates, including 15% AKI cases.

Accuracy: physician: 0.797, student: 0.574, AI: 0.568.

AI assistance improved recall and F1 scores: recall: 52.4% to 71.4%, F1: 37.7% to 46.1%.

In the AKI predicted group,
- recall increased while F1 decreased for physicians (recall: 36.4% to 60%, F1: 43.2% to 33.3%) and students (recall: 54.5% to 80%, F1: 44.4% to 36.9%).

For the non-AKI predicted group,
- both saw significant gains in recall and F1 with AI (physicians: recall 16.7% to 87.5%, F1: 18.2% to 66.7%; students: recall 44.4% to 75%, F1: 21.1% to 40%).

Thank you

jaesik.choi@kaist.ac.kr