GISC2020
Global ICT Standards Conference

뉴노멀시대 선도를 위한
ICT 표준의 역할

AI 신의성(Trustworthiness) 향상을 위한

설명가능한 AI (XAI) 기술 및 표준화 동향

이재호

서울시립대학교
2020-10-15

뉴 노멀 시대
선도를 위한
ICT 표준의
역할

Self-driving car

- Autonomous vs. automated
  - Automated:
    - control or operation by a machine
  - Autonomous:
    - acting alone or independently, self-governing

SAE (J3016) Automation Levels			
SAE Level	Name	Property	Execution of steering and acceleration/deceleration
Human driver monitors the driving environment			
0	No Automation		Human driver
1	Driver Assistance	Hands on	Human driver & system
2	Partial Automation	Hands off	System
Automated driving system monitors the driving environment			
3	Conditional Automation	Eyes off	System
4	High Automation	Mind off	System
5	Full Automation	Steering wheel optional	System

Trustworthiness

- 신뢰[신용]할 수 있음, 기댈[의지할] 수 있음. (Daum 영어사전)
- the quality of always being good, honest, sincere, etc. so that people can rely on you (Oxford Learner's Dictionaries)
- 안전·신뢰성: 시스템이 의도된 기능과 성능을 발휘할 수 있도록 안전성(safety), 신뢰성(reliability), 복구성(resilience), 보안성(security) 등을 포함하는 관심사. (TTA 정보통신용어사전)

Reliability

- 신뢰성, 신뢰도, 신뢰할 수 있음 (Daum 영어사전)
- the quality of being able to be trusted to do what somebody wants or needs (Oxford Learner's Dictionaries)
- 신뢰성/신뢰도: 명시된 기간 동안 주어진 조건 하에서 시스템이나 컴포넌트가 요구된 기능을 수행할 수 있는 능력 (TTA 정보통신용어사전)

Generic ‘de Jure’ definitions of trustworthiness

(JTC 1/AG 7, WG 13)

- Trust: degree to which a user or other stakeholder has confidence that a product or system will behave as intended
  - ISO/IEC 25010:2011(en) Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models
- Trustworthy ability to demonstrate authenticity, integrity and availability of ESI (3.1) over time
  - ISO/TR 15801:2017(en) Document management — Electronically stored information — Recommendations for trustworthiness and reliability
- Trustworthy stored electronically in an accurate, reliable and usable/readable manner, ensuring integrity over time
  - note 1 to entry: See ISO/TR 15801. ISO 18829:2017(en) Document management — Assessing ECM/EDRM implementations
- Trustworthy data (3.16) and related information (3.23) that is accurate, complete, relevant, readily understood by and available to those authorised users who need it to complete a task
  - ISO/IEC 19770-1:2017(en) Information technology — IT asset management — Part 1: IT asset management systems — Requirements
- likelihood that an entity will behave as expected. In the context of industrial automation, attributes of trustworthiness include reliability, security, and resiliency
  - IEC 62918, ed. 1.0 (2014-07) TC 45a
- "degree of confidence a stakeholder has that the system performs as expected with characteristics including safety, security, privacy, reliability and resilience in the face of environmental disruptions, human errors, system faults and attacks."
  - ISO/IEC FDIS 20924 IoT - Vocabulary SC41

IEC White Paper: Artificial intelligence across industries (2018)

- It covers current technological capabilities and provides a detailed description of some of the major existing and future challenges related to safety, security, privacy, trust and ethics that AI will have to address at the international level.
- Artificial intelligence challenges
  - 6.1 Social and economic challenges
    - Changes in decision-making
    - Advanced supply chain operations
  - 6.2 Data-related challenges
    - Selection of training data
    - Standardized data
  - 6.3 Algorithm-related challenges
    - Robustness
    - Transfer learning
    - Interpretability
    - Objective functions
  - 6.4 Infrastructure-related challenges
    - Hardware bottlenecks
    - Lack of platforms and frameworks
  - 6.5 Trustworthiness-related challenges
    - Trust
    - Privacy
    - Security
  - 6.6 Regulatory-related challenges
    - Liability
    - Privacy
    - Ethics

ISO/IEC/IEEE 15026 Systems and software assurance

- Part 1: Concepts and vocabulary
  - 3.1.1 assurance
    - grounds for justified confidence that a claim (3.1.4) has been or will be achieved
  - 3.1.4 claim
    - true-false statement about the limitations on the values of an unambiguously defined property — called the claim's property — and limitations on the uncertainty of the property's values falling within these limitations during the claim's duration of applicability under stated **conditions** (3.1.5)
    - A claim potentially contains the following:
      - property of the system-of-interest;
      - limitations on the value of the property associated with the claim (e.g., on its range);
      - limitations on the uncertainty of the property value meeting its limitations;
      - limitations on duration of claim's applicability;
      - duration-related uncertainty;
      - limitations on conditions associated with the claim; and
      - condition-related uncertainty.
  - 3.1.7 dependability
    - <of an item> ability to perform as and when required
    - Note 1 to entry: Dependability includes availability, reliability, recoverability, maintainability, and maintenance support performance, and, in some cases, other characteristics such as durability, safety and security.
    - Note 2 to entry: Dependability is used as a collective term for the time-related quality characteristics of an item.
    - [SOURCE: IEC 60050-192:2015, 192-01-22]

JTC 1/SC 42/WG 3: ISO/IEC 24028 Trustworthiness

- Trustworthiness
  - ability to meet stakeholders’ expectations in a verifiable way

    Note 1 to entry: Depending on the context or sector, and also on the specific product or service, data, and technology used, different characteristics apply and need verification to ensure stakeholders expectations are met.

    Note 2 to entry: Characteristics of trustworthiness include, for instance, reliability, availability, resilience, security, privacy, safety, accountability, transparency, integrity, authenticity, quality, usability.

    Note 3 to entry: Trustworthiness is an attribute that can be applied to services, products, technology, data and information as well as, in the context of governance, to organizations.

- Scope

  This document surveys topics related to trustworthiness in AI systems, including the following:
  - approaches to establish trust in AI systems through transparency, explainability, controllability, etc.;
  - engineering pitfalls and typical associated threats and risks to AI systems, along with possible
  - mitigation techniques and methods; and
  - approaches to assess and achieve availability, resiliency, reliability, accuracy, safety, security, and privacy of AI systems.

JTC 1/WG 13 Trustworthiness (2018)

- *Complete, improve and maintain the inventory (JTC 1 N14500) including the heat map as a JTC 1 standing document reflecting the landscape of trustworthiness in JTC 1, other ISO and IEC Committees, and other SDOs*
- *Complete terminology and description of characteristics and determine what type of document should be created.*
- *Develop horizontal deliverables such as frameworks, taxonomy and ontology for ICT trustworthiness for guiding trustworthiness efforts throughout JTC 1 and upon which other deliverables can be developed (beginning with ISO/IEC TS 24462, Ontology for ICT Trustworthiness Assessment)*

GISC2020
Global ICT Standards Conference

뉴 노멀 시대
선도를 위한
ICT 표준의
역할

GISC2020
Global ICT Standards Conference

뉴 노멀 시대
선도를 위한
ICT 표준의
역할

JTC 1/SC 41: ISO/IEC 30149 Trustworthiness principles

- Scope
  - This document provides principles for IoT trustworthiness based on ISO/IEC 30141 – IoT Reference Architecture.
- Base characteristics
  - Safety
  - Security
  - Privacy
  - Resilience
  - Reliability

Explainable Artificial Intelligence (XAI)

Explainable AI (XAI)

- ISO/IEC TR 24028: Overview of trustworthiness in artificial intelligence
  - An attempt to explain could offer multiple different, but equally valid modes of explanation, depending on whether stakeholders seek:
    - a causal understanding of how a result is arrived at,
    - an epistemic understanding of the knowledge on which the result is based, or
    - a justificatory understanding of the grounds in which the result is offered as being valid.
  - The subject of explanation could include the AI system itself and the result produced by the system.

- ISO/IEC 22989: Concepts and terminology
  - property of an AI system to express important factors influencing the AI system results in a way that humans can understand
  - ability to give an argument accounting for the reason for taking a course of action

- Other Sources:
  - Methods and techniques in the application of artificial intelligence technology (AI) such that the results of the solution can be understood by human experts (Wikipedia)
  - Explainable AI is a set of tools and frameworks to help you develop interpretable and inclusive machine learning models and deploy them with confidence (Google cloud XAI)
  - The DARPA Explainable AI (XAI) program aims to create a suite of machine learning techniques that:
    - Produce more explainable models, while maintaining a high level of learning performance (prediction accuracy); and
    - Enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners

Relevant Existing Standard Documents

- ISO/IEC CD 22989 Artificial intelligence — Concepts and terminology
- ISO/IEC TR 24028 Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence
  - 10.3 Explainability
    - 10.3.1 General
    - 10.3.2 Aims of explanation
    - 10.3.3 Ex-ante vs. ex-post explantation
    - 10.3.4 Approaches to explainability
    - 10.3.5 Modes of ex-post explanation
    - 10.3.6 Levels of explainability
    - 10.3.7 Evaluation of the explanations
- ISO/IEC CD TR 24030 Information technology — Artificial Intelligence (AI) — Use cases
  - Examples
    - Explainable Artificial Intelligence for Genomic Medicine
    - Leveraging AI to enhance adhesive quality
    - AI Situation Explanation Service for the Visually Impaired
    - Jet Engine Predictive Maintenance Service

뉴 노멀 시대
선도를 위한
ICT 표준의
역할

Justification

The case for explainable AI: how and why interpretability matters

- Giving users confidence in the system:
- Safeguarding against bias:
- Meeting regulatory standards or policy requirements:
- Improving system design:
- Assessing risk, robustness, and vulnerability:
- Understanding and verifying the outputs from a system:

Explainable AI: the basics [The Royal Society]
(https://royalsociety.org/-/media/policy/projects/explainable-ai/AI-and-interpretability-policy-briefing.pdf)

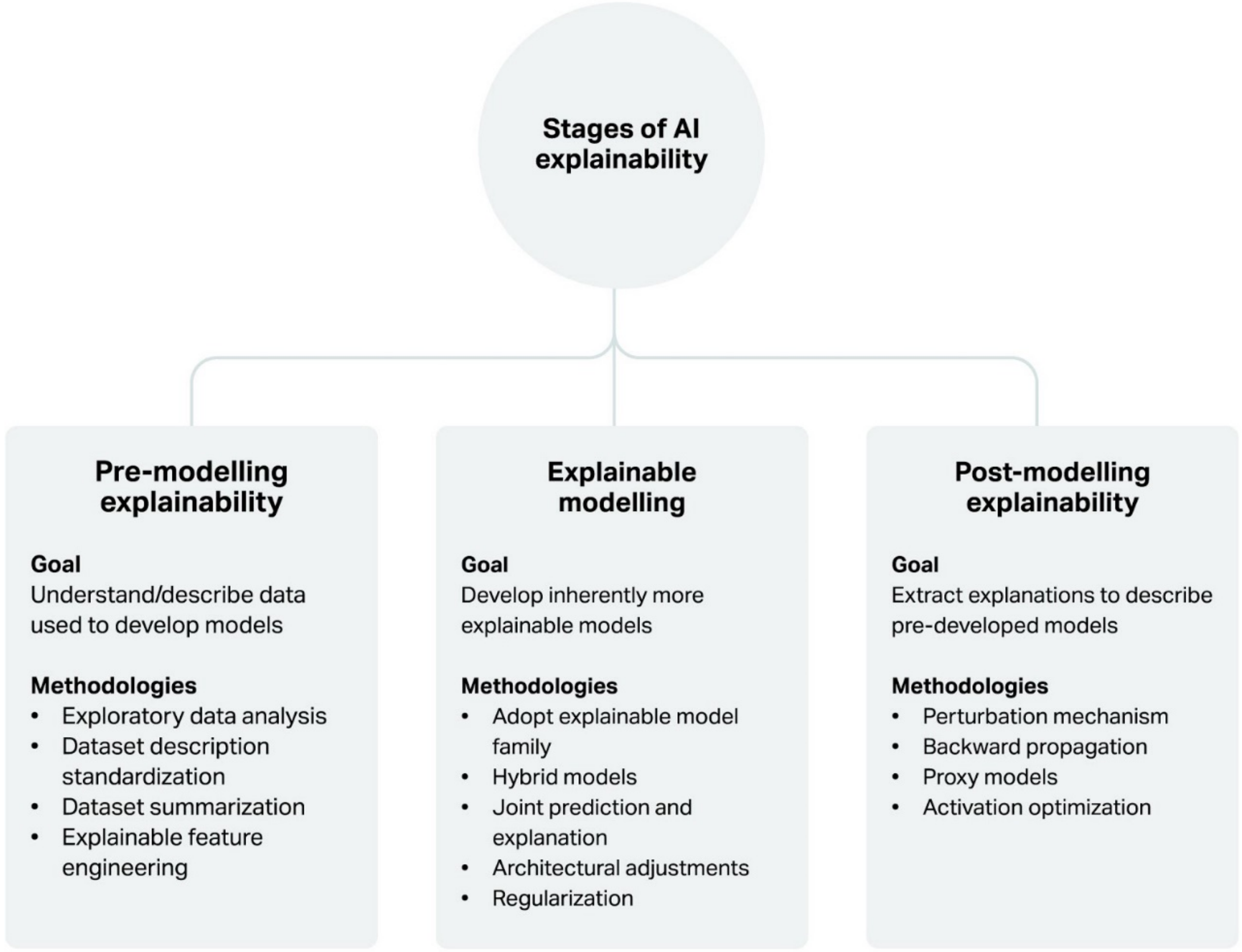
뉴 노멀 시대 선도를 위한 ICT 표준의 역할

Approaches

Type of explanation	Method
• Transparent details of what algorithm is being used	• Publishing the algorithm
• How does the model work?	**Inherently interpretable models** • Use models whose structure and function is easily understood by a human user, e.g. a short decision list. **Decomposable systems** • Structure the analysis in stages, with interpretable focus on those steps that are most important in decision-making. **Proxy models** • Use a second – interpretable – model which approximately matches a complex ‘black box’ system.
• Which inputs or features of the data are most influential in determining an output?	• Visualization or saliency mapping • Illustrate how strongly different input features affect the output from a system, typically performed for a specific data input.
• In an individual case, what would need to change to achieve a different output?	• Counterfactual (or example-based) explanations • Generate explanations focused on a single case, which identify the characteristics of the input data that would need to change in order to produce an alternative output.

Explainable AI: the basics [The Royal Society]

Three stages of AI explainability -- Bahador Khaleghi



Stages of AI explainability

Pre-modelling explainability

Goal
Understand/describe data used to develop models

Methodologies
- Exploratory data analysis
- Dataset description standardization
- Dataset summarization
- Explainable feature engineering

Explainable modelling

Goal
Develop inherently more explainable models

Methodologies
- Adopt explainable model family
- Hybrid models
- Joint prediction and explanation
- Architectural adjustments
- Regularization

Post-modelling explainability

Goal
Extract explanations to describe pre-developed models

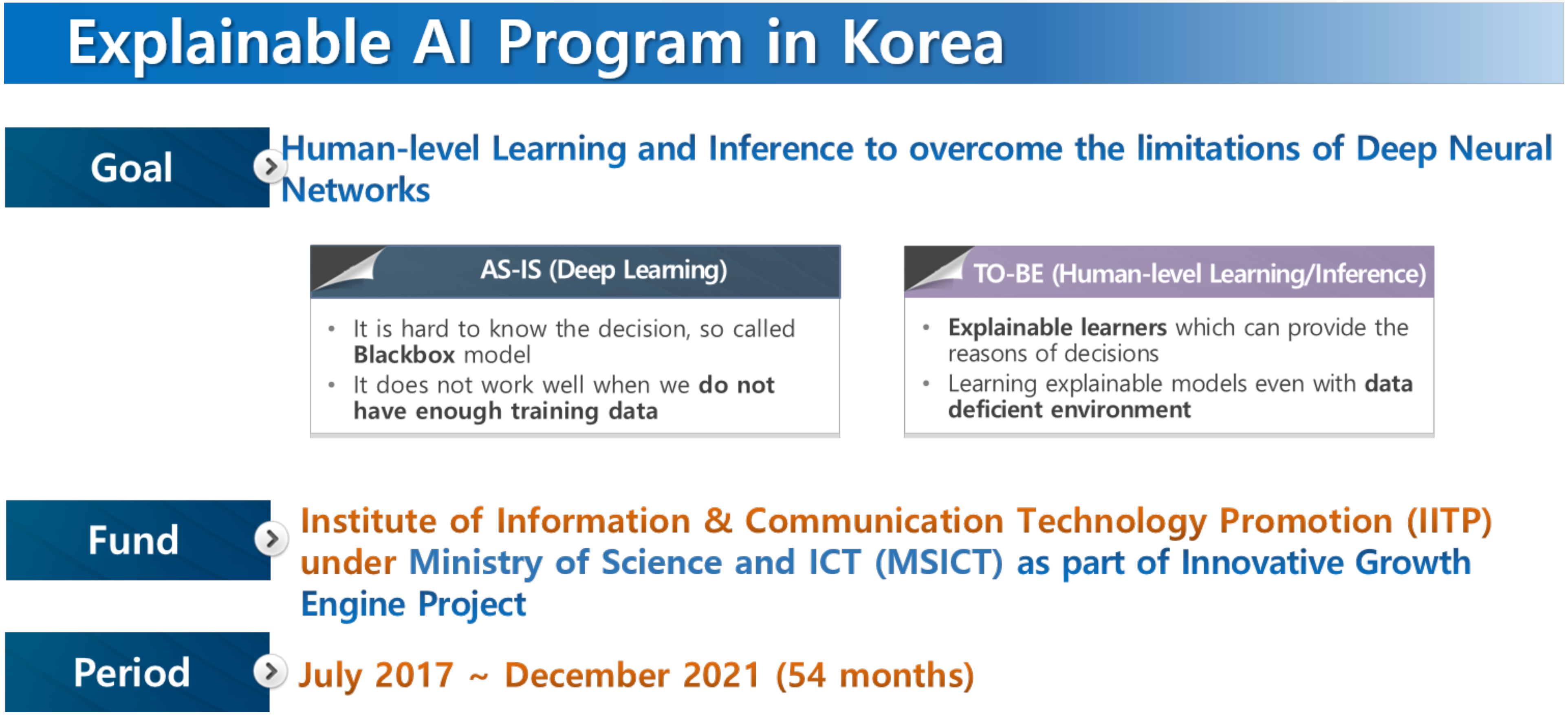
Methodologies
- Perturbation mechanism
- Backward propagation
- Proxy models
- Activation optimization

Regulations

- EU General Data Protection Regulation (GDPR)
  - GDPR Articles 13-15 and 21-22 outline requirements related to automated data processing and decision making.
  - When a decision is generated solely from automated processing (no human intervention), including profiling, the data subject has the right to receive an explanation of how the decision was rendered.
- US Federal Trade Commission (FTC) [17]
  - The FTC's law enforcement actions, studies, and guidance emphasize that the use of AI tools should be transparent, explainable, fair, and empirically sound, while fostering accountability. [Andrew Smith, FTC Bureau of Consumer Protection. Using Artificial Intelligence and Algorithms. April 8, 2020]

Explainable AI in Korea

- Explainable Artificial Intelligence(XAI) Center



Explainable AI Program in Korea

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**

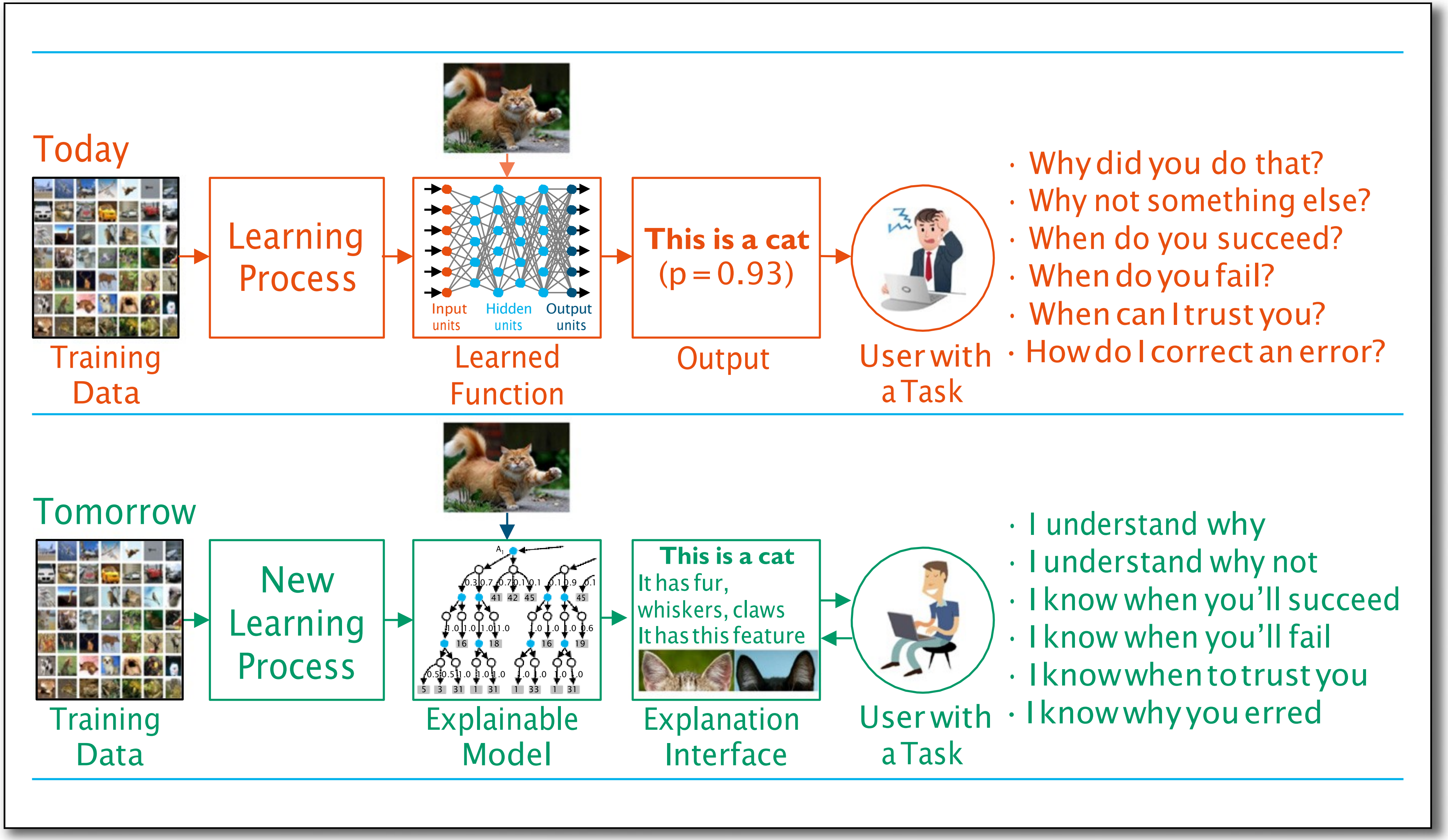
Fund › Institute of Information & Communication Technology Promotion (IITP) under Ministry of Science and ICT (MSICT) as part of Innovative Growth Engine Project

Period › July 2017 ~ December 2021 (54 months)

GISC2020
Global ICT Standards Conference

뉴 노멀 시대
선도를 위한
ICT 표준의
역할

DARPA XAI: Concept



Today

Training Data → Learning Process → Learned Function → This is a cat (p = 0.93) → Output → User with a Task

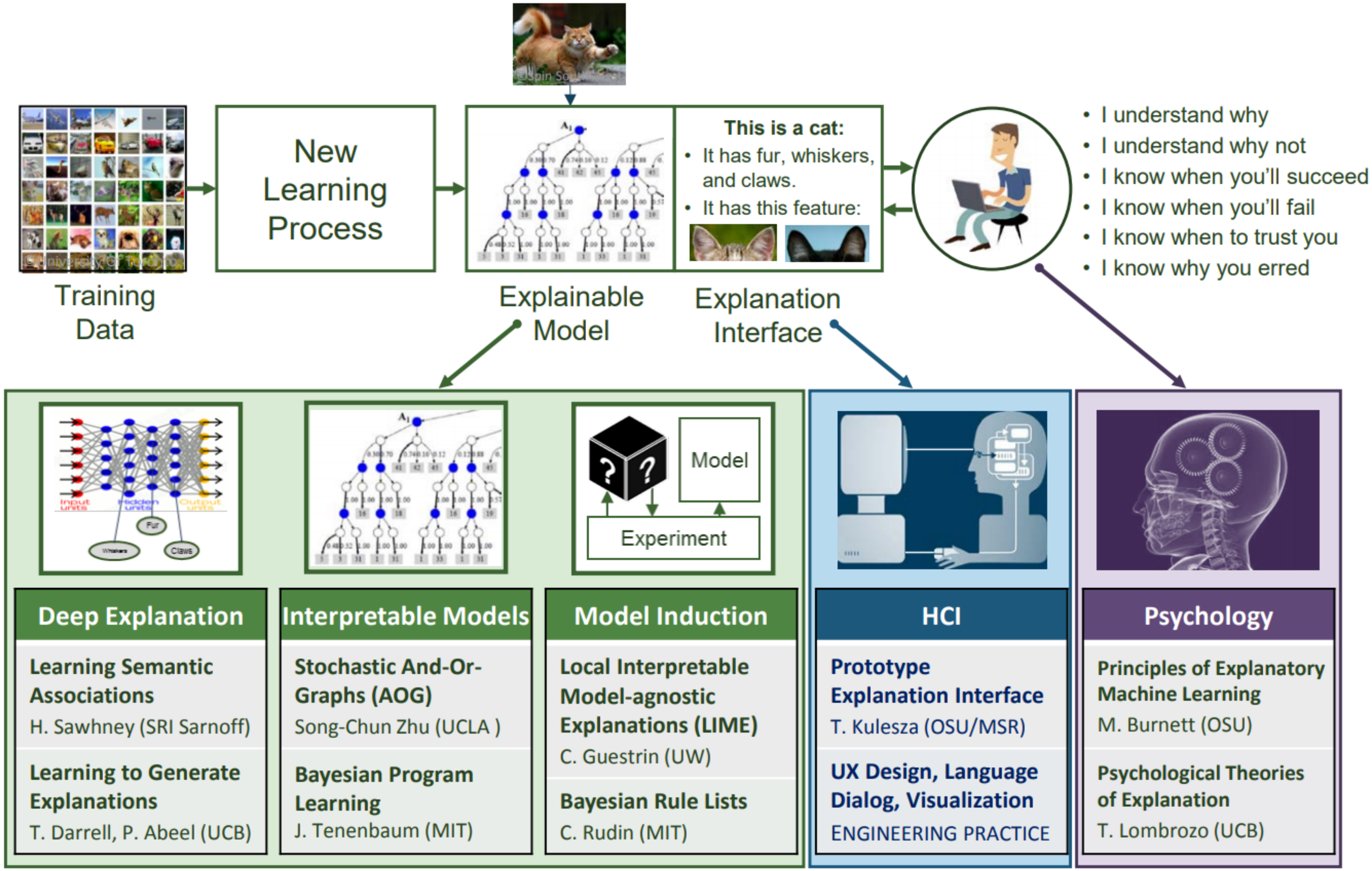
- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

Tomorrow

Training Data → New Learning Process → Explainable Model → This is a cat
It has fur, whiskers, claws
It has this feature
→ Explanation Interface ⇄ User with a Task

- I understand why
- I understand why not
- I know when you'll succeed
- I know when you'll fail
- I know when to trust you
- I know why you erred

Darpa XAI: Overview

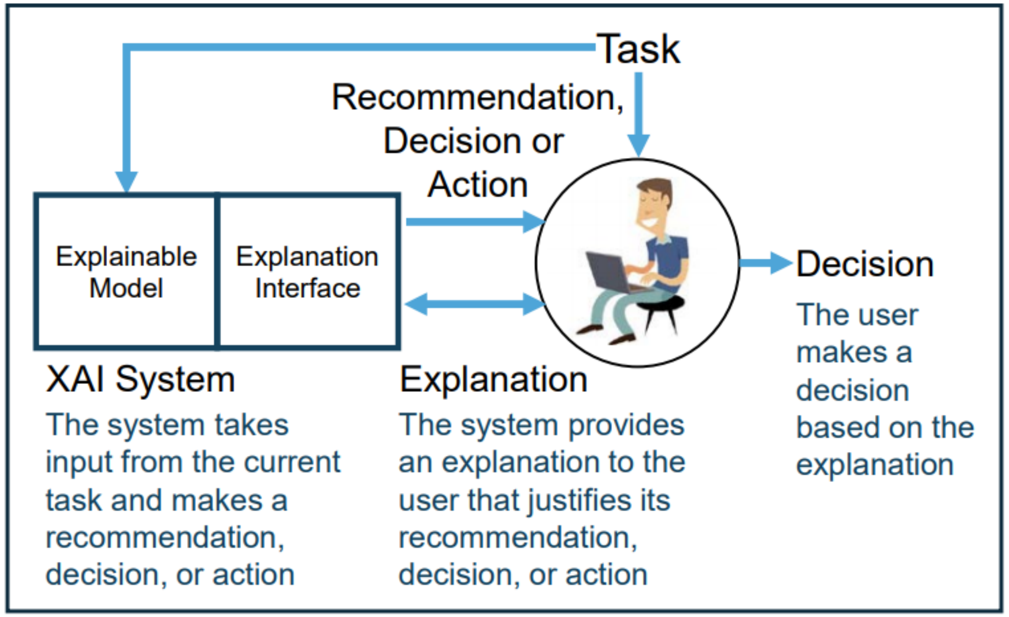


- I understand why
- I understand why not
- I know when you'll succeed
- I know when you'll fail
- I know when to trust you
- I know why you erred

Deep Explanation	Interpretable Models	Model Induction	HCI	Psychology
Learning Semantic Associations H. Sawhney (SRI Sarnoff)	**Stochastic And-Or-Graphs (AOG)** Song-Chun Zhu (UCLA)	**Local Interpretable Model-agnostic Explanations (LIME)** C. Guestrin (UW)	**Prototype Explanation Interface** T. Kulesza (OSU/MSR)	**Principles of Explanatory Machine Learning** M. Burnett (OSU)
Learning to Generate Explanations T. Darrell, P. Abeel (UCB)	**Bayesian Program Learning** J. Tenenbaum (MIT)	**Bayesian Rule Lists** C. Rudin (MIT)	**UX Design, Language Dialog, Visualization** ENGINEERING PRACTICE	**Psychological Theories of Explanation** T. Lombrozo (UCB)

DARPA XAI: Measuring Evaluation Effectiveness

Explanation Framework



Task

Recommendation, Decision or Action

Explainable Model

Explanation Interface

XAI System

The system takes input from the current task and makes a recommendation, decision, or action

Explanation

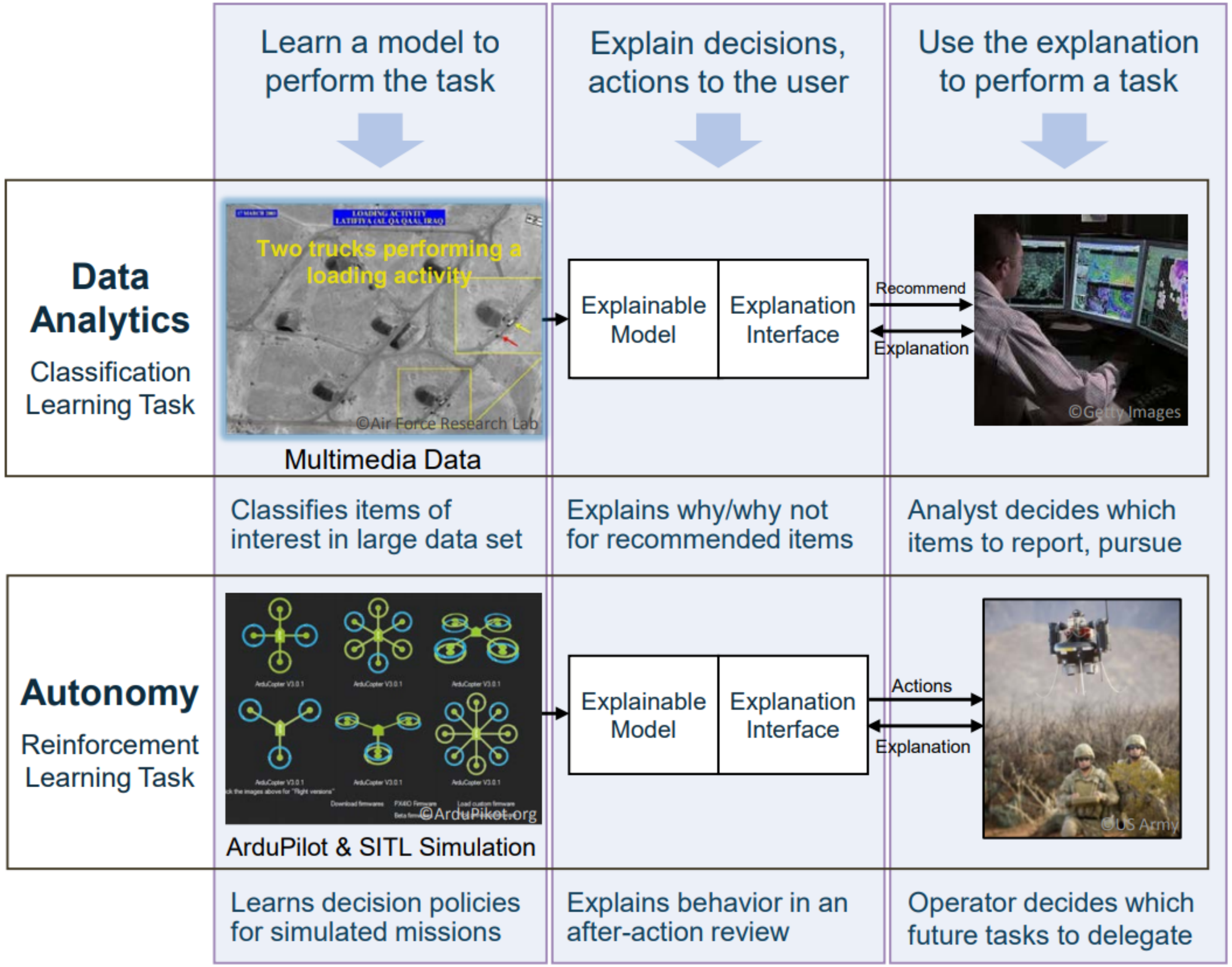
The system provides an explanation to the user that justifies its recommendation, decision, or action

Decision

The user makes a decision based on the explanation

Measure of Explanation Effectiveness
User Satisfaction
• Clarity of the explanation (user rating) • Utility of the explanation (user rating)
Mental Model
• Understanding individual decisions • Understanding the overall model • Strength/weakness assessment • 'What will it do' prediction • 'How do I intervene' prediction
Task Performance
• Does the explanation improve the user's decision, task performance? • Artificial decision tasks introduced to diagnose the user's understanding
Trust Assessment
• Appropriate future use and trust
Correctablity
• Identifying errors • Correcting errors • Continuous training

Darpa XAI: Challenge Problem Areas

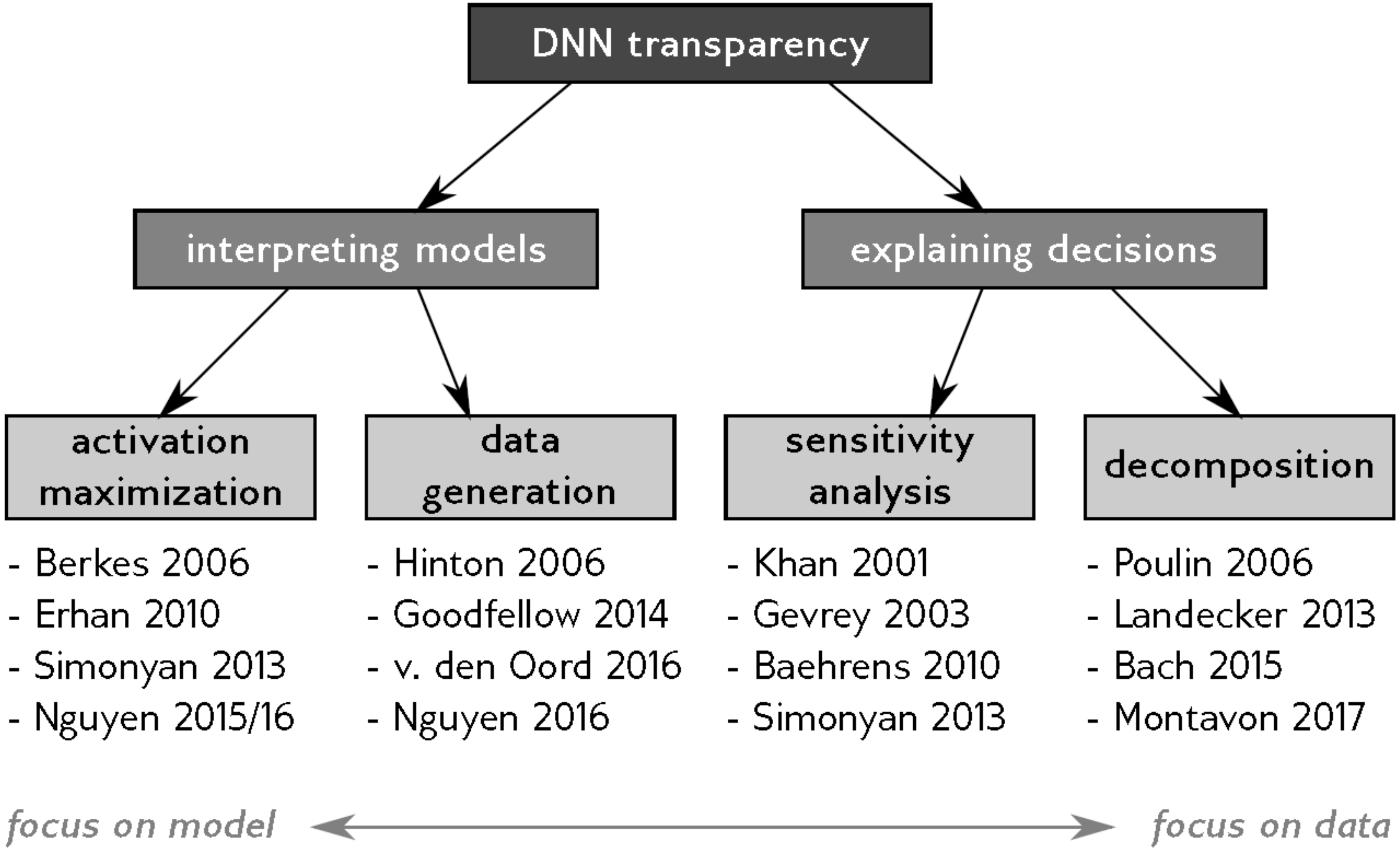


	Learn a model to perform the task	Explain decisions, actions to the user	Use the explanation to perform a task
Data Analytics Classification Learning Task	Multimedia Data	Explainable Model → Explanation Interface; Recommend / Explanation	
	Classifies items of interest in large data set	Explains why/why not for recommended items	Analyst decides which items to report, pursue
Autonomy Reinforcement Learning Task	ArduPilot & SITL Simulation	Explainable Model → Explanation Interface; Actions / Explanation	
	Learns decision policies for simulated missions	Explains behavior in an after-action review	Operator decides which future tasks to delegate

An analyst is looking for items of interest in massive multimedia data sets

An operator is directing autonomous systems to accomplish a series of missions

Deep Neural Network Transparency



DNN transparency

interpreting models

explaining decisions

activation maximization	data generation	sensitivity analysis	decomposition
- Berkes 2006	- Hinton 2006	- Khan 2001	- Poulin 2006
- Erhan 2010	- Goodfellow 2014	- Gevrey 2003	- Landecker 2013
- Simonyan 2013	- v. den Oord 2016	- Baehrens 2010	- Bach 2015
- Nguyen 2015/16	- Nguyen 2016	- Simonyan 2013	- Montavon 2017

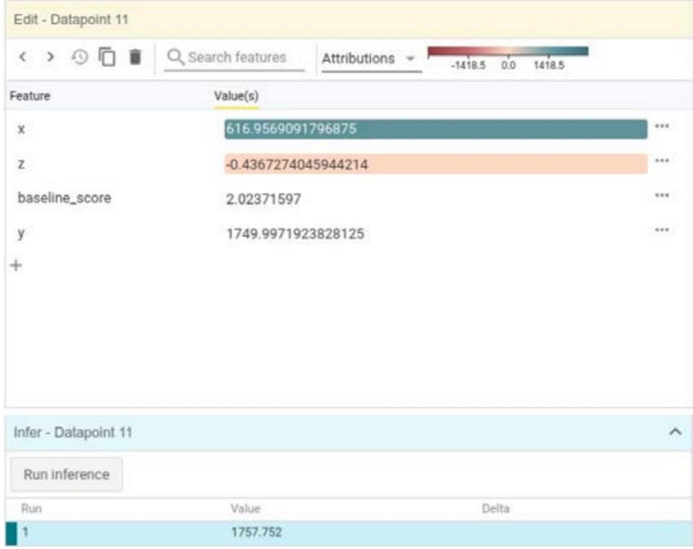
focus on model ⟷ *focus on data*

GISC2020
Global ICT Standards Conference

뉴 노멀 시대
선도를 위한
ICT 표준의
역할

Google Cloud Explanations

- AutoML Table(https://cloud.google.com/automl-tables/docs/features#ai-explanations/)
  - Feature Attributions
    - Shapley Values
    - Baselines and Counterfactuals
    - Attribution Methods
    - Aggregate Attributions
    - Attribution Limitations and Usage Considerations
  - Explanation Model Metadata
  - Visualizations with the What-If Tool



Edit - Datapoint 11

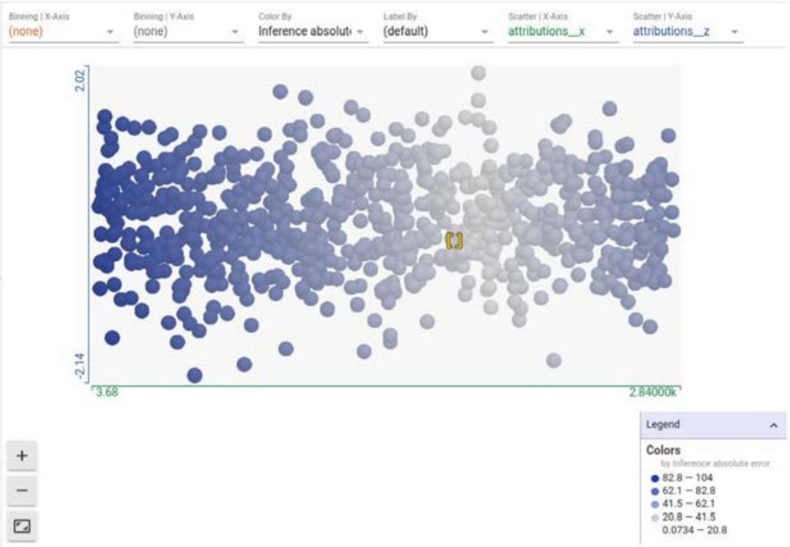
Search features | Attributions

Feature	Value(s)
x	616.9569091796875
z	-0.4367274045944214
baseline_score	2.02371597
y	1749.9971923828125

Infer - Datapoint 11

Run inference

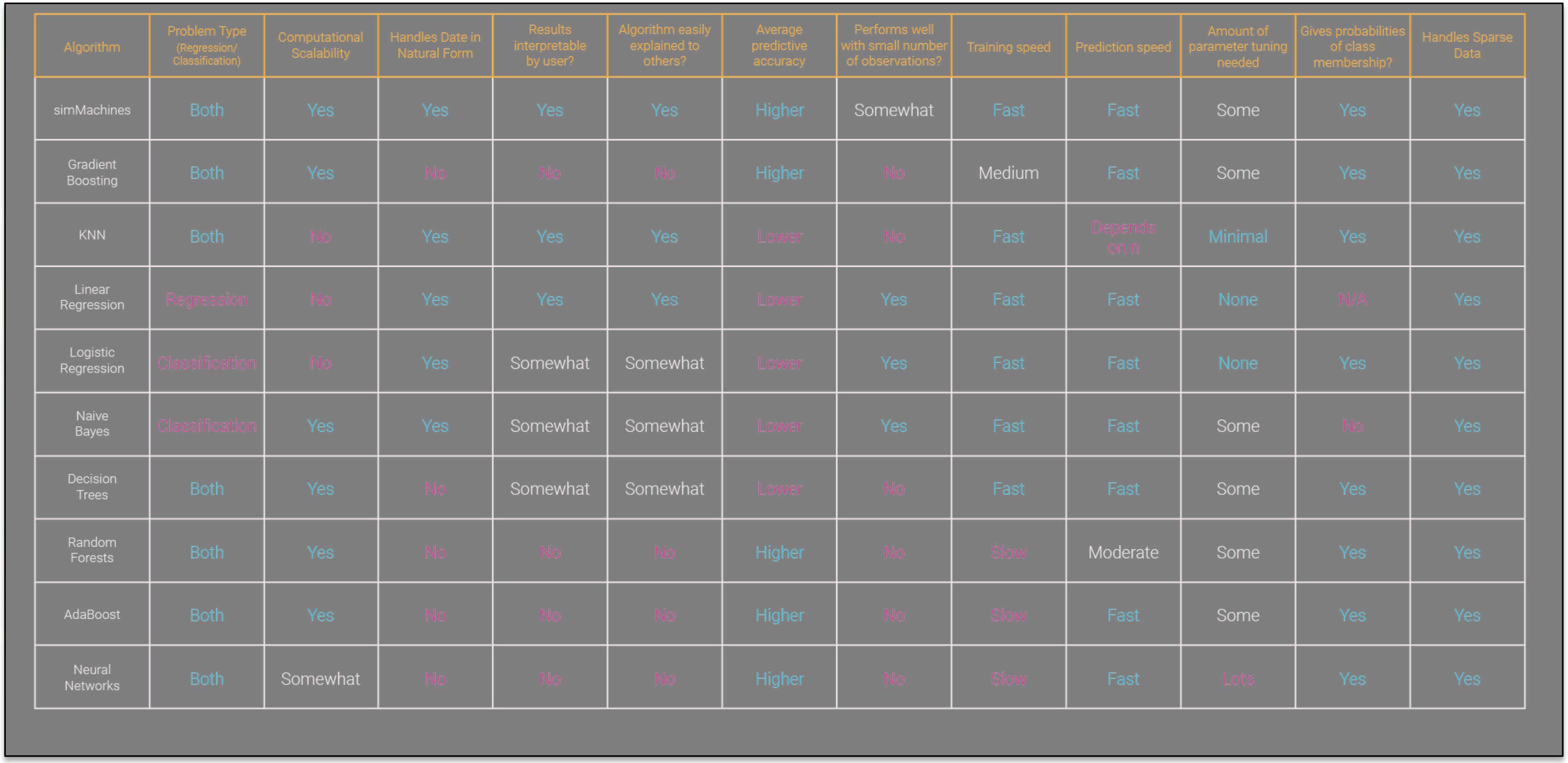
Run	Value	Delta
1	1757.752	



Binning | X-Axis
(none)
Binning | Y-Axis
(none)
Color By
Inference absolut
Label By
(default)
Scatter | X-Axis
attributions__x
Scatter | Y-Axis
attributions__z

simMachines (https://simmachines.com/)

- Transparent Machine Learning Predictions
  - similarity-based machine learning (nearest neighbor) method provides the Why behind every machine learning prediction.
  - Accurate, Justifiable, Actionable, Measurable prediction

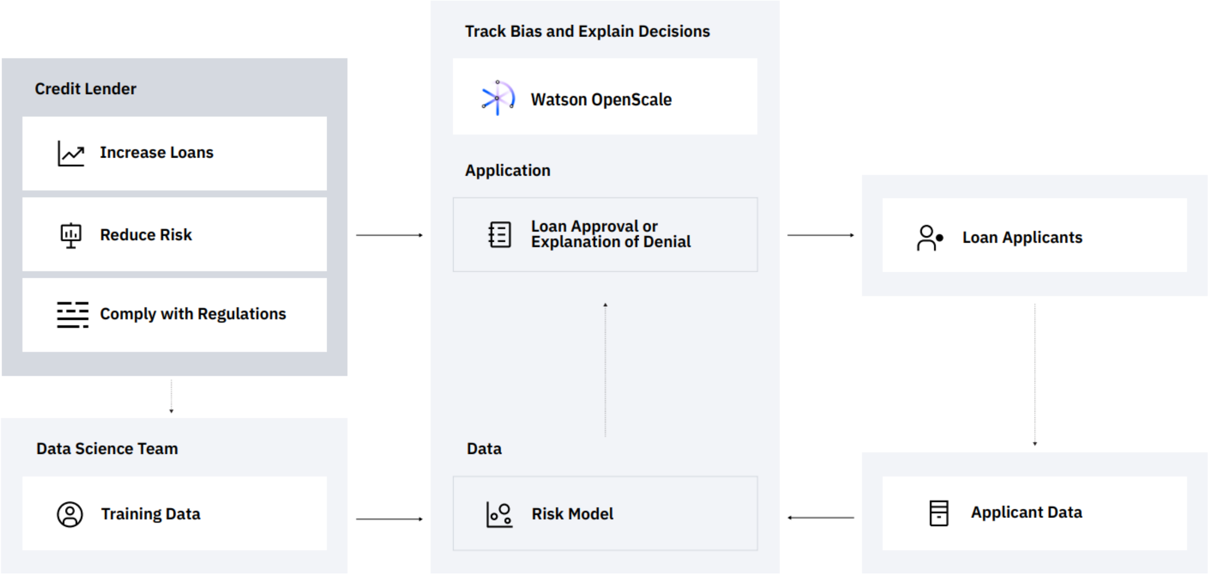


Algorithm	Problem Type (Regression/ Classification)	Computational Scalability	Handles Date in Natural Form	Results interpretable by user?	Algorithm easily explained to others?	Average predictive accuracy	Performs well with small number of observations?	Training speed	Prediction speed	Amount of parameter tuning needed	Gives probabilities of class membership?	Handles Sparse Data
simMachines	Both	Yes	Yes	Yes	Yes	Higher	Somewhat	Fast	Fast	Some	Yes	Yes
Gradient Boosting	Both	Yes	No	No	No	Higher	No	Medium	Fast	Some	Yes	Yes
KNN	Both	No	Yes	Yes	Yes	Lower	No	Fast	Depends on n	Minimal	Yes	Yes
Linear Regression	Regression	No	Yes	Yes	Yes	Lower	Yes	Fast	Fast	None	N/A	Yes
Logistic Regression	Classification	No	Yes	Somewhat	Somewhat	Lower	Yes	Fast	Fast	None	Yes	Yes
Naive Bayes	Classification	Yes	Yes	Somewhat	Somewhat	Lower	Yes	Fast	Fast	Some	No	Yes
Decision Trees	Both	Yes	No	Somewhat	Somewhat	Lower	No	Fast	Fast	Some	Yes	Yes
Random Forests	Both	Yes	No	No	No	Higher	No	Slow	Moderate	Some	Yes	Yes
AdaBoost	Both	Yes	No	No	No	Higher	No	Slow	Fast	Some	Yes	Yes
Neural Networks	Both	Somewhat	No	No	No	Higher	No	Slow	Fast	Lots	Yes	Yes

뉴 노멀 시대 선도를 위한 ICT 표준의 역할

IBM Watson OpenScale (https://www.ibm.com/kr-ko/cloud/watson-openscale)

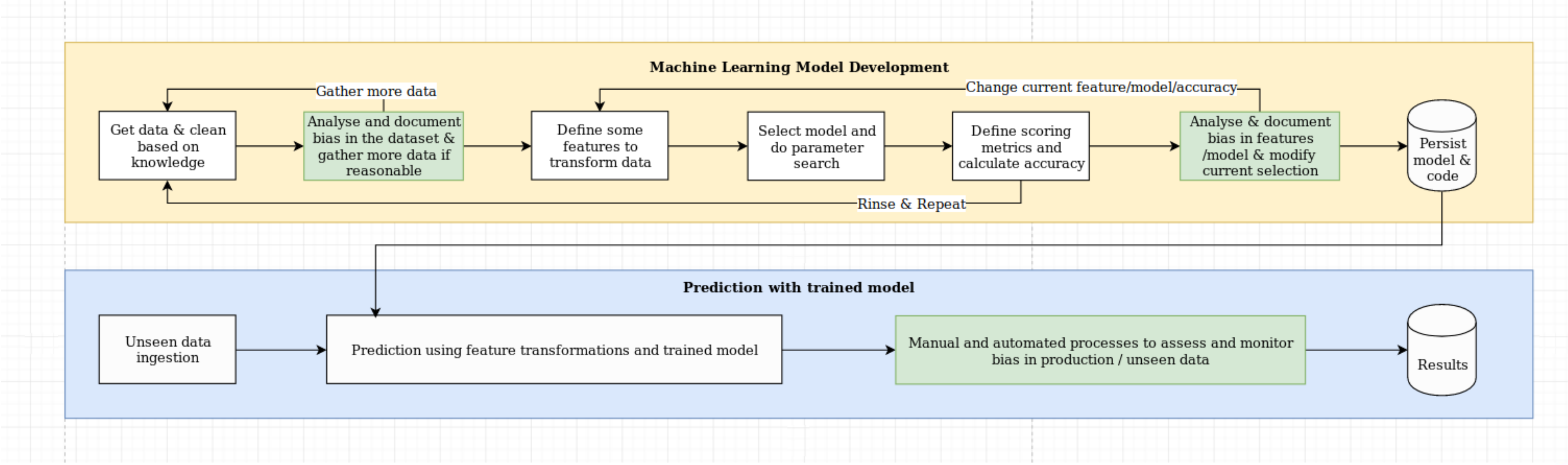
- the open platform for businesses to operationalize trusted AI and extend their deployments enterprise-wide.
  - Measure performance of production AI and its impact on business goals
  - Track actionable metrics and alerts in a single console
  - Enable the business user or project manager to understand AI outcomes
  - Apply business results to create a feedback loop that sustains AI outcomes
  - Govern and explain AI to maintain regulatory compliance
  - Automatically detect and mitigate harmful bias to improve outcomes
  - Accelerate the integration of AI into existing business applications



뉴 노멀 시대
선도를 위한
ICT 표준의
역할

XAI - An eXplainability toolbox for machine learning https://github.com/EthicalML/xai

- XAI is a Machine Learning library that is designed with AI explainability in its core.
- XAI contains various tools that enable for analysis and evaluation of data and models.
- The XAI library is maintained by The Institute for Ethical AI & ML, and it was developed based on the 8 principles for Responsible Machine Learning.
- XAI library is designed using the 3-steps of explainable machine learning, which involve 1) data analysis, 2) model evaluation, and 3) production monitoring.



Machine Learning Model Development

Gather more data

Change current feature/model/accuracy

Get data & clean based on knowledge

Analyse and document bias in the dataset & gather more data if reasonable

Define some features to transform data

Select model and do parameter search

Define scoring metrics and calculate accuracy

Analyse & document bias in features /model & modify current selection

Persist model & code

Rinse & Repeat

Prediction with trained model

Unseen data ingestion

Prediction using feature transformations and trained model

Manual and automated processes to assess and monitor bias in production / unseen data

Results

XAI - New Work Item Proposal (subject to change)

- Title:
  - Information technology – Artificial Intelligent – Requirements and guidelines for explainable AI systems

- Scope
  - This document will provide requirements for explainability for AI systems and guidelines to apply the techniques and model of Explainable Artificial Intelligence (XAI) to achieve the requirements. This document firstly provides
    - the desiderata for explainability of Artificial Intelligent systems from first principles and by analysing existing use cases, and
    - features and characteristics of techniques and models to realize explainability for AI systems.
  - The identified desiderata and features are then used to identify
    - requirements and guidelines for the explainability to achieve trustworthiness of AI systems.