

可解释人工智能

Explainable/Interpretable Artificial Intelligence
(XAI)

First

Interpretable and Robust AI in EEG Systems: *A Survey*

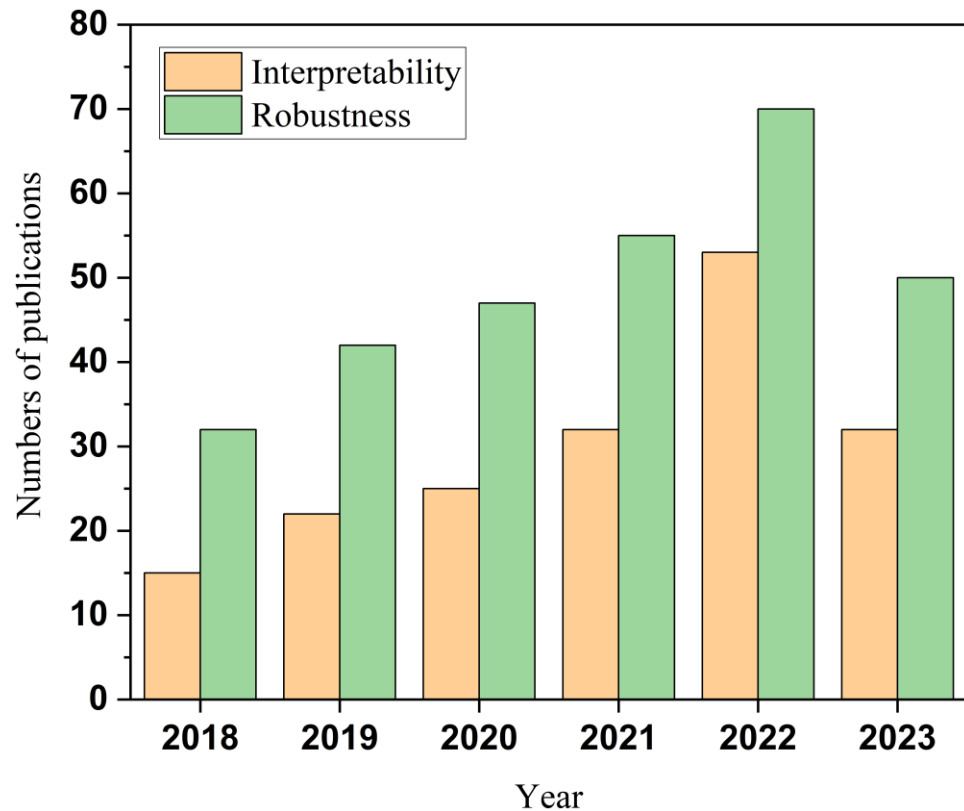
Interpretable

- Interpretability refers to understanding why and how the AI models make decisions and predictions.
- the interpretability allow researchers to gain insights into EEG dynamics and the link between brain states and cognitive functions, and also make it easier to identify potential biases and failure modes of EEG systems.
- From another point of view, the interpretability can foster user trust and acceptance of EEG systems, enabling users to build confidence in the validity and value of EEG systems.
- 理解模型的决策过程，指导解释原理性知识，提高对模型的置信。

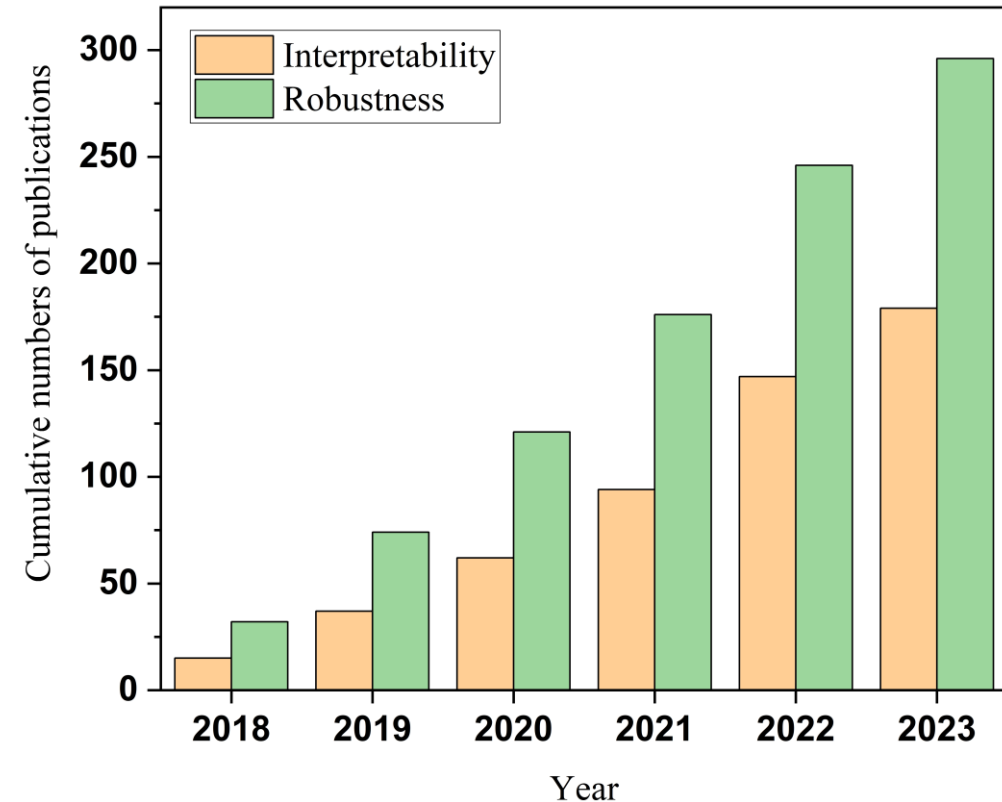
Robustness

- Robustness refers to the degree to which the decisions and predictions of AI models are free from attacks and perturbations.
- EEG data derived from brain tends to be noisy and variable across individuals, resulting in a lower signal-noise ratio (SNR).
- EEG signals are easily interfered by biological and environmental artifacts(muscle movements, eye blinks, heartbeat, electrical devices), and the same stimuli also evoke different EEG responses in different people which has unique neural rhythms.
- 解决干扰性问题

A surge of research interest (keywords)



AI with EEG



AI

Method Summary

Case 1: **MI Task**, model decision may pay more attention to **muscle movements** (noises) **rather than** EEG.

Sec. 3: **Interpretable** AI in EEG Systems

Sec. 3.1: Backpropagation-based Methods

反向传播

Sec. 3.2: Perturbation-based Methods

扰动

Sec. 3.3: Rule-based Methods

规则

Sec. 3.4: Discussion on Interpretable AI

Case 2: **Sleep State**, models identified that the signals of peripheral (外围) EEG channels generated by **regular eye movements during deep sleep** are highly correlated with sleep status even though these EEG signals **had long been overlooked**.

Local interpretability aims to explain **individual predictions** by illuminating **why a model correlates a specific EEG pattern with a particular condition**. **局部可解释性**旨在通过阐明为什么模型将特定EEG模式与特定条件相关联来解释个体预测

Global interpretability illuminates the **overall behavior of a model**, revealing **how it operates across multiple instances**. **全局可解释性**阐明了模型的整体行为，揭示了它如何在多个实例中运行

Techniques

局部可解释性指的是对某一个输入及其输出的理解,
全局可解释性指的是对整个模型整体的理解。

- Layer-wise Relevance Propagation (LRP)
- Deep Learning Important Features (Deep LIFT)
- Class Activation Mapping (CAM)
- Gradient-weighted Class Activation Mapping (Grad-CAM)
- random forest (RF)
- Fuzzy inference system (FIS)
- Bayesian system (BS)

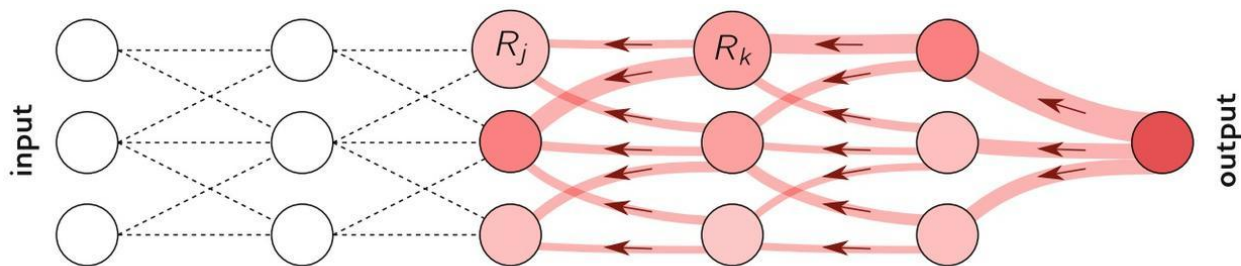
-
- Local Interpretable Model-Agnostic Explanations (LIME)
 - Shapley Additive Explanations (SHAP)

Interpetability Categories	Methods	Coverage	Explanation Type
Backpropagation-based Methods	LRP	Local/Global	Attribution
	DeepLIFT	Local/Global	Attribution
	CAM	Local	Attribution
	Grad-CAM	Local	Attribution
Perturbation-based Methods	LIME	Local	Attribution
	SHAP	Local	Attribution
Rule-based Methods	RF	Global	Decision Rules
	FIS	Global	Fuzzy Rules
	BS	Global	Bayesian Rules

Backpropagation-based Methods

事后方法

- Backpropagation-based methods **decompose** the model predictions by first **backpropagating the gradients from the predictions into input feature space** and then **visualizing the weights of these features** in raw EEG signals that contribute to predictions.



LRP归因：层间的反向传播归因法

LRP 的核心是利用反向传播将高层的相关性分值递归地传播到低层直至传播到输入

DeepLIFT 基于参考激活，能够将特征的重要性与预定义的参考点进行比较。其核心原理是计算每个输入特征的贡献分数。

(验证模型的预测逻辑是否符合生理学原理)

DeepLIFT 可以从模型预测中发现某些特征模式，以指导大脑研究。

Backpropagation-based Methods

- Class Activation Mapping (CAM)

通过可视化每个输入特征在最终分类决策中的重要性来生成热图

CAM是一种将CNN所看到或关注的内容可视化并为我们生成类输出的方法。

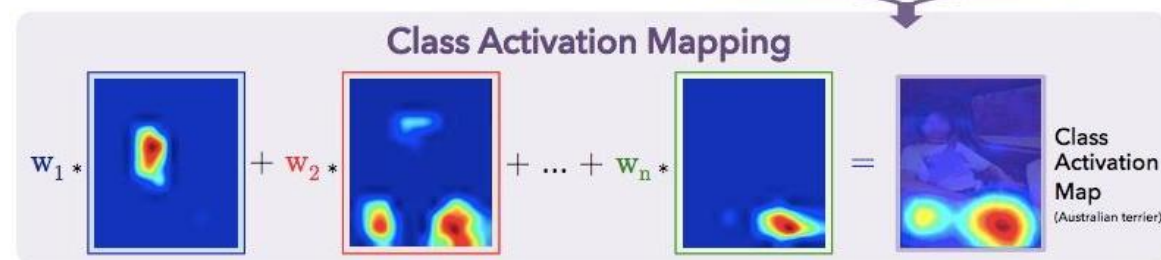
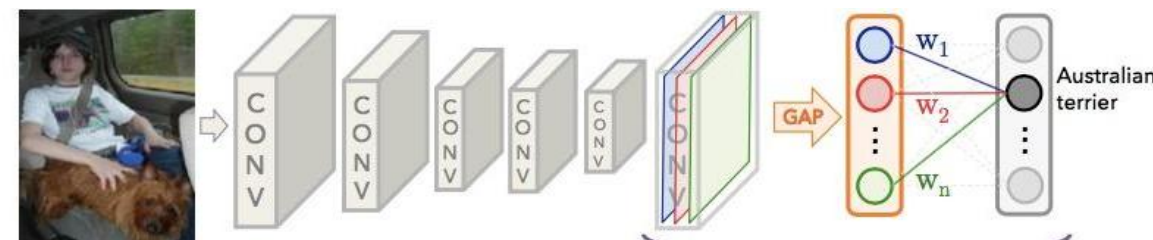
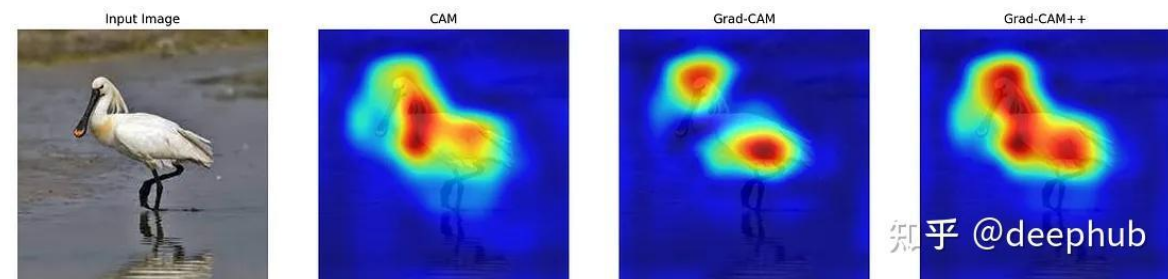
通过将图像传递给CNN，获得了相同图像的低分辨率特征图。

CAM的思想是，删除那些完全连接的神经网络，并用全局平均池化层代替它们，特征图中所有像素的平均值就是它的全局平均值。通过将GAP应用于所有特征映射将获得它们的标量值。

- Gradient-weighted Class Activation Mapping (GradCAM)

CAM的扩展

Grad-CAM背后的思想是，依赖于最后一个卷积层的特征映射中使用的梯度，而不是使用网络权重。这些梯度是通过反向传播得到的。



Perturbation-based Methods

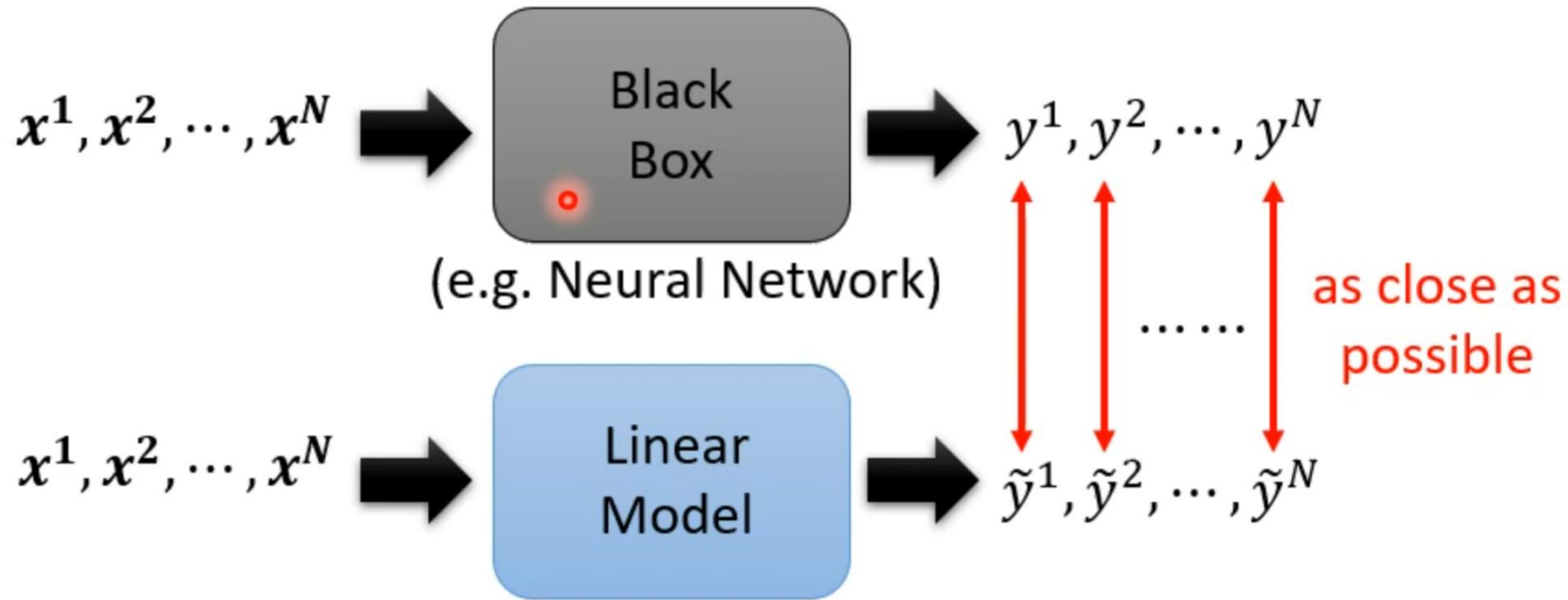
- perturb individual EEG samples and **observe the impact** on subsequent network neurons and predictions, trying to **reveal correlations** between samples and model outputs. (**post-hoc methods**)
- building local models to approximate the predictions of the original models based on perturbed inputs (**model-agnostic**) the local models establish the connection between biological features and original model predictions.

- **Interpretable Model-agnostic Explanations (LIME)**

LIME explains target model predictions by approximating them locally with interpretable models

- **Shapley Additive Explanation Values (SHAP)**

SHAP **quantifies the contribution** of each input features to prediction based on **the Shapley values** from game theory(博弈论)

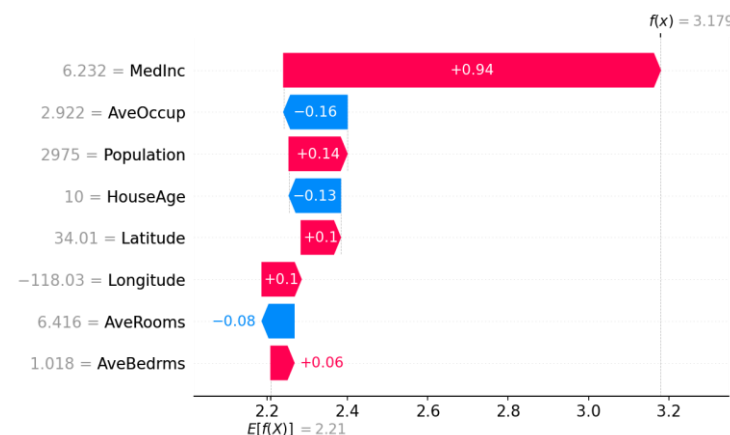
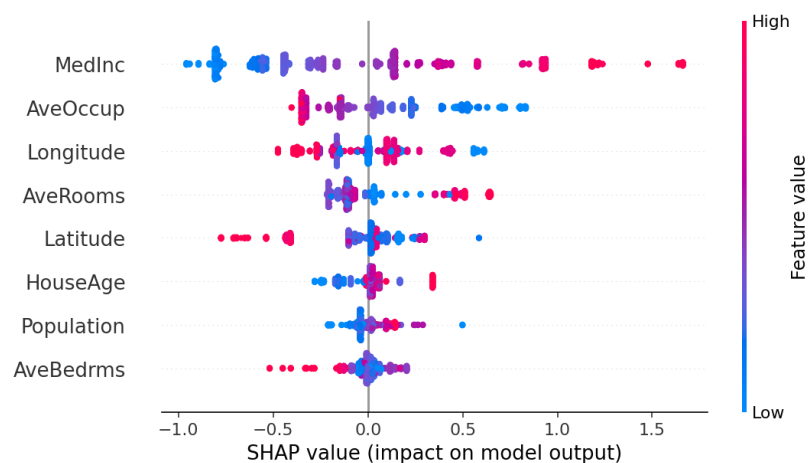


通过创建一个简化版的模型去解释整个复杂的模型

Local Interpretable **Model-Agnostic** Explanations (LIME)

SHAP

- **Local accuracy** ensures that the sum of **SHAP values for each input feature** and the expected model output equals the model prediction for a specific instance
- **Missingness** indicates that **if a feature is missing or has no impact on the model prediction, its SHAP value will be zero.**
- **Consistency** guarantees that if a **feature contributes more in a new model** compared to an old one, the SHAP value of that feature **should not decrease**
- SHAP 的核心是构建一个加性解释模型，将所有特征视为“贡献者”。每个特征的 SHAP 值表示其对模型输出的影响大小和方向。正值表示正向影响，负值表示负向影响。SHAP 的理论基础源于 Shapley 值，它起源于合作博弈论，用于**公平分配**多个参与者的贡献。



Rule-based Methods

- **Random Forest (RF)**: feature importance and decision paths
- **Bayesian System (BS)**: The BS uses Bayesian theorem to model the relationship between EEG features and prediction.
- **Fuzzy Inference System (FIS)**:

(控制论)

rule1: IF 购买意愿很强 THEN Action=买

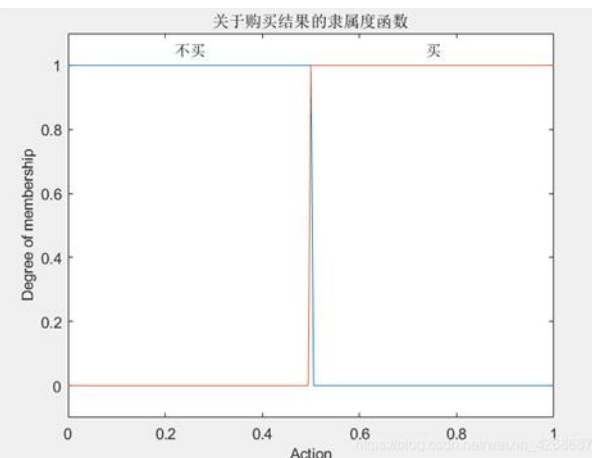
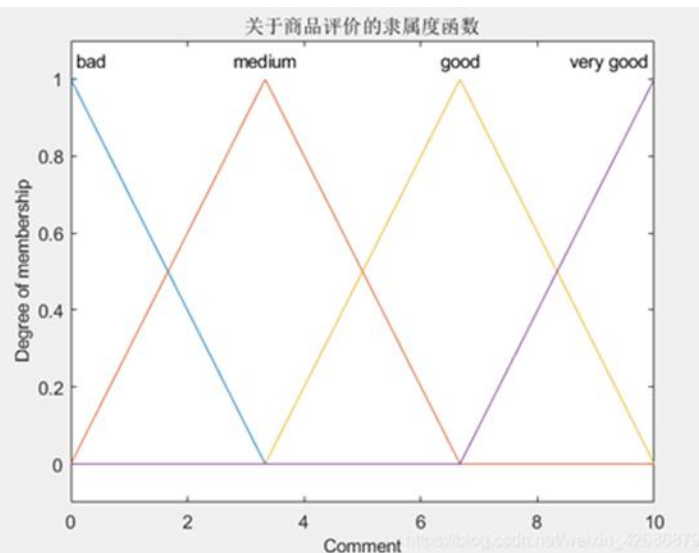
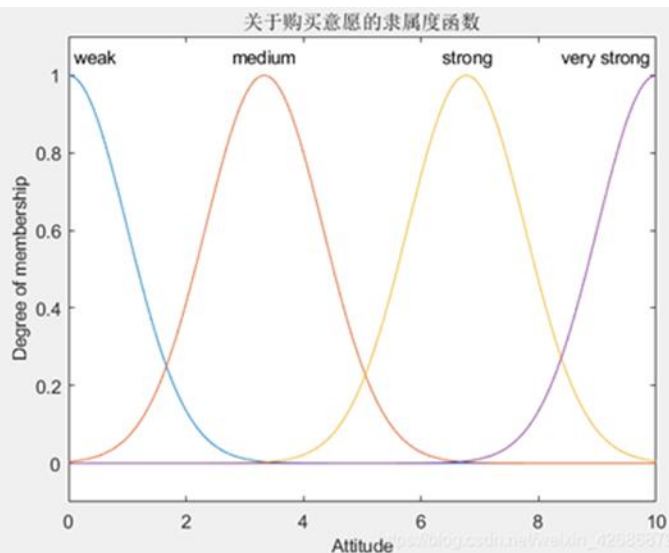
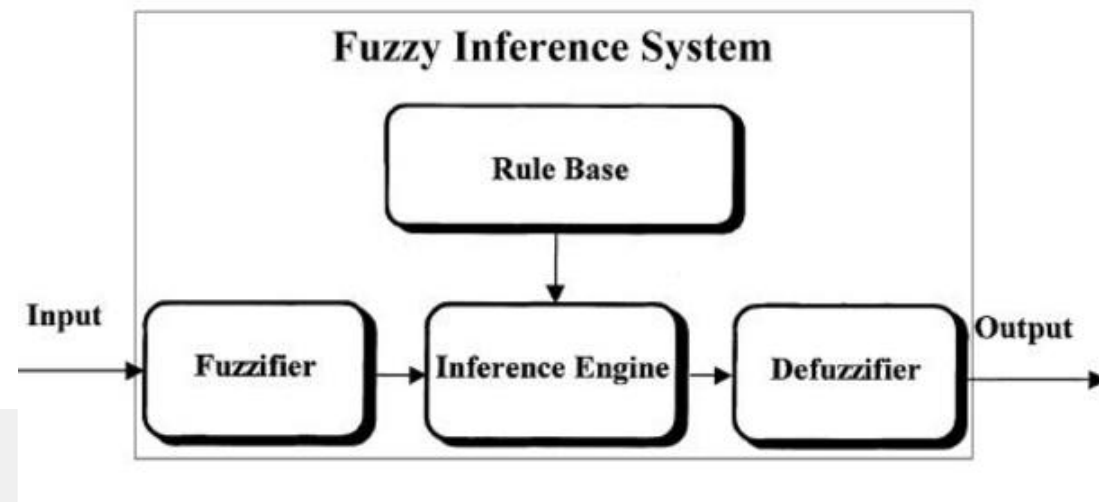
rule2: IF 商品评价极好且购买意愿为不为弱 THEN Action=买

rule3: IF 商品评价为差 THEN Action=不买

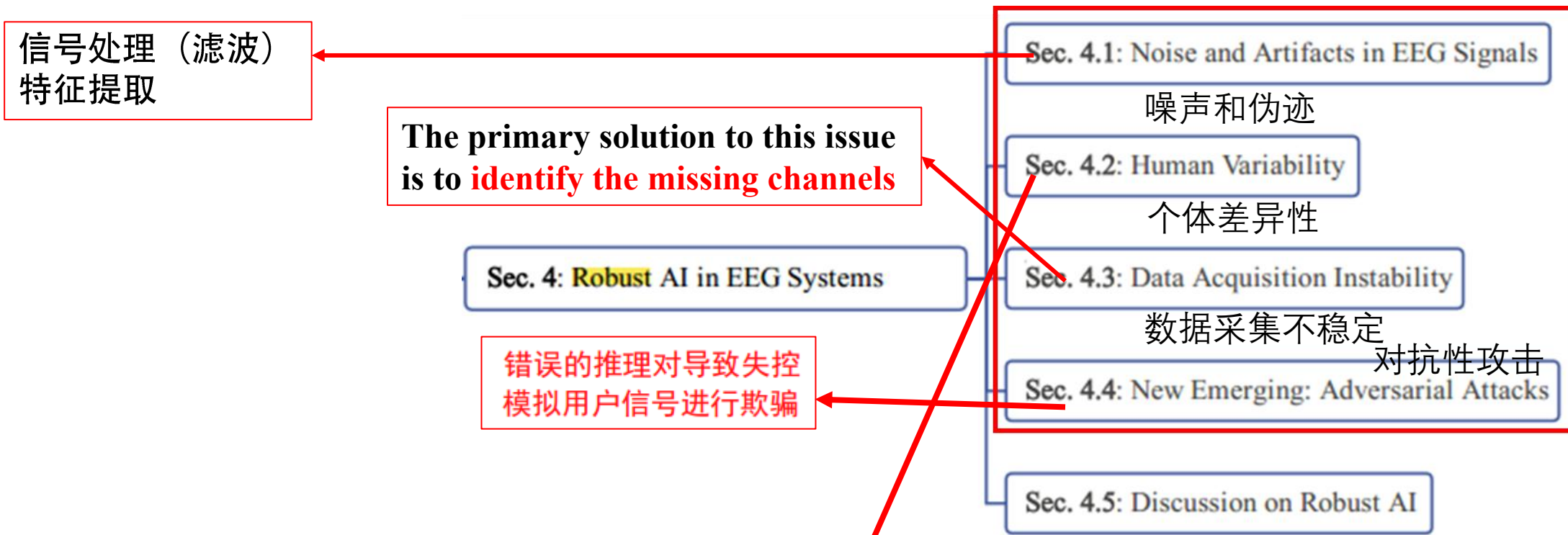
rule4: IF 如果商品评价不为差且购买意愿强 THEN Action=买

Fuzzy rules 是模糊逻辑中用于描述输入和输出之间关系的规则

通过对输入变量的隶属度进行模糊化, 将它们与规则进行匹配, 得到输出变量的隶属度, 最终通过解模糊化得到实际输出。



Method Summary



Undesirable Factors	Subcategory	Methods and Representative Works
Noise and Artifacts	External Noise Internal Artifacts	Traditional Signal Processing [130], [131] Models' Self-Robustness [132], [84]
Human Variability	Cross-subject Issues Cross-session Issues	Transfer Learning [133], [134], Dynamic Domain Adaptation [135] Transfer Learning [136], [137], Robust Feature Extraction [138], [135]
Data Acquisition Instability	Resistance Change Channel Missing & Broken	Attention Mechanism [139], [140] Missing Data Reconstruction [141], [142], [143], [144], [131]
Adversarial Attacks	Evasion & Manipulation	Adversarial Training [145], [146], [147], [148]

Future Direction

Interpretable

Prior Human Knowledge

通过已建立的**生理学原理** 指导（约束）模型关注相关特征

High-dimensional Feature Interpretation

缺乏对特征**为什么被分配特定贡献值**的洞察
提供**动态的特征描述**，而不仅仅是特征和预测之间的线性关系。

（通过模型的隐藏语义，如果高维特征可以被解释为噪声和本质特征之间的相似性，我们就可以知道模型的注意力是如何被吸引到噪声上的。）

Robust

Artificial Synthetic Data for Large Models

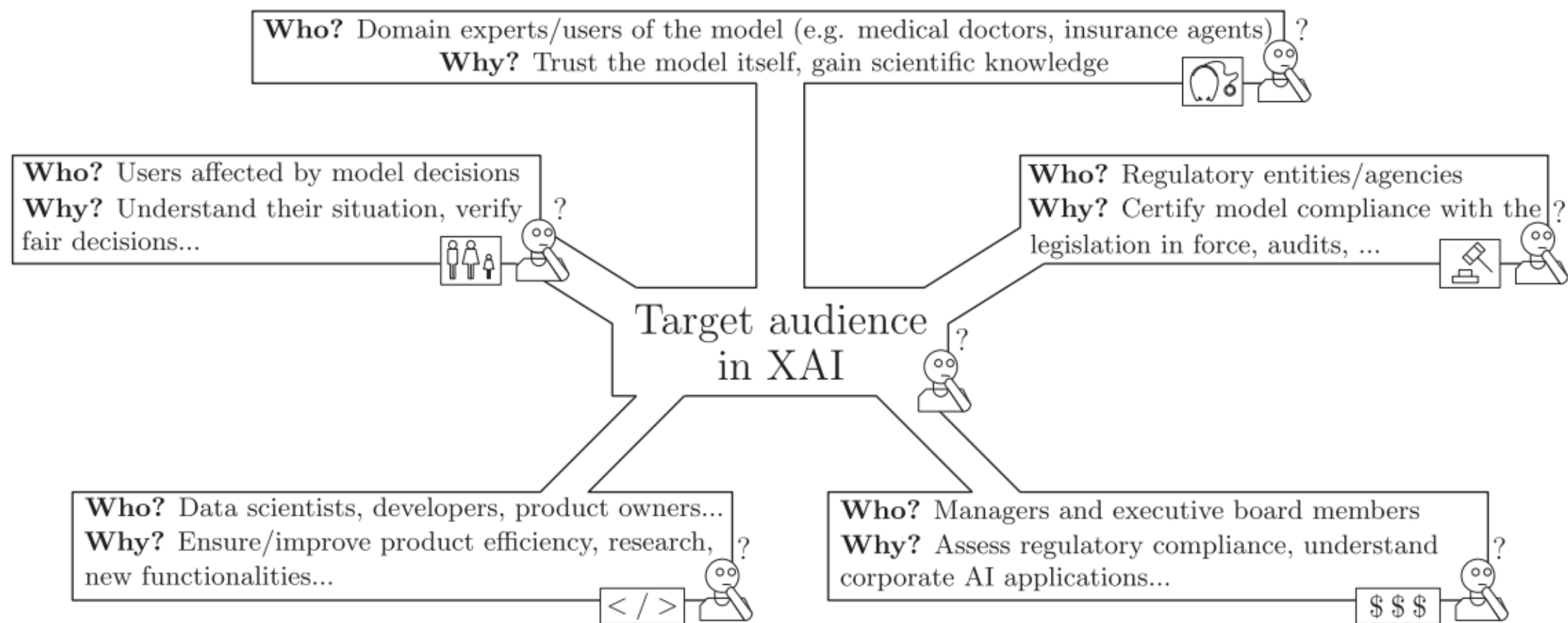
通过发展**数据增强**技术（足够的数据量）建立脑电大模型

Decoupling of EEG Signals for Robust Feature

解耦**主体身份信息**和**任务相关信息**

Second

Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI



What ?

“XAI will create a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners”.

用于解释的细节或理由完全取决于它们所呈现的观众。
解释是否使概念清晰或易于理解也完全取决于听众

Given a certain audience, explainability refers to the details and reasons a model gives to make its functioning clear or easy to understand.

模型是可以解释的，但模型的可解释性来自于模型本身的设计。

*Given an audience, an **explainable** Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand.*

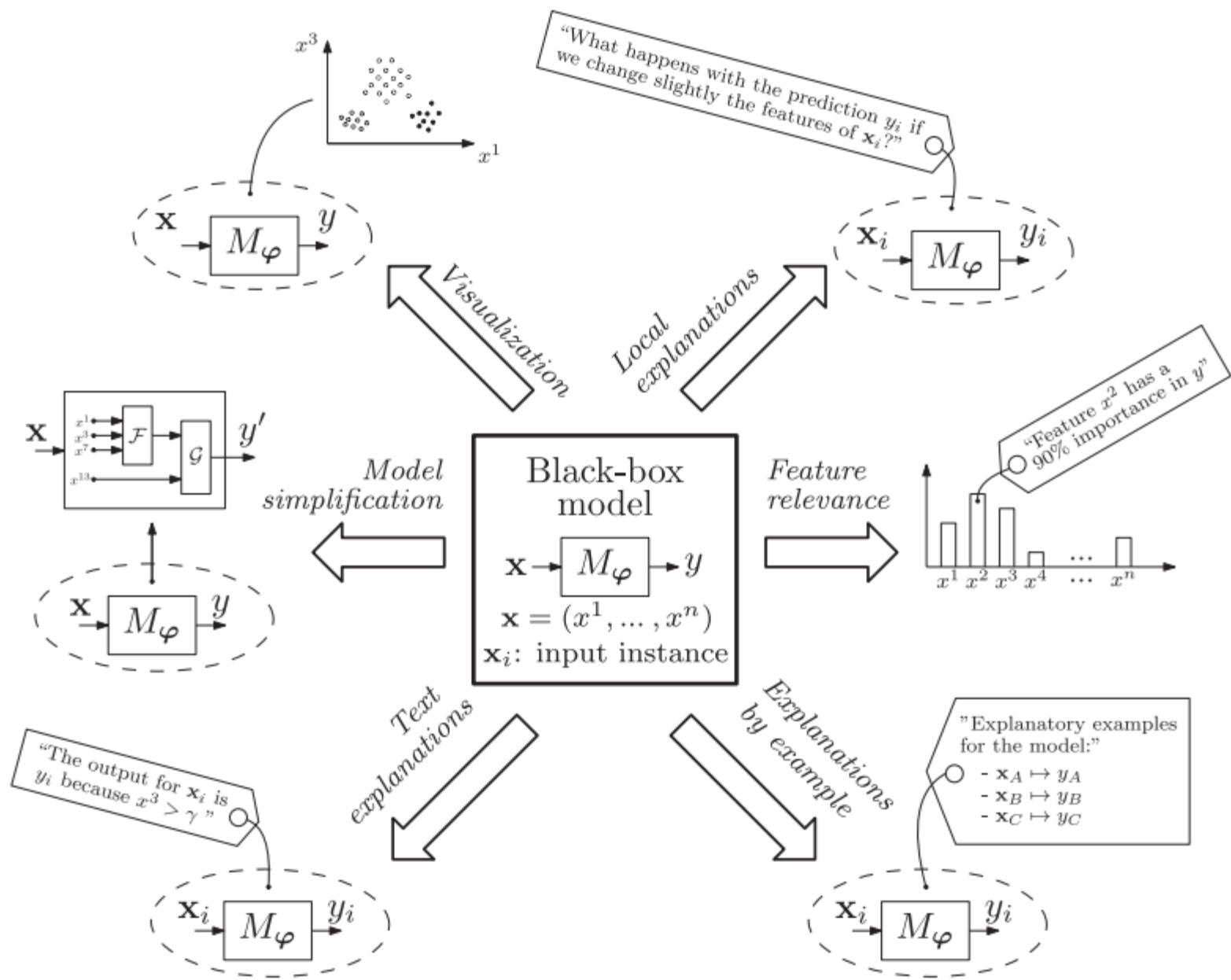
Why ?
What for (aim) ?

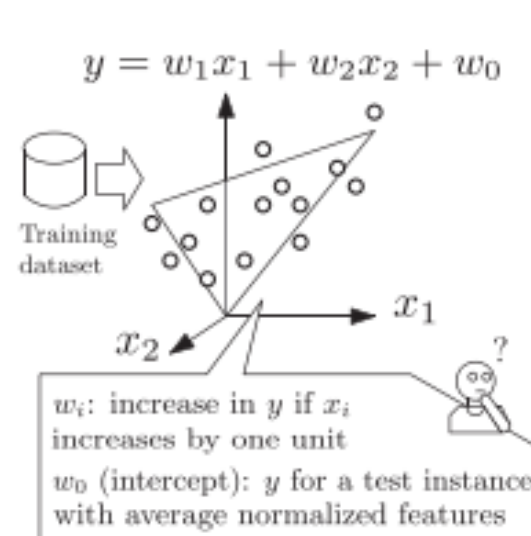
XAI Goal	Main target audience (Fig. 2)
Trustworthiness	Domain experts, users of the model affected by decisions
Causality	Domain experts, managers and executive board members, regulatory entities/agencies
Transferability	Domain experts, data scientists
Informativeness	All
Confidence	Domain experts, developers, managers, regulatory entities/agencies
Fairness	Users affected by model decisions, regulatory entities/agencies
Accessibility	Product owners, managers, users affected by model decisions
Interactivity	Domain experts, users affected by model decisions
Privacy awareness	Users affected by model decisions, regulatory entities/agencies

How ?

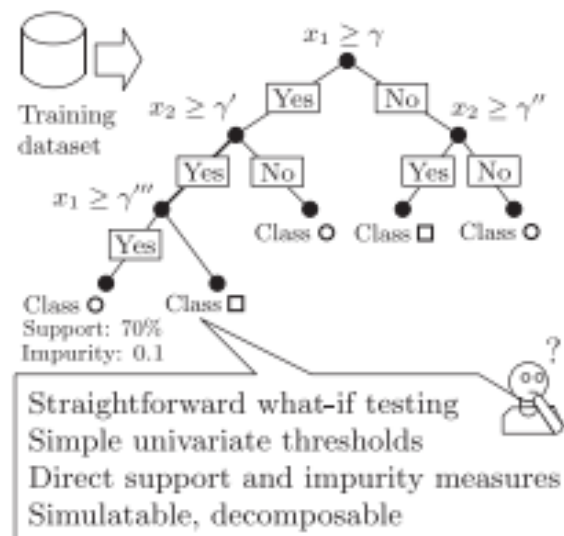
interpretable models (transparency)	
	Algorithmic transparency
	Decomposability
	Simulatability

model interpretability (post-hoc technique)	
	Text explanation
	Visualizations
	Local explanations
	Explanations by example
	Explanations by simplification
	Feature relevance

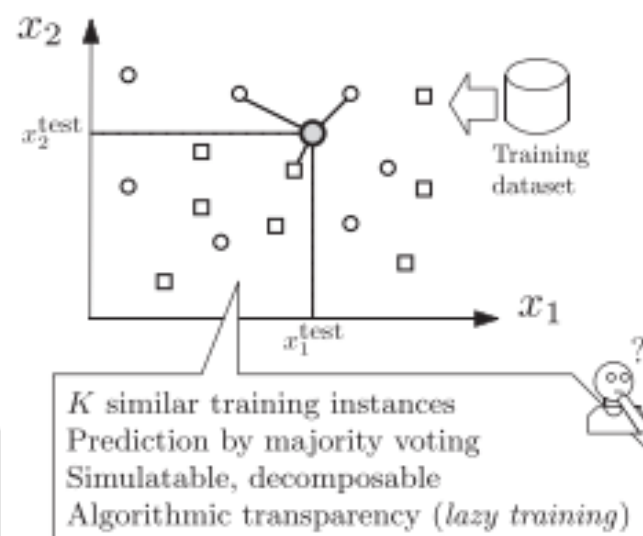




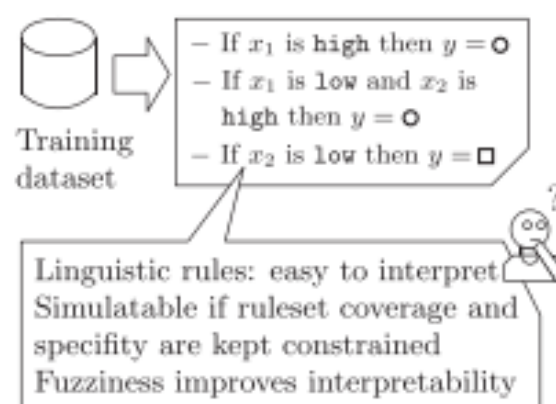
(a)



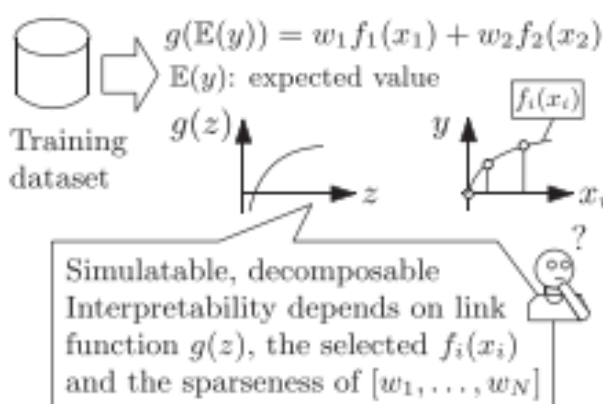
(b)



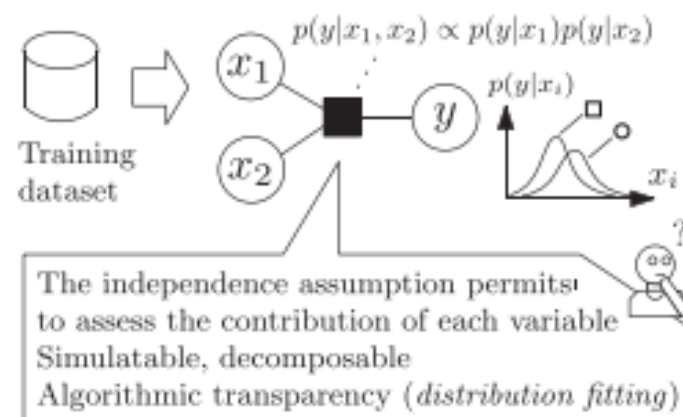
(c)



(d)

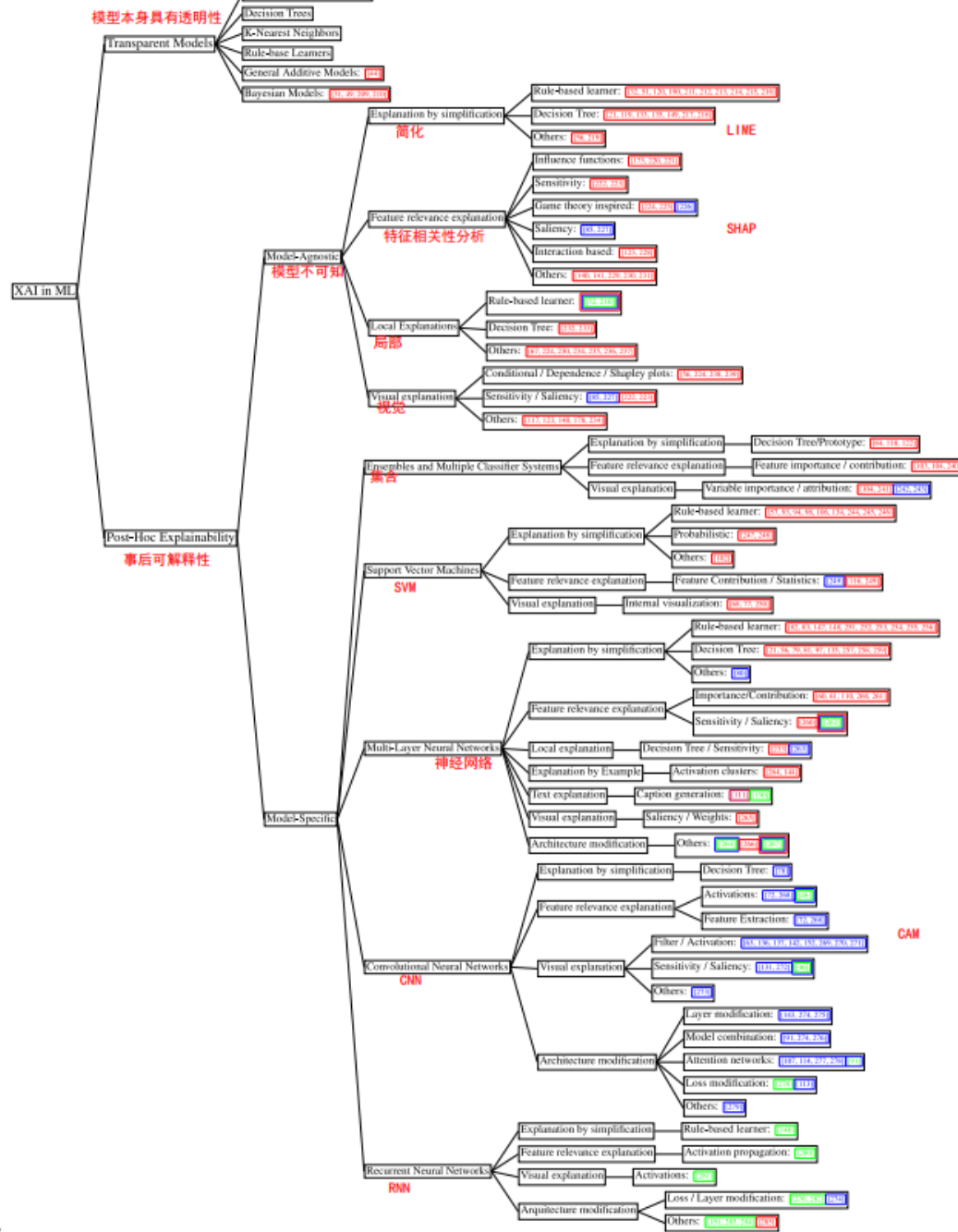


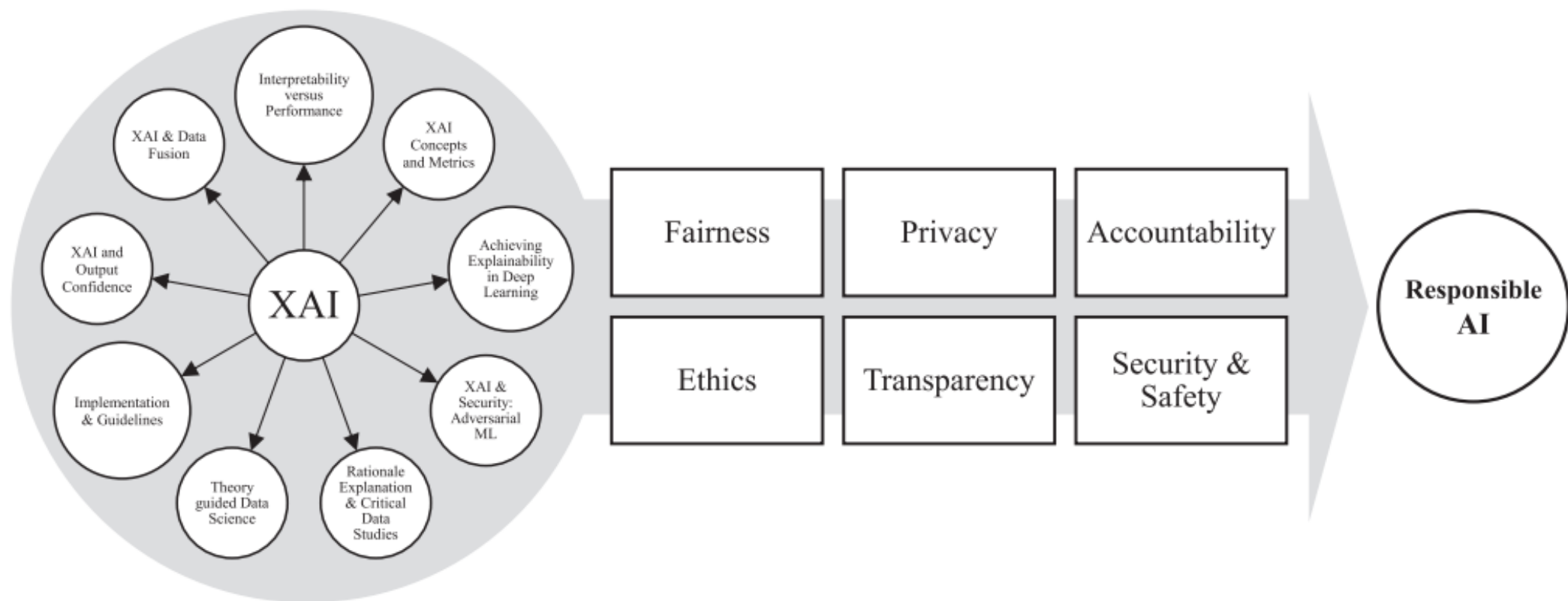
(e)



(f)

Model	Transparent ML Models			Post-hoc analysis
	Simulatability	Decomposability	Algorithmic Transparency	
Linear/Logistic Regression	Predictors are human readable and interactions among them are kept to a minimum	Variables are still readable, but the number of interactions and predictors involved in them have grown to force decomposition	Variables and interactions are too complex to be analyzed without mathematical tools	Not needed
Decision Trees	A human can simulate and obtain the prediction of a decision tree on his/her own, without requiring any mathematical background	The model comprises rules that do not alter data whatsoever, and preserves their readability	Human-readable rules that explain the knowledge learned from data and allows for a direct understanding of the prediction process	Not needed
K-Nearest Neighbors	The complexity of the model (number of variables, their understandability and the similarity measure under use) matches human naive capabilities for simulation	The amount of variables is too high and/or the similarity measure is too complex to be able to simulate the model completely, but the similarity measure and the set of variables can be decomposed and analyzed separately	The similarity measure cannot be decomposed and/or the number of variables is so high that the user has to rely on mathematical and statistical tools to analyze the model	Not needed
Rule Based Learners	Variables included in rules are readable, and the size of the rule set is manageable by a human user without external help	The size of the rule set becomes too large to be analyzed without decomposing it into small rule chunks	Rules have become so complicated (and the rule set size has grown so much) that mathematical tools are needed for inspecting the model behaviour	Not needed
General Additive Models	Variables and the interaction among them as per the smooth functions involved in the model must be constrained within human capabilities for understanding	Interactions become too complex to be simulated, so decomposition techniques are required for analyzing the model	Due to their complexity, variables and interactions cannot be analyzed without the application of mathematical and statistical tools	Not needed
Bayesian Models	Statistical relationships modeled among variables and the variables themselves should be directly understandable by the target audience	Statistical relationships involve so many variables that they must be decomposed in marginals so as to ease their analysis	Statistical relationships cannot be interpreted even if already decomposed, and predictors are so complex that model can be only analyzed with mathematical tools	Not needed
Tree Ensembles	✗	✗	✗	Needed: Usually <i>Model simplification</i> or <i>Feature relevance</i> techniques
Support Vector Machines	✗	✗	✗	Needed: Usually <i>Model simplification</i> or <i>Local explanations</i> techniques
Multi-layer Neural Network	✗	✗	✗	Needed: Usually <i>Model simplification</i> , <i>Feature relevance</i> or <i>Visualization</i> techniques
Convolutional Neural Network	✗	✗	✗	Needed: Usually <i>Feature relevance</i> or <i>Visualization</i> techniques
Recurrent Neural Network	✗	✗	✗	Needed: Usually <i>Feature relevance</i> techniques



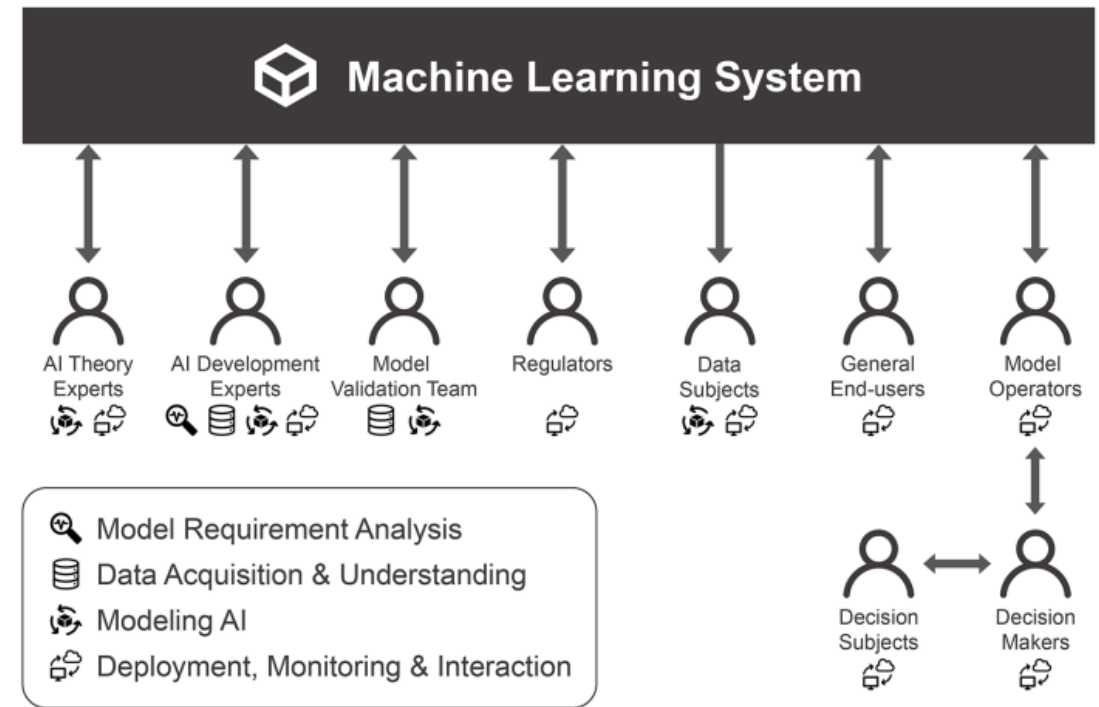
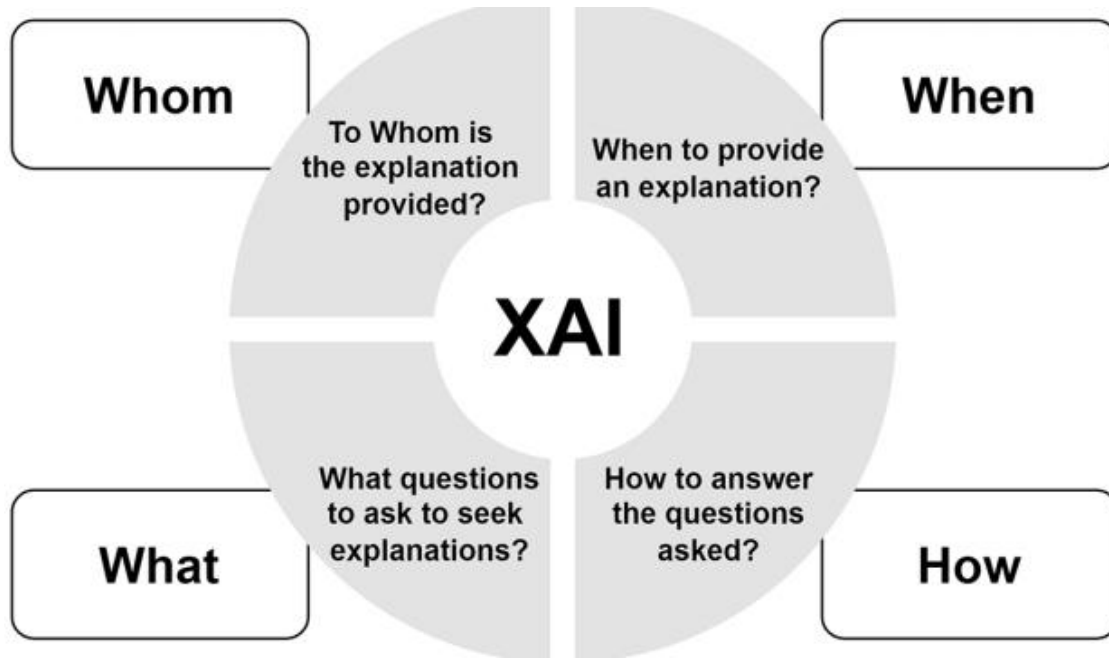


Third

A Roadmap of Explainable Artificial Intelligence: Explain to
Whom, When, What and How?

Summary

- Nine different stakeholders
 - Different stages of the AI system lifecycle
 - “what to explain”
 - Fore dimension of to whom, when, what, and how
 - XAI methods
 - bridge to connect stakeholders’ needs with XAI methods
 - Guideline : help stakeholders select the appropriate XAI method
-
- 针对不同的目标用户的不同需求，采用不同的方法方式呈现。



How Explanations: How does the system work as a whole?

Why Explanations: Why does the system make a particular decision?

Why-not Explanations: Why does the system not make a particular decision? What Explanations: What happens inside the system?

What-if Explanations: What would the system do if the input changes?

What-else Explanations: What else are the similar instances?

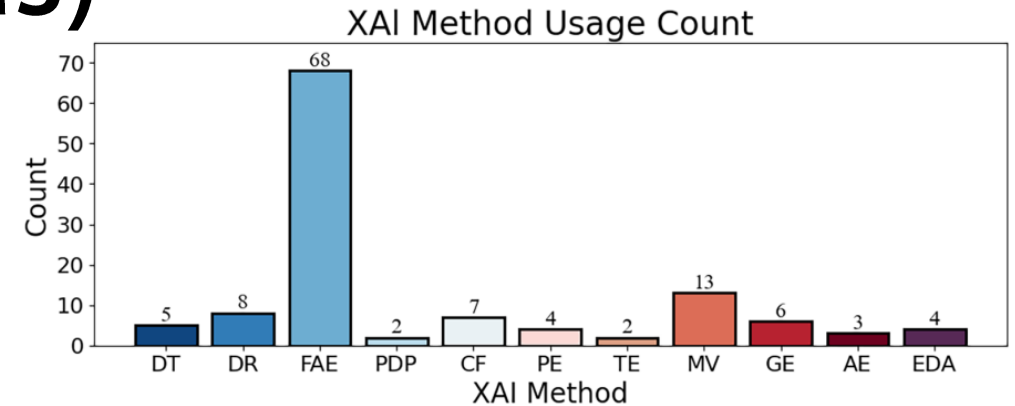
How-to Explanations: How to let the system make another particular decision?

How-still Explanations: How much of a perturbation can there be while maintaining the same decision?

Data Explanations: Ask for information about the data.

How to Explain (methods)

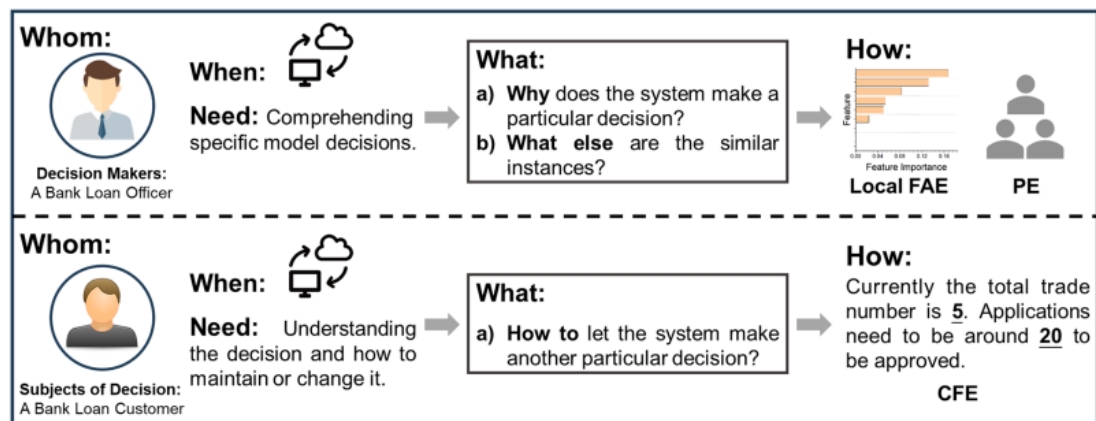
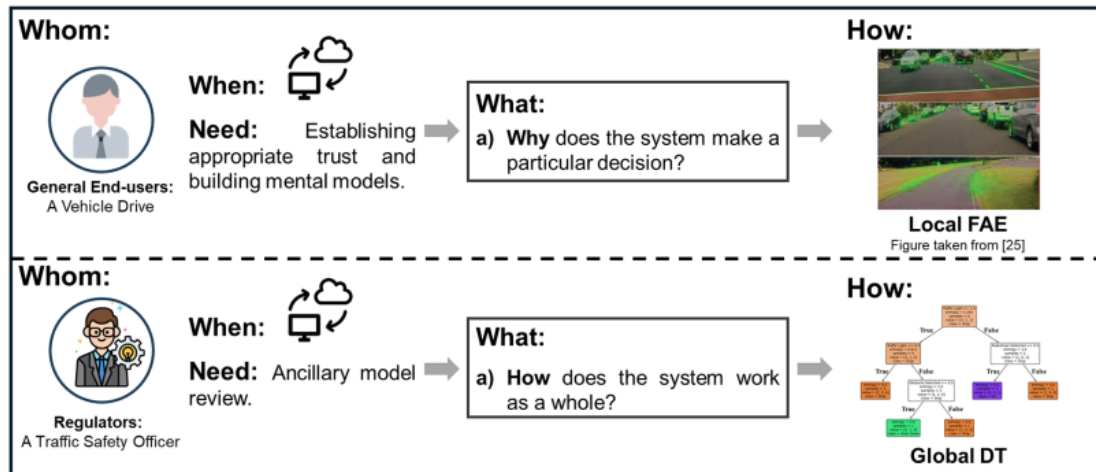
- ① Decision Trees (DTs) and Decision Rules (DRs)
- ② FAE : FAE provides the importance of each feature
- ③ PGP : provides a visual representation of how prediction results vary with different feature values
- ④ CF : CF explanations (CFE) and CF instance (CFI)
- ⑤ Prototype Explanation (PE)
- ⑥ Text Explanation (TE)
- ⑦ Model Visualization (MV)
- ⑧ Graph Explanation (GE)
- ⑨ Association Explanation (AE)
- ⑩ Exploratory Data Analysis (EDA)



"What to Explain" Question	XAI Methods
How explanations	Global DT, Global DR, Global FAE, GE
Why explanations	Local DT, Local DR, Local FAE, PE, TE, IpAE, OAE, Global DT, Global DR, CF, GE, ItAE
Why-not explanations	Local FAE, CF, Global DT, Global DR
What explanations	MV, GE, ItAE
What-if explanations	PDP, Global DT, Global DR, GE
What-else explanations	PE, CFI
How to be that explanations	CF, Local PDP, Global DT, Global DR
How to still be this explanations	Local DT, Local DR, PDP, Global DT, Global DR, PE
Data explanations	PE, EDA

Explain to Whom	When to Explain	Need	What to Explain	How to Explain
AI theory experts		Insight and understanding of the internal logic of complex ML models	What	<u>MV</u> , <u>GE</u> , <u>ItAE</u>
		Comparing analysis of multiple ML models	How, why, what	DT, DR, <u>FAE</u> , GE, PE, TE, AE, CF, <u>MV</u>
		Insight and understanding of datasets	Data	<u>PE</u> , <u>EDA</u>
		Analyzing potential errors, noise, and bias in the dataset	Data	PE, <u>EDA</u>
AI development experts		Assisting with feature selection	How, why	DT, DR, <u>FAE</u> , GE, PE, TE, AE, CF
		Optimizing model architecture and hyperparameters	How, why, what	DT, DR, <u>FAE</u> , GE, PE, TE, AE, CF, <u>MV</u>
		Checking the model's decisions	How, why, why-not, what, what-if, what-else, how-to, how-still	DT, DR, <u>FAE</u> , GE, PE, TE, AE, CF, <u>MV</u> , <u>PDP</u>
		Guiding model debugging and error refinement	How, why, why-not, what	DT, DR, <u>FAE</u> , GE, PE, TE, AE, CF, <u>MV</u>
		Adjusting the ML model to meet the user's expectations and needs	Why, why-not, what-if, how-to	DT, DR, Local <u>FAE</u> , PE, TE, AE, CF, GE, PDP
		Assessing the impact of dataset shift	How, why, data	DT, DR, <u>FAE</u> , GE, PE, TE, AE, CF, <u>EDA</u>
		Evaluating data suitability	Data	PE, <u>EDA</u>
Model validation team		Reviewing the ML model's decision logic	How, why, why-not, what-if, what-else, how-to, how-still	DT, DR, <u>FAE</u> , GE, PE, TE, AE, CF, PDP
		Determining compliance with regulations	How, why, why-not, what-if, what-else	DT, DR, <u>FAE</u> , GE, PE, TE, AE, CF, PDP
Model operators		Ensuring correct and efficient interaction	How, why, what-if, what-else	DT, DR, <u>FAE</u> , GE, PE, TE, AE, CF, PDP
		Comprehending specific model decisions	Why, why-not, what-else	DT, DR, Local <u>FAE</u> , PE, TE, AE, CF, GE
Decision makers		Deepening overall understanding of the ML model and improving decisions	How	Global DT, Global DR, Global <u>FAE</u> , GE
Regulators		Ancillary model review	How, why, why-not, what-if, what-else	DT, DR, <u>FAE</u> , GE, PE, TE, AE, CF, PDP
		Assisting in apportioning responsibility	Why	DT, DR, Local <u>FAE</u> , PE, TE, AE, CF, GE
Data subjects		Protecting personal data information	How, why	DT, DR, <u>FAE</u> , GE, PE, TE, AE, CF
Subjects of decision		Understanding the decision and how to maintain or change it	Why, why-not, what-if, what-else, how-to, how-still	DT, DR, Local <u>FAE</u> , PE, TE, AE, CF, GE, PDP
		Examining bias	How, why, why-not, what-if	DT, DR, <u>FAE</u> , GE, PE, TE, AE, CF, PDP
General end-users		Establishing appropriate trust and building mental models	How, why, why-not, what-if, what-else	DT, DR, <u>FAE</u> , GE, PE, TE, AE, CF, PDP
		Protecting personal data information	How, why	DT, DR, <u>FAE</u> , GE, PE, TE, AE, CF

Full Name	Abbreviation
Decision tree	DT
Decision rule	DR
Feature attribution explanation	FAE
Partial dependence plot	PDP
Counterfactual explanations/instance	CFE/CFI
Prototype explanation	PE
Text explanation	TE
Model visualization	MV
Graph explanation	GE
Input/Internal/output association explanation	IpAE/ItAE/OAE
Exploratory data analysis	EDA



任务分类1	任务分类1	任务分类2	任务分类2
C1	C2	C3	C4
S1	S2	S3	S4
用户A	用户B	用户B	用户A

$S1 \approx S4$
 $S2 \approx S3$
 $C1 \approx C2$
 $C3 \approx C4$

→

$C1+S1 \approx C1+S4 \approx C2+S4 \approx C2+S1$
 $C1+S2 \approx C2+S2 \approx C1+S3 \approx C2+S3$
 $C3+S2 \approx C4+S3 \approx C3+S3 \approx C4+S2$
 $C3+S1 \approx C4+S1 \approx C3+S4 \approx C4+S4$

