UNIVERSITÉ
LYON 1

21 Octobre 2024

XAI
Explainable Artificial Intelligence

Présenté par :

BÉRARD Thibaud
COLMANT Axel
GUILLAUME Jean-Luc
SAVIGNY Henri

Intelligence Artificielle & Cognition

Professeur : LEFEVRE Marie

Plan

I. Introduction : Contexte & Problématique

II. Techniques d’explicabilité

III. Applications et facteurs humains

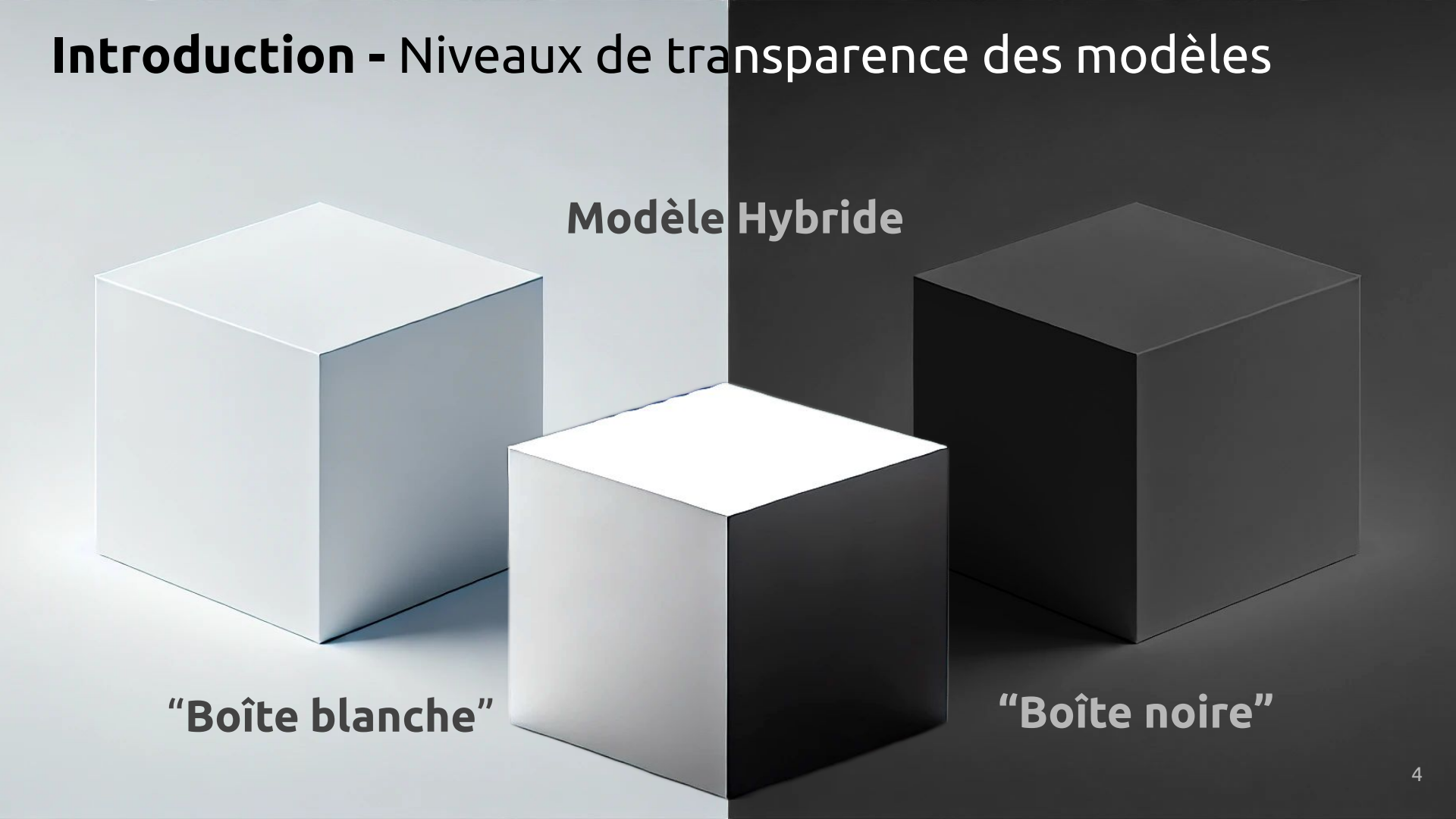
IV. Conclusion et perspectives

I

Introduction

1. Contexte
2. Problématiques

Introduction - Niveaux de transparence des modèles



Introduction - Problématiques

- Comment rendre l'IA **compréhensible** pour les humains ?
- Comment garantir la **confiance** dans des modèles d'IA opaques ?
- Comment éviter les biais et assurer des décisions **éthiques** et **équitables** ?

II

Techniques d’explicabilité

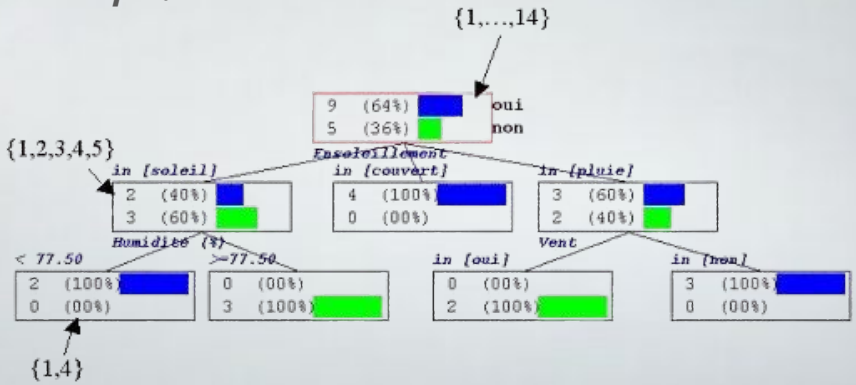
1. Catégorisation
2. Méthode de transparence
3. Méthode post-hoc
4. Taxonomie des techniques post-hoc

Techniques - Catégorisation

Modèle transparent

On conçoit des modèles d'IA intrinsèquement **interprétables**.

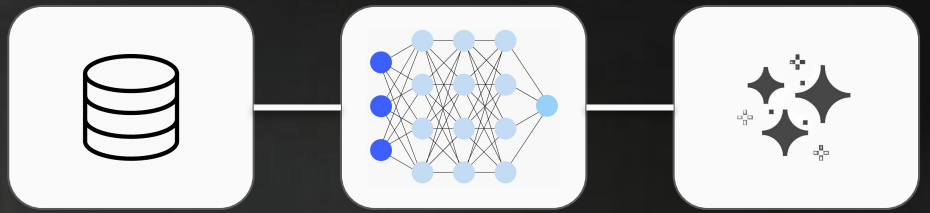
Exemple: *Arbre de décision*



Post-hoc

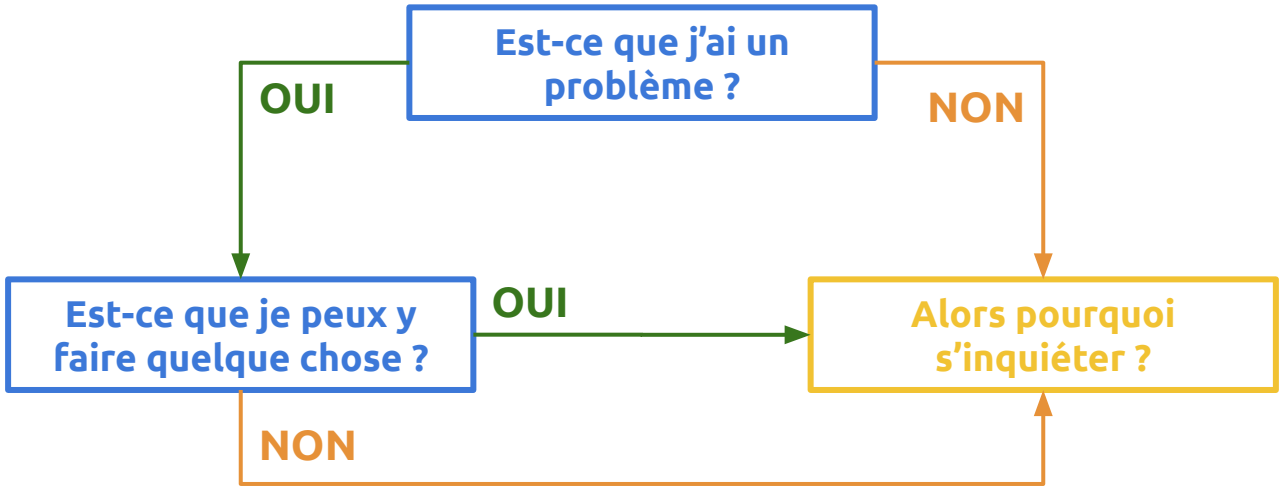
On analyse un modèle d'**IA déjà entraîné** pour comprendre ses décisions.

Classé en 3 catégories :



Techniques - Méthode de transparence (Exemple)

- Arbres de décision :

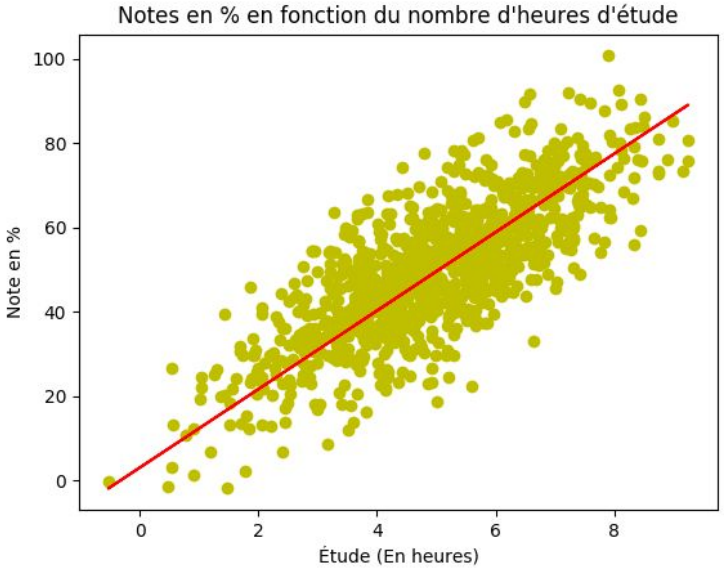


Techniques - Méthode de transparence (Exemple)

- Règles d'association :

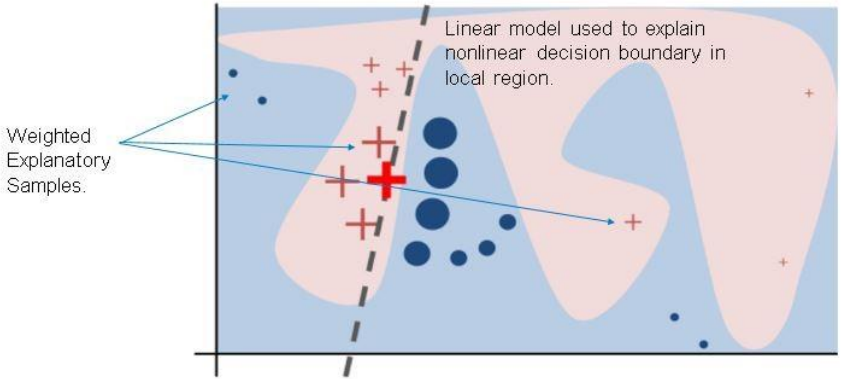
Techniques - Méthode de transparence (Exemple)

- Modèle linéaire :



Techniques - Méthode post-hoc (LIME)

Explique une décision spécifique en créant un modèle simple qui approxime le modèle complexe localement.



Left: Image of a bowl of bread. Middle and right: LIME explanations for the top 2 classes (bagel, strawberry) for image classification made by Google's Inception V3 neural network.

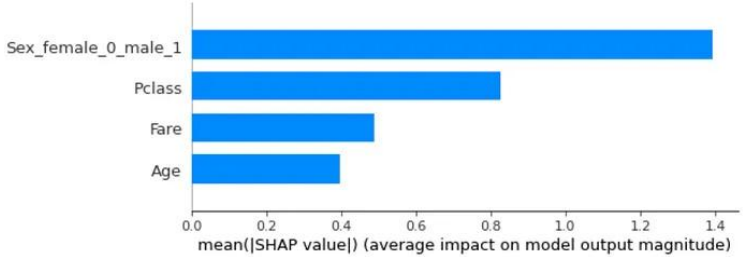
Techniques - Méthode post-hoc (SHAP)

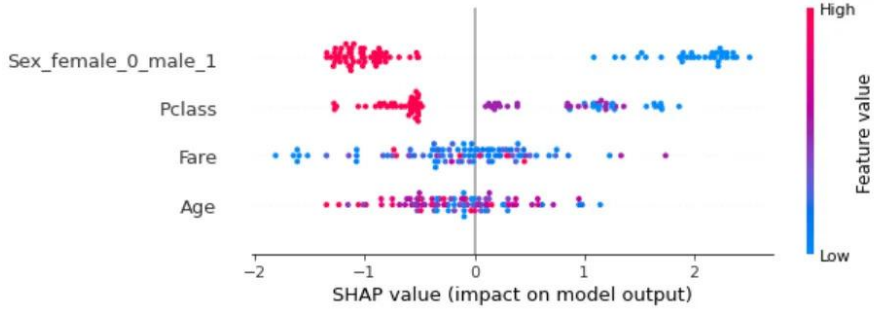
Calcule l'importance de chaque variable dans la décision de l'IA grâce aux SHAP values

- Perturbation des caractéristiques
- Calcul de la contribution moyenne
- Attribution des scores

Exemple SHAP : Titanic

	Age	Pclass	Fare	Sex
0	-0.135616	-0.532116	-0.089644	-0.771342
1	-0.101931	1.565495	-0.713686	2.148365
2	-0.343073	-0.516206	-0.288878	-1.171763
3	-1.012118	-0.567596	-0.402344	-0.853301
4	-0.263504	-0.857410	0.192260	1.900090
...	...	...	...	...

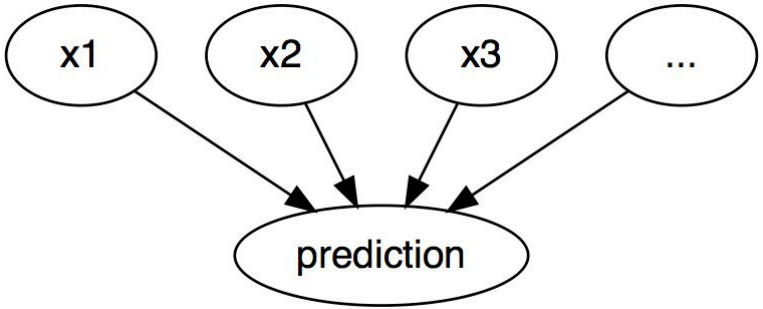




Techniques - Méthode post-hoc (Exemples contrefactuels)

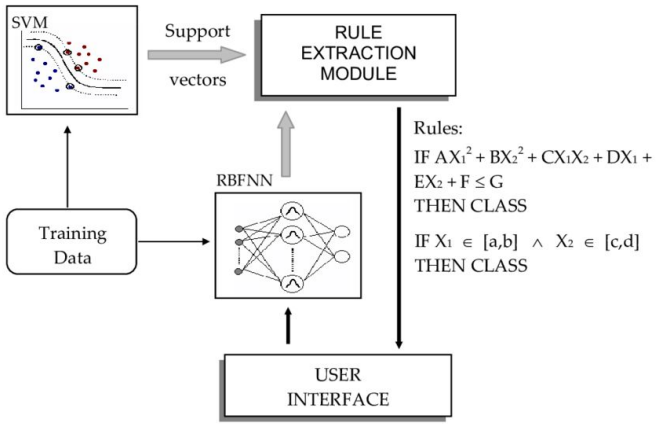
Comment modifier les données pour changer la prédiction ?

Explication contrefactuelles : décrit le plus petit changement sur les valeurs des caractéristiques qui change la prédiction vers une sortie prédéfinie.



Techniques - Méthode post-hoc (Extraction de règles)

Transformation du modèle en règles de décision lisibles par l'homme.



The rule extraction method from RBFNN and support vectors.

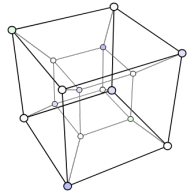
Techniques - Taxonomie des techniques post-hoc

- La taxonomie sert à **organiser** les méthodes d'explication

- **Objectif :**

  Catégoriser les techniques explicabilité en fonction de leurs caractéristiques afin de guider le choix de la méthode la plus appropriée pour une situation donnée.

Techniques - Taxonomie des techniques post-hoc



- **Représentable en 4 dimensions :**

  - Portée de l'explication :
    Globale (comportement général du modèle)
    Locale (explication d'une prédiction spécifique)

  - Niveau de détail
    Élevé (mécanismes internes du modèle)
    Faible (idée générale du fonctionnement)

  - Méthodes d'explication
    Sensibilité, gradients, exemples, règles, décomposition.

  - Type de données
    Images, texte, données tabulaires

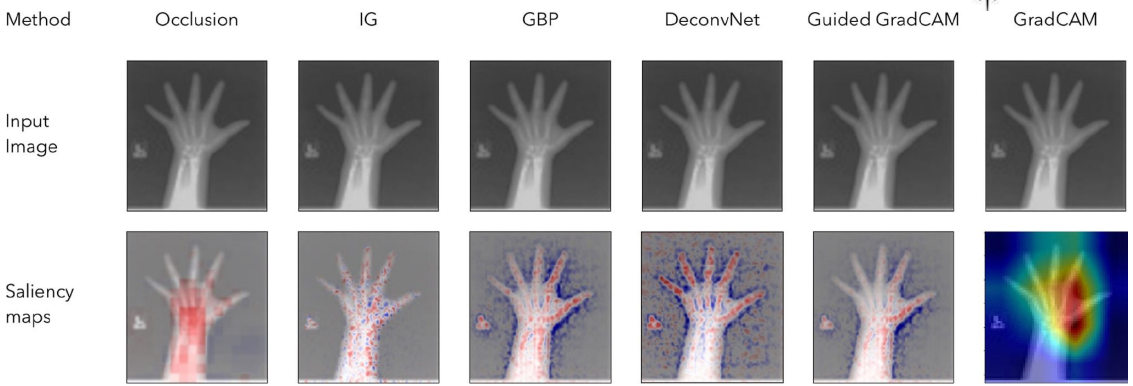
III

Applications

Applications - L'XAI au service de l'humain (Santé)

L'IA pour :

- Détecter des maladies
- Recommander des traitements
- Assister les chirurgiens
- …



Explainable AI in medical imaging: An overview for clinical practitioners – Saliency-based XAI approaches

L'XAI pour :

- Expliquer le **raisonnement** de l'IA
  Quelles données médicales ont influencé le diagnostic ?
- Renforcer la **confiance** du médecin
- Améliorer la **collaboration** humain-IA

Applications - L'XAI au service de l'humain (Finance)

L'IA pour :

- Détecter les fraudes
- Évaluer les risques
- Automatiser les décisions de crédit
- …

L'XAI pour :

- Expliquer des **facteur d'influence de décision** : *revenus, historique de crédit, …*
- Lutter contre les **discriminations** et promouvoir l'**inclusion** financière
- Renforcer la **confiance** des clients et des investisseurs

Applications - L'XAI au service de l'humain (Éducation) 

L'IA pour :

- Analyser les données de l'élève (réponses, interactions, etc.).
- Identifier les forces, les faiblesses et le style d'apprentissage de chaque élève.
- Proposer des exercices, des ressources et des activités adaptés.
- …

L'XAI pour :

- **Expliquer** les recommandations pédagogiques de l'IA.
- Aider l'élève à **comprendre** ses erreurs et à progresser.
- Promouvoir un apprentissage plus personnalisé et efficace.

Exemple : "L'XAI explique pourquoi l'IA recommande un exercice spécifique à l'élève, en fonction de ses difficultés et de son style d'apprentissage."

Applications - L'XAI au service de l'humain (Justice)

L'IA pour :

- Évaluer les risques de récidive
- Personnaliser les peines
- …

L'XAI pour :

- **Expliquer** les facteurs pris en compte : *Âge, antécédents, …*
- Détecter les **biais** potentiels de l'IA
- Garantir une justice **équitable**
- Renforcer la **confiance** du public dans le système

IV

Conclusion

Conclusion - Collaboration humain-IA

L’**IA** apporte sa puissance de calcul, sa rapidité et sa capacité à traiter de grandes quantités de données.

L'**Humain** apporte son expertise, son intuition et sa capacité de jugement

L’**XAI** facilite la communication et la compréhension mutuelle entre l'homme et l'IA

L’XAI permet de créer une synergie entre l'intelligence humaine et l'intelligence artificielle.

Sources

- Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI
  https://www.researchgate.net/publication/338184751_Explainable_Artificial_Intelligence_XAI_Concepts_Taxonomies_Opportunities_and_Challenges_toward_Responsible_AI

- “You just can’t go around killing people”, Explaining Agent Behavior to a Human Terminator
  https://openreview.net/pdf?id=0BdPwBncot

- Generating Global Policy Summaries for Renforcement Learning Agents Using Large Language Models
  https://drive.google.com/file/d/1XF3jb4B_6V7FCanZrpO-bWl4ZzRdzVIR/view

- Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning
  https://www.ijcai.org/proceedings/2019/0876.pdf

Sources

- Interpretable Machine Learning : A Guide for Making Black Box Models Explainable - *Christoph Molnar*
  https://christophm.github.io/interpretable-ml-book/

L'XAI : votre avis compte !

Faut-il se concentrer sur l'explicabilité des modèles ou sur la transparence des données utilisées pour les entraîner ?

L'XAI peut-elle réellement clarifier la responsabilité en cas de problème lié à l'IA ? Qui est responsable des décisions d'un système explicable ?

Peut-on garantir que les explications fournies par l'XAI sont objectives et impartiales, ou sont-elles toujours sujettes à interprétation ?

L'XAI risque-t-elle de créer une confiance aveugle en l'IA, en masquant ses limites potentielles ?

L'XAI sera-t-elle cruciale pour comprendre et contrôler une future AGI (Artificial General Intelligence) ?