



# — BIEN COMPRENDRE POUR BIEN DÉPLOYER: l'explicabilité des modèles d'IA, une nécessaire analyse

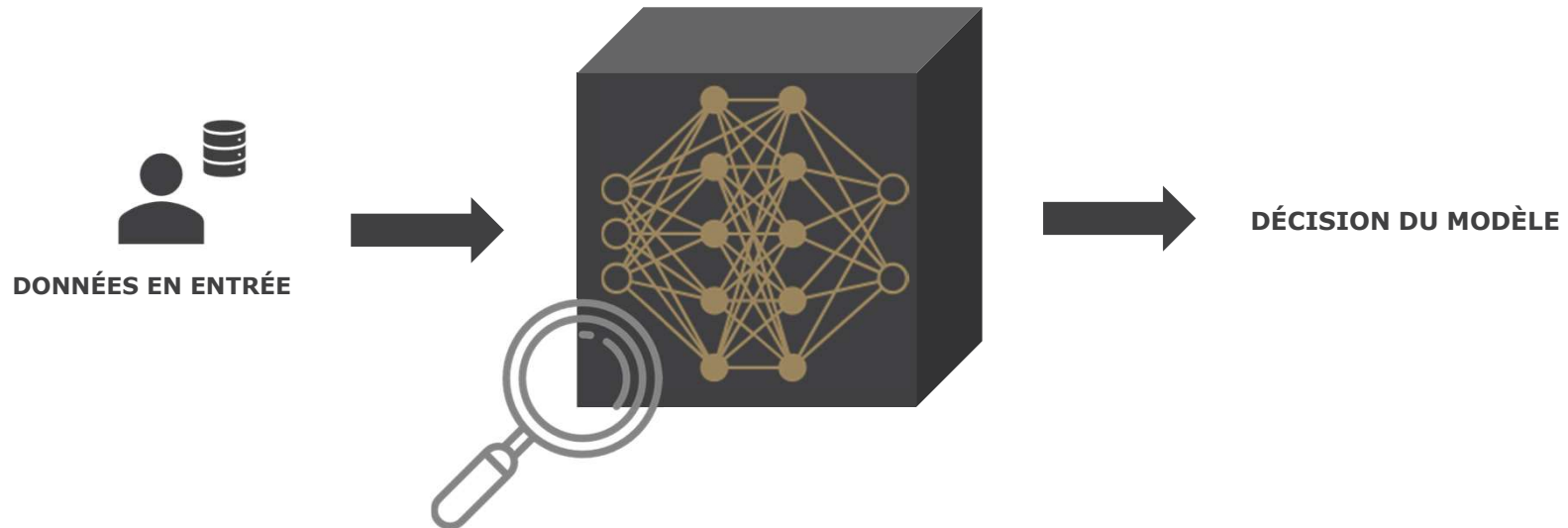
Sara MEFTAH

07/11/2024

# Explainable AI

CRUCIALE POUR OUVRIR LA BOITE NOIRE

---



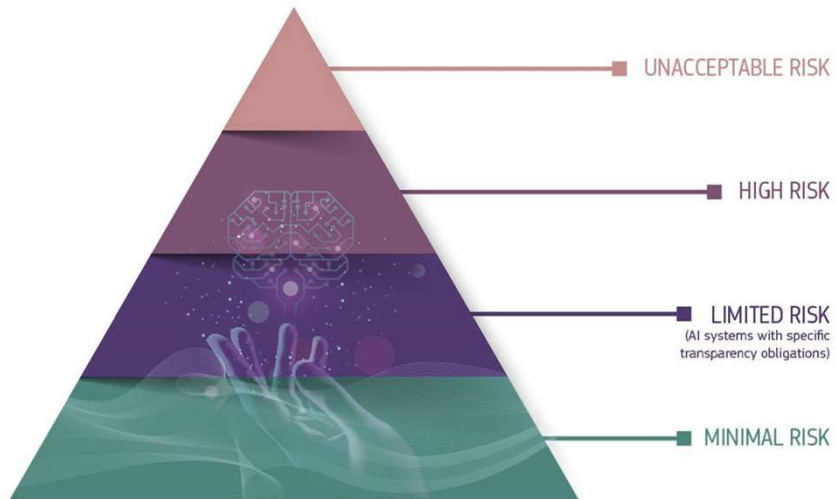
- Pourquoi le modèle a pris cette décision et pas une autre ?
- Peut-on améliorer les décisions de ce modèle ?
- Pourquoi le modèle se trompe sur certaines décisions ?



# Explainable AI

## DANS LE CONTEXTE DE L'AI ACT

### PLUSIEURS NIVEAUX DE RISQUE POUR LES SYSTÈMES D'IA



### LES SYSTÈMES D'IA À HAUT RISQUE

- **Transparence (Article 13)** : Les systèmes d'IA à haut risque doivent être transparents pour que les utilisateurs puissent comprendre et utiliser correctement les résultats produits.
- **Supervision humaine (Article 14)** : Les opérateurs humains doivent comprendre les capacités et limites des systèmes IA, tout en étant formés pour éviter un biais d'automatisation.
- **Système de gestion des risques (article 9)** : Les fournisseurs sont tenus d'identifier les risques et de mettre en œuvre des contrôles pour y faire face.

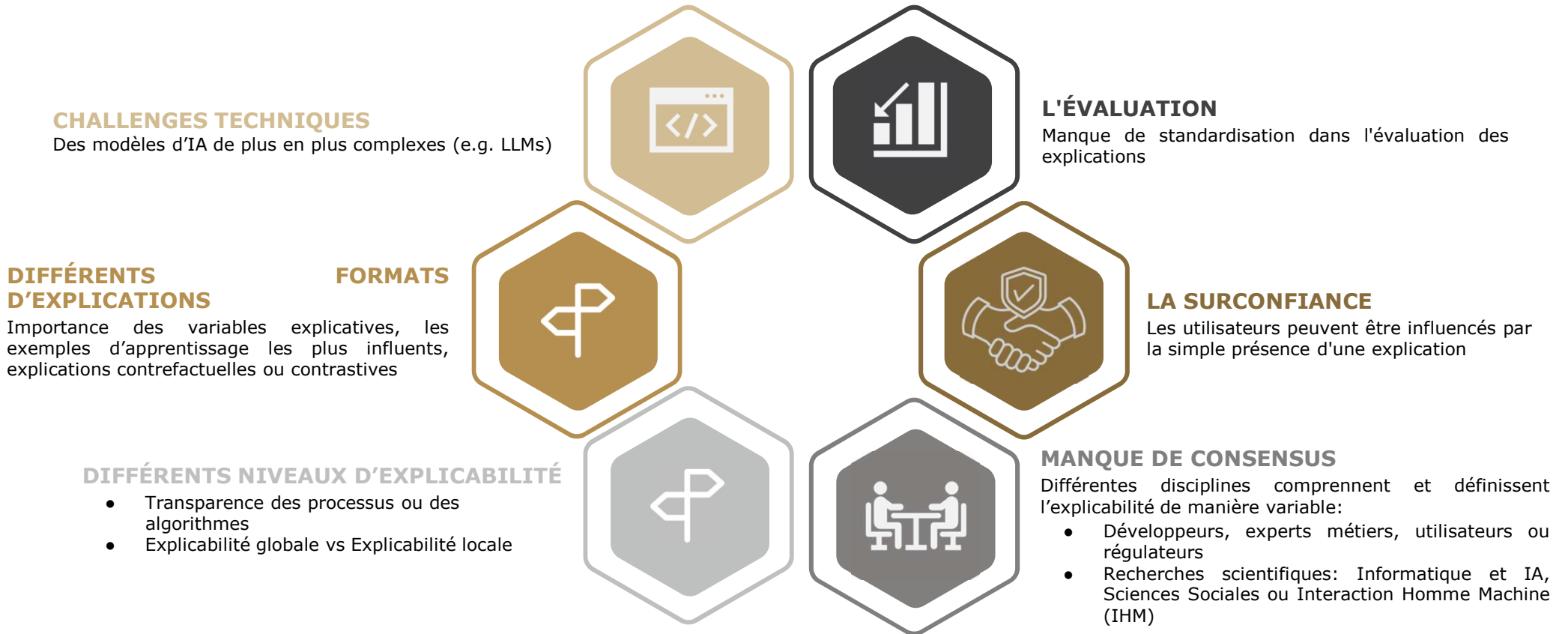
### AU-DELÀ DE LA CONFORMITÉ AVEC L'AI ACT

- Atténuation des risques
- Amélioration continue des performances
- Renforcement des pratiques éthiques de l'IA
- Détection des biais



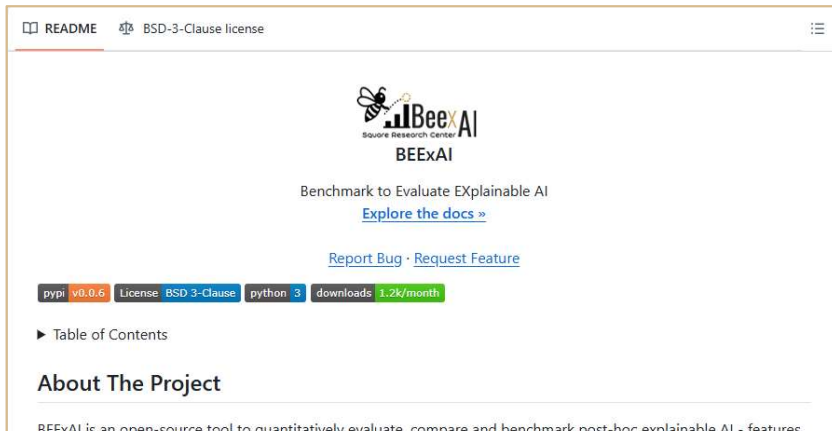
# Explainable AI

## LES BARRIÈRES À UN ENCADREMENT SOLIDE



# Explainable AI

LES TRAVAUX DU SQUARE RESEARCH CENTER



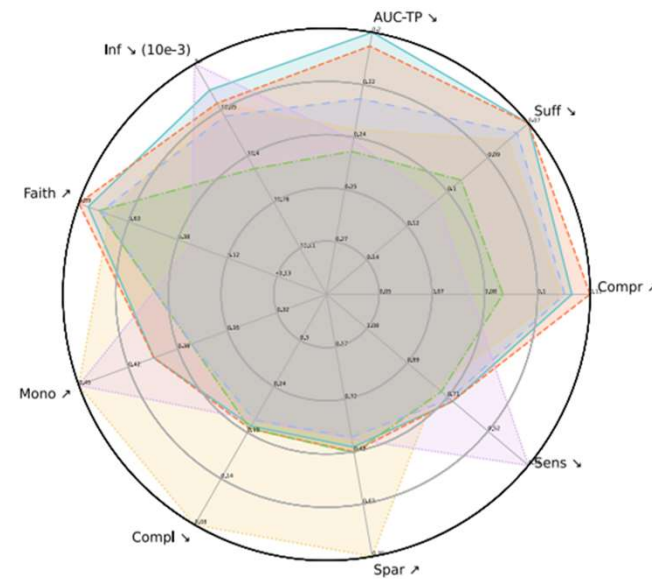
## BEEExAI: Benchmark to Evaluate Explainable AI

Samuel Sithakoul<sup>1,2</sup>, Sara Meftah<sup>1</sup>[0009-0009-3297-5682], and Clément Feutry<sup>1</sup>[0009-0009-6086-762X]

<sup>1</sup> Groupe Square Management, Square Research Center, 173 Avenue Achille Peretti, 92200 Neuilly-sur-Seine, France

<sup>2</sup> CentraleSupélec, 3 Rue Joliot Curie, 91190 Gif-sur-Yvette


**Abstract.** Recent research in explainability has given rise to numerous post-hoc attribution methods aimed at enhancing our comprehension of the outputs of black-box machine learning models.<sup>1</sup> However, evaluating the quality of explanations lacks a cohesive approach and a consensus





173 Avenue Achille Peretti  
92200 Neuilly-sur-Seine  
+33 (0)1 46 40 40 00

 Square Management

 @square\_managem