Interprétabilité des modèles prédictifs

A la lumière des enjeux éthiques et réglementaires, je reviendra sur le dilemme (si c'en est vraiment un) entre prédire et comprendre. Je distinguerai les modèles explicables des modèles interprétables. qui incluent les modèles causaux , ce qui permettra de faire le lien avec l'ASI.
On parle en effet beaucoup en ce moment de X-AI (explainable artificial intelligence) qui comprend des approches globales ou locales, agnostiques ou spécifiques sur l'importance des variables, des modèles de substitution pour les boites noires et d'autre part d'éthique, de confiance, d'équité des algorithmes.

Plan

1. Retour sur le dilemme comprendre-prédire
2. Nouvelles contraintes, nouveaux enjeux
3. Explicabilité (X-AI)
    a. Post-hoc interpretation
    b. Global versus local ; agnostique ou spécifique
    c. Importance des variables
4. Modèles interprétables
    a. Modèles simples
    b. Outils de visualisation IPA
    c. Causalité : réseaux bayesiens et graphes implicatifs
5. Ethique des algorithmes
    a. Biais des algos ou biais des données ?
    b. Fairness
6. Conclusion
    a. Faut-il renoncer aux boîtes noires ?
    b. Pour un permis de conduire les algorithmes ?

Analogie avec une automobile : fonctionnement réel boîte noire, mais normes, code de la route etc.