Title page

- Explainable AI is Dead, Long Live Explainable AI!
  - Titre aguicheur, aime pas, deuxième mieux
  - Hypothesis-driven Decision Support using Evaluative AI
- Papier pour aide à la décision
- Tim miller, chercheur australien, il publie sur des sujets assez varié en XAI
  - XAI 4 Multi agent, XAI 4 RL
  - Humain-AI interaction
  - AI assisted décision support

Introduction

Quick summary

- Argumente Changement paradigme XAI aide décision

Framework evaluative AI :

- Centrée sur l'humain
- Au dela des recommandation
- En evaluant les hypothèse du décideur
- **Mitiger Excès de confiance**

Over/Under reliance

Définition

- Aide à la décision, retrouve deux phénomènes
- Expliquer les défininions :
  - Excès de confiance : accepter les recommandations, même si faux
  - Manque de confiance : inverse : rejeter les recommendation, même si vrais
- == Automation bias :
  - **Pourquoi j'ai choisi ce papier**
  - La confiance excessive → Problématique
    - Au quotidient
      - correcteur d'orthographe
      - Recrutement ; Prêt banquaire ect
      - Enjeux plus sérieux : Unité de soin intensif, aviation, centrale nucléaire
      - Algo mesure du risque de violence conjugale → Excès de confiance → erreur d'estimation du danger
  - Echo domaine sciences cognitives licence -> manque un peu

Causes

- Manque engagement cognitif, l'esprit humain minimiser effort
- Ajout XAI qui explique
  - ~Biais de confirmation sur les explications : accepter ou rejeter

Solutions

- Solution 1 : forcer l'engagement cognitif
  - Généralement : forcer les gens donner décision avant machine
  - Pas giga efficace
  - Pas trop apprécié
- Solution 2 : Un paradigme shift en XAI 🤔😌💡
  - == papier
  - => Avant parler cela -> définir des critères plus claire

How we make decisions?

- naturellement : identifier option, comparer option, choisir option
- Des gens plus réflechis:
  - **dans notre cadre**, pour les system d'aide à la décision, résumé tout 4 points... DIAPO

10 cardinal decision issue

Bon system d'aide à la décision, besoin de

- Option: identifier, lister, réaliste/fesable
- Opinion & Possibilité : Proba et impact positif / negatif possible pour chaque options
- Compromis: comparer ce qu'on tout ce qu'on a dit au dessus
- Understand: comprendre le systeme d'aide à la décision

Does current decision support align with those criteria?

- System actuel ?? respectent ces critères

No explanotory information

- Cas classique d'automatisation des décisions : *décrire un peu*
- gens -> ignorer le system // soit accepter des mauvaise décisions
- Le décideur : Calibration de la confiance uniquement sur :
  - l'accuracy du model
  - Son expértise
- => Novice : se repose sur le systems // expert : utilise leur propre expertise
- Mais ne coche aucune des cases
  - X identifier les autres options probable
  - Opinion **uniquement** autour de la recommandation
  - X faire des compromis
  - X expliquabilité
- Mais est-ce que c'est quand même utile ?? OUI
  - D'accord -tout roule
  - Pas d'accord -> reconsidérer le choix
  - -> meilleurs décision
  - En pratique : non

With explanatory information

- Outil de XAI en plus
- Coche plus de case
  - -> Comprendre le modèle
  - -> faire des compromis : SHAP, counterfactual
  - X identifier les autres options probables
  - Jugement et possibilité uniquement autour de la recommandation
- => Toujours pour défendre la recommandation
- Est-ce que c'est quand même utile ?? OUI

    - Même raison que précédement :
        - Si pas d'accords -> regarder -> meilleurs décision
    - En pratique == pas le cas
- Un model interprétable coche uniquement la dernière case

Cognitive forçing

- Cognitive forcing : décideur donne décision avant machine
- Coche le plus de case paradigme actuel // toujours des problèmes
    - décideur voie plus d'option : forcé de les chercher
    - Toujours avec les outils XAI, on peut comprendre le modèle
    - et faire des compromis par exemple avec SHAP ou les counterfactual
- -> System toujours centré sur sa recommandation
- dès que centré autour de la recommandation == case partiellement coché
- => Sortir de ce paradigme de recommandation unique => evaluative AI

The evaluative AI framework

- Décrire : boucle, décideur -> HP -> feedback
- Le paradigme est inversé :
    - c'est la machine qui donne son avis sur la décision du *decision-maker*
    - Et non le décision maker qui donne son avis sur la décision de la machine

Properties

- Exemple d'interface
    - Potentiel mélanome ?
    - Vu sur toutes les options possible
    - intéraction avec l'utilisateur
    - hypothèse pour, hypothèse contre

Zoom on properties

- Naturellement leur modèle coche toutes les cases
- Option

- Trade-off
    - Réussi le mieux
    - Pour ou contre clair -> décideur bonne overview
    - Papier :"bon décideur" = personnes qui regarde les arguments qui vont contre leurs conclusions initiales
- => extrapolation sur de l'IRL
    - Décision complexe, type choix de stage, orientation
        - Regarde tous les pours et contre == bourbier
        - // fier a l'instincs et indentifier les contre serait plus efficace
    - les discutions IRL ?
        - Clé = être à l'écoute, tourner autour de l'opinion de l'autre sans forcément directement relate sur des pov personel

Limits

- Pourquoi les gens s'engagerai avec ce system et pas les autres méthode
    - Plus de controle
    - Proche de la manière dont on fait des décisions (identifier, comparer, choisir)
    - X : pas de preuve de ça dans le papier (en psychologie ça serait pas passé, jsp pour Humain-AI interaction)
- Méthode qui charge mentalement le décideur
    - X: toujours la moins aimé surement
    - Auteur se défend : quand même moins d'info

Mes critiques

- Les critères sont dur à différencier
    - Y'en a 10 de base, il en garde 6, 1 n'est jamais remplis, et 2 fusionne en 1 car proche (opinion et possibilité)
    - Des fois c'est dur de s'y retrouver, le tableau résumé est pas forcément accords avec que qui est dit dans le texte,
- Quand j'ai été voir la page wikipedia de l'automation bias, elle est assez remplis et l'autheur en parle pas du tout. Y'as pas mal d'autre facteur décrit et j'arrive pas à voir pourquoi y'a pas un mot dessus dans le papier
    - A la place l'intro parle du résonnement abductif pour appuyer son modèle comme un modèle proche de la manière naturel de la décision
    - Alors qu'il aurait eu la place car beaucoup de répétition dans son papier

Mes points forts du papier

- S'attaque à un vrais problème
- Avec une proposition forte, position pas facile à tenir
- Pas d'expérience pour appuyer l'évaluative AI
    - mais donne une liste exhaustive de piste de recherche dans la direction de l'évaluative AI

CONCLUSION

- auteur propose de changer de voix pour le XAI appliqué l'aide à la décision
- Qu'il faut arreter d'expliquer les recommendation et se focus sur l'utilisateur et ces hypothèses
- En se rapprochant de la manière dont on prends naturellement des décisions

---

--- Je garde pour les questions au cas où ---

Differences with cognitive forcing

- Apparament ça ressemblerai pas mal au technique de cognitive forçing
- Les auteurs essaye plusieurs fois de se différencier à travers le papier
- Ici la clé c'est que le décideur est en position de contrôle face à la machine, "machine in the loop"
- également que ça suit un chemin de décision plus naturelle

Long live XAI

- Le titre est pas vraiment clair au première abord mais il se défend
- L'auteur ne veut pas se séparer de l'XAI ou faire une refonte
- Il veut améliorer une petite branche de l'XAI
    - Evaluative AI $ \subset $ XAI
- XAI + approche basé sur la recommendation sont bien et adapté dans certains cas
    - Making decision at scale
- Il faudra toujours un model recommendation based pour n'importe quelle XAI technique
- Outil de XAI existant -> déjà adapté à l'evaluative AI
    - Counterfactuals
    - Feature importance (SHAP)
    - Wieghts of Evidence, case-based reasoning techniques