

INTELLIGENCE ARTIFICIELLE

**ADMINISTRATIONS, PROPOSEZ
VOS PROJETS D'EXPÉRIMENTATION
DANS LES SERVICES PUBLICS !**



Atelier AMI IA 2 #3 : Explicabilité des algorithmes

Jeudi 9 juillet, 14h-16h



Programme

14h00 - 14h50 : Introduction des enjeux et réflexions autour d'exemples

14h50-15h05 : Fondements théoriques de l'explicabilité

15h015 - 15h45 : Ateliers d'identification des enjeux d'explicabilité dans les projets

15h50 - 16h00 : Conclusion et mise en commun

Atelier AMI IA 2 #3 : Explicabilité des algorithmes



Séquence 1:

Introduction des enjeux et réflexions autour d'exemples

Les objectifs



Comprendre

un résultat, une décision, une prédiction, la motivation pour construire un système, un impact, un biais, des **choix**

Discuter

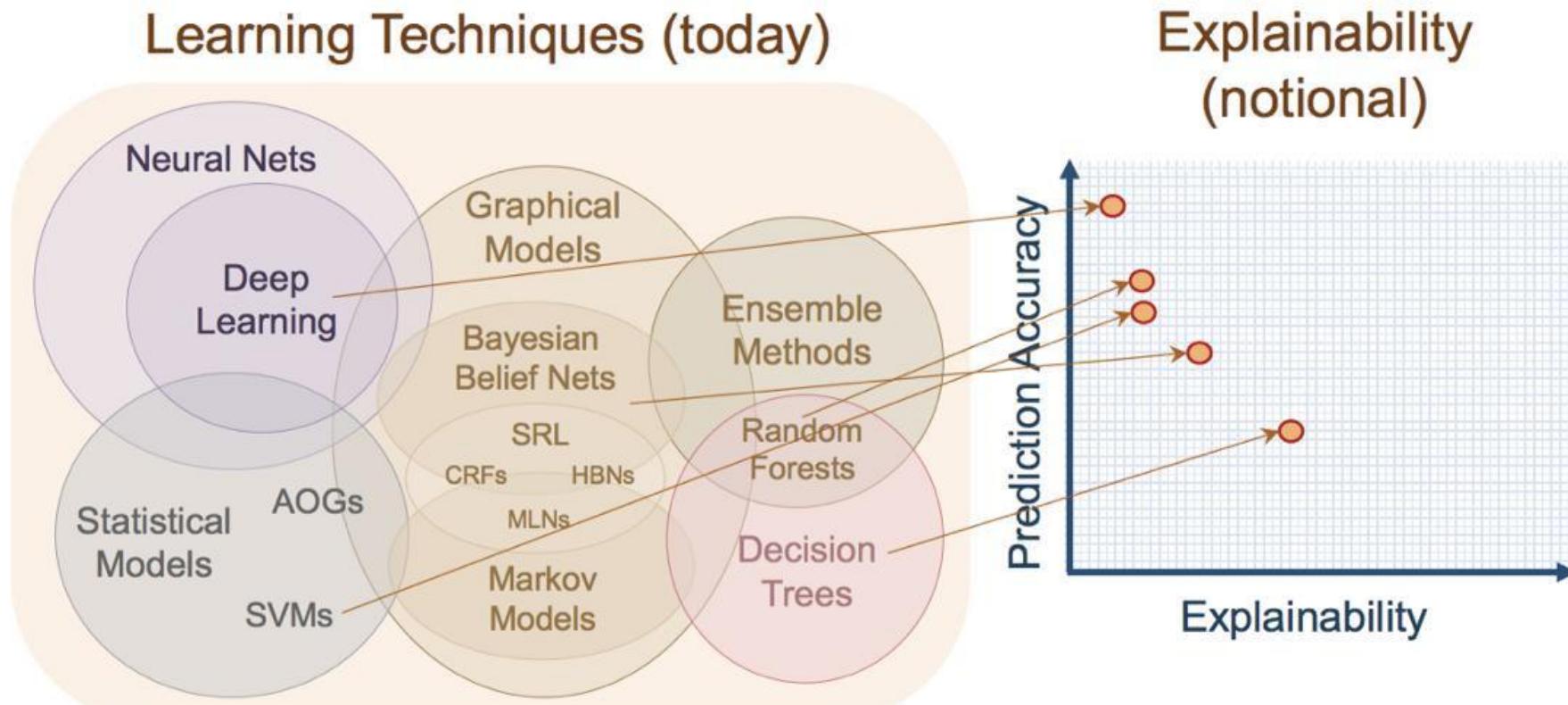
Contester

Pourquoi on en parle ?

	Principles et critères	Ethics of AI	Ethics of Data	Ethics of Interaction	Ethics of Intelligence	Ethics of Algorithm	Ethics of Design	Ethics of Impact	Ethics of Privacy	Ethics of Transparency	Ethics of Accountability	Ethics of Fairness, non-discriminates, justice	Ethics of Transparency, audited	Ethics of Auditing, accountability	Ethics of Implementability, interoperability	Ethics of Publicity, inclusion, social relevance	Ethics of Common policy tool	Ethics of Legislation framework, legal status of AI systems	Ethics of Responsibility, limited research funding	Ethics of public assessment, education about AI and its risks	Ethics of Employment	
Auteurs	Prates et al. (2018)	Bostrom et al. (2018)	Hwang et al. (2018)	Brockmeier et al. (2018)	Domingos et al. (2018)	Ghoshal et al. (2018)	Gómez et al. (2018)	Griffiths et al. (2018)	Hannak et al. (2018)	Hannak et al. (2018)	Hannak et al. (2018)	Hannak et al. (2018)	Hannak et al. (2018)	Hannak et al. (2018)	Hannak et al. (2018)	Hannak et al. (2018)	Hannak et al. (2018)	Hannak et al. (2018)	Hannak et al. (2018)	Hannak et al. (2018)	Hannak et al. (2018)	Hannak et al. (2018)
Key word	AI principles of the UN	Analysis of whose principles for the beneficial use of AI	Large application of whose principles	Comments on several implications of AI	Comments on several implications of AI	Comments on several implications of AI	Code of ethics developed by the professional community	Several codes of ethics developed by the professional community	Detailed description of ethical aspects in the context of AI	Detailed description of ethical aspects in the context of AI	Several guidelines related to the ethical use of AI	Several codes of ethics developed by the professional community	Several codes of ethics developed by the professional community	Several codes of ethics developed by the professional community	Several codes of ethics developed by the professional community	Several codes of ethics developed by the professional community	Several codes of ethics developed by the professional community	Several codes of ethics developed by the professional community	Several codes of ethics developed by the professional community			
Principles																						
accountability																						
fairness, non-discriminates, justice																						
transparency, audited																						
auditing, accountability																						
common good, sustainability, well-being																						
human oversight, control, auditing																						
implementability, interoperability																						
publicity, inclusion, social relevance																						
common policy tool																						
legislation framework, legal status of AI systems																						
responsibility, limited research funding																						
public assessment, education about AI and its risks																						
employment																						

The Ethics of AI ethics
An evaluation of Guidelines
Thilo Hagendorff, 2019

Pourquoi on en parle ?



Pourquoi on en parle ?



Expliquer les algorithmes publics

Introduction

Les algorithmes publics

À qui est-il destiné ?

Comment contribuer ?

Décision automatisée ou aide à la décision ?

Algorithmes du secteur public vs. algorithmes du secteur privé

Comment les administrations peuvent-elles "rendre des comptes" sur l'usage des algorithmes publics ?

Comment rendre des décisions "justes" à l'aide des algorithmes ?

Qui est concerné par la transparence des algorithmes ?

Quelles sont ces obligations en matière de transparence ?

Un outil pour tester vos connaissances

Les prochaines étapes

Les chantiers d'Etalab

3 - Le cadre juridique applicable

La loi pour une République numérique, et plus récemment le Réglement sur la protection des données à caractère personnel (RGPD) ont introduit de nouvelles dispositions concernant les algorithmes publics. Ces dispositions visent à introduire une **plus grande transparence** et une plus grande **redevabilité** de l'administration dans l'usage de ces systèmes, en particulier quand ils sont utilisés pour prendre des décisions.

Qui est concerné par la transparence des algorithmes ?

Le code des relations entre le public et l'administration (CRPA) précise le périmètre des administrations et des traitements concernés.



Si:

- vous êtes **une administration d'Etat, une collectivité, un organisme de droit public ou de droit privé intervenant dans le cadre d'une mission de service public** ([article L.300-2](#)),
- vous utilisez un **traitement algorithmique** (cf. la [définition](#) ci-dessus),
- à l'aide de ce traitement, vous prenez des **décisions administratives individuelles envers des personnes physiques ou morales, de droit public ou privé nommément désignées**,
- et que ce traitement n'est **pas couvert par l'un des secrets définis par la loi** ([2^e de l'article L.311-5](#))
, et notamment: délibérations du Gouvernement, défense nationale, conduite de la politique extérieure, sûreté de l'Etat, sécurité publique, sécurité des personnes ou des systèmes d'information, recherche et prévention d'infractions, etc.

L'importance du contexte

Pourquoi cette voiture est-elle rouge ?



- 1) Parce qu'elle renvoie une lumière dont la longueur d'onde est de 655 nm
- 2) Parce que quelqu'un l'a peinte en rouge
- 3) Parce que personne ne l'a peint en bleu

Les explications sont TOUJOURS contextuelles.

Les 5 questions à se poser

POURQUOI expliquer ?

POUR QUI expliquer ?

QUOI expliquer ?

QUAND expliquer ?

COMMENT expliquer ?

Explication globale (Booking.com)

- **Nos préférés (classement par défaut) :**

Le classement par défaut se présente sous la forme « Nos préférés ». Il se fait via un système de classement entièrement automatique. L'algorithme utilisé prend en compte de multiples critères incluant non seulement la popularité d'un fournisseur auprès de ses clients, les tarifs appliqués, mais aussi l'historique du service client et certaines informations liées aux réservations ainsi que le pourcentage de commission et le respect des délais de paiement de la commission.

- **Tarif le plus bas en premier :**

les offres s'afficheront dans l'ordre du tarif le plus bas en premier au tarif le plus élevé.

- **Note des commentaires clients et tarif :**

les offres s'afficheront dans l'ordre du meilleur rapport qualité/prix en premier au moins bon rapport qualité/prix.

Exemple: booking.com

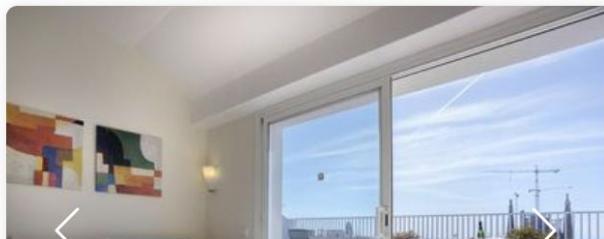
Explication globale (Homelidays.com)

Prix ▾ Chambres ▾ Confirmation Immédiate ▾

[Plus de filtres](#)

[Découvrez comment](#) les résultats de recherche sont classés

1 - 50 sur 300+



Consultée 165 fois au cours des 48 dernières heures

Friendly Rentals Las Terrazas

Appartement • 3 Ch • 1 Sdb • 5 Pers. • 70 m²

Comment les annonces sont-elles classées sur notre Site ?

×

Les annonces de location sont triées automatiquement à l'aide d'un algorithme tenant compte des critères sélectionnés par le vacancier sur la page de recherche. Les critères pouvant être sélectionnés sont les suivants :

Tri standard, Prix croissants, Prix décroissants, Appréciations les plus élevées, Disponibilités Récemment Mises à Jour, Nombre de chambres croissant, Nombre de chambres décroissant, Confirmation immédiate, Nombres d'Appréciations. Par ailleurs, le paramètre « Plus de filtres », la

selection d'une fourchette de prix, du nombre de chambres ou encore la selection d'une confirmation immédiate permettent d'inclure ou d'exclure les différentes options qui sauront répondre à vos besoins. Si aucune option spécifique de tri n'est sélectionnée, nous vous proposerons automatiquement par défaut le classement et l'affichage des annonces sur notre Site selon les caractéristiques de votre propriété, la qualité de l'expérience, et les frais de service payés par le Vacancier (Tri standard). Les caractéristiques de votre propriété sont évaluées selon plusieurs critères, tels que les retours des Vacanciers, les commodités offertes, et la situation géographique de votre propriété. La qualité de l'expérience est basée sur plusieurs critères tels que l'exactitude de votre calendrier, les délais de réponse du Propriétaire, le taux d'acceptation des réservations, la possibilité de réservation et de paiement en ligne, la cohérence des prix et l'expérience de qualité du séjour. Les frais de service versés par le Vacancier pour les réservations sont également un facteur de positionnement relatif aux propriétés ayant des offres similaires, en fonction des critères de pertinence décrits ci-dessus. Les résultats de recherche peuvent également apparaître sur l'application mobile dans un ordre différent que sur notre Site internet.

Explication locale (individuelle) - Facebook

Pourquoi vous voyez cette publication X

Découvrez comment gérer votre fil d'actualité X

Nous avons découvert que les gens voulaient en savoir plus sur le fonctionnement de leur fil d'actualité. Nous avons donc ajouté des fonctionnalités vous permettant de mieux comprendre et de mieux gérer ce que vous voyez.



Vous êtes amis avec **Antonio A. Casilli**

Cette publication de **Antonio A. Casilli** est populaire par rapport à d'autres publications que vous avez vues

Cette publication est considérée comme populaire, car de nombreuses personnes qui l'ont vue l'ont partagée

 Réactions	56
 Partages	20



Hubert Guillaud et **Benjamin Tincq** ont interagi avec cette publication

Simulateur - taxe d'habitation

The screenshot shows the 'Taxe d'Habitation' (Tax on Housing) simulator interface. On the left, a dark sidebar lists navigation options: 'Vos informations' (selected), 'Résultat', 'Abattements', 'Détail', and 'Réforme 2018'. The main content area has a light blue header 'COMPRENEZ VOTRE TAXE D'HABITATION !'. Below it, a note states: 'Cet outil a pour objectif de vous permettre de comprendre comment a été calculée votre taxe d'habitation 2017. Des erreurs peuvent subsister dans ce simulateur. La valeur sur votre avis d'imposition fait foi. Munissez-vous de votre avis d'imposition, qui contient des informations personnelles nécessaires au calcul.' The central form is divided into sections: 'MES INFORMATIONS' (with a note: 'Vous pourrez modifier les paramètres plus tard'), 'Ma situation' (allocations: ASPA, ASI, AAH; personal status: Veuf, Senior, Handicapé, Indigent), 'Mon habitation' (Département: AIN, Commune: ABERGEMENT CLEMENCIAZ, Valeur locative brute: 0), 'Résidence' (Principale, Secondaire, Dépendance résidence principale, Logement vacant), and 'Revenu fiscal de référence' (0, ISF checked). At the bottom, there are fields for 'Nombre de parts fiscales' (1) and 'Nombre de personnes à charge' (0).

COMPRENEZ VOTRE TAXE D'HABITATION !

Cet outil a pour objectif de vous permettre de comprendre comment a été calculée votre taxe d'habitation 2017. Des erreurs peuvent subsister dans ce simulateur. La valeur sur votre avis d'imposition fait foi. Munissez-vous de votre avis d'imposition, qui contient des informations personnelles nécessaires au calcul.

MES INFORMATIONS

Vous pourrez modifier les paramètres plus tard

Ma situation

Bénéficiaire de ces allocations

ASPA
 ASI
 AAH

Situation personnelle au 1er janvier 2017

Veuf
 Senior
 Handicapé
 Indigent

Revenu fiscal de référence

0
 ISF

RFR à blanc sur mon avis

Mon habitation

Département

AIN

Commune

ABERGEMENT CLEMENCIAZ

Valeur locative brute

0

Résidence

Principale
 Secondaire
 Dépendance résidence principale
 Logement vacant

Nombre de parts fiscales

1

Nombre de personnes à charge

0

Atelier AMI IA 2 #3 : Explicabilité des algorithmes



Séquence 2:

Fondements théoriques de l'explicabilité

A Multi-layered Approach for Interactive Black-box Explanations

Clément Henin & Daniel Le Métayer

July 9, 2020



IBEX

Context of explanation

Explanations in the black-box setting

Context of the explanation

Focus
Local
Global

Context of the explanation

Focus	Profil
Local	Technical expert
Global	Domain expert Auditor Lay user

Context of the explanation

Focus	Profil	Objective
Local	Technical expert	Improve
Global	Domain expert Auditor Lay user	Trust Challenge Action

Presentation of the black-box setting

F : the black-box function (spam classifier)

$$F : X \rightarrow Y \quad (1)$$

X : Input space (all possible emails)

Y : Output space (boolean spam or non-spam)

D : dataset representing the population (email dataset)

E : scope of the explanation

ex: $E = \{x_e\}$ (local explanation)

$E = D$ (global explanation)

Sampling: creation of emails used to inspect the model

Scope

$x_e = \text{"Hello, I am very happy to be at SRA. Clément Henin, PhD student at Inria"}$

Samples

$s_1 = \text{"Hello, I am very happy to be at SRA. Clément Henin, PhD student at Inria"}$

$s_2 = \text{"Hello, I am very happy to be at SRA."}$



SAMPLING

Sampling: creation of emails used to inspect the model

Scope

$x_e = \text{"Hello, I am very happy to be at SRA. Clément Henin, PhD student at Inria"}$

Population

$x_1 = \text{"Hello, If a machine is expected to be infallible, it cannot also be intelligent. Alan Turing"}$

$x_2 = \text{"Hello, Information is the resolution of uncertainty. Claude Shannon"}$

$x_3 = \text{"Hello, The theory has been developed for a hypothetical nervous system, or machine, called a perceptron. frank rosenblatt"}$

Samples

$s_1 = \text{"Hello, I am very happy to be at SRA. Clément Henin, PhD student at Inria"}$

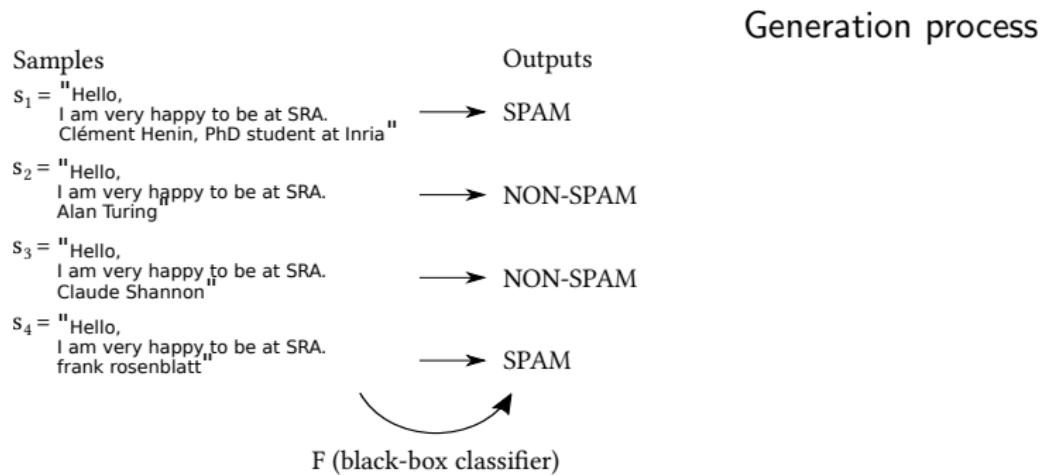
$s_2 = \text{"Hello, I am very happy to be at SRA. Alan Turing"}$

$s_3 = \text{"Hello, I am very happy to be at SRA. Claude Shannon"}$

$s_4 = \text{"Hello, I am very happy to be at SRA. frank rosenblatt"}$



Generation: analyse samples to create explanations



Generation: analyse samples to create explanations

Samples

$s_1 = \text{"Hello, I am very happy to be at SRA. Clément Henin, PhD student at Inria"}$

$s_2 = \text{"Hello, I am very happy to be at SRA. Alan Turing"}$

$s_3 = \text{"Hello, I am very happy to be at SRA. Claude Shannon"}$

$s_4 = \text{"Hello, I am very happy to be at SRA. frank rosenblatt"}$

Outputs

→ SPAM

→ NON-SPAM

→ NON-SPAM

→ SPAM

Generation process

- ▶ Rule based model (RBM)

F (black-box classifier)

Generation: analyse samples to create explanations

Samples

$s_1 = \text{"Hello, I am very happy to be at SRA. Clément Henin, PhD student at Inria"}$

$s_2 = \text{"Hello, I am very happy to be at SRA. Alan Turing"}$

$s_3 = \text{"Hello, I am very happy to be at SRA. Claude Shannon"}$

$s_4 = \text{"Hello, I am very happy to be at SRA. frank rosenblatt"}$

Outputs

→ SPAM

→ NON-SPAM

→ NON-SPAM

→ SPAM

F (black-box classifier)

Generation process

- ▶ Rule based model (RBM)
- ▶ Criteria 1: simplicity of the RBM
Number of rules
- ▶ Criteria 2: fidelity of the RBM
samples s.t. $\text{RBM}(s) = F(s)$

Generation: analyse samples to create explanations

Samples

$s_1 = \text{"Hello, I am very happy to be at SRA. Clément Henin, PhD student at Inria"}$

$s_2 = \text{"Hello, I am very happy to be at SRA. Alan Turing"}$

$s_3 = \text{"Hello, I am very happy to be at SRA. Claude Shannon"}$

$s_4 = \text{"Hello, I am very happy to be at SRA. frank rosenblatt"}$

Outputs

→ SPAM

→ NON-SPAM

→ NON-SPAM

→ SPAM

F (black-box classifier)

Generation process

- ▶ Rule based model (RBM)
- ▶ Criteria 1: simplicity of the RBM
Number of rules
- ▶ Criteria 2: fidelity of the RBM
samples s.t. $\text{RBM}(s) = F(s)$

RBM 1

If $\text{length}(\text{signature}) > 20$:

SPAM

Else:

NON-SPAM

rules = 1
fidelity = 75%

RBM 2

If $\text{length}(\text{signature}) > 20$ OR no capital letters in signature:

SPAM

Else:

NON-SPAM

rules = 2
fidelity = 100%

Thank you for your attention

Atelier AMI IA 2 #3 : Explicabilité des algorithmes



Séquence 3:

Ateliers d'identification des enjeux d'explicabilité dans les projets



Atelier AMI IA 2 #3 : Explicabilité des algorithmes

- Groupe 1 : Conseil d'Etat, DGGN : avec Kim
• Pour rejoindre la salle : <https://visio.incubateur.net/b/kim-ez9-zgq>
- Groupe 2 : IGN, INERI : avec Clément
• Pour rejoindre la salle : <https://visio.incubateur.net/b/kim-29x-rpn>
- Groupe 3: DGCL, DGCCRF, CRMNA : avec Soizic
• Pour rejoindre la salle : <https://visio.incubateur.net/b/kim-3ck-gw7>
- Groupe 4 : CHUB, DGS, IRSN : avec Simon
• Pour rejoindre la salle : <https://visio.incubateur.net/b/kim-nwd-uya>

Matrice des besoins en explication

Description du système

Responsable du système :

But du système :

Données en entrée :

Modalités d'interaction avec le système :

Niveau d'impact des décisions :

Degré d'automatisation :

Volume :

Les acteurs du système : (qui interagit avec le système ?)

Lister les acteurs dans les catégories suivantes :

- utilisateur/opérateur (professionnel ou quidam)
- personnes affectées par la décision
- responsable du système (doit-il rendre des comptes ?)
- Développeur
- auditeur externe

Niveau d'adhésion des acteurs (y-a-t-il des réticences à l'introduction du système?) :

Certains acteurs ont-ils intérêt à manipuler le système ?

Objectifs attendus des explications pour chaque acteur :

Lister les objectifs des acteurs dans les catégories suivantes :

- comprendre par curiosité / pour information
- acquérir des connaissances sur le domaine
- pour prendre leur décision / améliorer la confiance dans le système
- contester des décisions
- détecter un bug / améliorer le système
- manipuler le système (gaming, triche, optimisation)
- pouvoir justifier du bien-fondé des sorties à un autre acteur

Explications déjà en place :

Quelles explications sont déjà en place ? Comment sont-elles utilisées ?

Un exemple de Matrice des besoins en explication

Description du système

Responsable du système : Agence de la Biomédecine

But du système : Attribution d'un greffon cardiaque à un patient dans une liste de candidats en attente

Données en entrée : Données médicales du receveur et du donneur (âge, biologie, zones géographiques, ...)

Modalités d'interaction avec le système : Inscription par les médecins sur l'ordinateur de l'hôpital, notification sur téléphone du médecin de garde lorsque le greffon est proposé à un candidat

Niveau d'impact des décisions : Très important (allocation d'un traitement vital)

Degré d'automatisation : décision d'allocation automatique et irréversible, mais le médecin peut toujours refuser le greffon qui est proposé à un son patient

Volume : ~800 candidats ~500 greffes annuelles

Les acteurs du système : (qui interagit avec le système ?)

Lister les acteurs dans les catégories suivantes :

- utilisateur/opérateur (professionnel ou quidam)
Médecin (pro), infirmière (pro)
- personnes affectées par la décision
patient (quidam)
- responsable du système (doit-il rendre des comptes ?)
Agence de la Biomédecine
- développeur
Agence de la Biomédecine
- auditeur externe

Niveau d'adhésion des acteurs (y-a-t-il des réticences à l'introduction du système?) : Certains médecins étaient opposés car le système les empêche de prioriser les patients comme ils le souhaitent

Certains acteurs ont-ils intérêt à manipuler le système ? Les candidats à la greffe (pour être greffés plus vite), les médecins (pour greffer les patients qu'ils estiment prioritaires)

Objectifs attendus des explications pour chaque acteur :

Lister les objectifs des acteurs dans les catégories suivantes :

- comprendre par curiosité / pour information

Médecin (surtout internes et jeunes médecins)

- acquérir des connaissances sur le domaine

- pour prendre leur décision / améliorer la confiance dans le système

Les médecins transplantateurs veulent comprendre le système pour adapter leur prise en charge

- contester des décisions

Pas de contestation possible en l'état actuel

- détecter un bug / améliorer le système

Agence de la Biomédecine et médecins transplantateurs

- manipuler le système (gaming, triche, optimisation)

Certains médecins / les patients

- pouvoir justifier du bien-fondé des sorties à un autre acteur

Explications déjà en place :

Quelles explications sont déjà en place ? Comment sont-elles utilisées ? **Un cycle de formation (conférences, documents en ligne) a été proposé par l'agence de la biomédecine aux médecins / infirmières.**

Atelier AMI IA 2 #3 : Explicabilité des algorithmes



Séquence 4:

Conclusion et mise en commun