VERS UNE UTILISATION RESPONSABLE DES DONNÉES

Atelier datacraft
Labellisé Paris AI week (French Tech)
16 Novembre 2021

LE GROUPE DE TRAVAIL JUSQU'À AUJOURDHUI

AGENDA

01 Les composantes d’une IA de confiance

02 Focus sur la Fairness

03 Atelier

LES COMPOSANTES D'UNE IA DE CONFIANCE

CHARTE ETHIQUE

Trustworthy AI

[Read time: 45 min]

We aim to act responsibly, create and promote an AI that is lawful, ethical, inclusive user safety.

To that end, this document gathers a set of concrete guidelines, that are structured by the European Union (EU). The reason why we tackle AI trustworthiness by com components involved in the development of an AI system.

The primary audience of this document is AI practitioners, especially because it co level scale, before being tackled on a lower level - and technical - scale. As a resul

Content

- Human oversight
- Privacy and data governance
- Technical robustness and security
- Transparency and explicability
- Diversity, non discrimination and fairness
- Environmental and societal well-being
- Accountability
- Bibliography

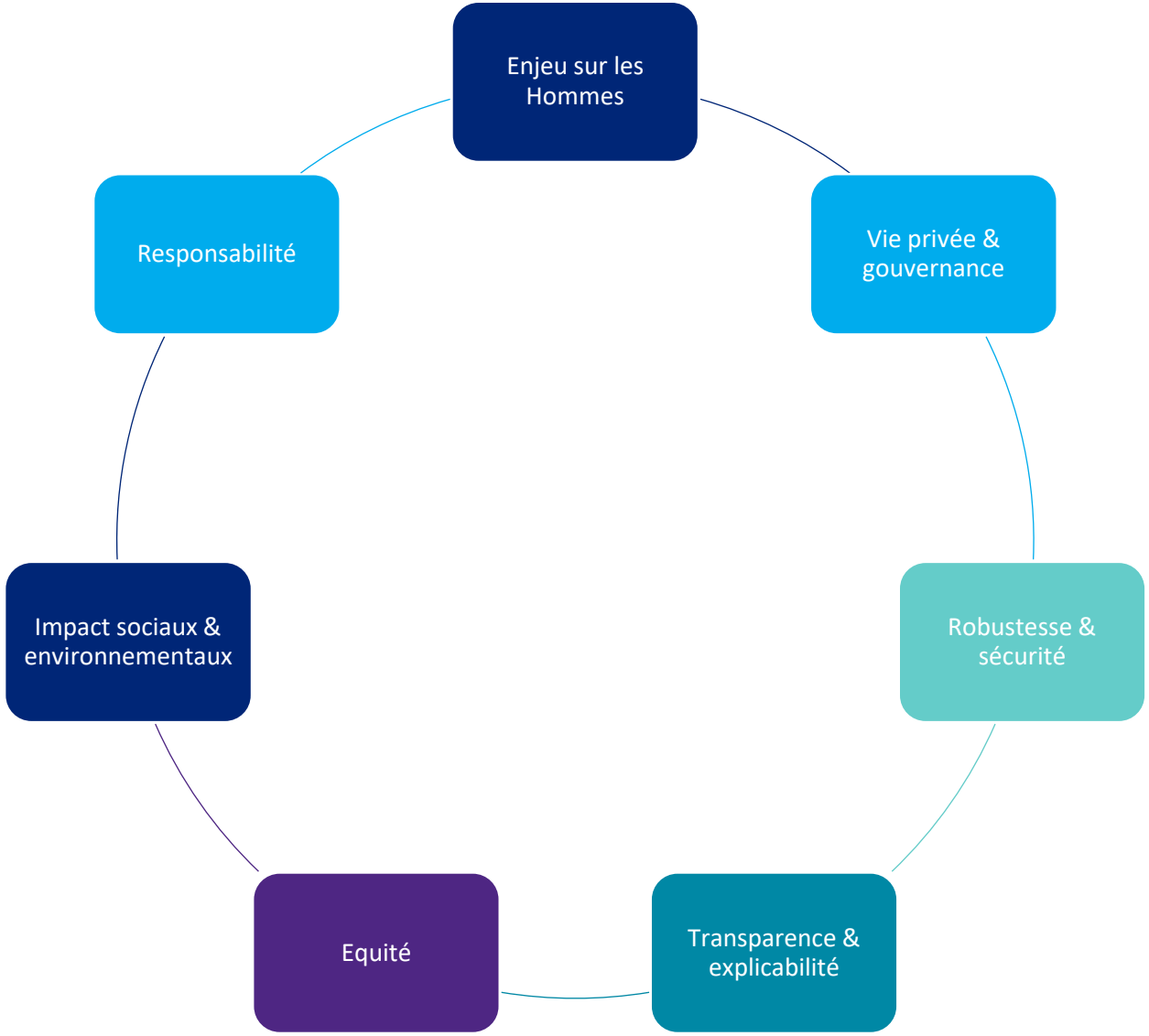
DANONE ONE PLANET. ONE HEALTH datacraft*

*« Créer une IA qui soit **légale**, **éthique**, **inclusive** et **fiable** »*

- Article divisé en plusieurs sections (temps de lecture : 45min)
- Structure inspirée des recommandations de l'UE
- A destination de **professionnels de la donnée**, et des **curieux** !

>> Lien <<

LES PILIERS D'UNE IA DE CONFIANCE



*L'IA de confiance repose sur **7 piliers***

CRÉER UNE IA ROBUSTE & SÉCURISÉE

Le cycle de vie des modèles d'IA

Data collection & pre-processing

- Considérer les biais d'acquisition (Qui ? Quand ? Comment ?)
- Documenter les traitements
- Prévenir l'ajout d'informations manifestement illégales, discriminatoires, …

Entrainement de l'IA

- Comprendre le fonctionnement des algorithmes
- Comprendre le biais associé aux métriques de performances
- Estimer l'incertitude associée à des prédictions

Mise en production de l'IA

- Assurer la reproductibilité et le traçage des prédictions
- Implémenter des boucles de rétroaction pour anticiper et corriger les dérives

La sécurité des modèles d'IA

Menaces & attaques

Données d'entraînement

Code du modèle

Exposition du modèle

Défense

White-hat surrogate

Authentication

Disparate impact analysis

Version control

CRÉER UNE IA DURABLE

Objectif : minimiser l'empreinte environnementale de l'IA

Une solution : mesurer, mesurer, mesurer, & partager
*Reporter les métriques liées à l'impact environnemental (**eq. carbone**) et les considérer autant que des métriques financières et/ou de performance des algorithmes*

1 **Estimer & reporter** les émissions de gaz à effet de serre produites par l'IA pour pouvoir dépenser moins, mieux, et « vert » (infrastructures et algorithmes)

2 **Privilégier une IA sobre :**
- Quelle est l'empreinte de mon modèle d'IA ?
- Un réseau de neurone est-il nécessaire ?
- Un réseau de neurones doit-il être réentraîné ? (*transfer learning*)
- Si oui, existe-il une version plus légère au prix d'une baisse raisonnable de performance : *distillation*, *alternatives*
- Les calculs peuvent-ils être opérés sur CPU, plutôt que GPU/TPU ?

CO2 impact, Green-Algorithms, CarbonAI

*« 1 BERT entraîné = 38 tours du monde en avion = 6500 ans de cerveau humain »**

* Source: https://aclanthology.org/P19-1355/

CRÉER UNE IA TRANSPARENTE ET EXPLICABLE

Explicabilité

Comprendre pourquoi un algorithme produit une prédiction

Motivation

- Remettre en question les prédictions
- Favoriser l'adoption des modèles d'IA
- Comprendre les composantes importantes de la décision

Méthodes statistiques
Souvent interprétables nativement (régressions, …)

Modèles de machine learning
Interprétations globale (au niveau du modèle) & locale (à l'individu)

Facilité d'interprétabilité

Transparence

Remettre en question la création de l'IA

1 **Documenter** et « rendre publique » la théorie sous-jacente du modèle d'IA (hypothèses de modélisations, algorithmes, …)

2 **Documenter** et « rendre publique » l'implémentation de l'IA

3 **Documenter** et « rendre publique » les données sur lesquelles l'IA a été entraînées

CRÉER UNE IA INCLUSIVE

Motivation: créer une IA « sans biais »

Ideation & design

- Être sensibilisé aux biais régulièrement induits par l'IA (genre, ethnie, publicité, traduction, …)
- Elaborer/designer un modèle d'IA de façon transparente et inclusive : identifier les **biais sociaux potentiels** a priori de son projet d'IA (population(s) vulnérable(s), …) et les **métriques pertinentes** pour les quantifier

Data collection & preprocessing

- Identifier les **biais techniques potentiels** a priori de son projet d'IA (historique de données, complétude des données, …)
- Garantir une représentativité de la population complète

Entraînement de l'IA

- Arbitrer entre performance statistique et « **performance éthique** »
- Si nécessaire, utiliser des traitements pré-entraînement sur les données, pendant l'entraînement sur la fonction de coût, ou post-entraînement
- Ces traitements peuvent être de nature statistique… Ou complètement manuelle

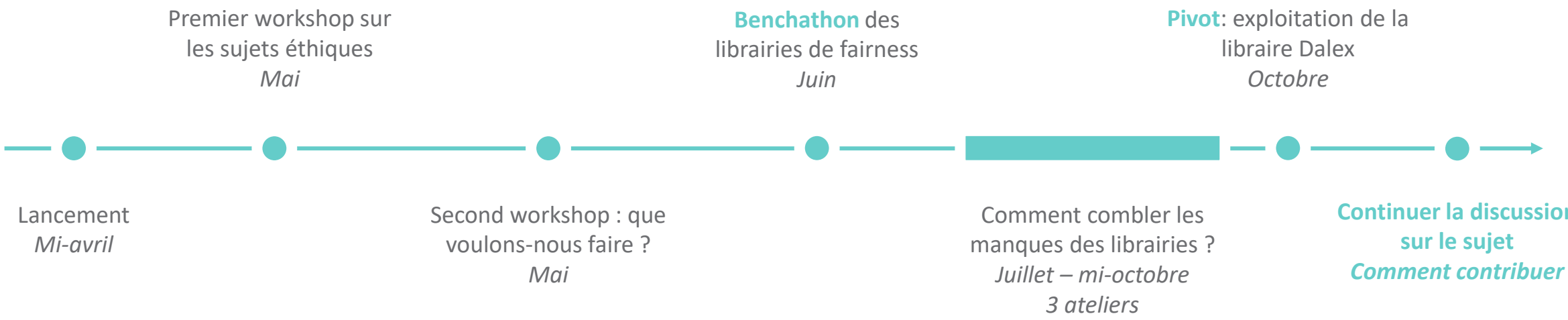
Mise en production

- Monitorer le modèle d'IA en production pour garantir le maintien des performances statistique et éthique
- Définir une stratégie de réentrainement claire, garantissant performances statistique et éthique

Focus de l'atelier d'aujourd'hui !

FOCUS SUR UNE IA INCLUSIVE

L'INITIATIVE IA DE CONFIANCE CHEZ DATACRAFT

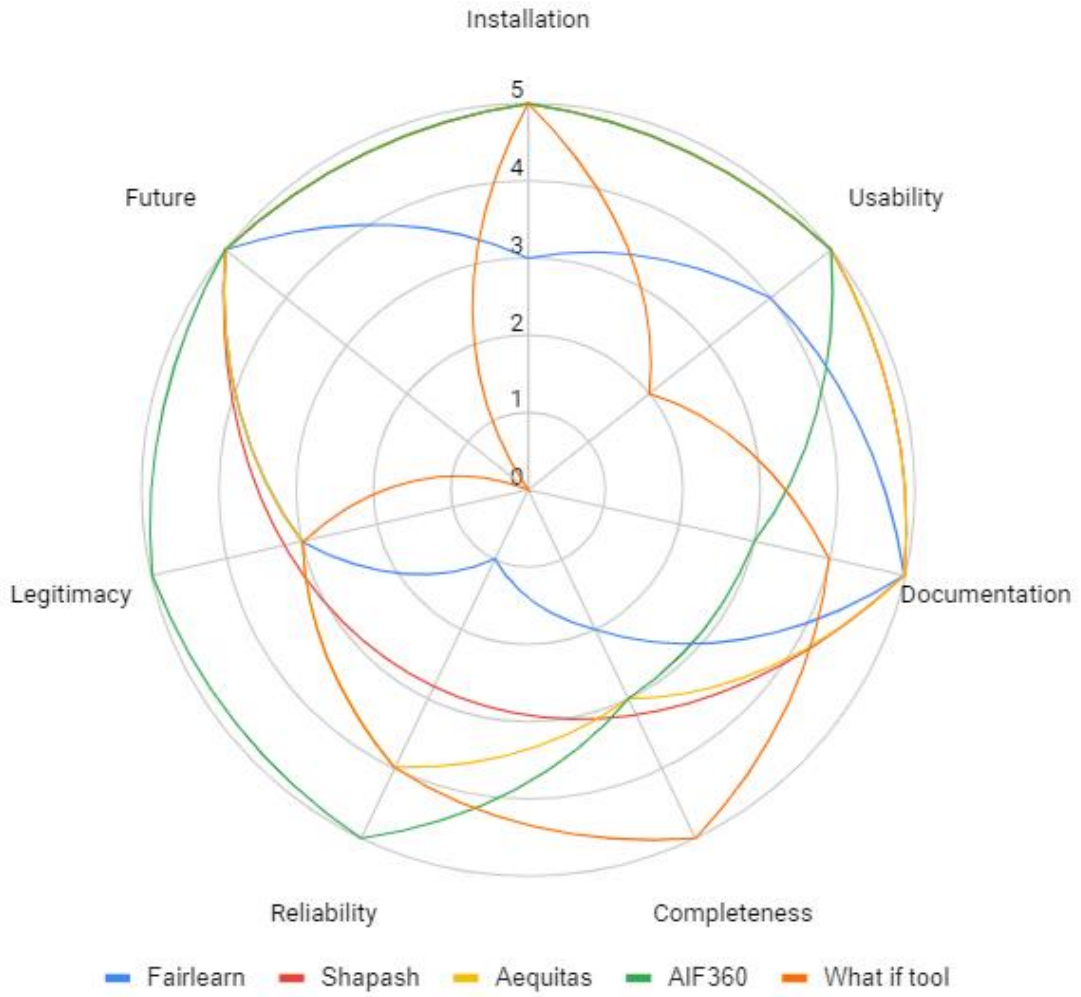


7 ateliers pour comprendre comment **garantir simplement** l'éthique des algorithmes

RÉSULTATS DU BENCHATHON*

- AIF360 : calculer des métriques et mitiger des biais
- Fairlearn : calculer des métriques et mitiger des biais
- Aequitas : audit et détection des biais
- What-if tool: analyse contrefactuelle
- Shapash : interprétabilité et transparence des modèles

Résultats exhaustifs ici



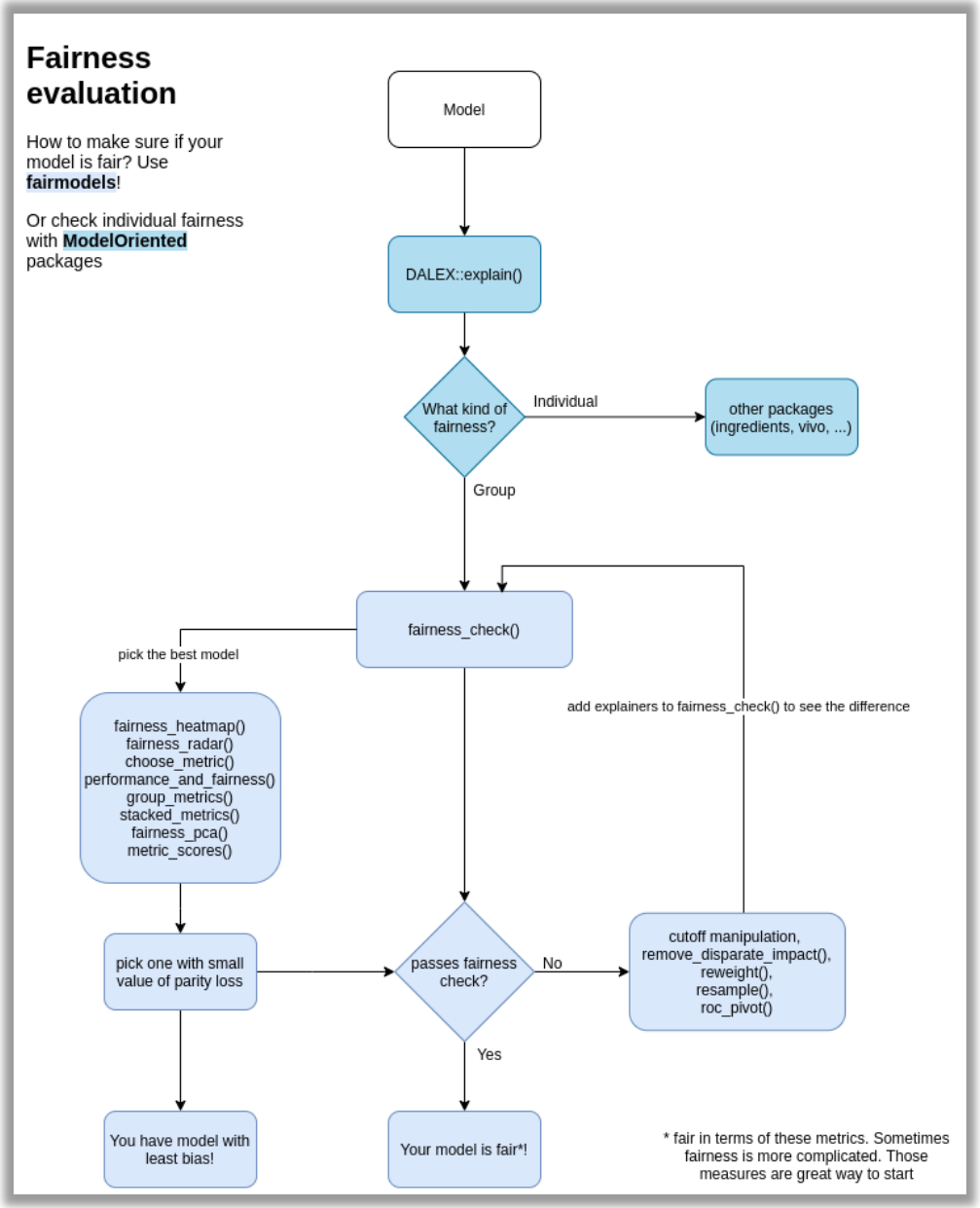
* Benchathon : atelier datacraft ayant pour but de benchmarker des outils de data science

DALEX

Un candidat naturel pour une prise en compte **simple et intuitive** de l'éthique dans les modèles d'IA

Avantages
- API simple d'utilisation (!)
- Plots intuitifs et instructifs
- Gestion de plusieurs modèles
- Nombreuses métriques
- Gestion de la régression

Inconvénients
- Peu de technique de mitigation (--> mais possible d'interfacer)
- Les tutoriels ne permettent pas de prendre du recul sur la démarche globale à adopter et se focalisent sur les outils (--> objectif de l'atelier d'aujourd'hui)



Quelle démarche adopter pour auditer la performance éthique d'une IA ? []

ATELIER

RÉFÉRENCES

Les références utilisées pour ces travaux sont disponibles ici