AI Fairness Guide

AI Fairness Guide est également disponible en version personnalisée selon vos besoins. Êtes-vous intéressé-e à un accompagnement professionnel pour résoudre vos défis éthiques? Contactez-nous sous info@ethix.ch.

Réalisable grâce au MIGROS Fonds pionnier

ethix

L'ÉTHIQUE AU COEUR DES ALGORITHMES

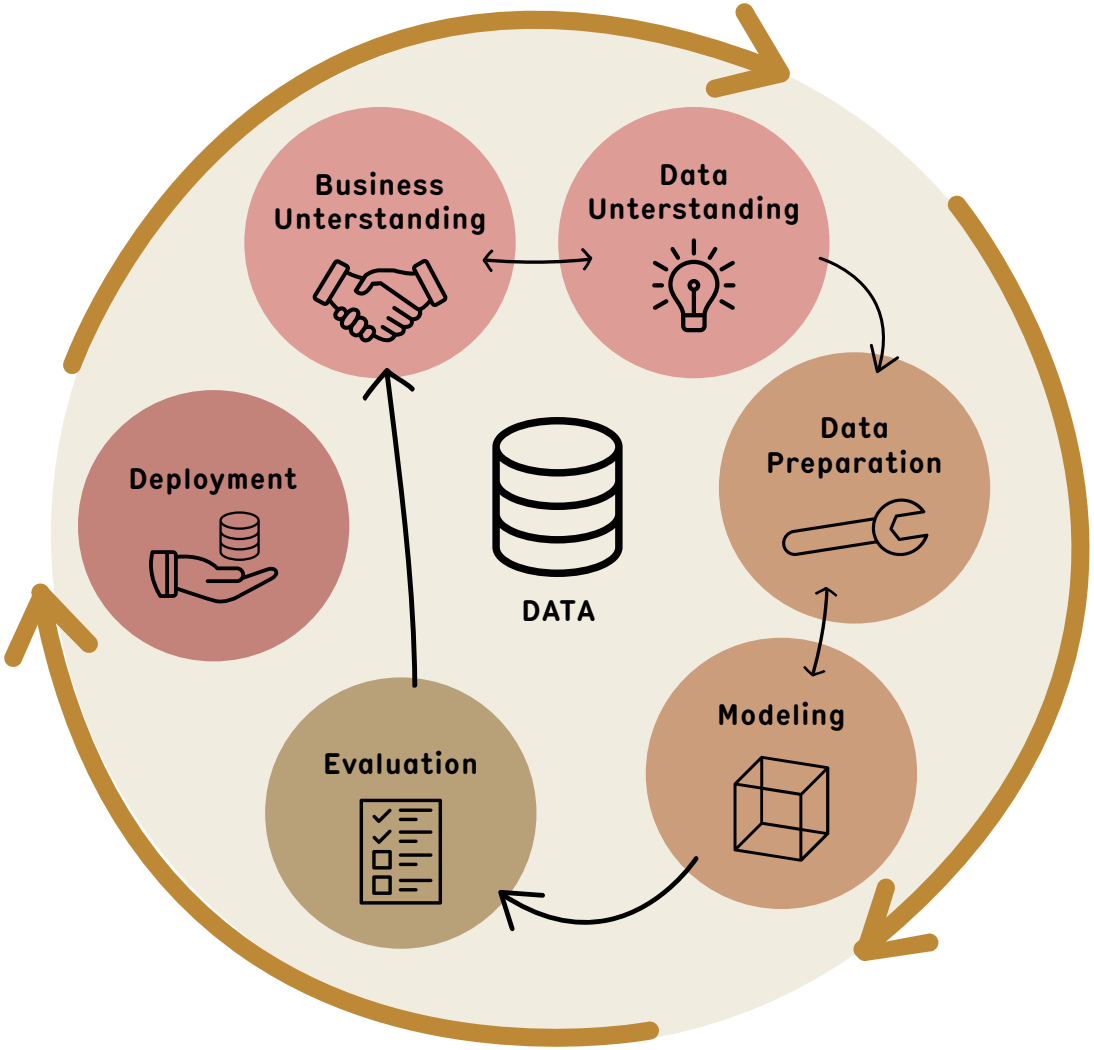
L'analyse automatisée de données ouvre de nouvelles opportunités économiques. Au-delà d'un effet de mode lié à l'utilisation à tout-va du terme d'*intelligence artificielle*, l'analyse de données peut être utile à tous les secteurs d'activités. Elle va permettre d'en apprendre plus sur ses clients, ses marchés et ses propres performances, mais elle va également rendre possible de nouveaux produits et services. Les entreprises à but commercial, mais aussi les institutions publiques et les organisations de la société civile peuvent faire usage de ces analyses de données.

Ce document s'adresse à tous ceux et celles qui programment, supervisent ou initient des projets d'analyse de données. Les artisans « data scientists », celles et ceux qui programment et configurent ces outils, s'engagent dans un voyage au long cours qui démarre avec la collecte des données pour se terminer lors de la mise en œuvre finale du système « intelligent », après avoir parcouru les étapes de nettoyage des données, de choix d'algorithme, de paramétrisation des systèmes, etc.

Dans ce document, nous proposons une checklist à destination des artisans, des « data scientists » qui produisent ce genre de système ainsi que des personnes qui les encadrent, afin de les aider à identifier les enjeux éthiques lors des différentes étapes de leur travail. Cette checklist peut s'utiliser de manière indépendante comme un fil conducteur pour évaluer l'ensemble du processus. Elle est le fruit d'échanges avec nos collègues de Swiss Statistical Design & Innovation (Swiss SDI) et de Datastory, deux sociétés suisses actives dans la data science. Elle s'appuie également sur le projet "Algo.Rules" de la fondation Bertelsmann (Allemagne).

AI Fairness Guide

ethix



Ce schéma résume de manière simplifiée les différentes étapes d'un processus d'analyse de données. Le présent document s'utilise comme une carte de navigation des différents défis éthiques à relever à chacune de ces étapes.
Il suit une approche chronologique du projet. Les différentes parties sont néanmoins indépendantes et peuvent être utilisées de manière autonome.
Ce document s'adresse en particulier aux équipes data.

Pour aller plus loin :

- Le projet "Algo.Rules" de la Fondation Bertelsmann: https://algorules.org/de/startseite

- L'article de Raji et al., "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing" (2020)

- Le papier scientifique "Getting into the engine room: a blueprint to investigate the shadowy steps of AI ethics", publié par J. Rochel et F. Evéquoz dans "AI & Society" (2020) https://link.springer.com/article/10.1007/s00146-020-01069-w

AI Fairness Guide

ethix

Etape 1
BUSINESS UNDERSTANDING

ACTIVITÉ CLARIFIER LES OBJECTIFS DU PROJET

Description

- Comprendre les besoins et les objectifs business du client :
  que vise le client avec ce projet ? Quelles sont les raisons qui motivent ces objectifs ?

- Définir quelles sont les parties prenantes impliquées dans le projet :
  qui poursuit quels objectifs dans ce projet ? Qui d'autre est concerné par le projet ?

Les questions à se poser

- Est-ce que les objectifs du projet portent potentiellement préjudice à des tiers ou à la société ? Par exemple :
  - Est-ce que les objectifs du projet impliquent des licenciements ?
  - Quel impact le projet a-t-il sur les utilisateur.trice.s/la société ?
  - Y'a-t-il un impact particulier sur des personnes vulnérables, par exemple des enfants ?

- Est-ce que des parties prenantes importantes du projet peuvent se sentir menacées par l'arrivée du projet / par l'implication de spécialistes données ? Par exemple: un-e employé-e dont la collaboration est essentielle dans le projet, mais qui pourrait se sentir menacé car le projet met en péril son emploi ?

- La responsabilité des différentes parties prenantes du projet est-elle claire ? Il y a-t-il un acteur.trice bénéficiant d'une vue d'ensemble sur le projet ? La responsabilité est-elle distribuée par étapes du projet ou en fonctiondes différents domaines ? Quels sont les rôles, droits et devoirs impliqués par cette responsabilité ?

Bonnes pratiques pour l'équipe data

- Connaître les objectifs business du client pour mieux estimer l'impact du projet sur des tiers / la société

- Sur la base d'expériences similaires, discuter avec le client pour lui faire comprendre les impacts potentiels du projet

- S'assurer de la bonne compréhension des parties prenantes du côté des clients

- Rédiger une charte éthique de projet qui pourra accompagner sa réalisation et qui pourra notamment inclure une discussion sur les valeurs qui sous-tendent la réalisation du projet, la culture et les valeurs de l'entreprise, les engagements pris ainsi que sur les lignes rouges que le projet ne devra pas franchir.

AI Fairness Guide

ethix

Etape 2
DATA UNDERSTANDING

ACTIVITÉ DÉTERMINER LES BESOINS DONNÉES

Description

- Comprendre les besoins en analyse de données pour le client
- Définir précisément les objectifs finaux de la récolte et de l'analyse des données

Les questions à se poser

- Est-ce qu'un recoupement des sources de données fait apparaître des données sensibles ? (p. ex. permettre d'identifier une personne et ses activités)
- Est-ce que les objectifs finaux de la récolte et de l'analyse de données sont clairs ?
- Est-ce que ces objectifs d'analyse de données sont alignés avec les objectifs business détaillés ci-dessus ? En d'autres termes, les données et leur exploitation prévue permettent-elles de répondre aux questions business pertinentes et uniquement à celles-ci ?
- Est-ce que ces objectifs et leurs conséquences sont compris de la même manière entre l'équipe data et le client (compréhension commune) ? Notamment, est-ce que le client détecte le potentiel impact du data mining pour sa société ?

Bonnes pratiques pour l'équipe data

- Connaître les contraintes légales ayant trait aux combinaisons des sources de données (par ex. donnés sensibles au sens du RGPD, risque d'apparition lors de la combinaison de données dérivées mettant en péril l'anonymat)
- Etablir des objectifs clairs et précis avec le client
- Prévoir de contrôler de façon itérative que l'évolution du projet conserve l'orientation de base définie

AI Fairness Guide

Etape 2
DATA UNDERSTANDING

ethix

ACTIVITÉ RÉCOLTE DES DONNÉES

Description
- Sélection et récolte des données

Les questions à se poser

- Qui a la charge de la récolte des données (client, entreprise tierce) ? Y'a-t-il un éventuel conflit d'intérêt ?
- Est-ce que les obligations légales/ bonnes pratiques sont respectées (consentement, LPD, GDPR) ?
- Les données sont-elles conservées de manière adéquate (sécurité, durée de stockage) ?

Bonnes pratiques pour l'équipe data

- Connaître le processus de récolte des données : expliquer au client les potentiels impacts lors d'une négligence dans ce domaine.
- Sensibiliser le client sur les risques lors de l'utilisation de données, notamment lorsque des données privées sont impliquées.
    - Par exemple identifier les données particulièrement délicates en terme de protection de la personnalité et prévoir de définir des sécurités pour interdire l'accès aux données brutes aux personnes non autorisées, mettre en place préventivement une politique de suppression de données non nécessaires, d'anonymisation ou de pseudonymisation, voire définir une date de « péremption » des données au-delà de laquelle elles seront supprimées.

AI Fairness Guide

Etape 2
DATA UNDERSTANDING

ethix

ACTIVITÉ DESCRIPTION DES DONNÉES

Description

- Description des sources de données (origine, actualité, fréquence de mise à jour, ...)
- Définition des données (quel type de donnée, limites associées à la donnée – ce qu'elle décrit, ce qu'elle ne décrit pas -, signification des catégories (classes), explication a priori de l'éventuelle absence de données, ...)

Les questions à se poser

- Est-ce que le client est conscient de son impact sur la qualité du projet lorsqu'une description des données est erronée (ex. mauvaise interprétation des résultats, conclusions abusives, manque de contexte) ?
- Quels sont les présupposés utilisés dans la catégorisation des données, en particulier pour une catégorisation qui serait discutable (par ex. genre = [hommes, femmes], ou également d'autres catégories) ?

Bonnes pratiques pour l'équipe data

- Sensibiliser le client quant à l'impact potentiel de cette classification
- Décrire les données de manière exhaustive, en mettant un soin particulier à expliciter les catégories et leurs présupposés

AI Fairness Guide

Etape 2
DATA UNDERSTANDING

ethix

ACTIVITÉ VÉRIFICATION DES DONNÉES

Description

- Contrôle qualité des données reçues
- Contrôle de l'information : est-ce que les données permettent de répondre aux objectifs identifiés en début de projet ?

Les questions à se poser

- Comment déterminer le moment opportun pour stopper les vérifications et « accepter » les données ? Après quelle quantité / qualité de tests ?
- Lors d'un contrôle qualité, comment choisir à quelles données accorder notre confiance et quelles données éliminer ?
- Quelles sont les mesures de sécurités mises en place pour le « data privacy » ?

Bonnes pratiques pour l'équipe data

- Contrôle qualité systématique des données, même si le fournisseur garantit cette qualité. (par ex. effectuer des statistiques descriptives pour chaque variable identifiée dans les données afin d'identifier des incohérences)
- Définir un « label » pour le contrôle de qualité des données, par ex. en définissant les sources fiables ou un benchmark de référence
- Décrire le benchmark normatif utilisé qui servira à juger de la qualité des données et identifier les biais potentiels dans le jeu de données

AI Fairness Guide

ethix

Etape 3
DATA PREPARATION

ACTIVITÉ DATA SELECTION – CLEANING – CONSTRUCTION – INTEGRATION

Description

- Processus principal pour la préparation des données
  - Sélection des données importantes
  - Nettoyage des sources de données
  - Gestion des données manquantes
  - Intégration des diverses sources de données
  - Etiquetage (labelisation) – dans une perspective de préparation des modèles d'entraînement et de test dans le cas d'un apprentissage supervisé

Les questions à se poser

- Est-ce que les données sont complètes et équilibrées ?
  - Est-ce que la génération de nouvelles données oriente le résultat final ?
- Est-ce que des biais sont présents dans les données ?
- Est-ce que la labélisation des données est faite de manière systématique ? La qualité de la labélisation est-elle suffisante ?

Bonnes pratiques pour l'équipe data

- Description complète des catégories (classes) et des concepts présents dans les données (compléter les descriptions faites à l'activité « data description »)
- Effectuer une génération de données pour compléter des données manquantes uniquement lorsque cela est indispensable. Documenter la méthode de génération de données. Garder à l'esprit que le résultat final peut être orienté de façon consciente ou inconsciente. Analyser les biais de modélisation potentiels lié à la génération de données, en les documentant de manière explicite et transparente
- Contrôler les effets de la correction des données (par ex. comparer modèles et résultats avec différentes corrections)
- Garder une transparence avec le client et le partenaire en documentant scrupuleusement les mesures de nettoyage des données
- Expliciter les dimensions normatives des techniques de préparation des données, en lien avec les objectifs du projet
- Documenter avec soin le cas des données manquantes ou inutilisables et les techniques utilisées pour les gérer ou les corriger
- Choisir avec soin les différentes méthodes de correction de la qualité des données afin de minimiser les biais, et décrire avec soin leurs implications potentielles (sur le plan normatif, et sur la suite du processus)

AI Fairness Guide

ethix

Etape 4
MODELING

ACTIVITÉ MODELING TECHNIQUE – TEST DESIGN – PARAMETERS – ASSESSMENT

Cette étape est fortement liée à la suivante (Évaluation) et les deux sont largement interdépendantes et peuvent être traitées de manière parallèle.

Description

- Création des modèles de Machine Learning
- Test et validation des modèles
- Documenter les buts et les effets attendus

Les questions à se poser

- Quels sont les critères pertinents pour évaluer la performance du modèle ?
  - Sur la base de l'état de l'art dans le domaine et pour le type de données considéré, métriques de performance à fixer
- Ces critères se rapportent-ils aux objectifs du projet ? Comment ? Il y a-t-il des contradictions entre les différents objectifs ? Si tel est le cas, quels critères privilégier ? Sur cette base, est-ce que le choix du modèle est optimal ? Quelles sont les alternatives ?
- Comment exprimer les divergences et définir les responsabilités lors de la validation du modèle avec le client ?

Bonnes pratiques pour l'équipe data

- Justifier le choix du modèle sur une base mathématique et sur la base de l'expérience (état de l'art)
- Expliciter le(s) objectif(s) recherchés en entrainant le modèle (métrique de performance)
- Expliciter les critères d'évaluation de la performance du modèle en fonction des objectifs du projet
- Documenter les arbitrages (tradeoffs) entre différents objectifs légitimes, qui pourraient entrer en concurrence lors de la sélection du modèle ou l'évaluation de sa performance
- Lister les implications, des impacts et des inconvénients liés au choix du modèle, ainsi qu'aux méthodes d'évaluation de la qualité du modèle
- Décrire les limites et vulnérabilités du modèle : Pourquoi est-il fait ? Qu'est-ce qu'il ne détecte pas ?
- Eventuellement, considérer des techniques d' « adversarial machine learning » pour identifier des vulnérabilités potentielles du modèle
- Effectuer une évaluation de l'importance du risque qui conjugue la probabilité de défaillance du système avec la gravité de cette défaillance.

AI Fairness Guide

Etape 5
EVALUATION

ethix

ACTIVITÉ EVALUATE RESULTS

Description

- Validation des résultats
- Définition de la suite du projet

Les questions à se poser

- Est-ce que les modèles sont stables ?
- Est-ce que les modèles sont évalués sur des données représentatives du monde réel ou seulement sur un échantillon peu représentatif ?
- Est-ce que les résultats de l'évaluation répondent au besoin du client ?

Bonnes pratiques

- Effectuer une validation croisée (cross-validation) pour évaluer la qualité du modèle
- Vérifier systématiquement les données d'évaluation (benchmark) (cf. étape 'Data Verification')
- Décrire précisément les données d'évaluation utilisées pour évaluer les résultats du modèle (cf. étapes 'Data Description' et suivantes)
- S'assurer que le modèle n'est pas biaisé pour un sous-ensemble particulier de données.
- Sur la base des résultats, préciser la description du champ d'application et les limitations du système. Organiser un espace de discussion avec le client quant au champ d'application du système, aux résultats qu'il permet (ou non) d'obtenir et le communiquer clairement aux parties prenantes (direction de projet).
- Prévoir une évaluation en continu avec le client (développement agile)

AI Fairness Guide

Etape 5
DEPLOYMENT

ethix

ACTIVITÉ DÉPLOIEMENT

Description

- Passage de l'évaluation (phase pilote) à l'exploitation « en production »
- Utilisation des modèles sur des données réelles (et non plus un échantillon comme dans les phases précédentes)
- Intégration de l'outil d'analyse données dans les processus de l'entreprise/organisation

Les questions à se poser

- Est-ce que des feedbacks sont donnés par des tiers pour faire évoluer le modèle (Active Learning) ? Dans ce cas, quelles précautions sont prises pour éviter des dérives du modèle suite à l'Active Learning, provoquées intentionnellement ou non ?
- L'utilisation finale de l'outil est-elle conforme aux objectifs initiaux ?
- L'outil a-t-il été bien intégré dans les processus de l'entreprise/organisation ? Est-ce que les responsabilités quant à cette utilisation finale sont définies ?

Bonnes pratiques

- Contrôler en continu les résultats obtenus, par la mise en place de métriques pour s'assurer de la stabilité du modèle
- Définir les responsabilités lors du déploiement : par ex. un expert métier identifie les résultats problématiques et un data scientist entraine et vérifie le modèle en repassant par le processus CRISP-DM décrit ici
- Informer les parties prenantes (y compris utilisateur.trice.s finaux si pertinent) des implications lors de l'utilisation du système (limitations identifiées, etc.)
- Prévoir des outils de signalement de résultats jugés problématiques par les utilisateur.trice.s finaux, et une procédure pour les analyser
- Mettre en place des outils de communication pour informer les utilisateur.trice.s du fonctionnement du modèle et s'assurer que les résultats obtenus sont compréhensibles

V1- 5/2021
ethix - Laboratoire d'éthique de l'innovation - www.ethix.ch