REINVENT ENGINEERING

CENTRALE NANTES

5 questions éthiques de l'IA

Jean-Pierre ELLOY, Morgan MAGNIN

jean-pierre.elloy@ec-nantes.fr, morgan.magnin@ec-nantes.fr

REINVENT ENGINEERING

CENTRALE NANTES

5 questions éthiques de l'IA

principes éthiques généraux

éthique des données

éthique du développement

éthique d'usage

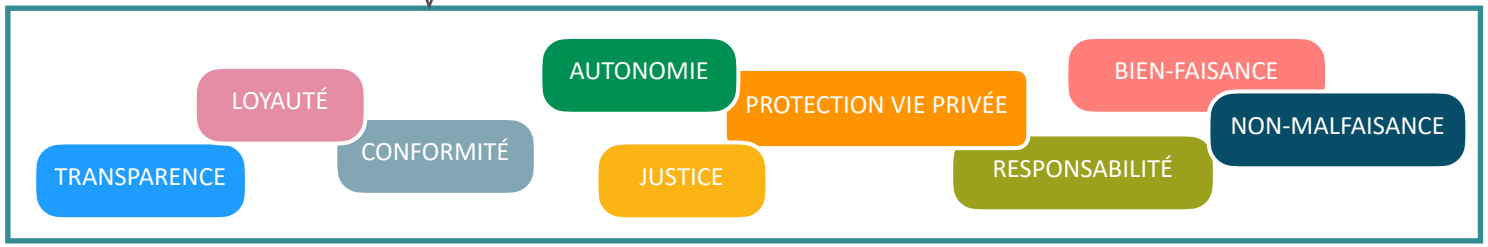
Jean-Pierre ELLOY, Morgan MAGNIN

jean-pierre.elloy@ec-nantes.fr, morgan.magnin@ec-nantes.fr

IA : déclinaison des **principes éthiques**

ensemble des principes éthiques applicables à l'IA …

… mais **comment** les appliquer ?



éthique

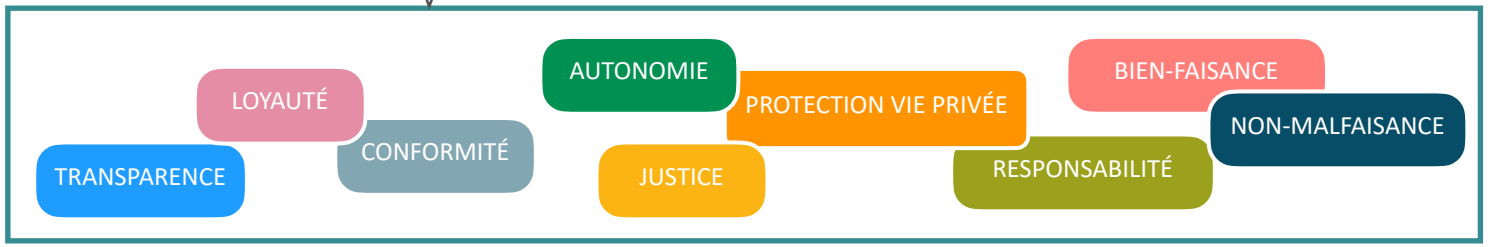
intelligence artificielle

… d'abord en commençant par identifier **à quels « moments »** le respect de ces principes est sensible

IA : déclinaison des **principes éthiques**

ensemble des principes éthiques applicables à l'IA …

… mais **comment** les appliquer ?



éthique

intelligence artificielle

… d'abord en commençant par identifier **à quels « moments »** le respect de ces principes est sensible

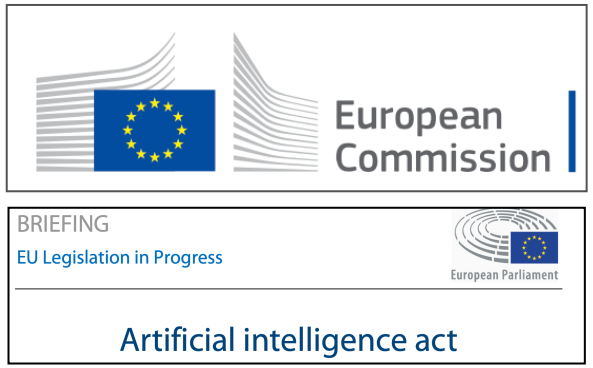
principes généraux : **droits** et **valeurs**

principes sur les **données** utilisées

principes dans le **développement**

principes d'**usage** des outils à base d'IA

5.1 IA : principes **éthiques généraux**



Principes à respecter par les SIA *"dignes de confiance"* :

- **bienfaisance** *(faire le bien)*
- **non-malfaisance** *(ne pas faire le mal)*
- **justice** *(être équitable)*
- **explicabilité** *(transparence du fonctionnement)*
- [autonomie *(préserver la liberté d'agir de l'homme)*]

Ces valeurs (sauf "explicabilité") sont celles de nos sociétés, et ne sont pas spécifiques à l'IA

***Droits** fondamentaux à respecter sans réserve :*

- *respect de la **dignité** humaine*
- ***liberté** des individus*
- *respect de **l'Etat** de droit*
- ***égalité**, non discrimination*
- *droit des **citoyens***

- Ces valeurs générales applicables à tout

Catalogue de risques à contrôler :

- ***risques inacceptables** (score social, identification biométrique, manipulation cognitive)*
- ***risques élevés** (sécurité des produits)*
- ***IA générative** (contenu illégal, données protégées)*
- ***risques mineurs** (transparence du fonctionnement, liberté d'acceptation/refus)*

5.2 IA : éthique **des données**

La **protection des données**
(personnelles, industrielles, publiques)

RGPD *réponse … française*

European Commission *réponse … européenne*

DATA ACT & DATA GOUVERNANCE ACT

Différentes perceptions de cette protection

- aux USA, la donnée est un produit commercial
- en Asie, la donnée est un élément de pouvoir
- en Europe, la donnée est sacralisée

5.2 IA : éthique des données

- **éthique extrinsèque**
  - Les **données sont « propres »**, mais leur collecte ou leur utilisation **peut porter préjudice**

exemple pour les données personnelles

Ethique **extrinsèque** des **données personnelles** > respect de **la vie privée**

mais … satisfaire ce principe provoque une tension entre :

- le besoin de données personnelles pour alimenter l'IA
- et la protection de la vie privée

pour résoudre cette tension > un PAI

Un **Processus d'Analyse d'Impact** doit identifier tous les risques induits par la divulgation des données privées, et comment on envisage de les réduire

Pour cela, un **Processus d'Analyse d'Impact** consiste à répondre à :

- quelles sont les informations personnelles collectées ?
- pourquoi le sont-elles ?
- comment sont-elles utilisées ?
- avec qui sont-elles partagées ?
- quels dispositifs ont été mis en place pour en protéger la divulgation ?
- combien de temps seront-elles conservées ?
- comment seront-elles supprimées ?

La **protection des données**
(personnelles, industrielles, publiques)

RGPD *réponse … française*

European Commission *réponse … européenne*

DATA ACT & DATA GOUVERNANCE ACT

Différentes perceptions de cette protection

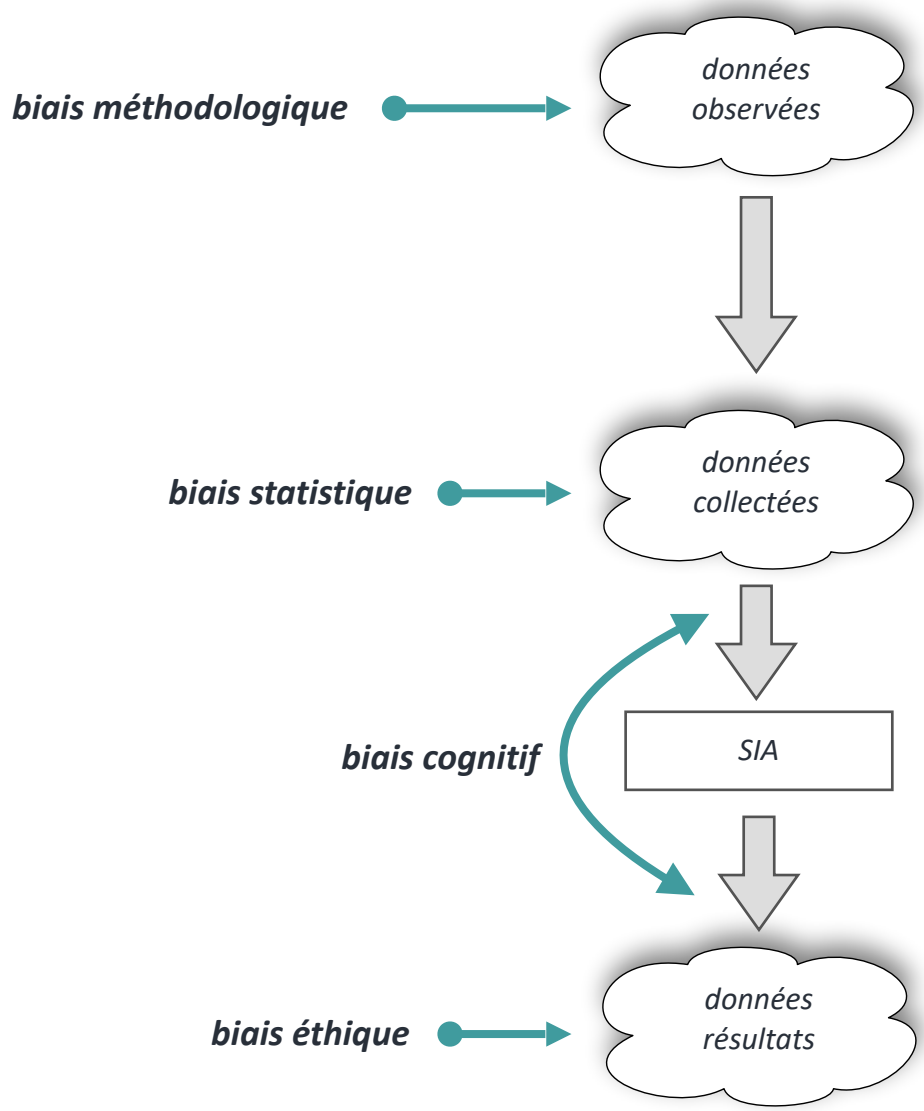
- aux USA, la donnée est un produit commercial
- en Asie, la donnée est un élément de pouvoir
- en Europe, la donnée est sacralisée

IA : éthique **des données**

- **éthique intrinsèque**

Les **données** peuvent être « **impropres** »

Pour être « **propres** », les données manipulées par un SIA doivent répondre aux critères de : **qualité** (absence de **biais**), **diversité** et **non-discrimination**, et **traçabilité**



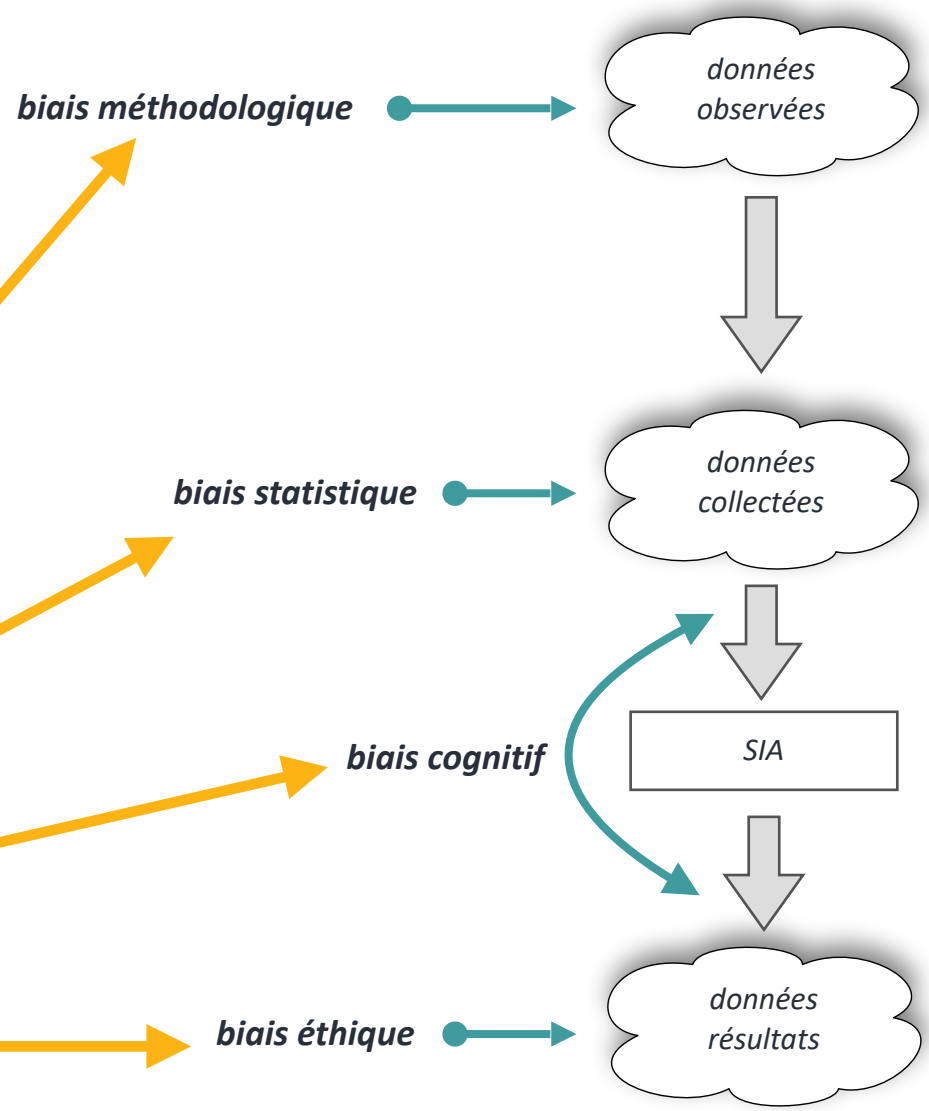
IA : éthique **des données**

- **éthique intrinsèque**

Les **données** peuvent être « **impropres** »

Pour être « **propres** », les données manipulées par un SIA doivent répondre aux critères de : **qualité** (absence de **biais**), **diversité** et **non-discrimination**, et **traçabilité**

- biais **méthodologique**
  … méthode de mesure et de sélection, **écart** entre *distribution* obtenue et distribution attendue
- biais **statistique**
  … **écart** par rapport à sa *valeur* exacte, avec son espérance mathématique
- biais **cognitif**
  … **écart** de *validation* entre celle de *l'intuition* (donc rapide) et celle de la *raison* (long)
- biais **éthique**
  … **écart** entre le résultat produit et un résultat socialement *juste*



5.3 IA : éthique du développement

Ces principes éthiques sont ceux du **numérique**, comme :

- l'**explicabilité**
- l'**interprétabilité** (*Interpretability*)
- la capacité (d'un système d'IA) à **rendre des comptes** (*Accountability*)
- la **transparence** (*Transparency*)
- la **fiabilité** (*Trustworthiness*)
- l'**équité** (*Fairness*)
- la **conformité** (respect des règles et de la législation)
- la **loyauté** (fidélité de la réalisation par rapport aux spécifications)
- l'identification des **responsabilités**

… qui créent, ensemble, la **confiance** en l'outil développé

5.3 IA : éthique du développement

Mais **des difficultés natives, propres à l'IA,** peuvent empêcher, ou gêner, la vérification de certaines propriétés :

Ces principes éthiques sont ceux du **numérique**, comme :

- l'**explicabilité**
- l'**interprétabilité** (*Interpretability*)
- la capacité (d'un système d'IA) à **rendre des comptes** (*Accountability*)
- la **transparence** (*Transparency*)
- la **fiabilité** (*Trustworthiness*)
- l'**équité** (*Fairness*)
- la **conformité** (respect des règles et de la législation)
- la **loyauté** (fidélité de la réalisation par rapport aux spécifications)
- l'identification des **responsabilités**

… qui créent, ensemble, la **confiance** en l'outil développé

5.3 IA : éthique **du développement**

Mais **des difficultés natives, propres à l'IA,** peuvent empêcher, ou gêner, la vérification de certaines propriétés :

Ces principes éthiques sont ceux du **numérique**, comme :

- l'**explicabilité**
- l'**interprétabilité** (*Interpretability*)
- la capacité (d'un système d'IA) à **rendre des comptes** (*Accountability*)
- la **transparence** (*Transparency*)
- la **fiabilité** (*Trustworthiness*)
- l'**équité** (*Fairness*)
- la **conformité** (respect des règles et de la législation)
- la **loyauté** (fidélité de la réalisation par rapport aux spécifications)
- l'identification des **responsabilités**

… qui créent, ensemble, la **confiance** en l'outil développé

« **apprendre sans comprendre** » est à la base des mécanismes d'apprentissage. C'est un apprentissage qui peut même être sans but (énoncé ou spécifié d'avance)

comment **expliquer les résultats** (transparence) d'un algorithme d'apprentissage qui évolue sans cesse pendant son utilisation

comment **vérifier la conformité (et la loyauté)** des résultats produits par un algorithme d'apprentissage… alors que la spécification des résultats attendus n'existe pas

en cas de faute, 3 acteurs sont impliqués : — le concepteur de l'algorithme, — l'utilisateur des résultats produits… — et le mécanisme d'apprentissage qui continue d'apprendre au cours de son usage

5.4 IA : éthique de son usage

LA DÉCLARATION DE MONTRÉAL POUR UN DÉVELOPPEMENT RESPONSABLE DE L'INTELLIGENCE ARTIFICIELLE
2018

1 - Principe de bien-être
- améliorer les conditions de vie, santé et conditions de travail
- permettre d'exercer ses capacités physiques et intellectuelles
- ne doit pas augmenter le stress, l'anxiété, le harcèlement

2 - Principe du respect de l'autonomie
- permettre à chacun de réaliser sa propre conception de la vie
- éviter les dépendances par une confusion entre IA et humains
- ne doit pas mettre en œuvre des mécanismes de surveillance, d'évaluation ou d'incitation contraignants

3 - Principe de protection de l'intimité et la vie privée
- protéger des espaces d'intimité non surveillés
- protéger l'intimité de la pensée et des émotions
- pas de profil personnalisé pour influencer le comportement
- avoir un contrôle étendu sur ses données personnelles

4 - Principe de solidarité
- favoriser les relations humaines et réduire l'isolement
- favoriser le travail collaboratif
- ne pas simuler des comportements cruels par des robots

5 - Principe de participation démocratique
- une IA qui peut affecter la qualité de vie des personnes doit être intelligible et justifiable
- le code des algorithmes doit être accessible aux autorités
- la découverte d'effets non prévus doit être signalée

6 - Principe d'équité
- ne pas créer de discriminations sociales, religieuses, ethniques..
- éliminer les relations de domination fondées sur la richesse, le pouvoir, la connaissance
- bénéficier économiquement à tous

7 - Principe d'inclusion de la diversité
- ne pas induire l'uniformisation ou la normalisation
- respecter les multiples expressions de toutes les diversités
- pour chaque catégorie de service, une offre d'IA diversifiée

8 - Principe de prudence
- restreindre la diffusion d'IA pouvant présenter un danger
- satisfaire des critères de fiabilité, sécurité intégrité... testés
- prévenir des risques d'usage néfaste des données et de l'IA

9 - Principe de responsabilité
- seuls des êtres humains peuvent être tenus pour responsables
- une décision qui affecte la vie doit être prise par une personne

10 - Principe de développement soutenable
- efficacité énergétique des infrastructures des IA
- filière de maintenance, réparation et recyclage
- lutter contre le gaspillage des ressources naturelles