

# IA POUR LA SOCIÉTÉ

## ÉTHIQUE, RESPONSABLE, DE CONFIANCE

---

Dr. Rémy Chaput

2023/02/28

Séminaire Abylsen

<https://rchaput.github.io/talk/abylsen-2023/>

# DE QUOI VA-T-ON PARLER ?

Intelligence Artificielle

(Apprentissage Machine)

Interactions entre systèmes d'IA et notre société

Tour d'horizon des problématiques, craintes, solutions possibles

Pas de détails techniques

# CONTEXTE

# L'IA, C'EST QUOI ?

- IA = “Intelligence” + “Artificielle”
- “Intelligence” = ??
- Une réaction intuitive : “Intelligence” = propre de l’humain
- Mais : “Artificiel” = ce qui n’est pas humain
- $\Rightarrow$  IA = Atteint le niveau d’un humain sans être humain
- $\Rightarrow$  “Effet IA” : redéfinition permanente  
“L’IA est tout ce qui n’a pas encore été résolu” (McCarthy, Minsky)

# QUELQUES EXEMPLES DE L'EFFET IA

- Jouer aux échecs, au Go, ...
- Reconnaissance automatique de caractères (texte)
- Traduction (Google Translate, DeepL)
- Aide à la décision
- Preuve formelle de théorèmes logiques

⇒ “C’est juste des maths !”

# BEAUCOUP DE MOTS-CLÉS

IA ...

- éthique
- responsable
- pour la société
- centrée sur l'humain
- sûre
- de confiance
- alignée sur les valeurs humaines
- bénéfique

etc.

# IMPACT DES SYSTÈMES SUR LA SOCIÉTÉ

- De plus en plus de systèmes déployés
- Démocratisation de ces systèmes
  - API disponibles (payantes ou non)
  - **Alaas** : *AI as a Service*
- Impact sur nos vies, notre environnement
- Ex: embauche automatique ; décision de crédit ; etc.

# INTÉRÊT DE LA SOCIÉTÉ POUR CES SYSTÈMES

- Intérêt du grand public d'une part
  - Scandales et controverses : [Tay](#), [Deepfakes](#), etc.
  - Applications intéressantes / amusantes : génération de texte (ChatGPT), images (Midjourney), etc.
  - Interrogations au sujet des futurs emplois
- Mais aussi des organisations
  - (inter)gouvernementales, à but non-lucratif, ...
  - *Nombreux* documents produits : [Déclaration de Montréal](#), [Principes d'Asilomar](#) etc.
  - Régulations en cours : [EU AI Act](#)



# PROBLÈMES LIÉS À L'IA

# PRINCIPALES CATÉGORIES DE PROBLÈMES

- Vie privée
- Discrimination / Biais
- Décision inexplicable
- Objectifs non-alignés

Mutuellement non exclusives

# VIE PRIVÉE

- Apprentissage Machine requiert *énormément* de données
- Traces sur l'Internet = beaucoup de données
- Jeux de données pas forcément utilisés avec consentement
- Systèmes qui collectent des données (abusives ?) pour la prise de décision

# VIE PRIVÉE

Exemple : assistants vocaux


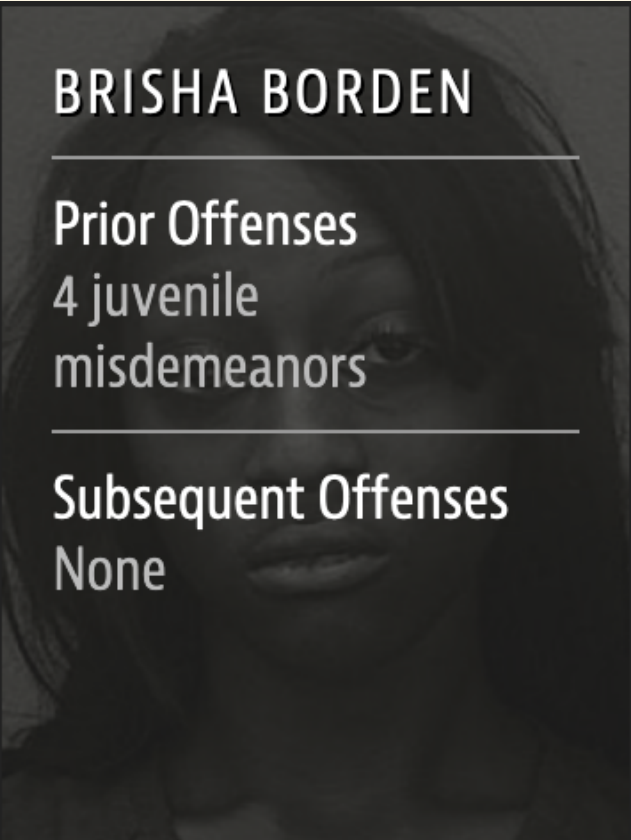
- Apple Siri, Amazon Alexa, Google Echo, ...
  - Et plus récemment les aspirateurs “autonomes”
- Collectent des données en permanence
  - Nécessaire pour leur fonctionnement
  - Mais pose des inquiétudes
- Peuvent transmettre ces données à de tierces personnes
  - Utile pour améliorer leur qualité de service
  - Mais données vocales sont extrêmement personnelles

# DISCRIMINATION / BIAIS

- Jeux de données pas toujours équilibrés
- Prédictions parfois faussées
- Modèles génératifs biaisés
- Populations sous-représentées ou pas prises en compte

# DISCRIMINATION / BIAIS

Exemple : COMPAS

	
<b>VERNON PRATER</b>	<b>BRISHA BORDEN</b>
<b>Prior Offenses</b> 2 armed robberies, 1 attempted armed robbery	<b>Prior Offenses</b> 4 juvenile misdemeanors
<b>Subsequent Offenses</b> 1 grand theft	<b>Subsequent Offenses</b> None
<b>LOW RISK</b> <b>3</b>	<b>HIGH RISK</b> <b>8</b>

Source: ProPublica

# DISCRIMINATION / BIAIS

Exemple : Larges modèles de langage(s)

- En majorité entraînés principalement sur l'anglais
- Certains modèles gèrent environ 100 langages
  - ... Mais il existe environ 7000 langages dans le monde
- Performances dégradées quand langage autre que l'anglais
- Les systèmes de traduction gèrent un peu plus de langages
  - ... Mais parfois avec des erreurs

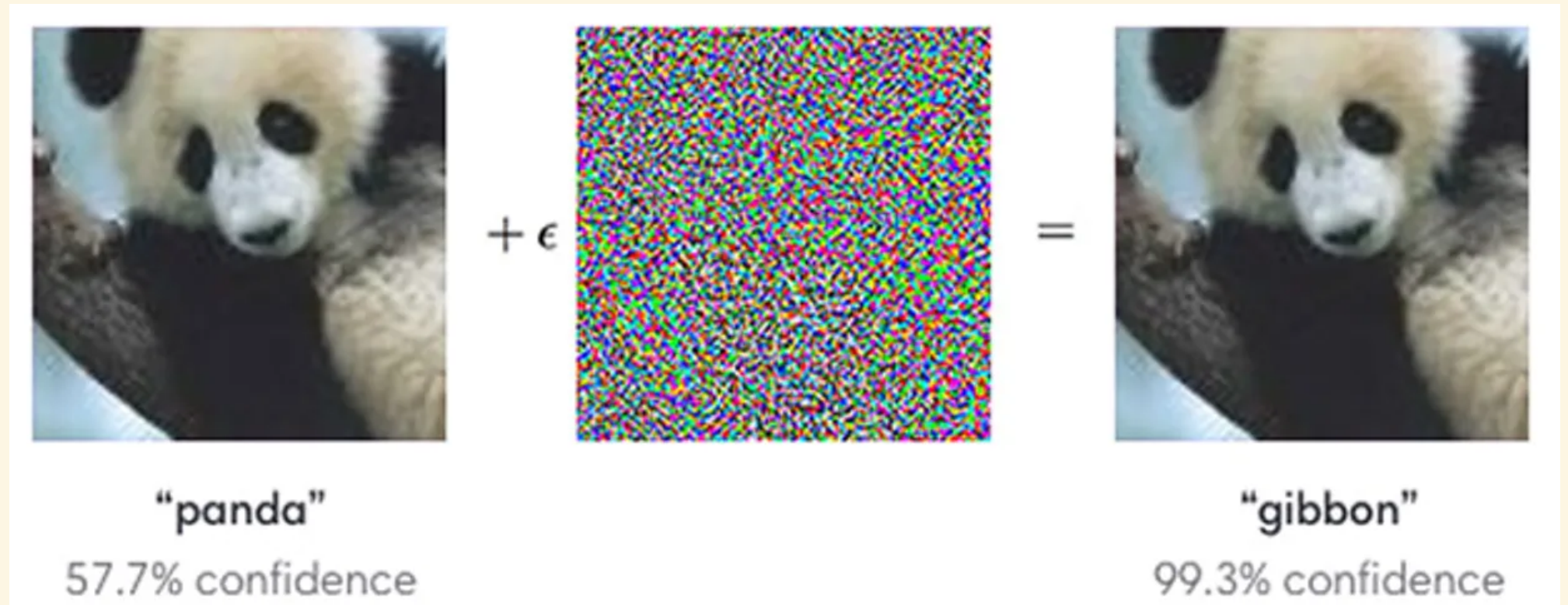
# DÉCISION INEXPLICABLE

- Utilisation de réseaux de neurones complexes (Millions de paramètres)
- Difficile (impossible ?) d'analyser *pourquoi* une décision a été prise
- Réseaux de neurones fonctionnent dans des espaces à très hautes dimensions (1000+)
  - Impossible pour nous de les imaginer
  - Donc de comprendre quelles corrélations sont apprises



# DÉCISION INEXPLICABLE

Exemple : images contradictoires



Source : OpenAI via IEEE

# OBJECTIFS NON-ALIGNÉS

- Problème du “Roi Midas” :
  - obtenir exactement ce que l’on a demandé
  - ... pas ce que l’on souhaite
- Exemple :
  - on veut réduire la production de  $\text{CO}^2$
  - le système propose de tuer tous les humains

# OBJECTIFS NON-ALIGNÉS

Pas un problème nouveau !

If we use, to achieve our purposes, a mechanical agency with whose operation we cannot interfere effectively ... we had better be quite sure that the purpose put into the machine is the purpose which we really desire.

(N. Wiener, 1960)

# OBJECTIFS NON-ALIGNÉS

Exemple : Coastrunner



Source : OpenAI

# DES DÉBUTS DE SOLUTION

# VIE PRIVÉE

- Chercheurs essaient de produire des systèmes qui nécessitent moins de données ou qui les anonymisent
  - Ex: Apprentissage fédéré
- Faire attention aux jeux de données utilisés ⇒ [Datasheets for Datasets](#)
- Déterminer les données dont vous avez absolument besoin et s'y limiter

# DISCRIMINATION / BIAIS

- Facile, retirer toutes les informations liées à un biais ?
- $\Rightarrow$  Nope
- Apprentissage machine très doué pour les corrélations
- Informations sur le nom, l'adresse, l'école, ... corrélient avec couleur de peau, sexe, ...
- $\Rightarrow$  Très difficile (impossible ?) de tout retirer !

# DISCRIMINATION / BIAIS

- Certains biais sont *nécessaires*
- Ex: embauche automatique sans entretien
  - Pas de bias = prendre un CV au hasard parmi l'ensemble des êtres humains de la planète (C. Tessier)
- ⇒ La question est de *choisir* les biais avec lesquels on est d'accord



# DISCRIMINATION / BIAIS

Exemple fictif : embauche automatique

- Entreprise composée d'anciens militaires
- Recrutent d'autres anciens militaires  $\Rightarrow$  camaraderie plus simple
- Jouent tous les soirs au football
- Décident d'utiliser un système d'IA pour automatiser le recrutement (trop de CV reçus)

$\Rightarrow$  Que va décider le système ?

$\Rightarrow$  Est-ce "normal" ? "acceptable" ?

# DISCRIMINATION / BIAIS

- Possibilité d'utiliser des systèmes d'IA pour analyser / détecter les biais
- Changer une donnée et observer l'impact sur le résultat
  - Ex: demande de crédit, si on diminue le salaire, le résultat passe de "OK" à "PAS OK"  $\Rightarrow$  normal ?
  - Même ex: on change le sexe de "H" à "F", le résultat passe de "OK" à "PAS OK"  $\Rightarrow$  pas normal ?
- Systèmes d'IA permettent d'automatiser ce processus

# DÉCISIONS INEXPLICABLES

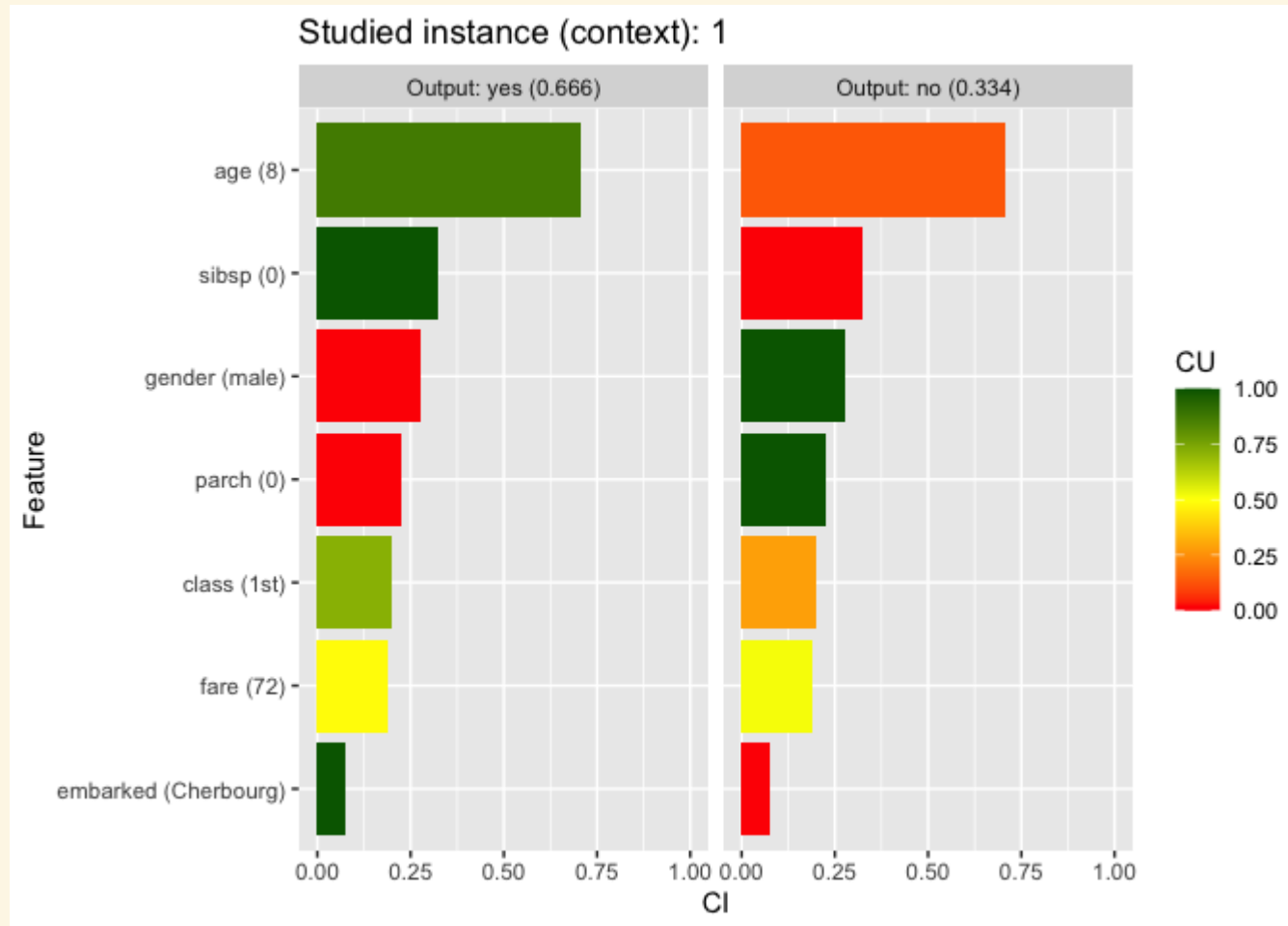
- Chercheurs essaient de produire des systèmes plus “interprétables”
- Possibilité d'utiliser des systèmes moins complexes, avec moins de variables
  - Ex: arbres de décision

# DÉCISIONS INEXPLICABLES

- Autre possibilité : expliquer à partir des données d'entrée
- Complètement agnostique du système utilisé
  - (+) Flexible, utilisable sur n'importe quel système
  - (-) L'explication est-elle vraiment fidèle ?

# DÉCISIONS INEXPLICABLES

Exemple : *Contextual Importance and Utility* (CIU)



# OBJECTIFS NON-ALIGNÉS

Nécessite à la fois des outils côté IA

Mais aussi une réflexion personnelle

Objectifs désirés

↕ Problème d'alignement *externe*

Objectifs exprimés

↕ Problème d'alignement *interne*

Objectifs satisfaits

# OBJECTIFS NON-ALIGNÉS

Côté IA (*interne*) :

- Surveiller les décisions du système
- Liens avec :
  - la robustesse
  - la sûreté (*AI Safety*)
  - la vérification formelle
  - la détection d'anomalie

# OBJECTIFS NON-ALIGNÉS

Côté concepteurs humains (*externe*) :

- Quelles sont les personnes impactées par le système ?
  - Utilisateurs
  - Mais plus globalement parties prenantes
- Quelles sont les valeurs importantes pour votre entreprise ?
- Possibilité d'utiliser des outils pour guider la réflexion



# CONCLUSION

# L'IA, UNE TECHNOLOGIE D'AVENIR

- Impressionnantes performances des systèmes d'IA
  - Surtout les plus récents
  - Résolution des tâches à un niveau quasi-humain (parfois meilleur)
- De plus en plus de tâches “résolues”
  - Vision par ordinateur, traitement du langage naturel, ...
- Augmentation des fonds investis dans les projets d'IA
  - Fonds publics ([Stratégie Nationale pour l'IA](#))
  - Fonds privés (OpenAI, DeepMind, Meta AI, startups ...)

# DES PROBLÈMES ÉPINEUX

- Vie privée ; Discrimination & Biais ; Décisions inexplicables ; Objectifs non-alignés
- Pas seulement techniques
- Plus le système est performant, plus l'impact est important

# PAS DE SOLUTION MAGIQUE

- Demande une part importante de réflexion humaine
- La réflexion doit souvent être continue dans le temps
  - *Avant* la conception du système
  - *Pendant* la conception
  - *Après* la conception / le déploiement

```
1 from ethical_ai import make_model_safe
2
3 safe_model = make_model_safe(MySuperAIModel())
4
5 >>> ModuleNotFoundError: No module named 'ethical_ai'
```

# MERCI DE VOTRE ATTENTION

Questions ?