Éthique et IA (v2 ?)

Éthique dans la conception

Remy Chaput

1er Octobre 2021

LIRIS - SyCoSMA

Contexte

Pourquoi parler d'éthique ?

- De plus en plus d'applications avec IA déployées

Pourquoi parler d'éthique ?

- De plus en plus d'applications avec IA déployées
- Un impact sur les humains
  - Vies individuelles
  - La société dans son ensemble

Pourquoi parler d'éthique ?

- De plus en plus d'applications avec IA déployées
- Un impact sur les humains
  - Vies individuelles
  - La société dans son ensemble
- Des questionnements sur les effets
  - Application systématique du *Machine Learning*
  - Cas omniprésent de la *fairness*
  - Et d'autres

Quel niveau d'éthique ?

Éthique pour la conception

- Codes de conduite, standards, processus de certifications
- Assure l'intégrité des développeurs et des utilisateurs

Quel niveau d'éthique ?

Éthique pour la conception

- Codes de conduite, standards, processus de certifications
- Assure l'intégrité des développeurs et des utilisateurs

Éthique par conception

- Algorithmes pour des "capacités éthiques"
- Le résultat est dans le comportement de l'agent

Quel niveau d'éthique ?

Éthique pour la conception

- Codes de conduite, standards, processus de certifications
- Assure l'intégrité des développeurs et des utilisateurs

Éthique par conception

- Algorithmes pour des "capacités éthiques"
- Le résultat est dans le comportement de l'agent

Éthique dans la conception

- Implications éthiques des systèmes d'IA
- Alors qu'ils intègrent la société

Quelques exemples de problèmes...

Les biais

- Évaluation de la probabilité de récidive (système COMPAS)
- Recrutement (partiellement) automatique

- Reproduisent les biais contenus dans le jeu de données
  - Biais historiques, biais par manque ou sur-représentation, ...
- Peut-on supprimer les biais ?

Les biais

- Évaluation de la probabilité de récidive (système COMPAS)
- Recrutement (partiellement) automatique

- Reproduisent les biais contenus dans le jeu de données
  - Biais historiques, biais par manque ou sur-représentation, ...
- Peut-on supprimer les biais ?
  - C. Tessier : Non, sinon on fait de l'aléatoire
  - Il faut choisir ce qui nous semble pertinent

Performance et vie privée

- Le *Machine Learning* promet de grandes performances
- Par ex., pour la détection de cancer

Performance et vie privée

- Le *Machine Learning* promet de grandes performances
- Par ex., pour la détection de cancer

- Mais requiert de grandes quantités de données

Performance et vie privée

- Le *Machine Learning* promet de grandes performances
- Par ex., pour la détection de cancer

- Mais requiert de grandes quantités de données

- $\Longrightarrow$ il y a un enjeu important entre la santé et la vie privée

Les solutions proposées

Des principes

- $\sim$ 100 documents proposés
- Différentes sources :
  - Entreprises privées
  - Agences gouvernementales
  - Monde académique
  - Organismes inter- ou supra-nationaux
  - Associations à but non lucratif
  - Associations professionnelles et sociétés savantes
  - etc.

Des principes

- $\sim$ 100 documents proposés
- Différentes sources :
  - Entreprises privées
  - Agences gouvernementales
  - Monde académique
  - Organismes inter- ou supra-nationaux
  - Associations à but non lucratif
  - Associations professionnelles et sociétés savantes
  - etc.

- 11 principes récurrents
  - **Transparence**
  - **Justice**, **équité** (*fairness*)
  - **Non-malveillance**
  - **Responsabilité**
  - **Vie privée**
  - Bénévolence
  - Liberté, autonomie
  - Confiance
  - Dignité
  - Viabilité (durabilité)
  - Solidarité

Comment les implémenter ?

Figure 1: Photo d'un développeur essayant d'implémenter des principes

Comment les implémenter ?

Les principes peuvent :

- Être trop abstrait, ne pas avoir de définition utilisable
  - Des opinions politiques, idéologiques ou philosophiques
  - Par ex., ”AI for social good”
- Avoir plusieurs définitions
  - Ou être interprétés différemment par des cultures différentes
- Être en conflit avec d’autres principes
  - De manière fondamentale ou seulement dans une situation donnée

Un exemple concret : l'équité

Un petit "jeu" :

- Comment répartir de l'argent de l'argent entre des étudiants ?
- Jeu de l'ultimatum
  - Une personne propose une somme ; Une autre accepte ou refuse
- 4 scénarios :
  - Français/Indien, Français/Français, Indien/Français, Indien/Indien

Un exemple concret : l'équité

Un petit "jeu" :

- Comment répartir de l'argent de l'argent entre des étudiants ?
- Jeu de l'ultimatum
  - Une personne propose une somme ; Une autre accepte ou refuse
- 4 scénarios :
  - Français/Indien, Français/Français, Indien/Français, Indien/Indien
- Non pas 1, mais 3 théories possibles de l'équité !

Un exemple concret : l'équité

Un petit "jeu" :

- Comment répartir de l'argent de l'argent entre des étudiants ?
- Jeu de l'ultimatum
  - Une personne propose une somme ; Une autre accepte ou refuse
- 4 scénarios :
  - Français/Indien, Français/Français, Indien/Français, Indien/Indien
- Non pas 1, mais 3 théories possibles de l'équité !
  - Équité formelle
    - Chacun doit recevoir la moitié

Un exemple concret : l'équité

Un petit "jeu" :

- Comment répartir de l'argent de l'argent entre des étudiants ?
- Jeu de l'ultimatum
  - Une personne propose une somme ; Une autre accepte ou refuse
- 4 scénarios :
  - Français/Indien, Français/Français, Indien/Français, Indien/Indien

- Non pas 1, mais 3 théories possibles de l'équité !
  - Équité formelle
    - Chacun doit recevoir la moitié
  - Équité globale ou de compensation
    - On veut réduire les inégalités globales
    - Les Indiens sont en moyenne plus pauvres que les Français
    - Donc on doit donner plus aux Indiens

Un exemple concret : l'équité

Un petit "jeu" :

- Comment répartir de l'argent de l'argent entre des étudiants ?
- Jeu de l'ultimatum
  - Une personne propose une somme ; Une autre accepte ou refuse
- 4 scénarios :
  - Français/Indien, Français/Français, Indien/Français, Indien/Indien

- Non pas 1, mais 3 théories possibles de l'équité !
  - Équité formelle
    - Chacun doit recevoir la moitié
  - Équité globale ou de compensation
    - On veut réduire les inégalités globales
    - Les Indiens sont en moyenne plus pauvres que les Français
    - Donc on doit donner plus aux Indiens
  - Équité locale
    - On veut égaliser les bénéfices de cette situation
    - On achète moins en France avec la même somme qu'en Inde
    - Donc on doit donner plus aux Français

Un exemple concret : l'équité

Un petit "jeu" :

- Comment répartir de l'argent de l'argent entre des étudiants ?
- Jeu de l'ultimatum
  - Une personne propose une somme ; Une autre accepte ou refuse
- 4 scénarios :
  - Français/Indien, Français/Français, Indien/Français, Indien/Indien

- Non pas 1, mais 3 théories possibles de l'équité !
  - Équité formelle
    - Chacun doit recevoir la moitié
  - Équité globale ou de compensation
    - On veut réduire les inégalités globales
    - Les Indiens sont en moyenne plus pauvres que les Français
    - Donc on doit donner plus aux Indiens
  - **Équité locale**
    - On veut égaliser les bénéfices de cette situation
    - On achète moins en France avec la même somme qu'en Inde
    - Donc on doit donner plus aux Français

Des certifications

- Certifier que le code et/ou le déroulement du code répondent à des critères

Des certifications

- Certifier que le code et/ou le déroulement du code répondent à des critères
- Ça marche pour certains critères (par ex., sécurité)

Des certifications

- Certifier que le code et/ou le déroulement du code répondent à des critères
- Ça marche pour certains critères (par ex., sécurité)
- … Mais pas pour l'éthique de manière générale

V. Dignum: "Responsible AI is more than ticking boxes"

Dans quelle direction va-t-on ?

Des tensions entre les principes

Tension

Un conflit entre des principes, valeurs ou buts

- Peuvent être fondamentales
  - i.e., une incompatibilité morale
  - On n'aura jamais les deux
- Ou bien contingentes
  - Résultent de contraintes technologiques actuelles
  - Par ex., le Deep Learning actuel est très efficace, mais peu interprétable

Pourquoi s'intéresser aux tensions ?

- Faire le pont entre les principes et la pratique
  - Les principes sont trop abstraits
  - Un cas individuel n'est pas généralisable
  - Mais les tensions permettent de regarder plusieurs cas

Pourquoi s'intéresser aux tensions ?

- Faire le pont entre les principes et la pratique
  - Les principes sont trop abstraits
  - Un cas individuel n'est pas généralisable
  - Mais les tensions permettent de regarder plusieurs cas
- Reconnaître les différences dans les valeurs
  - Considérer comment les principes sont interprétés dans différents groupes
  - Par ex., quelle définition utiliser pour équité ?

Pourquoi s'intéresser aux tensions ?

- Faire le pont entre les principes et la pratique
  - Les principes sont trop abstraits
  - Un cas individuel n'est pas généralisable
  - Mais les tensions permettent de regarder plusieurs cas
- Reconnaître les différences dans les valeurs
  - Considérer comment les principes sont interprétés dans différents groupes
  - Par ex., quelle définition utiliser pour équité ?
- Souligner les verrous scientifiques
  - Une tension n'est pas forcément un conflit fondamental
  - On peut chercher des solutions pour réduire le problème
  - Ex: des techniques de Deep Learning qui soient plus interprétables / explicables

Pourquoi s'intéresser aux tensions ?

- Faire le pont entre les principes et la pratique
  - Les principes sont trop abstraits
  - Un cas individuel n'est pas généralisable
  - Mais les tensions permettent de regarder plusieurs cas
- Reconnaître les différences dans les valeurs
  - Considérer comment les principes sont interprétés dans différents groupes
  - Par ex., quelle définition utiliser pour équité ?
- Souligner les verrous scientifiques
  - Une tension n'est pas forcément un conflit fondamental
  - On peut chercher des solutions pour réduire le problème
  - Ex: des techniques de Deep Learning qui soient plus interprétables / explicables
- Identifier le manque de connaissances
  - Qu'est-ce que signifient équité, autonomie, … ?
  - Qu'est-ce qui est techniquement possible ? Avec quels outils ? Quels coûts ?

Des exemples de tensions

- Utiliser les données pour améliorer la qualité et l'efficacité des services VS. respecter la vie privée et l'autonomie des individus

Des exemples de tensions

- Utiliser les données pour améliorer la qualité et l'efficacité des services VS. respecter la vie privée et l'autonomie des individus
- Utiliser des algorithmes pour prendre des décisions et prédictions plus précises VS. assurer un traitement équitable

Des exemples de tensions

- Utiliser les données pour améliorer la qualité et l'efficacité des services VS. respecter la vie privée et l'autonomie des individus
- Utiliser des algorithmes pour prendre des décisions et prédictions plus précises VS. assurer un traitement équitable
- Augmenter la personnalisation dans la sphère digitale VS. augmenter la solidarité et la citoyenneté

Des exemples de tensions

- Utiliser les données pour améliorer la qualité et l'efficacité des services VS. respecter la vie privée et l'autonomie des individus
- Utiliser des algorithmes pour prendre des décisions et prédictions plus précises VS. assurer un traitement équitable
- Augmenter la personnalisation dans la sphère digitale VS. augmenter la solidarité et la citoyenneté
- Utiliser l'automatisation pour rendre la vie plus facile VS. promouvoir la dignité et l'auto-actualisation

Comment identifier des tensions ?

Processus de réflexion et de questionnement

- Quand on utilise un système pour promouvoir une valeur, quels risques pour d'autres valeurs peuvent être introduits ?

Comment identifier des tensions ?

Processus de réflexion et de questionnement

- Quand on utilise un système pour promouvoir une valeur, quels risques pour d'autres valeurs peuvent être introduits ?
- Quand un groupe est avantagé par un système, quel autre groupe peut être désavantagé ?

Comment identifier des tensions ?

Processus de réflexion et de questionnement

- Quand on utilise un système pour promouvoir une valeur, quels risques pour d'autres valeurs peuvent être introduits ?
- Quand un groupe est avantagé par un système, quel autre groupe peut être désavantagé ?
- Quel compromis entre les avantages à court- et long-terme ?

Comment identifier des tensions ?

Processus de réflexion et de questionnement

- Quand on utilise un système pour promouvoir une valeur, quels risques pour d'autres valeurs peuvent être introduits ?
- Quand un groupe est avantagé par un système, quel autre groupe peut être désavantagé ?
- Quel compromis entre les avantages à court- et long-terme ?
- Quels futurs développements de l'IA peuvent améliorer ou au contraire menacer des valeurs ?

Comment résoudre les tensions

Pour une tension fondamentale :

- Il faut un choix
- Processus de réflexion
- Prise de décision démocratique, ...

Comment résoudre les tensions

Pour une tension fondamentale :

- Il faut un choix
- Processus de réflexion
- Prise de décision démocratique, ...

Pour une tension contingente :

- On peut attendre d'avoir de meilleurs outils
  - $\Longrightarrow$ nouvelle tension entre les gains à court-terme et long-terme
- Ou bien faire un compromis immédiatement
  - $\Longrightarrow$ même cas que la tension fondamentale

En résumé

Conclusion

- Les implications éthiques sont reconnues par la société

Conclusion

- Les implications éthiques sont reconnues par la société
- … Mais les solutions actuelles ne sont pas adéquates

Conclusion

- Les implications éthiques sont reconnues par la société
- … Mais les solutions actuelles ne sont pas adéquates

- Proposition de se focaliser sur les **tensions**

Conclusion

- Les implications éthiques sont reconnues par la société
- … Mais les solutions actuelles ne sont pas adéquates

- Proposition de se focaliser sur les **tensions**
- Pour les résoudre, il faut une discussion, un débat

Conclusion

- Les implications éthiques sont reconnues par la société
- … Mais les solutions actuelles ne sont pas adéquates

- Proposition de se focaliser sur les **tensions**
- Pour les résoudre, il faut une discussion, un débat
- Pas de “solution magique” ou universelle

Questions ?

References


Romina Boarini, Jean-François Laslier, and Stéphane Robin. “Interpersonal comparisons of utility in bargaining: evidence from a transcontinental ultimatum game”. In: *Theory and Decision* 67.4 (2009), pp 341–373.


Virginia Dignum. *Ethics in artificial intelligence: introduction to the special issue*. 2018.


Anna Jobin, Marcello Ienca, and Effy Vayena. “The global landscape of AI ethics guidelines”. In: *Nature Machine Intelligence* 1.9 (Sept. 2019), pp. 389–399.

References ii


Jess Whittlestone et al. “The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions”. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. AIES ’19. New York, NY, USA: Association for Computing Machinery, Jan. 27, 2019, pp. 195–200.