Point de vue algorithmique de l'éthique de l'IA : cas de l'Internet des objets et des données privées

Bruno DEFUDE et Sophie CHABRIDON, laboratoire Samovar, Télécom SudParis

Résumé: Après avoir introduit les principes pour un algorithme éthique (équité, transparence et redevabilité), cette présentation donne un aperçu de la recherche sur les aspects algorithmiques liés à l'interprétabilité des systèmes d'IA et notamment des réseaux de neurones profonds (deep learning). Les principales difficultés liées à l'interprétabilité sont la complexité de ces réseaux ainsi que la faible connaissance que l'on a parfois sur eux (dans le cas extrême, ce sont des boîtes noires dont on ne perçoit que les entrées et les sorties). Beaucoup de travaux sont menés actuellement sur ce domaine essayant par exemple de déterminer quelle portion d'une entrée pèse le plus dans une décision (quelle partie d'une image a permis de reconnaître un objet particulier) ou bien de comprendre le rôle de la structure du réseau (quel est le rôle d'une couche spécifique d'un réseau profond).

Nous prenons ensuite l'exemple de l'Internet des objets (IoT), exemple de technologies pénétrant de plus en plus notre vie quotidienne et pour lequel la prise en compte d'une dimension éthique dès la conception devient indispensable. Les caractéristiques techniques uniques de l'IoT en termes de dynamicité, très large échelle et volume de données collectées requièrent des solutions nouvelles. Devant la facilité de collecte de données privées, nous nous intéressons en particulier aux mécanismes permettant de prendre compte ces données privées dans les traitements, mais de manière non discriminatoire et en les protégeant, c'est-à-dire en manipulant des données chiffrées sans dévoiler leur contenu.

Références

S. Gambs, Quantifier et mesurer la discrimination
http://transparence.conf.citi-lab.fr/css/2017_style/img/files/quantifier_mesurer_discriminationSGAMBS.pdf

S. Kuper et al., Toward Scalable Verification for Safety-Critical Deep Networks
https://arxiv.org/abs/1801.05950

https://medium.com/@deepmindsafetyresearch/towards-robust-and-verified-ai-specification-testing-robust-training-and-formal-verification-69bd1bc48bda

W. Samek, K.R. Muller. Tutorial on interpretable machine learning, *GCPR'17 conference*
http://www.heatmapping.org/slides/2017_GCPR.pdf

B. Kim, F Doshi-Velez. Tutorial on interpretable machine learning, *ICML 2017*

B. Herman. The promise and peril of human evaluation for model interpretability (2017)
https://arxiv.org/abs/1711.07414

L. H. Gelpinet al. Explaining explanations: an overview of interpretability of machine
https://arxiv.org/abs/1806.00069

M. Drosou, H.V. Jagadish, E. Pitoura, and J. Stoyanovich. Diversity in Big Data: A Review. *Big Data Review*, **5(2)**, (2018)
https://doi.org/10.1089/big.2016.0054

The Ethics Guidelines for Trustworthy Artificial Intelligence
https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top

N. Kilbertus et al. Blind justice: Fairness with encrypted sensitive attributes, *35th Int. Conf. on Machine Learning 2018*
https://arxiv.org/abs/1806.03281

R. Kohavi, B. Becker. Census Income dataset. *UCI Machine Learning Repository*
https://archive.ics.uci.edu/ml/datasets/adult

Montréal declaration for a responsible development of Artificial Intelligence (2018)
https://www.montrealdeclaration-responsibleai.com

S. Tolan. Fair and Unbiased Algorithmic Decision Making: Current State and Future Challenges. *JRC Digital Economy Working Paper* (2018)
http://arxiv.org/abs/1901.04730

B. Turner. Extract of National Survey on Drug Use and Health (2015)
https://data.world/balexturner/drug-use-employment-work-absence-income-race-education

I. Zliobaitė, B. Custers. Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law*, **24(2)** (2016)
https://doi.org/10.1007/s10506-016-9182-5

Point de vue algorithmique de l'éthique de l'IA : cas de l'Internet des objets et des données privées

Bruno Defude

Samovar, Telecom SudParis

Transparence des algorithmes

- Equité (fairness)
  - Comportement biaisé (moteur de recherche de google)
  - Reproduction d'un biais présent dans les données (recommandation)
- Transparence difficile à définir (transparency)
- Redevabilité des algorithmes : « devoir de rendre compte » (accountability)
  - Respect de règles (notamment juridiques et éthiques)
  - Rendre intelligible la logique sous jacente du traitement
- http://binaire.blog.lemonde.fr/2017/12/16/algorithmes-au-dela-de-la-transparence-la-redevabilite/

Egalité, discrimination, équité [1]

- Principe d'égalité : les hommes et les femmes doivent être traités de la même manière indépendamment de leur sexe.
  Autrement dit, la valeur d'un individu est intrinsèque et ne doit pas être influencée par une caractéristique sensible (sexe, origine ethnique, . . .).
- Discrimination : peut-être vu comme une rupture du principe d'égalité.
- Equité : notion plus complexe que l'égalité qui essaye de corriger une situation injuste pouvant découler de l'application brutale du principe d'égalité.
  Par exemple, la discrimination positive peut être vue comme un mécanisme pour rétablir l'équité.

Différentes notions d'équité [1]

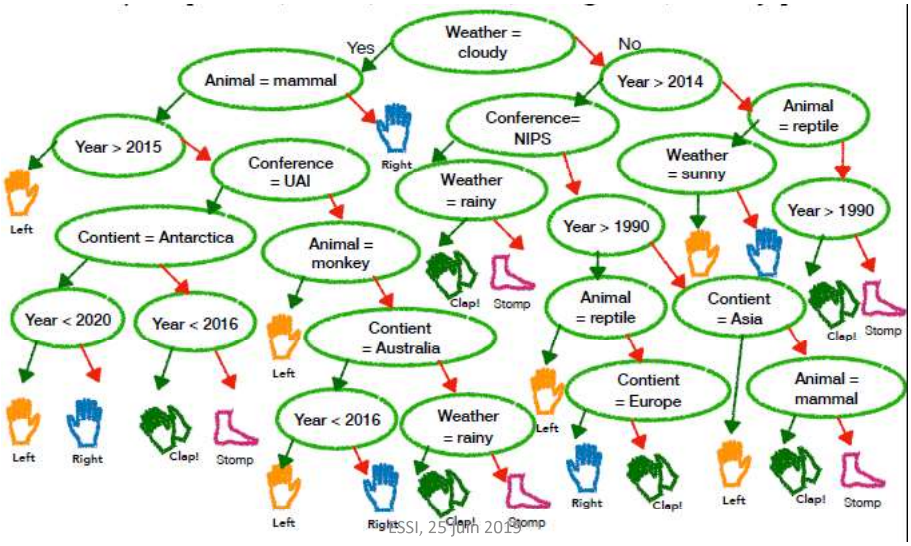
- Equité individuelle : deux individus dont les profils sont similaires à l'exception des attributs protégés devraient recevoir la même décision.
  Exemple : lors d'un recrutement pour un emploi, les candidats devraient être sélectionnés selon leurs compétences, indépendamment de leur sexe ou de leur origine ethnique.
  Problématique importante du choix de la mesure de similarité.
- Equité de groupe : les statistiques des décisions ciblant un groupe particulier devraient être approximativement les mêmes que dans la population globale.
  Exemple : si l'algorithme de recrutement doit produire une sortie binaire, accepté ou refusé, nous devons retrouver la même proportion d'hommes et de femmes acceptés.
- Difficulté : certaines études ont montré que certaines de ces métriques sont parfois incompatibles.

Transparence

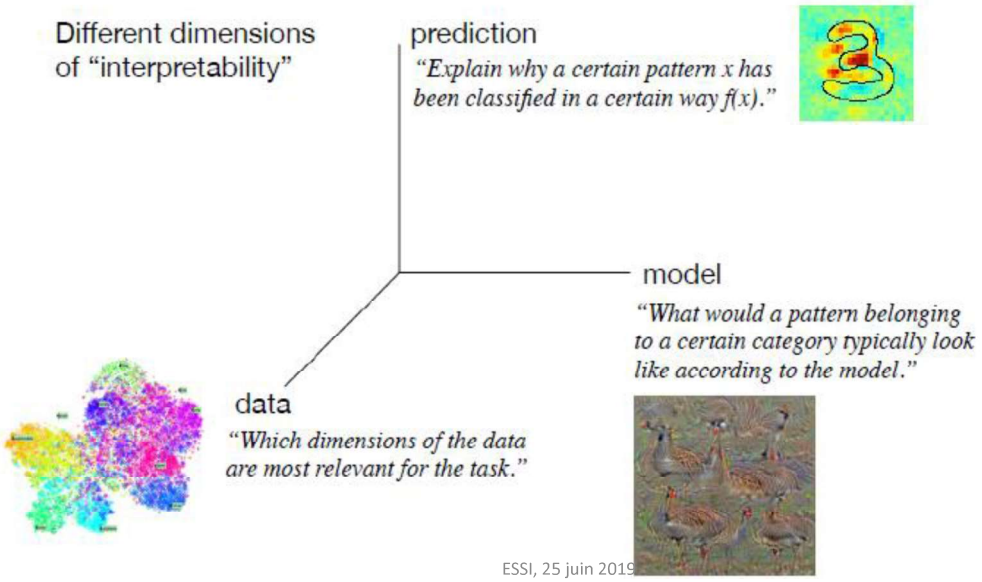
- Notion floue, mal définie
- Accès au code source permet elle la transparence ?
- Plutôt interprétation
  - Donner des explications à un humain
  - « bonnes » explications
  - Comprendre comment un système fonctionne (plus ou moins facile selon le système, ouvert à boite noire)
  - Tension entre interprétabilité (compréhensible par un humain) et complétude (rendre compte du système de la manière la plus précise possible)
  - Dilemmes éthiques [6]
    - When is it unethical to manipulate an explanation to better persuade users?
    - How do we balance our concerns for transparency and ethics with our desire for interpretability?

Interprétabilité pas seulement un problème pour RN-DL [5]

Entrée : [ICML, 2017, Australia, Kangaroo, Sunny]

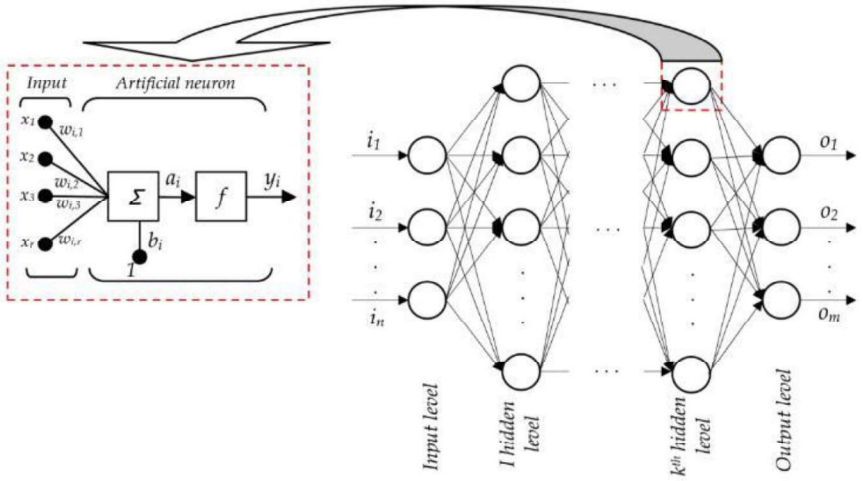


Dimensions de l'interprétabilité



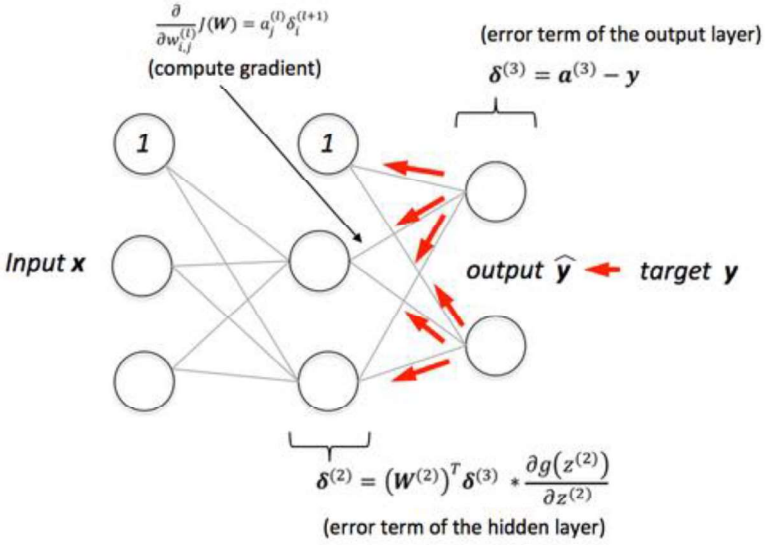
Extrait de [4]

Deep learning



http://www.intechopen.com/books/metallurgy-advances-in-materials-and-processes/artificial-intelligence-techniques-for-modelling-of-temperature-in-the-metal-cutting-process

Rétro-propagation



- Calcul itératif sur tout le training set jusqu'à convergence
- Erreur observée sur une entrée est rétro-propagée de la sortie vers l'entrée de manière à modifier les coefficients pour minimiser la fonction d'erreur

Explications dans les RN-DL (1)

- Complexité de ces systèmes les rend très peu intelligibles
  - ResNet, $5X10^7$ paramètres appris et 10^{10} opérations virgule flottante pour classifier une image!
- Besoin de réduire cette complexité
- Linear proxy models
  - LIME
  - Test un système boite noire en observant son comportement en perturbant les entrées : permet de construire un modèle linéaire local approximant le modèle au voisinage d'une entrée
  - Peut être utilisé pour identifier les régions d'une entrée qui ont le plus d'influence dans la prise de décision
  - Modèles de ce type sont prédictifs : peuvent être utilisés à la place du modèle complet

Explications dans les RN-DL (2)

- Decision trees
  - Décomposer un RN profond sous forme d'un arbre de décision (DeepRed)
  - Arbre peut être très grand (et donc peu interprétable)
  - Complexité en temps et en mémoire importante (passe pas à l'échelle)
- Automatic rule extraction
  - Approches décompositionnelles : travaillent au niveau du neurone pour extraire des règles simulant le comportement au niveau élémentaire (KT method), complexité exponentielle peu applicable sur des réseaux profonds
  - Approches « pédagogiques » : extraction des règles en observant le système comme une boite noire (relier entrées et sorties) : analyse de sensitivité ou méthodes d'échantillonage

Explications dans les RN-DL (3)

- Salience mapping
  - Test du réseau de manière répétitive en cachant une portion des entrées
  - Création d'une « image » (map) montrant quelles parties des entrées a véritablement de l'influence sur les sorties
  - Utilisation du gradient d'entrée ou de méthodes de décomposition (LRP)

Expliquer la classification d'une image avec LRP

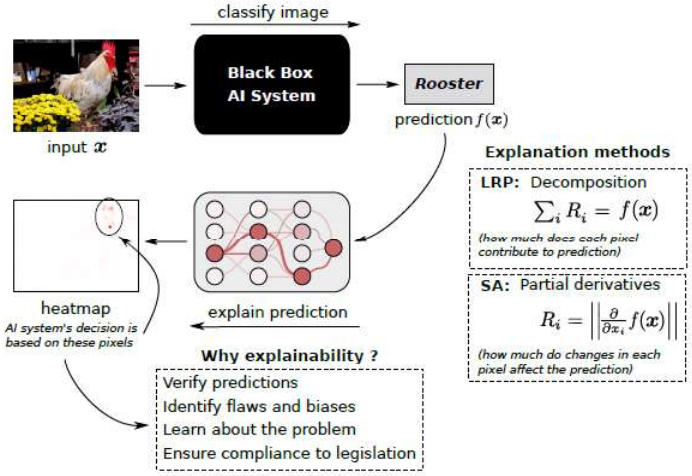


Fig. 1. Explaining predictions of an AI system. The input image is correctly classified as "rooster". In order to understand why the system has arrived at this decision, explanation methods such as SA or LRP are applied. The result of this explanation is an image, the heatmap, which visualizes the importance of each pixel for the prediction. In this example the rooster's red comb and wattle are the basis for the AI system's decision. With the heatmap one can verify that the AI system works as intended.

Extrait de [4]

Comprendre la structure des RN

- Identifier le rôle des couches
  - Tester leur capacité à traiter des problèmes différents du problème initial
  - Couche qui permet d'apprendre la « bonne » représentation des données d'entrée (en imagerie, feature extraction)
- Identifier le rôle d'un nœud à l'intérieur d'une couche
  - Visualiser les valeurs d'entrée qui maximise la valeur de sortie du nœud
  - Tester la capacité d'un nœud à résoudre une tache (network dissection)
- Pose aussi la question de la « bonne » structure du RN pour un problème donné
  - Quel nombre de couches ? Quels nombre de neurone par couche ?

Robustesse des algorithmes de ML

- L'introduction de bruit judicieusement choisi peut conduire à une erreur de classification

Au-delà de l'interprétabilité : la preuve [2]

- Contexte : nouveaux systèmes de contrôle aérien pilotés par un réseau de neurones (boite noire)
- Comment être sûr que la décision prise par la boite noire est correcte ?
- Comment être sûr qu'une propriété est bien respectée par la boite noire ?
- Idée : représenter un RN par un arbre où les feuilles sont les sorties du RN et les nœuds des conditions (fonction d'activation du neurone)
- RN de 6 niveaux avec 50 neurones par niveau représenté par un arbre avec 2^{300} feuilles!
- En fonction du problème, une grande partie de l'arbre n'est jamais explorée -> évaluation paresseuse guidée par le problème (2^{300} -> 2^{20})
- A permis de répondre à la question : est ce qu'un avion « inconnu » arrivant par la droite est détecté par le système? Et d'en avoir la preuve

Conclusion

- Domaine de recherche très actif
- Transparence mal définie
  - la transparence n'est pas forcément synonyme d'interprétabilité ou d'imputabilité
  - vérifier que l'exécution d'un programme corresponde au comportement intentionnel ou aux valeurs éthiques qui sont attendues comme la non-discrimination et l'équité
  - Lien fort avec la notion de loyauté (la propriété que le système se comporte comme promis)
- Interprétation des systèmes à améliorer
- Aller plus loin
  - Robustesse : mieux se comporter en cas d'attaque
  - Preuve : garantir le comportement d'un système d'IA

Références

- [1] S. Gambs, Quantifier et mesurer la discrimination, http://transparence.conf.citi-lab.fr/css/2017_style/img/files/quantifier_mesurer_discriminationSGAMBS.pdf
- [2] S. Kuper et al., Toward Scalable Verification for Safety-Critical Deep Networks, https://arxiv.org/abs/1801.05950
- [3] https://medium.com/@deepmindsafetyresearch/towards-robust-and-verified-ai-specification-testing-robust-training-and-formal-verification-69bd1bc48bda
- [4] W. Samek, K.R. Muller, Tutorial on interpretable machine learning, GCPR'17 conference, heatmapping.org
- [5] B. Kim, F Doshi-Velez, tutorial on interpretable machine learning, ICML 2017
- [6] B. Herman. The promise and peril of human evaluation for model interpretability, arXiv:1711.07414, 2017
- [7] L. H. Gelpin et al. Explaining explanations: an overview of interpretability of machine learning, arXiv:1806.00069v3, 2019

Institut Mines-Télécom

For an Ethical Use of the Internet of Things: favoring fair and unbiased algorithmic decisions while protecting privacy

Sophie Chabridon
SAMOVAR/ACMES

TELECOM SudParis

IP PARIS

Motivations

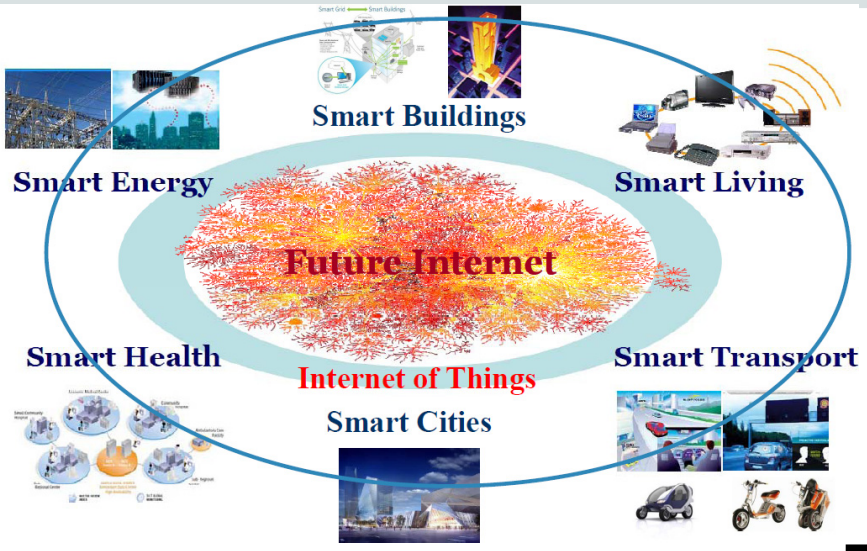
- Learning and decision-making algorithms enter our everyday life
- Concerns on fairness of algorithms, presence of bias, potential discrimination

- Case of the Internet of Things
  - Unprecedent size, dynamicity and ability to collect private data
  - Requires ethics by design

- Data-driven approach
  - Can diversity help to increase fairness?
  - Possible to ensure fairness while protecting privacy?

Motivations

Applications of the Internet of Things



IERC – European Cluster on the IoT

SUMMARY

1. Diversity
2. Privacy protection
3. Discussion

Institut Mines-Télécom

TELECOM SudParis
IP PARIS

2- Diversity

1- Diversity

Notion of diversity as proposed by [Drosou2018]

- Captures the quality of a collection of items, or of a composite item
- Ensures that different kinds of objects are present in the output of an algorithmic process
- Important for both ethical reasons (mitigate risks of exclusion or discrimination against minorities) and utilitarian reasons (to enable more powerful, accurate and engaging data analysis)
- Various measures based on distance, coverage, novelty…
- Helpful in data cleaning, selection, integration, preprocessing, ranking, result interpretation...

1- Diversity

Distance-based Measures

- Rely on a pair-wise distance measure between the elements of set *S*
- Can be defined with average or minimum
- Importance of the choice of distance measure
  - Euclidean, Jaccard, cosine, …

1- Diversity

Coverage-based Measures

- Coverage of a predefined set of aspects like topics, opinions, …
- Often measured probabilistically
- Diversity indicates how well S covers the considered aspects
- Can help counteract the tyranny of majority

1- Diversity

Novelty-based Measures

- Target difference with the past to reduce redundancy
- Select one element at a time (greedy approach)
- Can be combined with utility, popularity/unpopularity, serendipity

1- Diversity

Experiments

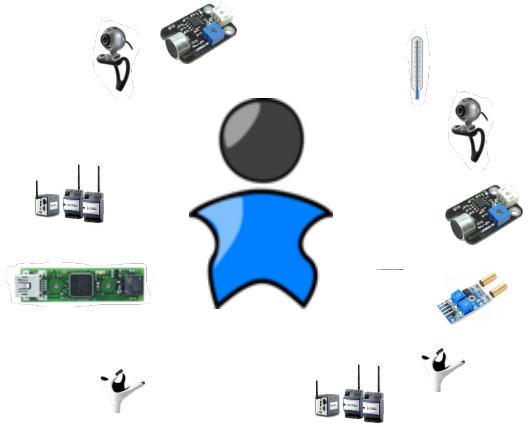
- Does increasing diversity also increase fairness?
- Experiments with 2 public datasets
  - Extract of 2015 Survey on Drug Use and Health [Turner 2017]
  - Census Income dataset (predict income to decide on loan attribution) [Kohavi1996]
- Pre-process dataset with simple distance-based diversity
- Target group fairness measured with p%−rule

Ratio between % of subjects having a certain sensitive attribute value assigned the positive decision and % of subjects not having that value also assigned the positive outcome should be no less than p%

- Work in progress
  - First results show increase in fairness
  - Also good performance on prediction precision

2- Privacy protection

- Impact on the acceptance of the IoT
- The user must be able to control what, why, when, where and how to share information



2- Privacy protection
Ensure non-discrimination

In EU, rights for private data protection (GDPR) and non-discrimination. How to comply with both?

- [Žliobaitė2016] shows contradiction between
  1) Ensure that data-driven decision making is not discriminatory
  2) Ensure minimum collection and storing of private data
- Intuition: Restricting access to sensitive information should prevent discrimination from happening
  - Limit direct discrimination
  - Not sufficient for indirect discrimination: legitimate variables correlated with sensitive characteristics
- Empirical and theoretical proof for linear regression models
- Guaranteeing non-discrimination requires sensitive data
- Sensitive attributes then not used as input for decision making

2- Privacy protection
Manipulate protected sensitive data

- Investigate technical solutions able to reason on sensitive attributes without violating privacy
- [Kilbertus2018]: fairness certification of decision-making algorithms while keeping sensitive attributes encrypted to both the regulator and the system provider
- Zero knowledge model: each party knows nothing except its input and output
- Relies on Secure Multi-party Computation
  - Computation is (i) a procedure to check the fairness of a model and certify it, (ii) a machine learning training procedure with fairness constraints, or (iii) a model evaluation to verify a decision.
  - 2 parties: modeler and regulator

3 - Discussion

3- Discussion

- Multidisciplinary work required for an ethical IoT
  - Distributed systems, cryptography, game theory, social sciences, ...
- Montréal declaration [Montréal2018]
  - Participatory work on guidelines definition for a responsible development of Artificial Intelligence
- EC expert group on Ethics [EC2018]
  - Requirements for Trustworthy AI from the earliest design phase: Accountability, Data Governance, Design for all, Governance of AI Autonomy (Human oversight), Non- Discrimination, Respect for Human, Autonomy, Respect for Privacy, Robustness, Safety, Transparency.

References

[Drosou2018] M. Drosou, H.V. Jagadish, E. Pitoura, and J. Stoyanovich. Diversity in Big Data: A Review. Big Data Review, 5(2), 2018.

[EC2018] Draft Ethics Guidelines for Trustworthy AI, https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

[Kilbertus2018] N. Kilbertus et al. Blind justice: Fairness with encrypted sensitive attributes, 35th Int. Conf. on Machine Learning 2018.

[Kohavi1996] R. Kohavi, B. Becker, Census Income dataset. UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/adult.

[Montréal2018] Montréal declaration for a responsible development of Artificial Intelligence. https://www.montrealdeclaration-responsibleai.com, 2018.

[Tolan2018] S. Tolan. Fair and Unbiased Algorithmic Decision Making: Current State and Future Challenges. JRC Digital Economy Working Paper, 2018.

[Turner2017] B. Turner, Extract of National Survey on Drug Use and Health 2015, https://data.world/balexturner/drug-use-employment-work-absence-income-race-education

[Žliobaitė2016] I. Žliobaitė, B. Custers. Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. Artificial Intelligence and Law, 24(2), 2016.