FAIR AI

2 Jours

Atelier pluridisciplinaire sur la justice sociale, l'équité et les discriminations dans les systèmes algorithmiques

18 et 19 novembre

LISIS Laboratoire Interdisciplinaire Sciences Innovations Sociétés

cnrs

centre internet et societe

LIX

Inria

aivancity

Atelier Jour 1

8h30 - 9h Accueil Café

9h - 9h15 Mots d'accueil
Tawhid Chtioui *(Président Fondateur d'AIVANCITY)*

9h15 - 9h30 Introduction de la journée
Bilel Benbouzid *(Professeur Affilié AIVANCITY, Maître de conférences en sociologie, Centre Internet et Société (Délégation CNRS),* LISIS (UGE)), **Ruta Binkyte-Sadauskiene** *(Doctorante en Équité dans les systèmes algorithmiques, Inria, LIX, École Polytechnique)*

9h30 - 12h Session 1: Méthodes de débiaisage statistique
Jean David Fermanian *(Professeur, Ensae-Crest) Apprentissage équitable avec bagging*
Robin Vogel *(Chercheur Postdoctoral, University of Edinburgh, Visual Computing Group (VICO)) Learning Fair Scoring Functions : Classement bipartite sous contraintes d'équité basées sur le ROC*
Nicolas Shreuder *(Chercheur Postdoctoral, Università di Genova (Italie), CREST, ENSAE***) et Evgenii Chzhen** *(Chargé de recherche, Université Paris-Saclay, CNRS) Classification avec abstention mais sans disparité*
Patrick Loiseau ***(****Chargé de recherches et Professeur chargé de cours, Inria, LIG, École Polytechnique) Sur la sélection équitable en présence de variance implicite et différentielle*
Patrice Bertail *(Professeur, Université Paris-Nanterre) Apprendre à partir de données biaisées : Une approche semi-paramétrique*

12h - 13h30 Pause déjeuner

13h30 - 14h30 Session 2: Méthodes causalité
Jean Michel Loubes (*Professeur, Université Toulouse 3, ANITI Chair Fairness in Machine Learning) Modeles counterfactuel pour des algorithmes plus justes*
Sami Zhioua *(Professeur chercheur, Inria, LIX, École Polytechnique) Le besoin de causalité pour garantir l'équité*

14h30 - 15h Pause café

15h - 17h Session 3: Défis et compromis
Virginie Do *(Doctorante en informatique, Université Paris Dauphine - PSL et Facebook AI Research) Équité bilatérale dans les classements via la dominance de Lorenz*
Cédric Langbort (*Professeur, University of Illinois at Urbana-Champaign (UIUC), coordinated science lab) Autour de la "fairness in the loop"*
Vincent Grari *(Doctorant, Sorbonne Université LIP6/CNRS) L'équité par l'apprentissage contradictoire*
Philippe Besse *(Professeur émérite, Université de Toulouse - INSA, Institut de Mathématiques - UMR CNRS 5219) Quels outils statistiques élémentaires pour répondre aux futures obligations réglementaires européennes sur l'IA?*

Atelier Jour 2

8h30 - 9h Accueil Café

9h - 10h Session 1 Deux cas classiques
Nathan Noiry *(Chercheur postdoctoral, Telecom Paris) L'équité en reconnaissance faciale*
Arthur Charpentier *(Professeur, Université du Québec à Montréal (UQAM)) Assurance, discrimination et équité, quel rôle* pour les actuaires

10h - 12h30 Session 2 Droit
Sarah Benichou *(Adjointe au directeur de la Promotion de l'égalité et de l'accès aux droits, Défenseur des droits)* et **Gaëtan Goldberg** *(Chargé de mission numérique, droits et libertés, Défenseur des droits) Lutte contre les discriminations algorithmiques : retour d'expérience du Défenseur des droits*
Ronan Pons *(Doctorant en droit, Université Toulouse 1 Capitole, Institut 3IA Toulouse (Aniti)) La Fairness algorithmique vue par le droit*
Gianclaudio Malgieri *(Professeur, EDHEC Business School) Justification des systèmes d'IA vers une responsabilité ex ante*
Brendan Berne *(Associé, Professeur de droit, Spencer West LLP, Rennes School of Business) La proposition de règlement européen sur l'IA : Un modèle pour un régime de réglementation éthique de l'IA ?*

12h30 - 13h30 Pause déjeuner

13h30 - 14h30 Session 3 Observation des Pratiques
Emilie Sirvent-Hien *(Responsable programme de recherche IA Responsable, Orange Innovation)*, **Christèle Tarnec** *(Ingénieur de Recherche en Éthique de l'IA, Orange Labs) Gouvernance IA et gestion des biais en entreprise : Pratiques et Limites*
Mariame Tighanimine *(Doctorante en sociologie du travail, LISE (CNAM-CNRS)) "La justice sociale part de motivations rationnelles. C'est l'irrationalité qui créée de l'injustice." Comment un groupe de chercheurs tente de l'intérieur de rendre l'intelligence artificielle "robustement bénéfique"*

14h30 - 15h30 Session Normativité et Réflexivité
Marc Fleurbaey *(Directeur de recherches, Paris School of Economics, CNRS) Le point de vue de l'économie normative*
Dominique Cardon *(Professeur, Sciences Po, Médialab)*, **Jean-Marie John-Mathews** *(Professeur Affilié AIVANCITY, Doctorant en interprétabilité et équité en ML, Université Paris-Saclay) De la réalité au monde. Une perspective critique sur l'équité en matière d'IA*

15h30 - 16h **Pause café**

16h - 17h Synthèse critique
Xavier Delaporte *(France Inter / L'Obs)*, ***Karine Gentelet*** *(Professeure, Université du Québec en Outaouais, titulaire de la Chaire Abéona-ENS-OBVIA)*, ***Hubert Guillaud*** (InternetActu.net)

À PROPOS

Cet atelier pluridisciplinaire organisé par le Centre Internet et Société (CIS), en partenariat avec le Laboratoire interdisciplinaire Sciences Innovations Sociétés (LISIS), vise à rassembler la communauté scientifique française autour de la recherche sur la justice sociale, l'équité et les discriminations dans les systèmes algorithmiques, domaine appelé aussi *Fairness in Machine Learning (FairML)*. La constitution de ce groupe rejoint les recommandations du Défenseur des Droits d'orienter la recherche vers le développement de connaissances plus adaptées au cadre juridique et politique français. Alors que la justice sociale se décide de manière contextuelle et pragmatique, la science dite de la « *fairness* dans le *machine learning* » reste cadrée par des repères politiques et juridiques anglo-saxons, en particulier nord-américains. C'est pourquoi une part importante du débat scientifique n'est pas directement transférable en Europe, en particulier en France, et ce pour au moins trois raisons.

Premièrement, le compromis, inversement proportionnel, dans le traitement des données, entre vie privée et équité pose des problèmes spécifiques en France. L'exemple le plus significatif, mais il y en a sans doute d'autres, est celui des variables ethniques. Essentielles à l'atténuation des biais algorithmiques de discrimination selon l'origine, celles-ci ne sont pas autorisées dans la collecte de données (sauf cas exceptionnel). Il faut chercher des méthodes plus adaptées pour détecter et atténuer ce type de discrimination algorithmique. Deuxièmement, les loi anti-discriminations s'appliquent de manière différente en Europe et aux États-Unis, notamment autour de la notion de discrimination indirecte. L'une des différences est l'absence, dans le droit communautaire, de seuils statistiques précis pour évaluer l'impact disproportionné de pratiques apparemment neutres, ce qui implique un raisonnement différent, déjà amorcé, sur la légalité des systèmes algorithmiques en Europe, et en France en particulier. Enfin, la troisième différence tient à la structuration de la recherche. Les jeux de « données-jouets » autour de la *fairness,* mises en commun par la communauté scientifique pour favoriser le dialogue et les collaborations, sont d'un intérêt limité pour débattre des systèmes algorithmiques utilisés en France. Le jeu de données Compas construits par le journal d'investigation Propublica, qui a permis de tester les différentes situations typiques d'indécidabilité en matière de justice prédictive, illustre un cas d'usage d'algorithme de prédiction de la récidive qui ne trouve aucune équivalence en France. Pourtant une bonne part du débat technique se structure autour de ce jeu de données. Ceci est bien connu de la littérature en sociologie des sciences, les *datasets* partagés par les chercheurs ne sont pas neutres. Ils produisent des effets structurants, souvent de manière invisible, sur la production des connaissances.

Ce sont ces limites qui nous invitent à proposer un atelier dans la perspective de favoriser un débat plus contributif aux spécificités françaises, tout en reprenant les axes thématiques majeurs de la recherche dans ce domaine: Économie normative et sens politique des métriques, Les perceptions de l'équité algorithmique, Méthodes pour construire des systèmes algorithmiques qui intègrent les contraintes d'équité, Les effets de long terme des contraintes d'équité sur les systèmes algorithmiques, Les instruments et pratique de l'audit en matière d'équité algorithmique, L'importance d'une réflexivité critique.

LES RÉSUMÉS

1

Apprentissage équitable avec bagging

Jean David Fermanian
Ensae-Crest

La question centrale de cet article est de savoir comment améliorer les algorithmes d'apprentissage supervisé avec une exigence d'équité garantissant que toute entrée sensible n'influence pas "injustement" le résultat de l'algorithme d'apprentissage. Pour atteindre cet objectif, nous procédons en trois étapes. D'abord, après avoir introduit plusieurs notions d'équité dans une approche uniforme, nous introduisons une notion plus générale à travers la définition de l'équité conditionnelle qui englobe la plupart des définitions d'équité bien connues. Ensuite, nous utilisons un ensemble de classificateurs binaires et continus pour obtenir une solution optimale pour un résultat prédictif équitable en utilisant une procédure de post-traitement connexe sans aucune transformation sur les données, ni sur les algorithmes d'apprentissage. Enfin, nous introduisons plusieurs critères pour mesurer l'équité de nos prédictions. Quelques données empiriques sont fournies pour illustrer notre approche.

Learning Fair Scoring Functions : Classement bipartite sous contraintes d'équité basées sur le ROC

Robin Vogel
University of Edinburgh, Visual Computing Group (VICO)

De nombreuses applications de l'IA impliquent de noter des individus en utilisant une fonction apprise de leurs attributs. Ces scores de risque prédictifs sont ensuite utilisés pour prendre des décisions selon que le score dépasse un certain seuil, qui peut varier selon le contexte. Le niveau de délégation accordé à de tels systèmes dans des applications critiques comme le crédit et le diagnostic médical dépendra fortement de la manière dont on pourra répondre aux questions d'équité. Dans cet article, nous étudions l'équité pour le problème de l'apprentissage de fonctions de notation à partir de données binaires étiquetées, une tâche d'apprentissage classique connue sous le nom de classement bipartite. Nous soutenons que la nature fonctionnelle de la courbe ROC, la mesure de référence de la précision du classement dans ce contexte, conduit à plusieurs façons de formuler les contraintes d'équité. Nous introduisons des familles générales de définitions d'équité basées sur l'AUC et sur les courbes ROC, et nous montrons que nos contraintes basées sur la courbe ROC peuvent être instanciées de telle sorte que les classificateurs obtenus par seuillage de la fonction de notation satisfont l'équité de classification pour une gamme désirée de seuils. Nous établissons des limites de généralisation pour les fonctions de notation apprises sous de telles contraintes, concevons des algorithmes d'apprentissage pratiques et démontrons la pertinence de notre approche par des expériences numériques sur des données réelles et synthétiques.

Classification avec abstention mais sans disparité

Nicolas Shreuder et Evgenii Chzhen
Université Paris-Saclay, CNRS, Università di Genova (Italie), CREST, ENSAE Paris

La classification avec abstention a gagné beaucoup d'attention ces dernières années car elle permet d'incorporer des décideurs humains dans le processus. Pourtant, l'abstention peut potentiellement amplifier les disparités et conduire à des prédictions discriminatoires. L'objectif de ce travail est de construire un algorithme de classification à usage général, qui est capable de s'abstenir de prédire, tout en évitant l'impact disparate. Nous formalisons ce problème comme une minimisation du risque sous des contraintes d'équité et d'abstention pour lesquelles nous dérivons la forme du classificateur optimal. Sur la base de ce résultat, nous proposons un algorithme de classification post-traitement, qui est capable de modifier n'importe quel classificateur basé sur le score en utilisant uniquement un échantillon non étiqueté. Nous établissons des garanties de risque d'échantillon fini, d'équité et d'abstention pour l'algorithme proposé. En particulier, nous montrons que les contraintes d'équité et d'abstention peuvent être atteintes indépendamment du classificateur initial tant qu'un nombre suffisant de données non étiquetées est disponible. La garantie de risque est établie en fonction de la qualité du classificateur initial. Notre schéma de post-traitement se réduit à un programme linéaire clairsemé permettant une implémentation efficace, que nous fournissons. Enfin, nous validons notre méthode de manière empirique en montrant que des taux d'abstention modérés permettent de contourner le compromis risque-équité.

Sur la sélection équitable en présence de variance implicite et différentielle

Patrick Loiseau
Inria, LIG, École Polytechnique

La discrimination dans les problèmes de sélection tels que l'embauche ou l'admission à l'université est souvent expliquée par un biais implicite du décideur contre les groupes démographiques défavorisés. Dans ce travail, nous considérons plutôt un modèle dans lequel le décideur reçoit une estimation bruyante de la qualité de chaque candidat, dont la variance dépend du groupe du candidat. Nous soutenons que cette variance différentielle est une caractéristique clé de nombreux problèmes de sélection. Ensuite, nous analysons la discrimination et le compromis équité-utilité potentiel qui découlent de cette variance différentielle.Plus précisément, nous considérons deux situations : dans la première, les variances du bruit sont inconnues du décideur qui choisit simplement les candidats ayant la meilleure qualité estimée indépendamment de leur groupe ; dans la seconde, les variances sont connues et le décideur choisit les candidats ayant la meilleure qualité attendue compte tenu de l'estimation bruyante. Nous montrons que les deux décideurs de base produisent de la discrimination. Nous étudions l'effet sur l'utilité de la sélection de l'imposition d'un mécanisme d'équité que nous appelons la règle Γ (il s'agit d'une extension de la règle classique des quatre cinquièmes et elle inclut également la parité démographique). Dans le premier cas (avec des variances inconnues), nous prouvons que, sous certaines conditions, l'imposition de la règle Γ augmente l'utilité de la sélection - il n'y a pas de compromis entre l'équité et l'utilité. Dans le deuxième cas (avec des variances connues), l'imposition de la règle Γ diminue l'utilité mais nous prouvons une limite sur la perte d'utilité due au mécanisme d'équité.

Apprendre à partir de données biaisées : Une approche semi-paramétrique

Patrice Bertail, S. Clémencon, Y. Guyonvarch and N. Noiry
Université Paris-Nanterre, Telecom Paris

Dans cet article, nous considérons des problèmes de minimisation du risque où la distribution (source) des observations d'apprentissage diffère de la distribution (cible) impliquée dans le risque que l'on cherche à minimiser mais est toujours définie sur le même espace mesurable . Notre objectif est de développer un cadre semi-paramétrique inspiré des idées d'économétrie et d'échantillonnage d'enquête, pour un tel problème spécifique d'apprentissage par transfert, lorsque des informations auxiliaires sur la population statistique cible sont disponibles, sous la forme d'attentes de fonctions connues. Sous l'hypothèse que la dérivée de Radon-Nikodym entre les deux distributions différentes (source et cible) appartient à une classe paramétrique, combinée à des conditions d'identifiabilité appropriées, nous montrons qu'un problème de minimisation du risque empirique (ERM) pondéré peut être formulé avec des poids (aléatoires). L'idée est simplement de trouver un estimateur des paramètres du modèle paramétrique qui correspond à l'information auxiliaire et d'utiliser le modèle pour repondérer les observations avec les paramètres estimés. Nous établissons une borne de généralisation prouvant que la solution au problème ERM pondéré ainsi construit atteint un taux d'apprentissage du même ordre que celui atteint en l'absence de tout biais d'échantillonnage. Au-delà de ces garanties théoriques, nous présentons des résultats numériques fournissant des preuves empiriques solides de la pertinence de l'approche.

Modeles counterfactuel pour des algorithmes plus justes

Jean-Michel Loubes
Université Toulouse 3, ANITI Chair Fairness in Machine Learning

Un algorithme "juste" ou "fair" doit pouvoir traiter de la même manière des individus similaires. Mais lorsque les individus appartiennent à des catégories différentes, cette notion de similarité devient difficile à établir : comment ,par exemple, comparer de manière équitable la taille d'un homme avec une femme ? Nous essaierons de fournir une réponse à l'aide d'explications "counterfactuelles" permettant ainsi de découvrir des mondes alternatifs où il devient possible de comparer des individus appartenant à des groupes différents.

Le besoin de causalité pour garantir l'équité

Sami Zhioua, Karima Maklouf, Carlos Pinzón, Castucia Palamidessi
Inria, COMETE, LIX, École Polytechnique

Plusieurs notions d'équité ont été définies et examinées au cours de la dernière décennie, telles que la parité statistique (Statistical Parity) et les quotas égalisées (Equalized Odds). Les notions d'équité les plus récentes, cependant, sont fondées sur la causalité et reflètent l'idée désormais largement acceptée selon laquelle l'utilisation de la causalité est nécessaire pour traiter de manière appropriée le problème de l'équité. Comme la majorité des notions d'équité causales sont définies en termes de quantités non observables (par exemple, les interventions et les contrefactuels), leur applicabilité dépend fortement de l'identifiabilité de ces quantités à partir des données observables. Dans cette intervention, nous illustrons ces deux subtilités (le besoin de causalité et le problème d'identifiabilité) à travers deux exemples concrèts. Nous pensons que cette illustration est nécessaire pour bien comprendre le role important que la causalité peut jouer dans le domaine de l'équité des systèmes de prise de décision automatisé.

Équité bilatérale dans les classements via la dominance de Lorenz

Virginie Do, Sam Corbett-Davies, Jamal Atif et Nicolas Usunier
Université Paris Dauphine - PSL et Facebook AI Research

Nous proposons une nouvelle approche pour déduire des listes classées de recommandations qui sont équitables à la fois envers les consommateurs du classement ("utilisateurs") et les "éléments" classés. Nous considérons que les recommandations sont équitables lorsqu'elles satisfont à l'efficacité (il n'est pas possible d'améliorer le sort d'une personne sans détériorer celui d'une autre) et à l'équité (elles sont maximalement redistributives à un niveau donné d'utilité globale). Ces deux conditions sont satisfaites lorsqu'une recommandation produit des profils d'utilité pour les utilisateurs et les éléments avec des courbes de Lorenz généralisées non dominées.
Nous y parvenons en trouvant des classements stochastiques qui maximisent une fonction de bien-être concave des profils d'utilité. La courbure de la fonction de bien-être et le poids relatif des utilisateurs et des articles contrôlent le degré de redistribution entre les utilités des utilisateurs et des articles. Par rapport aux algorithmes existants qui maximisent un compromis entre l'utilité de l'utilisateur et l'inégalité de l'utilité de l'article, nous fournissons un soutien théorique et expérimental que notre approche réalise des compromis supérieurs sur les problèmes avec des préférences bilatérales. Alors que les approches standard peuvent détruire l'utilité au lieu de la redistribuer, notre méthode redistribue efficacement l'utilité entre les utilisateurs et les articles.

Autour de la "fairness in the loup”

Cédric Langbort
University of Illinois at Urbana-Champaign (UIUC),
coordinated science lab

Depuis l'apparition récente de la notion de "fairness" comme un critère à part entière dans le développement d'algorithmes, nombre de travaux s'intéressent à la création de procédures satisfaisant un ou plusieurs critère(s) d'équité donné(s). En parallèle à ces travaux, une portion minoritaire mais croissante de la littérature a montré que, dans des situations comportant plusieurs étapes et/ou acteurs stratégiques (i.e., des situations où des 'boucles' d'interdépendances sont présentes), imposer localement une notion d'équité ne permet pas forcement de garantir un procédé "juste" aux yeux de tous et/ou a priori. Deux difficultés, au moins, se posent. D'une part, la structure d'une population, qui conditionne le critère d'équité à un instant/pour une opération donné(e), dépend des opérations, justes ou non, effectuées en amont: une séquence d'opérations "justes" indépendamment les unes des autres, n'est pas forcément "juste" par elle même, a priori, si les critères d'équité de la séquence ne sont pas convenablement calibrés. D'autre part, quand la structure de la population n'est pas donnée a priori, mais dépend des choix offerts aux individus, une unique opération "juste" peut tout de même résulter en des traitements différents. Dans cette présentation, je donnerai des exemples simples de ces situations, montrerai qu'elles sont liées conceptuellement, et discuterai une notion d'équité permettant d'appréhender de tels mécanismes en théorie. Quelques autres pistes connexes seront aussi discutées.

L'équité par l'apprentissage contradictoire

Vincent Grari
Sorbonne Université LIP6/CNRS

Ces dernières années, on a assisté à une augmentation spectaculaire de l'intérêt académique et sociétal pour l'apprentissage automatique équitable. Des travaux importants ont été réalisés sur la méthode d'équité de groupe pour imposer l'égalité des distributions sensibles générales. Pour atteindre cet objectif, de nombreuses approches contradictoires ont été introduites et ont montré des résultats très compétitifs sur des ensembles de données du monde réel.

Cependant, la plupart des stratégies d'équité dans les modèles d'apprentissage automatique se concentrent sur l'hypothèse que l'information sensible est observée. Or, cela n'est pas toujours possible en pratique. En raison de la protection de la vie privée et de diverses réglementations telles que le RGPD dans l'UE, de nombreux attributs personnels sensibles ne sont souvent pas collectés. Nous constatons un manque d'approches pour atténuer les biais dans de tels contextes difficiles, en particulier pour atteindre les objectifs classiques d'équité tels que la parité démographique et les chances égalisées. Nous discuterons des stratégies les plus récentes pour atteindre cet objectif.

Quels outils statistiques élémentaires pour répondre aux futures obligations réglementaires européennes sur l'IA?

Philippe Besse
Université de Toulouse - INSA, Institut de Mathématiques - UMR CNRS 5219

Suite à la publication du livre blanc pour une approche de l'IA basée sur l'excellence et la confiance, la Commission Européenne (CE) a publié de nombreuses propositions de textes réglementaires dont un *Artificial Intelligence Act (AI Act)* (2021) établissant des règles harmonisées sur l'intelligence artificielle (IA). Quels seront les conséquences et impacts de l'adoption à venir de ce texte du point de vue d'un mathématicien ou plutôt statisticien impliqué dans la conception de système d'intelligence artificielle (IA) à haut risque au sens de la CE ? Quels outils et méthodes vont permettre de répondre à l'obligation d'une analyse rigoureuse et documentée des données traitées, des performances, robustesse, résilience de l'algorithme, de son explicabilité, des risques, pour les droits fondamentaux, de biais discriminatoires ? Ces questions sont illustrées par un exemple numérique analogue à un score de crédit à la recherche d'un moins mauvais compromis entre toutes les contraintes. Nous concluons sur les avancées et limites de ce projet de règlement pour les systèmes d'IA à haut risque.

LES RÉSUMÉS

L'équité en reconnaissance faciale

Nathan Noiry
Telecom Paris

Depuis l'avènement des techniques d'apprentissage profond au début des années 2010, les systèmes de reconnaissance faciale ont atteint des niveaux de performance impressionnants. Malgré leur fiabilité biométrique, certains de ces systèmes sont biaisés et commettent statistiquement plus d'erreurs sur des sous-groupes spécifiques de la population (typiquement les afro-américains et les femmes). Dans cet exposé, je commencerai par rappeler les enjeux techniques de la reconnaissance faciale et discuterai des problèmes d'équité qu'ils impliquent. Je présenterai ensuite une solution de correction des biais de genre récemment développée.

Assurance, discrimination et équité, quel rôle pour les actuaires

Arthur Charpentier
Université du Québec à Montréal (UQAM)

En assurance "discrimination is not necessarily bad, equality not necessarily good" avait dit Bailey *et al.* (1970). L'équité actuarielle consiste à dire que la contribution (prime ou cotisation) juste doit être liée au risque associé. Or la r/glementation interdit certaines discrimination. Nous reviendrons ici sur les difficultés qui se posent aux actuaires dans un contexte de données massives, et insistant sur l'approche causale de la discrimination et de l'équité (en assurance, établir une relation causale toute directement à la prévention).

Lutte contre les discriminations algorithmiques : retour d'expérience du Défenseur des droits

Sarah Benichou et Gaëtan Goldberg
Défenseur des droits

La lutte contre les discriminations s'inscrit parmi les grands domaines de compétence du Défenseur des droits. Lorsqu'elle est le fruit de dispositifs algorithmiques, la discrimination est particulièrement difficile à identifier et à caractériser d'un point de vue juridique. A-travers une présentation de cas d'espèces et de jurisprudences étrangères, l'institution souhaite présenter ses recommandations pour prévenir les biais discriminatoires algorithmiques à la lumière notamment de la proposition de règlement sur l'IA de la Commission européenne.

2

La *Fairness* algorithmique vue par le droit

Ronan Pons
Université Toulouse 1 Capitole, Institut 3IA Toulouse (Aniti)

Depuis plusieurs années, le mouvement d'*algorithmic fairness* voit sa popularité grandir au sein des sciences sociales. Le droit ne fait pas exception à cette tendance mais plusieurs questions doivent être adressées avant que la discipline puisse s'approprier pleinement ce phénomène. Si la traduction de *fairness* en « équité » semble la plus parlante, elle n'est pas forcément la plus pertinente juridiquement. Il faut d'abord identifier les notions existantes en droit qui font écho à celle de *fairness*. Si ce lien est difficile à établir sous plusieurs aspects, le jeu en vaut la chandelle. En effet avec cette notion vient plusieurs outils. Des métriques qui pourraient s'avérer très intéressantes pour le droit, en particulier depuis l'élaboration par la Commission européenne d'une proposition de règlement européen sur les systèmes d'IA qui repose sur une logique de certification.

Justification des systèmes d'IA vers une responsabilité ex ante

Gianclaudio Malgieri
EDHEC Business School

L'équité dans l'IA n'est pas un objectif facile à atteindre, mais
il y a au moins trois étapes qui devraient être considérées systématiquement pour atteindre cet objectif et des recherches supplémentaires pourraient aider sur ces points. Premièrement, au niveau réglementaire, un cadre de "justification" (plutôt qu'une simple explication) pourrait être utile. Les utilisateurs et les développeurs d'IA devraient être invités à justifier pourquoi l'algorithme n'est "pas" discriminatoire, manipulateur, stigmatisant, etc. Cela serait possible grâce à un modèle de licence ex ante, partiellement approuvé par la proposition de loi européenne sur l'IA. Deuxièmement, au niveau des entreprises, deux outils pourraient être utiles : a) une analyse d'impact de l'IA (sur l'exemple du DPIA du GDPR et en s'inspirant également d'expériences étrangères, comme le modèle canadien "Algorithmic Impact Assessment"). Dans cette évaluation d'impact, les développeurs et les utilisateurs de l'IA devraient vérifier les éventuels préjudices ou autres utilisations abusives de leur IA dans leur contexte spécifique, analyser ces risques et proposer des mesures d'atténuation ; b) une conception participative devrait également être mise en place, en incluant des représentants des communautés touchées et des parties prenantes vulnérables et en veillant à ce qu'ils puissent exprimer leur point de vue dans le processus de conception.

2

La proposition de règlement européen sur l'IA : Un modèle pour un régime de réglementation éthique de l'IA ?

Brendan Berne
Spencer West LLP, Rennes School of Business

Dans le cadre de sa "stratégie européenne en matière de données", et conjointement avec son règlement sur la gouvernance des données ouvertes et sa directive sur les données ouvertes, l'Union européenne a proposé une pièce maîtresse du cadre réglementaire qu'elle est en train de mettre en place afin de favoriser l'adoption de l'IA sur son territoire et de faire face aux risques associés à certaines utilisations de cette technologie. This proposed regulation is intended to encourage the development of more research activities within its borders and establish a model set of rules for export to the rest of the world in its AI Regulation.
My presentation will lay out the EU vision of the future regulatory landscape and the mechanisms which the EU institutions intends to implement to ensure an ethical yet robust AI technology industry in the European Union.individuelle dans la conception des systèmes algorithmiques ? Ce que nous souhaitons montrer, c'est que l'enjeu de la recherche sur la justice sociale dans les machines prédictives réside dans la complexité de faire reposer les contraintes d'équité sur des grandeurs complexes tels que des "ensembles d'opportunités".

Gouvernance IA et gestion des biais en entreprise : Pratiques et Limites

Emilie Sirvent-Hien, Christèle Tarnec, Françoise Fessant, Thierry Nagellen
Orange Innovation

Le risque de biais algorithmique est reconnu et constitue un enjeu important pour les organisations au moment du déploiement massif de l'IA. La gouvernance s'organise avec 2 mouvements complémentaires Top-Down (des principes posés par le comex à l'application) et Bottom-up (des projets vers leurs responsables éthiques). Pour autant une fois le problème posé, les solutions ne sont pas toujours évidentes.

Nous présenterons à la fois la tendance en cours dans les entreprises pour organiser la gouvernance de l'IA et la gestion des biais (au moyen de grille d'auto-évaluation, d'outils de gestion de biais) et quelques limites que nous explorons en recherche (comme le manque de données et le recours au transfer learning, les réseaux IOT et l'équité distribuée).

2

“La justice sociale part de motivations rationnelles. C’est l’irrationalité qui créée de l’injustice.” Comment un groupe de chercheurs tente de l'intérieur de rendre l’intelligence artificielle “robustement bénéfique”

Mariame Tighanimine
LISE (CNAM-CNRS)

Cette intervention propose de présenter la structuration en cours d’un groupe de chercheurs en informatique et en mathématiques, agissant au sein et à la marge de l’arène professionnelle, à l’origine de réflexions et de productions scientifiques aussi bien théoriques qu’appliquées sur les questions de robustesse, de sécurité et finalement d’éthique des intelligences artificielles. Ils sont en réalité porteurs d’un projet ambitieux de refonte épistémologique de l’informatique qu’ils considèrent comme une “philosophie morale calculable” (Hoang & El Mhamdi, 2019) dont le développement est soumis à une *“deadline”* (Boström, 2014). Ainsi, répondre (techniquement) aux questions philosophiques posées par les outils informatiques et de calcul devient urgent. Chaque requête soumise à un algorithme est potentiellement considérée comme un dilemme moral que l’algorithme doit résoudre dans un temps contraint.
À partir d’un matériau de recherche constitué d’entretiens semi-directs, d’une observation participante en cours (commencée en mars 2020) et de l’analyse de données issues de ce terrain, nous verrons comment ce groupe auto-organisé tente de porter une critique “rationnelle” de l’intérieur (sur la nature des données, la non prise de conscience des enjeux épistémiques et sociaux de la recherche en optimisation...), et analyserons quelques réceptions par leurs pairs. Par ailleurs, nous nous demanderons dans quelle mesure ce collectif de chercheurs tend vers l’idéal durkheimien du groupe professionnel régulant l’activité professionnelle déstabilisatrice de l’ordre social, et au sein duquel doivent se constituer la morale et le droit professionnels.

Le point de vue de l’économie normative

Marc Fleurbaey
Paris School of Economics, CNRS

Les recherches sur l’équité dans les systèmes algorithmiques tendent à être envisagées comme une forme d’application de “ l’économie normative”. Dans cette présentation, je souhaite revenir sur cette tendance en discutant quelques travaux en sciences computationnelles qui mobilisent les développements marquants de l’économie normative inspirée par les philosophies de Rawls, Dworkin et Sen. Comment ces travaux en sciences computationnelles intègrent-ils la liberté et la responsabilité individuelle dans la conception des systèmes algorithmiques ? Ce que nous souhaitons montrer, c’est que l’enjeu de la recherche sur la justice sociale dans les machines prédictives réside dans la complexité de faire reposer les contraintes d'équité sur des grandeurs complexes tels que des « ensembles d’opportunités”.

2

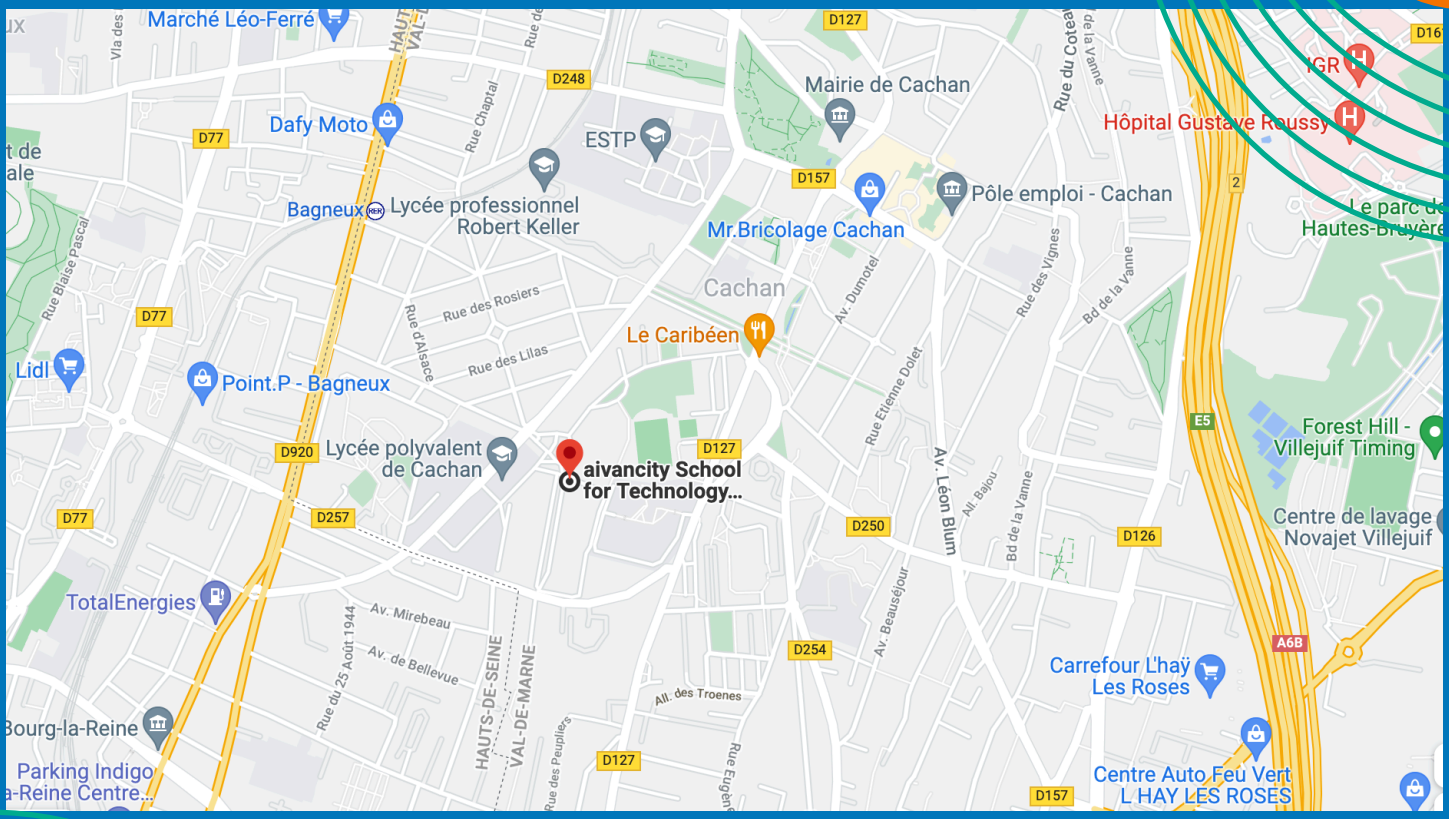
De la réalité au monde. Une perspective critique sur l'équité en matière d'IA

Dominique Cardon et Jean Marie John Mattews
Sciences Po, Médialab, Université Paris-Saclay

L'équité des décisions en matière d'intelligence artificielle (IA) est devenue un grand défi pour les gouvernements, les entreprises et les sociétés. Nous proposons une contribution théorique pour considérer l'éthique de l'IA en dehors des approches de haut niveau et descendantes, basée sur la distinction entre "réalité" et "monde" de Boltanski. Pour ce faire, nous théorisons le débat sur l'équité de l'IA et montrons que la critique de l'injustice du ML est « réaliste », c'est-à-dire ancrée dans une réalité déjà instituée fondée sur des catégories démographiques produites par les institutions. Deuxièmement, nous montrons que les limites des corrections d'équité « réalistes » conduisent à l'élaboration de « réponses radicales » à l'équité, c'est-à-dire des réponses qui changent radicalement le format des données. Troisièmement, nous montrons que la correction de l'équité évolue vers un « régime de domination » qui absorbe les critiques, et nous proposons quelques pistes théoriques et pratiques pour le développement ultérieur de l'éthique de l'IA. En utilisant un espace critique ad-hoc stabilisé par des tests de réalité aux côtés de l'algorithme, nous construisons un modèle de responsabilité partagée compatible avec la réponse radicale aux questions d'équité. Enfin, cet article montre l'apport fondamental des théories de la sociologie pragmatique, dans la mesure où elles offrent une perspective sociale et politique sur l'éthique de l'IA en donnant un rôle actif aux acteurs matériels tels que les formats de bases de données sur les débats éthiques. Dans un contexte où les données sont de plus en plus nombreuses, granulaires et comportementales, il est essentiel de renouveler notre conception de l'éthique de l'IA sur les algorithmes afin d'établir de nouveaux modèles de responsabilité qui prennent en compte les changements de paradigme informatique.

aivancity

Adresse: 57 Av. du Président Wilson, 94230 Cachan
(Entrez par le numéro 61, École Normale Supérieure de Cachan)



cnrs centre internet et societe

LLX

LISIS
Laboratoire Interdisciplinaire Sciences Innovations Sociétés

inria