De la promesse théorique aux paradoxes pratiques

Repenser l'équité algorithmique pour la prise de décision

Présenté par

Olivier Côté

Basé sur la lecture

Measure and Mismeasure of Fairness

Corbett-Davies et al. (2023)

Présentation finale

MAT998P — Équité et discrimination des modèles prédictifs

17 avril 2024

Des recommandations qui misent sur les bénéfices

Autorité des marchés financiers (2024)

« La segmentation des consommateurs [...] permet aux intervenants financiers d'offrir des produits et des services davantage **appropriés à la situation de chacun des groupes de consommateurs**

»

Des recommandations qui misent sur les bénéfices

Autorité des marchés financiers (2024)

« La segmentation des consommateurs [...] permet aux intervenants financiers d'offrir des produits et des services davantage **appropriés à la situation de chacun des groupes de consommateurs** [...] Toutefois, les **bénéfices** ne peuvent être atteints que par l'utilisation de données et de modèles ne présentant pas [...] de biais discriminatoires. »

La théorie peut éloigner du vrai problème

Selon Selbst et al. (2019), les concepts sociaux comme l'équité sont très contextuels et nuancés. On ne peut pas tout résoudre à travers des formalismes mathématiques (*Formalism trap*).

La théorie peut éloigner du vrai problème

Selon Selbst et al. (2019), les concepts sociaux comme l'équité sont très contextuels et nuancés. On ne peut pas tout résoudre à travers des formalismes mathématiques (*Formalism trap*).

Pour mieux aborder l'équité, Corbett-Davies et al. (2023) suggère l'approche **conséquentialiste**. Cette approche privilégie les impacts réels des décisions algorithmiques sur les individus, mettant l'accent sur le bien-être global plutôt que sur les définitions formelles.

Groupes protégés et décisions

Corbett-Davies et al. (2023) voient deux angles possibles pour l'étude de l'équité

- Effet des décisions sur les **groupes protégés** ;
- Effet des **groupes protégés** sur les décisions.

Groupes protégés et décisions

Corbett-Davies et al. (2023) voient deux angles possibles pour l'étude de l'équité

- Effet des décisions sur les **groupes protégés** ;
  - Différentes décisions peuvent entraîner des résultats variés pour différents groupes de personnes.

- Effet des **groupes protégés** sur les décisions.
  - Définitions visant à réduire l'impact direct et indirect de l'appartenance à un groupe peut avoir un effet, autant direct qu'indirect, sur les décisions.

Groupes protégés et décisions

Corbett-Davies et al. (2023) voient deux angles possibles pour l'étude de l'équité

- Effet des décisions sur les **groupes protégés** ;
  - Différentes décisions peuvent entraîner des résultats variés pour différents groupes de personnes.
  - **Équité de groupe**
- Effet des **groupes protégés** sur les décisions.
  - Définitions visant à réduire l'impact direct et indirect de l'appartenance à un groupe peut avoir un effet, autant direct qu'indirect, sur les décisions.
  - **Discrimination* directe et indirecte**

Groupes protégés et décisions

Corbett-Davies et al. (2023) voient deux angles possibles pour l'étude de l'équité

- Effet des décisions sur les **groupes protégés** ;
  - Différentes décisions peuvent entraîner des résultats variés pour différents groupes de personnes.
  - **Équité de groupe** ou **Impact disparate**
- Effet des **groupes protégés** sur les décisions.
  - Définitions visant à réduire l'impact direct et indirect de l'appartenance à un groupe peut avoir un effet, autant direct qu'indirect, sur les décisions.
  - **Discrimination* directe et indirecte** ou **Traitement disparate***

Groupes protégés et décisions

Corbett-Davies et al. (2023) voient deux angles possibles pour l'étude de l'équité

- Effet des décisions sur les **groupes protégés** ;
  - Différentes décisions peuvent entraîner des résultats variés pour différents groupes de personnes.
  - **Équité de groupe** ou **Impact disparate**
- Effet des **groupes protégés** sur les décisions.
  - Définitions visant à réduire l'impact direct et indirect de l'appartenance à un groupe peut avoir un effet, autant direct qu'indirect, sur les décisions.
  - **Discrimination* directe et indirecte** ou **Traitement disparate***

Les deux approches sont fondamentalement différentes, mais visent toutes à ce que les décisions ne soient **pas reliés aux groupes protégés**.

Définitions d'équité

Notation

Description	Notation	Généré par	Domaine	Ex. Diabète	Ex. Admission
Variables utilisées	X	$\mathcal{D}_X$	$\mathbb{R}^n$	Age, BMI	Résultats
Variables protégées	A	$\alpha(X)$	$\mathcal{A}$	Race	Genre
Décision binaire	$D = d(x)$	$\mathrm{P}(D = 1 \mid X = x)$	$\{0, 1\}$	Tester	Admettre
Budget décisionnel	b	—	$(0, 1)$	$b = 1$	$b < 1$
Résultat d'intérêt	Y	$Y(0)$ et $Y(1)$	$\{0, 1\}$	Diabète	Diplomation

Les règles de décisions faisables sont celles pour lesquelles $E(D) \leq b$.

Critères pour étudier les disparités

Limiter les **effets sur les disparités** implique toujours qu'une certaine indépendance soit respectée.

Nom du critère	Expression
Demographic parity	$D \perp\!\!\!\perp A$
Equalized false positive rates	$D \perp\!\!\!\perp A \vert Y = 0$
Counterfactual predictive parity	$Y(1) \perp\!\!\!\perp A \vert D = 0$
Counterfactual equalized odds	$D \perp\!\!\!\perp A \vert Y(1)$
Conditional principal fairness	$D \perp\!\!\!\perp A \vert Y(0), Y(1), X$

Critères pour l'effet sur les décisions

Une approche alternative pour considérer l'équité est d'étudier l'effet des groupes protégés sur les décisions.

Critères pour l'effet sur les décisions

Une approche alternative pour considérer l'équité est d'étudier l'effet des groupes protégés sur les décisions.

Cette approche est cohérente avec la notion de **traitement disparate** qu'on retrouve dans les différentes lois anti-discrimination.

Limiter l'effet sur les décisions

Limiter les **effets sur les décisions** peut se faire individuellement, ou en espérance.

Nom du critère	Expression	Possibilités	Mot clé relié
Blinding	$d(x,a) = d(x,a')$	$\forall x \in \mathcal{X}$ et $a, a' \in \mathcal{A}$	Discrimination directe
Counterfactual fairness	$E[D(a')\|X] = E[D\|X]$	$\forall a, a' \in \mathcal{A}$	Discrimination indirecte
Π-*fairness*	$E[D_{\Pi,A,a'}\|X] = E[D\|X]$	$\forall a, a' \in \mathcal{A}$	Discrimination indirecte

Limiter l'effet sur les décisions

Limiter les **effets sur les décisions** peut se faire individuellement, ou en espérance.

Nom du critère	Expression	Possibilités	Mot clé relié
Blinding	$d(x,a) = d(x,a')$	$\forall x \in \mathcal{X}$ et $a, a' \in \mathcal{A}$	Discrimination directe
Counterfactual fairness	$E[D(a')\|X] = E[D\|X]$	$\forall a, a' \in \mathcal{A}$	Discrimination indirecte
Π-*fairness*	$E[D_{\Pi,A,a'}\|X] = E[D\|X]$	$\forall a, a' \in \mathcal{A}$	Discrimination indirecte

Où Π est un ensemble de chemin allant de A à D

Limiter l'effet sur les décisions

Limiter les **effets sur les décisions** peut se faire individuellement, ou en espérance.

Nom du critère	Expression	Possibilités	Mot clé relié
Blinding	$d(x,a) = d(x,a')$	$\forall x \in \mathcal{X}$ et $a, a' \in \mathcal{A}$	Discrimination directe
Counterfactual fairness	$E[D(a')\|X] = E[D\|X]$	$\forall a, a' \in \mathcal{A}$	Discrimination indirecte
Π-*fairness*	$E[D_{\Pi, A, a'}\|X] = E[D\|X]$	$\forall a, a' \in \mathcal{A}$	Discrimination indirecte

Où Π est un ensemble de chemin allant de A à D , et $D_{\Pi,a,a'}$ est la décision lorsqu'on propage $A = a'$ à travers les chemins Π, $A = a$ à travers les autres chemins.

Exemple d'admission à l'université : Π-équité

Par abus de notation, on utilise les variables aléatoires comme une fonction de leurs parents.

L'ensemble Π est représenté sur la figure.

On a

$$D_{\Pi,a,a'} = D\{a, T[E(a'), M(a)]\}$$

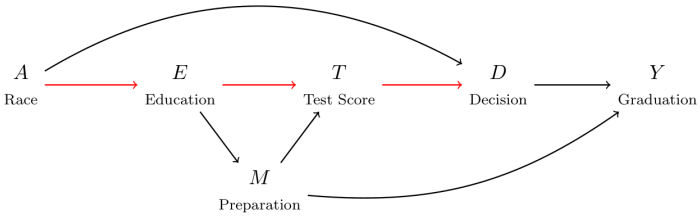


Figure 1 – Graphe causal de l'exemple d'admission à l'université de Corbett-Davies et al. (2023).

Calibration

La *calibration* stipule que la distribution estimée des risques est près de la distribution observée pour **tous les segments***.

Calibration

La *calibration* stipule que la distribution estimée des risques est près de la distribution observée pour **tous les segments***.

Généralement, limiter le **traitement disparate** sur les groupes protégés entraîne un manque de calibration.

Calibration

La *calibration* stipule que la distribution estimée des risques est près de la distribution observée pour **tous les segments***.

Généralement, limiter le **traitement disparate** sur les groupes protégés entraîne un manque de calibration.

Certains auteurs traite la *calibration* comme une contrainte d'équité, stipulant que les **risques estimés devraient avoir la même signification** pour tous les groupes protégés.

Limitations des définitions formelles d'équité

Des critères de décision optimaux pour certaines marges ne **capturent pas les variations** au sein de la population.

Limitations des définitions formelles d'équité

Des critères de décision optimaux pour certaines marges ne **capturent pas les variations** au sein de la population.

Simoiu et al. (2017) discute de l'**inframarginalité** et du problème d'appliquer des critères d'équité fait pour bien fonctionner aux marges, mais pas nécessairement dans les populations inframarginales.

Limitations des définitions formelles d'équité

Des critères de décision optimaux pour certaines marges ne **capturent pas les variations** au sein de la population.

Simoiu et al. (2017) discute de l'**inframarginalité** et du problème d'appliquer des critères d'équité fait pour bien fonctionner aux marges, mais pas nécessairement dans les populations inframarginales.

Une application rigide de règles équitables sur des mesures marginales **peut aggraver les disparités**.

Approche conséquentialiste pour l'équité

Des compromis dépendent de la situation

Pour l'exemple du **diabète**

- Possible de tester tout le monde ($b = 1$)
- Il n'y a pas d'effets secondaires ou indirects sur des tiers non concernés.
- Compromis entre **coût d'un test** et **bénéfice d'un test**.

Des compromis dépendent de la situation

Pour l'exemple du **diabète**

- Possible de tester tout le monde ($b = 1$)
- Il n'y a pas d'effets secondaires ou indirects sur des tiers non concernés.
- Compromis entre **coût d'un test** et **bénéfice d'un test**.

Pour l'exemple de l'**admission**

- Pas possible d'admettre tout le monde ($b < 1$)
- Les décisions ont une influence sur la réputation et l'écosystème universitaire.
- Compromis entre la **performance scolaire** et **diversité**.

Des compromis dépendent de la situation

Pour l'exemple du **diabète**

- Possible de tester tout le monde ($b = 1$)
- Il n'y a pas d'effets secondaires ou indirects sur des tiers non concernés.
- Compromis entre **coût d'un test** et **bénéfice d'un test**.

Pour l'exemple de l'**admission**

- Pas possible d'admettre tout le monde ($b < 1$)
- Les décisions ont une influence sur la réputation et l'écosystème universitaire.
- Compromis entre la **performance scolaire** et **diversité**.

Une approche conséquentialiste

On définit $v(y)$ comme étant le bénéfice d'avoir pris une décision positive $D = 1$ au lieu de $D = 0$ lorsque $Y = y$. Pour l'exemple du diabète

- $v(1)$ est le bénéfice de détecter le diabète.
- $-v(0)$ est le coût du test superflu.

Une approche conséquentialiste

On définit $v(y)$ comme étant le bénéfice d'avoir pris une décision positive $D = 1$ au lieu de $D = 0$ lorsque $Y = y$. Pour l'exemple du diabète

- $v(1)$ est le bénéfice de détecter le diabète.
- $-v(0)$ est le coût du test superflu.

Pour un individu avec $X = x$, l'utilité est

$$u_i(x) = E[v(Y)|X = x] = P(Y = 1|x)v(1) + P(Y = 0|x)v(0)$$

Une approche conséquentialiste

On définit $v(y)$ comme étant le bénéfice d'avoir pris une décision positive $D = 1$ au lieu de $D = 0$ lorsque $Y = y$. Pour l'exemple du diabète

- $v(1)$ est le bénéfice de détecter le diabète.
- $-v(0)$ est le coût du test superflu.

Pour un individu avec $X = x$, l'utilité est

$$u_i(x) = E[v(Y)|X = x] = P(Y = 1|x)v(1) + P(Y = 0|x)v(0)$$

On veut une règle de décision qui maximise les utilités individuelles

$$u(d^*) = \underset{d}{\text{argmax}}\, E[u_i(X)d(X)]$$

Prise de décision en l'absence de contraintes externes

En absence de contraintes externes,Corbett-Davies et al. (2023) proposent d'utiliser des **règles de seuil qui maximisent l'utilité** individuelle et collective.

Prise de décision en l'absence de contraintes externes

En absence de contraintes externes,Corbett-Davies et al. (2023) proposent d'utiliser des **règles de seuil qui maximisent l'utilité** individuelle et collective.

Les définitions formelles de l'**équité** peuvent être en **conflit** avec la maximisation de l'utilité **sans bénéfice clair** pour les individus.

Des compromis dépendent de la situation

Pour l'exemple du **diabète**

- Possible de tester tout le monde ($b = 1$)
- Il n'y a pas d'effets secondaires ou indirects sur des tiers non concernés.
- Compromis entre **coût d'un test** et **bénéfice d'un test**.

Pour l'exemple de l'**admission**

- Pas possible d'admettre tout le monde ($b < 1$)
- Les décisions ont une influence sur la réputation et l'écosystème universitaire.
- Compromis entre la **performance scolaire** et **diversité**.

Des compromis dépendent de la situation

Pour l'exemple du **diabète**

- Possible de tester tout le monde ($b = 1$)
- Il n'y a pas d'effets secondaires ou indirects sur des tiers non concernés.
- Compromis entre **coût d'un test** et **bénéfice d'un test**.

Pour l'exemple de l'**admission**

- Pas possible d'admettre tout le monde ($b < 1$)
- Les décisions ont une influence sur la réputation et l'écosystème universitaire.
- Compromis entre la **performance scolaire** et **diversité**.

Prise de décision en présence de contraintes externes

Tout comme avant, on désire optimiser l'utilité collective

$$u(d) = E[u(X)d(X)].$$

Cette fois, les décisions *faisables* doivent respecter $E[d(X)] \leq b < 1$. On dit alors qu'il y a des **contraintes externes**.

Dans l'exemple d'admission à l'université de Corbett-Davies et al. (2023), il y a consensus sur le fait qu'on aimerait maximiser la **performance académique** et la **diversité**.

Deux dimensions conflictuelles à respecter

Pour maximiser la **performance académique** et la **diversité**, on peut imaginer que les membres du comité de sélection ont une fonction d'utilité

$$u^*(d) = v\{E[d(X)\mathbf{m(X)}], E[d(X)1_{\{\alpha(X)=a_1\}}]\},$$

où $m(X)$ est notre estimation de la performance académique et $v(.,.)$ est croissante en ces deux arguments.

Un exemple de fonction $v(.,.)$ serait, pour $\lambda > 0$

$$v^*(a,b) = a + \lambda b,$$

Domination Pareto

Sans spécifier la forme de la fonction d'utilité $u^*(d)$, il est connu que les **deux dimensions conflictuelles doivent être maximisée**.

Sans même spécifier la forme exacte de u^*, une d politique qui pouvant être remplacée par une politique meilleure[1] d^* est dite **Pareto-dominée**.

1. qui augmente l'une des deux dimensions sans faire diminuer l'autre

Domination Pareto et utilité

D'après le théorème 17 de Corbett-Davies et al. (2023), on a

- « *any feasible decision policy satisfying* ***counterfactual equalized odds*** *is strongly Pareto dominated* »
- « *any feasible decision policy satisfying* ***conditional principal fairness*** *is strongly Pareto dominated* »
- « *any feasible decision policy satisfying* ***path-specific fairness*** *is strongly Pareto dominated* »

Domination Pareto et utilité (suite)

De la même manière, on a

« *…no* ***utility-maximizing decision-policy*** *satisfies counterfactual equalized odds, conditional principal fairness, or path-specific fairness* »

Domination Pareto et utilité (suite)

De la même manière, on a

« *…no* ***utility-maximizing decision-policy*** *satisfies counterfactual equalized odds, conditional principal fairness, or path-specific fairness* »

Poursuivre fermement une définition d'équité mène à des politiques sous-optimales.

Exemple numérique

Exemples : les données

Le jeu de données est un produit dérivé de `norauto` de Dutang and Charpentier (2020) et contient des informations (7 colonnes) sur les coûts de 10 000 assurés.

Exemples : les données

Le jeu de données est un produit dérivé de `norauto` de Dutang and Charpentier (2020) et contient des informations (7 colonnes) sur les coûts de 10 000 assurés.

Variable	Type	Domaine (ou quantité de niveaux)	Description
`id`	Entier	$\mathbb{N}^+$	Indicatrice de jeunes
`Male`	Booléen	$\{0, 1\}$	Indicatrice d'homme
`Young`	Booléen	$\{0, 1\}$	Indicatrice de ≤ 26 ans
`DistLimit`	Cat. ord.	(6 niveaux) {`12000km`, `16000km`, …}	Limite de distance inscrite au contrat
`GeoRegion`	Cat. nom.	(6 niveaux) {`High-`, `High+`, …}	Densité de la région
`Expo`	Numérique	$[0, 1]$	Exposition, en fraction d'une année
`ClaimAmount`	Numérique	$0 \cup \mathbb{R}^+$	Montant de réclamation (Unité non spécifiée)

Exemples : les données

Le jeu de données est un produit dérivé de `norauto` de Dutang and Charpentier (2020) et contient des informations (7 colonnes) sur les coûts de 10 000 assurés.

Variable	Type	Domaine (ou quantité de niveaux)	Description
`id`	Entier	$\mathbb{N}^+$	Indicatrice de jeunes
`Male`	Booléen	$\{0, 1\}$	Indicatrice d'homme
`Young`	Booléen	$\{0, 1\}$	Indicatrice de ≤ 26 ans
`DistLimit`	Cat. ord.	(6 niveaux) {`12000km`, `16000km`, …}	Limite de distance inscrite au contrat
`GeoRegion`	Cat. nom.	(6 niveaux) {`High-`, `High+`, …}	Densité de la région
`Expo`	Numérique	$[0, 1]$	Exposition, en fraction d'une année
`ClaimAmount`	Numérique	$0 \cup \mathbb{R}^+$	Montant de réclamation (Unité non spécifiée)

Exemple : l'objectif

L'objectif sera d'aider une compagnie avec une **capacité de 500 assurés** à **accepter des assurés** à **coûts** M **élevés**, leur clientèle cible, tout en ayant une bonne **parité de genre**.

Exemple : l'objectif

L'objectif sera d'aider une compagnie avec une **capacité de 500 assurés** à **accepter des assurés** à **coûts** M **élevés**, leur clientèle cible, tout en ayant une bonne **parité de genre**.

- Le budget b est de $500/10\,000 = 0.05$.

Exemple : l'objectif

L'objectif sera d'aider une compagnie avec une **capacité de 500 assurés** à **accepter des assurés** à **coûts M élevés**, leur clientèle cible, tout en ayant une bonne **parité de genre**.

- Le budget b est de $500/10\,000 = 0.05$.
- Une décision positive $D = 1$ est d'accepter l'assuré.

Exemple : l'objectif

L'objectif sera d'aider une compagnie avec une **capacité de 500 assurés** à **accepter des assurés** à **coûts M élevés**, leur clientèle cible, tout en ayant une bonne **parité de genre**.

- Le budget b est de $500/10\,000 = 0.05$.
- Une décision positive $D = 1$ est d'accepter l'assuré.
- L'indicatrice de coûts élevés est $Y_i = 1_{\{M_i \geq VaR_{0.95}(M)\}}$

Exemple : l'objectif

L'objectif sera d'aider une compagnie avec une **capacité de 500 assurés** à **accepter des assurés** à **coûts M élevés**, leur clientèle cible, tout en ayant une bonne **parité de genre**.

- Le budget b est de $500/10\,000 = 0.05$.
- Une décision positive $D = 1$ est d'accepter l'assuré.
- L'indicatrice de coûts élevés est $Y_i = 1_{\{M_i \geq VaR_{0.95}(M)\}}$
- La fonction d'utilité à maximiser est $u(X, A) = E(\widehat{Y}(X, A) \cdot \hat{d}(X, A)) - \lambda \cdot E[|1_{\{A=0\}} - 1_{\{A=1\}}| \cdot d(X, A)]$.

Exemple : la méthodologie

1. Entraîner trois modèles pour prédire l'occurence de coûts élevés `highcost`.
   - `best: highcost` $\sim$ `.`
   - `un: highcost` $\sim$ `. - Male`
   - `aw: highcost_mod` $\sim$ `. - Male` `*fairadapt`

Exemple : la méthodologie

1. Entraîner trois modèles pour prédire l'occurence de coûts élevés `highcost`.
   - `best: highcost` $\sim$ `.`
   - `un: highcost` $\sim$ `. - Male`
   - `aw: highcost_mod` $\sim$ `. - Male` `*fairadapt`

2. Pour chaque modèle, identifier les 500 hommes et 500 femmes les plus risqués.

Exemple : la méthodologie

1. Entraîner trois modèles pour prédire l'occurence de coûts élevés `highcost`.
   - `best: highcost` $\sim$ `.`
   - `un: highcost` $\sim$ `. - Male`
   - `aw: highcost_mod` $\sim$ `. - Male` `*fairadapt`
2. Pour chaque modèle, identifier les 500 hommes et 500 femmes les plus risqués.
3. Étudier les métriques d'équité selon la composition choisie.

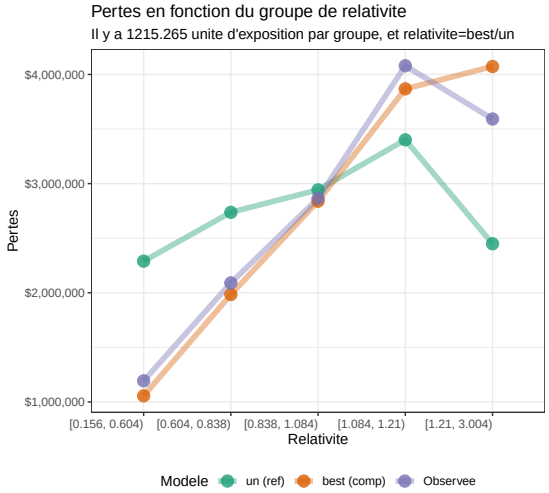
Exemple : les modèles

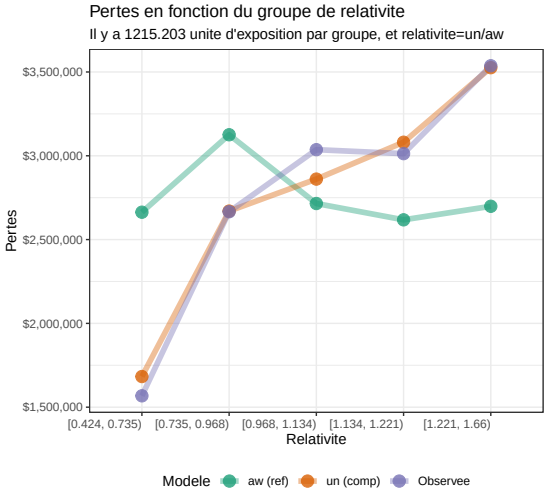
	Modèle			
Nom	`best`	`un`	`aw`	`average`
Variables	(X, D)	X	X^*	—
Restrictions	—	D	$(D, X \backslash X^*)$	(D, X)
Blinding ?	Non	Oui	Oui	Oui
Calibré ?	Oui	Non	Non	Non
AUC	0.619	0.611	0.597	0.500

Détails de modélisation :

- `lightgbm`;
- `fairadapt`* pour `aw`.
- `cv = 5`

Calibration des moyennes prédites





Calibration des probabilités prédites

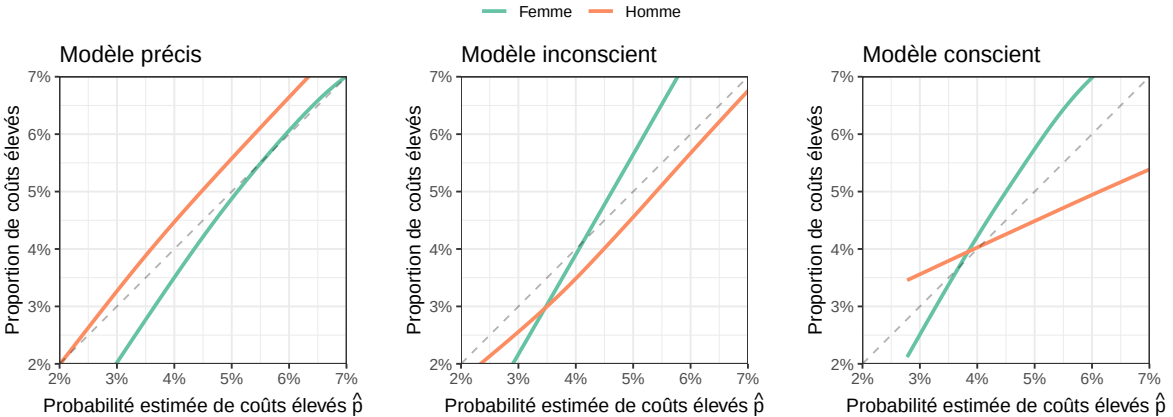
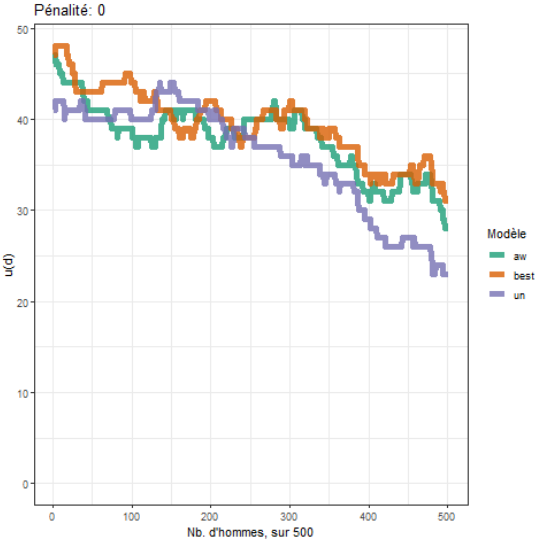


Figure 2 – GAM de Y_i en fonction de $\hat{p}_i$ pour les modèles précis (gauche), inconscient (milieu) ou conscient (droite).

On étudie l'utilité pour $\lambda \in \{0, 7.5, \dots 750\}$

$$u(X, A) = E[\hat{Y}(X, A) \cdot \hat{d}(X, A)] - \lambda \cdot E[|1_{\{A=0\}} - 1_{\{A=1\}}| \cdot d(X, A)]$$

Plus λ est élevée, plus les proportions des genres tend vers les proportions marginales.



Conclusion

C'est important de **choisir les principes éthiques en fonction des objectifs politiques** pour éviter les conséquences négatives inutiles.

- Déterminer les **objectifs conflictuels principaux**.
- **Formaliser ces objectifs** par une fonction d'utilité $u(d)$

Conclusion

C'est important de **choisir les principes éthiques en fonction des objectifs politiques** pour éviter les conséquences négatives inutiles.

- Déterminer les **objectifs conflictuels principaux**.
- **Formaliser ces objectifs** par une fonction d'utilité $u(d)$

Il est nécessaire d'évaluer les **conséquences réelles des décisions algorithmiques**, plutôt que de suivre des critères formels rigides.

- Déterminer les **coûts et les bénéfices** des décisions $v(Y)$
- S'assurer que les ajustements bénéficies aux bons groupes, et non aux **populations inframarginales**.

Bibliographie i

Autorité des marchés financiers (2024). Meilleures pratiques pour l'utilisation responsable de l'ia dans le secteur financier. Accessed : Febuary 28, 2024.

Corbett-Davies, S., Gaebler, J. D., Nilforoshan, H., Shroff, R., and Goel, S. (2023). The measure and mismeasure of fairness. *The Journal of Machine Learning Research*, 24(1) :14730–14846.

Dutang, C. and Charpentier, A. (2020). Package 'casdatasets'. *url : https ://www.openml.org/search*.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 59–68, New York, NY, USA. Association for Computing Machinery.

Simoiu, C., Corbett-Davies, S., and Goel, S. (2017). The problem of infra-marginality in outcome tests for discrimination.