Équité des classifieurs en présence de contraintes entre les attributs

Martin C. Cooper[1], Imane Bousdira[2]

[1] IRIT, University of Toulouse
[2] IRIT, Toulouse INP

Firstname.LastName@irit.fr

Résumé

En apprentissage automatique, une définition admise de l'équité d'une décision prise par un classifieur est qu'elle ne doit pas dépendre d'attributs protégés, tels que le genre. Cependant, lorsque des contraintes existent entre les attributs, de telles dépendances peuvent être cachées par celles-ci. Pour pallier ce problème, nous proposons de qualifier une décision d'équitable si elle possède une explication équitable. Nous définissons cette explication comme un impliquant premier de la décision qui ne contient aucun attribut protégé, en tenant compte des contraintes.
Trois définitions possibles de l'équité d'un classifieur peuvent être distinguées : pour chacune de ses décisions, (1) toutes les explications sont équitables, (2) il existe au moins une explication équitable, ou (3) toute modification des attributs protégés n'affecte pas le résultat. Ce document présente un résumé de notre travail [8], dans lesquel nous avons identifié les relations entre ces différentes définitions et étudié la complexité computationnelle de la vérification de l'équité des classifieurs.

Mots-clés

Équité, contraintes, explication abductive, complexité.

Abstract

In Machine Learning, an accepted definition of fairness of a decision taken by a classifier is that it should not depend on protected features, such as gender. Unfortunately, when constraints exist between features, such dependencies can be obscured by the constraints. To avoid this problem, we propose that a decision be considered fair if it has a fair explanation. We define this explanation as a prime-implicant reason for the decision that does not contain any protected feature, while taking constraints into account.
Three possible definitions of fairness of a classifier are that for all its decisions (1) there are only fair explanations, (2) there is at least one fair explanation, or (3) changing protected features does not change the outcome. This document presents a summary of our work [8], in which we identified the relationships between these different definitions of fairness and studied the computational complexity of testing fairness of classifiers.

Keywords

Fairness, constraints, abductive explanation, complexity.

1 Introduction

L'attention portée à l'équité a considérablement augmenté ces dernières années, en raison de l'usage croissant de l'intelligence artificielle dans divers secteurs. Les préoccupations relatives à l'équité se sont intensifiées avec l'introduction des systèmes d'IA dans des domaines critiques, tels que les décisions de recrutement [20]. Par ailleurs, le principe d'équité a été largement mis en avant dans le cadre de l'IA de confiance [13, 17]. Étant donné l'importance de l'équité, ce concept a été l'objet de nombreuses études, donnant lieu à de multiples définitions [6, 19]. L'une des définitions les plus fréquemment citées décrit l'équité comme l'absence de toute forme de favoritisme ou de discrimination envers un individu ou un groupe, en fonction de caractéristiques innées ou acquises [16]. Des définitions plus précises sont proposées dans la littérature, que Verma et Rubin [22] analysent dans leur article de synthèse.
Il est raisonnable de considérer qu'un classifieur est équitable si aucune de ses décisions n'est influencée par des attributs protégés, tels que l'origine ethnique ou le genre. Cependant, en présence de contraintes entre les attributs, cette définition simple doit être affinée. Les contraintes peuvent provenir de différentes origines. Elles peuvent résulter de l'encodage (encodage one-hot dans l'exemple 1) ou de restrictions liées au monde réel. En effet, ignorer ces contraintes peut altérer l'évaluation de l'équité.

Exemple 1. *Considérons une fonction $\kappa(m, f, g)$ qui renvoie 1 si un employé est éligible à une prime en fonction des attributs booléennes : m (homme), f (femme), g (objectifs atteints). Les attributs m et f sont protégés. Supposons qu'il existe une contrainte unique : $m \equiv \neg f$. Pour cette tâche, un modèle possible est : $\kappa(m, f, g) \equiv (m \wedge g) \vee (f \wedge g)$. Sans tenir compte de la contrainte, ce modèle pourrait être considéré comme inéquitable puisque $\kappa(0, 0, 1) \neq \kappa(1, 0, 1)$ (deux instances qui ne diffèrent que par l'attibut m ont des résultats différents). Cependant, en considérant la constraint, $\kappa(m, f, g) \equiv g$, ainsi, κ est clairement équitable.*

Le reste du document présente un aperçu des principaux résultats détaillés dans l'article [8]. La contribution majeure réside dans la prise en compte des contraintes entre les attributs lors de l'évaluation de l'équité, en proposant une nouvelle définition de l'équité en termes d'explications.

2 Préliminaires

Soit $\mathcal{F}$ un ensemble de n attributs et $\mathcal{K}$ un ensemble de (au moins deux) classes. Nous partons du principe que l'ensemble des attributs $\mathcal{F}$ est divisé en un ensemble $\mathcal{P}$ d'attributs protégés (tels que l'origine ethnique ou le genre) et un ensemble $\mathcal{N}$ d'attributs non protégés. $\mathbb{F}$ désigne l'espace des attributs, qui est le produit cartésien des domaines des n attributs. Il se peut que certains vecteurs d'attributs de $\mathbb{F}$ ne soient pas possibles, en raison d'un ensemble de contraintes noté $\mathcal{C}$. Nous désignons par $\mathbb{F}[\mathcal{C}]$ l'ensemble des instances de $\mathbb{F}$ qui satisfont les contraintes. Nous supposons que $\kappa : \mathbb{F}[\mathcal{C}] \to \mathcal{K}$ est un classifieur. Pour simplifier la notation, nous associons à chaque attribut un nombre compris entre 1 et n. Pour $\mathcal{S} \subseteq \mathcal{F}$ et $x \in \mathbb{F}$, la notation $x_{\mathcal{S}}$ désigne l'affectation partielle $\{(i, x_i) : i \in \mathcal{S}\}$ aux attributs de $\mathcal{S}$.
Dans le cadre de l'intelligence artificielle explicable (XAI), une notion formelle d'explication d'une décision a émergé [1, 2, 4, 5, 15, 18, 21], souvent désignée par explication abductive AXp (ou raison suffisante).

Définition 1. *Une* AXp faible *d'une décision associée à un couple classifieur-instance $\langle \kappa, x \rangle$ est un sous-ensemble S d'attributs tel que $\forall y \in \mathbb{F}[\mathcal{C}]$, $y_S = x_S \to \kappa(y) = \kappa(x)$. Une* AXp *est une AXp faible minimal pour l'inclusion.*

Dans la définition d'une AXp, les contraintes $\mathcal{C}$ sont prises en compte en ne considérant que les instances y (qui coïncident avec x sur S) qui satisfont les contraintes [3, 9, 10, 23]. Selon le contexte et pour simplifier la présentation, il peut être utile de considérer l'affectation partielle x_S, plutôt que l'ensemble S, comme l'explication abductive.

3 Équité des décisions en présence de contraintes

Étant donné une décision $\kappa(x) = c$ dans un espace d'attributs *sans contrainte* $\mathbb{F}$, nous pouvons considérer une AXp comme un impliquant premier de la décision : il s'agit d'une raison suffisante pour la décision qui n'est pas subsumée par une autre raison suffisante. Dans un espace d'attibuts sans contrainte, A subsume B si et seulement si $A \subseteq B$. En présence de contraintes $\mathcal{C}$, la subsomption est plus subtile : lorsqu'on explique une décision $\kappa(x) = c$, *A subsume B* si et seulement si l'affectation partielle x_B implique l'affectation partielle x_A dans $\mathbb{F}[\mathcal{C}]$ (c'est-à-dire, $\forall y \in \mathbb{F}[\mathcal{C}], (y_B = x_B) \to (y_A = x_A)$). *$A$ subsume strictement B* si et seulement si A subsume B, mais B ne subsume pas A. Cela conduit à la définition 2 suivante [7]. Les définitions d'une décision équitable sont introduites ci-après.

Définition 2. *Une* PI-explication *(explication impliquant-premier) d'une décision $\kappa(x) = c$ dans un espace d'attributs contraint $\mathbb{F}[\mathcal{C}]$ est une AXp qui n'est pas strictement subsumé par une autre AXp.*

Définition 3. *Une PI-explication est* équitable *si elle ne comporte aucun attribut protégé. Nous disons qu'une décision est* existentiellement équitable *si elle possède une PI-explication équitable. Nous disons qu'une décision est* universellement équitable *si toutes ses PI-explications sont équitables. Une décision est* inéquitable *si elle ne possède aucune PI-explication équitable.*

Dans l'exemple 1, toutes les décisions prises par le classifieur sont universellement équitables, car elles ont toutes la même PI-explication unique $\{g\}$.

4 L'impact des contraintes sur l'équité des décisions

Dans un premier temps, nous examinons le cas où il existe des contraintes entre les attributs protégés $\mathcal{P}$ et les attributs non protégés $\mathcal{N}$. L'exemple 2 montre que l'ignorance des contraintes peut amener à considérer comme inéquitable une décision universellement équitable. Il convient de noter que le fait de ne pas tenir compte des contraintes peut également faire passer une décision inéquitable pour universellement équitable.

Exemple 2. *Considérons une fonction $\kappa(e, m)$ qui renvoie 1 si une personne est éligible à une formation gratuite, en fonction de deux attributs : 'en emploi' (e) et 'en congé de maternité' (m). Supposons que $\kappa(e, m) \equiv (\neg e) \wedge (\neg m)$. Sans tenir compte des contraintes, la seule PI-explication pour la décision $\kappa(0, 0) = 1$ est $\{e, m\}$. Si m est un attribut protégé, alors cette PI-explication est inéquitable et, par conséquent, la décision paraît inéquitable. Cependant, il existe une contrainte selon laquelle une personne ne peut être en congé maternité que si elle est employé, donc $(e, m) = (0, 1)$ est impossible. Ainsi, dans $\mathbb{F}[\mathcal{C}]$, $\kappa \equiv \neg e$ et la seule PI-explication pour $\kappa(0, 0) = 1$ est $\{e\}$, qui est équitable.*

La Figure 1 présente une synthèse de l'impact de la prise en compte ou l'ignorance des contraintes dans les cas où elles n'apparaissent qu'au sein de $\mathcal{P}$ ou qu'au sein de $\mathcal{N}$.

5 Équité des classifieurs en présence de contraintes

Dans cette partie, nous nous intéressons à l'équité des classifieurs, plutôt qu'à celle d'une décision individuelle. À cet égard, nous examinons une définition existante et en proposons de nouvelles.

5.1 FTU contrainte comme mesure d'équité

Rappelons la définition de FTU (Fairness Through Unawareness), qui est une règle classique permettant d'évaluer l'équité des classifieurs dans des espaces d'attributs non contraints [11, 14, 12] :

$$\forall x, y \in \mathbb{F},\ x_{\mathcal{N}} = y_{\mathcal{N}} \ \to\ \kappa(x) = \kappa(y)$$

Lorsque des contraintes existent, nous ne pouvons pas appliquer directement FTU comme indiqué ci-dessus, notamment parce que les valeurs de $\kappa(y)$ peuvent être considérées comme indéfinies pour $y \notin \mathbb{F}[\mathcal{C}]$. Comme le montre l'exemple 1, lorsque les contraintes sont ignorées, un classifieur équitable peut être considéré à tort comme inéquitable. Ainsi, la définition de FTU dans $\mathbb{F}[\mathcal{C}]$ est la suivante.

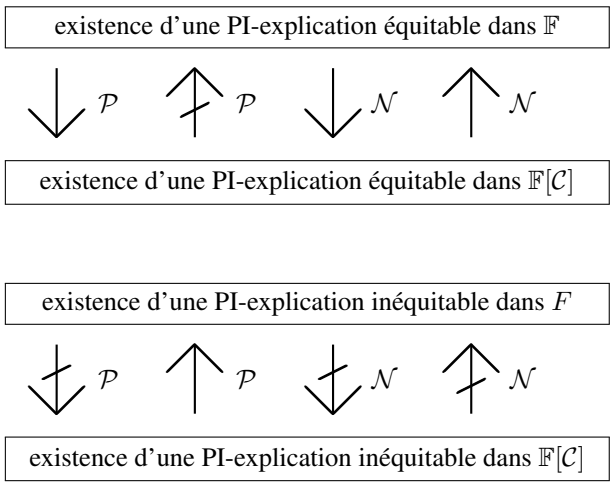

FIGURE 1 – La conservation (ou non) des PI-explications équitables/inéquitables de $\langle \kappa, x \rangle$ (pour $x \in \mathbb{F}[\mathcal{C}]$) lorsque les contraintes sont prises en compte (flèches vers le bas) ou ignorées (flèches vers le haut). $\mathcal{P}$ ou $\mathcal{N}$ signifie que les contraintes apparaissent uniquement au sein des attributs de $\mathcal{P}$ ou $\mathcal{N}$, respectivement.

Définition 4. *Un classifieur $\kappa : \mathbb{F}[\mathcal{C}] \to \mathcal{K}$ satisfait FTU dans $\mathbb{F}[\mathcal{C}]$* (FTU contrainte) *si*

$$\forall x, y \in \mathbb{F}[\mathcal{C}],\ x_{\mathcal{N}} = y_{\mathcal{N}} \ \rightarrow\ \kappa(x) = \kappa(y)$$

5.2 Mesures de l'équité basées sur les PI-explications

Conformément aux définitions d'une décision équitable données dans la Section 3, nous pouvons définir l'équité d'un classifieur, en se basant sur l'équité de ses décisions.

Définition 5. *Un classifieur $\kappa : \mathbb{F}[\mathcal{C}] \to \mathcal{K}$ satisfait*
- l'équité existentielle *si $\forall x \in \mathbb{F}[\mathcal{C}]$, $\langle \kappa, x \rangle$ a une PI-explication qui est équitable, et*
- l'équité universelle *si $\forall x \in \mathbb{F}[\mathcal{C}]$, toutes les PI-explications de $\langle \kappa, x \rangle$ sont équitables.*

6 Relations entre les définitions de l'équité des classifieurs

Nous résumons ci-dessous les principaux résultats issus de l'étude des relations entre les différentes définitions de l'équité des classifieurs.

- **FTU contrainte et équité existentielle**
  L'équité existentielle d'un classifieur κ implique que $\mathcal{N}$ est une AXp faible et donc que κ satisfait FTU dans $\mathbb{F}[\mathcal{C}]$. En revanche, FTU dans $\mathbb{F}[\mathcal{C}]$ n'implique pas l'équité existentielle.
- **Équité existentielle et équité universelle**
  Il est trivial qu'un classifieur satisfaisant l'équité universelle est existentiellement équitable. Cependant, l'équité existentielle n'implique pas l'équité universelle.

Nous introduisons par la suite des cas particuliers pour lesquels des définitions de l'équité s'avèrent équivalentes.

6.1 Absence de contraintes entre les attributs protégés et non protégés

Théorème 1. *Lorsqu'il n'existe aucune contrainte dans $\mathcal{C}$ entre les attributs protégés et non protégés, les notions suivantes d'équité d'un classifieur sont équivalentes :*

1. *équité universelle*
2. *équité existentielle*
3. *FTU contrainte (FTU dans $\mathbb{F}[\mathcal{C}]$)*

6.2 Contraintes lâches

Définition 6. *Soit un ensemble d'attributs $\mathcal{F}$ divisé en un ensemble d'attributs protégés $\mathcal{P}$ et un ensemble d'attributs non protégés $\mathcal{N}$. Un ensemble de contraintes $\mathcal{C}$ est* lâche *(par rapport à la partition $(\mathcal{P}, \mathcal{N})$ de l'ensemble des attributs) en $x \in \mathbb{F}[\mathcal{C}]$ si pour tout $p \in \mathcal{P}$, $x_{\{p\}}$ ne subsume pas strictement $x_{\mathcal{N}}$. L'ensemble de contraintes $\mathcal{C}$ est* lâche *si $\mathcal{C}$ est lâche en chaque $x \in \mathbb{F}[\mathcal{C}]$.*

Étant donné un ensemble de contraintes $\mathcal{C}$, la recherche d'affectations inéquitables qui susbsument strictement certaines affectations équitables peut être effectuée hors ligne en tant qu'étape préliminaire, indépendamment de tout classifieur, permettant de mieux comprendre les implications possibles liées à l'imposition de contraintes. Si de telles subsomptions n'existent pas, FTU contrainte garantit l'existence d'une PI-explication équitable.

Proposition 1. *Si l'ensemble de contraintes $\mathcal{C}$ est lâche, alors FTU dans $\mathbb{F}[\mathcal{C}]$ implique l'équité existentielle.*

6.3 Classifieurs détachés et contraintes

Définition 7. *Soit une instance $x \in \mathbb{F}[\mathcal{C}]$, la décision $\kappa(x) = c$ est* détachée *(par rapport aux contraintes $\mathcal{C}$) si l'ensemble $\mathcal{N}$ des attributs non protégés est une AXp faible de la décision $\kappa(x) = c$ qui n'est pas strictement subsumé par une AXp faible inéquitable. Un classifieur κ est* détaché *dans $\mathbb{F}[\mathcal{C}]$ si toutes ses décisions (pour $x \in \mathbb{F}[\mathcal{C}]$) sont détachées.*

Proposition 2. *Si un classifieur κ est détaché dans $\mathbb{F}[\mathcal{C}]$, alors κ est existentiellement équitable.*

7 Complexité computationnelle de la vérification de l'équité

Cette section présente les principaux résultats concernant la complexité de l'évaluation de l'équité. Nous supposons que κ est une fonction connue et que le calcul de $\kappa(z)$ pour $z \in \mathbb{F}[\mathcal{C}]$ peut être effectué en temps polynomial. Nous débutons par les tests d'équité existentielle et universelle pour une décision unique, à partir de laquelle nous déterminons la complexité de la vérification de l'équité d'un classifieur. Nous abordons ensuite la complexité du test de FTU contrainte qui est accessible aux solveurs SAT, avant de conclure par la complexité des tests pour évaluer si un ensemble de contraintes est lâche ou si un classifieur est détaché.

Proposition 3. *Le problème consistant à vérifier si une décision est existentiellement équitable appartient à* Σ_3^P. *Le problème consistant à vérifier si une décision est universellement équitable appartient à* Π_3^P.

Proposition 4. *Le problème consistant à vérifier l'équité existentielle d'un classifieur appartient à* Π_4^P. *Le problème consistant à vérifier l'équité universelle d'un classifieur appartient à* Π_3^P.

Proposition 5. *Le test de FTU contrainte pour un classifieur est coNP-complet.*

Nous pouvons énoncer le corollaire suivant du théorème 1, de la Proposition 1 et de la proposition 5.

Corollaire 1. *Lorsqu'il n'y a aucune contrainte dans* $\mathcal{C}$ *entre les attributs protégés et non protégés, le test de l'équité universelle d'un classifieur est coNP-complet. Lorsque l'ensemble de contraintes* $\mathcal{C}$ *est lâche, le test de l'équité existentielle d'un classifieur est également coNP-complet.*

Proposition 6. *Les problèmes consistant à vérifier si un ensemble de contraintes est lâche ou si un classifieur est détaché appartiennent à* Π_2^P.

8 Conclusion

L'équité est une exigence fondamentale pour les systèmes d'IA, dont l'influence croît dans des domaines critiques. L'évaluation de l'équité d'une décision particulière ou d'un classifieur dépend de la définition retenue de cette notion, étant donné la diversité des métriques existantes. Cette étude montre que l'ignorance des contraintes entre les attributs peut altérer de manière significative cette évaluation, notamment lorsqu'il existe des contraintes entre les attributs protégés et non protégés.

Les définitions proposées dans ce travail reposent sur l'existence d'explications ne contenant aucun attribut protégé. Nous considérons que cette approche est pertinente, car elle révèle les *raisons* sous-jacentes à chaque décision prise par le classifieur. Cependant, l'augmentation de la complexité computationnelle soulève certaines préoccupations. Pour y remédier, des conditions moins coûteuses à vérifier ont été identifiées.

Ce document présente un aperçu des principaux points abordés dans l'article [8], qui pose des bases théoriques novatrices sur l'équité formelle en présence de contraintes entre les attributs. L'approche théorique est enrichie par des perspectives pratiques, ouvrant ainsi plusieurs pistes pour la recherche future. L'article complet explore en profondeur ces éléments en fournissant des preuves, des exemples concrets et des explications détaillées.

Remerciements

Ce travail a été soutenu par le projet ForML ANR-23-CE25-0009.

Références

[1] Leila Amgoud and Jonathan Ben-Naim. Axiomatic foundations of explainability. In Luc De Raedt, editor, *IJCAI*, pages 636–642. ijcai.org, 2022. doi:10.24963/IJCAI.2022/90.

[2] Gilles Audemard, Steve Bellart, Louenas Bounia, Frédéric Koriche, Jean-Marie Lagniez, and Pierre Marquis. On the computational intelligibility of boolean classifiers. In Meghyn Bienvenu, Gerhard Lakemeyer, and Esra Erdem, editors, *KR*, pages 74–86, 2021. doi:10.24963/KR.2021/8.

[3] Gilles Audemard, Jean-Marie Lagniez, Pierre Marquis, and Nicolas Szczepanski. Deriving provably correct explanations for decision trees : The impact of domain theories. In *IJCAI*, pages 3688–3696. ijcai.org, 2024. URL : https://www.ijcai.org/proceedings/2024/408.

[4] Pablo Barceló, Mikaël Monet, Jorge Pérez, and Bernardo Subercaseaux. Model interpretability through the lens of computational complexity. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *NeurIPS*, 2020.

[5] Shahaf Bassan, Guy Amir, and Guy Katz. Local vs. global interpretability : A computational complexity perspective. In *ICML*. OpenReview.net, 2024.

[6] Reuben Binns. Fairness in machine learning : Lessons from political philosophy. In Sorelle A. Friedler and Christo Wilson, editors, *Conference on Fairness, Accountability and Transparency, FAT*, volume 81 of *Proceedings of Machine Learning Research*, pages 149–159. PMLR, 2018.

[7] Martin C. Cooper and Leila Amgoud. Abductive explanations of classifiers under constraints : Complexity and properties. In Kobi Gal, Ann Nowé, Grzegorz J. Nalepa, Roy Fairstein, and Roxana Radulescu, editors, *ECAI*, pages 469–476. IOS Press, 2023.

[8] Martin C. Cooper and Imane Bousdira. Fairness of classifiers in the presence of constraints between features. In *CP 2026*, 2026. arXiv:2605.00592.

[9] Martin C. Cooper and João Marques-Silva. On the tractability of explaining decisions of classifiers. In Laurent D. Michel, editor, *CP*, volume 210 of *LIPIcs*, pages 21 :1–21 :18. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021. doi:10.4230/LIPICS.CP.2021.21.

[10] Niku Gorji and Sasha Rubin. Sufficient reasons for classifier decisions in the presence of domain constraints. In *AAAI*, pages 5660–5667. AAAI Press, 2022. doi:10.1609/AAAI.V36I5.20507.

[11] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. The case for process fairness in learning : Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law*, 2016.

[12] Alexey Ignatiev, Martin C. Cooper, Mohamed Siala, Emmanuel Hebrard, and João Marques-Silva. Towards formal fairness in machine learning. In Helmut Simonis, editor, *CP*, volume 12333 of *LNCS*, pages 846–867. Springer, 2020.

[13] Davinder Kaur, Suleyman Uslu, Kaley J. Rittichier, and Arjan Durresi. Trustworthy artificial intelligence : A review. *ACM Comput. Surv.*, 55(2), 2022.

[14] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *NIPS*, pages 4066–4076, 2017.

[15] João Marques-Silva and Alexey Ignatiev. Delivering trustworthy AI through formal XAI. In *AAAI*, pages 12342–12350. AAAI Press, 2022. doi:10.1609/AAAI.V36I11.21499.

[16] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), 2021.

[17] High-Level Expert Group on Artificial Intelligence (AI HLEG) set up by the European Commission. Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, 2019.

[18] Sebastian Ordyniak, Giacomo Paesani, Mateusz Rychlicki, and Stefan Szeider. Explaining decisions in ML models : A parameterized complexity analysis. In Pierre Marquis, Magdalena Ortiz, and Maurice Pagnucco, editors, *KR*, 2024. doi:10.24963/KR.2024/53.

[19] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C. Parkes, and Yang Liu. How do fairness definitions fare ? : Examining public attitudes towards algorithmic definitions of fairness. In Vincent Conitzer, Gillian K. Hadfield, and Shannon Vallor, editors, *AAAI/ACM Conference on AI, Ethics, and Society, AIES*, pages 99–106. ACM, 2019.

[20] Candice Schumann, Jeffrey S. Foster, Nicholas Mattei, and John P. Dickerson. We need fairness and explainability in algorithmic hiring. In Amal El Fallah Seghrouchni, Gita Sukthankar, Bo An, and Neil Yorke-Smith, editors, *AAMAS*, pages 1716–1720. International Foundation for Autonomous Agents and Multiagent Systems, 2020.

[21] Andy Shih, Arthur Choi, and Adnan Darwiche. A symbolic approach to explaining bayesian network classifiers. In Jérôme Lang, editor, *IJCAI*, pages 5103–5111. ijcai.org, 2018. doi:10.24963/IJCAI.2018/708.

[22] Sahil Verma and Julia Rubin. Fairness definitions explained. In Yuriy Brun, Brittany Johnson, and Alexandra Meliou, editors, *Proceedings of the International Workshop on Software Fairness, FairWare@ICSE*, pages 1–7. ACM, 2018.

[23] Jinqiang Yu, Alexey Ignatiev, Peter J. Stuckey, Nina Narodytska, and João Marques-Silva. Eliminating the impossible, whatever remains must be true : On extracting and applying background knowledge in the context of formal explanations. In Brian Williams, Yiling Chen, and Jennifer Neville, editors, *AAAI*, pages 4123–4131. AAAI Press, 2023. doi:10.1609/AAAI.V37I4.25528.