On Formalizing Fairness in Prediction with Machine Learning

Pratik Gajane [1] **Mykola Pechenizkiy** [2]

Abstract

Machine learning algorithms for prediction are increasingly being used in critical decisions affecting human lives. Various fairness formalizations, with no firm consensus yet, are employed to prevent such algorithms from systematically discriminating against people based on certain attributes protected by law. The aim of this article is to survey how fairness is formalized in the machine learning literature for the task of prediction and present these formalizations with their corresponding notions of distributive justice from the social sciences literature. We provide theoretical as well as empirical critiques of these notions from the social sciences literature and explain how these critiques limit the suitability of the corresponding fairness formalizations to certain domains. We also suggest two notions of distributive justice which address some of these critiques and discuss avenues for prospective fairness formalizations.

1. Introduction

Discrimination refers to unfavourable treatment of people due to the membership to certain demographic groups that are distinguished by the attributes protected by law (henceforth, *protected attributes*). Discrimination, based on many attributes and in several domains, is prohibited by international legislation. Nowadays, machine learning algorithms are increasingly being used in high-impact domains such as credit, employment, education, and criminal justice which are prone to discrimination. The goal of fairness in prediction with machine learning is to design algorithms that make *fair* predictions devoid of discrimination.

The aim of this article is to survey how fairness is *formalized* in the machine learning literature and present these formalizations with their corresponding notions from the social sciences literature. The fairness formal-

izations in the machine learning literature correspond to the notions of *distributive justice* from the social sciences literature, as we discuss in Section 2. Since, some formalizations of fairness can be conflicting with others, the predictions produced by the algorithms using them would vastly differ as well. Therefore, from the practical point of view, it is important to study how fairness is formalized in the machine learning literature and the implications of various formalizations. To this end, **we present theoretical as well as empirical critiques of their corresponding notions from the social sciences literature.** The co-presentation is with the intention to assist in **determining the suitability of the existing formalizations of fairness in machine learning literature** and **building newer formalizations of fairness**. In Section 3, we **nominate two notions from the social sciences literature** which answer some of the critiques of the existing formalizations in the machine learning literature. Lastly, in Section 4, we discuss avenues for prospective fairness formalizations. We begin by formulating the problem of prediction with machine learning.

Mathematical formulation of prediction with machine learning: Let X, A and Z represent a set of individuals i.e. a *population*, protected attributes and remaining attributes respectively. Each individual can be assigned an outcome from a finite set Y. Some of the prediction outcomes are more beneficial or desirable than others. For an individual $x_i \in X$, let y_i be the true outcome (or label) to be predicted. A (possibly randomized) predictor can be represented by a mapping $\mathcal{H} : X \rightarrow Y$ from population X to the set of outcomes Y, such that $\mathcal{H}(x_i)$ is the predicted outcome for individual x_i. A group-conditional predictor consists of a set of mappings, one for each group of the population, $\mathcal{H} = \{\mathcal{H}_S\}$ for all $S \subset X$. For the sake of simplicity, assume that the groups induce a partition of the population.

2. What is fair? (Formalizations of fairness)

The first step in formalizing fairness in prediction with machine learning is to answer the following two questions:

- **Parity or preference?** : whether fairness means achieving parity or satisfying the preferences.
- **Treatment or impact?** : whether fairness is to be maintained in treatment or impact (results).

---

[1]Department of Computer Science, Montanuniversitat Leoben, Austria, pratik.gajane@unileoben.ac.at [2]Department of Computer Science, TU Eindhoven, the Netherlands, m.pechenizkiy@tue.nl.

Table 1. The surveyed formalizations of fairness

	Parity	Preference
Treatment	Unawareness Counterfactual measures	Preferred treatment
Impact	Group fairness Individual fairness Equality of opportunity	Preferred impact

Next, we will see the existing formalizations of fairness in the machine learning literature. Table 1 summarizes how they answer the questions presented above.

2.1. Fairness through unawareness

Any predictor which is not group-conditional satisfies this measure. Formally, it is defined as follows:

Definition 1 *(fairness through unawareness) A predictor is said to achieve fairness through unawareness if protected attributes are not explicitly used in the prediction process.*

A number of proposed predictors in the machine learning literature satisfy this measure (16; 30), while some don't (8; 22; 26). However, satisfying fairness through unawareness is not a sufficient condition to avoid discrimination when other background knowledge is available (35). Furthermore, some of the assumptions made during the construction of a predictor might not hold in real-life scenarios (9) which leads to discrimination even while satisfying this measure.

From the point of view of distributive justice, fairness through unawareness corresponds to the approach of being "blind" to counter discrimination. However, various discriminatory practices have been documented following race-blind approach in education, housing, credit, criminal justice system (7; 46). It has shown that, in the long run, race-blind approach is less efficient than race-conscious approach (18). Alternatively, some studies show that a blind approach can work for some specific tasks (21).

The above critiques challenge the suitability of fairness through unawareness to domains in which, protected attributes can be deduced from easily available non-protected attributes and structural barriers, which hinder the protected groups, are shown to be present by credible surveys.

2.2. Counterfactual measures

These measures model fairness through tools from causal interference and can be considered as a sub-class of causal measures (39; 31). Kusner et al. (28) recently introduced a counterfactual measure which can be defined as follows:

Definition 2 *A predictor $\mathcal{H}$ is counterfactually fair, given $Z = z$ and $A = a$, for all y and $a \neq a'$, iff*

$$\mathbb{P}\{\mathcal{H}_{A=a} = y \mid Z = z, A = a\} = \mathbb{P}\{\mathcal{H}_{A=a'} = y \mid Z = z, A = a\}$$

In the above definition, $\mathcal{H}_{A=a}$ is to be interpreted as the outcome of the predictor $\mathcal{H}$ if A had taken value a. This measure compares every individual with a different version of themselves and deems a predictor to be fair if its output remains the same when the protected attribute is flipped to its counterfactual value. A similar measure was introduced independently by Kilbertus et al. (27).

In the literature of social sciences, the closest correspondent to these measures is the theory for counterfactual reasoning given by Lewis (29). There has been research to indicate that counterfactual reasoning is susceptible to hindsight bias (36; 41) and outcome bias (i.e. evaluating the quality of a decision when its outcome is already known) (5). Moreover, it has been argued that counterfactual reasoning may negatively influence the process of identifying causality (1; 10).

These critiques bring into question the suitability of counterfactual measures for potential domains for prediction using machine learning like health-care system or judicial system where the above-mentioned biases are frequently observed.

2.3. Group fairness (Statistical/demographic parity)

Group fairness imposes the condition that the predictor should predict a particular outcome for individuals across groups with *almost* equal probability.

Definition 3 *(Group fairness) A predictor $\mathcal{H} : X \rightarrow Y$ achieves group fairness with bias ϵ with respect to groups $S, T \subseteq X$ and $O \subseteq A$ being any subset of outcomes iff*

$$|\mathbb{P}\{\mathcal{H}(x_i) \in O \mid x_i \in S\} - \mathbb{P}\{\mathcal{H}(x_j) \in O \mid x_j \in T\}| \leq \epsilon$$

From the above definition it is clear that, group fairness imposes the condition of statistical and demographic parity on the predictor. Unlike some of the other formalizations of fairness, group fairness is independent of the "ground truth" i.e. the label information. This is useful when reliable ground truth information is not available e.g. in domains like employment, housing, credit and criminal justice, discrimination against protected groups has been well-documented (33; 47). Alternatively, in the cases where disproportionality in the respective outcomes can be justified by using non-protected attributes (which don't merely serve as a proxy for protected attributes), imposing statistical parity leads to incorrect outcomes and may amount to discrimination against qualified candidates (30). Another deficiency of group fairness is that the predictor is not stipulated to select the most "qualified" individuals within the groups as long as it maintains statistical parity (16).

The formalization of group fairness follows from the notion of *collectivist egalitarianism* for distributive justice. In practice, the biggest (in terms of the number of people affected) implementation of group fairness is the application of affirmative action (12) in India and USA to address discrimination on the basis of caste (15), race and gender.

See Weisskopf (48) for arguments made for and against affirmative action polices in both India and the USA. Two of the standard objections to group fairness are: it is not meritocratic and it reduces efficiency.

The underlying assumption behind the first claim is that the allocation of social benefits without affirmative action is meritocratic. However, several studies (13; 6; 34) have confirmed discrimination on the basis of protected attributes. For the second claim, Holzer and Neumark (24) conclude on the basis of several studies that "the empirical case against Affirmative Action on the grounds of efficiency is weak at best". In India, a study by Deshpande and Weisskopf (14) found no evidence of loss in efficiency because of affirmative action policies. Nonetheless, deficiencies mentioned earlier limit the applicability of group fairness.

2.4. Individual fairness

Individual fairness ascertains that a predictor is fair if it produces similar outputs for similar individuals.

Definition 4 *(Individual fairness) A predictor achieves individual fairness iff $\mathcal{H}(x_i) \approx \mathcal{H}(x_j) \mid d(x_i, x_i) \approx 0$ where $d : X \times X \to \mathbb{R}$ is a distance metric for individuals.*

Several works including Dwork et al. (16) and Luong et al. (30) use this notion. The notion of individual fairness can be captured by (D, d)-Lipschitz property which states that $D(\mathcal{H}(x_i)_Y, \mathcal{H}(x_j)_Y) \leq d(x_i, x_j)$ where D is a distance measure for distributions. Furthermore, Dwork et al. (16) prove that any predictor satisfying (D, d)-Lipschitz property also achieves statistical parity with certain bias.

In the social sciences literature, this formalization is equivalent to *individualism egalitarianism*. According to Sacksteder (42), this is the formal principle of justice. This notion delegates the responsibility of ensuring fairness from the predictor to the distance metric. If the distance metric uses the protected attributes directly or indirectly to compute the distance between two individuals, a predictor satisfying Definition 4 could still be discriminatory. Therefore, the potency of this notion of fairness to prevent discrimination depends largely upon the distance metric used. Hence, individual fairness as stated above, can not be considered suitable for domains where reliable and non-discriminating distance metric is not available [1].

2.5. Equality of opportunity

In the literature of machine learning, the formalization of equality of opportunity was introduced by Hardt et al. (22). An equivalent formalization was also proposed concurrently and independently by Zafar et al. (50). To formalize it, let us consider the case of binary outcomes with a single beneficial

---

[1] Dwork et al. (16) have provided some approaches to build distance metrics.

outcome $y = 1$.

Definition 5 *(Equal opportunity) A predictor is said to satisfy equal opportunity with respect to group S iff $\mathbb{P}\{\mathcal{H}(x_i) = 1 \mid y_i = 1, x_i \in S\} = \mathbb{P}\{\mathcal{H}(x_j) = 1 \mid y_j = 1, x_j \in X \setminus S\}$.*

It can be considered as a stipulation which states that the true positive rate should be the same for all the groups. An equivalent notion proposed by Zafar et al. (50), called *disparate mistreatment*, asks for the equivalence of misclassification rates across the groups.

In the social sciences literature, the corresponding notion was presented by Rawls (40). An essay by Arneson (3) states that equality of opportunity would not be able to cope with the problems of *stunted ambition* and *selection by bigotry*. The notion of equality of opportunity has also been criticized for not considering the effect of discrimination due to protected attributes like gender (32) and race (45). which have been shown to affect one's access to opportunities in domains such as education, business, politics in many parts of the world (25). The exclusion of attributes like race and gender from the list of attributes deemed to be affecting an individual's life prospects in the notion of equality of opportunity thus calls into question its suitability to the domains in which there exists vast evidence that such attributes do indeed affect one's prospects.

2.6. Preference-based fairness

Zafar et al. (49) introduce two preference-based formalizations of fairness. In order to provide the definitions for the same, the authors first introduce the notion of *group benefit* which is defined as the expected proportion of individuals in the group for whom the predictor predicts the beneficial outcome. Group benefit can also be defined as the expected proportion of individuals from the group who receive the beneficial output for whom the true label is the same. Based on the above notion of group benefit, Zafar et al. (49) provide following two fairness formalizations.

Definition 6 *(Preferred treatment) A group-conditional predictor is said to satisfy preferred treatment if each group receives more benefit from their respective predictor than they would have received from any other predictor i.e.*

$$\mathbb{B}_S(\mathcal{H}_S) \geq \mathbb{B}_S(\mathcal{H}_T) \qquad \text{for all } S, T \subset X$$

Definition 7 *(Preferred impact) A predictor $\mathcal{H}$ is said to have preferred impact as compared to another predictor $\mathcal{H}'$ if $\mathcal{H}$ offers at-least as much benefit as $\mathcal{H}'$ for all the groups.*

$$\mathbb{B}_S(\mathcal{H}) \geq \mathbb{B}_S(\mathcal{H}') \qquad \text{for all } S \subset X$$

If a classifier is not group-conditional then, it by default satisfies preferred treatment. In certain applications, there might not be a single universally accepted beneficial outcome. It is possible that a few individuals from a group may prefer another outcome than the one preferred by the ma-

jority of the group. In order to alleviate their concerns, the collectivist definition of group benefit needs to be extended to account for individual preferences.

In the social sciences literature, the above notion corresponds to *envy-freeness* (4). This notion of fairness is attractive because it can be defined in terms of ordinal preference relations of the utility values of the predictors. On the other hand, Holcombe (23) show that freedom from envy is neither necessary nor sufficient for fairness. For many real-world problems, one needs to find fair and efficient solutions amongst the groups. An efficient solution ensures the greatest possible benefit to the groups. In decision making problems, like the domain applications of prediction with machine learning, it can be formally expressed by the notion of *Pareto-efficiency*. However, deciding whether there is a Pareto-efficient envy-free allocation is computationally very hard even with simple additive preferences (11).

These critiques indicate that the suitability of such envy-free formalizations is limited only to the domains where such an effective and envy-free allocation can be computed easily.

3. Prospective notions of fairness

In this section, we describe two prospective notions of fairness which have not been considered in the literature of machine learning so far. Our intent is to address the critique that many of the past formalizations, as seen in Section 2, do not offset for the fact that social benefits are being allocated unequally by the algorithms among the people owing to the attributes they had no say in.

- **Equality of resources:** Dworkin (17) propose the notion of *equality of resources* in which unequal distribution of social benefits is only considered fair when it results from the intentional decisions and actions of the concerned individuals. Equality of resources is *ambition-sensitive* i.e. each individual's ambitions and choices that follow them ascertains the benefits they receive and *endowment-insensitive* i.e. each individual's unchosen circumstances including the natural endowments should be offset. In the second property, equality of resources differs from equality of opportunity as the latter considers differences in natural endowments (including the protected attributes such as sex) as facts of nature which need not be adjusted to achieve fairness.

- **Equality of capability of functioning:** Sen (44) extends the insight that people should not be held responsible for attributes they had no say in to include personal attributes which cause difficulty in developing *functionings*. Functionings are states of "being and doing", that is, various states of existence and activities that an individual can undertake. Sen (43; 44) argue that variations related to the protected attributes like

age, sex, gender, race, caste give individuals unequal powers to achieve goals even when they have the same opportunities. In order to equalize capabilities, people should be compensated for their unequal powers to convert opportunities into functionings. To this point, it sounds similar to quality of resources described above. Crucially however, the notion of equality of capability calls for addressing inequalities due to social endowments (e.g. gender) as well as natural endowments (e.g. sex) , in contrast to the equality of resources (37).

One of the main strengths of this notion of fairness that it is flexible which allows it to be developed and applied in many different ways (2). Indeed, this notion has been used in the foundations of human development paradigm by the United Nations (19; 20). One of the major criticism of Equality of capability theory concerns the failure to identify of valuable capabilities (38). Another criticism is that the informational requirement of this approach can be very high (2). The second criticism applies to equality of resources as well and it makes exact mathematical formalizations of these notions a potentially difficult problem. However, the suitability of these prospective formalizations (unlike the current formalizations) to domains in which natural endowments or social endowments or both impede an individual's prospect to receive social benefits makes the open problem of formalizing them worthwhile. We intend this article to serve as a call for machine learning experts to work on formalizing them.

4. Discussion and further directions

As the field of fairness in machine learning prediction algorithms is evolving rapidly, it is important for us to analyze the fairness formalizations considered so far. To this end, we juxtaposed the fairness notions previously considered in the machine learning literature with their corresponding theories of distributive justice in the social sciences literature. We saw theoretical critique and analysis of these fairness notions from the social sciences literature. Such critiques of the formalizations and experimental studies of their use in large-scale practice serve as guiding principles while choosing the fairness formalizations to use in particular domains.

We also proposed two prospective notions of fairness, which have been studied extensively in the social sciences literature. Of course, we do not claim that these notions will serve as panacea for all the critiques of the current notions. Our intention is to initiate a discussion about fairness formalizations in prediction with machine learning which recognize that - **the problem of fair prediction cannot be addressed without considering social issues such as unequal access to resources and social conditioning. While these factors are difficult to quantify and formalize mathematically, it is important to acknowledge their impact and attempt to incorporate them in fairness formalizations.**

References

[1] 1997. Counterfactual Thinking: The Intersection of Affect and Function. Advances in Experimental Social Psychology, Vol. 29. Academic Press, 1 – 59.

[2] Sabina Alkire. 2002. *Valuing Freedoms: Sen's Capability Approach and Poverty Reduction*. Oxford University Press.

[3] Richard J. Arneson. 1999. Against Rawlsian Equality of Opportunity. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 93, 1 (1999), 77–112.

[4] Christian Arnsperger. 1994. Envy-Freeness and Distributive Justice. *Journal of Economic Surveys* 8, 2 (June 1994), 155–186.

[5] Jonathan Baron and John Hershey. 1988. Outcome Bias in Decision Evaluation. 54 (05 1988), 569–79.

[6] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review* 94, 4 (September 2004), 991–1013.

[7] Eduardo Bonilla-Silva. 2013. *Racism without Racists: Color-Blind Racism and the Persistence of Racial Inequality in the United States* (4th ed.). Rowman & Littlefield Publishers.

[8] Toon Calders and Sicco Verwer. 2010. Three Naive Bayes Approaches for Discrimination-free Classification. *Data Min. Knowl. Discov.* 21, 2 (Sept. 2010), 277–292.

[9] Toon Calders and Indrė Žliobaitė. 2013. *Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures*. Springer Berlin Heidelberg, Berlin, Heidelberg, 43–57.

[10] A. P. Dawid. 2000. Causal Inference without Counterfactuals. *J. Amer. Statist. Assoc.* 95, 450 (2000), 407–424.

[11] Bart de Keijzer, Sylvain Bouveret, Tomas Klos, and Yingqian Zhang. 2009. *On the Complexity of Efficiency and Envy-Freeness in Fair Division of Indivisible Goods with Additive Preferences*. Springer Berlin Heidelberg, Berlin, Heidelberg, 98–110.

[12] Ashwini Deshpande. 2013. *Affirmative Action in India*. Oxford University Press, New Delhi.

[13] Ashwini Deshpande and Katherine Newman. 2007. Where the Path Leads: The Role of Caste in Post-University Employment Expectations. *Economic and Political Weekly* 42, 41 (2007), 4133–4140.

[14] Ashwini Deshpande and Thomas E. Weisskopf. 2016. Affirmative Action and Productivity in the Indian Railways. *The Review of Black Political Economy* 43, 2 (01 Jun 2016), 245–251.

[15] Louis Dumont. 1980. *Homo Hierarchicus: The Caste System and Its Implications*. The University of Chicago Press.

[16] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)*.

[17] Ronald Dworkin. 1981. What is Equality? Part 2: Equality of Resources. *Philosophy and Public Affairs* 10, 4 (1981), 283–345.

[18] Roland Fryer, G. Loury, and T. Yuret. 2008. An Economic Analysis of Color-Blind Affirmative Action. *Journal of Law, Economics, and Organization* 24, 2 (2008), 319–355.

[19] Sakiko Fukuda-Parr. 2003. THE HUMAN DEVELOPMENT PARADIGM: OPERATIONALIZING SEN'S IDEAS ON CAPABILITIES. *Feminist Economics* 9, 2-3 (2003), 301–317.

[20] Sakiko Fukuda-Parr and A.K. Shiva Kumar. 2003. *Readings in human development : concepts, measures and policies for a development paradigm*. Oxford University Press.

[21] Claudia Goldin and Cecilia Rouse. 2000. Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians. *American Economic Review* 90, 4 (September 2000), 715–741.

[22] Moritz Hardt, Eric Price, , and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*.

[23] Randall G. Holcombe. 1997. Absence of Envy Does Not Imply Fairness. *Southern Economic Journal* 63, 3 (1997), 797–802.

[24] Harry Holzer and David Neumark Neumark. 2000. Assessing Affirmative Action. *Journal of Economic Literature* 38, 3 (2000), 483–568.

[25] Sarah Iqbal. 2015. Women, business, and the law 2016 : getting to equal. (2015).

[26] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-Aware Classifier with Prejudice Remover Regularizer. In *Proceedings of the 2012 European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part II (ECML PKDD'12)*. 35–50.

[27] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding Discrimination through Causal Reasoning. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 656–666.

[28] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4066–4076.

[29] David Lewis. 1973. Causation. *Journal of Philosophy* 70, 17 (1973), 556–567.

[30] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. 2011. k-NN As an Implementation of Situation Testing for Discrimination Discovery and Prevention. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11)*.

[31] Razieh Nabi and Ilya Shpitser. 2018. Fair Inference on Outcomes. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*.

[32] Susan Moller Okin. 1991. Justice, Gender, and the Family. *Philosophy and Public Affairs* 20, 1 (1991), 77–97.

[33] Devah Pager and Hana Shepherd. 2008. The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets. *Annual Review of Sociolgy* 34 (2008), 181–209.

[34] Devah Pager and Bruce Western. 2006. Race at Work: Realities of Race and Criminal Record in the New York City Job Market. (2006 2006).

[35] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware Data Mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '08)*.

[36] John V. Petrocelli and Steven J. Sherman. 2010. Event detail and confidence in gambling: The role of counterfactual thought reactions. *Journal of Experimental Social Psychology* 46, 1 (2010), 61 – 72.

[37] Roland Pierik and Ingrid Robeyns. 2007. Resources versus Capabilities: Social Endowments in Egalitarian Theory. *Political Studies* 55, 1 (2007), 133–152.

[38] Mozaffar Qizilbash. 1996. Capabilities, wellbeing and human development: A survey. *The Journal of Development Studies* 33, 2 (1996), 143–162.

[39] Bilal Qureshi, Faisal Kamiran, Asim Karim, and Salvatore Ruggieri. 2016. Causal Discrimination Discovery Through Propensity Score Analysis. *CoRR* abs/1608.03735 (2016).

[40] John Rawls. 1971. *A Theory of Justice*. Harvard University Press.

[41] Neal J. Roese and James M. Olson. 1996. Counterfactuals, causal attributions, and the hindsight bias: A conceptual integration. *Journal of Experimental Social Psychology* 32, 3 (1 1 1996), 197–227.

[42] William Sacksteder. 1964. The Idea of Justice and the Problem of Argument. Chaim Perelman. *Ethics* 75, 1 (1964), 66–67.

[43] Amartya Sen. 1990. Justice: Means versus Freedoms. *Philosophy and Public Affairs* 19, Spring (1990).

[44] Amartya Sen. 1992. *Inequality Reexamined*. Clarendon Press, Oxford. New York: Russell Sage Foundation, and Cambridge. MA: Harvard University Press, 1992; Italian translation: Il Mulino, 1994; French translation: Seuil, 2000.

[45] Seana Valentine Shiffrin. 2004. Race and Ethnicity, Race, Labor, and the Fair Equality of Opportunity Principle. *Fordham Law Review* 72 (2004). Issue 5.

[46] Andrew Taslitz. 2007. Racial Blindsight: The Absurdity of Color-Blind Criminal Justice. *Ohio State Journal of Criminal Law* (2007).

[47] Kaveh Waddell. 2016. How Algorithms Can Bring Down Minorities' Credit Scores. (2 December 2016).

[48] Thomas E. Weisskopf. 2004. *Affirmative action in the United States and India : a comparative perspective / Thomas E. Weisskopf*. Routledge London ; New York. xvi, 286 p. ; pages.

[49] M. Zafar, I. Valera, M. Gomez Rodriguez, K. P. Gummadi, and A. Weller. 2017. From Parity to Preference-based Notions of Fairness in Classification. *ArXiv e-prints* (June 2017).

[50] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification Without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. 1171–1180.