Algorithmic fairness

Salvatore Ruggieri, salvatore.ruggieri@unipi.it

Fairness

Increasingly sophisticated algorithms from Artificial Intelligence (AI) and Machine Learning (ML) support knowledge discovery from big data of human activity.

They enable the extraction of patterns and profiles of human behavior which are able to make extremely accurate predictions. Decisions are then being partly or fully delegated to such algorithms for a wide range of socially sensitive tasks: personnel selection and wages, credit scoring, criminal justice, assisted diagnosis in medicine, personalization in schooling, sentiment analysis in texts and images, people monitoring through facial recognition, news recommendation, community bulding in social networks, dynamic pricing of services and products.

The benefits of algorithmic-based decision making cannot be neglected, e.g., procedural regularity – same procedure applied to each data subject. However, automated decisions based on profiling or social sorting may be biased[1] for several reasons. Historical data may contain human (cognitive) bias and discriminatory practices that are endemic, to which the algorithms assign the status of general rules. Also, the usage of AI/ML models reinforces such practices because data about model's decisions become inputs in subsequent model construction (feedback loops).

Algorithms may wrongly interpret spurious correlations in data as causation, making predictions based on ungrounded reasons. Moreover, algorithms pursue the optimization of quality metrics, such as accuracy of predictions, that favor precision over the majority of people against small groups. Finally, the technical process of designing and deploying algorithms is not yet mature and standardized. Rather, it is full of small and big decisions (sometimes, trial and error steps) that may hide bias, such as selecting non-representative data, performing overspecialization of the models, ignoring socio-technical impacts, or using models in deployment contexts they are not tested for. These risks are exacerbated by the fact that the AI/ML models are complex for human understanding, or not even intelligible, sometimes they are based on randomness or time-dependent non-reproducible conditions.[2]

Algorithmic fairness is the absence of bias in automated decision making. The

[1] Eirini Ntoutsi et al. 'Bias in data-driven Artificial Intelligence systems - An introductory survey'. In: *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 10.3, 2020.

[2] Joshua A. Kroll et al. 'Accountable Algorithms'. In: *U. of Penn. Law Review* 165, 2017, pp. 633–705.

most relevant case of bias is discrimination against protected-by-law social groups in decision making (see book entry *Discrimination data analysis*).

Fair algorithms are designed with the purpose of preventing biased decisions in algorithmic decision making. The naïve approach of deleting attributes that denote protected groups from the original dataset (fairness through unawareness) does not prevent a model from indirectly learning discriminatory decisions, since other attributes (called redundant encodings) that are strongly correlated with them could be used as proxies by the AI/ML algorithms. Restrictions on automated decision-making are provided by the EU General Data Protection Regulation, which states (article 22) "the right not to be subject to a decision based solely on automated processing".

Moreover, (recital 71) "in order to ensure fair and transparent processing in respect of the data subject [...] the controller should use appropriate mathematical or statistical procedures [...] to prevent, inter alia, discriminatory effects on natural persons". Several initiatives have started to audit, standardize and certify algorithmic fairness, such as the ICO Draft on AI Auditing Framework[3], the draft IEEE P7003™ Standard on Algorithmic Bias Considerations[4], and the IEEE Ethics Certification Program for Autonomous and Intelligent Systems[5].

Fairness measures

The design of algorithms assumes requirements on the expected functionalities (what the system is supposed to do) and qualities (what the system is supposed to be). Fairness is a non-functional requirement on algorithms. In order to enforce fairness, or even test it, designers need a measurable definition of fairness. Quantitative definitions have been introduced in philosophy, economics, and machine learning in the last 50 years,[6,7] with more than 20 different definitions of fairness appeared thus far in the computer science literature.[8,9]

[3] https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-consultation-on-the-draft-ai-auditing-framework-guidance-for-organisations

[4] https://standards.ieee.org/project/7003.html

[5] https://standards.ieee.org/industry-connections/ecpais.html

[6] Ben Hutchinson, Margaret Mitchell. '50 Years of Test (Un)fairness: Lessons for Machine Learning'. In: *FAT*. ACM, 2019, pp. 49–58.

[7] Reuben Binns. 'Fairness in Machine Learning: Lessons from Political Philosophy'. In: *FAT*. vol. 81. Proceedings of Machine Learning Research. PMLR, 2018, pp. 149–159.

[8] Ninareh Mehrabi et al. 'A Survey on Bias and Fairness in Machine Learning'. In: *CoRR* abs/1908.09635, 2019.

[9] Indre Zliobaite. 'Measuring discrimination in algorithmic decision making'. In: *Data Min. Knowl.*

Existing fairness definitions can be categorized[10] into: (i) "predicted outcome", (ii) "predicted and actual outcome", (iii) "predicted probabilities and actual outcome", (iv) "similarity based", and (v) "causal reasoning". We present them by using a statistical notation. Let G be a protected ground (or sensitive attribute), say gender, with values f (protected group) and m (unprotected group). Let d be the decision of the algorithm, with values $+$ (positive) and $-$ (negative). Such a decision may come with a score S, a number between 0 and 1 stating the confidence in the prediction. Finally, let T be the ground-truth decision, if available (we will discuss this concept later on).

"Predicted outcome" definitions solely rely on an algorithm's decisions. For example, *group fairness* (also called, *demographic parity* or *statistical parity*) requires the probabilities of positive decisions for protected and unprotected groups to be the same, or, in formula $P(d = + \mid G = f) = P(d = + \mid G = m)$. If it is less likely that females are assigned a positive decision than male, this is interpreted as unfairness. This measure originates from studies on disparate impact discrimination.[11]

"Predicted and actual outcome" definitions combine an algorithm's decision with the true decision. For instance, *predictive parity* requires the probability of correct positive predictions to be the same for both groups, namely $P(T = + \mid d = +,\ G = f) = P(T = + \mid d = +,\ G = m)$.

"Predicted probabilities and actual outcome" definitions refer to the predicted probabilities instead of the predicted outcomes. For example, *well calibration* requires the probability of positive predictions to be equally calibrated for both groups, namely $P(T = + \mid S = s,\ G = f) = P(T = + \mid S = s,\ G = m) = s$. Calibration is a desirable property of scoring algorithms, meaning that the probability of positive decision is proportional to the scoring value, i.e., $P(T = + \mid S = s)$.

The previous definitions do not consider attributes used by algorithms in decision making, except G. This may hide unfairness if the data distribution differs among groups, e.g., if male applicants are less qualified than females in the available data.

"Similarity based" definitions (or *individual fairness*) instead employ other attributes which are relevant for decision making. They originate from the formal equality principle to "treat like cases as like", also known, in its negative form, as *disparate treatment* discrimination. They assume a distance function $d(\mathbf{x}, \mathbf{y})$ which

Discov. 31.4, 2017, pp. 1060–1089.

[10] Sahil Verma, Julia Rubin. 'Fairness definitions explained'. In: *FairWare@ICSE*. ACM, 2018, pp. 1–7.

[11] Andrea Romei, Salvatore Ruggieri. 'A multidisciplinary survey on discrimination analysis'. In: *Knowledge Eng. Review* 29.5, 2014, pp. 582–638.

measures the dissimilarity between vector of individuals' characteristics $\mathbf{x}$ and $\mathbf{y}$, except G. It consists of a non-negative real number, close to 0 when the two individuals are highly similar or 'near' each other's, and becoming larger the more they differ. *Fairness through awareness* states that $d(\mathbf{x}, \mathbf{y})$ bounds the distance between decisions $Y_{\mathbf{x}}$ and $Y_{\mathbf{y}}$ (or between scores $S_{\mathbf{x}}$ and $S_{\mathbf{y}}$) for individuals $\mathbf{x}$ and $\mathbf{y}$. For instance, if two individuals are equally qualified for obtaining a loan (zero distance), then there should be no difference between their predictions or scores – independently from their protected attribute G.

Finally, "causal reasoning" definitions are based on statistical models of cause-effect, such as directed acyclic graphs which capture relations between attributes and their impact on the decision outcomes. For example, counterfactual fairness verifies whether the decision d in the causal graph is not affected by a change of the protected attribute G. Causal graphs can be (partially) inferred from data but need domain expert interventions for their validation.

Despite the many definitions of fairness proposed, the choice of the most appropriate measure in a given application context is left open,[12] and it can only be made as the result of involving humans in the loop.[13] For instance, which individual's characteristics should be considered in the similarity-based measures? And, how should the distance measure be defined beyond blindly choosing one offered by the tool at hand? Another critical point for measures which rely on the true decision T, is how to collect ground-truth. Very few datasets have such an information. The COMPAS dataset[14] for the analysis of recidivism risk scoring approximates the ground-truth by a 2-year look-ahead of the behavior of a defendant. In most cases, however, there is no counterfactual to look at. We do not know, for instance, if an applicant with rejected loan request would have repaid the loan. Practitioners may be tempted to use data on past loans as ground-truth. But such data is biased, as only accepted loans are monitored. Hence, the risk is to reinforce the bias already present in past data.

Fair algorithms

Four non-mutually exclusive strategies can be devised for fairness-by-design of AI/ML

[12] Alexandra Chouldechova, Aaron Roth. 'A snapshot of the frontiers of fairness in machine learning'. In: *Commun. ACM* 63.5, 2020, pp. 82–89.

[13] Nripsuta Ani Saxena et al. 'How do fairness definitions fare? Testing public attitudes towards three algorithmic definitions of fairness in loan allocations'. In: *Artif. Intell.* 283, 2020, p. 103238.

[14] https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis

models.

Pre-processing approaches. The first strategy consists of a controlled sanitization of the data used to train an AI/ML model with respect to specific biases. The methodologies are similar to the ones used for preserving privacy in data release. They include changing historical decision outcomes to correct biased decisions, reweighting or sampling data to balance data distributions, or abstracting values to ranges to mask bias in some specific contexts. Pre-processing approaches allow for obtaining less biased data, which can be released (as in the case of official statistics) or used for training AI/ML models. The latter case alleviates for the lack of ground-truth availability (the true decision T). Another advantage of pre-processing approaches is that they are independent of the AI/ML model and algorithm at hand. A pertinent legal question is whether modifications of data could be considered lawful, especially in the case of personal data.

In-processing approaches. The second strategy is to modify the AI/ML algorithm, by incorporating fairness criteria in model construction, such as regularizing the optimization objective with a fairness measure. Impossibility results,[15] however, state that any two statistical fairness measures cannot be satisfied at the same time. Hence, the strategy must be specific of a given fairness measure. There is a fast growing adoption of in-processing approaches in many AI/ML problems other than in the original setting of classification, including ranking, clustering, community detection, influence maximization, distribution/allocation of goods, and models on non-structured data such as natural language texts and images. In dynamic settings such as online learning, bandit learning, and reinforcement learning, the actions of the fair algorithms feed back into the data it observes subsequently, thus taking into account feedback loops. In causal fairness, the decision-making model is trained to exploit causal dependencies between people's characteristics and decisions, instead of correlations.

An area somehow in the middle between pre-processing and in-processing approaches is fair representation learning, where the model inferred from data is not used directly for decision making, but rather as intermediate knowledge. For example, fair variational auto encoders map texts into a latent space (which can be used for text classification or machine translation) by constraining the mapping not to be subject to a sensitive attribute. This could be done by adversarial de-biasing: train at the same time a representation model and an adversarial model that tries and reconstruct from the representation a sensitive attribute, in such a way that the adversarial

[15] Alexandra Chouldechova. 'Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments'. In: *Big Data* 5.2, 2017, pp. 153–163.

model fails its goal.

Post-processing approaches. The third strategy is to post-process the AI/ML model once it has been computed, so to identify and remove unfair decision paths. This can be achieved also by involving human experts (human-in-the-loop[16]) in the exploration and interpretation of the model (if it is interpretable) or of the model's decisions. Post-processing approaches consist of altering the model's internals, for instance by correcting the confidence of classification rules, or the probabilities of Bayesian models, or by simply re-training the model with adjusted parameters. Post-processing becomes necessary for tasks for which there is no in-processing approach explicitly designed for the fairness requirement at hand.

Prediction-time approaches. The last strategy assumes no change in the construction of AI/ML models, but rather correcting their predictions at run-time. Proposed approaches include promoting, demoting or rejecting predictions close to the decision boundary, differentiating the decision boundary itself over different social groups, or wrapping a fair classifier on top of a black-box base classifier. Such approaches may be applied to legacy software, including non-AI/ML algorithms, that cannot be replaced by in-processing or changed by post-processing approaches.

Fairness auditing, discrimination discovery, explainability

The goal of auditing AI/ML models is to test fairness of their decisions. Approaches can vary widely on the basis of the information available. On one extreme, there are mathematical proofs of fairness using formal methods.[17]

This requires the disclosure of the AI/ML model and its representability in some formal language. If any of the two is not possible, but the model can be queried at will, inferences on the model decision distribution can be made. This line of research share methods with the area of explainability[18,19] of AI/ML black boxes (see book entry *Explainability*). Both provide surrogate models, feature importance weights, decisions rules, or other outputs that describe the behavior of the AI/ML

[16] Bettina Berendt, Sören Preibusch. 'Toward Accountable Discrimination-Aware Data Mining: The Importance of Keeping the Human in the Loop - and Under the Looking Glass'. In: *Big Data* 5.2, 2017, pp. 135–152.

[17] Aws Albarghouthi. 'Fairness: A Formal-Methods Perspective'. In: *SAS*. vol. 11002. Lecture Notes in Computer Science. Springer, 2018, pp. 1–4.

[18] Riccardo Guidotti et al. 'A Survey of Methods for Explaining Black Box Models'. In: *ACM Comput. Surv.* 51.5, 2019, 93:1–93:42.

[19] Alejandro Barredo Arrieta et al. 'Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI'. in: *Inf. Fusion* 58, 2020, pp. 82–115.

model globally, at the decision boundary, or for a specific input instances.

Here, the interest is in detecting the influence of protected attributes on a model decision. Finally, there are cases when the algorithm cannot be accessed at all (such as in criminal justice or in tax assessment or in web user profiling), but its input/outputs can be collected. Here, the approaches for discrimination discovery from data[20] can be applied.

An intriguing question is how to compare the relative (un)fairness of two competing AI/ML models, in particular, when they refer to different fairness measures. A possible answer relies on using inequality indices from economics and social welfare.[21]

[20] Andrea Romei, Salvatore Ruggieri. Ibid.

[21] Till Speicher et al. 'A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual & Group Unfairness via Inequality Indices'. In: *KDD*. ACM, 2018, pp. 2239–2248.