Opening the black box: a primer for anti-discrimination

Salvatore Ruggieri*, Fosca Giannotti, Riccardo Guidotti,
Anna Monreale, Dino Pedreschi, and Franco Turini

KDD-Lab, ISTI-CNR and University of Pisa, Italy

*Corresponding author: salvatore.ruggieri@unipi.it

SOMMARIO

L'uso pervasivo dell'Intelligenza Artificiale (AI) nella moderna società dell'informazione richiede di controbilanciare il potere decisionale ad essa delegato con opportune valutazioni del rischio. In questo lavoro consideriamo il rischio di decisioni discriminatorie e le metodologie per condurre audit e per progettare modelli di AI che siano fair by-design. In particolare, studiamo le relazioni tra l'analisi di non-discriminazione e l'analisi delle spiegazioni delle decisioni prese da un modello di AI, mostrando come queste ultime siano una generalizzazione delle prime.

ABSTRACT

The pervasive adoption of Artificial Intelligence (AI) models in the modern information society, requires counterbalancing the growing decision power demanded to AI models with risk assessment methodologies. In this paper, we consider the risk of discriminatory decisions and review approaches for discovering discrimination and for designing fair AI models. We highlight the tight relations between discrimination discovery and explainable AI, with the latter being a more general approach for understanding the behavior of black boxes.

KEYWORDS: Artificial Intelligence, Bias and discrimination, Explainable AI.

1. AI risks

Increasingly sophisticated Artificial Intelligence (AI[1]) algorithms support knowledge discovery from big data of human activity. They enable the extraction of patterns and profiles of human behavior (*AI models*) which are able to make extremely accurate predictions. AI models are thus becoming the backbone of private business and public policy decision-making processes in the modern information society. Decisions are being partly or fully delegated to them 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, friend suggestion in social networks, dynamic pricing of services and products, etc.

Although the benefits of AI cannot be neglected, AI-driven decisions based on profiling or social sorting may be biased[2] for several reasons. Historical data may contain human (cognitive) bias and discriminatory practices that are endemic in reality, to which the AI algorithm assigns the status of general rules. Also, the usage of AI models reinforces such practices because data about model's decisions become inputs in subsequent model construction (*feedback loops*). AI algorithms may wrongly interpret spurious correlations in data as causation, making predictions based on ungrounded reasons. Moreover, such 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 AI models 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, using models in deployment contexts they are not tested for, etc.[3] These risks are exacerbated by the fact that the AI models are complex for human understanding, or not even intelligible, sometimes they are based on randomness or time-dependent non-reproducible conditions.[4]

To counterbalance the increasing power granted to AI models, methodologies and tools for making them accountable are deemed necessary as legal and ethical

---

[1] S. SAMOILI et al. *Defining Artificial Intelligence. Towards an operational definition and taxonomy of Artificial Intelligence*. EUR 30117 EN, Publications Office of the European Union, JRC118163, 2020.
[2] E. NTOUTSI et al. *Bias in data-driven Artificial Intelligence systems - An introductory survey*. In: *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 10.3, 2020.
[3] D. DANKS, A. J. LONDON. *Regulating Autonomous Systems: Beyond Standards*. In: *IEEE Intell. Syst.* 32.1, 2017, pp. 88–91.
[4] J. A. KROLL et al. *Accountable Algorithms*. In: *U. of Penn. Law Review* 165, 2017, pp. 633–705.

requirements. Several initiatives[5] have started towards such an objective, such as the ICO Draft on AI Auditing Framework[6], the EU Ethics guidelines for trustworthy AI[7], the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems[8], including the IEEE P7000™ Standard drafts, the IEEE Ethics Certification Program for Autonomous and Intelligent Systems[9], and the Council of Europe study on Algorithms and Human Rights.[10]

In this paper, we consider the risk of discriminatory decisions and review approaches for discovering discrimination and for designing fair AI models. We highlight the tight relations between discrimination discovery and explainable AI, with the latter being a more general approach for understanding the behavior of black boxes.

2. Discrimination discovery and fairness in AI

Discrimination refers to an unjustified distinction of individuals based on their membership, or perceived membership, in a certain group or category. Some groups, traditionally subject to discrimination, are explicitly listed as "protected groups" by national and international human rights laws. Justified distinctions are exceptions explicitly admitted by law. The problems of assessing the presence, extent, nature, and trends of discrimination and of preventing discrimination in (possibly automated) decision-making have been investigated[11] from a social, legal, economic, and, in the last decade, from a computer science perspective.

Discrimination discovery consists in the actual discovery of discriminatory situations and practices hidden in a dataset of historical decision records, such as those generated by AI-based automated decision-making. The aim is to extract contexts of possible discrimination supported by legally grounded quantitative measures of the degree of discrimination suffered by protected-by-law groups in such contexts. Reasoning on the extracted contexts can support all the actors in an argument about possible discriminatory behaviors. The AI designer can use them both to assess AI models before deployment, or to argument against allegations of discriminatory

[5] https://www.oecd.org/going-digital/ai/initiatives-worldwide
[6] https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-consultation-on-the-draft-ai-auditing-framework-guidance-for-organisations
[7] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
[8] https://ethicsinaction.ieee.org
[9] https://standards.ieee.org/industry-connections/ecpais.html
[10] COMMITTEE OF EXPERTS ON INTERNET INTERMEDIARIES (MSI-NET). *Algorithms and Human Rights*. Council of Europe, 2018.
[11] A. ROMEI, S. RUGGIERI. *A multidisciplinary survey on discrimination analysis*. In: *Knowledge Eng. Review* 29.5, 2014, pp. 582–638.

behavior. A complainant in a case can use them to find specific situations in which there is a *prima facie* evidence of discrimination against groups she belongs to. Control authorities can base the fight against discrimination on a formalized process of intelligent data analysis.

The original paper[12] introducing discrimination discovery proposes the automatic extraction from a dataset of historical decision records, of classification rules of the form: PREMISES → DECISION. A rule is weighted by a confidence measure, stating the probability of the decision given the premises of the rule. For instance, the rule

RACE=AFR-AM., CITY=NYC → CREDIT=BAD [conf=0.75]

states that African-American applicants from NYC are assigned bad credit with a 75% probability. Three kinds of facts (items) are used in decision rules: potentially discriminatory items, such as RACE=AFR-AM., (potentially[13]) non-discriminatory items, such as CITY=NYC, and decision items, such as CREDIT=BAD. The potentially discriminatory items are specified with reference to a legal framework, to denote some designated groups of people protected by the anti-discrimination laws. The non-discriminatory items define the context where a discriminatory decision may take place - here, the set of applicants from the city of NYC.

In which circumstances does an extracted rule reveal a (possibly unintentional) discriminatory decision strategy? The idea here is to measure the discrimination of a rule by the gain of confidence due to the presence of the potentially discriminatory items in the premise of the rule. In the above example, we compare the 0.75 confidence of the rule RACE=AFR-AM., CITY=NYC → CREDIT=BAD with the confidence of the rule obtained by negating the first item, i.e., RACE≠AFR-AM., CITY=NYC → CREDIT=BAD. If, e.g., the confidence of the latter rule is 0.25, then we conclude that African-American applicants in NYC have a probability of being assigned bad credit which is 3 times larger than applicants from other social groups of NYC.

The ratio between the two confidences is called the risk ratio. It is a quantitative measure of *disparate impact* (or *group discrimination*) over the protected group by the decision-making process producing the historical data in input to the discrimination discovery analysis. In addition to risk ratio, risk difference and other algebraic variants,

[12] D. PEDRESCHI, S. RUGGIERI, F. TURINI. Discrimination-aware data mining. In: *KDD*. ACM, 2008, pp. 560–568.

[13] "Potentially" because such items may be correlated with discriminatory ones. For instance, the zip code of a neighborhood whose vast majority of inhabitants is African-American, can hide forms of indirect discrimination – *redlining* in the example of spatial segregation.

more than 20 measures[14] have been proposed to account for providing statistical confidence intervals, for dealing with multiple unprotected groups and/or multi-valued decisions (e.g., ratings), for continuous decision values (e.g., wages), for confounding factors that may justify large values of discrimination measures (such as genuine occupational requirements), and for eliminating spurious correlations in favor of causality of conclusions.

Decision rules can be ranked based on a reference measure of disparate impact to highlight the top-k contexts with the highest disproportionate burden imposed on protected groups. Very few legal cases exist, however, that refer to specific measures, e.g., the *fourth-fifth rule*[15] and the *Castaneda rule*[16] in the U.S. Unfortunately, the choice of the reference measure is non-trivial, and often, it is an under-evaluated aspect of the process. For instance, it has been shown[17] that the top-k contexts ranked accordingly to different measures may differ considerably.

The rule-based approach does not account for *disparate treatment* (or *individual discrimination*). In fact, there is no control within a context (e.g., applicants from NYC) of the characteristics of individuals in the protected group (e.g., African-American) as opposed to all other applications in that context. Approaches for individual discrimination discovery[18] rely on a distance measure. $d(\mathbf{x},\mathbf{y})$ measures the dissimilarity between vector of individuals' characteristics $\mathbf{x}$ and $\mathbf{y}$. 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. According to the formal equality principle to "treat like cases as like", the distance measure is used to compare the similarity of decisions for similar $\mathbf{x}$ and $\mathbf{y}$. For example, we may look for an individual $\mathbf{x}$ of a protected group with negative decision and such that most of its close neighbors not in the protected group are assigned a positive decision. A critical aspect is the definition of the distance function:[19] which individual characteristics are sufficient and necessary for taking a decision? what is the mathematical counterpart of similarity between individuals in different application contexts (personnel selection, credit scoring, etc.)?

---

[14] I. ZLIOBAITE. *Measuring discrimination in algorithmic decision making.* In: *Data Min. Knowl. Discov.* 31.4, 2017, pp. 1060–1089.
[15] Griggs v. Duke Power Co., 401 U.S. 424 (1971).
[16] Castaneda v. Partida, 430 U.S. 482 (1977).
[17] D. PEDRESCHI, S. RUGGIERI, F. TURINI. A study of top-k measures for discrimination discovery. In: *SAC*. ACM, 2012, pp. 126–131.
[18] L. ZHANG, Y. WU, X. WU. *Causal Modeling-Based Discrimination Discovery and Removal: Criteria, Bounds, and Algorithms*. In: *IEEE Trans. Knowl. Data Eng.* 31.11, 2019, pp. 2035–2050.
[19] A. CHOULDECHOVA, A. ROTH. *A snapshot of the frontiers of fairness in machine learning.* In: *Commun. ACM* 63.5, 2020, pp. 82–89.

A parallel stream of research in the AI community, particularly in machine learning, has been focusing on the design of AI models that account for non-discrimination by design: *fair AI models*. In fact, the naïve approach of deleting attributes that denote protected groups from the data used for training AI models (*fairness through unawareness*) does not prevent an AI algorithm from indirectly learning discriminatory decisions,[20] since other attributes that are strongly correlated with them could be used as proxies. Four non mutually-exclusive strategies have been considered for embedding fairness in AI models.[21] Pre-processing approaches consist of a controlled distortion of the training data with the intent to remove the bias in such data. In-processing approaches design (variants of) learning algorithms that optimize objective functions accounting for both accuracy and fairness of predictions. Post-processing approaches modify the AI model once it has been extracted with standard techniques, in order to identify and remove biased rules from it. Finally, run-time approaches act at prediction time by correcting predictions to keep proportionality of decisions among protected and unprotected groups.

As for discrimination discovery, a large number of quantitative definitions of fairness have been proposed, often rediscovering notions from other sciences.[22] The choice of the most appropriate measure of fairness in a given application context is left open,[23] and it can only be made as the result of multi-disciplinary collaboration.

3. Explainable AI

Explainability (or *explainable AI*[24,25]) refers to the extent the internal mechanics of an AI model can be explained in understandable terms to a human. It is often used interchangeably with interpretability. Explainability allows the AI designer to verify that AI models work as expected, in particular in compliance with the law,[26] making it possible to debug and improve AI models during development.

---

[20] I. ZLIOBAITE, B. CUSTERS. *Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models*. In: *Artif. Intell. Law* 24.2, 2016, pp. 183–201.
[21] N. MEHRABI et al. *A Survey on Bias and Fairness in Machine Learning*. In: *CoRR* abs/1908.09635, 2019.
[22] B. HUTCHINSON, M. MITCHELL. 50 Years of Test (Un)fairness: Lessons for Machine Learning. In: *FAT*. ACM, 2019, pp. 49–58.
[23] M. SRIVASTAVA, H. HEIDARI, A. KRAUSE. Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. In: *KDD*. ACM, 2019, pp. 2459–2468.
[24] R. GUIDOTTI et al. *A Survey of Methods for Explaining Black Box Models*. In: *ACM Comput. Surv.* 51.5, 2019, 93:1–93:42.
[25] A. B. ARRIETA et al. *Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI*. in: *Inf. Fusion* 58, 2020, pp. 82–115.
[26] G. MALGIERI, G. COMANDE´. *Why a Right to Legibility of Automated Decision-Making Exists in the GDPR*. in: *International Data Privacy Law* 7.4, 2017, pp. 243–265.

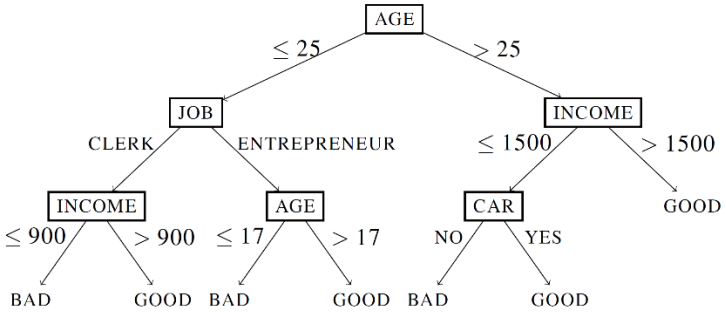
Two main streams of research are being pursued in the area of explainable AI. One includes approaches for understanding the *global* logic of an AI model by building an interpretable surrogate model able to mimic the obscure decision system (a form of reverse engineering). Interpretable models take the form[27] of decision rules, probabilistic (Bayesian) models, linear regression models, decision trees. The second stream focuses on the *local* behavior of a model, searching for an explanation of the decision made for a specific individual. Approaches can be also categorized as *model-dependent* or *model-agnostic*. The former methods apply to specific AI models. For instance, saliency masks (the regions of an image that are mainly responsible for the decision) are specific of (deep) neural networks. The latter methods apply to any AI model of a given family (classification, regression, ranking, clustering, etc.).

The blooming line of local model-agnostic explanations started with the LIME method.[28] The main idea is to randomly perturb the characteristics of an individual whose decision has to be explained, by generating several neighbor individuals. Starting from the decisions of the black box on the neighborhood, a local interpretable model can detect which characteristics mostly affect the decision value for the individual. LIME adopts a linear model able to weight features based on their importance in determining the decision. We will consider a variant of this approach next. Interestingly, we observe that local approaches rely on a distance function $d(\mathbf{x},\mathbf{y})$ between individuals' characteristics – as in the case of discrimination discovery for disparate treatment.

In general, an explanation can be derived through three forms of reasonings: [29] abduction (what is the most plausible reason given the data?), counterfactual reasoning (what would have happened for different data?), and prospective reasoning (what will happen for given data?). Regarding the central question "what is an explanation?", researchers can build on[30] philosophy, cognitive science, and social psychology. However, there is still no general consensus in the AI community. Counterfactual explanations, also called contrastive explanations, are deemed suitable[31] for a lay man. They show which characteristics of an individual should be changed in order to change the black box decision. This would allow, for

[27] C. MOLNAR. *Interpretable Machine Learning*. Lulu.com, 2019.
[28] M. T. RIBEIRO, S. SINGH, C. GUESTRIN. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: *KDD*. ACM, 2016, pp. 1135–1144.
[29] R. R. HOFFMAN, G. KLEIN. *Explaining Explanation, Part 1: Theoretical Foundations*. In: *IEEE Intelligent Systems* 32.3, 2017, pp. 68–73.
[30] T. MILLER. *Explanation in Artificial Intelligence: Insights from the social sciences*. In: *Artificial Intelligence* 267, 2019, pp. 1 –38.
[31] B. D. MITTELSTADT, C. RUSSELL, S. WACHTER. Explaining Explanations in AI. in: *FAT*. ACM, 2019, pp. 279–288.

Figure 1: Example of decision tree locally mimicking an AI black box model.



example, a bank customer whose loan application has been rejected to improve her application in an actionable recourse.[32]

The LORE[33] method provides both factual and counterfactual explanations in the form of decision rules. Consider a loan applicant with characteristics:

$$\mathbf{x} = \{ \text{AGE}=22, \text{JOB}=\text{CLERK}, \text{INCOME}=800, \text{CAR}=\text{NO} \}$$

whose application has been denied. A possible factual rule is:

$$\text{AGE} \leq 25, \text{JOB}=\text{CLERK}, \text{INCOME} \leq 900 \rightarrow \text{CREDIT}=\text{BAD}$$

This rule is obtained with a local model-agnostic approach, in the style of LIME, by generating a neighborhood of individuals close to the characteristics of **x**, with a balanced proportion of decisions CREDIT=BAD and CREDIT=GOOD. Starting from this dataset, an abductive reasoning approach is followed in search of the most plausible reason for the decision. This is done in two steps. First, data is generalized, through induction, to a decision tree (see Figure 1), where decision nodes distinguish the behavior of the AI black box model based on conditions on data values. The decision tree codes the possible reasons for the decisions of the AI model in the neighborhood of **x**. The second step consists of choosing the path in the decision tree satisfied as **x** as the most plausible reason for the black box decision. For the running example, it is the leftmost path of the decision tree in Figure 1. The factual rule above is a textual rewriting of such a path.

[32] S. VENKATASUBRAMANIAN, M. ALFANO. The philosophical basis of algorithmic recourse. In: *FAT**. ACM, 2020, pp. 284–293.
[33] R. GUIDOTTI et al. *Factual and Counterfactual Explanations for Black Box Decision Making*. In: *IEEE Intell. Syst.* 34.6, 2019, pp. 14–23.

Interestingly, counterfactual rules can be obtained from the same decision tree by following the paths that end in good credit decisions:

$$\underline{\text{AGE} > 25}, \underline{\text{INCOME} > 1500} \rightarrow \text{CREDIT}=\text{GOOD}$$

$$\underline{\text{AGE} > 25}, \text{INCOME} \leq 1500, \underline{\text{CAR}=\text{YES}} \rightarrow \text{CREDIT}=\text{GOOD}$$

$$\text{AGE} \leq 25, \text{JOB}=\text{CLERK}, \underline{\text{INCOME} > 900} \rightarrow \text{CREDIT}=\text{GOOD}$$

$$\text{AGE} \leq 25, \underline{\text{JOB}=\text{ENTREPRENEUR}} \rightarrow \text{CREDIT}=\text{GOOD}$$

Underlined conditions are not met by the applicant, who should change her characteristics **x** in order to reverse the black box decision. In particular, the first and the second counterfactual rules require the applicant to increase her age, which is not actionable. Hence, these rules should be filtered out. The third rule asks for greater income. The last rule requires a different job.

4. Closing the gap: explanations for discrimination discovery

What is the relation between discrimination discovery and explainable AI? Intuitively, discrimination discovery is a form of explanation of the decisions recorded in a dataset which focuses explicitly on individuals or groups protected by the law. The objective is to find out if a negative decision occurs disproportionately more often in comparison to unprotected social groups (group discrimination) or in comparison to individuals with similar characteristics (individual discrimination).

Let us focus on individual discrimination. Assume to use LORE for obtaining an explanation of the negative decision for an individual **x**, and that the factual rule returned is:

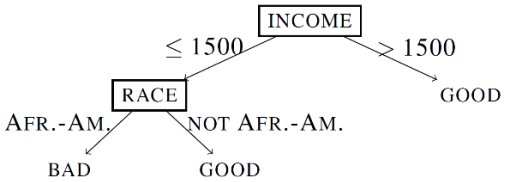
$$\text{RACE}=\text{AFR-AM.} \rightarrow \text{CREDIT}=\text{BAD}\ [\text{conf}=0.75]$$

and the counterfactual rule is:

$$\text{RACE}\neq\text{AFR-AM.} \rightarrow \text{CREDIT}=\text{GOOD}\ [\text{conf}=0.75]$$

where the confidence of the rules is also shown. Now, these rules have to be interpreted locally, i.e., among the individuals close to **x**, 75% of those who are African-American are assigned a bad credit score and 75% of those who are not are assigned a good credit score (or, equivalently, 25% of them are assigned a bad credit score). The risk ratio of confidence for bad credit is $0.75/0.25 = 3$, which appears sufficiently high for further investigation.

Figure 2: Decision tree explanations for discrimination discovery.



This example is very close to an earlier approach[34] for individual discrimination discovery, with the difference that the neighborhood is synthetically generated in the explanation approach whilst it was computed from existing data[35] in the discrimination discovery approach.

Assume now that LORE returns a slightly more intricate factual rule:

$$\text{RACE}=\text{AFR-AM.}, \text{INCOME}\leq 1500 \rightarrow \text{CREDIT}=\text{BAD}\ [\text{conf}=0.75]$$

It reads that, among the individuals close to **x**, those who are African-American and with income smaller or equal than 1500 are (in 75% of cases) assigned a bad credit score. The individual **x** belongs to such a group as well, and the rule provides a plausible reason for the black box decision. The factual rule may result from the decision tree[36] in Figure 2, from which we would also derive the following counterfactual rule:

$$\text{RACE}\neq\text{AFR.-AM.}, \text{INCOME} \leq 1500 \rightarrow \text{CREDIT}=\text{GOOD}\ [\text{conf}=0.75]$$

whose confidence is also shown. From this rule, we derive that the risk ratio of individuals close to **x** and having INCOME $\leq$ 1500, is $0.75/0.25 = 3$. This is a stronger form of discrimination discovery compared to the first case of this section. In fact, the pair of factual and counterfactual rules allows for detecting a specific context (INCOME $\leq$ 1500) where individual discrimination occurs. The context is specific in

[34] B. L. THANH, S. RUGGIERI, F. TURINI. k-NN as an implementation of situation testing for discrimination discovery and prevention. In: *KDD*. ACM, 2011, pp. 502–510.
[35] The advantage of this choice is that it does not require the availability of a black box to query, but only of a dataset of its past decisions.
[36] This is not the only possibility, as other decision trees may originate the same rule.

the sense that for the generality of the neighbors of **x**, the risk ratio measure could be lower and then not worth investigating further the case.

The above reasoning can be easily generalized to decision trees where the protected group attribute (race, gender, disability, etc.) occurs [37] only in the last decision nodes. Under this condition, the path from the root to the last decision node defines the context of local discrimination, and the counterfactual rule provide the information needed for computing the discrimination measure (risk ratio, risk difference, etc.).

The key finding that individual discrimination discovery is a "special case" of explainability is not specific of the LORE approach. By using the LIME method, for instance, we would obtain an explanation in terms of importance weights of individuals' attributes. Such weights are inferred through statistical regression of the black box decision based on individual's attributes. If the weight of the attribute denoting membership to a protected group is non-zero (and this is statistically significant), this means that the membership to a protected group significantly affects the decision of the black box. This is the typical approach used in economics for the analysis of discrimination in labor data[38].

There is also a (more direct) parallel between group discrimination discovery and global explanations. Recall that an approach for explaining the global behavior of a black box model is to build an interpretable model that mimics its decisions. Rule-based models, in particular, are interpretable models which consists of a set of decision rules together with a voting mechanism for making predictions based on such rules. If the rule set includes rules contrasting the decisions, in a same context, for the protected and the unprotected groups, we can directly compute from them the discrimination measures for that context.

5. Conclusion

We argued that explainable AI is, on the technical side, a more general problem than discrimination discovery. This fact does not solve the key issues in making technological solutions ready for deployment. Individual characteristics considered in the analyses should be relevant and complete for the decision at hand. Relevant means the characteristics recorded in the data are legally grounded for making the decision. Complete means that all legally grounded characteristics are recorded in the data.

[37] Decision tree building algorithms can be easily adapted to force the last decision node to test membership to a protected group.
[38] A. ROMEI, S. RUGGIERI. Ibid.

What is relevant and complete is domain specific, probably with no universally agreed answer, and, in practical situations, rarely available in collected data (e.g., for data protection limitations). In the case of group discrimination / global explanations, in addition, the approaches are parametric to a quantitative measure of discrimination which, in practical situations, should be grounded on legal basis. In the case of individual discrimination / local explanations, a distance measure between individual's characteristics has also to be chosen. How to translate differences between individuals into a mathematical formula is something that data scientists should not do by themselves.

Acknowledgements

This work is partially supported by the European Community H2020 programme under the funding schemes: MCSA-ITN G.A. 860630 *NoBIAS "Artificial Intelligence without Bias"* and the ERC-2018-ADG G.A. 834756 *"XAI: Science and technology for the eXplanation of AI decision making"*.