Understanding Fairness Requirements for ML-based Software

***Abstract*—Today's technologies are becoming more and more pervasive and advanced software systems can replace human beings in many different tasks. This is especially true in the case of automated decision-making systems based on machine learning (ML). Important ethical implications arise when such decision systems are used in sensitive contexts (e.g., justice or loans). The elicitation of these implications, that is, of the ethical requirements behind ML-based systems is a new challenge we must address to avoid societal risks. This is particularly urgent for fairness since this notion lacks a precise and commonly accepted definition, thus hampering its assessment. The goal of this paper is to give a comprehensive definition of fairness, present a unified taxonomy of alternative interpretations, define a new decision tree that can guide the choice of the correct interpretation, and carry out a preliminary assessment with experiments in a real-world context.**

***Index Terms*—fairness, non-functional requirements, machine learning**

I. Introduction

Fairness can be defined as *the absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics* [1], but this definition can easily accommodate different interpretations. Philosophers have been debating about various definitions for centuries [2]–[4], psychologists have been trying to measure how people perceive fairness for decades [5], and now computer scientists are trying to embed it in computing and decision making systems. Since nowadays we outsource complex actions and decisions to computer systems, assessing fairness is not just a matter of judging human decisions, but it is the problem of analyzing how complex software systems make decisions. We must avoid *algorithmic* [6] bias, that is, we must avoid that used algorithms do not work properly and hamper the fairness of the decision.

Ex-post analysis of existing automated decision systems revealed serious fairness violations. For example, ProPublica highlighted that the COMPAS Recidivism Algorithm, used by officials to predict the likelihood of criminal defendants to commit other crimes, was unfair to certain subjects[1]: black people were wrongly assigned a higher risk of recidivism than white people under the same conditions. Dastin [7] analyses the experimental, artificial intelligence-based system used by Amazon to rank curricula and select candidates to hire. Since the system was trained on the resumes submitted in the previous ten years with a significant predominance of male candidates (a well-known gender imbalance in the technology industry), it learned to prefer male candidates.

Although fairness is a highly desirable quality in society, it can be surprisingly difficult to achieve in practice. It is particularly challenging because of its uncertainty, domain dependence, and lack of awareness and regulations. The semantics of fairness can be different and is often context-dependent. For example, Amazon might have wanted to hire the same number of women and men, or have the number of hired women and men be proportional to the applicants' gender. Both could be "appropriate" interpretations, but with significant differences. The community needs proper means to reason on, specify, measure, and compare different notions of fairness. Note that the *impossibility theorem* [8] states that one cannot satisfy three different definitions at the same time, with one exception: the algorithm is a perfect predictor and there is no prevalent difference between groups.

Because of these problems, this paper aims to promote deliberation on the fairness of automated decision systems by means of properly stated requirements. One should systematically elicit these requirements by reasoning on the different connotations, expressing them rigorously, and evaluating their consequences on the future decisions before the system is implemented and used in practice. The proposed conceptual framework focuses on *statistical* fairness interpretations and merge the definitions presented by Verma and Rubin [9] and by Saleiro et al. [10]. We do not consider other definitions, like counterfactual [11] or causal ones [12], because they use the concept of causal graphs, and do not involve any prediction algorithm. We want to foster the idea of *fairness by design* to treat fairness in a sound way and prevent a-posteriori recovery actions to fix discovered violations. The framework unifies the different definitions in a single taxonomy and sketches a decision tree to help (requirements) experts identify the right definition and understand the differences.

The rest of this paper is organized as follows. Section II surveys related approaches. Section III introduces our taxonomy of statistical fairness interpretations. Section IV describes the decision tree we defined to help experts identify the definitions of interest. Section V concludes the paper.

II. Related Work

Data science ethics [14] is already addressing moral problems related to data. Software engineering should widen the scope and understand how ethical issues have an impact on the different phases of the development process (*ethics-aware software engineering* [15]).

The work presented in this paper is partially supported by project ICTD4Dev, funded by AICS (Italian Agency for Development Cooperation).

[1]https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

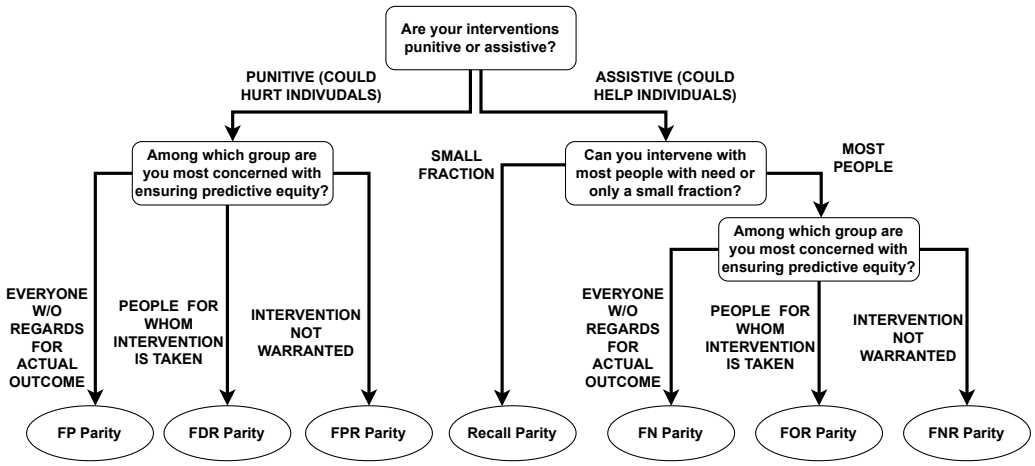

Fig. 1. Fairness tree presented [13].

Focusing on requirements elicitation, Habibullah and Horkoof [16] point out the prominent role of fairness and the lack of a common ground knowledge for machine learning (ML) systems. Also Brun and Meliou [17] say: *"The creation of a catalog of fairness definitions and guidelines for when each definition should be used is an open challenge in the engineering of fair software. Of the dozens of existing definitions, each depends on and is appropriate for certain contexts."*

Verma and Rubin [9] are the first to present a comprehensive taxonomy of statistical fairness definitions. Proposed definitions are based on the idea of *protected attribute*, that is, a characteristic for which one wants to assess non-discrimination (e.g., religion, race, and sex). This work does not include all available definitions and there is no guidance to help the expert decide and choose the right interpretation.

Similarly, Saleiro et al. [13] propose the *fairness tree* of Figure 1, that is, a decision tree that organizes the most common definitions of (statistical) fairness. Given a specific context and goal, it helps select the optimal and appropriate definition by asking the expert whether the system is supposed to be punitive or assistive, that is, it may hurt or help individuals. In the former case, fairness should strictly refer to individuals with a specific value for the protected attribute, with the risk of ignoring someone who should be in (false negatives, FNs). In the latter case, the recipients should include all eligible subjects, with the risk of including someone who should not be part of the group (false positives, FPs). If one takes a positive attitude, the decision is about the number of affected people: only a few or as many as possible. In both cases, the last question refers to the groups the expert is interested in ensuring fairness: the whole past and future population, those who will be predicted positively (future population), and the past decisions (used to test the system), respectively. This tree does not consider definitions that do not only involve FPs or FNs, ignores some known definitions, and the categorization of intervention is purely binary, no mixed or neutral options are considered.

There are also systems that measure algorithmic fairness given the different definitions. Saleiro et al. [13] also developed *Aequitas*, an open-source bias and fairness audit toolkit that allows users to seamlessly test models for several bias and fairness definitions. *AI Fairness 360* [18] is another open-source framework that measures algorithmic fairness through statistical measures, similarly *Fairlearn* [19] both shares some metrics with the two previous systems and provides additional statistical definitions. Although these solutions help identify, reason on, and measure fairness, they tend to be partial, since they only focus on some interpretations. Each system has its own specific approach, which does not help the expert reason on how fair a system is and creates confusion when comparing results from different systems.

III. Fairness Taxonomy

To formalize fairness, we assume that each individual a system is supposed to manage is represented by a tuple $I = \langle a_1, a_2, a_3, ..., a_n \rangle$, with n attributes a_i. One of these attributes, g ($g \in I$), is the protected attribute that one wants to use for the analysis. It can only assume two values, which are used to characterize discriminated (*discr*) and privileged (*priv*) individuals. If g is not intrinsically binary, there must be a user-defined way to group the different values. X ($X \subseteq I$) identifies the set of attributes used by the prediction algorithm. L ($L \subseteq X$) refers to the legitimate attributes that can affect the outcome and the set can be empty.

To know both the actual conclusion and the outcome produced by the system, we consider subjects for which we have a decision, no matter how it was produced, and the prediction of an ML-based classification algorithm. Thus, y ($y \in I$) is the known boolean classification, and $\hat{y}$ is the outcome produced by the decision-making system.

If we applied this definition to the well-known German Credit problem [20], the individuals are the persons who asked for credit. g is the *sex* of each person (*female* or *male*), X contains *credit amount*, *savings*, *job*, and other attributes, L only contains *savings*, and we discriminated between *low* and the other values, y is *true* (1) if the bank gave the loan,

and *false* (0) otherwise; similarly for $\hat{y}$. We can now use this example to exemplify the different definitions we present.

A. Taxonomy

Our taxonomy takes into account the statistical definitions of fairness borrowed from [9] and [10]. For each considered interpretation, we provide a precise definition, explain it by means of the running example, and give a mathematical formulation. These formulations, summarized in Table I, exploit the confusion matrix generated by the prediction algorithm [21], where TP (true positive) and TN (true negative) are the cardinalities of the sets of correct predictions ($y = \hat{y}$) in the dataset with a positive or negative outcome, respectively. In contrast, FP (false positive) and FN (false negative) are the cardinalities of positive (negative) predictions in the dataset for which the prediction is not the same as the decision ($y \neq \hat{y}$).

The mathematical formulation calculates a ratio between two probabilities. For example, if we consider *Group Fairness*, the numerator computes the probability of having a positive prediction ($\hat{y} = 1$) for the individuals in the discriminated group ($g = discr$), while the denominator computes the same probability in the privileged group ($g = priv$)[2]. The ideal value is 1, which means both groups are treated equally. If the value is between 0 and $1 - t$, the discriminated group is treated unfairly, whereas if the value is greater than $1 + t$, the privileged group is treated unfairly. Parameter t is a threshold value that must be set by the user. Note that when the definition considers only FN, the treatment of the discriminated group is at the denominator and the privileged one at the numerator; if it considers FP, TN, or TP the fraction is flipped to set the same threshold to evaluate all formulas consistently.

The taxonomy classifies the definitions based on similarities. Considered definitions only address one attribute at a time, and we do not study the so-called *intersectional bias*, which will be part of our future work.

The expert starts from a hypothesis based on the two values of the protected attribute. In our examples, females are assumed to be the discriminated group and males are the privileged one. Each definition then compares the treatment of the discriminated group with respect to the privileged one.

The first set of definitions comprises the ones already proposed by Verma and Rubin [9] that are only based on the results produced by the system and that do not consider past knowledge. These definitions are:

(1) Group Fairness [9] requires that discriminated and privileged individuals have the same probability of a positive outcome. For example, the percentage of females and males in the dataset who are predicted to have the loan should be the same.

(2) Conditional Statistical Parity [9] requires that discriminated and privileged individuals, with the same values for the legitimate attributes L, have the same probability of a positive outcome. For example, the percentage of females and males who are predicted to have the loan, and have the same *savings* level, should be the same.

[2]Formulae 9 to 12 combine of two values of the confusion matrix, thus the result is a product of the ratios.

TABLE I
DEFINITIONS OF STATISTICAL FAIRNESS PART OF THE TAXONOMY.

	Name	Formal Definition
1	Group Fairness	$\frac{P(\hat{y}=1\|g=discr)}{P(\hat{y}=1\|g=priv)}$
2	Conditional Statistical Parity	$\frac{P(\hat{y}=1\|g=discr,L=l)}{P(\hat{y}=1\|g=priv,L=l)}$
3	Predictive Equality	$\frac{P(\hat{y}=1\|y=0,g=discr)}{P(\hat{y}=1\|y=0,g=priv)}$
4	Predictive Parity	$\frac{P(y=0\|\hat{y}=1,g=discr)}{P(y=0\|\hat{y}=1,g=priv)}$
5	FP/GS Parity	$\frac{P(\hat{y}=1,y=0\|g=discr)}{P(\hat{y}=1,y=0\|g=priv)}$
6	Equal Opportunity	$\frac{P(\hat{y}=0\|y=1,g=priv)}{P(\hat{y}=0\|y=1,g=discr)}$
7	FOR Parity	$\frac{P(y=1\|\hat{y}=0,g=priv)}{P(y=1\|\hat{y}=0,g=discr)}$
8	FN/GS Parity	$\frac{P(\hat{y}=0,y=1\|g=priv)}{P(\hat{y}=0,y=1\|g=discr)}$
9	Treatment Equality	$\frac{P(\hat{y}=0\|y=1,g=priv)}{P(\hat{y}=1\|y=0,g=priv)} * \frac{P(\hat{y}=1\|y=0,g=discr)}{P(\hat{y}=0\|y=1,g=discr)}$
10	Conditional Use Accuracy Equality	$\frac{P(\hat{y}=1\|y=1,g=discr)}{P(\hat{y}=1\|y=1,g=priv)} * \frac{P(\hat{y}=0\|y=0,g=discr)}{P(\hat{y}=0\|y=0,g=priv)}$
11	Overall Accuracy Equality	$\frac{P(\hat{y}=y=1,g=discr)}{P(\hat{y}=y=1,g=priv)} * \frac{P(\hat{y}=y=0,g=discr)}{P(\hat{y}=y=0,g=priv)}$
12	Equalized Odds	$\frac{P(\hat{y}=1\|y=1,g=discr)}{P(\hat{y}=1\|y=1,g=priv)} * \frac{P(\hat{y}=1\|y=0,g=discr)}{P(\hat{y}=1\|y=0,g=priv)}$

The second set comprises definitions based on actual and predicted outcomes, that is, known decisions (in the dataset) and results produced by the ML-based system. In this case, the dataset is not only used for training but also for assessing the actual fairness. We speculate on both correct and wrong predictions, that is, predictions that are the same as available decisions and those that are not.

If one focuses on wrong predictions, we have definitions that exploit false positives (FP), false negatives (FN) or both sets. The interpretations based on FPs are:

(3) Predictive Equality [9], which is equal to **FPR Parity** [10], requires that discriminated and privileged individuals with a negative decision have the same probability of having a positive prediction. For example, females and males who did not get the loan should have equal probability to be predicted to have it.

(4) Predictive Parity [9], which is also called **FDR Parity** [10], requires that discriminated and privileged individuals who are predicted to have a positive outcome should have the same probability of actually having a negative decision. For example, females and males who are predicted to have the loan should have an equal probability that the available decisions state they did not received it.

(5) FP/GS[3] Parity [10] requires that discriminated and privileged individuals with a negative decision have the same probability of a positive prediction. For example, females and males should have equal probability to be predicted to have the credit and actually did not have it (in the dataset).

[3]GS stands for Group Size.

Those based on FNs are:

(6) Equal Opportunity [9], which is equal to **FNR Parity and 1-Recall Parity** [10], requires that discriminated and privileged individuals who have a positive decision should have an equal probability of a negative prediction. For example, females and males who had the loan should have equal probability to be predicted to not have it.

(7) FOR Parity [10] requires that discriminated and privileged individuals who are predicted to have a negative outcome should have an equal probability of a positive decision (in the dataset). For example, females and males who are predicted to not have the credit should have the same probability of a positive decision.

(8) FN/GS Parity [10] requires that discriminated and privileged individuals with a positive decision have equal probability to be predicted to have a negative outcome. For example, females and males should have equal probability to be predicted to not have the credit and actually got it.

If we are interested in combining FPs and FNs, there is only one definition we can exploit:

(9) Treatment Equality [9], which is the combination of **FN/GS** and **FP/GS Parity**, requires that the ratio between wrong and correct predictions is the same for privileged and discriminated individuals. For example, the percentage of erroneous predictions is the same for females and males.

If we focus on correct predictions, the definitions exploit TPs and TNs:

(10) Conditional Use Accuracy Equality [9], which is the combination of **Predictive Parity** and **FOR Parity**, requires that discriminated and privileged individuals with a positive decision have an equal chance of being predicted to have a positive outcome. Similarly, discriminated and privileged individuals with a negative decision should have an equal chance of being predicted to have a negative outcome. For example, females and males who had a credit should have equal probability to be predicted to have it, and females and males that did not have a credit should have equal probability to be predicted to not have it.

(11) Overall Accuracy Equality [9] requires that discriminated and privileged individuals have an equal probability of a positive decision and of a positive prediction. They should also have an equal probability of a negative decision and of a negative prediction. For example, females and males should have equal probability to have the credit and be predicted to have it, and they should have equal probability to not have the credit and be predicted to not have it.

In the end, if one wanted to combine wrong and correct predictions, s/he can predicate on TPs and FPs:

(12) Equalized Odds [9], which is the combination of **Predictive Equality** and **Equal Opportunity**, requires that discriminated and privileged individuals who have a positive decision should have an equal probability of a positive prediction. Similarly, discriminated and privileged groups who have a negative decision should have an equal probability of a positive prediction. For example, females and males that had the credit should have the same probability of being predicted to have it, and females and males who did not have the credit should have the same probability of being predicted positively.

TABLE II
MAPPING BETWEEN DEFINITIONS IN [9] AND DEFINITIONS IN [10].

Definitions in [9]	**Definitions in [10]**
Group Fairness	-
Conditional Statistical parity	-
Predictive Equality	FPR Parity
Predictive Parity	FDR Parity
Equal Opportunity	FNR Parity, 1-Recall Parity
Equalized Odds	-
Conditional use accuracy equality	-
Overall accuracy equality	-
Treatment equality	-
-	FOR Parity
-	FN/GS Parity
-	FP/GS Parity

Table II summarises all definitions and provides a mapping between the definitions presented in [9] and those proposed in [10]. Note that there are three cases where the same interpretation has two different names: for example, *Equal Opportunity* is the same as *FNR Parity* and is also complementary to *Recall Parity*.

Besides deciding which interpretation one should adopt, there is also the problem of choosing the right threshold to interpret obtained results. Different domains usually call for different thresholds and the identification of the right value can be incremental and depends on many different factors.

B. A first assessment

The different interpretations presented above do not simply provide the same outcome from different angles. To witness the importance, and impact, of the different options, Table III presents the experimental results we obtained by applying the different definitions to the data available for the aforementioned German Credit problem [20]. We started from the available dataset and adopted a threshold t equal to 0.15 (i.e., a 15% deviation is small enough to understand whether there is a discrimination between women and men). We applied holdout evaluation, that is, we split the dataset in two parts ($2/3$ training set, and $1/3$ test set), and trained a binary classifier (random forest) on the training set. We then tested the resulting decision making system on the test set. Different formulae gave different results. It is clear that the same system can be considered fair (six definitions have values between .85 and 1.15), unfair and privilege women (five definitions have values between 0 and .85), or unfair and privilege men (one definition has a value equal to 1.15).

Even if we only consider a single dataset, different interpretations lead to significantly different results. The assessment cannot only be done ex-post by picking a definition randomly. The adoption of a conscious approach is key and a coherent treatment throughout the whole development process is mandatory to get significant and consistent results. We must reason on the fairness of ML-based systems from the

TABLE III
FAIRNESS VALUES COMPUTED ON THE GERMAN CREDIT DATA.

Name	Value	Discr. Group
Group Fairness	0.866	-
Conditional Statistical Parity	0.667	woman
Predictive Equality	0.757	-
Predictive Parity	0.993	-
FP/GS Parity	0.867	-
Equal Opportunity	0.534	woman
FOR Parity	1.150	man
FN/GS Parity	0.577	woman
Treatment Equality	0.867	-
Conditional Use Accuracy Equality	1.109	-
Overall Accuracy Equality	0.623	woman
Equalized Odds	0.694	woman

beginning and we need proper means to identify the right interpretation, select the right definition for automated assessment, and maybe conduct *what-if* analyses by working with different hypotheses and comparing obtained results.

IV. FAIRNESS DECISION TREE

Figure 2 presents our fairness decision tree, which aims to extend and update the original proposal by Saleiro et al. (Figure 1). Our goal is to widen the scope of original tree and offer experts a readily-available tool to let them reason on and decide about the fairness interpretations they consider most suitable for their ML-based systems. The proposed decision tree is open and can further be extended if new needs and definitions arise. Note that the different dashed lines show the relationships between fairness definitions as specified in Section III.

The idea of *fairness by design* and the multiple definitions of Section III impose that experts be guided through the steps needed to move from a general idea to a proper interpretation and its operationalization. The right interpretation depends on the problem at hand, but some general guidelines can be envisaged and materialized through a decision tree that guides the expert to both identify the best definition and reason on similar, related, formulations in a conscious and rational manner. At most four questions are enough to let the decision tree suggest the best definition. As in the original proposal, the leaves identify the possible, alternative definitions of fairness.

The first question experts are supposed to answer (**A. Is past knowledge relevant?**) is key to decide about the importance of obtained predictions with respect to past decisions on similar subjects —maybe obtained without any automated system. If the answer is *yes*, there is the intention to reason on the trade-off between past experiences and ML-based predictions. If the answer is *no*, past decisions are not relevant. In other words, this question is about the option of confronting ML-based predictions against domain knowledge.

If only predictions are of interest, then the decision is simple and the expert should focus on legitimate attributes (**B. Are predictions based on legitimate attributes?**). The definition introduced in Section III says that a set of legitimate attributes (L) may affect the outcome. If only the ML-based prediction matters and legitimate attributes exist (L is not empty), the expert is interested in *Conditional Statistical Parity*. If L were empty, *Group Fairness* would be the proper interpretation.

If we are interested in predictions and past knowledge, the second question (**C. Which type of predictions are you interested in?**) is about the different types of predictions. They could be *wrong* predictions when the predicted outcome is not equal to the decision embedded in the past knowledge, or *correct* if the values are equal. The question can have three possible answers: *wrong* if the expert only focuses on incorrect predictions, *correct* if s/he focuses on correct ones, and *both* if both correct and incorrect predictions are of interest.

If only wrong predictions matter, the next question is about the type of error: negative predictions that should have been positive or the opposite, that is, positive predictions that should have been negative. Again, the possible outcomes are: *positive*, *negative*, and *both*.

If one decided for *positive* or *negative* predictions, then the next question refers to the impact past decisions (acquired knowledge) have on predictions (**E. How conservative should decisions be?**). The idea is to combine the two aspects. The question wants to make the analyst reflect on the *conservatism* of decisions. *High* means that past decisions are more important than predictions so there is high conservatism, *Low* means the expert is willing to consider that the automated prediction is more important than past knowledge, and *Medium* is a compromise. Thus, definitions (3), (4), and (5) analyze, in inverse order of conservatism, wrong positive predictions, and definitions (6), (7) and (8) wrong negative predictions. Definition (9) is the only definition that compares both error types.

If the interest is for *correct* predictions (question C), the next question (**F. Do you weigh more predictions or past decisions?**) asks the expert how to balance the two contributions while assessing fairness. Definitions (10) and (11) refer to correct outcomes and weigh predictions and decisions differently.

Finally, if one were interested in *both* prediction types (question C), and wanted to reason on both of them at the same type, definition (12) combines wrong and correct predictions.

Note that the subtree routed in node D, with only edges *positive* and *negative*, is the same as the original proposal [13], but we do not consider Recall Parity since, as Table II suggests, it is already covered with *Equal Opportunity*, which is $1 - Recall\ Parity$.

We can now briefly try to use the decision tree with our German Credit example. We assume that we are interested in assessing fairness between women and men, and we can envision at least three cases. If we are only concerned with the results obtained from the ML-based decision system, no matter the correctness of produced predictions, we could add that *savings* must be greater than *low*, that is, we have a legitimate attribute, then the right definition is (2) *Conditional Statistical Parity*. If the idea of fairness should analyze all those cases where the system would have granted the loan (positive prediction), but the past knowledge would have not,

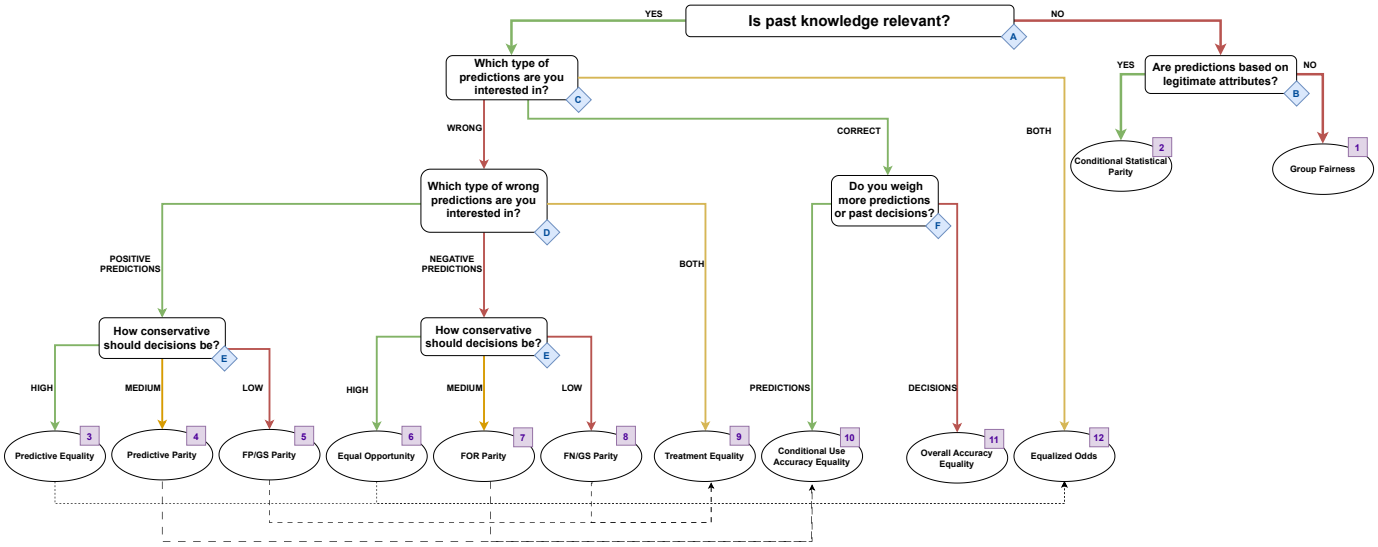

Fig. 2. Fairness Decision Tree (dotted lines highlight dependencies between different definitions).

then we should adopt (3) *Predictive Equality*. If we wanted to combine past knowledge and predictions —both correct and wrong ones, then the right decision is (12) *Equalized Odds*. Note that, given the available dataset, two, out of three, definitions would lead to unfairness towards women.

V. Conclusions and Future Work

Complex decisions in sensitive contexts, for example justice and bank loans, are more and more outsourced to automated ML-based decision-making systems. This paper focuses on the fairness —as non-functional requirement— of these systems and claims that it cannot simply be evaluated ex-post: it must be analyzed and specified properly at the beginning of the software process. To do this, the paper proposes a unified taxonomy of common statistical interpretations of fairness and a decision tree to help experts select the right interpretation and understand the implications of the different choices.

Our future work comprises the extension of our analysis to further definitions that are not based on statistics, the study of intersectional bias, that is, the analysis of multiple attributes at the same time, and a thorough assessment on real cases.

References

[1] N. A. Saxena, K. Huang, E. DeFilippis, G. Radanovic, D. C. Parkes, and Y. Liu, "How Do Fairness Definitions Fare? Testing Public Attitudes Towards Three Algorithmic Definitions of Fairness in Loan Allocations," *Artificial Intelligence*, vol. 283, p. 103238, 2020.

[2] P. Neal, "Justice as Fairness: Political or Metaphysical?" *Political Theory*, vol. 18, no. 1, pp. 24–50, 1990.

[3] I. Boran, "Benefits, Intentions, and the Principle of Fairness," *Canadian Journal of Philosophy*, vol. 36, no. 1, p. 95–115, 2006.

[4] M. Fleurbaey, *Fairness, responsibility, and welfare*. OUP Oxford, 2008.

[5] J. Jetten and K. Peters, *The Social Psychology of Inequality*. Springer International Publishing, 01 2019.

[6] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Survey*, vol. 54, no. 6, pp. 115:1–115:35, 2022.

[7] J. Dastin, "Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women," in *Ethics of data and analytics*. Auerbach Publications, 2018, pp. 296–299.

[8] A. Chouldechova, "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, 2017.

[9] S. Verma and J. Rubin, "Fairness Definitions Explained," in *Proceedings of the International Workshop on Software Fairness*. ACM, 2018, pp. 1–7.

[10] P. Saleiro, B. Kuester, A. Stevens, A. Anisfeld, L. Hinkson, J. London, and R. Ghani, "Aequitas: A Bias and Fairness Audit Toolkit," *CoRR*, vol. abs/1811.05577, 2018.

[11] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual Fairness," *Advances in neural information processing systems*, vol. 30, 2017.

[12] S. Galhotra, Y. Brun, and A. Meliou, "Fairness Testing: Testing Software for Discrimination," in *Proceedings of the 2017 11th Joint meeting on foundations of software engineering*, 2017, pp. 498–510.

[13] P. Saleiro, K. T. Rodolfa, and R. Ghani, "Dealing with Bias and Fairness in Data Science Systems: A Practical Hands-on Tutorial," in *The 26th Conference on Knowledge Discovery and Data Mining*. ACM, 2020, pp. 3513–3514.

[14] L. Floridi and M. Taddeo, "What is Data Ethics?" *Philosophical Transactions of The Royal Society A Mathematical Physical and Engineering Sciences*, vol. 374, pp. 1–9, 12 2016.

[15] F. B. Aydemir and F. Dalpiaz, "A Roadmap for Ethics-aware Software Engineering," in *Proceedings of the International Workshop on Software Fairness*. ACM, 2018, pp. 15–21.

[16] K. M. Habibullah and J. Horkoff, "Non-functional Requirements for Machine Learning: Understanding Current Use and Challenges in Industry," in *29th International Requirements Engineering Conference*. IEEE, 2021, pp. 13–23.

[17] Y. Brun and A. Meliou, "Software Fairness," in *Proceedings of the 2018 Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. ACM, 2018, pp. 754–759.

[18] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović *et al.*, "AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4–1, 2019.

[19] S. Bird, M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker, "Fairlearn: A Toolkit for Assessing and Improving Fairness in AI," *Microsoft, Technology Repository*, 2020.

[20] M. Lichman *et al.*, "UCI Machine Learning Repository," 2013.

[21] F. J. Provost and R. Kohavi, "Guest Editors' Introduction: On Applied Research in Machine Learning," *Machine Learning*, vol. 30, no. 2-3, pp. 127–132, 1998.