Discrimination-Aware Process Mining: a Discussion

Timo Pohl, Mahnaz Sadat Qafari, and Wil M. P. van der Aalst

Process and Data Science Chair (PADS)
RWTH Aachen University Aachen, Germany
timo.pohl@rwth-aachen.de, {m.s.qafari,wvdaalst}@pads.rwth-aachen.de

Abstract. Organizations increasingly use process mining techniques to gain insight into their processes. Process mining techniques can be used to monitor and/or enhance processes. However, the impact of processes on the people involved, in terms of unfair discrimination, has not been studied. Another neglected area is the impact of applying process mining techniques on the fairness of processes. In this paper, we overview and categorize the existing fairness concepts in machine learning. Moreover, we summarize the areas where fairness is relevant to process mining and provide an approach to applying existing fairness definitions in process mining. Finally, we present some of the fairness-related challenges in processes.

Keywords: Process mining · fairness · discrimination.

1 Introduction

Organizations interact with and affect people, such as customers, employees, or stockholders in many forms. They operate in various sensitive environments such as education, employment, healthcare, and finance. Processes taking place in such sensitive environments often have important and life-changing effects on the people involved. Moreover, such processes typically involve several decision-makings which are performed by human resources or (supported by) machine learning algorithms trained on historical data. These decision-makings are one of many factors that make processes vulnerable to various forms of discrimination. See [20] for real-life examples of discriminatory outcomes produced by algorithmic decision makers. As the impact of processes on the people involved can be very drastic, it is crucial to be able to identify instances of discrimination within processes in order to minimize negative impacts.

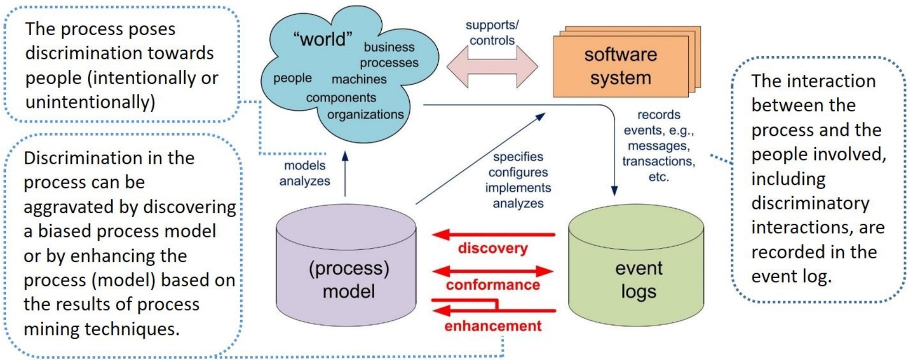
Process mining is a set of techniques that combine data science with model-based process analysis to enable the understanding and improvement of operational processes. Even though the concept of responsible data science has been investigated in process mining related literature, [2,4], to the best of our knowledge, in this area, [23] is the only work dedicated to fairness. In this work, making fair conclusions, which is one of the main aspects of fairness in process mining, is investigated. Here, we mainly focus on another main aspect of fairness in process mining: detecting unfair discrimination against cases and resources. Intuitively, (unfair) discrimination is the act of treating similar individuals in the same situation differently based on one or more protected attributes, such as ethnicity, race, gender, (dis)ability, or sexual orientation [11].

Typically, process mining techniques are categorized into three types: process discovery, conformance checking, and process enhancement. Figure 1 (adapted from [1]) shows the interaction between the processes, the environment they take place in, and the process mining techniques. Processes impact their environment, which may intentionally or unintentionally pose discrimination towards the people in their environment. This discrimination might have stemmed from the process itself, its resource(s), or learned from the historical data. The interaction between the process and its environment is captured by the information systems and manifests itself in the event log. The discrimination level of the process can get aggravated by applying the results of process mining techniques on event logs containing discrimination.

Fairness is a context-sensitive concept. Consequently, there is a huge number of fairness definitions in the literature, some of which are in contradiction with each other [8]. Furthermore, there is a lack of consensus, both in academia and society, on which definition of fairness is the correct one [14,16]. This makes it hard to decide on the proper definition of fairness to audit a process.

This issue is aggravated when there are multiple human entities with different roles and desires in an organization as each one may entail a different notion of fairness. Therefore, in this paper, we categorize the fairness concepts and definitions based on their properties such that it makes it easier for the user to select the appropriate one. We discuss some of the applications and challenges of applying fairness in process mining. Moreover, we elaborate on a mapping between the existing techniques to measure fairness and process mining.

The rest of the paper is organized as follows. In Section 2, we provide a brief overview of fairness considerations in literature and describe common fairness



The process poses discrimination towards people (intentionally or unintentionally)

Discrimination in the process can be aggravated by discovering a biased process model or by enhancing the process (model) based on the results of process mining techniques.

The interaction between the process and the people involved, including discriminatory interactions, are recorded in the event log.

Fig. 1: Process evolution diagram; positioning of the three main types of process mining [1]. Some of the areas where discrimination can play a role in process mining are shown in this picture.

concepts and measures. In Section 3, we discuss the possible applications of fairness in process mining. In Section 4 we elaborate on mapping the existing fairness definitions to process mining. Finally, in Section 5, we conclude and provide directions for future work.

2 Literature Review

In this section, we provide an overview of the fairness understandings, concepts, and measures defined in the literature. We start by defining relevant terms and concepts. Then, we present a taxonomy that provides an overview of various fairness measures. Due to the extensive amount of fairness-related scientific literature, we present only concepts and measures potentially relevant to the area of process mining. For a comprehensive overview of fairness research, we refer interested readers to [8,16,20,25].

2.1 Relevant Terms

Here, we briefly discuss terms relevant to fairness. For a detailed discussion on such fairness fundamentals, we refer readers to chapter 3 of [10].

- *Discrimination:* The word *discrimination* means “to divide”, “separate”, “distinguish”, which is exactly the goal of classification. Therefore, discrimination itself is not necessarily unjust or unfair. However, discrimination is considered unfair if individuals receive harmful treatment based on their membership to a specific group [3].
- *Protected groups/ Sensitive attributes.* A protected group is a subgroup of the population. The attributes indicating if an individual belongs to a protected group are called *sensitive attributes.*
- *Outcome.* Outcome is an attribute that captures an aspect of the system that is supposed to be fair. It is important to note that not just the outcome is context-dependent, but also its desirability. For example, in a hospital context, less waiting time for visiting a doctor is more desirable while more waiting time (up to a threshold) between an elaborate surgery and the discharge of a patient is more desirable.

2.2 Fairness Taxonomy

In this subsection, we present fairness concepts and measures defined in machine learning literature. We structure these measures in a taxonomy (Figure 2), which is an extension of the fairness tree presented by Saleiro et al. in [25]. The fairness definitions in machine learning can be conceptually divided into *group fairness* definitions and *individual fairness* definitions[7,24,28].

Group Fairness assesses the (approximate) parity of some statistical measure across all demographic sub-populations [7,15]. The group fairness measures are further divided into three categories: disparate distribution, disparate representation, and disparate error.

1. *Distribution-based fairness.* Here, the main idea is that the distribution of the predictions should be similar across all subgroups [21]. Another example of distribution-based fairness measures is proposed in [9].
2. *Measures assessing representation.* The fairness measures in this category are based on the representation of the various subgroups in the outcome of a classifier or a subset selection method [13]. Based on the application and context, the measures are further divided into the following two categories.
   - *Coverage-based fairness.* In this category of measures, the main concern is either having the same number of people from each group or having a number proportional to their relative representation in the whole population in the selected/sampled groups [25].
   - *Ranking-based fairness* [17] defines measures for assessing representation tailored for scenarios in which individuals are ranked according to some predicted score. It also assumes a notion of ground truth which indicates the correct ordering. In essence, this definition requires that every subgroup has an equal representation in the top-n candidates in both rankings, one ranked by ground truth, the other by predicted score. Another example of ranking-based fairness is defined in [27]. Here, the fairness criterion is that the number of protected elements in the top-n candidates (for every n) is the same number that would be expected if the top-n candidates were picked at random from the overall population.
3. *Measures assessing error.* This group of measures assesses the discrimination made via errors made by the predictor and requires the existence of some predicted value, as well as a notion of ground truth [6,16]. These measures are further subdivided into three contextual categories: *assistive*, *punitive*, and *neutral*.
   - *Assistive context.* In this context, a positive classification is assumed to bring benefits to the individuals, therefore false negatives are more undesirable in terms of fairness than false positives.
   - *Punitive context.* This context is exactly the other way around, i.e., a positive classification is assumed to bring negative consequences for the individuals. Hence, false positives are more undesirable in terms of fairness than false negatives.
   - *Neutral contexts.* Here, we assume that false negatives and false positives are equally undesirable.

The main advantage of group fairness definitions is their simplicity. They can be easily explained and verified [8]. However, their main drawback comes from the fact that this category of fairness definitions provides guarantees only to "average" members of the protected groups. Consequently, they do not provide guarantees to individuals or subgroups within the protected groups. Moreover, some of these measures can be at odds with one another [8].

It is important to note, that group fairness measures based on parity require some assumptions. The main assumption is that differences between groups are due solely to unwarranted bias and that all warranted differences have been eliminated (for example by removing them from the data) [15]. This includes the assumption that the reasons for existing differences do not lie in the choices of individuals but in factors outside of their control [15]. If these assumptions apply, these measures can help in correcting the unjust bias. However, if these assumptions do not apply, enforcing them can lead to outcomes that are unfair from the perspective of an individual or it can lead to a form of reverse discrimination towards the rest of the population [16].

Individual Fairness assesses the similarity of the outcome of pairs of similar individuals ignoring their differences in terms of protected attributes [9]. Two main techniques for assessing individual fairness are *similarity-based fairness* and *counterfactual fairness* (highlighted in yellow in Figure 2).

1. *Similarity-based fairness* assesses individual fairness by using two similarity metrics. The first metric estimates the similarity of two individuals. The second metric estimates the similarity of the outcomes that two individuals received. To assess the fairness from individual A's perspective, one simply matches A to the most similar individual(s) in the data. Then, the similarity of the two individuals is compared with the similarity of their outcomes. By doing this for every individual in our data, we can measure how similarly similar individuals are treated. Examples of similarity-based fairness measures can be found in [9,28].
2. *Counterfactual fairness* is formulated in the context of fair classifications. The main idea is to investigate the question of "how would the prediction change if the protected attribute of an individual were different"[12]. Under this approach, a decision is considered fair towards an individual if the outcome of the decision is the same in (a) the actual world and (b) a counterfactual world where the individual belonged to a different demographic group [19]. Counterfactual fairness can also be used to assess group fairness. By studying in what direction the prediction changes when changing protected attributes, it is possible to infer which groups are given preferential outcome(s). For example, if by changing the group membership from G to G', the prediction always changes from a negative to a positive outcome, this indicates discrimination against either group G or G'.

The main advantage of individual fairness definitions is their semantics, as they provide guarantees to individuals and not average members. However, they require making significant assumptions. For example, similarity-based fairness measures are built on similarity measures, the definition of which can require a large amount of domain knowledge that even domain experts rarely possess.

3 Fairness Applications in Process Mining

Fairness has three key applications in process mining. In the following, we briefly discuss each application and provide promising lines of research for each one. It is worth noting that fairness is not relevant in all processes. For example, in fully automated processes with no human involvement, fairness does not play a role. The relevance of fairness to a process depends on (1) how much it involves

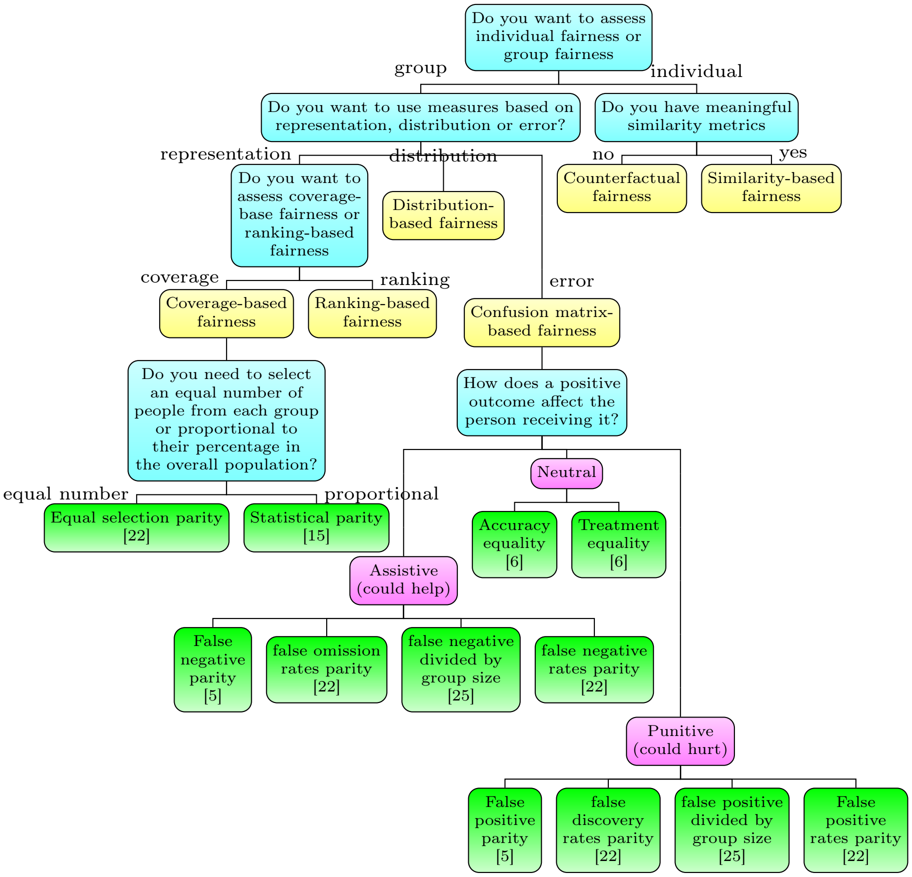

Do you want to assess individual fairness or group fairness
group
individual
Do you want to use measures based on representation, distribution or error?
Do you have meaningful similarity metrics
representation
distribution
no
yes
Counterfactual fairness
Similarity-based fairness
Do you want to assess coverage-base fairness or ranking-based fairness
Distribution-based fairness
coverage
ranking
error
Coverage-based fairness
Ranking-based fairness
Confusion matrix-based fairness
Do you need to select an equal number of people from each group or proportional to their percentage in the overall population?
How does a positive outcome affect the person receiving it?
Neutral
equal number
proportional
Equal selection parity [22]
Statistical parity [15]
Accuracy equality [6]
Treatment equality [6]
Assistive (could help)
False negative parity [5]
false omission rates parity [22]
false negative divided by group size [25]
false negative rates parity [22]
Punitive (could hurt)
False positive parity [5]
false discovery rates parity [22]
false positive divided by group size [25]
False positive rates parity [22]

Fig. 2: Taxonomy of Fairness measures

humans (for example, as cases or resources) and (2) how strong the impact of the process on the lives of the involved people is.

Discrimination in processes. Processes involve at least two discrimination-relevant entities: *resources* and *cases*. The general idea is, that cases cannot directly influence the process but may suffer from discrimination. From a case perspective, waiting times in and between activities, the number of re-do's, the success rate, occurrences of deviations, and the allocation of resources are some examples of possible outcomes (as defined in section 2). Resources, in turn, can cause discrimination by making biased decisions. However, they can also be affected by discrimination. From a resource perspective, possible outcomes include the assigned workload and the complexity of the assigned tasks. Some of the interesting lines of research concerned with discrimination in processes are:

- developing process mining specific measures to assess the level of discrimination in event logs and process models,
- developing methods for on-time monitoring of fairness in a process so that process owners can react on time and prevent unfair discrimination, and
- providing methods to improve/enhance fairness in a process by reducing the discrimination level in the event log or process model.

Making fair conclusions. [1] Root cause analysis is one of the main steps before designing re-engineering steps to enhance a process. Traditionally, root cause analysis is performed using machine learning techniques that are based on pattern recognition and correlation. However, correlation does not necessarily imply causation. Thus, applying these results, especially when affecting people (e.g., by blaming, firing, promoting), can result in unfairness. For example, in a hospital, is it fair to say that the cardiac surgeon with the highest mortality rate among his/her patients is the worst surgeon? Or is he/she the most experienced one who gets the hardest cases? Several factors must be considered to infer causal relationships. Possible reasons for situations where correlation does not imply causality include the Simpson-paradox [26] and (sampling) bias in the data. Two interesting lines of research for this application are:

- providing methods to distinguish causality from mere correlation, and
- providing methods for evaluating the extent to which a particular cause is responsible for an effect (outcome)

Impact of process mining techniques on fairness. There are several algorithms and heuristics for performing process mining tasks, each of which can be fine-tuned by adjusting various parameters. These methods have been developed to optimize various metrics, but not fairness. Moreover, some process mining techniques could distort the results of a fairness analysis. For example, it is a commonly used rule of thumb, that the discovered model should be able to explain 80 percent of the cases in the event log. However, how this filtering step affects the results of a fairness analysis, has not been studied. In general, any process mining technique that its process analysis pipeline involves filtering, ranking, or decision making (e.g., in the form of clustering or classification) is prone to causing or amplifying discrimination. Promising lines of research in this area include:

- investigating the effect of process mining techniques in terms of the possibility of causing/reinforcing discrimination,
- developing fairness-aware quality measures for process models and event logs,
- investigating the effect that applying confidentiality preserving techniques has on the fairness of event logs, and
- providing methods to find and remove the root cause of discrimination in processes.

[1] Even though this aspect of fairness is not the main focus of this paper, we mention it for completeness.

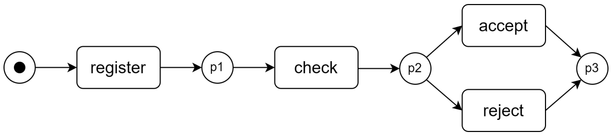


Fig. 3: An example of a simple process.

event identifier	case id	activity name	timestamp	resource	gender
e_1	1	register	19.03.3019	Alice	female
e_2	1	check	20.03.3019	Alice	female
e_3	1	reject	22.03.3019	Bob	female
e_4	2	register	22.03.3019	Sara	male
e_5	2	check	24.03.3019	Bob	male
e_6	2	accept	27.03.3019	Bob	male

Table 1: An event log with two cases for the process shown in Figure 3.

4 Mapping Existing Fairness Definitions to Process Mining

Many fairness definitions and measures in fair machine learning have been defined in the context of classification ([8,16], also see table 2 in [20]). Most of these classification-based measures require the following inputs that are not always clearly defined in a process mining context:

1. a *dataset*, in tabular form, containing one or more sensitive attributes and possibly some descriptive attributes,
2. a *model* to analyze its outcome in terms of fairness. In a classification context, the outcome corresponds to the prediction made by the classifier.
3. a notion of *ground truth* is needed for measures assessing the errors made by the model. Such ground truth indicates how things should have been in a fair and ideal world, which in a classification context, corresponds to the ideal predictions.

To measure fairness in process mining, we are interested in assessing the fairness of the process (corresponding to model), in terms of its manifestation (analogous to outcome), compared to how it should have been (analogous to ground truth). We can assume that the ground truth is provided by a domain expert or can be computed (approximated) using a normative model. To be able to assess the discrimination level using the techniques mentioned in Section 2, we need to extract the data in a tabular form. Here, we briefly mention how to extract a data table from an event log. An event log is a collection of events, where each event refers to the occurrence of a specific activity at a specific point in time, for a specific case (identified with a specific case identifier). A *case* is defined as the chronologically ordered sequence of events with the same case identifier in the event log. An example of a simple process is shown in Figure 3. Table 1

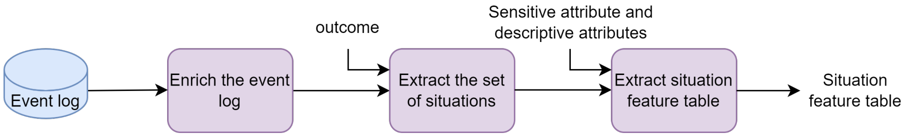

outcome
Sensitive attribute and descriptive attributes
Event log
Enrich the event log
Extract the set of situations
Extract situation feature table
Situation feature table

Fig. 4: The steps of extracting a situation feature table from an event log.

shows an event log with two cases $t_1 = \langle e_1, e_2, e_3 \rangle$ and $t_2 = \langle e_4, e_5, e_6 \rangle$ for the process in Figure 3. To turn an event log to a tabular data, we use the method explained in [23]. This method involves three steps: 1) enriching the event log, 2) extracting a set of outcome-sensitive prefixes of the cases in the event log, and 3) extracting the tabular data called *situation feature table* (Figure 4). In the following, we explain these three steps in more detail.

Enriching the event log. In this step, the event log is enriched with several derivative attributes extracted from the event log and possibly other sources. For example, we may add the decision made in a choice place as an attribute to the event that happened just before that choice place. More examples of attributes that can be used to enrich the event log include the event duration, waiting time for each event, throughput time of a case, the duration of a case on a normative model, or some ground truth indicated by a process expert.

Extracting the set of situations. In this step, we map each case to multiple prefixes of it, where each prefix ends with the occurrence of the outcome. These prefixes are called *situations*. Examples of situations include:

- If the outcome is a decision made in a choice place, each situation corresponds to the prefix of a case recorded before that place. For example, in the process of Figure 3, if the outcome is the choice made in $p2$, then the two cases in Table 1 are mapped to the situations $s_1 = \langle e_1, e_2 \rangle$ and $s_2 = \langle e_4, e_5 \rangle$.
- If the outcome is an event attribute of a group of events, then each situation is a prefix of a case in the event log where the prefix ends with one of the events of that group. For example, in the process of Figure 3, if the outcome is the duration of the event with activity name "check", then the two cases in the Table 1 are mapped to two situations $s_1 = \langle e_1 \rangle$ and $s_2 = \langle e_4 \rangle$.
- If the outcome is a case-level attribute, then each situation corresponds to a case. For example, in the process of Figure 3, if the outcome is the "throughput time", then the two cases in the Table 1 are mapped to two situations $s_1 = \langle e_1, e_2, e_1 \rangle$ and $s_2 = \langle e_4, e_5, e_6 \rangle$.

Extracting the situation feature table. In the third step, tabular data is extracted from the set of the situations in the previous step. The resulting table is called a *situation feature table*. The set of features extracted from the set of situations includes sensitive attributes and the outcome (and possibly the ground truth). This tabular data can be used to measure the level of discrimination. An example

duration-register	resource-check	gender	$p2$-choice
1 day	Alice	female	reject
2 days	Bob	male	accept

Table 2: A situation feature table extracted from the event log in Table 1 in which the outcome is the choice made at $p2$ and the sensitive attribute is "gender".

of a situation feature table extracted from the event log in Table 1 is shown in Table 2 in which the outcome is the choice made in place $p2$ (Figure 3) and the descriptive attributes are the duration of the event with activity name "register", the resource of the event with activity name "check", and the "gender". In this example "gender" is the sensitive attribute.

5 Conclusion

Organizations operate in many important areas of life, sometimes with a life-changing impact on people. This makes inspecting their impact in terms of discrimination (as one aspect of unfairness) an important topic. However, the fairness aspects of processes have rarely been considered in literature. In this paper, we discussed the placement of fairness in the process mining realm.

We discussed fairness primarily in terms of equal and non-discriminatory treatment of individuals and groups and provided an overview of various fairness definitions to detect discrimination. We presented these definitions in a structured way using a taxonomy. Furthermore, we discussed three potential key contributions that fairness can have in process mining, again with a focus on discrimination. We also provided an approach on how to map existing fairness definitions to process mining by using situation feature tables.

In conclusion, the main question one should ask before enhancing a process with fairness-related objectives is whether the differences between groups or individuals are the result of an unjust bias towards them and whether this bias needs to be corrected. Not all cases of discrimination are unfair. A methodology to quantify the explainable and illegal discrimination in data has been presented in [18]. Moreover, to assess the fairness of a system, it is crucial to be able to justify the selected fairness measurement from a moral perspective. Therefore, it is important to consider the assumptions behind each measure. For example, in similarity-based fairness measures, it is assumed that the similarity metric expresses ground truth (or the best available approximation of it) [9]. Also, the assumptions connected to statistical parity have been discussed in great detail in the academic literature [9,15,24]. Another point to note while planning to enhance a process with fairness objectives is that the costs (such as reduction in accuracy) are often immediately realized, whereas its benefits are usually not immediate and less tangible [8].

Acknowledgment

We thank Alexander von Humboldt (AvH) Stiftung for supporting our research.

References

1. v.d. Aalst, W.: Process mining: data science in action, vol. 2. Springer (2016)
2. v.d. Aalst, W.: Responsible data science: using event data in a "people friendly" manner. In: International Conference on Enterprise Information Systems. pp. 3–28. Springer (2016)
3. v.d. Aalst, W.: Responsible data science: Using event data in a "people friendly" manner. In: Enterprise Information Systems. pp. 3–28 (06 2017)
4. v.d. Aalst, W.: Responsible Data Science in a Dynamic World: The Four Essential Elements of Data Science, pp. 3–10. Deutsche Nationalbibliothek (2019)
5. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias. In: Ethics of Data and Analytics, pp. 254–264. Auerbach Publications (2016)
6. Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A.: Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research **50**(1), 3–44 (2021)
7. Binns, R.: On the apparent conflict between individual and group fairness. In: Hildebrandt, M., Castillo, C., Celis, L.E., Ruggieri, S., Taylor, L., Zanfir-Fortuna, G. (eds.) FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020. pp. 514–524. ACM (2020)
8. Chouldechova, A., Roth, A.: A snapshot of the frontiers of fairness in machine learning. Commun. ACM **63**(5), 82–89 (apr 2020)
9. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. p. 214–226. ITCS '12, Association for Computing Machinery, New York, NY, USA (2012)
10. Ekstrand, M.D., Das, A., Burke, R., Diaz, F.: Fairness and discrimination in information access systems. CoRR **abs/2105.05779** (2021), `https://arxiv.org/abs/2105.05779`
11. Fibbi, R., Midtbøen, A.H., Simon, P.: Concepts of Discrimination, pp. 13–20. Springer International Publishing, Cham (2021)
12. Garg, S., Perot, V., Limtiaco, N., Taly, A., Chi, E.H., Beutel, A.: Counterfactual fairness in text classification through robustness. In: Conitzer, V., Hadfield, G.K., Vallor, S. (eds.) Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2019, Honolulu, HI, USA, January 27-28, 2019. pp. 219–226. ACM (2019)
13. Grabowicz, P.A., Perello, N., Mishra, A.: Marrying fairness and explainability in supervised learning. In: 2022 ACM Conference on Fairness, Accountability, and Transparency. p. 1905–1916. FAccT '22, Association for Computing Machinery (2022)
14. Harrison, G., Hanson, J., Jacinto, C., Ramirez, J., Ur, B.: An empirical study on the perceived fairness of realistic, imperfect machine learning models. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. p. 392–402. FAT* '20, Association for Computing Machinery, New York, NY, USA (2020)

15. Hertweck, C., Heitz, C., Loi, M.: On the moral justification of statistical parity. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. p. 747–757. FAccT '21, Association for Computing Machinery, New York, NY, USA (2021)
16. Hutchinson, B., Mitchell, M.: 50 years of test (un)fairness: Lessons for machine learning. In: danah boyd, Morgenstern, J.H. (eds.) Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019. pp. 49–58. ACM (2019)
17. Jones, M.B.: Moderated regression and equal opportunity. Educational and Psychological Measurement **33**(3), 591–602 (1973)
18. Kamiran, F., Žliobaitė, I.: Explainable and non-explainable discrimination in classification. In: Discrimination and Privacy in the Information Society, pp. 155–170. Springer (2013)
19. Kusner, M.J., Loftus, J.R., Russell, C., Silva, R.: Counterfactual fairness. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 4066–4076 (2017)
20. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR) **54**(6), 1–35 (2021)
21. Pfohl, S.R., Foryciarz, A., Shah, N.H.: An empirical characterization of fair machine learning for clinical risk prediction. Journal of Biomedical Informatics **113**, 103621 (jan 2021)
22. Powers, D.M.: Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061 (2020)
23. Qafari, M.S., v.d. Aalst, W.: Fairness-aware process mining. In: OTM Confederated International Conferences" On the Move to Meaningful Internet Systems". pp. 182–192. Springer (2019)
24. Räz, T.: Group fairness: Independence revisited. In: Elish, M.C., Isaac, W., Zemel, R.S. (eds.) FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021. pp. 129–137. ACM (2021)
25. Saleiro, P., Kuester, B., Stevens, A., Anisfeld, A., Hinkson, L., London, J., Ghani, R.: Aequitas: A bias and fairness audit toolkit. CoRR **abs/1811.05577** (2018)
26. Simpson, E.H.: The interpretation of interaction in contingency tables. Journal of the Royal Statistical Society: Series B (Methodological) **13**(2), 238–241 (1951)
27. Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., Baeza-Yates, R.: Fa*ir: A fair top-k ranking algorithm. CoRR p. 1569–1578 (2017)
28. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28. p. III–325–III–333. ICML'13, JMLR.org (2013)