Fair Engineering of Machine Learning Systems – Lessons Learned From a Literature Review

Julian Sengewald
TU Dortmund
julian.sengewald@tu-dortmund.de

Richard Lackes
TU Dortmund
richard.lackes@tu-dortmund.de

Abstract

With the growing prevalence of AI algorithms and their use to prepare and even execute decisions, there is increasing debate about whether the results of machine learning systems tend to be fairer or more unfair. When faced with engineering a fair machine learning solution in practice, trade-offs arise between conflicting fairness notions. We conduct a literature review on this topic. The results of our review indicate that a slight consensus exists that the human concept of fairness is much broader than what lies in the scope of current fairness metrics. We discuss the context of judging fairness metrics. We also find that, albeit much research already has been done, there is room for improvement when seeking to generalize the findings across different scenarios.

1. Introduction

Because of documented misbehavior in machine learning algorithms, the topic of algorithmic fairness has attracted much attention in recent years. For example, in healthcare applications [1], crime prediction [2], or ad-delivery [3]. All these cases have raised a significant debate about algorithmic fairness in research. For example, research was conducted on synthesizing the causes of unfairness in machine learning [4], algorithmic measurement of fairness [5], or optimization methods to achieve a certain notion of algorithmic fairness [6].

2. Background

2.1. Machine learning

For clarification, the concept of machine learning systems should first be formally specified to be able to define the various fairness specifications on this basis precisely. The machine learning system $h(\cdot)$ will allocate a benefit to an individual instance x if $h(x) = 1$. Additional information is carried by the real class label y where $y(x) = 0$ is ineligible and $y(x) = 1$ is eligible. If the machine learning system predicts $h(x) = 1$, but $y(x) = 0$, then the system produces a false positive (FP), whereas $h(x) = 0$, but y(x)=1, is a false negative (FN). In the remaining cases, the prediction is correct. Often, the predictions that the machine learning system takes are defined on a probability domain, i.e., $h(x) \in (0,1)$, which can be interpreted as a score. In such cases, instances are classified as belonging to the eligible outcome if the score exceeds a predefined threshold, i.e., $h(x) > \tau$.

2.2. Gateway and selection decisions

We classify two types of decision-making that one can find in problems where machine learning may be applied, and fairness is a concern: (1) gateway decisions (2) selection decisions. A gateway decision would be characterized by having to decide about the treatment of a particular instance. Depending on the decision at the gateway, the instance would experience completely different treatments (e.g., bail or no bail decision). In such applications, we are primarily concerned with the quality and the costs and harms of a wrong decision. The costs of an FP and FN are determined by the wrong submission to a certain branch of a treatment process. Selection problems are due to limited resources, such that even when $h(x) = 1$ not every instance receives the benefit (e.g., resume selection for job interviews). If there were infinite resources, there would be no classification costs (and no selection problem). Thus, in selection problems, the cost of an FP is mainly defined by the fact that an FN cannot receive the benefit. Distinguishing between those two types of decision problems may help to understand situations of unfairness.

URI: https://hdl.handle.net/10125/79942
978-0-9981331-5-7

2.3. Fairness metrics

Fairness can be defined either at the individual level or at the group level [7]. We are concerned with fairness at the group level. Technical fairness measures can quantify systematic biases in machine learning systems that lead to disproportionately harming one group. Different fairness models can be defined with this configuration:

- Demographic parity (also known as statistical parity): equal allocation of the benefit, e.g. [8]
- Equalized odds: equal true positive rate and equal false-positive rate across groups [9]
- Equal opportunity requires an equal true positive rate across groups [9]

The precise mathematical definitions of these metrics and their components are given in Table 6. An example of the meaning of a fairness metric is $P(h(x) = 1 | x \in g_i)$, which is the probability of how often the machine learning system will allocate a benefit to the group i.

2.4. Case study

The case study illustrates practical problems when engineering a fair machine learning solution, which motivated the following literature review. We used the German credit dataset from the UCI machine learning repository for the case study. The task of the case study is to predict failed/non-failed credits according to a set of input attributes. This problem is modeled using logistic regression. Suppose there is only enough capacity ϕ at the bank to process 100 credits. The bank would decrease the score predicted by the machine learning model and grant credits to the 100 top applicants. The resulting threshold is then τ (e.g., 0.93). Suppose that there are two groups $g_1 = \{x: x_{Age} \leq 25\}$ and $g_2 = \{x: x_{Age} > 25\}$. The corresponding fairness metrics are given in Table 1.

Table 1. Equal group threshold

Fairness metric	Groups	
	g_1	g_2
$P(h(x) = 1 \mid x \in g_i)$	0.02	0.12
$P(h(x) = 1 \mid y(x) = 1, x \in g_i)$	0.04	0.16
$P(h(x) = 1 \mid y(x) = 0, x \in g_i)$	0.00	0.02

The result of this procedure is very unfair for the younger group of credit applicants. The reason for this lies in the distribution of the score across groups (Figure 1). The distribution of scores for younger people is shifted to the left.

A problem here is, of course, that the loan default yes/no plays a role as well as the loan amount. The defaulted loans of the young are only 82% of the loan amount of defaulted loans of the old on average.

This could justify a group-specific threshold because the financial risk for the bank is lower for the young (to make things easier without considering repayment/interest rates). Also, equal access to financing is a concern in society because younger people may not have had so much opportunity in their lives yet to develop the financial strength necessary to withstand a credit application check and are otherwise left out. Addressing this issue, we set a group-specific threshold τ'_g where we spread the resources proportionally according to a convex combination of the group "eligibility"-rate $\phi * P(x \in g_i | y(x) = 1)$. Note that this also meant that $\tau'_2 > \tau$ and $\tau'_1 < \tau$. The corresponding fairness metrics are reported in Table 2. While the situation for the young has improved, the older are now slightly worse off than before.

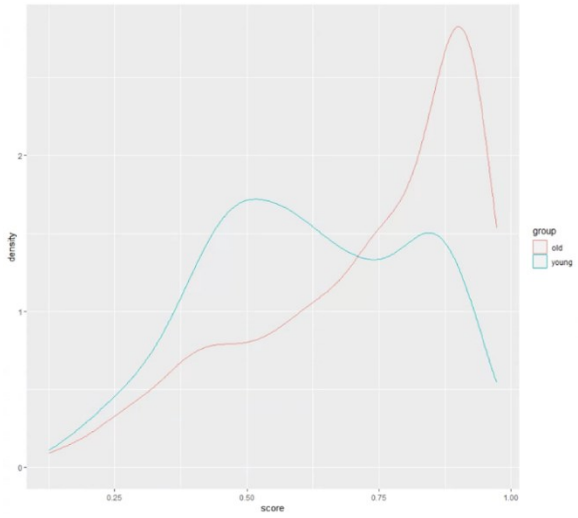

Figure 1. Distribution of scores

Table 2. Unequal group threshold

Fairness metric	Groups	
	g_1	g_2
$P(h(x) = 1 \mid x \in g_i)$	0.06	0.11
$P(h(x) = 1 \mid y(x) = 1, x \in g_i)$	0.09	0.14
$P(h(x) = 1 \mid y(x) = 0, x \in g_i)$	0.01	0.02

A third suggestion would be to skip the age attribute from the machine learning model. Indeed, the overall fairness situation improves, but overall performance goes down because more ultimately defaulting credits will be classified as non-defaulting credits. Similar considerations can also be done for

different machine learning algorithms, yielding different fairness and overall predictive performance.

Since there the fairness metrics pose a trade-off, and we ask which fairness metric does correlate most with layman perception of fairness. We formulate our research questions (RQ):

RQ1: What is the current state on which fairness metrics will be regarded as the fairest by the public?

RQ2: What contextual factors are important for implementing human notions of fairness into fairness metrics?

To investigate this subject, we conducted a review of the related literature.

3. Research methodology

3.1. Search query

We employ a variety of compositions of search strings [("fairness" OR "justice") AND "judg*" OR "perce*") AND ("machine learning" OR "artificial intelligence" OR "algorithmic decision making")]. The search was conducted for the database fields abstract and title.

3.2. Time period and other search criteria

We chose the years from 2016 to 2021 as the time period for this research. This is because the topic of algorithmic decision-making would not have been generally understood in the general public population. If the database offered the option only to include peer-reviewed research, we chose that option; otherwise, we used peer review as an inclusion criterion for the search results.

3.2. Inclusion criteria

As inclusion criteria for all search results from the primary search query, we choose:

- The title, abstract, or introduction of a paper must be related to the perception of algorithmic fairness.
- The paper is a peer-reviewed research article (including conference proceedings).
- The paper is about empirical research, not a technological artifact, algorithm, method, or philosophical discussion.

Table 3. Databases searched

	Found	Included	Backward	Forward	Total
ACM	17	11	4	8	23
IEEE/AIS	0	-	-	-	-
Total	17	11	4	8	23

All those inclusion criteria must hold for a paper to be included in the primary database. On selected papers, we conducted a forward and backward search. The result of this whole process is given in Table 3.

The databases we have chosen reflect the associations related to the community of information systems.

4. Results

4.1. Overview

We identify the following meta topics in the literature as summarized in Table 4:

Table 4. Meta-topics in literature

Fairness Metrics	[10–15]
Transparency	[16–19]
Use of sensitive attributes	[14, 20–22]
Human vs. ADM	[10, 12, 19, 23–28]
Methodology	[29, 30]

The studies also differed in the scenarios that were considered. However, most studies dealt with problems in legal justice (esp. risk of reoffence prediction), as shown in Table 5.

Most studies considered gateway decisions; only a few studies concerned resource allocation and selection: [13, 27, 28].

In the following, we present more detailed results:

Table 5. Overview of scenarios

Work	
Hiring (resume selection)	[26, 27]
Evaluation, promotion	[17, 27]
Task scheduling	[27]
Training	[10, 26]
Justice	
Small offences (parking tickets)	[25]
Starting prosecution/lawsuit	[25]

Risk of re-offense (Granting parole/bail)	[11, 12, 15, 20, 22, 31]
Child protection	[14, 22]
Education	[23, 30]
Health	
Diagnosis and Treatment	[11, 22, 24, 25, 30]
Fitness recommendations	[25]
Autonomous driving	[32]
Media (News recommendation)	[25]
Account blocking (banking, social platforms)	[17, 25]
Banking Loan	[13, 17, 22]
Social welfare	[22, 30]
None/not-classified	[33, 34]/[28]

4.2. Fairness metrics

First, we matched the fairness metrics in each study to the closest fairness metric according to our classification. The aggregated result is depicted in Table 6.

Table 6. Overview of fairness metrics

Fairness Metric	**Freq-uency**
Demographic Parity (=DP) $P(h(x) = 1 \mid x \in g_i) = \vartheta$	4 [11–14]
Equal error rates (=EER) $P(h(x) = 1 \mid y = 0, x \in g_i) = \vartheta_{error,0}$ $P(h(x) = 0 \mid y = 1, x \in g_i) = \vartheta_{error,1}$	5 [10–12, 14]
Equal false positive rate (=FPR) $P(h(x) = 1 \mid y = 0, x \in g_i) = \vartheta_0$	1 [11]
Equal false negative rate (=FNR) $P(h(x) = 0 \mid y = 1, x \in g_i) = \vartheta_1$	1 [11]
Equalized odds (EO)	2 [13, 14]
Equal accuracy (=Acc) *Equal error rates imply equal accuracy*	4

The matching of the studies to the corresponding fairness metrics was carried out as depicted in Table 6. We matched demographic parity with the following denominations in the papers =DP [11], equal outcomes [12], equal resource allocation [13], statistical parity [14]. We matched =ERR with the following definitions in the corresponding papers: equal error rates [10], EP [11], equalized odds (equal FPR and FNR) [14]. The term accuracy was used in [12]. Equal error rates imply equal accuracy and vice versa; hence, this implies equal accuracy if one favors equal error rates. Therefore, the matching of equal accuracy and equal error rates is the same. = FPR and =FNR, we matched FNP and FPP [13]. For equalized odds, we matched the equalized odds from [14] and equal rates from [13]. The latter considers a decision in which a limited number of resources is split proportionally to repayment ability [13]. This implies that the decision is independent of group membership but conditioned on the true outcome and. In our view, this closely matched the equal opportunity and equalized odd definition of fairness, which measure meritocratic and non- meritocratic allocation. Although it should be bearded that [13] studied a problem for allocating a continuous benefit, equalized odds were initially proposed for binary outcomes. Further, they consider individualized instead of group fairness [13]. It may be that fairness at the individual level is considered differently than at the group level; but this may also depend on the amount of information given. The way in which the study is conducted, there was no difference between the two compared individuals except their ethnicity (group membership) and repayment ability. Concerning our research questions, we looked at each study, comparing pairwise the metrics under consideration and counted how often they performed better (i.e., preferred by a higher number of people) against the remaining metrics. Demographic parity was in 57% (4/7) pairwise comparison the most preferred metric [11–14]. ERR was in 33% of the pairwise comparisons (3/9), the most preferred metric [11, 12]. =FPR was in 83% (5/6) cases the most preferred metric [11, 12]. =FNR was in 25% (1/4) of the cases preferred [11], but this could be due to the framing (see also Section 4.3). Finally, considering the comparison of equalized odds with DP, the former was always preferred [13, 14]. One study contained metrics that we did not find in other studies [11], and they were also the least preferred; those metrics were excluded. We also checked the qualitative results of the studies we reviewed. In qualitative interviews, experimental subjects were not always willing to sacrifice overall accuracy for increased fairness (exceptions include if a larger or more disadvantaged group would benefit) [14]. A slight preference for favoring the disadvantaged group (affirmative action) was also found in other contexts [13]. Hence, one should also take such aspects as group size and disadvantage level into account. Such aspects may be helpful when developing new fairness metrics. How can these results be interpreted? First, EO and =FPR appear in conjunction to be the most favored metrics. The next most favored is DP. Albeit DP would imply that we would not have a decision and thus a machine learning problem. So, one needs to be wary of overinterpreting that result. Also, the way of aggregation can affect the ranking. We aggregated over pairwise comparison within a separate analysis in

the literature we examined. A caveat is that this could give a single study much weight if this study conducted many comparisons.

4.3. Experimental procedures

After having reviewed the experimental procedures, we identified four criteria that could contribute to the further comparability and futility of such studies in the future:

- Framing costs of wrong decisions.
- Visualization of scenarios.
- Availability of no-choice option.
- Defining the target population

The framing of the costs of a wrong decision might explain two seemingly contradictory results concerning the preference for equalizing FP. For instance, in [11] the costs of FP and FN were disproportionate in the two scenarios examined (granting parole vs. diagnostic analysis in healthcare). The cost of a FP and FN are difficult to compare when predicting the risk of reoffence for the purpose of granting parole, because in the case of a FP, the cost of inaccuracy is borne by the convict, but the cost of an FN is wholly borne by the society. The description of the scenario given "[…] *A defendant falsely predicted to reoffend can unjustly face longer sentences, while a defendant falsely predicted not to reoffend may commit a crime that was preventable*" [11] also (over-)emphasized this aspect in comparison to the case of health risk prediction where *"[…] a patient falsely diagnosed with high risk of cancer may unnecessarily undergo high-risk and costly medical treatments, while a patient falsely labeled as low-risk for cancer may face a lower chance of survival*" [11]. We expect non-medical specialists to struggle to balance FN and FP costs in the healthcare scenario, whereas in the scenario related to crime, they may regard FN worse than FP as they may be affected firsthand by an FN. Other research related to machine learning applied to justice administration sought to frame FN and FP more comparable by considering bail/no-bail decisions for non-violent crimes and not mentioning the possibility of committing further crimes for FN [12]. Differences may exist due to the samples' differing compositions. Healthcare scenarios are also sensitive towards institutional cross-country differences, such as the existence/coverage of public insurances. Apart from institutional differences, there are also cross-cultural differences [32]. However, most studies have been conducted with samples obtained from Amazon Mechanical Turk (MTurk) possessing a similar composition in the studies under consideration.

Visualization of scenarios might be another point to consider when designing experiments. Visualizations in the reviewed literature can be divided into instance-based and aggregated depictions. Instance-based representation can be, for example, binary, in which a single instance received or did not receive the benefit [10], or pairwise, in which the classification of two individuals are compared [13], or depict multiple instances [11], or be supported by pictorial depiction [11, 14], including the use of confusion matrices [14]. Aggregated depiction can be based on a multi-metric [12] or supported by a diagram [12]. In sum, many visualization types have been used in experimental studies. Some researchers, however, pointed out that the cognitive load incurred by complicated visualization practices, e.g., multiple instances, could affect comparability because experimental subjects do not fully grasp the actual situation [12]. Detailed pictorial depictions of multi-instance situations require the experimental subjects to mentally calculate fairness metrics, whereas, in single/multi-metric representations, the aggregation has already been done.

Furthermore, photographs, when presenting the experimental vignettes, influenced how female experimental subjects judged fairness [35]. This is actually of importance because of the plenty of results about how perceived fairness can be affected by demographics (e.g., age), and the domain of decision can also affect the effect of demographics [30] At the same time, the demographics of the experimental participants themselves did not affect the unfairness perception of using a demographic attribute [20]. Another question is which visual best aids comprehension. First, when using a confusion matrix, it seems better to use a contextualized one (i.e., giving actual names to positive and negative outcomes) [36]. Second, for the task of comparing situations, contextualized confusion matrices are understood as well as bar charts [36]. Visualization improves experimental participants' comprehension but employing pictures should be carefully considered as it impacts the ratings.

We discovered only one study that included a no-choice option [11]; the majority of studies forced a choice between a predefined list of fairness metrics. On the one hand, it makes sense from a practical standpoint to evaluate "established" fairness metrics and then choose the one that best correlates with the human judgment of fairness. But, since we do not (yet) know which fairness metric is most suitable for human perceptions of fairness, one cannot know beforehand if the list of fairness metrics is exhaustive. This is a striking point because, in the study that included the "no option preferred"/"do not know" category, it was

a relatively often chosen category [11]. For example, research on survey methodology points out that the "do not know" option for attitudes is attributed to ambivalence and ambiguity [37]. Hence, the omission of the "no option preferred"/"do not know" category could seriously affect results. Nonetheless, careful consideration is required because "do not know" was also found increasing satisficing behavior [37].

All studies either measured fairness preference elicitation either by using a Likert scale, e.g. [10, 12], or binary choice between two alternatives, e.g. [11]. In terms of responses, there seems no substantial difference between the two measurement scales [30].

A general question is if crowdsourcing of fairness perceptions is desirable. First, many research studies use platforms like AMT and obtain skewed samples of the general population (e.g., age [11, 22]). However, the impact of such samples can be reduced statistically [38] or with in-experiment stimuli. Such in-experiment stimuli can be used before conducting the fairness perception measurement by providing deliberate exposure to varied viewpoints and diversity; this shifts a small group's vote to more closely representing the majority vote [20, 31]. Hence, crowdsourcing results can be applied to a larger population, given that one stimulates diversity in thinking. However, prior literature also points out that algorithmic fairness is all about that algorithms, machine learning, and AI work well for minorities and disadvantaged groups in society [15]. It may be questioned if majority votes are the best course of action for future research on fairness perceptions. Henceforth, more research into the perceptions of fairness among marginalized populations may be critical, e.g. [24]. For example, we may employ student populations because they are more likely to understand the implications of algorithms used in the recruitment process as they are affected firsthand. Though, one must remember that students make up most prospective hires and that people still change careers at a later age. Prior studies of in-sample differences in demographic effects are also mixed [10, 22]. Hence, an important question is who the target population is when doing crowdsourced design of ethical AI systems.

4.4. The context of judging fairness metrics

We discussed the fairness perceptions of fairness metrics in the preceding sections, but the context of fairness judgment also needs to be considered for understanding the limits and potential future avenues for research. Since all experiments usually involve asking an experimental subject to judge a situation affecting a group of individuals, it seems natural to consider the contextual effects of who you ask to judge whose allocation.

First, the recipient's attributes should be considered (e.g., age, gender, or ethnicity). Overall, some demographic attributes such as age [14, 20, 30], health status [30], criminal history [30], having children [30] seem more fair or acceptable for model inclusion than ethnicity and gender [14, 20]. This can be situation-specific but primarily not dependent on the relationship between the recipient and the fairness judger (except political partisanship) [30].

Prior experience with AI-based ADM increases its fairness perception [19]. In division tasks, the subject's outcome compared to the outcome is seen to be less fair as compared to human decision-making as a group; the more the subject knows about the algorithm (computer programming) and the greater their interpersonal power [28]). The latter aligns with another finding that revealed that mathematics and natural science majors were less inclined towards protesting against ADM [23]. These findings imply that education in machine learning and ADM makes humans believe more in technology. This is not necessarily a good thing, given the documented AI misbehavior [3, 39] and that ML educated are also the ones that typically are the ones that apply ML.

On the other hand, self-perceived marginalization reduces the fairness perception [10]. In addition, differences depend on prior expectations concerning the outcome. Individuals who do not receive a benefit allocated by ADM but anticipate qualifying for it have a more pronounced perception of unfairness [10]. The management of (unwarranted) expectations seems therefore also crucial in ADM. Also, prior distrust in the domain where ADM is deployed may reduce fairness perception of ADM in comparison to human decision-making [24]. Hence, that is a similar phenomenon as self-perceived marginalization, which can reduce fairness perceptions [10].

Another approach to fairness is considering several dimensions at the same time. Such a multi-dimensional study was proposed to evaluate characteristics of a person (circumstances) that should not affect the amount of benefit (utility) they receive given the same level of meritocracy [22]. This augments the EO metric to include affiliation with several groups and an individualized utility quantification of the received benefit. They found augmented EO increases utility perceptions. Another vein of literature looked at the used features and studies if properties of these features explain fairness judgment [22]. Interestingly, a set of features collectively is predictive of fairness perception of ML decision outcomes across situations, suggesting that feature properties explain situational fairness

perception [22]. Secondly, looking at the single most predictive property, humans mainly evaluate the relevance and truthfulness of a feature [22]. Other literature found unrelated demographic attributes were not acceptable, even if they increased accuracy [21]. This suggests that the situational relevance of demographic attributes is critical for experiments on the perception of fairness, including those on metrics.

A contextual factor affecting the ratings of the fairness metrics could be the domain in which algorithmic decision-making (ADM) is applied. Human decisions are considered fairer even in tasks that usually require human skills and allow for human biases (work assignment/scheduling, hiring, work evaluation), as found in survey experiments [27]. Qualitative results from the previous study hint that humans perceive human decision-making as more fair because it may consider nuanced contextual factors (e.g., holiday plans) and be less sensitive to errors [27]. Similar results were also obtained in laboratory experiments on division tasks (e.g., sharing rent, good division) where ADM was perceived as less fair than human decisions, where a group discussion achieved the latter. It was noted that humans' perception of fairness is often rather holistic and comprises altruism/pro-social behavior. However, the capability of holistic perspective-taking was not attributed to ADM, possibly explaining why they were perceived as not fair [28]. ADM is also perceived as less fair if ADM is done too extensively (as compared to partial ADM) [26] or done in high stake situations such as criminal justice [12, 25] and resume screening [26].

In contrast, for the scenario of university admission, ADM was perceived to be fairer than human decision-making [23]. This could be because the decision-making attributes employed were perceived as relevant properties [22]. Similarly, ADM was preferred when asked about the general fairness of ADM vs. human decision-making [34]. Those studies were non-MTurk samples conducted in Germany and Netherlands [23, 34]. Another factor for ADM's perceived fairness here may be that algorithmic unfairness problems have received less attention in Europe than in the US, where most research has been done.

Research on other than the before-mentioned scenarios did not find a significant difference between humans and ADM in health-related issues and the media [25]. An explanation may be that those scenarios are inherently different from work-related scenarios, or that the scenario description was overall shorter (4-10 words [25]) than in other studies (33-76 words [27]), or the experiment was not facilitated by support staff [28].

Concerning the experimental procedures, we noted there are also some problems with scenario settings. As revealed by answers to open-ended questions, humans could misinterpret judging the fairness of the overall situation instead of comparing the outcome as produced by either a machine (e.g., judging short notice, or that a process is fair because everybody is subject to the process, or the transparency of the process) [27].

In sum, there seems to be some evidence that humans perceive ADM as less fair because it does not comprise all aspects perceived as necessary. The case of algorithmic discrimination was not raised, while the problem of algorithmic sensitivity towards errors was. Interestingly, while ADM is praised for its capability of procedural fairness and treating everyone the same, this was not necessarily what most experimental subjects perceived as fair. Instead, there seems to be a preference for exceptions to the general rule, even though humans agree that this would constitute a deviation from the principle of procedural fairness treating everyone the same. An absence of concerns that machines could be discriminatory was also noted in previous qualitative work [33].

In finding a fairness metric that best matches human notions of fairness, a perceived less fairness of ADM could affect the ratings because humans distrust algorithms in general. So, based on our findings thus far, the answer to the question of whether ADM is regarded to be more fair by experimental participants than human decision making is best summarized informally by "Yes, ADM is fair according to what you mean by fairness, but it is not really what I mean by fairness". Hellberger and Araujo put this as "fairness is not justice" referring to many other aspects that humans find just just [34].

But the broader concept of fairness as understood by humans that emerged from the review also benefits from the perspective of IS and management researchers. The wish for the availability of human intervention fits into the picture of what is known in IS and management science from the literature on algorithm aversion [40]. Human intervention on ADM by single actors was also documented in the public sector, what they denominated as upstreaming done by "street-level bureaucrats" [41]. To this related is the issue of employees' fairness perceptions in hiring (which was not included in our initial literature search). An ethnographic study accompanying a rollout of an AI hiring system at a large company and documented many examples where human interventions on the "neutral" algorithm were requested: lowering the threshold for the previous intern, letting applicants pass that were just on par off with the critical threshold, or allowing for different

thresholds across countries because application numbers were different [42]. All these interventions result in an unequal process because the threshold was different for different instances and did occur through human intervention and not the algorithm. So, while the availability of human control over algorithms might increase the adoption of algorithmic decision-making, there also might be a risk for manipulation by single agents from what they perceive as fair. This problem is also related to "fairwashing" of machine learning models due to their intransparency [43].

Another topic is the role of ML and software developers in ensuring fairness [41, 44, 45]

To summarize, the seemingly innocent question of "Human or AI" involves many future research opportunities, such as developing a conceptualization of fairness and system design.

6. Findings from our review

Humans have very complicated perceptions about what constitutes fairness in a particular situation. Moreover, these perceptions include considerations that are not covered by the fairness metrics.

Fairness perception can be improved if the possibility of human intervention or overwriting of ADM is included in the process. However, there are risks of (involuntarily) manipulation of the ADM through human actors.

7. Implications for research

We list the implications for future research:

- Current fairness metrics may not be exhaustive.
- Scenarios are sensitive to many factors. Therefore, there may be a need for a scenario bank containing calibrated and parametrized situations (e.g., similarly as the information systems community already uses the Inter-Nomological Network for identifying construct identity [46]).
- There is a need for a better conceptualization of fairness preferences in algorithmic decision-making.
- Enhancing fairness also means thinking about the costs of wrong decisions carefully. Our taxonomy of gateway and selection decisions may be helpful.
- There are few studies on fairness perceptions in algorithmic hiring, even though this is a topic of interest for the broader IS community [39, 42].

8. Conclusion

Recently, the concept of algorithmic fairness has gained traction. However, what is the most preferred metric of fairness? A few studies have been undertaken to crowdsource fairness perceptions to determine the statistic with the highest association with layperson fairness perceptions. We summarized the current literature on that topic. We aimed to provide an overview that other researchers might utilize to perform similar crowdsourcing experiments. For this, we also reviewed the experimental procedures because, to the best of our knowledge, as the topic is relatively new, not so much is known yet about how to do research intersecting machine learning and human perceptions.

Additionally, we explored some of the drawbacks to such undertakings. That is, we discussed the ethical implications of crowdsourcing fairness perceptions. Here, it is essential to address the target population to ensure algorithmic fairness. Furthermore, we also discussed the circumstances and possible dangers of human intervention in ethical machine learning.

9. References

[1] Obermeyer, Z., B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations", Science, 366(6464), 2019, pp. 447–453.

[2] Angwin, J., J. Larson, L. Kirchner, and S. Mattu, "Machine Bias", ProPublica, 23.05.2016.

[3] Sweeney, L., "Discrimination in Online Ad Delivery: Google Ads, Black Names and White Names, Racial Discrimination, and Click Advertising", ACM Queue, 11(3), 2013, pp. 10–29.

[4] Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning", acm computing surveys, 54(6), 2021, Article 115.

[5] Herington, J., "Measuring Fairness in an Unfair World", in Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2020. Association for Computing Machinery: New York, NY, USA.

[6] Haas, C., "The Price of Fairness: A Framework to Explore Trade-Offs in Algorithmic Fairness", ICIS 2019 Proceedings, 19, 2019.

[7] Hutchinson, B. and M. Mitchell, "50 Years of Test (Un)fairness: Lessons for Machine Learning", in Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019. Association for Computing Machinery: Atlanta, GA, USA.

[8] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian,

"Certifying and Removing Disparate Impact", in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015. Association for Computing Machinery: Sydney, NSW, Australia.

[9] Hardt, M., E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning", in Proceedings of the 30th International Conference on Neural Information Processing Systems. 2016. Curran Associates Inc: Barcelona, Spain.

[10] Wang, R., F.M. Harper, and H. Zhu, "Factors Influencing Perceived Fairness in Algorithmic Decision-Making", in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, R. Bernhaupt, Editor, CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu HI USA, 25 04 2020 30 04 2020. 2020. Association for Computing Machinery: New York, NY, United States.

[11] Srivastava, M., H. Heidari, and A. Krause, "Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning", in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019. Association for Computing Machinery: Anchorage, AK, USA.

[12] Harrison, G., J. Hanson, C. Jacinto, J. Ramirez, and B. Ur, "An Empirical Study on the Perceived Fairness of Realistic, Imperfect Machine Learning Models", in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020. Association for Computing Machinery: New York, NY, USA.

[13] Saxena, N.A., 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, 283, 2020.

[14] Cheng, H.-F., L. Stapleton, R. Wang, P. Bullock, A. Chouldechova, Z.S.S. Wu, and H. Zhu, "Soliciting Stakeholders' Fairness Notions in Child Maltreatment Predictive Systems", in Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Y. Kitamura, A. Quigley, K. Isbister, T. Igarashi, P. Bjørn, and S. Drucker, Editors, CHI '21: CHI Conference on Human Factors in Computing Systems, Yokohama Japan, 08 05 2021 13 05 2021. 05062021. ACM: New York, NY, USA.

[15] Mohammad, Y., H. Hoda, and K. Andreas, "A Human-in-the-loop Framework to Construct Context-dependent Mathematical Formulations of Fairness", in AAAI/ACM Conference on AI, Ethics, and Society (AIES 2021). 2019.

[16] Dodge, J., Q.V. Liao, Y. Zhang, R.K.E. Bellamy, and C. Dugan, "Explaining models", in IUI '19, W.-T. Fu, S. Pan, O. Brdiczka, P. Chau, and G. Calvary, Editors, IUI '19: 24th International Conference on Intelligent User Interfaces, Marina del Ray California, 17 03 2019 20 03 2019. 2019. ACM: New York (NY).

[17] Binns, R., M. van Kleek, M. Veale, U. Lyngs, J. Zhao, and N. Shadbolt, "'It's Reducing a Human Being to a Percentage'", in Engage with CHI: CHI 2018 : proceedings of the 2018 CHI Conference on Human Factors in Computing Systems : April 21 -26, 2018, Montréal, QC, Canada, R. Mandryk and M. Hancock, Editors, CHI '18: CHI Conference on Human Factors in Computing Systems, Montreal QC Canada, 21 04 2018 26 04 2018. 2018. The Association for Computing Machinery: New York, New York.

[18] Kizilcec, R.F., "How Much Information?", in #chi4good: CHI 2016 : San Jose, CA, USA, May 7-12 : proceedings : the 34th Annual CHI Conference on Human Factors in Computing Systems : San Jose Convention Center, J. Kaye, A. Druin, C. Lampe, D. Morris, and J.P. Hourcade, Editors, CHI'16: CHI Conference on Human Factors in Computing Systems, San Jose California USA, 07 05 2016 12 05 2016. 2016. The Association for Computing Machinery: New York, New York.

[19] Schöffer, J., Y. Machowski, and N. Kühl, "A Study on Fairness and Trust Perceptions in Automated Decision Making", in Joint Proceedings of the ACM IUI 2021 Workshops, April 13–17, 2021, College Station, USA, Online, 2021. 2021. CEUR Workshop Proceedings (CEUR-WS): Online.

[20] van Berkel, N., J. Goncalves, D. Hettiachchi, S. Wijenayake, R.M. Kelly, and V. Kostakos, "Crowdsourcing Perceptions of Fair Predictors for Machine Learning", Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 2019, pp. 1–21.

[21] Grgić-Hlača, Nina, Zafar, Muhammad Bilal, Gummadi, Krishna P, and A. Weller, "The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making", 1/1/2016.

[22] Albach, M. and Wright, James, R., "The Role of Accuracy in Algorithmic Process Fairness Across Multiple Domains", in Proceedings of the 22nd ACM Conference on Economics and Computation. 2021. Association for Computing Machinery.

[23] Marcinkowski, F., K. Kieslich, C. Starke, and M. Lünich, "Implications of AI (un-)fairness in higher education admissions", in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. 2020. Association for Computing Machinery: Barcelona, Spain.

[24] Lee, M.K. and K. Rich, "Who Is Included in Human Perceptions of AI?: Trust and Perceived Fairness around Healthcare AI and Cultural Mistrust", in Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.

[25] Araujo, T., N. Helberger, S. Kruikemeier, and C.H. de Vreese, "In AI we trust? Perceptions about automated decision-making by artificial intelligence", AI & SOCIETY, 35(3), 2020, pp. 611–623.

[26] Langer, M., C.J. König, and M. Papathanasiou, "Highly automated job interviews: Acceptance under the

influence of stakes", International Journal of Selection and Assessment, 27(3), 2019, pp. 217–234.

[27] Lee, M.K., "Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management", Big Data & Society, 5(1), 2018, 1-16.

[28] Lee, M.K. and S. Baykal, "Algorithmic Mediation in Group Decisions", in CSCW'17: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing : Feb. 25-Mar. 1, 2017, Portland, OR, USA, C.P. Lee, S. Poltrock, L. Barkhuus, M. Borges, and W. Kellogg, Editors, CSCW '17: Computer Supported Cooperative Work and Social Computing, Portland Oregon USA. 2017. ACM Association for Computing Machinery: New York, NY.

[29] Hannan, J., H.-Y. Winnie Chen, and K. Joseph, "Who Gets What, According to Whom? An Analysis of Fairness Perceptions in Service Allocation", in Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. 2021. Association for Computing Machinery: Virtual Event, USA.

[30] Georg, A., S. Ivan, L. Florian, W. Claudia, and S. Markus, "The FairCeptron: A Framework for Measuring Human Perceptions of Algorithmic Fairness", in Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization. 2021. Association for Computing Machinery.

[31] van Berkel, N., J. Goncalves, D. Russo, S. Hosio, and M.B. Skov, "Effect of Information Presentation on Fairness Perceptions of Machine Learning Predictors", in Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Y. Kitamura, A. Quigley, K. Isbister, T. Igarashi, P. Bjørn, and S. Drucker, Editors, CHI '21: CHI Conference on Human Factors in Computing Systems, Yokohama Japan, 08 05 2021 13 05 2021. 05062021. ACM: New York, NY, USA.

[32] Awad, E., S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan, "The Moral Machine experiment", Nature, 563(7729), 2018, pp. 59–64.

[33] Woodruff, A., S.E. Fox, S. Rousso-Schindler, and J. Warshaw, "A Qualitative Exploration of Perceptions of Algorithmic Fairness", in Mandryk, Hancock et al. (Hg.) – Proceedings of the 2018 CHI.

[34] Helberger, N., T. Araujo, and C.H. de Vreese, "Who is the fairest of them all? Public attitudes and expectations regarding automated decision-making", Computer Law & Security Review, 39, 2020, p. 105456.

[35] Mallari, K., K. Inkpen, P. Johns, S. Tan, D. Ramesh, and E. Kamar, "Do I Look Like a Criminal? Examining how Race Presentation Impacts Human Judgement of Recidivism", in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, R. Bernhaupt, Editor, CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu HI USA, 25 04 2020 30 04 2020. 2020. Association for Computing Machinery: New York, NY, United States.

[36] Shen, H., H. Jin, Á.A. Cabrera, A. Perer, H. Zhu, and J.I. Hong, "Designing Alternative Representations of Confusion Matrices to Support Non-Expert Public Understanding of Algorithm Performance", Proceedings of the ACM on Human-Computer Interaction, 4(CSCW2), 2020, Article 153.

[37] Krosnick, J.A. and S. Presser, "Chapter 9: Question and Questionnaire Design", in Handbook of Survey Research, P. Marsden and J. Wright, Editors. 2010. Emerald Group Publishing Limited.

[38] Chen, J.K.T., R.L. Valliant, and M.R. Elliott, "Calibrating non-probability surveys to estimated control totals using LASSO, with an application to political polling", Journal of the Royal Statistical Society: Series C (Applied Statistics), 68(3), 2019, pp. 657–681.

[39] Köchling, A. and M.C. Wehner, "Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development", Business Research, 13(3), 2020, pp. 795–848.

[40] Dietvorst, B.J., J.P. Simmons, and C. Massey, "Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them", Management Science, 64(3), 2016, pp. 1155–1170.

[41] Veale, M., M. van Kleek, and R. Binns, "Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making", in Proceedings of the 2018 chi conference on human factors in computing systems.

[42] van den Broek, E., A. Sergeeva, and M. and Huysman, "Hiring Algorithms: An Ethnography of Fairness in Practice", ICIS 2019 Proceedings. 6., 2019.

[43] Ulrich Aivodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp, "Fairwashing: the risk of rationalization", in Proceedings of the 36th International Conference on Machine Learning, K. Chaudhuri and Ruslan Salakhutdinov, Editors. 2019. PMLR: Proceedings of Machine Learning Research.

[44] Holstein, K., J. Wortman Vaughan, H. Daumé III, M. Dudik, and H. Wallach, "Improving fairness in machine learning systems: What do industry practitioners need?", in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019.

[45] Cowgill, B., F. Dell'Acqua, and S. Matz, "The Managerial Effects of Algorithmic Fairness Activism", AEA Papers and Proceedings, 110, 2020, pp. 85–90.

[46] Larsen, K. and C.H. Bong, "A Tool for Addressing Construct Identity in Literature Reviews and Meta-Analyses", MIS Quarterly, 40, 2016, 529-551; A1.