“☑ Fairness Toolkits, A Checkbox Culture?” On the Factors that Fragment Developer Practices in Handling Algorithmic Harms [Supplementary Material]

Agathe Balayn, Mireia Yurrita, Jie Yang, Ujwal Gadiraju
{a.m.a.balayn,m.yurritasemperena,j.yang-3,u.k.gadiraju}@tudelft.nl
Delft University of Technology
the Netherlands

ACM Reference Format:
Agathe Balayn, Mireia Yurrita, Jie Yang, Ujwal Gadiraju. 2023. “☑ Fairness Toolkits, A Checkbox Culture?” On the Factors that Fragment Developer Practices in Handling Algorithmic Harms [Supplementary Material]. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES ’23), August 8–10, 2023, Montréal, QC, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3600211.3604674

1 DETAILS ON THE METHODOLOGY

1.1 Interview participants

Table 1 introduces the distribution of participants to our interviews.

Table 1: Background of the participants in our study. Note that some participants reported multiple educational backgrounds.

Dimension	Values (and number)
Demographic information	
Nationality	US (6), Netherlands (6), India (4), Iran (2), Russia (2), Romania (2), Sint Maarten (1), Canada (1), Brazil (1), Slovakia (1), Poland (1), Greece (1), Spain (1), Ukraine (1)
Gender	male (24), female (6)
Highest education	BSc (2), MSc (21), PhD (7)
Experience with machine learning	
Work type	applications (14), research (8), both (8)
Application domain	healthcare (4), finance (3), recommender systems (related to human resources) (3), predictive maintenance (1), others
Education	computer science (25), mechanical engineering (3), business or economics (3), sociology (1), psychology (1), accountant ethics and compliance (1)
Years of experience	2 or less (13); 3 to 5 (15), 15 (2)
Experience with algorithmic fairness	
Years of experience	18 (1), 3 (3), 2 (7); 1 (2), 0.5 (7); 0 (10)
Type of experience	long-term research (6), short-term research (4), frequent use (7), irregular use (3), none (10)
Toolkit	no exp. then FairLearn (5), no exp. then AIF360 (5), exp. with FairLearn (11), exp. with AIF360 (9)

1.2 Interview use-cases

Table 2 introduces the harms we included in the two use-cases.

1.3 Questions asked to the participants during the interviews

Questions on background experience. We started the interviews by giving a brief overview of our research to the participants, and by questioning them about their background (demographics and machine learning experience). Once all required tasks were completed by the participants, we asked final questions about their fairness experiences, how they learned and work with algorithmic fairness/harms, and reasons for using a certain toolkit, as well as their broader knowledge of the responsible machine learning field. We made sure not to ask any question related to their algorithmic fairness experience at the beginning of the interviews not to bias them towards thinking of particular topics.

Questions on higher-level reflections. At the end of the interviews, we also asked general reflection questions about any other considerations they might have when building models, any additional harm they could envision, their experiences with the fairness toolkits that we had introduced (for practitioners who previously did not know these toolkits) and potential changes they would like to see in these toolkits, about algorithmic fairness and whether it can be solved as well as on the limits of fairness metrics and mitigation methods (when not mentioned earlier), about their responsibility in considering algorithmic harms, and about any other wish, doubt, or remark.

Questions on the process. While the participants were working on the tasks, we asked them questions about their process, in order to understand the reasons for performing each exploration activity, the thoughts they had when seeing the results of an exploration, and the actions they would take based on these results, as well as to make sure they had not forgotten any activity. We especially questioned them on activities that might have a connection to algorithmic harms (e.g., observing data distributions and rebalancing the dataset based on the target labels). After the two tasks in the case of the participants inexperienced with toolkits (not to bias the participants towards certain reflections when looking at the second task), and after the first task for the other participants, we further questioned them on the algorithmic harms they had not investigated (whether they usually consider them, why or why not, how they would handle them) during their exploration of both tasks, and on the harms that could be resulting from the activities they mentioned. We identified the harms we posed questions on through our analysis of the literature (Table 2), and we also coded any other harm they could mention. We made sure to first ask vague questions (e.g., what can be issues with the activity of labeling data with crowd

AIES ’23, August 8–10, 2023, Montréal, QC, Canada

ACM ISBN 979-8-4007-0231-0/23/08.
https://doi.org/10.1145/3600211.3604674

Table 2: Examples of potential harms introduced in the two use-cases presented to participants.

Category	Task 1: Hospital readmissions	Task 2: Medical services utilization
Distributive unfairness		
Biased dataset causing unfairness	High imbalance for various potentially sensitive attributes (e.g., `race`: 74% Caucasian, 20% African American and the rest divided in 4 other categories).	High unbalance of `race` (white at 80%, others at 20%).
Sensitive attributes	"Classic" sensitive attributes (e.g., `gender`, `race`), and other, rarer, potentially sensitive ones (e.g., `marital status`, `weight`). Proxies (`region` was synthesized to be highly correlated with race).	Same with `race`, `sex`, `age`, and question of `marital status`, `military service`. Proxies (e.g., `race` highly correlated with `poverty status`).
Conceptual limitations of metrics	Consequences of the model output not only for the patients but also for their family, not measurable.	Consequences of the model output not only for the insured but also for their family, not measurable.
Harmful datasets		
Inappropriate attributes	Utility and ethics of using the `marital status` to predict hospital readmissions.	Same for `marital status`, and `military service status`.
Inappropriate attribute encoding	`Gender` encoded as binary, `age` encoded into three categories.	`Race` encoded as binary (white, non-white).
Desirability of the ML model		
Task encoding desirability	Over-simplified and potentially irrelevant target labels (unjustified threshold of 30 days).	Potentially unethical task where insurance prices would be computed based on estimation of medical services utilization.
Impact of technical ML activities onto harms (especially unfairness)		
Missing data	Synthetically introduced to correlate with specific values of the `weight` and `medical speciality` attributes.	21% of synthetically introduced missing values for the `weight` attributes with primarily values corresponding to `gender` female, which would lead to gender imbalance if the corresponding records were droppped.
Outliers	Synthetic injection of outliers in the number of `lab procedures` attribute	Outliers introduced within one synthetic attribute corresponding to an aggregation of several other attributes.
Duplicates	No visible duplicates.	20% of synthetically introduced duplicates, that would decrease dataset size consequently as well as create certain target label inbalance if dropped.

workers), before going onto more specific questions (e.g., what do you think of potentially poor labor conditions of crowd workers), so as to see to what extent the practitioners actively think about these harms.

1.4 Other materials

Tutorial. The tutorial consisted in presenting the concept of algorithmic fairness, the ways different fairness definitions are computed and different mitigation methods are applied (concepts of data pre-processing, model in-processing, and output post-processing), as well as illustrating the use of one of the toolkits to apply these definitions and mitigation methods. We gave the tutorial with a third use-case dealing with the prediction of credits default [1, 2]. This use-case was chosen for its popularity within tutorials on algorithmic fairness and toolkits, so as to be as close as possible to what a machine learning practitioner might see first when learning about algorithmic fairness.

To give the tutorial, we shared our screen with the participants, showing a Jupyter notebook we had prepared with these concepts and examples of application of the tools on the credits default dataset. We especially presented the computation of some of the metrics on a simple logistic regression classifier, and on the same classifier to which various mitigation methods (e.g., the threshold optimizer and grid search algorithms of FairLearn, as well as the reweighing and prejudice remover algorithms of AIF360) are applied. We made sure to answer any question the participants had during the tutorial and later when provided with their second task. At the end of the tutorial whose aim was to give the participants a basic introduction to algorithmic fairness and toolkits, we asked for verbal validation from the participants to confirm we achieved our goal.

Notebooks. When working on these tasks, we made sure to reassure the participants that they did not have to code the entire exploration they would perform (only if they wished to), but they could also simply speak out-loud and report on what they would do. We had already prepared additional notebooks with code snippets that the participants might want to use, and we shared these snippets with them whenever they would mention a certain exploration activity that would correspond to the snippet. This allowed to reduce the complexity of the session for the participants, to accelerate the process, as well as to see them reflect about concrete results of the exploration activities.

Pilot Studies. Before performing the interviews, we performed two pilot studies with practitioners working at our institution. These two studies allowed us to check for the understandablity of the tasks, to refine our questions to prompt about the different harms, to better time each task, and identify relevant reflection questions, as well as to make sure that we had prepared enough code snippets to help the practitioners.

2 ADDITIONAL RESULTS

2.1 Results on the fairness toolkits

Table 3 introduces the properties of the fairness toolkits (functional and non-functional requirements) that practitioners reported as important when choosing which of the available toolkits to adopt.

2.2 Results on practices and variable rationales and factors

Table 4 describes the types of rationales our participants express when handling algorithmic harms. These rationales hint at different factors that impact the practices. Table 5 describes the types of challenges and impossibilities our participants envision when handling algorithmic fairness, showing the diversity practitioners have in the way of thinking about these problems.

Table 3: Properties of toolkits highlighted by the practitioners.

Property	Example	Comparison and contradictions
Compatibility with coding frameworks	*P3 “FairLearn is natural to use for those who work with scikit learn because it is the same API. But at the same time, there is also a lot of models working with a really huge amount of data, so they’re using MLLeap, SparkMLleaP, here FairLearn will be much harder to implement.”*	*P13 “AIF is a really good library because it has Scikit learn. This library has this kind of compatibility with the pipelines that I already use.”*
Compatibility with production	*P12 “it is not being updated very often. it has the dependencies of older versions of Scikit where something was changed and so on. So it is not very perfectly maintained, so this is one thing. So every time you add something to your production, you’ll something that will be Updated often or don’t have very many dependencies.”*	-
Maintenance	*P3 “if I want to use something, I look in which stage it is. Although AIF360 has a huge amount of stars, the amount of issues shows it is less handled than FairLearn. So I prefer FairLearn because I know that if there is a bug, it will be fixed earlier.” P13 “First, it’s taken care of by other people, not by the company that I am.”*	-
Open source	*P28 “In my next models that I will train, if there is a free (open source) tool, I will check it out and try to apply it to get more insight about how the tool works.”*	-
Ease of use /extension	*P1 “AIF360 uses this ridiculous data structure that Doesn’t allow you to. I mean, have you tried to put in your own data set in F 360? How easy was it?”*	AIF360 is mentioned as more complex to apply than FairLearn.
Functionalities	*P7 “AIF360 is more complete because it has most of the FairLearn functionalities and a few more mitigation algorithms for the group fairness and individual fairness, that is very new” P21 “FairLearn is somewhat more limited in terms of fairness enhancement because it doesn’t have anything that affect the model during training”*	FairLearn is often mentioned to have less metrics and mitigation methods available, yet one practitioner mentions its advantage in presenting disaggregated metrics.
Adaptability to algorithms and tasks	*P2 “it’s designed for tabular data mostly so there are a lot of different types of data, it’s a work in progress.” P7 “in the financial industry, some of those techniques that are published as a paper or announced in some standard packages and libraries, may not be very applicable for your problem”*	Mentioned for both toolkits.
Learning curve	*P21 “ for fairness, you’ll have a lot of problems during your job and you can’t have someone who, you’re hiring and they will need a week or two to learn the toolkit. And imagine how many other problems you’ll have. The learning curve is quite high.”*	-
Transparent implementation	*P3 “ Fair learn was more natural because it’s simpler and the majority of the things there are not black boxes. With AIF360, there are lots of things based on threshold optimizer or things that already are machine learning models, biased as well. So I would prefer to work with something that is more transparent.”*	FairLearn would be more transparent (only one practitioner discussed this point).
Documentation	*P6 “ they invested a lot in their tutorials and and all the other their guides and that that was really nice to see. and they made it very easy to use.”*	FairLearn tutorials are often mentioned *P29 “An issue I have with AIF360, they don’t have a lot of documentation on how to do this.”*, but one participant mentions preferring AIF360. *P21 “AIF is definitely better with a lot more guidance materials.”*
Socio-technical considerations	*P29 “Our choices are more deliberate about what we encourage or not, because there is this danger of giving people many tools and not educating them about what they mean. That’s a big limitation of AIF360: if you use this tool with some definitions of fairness, then you will be able to solve your problems with very business solutions.”*	FairLearn argued to provide more socio-technical information.

REFERENCES

[1] Professor Dr. Hans Hofmann. 1994. *Statlog (German Credit Data) Data Set.* https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

[2] I-Cheng Yeh. 2016. *default of credit card clients Data Set.* https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

Table 4: Conceptions around algorithmic fairness metrics and mitigation methods, and their handling. These various conceptions and practices reveal the fragmentation that takes place across practitioners around algorithmic harms.

Conception	Example
Rationales for selecting metrics	
All available metrics (P2, P9, P10, P11, P14, P16, P18, P19, P26, P27, P28)	*P2 “because this model will work in hospital with patients where fairness is important, we check all the group fairness metrics of FairLearn.”*
Prioritizing group accuracy or group output distribution metrics based on use-case type (e.g., distribution of resources, hiring) (P1, P3, P13, P21, P25, P27, P28, P29)	*P1 “I think it’s quite important that the model is accurate for people if particular resources are being distributed, like whether you actually receive care or something. So it really depends. In some cases, you really care about whether the model is accurate. In some cases you care more about whether the same proportion of people get a particular resource.”*
Prioritizing specific group accuracy metrics based on the weighing of different errors (P1, P2, P4, P6, P12, P13, P19, P28, P29)	*P6 “False negatives and false positives are both damaging. I’d have to really think of the costs of those two sides, that informs what fairness criteria you would choose.”*
Involving external information (experts or laws) (P1, P4, P6, P8, P12, P19, P22, P28, P29)	*P8 “Depending on domain knowledge, you want to know what metric you want to look at. Just by myself, I wouldn’t really have an idea what would be in this case the best metric. A doctor would know. This is either some legal stuff or just some ethical stuff that we want to make sure that’s OK. ”*
Using their own intuition	*P11 “I know there are a million different metrics. I would compute statistical parity for sure. And then I would probably go down the list.”*
Judging when the metrics values are satisfying	
Acceptability for the data subjects	*P29 “Absolute fairness is not possible to achieve. So it could be like: yes, there is some disparity, but let’s say the impacted communities sort of feels fine about that.”*
Acceptability for the model requesters	*P19 “I don’t think it’s possible to remove the entire unfairness. But I think that’s all dependent on the people that they’re making the model for, and how they react to it.”*
Acceptability for experts	*P6 “There’s a question of what is an acceptable difference in performance and I think it’s a difficult question to answer, and that’s something you want to talk to all the stakeholders about.”*
Rationales for selecting mitigation methods	
No mitigation can/should be done because the data represents the world (be it unfair or not)	*P23 “some of them comes by nature, like the data given the situation happening in the real world. So you get that bias into data, and that’s not something you can change actually, it’s by nature happening.”*
Based on image it brings to the company	*P13 “[talking about post-processing methods that flip certain model outputs] They kind of imply a bias in the process. It would be a problem for the company to say that they are doing this: if I am a company and I am saying publicly that I am imputing bias on my model, how would society react to it?”*
By experimenting	*P21 “try out a few of those algorithms which are still applicable, see if they actually maybe work better.”*
Preference for not simulating new data	*P22 “if possible, we want to re-sample the data instead of simulating data. I typically prefer if they can get the data from the source corrected, as much as we can.”*
Preference for changing the data (P9, P15, P16, P19, P20, P24)	*P9 “if you can get fair data or balanced data, that is one of the best ways to make sure that your classifier is going to be accurate on all all types and all representations of people. Ultimately, like more data has always been the best way to make a machine learning model more accurate.”*
Admitting not knowing how to choose, or having to read further the documentation	*P11 “I would just like read up on it so that I know about this strategy is better.”*
Mentioned limitations of the metrics	
No limitation envisioned	*P19 “I think for fairness these metrics work well.”*
Limitations of certain metrics said to be fulfilled by others (P8, P10, P21, P24)	When asked whether one metric such as demographic parity is enough, they answer no but instead they can use another metric like equalised odds.
Limited to account for exploitation of outputs by decision-makers	*P3 “it reminds me as well of this famous child benefit scandal, when the problem was not a model per say, but the problem was also the people who were using these predictions. They were literally doing this manual post processing of predictions according to their beliefs.”*
Dangers of fairness metrics to be used as checkboxes (P3, P6, P9, P13, P29)	*P6 “It’s easy to think: we checked the fairness box because we implemented this specific library, or this constraint when really fairness is a much broader topic.”*
Dangers of fairness metrics to remove critical attitude (P3, P6, P9, P13, P29)	*P13 “Responsible AI is an AI which is built with high quality processes, not only regarding fairness, but regarding using the best metrics, not doing something like “My metric is good, so my model is good”. No. Have a critical point of view.”*
Mentioned limitations of the mitigation methods	
Non-applicability to certain types of tasks / algorithms	*P7 “we needed to mix up some approaches in order to customize them and modify them. In some cases, there is absolutely no methodologies to tackle individual fairness mitigation, that can be applied on the loan adjudication use case.”*
Impact of one method on different fairness metrics	*P21 “Optimizing for one type of fairness will suddenly make another type of fairness worse. If I optimize for fairness between individuals, it’s possible that the fairness between groups will suffer.”*
Does not fix structural causes of injustice	*P2 “I think about demographic parity, about making the decisions equal for everyone in population. It depends a lot on the way you do this, because you can also positively discriminate to get these outcomes, and it differs by use case if this would be fair. Or you can get a population fair by making the model work less good for the majority group and then it would be demographic parity. I wouldn’t consider that fair.”*
Approach might not be ethical	*P1 “One thing that people very commonly do is use different decision thresholds. The ones that I was talking about earlier for different groups, and that’s a very easy way to get different selection rates, but what does it imply in practice? What this really means is that you literally put people to a different standard. And then whether that’s justifiable or not, it really depends on the scenario.”*
Biases users to take technical mitigation approaches when they might need to be structural	*P29 “If you find some disparity, what does that mean in the real world? Then what is the intervention you take? If you don’t understand the harm, you can’t take an intervention to stop the harm. That part is very important because there are plenty of cases where there’s an intervention that isn’t technical.”*

Table 5: Examples of impossibilities mentioned by practitioners along their process, that reveal fragmentation of practices across practitioners, and various types of factors impacting practices.

Type	Example
Inherent statistical and theoretically clashing impossibility around algorithmic fairness and absence of harms	
if considering all sensitive proxies	*P21 “We are going into territory where fairness becomes almost impossible, because it could be that Medicare and Medicaid are a proxy for demographic features: whether minorities are more likely to take Medicare and Medicaid.”*
because of all attributes being possibly sensitive	*P17 “I guess the only one that society has said it's OK to be biased on is smoking because it is probably the only one that you have conscious decision you can make about although you could argue that depending on where you're born, it is probably different probabilities.”*
simultaneously for multiple metrics	*P21 “optimizing for one type of fairness will suddenly make another type of fairness worse. if I optimize for fairness between individuals, it's possible that the fairness between groups will suffer, but also even one level lower, if I optimize for predictive parity, it's possible that the disparate impact will suffer.”*
theoretically clashing objectives around algorithmic fairness and absence of harms (e.g., privacy around data attributes and their encoding, fairness, and accuracy)	*P9 “Is the dataset collected in a way that had the informed consent of people in the data set? Or are we collecting hospital records and using that data to do something that patients were not made aware of? This healthcare case is sort of limited with what you can do because you're under health care data constraints like HIPAA.”*
Theoretically clashing objectives around the use of machine learning and the absence of harms	Employing machine learning itself might be the subject of trade-off, as it might be useful for various stakeholders to deploy a machine learning model, but this model would require privacy-infringing data (P19), or might negatively impact the environment (P28).
Objectives statistically clashing with harms	See below for these objectives.
Requirements on model objectives: Typically no requirement on algorithmic fairness and other harms	*P7 “For example, we had a company involved in paper recycling. In that case, we definitely need to make sure that the amount of data that we are requesting or any other request that we have from the client wouldn't have any side effect on the environment.”*
Requirements on system infrastructure: Deployment requirements such as easiness of deployment, easiness of update, *and easiness of monitoring*, and running time	*P29 “do you want it to be a simple model so that you could retrain it properly? Do you want something that's very small, so you can deploy it on like a AWS or on Azure” P3 “The simpler is the model, the easier it will be to deploy, the easier it will be to monitor, and the easier will be to retrain”*
Computational power in relation to environmental impact (only 2 practitioners)	*P15 “We have like 20,000 GPUs and it gives a very high accuracy like human level. On the flip side, you have this much power budget and then how do you obtain this same accuracy within any alternative algorithm? Can you achieve the same with much less compute power?”*
Data constraints: Availability of data samples/attributes, feasibility of collecting new data records, feasibility of collecting new data attributes // impact training dataset, choice of algorithmic, resulting model performance	*P5 “after I do this, one of the first things that I would consider doing is to see whether This data set is sufficient enough For running a model. sufficiency test comes from 2 perspectives. One is What kind of Choice of model that I want to use. if the data set is not large enough, I cannot use a neural network, I would End up using a Linear kind of a model which would basically have its own limitations. I would want to be Clear of that.*
Impossibility due to the complexity of the concept of fairness	
due to the complexity of the concept of fairness	*P6 “I don't think you can reach a fair model because it's hard to measure.”*
due to the complexity in accounting for the impact on other stakeholders	*P25 “[would you consider how different people might be Affected by the same output?] Should be considered, but I don't have a way to consider it in terms of improving the model.”*
due to the complexity in accounting for the impact on individuals	*P18 “This is definitely something that we should take into account. I'm not really sure how to take those into account. Maybe we could add the number of children or add more features in the data to make sure that these decisions actually. To account for those specific differences, I think that's really hard and really subjective.”*
Impossibility due to the subjectivity of the concept of fairness	
Some practitioners seem to think that despite their subjectivity, there is in theory one appropriate solution that could be defined for a certain context or at a certain level (e.g., a single country)	*P28 “I wouldn't say that someone has to have a different insurance premium when we talk about sex or race. So we would make those variables as protected. I would also say potentially age since at the end of the day, if you make it a constant that will make lives for people easier. But I think our society accepts the fact that there are different premiums if you are older. If you are in your working years or if you are young adult or you were just recently born.”*
As fairness is subjective (e.g., on the culture-level or individual-level), it is difficulty or impossible to envision a one-size-fits-all approach at any level	*P6 “I think you can ever say that you're absolutely fair. And I don't think you can ever agree between two people what their definition of fairness is. So I don't think you can reach it and I think it's because it's hard to measure and it's hard to agree what the criterion should be.”*
As interests are clashing across types of stakeholders, it is impossible for all to be satisfied simultaneously	*P21 “Ultimately, everyone cares for a model that performs well. The problem is that a model which performs well for the hospital is not necessarily a model that will perform well for the Asian people who go to that hospital.”*
Subjectivity not only for algorithmic fairness but also for harms like feature encoding	*P16 “[talking about gender being binary in our dataset] I believe that everyone can be whatever he or she or they want to be. So the data itself should respond on this society request. So I mean it is a science request and we have very complex society. And if we have an issue with describing ourselves, we need to somehow mitigate it.”*
Impossibility due to the "limits" of the practitioners (assuming algorithmic fairness is reachable in theory)	
due to limited knowledge of the practitioners and lack of guidance / regulations	*P21 “ a person who would like to learn how to build a model and is confronted with a choice of 17 different mitigation techniques will know which one to choose? Probably not” P29 “This healthcare case is sort of limited with what you can do because you're under health care data constraints like HIPAA, but I think there's a lot of other use cases where there is no regulation about what companies can do with the data they collect, and that led to a lot of issues.”*
due to biases of the practitioners and domain experts	*P8 “most of the time with the help of someone having domain knowledge because even though it could be that an expert has some unknown bias thinking “oh, we should probably look into this group”, it is also domain knowledge.”.*
due to biases of the tool developers	*P16 “someone decided that we'll go this way with these metrics. Because of different cultures, let's say a group of people who decide that equality between men and female is irrelevant, what we will do with this toolkit?”*
due to lack of tools available for the practitioners	*P11 “I would make weights protected. It's a bit tricky 'cause it's continuous, and I don't know if there are fairness metrics for that.”*
(Process) Impossibility due to the lack of incentives and time given to the practitioners from their company or model requesters	*P14 “the other challenges is that, as I told you, from a legality compliance and from the organization perspective, the appreciation should be there for you to spend the time. I don't feel like it's still there.” P22 “Everybody has deadlines and this is going to add to the work. But it is important in the long run.”*
Handling impossibilities	
Making the least-bad choice (with intuition or external inputs)	*P30 “ if I decide to optimise for demographic parity or equalised odds, it's impossible to optimise for everything, so I need to pick up specific metric that I'm going to look.” P21 “This boils down to making a rational, reasonable choice of what are we trying to optimize at the early stages? And then you know, keeping in mind that making some sort of fairness metric better, even a lot better, it can still negatively influence other metrics.”*
Neglecting the issue and focusing on model performance	*P18 “This would not really be of my concern as in having to include, for sex, I don't know, 20 categorical options. Because I feel like at the end of the day, we're not doing politics here, but we're trying to solve a problem. But if the results that we obtain are really poor because of the fact that we did not take into account these attributes or variables, then we should include them.”*
Not accounting for limitations of fairness metrics because they are better than nothing	*P8 “if you don't depend on metrics then how are you going to evaluate your model? You need to have at least some metrics to be able to say a) my model is fine, and b) my model doesn't have any harmful applications.”*