A Research Ethics and Social Impact

Positionality statement. We, the authors of this manuscript, identify with different genders and hail from different continents. We work at universities, have training in computer science and human-computer interaction, practical experience with machine learning and data science projects, bear a strong interest in critical machine learning literature, and have conducted several empirical studies with machine learning developers across the world. We are motivated by our belief that machine learning practice can be made more responsible by fostering reflections of machine learning stakeholders on the harmful impacts of machine learning. We acknowledge our positionality and the impact it might have on our study setup and the analysis of the interview transcripts. We did our best to accurately report and fairly account for all opinions of the study participants (e.g., by further discussing the findings with researchers external to the author list). While abstaining from emitting validity judgments about the interview transcripts, when relevant literature exists (especially critical machine learning literature), we added reflections on the opinions and practices of the participants based on this literature. None of the authors was acquainted with any study participants before the study, we are external to the organizations in which the study participants work, and neither we nor the participants had any stake in the interviews. In our discussions, we encouraged the study participants to freely express their opinions on the study topic reassuring them that there is no right or wrong position.

Ethical considerations statements and limitations. Concerning the participation of the ML developers in our study, we ensured that our study was reviewed and accepted by the ethics committee of our institution. Our participants were informed about the study and its potential implications, and they signed a consent form before their participation. We handled the study data according to what was stipulated in the ethics review form and the consent form. We encouraged participants to reflect on the potential confidentiality of the information they discussed –the use of common, public, datasets and use cases helped to mitigate confidentiality risks. In terms of the population of ML stakeholders we surveyed, we acknowledge the limited sample of participants we could feasibly interview for the scope of this work. Particularly, as discussed in the core of the paper, we only focused on ML developers, leaving out of consideration any other stakeholder in the ML supply chain. Besides, the ML developers we interviewed present a strong gender imbalance (due to the skew of ML developers in the world), and are all employed within Europe, the US, or Canada —leaving out of consideration any difference with ML perceptions and practices in other areas of the world—. They were raised in different countries across Europe, North and South America, and Asia and hold different ethnicities, another imbalance. As perceptions and awareness of the harmful impact of machine learning might vary based on participants' background and lived experiences, we argue that our study would merit being replicated with other participants in the future. Our focus on a single domain of application with a single type of machine learning model might have also skewed our participants towards reflecting on certain harmful impacts more than others and on disclosing only parts of their practices. Hence, we also encourage replication of our study across contexts in the future.

Adverse impact statement. While we do not envision a strong adverse impact on our study participants (cf. the above discussion on handling confidentiality), we do imagine that our research might lead to an impact on the perceptions of the harmful impact of machine learning on our participants and on other ML developers, and later on to potentially influence their practices. It might also lead to changes in the attitudes and practices of organizations

employing ML developers, deploying ML models, or regulating these models, concerning the harmful impact of machine learning. While we hope for positive changes, this might also lead to decreased disclosure intention around machine learning models — that we hope could and should be handled in future regulations.

B Additional details about our method

B.1 Questions asked during the interviews

Pilot Studies. Before performing the interviews, we performed two pilot studies with developers working at our institution. It allowed us to check for the understandablity of the tasks, to refine our questions to prompt about different types of ML harm, to better time each task, as well as to make sure that we had prepared enough code snippets to help the developers inclined to use our notebook.

Questions on background experience. We started the interviews by questioning the participants 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, 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. During the task, we asked about their process (e.g., the thoughts they had when seeing results of an exploration, and the follow-up actions they would take) to understand the reasons for performing each activity and make sure they had not forgotten any activity. After the task, we further questioned them on the algorithmic harms they had not investigated (whether they usually consider them, why or why not) during their exploration, and on the harms that could be resulting from the activities they mentioned. We identified harm to question through our analysis of the literature (Section 2.1.1), and we 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 workers), before going onto more specific questions (e.g., what do you think of potentially poor labor conditions of crowd workers), to see to what extent the developers actively think about these harms.

B.2 Other materials

Table 1 lists demographic information about the participants, and Table 2 provide additional details about the use-case we crafted and the dataset we transformed to make sure to include specific issues that could relate to harmful impacts of the subsequent ML model.

Resulting themes and codes. The coding process resulted in 13 high-level code categories (e.g., data schema considerations) with 3 to 6 intermediate levels of codes per category (e.g., sensitive attributes, inappropriate

Table 1. Background of the participants in our study.

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 for 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. (10), exp. with FairLearn (11), exp. with AIF360 (9)

Table 2. Examples of potential harm in the use-case.

Category	Task: Hospital readmissions
Desirability of the ML model	
Task encoding desirability	Over-simplified and potentially irrelevant target labels (unjustified threshold of 30 days).
Distributive unfairness	
Biased dataset causing unfairness	High imbalance for various potentially sensitive attributes (e.g., `race`: 74% Caucasian, 20% African American and 4 other categories).
Sensitive attributes	"Classic" sensitive attributes (e.g., `gender`), and rarer potentially sensitive ones (e.g., `marital status`). Proxies (`region` synthesized to be highly correlated with race).
Conceptual limitations	Consequences of the model output not only for the patients but also for their family.
Harmful datasets	
Attribute information	Utility and ethics of using the `marital status` to predict hospital readmissions.
Encoding	`Gender` encoded as binary, `age` encoded into three categories.
Impact of various technical ML activities onto these harms	
Missing data	Synthetically introduced to correlate with specific values of the `weight` and `medical speciality` attributes.
Outliers	Synthetic injection of outliers in the number of `lab procedures` attribute
Duplicates	No visible duplicates.

attributes), and 8 to 34 finer-granularity codes (e.g., automatic or expert-supported identification of attributes) that represent the different response declinations. In total, this represents 276 finer-granularity codes.

C Detailed Results

C.1 On the Harmful Impact of ML systems

Figure 1 represents the types of harmful impact of ML systems identified in the literature and across the interviews with ML developers (algorithmic fairness can be viewed as a limited subset of distributive fairness).

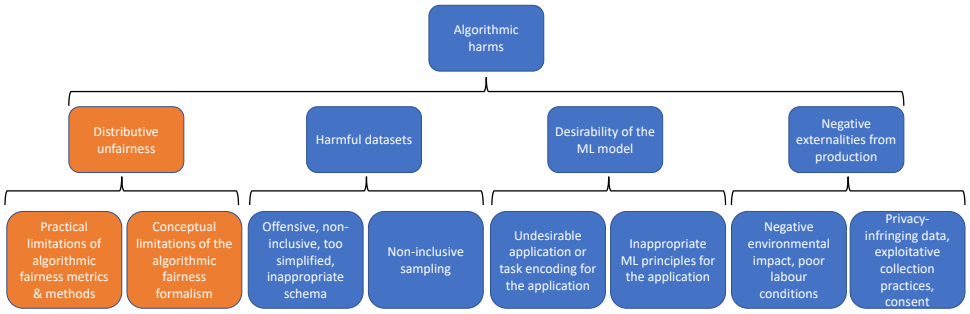

Fig. 1. Taxonomy of ML harmful impact investigated in our study. In orange we represent **the limitations of algorithmic fairness**, i.e., the current, flawed, solution to distributive unfairness, and in blue we represent the **other types of harm.**

We list in Table 3, Table 4, Table 5 the different categories and sub-categories of ML harmful impact discussed by the ML developers.

Table 3. The various conceptions of one macro-category of harms: around the ideal output distribution (i.e., distributive fairness). We do not include when the practitioners are not aware of or lacking precise information to discuss the harm, as this applies for each of these harms.

Harm	Conception	Example
Output distribution (distributive fairness)		
Ideal distribution & distributive fairness	No mitigation can/should be done because the data represents the world (be it unfair or not)	*P23 "some of them come 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."*
	Distribution representative of the real population	*P5 " what is the statistical characteristics of the real world scenario and what are the statistical characteristics of the scenario that you see here. When I say statistical characteristics, I'm actually speaking about this set of data across parameters. I focus on protected category variables."*
	Equal accuracy across sensitive features via equal distribution	*P28 "if you want to have the same probability of giving a correct answer for all societal groups, you need to be training with the dataset that is one divided by the number of social groups that are considered."*
	Middle ground: none of the two distributions is feasible to collect	*P11 "For all of these distributions, I would consult either a specialist or literature from medicine to see from all the hospital patients or just diabetes patients: does the distribution look somewhat like that?"*
	Ambiguous judgement of acceptable slack	*P28 " I would say the data static between female and male is quite balanced. You can try to make it 50, 50, but it might be the case that make it 50, 50 doesn't change much in the accuracy of the whole model because it's quite similar the number of data points."*
	Historical biases in joint distributions	*P2 "I would also look at the selection rates in historical data. Has it really been unfair in history? And do we have to fix?"* (P2, P11, P12, P20, P21, P23)
	Rare consideration of intersectionality	*P21 "checking whether we have any groups that are specifically underrepresented if we take a look at the combination of the demographic features, that's possibly something to take into account."*

Table 4. The various conceptions of two macro-categories of harms: around the desirability of the system and the development process. We do not include when the practitioners are not aware of or lacking precise information to discuss the harm, as this applies for each of these harms.

Harm	Conception	Example
Desirability of the system		
Goal of the system	Broad ethical considerations (society)	*P28 "We need to look at the bigger picture to see if our work is ethical. And that can go for the carbon footprint, the sustainability, the impact this may have in the labour market, and in warfare."*
	Morality (society)	*P17 "That's a big problem. Everybody as they get older, they have more health costs, so that'd be price gauging, the hot button issue of building based on pre-existing conditions. For health insurance, I think that's unethical."*
	Utility for the organization	*P16 "It's appropriate and relevant for the business. They want to save money or to reduce time of the workers."*
	Impact on the organization	*P25 "even the organisation where the model was employed might be affected."*
	Impact on society and "silenced" individuals	*P6 "we might ask what are the consequences of some people having access to this model and others not? Some might say this will have knock on effects in a broader scope where there are bigger consequences, where people of some descent might not trust us. So in the overall picture, it's a harm to society for us to deploy it."*
Employing ML	Appropriateness/ethics	*P1 "I would question whether we should be using ML at all? question all the assumptions that are being made."*
	Complexity & flexibility	*P3 "Everybody is afraid of changing something [with deep learning models] because if you change this, it breaks this. So we usually start with: what was the problem that you are trying to solve? could it be solved by simple query or simple statistical model, or by business rules and statistical model? If not, by machine learning? It's not about amplifying the buzz and having AI everywhere. It's about the real value of using it."*
	Right to explanations	*P27 "at least if a computer tells the person you're not getting a loan, explain why."*
Automation mode	Removing human bias collaboratively	*P27 "cause people can also have biases. It should be a doctor and in addition, this model. I don't think we should just believe the output of the model, but things should be used hand in hand with an expert."*
	Suggesting to human decider	*P4 "It's possible to automate, but it's not wise to let the model do all the work. It's important to have another medical professional opinion complementary to the model. But, building a model, if it is good, could help yield insights for the doctors to be more aware of things that they did not know before."*
	First filtering tool	*P29 "Do I think the hospital can fully automate this? No, I think you can use it as a recommendation or triage tool. You don't have unlimited healthcare resources, unlimited doctor availability, so it's sort of a triage."*
Task design	Meaningfulness	*P1 "Think whether the problem was formulated in a way that makes sense, for example why is 30 days the cut off? Is there something specific about these dates or was it just chosen out of the data?"*
	Alignment with goal	*P17 "A better way would be pay per probability, so if there's a 0% chance they're getting re-admitted, we're going to pay you more, but as there's like a 50% chance, we're going to pay you a little less, and 100% chance, we'll put the full penalty."*
	Informativeness	*P17 "we're just trying to classify you and say "are you someone that is going to use a lot of health care services or not?" I wouldn't do it this way. You're not going to get a lot of information. I'd rather use a regression."*
Development process		
Labor	Crowd exploitation	*P1 "Crowdsourcing is very important from an exploitative point of view."*
Environmental impact	Only around training	*P8 "You need a big amount of CPU time, GPU time, to train a big model. It's bad energy-wise."*
	Training and inference	*P15 " it is a very big growing problem in the whole computer science community because you have these very big models like GPT 3 which all the big companies are doing. But then you need a whole lot of compute power for them, so these are not the things that run on like one GPU or my computer."*
	Only for large deep learning models	*P9 "From my understanding, that only happens at the scale of a really large language model, the things which literally have like trillions of parameters."*
	Balancing with benefits of the application	*P4 "I have thought about this before in terms of climate AI. I have read that training a model to tackle AI is actually counterproductive because it harms the environment."*
	Scale: Not relevant as models are beneficial	*P2 "I wouldn't consider that. I think automating anything would make stuff more efficient, so I think it would save energy somewhere else."*
	Not relevant as other systems are worse	*P8 'There are better ways than reducing model training to improve the environment."*
Privacy	Consent for data use	*P18 "You need to make sure that everyone is ok with data being collected and used."P19 "look at whether the Clients are OK with their information being shared like this."*
	Anonymisation of data subjects	*P7 "Since the data are not publicly available, we need to take care of masking the data set not to release any personal information, not to release any sensitive information within the training."*
Team	Resource sharing	*P6 "This was a university cluster that we shared with others. I didn't want to hog the whole cluster for myself."*

Table 5. The various conceptions of one macro-category of harms: around the dataset schema and its population. We do not include when the practitioners are not aware of or lacking precise information to discuss the harm, as this applies for each of these harms.

Harm	Conception	Example
Dataset schema		
Feature desirability	Relevance through causal relation or correlation	*P5 "I would primarily try and understand what's the merit in using these numbers. Without a specification on the positive correlation, or the causality link to the outcome, it may not merit being used."*
	Use-case dependence	*P1 "This is tricky because it may or not make sense depending on what you're using this model for."*
	Acceptability as proxy	*P1 "it would be better to have a feature for your socioeconomic status. But race could be a proxy."*
	Completeness	*P13 "My first thought would be that the dataset doesn't have a bunch of information regarding the patient exams. I think it would be cool to include it to be more precise regarding the target feature."*
Feature sensitivity	*Sensitivity based on:* * Regulations	*P7 "In the credit adjudication use-case [..], one of the regulations was that the sensitive features should not be used as a predictor in the training of the model."*
	* Ethicality (sensitivity, relevance, offensiveness)	*P13 "If I use gender to try to predict something that is not related to gender, for example whether this person would be a good employee, the sensitive features to predict these labels, that would be bad."*
	* Exception if causally related to target label	*P13 "I don't know if race or gender is important to predict the diabetes. If this feature would be important for this problem, it wouldn't be a sensitive feature.".*
	* Exception if causally related to target label and volitional	*P17 [looking at dataset features: e.g., demographic, military service, poverty status, heart diseases, etc.] I wouldn't want to be biased on any of them. The only one that society has said it's OK to be biased on is smoking because it is probably the only one on which you can make a conscious decision."*
	Confusion with * privacy infringing features	*P15 "I would think that there are personal information. I mean their history, their age, gender and all those things apart from the things that hospital needs to note down."*
	* a parameter of a tool that would (magically) avoid discrimination	*P30 "Marital status and region: those are things that could be removed. And protected that would be more the tricky ones like sex, employment status. I'm curious to see if there will be a difference between protecting a sample and removing it."*
	Forbidden to: * use for decision making	*P7 "The sensitive features should not be used as a predictor in the training of the model."*
	* receive high feature importance for the model	*P2 "I would check which coefficients have the highest weight. Just to see on what attributes is the model predicting on, And those shouldn't be the sensitive attributes."*
	* display model disparity across this feature	*P12 "your boss just asks you to make a classifier that works fairly for some feature."*
	Sensitive proxy: any attribute correlated to a sensitive attribute	*P3 "Getting back to the financial use-case, if you know the ZIP codes, it could be really sensitive features as well because ZIP code could predict for example your economic status."*
	Sensitive proxy: not accounted due to impossibility to "unbias" the model for all attributes	*P21 "We are going into territory where fairness becomes almost impossible, because it could well be that Medicare and Medicaid are a proxy for demographic features: whether minorities are, for example more likely to take Medicare and Medicaid."*
Encoding meaningfulness	Silenced "values" (i.e., individuals)	*P15 "You would also have other races, there's not just two races. Then those kind of communities, for instance. Also for gender, I would say that to include more other genders."*
	Doubtfully aggregated values	*P20 "It's white and non white here. From the start, it's a bad feature. The people that are not white also are different between them. This should have been a category feature with all the races that are here."*
	Informativeness of values	*P27 ""Other" isn't really informative here. You see, ideally you don't want other and missing and all that. Those kind of values in your data. This is really not informative."*
	Correctness of values	*P1 "Let's look at the race column. We have mostly Caucasians, a bit of African American, unknown, Hispanic, other, Asian. Always interesting to see how race is Hispanic: that's not a race, it's just false."*
	Concept representation & measurement errors	*P1 "I would want to know how this data was collected. Like who determines the race and gender columns?" P24 "I will try to understand what each column means, and whether or not there have been mistakes in encoding the data and maybe reach out to the people responsible and say hey, what's up?"*

C.2 On Developers' Workflow, and Goals and Factors in Tension with ML Harmful Impact

In Figure 2, we show the workflow followed by ML developers to tackle the harmful impact of ML systems.

Table 6 presents the types of goals participants tackle in relation to harms, and Table 7 lists the external factors that participants might perceive in tension with harms, and might or not decide to trade-off with certain harms.

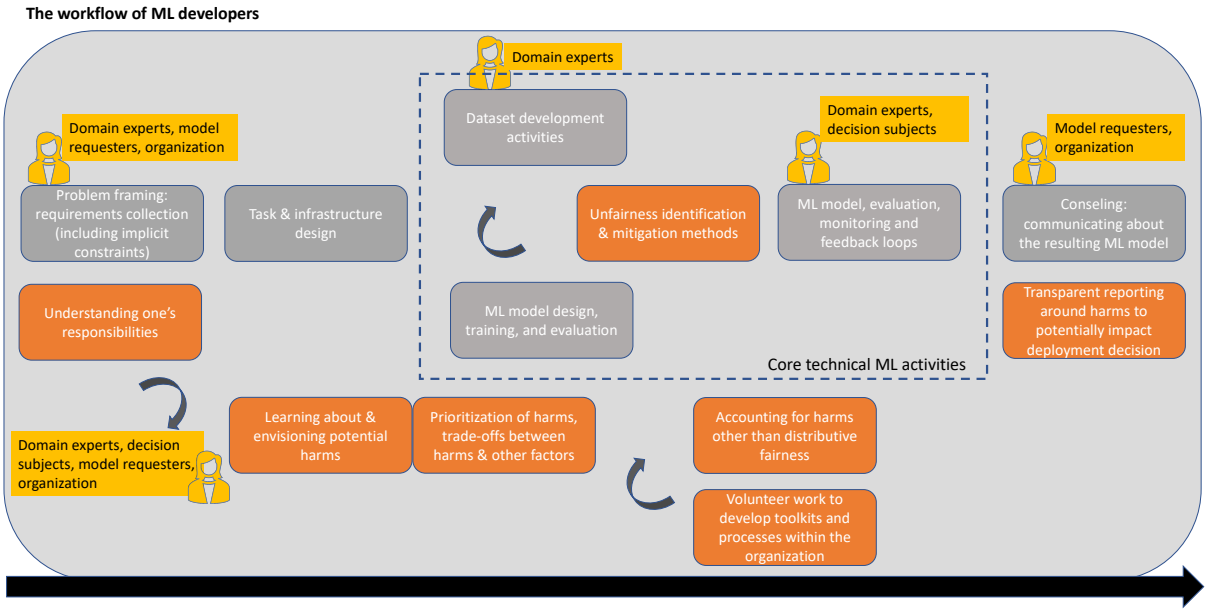

Fig. 2. The workflow followed by ML developers when considering harms across their ML system. In grey, the typical activities of the ML lifecycle performed by ML developers (these activities can impact harms), in orange the activities of the ML lifecycle that developers perform to explicitly handle harms, and in yellow the stakeholders potentially involved in these activities.

C.3 Zoom on Algorithmic Fairness

A comprehensive analysis of the concepts related to algorithmic fairness can be found here, with summaries of their practices related to fairness metrics in Table 9, and fairness mitigation methods in Table 10, as well as how they handle via (simpler) approaches sensitive features and data distributions in Table 8.

C.4 On the Sources of Harmful Impact: the perceived impact of the activities of the ML lifecycle

Tables 11, 12 describe how the participants conceived the activities in the ML lifecycle, in relation to ML harmful impact.

Received 15 March 2025; accepted 15 May 2025

Table 6. Goals formulated by ML developers along the interview sessions.

Type	Example
Modes of handling harms and potential impossibilities	
Not deploying the system (P1, P29, P17)	*P1 "if you really need the mitigation approaches for the model to be accurate or have a good selection rate, you should always question whether machine learning makes sense to use in this scenario."*
Making issues transparent for the decision makers to make the informed choice to deploy	*P6 "That would be a conversation I would have with the hospital. I could say where we're confident, and where we're not confident."*
Making issues transparent for the decision makers to account for it in deployment	*P20 "I would certainly voice my concerns towards the Fairness of a problem and how people plan to solve it"*
Not accounting for the specific issue	*P6 "There's a question of what is the current performance. We're comfortable deploying something if it improves the baseline performance, maybe it's OK if the data is not perfect."*
Mitigating this issue instead of prioritizing another objective	*P17 "I think they could automate it. But it's just those other concerns that I've addressed. You need to understand how it's affecting people and what you could do if you were getting really poor performance on one of our smaller subsets."*
Examples of rationales for prioritization of harms and other objectives	
Making the least-bad choice around impossibility (with intuition or external inputs)	*P30 " if I decide for example, to optimise for demographic parity or equalised odds. Once again, it's impossible to optimise for everything, so I need to pick up specific metric that I'm going to look." P21 "This ultimately boils down to being able to make a rational, reasonable choice of what are we actually 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."*
Compromising on certain aspects hoping to solve other issues	P2 mention that an attribute is sensitive when it should not be used for decision making, but considers that one can train a model with it as long as the model does not learn to rely too extensively on it. Some practitioners recognize that one cannot aim for equal data distributions across groups and that a middle ground is acceptable.
Neglecting the issue to focus on other objectives such as model performance	*P18 "This would not 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 (impossible?) 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."*
Judging when the metrics values are satisfying	
Ambiguous	*P2 "the difference between African American and Caucasian, their balanced accuracy is pretty equal. I think false negative rate is also pretty good. So, I think this model is for them about equal. So I would not be worried too much about these numbers."*
Value higher than (human) baseline	*P6 "We're comfortable deploying something if it improves the baseline performance."*
When one has tried mitigating as much as possible	*P20 "I strongly believe that there is no way we could achieve absolute fairness because we are biased by nature. You should try your best, and you stop when you run out of ideas and after you've done your best."*
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."*

Table 7. Other factors that might impact harms (in grey the ones that are accurately envisioned in relation to harms).

Type	Example
Requirements on model objectives	
Accuracy, type of output, inference time // impact choice of algorithm	*P15 " do I want the probability of hospital readmissions? —I would guess that is what I want then probability-based classifiers are good."*
Model explainability for decision-maker	*P8 "For the choice of algorithm, definitely in such a hospital case, you would prefer a non black box algorithm so you can have a look at: how does every feature influence my results?"*
Rare consideration of model explainability for data subject	*P27 "You should not base the output only on the model. It should also be an expert, so that's not a black box who tells the person "you're not getting a loan" and that person would be really confused of why."*
Necessity to trade-off these requirements	*P2 "I would first check a lot of different classification models. And check, which one has the highest AUC value. On that I would choose the model, but if there is a more explainable model that just lacks a bit of accuracy or AUC, then I would choose that one over the bigger models that are not that explainable. "*
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 for deployment requirements and cost // impact algorithmic choice, dataset size, and trade-off with model accuracy	*P29 "Do you want something that's very small, so you can deploy it on like a AWS or on Azure?"*
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?"*
Requirements on the development processes	
Time pressure	*P22 "Everybody has deadlines and this is going to add to the work. But it is important in the long run."*
Data constraints	
Availability of data samples/attributes, feasibility of collecting new data records or 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.*
Data types impact choice of algorithm	*P25 "There are algorithms which actually take both (continuous and categorical). You can input the range value as well and then feed categorical data as well and then also the model will work. Otherwise, these range values again need to be converted into categorical manually"*.
Using certain features for training higher accuracy/fairness models, opposed to feasibility and practicality constraints	*P6 "Right now, we have 100,000 records. If we decide that we want another feature, we have to wait a long time before we get all the data on that feature again. So we always try our best and see if it's good enough."*
Trading-off the appropriateness of the target label with the above data constraints	*P1 "In machine learning, you will often see that people choose a target label based on what happens to be available or what's easy to get rather than when you think about more statistical inference and stuff like that, then it's typically much more well thought out. Many of the issues with fairness can come from mismeasurement."*
Inherent statistical and theoretically clashing impossibility around algorithmic fairness and absence of harms	
Inherent statistical impossibility in reaching algorithmic fairness if considering all sensitive proxies	*P21 "We are going into territory where fairness becomes almost impossible, because it could well be that Medicare and Medicaid are a proxy for demographic features: whether minorities are, for example more likely to take Medicare and Medicaid."*
Inherent statistical impossibility in reaching algorithmic fairness 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."*
Inherent statistical impossibility in reaching algorithmic fairness 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)	Impossibility in reaching or measuring algorithmic fairness without accessing sensitive attributes traded off with the law forbidding to exploit these attributes *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 they 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."* 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).

Table 8. Practices around data issues towards algorithmic fairness (sensitive features and data distributions): simple approaches to identify and to handle them (in grey explicit trade-offs).

Conception	Example
Identification of sensitive features	
Mandatory according to external entity (guidelines, regulations, client, model owner)	*P11 "I know that these are legally defined. So, the EU for example, has a guideline on what are sensitive attributes. I will look at that as a baseline. Anything that's in there is protected or sensitive.*
Based on the existence of human discrimination on certain attributes	*P11 "Weight: obesity is common among people that have diabetes, so perhaps people are misjudged by doctors if they are comparatively thin for a diabetic person."*
Based on intuition	*P16 "I would say, the most obvious sensitive features are race and sex. But also status of veteran is important for me. It can also be kind of sensitive." P21 "by definition, protected groups are minorities."*
Based on background experience/-knowledge	*P3 " I already see the alarms such as race, gender and age as well."*
Based on personal reflection	*P21 "What is for me important to consider is just thinking where that data comes from, or trying to imagine what could have influenced the initial fairness of the data. Obviously, people from specific background are less willing to answer some questions, maybe in some geographical region where the data was collected, or at some time when it was collected a group was underrepresented."*
Based on information collected from other stakeholders or from the literature	*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."*
Combination of the above	*P 4 "Of course, the law might not cover everything that could be sensitive, so I would also go into data and think for each feature about whether this is something that could lead to bias?"*
Identification of proxies based on intuition	*P16 " Pregnant status would be very sensitive because it's related to the sex".*
Identification of proxies based on statistical tests	*P28 "I will check what is the correlation of each variable to each other. Basically, having a correlation matrix and checking if there is a higher local relation to those that we have protected.*
Identification of proxies: ambiguous correlation threshold definition	*P28 "Marital status. It's quite a big negative correlation. Age, there's a decent correlation. I would consider something as positive or negative correlated when it's magnitude is higher than 0.25. That's a value that I take from personal experience with my own research."*
Handling of sensitive features	
Dropping attributes: because they are forbidden/sensitive	*P7 "We had to remove the sensitive features in the training set, and then feed the training set into the modeling and model training."*
Dropping attributes: to train "unbiased" model	*P3 "I also make sure that if even I decide to drop these sensitive features, there is no more of this information ingrained somewhere in the data.*
Dropping attributes: not appropriate due to proxies	*P17 "You could argue you get rid of race and sex and just make your models blind to this sort of stuff. But it might not be truly blind because you can have like satellite features. Or like indirectly related features."*
Dropping attributes: not appropriate when they are informative of the target label	*P16 "I would see again the correlation between these attributes and target columns. I expect to see some correlation between some of them. We could keep it as it is, and we will understand the importance of different features later."*
Dropping attributes: not appropriate to monitor algorithmic fairness	*P10 "These are my sensitive attributes. it's important to leave those in. I keep it just to check if it has a weird distribution."*
Handling undesired data distribution	
Grouping the values that are too underrepresented into a larger group (P2, P8, P28)	*P8 "other groups, for instance, these bottom four are really low in number, so in order to get some insightful results, you might want to group them."*
Leaving out under-represented populations (P2, P6, P15, P21, P25)	*P15 "if I have to make a model out of this, then you have to account that the dataset itself has very few points for this category. So accounting for all of those things, I would leave out some percentage of data set which is not representational in a way."*
Dropping the attributes which display problematic distributions	*P23 "For example for some variables, if it's very biased, you should avoid using those."*
Transforming set of samples: Collecting additional data, artificially augmenting data, undersampling (P20, P25)	Naturally, all practitioners discussed the possibility to collect more samples, and some mentioned avoiding undersampling not to lose information.
Strategy depends on amount of data	*P2 "If there's only 3 Asians in the whole dataset, it wouldn't make sense to make up for that: it is not enough data to equalise over this. So I would only equalise over Caucasian and African American. Or maybe even combine others as the minority group and have Caucasian as the majority grouP"*

Table 9. Conceptions and practices around algorithmic fairness metrics.

Conception	Example
Used notions	
Group accuracy (e.g., equalized odds)	*P28 "I would compare accuracy for the races "0" and "1", and see whether the results are similar."*
Group output distributions	*P22 "I look for statistical parity and disparate impact because those are not dependent on the target."*
Individual fairness	*P21 "We can have fairness between groups, not necessarily meaning that similar individuals will get the same outcome."*
Reasoning for selecting metrics	
All 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."*
Metrics applicable for both data and outputs (output distribution based)	*P13 "I chose disparate impact ratio because it is a metric that can be applied before and after the training of a model."*
Prioritizing group accuracy or group output distribution metrics based on data correctness	*P15 "demographic fairness is very important. But sometimes, you pick a very obscure data set, then demographic fairness is not the answer if your dataset or representation is fundamentally not correct."*
Prioritizing group accuracy or output distribution metrics based on existence of causal relations between sensitive and target attributes	*P6 "Demographic parity wouldn't be used because it's possible that because of many factors, Caucasian people should be discharged at a higher or lower rate than African American, and so we don't want those to be set to be equal. We want the error rates to be roughly the same, not the selection rates."*
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."*
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 reflect underlying injustice (P1, P2, P3, P9, P18)	*P9 "In the college admission example, due to historical factors, we see correlations between certain races and socioeconomic classes, and between certain socioeconomic classes and education. Should people of different races be given equivalent outcomes? I don't think you can say yes. You have to consider and fix the underlying factors first. You can't just fix it at this top level and expect it to be done. So I can't call demographic parity enough."*
Limited to reflect certain notions of fairness	*P6 "I'm sure that if we look at the broad range of people, people have views on fairness that are defined on very different criteria than the ones that we can see in these numbers."*
Limited to account for the impact on other stakeholders	*P19 "it depends on the situation, but mostly it's not only me who could be affected, but people around me can also be indirectly affected by whatever it is. In the case of health, if I was to be discharged without being supposed to, I would be directly affected, but also my family or people that I'm surrounded by."*
Limited to account for individual outcomes (impact of outputs on each individual)	*P18 "If I don't get a credit score, it's no problem because I'm young, I have a lot of opportunities ahead for myself, but then if I were to be 50 and I'm trying to get a credit and if I'm not allowed to get one and I have 4 kids and I know I'm gonna be homeless, then maybe it's worthwhile giving me the credit, because then I'm gonna have a lot of other issues."*
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 also 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."*

Table 10. Conceptions around algorithmic fairness mitigation (in grey explicit trade-offs).

Conception	Example
Used methods	
Manual data rebalancing or attribute dropping	See Table 11.
Scoping out populations (P2, P9, P15, P25)	*P9 "I understand that most people are over the age of 60. So you can choose to limit the scope of your classifier and use this one on people who are over 60. that's one way of making sure that you're not having false positives or false negatives on these underrepresented data."*
Modeling a new task (P4, P6, P15, P17, P28)	*P6 "we actually have enough data that we might be able to train separate models. So you might not even use the normal FairLearn strategy, which is to train one model that works well across populations."*
Data preprocessing method	*P22 "we would use some of this re-weighing or adversarial debiasing kind of techniques."* (reweighing P2, P4, P11, P12, P15, P16, P21, P22, P23, P24, correlation remover P2, P3, P4, P12, P29)
In-processing method (P2, P8, P11, P12, P15, P16, P21, P24, P29, P7, P19, P17)	*P2 "After [computing fairness metrics], I would do some in-processing mitigation."* (e.g., grid search and Lagrangian classifier)
Post-processing method (P1, P2, P3, P12, P21, P29)	*P3 "You have threshold optimizer. So for example, for logistic regression, the decision threshold by default is 0.5, and you also can play a little bit with the threshold that defines whether this data point belongs to this class or to that class."*
Reduction method	*P6 "what we've done internally, it is doing this reductions approach in FairLearn."*
Selection	
Based on speed	*P6 "the major downside to the reduction approach is that it can take a long time."*
Based on amount of available data	*P6 "we actually have enough data that we might be able to train separate models."*
Based on applicability to specific model	*P12 "the cons are that they are not model agnostic. So that means that it depends on each kind of model you apply. You'll need to know all of them where they can be applied."*
Based on compatibility with deployment constraints	*P12 "When you are in production, in some cases, you won't be able to do a lot of changes. So post processing is good, you're just changing the labels and given a minimal loss of accuracy, you may just make it fair."*
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 mitigation methods	
Non-applicability to certain types of tasks / algorithms	*P7 "we needed to somehow 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."*
Inadapted solution to the cause of the unfairness	*P29 "When they were trying to test out a model to allocate poverty benefits to low income individuals, especially for food banks, Hispanic applicants were being rejected at a higher rate, and that's just because these applicants actually aren't fluent in English. They're having trouble with the application form, and so the solution to make this system more fair: just offer the form in Spanish."*
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 11. Summary (part 1) of the ways the activities performed during the machine learning lifecycle are conceived in relation to harms (in green) and other trade-off (in grey), and handled (in red), potentially influenced by other factors.

Activity	Conception	Example
Data duplicates	No envisioned harm	*P10 "I would delete one or the other, because I don't think it would make any effect."*
	Percentage of duplicates within dataset	*P4 "It's important to have them because they represent the distribution. But it depends: if there's a lot of the same occasions, you might want to trim it down a bit."*
	Removing all duplicates (P10, P20, P25)	No awareness of the different natures of duplicates (real or apparent) *P10 "I would delete one or the other, because I don't think it would make any effect or any changes."*
	Understanding the nature of duplicates to handle them	*P2 "it depends also on the use case. Why are there duplicates? How do those duplicates get into the data? It could be really similar people and then you would leave them ."*
Data outliers	No envisioned harm	None mentioned when prompted
	Cause of dataset biases and algorithmic unfairness (only P5)	*P5 "I would be cautious of eliminating outliers as it can cause bias. I would focus on statistical characteristics to know what's the proportion of outliers. "*
	Cause of population silences (only P21)	*P21 "I would look at whether we have any important outliers in the data. What could be a problem is say you know five people in this big dataset of 100,000 records spent in hospital 100 days and you know all the others spent less than 20. Then you know the question would be whether the model that I built is at all applicable to such people. I would say probably not so maybe it's best to remove records which seem to have very strong outliers. And have that caveat that you know the model shouldn't be applied in some very rare cases."*
	Indirect sign of deployment issues, in turn causing potential algorithmic unfairness (only P6)	*P6 "it is useful to look at the distribution and see if there are outliers, but only as a way to detect if there is input issues. If someone is listed as being 10 pounds, then you know that's an issue where someone entered it wrong and then I'd look at why was this entered in wrong? Is there a manual process somewhere in the chain that this is the result of? Now that I've been confronted with this fact that there's manually entered data, then I'd have to go back and think about what are the consequences of that at inference time?"*
	Dataset size, impact of removing outliers on model accuracy with or without experiment, impact of outlier handling in deployment	*P28 "deleting points just because they are outliers, that's not the right approach, because those outliers could be those that have the most information, while the ones that are located in the median in this case, or the mean, they are more common and provide less information." P9 "What I usually end up doing is training a classifier on the data with and without the outlier. Then I defer the problem to once I have more information about how the dataset has been trained with and without the outlier. My approach would be to consult a textbook." P19 "I would also check percentage of the outliers, if the outliers are less than 10% of all the data, I would discard them. If it's a little more, then I would let them and use a model that is good with outliers."*
	Understanding provenance to handle outliers	*P2 "If you have weird outliers, I would look at those rows because they're often something parsed wrongly. Then you can remove those. If there's enough data and there are some outliers, they could just be outliers, so we would keep them in. If I cannot explain why it has to be removed, then I won't remove it."*
	Adopting one of the three default approaches in any case	*P18 "If we're talking about use cases where the outliers are really Purely of an anomalous nature, you can just get rid of them. For example, having a person in our data set being 400 years old. Well, that's to my estimate, at least unlikely. So just can remove that entry because It's not really reliable."*
Missing values	No envisioned harm	None mentioned when prompted.
	Cause of dataset biases and algorithmic unfairness (only P21)	*P21 "I wouldn't drop them. People from specific backgrounds are less willing to answer some demographic questions. For instance, people from some minority group would be less willing to admit that they are using state insurance. If not dropping, I would say imputation. That depends how much time we have."*
	Cause population silences	Only P29.
	Depends on dataset size	*P2 "Depends on how much is missing. I'd impute it if there's not a lot of data missing. I'd impute it with the most recurring value for categorical columns, and for numeric data, you have regression models."*
	Stakes of the system	*P15 "It's not related to diagnosis but to re-admissions. If this problem is critical, here it's an important model for the public healthcare system, I wouldn't introduce averaging or some interpolation for imputing the missing data, because it has to be as accurate as possible. Then, I should remove the whole data point."*
	Handling by dropping records or imputing them or dropping attribute, depending on other factors, or taking one default approach	*P11 "I would look at which columns have excessive amount of missing values like one third, then I would remove this variable from the dataset. After removing columns that have a lot of missing values, I would remove all rows that have missing values so that this dataset has no missing values. The data is quite big (over 100,000 records), so if we have to remove two or three variables with missing values and then we will remove all other rows that contain any NaN, we still have quite large datasets."*
Data distribution shifts	No envisioned shift & harm	Most practitioners have not expressed any concern around data distribution shifts along their process.
	Ensuring the populations seen in deployment are represented in training	*P15 "Is this really representational of the general situation of diabetes? For instance, sometimes these things are taken from very specific hospitals, very specific region, and that region might have very specific distribution of diabetes. It's not representative of the entire country."* (P15, P22, P25, P29)
	Ensuring the model is adapted to any distribution shift happening after deployment	*P3 "I'm thinking immediately how this model will be deployed, how often it should be retrained. Usually, the biggest problem is a huge difference between production and training data. When you get more sensitive medical devices, the way the data is distributed also changes, because the bad quality medical devices will have much more noisy data and if you optimise everything and you deliberately make sure that this data will be processed in this way, then you will be literally fucked up if the quality of medical devices will be better."*

Table 12. Summary (part 2) of the ways the activities performed during the machine learning lifecycle are conceived in relation to harms (in green) and other trade-off (in grey), and handled (in red), potentially influenced by other factors.

	Conception	Example
Preprocessing	No envisioned harm	None mentioned when prompted about activities mentioned in the interview like dataset merging, feature engineering (e.g., reduction, normalisation and standardization), data format preparation (e.g., transforming string data into a one-hot encoding), data balancing, and data splitting.
	Cause of dataset biases and algorithmic unfairness (only for data splitting, data label rebalancing –P1, P29, P30–, and data annotations)	*P5 " Training-test split, I would prefer to make it absolute, looking at it in terms of proportion. The split is going to be random and the split may not be an unbiased split, so that is something that I would standardize." P11 "if we have this re-admit that is a false negative committed by the humans that decided. That's exactly what you want to avoid that the model repeats this behavior. If this proportion fits with what medical experts say, then it might be fine. It's like a cognitive bias, so I would look at these kinds of variables. And make sure that it's all representative and makes sense to experts."*
	Model accuracy & data-model compatibility	*P25 "There are algorithms which take both. You can input the range value and then feed categorical data. Otherwise, these range values need to be converted into categorical manually."*
Data labeling	Impact on model accuracy but no envisioned harm	*P15 "Labels are very important: the source of annotation can be noisy. The label itself can be noisy, so there can be misinterpretation of: OK I am a labeler and how do I interpret this?"*
	Cause of dataset biases and algorithmic unfairness (label unavoidable subjectivity)	*P20 "This is a very important source of bias, because if it's not something objective like doctors looking at X rays but something like insurance, and people manually label this based on their experience, they're 100% introducing bias. Maybe someone which is a minority would take into account bias more. But anything that is subjectively labeled is inherently biased. Because I think all the people are inherently biased."*
	Label "quality" vs quantity	*P9 "There is a very large graph of everywhere that you can have a fairness issue in a machine learning pipeline and labeling was one of them. So I think when it comes to something like Amazon Mechanical Turk, you have to decide for yourself whether the possible biases of the people labeling your data are more important. Ultimately there's a threshold: are those things more important for your use case than having massive amounts of labeled data which is something that Mechanical Turk can provide you?"*
	Improving "quality" with the labelers	*P24 "I acknowledge that there can be labeling bias. And this is again Specific on the case. in the hospital, I think I would reach out to the doctors who actually labeled the patients ."*
	No action due to unavoidable subjectivity	*P5 "I need a comfort on the quality of data. Once I have a reasonable comfort, I'll go ahead because there's no end point to trying to understand data labeling or data annotation, there will always be bias."*
Model building	No envisioned harm	*P25 "In terms of building the model, considering fairness? Didn't we consider all of these things already? we removed all the features, stuff like that. The next step after cleaning everything is model building."*
	No model impact on harms because it only comes from data	*P2 "I don't think that giving a parameter a certain value can lead to harmful implications. I think it's mostly caused by the data, not really by the model."*
	Cause of algorithmic unfairness	*P5 "there may be models where you choose hyperparameters. And the choice may induce bias. I'd look at a grid search. There is a functionality that's available in FairLearn to search all combinations of my dataset/model. And run them to know which has a higher propensity of bias. There may be impact caused by multiple other factors including the batch size, the epochs, the learning rate".*
	No awareness but benefice of the doubt	*P4 "In the model selection for sure. For hyperparameters like learning rate, I can't see the connection between how it might harm people because it just influences accuracy and other things. But I'm also hesitant to say it doesn't affect it at all because I feel you never know with these things, so you should always be cautious."*
	Accuracy, explainability, privacy, expected output type, cost of training, maintenance	*P3 "For me, 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. So if there is a choice between doing something with deep learning and doing something with logistic regression with properly engineered features. I'm gonna go with logistic regression, because it will be just easier and less expensive to run in prod."*
	Algorithmic fairness as the second stage of model building	*P9 "The first iteration will always be to investigate even the feasibility of the accuracy, 'cause the second you start trying to incorporate other things like privacy or fairness into your models, you will immediately start making accuracy tradeoffs like in privacy. It's almost by definition 'cause you're introducing noise."*
Model evaluation	Selecting various accuracy metrics by default	*P6 " I think there's the standard stuff, right? There's a confusion matrix. There's the Roc curve and the area under the curve. There's precision and accuracy plots. I would start making those."*
	Selecting various accuracy metrics based on judgment of errors	*P4 "I would train the model on the data et for whatever I've balanced on and just see the performance like accuracy, recall, precision. Depending on the use case, one metric might be better than the other. I would try to figure out whether a false positive or false negative is less worse and then figure out the metric."*
	Accounting for feature meaningfulness	*P2 " I would cheque which Coefficients have the highest weight. Just to see on what attributes is the model predicting on? And those shouldn't be the sensitive attributes."*
	Accounting for algorithmic fairness when the use-case is sensitive	*P9 "when we talk about automating a task, you can create an arguably false dichotomy between sensitive tasks and insensitive tasks or tasks which maybe require you to actually mind responsible AI concepts. For example, you're going to pay far more attention if you're trying to automate something in college admissions, versus trying to use machine learning to automate the protocol for handwriting recognition."*
	Accounting for fairness if use-case involves people	*P2 " when the use case is about making decisions for people, especially when it's for obscure groups of people. Fairness issues can really disturb groups in society."*
	Accounting for algorithmic fairness without knowing the concept	*P28 "accuracy is only a certain perspective. The performance of the model can say it's 99%, but it's not telling you how accurate it is for different groups of society. Perhaps, for instance, it could be very inaccurate for African Americans, very accurate for caucasian, and that's not reflected only in accuracy."*
	Representativity of the test set	*P6 "When we evaluate accuracy on subgroups: do we have enough data to say that we have that accuracy? False confidence is a big danger."*