Challenges in Algorithmic Fairness when using Multi-Party Computation Models

Colette Wibaut, Vincent Dunning, and Marie Beth van Egmond

TNO, Anna van Buerenplein 1, 2595 DA, Den Haag, NL
colette.wibaut@tno.nl
vincent.dunning@tno.nl
http://www.tno.nl

Abstract. While the topics of Secure Multi-Party Computation (MPC) and Algorithmic Fairness (or in short, fairness) are essential in the area of Responsible AI, they are typically researched separately. However, when multiple parties train a model in a privacy-preserving manner, fairness of the model is not guaranteed in any way. In fact, the multiple parties involved could run into several challenges when wanting to measure and mitigate unfairness. We reflect on the existing technical solutions in this field and identify three practical challenges. First of all, when computing with multiple parties, the focus lies on the computed result of a mathematical model, the *output.* Fairness assessments also cover the *outcome* of a model, i.e. what the output entails in deployment. Without proper agreements, the individual parties in an MPC setting could act differently upon the output and have a conflicting definition of fairness. Secondly, in a multi-party setting, the data is distributed. Therefore, a difference can arise between *global fairness* (evaluated across all data) and *local fairness* (across local data). Finally, fairness is not a static measure. All sorts of feedback loops can occur, some directly affecting the model when it is retrained with new data. Working with multiple parties could make this even more problematic, because each party can have feedback loops in their own system which influence the total system and fairness for others as well. In this position paper, we hope to pave the way for integrating fairness challenges into future MPC studies, an important new field of research.

Keywords: Algorithmic fairness · Secure Multi-Party Computation · Privacy Enhancing Technologies · Data security · Local and global fairness · Feedback loops

1 Introduction

Motivation for this position paper Both the use of Secure Multi-Party Computation (MPC) - or more broadly, Privacy-Enhancing Technologies (PETs) - and techniques for Algorithmic Fairness (or in short, fairness) are important and upcoming research topics in the research area of Responsible AI. A new paradigm of being able to get the insights but not sharing data is being researched and deployed. At the same time, awareness for the need of fair models is growing.

In general, MPC and fairness pursue similar goals: an ethical way of working with data. Both research areas contribute to the seven key requirements on trustworthy AI, set by the high-level expert group of the European Commission on AI [11], especially *privacy and data governance*, as well as *diversity, non-discrimination and fairness*. However, when we zoom in, the concepts of fairness and privacy can be contradictory.

First of all, measuring fairness can cause privacy issues. For example, to be able to assess a model's fairness with respect to ethnicity, one needs to use data on ethnic background. This is very sensitive data that needs to be protected. An overview of techniques to measure fairness without the use of these kind of sensitive features is given by Ashurst [2]. Note that MPC is actually mentioned as a solution here. However, this is out of scope for this article.

On the other hand, when protecting the privacy of the input of a model by using MPC, the model is not yet protected from being unfair. In his article [6], Calvi has recently started the debate on the potential 'unfair side of PETs'. However, it does not address the challenge on measuring fairness in a setting where the input data is protected, which we do in this paper.

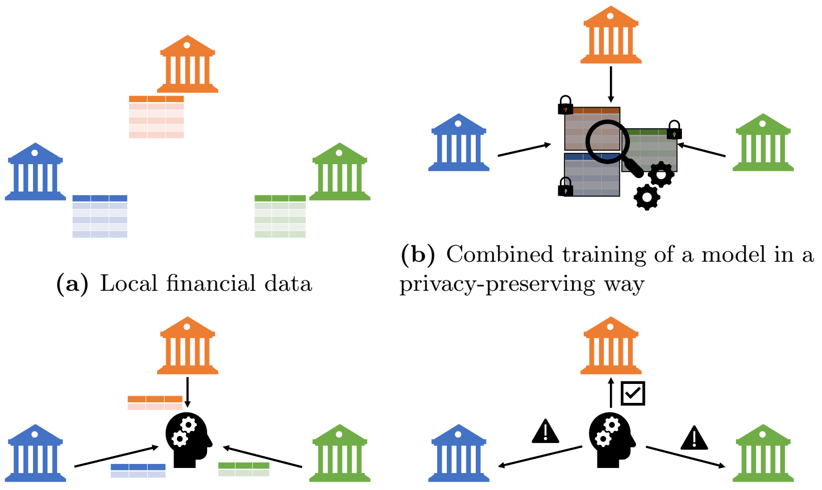
In sections 1.1 and 1.2 we first give short introductions on MPC and fairness. In chapter 2 we reflect on some existing strategies on fairness in mainly federated learning settings. In chapter 3 we will describe three potential challenges one could run into in practice, when one wants to assess fairness in a multi-party setting. Finally, in chapter 4 we conclude and discuss potential avenues for future research.

1.1 Secure Multi-Party Computation

Secure Multi-Party Computation (MPC) enables computation on joint datasets of multiple parties without revealing anything about the individual datasets. MPC refers to different techniques that enable parties to collaborate on (sensitive) data with strengthened guarantees about the privacy of the data. Examples of this are techniques such as homomorphic encryption and secret sharing. A broader category that includes MPC is the so-called Privacy-Enhancing Technologies (PETs). Readers familiar with PETs might also recognize other techniques such as differential privacy, zero-knowledge proofs, synthetic data and federated learning as PETs.

The main characteristic of an MPC protocol is that it relies on some form of encryption to hide the actual data, but does contain a structure that allows computations to be performed on the encrypted data. After this computation has been performed, the hidden result of the computation can be converted back to plain text, where only the outcome of the computation is revealed, nothing else. This allows parties to get insights on their distributed data, without needing to reveal their sensitive inputs.

One domain where adoption of MPC has been proposed in recent years is finance [3]. While the original catalysts for this increase were cryptocurrencies, nowadays also traditional financial institutions are investigating the potential of MPC to broaden their capabilities. A prominent field of use of MPC in the future might be to analyse financial risks with a group of banks [20]. We will use a specific example of this throughout this article.



(a) Local financial data

(b) Combined training of a model in a privacy-preserving way

(c) Result is a global model that each party can query for credit risk scores

(d) Parties get high- or low-risk classification from the model

Fig. 1: Collaborative Financial Data Analysis using MPC

Example: MPC for collaborative analysis of financial data Throughout this paper we will use an example MPC setting to explain our ideas. We consider a group of banks which collaboratively analyses financial data in order to get better insights into the behaviour of their customers.

An overview of the scenario can be found in Figure 1. Three banks collaboratively train a model using their local data. They do this by combining their respective datasets using MPC, ensuring that they *only* see the resulting credit risk score model and learn nothing about the transactions of customers of other parties. After training the model, it can be used to compute credit risk locally. We assume the output of this model is a classification of an individual into one of two categories: high risk or low risk. It is important to note that the model computes the credit risk for each customer, but does not dictate what a financial institution can or needs to do with this classification. In fact, this can often not even be discussed by the banks due to regulations such as competition law.

MPC models are designed to ensure that sensitive data, such as that held by banks, remains private and secure. However, besides considerations on privacy, fairness concerns are also remarked, like in [8]. In this setting for example, banks should also ensure that no groups are discriminated in obtaining a loan. Doing this in the right way, becomes more difficult when the data is distributed among multiple parties. In chapter 2 and 3 we will go into this, but first we will introduce the concept of fairness in general.

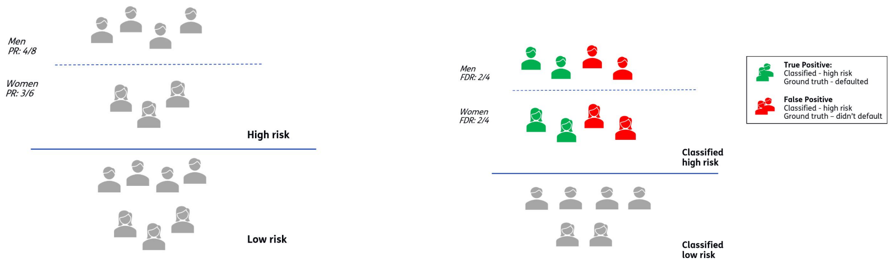
1.2 Introduction to fairness

Fairness is described by Mehrabi et al. as 'the absence of any prejudice or favouritism towards an individual or a group based on their inherent or acquired characteristics' [21]. As we are increasingly involving AI in decision-making processes, the risk of prejudice intertwining in automated systems is looming. Since these systems may be used in life-impacting scenarios, such as obtaining a loan in the example in Figure 1, it is important to investigate the fairness of their output and decisions that are made using it. For example, we don't want bank clients from a particular gender to be misjudged as a high-risk case more often than another gender. Especially, if this leads to them being denied a loan more often, which can be seen as discrimination.

Unfairness can be caused by bias, appearing in the training data, model, or use of the model [21, 18]. Similar to humans only being able to cancel out some heuristics about the world around us, 'bias-free' AI does not exist either. For an AI to learn a pattern, it will pick up on tendencies in the data that can ultimately be seen as some sort of bias. The ultimate goal can therefore only be to assess and mitigate the unwanted biases (i.e., those which entail discrimination or other undesirable effects). However, there is not one way to measure if a model is biased towards a specific group or based on a discriminating characteristic, so the process starts by choosing the right approach.

The focus of this paper will be *group fairness*, where you want to treat different groups similarly [21]. To start a group fairness assessment, one has to define the subgroups and the characteristics that define their differences, called sensitive attributes [24]. Next, it is important to make a distinction between the *output* and *outcome* of a system in which a model is used. The *output* is the result of the model computation, the *outcome* is what is done with that result by the party applying it. In our example, the output of the classification model is telling the bank which clients fall in the high and low category of defaulting on their loan. The outcome here is the bank choosing not to give out loans to clients in the high risk group. An alternative outcome could be that the bank decides to give all clients with a low risk classification an extra amount of credit. Note that it is possible to get to both outcomes from the same output, but it would be considered bad practice if all banks in the MPC setting had such a different intention on how to use the model. The chosen way of using the output, the outcome, is an important factor in defining how to measure fairness, because this defines the way the clients are treated. The measurement itself will be done with fairness metrics over the output of the model.

Fairness metrics can provide insight in the discrepancy of the output for different groups. Statistical (or demographic) parity on gender, for example, would mean in this case that an equal rate of men and women are predicted to be high risk, as shown in Figure 2a.



(a) Demographic parity **(b)** False discovery rate parity

Fig. 2: Two possible metrics for fairness. The model classifies each individual high or low risk, which are put in one figure for fairness assessment. Demographic parity is met in a), where men and women are classified as high risk in an equal ratio. False discovery rate parity is met in b), where the relative number of false positives is equal for both subgroups.

Nevertheless, statistical parity does not always satisfy equal treatment for all groups. There might be an equal number of people classified as high risk defaulter for both genders, but if the number of false positives (incorrectly classified as defaulters) for men is much higher, one can debate how fair the system is. A vast number of alternative fairness metrics are proposed in research, such as 'false positive parity' (false discovery rate) to fit this case [18, 25]. This paper will not discuss them all, but overviews can be found in e.g. [24, 21, 18, 25]. The example shows that metrics might be contradicting each other, and are even at a trade-off[5]. Saleiro et al. and Ruf and Detyniecki provide practical ways to navigate the different options, in the form of a Fairness tree (Aequitas) and Fairness Compass respectively [25, 17, 24]. Both tools help selecting which metric fits which model outcome, with questions such as whether or not the 'intervention is punitive or assistive', whether you want it to fit a certain representation policy and whom you think is most vulnerable in the situation. This illustrates that the intention for deployment of the model is determinant for the right fairness metric. It highlights the importance for the outcome to be clearly defined in order to choose a suitable fairness metric.

An important final note, however, is that fairness is not just a metric and that it is not something static [7, 9]. A fairness assessment only holds for the time, data, model, situation and usage as defined. For example, the predictor-prediction relationship of the model can change, or the data distribution can change when retraining with newly acquired data. Moreover, using a model with a different goal than what it was designed for - such as a group risk assessment used on an individual - increases the risk of flawed predictions. These examples illustrate fair AI also includes an investigation and documentation of the goals, (ground truth) data, usage (environment), time, policy, regulations, etc. It is essential that continuous monitoring and evaluation of fairness takes place, throughout the whole AI life cycle.

On top of that, one should realise that bias can't be completely mitigated, so fairness is also a matter of decision-making on what is important. This starts with agreements on the goals and usage of a system and requires well-informed, responsible and justifiable decisions. These agreements are necessary to measure suitable bias risk, as well as to prevent unfairness by misuse of the model. Nonetheless, in an MPC setting, it can be difficult to exactly define and control the important factors in such agreements with other parties. This results in the challenges identified in this paper, which concern both the metric definition, as well as the bigger context of fairness.

2 Fairness and Privacy-Enhancing Technologies

For an MPC model, the data is by definition distributed among multiple parties. This can make it challenging to measure and mitigate fairness, because not all data is available to everyone. Especially the goal of measuring fairness in an MPC setting has not been discussed a lot in literature yet.[1] In this chapter, we will first discuss existing strategies for achieving fairness when using Privacy-Enhancing Technologies in general, specifically in Federated Learning (FL).

FL is a way of training machine learning models on decentralized data, by exchanging only parameters of the model. While FL and MPC are related in the sense that they both allow parties to collaboratively train a machine learning model, there are some differences that might render existing fairness strategies incompatible. The main difference is that in FL only parameters are exchanged, while when using MPC, all data is exchanged in encrypted form. FL typically accepts leakage of statistical information through model weights, which can be extended to sharing statistics to remove bias from the data or the model. With MPC, this is more difficult as nothing is allowed to leak during the computation.

2.1 Existing technical strategies (mainly for federated learning)

There are numerous strategies to achieve fairness in machine learning models that also apply to models trained across multiple parties. Typically, these can be categorized in three strategies: pre-processing, in-processing and post-processing.

Pre-processing Pre-processing usually comes down to rebalancing the training dataset in order to remove bias across subgroups such that the model will not train on those biases. In 2020, Abay et al. [1] proposed two methods to apply reweighing in a federated learning context in two ways:

i Each party can locally apply reweighing. This is efficient and fully preserves the privacy of the parties, but lacks a global view of the weights that should be assigned to get a globally fair dataset.

[1] Note that there is also a different definition of fairness in MPC protocols that is not the topic of this paper. Namely, fairness in MPC can refer to a security model in MPC protocols where either all the parties receive the outcome, or none receive the outcome.

ii If the parties are willing to share sensitive attributes and sampling counts with noise, they can use differential privacy to communicate the local weights with each other, at the cost of some more communication and information leakage for global reweighing.

While these are conceptually simple and effective ways to mitigate bias for virtually any model and PET, they only apply to the training data. Therefore, potential fairness issues during the inference phase when the model is actually used cannot be prevented by pre-processing alone. Furthermore, collaboratively reweighing could prove difficult in an MPC protocol where nothing should leak during the computation. With federated learning, leaking (some) statistical information is more common since the weights of the local models are allowed to be shared as well. Pessach et al. [23] recently proposed the first solution for a privacy-preserving pre-process mechanism in an MPC setting, where distances between the distributions of attributes of two groups are decreased on federated data.

In-processing Instead of altering the training data, in-processing techniques aim to make the model fair during the training phase. In this strategy, a certain fairness objective is added to the training process that can be optimized for. One prominent example in regular centralized machine learning is prejudice remover. Intuitively, this adds a fairness metric to the loss function of a training procedure such that the loss is altered in a way that it punishes models that are overfitted and biased towards a certain sensitive feature. Again, Abay et al. [1] proposed a way to perform prejudice removal in federated learning. In the straightforward way, each party simply uses the prejudice remover during their local training step, after which the aggregation step remains untouched. Similar strategies have for example been proposed by [26, 27, 10, 13]. This is also known as local debiasing. While local debiasing is easy to apply in a federated learning setting, it can be hard to tune the parameters without leaking sensitive information, such that the model remains accurate enough while mitigating unwanted bias. It is expected that similar trade-offs will be observed with other in-processing methods. Furthermore, it is not yet clear how similar strategies can be applied to other MPC. Therefore, this approach was extended by Ezzeldin et al. in 2023 [12]. Conceptually, they additionally let the parties assess the fairness of the global model towards their local datasets and update their aggregation weight accordingly. Intuitively, parties which are more in line with global fairness will have a higher weight during the next aggregation round.

Post-processing Perhaps the most prominent example of post-processing is the landmark paper on equality of opportunity by Hardt, Price and Srebo [16] where a model is first trained using a regular training process, after which the model is adjusted to be fair by analysing ROC curves. With this strategy, the training phase remains untouched and thus is it likely to be supported in more PET settings compared to the in-processing techniques. However, access to the predicted label, sensitive attribute and target label is assumed, which might not always be the case in a PET setting. A post-processing solution specifically tailored to federated learning was proposed in 2021 by Luo et al. [19]. They let each party share statistics about their dataset to the central server, who can compute the global distributions. After that, virtual data points are sampled from these distributions and used to adjust the model. This will be difficult to achieve with MPC, as parties would need to sample from secret weight distributions.

3 Challenges in mitigating fairness issues in an MPC setting

In this paper, we reflect on three challenges that could occur when wanting to assess fairness while computing securely with multiple parties. We consider the example of three banks doing a multi-party computation, as described in section 1.1. We look at the fairness between two gender groups; men and women.

The three challenges could be considered in a chronological order. Firstly, multiple parties need to agree upon a common goal, or at least should be aware that different goals require different fairness metrics. When the goal is determined, a difference can occur between local and global fairness on that metric, the second challenge. Finally, to maintain fairness, the multiple parties need to be aware of individual feedback loops in the system, which can pollute the model and its data for others.

3.1 Challenge 1: a common goal

Consider the example of the three banks and the introduction to fairness (section 1.1 and 1.2): financial institutions are collaboratively training a credit risk model, classifying clients into high and low risk. The banks should not be discriminating based on characteristics such as gender, so a fairness assessment is relevant. To start this assessment with a fairness metric, it is necessary to know the intentions of the banks regarding the *outcome* of the model. Even though a common agreement on this exact outcome would be desirable, this might not be documented and complied to exactly. The discussion more often concerns the output of the model, plus it might be hard to reach perfectly similar deployment. This means that a situation is plausible in which two banks use the same output for a different outcome. For example, if one bank would choose to deny the high risk clients a loan (punitive), whereas another bank gives the low risk clients a higher loan (assistive). These outcomes require different fairness metrics. Considering that this is an example of bad practice for the usage of a model, it is still not the only problem. Even in the same usage setting, each bank might have their own policy or world view they want to adhere to, also resulting in a different suitable fairness metric.

Now, one could argue that each party ought to measure their own fairness, according to their own goal for the model output. A first question this raises, is whether or not the metric is still generally discussed, or left for each bank to choose for themselves. This is especially relevant in the scenario that ‘unfairness’ is found for one of these parties. What does that mean for the others? Is the issue the local data distribution, the model, the metric? In order to investigate the root of the problem and its consequences for others, the whole setting and assessment should be disclosed. Then the next problem is figuring out possible technical adjustments to the model to solve the fairness issue. Modifying the model might be the solution for one party, yet creates new problems for others. How to return to the drawing board with all the necessary information, representing each parties best interest?

Assessing a fairness metric at once for all the data, is technically possible as a secure computation. However, one is unable to answer crucial questions to choose a fairness metric, as illustrated earlier. Take the example with different outcomes for the same MPC output. For the assistive outcome, you will want to check the group rates for correct low-risk classification (true positives). Men and women should be relatively equally given the extra credit. For the punitive outcome, it is more important to look at the false positives in the high-risk class (false discovery rate from Figure 2b). Here, nor men or women should be more often wrongly denied a loan than the other. The example shows how different fairness metrics would apply. Again, also besides this undesirable setting, these disagreements on the definition of fairness can exist. For instance, if one of the banks has the ambition to reach a 50/50 division of issuing loans in terms of gender. This would demand a new perspective on which fairness metric fits best. Summarizing, because of their frequent incompatibility, there is no possible 'general' fairness for all parties if these parties do not use the model with the same intention.

Even in a setting where the banks agree on one way to use the model, one outcome and one fairness metric, there are still some unanswered questions left. Firstly, a general outcome might not hold for each individual party, because fairness is relative to data distribution. The further equal setting would make it easier to discover that the distribution is at the root of this problem, but doesn't solve it. It is possibly even undesirable to use a general measurement for the outcome over a specific subgroup (covered in Challenge 2). Secondly, there is a matter of ownership: who is responsible for the fairness assessment? Can one or each bank be part of that, or should an external party be involved as overlooking eye? The first may be difficult to choose, the second is not ideal in a PET scenario where no one is supposed to oversee all data. Note that while this section is highlighted with an example from the financial sector, the challenge is much broader and will also occur in other settings such as the medical domain. For example, a similar challenge occurs in an MPC setting where a model is used to prioritize patients at general practitioners and one GP uses a positive advice of the model to prioritize patients while another GP uses a negative advice of the model to further delay seeing a patient.

Takeaway Documentation of usage and goal of a model are crucial to facilitate and monitor fairness. To make sure a model is used in the correct manner, it should have a clear manual. To measure fairness with a metric, the goal of the model should be clear, both in output and outcome. In a PET setting, these agreements do not only account for one party but affect all stakeholders involved in the collaborative computation. The issues for the banks illustrates how a collaboration in an MPC setting makes it hard to assess fairness for each party themselves, but also as a collective. An MPC setting requires more elaborate agreements. In order to perform a general fairness assessment, different parties should be aligned to the level of using the model in the exact same manner, for the exact same outcome and with the same intentions and policies. It is a challenge to get to such an alignment, including responsibilities, and one can wonder if it is desirable if data distributions differ per party. Measuring fairness individually would require full disclosure on goals, usage, but also data distribution in order to meaningfully investigate a case of unfairness. This conflicts with the objective of MPC. Further research should look for the right way to an agreement, considering all these factors.

3.2 Challenge 2: global vs. local fairness

As explained in section 1.2 and the first challenge, choosing a fairness metric is an essential first step. But when a fairness metric is chosen, the way fairness is measured on distributed data is not straightforward. One practical challenge that the multiple parties can run into is the difference between *global* and *local* fairness. These concepts have been described mainly in the context of federated learning, for example in [15] and [26]. In the context of MPC, local fairness means the fairness is measured over the datasets of one single party and global fairness is measured over the entire dataset. We will now illustrate that if one has local or global fairness, the other one is not automatically achieved.

Global but no local fairness Suppose the three banks, in the example in section 1.1, have agreed upon using the metric *false discovery rate (FDR)* parity for men and women. In Figure 2b, we have seen an example where 8 persons in total are classified as high risk. We see that that the FDR is equal for men and women (50%) over the entire dataset. Therefore, the model is globally fair. However, if we look at the population of the individual banks A and B in Figure 3, we get a different view. For bank A we see that the FDR for women is 0 and for men is 1. Therefore, there is no *local* fairness for this metric at bank A, and for bank B vice versa. What the consequences of this difference are in practice, depends on the context. But in general, one should be aware that a model can be fair on a global data set, while it is not on the local data set. Local fairness can be desirable to know that a bank treats different groups of their clients equally.

Intersectional fairness The example might seem extreme, but note that when data distributions among parties differ, the differences can actually occur because of different outcomes of the same model. If for example bank A and bank B have populations of different ages and age and gender are related in the models prediction, this can cause different predictions among those features. Therefore, this topic is connected to the topic of *intersectional fairness.* In the case that bank A only has younger people and bank B only has older people, the model is fair on gender and fair on age, but it is not fair on the intersection of those

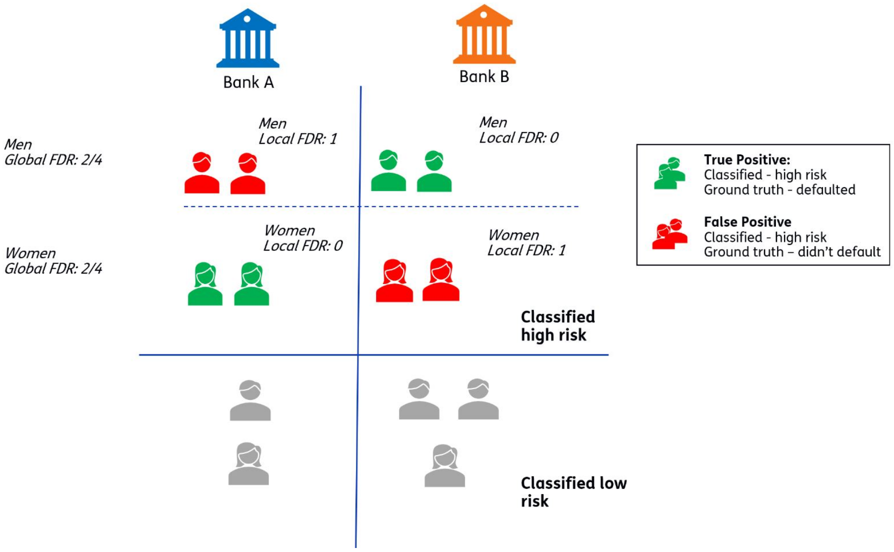


Fig. 3: Example of global but no local fairness.

two: young women (FDR = 0) are treated differently than young men (FDR=1). Therefore, the mitigation of this challenge should also be sought in this field of research, for example in the survey by Gohar [14].

Local but no global fairness While global fairness does not guarantee local fairness, also local fairness does not guarantee global fairness. Consider the example in Figure 4. Suppose bank A measures the fairness on their dataset, it concludes that the FDR is 0 for both men and women. Bank B will see that the FDR is 0.5 at both subgroups and therefore the fairness metric is met. However, when we look at the full dataset, we see that for men the FPR is 1/3 and for women it is 1/6. Therefore, the model is not globally fair. Again, it depends on the context what is desirable here. But in general, if party A and B only check for fairness on their own dataset, they might miss the fact that the model is not globally fair. And the fact that men and women are treated differently in general (say, measured over all banks in the Netherlands) seems undesirable. Note that in this example, the difference is caused by an unequal distribution of men and women amongst the banks, so again different data distribution plays a role.

Takeaway What a fairness metric means in a multi-party setting is not always clear due to the difference between local and global fairness. The two concepts can even be contradictory. The literature on local and global fairness in federated learning, i.e. [15, 26], offers some technical insights in how these concepts relate and how to reach one of the two or both. However, there seems to be no clear vision yet on which of the two is desirable in what context in practice. When designing a PET model with multiple parties, one should consider the way of measuring fairness and start a (partly non-technical) discussion on what global without local (or vice versa) fairness means. Also, one should be aware that a

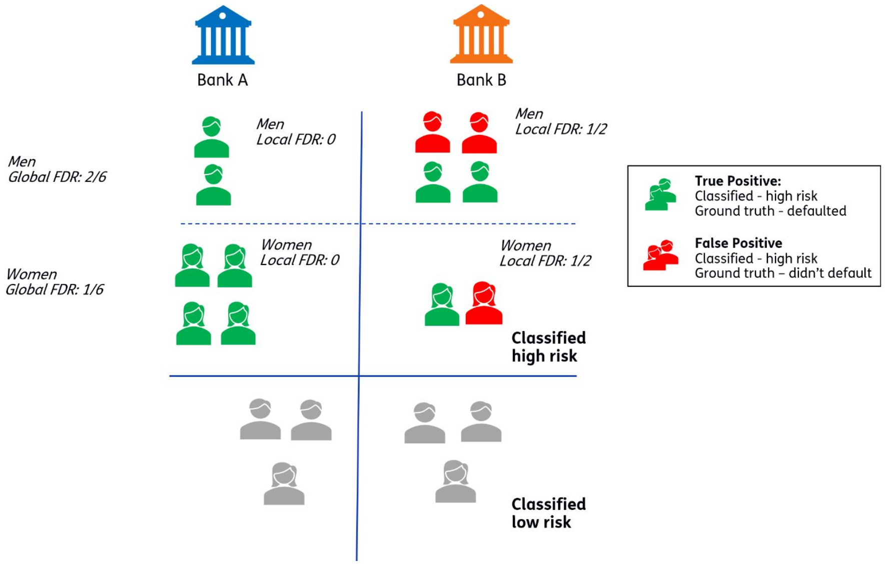


Fig. 4: Example of local but no global fairness

non-random distribution of datasets among the parties can cause differences in local versus global fairness.

3.3 Challenge 3: feedback loops

As was noted before, using a model with a different objective than what it was designed for, increases the risk of flawed predictions and possible unfairness. A model that was designed for classification on groups should not be used for an individual case, and vice versa. When retraining the model later, the wrong classification from this misused model will end-up in the new dataset that is used. This is an example of how a (polluting) feedback loop can occur, in which the input of a system changes over time, changing the system itself. These loops do not only occur through misuse of a model, but also naturally due to time, environment and usage. It can affect the different elements of an AI pipeline, from features to user. These feedback loops can cause bias to sneak into the data, and a static fairness measurement will not hold. Other examples, as described in Pagan et al. [22] and applied to our example, are:

1. a sampling feedback loop, where the decision whom to issue a loan might cause one of the gender groups to not apply for a loan any more;
2. an individual feedback loop, if the requester of the loan decides to spend less money because it knows that it is denied a loan multiple times (assuming 'money spent' is a predictor);
3. a feature feedback loop, where a predictor of repaying a loan is the risk classification a person has received before. In other words, if being classified

as high risk defaulter, increases the chances of being classified as such in a later iteration;
4. a machine learning model feedback loop, where only information about those people who actually received a loan will be available to learn from;
5. an outcome feedback loop, if a bank would use the model to apply a higher interest rate for those in the high risk category, this might increase their chances of defaulting.

This is not an exhaustive list, but exhibits forms in which the input and output of a model can be affected after deployment. As said before, fairness is not static, and thus with these feedback loops a momentary fairness evaluation will not hold over time [9]. A multi-party setting makes it extra difficult to track all these loops and their effects for each of the parties. Some feedback loops perform on a more general level such as the sampling feedback loops. This loop possibly holds for all parties simultaneously, making it a bit more insightful. Nonetheless, it could differ per party, when context specific factors influence the sample locally. This makes it difficult to notice the effects for the other parties in the computation. Other feedback loops occur when retraining a model, such as the ML model and outcome feedback loops. These loops are especially threatening in an MPC setting, where the data of one party affects the total system for each party. If one of the parties causes one of these loops to occur, it affects the data of the whole group. A distribution shift in the data would require a new assessment and fitting actions to maintain fairness.

Concluding, feedback loops are not uniquely occurring in the MPC setting. However, the MPC setting makes it more important, yet harder to monitor their effect on fairness, because multiple parties are involved. In all of these systems, feedback loops can appear. These will affect the total dataset in a new retraining iteration, possibly shifting the model that is used by everyone in the collaboration. When one is unaware of what is happening in other systems, it is also more difficult to detect and monitor possible influences of loops it entails.

Takeaway Different feedback loops can occur in the separate systems of each party, possibly affecting the total model and fairness for others. The good news about this challenge is that it is relatively easy to overcome. As in the earlier issues, it starts again with alignment on the goal, deployment and no-go's of the model outcome. Part of the issues can be mitigated when everyone uses the model only by the manual of its design, and decision-making factors (outcome) are not part of the models parameters (features). When the model is retrained with newly acquired data, this should be the same in all possible aspects and predictors should not be influenced by an individual party's usage.

Part of these loops might technically be inevitable. The most important thing to do, is to periodically monitor and evaluate the model for bias sneaking in. The mentioned agreement in this case is even more important, because it facilitates an investigation of where and how they may occur. In that way, one can also take appropriate countermeasures, such as taking an extra random sample when you know the retraining dataset will be skewed after the deployment of the model. Still, it also opens the discussion for questions such as whether or not the use of a ML model is desirable if they have these recurring effects. Finally, we can toss the debate on responsibility once more for such overarching monitoring and evaluation. Will each party investigate their own system? Which information needs to be shared with the others?

4 Discussion

In this paper, we have reflected upon the challenges around (measuring) algorithmic fairness when using MPC models. We have concluded that technical solutions in the context of PETs and fairness are mainly focused on federated learning, and not yet on MPC protocols. We identified three practical challenges that can occur in practice. The key takeaway is that there need to be agreements between the parties upon what fairness means for them. Furthermore:

- To perform one general fairness assessment for all parties involved in MPC, multiple criteria need to be met. The different parties should be aligned to the level of using the model in the exact same manner, for the exact same outcome and under the same intentions and policies. Individual assessment poses the unsolvable situation in which one party finds unfairness and is not able to change the model.
- What a fairness metric means in a multi-party setting depends on the data that is used. In a multi-party setting, this means that in case the federated data is different, different fairness metrics can be applicable.
- When multiple parties use a model in their own system, it is inevitable that feedback-loops will occur. To monitor, evaluate and act upon their effects on the fairness of a model effectively, an alignment and insight on the usage of each party is needed.

In future research, we hope that technical solutions can be found to some of the challenges around fairness assessment in an MPC setting. Still, there are both technical and non-technical open questions, such as:

- How can fairness mitigation techniques (pre-, in- and post-processing) in Federated Learning be applied in other MPC protocols?
- How can multiple parties agree upon the way they measure fairness and how they act upon it? The field of research in PETs and Data Spaces might provide guidelines for building agreements around MPC protocols [4].
- Which mitigation techniques in intersectional fairness can be applied to the issue of local vs. global fairness in MPC protocols?
- Which techniques in achieving local and/or global fairness in Federated Learning can be applied in an MPC protocol?
- How to operate on possible feedback loops of separate systems in the MPC, facilitating the evaluation and monitoring of their effects on fairness?

With this position paper, we hope to have paved the way for a promising new research field for the necessary integration of fairness techniques into MPC research.

Acknowledgments. The research activities that led to this paper were supported by TNO's Appl.AI programme.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abay, A., Zhou, Y., Baracaldo, N., Rajamoni, S., Chuba, E., Ludwig, H.: Mitigating bias in federated learning. arXiv preprint arXiv:2012.02447 (2020)
2. Ashurst, C., Weller, A.: Fairness without demographic data: A survey of approaches. In: Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization. pp. 1–12 (2023)
3. Baum, C., Chiang, J.H.y., David, B., Frederiksen, T.K.: Sok: Privacy-enhancing technologies in finance. Cryptology ePrint Archive (2023)
4. BDVA, DSC, C.: Leveraging the benefits of combining data spaces and privacy enhancing technologies (March 2024), https://bdva.eu/news/bdva-and-coe-dsc-joint-white-paper-on-combining-data-spaces-and-pets/
5. Braun, C.: Fairness in machine learning (Jan 2024), https://dida.do/blog/fairness-in-ml
6. Calvi, A., Malgieri, G., Kotzinos, D.: The unfair side of privacy enhancing technologies: addressing the trade-offs between pets and fairness. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. p. 2047–2059. FAccT '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3630106.3659024, https://doi.org/10.1145/3630106.3659024
7. D'Amour, A., Srinivasan, H., Atwood, J., Baljekar, P., Sculley, D., Halpern, Y.: Fairness is not static: deeper understanding of long term fairness via simulation studies. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. pp. 525–534 (2020)
8. Das, S., Stanton, R., Wallace, N.: Algorithmic fairness. Annual Review of Financial Economics **15**(1), 565–593 (2023)
9. Deldjoo, Y., Jannach, D., Bellogín, A., Difonzo, A., Zanzonelli, D.: Fairness in recommender systems: research landscape and future directions. User Modeling and User-Adapted Interaction **34**, 1–50 (04 2023). https://doi.org/10.1007/s11257-023-09364-z
10. Du, W., Xu, D., Wu, X., Tong, H.: Fairness-aware agnostic federated learning. In: Proceedings of the 2021 SIAM International Conference on Data Mining (SDM). pp. 181–189. SIAM (2021)
11. European Commission, Directorate-General for Communications Networks, C., Technology: Ethics guidelines for trustworthy ai (2019), https://data.europa.eu/doi/10.2759/346720
12. Ezzeldin, Y.H., Yan, S., He, C., Ferrara, E., Avestimehr, A.S.: Fairfed: Enabling group fairness in federated learning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37(6), pp. 7494–7502 (June 2023)
13. Gálvez, B.R., Granqvist, F., van Dalen, R., Seigel, M.: Enforcing fairness in private federated learning via the modified method of differential multipliers. In: NeurIPS 2021 Workshop Privacy in Machine Learning (2021)

14. Gohar, U., Cheng, L.: A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges. arXiv preprint arXiv:2305.06969 (2023)
15. Hamman, F., Dutta, S.: Demystifying local and global fairness trade-offs in federated learning using information theory. In: Federated Learning and Analytics in Practice: Algorithms, Systems, Applications, and Opportunities (2023)
16. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29** (2016)
17. Aequitas tool
18. Korteling, W., Drie, R.v., Veenman, C.: Fair ai: State-of-the-art overview of the literature (December 2022), https://publications.tno.nl/publication/34640564/UJuRDe/TNO-2023-R10060.pdf
19. Luo, M., Chen, F., Hu, D., Zhang, Y., Liang, J., Feng, J.: No fear of heterogeneity: Classifier calibration for federated learning with non-iid data. Advances in Neural Information Processing Systems **34**, 5972–5984 (2021)
20. Maxwell, N.: Innovation and discussion paper: Case studies of the use of privacy preserving analysis to tackle financial crime (January 2021), https://www.future-fis.com
21. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. **54**(6) (jul 2022). https://doi.org/10.1145/3457607, https://doi.org/10.1145/3457607
22. Pagan, N., Baumann, J., Elokda, E., De Pasquale, G., Bolognani, S., Hannák, A.: A classification of feedback loops and their relation to biases in automated decision-making systems. In: Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization. pp. 1–14 (2023)
23. Pessach, D., Tassa, T., Shmueli, E.: Fairness-driven private collaborative machine learning. ACM Transactions on Intelligent Systems and Technology **15**(2), 1–30 (2024)
24. Ruf, B., Detyniecki, M.: Towards the right kind of fairness in ai. arXiv preprint arXiv:2102.08453 (2021)
25. Saleiro, P., Kuester, B., Stevens, A., Anisfeld, A., Hinkson, L., London, J., Ghani, R.: Aequitas: A bias and fairness audit toolkit. CoRR **abs/1811.05577** (2018), http://arxiv.org/abs/1811.05577
26. Wang, G., Payani, A., Lee, M., Kompella, R.: Mitigating group bias in federated learning: Beyond local fairness. arXiv preprint arXiv:2305.09931 (2023)
27. Zhang, D.Y., Kou, Z., Wang, D.: Fairfl: A fair federated learning approach to reducing demographic bias in privacy-sensitive classification models. In: 2020 IEEE International Conference on Big Data (Big Data). pp. 1051–1060. IEEE (2020)