Universidade Federal de Minas Gerais

Department of Computer Science

Undergraduate Thesis, Part I

Relations between Causality, Fairness, Privacy, Accuracy, Information Flow and Explainability in Machine Learning

Type: Scientific

Abstract

Recent years have witnessed an enormous advance in the area of Machine Learning, reflected by the popularity of Artificial Inteligence systems. For most of the history of machine learning research, the main goal was the development of machine learning algorithms that led to more accurate models, but it is now very clear that there are many other important areas to develop. We want models to be fair to unprivileged groups in society, to not reveal private information used in the model training, to provide comprehensible explanations to humans in order to help identifying causal relationships, among many relevant goals other than simply improving model accuracy. This work reviews the literature for the identified relationships among these concepts in Machine Learning.

Supervisor:

Prof. Mário Sérgio Alvim

Thesis written by:

Artur Gaspar da Silva

Academic Semester 2024/1

CONTENTS

I. Introduction

Recent research [27] [1] [23] [24] indicates numerous tensions and synergies between many concepts that surround the Machine Learning literature, including Fairness, Privacy, Accuracy and Explainability. For instance, there is an inherent tradeoff between Fairness and Accuracy such that, depending on the data distribution, it might be impossible to develop a model that achieves acceptable values for both fairness and accuracy if we consider some reasonable fairness metrics [61]. Also, there has been some work on introducing Causality concepts into the discussion, for instance, to develop better fairness metrics [54]. It has also been suggested to use naturally interpretable models (as explanations for more complex models) for auditing systems and checking if they are fair, although this might lead to problems [88]. This area of research is especially relevant nowadays, given the importance that Machine Learning and Artificial Intelligence systems have: we now have computational systems that are part of processes of making decisions with big impacts on people's lives, for instance, recidivism prediction [8], loan approvals [69], hiring decisions [43], and others.

In this first part of the Undergraduate Thesis (POC I), we provide a concise review of the literature on the topics presented. The main goal is to provide a solid basis for future work on these topics and identify the known connections among them. Section II discusses how and when this work was developed; Section III discusses the expected results for POC I and POC II; Section IV discusses the already developed theoretical work on these areas; Section V discusses connections between them found in the literature; Section VI provides conclusions and possible future lines of work.

II. Methodology

The methodology applied to this project consists, in general, of reading as many papers on the subject as possible in order to gather what has been produced recently. Also, for developing the necessary theoretical background, the classical book Causality [59], by Judea Pearl, was of utmost importance.

The first two weeks (03/17/2024-03/31/2024) consisted of contacting the advisor, preparing the themes and writing the intial proposal; the next two months (04/01/2024-06/27/2024) consisted of reading the Causality book by Judea Pearl, reading papers on the relevant topics, and preparing the partial pitch; the last month (06/28/2024-07/31/2024) was dedicated to writing the final report and preparing the final pitch, as well as reading more papers on the relevant topics.

III. Expected Results

For the first part of the project (POC I), the expected result is an extensive review of the literature on Causality, Fairness, Privacy, Accuracy and Interpretability in Machine Learning, and the relationships between these concepts. For the second part (POC II), the expected results are the reproduction and verification of the viability of applying the theoretical framework of Quantitative Information Flow to these concepts, with the possibility of developing new theoretical results. This document shows the results of the first part, POC I.

IV. Theoretical Background for Individual Concepts

First, we provide a general background on some introductory concepts in the areas of Machine Learning, Causality, Fairness, Privacy, Explainability in Machine Learning, and also Quantitative Information Flow (QIF).

A. Machine Learning and scenario considered

Machine Learning is the field of study that focuses on developing methods of learning general patterns from limited data. In recent years, important advances have been observed in Machine Learning research, and also the popularity and applications of some methods have increased significantly. One example is the improvement of Convolutional Neural Network architectures, and the popularity of Generative Models. Also, some recent research is focused on the theory behind such machine learning methods [17] [34], and statistical learning in general [75]. Part of the goal of this formalization is to provide more qualitative guarantees in regard not only to accuracy, but also fairness, privacy, interpretability, and other important qualities. We discuss these different goals in the next four subsections.

In general, we consider *supervised learning* problems: in this scenario, a *machine learning algorithm* is an algorithm that receives many data points, which we call *training data*, and outputs a *model*. This model is itself an algorithm that receives a data point with some information omitted, encoded in what we call the *target variable*, and outputs a guess of the omitted information, which we call the *model prediction*. The model is then evaluated with other data points, ideally not the same ones used for training the model. This is called supervised learning because the algorithm has access to the target variable during the training process, which is not the case for unsupervised learning.

Figure 1 shows how the other concepts are related to machine learning in this context: we usually can assume that the training data is generated by some causal process, which can be modeled by a causal model; possible privacy attacks include performing sensitive information inference on the training data and on the

model itself; we usually measure how accurate and fair a system is by analyzing its predictions for many data points; we can also obtain local and global explanations for complex models by this type of analysis.

B. Accuracy

Accuracy is the notion of how close some estimate is to the true value we are estimating. In the context of Machine Learning, it represents how close the predictions of a given model are to the real value of the variable the model aims to predict. For binary classification (the scenario in which the target value has only two possible values), accuracy is defined in machine learning as $\frac{TP+TN}{TP+TN+FP+FN}$, where:

1) TP is the number of True Positives: how many predictions were labeled as True and were really True.
2) TN is the number of True Positives: how many predictions were labeled as True but were actually False.
3) FP is the number of True Positives: how many predictions were labeled as False and were really False.
4) FN is the number of True Positives: how many predictions were labeled as False but were actually True.

So, the usual notion of accuracy is the proportion of the predictions from the model that were correct for the available data. This generalizes to multiclass classification problems (in which the target variable has a finite number of possible values) by considering the proportion of times that the model's prediction was the correct class. *Regression* problems are the ones in which the target variable has an infinite number of possible values but can be codified as a vector of numbers: for instance, the value of some building at two different times. We can see how accurate a regression model is in many ways. For instance, by computing the square difference between the prediction value and the real value, summed for all training data points.

Note that during the training phase, the Machine Learning algorithm usually has the goal of providing the model that provides the best possible accuracy in the training data, among the possible models supported. If there is enough freedom among the possible models outputted from such an algorithm, then it might be possible to obtain a model that has very good accuracy, but if we try to use this same model on data other than the training data, this same model performs very badly. For instance, if we try to fit the data presented in 2a with an arbitrary degree polynom, then it's possible to fit the data perfectly, as shown in 2b. In this situation, the model obtained by the second-degree polynomial presented in 2c would, in general, be more useful, even though its accuracy on the training data is not 100% as in 2b. This problem is called *overfitting*, and is usually avoided by limiting how powerful the model can be, in conjuction with verifying accuracy values with data points not present in training dataset, which we call the *testing data*. Also, note that to evaluate the model, the test data should also have the target value of each data point.

The classical goal of Machine Learning is to provide models with good *test accuracy*, the accuracy evaluated for the test data, and a Machine Learning algorithm that produces models such that good results on the training data reflet on good results on testing data are said to *generalize* well. The idea is that the model captures a more general abstract concept than the simple memorization of the training data, as illustrated by figure 2d.

However, note that accuracy does not always represent how accurately the model fulfills all the goals of the relevant stakeholders. This is because, with the growth of Machine Learning applications in real-life scenarios and the impact on society as a whole, there has been a growing focus on aspects other than simply maximizing the accuracy. This is similar to the way that we are sometimes worried not only about how safe a cryptographic algorithm is but also about how computationally efficient it is.

One classical binary classification example presented in figure 3a is the concern with *sensitivity* and *specificity*: how many of the data points with target value equal to 1 are predicted by the model to have target value 1 and how much of the poins with target value 0 are predicted to have value 1 instead, respectively. One example in which this distinction is very important is in medical diagnosis: imagine that the model is predicting if someone has a deadly disease and a simple and low-risk treatment, a false negative is worse for the patient than a false positive diagnosis of having the disease. Other examples of notions to be considered, presented in 3b, are *precision* and *recall*: how many of the data points with predicted value 1 really do have target value 1 and how many of the data points with target value 1 the model correctly predicted the target value, respectively.

We can then create machine learning models that optimize for something other than accuracy. For instance, a combination of sensitivity and specificity that better suits our needs. The function that we aim to optimize with our model is sometimes called the *objective function*, and we can encode many different goals in these functions, with one classical example is adding a *regularization term* that penalizes more complex models, which sometimes helps in avoiding overfit.

Usually, there are other important goals to keep in mind when developing a machine learning algorithm:

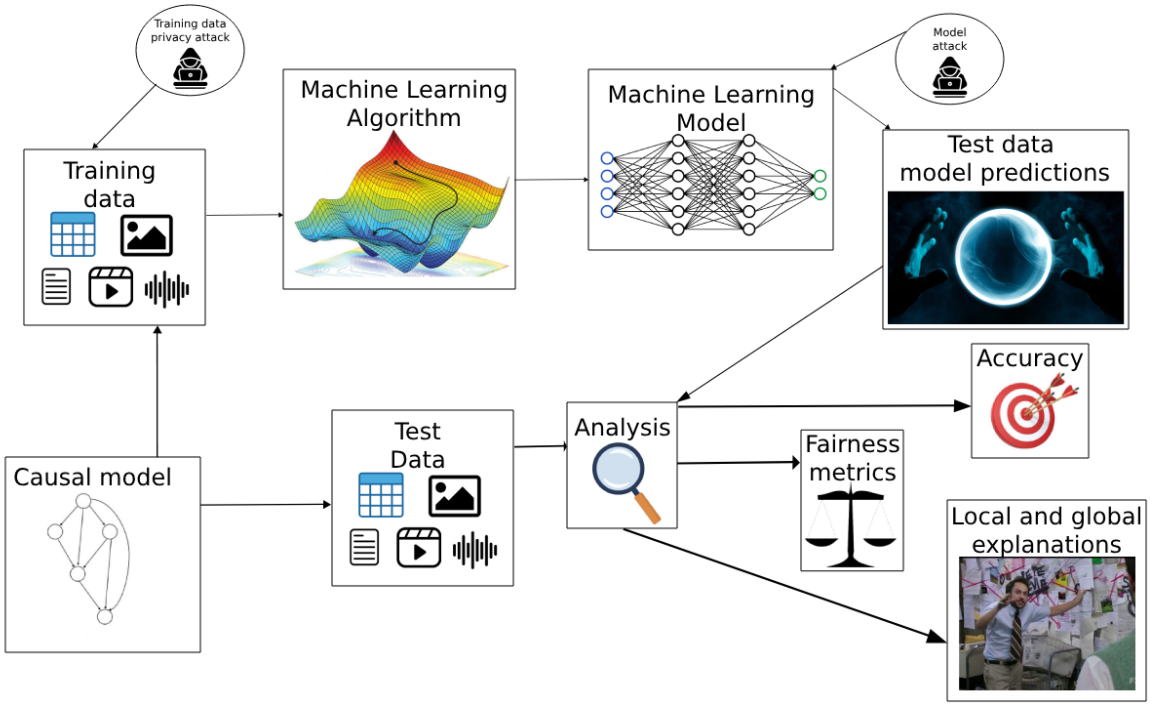


Fig. 1: Figure representing the supervised learning scenario we consider throughout this work.

the fairness goal aims to reduce unfair disparities in treatment of different individuals by machine learning models, the privacy goal aims to reduce the chance of someone discovering confidential or private data without authorization, and the explainability goal aims to improve the comprehension of how complex models work. Causality and Quantitative Information Flows are not goals by themselves, but methods for the representation and pursuit of other goals.

C. Fairness

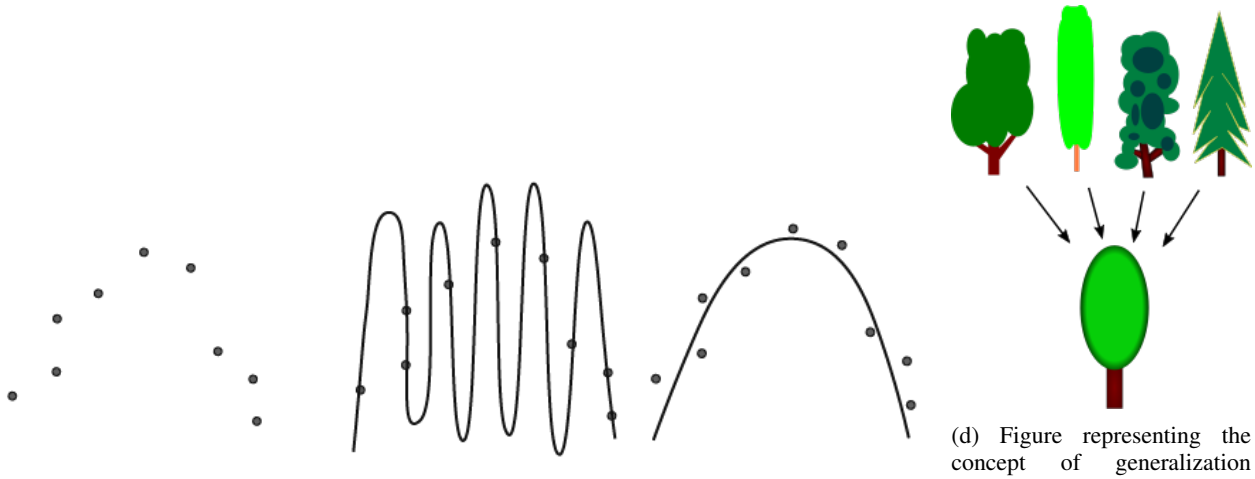
In the context of Machine Learning, fairness refers to the reduction, as much as possible, of *algorithmic bias*, the bias introduced by algorithmic decisions. This bias might have a big social impact because this can expand existing unfair discrimination in society, as machine learning algorithms are being used to make more and more important decisions. One famous example is the COMPAS recidivism algorithm, which has been used by the United States courts to estimate how likely someone is to reoffend in the future. It was revealed [8] that this tool was heavily biased against black people.

For binary classification, we will say that the result is *positive* for a data point if it benefits the person represented by that data point, and *negative* otherwise. We will say that the *unprivileged group* is the group of people affected negatively by the bias, and the *privileged group* is the other group of people.

Such biases can happen because of many factors. The algorithm itself might introduce bias, or the data might be biased. The data may have been collected in a biased way (in the COMPAS example, this would be the case if reincidivism data was collected more for black reincidivists than for white), or the data might be simply reinforcing some bias in society.

Also, the bias in society might be such that the data is in disagreement with reality (the unprivileged group's true values for the target variable would affect them in the same way as the privileged group), or it is in agreement with reality because of structural biases in society. For instance, if the prediction of the algorithm is whether someone will have good grades if accepted to some university, people in the unprivileged group might not have had as good oportunities in life as people in the privileged group, so the data is correct when it says that those people will have worse grades. Even so, the results might still be considered unfair, this depends on the notion of fairness we consider. All of these unfairness possibilities can be further divided into other types of unfairness, as was done in [56]. Image 4 illustrates where unfairness might come from, and we summarize below the ways in which unfairness might be introduced:

1) Algorithm results do not reflect the data.
   a) The algorithm might optimize for the majority only, achieving good overall accuracy even though it's mostly wrong for minorities. This can



(a) Figure representing data points.

(b) Figure representing overfitted data.

(c) Figure representing non overfitted data.

(d) Figure representing the concept of generalization using trees. Obtained from Wikipedia.

Fig. 2: Figures representing the goal of generalization and the problem of overfitting.

be considered a type of Aggregation Bias.

b) Systematic errors in the algorithm, that lead to biased estimation.

2) The data can be biased, not reflecting the reality.

a) We can have structural biases in society, such that people in unprivileged groups do not have the necessary opportunities, but if they were treated similarly to the privileged group by society, they would have similar results. For instance, an unprivileged group that doesn't have good educational opportunities will have worse scores on exams because of historical discrimination, and although just looking at whether someone is in this group could lead to a good accuracy, it might only perpetuate current unfair biases in society.

b) The bias can also be introduced in a way that people in the unprivileged group were misclassified before the data was collected. For instance, maybe capable people in an unprivileged group usually don't get a job even though they are actually as capable as the unprivileged group.

c) Data collection doesn't reflect the reality: Measurement bias (for instance, COMPASS used friend/family arrests as a proxy for a risk score present in the dataset), Omitted Variable Bias (this violates assumptions of some learning models, for instance linear regression models usually assumes error terms uncorrelated with the parameters considered in the regression), Representation/Sampling Bias (biased sampling lacking the diversity of the population), Simpson's Paradox (if we don't have data on a confounder, correlations might be spurious [59]).

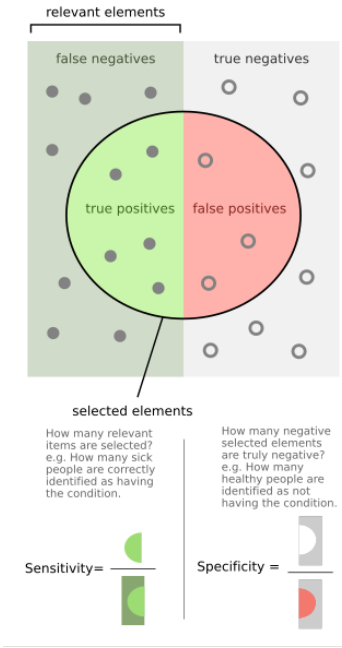
d) If the data is collected on a group fundamentally distinct from the one where it will be used, for instance another population (Population Bias) or the same population but at another time (Temporal Bias), unfair bias might be introduced.

e) Data that relies on people's opinion is prone to many biases: Social Bias (people do what others are doing), Self-Selection Bias (people think that everyone agrees with them), and many others.

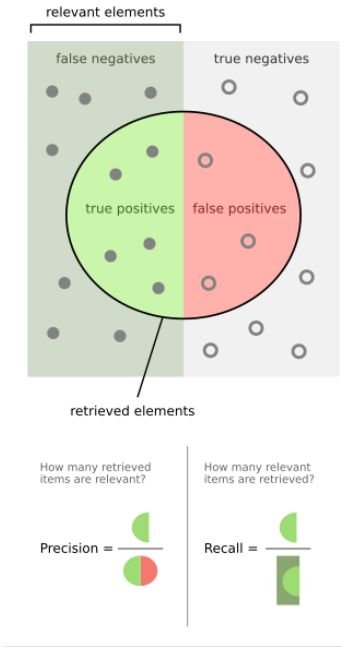
3) Data might depend on the algorithm's previous output: Presentation Bias (the user is presented to some selected advertisements, for instance), Ranking Bias (search engines ordering results in a biased way), Popularity Bias (more popular items are shown more). This might strengthen biases through time.

4) Finally, the circumstances can change through time, either by the influence of the algorithm itself or other factors, which can worsen the quality of algorithms previously considered able to provide good results (Emergent Bias).

Besides biased data and deliberate bias in the algorithm, such that the results of the algorithm do not reflect the data, it is also possible to introduce bias because the algorithm might prioritize making correct predictions for the majority of the population, if it can't make correct predictions for both the majority and the minority. Another possibility is that the prediction might depend on past decisions of the algorithm, and we only know the result if the result provided is positive (for instance, we only know if someone will reincide if we release them). In this type of scenario, according to Learning Theory, it's important to take suboptimal decisions to *explore* different options and gather more data [22] (found in arxiv.org), which might be considered



(a) Figure representing the concepts of sensitivity and specificity. Obtained from Wikipedia.



(b) Figure representing the concepts of precision and recall. Obtained from Wikipedia.

Fig. 3: Representation of some notions other than accuracy.

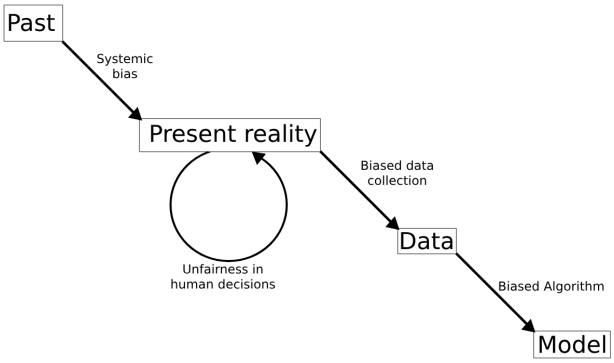


Fig. 4: Figure representing some of the main possible sources of unfairness.

unethical as it might have a big cost to society (releasing someone that's probably going to commit more crimes) or to the individual (not giving a life-saving drug to some patients as an experiment to see the survival rates for that specific group).

Many different notions of algorithmic fairness have been developed, and some are not compatible [3]. Initially, the notions of fairness could be grouped into two main types: statistical and individual definitions of fairness [22]. Statistical (group) notions of fairness require some statistical metrics to be similar for certain demographic groups, and individual notions enforce constraints on pairs of individuals, for instance requiring similar individuals to be treated similarly. Many problems with statistical notions and why they, in general, don't provide good individual guarantees are presented in [24] [39]. For instance, one such problem is satisfying the constraints for two protected attributes individually but not for combinations of these attributes. One problem with both individual and group notions is *composition*: it is not always the case that satisfying fairness constraints in individual, isolated, components of a system imply that fairness constraints will be satisfied for the whole system [25]. Finally, there are also causal approaches to fairness notions, which we will discuss more in Section V. In general, it is not possible to satisfy some of the main notions of fairness at the same time [35] [11] [85], and which fairness notion to use depends a lot on the specific goals of each different system. We will now define some of the main notions of fairness. We consider that Y is the binary target variable, with $Y = 1$ as the positive result and $Y = 0$ as the negative one; A is the binary sensitive attribute, with $A = 1$ for the privileged group and $A = 0$ for the unprivileged group; $\hat{Y}$ is the model prediction of the target variable value; X is a set of legitimate factors that can be used for classification.

Definition 1 (Equal Opportunity Difference)**.** *We define* Equal Opportunity Difference *as* $P(\hat{Y} = 1|A = 1, Y = 1) - P(\hat{Y} = 1|A = 0, Y = 1)$. *Equal Opportunity is*

satisfied *if the Equal Opportunity Difference is equal to zero.*

Definition 2 (Statistical Disparity)**.** *We define* Statistical Disparity*, also known as* Demographic Disparity*, as* $P(\hat{Y}=1|A=1)-P(\hat{Y}=1|A=0)$*. Statistical (Demographical)* Parity *is* satisfied *if the Statistical Disparity is equal to zero.*

Definition 3 (Conditional Statistical Disparity)**.** *We define* Conditional Statistical Disparity*, conditioned on* x*, as* $P(\hat{Y}=1|A=1,X=x)-P(\hat{Y}=1|A=0,X=x)$*. Conditional Statistical* Parity *is* satisfied *if the Conditional Statistical Disparity is equal to zero.*

A possible general definition of *individual fairness* notions is that an algorithm is considered fair if it gives similar outcomes to similar individuals, according to similarity notions relevant to the specific scenario considered.

The techniques developed to reduce unfairness in algorithmic decision-making can be divided into *pre-processing*, *in-processing* and *post-processing*. Pre-processing techniques modify the training data to remove biases present there. In-processing techniques modify the learning algorithm itself, for instance, by changing the objective learning function to include not only accuracy but also adding to it some statistical fairness metric, or including some constraint that it has to satisfy. Post-processing techniques act after the model is trained to reduce the unfairness in the decisions made by such a model.

In general, just removing the variables that would be considered unfair to use directly to classify an individual is not enough to guarantee fairness. As illustrated in image 5, the variables we would remove might be highly correlated to other variables, which could be used by the model to discriminate almost as if we hadn't removed any variable. Also, even if the machine learning model itself didn't use any sensitive variables or correlated attributes for the predictions, we still need to collect this sensitive data to be able to measure how unfair the model is.

D. Privacy

In the context of Machine Learning, a privacy-preserving algorithm is one that does not allow information considered private/sensitive to be obtained by unauthorized parties. According to the terminology presented in [46], this is called *private ML*. The private information to be protected can be the data used to train the model or the model parameters and structure itself. It is also possible to use Machine Learning to enhance privacy, *ML enhanced Privacy Protection*, or to serve as an attack tool, *ML-based Privacy Attack*. We will focus on private ML.

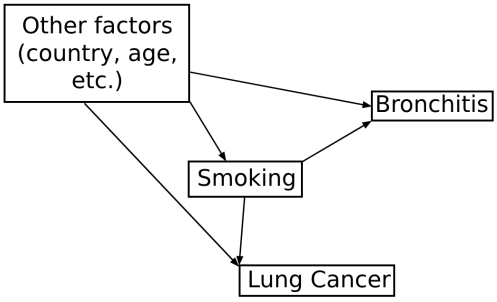


Fig. 5: Figure representing potential correlations between sensitive variables and other factors: if the disease status is a sensible information that could be used for unfair discrimination, then removing this information might not be enough to avoid unfair discrimination, as smoking, age and other factors combined might lead to unfair discrimination almost as if the model had direct access to the sensitive values.

We call *adversary* the agent that wants to discover the private information, and *secret* the private information itself. There are some possible goals of the adversary: she might wish to recover the model itself by trying to approximate the function that represents the model, to recover some feature or statistical property of the dataset, to discover whether some individual data point is present in the dataset, or even recover the exact values of individual samples in the dataset. We distinguish between the *White-Box access* and *Black-Box access* scenarios as the situation in which the adversary has or does not have full access to the trained model and their parameters, respectively.

Model Extraction Attacks assume an adversary with black-box access to the model, and no prior knowledge of the model parameters and training data. Some approaches to model extraction are presented in [72]. Although the most efficient attacks rely on confidence values, attacks that rely only on the output class labels are also presented. Other works focused on estimating hyperparameters [79] for an adversary that knows the training dataset and the Machine Learning algorithm. Notice that recovering the model can help in the development of attacks against the training data, even if the adversary doesn't have prior knowledge of the model.

Feature estimation attacks focus on recovering statistical properties or features of the training dataset, for an adversary that does not know the training data or the its distribution. Some attacks are presented in [29] both for black-box and white-box model access. As expected, the white-box attacks lead to better attacks, and they provide examples for recovering individual sensitive information from marital happiness answers in two different datasets

and white-box attacks for recovering images in a face recognition model. The results lead to the possibility of (almost) identifying someone with only their name and the face recognition model and to the possibility of identifying the answer to a sensitive topic on a supposedly anonymous questionnaire (whether the person answering ever cheated on their significant other) with very high precision and a recall bigger than 20%.

The attacks presented in [10] aims to recover statistical information about the training dataset by the use of many models trained on different datasets and a meta-classifier that identifies if a given model was trained in a dataset with some desired statistical property. This is a white-box scenario, and they focused on attacking Support Vector Machines and Hidden Markov Models. Another common type of attack is the Membership Inference Attack, reviewed in [38], in which the adversary aims to infer whether a given data point was used to train a given model or not. One approach is presented in [19], in which the adversary can sample from the original data distribution and has black-box access to the target model. It works by training many "shadow models" by using the data distribution, half with the target data point and half without it, then performing some computations based on confidence scores to estimate how likely the real model is to have used the relevant data point.

We will now focus on methods of protecting the training data from privacy attacks. The main methods we will mention are Differential Privacy, Local Differential Privacy, and Homomorphic Encryption.

Differential Privacy (DP) is one of the most important definitions of quantifying privacy: the idea of DP is to define how hard it should be to distinguish one dataset from another that differs by at most any one individual data point. More generally, we can define how hard it should be to distinguish an element from a neighboring element, such that in the database example, the elements are databases and the neighboring relation is such that two databases are neighbors if and only if they differ by at most one data point. A (possibly randomized) algorithm that aims to obtain a dataset that is protected according to the DP definition is called a *Differential Privacy Mechanism*. The formal definition is presented below.

Definition 4 (ϵ-Differential Privacy)**.** *A randomized algorithm from $\mathcal{X}$ to $\mathcal{Z}$ satisfies ϵ-Differential Privacy, for $\epsilon > 0$, if for every $x, x' \in \mathcal{X}$ such that x is a neighgboring element of x', and for every $S \subseteq \mathcal{Z}$:*

$$P(\mathcal{K}(x) \in S) \leq e^{\epsilon} P(\mathcal{K}(x') \in S)$$

The parameter ϵ can be interpreted as how close we require the probabilities of neighboring datasets to be, such that smaller values of ϵ lead to stronger requirements. In some practical applications, the Differential Privacy restriction is too strong. Thus, we have a relaxed definition for Differential Privacy, in which the parameter δ can be interpreted as the probability that the DP guarantee will not be satisfied.:

Definition 5 ((ϵ, δ)-Differential Privacy)**.** *A randomized algorithm from $\mathcal{X}$ to $\mathcal{Z}$ satisfies (ϵ, δ)-Differential Privacy, for $\epsilon > 0$ and $\delta \in [0, 1]$, if for every $x, x' \in \mathcal{X}$ such that x is a neighgboring element of x'.*

$$P(\mathcal{K}(x) \in S) \leq e^{\epsilon} P(\mathcal{K}(x') \in S) + \delta$$

Local Differential Privacy (LDP) is a concept important when the data collector is not assumed to be trusted and consists of the same restrictions as Differential Privacy, but with $\mathcal{X}$ as the set of possible individual values and the neighboring relation being such that all possible values are neighbors. This algorithm should be applied locally by the owner of each data point.

The idea is that individuals apply noise locally in a way such that the data collector can still obtain the desired results with the noisy data. The classical example is the scenario in which we want to discover how many people do some illegal activity (for instance, use illegal drugs) in a given region. The person answering might be tempted to lie to not let this sensitive information leak. But if we tell them to toss two coins, such that if the first one comes up heads, they answer truthfully, but if not, then the answer should be "yes" if the second coin came up heads and "no" otherwise. We will know that approximately $\frac{1}{2}$ of the answers is not the true answer of the person, such that half of these are "yes" and half "no". We can then remove 25% of the total number of answers from the number of "yes" and another 25% from the total number of "no" to estimate the real distribution. [81] delves into many mechanisms, information metrics, and applications relevant to Local Differential Privacy, including machine learning on private data.

There are some common misconceptions about what exactly are the assumptions of Local Differential Privacy. Notice, for instance, that if the data points of two individuals are known to be highly correlated (for instance, genetic data for two siblings), then even if their data points after the application of the LDP mechanism do satisfy (δ, ϵ)-LDP, the tuple of these two data points may not satisfy (δ, ϵ)-LDP, which can improve the inferences that an adversary can make. Imagine the extreme case: if we know that all data points are equal and the LDP mechanism gives a higher probability of not changing the data point, then the adversary can discover the common value of all data points simply by looking at the most common value after the mechanism is applied. This can also be a problem for Differential Privacy, as shown in [47].

The possible dependencies among data points have led to some different definitions of what exactly are the

assumptions of LDP, for instance, that all data points are independent, or that the adversary knows all data points but one. [73] discusses how a causal interpretation can help in uncovering the meaning of each LDP assumption, which are or not equivalent, and also compares with potential causal notions of LDP. We discuss this more at V, subsection "Privacy $\times$ Causality".

Federated Learning is another method that can help improve the privacy of individuals. The idea is that each individual trains a Machine Learning model locally, and shares information with a centralized server to improve a global model. The privacy risks are reduced because no individual data point and no individual user updates to the model are stored in the server. But still, without extra preparations, it might be possible to attack individual data points, as explored in [80].

Finally, *Homomorphic Encryption* is a form of encryption that allows computations to be done without decrypting the data, just the result is decrypted. *Secure Multi-Party Computation* is also an option if there are multiple parties responsible for this computation. The major drawback of these methods is the significant additional computational cost. Homomorphic Encryption and Secure Multi-Party Computation have been proposed for frequency estimation [84], Deep Learning [37] [32], and others.

E. Explainability

Explainability, loosely defined, concerns the ability to assign meaning to why some model took one or more decisions, in a way that can be interpreted by humans. There is currently no consensus for a precise definition of the term, and many papers argue about distinctions between interpretability and explainability. For instance, [31] defines interpretability as the property of being able to describe the internal working of systems to humans and *completeness* as the property of being able to accurately describe the operation of the system, and an explainable system has both properties at an acceptable level. [65] draws this distinction differently: they define transparency as the higher-level explanation given by the designers of the system of their choices of architecture, algorithm, and hyperparameters, while interpretations are defined as answers to the question "what does the model bases its decision on?", and explanations as the combination of interpretations and contextual information from domain knowledge. In general, it is considered necessary that an explanation both explains the inner workings of a system accurately and is comprehensible enough for humans with the relevant domain knowledge. There is a trade-off between these two goals, as we want not only to reach a balance between simplicity and accuracy, but also want important biases in the model to be evident in the explanation, as mentioned in [31]. This is represented by Figure 6.

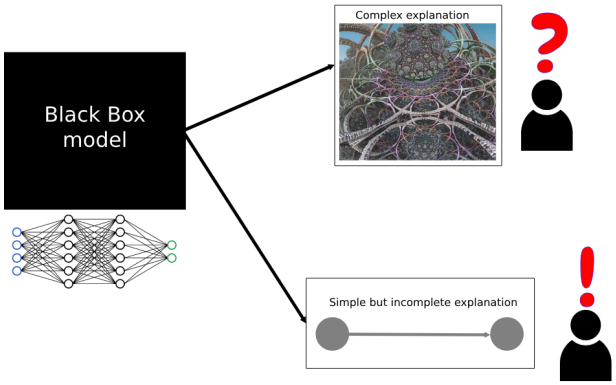


Fig. 6: Figure representing two important aspects to consider when developing explanations for black-box models: how clear the explanation is for humans and how representative it is of the real model behavior.

Being able to provide good explanations for the behavior of a system has many applications, ranging from helping developers debug existing models to aiding in solving legal issues regarding whether the system is treating some minority group unfairly. There are some laws being developed to require explanations to be given when necessary, following the idea of the "Right to Explanation", under which people would be able to require exaplanations for decisions influenced by automated systems that affect their life. The European Union "General Data Protection Regulation" and the French "Loi pour une République numérique" are examples of laws that partially include the desired "Right to Explanation". Although the movement for requiring explanations has gained strength in the last few years, there are also arguments against requiring a comprehensible explanation to always be provided, as this might lead to a smaller accuracy than if no explanation is required, which can be harmful, for instance, for medical decisions [48].

There are some classifications of the existing approaches to explainability, but we focus on the classification defined in [44]:

1) We can divide the approaches based on whether they are "local" or "global":
   a) *Local Explanations* focus on explaining individual decisions.
   b) *Global Explanations* focus on explaining the overall workings of the model.
2) We can also divide the approaches into "Model Agnostic" and "Model Specific":
   a) *Model-Agnostic Explanations* are methods that don't depend on the inner workings of the model.

   b) *Model-Specific Explanations* refer to methods developed to work only for a specific group of models, for instance, Grad-CAM and Shap-CAM are specific for Convolutional Neural Networks.
3) We can divide explainability approaches based on the data types these methods deal with, for instance tabular, text, image, or graph data.
4) Finally, we can divide approaches according to the purpose of the explanations, for instance:
   a) Creating alternative intrinsicly interpretable models.
   b) Explaining complex black-box models.
   c) Enhance how fair a model is.
   d) Test the sensitivity of predictions.
   e) Etc..

LIME [63] and SHAP [49] are examples of methods that provide local explanations and can be used to derive global explanations, both are Model-Agnostic methods that deal with image, text, or tabular data, and can serve many of the explanation purposes mentioned. Figure 7 shows an example of a model-specific explanation method for Neural Networks based on "expected gradients", from the "shap" python package. Notice that this method shows not only what influenced the final decision, but also what influenced the model to not "choose" each of the possible incorrect predictions. The model used is a Convolutional Neural Network with one output value per digit representing how much the model "believes" that the input image represents that digit.

As discussed at [13], explanations may be useful to data scientists, business owners, model risk analysts, regulators, and consumers, each with different goals in mind. The concerns that might be reduced by the use of explainable models include correctness (only variables relevant should be used in the final decisions and we should not use spurious correlations incorrectly), robustness (the model should not be susceptible to small perturbations), bias (the model should not be biased against specific subgroups), improvement (we might want to improve the model, and explanations can aid in this goal), transferability (the model should be useful in populations other than the one used to train and test the data) and human comprehensibility (this can aid an expert or even a non-expert in using and even trusting the results provided by the model).

The paper also mentions some criteria for evaluating explanations, which include how comprehensible the explanation is, how they accurately capture the models they aim to explain, how accurately they can be used to predict other outcomes of the model, how they scale to larger and more complex models, and how restrictive they are on the type of accepted model, some of these notions are further explored in [21]. They also evaluate explanations by example (for instance, counterfactuals [77] that provide examples with small changes to an input that can modify the output), and explanations by simplification (approximating a complex model by a simpler one). These approaches are different from SHAP, for instance, as it is based on game theory-based feature importance concepts. LIME can be considered to provide explanations by simplification, by locally approximating the model to a linear model. As mentioned in [65], explanations can also be used to enhance scientific research in natural sciences.

There are also inherently interpretable models, as their inner workings may be easy for humans to understand. Some examples mentioned in [13] are Linear Logistic Regression, Decision Trees, K-Nearest Neighbors (KNN), Rule-Based Learning, Generalized Additive Models (GAMs), and Bayesian Networks. Some researchers criticize this claim for some of these models, for instance, [45] points out that linear models are not necessarily easily interpretable, as they sometimes rely on un-interpretable and heavily-engineered features.

Other criticisms of the current development of explainability approaches were made: for instance, [41] argues that explainability is always a means to an end, and by requiring explanations to our models we may be significantly restricting the space of possibilities of dealing with the problems we need to face. For instance, there are approaches to verify if a model is fair that does not rely on explanations of the inner workings of the model (and in this case, relying on explanations might lead to other problems, as mentioned in [88]). Another possible problem with explainability is that if we rely on human opinions we might end up with methods that are persuasive, instead of accurate, as these two properties might not be fully aligned.

F. Causality

Most of our discussion of causality will be based on the work of Judea Pearl [59]. There are many alternative notions of causality, but most of them are already discussed in [59].

Judea Pearl divides causality into three levels (sometimes called *The Ladder of Causation*), illustrated by Figure 8, according to the type of question we want to answer:

1) *The first level of causation* deals with questions that can be answered by looking at the data. According to Pearl, this kind of question can be written in terms of what we see: for instance, given that we see that the floor of an entire street is wet then it probably rained just before. The notation used by Pearl is the usual probabilistic notation, $P(Y = y|X = x)$, abbreviated as $P(y|x)$: this means that

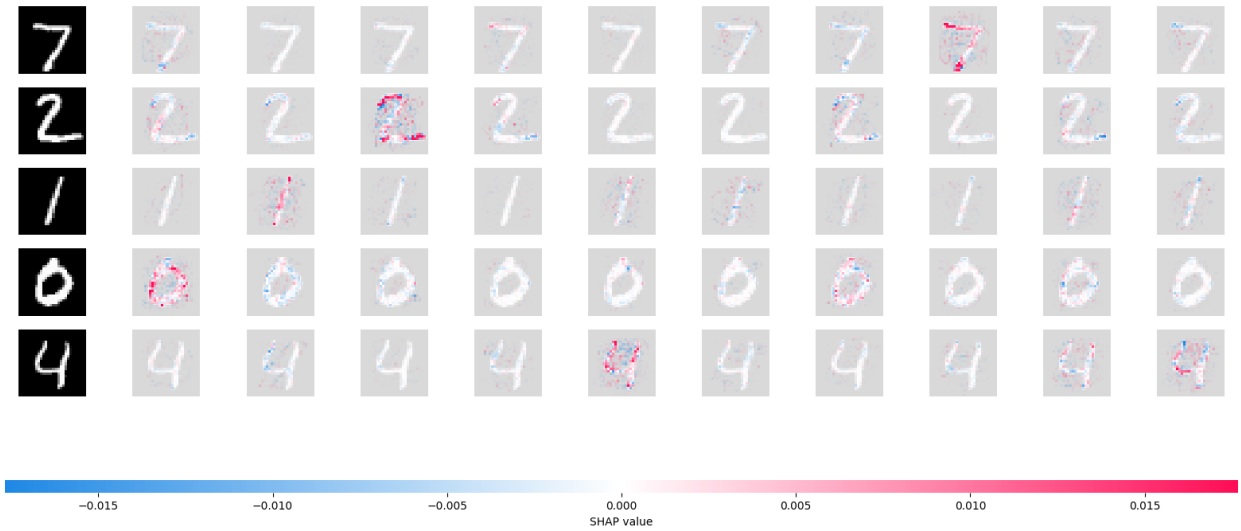


Fig. 7: Figure representing explanations based on "expected gradients" for MNIST examples, showing what pixels influenced positively (red) or negatively (blue) in the prediction. Each line represents a data point and each column a digit, such that the image in a given line and column shows which pixels influenced the output prediction of that data point representing the correspondent digit.

we observed $X = x$ and want to know how likely it is for Y to have the value y.

2) *The second level of causation* concerns questions of the type: "what's the consequence of some specific action". For instance: "if we make a street wet by throwing water manually into it, what is the probability that it rained?". The difference between seeing and doing is the fundamental distinction between Pearl's first and second levels of causation. Not only is the meaning completely different, but it's also impossible to answer questions at the second level of causation by only looking at the data: extra assumptions are necessary. In our example, if we look only at the data what we see is a strong correlation between raining and the floor getting wet, and in principle, we have no idea which one causes which. There are also scenarios in which two variables are strongly correlated but neither one causes the other. For instance, in some places, the number of ice cream sales is strongly correlated with the number of deaths by shark attacks, but clearly, that's because the times of the year when more people go to the beach are the same as when people buy more ice creams. The notation used by Pearl is $P(Y = y|do(X = x))$, which can be abbreviated as $P(y|do(x))$ or $P(y|\hat{x})$, which can be interpreted as how likely Y is to have value y when we set the value of X to x manually, and when we say "manually", we mean by an external intervention, that is not causally affected by the variables in the system.

3) *The third level of causation* regards questions of the type: "what would have happened had something been different?". For instance, "what would be the average temperature of Earth had the Industrial Revolution never happened?". Note that we observe something that happened, and then we think what would have happened if we changed something that already happened. This is what Pearl calls a counterfactual question: it requires imagining alternative worlds. To deal with this kind of question, Pearl proposes *Structural Causal Models* (SCMs), which consider deterministic relationships between variables and add all the uncertainty to the values of unobserved variables. The notation used by Pearl for the counterfactual notion is $P(Y_{x'} = y|X = x)$, which can be interpreted as how likely it is for Y to have value y in the world that we manually set X to x', given that we observed that the value of X is x.

Although the examples for the first level of causation usually involve some form of conditional probability, Pearl states that anything that can be computed from the joint distribution belongs to the first level of causation. This includes basically all of the usual machine learning approaches.

Pearl claims that any method of solving questions at the second level of causation must rely on assumptions beyond the data. One way of structuring many of the relevant assumptions is by using a directed acyclic graph in which each vertex is a variable and each edge represents the causal directions between two variables.

Many other types of assumptions can be made, Pearl discusses them in detail at [59].

Also, in the way that Pearl defines counterfactuals, there is a fundamental distinction between counterfactuals and actions with observations: in his notation, when we write the probability of some variable reaching some value given that there was an action and an observation, we interpret this as if the observation came after the action. This is different from the counterfactual notation, in which the observation comes before the action. This means that $P(Y_{x'} = y|X = x)$ is different from $P(Y = y|do(X = x'), X = x)$, as the latter should be interpreted as the probability that Y has a value y when we manually set X to x' and then observe that X has value x, which doesn't make sense if $x \neq x'$ (we shouldn't be able to condition on impossible scenarios). This difference of interpretation can lead to confusion and was mentioned in the second part of the "question to author" at [59, Subsection 11.7.2].

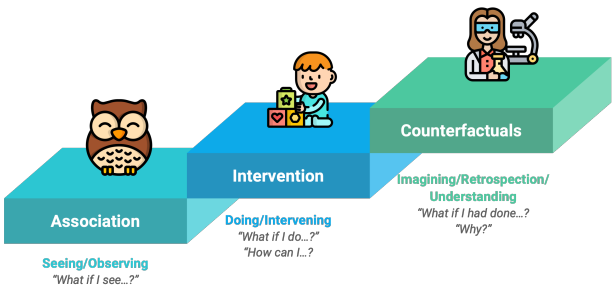


Fig. 8: Figure representing the three levels of causation.

In general, Judea Pearl's approach to causality assumes a Directed Acyclic Graph (DAG) for the relations between variables. This leads to what is called *Recursive Models* in [59]. Many results apply only to this type of model, but some generalizations are available and mentioned in [59]. Non-recursive models can be used for representing feedback loops: for instance, price and demand in economic models. This type of model has been widely used in economics in the form of Structural Equations, as presented in [59, Subsection 1.4.1]. Pearl also argues extensively about the possible causal interpretation of such equations and defines Structural Causal Models as models in which each variable has an equation that determines its value. This type of model can be used for answering questions at the third level of causation.

We will now discuss "Causal Discovery" and "Causal Inference", concepts illustrated by Figure 9.

Pearl presents many theoretical results about *Causal Discovery* on [59, Chapter 2], which revolves around the discovery of causal relations when we have access only to data. If two causal structures are capable of generating the same joint distributions, then we will be unable to distinguish between them if we observe only data. Also, sometimes a distribution is unstable for some models, in the sense that although the model can generate this distribution, it can only do so for some very specific configuration of the parameters. For instance, if we have A and B as the outcome of two independent fair coin throws (1 if heads and 0 otherwise, for instance), and C as the XOR between them. In the resulting joint distribution of the three variables, each pair of variables will be marginally independent but dependent if we condition on the third variable. This can be generated by three different causal structures, but only one is stable to small changes in the model (for instance, to small changes in the probability of each coin).

With the assumption that the observed distribution is stable in respect to the underlying causal model, and based on the principle of Occam's Razor, Pearl proceeds to define algorithms to recover as much information as is possible with only the data. In general, it is impossible to distinguish some relations: for instance, $A \rightarrow B$ (meaning that A causes B) is indistinguishable from $A \leftarrow B$ or $A \leftarrow U \rightarrow B$ for an unobservable U, as all three of them can generate exactly the same distributions on A, B in a stable way, depending on the model's parameters. Notice that $A \leftarrow U \rightarrow B$ can also generate distributions in which A and B are independent: we can, for instance, set U as the result of a fair four-sided dice with values $\{0, 1, 2, 3\}$, X as an indicator variable of the parity of the result (1 if it is odd and 0 if it is pair) and Y as an indicator of whether the result is bigger than 1 (0 if it is not, 1 if it is), in this case X will be independent of Y even though there is an unobservable confounder U. But this is not stable, in our example if we change slightly the probability distribution of U, for instance by increasing the probability of the outcome 0 and decreasing all others equally, then X and Y become dependent, $P(X = 0|Y = 0) \neq P(X = 0)$. Pearl provides an extensive analysis of stability and how causal relations are reflected as statistical dependencies between variables in the data.

Causal Inference regards the problem of inferring attributes of a causal model given its structure. In this situation, we have some causal model a priori and want to estimate quantities such as $P(Y = y|do(X = x))$. Pearl introduces the *do calculus*, which provides a way of deriving expressions for such quantities that do not contain *do* operator and can thus be estimated from data. The *do* calculus is based on three basic inference rules, which were shown to be necessary and sufficient, in the sense that a causal effect based on the *do* operator is identifiable by observing only the data and the assumption of the DAG structure of the underlying model if and only if we can derive an expression without the *do* operator using only these three rules. Many other

quantities can be defined with Pearl's framework, for instance, there are also definitions for identifying the results of dynamic plans, in which one action comes after others and might depend on the results of previous actions, which is further discussed in [59, Section 4.4].

In [59, Section 4.5], Pearl dives into two other causal concepts: *Direct and Indirect Effects*, which regard the effects that some variable X have on another variable Y that do or do not depend on other variables. For instance, smoking causes tar deposits in the lungs and also cancer, but we can think of how much of the effect of smoking on the probability of cancer is due to tar deposits, and how much is due to other factors. *Natural Direct Effects* are "average" estimations of direct effects for the different values the variables can assume, as the way that Pearl defines Direct and Indirect Effects is dependent on the exact values of the variables in question.

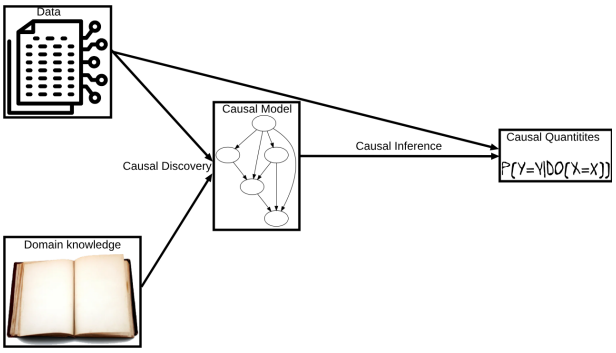


Fig. 9: Figure representing the difference between causal Discovery and Causal Inference.

Another contribution of the causality framework presented in [59] is the causal approach to *confounding*. In some situations, it is necessary to condition on some variables to account for possible spurious correlations that might arise due to confounding, but in other situations controlling for some variables would actually create new spurious correlations. This also depends on what we want to compute, but if the goal is computing the effect of some intervention, then the do-calculus can be used. Pearl argues in [59, Chapter 6] in favor of the causal approaches to confounding.

Counterfactual statements are defined in terms of simpler axioms, [59, Chapter 7] further discusses many results about counterfactuals. These results include how to use a Structural Causal Model to estimate the value of counterfactual statements, or how to check if this is even possible to estimate only with the SCM and data. [59, Chapter 8] delves into the problem of bounding values of expressions that can not be computed exactly. Pearl also discusses how some extra assumptions and tools can help in the estimation of expressions with the do operator and counterfactual expressions: for instance, some causal quantities are easier to estimate if we can experimentally change the value of some variables via external interventions.

Finally, there are some subtleties to the meaning of causation in different settings. Pearl discusses many notions of causation besides the ones we already discussed (such as direct and indirect effects), including:

1) We say that $X = 1$ was a *sufficient cause* of $Y = 1$ if when we are in a situation in which $X = 0$ and $Y = 0$ we expect that manipulating the value of X so it becames 1 is likely to change the value of Y to 1. In Pearl's notation $P(Y_{X=1} = 1|X = 0, Y = 0) \approx 1$, and $P(Y_{X=1} = 1|X = 0, Y = 0)$ is called the *Probability of Sufficiency*.
2) We say that $X = 1$ was a emphnecessary cause of $Y = 1$ if when we are in a situation in which $X = 1$ and $Y = 1$ we expect that manipulating the value of X so it becomse 0 is likely to change the value of Y to 0, in Pearl's notation $P(Y_{X=0} = 0|X = 1, Y = 1) \approx 1$, and $P(Y_{X=0} = 0|X = 1, Y = 1)$ is called the *Probability of Necessity*.
3) There is also the *Probability of Necessity and Sufficiency*, defined as the chance that manipulating the value of X so it becomes 1 will set Y to 1 and manipulating X so it becomes 0 will set Y to 0, written as $P(Y_{X=1} = 1, Y_{X=0} = 0)$.
4) Pearl also defines the *Probability of Disablement* as $P(Y_{X=0} = 0|Y = 1)$.
5) The *Probability of Enablement* is defined as $P(Y_{X=1} = 1|Y = 0)$.

Many relations between these values, bounds, and conditions for identification are presented in [59, Chapter 9].

The notion of "Actual Cause" is defined as an alternative to the sufficient and necessary notions of causation. Pearl mentions that necessary causation is closer to *token-level*, more individual than generic, as it conditions on events that really happened, while sufficient causation is closer to *type-level*, more generic than individual, as the events we condition on are less specific and related to an alternative imaginary scenario. The "Actual Cause" is intended to be token-level, to define what actually caused something. It is defined in terms of "Causal Beams" and "Sustenance".

We say that $X = x$ *causally sustains* $Y = y$ in $U = u$ (representing the uncertainty, the unobservable factors) relative to contingencies in W if and only if: $X = x$ and $Y = y$ under $U = u$; for any w we get $Y = y$ under $U = u$ and interventions that set $X = x$ and $W = w$; and we get $Y = y' \neq y$ under $U = u$ and interventions that set $X = x', W = w'$ for some $x' \neq x$ and some w'. In other words, $X = x$ causally sustains $Y = y$ in the circumstances $U = u$ relative to W when

$X = x$ and $Y = y$ in the scenario $U = u$, the value of Y never changes if we change the value of W and keep the value of X, and the value of Y can change if we change the values of X and W, keeping everything else as it is with $U = u$. This means that in the situation that actually happened $U = u$, then $X = x$ is enough to sustain $Y = y$ under interventions on W, but if we intervene to change X there will be a scenario in which intervening in W changes Y. If $W = \emptyset$, then $X = x$ causally sustains $Y = y$ in $U = u$ relative to W if in $U = u$ we have $X = x$ and $Y = y$, but it is possible to change Y by changing X.

A *Causal Beam* is a causal model defined in terms of circumstances $U = u$ and another causal model, such that the parents of each node in the new model are sufficient to maintain the value of the node regardless of the value of changes on the other parent's values, and that it is possible to change the value of other parents and the new parents to change the value of the node. If changing the value of only the new parents is enough to change the value of the node for every node, then the Causal Beam is considered a *Natural Beam*. Natural Beams represent the simplified version of the model that represents the actual scenario $U = u$, such that the parents of each node are enough to sustain the value of the node regardless of changes in other variables, and are also capable of changing the value of the node by themselves.

Pearl notes that in the definition of Causal Models provided, the parents of a node are defined in a way that makes the functions of the model non-trivial regarding all their arguments and all possible circumstances u, but when we consider a specific value of U we can simplify the model further. The example introduced by Pearl considers $f_i(x_1, x_2, u) = ax_1 + bux_2$ as the function that defines the value of variable V_i in the original model: in this scenario when $u = 0$ the value of X_2 becomes irrelevant, so we can simplify the model by defining $f_i(x_1) = ax_1$. The variable V_i will then have only X_1 as a parent, instead of both X_1 and X_2.

Finally, we say that $X = x$ is an *Actual Cause* of $Y = y$ in the state $U = u$ if and only if there is a natural beam under circumstances $U = u$ such that if intervene with $X = x$ then $Y = y$ and if we intervene with $X = x'$ for some $x' \neq x$ we get $Y \neq y$. This represents token-level causation, whether or not $X = x$ actually caused $Y = y$ in the real scenario $U = u$. As with many of the results presented, these definitions assume we have a full description of the causal model. The notion of Actual Cause is further discussed in [59, Chapter 10].

G. Quantitative Information Flow

The area of Quantitative Information Flow deals with methods of measuring information leakage from systems. This estimation is important to consider when developing real systems, as some information leakages are acceptable. For instance, whenever someone tries to authenticate with a username and password, but incorrectly guesses the password, some information leaks about the real value of the password: we now at least know that it's not the one that they tried. But we usually agree that this is acceptable while revealing the real password whenever someone makes an incorrect guess is unacceptable. How to adequately quantify the amount of information leaked from a system might depend on the goals of the people involved and on the information they have before the system executes. We thus need to first define some important notions, illustrated by Figure 10, before proceeding:

1) *Adversary* is an agent that tries to gain something with the information that leaks from the system.
2) *Secret* is the non-public data that the system processes.
3) *Prior Disribution* is the probability distribution on secrets that represents the knowledge of the adversary before the system runs, how likely the secret is each possible value according to the adversary's prior knowledge.
4) *Posterior Distribution* is the hyper-distribution, a probability distribution on probability distributions, on secrets. It represents the knowledge of the adversary after the system runs: ignoring some technicalities, we can consider this hyper-distribution to have one distribution per observable system output representing the adversary's knowledge of the secret for each possible observable system output.
5) A information-theoretical *Channel* is a representation of the system, which encodes the distributions of possible observable values output from the system, which might depend on the secret value.
6) The set $\mathcal{X}$ represents the set of possible values of the secrets.
7) The set $\mathcal{Y}$ represents the possible values of observable outputs of the system.
8) The set $\mathcal{W}$ represents the possible values of actions the adversary might take.

We consider the g-vulnerability framework, introduced in [6]. This framework introduces new definitions, which we present more formally:

Definition 6 (Gain Function)**.** *A* Gain Function *is a function $g : \mathcal{W} \times \mathcal{X} \to \mathbb{R}$, such that $g(w, x)$ defines the gain of the adversary if she takes the action w when the secret value is actually x.*

We do not have a loss for the system owner because we consider a zero-sum game: the gain of the adversary is exactly the loss of the people responsible for the system.

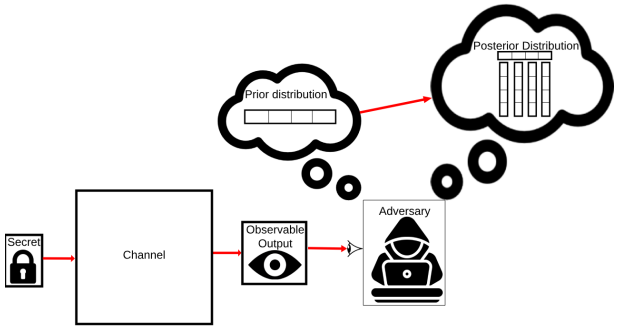


Fig. 10: Figure representing the scenario we consider in Quantitative Information Flow.

Definition 7 (Prior g-Vulnerability)**.** *The* Prior Vulnerability *of the system is defined as the average gain of the adversary if she takes the action that maximizes her expected gain, according to the distribution on secrets that represents her prior knowledge. Given a prior distribution on secrets π and a gain function g, this is defined as:*

$$V_g(\pi) = \max_{w \in \mathcal{W}} \sum_{x \in \mathcal{X}} \pi_x g(w, x)$$

Definition 8 (Posterior g-Vulnerability)**.** *The* Posterior Vulnerability *of the system is defined in the same way as the prior vulnerability, but considering the hyper-distribution that represents the posterior knowledge. Given a prior distribution on secrets π and a channel C and a gain function g, this is defined as:*

$$V_g[\pi \triangleright C] = \sum_{y \in \mathcal{Y}} \max_{w \in \mathcal{W}} \sum_{x \in \mathcal{X}} \pi_x C_{x,y} g(w, x)$$

The posterior g-vulnerability represents what will be the expected gain of the adversary after the system runs, but the estimative is made according to the knowledge the adversary has before the system runs

Definition 9 (Additive Leakage)**.** *The* Additive Leakage *can be defined as the difference between posterior and prior vulnerability:*

$$\mathcal{L}_g^+(\pi, C) = V_g[\pi \triangleright C] - V_g(\pi)$$

Definition 10 (Multiplicative Leakage)**.** *The* Multiplicative Leakage *is the result of dividing the posterior vulnerability by the prior vulnerability values:*

$$\mathcal{L}_g^\times(\pi, C) = \frac{V_g[\pi \triangleright C]}{V_g(\pi)}$$

There are many valuable theoretical results about channels regarding the relationships between prior and posterior g-vulnerabilities. We mention some of these results:

1) [6, Chapter 7] shows results about the *capacity* of a channel, which is the maximum possible (additive or multiplicative) leakage that can happen through a channel if we fix either the prior, the gain function or neither.
2) [6, Chapter 9] shows results about *refinement* of channels: in short, a channel is strictly better (for all priors and gain functions) than another channel in respect to the posterior vulnerability if and only if it can be written as a post-processing of this other channel.
3) [6, Chapter 10] presents the notion of *Dalenious vulnerability*: it might be the case that the adversary is interested in a secret other than the one considered in the system, and can obtain information about this other system via a known joint distribution between this other secret and the secret that the system considers. In this case, a channel is also strictly better than another in respect to Dalenious leakage, for any such joint distribution and gain function, if and only it can be written as a post-processing of this other channel.
4) [6, Chapter 11] discusses the axiomatic characterization of the notion of vulnerability, and even how some results can be obtained by different axioms that consider the worst-case scenario instead of the average gains of the adversary.

Even though most of the work on Quantitative Information Flow was developed with a stronger focus on measuring how system a system is, the g-vulnerability framework can be considered a general notion for quantifying information flow. This means that, in the future, we might be able to use it in the areas explored in this work, as some can be viewed in terms of information flows.

V. Contributions: A Review of the Relations Between the Concepts

In this section we mention comparisons found in the literature between these areas of research, and discuss some ideas for unexplored relations in the literature.

A. Accuracy × Fairness

Some results that indicate an inherent trade-off between fairness and accuracy in machine learning:

1) [61] shows that there are trade-offs between Equal Opportunity Difference and accuracy such that, depending on the data distribution, it might be impossible to achieve both perfect Equal Opportunity Difference and non-trivial accuracy. It also shows some other theoretical results that associate EOD and accuracy, such as sufficient conditions for the existence of non-trivially accurate predictors that

lead to zero Equal Opportunity Difference and non-trivial accuracy and algebraic and geometric properties of the feasible values of Equal Opportunity Difference and accuracy.

2) [1] provides methods for computing the best possible accuracy given some level of fairness, for a general notion of fairness that encompasses many common metrics. They devise an algorithm for solving a constrained linear optimization problem that minimizes the error subject to fairness constraints and provide experimental results for some datasets, including the COMPAS [8] dataset. One interesting result is that for some datasets (such as the Compas dataset), it is possible to reduce Equal Opportunity Difference without changing a lot the accuracy, but for some datasets (such as the Dutch Census Dataset [57] with gender as the protected attribute and the goal is if someone has a prestigious occupation).
3) [40] shows that it is always possible for an adversary to corrupt the data available for a learning algorithm in a way that the algorithm is unfair, and sometimes it is possible to do so without changing the accuracy.
4) [74] Provides a method of finding the full Pareto front of accuracy versus fairness. Their approach is based on a genetic algorithm, and the notion of fairness that they consider is the False Positive Rate for avoiding disparate mistreatment, but it is possible to use most metrics available in the literature.
5) [30] provides an empirical analysis of some of the existing fairness-enhancing methods for machine learning, showing that the results are influenced a lot by the fairness notion used and also by the dataset.

B. Accuracy × Privacy

If we consider obfuscation methods for privacy (such as Differential-Privacy and Local Differential Privacy mechanisms), in general, bigger privacy constraints imply a smaller accuracy when training Machine Learning Models. There are also homomorphic encryption and secure multi-party computation approaches, which have computational complexity as a major challenge instead of accuracy. In fact, some predictions can be made without any loss of accuracy at all, as shown in [84] for the naive Bayes classifier. Notice that in the Local Differential Privacy context, we know exactly how the noise is applied and thus might be able to reverse some of the effect of the noise when training the model, and in general, the effectiveness of this reversion determines how much the accuracy of the model will be affected.

1) [83] provides an overview of the use of Local Differential Privacy in general, but also further discusses machine learning in the local different privacy scenario, both for unsupervised and supervised learning.
2) [87] compares Local Differential Privacy and Federated Learning approaches empirically, concluding that LDP approaches consume less computational resources on the client side, benefit from a large user population, and can be reused for other tasks, while the data transferred to the server for the Federated Learning approach is specific for one inference task.
3) [62] proposes a method for applying Local Differential Privacy to high-dimensional data and evaluates it empirically, showing that in general stronger privacy also implies smaller accuracy.
4) [47] explores differential privacy in the setting in which the individual data points are correlated in a way that can be explored by the adversary, and investigates how to provide better privacy guarantees in this scenario. This might be important to consider in the machine learning scenario as well in the future, as the adversary might take advantage of such correlations in this type of situation as well.
5) [84] discusses how to use homomorphic encryption to preserve privacy without loss of accuracy. The classification algorithm used as an example is the naive Bayes classifier,

C. Accuracy × Explainability

If the Machine Learning model is inherently interpretable by humans, then usually it is less accurate than more complex models (the most common example are the deep convolutional neural networks). Also, explainability methods might help the model developer in identifying problems and improving the model quality, which might include improving the accuracy [16]. Also, for simpler models, it might be possible to keep good accuracy and explainability levels.

1) [60] discusses some arguments for and against this dilemma and points out some options of where the focus should be when developing and implementing Machine Learning systems.
2) [48] argues that in some situations it might be better to accept results without an explanation when these results can be empirically verified to lead to better results than well-explained results. Two important points raised by the paper are: “The opacity, independence from an explicit domain model, and lack of causal insight associated with some powerful machine learning approaches are not radically different from routine aspects of medical decision-making” and “In medicine, the ability to intervene effectively in the world by exploiting causal relationships often derives from experience

and precedes clinicians' ability to understand why interventions work."

3) [7] discusses the known trade-offs between accuracy and explainability for some distinct scenarios and models, the areas that can benefit the most from explainable ML, and a general review of explainability.
4) [12] and [2] provide an empirical study on whether explanations for AI predictions do or do not improve the accuracy of human decision-making, and the results indicate that this might not be the case.
5) [76] provides an empirical study to evaluate the opinion of the general public on whether it is worth trading explainability for accuracy in healthcare and non-healthcare scenarios. The conclusion is that in healthcare accuracy is favored, and in other scenarios, explainability is valued equally to or more than accuracy.

D. Accuracy × Causality

Accuracy is about how much a machine learning model makes correct guesses, Causality is about studying the causal relationships between variables. We can say that the problem of getting good accuracy lies on the first level of causation, as we are concerned about the data, although we could, for instance, use causality concepts to transfer what was learned with the data of one population to another. In some situations, algorithms based on causal notions can be even more accurate than purely associative algorithms.

1) [64] presents a counterfactual-based algorithm for medical diagnosis when there are many possible causes for a patient's symptoms. Their algorithm results in better accuracy than some classical Machine Learning algorithms not based on causal notions.
2) [68] discusses many relations between causal learning and machine learning, including references to works that discuss the idea of transferring results to different populations without losing too much accuracy.
3) [86] presents a method of selecting features with causal analysis before training a machine learning model based on causal concepts and shows that this approach can improve the accuracy of network intrusion detection when compared to classical machine learning methods.

E. Accuracy × QIF

Accuracy can be viewed as a form of utility, which represents how useful a system is, which might be affected by how vulnerable it is. It is trivial to develop a system without any vulnerability, for instance, a system that always outputs 0 no matter what is the input. But it's not useful, so we can consider that QIF is related to accuracy when we consider the former as the study of information leakage from a system and the latter as a form of measuring utility.

F. Fairness × Privacy

Some impossibility results in the literature argue why it is impossible to achieve both privacy and group fairness constraints under non-trivial accuracy [23]. We also have positive results that show how stasifying privacy constraints can help mitigate group unfairness [52] [51] [9], and we also have similar positive results results for individual fairness notions [24].

1) [52] aims to provide theoretical results that relate how training a machine learning model on a dataset in which local differential privacy mechanisms were applied can impact the fairness of this model if the data-gatherer does not try to reverse the applied noise. The results are provided for a simplified machine learning model, from a theoretical perspective.
2) [51] aims to evaluate experimentally how training a machine learning model on a multi-dimensional data set in which local differential privacy mechanisms were applied can impact the fairness of the model, again if the data-gatherer does not try to reverse the noise.
3) [9] executes an empirical evaluation of the impact of many Local Differential Privacy mechanisms on fairness when we do not try to reverse the applied noise, including a new privacy budget allocation scheme based on the domain size of sensitive attributes.
4) [24] introduces a notion of individual fairness that is a generalization of differential privacy and explores the relationship between this notion of fairness, differential privacy, and statistical disparity.
5) [23] presents theoretical results that show the impossibility of having a classifier that is not trivially accurate and at the same time satisfies $(\epsilon, 0)$-differential privacy and equality of opportunity constraints. The paper also shows a PAC learner for an approximate fairness definition they provide.
6) [36] compiles results that relate the concepts of privacy and fairness, and also causal discovery.
7) [28] discusses the use of Homomorphic Encryption in combination with fair representation learning [85] to provide both privacy and fairness guarantees, respectively, when training machine learning models. Then the paper discusses how to provide local and global explanations for the model.

G. Fairness × Explainability

Fairness evaluation is a goal usually mentioned when the utility of explainability studies is discussed in the

literature [66]. There are also some negative results in this respect, that indicate that in some situations, it is possible to provide convincingly fair explanations for decisions that were actually made in an unfair way [88].

1) [66] discusses how not providing explanations for decisions can threaten accountability, bias avoidance, and transparency when evaluating and mitigating unfairness in ML-based decisions.
2) [88] shows empirical results of training a baseline, a fair, an unfair, and a random model on the COMPAS [8] dataset and then finding possible explanations for the results provided by the model. The paper explores a space of explanations that are simple, verifiable, and relevant according to definitions they provide for these terms and shows that it's possible to provide valid justifications for the decisions for the majority of the models. They also explore the situation in which explanations are provided for many data points and then analyzed to verify for consistency, sufficiency, and uniqueness (which they define formally in the paper), and achieve similar results: it can be possible to provide justifications for almost any decision that seem valid, even if the justification does not reflect how the model really decides.
3) [28] discusses the use of Homomorphic Encryption in combination with fair representation learning [85] to provide both privacy and fairness guarantees, respectively, when training machine learning models. Then the paper discusses how to provide local and global explanations for the model.

H. Fairness × Causality

Fairness might be considered a causal concept: we want to know whether the outcome was or was not caused by an unfair process. The causal distinction might be important because, for instance, we might consider some causes fair and others unfair for the decision of not hiring someone. If the cause of the decision was the educational background and work experience of the person, it is usually considered fair, but if it is the skin color of the person it is usually considered unfair. There are many different definitions of causal notions of fairness, each with its own specific meaning, as with causal notions in general.

1) [59] mentions notions of sex discrimination in college admissions and discrimination in hiring. Pearl also argues why the causal interpretation should be used instead of purely statistical notions.
2) [55] and [70] survey many of the causal definitions of fairness present in the literature, [55] also classifies them according to Pearl's levels of causation and guidelines for which situations best suit each notion.
3) [67] shows a causal interpretation of notions of fairness and approaches the conflict between different fairness notions from this perspective. One important note is that this paper does not use the correct notion of equalized odds, it reverses Y and $\hat{Y}$: the correct notion is that the prediction is independent of the sensitive attribute given the ground truth, but they consider a definition such that the output is independent of the sensitive attribute given the prediction. The correct notion of equalized odds is what they call predictive parity. *This paper was found in arxiv.org*.
4) [53] compiles major identifiability results from the causality literature and also discusses how they can be used in practice to obtain causal knowledge from data.

I. Fairness × QIF

One possibility of measuring fairness with Quantitative Information Flow notions is to measure the *reverse flow*: instead of looking at how much observing the output of a model helps in estimating input values (this would be a privacy concern), we measure how much observing the sensitive attribute helps in estimating output values. This idea is explored in [5] and [58].

J. Privacy × Explainability

One of the main questions about the relationship between Privacy and Explainability is whether explainability is affected when privacy-preserving mechanisms are applied. Some results indicate that this is not always the case, but the conceptual differences are discussed times in more detail throughout the literature.

1) [15] evaluates the effect of applying privacy protection mechanisms on Shapley values, which can be used for explanation purposes, both local and global. The conclusion is that applying privacy-protection mechanisms does not affect significatively the quality of explanations based on Shapley values.
2) [33] discusses the legislation of privacy and explainability, arguing that we should not have both in the law, as they claim that demanding explanations can inevitably lead to demands on the visualization of the training data itself.
3) [27] points some conflicts and synergies that involve the two concepts: the conflicts include the conceptual difference between demanding more transparency in respect to the process and demanding secrecy in respect to the training data, which is part of the training process of the model, the possibility of using data obtained from explainability methods to infer sensitive information. The synergies involve the possibility of using information

obtained from explainability methods to improve the privacy of the model, and the fact that some results in the literature claim to have achieved good explainability for some models while keeping acceptable privacy constraints.

4) [28] discusses the use of Homomorphic Encryption in combination with fair representation learning [85] to provide both privacy and fairness guarantees, respectively, when training machine learning models. Then the paper discusses how to provide local and global explanations for the model.

K. Privacy × Causality

Regarding the relationship between privacy and causality, one approach in the literature is to try to provide privacy for sensitive data by the addition of noise, and at the same time allow causal discovery and causal inference based on the noisy data. Another relation noted in the literature is that models based on causal relationships between variables are more resilient to privacy attacks, such that privacy-protection mechanisms are more effective for this type of model. Another possibility is modeling differential privacy as a causal, instead of purely statistical, property.

1) [14] discusses local differential privacy in the context of causal discovery: how local differential privacy mechanisms affect the task of causal discovery. The conclusion is that the geometric mechanism has a smaller effect than the methods based on generalizations of the Randomized Response mechanism, as they tend to degrade less the original correlations present in the data. It seems that this paper also doesn't try to recover the original distribution from the noisy one before trying causal discovery...
2) [71] shows that models that use only the parents of the target variable, which they call causal models, provide stronger ϵ-differential privacy guarantees, are provably more robust to Membership Inference Attacks, generalize better to other distributions based on the same causal structure.
3) [73] discusses how differential privacy can be interpreted as a causal notion and what are the benefits of this interpretation. They analyze some possible causal and associational interpretations of the intuitive idea of Differential Privacy. The causal interpretations consist of interventions on all data points or one data point, and they distinguish between an individual's data point and the real value of the individual. They show the equivalence between many of the definitions they present and the unidirectional implications of two of them. In summary:
   a) Differential privacy constraints on an algorithm that takes many data points as input can be interpreted as requiring that the output value of this algorithm does not change if we change only one data point. This will be considered the definition of Differential Privacy.
   b) This is equivalent to requiring that for any input distributions the probability of any outcome of the algorithm is similar if one data point was observed to be different and the rest equal.
   c) This is equivalent to requiring that for any distribution with independent data points such that any individual data point value is possible, the probability of any outcome of the algorithm is the same if one data point was observed to be different.
   d) Requiring the same as above without the assumption of independent data points is a stronger requirement than differential privacy (it implies DP, but not the opposite).
   e) This is also equivalent to causal notions that require that intervening to change one value and *intervening* to keep the others constant do not change much the probability of any outcome. This holds regardless of whether we consider that for an algorithm to satisfy differential privacy, this must hold for all distributions or for only one (as the interventions are done on *all* variables).
   f) This is also equivalent to the causal notion that interevening on one variable should not change much the probability of each output if we consider this must hold for all probability distributions.
   g) If we consider the previous definition but without requiring to hold for all distributions, then for technical reasons of specific zero probability distributions, this is weaker than differential privacy (DP implies this definition, but not the opposite).
4) [82] discusses new methods to pursue causal discovery and at the same time maintain differential privacy guarantees. The core of the idea considered is to obtain some independence information from the original data in order to reduce the privacy budget necessary for differential privacy. The focus of this paper is classical differential privacy instead of local differential privacy.
5) [42] discusses causal inference instead of causal discovery under differential privacy guarantees, for the causal discovery framework of the Additive Noise Model.

L. Privacy × QIF

As QIF aims to quantify the flow of information, we can naturally consider applying it to the privacy scenario, by measuring how much information leaks about an arbitrary individual. One general idea is that differential

privacy is a worst-case notion and g-vulnerability is an average-case notion, and there are results [6] showing that differential privacy implies bounds on leakage under arbitrary gain functions and prior distributions, but not the opposite.

1) [4] discusses the relations between differential privacy and g-vulnerability, including bounds on g-leakage as a function of the ϵ parameter of differential privacy, and the fact that there is no bound on differential privacy as a function of the g-vulnerability.
2) [26] discusses how the ϵ parameter of Differential Privacy is related to max-case g-vulnerability: e^{ϵ} is exactly the multiplicative max-case channel capacity under a fixed prior. This work also discusses many other theoretical results relating g-vulnerability notions with differential privacy. *Found in arxiv.org*.

M. Explainability × Causality

Causal models are considered inherently interpretable, as the causal relationships between variables are explicitly represented in the model. Many machine learning explanation methods are based only on the data and thus based on the correlations between variables, so they usually lack a causal basis, which is a recurrent criticism of such methods. So, another possible relation between the two concepts is to apply causal tools and measures to improve the quality of the machine learning explanations.

1) [78] discusses how machine learning models in combination with explainability methods may not capture all the relevant correlations between the variables. As mentioned in the paper these explanations are *local to the model*, and the paper shows that important correlations in the data might not be captured by ML models trained with the data. *This paper was found in* arxiv.org.
2) [50] discusses a method of providing causal-based explanations for some types of user queries. The idea is to estimate a causal graph from the data and answer queries of why some effect is observed by estimating the causal effect of each variable. The causal notion used is Database Causality, an extension of Pearl's Actual Causality.
3) [20] provides a broad literature review of the relations between causality and explainability. They identify three main connections between the two concepts present in the literature:
   a) Some works present critics of Explainable AI through a causal perspective, which involves mainly the fact that many explainability approaches lack a causal interpretation, others discuss problems with non-causal explanations in general, and others discuss the adequate form of presenting explanations.
   b) Some works support explainability as a basis for further causal investigation, which might be done empirically.
   c) Finally, some works adhere to the idea of using causal tools to support machine learning model explanations, for instance, the use of the do operator, the estimation of the probability of necessity and the probability of sufficiency, etc.. Causal counterfactual explanations are also considered, and some works consider simply providing a coherent causal model of a system to be a form of explaining the system itself.

N. Explainability × QIF

It might be possible to create an explanation method based on how much information flows through each input to the outputs of a machine learning model. The relation between these two concepts remains largely unexplored. There is, though, an approach by [18, Part III] that uses an alternative definition of information based on Sigma algebras and brings concepts from measure theory to provide some theoretical results involving deep learning.

O. Causality × QIF

The relation between these two concepts still remains largely unexplored. One idea to explore is to try using QIF ideas to estimate what variable will reveal more information according to some specific goal, which might help computationally in deciding which variables to observe.

VI. Conclusions and future work

We can summarize our findings in the following way:

1) **Accuracy × Fairness:** There is an inherent trade-off between accuracy and some notions of fairness such as Equal Opportunity Difference and Statistical Parity.
2) **Accuracy × Privacy:** Privacy by addition of noise usually reduces accuracy. Privacy by homomorphic encryption does *not* affect accuracy at the cost of greater computational complexity.
3) **Accuracy × Explainability:** In many situations, more interpretability implies less accuracy (as more compex models are harder to explain/interpret).
4) **Accuracy × Causality:** Causality might help in the transference of machine learning model results between distinct populations, and the development with causal basis can reduce how sensible it is to small changes in the data used to train the model.
5) **Accuracy × QIF:** accuracy can be seen as a form of utility, which usually has a trade-off with information leakage, measured by QIF.

6) **Fairness** $\times$ **Privacy:** If the privacy-preserving mechanism is based on noise, then it might help in reducing unfairness according to some measures. Most results use the noisy distribution directly, without first trying to recover the original distribution.
7) **Fairness** $\times$ **Explainability:** One possibility for checking whether a system is fair is to demand explanations for the system's working, but sometimes it is possible to deceive this verification.
8) **Fairness** $\times$ **Causality:** there are many distinct causal fairness notions, each with its own meaning and class of applicable situations.
9) **Fairness** $\times$ **QIF:** Fairness might be measured by *reverse* information flow, how much one can infer from the output by observing the sensitive values of the input instead of how much can be inferred from the input from observing the output.
10) **Privacy** $\times$ **Explainability:** Conceptually, more privacy implies less explaination power as the data points can not be used in explanations. In practice, however, privacy constraints do not seem to have an impact on some of the classical XAI methods.
11) **Privacy** $\times$ **Causality:** Causal interpretations were used to disentangle confusions about Differential Privacy assumptions. We can use causal notions to improve the use of differential privacy budget and some methods make causal discovery/inference harder.
12) **Privacy** $\times$ **QIF:** It is possible to use the Quantitative Information Flow g-vulnerability framework to model private information leakage, and there are theoretical results that show relations with differential privacy, with equality for max-vulnerability.
13) **Explainability** $\times$ **Causality:** In many situations, causal explanations the final goal and although classical explainability methods are based only on the data, it might be possible to use them, as basis for further empirical explanation.
14) **Explainability** $\times$ **QIF:** The relationship between these two is largely unexplored. Maybe it's possible to provide explanations based on how much information flows from each feature to the output?
15) **Causality** $\times$ **QIF:** The releationship between these two is largely unexplored too.

There are some possible avenues for future works:

1) Using Quantitative Information Flow notions to develop explainability methods.
2) Using Quantitative Information Flow notions to quantify the flow of information for all types of machine learning attacks.
3) Explore what happens with the Privacy $\times$ Fairness and Privacy $\times$ Accuracy trade-offs under noisy distributions if we first try to recover the original distribution.
4) Exploring the relationship between Quantitative Information and causality, maybe by measuring the information flow between variables to see which variable should be observed to increase the expected information gain on other variables, or other relations.
5) Discover which variables to add noise in order to hinder individual sensitive attribute inference under the least possible privacy budget expense.
6) Using Quantitative Information Flow to develop privacy results similar to the ones obtained with Differential Privacy, maybe by developing new and useful gain functions.
7) Combining existing results with modern Learning Theory, especially by using theoretical results related to Deep Learning.

REFERENCES

[1] Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., Wallach, H.: A reductions approach to fair classification. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 60–69. PMLR (10–15 Jul 2018), https://proceedings.mlr.press/v80/agarwal18a.html

[2] Alufaisan, Y., Marusich, L.R., Bakdash, J.Z., Zhou, Y., Kantarcioglu, M.: Does explainable artificial intelligence improve human decision-making? In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 6618–6626 (2021)

[3] Alves, G., Bernier, F., Couceiro, M., Makhlouf, K., Palamidessi, C., Zhioua, S.: Survey on fairness notions and related tensions. EURO journal on decision processes **11**, 100033 (2023)

[4] Alvim, M.S., Andrés, M.E., Chatzikokolakis, K., Degano, P., Palamidessi, C.: On the information leakage of differentially-private mechanisms. Journal of Computer Security **23**(4), 427–469 (2015)

[5] Alvim, M., Fernandes, N., Nogueira, B., Palamidessi, C., Silva, T.: On the duality of privacy and fairness (extended abstract). In: International Conference on AI and the Digital Economy (CADE 2023). Institution of Engineering and Technology, United Kingdom (2023). https://doi.org/10.1049/icp.2023.2563, 9th International Conference on AI and the Digital Economy, CADE 2023 ; Conference date: 26-06-2023 Through 28-06-2023

[6] Alvim, M., Chatzikokolakis, K., McIver, A., Morgan, C., Palamidessi, C., Smith, G.: The Science of Quantitative Information Flow. Information Security and Cryptography, Springer International Publishing (2020), https://books.google.com.br/books?id=jJH-DwAAQBAJ

[7] Angelov, P.P., Soares, E.A., Jiang, R., Arnold, N.I., Atkinson, P.M.: Explainable artificial intelligence: an analytical review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **11**(5), e1424 (2021)

[8] Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. URL https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing (2019)

[9] Arcolezi, H.H., Makhlouf, K., Palamidessi, C.: (local) differential privacy has no disparate impact on fairness. In: IFIP Annual Conference on Data and Applications Security and Privacy. pp. 3–21. Springer (2023)

[10] Ateniese, G., Mancini, L.V., Spognardi, A., Villani, A., Vitali, D., Felici, G.: Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. International Journal of Security and Networks **10**(3), 137–150 (2015)

[11] Bell, A., Bynum, L., Drushchak, N., Zakharchenko, T., Rosenblatt, L., Stoyanovich, J.: The possibility of fairness: Revisiting the impossibility theorem in practice. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. pp. 400–422 (2023)

[12] Bell, A., Solano-Kamaiko, I., Nov, O., Stoyanovich, J.: It's just not that simple: an empirical study of the accuracy-explainability trade-off in machine learning for public policy. In: Proceedings of the 2022 ACM conference on fairness, accountability, and transparency. pp. 248–266 (2022)

[13] Belle, V., Papantonis, I.: Principles and practice of explainable machine learning. Frontiers in big Data **4**, 688969 (2021)

[14] Binkyte, R., Pinzón, C.A., Lestyán, S., Jung, K., Arcolezi, H.H., Palamidessi, C.: Causal discovery under local privacy. In: Causal Learning and Reasoning. pp. 325–383. PMLR (2024)

[15] Bozorgpanah, A., Torra, V., Aliahmadipour, L.: Privacy and explainability: The effects of data protection on shapley values. Technologies **10**(6), 125 (2022)

[16] Burkart, N., Huber, M.F.: A survey on the explainability of supervised machine learning. Journal of Artificial Intelligence Research **70**, 245–317 (2021)

[17] Calin, O.: Deep learning architectures. Springer (2020)

[18] Calin, O.: Deep learning architectures. Springer (2020)

[19] Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., Tramer, F.: Membership inference attacks from first principles. In: 2022 IEEE Symposium on Security and Privacy (SP). pp. 1897–1914. IEEE (2022)

[20] Carloni, G., Berti, A., Colantonio, S.: The role of causality in explainable artificial intelligence. arXiv preprint arXiv:2309.09901 (2023)

[21] Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning interpretability: A survey on methods and metrics. Electronics **8**(8), 832 (2019)

[22] Chouldechova, A., Roth, A.: The frontiers of fairness in machine learning. arXiv preprint arXiv:1810.08810 (2018)

[23] Cummings, R., Gupta, V., Kimpara, D., Morgenstern, J.: On the compatibility of privacy and fairness. In: Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization. p. 309–315. UMAP'19 Adjunct, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3314183.3323847, https://doi.org/10.1145/3314183.3323847

[24] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. p. 214–226. ITCS '12, Association for Computing Machinery, New York, NY, USA (2012). https://doi.org/10.1145/2090236.2090255, https://doi.org/10.1145/2090236.2090255

[25] Dwork, C., Ilvento, C.: Fairness under composition. arXiv preprint arXiv:1806.06122 (2018)

[26] Fernandes, N., McIver, A., Sadeghi, P.: Explaining epsilon in local differential privacy through the lens of quantitative information flow. arXiv preprint arXiv:2210.12916 (2022)

[27] Ferry, J., Aïvodji, U., Gambs, S., Huguet, M.J., Siala, M.: Sok: Taming the triangle - on the interplays between fairness, interpretability and privacy in machine learning. ArXiv **abs/2312.16191** (2023), https://api.semanticscholar.org/CorpusID:266573131

[28] Franco, D., Oneto, L., Navarin, N., Anguita, D.: Toward learning trustworthily from data combining privacy, fairness, and explainability: an application to face recognition. Entropy **23**(8), 1047 (2021)

[29] Fredrikson, M., Jha, S., Ristenpart, T.: Model inversion attacks that exploit confidence information and basic countermeasures. In: Proceedings of the 22nd ACM SIGSAC conference on computer and communications security. pp. 1322–1333 (2015)

[30] Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P., Roth, D.: A comparative study of fairness-enhancing interventions in machine learning. In: Proceedings of the conference on fairness, accountability, and transparency. pp. 329–338 (2019)

[31] Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). pp. 80–89. IEEE (2018)

[32] Goswami, M.J.: Privacy-preserving deep learning using secure multi-party computation. International IT Journal of Research, ISSN: 3007-6706 **2**(2), 50–55 (2024)

[33] Grant, T.D., Wischik, D.J.: Show us the data: Privacy, explainability, and why the law can't have both. Geo. Wash. L. Rev. **88**, 1350 (2020)

[34] Grohs, P., Kutyniok, G.: Mathematical aspects of deep learning. Cambridge University Press (2022)

[35] Hellman, D.: Measuring algorithmic fairness. Virginia Law Review **106**(4), 811–866 (2020)

[36] Henao, C.P.: Exploring fairness and privacy in machine learning. Ph.D. thesis, Institut Polytechnique de Paris (2023)

[37] Hesamifard, E., Takabi, H., Ghasemi, M.: Cryptodl: Deep neural networks over encrypted data. arXiv preprint arXiv:1711.05189 (2017)

[38] Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P.S., Zhang, X.: Membership inference attacks on machine learning: A survey. ACM Computing Surveys (CSUR) **54**(11s), 1–37 (2022)

[39] Kearns, M., Neel, S., Roth, A., Wu, Z.S.: Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In: International conference on machine learning. pp. 2564–2572. PMLR (2018)

[40] Konstantinov, N., Lampert, C.H.: On the impossibility of fairness-aware learning from corrupted data. In: Algorithmic Fairness through the Lens of Causality and Robustness workshop. pp. 59–83. PMLR (2022)

[41] Krishnan, M.: Against interpretability: a critical examination of the interpretability problem in machine learning. Philosophy & Technology **33**(3), 487–502 (2020)

[42] Kusner, M.J., Sun, Y., Sridharan, K., Weinberger, K.Q.: Private causal inference. In: Artificial Intelligence and Statistics. pp. 1308–1317. PMLR (2016)

[43] Li, L., Lassiter, T., Oh, J., Lee, M.K.: Algorithmic hiring in practice: Recruiter and hr professional's perspectives on ai use in hiring. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. pp. 166–176 (2021)

[44] Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable ai: A review of machine learning interpretability methods. Entropy **23**(1), 18 (2020)

[45] Lipton, Z.C.: The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue **16**(3), 31–57 (2018)

[46] Liu, B., Ding, M., Shaham, S., Rahayu, W., Farokhi, F., Lin, Z.: When machine learning meets privacy: A survey and outlook. ACM Computing Surveys (CSUR) **54**(2), 1–36 (2021)

[47] Liu, C., Chakraborty, S., Mittal, P.: Dependence makes you vulnberable: Differential privacy under dependent tuples. In: NDSS. vol. 16, pp. 21–24 (2016)

[48] London, A.J.: Artificial intelligence and black-box medical decisions: accuracy versus explainability. Hastings Center Report **49**(1), 15–21 (2019)

[49] Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)

[50] Ma, P., Ding, R., Wang, S., Han, S., Zhang, D.: Xinsight: explainable data analysis through the lens of causality. Proceedings of the ACM on Management of Data **1**(2), 1–27 (2023)

[51] Makhlouf, K., Arcolezi, H.H., Zhioua, S., Brahim, G.B., Palamidessi, C.: On the impact of multi-dimensional local differential privacy on fairness. Data Mining and Knowledge Discovery pp. 1–24 (2024)

[52] Makhlouf, K., Stefanovic, T., Arcolezi, H.H., Palamidessi, C.: A systematic and formal study of the impact of local differential privacy on fairness: Preliminary results (2024), https://arxiv.org/abs/2405.14725

[53] Makhlouf, K., Zhioua, S., Palamidessi, C.: Identifiability of causal-based ml fairness notions. In: 2022 14th international conference on computational intelligence and communication networks (CICN). pp. 1–8. IEEE (2022)

[54] Makhlouf, K., Zhioua, S., Palamidessi, C.: Survey on causal-based machine learning fairness notions (2022)

[55] Makhlouf, K., Zhioua, S., Palamidessi, C.: When causality meets fairness: A survey. Journal of Logical and Algebraic Methods in Programming p. 101000 (2024)

[56] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM computing surveys (CSUR) **54**(6), 1–35 (2021)

[57] Meroño-Peñuela, A., Ashkpour, A., Guéret, C., Schlobach, S.: Cedar: the dutch historical censuses as linked open data. Semantic Web **8**(2), 297–310 (2017)

[58] Nogueira, B.D., et al.: On the relation of privacy and fairness through the lenses of quantitative information flow (2023)

[59] Pearl, J.: Causality: Models, Reasoning and Inference. Cambridge University Press, USA, 2nd edn. (2009)

[60] Petkovic, D.: It is not "accuracy vs. explainability"—we need both for trustworthy ai systems. IEEE Transactions on Technology and Society **4**(1), 46–53 (2023)

[61] Pinzón, C., Palamidessi, C., Piantanida, P., Valencia, F.: On the incompatibility of accuracy and equal opportunity. Machine Learning (May 2023). https://doi.org/10.1007/s10994-023-06331-y, https://doi.org/10.1007/s10994-023-06331-y

[62] Ren, X., Yu, C.M., Yu, W., Yang, S., Yang, X., McCann, J.A., Philip, S.Y.: Lopub: high-dimensional crowdsourced data publication with local differential privacy. IEEE Transactions on Information Forensics and Security **13**(9), 2151–2166 (2018)

[63] Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)

[64] Richens, J.G., Lee, C.M., Johri, S.: Improving the accuracy of medical diagnosis with causal machine learning. Nature communications **11**(1), 3923 (2020)

[65] Roscher, R., Bohn, B., Duarte, M.F., Garcke, J.: Explainable machine learning for scientific insights and discoveries. Ieee Access **8**, 42200–42216 (2020)

[66] Rueda, J., Rodríguez, J.D., Jounou, I.P., Hortal-Carmona, J., Ausín, T., Rodríguez-Arias, D.: "just" accuracy? procedural fairness demands explainability in ai-based medical resource allocations. AI & society pp. 1–12 (2022)

[67] Saravanakumar, K.K.: The impossibility theorem of machine fairness–a causal perspective. arXiv preprint arXiv:2007.06024 (2020)

[68] Schölkopf, B.: Causality for machine learning. In: Probabilistic and causal inference: The works of Judea Pearl, pp. 765–804 (2022)

[69] Sheikh, M.A., Goel, A.K., Kumar, T.: An approach for prediction of loan approval using machine learning algorithm. In: 2020 international conference on electronics and sustainable communication systems (ICESC). pp. 490–494. IEEE (2020)

[70] Su, C., Yu, G., Wang, J., Yan, Z., Cui, L.: A review of causality-based fairness machine learning. Intelligence & Robotics **2**(3), 244–274 (2022)

[71] Tople, S., Sharma, A., Nori, A.: Alleviating privacy attacks via causal learning. In: International Conference on Machine Learning. pp. 9537–9547. PMLR (2020)

[72] Tramèr, F., Zhang, F., Juels, A., Reiter, M.K., Ristenpart, T.: Stealing machine learning models via prediction {APIs}. In: 25th USENIX security symposium (USENIX Security 16). pp. 601–618 (2016)

[73] Tschantz, M.C., Sen, S., Datta, A.: Sok: Differential privacy as a causal property. In: 2020 IEEE Symposium on Security and Privacy (SP). pp. 354–371. IEEE (2020)

[74] Valdivia, A., Sánchez-Monedero, J., Casillas, J.: How fair can we go in machine learning? assessing the boundaries of accuracy and fairness. International Journal of Intelligent Systems **36**(4), 1619–1643 (2021)

[75] Vapnik, V.N.: An overview of statistical learning theory. IEEE transactions on neural networks **10**(5), 988–999 (1999)

[76] van der Veer, S.N., Riste, L., Cheraghi-Sohi, S., Phipps, D.L., Tully, M.P., Bozentko, K., Atwood, S., Hubbard, A., Wiper, C., Oswald, M., et al.: Trading off accuracy and explainability in ai decision-making: findings from 2 citizens' juries. Journal of the American Medical Informatics Association **28**(10), 2128–2138 (2021)

[77] Verma, S., Dickerson, J., Hines, K.: Counterfactual explanations for machine learning: A review. arXiv preprint arXiv:2010.10596 **2**, 1 (2020)

[78] Vowels, M.J.: Trying to outrun causality with machine learning: Limitations of model explainability techniques for identifying predictive variables. stat **1050**, 22 (2022)

[79] Wang, B., Gong, N.Z.: Stealing hyperparameters in machine learning. In: 2018 IEEE symposium on security and privacy (SP). pp. 36–52. IEEE (2018)

[80] Wang, Z., Song, M., Zhang, Z., Song, Y., Wang, Q., Qi, H.: Beyond inferring class representatives: User-level privacy leakage from federated learning. In: IEEE INFOCOM 2019-IEEE conference on computer communications. pp. 2512–2520. IEEE (2019)

[81] Xiong, X., Liu, S., Li, D., Cai, Z., Niu, X.: A comprehensive survey on local differential privacy. Security and Communication Networks **2020**(1), 8829523 (2020)

[82] Xu, D., Yuan, S., Wu, X.: Differential privacy preserving causal graph discovery. In: 2017 IEEE Symposium on Privacy-Aware Computing (PAC). pp. 60–71. IEEE (2017)

[83] Yang, M., Guo, T., Zhu, T., Tjuawinata, I., Zhao, J., Lam, K.Y.: Local differential privacy and its applications: A comprehensive survey. Computer Standards & Interfaces p. 103827 (2023)

[84] Yang, Z., Zhong, S., Wright, R.N.: Privacy-preserving classification of customer data without loss of accuracy. In: Proceedings of the 2005 SIAM International Conference on Data Mining. pp. 92–102. SIAM (2005)

[85] Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: International conference on machine learning. pp. 325–333. PMLR (2013)

[86] Zeng, Z., Peng, W., Zhao, B.: Improving the accuracy of network intrusion detection with causal machine learning. Security and Communication Networks **2021**(1), 8986243 (2021)

[87] Zheng, H., Hu, H., Han, Z.: Preserving user privacy for machine learning: Local differential privacy or federated machine learning? IEEE Intelligent Systems **35**(4), 5–14 (2020)

[88] Zhou, J., Joachims, T.: How to explain and justify almost any decision: Potential pitfalls for accountability in ai decision-making. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. p. 12–21. FAccT '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3593013.3593972, https://doi.org/10.1145/3593013.3593972