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

Type: Scientific
Advisor: Mário Sérgio Alvim

Artur Gaspar da Silva

05/04/2024

Abstract

Do que podemos falar aqui?

1 Trocar acurácia por prediction strength?

2 Introduction

Recent resarch[6][1][4][5] 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[10]. Also, there has been some work on introducing Causality concepts into the discussion, for instance, to develop better fairness metrics[8]. It has also been suggested to use naturally-interpretable models (as explanations to more complex models) for auditing systems and checking if they are fair, although this might lead to problems[12]. 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[3], loan approvals[11], hiring decisions[7], and others.

In this first part of the Undergraduate Thesis, 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 3 discusses the already developed theoretical work on these areas, section 4 discusses connections between them found in the literature, section 5 provides the contributions of this analysis in the form of a higher-level discussion of what we can gather from the literature and the meaning of these results, and section 6 provides conclusions and possible future lines of work.

3 Theoretical Reference: 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).

3.1 Machine Learning

Machine Learning is the field of study that focuses on developing methods of learning patterns from data in a way that generalizes to situations not present in the data. In recent years, significative advances have been observed in Machine Learning research, and also the popularity and applications of some methods increased

significantly. One example is the improvement of Neural Network Architectures[**?**], and the popularity of Generative Models[**?**]. Also, some recent research is focused on the theory behind such machine learning methods[**?**], and statistical learning in general [**?**]. Part of the goal of the formalization of such theory is to provide more qualitative guarantees, for instance, in regard not only to accuracy, but also fairness, privacy and interpretability. We discuss these different goals in the next four subsections.

In general, we consider **supervised learning** problems: a machine learning model 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 without the target value and outputs a prediction of the target value, which represents one or more variable of interest. This is callled supervised learning because the algorithm has acces to the target variable during the training process.

3.2 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. 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 correct. 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 access how accurate a regression model is in many ways, for instance the square difference between the prediction value and the real value, added through all training data points.

One important point to consider is that it is better to evaluate how accurate the system is with data different from the data used for training. This is because 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, then it might be possible to obtain a model that has a very good accuracy, but if we try to use this same model on other data, this same model performs very badly. For instance, if all functions from the training data to the output are allowed, then a model that simply memorizes the training data and outputs the correct result by looking at the memorized data and outputs a random answer if the input is not in the training data will have one hundred percent accuracy in the training data, but we can obviously provide no guarantee for data points not present. This problem is called **overfitting**, and is usually avoided by limiting how powerfull the model can be, in conjuction with verifying real accuracy values with data other than the training data, and we call this the **testing data**. Also, notice 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*, and an 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. However, with the groth of Machine Learning applications in real-life scenarios and the impact on society as a whole, there has been a crescent focus on aspects other than simply maximizing the accuracy, in a similar way that we usually are worried only about how fast a cryptographic algorithm is but also about how safe it is.

3.3 Fairness

Dava pra citar Hellman D. Measuring algorithmic fairness pra falar que "Indeed, the general consensus is that the meaning and implication of each approach highly depends on the context and consequential decisions associated with an intelligent system"

Citar Pushing the limits of fairness impossibility: Who's the fairest of them all? De que em geral não dá pra satisfazer as noções principais de fairness ao mesmo tempo, mas dá pra tentar encontrar meio termos e otimizar uma em função da outra. The Possibility of Fairness: Revisiting the Impossibility Theorem in Practice parece melhor porque é mais empírico. The impossibility of "fairness": a generalized impossibility result for decisions fortalece isso um pouco e foi escrito por um cara de harward mas ficou só no arxiv.

In the context of Machine Learning, fairness refers to the reduction, as much as possible, of **algorithmic bias**. Algorithm Bias is the bias introduced by algorithmic decisions. This bias might have a big social impact, as this can further existing unfair discrimination in society, as machine learning algorithms are being used to make more and more important decisions. One famouse example is the COMPAS recidivism algorithm, that has been used by the United States courts to estimate how likely someone is to reoffend in the future. It was revealed [3] that this tool was heavily biased against black people.

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 **unpriviledged group** is the group of people affected negatively by the bias, and the **priviledged group** is the other group of people.

Such biases can happen because of many factors. The algorithm itself might be introducing bias, or the data might be biased. The data might have been collected in a biased way (in the compass 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 of the society.

Also, the bias in society might be such that the data is in disagreement with reality (the unpriviledged group true values for the target variable would affect them in the same way as the priviledged group), or it is in agreement with reality because of structural biases in society. For instance, if the prediction of the algorithm is whether or not someone will have good grades if accepted to some university, people in the unpriviledged group might not have had as good oportunities in life as people in the priviledged group, so the data is correct when it says that those people will have worse grades. Even though, the results might still be considered unfair: this depends on the notion of fairness we consider. All of these unfairness possibilities can be futther divided into other types of unfairness, as was done in [**?**].

Besides deliberate bias in the algorithm, such that the results of the algorithm do not reflect the data, and biased 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 [**?**], which might be considered 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 [**?**]. Initially, the notions of fairness could be grouped into two main types: statistical and individual definitions of fairness[**?**]. Statistical (group) notions of fairness require some statistical metric 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 [**?**][**?**]. Some of these problems include: satisfying the constraints for two protected attributes individually but not to 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 [**?**]. Finally, there are also causal approaches to fairness notions, which we will discuss more in Section 4.

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 modifies the learning algorithm iteself, 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.

We summarize below the ways in which unfairness might be introduced:

1. Algorithm results does 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 be considered a type of Aggregation Bias.
   (b) Systematic errors in the algorithm, that leads to biased estimation.
2. The Data can be biased, not reflecting the reality.
   (a) We can have a structural biases in society, such that people in unpriviledged groups do not have the necessary oportunities, but if they were treated similary to the priviledged group by society, they would have similar results. For instance, an unpriviledged group that doesn't have good education oportunities 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 be only perpetuate current unfair biases in society.
   (b) The bias can also be introduced in a way that people in the unpriviledged group were misclassified before the data was colected, for instance maybe capable people in an unpriviledged group usually don't get a job even though they are actually as capable as the unpriviledged 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), Ommited 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 [9]).
   (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 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 itens are shown more). This might strengthen biases through time.
4. Finally, the circunstances can change through time, either by influence of the algorithmt itself or other factors, which can worsen the quality of algorithms previously considered to provide good results (Emergent Bias).

3.4 Privacy

Privacy references: https://dl.acm.org/doi/abs/10.1145/3436755 and https://ieeexplore.ieee.org/document/9433648. The second one doesn't add much.

WMLMPASAO: When Machine Learning Meets Privacy: A Survey And Outlook

Membership Inference Attacks Against Machine Learning Models

Privacy-preserving classification of customer data without loss of accuracy

CryptoDL: Deep Neural Networks over Encrypted Data

About dependence between individuals: Dependence Makes You Vulnerable: Differential Privacy Under Dependent Tuples

Falar de DP e LDP também, são importantes: A Comprehensive Survey on Local Differential Privacy

Falar de Federated Learning também!

In the context of Machine Learning, a privacy-preserving algorithm is one that doesn't allow information considered private/sensitive to be obtained by unauthorized parties. This has been named *private ML*[**?**], and the private data to be protected can be the data using to train the model or the model parameters and structure itself. It is also possible to use Machine Learning to enhance privacy, or to serve as an attack tool. We focus on *private ML*.

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 (Model Extraction Attack) by trying to approximate the function that represents the model, to recover some feature or statistical property of the dataset (Feature Estimation Attack), to discover whether some individual data point is present in the dataset (Membership Inference Attack), or even recover the exact values of individual samples in the datset (Model Memorization Attack). 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.

One approach to improving privacy in Machine Learning is by encrypting the data or the model in a way such that the computations can be done with the encrypted data/model, and just the result is decrypted. Secure Multi-Party Computation is also an option if there are multiple parties responsible for this computation. Other option is to obfuscate the data or the model, by introducing pertubation in a way that protects the private information in a way that (hopefully) doesn't worsen much the results. There is also the aggregation approach, that focuses on multy-party computations such that the data of one party remains private to other parties, and is used by aggregation learning, for instance. Finally, it's possible to develop *back-door* attacks: malicious modifications to the model might allow an adversary to infer private information, if she provides the specific inputs that are designed to trigger that. All of these are listed in more detail in [**?**].

3.5 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, [**?**] 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. The paper [**?**] draws this distinction in a different way: 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 we also want important biases in the model to be evident in the explanation, as mentioned in [**?**].

There are some classifications of the existing approaches to explainability, but we focus on the classification defined in [**?**]. We can divide the approaches based on whether they are **local** or **global**: local explanations focus on explaining individual decisions, and global explanations focus on explaining the overall workings of the model. LIME [**?**] and SHAP [**?**] are examples of methods that provide local explanations and can be used to derive global explanations. We can also divide the approaches into **Model Agnostic** and **Model Specific**: the former refers to methods that don't depend on the inner workings of the model (for instance, SHAP [**?**] and LIME [**?**]), the latter 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). Finally, we can divide explainability approaches on the data types these methods deal with and the purpose of the explanations.

As discussed at [**?**], explanations may be usefull to data scientists, business owners, model risk analysts, regulators and consummers, 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 pertubations), 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 criterias for evaluating explanations, that 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 [**?**]. They also evaluate explanations by example (for instance, counterfactuals [**?**] 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-theoretic 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 [**?**], explanations can also be used to enhance scientific research in natural sciences.

Some models are inheretely interpretable, as their inner working is easier to understand for humans. Some examples mentioned in [**?**] are: Linear
Logistic Regression, Decision Trees, K-Nearest Neighbors (KNN), Rule-based learning, Generalized Additive Models (GAMs) and Bayesian networks. Some researchers criticise this claim for some of these models, for instance, [**?**, **?**] points out that linear models are not necessarily easily interpretable, as they sometimes rely on un-interpretable and heavily-engineered features.

Other criticisms to the current development of explainability approaches were made: for instance, [**?**] argues that explainability is always a means to an end, and by requiring explanations to our models we may be significaltly 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 do 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 [**?**]). Another possible problem with explainability is that if we rely on humans opinions we might end up with methods that are *persuasive*, instead of *accurate*, as these two properties might not be fully aligned.

An in-depth survey on machine learning explainability can be found at [**?**], with some definitions regarding ontology and the philosophical meaning of explanations and knowledge sharing.

Esse paper parece discutir porque black-box pode ser aceitável em ML pra decisões médicas ainda por cima: Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability

3.6 Causality

Most of our discussion of causality will be based on the work of Judea Pearl [9]. There are many alternative notions of causality, but most of them are already discussed in [9]. Judea Pearl divides causality into three levels, according to the type of question we want to answer.

The first level of causation deals with questions that can be answered by looking at the data. For instance, questions that can be answered by Bayesian networks, deep neural networks or other machine learning algorithms. 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. Although the examples 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. The notation used by Pearl is the usual probabilistic notation, $P(Y = y|X = x)$, abreviated as $P(y|x)$, this means that we observed $X = x$ and want to know how likely it is for Y to have the value y.

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 completelly different, 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 of when people buy more ice creams. Pearl proclaims 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. There are many other types of assumptions that can be made, Pearl discusses them in detail at [9]. The notation used by Pearl is $P(Y = y|do(X = x))$, which can be abreviated 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.

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 make an observation about something that happened, 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 adds all the uncertainty to the values of unobserved variables. Also, in the way the 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. 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. This is different from $P(Y = y|do(X = x'), X = x)$, as this 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 subsection 11.7.2 of [9].

In general, Judea Pearl's approach to causality assumes an Directed *Acyclic* Graph (DAG) for the relations between variables. This is what is called **recursive models** in [9]. Many results apply only to this type of models, but some generalizations are available and mentioned in [9]. 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 subsection 1.4.1 of [9]. Pearl also argues extensively about the possible causal interpretation of such equations, and defines Structural Causal Models as modens in which each variable has an equation that determines its value. This type of model can be used for answering questions of the third level of causation.

Pearl presents many results about **Causal Discovery** on the second chapter of [9]. Many theoretical results are presented of what can and can not be done regarding 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 model, 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 marginalyl 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 assumptions 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, $A \to B$ (meaning that A causes B) is indistinguishable from $A \leftarrow B$ or $A \leftarrow U \to B$ 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 $A \leftarrow U \to B$ 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 increase the probability of the outcome 0, 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 infering attributes of a causal model given it's strctured. In this situation, we have some causal model a priori, and want to estimate quantities such as the $P(Y = y|do(X = x))$. Pearl introduces the *do* **calculus**, which provides a way of deriving expressions for such quantities that do not depend on a 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 suficient, 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 the identifying the results of dynamic plans, in which one action comes after others and might depend on the results of previous actions. This is further discussed in section 4.4 of [9]. In section 4.5, Pearl dives into two other causal concepts: **Direct and Indirect Effects**, which regard the effects that some variable X have on other 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 more direc 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 of the exact values of the variables in question.

Another contribution of the causality framework presented in [9] 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 arrise due to confounding, but in other situations controling 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 chapter 6 of [9] in favor of the causal approaches to confounding.

Counterfactual statements are defined in terms of simpler axioms, and Chapter 7 of [9] further discusses many results about counterfactuals. These results include how to use an Structural Causal Model to estimate the value counterfactual statements, or how to check if this is even possible to estimate only with the SCM and data. Chapter 8 delves into the problem of bounding values of expressions that can not be computed directly. 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, if we can experimentally change the value of some variables via external interventions then some effects are easier to estimate.

Finally, there are some subtleties to the meaning of causation in different settings. Pearl discusses many notions of causation, including the probability of necessity, the probability of sufficiency, the natural direct effect (which we already mentined), the actual cause, and others. 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**. We say that $X = 1$ was a necessary 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**. 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 $P(Y_{X=1} = 1, Y_{X=0} = 0)$. Pearl also defines the **Probability of Disablement** as $P(Y_{X=0} = 0|Y = 1)$ and the **Probability of Enablement** as $P(Y_{X=1} = 1|Y = 0)$. Many relations between these values, bounds and conditions for identification are presented in Chapter 9 of [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 happenned, 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 we have: $X = x$ 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 circunstances $U = u$ 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 Y, keeping everyting else as it is with $U = u$. This means that in the situation that actually happenned $U = u$, then $X = x$ is enough to *sustain* $Y = y$ under inverventions 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 circunstances $U = u$ and another causal model, such that the parents of each node in the new model are sufficient to entain 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 circunstances 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 $f_i(x_1) = ax_1$.

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 circunstances $U = u$ 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 a 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 Chapter 10 of [9].

3.7 Quantitative Information Flow

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

We define an **adversary** as an agent that tries to gain something with the information that leaks from the system, the **secret** as the data that the system processes, the **prior** as the distribution on secrets that represents the knowledge of the adversary before the system is run, and the **posterior** as the hyper-distribution (a distribution on distributions) on secrets that represents the knowledge of the adversary after the system is run. We use an information-theoretic **channel** to represent the distributions of observable values outputed from the system, which might depend on the secret value. The set $\mathcal{X}$ represents the set of possible values of the secrets, the set $\mathcal{Y}$ represents the possible values of observable outputs of the system, and the set $\mathcal{W}$ represents the possible values of actions the adversary might take.

We consider the g-vulnerability framweork, introduced in [2]. We define a gain 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 actual secret value is x. We consider a zero-sum game: the gain of the adversary is exactly the loss of the people responsible for the system. The **prior vulnerability** of the system is defined as the average gain of the adversary if

she takes the action that maximizes her gain, according to the prior distribution on secrets that represents her knowledge of the secret. The **posterior vulnerability** of the system is defined in the same way, but considering the hyper-distribution that represents the posterior knowledge. Notice that each distribution in the posterior hyper-distribution represents the knowledge of the adversary after some of the possible observations after the system is run, so it represents what we, before the system runs, estimate will be the future values of the adversarial knowledge after the system runs. The **additive leakage** can be defined as the difference between posterior and prior vulnerability, and the **multiplicative leakage** as the result of dividing of the posterior vulnerability by the prior vulnerability values. As teorema leak ¿= 1 ¿= 0

There are some valuable theoretical results about channels regarding the relationships between prior and posterior g-vulnerabilities. Chapter 7 of [2] shows results about the *capacity* of a channel, which is the maximum possible (aditive or multiplicative) leakage that can happen through a channel if we fix either the prior, the gain function or neither. Chapter 9 discusses results about *refinement* of channels: in short, a channel is strictly better (for all priors and gain functions) to another channel in respect to the posterior vulnerability if and only if it can be written as a post-processing of this other channel. Chapter 10 discusses 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. Chapter 11 [2] 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.

4 Theoretical Reference: relations between the concepts

In this section we mention comparisons found in the literature between these areas of research.

4.1 Accuracy × Fairness

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

1. [10] shows that there are trade-offs between Equal Oportunity Difference and accuracy such that, depending on the data distribution, it might impossible to achieve both perfect Equal Oportunity Difference and non-trivial accuracy. It also shows some other theoretical results that associate EOD and accuracy, such as sufficient conditions for the existance 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 [3] dataset. One interesting results is that for some datasets (such as the Compas dataset), it is possible to reduce Equal Opportunity Difference without changing much the accuracy, but for some datasets (such as the Dutch Census Dataset[**?**] with gender as the protected attribute and the goal is if someone has a prestigious occupation) Dutch Census deve ser esse: https://www.researchgate.net/profile/Eric-Schulte-Nordholt/publication/286099390_The_Dutch_Census_2011/links/58a3228b458515d15fd942d2/The-Dutch-Census-2011.pdf

3. [**?**] 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. [**?**] Provides a method of finding the full Paretto front of accuracy versus fairness. Their approach is based on a genetic algorithm, and the notion of fairness that they consider is False-Positive Rate for avoiding disparate mistreatment, but it is possible to use most metrics available in the literature.

5. [?] provides 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.

4.2 Accuracy × Privacy

If we consider obfuscation methods for privacy (such as Differential-Privacy and Local Differential Privacy mechanisms), in general bigger privacy constraints imply in 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 the accuracy. In fact, some predictions can be made without any loss of accuracy at all, as shown in [?] for the naive Bayes classifier. Notice that in the Local Differential Privacy context, we know exactly how the noise is applied an thus might be able to reverse some 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. [?] 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. [?] compares Local Differential Privacy and Federated Learning approaches empirically, concluding that LDP approaches consume less computational resources on the client-side, benefits from a large user population and can be reused for other tasks, while the data transfered to the server for the Federated Learning approach is specific for one inference task.
3. [?] 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. [?] 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 scenarios. This might be important to consider in the machine learning scenario as well in the future, as the advesary might take advantage of such correlations in this type of situation as well.

4.3 Accuracy × Explainability

If the Machine Learning model is inherentely 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 [?]. Also, for simpler models it might be possible to keep good accuracy and explainability levels.

1. [?] 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. [?] 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. [?] 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. [?] and [?] 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. [?] 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.

4.4 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 causality concepts could be used to transfer what was learned with the data of one population to another, for instance. Another possible relation is evaluating how accurate are causal discovery and causal inference algorithms. In general, these two concepts are unrelated.

4.5 Accuracy × QIF

Acuracy 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.

4.6 Fairness × Privacy

There are some impossibility results in the literature that argue why it is impossible to achieve both pricacy and group fairness constraints under non-trivial accuracy[**?**]. We also have positive results that show how stasifying privacy constraints can help mitigating group fairness[**?**][**?**][**?**], and we also have similar positive results results for individual fairness notions[5].

1. [**?**] 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 persepctive.

2. [**?**] 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. [**?**] 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. [5] 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. [**?**] present 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.

4.7 Fairness × Explainability

Some people try to judge fairness based on the explanation, we can mention the How to Justify Almost Anything aqui, que pra auditar pode não ser uma boa ideia, só isso mesmo talvez.

1. [**?**] discusses how not providing explanations for decision can threaten accountability, bias avoidance and transparency when evaluating and mitigating unfairness in ML-based decisions.

4.8 Fairness × Causality

Citar paper Karima sobre noções causais de fairness. Citar tmb THE IMPOSSIBILITY THEOREM OF MACHINE FAIRNESS A CAUSAL PERSPECTIVE, esse segundo tem um problema de considerar equalized odds ao contrário.

4.9 Fairness × QIF

Citar trabalho do Bruno sobre QIF ao contrário como noção de fairness talvez

4.10 Privacy × Explainability

No idea!

4.11 Privacy × Causality

No idea! Maybe be private to causal discovery? Maybe be private but allow causal discovery? I think that there is a paper by Sylvia about this...

4.12 Privacy × QIF

QIF was desined to work with privacy, quantifying how much the sensitive information is leaking is in a way quantifying privacy.

4.13 Explainability × Causality

Causality is inheretely easier to explain.

4.14 Explainability × QIF

No idea! Maybe a way to see which variables leak more information?

4.15 Causality × QIF

I don't think that the paper I saw was published, but maybe we can mention that.

5 Contributions

Citar artigos lidos e o livro de causalidade, falando que as atividades consistiram principalmente de entender um pouco mais do assunto, e dar uma discutida resumida das coisas que foram vistas.

6 Conclusions and future work

Aqui dá pra falar sobre quais dos caminhos parecem mais promissores pra fazer uma pesquisa futura...

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] 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

[3] 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)

[4] 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

[5] 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

[6] 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

[7] 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)

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

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

[10] 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

[11] 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)

[12] 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