Privacy and AI Ethics – Understanding the convergences and tensions for the responsible development of machine learning

Funding[1]

Contributions Program 2020-2021 of the Office of the Privacy Commissioner of Canada

Principal investigator :

Sébastien Gambs UQAM gambs.sebastien@uqam.ca

Authors[2]

Ulrich Aïvodji	UQAM	aivodji.ulrich@uqam.ca
Céline Castets-Renard	University of Ottawa	celine.castets-renard@uottawa.ca
Ignacio Cofone	McGill University	ignacio.cofone@mcgill.ca
Sébastien Gambs	UQAM	gambs.sebastien@uqam.ca
Aude Marie Marcoux	UQAM	marcoux.aude_marie@courrier.uqam.ca
Dominic Martin	UQAM	martin.dominic@uqam.ca

Contributors[3]

Olga Abimana	University of Ottawa
Louis Béziaud	UQAM – Université de Rennes 1
Brandon Bonan	McGill University
Sophie Gagné-Landmann	McGill University
Edynne Grand-Pierre	University of Ottawa
Ana Qarri	McGill University

[1]The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Office of the Privacy Commissioner of Canada.

[2]The authors are listed by alphabetical order. All the authors have contributed equally to the report.

[3]We thank the following persons for their contributions to the research conducted in this project.

Contents

1 Introduction

Context. The democratization of mobile systems and the development of information technologies have been accompanied by a massive increase in the amount and the diversity of data collected about individuals, often referred to as Big Data. Furthermore, in Machine Learning, the Deep Learning revolution [LBH15] coupled with the access to Big Data has enabled a "quantum leap" in the prediction power in many domains, which has led to the possibility to realize inferences with an unprecedented level of accuracy and details. This success of machine learning models is such that they are now ubiquitous in our society. For instance, machine learning-based systems are now used in banking for assessing the risk associated with loan applications, in hiring systems to assess the quality of an applicant [FTT12] and in predictive justice to quantify the recidivism risk of an inmate [Ele16]. However, the widespread use of machine learning models also raises serious privacy and ethical issues, especially if their predictions are put into action in domains in which they can significantly affect individuals [AGG18].

Privacy issues in machine learning. With respect to privacy, in addition to the inferences that can be made from the data itself, it is also important to understand how much the output of the learning algorithm itself (*e.g.*, the model) leaks information about the input data it was trained on. For instance, new attacks have been recently developed against machine learning models in which the training data can be reconstructed from the model [FLJ+14] either in the white box setting (*i.e.*, in which the description of the model is known) or the black box setting (*i.e.*, in which it is only possible to interact with the model by querying an API with a particular input to receive the associated output). Another possible inference attack against a machine learning model is a membership attack [SSSS17] in which the objective of the adversary is to be able to predict whether the profile of a particular individual (which is known to him) was in the dataset used to train the model. Generally, this inference is deemed problematic if revealing the membership of the profile to this database leads to learning of sensitive information (*e.g.*, that the individual is part of a cohort of patients for a disease). This line of research is still in its infancy, and much work remains to be done, in particular with respect to how to prevent these inference attacks, even if some preliminary protection methods [YGFJ18] have been proposed based on differential privacy [DR+14].

Ethical issues in machine learning. In addition to privacy, the machine learning community has also started recently to investigate ethical issues such as the fairness, accountability and transparency of machine learning models through the organization since 2014, of an annual specialized workshop dedicated to this issue[4] and more recently through the creation of the FAccT conference[5], which follows a highly multidisciplinary approach to address some of the ethical challenges highlighted in the following. Among other things, this work has led to fundamental questions about the ways one can define a fair algorithm, as well as the meaning

[4] http://www.fatml.org
[5] https://facctconference.org

of social justice in an advanced capitalist economy [Raw01]. The European lawmakers have also enacted some measures that asked to provide for more accountability [CR19].

In recent years, several initiatives [FC19][6,7,8,9] were launched to propose design principles and guidelines for the responsible development of artificial intelligence. However, *very few research works have explored the tensions, but also convergences that can emerge when addressing jointly the privacy and ethical challenges when designing and deploying machine learning models.* For example, to be able to audit a machine learning model for potential biases, it is often easier (but not necessarily mandatory) to have access to its structure or at least an approximation, thus highlighting the strong link between transparency and fairness. In addition, some research has proposed to use anonymization methods as a way to enhance fairness as a side effect (*e.g.*, [ABG+21]), thus showing a positive connection between privacy and fairness.

A fundamental open question is to investigate when the achievements of these different objectives results in a positive sum game. Indeed, as shown by recent work [MSDH19], aiming for interpretability can open the door to inference attacks against privacy. In addition, being able to quantify the level of discrimination of a particular machine learning model usually requires the collection of sensitive data, such as the attributes that could lead to discrimination, which is clearly in tension with privacy. Finally, it is possible that, under the excuse of making its models more interpretable and transparent, a company might be tempted to perform fairwashing, which can be defined as promoting the false impression that the models used by the company respect some particular ethical values while it might not be the case. An example of such a risk has been studied in [AAF+19] in which the authors demonstrate that it is possible to use black-box explanation to rationalize the decisions of a predictive model that is particularly discriminating towards a subgroup of the population.

Objective and organization of the report. We believe that to be able to understand how to best address privacy and ethics responsibly when developing machine learning models, we need first to have a clear view of how these concepts interact with each other in a positive as well as negative manner. The objective of our report is precisely to investigate this question by following an interdisciplinary approach at the crossroads of computer science, law and ethics.

The outline of the report is the following. First, in Section 2, we briefly describe the background notions of machine learning before reviewing in Section 3 the main conceptions of privacy that exist in the literature as well as the main privacy attacks that have been developed in recent years against machine learning models before discussing the relationship between privacy and innovation in Section 4. Afterwards, each of the following section will be dedicated to explore and discuss the intersection between privacy and other ethical issues such as Fairness (Section 5), Transparency (Section 6), Accountability (Section 7), Security (Section 8.1) and the Right to Erasure (Section 8.2).

[6]Human Rights in the Age of AI, Access Now (2018).

[7]Ethics Guidelines for Trustworthy AI, High-Level Expert Group on AI set up by the European Commission, 8 April 2019.

[8]Microsoft AI Principles (2018).

[9]Déclaration de Montréal (2018).

For each of this ethical issue, we will discuss first how it has been conceptualized in the field of law and ethics before describing how this issue has been formalized and addressed within the field of machine learning. Afterwards, we will highlight the main convergences and tensions between this ethical issue and privacy within the context of machine learning by trying to draw generic observations as well as possible remediations to solve the tensions. Finally, we discuss in Section 9 the new issue of *ethics washing*, which referred to promoting the false impression that a machine learning model respect some ethical values while it might not be the case before concluding the report in Section 10.

2 Machine learning fundamentals

Artificial intelligence, machine learning and deep learning. *Artificial Intelligence* is usually defined in a broad manner as the area of computer science that develop theories and algorithms with the objective of providing machines with the capacity to simulate the behaviour of human with respect to some tasks. Expert systems that encode the knowledge of experts in the form of if-then-rules, which have been used in the medical domain and algorithms that simulate a player of a particular game based on heuristics for exploring the search space efficiently, are some examples of early research topics in artificial intelligence.

Machine Learning is a subfield of Artificial Intelligence, in which the objective is to give the ability to machine to learn from examples, rather than encoding rules directly. Finally, *Deep Learning* is itself a specific subdomain of Machine Learning that targets the training of learning algorithms and architectures called "deep neural networks". Deep neural networks are a form of neural networks based on many layers (in contrast to a few for early implementations of neural networks) that have lead to important breakthroughs in tasks in domains such as computer vision, natural language processing that were previously considered difficult.

To quote Gary Marcus & Ernest Davis in their book *Rebooting AI* [MD19]: "A handy way to think about the relation between deep learning, machine learning, and AI is this Venn diagram. AI includes machine learning, but also includes, for example, any necessary algorithm or knowledge that is hand-coded or built by traditional programming techniques rather than learned. Machine learning includes any technique that allows a machine to learn from data; deep learning is the best-known of those techniques, but not the only one."

One of the most common definitions of machine learning is due to Tom Mitchell [Mit97]: "A computer program is said to learn from experience E with respect to some class of tasks T, and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E". Generally speaking, one can distinguish three types of tasks T, namely *supervised learning*, *unsupervised learning* and *reinforcement learning* [Mur21].

Supervised learning. Supervised learning aims at learning a function f (*i.e.*, typically a machine learning model) mapping an input x (*e.g.*, the profile of an individual) of a particular feature space X to an output y (*e.g.*, a prediction with respect to the input profile) of another domain Y. The components of a model are often referred to as *parameters* or *weights*, and the

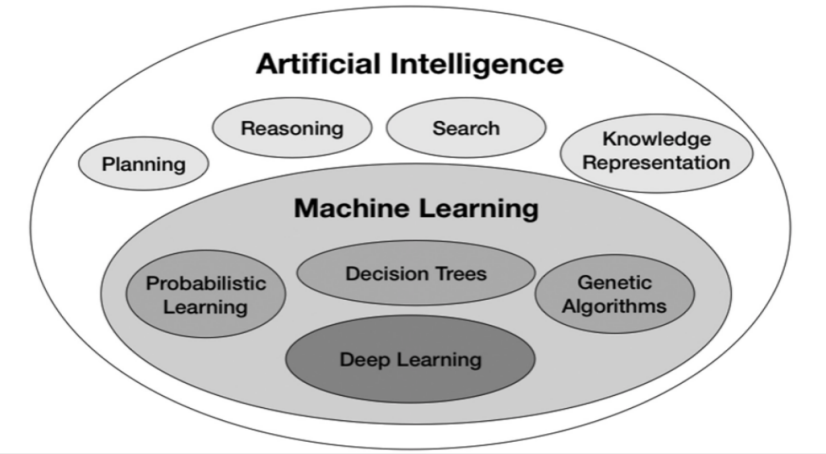


Figure 1: Illustration of artificial intelligence and its subdomains taken from [MD19].

parameters of the algorithm used to find a model are called *hyperparameters*. When Y is a categorical domain, the learning task is called *classification* while if Y is a real-valued domain, it is called a *regression*.

The prediction of whether or not someone applying for a loan at the bank will be able to reimburse it is an example of a (binary) classification task while inferring the risk of recidivism (*e.g.*, on a scale of 1 to 10) of an inmate that is preventively released is a form of regression. For both classification and regression tasks, the associated experience E is a training phase, in which a set of pairs (x, y), hereafter referred to as training data, of inputs x and their associated outputs y –usually referred to as labels– are used to build a model that can reliably predict the output for any data that comes from X.

The performance P of the learning process is usually measured by a training error quantifying how close the predicted output $\hat{y}$ is to the real output y. In addition to having a small training error (which could be obtained, for instance, if the model memorizes the training data), a key property in machine learning is to be able to generalize to new pairs (x, y). *Generalization* is generally estimated by evaluating the accuracy of the model prediction on a previously unseen dataset often called the test set.

Unsupervised and reinforcement learning. Unsupervised learning aims at discovering patterns in unlabelled data (*i.e.,* data samples with no particular outputs/labels associated). Compared to supervised learning, the experience E associated with unsupervised learning can be diverse and includes problems like finding clusters in the data, finding low dimensional representation of the data or simply understanding the data distribution. Finally, reinforcement learning aims at training agents to identify the best sequence of actions to take in a given environment by optimizing a cumulative reward. The experience E associated with such types of tasks consists of iteratively learning to distinguish good actions from bad ones through a

trial-and-error process.

In this report, we will mainly focus on the privacy and ethical issues related to supervised learning.

3 Privacy and Machine Learning

3.1 Conceptions of privacy

Privacy has different meanings for legal scholars, which can be grouped into three main categories. Some refer to privacy as *restricting access*, in which privacy is reduced when more people have access to one's personal information or the same number of people have access to less of one's personal information. Others define privacy as *control*, in which increasing privacy is increasing the level of control over one's information, thus also associated privacy with autonomy. Finally, others view privacy as *contextual integrity*, in which privacy is preserved to the extent that personal data is transferred in compliance with informational norms that exist in society [Nis09]. More importantly, under all three conceptions, privacy is about setting the norms for the collection, use and dissemination of people's personal information. Focusing on disclosure while ignoring collection and use is thus a reductive and, from the point of view of legal scholars, incorrect view of privacy.

In this report, the concept of privacy will be considered in the light of the Canadian context but also with respect to the European Union law, especially the *General Data Protection Regulation* (GDPR). With respect to Canada, we will mainly focus on two federal privacy laws: the *Personal Information Protection and Electronic Documents Act* (PIPEDA), which covers the personal information-handling practices of businesses and the private sector, and the *Privacy Act*, which covers the personal information-handling practices of federal government departments and agencies in the public sector, as well as the constitutional law in the form of informational privacy protected in section 8 of the *Canadian Charter of Rights and Freedoms*.

Finally, our analysis builds on the premise that we need to protect privacy for various reasons [DeC97, vdHBPW20, Vél21]. These include, but are not limited to, preventing harm as unrestricted access to personal data — bank account, social media profile, whereabouts, and so on — can lead to harming people (*e.g.*, profiling, targeted advertisement, identity theft, ...), either intentionally or not. Privacy is also necessary for protecting freedom and autonomy. Indeed, a lack of privacy may expose people to entities that will constraint their choices and bring them to make decisions they would not have otherwise made. Privacy is also needed for securing informational justice and social power. More precisely, people must be able to have some control over their personal information to avoid being misrepresented, to be empowered when dealing with other people, when negotiating contracts, and so on. Finally, privacy is a fundamental requirement for preserving human dignity as a lack of privacy can impose a psychological burden on people but also lead them to make different choices and even lose their sense of themselves. One may think, for instance, of people that are subjected to mass surveillance or any other form of illegitimate surveillance.

3.2 Privacy attacks against machine learning models

Hereafter by the term *adversary*, we will refer to any entity that aims at recovering the personal data of an individual without the latter's explicit consent. In practice, this broad definition can cover a wide class of potential attackers. To realize his objective, the adversary will perform a privacy attack (also called an *inference attack*) against a machine learning model. The vast majority of works that have investigated privacy attacks in machine learning have considered almost exclusively the supervised learning setting and, more precisely, classification tasks. Inference attacks against machine learning models include *membership inference* [SSSS17], *property inference* [AMS+15], *model inversion* [FJR15], *model extraction* [TZJ+16] and *training data reconstruction* [CLE+19]. We will briefly review these families of inference attacks in the following.

Membership inference. Membership inference attacks against machine learning models have been introduced by Shokri, Stronati, Song and Shmatikov [SSSS17]. Consider a data record x and a target model M trained over a training dataset D. A membership inference attack consists of predicting if x belongs to D while only observing the predictions of M. For instance, the authors demonstrated the possibility for an adversary to assess a data subject's presence in a medical dataset (*e.g.*, cancer patients), highlighting the potential privacy damage this type of attack can cause.

Property inference. Property inference attacks involve training a meta-classifier to detect if the target model has a given property P [AMS+15]. For instance, property inference attacks have been used to learn that people from a particular ethnic group produce a speech recognition system's training set or that the data used to train a particular type of network traffic classifier come from a specific type of traffic. Property inference can be viewed as a generalization of membership inference, for which the property to infer is "Does D contain x ?".

Model inversion. Model inversion attacks aim at predicting, given a target model M and an output class y, the sensitive hidden features of inputs x such that $M(x) = y$. As a result, the adversary will learn the average of the inputs that belong to the class y (*i.e.*, an average representative of the class). Initially, model inversion has been used [FJR15] (1) to infer if participants to a survey have admitted to cheating on their partner and (2) to reconstruct people's faces by inverting a facial recognition system.

Reconstruction of the training data. Training data reconstruction attacks are similar to model inversion attacks. However, instead of recovering global characteristics of profiles belonging to a particular class, they aim to reconstruct the original training records. For instance, recent works have shown that one can recover private data such as credit card numbers from language models [CLE+19] and that this risk is exacerbated in larger language models such GPT-2 [CTW+20].

Model extraction. Model extraction attacks aim at inferring, given a target model M and its predictions for a chosen set of inputs, the parameters and/or hyperparameters of the model. When performing model extraction, the objective of the adversary can be either to optimize the accuracy of the model (*e.g.*, if he wants to sell the use of the model afterwards) or its fidelity with respect to the target model (*e.g.*, if the model extraction is a preliminary step before conducting another attack). Tramer, Zhang, Juels, Reiter and Ristenpart [TZJ+16] have demonstrated the effectiveness of such attacks by reconstructing several machine learning models after querying online Machine-Learning-as-a-Service (MAAS) platforms. Wang and Cong [WG18] have also proposed attacks to steal hyperparameters of several machine learning models.

4 Privacy and Innovation

Data access and data collection, which include consumer data and other personal data, are often presented as an important factor for innovation in machine learning or innovation in the technology industry as a whole. This suggests a tension between privacy and innovation, but a careful analysis of the literature advises great caution before endorsing this view, for the protection of privacy may not undermine technological innovation as much as it seems. In the following, we discuss the business model of some technological corporations and the role that private data and advertising plays in that model before explaining the tension, or the lack thereof, between privacy and innovation.

The business models of companies in the technology industry often share a common feature: offering free services to a wide user base in order to collect personal data and convert this resource into a money stream [Gal17, Vél21]. Data monetization can be performed *directly* if the data is sold directly to other parties, or *indirectly* if it is used to generate other products or services [Lan20]. Companies such as Alphabet, the holding company of Google, or Facebook, use the data collected on their users to build finely-grained personal profiles, which are then used to sell targeted advertisements. In addition, the practice of collecting and accumulating private data seems pervasive in the technology industry as a whole [Zub20], as well as other sectors: insurances, finance or banking, the health sector [BvdH15], and so on.

Using private data raises issues of privacy, but it is also believed to be an important vector for business productivity and innovation [noa16, noa19, Ng18]. This claim can be interpreted in at least two different ways with respect to innovation in machine learning. First, the advertisement business is very profitable for some technological corporations, such as Alphabet and Facebook, which are also leaders in developing this technology. In fact, these companies are among the wealthiest companies in the world, both in terms of their market value and their access to liquidity [Mar17]. This provides important economic resources to invest in research and development, and to develop new innovative products.

While it is plausible that access to liquidity helps some technological companies innovate, one may wonder if personal data and privacy-invading advertisements are a necessary part of that equation. Lobbyists or representatives for big technology companies often argue that

strong privacy laws would hurt the advertisement market and curb innovation, but there is sparse empirical evidence to back up this claim [Vél21, Wei19]. For instance, the technological industry sells different kinds of advertising, with some forms of ads not being as privacy-invasive as others. Companies could rely more on context — rather than personal profiles – to target their audience: ads on cars can be shown to people that have searched "car" or visited car websites. This would avoid targeting users based on personal attributes such as gender, age, race or political beliefs. There no reason to believe that more contextual advertising is less effective or less profitable, but this would reduce the need to collect personal data and build consumer profiles.

The second interpretation of the claim that private data is a vector for business innovation suggests that this data is important to train machine learning models. One may think, for instance, of the algorithms that Facebook or Alphabet uses to deliver ads [Fac20], personal vocal assistants such as Apple's Siri or Amazon's Alexa, or conversational robots. In fact, it is often claimed that some companies lead in their specific market precisely because they have access to that data. According to this second interpretation, access to personal data is not a vector of innovation because it increases spending power in research and development, but rather because access to private data is a necessary ingredient for developing new machine learning models.

However, there are also reasons to question these claims. While it is true that personal assistants or recommendation algorithms are challenging to build without access to private data on consumers, this seems to be the exception rather than the rule. Many of the most popular machine learning applications do not require private data collected by business firms. One may think, for instance, of the GPT-3 model developed by the OpenAI corporation that uses deep learning to produce human-like texts. The model was trained on an improved version of the CommonCrawl dataset that comprises a large amount of public web pages [BMR+20]. Additional examples include other natural language processing applications, translation applications such as Google Translate, computer vision and most machine learning models used in the health sector [FDC20].

5 Privacy and Fairness

5.1 Conceptions of fairness

The notion of fairness strongly overlaps with the notion of "equality" and "equity" while still being different from these two other normative ideals in some respects. For instance, the political philosopher John Rawls introduced a theory of social justice labelled "Justice as Fairness" promoting the importance of freedom and equality in society [Raw99]. This theory claims that some inequalities – which possibly includes higher income for the most talented individuals in society – could be fair if they are advantageous to the least well-off. Other contemporary ethical theories can permit some forms of reverse or positive discrimination against groups that have been treated unfairly in the past. In this respect, policies that promote hiring more women or members of visible minority groups can be fair, even if these policies

imply the unequal treatment of some individuals [Gol15].

In the field of law, "fairness" can have different interpretations, such as "loyalty" or "equity". An example of "loyalty" can be given in contractual matters. Fairness is a component of good faith (*bona fide*), enshrined by judges (common law) or the legislator (Civil Code of Québec). This obligation of good faith makes it possible to integrate into contractual life a moral dimension comprising duties of loyalty, collaboration and information. Moreover, it should be noted that in 2016, the French lawmaker created some new "loyalty" obligations for online platforms, which are integrated into the Consumer Code. It mainly consists of providing consumers with fair, clear and transparent information on their services' terms and conditions. The services targeted are those referencing, ranking and dereferencing content, goods or services that rely on algorithmic systems.

Equity is the moderating principle of objective law (laws and administrative regulations) according to which everyone is entitled to fair, equal and reasonable treatment. In certain limited cases, the law makes room for the notion of equity by leaving it to the judge to determine "*ex aequo et bono*" (according to what is fair and good), by setting aside legal rules when he believes that their strict application would have unequal or unreasonable consequences.

It is also important to distinguish between equality and equity. Equality in law means treating all people the same, regardless of their circumstances. The objective of legal equity is to ensure that everyone is treated fairly, equally and reasonably, according to their circumstances. This principle is used when the strict application of (legal) rules would result in unfair consequences to one of the parties. Equity, therefore, allows for the implementation of corrective measures that can lead to affirmative action. Fairness in machine learning interacts with two forms of prohibited discrimination. Direct discrimination happens when a decision model makes a prohibited classification. Indirect discrimination happens when a decision disproportionately disadvantages members of a protected category.

5.2 Fairness in machine learning

The discrimination that can occur due to the use of prediction made by machine learning models deployed in automatic decision systems can result from many causes, such as an error in the measurement process that led to the collection of the training data or an error made by the classifier on particular sub-groups of the population. More often, discrimination arises because the training data is inherently biased for historical and societal reasons and that the classifier learns to reproduce this negative bias. If a dataset possesses a strong bias towards a particular protected group of the population (*e.g.*, an ethnic group or a vulnerable minority) that can easily be detected, a naïve solution would consist in simply removing the sensitive attribute from the training data, thus avoiding *direct discrimination*.

However, indirect discrimination is still possible in this case due to the correlations that exist between the sensitive attribute and other attributes. In particular, some attributes strongly correlated with the sensitive attribute could act as *proxies* even if this attribute is removed from the data. Furthermore, there are many ways [FSV16, SG19] in which (un)fairness can be introduced in machine learning models, from which data-related issues are just a fraction

of them. Algorithmic design choices such as model compression [HCC+19, HMC+20], early stopping techniques [AH20, JZTM20] or how data confidentiality issues are handled [BS19] can all affect machine learning models' performances on different sub-groups of a population.

Several notions of fairness have emerged in recent years in machine learning to quantify and formalize this concept as well as to develop fairness-enhancement methods [Nar18, BHJ+18, VR18, CR18]. Most of these notions can be categorized into three families. First, approaches based on *group fairness* [CV10, Cho17, KMR17, HPS+16, ZVGRG17] aim at producing machine learning models that approximate parity for given statistical measure (*e.g.*, false positive or false negative rates) across a given set of subgroups of the population, defined by the sensitive attributes. Second, the rationale behind techniques implementing *individual fairness* [DHP+12, JKMR16] is that machine learning models should output the same decisions for similar individuals, for a given definition of similarity. For instance, if two profiles are identical, except for a sensitive attribute such as gender or race, the machine learning model should output the same prediction. Finally, fairness-aware methods relying on *causal fairness* [KCP+17, KLRS17, NS18] leverage on causal assumptions to estimate the effects of sensitive attributes (*e.g.*, gender or race) on other attributes as well as to design machine learning models that are constrained to exhibit a tolerable level of discrimination with respect to these sensitive attributes.

5.3 Convergences

Fairness as a fundamental principle of privacy legislation. From a legal point of view, several convergences between fairness and privacy are perceptible. First, personal data shall be processed lawfully, fairly and in a transparent manner in relation to the data subject ("lawfulness, fairness and transparency") (see GDPR, article 5). The principles of fair and transparent processing require that the data subject be informed of the existence of the processing operation and its purposes. The controller should provide the data subject with any further information necessary to ensure fair and transparent processing, taking into account the specific circumstances and context in which the personal data are processed.

Provisions for automated decision systems. There is also a convergence when it comes to evaluating personal aspects relating to a natural person that is based on automated processing. Such automated decision systems are used to make predictions, recommendations or decisions about individuals that could have significant impacts on them.

The GDPR enacts the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention (GDPR, article 22). GDPR also integrates protection against "profiling", which consists of any form of automated processing of personal data evaluating the personal aspects relating to a natural person, in particular, to analyze or predict aspects concerning the data subject's performance at work, economic situation, health, personal preferences

or interests, reliability or behaviour, location or movements, where it produces legal effects concerning him or her or similarly significantly affects him or her. This definition matches the kind of prediction made by machine learning models. In contrast to GDPR, Bill C-11 ignores and does not discuss the issue of profiling.

In the GDPR, decision-making based on such processing, including profiling, is exceptionally allowed when expressly authorized by Union or Member State law to which the controller is subject, including for fraud and tax-evasion monitoring and prevention purposes conducted in accordance with the regulations, standards and recommendations of Union institutions or national oversight bodies and to ensure the security and reliability of a service provided by the controller, or necessary for the entering or performance of a contract between the data subject and a controller, or when the data subject has given his or her explicit consent (GDPR, article 22.2). When these three exceptions apply, there is a tension between fairness and privacy in these particular cases.

Recital 71 of the GDPR gives more details about profiling by machine learning, which confirms the convergence between fairness and privacy:

> *"In order to ensure fair and transparent processing in respect of the data subject, taking into account the specific circumstances and context in which the personal data are processed, the controller should use appropriate mathematical or statistical procedures for the profiling, implement technical and organisational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimised, secure personal data in a manner that takes account of the potential risks involved for the interests and rights of the data subject and that prevents, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or that result in measures having such an effect. Automated decision-making and profiling based on special categories of personal data should be allowed only under specific conditions."*

.

Privacy impact assessment could be extended to other ethical issues. The provisions about data protection impact assessment are also a way of convergence between fairness and privacy. Bill C-11 does not clearly provide an obligation to conduct a data protection impact assessment and only states that "every organization must implement a privacy management program that includes the organization's policies, practices and procedures put in place to fulfill its obligations under this Act, including policies, practices and procedures respecting". In comparison, the GDPR enacts that "where processing operations are likely to result in a high risk to the rights and freedoms of natural persons, the controller should be responsible for the carrying-out of a data protection impact assessment to evaluate, in particular, the origin, nature, particularity and severity of that risk" (GDPR, article 35).

As indicated in the Guidelines of the Article 29 Data Protection Working Party on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679, the reference to "the rights and freedoms" of data subjects primarily concerns the rights to data protection and privacy but may also involve other fundamental rights such as freedom of speech, freedom of thought, freedom of movement, prohibition of discrimination, right to liberty, conscience and religion. The Data Protection Impact Assessment is an Algorithmic Impact Assessment that allows considering some values of fairness, such as equity and non-discrimination in a machine learning system.

5.4 Tensions

Measuring and improving fairness often requires access to sensitive attributes. The first tension with privacy is that implementing fairness requires access, in most cases, to sensitive attributes. Blocking sensitive information is very rarely conducive to fair models. Indeed, to build fair models, it is often necessary to collect sensitive information on individuals such as gender, race and age, as fairness-aware machine learning algorithms require to have this information on a wide range of sensitive data to inspect models [TFVH20, Aga20]. This inevitably introduces a privacy risk when training those models due to the collection of sensitive data [VB17, JKM$^+$19, KGK$^+$18].

Thus, the main challenge is (1) to design fair models without collecting unnecessary sensitive attributes that pose risks for individuals and (2) to prevent the data collected from being disclosed. For example, when training a recidivism prediction application, participation in the training data (*e.g.*, as revealed through a membership inference attack) means that an individual has committed a crime – which is sensitive information. Therefore, fairness and membership privacy are both needed for the use of machine learning to be ethical [CS20].

Privacy and anti-discrimination laws can collide. There is a tension between direct and indirect discrimination in machine learning under anti-discrimination law [Cof19]. For instance, a classifier could illegally disadvantage a protected category, but ignoring data about protected categories can also lead to indirect discrimination. Suppose a machine learning algorithm is "race-blind" (*i.e.*, the race attribute is removed from the data) to avoid direct discrimination. In that case, it may be impossible to determine whether the output is indirectly discriminatory on the basis of race [WMR21]. Using the information on membership of a protected category to treat its members differently could give rise to a direct discrimination challenge (which the law forbids), but just using race in predicting recidivism should not by itself do this, as how the algorithm employs the racial classification matters [Hel20].

This tension forms a paradox in regulating information to prevent algorithmic discrimination: "To avoid disparate treatment, the protected category attributes cannot be considered but to avoid disparate impact, they must be considered." [Cof19]. Thus, sensitive attributes must be considered in the training process to avoid proxy discrimination [PS19]. Otherwise, the challenge of blocking sensitive attributes to prevent discrimination is identifying and blocking an endless list of proxies for legally protected categories, which raises two problems [Cof19].

The first problem is that one may never cease to find attributes that are predictive of each other, and one may not know in advance which those proxies are. The second problem is that those proxies could also contain valuable and legitimate information. Thus, blocking them may reduce accuracy and be self-defeating by reducing the ability to detect bias. Other recent studies have shown that differentially-private models have a larger drop of accuracy on underrepresented subgroups [BS19].

Possible remediations. Complying with the principles of data minimization and purpose limitation of data protection law (even in jurisdictions in which they are not mandated) can reduce these risks and the consequent tension between privacy and fairness. Another method is to have the data stored by a trusted third party [VB17]. However, while this might help appease users in that they are giving their data to a trusted entity in lieu of those that employ the algorithms, they are still sharing this sensitive data. This prompts privacy concerns that models might still be vulnerable to privacy attacks [JKM$^+$19]. Employing differential privacy [DMNS06, DR$^+$14] has been proposed as a mitigation strategy to address this [JKM$^+$19] although others suggest that it is still vulnerable to state-of-the-art inference attacks [JE19].

Jointly achieving privacy and fairness without using sensitive attributes. As we mentioned previously, removing sensitive attributes rarely leads to a reduction in discrimination due to proxies [PS19]. However, researchers have proposed methods for pre-processing data, anonymizing or encrypting sensitive attributes to achieve fairness [KC12, KGK$^+$18, ML20]. The relation of pre-processing methods to privacy can be described as follows: "since both fairness and privacy can be enhanced by removing or obfuscating the sensitive information, with the adversary objective of minimal data distortion" [PS20]. This measure can simultaneously increase privacy and fairness [Cof19].

Another line of work concerns the design of techniques that avoid using sensitive attributes by implementing the Rawlsian Max-Min fairness principle [Raw01], which consists of maximizing the utility for the most disadvantaged group. These techniques can be categorized into two categories depending on how the most disadvantaged group is taken into account without using sensitives attributes, namely, distributionally robust optimization (DRO) [HSNL18] and adversarial reweighted learning (ARL) [LBC$^+$20]. DRO-based techniques implement fairness by minimizing the worst-case training loss over a set of test distributions chosen around the training distribution and aiming to mimic the protected groups one may encounter in reality. On the other hand, ARL-based techniques implement fairness by improving the performances of a model on computationally-identifiable subgroups [HJKRR18]. More precisely, the approach consists of a minimax game between a *learner* that seeks to optimize the training loss and an *adversary* that seeks to find computationally-identifiable regions with high loss and improve the performances of the learner on these regions. Another practical solution to deal with the direct and indirect discrimination paradox explained above is to control a dependent attribute [KGK$^+$18, TTK18].

There are also legal approaches to the paradox, such as data minimization and privacy

by design. Thus, law can help in making algorithms more responsible and less discriminatory, meaning that privacy protection can contribute to the objectives of antidiscrimination law [Cof19, Ish19]. However, an adversary can still infer private information, including the sensitive attribute of the user, from the output of machine learning models trained with fairness-enhancing techniques that are agnostic to the sensitive attribute [JKM+19]. Thus, approaches that do not use the sensitive attributes can be viewed as a form of data minimization and will need to be co-designed with differential privacy to offer protection against privacy inference attacks. For instance, a recent work has achieved this objective by combining secure multiparty computation with differential privacy [JKM+19].

6 Privacy and Transparency

6.1 Conceptions of transparency

The proposed Consumer Privacy Protection Act (Bill C-11, reforming PIPEDA) defines an automated decision system as any technology that assists or replaces the judgement of human decision-makers using techniques such as rules-based systems, regression analysis, predictive analytics, machine learning, deep learning or neural nets. Regarding the risks and issues for data protection, specific rules have to be enacted. Bill C-11 states that "If the organization has used an automated decision system to make a prediction, recommendation or decision about the individual, the organization must, on request by the individual, provide them with an explanation of the prediction, recommendation or decision and of how the personal information that was used to make the prediction, recommendation or decision was obtained" (article 63(2)).

The GDPR also requires that suitable safeguards are enacted in many cases, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision. These safeguards show that the aim of the legislator is to preserve a certain level of data protection, even when the use of machine learning is not fair but legitimized by certain objectives. However, in this case, both transparency and privacy are called into question, and there is finally no tension between these two values. The tension is located between these two values taken together and the objectives of general interest.

6.2 Transparency in machine learning

Techniques to promote transparency in machine learning can be broadly categorized into three families. In particular, *transparent-box design* techniques promotes the design of inherently interpretable models (*e.g.*, linear models, decision trees, rule lists or rule sets) while *post-hoc explanation* techniques [GMR+18] aim at explaining the outcomes of black-box models (*e.g.*, deep neural networks or ensemble methods) through an interpretable surrogate model. In addition, *hybrid approaches* [Wan19, PWH20] promote the co-training of black-box models and their interpretable substitutes. The main objective of hybrid approaches is to help decision-

makers select the model to be used for a prediction task in full consideration of the different alternatives' performance trade-offs (*i.e.*, interpretable as well as black-box models).

Transparent box design focuses on the problem of designing models that are human-understandable. The notion of understandability is often expressed through criteria such as simulatability, decomposability and algorithmic transparency [Lip18]. In a nutshell, simulatability corresponds to the possibility for a model to be described syntactically and contemplated at once by a human, while composability refers to the fact that all the components of the model (*e.g.*, attributes, antecedents of rule lists or weights of linear models) are understandable by a human. Finally, algorithmic transparency can be defined by the ability to prove the correctness of the training process (*e.g.*, convergence to a unique well-behaving model).

Post-hoc explanation aims at explaining how black-box ML models produce their outcomes through different forms of explanations [GMR$^+$18, ADRDS$^+$20]. In particular, current post-hoc explanation techniques include global explanations, local explanations, example-based explanations and gradient-based attribution techniques. Global explanations explain the complete logic of the black-box model by training a surrogate model that is interpretable by design (*e.g.*, linear models, rule-based models or decision trees) while maximizing its fidelity with respect to the predictions of the black-box. Local explanations only aim at explaining a single decision by approximating the black-box in the vicinity of the input profile through an interpretable model. Example-based explanations produce particular data points to explain either the black-box model's behaviour or its training data distribution. Finally, gradient-based attribution techniques leverage the inputs' gradients to provide a relevance score of the features with respect to the values outputted.

6.3 Convergence and tension

Transparency is a fundamental requirement of privacy legislation. From a legal point of view, a convergence exists between transparency and privacy regarding the fact that any personal data processing should be transparent to individuals with respect to how personal data concerning them are collected, used, consulted or otherwise processed and to what extent the personal data are or will be processed [Cof21, CS19]. The principle of transparency requires that any information and communication relating to the processing of those personal data be easily accessible and easy to understand and that clear and plain language be used. Especially, individuals should be made aware of risks, rules, safeguards and rights in relation to the processing of personal data and how to exercise their rights in relation to such processing, including machine learning. The principle of transparency means that transparent information, communication and modalities for the exercise of the rights of the data subject should be fulfilled. The convergence is perfect as transparency is the preferred method to guarantee the respect of privacy rights of individuals.

As for AI to manipulate or influence individuals, we can note that the proposed regulation on AI published in April 2021 by the European Commission provides specific rules of transparency regarding certain manipulation processes such as "deep fakes". Article 52(3) states that: "Users of an AI system that generates or manipulates image, audio or video content that appreciably

resembles existing persons, objects, places or other entities or events and would falsely appear to a person to be authentic or truthful ('deep fake'), shall disclose that the content has been artificially generated or manipulated".

Moreover, some AI system are prohibited. Article 5(1) prohibits: "The placing on the market, putting into service or use of an AI system that deploys subliminal techniques beyond a person's consciousness in order to materially distort a person's behavior in a manner that causes or is likely to cause that person or another person physical or psychological harm".

Explanations provide additional information that privacy attacks can exploit. Providing explanations necessarily reveal more information than simply outputting the prediction of the model. In particular, black-box models have been shown to be vulnerable to inference attacks exploiting post-hoc explanations. In this setting, the adversary has access to both the predictions and the explanations of the target black-box model. For instance, recent works have shown that gradient-based explanation can be used to perform high-fidelity model extraction on two-layer neural networks [MSDH19] or that explanation techniques can be leveraged to perform membership inference as well as near-complete dataset reconstruction[SSZ19]. Further, Aïvodji, Bolot and Gambs [ABG20] have demonstrated that counterfactual explanations can be used to perform model agnostic and high-fidelity model extraction under low budget in terms of queries to the black-box.

This vulnerability of black-box models to inference attacks is often connected to *over-fitting*, which describes a situation in which the error of a model on its training is significantly lower than its error on the test set. This corresponds to a situation in which the model has memorized examples of the training set but is not able to generalize to new instances such as ones from the test set. Most membership inference attacks exploit this phenomenon to verify if a particular profile belongs to a target model's training set by observing the confidence in its prediction. Intuitively, models designed using transparent box approaches might also seem vulnerable because their descriptions are more likely to be public information. However, in practice, they are less prone to privacy inference attacks than complex black-box models, usually because these models' simplicity makes them less able to memorize examples.

6.4 Case study 1: Model reconstruction from counterfactual explanations

In this section, we present a concrete example of tension between privacy and explainability in which an adversary leverage counterfactual explanations to perform a model extraction attack [ABG20].

6.4.1 Model extraction

As briefly introduced in Section 3.2, a model extraction attack is an privacy attack in which an adversary $\mathcal{A}$ obtains a surrogate model $\mathcal{S}_{\mathcal{A}}$ that is similar to the targeted model $\mathcal{B}$. The precise meaning of the *similarity* depends on the adversary's objective, while the success of the attack depends on the adversary's capabilities.

Adversary objective. Previous works [ASJ+19, JCB+20] have considered two main categories of model extraction attacks depending on the goal of the adversary, namely *accuracy-based* and *fidelity-based* model extraction attacks. In accuracy-based model extraction attacks, also known a *theft-motivated model extraction* attacks [JCB+20], the adversary aims at learning a surrogate model $\mathcal{S}_\mathcal{A}$ whose accuracy is as close as possible to that of the target's model $\mathcal{B}$. Typically here, model extraction provides a financial benefit to the adversary as he can use the surrogate model as a substitute for the commercial API of his target. In fidelity-based model extraction attacks, also known as *reconnaissance-motivated model extraction* attacks [JCB+20], the objective of the adversary is to build a surrogate model $\mathcal{S}_\mathcal{A}$ maximizing the *fidelity* with the target's model $\mathcal{B}$. The fidelity $\text{Fid}(\mathcal{S}_\mathcal{A})$ of the surrogate is defined as its accuracy relative to $\mathcal{B}$ over a reference set $X_r \subset \mathcal{X}$ [CS96]:

$$\text{Fid}(\mathcal{S}_\mathcal{A}) = \frac{1}{|X_r|} \sum_{x \in X_r} \mathbb{I}(\mathcal{S}_\mathcal{A}(x) = \mathcal{B}(x)). \tag{1}$$

In this context, a model extraction attack is often a first step towards mounting other attacks such as a model inversion attacks [FLJ+14, FJR15] or adversarial examples discovery [SZS+13, GSS14, PMG+17].

A particular case of fidelity-based model extraction attack, known as *functionally equivalent extraction*, occurs when the adversary is able to build a surrogate $\mathcal{S}_\mathcal{A}$ matching the predictions of the target's model $\mathcal{B}$ over the whole input space (*i.e.*, $\forall x \in \mathcal{X}, \mathcal{S}_\mathcal{A}(x) = \mathcal{B}(x)$). As pointed out in [JCB+20], functionally equivalent extraction attacks require model-specific techniques. In contrast, both accuracy-based and fidelity-based model extraction attacks generally rely on the flexibility of learning-based approaches, making them more generic. In the latter case, the target's model $\mathcal{B}$ is used as a labeling oracle by the adversary.

Adversary capabilities. Following the taxonomy introduced in [JCB+20], we describe the adversary capabilities around three axes, namely the *domain knowledge*, the *deployment knowledge* and the *model access*. Domain knowledge corresponds to the adversary's prior information on the task of the target model. For learning-based approaches, a common assumption is that the adversary knows as much about the task as the designer of the target model. Deployment knowledge refers to the adversary's knowledge of the target model's characteristics (*e.g.*, architecture, training dataset, training algorithm, hyperparameters, ...). Finally, the model access indicates how the adversary interacts with the target's model and the form of information extracted from these interactions. More precisely, this includes both the number of queries the adversary is allowed to make to the target's model and the type of the model's output (*e.g.*, labels, probabilities, gradients, counterfactual explanations,...).

6.4.2 Counterfactual explanation

Counterfactual explanations [WMR17, LLS+17, LLM+17, TSHL17, GCV+18, Rus19, USL19, JKV+19, PBK20, MST20, KBBV20] are data instances that are close to the input instance to be explained but whose model predictions are different from that of the input instance. More precisely, given a black-box model $\mathcal{B}$, an original input x_0, its predicted outcome $y_0 = \mathcal{B}(x_0)$

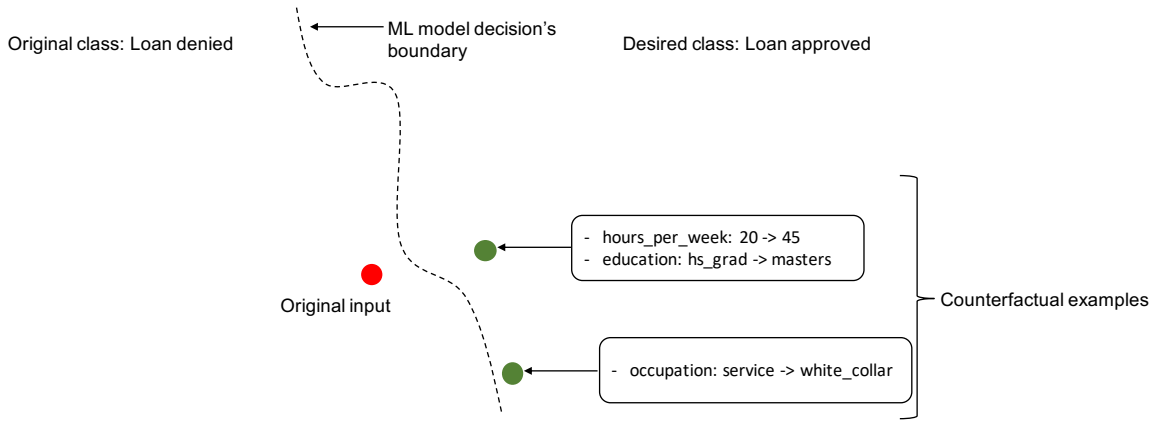


Figure 2: Illustration of a counterfactual explanation scenario. Given an original instance for which the model predicts the *loan denied* class, a counterfactual explanation framework provides different instances that are close to the original one but belong to the desired class (*loan approved* here). An individual asking for an explanation can thus see which aspects of his profile he may try to change to yield the desired outcome.

and a desired outcome $y \neq y_0$, a counterfactual explanation $c(x_0)$ for the input x_0 is usually obtained by solving the following optimization problem:

$$c(x_0) = \underset{c}{\operatorname{argmin}}\ L(\mathcal{B}(c), y) + |c - x_0|, \tag{2}$$

in which $L(\mathcal{B}(c), y)$ ensures that the obtained counterfactual $c(x_0)$ has a different prediction from that of the original input x_0 while the second term $|c - x_0|$ helps in obtaining a counterfactual close to the original instance. Figure 2 illustrates a counterfactual explanation scenario while Table 1 provides concrete examples of counterfactual explanations obtained on a real world dataset, namely Adult Income [FA10].

	Age	**Workclass**	**Education**	**Marital status**	**Relationship**	**Occupation**	**Race**	**Gender**	**Capital gain**	**Capital loss**	**Hours per week**
Original input (outcome: $\leq$ 50K)	33	Private	Assoc-acdm	Married	Own-child	Professional	White	Female	0	0	40
Counterfactuals (outcome: $>$ 50K)	-	-	Doctorate	-	-	-	-	-	33703	-	39
	-	-	-	-	-	White-collar	-	-	99985	4333	-

Table 1: Examples of counterfactuals obtained on Adult Income [FA10] dataset. The task is to predict whether an individual earns more than 50,000$ per year. The top row corresponds to the different features of the input instance. The second row depicts the data instance to be explained as well as its original outcome. Finally, the last two rows are examples of counterfactuals generated to explain the original input. Dashed marks refer to features that are unchanged.

Diverse counterfactuals. To be more actionable, counterfactual explanation frameworks often generate for each input instance, several counterfactuals covering a diverse range of

possibilities instead of the single closest one [WMR17]. Providing diverse counterfactuals allows users to decide the most efficient way by which they can influence their profile to obtain the desired outcome. At the same time, on the privacy side, it also leaks more information to the adversary and enables him to mount a more powerful attack. In this study, we rely on the *DiCE* framework [MST20] to implement the explanation API of the target models. Nonetheless, the proposed attack is generic enough to work with any counterfactual explanation framework.

In a nutshell, DiCE aims to find valid and actionable counterfactual examples by solving the following optimization problem:

$$\begin{aligned} C(x_0) = \quad & \underset{c_1,\ldots,c_k}{\operatorname{argmin}} \ \tfrac{1}{k}\Sigma_{i=1}^{k} L(\mathcal{B}(c_i), y) + \tfrac{\lambda_1}{k}\Sigma_{i=1}^{k}|c_i - x_0| \\ & -\lambda_2 \operatorname{dpp\_diversity}(c_1, \ldots, c_k), \end{aligned} \tag{3}$$

in which $\mathcal{B}$ is the black-box model, x_0 is the original input to be explained, $y \neq \mathcal{B}(x_0)$ is the desired outcome, c_i is a counterfactual example and k is the number of counterfactuals to return. The loss function $L(\mathcal{B}(c_i), y)$ ensures that each of the counterfactuals has a different outcome than that of the original input x_0 while $|c_i - x_0|$ leads to the counterfactual being close to the original input. Finally, dpp_diversity$(\cdot)$ is the diversity metric while $\lambda_1 \in \mathbb{R}^+$ and $\lambda_2 \in \mathbb{R}^+$ are the hyperparameters used to balance the proximity and diversity. More precisely, the larger λ_1 is, the closer the counterfactuals will be to to the query instance. Similarly, the larger λ_2 is, the more diverse the counterfactuals return will be diverse.

Hereafter, we will investigate the success rate of explanation-based model reconstruction with both single and diverse counterfactuals.

6.4.3 Model extraction from counterfactual explanations

In this section, we first frame the generic problem of explanation-based model extraction before presenting the particular case of counterfactual explanation, which is our focus. Afterwards, we describe the different adversarial models investigated in our work before describing their corresponding model extraction attacks.

Problem formulation. As illustrated in Figure 3, in an explanation-based model extraction attack, the adversary leverages both the predictions and the explanations of the target model to build the surrogate model.

Definition 1 (Explanation-based model extraction) *Given a target model $\mathcal{B}$, its prediction API $\mathcal{B}(\cdot)$ as well as its explanation API $\mathcal{E}(\cdot)$, both available in a black-box setting, a set of data points $x_1, \ldots, x_n$, the explanation-based extraction attack consists in using both the explanations and the predictions of the target model to build a surrogate $S_{\mathcal{A}} \approx \mathcal{B}$, using an attack process $\psi(\cdot)$.*

In the particular context of counterfactual explanations, the explanation API $\mathcal{E}(\cdot)$ returns for each data point x_i its corresponding counterfactual explanation $c(x_i)$ along with its associated outcome $\overline{y_i}$. In the case of diverse counterfactuals, the explanation API will return a set $C(x_i)$ of counterfactual examples instead of a single one.

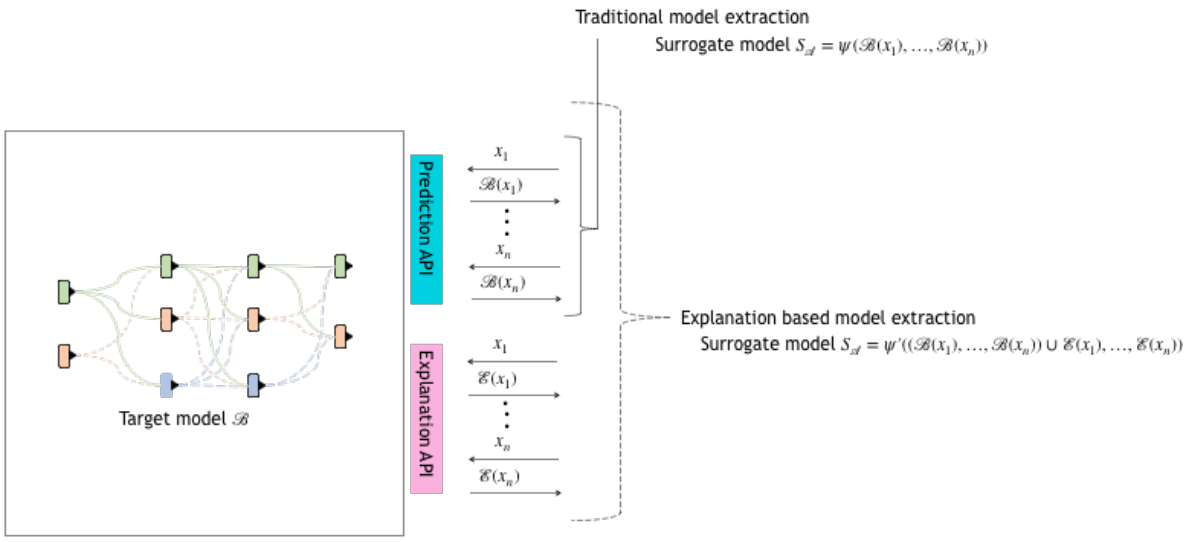


Figure 3: Illustration of a traditional model extraction attack and an explanation-based model extraction. In the former, the adversary relies on the predictions $\mathcal{B}(x_1), \ldots, \mathcal{B}(x_n)$ of the target model $\mathcal{B}$ to build the surrogate model $S_{\mathcal{A}}$ using a process $\psi(\cdot)$, while in the later, the adversary combines the predictions $\mathcal{B}(x_1), \ldots, \mathcal{B}(x_n)$ and the explanations $\mathcal{E}(x_1), \ldots, \mathcal{E}(x_n)$ of the target model $\mathcal{B}$ to generate the surrogate $S_{\mathcal{A}}$ using another process $\psi'(\cdot)$.

6.4.4 Attack description

Adversary model. We are interested in a fidelity-based extraction attack (also called reconnaissance-motivated extraction attack) in which the adversary $\mathcal{A}$ will rely on both the predictions and the counterfactual explanations of the target model to conduct his attack. Similarly to [JCB+20], we will assume that the adversary knows as much about the task as the designer of the target model in terms of domain knowledge. As for the model access, the adversary will have black-box access to the target model's predictions and counterfactual explanations. We also assume a bound on the number of queries that the adversary is allowed to make. Each query to the explanation API $\mathcal{E}(\cdot)$ returns one or more counterfactual explanations depending on the diversity criteria. Finally, for the deployment knowledge, we consider different scenarios according to (1) the knowledge of the training data distribution, which can be *known*, *partially known* (*e.g.*, knowledge of the marginal distribution) or *unknown*, (2) the knowledge of the target model architecture (*known* or *unknown*) and (3) the use of the training data by the explanation algorithm (*used* or *unused*).

Attack strategy. To conduct his attack, the adversary first builds his *attack set* $D_{\mathcal{A}}$ according to his knowledge of the distribution of the target model's training data. Then, for each data point $x \in D_{\mathcal{A}}$, he sends a query to both the prediction API $\mathcal{B}(\cdot)$ and the explanation API $\mathcal{E}(\cdot)$ of the target model. Finally, $\mathcal{A}$ trains the surrogate model $S_{\mathcal{A}}$ according to his knowledge of the target model's architecture, by using a *transfer set* $\mathcal{T}_{\mathcal{A}} = \{D_{\mathcal{A}}, \mathcal{B}(D_{\mathcal{A}})\} \cup \mathcal{E}(D_{\mathcal{A}})$ consisting

of both the outputs of the prediction and explanation APIs.

In traditional model extraction attacks, the transfer set $\mathcal{T}_{\mathcal{A}}$ of the adversary can be imbalanced due to the unequal distribution of classes within the dataset. As a result, there may be a significant difference between the class-based accuracy of the surrogate model $S_{\mathcal{A}}$ and the target model $\mathcal{B}$ [ASJ+19]. In contrast, counterfactual explanations-based model extractions attack do not suffer from such limitations as the attack set is balanced by construction since each instance is followed by its corresponding counterfactual explanation.

6.4.5 Experimental setting

In this section, we report on the performances of counterfactual explanations-based model extraction attacks when evaluated on real datasets.

Datasets. We have conducted our experiments on three public datasets that are extensively used in the *FaccT* (Fairness, Accountability, and Transparency) literature, namely *Adult Income* [FA10], *COMPAS* [ALMK16] and *Default Credit* [FA10].

- In a nutshell, the Adult Income dataset contains information about individuals collected from the 1994 U.S. census. The dataset contains 48,842 individuals, each described by 11 attributes. The underlying classification task is to predict whether or not an individual makes more than 50,000$ per year in terms of income.

- The COMPAS dataset gathers records from criminal offenders in Florida during 2013 and 2014. The dataset contains 7,214 individuals, each described by 8 attributes. The classification task considered is to predict whether a subject will re-offend within two years after being released.

- Finally, the Default Credit dataset is composed of information on Taiwanese credit card users. The dataset contains 29,986 individuals, each described by 23 attributes, while the classification task is to predict whether a user will default in his payments.

Evaluation metrics. Our main objective is to conduct a reconnaissance-motivated model extraction attack. As such, we will use the fidelity metric as our primary evaluation metric for the success of the attack. Nonetheless, we will also report on the accuracy of the surrogate.

Black-box models. Each dataset is split into three subsets, namely the *training sets* (67%), the *testing sets* (16.5%) and the *attack pools* (16.5%). The black-box models are learned on the training sets. The testing sets are used to evaluate (1) the accuracy of both black-box models and surrogates models and (2) the fidelity of the surrogate model relative to the target black-box model. The attack pools are used only for the scenario in which the adversary is assumed to know the data distribution. For both Adult Income and COMPAS, the target models are Multi-Layer Perceptrons (MLPs) with two hidden layers, with respectively 75 and 50 neurons. For Default Credit, the target model is a MLP with one hidden layer of 50 neurons.

For all the three target models, we have used the L1 regularization (with $\lambda = 0.001$), the RMSprop optimizer [TH12], the rectifier activation function (*ReLu*) for hidden layers, the

Sigmoid activation function for output layers and train the models for 100 epochs. Table 2 summarizes the accuracy of the three black-box models on their training and test sets.

Dataset	**Training Set**	**Test Set**
Adult Income	85.36	84.70
COMPAS	69.00	66.30
Default Credit	81.10	80.70

Table 2: Performances of the black-box models. Columns report the accuracy of the black-box models on their training set and test set.

Scenarios investigated. We consider five different counterfactual-based model extraction scenarios, namely (S1) single counterfactual with known training data distribution, (S2) single counterfactual with partially known training data distribution, (S3) single counterfactual with unknown training data distribution, (S4) multiple counterfactuals with known training data distribution and (S5) impact of the proximity and diversity metrics on the performances of the model extraction. The first three scenarios are variants of the same setting in which the explanation API only provides a single counterfactual explanation per query, but under different assumptions on the adversary knowledge on the distribution of the training data of the target model. The objective of the last two scenarios is to study the impact on the success rate of the extraction attack of having access to multiple and diverse counterfactual explanations per query.

For all five scenarios, the performances are evaluated according to the adversary's knowledge on the architecture of the target model and whether or not the explanation API uses the training data. When the adversary does not know the target model's architecture, we imagine that typically the adversary will have a trial-and-error strategy in which different architectures will be tried with the one maximizing the fidelity of the surrogate being kept at the end. In our experiments, we simulate this situation with an adversary that tries 5 different architectures, which we describe in Table 3. Remark that since the surrogate training is done offline once the transfer set has been built, the adversary is only limited in its exploration by its computational resources and the time he is willing to dedicate to this exploration. In particular, if he has the sufficient resources, he might even use advanced techniques for exploring the space of possible architectures such as *Neural Architecture Search* [EMH19] to maximize the fidelity of the surrogate model.

Hereafter, we detail each of the five scenarios.

(S1) Single counterfactual with known training data distribution. In this scenario, the adversary directly uses the attack pool as his attack set $D_{\mathcal{A}}$. More precisely, he selects a subset $Q_{\mathcal{A}}$ of $D_{\mathcal{A}}$ to query the target model and construct his transfer set $\mathcal{T}_{\mathcal{A}} = \{Q_{\mathcal{A}}, \mathcal{B}(Q_{\mathcal{A}})\} \cup \mathcal{E}(Q_{\mathcal{A}})$. In the experiments conducted, we have considered different values $|Q_{\mathcal{A}}| \in \{100, 250, 500, 1000\}$ for the number of queries to study its effect on the attack's performance. For each value of $|Q_{\mathcal{A}}|$, the experiment is repeated over 10 random sampling

of $Q_{\mathcal{A}}$ and the average fidelity and accuracy of the surrogate are reported. Additionally, we compare the performances of the surrogate with a baseline model trained using the complete attack pool $D_{\mathcal{A}}$ and the predictions $\mathcal{B}(D_{\mathcal{A}})$ of the target model.

(S2) Single counterfactual with partially known training data distribution. Here, the adversary is assumed to know the marginal distribution of the attributes of the training set. To perform his attack, in this scenario, the adversary builds an attack set $D_{\mathcal{A}}$, composed of data points sampled according to the marginal distribution of the attributes. The rest of the attack is similar to the process described above for (S1).

(S3) Single counterfactual with unknown training data distribution. This scenario is similar to (S2) except that the distribution of the training data of the target model is unknown. As a consequence, the attack set $D_{\mathcal{A}}$ is generated simply by uniformly sampling data points from the input space. Clearly, this can sometimes lead to the generation of unrealistic data points.

(S4) Multiple counterfactuals with known training data distribution. In this scenario, the same configuration used in (S1) is considered, but the number k of counterfactuals provided by the explanation API is increased. More precisely, the attack performances are studied for k in the range $\{3, 5, 7\}$. For each of these settings, the default values for the proximity and diversity hyperparameters are used (*i.e.*, $\lambda_1 = 0.5$ and $\lambda_2 = 1.0$).

(S5) Impact of the proximity and diversity on the performances of the model extraction. In this scenario, the impact of proximity and diversity on the surrogate model's performance is explored. For the sake of simplicity, we focus on the setting in which the adversary knows the data distribution and the training data is used by the explanation API since the results are similar in both cases. We set $|Q_{\mathcal{A}}| = 1000$, $k = 5$, $\lambda_1 \in \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$ and $\lambda_2 \in \{1.0, 1.5, 2.0, 2.5, 3.0\}$.

All our experiments were run on an Intel Core i7 (2.90 GHz, 16GB of RAM) laptop.

	Hidden layers	**Hidden activation**	**Output activation**	**Loss**	**Optimizer**	**Regularizer**	**Epochs**
Arch 1	$100, 50$	ReLu	Sigmoid	Binary cross-entropy	RMSprop	$L_1(0.001)$	100
Arch 2	$100, 50$	ReLu	Sigmoid	Binary cross-entropy	Adam	$L_1(0.01)$	20
Arch 3	$200, 100, 50, 25$	ReLu	Sigmoid	Binary cross-entropy	RMSprop	$L_1(0.01)$	20
Arch 4	$200, 100, 50, 25$	ReLu	Sigmoid	Binary cross-entropy	Adam	$L_1(0.01)$	20
Arch 5	$100, 75, 50$	ReLu	Sigmoid	Binary cross-entropy	RMSprop	$L_1(0.001)$	100
Arch 6	$100, 75, 50$	ReLu	Sigmoid	Binary cross-entropy	Adam	$L_1(0.01)$	20

Table 3: Architectures of the models used across the experiments. For both Adult Income and COMPAS datasets, we use **Arch 5** as the target model architecture, the adversary uses the remaining architectures as candidate architectures when the target model architecture is unknown. For the Default Credit dataset, **Arch 1** is used as the target model architecture and the remaining when the target model architecture is unknown.

Dataset	Target model Architecture	Training data used by $\mathcal{E}(\cdot)$	100 Queries	250 Queries	500 Queries	1000 Queries	Baseline Model
Adult Income	known	yes	89.02/81.05	92.06/82.94	93.21/83.26	94.22/83.68	81.28/76.06
		no	89.39/81.47	91.78/82.87	92.17/82.74	94.84/83.88	
	unknown	yes	89.27/81.11	92.42/83.18	93.62/83.52	94.65/83.88	81.28/76.06
		no	88.99/81.24	92.21/83.05	93.40/83.28	94.89/83.97	
COMPAS	known	yes	87.13/66.19	91.29/67.30	92.17/67.11	92.85/66.97	71.42/61.09
		no	87.91/65.81	89.57/65.86	92.62/66.49	93.92/66.50	
	unknown	yes	88.13/66.50	90.81/67.26	92.00/67.16	92.36/66.90	85.04/64.95
		no	89.12/66.16	90.08/66.03	92.91/66.66	93.43/66.49	
Default Credit	known	yes	97.09/80.22	97.93/80.55	98.31/80.63	98.57/80.52	88.52/77.86
		no	97.15/80.20	97.77/80.34	97.77/80.34	98.28/80.48	
	unknown	yes	97.08/80.12	97.99/80.57	98.39/80.58	98.39/80.58	88.52/77.86
		no	96.90/80.15	97.52/80.38	97.90/80.4	98.03/80.43	

Table 4: Performances (fidelity/accuracy) of the model extraction attack in scenario (S1) for Adult Income, COMPAS, and Default Credit datasets. For each of the query scenarios, we report on the performances (averaged over 10 extraction attacks) of the surrogate model. The column of the baseline model correspond to the fidelity/accuracy of the surrogate model obtained using the whole attack pool $D_{\mathcal{A}}$ to conduct a traditional model extraction attack.

6.4.6 Experimental results

(S1) Single counterfactual with known training data distribution. Table 4 summarizes the results obtained for scenario (S1). The attack is evaluated on Adult Income, COMPAS and Default Credit datasets. Overall, for all these three datasets, we observe that with only 250 queries, our attack reaches a fidelity of 90%. This fidelity is higher than that of the baseline, which is a traditional model extraction attack with 8059 queries for Adult Income, 1192 queries for COMPAS and 4948 queries for Default Credit. We also observed that as the number of queries increases, both the fidelity and the accuracy of the surrogate also improve. With only 1000 queries, our attack already reaches a fidelity of 94% on Adult Income, 93% on COMPAS and 98% on Default Credit and an accuracy matching that of the target model (as measured on its test set) on all three datasets. Moreover, an interesting finding of our study is that the knowledge of the target model architecture and the use of the training data by the explanation API does not lead to a significant advantage with respect to the attack's success.

(S2) Single counterfactual with partially known training data distribution. Table 5 displays the results obtained for scenario (S2). Here, for the sake of simplicity, we have only performed the experiments on the Adult Income dataset. The results demonstrate that an adversary who only knows the features' marginal distribution can still perform a powerful model extraction attack. In particular, with 1000 queries, the surrogate model $S_{\mathcal{A}}$ still reaches a fidelity of 93% and an accuracy close to that of the target model on the test set.

(S3) Single counterfactual with unknown training data distribution. Table 6 describes the performance of our attack for scenario (S3). Similarly to (S2), we focus on the Adult Income dataset. Overall, the results show that even without knowing the data.

Dataset	Target model Architecture	Training data used by $\mathcal{E}(\cdot)$	100 Queries	250 Queries	500 Queries	1000 Queries
Adult Income	known	yes	86.19/79.47	89.05/81.37	91.70/82.84	92.95/83.30
		no	86.48/79.82	89.54/81.77	91.74/82.84	92.60/83.20
	unknown	yes	86.22/79.46	90.01/81.84	92.14/83.12	92.97/83.4
		no	86.22/79.83	90.02/81.94	92.13/83.09	93.54/83.65

Table 5: Performances (fidelity/accuracy) of the model extraction attack in scenario (S2) for Adult Income. For each of the query scenarios, we report on the performances (averaged over 10 extraction attacks) of the surrogate model.

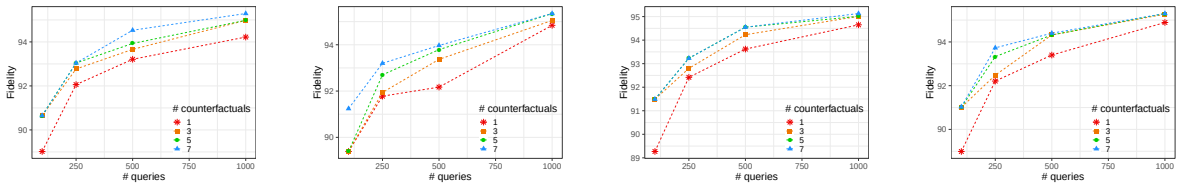
Dataset	Target model Architecture	Training data used by $C(\cdot)$	100 Queries	250 Queries	500 Queries	1000 Queries
Adult Income	known	yes	82.30/75.90	83.28/77.12	84.46/78.25	85.06/78.58
		no	82.31/76.11	82.63/76.78	83.74/77.57	83.74/77.57
	unknown	yes	81.98/75.48	84.31/77.38	85.75/78.79	85.75/78.79
		no	81.58/75.59	83.37/77.15	84.60/78.28	84.61/78.25

Table 6: Performances (fidelity/accuracy) of the model extraction attack in scenario (S3) for Adult Income. For each of the query scenarios, we report on the performances (averaged over 10 extraction attacks) of the surrogate model.

distribution, the adversary can build a surrogate model performing better than the one obtained using a traditional extraction attack with 8× more labels and with full knowledge of the data distribution. However, compared to the fidelity of counterfactual-based extraction attacks with partial knowledge (respectively full knowledge) of the data distribution, the surrogate's fidelity decreases by 7.79% (respectively 9.14%).

(S4) Multiple counterfactuals with known training data distribution. Figure 4 describes the impact of the number of counterfactuals provided for each query on the performance of the extraction attack. Overall, we can observe that the fidelity of the surrogate improves as the number of counterfactuals increases. Besides, the performances of the surrogate model when the adversary does not use the architecture of the target model (Figures 4c and 4d) are slightly better than the performances of the surrogates trained using the same architecture as the target model (Figures 4a and 4b). These results also corroborate our previous findings that the target model architecture's knowledge does not provide a significant advantage to the adversary. Note that if the training data is used by the explanation API, this seems to give the adversary a small advantage in the lower query budget regime ($|Q_{\mathcal{A}}| \leq 500$). However, in higher query budget regimes ($|Q_{\mathcal{A}}| > 500$), it does not provide a significant advantage to the adversary.

(S5) Impact of proximity and diversity on the performance of the model extraction attack. Figure 5 summarizes the results obtained for scenario (S5) on the Adult



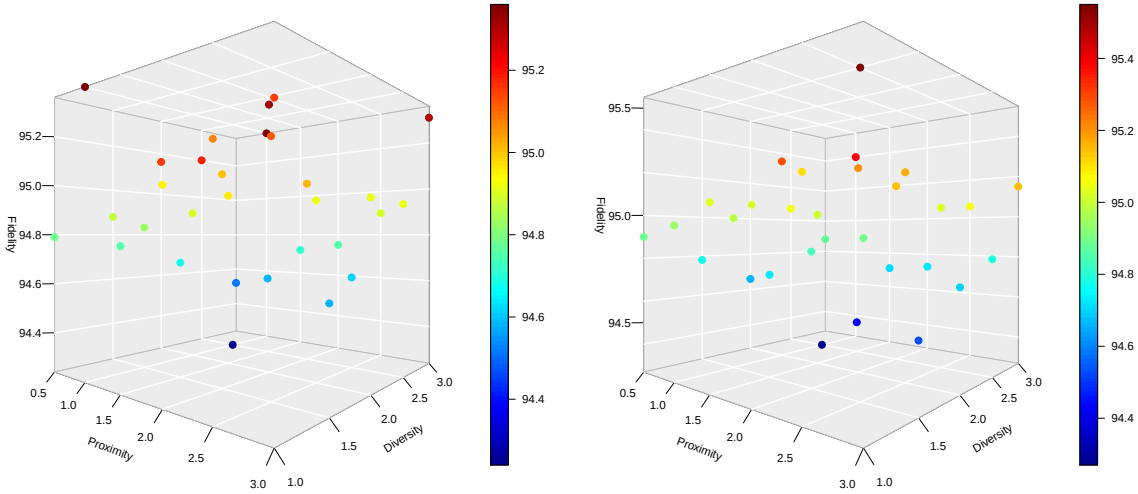
(a) Arch. known, training data used by $\mathcal{E}(\cdot)$

(b) Arch. known, training data not used by $\mathcal{E}(\cdot)$

(c) Arch. unknown, training data used by $\mathcal{E}(\cdot)$

(d) Arch unknown, training data not used by $\mathcal{E}(\cdot)$

Figure 4: Performances (*i.e.*, fidelity) of the model extraction attack in scenario (S4) for Adult Income. Results demonstrate the impact of the number of counterfactual explanations per query on the extraction attack's fidelity.



(a) Architecture known

(b) Architecture unknown

Figure 5: Performances (fidelity) of the model extraction attack in scenario (S5) for Adult Income dataset. The results show the impact of the proximity and the diversity metrics on the fidelity of the surrogate.

Income dataset. Overall, the higher we set the constraints, the more likely the surrogate found will be of high fidelity. Similar to our previous observations, the knowledge of the target model architecture does not provide a significant advantage.

Summary of the results. Consistently across the experiments, we have observed that counterfactual explanations can be leveraged by an adversary with a limited query budget to perform high-fidelity and high-accuracy model extractions. In particular, when the adversary has

partial or complete knowledge of the data distribution, he can obtain a high-fidelity and a high-accuracy surrogate with only 500 queries. In contrast, when the data distribution is unknown, the surrogate performances are lower as expected. However, even in this restricted setting, the surrogate obtained with our attack still performs better than the surrogate generated using traditional model extraction attacks with full knowledge of the data distribution. Additionally, experiments with multiple and diverse counterfactuals demonstrate that this requirement leads to better performances of the model extraction attacks.

7 Privacy and Accountability

There is a high level of convergence and a few tensions between policies and organizational practices that would enforce privacy and improve accountability in machine learning. More accountability seems generally desirable, but it is still unclear what are the best policies or organizational practices to held people more accountable when they develop or use machine learning applications, or how accountability should be specified.

7.1 Conceptions of accountability

Accountability is often used as an umbrella term for various features or desirable aspects of our social arrangements, including transparency, responsibility, answerability, attributability, and the proper auditing and sanctioning of algorithmic decision-makers [Sho11, Smi12, Esh14]. However, there is often a lack of distinctions between those concepts. We define accountability, or the state of being accountable, as follows: there is accountability when an individual or a group of individuals A provides a justification or an explanation about O — a political decision, a policy, the functioning of a new product or service, etc. — to another individual or group B. The individual or group B that will receive the explanation must have some sanctioning power over A in the sense that B must be able to impose a punishment or corrective actions if it considers the account to be unsatisfactory [Bin18, Bov10, Mul00].

This definition is inclusive and captures many types of accountability that can be specified differently [Wie, Wie]. For instance, various groups of individuals can be considered accountable to different forums: an elected representative may be accountable to its constituents, a corporation may be accountable to its shareholder, and so on. People may also have different expectations about O, the object of accountability and different types of sanctioning power may be established for B. Typical mechanisms of accountability include elections, independent review processes, an ombudsman position within an organization or a public agency [Adl, Gil] and various internal supervisory procedures [QM].

7.2 Accountability in machine learning

Many authors have pointed out the fact that the accountability mechanisms and legal standards that govern decision processes in machine learning have not kept pace with AI technology [AC18, Bin18, KBF$^+$16]. To name a few examples, Reisman, Schultz, Crawford and Whittaker [RSCW]

claim in a popular report that "public agencies urgently need a practical framework to assess automated decision systems and to ensure public accountability". The Fairness, Accountability, and Transparency (FAccT) movement gained momentum in the machine learning community given the challenges of ensuring non-discrimination with AI applications, as well as due process, and understandability in decision-making [DFA$^+$, DF]. Calls for regulation and auditing mechanisms, for standards, codes of conduct, certifications, and data protection impact assessments for algorithmic accountability in organizations, for the development of technical methods to ensure explainability as a tool to achieve accountability have grown significantly in the past years in diverse streams of work.

Various questions or challenges need to be addressed before there is more accountability in machine learning. These questions are both conceptual and practical. For instance, who should be made more accountable: the organization that produce or use machine learning, or the legislators that should introduce new regulations? In addition, as current tools available to courts and legislators were mainly designed to oversee human decisions, is there a need for additional legal provisions to deal with new technological developments?

There is also an issue with the burden of proof when decisions are made by machine learning models. Indeed, an ethical challenge related to accountability is the "concept of distributed moral responsibility" [Flo16] or "diluted responsibility" [Cha17] or, stated in other words, the problem of the multiple hands. As Floridi [Flo16] said it: "[T]oo often "distributed" turns into "diffused": everybody's problem becomes nobody's responsibility".

Another set of questions concerns the type of forum that should receive the accounts from private and public organizations, the type of explanation that should be provided, and the proper sanctioning power for this forum. In the absence of any control or sanction mechanisms and/or regulatory body directly entitled to ensure accountability in the field of machine learning, in the absence of any kind of right to explanation, and in the absence of any ethical and/or legal basis that would be required to justify demands for explanations about algorithmic decision-making, on which basis people affected by algorithmic decisions could claim for justifications?

Accountability for content moderation purposes. The difficulty with AI for content moderation purposes is that these kinds of tools are implemented by companies like Google or Facebook whose processes are not transparent. The criteria for moderation are not known and, as more and more governments seek to make them responsible for illegal content, the tendency is to over-moderate content, (sometimes where the content is only "inappropriate") at the expense of freedom of expression. There are also tendencies to delete content from certain communities or opinions and there is a total lack of transparency and evidence on the subject. Moreover, even if several governments try to make these companies more accountable, the "moderation subject" is not universal as the freedom of speech is differently protected and limited, especially in the US when the Big Tech are American companies.

7.3 Convergences

Legal obligations of providing accountability under privacy legislations. From a legal approach, privacy and accountability may converge if one considers that privacy laws may themselves incorporate accountability obligations. Therefore, compliance with such legislation, such as the GDPR, implies documenting the respect of privacy rules, and this obligation to document can be added to similar obligations stemming from other legislation, such as banking legislation, for example. Thus, the GDPR integrates accountability tools, such as the data controller's obligation to carry out an assessment of the impact of the envisaged processing operations on the protection of personal data (Data Protection Impact Assessment). Another example is that each controller shall maintain a record of processing activities under its responsibility. Moreover, compliance with accountability obligations can be done while respecting privacy legislation, as it is most often not required to reveal personal data.

Another challenging question related to accountability concerns the increasing use of privacy-invasive inferences and thus, the justification for the use of inferences. Wachter and Mittelstadt [WM19] argue that machine learning draws non-intuitive and unverifiable inferences and predictions about the behaviors, preferences, sensitive attributes, and private lives of people, and that these inferences create new opportunities for discriminatory, biased, and privacy-invasive profiling and decision-making. To close the various accountability gaps that exist with new AI technologies, and to promote justification of inferences, they suggest a "new right to reasonable inferences". This new right would be "applicable to 'high risk' inferences that cause damage to privacy or reputation, or have low verifiability in the sense of being predictive or opinion-based while being used for important decisions. This right would require ex-ante justification to be given by the data controller to establish whether an inference is reasonable".

According to Wachter and Mittelstadt [WM19], this disclosure would address "(1) why certain data are normatively acceptable bases to draw inferences; (2) why these inferences are normatively acceptable and relevant for the chosen processing purpose or type of automated decision; and (3) whether the data and methods used to draw the inferences are accurate and statistically reliable. An ex-post mechanism would allow data subjects to challenge unreasonable inferences, which can support challenges against automated decisions exercised under Article 22(3) of the GDPR".

This type of disclosure appears to be an interesting step towards the respect of individual and group privacy as de-identification is still imperfect. Indeed, de-identified data can still be identifiable in the absence of regulations or technical safeguards to prevent re-identification of de-identified information. Not only de-identified data can reveal identifiable individual information but can also reveal harmful trends and information about members of groups if for example those de-identified data when processed by an AI system is used in a discriminatory manner, even unintentionally. Thus, the regulation of de-identified data appears essential for preventing AI bias [Pau17, JDHF19, Cof19].

However, while this "right to reasonable inferences" is an interesting concept in theory, it appears to be difficult to implement in a legal text and in practice. Indeed in law, vague

notions such as "reasonable" are subject to interpretation and this will be the case here even more so since inferences are not always visible and traceable in the absence of transparency. What should we consider as "reasonable", seems difficult to pinpoint *in abstracto* as it depends of the context. Thus, this notion leaves too much room for uncertainty in AI inferences to encourage its consecration in law.

Moreover, the concrete implementation of this proposal is difficult to imagine in the absence of real capacities to control and verify that the inferences are not "unreasonable". Most of the inferences are not known to the persons concerned. Consequently, it would be difficult to say whether or not it is "reasonable". It would be more useful to enshrine a right not to suffer negative consequences of inferences or decision-making by algorithms, unless there is an explanation and justification by a human being of the unfavorable decision thus made.

Accountability and public reason. From an ethical perspective, some convergences between privacy and accountability can be pointed out, such as the possibility that accountability approaches can help in demonstrating that a company has respected a particular privacy requirement. In addition, in response to the lack of the accountability mechanisms and legal standards that could govern decision processes in machine learning, Binns [Bin18] argues that the notion of public reason – in brief, the idea that rules, decisions, and outcomes need to be justifiable by common principles – might reasonably fill this gap. This argument echoes one of the two concepts of accountability defined by Bovens [Bov10]: that accountability as a virtue (a normative concept or a set of standards) as well as as a mechanism (a descriptive concept seen as a social mechanism ensuring one can be held accountable).

Both concepts are important in democracy but for different reasons. Accountability as a virtue is important because it provides legitimacy, while accountability as a mechanism is instrumental in achieving accountable governance, and thus contributes to the legitimacy of public governance. Therefore, Binns [Bin18] normative argument drawing from political philosophy stipulates that the notion of public reason "is an answer to the problem of reasonable pluralism in the context of algorithmic decision making". In other words, if "decision-maker" and "decision-subject" disagree over the adequacy of the justifications provided for a decision made from the use of personal data, that conflict could be resolved by referring to the grand principles of the public reason.

Connections between transparency, accountability and privacy. As Binns [Bin18] puts it, a potential challenge of his model of accountability is raised by the opacity of some algorithmic decision-making systems that may lead to "algocracy", in which "the legitimacy of public decision-making processes" is thwarted by "the opacity of certain algocratic governance systems" [Dan16]: "Considering the wide range of machine learning methods available, there are trade-offs to be made between interpretability and accuracy ". In addition to privacy, this also highlights the strong connection between transparency and accountability in the sense that the former is often a prerequisite for the latter.

On this topic, other authors endorse a stronger position such as Ananny and Crawford [AC18]

who argued that "if a system is so complex that even those with total views into it are unable to describe its failures and successes, then accountability models might focus on the whether the system . . . should be built at all". However, as Binns [Bin18] put it, accounting for a system can be more than explaining its outputs. Indeed, sometimes what matters will not be how a system arrived at a certain output, but what goals it is supposed to serve, which inputs were involved in, etc. This is the reason why a crucial element of accountability, in terms of privacy, is design. Machine learning models should be designed for privacy by implementing appropriate technical and organizational measures prior to and during all phases of collection and processing in such a way of ensuring the respect of privacy.

Practical implementation of accountability contributing to privacy. For accountability to really enhance privacy protection and meet expectations, some authors called for the need to translate it not only in general principles but into practical measures that take into account its plural dimensions [Ben95, BLM15]. Thus, Butin and Le Métayer [BLM15] developed a systematic approach covering the entire life cycle of personal data ("end-to-end accountability"), considering three types of accountability: accountability of policy (*i.e.*, a given organization should be able to prove it has a clear and defined privacy policy), accountability of procedures (*i.e.*, a given organization should be able to demonstrate that its procedures are sufficient for the implementation of its privacy policy) and accountability of practice (*i.e.*, a given organization should be able to demonstrate its privacy policies have effectively been met. As stated by Butin and Le Métayer: "Roughly speaking, the first type of accountability is purely declarative and provides at best a form of legal guarantee (binding commitment); the second type adds guarantees at the organizational level but only the third type can deliver the full promises of accountability". Despite the lack of – and above – legal accountability mechanisms, this end-to-end accountability, also referred to as "accountability by design" may prove to ensure better privacy protection related to machine learning models using personal data.

Privacy by design should definitively be part of the solution, but as argued by Guagnin, Hempel and Ilten, "it cannot provide for the discourse that is missing in the current regime. Both regimes – the legal paradigm regime and a possible technological paradigm regime, black-box rules and logics and conceal them behind complex artefacts – impervious to the public. We argue that this is exactly what needs to change: we envision a regime governed by a "discourse paradigm". This is where accountability comes into play" [GHI12].

7.4 Tensions

Disclosure of private information with heightened transparency. An individual that is held to account must be more transparent about its practices. The main tension between privacy and accountability comes from the risk that heightened transparency could lead to the disclosure of private information. Consider a case wherein a company would use machine learning to screen job applicants and would be held to account about the non-discriminatory nature of its selection process. The company may want to provide data about the applicants

profiles to show there is no bias, but this data may allow to identify some candidates. One impact of having more accountability in this case would be the disclosure of personal information on specific job applicants. An obligation of account-giving may also place an organization in a difficult position regarding the protection of non-personal information that must not be made public, such as a trade-secrecy, proprietary technologies or other internal information.

In an analysis of the limitations of transparency, Mike Ananny and Kate Crawford [AC18] suggest that it can do great harm if implemented without a notion of why some part of a system should be revealed: it may "threaten privacy and inhibit honest conversation" and "expose vulnerable individuals or groups to intimidation by powerful and potentially malevolent authorities".

However, the risks of disclosure with more accountability should not be exaggerated, for a careful specification of the type of accountability that is expected can often mitigate these risks. First, there are different types of accountability depending on the forums that is responsible for receiving an account [Bov, Wie]. In particular, *political accountability* corresponds to the situation in which the civil servants in a public organization have to account for their practices to political superiors. In this case, the obligation of account-giving is the consequence of the delegation of power from citizens to their political representatives. *Legal accountability* typically involves the legal obligations of an agent and a judges or jury will act as a forum. There is also a wide range of quasi-legal forums that exercise independent administrative and financial supervision or control, and these would create a form of *Administrative accountability*. *Professional accountability* concerns the relations between a professional and his peer group or professional association. Finally, there are more direct accountability relations between public or private agencies and their clients, citizens, or even civil society, which we may refer to (slightly differently than Boven's definition) as *social accountability*.

Political and social types of accountability are more public in nature given the type of forum and the delegation of power that are involved. These forums may also involve larger group of individuals, which make it even more difficult to contain or control the information that they receive. If proper account-giving is likely to involve sensitive information, then it may be preferable to favour other types of accountability, such as administrative, professional or legal accountability. With administrative accountability, a forum may deliberate behind close doors, control the information that is made public or at least insure that there is no violation of privacy. It is also easier for the member of these forums to comply with various forms of non-disclosure agreements, which can provided additional guarantees.

One option to protect against the disclosure of private information comes from the type of forum of accountability that is involved, another option is to specify the type of justification or explanation that ought to be provided. In the field of explainable AI (XAI), as least four approaches to make algorithms intelligible can be outlined: (1) *explaining the model* behind an algorithm, (2) *explaining the outcomes* of an algorithmic decision or process, (3) *inspecting the black box* and (4) *creating a transparent box model* [GMR$^+$18, And].

If there is a risk that account-giving will violate privacy in making private information accessible, as that was the case with the algorithm for selecting job applicants mentioned above, then an option is to favour the approaches (1), (3) or (4). Indeed, these approaches focus either

on the global logic of a system, inspecting the inner functioning of a system or making a system open or transparent. Therefore, it is easier to engage in these forms of accountability without disclosing real inputs or outputs. In some cases, this may be sufficient to protect personal data about people if this is part of the information flow of the system.

However, the types of explanation that are more likely to protect personal information, usually also shed light on the inner functioning of an algorithm, which may create a risk for protecting trade secrecy or information about internal processes. In this situation, another option may be to favour approach (2) mentioned above: typically, providing an account on an algorithms based on its outcome makes it easier to protect sensitive information about its inner functioning. Due to this reason, some organizations will favour black-box explanations based on inputs and outputs when attempting to provide an account about some of their algorithms (see also the previous section on transparency).

Accumulation of private information. There is also an inherent tension between privacy and the need to record a lot of information (*e.g.*, in the form of logs) to provide an account of one's practices. The first tension discussed above concerns the risk of disclosure to external parties, but even if these risks are mitigated, more accountability may lead to the accumulation of more information that could be used internally even if it is not disclosed externally. Other risks include data breaches, where external — or even internal — parties would access internal data illicitly.

Although these risks are real, they can also be prevented in different ways. First, public and private organizations should have good data management practices whether they are being held to account or not. This includes proper security measures to prevent unwanted access to any internal data, but also short and long term data management plans. An organization should not accumulate data needlessly and it should destroy this data when is it not needed anymore. Therefore, it is not clear that making organizations more accountable would necessarily lead to more data accumulation, at least not if it is properly managed.

8 Privacy and Data Protection

8.1 Privacy and Security

Security in machine learning. The security of a system is often analyzed through the lens of the CIA (Confidentiality, Integrity, Availability) model [PP12]. Confidentiality aims at protecting the content of private information (*e.g.*, email or file) from being disclosed to an unauthorized entity while integrity ensures that the behaviour of the system or data that it stores will not be altered by the adversary. Finally, availability is the property that the system remains accessible to its legitimate users, even under extreme circumstances such as a natural disaster or a distributed-denial-of-service (DDoS) attack.

In addition to the privacy attacks discussed in Section 3, machine learning models are also vulnerable to security attacks targeting confidentiality, integrity or availability. In particular, model extraction attacks aiming for accuracy can be viewed as attacks targeting confidentiality.

When performing such an attack, the adversary seeks to obtain a high-accuracy surrogate that can be used instead of the commercial API of the target model. In contrast, model extraction attacks optimizing fidelity as the key metric are classified as privacy attacks as they allow the adversary to learn information (*e.g.*, the parameters or structure of the model) that can be used as the first steps to mount more powerful attacks.

In the field of adversarial learning, attacks targeting integrity include adversarial examples [BCM+13, SZS+13], which aims to alter a machine learning model's behaviour at inference time, and data poisoning [NBC+08, JOB+18], which occur at training time to influence the structure of the model learned. Finally, sponge examples [SZB+20] are the only documented attacks to date that target the availability of the machine learning model. In such an attack, the adversary crafts input data that have the property to increase both the target model's inference time and energy consumption.

Convergence : Security as a necessary condition for privacy. From a legal perspective, the convergence between privacy and security is obvious, as personal data should be processed in a manner that ensures appropriate security and confidentiality of the personal data, including for preventing unauthorized access to or use of personal data and the equipment used for the processing. This obligation to provide security means it is very important to protect privacy and, more broadly, fundamental rights. If not addressed in an appropriate and timely manner, a personal data breach may result in physical, material or non-material damage to individuals such as loss of control over their personal data or limitation of their rights, discrimination, identity theft or fraud, financial loss, unauthorized reversal of pseudonymization, damage to reputation, loss of confidentiality of personal data protected by professional secrecy or any other significant economic or social disadvantage to the individual concerned. For this reason, in case of breach of security, some safeguards must be implemented, such as the notification of the breach to the individual whose data has been stolen in order to mitigate the harm or risk of harm to the individual that could result from the breach.

Tension : Defense mechanisms addressing security risks can increase the success of privacy attacks. In machine learning, works in security and privacy are often done separately. As a result, the interaction between privacy and security is often not considered in existing defense techniques to mitigate security risks. For instance, Song, Shokri and Mittal [SSM19] have shown that defense techniques designed to protect machine learning models against adversarial examples [BCM+13, SZS+13] also make them more vulnerable to membership inference [SSSS17].

8.2 Privacy and Right to Erasure

Convergence : Implementing the right to erasure in machine learning. *Machine unlearning* [CY15, GGVZ19, BCCC+19] is the process by which a machine model M trained with a dataset D that contains the record x of a particular individual, can comply with the "Right to erasure" of the latter by removing x from D as well as its contribution to the

parameters of the model M. One approach to perform machine unlearning is the so-called SISA (Sharded, Isolated, Sliced and Aggregated training) technique, in which the model designer first partitions the original training dataset D into n disjoint subsets $D_1, D_2, \ldots, D_n$, before training M as an ensemble of n sub-models $M_1, M_2, \ldots, M_n$, using the partitions. Since x only contributes to one sub-model, implementing the right to erasure only involves retraining a small part of the overall system.

Tension between privacy and the right to erasure. From a legal perspective, a tension appears between privacy and the rights of freedom of expression and information regarding the right of the data subject to obtain from the controller the erasure of personal data without undue delay in specific circumstances. This can happen, for instance, if the personal data is no longer necessary in relation to the purposes for which it was collected or otherwise processed. However, by exception, this right of erasure shall not apply to the extent that processing is necessary for exercising the right of freedom of expression and information (see GDPR, article 17). One of the challenges for machine learning is to sort through the data as soon as an exception applies.

Another ethical issues with machine learning systems making predictions about an individual's future based only on their past data, is that such system assumes no change in individual or societal behavior. In particular, if there is a past error, the future is considered with an unfavorable prediction. From a legal point of view, if we assume a predictive tool for recidivism based on the past, an individual who has already had difficulties with the justice system will be badly rated, especially if he has been sentenced to a prison term. This contravenes the principles of the right to be forgotten when sentences have been served, as well as the right to social reintegration.

Tension between machine unlearning and privacy. A recent work [CZW+20] has shown that satisfying the right to erasure of a user through machine unlearning can cause the machine learning model to become vulnerable to membership inference. More precisely, the attack proposed by the authors exploits the difference in the predictions of the original model and the unlearned model. The attack runs in two phases. During the first phase, the adversary collects the prediction and the confidence value of the original model for a given target user, while during the second phase, the same query is made but on the unlearned model. The information collected is then used to train a meta-classifier that can distinguish between members and non-members of the model's original training set.

Another issue to consider is that given the computational cost of machine unlearning and the impact it might have on the performance of a machine learning model, a model designer can be tempted to organize the records in the partitions by grouping data subjects that are more likely to claim their right to erasure together. From this perspective, the right to erasure can be seen as a catalyst to anti-data minimization practices, in which the model designer will collect more data to infer users that are more likely to require machine unlearning.

9 Privacy and Ethics Washing

Another important issue that can arise in the development of ethically-aligned machine learning is *ethics washing*, which corresponds to promoting the false impression of respecting ethical values while it might not be the case [Wag18, WD19, YHP19, Gre19, Flo19]. The fact that private or public organizations can use marketing or communication techniques to promote a positive perception of their practices regarding sensitive ethical matters in society is not a new issue per se. For instance, public corporations often engage in what is commonly referred to as "greenwashing", which promotes an inaccurate perception of their commitment to protecting the environment. Within the big tech industry, the application of similar as well as new techniques to promote a positive ethical perception of machine learning applications can be observed, leading to ethics washing.

In this section, our main objective is to define how the notion of ethics washing applies to the context of machine learning and to explore the associated technical, legal and ethical issues, especially for ensuring proper accountability in the development and use of machine learning applications. More precisely, we will first review different forms of ethics washing in machine learning, illustrating them with examples. For instance, we will discuss (1) the concept of privacy washing that arises in ML when a model producer falsely claims that its model protects the privacy of the users, (2) the risk of metrics' cherry-picking in algorithmic fairness, and (3) how fairwashing can arise when post-hoc explanation techniques are used to hide the harmful behaviour of black-box models. Then, we analyze why ethics washing is precisely an ethical issue from an ethical perspective. Afterwards, we summarize why the plethora of available ethical guidelines are not enough to adequately address ethics washing and why regulations will ultimately be needed. Finally, we discuss two possible approaches that could be explored to limit the possibility of conducting ethics washing and to improve accountability of ML models.

Ethics washing in machine learning. The concept of *ethics washing* has been coined by Wagner and Delacroix to characterize strategies that can prevent or influence legal regulations [WD19]. In this work, we define *ethics washing in machine learning* as a claim with respect to an ethical aspect of the ML model whereby a corporation or another organization (1) promotes a positive ethical perception of a practice, product, aim, policy, ... related to the ML model to the public, a governmental agency or another individual within the group and (2) such that this perception is inaccurate in the sense that the real practice of the company is less ethical than what is being promoted. This definition is very similar to what has been discussed by [Wag18] and [Met19]. The idea of ethics washing is also close from what [Haw12] refers to as "ethical chic" and what David Vogel calls the "market for virtues" [Vog07]. While ethics washing is often an exercise of public relations [Bie20], it does not always have to be so. For instance, this can happen if an organization's message aims its own members, as opposed to people outside the group. One may think, for example, of the positive image that Google and big technological corporations want to promote to the public and governmental agencies, but also their own employees. Hereafter, we describe some plausible manifestations of ethics washing in machine learning in different contexts.

9.1 Privacy washing

Differential privacy [DMNS06] is a privacy model proposed by Cynthia Dwork and collaborators that aims at preventing the inferences that can be performed with respect to individuals whose records are stored on a dataset by limiting the influence of any record on the output of a computation performed on this dataset. In contrast to other privacy models such as *k-anonymity*, *l-diversity* or *t-closeness* [Swe02, ADJM07, LLV07], the guarantees provided by differential privacy on the information leakage holds regardless of the attacker's background knowledge.

The level of protection provided by a differentially-private mechanism is usually a function of a parameter ϵ. More precisely, the smaller the parameter ϵ, the better the protection offered by the mechanism. Thus, differential privacy enables to formally and easily evaluate *privacy-utility* trade-offs by observing how a particular utility measure evolves when changing the value of ϵ. However, when differential privacy is incorporated in the training of a ML model, a given value of ϵ can have different meanings. For instance, [JE19] have observed a significant gap between the theoretical privacy loss of differentially-private ML algorithms and the effective success rate of privacy inference attacks, such as membership inference attacks [SSSS17] and attribute inference attacks [YGFJ18].

In the context of machine learning, **privacy washing** can be achieved by simply advertising (1) the use of differential privacy without specifying neither the differentially-private mechanism nor the value of the parameter ϵ or (2) the use of a particular privacy mechanism with parameter ϵ that provides meaningless privacy guarantees. Concretes examples of privacy washing by big techs have been mainly observed in the context of data analytics. For instance, Apple has been criticized for irresponsible and potentially insecure implementation of differential privacy [TKB$^+$17, Gre17].

While currently differential privacy has been mainly used in practice for performing simple analytics tasks in a privacy-preserving manner, in the future it is possible that a company might be tempted to use it for the training of ML models that it will publicly release[10]. In such a case, the value of the privacy parameter will be crucial to determine the real protection provided to individual and to avoid privacy washing.

9.2 Fairness washing

Given a black-box model that exhibits discrimination against a particular subgroup of the population according to a specific fairness definition, a simple way to perform ethical washing, which we coined as **fairness metrics' cherry-picking**, is to choose from the myriad of fairness metrics, another metric that would seem not to discriminate against the considered group. There are currently many different ways to define and quantify the fairness of a ML model, and there is no consensus in the community on which fairness notion should be used for a particular prediction task. In addition, some of these fairness notions are incompatible in the

[10]While most of the big tech companies usually do not share the data of their users due to privacy and confidentiality reasons, on the other hand, it is common for them in the machine learning context to share the by-products of their analysis such as ML models

sense that optimizing a particular fairness metric can have a detrimental effect on another one. In particular, requiring that a particular ML model exhibits group fairness is likely to prevent the possibility of reaching individual fairness and *vice-versa*.

This lack of consensus on what fairness definition is appropriate in a particular context opens the door for the black-box model producer to choose the fairness metric that is the most favorable to him. For instance, he could confidently claim that his model is not discriminating while it is the case according to the original metric. Real-world examples of such practice include the cases of automated hiring systems such as Pymetrics and HireVue that explicitly claim, while using only demographic parity, that their black-box hiring systems implement bias discovery and mitigation [SMDE20].

9.3 Explainability and fairwashing

Post-hoc explanation techniques are often plebiscited as a way to achieve accountable machine learning. However, recently a growing body of research has exposed the issue that these techniques can be arbitrarily manipulated to gain the trust of the users. For example, the explanations can give the impression that the black-box models exhibit non-discriminatory patterns while it might not be the case. In particular, most of the explanation methods are flexible in the manner that they explain a black-box model in the sense that they are usually many different and diverse explanations, all with similar fidelity with respect to the predictions of the ML model, that can be generated. This allows to manipulate the explanations to perform ethics washing. In Table 7, we summarize recent works that have investigated the risk that post-hoc explanation techniques can be intentionally used to fool users.

Concept	Reference	Agnostic	Global	Local	Feature importance	Example	Visualization	Text
Fairwashing	[AAF+19]	✓	✓	✓	×	×	×	✓
Stealthily Biased Sampling	[FHM19]	✓	✓	×	×	✓	×	×
Misleading Saliency Maps	[HJM19] [DAA+19] [APD+20]	×	×	✓	✓	×	✓	×
Unjustified Counterfactual Explanations	[LLM+19]	✓	×	✓	×	✓	×	×
Public Relations attack	[MT19]	✓	×	✓	×	×	×	×
Scaffolding	[SHJ+19]	✓	×	✓	×	×	×	×
Misleading black-box explanations	[LB19]	✓	✓	×	×	×	×	✓

Table 7: Summary of works related to ethics washing in post-hoc explainability

For instance, Aïvodji, Arai, Fortineau, Gambs, Hara and Tapp raised the awareness of the risk of *fairwashing* [AAF+19] through **global and local explanations' manipulation**, which is the possibility that post-hoc explanation techniques could be used to provide cover for unfair black-box ML models. They coined the process by which this fraud can be performed as *rationalization* and devised `LaundryML`, an algorithm that can systematically rationalize black-box models' decision through global or local explanations. Given access to an unfair black-box model $\mathcal{B}$, `LaundryML`, produces an ensemble of interpretable surrogate models that are fairer than $\mathcal{B}$ according to a predefined notion of fairness. To realize this, the objective function of the learning algorithm is modified to minimize the misclassification error while

improving fairness. The learning algorithm searches over the space of potential interpretable models and enumerates fair yet faithful ones. These interpretable models can then be used to under-report the degree of the unfairness of the original black-box model.

[LB19] have also investigated the possibility that black-box models can be explained with high fidelity by global interpretable models whose features are very different from that of the black-box and look innocuous. More precisely, they propose a technique to generate misleading explanations by maximizing the fidelity of the interpretable surrogate models while favouring models composed of features that users believe are appropriate, in contrast to models whose features can be considered problematic by users (*e.g.*, because they could be considered discriminatory).

Slack, Hilgard, Jia, Singh and Lakkaraju have demonstrated that variants of LIME [RSG16] and SHAP [LL17], two popular post-hoc local explanation techniques, can be manipulated to underestimate the unfairness of black-box ML models [SHJ+19]. Following the same line of work, Merrer and Tredan have demonstrated that a malicious model producer can always craft a fake local explanation to hide the use of discriminatory features [MT19]. In addition, they showed that it is impossible to detect such manipulation if the user can only make a limited number of queries to the model as the detection requires an exhaustive search of the input space, which is unlikely to be feasible for most real-world datasets.

Laugel, Lesot, Marsala, Renard and Detyniecki have highlighted a risk related to the use of counterfactual explanation [LLM+19], which is a form of **example-based explanations' manipulation**. Counterfactual explanations are usually used to help individuals to understand how they can modify attributes of their profiles that are under their control to change the outcome of the machine learning model by showing instance matching that criteria [WMR17]. To formalize the notion of manipulation in this context, the authors defined the concept of justification relative to counterfactual explanations. Given a black-box model $\mathcal{B}$ and a counterfactual c_e produced to explain a decision of $\mathcal{B}$, c_e is justified if it lies within a cluster (*i.e.*, an ensemble of instances) in which there exists an actual instance of the training set of $\mathcal{B}$. This means that the counterfactual is actually back-up by ground-truth data. Using this criterion, the authors have demonstrated that the risk of unjustified counterfactual is actually high in post-hoc counterfactual explanations.

In the same context of example-based explanations' manipulation, Fukuchi, Hara and Maehara have also introduced the risk of *stealthily biased sampling* [FHM19], which occurs when a model producer explains the behaviour of its black-box model by sampling a subset S of its training dataset D. In this setting, a dishonest model producer can sample S in such a way that (1) S is fairer than D, for a given definition of fairness, and (2) it is hard to distinguish the distribution of S from the underlying distribution P of D. The authors also prove the hardness of detecting this fraud by showing that it would be hard for the most powerful detector (*i.e.*, one that has access to the underlying distribution P) to distinguish S from P with a Kolmogorov–Smirnov test, which is a classical test in statistics to assess whether two distributions are identical or not [MJ51].

Visualization-based explanation techniques can be also be manipulated, as shown by recent work on **saliency map based explanations' manipulation** [HJM19, DAA+19]. More

precisely, Heo, Joo and Moon have shown that these types of explanations are vulnerable to the so-called *adversarial model manipulation* [HJM19]. Given a black-box model $\mathcal{B}$ and a target saliency map h_t, adversarial model manipulation aims at forcing the saliency map of any input to be similar to h_t. This goal is achieved by fine-tuning $\mathcal{B}$ with a training objective that penalizes the original training objective of $\mathcal{B}$ with a term involving h_t. The resulting model $\mathcal{B}'$ displays an accuracy close to that of $\mathcal{B}$ but produces for any input a saliency map that is similar to h_t. In a similar direction, Anders, Pasliev, Dombrowski, Müller and Kessel have further demonstrated that the model can be manipulated such that it has perfect fidelity with the black-box model $\mathcal{B}$ while reproducing arbitrary saliency maps [APD+20]. Following the same line of research, Dombrowski, Alber, Anders, Ackermann, Müller and Kessel [DAA+19] have demonstrated that saliency map based techniques are also vulnerable to *adversarial input manipulation*.

9.4 Case study 2 : Characterization of the risk of fairwashing

In this section, we will present a concrete example of the tension between fairness and explainability. In particular, we will discuss how fairwashing is possible in the context of an unfair black-box model that will be explained by a fairer model through post-hoc explanations' manipulation. However, to realize this, the post-hoc explanation model must produce different predictions than the original black-box on some inputs, leading to a decrease in the fidelity imposed by the difference in unfairness. A more detailed version of this case study can be found in [AAGH21].

9.4.1 Setting and problem formulation

Notations. Let $X \in \mathcal{X} \subset \mathbb{R}^n$ denote a feature vector, $Y \in \mathcal{Y} = \{0, 1\}$ its associated binary label (for simplicity we assume a binary classification setup without loss of generalization) and $G \in \mathcal{G} = \{0, 1\}$ a feature defining a group membership (*e.g.*, with respect to a sensitive attribute) for every data point sampled from $\mathcal{X}$. In addition, we assume that $\mathcal{B} : \mathcal{X} \to \mathcal{Y}$ refers to a black-box classifier of a particular model class $\mathcal{B}$ (*e.g.*, neural network or ensemble model) mapping any input $X \in \mathcal{X}$ to its associated prediction $\hat{Y} \in \mathcal{Y}$. Finally, let $e : \mathcal{X} \to \mathcal{Y}$ be a global explanation model from a particular model class $\mathcal{E}$ (*e.g.*, linear model, rule list or decision tree) designed to explain $\mathcal{B}$.

We train a classifier f (black-box model or explanation model) by minimizing its average loss $L_{\mathcal{D}}(f)$, for a given loss function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ (*e.g.*, cross entropy), over a dataset of interest $\mathcal{D}$ (*e.g.*, training set or suing group), which means $L_{\mathcal{D}}(f) = \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}}[l(f(x_i), y_i)]$. In addition, the performance of f is also measured in terms of its unfairness $\texttt{unf}_{\mathcal{D}}(f)$ on $\mathcal{D}$.

In this work, we focus on *statistical notions of fairness* [CV10, Cho17, CDPF+17, HPS+16], which require a model to exhibit approximate parity according to a statistical measure across the different groups defined by the group membership G. In particular, we consider four different statistical notions of fairness, namely statistical parity [DHP+12, CV10, KAAS12, FFM+15, Zli15], predictive equality [Cho17, CDPF+17], equal opportunity [HPS+16] and

equalized odds [Cho17, KMR17, HPS+16, ZVGRG17]. The definitions of these fairness metrics are listed in Table 8.

Table 8: Summary of the different statistical notions of fairness considered.

Fairness notion	**Definition**
Statistical Parity (Δ_{SP})	$\lvert P(\hat{Y}=1\lvert G=0) - P(\hat{Y}=1\lvert G=1)\rvert$
Predictive Equality (Δ_{PE})	$\lvert P(\hat{Y}=1\lvert Y=0,G=0) - P(\hat{Y}=1\lvert Y=0,G=1)\rvert$
Equal Opportunity (Δ_{EOpp})	$\lvert P(\hat{Y}=1\lvert Y=1,G=0) - P(\hat{Y}=1\lvert Y=1,G=1)\rvert$
Equalized Odds (Δ_{EOdds})	$\lvert P(\hat{Y}=1\lvert Y=1,G=0) - P(\hat{Y}=1\lvert Y=1,G=1)\rvert$ and $\lvert P(\hat{Y}=1\lvert Y=0,G=0) - P(\hat{Y}=1\lvert Y=0,G=1)\rvert$

Problem formulation. Our investigation is motivated by the previous work from [AAF+19], who have defined fairwashing in model and outcome explanations as a manipulation exercise in which high-fidelity and fairer explanations can be designed to explain unfair black-box models.

Definition 2 (Global explanation fidelity) Let $\mathcal{B}$ be a black-box model, e a global explanation model for $\mathcal{B}$ and X a set of data instances. Following the definition in [CS96], the fidelity of e with respect to $\mathcal{B}$ on X is expressed as:

$$\mathsf{fidelity}(e) = \frac{1}{|X|} \sum_{x \in X} \mathbb{I}(e(x) = \mathcal{B}(x)).$$

Definition 3 (Global fairwashing attack) Let $\mathcal{B}$ be a black-box model and X_{sg} a set of data instances hereafter referred to as suing group. A global fairwashing attack consists in finding an interpretable global model $e = p(\mathcal{B}, X_{sg})$ derived from the black-box $\mathcal{B}$ and the suing group X_{sg} using some attack process $p(\cdot,\cdot)$, such that e is fairer than $\mathcal{B}$ for a given fairness metric.

[AAF+19] devised `LaundryML`, an algorithm that can systematically fairwash unfair black-box models' decisions through both global and local explanations. `LaundryML` is a constrained model enumeration technique [HM17] that searches for explanation models maximizing the fidelity while minimizing the unfairness for a given unfair black-box model.

While [AAF+19] focused on the search of high-fidelity explanation models to perform fairwashing attacks, our work goes a step further by determining the fidelity-unfairness trade-offs of those attacks. This allows for a better characterization of the *manipulability* of the explanations. For this purpose, we will compute the set of Pareto optimal explanation models describing all the achievable fidelity-unfairness trade-offs, by solving the following problem:

$$\begin{aligned} \text{minimize} \quad & L_{\mathcal{D}_{sg}}(e) \\ \text{subject to} \quad & \texttt{unf}_{\mathcal{D}_{sg}}(e) \leq \epsilon, \end{aligned} \tag{4}$$

in which e is the explanation model, $\mathcal{D}_{sg} = \{X_{sg}, \mathcal{B}(X_{sg})\}$ is formed by the suing group and the prediction of the black-box model $\mathcal{B}$ on the suing group, while ϵ is the value of the unfairness constraint.

9.4.2 Experimental evaluation

The main objective of this section is to demonstrate empirically that (1) the incurred error in the fidelity imposed by fairwashing can be very small (Section 9.4.3), (2) fairwashing can generalize beyond suing groups (Section 9.4.4) and (3) transfer across black-box models (Section 9.4.5).

Datasets. We have investigated a real-world data commonly used in the fairness literature, namely `Adult Income`.

Adult Income. The UCI Adult Income [FA10] dataset contains demographic information about $48,842$ individuals from the 1994 U.S. census. The associated classification task consists in predicting whether a particular individual earns more than $50,000$\$ per year. We used `gender` (`Female`, `Male`) as group membership.

Preprocessing. Before running the experiments, the dataset is split into three subsets, namely the *training set* (67%), the *suing group* (16.5%) and the *test set* (16.5%). Overall, we created 10 different samplings of the three subsets using different random seeds, the results of all the experiments being averaged over these 10 samples. The training set is directly used to train the black-box models, while the suing group dataset is used to prepare the explanation models as well to evaluate their fidelity-unfairness trade-offs. Finally, the test set is used to assess the accuracy of the black-box models as well as the generalization of the explanation models beyond their suing groups. For all models (*i.e.*, black-boxes and explanation models), we used a one-hot encoding of the features of the dataset.

Black-box models. We have trained four different types of black-box models on the dataset, namely a Deep Neural Network (DNN), a Random Forest (RF) [Bre01], a AdaBoost classifier [FS97] and a XgBoost classifier [CG16]. To tune the hyperparameters of these models, during their training, we performed a hyperparameter search with 25 iterations using `HyperOpt` [BYC13].

Explanation models. We solved the optimisation problem defined in Equation 4 for logistic regression model class. In particular, we used the exponentiated gradient technique [ABD$^+$18], which is a model agnostic technique to train any classifier under fairness constraints (we use its implementation in the `Fairlearn` library [BDE$^+$20]).

9.4.3 Experiment 1: fidelity-unfairness trade-offs in fairwashing

The main objective of this experiment is to characterize the fidelity-unfairness trade-offs incurred by fairwashing when using different fairness metrics and black-box models.

Setup. Given a suing group X_{sg}, for each black-box model $\mathcal{B}$ and each fairness metric m, the Pareto fronts are obtained by first sweeping over 300 values of fairness coefficients $\epsilon_m \in [0, 1]$. Afterwards, for each value of ϵ_m, an explanation model e_{ϵ_m} is trained to satisfy the unfairness

constraint ϵ_m on X_{sg}, by solving the problem in Equation 4 for logistic regression. Then, its effective unfairness and fidelity (with respect to $\mathcal{B}$) on X_{sg} are returned. Finally, the set of non-dominated points is computed.

Results. Top rows in Figure 6 show the fidelity-unfairness trade-offs of fairwashed logistic regression explainers found for the four black-box models, respectively on Adult Income, for members of the suing group, using four different fairness metrics: equalized odds, equal opportunity, predictive equality and statistical parity.

Consistently over all these results, we observe that the fairwashed explanation models found for the suing groups were significantly less unfair than the black-box models while maintaining high fidelity. More precisely, for any combination of fairness metric m and black-box model $\mathcal{B}$, a fairwashed logistic regression displays an unfairness less than 50% of the unfairness of $\mathcal{B}$ while maintaining a fidelity greater than 90%.

9.4.4 Experiment 2: generalization of fairwashing beyond suing groups

The main objective of this experiment is to assess the generalization of a fairwashed explanation models beyond the suing group. This generalization indicates whether or not fairwashing is an attack that has to be tailored for a particular subset of data instances or whether it is more generic in its scope and thus also more problematic.

Setup. We used the same experimental setup as in `Experiment 1`. However, the unfairness and fidelity of the explanation model are computed on a test set X_{test} such that $X_{sg} \cap X_{test} = \emptyset$. The role of X_{test} is to mimic non-members of the suing group X_{sg}, which is targeted by the fairwashed explanation model.

Results. Bottom rows in Figure 6 show the fidelity-unfairness trade-offs of fairwashed logistic regression explainers found for the four black-box models on Adult Income, for non-members of the suing group, using four different fairness metrics: equalized odds, equal opportunity, predictive equality and statistical parity.

Overall, the results show that the explanation models designed for a particular suing group generalize well also to non-members of that suing group by achieving similar fidelity-unfairness trade-offs. The small gap between results on members of the suing group X_{sg} and those on non-members X_{test} can be explained by the fact that methods for learning fair classifiers are usually not robust to perturbations in the training distribution [HV19], resulting in fairness violations when evaluated on test sets. In fact, the fairwashing attack defined in Equation 4 is equivalent to a problem of training an interpretable model under a fairness constraint, in which the training pair (X, Y) is formed by the suing group X_{sg} and the predictions $\mathcal{B}(X_{sg})$ of the black-box $\mathcal{B}$. As a result, the issue of generalizing beyond the suing group can be reduced to the problem of generalizing fairness beyond the training set. In our scenario, the fairness violations are small enough to favor high-fidelity explanation models on the test set. Nonetheless, with

the development of robust fairness-enhancing techniques (*e.g.*, [MDJ+20]), it is reasonable to expect that this small gap in fidelity could be narrowed further in the future.

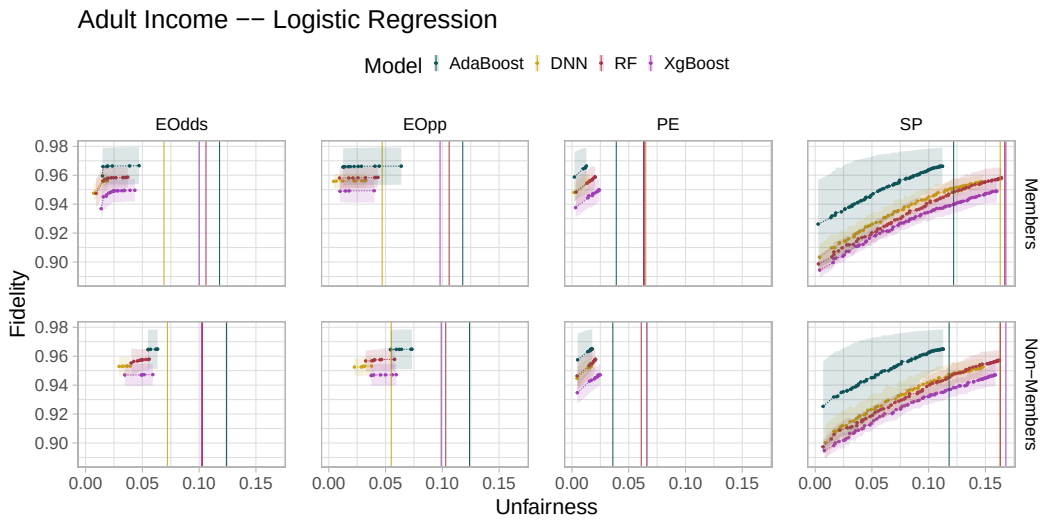


Figure 6: Fidelity-unfairness trade-off of fairwashing attacks for equalized odds, equal opportunity, predictive equality and statistical parity metrics on Adult Income, using logistic regression models as explanation models. Vertical lines denote the unfairness of the black-box models. Results are averaged over 10 fairwashing attacks.

9.4.5 Experiment 3: transferability of fairwashing

The objective of this experiment is to verify whether an explanation model specifically designed for a particular black-box model (called the teacher model) can be used to fairwash decisions of other black-box models (here the student models). The motivation for this experiment is similar to the study of transferability in the context of adversarial learning, which shows that it is not so much the characteristics of the black-box model rather than that of the dataset and the classification task that makes the attack possible.

Setup. Given a suing group X_{sg}, a teacher black-box model $\mathcal{B}_{teacher}$, a fairness metric m, its associated fairness constraint ϵ_m and a set of student black-box models $\mathcal{B}^i_{student}$, with $i = 1, \ldots, n$, an explanation model e_{ϵ_m} is trained to satisfy the unfairness constraint ϵ_m on X_{sg}, by solving the problem in Eq 4 for logistic regression. Afterwards, the unfairness and fidelity of e_{ϵ_m} are evaluated with respect to each of the student black-box models $\mathcal{B}^i_{student}$ on X_{sg}. For this experiment, we considered four black-box models (AdaBoost, DNN, RF and XgBoost). First, we fixed one model as the teacher black-box model and used the remaining ones as the

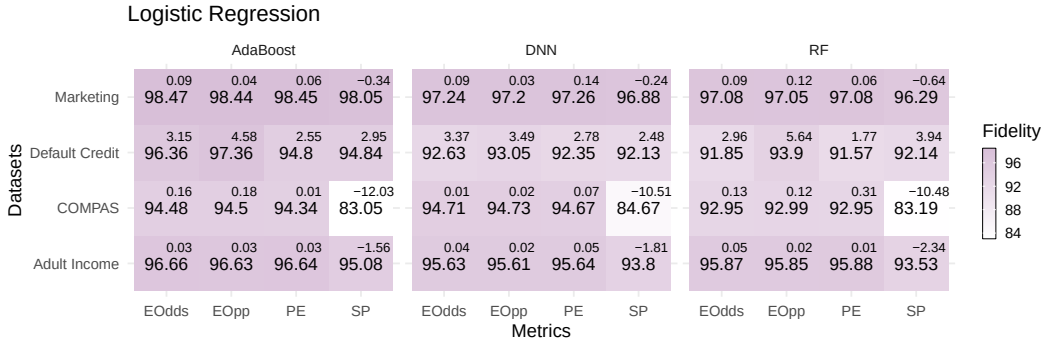


Figure 7: Fidelity of the fairwashed logistic regression models that are 50% less unfair than the black-box models they are explaining. Results (averaged over 10 fairwashing attacks) are shown for AdaBoost, DNN and RF black-box models, for all datasets and fairness metrics. The content of each cell is in the form of x^y, in which x represents the fidelity of the fairwashed explanation model, and y its percentage change with respect to the fidelity of the unconstrained explainer, used here as a baseline.

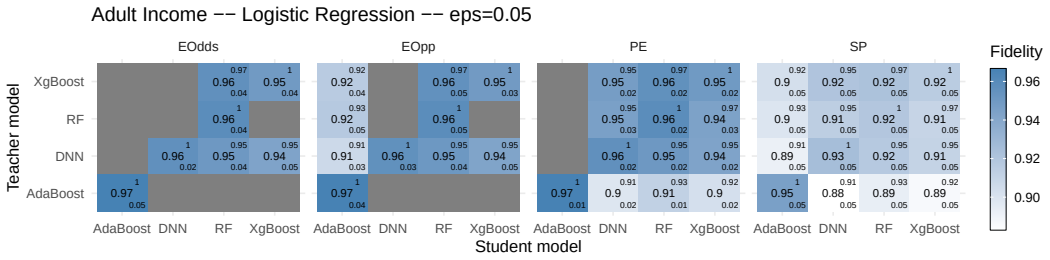


Figure 8: Analysis of the transferability of fairwashing attacks for equalized odds, equal opportunity, predictive parity and statistical parity on Adult Income, for $\epsilon = 0.05$, and for logistic regression explanation models. The result in each cell is in the form of x_z^y, in which y denotes the label agreement between the teacher black-box model and the student black-box model, x is the fidelity of the fairwashed explanation model and z is its unfairness. Blank cells denotes the absence of transferability for the unfairness constraint imposed. Results are averaged over 10 fairwashing attacks.

student black-box models. We conducted the experiments for all four possible combinations of the teacher-student models. We evaluated the transferability for an unfairness constraint of 0.05.

Results. Figure 8 displays the fidelity and unfairness of fairwashed logistic regression explainers with respect to both the teacher and the set of student black-box models on the suing group X_{sg}. Results are shown for Adult Income, for all fairness metrics.

Overall, our results demonstrate that fairwashing can transfer with high fidelity for most models. For instance, on Adult Income with an equal opportunity constraint set to 0.05, a fairwashed logistic regression explainer that had a fidelity of 96% for a DNN teacher model successfully transferred to AdaBoost, RF and XgBoost student models with a fidelity of 91%, 95% and 94% respectively (*c.f.*, Figure 8).

9.4.6 Discussion

In this case study, we have characterized the manipulability power of fairwashing attacks by analyzing their fidelity-unfairness trade-offs in diverse situations. In particular, we have demonstrated for different fairness metrics and black-box models that (1) fairwashed explanation models can exhibit significantly low unfairness while having a high fidelity to the black-box, (2) fairwashed explanation models can generalize beyond the suing group and (3) fairwashing attacks can transfer across black-box models.

The first lesson to draw from our investigation is that *relying on the fidelity as a proxy for the quality of a post-hoc explanation can be misleading* as a fairwashed explanation model can exhibit high fidelity while being significantly less unfair than the black-box model being explained. In addition, the results obtained for the generalization of fairwashed models beyond suing groups demonstrate that a *fairwashed explanation model can also rationalize subsequent unfair decisions made by the original black-box model for free.* This fact preclude the possibility of designing fairwashing detection techniques that leverage on the instability of the unfairness with respect to variations in the suing group. Indeed, such technique will most likely fail against fairwashed explanation models designed using stable fair classification algorithms [HV19, MDJ+20].

Furthermore, while generalizing beyond the suing group enables a dishonest black-box model producer to reuse its fairwashed explanation models for subsequent unfair decisions, the transferability property shown in the experiments could help the latter *to use a fairwashed explanation models to rationalize unfair decisions for other black-box models than the one it was designed for.* As a direct consequence, a model producer accused of deploying a black-box model providing unfair decisions could develop after *ex post facto* another black-box model to rationalize the unfair decisions of the first one. These results also suggest that auditing a black-box model for fairwashing should go beyond the model itself and consider the data distribution as well as the learning task, and consider the Rashomon set of the explainer.

10 Conclusion

Despite the fact that the interactions between privacy and ethics remain to be fully characterized in the domain of machine learning, in this report we have tried to partially address this gap by investigating some of the convergences and tensions between privacy and other ethical values that should be integrated into real-life machine learning applications. For example, we have seen that some implementations of fairness and security are in contradiction with privacy, securing against vulnerabilities targeting machine learning models' integrity can potentially harm privacy, and making automated decisions more transparent through post-hoc explanations can lead to powerful privacy attacks. We have also discussed the convergences and tensions that exist between privacy and other values from a normative perspective, such as how the right to erasure can conflict with privacy and how the former can even foster practices at odds with data minimization.

To address these issues, we believe that a partial solution would be to enshrine a liability of the processor that would be especially relevant in machine learning, as most of the controllers do not deploy themselves AI systems but only use them. To guarantee efficient protection of the data subject, a chain of liability from controllers to processors must be created. Another recommendation would be to enshrine specific rules for profiling. At this stage, Bill C-11 does not provide for any provisions, although measures would be necessary to establish a balance between privacy and machine learning.

Furthermore, designing and troubleshooting a system that simultaneously exhibits all of these properties would require a combination of expertise that is quite rare in today's data science job market. Therefore, it is crucial to consider separation of concerns approaches favouring collaboration instead of over-specialization. For example, an algorithmic fairness expert and a data privacy expert should be able to collaborate to design a machine learning model that is simultaneously fair and robust against privacy inference attacks. Not only does such an approach allows rapid development, but it also makes it possible to better benefit from the advancement of the state of knowledge in the various fields involved. We have already described existing contributions promoting such an approach through different convergences we highlighted in this work.

Finally, we also recommend that AI decisions should be supervised by a human whenever they are likely to affect the interests of individuals, as a group or individually. A particular challenge being to verify the quality and the effectiveness of the human control, as well as the moment of its intervention (before the decision is taken by the machine, afterwards, following the whole process). The question of human control is essential, but we must be concerned with the concrete modalities of its implementation to guarantee the protection of individuals' interests. It is also necessary that the human control is effective and that the human is not influenced by the machine, which must be verified. Stated more concretely, the human intervention has to be a true way of oversighting.

References

[AAF+19] Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. Fairwashing: the risk of rationalization. In *International Conference on Machine Learning*, pages 161–170. PMLR, 2019.

[AAGH21] Ulrich Aïvodji, Hiromi Arai, Sébastien Gambs, and Satoshi Hara. Characterizing the risk of fairwashing. In *Proceedings of the 35th Conference on Neural Information Processing Systems, NeurIPS 2021*, 2021.

[ABD+18] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning*, pages 60–69, 2018.

[ABG20] Ulrich Aïvodji, Alexandre Bolot, and Sébastien Gambs. Model extraction from counterfactual explanations. *arXiv preprint arXiv:2009.01884*, 2020.

[ABG+21] Ulrich Aïvodji, François Bidet, Sébastien Gambs, Rosin Claude Ngueveu, and Alain Tapp. Local data debiasing for fairness based on generative adversarial training. *Algorithms*, 14(3):87, 2021.

[AC18] Mike Ananny and Kate Crawford. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *new media & society*, 20(3):973–989, 2018.

[ADJM07] Machanavajjhala Ashwin, Kifer Daniel, Gehrke Johannes, and Venkitasubramaniam Muthuramakrishnan. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1):1–52, 2007.

[Adl] Michael Adler. A socio-legal approach to administrative justice. 25(4):323–352.

[ADRDS+20] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[Aga20] Sushant Agarwal. Trade-offs between fairness, interpretability, and privacy in machine learning. Master's thesis, University of Waterloo, 2020.

[AGG18] Ajay Agrawal, Joshua Gans, and Avi Goldfarb. *Prediction machines: the simple economics of artificial intelligence*. Harvard Business Press, 2018.

[AH20] Chirag Agarwal and Sara Hooker. Estimating example difficulty using variance of gradients. *arXiv preprint arXiv:2008.11600*, 2020.

[ALMK16] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica, May*, 23, 2016.

[AMS+15] Giuseppe Ateniese, Luigi V Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali, and Giovanni Felici. 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.

[And] Leighton Andrews. Public administration, public leadership and the construction of public value in the age of the algorithm and 'big data'. 97(2):296–310. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/padm.12534.

[APD+20] Christopher J Anders, Plamen Pasliev, Ann-Kathrin Dombrowski, Klaus-Robert Müller, and Pan Kessel. Fairwashing explanations with off-manifold detergent. *arXiv preprint arXiv:2007.09969*, 2020.

[ASJ+19] Buse Gul Atli, Sebastian Szyller, Mika Juuti, Samuel Marchal, and N Asokan. Extraction of complex dnn models: Real threat or boogeyman? *arXiv preprint arXiv:1910.05429*, 2019.

[BCCC+19] Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. *arXiv preprint arXiv:1912.03817*, 2019.

[BCM+13] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

[BDE+20] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in ai. *Microsoft Technical Report MSR-TR-2020-32*, 2020.

[Ben95] C Bennett. Implementing privacy codes of practice: A report to the canadian standards association. *Rexdale: CSA*, 1995.

[BHJ+18] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 0(0):0049124118782533, 2018.

[Bie20] Elettra Bietti. From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 210–219, Barcelona, Spain, January 2020. Association for Computing Machinery.

[Bin18] Reuben Binns. Algorithmic accountability and public reason. *Philosophy & technology*, 31(4):543–556, 2018.

[BLM15] Denis Butin and Daniel Le Métayer. A guide to end-to-end privacy accountability. In *2015 IEEE/ACM 1st International Workshop on TEchnical and LEgal aspects of data pRivacy and SEcurity*, pages 20–25. IEEE, 2015.

[BMR+20] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[Bov] Mark Bovens. Analysing and assessing accountability: A conceptual framework1. 13(4):447–468. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0386.2007.00378.x.

[Bov10] Mark Bovens. Two concepts of accountability: Accountability as a virtue and as a mechanism. *West European Politics*, 33(5):946–967, 2010.

[Bre01] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[BS19] Eugene Bagdasaryan and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. *arXiv preprint arXiv:1905.12101*, 2019.

[BvdH15] Koen Bruynseels and Jeroen van den Hoven. How to do things with personal big biodata. *Social dimensions of privacy: interdisciplinary perspectives*, page 122, 2015.

[BYC13] James Bergstra, Daniel Yamins, and David Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proceedings of the 30th International Conference on Machine Learning*, pages 115–123, 2013.

[CDPF+17] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806, 2017.

[CG16] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[Cha17] Louis Chartrand. Agencéité et responsabilité des agents artificiels. *Éthique publique. Revue internationale d'éthique sociétale et gouvernementale*, 19(2), 2017.

[Cho17] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[CLE+19] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pages 267–284, 2019.

[Cof19] Ignacio N Cofone. Algorithmic discrimination is an information problem. *Hastings LJ*, 70:1389, 2019.

[Cof21] Ignacio Cofone. Ai and judicial decision-making. *Artificial Intelligence and the Law in Canada (Toronto: LexisNexis Canada, 2021)*, 2021.

[CR18] Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

[CR19] Celine Castets-Renard. Accountability of algorithms in the gdpr and beyond: A european legal framework on automated decision-making. *Fordham Intell. Prop. Media & Ent. LJ*, 30:91, 2019.

[CS96] Mark Craven and Jude W Shavlik. Extracting tree-structured representations of trained networks. In *Advances in neural information processing systems*, pages 24–30, 1996.

[CS19] Ignacio N Cofone and Katherine J Strandburg. Strategic games and algorithmic secrecy. *McGill LJ*, 64:623, 2019.

[CS20] Hongyan Chang and Reza Shokri. On the privacy risks of algorithmic fairness. *arXiv preprint arXiv:2011.03731*, 2020.

[CTW+20] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. *arXiv preprint arXiv:2012.07805*, 2020.

[CV10] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

[CY15] Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pages 463–480. IEEE, 2015.

[CZW+20] Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. When machine unlearning jeopardizes privacy. *arXiv preprint arXiv:2005.02205*, 2020.

[DAA+19] Ann-Kathrin Dombrowski, Maximilian Alber, Christopher J Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. *arXiv preprint arXiv:1906.07983*, 2019.

[Dan16] John Danaher. The threat of algocracy: Reality, resistance and accommodation. *Philosophy & Technology*, 29(3):245–268, 2016.

[DeC97] Judith Wagner DeCew. *In pursuit of privacy: Law, ethics, and the rise of technology*. Cornell University Press, 1997.

[DF] Nicholas Diakopoulos and Sorelle Friedler. We need to hold algorithms accountable—here's how to do it.

[DFA+] Nicholas Diakopoulos, Sorelle Friedler, Marcelo Arenas, Solon Barocas, Michael Hay, Bill Howe, H. V. Jagadish, Kris Unsworth, Arnaud Sahuguet, Suresh Venkatasubramanian, Christo Wilson, Cong Yu, and Bendert Zevenbergen. Principles for accountable algorithms and a social impact statement for algorithms :: FAT ML.

[DHP+12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[DR+14] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[Ele16] Electronic Privacy Information Center. EPIC - Algorithms in the Criminal Justice System. `https://epic.org/ai/criminal-justice/index.html`, 2016.

[EMH19] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *Journal of Machine Learning Research*, 20:1–21, 2019.

[Esh14] Andrew Eshleman. Moral responsibility. 2014.

[FA10] Andrew Frank and Arthur Asuncion. Uci machine learning repository [http://archive. ics. uci. edu/ml]. irvine, ca: University of california. *School of information and computer science*, 213:2–2, 2010.

[Fac20] Facebook. How Does Facebook Use Machine Learning to Deliver Ads?, June 2020.

[FC19] Luciano Floridi and Josh Cowls. A unified framework of five principles for ai in society. *Issue 1.1, Summer 2019*, 1(1), 2019.

[FDC20] Munira Ferdous, Jui Debnath, and Narayan Ranjan Chakraborty. Machine learning algorithms in healthcare: A literature survey. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE, 2020.

[FFM+15] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268, 2015.

[FHM19] Kazuto Fukuchi, Satoshi Hara, and Takanori Maehara. Pretending fair decisions via stealthily biased sampling. *arXiv preprint arXiv:1901.08291*, 2019.

[FJR15] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333, 2015.

[FLJ+14] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *23rd {USENIX} Security Symposium ({USENIX} Security 14)*, pages 17–32, 2014.

[Flo16] Luciano Floridi. Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2083):20160112, 2016.

[Flo19] Luciano Floridi. Translating principles into practices of digital ethics: Five risks of being unethical. *Philosophy & Technology*, pages 1–9, 2019.

[FS97] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[FSV16] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.

[FTT12] Evanthia Faliagka, Athanasios Tsakalidis, and Giannis Tzimas. An integrated e-recruitment system for automated personality mining and applicant ranking. *Internet research*, 2012.

[Gal17] Scott Galloway. *The four: the hidden DNA of Amazon, Apple, Facebook and Google*. Random House, 2017.

[GCV+18] Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen, and Freddy Lecue. Interpretable credit application predictions with counterfactual explanations. *arXiv preprint arXiv:1811.05245*, 2018.

[GGVZ19] Antonio Ginart, Melody Y Guan, Gregory Valiant, and James Zou. Making ai forget you: Data deletion in machine learning. *arXiv preprint arXiv:1907.05012*, 2019.

[GHI12] Daniel Guagnin, Leon Hempel, and Carla Ilten. Bridging the gap: We need to get together. In *Managing privacy through accountability*, pages 102–124. Springer, 2012.

[Gil] Sharon Gilad. Accountability or expectations management? the role of the ombudsman in financial regulation. 30(2):227–253.

[GMR+18] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[Gol15] Alan H Goldman. *Justice and reverse discrimination*. Princeton University Press, 2015.

[Gre17] Andy Greenberg. How one of apple's key privacy safeguards falls short, 2017.

[Gre19] Ben Green. "good" isn't good enough. 2019. Presented at AI for Social Good workshop at NeurIPS (2019).

[GSS14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[Haw12] Fran Hawthorne. *Ethical chic: The inside story of the companies we think we love*. Beacon Press, 2012.

[HCC+19] Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*, 2019.

[Hel20] Deborah Hellman. Measuring algorithmic fairness. *Va. L. Rev.*, 106:811, 2020.

[HJKRR18] Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.

[HJM19] Juyeon Heo, Sunghwan Joo, and Taesup Moon. Fooling neural network interpretations via adversarial model manipulation. *arXiv preprint arXiv:1902.02041*, 2019.

[HM17] Satoshi Hara and Takanori Maehara. Enumerate Lasso solutions for feature selection. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 1985–1991, 2017.

[HMC+20] Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058*, 2020.

[HPS+16] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[HSNL18] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.

[HV19] Lingxiao Huang and Nisheeth Vishnoi. Stable and fair classification. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2879–2890, 2019.

[Ish19] Kaori Ishii. Comparative legal study on privacy and personal data protection for robots equipped with artificial intelligence: looking at functional and technological aspects. *AI & SOCIETY*, 34(3):509–533, 2019.

[JCB+20] Matthew Jagielski, Nicholas Carlini, David Berthelot, Alex Kurakin, and Nicolas Papernot. High accuracy and high fidelity extraction of neural networks. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*, 2020.

[JDHF19] Kerina Jones, Helen Daniels, Sharon Heys, and David Vincent Ford. Toward an ethically founded framework for the use of mobile phone call detail records in health research. *JMIR Mhealth Uhealth*, 7, 2019.

[JE19] Bargav Jayaraman and David Evans. Evaluating differentially private machine learning in practice. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pages 1895–1912, 2019.

[JKM+19] Matthew Jagielski, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi-Malvajerdi, and Jonathan Ullman. Differentially private fair learning. In *International Conference on Machine Learning*, pages 3000–3008. PMLR, 2019.

[JKMR16] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, pages 325–333, 2016.

[JKV+19] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*, 2019.

[JOB+18] Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 19–35. IEEE, 2018.

[JZTM20] Ziheng Jiang, Chiyuan Zhang, Kunal Talwar, and Michael C Mozer. Characterizing structural regularities of labeled data in overparameterized models. *arXiv e-prints*, pages arXiv–2002, 2020.

[KAAS12] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50, 2012.

[KBBV20] Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions. In *International Conference on Artificial Intelligence and Statistics*, pages 895–905, 2020.

[KBF+16] Joshua A Kroll, Solon Barocas, Edward W Felten, Joel R Reidenberg, David G Robinson, and Harlan Yu. Accountable algorithms. *U. Pa. L. Rev.*, 165:633, 2016.

[KC12] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

[KCP+17] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666, 2017.

[KGK+18] Niki Kilbertus, Adrià Gascón, Matt Kusner, Michael Veale, Krishna Gummadi, and Adrian Weller. Blind justice: Fairness with encrypted sensitive attributes. In *International Conference on Machine Learning*, pages 2630–2639. PMLR, 2018.

[KLRS17] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

[KMR17] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

[Lan20] Douglas B. Laney. Data Monetization: New Value Streams You Need Right Now. *Forbes*, June 2020. Section: CIO Network.

[LB19] Himabindu Lakkaraju and Osbert Bastani. " how do i fool you?": Manipulating user trust via misleading black box explanations. *arXiv preprint arXiv:1911.06473*, 2019.

[LBC+20] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. In *NeurIPS*, 2020.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[Lip18] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[LL17] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.

[LLM+17] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. Inverse classification for comparison-based interpretability in machine learning. *arXiv preprint arXiv:1712.08443*, 2017.

[LLM+19] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. The dangers of post-hoc interpretability: unjustified counterfactual explanations. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 2801–2807. AAAI Press, 2019.

[LLS+17] Michael T Lash, Qihang Lin, Nick Street, Jennifer G Robinson, and Jeffrey Ohlmann. Generalized inverse classification. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 162–170. SIAM, 2017.

[LLV07] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 106–115. IEEE, 2007.

[Mar17] Dominic Martin. Frankenstein 2.0. In Catherine Régis, Karim Benyekhlef, and Daniel Weinstock, editors, *Sauvons la justice!*, pages 109–114. Montréal, del busson Éditeur edition, 2017.

[MD19] Gary Marcus and Ernest Davis. *Rebooting AI: Building Artificial Intelligence We Can Trust*. Pantheon, New York, September 2019.

[MDJ+20] Debmalya Mandal, Samuel Deng, Suman Jana, Jeannette Wing, and Daniel J Hsu. Ensuring fairness beyond the training data. In *Advances in Neural Information Processing Systems*, 2020.

[Met19] Thomas Metzinger. Eu guidelines: Ethics washing made in europe. `https://www.tagesspiegel.de/politik/eu-guidelines-ethics-washing-made-in-europe/24195496.html`, 2019. Accessed: 2019-10-09.

[Mit97] TM Mitchell. Machine learning, mcgraw-hill higher education. *New York*, 1997.

[MJ51] Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

[ML20] Abdul Majeed and Sungchang Lee. Attribute susceptibility and entropy based data anonymization to improve users community privacy and utility in publishing data. *Applied Intelligence*, pages 1–20, 2020.

[MSDH19] Smitha Milli, Ludwig Schmidt, Anca D Dragan, and Moritz Hardt. Model reconstruction from model explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 1–9, 2019.

[MST20] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, 2020.

[MT19] Erwan Le Merrer and Gilles Tredan. The bouncer problem: Challenges to remote explainability. *arXiv preprint arXiv:1910.01432*, 2019.

[Mul00] Richard Mulgan. 'accountability': An ever-expanding concept? *Public administration*, 78(3):555–573, 2000.

[Mur21] Kevin P. Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2021.

[Nar18] Arvind Narayanan. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*, 2018.

[NBC+08] Blaine Nelson, Marco Barreno, Fuching Jack Chi, Anthony D Joseph, Benjamin IP Rubinstein, Udam Saini, Charles A Sutton, J Doug Tygar, and Kai Xia. Exploiting machine learning to subvert your spam filter. *LEET*, 8:1–9, 2008.

[Ng18] A Ng. How artificial intelligence and data add value to businesses. *McKinsey Global Institute, New York*, 2018.

[Nis09] Helen Nissenbaum. *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press, 2009.

[noa16] From Data to Disruption: Innovation Through Digital Intelligence - SPONSORED CONTENT FROM IBM. *Harvard Business Review*, December 2016. Section: Technology.

[noa19] From innovation to monetization: The economics of data-driven transformation. Technical report, MIT Technology Review, January 2019.

[NS18] Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[Pau17] Kim Pauline. Data-driven discrimination at work. *William and Mary Law Review*, 58, 2017.

[PBK20] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. Learning model-agnostic counterfactual explanations for tabular data. In *Proceedings of The Web Conference 2020*, pages 3126–3132, 2020.

[PMG+17] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017.

[PP12] Charles P Pfleeger and Shari Lawrence Pfleeger. *Analyzing computer security: A threat/vulnerability/countermeasure approach*. Prentice Hall Professional, 2012.

[PS19] Anya ER Prince and Daniel Schwarcz. Proxy discrimination in the age of artificial intelligence and big data. *Iowa L. Rev.*, 105:1257, 2019.

[PS20] Dana Pessach and Erez Shmueli. Algorithmic fairness. *arXiv preprint arXiv:2001.09784*, 2020.

[PWH20] Danqing Pan, Tong Wang, and Satoshi Hara. Interpretable companions for black-box models. In *International Conference on Artificial Intelligence and Statistics*, pages 2444–2454. PMLR, 2020.

[QM] Shengnan Qiu and Gillian Macnaughton. Mechanisms of accountability for the realization of the right to health in china. 19(1):279–292.

[Raw99] John Rawls. A theory of justice, rev. ed, 1999.

[Raw01] John Rawls. *Justice as fairness: A restatement*. Harvard University Press, 2001.

[RSCW] Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker. Algorithmic impact assessments: A practical framework for public agency accountability.

[RSG16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " 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*, pages 1135–1144, 2016.

[Rus19] Chris Russell. Efficient search for diverse coherent explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 20–28, 2019.

[SG19] Harini Suresh and John V Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2019.

[SHJ+19] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. How can we fool lime and shap? adversarial attacks on post hoc explanation methods. *arXiv preprint arXiv:1911.02508*, 2019.

[Sho11] David Shoemaker. Attributability, answerability, and accountability: Toward a wider theory of moral responsibility. *Ethics*, 121(3):602–632, 2011.

[SMDE20] Javier Sánchez-Monedero, Lina Dencik, and Lilian Edwards. What does it mean to'solve'the problem of discrimination in hiring? social, technical and legal perspectives from the uk on automated hiring systems. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 458–468, 2020.

[Smi12] Angela M Smith. Attributability, answerability, and accountability: In defense of a unified account. *Ethics*, 122(3):575–589, 2012.

[SSM19] Liwei Song, Reza Shokri, and Prateek Mittal. Privacy risks of securing machine learning models against adversarial examples. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 241–257, 2019.

[SSSS17] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017.

[SSZ19] Reza Shokri, Martin Strobel, and Yair Zick. On the privacy risks of model explanations. *arXiv preprint arXiv:1907.00164*, 2019.

[Swe02] Latanya Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):571–588, 2002.

[SZB+20] Ilia Shumailov, Yiren Zhao, Daniel Bates, Nicolas Papernot, Robert Mullins, and Ross Anderson. Sponge examples: Energy-latency attacks on neural networks. *arXiv preprint arXiv:2006.03463*, 2020.

[SZS+13] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[TFVH20] Cuong Tran, Ferdinando Fioretto, and Pascal Van Hentenryck. Differentially private and fair deep learning: A lagrangian dual approach. *arXiv preprint arXiv:2009.12562*, 2020.

[TH12] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

[TKB+17] Jun Tang, Aleksandra Korolova, Xiaolong Bai, Xueqiang Wang, and Xiaofeng Wang. Privacy loss in apple's implementation of differential privacy on macos 10.12. *arXiv preprint arXiv:1709.02753*, 2017.

[TSHL17] Gabriele Tolomei, Fabrizio Silvestri, Andrew Haines, and Mounia Lalmas. Interpretable predictions of tree-based ensembles via actionable feature tweaking. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 465–474, 2017.

[TTK18] Ugur Turan, Ismail H Toroslu, and Murat Kantarcioglu. Graph based proactive secure decomposition algorithm for context dependent attribute based inference control problem. *arXiv preprint arXiv:1803.00497*, 2018.

[TZJ+16] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction apis. In *25th {USENIX} Security Symposium ({USENIX} Security 16)*, pages 601–618, 2016.

[USL19] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 10–19, 2019.

[VB17] Michael Veale and Reuben Binns. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4(2):2053951717743530, 2017.

[vdHBPW20] Jeroen van den Hoven, Martijn Blaauw, Wolter Pieters, and Martijn Warnier. Privacy and Information Technology. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2020 edition, 2020.

[Vél21] Carissa Véliz. *Privacy is Power: Why and how You Should Take Back Control of Your Data*. London: Bantam Press, 2021.

[Vog07] David Vogel. *The market for virtue: The potential and limits of corporate social responsibility*. Brookings Institution Press, 2007.

[VR18] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE, 2018.

[Wag18] Ben Wagner. Ethics As An Escape From Regulation. From "Ethics-Washing" To Ethics-Shopping? In Emre Bayamiloglu, Irina Baraliuc, Liisa Janssens, and Mireille Hildebrandt, editors, *Being Profiled: Cogitas Ergo Sum*. Amsterdam University Press, 2018.

[Wan19] Tong Wang. Gaining free or low-cost interpretability with interpretable partial substitute. In *International Conference on Machine Learning*, pages 6505–6514. PMLR, 2019.

[WD19] Ben Wagner and Sylvie Delacroix. Constructing a mutually supportive interface between ethics and regulation. *Available at SSRN 3404179*, 2019.

[Wei19] Gabriel Weinberg. What if we all just sold non-creepy advertising? *The New York Times*, 2019.

[WG18] Binghui Wang and Neil Zhenqiang Gong. Stealing hyperparameters in machine learning. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 36–52. IEEE, 2018.

[Wie] Maranke Wieringa. What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 1–18. Association for Computing Machinery.

[WM19] Sandra Wachter and Brent Mittlestadt. A right to reasonable inferences: Rethinking data protection law in the age of big data and ai. *Columbia Business Law Review*, 2, 2019.

[WMR17] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[WMR21] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Bias preservation in machine learning: The legality of fairness metrics under eu non-discrimination law. *West Virginia Law Review, Forthcoming*, 2021.

[YGFJ18] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282. IEEE, 2018.

[YHP19] Karen Yeung, Andrew Howes, and Ganna Pogrebna. Ai governance by human rights-centred design, deliberation and oversight: An end to ethics washing. *The Oxford Handbook of AI Ethics, Oxford University Press (2019)*, 2019.

[Zli15] Indre Zliobaite. On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723*, 2015.

[Zub20] Shoshana Zuboff. *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. New York: PublicAffairs, 2020.

[ZVGRG17] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180. International World Wide Web Conferences Steering Committee, 2017.