Equal Protection Under the Algorithm: A Legal-Inspired Framework for Identifying Discrimination in Machine Learning

Sucheta Soundarajan [1] **Daniel L. Clausen** [2]

Abstract

Within the field of ethical machine learning, an area of special concern is the possibility of machine learning algorithms discriminating against groups of people in unethical ways, such as targeting advertisements based on race. In this paper, we propose a framework based on long-standing U.S. legal principles to determine whether the targeting of a group should be viewed with suspicion. Unlike existing work, we are focused on the case when the group is not correlated with known 'protected features', or such data is unavailable.

1. Introduction

The past decade has seen a profusion of machine learning techniques throughout society, reaching into nearly all aspects of our lives, communities, and world. One area of great concern is illegal or unethical discrimination in the decisions made by AI systems. How can one ensure that advertisements, news stories, and similar recommendations are not inappropriately targeted towards individuals on the basis of race, religion, or other protected categories?

Existing work uses the approach of defining a set of 'protected' features on which decisions should not be made (e.g., race or gender), and disallows algorithms from using these or correlated features. This strategy has been prominent in a number of algorithmic approaches. However, this approach suffers from important drawbacks:

First, data corresponding to protected categories may be unavailable, so correlated proxy variables cannot be identified.

Second, it is possible that an action is not discriminatory with respect to a protected class as a whole, but could be discriminatory with respect to a subclass.

[1]Department of Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY, USA [2]BurgherGray LLP, New York, NY, USA. Correspondence to: Sucheta Soundarajan <susounda@syr.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

Third, it is important to observe that protected categories do not simply exist as a contextless laundry list of labels. Rather, legal systems have determined that discriminating on the basis of certain categories is dangerous because of their historical and societal context. However, because legal systems generally favor simple and clear descriptions, it is possible that there are other categories that, while not currently legally protected, are equally deserving of protection. For example, while men as a group have not faced historic discrimination in business, it is possible that men of below-average height are discriminated against. With the advent of machine learning algorithms that can identify classes using combinations of features, it is possible that we can identify more complex classes that should be protected.

Existing research on fairness in machine learning draws from the equal protection laws of the United States, and in particular, from the *disparate impact test* from US employment law (Feldman et al., 2015)). However, this body of law is rich and deep, with a multitude of other tests from which researchers may draw. Our goal is to use that same body of legal standards to propose an algorithmic framework for determining if a targeted group may *deserve* protection.

Our framework is unsupervised, operating without predefined protected features. Unlike existing work, our purpose is explicitly *not* to determine whether discrimination has occurred with respect to protected features, but rather to determine whether discrimination with respect to a particular class of individuals is suspicious and *should* be subject to further examination. It is possible (even likely) that the categories identified by this framework will not match the specific suspect classes already identified in the law; and indeed, that is not our goal. The law is limited to what can be described in natural language, but ethical machine learning may demand that we go beyond current legal standards.

2. Related Work

In recent years, a great deal of attention has been paid to the problem of detecting unethical discrimination in decisions made by machine learning algorithms. Much of this work has assumed knowledge of certain 'protected' features (such as race), on which decisions should not be based. For

example, Hardt, et al. show how to adjust a predictor to remove discrimination based on such features (Hardt et al., 2017). This type of balance requires that the predicted label for an individual is independent of that individual's protected features, conditioned on the true label of the individual (Hardt et al., 2017; Chouldechova, 2017). The concept of demographic parity is fairly widespread (Zemel et al., 2013). Work in this area generally recognizes that even if protected features are not explicitly included, the feature set may contain correlated features, leading to problematic outcomes (Barocas et al., 2017; Ajunwa et al., 2016).

Some existing literature has considered extending existing legal principles to the machine learning setting, with a particular focus on the *disparate impact* legal test. For example, Feldman, et al. propose a test to determine whether an algorithmic decision has a disparate impact with respect to different protected groups (Feldman et al., 2015).

3. Legal Standards For Discrimination

Historically, the United States has seen a great deal of invidious discrimination on the basis of personal characteristics such as race, religion, gender, or poverty. The U.S. Constitution is vague as to what forms of discrimination are permissible, and when. In a typical process, the federal or a state legislature will create a law, policy, or some other action, which may then be challenged by a group or individual as being unconstitutionally discriminatory towards some group. The U.S. Supreme Court has developed a host of standards and tests (the *judicial review standards*) for examining such actions. These tests operate by first determining how 'suspect' the targeted group is (i.e., the level of protection needed), and then examining whether the discrimination is legally permissible.

A *suspect class* is defined as a class of people who, among other characteristics, may have (1) faced historic discrimination, possibly due to harmful stereotypes, (2) possess an immutable or highly-visible trait (such as skin color), and (3) lack power to protect themselves politically. Race, nationality, and religion are known suspect classes. If a class has been identified as suspect, then in order for a discriminatory action to be permissible, it must pass the strict scrunity test. To pass, the action must be (1) motivated by a compelling governmental interest (e.g., national security), and (2) must be narrowly tailored to achieve that interest.

4. Proposed Framework

Many of the legal concepts used in the judicial review standards can be neatly transferred to machine language terms. Here, we discuss how the pertinent concepts described in the **Legal Standards For Discrimination** section may be adapted to the machine learning setting. This framework is applicable to settings in which there are three sets of parties: the individuals being targeted for some treatment, the organizations that have an interest in a successful treatment, and a middleman organization that controls the treatment and targeting.

We again emphasize that our goal is *not* to show how one can infer whether a particular class of individuals is a class that is currently protected by US law, but rather to extend existing legal tests to the machine learning domain to (1) determine whether a class being targeted should be considered a suspect class and therefore deserving of protection, and (2) evaluate whether the treatment applied to that class is appropriate. When possible, for each factor within our test, we describe how it can be evaluated algorithmically or by a human.

Our framework consists of two parts: (1) identifying whether a group is a *suspect class*, and (2) if a classification is suspect, then determining whether the discriminatory treatment taken with respect to that class is appropriate.

Note that our goal in this work, rather than define a specific algorithm, is to provide an outline of a framework that shows how other parts of U.S. equal protection law may apply to machine learning. There are many types of applications where this work may be relevant, and it is not possible to create a single algorithm that is applicable to all of them. Rather, we provide a high-level description of how one can define an appropriate algorithm.

4.1. Terminology

A *treatment* is a targeting action, such as showing an ad. A *user* is an individual. A *class* is a set of users, and a *targeted class* is a class that receives a treatment. An *aggregator* is a middleman organization that performs the treatment. A *producer* is an organization on whose behalf the aggregator targets users. For example, users could represent individuals browsing the web, producers can be companies that use ads to sell products, and an aggregator could be Google or Facebook Ads. It is possible that the aggregator, the producer, or both determine which users to target.

4.2. Applicability

For this framework to be applied, the following are required:

Classification or Clustering Applications: This framework proposes standards that are appropriate for classification or clustering tasks in which users are grouped together into distinct groups to receive a treatment. Additionally, for the first test (identifying whether a class is suspect), it is necessary to have data that cuts across multiple targeting decisions, whether historically (e.g., the decisions are actually the same decision made at different points in time), or cutting across different targeting campaigns.

Presence of a Targeting Aggregator: This framework is meant to be used by aggregators that have an interest in balancing the interests of its users (the targeted individuals) with the interests of its clients (the producers), and have access to sufficient quantities and types of data to perform the analysis required by our framework. Practically speaking, these tests should not be left to those who stand to profit from unethical behavior; additionally, testing whether a class is suspect requires either historical or cross-cutting data to determine whether that class has faced a pattern of discrimination and whether the class is mutable. As in the legal world, we do not suggest a specific threshold for distinguishing between a suspect and non-suspect class. Rather, we suggest use of the listed factors as guidelines to flag a class of individuals as being potentially-suspect, and subject to further review (using, e.g., existing tests for disparate impact (Feldman et al., 2015; Barocas & Selbst, 2017)).

4.3. Identifying Suspect Classes

The first part of our test evaluates whether a targeted class should be flagged as suspect. It contains two factors: **Discriminatory Pattern** and **Mutability**.

Discriminatory Pattern: To what extent is there a discriminatory pattern with respect to this set of individuals? Discrimination with respect to a class is more likely to be problematic if individuals within that class have historically been grouped together for purposes of making other decisions, particularly ethically-problematic targeting decisions.

Algorithmic Evaluation: (1) If protected attributes are available, and the targeted class is very similar to a class that is known to have been historically discriminated against (e.g., one of the known-suspect classes based on race), or if the algorithm uses features based on these characteristics, then the class is automatically suspect. To determine whether this factor is met, the aggregator can maintain a list of protected attributes, and determine whether the targeted group disproportionately possesses one (or more) of the same suspect attributes. If so, the class is flagged as suspect. There are many ways to perform this so-called 'disparate impact' test (Feldman et al., 2015; Barocas & Selbst, 2017).

(2) If protected attributes like race or religion are not available, but other treatments have been applied to the same or a similar targeted class, then the class is suspect. For example, if a university is purchasing web advertisements and has decided to target or exclude some set of individuals, then if other universities have similarly decided to target or exclude the same group, this may be ethically dubious. This can be evaluated through use of cross-campaign data (if available). For example, if an advertising campaign is controlled or conducted by an aggregator like Facebook or Google rather than the individual producers that are purchasing the ads, that aggregator can compare the targeted class to each of the classes targeted in other campaigns; if the current targeted class has high overlap with many other targeted classes, then the class is suspect.

Mutability: To what extent do individuals move in or out of the targeted set? If a class is targeted, and individuals in that class are unable to move into or out of that class (or a subclass), then again potential ethical problems arise. Problems may arise in both directions: individuals should be able to both move out of classes that are negatively targeted (e.g., high-interest loans) and into classes that are beneficially targeted (e.g., low-interest loans). It is important to note that a class is not mutable simply because an individual can theoretically move in and out of it: this movement needs to actually happen with sufficient regularity.

Algorithmic Evaluation: (1) If longitudinal data is available, it is easy to determine whether individuals actually do move in and out of the targeted class, and with what frequency. Note that if this approach is used, even if the class as a whole is determined to be mutable, it is important to determine whether there are sub-classes of the targeted class that are themselves immutable. For example, the class of low-income individuals might appear to be mutable, because college students are often low-income but move out of the class once they graduate. However, this class contains large sub-classes that are themselves much less mutable (e.g., people trapped in the 'cycle of poverty').

(2) If the features of the class are interpretable, one could label individual features as mutable (e.g., hobbies) or immutable (e.g., gender), and determine whether the class is based on features that are themselves mutable or immutable.

4.4. Identifying Suspect Treatments

The second prong evaluates whether the discrimination itself is suspicious. It contains the **Compelling Interest**, **Narrow Tailoring**, and **Powerlessness** factors.

Compelling Interest: Is the reason for the discrimination compelling? In the legal domain, the government balances its own compelling interests (e.g., national defense, providing education, etc.) with the compelling interests of its citizens (e.g., free speech, free exercise of religion, etc.). In the machine learning domain, some decisions (e.g., medical applications) are clearly compelling. In other cases, the analysis is murkier: a store may view increasing its profits by a small amount to be a 'compelling interest', while the individuals being targeted by the store's advertisements may disagree. In this prong of the test, we propose balancing the interests of the producer against possible harms to the users.

Algorithmic Evaluation: Whether or not an interest is 'compelling' is subjective, and this prong is best evaluated by a human. One possible algorithmic option is to compare the expected benefit to the producer to the suspectness of the

targeted class. For example, if the producer expects a certain success rate (e.g., clickthrough rate) when targeting an extremely-suspect class (based on the factors above); but by targeting a broader and substantially less suspect class (as measured by the factors above), the expected success rate drops only slightly, then it is better to target the less-suspect class. The aggregator can manage this by setting its prices based on the suspectness of the targeted class.

Powerlessness: To what extent are individuals in the set aware that they are being targeted, and are they able to opt-out of or modify the targeting? In the judicial review standards, this particular factor is a characteristic of the class, but in the machine learning setting, it makes the most sense to view this as a feature of the entire system.

Algorithmic Evaluation: If it is clear to users that they are being targeted, and they understand why they are being targeted and can make efforts to change their class membership, then the discrimination is somewhat, though not completely, less problematic. Note that the ability to opt-out or modify the targeting must be real and meaningful: for example, in the web advertising context, if poor minority users are continually shown ads for payday lenders, for-profit colleges, online gambling, etc., and opting out of one ad just results in a different similar ad being shown, their ability to opt out does not translate into actual power.

Narrow Tailoring: If the features provided are interpretable, then one can ask whether the treatment is clearly connected with the features of the class. In the legal setting, laws typically function to restrict the behavior of individuals, and thus the courts apply a *least restrictive means* test to ensure that the restriction is as minimal as possible while still accomplishing the compelling governmental interest. In a machine learning setting, decisions are typically not about 'restrictions', so this standard does not identically transfer. We suggest instead evaluating whether the treatments (e.g., an ad campaign) are narrowly tailored to the features of the class. If not, then the discrimination could be inappropriate. For example, if it is determined that an individual prefers a certain type of news article, then it is appropriate to show them that type of article; however, it may be less appropriate to infer the type of loan they should receive.

Algorithmic Evaluation: This factor is difficult to analyze algorithmically, and inapplicable to cases where the features are not interpretable. We recommend that if a treatment fails the previous two factors, then the aggregator require the producer to use interpretable features, with justification for why those features are narrowly tailored to the treatment.

4.5. Correcting Ethically Concerning Treatments

The primary intention behind this framework is to implement our proposed tests algorithmically to automatically flag potentially-discriminatory treatments, and then have a human evaluate each flagged treatment. However, we acknowledge that when dealing with societal-scale applications (like online ads), it may not be practical for a human to look at every flagged treatment.

From an algorithmic perspective, if a treatment is found to be ethically problematic, then assuming that the class and treatment cannot be modified, the easiest corrections to make would be to either (1) address the *Powerlessness* criterion by making the targeted individuals aware that they are being targeted, and allowing them to meaningfully opt out of the treatment, or (2) if applicable, apply competing treatments (e.g., show ads for both high- and low-interest loans: even if the users are not eligible for these loans, simply knowing that they exist may help balance the effect of being shown ads for high-interest loans).

5. Example of Framework Application

Suppose that a payday lender wishes to target ads to a set of individuals, and does not explicitly use features like race, income, family background, etc. Instead, the lender uses a complex combination of which sites an individual visits, the times of day that they are usually active online, and other features. Suppose that this set of individuals has been targeted by other lenders, though perhaps these other lenders identified the class by using different features (and so a **Discriminatory Pattern** exists), but individuals are able to move in and out of the class with regularity (and so the class is **Mutable**). Because there has been a discriminatory pattern, the discrimination should be examined further.

Once this scenario is flagged for review, ideally, a human would determine that this particular targeting may trigger the **Compelling Interest** test, as many payday lenders are known to be predatory and can cause great harm to their customers. To mitigate these problems, targeted individuals should be able to opt out of the treatment (**Powerlessness**) and be shown competing ads for other, more standard loans.

6. Conclusion

In this paper, we have presented a 2-part framework for determining (1) whether a targeted class should be considered suspect, and (2) whether the treatment applied to a suspect class should be considered suspect. Unlike existing work, which typically assumes that certain protected attributes are known and available, we provide tests to determine whether any targeted class should be considered suspect. There is clearly much work left to be done with respect to this framework. Most importantly, we have provided general guidelines, but we must determine specific algorithmic details before the framework can be applied.

References

Ajunwa, Ifeoma, Freidler, Sorelle, Scheidegger, Carlos, and Venkatasubramanian, Suresh. Hiring by algorithm: Predicting and preventing disparate impact. *SSRN*, 2016.

Barocas, Solon and Selbst, Andrew D. Big data's disparate impact. *California Law Review*, 2017.

Barocas, Solon, Bradley, Elizabeth, Honavar, Vasant, and Provost, Foster. Big data, data science, and civil rights. *CoRR*, 2017.

Chouldechova, Alexandra. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Corr*, 2017.

Feldman, Michael, Friedler, Sorelle A., Moeller, John, Scheidegger, Carlos, and Venkatasubramanian, Suresh. Certifying and removing disparate impact. In *KDD*, 2015.

Hardt, Moritz, Price, Eric, and Srebro, Nathan. Equality of opportunity in supervised learning. In *NIPS*, 2017.

Zemel, Rich, Wu, Yu, Swersky, Kevin, Pitassi, Toni, and Dwork, Cynthia. Learning fair representations. *Journal of Machine Learning Research*, 2013.