Model Multiplicity: Opportunities, Concerns, and Solutions

EMILY BLACK, Carnegie Mellon University, USA
MANISH RAGHAVAN, Harvard University, USA
SOLON BAROCAS, Microsoft Research, USA

Recent scholarship has brought attention to the fact that there often exist multiple models for a given prediction task with equal accuracy that differ in their individual-level predictions or aggregate properties. This phenomenon—which we call *model multiplicity*—can introduce a good deal of flexibility into the model selection process, creating a range of exciting opportunities. By demonstrating that there are many different ways of making equally accurate predictions, multiplicity gives practitioners the freedom to prioritize other values in their model selection process without having to abandon their commitment to maximizing accuracy. However, multiplicity also brings to light a concerning truth: model selection on the basis of accuracy alone—the default procedure in many deployment scenarios—fails to consider what might be meaningful differences between equally accurate models with respect to other criteria such as fairness, robustness, and interpretability. Unless these criteria are taken into account explicitly, developers might end up making unnecessary trade-offs or could even mask intentional discrimination. Furthermore, the prospect that there might exist another model of equal accuracy that flips a prediction for a particular individual may lead to a crisis in justifiability: why should an individual be subject to an adverse model outcome if there exists an equally accurate model that treats them more favorably? In this work, we investigate how to take advantage of the flexibility afforded by model multiplicity while addressing the concerns with justifiability that it might raise?

CCS Concepts: • **Social and professional topics → Computing / technology policy**; • **Computing methodologies → Machine learning**; • **Theory of computation → Machine learning theory**; • **General and reference → Evaluation**; **Performance**.

Additional Key Words and Phrases: Model multiplicity, predictive multiplicity, procedural multiplicity, fairness, discrimination, recourse, arbitrariness

ACM Reference Format:
Emily Black, Manish Raghavan, and Solon Barocas. 2022. Model Multiplicity: Opportunities, Concerns, and Solutions. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3531146.3533149

1 INTRODUCTION

How do model makers select which model to deploy for a given prediction task? Even after a practitioner translates a decision into a prediction task (e.g. casting the task of determining who is "creditworthy" as predicting whether applicants are likely to default on a loan), there are myriad decisions made about how to make a model, all of which may influence its ultimate behavior: what model type should be used (from simple linear models to more complex random forests and neural networks); what factors should be considered as inputs to the model; how many times should the model iterate through the training data to learn the patterns therein? How should model makers select between all the possible models that could have been created for a prediction task?

The standard answer to this question is to choose the model that maximizes accuracy. Using maximum accuracy as a decision criterion for model selection may suggest that there is *one* model with the best accuracy, a common assumption

Manuscript submitted to ACM

in the technical literature. However, recent work has reminded us that there are usually multiple models with equivalent accuracy but significantly different properties. For example, Rodolfa et al. [52] have demonstrated that maximally accurate models can produce varying degrees of demographic disparities; D'Amour et al. [23] have shown that models with the same accuracy can be more or less robust; Chen et al. [15] have shown that it is possible to create interpretable models with the same accuracy as neural networks.

We call this phenomenon *model multiplicity*: when models with equivalent accuracy for a certain prediction task differ in terms of their *internals*—which determine a model's decision process—and their predictions. The existence of model multiplicity presents exciting opportunities because it offers model makers the flexibility to prioritize, and optimize for, desirable properties at no cost to accuracy, contrary to some conventional wisdom [7, 16, 28, 42, 57, 60, 62]. The existence of equally accurate models that differ along other axes, including fairness, interpretability, and robustness, allows for model selection to be guided by these other desiderata alongside accuracy. For example, as much recent work in algorithmic fairness has demonstrated, it is often possible to improve the fairness of models with *no* cost to accuracy [19, 28, 52, 61]. Model multiplicity can also improve individual experiences with automated decision making by allowing practitioners to create models that make recourse easier (e.g., by limiting the use of features to only those that are mutable). While the freedom that model multiplicity affords is broad, in this paper we largely focus on its implications with respect to the fairness of a model and the ability for people subject to the model to seek recourse. We also show that model multiplicity has legal implications—which we study in the context of lending—because it places pressure on model developers to search for and adopt the least discriminatory model among those that are equally accurate.

However, along with these benefits comes a potentially surprising revelation: given that there are multiple models for a prediction task with equivalent accuracy, selecting models on the basis of accuracy alone—the default procedure in many deployment scenarios—does not lead to a selection of one unique model best suited for the task. Model selection on the basis of accuracy alone is an underspecified [23] selection process. Unless other considerations are explicitly incorporated into the model development process, model developers selecting models on the basis of accuracy are unlikely to happen upon the model, among all those which are equally accurate, that best addresses those considerations (e.g., minimizes disparate impact).

Further, model multiplicity undermines the justification that we can offer individuals for being subject to any adverse decision process or outcome. Consider the situation where an individual is denied a loan, yet there exists an equally accurate model which would have recommended acceptance. Why must they be subject to the model that rejected them and not an equally accurate, and thus equally viable, one that does not? The fact that such high-stakes decisions may come down to arbitrary choices on the part of model developers may be quite unsettling—and may even conflict with the expectations of the laws that govern such decision making. Thus, while model multiplicity allows for greater choice in the model selection process, it also imposes an additional burden on model developers to put that freedom of choice to good use and to justify how they reach their decisions.

In this paper, we attempt to answer the following question: *how do we take advantage of model multiplicity while addressing its concerning implications?* To do so, we propose a process by which model developers can specify, justify, and document a wider set of behaviors which qualify a model for use in a specific context to guide the model selection process. Concretely, we present three main contributions: (1) a principled understanding of the relationships between multiplicity, accuracy, and variance, providing intuition for why multiplicity may actually *increase* with accuracy, backed by theoretical results deferred to Appendix A; (2) connections between the technical aspects of model multiplicity and their legal implications; and (3) a set of policy recommendations for how to take advantage of model multiplicity while addressing the concerns it raises. Ultimately, we hope that the explicit recognition of model multiplicity, along with legal

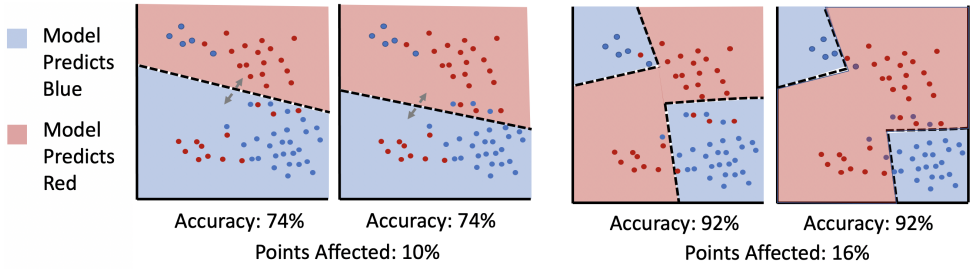

Fig. 1. A stylized graphic displaying how models with higher accuracy can actually lead to *more* model multiplicity: On the left, we show a simple linear model on some data, with accuracy of approximately 75%. On the right, we show a more complex model which fits the data better and reaches 92% accuracy. In order to achieve a better fit to the data, the more complex model has a more complex decision surface. In having a more complex decision surface, there are more opportunities for shifts in decision surface to take place in reaction to changes in the training process, and thus there are more points in the distribution that are susceptible to a change in prediction.

requirements preventing discrimination, will lead policy makers and practitioners to hold models to a higher standard on axes beyond accuracy—and restore the justifiability of model decisions.

The rest of this paper proceeds as follows: in Section 2, we provide an overview of model multiplicity and situate it in the existing literature. Section 3 explores the relationship between multiplicity and accuracy, connecting multiplicity to standard ideas from machine learning theory. Sections 4 and 5 articulate the potential benefits and harms respectively of model multiplicity, drawing connections to the law. In Section 6, we provide recommendations for a model development process that explicitly accounts for multiplicity.

2 DEFINING MULTIPLICITY

Model multiplicity occurs when models with equivalent *accuracy* for a certain prediction task differ in terms of their internals—i.e. the decision surface of the a model—[1] or their predictions. In this section, we define model multiplicity in more detail, beginning by describing the setting in which we consider model multiplicity, our definition of model accuracy, key terms for the paper, and, finally, the definitions of the components of model multiplicity: *procedural* and *predictive* multiplicity.

2.1 Preliminary Definitions

Setting. In this paper, we focus on classification models, although the main insights of this work apply to the regression setting as well. A classification model predicts to which class, or category, an input x belongs, from some pre-set collection of categories. For example, predicting whether an individual will default on their loan is a classification task. Classification models have *decision surfaces* which delineate between different classes in the model's input space (see Figure 1). We will focus on the case where there are only two classes (also known as binary classification).

Accuracy. Broadly, accuracy is a measure of how well a model's predictions match the underlying labels in the data. Importantly, model developers cannot know how accurate the model is on all possible model inputs (e.g., over all possible loan applicants); accuracy must be estimated on available data. In practice, there are a variety of measures of accuracy; for simplicity, we will take accuracy to mean the fraction of predictions for which the model is correct. When we refer to several models exhibiting *equivalent* accuracy, they may not have exactly the same accuracy, but accuracy that is functionally indistinguishable (e.g., an accuracy of 97.8989 and 97.8990).[2]

[1] Intuitively, model internals, or the decision surface, can be understood to fully specify the *process* by which a model makes its decisions.
[2] We note that what levels of accuracy are functionally indistinguishable may depend on the context in which the model is used.

2.2 Procedural and Predictive Multiplicity

Model multiplicity describes how models for a given prediction task can differ even when they exhibit equal accuracy. We draw attention to two ways in which models differ despite equal accuracy: in their internals, or *procedural multiplicity*, and in their predictions, or *predictive multiplicity*.

Procedural Multiplicity. Procedural multiplicity refers to the phenomenon where several models for a given prediction task have equivalent accuracy, yet differ in their model internals. More technically, procedural multiplicity occurs when models which have the same accuracy exhibit some difference in their decision surface, as this changes the way in which a model's inputs are combined to reach a conclusion. In other words, procedural multiplicity describes the situation where models of equal accuracy differ in the *process* by which they reach a given prediction. One example of a difference in the model's internals is the use of various input *features* into a model's decision for a given prediction: for example, one model may use gender as a feature to make loan granting decisions; another may not. Another example of a difference in model internals is a difference in *model class*. For example, a random forest model and a linear model may have equivalent accuracy for a certain task, but likely vary in the way they reach each prediction. One way procedural multiplicity can become apparent to model subjects is when equally accurate models produce qualitatively different explanations for the same decision. In one example from Anders et al., two credit scoring models make the exact same predictions on every point, but one model justifies its decision on the basis of gender, while the other relies on income and tax payments [4].

Predictive Multiplicity. Predictive multiplicity refers to the phenomenon where models with equivalent accuracy for a certain task differ in their predictions (i.e., two models predict different classes for the same input). Predictive multiplicity, like accuracy, is measured on the labeled data available to a model developer: given a prediction task, models that exhibit predictive multiplicity have equal accuracy but predict different classes for some data points in the training or test set.[3] Thus, model developers cannot measure the full extent of disagreement between any two models, but can only estimate it based on available data.

Relationship Between Procedural and Predictive Multiplicity. Note that differences in model predictions on a certain input require differences in decision boundaries, implying predictive multiplicity is a special case of procedural multiplicity. The converse does not hold: two models with the same prediction on a given point may still exhibit variation in the process by which that outcome was reached [4, 11]. However, we draw attention to predictive multiplicity on its own due to its unique normative and legal implications. Throughout this paper, when we refer to procedural multiplicity, we refer to the aspects of procedural multiplicity that occur even in the absence of (observed) predictive multiplicity: that is, models with equal accuracy with different decision processes that do not necessarily manifest in different predictions on the available data. Of course, any change to a model's decision boundary, and thus any two models exhibiting procedural multiplicity, will differ on *some* potential input point; but if no such input is present in the data, then this difference will not result in observable predictive multiplicity.

2.3 Sources of Multiplicity

When creating a model for a given learning problem, every decision point a model developer faces along the model building pipeline serves as a fork, where each potential choice may lead to multiplicitous models. In the context of this paper, we define a learning problem to be the prediction of a pre-defined target. While there may be further multiplicity-like problems stemming from the various ways that a nebulous real-world goal may be translated into predicting a specific

[3]There is disagreement in the literature on this definition: for example, Marx et al. [40] define predictive multiplicity only on a model's training set.

target [44], we view these as out of scope for this paper. However, all modeling decisions made once the prediction target is set are within-scope and possible sources of model multiplicity.

Decisions that can result in multiplicity include choosing what information should be included as input to the model [26], which points are included in the training set [10], which model class should be adopted [15], what random numbers the model's parameters are initialized with [10, 41], among many others [40]. Through these choices, the model developer creates one model, but each other choice they could have taken may have lead to a model that would have performed with similar accuracy. In theory, the sources of multiplicity are infinite, as there are infinite possible modeling choices. In practice, however, the range of choices is restricted by practical (including budgetary) constraints.

2.4 Aggregate and Individual Effects

At a high level, model multiplicity can result in differences between models at the *aggregate* level or at the *individual* level. By aggregate effects, we refer to differences in global model properties between multiplicitous models (e.g., satisfaction of group-level fairness criteria (such as equal selection rates across different demographic groups)). By individual effects, we refer to the way in which differences between models of equal accuracy impact individuals' experience with the model, including differences in individual predictions or explanations of those predictions. Aggregate and individual effects are not disjoint categories of model behavior, as some forms of model multiplicity may impact both aggregate-level and individual-level outcomes. Often, however, individual effects do not manifest at the aggregate level. For example, differences in individual predictions may not impact the overall treatment of any demographic group. We therefore find that making these two perspectives explicit helps to better understand the overall impacts of multiplicity.

2.5 Arbitrariness Versus Randomness

In this work, we draw a distinction between arbitrariness and randomness in selection processes. By an arbitrary selection process, we mean a completely unconsidered decision—one that is made without thought or perhaps even without knowledge that a choice was being made. By a random selection process, we mean a decision which is *purposefully* left to chance. We draw this distinction to stress that a random selection process is predicated on a conscious choice to employ this selection method: as Perry and Zarsky [46] write, "the decision to opt for chance must be reasoned."[4]

2.6 Related Work

Model multiplicity has been recognized in the machine learning literature, though not always under the same name, starting with Breiamn's characterization of the "Rashomon Effect" [13]. For example, Dong and Rudin [26] and Fisher et al. [32] demonstrate procedural multiplicity in feature importance, showing that models relying on different sets of features can reach the same accuracy; Black et al. [11] and Mehrer et al. [41] have shown that deep models with similar accuracies relying on the *same features* may still combine those features in different ways to reach a given output. Recent studies also provide evidence for predictive multiplicity: Marx et al. [40], who introduced the term, focus on its effects at the individual level (equally accurate models can make different predictions for individuals), while others have demonstrated its effects at the aggregate level (equally accurate models can have different properties, including fairness and robustness) [23, 51]. A recent line of work has sought to quantify and mitigate model multiplicity in a variety of settings [10–12, 20, 45, 49, 55]. Our work builds upon and synthesizes this technical foundation to understand the relationship between model multiplicity, complexity, and error, as discussed in Section 3, and to relate the wide range of effects of model multiplicity to the law.

[4] As Perry and Zarsky [46] describe in their work, there are many situations where random (not arbitrary) selection is justifiable: for example, allocating a scarce, indivisible resource among many with equally strong claims—such as allocating public housing among equally needy applicants.

Some legal scholars have also begun to consider the possibility of model multiplicity and its implications, though this discussion is largely focused around predictive, and not procedural, multiplicity. Kim [35] has argued that predictive multiplicity means that certain interventions aimed a reducing disparate impact "do not require special legal justification" as the lack of one "correct" model means that there is "no clear baseline" against which any departures might be challenged. Kim points out that it simply does not make sense to say that someone has been unfairly denied a job that they would have otherwise secured if not for the attempt to reduce disparate impact because there is nothing that entitles anyone to having a particular model chosen over an equally accurate alternative. On this account, multiplicity provides developers with the freedom to choose the model among those with equal accuracy that exhibits the least disparate impact without raising concerns with disparate treatment. However, this work does not address the concerns that model multiplicity may raise. Creel and Hellman [22] briefly note the unsettling implications of predictive multiplicity with respect to arbitrariness in algorithmic decision making, but ultimately argue that arbitrariness is only a problem in algorithmic decision making when there is an algorithmic monoculture that locks an individual out from certain opportunities across the board (e.g., when all lenders use the same algorithm and thus all reach similarly adverse decisions for a particular individual). Contra Creel and Hellman [22], many legal scholars have been calling for legal protections, inspired by due process principles and practices, to address the potential arbitrariness of algorithmic decision making more generally, even in the absence of an algorithmic monoculture [17, 18, 21]. In contrast to prior work, we address both the benefits and the concerns of procedural and predictive multiplicity, and we provide concrete recommendations for how to take advantage of the benefits of model multiplicity in practice, without falling prey to the concerns that it might provoke.

3 ACCURACY AND MODEL MULTIPLICITY

By default, accuracy is the primary measure by which machine learning systems are evaluated. This focus is pervasive throughout machine learning scholarship and practice [8], perhaps best evidenced by the Common Task Framework [27], through which independent researchers compare predictive performance on common datasets. But accuracy plays a larger role in model development than evaluation alone: accuracy is typically the main or sole criterion used for model *selection*. When deciding which of many possible models to deploy, a practitioner will often choose the most accurate one.

The idea that model selection can be reduced to accuracy-maximization rests on a pair of premises: that accuracy is the primary measure of how "good" a model is, and that, for a given task, models that maximize accuracy do not differ meaningfully from one another. In other words, if accuracy-maximization leads to a unique or near-unique optimal model, then no other criteria need be used in model selection. Even if we accept that accuracy should be the primary evaluation criterion (setting aside, for now, properties like fairness, robustness, and interpretability that might be perceived as crucial to model performance in practice), evidence suggests that accuracy-maximizing models are not unique [10–12, 23, 40, 45, 51]. And yet, the intuition that there exists a unique "correct" model, and that accuracy-maximization should ultimately discover it, remains pervasive [35, 39, 50]. In what follows, we trace the roots of this intuition and offer a theoretical basis for why, as machine learning becomes more sophisticated, we should expect accuracy-maximization to yield *more* multiplicity, rather than less. As a result, accuracy is an incomplete basis for model selection. We focus here on predictive multiplicity, though it may be possible to derive analogous results for procedural multiplicity as well.

Does accuracy-maximization reduce predictive multiplicity? Our intuition that accuracy-maximization should lead to little or no predictive multiplicity comes from the idea that there exists a single "best" or "correct" predictor (known as the Bayes optimal predictor [56]), and increasingly sophisticated models will converge to this optimal predictor. In general, Bayes optimal models are unique, and it may be tempting to apply this intuition more broadly: we might believe that even when our models aren't Bayes-optimal, the maximally accurate model for a given dataset is near-unique. We can make this

idea rigorous: Theorem A.2, included in the appendix, demonstrates as the error of a model approaches that of the Bayes optimal predictor, the model must approach the Bayes optimal predictor.[5] In other words, as models get more accurate, they must converge to one another in the limit. Results like these can lead to a slippage in intuition—Bayes optimal predictors are unique, so the best predictor we can build should also be unique. And yet, empirical evidence seems to suggest the opposite: developing more accurate models can often lead to *more* multiplicity (see Figure 1 for an example) [10, 41]. While this might appear to contradict Theorem A.2, in reality, models are sufficiently far from Bayes optimal, leaving plenty of room for multiplicity. To derive a more nuanced view, we turn to standard bias-variance decompositions of error.

Bias, variance, and multiplicity. Conceptually, errors in machine learning systems come from three sources: bias, variance, and irreducible noise [25, 33]. This decomposition helps us understand fundamental trade-offs in machine learning: more expressive and sophisticated machine learning techniques (such as deep learning) have less bias because the average model can more accurately approximate the Bayes optimal predictor than less expressive techniques (such as linear regression); but this increased expressivity comes at the cost of high variance, since any particular model is much more sensitive to random choices in the model development pipeline. Crucially, as Theorem A.3 shows, multiplicity is tightly related to variance. To the extent that increased accuracy is achieved through increased model complexity (and therefore variance), we should therefore expect to see *more* predictive multiplicity, as noted in Corollary A.4. Thus, accuracy is not an antidote to multiplicity, and model selection cannot simply be reduced to accuracy-maximization. Instead, we must explicitly consider and deal with multiplicity, beginning with an understanding of the benefits and challenges it brings.

4 OPPORTUNITIES

By shattering the intuition that there is *one* most accurate—and therefore correct—model, multiplicity can introduce much more freedom into the model selection process. On the aggregate level, this means that model developers can express preferences over values beyond accuracy *at no cost to accuracy*, including with respect to properties like fairness, robustness, and interpretability, among others. This same flexibility manifests on an individual level: to illustrate this point, we focus on the ability to improve the recourse available to the individuals subject to a model's adverse decisions. This section will consider the benefits at both levels.

4.1 Aggregate Benefits: Flexibility

By demonstrating that there are many different ways of making equally accurate predictions, multiplicity gives model developers the flexibility to prioritize other values in their model selection process without having to abandon their commitment to maximizing accuracy. While this benefit is broad, we focus in particular on its implications for fairness. In fact, as we'll discuss in this section, the flexibility afforded by multiplicity is particularly relevant to the law because it creates legal pressure for model developers to reduce avoidable disparate impact in their deployed models. We demonstrate this flexibility—and its connections to the law—through both procedural and predictive multiplicity.

Procedural Multiplicity. Model developers can leverage procedural multiplicity to ensure that a model has desirable model internals without sacrificing accuracy. As shown in prior work, model developers might exploit procedural multiplicity to select a model class that is more robust or interpretable than other model classes of equal accuracy [23, 53]. This is far from an exhaustive list, as procedural multiplicity creates the possibility for *any* quality of a model's decision process to be prioritized at potentially no cost to accuracy. However, in the context of fairness, the possibility that replacing or removing certain features from a model may not affect its accuracy is particularly relevant. If there are certain features that

[5] This result holds as long as data points are more predictable than 50-50 coin flips.

are perceived as a normatively objectionable basis for decision making, procedural multiplicity suggests—and research has demonstrated empirically [9, 26]—that model developers can remove these from their models while still potentially achieving the same level of accuracy in their predictions. For example, features may be normatively objectionable because they are protected attributes such as race or sex or because they are proxies for such attributes, such as zip code. Discrimination law imposes exactly these kinds of constraints on model developers in certain regulated domains via a prohibition on so-called "disparate treatment." For example, the Equal Credit Opportunity Act (ECOA) prohibits the consideration of race, sex, age, and a number of other legally protected attributes in lending decisions [2, 30]. Thus, lenders using machine learning to develop credit scoring models are understood to be legally prohibited from including these features in their models. While this prohibition is designed to prevent lenders from relying on features that have served as the basis for discriminatory decision making in the past, it is also designed to encourage lenders to find other features that serve their goals at least as well. Procedural multiplicity demonstrates that it may be technically possible to do so, putting to bed the idea that there is only ever one set of features that would allow model developers to achieve some level of accuracy in their decision making.

Predictive Multiplicity. While procedural multiplicity gives model developers the flexibility to incorporate their normative preferences into the model's decision-making process, predictive multiplicity allows model developers to impose their preferences on the model's predictions—potentially without impacting accuracy. In the context of fairness, predictive multiplicity creates the possibility to minimize differences in prediction-based metrics across groups, notably differential validity (i.e., differences in accuracy) and disparate impact (i.e., differences in model predictions). Rodolfa et al. [51] show that this is possible in practice across a wide range of real-world applications, including such high-stakes domains as criminal justice, housing, and education. Similarly, algorithmic hiring companies such as HireVue require that their models return similar distributions of predictions across demographic groups, and claim that this has little impact on predictive accuracy [38].

This aspect of predictive multiplicity speaks directly to so-called "disparate impact" doctrine, which imposes liability on model developers for avoidable disparities in the rate at which members of legally protected groups obtain the desired outcome from a decision-making process. As the official commentary on ECOA states, the law "prohibit[s] a creditor practice that is discriminatory in effect because it has a disproportionately negative impact on a prohibited basis, even though the creditor has no intent to discriminate and the practice appears neutral on its face, unless the creditor practice meets a legitimate business need that cannot reasonably be achieved as well by means that are less disparate in their impact" [2].[6] To appreciate what this means in the context of a lender employing machine learning, imagine that the "creditor practice" in question is the use of a machine learning model developed to predict default and that the lender's primary "business need" is predicting default as accurately as possible so as to make appropriate lending decisions. The official commentary suggests that even when machine learning has been adopted for this purpose, involves no legally proscribed features, and demonstrates a high degree of accuracy, lenders still face liability if they fail to adopt whatever alternative "means" might exist for achieving their same goal but with smaller disparities in outcomes across legally protected groups. Predictive multiplicity suggests that there may exist such alternative "means" because there may be a different model of equivalent accuracy that generates less disparate impact. The disparate impact doctrine can thus be interpreted to say that predictive multiplicity creates legal risk for those who fail to adopt the least discriminatory model among those that are equally accurate [47, 48].

[6] While disparate impact is not written directly into ECOA, the formal guidance suggests that it is understood to apply under the law.

4.2 Individual Benefits: Improved Possibilities for Recourse

Multiplicity can also provide the flexibility necessary to improve individuals' experience of model procedures and their outcomes. To illustrate this point in the context of fairness, we explain how procedural and predictive multiplicity can improve a person's capacity to achieve recourse—that is, to obtain a more desirable outcome after receiving an adverse decision.

Procedural Multiplicity. Recent scholarship has suggested that one of the important functions of providing explanations of model decisions is to help people subject to an adverse decision understand how they might obtain a more favorable prediction in the future [59]. In the United States, the Fair Credit Reporting Act (FCRA) and ECOA both require that lenders explain their decisions to consumers who were unsuccessful in their applications for credit [30, 31]. Both laws compel lenders to provide so-called "adverse action notices" that state the "principle reasons" for adverse decisions, on the belief that doing so may help consumers more effectively navigate the process of obtaining credit in the future [6, 54]. Scholars have suggested that lenders might comply with these requirements by offering counterfactual explanations that point out, for example, that an applicant would have been successful if their annual income had been $10,000 higher [58]. In light of such an explanation, the consumer might look for ways to increase their income and then reapply for a loan. However, as prior work has pointed out, such explanations may not facilitate recourse if the highlighted factors are immutable and thus cannot be acted upon by the consumer [58]. Explanations only facilitate recourse if they suggest changes to features that consumers can actually execute in practice. This insight has motivated a good deal of recent research focused on developing methods to produce explanations that suggest viable and efficient paths to future success [34]. Procedural multiplicity suggests that there is another—and more direct—way to achieve these same goals. Rather than searching for different possible explanations of the decisions of a fixed model that would be easiest for consumers to act upon (the current focus in the literature on recourse), model developers could exploit procedural multiplicity to find models that exhibit the same degree of accuracy but differ in the degree to which they rely on features known to be difficult or impossible for people to change. Thus, procedural multiplicity gives model developers a way to take recourse into account in the model development process, not just in deciding which techniques to rely on when explaining a model's decisions.

Predictive Multiplicity. As algorithms have been adopted in a growing range of high-stakes decisions, scholars have begun to worry about the possible harms of an *algorithmic monoculture* [22, 36]. For example, if several lenders all converge on one credit scoring model (and thus on the same predictions of default for each applicant), consumers who were rejected by one lender may find that they have no better luck when they submit an application to other lenders. This, too, is a problem of recourse, but at the level of an entire domain of decision making, rather than at the level of a model. Predictive multiplicity may serve as a natural bulwark against this worrisome possibility: even if lenders all maximize prediction accuracy, they may still end up with models that produce different individual-level predictions. The perhaps surprising benefit of predictive multiplicity is that, even when models are selected on the basis of accuracy alone, there will be inherent heterogeneity in the models selected by different firms [22].

5 CONCERNS

While procedural and predictive multiplicity gives us the flexibility to prioritize values beyond accuracy, this very same flexibility can be cause for serious concern. The fact that we can choose among many possible models with equivalent accuracy can lead to problems of underspecification and to arbitrariness in decision making. Selecting models on the basis of accuracy alone can obfuscate large differences between multiplicitous models that we might actually care about, but have failed to explicitly integrate into the set of considerations that go into the model development process. Perhaps even more importantly, model multiplicity also means that accuracy alone is an insufficient justification for why one

model was chosen over another equally viable (i.e., accurate) alternative. In this section, we consider the concerning implications of model multiplicity and how the law bears on some of these concerns.

5.1 Aggregate Concerns: Underspecification

As we've shown, model multiplicity gives model developers the option to prioritize values beyond accuracy, since models with equal accuracy can have quite different aggregate- and individual-level effects. This also means, however, that failing to consider what other behaviors may be desired, and continuing to choose models on the basis of accuracy alone, leaves model behavior on axes other than accuracy up to an arbitrary choice: without explicitly specifying what behaviors a model should exhibit—such as fairness, robustness, and interpretability—and optimizing for them, it is unlikely that a model will naturally exhibit such behaviors. D'Amour et al. [23] call this the problem of underspecification.[7] Underspecification reveals that we need to make our desired model properties explicit if we want our models to exhibit them.

Procedural Multiplicity. Procedural multiplicity can give rise to three rather serious problems. First, as mentioned, selecting a model on the basis of accuracy does not guarantee that it will exhibit other desirable properties. Second, because it may be possible for models with different internals to still generate the same set of predictions, changes made to the internals of a model may not have the anticipated effect on predictions. Third, procedural multiplicity can be leveraged to remove anything from the decision-making process that would raise legal or normative concerns (e.g., legally protected or otherwise controversial features), while preserving a troubling, but avoidable, outcome (e.g., disparate impact). We focus on the second two concerns.

First, procedural multiplicity means that removing features proscribed by discrimination law may do nothing to reduce disparities in predictions, which may have been the explicit intent of such an intervention. As discussed earlier, discrimination law imposes strict prohibitions on the use of certain characteristics in decision making across a range of high-stakes domains, including lending. These prohibitions on "disparate treatment" were put in place to protect people who possess these characteristics from systematically worse treatment than others (hence the term "protected characteristics"). Procedural multiplicity undermines these protections because it opens up the possibility that people with these characteristics might be subject to the same disfavorable predictions without directly considering these characteristics [24]. While disparate impact doctrine has developed, in part, in recognition of the potentially limited efficacy on prohibitions on disparate treatment [5]—placing demands on decision makers to be able to justify disparities in model predictions, even if they haven't considered any protected characteristics—calls for procedural interventions remain commonplace. For example, Black et al. [9] observe this phenomenon in debates about the design and use of risk assessment tools in the criminal justice system, where procedural interventions recommended by experts and advocates, such as removing nonviolent arrests from the criminal history considered by the tools, seem to be suggested with the expectation that they will reduce racial disparities in tools' predictions. Procedural multiplicity means that there is no guarantee that these changes will have the desired effects on model predictions.

Second, given that there might be many ways to develop a model that generates the same predictions, developers could search for models that seem to be more palatable from a procedural perspective (e.g., because they don't involve legally proscribed or otherwise controversial features) but display the same worrisome predictive behavior. Objecting to these predictions might be more difficult when the process that generates them seems benign or perhaps even desirable. This is not just a hypothetical concern; recent work has shown that it is possible to create two models with *exactly the same predictions* that rely on *completely different features to make up their decision* [4, 11]. This suggests that not only

[7] We note that there is a subtle difference between *underspecification*, where a model developer fails to fully articulate and incorporate their full set of behavioral desiderata into the model building process (as D'Amour et al. [23] show in the case of model robustness) and *mis-specification*, where a model developer chooses the wrong target to optimize for (as Obermeyer et al. [43] demonstrate in a healthcare system's choice to use healthcare costs as a proxy for healthcare needs).

might procedural interventions fail to have their intended effects on predictions, but that procedural multiplicity can be exploited adversarially to develop a compelling justification for whatever disparities in predictions that model developers might like to preserve. While this possibility, often referred to as *proxy discrimination* [24], is well-studied in the literature, we note that it is a result of procedural multiplicity.

Taken together, these observations about procedural multiplicity highlight the need for model developers—or those seeking to influence or regulate their choices—to fully specify the kinds of predictions that they would like models to generate. Unless these are optimized for explicitly, there is no reason why maximizing accuracy or making procedural interventions will lead to the desired model behavior.

Predictive Multiplicity. The reality of predictive multiplicity highlights that model selection on the basis of accuracy does not guarantee the desired prediction-based behaviors beyond accuracy. Specifically, in the context of fairness, predictive multiplicity tells us that there may be several equally accurate models that each vary in the degree to which accuracy, selection rates, or other fairness metrics differ across groups. Unless this is made an explicit consideration in the model development process, the chosen model can be an arbitrarily bad pick with respect to fairness metrics among those that are all equally accurate.

5.2 Individual Level Concerns: Loss of Justifiability

Model multiplicity also creates serious challenges for justifying the ultimate choice of model, given that different choices can result in more or less favorable situations and predictions for any given individual. Globally, this raises a fundamental question: what justification is there for subjecting a particular person to an adverse *model procedure* or *model prediction* if that person would have received more favorable treatment under a different, but equally accurate model? This section will consider the crisis of justifiability brought about by both procedural and predictive multiplicity and again discuss how the law bears on this challenge.

Procedural Multiplicity. As discussed, procedural multiplicity admits the possibility of creating models with very different internals, even if they all exhibit the same degree of accuracy and all result in the same predictions. This increased flexibility, however, leads to a difficulty in justifying why a particular way of reaching the prediction is necessary. We again focus on the example of recourse: we previously suggested that predictive multiplicity is desirable when it allows developers to favor models with internals that would make recourse easier (e.g., selecting models with features that people would find less challenging to change). Yet, for any given individual, there might exist an alternative model with identical predictions that would have given the individual an easier path to recourse. Consider a scenario in which a lender offers an explanation for an adverse decision that an applicant for credit would find challenging to act on. Even if the applicant accepts that this is a valid explanation for their adverse prediction and the easiest of all possible explanations for the applicant to act on, the applicant might nevertheless ask: why did the lender choose the model that makes recourse more difficult for me instead of the model that would have made recourse easier for me, given that both would have resulted in the same predictions? The applicant might ask more generally: why must I be subject to this model rather than the other? Procedural multiplicity makes it challenging to answer these questions because accuracy alone cannot justify the ultimate choice of model.

Predictive Multiplicity. Predictive multiplicity can be just as unsettling when it comes to the justifiability of decisions because individuals might receive favorable predictions under some models and unfavorable predictions under others, even when all of these models are equally accurate. To illustrate this point, consider a situation in which there are two models that exhibit the same accuracy, but only one of which would instruct a lender to grant an applicant's request for credit. If the lender happens to choose the one that denies the applicant's request, how would the lender justify its adverse decision, given that the lender could have just as easily chosen the other model? This line of questioning is unsettling

because it reveals that choosing a model based on accuracy alone is akin to choosing arbitrarily between more or less favorable predictions for certain individuals.

The disquieting prospect that consumers' access to credit might rest on decisions made without adequate care was one of the main concerns that motivated the passage of FCRA and ECOA, both of which target arbitrariness in lending decisions. The legislative record suggests the FCRA was designed to "protect consumers from inaccurate or arbitrary information in a consumer report which is being used as a factor in determining an individual's eligibility for credit, insurance, or employment" [1]. It seeks to do this by requiring that lenders adopt reasonable procedures to ensure the "accuracy, relevancy, and proper utilization" of the information in credit reports. In regulating the information that goes into high-stakes decision making, FCRA seems to be designed to guard against capricious, sloppy, and otherwise faulty decision making. As discussed earlier, ECOA requires lenders facing a disparate impact claim to demonstrate that "the creditor practice meets a legitimate business need"; in practice, this is often accomplished by demonstrating that their credit scoring models reasonably accurately predict default. In other words, absent some justification for assessing applicants for credit in a manner that generates a disparate impact, lenders will be found liable for discrimination. Finally, both FRCA and ECOA require lenders to provide adverse action notices, on the belief that having to justify their decisions will cause lenders to be less arbitrary in their decision making [54]. Note that lenders are only required to justify their particular way of making decisions when they face a disparate impact charge. Absent any identified disparate impact, FCRA and ECOA only require that lenders provide an explanation for any particular decision, not a justification for the manner in which they make decisions. Yet it is possible to interpret this more modest requirement as an *indirect* way of trying to ensure that there are good justifications for why lenders make decisions the way that they do. For example, if the proffered reason for an adverse decision is something that seems to lack face validity as a predictor of default, then consumers might question whether the basis for decision making is well justified (namely because it seems unlikely that predictions of default on that basis would be accurate) [54]. These laws are obviously both premised on the idea that there should be good reasons for the manner in which lenders go about making their highly-consequential decisions.[8] The problem with predictive multiplicity is that it makes avoiding arbitrariness difficult even when lenders seek maximally accurate predictions.

Accuracy has traditionally provided a justification for model selection because it was assumed that there must be one unique model of maximally achievable accuracy. If selecting on the basis of accuracy leaves model developers with only one choice, then, according to this thinking, the ultimate choice must be justified. Multiplicity reveals this assumption to be false. While we might welcome the fact that selecting models on the basis of accuracy does not limit developers' choices to just one option, we should also recognize the threat that it poses to the justifications that we can now offer for the ultimate choice of model. Just as selecting on the basis of accuracy does not entitle anyone to a specific prediction [35], selecting on the basis of accuracy need not condemn anyone to a specific prediction. Whatever the chosen model, there always exists an alternative model of equal accuracy that would reverse an individual's prediction.[9] And any given individual might ask: why was one model chosen over the other? Model multiplicity means that we have lost a fundamental basis for justification that needs to be replaced.

This is well reflected in the worries expressed by Citron and Pasquale [18] when they point to a "a study of 500,000 files [in which] 29% of consumers had credit scores that differed by at least 50 points between the three credit bureaus." They argue that "[b]arring some undisclosed, divergent aims of the bureaus, these variations suggest a substantial proportion of

[8]While Creel and Hellman [22] suggest that arbitrariness in decision making is only a problem when there is no alternative decision maker to whom a person can turn after receiving an adverse decision, these legal requirements seem to be designed to guard against arbitrariness in the decision making of private actors whether or not there are alternatives in the marketplace.

[9]In theory, such a model always exists; whether one could reasonably be found in practice depends on both the model developer's choices and the individual in question.

arbitrary assessments" [18]. If we assume that the three credit bureaus all have access to similar information, that they are all seeking to predict default, and that they each have the means to achieve similar accuracy in their predictions, then much of the resulting divergence in scores for particular individuals is likely the result of predictive multiplicity. Rather than accepting this as an unavoidable or even desirable effect of the heterogeneity naturally engendered by predictive multiplicity, Citron and Pasquale argue that the divergence is evidence of arbitrariness, on the likely belief that if the bureaus had good reasons for choosing their credit scoring models, the models would not return different predictions. Accuracy is no longer a sufficiently good reason because selecting models on that basis cannot supply one correct answer; there now remains an unaddressed degree of arbitrariness. In a perhaps surprising reversal, what we described earlier as a welcome guard against algorithmic monoculture is here presented as a threat to justifiability: why must any individual be subject to the chosen model when an equally accurate alternative exits that would have given the individual a more desirable prediction?

In order to recover justifiability of model decisions, accuracy can no longer be used as the reason why a particular model was chosen in high-stakes applications. There must be additional criteria used to determine whether a model performs sufficiently well for high-stakes deployment, and why one model—and therefore its decision process and predictions—should have been chosen over an equally accurate alternative.

6 SOLUTIONS

The problems arising from model multiplicity underscore the need for a more careful model selection process that explicitly takes multiplicity into account. A core component of the risks imposed by model multiplicity is that there is no *thought* given to the selection of the model among apparently equally viable choices: the selection is arbitrary, as it occurs without admission or even knowledge that a choice is being made. In order to take advantage of model multiplicity while making justifiable model decisions, we must create a non-arbitrary method of choosing between high-accuracy models that specifies the behaviors we wish to see in the model, and documents the reasoning behind the choices made. Towards this goal, we can make explicit and justify a set of criteria for acceptable model procedural and predictive behavior beyond accuracy alone, document these criteria and justifications, and then only consider models that meet these criteria. We call this set of criteria the *meta-rule*. As there may still be multiplicitous models that all satisfy the meta-rule, we suggest ways to further choose between models with differing individual predictions to prevent arbitrariness that satisfy a given meta-rule via various prediction aggregation techniques. Importantly, all of these choices—the meta-rule and the aggregation technique—must be documented and justified. This, ultimately, serves as a justification for why an individual is subject to a certain model decision.

6.1 Meta-Rules

Using a meta-rule provides a reasoned way to choose amongst multiplicitious models: the explicit consideration of what model behaviors make up the meta-rule may provide model developers greater clarity on how to optimize for these behaviors during the model building process, preventing issues of underspecification discussed in Section 5.1. For example, for a loan prediction model, a meta-rule may be: the model must have over 95% accuracy, rely on features only available in the individual's recent banking activity, and have near-equal true positive rate across demographic groups. Moreover, the documentation of these decisions and the reasoning behind them can serve as a justification for the model.

The restriction to only consider models that satisfy all criteria of a meta-rule reduces the set of multiplicitious models which all reach similar accuracy on a given prediction task to a smaller set—which, crucially, all satisfy certain specifications for what it means to be an acceptable model in a given context. Importantly, a meta-rule should specify the *actual behaviors desired*: if minimal racial disparity is preferable subject to maximal accuracy, this should be enforced through an explicit outcomes-based constraint, rather than a procedural constraint that stakeholders may expect to reach such an outcome.

A meta-rule should be deliberated over and documented, with justifications for each qualification on the model. Put together, the explanations of the desiderata within a meta-rule constitute the justification behind a decision from a model that satisfies such desiderata. This is because the meta-rule compels model developers to *explicitly* consider the differences that may exist between multiplicitous models and decide on the criteria that are relevant to the application that would disqualify a model (even with high accuracy), instead of choosing arbitrarily.

In practice, the ability to explore the space of equally accurate models in order to find one which satisfies a meta-rule may be constrained by the model developer's ability to experiment with different design choices, which in turn may be influenced by restrictions on the amount of time and money that they can spend on the exploration process. Following Selbst and Barocas [54], the meta-rule should also document the practical constraints that developers face (e.g. available funding, available talent, available data, etc.) in their model selection process, as this ultimately influences the breadth with which they may search for multiplicitous models. Doing so would help to justify the choice of model among a potentially infinite set of alternatives, while also providing the necessary information for others (e.g., the person subject to the decision, an auditor, a regulator, etc.) to assess whether the efforts to find more desirable alternatives were reasonably exhaustive under these constraints.

Further, the very question of how to search among equally accurate models is only beginning to be addressed by the research literature. While some dimensions of exploration may be costly, such as collecting more data to explore alternate features to include in a model, the most common method of exploration—hyperparameter variation [40, 51]—is already standard machine learning practice. Whereas model developers currently explore a range of possible models through hyperparameter tuning and select one that maximizes accuracy, a meta-rule would require that the model developer maintain a set of models that satisfy the meta-rule.

In theory, and often even in practice [52], it is unlikely there is only *one* model which satisfies a meta-rule. As we discuss in the next section, we can account for residual differences in prediction between models which satisfy the meta-rule with model aggregation techniques.

6.2 Aggregation Techniques

Given a set of models that all satisfy the meta-rule, how might a decision maker choose among models? In fact, there are several ways to produce a single model from a set of equally "good" models. Here, we focus on three such techniques, and, in particular, we demonstrate that each may be appropriate for use in different contexts. Let $\mathcal{M}$ be a distribution over models that satisfy decision makers' meta-rule.[10] The techniques that follow require that the model developer can construct a *random sample* from $\mathcal{M}$, as opposed to enumerating all of the models in $\mathcal{M}$. The three aggregation techniques we consider are **mode aggregation**, **randomized predictions**, and **random model selection**. Importantly, these techniques help to restore justifiability because they each involve deliberating over how to choose between multiplicitious models.

- **Mode aggregation** [11]: The mode predictor $\overline{m}$ aggregates models from the model distribution $\mathcal{M}$ by outputting the majority vote over the models $m \in \mathcal{M}$ for each example x. Formally, in the case of binary classification, this is

$$\overline{m}(x) \triangleq \begin{cases} 1 & \Pr_{m \sim \mathcal{M}}[m(x) = 1] \geq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}.$$

  Note that the mode predictor $\overline{m}$ is the one that minimizes the expected disagreement between itself and a randomly chosen model $m \sim \mathcal{M}$.

[10] In practice, $\mathcal{M}$ may be constrained by the decision maker's ability to experiment with different design choices (e.g., type of model, random seed, etc.).

- **Randomized Predictions**: Under randomized prediction, the decision maker uses the classifier m^{rand} that, for each example x, randomly samples a model $m \sim \mathcal{M}$ and outputs $m(x)$. Formally, this is

$$\Pr[m^{\text{rand}}(x) = y] \triangleq \Pr_{m \sim \mathcal{M}}[m(x) = y].$$

- **Random Model Selection**: Under random model selection, the decision maker randomly samples some $m \sim \mathcal{M}$ and applies that m to every decision subject. Note that random model selection differs from an arbitrary selection in that the randomness (and the act of choosing) is made explicit [46].

Each of these techniques, alongside documentation of the reasons why a given method was chosen, provides a justifiable way to resolve multiplicity in different contexts. When decisions are made by a centralized authority, the decision maker's objective may be to resolve multiplicity by providing a consistent predictor (i.e., contains no explicit randomness) that minimizes multiplicity across the model distribution $\mathcal{M}$: for example, the government has a special legal burden to ensure consistency in decision making [14]. In such cases, the mode predictor best achieves these goals: it is the model that minimizes multiplicity compared to the model distribution $\mathcal{M}$.[11] Recent work has shown that, beyond stabilizing model predictions, mode aggregation also results in more stable model explanations, and thus suggests that models which return the mode over a random sample of similar models have more stable internals than individual models [11].

On the other hand, consider decisions that are low-stakes and frequent, such as choosing which advertisement to show to a user. Suppose 70% of models in the distribution $\mathcal{M}$ predict that a user x prefers credit card ads, and 30% predict that x prefers ads for cars. While the mode predictor would resolve this multiplicity by always showing x an ad for credit cards, under randomized prediction, the model will show the user credit card ads 70% of the time and car ads the other 30% of the time.[12] Of course, there are plenty of applications where such randomized predictions are undesirable; but in applications where decisions are low-stakes and repeated, this randomized sampling might give a person outcomes that better reflect the uncertainty contained in the model distribution.

Finally, there are cases where society would prefer that the model developer simply samples a random model $m \sim \mathcal{M}$ and always uses m. For example, consider an application like hiring or lending where multiple private actors make independent decisions. We may not want explicit randomness through sampling in these decisions, but if each supposedly independent actor uses the same mode predictor, then decision making effectively becomes a monoculture, which can have negative impacts both for individuals' recourse and social welfare [22, 36]. To prevent this, we might prefer that each model developer independently choose its own random model $m \sim \mathcal{M}$. And while random model selection may seem like the de facto resolution of predictive multiplicity in practice, private model developers may end up converging on the same models for a variety of reasons, including third-party vendors selling the same tools to multiple clients [48] or centralized evaluation (e.g., credit scores).

Crucially, all three of these methods mitigate arbitrariness since choosing among them *requires considering and deliberating between the different options*. By requiring model developers to document the model building process—and their ultimate decision on how to address remaining multiplicity—we can reach a justification for why a model's decision process and predictions are the way that they are [54].

[11] Black et al. [11] provides an evaluation of this approach, including theoretical guarantees on the consistency of mode-aggregated decisions.
[12] Randomized prediction is often used in the fairness literature to ensure that individuals or groups have similar probabilities of receiving an given outcome from a classifier [3, 29]. Our use of randomized prediction ensures that an individual has a chance of getting any outcome available to them under some $m \sim \mathcal{M}$.

7 CONCLUSION

Our work considers the implications of *model multiplicity*, the phenomenon of multiple models with equal accuracy for a given prediction task exhibiting different individual predictions or aggregate properties. We show that model multiplicity leads to increased flexibility—and perhaps even legal pressure—to prioritize fairness, robustness, and interpretability, among other values, in the model building process. However, this increased flexibility also leads to the risk of avoidable discrimination and to a lack of justification for model decisions when the model is chosen on the basis of accuracy alone. While this work does not serve as a complete exploration of the impact that predictive multiplicity may have on law and policy, we hope that by bringing attention to model multiplicity that we can add to the momentum to take advantage of the opportunities that it creates and head off the resistance that it could provoke.

ACKNOWLEDGMENTS

The authors thank Pauline Kim, Ashesh Rambachan, Aaron Rieke, Andrew Selbst, and members of FATE research group at Microsoft Research for their helpful comments on this work.

FUNDING/SUPPORT

Most of this research was conducted while Emily Black was an intern at Microsoft Research. This work was also supported by an Amazon Graduate Research Fellowship and the Center for Research on Computation and Society (CRCS) at the Harvard John A. Paulison School of Engineering and Applied Sciences. Solon Barocas is a full-time employee of Microsoft Research.

REFERENCES

[1] 1970. 116 Cong. Reg. 36572.
[2] 2011. Comment for 1002.6 - Rules Concerning Evaluation of Applications. https://www.consumerfinance.gov/rules-policy/regulations/1002/interp-6/.
[3] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. In *International Conference on Machine Learning*. 60–69.
[4] Christopher Anders, Plamen Pasliev, Ann-Kathrin Dombrowski, Klaus-Robert Müller, and Pan Kessel. 2020. Fairwashing explanations with off-manifold detergent. In *International Conference on Machine Learning*. PMLR, 314–323.
[5] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *California Law Review* 104 (2016), 671–732.
[6] Solon Barocas, Andrew D Selbst, and Manish Raghavan. 2020. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 80–89.
[7] Dimitris Bertsimas, Arthur Delarue, Patrick Jaillet, and Sebastien Martin. 2019. The price of interpretability. *arXiv preprint arXiv:1907.03419* (2019).
[8] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2021. The values encoded in machine learning research. *arXiv preprint arXiv:2106.15590* (2021).
[9] Emily Black, Solon Barocas, Alexandra Chouldechova, Logan Koepke, Kristian Lum, Michael Madaio, and Sarah Riley. 2022. Reducing Racial Disparity Through Procedural Interventions: Conceptions and Outcomes.
[10] Emily Black and Matt Fredrikson. 2021. Leave-one-out Unfairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 285–295.
[11] Emily Black, Klas Leino, and Matt Fredrikson. 2022. Selective Ensembles for Consistent Predictions. In *International Conference on Learning Representations*.
[12] Emily Black, Zifan Wang, Matt Fredrikson, and Anupam Datta. 2022. Consistent Counterfactuals for Deep Models. In *International Conference on Learning Representations*.
[13] Leo Breiman. 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science* 16, 3 (2001), 199–231.
[14] Lisa Schultz Bressman. 2003. Beyond accountability: Arbitrariness and legitimacy in the administrative state. *NYUL Rev.* 78 (2003), 461.
[15] Chaofan Chen, Kangcheng Lin, Cynthia Rudin, Yaron Shaposhnik, Sijia Wang, and Tong Wang. 2018. An interpretable model with globally consistent explanations for credit risk. *arXiv preprint arXiv:1811.12615* (2018).
[16] Irene Y. Chen, Fredrik D. Johansson, and David A. Sontag. 2018. Why Is My Classifier Discriminatory?. In *Advances in Neural Information Processing Systems*. 3543–3554.
[17] Danielle Keats Citron. 2007. Technological due process. *Wash. L Rev.* 85 (2007), 1249.

[18] Danielle Keats Citron and Frank Pasquale. 2014. The scored society: Due process for automated predictions. *Wash. L. Rev.* 89 (2014), 1.

[19] A Feder Cooper and Ellen Abrams. 2021. Emergent Unfairness in Algorithmic Fairness-Accuracy Trade-Off Research. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 46–54.

[20] Amanda Coston, Ashesh Rambachan, and Alexandra Chouldechova. 2021. Characterizing Fairness Over the Set of Good Models Under Selective Labels. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*. PMLR, 2144–2155.

[21] Kate Crawford and Jason Schultz. 2014. Big data and due process: Toward a framework to redress predictive privacy harms. *BCL Rev.* 55 (2014), 93.

[22] Kathleen Creel and Deborah Hellman. 2021. The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision Making Systems. *Virginia Public Law and Legal Theory Research Paper* 2021-13 (2021).

[23] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. 2020. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395* (2020).

[24] Anupam Datta, Matt Fredrikson, Gihyuk Ko, Piotr Mardziel, and Shayak Sen. 2017. Proxy Discrimination in Data-Driven Systems. *arXiv preprint arXiv:1707.08120* (2017).

[25] Pedro Domingos. 2000. A unified bias-variance decomposition. In *Proceedings of 17th International Conference on Machine Learning*. 231–238.

[26] Jiayun Dong and Cynthia Rudin. 2019. Variable importance clouds: A way to explore variable importance for the set of good models. *arXiv preprint arXiv:1901.03209* (2019).

[27] David Donoho. 2017. 50 years of data science. *Journal of Computational and Graphical Statistics* 26, 4 (2017), 745–766.

[28] Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush Varshney. 2020. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *International Conference on Machine Learning*. PMLR, 2803–2813.

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

[30] Equal Credit Opportunities Act, Public Law 93-495 1974. Codified at 15 U.S.C. § 1691, et seq.

[31] Fair Credit Reporting Act, Public Law 91-508 1970. Codified at 15 U.S.C. § 1681, et seq.

[32] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. 2019. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *J. Mach. Learn. Res.* 20, 177 (2019), 1–81.

[33] Stuart Geman, Elie Bienenstock, and René Doursat. 1992. Neural networks and the bias/variance dilemma. *Neural computation* 4, 1 (1992), 1–58.

[34] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2020. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv preprint arXiv:2010.04050* (2020).

[35] Pauline T. Kim. 2022. Race-aware algorithms: Fairness, nondiscrimination and affirmative action. *California Law Review* 110 (2022).

[36] Jon Kleinberg and Manish Raghavan. 2021. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences* 118, 22 (2021).

[37] Ron Kohavi, David H Wolpert, et al. 1996. Bias plus variance decomposition for zero-one loss functions. In *ICML*, Vol. 96. 275–83.

[38] Loren Larsen. 2019. *Resumes, Robots, and Racism: The Truth about AI in Hiring*. HireVue.

[39] David Lehr and Paul Ohm. 2017. Playing with the data: what legal scholars should learn about machine learning. *UCDL Rev.* 51 (2017), 653.

[40] Charles T. Marx, Flávio P. Calmon, and Berk Ustun. 2020. Predictive Multiplicity in Classification. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, 6765–6774.

[41] Johannes Mehrer, Courtney J Spoerer, Nikolaus Kriegeskorte, and Tim C Kietzmann. 2020. Individual differences among deep neural network models. *Nature communications* 11, 1 (2020), 1–12.

[42] Aditya Krishna Menon and Robert C Williamson. 2018. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*. PMLR, 107–118.

[43] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[44] Samir Passi and Solon Barocas. 2019. Problem formulation and fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 39–48.

[45] Martin Pawelczyk, Klaus Broelemann, and Gjergji. Kasneci. 2020. On Counterfactual Explanations under Predictive Multiplicity. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI) (Proceedings of Machine Learning Research)*.

[46] Ronen Perry and Tal Z Zarsky. 2014. May the Odds Be Ever in Your Favor: Lotteries in Law. *Ala. L. Rev.* 66 (2014), 1035.

[47] Manish Raghavan and Solon Barocas. 2019. Challenges for mitigating bias in algorithmic hiring. *Brookings. Retrieved February* 25 (2019), 2020.

[48] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 469–481.

[49] Xavier Renard, Thibault Laugel, and Marcin Detyniecki. 2021. Understanding Prediction Discrepancies in Machine Learning Classifiers. *arXiv preprint arXiv:2104.05467* (2021).

[50] Michael L Rich. 2016. Machine learning, automated suspicion algorithms, and the fourth amendment. *University of Pennsylvania Law Review* (2016), 871–929.

[51] Kit T Rodolfa, Hemank Lamba, and Rayid Ghani. 2021. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nature Machine Intelligence* 3, 10 (2021), 896–904.

[52] Kit T Rodolfa, Erika Salomon, Lauren Haynes, Iván Higuera Mendieta, Jamie Larson, and Rayid Ghani. 2020. Case study: predictive fairness to reduce misdemeanor recidivism through social service interventions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 142–153.
[53] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.
[54] Andrew D Selbst and Solon Barocas. 2018. The intuitive appeal of explainable machines. *Fordham L. Rev.* 87 (2018), 1085.
[55] Lesia Semenova, Cynthia Rudin, and Ronald Parr. 2019. A study in Rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. *arXiv preprint arXiv:1908.01755* (2019).
[56] Shai Shalev-Shwartz and Shai Ben-David. 2014. *Understanding machine learning: From theory to algorithms*. Cambridge university press.
[57] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. 2019. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*.
[58] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable Recourse in Linear Classification. In *Conference on Fairness, Accountability, and Transparency*. 10–19.
[59] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2018. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology* 31, 2 (2018), 841–887.
[60] Tong Wang. 2019. Gaining free or low-cost interpretability with interpretable partial substitute. In *International Conference on Machine Learning*. PMLR, 6505–6514.
[61] Michael Wick, swetasudha panda, and Jean-Baptiste Tristan. 2019. Unlocking Fairness: a Trade-off Revisited. In *Advances in Neural Information Processing Systems*, Vol. 32.
[62] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. 2019. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*. PMLR, 7472–7482.

A A FORMAL MODEL OF MULTIPLICITY

Here, we formalize the relationship between individual-level disagreement in models and standard formulations of the bias-variance trade-off. Standard models of the bias-variance trade-off decompose loss into three components: bias, variance, and noise (e.g., [37]). Bias refers to the difference between the mean predictor (or in the case of 0-1 loss, which is simplest, the mode predictor) and the Bayes optimal predictor. Variance refers to the difference between any particular model and the mode predictor. Noise refers to the expected loss of the Bayes optimal model. For simplicity, we focus on the case of binary classification with 0-1 loss, though similar approaches could be used to characterize continuous models.

We begin by providing basic definitions in Section A.1. We explore the relationship between predictive multiplicity and accuracy in Section A.2, showing a fairly loose connection. We conclude by showing a tighter connection between predictive multiplicity and variance in Section A.3.

A.1 Basic Definitions

Suppose binary classification models come from some fixed distribution $\mathcal{M}$ (e.g., the distribution induced by random seeds, inclusion of data, etc.). Let $\mathcal{D}$ be the distribution over data. In a slight abuse of notation, we will denote a random data point as $x \sim \mathcal{D}$, and a random (data point, label) pair as $(x,y) \sim \mathcal{D}$. Let m^* be the Bayes optimal model, and let $\overline{m}$ be the mode predictor, defined as

$$\overline{m}(x) \triangleq \begin{cases} 1 & \Pr_{m\sim\mathcal{M}}[m(x)=1] \geq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases},$$

i.e., $\overline{m}$ assigns x the most probable label over the distribution of models $\mathcal{M}$. The 0-1 loss is defined as

$$L(y_1,y_2) \triangleq \begin{cases} 0 & y_1 = y_2 \\ 1 & \text{otherwise} \end{cases}.$$

For a given data point x, define

$$N(x) \triangleq \mathbb{E}_{y|x}\left[L(m^*(x),y)\right] \quad \text{(noise)}$$

$$B(x) \triangleq L(\overline{m}(x),m^*(x)) \quad \text{(bias)}$$

$$V_m(x) \triangleq L(m(x),\overline{m}(x)) \quad \text{(variance)}$$

Note that of these three, variance is the only one that depends on the particular model m. The expected error (also known as the "loss") of a model m on dataset $\mathcal{D}$ is

$$\text{err}(m,\mathcal{D}) \triangleq \mathbb{E}_{(x,y)\sim D}[L(m(x),y)].$$

Define the disagreement between two models m_1 and m_2 as

$$d(m_1,m_2) \triangleq \Pr_{x\sim\mathcal{D}}[m_1(x) \neq m_2(x)] = \mathbb{E}_{x\sim\mathcal{D}}[L(m_1(x),m_2(x))],$$

i.e., $d(m_1,m_2)$ is the probability that m_1 and m_2 disagree on a randomly drawn data point. Intuitively, predictive multiplicity flips decisions for more people as d grows. Note that $d(\cdot,\cdot)$ is symmetric and satisfies the triangle inequality:[13]

$$d(m_1,m_2) \leq d(m_1,m_3)+d(m_2,m_3)$$

A natural way to formalize predictive multiplicity for a distribution $\mathcal{M}$ over models is the expected pairwise disagreement over the distribution of models: if two models are randomly selected, how many points do they disagree on in expectation?

$$I(\mathcal{M}) \triangleq \mathbb{E}_{m_1,m_2\sim\mathcal{M}}[d(m_1,m_2)].$$

Thus, I is a formal measure of the predictive multiplicity present in a distribution $\mathcal{M}$ over models. Note that I is task-specific, since it depends on the data distribution $\mathcal{D}$. For the remainder of this paper, we will assume that $\mathcal{M}$ refers to the Rashomon set, i.e., all models in $\mathcal{M}$ have the same error L^*.

A.2 Predictive Multiplicity and Accuracy

Here, we present results relating predictive multiplicity to error. If all models in $\mathcal{M}$ have error L^*, Theorem A.1 upper-bounds predictive multiplicity $I(\mathcal{M})$ by $2L^*$. We will revisit this bound in Section A.3 to derive a tighter bound. Theorem A.2 shows that under certain assumptions, as L^* decreases (models in $\mathcal{M}$ become more accurate), predictive multiplicity approaches 0.

Theorem A.1.

$$I(\mathcal{M}) \leq 2L^*$$

[13] In fact, d is a metric, since it is nonnegative and $d(m,m)=0$.

Proof. We formalize the observation that two models that only make mistakes with probability p can only disagree with one another with probability at most $2p$:

$$\begin{aligned}
&\Pr_{x\sim\mathcal{D}}[m_1(x)\neq m_2(x)]\\
&\leq \Pr_{(x,y)\sim\mathcal{D}}[(m_1(x)=y\cap m_2(x)\neq y)\cup(m_1(x)\neq y\cap m_2(x)=y)]\\
&= \Pr_{(x,y)\sim\mathcal{D}}[m_1(x)=y\cap m_2(x)\neq y]+\Pr_{(x,y)\sim\mathcal{D}}[m_1(x)\neq y\cap m_2(x)=y]\\
&\leq \Pr_{(x,y)\sim\mathcal{D}}[m_2(x)\neq y]+\Pr_{(x,y)\sim\mathcal{D}}[m_1(x)\neq y]\\
&= 2L^*
\end{aligned}$$

Since this holds for any $m_1, m_2 \in \mathcal{M}$, it holds in expectation over for random $m_1, m_2 \sim \mathcal{M}$. □

Note that this characterization is essentially tight: consider the case where the true label is always $y=1$ and $\mathcal{M}$ assigns equal probability to each of k models, where $m_i(x_i)=0$ and $m_i(x)=1$ for all $x\neq x_i$. Then, each model makes exactly one error (on x_i), and models m_i and m_j disagree in exactly two points (x_i and x_j). Thus, the expected number of disagreements between two randomly selected models is $2L^*(1-1/k)$, taking into account the probability that the same model is selected twice.

Uniqueness. Next, we show that optimal models are in some sense unique. Our intuition is that the Bayes-optimal model is unique. This isn't strictly true: if $\Pr[y=1]=1/2$ given an x, then all models are Bayes-optimal, since they all have loss $1/2$. But our intuition should still hold for "predictable" problems, where y can be predicted from x better than random chance.

Assume $|\Pr_{y\sim\mathcal{D}|x}[y=1]-1/2|>c$, i.e., y can be predicted better than 50-50 chance for every x. We will show that predictive multiplicity goes to 0 as model performance approaches the Bayes risk. As before, let L^* be the loss of any model in $\mathcal{M}$. Then, the following theorem shows that as the models in $\mathcal{M}$ approach the Bayes risk R on $\mathcal{D}$, predictive multiplicity goes to 0.

Theorem A.2. *If* $|\Pr_{y\sim\mathcal{D}|x}[y=1]-1/2|>c$, *then*

$$I(\mathcal{M}) \leq \frac{L^*-R}{c}.$$

Proof.

$$
\begin{aligned}
&\mathbb{E}_{(x,y)\sim\mathcal{D}}[L(m(x),y)] \\
&= \Pr_{(x,y)\sim\mathcal{D}}[m(x)\neq y] \\
&= \Pr_{(x,y)\sim\mathcal{D}}[m^*(x)\neq y] + \Pr_{(x,y)\sim\mathcal{D}}[m(x)\neq m^*(x)] \\
&\quad -2\Pr_{(x,y)\sim\mathcal{D}}[m(x)\neq m^*(x)\cap m^*(x)\neq y] \\
&= R + d(m,m^*) \\
&\quad -2\Pr_{(x,y)\sim\mathcal{D}}[m(x)\neq m^*(x)]\Pr_{(x,y)\sim\mathcal{D}}[m^*(x)\neq y \mid m(x)\neq m^*(x)] \\
L^* &= R + d(m,m^*)\left(1-2\Pr_{(x,y)\sim\mathcal{D}}[m^*(x)\neq y \mid m(x)\neq m^*(x)]\right) \\
d(m,m^*) &= \frac{L^*-R}{1-2\Pr_{(x,y)\sim\mathcal{D}}[m^*(x)\neq y \mid m(x)\neq m^*(x)]} \\
&\leq \frac{L^*-R}{1-2(1/2-c)} \qquad (m^* \text{ is wrong with probability at most } 1/2-c \text{ by assumption}) \\
&= \frac{L^*-R}{2c}
\end{aligned}
$$

Thus, the distance between any $m\sim\mathcal{M}$ and the Bayes optimal model m^* is bounded. We can use this to bound predictive multiplicity as follows:

$$
\begin{aligned}
I(\mathcal{M}) &= \mathbb{E}_{m_1,m_2\sim\mathcal{M}}[d(m_1,m_2)] \\
&\leq \mathbb{E}_{m_1,m_2\sim\mathcal{M}}\left[d(m_1,m^*)+d(m_2,m^*)\right] \\
&\leq \frac{2(L^*-R)}{2c} \\
&= \frac{L^*-R}{c}
\end{aligned}
$$

□

A.3 Predictive Multiplicity and Variance

Next, we show a tight connection between predictive multiplicity and variance. Theorem A.3 shows that predictive multiplicity and variance are within a factor of 2 of one another. As a corollary, we show that reducing loss can *increase* predictive multiplicity (Corollary A.4). Theorem A.5 uses this result to sharpen the bound in Theorem A.1.

Let the expected variance of a model distribution $\mathcal{M}$ be

$$V(\mathcal{M}) \triangleq \mathbb{E}_{m\sim\mathcal{M},x\sim\mathcal{D}}[V_m(x)].$$

Theorem A.3.

$$\frac{1}{2}V(\mathcal{M}) \leq I(\mathcal{M}) \leq 2V(\mathcal{M})$$

Proof. We begin with an upper bound on predictive multiplicity.

$$\begin{aligned}
I(\mathcal{M}) &= \mathbb{E}_{m_1,m_2\sim\mathcal{M}}[d(m_1,m_2)] \\
&\leq \mathbb{E}_{m_1,m_2\sim\mathcal{M}}[d(m_1,\overline{m})+d(m_2,\overline{m})] \\
&= 2\mathbb{E}_{m\sim\mathcal{M}}[d(m,\overline{m})] \\
&= 2\mathbb{E}_{m\sim\mathcal{M}}[\mathbb{E}_{x\sim\mathcal{D}}[L(m,\overline{m})]] \\
&= 2V(\mathcal{M})
\end{aligned}$$

We derive a lower bound with the following observation: if m disagrees with the mode predictor $\overline{m}$ on an instance x, then m must disagree on x with at least half of the models in $\mathcal{M}$.[14] Formally, we can write this as

$$m(x)\neq\overline{m}(x) \Longrightarrow \Pr_{m'\sim\mathcal{M}}[m(x)\neq m'(x)] \geq \frac{1}{2},$$

which implies

$$\Pr_{x\sim\mathcal{D}}[m(x)\neq\overline{m}(x)] \leq 2\Pr_{m'\sim\mathcal{M},x\sim\mathcal{D}}[m(x)\neq m'(x)]. \tag{1}$$

Using this, we have

$$\begin{aligned}
V(\mathcal{M}) &= \mathbb{E}_{m\sim\mathcal{M},x\sim\mathcal{D}}[V_m(x)] \\
&= \mathbb{E}_{m\sim\mathcal{M}}\left[\Pr_{x\sim\mathcal{D}}[m(x)\neq\overline{m}(x)]\right] \\
&\leq 2\mathbb{E}_{m\sim\mathcal{M}}\left[\Pr_{m'\sim\mathcal{M},x\sim\mathcal{D}}[m(x)\neq m'(x)]\right] && \text{(by (1))} \\
&= 2\mathbb{E}_{m,m'\sim\mathcal{M}}\left[\Pr_{x\sim\mathcal{D}}[m(x)\neq m'(x)]\right] \\
&= 2\mathbb{E}_{m,m'\sim\mathcal{M}}\left[d(m,m')\right] \\
&= 2I(\mathcal{M})
\end{aligned}$$

Putting this together, we have

$$\frac{1}{2}V(\mathcal{M}) \leq I(\mathcal{M}) \leq 2V(\mathcal{M}).$$

□

This shows a fairly tight connection between model variance and predictive multiplicity, which can help our intuition in a few ways. First, we see that increasing accuracy by increasing variance and reducing bias (e.g., using a more complex model class) can actually *increase* predictive multiplicity, consistent with empirical findings [10, 12]. Second, we see that efforts to decrease model variance (e.g., more data) should *reduce* predictive multiplicity. This yields the following result:

Corollary A.4. *Reducing loss by decreasing bias and increasing variance can increase predictive multiplicity.*

Deriving a tighter relationship between predictive multiplicity and accuracy. We can use this insight on the connection between predictive multiplicity and accuracy to improve the bound in Theorem A.1. As before, let L^* be the loss of any model in $\mathcal{M}$. Let R be the Bayes risk, and let B be the bias of $\mathcal{M}$ (i.e., the error of the mode predictor $\overline{m}$).

[14] By "half," we mean models that account for at least half the probability mass of $\mathcal{M}$.

Theorem A.5.

$$I(\mathcal{M}) \leq 2[L^* - R(1-2B)]$$

Proof. We begin with decomposition of error into noise, bias, and variance from [25].

$$\begin{aligned}
L^* &= \mathbb{E}_{(x,y)\sim\mathcal{D},m\sim\mathcal{M}}[L(m(x),y)] \\
&= (2\cdot \Pr_{x\sim D}[\overline{m}(x) = m^*(x)] - 1)\mathbb{E}_{x\sim D}[N(x)] + \mathbb{E}_{x\sim\mathcal{D}}[B(x)] \\
&+ \mathbb{E}_{x\sim\mathcal{D},m\sim\mathcal{M}}\left[(-1)^{\mathbb{1}_{\overline{m}(x)\neq m^*(x)}} V_m(x)\right] \\
&= (2(1-d(\overline{m},m^*)-1)R + d(\overline{m},m^*) + \mathbb{E}_{x\sim\mathcal{D},m\sim\mathcal{M}}[(1-2B(x))V_m(x)] \\
&= (1-2B)R + B + V(\mathcal{M}) - 2\mathbb{E}_{x\sim\mathcal{D},m\sim\mathcal{M}}[B(x)V_m(x)] \\
&= (1-2B)R + B + V(\mathcal{M}) - 2\Pr_{x\sim\mathcal{D},m\sim\mathcal{M}}[B(x)=1 \cap V_m(x)=1] \\
&= (1-2B)R + B + V(\mathcal{M}) - 2\Pr_{x\sim\mathcal{D}}[B(x)=1]\Pr_{m\sim\mathcal{M}}[V_m(x)=1 \mid B(x)=1]
\end{aligned}$$

Note that for any x, $\Pr_{m\sim\mathcal{M}}[V_m(x)=1] \leq \frac{1}{2}$ because by definition of the mode predictor, the probability a random model disagrees with the mode predictor on a given x as at most a half. Since this holds for every x, this is true conditioned on $B(x)=1$, so

$$\Pr_{m\sim\mathcal{M}}[V_m(x)=1 \mid B(x)=1] \leq \frac{1}{2}.$$

Thus, we have

$$\begin{aligned}
L^* &= (1-2B)R + B + V(\mathcal{M}) - 2\Pr_{x\sim\mathcal{D}}[B(x)=1]\Pr_{m\sim\mathcal{M}}[V_m(x)=1 \mid B(x)=1] \\
&\geq (1-2B)R + B + V(\mathcal{M}) - \Pr_{x\sim\mathcal{D}}[B(x)=1] \\
&= (1-2B)R + B + V(\mathcal{M}) - B \\
&= (1-2B)R + V(\mathcal{M}) \\
&\geq (1-2B)R + \frac{1}{2}I(\mathcal{M}) && \text{(By Theorem A.3)}
\end{aligned}$$

Rearranging yields the desired result:

$$I(\mathcal{M}) \leq 2[L^* - R(1-2B)].$$

□

Note that $B \leq \frac{1}{2}$, since a model that deterministically predicts the more likely class achieves loss at most $\frac{1}{2}$. As a result, Theorem A.5 immediately implies Theorem A.1.