Machine learning models, or algorithms trained on historical data, can be harmful when they are used to make decisions about people. From underfitting to overfitting, from bugs in code to biased training data, machine learning engineers and other stakeholders alike are concerned with the capacity of these algorithms to make problematic decisions while going unnoticed by their human supervisors, escaping scrutiny due to their increasing complexity and incomprehensibility. In recent years, this has led to a flurry of technical research on methods to anticipate, detect, or prevent undesirable behavior, under the umbrella terms of “fair machine learning” and “explainable artificial intelligence.” But these efforts have not gone without criticism, especially from social science disciplines quick to point out problems with a techno-solutionist approach. In this thesis, we concern ourselves with trying to articulate what exactly we can learn from quantitative methods proposed to address this problem and what we should do with that information. In the first Part, which centers on a family of “explainability” techniques known as feature importance, we quantitatively and qualitatively evaluate specific algorithms based on game theoretic principles, and interrogate the philosophy behind the overall methodology. The next Part articulates and draws attention to specific machine learning harms and failure modes that must be treated as public policy issues and offer suggestions on how to do so. For the third Part, we derive novel, data-dependent, group-specific generalization bounds on the performance of machine learning models trained on heterogeneous data, which offers a framework to understand the data and modeling conditions which cause performance disparities across those groups.
Explainability, fairness, and evaluation in machine learning: From theory to policy and back

by

Indra Elizabeth Kumar

B. A., Scripps College, 2016
M. S., University of Massachusetts, 2019

A dissertation submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in the Department of Computer Science at Brown University

Providence, Rhode Island
May 2024
This dissertation by Indra Elizabeth Kumar is accepted in its present form by the Department of Computer Science as satisfying the dissertation requirement for the degree of Doctor of Philosophy.

Date ________________

Suresh Venkatasubramanian, Director

Recommended to the Graduate Council

Date ________________

Timothy Edgar, Reader

Date ________________

Stephen Bach, Reader

Approved by the Graduate Council

Date ________________

Thomas A. Lewis
Dean of the Graduate School
Acknowledgements

I am grateful to the many wonderful people and institutions that supported this work, who I’ve done my best to enumerate here.

First, I want to thank Suresh for being everything I could have hoped for in an advisor. Your guidance and support have been my greatest assets throughout my research career thus far. Thank you for trusting me to make my own decisions, pushing me to do my best work, and prioritizing my wellbeing. I would also like to thank the rest of my committee for their thoughtful feedback on this work, as well as the many other members of the faculty and staff at the Utah School of Computing, Brown Data Science Institute, and Brown Computer Science Department who have taught me things, provided support, and just generally helped me out over the years. I feel lucky to have been affiliated with both institutions during my time as a graduate student.

This dissertation would not have been possible without my excellent coauthors on the work it contains; it has been an honor to collaborate with and learn from all of you. Similarly, to my fellow graduate students at the U who started this journey with me, and the ones who welcomed or joined me at Brown: your passion, intellect, and kindness have all been a constant source of inspiration to me. Thank you for being my community and support system, and I hope we will stay lifelong friends. Thank you, additionally, to the GLO, for everything you do to support us as workers.

I would also like to thank the Utah chapter of the ARCS Foundation, which generously supported the first year of my Ph.D. with the Noel de Nevers Memorial Fellowship, and Arthur AI, which supported me as a Machine Learning Summer Research Fellow.

Finally, to my family, friends-who-are-basically-family, and Kevin: Thank you for existing, sticking with me through the lowest and highest points of my life and career, bringing me joy, keeping me alive, and, of course, believing in me.
Contents

List of Tables vii
List of Figures viii
Introduction 1

I Understanding feature importance and its limitations 3

1 On the limitations of Shapley value explanations 4
  1.1 Background: Using Shapley values for explanation 5
  1.2 Mathematical critiques 8
  1.3 Human-centered concerns 14

2 Quantifying the limitations of Shapley value explanations 20
  2.1 Shapley residuals 20
  2.2 Residuals for explanation methods 25
  2.3 Discussion and limitations 29

3 Using the lens of feminist epistemology to critique feature importance 32
  3.1 Background: Feminist epistemology and philosophy of science 33
  3.2 Values in feature importance 35
  3.3 Towards pluralistic, contextual, and interactive approaches to explanation 43

II Public policy problems in machine learning 49

4 Applying the Equal Credit Opportunity Act to machine learning models 50
  4.1 Background: Fair lending law and regulation 51
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.2 Fair lending and algorithms</td>
<td>55</td>
</tr>
<tr>
<td>4.3 Discrimination risks in models trained on loan repayment data</td>
<td>62</td>
</tr>
<tr>
<td>4.4 Regulatory opportunities</td>
<td>67</td>
</tr>
<tr>
<td>5 Challenging the fallacy of AI functionality in public policy</td>
<td>70</td>
</tr>
<tr>
<td>5.1 Background: Related work</td>
<td>71</td>
</tr>
<tr>
<td>5.2 The functionality assumption</td>
<td>72</td>
</tr>
<tr>
<td>5.3 Taxonomy of failures</td>
<td>75</td>
</tr>
<tr>
<td>5.4 Interventions</td>
<td>84</td>
</tr>
<tr>
<td>III Provable guarantees in fair machine learning</td>
<td>91</td>
</tr>
<tr>
<td>6 Analyzing the Regularizing Effects of Group-Fair Training on Shared Models</td>
<td>92</td>
</tr>
<tr>
<td>6.1 Background: Preliminaries and related work</td>
<td>93</td>
</tr>
<tr>
<td>6.2 Bounding generalization error in fair training</td>
<td>98</td>
</tr>
<tr>
<td>6.3 Experiments</td>
<td>106</td>
</tr>
<tr>
<td>6.4 Discussion</td>
<td>108</td>
</tr>
<tr>
<td>Bibliography</td>
<td>116</td>
</tr>
</tbody>
</table>
List of Tables

1.1 Proposed value function $v_{f,x}$ for each method, compared with the quantity $\hat{v}_{f,x}$ the algorithm actually approximates. The interventional distribution $D$ used depends on the method (i.e., for KernelSHAP it is the observational joint distribution of $X$). 7

2.1 KernelSHAP game for Example 1 - the input $(1,1,1)$ to $f(x_1, x_2, x_3) = x_1 + 2x_2x_3$ where $x_i$ are iid $\mathcal{N}(0,1)$ features. 22

5.1 Failure Taxonomy 76

6.1 Sample sizes $m_{1:3}$, parameter vectors $\beta_{1:3}$, and Monte-Carlo empirical Rademacher averages (MCERA) for both $\mathcal{H}$ and $\hat{\mathcal{H}}_{1:3}$. 106

6.2 Data generating parameters for logistic regression experiments. 106
List of Figures

1.1 Visualizing a game and its gradient. .................................................. 5
1.2 Samples that might be drawn to estimate $E[f(1, Y)]$ and $E[f(X, 2)]$ to explain $f(1, 2)$ for some function $f$, given correlated Gaussian distributions for $X$ and $Y$, depending on whether the expectation is taken over $X|Y = 2$ and $Y|X = 1$ (left) or $X$ and $Y$ (right) ................................................................. 11

2.1 Visualizing the decomposition of a game and its residuals. .................... 25
2.2 Shapley values and residuals on a decision tree for the Occupancy Detection task .......................................................... 28
2.3 Geometric representation of the KernelSHAP game for $f(10, 320)$, where arrows to the right indicate inclusion of the light feature and arrows down indicate inclusion of the hour feature. ........................................ 29

6.1 Visualization of unrestricted class $\mathcal{H}$, theoretical restricted class $\mathcal{H}^*_i$, and samples of empirical restricted class $\hat{\mathcal{H}}_i$ (varying $z_i$). One possible empirical malfare minimizer $\hat{h}$ (contained by $\hat{\mathcal{H}}_i$ and $\mathcal{H}^*_i$ with high probability), as well as the true malfare minimizer $h^*$ (which may fall outside of $\mathcal{H}^*_i$ or $\hat{\mathcal{H}}_i$ due to overfitting to groups other than $i$) are also shown. ................................................................. 101
6.2 Rademacher average samples in the parameter space of $\hat{\mathcal{H}}_i$ for each group $i \in \{1, 2, 3\}$. 105
6.3 Average test risk of pooled and separately trained models on three groups (see table 6.2). ................................................................. 107
6.4 Generalization error bounds derived from original hypothesis class $\mathcal{H}$ and restricted hypothesis classes $\hat{\mathcal{H}}_i$, compared with shared model $\hat{h}$ train-test gap over 7 independent runs, with quartiles and median trend lines. ........................................ 107
Introduction

Machine learning models, or algorithms trained on historical data, can be harmful when they are used to make decisions about people. They can make mistakes if deployed in high-stakes situations, resulting in misdiagnosis, denial of critical resources and benefits, or even incarceration. They can learn to treat people differently based on characteristics we would really prefer them not to use, like race or gender. From underfitting to overfitting, from bugs in code to biased training data, machine learning engineers and other stakeholders alike are concerned with the capacity of these algorithms to make problematic decisions while going unnoticed by their human supervisors, escaping scrutiny due to their increasing complexity and incomprehensibility.

In recent years, this has led to a flurry of technical research on methods to anticipate, detect, or prevent undesirable behavior, under the umbrella terms of “fair machine learning” and “explainable artificial intelligence.” But these efforts have not gone without criticism, especially from social science disciplines quick to point out problems with the techno-solutionist approach that many computer scientists assume by default: not every social problem can be fixed with computer-science thinking.

In this thesis, we broadly attempt to articulate what exactly can be learned from quantitative methods proposed to detect and prevent undesirable behavior in machine learning models and how users and regulators can act on that information.

The first Part of this thesis concerns itself with the still-exploding body of work in explainable AI. In particular, we are interested in the usage of “post-hoc” explanations in the form of feature importance metrics, which assign a notion of absolute or relative importance to the input features in a model in an attempt to understand its internal logic. In the first chapter, we quantitatively and qualitatively evaluate specific post-hoc explanation algorithms based on the game theoretic notion of Shapley values, finding both mathematical and human-centered reasons to be skeptical of their usefulness. In the second chapter, we expand on this analysis by proposing Shapley residuals, a measure to quantify the limitations of Shapley value based explanations, and illustrate the type of missing information about variable interactions that they capture. In the third chapter, we use the
lens of feminist epistemology to interrogate and critique the philosophy behind the methodology implicit in feature importance work.

Part II articulates and draws attention to machine learning harms that must be treated as public policy issues, and offers suggestions on how to do so. In chapter 4, we describe the relationship between fair ML methods and U.S. fair lending policy, and use this to point out the conditions under which regulators must take action to prevent discrimination in algorithms. In chapter 5, we describe a tactical misstep often made in the regulation of AI, which we call the “functionality fallacy,” or the tendency for policymakers to assume that the AI they are tasked with regulating are “too smart,” when in fact this false perspective results in the deemphasis of real, contemporary problems with the technology. We also develop a two-level taxonomy of AI failures, based on an iterative coding procedure over hundreds of real cases, to provide language for future discussions and to demonstrate the wide scope of the problem.

Finally, in response to the trend across disciplines of using increasingly complex models which may overfit on small populations, in the final Part we derive novel group-specific generalization bounds on the performance of machine learning models trained on heterogeneous data. When the best hypothesis for a prediction problem on a minority group population differs significantly from that of the majority group, a model from a sufficiently simple hypothesis class optimized for a traditional loss function will prioritize fitting the majority group well and generalize poorly to the minority group. On the other hand, a model complex enough to learn the difference between the groups may end up overfitting to the minority group due to its smaller sample size, and thus generalize poorly on that group upon deployment. Our approach utilizes empirical Rademacher averages for learning over a hypothesis class which is less likely to overfit to the minority group. The resulting error bound offers better guarantees for group-level generalization performance in certain cases.
Part I

Understanding feature importance and its limitations
Chapter 1

On the limitations of Shapley value explanations

One of the more prominent tools in the feature importance literature has been the Shapley value, a method for additively attributing value among players of a cooperative game. In this setting, the “players” are the features used by the model, and the game is the prediction of the model. A variety of methods to assign feature influence using the Shapley value have been developed [260, 395, 271, 103, 283, 157, 3].

In this chapter, we demonstrate that applying the Shapley value to the problem of feature importance introduces mathematically formalizable properties which may not align with what we would expect from an explanation. Taking a human-centric perspective, we evaluate Shapley-value-based explanations through established frameworks of what people expect from explanations, and find them wanting. We find that the game theoretic problem formulation of Shapley-value-based explanations do not match the proposed use cases for its solution, and thus caution against their usage.

Further, we propose a method to quantify the extent to which Shapley value explanation deviates from their natural interpretation. We utilize an interpretation of Shapley values as the result of an orthogonal projection between vector spaces to calculate a residual representing the kernel component of that projection. We provide an algorithm for computing these residuals, characterize different modeling settings based on the value of the residuals, and demonstrate that they capture information about model predictions that Shapley values cannot. Shapley residuals can thus act as a warning to practitioners against overestimating the degree to which Shapley-value-based explanations give them insight into a model.
1.1 **Background:** Using Shapley values for explanation

Cooperative game theory. A cooperative game consists of $d$ players and a value function $v : 2^d \to \mathbb{R}$ where $[d] \triangleq \{1, \ldots, d\}$. The quantity $v(S)$ represents the value of the game for a coalition of players $S \in N \triangleq 2^d$. This value function represents how much collective payoff a set of players can gain by “cooperating” as a set. At a high level, the Shapley value is one way to allocate the total value of the grand coalition, $v(\{1, 2, \ldots, N\})$, between the individual players. It is based on trying to answer the question: how much does player $i$ contribute to the coalition?

It will be useful for us to visualize a game as a function over the vertices of a $d$-dimensional hypercube. Each coordinate corresponds to the presence or absence of a certain player, and each vertex corresponds to a subset of players. Specifically, we can think of the set $N$ as the $d$-dimensional hypercube $G = (V = N, E)$ with each vertex labeled by a set $S \subseteq [d]$ and edges between sets $S$ and $S \cup \{i\}$ for all $i \in [d], S$. We depict this interpretation in Figure 1.1(a).

Let $\mathbb{R}^V$ be the space of functions from $V$ to $\mathbb{R}$ and let $\mathbb{R}^E$ be the space of functions from $E$ to $\mathbb{R}$. In particular, the game $v$ is an element of $\mathbb{R}^V$. The differential operator $\nabla : \mathbb{R}^V \to \mathbb{R}^E$ is then defined as

$$\nabla v(S, S \cup \{i\}) = v(S \cup i) - v(S)$$

for any $v \in \mathbb{R}^V$. Essentially $\nabla$ is a discrete gradient operator on $G$, mapping functions on vertices to functions on edges (see Figure 1.1(b)). The value $\nabla v(S, S \cup \{i\})$ is also referred to as the marginal contribution of player $i$ with respect to set $S$.

The Shapley value Intuitively, the Shapley value can be understood as a weighted average of a player’s marginal contributions to every possible subset of players.
Definition 1. Let $\Pi$ be the set of permutations of the integers up to $N$, and given $\pi \in \Pi$ let $S_{i,\pi} = \{j : \pi(j) < \pi(i)\}$ represent the players preceding player $i$ in $\pi$. The Shapley value of player $i$ is then

$$\phi_i(v) = \frac{1}{N!} \sum_{\pi \in \Pi} \nabla v(S, S \cup \{i\})$$ (1.1)

This can be rewritten in terms of the unique subsets $S \subseteq \{1, 2, ..., N\}$ and the number of permutations for which some ordering of $S$ immediately precedes player $i$:

$$\phi_i(v) = \frac{1}{N!} \sum_{S \subseteq \{1, 2, ..., N\} \mid S \mid} |S|!(N - |S| - 1)! \nabla v(S, S \cup \{i\})$$ (1.2)

Theorem 1 (Shapley values[370]). The Shapley values $\phi_i(v), i \in [d]$ are the unique values satisfying the properties

Efficiency: $\sum_{i=1}^{d} \phi_i(v) = v([d])$.

Dummy: If $v(S \cup \{i\}) = v(S)$ for all $S \subset [d] \setminus \{i\}$, then $\phi_i(v) = 0$.

Symmetry: If $v(S \cup \{i\}) = v(S \cup \{j\})$ for all $S \subset [d] \setminus \{i, j\}$, then $\phi_i(v) = \phi_j(v)$.

Linearity: If $v, v'$ are two games on $d$ players, then $\phi_i(\alpha v + \alpha' v') = \alpha \phi_i(v) + \alpha' \phi_i(v')$.

Shapley values for feature importance Several methods have been proposed to apply the Shapley value to the problem of feature importance. Given a model $f(x_1, x_2, ..., x_d)$, the features from 1 to $d$ can be considered players in a game in which the payoff $v$ is some measure of the importance or influence of that subset. The Shapley value $\phi_v(i)$ can then be viewed as the “influence” of $i$ on the outcome. In this section, we describe methods which consist of defining a value function $v_f$ with respect to a model $f$, and computing (or approximating) the resulting Shapley values. We will use the following notation:

$X$: a multivariate random variable $\{X_1, X_2, ..., X_d\}$

$x$: a set of values $\{x_1, x_2, ..., x_d\}$

$X_S$: the set of random variables $\{X_i : i \in S\}$

$x_S$: the set of values $\{x_i : i \in S\}$

Value functions Shapley values have a fairly long history in the context of feature importance. Kruskal [247] and Lipovetsky and Conklin [260] proposed using the Shapley value to analyze global feature importance in linear regression by using the value function $v_f(S)$ to represent the $R^2$ of a linear model $f$ built on predictors $S$, to decompose the variance explained additively between the features. Owen and Prieur [314] applied the Shapley value to the problem of sensitivity analysis, where the total variance of a function is the quantity of interest.
Many recently proposed “local” methods \cite{346,269,271} define a value function $v_{f,x} : 2^d \to \mathbb{R}$ that depends on a specific data instance $x$ to explain how each feature contributes to the output of the function on this instance. The value of the grand coalition, in this setting, is the prediction of the model at $x$: $v_{f,x}(D) = f(x)$. In addition, to use Shapley values as an “explanation” of the (grand coalition of) features in this way, these methods also need to specify how $v_{f,x}$ acts on proper subsets of the features.

The definitions of Shapley sampling values \cite{395}, as well as SHAP values \cite{269}, are derived from defining $v_{f,x}(S)$ as the conditional expected model output on a data point when only the features in $S$ are known:

$$v_{f,x}(S) = E[f(X)|X_S = x_S] = E_{X_S|x_S}[f(x_S, X_{\bar{S}})] \quad (1.3)$$

Quantitative Input Influence (QII) \cite{103} draws on ideas from causal inference to propose simulating an intervention on the features not in $S$, thus breaking correlations with the features in $S$:

$$v_{f,x}(S) = E_D[f(x_S, X_{\bar{S}})] \quad (1.4)$$

where the distribution $D$ is derived from the product of the marginal distributions of the features in $\bar{S}$. The approach of using a distribution other than that of the original data was further generalized by \cite{283}, who also propose the Formulate, Approximate, Explain (FAE) framework, so as to unify a number of different approaches to Shapley value explanations.

<table>
<thead>
<tr>
<th>Method</th>
<th>$v_{f,x}(S)$</th>
<th>$\hat{v}_{f,x}(S)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>KernelSHAP, SHapley sampling values</td>
<td>$E_{X_S</td>
<td>x_S}[f(x_S, X_{\bar{S}})]$</td>
</tr>
<tr>
<td>QII, FAE, Interventional TreeSHAP</td>
<td>$E_D[f(x_S, X_{\bar{S}})]$</td>
<td>$E_D[f(x_S, X_{\bar{S}})]$</td>
</tr>
<tr>
<td>Conditional TreeSHAP, Frye et al. \cite{157}, Aas et al. \cite{3}</td>
<td>$E_{X_S</td>
<td>x_S}[f(x_S, X_{\bar{S}})]$</td>
</tr>
</tbody>
</table>

Table 1.1: Proposed value function $v_{f,x}$ for each method, compared with the quantity $\hat{v}_{f,x}$ the algorithm actually approximates. The interventional distribution $D$ used depends on the method (i.e., for KernelSHAP it is the observational joint distribution of $X$).

**Algorithms** Methods based on the same value function can differ in their mathematical properties based on the assumptions and computational methods employed for approximation. TreeSHAP \cite{271}, an efficient algorithm for calculating SHAP values on additive tree-based models such as random forests and gradient boosting machines, can estimate $E_{X_S|x_S}[f(x_S, X_{\bar{S}})]$ by observing what proportion of the samples in the training set matching the condition $x_S$ fall into each leaf node, a method which does not rely on a feature independence assumption. In the algorithm for KernelSHAP \cite{269}, conditional expectations are estimated by assuming feature independence; samples of the features in $\bar{S} = D \setminus S$ are drawn from the marginal joint distribution of these variables. This
effectively approximates an expectation over an interventional distribution instead, though in a slightly different way from QII.

In Table 1.1, we categorize each method based on how they define a value function $v_{f,x}(S)$ and how they estimate that value function $\hat{v}_{f,x}(S)$. In the rest of the paper, we will refer to these value functions as either interventional or conditional based on the estimation method. That is to say, KernelSHAP, Shapley sampling values, QII, and FAE are interventional methods, while TreeSHAP as well as some other algorithms we will introduce later are conditional.

1.2 Mathematical critiques

We now present a number of mathematically articulated problems that arise when we attempt to interpret Shapley values as feature importance measures. These problems arise from the estimation procedures that are in use as well as the fundamental axiomatic structure of Shapley values.

Conditional versus interventional distributions

A fundamental difference between the interventional and conditional value functions is revealed by what we call the indirect influence debate. Suppose $f$ is defined with domain $\mathbb{R}^d$, but for a certain feature $i$, $f(x) = f(x')$ whenever $x_j = x'_j$ for all $j \neq i$; that is to say, intervening on the value of $x_i$ alone does not change the output of $f$. We call this a variable with no interventional effect.

Should a feature with no interventional effect be considered an “input” to this function? We could define a new function $f'$ with domain $\mathbb{R}^{d-1}$ to perfectly capture the output, so perhaps not. What if, in the relevant input space, $x_i$ is a statistical proxy for some $x_j$ which does affect the output of $f$? Shapley value based feature importance methods must grapple with these choices.

Adler et al. [10] take the information-theoretic position that “the information content of a feature can be estimated by trying to predict it from the remaining features.” This perspective can help diagnose situations where an undesirable proxy variable is being used by a model, as in the classic case of redlining. While Adler et al. go on to analyze how the accuracy of a model depends on indirect information, the conditional value function aligns with this information-theoretic principle as well: If a certain feature $i$ can help predict the features in $\bar{S}$, then the quantities $v_{f,x}(S \cup i) = E[f(X)|X_{S \cup i} = x_{S \cup i}]$ and $v_{f,x}(S) = E[f(X)|X_S = x_S]$ may be meaningfully different, meaning that the marginal contribution of feature $i$ is nonzero. For this reason the Shapley value of the conditional value function may attribute influence to features with no interventional effect, a positive thing from the perspective of Adler et al..

Merrick and Taly [283], on the other hand, criticize the capacity to attribute indirect influence as being paradoxical, and show that interventional methods will never attribute attribute influence
to an $x_i$ which has no interventional effect on $f$, which they see as a desirable property.

Unfortunately, the decision between the two types of value functions is a catch-22. Both methods introduce serious issues: Choosing a conditional method requires further modeling of how the features are interrelated, while choosing an interventional method induces an “out-of-distribution” problem.

**Issues with conditional distributions** The conditional value function induces two major difficulties. First, the exact computation of the Shapley value for a conditional value function would require knowledge of $2^d$ different multivariate distributions, and so a significant amount of approximation or modeling is necessary. Second, since influence can be computed on an arbitrarily large set of features, it becomes necessary to choose a set that is meaningful because the explanations may change based on which features are considered.

Solutions have been proposed to deal with the computational complexity of this problem. The TreeSHAP algorithm estimates the conditional expectations of any tree ensemble directly, without sampling, using information computed during model training. The algorithm utilizes information about the training instances which fall into each leaf node to model each conditional distribution. It is not, however, set up to attribute influence to variables without an interventional effect, as the trees contain no information about the distribution of variables not in the model.

For arbitrary types of models, estimating the conditional expectations requires a substantial amount of additional modeling of relationships in the data which are not necessarily captured by the model that one is trying to explain. Aas et al. [3] and Frye et al. [157] have developed methods that aim to generate in-distribution samples for the relevant calculations.

Even if computational issues are resolved, there are additional inconsistencies introduced by the capacity of the Shapley value to attribute influence to an arbitrarily large feature set given a single function. The modeler must decide which features count as players in the cooperative game and which are redundant, and since the problem definition posits that the attributions add up to the value of $f(x)$, this choice can affect the resulting explanations.

Consider the addition of a redundant variable $C$ to a dataset with two features, $A$ and $B$, so that $P(X_C = X_B) = 1$. Suppose a model $f$ is trained on all three features. Intuitively, the features $B$ and $C$ should be equally informative to the model and so should have the same Shapley value under the conditional value function. Formally, the following properties will hold:
so this means $v_{f,x}(B) = v_{f,x}(C)$ and $v_{f,x}(AB) = v_{f,x}(AC) = v_{f,x}(ABC)$. Therefore, for any data instance $x$,

$$
\phi_v(A) = \frac{1}{3} \Delta_v(A, \emptyset) + \frac{2}{3} \Delta_v(A, BC) \\
\phi_v(B) = \phi_v(C) = \frac{1}{3} \Delta_v(B, \emptyset) + \frac{1}{6} \Delta_v(B, A)
$$

Now consider what would happen if we defined a new function $f'(x_A, x_B) = f(x_A, x_B, x_B)$. For any data instance, since $x_B = x_C$, $f'(x) = f(x)$. It is effectively the same model for all in-distribution data points, so the games $v_{f,x}$ and $v'_{f,x}$ are the same for all subsets of variables. Yet if we choose to limit the scope of our explanation to two variables instead of three, the attribution for both $A$ and $B$ will come out to be different:

$$
\phi'_{v}(A) = \frac{1}{2} \Delta_v(A, \emptyset) + \frac{1}{2} \Delta_v(A, BC) \\
\phi'_{v}(B) = \frac{1}{2} \Delta_v(B, \emptyset) + \frac{1}{2} \Delta_v(B, A)
$$

Notice that $\phi'_v(B)$ is neither equal to $\phi_v(B)$, its assigned influence in the 3-variable setting, nor $\phi_v(B) + \phi_v(C)$, the “total” influence of the two identical variables in the 3-variable setting. The relative apparent importances of $A$ and $B$ thus depend on whether $C$ is considered to be a third feature, even though the two functions are effectively the same.

It is not obvious whether two statistically related features should be considered as separate “players” in the cooperative game, yet this choice has an impact on the output of these additive explanation models. Suppose, for instance, that $B$ is a sensitive feature, and $C$ is a non-sensitive feature that happens to perfectly correlate with it. Two different “fairness” audits of the same function would come out with quantitatively different results.

Frye et al. [157] propose to a solution to the problem in terms of incorporating causal knowledge:

...If $x_i$ is known to be the deterministic causal ancestor of $x_j$, one might want to attribute all the importance to $x_i$ and none to $x_j$. 

$$
E[f(X)|X_B, X_C] = E[f(X)|X_B] \\
= E[f(X)|X_C] \\
E[f(X)|X_A, X_B, X_C] = E[f(X)|X_A, X_B] \\
= E[f(X)|X_A, X_C]
$$
They propose not only discounting fully redundant variables which are causal descendants of other variables in the model, but relaxing the symmetry axiom which uniquely defines the Shapley value. Instead of averaging marginal contributions over every permutation, they suggest defining a quasivalue which considers only certain permutations; for example, orderings which place causal ancestors before their descendants.

In this framework, fully redundant features will receive zero attribution and will not change the resulting value of the remaining features. For instance, in the above example, if variable $C$ were known to be a causal descendant of $B$, the Asymmetric Shapley Values of $A$ and $B$ under $f'$ will be the same as they were under $f$.

A fully specified causal model is not required to use this method: they “span the data-agnosticism continuum in the sense that they allow any knowledge about the data, however incomplete, to be incorporated into an explanation of the model’s behaviour.” The results in Frye et al. [157] demonstrate, however, the sensitivity of the game theoretic approach to the amount of prior knowledge about the relative agency of each feature, which we consider a significant limitation of the approach.

There are thus both practical and epistemological challenges with computing the Shapley values of games with a conditional value function.

**Issues with interventional distributions** Conditional value functions introduce undesirable complexities to the feature importance problem, so those inclined against methods with the capacity for attributing indirect influence may prefer the methods interventional value functions instead.
These methods, however, are highly sensitive to properties of the model which are not relevant to what it has learned about the data it was trained on.

Methods which use an interventional value function fundamentally rely on evaluating a model on out-of-distribution samples (Figure 1.2). Consider, for example, a model trained on a data set with three features: \( X_1 \) and \( X_2 \), both \( N(0,1) \), and an engineered feature \( X_3 = X_1X_2 \). To calculate \( v_{f,x}(\{1,2\}) \) for some \( x = \{x_1, x_2, x_3\} \), we would have to estimate \( E(f(x_1, x_2, X_3)) \) over some distribution for \( X_3 \) which does not depend on \( x_1 \) or \( x_2 \). Therefore we will almost certainly have to evaluate \( f \) on some sample \( \{x_1, x_2, x'_3\} \) which does not respect \( x'_3 = x_1x_2 \) - thus, it is well outside the domain of the actual data distribution. The model \( f \) has never seen an example like this in training, and has therefore not learned much about this part of the feature space. Its predictions on this feature space are not necessarily relevant to the task of explaining an in-distribution sample, yet the explanations will be affected by them.

This “out-of-distribution” phenomenon has been explored recently by Hooker and Mentch [204], who show why “permutation-based” methods to evaluate feature importance can be highly misleading: when values are substituted into feature set \( \bar{S} \) that are unlikely or impossible when conditioned on feature set \( S \), the model \( f \) is forced to extrapolate to an unseen part of the feature space. They show that these feature importance methods are highly sensitive to the way in which the model extrapolates to these edge cases, which is undesirable information for a model “explanation” to capture.

Slack et al. [379] demonstrate how to exploit this sensitivity by devising models which illegally discriminate on some protected feature for in-distribution samples, but exhibit different behavior on the out-of-distribution samples used by KernelSHAP so as to simulate “fairness” in the resulting explanations. By manipulating the model’s behavior on unfamiliar parts of the feature space, they can twist the explanations on the familiar part to their will.

These challenges illustrate that intervening on a subset of features of a data case before applying a model trained on a sample from a certain distribution is inherently misleading.

Additivity constraints

In addition to the problems demonstrated above, which have to do with the choice between two families of value functions, we also identify problems which are common to both. These are linked to the axiomatic underpinnings of Shapley values.

For any two of the axioms that define the Shapley value, there exists an alternative attribution between players which satisfies those two but not the other; the Shapley value is therefore only unique because it satisfies all three. Since the notion of the sum of two games is not especially meaningful, the Additivity axiom has been described by game theorists as “mathematically convenient” and
“not nearly so innocent as the other two” [312]. The choice to constrain the value to be unique in this way has implications for what kinds of models can be explained intuitively by the Shapley value. Even in simple cases where feature independence renders the interventional versus conditional debate irrelevant, we find the Shapley value conceptually limited for non-additive models.

The Shapley value seems to intuitively align with what is considered important in an additive setting. Consider applying any of the expectation value functions to $f(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d$ where the features $X_i$ are independent. For any subset $S$,

$$v_{f,x}(S) = E_{X_S|X_S}[f(x_S, X_{\bar{S}})]$$

$$= f(x_S, E[X_{\bar{S}}])$$

$$= \sum_{j \in S} \beta_j x_j + \sum_{j \in S} \beta_j E[X_j] + \beta_0$$

so the marginal contribution for feature $i \notin S$ is

$$\sum_{j \in S \cup i} \beta_j x_j + \sum_{j \in S \cup i} \beta_j E[X_j] + \beta_0$$

$$- \left( \sum_{j \in S} \beta_j x_j + \sum_{j \in S} \beta_j E[X_j] + \beta_0 \right)$$

$$= \sum_{j \in S} \beta_j x_j + \beta_i x_i + \sum_{j \in S \cup i} \beta_j E[X_j] + \beta_0$$

$$- \left( \sum_{j \in S} \beta_j x_j + \beta_i E[X_i] + \sum_{j \in S \cup i} \beta_j E[X_j] + \beta_0 \right)$$

$$= \beta_i (x_i - E[X_i])$$

In this way, the Shapley value is supported by the common intuition that coefficient size, if variables are appropriately scaled, signals importance in a linear model.

The additivity axiom is aligned with additive models in another way: the games resulting from two models sum to the expectation game of the sum of the two models. This seems reasonable when the models are additive in the first place. But imagine if the additivity constraint were relaxed. We could use an alternative attribution $\psi$ which satisfies the other two axioms: $\psi: \psi(i) = v(i)$ for $i \in U$ and $\psi(i) = \frac{1}{|U|} (v(D) - \sum_{j \in U} v(j))$ where $U$ is the set of dummy features. Using the expectation value function in this setting, any feature which did not satisfy $\beta_i (x_i - E[X_i]) = 0$ would get the same attribution. In this sense the additivity constraint seems necessary for a game-based feature attribution to provide any meaningful quantities about an additive model. Under an interventional interpretation of the attribution — using the values to assess which data changes produce the largest model prediction change — this is not a helpful property.
Under an interventional interpretation, Shapley values are as uninformative for non-additive models as this alternative attribution is for linear ones. For instance, any value function which always evaluates to 0 except on the grand coalition will evenly distribute influence among players. Consider a model given by $f(x) = \Pi_{j=1}^d x_d$ where the features are independent and centered at 0. Then for any subset $S$,

$$v_{f,x}(S) = E[f(X_S, X_{\bar{S}})|X_S = x_S]$$

$$= E[\prod_{j=1}^d X_d|X_S = x_S]$$

$$= \prod_{j=1}^d E[X_j|X_S = x_S]$$

$$= \left(\prod_{j \in S} x_j \right) \left(\prod_{j \in \bar{S}} E[x_j]\right)$$

which, since $E[x_j]$ is 0, is always 0 unless $S = D$. Then the Shapley value for every feature $i$ is $\frac{1}{d}f(x)$, regardless of the value $x_i$. Even if, for instance, the magnitude of one of the variables is much higher than the other. This property will, in fact, hold for all multiplicative functions of independently distributed, zero-centered data. Thus, while Shapley values are touted for their “model-agnostic” quality, under the lens of a particular interpretation, this is not the case.

### 1.3 Human-centered concerns

In this section we turn to the human side of the interaction between feature importance methods and the people who use them. This perspective is closer in spirit to the “human-grounded metrics” that Doshi-Velez and Kim [116] describe in comparison with the “functionally-grounded evaluation” of the previous section.

We use the framework set out by Selbst and Barocas [363], who argue that there are three general motivations behind the call for explanations in AI:

The first is a fundamental question of autonomy, dignity, and personhood. The second is a more instrumental value: educating the subjects of automated decisions about how to achieve different results. The third is a more normative question—the idea that explaining the model will allow people to debate whether the model’s rules are justifiable.

In this section, we attempt to reconcile the Shapley value feature importance formalization of machine learning “explanations” with these three goals. We argue that the theoretical properties
of the Shapley value are not naturally well-suited to any one of these objectives. While we focus here on these issues in the context of Shapley values, many of these critiques also apply to other explanatory methods.

**Explanations as contrastive statements** The presence of the phrase “right to explanation” in the GDPR illustrates the sense many of us have that it is inherently unethical to make decisions about an individual without providing an explanation, in a way that Selbst and Barocas [363] argue has more to do with “procedural justice” than “wanting an explanation for the purpose of vindicating certain specific empowerment or accountability goals.”

It is not immediately clear how to formally evaluate a method that provides explanations merely because it should, rather than to improve on a particular metric or task. In this setting, Doshi-Velez and Kim suggest the empirical approach of running user tests where humans are provided with explanations and they evaluate their “quality”. But in fact, what humans consider a good explanation has been studied extensively in the social sciences, leading to several formal theories of how humans generate and select explanations.

Miller [285] provides an overview of this literature. One of his major findings is that the way humans explain phenomena to each other is through contrastive statements:

People do not explain the causes for an event per se, but explain the cause of an event relative to some other event that did not occur; that is, an explanation is always of the form “Why P rather than Q?”, in which P is the target event and Q is a counterfactual contrast case that did not occur.

He attributes this insight to work by Lipton [261]. More recently, a similar argument has been made by Merrick and Taly [283], referencing earlier work by Kahneman and Miller [223].

We now outline different ways in which Shapley values can be interpreted as contrastive explanations.

**Shapley value sets as a single contrastive statement** The above-mentioned research supports the hypothesis that people ask for explanations when the outcome, P, is “unexpected” compared to the outcome Q. In this sense, we can interpret Shapley-based explanations as a contrastive statement where the outcome to be explained is $v(D)$ and the foil – the counterfactual case which did not happen – is implicitly set to be $v(\emptyset)$. In the “local” settings described earlier, $v(D)$ is $f(x)$ and $v(\emptyset)$ is $E(f(x))$:

$$f(x) = E(f(x)) + \phi_1 + \phi_2 + \ldots + \phi_d$$

Thus, the Shapley values can be thought of as a set of answers to the question, “Why $f(x)$ rather than $E(f(x))$?”
While the expected value of a function seems like a natural foil to an “unexpected” \( f(x) \), due to the properties of the expectation, there may not be a scenario in the data space of \( X \) with the outcome \( E(f(x)) \). Thus, the expected value may not be “expected” by anyone with a reasonable understanding of the situation at hand at all.

If we are willing to consider intervention distributions (Section 1.1), then the framework provided by Merrick and Taly [283] provides a slightly different contrastive explanation: in their setting, the Shapley value assignment can be thought of as a set of answers to the question, “Why \( f(x) \) rather than \( f(r) \)?”, where \( r \) is chosen from the reference distribution. This of course requires the specification of the reference distribution and carries with it the estimation issues described above in Section 1.2.

**Marginal contributions as contrastive statements** An alternate way to consider Shapley value-based methods as contrastive statements is by examining the marginal contribution of features. The set of marginal contributions of each feature \( i \), which are averaged in a certain way over all subsets \( S \) to calculate the Shapley value, can be thought of as a set of contrastive explanations. Each quantity \( \Delta(i, S) \) represents a contrastive explanation for why feature \( i \) is important: “Why choose a model with \( S \) and \( i \) rather than a model with just \( S \)? Because it improves \( v \) by \( \Delta(i, S) \) amount.” This quantity is an important part of stepwise selection, a modeling procedure in which features which increase the accuracy of a model are successively added to the modeling set.

Note that regardless of what order features were actually added to the model in, all permutations are considered when the Shapley value is calculated. It is not clear that taking an average of quantities representing “all possible contrastive explanations” for a certain set of foils is a sensible way to summarize information. Instead, Miller [285] argues that humans are selective about explanations: certain contrasts are more meaningful than others. An example of this is the difference between necessary and sufficient causes:

Lipton argues that necessary causes are preferred to sufficient causes. For example, consider mutations in the DNA of a particular species of beetle that cause its wings to grow longer than normal when kept in certain temperatures. Now, consider that there are two such mutations, \( M_1 \) and \( M_2 \), and either is sufficient to cause the mutation. To contrast with a beetle whose wings would not change, the explanation of temperature is preferred to either of the mutations \( M_1 \) or \( M_2 \), because neither \( M_1 \) nor \( M_2 \) are individually necessary for the observed event; merely that either \( M_1 \) or \( M_2 \). In contrast, the temperature is necessary, and is preferred, even if we know that the cause was \( M_1 \).

Consider, without specifying how to quantify the importance \( v \) of a feature coalition, computing some kind of allocation for each feature to analyze the positive classification of a beetle with
longer wings. Lipton’s argument above suggests that since all “yes” cases share a property $T$, a contrastive statement highlighting this is more relevant than comparisons based on $M_1$ or $M_2$. This is fundamentally at odds with the idea that the “yes” prediction should be split additively between different coalitions of $M_1$, $M_2$ and $T$, a property induced by the notion of the Shapley value.

Using Shapley-valued based methods to enable action  One motivation for “explaining” a function is to enable individuals to figure out how to achieve a desirable outcome. For example, one might allow an individual to query the model for a specific contrastive explanation in which the person $p$’s outcome, $f(p)$, is compared with a person $q$ with desirable outcome $f(q) = Q$ determined by the user, such that the user might be able to alter their own situation to approximate $q$. This setup has been formalized as the “counterfactual explanation” problem by Wachter et al. [425] (with an analysis of hidden assumptions by Barocas et al. [29]). Ustun et al. [415] further specify a way to model this problem by searching for changes within characteristics which are actually mutable; they call this the “actionable recourse” problem (with a corresponding analysis by Venkatasubramanian and Alfano [419]).

Unlike these methods, Shapley value based frameworks do not explicitly attempt to provide guidance how a user might alter one’s behavior in a desirable way. Further, observing that a certain feature carries a large influence over the model does not necessarily imply that changing that feature (even significantly) will change the outcome favorably.

Suppose, in a very simple nonlinear example, that a univariate model is defined as $f(x) = 2 - (x - 1)^2$, for some $X \sim N(0, 1)$. A person for whom $x = 1$ will get $f(1) = 2$, and $E(X) = 0$, so the Shapley value for this person’s single input is then $\phi(x) = 2$. Suppose they were hoping for an even higher score. The fact that the value is positive, along with the general knowledge that 1 is a bit high with respect to an average value of $X$, might make this person think that increasing their $x$ value even more will increase their score – but it will not.

This problem stems from the fact that the contrastive quantity $E(f(x))$ is not desirable, but even if $v(\emptyset)$ is chosen to be some desirable outcome $f(q)$ of some $q$, such as in Merrick and Taly [283], the Shapley values themselves do not correspond to specific actions: the interventional effect of changing one input from $x$ to that from $q$ is just one of the marginal contributions that are averaged together to form the Shapley value of that input, as we discussed in Section 1.3.

Shapley-based explanations for normative evaluation  Shapley-value-based explanations are primarily used for purposes of normative evaluation: deciding whether a model’s behavior is acceptable [42]. This is done either at the development stage, to help a human evaluate a model, or at the decision-making stage, to help a human evaluate a specific decision made by a model. In this section we explore how the information content of the Shapley value is insufficient for evaluation.
We marshal evidence to make three points. Firstly, data scientists do not have a clear mental model of what insights Shapley-value-based analysis brings. Secondly, in the face of this uncertainty, they tend to rely on narrative and confirmation biases. Thirdly, even if they do understand the analysis, it is not obvious that it can be operationalized for specific evaluation tasks.

Since there is no standard procedure for converting Shapley values into a statement about a model’s behavior, developers rely on their own mental model of what the values represent. Kaur et al. [231] conducted a contextual inquiry and survey of data scientists to observe their interpretation of interpretability tools including the SHAP Python package. They found that many participants did not have an accurate mental model of what a SHAP analysis represents, yet used them to make decisions on whether the model was ready for deployment, over-trusting and misusing the tool.

Using feature importance during model development in this way is ripe for narrative and confirmation biases. Passi and Jackson [322] conducted ethnographic fieldwork with a corporate data science team and described situations in which applying intuition to feature importance was a key component of the model development cycle. In one instance, when developers communicated the results of a modeling effort to project managers, the stakeholders immediately decided it was “useful” based entirely on the feature importance list:

Certain highly-weighted features matched business intuitions, and everyone in the meeting considered this a good thing. …Regarding counter-intuitive feature importances, [a data scientist] reminded [the stakeholders] that machine-learning models do not approach data in the same way humans do. He pointed out that models use “a lot of complex math” to tell us things that we may not know or fully understand.

This suggests that even when an individual lacks a correct mental model of the meaning of Shapley values, they may use them to justify their evaluation anyway, whether or not this analysis is well-founded.

In support of this hypothesis, empirical studies have shown that interpretability is not always helpful in task-specific settings. Poursabzi-Sangdeh et al. [330], for instance, demonstrated that “interpretable” models may not be easier to evaluate:

Participants who were shown a clear model with a small number of features were better able to simulate the model’s predictions. However, contrary to what one might expect when manipulating interpretability, we found no improvements in the degree to which participants followed the model’s predictions when it was beneficial to do so. Even more surprisingly, increased transparency hampered people’s ability to detect when the model makes a sizable mistake and correct for it, seemingly due to information overload.
This suggests that common intuition for the benefits of interpretability (and the types of questions it can help answer) may be based on faulty assumptions, and these questions should instead be concretely specified and tested. For instance, data scientists might want to know:

- Whether an error was made at any point in the data processing pipeline for a certain feature
- Whether the model is acting upon spurious correlations or other artifacts of training data
- Whether the model exhibits inappropriate biases
- Whether the model’s accuracy will improve if a certain feature is included or excluded

While Shapley-value-based methods might help qualitatively inform investigations that lead to answers to these questions, it is not clear that they provide direct answers to any specific question related to the points of interest above. Weerts et al. [429], for instance, conducted a human-grounded evaluation of SHAP and did not find evidence that it helped users assess the correctness of predictions.
Chapter 2

Quantifying the limitations of Shapley value explanations

This chapter is also adapted from joint work with Carlos Scheidegger, Suresh Venkatasubramanian, and Sorelle Friedler [249].

2.1 Shapley residuals

Motivated by the previous chapter, we introduce Shapley Residuals, vector-valued objects that capture a specific type of quantitative information lost by Shapley values. Shapley residuals can be associated with individual variables, as well as with sets of variables. When the residual of a feature exhibits a large norm, the associated Shapley value should be taken with skepticism: the resulting importance is not just due to the variable acting by itself. On the other hand, if a residual is small, most of the effect of the variable on the model is explainable by the variable acting independently (we make these statements precise in Section 2.1). The Shapley residual, then, communicates important details about what the explanation actually represents.

To build an intuition for why this is an important problem, consider an algorithm which makes admissions decisions purely on the basis of gender and department: \( f(g, d) = g + d - 2dg \), where \( g = -1 \) if the applicant is male and \( g = 1 \) otherwise, and there are two departments, represented by \( d = 1 \) and \( d = -1 \). In this contrived scenario, the applicant is admitted if \( f(g, d) > 0 \) (which only happens when \( g \) and \( d \) have different signs) and is rejected otherwise. Clearly, the admissions decision is affected by gender—yet if each of the two variables are distributed with mean 0, the KernelSHAP values [269] which are supposed to explain the decision \( f(1, 1) = 0 \) are both 0, since to
compute the Shapley value, each features’ univariate and interaction influences are averaged together and cancel each other out. In this way, the computation of the Shapley value has implicitly obscured a discriminatory effect, and the corresponding nonzero Shapley residuals would demonstrate that the Shapley values are not telling the whole story.

To more precisely describe what Shapley residuals capture, consider the following two motivating scenarios. First, suppose a practitioner uses Shapley values to determine the effect of data interventions on model outcomes. Consider two models $f_1$ and $f_2$. In a real-world scenario, the practitioner will often only have black-box access to such models, and the models will often be significantly more complex. Here, we use these simple models:

$$f_1(x_1, x_2, x_3) = x_1 + x_2 + x_3$$
$$f_2(x_1, x_2, x_3) = x_1 + 2x_2x_3$$

Suppose the practitioner seeks to explain the output $f_1(1, 1, 1) = 3$ or $f_2(1, 1, 1) = 3$, using KernelSHAP to compute local feature importances. For both models, the Shapley values of $x_1$, $x_2$, and $x_3$ are all 1. Despite that, intervening by increasing the value of $x_2$ changes $f_2$ more than increasing the value of $x_1$; in $f_1$, this clearly does not happen. The Shapley residuals for all variables in $f_1$ are zero, indicating that variables in $f_1$ do not interact (as we prove in Section 2.1). The Shapley residuals for $x_2$ and $x_3$ in $f_2$, on the other hand, are nonzero, while the Shapley residual of $x_1$ is still zero. Finally, the Shapley residual for the set of variables $\{x_2, x_3\}$ is also zero. As we show in Section 2.1, these statements imply the following behavior for variables of $f_2$: $x_1$ has no interactions with other variables (its residual is zero); $x_2$ and $x_3$ interact with other variables (their residuals are non zero); $x_2$ and $x_3$ only interact with each other (the residual of the set $\{x_2, x_3\}$ is zero). Thus, access to Shapley residuals gives warning that intervening on $x_2$ or $x_3$ in $f_2$ could act differently than $x_1$ due to an interaction between $x_2$ and $x_3$.

In the second scenario, consider a data generating distribution where $\alpha$ controls the correlation between two features in $X$ and a regression target $y$:

$$(X, y) \sim \mathcal{N}\left(\begin{bmatrix} 0 & 0 \\ 1 & \alpha \\ \alpha & 1 \end{bmatrix}, \langle X, (3, 1) \rangle \right).$$

We examine a regression model $f(x_1, x_2) = \beta_1 x_1 + \beta_2 x_2$ determined via linear least squares. Assume access to infinitely many IID samples from $(X, y)$, $\beta = (3, 1)$. Suppose a practitioner wanted to explain the output of $f(1, 1) = \beta_1 + \beta_2$, this time using Conditional Expectation SHAP [398]. The Shapley values are $\beta_1 + \alpha(\beta_2 - \beta_1)/2$ for $x_1$ and $\beta_2 + \alpha(\beta_1 - \beta_2)/2$ for $x_2$. When $\alpha \approx 0$, Shapley values correspond to model weights $\beta_1, \beta_2$, and support a (valid) interventional interpretation that changing $x_1$ yields a larger change to the output of $f$ than does $x_2$. However, if $\alpha \approx 1$, Shapley
Table 2.1: KernelSHAP game for Example 1 - the input \((1,1,1)\) to \(f(x_1, x_2, x_3) = x_1 + 2x_2x_3\) where \(x_i\) are iid \(\mathcal{N}(0,1)\) features.

<table>
<thead>
<tr>
<th>S</th>
<th>Hypercube Coordinate</th>
<th>(v(S)) Definition for explaining ((1,1,1)) with KernelSHAP</th>
<th>(v(S)) Value given i.i.d. (x_i \sim \mathcal{N}(0,1))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\emptyset)</td>
<td>(0,0,0)</td>
<td>(E[f(x)])</td>
<td>0</td>
</tr>
<tr>
<td>({x_1})</td>
<td>(1,0,0)</td>
<td>(E[f(x)</td>
<td>x_1 = 1)</td>
</tr>
<tr>
<td>({x_2})</td>
<td>(0,1,0)</td>
<td>(E[f(x)</td>
<td>x_2 = 1)</td>
</tr>
<tr>
<td>({x_3})</td>
<td>(0,0,1)</td>
<td>(E[f(x)</td>
<td>x_3 = 1)</td>
</tr>
<tr>
<td>({x_1, x_2})</td>
<td>(1,1,0)</td>
<td>(E[f(x)</td>
<td>x_1 = 1, x_2 = 1)</td>
</tr>
<tr>
<td>({x_1, x_3})</td>
<td>(1,0,1)</td>
<td>(E[f(x)</td>
<td>x_1 = 1, x_3 = 1)</td>
</tr>
<tr>
<td>({x_2, x_3})</td>
<td>(0,1,1)</td>
<td>(E[f(x)</td>
<td>x_2 = 1, x_3 = 1)</td>
</tr>
<tr>
<td>({x_1, x_2, x_3})</td>
<td>(1,1,1)</td>
<td>(E[f(x)</td>
<td>x_1 = 1, x_2 = 1, x_3 = 1)</td>
</tr>
</tbody>
</table>

values do not support this interpretation. A practitioner employing Shapley values alone lacks the information to distinguish these scenarios. Shapley residuals provide useful diagnostic information; the norm of the residuals for \(x_1\) and \(x_2\) is exactly linearly proportional to \(\alpha\).

In these simple scenarios, it is clear that Shapley residuals capture, respectively, variable interactions and mismatches between dependent features in the data and independent variables in the model. As we show in Section 2.2, these observations apply to real-world scenarios as well.

**Partial gradients on the hypercube** Recall our definition of the \(\nabla\) operator: Let \(\mathbb{R}^V\) be the space of functions from \(V\) to \(\mathbb{R}\) and let \(\mathbb{R}^E\) be the space of functions from \(E\) to \(\mathbb{R}\). In particular, the game \(v\) is an element of \(\mathbb{R}^V\). The differential operator \(\nabla : \mathbb{R}^V \to \mathbb{R}^E\) is then defined as

\[
\nabla v(S, S \cup \{i\}) = v(S \cup i) - v(S)
\]

for any \(v \in \mathbb{R}^V\).

We will also define a partial gradient \(\nabla_i : \mathbb{R}^V \to \mathbb{R}^E\):

\[
\nabla_i u(S, S \cup \{j\}) = \begin{cases} u(S \cup j) - u(S) & i = j \\ 0 & \text{otherwise} \end{cases}
\]

Intuitively, \(\nabla_i\) evaluates a gradient for edges corresponding to the insertion of \(i\), and takes the value 0 everywhere else. On the hypercube, only edges on the \(i\)th axis of \(\nabla_i v\) will take a nonzero value. See the Edge Space portion of Figure 2.1(a) for an illustration of this procedure on the running example.

**Geometric characterization of Shapley values** A geometric interpretation of Shapley values dates back at least to Kleinberg and Weiss [241], showing they can be expressed in terms of projections...
from the space of games to the space of cooperative games with independently contributing players. A key advance was made by Stern and Tettenhorst [392], building on earlier work by Candogan et al. [61] who proposed viewing the game as a scalar function defined on the hypercube and studying its discrete gradient. To understand this advance, we first introduce a special class of games.

**Inessential games** Let $I$ denote the space of games $v$ such that for all $S \subseteq [d]$, $v(S) = \sum_{i \in S} v(\{i\})$. $I$ is called the space of inessential games. Intuitively an inessential game is one in which the player interactions are simple and additive: every player adds a fixed value $v(\{i\})$ to a coalition $S$ independent of the composition of $S$. Inessentiality is a key feature of what makes Shapley values attractive for feature importance – if each contribution is fixed and combines additively, we have a natural interpretation for how much each feature contributes to the overall model output. Specifically, if a game is inessential, it then follows that the Shapley value for player $i$ is $v(\{i\})$. In our running example using KernelSHAP, this is $E[f(x)|x_i = 1]$, the contribution (averaged over other variables) of the variable $x_i$.

In general though, a game might not be inessential. The key insight of Stern and Tettenhorst [392] was to express inessentiality of games in terms of gradients on the hypercube.

**Proposition 1** ([392, Prop 3.3]). The game $v$ is inessential if and only if for each $i \in [d]$ there exists $v_i \in \mathbb{R}^V$ such that $\nabla_i v = \nabla_i v_i$.

The main result by Stern and Tettenhorst [392] is a decomposition of an arbitrary game $v$ into games that are “close to being inessential” and allow extraction of Shapley values. If $v$ is not inessential, we cannot be sure to find $v_i$ such that $\nabla_i v = \nabla_i v_i$, but we can find the “closest” such $v_i$ as the solution to the least squares problem

$$\min_{x \in \mathbb{R}^V, x(\emptyset) = 0} \|\nabla x - \nabla v\|$$

**Theorem 2** (Stern and Tettenhorst [392]). Given a game $v$, let $v_i$ be defined as above. Then

1. $\sum v_i = v$
2. If $v(S \cup \{i\}) = v(S)$ for all $S \subseteq [d]$, then $v_i = 0$
3. For any $\alpha, \alpha' \in \mathbb{R}$ and games $v, v'$, $(\alpha v + \alpha' v')_i = \alpha v_i + \alpha' v'_i$
4. If $\pi$ is a permutation of $[d]$ and $\pi \circ v$ is the game $\pi \circ v(S) = v(\pi(S))$, then $(\pi \circ v)_i = v_{\pi(i)}$

Consider the mapping $\phi(v)(S) = \sum_{i \in S} v_i([d])$. The above result implies this is a Shapley mapping and therefore $\phi_i(v) = v_i([d])$ are the Shapley values of $v$. We illustrate the construction in Figure 2.1(a).
**Shapley Residuals**  The inessentiality of a game is inextricably linked to the meaningfulness of Shapley values for the reasons given above. The idea we explore now is the converse: can the degree to which a game is not inessential provide insights into where Shapley values are not able to capture feature influence?

By the fundamental theorem of linear algebra, we can write

\[
\nabla_i v = \nabla v_i + r_i
\]

where \( r_i \) is orthogonal to \( \nabla v_i \). This allows us to interpret \( r_i \) (a vector with one value for each edge of the hypercube) as a measure of deviation from inessentiality, because by Proposition 1, this vector is identically 0 if and only if the game is inessential.

We can generalize these ideas further to subsets of players. We begin with a generalized notion of inessentiality:

**Definition 2.** The game \( v \) is inessential relative to \( S \) if \( v(C) = v(S) + v(C \setminus S) \) for all \( S \) and \( C \) such that \( S \subset C \subset [d] \).

That is, each coalition containing \( S \) obtains a value equal to the subcoalition \( S \) working separately from \( C \setminus S \); in this sense, inessentiality with respect to \( S \) can speak to the lack of interactions between \( S \) and its complement. In addition, inessentiality relative to a single player \( i \) is the same as inessentiality relative to the singleton set \( \{i\} \).

Next, we generalize the notion of a partial derivative.

**Definition 3.** For a subset \( S \subset [d] \), let \( \nabla_S : \mathbb{R}^V \to \mathbb{R}^E \) be the operator \( \nabla_S = \sum_{i \in S} \nabla_i \), or

\[
\nabla_S u(C, C \cup \{j\}) = \begin{cases} 
\nabla u(C, C \cup \{i\}) & \text{if } i = j \text{ and } i \in S, \\
0 & \text{otherwise}.
\end{cases}
\]

We can now prove a result similar to Proposition 1 for relative inessentiality.

**Proposition 2.** The game \( v \) is inessential relative to \( S \) if and only if there exists \( v_S \in \mathbb{R}^V \) such that \( \nabla_S v = \nabla v_S \).

To understand the limits of Shapley values, we propose to quantify the degree of deviation from inessentiality with the following definition:

**Definition 4 (Shapley Residuals).** We call \( r_i = \nabla_i v - \nabla v_i \) the Shapley Residual of player \( i \). Analogously, \( r_S = \sum_{i \in S} r_i \) is the Shapley Residual of set \( S \).

Shapley Residuals are a novel diagnostic tool for feature importance, and enjoy a number of relevant properties.
Proposition 3. If \( v \) is inessential, then \( v \) is inessential relative to all \( i \in [d] \) and all subsets \( S \subset [d] \).
If \( v \) is inessential with respect to each player of \( i, j, \ldots, z \) then \( v \) is inessential relative to the set \( \{i, j, \ldots, z\} \).

The proof of this proposition is in the appendix. The following corollaries are straightforward.

Corollary 1. \( v \) is inessential iff \( r_i = 0 \) for each \( i \in [d] \).

Corollary 2. \( v \) is inessential relative to \( S \) iff \( r_S = \sum_{i \in S} r_i = 0 \).

This allows us to interpret \( \sum_{i \in N} ||r_i||^2 \) as the deviation from inessentiality of \( v \) and \( ||\sum_{i \in S} r_i||^2 \) as the deviation from inessentiality of \( v \) relative to \( S \).

In this paper we will focus on the computation and evaluation of residuals with respect to individual players i.e \( r_S \) for \( S = \{i\} \). Figure 2.1(b) illustrates the construction of residuals. Algorithm 1 describes how to compute residuals.\(^1\)

2.2 Residuals for explanation methods

We have established that the norm of the residual \( r_i \) characterizes the degree to which the value function \( v \) is not inessential with respect to the player \( i \). We now show how to interpret this when

\[^1\text{We can take an unconstrained minimum here and subtract } v_i(\emptyset) \text{ at the end because adding a constant value to } v \text{ does not change } \nabla v.\]
Algorithm 1: Exactly calculate the $i$th Shapley value and Shapley residual of $v$

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Compute $\nabla_i v$</td>
</tr>
<tr>
<td>2.</td>
<td>Solve $v_i = \arg\min_{x \in \mathbb{R}^V}</td>
</tr>
<tr>
<td>3.</td>
<td>Compute $\nabla v_i$</td>
</tr>
<tr>
<td>4.</td>
<td>Return Shapley residual $r_i = \nabla_i v - \nabla v_i$</td>
</tr>
<tr>
<td>5.</td>
<td>Return Shapley value $\phi_i = v_i(S) - v_i(\emptyset)$ where $S$ is the set of all players</td>
</tr>
</tbody>
</table>

Attributing feature importance via Shapley values for two popular methods. As has been noted, the different methods for Shapley value-based explanation (whether local or global) all reduce to a specific choice for the game $v$, at which point the Shapley values of $v$ are estimated and returned [398, 248, 283].

The definitions of Shapley sampling values [395], as well as SHAP values [269], are derived from defining $v$ as the conditional expected model output on a data point when only the features in $S$ are known: $v^\text{cond}_{f,x}(S) = E[f(X)|X_S = x_S]$. We call this Conditional Expectation SHAP after Sundararajan and Najmi [398].

The Interventional SHAP value function, which defines KernelSHAP, is derived from defining $v$ by taking an expectation of $f$ over the joint distribution of $\bar{S}$ while fixing the feature values from $S$: $v^\text{int}_{f,x}(S) = E[f(|x_S, \bar{X}_S|)]$. Notably, the two values are the same if the features in $\bar{S}$ are independent from those in $S$.

We will show that the residual $r_S$ captures the degree to which interactions between the features in $S$ and its complement arise in the model or in the data, depending on which form of Shapley-based feature importance is used to define the value function $v$.

Interventional SHAP: Recall the problem of explaining two models where $f_1(x_1, x_2, x_3) = x_1 + x_2 + x_3$ and $f_2(x_1, x_2, x_3) = x_1 + 2x_2x_3$. Note that in the first model all three variables contribute independently to the model output, whereas in the second model the variables $x_2$ and $x_3$ interact in their contribution. We can compute the associated residuals $r_1, r_2, r_3$ and their norms for these two models. For the first one, all residuals are identically zero. However, in the second model if $x_2$ and $x_3$ are nonzero for a certain input, they will have a nonzero residual. In other words, the residual captures feature interactions in the model. Our first result, which we prove in the appendix, shows that this intuition can be made precise.

Lemma 1. Let $f : X = \{X_1, X_2, ..., X_d\} \to Y$ be a multivariate function. Suppose $f$ can be decomposed as $f(x) = g(x_S) + h(x_{\bar{S}})$, for some functions $g : \{X_j : j \in S\} \to Y$ and $h : \{X_j : j \notin S\} \to Y$. Let $z = \{z_1, z_2, ..., z_n\} \in X$. Then $v^\text{int}_{f,z}$ is relatively inessential with respect to the set $S$.

This is important because if the model really does decompose additively for a certain variable $i$, the practitioner understands what to expect when variable $i$ is perturbed. The Interventional
Shapley residuals thus quantify the extent to which the SHAP values must be augmented with more information to capture interaction effects in the model.

**Conditional Expectation SHAP** As the residual for Interventional SHAP can be thought of as detecting feature interactions in a model, the residuals of Conditional Expectation SHAP can detect feature interactions in the data. Let $X \sim \mathcal{N}([0,0]^T, \Sigma)$ for $\Sigma = \begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix}$, and let $Y = f(X) = \beta^T X$ (note that ordinary least squares will recover $f$ in the limit of infinite data). Given input $x_1, x_2$, the SHAP values of $f$ are $\phi_1 = \beta_1 x_1 + \alpha \frac{\beta_2 x_1 - \beta_1 x_2}{2}$, $\phi_2 = \beta_2 x_2 + \alpha \frac{-\beta_1 x_2 + \beta_2 x_1}{2}$. That is to say, they are linearly dependent on the correlation between the two variables. In particular, consider explaining the input $[1,1]$ to the function $x_1 + 3 x_2$; the SHAP values are $\phi_1 = 1 + \alpha$ and $\phi_2 = 3 - \alpha$ and the residuals are both $2\alpha$. Notably, as interaction between variables increases in the data (measured by $\alpha$), the residual increases and the SHAP values deviate further and further from the coefficients of the actual model. We can make this intuition precise.

**Lemma 2.** Let $f : X = \{X_1, X_2, \ldots, X_d\} \rightarrow Y$ be a multivariate function. Suppose $f$ can be decomposed as $f(x) = g(x_S) + h(x_{\bar{S}})$, for some functions $g : \{X_j : j \in S\} \rightarrow Y$ and $h : \{X_j : j \notin S\} \rightarrow Y$. Let $z = \{z_1, z_2, \ldots, z_n\} \in X$. Suppose further that all $X_j : j \in S$ are distributed independently from all $X_j : j \notin S$. Then $v_{f,z}^{\text{Cond}}$ is relatively inessential with respect to set $S$.

The residual on Conditional Expectation SHAP thus quantifies the extent to which an interpretation of the SHAP values can be interpreted as interventional, because depending on the causal structure of the data, correlated features could imply that perturbing a feature $i$ could result in the perturbation of a different feature as well.

**Inspecting Shapley residuals in practice** Shapley residuals are vectors in the same space as gradients, and are generally high-dimensional entities; a full study of their properties remains an important topic for future work. The characterization in this section shows that the norm of the residual vectors captures important limitations of Shapley values. Thus, our experiments use the scaled norm of the residual vectors, defined to be the norm of the residual vector divided by the norm of the discrete gradient vector. Normalized residuals make them easier to compare across experiments.

**Experiments** Having theoretically justified Shapley residuals in previous sections, we now focus on illustrating what these residuals can help us understand about models on a real-world dataset. Throughout, we use our own implementation of KernelSHAP to calculate the exact Shapley values.
and residuals (see Algorithm 1 in Section 2.1). Some additional experiments can be found in the appendix.

On comparisons to other feature importance methods We note that Shapley residuals are not a feature importance evaluation method, nor are they an “explanation method” in and of themselves. Rather, they are a quantification of the (valuable) information lost by Shapley values. A direct comparison of different feature influence evaluation methods makes sense when there is a clear objective to compare against. Such an objective doesn’t really exist here. Rather, we choose to provide an internal validation that lays out the mathematical foundation on which the method rests. This allows a user to decide the context in which to employ one method or another. For example, as we discussed in Section 2.3, Shapley-Taylor indices and Shapley residuals appear to capture different kinds of interactions that are potentially of interest to a user. There is no meaningful way to compare them in a vacuum because one is not “better” than another.

![Figure 2.2: Shapley values and residuals on a decision tree for the Occupancy Detection task](image)

Variable Interactions in Occupancy Detection Consider the Shapley values and residuals for an occupancy detection dataset (20,560 instances) used to predict whether an office room is occupied. The 7 attributes include a date stamp for an hour and day of the week. A decision tree model with maximum depth 3 is trained on 75% of the data using the features light and hour. When evaluated on the remaining test set, the ROC-AUC for this decision tree is 0.991. We then calculate the Shapley values and residuals (using 50 randomly sampled background rows from the test set) for 1000 randomly sampled test instances. The results for the variable “light” are shown in Figure 2.2.

---

2Code is provided in the supplementary material

3https://archive.ics.uci.edu/ml/datasets/Occupancy+Detection
The reason that the cluster of points in the middle has a high residual is illustrated in Figure 2.2(c). Calculating the expected prediction while fixing a light value of 320, unlike most other possible values, results in a mix of low and high predictions. These average to 0.4, while both the overall expectation and particular prediction for occupancy probability for those points are 0.25.

\[
\begin{align*}
E[f(H, L)] &= .24 \quad \begin{array}{c}+ .01 \end{array} \quad E[f(10, L)] = .25 \\
+ .16 \downarrow &\quad \text{E}[f(H, 320)] = .40 \quad \begin{array}{c}− .39 \end{array} \quad f(10, 320) = .01
\end{align*}
\]

Figure 2.3: Geometric representation of the KernelSHAP game for \(f(10, 320)\), where arrows to the right indicate inclusion of the light feature and arrows down indicate inclusion of the hour feature.

Specifically, the KernelSHAP game for \(f(H, L) = P(\text{occupant} = T)\) for \(L = 320\) and \(H = 10\) is shown in Figure 2.3. \(L = 320\) is a positive indicator of occupancy if \(H\) is unknown (+.16) but is a “negative” indicator of occupancy if \(H\) is known to be 10 (-.24), due to the interactions in the model in this area of the feature space. The light Shapley value is close to 0 for points in this range, then, because it is the average of a positive and negative number – not because it is of “low importance” – and the non-inessentiality of this feature is what is being captured by the residual.

2.3 Discussion and limitations

**Relationship with Other Interaction Indices** [400], similarly recognizing that Shapley values lose information about interactions, proposed Shapley-Taylor Interaction Indices, a generalization of Shapley values which attributes influence among interaction terms. Specifically, the Shapley-Taylor explanation for \(x\) of order \(k\) assigns values \(I_k^S\) to subsets of features \(S\) of size \(|S| \leq k\) such that \(\sum I_k^S = f(x)\). The terms for the subsets for which \(|S| < k\) represent a discrete Taylor series around \(v(\emptyset)\). When \(|S| = k\), \(I_k^S\) is defined similarly to the Shapley value: a discrete derivative averaged over permutations.

Our residuals capture fundamentally different information about interactions than Shapley-Taylor. Consider some subset \(S\) for which \(|S| < k\). Our residual \(r_S\) is 0 when the marginal value of adding \(S\) to a coalition \(W\) is constant with respect to sets for which \(W \cap S = \emptyset\); in other words, it is about the presence or absence of interactions of \(S\) with other variables. The Shapley-Taylor interaction index \(I_k^S\), on the other hand, is 0 when \(v(S)\) can be inferred from the values of \(v(W)\) for \(\{W \subset S\}\). The Taylor terms of the Shapley-Taylor explanation thus capture information about how the players in \(S\) interact with each other when no other variables are involved. For instance, if for a certain game \(v(\{i\}) + v(\{j\}) = v(\{i, j\})\), this means that the term \(v(\{i, j\})\) provides no interaction information
about the two players, and $I^k_{\{i,j\}}$ for explanation sizes $k > 2$ will be 0.

However, the Taylor indices for a coalition $S$ say nothing about whether the variables within $S$ interact once a player outside of $S$ is involved. Consider a three-player game between $a, b,$ and $c$, where $v(\{a\}) + v(\{b\}) = v(\{a, b\})$ and $v(\{a\}) + v(\{c\}) = v(\{a, c\})$; this would make $I^k_{\{a, b\}}$ and $I^k_{\{a, c\}}$ equal to 0, implying that $a$ and $b$ do not interact, and $a$ and $c$ do not interact. But it could be that $v$ is not relatively inessential with respect to $a$. If $v(\{a, b, c\}) - v(\{b, c\}) \neq v(\{a, c\}) - v(\{c\})$, then $a$’s relative contribution with respect to $c$ changes once $b$ is involved. This constitutes an interaction between $a$ and $c$ that is not described by the Shapley-Taylor index for $\{a, c\}$, but is rather captured by the third-order interaction of $\{a, b, c\}$. In this scenario, our residuals would show $r_a \neq 0$, alerting us to the fact that $a$ interacts with $\{b, c\}$; additionally, it would have $r_{\{a, b\}} \neq 0$, alerting us to the fact that $\{a, b\}$ interacts with $c$.

In general, since $r_i$ captures information about all of $i$’s interactions, we can state the following connection between the two Shapley extensions:

**Lemma 3.** Given subset $S$, $|S| < k$, if $\exists i \in S \text{ s.t } r_i = 0$, then the Shapley-Taylor index $I^k_S = 0$.

We have focused our attention on Shapley-Taylor interaction indices because of their proposed use for explanations. It should be noted that they (as well as the Shapley interaction index proposed in [271]) are special cases of a general class of interaction indexes investigated in a long line of work starting with [316] and surveyed in [158] (including the Grabisch-Roubens [172] Shapley and Banzhaf interaction indices). All of these differ from Shapley residuals – the latter are meant to represent the information lost when computing singleton Shapley values, not their higher-dimensional extensions, which are based on a different notion of a derivative.

**Usability considerations** Our motivation for this work is to contribute further to the theoretical foundation of Shapley-value-based feature importance measures and, critically, to introduce Shapley residuals to quantify missing importance. Our goal is that residuals be a warning attached to specific Shapley values and thus alert practitioners to model complexities and importances that have previously gone unattended. Further research is needed to investigate whether these residuals can be effectively utilized by humans to make better decisions about their models. An empirical, human-centered investigation is critical because, like Shapley values themselves, the meaning of these residuals may be hard for practitioners to understand, and therefore errors in the interpretation of these residuals may cause unanticipated negative consequences.

**Performance considerations** SHAP implementations provide a partial evaluation of the game vector [269, 268], which provides analysts with results even in high-dimensional settings. Unfortunately, there is no assumption-free provable bound on the relationship between partially evaluated
game vectors and actual Shapley values. Our goal here is more precisely characterize the information not conveyed by Shapley values and so we always compute the full vector. Thus, the runtime is ultimately exponential in the number of variables to analyze. This currently limits the number of variables for which Shapley residuals can be practically computed to 20 to 30 (with corresponding vectors of length between a million and a billion elements). Since the derivative operator is sparse and well-conditioned, the least squares problem is efficiently solved by the LSQR method [319]. Still, in future work, we hope to efficiently identify whether a particular residual is nonzero, and approximate properties of residuals which capture the entirety of non-linear interactions of a particular feature.

**Conclusion**  A goal in interpretable machine learning, and within Shapley-value-based feature importance, is to give a rigorous theoretical foundation to interpretability notions so that practitioners can better understand the impacts of their models. This is especially important in contexts where models make high-stakes decisions about people, e.g., via criminal risk assessments and interview screening algorithms. We believe people have the right to understand those decisions, and particularly which features were important for the decision. Putting such feature importance measurements on solid theoretical grounds is important for the validity of these feature importance claims. Their validity is an important part of the ethics of algorithms as societal interventions.
Chapter 3

Using the lens of feminist epistemology to critique feature importance

In recent years, the number of new methods for measuring feature importance for machine learning (ML) models has exploded, leaving ML practitioners spoilt for choice. As black-box algorithms with inscrutable inner mechanisms are increasingly used for crucial decisions, demands for greater transparency and accountability have increased, leading to legal requirements for explanation like that in the European Union’s General Data Protection Regulation (GDPR). Media coverage and public awareness of potentially problematic algorithmic decisions have also increased, forcing ML practitioners to come up with improved and novel ways to explain their models. In this paper, we focus on a subclass of explanation methods, known as feature importance or feature attribution methods. Broadly construed, these assign a quantitative measure of importance to the inputs to a function. They are a popular way of approaching demands for explainability, as they fit the human tendency to attribute outcomes to specific factors.

As feature importance methods proliferate, practitioners struggle to decide which methods to use, and creators of methods strive to demonstrate that their methods are superior to others. However, much of the evaluation of various methods’ pros and cons proceeds in an implicitly value-neutral context: rarely do participants in this discourse pause to consider if certain supposed theoretical or practical virtues would be virtues in all social or applicational contexts.\(^1\) Papers introducing new

\(^{1}\text{The exceptions to this tend to occur in interdisciplinary venues like the FAccT conference [29], or in venues for human-computer interaction professionals [230, 201].}
feature importance methods state that certain properties of the proposed methods are “desirable” without indicating the contexts or uses for which those properties are desirable [399, 270]. Stating a property as “desirable” without further elaboration elides questions such as: Desirable to whom? For what purposes? When applied to what types of data?

In this chapter, adapted from joint work with Leif Hancox-Li [179], we argue that a number of popular feature importance methods implicitly contain values that are counter to the kind of pluralistic, contextual, and interactional view of epistemology advocated by many feminist epistemologists. The existence of these values is not problematic in itself unless one adheres to the ideal of a “view from nowhere” [293]. However, the partiality of feature importance methods is often not acknowledged by the creators or users of those methods. Instead, debates focus on various apparently value-neutral properties of the methods, to the point where some aspire to create universal benchmarks to evaluate all explanation methods [117, 359].

In the next section, we provide a brief overview of feminist epistemology, in particular, the frameworks of situated knowledges and standpoint epistemology. Then we describe the various ways in which values counter to feminist epistemology may be embedded in feature importance methods. Finally, we offer some suggestions on how ML research on explanations can include more feminist values.

3.1 **Background: Feminist epistemology and philosophy of science**

Against the unitary view from nowhere that is assumed in much of science and mainstream analytic epistemology, feminist epistemologists have argued for understanding science through the idea of situated knowledges. Instead of aiming for one correct, objective view of things, they argue that we should instead accept that knowers are social beings, who bring perspectives to each issue that are conditioned by their social experiences. In this view, there is no “god trick” that lets us know from the point of view of the unmarked [180]. It is not, however, a completely relativistic view—we can still speak of views being “less false”, even if we cannot speak of truth or truth-likeness [181].

Feminist epistemologists came to this view as a reaction against what they saw as problematic claims to unitary knowledge in professions, like the natural sciences, where largely white and male knowers have constructed one way of seeing as the “objective” one, while alternative ways of knowing, many propounded by marginalized groups in science, are ignored [233]. Feminist epistemologists have described numerous examples of how the natural sciences could have benefited from accommodating more ways of knowing [232, 274]. Indeed, some argue that subjugated knowledges, that is, the perspectives of those with the least power, have an advantage in providing more empirically accurate
and comprehensive accounts of the world [181, 444]. Given the non-representative nature of machine learning practitioners' demographics [60], we should be similarly wary of claims by ML practitioners about certain methods being universally or objectively better than others.

Indeed, compared to many of the natural sciences examined by feminist epistemologists, there may be even greater reason to be wary of claims of value-neutrality in ML, because much of the value of ML explanations comes from their utility in practical contexts. The same type of explanation may be useful in one type of practical application but not another. This is unlike many hypotheses in the natural sciences, which are often evaluated not solely by their practical utility, but also by how they fit with existing scientific theory as part of a coherent, realistic description of the world.

In contrast to the view from nowhere, the situated knowledge framework focuses on the pragmatics of acquiring knowledge in particular contexts. This kind of contextual thinking includes considering how traditional epistemic goals intersect with social goals. As Sandra Harding points out, “The kinds of explanations favored by modern science have not always been the most effective ones for all projects—for example, for achieving environmental balance or preventing chronic bodily malfunctions” [181]. In the case of ML explanations, the intertwining of epistemic and social goals is particularly striking, since the demand for an explanation is often not just a demand for knowledge of how the algorithm works, but also part of an ethical demand for greater accountability or transparency [118, 334].

Another aspect of feminist epistemology that is particularly relevant to ML explanations is its implications for formal frameworks. As Donna Haraway argues, “rational knowledge does not pretend to disengagement... to be fully self-contained or fully formalizable. Rational knowledge is a process of ongoing critical interpretation among ‘fields’ of interpreters and decoders” [180]. In reducing a complex system to one formal framework, we often reduce the possible meanings it can have and the possible relations it can have to other systems. Focusing on formalisms also distracts from the meanings we attach to formal symbols or formulae, even though the same formalism may take on different meanings in different social or applicational contexts.

Haraway also advocates for a critical stance on boundaries and objects. By paying attention to how objects of study are partly constructed by social processes, we can also see how the boundaries that define those objects “materialize in social interaction” [180]. Similarly, in taking a feminist approach to ML explanations, we should critically question how we draw the boundaries around what the realm of “explanation” is. Are we defining the realm of legitimate study of explanations in an unnecessarily narrow way? Are there possible agents who can help construct explanations or be

---

2For example, a feature-highlighting explanation may provide an actionable guide to a decision subject on how to change an algorithmic decision in some contexts, but not in others, depending on what factors the explanation recommends that the subject change [29]. Some factors are impossible to change due to physical impossibility or environmental factors. However, the same feature-highlighting explanation that is unhelpful to a decision subject may be useful to a data scientist doing feature selection or feature engineering.
part of explanations, that we have left out of our idealized view of explanations?

Finally, feminist epistemology also encourages a more interactive way of how knowledge is created, considering the object of knowledge not just as a passive object with static properties, but also as an agent that can enter in conversation with the knower. In this view, “[t]he world neither speaks itself nor disappears in favor of a master decoder” [180]. For example, Evelyn Fox Keller attributes Barbara McClintock’s successful insight into maize genetics to McClintock’s willingness to erode the boundaries between subject (the scientist) and object (the maize plants) [233]. Keller also attributes McClintock’s openness to questioning the central dogma of molecular biology to her lack of investment in the idea of the passivity of nature. A related idea, from black feminist epistemology, is that of using dialogue to assess knowledge claims—interacting with the object of your knowledge rather than observing it from a detached distance [81].

In short, feminist epistemology provides a more pluralistic, contextual, interactional, and critical corrective to various traditionally “masculine” ways of doing science, a corrective that can also be applied to how we create, use, and evaluate feature importance methods in ML. As we’ll see in the next section, the apparently objective endeavor of figuring out how to attribute feature importances contains various implicit values, many of which are counter to the principles of feminist epistemology.

3.2 Values in feature importance

We define feature importance to be any quantitative assignment of importance or influence to each input feature of some model learned from data. How the notion of “importance” is formalized varies between methods, and can be defined in reference to either the entire model and its training procedure, or to the model’s prediction on one particular input.

To provide broader context for our more general claim that feature importance methods can be inflected with particular epistemic values, a look at the history of feature importance methods is instructive. This history shows how contingent choices influenced by practitioners’ social environments led to the current state of affairs. In particular, the widespread modern emphasis on predictive accuracy as the preeminent virtue of feature importance comes from instrumentalist epistemic values championed by the likes of Leo Breiman.

Informal descriptions of the utility of feature importance often cite the high dimensionality of useful models, implying that they could be made smaller [253, 74, 175]. In the popular textbook “Elements of Statistical Learning,” for instance, Hastie, Tibshirani and Friedman write, “In data mining applications the input predictor variables are seldom equally relevant. Often only a few of them have substantial influence on the response; the vast majority are irrelevant and could just as well have not been included. It is often useful to learn the relative importance or contribution of each
input variable in predicting the response” [186]. Indeed, in the traditional, statistical data-modeling setting, feature importance is often described in terms of feature selection (such as in the case of F-tests), with the ultimate goal being a simplified model that is more likely to describe a real phenomenon. Data modeling in this sense has an emphasis on description, that is, the objective of scientific truth.

In describing a change of direction away from data modeling, Matthew Jones argues that modern research in data science is born “more from an engineering culture of predictive utility than from a scientific culture of truth” [222], and the same can be said of feature importance. This can be traced to Leo Breiman’s novel proposal of using feature importance for purposes other than feature selection. Breiman argued against “assuming that the data are generated by a given stochastic data model” in favor of a new algorithmic culture [52].

Breiman proposed a new, influential view of data analysis which de-emphasized the importance of developing probabilistic models of underlying data generating processes. He argued that a prediction-centered view was the right way to investigate the relationship between the response and predictor variables, and that black-box algorithmic models can provide more reliable and interesting information than weakly predictive data models can: “The goal is not interpretability, but accurate information. Interpretability is a secondary goal that can be finessed.” Regardless, he proposed a general way to measure the importance of predictors in a model: “My definition of variable importance is based on prediction. A variable might be considered important if deleting it seriously affects prediction accuracy” [52].

This logic makes sense on its face, but Breiman never worked towards providing a “satisfactory theoretical definition” [52] of importance. The mechanism of variable deletion, a step in the process of feature selection, was the basis of the analysis; but for Breiman, the advantage of using trees and ensembles of trees was exactly that it allowed one to avoid doing feature selection: “Existing methods, he and his collaborator Meisel argued in a report to the US Air Force, required the analyst to choose, without reference to the data, a means for reducing its dimensionality” [221]. Breiman’s version of feature importance was not intended to help create more veridical models of reality by deleting extraneous information, and this is precisely the reason that permutation feature importance, his approximation of the definition he intuited, has been shown to be problematic [203]. Breiman wanted to calculate the change in prediction accuracy under the removal of information, but the removal is simulated instead of actually tested because there was no intention of actually removing the feature.

Yet permute-and-predict notions of importance persisted throughout recent algorithmic history. Hooker and Mentch observed that this was because of their pragmatic properties: “they are each computationally cheap, requiring \(O(N)\) operations, apply to the \(f(x)\) derived from any learning method, have no tuning parameters, and in relying only on averages, they are statistically very
stable. Moreover, the approach is readily understood and easily explained to users in applied areas” [203]. Permutation feature importance influenced many ideas throughout the history of the problem of calculating feature importance for black-box models [19, 149, 99, 10]. As argued in [119], accuracy is not necessarily a value-neutral evaluation criterion; the popularity of black-box algorithmic models reflect a value-laden disciplinary shift in the field of machine learning, and this is paralleled in the history of feature importance.

While instrumentalism of the kind advocated by Breiman may not be particularly feminist or anti-feminist, it’s certainly an epistemic value that became dominant due to historically contingent factors. Replies by Brad Efron and Bruce Hoadley in [52] argued against Breiman’s conception of feature importance and indicate a possible alternate path for feature importance that failed to become the dominant conception. Absent the social factors that influenced Breiman’s thinking and that facilitated his stature in the discipline, perhaps we would have discussions of feature importance that have a more pluralistic tone, rather than discussions that over-emphasize the same “virtues”. The anti-pluralism inherent in Breiman’s championing of predictive accuracy can be seen as counter to feminist epistemology. In addition, some feminists have regarded instrumentalist values as counter to feminism.\footnote{See [195] for an example of this view, and, as a counterpoint, [9] for why anti-positivism or anti-instrumentalism may be unhelpful for feminist goals.}

This historical preamble provides crucial context for what follows, by demonstrating that the epistemic values favored today are the product of historical processes influenced by social factors. In the following sections, we describe more epistemic values that inflect modern feature importance methods, arguing that these values conflict with feminist epistemology.

**Universality as an epistemic value** One aspect of how values enter into feature importance measures is the ways in which these are derived and evaluated. Many methods are arrived at axiomatically, and many proposed evaluation standards pin their hopes on a universal “benchmark” or “gold standard” [117, 359]. We argue that the axiomatic approach towards derivation and the “gold standard” narrative for evaluation embed universal, anti-pluralist values, which go against feminist epistemic values [9, 111]. At the same time, human-centered studies on how explanation methods are used in real-world contexts demonstrate that these methods are used for very different purposes and audiences, cautioning against applying context-free desiderata or evaluation criteria [201, 200].

Many proposals of new explanation techniques take an axiomatic approach, where they specify desirable properties for explanations, then construct or derive a method that has those properties [270, 399, 381, 70, 377, 41]. They also often cite the fact that a method “uniquely” satisfies some set of axioms as a reason to accept it, such as in local explanations based on Shapley values [270, 401].
The Shapley value is the solution to a fundamental problem in game theory, which is uniquely defined with respect to certain axiomatic conditions put forth by many authors as fundamentally desirable. However, even in the original context of game theory, some of these axioms were thought by mathematicians to be “mathematically convenient” [312] and “not so innocent” [265]. By the former, they meant that the solution is constrained to be unique only if it is required to satisfy all of the axioms, which makes it tempting to say that the axioms are desirable. Despite this, the creators of SHAP, a popular feature importance method based on Shapley values, explicitly position their method as embodying what they hope will be universally desirable principles for all explanation methods: “The thread of unity that SHAP weaves through the literature is an encouraging sign that common principles about model interpretation can inform the development of future methods” [270]. Similar remarks on unity can be found in further elaborations on SHAP [99].

There are some understandable reasons for taking an axiomatic approach. As [399] argues, when evaluating an explanation method, it can be unclear if a provided explanation looks wrong because the model itself is wrong, or if it looks wrong because the explanation method is at fault. The axiomatic approach sidesteps these empirical ambiguities by providing a guarantee that the method has certain ostensibly desirable mathematical properties, regardless of its empirical performance. However, proponents of axiomatic approaches position their methods as providing “intuitively” desirable properties [99, 399]. Implicit in this way of justifying axiomatic properties is that the intuition referred to is universally held or applicable to all situations. This ignores the fact that those properties may not be the most appropriate ones in every context of application.

A more pluralistic view of explanatory virtues would cast a critical eye on the idea of universally desirable properties and the ability of a researcher to know what those are a priori. User studies back up the idea that explanation methods are used in very different contexts, for different purposes [201, 200]. In the absence of user studies comparing the usefulness of explanation methods for different purposes, it’s not obvious that a method that is more effective for communicating with non-technical stakeholders is necessarily also more effective for helping data scientists debug a model. As a concrete example, one user study found that SHAP is not very helpful for helping data scientists spot obvious problems with the data [230], but another user study found that SHAP is good at producing explanations that fit human intuition [271]. As we point out in Section 3.3, many papers on ML feature importance methods evaluate the methods on “tasks” that are untied to any specific pragmatic goal, even as the feminist epistemology literature has long pointed out the problems with “supposedly universal principles drawn from bounded artificial examples” [9].

In the broader context of ML research in general, others have warned of the pitfalls of having benchmarks that determine what “success” or “progress” is [119, 106]. Even the organizers of ML...

---

4 See below for examples of specific axioms that are unlikely to be intuitive or valid in many real-world situations.
competitions admit that competing on fixed metrics can lead to a kind of meaningless optimization after a while [135]. These broader lessons also apply to ML explanations. Favoring a fixed number of benchmarks for ML explanations risks channelizing new methods into a few directions, privileging some goals or audiences over others. Without broader participation from different stakeholders in the design of ML explanations, this runs the risk of not just restricting directions of research, but also creating explanation methods that privilege the interests of certain groups over others.

**Additively partitioning influence**  We’ve argued that the idea of a universally applicable explanation method, based on universally desirable axioms, is contrary to pluralism. In addition, some axioms that are touted as obviously desirable contain their own embedded values. One type of axiom that is common in axiomatic approaches towards explanations is an axiom that encodes the additivity of feature importances.

The idea behind this notion is that the individual contributions of each feature, as measured by the feature importance method, should add up to the interventional effect of all the features taken together, thus partitioning the total effect of the feature set among its components. This axiom either goes unnamed [24] or appears under different names in different papers, such as “completeness” [399], “summation-to-delta” [373], or “local accuracy” [270]. In particular, given a machine learning model $f$ which computes a prediction for a specific instance $x$, the axioms can be described as the following.

**Axiom 1 (Completeness).** Given feature attributions $\phi_1, \phi_2, \ldots, \phi_n$ corresponding to features 1 through $n$ for a certain input $x$ to a function $f$, $f(x) = f(x') + \sum_{i=1}^{n} \phi_i$ for some baseline input $x'$.

**Axiom 2 (Local Accuracy).** Given feature attributions $\phi_1, \phi_2, \ldots, \phi_n$ corresponding to features 1 through $n$ for a certain input $x$ to a function $f$ and its simplified features $h_i(x)$ representing the “presence” of feature $i$, $f(x) = \phi_0 + \sum_{i=1}^{n} \phi_i h_i(x)$ for some baseline $\phi_0$.

These two axioms are equivalent if we take $\phi_0 = f(x')$ and define $h_i(x)$ to be 1 for all $i$.

A stronger desideratum, which requires the sums of individual contributions to be proportional to the joint effect for each subset of features, is called faithfulness [41]. Faithfulness is proposed as a metric, rather than an axiom, which measures closeness to an ideal: a perfect faithfulness score implies completeness and local accuracy.

**Definition 5 (Faithfulness).** Given feature attributions $\phi_1, \phi_2, \ldots, \phi_n$ corresponding to features 1 through $n$ for a certain input $x$ to a function $f$, for any subset $S$ of features, let $x_S$ be equal to $x$ with the features not in $S$ replaced with those from some baseline vector $x'$. Then faithfulness is defined as

$$\text{corr}_{S \subseteq \{n\}} \left( \sum_{i \in S} \phi_i, f(x) - f(x_S) \right)$$
Axioms like completeness, local accuracy, and faithfulness align with the natural properties of coefficient size in linear models of independently distributed features. Given a linear model $f(x) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$, it is easy to see what happens when we change the features in some set $S$ to be taken from some baseline vector $x'$: the difference between $x$ and $x_S$ is exactly $\sum_{i \in S} \beta_i (x_i - x'_i)$.

Thus, the feature importance term $\phi_i = \beta_i (x_i - x'_i)$, where $x'$ is taken to be the empirical mean of the data, satisfies Completeness and Local Accuracy, as well as minimizing Faithfulness, and is in fact the value assigned by SHAP for this model. In this linear setting, $\phi_i$ represents the univariate effect of perturbing feature $i$ from its expectation to the value represented in the input.

If a user interprets feature importance to represent univariate effects in all modeling settings, however, this property is fundamentally misleading in many applications of complex ML models. Often, input features are highly correlated, and ML models learn to infer complex interactions between them. Additivity axioms seek to impose an ideal which does not hold in the real world. Imposing additivity in real-world conditions thus attempts to optimally simplify real interactions in the data in order to fulfill some ideal of analytic and cognitive tractability. In doing so, researchers privilege achieving some kind of mathematical ideal over modeling the genuine complexity in the data.

This is not to say that it is always bad to impose any of these axioms. Rather, they serve as examples where a value-laden choice is made, namely, to stipulate a mathematical ideal to obtain certain analytic goals as opposed to creating new tools to handle complexity at face value. This starting from what is formally tractable means that “what is deemed interpretable comes to affect what is presented as explainable, and thus what can be considered meaningful in an algorithmic system” [38]. Other potential sources of meaning are foreclosed.

There might be certain purposes for which favoring a mathematical ideal is appropriate. But for other purposes, such as understanding complex interactions between features without over-simplifying, it is less appropriate. The value-laden nature of this design choice in creating explanation methods is hidden behind the apparently objective and “rigorous” approach of starting from axioms and proving uniqueness theorems based on them.

It is important to acknowledge that the idealized nature of the properties discussed above aren’t just a mathematical ideal, but also a cognitive ideal: it is thought that it’s easier for people to understand and use explanations if the feature attributions behave as described in the relevant axioms. But again, if the actual features do not behave as stipulated in the axioms, then the cognitive ideal merely gives the appearance of understandability. This is potentially dangerous

---

5The related axiom of linearity in [399], that linear combinations of models should result in linear combinations of feature attributions, is justified “intuitively.” Linearity is implied by how we would want coefficients to combine if we were talking about linear regression. This “intuition” thus again presumes that coefficient size is a reasonable way to explain linear models.
when it leads to data scientists having unjustified confidence in their model and thus to premature intents to deploy, as found in [230]. In addition, to the extent that there are cognitive limitations on the kinds of explanations that can be understood, this doesn’t mean that the correct way to impose those limitations is by the deductive approach of starting from axioms that translate those limitations into mathematics. While the deductive approach may be more in line with conventions of scholarship in machine learning, perhaps more inductive approaches that start with empirical work on what users need in specific contexts, rather than using user experiments to validate a deductive approach after the fact, would lead to different types of explanation methods.

**Epistemic values in the data-versus-model debate** Another way in which epistemic values may enter into particular definitions of a feature importance method is the extent to which the explanation emphasizes structure in the data, as opposed to structure in the model. This contrast is pertinent in the ongoing debate about the appropriate type of probability distribution to use in computing various feature importance measures, including (but not limited to) SHAP [248]. The choice is between using “interventional” distributions, which ignore correlations between input variables, and using “observational” distributions, which take those correlations into account.

Specifically, given an input $x$ to a function $f$ and a specific feature set $S$, the question is how to quantify the effect of removing information about $S$ by looking at the value of the quantity $f(x) - f(x_S)$, where $x_S$ is $x$ but with the features in $S$ replaced by samples from some distribution. One option is to intervene on those features, using (for instance) their marginal distributions or a fixed baseline value, as in [103, 283]; the other is to condition on the features not in $S$ and sample from what the values in $S$ might be from the other features in $S$, based on correlations in the observational data.

The authors of SHAP recommend using interventional distributions if we want the explanation to be “true to the model”, and observational distributions if we want the explanation to be “true to the data” [71]. In short, we should use interventional distributions if we want to know how the model works independently of the structure of the particular data it’s used on. On the other hand, they argue, if our interest is in a “natural mechanism in the world,” then the explanation derived from observational distributions captures the important causal features more accurately.

Going back to the importance of contextual thinking in feminist values, we can see that the “true to the model” option abstracts the model away from its application context, which is particularly unhelpful in assessing its impact on real populations. The authors’ justification for interventional distributions in this option is illuminating, as they use an example that is particularly ill-suited for interventional distributions once we carefully consider the context of application.

To compare interventional and observational distributions, the authors of [71] consider a model that predicts whether an applicant will default on a loan. They then test how useful the interventional
and observational SHAP values would be to potential applicants, by considering how effectively an applicant can modify their risk by setting a feature to its mean. However, this way of testing is decontextualized from the reality of the deployment context, because in the real world, it’s often impractical for most applicants to be able to change only one feature without also changing others [29]. Based on this method of testing, the authors conclude that the interventional distribution is better if we want to be “true to the model,” and in addition that “[b]eing true to the model is the best choice for most applications of explainable AI, where the goal is to explain the model itself.” This statement reveals the model-centered orientation that is an implicit epistemic value in their research. A countervailing perspective might consider that many real-world applications work on data that’s highly correlated, in contexts where individuals are often not able to change the value of one feature at a time. The authors implicitly center understanding the model independently of application context. In contrast, feminist epistemologists often criticize the use of simple examples and advocate considering real-life examples in all their complexity [9].

Other researchers have taken sides in this debate in the context of their own work. The authors of [399], for instance, argue that implementation invariance is a desirable property of feature importance: if two different models practically describe the same function, their feature attributions should be the same. Asserting that the mechanics of how the inputs relate to the outputs are not important reflects the rejection of data modeling as understood in [52]. The authors of [399] additionally assert that we should not be able to attribute influence to features not in the model, and instead only explain the model itself. Merrick and Taly share this view, calling features not in the model “irrelevant” [283]. This mathematically implies the choice of interventional distributions in SHAP, as any influence notion that can be captured through correlation could assign influence to unmodelled variables. Although individual counterarguments to these ideas have been pointed out [203, 379], they remain a persistent trend in the explainable ML literature.

**Computational modularity as an epistemic value** Another way in which post-hoc feature importance methods contain embedded values is in design decisions that abstract from real-world complexities and hide implementation details behind an easy-to-use interface, all in a plug-and-play package. Previous research on values in computer science has identified this as a key part of computational culture [273, 168], and indeed, this kind of modular design, where details are hidden except on a need-to-know basis, is part of what constitutes good design in programming [134]. Other researchers have cautioned against the same tendency when it comes to “fairness” solutions

---

6 In particular, the credit data that the authors use in their experiment is probably highly correlated and therefore hard for decision subjects to intervene on “cleanly” in the way the authors do in the experiment. The original dataset used in [71] has been removed from public view, but given that it’s in the context of credit, the criticisms of [29] on how feature-highlighting explanations of credit models are hard for decision subjects to act on are probably applicable to it.
in machine learning, calling it the “Portability Trap” [365]. Lucy Suchman has characterized this tendency as “construing technical systems as commodities that can be stabilized and cut loose from the sites of their production long enough to be exported en masse to the sites of their use” [396]. Assuming that a solution is portable in this way is possible only if you think contextual factors (e.g. the circumstances that the model and explanation are used in) are less important than the solution itself.

In the context of feature importance methods, hiding the details is not a value-neutral decision, especially when the very virtues of the methods themselves are under debate. User research has shown that many data scientists interpret SHAP outputs in a relatively uncritical manner. Many are unable to describe the visualizations produced by these methods, even though they trust them [230]. The highly “usable” nature of the application programming interface (API), which allows data scientists to quickly produce visualizations without having to understand how things work under the hood, encourages this combination of high trust in results and poor understanding of methods. This relates to our account in Section 3.2 of how black-box predictive modeling culture arose out of Breiman’s government contracting milieu, where obtaining results was prioritized over understanding. Ironically, this push for obtaining results has been extended to the case of explanation methods, which ostensibly exist in order to improve understanding.

3.3 Towards pluralistic, contextual, and interactive approaches to explanation

Based on examples of overly universal approaches to ML explanations, and on the values espoused by feminist epistemologists, we suggest some preliminary steps towards creating explanations that are more consonant with feminist epistemic values. Much of the following advice is not necessarily specific to feature importance methods: many parts are also applicable to other aspects of ML. In part, this reflects the fact that ML as a field has not paid much attention to feminist epistemology, so that there are ample opportunities to apply feminist epistemological methods to almost any aspect of ML.

Incorporating subjugated points of view In line with the idea that the viewpoints of the marginalized may have a crucial critical edge over the perspectives of the powerful, we should place more emphasis on ways for subjugated populations to participate in designing ML explanations. Participation was identified as a crucial quality for feminist design in HCI by Shaowen Bardzell, who argued that since knowers are not substitutable for one another, “ongoing participation and dialogue among designers and users can lead to valuable insights that could not be achieved scientifically”
Participatory machine learning practices [275, 385] could be extended from participating in algorithm design to participating in the design of explanations, albeit with an awareness that participatory design is not a cure-all [382].

Incorporating the perspectives of marginalized populations is particularly important when they are stakeholders in the decisions being explained, for example, if they are the subjects of automatic allocations of resources like health or housing. Given the demographics of ML practitioners [60], the designers of algorithms making such decisions are unlikely to have a range of social experiences similar to the decision subjects. Leaving design decisions about end-user explanations up to data scientists would thus likely exclude subjugated points of view.

**Evaluating explanations contextually** One important recommendation in feminist epistemology is to acknowledge the situatedness of knowledge—that different knowers have different perspectives. A situated investigation, then, is one that “forefronts the details of the context of specific people and places at particular points in time, rather than trying to study a system or question with an abstract approach removed from social context” [229]. Elish and boyd [126] argue that the act of developing machine learning models itself should be framed as a situated practice, since humans are engaged in personally interpreting data during the development process. So, too, should any attempt to incorporate explanations into a machine learning system.

In evaluating explanations in a feminist spirit, we should consider various contextual factors that affect how well an explanation can perform its social function, such as how explanations are interpreted by stakeholders, unintended consequences, how explanations perform in realistic applications as opposed to toy datasets or problems, and what contexts exist that could cause an explanation method to fail. By bringing in contextual considerations, feminist epistemology enlarges the problem definition of traditional analytic epistemology to include social and ethical considerations. Similarly, one feminist direction that ML explanations could expand in is enlarging traditional problem definitions for work on explanations. Currently, a large proportion of work follows this template [399, 270, 24, 373]:

1. Propose a new method of generating explanations for black-box models.

2. Show that the new method satisfies certain desirable properties. Often, these properties are defined by axioms and the method is proven to satisfy them. Sometimes, the method is shown to satisfy them by carrying out computational experiments. Rarely are the allegedly desirable properties validated by considering what uptake they’ll get in different social contexts and different audiences.

3. Apply the method to one or more datasets. Use it to explain predictions trained and made on the dataset(s). Validate that these explanations “make sense.”
(3) Is typically the last step of evaluation, but occasionally, researchers include an additional step of conducting user studies on the explanation method. Test subjects are often recruited via general surveying sites like Amazon’s Mechanical Turk, and asked to evaluate explanations with little to no context about why the predictions are being made and what purposes the explanations would serve in a real-world applications [346, 272, 378, 372]. Test subjects are typically asked to evaluate the explanation on relatively context-free properties such as using the provided explanations to improve the algorithm or simulating the model’s predictions, without explaining what the predictions are being used for [346, 330].

However, it is hard to evaluate the virtues of explanations without considering the context in which they are meant to operate—the audience, the goals of the explanation, the type of model or data they are applied to, etc. An explanation that best helps data scientists improve a model is not necessarily the best type of explanation for a customer whose request has been denied by an algorithm—if, for example, the most important features according to the explanation are ones that the customer cannot change [29]. For similar reasons, explanations that are tailored to satisfy legal requirements like GDPR are also not necessarily the most helpful explanations for decision subjects or model builders [124]. Model builders may benefit more from more sophisticated or technical explanations that are harder for general audiences to understand.

To apply lessons from feminist epistemology to ML explanations, a first step would be to evaluate explanations in a more context-sensitive manner. Axioms should be considered in relation to how desirable they are in specific contexts. User studies on different types of consumers of explanations, rather than on generic consumers, would help. In line with the epistemic advantages that may accrue to marginalized persons, user studies focusing on groups marginalized by algorithms may be particularly helpful in surfacing contextual harms that might be overlooked by more privileged groups. Benchmarks or gold standards tend to impose universal, context-insensitive norms, so they should be used with extreme caution or avoided, as we argued in section 3.2. While using benchmarks may foster an impression that the results are more objective, testing explanation methods on users trying to solve real problems, real (and novel) datasets, all embedded in real social contexts, would provide more comprehensive and nuanced information on the practicalities of implementing the methods. De-contextualizing explanation methods allows for easier comparison across contexts, but also abstracts away from social factors that potentially affect the methods’ usefulness.

**Encouraging seamfulness and pluralism** Designers have proposed the idea of “seamful” design as a way to encourage technology users to avoid over-trusting one interpretation and keep multiple interpretations in mind [67]. These ideas can also be applied when we design visualizations for ML explanations. As an antidote against data scientists over-trusting explanations [230], creators of explanatory outputs could design for the following effects [366]:

1. Purposefully block the most obvious interpretations of any visualizations, especially if they are misleading, in order to stimulate new ways of understanding the visualizations. One example where this would be helpful is in designing visualizations of log-odds probabilities to thwart the more “obvious” way of reading those quantities in a more intuitive probability space. Another possibility is to design feature attribution charts in a way that reminds viewers that the separate “bars” displayed for each feature aren’t actually separate in reality, due to strong interactions and correlations between features.

2. Downplay the explanation’s authority. We should work on ways to indicate how charts may be inaccurate or misleading. For example, perhaps feature attributions for strongly correlated features can be presented in a more “uncertain” manner relative to feature attributions for less correlated features. Here, the uncertainty quantification literature could be particularly helpful. Alternatively, instead of presenting feature importances in the familiar forms of bar charts or a list of numbers, they could be presented in less familiar ways, providing room for reinterpretation of what they mean and thwarting tendencies to understand them in the same way one understands other bar charts. In an example described in [366], the frequencies of various email categories were represented by changes in a plant’s shape. By presenting the data in an unfamiliar form, the artists invite active interpretation and reflection, subverting naively straightforward interpretations.

3. Design explanatory visualizations in a way that is ambiguous and thwarts any one authoritative interpretation. One concrete suggestion for doing so is to integrate the spatial distortions that dimensionality reduction algorithms produce directly into cluster visualizations [38]. While this might seem to be at odds with the highly quantitative nature of most explanation methods, the presence of numbers attached to particular features doesn’t mean that those numbers are accurate or adequate representations of the features’ actual importances, especially when the explanation method is tailored to satisfy axioms that don’t hold in the real world (see Sections 3.2 and 3.2). Given the uncertain relationships between those numbers and the actual features in the data, visualizing them as though they are certain and have unambiguous importance values is misleading. For example, one can imagine an interface that includes multiple explanatory accounts of a model and helps users see the differences between them. In contrast, we currently have multiple, discrete explanation methods that each present their own seemingly authoritative accounts, hiding the uncertainty that is inherent in each of them.

---

7SHAP, a popular post-hoc explanation method, displays feature importance values for classification algorithms in a log-odds space [267].
Acknowledging the role of problem definition  Section 3.3 outlined a “recipe” for writing a paper about a new ML explanation method. That this “recipe” exists suggests some kind of common understanding about what the “problem” of ML explanation is, and what counts as a solution to it. One of the lessons of feminist epistemology is to question how boundaries are drawn around objects. The recipe described suggests a certain limited view of what an explanation is, which largely leaves out social context except at the possible but infrequent step of conducting a user study. If we instead question the boundaries of the problem, new research projects open up.

The field of human-computer interaction (HCI) has helped in expanding the boundaries of ML explanation research by doing better user studies [230] and exploring the social functions of ML explanations [200, 201]. In other parts of ML, HCI has also added new perspectives by studying how improving user interfaces can improve ML systems even without intervening on the algorithm [360]. Others have studied how co-creating models with end users leads to more more explainable models and more generalizable models [447].

The increased contributions to this field from HCI is a positive trend that will hopefully accelerate the incorporation of feminist epistemic values into ML explanations. One can only hope for more work that integrates traditional technical work with human-centered approaches, as advocated in [417].

Using formalisms critically  As explained in Section 3.1, feminist epistemology has long taken a critical stance towards fully formalized systems, instead emphasizing the interactive nature of knowledge creation and the importance of exploring multiple possible meanings. Other researchers have remarked on the tendency of ML to over-value formal proofs [262], and the doubts we raise above about axiomatic approaches to explanation can be seen as following this line of thought.

However, jettisoning formalisms altogether is not a solution either, since explaining a formally defined ML algorithm requires engaging with formalisms at some level. Rather, a feminist approach would use formalisms when appropriate, with a careful attention to what possibilities a particular formal framework excludes or renders unsayable, and what it inadvertently favors as prescriptive [13]. When one does use a formal framework, its limitations and idealizations should be clearly acknowledged.

Using interactive approaches to explanation  As explained in Section 3.1, feminist epistemology emphasizes interactional ways of knowing, where the object of study and the knower are engaged in a back-and-forth conversation, rather than a one-directional process where the knower decodes a passive object. Most mainstream ML explanation methods require minimal interaction between the user and the algorithm before creating the explanation, conforming more to the latter mold of an active knower and a passive object. The user typically enters the features they are interested in
and the type of explanation they want (e.g. local or global), and gets a static visualization or some numbers in return.

HCI research suggests there might be more interactive ways of creating explanations. One example is Gamut, a visual analytics system that provides an interactive interface to support the interpretation of generalized additive models [200]. Users found Gamut to be helpful and wanted to use it to understand their own data. One possible way in which interactivity could be extended beyond what Gamut did is by having users interactively design explanation methods, rather than just having an interactive interface for generating pre-determined types of explanations.

**Conclusion** We’ve attempted to apply feminist epistemological values to ML feature importance methods, diagnosing how the ways we construct and evaluate these methods may insufficiently incorporate these values. Many explanation methods over-emphasize theoretical properties that may seem desirable in idealized contexts but have questionable usefulness in many applicational contexts. Popular feature importance methods also embed strong instrumentalist values that arose from a government contracting and defense milieu. Practitioners value computational modularity in the form of plug-and-play software packages for generating explanations. All of these tendencies—de-emphasizing real-world data, instrumentalism, and computational modularity—are value-laden and should not be considered to be objectively desirable in all contexts.

In particular, we have argued that popular methods of constructing and evaluating feature importance methods do not sufficiently incorporate feminist epistemological values such as context-sensitivity, critically interrogating boundaries, taking the expertise of subjugated knowers seriously, pluralism, and interactional ways of knowing. This is not to say that these popular methods are useless, but to point to the uniformity of values that they embody, and suggest further possibilities that fall outside the existing paradigm. To this end, we’ve suggested some directions in which future research on explanations can proceed, guided by lessons from feminist epistemology. In our view, we can incorporate feminist values by better incorporating subjugated points of view into the process of designing explanations, evaluating explanations more contextually, resisting the temptation to create benchmarks, thinking critically about how problem definitions favor some values over others, using formalisms with a critical eye, and working on interactive approaches to constructing explanations.

Our proposals mirror recent calls to move towards more human-centered approaches to machine learning [417]. To this end, it’s heartening that user studies on ML explanations have started appearing in major HCI conferences [230, 201]. However, at mainstream, “technical” ML conferences, this kind of work is still largely relegated to special workshops rather than being in the main program. As ML expands its reach into human life, we hope that ML researchers will also correspondingly acknowledge that this broader reach necessitates allowing a broader set of methodologies into its boundaries and including more diverse approaches towards evaluating explanations.
Part II

Public policy problems in machine learning
Chapter 4

Applying the Equal Credit Opportunity Act to machine learning models

Credit is an essential component of financial well-being for Americans, and unequal access to it is a significant factor in the economic disparities between demographic groups that exist today. For this reason, it is critical to make sure the American lending ecosystem is free of discrimination. In America, there are laws in place which specifically ban discrimination in lending, as well as agencies charged with enforcing them. Today, machine learning (ML) algorithms (sometimes trained on “nontraditional” data) are increasingly being used to allocate access to credit. A vast body of research has demonstrated that ML algorithms can encode many different versions of “unfairness,” thus raising the concern that banks and other financial institutions could—potentially unwittingly—engage in illegal discrimination through the use of this technology.

The nebulous threat of “algorithmic discrimination” poses a challenge to federal regulators, who must decide how, if at all, to update their enforcement practices or issue new guidance in light of these concerns [308], which are often articulated by computer scientists in the abstract and not in terms of the actual practices, data, and algorithms used in this sector. Meanwhile, without specific guidance from regulators, researchers and practitioners who want to study or apply fair ML in this particular setting lack a clear picture of the kinds of tools and metrics that will be useful, legal, and practical for detecting and correcting unfairness in algorithms in this setting.

For these reasons, this chapter, adapted from Kumar, Hines, and Dickerson [250] aims to orient the conversation around fair ML research in the context of predicting credit risk from both perspectives.
We discuss methods proposed by the ML community to measure unfairness in algorithms, and
determine the extent to which they may relate to the principles of the Equal Credit Opportunity Act
(ECOA) and the goals of the federal agencies discussed above. Keeping these metrics in mind, we
contextualize results from fair ML research in the consumer credit setting, and identify specific fair
lending risks throughout different parts of a machine learning system’s development. By analyzing
how these mechanisms are likely to play out in the credit setting, we can be more specific about the
kinds of problems regulators should anticipate and address, rather than repeating the folk wisdom
of “bias in, bias out.” Finally, we discuss specific opportunities for regulators to use their authority
to encourage fair ML practices.

4.1 Background: Fair lending law and regulation

Anti-discrimination legislation

The issue of discrimination in credit lending decisions is not novel to the algorithmic setting. While
lending has been around for centuries, Americans increasingly began to rely on consumer credit to
finance large purchases in the 1950’s and 60’s [351]. During this period, individual loan officers and
specialists were responsible for subjectively determining whether a loan applicant was creditworthy;
numerical methods for estimating credit risk existed but were not widely or systematically used
[78]. This presented a risk of intentional discrimination due to personal bias. Additionally, some
codified lending policies in effect at the time clearly disadvantaged women and minorities: During
congressional hearings, testimonies cited practices such as requiring single women to provide a male
co-signer for a mortgage loan [351, 152].

In the spirit of implementing ideas from the civil rights legislation of the 60’s, which did not
directly address lending, ECOA was passed in 1974 to ensure that all Americans were treated
fairly in a system that determined so much of their economic success. It prohibits creditors from
“discriminat[ing] against any applicant, with respect to any aspect of a credit transaction on the
basis of race, color, religion, national origin, sex or marital status,” among other factors [414]. The
law applies to any organization that extends credit, including loans and credit cards.

The Fair Housing Act, also known as Title VIII of the Civil Rights Act of 1968, prohibits
discrimination in housing on the basis of several protected characteristics, and applies to mortgage
providers. The U.S. Department of Housing and Urban Development (HUD) enforces the Fair
Housing Act, and has specified narrow rules making disparate impact litigation difficult; partly
because of this, mortgage algorithms are not our main focus in this paper.
Data collection rules

At the time of its passing, the ECOA gave the Board of Governors of the Federal Reserve Board (FRB) rulemaking authority to implement the law; this set of rules is known as Regulation B. Regulation B specifically prohibits the collection of information about protected characteristics: “A creditor shall not inquire about the race, color, religion, national origin, or sex of an applicant or any other person in connection with a credit transaction” [344]. Credit transactions, here, can include things like consumer credit, business credit, mortgage loans, and refinancing.

A glaring set of exceptions to this rule are in cases where the Home Mortgage Disclosure Act (HMDA) applies. Passed in 1975, the act requires certain financial institutions to provide mortgage data to the public, and in particular requires lenders to collect and report race and gender information of mortgage applications. The act was drafted in response to the practice of redlining, in which lenders would explicitly identify geographic regions and neighborhoods that they would not lend to because they were inhabited by people of color. This information is used to identify indicators of mortgage discrimination and encourage lenders to comply with ECOA [404].

In the non-mortgage setting, Regulation B contains an additional exception to the ban on collecting protected characteristics: when the information is explicitly collected for self-testing, which is defined as any inquiry “designed and used specifically to determine the extent or effectiveness of a creditor’s compliance with the Act” [344]. In doing so, lenders must make it clear to the applicant that providing the information is voluntary. However, this practice is very uncommon; Slaughter et al. [380] speculate that this is because of a “fear that their collection of the data will validate or exacerbate claims that their decisions are biased.” Self-testing might also be disincentivized if corporations believe that the data itself would ultimately benefit plaintiffs in a potential disparate impact suit.

It may seem counter-intuitive that HMDA requires the collection of sensitive information while ECOA bans it. In fact, both HMDA’s requirement for collecting sensitive information and ECOA’s ban on it are controversial. Some argue that the existence of HMDA provides an important basis of evidence for lawsuits or that the policy itself caused lenders to curb their own discriminatory practices, and thus a similar provision should be in place for non-mortgage lenders [404, 5, 48, 66]. Others, especially banks, have argued that HMDA is unfair, costly, and leads to false accusations of illegal discrimination [217].

The Federal Reserve Board, which was responsible for enforcing ECOA until the Consumer Financial Protection Bureau (CFPB) was established, has considered removing the ban on the collection of protected information several times since the law was originally passed. In 2003, it rejected a proposal to lift the ban and mandate the collection of certain sensitive information [303]. The first reason it cited was that creditors might use this information for discriminatory purposes;
however, many members of Congress, consumer advocates and researchers found this unconvincing [436]. The second was that “many creditors would elect not to collect the data while those that did collect it would use inconsistent standards, criteria and methods. Consequently, the data would be of questionable utility because there would be no assurance of its accuracy nor would there be any way to compare it from creditor to creditor” [404]. The U.S. Government Accountability Office found in 2008 that while such a mandatory data collection could provide benefits to researchers and regulators, it could be costly or difficult for the lenders themselves [436]; Bogen et al. [48] suggest that the failure to implement such measures has largely been due to pressure from banks, which Taylor [404] found were the overwhelming dissenting voice in responses to the FRB’s request for comments on their proposal.

**Fair lending in practice**

The two major discrimination doctrines which are relevant to fair lending law today are disparate treatment and disparate impact. Disparate treatment applies when individuals are explicitly treated differently on a prohibited basis. Under disparate impact doctrine, on the other hand, a creditor may be found to have illegally discriminated against a protected class if the effect of the practice adversely impacts that group even if the policy in question was facially neutral. The Supreme Court has found that the disparate impact is cognizable under the FHA [407], but has not made a similar ruling about ECOA. However, the court’s language in Inclusive Communities [405], relevant case law [33], and the CFPB’s official interpretation of Regulation B [83] all support the general consensus that disparate impact theory is cognizable under ECOA. Federal courts have consistently upheld this since the 1980s [66].

Plaintiffs rely on burden-shifting schemes for establishing a *prima facie* claim under both theories, which can then be rebutted by the defendant. For a disparate treatment case, most circuit courts have found that a modification of the McDonnell-Douglas test, originally developed by the Supreme Court in an employment discrimination case [278], can be applied to an ECOA claim—but there is no official nationwide rule on the issue [33]. In a prima facie disparate impact claim [2], a plaintiff must point to a specific policy or action taken by the defendant that had a disproportionately adverse impact on members of a protected class. The defendant may respond by arguing there is a legitimate business necessity for the policy. Then, the plaintiff can respond by arguing there was a less discriminatory alternative that the defendant refused to use.

In a disparate impact claim, expert statistical testimony is necessary to demonstrate that an adverse impact exists and is disproportionately felt by members of a protected class [33]. Again, we lack official Supreme Court guidance on how exactly to go about this under ECOA. In employment discrimination cases, however, the ratio of the proportion of the protected class that receives a
favorable outcome and the proportion of the control class is used; the oft-cited “80% rule” is related
to this statistic [131]. A related metric with precedence in the credit setting is the standardized
mean difference of outcomes for two groups [177]. However, it is usually insufficient to simply
compare the approval rates of two groups of applicants; since information related to creditworthiness
is generally available, higher courts generally expect that plaintiffs will compare the selection
rates of qualified applicants [33]. For this reason, statistical evidence which controls for drivers of
creditworthiness—such as a conditional marginal effects test—are seen as more appropriate by federal
agencies [86].

It is difficult for plaintiffs to find evidence that an individual lending decision was discriminatory,
especially in the non-mortgage setting where sensitive attribute data about applicants is generally
unavailable; some lower courts have historically acknowledged this [404]. Bogen et al. [48] point
out that “one of the few, robust public studies on credit scores and discrimination in the United
States was performed by the FRB in 2007, at the direction of Congress. To conduct its analysis, the
FRB created a database that, for the first time, combined sensitive attribute data collected by the
Social Security Administration with a large, nationally representative sample of individuals’ credit
records... this unusual undertaking would not have been possible without significant governmental
time and resources.” Interestingly, the CFPB has worked around this data limitation in some of its
enforcement actions by imputing racial information using Bayesian Improved Surname Geocoding
(BISG) to amass evidence of disparate impact [16]. (On the other hand, in the mortgage setting
where data is available, HMDA data alone cannot prove or disprove discrimination, and the results
of discrimination studies using HMDA data are usually contentious [318].)

An important precedent is, of course, the general acceptance of traditional credit scores as a
basis of loan underwriting. Like the machine learning algorithms which are the focus of this paper,
credit scores are functions of data which are meant to provide a quantitative basis on which to
make a lending decision. As of yet, there have not been successful challenges against credit scores
using disparate impact theory [210]. A combination of factors has contributed to this, but one
seems to be that official CFPB interpretations of ECOA and OCC guidance on models are fairly
generous as to what counts as a business necessity and relation to creditworthiness, respectively
[33]. Further complicating this matter is the fact that creditors tend to (credibly) argue that their
scoring methods tend to expand credit to minority applicants when compared to other methods.
The FRB bolstered the credit score’s ubiquity in their analysis of the 2007 database: they claimed
that while credit scores have a “differential effect” [304], they did not “produce a disparate impact”
[22] because credit characteristics do not act as “proxies for race and ethnicity” according to their
own definition (which we will discuss the limitations of later).
4.2 Fair lending and algorithms

In this section, we argue that federal agencies have indicated an interest in regulating discriminatory algorithms, and that certain ideas from the fair machine learning literature should have relevance for that regulation.

Agency communications on fair lending in algorithms

Here, we analyze recent messaging from several federal agencies on the threat of algorithmic fairness in finance and credit. These agencies are generally allowed to operate independently, but many have been known to act cooperatively and take a unified stance on the interpretation of the law [326]. The OCC, FRB, FDIC, and CFPB recently issued a rare joint request for information regarding the use of artificial intelligence (AI) in financial services, inquiring, among other things, whether banks and other interested parties feel that additional regulatory guidance on the matter is necessary [308]. Their response to the threat of algorithmic discrimination will be highly influential since, as Alex Engler has argued, “major legislative changes to AI oversight seem unlikely in the near future, which means that regulatory interventions will set precedent for the government’s approach to protecting citizens from AI harms” [128].

The Consumer Financial Protection Bureau (CFPB)  The CFPB was created by the Dodd-Frank Wall Street Reform and Consumer Protection Act in 2011. It was designed to consolidate responsibilities from several other agencies such as the Federal Reserve, FTC, and FDIC, to write and enforce rules for both bank and non-bank financial institutions. It has situated itself as being well-posed to tackle new regulatory challenges introduced by technology. The CFPB’s internal “Office of Competition and Innovation,” dedicated in part to addressing these challenges, has taken initiatives such as holding tech sprints, issuing no-action letters (NALs), and developing compliance assistance sandboxes.

The most relevant action the CFPB has taken with respect to algorithmic discrimination was its NAL to fintech lending company Upstart in 2017. Upstart provided detailed public (and some private) information about its underwriting process with the bureau and requested a formal statement from the CFPB that they would not trigger any enforcement action [413]. The CFPB granted the NAL. Part of the terms of the letter stipulated that Upstart would send the CFPB updates “regarding the loan applications it receives, how it decides which loans to approve, and how it will mitigate risk to consumers, as well as information on how its model expands access to credit for traditionally underserved populations” to “further its understanding of how these types of practices impact access to credit generally and for traditionally underserved populations, as well as the application of compliance management systems for these emerging practices” [84].
The CFPB has been criticized for this move because of Upstart’s usage of educational data in its algorithm. Several Democratic senators wrote a letter to the CFPB arguing that using this information could result in discrimination against minorities, and further argued that NALs should not be issued to provide immunity from ECOA in general [54]. A group of advocacy organizations expressed concern that the Upstart NAL was issued without “fully accounting for certain aspects of the company’s model that have long been recognized as having a disparate impact on borrowers of color” and pointed out that the CFPB did not attempt to replicate the company’s fair lending analysis [8].

Under the Biden administration, the CFPB is expected to enforce fair lending laws more aggressively than it did under Trump. Biden “has pinpointed the agency as a key weapon in his arsenal to address racial disparities in access to loans, capital and credit” [122]. To do this, the CFPB is likely to utilize disparate impact doctrine as it did under Obama, even in cases “where disparate racial impact was unintentional” [284]. This proactive regulatory behavior will be partly aimed directly at the algorithmic discrimination issue. Biden’s CFPB Director, Rohit Chopra, has repeatedly remarked that the agency will “closely watch for digital redlining, disguised through so-called neutral algorithms, that may reinforce the biases that have long existed” [76].

The Federal Trade Commission (FTC)  The FTC is tasked with protecting consumers in the United States, and thus shares the power to enforce ECOA with the CFPB. In particular, they are responsible for the regulation of non-bank financial service providers.

The FTC has positioned itself as particularly concerned with algorithmic discrimination. In 2016, it acknowledged the potential of alternative credit scores to help expand credit to populations previously deemed unscoreable, such as consumers without a credit history but that nonetheless pay their rent on time or own a car; it also pointed out that algorithmic credit decisions with a disparate impact on a protected class could violate ECOA, noting that it has taken enforcement action using this doctrine in a mortgage case [137]. A 2021 blog post by staff at the FTC’s Bureau of Consumer Protection posited that “apparently neutral technology can produce troubling outcomes – including discrimination by race or other legally protected classes” and indicated that the FTC Act’s prohibition of unfair or deceptive practices would include the sale or use of racially biased algorithm [219].

Most recently, Commissioner Rebecca Slaughter published a report on “Algorithms and Economic Justice” [380], stating that ECOA “can and should be aggressively applied” to threats of algorithmic discrimination. Notably, Slaughter expressed a personal opinion that “as with mortgage data, all other kinds of credit should be monitored by creditors consciously for disparities on the basis of protected status,” and advocated for the collection of protected class data to enable firms to self-test their algorithms for fairness under Regulation B’s existing exceptions.
The Office of the Comptroller of the Currency  The OCC, which was established by the National Currency Act of 1863, is meant to regulate and charter the nation’s banks. In particular, they make sure national banks and federal savings associations “operate in a safe and sound manner, provide fair access to financial services, treat customers fairly, and comply with applicable laws and regulations” [299]. While the OCC has traditionally only regulated traditional banks, the question of whether it should be in charge of regulating “fintech” companies has been fraught and remains legally unresolved [374].

The OCC notably issues guidance to banks on how to reduce risk in the development and use of mathematical models. In 2011, together with the FRB, the OCC issued Supervision and Regulation Letter 11-7 (SR 11-7), entitled Supervisory Guidance on Model Risk Management. The document describes “key aspects of an effective model risk management framework, including robust model development, implementation, and use; effective validation; and sound governance, policies, and controls” [302]. While this document does not explicitly mention illegal discrimination as a risk, in 2016, then-Comptroller Thomas J. Curry suggested that ECOA violations could be construed as such [101]:

New companies and companies deploying new technology should understand and ensure their products and services comply with existing laws, such as the Equal Credit Opportunity Act ... Lenders who operate without considering these questions may be accruing underappreciated financial risks and reputational liabilities.

Measuring unfairness in an algorithmic lending context

Recent work in machine learning has attempted to measure and mitigate discrimination in predictive models. In this section, we analyze how several technical measures of fairness align with the principles which are implied by ECOA and the regulatory bodies which concern themselves with enforcing it. We emphasize that none of these definitions can exactly capture whether a decision does, or does not, violate anti-discrimination law. However, many of the proposed metrics are consistent with a long history of testing for discrimination [211]. We therefore suggest they might be useful as evidence in litigation, internal auditing, or as metrics with which to optimize a fair model. Importantly, the theoretical properties of mathematical fairness metrics which we discuss here, such as how they relate to different data and modeling conditions, can be—and have been—formally studied. These studies, which we situate in the context of credit discrimination in the following section, provide intuition about which practices in algorithmic lending are likely to be problematic.

Throughout this section, we will refer to the framework of Friedler et al. [155], in which true qualities of individuals—the ideal basis for the decision-making process being learned—are referred to as construct features. In the credit setting, proposed construct features might include qualities such
as trustworthiness, reliability, and financial stability. The quality that the algorithm is trying to predict is known as the construct decision; in the credit setting, this is often described or referred to as creditworthiness.

In a perfect world, underwriters would have access to these construct features and use them to build a model to estimate creditworthiness. In reality, it is impossible to directly measure the qualities that define a strong credit applicant. Instead, algorithm designers must use approximations of the construct features, called observed features, as inputs. To use supervised learning methods, the algorithm designers must have access to a historical dataset of observed features and an associated historical measurement of the construct outcome for each row; this is called the observed decision or target variable. In the credit setting, the observed version of creditworthiness could be defined as (for instance) whether or not a historical applicant ended up defaulting on a loan they were issued within a certain amount of time; Barocas and Selbst [27] have pointed out that these modeling decisions are subjective choices that must be defended when making a business necessity defense. We will see that different assumptions about how these observational processes admit different intuitions for the appropriateness of fairness methods.

Fairness as blindness with proxy removal Perhaps the most lenient implementation of ECOA would be to say that any data-driven scoring system optimized to predict some business-necessity-related outcome should be presumed nondiscriminatory as long as its inputs are not protected attributes or “proxies” for them. An oft-cited example of what is meant by a “proxy” is the use of zip codes for racial discrimination in the illegal practice of redlining. The OCC has also pointed out that a person’s primary language being Spanish is a proxy for racial or ethnic groups [307], and differential treatment based on this feature has resulted in discrimination enforcement in the past by the CFPB [85].

The main problem with this logic is the difficulty of defining and identifying what a problematic proxy is, especially in light of advances in machine learning. An expansive definition of a proxy might include any feature statistically related in some way to a protected class, but the history of ECOA and its enforcement (in particular the generous guidelines around business necessity) generally support the usage of features with this quality.

More narrow definitions of “proxy” often involve the relationship between a variable’s protected-class-relatedness and its predictiveness for the task at hand. “Proxy discrimination” is defined by Prince and Schwarcz [331] as a special case of disparate impact, when a variable is predictive of an outcome because it is correlated with a sensitive attribute. An example that instantiates this line of reasoning appears in the FRB’s 2007 study, which posits that “a credit characteristic that derives its predictiveness solely by functioning as a proxy for demographics would not predict performance in a model that was estimated in a demographically neutral environment, where demographics
are controlled for or where the estimation sample is limited to a single demographic group” [22] and goes on to argue that if credit characteristics are still predictive in a demographically neutral environment they do not cause disparate impact. Similarly, Bartlett et al. [32] posit that “scoring or pricing on a proxy variable that has significant residual correlation with race or ethnicity after orthogonalizing with respect to hidden fundamental credit-risk variables is illegitimate.”

These definitions unfortunately induce thorny questions about how to quantify the slippery concepts of a variable’s predictiveness and protected-class-relatedness, which are again made less well-defined by advances in machine learning. The FRB study focused on linear models, where a variable’s coefficient can act as a notion of importance. However, in the case of complex and nonlinear ML models, the question of how to measure predictiveness of an individual variable is the center of a long-standing debate [430]. For instance, some have argued that a variable is important if the learned model’s output is sensitive to the input in some measurable way [51]; some argue a variable’s influence should be measured in terms of how much it improves a model within a class [148]; some argue the variable’s importance with respect to every subset of the other variables is important [270]. This question cannot be answered without encoding implicit epistemic values [179], and remains not well defined within the community at large, much less in the context of a business necessity defense.

As for protected-class-relatedness, the focus on identifying the “proxy-ness” of a single variable ignores the fact that several variables taken together, especially if a complex nonlinear model is used, can be very related to protected class information even when the individual variables are not. Gillis [167] found that by training a race prediction model on HMDA variables, they are collectively “more predictive of race than zip code.” For these reasons, Gillis concludes that ultimately, rather than focusing on the elimination of individual proxies from credit risk models, regulatory agencies should measure the fairness of a machine learning model in terms of outcomes. The rest of the definitions mentioned in this section all at least partly judge a model by its predictions.

**Equality of outcomes**

As we discussed in Section 1, comparing rates of positive and negative outcomes across groups is often used as evidence of disparate impact in employment discrimination. In ML, equal outcome rates across groups is often called demographic parity. Returning to the framework of Friedler et al. [155], if we believe that each subpopulation is similar in the construct feature space, we should assert that any differences in outcomes under an algorithm are discriminatory. However, as we also discussed in Section 1, credit scores are rarely challenged in courts even though they differ across groups due to a desire to compare outcome rates among qualified applicants.

This intuition—that a disparate impact analysis may look different when conducted on qualified
applicants versus overall applicants—directly corresponds to the worldview that relevant construct features may differ across groups. The persistence of this assumption in the credit setting raises the question of which differently distributed construct features courts and regulators consider to nevertheless be legitimate bases for decision-making, and which they do not. While perhaps no unbiased individual would assert that qualities such as trustworthiness differ in protected groups, they may feel that there could be cultural differences, or differences caused by structural discrimination, or differences induced by the self-selection of applicants, that are nonetheless valid bases for loan approval.

Fairness metrics based on raw outcome rates, then, should be of more concern in a discrimination case if a model relies on data that are meant to be predictive for different reasons than traditional data are. In other words, if the predictive utility of some data is explained by its association with constructs that should not differ across groups (such as trustworthiness), differences in raw outcomes should be considered less justifiable from a business necessity perspective; we discuss this situation further in the next section.

**Comparing predicted and actual outcomes** A large body of work in fair ML focuses on equalizing some statistic relating actual and predicted outcomes across different demographic subgroups. We attempt to shed insight on how each might be relevant to a fair lending case. Some of these qualities cannot be simultaneously satisfied by any decision procedure except under specific circumstances [240], and judging the relative importance of each metric forces us to surface our worldview assumptions and moral intuitions about when disparate outcomes are wrong.

Metrics related to sufficiency measure the extent to which the classifier’s score is equally predictive for different groups: If a model satisfies “sufficiency,” given the score, outcome is independent from the sensitive attribute. This implies further information about the sensitive attribute will not improve the model’s accuracy—the score is sufficient. Barocas et al. [28] point out that “sufficiency often comes for free (at least approximately) as a consequence of standard machine learning practices”, as all available predictive information should be exploited by an optimal model. The oft-cited metric of group calibration measures a model’s closeness to a stronger condition which implies sufficiency.

Sufficiency aligns with a specific kind of moral logic. Legal scholar Deborah Hellman has argued that enforcing sufficiency would uphold the intuitive notion that “everyone is entitled to be treated by the most accurate test available (or feasible, or imaginable)” [189]: if there were information in the data that could have helped predict an outcome, it should have been utilized. However, this interpretation of “fairness” does not align well with the disparate impact doctrine, which is triggered by the distribution of outcomes. Further, emphasizing a classifier’s decision-making skills does not take into account the differing relative badness of false positives (qualified applicants denied credit) and false negatives (unqualified applicants receiving credit). In the credit setting, anti-discrimination
law is much more concerned with the former.

On the other hand, separation, otherwise known as equalized odds, allows correlation between the score and the sensitive attribute to the extent that it is justified by the target variable. It requires that score distributions be equal between protected and unprotected individuals within the groups of qualified and unqualified individuals. Kozodoi et al. [243] argue that separation is a good measure of fairness for credit because it “accounts for the imbalanced misclassification costs of the customer, and, as these imbalanced costs also exist for the financial institution, separation is also able to consider the interests of the loan market.” The relaxation of separation that qualified individuals from each group receive credit at the same rate is called equality of opportunity [182]. This closely aligns with ECOA precedent suggesting only the acceptance rates of “qualified applicants” should be compared. Hellman [189] argues that the ratio of false positives and false negatives is a normatively meaningful statistic that should be equalized across groups, corresponding to a different relaxation of separation. In the credit setting, while we generally think of access to credit as a uniformly positive thing, if an applicant gets a loan they cannot pay back it is ultimately bad for them, perhaps indicating that we should balance the risk of false positives with false negatives.

Famously, sufficiency and separation cannot generally be achieved simultaneously [77]. This relates to the generally accepted fact that rates of qualified and unqualified applicants (as defined by the data) may differ across groups. Essentially, if the input data contains associations with group membership, a sufficient classifier will learn that the groups should have differing score distributions in order to be as predictive as possible—thus violating separation.

Further, while we can certainly relate these statistics to the principles of nondiscrimination, they are only meaningful to the extent that the observed data are meaningful. If the observed decisions were generated in a historically discriminatory or otherwise problematic manner, fairness with respect to those decisions does not imply fairness with respect to the “true” or desired variable of creditworthiness. This is why Wachter et al. [426] call metrics in this group “bias-preserving.” We discuss when this is likely to happen in the credit setting in the next section.

**Causal and counterfactual reasoning** There have been many attempts to measure illegal discrimination using causal reasoning. The term “causal inference” refers to a broad spectrum of methods and perspectives [289, 324], but in essence, the goal of applying it to discrimination attempts to answer the question, “Does a protected attribute cause a particular decision outcome?” To answer this question with causal logic, one might turn to comparing an actual outcome to a certain “what-if” scenario called a counterfactual. For instance, to determine whether a system discriminated against a black individual who was denied a loan, one might try to estimate what would have happened if the individual were white. To do this analysis, practitioners assert or discover a model of the different cause and effect relationships between relevant variables and use
them to make inferences about the counterfactual scenario.

Using causal models, one can additionally attempt to distinguish between “direct” and “indirect” effects of a sensitive attribute. Some methods for learning fair models involve measuring “effects of [sensitive attributes] that are mediated by other attributes, keeping only those effects carried along paths deemed fair” [205]. This work posits that features like gender or race may be causally related to information that one might assert is nonetheless a valid basis for decision-making, such as GPA or department choice in the setting of graduate admissions.

Causal reasoning methods seem to closely match language used to describe discrimination in the abstract, as proponents of causal inference often point out [28]. Causal logic is also often used by humans in practice for moral reasoning about decision-making in general [238]. However, applying this line of thinking to discrimination in algorithms suffers from both conceptual and practical limitations. On a practical level, specifying causal models requires making assumptions that cannot be validated by observational criteria and introduce complicated questions about how to understand relationships between human categories [28]. A full treatment of the conceptual critique is beyond the scope of this paper, but we very roughly summarize a line of work from Kohler-Haussman and Hu here: because discrimination is a “thick ethical concept,” which both describes and evaluates actions, it cannot be defined in terms of a causal model [242]. Further, traits modeled as direct or indirect “effects” of social categories are are often in fact constitutive features of those categories and relate to what makes discrimination distinctly morally problematic [207]. For these reasons, fair ML research based on measuring or improving causal and counterfactual metrics of fairness are unlikely to easily translate to enforcement or compliance with anti-discrimination law in lending.

4.3 Discrimination risks in models trained on loan repayment data

Machine learning algorithms complicate the interpretation of fair lending law by blurring the line of what it means for a policy to be facially neutral: even if an algorithm does not have access to protected class information, it may have been intentionally or unintentionally trained in a way that makes the ultimate policy not-so-neutral. In this section, we use results from fair ML literature – largely quantified in terms of the fairness metrics discussed previously – to determine where and how bias is likely to occur in a credit modeling setting: Firstly, if a model is trained primarily on data pertaining to a certain demographic group, that model may perform disproportionately well on that demographic group compared to others; this manifests through the problem of credit invisibility. Secondly, if the observed features used to train the model introduce group skew from the “true” process being modeled, a model may pick up on or exacerbate these effects; these concerns are raised
in a novel way by models trained on alternative data. Thirdly, the extent to which the first two issues introduce disparities through learning are affected by other modeling choices.

**Sampling processes and credit invisibility**

Algorithms developed with ML techniques improve when exposed to more and more historical data. Intuitively, if training data is less available for some subpopulation of individuals, a model trained on the whole population may have performance disparities when evaluated on the groups individually. Further, since many performance metrics which are used to optimize models are constructed as averages of some kind of error-based cost across the population, these metrics are primarily driven by the model’s performance on the majority class.

Both theoretical and empirical work provide evidence for the general principle that a group’s under-representation in a data set can lead to group fairness disparities. Chen et al. [72] show that a learning procedure’s expected performance disparities over a distribution can be additively decomposed into bias, variance, and noise components, and note that disparities caused by a difference in variances can be caused by differences in sample sizes across the groups. Buolamwini and Gebru [58]’s seminal work on performance disparities in gender classification models across skin tones found that popular facial analysis benchmarking datasets are overwhelmingly white and male, and discovered that many commercial facial analysis software systems were disproportionately wrong on darker females.

This source of unfairness in machine learning is relevant in the credit setting because of the effects of credit invisibility. In America, millions of people are “unscorable” because of their lack of credit history, and therefore face barriers to accessing credit. The CFPB recently found that income is strongly correlated with having a scored credit record, and that “Blacks and Hispanics are more likely than Whites or Asians to be credit invisible or to have unscored credit records” [305]. This matters because individuals who have never accessed credit are inherently missing from credit-report-based datasets that could be used for the supervised learning of creditworthiness. Even if they are “scorable,” individuals who have historically applied for and were rejected from loans are also by definition missing from training datasets based on the outcomes of those particular loans. Recall that the data used to train a ML model must contain both observed features and observed decisions such as whether a historical applicant ended up defaulting on a loan. If it is not known whether an individual would have defaulted on a loan, they cannot be directly included in the supervised learning problem; therefore, we should expect low performance on those subgroups in models trained directly on historically issued loans, making them less fair from a statistical group fairness perspective.

If, however, we are concerned about the equality of outcomes of models trained on historically
issued loans, the “fairness” narrative of training on selectively labelled data may be different. Rambachan and Roth [342] suggest that under certain conditions, if a prior selection policy was biased against a certain group, a machine learning model trained on approved applicants disproportionately favor that group.

Of course, credit modelers know that they are missing information about applicants who were denied loans. Proceeding to only analyze the accepted applicants is called the “known good-bad” approach, but creditors usually attempt to incorporate information about the rejected applicants into their model; according to the FDIC, in the bank-issued credit card sector, “certain inferences are made to break down the rejected applicants into good and bad accounts. This procedure, known as reject inferencing, makes certain assumptions on how rejected applicants would have performed had they been accepted and attempts to mitigate any accept-only bias of the sample” [306]. Recent work has suggested that different methods for reject inferencing may have different fairness properties [89].

There is no simple solution to correcting for the known problems induced by sampling biases. Critically, measuring (and optimizing for) the group fairness of models on the limited labelled data available using the “known good-bad” approach will produce misleading or harmful results [88, 224]. Further research studying the conditions when this hidden fairness problem arises will provide intuition with which to interpret the methodologies employed by particular lenders.

**Observational bias and implications for alternative data**

The nature of the specific data attributes, or features, used as inputs to a machine learning model also have an effect on the fairness of that algorithm. Recall from the framework proposed by Friedler et al. [155] that an algorithm being learned by the supervised learning process is a function from the observed feature space to the observed decision space. The observational processes which imperfectly capture the construct features and decision can thus add group skew to the “true” relationship between construct features and construct decisions in the resulting algorithm, even when protected class attributes are not directly accessible to the algorithm through the observed features.

One way this can happen is when the observed decisions were generated by an explicitly discriminatory process, thus skewing the mapping from the construct decision space to the observed decision space. An example of this would be using performance reviews made by an individual with a personal bias against women as the target variable of a hiring model. If the goal of the algorithm is to make nondiscriminatory decisions, this is a poor choice of observed decision. Additionally, as we stated in the previous section, if the target variable itself is biased, fairness metrics which rely on “true” labels in the training data will be misleading.
The “observed decision” of loan repayment has generally been treated as an appropriate measurement of creditworthiness for the purposes of a business necessity defense [33, 27]. However, if factors involved in certain instances of loan repayment are unlikely to generalize to future conditions, this may present a challenge for that argument. For instance, suppose pandemic-induced conditions disproportionately caused a certain protected group of people to default at a higher rate than others. Since the pandemic conditions may not repeat themselves in the future, the measured default variable during this time may not be relevant from a business necessity perspective, and the statistical fairness criteria cannot bolster the business necessity argument.

Another way group differences can manifest in an algorithm is if groups who are similar in the construct feature space appear different in the observed feature space because the corresponding observational process is affected by cultural differences or discrimination. Xiang [446] provides an example in the context of auto insurance, in which insurers would like to measure the construct feature quality of risk aversion. In this scenario, we suppose risky non-Asian drivers would choose to drive red cars more often than low-risk drivers because they are perceived as flashy or ostentatious. But it is possible that Asian drivers who drive red cars do so because red is considered a lucky color, and are no riskier than non-red-car-driving Asians. Kiviat [239] calls such data, which “improperly confliate[s] morally distinct situations and behaviors,” morally heterogeneous–and finds that Americans often think using this kind of data in decision-making can be “unfair.”

To determine whether observational processes are inducing bias in a credit modeling dataset, we need to take a stance on whether or not group differences can preexist in the construct feature space. Credit scores, for example, have repeatedly shown to be differently distributed across groups [65], yet this data is seen by regulators as being related enough to creditworthiness to suffice as a basis for underwriting despite the resulting disparate outcomes. In other words, it is implicitly being touted as a valid, low-bias measurement of a relevant construct feature, such as financial stability. Friedler et al. [155] call this assumption “What you See is What You Get”–the idea that any group disparities seen in observed data are due to group disparities in construct features and are therefore an appropriate basis for decision-making. Kiviat [239] has shown that data which have a “logical relatedness” to a task at hand is generally seen as fair to use for that task. Again, if evidence arises that an observational process in the training data is erroneous or generated by discriminatory processes, the claim that the data is related to creditworthiness is weaker.

For this reason it is important to vet novel, “alternative” data sources for measurement validity with respect to construct features and potential for group skew. "Alternative data” refers to information that lenders may use for credit decisions but that is “not typically found in the consumer’s credit files,” including data regarding recurring payments for utilities and rent, or cash flow data regarding deposit accounts [8]. This strategy is gaining significant traction; Jagtiani and Lemieux [216] has uncovered evidence that online lenders are increasingly using non-traditional
data to underwrite their loans. Turning to alternative data sources is meant to address the “thin file” problem of unscored and underscored credit applicants, and in some cases this may be an appropriate solution to that problem. FinRegLab found that cash flow data provided “independent predictive value across all [demographic] groups” for credit risk and loan performance [145], thus appealing to the concept of sufficiency.

However, other variables have been controversial, such as educational data in the case of Upstart. Hurley and Adebayo [210] have reported that all kinds of data—social media profiles, technology usage, and “how quickly a user scrolls through terms of service”—have been used for underwriting purposes by fintech companies. In general, data should receive heightened scrutiny if, as some of these features seem to be, they are attempts to measure construct features that should not vary across groups, such as personality traits or intelligence.

**Model complexity**

The degree to which the sampling and modeling problems described above actually affect a machine learning algorithm depends on the chosen model class and training procedure. In particular, it relates to a model type’s capacity, which measures how well it can capture complex patterns in data. For example, new, “powerful” ML tools like gradient boosting and deep learning are high-capacity, whereas traditional linear models are low-capacity. These differences are salient in a fairness sense: Low-capacity models on data which is disparately predictive between classes may result in low cost-based fairness. On the other hand, high-capacity models on predictive data can be have more unequal outcomes than simple models if there is bias in the labels.

Low capacity models on disparately predictive data can discriminate in a “statistical group fairness” sense. Chen et al. [72]’s decomposition of statistical group fairness measures shows that differences in a model’s bias can cause group fairness to deteriorate, which happens when “the chosen model class is not flexible enough to fit both groups well.” For this reason, a sufficiently complex model trained on culturally diverse data could be “less” biased than a simple one (for instance, able to capture the different meanings of Red in the car example). In theory, this could present a problem for banks, who traditionally use simpler models such as logistic regression [22, 27] if they apply these models to alternative, “morally heterogeneous” data as in the car example. Interestingly, however, studies have show that advanced modeling techniques to more traditional data does not necessarily improve outcomes [43], so low model capacity may not be acting as a source of bias in this context.

On the other hand, high-capacity models on very biased but predictive historical data can amplify discrimination more than a low-capacity model can. For instance, a sufficiently complex model trained on biased hiring data could be more biased than a simple one, by being able to more
precisely pick up on gender through resume items using combinations of words instead of single words. This presents a risk in the case of new, more cutting-edge “fintech” companies which are more likely to be experimenting with high-capacity models such as gradient boosting and random forests. Feldman et al. [141] has shown that outcome rate disparities of a model are mathematically connected to how predictive the input data are of a protected attribute, and there is also evidence applying data mining to HMDA \textit{is more predictive of race than zip code} [167]. For these reasons, the accidental encoding of racial information in a high-capacity model trained on biased data is a real danger in the credit setting.

4.4 Regulatory opportunities

This paper has demonstrated that ML fairness research suggests that machine learning and alternative data present fair lending risks that should be of concern to regulators. Now we tackle the implications for the enforcement of fair lending regulation. In particular, we present two broad strategies that regulators could pursue to identify and mitigate the fairness risks identified in Section 2.

Expanding the collection and analysis of protected class attributes

Protected class information on loan applicants is necessary to effectively measure and mitigate unfairness, which as Section 2 has argued, is a real threat—yet it is still legally risky to collect. When the CFPB was first established, Taylor [404] suggested they were well-positioned to remove the general ban preventing creditors to collect protected class information; in theory, they can directly amend or change Regulation B. In practice this would be politically difficult, for reasons discussed in Section 1.1.2. However, the CFPB has made steps to increase the amount of data available in this space by changing the requirements in the context of small business loans. Further action incentivizing or requiring the collection of protected class data would enable interventions for detecting and preventing discrimination as well as expanding access to credit.

The first benefit of this data is unrelated to machine learning specifically: protected class information on applicants for loans would enable external oversight of lenders as HMDA data does in the mortgage setting. As Regulation C states, a major purpose of the HMDA data requirement is to “assist in identifying potential discriminatory lending patterns and enforcing antidiscrimination statutes;” as mentioned previously, many have argued that a similar provision would be helpful for the same tasks [436, 404]. The CFPB would no longer have to rely on BISG or related methods to impute sensitive data for their audits.

The collection of protected class attributes in credit data also expands the range of tools for developing fair models available to algorithm developers. Many of the quantitative fairness
frameworks discussed in Section 2 lend themselves to an optimization problem: training or modifying a model to be fair(er). Some of these tools offer interventions to modify the training dataset, the way the model is trained, or tweaking the model after it has been learned in the usual way [154]. Access to protected class information in the training dataset is generally required to implement most of these methods, although workarounds have been proposed [73].

The act of directly forcing an algorithm to conform to fairness metrics raises potential legal issues. Methods that require access to a sensitive attribute of an applicant at decision time are unlikely to gain traction in the credit setting, since ECOA specifically prohibits using the protected class of an individual in a credit decision. This includes options such as training different models for different protected classes, which may be legal in other contexts [189].

What remains less clear is whether interventions which have access to historical protected class information at training time could be legal or even encouraged. Some scholars have expressed concern that preemptively modifying an algorithm for fairness could be considered disparate treatment [27], or analogous to racial quotas [446]. However, Kim [234]’s detailed treatment of the issue describes many situations in which race-conscious decision-making is not considered disparate treatment under anti-discrimination law. She concludes that techniques “more accurately understood as removing bias from processes,” such as efforts to correct biased input data or formulate a fair problem specification, are legally permissible, whereas methods that more closely resemble a quota system will likely trigger close legal scrutiny. The qualitative differences between specific bias mitigation methods are therefore important to describe and evaluate, but there appears to be a legal path forward for regulators to encourage some of these strategies. Guidance from agencies on this issue is currently sorely lacking, and must be addressed.

There are other considerations at hand here, as modifying models to fit fairness criteria in practice can introduce other, non-legal problems. Friedler et al. [154] benchmarked several against a variety of fairness metrics on existing datasets, and found that they tended to be brittle and sensitive to fluctuations in dataset composition, highlighting the importance of careful experimental design when drawing conclusions about fairness. Studies employing economic methods have also shown that the long-term effects of enforcing fairness have implications for social welfare [263, 206]; this relates to a broader discussion of balancing nondiscrimination with economic efficiency that is outside the scope of this paper.

Even if regulators or lenders are uncomfortable with making their traditional credit risk models fairness-aware, ECOA specifically allows special purpose credit programs to be targeted at expanding access to credit to traditionally underserved populations. Unfortunately, few lenders have taken advantage of this allowance. Protected class data could provide insight into how to effectively underwrite credit to those populations.
Managing discrimination risk as model risk

Several organizations [16] have suggested that one way regulators can use their authority to mitigate discrimination risk is by treating it as any other kind of model risk and applying the relevant guidelines and standards to the development of models in the financial space. For instance, to apply SR 11-7 to discrimination risk, some suggest “the Agencies should ensure that financial institutions have appropriate Compliance Management Systems that effectively identify and control risks related to AI systems, including the risk of discriminatory or inequitable outcomes for consumers” [8].

Scholarly analyses of discrimination support the idea that developers should be held liable for unintentional algorithmic discrimination. In his theoretical treatment of algorithmic discrimination, legal scholar Tal Zarsky considers the negligent or reckless usage of biased data to be a form of intentional implicit discrimination, in which “firms’ failure to act and prevent discrimination” is a form of intent. For this reason, he argues that “such behavior should be actively countered” and that “additional policy discussions must establish the proper standard of care this normative justification calls for on behalf of the scorers” [449].

SR 11-7 emphasizes that model risk guidance emphasizes that “risk assessment should be conducted by independent actors within the institution or a third party.” If financial institutions are pressured to do more self-testing using gathered or approximated protected class data, and follow this guidance, the team who developed a model may be informed that their models have undesirable fairness metrics. They may be able to use this information to develop a less discriminatory alternative without directly using protected class data. For instance, simply changing a model’s overall acceptance threshold can influence the fairness statistics of a model [8]. This “indirect” optimization may be more legally defensible than the direct optimization strategies discussed previously.

The OCC’s model risk guidance also recommends continuous monitoring of models in deployment. Monitoring models for correctness is important, but developers can also monitor their models for changes in the incoming data that could affect the fairness dynamics described above. For instance, macroeconomic changes may affect changes in the underlying demographic composition of applicants, which will in turn affect observable fairness characteristics.

Specific federal guidance on how to responsibly manage bias risk could be developed based on the fair ML results discussed in this paper. For instance, the agencies could recommend that bias risk be considered and estimated when a developer chooses how to conduct reject inferencing, as discussed in Section 2. NIST has already taken steps towards developing a framework to mitigate bias risk in general; the Agencies could build off of their work or develop a credit-specific framework in parallel.
Chapter 5

Challenging the fallacy of AI functionality in public policy

As one of over 20,000 cases falsely flagged for unemployment benefit fraud by Michigan’s MIDAS algorithm [68], Brian Russell had to file for bankruptcy, undermining his ability to provide for his two young children. The state finally cleared him of the false charges two years later [125]. RealPage, one of several automated tenant screening tools producing “cheap and fast—but not necessarily accurate—reports for an estimated nine out of 10 landlords across the country”, flagged Davone Jackson with a false arrest record, pushing him out of low income housing and into a small motel room with his 9-year-old daughter for nearly a year [237, 236]. Josiah Elleston-Burrell had his post-secondary admissions potentially revoked [254, 235], Robert Williams was wrongfully arrested for a false facial recognition match [194], Tammy Dobbs lost critical access to healthcare benefits [257]. The repercussions of AI-related functionality failures in high stakes scenarios cannot be overstated, and the impact reverberates in real lives for weeks, months and even years.

Despite the current public fervor over the great potential of AI, many deployed algorithmic products do not work. AI-enabled moderation tools regularly flag safe content [244, 317, 178], teacher assessment tools mark star instructors to be fired [350, 320], hospital bed assignment algorithms prioritize healthy over sick patients [298], and medical insurance service distribution and pricing systems gatekeep necessary care-taking resources [257, 350]. Deployed AI-enabled clinical support tools misallocate prescriptions [402], misread medical images [153, 297], and misdiagnose [394, 441]. The New York MTA’s pilot of facial recognition had a reported 100% error rate, yet the program moved forward anyway [39]. Some of these failures have already proven to disproportionately impact some more than others: moderation tool glitches target minoritized groups [110]; facial
recognition tools fail on darker skinned female faces [58]; a hospital resource allocation algorithm’s misjudgements will mostly impact Black and lower income patients [298]. However, all failures in sum reveal a broader pattern of a market saturated with dysfunctional, deployed AI products.

Importantly, the hype is not limited to AI’s boosters in corporations and the technology press; scholars and policymakers often assume functionality while discussing the dangers of algorithmic systems as well. In fact, many of the current critiques, policy positions and interventions in algorithmic accountability implicitly begin from the premise that such deployed algorithmic systems work, echoing narratives of super-human ability [146], broad applicability [339], and consistency [328], espoused in corporate marketing materials, academic research papers and in mainstream media. These proposals thus often fall short of acknowledging the functionality issues in AI deployments and the role of the lack of functional safety in contributing to the harm perpetuated by these systems.

Although notions of accuracy and product expectations are stakeholder-dependent and can be contested, the assessment of such claims are often easier to empirically measure, grounding the discussion of harm in a way that is challenging to repudiate.

As an overlooked aspect of AI policy, functionality is often presented as a consideration secondary to other ethical challenges. In this chapter, adapted from Raji, Kumar, Horowitz, and Selbst [341], we argue that it is a primary concern that often precedes such problems. We start by calling out what we perceive to be a functionality assumption, prevalent in much of the discourse on AI risks. We then argue that this assumption does not hold in a large set of cases. Drawing on the AI, Algorithmic and Automation Incident and Controversy Repository (AAAIRC), we offer a taxonomy of the ways in which such failures can take form and the harms they cause, which differ from the more commonly cited critiques of AI. We then discuss the existing accountability tools to address functionality issues, that are often overlooked in AI policy literature and in practice, due in large part to this assumption of functionality.

5.1 **Background: Related work**

A review of past work demonstrates that although there is some acknowledgement that AI has a functionality problem, little has been done to systematically discuss the range of problems specifically associated with functionality.

Recent work details that the AI research field suffers from scientific validity and evaluation problems [176, 120]. Kapoor and Narayanan [228] have demonstrated reproducibility failures in published work on predicting civil wars. Liao et al. [259] found that advances in machine learning often “evaporate under closer scrutiny or turn out to be less widely applicable than originally hoped.”

There is also some work demonstrating that AI products are challenging to engineer correctly
in practice. In a survey of practitioners, Wan et al. [427] describe how developers often modify traditional software engineering practices due to unique challenges presented by ML, such as the increased effort required for testing and defining requirements. They also found that ML practitioners "tend to communicate less frequently with clients" and struggle to make accurate plans for the tasks required in the development process. Sculley et al. [362] have additionally argued that ML systems "have a special capacity for incurring technical debt."

Other papers discuss how the AI label lends itself to inflated claims of functionality that the systems cannot meet. Kaltheneuer et al. [225] and Broussard [53] critique hyped narratives pushed in the AI industry, joined by many similar domain-specific critiques [406, 36, 391, 384, 335, 35]. Narayanan [295] recently popularized the metaphor of "snake oil" as a description of such AI products, raising concerns about the hyperbolic claims now common on the market today. Richardson [348] has noted that despite the "intelligent" label, many deployed AI systems used by public agencies involve simple models defined by manually crafted heuristics. Similarly, Raji et al. [339] argue that AI makes claims to generality while modeling behaviour that is determined by highly constrained and context-specific data. In a study of actual AI policy discussions, Kraft et al. [245] found that policymakers often define AI with respect to how human-like a system is, and concluded that this could lead to deprivatizing issues more grounded in reality.

Finally, Vinsel [420] has argued that even critics of technology often hype the very technologies that they critique, as a way of inflating the perception of their dangers. He refers to this phenomenon as "criti-hype"—criticism which both needs and feeds on hype. As an example, he points to disinformation researchers, who embrace corporate talking points of a recommendation model that can meaningfully influence consumer behavior to the point of controlling their purchases or voting activity—when in actuality, these algorithms have little ability to do either [212, 40, 354, 166, 192]. Even the infamous Cambridge Analytica product was revealed to be "barely better than chance at applying the right [personality] scores to individuals", and the company accused explicitly of "selling snake oil" [192].

### 5.2 The functionality assumption

It is unsurprising that promoters of AI do not tend to question its functionality. More surprising is the prevalence of criti-hype in the scholarship and political narratives around automation and machine learning—even amidst discussion of valid concerns such as trustworthiness, democratization, fairness, interpretability, and safety. These fears, though legitimate, are often premature "wishful worries"—fears that can only be realized once the technology works, or works "too well", rather than being grounded in a reality where these systems do not always function as expected [420].
In this section, we discuss how criti-hype in AI manifests as an unspoken assumption of functionality. The functionality of AI systems is rarely explicitly mentioned in AI principle statements, policy proposals, and AI ethics guidelines. In a recent review of the landscape of AI ethics guidelines, Jobin et al. [220] found that few acknowledge the possibility of AI not working as advertised. In guidelines about preventing malfeasance, the primary concern is malicious use of supposedly functional AI products by nefarious actors. Guidelines around “trust” are geared towards eliciting trust in AI systems from users or the public, implying that trusting these AI products would be to the benefit of these stakeholders and allow AI to “fulfill its world changing potential” [220]. Just one guideline of the hundreds reviewed in the survey “explicitly suggests that, instead of demanding understandability, it should be ensured that AI fulfills public expectations” [220]. Similarly, the U.S. National Institute of Standards and Technology (NIST) seeks to define “trustworthiness” based primarily on how much people are willing to use the AI systems they are interacting with [390]. This framing puts the onus on people to trust in systems, and not on institutions to make their systems reliably operational, in order to earn that trust [7, 56]. NIST’s concept of trust is also limited, citing the “dependability” section of ISO/IEEE/IEC standards [213], but leaving out other critical concepts in these dependability engineering standards that represent basic functionality requirements, including assurance, claim veracity, integrity level, systematic failure, or dangerous condition. Similarly, the international trade group, the Organisation for Economic Co-operation and Development (OECD), mentions “robustness” and “trustworthy AI” in their AI principles but makes no explicit mention of expectations around basic functionality or performance assessment [448].

The ideal of “democratizing” AI systems, and the resulting AI innovation policy, is another effort premised on the assumed functionality of AI. This is the argument that access to AI tooling and AI skills should be expanded [397, 23, 161, 185]—with the corollary claim that it is problematic that only certain institutions, nations, or individuals have access to the ability to build these systems [14]. A recent example of democratization efforts was the global push for the relaxation of oversight in data sharing in order to allow for more innovation in AI tool development in the wake of the COVID-19 pandemic [18, 313, 123, 432, 281]. The goal of such efforts was to empower a wider range of non-AI domain experts to participate in AI tool development. This policy impact was long lasting and informed later efforts such as the AI National Resource (AINR) effort in the US [196] and the National Medical Imaging Platform (NMIP) executed by National Health Services (NHS) in the UK [251]. In this flurry of expedited activity, some parallel concerns were also raised about how the new COVID-19 AI tools would adequately address cybersecurity, privacy, and anti-discrimination challenges [246, 112], but the functionality and utility of the systems remained untested for some time [188, 353, 445, 423].

An extremely premature set of concerns are those of an autonomous agent becoming so intelligent that humans lose control of the system. While it is not controversial to claim that such concerns are
far from being realized [100, 21, 332], this fear of misspecified objectives, runaway feedback loops, and AI alignment presumes the existence of an industry that can get AI systems to execute on any clearly declared objectives, and that the main challenge is to choose and design an appropriate goal. Needless to say, if one thinks the danger of AI is that it will work too well [368], it is a necessary precondition that it works at all.

The fear of hyper-competent AI systems also drives discussions on potential misuse [55]. For example, expressed concerns around large language models centers on hyped narratives of the models’ ability to generate hyper-realistic online content, which could theoretically be used by malicious actors to facilitate harmful misinformation campaigns [431, 388]. While these are credible threats, concerns around large language models tend to dismiss the practical limitations of what these models can achieve [36], neglecting to address more mundane hazards tied to the premature deployment of a system that does not work [132, 416]. This pattern is evident in the EU draft AI regulation [15], where, even as the legislation does concern functionality to a degree, the primary concerns—questions of “manipulative systems,” “social scoring,” and “emotional or biometric categorization”—“border on the fantastical” [418, p. 98].

A major policy focus in recent years has been addressing issues of bias and fairness in AI. Fairness research is often centered around attempting to balance some notion of accuracy with some notion of fairness [154, 147, 142]. This research question presumes that an unconstrained solution without fairness restrictions is the optimal solution to the problem. However, this intuition is only valid when certain conditions and assumptions are met [286, 156, 435], such as the measurement validity of the data and labels. Scholarship on fairness also sometimes presumes that unconstrained models will be optimal or at least useful. Barocas and Selbst [27, p. 707] argued that U.S. anti-discrimination law would have difficulty addressing algorithmic bias because the “nature of data mining” means that in many cases we can assume the decision is at least statistically valid. Similarly, as an early example of technical fairness solutions, Feldman et al. [143] created a method to remove disparate impact from a model while preserving rank, which only makes sense if the unconstrained system output is correct in the first place. Industry practitioners then carry this assumption into how they approach fairness in AI deployments. For example, audits of AI hiring tools focus primarily on ensuring an 80% selection rate for protected classes (the so-called 4/5ths rule) is satisfied, and rarely mention product validation processes, demonstrating an assumed validity of the prediction task [335, 437, 127].

Another dominant theme in AI policy developments is that of explainability or interpretability. The purpose of making models explainable or interpretable differs depending on who is seen as needing to understand them. From the engineering side, interpretability is usually desired for debugging purposes [42], so it is focused on functionality. But on the legal or ethical side, things look different. There has been much discussion about whether the GDPR includes a “right to
explanation” and what such a right entails [226, 364, 424, 124]. Those rights would serve different purposes. To the extent the purpose of explanation is to enable contestation [227], then functionality is likely included as an aspect of the system subject to challenge. To the extent explanation is desired to educate consumers about how to improve their chances in the future [29], such rights are only useful when the underlying model is functional. Similarly, to the extent regulators are looking into functionality, explanations aimed at regulators can assist oversight, but typically explanations are desired to check the basis for decisions, while assuming the systems work as intended.

Not all recent policy developments hold the functionality assumption strongly. The Food and Drug Administration (FDA) guidelines for AI systems integrated into software as a medical device (SaMD) has a strong emphasis on functional performance, clearly not taking product performance as a given [150]. The draft AI Act in the EU includes requirements for pre-marketing controls to establish products’ safety and performance, as well as quality management for high risk systems [418]. These mentions suggest that functionality is not always ignored outright. Sometimes, it is considered in policy, but in many cases, that consideration lacks the emphasis of the other concerns presented.

5.3 Taxonomy of failures

Functionality can be difficult to define precisely. The dictionary definition of “fitness for a product’s intended use” [301] is useful, but incomplete, as some intended uses are impossible. Functionality could also be seen as a statement that a product lives up to the vendor’s performance claims, but this, too, is incomplete; specifications chosen by the vendor could be insufficient to solve the problem at hand. Another possible definition is ”meeting stakeholder expectations” more generally, but this is too broad as it sweeps in wider AI ethics concerns with those of performance or operation.

Lacking a perfectly precise definition of functionality, in this section we invert the question by creating a taxonomy that brings together disparate notions of product failure. Our taxonomy serves several other purposes, as well. Firstly, the sheer number of points of failure we were able to identify illustrates the scope of the problem. Secondly, we offer language in which to ground future discussions of functionality in research and policy. Finally, we hope that future proposals for interventions can use this framework to concretely illustrate the way any proposed interventions might work to prevent different kinds of failure.

Methodology To challenge the functionality assumption and demonstrate the various ways in which AI doesn’t work, we developed a taxonomy of known AI failures through the systematic review of case studies. To do this, we partly relied on the AI, Algorithmic and Automation Incident and Controversy Repository (AIAAIC) spreadsheet crowdsourced from journalism professionals [69]. Out of a database of over 800 cases, we filtered the cases down to a spreadsheet of 283 cases from 2012 to
2021 based on whether the technology involved claimed to be AI, ML or data-driven, and whether the harm reported was due to a failure of the technology. In particular, we focused on describing the ways in which the artifact itself was connected to the failure, as opposed to infrastructural or environmental "meta" failures which caused harm through the artifact. We split up the rows in the resulting set and used an iterative tagging procedure to come up with categories that associate each example with a different element or cause of failure. We updated, merged, and grouped our tags in meetings between tagging sessions, resulting in the following taxonomy. We then chose known case studies from the media and academic literature to illustrate and best characterize these failure modes.

Here, we present a taxonomy of AI system failures and provide examples of known instances of harm. Many of these cases are direct refutations of the specific instances of the functionality assumptions in the previous section.

Table 5.1: Failure Taxonomy

<table>
<thead>
<tr>
<th>Impossible Tasks</th>
<th>Conceptually Impossible</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engineering Failures</td>
<td>Design Failures</td>
</tr>
<tr>
<td></td>
<td>Implementation Failures</td>
</tr>
<tr>
<td></td>
<td>Missing Safety Features</td>
</tr>
<tr>
<td>Post-Deployment Failures</td>
<td>Robustness Issues</td>
</tr>
<tr>
<td></td>
<td>Failure under Adversarial Attacks</td>
</tr>
<tr>
<td></td>
<td>Unanticipated Interactions</td>
</tr>
<tr>
<td>Communication Failures</td>
<td>Falsified or Overstated Capabilities</td>
</tr>
<tr>
<td></td>
<td>Misrepresented Capabilities</td>
</tr>
</tbody>
</table>

**Impossible Tasks**

In some situations, a system is not just "broken" in the sense that it needs to be fixed. Researchers across many fields have shown that certain prediction tasks cannot be solved with machine learning. These are settings in which no specific AI developed for the task can ever possibly work, and a functionality-centered critique can be made with respect to the task more generally. Since these general critiques sometimes rely on philosophical, controversial, or morally contested grounds, the arguments can be difficult to leverage practically and may imply the need for further evidence of failure modes along the lines of our other categories.

**Conceptually Impossible** Certain classes of tasks have been scientifically or philosophically "debunked" by extensive literature. In these cases, there is no plausible connection between observable
data and the proposed target of the prediction task. This includes what Stark and Hutson call “physiognomic artificial intelligence,” which attempts to infer or create hierarchies about personal characteristics from data about their physical appearance [391]. Criticizing the EU Act’s failure to address this inconvenient truth, Veale and Borgesius [418] pointed out that “those claiming to detect emotion use oversimplified, questionable taxonomies; incorrectly assume universality across cultures and contexts; and risk ‘[taking] us back to the phrenological past’ of analysing character traits from facial structures.”

A notorious example of technology broken by definition are attempts to infer “criminality” from a person’s physical appearance. A paper claiming to do this “with no racial bias” was announced by researchers at Harrisburg University in 2020, prompting widespread criticism from the machine learning community [159]. In an open letter, the Coalition for Critical Technology note that the only plausible relationship between a person’s appearance and their propensity to commit a crime is via the biased nature of the category of “criminality” itself [151]. In this setting, there is no logical basis with which to claim functionality.

**Practically Impossible** There can be other, more practical reasons for why a machine learning model or algorithm cannot perform a certain task. For example, in the absence of any reasonable observable characteristics or accessible data to measure the model goals in question, attempts to represent these objectives end up being inappropriate proxies. As a construct validity issue, the constructs of the built model could not possibly meaningfully represent those relevant to the task at hand [214, 215].

Many predictive policing tools are arguably practically impossible AI systems. Predictive policing attempts to predict crime at either the granularity of location or at an individual level [144]. The data that would be required to do the task properly—accurate data about when and where crimes occur—does not and will never exist. While crime is a concept with a fairly fixed definition, it is practically impossible to predict because of structural problems in its collection. The problems with crime data are well-documented—whether in differential victim crime reporting rates [17], selection bias based on policing activities [266, 130], dirty data from periods of recorded unlawful policing [349], and more.

Due to upstream policy, data or societal choices, AI tasks can be practically impossible for one set of developers and not for another, or for different reasons in different contexts. The fragmentation, billing focus, and competing incentives of the US healthcare system have made multiple healthcare-related AI tasks practically impossible [12]. US EHR data is often erroneous, miscoded, fragmented, and incomplete [199, 198], creating a mismatch between available data and intended use [165]. Many of these challenges appeared when IBM attempted to support cancer diagnoses. In one instance, this meant using synthetic as opposed to real patients for oncology prediction data, leading to
“unsafe and incorrect” recommendations for cancer treatments [356]. In another, IBM worked with MD Anderson to work on leukemia patient records, poorly extracting reliable insights from time-dependent information like therapy timelines—the components of care most likely to be mixed up in fragmented doctors’ notes [394, 376].

**Engineering Failures**

Algorithm developers maintain enormous discretion over a host of decisions, and make choices throughout the model development lifecycle. These engineering choices include defining problem formulation [321], setting up evaluation criteria [323, 259], and determining a variety of other details [322, 290]. Failures in AI systems can often be traced to these specific policies or decisions in the development process of the system.

**Model Design Failures**  Sometimes, the design specifications of a model are inappropriate for the task it is being developed for. For instance, in a classification model, choices such as which input and target variables to use, whether to prioritize accepting true positives or rejecting false negatives, and how to process the training data all factor into determining model outcomes. These choices are normative and may prioritize values such as efficiency over preventing harmful failures [258, 113].

In 2014, BBC Panorama uncovered evidence of international students systematically cheating on English language exams run by the UK’s Educational Testing Service by having others take the exam for them. The Home Office began an investigation and campaign to cancel the visas of anyone who was found to have cheated. In 2015, ETS used voice recognition technology to identify this type of cheating. According to the National Audit Office [309],

ETS identified 97% of all UK tests as “suspicious”. It classified 58% of 58,459 UK tests as “invalid” and 39% as “questionable”. The Home Office did not have the expertise to validate the results nor did it, at this stage, get an expert opinion on the quality of the voice recognition evidence. ... but the Home Office started cancelling visas of those individuals given an “invalid” test.

The staggering number of accusations obviously included a number of false positives. The accuracy of ETS’s method was disputed between experts sought by the National Union of Students and the Home Office; the resulting estimates of error rates ranged from 1% to 30%. Yet out of 12,500 people who appealed their immigration decisions, only 3,600 won their cases—and only a fraction of these were won through actually disproving the allegations of cheating. This highly opaque system was thus notable for the disproportionate amount of emphasis that was put into finding cheaters rather than protecting those who were falsely accused. Although we cannot be sure
the voice recognition model was trained to optimize for sensitivity rather than specificity, as the head of the NAO aptly put, "When the Home Office acted vigorously to exclude individuals and shut down colleges involved in the English language test cheating scandal, we think they should have taken an equally vigorous approach to protecting those who did not cheat but who were still caught up in the process, however small a proportion they might be" [309]. This is an example of a system that was not designed to prevent a particular type of harmful failure.

Model Implementation Failures  Even if a model was conceptualized in a reasonable way, some component of the system downstream from the original plan can be executed badly, lazily, or wrong. In 2011, the state of Idaho attempted to build an algorithm to set Medicaid assistance limits for individuals with developmental and intellectual disabilities. When individuals reported sudden drastic cuts to their allowances, the ACLU of Idaho tried to find out how the allowances were being calculated, only to be told it was a trade secret. The class action lawsuit that followed resulted in a court-ordered disclosure of the algorithm, which was revealed to have critical flaws. According to Richard Eppink, Legal Director of the ACLU of Idaho [389],

There were a lot of things wrong with it. First of all, the data they used to come up with their formula for setting people’s assistance limits was corrupt. They were using historical data to predict what was going to happen in the future. But they had to throw out two-thirds of the records they had before they came up with the formula because of data entry errors and data that didn’t make sense.

Data validation is a critical step in the construction of a ML system, and the team that built the benefit system chose to use a highly problematic dataset to train their model. For this reason, we consider this to be an implementation failure.

Another way that failures can be attributed to poor implementation is when a testing framework was not appropriately implemented. One area in which a lack of sufficient testing has been observed in the development of AI is in the area of clinical medicine. Nagendran et al. [294] systematically examined the methods and claims of studies which compared the performance of diagnostic deep learning computer vision algorithms against that of expert clinicians. In their literature review, they identified 10 randomized clinical trials and 81 non-randomized clinical trials. Of the 81 non-randomized studies, they found the median number of clinical experts compared to the AI was 4, full access to datasets and code were unavailable in over 90% of studies, the overall risk of bias was high, and adherence to reporting standards were suboptimal, and therefore poorly substantiate their claims. Similarly, the Epic sepsis prediction model, a product actually implemented at hundreds of hospitals, was recently externally validated by Wong et al. [441], who found that the model had poor calibration to other hospital settings and discriminated against under-represented demographics.
These results suggest that the model’s testing prior to deployment may have been insufficient to estimate its real-world performance. Notably, the COVID-19 technology which resulted from innovation policy and democratization efforts mentioned above was later shown to be completely unsuitable for clinical deployment after the fact [188, 353, 445, 423].

**Missing Safety Features** Sometimes model failures are anticipated yet difficult to prevent; in this case, engineers can sometimes take steps to ensure these points of failure will not cause harm. In 2014, a Nest Labs smoke and carbon monoxide detector was recalled [438]. The detector had a feature which allowed the user to turn it off with a “wave” gesture. However, the company discovered in testing that under certain circumstances, the sensor could be unintentionally deactivated. Detecting a wave gesture with complete accuracy is impossible, and Google acknowledges factors that contribute to the possibility of accidental wave triggering for its other home products [170]. However, the lack of a failsafe to make sure the carbon monoxide detector could not be turned off accidentally made the product dangerous.

In the same way, the National Transportation Safety Board (NTSB) cited a lack of adequate safety measures—such as “a warning/alert when the driver’s hands are off the steering wheel”, “remote monitoring of vehicle operators” and even the companies’ “inadequate safety culture”—as the probable causes in at least two highly publicized fatal crashes of Uber [47, 433] and Tesla [45, 46] self-driving cars. As products in public beta-testing, this lack of functional safeguards was considered to be an even more serious operational hazard than any of the engineering failures involved (such as the vehicle’s inability to detect an incoming pedestrian [47] or truck [45]).

This category also encompasses algorithmic decision systems in critical settings that lack a functional appeals process. This has been a recurring feature in algorithms which allocate benefits on behalf of the government [133]. Not all of these automated systems rely on machine learning, but many have been plagued by bugs and faulty data, resulting in the denial of critical resources owed to citizens. In the case of the Idaho data-driven benefit allocation system, even the people responsible for reviewing appeals were unable to act as a failsafe for the algorithm’s mistakes: “They would look at the system and say, ‘It’s beyond my authority and my expertise to question the quality of this result’ ” [256].

**Deployment Failures**

Sometimes, despite attempts to anticipate failure modes during the design phase, the model does not “fail” until it is exposed to certain external factors and dynamics that arise after it is deployed.

**Robustness Issues** A well-documented source of failure is a lack of robustness to changing external conditions. Liao et al. [259] have observed that the benchmarking methods used for evaluation
in machine learning can suffer from both internal and external validity problems, where “internal validity refers to issues that arise within the context of a single benchmark” and “external validity asks whether progress on a benchmark transfers to other problems.” If a model is developed in a certain context without strong evaluation methods for external validity, it may perform poorly when exposed to real-world conditions that were not captured by the original context. For instance, while many computer vision models developed on ImageNet are tested on synthetic image perturbations in an attempt to measure and improve robustness, but Taori et al. [403] have found that these models are not robust to real-world distribution shifts such as a change in lighting or pose.

Robustness issues are also of dangerous consequence in language models. For example, when large language models are used to process the queries of AI-powered web search [296], the models’ fragility to misspellings [288, 333], or trivial changes to format [35] and context [36] can lead to unexpected results. In one case, a large language model used in Google search could not adequately handle cases of negation [132] – and so when queried with “what to do when having a seizure”, the model alarmingly sourced the information for what not to do, unable to differentiate between the two cases [416].

**Failure under Adversarial Attacks**

Failures can also be induced by the actions of an adversary—an actor deliberately trying to make the model fail. Real-world examples of this often appear in the context of facial recognition, in which adversaries have some evidence that they can fool face-detection systems with, such as 3d-printed masks [325] or software-generated makeup [174]. Machine learning researchers have studied what they call “adversarial examples,” or inputs that are designed to make a machine learning model fail [169]. However, some of this research has been criticized by its lack of a believable threat model—in other words, not focusing on what real-world “adversaries” are actually likely to do [310].

**Unanticipated Interactions**

A model can also fail to account for uses or interactions that it was not initially conceived to handle. Even if an external actor or user is not deliberately trying to break a model, their actions may induce failure if they interact with the model in a way that was not planned for by the model’s designers. For instance, there is evidence that this happened at the Las Vegas Police Department:

As new records about one popular police facial recognition system show, the quality of the probe image dramatically affects the likelihood that the system will return probable matches. But that doesn’t mean police don’t use bad pictures anyway. According to documents obtained by Motherboard, the Las Vegas Metropolitan Police Department (LVMPD) used “non-suitable” probe images in almost half of all the facial recognition
searches it made last year, greatly increasing the chances the system would falsely identify suspects, facial recognition researchers said. [136]

This aligns with reports from Garvie [162] about other police departments inappropriately uploading sketch and celebrity photos to facial recognition tools. It is possible for designers to preempt misuse by implementing instructions, warnings, or error conditions, and failure to do so creates a system that does not function properly.

**Communication Failures**

As with other areas of software development, roles in AI development and deployment are becoming more specialized. Some roles focus on managing the data that feeds into models, others specialize in modeling, and others optimally engineer models for speed and scale [104]. There are even those in "analytics translator" roles – managers dedicated to acting as communicators between data science work and non-technical business leaders [190]. And, of course, there are salespeople. Throughout this chain of actors, potential miscommunications or outright lies can happen about the performance, functional safety or other aspects of deployed AI/ML systems. Communication failures often co-occur with other functional safety problems, and the lack of accountability for false claims – intentional or otherwise – makes these particularly pernicious and likely to occur as AI hype continues absent effective regulation.

**Falsified or Overstated Capabilities**

To pursue commercial or reputational interests, companies and researchers may explicitly make claims about models which are provably untrue. A common form of this are claims that a product is "AI", when in fact it mainly involves humans making decisions behind the scenes. While this in and of itself may not create unsafe products, expectations based on unreasonable claims can create unearned trust, and a potential over-reliance that hurts parties who purchase the product. As an example, investors poured money into ScaleFactor, a startup that claimed to have AI that could replace accountants for small businesses, with the exciting (for accountants) tagline "Because evenings are for families, not finance" [218]. Under the hood, however,

Instead of software producing financial statements, dozens of accountants did most of it manually from ScaleFactor’s Austin headquarters or from an outsourcing office in the Philippines, according to former employees. Some customers say they received books filled with errors, and were forced to re-hire accountants, or clean up the mess themselves. [218]
Even large well-funded entities misrepresent the capabilities of their AI products. Deceptively constructed evaluation schemes allow AI product creators to make false claims. In 2018, Microsoft created machine translation with "equal accuracy to humans in Chinese to English translations" [410]. However, the study used to make this claim (still prominently displayed in press release materials) was quickly debunked by a series of outside researchers who found that at the document-level, when provided with context from nearby sentences, and/or compared to human experts, the machine translation model did not indeed achieve equal accuracy to human translators [409, 255]. This follows a pattern seen with machine learning products in general, where the advertised performance on a simple and static data benchmark, is much lower than the performance on the often more complex and diverse data encountered in practice.

**Misrepresented Capabilities** A simple way to deceive customers into using prediction services is to sell the product for a purpose you know it can’t reliably be used for. In 2018, the ACLU of Northern California revealed that Amazon effectively misrepresented capabilities to police departments in selling their facial recognition product, Rekognition. Building on previous work [58], the ACLU ran Rekognition with a database of mugshots against members of U.S. Congress using the default setting and found 28 members falsely matched within the database, with people of color shown as a disproportionate share of these errors [387]. This result was echoed by Raji and Buolamwini [336] months later. Amazon responded by claiming that for police use cases, the threshold for the service should be set at either 95% or 99% confidence [442]. However, based on a detailed timeline of events [6], it is clear that in selling the service through blog posts and other campaigns that thresholds were set at 80% or 85% confidence, as the ACLU had used in its investigation. In fact, suggestions to shift that threshold were buried in manuals end-users did not read or use – even when working in partnership with Amazon. At least one of Amazon’s police clients also claimed being unaware of needing to modify the default threshold [282].

The hype surrounding IBM’s Watson in healthcare represents another example where a product that may have been fully capable of performing *specific* helpful tasks was sold as a panacea to health care’s ills. As discussed earlier, this is partially the result of functional failures like practical impossibility – but these failures were coupled with deceptively exaggerated claims. The backlash to this hype has been swift in recent years, with one venture capitalist claiming "I think what IBM is excellent at is using their sales and marketing infrastructure to convince people who have asymmetrically less knowledge to pay for something" [440]. At Memorial-Sloan Kettering, after $62 million dollars spent and may years of effort, MD Anderson famously cancelled IBM Watson contracts with no results to show for it [193].

This is particularly a problem in the context of algorithms developed by public agencies – where the AI systems can be adopted as symbols for progress, or smokescreens for undesirable policy
outcomes, and thus liable to inflated narratives of performance. Green [173] discusses how the celebrated success of “self-driving shuttles” in Columbus, Ohio omits its marked failure in the lower-income Linden neighborhood, where residents were now locked out of the transportation apps due to a lack of access to a bank account, credit cards, a data plan or Wi-Fi. Similarly, Eubanks [133] demonstrates how a $1.4 billion contract with a coalition of high-tech companies led an Indiana governor to stubbornly continue a welfare automation algorithm that resulted in a 54% increase in the denials of welfare applications.

5.4 Interventions

The challenge of dealing with an influx of fraudulent or dysfunctional products is one that has plagued many industries, including food safety [44], medicine [20, 34], financial modeling [375], civil aviation [191] and the automobile industry [292, 421]. In many cases, it required the active advocacy of concerned citizens to lead to the policy interventions that would effectively change the tide of these industries. The AI field seems to now be facing this same challenge.

Thankfully, as AI operates as a general purpose technology prevalent in many of these industries, there already exists a plethora of governance infrastructure to address this issue in related fields of application. In fact, healthcare is the field where AI product failures appear to be the most visible, in part due to the rigor of pre-established evaluation processes [443, 264, 352, 37]. Similarly, the transportation industry has a rich history of thorough accident reports and investigations, through organizations such as the National Transportation and Safety Board (NTSB), who have already been responsible for assessing the damage from the few known cases of self-driving car crashes from Uber and Tesla [183].

In this section, we specifically outline the legal and organizational interventions necessary to address functionality issues in general context in which AI is developed and deployed into the market. In broader terms, the concept of functional safety in engineering design literature [386, 355] well encapsulates the concerns articulated in this paper—namely that a system can be deployed without working very well, and that such performance issues can cause harm worth preventing.

Legal/Policy Interventions

The law has several tools at its disposal to address product failures to work correctly. They mostly fall in the category of consumer protection law. This discussion will be U.S.-based, but analogues exist in most jurisdictions.
**Consumer Protection** The Federal Trade Commission is the federal consumer protection agency within the United States with the broadest subject matter jurisdiction. Under Section 5 of the FTC Act, it has the authority to regulate “unfair and deceptive acts or practices” in commerce [138]. This is a broad grant authority to regulate practices that injure consumers. The authority to regulate deceptive practices applies to any material misleading claims relating to a consumer product. The FTC need not show intent to deceive or that deception actually occurred, only that claims are misleading. Deceptive claims can be expressed explicitly—for example, representation in the sales materials that is inaccurate—or implied, such as an aspect of the design that suggests a functionality the product lacks [202, 184]. Many of the different failures, especially impossibility, can trigger a deceptive practices claim.

The FTC’s ability to address unfair practices is wider-ranging but more controversial. The FTC can reach any practice “likely to cause substantial injury to consumers[,] not reasonably avoidable by consumers themselves and not outweighed by countervailing benefits to consumers” [138]. Thus, where dysfunctional AI is being sold and its failures causes substantial harm to consumers, the FTC could step in. Based on the FTC’s approach to data security, in which the Commission has sued companies for failing to adequately secure consumer data in their possession against unknown third-party attackers [279], even post-deployment failures—if foreseeable and harmful—can be included among unfair practices, though they partially attributable to external actors.

The FTC can use this authority to seek an injunction, requiring companies to cease the practice. Formally, the FTC does not have the power to issue fines under its Section 5 authority, but the Commission frequently enters into long-term consent decrees with companies that it sues, permitting continuing jurisdiction, monitoring, and fines for future violations [393, 140]. The Commission does not have general rulemaking authority, so most of its actions to date have taken the form of public education and enforcement. The Commission does, however, have authority to make rules regarding unfair or deceptive practices under the Magnuson-Moss Warranty Act. Though it has created no new rules since 1980, in July 2021, the FTC voted to change internal agency policies to make it easier to do so [139].

Other federal agencies also have the ability to regulate faulty AI systems, depending on their subject matter. The Consumer Product Safety Commission governs the risks of physical injury due to consumer products. They can create mandatory standards for products, can require certifications of adherence to those rules, and can investigate products that have caused harm, leading to bans or mandatory recalls [82]. The National Highway Safety Administration offers similar oversight for automobiles specifically. The Consumer Finance Protection Bureau can regulate harms from products dealing with loans, banking, or other consumer finance issues [1].

In addition to various federal agencies, all states have consumer protection statutes that bar deceptive practices and many bar unfair practices as well, like the FTC Act [62]. False advertising
laws are related and also common. State attorneys general often take active roles as enforcers of those laws [80]. Of course, the efficacy of such laws varies from state to state, but in principle, they become another source of law and enforcement to look to for the same reasons that the FTC can regulate under Section 5. One particular state law worth noting is California’s Unfair Competition Law, which allows individuals to sue for injunctive relief to halt conduct that violates other laws, even if individuals could not otherwise sue under that law [453].

It is certainly no great revelation that federal and state regulatory apparatuses exist. Rather, our point is that while concerns about discrimination and due process can lead to difficult questions about the operation of existing law and proposals for legal reform, thinking about the ways that AI is not working makes it look like other product failures that we know how to address. Where AI doesn’t work, suddenly regulatory authority is easy to find.

**Products Liability Law** Another avenue for legal accountability may come from the tort of products liability, though there are some potential hurdles. In general, if a person is injured by a defective product, they can sue the producer or seller in products liability. The plaintiff need not have purchased or used the product; it is enough that they were injured by it, and the product has a defect that rendered it unsafe.

It would stand to reason that a functionality failure in an AI system could be deemed a product defect. But surprisingly, defective software has never led to a products liability verdict. One commonly cited reason is that products liability applies most clearly to tangible things, rather than information products, and that aside from a stray comment in one appellate case [439], no court has actually ruled that software is even a “product” for these purposes [59, 129]. This would likely not be a problem for software that resides within a physical system, but for non-embodied AI, it might pose a hurdle. In a similar vein, because most software harms have typically been economic in nature, with, for example, a software crash leading to a loss of work product, courts have rejected these claims as “pure economic loss” belonging more properly in contract law than tort. But these mostly reflect courts’ anxiety with intangible *injuries*, and as AI discourse has come to recognize many concrete harms, these concerns are less likely to be hurdles going forward [75].

Writing about software and tort law, Choi [75] identifies the complexity of software as a more fundamental type of hurdle. For software of nontrivial complexity, it is provably impossible to guarantee bug-free code. An important part of products liability is weighing the cost of improvements and more testing against the harms. But as no amount of testing can guarantee bug-free software, it will difficult to determine how much testing is enough to be considered reasonable or non-negligent [75, 209]. Choi analogizes this issue to car crashes: car crashes are inevitable, but courts developed the idea of crashworthiness to ask about the car’s contribution to the total harm, even if the initial injury was attributable to a product defect [75]. While Choi looks to crashworthiness as a solution,
the thrust of his argument is that software can cause exactly the type of injury that products liability
aims to protect us from, and doctrine should reflect that.

While algorithmic systems have a similar sort of problem, the failure we describe here are more
basic. Much as writing bug-free software is impossible, creating a model that handles every corner
case perfectly is impossible. But the failures we address here are not about unforeseeable corner
cases in models. We are concerned with easier questions of basic functionality, without which a system
should never have been shipped. If a system is not functional, in the sense we describe, a court should
have no problem finding that it is unreasonably defective. As discussed above, a product could be
placed on the market claiming the ability to do something it cannot achieve in theory or in practice,
or it can fail to be robust to unanticipated but foreseeable uses by consumers. Even where these
errors might be difficult to classify in doctrinally rigid categories of defect, courts have increasingly
been relying on “malfunction doctrine,” which allows for circumstantial evidence to be used as proof
of defect where “a product fails to perform its manifestly intended function.” [345]. Courts are
increasingly relying on this doctrine and it could apply here [315, 164]. Products liability could
especially easily apply to engineering failures, where the error was foreseeable and an alternative,
working version of the product should have been built.

**Warranties** Another area of law implicated by product failure is warranty law, which protects the
purchasers of defunct AI and certain third parties who stand to benefit from the sale. Sales of goods
typically come with a set of implied warranties. The implied warranty of merchantability applies to
all goods and states, among other things, that the good is “fit for the ordinary purposes for which
such goods are used” [411]. The implied warranty of fitness for particular purpose applies when a
seller knows that the buyer has a specific purpose in mind and the buyer is relying on the seller’s
skill or judgment about the good’s fitness, stating that the good is fit for that purpose[412]. Defunct
AI will breach both these warranties. The remedy for such a breach is limited to contract damages.
This area of law is concerned with ensuring that purchasers get what they pay for, so compensation
will be limited roughly to value of the sale. Injuries not related to the breach of contract are meant
to be worked out in tort law, as described above.

**Fraud** In extreme cases, the sale of defunct AI may constitute fraud. Fraud has many specific
meanings in law, but invariably it involves a knowing or intentional misrepresentation that the
victim relied on in good faith. In contract law, proving that a person was defrauded can lead to
contract damages. Restitution is another possible remedy for fraud. In tort law, a claim of fraud can
lead to compensation necessary to rectify any harms that come from the fraud, as well as punitive
damages in egregious cases. Fraud is difficult to prove, and our examples do not clearly indicate
fraud, but it is theoretically possible if someone is selling snake oil. Fraud can lead to criminal
liability as well.

**Other Legal Avenues Already Being Explored**  Finally, other areas of law that are already involved in the accountability discussion, such as discrimination and due process, become much easier cases to make when the AI doesn’t work. Disparate impact law requires that the AI tool used be adequately predictive of the desired outcome, before even getting into the question of whether it is too discriminatory or not [27]. A lack of construct validity would easily subject a model’s user to liability. Due process requires decisions to not be arbitrary, and AI that doesn’t work loses its claim to making decisions on a sound basis [79]. Where AI doesn’t work, legal cases in general become easier.

**Organizational interventions**

In addition to legal levers, there are many organizational interventions that can be deployed to address the range of functionality issues discussed. Due to clear conflicts of interest, the self-regulatory approaches described are far from adequate oversight for these challenges, and the presence of regulation does a lot to incentivise organizations to take these actions in the first place. However, they do provide an immediate path forward in addressing these issues.

**Internal Audits & Documentation**  After similar crises of performance in fields such as aerospace, finance and medicine, such processes evolved in those industries to enforce a new level of introspection in the form of internal audits. Taking the form of anything from documentation exercises to challenge datasets as benchmarks, these processes raised the bar for deployment criteria and matured the product development pipeline in the process [338]. The AI field could certainly adopt similar techniques for increasing the scrutiny of their systems, especially given the nascent state of reflection and standardization common in ML evaluation processes [259]. For example, the “Failure modes, effects, and diagnostic analysis (FMEDA)” documentation process from the aerospace industry could support the identification of functional safety issues prior to AI deployment [338], in addition to other resources from aerospace (such as the functional hazard analyses (FHA) or Functional Design Assurance Levels (FDALS)).

Ultimately, internal audits are a self-regulatory approach—though audits conducted by independent second parties such as a consultancy firm could provide a fresh perspective on quality control and performance in reference to articulated organizational expectations [340]. The challenge with such audits, however, is that the results are rarely communicated externally and disclosure is not mandatory, nor is it incentivized. As a result, assessment outcomes are mainly for internal use only, often just to set internal quality assurance standards for deployment and prompt further engineering reflection during the evaluation process.
Product Certification & Standards  A trickier intervention is the avenue of product certification and standards development for AI products. This concept has already made its way into AI policy discourse; CEN (European Committee for Standardisation) and CENELEC (European Committee for Electrotechnical Standardisation), two of three European Standardisation Organisations (ESOs) were heavily involved in the creation of the EU’s draft AI Act [418]. On the U.S. front, industry groups IEEE and ISO regularly shape conversations, with IEEE going so far as to attempt the development of a certification program [163, 187]. In the aviation industry, much of the establishment of engineering standards happened without active government intervention, between industry peers [338]. These efforts resemble the Partnership on AI’s attempt to establish norms on model documentation processes [337]. Collective industry-wide decision-making on critical issues can raise the bar for the entire industry and raise awareness within the industry of the importance of handling functionality challenges. Existing functional safety standards from the automobile (ISO 26262), aerospace (US RTCA DO-178C), defense (MIL-STD-882E) and electronics (IEEE IEC 61508 / IEC 61511) industries, amongst others, can provide a template on how to approach this challenge within the AI industry.

Other Interventions  There are several other organizational factors that can determine and assess the functional safety of a system. As a client making decisions on which projects to select, or permit for purchase, it can be good to set performance related requirements for procurement and leverage this procurement process in order to set expectations for functionality [383, 291, 347, 357]. Similarly, cultural expectations for safety and engineering responsibility impact the quality of the output from the product development process – setting these expectations internally and fostering a healthy safety culture can increase cooperation on other industry-wide and organizational measures [355]. Also, as functionality is a safety risk aligned with profit-oriented goals, many model logging and evaluation operations tools are available for organizations to leverage in the internal inspection of their systems – including tools for more continuous monitoring of deployed systems [343, 369].

Conclusion: The Road Ahead

We cannot take for granted that AI products work. Buying into the presented narrative of a product with at least basic utility or an industry that will soon enough “inevitably” overcome known functional issues causes us to miss important sources of harm and available legal and organizational remedies. Although functionality issues are not completely ignored in AI policy, the lack of awareness of the range in which these issues arise leads to the problems being inadequately emphasized and poorly addressed by the full scope of accountability tools available. The fact that faulty AI products are on the market today makes this problem particularly urgent. Poorly vetted products permeate
our lives, and while many readily accept the potential for harms as a tradeoff, the claims of the products’ benefits go unchallenged. But addressing functionality involves more than calling out demonstrably broken products. It also means challenging those who develop AI systems to better and more honestly understand, explore, and articulate the limits of their products prior to their release into the market or public use. Adequate assessment and communication of functionality should be a minimum requirement for mass deployment of algorithmic systems. Products that do not function should not have the opportunity to affect people’s lives.
Part III

Provable guarantees in fair machine learning
Chapter 6

Analyzing the Regularizing Effects of Group-Fair Training on Shared Models

It is well-known that learned models can have performance or outcome disparities on underrepresented or disadvantaged groups in a distribution [57, 298]. Research suggests that these disparities are the result of a complex interaction between the training procedure, model class, and training data [72].

Group-based welfare-centric machine learning attempts to mitigate disparities by optimizing aggregations of per-group risk values, rather than average overall loss. In other words, the task is to approximate \( \arg\min_{h \in H} W(R(h, D_1), \ldots, R(h, D_g)) \) for some malfare function \( W(\cdot) \), where \( R(h, D_i) \) is the risk (average loss) of group \( i \) under model \( h \). Such objectives produce models that fairly compromise among groups in various ways. The malfare function determines the fairness concept; for example, \( w \)-weighted risk minimization is equivalent to optimizing utilitarian malfare \( W_1(S; w) = S \cdot w \), and taking \( \Lambda(\cdot) \) to be the maximum produces the minimax-optimal \( h^* \), a.k.a., the egalitarian or Rawlsian fair model.

However, training with empirical risk is susceptible to “overfitting to fairness,” wherein models overfit small or high-risk minority groups. Cousins [90, 91, 93] shows that generalization error (overfitting) of the overall objective decreases with each group’s sample size, but the current SOTA generalization bounds for group \( i \) depend only on group \( i \)’s sample size. We address this discrepancy; in particular, we show that in fair learning, each group \( i \) effectively learns over a “restricted class” of models that are reasonably likely given the training data for all groups \( j \neq i \), thus we bound their generalization error via Rademacher averages of the restricted class, improving over existing
bounds based on the original hypothesis class.

We begin by introducing notation and preliminary concepts (section 6.1) and situating our approach with respect to existing literature (section 6.1). We derive group-specific bounds on the generalization error of jointly trained models, which benefit from the larger sample size of the majority group (section 6.2). These techniques also translate to improved bounds on the generalization error of the malfare objective itself. Additionally, we experimentally verify our methods on synthetic linear and logistic regression tasks, finding that our bounds better describe the overfitting behavior of fair-learning methods than SOTA analysis (section 6.3). Our analysis allows us to resolve key real-world problems, such as when multiple groups benefit from pooling data to train a single (shared) model. All proofs are relegated to section 6.4.

This chapter has been adapted from Cousins, Kumar, and Venkatasubramanian [98].

6.1 Background: Preliminaries and related work

Preliminaries

We assume a standard supervised learning setting. Given domain label space \( \mathcal{Y} \), domain \( \mathcal{X} \), and codomain \( \mathcal{Y}' \), we assume a hypothesis class \( \mathcal{H} \subseteq \mathcal{X} \rightarrow \mathcal{Y}' \) and loss function \( \ell : \mathcal{Y}' \times \mathcal{Y} \rightarrow \mathbb{R} \). Now, suppose a sample \( (x, y) \) = \( z \in (\mathcal{X} \times \mathcal{Y})^m \) or instance distribution \( \mathcal{D} \) over \( \mathcal{X} \times \mathcal{Y} \). We define the empirical risk of hypothesis \( h \) as

\[
\hat{R}(h, z) = \frac{1}{m} \sum_{i=1}^{m} \ell(h(x_i), y_i),
\]

and the true risk over the distribution \( \mathcal{D} \) as

\[
R(h, \mathcal{D}) = \mathbb{E}_{(x,y)\sim \mathcal{D}}[\ell(h(x), y)].
\]

A standard supervised learning task then identifies the empirical risk minimizer

\[
\hat{h} = \arg\min_{h \in \mathcal{H}} \hat{R}(h, z)
\]

as a proxy for the true risk minimizer

\[
h^* = \arg\min_{h \in \mathcal{H}} R(h, \mathcal{D}).
\]

This framework encapsulates simple supervised settings where \( \mathcal{Y} = \mathcal{Y}' \), such as least-squares regression or hard binary classification, but it also contains more sophisticated supervised learning settings, like probabilistic classification or conditional density estimation.
Group-Fair Learning  This work considers group-fair learning, in which we assume not one instance distribution $D$, but rather $g$ per-group instance distributions $D_{1:g}$, and per-group samples $z_i \sim D_i^{m_i}$, where $m_i$ is the sample size for group $i$. The distribution $D_i$ encapsulates the situations encountered by members of each group $i$, which may vary in $X$ (situations encountered by each group), as well as their conditional labels $Y|X$ (responses or labels to a given situation).

To treat groups fairly, we consider objectives that consider the risk of all groups. In particular, we assume a cardinal malfare function $\mathcal{M}(\cdot): \mathbb{R}^g \to \mathbb{R}$, and we then seek the empirical malfare minimizer

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \mathcal{M}(i \mapsto \hat{R}(h, z_i))$$

as a proxy for the true malfare minimizer

$$h^* = \arg\min_{h \in \mathcal{H}} \mathcal{M}(i \mapsto R(h, D_i)) .$$

On Malfare Functions  The choice of malfare function $\mathcal{M}(\cdot)$ directly encodes how one wishes to make tradeoffs between various groups at various levels of risk. The malfare function is thus a fundamental fair-learning hyperparameter that must be selected to achieve a modeler's desired fairness properties, i.e., choosing a malfare function is equivalent to choosing a fairness concept.

Two popular choices are the utilitarian malfare (weighted average), which generally weights the risk of each group proportional to their size, and the egalitarian malfare, which seeks to lift up the most disadvantaged groups by minimizing the maximum risk. These are in some sense two extremes of a spectrum: utilitarian malfare only weights groups, and does not distinguish between high-risk and low-risk groups (no equitable redistribution), whereas egalitarian malfare considers only the risk of each group, offering preferential treatment to those most in need (no consideration of non-minimal groups). It is known that both of the above malfare functions belong to a general class of such functions.

Definition 1 (Power-Mean Malfare). Suppose some $p \geq 1$, positive probability measure $w \in \triangle_g$, and nonnegative risk vector $S \in \mathbb{R}_{0+}^g$. We define the weighted power-mean as

$$M_p(S; w) \doteq \left( \sum_{i=1}^g w_i S_i^p \right)^{1/p} , \quad M_{\infty}(S; w) \doteq \max_{i \in 1, \ldots, g} S_i . \quad (6.1)$$

Both utilitarian and egalitarian malfare arise as power-mean special-cases $p = 1$ and $p = \infty$, respectively.
The power-mean class is axiomatically justified [105, 171, 90, 93], which motivates its use in a variety of learning and allocation settings [26, 95, 422, 96, 97]. Fairness and robustness are closely linked, and Cousins [92] also motivates power-means, as well as Gini malfare, and other malfare classes, from the perspective of robustness. This work is neutral to the choice of malfare function; we only seek to show that our methods may be applied to any malfare concept that meets certain broad criteria.

We generally assume that $W(\cdot)$ is monotonic, i.e., that increasing any group’s risk never decreases malfare. Furthermore, convex malfare functions are convenient for optimization, and in section 6.2 we utilize this property to efficiently bound Rademacher averages. Finally, several of our bounds have algebraically convenient corollaries if we assume Lipschitz continuity, i.e., small changes to risk yield small changes to malfare. The power-mean malfare family, as well as other malfare classes, such as the Gini class [434, 160] or the utilitarian-maximin class [107, 49, 361], each arise uniquely from their own sets of axioms. Each assume some type of monotonicity, transfer principles, such as the Pigou-Dalton [327, 102], which incentivize equitable redistribution of harm and give rise to convexity, as well as some concept of continuity, which coupled with functional analysis of the resultant class, give rise to Lipschitz continuity. Our criteria for malfare functions are thus quite reasonable.

**Statistical Background** The Rademacher average is a key statistical tool used to bound the supremum deviation of empirical means from their expectations [30]. Denote the loss class $\ell \circ H \doteq \{(x, y) \mapsto \ell(h(x), y) \mid h \in H\}$, and define Rademacher averages as follows.

**Definition 2** (Rademacher Averages). Let $\sigma_1, \ldots, \sigma_m$ be a vector of $m$ i.i.d. $\text{Unif}(\pm 1)$ random variables. The empirical Rademacher average is then

$$ \hat{R}_m(\ell \circ H, z) = \mathbb{E}_\sigma \left[ \sup_{h \in H} \frac{1}{m} \sum_{i=1}^m \sigma_i \ell(h(x_i), y_i) \right], $$

i.e., the maximum correlation of any $h \in H$ with noise on a sample $z \in (X \times Y)^m$, and the Rademacher average is its expectation over i.i.d. samples from $D$, i.e.,

$$ R_m(\ell \circ H, D) = \mathbb{E}_{z \sim D^m} \left[ \hat{R}_m(\ell \circ H, z) \right]. $$

Assuming bounded loss range $r$, let $\varepsilon_i = r \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}$ and $\eta_i \doteq 2\hat{R}_{m_i}(\ell \circ H, z_i) + 2\varepsilon_i$. For any failure probability $\delta$, $h \in H$, and group $i$, sampling error is bounded as

$$ \mathbb{P}_{z_i \sim D_i^m} \left( \left| \hat{R}(h, z_i) - R(h, D_i) \right| > \varepsilon_i \right) < 2\delta. \quad (6.2) $$
Moreover, considering all \( h \in \mathcal{H} \) simultaneously, we have for each group \( i \) that
\[
P_{z_i \sim D_i} \left( 2 \mathbb{R}_{m}(\ell \circ \mathcal{H}, D_i) > 2 \hat{R}_{m}(\ell \circ \mathcal{H}, z_i) + \varepsilon_i \right) \cup \sup_{h \in \mathcal{H}} \left| \hat{R}(h, z_i) - R(h, D_i) \right| > \hat{\eta}_i \right) < 3\delta . \tag{6.3}
\]

Equations (6.2) and (6.3) are used throughout for various hypotheses and hypothesis classes, both in the above forms, and as 1-tailed variants. These “textbook results” are now standard in learning theory\(^1\) [367, 287].

The quantity \( |\hat{R}(h, z_i) - R(h, D_i)| \) of (6.2) is the absolute deviation between the empirical risk and the expected risk for each individual \( h \in \mathcal{H} \), and it bounds the estimation error (i.e., error due to sampling) of any such function. The quantity \( \sup_{h \in \mathcal{H}} |\hat{R}(h, z_i) - R(h, D_i)| \) of (6.3) is known as the supremum deviation over the loss class \( \ell \circ \mathcal{H} \), and it bounds the generalization error, both due to sampling error and due to selection bias (training), of the learned \( \hat{h} \).

From (6.3) and a union-bound over groups, following Cousins [90, 91, 93], we probabilistically bound each group’s generalization error (training-true risk gap) as
\[
P_{z_i \sim D_i} \left( \forall i : \left| \hat{R}(\hat{h}, z_i) - R(\hat{h}, D_i) \right| \leq \hat{\eta}_i \right) \geq 1 - 3g\delta . \tag{6.4}
\]

Moreover, if \( \Lambda(S) \) is monotonically increasing in \( S \), then the malfare generalization error obeys
\[
P_{z_i \sim D_i} \left( \begin{array}{c}
\Lambda(i \mapsto \hat{R}(\hat{h}, z_i) - \hat{\eta}_i) \\
\leq \Lambda(i \mapsto R(\hat{h}, D_i)) \quad \text{(Empcl. LB)}
\end{array} \right) \leq \Lambda(i \mapsto \hat{R}(\hat{h}, D_i) + \hat{\eta}_i) \quad \text{(Empcl. UB)} \geq 1 - 3g\delta , \tag{6.5}
\]
i.e., the true malfare of \( \hat{h} \) is sandwiched by upper and lower bounds in terms of empirical malfare.

Finally, using also a union bound over (6.2), the gap between the true risk of the empirical malfare minimizer \( \hat{h} \) and the true malfare minimizer \( h^* \) is
\[
P_{z_i \sim D_i} \left( \begin{array}{c}
\Lambda(i \mapsto R(\hat{h}, D_i) - \hat{\eta}_i) \\
\leq \Lambda(i \mapsto R(h^*, D_i) + \varepsilon_i)
\end{array} \right) \geq 1 - 5g\delta . \tag{6.6}
\]

**Related work**

This work follows others in group-based welfare-centric fair machine learning. This often takes the form of Rawlsian or egalitarian learning, also known as minimax fair learning, wherein \( \Lambda(\cdot) \) is the maximum function, and the goal is to minimize the maximum (over groups) average loss [109, 371, 4, 276, 252, 87, 371, 115], which is a form of distributionally robust optimization [208, 203].

\(^1\)Constants vary between sources, depending on definitions and derivations. Our probabilistic statements use 2-tailed Hoeffding [1963] bounds and 3 applications of McDiarmid’s [1989] inequality, with optimal bounded differences, for Rademacher averages with no absolute value inside the supremum.
Most such works only consider performance over the training set, but the Seldonian learner framework \cite{408} explicitly requires trained models be probably approximately optimal w.r.t. some constrained nonlinear objective. Similarly, the fair-PAC learning framework \cite{90, 93} considers malfare minimization with power-mean objectives.

Due to the nonlinearity of $W(\cdot)$, existing work bounds generalization error separately for each group $j$, and applies assumed Lipschitz or Hölder continuity properties of $W(\cdot)$ to bound the overall objective \cite{90, 91, 93}. In this work, we show sharper bounds on the generalization error of malfare objectives, but we also seek to bound each group’s generalization error.

**Multitask learning** There is overlap between group fair learning (GFL) and multitask learning (MTL). This work shows that GFL reduces generalization error for all groups (particularly smaller groups), which is essentially the motivation for MTL. In both cases, we have $g$ distributions (per-group in GFL, per-task in MTL) and some objective that considers each distribution through $R(h, D_i)$. To our knowledge, there is no published work in multitask learning on objectives that treat tasks nonlinearly, i.e., the objective is always \cite{63, 451, 452}

$$\hat{h} = \arg\min_{h \in H} \sum_{i=1}^{g} \frac{1}{m_i} \sum_{j=1}^{m_i} \ell_i(h(x_{i,j}), y_{i,j}).$$

Existing MTL analysis bounds generalization error by considering all data at once \cite{450, 452}; assuming $m$ samples each for $g$ groups, VC dimension, Rademacher averages, etc. bound total estimation error as $O(\sqrt{\ln \frac{1}{m} g})$. Such methods do not apply in our setting, as we seek per-group generalization bounds and treat nonlinear objectives, thus ultimately we do not expect bounds of this order.

**To pool or not to pool** Some work directly addresses the tradeoff between training pooled versus separate models for groups. Dwork et al. \cite{121} define the cost-of-coupling as

$$\max_{\mathcal{D}} \left( \min_{h \in H} \sum_{i=1}^{g} R(h, D_i) - \sum_{i=1}^{g} \min_{h \in H} R(h, D_i) \right),$$

i.e., worst-case difference between the sum risk of the optimal shared model $\hat{h}$, vs. sum risk of optimal per-group models $\hat{h}_{1:g}$. When this quantity is positive, training with pooled data may require tradeoffs in accuracy across groups. They then introduce transfer learning methods to train per-group classifiers $\hat{h}_{1:g}$ while leveraging available data where appropriate. Similarly to our work, this results in improved VC-theoretic groupwise bounds on generalization error than fully separated training. However, the goal of our learning framework is still to learn a joint model, avoiding thorny questions of disparate treatment. Wang et al. \cite{428} also examine the tradeoff, where the metric of
interest or benefit of splitting is based on an egalitarian notion of fairness. They largely focus on the infinite-samples or known-distributions settings; however, they provide VC-theoretic generalization bounds on the benefit of splitting. These are necessarily worst-case (over possible distributions), and specific to binary classification, whereas we provide data-dependent Rademacher average bounds applicable to a broad range of supervised and unsupervised settings.

6.2 Bounding generalization error in fair training

The generalization error analysis of section 6.1 does not take into account the fact that learning is not equally likely to produce any \( h \in \mathcal{H} \). In this section, we present a sharper analysis that reflects this, both in per-group generalization error bounds, and in the overall generalization error of a malfare objective.

Our approach is to take the core idea of localization [31] — restricting the function class of interest to a subset that with high probability contains the function that will be learned — and generalize it to apply in multi-group fair learning settings. In Section 6.2 we argue that, for each group \( i \), with high probability, the learned function \( \hat{h} \) belongs to some \( \mathcal{H}^*_i \subseteq \mathcal{H} \). We bound generalization error over \( \mathcal{H}^*_i \), with \( \hat{R}_{m,i}(\mathcal{H}^*_i, z_i) \), where often \( \hat{R}_{m,i}(\mathcal{H}^*_i, z_i) \ll \hat{R}_{m,i}(\mathcal{H}, z_i) \). The analysis depends on the group index \( i \), since while analyzing group \( i \), we can treat the training samples \( z_j \) as constant for each \( j \neq i \), but the class \( \mathcal{H}^*_i \) must not depend on \( z_i \) for vital technical reasons (see proof of theorem 4; we require \( \mathcal{H}^*_i \) to be established independently from \( z_i \) in \( \hat{R}_{m,i}(\mathcal{H}^*_i, z_i) \)). We thus establish a theoretical hypothesis class that directly depends on the training sample for each \( j \neq i \), but depends on the distribution for group \( i \) instead of its training sample.

Unfortunately, \( \mathcal{H}^*_i \) is a theoretical object (not actually known, as it depends on \( D_i \)). Thus, we have little recourse but to relax to dependence on purely empirical quantities. We thus establish in Section 6.2 an empirical class \( \hat{\mathcal{H}}_i \), which depends on \( z_i \) instead of \( D_i \). At first glance this seems to violate core statistical precepts, but through careful construction, we show that \( \hat{\mathcal{H}}_i \) acts merely as a probabilistic proxy for \( \mathcal{H}^*_i \).

Finally, we must actually estimate the relevant Rademacher bounds. In Section 6.2 we illustrate how this can be done for linear hypothesis classes using Monte-Carlo Rademacher averaging.

Theoretical Restricted Classes

When bounding the generalization error of group \( i \), we want to construct a restricted hypothesis class leveraging information given by the remaining group samples, in particular their empirical risks. However, we can’t directly use the group \( i \) sample \( z_i \), so instead we bound empirical risk \( \hat{R}(h, z_i) \) in terms of \( R(h, D_i) \). Intuitively, we want this restricted class to be the set of all \( h \in \mathcal{H} \)
that could reasonably be the function we learn from all data (the empirical malfare minimizer $\hat{h}$), where the restricted class is constructed after observing only the data $z_j$ for all groups $j \neq i$.

Similar techniques are common in learning theory and the study of localization, where a theoretical class is constructed based on the (unknown) distribution(s), and subsequently an empirical class that is with high probability a superset which can be built from the data. Our approach, however, is unique in that it is in some sense half-empirical, as the theoretical class depends on the distribution $D_i$ of one group, and the samples $z_j$ from all groups $j \neq i$. We do this instead of constructing a "fully theoretical" class using only the distributions $D_1: g$, as well as an empirical variant based on all training samples $z_1: g$, which would be substantially larger.

First, let $\varepsilon_i \doteq r \sqrt{\frac{\ln \frac{1}{\delta_i}}{2m_i}}$ and $\eta_i \doteq 2R_{m_i}(\ell \circ \mathcal{H}, D_i) + \varepsilon_i$, where $m_i$ is the sample size for group $i$ and $r$ is the range of loss values in $\ell \circ \mathcal{H}$. Recall that the empirical malfare minimization task is to select

$$
\hat{h} \doteq \arg\min_{h' \in \mathcal{H}} \mathcal{M}(j \mapsto \hat{R}(h', z_j))
$$

but since we can't observe sample $i$ yet, we (pessimistically) upper-bound the objective value (w.h.p.) as

$$
\inf_{h' \in \mathcal{H}} \mathcal{M}(j \mapsto \hat{R}(h', z_j); w) \leq \\
\inf_{h' \in \mathcal{H}} \mathcal{M}(j \mapsto \begin{cases} 
\hat{R}(h', z_j) & j \neq i \\
R(h', D_i) + \varepsilon_i & j = i 
\end{cases}),
$$

and (optimistically) lower-bound the empirical malfare of all $h \in \mathcal{H}$, w.h.p. simultaneously, as

$$
\mathcal{M}(j \mapsto \hat{R}(h, z_j); w) \geq \\
\mathcal{M}(j \mapsto \begin{cases} 
\hat{R}(h, z_j) & j \neq i \\
R(h, D_i) - \eta_i & j = i
\end{cases}).
$$

Via this analysis, we then construct our theoretical class, which with high probability shall contain the empirical malfare minimizer $\hat{h}$, as the subset $\mathcal{H}^*_i \subseteq \mathcal{H}$ constrained to $h$ such that

$$
\mathcal{M}(j \mapsto \begin{cases} 
\hat{R}(h, z_j) & j \neq i \\
R(h, D_i) - \eta_i & j = i
\end{cases}) \leq \\
\inf_{h' \in \mathcal{H}} \mathcal{M}(j \mapsto \begin{cases} 
\hat{R}(h', z_j) & j \neq i \\
R(h', D_i) + \varepsilon_i & j = i
\end{cases}). \quad (6.7)
$$
This construction is valid (formalized in theorem 3), as we took any \( h \) that optimistically could outperform a pessimistic estimate of the empirical objective.

The LHS is a “best case” estimate of the empirical malfare of a candidate hypothesis, whereas the RHS is a “worst case” estimate of the minimal empirical malfare, because we want our restricted hypothesis class to be large enough to contain any \( h \in H \) that might be the empirical malfare minimizer. In particular, the LHS uses a Rademacher average bound (6.3), as the bound must apply to all \( h \in H \), but a simple tail-bound term (6.2) suffices on the RHS, as we are comparing to a bound involving some specific \( h' \) (not dependent on the data \( z_i \)).

Intuitively, for utilitarian malfare, \( \hat{H}_i \) describes models that definitely perform well for groups \( j \neq i \), and will probably perform well for group \( i \). Some malfare functions, such as power-means, are undefined for negative risk values, and the LHS risk lower bounds \( R(h, D_i) - \eta_i \) may be negative. However, if we assume risk (or loss) is nonnegative, we may use the risk lower bound \( \max(0, R(h, D_i) - \eta_i) \), which preserves convexity, continuity, and even differentiability if \( p < \infty \) except around the point \( 0 \).

Observe now that, conditioning on \( z_j \) for each \( j \neq i \), with high probability over choice of \( z_i \), empirical malfare minimization yields some \( \hat{h} \in H_i^* \). Therefore, for all intents and purposes, learning occurs over \( H_i^* \), and we may thus use Rademacher averages over this restricted class to bound generalization error for group \( i \). Formally put, we have the following result.

**Theorem 3** (Theoretical Group-Regularized Malfare Bounds). Suppose a monotonic malfare function \( M(\cdot) : \mathbb{R}^g \to \mathbb{R} \), hypothesis class \( H \subseteq \mathcal{X} \to \mathcal{Y} \), loss function \( \ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R} \), per-group distributions \( D_{1:g} \) over \( \mathcal{X} \times \mathcal{Y} \), and per-group samples \( z_{1:g} \), with \( z_j \sim D_j^{m_j} \) for each group \( j \). Fix any group index \( i \), and take \( H_i^* \) defined as in (6.7). The following then hold:

1. With probability at least \( 1 - 2\delta \) over choice of \( z_i \), it holds that \( \hat{h} \in H_i^* \).
2. With probability at least \( 1 - 4\delta \) over choice of \( z_i \),

\[
\left| R(\hat{h}, D_i) - \hat{R}(\hat{h}, z_i) \right| \leq 2\mathbb{K}_{m_i} (\ell \circ H_i^*, D_i) + \varepsilon_i .
\]

**Empirical Restricted Classes**

\( H_i^* \) is an object only of theoretical interest (it is not actually known, since it depends on \( D_i \)). Consequently, without more information, \( \mathbb{K}_{m_i} (\ell \circ H_i^*, D_i) \), and thus the bounds of theorem 3, can not be computed. We remedy this issue here, relaxing dependence on the distribution \( D_i \) by replacing it with dependence on the training sample \( z_i \) and thus establishing a new *empirically restricted hypothesis class*.

Note that we can’t simply substitute \( \hat{R}(h, z_i) \) for \( R(h, D_i) \), as theorem 3 clearly requires the restricted hypothesis class \( H_i^* \) to be fixed before observing the training data \( z_i \). We account for this
Figure 6.1: Visualization of unrestricted class $\mathcal{H}$, theoretical restricted class $\mathcal{H}_*^i$, and samples of empirical restricted class $\hat{\mathcal{H}}_i$ (varying $z_i$). One possible empirical malfare minimizer $\hat{h}$ (contained by $\mathcal{H}_i$ and $\mathcal{H}_*^i$ with high probability), as well as the true malfare minimzer $h^*$ (which may fall outside of $\mathcal{H}_*^i$ or $\mathcal{H}_i$ due to overfitting to groups other than $i$) are also shown.

by indirectly using $\hat{R}(h, z_i)$ to bound $R(h, D_i)$. In particular, take $\hat{\eta}_i = 2\hat{R}_{m_i}(\ell \circ \mathcal{H}, z_i) + 2\varepsilon_i$, and take $\varepsilon_i = r\sqrt{\frac{\ln \frac{1}{\delta}}{2m}}$, as in (6.3). Now, we construct our empirical class $\hat{\mathcal{H}}_i$, which with high probability shall contain the theoretical class $\mathcal{H}_*^i$, as the subset $\hat{\mathcal{H}}_i \subseteq \mathcal{H}$ constrained to $h$ such that

$$\inf_{h' \in \mathcal{H}} \mathcal{M} \left( j \mapsto \begin{cases} j \neq i & \hat{R}(h, z_j) \\ j = i & \hat{R}(h, z_i) - 2\hat{\eta}_i \end{cases} \right) \leq \inf_{h' \in \mathcal{H}} \mathcal{M} \left( j \mapsto \begin{cases} j \neq i & \hat{R}(h', z_j) \\ j = i & \hat{R}(h', z_i) + 2\varepsilon_i \end{cases} \right).$$

(6.8)

Note that (6.8) matches (6.7), except risks and Rademacher averages are bounded in terms of their empirical counterparts. In particular, on the LHS, w.h.p., for all $h \in \mathcal{H}$ it holds $\hat{R}(h, z_i) - 2\hat{\eta}_i \leq R(h, D_i) - \eta_i$, and on the RHS, w.h.p., $\hat{R}(h', z_i) + 2\varepsilon_i \geq R(h', D_i) + \varepsilon_i$. Figure 6.1 visualizes the difference between $\hat{\mathcal{H}}_i$ and $\mathcal{H}_*^i$, as well as other key players.

Observe now that, with high probability, $\mathcal{H}_*^i \subseteq \hat{\mathcal{H}}_i$, therefore we can employ the theoretical properties of $\mathcal{H}_*^i$ while being able to compute everything from a sample using $\hat{\mathcal{H}}_i$. The following theorem makes precise this statement, and should be viewed as an empirical counterpart to theorem 3.
**Theorem 4** (Empirical Group-Regularized Malfare Bounds). Suppose as in theorem 3. The following then hold for $\hat{H}_i$ defined as in (6.8).
1. With probability at least $1 - 4\delta$ over choice of $z_i$, it holds that $\hat{h} \in H_i^* \subseteq \hat{H}_i$.
2. With probability at least $1 - 6\delta$, it holds that
   \[
   \left| R(\hat{h}, D_i) - \hat{R}(\hat{h}, z_i) \right| \leq 2\tilde{R}_{m_i}(\ell \circ \hat{H}_i, z_i) + 2\varepsilon_i .
   \]

Theorem 4 satisfies our primary goal of showing per-group generalization bounds for fair learning that leverage information from other groups. In particular, when $\hat{H}_i \subset H_i$, we obtain sharper generalization bounds, which quantifies the intuition that training a shared model is less susceptible to overfitting than training per-group models. Theorem 4 part 2 should be contrasted with (6.4), which gives a similar guarantee using Rademacher averages of the unrestricted class $H_i$. Corollary 3 now applies these bounds to improve the state-of-the-art generalization guarantees for (nonlinear) malfare objectives, which would otherwise depend on Rademacher averages of $\hat{H}_i$ rather than $H_i$, cf. (6.5).

**Corollary 3** (Empirical Malfare Generalization Bounds). Suppose as in theorem 4. Suppose also that there exists some $\lambda > 0$ and norm $\|\cdot\|_M$ such that $M(\cdot)$ is $\lambda\|\cdot\|_M$ Lipschitz continuous, i.e.,
\[
\forall S, S': M(S + S') \leq M(S) + \lambda\|S'\|_M .
\]
We then have:
1. With probability at least $1 - 5g\delta$, the true malfare of $\hat{h}$ is bounded by
   \[
   M(j \mapsto R(\hat{h}, D_j)) \leq M(j \mapsto R(\hat{h}, z_j) + 2\tilde{R}_{m_j}(\hat{H}_j, z_j) + 2\varepsilon_j)
   \leq M(j \mapsto R(\hat{h}, z_j)) + \lambda\|j \mapsto 2\tilde{R}_{m_j}(\hat{H}_j, z_j) + 2\varepsilon_j\|_M .
   \]
2. With probability at least $1 - 6g\delta$, we bound the suboptimality of $\hat{h}$ as
   \[
   M(j \mapsto R(\hat{h}, D_j)) \leq M(j \mapsto R(h^*, D_j)) + 2\tilde{R}_{m_j}(\hat{H}_j, z_j) + 3\varepsilon_j.
   \]
   \[
   \Rightarrow \left| M(j \mapsto R(h^*, D_j)) - M(j \mapsto R(\hat{h}, D_j)) \right| \leq \lambda\|j \mapsto 2\tilde{R}_{m_j}(\hat{H}_j, z_j) + 3\varepsilon_j\|_M .
   \]

The first inequality of parts 1 & 2 of corollary 3 is sharper, but the second is generally more analytically convenient. In particular, any power-mean malfare function $\Lambda_p(\cdot; w)$ obeys
\[
\Lambda_p(S + S'; w) - \Lambda_p(S; w) \leq \Lambda_p(S'; w) \leq \|S'\|_\infty ,
\]
thus we bound malfare generalization error in terms of the generalization error of each group.

Naturally, one may ask how sharp this localization strategy is. We now show an example where theorem 4 improves slow $O\left(\frac{1}{\sqrt{m}}\right)$ convergence rates to fast $O\left(\frac{1}{m}\right)$ convergence rates. Consider
unit-range 0-dimensional linear regression, i.e., mean estimation under square loss \( \ell \), with \( g = 1 \). Thus we have

\[
\ell \circ \mathcal{H} = \left\{ \ell(h_c(x), y) = (c - y)^2 \ \bigg| \ c \in [-r, r] \right\}
\]

with \( r = 1 \). Take constant probability distribution \( D = 0 \), thus \( y = 0 \). From random walk theory, we have

\[
\hat{R}^2_m(\ell \circ \hat{H}, y) = \sum_{i=1}^{m} \sigma_i c^2
\]

\[
= \frac{r^2}{2} \sum_{i=1}^{m} \sigma_i \approx r^2 \sqrt{\frac{1}{2\pi m}}
\]

To construct \( \hat{H} \), observe that we have \( \hat{R}(c, y) = c^2 \), thus via (6.8) we restrict s.t. \( c^2 \leq 4\hat{R}^2_m(\ell \circ \mathcal{H}, y) + 6\varepsilon \approx \sqrt{\frac{s}{\pi m}} + 6\sqrt{\frac{\ln \frac{1}{2m}}{2m}} \implies |c| \leq r \in \Theta \sqrt{\frac{1}{m}} \). We thus have

\[
\hat{R}^2_m(\ell \circ \hat{H}, y) \approx r^2 \sqrt{\frac{1}{2\pi m}} \in \Theta \left( \frac{1}{m} \right)
\]

which asymptotically improves \( \hat{R}^2_m(\ell \circ \mathcal{H}, y) \approx \sqrt{\frac{1}{2\pi m}} \).

**Monte-Carlo Rademacher Averages of Linear Hypothesis Classes**

We now present a method to estimate Rademacher averages for linear hypothesis classes using Monte-Carlo sampling. We start by noting that, in general if \( \ell(\hat{y}, y) = f(g(\hat{y}, y)) \) and \( f \) is \( \lambda \)-Lipschitz-continuous, then we have, for any \( z \in (\mathcal{X} \times \mathcal{Y})^m \), that

\[
\hat{R}^2_m(\ell \circ \mathcal{H}, z) \leq \lambda \hat{R}^2_m(g \circ \mathcal{H}, z)
\]  

(6.10)

For this reason, we formulate the Rademacher averages of both linear least-squares regression and logistic regression as follows. Take \( \mathcal{H} = \{ h_\beta(x) = \beta \cdot x \ \big| \ \beta \in B \} \) and loss function \( \ell(\hat{y}, y) = f(g(\hat{y}, y)) \), where for least-squares regression, \( g(\hat{y}, y) = \hat{y} - y \) and \( f(u) = u^2 \). This is \( \lambda \)-Lipschitz continuous, assuming bounded \( B, \mathcal{X}, \) and \( \mathcal{Y} \), with \( \lambda = 2 \sup_{B, \mathcal{X}, \mathcal{Y}} |x \cdot \beta - y| \). For logistic regression, in which \( \mathcal{Y} = \pm 1 \), we have \( g(\hat{y}, y) = \hat{y} \cdot y \) and \( f(u) = \ln(1 + \exp(u)) \), which is 1-Lipschitz.

\( ^2 \)In practice, we compute the Lipschitz constant over \( \mathcal{H} \), rather than over \( \mathcal{H}_i \subseteq \mathcal{H} \), which would require computing the diameter of \( \mathcal{H}_i \) or bounding the range of \( g \circ \mathcal{H}_i \).
Estimation  Standard methods for bounding Rademacher averages of linear regression classes start by bounding the Rademacher average of \( \mathcal{H} \) itself [367]. However, this method is loose [94], and seems especially so for irregular weight spaces (i.e., those not defined by simple \( p \)-norms), which known analytic methods can not handle.

Instead, we directly estimate the Rademacher average of the function family \( g \circ \mathcal{H} \) directly using Monte-Carlo estimation. That is to say, given sampled Rademacher random variables \( \sigma \in (\pm 1)^{n \times m} \)
and data sample \( z \in (\mathcal{X} \times \mathcal{Y})^m \), we compute \( \hat{\mathbf{R}}_m^n(g \circ \mathcal{H}, z; \sigma) \triangleq \frac{1}{n} \sum_{k=1}^n \sup_{\beta \in W} \frac{1}{m} \sum_{j=1}^m \sigma_{k,j} g(x_j, \beta, y_j) \). This fully data-dependent method gracefully tolerates arbitrary data distributions and parameter spaces, and is loose only in a small amount of Monte-Carlo error and the contraction inequality [94]. In practice, we use \( \lambda \hat{\mathbf{R}}_m^n(g \circ \mathcal{H}, z_i; \sigma) \) as a plug-in estimate of \( \Lambda \hat{\mathbf{R}}_m^n(g \circ \mathcal{H}, z_i) \), which then bounds \( \hat{\mathbf{R}}_m^\ell(g \circ \mathcal{H}, z_i) \) via (6.10). We similarly estimate and bound Rademacher averages over our restricted hypothesis classes as \( \lambda \hat{\mathbf{R}}_m^\ell(g \circ \mathcal{H}_i, z_i; \sigma) \).

**Lemma 4** (Convex Optimization for Monte-Carlo Rademacher Averages). Suppose the parameter space \( \mathcal{B} \) of \( \mathcal{H} \) is a convex set, loss \( \ell(h_\beta(x), y) \) is convex in \( \beta \in \mathcal{B} \) for all \( x \in \mathcal{X}, y \in \mathcal{Y} \), and malfare \( \Lambda(\cdot) : \mathbb{R}^\beta \to \mathbb{R} \) is quasiconvex and monotonically increasing in each argument. Then the parameter spaces of \( \mathcal{H}_i \) and \( \mathcal{H}_i^* \) are convex sets.

Moreover, if \( g \circ \mathcal{H} \) is an affine function family, then \( \hat{\mathbf{R}}_m^n(g \circ \mathcal{H}_i, z_i; \sigma) \) reduces to maximizing a linear function over a convex set. Similarly, if we strengthen the quasiconvexity assumption on \( \Lambda(\cdot) \) to convexity, then EMM reduces to minimizing a convex objective over the convex set \( \mathcal{B} \).

All code used to generate the results in this paper are available upon request. The computation of each supremum in (6.11), i.e., \( \hat{\mathbf{R}}_m^n(g \circ \mathcal{H}, z_i; \sigma) \) and \( \hat{\mathbf{R}}_m^n(g \circ \mathcal{H}_i, z_i; \sigma) \) optimize linear functions of \( \beta \). However, since the restricted hypothesis constraints are convex functions (see lemma 4) over the parameter space, we need to use solvers that can handle nonlinear convex constraints. For this reason, we use either the ECOS [114] or SCS [300] algorithms available in CVXPY [108, 11]. These algorithms are also able to compute the upper bound of the restricted hypothesis class constraint itself, which minimizes an objective which is dependent on the loss \( \ell \).

Visualizing \( \mathcal{H}_i \) in least-squares regression  For least-squares regression, under utilitarian malfare, the restricted hypothesis constraint of \( \mathcal{H}_i \) is an ellipsoid (under egalitarian welfare, it is an intersection of ellipsoids). We visualize a simple example in fig. 6.2, with parameters and results described in table 6.1.

Taking \( \mathcal{B} \triangleq \{ \beta \in \mathbb{R}^2 ||\beta||_1 \leq 1 \} \) to be the unit \( \ell_1 \) ball, we sample \( (x, y) \) as \( x \sim \text{Unif}([-1, 1]^2) \), \( y = x \cdot \beta + \text{Unif}([-1, 1]) \), where each group has slightly different data generating parameters \( \beta \). In fig. 6.2, taking \( \delta = 0.1 \), we plot the \( n = 100 \) values of \( \beta \) which realize each supremum of (6.11) for
Figure 6.2: Rademacher average samples in the parameter space of \( \hat{H}_i \) for each group \( i \in \{1, 2, 3\} \).
Table 6.1: Sample sizes $m_{1:3}$, parameter vectors $\beta_{1:3}$, and Monte-Carlo empirical Rademacher averages (MCERA) for both $H$ and $\hat{H}_{1:3}$.

<table>
<thead>
<tr>
<th>Group ID</th>
<th>$m_i$</th>
<th>True $\beta$</th>
<th>MCERA $H$</th>
<th>$\hat{H}_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6500</td>
<td>(0.3,0.3)</td>
<td>0.047</td>
<td>0.046</td>
</tr>
<tr>
<td>2</td>
<td>3000</td>
<td>(-0.1,0.1)</td>
<td>0.075</td>
<td>0.046</td>
</tr>
<tr>
<td>3</td>
<td>500</td>
<td>(0.3,0)</td>
<td>0.183</td>
<td>0.135</td>
</tr>
</tbody>
</table>

some Rademacher sample $\sigma_k$. These points necessarily lie on either (the corner of) the $\ell_1$ constraint boundary of $B$ or the restricted hypothesis constraint boundary of $\hat{H}_i$, illustrated by the concentric ellipses, which represent constant upper-bounds of weighted utilitarian malfare over the whole dataset and are centered around $\hat{h}$.

Note that for the smallest group, 3, the fact that $\hat{h}$ must perform well on the other two groups under weighted utilitarian malfare shrinks $\hat{H}_3$ significantly. However, the generalization bound over the largest group, 1, is not significantly improved when taken over $\hat{H}_1$.

### 6.3 Experiments

We illustrate the utility of our results with some experiments. Our approach is to construct an example dataset where we can demonstrate a clear benefit (to minority groups) to pooled training, and then show how our refined generalization bounds are in fact sharper than standard Rademacher bounds. We do this by assuming that the individual distributions of the groups are similar enough that, for underrepresented minority groups, pooled training reduces generalization error.

Our experiments are based on a binary logistic regression task with 3 groups. Suppose the unit $\ell_\infty$ ball domain, i.e., $\mathcal{X} = [-1,1]^{15}$, binary label space $\mathcal{Y} = \pm 1$, and parameter space $B = \{\beta \in \mathbb{R}^{15} \mid \|\beta\|_1 \leq 15\}$. For each group $i$, we generate samples $(x, y)$ with $x \sim \text{Unif}(\mathcal{X})$, $\mathbb{P}(y = 1) = \text{logistic}(x \cdot \beta_i + \xi)$, with noise $\xi \sim \mathcal{N}(0,0.1)$, for $\text{logistic}(u) = \frac{1}{1 + \exp(-u)}$.

We assume groupwise data generating parameters and a constant proportional composition of the full training sample as in table 6.2. Notably, the data generating model for groups 1 and 3 are very similar, but there is always much more data available for group 1.

Table 6.2: Data generating parameters for logistic regression experiments.

<table>
<thead>
<tr>
<th>Data proportion</th>
<th>True parameters $\beta_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>75%</td>
</tr>
<tr>
<td>Group 2</td>
<td>20%</td>
</tr>
<tr>
<td>Group 3</td>
<td>5%</td>
</tr>
</tbody>
</table>
Figure 6.3: Average test risk of pooled and separately trained models on three groups (see table 6.2).

Figure 6.4: Generalization error bounds derived from original hypothesis class $\mathcal{H}$ and restricted hypothesis classes $\hat{\mathcal{H}}_i$, compared with shared model $\hat{h}$ train-test gap over 7 independent runs, with quartiles and median trend lines.
In fig. 6.3, we plot the average test risk of each group \( i \) over 7 independent runs for malfare-minimizing models \( \hat{h} \) or for risk-minimizing models \( \hat{h}_i \) as a function of total training sample size, where test risk is computed from a held-out test set with 20,000 samples for each group. We observe that pooled models almost always have lower per-group test risks than the separately-trained models \( \hat{h}_i \) on the minority groups (2 and 3), which we attribute to the regularizing effect of pooled training overcoming the small discrepancies between the data generating parameters of each group (see table 6.2). While the above describes small-sample behavior, for sufficient sample sizes, per-group models should dominate shared models, and we do observe this for group 3 with the maximum sample size of \( 32768 \cdot 0.05 \approx 1638 \).

We then compute the bounds derived from Monte-Carlo Rademacher averages (with \( \delta = 0.1 \)) over samples \( z_i \) over both \( \mathcal{H} \) and \( \hat{\mathcal{H}}_i \) for each \( i \) (fig. 6.4). Since the bounds derived from Rademacher averages over \( \mathcal{H} \) essentially function as bounds on the generalization error of the separately trained models, the fact that the bound over \( \hat{\mathcal{H}}_i \) is tighter correctly suggests that pooled training is better for the minority groups in this scenario, especially when using egalitarian training.

In the utilitarian case, we see that initially \( \hat{\mathcal{H}}_i \) bounds match \( \mathcal{H} \) bounds, but for sufficiently large sample sizes, they diverge. In the egalitarian case, \( \hat{\mathcal{H}}_i \) bounds are always better than \( \mathcal{H} \) bounds, and they appear to decay at an asymptotically greater rate (slope on the log-log plot), reaching an order of magnitude improvement in the case of the largest group (group 1). This suggests that our bounds characterize generalization error substantially more sharply than the naïve method.

### 6.4 Discussion

We show that fair learning, like multitask learning, has a regularizing effect, reducing overfitting to each group as compared to per-group models trained solely on their data. Concretely, we show that, from the perspective of each group, fair-learning (empirical malfare minimization) effectively occurs over some restricted hypothesis class, and we the bound generalization error of each group’s risk in terms of their Rademacher averages over these restricted classes. This technique yields refined generalization bounds, not just for the overall learning, task, but also for the risk of each individual group.

Such bounds are of particular importance in learning settings where minority groups often suffer poor model performance [280], such as medical ML [298] and facial recognition [57, 64]. Moreover, in critical systems, having provable guarantees on the generalization error of each task, rather than just the overall generalization error, can greatly improve reliability and user trust. This is also valuable in multi-task learning settings, where task fairness and task-specific bounds are of interest, e.g., in distributionally-robust LLMs [311].
While the contributions of this paper are theoretical, our setting is practically motivated. Understanding the generalization error of each group allows modelers to make better-informed decisions, particularly regarding minority groups. Generalization bounds for a group-specific model \( \hat{h}_i \) and a shared model \( \hat{h} \) can be used to bound risk for group \( i \), which can be used for model selection (i.e., group \( i \) can select between \( \hat{h} \) and \( \hat{h}_i \) with confidence). It is known that, given infinite data, individual models are always preferable, and the degree of suboptimality of a shared model can be bounded using transfer learning techniques; however, for data-hungry models, in particular with sparse data for minority groups, a better understanding of the interplay between generalization error and the negative impacts of majority group data on minority group performance are vital.

We also envision more sophisticated applications of our bounds. For example, if some smaller groups are more similar to minority group \( i \) than a majority group, a shared model \( \hat{h} \) optimizing, say, utilitarian malfare, may perform poorly for group \( i \), but perhaps a better-performing \( \hat{h}' \) would arise from optimizing a more egalitarian malfare function (i.e., higher \( p \) power-mean), or one that emphasizes similar groups (through the weights vector \( w \)). Group-fair learning methods can be combined with other aspects of model selection, such as feature and hyperparameter selection, where the bias-variance tradeoff plays a significant role. Our bounds indicate that we can provably learn a more complex shared model without overfitting, and our analysis enables rigorous model selection guarantees, both for individual group risks and for malfare objectives. We are hopeful that future work explores these model-search questions and other applications of our methods.
Proofs

We now show theorem 3.

**Theorem 3** (Theoretical Group-Regularized Malfare Bounds). Suppose a monotonic malfare function \( \mathcal{W} : \mathbb{R}^g \to \mathbb{R} \), hypothesis class \( \mathcal{H} \subseteq \mathcal{X} \to \mathcal{Y} \), loss function \( \ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R} \), per-group distributions \( \mathcal{D}_{1:g} \) over \( \mathcal{X} \times \mathcal{Y} \), and per-group samples \( z_{1:g} \), with \( z_j \sim \mathcal{D}_j^{m_j} \) for each group \( j \). Fix any group index \( i \), and take \( \mathcal{H}_i^* \) defined as in (6.7). The following then hold.

1. With probability at least \( 1 - 2\delta \) over choice of \( z_i \), it holds that \( \hat{h} \in \mathcal{H}_i^* \).
2. With probability at least \( 1 - 4\delta \) over choice of \( z_i \),

\[
\left| R(h, \mathcal{D}_i) - \hat{R}(\hat{h}, z_i) \right| \leq 2 \mathbb{R}_{m_i}(\ell \circ \mathcal{H}_i^*, \mathcal{D}_i) + \varepsilon_i .
\]

**Proof.** We begin by proving part 1 and then we prove part 2 as a consequence.

We now show part 1. Recall that \( \eta_i \equiv 2 \mathbb{R}_{m_i}(\ell \circ \mathcal{H}_i^*, \mathcal{D}_i) + \varepsilon_i \). With probability at least \( 1 - 2\delta \), for all \( h \in \mathcal{H} \), it holds that

\[
\left| R(h, \mathcal{D}_i) - \hat{R}(\hat{h}, z_i) \right| \leq \eta_i .
\]

This is a textbook application of McDiarmid’s bounded difference inequality, using twice the Rademacher average to bound the expected supremum deviation, i.e., the upper and lower tails of (6.3).

We could use this directly to show a weaker version of the result, however to show the stated form, we need only one tail of the above, which is used to bound generalization error of the (unknown) \( \hat{h} \), and also one tail of the simple Hoeffding’s inequality tail bound (6.2).

Now, suppose some arbitrary but fixed \( h' \) that realizes the infimum of (6.7), i.e.,

\[
\arg\min_{h' \in \mathcal{H}} \left\{ j \mapsto \begin{cases} j \neq i & \hat{R}(h', z_j) \\ j = i & R(h', \mathcal{D}_i) + \varepsilon_i \end{cases} \right\}
\]

(technically, \( h' \) may be in \( \mathcal{H} \) or a limit of a sequence of functions in \( \mathcal{H} \)). Recalling \( \varepsilon_i \equiv r \sqrt{\frac{\log \frac{1}{\delta}}{m_i}} \), we obtain by Hoeffding’s inequality that, with probability at least \( 1 - 2\delta \), it holds

\[
\left| R(h', \mathcal{D}_i) - \hat{R}(h', z_i) \right| \leq \varepsilon_i .
\]
Therefore, when these bounds hold, we have

$$
\mathcal{M} \left( j \mapsto \begin{cases} 
  j \neq i & \hat{R}(\hat{h}, z_j) \\
  j = i & \hat{R}(\hat{h}, D_i) - \eta_i 
\end{cases} \right) \leq \mathcal{M} \left( j \mapsto \hat{R}(\hat{h}, z_j) \right)
$$

By Definition

$$
= \inf_{h' \in \mathcal{H}} \mathcal{M} \left( j \mapsto \begin{cases} 
  j \neq i & \hat{R}(h', z_j) \\
  j = i & R(h', D_i) + \epsilon_i 
\end{cases} \right). \quad \text{W.h.p.: } \hat{R}(h', z_i) \leq R(h', D_i) + \epsilon_i
$$

Monotonicity of $\mathcal{M}(\cdot)$

We may thus conclude with probability at least $1 - 4\delta$ that $\hat{h} \in \mathcal{H}_i^*$ (by definition). However, observe that both the McDiarmid (Rademacher) and Hoeffding bounds required only one tail each, and thus a more careful analysis yields the guarantee with probability at least $1 - 2\delta$.

We now show part 2. By part 1, we have that $\hat{h} \in \mathcal{H}_i^*$ with probability at least $1 - 2\delta$. Then we apply the standard 2-tailed Rademacher bound with McDiarmid’s inequality over the restricted class $\mathcal{H}_i^*$, i.e., we have

$$
\mathbb{P}_{z_i \sim D_i^{m_i}} \left( \sup_{h \in \mathcal{H}_i^*} \left| R(h, D_i) - \hat{R}(\hat{h}, z_i) \right| \leq \eta_i \right) \leq 1 - 2\delta.
$$

The union bound then yields the desideratum.

We now show theorem 4.

**Theorem 4 (Empirical Group-Regularized Malfare Bounds).** Suppose as in theorem 3. The following then hold for $\hat{H}_i$ defined as in (6.8).

1. With probability at least $1 - 4\delta$ over choice of $z_i$, it holds that $\hat{h} \in \mathcal{H}_i^* \subseteq \hat{H}_i$.

2. With probability at least $1 - 6\delta$, it holds that

$$
\left| R(\hat{h}, D_i) - \hat{R}(\hat{h}, z_i) \right| \leq 2\mathbb{R}_{m_i}(\ell \circ \hat{H}_i, z_i) + 2\epsilon_i.
$$

**Proof.** We begin by proving part 1, and we then show part 2 as a consequence.

We now show part 1. First, we apply part 1 of theorem 3 (2 tails). We will then argue that

$$
\mathbb{P}_{z_i \sim D_i^{m_i}} \left( \mathcal{H}_i^* \subseteq \hat{H}_i \right) \geq 1 - 2\delta,
$$

which holds for similar reasons (a 1-tail Rademacher bound for $\mathcal{H}_i$, and a 1-tail Hoeffding bound for $\hat{h}$, both the opposite tails bounded in part 1 of theorem 3). The result then follows via union bound.
In particular, recall (6.7)

\[
\mathcal{H}_i^* = \left\{ h \in \mathcal{H} \mid \mathbb{M} \left( j \mapsto \begin{cases} j \neq i & \hat{R}(h, z_j) \\ j = i & R(h, D_i) - \eta_i \end{cases} \right) \leq \inf_{h' \in \mathcal{H}} \mathbb{M} \left( j \mapsto \begin{cases} j \neq i & \hat{R}(h', z_j) \\ j = i & R(h', D_i) + \varepsilon_i \end{cases} \right) \right\},
\]

and also (6.8)

\[
\hat{\mathcal{H}}_i = \left\{ h \in \hat{\mathcal{H}} \mid \mathbb{M} \left( j \mapsto \begin{cases} j \neq i & \hat{R}(h, z_j) \\ j = i & \hat{R}(h, z_i) - 2\tilde{\eta}_i \end{cases} \right) \leq \inf_{h' \in \mathcal{H}} \mathbb{M} \left( j \mapsto \begin{cases} j \neq i & \hat{R}(h', z_j) \\ j = i & \hat{R}(h', z_i) + 2\varepsilon_i \end{cases} \right) \right\}.
\]

Now, observe that by McDiarmid's inequality, by essentially the same argument as in (6.3), it holds that

\[
\mathbb{P}_{z \sim \mathcal{Z}^m}(\sup_{h \in \mathcal{H}} \hat{R}(h, z_i) - R(h, D_i) + \eta_i > 2\tilde{\eta}_i) = \mathbb{P}_{z \sim \mathcal{Z}^m}(\sup_{h \in \mathcal{H}} \hat{R}(h, z_i) - R(h, D_i) + 2\mathbb{R}_m(\ell \circ \mathcal{H}, D_i) > 4\mathbb{R}_m(\ell \circ \mathcal{H}, z_i) + 3\varepsilon_i) < \delta.
\]

We thus have that, with probability at least $1 - \delta$, for all $h \in \mathcal{H}_i^*$,

\[
\mathbb{M} \left( j \mapsto \begin{cases} j \neq i & \hat{R}(h, z_j) \\ j = i & \hat{R}(h, z_i) - 2\tilde{\eta}_i \end{cases} \right) \leq \mathbb{M} \left( j \mapsto \begin{cases} j \neq i & \hat{R}(h, z_j) \\ j = i & R(h, D_i) - \eta_i \end{cases} \right),
\]

and similarly, with probability at least $1 - \delta$ by the Hoeffding bound (6.2) on $h'$, we have

\[
\inf_{h' \in \mathcal{H}} \mathbb{M} \left( j \mapsto \begin{cases} j \neq i & \hat{R}(h', z_j) \\ j = i & R(h', D_i) + \varepsilon_i \end{cases} \right) \leq \inf_{h' \in \mathcal{H}} \mathbb{M} \left( j \mapsto \begin{cases} j \neq i & \hat{R}(h', z_j) \\ j = i & \hat{R}(h', z_i) + 2\varepsilon_i \end{cases} \right),
\]

where both steps apply monotonicity of $\mathbb{M}(\cdot)$.

From this, we may conclude that, with probability at least $1 - 2\delta$, for each $h \in \mathcal{H}$, if the constraint in (6.7) is satisfied, then the constraint (6.8) is satisfied, thus $\mathcal{H}_i^* \subseteq \hat{\mathcal{H}}_i$. The union bound over all tail bounds above then yields part 1.

We now show part 2. This result essentially follows the structure of part 2 of theorem 3. However, we now start with part 1 above, which allows us to conclude that $\hat{h} \in \hat{\mathcal{H}}_i$ with probability at least $1 - 4\delta$, and then apply the standard empirical Rademacher bounds, i.e., 2 tails of (6.3) (we require only the upper and lower bounds to the supremum deviation, not the bound on the Rademacher average itself), to $\hat{\mathcal{H}}_i$ (rather than to $\mathcal{H}_i^*$). Taking the union bound over all events then yields the desideratum.

We now show corollary 3.
Corollary 3 (Empirical Malfare Generalization Bounds). Suppose as in theorem 4. Suppose also that there exists some $\lambda > 0$ and norm $\| \cdot \|_{\mathcal{M}}$ such that $\mathcal{M}(\cdot)$ is $\lambda \| \cdot \|_{\mathcal{M}}$ Lipschitz continuous, i.e., $\forall S, S': \mathcal{M}(S + S') \leq \mathcal{M}(S) + \lambda \| S' \|_{\mathcal{M}}$. We then have:

1. With probability at least $1 - 5g\delta$, the true malfare of $\hat{h}$ is bounded by

$$
\mathcal{M} \left( j \mapsto R(\hat{h}, D_j) \right)
\leq \mathcal{M} \left( j \mapsto \hat{R}(\hat{h}, z_j) + 2\hat{k}_{m_j}(\hat{h}_j, z_j) + 2\varepsilon_j \right)
\leq \mathcal{M} \left( j \mapsto \hat{R}(\hat{h}, z_j) \right) + \lambda \left\| j \mapsto 2\hat{k}_{m_j}(\hat{h}_j, z_j) + 2\varepsilon_j \right\|_{\mathcal{M}} .
$$

2. With probability at least $1 - 6g\delta$, we bound the suboptimality of $\hat{h}$ as

$$
\mathcal{M} \left( j \mapsto R(\hat{h}, D_j) \right)
\leq \mathcal{M} \left( j \mapsto R(h^*, D_j) + 2\hat{k}_{m_j}(\hat{h}_j, z_j) + 3\varepsilon_j \right)
= \mathcal{M} \left( j \mapsto R(h^*, D_j) \right) - \mathcal{M} \left( j \mapsto R(\hat{h}, D_j) \right)
\leq \lambda \left\| j \mapsto 2\hat{k}_{m_j}(\hat{h}_j, z_j) + 3\varepsilon_j \right\|_{\mathcal{M}} .
$$

Proof. For both results, we apply part 2 of theorem 4 to each group $i$, which by union bound gives a result with probability at least $1 - 6g\delta$. However, careful accounting reveals that we only require one tail of the final Rademacher bound of theorem 4 part 2, i.e., we require $R(\hat{h}, D_j) \leq \hat{R}(\hat{h}, z_j) + 2\hat{k}_{m_j}(\hat{h}_j, z_j) + 2\varepsilon_j$, but not $\hat{R}(\hat{h}, z_j) \leq R(\hat{h}, D_j) + 2\hat{k}_{m_j}(\hat{h}_j, z_j) + 2\varepsilon_j$, thus we begin with tail bounds that hold with probability at least $1 - 5g\delta$.

We now show part 1. Subject to all tail bounds holding, we have for all $j \in 1, \ldots, g$ that $R(\hat{h}, D_j) \leq \hat{R}(\hat{h}, z_j) + 2\hat{k}_{m_j}(\hat{h}_j, z_j) + 2\varepsilon_j$, thus by monotonicity of $\mathcal{M}(\cdot)$, we have

$$
\mathcal{M} \left( j \mapsto R(\hat{h}, D_j) \right) \leq \mathcal{M} \left( j \mapsto \hat{R}(\hat{h}, z_j) + 2\hat{k}_{m_j}(\hat{h}_j, z_j) + 2\varepsilon_j \right) .
$$

Applying the Lipschitz property then yields the final portion of part 1.

We now show part 2. First, observe that $\mathcal{M} \left( j \mapsto R(h^*, D_j) \right) \leq \mathcal{M} \left( j \mapsto R(\hat{h}, D_j) \right)$ by definition. For the remaining inequality, we introduce one new tail bound for each group $j$, in particular, a 1-tail Hoeffding bound of

$$
\mathbb{P}_{z_j \sim D^m_j} \left( R(h^*, D_j) \leq \hat{R}(h^*, z_j) + \varepsilon_j \right) \geq 1 - \delta .
$$

This seems familiar, but it is not quite the same as the 2-tail Hoeffding bound on each $R(h^*, D_j)$ used by theorems 3 and 4, thus this tail bound must be counted separately. Now, we substitute into the result of part 1, again applying monotonicity, to get

$$
\mathcal{M} \left( j \mapsto R(\hat{h}, D_j) \right) \leq \mathcal{M} \left( j \mapsto R(h^*, D_j) + 2\hat{k}_{m_j}(\hat{h}_j, z_j) + 3\varepsilon_j \right) .
$$
Applying the Lipschitz property then yields the final portion of part 2. By union bound, we may conclude the result with probability at least $1 - 6g\delta$. □

We now show lemma 4.

**Lemma 4** (Convex Optimization for Monte-Carlo Rademacher Averages). Suppose the parameter space $B$ of $H$ is a convex set, loss $f(h_\beta(x), y)$ is convex in $\beta \in B$ for all $x \in X$, $y \in Y$, and malfare $M(\cdot) : \mathbb{R}^g \to \mathbb{R}$ is quasiconvex and monotonically increasing in each argument. Then the parameter spaces of $\hat{H}_i$ and $\hat{H}^*_i$ are convex sets.

Moreover, if $g \circ H$ is an affine function family, then $\hat{R}^n_m(g \circ \hat{H}_i, z; \sigma)$ reduces to maximizing a linear function over a convex set. Similarly, if we strengthen the quasiconvexity assumption on $M(\cdot)$ to convexity, then EMM reduces to minimizing a convex objective over the convex set $B$.

**Proof.** We first show that the restricted parameter spaces of $\hat{H}^*_i$ and $\hat{H}_i$ are convex sets.

The crux of this result is to show that $M(j \mapsto f_j(\beta))$ is quasiconvex, where $f_j(x)$ represents $\hat{R}(h_\beta, z_j) - c_j$ or $R(h_\beta, D_j) - c_j$ for some constant $c \in \mathbb{R}^g$. This indeed holds, so long as $f_j(\beta)$ is quasiconvex. First note that convexity of loss immediately implies convexity of (empirical) risk. Now, by standard compositional rules, since we assume $M(\cdot)$ to be quasiconvex and monotonic, we conclude that $M(j \mapsto f_j(\beta))$ is quasiconvex in $\beta \in B$.

Now, converting $H$ to $B$, observe that the parameter spaces associated with both $\hat{H}^*_i$ in (6.7)

$$
\left\{ \beta \in B \mid M \left( j \mapsto \begin{cases} j \neq i \quad \hat{R}(h_\beta, z_j) \\ j = i \quad \hat{R}(h_\beta, D_i) - \eta_i \end{cases} \right) \leq \inf_{h' \in \hat{H}} M \left( j \mapsto \begin{cases} j \neq i \quad \hat{R}(h', z_j) \\ j = i \quad \hat{R}(h', D_i) + \epsilon_i \end{cases} \right) \right\},
$$

and also $\hat{H}_i$ in (6.8)

$$
\left\{ \beta \in B \mid M \left( j \mapsto \begin{cases} j \neq i \quad \hat{R}(h_\beta, z_j) \\ j = i \quad \hat{R}(h_\beta, z_i) - 2\eta_i \end{cases} \right) \leq \inf_{h' \in \hat{H}} M \left( j \mapsto \begin{cases} j \neq i \quad \hat{R}(h', z_j) \\ j = i \quad \hat{R}(h', z_i) + 2\epsilon_i \end{cases} \right) \right\},
$$

are subsets of the convex set $B$. In particular, the RHS of the condition is constant in $\beta$, and as above, the LHS is quasiconvex in $\beta$, thus both restricted parameter spaces are convex sets.

Now, note that once we determine the parameter space to be convex, Monte-Carlo Rademacher averages can be efficiently computed via standard convex optimization techniques, e.g., first-order methods to maximize a linear objective on a convex set. Key to this observation is that we assumed $g \circ H$ to be an affine function family, thus even after multiplying terms by $\pm 1$ in the Monte-Carlo Rademacher average (6.11), the objective of the supremum remains convex.

Finally, observe that if $M(\cdot)$ is convex and monotonically increasing, then the EMM objective is also convex. This follows from standard compositional rules, see discussion following Boyd and
Vandenberghe [50] equation (3.15). EMM then reduces to minimizing a convex function on a convex set.
Bibliography


[8] ACLU and other organizations. Re: Request for information and comment on financial institutions’ use of artificial intelligence, including machine


[127] Alex Engler. Independent auditors are struggling to hold AI companies accountable. fastcompany, 2021.


[156] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making.


132


[284] Andrew Michaelson, Brian Thavarajah, and Margaret McPherson. A revived disparate impact doctrine under Biden’s CFPB.


[345] Restatement (Third) of Torts: Products Liability § 3.


[362] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. Hidden technical


[381] Jakub Sliwinski, Martin Strobel, and Yair Zick. Axiomatic characterization of data-driven


