

Transparency and Interpretability

Responsible Data Science
DS-UA 202 and DS-GA 1017

Instructors: Julia Stoyanovich and George Wood

This reader contains links to online materials and excerpts from selected articles on transparency and interpretability. For convenience, the readings are organized by course week. Please note that some excerpts end in the middle of a section. Where that is the case, the partial section is not required reading.

Week 10: Auditing black-box models	3
Tulio Ribeiro, Singh, and Guestrin (2016) “Why should I trust you? Explaining the predictions of any classifier” <i>ACM KDD 2016: 1135-1144</i>	4
Datta, Sen, and Zick (2016) “Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems,” <i>IEEE SP 2016:598-617</i>	14
Lundberg and Lee (2017) “A unified approach to interpreting model predictions,” <i>NIPS 2017:4765-4774</i>	34
Week 11: Discrimination in on-line ad delivery	44
Sweeney (2013) “Discrimination in online ad delivery” <i>CACM 2013 56(5): 44-54</i>	45
Datta, Tschantz, and Datta(2015) “Automated experiments on ad privacy settings” <i>PoPETs 2015 (1): 92-112</i>	56
Ali, Sapiezynski, Bogen, Korolova, Mislove, Rieke (2019) “Discrimination through optimization: How Facebook’s ad delivery can lead to biased outcomes” <i>ACM CSCW 2019</i>	77

Weeks 12, 13: Interpretability	107
Selbst and Barocas (2018) “The intuitive appeal of explainable machines” <i>Fordham Law Review</i> 2018 (87): 1085-1139	108
Stoyanovich, Van Bavel, West (2020) “The imperative of interpretable ma- chines” <i>Nature Machine Intelligence</i> :	163
Stoyanovich and Howe (2019) “Nutritional labels for data and models” <i>IEEE Data Engineering Bulletin</i> : 42(3): 13-23	166

Week 10: Auditing black-box models

“Why Should I Trust You?”

Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu

ABSTRACT

Despite widespread adoption, machine learning models remain mostly black boxes. Understanding the reasons behind predictions is, however, quite important in assessing *trust*, which is fundamental if one plans to take action based on a prediction, or when choosing whether to deploy a new model. Such understanding also provides insights into the model, which can be used to transform an untrustworthy model or prediction into a trustworthy one.

In this work, we propose LIME, a novel explanation technique that explains the predictions of *any* classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction. We also propose a method to explain models by presenting representative individual predictions and their explanations in a non-redundant way, framing the task as a submodular optimization problem. We demonstrate the flexibility of these methods by explaining different models for text (e.g. random forests) and image classification (e.g. neural networks). We show the utility of explanations via novel experiments, both simulated and with human subjects, on various scenarios that require trust: deciding if one should trust a prediction, choosing between models, improving an untrustworthy classifier, and identifying why a classifier should not be trusted.

1. INTRODUCTION

Machine learning is at the core of many recent advances in science and technology. Unfortunately, the important role of humans is an oft-overlooked aspect in the field. Whether humans are directly using machine learning classifiers as tools, or are deploying models within other products, a vital concern remains: *if the users do not trust a model or a prediction, they will not use it*. It is important to differentiate between two different (but related) definitions of trust: (1) *trusting a prediction*, i.e. whether a user trusts an individual prediction sufficiently to take some action based on it, and (2) *trusting a model*, i.e. whether the user trusts a model to behave in reasonable ways if deployed. Both are directly impacted by

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD 2016 San Francisco, CA, USA

© 2016 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ISBN 978-1-4503-4232-2/16/08...\$15.00

DOI: <http://dx.doi.org/10.1145/2939672.2939778>

how much the human understands a model’s behaviour, as opposed to seeing it as a black box.

Determining trust in individual predictions is an important problem when the model is used for decision making. When using machine learning for medical diagnosis [6] or terrorism detection, for example, predictions cannot be acted upon on blind faith, as the consequences may be catastrophic.

Apart from trusting individual predictions, there is also a need to evaluate the model as a whole before deploying it “in the wild”. To make this decision, users need to be confident that the model will perform well on real-world data, according to the metrics of interest. Currently, models are evaluated using accuracy metrics on an available validation dataset. However, real-world data is often significantly different, and further, the evaluation metric may not be indicative of the product’s goal. Inspecting individual predictions and their explanations is a worthwhile solution, in addition to such metrics. In this case, it is important to aid users by suggesting which instances to inspect, especially for large datasets.

In this paper, we propose providing explanations for individual predictions as a solution to the “trusting a prediction” problem, and selecting multiple such predictions (and explanations) as a solution to the “trusting the model” problem. Our main contributions are summarized as follows.

- LIME, an algorithm that can explain the predictions of *any* classifier or regressor in a faithful way, by approximating it locally with an interpretable model.
- SP-LIME, a method that selects a set of representative instances with explanations to address the “trusting the model” problem, via submodular optimization.
- Comprehensive evaluation with simulated and human subjects, where we measure the impact of explanations on trust and associated tasks. In our experiments, non-experts using LIME are able to pick which classifier from a pair generalizes better in the real world. Further, they are able to greatly improve an untrustworthy classifier trained on 20 newsgroups, by doing feature engineering using LIME. We also show how understanding the predictions of a neural network on images helps practitioners know when and why they should not trust a model.

2. THE CASE FOR EXPLANATIONS

By “explaining a prediction”, we mean presenting textual or visual artifacts that provide qualitative understanding of the relationship between the instance’s components (e.g. words in text, patches in an image) and the model’s prediction. We

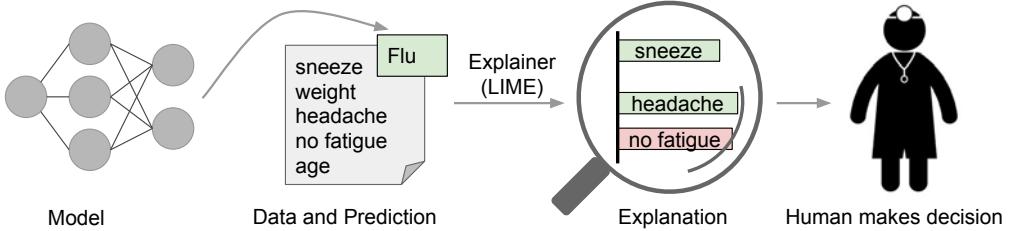


Figure 1: Explaining individual predictions. A model predicts that a patient has the flu, and LIME highlights the symptoms in the patient’s history that led to the prediction. Sneeze and headache are portrayed as contributing to the “flu” prediction, while “no fatigue” is evidence against it. With these, a doctor can make an informed decision about whether to trust the model’s prediction.

argue that explaining predictions is an important aspect in getting humans to trust and use machine learning effectively, if the explanations are faithful and intelligible.

The process of explaining individual predictions is illustrated in Figure 1. It is clear that a doctor is much better positioned to make a decision with the help of a model if intelligible explanations are provided. In this case, an explanation is a small list of symptoms with relative weights – symptoms that either contribute to the prediction (in green) or are evidence against it (in red). Humans usually have prior knowledge about the application domain, which they can use to accept (trust) or reject a prediction if they understand the reasoning behind it. It has been observed, for example, that providing explanations can increase the acceptance of movie recommendations [12] and other automated systems [8].

Every machine learning application also requires a certain measure of overall trust in the model. Development and evaluation of a classification model often consists of collecting annotated data, of which a held-out subset is used for automated evaluation. Although this is a useful pipeline for many applications, evaluation on validation data may not correspond to performance “in the wild”, as practitioners often overestimate the accuracy of their models [21], and thus trust cannot rely solely on it. Looking at examples offers an alternative method to assess truth in the model, especially if the examples are explained. We thus propose explaining several representative individual predictions of a model as a way to provide a global understanding.

There are several ways a model or its evaluation can go wrong. Data leakage, for example, defined as the unintentional leakage of signal into the training (and validation) data that would not appear when deployed [14], potentially increases accuracy. A challenging example cited by (**author?**) [14] is one where the patient ID was found to be heavily correlated with the target class in the training and validation data. This issue would be incredibly challenging to identify just by observing the predictions and the raw data, but much easier if explanations such as the one in Figure 1 are provided, as patient ID would be listed as an explanation for predictions. Another particularly hard to detect problem is dataset shift [5], where training data is different than test data (we give an example in the famous 20 newsgroups dataset later on). The insights given by explanations are particularly helpful in identifying what must be done to convert an untrustworthy model into a trustworthy one – for example, removing leaked data or changing the training data to avoid dataset shift.

Machine learning practitioners often have to select a model from a number of alternatives, requiring them to assess the relative trust between two or more models. In Figure

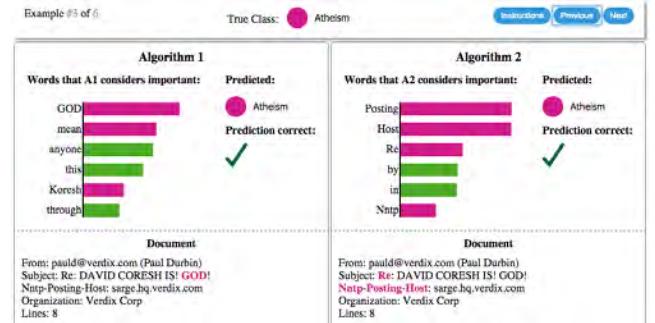


Figure 2: Explaining individual predictions of competing classifiers trying to determine if a document is about “Christianity” or “Atheism”. The bar chart represents the importance given to the most relevant words, also highlighted in the text. Color indicates which class the word contributes to (green for “Christianity”, magenta for “Atheism”).

2, we show how individual prediction explanations can be used to select between models, in conjunction with accuracy. In this case, the algorithm with higher accuracy on the validation set is actually much worse, a fact that is easy to see when explanations are provided (again, due to human prior knowledge), but hard otherwise. Further, there is frequently a mismatch between the metrics that we can compute and optimize (e.g. accuracy) and the actual metrics of interest such as user engagement and retention. While we may not be able to measure such metrics, we have knowledge about how certain model behaviors can influence them. Therefore, a practitioner may wish to choose a less accurate model for content recommendation that does not place high importance in features related to “clickbait” articles (which may hurt user retention), even if exploiting such features increases the accuracy of the model in cross validation. We note that explanations are particularly useful in these (and other) scenarios if a method can produce them for *any* model, so that a variety of models can be compared.

Desired Characteristics for Explainers

We now outline a number of desired characteristics from explanation methods.

An essential criterion for explanations is that they must be **interpretable**, i.e., provide qualitative understanding between the input variables and the response. We note that interpretability must take into account the user’s limitations. Thus, a linear model [24], a gradient vector [2] or an additive model [6] may or may not be interpretable. For example, if

hundreds or thousands of features significantly contribute to a prediction, it is not reasonable to expect any user to comprehend why the prediction was made, even if individual weights can be inspected. This requirement further implies that explanations should be easy to understand, which is not necessarily true of the features used by the model, and thus the “input variables” in the explanations may need to be different than the features. Finally, we note that the notion of interpretability also depends on the target audience. Machine learning practitioners may be able to interpret small Bayesian networks, but laymen may be more comfortable with a small number of weighted features as an explanation.

Another essential criterion is **local fidelity**. Although it is often impossible for an explanation to be completely faithful unless it is the complete description of the model itself, for an explanation to be meaningful it must at least be *locally faithful*, i.e. it must correspond to how the model behaves in the vicinity of the instance being predicted. We note that local fidelity does not imply global fidelity: features that are globally important may not be important in the local context, and vice versa. While global fidelity would imply local fidelity, identifying globally faithful explanations that are interpretable remains a challenge for complex models.

While there are models that are inherently interpretable [6, 17, 26, 27], an explainer should be able to explain *any* model, and thus be **model-agnostic** (i.e. treat the original model as a black box). Apart from the fact that many state-of-the-art classifiers are not currently interpretable, this also provides flexibility to explain future classifiers.

In addition to explaining predictions, providing a **global perspective** is important to ascertain trust in the model. As mentioned before, accuracy may often not be a suitable metric to evaluate the model, and thus we want to *explain the model*. Building upon the explanations for individual predictions, we select a few explanations to present to the user, such that they are representative of the model.

3. LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS

We now present Local Interpretable Model-agnostic Explanations (**LIME**). The overall goal of LIME is to identify an **interpretable** model over the *interpretable representation* that is **locally faithful** to the classifier.

3.1 Interpretable Data Representations

Before we present the explanation system, it is important to distinguish between features and interpretable data representations. As mentioned before, **interpretable** explanations need to use a representation that is understandable to humans, regardless of the actual features used by the model. For example, a possible *interpretable representation* for text classification is a binary vector indicating the presence or absence of a word, even though the classifier may use more complex (and incomprehensible) features such as word embeddings. Likewise for image classification, an *interpretable representation* may be a binary vector indicating the “presence” or “absence” of a contiguous patch of similar pixels (a super-pixel), while the classifier may represent the image as a tensor with three color channels per pixel. We denote $x \in \mathbb{R}^d$ be the original representation of an instance being explained, and we use $x' \in \{0, 1\}^{d'}$ to denote a binary vector for its interpretable representation.

3.2 Fidelity-Interpretability Trade-off

Formally, we define an explanation as a model $g \in G$, where G is a class of potentially *interpretable* models, such as linear models, decision trees, or falling rule lists [27], i.e. a model $g \in G$ can be readily presented to the user with visual or textual artifacts. The domain of g is $\{0, 1\}^{d'}$, i.e. g acts over absence/presence of the *interpretable components*. As not every $g \in G$ may be simple enough to be interpretable - thus we let $\Omega(g)$ be a measure of *complexity* (as opposed to *interpretability*) of the explanation $g \in G$. For example, for decision trees $\Omega(g)$ may be the depth of the tree, while for linear models, $\Omega(g)$ may be the number of non-zero weights.

Let the model being explained be denoted $f : \mathbb{R}^d \rightarrow \mathbb{R}$. In classification, $f(x)$ is the probability (or a binary indicator) that x belongs to a certain class¹. We further use $\pi_x(z)$ as a proximity measure between an instance z to x , so as to define locality around x . Finally, let $\mathcal{L}(f, g, \pi_x)$ be a measure of how unfaithful g is in approximating f in the locality defined by π_x . In order to ensure both **interpretability** and **local fidelity**, we must minimize $\mathcal{L}(f, g, \pi_x)$ while having $\Omega(g)$ be low enough to be interpretable by humans. The explanation produced by **LIME** is obtained by the following:

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \quad \mathcal{L}(f, g, \pi_x) + \Omega(g) \quad (1)$$

This formulation can be used with different explanation families G , fidelity functions \mathcal{L} , and complexity measures Ω . Here we focus on sparse linear models as explanations, and on performing the search using perturbations.

3.3 Sampling for Local Exploration

We want to minimize the locality-aware loss $\mathcal{L}(f, g, \pi_x)$ without making any assumptions about f , since we want the explainer to be **model-agnostic**. Thus, in order to learn the local behavior of f as the interpretable inputs vary, we approximate $\mathcal{L}(f, g, \pi_x)$ by drawing samples, weighted by π_x . We sample instances around x' by drawing nonzero elements of x' uniformly at random (where the number of such draws is also uniformly sampled). Given a perturbed sample $z' \in \{0, 1\}^{d'}$ (which contains a fraction of the nonzero elements of x'), we recover the sample in the original representation $z \in \mathbb{R}^d$ and obtain $f(z)$, which is used as a *label* for the explanation model. Given this dataset \mathcal{Z} of perturbed samples with the associated labels, we optimize Eq. (1) to get an explanation $\xi(x)$. The primary intuition behind LIME is presented in Figure 3, where we sample instances both in the vicinity of x (which have a high weight due to π_x) and far away from x (low weight from π_x). Even though the original model may be too complex to explain globally, LIME presents an explanation that is locally faithful (linear in this case), where the locality is captured by π_x . It is worth noting that our method is fairly robust to sampling noise since the samples are weighted by π_x in Eq. (1). We now present a concrete instance of this general framework.

3.4 Sparse Linear Explanations

For the rest of this paper, we let G be the class of linear models, such that $g(z') = w_g \cdot z'$. We use the locally weighted square loss as \mathcal{L} , as defined in Eq. (2), where we let $\pi_x(z) = \exp(-D(x, z)^2 / \sigma^2)$ be an exponential kernel defined on some

¹For multiple classes, we explain each class separately, thus $f(x)$ is the prediction of the relevant class.

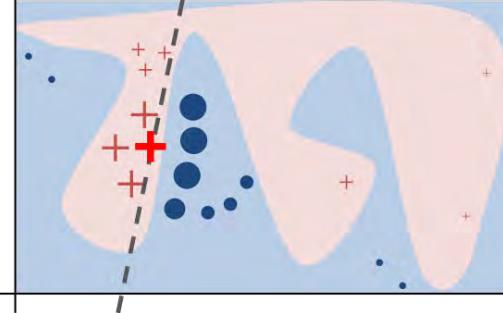


Figure 3: Toy example to present intuition for LIME. The black-box model’s complex decision function f (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using f , and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.

distance function D (e.g. cosine distance for text, L_2 distance for images) with width σ .

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) (f(z) - g(z'))^2 \quad (2)$$

For text classification, we ensure that the explanation is **interpretable** by letting the *interpretable representation* be a bag of words, and by setting a limit K on the number of words, i.e. $\Omega(g) = \infty \mathbb{1}[\|w_g\|_0 > K]$. Potentially, K can be adapted to be as big as the user can handle, or we could have different values of K for different instances. In this paper we use a constant value for K , leaving the exploration of different values to future work. We use the same Ω for image classification, using “super-pixels” (computed using any standard algorithm) instead of words, such that the interpretable representation of an image is a binary vector where 1 indicates the original super-pixel and 0 indicates a grayed out super-pixel. This particular choice of Ω makes directly solving Eq. (1) intractable, but we approximate it by first selecting K features with Lasso (using the regularization path [9]) and then learning the weights via least squares (a procedure we call K-LASSO in Algorithm 1). Since Algorithm 1 produces an explanation for an individual prediction, its complexity does not depend on the size of the dataset, but instead on time to compute $f(x)$ and on the number of samples N . In practice, explaining random forests with 1000 trees using scikit-learn (<http://scikit-learn.org>) on a laptop with $N = 5000$ takes under 3 seconds without any optimizations such as using gpus or parallelization. Explaining each prediction of the Inception network [25] for image classification takes around 10 minutes.

Any choice of interpretable representations and G will have some inherent drawbacks. First, while the underlying model can be treated as a black-box, certain interpretable representations will not be powerful enough to explain certain behaviors. For example, a model that predicts sepia-toned images to be *retro* cannot be explained by presence of absence of super pixels. Second, our choice of G (sparse linear models) means that if the underlying model is highly non-linear even in the locality of the prediction, there may not be a faithful explanation. However, we can estimate the faithfulness of

Algorithm 1 Sparse Linear Explanations using LIME

```

Require: Classifier  $f$ , Number of samples  $N$ 
Require: Instance  $x$ , and its interpretable version  $x'$ 
Require: Similarity kernel  $\pi_x$ , Length of explanation  $K$ 

 $\mathcal{Z} \leftarrow \{\}$ 
for  $i \in \{1, 2, 3, \dots, N\}$  do
     $z'_i \leftarrow \text{sample\_around}(x')$ 
     $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$ 
end for
 $w \leftarrow \text{K-Lasso}(\mathcal{Z}, K)$   $\triangleright$  with  $z'_i$  as features,  $f(z)$  as target
return  $w$ 

```

the explanation on \mathcal{Z} , and present this information to the user. This estimate of faithfulness can also be used for selecting an appropriate family of explanations from a set of multiple interpretable model classes, thus adapting to the given dataset and the classifier. We leave such exploration for future work, as linear explanations work quite well for multiple black-box models in our experiments.

3.5 Example 1: Text classification with SVMs

In Figure 2 (right side), we explain the predictions of a support vector machine with RBF kernel trained on unigrams to differentiate “Christianity” from “Atheism” (on a subset of the 20 newsgroup dataset). Although this classifier achieves 94% held-out accuracy, and one would be tempted to trust it based on this, the explanation for an instance shows that predictions are made for quite arbitrary reasons (words “Posting”, “Host”, and “Re” have no connection to either Christianity or Atheism). The word “Posting” appears in 22% of examples in the training set, 99% of them in the class “Atheism”. Even if headers are removed, proper names of prolific posters in the original newsgroups are selected by the classifier, which would also not generalize.

After getting such insights from explanations, it is clear that this dataset has serious issues (which are not evident just by studying the raw data or predictions), and that this classifier, or held-out evaluation, cannot be trusted. It is also clear what the problems are, and the steps that can be taken to fix these issues and train a more trustworthy classifier.

3.6 Example 2: Deep networks for images

When using sparse linear explanations for image classifiers, one may wish to just highlight the super-pixels with positive weight towards a specific class, as they give intuition as to why the model would think that class may be present. We explain the prediction of Google’s pre-trained Inception neural network [25] in this fashion on an arbitrary image (Figure 4a). Figures 4b, 4c, 4d show the superpixels explanations for the top 3 predicted classes (with the rest of the image grayed out), having set $K = 10$. What the neural network picks up on for each of the classes is quite natural to humans - Figure 4b in particular provides insight as to why acoustic guitar was predicted to be electric: due to the fretboard. This kind of explanation enhances trust in the classifier (even if the top predicted class is wrong), as it shows that it is not acting in an unreasonable manner.

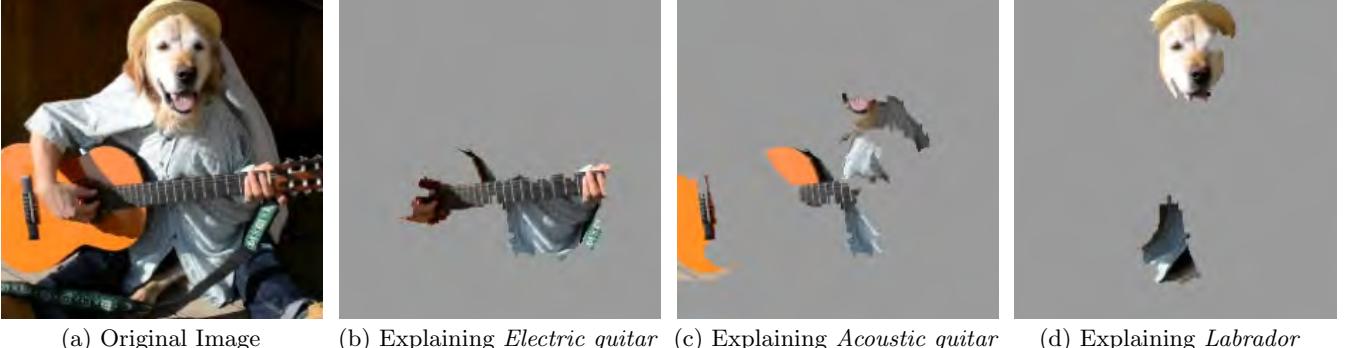


Figure 4: Explaining an image classification prediction made by Google’s Inception neural network. The top 3 classes predicted are “Electric Guitar” ($p = 0.32$), “Acoustic guitar” ($p = 0.24$) and “Labrador” ($p = 0.21$)

4. SUBMODULAR PICK FOR EXPLAINING MODELS

Although an explanation of a single prediction provides some understanding into the reliability of the classifier to the user, it is not sufficient to evaluate and assess trust in the model as a whole. We propose to give a global understanding of the model by explaining a set of individual instances. This approach is still model agnostic, and is complementary to computing summary statistics such as held-out accuracy.

Even though explanations of multiple instances can be insightful, these instances need to be selected judiciously, since users may not have the time to examine a large number of explanations. We represent the time/patience that humans have by a budget B that denotes the number of explanations they are willing to look at in order to understand a model. Given a set of instances X , we define the **pick step** as the task of selecting B instances for the user to inspect.

The pick step is not dependent on the existence of explanations - one of the main purpose of tools like Modeltracker [1] and others [11] is to assist users in selecting instances themselves, and examining the raw data and predictions. However, since looking at raw data is not enough to understand predictions and get insights, the pick step should take into account the explanations that accompany each prediction. Moreover, this method should pick a diverse, representative set of explanations to show the user – i.e. non-redundant explanations that represent how the model behaves globally.

Given the explanations for a set of instances X ($|X| = n$), we construct an $n \times d'$ *explanation matrix* \mathcal{W} that represents the local importance of the interpretable components for each instance. When using linear models as explanations, for an instance x_i and explanation $g_i = \xi(x_i)$, we set $\mathcal{W}_{ij} = |w_{g_{ij}}|$. Further, for each component (column) j in \mathcal{W} , we let I_j denote the *global* importance of that component in the explanation space. Intuitively, we want I such that features that explain many different instances have higher importance scores. In Figure 5, we show a toy example \mathcal{W} , with $n = d' = 5$, where \mathcal{W} is binary (for simplicity). The importance function I should score feature f_2 higher than feature f_1 , i.e. $I_2 > I_1$, since feature f_2 is used to explain more instances. Concretely for the text applications, we set $I_j = \sqrt{\sum_{i=1}^n \mathcal{W}_{ij}}$. For images, I must measure something that is comparable across the super-pixels in different images,

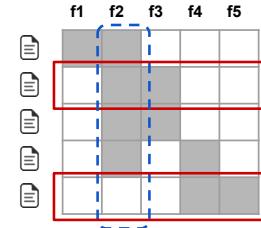


Figure 5: Toy example \mathcal{W} . Rows represent instances (documents) and columns represent features (words). Feature f_2 (dotted blue) has the highest importance. Rows 2 and 5 (in red) would be selected by the pick procedure, covering all but feature f_1 .

Algorithm 2 Submodular pick (SP) algorithm

Require: Instances X , Budget B

```

for all  $x_i \in X$  do
     $\mathcal{W}_i \leftarrow \text{explain}(x_i, x'_i)$             $\triangleright$  Using Algorithm 1
end for
for  $j \in \{1 \dots d'\}$  do
     $I_j \leftarrow \sqrt{\sum_{i=1}^n |\mathcal{W}_{ij}|}$     $\triangleright$  Compute feature importances
end for
 $V \leftarrow \{\}$ 
while  $|V| < B$  do            $\triangleright$  Greedy optimization of Eq (4)
     $V \leftarrow V \cup \arg\max_i c(V \cup \{i\}, \mathcal{W}, I)$ 
end while
return  $V$ 

```

such as color histograms or other features of super-pixels; we leave further exploration of these ideas for future work.

While we want to pick instances that cover the important components, the set of explanations must not be redundant in the components they show the users, i.e. avoid selecting instances with similar explanations. In Figure 5, after the second row is picked, the third row adds no value, as the user has already seen features f_2 and f_3 - while the last row exposes the user to completely new features. Selecting the second and last row results in the coverage of almost all the features. We formalize this non-redundant coverage intuition in Eq. (3), where we define coverage as the set function c that, given \mathcal{W} and I , computes the total importance of the features that appear in at least one instance in a set V .

$$c(V, \mathcal{W}, I) = \sum_{j=1}^{d'} \mathbb{1}_{[\exists i \in V : w_{ij} > 0]} I_j \quad (3)$$

The pick problem, defined in Eq. (4), consists of finding the set $V, |V| \leq B$ that achieves highest coverage.

$$\text{Pick}(\mathcal{W}, I) = \underset{V, |V| \leq B}{\operatorname{argmax}} c(V, \mathcal{W}, I) \quad (4)$$

The problem in Eq. (4) is maximizing a weighted coverage function, and is NP-hard [10]. Let $c(V \cup \{i\}, \mathcal{W}, I) - c(V, \mathcal{W}, I)$ be the marginal coverage gain of adding an instance i to a set V . Due to submodularity, a greedy algorithm that iteratively adds the instance with the highest marginal coverage gain to the solution offers a constant-factor approximation guarantee of $1 - 1/e$ to the optimum [15]. We outline this approximation in Algorithm 2, and call it **submodular pick**.

5. SIMULATED USER EXPERIMENTS

In this section, we present simulated user experiments to evaluate the utility of explanations in trust-related tasks. In particular, we address the following questions: (1) Are the explanations faithful to the model, (2) Can the explanations aid users in ascertaining trust in predictions, and (3) Are the explanations useful for evaluating the model as a whole. Code and data for replicating our experiments are available at <https://github.com/marcotcr/lime-experiments>.

5.1 Experiment Setup

We use two sentiment analysis datasets (*books* and *DVDs*, 2000 instances each) where the task is to classify product reviews as positive or negative [4]. We train decision trees (**DT**), logistic regression with L2 regularization (**LR**), nearest neighbors (**NN**), and support vector machines with RBF kernel (**SVM**), all using bag of words as features. We also include random forests (with 1000 trees) trained with the average word2vec embedding [19] (**RF**), a model that is impossible to interpret without a technique like LIME. We use the implementations and default parameters of scikit-learn, unless noted otherwise. We divide each dataset into train (1600 instances) and test (400 instances).

To explain individual predictions, we compare our proposed approach (**LIME**), with **parzen** [2], a method that approximates the black box classifier globally with Parzen windows, and explains individual predictions by taking the gradient of the prediction probability function. For parzen, we take the K features with the highest absolute gradients as explanations. We set the hyper-parameters for parzen and LIME using cross validation, and set $N = 15,000$. We also compare against a **greedy** procedure (similar to **(author?)** [18]) in which we greedily remove features that contribute the most to the predicted class until the prediction changes (or we reach the maximum of K features), and a **random** procedure that randomly picks K features as an explanation. We set K to 10 for our experiments.

For experiments where the pick procedure applies, we either do random selection (random pick, **RP**) or the procedure described in §4 (submodular pick, **SP**). We refer to pick-explainer combinations by adding RP or SP as a prefix.

5.2 Are explanations faithful to the model?

We measure faithfulness of explanations on classifiers that are by themselves interpretable (sparse logistic regression

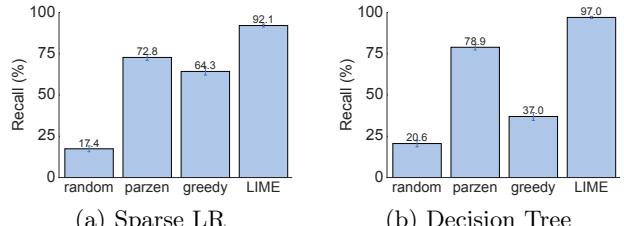


Figure 6: Recall on truly important features for two interpretable classifiers on the books dataset.

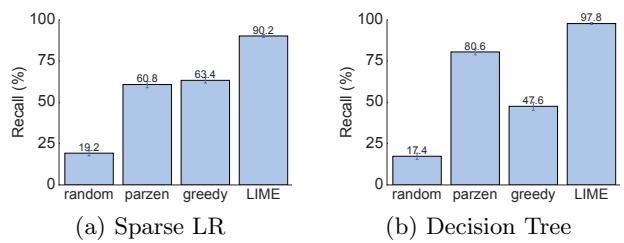


Figure 7: Recall on truly important features for two interpretable classifiers on the DVDs dataset.

and decision trees). In particular, we train both classifiers such that the maximum number of features they use for any instance is 10, and thus we know the *gold* set of features that the are considered important by these models. For each prediction on the test set, we generate explanations and compute the fraction of these *gold* features that are recovered by the explanations. We report this recall averaged over all the test instances in Figures 6 and 7. We observe that the greedy approach is comparable to parzen on logistic regression, but is substantially worse on decision trees since changing a single feature at a time often does not have an effect on the prediction. The overall recall by parzen is low, likely due to the difficulty in approximating the original high-dimensional classifier. LIME consistently provides $> 90\%$ recall for both classifiers on both datasets, demonstrating that LIME explanations are faithful to the models.

5.3 Should I trust this prediction?

In order to simulate trust in individual predictions, we first randomly select 25% of the features to be “untrustworthy”, and assume that the users can identify and would not want to trust these features (such as the headers in 20 newsgroups, leaked data, etc). We thus develop *oracle* “trustworthiness” by labeling test set predictions from a black box classifier as “untrustworthy” if the prediction changes when untrustworthy features are removed from the instance, and “trustworthy” otherwise. In order to simulate users, we assume that users deem predictions untrustworthy from LIME and parzen explanations if the prediction from the linear approximation changes when all untrustworthy features that appear in the explanations are removed (the simulated human “discounts” the effect of untrustworthy features). For greedy and random, the prediction is mistrusted if any untrustworthy features are present in the explanation, since these methods do not provide a notion of the contribution of each feature to the prediction. Thus for each test set prediction, we can evaluate whether the simulated user trusts it using each explanation method, and compare it to the trustworthiness oracle.

Using this setup, we report the F1 on the trustworthy

Table 1: Average F1 of *trustworthiness* for different explainers on a collection of classifiers and datasets.

	Books				DVDs			
	LR	NN	RF	SVM	LR	NN	RF	SVM
Random	14.6	14.8	14.7	14.7	14.2	14.3	14.5	14.4
Parzen	84.0	87.6	94.3	92.3	87.0	81.7	94.2	87.3
Greedy	53.7	47.4	45.0	53.3	52.4	58.1	46.6	55.1
LIME	96.6	94.5	96.2	96.7	96.6	91.8	96.1	95.6

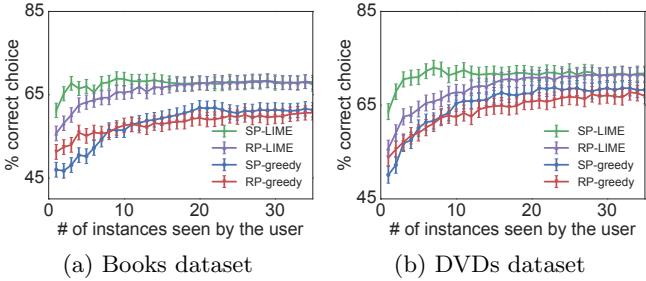


Figure 8: Choosing between two classifiers, as the number of instances shown to a simulated user is varied. Averages and standard errors from 800 runs.

predictions for each explanation method, averaged over 100 runs, in Table 1. The results indicate that LIME dominates others (all results are significant at $p = 0.01$) on both datasets, and for all of the black box models. The other methods either achieve a lower recall (i.e. they mistrust predictions more than they should) or lower precision (i.e. they trust too many predictions), while LIME maintains both high precision and high recall. Even though we artificially select which features are untrustworthy, these results indicate that LIME is helpful in assessing trust in individual predictions.

5.4 Can I trust this model?

In the final simulated user experiment, we evaluate whether the explanations can be used for model selection, simulating the case where a human has to decide between two competing models with similar accuracy on validation data. For this purpose, we add 10 artificially “noisy” features. Specifically, on training and validation sets (80/20 split of the original training data), each artificial feature appears in 10% of the examples in one class, and 20% of the other, while on the test instances, each artificial feature appears in 10% of the examples in each class. This recreates the situation where the models use not only features that are informative in the real world, but also ones that introduce spurious correlations. We create pairs of competing classifiers by repeatedly training pairs of random forests with 30 trees until their validation accuracy is within 0.1% of each other, but their test accuracy differs by at least 5%. Thus, it is not possible to identify the *better* classifier (the one with higher test accuracy) from the accuracy on the validation data.

The goal of this experiment is to evaluate whether a user can identify the better classifier based on the explanations of B instances from the validation set. The simulated human marks the set of artificial features that appear in the B explanations as untrustworthy, following which we evaluate how many total predictions in the validation set should be trusted (as in the previous section, treating only marked features as untrustworthy). Then, we select the classifier with

fewer untrustworthy predictions, and compare this choice to the classifier with higher held-out test set accuracy.

We present the accuracy of picking the correct classifier as B varies, averaged over 800 runs, in Figure 8. We omit SP-parzen and RP-parzen from the figure since they did not produce useful explanations, performing only slightly better than random. LIME is consistently better than greedy, irrespective of the pick method. Further, combining submodular pick with LIME outperforms all other methods, in particular it is much better than RP-LIME when only a few examples are shown to the users. These results demonstrate that the trust assessments provided by SP-selected LIME explanations are good indicators of generalization, which we validate with human experiments in the next section.

6. EVALUATION WITH HUMAN SUBJECTS

In this section, we recreate three scenarios in machine learning that require trust and understanding of predictions and models. In particular, we evaluate LIME and SP-LIME in the following settings: (1) Can users choose which of two classifiers generalizes better (§ 6.2), (2) based on the explanations, can users perform feature engineering to improve the model (§ 6.3), and (3) are users able to identify and describe classifier irregularities by looking at explanations (§ 6.4).

6.1 Experiment setup

For experiments in §6.2 and §6.3, we use the “Christianity” and “Atheism” documents from the 20 newsgroups dataset mentioned beforehand. This dataset is problematic since it contains features that do not generalize (e.g. very informative header information and author names), and thus validation accuracy considerably overestimates real-world performance.

In order to estimate the real world performance, we create a new *religion dataset* for evaluation. We download Atheism and Christianity websites from the DMOZ directory and human curated lists, yielding 819 webpages in each class. High accuracy on this dataset by a classifier trained on 20 newsgroups indicates that the classifier is generalizing using semantic content, instead of placing importance on the data specific issues outlined above. Unless noted otherwise, we use SVM with RBF kernel, trained on the 20 newsgroups data with hyper-parameters tuned via the cross-validation.

6.2 Can users select the best classifier?

In this section, we want to evaluate whether explanations can help users decide which classifier generalizes better, i.e., which classifier would the user deploy “in the wild”. Specifically, users have to decide between two classifiers: SVM trained on the original 20 newsgroups dataset, and a version of the same classifier trained on a “cleaned” dataset where many of the features that do not generalize have been manually removed. The original classifier achieves an accuracy score of 57.3% on the *religion dataset*, while the “cleaned” classifier achieves a score of 69.0%. In contrast, the test accuracy on the original 20 newsgroups split is 94.0% and 88.6%, respectively – suggesting that the worse classifier would be selected if accuracy alone is used as a measure of trust.

We recruit human subjects on Amazon Mechanical Turk – by no means machine learning experts, but instead people with basic knowledge about religion. We measure their ability to choose the better algorithm by seeing side-by-side explanations with the associated raw data (as shown in Figure 2). We restrict both the number of words in each explanation (K) and the number of documents that each

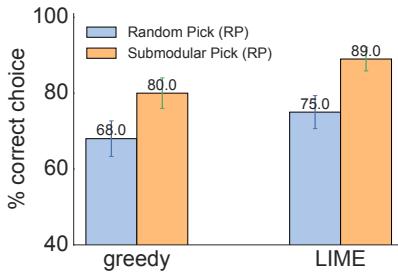


Figure 9: Average accuracy of human subject (with standard errors) in choosing between two classifiers.

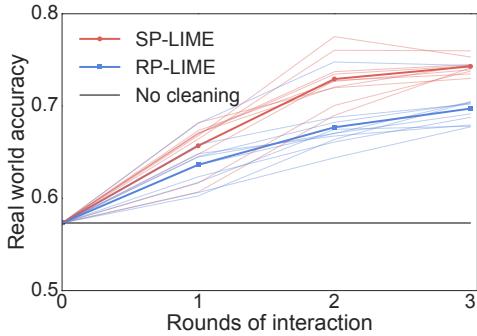


Figure 10: Feature engineering experiment. Each shaded line represents the average accuracy of subjects in a path starting from one of the initial 10 subjects. Each solid line represents the average across all paths per round of interaction.

person inspects (B) to 6. The position of each algorithm and the order of the instances seen are randomized between subjects. After examining the explanations, users are asked to select which algorithm will perform best in the real world. The explanations are produced by either greedy (chosen as a baseline due to its performance in the simulated user experiment) or LIME, and the instances are selected either by random (RP) or submodular pick (SP). We modify the greedy step in Algorithm 2 slightly so it alternates between explanations of the two classifiers. For each setting, we repeat the experiment with 100 users.

The results are presented in Figure 9. Note that all of the methods are good at identifying the better classifier, demonstrating that the explanations are useful in determining which classifier to trust, while using test set accuracy would result in the selection of the wrong classifier. Further, we see that the submodular pick (SP) greatly improves the user’s ability to select the best classifier when compared to random pick (RP), with LIME outperforming greedy in both cases.

6.3 Can non-experts improve a classifier?

If one notes that a classifier is untrustworthy, a common task in machine learning is feature engineering, i.e. modifying the set of features and retraining in order to improve generalization. Explanations can aid in this process by presenting the important features, particularly for removing features that the users feel do not generalize.

We use the 20 newsgroups data here as well, and ask Amazon Mechanical Turk users to identify which words from the explanations should be removed from subsequent training, for the worse classifier from the previous section (§6.2). In each round, the subject marks words for deletion after observing

$B = 10$ instances with $K = 10$ words in each explanation (an interface similar to Figure 2, but with a single algorithm). As a reminder, the users here are not experts in machine learning and are unfamiliar with feature engineering, thus are only identifying words based on their semantic content. Further, users do not have any access to the *religion* dataset – they do not even know of its existence. We start the experiment with 10 subjects. After they mark words for deletion, we train 10 different classifiers, one for each subject (with the corresponding words removed). The explanations for each classifier are then presented to a set of 5 users in a new round of interaction, which results in 50 new classifiers. We do a final round, after which we have 250 classifiers, each with a path of interaction tracing back to the first 10 subjects.

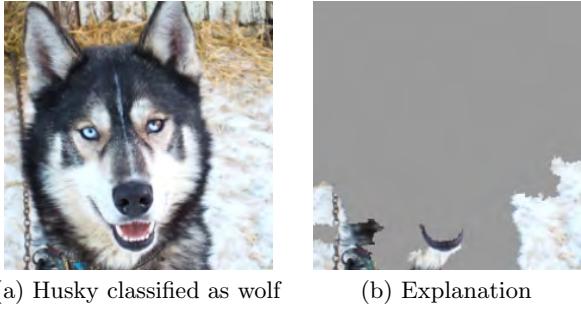
The explanations and instances shown to each user are produced by **SP-LIME** or **RP-LIME**. We show the average accuracy on the *religion* dataset at each interaction round for the paths originating from each of the original 10 subjects (shaded lines), and the average across all paths (solid lines) in Figure 10. It is clear from the figure that the crowd workers are able to improve the model by removing features they deem unimportant for the task. Further, **SP-LIME** outperforms **RP-LIME**, indicating selection of the instances to show the users is crucial for efficient feature engineering.

Each subject took an average of 3.6 minutes per round of cleaning, resulting in just under 11 minutes to produce a classifier that generalizes much better to real world data. Each path had on average 200 words removed with **SP**, and 157 with **RP**, indicating that incorporating coverage of important features is useful for feature engineering. Further, out of an average of 200 words selected with **SP**, 174 were selected by at least half of the users, while 68 by *all* the users. Along with the fact that the variance in the accuracy decreases across rounds, this high agreement demonstrates that the users are converging to similar *correct* models. This evaluation is an example of how explanations make it easy to improve an untrustworthy classifier – in this case easy enough that machine learning knowledge is not required.

6.4 Do explanations lead to insights?

Often artifacts of data collection can induce undesirable correlations that the classifiers pick up during training. These issues can be very difficult to identify just by looking at the raw data and predictions. In an effort to reproduce such a setting, we take the task of distinguishing between photos of Wolves and Eskimo Dogs (huskies). We train a logistic regression classifier on a training set of 20 images, hand selected such that all pictures of wolves had snow in the background, while pictures of huskies did not. As the features for the images, we use the first max-pooling layer of Google’s pre-trained Inception neural network [25]. On a collection of additional 60 images, the classifier predicts “Wolf” if there is snow (or light background at the bottom), and “Husky” otherwise, regardless of animal color, position, pose, etc. We trained this *bad* classifier intentionally, to evaluate whether subjects are able to detect it.

The experiment proceeds as follows: we first present a balanced set of 10 test predictions (without explanations), where one wolf is not in a snowy background (and thus the prediction is “Husky”) and one husky is (and is thus predicted as “Wolf”). We show the “Husky” mistake in Figure 11a. The other 8 examples are classified correctly. We then ask the subject three questions: (1) Do they trust this algorithm



(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model’s prediction in the “Husky vs Wolf” task.

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Table 2: “Husky vs Wolf” experiment results.

to work well in the real world, (2) why, and (3) how do they think the algorithm is able to distinguish between these photos of wolves and huskies. After getting these responses, we show the same images with the associated explanations, such as in Figure 11b, and ask the same questions.

Since this task requires some familiarity with the notion of spurious correlations and generalization, the set of subjects for this experiment were graduate students who have taken at least one graduate machine learning course. After gathering the responses, we had 3 independent evaluators read their reasoning and determine if each subject mentioned snow, background, or equivalent as a feature the model may be using. We pick the majority to decide whether the subject was correct about the insight, and report these numbers before and after showing the explanations in Table 2.

Before observing the explanations, more than a third trusted the classifier, and a little less than half mentioned the snow pattern as something the neural network was using – although all speculated on other patterns. After examining the explanations, however, almost all of the subjects identified the correct insight, with much more certainty that it was a determining factor. Further, the trust in the classifier also dropped substantially. Although our sample size is small, this experiment demonstrates the utility of explaining individual predictions for getting insights into classifiers knowing when not to trust them and why.

7. RELATED WORK

The problems with relying on validation set accuracy as the primary measure of trust have been well studied. Practitioners consistently overestimate their model’s accuracy [21], propagate feedback loops [23], or fail to notice data leaks [14]. In order to address these issues, researchers have proposed tools like Gestalt [20] and Modeltracker [1], which help users navigate individual instances. These tools are complementary to LIME in terms of explaining models, since they do not address the problem of explaining individual predictions. Further, our submodular pick procedure can be incorporated in such tools to aid users in navigating larger datasets.

Some recent work aims to anticipate failures in machine

learning, specifically for vision tasks [3, 29]. Letting users know when the systems are likely to fail can lead to an increase in trust, by avoiding “silly mistakes” [8]. These solutions either require additional annotations and feature engineering that is specific to vision tasks or do not provide insight into why a decision should not be trusted. Furthermore, they assume that the current evaluation metrics are reliable, which may not be the case if problems such as data leakage are present. Other recent work [11] focuses on exposing users to different kinds of mistakes (our pick step). Interestingly, the subjects in their study did not notice the serious problems in the 20 newsgroups data even after looking at many mistakes, suggesting that examining raw data is not sufficient. Note that (**author?**) [11] are not alone in this regard, many researchers in the field have unwittingly published classifiers that would not generalize for this task. Using LIME, we show that even non-experts are able to identify these irregularities when explanations are present. Further, LIME can complement these existing systems, and allow users to assess trust even when a prediction seems “correct” but is made for the wrong reasons.

Recognizing the utility of explanations in assessing trust, many have proposed using interpretable models [27], especially for the medical domain [6, 17, 26]. While such models may be appropriate for some domains, they may not apply equally well to others (e.g. a supersparse linear model [26] with 5 – 10 features is unsuitable for text applications). Interpretability, in these cases, comes at the cost of flexibility, accuracy, or efficiency. For text, EluciDebug [16] is a full human-in-the-loop system that shares many of our goals (interpretability, faithfulness, etc). However, they focus on an already interpretable model (Naive Bayes). In computer vision, systems that rely on object detection to produce candidate alignments [13] or attention [28] are able to produce explanations for their predictions. These are, however, constrained to specific neural network architectures or incapable of detecting “non object” parts of the images. Here we focus on general, model-agnostic explanations that can be applied to any classifier or regressor that is appropriate for the domain - even ones that are yet to be proposed.

A common approach to model-agnostic explanation is learning a potentially interpretable model on the predictions of the original model [2, 7, 22]. Having the explanation be a gradient vector [2] captures a similar locality intuition to that of LIME. However, interpreting the coefficients on the gradient is difficult, particularly for confident predictions (where gradient is near zero). Further, these explanations approximate the original model *globally*, thus maintaining local fidelity becomes a significant challenge, as our experiments demonstrate. In contrast, LIME solves the much more feasible task of finding a model that approximates the original model *locally*. The idea of perturbing inputs for explanations has been explored before [24], where the authors focus on learning a specific *contribution* model, as opposed to our general framework. None of these approaches explicitly take cognitive limitations into account, and thus may produce non-interpretable explanations, such as a gradients or linear models with thousands of non-zero weights. The problem becomes worse if the original features are nonsensical to humans (e.g. word embeddings). In contrast, LIME incorporates interpretability both in the optimization and in our notion of *interpretable representation*, such that domain and task specific interpretability criteria can be accommodated.

8. CONCLUSION AND FUTURE WORK

In this paper, we argued that trust is crucial for effective human interaction with machine learning systems, and that explaining individual predictions is important in assessing trust. We proposed LIME, a modular and extensible approach to faithfully explain the predictions of *any* model in an interpretable manner. We also introduced SP-LIME, a method to select representative and non-redundant predictions, providing a global view of the model to users. Our experiments demonstrated that explanations are useful for a variety of models in trust-related tasks in the text and image domains, with both expert and non-expert users: deciding between models, assessing trust, improving untrustworthy models, and getting insights into predictions.

There are a number of avenues of future work that we would like to explore. Although we describe only sparse linear models as explanations, our framework supports the exploration of a variety of explanation families, such as decision trees; it would be interesting to see a comparative study on these with real users. One issue that we do not mention in this work was how to perform the pick step for images, and we would like to address this limitation in the future. The domain and model agnosticism enables us to explore a variety of applications, and we would like to investigate potential uses in speech, video, and medical domains, as well as recommendation systems. Finally, we would like to explore theoretical properties (such as the appropriate number of samples) and computational optimizations (such as using parallelization and GPU processing), in order to provide the accurate, real-time explanations that are critical for any human-in-the-loop machine learning system.

Acknowledgements

We would like to thank Scott Lundberg, Tianqi Chen, and Tyler Johnson for helpful discussions and feedback. This work was supported in part by ONR awards #W911NF-13-1-0246 and #N00014-13-1-0023, and in part by TerraSwarm, one of six centers of STARnet, a Semiconductor Research Corporation program sponsored by MARCO and DARPA.

9. REFERENCES

- [1] S. Amershi, M. Chickering, S. M. Drucker, B. Lee, P. Simard, and J. Suh. Modeltracker: Redesigning performance analysis tools for machine learning. In *Human Factors in Computing Systems (CHI)*, 2015.
- [2] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11, 2010.
- [3] A. Bansal, A. Farhadi, and D. Parikh. Towards transparent systems: Semantic characterization of failure modes. In *European Conference on Computer Vision (ECCV)*, 2014.
- [4] J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Association for Computational Linguistics (ACL)*, 2007.
- [5] J. Q. Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset Shift in Machine Learning*. MIT, 2009.
- [6] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Knowledge Discovery and Data Mining (KDD)*, 2015.
- [7] M. W. Craven and J. W. Shavlik. Extracting tree-structured representations of trained networks. *Neural information processing systems (NIPS)*, pages 24–30, 1996.
- [8] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck. The role of trust in automation reliance. *Int. J. Hum.-Comput. Stud.*, 58(6), 2003.
- [9] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.
- [10] U. Feige. A threshold of $\ln n$ for approximating set cover. *J. ACM*, 45(4), July 1998.
- [11] A. Groce, T. Kulesza, C. Zhang, S. Shamasunder, M. Burnett, W.-K. Wong, S. Stumpf, S. Das, A. Shinsel, F. Bice, and K. McIntosh. You are the only possible oracle: Effective test selection for end users of interactive machine learning systems. *IEEE Trans. Softw. Eng.*, 40(3), 2014.
- [12] J. L. Herlocker, J. A. Konstan, and J. Riedl. Explaining collaborative filtering recommendations. In *Conference on Computer Supported Cooperative Work (CSCW)*, 2000.
- [13] A. Karpathy and F. Li. Deep visual-semantic alignments for generating image descriptions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [14] S. Kaufman, S. Rosset, and C. Perlich. Leakage in data mining: Formulation, detection, and avoidance. In *Knowledge Discovery and Data Mining (KDD)*, 2011.
- [15] A. Krause and D. Golovin. Submodular function maximization. In *Tractability: Practical Approaches to Hard Problems*. Cambridge University Press, February 2014.
- [16] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In *Intelligent User Interfaces (IUI)*, 2015.
- [17] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics*, 2015.
- [18] D. Martens and F. Provost. Explaining data-driven document classifications. *MIS Q.*, 38(1), 2014.
- [19] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Neural Information Processing Systems (NIPS)*. 2013.
- [20] K. Patel, N. Bancroft, S. M. Drucker, J. Fogarty, A. J. Ko, and J. Landay. Gestalt: Integrated support for implementation and analysis in machine learning. In *User Interface Software and Technology (UIST)*, 2010.
- [21] K. Patel, J. Fogarty, J. A. Landay, and B. Harrison. Investigating statistical machine learning as a tool for software development. In *Human Factors in Computing Systems (CHI)*, 2008.
- [22] I. Sanchez, T. Rocktaschel, S. Riedel, and S. Singh. Towards extracting faithful and descriptive representations of latent variable models. In *AAAI Spring Symposium on Knowledge Representation and Reasoning (KRR): Integrating Symbolic and Neural Approaches*, 2015.
- [23] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, and J.-F. Crespo. Hidden technical debt in machine learning systems. In *Neural Information Processing Systems (NIPS)*. 2015.
- [24] E. Strumbelj and I. Kononenko. An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 11, 2010.
- [25] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [26] B. Ustun and C. Rudin. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 2015.
- [27] F. Wang and C. Rudin. Falling rule lists. In *Artificial Intelligence and Statistics (AISTATS)*, 2015.
- [28] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, 2015.
- [29] P. Zhang, J. Wang, A. Farhadi, M. Hebert, and D. Parikh. Predicting failures of vision systems. In *Computer Vision and Pattern Recognition (CVPR)*, 2014.

Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems

Anupam Datta Shayak Sen Yair Zick
 Carnegie Mellon University, Pittsburgh, USA
 {danupam, shayaks, yairzick}@cmu.edu

Abstract—Algorithmic systems that employ machine learning play an increasing role in making substantive decisions in modern society, ranging from online personalization to insurance and credit decisions to predictive policing. But their decision-making processes are often opaque—it is difficult to explain why a certain decision was made. We develop a formal foundation to improve the transparency of such decision-making systems. Specifically, we introduce a family of *Quantitative Input Influence (QII)* measures that capture the degree of influence of inputs on outputs of systems. These measures provide a foundation for the design of transparency reports that accompany system decisions (e.g., explaining a specific credit decision) and for testing tools useful for internal and external oversight (e.g., to detect algorithmic discrimination).

Distinctively, our *causal* QII measures carefully account for correlated inputs while measuring influence. They support a general class of transparency queries and can, in particular, explain decisions about individuals (e.g., a loan decision) and groups (e.g., disparate impact based on gender). Finally, since single inputs may not always have high influence, the QII measures also quantify the *joint influence* of a set of inputs (e.g., age and income) on outcomes (e.g. loan decisions) and the *marginal influence* of individual inputs within such a set (e.g., income). Since a single input may be part of multiple influential sets, the average marginal influence of the input is computed using principled aggregation measures, such as the Shapley value, previously applied to measure influence in voting. Further, since transparency reports could compromise privacy, we explore the transparency-privacy tradeoff and prove that a number of useful transparency reports can be made differentially private with very little addition of noise.

Our empirical validation with standard machine learning algorithms demonstrates that QII measures are a useful transparency mechanism when black box access to the learning system is available. In particular, they provide better explanations than standard associative measures for a host of scenarios that we consider. Further, we show that in the situations we consider, QII is efficiently approximable and can be made differentially private while preserving accuracy.

I. INTRODUCTION

Algorithmic decision-making systems that employ machine learning and related statistical methods are ubiquitous. They drive decisions in sectors as diverse as Web services, health-care, education, insurance, law enforcement and defense [1], [2], [3], [4], [5]. Yet their decision-making processes are often opaque. *Algorithmic transparency* is an emerging research area aimed at explaining decisions made by algorithmic systems.

The call for algorithmic transparency has grown in intensity as public and private sector organizations increasingly use large volumes of personal information and complex data analytics systems for decision-making [6]. Algorithmic transparency provides several benefits. First, it is essential to enable identification of harms, such as discrimination, introduced by algorithmic decision-making (e.g., high interest credit cards targeted to protected groups) and to hold entities in the decision-making chain accountable for such practices. This form of accountability can incentivize entities to adopt appropriate corrective measures. Second, transparency can help detect errors in input data which resulted in an adverse decision (e.g., incorrect information in a user's profile because of which insurance or credit was denied). Such errors can then be corrected. Third, by explaining why an adverse decision was made, it can provide guidance on how to reverse it (e.g., by identifying a specific factor in the credit profile that needs to be improved).

Our Goal. While the importance of algorithmic transparency is recognized, work on computational foundations for this research area has been limited. This paper initiates progress in that direction by focusing on a concrete algorithmic transparency question:

How can we measure the influence of inputs (or features) on decisions made by an algorithmic system about individuals or groups of individuals?

Our goal is to inform the design of transparency reports, which include answers to transparency queries of this form. To be concrete, let us consider a predictive policing system that forecasts future criminal activity based on historical data; individuals high on the list receive visits from the police. An individual who receives a visit from the police may seek a transparency report that provides answers to *personalized transparency queries* about the influence of various inputs (or features), such as race or recent criminal history, on the system's decision. An oversight agency or the public may desire a transparency report that provides answers to *aggregate transparency queries*, such as the influence of sensitive inputs (e.g., gender, race) on the system's decisions concerning the entire population or about systematic differences in decisions

among groups of individuals (e.g., discrimination based on race or age). These reports can thus help identify harms and errors in input data, and provide guidance on what input features to work on to modify the decision.

Our Model. We focus on a setting where a *transparency report* is generated with black-box access to the decision-making system¹ and knowledge of the input dataset on which it operates. This setting models the kind of access available to a private or public sector entity that proactively publishes transparency reports. It also models a useful level of access required for internal or external oversight of such systems to identify harms introduced by them. For the former use case, our approach provides a basis for design of transparency mechanisms; for the latter, it provides a formal basis for testing. Returning to our predictive policing system, the law enforcement agency that employs it could proactively publish transparency reports, and test the system for early detection of harms like race-based discrimination. An oversight agency could also use transparency reports for post hoc identification of harms.

Our Approach. We formalize transparency reports by introducing a family of *Quantitative Input Influence (QII)* measures that capture the degree of influence of inputs on outputs of the system. Three desiderata drove the definitions of these measures.

First, we seek a formalization of a *general* class of transparency reports that allows us to answer many useful transparency queries related to input influence, including but not limited to the example forms described above about the system’s decisions about individuals and groups.

Second, we seek input influence measures that appropriately account for *correlated inputs*—a common case for our target applications. For example, consider a system that assists in hiring decisions for a moving company. Gender and the ability to lift heavy weights are inputs to the system. They are positively correlated with each other and with the hiring decisions. Yet transparency into whether the system uses the weight lifting ability or the gender in making its decisions (and to what degree) has substantive implications for determining if it is engaging in discrimination (the business necessity defense could apply in the former case [7]). This observation makes us look beyond correlation coefficients and other associative measures.

Third, we seek measures that appropriately quantify input influence in settings where any input by itself does not have significant influence on outcomes but a set of inputs does. In such cases, we seek measures of *joint influence* of a set of inputs (e.g., age and income) on a system’s decision (e.g., to serve a high-paying job ad). We also seek measures of *marginal influence* of an input within such a set (e.g., age) on the decision. This notion allows us to provide finer-grained

transparency about the relative importance of individual inputs within the set (e.g., age vs. income) in the system’s decision.

We achieve the first desideratum by formalizing a notion of a *quantity of interest*. A transparency query measures the influence of an input on a quantity of interest. A quantity of interest represents a property of the behavior of the system for a given input distribution. Our formalization supports a wide range of statistical properties including probabilities of various outcomes in the output distribution and probabilities of output distribution outcomes conditioned on input distribution events. Examples of quantities of interest include the conditional probability of an outcome for a particular individual or group, and the ratio of conditional probabilities for an outcome for two different groups (a metric used as evidence of disparate impact under discrimination law in the US [7]).

We achieve the second desideratum by formalizing *causal QII* measures. These measures (called *Unary QII*) model the difference in the quantity of interest when the system operates over two related input distributions—the real distribution and a hypothetical (or counterfactual) distribution that is constructed from the real distribution in a specific way to account for correlations among inputs. Specifically, if we are interested in measuring the influence of an input on a quantity of interest of the system behavior, we construct the hypothetical distribution by retaining the marginal distribution over all other inputs and sampling the input of interest from its prior distribution. This choice breaks the correlations between this input and all other inputs and thus lets us measure the influence of this input on the quantity of interest, independently of other correlated inputs. Revisiting our moving company hiring example, if the system makes decisions only using the weightlifting ability of applicants, the influence of gender will be zero on the ratio of conditional probabilities of being hired for males and females.

We achieve the third desideratum in two steps. First, we define a notion of joint influence of a set of inputs (called *Set QII*) via a natural generalization of the definition of the hypothetical distribution in the Unary QII definition. Second, we define a family of *Marginal QII* measures that model the difference on the quantity of interest as we consider sets with and without the specific input whose marginal influence we want to measure. Depending on the application, we may pick these sets in different ways, thus motivating several different measures. For example, we could fix a set of inputs and ask about the marginal influence of any given input in that set on the quantity of interest. Alternatively, we may be interested in the average marginal influence of an input when it belongs to one of several different sets that significantly affect the quantity of interest. We consider several marginal influence aggregation measures from cooperative game theory originally developed in the context of influence measurement in voting scenarios and discuss their applicability in our setting. We also build on that literature to present an efficient approximate algorithm for computing these measures.

Recognizing that different forms of transparency reports may be appropriate for different settings, we generalize our QII measures to be parametric in its key elements: the intervention

¹By “black-box access to the decision-making system” we mean a typical setting of software testing with complete control of inputs to the system and full observability of the outputs.

used to construct the hypothetical input distribution; the quantity of interest; the difference measure used to quantify the distance in the quantity of interest when the system operates over the real and hypothetical input distributions; and the aggregation measure used to combine marginal QII measures across different sets. This generalized definition provides a structure for exploring the design space of transparency reports.

Since transparency reports released to an individual, regulatory agency, or the public might compromise individual privacy, we explore the possibility of answering transparency queries while protecting differential privacy [8]. We prove bounds on the sensitivity of a number of transparency queries and leverage prior results on privacy amplification via sampling [9] to accurately answer these queries.

We demonstrate the utility of the QII framework by developing two machine learning applications on real datasets: an income classification application based on the benchmark adult dataset [10], and a predictive policing application based on the National Longitudinal Survey of Youth [11]. Using these applications, we argue, in Section VII, the need for causal measurement by empirically demonstrating that in the presence of correlated inputs, observational measures are not informative in identifying input influence. Further, we analyze transparency reports of individuals in our dataset to demonstrate how Marginal QII can provide insights into individuals' classification outcomes. Finally, we demonstrate that under most circumstances, QII measures can be made differentially private with minimal addition of noise, and can be approximated efficiently.

In summary, this paper makes the following contributions:

- A formalization of a specific algorithmic transparency problem for decision-making systems. Specifically, we define a family of Quantitative Input Influence metrics that accounts for correlated inputs, and provides answers to a general class of transparency queries, including the absolute and marginal influence of inputs on various behavioral system properties. These metrics can inform the design of transparency mechanisms and guide proactive system testing and posthoc investigations.
- A formal treatment of privacy-transparency trade-offs, in particular, by construction of differentially private answers to transparency queries.
- An implementation and experimental evaluation of the metrics over two real data sets. The evaluation demonstrates that (a) the QII measures are *informative*; (b) they remain *accurate* while preserving differential privacy; and (c) can be *computed* quite quickly for standard machine learning systems applied to real data sets.

II. UNARY QII

Consider the situation discussed in the introduction, where an automated system assists in hiring decisions for a moving company. The input features used by this classification system are : *Age*, *Gender*, *Weight Lifting Ability*, *Marital Status* and *Education*. Suppose that, as before, weight lifting ability is

strongly correlated with gender (with men having better overall lifting ability than woman). One particular question that an analyst may want to ask is: "What is the influence of the input *Gender* on positive classification for women?". The analyst observes that 20% of women are approved according to his classifier. Then, he replaces every woman's field for gender with a random value, and notices that the number of women approved does not change. In other words, an *intervention* on the *Gender* variable does not cause a significant change in the classification outcome. Repeating this process with *Weight Lifting Ability* results in a 20% increase in women's hiring. Therefore, he concludes that for this classifier, *Weight Lifting Ability* has more influence on positive classification for women than *Gender*.

By breaking correlations between gender and weight lifting ability, we are able to establish a causal relationship between the outcome of the classifier and the inputs. We are able to identify that despite the strong correlation between a negative classification outcome for women, the feature gender was not a cause of this outcome. We formalize the intuition behind such causal experimentation in our definition of Quantitative Input Influence (QII).

We are given an algorithm \mathcal{A} . \mathcal{A} operates on inputs (also referred to as *features* for ML systems), $N = \{1, \dots, n\}$. Every $i \in N$, can take on various *states*, given by X_i . We let $\mathcal{X} = \prod_{i \in N} \mathcal{X}_i$ be the set of possible feature state vectors, let \mathcal{Z} be the set of possible outputs of \mathcal{A} . For a vector $\mathbf{x} \in \mathcal{X}$ and set of inputs $S \subseteq N$, $\mathbf{x}|_S$ denotes the vector of inputs in S . We are also given a probability distribution π on \mathcal{X} , where $\pi(\mathbf{x})$ is the probability of the input vector \mathbf{x} . We can define a marginal probability of a set of inputs S in the standard way as follows:

$$\pi_S(\mathbf{x}|_S) = \sum_{\{\mathbf{x}' \in \mathcal{X} | \mathbf{x}'|_S = \mathbf{x}|_S\}} \pi(\mathbf{x}') \quad (1)$$

When S is a singleton set $\{i\}$, we write the marginal probability of the single input as $\pi_i(x)$.

Informally, to quantify the influence of an input i , we compute its effect on some *quantity of interest*; that is, we measure the difference in the quantity of interest, when the feature i is changed via an intervention. In the example above, the quantity of interest is the fraction of positive classification of women. In this paper, we employ a particular interpretation of "changing an input", where we replace the value of every input with a random independently chosen value. To describe the replacement operation for input i , we first define an expanded probability space on $\mathcal{X} \times \mathcal{X}$, with the following distribution:

$$\tilde{\pi}(\mathbf{x}, \mathbf{u}) = \pi(\mathbf{x})\pi(\mathbf{u}). \quad (2)$$

The first component of an expanded vector (\mathbf{x}, \mathbf{u}) , is just the original input vector, whereas the second component represents an independent random vector drawn from the same distribution π . Over this expanded probability space, the random variable $X(\mathbf{x}, u_i) = \mathbf{x}$ represents the original feature vector.

The random variable $X_{-i}U_i(\mathbf{x}, \mathbf{u}) = \mathbf{x}|_{N \setminus \{i\}}u_i$, represents the random variable with input i replaced with a random sample. Defining this expanded probability space allows us to switch between the original distribution, represented by the random variable X , and the *intervened distribution*, represented by $X_{-i}U_i$. Notice that both these random variables are defined from $\mathcal{X} \times \mathcal{X}$, the expanded probability space, to \mathcal{X} . We denote the set of random variables of the type $\mathcal{X} \times \mathcal{X} \rightarrow \mathcal{X}$ as $\mathfrak{R}(\mathcal{X})$.

We can now define probabilities over this expanded space. For example, the probability over X remains the same:

$$\begin{aligned}\Pr(X = \mathbf{x}) &= \sum_{\{(\mathbf{x}', \mathbf{u}') | \mathbf{x}' = \mathbf{x}\}} \tilde{\pi}(\mathbf{x}', \mathbf{u}') \\ &= \left(\sum_{\{\mathbf{x}' | \mathbf{x}' = \mathbf{x}\}} \pi(\mathbf{x}') \right) \left(\sum_{\mathbf{u}'} \pi(\mathbf{u}') \right) \\ &= \pi(\mathbf{x})\end{aligned}$$

Similarly, we can define more complex quantities. The following expression represents the expectation of a classifier c evaluating to 1, when i is randomly intervened on:

$$\mathbb{E}(c(X_{-i}U_i) = 1) = \sum_{\{(\mathbf{x}, \mathbf{u}) | c(\mathbf{x}_{N \setminus i}u_i) = 1\}} \tilde{\pi}(\mathbf{x}, u_i).$$

Observe that the expression above computes the probability of the classifier c evaluating to 1, when input i is replaced with a random sample from its probability distribution $\pi_i(u_i)$.

$$\begin{aligned}&\sum_{\{(\mathbf{x}, \mathbf{u}) | c(\mathbf{x}_{N \setminus i}u_i) = 1\}} \tilde{\pi}(\mathbf{x}, u_i) \\ &= \sum_{\mathbf{x}} \pi(\mathbf{x}) \sum_{\{u'_i | c(\mathbf{x}_{N \setminus i}u'_i) = 1\}} \sum_{\{\mathbf{u} | u_i = u'_i\}} \pi(\mathbf{u}) \\ &= \sum_{\mathbf{x}} \pi(\mathbf{x}) \sum_{\{u'_i | c(\mathbf{x}_{N \setminus i}u'_i) = 1\}} \pi_i(u'_i)\end{aligned}$$

We can also define conditional distributions in the usual way. The following represents the probability of the classifier evaluating to 1 under the randomized intervention on input i of X , given that X belongs to some subset $\mathcal{Y} \subseteq \mathcal{X}$:

$$\mathbb{E}(c(X_{-i}U_i) = 1 | X \in \mathcal{Y}) = \frac{\mathbb{E}(c(X_{-i}U_i) = 1 \wedge X \in \mathcal{Y})}{\mathbb{E}(X \in \mathcal{Y})}.$$

Formally, for an algorithm \mathcal{A} , a *quantity of interest* $Q_{\mathcal{A}}(\cdot) : \mathfrak{R}(\mathcal{X}) \mapsto \mathbb{R}$ is a function of a random variable from $\mathfrak{R}(\mathcal{X})$.

Definition 1 (QII). For a quantity of interest $Q_{\mathcal{A}}(\cdot)$, and an input i , the Quantitative Input Influence of i on $Q_{\mathcal{A}}(\cdot)$ is defined to be

$$\iota^{Q_{\mathcal{A}}}(i) = Q_{\mathcal{A}}(X) - Q_{\mathcal{A}}(X_{-i}U_i).$$

In the example above, for a classifier \mathcal{A} , the quantity of interest, the fraction of women (represented by the set $\mathcal{W} \subseteq \mathcal{X}$) with positive classification, can be expressed as follows:

$$Q_{\mathcal{A}}(\cdot) = \mathbb{E}(\mathcal{A}(\cdot) = 1 | X \in \mathcal{W}),$$

and the influence of input i is:

$$\iota(i) = \mathbb{E}(\mathcal{A}(X) = 1 | X \in \mathcal{W}) - \mathbb{E}(\mathcal{A}(X_{-i}U_i) = 1 | X \in \mathcal{W}).$$

When \mathcal{A} is clear from the context, we simply write Q rather than $Q_{\mathcal{A}}$. We now instantiate this definition with different quantities of interest to illustrate the above definition in three different scenarios.

A. QII for Individual Outcomes

One intended use of QII is to provide personalized transparency reports to users of data analytics systems. For example, if a person is denied a job application due to feedback from a machine learning algorithm, an explanation of which factors were most influential for that person's classification can provide valuable insight into the classification outcome.

For QII to quantify the use of an input for individual outcomes, we define the quantity of interest to be the classification outcome for a particular individual. Given a particular individual \mathbf{x} , we define $Q_{\text{ind}}^{\mathbf{x}}(\cdot)$ to be $\mathbb{E}(c(\cdot) = 1 | X = \mathbf{x})$. The influence measure is therefore:

$$\iota_{\text{ind}}^{\mathbf{x}}(i) = \mathbb{E}(c(X) = 1 | X = \mathbf{x}) - \mathbb{E}(c(X_{-i}U_i) = 1 | X = \mathbf{x}) \quad (3)$$

When the quantity of interest is not the probability of positive classification but the classification that \mathbf{x} actually received, a slight modification of the above QII measure is more appropriate:

$$\begin{aligned}\iota_{\text{ind-act}}^{\mathbf{x}}(i) &= \mathbb{E}(c(X) = c(\mathbf{x}) | X = \mathbf{x}) \\ &\quad - \mathbb{E}(c(X_{-i}U_i) = c(\mathbf{x}) | X = \mathbf{x}) \\ &= 1 - \mathbb{E}(c(X_{-i}U_i) = c(\mathbf{x}) | X = \mathbf{x}) \\ &= \mathbb{E}(c(X_{-i}U_i) \neq c(\mathbf{x}) | X = \mathbf{x})\end{aligned} \quad (4)$$

The above probability can be interpreted as the probability that feature i is pivotal to the classification of $c(\mathbf{x})$. Computing the average of this quantity over X yields:

$$\begin{aligned}&\sum_{\mathbf{x} \in \mathcal{X}} \Pr(X = \mathbf{x}) \mathbb{E}(i \text{ is pivotal for } c(X) | X = \mathbf{x}) \\ &= \mathbb{E}(i \text{ is pivotal for } c(X)).\end{aligned} \quad (5)$$

We denote this average QII for individual outcomes as defined above, by $\iota_{\text{ind-avg}}(i)$, and use it as a measure for importance of an input towards classification outcomes.

B. QII for Group Outcomes

As in the running example, the quantity of interest may be the classification outcome for a set of individuals. Given a group of individuals $\mathcal{Y} \subseteq \mathcal{X}$, we define $Q_{\text{grp}}^{\mathcal{Y}}(\cdot)$ to be $\mathbb{E}(c(\cdot) = 1 | X \in \mathcal{Y})$. The influence measure is therefore:

$$\iota_{\text{grp}}^{\mathcal{Y}}(i) = \mathbb{E}(c(X) = 1 | X \in \mathcal{Y}) - \mathbb{E}(c(X_{-i}U_i) = 1 | X \in \mathcal{Y}) \quad (6)$$

C. QII for Group Disparity

Instead of simply classification outcomes, an analyst may be interested in more nuanced properties of data analytics systems. Recently, disparate impact has come to the fore as a measure of unfairness, which compares the rates of positive classification within protected groups defined by gender or race. The ‘80% rule’ in employment which states that the rate of selection within a protected demographic should be at least 80% of the rate of selection within the unprotected demographic. The quantity of interest in such a scenario is the ratio in positive classification outcomes for a protected group \mathcal{Y} from the rest of the population $\mathcal{X} \setminus \mathcal{Y}$.

$$\frac{\mathbb{E}(c(X) = 1 | X \in \mathcal{Y})}{\mathbb{E}(c(X) = 1 | X \notin \mathcal{Y})}$$

However, the ratio of classification rates is unstable at low values of positive classification. Therefore, for the computations in this paper we use the difference in classification rates as our measure of group disparity.

$$Q_{\text{disp}}^{\mathcal{Y}}(\cdot) = |\mathbb{E}(c(\cdot) = 1 | X \in \mathcal{Y}) - \mathbb{E}(c(\cdot) = 1 | X \notin \mathcal{Y})| \quad (7)$$

The QII measure of an input group disparity, as a result is:

$$\iota_{\text{disp}}^{\mathcal{Y}}(i) = Q_{\text{disp}}^{\mathcal{Y}}(X) - Q_{\text{disp}}^{\mathcal{Y}}(X_{-i}U_i). \quad (8)$$

More generally, group disparity can be viewed as an association between classification outcomes and membership in a group. QII on a measure of such association (e.g., group disparity) identifies the variable that causes the association in the classifier. *Proxy variables* are variables that are associated with protected attributes. However, for concerns of discrimination such as *digital redlining*, it is important to identify which proxy variables actually introduce group disparity. It is straightforward to observe that features with high QII for group disparity are proxy variables, and also cause group disparity. Therefore, QII on group disparity is a useful diagnostic tool for determining discrimination. The use of QII in identifying proxy variables is explored experimentally in Section VII-B. Note that because of such proxy variables, simply ensuring that protected attributes are not input to the classifier is not sufficient to avoid discrimination (see also [12]).

III. SET AND MARGINAL QII

In many situations, intervention on a single input variable has no influence on the outcome of a system. Consider, for example, a two-feature setting where features are age (A) and income (I), and the classifier is $c(A, I) = (A = \text{old}) \wedge (I = \text{high})$. In other words, the only datapoints that are labeled 1 are those of elderly persons with high income. Now, given a datapoint where $A = \text{young}, I = \text{low}$, an intervention on either age or income would result in the same classification. However, it would be misleading to say that neither age nor income have an influence over the outcome: changing both the states of income and age would result in a change in outcome.

Equating influence with the *individual* ability to affect the outcome is uninformative in real datasets as well: Figure 1 is a histogram of influences of features on outcomes of individuals for a classifier learnt from the adult dataset [13]². For most individuals, all features have zero influence: changing the state of one feature alone is not likely to change the outcome of a classifier. Of the 19537 datapoints we evaluate, more than half have $\iota^{\mathbf{x}}(i) = 0$ for all $i \in N$. Indeed, changes to outcome are more likely to occur if we intervene on *sets of features*. In order to get a better understanding of the influence of a feature $i \in N$, we should measure its effect when coupled with interventions on other features. We define the influence of a set of inputs as a straightforward extension of the influence of individual inputs. Essentially, we wish the influence of a set of inputs $S \subseteq N$ to be the same as when the set of inputs is considered to be a single input; when intervening on S , we draw the states of $i \in S$ based on the joint distribution of the states of features in S , $\pi_S(\mathbf{u}_S)$, as defined in Equation (1).

We can naturally define a distribution over $\mathcal{X} \times \prod_{i \in S} \mathcal{X}_i$, naturally extending (2) as:

$$\tilde{\pi}(\mathbf{x}, \mathbf{u}_S) = \pi(\mathbf{x})\pi_S(\mathbf{u}_S). \quad (9)$$

We also define the random variable $X_{-S}U_S(\mathbf{x}, \mathbf{u}_S) = \mathbf{x}|_{N \setminus S}\mathbf{u}_S$; $X_{-S}(\mathbf{x}, \mathbf{u}_S)$ has the states of features in $N \setminus S$ fixed to their original values in \mathbf{x} , but features in S take on new values according to \mathbf{u}_S .

Definition 2 (Set QII). For a quantity of interest Q , and an input i , the Quantitative Input Influence of set $S \subseteq N$ on Q is defined to be

$$\iota^Q(S) = Q(X) - Q(X_{-S}U_S).$$

Considering the influence of a set of inputs opens up a number of interesting questions due to the interaction between inputs. First among these is how does one measure the *individual effect* of a feature, given the measured effects of interventions on sets of features. One natural way of doing so is by measuring the *marginal effect* of a feature on a set.

²The adult dataset contains approximately 31k datapoints of users’ personal attributes, and whether their income is more than \$50k per annum; see Section VII for more details.

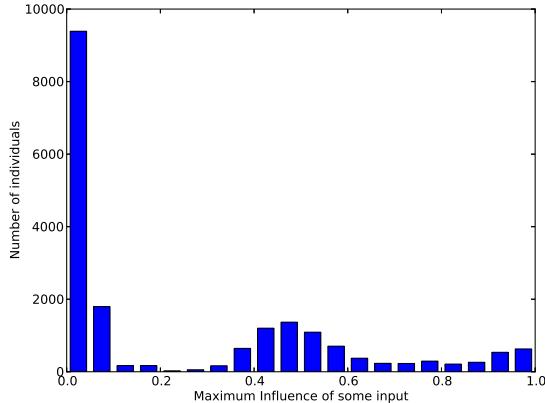


Fig. 1: A histogram of the highest specific causal influence for some feature across individuals in the adult dataset. Alone, most inputs alone have very low influence.

Definition 3 (Marginal QII). For a quantity of interest Q , and an input i , the Quantitative Input Influence of input i over a set $S \subseteq N$ on Q is defined to be

$$\iota^Q(i, S) = Q(X_{-S} U_S) - Q(X_{-S \cup \{i\}} U_{S \cup \{i\}}).$$

Notice that marginal QII can also be viewed as a difference in set QIIs: $\iota^Q(S \cup \{i\}) - \iota^Q(S)$. Informally, the difference between $\iota^Q(S \cup \{i\})$ and $\iota^Q(S)$ measures the “added value” obtained by intervening on $S \cup \{i\}$, versus intervening on S alone.

The marginal contribution of i may vary significantly based on S . Thus, we are interested in the *aggregate marginal contribution* of i to S , where S is sampled from some natural distribution over subsets of $N \setminus \{i\}$. In what follows, we describe a few measures for aggregating the marginal contribution of a feature i to sets, based on different methods for sampling sets. The primary method of aggregating the marginal contribution is the Shapley value [14]. The less theoretically inclined reader can choose to proceed to Section V without a loss in continuity.

A. Cooperative Games and Causality

In this section, we discuss how measures from the theory of cooperative games define measures for aggregating marginal influence. In particular, we observe that the Shapley value [14] is characterized by axioms that are natural in our setting. However, other measures may be appropriate for certain input data generation processes.

Definition 2 measures the influence that an intervention on a set of features $S \subseteq N$ has on the outcome. One can naturally think of Set QII as a function $v : 2^N \rightarrow \mathbb{R}$, where $v(S)$ is the influence of S on the outcome. With this intuition in mind, one can naturally study influence measures using *cooperative game theory*, and in particular, prevalent influence measures in cooperative games such as the Shapley value, Banzhaf index and others. These measures can be thought of as *influence*

aggregation methods, which, given an influence measure $v : 2^N \rightarrow \mathbb{R}$, output a vector $\phi \in \mathbb{R}^n$, whose i -th coordinate corresponds in some natural way to the aggregate influence, or aggregate causal effect, of feature i .

The original motivation for game-theoretic measures is *revenue division* [15, Chapter 18]: the function v describes the amount of money that each subset of players $S \subseteq N$ can generate; assuming that the set N generates a total revenue of $v(N)$, how should $v(N)$ be divided amongst the players? A special case of revenue division that has received significant attention is the measurement of voting power [16]. In voting systems with multiple agents with differing weights, voting power often does not directly correspond to the weights of the agents. For example, the US presidential election can roughly be modeled as a cooperative game where each state is an agent. The weight of a state is the number of electors in that state (i.e., the number of votes it brings to the presidential candidate who wins that state). Although states like California and Texas have higher weight, swing states like Pennsylvania and Ohio tend to have higher power in determining the outcome of elections.

A voting system is modeled as a cooperative game: players are voters, and the value of a coalition $S \subseteq N$ is 1 if S can make a decision (e.g. pass a bill, form a government, or perform a task), and is 0 otherwise. Note the similarity to classification, with players being replaced by features. The game-theoretic measures of revenue division are a measure of *voting power*: how much influence does player i have in the decision making process? Thus the notions of voting power and revenue division fit naturally with our goals when defining aggregate QII influence measures: in both settings, one is interested in measuring the aggregate effect that a single element has, given the actions of subsets.

A revenue division should ideally satisfy certain desiderata. Formally, we wish to find a function $\phi(N, v)$, whose input is N and $v : 2^N \rightarrow \mathbb{R}$, and whose output is a vector in \mathbb{R}^n , such that $\phi_i(N, v)$ measures some quantity describing the overall contribution of the i -th player. Research on fair revenue division in cooperative games traditionally follows an axiomatic approach: define a set of properties that a revenue division should satisfy, derive a function that outputs a value for each player, and argue that it is the unique function that satisfies these properties.

Several canonical fair cooperative solution concepts rely on the fundamental notion of *marginal contribution*. Given a player i and a set $S \subseteq N \setminus \{i\}$, the marginal contribution of i to S is denoted $m_i(S, v) = v(S \cup \{i\}) - v(S)$ (we simply write $m_i(S)$ when v is clear from the context). Marginal QII, as defined above, can be viewed as an instance of a measure of marginal contribution. Given a permutation $\pi \in \Pi(N)$ of the elements in N , we define $P_i(\sigma) = \{j \in N \mid \sigma(j) < \sigma(i)\}$; this is the set of i ’s *predecessors* in σ . We can now similarly define the marginal contribution of i to a permutation $\sigma \in \Pi(N)$ as $m_i(\sigma) = m_i(P_i(\sigma))$. Intuitively, one can think of the players sequentially entering a room, according to some ordering σ ; the value $m_i(\sigma)$ is the marginal contribution that i has to whoever is in the room when she enters it.

Generally speaking, game theoretic influence measures specify some reasonable way of aggregating the marginal contributions of i to sets $S \subseteq N$. That is, they measure a player's *expected marginal contribution* to sets sampled from some distribution \mathcal{D} over 2^N , resulting in a payoff of

$$\mathbb{E}_{S \sim \mathcal{D}}[m_i(S)] = \sum_{S \subseteq N} \Pr_{\mathcal{D}}[S] m_i(S).$$

Thus, fair revenue division draws its appeal from the degree to which the distribution \mathcal{D} is justifiable within the context where revenue is shared. In our setting, we argue for the use of the Shapley value. Introduced by the late Lloyd Shapley, the Shapley value is one of the most canonical methods of dividing revenue in cooperative games. It is defined as follows:

$$\varphi_i(N, v) = \mathbb{E}_\sigma[m_i(\sigma)] = \frac{1}{n!} \sum_{\sigma \in \Pi(N)} m_i(\sigma)$$

Intuitively, the Shapley value describes the following process: players are sequentially selected according to some randomly chosen order σ ; each player receives a payment of $m_i(\sigma)$. The Shapley value is the expected payment to the players under this regime. The definition we use describes a distribution over permutations of N , not its subsets; however, it is easy to describe the Shapley value in terms of a distribution over subsets. If we define $p[S] = \frac{1}{n} \frac{1}{\binom{n-1}{|S|}}$, it is a simple exercise to show that

$$\varphi_i(N, v) = \sum_{S \subseteq N} p[S] m_i(S).$$

Intuitively, $p[S]$ describes the following process: first, choose a number $k \in [0, n-1]$ uniformly at random; next, choose a set of size k uniformly at random.

The Shapley value is one of many reasonable ways of measuring influence; we provide a detailed review of two others — the *Banzhaf index* [17], and the *Deegan-Packel index* [18] — in Appendix A.

B. Axiomatic Treatment of the Shapley Value

In this work, the Shapley value is our function of choice for aggregating marginal feature influence. The objective of this section is to justify our choice, and provide a brief exposition of axiomatic game-theoretic value theory. We present the axioms that define the Shapley value, and discuss how they apply in the QII setting. As we show, by requiring some desired properties, one arrives at a game-theoretic influence measure as the *unique* function for measuring information use in our setting.

The Shapley value satisfies the following properties:

Definition 4 (Symmetry (Sym)). We say that $i, j \in N$ are *symmetric* if $v(S \cup \{i\}) = v(S \cup \{j\})$ for all $S \subseteq N \setminus \{i, j\}$. A value ϕ satisfies *symmetry* if $\phi_i = \phi_j$ whenever i and j are symmetric.

Definition 5 (Dummy (Dum)). We say that a player $i \in N$ is a *dummy* if $v(S \cup \{i\}) = v(S)$ for all $S \subseteq N$. A value ϕ satisfies the *dummy* property if $\phi_i = 0$ whenever i is a dummy.

Definition 6 (Efficiency (Eff)). A value satisfies the *efficiency* property if $\sum_{i \in N} \phi_i = v(N)$.

All of these axioms take on a natural interpretation in the QII setting. Indeed, if two features have the same probabilistic effect, no matter what other interventions are already in place, they should have the same influence. In our context, the dummy axiom says that a feature that never offers information with respect to an outcome should have no influence. In the case of specific causal influence, the efficiency axiom simply states that the total amount of influence should sum to

$$\begin{aligned} & \Pr(c(X) = c(\mathbf{x}) \mid X = \mathbf{x}) - \Pr(c(X_{-N}) = c(\mathbf{x}) \mid X = \mathbf{x}) \\ &= 1 - \Pr(c(X) = c(\mathbf{x})) = \Pr(c(X) \neq c(\mathbf{x})). \end{aligned}$$

That is, the total amount of influence possible is the likelihood of encountering elements whose evaluation is not $c(\mathbf{x})$. This is natural: if the vast majority of elements have a value of $c(\mathbf{x})$, it is quite unlikely that changes in features' state will have any effect on the outcome whatsoever; thus, the total amount of influence that can be assigned is $\Pr(c(X) \neq c(\mathbf{x}))$. Similarly, if the vast majority of points have a value different from \mathbf{x} , then it is likelier that a random intervention would result in a change in value, resulting in more influence to be assigned.

In the original paper by [14], it is shown that the Shapley value is the only function that satisfies (Sym), (Dum), (Eff), as well as the additivity (Add) axiom.

Definition 7 (Additivity (Add)). Given two games $\langle N, v_1 \rangle, \langle N, v_2 \rangle$, we write $\langle N, v_1 + v_2 \rangle$ to denote the game $v'(S) = v_1(S) + v_2(S)$ for all $S \subseteq N$. A value ϕ satisfies the *additivity* property if $\phi_i(N, v_1) + \phi_i(N, v_2) = \phi_i(N, v_1 + v_2)$ for all $i \in N$.

In our setting, the additivity axiom makes little intuitive sense; it would imply, for example, that if we were to multiply Q by a constant c , the influence of i in the resulting game should be multiplied by c as well, which is difficult to justify.

[19] offers an alternative characterization of the Shapley value, based on the more natural *monotonicity* assumption, which is a strong generalization of the dummy axiom.

Definition 8 (Monotonicity (Mono)). Given two games $\langle N, v_1 \rangle, \langle N, v_2 \rangle$, a value ϕ satisfies *strong monotonicity* if $m_i(S, v_1) \geq m_i(S, v_2)$ for all S implies that $\phi_i(N, v_1) \geq \phi_i(N, v_2)$, where a strict inequality for some set $S \subseteq N$ implies a strict inequality for the values as well.

Monotonicity makes intuitive sense in the QII setting: if a feature has consistently higher influence on the outcome in one setting than another, its measure of influence should increase. For example, if a user receives two transparency reports (say, for two separate loan applications), and in one report gender had a consistently higher effect on the outcome than in the other, then the transparency report should reflect this.

Theorem 9 ([19]). *The Shapley value is the only function that satisfies (Sym), (Eff) and (Mono).*

To conclude, the Shapley value is a *unique* way of measuring aggregate influence in the QII setting, while satisfying a set of very natural axioms.

IV. TRANSPARENCY SCHEMAS

We now discuss two generalizations of the definitions presented in Section II, and then define a transparency schema that map the space of transparency reports based on QII.

a) Intervention Distribution: In this paper we only consider randomized interventions when the interventions are drawn independently from the priors of the given input. However, depending on the specific causal question at hand, we may use different interventions. Formally, this is achieved by allowing an arbitrary intervention distribution π^{inter} such that

$$\tilde{\pi}(\mathbf{x}, \mathbf{u}) = \pi(\mathbf{x})\pi^{\text{inter}}(\mathbf{u}).$$

The subsequent definitions remain unchanged. One example of an intervention different from the randomized intervention considered in the rest of the paper is one held constant at a vector \mathbf{x}_0 :

$$\pi_{\mathbf{x}_0}^{\text{inter}}(\mathbf{u}) = \begin{cases} 1 & \text{for } \mathbf{u} = \mathbf{x}_0 \\ 0 & \text{o.w.} \end{cases}$$

A QII measure defined on the constant intervention as defined above, measures the influence of being different from a default, where the default is represented by \mathbf{x}_0 .

b) Difference Measure: A second generalization allows us to consider quantities of interest which are not real numbers. Consider, for example, the situation where the quantity of interest is an output probability distribution, as in the case in a randomized classifier. In this setting, a suitable measure for quantifying the distance between distributions can be used as a difference measure between the two quantities of interest. Examples of such difference measures include the KL-divergence [20] between distribution or distance metrics between vectors.

c) Transparency Schema: We now present a transparency schema that maps the space of transparency reports based on QII measures. It consists of the following elements:

- A *quantity of interest*, which captures the aspect of the system we wish to gain transparency into.
- An *intervention distribution*, which defines how a counterfactual distribution is constructed from the true distribution.
- A *difference measure*, which quantifies the difference between two quantities of interest.
- An *aggregation technique*, which combines marginal QII measures across different subsets of inputs (features).

For a given application, one has to appropriately instantiate this schema. We have described several instances of each schema element. The choices of the schema elements are guided by the particular causal question being posed. For instance, when the question is: “Which features are most important for group disparity?”, the natural quantity of interest

is a measure of group disparity, and the natural intervention distribution is using the prior as the question does not suggest a particular bias. On the other hand, when the question is: “Which features are most influential for person A’s classification as opposed to person B?”, a natural quantity of interest is person A’s classification, and a natural intervention distribution is the constant intervention using the features of person B. A thorough exploration of other points in this design space remains an important direction for future work.

V. ESTIMATION

While the model we propose offers several appealing properties, it faces several technical implementation issues. Several elements of our work require significant computational effort; in particular, both the probability that a change in feature state would cause a change in outcome, and the game-theoretic influence measures are difficult to compute exactly. In the following sections we discuss these issues and our proposed solutions.

A. Computing Power Indices

Computing the Shapley or Banzhaf values exactly is generally computationally intractable (see [21, Chapter 4] for a general overview); however, their probabilistic nature means that they can be well-approximated via random sampling. More formally, given a random variable X , suppose that we are interested in estimating some determined quantity $q(X)$ (say, $q(X)$ is the mean of X); we say that a random variable q^* is an ε - δ approximation of $q(X)$ if

$$\Pr[|q^* - q(X)| \geq \varepsilon] < \delta;$$

in other words, it is extremely likely that the difference between $q(X)$ and q^* is no more than ε . An ε - δ approximation scheme for $q(X)$ is an algorithm that for any $\varepsilon, \delta \in (0, 1)$ is able to output a random variable q^* that is an ε - δ approximation of $q(X)$, and runs in time polynomial in $\frac{1}{\varepsilon}, \log 1/\delta$.

[22] show that when $\langle N, v \rangle$ is a *simple* game (i.e. a game where $v(S) \in \{0, 1\}$ for all $S \subseteq N$), there exists an ε - δ approximation scheme for both the Banzhaf and Shapley values; that is, for $\phi \in \{\varphi, \beta\}$, we can guarantee that for any $\varepsilon, \delta > 0$, with probability $\geq 1 - \delta$, we output a value ϕ_i^* such that $|\phi_i^* - \phi_i| < \varepsilon$.

More generally, [23] observe that the number of i.i.d. samples needed in order to approximate the Shapley value and Banzhaf index is parametrized in $\Delta(v) = \max_{S \subseteq N} v(S) - \min_{S \subseteq N} v(S)$. Thus, if $\Delta(v)$ is a bounded value, then an ε - δ approximation exists. In our setting, coalitional values are always within the interval $[0, 1]$, which immediately implies the following theorem.

Theorem 10. *There exists an ε - δ approximation scheme for the Banzhaf and Shapley values in the QII setting.*

B. Estimating Q

Since we do not have access to the prior generating the data, we simply estimate it by observing the dataset itself. Recall that \mathcal{X} is the set of all possible user profiles; in this

case, a dataset is simply a multiset (i.e. possibly containing multiple copies of user profiles) contained in \mathcal{X} . Let \mathcal{D} be a finite multiset of \mathcal{X} , the input space. We estimate probabilities by computing sums over \mathcal{D} . For example, for a classifier c , the probability of $c(X) = 1$.

$$\hat{\mathbb{E}}_{\mathcal{D}}(c(X) = 1) = \frac{\sum_{\mathbf{x} \in \mathcal{D}} \mathbb{1}(c(\mathbf{x}) = 1)}{|\mathcal{D}|}. \quad (10)$$

Given a set of features $S \subseteq N$, let $\mathcal{D}|_S$ denote the elements of \mathcal{D} truncated to only the features in S . Then, the intervened probability can be estimated as follows:

$$\hat{\mathbb{E}}_{\mathcal{D}}(c(X_{-S}) = 1) = \frac{\sum_{\mathbf{u}_S \in \mathcal{D}|_S} \sum_{\mathbf{x} \in \mathcal{D}} \mathbb{1}(c(\mathbf{x}|_{N \setminus S} \mathbf{u}_S) = 1)}{|\mathcal{D}|^2}. \quad (11)$$

Similarly, the intervened probability on individual outcomes can be estimated as follows:

$$\hat{\mathbb{E}}_{\mathcal{D}}(c(X_{-S}) = 1 | X = \mathbf{x}) = \frac{\sum_{\mathbf{u}_S \in \mathcal{D}_S} \mathbb{1}(c(\mathbf{x}|_{N \setminus S} \mathbf{u}_S) = 1)}{|\mathcal{D}|}. \quad (12)$$

Finally, let us observe group disparity:

$$|\hat{\mathbb{E}}_{\mathcal{D}}(c(X_{-S}) = 1 | X \in \mathcal{Y}) - \hat{\mathbb{E}}_{\mathcal{D}}(c(X_{-S}) = 1 | X \notin \mathcal{Y})|$$

The term $\hat{\mathbb{E}}_{\mathcal{D}}(c(X_{-S}) = 1 | X \in \mathcal{Y})$ equals

$$\frac{1}{|\mathcal{Y}|} \sum_{\mathbf{x} \in \mathcal{Y}} \sum_{\mathbf{u}_S \in \mathcal{D}_S} \mathbb{1}(c(\mathbf{x}|_{N \setminus S} \mathbf{u}_S) = 1),$$

Thus group disparity can be written as:

$$\begin{aligned} & \left| \frac{1}{|\mathcal{Y}|} \sum_{\mathbf{x} \in \mathcal{Y}} \sum_{\mathbf{u}_S \in \mathcal{D}_S} \mathbb{1}(c(\mathbf{x}|_{N \setminus S} \mathbf{u}_S) = 1) \right. \\ & \left. - \frac{1}{|\mathcal{D} \setminus \mathcal{Y}|} \sum_{\mathbf{x} \in \mathcal{D} \setminus \mathcal{Y}} \sum_{\mathbf{u}_S \in \mathcal{D}_S} \mathbb{1}(c(\mathbf{x}|_{N \setminus S} \mathbf{u}_S) = 1) \right|. \end{aligned} \quad (13)$$

We write $\hat{Q}_{\text{disp}}^{\mathcal{Y}}(S)$ to denote (13).

If \mathcal{D} is large, these sums cannot be computed efficiently. Therefore, we approximate the sums by sampling from the dataset \mathcal{D} . It is possible to show using the Hoeffding bound [24], partial sums of n random variables X_i , within a bound Δ , can be well-approximated with the following probabilistic bound:

$$\Pr \left(\left| \frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}X_i) \right| \geq \varepsilon \right) \leq 2 \exp \left(\frac{-2n\varepsilon^2}{\Delta} \right)$$

Since all the samples of measures discussed in the paper are bounded within the interval $[0, 1]$, we admit an ε - δ approximation scheme where the number of samples n can be chosen to be greater than $\log(2/\delta)/2\varepsilon^2$. Note that these bounds are independent of the size of the dataset. Therefore, given an efficient sampler, these quantities of interest can be approximated efficiently even for large datasets.

VI. PRIVATE TRANSPARENCY REPORTS

One important concern is that releasing influence measures estimated from a dataset might leak information about individual users; our goal is providing accurate transparency reports, without compromising individual users' private data. To mitigate this concern, we add noise to make the measures differentially private. We show that the sensitivities of the QII measures considered in this paper are very low and therefore very little noise needs to be added to achieve differential privacy.

The *sensitivity* of a function is a key parameter in ensuring that it is differentially private; it is simply the worst-case change in its value, assuming that we change a single data point in our dataset. Given some function f over datasets, we define the sensitivity of a function f with respect to a dataset \mathcal{D} , denoted by $\Delta f(\mathcal{D})$ as

$$\max_{\mathcal{D}'} |f(\mathcal{D}) - f(\mathcal{D}')|$$

where \mathcal{D} and \mathcal{D}' differ by at most one instance. We use the shorthand Δf when \mathcal{D} is clear from the context.

In order to not leak information about the users used to compute the influence of an input, we use the standard Laplace Mechanism [8] and make the influence measure differentially private. The amount of noise required depends on the sensitivity of the influence measure. We show that the influence measure has low sensitivity for the individuals used to sample inputs. Further, due to a result from [9] (and stated in [25]), sampling amplifies the privacy of the computed statistic, allowing us to achieve high privacy with minimal noise addition.

The standard technique for making any function differentially private is to add Laplace noise calibrated to the sensitivity of the function:

Theorem 11 ([8]). *For any function f from datasets to \mathbb{R} , the mechanism \mathcal{K}_f that adds independently generated noise with distribution $\text{Lap}(\Delta f(\mathcal{D})/\epsilon)$ to the k output enjoys ϵ -differential privacy.*

Since each of the quantities of interest aggregate over a large number of instances, the sensitivity of each function is very low.

Theorem 12. *Given a dataset \mathcal{D} ,*

- 1) $\Delta \hat{\mathbb{E}}_{\mathcal{D}}(c(X) = 1) = \frac{1}{|\mathcal{D}|}$
- 2) $\Delta \hat{\mathbb{E}}_{\mathcal{D}}(c(X_{-S}) = 1) \leq \frac{2}{|\mathcal{D}|}$
- 3) $\Delta \hat{\mathbb{E}}_{\mathcal{D}}(c(X_{-S}) = 1 | X = \mathbf{x}) = \frac{1}{|\mathcal{D}|}$
- 4) $\hat{Q}_{\text{disp}}^{\mathcal{Y}}(S) \leq \max \left\{ \frac{1}{|\mathcal{D} \cap \mathcal{Y}|}, \frac{1}{|\mathcal{D} \setminus \mathcal{Y}|} \right\}$

Proof. We examine some cases here. In Equation 10, if two datasets differ by one instance, then at most one term of the summation will differ. Since each term can only be either 0 or 1, the sensitivity of the function is

$$\Delta \hat{\mathbb{E}}_{\mathcal{D}}(c(X) = 1) = \left| \frac{0}{|\mathcal{D}|} - \frac{1}{|\mathcal{D}|} \right| = \frac{1}{|\mathcal{D}|}.$$

Similarly, in Equation 11, an instance appears $2|\mathcal{D}| - 1$ times, once each for the inner summation and the outer summation, and therefore, the sensitivity of the function is

$$\Delta \hat{\mathbb{E}}_{\mathcal{D}}(c(X_{-S}) = 1) = \frac{2|\mathcal{D}| - 1}{|\mathcal{D}|^2} \leq \frac{2}{|\mathcal{D}|}.$$

For individual outcomes (Equation (12)), similarly, only one term of the summation can differ. Therefore, the sensitivity of (12) is $1/|\mathcal{D}|$.

Finally, we observe that a change in a single element \mathbf{x}' of \mathcal{D} will cause a change of at most $\frac{1}{|\mathcal{D} \cap \mathcal{Y}|}$ if $\mathbf{x}' \in \mathcal{D} \cap \mathcal{Y}$, or of at most $\frac{1}{|\mathcal{D} \setminus \mathcal{Y}|}$ if $\mathbf{x}' \in \mathcal{D} \setminus \mathcal{Y}$. Thus, the maximal change to (13) is at most $\max \left\{ \frac{1}{|\mathcal{Y}|}, \frac{1}{|\mathcal{D} \setminus \mathcal{Y}|} \right\}$. \square

While the sensitivity of most quantities of interest is low (at most a $\frac{2}{|\mathcal{D}|}$), $\hat{Q}_{\text{disp}}^{\mathcal{Y}}(S)$ can be quite high when $|\mathcal{Y}|$ is either very small or very large. This makes intuitive sense: if \mathcal{Y} is a very small minority, then any changes to its members are easily detected; similarly, if \mathcal{Y} is a vast majority, then changes to protected minorities may be easily detected.

We observe that the quantities of interest which exhibit low sensitivity will have low influence sensitivity as well: for example, the local influence of S is $\mathbb{1}(c(\mathbf{x}) = 1) - \hat{\mathbb{E}}_{\mathcal{D}}(c(X_{-S}) = 1) | X = \mathbf{x})$; changing any $\mathbf{x}' \in \mathcal{D}$ (where $\mathbf{x}' \neq \mathbf{x}$) will result in a change of at most $\frac{1}{|\mathcal{D}|}$ to the local influence.

Finally, since the Shapley and Banzhaf indices are normalized sums of the differences of the set influence functions, we can show that if an influence function ι has sensitivity $\Delta\iota$, then the sensitivity of the indices are at most $2\Delta\iota$.

To conclude, all of the QII measures discussed above (except for group parity) have a sensitivity of $\frac{\alpha}{|\mathcal{D}|}$, with α being a small constant. To ensure differential privacy, we need only need add noise with a Laplacian distribution $\text{Lap}(k/|\mathcal{D}|)$ to achieve 1-differential privacy.

Further, it is known that sampling amplifies differential privacy.

Theorem 13 ([9], [25]). *If \mathcal{A} is 1-differentially private, then for any $\epsilon \in (0, 1)$, $\mathcal{A}'(\epsilon)$ is 2ϵ -differentially private, where $\mathcal{A}'(\epsilon)$ is obtained by sampling an ϵ fraction of inputs and then running \mathcal{A} on the sample.*

Therefore, our approach of sampling instances from \mathcal{D} to speed up computation has the additional benefit of ensuring that our computation is private.

Table I contains a summary of all QII measures defined in this paper, and their sensitivity.

VII. EXPERIMENTAL EVALUATION

We demonstrate the utility of the QII framework by developing two simple machine learning applications on real datasets. Using these applications, we first argue, in Section VII-A, the need for causal measurement by empirically demonstrating that in the presence of correlated inputs, observational measures are not informative in identifying which inputs were

actually used. In Section VII-B, we illustrate the distinction between different quantities of interest on which Unary QII can be computed. We also illustrate the effect of discrimination on the QII measure. In Section VII-C, we analyze transparency reports of three individuals to demonstrate how Marginal QII can provide insights into individuals' classification outcomes. Finally, we analyze the loss in utility due to the use of differential privacy, and provide execution times for generating transparency reports using our prototype implementation.

We use the following datasets in our experiments:

- **adult** [10]: This standard machine learning benchmark dataset is a subset of US census data that classifies the income of individuals, and contains factors such as age, race, gender, marital status and other socio-economic parameters. We use this dataset to train a classifier that predicts the income of individuals from other parameters. Such a classifier could potentially be used to assist credit decisions.
- **arrests** [11]: The National Longitudinal Surveys are a set of surveys conducted by the Bureau of Labor Statistics of the United States. In particular, we use the National Longitudinal Survey of Youth 1997 which is a survey of young men and women born in the years 1980-84. Respondents were ages 12-17 when first interviewed in 1997 and were subsequently interviewed every year till 2013. The survey covers various aspects of an individual's life such as medical history, criminal records and economic parameters. From this dataset, we extract the following features: age, gender, race, region, history of drug use, history of smoking, and history of arrests. We use this data to train a classifier that predicts history of arrests to aid in predictive policing, where socio-economic factors are used to decide whether individuals should receive a visit from the police. This application is inspired by a similar application in [26].

The two applications described above are hypothetical examples of decision-making aided by machine learning that use potentially sensitive socio-economic data about individuals, and not real systems that are currently in use. We use these classifiers to illustrate the subtle causal questions that our QII measures can answer.

We use the following standard machine learning classifiers in our dataset: Logistic Regression, SVM with a radial basis function kernel, Decision Tree, and Gradient Boosted Decision Trees. Bishop's machine learning text [27] is an excellent resource for an introduction to these classifiers. While Logistic Regression is a linear classifier, the other three are nonlinear and can potentially learn very complex models. All our experiments are implemented in Python with the numpy library, and the scikit-learn machine learning toolkit, and run on an Intel i7 computer with 4 GB of memory.

A. Comparison with Observational Measures

In the presence of correlated inputs, observational measures often cannot identify which inputs were causally influential. To illustrate this phenomena on real datasets, we train two

Name	Notation	Quantity of Interest	Sensitivity
QII on Individual Outcomes (3)	$\iota_{\text{ind}}(S)$	Positive Classification of an Individual	$1/ \mathcal{D} $
QII on Actual Individual Outcomes (4)	$\iota_{\text{ind-act}}(S)$	Actual Classification of an Individual	$1/ \mathcal{D} $
Average QII (5)	$\iota_{\text{ind-avg}}(S)$	Average Actual Classification	$2/ \mathcal{D} $
QII on Group Outcomes (6)	$\iota_{\text{grp}}^{\mathcal{Y}}(S)$	Positive Classification for a Group	$2/ \mathcal{D} \cap \mathcal{Y} $
QII on Group Disparity (8)	$\iota_{\text{disp}}^{\mathcal{Y}}(S)$	Difference in classification rates among groups	$2 \max(1/ \mathcal{D} \setminus \mathcal{Y} , 1/ \mathcal{D} \cap \mathcal{Y})$

TABLE I: A summary of the QII measures defined in the paper

classifiers: (A) where gender is provided as an actual input, and (B) where gender is not provided as an input. For classifier (B), clearly the input *Gender* has no effect and any correlation between the outcome and gender is caused via inference from other inputs. In Table II, for both the adult and the arrests dataset, we compute the following observational measures: Mutual Information (MI), Jaccard Index (Jaccard), Pearson Correlation (corr), and the Disparate Impact Ratio (disp) to measure the similarity between Gender and the classifiers outcome. We also measure the QII of Gender on outcome. We observe that in many scenarios the observational quantities do not change, or sometimes increase, from classifier A to classifier B, when gender is removed as an actual input to the classifier. On the other hand, if the outcome of the classifier does not depend on the input *Gender*, then the QII is guaranteed to be zero.

B. Unary QII Measures

In Figure 2, we illustrate the use of different Unary QII measures. Figures 2a, and 2b, show the Average QII measure (Equation 5) computed for features of a decision forest classifier. For the income classifier trained on the adult dataset, the feature with highest influence is *Marital Status*, followed by *Occupation*, *Relationship* and *Capital Gain*. Sensitive features such as *Gender* and *Race* have relatively lower influence. For the predictive policing classifier trained on the arrests dataset, the most influential input is *Drug History*, followed by *Gender*, and *Smoking History*. We observe that influence on outcomes may be different from influence on group disparity.

QII on group disparity: Figures 2c, 2d show influences of features on group disparity for two different settings. The figure on the left shows the influence of features on group disparity by Gender in the adult dataset; the figure on the right shows the influence of group disparity by Race in the arrests dataset. For the income classifier trained on the adult dataset, we observe that most inputs have negative influence on group disparity; randomly intervening on most inputs would lead to a reduction in group disparity. In other words, a classifier that did not use these inputs would be fairer. Interestingly, in this classifier, marital status and not sex has the highest influence on group disparity by sex.

For the arrests dataset, most inputs have the effect of increasing group disparity if randomly intervened on. In particular, *Drug history* has the highest positive influence on disparity in arrests. Although Drug history is correlated with race, using it reduces disparate impact by race, i.e. makes fairer decisions.

In both examples, features correlated with the sensitive attribute are the most influential for group disparity according to the sensitive attribute instead of the sensitive attribute itself. It is in this sense that QII measures can identify proxy variables that cause associations between outcomes and sensitive attributes.

QII with artificial discrimination: We simulate discrimination using an artificial experiment. We first randomly assign ZIP codes to individuals in our dataset. Then to simulate systematic bias, we make an f fraction of the ZIP codes discriminatory in the following sense: All individuals in the protected set are automatically assigned a negative classification outcome. We then study the change in the influence of features as we increase f . Figure 3a, shows that the influence of *Gender* increases almost linearly with f . Recall that *Marital Status* was the most influential feature for this classifier without any added discrimination. As f increases, the importance of *Marital Status* decreases as expected, since the number of individuals for whom *Marital Status* is pivotal decreases.

C. Personalized Transparency Reports

To illustrate the utility of personalized transparency reports, we study the classification of individuals who received potentially unexpected outcomes. For the personalized transparency reports, we use decision forests.

The influence measure that we employ is the Shapley value, with the underlying cooperative game defined over the local influence Q . In more detail, $v(S) = \iota^{Q_A}(S)$, with Q_A being $\mathbb{E}[c(\cdot) = 1 | X = \mathbf{x}]$; that is, the marginal contribution of $i \in N$ to S is given by $m_i(S) = \mathbb{E}[c(X_{-S}) = 1 | X = \mathbf{x}] - \mathbb{E}[c(X_{-S \cup \{i\}}) = 1 | X = \mathbf{x}]$.

We emphasize that some features may have a negative Shapley value; this should be interpreted as follows: a feature with a high positive Shapley value often increases the certainty that the classification outcome is 1, whereas a feature whose Shapley value is negative is one that increases the certainty that the classification outcome would be zero.

Mr. X: The first example is of an individual from the adult dataset, who we refer to as Mr. X, and is described in Figure 4a. He is deemed to be a low income individual, by an income classifier learned from the data. This result may be surprising to him: he reports high capital gains (\$14k), and only 2.1% of people with capital gains higher than \$10k are reported as low income. In fact, he might be led to believe that his classification may be a result of his ethnicity or country of origin. Examining his transparency report in Figure 4b, however, we find that the the most influential features that led

		logistic		kernel svm		decision tree		random forest	
		adult	arrests	adult	arrests	adult	arrests	adult	arrests
MI	A	0.045	0.049	0.046	0.047	0.043	0.054	0.044	0.053
MI	B	0.043	0.050	0.044	0.053	0.042	0.051	0.043	0.052
Jaccard	A	0.501	0.619	0.500	0.612	0.501	0.614	0.501	0.620
Jaccard	B	0.500	0.611	0.501	0.615	0.500	0.614	0.501	0.617
corr	A	0.218	0.265	0.220	0.247	0.213	0.262	0.218	0.262
corr	B	0.215	0.253	0.218	0.260	0.215	0.257	0.215	0.259
disp	A	0.286	0.298	0.377	0.033	0.302	0.335	0.315	0.223
disp	B	0.295	0.301	0.312	0.096	0.377	0.228	0.302	0.129
QII	A	0.036	0.135	0.044	0.149	0.023	0.116	0.012	0.109
QII	B	0	0	0	0	0	0	0	0

TABLE II: Comparison of QII with associative measures. For 4 different classifiers, we compute metrics such as Mutual Information (MI), Jaccard Index (JI), Pearson Correlation (corr), Group Disparity (disp) and Average QII between Gender and the outcome of the learned classifier. Each metric is computed in two situations: (A) when Gender is provided as an input to the classifier, and (B) when Gender is not provided as an input to the classifier.

to his negative classification were Marital Status, Relationship and Education.

Mr. Y: The second example, to whom we refer as Mr. Y (Figure 5), has even higher capital gains than Mr. X. Mr. Y is a 27 year old, with only Preschool education, and is engaged in fishing. Examination of the transparency report reveals that the most influential factor for negative classification for Mr. Y is his Occupation. Interestingly, his low level of education is not considered very important by this classifier.

Mr. Z: The third example, who we refer to as Mr. Z (Figure 6) is from the arrests dataset. History of drug use and smoking are both strong indicators of arrests. However, Mr. X received positive classification by this classifier even without any history of drug use or smoking. On examining his classifier, it appears that race, age and gender were most influential in determining his outcome. In other words, the classifier that we train for this dataset (a decision forest) has picked up on the correlations between race (Black), and age (born in 1984) to infer that this individual is likely to engage in criminal activity. Indeed, our interventional approach indicates that this is not a mere correlation effect: race is actively being used by this classifier to determine outcomes. Of course, in this instance, we have explicitly offered the race parameter to our classifier as a viable feature. However, our influence measure is able to pick up on this fact, and alert us of the problematic behavior of the underlying classifier. More generally, this example illustrates a concern with the black box use of machine learning which can lead to unfavorable outcomes for individuals.

D. Differential Privacy

Most QII measures considered in this paper have very low sensitivity, and therefore can be made differentially private with negligible loss in utility. However, recall that the sensitivity of influence measure on group disparity $\iota_{\text{disp}}^{\mathcal{Y}}$ depends on the size of the protected group in the dataset \mathcal{D} as follows:

$$\iota_{\text{disp}}^{\mathcal{Y}} = 2 \max \left(\frac{1}{|\mathcal{D} \setminus \mathcal{Y}|}, \frac{1}{|\mathcal{D} \cap \mathcal{Y}|} \right)$$

For sufficiently small minority groups, a large amount of noise might be required to ensure differential privacy, leading

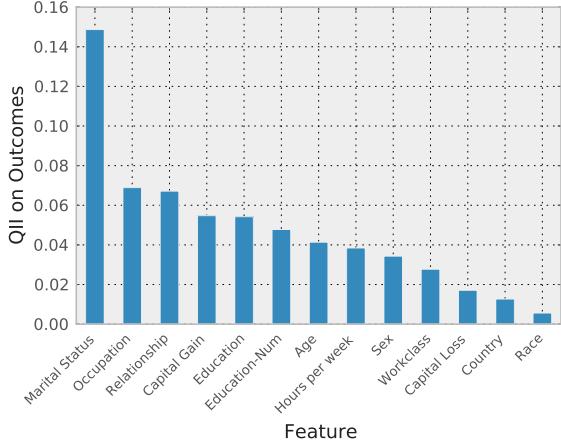
to a loss in utility of the QII measure. To estimate the loss in utility, we set a noise of 0.005 as the threshold of noise at which the measure is no longer useful, and then compute fraction of times noise crosses that threshold when Laplacian noise is added at $\epsilon = 1$. The results of this experiment are as follows:

\mathcal{Y}	Count	Loss in Utility
Race: White	27816	2.97×10^{-14}
Race: Black	3124	5.41×10^{-14}
Race: Asian-Pac-Islander	1039	6.14×10^{-5}
Race: Amer-Indian-Eskimo	311	0.08
Race: Other	271	0.13
Gender: Male	21790	3.3×10^{-47}
Gender: Female	10771	3.3×10^{-47}

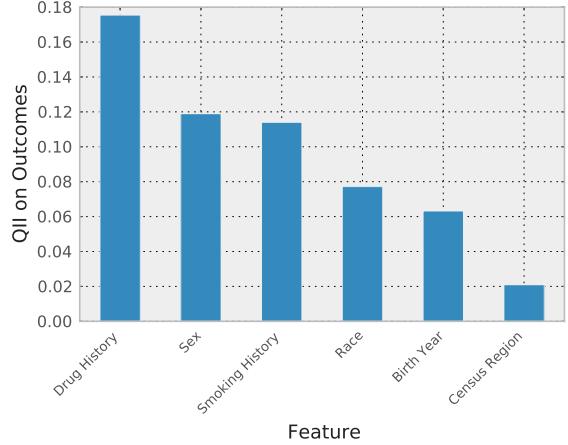
We note that for most reasonably sized groups, the loss in utility is negligible. However, the Asian-Pac-Islander, and the Amer-Indian-Eskimo racial groups are underrepresented in this dataset. For these groups, the QII on Group Disparity estimate needs to be very noisy to protect privacy.

E. Performance

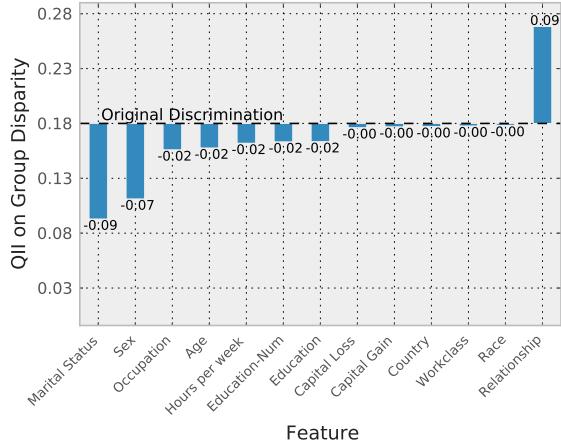
We report runtimes of our prototype for generating transparency reports on the adult dataset. Recall from Section VI that we approximate QII measures by computing sums over samples of the dataset. According to the Hoeffding bound to derive an (ϵ, δ) estimate of a QII measure, at $\epsilon = 0.01$, and $n = 37000$ samples, $\delta = 2 \exp(-n\epsilon^2) < 0.05$ is an upper bound on the probability of the output being off by ϵ . Table III shows the runtimes of four different QII computations, for 37000 samples each. The runtimes of all algorithms except for kernel SVM are fast enough to allow real-time feedback for machine learning application developers. Evaluating QII metrics for Kernel SVMs is much slower than the other metrics because each call to the SVM classifier is very computationally intensive due to a large number of distance computations that it entails. We expect that these runtimes can be optimized significantly. We present them as proof of tractability.



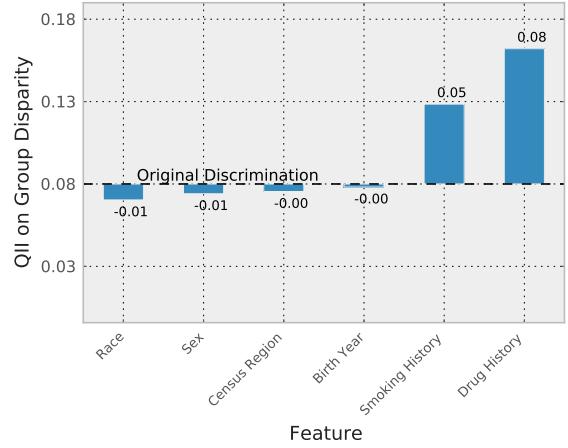
(a) QII of inputs on Outcomes for the `adult` dataset



(b) QII of inputs on Outcomes for the `arrests` dataset



(c) QII of Inputs on Group Disparity by Sex in the `adult` dataset



(d) Influence on Group Disparity by Race in the `arrests` dataset

Fig. 2: QII measures for the `adult` and `arrests` datasets

	logistic	kernel-svm	decision-tree	decision-forest
QII on Group Disparity	0.56	234.93	0.57	0.73
Average QII	0.85	322.82	0.77	1.12
QII on Individual Outcomes (Shapley)	6.85	2522.3	7.78	9.30
QII on Individual Outcomes (Banzhaf)	6.77	2413.3	7.64	10.34

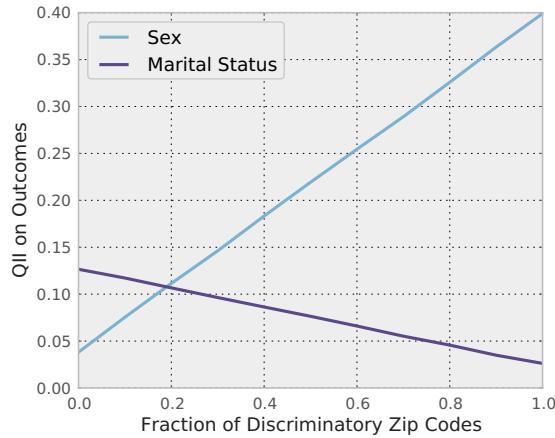
TABLE III: Runtimes in seconds for transparency report computation

VIII. DISCUSSION

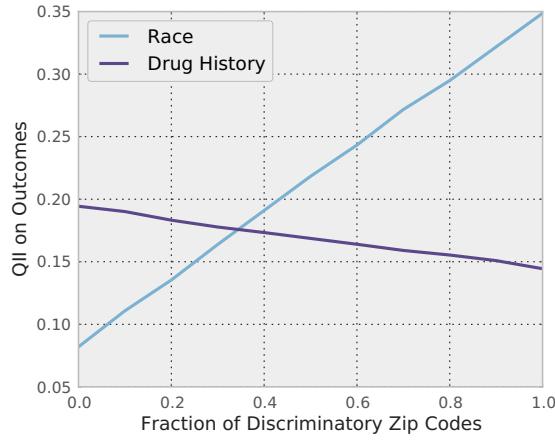
A. Probabilistic Interpretation of Power Indices

In order to quantitatively measure the influence of data inputs on classification outcomes, we propose causal interventions on sets of features; as we argue in Section III, the aggregate marginal influence of i for different subsets of features is a natural quantity representing its influence. In order to aggregate the various influences i has on the outcome, it is natural to define some probability distribution over (or equivalently, a weighted sum of) subsets of $N \setminus \{i\}$, where $\Pr[S]$ represents the probability of measuring the marginal contribution of i to S ; $\Pr[S]$ yields a value $\sum_{S \subseteq N \setminus \{i\}} m_i(S)$.

For the Banzhaf index, we have $\Pr[S] = \frac{1}{2^{n-1}}$, the Shapley value has $\Pr[S] = \frac{k!(n-k-1)!}{n!}$ (here, $|S| = k$), and the Deegan-Packel Index selects minimal winning coalitions uniformly at random. These choices of values for $\Pr[S]$ are based on some natural assumptions on the way that players (features) interact, but they are by no means exhaustive. One can define other sampling methods that are more appropriate for the model at hand; for example, it is entirely possible that the only interventions that are possible in a certain setting are of size $\leq k+1$, it is reasonable to aggregate the marginal influence



(a) Change in QII of inputs as discrimination by Zip Code increases in the adult dataset



(b) Change in QII of inputs as discrimination by Zip Code increases in the arrests dataset

Fig. 3: The effect of discrimination on QII.

of i over sets of size $\leq k$, i.e.

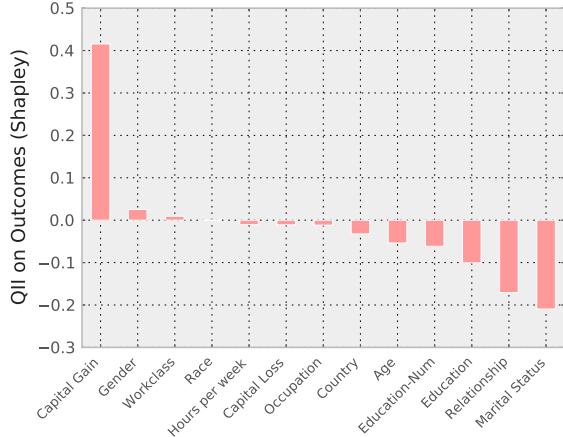
$$\Pr[S] = \begin{cases} \frac{1}{\binom{n-1}{|S|-1}} & \text{if } |S| \leq k \\ 0 & \text{otherwise.} \end{cases}$$

The key point here is that one must define *some* aggregation method, and that choice reflects some normative approach on how (and which) marginal contributions are considered. The Shapley and Banzhaf indices do have some highly desirable properties, but they are, first and foremost, *a-priori* measures of influence. That is, they do not factor in any assumptions on what interventions are possible or desirable.

One natural candidate for a probability distribution over S is some natural extension of the prior distribution over the dataset; for example, if all features are binary, one can identify a set with a feature vector (namely by identifying each $S \subseteq N$ with its indicator vector), and set $\Pr[S] = \pi(S)$ for all $S \subseteq N$.

Age	23
Workclass	Private
Education	11th
Education-Num	7
Marital Status	Never-married
Occupation	Craft-repair
Relationship	Own-child
Race	Asian-Pac-Islander
Gender	Male
Capital Gain	14344
Capital Loss	0
Hours per week	40
Country	Vietnam

(a) Mr. X's profile



(b) Transparency report for Mr. X's negative classification

Fig. 4: Mr. X

If features are not binary, then there is no canonical way to transition from the data prior to a prior over subsets of features.

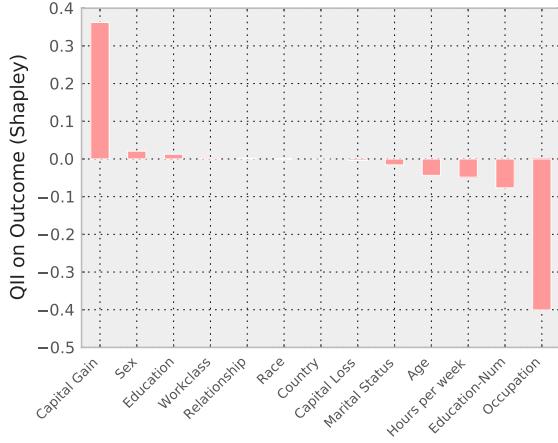
B. Fairness

Due to the widespread and black box use of machine learning in aiding decision making, there is a legitimate concern of algorithms introducing and perpetuating social harms such as racial discrimination [28], [6]. As a result, the algorithmic foundations of fairness in personal information processing systems have received significant attention recently [29], [30], [31], [12], [32]. While many of the algorithmic approaches [29], [31], [32] have focused on group parity as a metric for achieving fairness in classification, Dwork et al. [12] argue that group parity is insufficient as a basis for fairness, and propose a similarity-based approach which prescribes that similar individuals should receive similar classification outcomes. However, this approach requires a similarity metric for individuals which is often subjective and difficult to construct.

QII does not suggest any normative definition of fairness. Instead, we view QII as a diagnostic tool to aid fine-grained fairness determinations. In fact, QII can be used in the spirit of the similarity based definition of [12]. By comparing the personalized privacy reports of individuals who are *perceived*

Age	27
Workclass	Private
Education	Preschool
Education-Num	1
Marital Status	Married-civ-spouse
Occupation	Farming-fishing
Relationship	Other-relative
Race	White
Gender	Male
Capital Gain	41310
Capital Loss	0
Hours per week	24
Country	Mexico

(a) Mr. Y's profile



(b) Transparency report for Mr. Y's negative classification

Fig. 5: Mr. Y.

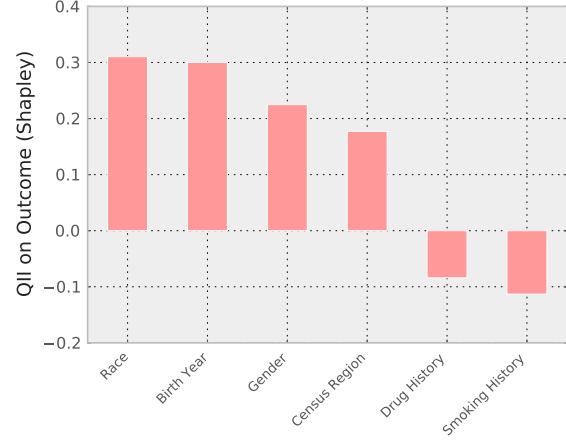
to be similar but received different classification outcomes, and identifying the inputs which were used by the classifier to provide different outcomes. Additionally, when group parity is used as a criteria for fairness, QII can identify the features that lead to group disparity, thereby identifying features being used by a classifier as a proxy for sensitive attributes.

The determination of whether using certain proxies for sensitive attributes is discriminatory is often a task-specific normative judgment. For example, using standardized test scores (e.g., SAT scores) for admissions decisions is by and large accepted, although SAT scores may be a proxy for several protected attributes. In fact, several universities have recently announced that they will not use SAT scores for admissions citing this reason [33], [34]. Our goal is not to provide such normative judgments. Rather we seek to provide fine-grained transparency into input usage (e.g., what's the extent to which SAT scores influence decisions), which is useful to make determinations of discrimination from a specific normative position.

Finally, we note that an interesting question is whether providing a sensitive attribute as an input to a classifier is fundamentally discriminatory behavior, even if QII can show that the sensitive input has no significant impact on the

Birth Year	1984
Drug History	None
Smoking History	None
Census Region	West
Race	Black
Gender	Male

(a) Mr. Z's profile



(b) Transparency report for Mr. Z's positive classification

Fig. 6: Mr. Z.

outcome. Our view is that this is a policy question and different legal frameworks might take different viewpoints on it. At a technical level, from the standpoint of information use, the two situations are identical: the sensitive input is not really used although it is supplied. However, the very fact that it was supplied might be indicative of an intent to discriminate even if that intended goal was not achieved. No matter what the policy decision is on this question, QII remains a useful diagnostic tool for discrimination because of the presence of proxy variables as described earlier.

IX. RELATED WORK

A. Quantitative Causal Measures

Causal models and probabilistic interventions have been used in a few other settings. While the form of the interventions in some of these settings may be very similar, our generalization to account for different quantities of interests enables us to reason about a large class of transparency queries for data analytics systems ranging from classification outcomes of individuals to disparity among groups. Further, the notion of marginal contribution which we use to compute responsibility does not appear in this line of prior work.

Janzing et al. [35] use interventions to assess the causal importance of relations between variables in causal graphs; in order to assess the causal effect of a relation between two variables, $X \rightarrow Y$ (assuming that both take on specific values $X = x$ and $Y = y$), a new causal model is constructed, where the value of X is replaced with a prior over the possible values of X . The influence of the causal relation is defined as the KL-

Divergence of the joint distribution of all the variables in the two causal models with and without the value of X replaced. The approach of the intervening with a random value from the prior is similar to our approach of constructing X_{-S} .

Independently, there has been considerable work in the machine learning community to define importance metrics for variables, mainly for the purpose of feature selection (see [36] for a comprehensive overview). One important metric is called Permutation Importance [37], which measures the importance of a feature towards classification by randomly permuting the values of the feature and then computing the difference of classification accuracies before and after the permutation. Replacing a feature with a random permutation can be viewed as a sampling the feature independently from the prior.

There exists extensive literature on establishing causal relations, as opposed to quantifying them. Prominently, Pearl's work [38] provides a mathematical foundation for causal reasoning and inference. In [39], Tian and Pearl discuss measures of causal strength for individual binary inputs and outputs in a probabilistic setting. Another thread of work by Halpern and Pearl discusses actual causation [40], which is extended in [41] to derive a measure of responsibility as degree of causality. In [41], Chockler and Halpern define the responsibility of a variable X to an outcome as the amount of change required in order to make X the counterfactual cause. As we discuss in Appendix A-B, the Deegan-Packel index is strongly related to causal responsibility.

B. Quantitative Information Flow

One can think of our results as a causal alternative to *quantitative information flow*. Quantitative information flow is a broad class of metrics that quantify the information leaked by a process by comparing the *information* contained before and after observing the outcome of the process. Quantitative Information Flow traces its information-theoretic roots to the work of Shannon [42] and Rényi [43]. Recent works have proposed measures for quantifying the security of information by measuring the amount of information leaked from inputs to outputs by certain variables; we point the reader to [44] for an overview, and to [45] for an exposition on information theory. Quantitative Information Flow is concerned with information leaks and therefore needs to account for correlations between inputs that may lead to leakage. The dual problem of transparency, on the other hand, requires us to destroy correlations while analyzing the outcomes of a system to identify the causal paths for information leakage.

C. Interpretable Machine Learning

An orthogonal approach to adding interpretability to machine learning is to constrain the choice of models to those that are interpretable by design. This can either proceed through regularization techniques such as Lasso [46] that attempt to pick a small subset of the most important features, or by using models that structurally match human reasoning such as Bayesian Rule Lists [47], Supersparse Linear Integer Models [48], or Probabilistic Scaling [49]. Since the choice

of models in this approach is restricted, a loss in predictive accuracy is a concern, and therefore, the central focus in this line of work is the minimization of the loss in accuracy while maintaining interpretability. On the other hand, our approach to interpretability is forensic. We add interpretability to machine learning models after they have been learnt. As a result, our approach does not constrain the choice of models that can be used.

D. Experimentation on Web Services

There is an emerging body of work on systematic experimentation to enhance transparency into Web services such as targeted advertising [50], [51], [52], [53], [54]. The setting in this line of work is different since they have restricted access to the analytics systems through publicly available interfaces. As a result they only have partial control of inputs, partial observability of outputs, and little or no knowledge of input distributions. The intended use of these experiments is to enable external oversight into Web services without any cooperation. Our framework is more appropriate for a transparency mechanism where an entity proactively publishes transparency reports for individuals and groups. Our framework is also appropriate for use as an internal or external oversight tool with access to mechanisms with control and knowledge of input distributions, thereby forming a basis for testing.

E. Game-Theoretic Influence Measures

Recent years have seen game-theoretic influence measures used in various settings. Datta et al. [55] also define a measure for quantifying feature influence in classification tasks. Their measure does not account for the prior on the data, nor does it use interventions that break correlations between sets of features. In the terminology of this paper, the quantity of interest used by [55] is the ability of changing the outcome by changing the state of a feature. This work greatly extends and generalizes the concepts presented in [55], by both accounting for interventions on sets, and by generalizing the notion of influence to include a wide range of system behaviors, such as group disparity, group outcomes and individual outcomes.

Game theoretic measures have been used by various research disciplines to measure influence. Indeed, such measures are relevant whenever one is interested in measuring the marginal contribution of variables, and when sets of variables are able to cause some measurable effect. Lindelauf et al. [56] and Michalak et al. [57] use game theoretic influence measures on graph-based games in order to identify key members of terrorist networks. Del Pozo et al. [58] and Michalak et al. [59] use similar ideas for identifying important members of large social networks, providing scalable algorithms for influence computation. Bork et al. [60] use the Shapley value to assign importance to protein interactions in large, complex biological interaction networks; Keinan et al. [61] employ the Shapley value in order to measure causal effects in neurophysical models. The novelty in our use of the game theoretic power indices lies in the conception of a cooperative game via a valuation function $\iota(S)$, defined by an randomized intervention

on inputs S . Such an intervention breaks correlations and allows us to compute marginal causal influences on a wide range of system behaviors.

X. CONCLUSION & FUTURE WORK

In this paper, we present QII, a general family of metrics for quantifying the influence of inputs in systems that process personal information. In particular, QII lends insights into the behavior of opaque machine learning algorithms by allowing us to answer a wide class of transparency queries ranging from influence on individual causal outcomes to influence on disparate impact. To achieve this, QII breaks correlations between inputs to allow causal reasoning, and computes the marginal influence of inputs in situations where inputs cannot affect outcomes alone. Also, we demonstrate that QII can be efficiently approximated, and can be made differentially private with negligible noise addition in many cases.

An immediate next step in this line of work is to explore adoption strategies in the many areas that use personal information to aid decision making. Areas such as healthcare [3], predictive policing [1], education [4], and defense [5] all have a particularly acute need for transparency in their decision making. It is likely that specific applications will guide us in our choice of a QII metric that is appropriate for that scenario, which includes a choice for our game-theoretic power index.

We have not considered situations where inputs do not have well understood semantics. Such situations arise often in settings such as image or speech recognition, and automated video surveillance. With the proliferation of immense processing power, complex machine learning models such as deep neural networks have become ubiquitous in these domains. Defining transparency and developing analysis techniques in such settings is important future work.

REFERENCES

- [1] W. L. Perry, B. McInnis, C. C. Price, S. C. Smith, and J. S. Hollywood, *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations*. RAND Corporation, 2013.
- [2] T. Alloway, "Big data: Credit where credits due," <http://www.ft.com/cms/s/0/7933792e-a2e6-11e4-9e06-00144feab7de.html>.
- [3] T. B. Murdoch and A. S. Detsky, "The inevitable application of big data to health care," <http://jama.jamanetwork.com/article.aspx?articleid=1674245>.
- [4] "Big data in education," <https://www.edx.org/course/big-data-education-teacherscollege-bde1x>.
- [5] "Big data in government, defense and homeland security 2015 - 2020," <http://www.prnewswire.com/news-releases/big-data-in-government-defense-and-homeland-security-2015--2020.html>.
- [6] J. Podesta, P. Pritzker, E. Moniz, J. Holdern, and J. Zients, "Big data: Seizing opportunities, preserving values," Executive Office of the President - The White House, Tech. Rep., May 2014.
- [7] "E.G. Griggs v. Duke Power Co., 401 U.S. 424, 91 S. Ct. 849, 28 L. Ed. 2d 158 (1977)."
- [8] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proceedings of the Third Conference on Theory of Cryptography*, ser. TCC'06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 265–284. [Online]. Available: http://dx.doi.org/10.1007/11681878_14
- [9] S. Kasiviswanathan, H. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What can we learn privately?" in *Proceedings of the 49th IEEE Symposium on Foundations of Computer Science (FOCS 2008)*, Oct 2008, pp. 531–540.
- [10] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: <http://archive.ics.uci.edu/ml>
- [11] "National longitudinal surveys," <http://www.bls.gov/nls/>.
- [12] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS 2012)*, 2012, pp. 214–226.
- [13] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: <http://archive.ics.uci.edu/ml>
- [14] L. Shapley, "A value for n -person games," in *Contributions to the Theory of Games*, vol. 2, ser. Annals of Mathematics Studies, no. 28. Princeton University Press, 1953, pp. 307–317.
- [15] M. Maschler, E. Solan, and S. Zamir, *Game Theory*. Cambridge University Press, 2013.
- [16] L. S. Shapley and M. Shubik, "A method for evaluating the distribution of power in a committee system," *The American Political Science Review*, vol. 48, no. 3, pp. 787–792, 1954.
- [17] J. Banzhaf, "Weighted voting doesn't work: a mathematical analysis," *Rutgers Law Review*, vol. 19, pp. 317–343, 1965.
- [18] J. Deegan and E. Packel, "A new index of power for simple n -person games," *International Journal of Game Theory*, vol. 7, pp. 113–123, 1978.
- [19] H. Young, "Monotonic solutions of cooperative games," *International Journal of Game Theory*, vol. 14, no. 2, pp. 65–72, 1985.
- [20] S. Kullback and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.
- [21] G. Chalkiadakis, E. Elkind, and M. Wooldridge, *Computational Aspects of Cooperative Game Theory*. Morgan and Claypool, 2011.
- [22] Y. Bachrach, E. Markakis, E. Resnick, A. Procaccia, J. Rosenschein, and A. Saberi, "Approximating power indices: theoretical and empirical analysis," *Autonomous Agents and Multi-Agent Systems*, vol. 20, no. 2, pp. 105–122, 2010.
- [23] S. Maleki, L. Tran-Thanh, G. Hines, T. Rahwan, and A. Rogers, "Bounding the estimation error of sampling-based shapley value approximation with/without stratifying," *CoRR*, vol. abs/1306.4265, 2013.
- [24] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, March 1963. [Online]. Available: <http://www.jstor.org/stable/2282952>
- [25] N. Li, W. H. Qardaji, and D. Su, "Provably private data anonymization: Or, k -anonymity meets differential privacy," *CoRR*, vol. abs/1101.2604, 2011. [Online]. Available: <http://arxiv.org/abs/1101.2604>
- [26] Z. Jelveh and M. Luca, "Towards diagnosing accuracy loss in discrimination-aware classification: An application to predictive policing," *Fairness, Accountability and Transparency in Machine Learning*, vol. 26, no. 1, pp. 137–141, 2014.
- [27] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.
- [28] S. Barocas and H. Nissenbaum, "Big data's end run around procedural privacy protections," *Communications of the ACM*, vol. 57, no. 11, pp. 31–33, Oct. 2014.
- [29] T. Calders and S. Verwer, "Three naive bayes approaches for discrimination-free classification," *Data Mining and Knowledge Discovery*, vol. 21, no. 2, pp. 277–292, 2010. [Online]. Available: <http://dx.doi.org/10.1007/s10618-010-0190-x>
- [30] A. Datta, M. Tschantz, and A. Datta, "Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination," in *Proceedings on Privacy Enhancing Technologies (PoPETs 2015)*, 2015, pp. 92–112.
- [31] T. Kamishima, S. Akaho, and J. Sakuma, "Fairness-aware learning through regularization approach," in *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW 2011)*, 2011, pp. 643–650.
- [32] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *Proceedings of the 30th International Conference on Machine Learning (ICML 2013)*, 2013, pp. 325–333.
- [33] G. W. University, "Standardized test scores will be optional for gw applicants," 2015. [Online]. Available: <https://gwtoday.gwu.edu/standardized-test-scores-will-be-optional-gw-applicants>
- [34] The National Center for Fair and Open Testing, "850+ colleges and universities that do not use sat/act scores to admit substantial numbers of students into bachelor degree programs," 2015. [Online]. Available: <http://www.fairtest.org/university/optional>

- [35] D. Janzing, D. Balduzzi, M. Grosse-Wentrup, and B. Schölkopf, "Quantifying causal influences," *Ann. Statist.*, vol. 41, no. 5, pp. 2324–2358, 10 2013.
- [36] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003. [Online]. Available: <http://dl.acm.org/citation.cfm?id=944919.944968>
- [37] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001. [Online]. Available: <http://dx.doi.org/10.1023/A:1010933404324>
- [38] J. Pearl, *Causality: Models, Reasoning and Inference*, 2nd ed. New York, NY, USA: Cambridge University Press, 2009.
- [39] J. Tian and J. Pearl, "Probabilities of causation: Bounds and identification," *Annals of Mathematics and Artificial Intelligence*, vol. 28, no. 1-4, pp. 287–313, 2000.
- [40] J. Halpern and J. Pearl, "Causes and explanations: A structural-model approach. part i: Causes," *The British journal for the philosophy of science*, vol. 56, no. 4, pp. 843–887, 2005.
- [41] H. Chockler and J. Halpern, "Responsibility and blame: A structural-model approach," *Journal of Artificial Intelligence Research*, vol. 22, pp. 93–115, 2004.
- [42] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948. [Online]. Available: <http://dx.doi.org/10.1002/j.1538-7305.1948.tb01338.x>
- [43] A. Rényi, "On measures of entropy and information," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. Berkeley, Calif.: University of California Press, 1961, pp. 547–561. [Online]. Available: <http://projecteuclid.org/euclid.bsmsp/1200512181>
- [44] G. Smith, "Quantifying information flow using min-entropy," in *Proceedings of the 8th International Conference on Quantitative Evaluation of Systems (QEST 2011)*, 2011, pp. 159–167.
- [45] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.
- [46] R. Tibshirani, "Regression shrinkage and selection via the lasso: a retrospective," *Journal of the Royal Statistical Society Series B*, vol. 73, no. 3, pp. 273–282, 2011. [Online]. Available: <http://EconPapers.repec.org/RePEc:bla:jorss:v:73:y:2011:i:3:p:273-282>
- [47] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan, "Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model," *Ann. Appl. Stat.*, vol. 9, no. 3, pp. 1350–1371, 09 2015. [Online]. Available: <http://dx.doi.org/10.1214/15-AOAS848>
- [48] B. Ustun, S. Trac, and C. Rudin, "Supersparse linear integer models for interpretable classification," *ArXiv e-prints*, 2013. [Online]. Available: <http://arxiv.org/pdf/1306.5860v1>
- [49] S. Rping, "Learning interpretable models." Ph.D. dissertation, Dortmund University of Technology, 2006, <http://d-nb.info/997491736>.
- [50] S. Guha, B. Cheng, and P. Francis, "Challenges in measuring online advertising systems," in *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, ser. IMC '10. New York, NY, USA: ACM, 2010, pp. 81–87.
- [51] P. Barford, I. Canadi, D. Krushevskaja, Q. Ma, and S. Muthukrishnan, "Adscape: Harvesting and analyzing online display ads," in *Proceedings of the 23rd International Conference on World Wide Web*, ser. WWW '14. New York, NY, USA: ACM, 2014, pp. 597–608.
- [52] M. Lécuyer, G. Ducoffe, F. Lan, A. Papancea, T. Petsios, R. Spahn, A. Chaintreau, and R. Geambasu, "Xray: Enhancing the web's transparency with differential correlation," in *Proceedings of the 23rd USENIX Conference on Security Symposium*, ser. SEC'14. Berkeley, CA, USA: USENIX Association, 2014, pp. 49–64.
- [53] A. Datta, M. C. Tschantz, and A. Datta, "Automated experiments on ad privacy settings," *PoPETs*, vol. 2015, no. 1, pp. 92–112, 2015.
- [54] M. Lécuyer, R. Spahn, Y. Spiliopoulos, A. Chaintreau, R. Geambasu, and D. Hsu, "Sunlight: Fine-grained targeting detection at scale with statistical confidence," in *Proceedings of the 22Nd ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '15. New York, NY, USA: ACM, 2015, pp. 554–566.
- [55] A. Datta, A. Datta, A. Procaccia, and Y. Zick, "Influence in classification via cooperative game theory," in *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI 2015)*, 2015, pp. 511–517.
- [56] R. Lindelauf, H. Hamers, and B. Husslage, "Cooperative game theoretic centrality analysis of terrorist networks: The cases of jemaah islamiyah and al qaeda," *European Journal of Operational Research*, vol. 229, no. 1, pp. 230–238, 2013.
- [57] T. Michalak, T. Rahwan, P. Szczepanski, O. Skibski, R. Narayanan, M. Wooldridge, and N. Jennings, "Computational analysis of connectivity games with applications to the investigation of terrorist networks," in *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI 2013)*, 2013, pp. 293–301.
- [58] M. del Pozo, C. Manuel, E. González-Arangüena, and G. Owen, "Centrality in directed social networks. a game theoretic approach," *Social Networks*, vol. 33, no. 3, pp. 191–200, 2011.
- [59] T. Michalak, K. Aaditha, P. Szczepanski, B. Ravindran, and N. Jennings, "Efficient computation of the shapley value for game-theoretic network centrality," *Journal of Artificial Intelligence Research*, vol. 46, pp. 607–650, 2013.
- [60] P. Bork, L. Jensen, C. von Mering, A. Ramani, I. Lee, and E. Marcott, "Protein interaction networks from yeast to human," *Current Opinions in Structural Biology*, vol. 14, no. 3, pp. 292–299, 2004.
- [61] A. Keinan, B. Sandbank, C. Hilgetag, I. Meilijson, and E. Ruppin, "Fair attribution of functional contribution in artificial and biological networks," *Neural Computation*, vol. 16, no. 9, pp. 1887–1915, September 2004.
- [62] M. Malawski, "Equal treatment, symmetry and banzhaf value axiomatizations," *International Journal of Game Theory*, vol. 31, no. 1, pp. 47–67, 2002.

APPENDIX A ALTERNATIVE GAME-THEORETIC INFLUENCE MEASURES

In what follows, we describe two alternatives to the Shapley value used in this work. The Shapley value makes intuitive sense in our setting, as we argue in Section III-B. However, other measures may be appropriate for certain input data generation processes. In what follows we revisit the Banzhaf index, briefly discussed in Section III-A, and introduce the readers to the *Deegan-Packel index*, a game-theoretic influence measure with deep connections to a formal theory of responsibility and blame [41].

A. The Banzhaf Index

Recall that the Banzhaf index, denoted $\beta_i(N, v)$ is defined as follows:

$$\beta_i(N, v) = \frac{1}{2^{n-1}} \sum_{S \subseteq N \setminus \{i\}} m_i(S).$$

The Banzhaf index can be thought of as follows: each $j \in N \setminus \{i\}$ will join a work effort with probability $\frac{1}{2}$ (or, equivalently, each $S \subseteq N \setminus \{i\}$ has an equal chance of forming); if i joins as well, then its expected marginal contribution to the set formed is exactly the Banzhaf index. Note the marked difference between the probabilistic models: under the Shapley value, we sample *permutations* uniformly at random, whereas under the regime of the Banzhaf index, we sample sets uniformly at random. The different sampling protocols reflect different normative assumptions. For one, the Banzhaf index is not guaranteed to be efficient; that is, $\sum_{i \in N} \beta_i(N, v)$ is not necessarily equal to $v(N)$, whereas it is always the case that $\sum_{i=1}^n \varphi_i(N, v) = v(N)$. Moreover, the Banzhaf index is more biased towards measuring the marginal contribution of i to sets of size $\frac{n}{2} \pm O(\sqrt{n})$; this is because the expected size of a randomly selected set follows a binomial distribution $B(n, \frac{1}{2})$. On the other hand, the Shapley value is equally likely to measure the marginal contribution of i to sets of any size $k \in \{0, \dots, n\}$, as i is equally likely to be in any one position in a randomly selected permutation σ (and, in particular, the set of i 's predecessors in σ is equally likely to have any size $k \in \{0, \dots, n-1\}$).

Going back to the QII setting, the difference in sampling procedure is not merely an interesting anecdote: it is a significant modeling choice. Intuitively, the Banzhaf index is more appropriate if we assume that large sets of features would have a significant influence on outcomes, whereas the Shapley value is more appropriate if we assume that even small sets of features might cause significant effects on the outcome. Indeed, as we mention in Section VIII, aggregating the marginal influence of i over sets is a significant modeling choice; while using the measures proposed here is perfectly reasonable in many settings, other aggregation methods may be applicable in others.

Unlike the Shapley value, the Banzhaf index is not guaranteed to be efficient (although it does satisfy the symmetry and dummy properties). Indeed, [62] shows that replacing the efficiency axiom with an alternative axiom, uniquely

characterizes the Banzhaf index; the axiom, called *2-efficiency*, prescribes the behavior of an influence measure when two players merge. First, let us define a *merged game*; given a game $\langle N, v \rangle$, and two players $i, j \in N$, we write $T = \{i, j\}$. We define the game \bar{v} on $N \setminus T \cup \{\bar{T}\}$ as follows: for every set $S \subseteq N \setminus \{i, j\}$, $\bar{v}(S) = v(S)$, and $\bar{v}(S \cup \{\bar{T}\}) = v(S \cup \{i, j\})$, note that the added player \bar{T} represents the two players i and j who are now acting as one. The 2-Efficiency axiom states that influence should be invariant under merges.

Definition 14 (2-Efficiency (2-EFF)). Given two players $i, j \in N$, let \bar{v} be the game resulting from the merge of i and j into a single player \bar{T} ; an influence measure ϕ satisfies 2-Efficiency if $\phi_i(N, v) + \phi_j(N, v) = \phi_{\bar{T}}(N \setminus \{i, j\} \cup \{\bar{T}\}, \bar{v})$.

Theorem 15 ([62]). *The Banzhaf index is the only function to satisfy (Sym), (D), (Mono) and (2-EFF).*

In our context, 2-Efficiency can be interpreted as follows: suppose that we artificially treat two features i and j as one, keeping all other parameters fixed; in this setting, 2-efficiency means that the influence of merged features equals the influence they had as separate entities.

B. The Deegan-Packel Index

Finally, we discuss the *Deegan-Packel index* [18]. While the Shapley value and Banzhaf index are well-defined for any coalitional game, the Deegan-Packel index is only defined for *simple games*. A cooperative game is said to be simple if $v(S) \in \{0, 1\}$ for all $S \subseteq N$. In our setting, an influence measure would correspond to a simple game if it is binary (e.g. it measures some threshold behavior, or corresponds to a binary classifier). The binary requirement is rather strong; however, we wish to draw the reader's attention to the Deegan-Packel index, as it has an interesting connection to *causal responsibility* [41], a variant of the classic Pearl-Halpern causality model [40], which aims to measure the degree to which a single variable causes an outcome.

Given a simple game $v : 2^N \rightarrow \{0, 1\}$, let $\mathcal{M}(v)$ be the set of *minimal winning coalitions*; that is, for every $S \in \mathcal{M}(v)$, $v(S) = 1$, and $v(T) = 0$ for every strict subset of S . The Deegan-Packel index assigns a value of

$$\delta_i(N, v) = \frac{1}{|\mathcal{M}(v)|} \sum_{S \in \mathcal{M}(v) : i \in S} \frac{1}{|S|}.$$

The intuition behind the Deegan-Packel index is as follows: players will not form coalitions any larger than what they absolutely have to in order to win, so it does not make sense to measure their effect on non-minimal winning coalitions. Furthermore, when a minimal winning coalition is formed, the benefits from its formation are divided equally among its members; in particular, small coalitions confer a greater benefit for those forming them than large ones. The Deegan-Packel index measures the expected payment one receives, assuming that every minimal winning coalition is equally likely to form. Interestingly, the Deegan-Packel index corresponds nicely to the notion of responsibility and blame described in [41].

Suppose that we have a set of variables X_1, \dots, X_n set to x_1, \dots, x_n , and some binary effect $f(x_1, \dots, x_n)$ (written as $f(\mathbf{x})$) occurs (say, $f(\mathbf{x}) = 1$). To establish a causal relation between the setting of X_i to x_i and $f(\mathbf{x}) = 1$, [40] require that there is some set $S \subseteq N \setminus \{i\}$ and some values $(y_j)_{j \in S \cup \{i\}}$ such that $f(\mathbf{x}_{-S \cup \{i\}}, (y_j)_{j \in S \cup \{i\}}) = 0$, but $f(\mathbf{x}_{-S}, (y_j)_{j \in S}) = 1$. In words, an intervention on the values of both S and i may cause a change in the value of f , but performing the same intervention just on the variables in S would not cause such a change. This definition is at the heart of the marginal contribution approach to interventions that we describe in Section III-A. [41] define the responsibility of i for an outcome as $\frac{1}{k+1}$, where k is the size of the smallest set S for which the causality definition holds with respect to i . The Deegan-Packel index can thus be thought of as measuring a similar notion: instead of taking the overall minimal number of changes necessary in order to make i a direct, counterfactual cause, we observe all minimal sets that do so. Taking the average responsibility of i (referred to as *blame* in [41]) according to this variant, we obtain the Deegan-Packel index.

Example 16. Let us examine the following setup, based on Example 3.3 in [41]. There are $n = 2k + 1$ voters (n is an odd number) who must choose between two candidates, Mr. B and Mr. G ([41] describe the setting with $n = 11$). All voters elected Mr. B , resulting in an n -0 win. It is natural to ask: how responsible was voter i for the victory of Mr. B ? According to [41], the degree of responsibility of each voter is $\frac{1}{k+1}$. It will require that i and k additional voters change their vote in order for the outcome to change. Modeling this setup as a cooperative game is quite natural: the voters are the players $N = \{1, \dots, n\}$; for every subset $S \subseteq N$ we have

$$v(S) = \begin{cases} 1 & \text{if } |S| \geq k + 1 \\ 0 & \text{otherwise.} \end{cases}$$

That is, $v(S) = 1$ if and only if the set S can change the outcome of the election. The minimal winning coalitions here are the subsets of N of size $k + 1$, thus the Deegan-Packel index of player i is

$$\begin{aligned} \delta_i(N, v) &= \frac{1}{|\mathcal{M}(v)|} \sum_{S \in \mathcal{M}(v): i \in S} \frac{1}{|S|} \\ &= \frac{1}{\binom{n}{k+1}} \binom{n}{k} \frac{1}{k+1} = \frac{1}{n-k} = \frac{1}{k+1} \end{aligned}$$

We note that if one assumes that all voters are equally likely to prefer Mr. B over Mr. G , then the blame of voter i would be computed in the exact manner as the Deegan-Packel index.

A Unified Approach to Interpreting Model Predictions

Scott M. Lundberg

Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98105
slund1@cs.washington.edu

Su-In Lee

Paul G. Allen School of Computer Science
Department of Genome Sciences
University of Washington
Seattle, WA 98105
suinlee@cs.washington.edu

Abstract

Understanding why a model makes a certain prediction can be as crucial as the prediction’s accuracy in many applications. However, the highest accuracy for large modern datasets is often achieved by complex models that even experts struggle to interpret, such as ensemble or deep learning models, creating a tension between *accuracy* and *interpretability*. In response, various methods have recently been proposed to help users interpret the predictions of complex models, but it is often unclear how these methods are related and when one method is preferable over another. To address this problem, we present a unified framework for interpreting predictions, SHAP (SHapley Additive exPlanations). SHAP assigns each feature an importance value for a particular prediction. Its novel components include: (1) the identification of a new class of additive feature importance measures, and (2) theoretical results showing there is a unique solution in this class with a set of desirable properties. The new class unifies six existing methods, notable because several recent methods in the class lack the proposed desirable properties. Based on insights from this unification, we present new methods that show improved computational performance and/or better consistency with human intuition than previous approaches.

1 Introduction

The ability to correctly interpret a prediction model’s output is extremely important. It engenders appropriate user trust, provides insight into how a model may be improved, and supports understanding of the process being modeled. In some applications, simple models (e.g., linear models) are often preferred for their ease of interpretation, even if they may be less accurate than complex ones. However, the growing availability of big data has increased the benefits of using complex models, so bringing to the forefront the trade-off between accuracy and interpretability of a model’s output. A wide variety of different methods have been recently proposed to address this issue [5, 8, 9, 3, 4, 1]. But an understanding of how these methods relate and when one method is preferable to another is still lacking.

Here, we present a novel unified approach to interpreting model predictions.¹ Our approach leads to three potentially surprising results that bring clarity to the growing space of methods:

1. We introduce the perspective of viewing *any* explanation of a model’s prediction as a model itself, which we term the *explanation model*. This lets us define the class of *additive feature attribution methods* (Section 2), which unifies six current methods.

¹<https://github.com/slundberg/shap>

2. We then show that game theory results guaranteeing a unique solution apply to the *entire class* of additive feature attribution methods (Section 3) and propose *SHAP values* as a unified measure of feature importance that various methods approximate (Section 4).
3. We propose new SHAP value estimation methods and demonstrate that they are better aligned with human intuition as measured by user studies and more effectively discriminate among model output classes than several existing methods (Section 5).

2 Additive Feature Attribution Methods

The best explanation of a simple model is the model itself; it perfectly represents itself and is easy to understand. For complex models, such as ensemble methods or deep networks, we cannot use the original model as its own best explanation because it is not easy to understand. Instead, we must use a simpler *explanation model*, which we define as any interpretable approximation of the original model. We show below that six current explanation methods from the literature all use the same explanation model. This previously unappreciated unity has interesting implications, which we describe in later sections.

Let f be the original prediction model to be explained and g the explanation model. Here, we focus on *local methods* designed to explain a prediction $f(x)$ based on a single input x , as proposed in LIME [5]. Explanation models often use *simplified inputs* x' that map to the original inputs through a mapping function $x = h_x(x')$. Local methods try to ensure $g(z') \approx f(h_x(z'))$ whenever $z' \approx x'$. (Note that $h_x(x') = x$ even though x' may contain less information than x because h_x is specific to the current input x .)

Definition 1 Additive feature attribution methods have an explanation model that is a linear function of binary variables:

$$g(z') = \phi_0 + \sum_{i=1}^M \phi_i z'_i, \quad (1)$$

where $z' \in \{0, 1\}^M$, M is the number of simplified input features, and $\phi_i \in \mathbb{R}$.

Methods with explanation models matching Definition 1 attribute an effect ϕ_i to each feature, and summing the effects of all feature attributions approximates the output $f(x)$ of the original model. Many current methods match Definition 1, several of which are discussed below.

2.1 LIME

The *LIME* method interprets individual model predictions based on locally approximating the model around a given prediction [5]. The local linear explanation model that LIME uses adheres to Equation 1 exactly and is thus an additive feature attribution method. LIME refers to simplified inputs x' as “interpretable inputs,” and the mapping $x = h_x(x')$ converts a binary vector of interpretable inputs into the original input space. Different types of h_x mappings are used for different input spaces. For bag of words text features, h_x converts a vector of 1’s or 0’s (present or not) into the original word count if the simplified input is one, or zero if the simplified input is zero. For images, h_x treats the image as a set of super pixels; it then maps 1 to leaving the super pixel as its original value and 0 to replacing the super pixel with an average of neighboring pixels (this is meant to represent being missing).

To find ϕ , LIME minimizes the following objective function:

$$\xi = \arg \min_{g \in \mathcal{G}} L(f, g, \pi_{x'}) + \Omega(g). \quad (2)$$

Faithfulness of the explanation model $g(z')$ to the original model $f(h_x(z'))$ is enforced through the loss L over a set of samples in the simplified input space weighted by the local kernel $\pi_{x'}$. Ω penalizes the complexity of g . Since in LIME g follows Equation 1 and L is a squared loss, Equation 2 can be solved using penalized linear regression.

2.2 DeepLIFT

DeepLIFT was recently proposed as a recursive prediction explanation method for deep learning [8, 7]. It attributes to each input x_i a value $C_{\Delta x_i \Delta o}$ that represents the effect of that input being set to a reference value as opposed to its original value. This means that for DeepLIFT, the mapping $x = h_x(x')$ converts binary values into the original inputs, where 1 indicates that an input takes its original value, and 0 indicates that it takes the reference value. The reference value, though chosen by the user, represents a typical uninformative background value for the feature.

DeepLIFT uses a "summation-to-delta" property that states:

$$\sum_{i=1}^n C_{\Delta x_i \Delta o} = \Delta o, \quad (3)$$

where $o = f(x)$ is the model output, $\Delta o = f(x) - f(r)$, $\Delta x_i = x_i - r_i$, and r is the reference input. If we let $\phi_i = C_{\Delta x_i \Delta o}$ and $\phi_0 = f(r)$, then DeepLIFT's explanation model matches Equation 1 and is thus another additive feature attribution method.

2.3 Layer-Wise Relevance Propagation

The *layer-wise relevance propagation* method interprets the predictions of deep networks [1]. As noted by Shrikumar et al., this method is equivalent to DeepLIFT with the reference activations of all neurons fixed to zero. Thus, $x = h_x(x')$ converts binary values into the original input space, where 1 means that an input takes its original value, and 0 means an input takes the 0 value. Layer-wise relevance propagation's explanation model, like DeepLIFT's, matches Equation 1.

2.4 Classic Shapley Value Estimation

Three previous methods use classic equations from cooperative game theory to compute explanations of model predictions: Shapley regression values [4], Shapley sampling values [9], and Quantitative Input Influence [3].

Shapley regression values are feature importances for linear models in the presence of multicollinearity. This method requires retraining the model on all feature subsets $S \subseteq F$, where F is the set of all features. It assigns an importance value to each feature that represents the effect on the model prediction of including that feature. To compute this effect, a model $f_{S \cup \{i\}}$ is trained with that feature present, and another model f_S is trained with the feature withheld. Then, predictions from the two models are compared on the current input $f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)$, where x_S represents the values of the input features in the set S . Since the effect of withholding a feature depends on other features in the model, the preceding differences are computed for all possible subsets $S \subseteq F \setminus \{i\}$. The Shapley values are then computed and used as feature attributions. They are a weighted average of all possible differences:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)]. \quad (4)$$

For Shapley regression values, h_x maps 1 or 0 to the original input space, where 1 indicates the input is included in the model, and 0 indicates exclusion from the model. If we let $\phi_0 = f_\emptyset(\emptyset)$, then the Shapley regression values match Equation 1 and are hence an additive feature attribution method.

Shapley sampling values are meant to explain any model by: (1) applying sampling approximations to Equation 4, and (2) approximating the effect of removing a variable from the model by integrating over samples from the training dataset. This eliminates the need to retrain the model and allows fewer than $2^{|F|}$ differences to be computed. Since the explanation model form of Shapley sampling values is the same as that for Shapley regression values, it is also an additive feature attribution method.

Quantitative input influence is a broader framework that addresses more than feature attributions. However, as part of its method it independently proposes a sampling approximation to Shapley values that is nearly identical to Shapley sampling values. It is thus another additive feature attribution method.

3 Simple Properties Uniquely Determine Additive Feature Attributions

A surprising attribute of the class of additive feature attribution methods is the presence of a single unique solution in this class with three desirable properties (described below). While these properties are familiar to the classical Shapley value estimation methods, they were previously unknown for other additive feature attribution methods.

The first desirable property is *local accuracy*. When approximating the original model f for a specific input x , local accuracy requires the explanation model to at least match the output of f for the simplified input x' (which corresponds to the original input x).

Property 1 (Local accuracy)

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^M \phi_i x'_i \quad (5)$$

The explanation model $g(x')$ matches the original model $f(x)$ when $x = h_x(x')$, where $\phi_0 = f(h_x(\mathbf{0}))$ represents the model output with all simplified inputs toggled off (i.e. missing).

The second property is *missingness*. If the simplified inputs represent feature presence, then missingness requires features missing in the original input to have no impact. All of the methods described in Section 2 obey the missingness property.

Property 2 (Missingness)

$$x'_i = 0 \implies \phi_i = 0 \quad (6)$$

Missingness constrains features where $x'_i = 0$ to have no attributed impact.

The third property is *consistency*. Consistency states that if a model changes so that some simplified input's contribution increases or stays the same regardless of the other inputs, that input's attribution should not decrease.

Property 3 (Consistency) *Let $f_x(z') = f(h_x(z'))$ and $z' \setminus i$ denote setting $z'_i = 0$. For any two models f and f' , if*

$$f'_x(z') - f'_x(z' \setminus i) \geq f_x(z') - f_x(z' \setminus i) \quad (7)$$

for all inputs $z' \in \{0, 1\}^M$, then $\phi_i(f', x) \geq \phi_i(f, x)$.

Theorem 1 *Only one possible explanation model g follows Definition 1 and satisfies Properties 1, 2, and 3:*

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} [f_x(z') - f_x(z' \setminus i)] \quad (8)$$

where $|z'|$ is the number of non-zero entries in z' , and $z' \subseteq x'$ represents all z' vectors where the non-zero entries are a subset of the non-zero entries in x' .

Theorem 1 follows from combined cooperative game theory results, where the values ϕ_i are known as Shapley values [6]. Young (1985) demonstrated that Shapley values are the only set of values that satisfy three axioms similar to Property 1, Property 3, and a final property that we show to be redundant in this setting (see Supplementary Material). Property 2 is required to adapt the Shapley proofs to the class of additive feature attribution methods.

Under Properties 1-3, for a given simplified input mapping h_x , Theorem 1 shows that there is only one possible additive feature attribution method. This result implies that methods not based on Shapley values violate local accuracy and/or consistency (methods in Section 2 already respect missingness). The following section proposes a unified approach that improves previous methods, preventing them from unintentionally violating Properties 1 and 3.

4 SHAP (SHapley Additive exPlanation) Values

We propose SHAP values as a unified measure of feature importance. These are the Shapley values of a conditional expectation function of the original model; thus, they are the solution to Equation

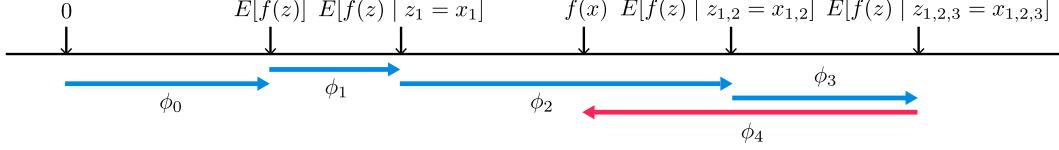


Figure 1: SHAP (SHapley Additive exPlanation) values attribute to each feature the change in the expected model prediction when conditioning on that feature. They explain how to get from the base value $E[f(z)]$ that would be predicted if we did not know any features to the current output $f(x)$. This diagram shows a single ordering. When the model is non-linear or the input features are not independent, however, the order in which features are added to the expectation matters, and the SHAP values arise from averaging the ϕ_i values across all possible orderings.

8, where $f_x(z') = f(h_x(z')) = E[f(z) | z_S]$, and S is the set of non-zero indexes in z' (Figure 1). Based on Sections 2 and 3, SHAP values provide the unique additive feature importance measure that adheres to Properties 1-3 and uses conditional expectations to define simplified inputs. Implicit in this definition of SHAP values is a simplified input mapping, $h_x(z') = z_S$, where z_S has missing values for features not in the set S . Since most models cannot handle arbitrary patterns of missing input values, we approximate $f(z_S)$ with $E[f(z) | z_S]$. This definition of SHAP values is designed to closely align with the Shapley regression, Shapley sampling, and quantitative input influence feature attributions, while also allowing for connections with LIME, DeepLIFT, and layer-wise relevance propagation.

The exact computation of SHAP values is challenging. However, by combining insights from current additive feature attribution methods, we can approximate them. We describe two model-agnostic approximation methods, one that is already known (Shapley sampling values) and another that is novel (Kernel SHAP). We also describe four model-type-specific approximation methods, two of which are novel (Max SHAP, Deep SHAP). When using these methods, feature independence and model linearity are two optional assumptions simplifying the computation of the expected values (note that \bar{S} is the set of features not in S):

$$f(h_x(z')) = E[f(z) | z_S] \quad \text{SHAP explanation model simplified input mapping} \quad (9)$$

$$= E_{z_S | z_S}[f(z)] \quad \text{expectation over } z_{\bar{S}} | z_S \quad (10)$$

$$\approx E_{z_{\bar{S}}}[f(z)] \quad \text{assume feature independence (as in [9, 5, 7, 3])} \quad (11)$$

$$\approx f([z_S, E[z_{\bar{S}}]]). \quad \text{assume model linearity} \quad (12)$$

4.1 Model-Agnostic Approximations

If we assume feature independence when approximating conditional expectations (Equation 11), as in [9, 5, 7, 3], then SHAP values can be estimated directly using the Shapley sampling values method [9] or equivalently the Quantitative Input Influence method [3]. These methods use a sampling approximation of a permutation version of the classic Shapley value equations (Equation 8). Separate sampling estimates are performed for each feature attribution. While reasonable to compute for a small number of inputs, the Kernel SHAP method described next requires fewer evaluations of the original model to obtain similar approximation accuracy (Section 5).

Kernel SHAP (Linear LIME + Shapley values)

Linear LIME uses a linear explanation model to locally approximate f , where local is measured in the simplified binary input space. At first glance, the regression formulation of LIME in Equation 2 seems very different from the classical Shapley value formulation of Equation 8. However, since linear LIME is an additive feature attribution method, we know the Shapley values are the only possible solution to Equation 2 that satisfies Properties 1-3 – local accuracy, missingness and consistency. A natural question to pose is whether the solution to Equation 2 recovers these values. The answer depends on the choice of loss function L , weighting kernel $\pi_{x'}$ and regularization term Ω . The LIME choices for these parameters are made heuristically; using these choices, Equation 2 does not recover the Shapley values. One consequence is that local accuracy and/or consistency are violated, which in turn leads to unintuitive behavior in certain circumstances (see Section 5).

Below we show how to avoid heuristically choosing the parameters in Equation 2 and how to find the loss function L , weighting kernel $\pi_{x'}$, and regularization term Ω that recover the Shapley values.

Theorem 2 (Shapley kernel) *Under Definition 1, the specific forms of $\pi_{x'}$, L , and Ω that make solutions of Equation 2 consistent with Properties 1 through 3 are:*

$$\begin{aligned}\Omega(g) &= 0, \\ \pi_{x'}(z') &= \frac{(M-1)}{(M \text{ choose } |z'|)|z'|(M-|z'|)}, \\ L(f, g, \pi_{x'}) &= \sum_{z' \in Z} [f(h_x(z')) - g(z')]^2 \pi_{x'}(z'),\end{aligned}$$

where $|z'|$ is the number of non-zero elements in z' .

The proof of Theorem 2 is shown in the Supplementary Material.

It is important to note that $\pi_{x'}(z') = \infty$ when $|z'| \in \{0, M\}$, which enforces $\phi_0 = f_x(\emptyset)$ and $f(x) = \sum_{i=0}^M \phi_i$. In practice, these infinite weights can be avoided during optimization by analytically eliminating two variables using these constraints.

Since $g(z')$ in Theorem 2 is assumed to follow a linear form, and L is a squared loss, Equation 2 can still be solved using linear regression. As a consequence, the Shapley values from game theory can be computed using weighted linear regression.² Since LIME uses a simplified input mapping that is equivalent to the approximation of the SHAP mapping given in Equation 12, this enables regression-based, model-agnostic estimation of SHAP values. Jointly estimating all SHAP values using regression provides better sample efficiency than the direct use of classical Shapley equations (see Section 5).

The intuitive connection between linear regression and Shapley values is that Equation 8 is a difference of means. Since the mean is also the best least squares point estimate for a set of data points, it is natural to search for a weighting kernel that causes linear least squares regression to recapitulate the Shapley values. This leads to a kernel that distinctly differs from previous heuristically chosen kernels (Figure 2A).

4.2 Model-Specific Approximations

While Kernel SHAP improves the sample efficiency of model-agnostic estimations of SHAP values, by restricting our attention to specific model types, we can develop faster model-specific approximation methods.

Linear SHAP

For linear models, if we assume input feature independence (Equation 11), SHAP values can be approximated directly from the model's weight coefficients.

Corollary 1 (Linear SHAP) *Given a linear model $f(x) = \sum_{j=1}^M w_j x_j + b$: $\phi_0(f, x) = b$ and*

$$\phi_i(f, x) = w_j(x_j - E[x_j])$$

This follows from Theorem 2 and Equation 11, and it has been previously noted by Štrumbelj and Kononenko [9].

Low-Order SHAP

Since linear regression using Theorem 2 has complexity $O(2^M + M^3)$, it is efficient for small values of M if we choose an approximation of the conditional expectations (Equation 11 or 12).

²During the preparation of this manuscript we discovered this parallels an equivalent constrained quadratic minimization formulation of Shapley values proposed in econometrics [2].

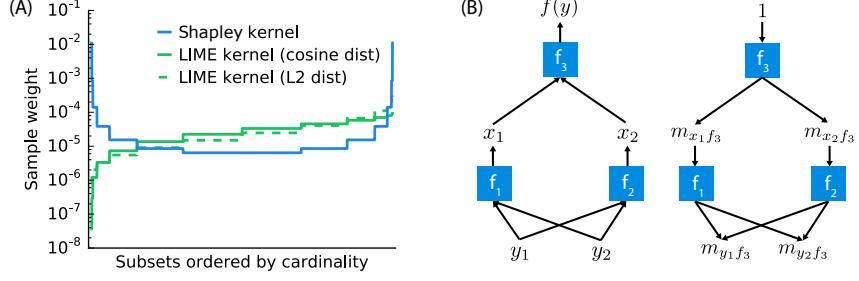


Figure 2: (A) The Shapley kernel weighting is symmetric when all possible z' vectors are ordered by cardinality there are 2^{15} vectors in this example. This is distinctly different from previous heuristically chosen kernels. (B) Compositional models such as deep neural networks are comprised of many simple components. Given analytic solutions for the Shapley values of the components, fast approximations for the full model can be made using DeepLIFT’s style of back-propagation.

Max SHAP

Using a permutation formulation of Shapley values, we can calculate the probability that each input will increase the maximum value over every other input. Doing this on a sorted order of input values lets us compute the Shapley values of a max function with M inputs in $O(M^2)$ time instead of $O(M2^M)$. See Supplementary Material for the full algorithm.

Deep SHAP (DeepLIFT + Shapley values)

While Kernel SHAP can be used on any model, including deep models, it is natural to ask whether there is a way to leverage extra knowledge about the compositional nature of deep networks to improve computational performance. We find an answer to this question through a previously unappreciated connection between Shapley values and DeepLIFT [8]. If we interpret the reference value in Equation 3 as representing $E[x]$ in Equation 12, then DeepLIFT approximates SHAP values assuming that the input features are independent of one another and the deep model is linear. DeepLIFT uses a linear composition rule, which is equivalent to linearizing the non-linear components of a neural network. Its back-propagation rules defining how each component is linearized are intuitive but were heuristically chosen. Since DeepLIFT is an additive feature attribution method that satisfies local accuracy and missingness, we know that Shapley values represent the only attribution values that satisfy consistency. This motivates our adapting DeepLIFT to become a compositional approximation of SHAP values, leading to Deep SHAP.

Deep SHAP combines SHAP values computed for smaller components of the network into SHAP values for the whole network. It does so by recursively passing DeepLIFT’s multipliers, now defined in terms of SHAP values, backwards through the network as in Figure 2B:

$$m_{x_j f_3} = \frac{\phi_i(f_3, x)}{x_j - E[x_j]} \quad (13)$$

$$\forall_{j \in \{1,2\}} \quad m_{y_i f_j} = \frac{\phi_i(f_j, y)}{y_i - E[y_i]} \quad (14)$$

$$m_{y_i f_3} = \sum_{j=1}^2 m_{y_i f_j} m_{x_j f_3} \quad \text{chain rule} \quad (15)$$

$$\phi_i(f_3, y) \approx m_{y_i f_3}(y_i - E[y_i]) \quad \text{linear approximation} \quad (16)$$

Since the SHAP values for the simple network components can be efficiently solved analytically if they are linear, max pooling, or an activation function with just one input, this composition rule enables a fast approximation of values for the whole model. Deep SHAP avoids the need to heuristically choose ways to linearize components. Instead, it derives an effective linearization from the SHAP values computed for each component. The *max* function offers one example where this leads to improved attributions (see Section 5).

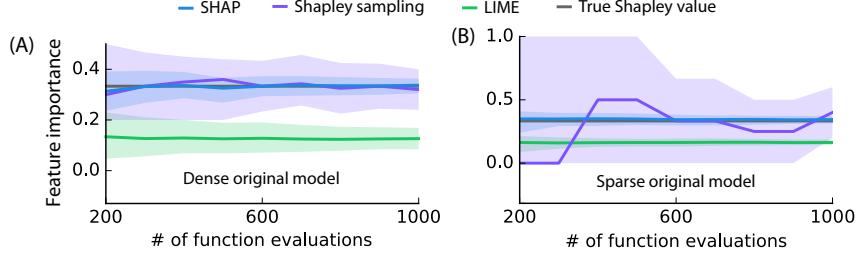


Figure 3: Comparison of three additive feature attribution methods: Kernel SHAP (using a debiased lasso), Shapley sampling values, and LIME (using the open source implementation). Feature importance estimates are shown for one feature in two models as the number of evaluations of the original model function increases. The 10th and 90th percentiles are shown for 200 replicate estimates at each sample size. (A) A decision tree model using all 10 input features is explained for a single input. (B) A decision tree using only 3 of 100 input features is explained for a single input.

5 Computational and User Study Experiments

We evaluated the benefits of SHAP values using the Kernel SHAP and Deep SHAP approximation methods. First, we compared the computational efficiency and accuracy of Kernel SHAP vs. LIME and Shapley sampling values. Second, we designed user studies to compare SHAP values with alternative feature importance allocations represented by DeepLIFT and LIME. As might be expected, SHAP values prove more consistent with human intuition than other methods that fail to meet Properties 1-3 (Section 2). Finally, we use MNIST digit image classification to compare SHAP with DeepLIFT and LIME.

5.1 Computational Efficiency

Theorem 2 connects Shapley values from game theory with weighted linear regression. Kernel SHAP uses this connection to compute feature importance. This leads to more accurate estimates with fewer evaluations of the original model than previous sampling-based estimates of Equation 8, particularly when regularization is added to the linear model (Figure 3). Comparing Shapley sampling, SHAP, and LIME on both dense and sparse decision tree models illustrates both the improved sample efficiency of Kernel SHAP and that values from LIME can differ significantly from SHAP values that satisfy local accuracy and consistency.

5.2 Consistency with Human Intuition

Theorem 1 provides a strong incentive for all additive feature attribution methods to use SHAP values. Both LIME and DeepLIFT, as originally demonstrated, compute different feature importance values. To validate the importance of Theorem 1, we compared explanations from LIME, DeepLIFT, and SHAP with user explanations of simple models (using Amazon Mechanical Turk). Our testing assumes that good model explanations should be consistent with explanations from humans who understand that model.

We compared LIME, DeepLIFT, and SHAP with human explanations for two settings. The first setting used a sickness score that was higher when only one of two symptoms was present (Figure 4A). The second used a max allocation problem to which DeepLIFT can be applied. Participants were told a short story about how three men made money based on the maximum score any of them achieved (Figure 4B). In both cases, participants were asked to assign credit for the output (the sickness score or money won) among the inputs (i.e., symptoms or players). We found a much stronger agreement between human explanations and SHAP than with other methods. SHAP’s improved performance for max functions addresses the open problem of max pooling functions in DeepLIFT [7].

5.3 Explaining Class Differences

As discussed in Section 4.2, DeepLIFT’s compositional approach suggests a compositional approximation of SHAP values (Deep SHAP). These insights, in turn, improve DeepLIFT, and a new version

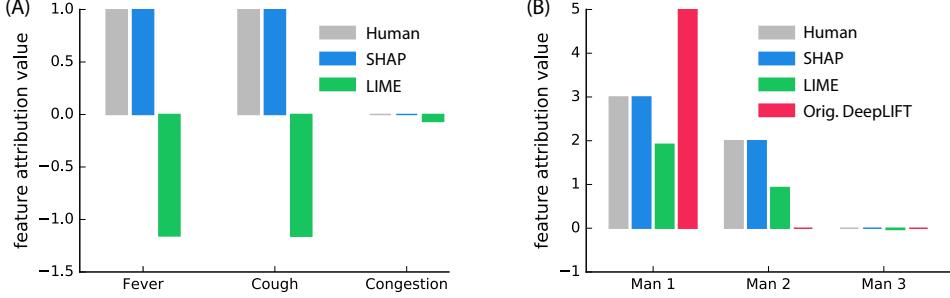


Figure 4: Human feature impact estimates are shown as the most common explanation given among 30 (A) and 52 (B) random individuals, respectively. (A) Feature attributions for a model output value (sickness score) of 2. The model output is 2 when fever and cough are both present, 5 when only one of fever or cough is present, and 0 otherwise. (B) Attributions of profit among three men, given according to the maximum number of questions any man got right. The first man got 5 questions right, the second 4 questions, and the third got none right, so the profit is \$5.

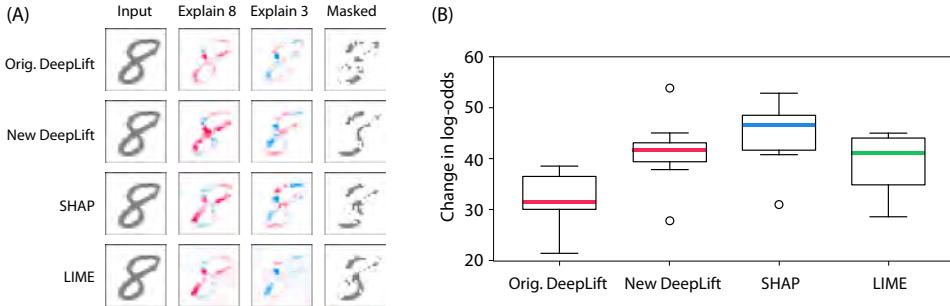


Figure 5: Explaining the output of a convolutional network trained on the MNIST digit dataset. Orig. DeepLIFT has no explicit Shapley approximations, while New DeepLIFT seeks to better approximate Shapley values. (A) Red areas increase the probability of that class, and blue areas decrease the probability. Masked removes pixels in order to go from 8 to 3. (B) The change in log odds when masking over 20 random images supports the use of better estimates of SHAP values.

includes updates to better match Shapley values [7]. Figure 5 extends DeepLIFT’s convolutional network example to highlight the increased performance of estimates that are closer to SHAP values. The pre-trained model and Figure 5 example are the same as those used in [7], with inputs normalized between 0 and 1. Two convolution layers and 2 dense layers are followed by a 10-way softmax output layer. Both DeepLIFT versions explain a normalized version of the linear layer, while SHAP (computed using Kernel SHAP) and LIME explain the model’s output. SHAP and LIME were both run with 50k samples (Supplementary Figure 1); to improve performance, LIME was modified to use single pixel segmentation over the digit pixels. To match [7], we masked 20% of the pixels chosen to switch the predicted class from 8 to 3 according to the feature attribution given by each method.

6 Conclusion

The growing tension between the accuracy and interpretability of model predictions has motivated the development of methods that help users interpret predictions. The SHAP framework identifies the class of additive feature importance methods (which includes six previous methods) and shows there is a unique solution in this class that adheres to desirable properties. 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.

We presented several different estimation methods for SHAP values, along with proofs and experiments showing that these values are desirable. Promising next steps involve developing faster model-type-specific estimation methods that make fewer assumptions, integrating work on estimating interaction effects from game theory, and defining new explanation model classes.

Acknowledgements

This work was supported by a National Science Foundation (NSF) DBI-135589, NSF CAREER DBI-155230, American Cancer Society 127332-RSG-15-097-01-TBG, National Institute of Health (NIH) AG049196, and NSF Graduate Research Fellowship. We would like to thank Marco Ribeiro, Erik Štrumbelj, Avanti Shrikumar, Yair Zick, the Lee Lab, and the NIPS reviewers for feedback that has significantly improved this work.

References

- [1] Sebastian Bach et al. “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation”. In: *PloS One* 10.7 (2015), e0130140.
- [2] A Charnes et al. “Extremal principle solutions of games in characteristic function form: core, Chebychev and Shapley value generalizations”. In: *Econometrics of Planning and Efficiency* 11 (1988), pp. 123–133.
- [3] Anupam Datta, Shayak Sen, and Yair Zick. “Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems”. In: *Security and Privacy (SP), 2016 IEEE Symposium on*. IEEE. 2016, pp. 598–617.
- [4] Stan Lipovetsky and Michael Conklin. “Analysis of regression in game theory approach”. In: *Applied Stochastic Models in Business and Industry* 17.4 (2001), pp. 319–330.
- [5] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “Why should i trust you?: Explaining the predictions of any classifier”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2016, pp. 1135–1144.
- [6] Lloyd S Shapley. “A value for n-person games”. In: *Contributions to the Theory of Games* 2.28 (1953), pp. 307–317.
- [7] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. “Learning Important Features Through Propagating Activation Differences”. In: *arXiv preprint arXiv:1704.02685* (2017).
- [8] Avanti Shrikumar et al. “Not Just a Black Box: Learning Important Features Through Propagating Activation Differences”. In: *arXiv preprint arXiv:1605.01713* (2016).
- [9] Erik Štrumbelj and Igor Kononenko. “Explaining prediction models and individual predictions with feature contributions”. In: *Knowledge and information systems* 41.3 (2014), pp. 647–665.
- [10] H Peyton Young. “Monotonic solutions of cooperative games”. In: *International Journal of Game Theory* 14.2 (1985), pp. 65–72.

Week 11: Discrimination in on-line ad delivery

DOI:10.1145/2447976.2447990

 Article development led by **queue.acm.org**

Google ads, black names and white names, racial discrimination, and click advertising.

BY LATANYA SWEENEY

Discrimination in Online Ad Delivery

DO ONLINE ADS suggestive of arrest records appear more often with searches of black-sounding names than white-sounding names? What is a black-sounding name or white-sounding name, anyway? How do you design technology to reason about societal consequences like structural racism? Let's take a scientific dive into online ad delivery to find answers.

"Have you ever been arrested?" Imagine this question appearing whenever someone enters your name in a search engine. Perhaps you are in competition for an award or a new job, or maybe you are in a position of trust, such as a professor or a volunteer. Perhaps you are dating or engaged in any one of hundreds of circumstances for which someone wants to learn more about you online. Appearing alongside your accomplishments is an advertisement implying you may have a criminal record, whether you actually have one or not. Worse, the ads may not appear for your competitors.

Employers frequently ask whether applicants have ever been arrested or charged with a crime, but if an employer disqualifies a job applicant based solely upon information indicating an arrest record, the company may face legal consequences. The U.S. Equal Employment Opportunity Commission (EEOC) is the federal agency charged with enforcing Title VII of the Civil Rights Act of 1964, a law that applies to most employers, prohibiting employment discrimination based on race, color, religion, sex, or national origin, and extended to those having criminal records.^{5,11} Title VII does not prohibit employers from obtaining criminal background information, but a blanket policy of excluding applicants based solely upon information indicating an arrest record can result in a charge of discrimination.

To make a determination, the EEOC uses an adverse impact test that measures whether certain practices, intentional or not, have a disproportionate effect on a group of people whose defining characteristics are covered by Title VII. To decide, you calculate the percentage of people affected in each group and then divide the smaller value by the larger to get the ratio and compare the result to 80. If the ratio is less than 80, then the EEOC considers the effect disproportionate and may hold the employer responsible for discrimination.⁶

What about online ads suggesting someone with your name has an arrest record? Title VII only applies if you have an arrest record and can prove the employer inappropriately used the ads.

Are the ads commercial free speech—a constitutional right to display the ad associated with your name? The First Amendment of the U.S. Constitution protects advertising, but the U.S. Supreme Court set out a test for assessing restrictions on commercial speech, which begins by determining whether the speech is misleading.³ Are online ads suggesting the existence of an arrest record misleading if no one by that name has an arrest record?



Figure 1. Ads from a Google search of three different names beginning with first name "Latanya."

Ads related to Latanya Farrell ⓘ
Latanya Farrell Arrested?
www.instantcheckmate.com/
 1) Enter Name and State. 2) Access Full Background Checks Instantly.

Latanya Farrell
www.publicrecords.com/
 Public Records Found For: Latanya Farrell. View Now.

(a)

LATANYA FARRELL
 40 Lexington Rd
 West Hartford, CT 06119
 DOB: Jun 10, 1972 (40 years old)

Criminal History
 This section contains possible citation, arrest, and criminal records for the subject of this report.
 While our database does contain hundreds of millions of arrest records, different counties have different rules regarding what information they will and will not release.
 We share with you as much information as we possibly can, but a clean slate here should not be interpreted as a guarantee that Latanya Farrell has never been arrested; it simply means that we were not able to locate any matching arrest records in the data that is available to us.

Possible Matching Arrest Records			
Name	County and State	Offenses	View Details
No matching arrest records were found.			

(b)

Ads by Google
Latanya Sweeney, Arrested?
 1) Enter Name and State. 2) Access Full Background Checks Instantly.
www.instantcheckmate.com/

Latanya Sweeney
 Public Records Found For: Latanya Sweeney. View Now.
www.publicrecords.com/

La Tanya
 Search for La Tanya Look Up Fast Results now!
www.ask.com/La+Tanya

(c)

LATANYA SWEENEY
 1420 Centre Ave
 Pittsburgh, PA 15219
 DOB: Oct 27, 1969 (53 years old)

Criminal History
 This section contains possible citation, arrest, and criminal records for the subject of this report.
 While our database does contain hundreds of millions of arrest records, different counties have different rules regarding what information they will and will not release.
 We share with you as much information as we possibly can, but a clean slate here should not be interpreted as a guarantee that Latanya Sweeney has never been arrested; it simply means that we were not able to locate any matching arrest records in the data that is available to us.

Possible Matching Arrest Records			
Name	County and State	Offenses	View Details
No matching arrest records were found.			

(d)

Assume the ads are free speech: what happens when these ads appear more often for one racial group than another? Not everyone is being equally affected by free speech. Is that free speech or racial discrimination?

Racism, as defined by the U.S. Commission on Civil Rights, is “any attitude, action, or institutional structure which subordinates a person or group because of their color.”¹⁶ Racial discrimination results when a person or group of people is treated differently based on their racial origins, according to the Panel on Methods for Assessing Discrimination of the National Research Council.¹² Power is a necessary precondition, for it depends on the ability to give or withhold benefits, facilities, services, and opportunities from someone who should be entitled to them and is denied on the basis of race. Institutional or structural racism, as defined in *The Social Work Dictionary*, is a system of procedures/patterns whose effect is to foster discriminatory outcomes or give preferences to members of one group over another.¹

These considerations frame the relevant socio-legal landscape. Now we turn to whether online ads suggestive of arrest records appear more often for one racial group than another among a sample of racially associated names, and if so, how technology can solve the problem.

The Pattern

What is the suspected pattern of ad delivery? Here is an overview using real-world examples.

Earlier this year, a Google search for *Latanya Farrell*, *Latanya Sweeney*, and *Latanya Lockett* yielded ads and criminal reports like those shown in Figure 1. The ads appeared on Google.com (Figure 1a, 1c) and on a news website, Reuters.com, to which Google supplies ads (Figure 1c). All the ads in question linked to instantcheckmate.com (Figure 1b, 1d). The first ad implied *Latanya Farrell* might have been arrested. Was she? Clicking on the link and paying the requisite fee revealed the company had no arrest record for her or *Latanya Sweeney*, but there is a record for *Latanya Lockett*.

In comparison, searches for *Kristen Haring*, *Kristen Sparrow*, and *Kristen Lindquist* did not yield any instant-

checkmate.com ads, even though the company's database reported having records for all three names and arrest records for Sparrow and Lindquist.

Searches for *Jill Foley*, *Jill Schneider*, and *Jill James* displayed instantcheckmate.com ads with neutral copy; the word *arrest* did not appear in the ads even though arrest records for all three names appeared in the company's database. Figure 2 shows ads appearing on Google.com and Reuters.com and criminal reports from instantcheckmate.com for the first two names.

Finally, we considered a proxy for race associated with these names. Figure 3 shows racial distinction in Google image search results for *Latanya*, *Latisha*, *Kristen*, and *Jill*, respectively. The faces associated with *Latanya* and *Latisha* tend to be black, while white faces dominate the images of *Kristen* and *Jill*.

These handpicked examples describe the suspected pattern: ads suggesting arrest tend to appear with names associated with blacks, and neutral or no ads appear with names associated with whites, regardless of whether the company placing the ad has an arrest record associated with the name.

Google Adsense

Who generates the ad's text? Who decides when and where an ad will appear? What is the relationship among Google, a news website such as Reuters, and Instant Checkmate in the previous examples? An overview of Google AdSense, the program that delivered the ads, provides the answers.

In printed newspapers, everyone who reads the publication sees the same ad in the same space. Online ads can be tailored to the reader's search criteria, interests, geographical location, and so on. Any two readers (or even the same reader returning to the same website) might view different ads.

Google AdSense is the largest provider of dynamic online advertisements, placing ads for millions of sponsors on millions of websites.⁹ In the first quarter of 2011, Google earned \$2.43 billion through Google AdSense.¹⁰ Several different advertising arrangements exist, but for simplicity this article describes only those features of Google AdSense specific to the Instant Checkmate ads in question.

Figure 2. Ad from a search of three different names beginning with the first name "Jill."

(a) Ad related to Jill Schneider

Jill Schneider Art
www.posters2prints.com/
Custom Frame Prints and Canvas. Shop Now, SAVE Big + Free Shipping!

We Found Jill Schneider
www.intelius.com/
Current Phone, Address, Age & More. Instant & Accurate Jill Schneider
10,256 people +1'd this page
Reverse Lookup - Reverse Cell Phone Directory - Date Check - Property Records

Located: Jill Schneider
www.instantcheckmate.com/
Information found on Jill Schneider Jill Schneider found in database.

(b) Instant Checkmate profile for Jill Schneider

JILL SCHNEIDER
1707 70th St
Kansas City, MO 64118
DOB: Mar 31, 1969 (43 years old)

Certified

Personal: Name, aliases, birthdate, phone numbers, etc.

Location: Detailed address history and related data, maps, etc.

Related Persons: Known family members, business associates, roommates, etc.

Marriage / Divorce: Marriage and divorce records on file...

Criminal History: Arrest records, speeding tickets, mugshots, etc.

Licenses: FAA licenses, DEA licenses, Other Licenses, etc.

Sex Offenders: Sex offenders living near Jill Schneider's primary location.

Criminal History

This section contains possible citation, arrest, and criminal records for the subject of this report. While our database does contain hundreds of millions of arrest records, different counties have different rules regarding what information they will and will not release.

We share with you as much information as we possibly can, but a clean slate here should not be interpreted as a guarantee that Jill Schneider has never been arrested; it simply means that we were not able to locate any matching arrest records in the data that is available to us.

Possible Matching Arrest Records

Name	County and State	Offenses	View Details
1 Jill E Schneider	WI Admin Office of Courts(CM) disposition	Criminal/traffic	View Details
2 Jill E Schneider	WI Admin Office of Courts(CM)	Criminal/traffic	View Details
3 Jill E Schneider	WI Admin Office of Courts(CM) disposition	Criminal/traffic	View Details
4 Jill E Schneider	WI Admin Office of Courts(CM)	Criminal/traffic	View Details

(c) Ad related to Jill James

Located: Jill James
www.instantcheckmate.com/
Information found on Jill James Jill James found in database.

(d) Instant Checkmate profile for Jill James

JILL JAMES
105 Seabreeze Ct
Cary, NC 27513
DOB: May 31, 1958 (54 years old)

Certified

Personal: Name, aliases, birthdate, phone numbers, etc.

Location: Detailed address history and related data, maps, etc.

Related Persons: Known family members, business associates, roommates, etc.

Marriage / Divorce: Marriage and divorce records on file...

Criminal History: Arrest records, speeding tickets, mugshots, etc.

Licenses: FAA licenses, DEA licenses, Other Licenses, etc.

Sex Offenders: Sex offenders living near Jill James's primary location.

Criminal History

This section contains possible citation, arrest, and criminal records for the subject of this report. While our database does contain hundreds of millions of arrest records, different counties have different rules regarding what information they will and will not release.

We share with you as much information as we possibly can, but a clean slate here should not be interpreted as a guarantee that Jill James has never been arrested; it simply means that we were not able to locate any matching arrest records in the data that is available to us.

Possible Matching Arrest Records

Name	County and State	Offenses	View Details
1 Jill B James	NC Admin Office of Courts demographic criminal	Criminal/traffic	View Details
2 Jill James	NC Admin Office of Courts demographic criminal	Criminal/traffic	View Details
3 Jill James	Individual NC courts	Criminal/traffic	View Details
4 Jill B James	Individual NC courts	Criminal/traffic	View Details
5 Jill Pate James	Individual NC courts	Criminal/traffic	View Details
6 Jill Pate James	NC Admin Office of Courts demographic criminal	Criminal/traffic	View Details
7 Jill Kelly James	NC Admin Office of Courts demographic criminal	Criminal/traffic	View Details
8 Jill Kelly James	Individual NC courts	Criminal/traffic	View Details
9 Jill Rosamond James	NC Admin Office of Courts demographic infractions	Criminal/traffic	View Details
10 Jill Rosamond James	NC Admin Office of Courts demographic criminal	Criminal/traffic	View Details

When a reader enters search criteria in an enrolled website, Google AdSense embeds into the Web page of results ads believed to be relevant to the search. Figures 1 and 2 show ads delivered by Google AdSense in response to various *firstname lastname* searches.

An advertiser provides Google with search criteria, copies of possible ads to deliver, and a bid to pay if a reader clicks the delivered ad. (For convenience, this article conflates Google AdSense with the related Google Adwords.) Google operates a real-time auction across bids for the same search criteria based on a “quality score” for each bid. A quality score includes many factors such as the past performance of the ad and characteristics of the company’s website.¹⁰ The ad having the highest quality score appears first, the second-highest second, and so on, and Google may elect not to show any ad if it considers the bid too low or if showing the ad exceeds a threshold (For example, a maximum account total for the advertiser). The Instant Checkmate ads in figures 1 and 2 often appeared first among ads, implying Instant Checkmate ads had the highest quality scores.

A website owner wanting to “host” online ads enrolls in AdSense and modifies the website to send a user’s search criteria to Google and to display returning ads under a banner “Ads by Google” among search results. For example, Reuters.com hosts AdSense, and entering *Latanya Sweeney* in the

search bar generated a new Web page with ads under the banner “Ads by Google” (Figure 1c).

There is no cost for displaying an ad, but if the user actually clicks on the ad, the advertiser pays the auction price. This may be as little as a few pennies, and the amount is split between Google and the host. Clicking the *Latanya Sweeney* ad on Reuters.com (Figure 1c) would cause Instant Checkmate to pay its auction amount to Google, and Google would split the amount with Reuters.

Search Criteria

What search criteria did Instant Checkmate specify? Will ads be delivered for made-up names? Ads displayed on Google.com allow users to learn why a specific ad appeared. Clicking the circled “i” in the ad banner (for example, Figure 1c) leads to a Web page explaining the ads. Doing so for ads in figures 1 and 2 reveals that the ads appeared because the search criteria matched the exact first- and last-name combination searched.

So, the search criteria must consist of both first and last names; and the names should belong to real people because a company presumably bids on records it sells.

The next steps describe the systematic construction of a list of racially associated first and last names for real people to use as search criteria. Neither Instant Checkmate nor Google are presumed to have used such a list.

Rather, the list provides a qualified sample of names to use in testing ad-delivery systems.

Black- and White-Identifying Names

Black-identifying and white-identifying first names occur with sufficiently higher frequency in one race than the other.

In 2003 Marianne Bertrand and Sendhil Mullainathan of the National Bureau of Economic Research (NBER) conducted an experiment in which they provided resumes to job posts that were virtually identical, except some of the resumes had black-identifying names and others had white-identifying names. Results showed white names received 50% more interviews.²

The study used names given to black and white babies in Massachusetts between 1974 and 1979, defining black-identifying and white-identifying names as those that have the highest ratio of frequency in one racial group to frequency in the other racial group.

In the popular book *Freakonomics*, Steven Levitt and Stephen Dubner report the top 20 whitest- and blackest-identifying girl and boy names. The list comes from earlier work by Levitt and Roland Fryer, which shows a pattern change in the way blacks named their children starting in the 1970s.⁷ It was compiled from names given to black and white children recorded in California birth records from 1961–2000 (more than 16 million births).

To test ad delivery, I combined the lists from these prior studies and added two black female names, *Latanya* and *Latisha*. Table 1 lists the names used here, consisting of eight for each of the categories: white female, black female, white male, and black male from the Bertrand and Mullainathan study (first row in Table 1); and the first eight names for each category from the Fryer and Levitt work (second row in Table 1). *Emily*, a white female name, *Ebony*, a black female name, and *Darnell*, a black male name, appear in both rows. The third row includes the observation shown in Figure 3. Removing duplicates leaves a total of 63 distinct first names.

Full Names of Real People

Web searches provide a means of locating and harvesting a real person’s first and last name (full name) by sampling

Table 1. Black-identifying names and white-identifying first names.

	White Female	Black Female	White Male	Black Male
(a)	Allison	Aisha	Brad	Darnell
	Anne	Ebony	Brendan	Hakim
	Carrie	Keisha	Geoffrey	Jermaine
	Emily	Kenya	Greg	Kareem
	Jill	Latanya	Brett	Jamal
	Laurie	Lakisha	Jay	Leroy
	Kristen	Latoya	Matthew	Rasheed
	Meredith	Tamika	Neil	Tremayne
(b)	Molly	Imani	Jake	DeShawn
	Amy	Ebony*	Connor	DeAndre
	Claire	Shanice	Tanner	Marquis
	Emily*	Aaliyah	Wyatt	Darnell*
	Katie	Precious	Cody	Terrell
	Madeline	Nia	Dustin	Malik
	Katelyn	Deja	Luke	Trevon
	Emma	Diamond	Jack	Tyrone
(c)		Latanya Latisha		

names of professionals appearing on the Web; and sampling names of people active on social media sites and blogs (netizens).

Professionals often have their own Web pages that list positions and describe prior accomplishments. Several professions have degree designations (for example, Ph.D., M.D., J.D., or MBA) associated with people in that profession. A Google search for a first name and a degree designation can yield lists of people having that first name and degree.

The next step is to visit the Web page associated with each full name, and if an image is discernible, record whether the person appears black, white, or other.

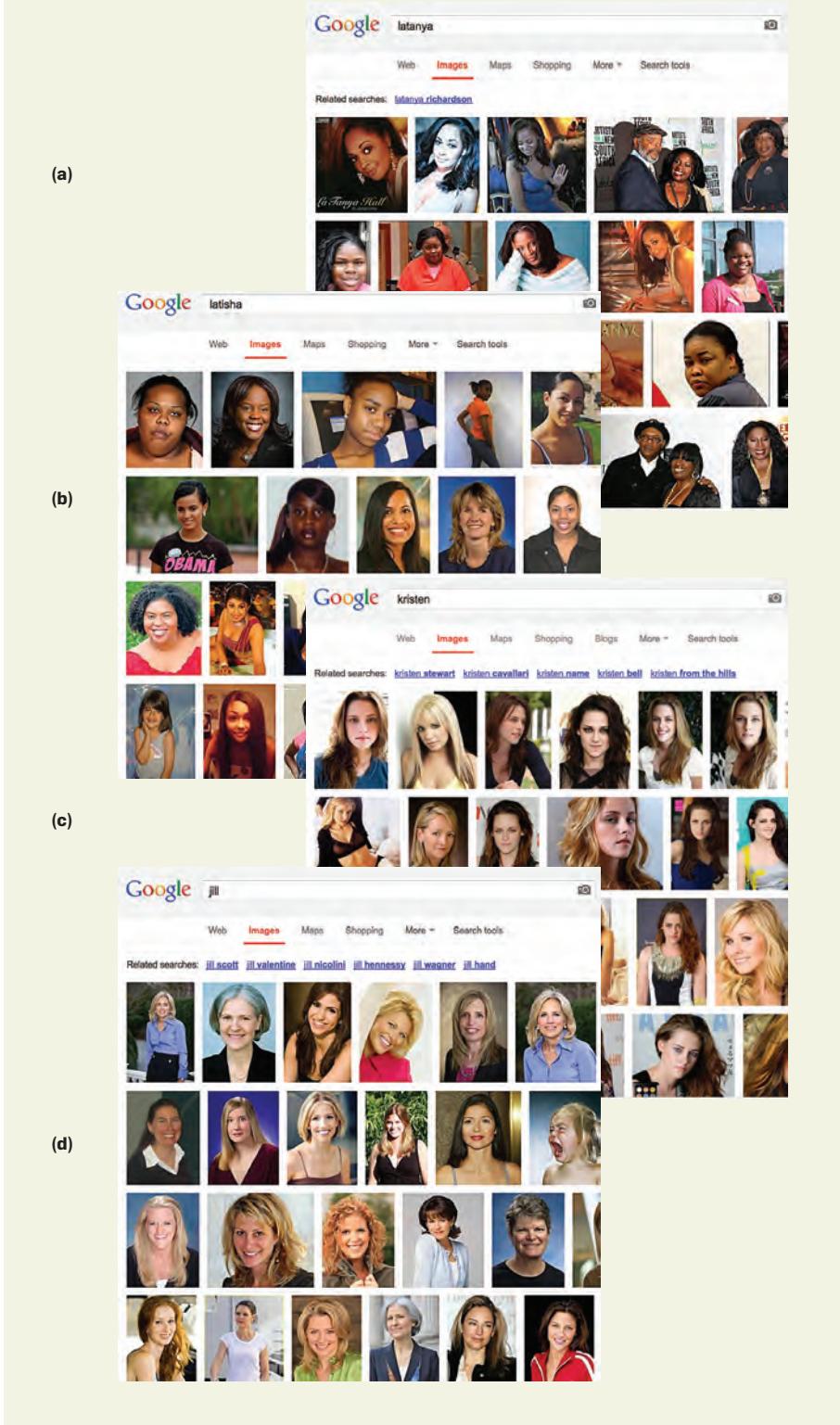
Here are two examples from my test. A Google search for *Ebony PhD* revealed links for real people having *Ebony* as a first name—specifically, *Ebony Bookman*, *Ebony Glover*, *Ebony Baylor*, and *Ebony Utley*. I harvested the full names appearing on the first three pages of search results, using searches with other degree designations to find at least 10 full names for *Ebony*. Clicking on the link associated with *Ebony Glover* displayed an image.⁸ The *Ebony Glover* in this study appeared black.

Similarly, search results for *Jill PhD* listed professionals whose first name is *Jill*. Visiting links yielded Web pages with more information about each person. For example, *Jill Schneider*'s Web page had an image showing that she is white.¹⁴

PeekYou searches were used to harvest a sample of full names of netizens having racially associated first names. The website peekyou.com compiles online and offline information on individuals—thereby connecting residential information with Facebook and Twitter users, bloggers, and others—then assigns its own rating to reflect the size of each person's online footprint. Search results from peekyou.com list people having the highest score first, and include an image of the person.

A PeekYou search of *Ebony* listed *Ebony Small*, *Ebony Cams*, *Ebony King*, *Ebony Springer*, and *Ebony Tan*. A PeekYou search for *Jill* listed *Jill Christopher*, *Jill Spivack*, *Jill English*, *Jill Pantoza*, and *Jill Dobson*. After harvesting these and other full names, I reported the race of the person if discernible.

Figure 3. Image search results for first names Latanya, Latisha, Kirsten, and Jill.



Armed with the approach just described, I harvested 2,184 racially associated full names of people with an online presence from September 24 through October 22, 2012. Most images associated with black-identifying names were of black people (88%),

and an even greater percentage of images associated with white-identifying names were of white people (96%).¹⁵

Google searches of first names and degree designations were not as productive as first name lookups on PeekYou. On Google, white male

names, *Cody*, *Connor*, *Tanner*, and *Wyatt* retrieved results with those as last names rather than first names; the black male name, *Kenya*, was confused with the country; and black names *Aaliyah*, *Deja*, *Diamond*, *Hakim*, *Malik*, *Marquis*, *Nia*, *Precious*, and *Rasheed* retrieved fewer than 10 full names. Only *Diamond* posed a problem with PeekYou searches, seemingly confused with other online entities. *Diamond* was therefore excluded from further consideration.

Some black first names had perfect predictions (100%): *Aaliyah*, *DeAndre*, *Imani*, *Jermaine*, *Lakisha*, *Latoya*, *Malik*, *Tamika*, and *Trevon*. The worst predictors of blacks were *Jamal* (48%) and *Leroy* (50%). Among white first names, 12 of 31 names made perfect predictions: *Brad*, *Brett*, *Cody*, *Dustin*, *Greg*, *Jill*, *Katelyn*, *Katie*, *Kristen*, *Matthew*, *Tanner*, and *Wyatt*; the worst predictors of whites were *Jay* (78%) and *Brendan* (83%). These findings strongly support the use of these names as racial indicators in this study.

Sixty-two full names appeared in the list twice even though the people were not necessarily the same. No name appeared more than twice. Overall, Google and PeekYou searches tended to yield different names.

Ad Delivery

With this list of names suggestive of race, I was ready to test which ads appear when these names are searched. To do this, I examined ads delivered on two sites, Google.com and Reuters.com, in response to searches of each full name, once at each site. The browser's cache and cookies were cleared before each search, and copies of Web pages received were preserved. Figures 1, 2, 5, and 6 provide examples.

From September 24 through October 23, 2012, I searched 2,184 full names on Google.com and Reuters.com. The searches took place at different times of day, different days of the week, with different IP and machine addresses operating in different parts of the United States using different browsers. I manually searched 1,373 of the names and used automated means¹⁷ for the remaining 812 names. Here are nine observations.

1. Fewer ads appeared on Google.com than Reuters.com—about five times

Of the more than 2,000 names searched, 78% had at least one ad for public records about the person being searched.

fewer. When ads did appear on Google.com, typically only one ad showed, compared with three ads routinely appearing on Reuters.com. This suggests Google may be sensitive to the number of ads appearing on Google.com.

2. Of 5,337 ads captured, 78% were for government-collected information (public records) about the person whose name was searched. Public records in the U.S. often include a person's address, phone number, and criminal history. Of the more than 2,000 names searched, 78% had at least one ad for public records about the person being searched.

3. Four companies had more than half of all the ads captured. These companies were Instant Checkmate, PublicRecords (which is owned by Intelius), PeopleSmart, and PeopleFinders, and all their ads were selling public records. Instant Checkmate ads appeared more than any other: 29% of all ads. Ad distribution was different on Google's site; Instant Checkmate still had the most ads (50%), but Intelius.com, while not in the top four overall, had the second most ads on Google.com. These companies dominate the advertising space for online ads selling public records.

4. Ads for public records on a person appeared more often for those with black-associated names than white-associated names, regardless of company. PeopleSmart ads appeared disproportionately higher for black-identifying names—41% as opposed to 29% for white names. PublicRecords ads appeared 10% more often for those with black first names than white. Instant Checkmate ads displayed only slightly more often for black-associated names (2% difference). This is an interesting finding and it spawns the question: Public records contain information on everyone, so why more ads for black-associated names?

5. Instant Checkmate ads dominated the topmost ad position. They occupied that spot in almost half of all searches on Reuters.com. This suggests Instant Checkmate offers Google more money or has higher quality scores than do its competitors.

6. Instant Checkmate had the largest percentage of ads in virtually every first-name category, except for *Kristen*, *Connor*, and *Tremayne*. For those names, Instant Checkmate had uncharacteristically fewer ads (less than 25%). Pub-

licRecords had ads for 80% of names beginning with *Tremayne*, and *Connor*, and 58% for *Kristen*, compared to 20% and less for Instant Checkmate. Why the underrepresentation in these first names? During a conference call with company's representatives, they asserted that Instant Checkmate gave the same ad text to Google for groups of last names (not first names).

7. *Almost all ads for public records included the name of the person, making each ad virtually unique, but beyond personalization, the ad templates showed little variability.* The only exception was Instant Checkmate. Almost all PeopleFinder ads appearing on Reuters.com used the same personalized template. PublicRecords used five templates and PeopleSmart seven, but Instant Checkmate used 18 different ad templates on Reuters.com. Figure 4 enumerates ad templates for frequencies of 10 or more for all four companies (replace *fullname* with the person's first and last name).

While Instant Checkmate's competitors also sell criminal history information, only Instant Checkmate ads used the word *arrest*.

8. *A greater percentage of Instant Checkmate ads using the word "arrest" appeared for black-identifying first names than for white first names.* More than 1,100 Instant Checkmate ads appeared on Reuters.com, with 488 having black-identifying first names; of these, 60% used *arrest* in the ad text. Of the 638 ads displayed with white-identifying names, 48% used *arrest*. This difference is statistically significant, with less than a 0.1% probability that the data can be explained by chance (chi-square test: $X^2(1)=14.32, p < 0.001$). The EEOC's and U.S. Department of Labor's adverse impact test for measuring discrimination is 77 in this case, so if this were an employment situation, a charge of discrimination might result. (The adverse impact test uses the ratio of neutral ads, or 100 minus the percentages given, to compute disparity: $100-60=40$ and $100-48=2$; dividing 40 by 2 equals 77.)

The highest percentage of neutral ads (where the word *arrest* does not appear in ad text) on Reuters.com were those for *Jill* (77%) and *Emma* (75%), both white-identifying names. Names receiving the highest percentage of ads with *arrest* in the text were *Darnell*

(84%), *Jermaine* (81%), and *DeShawn* (86%), all black-identifying first names. Some names appeared counter to this pattern: *Dustin*, a white-identifying name, generated *arrest* ads in 81% of searches; and *Imani*, a black-identifying name, resulted in neutral ads in 75% of searches.

9. *Discrimination results on Google's site were similar, but, interestingly, ad text and distributions were different.* While the same neutral and *arrest* ads having dominant appearances on Reuters.com also appeared frequently on Google.com, Instant Checkmate ads on Google included an additional 10 templates, all using the word *criminal* or *arrest*.

More than 400 Instant Checkmate ads appeared on Google, and 90% of these were suggestive of *arrest*, regardless of race. Still, a greater percentage of Instant Checkmate ads suggestive of *arrest* displayed for black-associated first names than for whites. Of the 366

ads that appeared for black-identifying names, 92% were suggestive of *arrest*. Far fewer ads displayed for white-identifying names (66 total), but 80% were suggestive of *arrest*. This difference in the ratios 92 and 80 is statistically significant, with less than a 1% probability that the data can be explained by chance (chi-square test: $X^2(1)=7.71, p < 0.01$). The EEOC's adverse impact test for measuring discrimination is 40%, so if this were employment, a charge of discrimination might result. (The adverse impact test gives $100-92=8$ and $100-80=20$; dividing 8 by 20 equals 40.)

A greater percentage of Instant Checkmate ads having the word *arrest* in ad text appeared for black-identifying first names than for white-identifying first names within professional and netizen subsets, too. On Reuters.com, which hosts Google AdSense ads, a black-identifying name was 25% more likely to generate an ad suggestive of an *arrest* record.

Figure 4. Template for ads for public records on Reuters for frequencies less than 10. Full list is available.¹⁵

instantcheckmate	Peoplesmart
382 Located: fullname Information found on fullname fullname found in database.	87 We found: fullname 1) Get Aisha's Background Report 2) Current Contact Info—Try Free!
96 We found fullname Search Arrests, Address, Phone, etc. Search records for fullname.	105 We found: fullname 1) Contact fullname—Free Info! 2) Current Address, Phone & More.
40 Background of fullname Search Instant Checkmate for the Records of fullname	348 We found: fullname 1) Contact fullname—Free Info! 2) Current Phone, Address & More.
17 fullname's Records 1) Enter Name and State. 2) Access Full Background Checks Instantly.	
Publicrecords	
195 fullname: Truth Arrests and Much More. Everything About fullname	570 fullname Public Records Found For: fullname. View now.
67 fullname Truth Looking for fullname? Check fullname's Arrests	128 fullname Public Records Found For: fullname. Search now.
176 fullname, Arrested? 1) Enter Name and State. 2) Access Full Background Checks Instantly.	13 Records: fullname Database of all lastname's in the Country. Search now.
55 fullname Located Background Check, Arrest Records, Phone, & Address. Instant, Accurate	56 fullname We have Public Records For: fullname. Search Now.
62 Looking for fullname? Comprehensive Background Report and More on fullname	
Peoplefinders	
	523 We found fullname Current Address, Phone and Age. Find fullname, Anywhere.

Figure 5. Senator Claire McCaskill's campaign ad appeared next to an ad using the word "arrest."

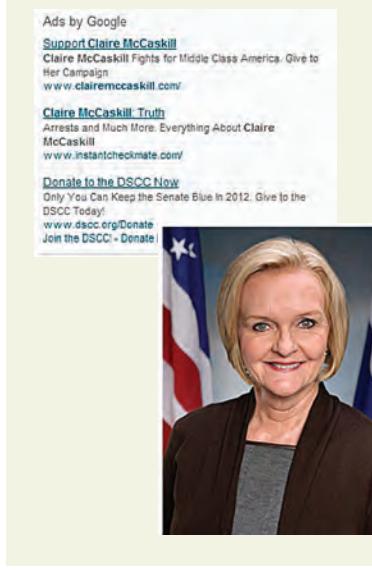


Figure 6. An assortment of ads appearing for Latisha Smith.



These findings reject the hypothesis that no difference exists in the delivery of ads suggestive of an arrest record based on searches of racially associated names.

Additional Observations

The people behind the names used in this study are diverse. Political figures included Maryland State Representatives Aisha Braveboy (arrest ad) and Jay Jacobs (neutral ad); Jill Biden (neutral ad), wife of U.S. Vice President Joe Biden; and Claire McCaskill, whose campaign ad for the U.S. Sen-

ate in Missouri appeared alongside an Instant Checkmate ad using the word *arrest* (Figure 5). Names mined from academic websites included graduate students, staff, and accomplished academics, such as Amy Gutmann, president of the University of Pennsylvania. Dustin Hoffman (arrest ad) was among names of celebrities used. A smorgasbord of athletes appeared, from local to national fame (assorted neutral and arrest ads). The youngest person whose name was used in the study was a missing 11-year-old black girl.

More than 1,100 of the names harvested for this study were from PeekYou, with scores estimating the name's overall presence on the Web. As expected, celebrities get the highest scores of 10s and 9s. Only four names used here had a PeekYou score of 10, and 12 had a score of 9, including Dustin Hoffman. Only two ads appeared for these high-scoring names; an abundance of ads appeared across the remaining spectrum of PeekYou scores. We might presume that the bid price needed to display an ad is greater for more popular names with higher PeekYou scores. Knowing that very few high-scoring people were in the study and that ads appeared across the full spectrum of PeekYou scores reduces concern about variations in bid prices.

Different Instant Checkmate ads sometimes appeared for the same person. About 200 names had Instant Checkmate ads on both Reuters.com and Google.com, but only 42 of these names received the same ad. The other 82% of names received different ads across the two sites. At most, three distinct ads appeared across Reuters.com and Google.com for the same name. Figure 6 shows the assortment of ads appearing for *Latisha Smith*. Having different possible ad texts for a name reminds us that while Instant Checkmate provided the ad texts, Google's technology selected among the possible texts in deciding which to display. Figure 6 shows ads both suggestive of arrest and not, though more ads appear suggestive of arrest than not.

More About the Problem

Why is this discrimination occurring? Is Instant Checkmate, Google, or society to blame? We do not yet know. Google understands that an advertiser

may not know which ad copy will work best, so the advertiser may provide multiple templates for the same search string, and the "Google algorithm" learns over time which ad text gets the most clicks from viewers. It does this by assigning weights (or probabilities) based on the click history of each ad. At first, all possible ad texts are weighted the same and are equally likely to produce a click. Over time, as people tend to click one ad copy over others, the weights change, so the ad text getting the most clicks eventually displays more frequently.

Did Instant Checkmate provide ad templates suggestive of arrest disproportionately to black-identifying names? Or did Instant Checkmate provide roughly the same templates evenly across racially associated names but users clicked ads suggestive of arrest more often for black-identifying names? As mentioned earlier, during a conference call with the founders of Instant Checkmate and their lawyer, the company's representatives asserted that Instant Checkmate gave the same ad text to Google for groups of last names (not first names) in its database; they expressed no other criteria for name and ad selection.

This study is a start, but more research is needed. To preserve research opportunities, I captured additional results for 50 hits on 2,184 names across 30 Web sites serving Google Ads to learn the underlying distributions of ad occurrences per name. While analyzing the data may prove illuminating, in the end the basic message presented in this study does not change: there is discrimination in delivery of these ads.

Technical Solutions

How can technology solve this problem? One answer is to change the quality scores of ads to discount for unwanted bias. The idea is to measure real-time bias in an ad's delivery and then adjust the weight of the ad accordingly at auction. The general term for Google's technology is *ad exchange*. This approach generalizes to other ad exchanges (not just Google's); integrates seamlessly into the way ad exchanges operate, allowing minimal modifications to harmonize ad deliveries with societal norms; and, works regardless of the cause of the discrimi-

nation—advertiser bias in placing ads or society bias in selecting ads.

Discrimination, however, is at the heart of online advertising. Differential delivery is the very idea behind it. For example, if young women with children tend to purchase baby products and retired men with bass boats tend to purchase fishing supplies, and you know the viewer is one of these two types, then it is more efficient to offer ads for baby products to the young mother and fishing rods to the fisherman, not the other way around.

On the other hand, not all discrimination is desirable. Societies have identified groups of people to protect from specific forms of discrimination. Delivering ads suggestive of arrest much more often for searches of black-identifying names than for white-identifying names is an example of unwanted discrimination, according to American social and legal norms. This is especially true because the ads appear regardless of whether actual arrest records exist for the names in the company's database.

The good news is that we can use the mechanics and legal criteria described earlier to build technology that distinguishes between desirable and undesirable discrimination in ad delivery. Here I detail the four key components:

1. Identifying Affected Groups. A set of predicates can be defined to identify members of protected and comparison groups. Given an ad's search string and text, a predicate returns *true* if the ad can impact the group that is the subject of the predicate and returns *false* otherwise. Statistics of baby names can identify first names for constructing race and gender groups and last names for grouping some ethnicities. Special word lists or functions that report degree of membership may be helpful for other comparisons.

In this study, ads appeared on searches of full names for real people, and first names assigned to more black or white babies formed groups for testing. These *black* and *white* predicates evaluate to *true* or *false* based on the first name of the search string.

2. Specifying the Scope of Ads to Assess. The focus should be on those ads capable of impacting a protected group in a form of discrimination prohibited by law or social norm. Protec-

Discrimination is at the heart of online advertising. Differential delivery is the very idea behind it.

tion typically concerns the ability to give or withhold benefits, facilities, services, employment, or opportunities. Instead of lumping all ads together, it is better to use search strings, ad texts, products, or URLs that display with ads to decide which ads to assess.

This study assessed search strings of first and last names of real people, ads for public records, and ads having a specific display URL (instantcheckmate.com), the latter being the most informative because the adverse ads all had the same display URL.

Of course, the audience for the ads is not necessarily the people who are the subject of the ads. In this study, the audience is a person inquiring about the person whose name is the subject of the ad. This distinction is important when thinking about the identity of groups that might be impacted by an ad. Group membership is based on the ad's search string and text. The audience may resonate more with a distinctly positive or negative characterization of the group.

3. Determining Ad Sentiment. Originally associated with summarizing product and movie reviews, sentiment analysis is an area of computer science that uses natural-language processing and text analytics to determine the overall attitude of a writing.¹³ Sentiment analysis can measure whether an ad's search string and accompanying text has positive, negative, or neutral sentiment. A literature search does not find any prior application to online ads, but a lot of research has been done assessing sentiment in social media (sentiment140.com, for example, reports the sentiment of tweets, which like advertisements have limited words).

In this study, ads containing the word *arrest* or *criminal* were classified as having negative sentiment; ads without those words were classified as neutral.

4. Testing for Adverse Impact. Consider a table where columns are comparative groups, rows are sentiment, and values are the number of ad impressions (the number of times an ad appears, though the ad is not necessarily clicked). Ignore neutral ads. Comparing the percentage of ads having the same positive or negative sentiment across groups reveals the degree to which one group may be impacted more or less by the ad's sentiment.

Table 2. Negative and neutral sentiments of black and white groups.

	Black	White
Negative	291	60%
Neutral	197	40%
Positive		
Totals	488	638

A chi-square test can determine statistical significance, and the adverse impact test used by the EEOC and the U.S. Department of Labor can alert whether in some circumstances legal risks may result.

In this study the groups are black and white, and the sentiments are negative and neutral. Table 2 shows a summary chart. Of the 488 ads that appeared for the black group, 291 (or 60%) had negative sentiment. Of the 638 ads displayed for the white group, 308 (or 48%) had negative sentiment. The difference is statistically significant ($\chi^2(1)=14.32, p < 0.001$) and has an adverse impact measure of (40/52), or 77%.

An easy way of incorporating this analysis into an ad exchange is to decide which bias test is critical (for example, statistical significance or adverse impact test) and then factor the test result into the quality score for the ad at auction. For example, if we were to modify the ad exchange not to display any ad having an adverse impact score of less than 80, which is the EEOC standard, then arrest ads for blacks would sometimes appear, but would not be overly disproportionate to whites, regardless of advertiser or click bias.

Though this study served as an example throughout, the approach generalizes to many other forms of discrimination and combats other ways ad exchanges may foster discrimination.

Suppose female names tend to get neutral ads such as "Buy now," while male names tend to get positive ads such as "Buy now. 50% off!" Or suppose black names tend to get neutral ads such as "Looking for Ebony Jones," while white names tend to get positive ads such as "Meredith Jones. Fantastic!" Then the same analysis would suppress some occurrences of the positive ads so as not to foster a discriminatory effect.

This approach does not stop the appearance of negative ads for a store

placed by a disgruntled customer or ads placed by competitors on brand names of the competition, unless these are deemed to be protected groups.

Nonprotected marketing discrimination can continue even to protected groups. For example, suppose search terms associated with blacks tend to get neutral ads for some music artists, while those associated with whites tend to get neutral ads for other music artists. All ads would appear regardless of the disproportionate distribution because the ads are not subject to suppression.

As a final example, this approach allows everyone to be negatively impacted as long as the impact is approximately the same. Suppose all ads for public records on all names, regardless of race, were equally suggestive of arrest and had almost the same number of impressions; then no ads suggestive of arrest would be suppressed.

Computer scientist Cynthia Dwork and her colleagues have been working on algorithms that assure racial fairness.⁴ Their general notion is to ensure similar groups receive similar ads in proportions consistent with the population. Utility is the critical concern with this direction because not all forms of discrimination are bad, and unusual and outlier ads could be unnecessarily suppressed. Still, their research direction looks promising.

In conclusion, this study demonstrates that technology can foster discriminatory outcomes, but it also shows that technology can thwart unwanted discrimination.

Acknowledgments

The author thanks Ben Edelman, Claudine Gay, Gary King, Annie Lewis, and weekly Topics in Privacy participants (David Abrams, Micah Altman, Merce Crosas, Bob Gelman, Harry Lewis, Joe Pato, and Salil Vadhan) for discussions; Adam Tanner for first suspecting a pattern; Diane Lopez and Matthew Fox in Harvard's Office of the General Counsel for making publication possible in the face of legal threats; and Sean Hooley for editorial suggestions. Data from this study is available at foreverdata.org and the IQSS Dataverse Network. Supported in part by NSF grant CNS-1237235 and a gift from Google, Inc.

Related articles on queue.acm.org

Modeling People and Places with Internet Photo Collections

David Crandall, Noah Snavely
<http://queue.acm.org/detail.cfm?id=2212756>

Interactive Dynamics for Visual Analysis

Jeffrey Heer, Ben Shneiderman
<http://queue.acm.org/detail.cfm?id=2146416>

Social Perception

James L. Crowley
<http://queue.acm.org/detail.cfm?id=1147531r>

References

1. Barker R. *The Social Work Dictionary* (5th ed.). NASW Press, Washington, DC, ss, 2003.
2. Bertrand, M. and Mullainathan, S. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. NBER Working Paper No. 9873, 2003; <http://www.nber.org/papers/w9873>.
3. *Central Hudson Gas & Electric Corp. v. Public Service Commission of New York*. Supreme Court of the United States, 447 U.S. 557, 1980.
4. Dwork, C., Hardt, M., et al. 2011. Fairness through awareness. arXiv:1104.3913; <http://arxiv.org/abs/1104.3913>.
5. Equal Employment Opportunity Commission. Consideration of arrest and conviction records in employment decisions under Title VII of the Civil Rights Act of 1964. Washington, DC, 915.002, 2012. http://www.eeoc.gov/laws/guidance/arrest_conviction.cfm.
6. Equal Employment Opportunity Commission. Uniform guidelines on employee selection procedures. Washington, DC, 1978.
7. Fryer, R. and Levitt, S. The causes and consequences of distinctively black names. *The Quarterly Journal of Economics* 59, 3 (2004); <http://pricetheory.uchicago.edu/levitt/Papers/FryerLevitt2004.pdf>.
8. Glover, E.; <http://www.physiology.emory.edu/FIRST/ebonyc2.htm> (archived at <http://foreverdata.org/onleads>).
9. Google AdSense; <http://google.com/adsense>.
10. Google. Google announces first quarter 2011 financial results; http://investor.google.com/earnings/2011/Q1_google_earnings.html.
11. Harris, P. and Keller, K. Ex-offenders need not apply: The criminal background check in hiring decisions. *Journal of Contemporary Criminal Justice* 21, 1 (2005), 6-30.
12. Panel on Methods for Assessing Discrimination, National Research Council. Measuring racial discrimination. National Academy Press, Washington, DC, 2004.
13. Pang, B. and Lee, L. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics* (2004).
14. Schneider, J. <http://www.lehigh.edu/bio/jill.html> (Archived at <http://foreverdata.org/onleads>).
15. Sweeney, L. Discrimination in online ad delivery (2013). (For details, see full technical report at <http://ssrn.com/abstract=2208240>. Data, including Web pages and ads, archived at <http://foreverdata.org/onleads>).
16. U.S. Commission on Civil Rights. Racism in America and how to combat it. Washington, DC, 1970.
17. WebShot Command Line Server Edition. Version 1.9.1.1; <http://www.websitesscreenshots.com/>.

Latanya Sweeney (latanya@fas.harvard.edu) is professor of government and technology in residence at Harvard University. She creates and uses technology to assess and solve societal, political, and governance problems and teaches others how to do the same. She is also founder and director of the Data Privacy Lab at Harvard.

Amit Datta*, Michael Carl Tschantz, and Anupam Datta

Automated Experiments on Ad Privacy Settings

A Tale of Opacity, Choice, and Discrimination

Abstract: To partly address people's concerns over web tracking, Google has created the Ad Settings webpage to provide information about and some choice over the profiles Google creates on users. We present AdFisher, an automated tool that explores how user behaviors, Google's ads, and Ad Settings interact. AdFisher can run browser-based experiments and analyze data using machine learning and significance tests. Our tool uses a rigorous experimental design and statistical analysis to ensure the statistical soundness of our results. We use AdFisher to find that the Ad Settings was opaque about some features of a user's profile, that it does provide some choice on ads, and that these choices can lead to seemingly discriminatory ads. In particular, we found that visiting webpages associated with substance abuse changed the ads shown but not the settings page. We also found that setting the gender to female resulted in getting fewer instances of an ad related to high paying jobs than setting it to male. We cannot determine who caused these findings due to our limited visibility into the ad ecosystem, which includes Google, advertisers, websites, and users. Nevertheless, these results can form the starting point for deeper investigations by either the companies themselves or by regulatory bodies.

Keywords: blackbox analysis, information flow, behavioral advertising, transparency, choice, discrimination

DOI 10.1515/popets-2015-0007

Received 11/22/2014; revised 2/18/2015; accepted 2/18/2015.

1 Introduction

Problem and Overview. With the advancement of tracking technologies and the growth of online data aggregators, data collection on the Internet has become a

serious privacy concern. Colossal amounts of collected data are used, sold, and resold for serving targeted content, notably advertisements, on websites (e.g., [1]). Many websites providing content, such as news, outsource their advertising operations to large third-party ad networks, such as Google's DoubleClick. These networks embed tracking code into webpages across many sites providing the network with a more global view of each user's behaviors.

People are concerned about behavioral marketing on the web (e.g., [2]). To increase transparency and control, Google provides Ad Settings, which is “a Google tool that helps you control the ads you see on Google services and on websites that partner with Google” [3]. It displays inferences Google has made about a user's demographics and interests based on his browsing behavior. Users can view and edit these settings at

<http://www.google.com/settings/ads>

Yahoo [4] and Microsoft [5] also offer personalized ad settings.

However, they provide little information about how these pages operate, leaving open the question of how completely these settings describe the profile they have about a user. In this study, we explore how a user's behaviors, either directly with the settings or with content providers, alter the ads and settings shown to the user and whether these changes are in harmony. In particular, we study the degree to which the settings provides transparency and choice as well as checking for the presence of discrimination. Transparency is important for people to understand how the use of data about them affects the ads they see. Choice allows users to control how this data gets used, enabling them to protect the information they find sensitive. Discrimination is an increasing concern about machine learning systems and one reason people like to keep information private [6, 7].

To conduct these studies, we developed AdFisher, a tool for automating randomized, controlled experiments for studying online tracking. Our tool offers a combination of automation, statistical rigor, scalability, and explanation for determining the use of information by web advertising algorithms and by personalized ad settings, such as Google Ad Settings. The tool can simulate having a particular interest or attribute by visiting web-

*Corresponding Author: Amit Datta: Carnegie Mellon University, E-mail: amitdatta@cmu.edu

Michael Carl Tschantz: International Computer Science Institute, E-mail: mct@icsi.berkeley.edu

Anupam Datta: Carnegie Mellon University, E-mail: danupam@cmu.edu

pages associated with that interest or by altering the ad settings provided by Google. It collects ads served by Google and also the settings that Google provides to the simulated users. It automatically analyzes the data to determine whether statistically significant differences between groups of agents exist. AdFisher uses machine learning to automatically detect differences and then executes a test of significance specialized for the difference it found.

Someone using AdFisher to study behavioral targeting only has to provide the behaviors the two groups are to perform (e.g., visiting websites) and the measurements (e.g., which ads) to collect afterwards. AdFisher can easily run multiple experiments exploring the causal connections between users' browsing activities, and the ads and settings that Google shows.

The advertising ecosystem is a vast, distributed, and decentralized system with several players including the users consuming content, the advertisers, the publishers of web content, and ad networks. With the exception of the user, we treat the entire ecosystem as a blackbox. We measure simulated users' interactions with this blackbox including page views, ads, and ad settings. Without knowledge of the internal workings of the ecosystem, we cannot assign responsibility for our findings to any single player within it nor rule out that they are unintended consequences of interactions between players. However, our results show the presence of concerning effects illustrating the existence of issues that could be investigated more deeply by either the players themselves or by regulatory bodies with the power to see the internal dynamics of the ecosystem.

Motivating Experiments. In one experiment, we explored whether visiting websites related to substance abuse has an impact on Google's ads or settings. We created an experimental group and a control group of agents. The browser agents in the experimental group visited websites on substance abuse while the agents in the control group simply waited. Then, both groups of agents collected ads served by Google on a news website.

Having run the experiment and collected the data, we had to determine whether any difference existed in the outputs shown to the agents. One way would be to intuit what the difference could be (e.g. more ads containing the word "alcohol") and test for that difference. However, developing this intuition can take considerable effort. Moreover, it does not help find unexpected differences. Thus, we instead used machine learning to automatically find differentiating patterns in the data. Specifically, AdFisher finds a classifier that can pre-

dict which group an agent belonged to, from the ads shown to an agent. The classifier is trained on a subset of the data. A separate test subset is used to determine whether the classifier found a statistically significant difference between the ads shown to each group of agents. In this experiment, AdFisher found a classifier that could distinguish between the two groups of agents by using the fact that only the agents that visited the substance abuse websites received ads for Watershed Rehab.

We also measured the settings that Google provided to each agent on its Ad Settings page after the experimental group of agents visited the webpages associated with substance abuse. We found no differences (significant or otherwise) between the pages for the agents. Thus, information about visits to these websites is indeed being used to serve ads, but the Ad Settings page does not reflect this use in this case. Rather than providing transparency, in this instance, the ad settings were *opaque* as to the impact of this factor.

In another experiment, we examined whether the settings provide *choice* to users. We found that removing interests from the Google Ad Settings page changes the ads that a user sees. In particular, we had both groups of agents visit a site related to online dating. Then, only one of the groups removed the interest related to online dating. Thereafter, the top ads shown to the group that kept the interest were related to dating but not the top ads shown to the other group. Thus, the ad settings do offer the users a degree of choice over the ads they see.

We also found evidence suggestive of *discrimination* from another experiment. We set the agents' gender to female or male on Google's Ad Settings page. We then had both the female and male groups of agents visit webpages associated with employment. We established that Google used this gender information to select ads, as one might expect. The interesting result was how the ads differed between the groups: during this experiment, Google showed the simulated males ads from a certain career coaching agency that promised large salaries more frequently than the simulated females, a finding suggestive of discrimination. Ours is the first study that provides statistically significant evidence of an instance of discrimination in online advertising when demographic information is supplied via a transparency-control mechanism (i.e., the Ad Settings page).

While neither of our findings of opacity or discrimination are clear violations of Google's privacy policy [8] and we do not claim these findings to generalize or imply widespread issues, we find them concerning and warranting further investigation by those with visibility into

the ad ecosystem. Furthermore, while our finding of discrimination in the non-normative sense of the word is on firm statistical footing, we acknowledge that people may disagree about whether we found discrimination in the normative sense of the word. We defer discussion of whether our findings suggest unjust discrimination until Section 7.

Contributions. In addition to the experimental findings highlighted above, we provide AdFisher, a tool for *automating* such experiments. AdFisher is structured as a Python API providing functions for setting up, running, and analyzing experiments. We use Selenium to drive Firefox browsers and the scikit-learn library [9] for implementations of classification algorithms. We use the SciPy library [10] for implementing the statistical analyses of the core methodology.

AdFisher offers *rigor* by performing a carefully designed experiment. The statistical analyses techniques applied do not make questionable assumptions about the collected data. We base our design and analysis on a prior proposal that makes no assumptions about the data being independent or identically distributed [11]. Since advertisers update their behavior continuously in response to unobserved inputs (such as ad auctions) and the experimenters' own actions, such assumptions may not always hold. Indeed, in practice, the distribution of ads changes over time and simulated users, or *agents*, interfere with one another [11].

Our automation, experimental design, and statistical analyses allow us to *scale* to handling large numbers of agents for finding subtle differences. In particular, we modify the prior analysis of Tschantz et al. [11] to allow for experiments running over long periods of time. We do so by using *blocking* (e.g., [12]), a nested statistical analysis not previously applied to understanding web advertising. The blocking analysis ensures that agents are only compared to the agents that start out like it and then aggregates together the comparisons across blocks of agents. Thus, AdFisher may run agents in batches spread out over time while only comparing those agents running simultaneously to one another.

AdFisher also provides *explanations* as to how Google alters its behaviors in response to different user actions. It uses the trained classifier model to find which features were most useful for the classifier to make its predictions. It provides the top features from each group to provide the experimenter/analyst with a qualitative understanding of how the ads differed between the groups.

To maintain statistical rigor, we carefully circumscribe our claims. We only claim statistical soundness of our results: if our techniques detect an effect of the browsing activities on the ads, then there is indeed one with high likelihood (made quantitative by a p-value). We do not claim that we will always find a difference if one exists, nor that the differences we find are typical of those experienced by users. Furthermore, while we can characterize the differences, we cannot assign blame for them since either Google or the advertisers working with Google could be responsible.

Contents. After covering prior work next, we present, in Section 3, privacy properties that our tool AdFisher can check: nondiscrimination, transparency, and choice. Section 4 explains the methodology we use to ensure sound conclusions from using AdFisher. Section 5 presents the design of AdFisher. Section 6 discusses our use of AdFisher to study Google's ads and settings. We end with conclusions and future work.

Raw data and additional details about AdFisher and our experiments can be found at

<http://www.cs.cmu.edu/~mtschant/ife/>

AdFisher is freely available at

<https://github.com/tadatitam/info-flow-experiments/>

2 Prior Work

We are not the first to study how Google uses information. The work with the closest subject of study to ours is by Wills and Tatar [13]. They studied both the ads shown by Google and the behavior of Google's Ad Settings (then called the "Ad Preferences"). Like us, they find the presence of opacity: various interests impacted the ads and settings shown to the user and that ads could change without a corresponding change in Ad Settings. Unlike our study, theirs was mostly manual, small scale, lacked any statistical analysis, and did not follow a rigorous experimental design. Furthermore, we additionally study choice and discrimination.

Other related works differ from us in both goals and methods. They all focus on how visiting webpages change the ads seen. While we examine such changes in our work, we do so as part of a larger analysis of the interactions between ads and personalized ad settings, a topic they do not study.

Barford et al. come the closest in that their recent study looked at both ads and ad settings [14]. They do so in their study of the "adscape", an attempt to understand each ad on the Internet. They study each ad

individually and cast a wide net to analyze many ads from many websites while simulating many different interests. They only examine the ad settings to determine whether they successfully induced an interest. We rigorously study how the settings affects the ads shown (choice) and how behaviors can affect ads without affecting the settings (transparency). Furthermore, we use focused collections of data and an analysis that considers all ads collectively to find subtle causal effects within Google’s advertising ecosystem. We also use a randomized experimental design and analysis to ensure that our results imply causation.

The usage study closest to ours in statistical methodology is that of Tschantz et al. [11]. They developed a rigorous methodology for determining whether a system like Google uses information. Due to limitations of their methodology, they only ran small-scale studies. While they observed that browsing behaviors could affect Ad Settings, they did not study how this related to the ads received. Furthermore, while we build upon their methodology, we automate the selection of an appropriate test statistic by using machine learning whereas they manually selected test statistics.

The usage study closest to ours in terms of implementation is that of Liu et al. in that they also use machine learning [15]. Their goal is to determine whether an ad was selected due to the content of a page, by using behavioral profiling, or from a previous webpage visit. Thus, rather than using machine learning to select a statistical test for finding causal relations, they do so to detect whether an ad on a webpage matches the content on the page to make a case for the first possibility. Thus, they have a separate classifier for each interest a webpage might cover. Rather than perform a statistical analysis to determine whether treatment groups have a statistically significant difference, they use their classifiers to judge the ratio of ads on a page unrelated to the page’s content, which they presume indicates that the ads were the result of behavioral targeting.

Lécuyer et al. present XRay, a tool that looks for correlations between the data that web services have about users and the ads shown to users [16]. While their tool may check many changes to a type of input to determine whether any of them has a correlation with the frequency of a single ad, it does not check for causation, as ours does.

Engelhardt et al. study filter bubbles with an analysis that assumes independence between observations [17], an assumption we are uncomfortable making. (See Section 4.4.)

Guha et al. compare ads seen by three agents to see whether Google treats differently the one that behaves differently from the other two [18]. We adopt their suggestion of focusing on the title and URL displayed on ads when comparing ads to avoid noise from other less stable parts of the ad. Our work differs by studying the ad settings in addition to the ads and by using larger numbers of agents. Furthermore, we use rigorous statistical analyses. Balebako et al. run similar experiments to study the effectiveness of privacy tools [19].

Sweeney ran an experiment to determine that searching for names associated with African-Americans produced more search ads suggestive of an arrest record than names associated with European-Americans [20]. Her study required considerable insight to determine that suggestions of an arrest was a key difference. Ad-Fisher can automate not just the collection of the ads, but also the identification of such key differences by using its machine learning capabilities. Indeed, it found on its own that simulated males were more often shown ads encouraging the user to seek coaching for high paying jobs than simulated females.

3 Privacy Properties

Motivating our methodology for finding causal relationships, we present some properties of ad networks that we can check with such a methodology in place. As a fundamental limitation of science, we can only prove the existence of a causal effect; we cannot prove that one does not exist (see Section 4.5). Thus, experiments can only demonstrate violations of nondiscrimination and transparency, which require effects. On the other hand, we can experimentally demonstrate that effectful choice and ad choice are complied with in the cases that we test since compliance follows from the existence of an effect. Table 1 summarizes these properties.

3.1 Discrimination

At its core, *discrimination* between two classes of individuals (e.g., one race vs. another) occurs when the attribute distinguishing those two classes causes a change in behavior toward those two classes. In our case, discrimination occurs when membership in a class causes a change in ads. Such discrimination is not always bad (e.g., many would be comfortable with men and women receiving different clothing ads). We limit our discuss-

Property Name	Requirement	Causal Test	Finding
Nondiscrimination	Users differing only on protected attributes are treated similarly	Find that presence of protected attribute causes a change in ads	Violation
Transparency	User can view all data about him used for ad selection	Find attribute that causes a change in ads, not in settings	Violation
Effectful choice	Changing a setting has an effect on ads	Find that changing a setting causes a change in ads	Compliance
Ad choice	Removing an interest decreases the number ads related to that interest	Find setting causes a decrease in relevant ads	Compliance

Table 1. Privacy Properties Tested on Google’s Ad Settings

sion of whether the discrimination we found is unjust to the discussion section (§7) and do not claim to have a scientific method of determining the morality of discrimination.

Determining whether class membership causes a change in ads is difficult since many factors not under the experimenter’s control or even observable to the experimenter may also cause changes. Our experimental methodology determines when membership in certain classes causes significant changes in ads by comparing many instances of each class.

We are limited in the classes we can consider since we cannot create actual people that vary by the traditional subjects of discrimination, such as race or gender. Instead, we look at classes that function as surrogates for those classes of interest. For example, rather than directly looking at how gender affects people’s ads, we instead look at how altering a gender setting affects ads or at how visiting websites associated with each gender affects ads.

3.2 Transparency

Transparency tools like Google Ad Settings provide online consumers with some understanding of the information that ad networks collect and use about them. By displaying to users what the ad network may have learned about the interests and demographics of a user, such tools attempt to make targeting mechanisms more transparent.

However the technique for studying transparency is not clear. One cannot expect an ad network to be *completely transparent* to a user. This would involve the tool displaying all other users’ interests as well. A more reasonable expectation is for the ad network to display any inferred interests about that user. So, if an ad network has inferred some interest about a user and is serving

ads relevant to that interest, then that interest should be displayed on the transparency tool. However, even this notion of transparency cannot be checked precisely as the ad network may serve ads about some other interest correlated with the original inferred interest, but not display the correlated interest on the transparency tool.

Thus, we only study the extreme case of the lack of transparency — *opacity*, and leave complex notions of transparency open for future research. We say that a transparency tool has opacity if some browsing activity results in a significant effect on the ads served, but has no effect on the ad settings. If there is a difference in the ads, we can argue that prior browsing activities must have been tracked and used by the ad network to serve relevant ads. However, if this use does not show up on the transparency tool, we have found at least one example which demonstrates a lack of transparency.

3.3 Choice

The Ad Settings page offers users the option of editing the interests and demographics inferred about them. However, the exact nature of how these edits impact the ad network is unclear. We examine two notions of choice.

A very coarse form is *effectful choice*, which requires that altering the settings has some effect on the ads seen by the user. This shows that altering settings is not merely a “placebo button”: it has a real effect on the network’s ads. However, effectful choice does not capture whether the effect on ads is meaningful. For example, even if a user adds interests for cars and starts receiving fewer ads for cars, effectful choice is satisfied. Moreover, we cannot find violations of effectful choice. If we find no differences in the ads, we cannot conclude that users do not have effectful choice since it could be

the result of the ad repository lacking ads relevant to the interest.

Ideally, the effect on ads after altering a setting would be meaningful and related to the changed setting. One way such an effect would be meaningful, in the case of removing an inferred interest, is a decrease in the number of ads related to the removed interest. We call this requirement *ad choice*. One way to judge whether an ad is relevant is to check it for keywords associated with the interest. If upon removing an interest, we find a statistically significant decrease in the number of ads containing some keywords, then we will conclude that the choice was respected. In addition to testing for compliance in ad choice, we can also test for a violation by checking for a statistically significant increase in the number of related ads to find egregious violations. By requiring the effect to have a fixed direction, we can find both compliance and violations of ad choice.

4 Methodology

The goal of our methodology is to establish that a certain type of input to a system causes an effect on a certain type of output of the system. For example, in our experiments, we study the system of Google. The inputs we study are visits to content providing websites and users' interactions with the Ad Settings page. The outputs we study are the settings and ads shown to the users by Google. However, nothing in our methodology limits ourselves to these particular topics; it is appropriate for determining I/O properties of any web system. Here, we present an overview of our methodology; Appendix B provides details of the statistical analysis.

4.1 Background: Significance Testing

To establish causation, we start with the approach of Fisher (our tool's namesake) for significance testing [21] as specialized by Tschantz et al. for the setting of online systems [11]. Significance testing examines a *null hypothesis*, in our case, that the inputs do not affect the outputs. To test this hypothesis the experimenter selects two values that the inputs could take on, typically called the *control* and *experimental treatments*. The experimenter applies the treatments to *experimental units*. In our setting, the units are the browser agents, that is, simulated users. To avoid noise, the experimental units should initially be as close to identical as possible as

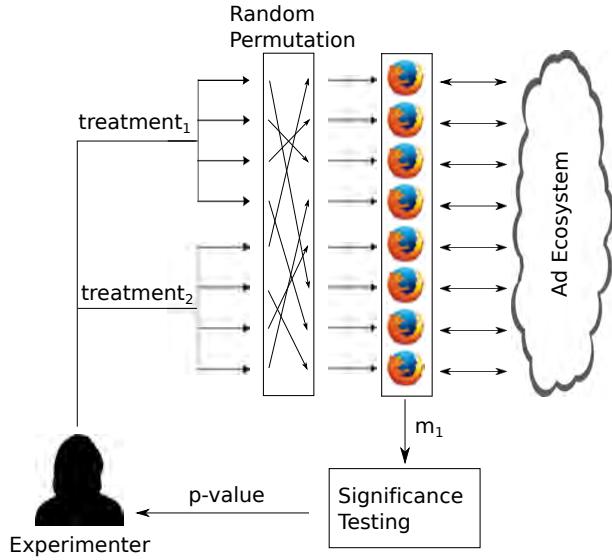


Fig. 1. Experimental setup to carry out significance testing on eight browser agents comparing the effects of two treatments. Each agent is randomly assigned a treatment which specifies what actions to perform on the web. After these actions are complete, they collect measurements which are used for significance testing.

far as the inputs and outputs in question are concerned. For example, an agent created with the Firefox browser should not be compared to one created with the Internet Explorer browser since Google can detect the browser used.

The experimenter randomly applies the experimental (control) treatment to half of the agents, which form the experimental (control) group. (See Figure 1.) Each agent carries out actions specified in the treatment applied to it. Next, the experimenter takes measurements of the outputs Google sends to the agents, such as ads. At this point, the experiment is complete and data analysis begins.

Data analysis starts by computing a *test statistic* over the measurements. The experimenter selects a test statistic that she suspects will take on a high value when the outputs to the two groups differ. That is, the statistic is a measure of distance between the two groups. She then uses the *permutation test* to determine whether the value the test statistic actually took on is higher than what one would expect by chance unless the groups actually differ. The permutation test randomly permutes the labels (control and experimental) associated with each observation, and recomputes a hypothetical test statistic. Since the null hypothesis is that the inputs have no effect, the random assignment should have no

effect on the value of the test statistic. Thus, under the null hypothesis, it is unlikely that the actual value of the test statistic is larger than the vast majority of hypothetical values.

The *p-value* of the permutation test is the proportion of the permutations where the test statistic was greater than or equal to the actual observed statistic. If the value of the test statistic is so high that under the null hypothesis it would take on as high of a value in less than 5% of the random assignments, then we conclude that the value is *statistically significant* (at the 5% level) and that causation is likely.

4.2 Blocking

In practice, the above methodology can be difficult to use since creating a large number of nearly identical agents might not be possible. In our case, we could only run ten agents in parallel given our hardware and network limitations. Comparing agents running at different times can result in additional noise since ads served to an agent change over time. Thus, with the above methodology, we were limited to just ten comparable units. Since some effects that the inputs have on Google’s outputs can be probabilistic and subtle, they might be missed looking at so few agents.

To avoid this limitation, we extended the above methodology to handle varying units using *blocking* [12]. To use blocking, we created *blocks* of nearly identical agents running in parallel. These agents differ in terms their identifiers (e.g., process id) and location in memory. Despite the agents running in parallel, the operating system’s scheduler determines the exact order in which the agents operate. Each block’s agents were randomly partitioned into the control and experimental groups. This randomization ensures that the minor differences between agents noted above should have no systematic impact upon the results: these differences become noise that probably disappears as the sample size increases. Running these blocks in a staged fashion, the experiment proceeds on block after block. A modified permutation test now only compares the actual value of the test statistic to hypothetical values computed by reassignments of agents that respect the blocking structure. These reassignments do not permute labels across blocks of observations.

Using blocking, we can scale to any number of agents by running as many blocks as needed. However, the computation of the permutation test increases exponentially with the number of blocks. Thus, rather than

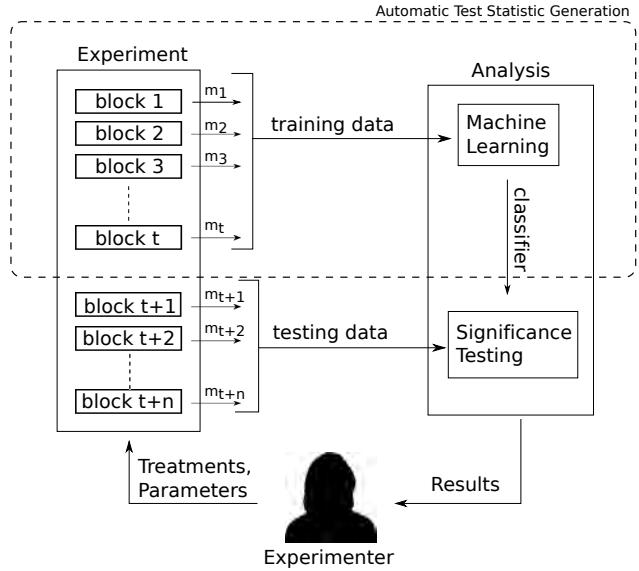


Fig. 2. Our experimental setup with training and testing blocks. Measurements from the training blocks are used to build a classifier. The trained classifier is used to compute the test statistic on the measurements from the testing blocks for significance testing.

compute the exact p-value, we estimate it by randomly sampling the possible reassignments. We can use a confidence interval to characterize the quality of the estimation [12]. The p-values we report are actually the upper bounds of the 99% confidence intervals of the p-values (details in Appendix B).

4.3 Selecting Test Statistics

The above methodology leaves open the question of how to select the test statistic. In some cases, the experimenter might be interested in a particular test statistic. For example, an experimenter testing ad choice could use a test statistic that counts the number of ads related to the removed interest. In other cases, the experimenter might be looking for *any* effect. AdFisher offers the ability to automatically select a test statistic. To do so, it partitions the collected data into training and testing subsets, and uses the training data to train a classifier. Figure 2 shows an overview of AdFisher’s workflow.

To select a classifier, AdFisher uses 10-fold cross validation on the training data to select among several possible parameters. The classifier predicts which treatment an agent received, only from the ads that get served to that agent. If the classifier is able to make this prediction with high accuracy, it suggests a systematic difference between the ads served to the two groups

that the classifier was able to learn. If no difference exists, then we would expect the number to be near the guessing rate of 50%. AdFisher uses the accuracy of this classifier as its test statistic.

To avoid the possibility of seeing a high accuracy due to overfitting, AdFisher evaluates the accuracy of the classifier on a testing data set that is disjoint from the training data set. That is, in the language of statistics, we form our hypothesis about the test statistic being able to distinguish the groups before seeing the data on which we test it to ensure that it has predictive power. AdFisher uses the permutation test to determine whether the degree to which the classifier’s accuracy on the test data surpasses the guessing rate is statistically significant. That is, it calculates the p-value that measures the probability of seeing the observed accuracy given that the classifier is just guessing. If the p-value is below 0.05, we conclude that it is unlikely that classifier is guessing and that it must be making use of some difference between the ads shown to the two groups.

4.4 Avoiding Pitfalls

The above methodology avoids some pitfalls. Most fundamentally, we use a statistical analysis whose assumptions match those of our experimental design. Assumptions required by many statistical analyses appear unjustifiable in our setting. For example, many analyses assume that the agents do not interact or that the ads are independent and identically distributed (e.g., [14, 17]). Given that all agents receive ads from the same pool of possible ads governed by the same advertisers’ budgets, these assumptions appear unlikely to hold. Indeed, empirical evidence suggests that it does not [11]. The permutation test, which does not require this assumption, allows us to ensure statistical soundness of our analysis without making these assumptions [22].

Our use of randomization implies that many factors that could be confounding factors in an unrandomized design become noise in our design (e.g., [12]). While such noise may require us to use a large sample size to find an effect, it does not affect the soundness of our analysis.

Our use of two data sets, one for training the classifier to select the test statistic and one for hypothesis testing ensures that we do not engage in overfitting, data dredging, or multiple hypothesis testing (e.g., [23]). All these problems result from looking for so many possible patterns that one is found by chance. While we look for many patterns in the training data, we only check for one in the testing data.

Relatedly, by reporting a p-value, we provide a quantitative measure of the confidence we have that the observed effect is genuine and not just by chance [24]. Reporting simply the classifier accuracy or that some difference occurred fails to quantify the possibility that the result was a fluke.

4.5 Scope

We restrict the scope of our methodology to making claims that an effect exists with high likelihood as quantified by the p-value. That is, we expect our methodology to only rarely suggest that an effect exists when one does not.

We do not claim “completeness” or “power”: we might fail to detect some use of information. For example, Google might not serve different ads upon detecting that all the browser agents in our experiment are running from the same IP address. Despite this limitation in our experiments, we found interesting instances of usage.

Furthermore, we do not claim that our results generalize to all users. To do so, we would need to take a random sample of all users, their IP addresses, browsers, and behaviors, which is prohibitively expensive. We cannot generalize our results if for example, instead of turning off some usage upon detecting our experiments, Google turns it on. While our experiments would detect this usage, it might not be experienced by normal users. However, it would be odd if Google purposefully performs questionable behaviors only with those attempting to find it.

While we use webpages associated with various interests to simulate users with those interests, we cannot establish that having the interest itself caused the ads to change. It is possible that other features of the visited webpages causes change - a form of confounding called “profile contamination” [14], since the pages cover other topics as well. Nevertheless, we have determined that visiting webpages associated with the interest does result in seeing a change, which should give pause to users visiting webpages associated with sensitive interests.

Lastly, we do not attempt to determine how the information was used. It could have been used by Google directly for targeting or it could have been used by advertisers to place their bids. We cannot assign blame. We hope future work will shed light on these issues, but given that we cannot observe the interactions between Google and advertisers, we are unsure whether it can be done.

5 AdFisher

In this section, we describe AdFisher - a tool implementing our methodology. AdFisher makes it easy to run experiments using the above methodology for a set of treatments, measurements, and classifiers (test statistics) we have implemented. AdFisher is also extensible allowing the experimenter to implement additional treatments, measurements, or test statistics. For example, an experimenter interested in studying a different online platform only needs to add code to perform actions and collect measurements on that platform. They need not modify methods that randomize the treatments, carry out the experiment, or perform the data analysis.

To simulate a new person on the network, AdFisher creates each agent from a fresh browser instance with no browsing history, cookies, or other personalization. AdFisher randomly assigns each agent to a group and applies the appropriate treatment, such as having the browser visit webpages. Next, AdFisher makes measurements of the agent, such as collecting the ads shown to the browser upon visiting another webpage. All of the agents within a block execute and finish the treatments before moving on to collect the measurements to remove time as a factor. AdFisher runs all the agents on the same machine to prevent differences based on location, IP address, operating system, or other machine specific differences between agents.

Next, we detail the particular treatments, measurements, and test statistics that we have implemented in AdFisher. We also discuss how AdFisher aids an experimenter in understanding the results.

Treatments. A treatment specifies what actions are to be performed by a browser agent. AdFisher automatically applies treatments assigned to each agent. Typically, these treatments involve invoking the Selenium WebDriver to make the agent interact with webpages.

AdFisher makes it easy to carry out common treatments by providing ready-made implementations. The simplest stock treatments we provide set interests, gender, and age range in Google’s Ad Settings. Another stock treatment is to visit a list of webpages stored on a file.

To make it easy to see whether websites associated with a particular interest causes a change in behavior, we have provided the ability to create lists of webpages associated with a category on Alexa. For each category, Alexa tracks the top websites sorted according to their traffic rank measure (a combination of the number of

users and page views) [25]. The experimenter can use AdFisher to download the URLs of the top webpages Alexa associates with an interest. By default, it downloads the top 100 URLs. A treatment can then specify that agents visit this list of websites. While these treatments do not correspond directly to having such an interest, it allows us to study how Google responds to people visiting webpages associated with those interests.

Often in our experiments, we compared the effects of a certain treatment applied to the experimental group against the *null treatment* applied to the control group. Under the null treatment, agents do nothing while agents under a different treatment complete their respective treatment phase.

Measurements. AdFisher can currently measure the values set in Google’s Ad Settings page and the ads shown to the agents after the treatments. It comes with stock functionality for collecting and analyzing text ads. Experimenters can add methods for image, video, and flash ads.

To find a reasonable website for ad collection, we looked to news sites since they generally show many ads. Among the top 20 news websites on alexa.com, only five displayed text ads served by Google: theguardian.com/us, timesofindia.indiatimes.com, bbc.com/news, reuters.com/news/us and bloomberg.com. AdFisher comes with the built-in functionality to collect ads from any of these websites. One can also specify for how many reloads ads are to be collected (default 10), or how long to wait between successive reloads (default 5s). For each page reload, AdFisher parses the page to find the ads shown by Google and stores the ads. The experimenter can add parsers to collect ads from other websites.

We run most of our experiments on Times of India as it serves the most (five) text ads per page reload. We repeat some experiments on the Guardian (three ads per reload) to demonstrate that our results are not specific to one site.

Classification. While the experimenter can provide AdFisher with a test statistic to use on the collected data, AdFisher is also capable of automatically selecting a test statistic using machine learning. It splits the entire data set into training and testing subsets, and examines a training subset of the collected measurements to select a classifier that distinguishes between the measurements taken from each group. From the point of view of machine learning, the set of ads collected by

an agent corresponds to an *instance* of the concept the classifier is attempting to learn.

Machine learning algorithms operate over sets of *features*. AdFisher has functions for converting the text ads seen by an agent into three different feature sets. The *URL feature set* consists of the URLs displayed by the ads (or occasionally some other text if the ad displays it where URLs normally go). Under this feature set, the feature vector representing an agent’s data has a value of n in the i th entry iff the agent received n ads that display the i th URL where the order is fixed but arbitrary.

The *URL+Title feature set* looks at both the displayed URL and the title of the ad jointly. It represents an agent’s data as a vector where the i th entry is n iff the agent received n ads containing the i th pair of a URL and title.

The third feature set AdFisher has implemented is the *word feature set*. This set is based on word stems, the main part of the word with suffixes such as “ed” or “ing” removed in a manner similar to the work of Balebako et al. [19]. Each word stem that appeared in an ad is assigned a unique id. The i th entry in the feature vector is the number of times that words with the i th stem appeared in the agent’s ads.

We explored a variety of classification algorithms provided by the scikit-learn library [9]. We found that logistic regression with an L2 penalty over the URL+title feature set consistently performed well compared to the others. At its core, logistic regression predicts a class given a feature vector by multiplying each of the entries of the vector by its own weighting coefficient (e.g., [26]). It then takes the sum of all these products. If the sum is positive, it predicts one class; if negative, it predicts the other.

While using logistic regression, the training stage consists of selecting the coefficients assigned to each feature to predict the training data. Selecting coefficients requires balancing the training-accuracy of the model with avoiding overfitting the data with an overly complex model. We apply 10-fold cross-validation on the training data to select the regularization parameter of the logistic regression classifier. By default, AdFisher splits the data into training and test sets by using the last 10% of the data collected for testing.

Explanations. To explain how the learned classifier distinguished between the groups, we explored several methods. We found the most informative to be the model produced by the classifier itself. Recall that logistic regression weighted the various features of the instances

with coefficients reflecting how predictive they are of each group. Thus, with the URL+title feature set, examining the features with the most extreme coefficients identifies the URL+title pair most used to predict the group to which agents receiving an ad with that URL+title belongs.

We also explored using simple metrics for providing explanations, like ads with the highest frequency in each group. However, some generic ads get served in huge numbers to both groups. We also looked at the proportion of times an ad was served to agents in one group to the total number of times observed by all groups. However, this did not provide much insight since the proportion typically reached its maximum value of 1.0 from ads that only appeared once. Another choice we explored was to compute the difference in the number of times an ad appears between the groups. However, this metric is also highly influenced by how common the ad is across all groups.

6 Experiments

In this section, we discuss experiments that we carried out using AdFisher. In total, we ran 21 experiments, each of which created its own testing data sets using independent random assignments of treatments to agents. We analyze each test data set only once and report the results of each experiment separately. Thus, we do not test multiple hypotheses on any of our test data sets ensuring that the probability of false positives (p-value) are independent with the exception of our analyses for ad choice. In that case, we apply a Bonferroni correction.

Each experiment examines one of the properties of interest from Table 1. We found violations of nondiscrimination and data transparency and cases of compliance with effectful and ad choice. Since these summaries each depend upon more than one experiment, they are the composite of multiple hypotheses. To prevent false positives for these summaries, for each property, we report p-values adjusted by the number of experiments used to explore that property. We use the Holm-Bonferroni method for our adjustments, which is uniformly more powerful than the commonly used Bonferroni correction [27]. This method orders the component hypotheses by their unadjusted p-values applying a different correction to each until reaching a hypothesis whose adjusted value is too large to reject. This hypoth-

esis and all remaining hypotheses are rejected regardless of their p-values. Appendix C provides details.

Table 2 in Appendix A summarizes our findings.

6.1 Nondiscrimination

We use AdFisher to demonstrate a violation in the nondiscrimination property. If AdFisher finds a statistically significant difference in how Google treats two experimental groups, one consisting of members having a protected attribute and one whose members do not, then the experimenter has strong evidence that Google discriminates on that attribute. In particular, we use AdFisher’s ability to automatically select a test statistic to check for possible differences to test the null hypothesis that the two experimental groups have no differences in the ads they receive.

As mentioned before, it is difficult to send a clear signal about any attribute by visiting related webpages since they may have content related to other attributes. The only way to send a clear signal is via Ad Settings. Thus, we focus on attributes that can be set on the Ad Settings page. In a series of experiments, we set the gender of one group to female and the other to male. In one of the experiments, the agents went straight to collecting ads; in the others, they simulated an interest in jobs. In all but one experiment, they collected ads from the Times of India (TOI); in the exception, they collected ads from the Guardian. In one experiment, they also visited the top 10 websites for the U.S. according to alexa.com to fill out their interests.¹ Table 3 in Appendix A summarizes results from these experiments.

AdFisher found a statistically significant difference in the ads for male and female agents that simulated an interest in jobs in May, 2014. It also found evidence of discrimination in the nature of the effect. In particular, it found that females received fewer instances of an ad encouraging the taking of high paying jobs than males. AdFisher did not find any statistically significant differences among the agents that did not visit the job-related pages or those operating in July, 2014. We detail the experiment finding a violation before discussing why we think the other experiments did not result in significant results.

Gender and Jobs. In this experiment, we examine how changing the gender demographic on Google Ad Settings affects the ads served and interests inferred for

agents browsing employment related websites. We set up AdFisher to have the agents in one group visit the Google Ad Settings page and set the gender bit to female while agents in the other group set theirs to male. All the agents then visited the top 100 websites listed under the Employment category of Alexa². The agents then collect ads from Times of India.

AdFisher ran 100 blocks of 10 agents each. (We used blocks of size 10 in all our experiments.) AdFisher used the ads of 900 agents (450 from each group) for training a classifier using the URL+title feature set, and used the remaining 100 agents’ ads for testing. The learned classifier attained a test-accuracy of 93%, suggesting that Google did in fact treat the genders differently. To test whether this response was statistically significant, AdFisher computed a p-value by running the permutation test on a million randomly selected block-respecting permutations of the data. The significance test yielded an adjusted p-value of < 0.00005.

We then examined the model learned by AdFisher to explain the nature of the difference. Table 4 shows the five URL+title pairs that the model identifies as the strongest indicators of being from the female or male group. How ads for identifying the two groups differ is concerning. The two URL+title pairs with the highest coefficients for indicating a male were for a career coaching service for “\$200k+” executive positions. Google showed the ads 1852 times to the male group but just 318 times to the female group. The top two URL+title pairs for the female group was for a generic job posting service and for an auto dealer.

The found discrimination in this experiment was predominately from a pair of job-related ads for the same service making the finding highly sensitive to changes in the serving of these ads. A closer examination of the ads from the same experimental setup ran in July, 2014, showed that the frequency of these ads reduced from 2170 to just 48, with one of the ads completely disappearing. These 48 ads were only shown to males, continuing the pattern of discrimination. This pattern was recognized by the machine learning algorithm, which selected the ad as the second most useful for identifying males. However, they were too infrequent to establish statistical significance. A longer running experiment with more blocks might have succeeded.

¹ <http://www.alexa.com/topsites/countries/US>

² <http://www.alexa.com/topsites/category/Top/Business/Employment>

6.2 Transparency

AdFisher can demonstrate violations of individual data use transparency. AdFisher tests the null hypothesis that two groups of agents with the same ad settings receives ads from the same distribution despite being subjected to different experimental treatments. Rejecting the null hypothesis implies that some difference exists in the ads that is not documented by the ad settings.

In particular, we ran a series of experiments to examine how much transparency Google’s Ad Settings provided. We checked whether visiting webpages associated with some interest could cause a change in the ads shown that is not reflected in the settings.

We ran such experiments for five interests: substance abuse, disabilities, infertility³, mental disorders⁴, and adult websites⁵. Results from statistical analysis of these experiments are shown in Table 5 of Appendix A.

We examined the interests found in the settings for the two cases where we found a statistically significant difference in ads, substance abuse and disability. We found that settings did not change at all for substance abuse and changed in an unexpected manner for disabilities. Thus, we detail these two experiments below.

Substance Abuse. We were interested in whether Google’s outputs would change in response to visiting webpages associated with substance abuse, a highly sensitive topic. Thus, we ran an experiment in which the experimental group visited such websites while the control group idled. Then, we collected the Ad Settings and the Google ads shown to the agents at the Times of India. For the webpages associated with substance abuse, we used the top 100 websites on the Alexa list for substance abuse⁶.

AdFisher ran 100 blocks of 10 agents each. At the end of visiting the webpages associated with substance abuse, none of the 500 agents in the experimental group had interests listed on their Ad Settings pages. (None of the agents in the control group did either since the settings start out empty.) If one expects the Ad Settings page to reflect all learned inferences, then he would not anticipate ads relevant to those website visits given the lack of interests listed.



Fig. 3. Screenshot of an ad with the top URL+title for identifying agents that visited webpages associated with substance abuse

However, the ads collected from the Times of India told a different story. The learned classifier attained a test-accuracy of 81%, suggesting that Google did in fact respond to the page visits. Indeed, using the permutation test, AdFisher found an adjusted p-value of < 0.00005 . Thus, we conclude that the differences are statistically significant: Google’s ads changed in response to visiting the webpages associated with substance abuse. Despite this change being significant, the Ad Settings pages provided no hint of its existence: the transparency tool is opaque!

We looked at the URL+title pairs with the highest coefficients for identifying the experimental group that visited the websites related to substance abuse. Table 6 provides information on coefficients and URL+titles learned. The three highest were for “Watershed Rehab”. The top two had URLs for this drug and alcohol rehab center. The third lacked a URL and had other text in its place. Figure 3 shows one of Watershed’s ads. The experimental group saw these ads a total of 3309 times (16% of the ads); the control group never saw any of them nor contained any ads with the word “rehab” or “rehabilitation”. None of the top five URL+title pairs for identifying the control group had any discernible relationship with rehab or substance abuse.

These results remain robust across variations on this design with statistical significance in three variations. For example, two of these ads remain the top two ads for identifying the agents that visited the substance abuse websites in July using ads collected from the Guardian.

One possible reason why Google served Watershed’s ads could be *remarketing*, a marketing strategy that encourages users to return to previously visited websites [28]. The website thewatershed.com features among the top 100 websites about substance-abuse on Alexa, and agents visiting that site may be served Watershed’s ads as part of remarketing. However, these users cannot see any changes on Google Ad Settings despite Google having learnt some characteristic (visited thewatershed.com) about them and serving ads relevant to that characteristic.

Disabilities. This experiment was nearly identical in setup but used websites related to disabilities instead of

³ http://www.alexa.com/topsites/category/Top/Health/Reproductive_Health/Infertility

⁴ http://www.alexa.com/topsites/category/Top/Health/Mental_Health/Disorders

⁵ <http://www.alexa.com/topsites/category/Top/Adult>

⁶ http://www.alexa.com/topsites/category/Top/Health/Addictions/Substance_Abuse

substance abuse. We used the top 100 websites on Alexa on the topic.⁷

For this experiment, AdFisher found a classifier with a test-accuracy of 75%. It found a statistically significant difference with an adjusted p-value of less than 0.00005.

Looking at the top ads for identifying agents that visited the webpages associated with disabilities, we see that the top two ads have the URL www.abilitiesexpo.com and the titles “Mobility Lifter” and “Standing Wheelchairs”. They were shown a total of 1076 times to the experimental group but never to the control group. (See Table 7.)

This time, Google did change the settings in response to the agents visiting the websites. None of them are directly related to disabilities suggesting that Google might have focused on other aspects of the visited pages. Once again, we believe that the top ads were served due to remarketing, as abilitiesexpo.com was among the top 100 websites related to disabilities.

6.3 Effectful Choice

We tested whether making changes to Ad Settings has an effect on the ads seen, thereby giving the users a degree of choice over the ads. In particular, AdFisher tests the null hypothesis that changing some ad setting has no effect on the ads.

First, we tested whether opting out of tracking actually had an effect by comparing the ads shown to agents that opted out after visiting car-related websites to ads from those that did not opt out. We found a statistically significant difference.

We also tested whether removing interests from the settings page actually had an effect. We set AdFisher to have both groups of agents simulate some interest. AdFisher then had the agents in one of the groups remove interests from Google’s Ad Settings related to the induced interest. We found statistically significant differences between the ads both groups collected from the Times of India for two induced interests: online dating and weight loss. We describe one in detail below.

Online Dating. We simulated an interest in online dating by visiting the website www.midsummerseve.com/, a website we choose since it sets Google’s ad setting for “Dating & Personals” (this site no longer affects

the setting). AdFisher then had just the agents in the experimental group remove the interest “Dating & Personals” (the only one containing the keyword “dating”). All the agents then collected ads from the Times of India.

AdFisher found statistically significant differences between the groups with a classifier accuracy of 74% and an adjusted p-value of < 0.00003 . Furthermore, the effect appears related to the interests removed. The top ad for identifying agents that kept the romantic interests has the title “Are You Single?” and the second ad’s title is “Why can’t I find a date?”. None of the top five for the control group that removed the interests were related to dating (Table 9). Thus, the ad settings appear to actually give users the ability to avoid ads they might dislike or find embarrassing. In the next set of experiments, we explicitly test for this ability.

We repeated this experiment in July, 2014, using the websites relationshipsurgery.com and datemypet.com, which also had an effect on Ad Settings, but did not find statistically significant differences.

6.4 Ad Choice

Whereas the other experiments tested merely for the presence of an effect, testing for ad choice requires determining whether the effect is an increase or decrease in the number of relevant ads seen. Fortunately, since AdFisher uses a one-sided permutation test, it tests for either an increase or a decrease, but not for both simultaneously, making it usable for this purpose. In particular, after removing an interest, we check for a decrease to test for compliance using the null hypothesis that either no change or an increase occurred, since rejecting this hypothesis would imply that a decrease in the number of related ads occurred. To check for a violation, we test for the null hypothesis that either no change or a decrease occurred. Due to testing two hypotheses, we use an adjustment to the p-value cutoff considered significant to avoid finding significant results simply from testing multiple hypotheses. In particular, we use the standard Bonferroni correction, which calls for multiplying the p-value by 2 (e.g., [29]).

We ran three experiments checking for ad choice. The experiments followed the same setup as the effectful choice ones, but this time we used all the blocks for testing a given test statistic. The test statistic counted the number of ads containing keywords. In the first, we again test online dating using relationshipsurgery.com and datemypet.com. In particular, we found that re-

⁷ <http://www.alexameter.com/topsites/category/Top/Society/Disabled>

moving online dating resulted in a significant decrease (p-value adjusted for all six experiments: 0.0456) in the number of ads containing related keywords (from 109 to 34). We detail the inconclusive results for weight loss below.

Weight Loss. We induced an interest in weight loss by visiting dietingsucks.blogspot.com. Afterwards, the agents in the experimental group removed the interests “Fitness” and “Fitness Equipment and Accessories”, the only ones related to weight loss. We then used a test statistic that counted the number of ads containing the keyword “fitness”. Interestingly, the test statistic was higher on the group with the interests removed, although not to a statistically significant degree. We repeated the process with a longer keyword list and found that removing interests decreased test statistic this time, but also not to a statistically significant degree.

7 Discussion and Conclusion

Using AdFisher, we conducted 21 experiments using 17,370 agents that collected over 600,000 ads. Our experiments found instances of discrimination, opacity, and choice in targeted ads of Google. Discrimination, is at some level, inherent to profiling: the point of profiling is to treat some people differently. While customization can be helpful, we highlight a case where the customization appears inappropriate taking on the negative connotations of discrimination. In particular, we found that males were shown ads encouraging the seeking of coaching services for high paying jobs more than females (§6.1).

We do not, however, claim that any laws or policies were broken. Indeed, Google’s policies allow it to serve different ads based on gender. Furthermore, we cannot determine whether Google, the advertiser, or complex interactions among them and others caused the discrimination (§4.5). Even if we could, the discrimination might have resulted unintentionally from algorithms optimizing click-through rates or other metrics free of bigotry. Given the pervasive structural nature of gender discrimination in society at large, blaming one party may ignore context and correlations that make avoiding such discrimination difficult. More generally, we believe that no scientific study can demonstrate discrimination in the sense of *unjust discrimination* since science cannot demonstrate normative statements (e.g., [30])

Nevertheless, we are comfortable describing the results as “discrimination”. From a strictly scientific view point, we have shown discrimination in the non-normative sense of the word. Personally, we also believe the results show discrimination in the normative sense of the word. Male candidates getting more encouragement to seek coaching services for high-paying jobs could further the current gender pay gap (e.g., [31]). Thus, we do not see the found discrimination in our vision of a just society even if we are incapable of blaming any particular parties for this outcome.

Furthermore, we know of no justification for such customization of the ads in question. Indeed, our concern about this outcome does not depend upon how the ads were selected. Even if this decision was made solely for economic reasons, it would continue to be discrimination [32]. In particular, we would remain concerned if the cause of the discrimination was an algorithm ran by Google and/or the advertiser automatically determining that males are more likely than females to click on the ads in question. The amoral status of an algorithm does not negate its effects on society.

However, we also recognize the possibility that no party is at fault and such unjust effects may be inadvertent and difficult to prevent. We encourage research developing tools that ad networks and advertisers can use to prevent such unacceptable outcomes (e.g., [33]).

Opacity occurs when a tool for providing transparency into how ads are selected and the profile kept on a person actually fails to provide such transparency. Our experiment on substance abuse showed an extreme case in which the tool failed to show any profiling but the ad distributions were significantly different in response to behavior (§6.2). In particular, our experiment achieved an adjusted p-value of < 0.00005 , which is 1000 times more significant than the standard 0.05 cutoff for statistical significance. This experiment remained robust to variations showing a pattern of such opacity.

Ideally, tools, such as Ad Settings, would provide a complete representation of the profile kept on a person, or at least the portion of the profile that is used to select ads shown to the person. Two people with identical profiles might continue to receive different ads due to other factors affecting the choice of ads such as A/B testing or the time of day. However, systematic differences between ads shown at the same time and in the same context, such as those we found, would not exist for such pairs of people.

In our experiments testing transparency, we suspect that Google served the top ads as part of remarketing, but our blackbox experiments do not determine whether

this is the case. While such remarketing may appear less concerning than Google inferring a substance abuse issue about a person, its highly targeted nature is worrisome particularly in settings with shared computers or shoulder surfing. There is a need for a more inclusive transparency/control mechanism which encompasses re-marketed ads as well. Additionally, Google states that “we prohibit advertisers from remarketing based on sensitive information, such as health information” [28]. Although Google does not specify what they consider to be “health information”, we view the ads as in violation of Google’s policy, thereby raising the question of how Google should enforce its policies.

Lastly, we found that Google Ad Settings does provide the user with a degree of choice about the ads shown. In this aspect, the transparency/control tool operated as we expected.

Our tool, AdFisher, makes it easy to run additional experiments exploring the relations between Google’s ads and settings. It can be extended to study other systems. Its design ensures that it can run and analyze large scale experiments to find subtle differences. It automatically finds differences between large data sets produced by different groups of agents and explains the nature of those differences. By completely automating the data analysis, we ensure that an appropriate statistical analysis determines whether these differences are statistically significant and sound conclusions.

AdFisher may have cost advertisers a small sum of money. AdFisher never clicked on any ads to avoid per click fees, which can run over \$4 [34]. Its experiments may have caused per-impression fees, which run about \$0.00069 [35]. In the billion dollar ad industry, its total effect was about \$400.

8 Future Work

We would like to extend AdFisher to study information flow on other advertising systems like Facebook, Bing, or Gmail. We would also like to analyze other kinds of ads like image or flash ads. We also plan to use the tool to detect price discrimination on sites like Amazon or Kayak, or find differences in suggested posts on blogs and news websites, based on past user behavior. We have already mentioned the interesting problem of how ad networks can ensure that their policies are respected by advertisers (§7).

We also like to assign blame where it is due. However, doing so is often difficult. For example, our view on

blame varies based on why females were discriminated against in our gender and jobs experiment. If Google allowed the advertiser to easily discriminate, we would blame both. If the advertiser circumvented Google’s efforts to prevent such discrimination by targeting correlates of gender, we would blame just the advertiser. If Google decided to target just males with the ad on its own, we would blame just Google. While we lack the access needed to make this determination, both Google and the advertiser have enough information to audit the other with our tool.

As another example, consider the results of opacity after visiting substance abuse websites. While we suspect, remarketing is the cause, it is also possible that Google is targeting users without the rehab center’s knowledge. In this case, it would remain unclear as to whether Google is targeting users as substance abusers or due to some other content correlated with the webpages we visited to simulate an interest in substance abuse. We would like to find ways of controlling for these confounding factors.

For these reasons, we cannot claim that Google has violated its policies. In fact, we consider it more likely that Google has lost control over its massive, automated advertising system. Even without advertisers placing inappropriate bids, large-scale machine learning can behave in unexpected ways. With this in mind, we hope future research will examine how to produce machine learning algorithms that automatically avoid discriminating against users in unacceptable ways and automatically provide transparency to users.

Acknowledgements. We thank Jeannette M. Wing for helpful discussions about this work. We thank Augustin Chaintreau, Roxana Geambasu, Qiang Ma, Latanya Sweeney, and Craig E. Wills for providing additional information about their works. We thank the reviewers of this paper for their helpful comments. This research was supported by the National Science Foundation (NSF) grants CCF0424422 and CNS1064688. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government or any other entity.

References

- [1] J. R. Mayer and J. C. Mitchell, “Third-party web tracking: Policy and technology,” in *IEEE Symposium on Security and*

- Privacy*, 2012, pp. 413–427.
- [2] B. Ur, P. G. Leon, L. F. Cranor, R. Shay, and Y. Wang, "Smart, useful, scary, creepy: Perceptions of online behavioral advertising," in *Proceedings of the Eighth Symposium on Usable Privacy and Security*. ACM, 2012, pp. 4:1–4:15.
 - [3] Google, "About ads settings," <https://support.google.com/ads/answer/2662856>, accessed Nov. 21, 2014.
 - [4] Yahoo!, "Ad interest manager," https://info.yahoo.com/privacy/us/yahoo/opt_out/targeting/details.html, accessed Nov. 21, 2014.
 - [5] Microsoft, "Microsoft personalized ad preferences," <http://choice.microsoft.com/en-us/opt-out>, accessed Nov. 21, 2014.
 - [6] Executive Office of the President, "Big data: Seizing opportunities, preserving values," Posted at http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf, 2014, accessed Jan. 26, 2014.
 - [7] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, S. Dasgupta and D. Mcallester, Eds., vol. 28. JMLR Workshop and Conference Proceedings, May 2013, pp. 325–333. [Online]. Available: <http://jmlr.org/proceedings/papers/v28/zemel13.pdf>
 - [8] Google, "Privacy policy," <https://www.google.com/intl/en/policies/privacy/>, accessed Nov. 21, 2014.
 - [9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
 - [10] E. Jones, T. Oliphant, P. Peterson et al., "SciPy: Open source scientific tools for Python," 2001, <http://www.scipy.org/>.
 - [11] M. C. Tschantz, A. Datta, A. Datta, and J. M. Wing, "A methodology for information flow experiments," ArXiv, Tech. Rep. arXiv:1405.2376v1, 2014.
 - [12] P. Good, *Permutation, Parametric and Bootstrap Tests of Hypotheses*. Springer, 2005.
 - [13] C. E. Wills and C. Tatar, "Understanding what they do with what they know," in *Proceedings of the 2012 ACM Workshop on Privacy in the Electronic Society*, 2012, pp. 13–18.
 - [14] P. Barford, I. Canadi, D. Krushevskaja, Q. Ma, and S. Muthukrishnan, "Adscape: Harvesting and analyzing online display ads," in *Proceedings of the 23rd International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2014, pp. 597–608.
 - [15] B. Liu, A. Sheth, U. Weinsberg, J. Chandrashekhar, and R. Govindan, "AdReveal: Improving transparency into online targeted advertising," in *Proceedings of the Twelfth ACM Workshop on Hot Topics in Networks*. ACM, 2013, pp. 12:1–12:7.
 - [16] M. Lécuyer, G. Ducoffe, F. Lan, A. Papancea, T. Petsios, R. Spahn, A. Chaintreau, and R. Geambasu, "XRay: Increasing the web's transparency with differential correlation," in *Proceedings of the USENIX Security Symposium*, 2014.
 - [17] S. Englehardt, C. Eubank, P. Zimmerman, D. Reisman, and A. Narayanan, "Web privacy measurement: Scientific principles, engineering platform, and new re-
 - sults," Manuscript posted at <http://randomwalker.info/publications/WebPrivacyMeasurement.pdf>, 2014, accessed Nov. 22, 2014.
 - [18] S. Guha, B. Cheng, and P. Francis, "Challenges in measuring online advertising systems," in *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, 2010, pp. 81–87.
 - [19] R. Balebako, P. Leon, R. Shay, B. Ur, Y. Wang, and L. Cranor, "Measuring the effectiveness of privacy tools for limiting behavioral advertising," in *Web 2.0 Security and Privacy Workshop*, 2012.
 - [20] L. Sweeney, "Discrimination in online ad delivery," *Commun. ACM*, vol. 56, no. 5, pp. 44–54, 2013.
 - [21] R. A. Fisher, *The Design of Experiments*. Oliver & Boyd, 1935.
 - [22] S. Greenland and J. M. Robins, "Identifiability, exchangeability, and epidemiological confounding," *International Journal of Epidemiology*, vol. 15, no. 3, pp. 413–419, 1986.
 - [23] T. M. Mitchell, *Machine Learning*. McGraw-Hill, 1997.
 - [24] D. D. Jensen, "Induction with randomization testing: Decision-oriented analysis of large data sets," Ph.D. dissertation, Sever Institute of Washington University, 1992.
 - [25] Alexa, "Is popularity in the top sites by category directory based on traffic rank?" <https://support.alexa.com/hc/en-us/articles/200461970>, accessed Nov. 21, 2014.
 - [26] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
 - [27] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979.
 - [28] Google, "Google privacy and terms," <http://www.google.com/policies/technologies/ads/>, accessed Nov. 22, 2014.
 - [29] H. Abdi, "Bonferroni and Šidák corrections for multiple comparisons," in *Encyclopedia of Measurement and Statistics*, N. J. Salkind, Ed. Sage, 2007.
 - [30] D. Hume, *A Treatise of Human Nature: Being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects*, 1738, book III, part I, section I.
 - [31] Pew Research Center's Social and Demographic Trends Project, "On pay gap, millennial women near parity — for now: Despite gains, many see roadblocks ahead," 2013.
 - [32] T. Z. Zarsky, "Understanding discrimination in the scored society," *Washington Law Review*, vol. 89, pp. 1375–1412, 2014.
 - [33] R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *Proceedings of the 30th International Conference on Machine Learning*, ser. JMLR: W&CP, vol. 28. JMLR.org, 2013, pp. 325–333.
 - [34] Adgooroo, "Adwords cost per click rises 26% between 2012 and 2014," <http://www.adgooroo.com/resources/blog/adwords-cost-per-click-rises-26-between-2012-and-2014/>, accessed Nov. 21, 2014.
 - [35] L. Olejnik, T. Minh-Dung, and C. Castelluccia, "Selling off privacy at auction," in *Network and Distributed System Security Symposium (NDSS)*. The Internet Society, 2013.
 - [36] C. J. Clopper and E. S. Pearson, "The use of confidence or fiducial limits illustrated in the case of the binomial," *Biometrika*, vol. 26, no. 4, pp. 404–413, 1934.

A Tables

Table 2 summarizes the results. Table 3 covers the discrimination experiments with Table 4 showing the top ads for experiment on gender and jobs. Table 5 covers the opacity experiments with Table 6 showing the top ads for the substance-abuse experiment and Table 7 showing them for the disability experiment. Table 8 show the experiments for effectful choice with Table 9 showing the tops ads for online dating. Tables 10 and 11 cover ad choice.

B Details of Methodology

Let the units be arranged in a vector \vec{u} of length n . Let \vec{t} be a *treatment vector*, a vector of length n whose entries are the treatments that the experimenter wants to apply to the units. In the case of just two treatments, \vec{t} can be half full of the first treatment and half full of the second. Let a be an *assignment* of units to treatments, a bijection that maps each entry of \vec{u} to an entry in \vec{t} . That is, an assignment is a permutation on the set of indices of \vec{u} and \vec{t} .

The result of the experiment is a vector of observations \vec{y} where the i th entry of \vec{y} is the response measured for the unit assigned to the i th treatment in \vec{t} by the assignment used. In a randomized experiment, such as those AdFisher runs, the actual assignment used is selected at random uniformly over some set of possible assignments \mathcal{A} .

Let s be a test statistic of the observations of the units. That is $s : \mathcal{Y}^n \rightarrow \mathcal{R}$ where \mathcal{Y} is the set of possible observations made over units, n is the number of units, and \mathcal{R} is the range of s . We require \mathcal{R} to be ordered numbers such as the natural or real numbers. We allow s to treat its arguments differently, that is, the order in which the observations are passed to s matters.

If the null hypothesis is true, then we would expect the value of s to be the same under every permutation of the arguments since the assignment of units to treatments should not matter under the null hypothesis. This reasoning motivates the permutation test. The value produced by a (one-tailed signed) permutation test given observed responses \vec{y} and a test statistic s is

$$\frac{|\{a \in \mathcal{A} \mid s(\vec{y}) \leq s(a(\vec{y}))\}|}{|\mathcal{A}|} = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} I[s(\vec{y}) \leq s(a(\vec{y}))] \quad (1)$$

where the assignments in \mathcal{A} only swaps nearly identical units and $I[\cdot]$ returns 1 if its argument is true and 0 otherwise.

Blocking. For the blocking design, the set of units \mathcal{U} is partitioned into k blocks \mathcal{B}_1 to \mathcal{B}_k . In our case, all the blocks have the same size. Let $|\mathcal{B}_i| = m$ for all i . The set of assignments \mathcal{A} is equal to the set of functions from \mathcal{U} to \mathcal{U} that are permutations not mixing up blocks. That is, a such that for all i and all u in \mathcal{B}_i , $a(u) \in \mathcal{B}_i$. Thus, we may treat \mathcal{A} as k permutations, one for each \mathcal{B}_i . Thus, \mathcal{A} is isomorphic to $\times_{i=1}^k \Pi(\mathcal{B}_i)$ where $\Pi(\mathcal{B}_i)$ is the set of all permutations over \mathcal{B}_i . Thus, $|\times_{i=1}^k \Pi(\mathcal{B}_i)| = (m!)^k$. Thus, (1) can be computed as

$$\frac{1}{(m!)^k} \sum_{a \in \times_{i=1}^k \Pi(\mathcal{B}_i)} I[s(\vec{y}) \leq s(a(\vec{y}))] \quad (2)$$

Sampling. Computing (2) can be difficult when the set of considered arrangements is large. One solution is to randomly sample from the assignments \mathcal{A} . Let \mathcal{A}' be a random subset of \mathcal{A} . We then use the approximation

$$\frac{1}{|\mathcal{A}'|} \sum_{a \in \mathcal{A}'} I[s(\vec{y}) \leq s(a(\vec{y}))] \quad (3)$$

Confidence Intervals. Let \hat{P} be this approximation and p be the true value of (2). p can be understood as the frequency of arrangements that yield large values of the test statistic where *largeness* is determined to be at least as large as the observed value $s(\vec{y})$. That is, the probability that a randomly selected arrangement will yield a large value is p . \hat{P} is the frequency of seeing large values in the $|\mathcal{A}'|$ sampled arrangements. Since the arrangements in the sample were drawn uniformly at random from \mathcal{A} and each draw has probability p of being large, the number of large values will obey the binomial distribution. Let us denote this value as L . and $|\mathcal{A}'|$ as n . Since $\hat{P} = L/n$, $\hat{P} * n$ also obeys the binomial distribution. Thus,

$$\Pr[\hat{P} = \hat{p} \mid n, p] = \binom{n}{\hat{p}n} p^{\hat{p}n} (1-p)^{(1-\hat{p})n} \quad (4)$$

Thus, we may use a binomial proportion confidence interval. We use the Clopper-Pearson interval [36].

Test Statistic. The statistic we use is based on a classifier c . Let $c(y_i) = 1$ mean that c classifies the i th observation as having come from the experimental group and $c(y_i) = 0$ as from the control group. Let $\neg(0) = 1$ and $\neg(1) = 0$. Let \vec{y} be ordered so that all of the exper-

Property	Treatment	Other Actions	Source	When	Length (hrs)	# ads	Result
Nondiscrimination	Gender	-	TOI	May	10	40,400	Inconclusive
	Gender	Jobs	TOI	May	45	43,393	Violation
	Gender	Jobs	TOI	July	39	35,032	Inconclusive
	Gender	Jobs	Guardian	July	53	22,596	Inconclusive
	Gender	Jobs & Top 10	TOI	July	58	28,738	Inconclusive
Data use transparency	Substance abuse	-	TOI	May	37	42,624	Violation
	Substance abuse	-	TOI	July	41	34,408	Violation
	Substance abuse	-	Guardian	July	51	19,848	Violation
	Substance abuse	Top 10	TOI	July	54	32,541	Violation
	Disability	-	TOI	May	44	43,136	Violation
	Mental disorder	-	TOI	May	35	44,560	Inconclusive
	Infertility	-	TOI	May	42	44,982	Inconclusive
	Adult websites	-	TOI	May	57	35,430	Inconclusive
Effectful choice	Opting out	-	TOI	May	9	18,085	Compliance
	Dating interest	-	TOI	May	12	35,737	Compliance
	Dating interest	-	TOI	July	17	22,913	Inconclusive
	Weight loss interest	-	TOI	May	15	31,275	Compliance
	Weight loss interest	-	TOI	July	15	27,238	Inconclusive
Ad choice	Dating interest	-	TOI	July	1	1,946	Compliance
	Weight loss interest	-	TOI	July	1	2,862	Inconclusive
	Weight loss interest	-	TOI	July	1	3,281	Inconclusive

Table 2. Summary of our experimental results. Ads are collected from the Times of India (TOI) or the Guardian. We report how long each experiment took, how many ads were collected for it, and what result we concluded.

Treatment	Other visits	Measurement	Blocks	# ads (# unique ads)		Accuracy	Unadj. p-value	Adj. p-value
				female	male			
Gender	Jobs	TOI, May	100	21,766 (545)	21,627 (533)	93%	0.0000053	0.0000265*
Gender	Jobs	Guardian, July	100	11,366 (410)	11,230 (408)	57%	0.12	0.48
Gender	Jobs & Top 10	TOI, July	100	14,507 (461)	14,231 (518)	56%	0.14	n/a
Gender	Jobs	TOI, July	100	17,019 (673)	18,013 (690)	55%	0.20	n/a
Gender	-	TOI, May	100	20,137 (603)	20,263 (630)	48%	0.77	n/a

Table 3. Results from the discrimination experiments sorted by unadjusted p-value. TOI stands for Times of India. * denotes statistically significant results under the Holm-Bonferroni method.

Title	URL	Coefficient	appears in agents		total appearances	
			female	male	female	male
Top ads for identifying the simulated female group						
Jobs (Hiring Now)	www.jobsinyourarea.co	0.34	6	3	45	8
4Runner Parts Service	www.westernpatoyotaservice.com	0.281	6	2	36	5
Criminal Justice Program	www3.mc3.edu/Criminal+Justice	0.247	5	1	29	1
Goodwill - Hiring	goodwill.careerboutique.com	0.22	45	15	121	39
UMUC Cyber Training	www.umuc.edu/cybersecuritytraining	0.199	19	17	38	30
Top ads for identifying agents in the simulated male group						
\$200k+ Jobs - Execs Only	careerchange.com	-0.704	60	402	311	1816
Find Next \$200k+ Job	careerchange.com	-0.262	2	11	7	36
Become a Youth Counselor	www.youthcounseling.degreeleap.com	-0.253	0	45	0	310
CDL-A OTR Trucking Jobs	www.tadiviers.com/OTRJobs	-0.149	0	1	0	8
Free Resume Templates	resume-templates.resume-now.com	-0.149	3	1	8	10

Table 4. Top URL+titles for the gender and jobs experiment on the Times of India in May.

Treatment	Other visits	Measurement	# ads (# unique ads)		Accuracy	Unadj. p-value	Adj. p-value
			experimental	control			
Substance abuse	-	TOI, May	20,420 (427)	22,204 (530)	81%	0.0000053	0.0000424*
Substance abuse	-	TOI, July	16,206 (653)	18,202 (814)	98%	0.0000053	0.0000371*
Substance abuse	Top 10	TOI, July	15,713 (603)	16,828 (679)	65%	0.0000053	0.0000318*
Disability	-	TOI, May	19,787 (546)	23,349 (684)	75%	0.0000053	0.0000265*
Substance abuse	-	Guardian, July	8,359 (242)	11,489 (319)	62%	0.0075	0.03*
Mental disorder	-	TOI, May	22,303 (407)	22,257 (465)	59%	0.053	0.159
Infertility	-	TOI, May	22,438 (605)	22,544 (625)	57%	0.11	n/a
Adult websites	-	TOI, May	17,670 (602)	17,760 (580)	52%	0.42	n/a

Table 5. Results from transparency experiments. TOI stands for Times of India. Every experiment for this property ran with 100 blocks. * denotes statistically significant results under the Holm-Bonferroni method.

Title	URL	Coefficient	appears in agents		total appearances	
			control	experi.	control	experi.
Top ads for identifying agents in the experimental group (visited websites associated with substance abuse)						
The Watershed Rehab	www.thewatershed.com/Help	-0.888	0	280	0	2276
Watershed Rehab	www.thewatershed.com/Rehab	-0.670	0	51	0	362
The Watershed Rehab	Ads by Google	-0.463	0	258	0	771
Veteran Home Loans	www.vamortgagecenter.com	-0.414	13	15	22	33
CAD Paper Rolls	paper-roll.net/Cad-Paper	-0.405	0	4	0	21
Top ads for identifying agents in control group						
Alluria Alert	www.bestbeautybrand.com	0.489	2	0	9	0
Best Dividend Stocks	dividends.wyattresearch.com	0.431	20	10	54	24
10 Stocks to Hold Forever	www.streetauthority.com	0.428	51	44	118	76
Delivery Drivers Wanted	get.lyft.com/drive	0.362	22	6	54	14
VA Home Loans Start Here	www.vamortgagecenter.com	0.354	23	6	41	9

Table 6. Top URL+titles for substance abuse experiment on the Times of India in May.

Title	URL	Coefficient	appears in agents		total appearances	
			control	experi.	control	experi.
Top ads for identifying agents in the experimental group (visited websites associated with disability)						
Mobility Lifter	www.abilitiesexpo.com	-1.543	0	84	0	568
Standing Wheelchairs	www.abilitiesexpo.com	-1.425	0	88	0	508
Smoking MN Healthcare	www.stillaproblem.com	-1.415	0	24	0	60
Bike Prices	www.bikesdirect.com	-1.299	0	24	0	79
\$19 Car Insurance - New	auto-insurance.quotelab.com/MN	-1.276	0	6	0	9
Top ads for identifying agents in control group						
Beautiful Women in Kiev	anastasiadate.com	1.304	190	46	533	116
Melucci DDS	AdsbyGoogle	1.255	4	2	10	6
17.2% 2013 Annuity Return	advisorworld.com/CompareAnnuities	1.189	30	5	46	6
3 Exercises To Never Do	homeworkoutrevolution.net	1.16	1	1	3	1
Find CNA Schools Near You	cna-degrees.courseadvisor.com	1.05	22	0	49	0

Table 7. Top URL+titles for disability experiment on the Times of India in May.

Experiment	blocks	# ads (# unique ads)			accuracy	Unadj. p-value	Adj. p-value
		removed/opt-out	keep/opt-in	total			
Opting out	54	9,029 (139)	9,056 (293)	18,085 (366)	83%	0.0000053	0.0000265*
Dating (May)	100	17,975 (518)	17,762 (457)	35,737 (669)	74%	0.0000053	0.0000212*
Weight Loss (May)	83	15,826 (367)	15,449 (427)	31,275 (548)	60%	0.041	0.123
Dating (July)	90	11,657 (727)	11,256 (706)	22,913 (1,014)	59%	0.070	n/a
Weight Loss (July)	100	14,168 (917)	13,070 (919)	27,238 (1,323)	52%	0.41	n/a

Table 8. Results from effectful choice experiments using the Times of India sorted by unadjusted p-value. * denotes statistically significant results under the Holm-Bonferroni method.

Title	URL	Coefficient	appears in agents		total appearances	
			kept	removed	kept	removed
Top ads for identifying the group that kept dating interests						
Are You Single?	www.zoosk.com/Dating	1.583	367	33	2433	78
Top 5 Online Dating Sites	www.consumer-rankings.com/Dating	1.109	116	10	408	13
Why can't I find a date?	www.gk2gk.com	0.935	18	3	51	5
Latest Breaking News	www.onlineinsider.com	0.624	2	1	6	1
Gorgeous Russian Ladies	anastasiadate.com	0.620	11	0	21	0
Top ads for identifying agents in the group that removed dating interests						
Car Loans w/ Bad Credit	www.car.com/Bad-Credit-Car-Loan	-1.113	5	13	8	37
Individual Health Plans	www.individualhealthquotes.com	-0.831	7	9	21	46
Crazy New Obama Tax	www.endofamerica.com	-0.722	19	31	22	51
Atrial Fibrillation Guide	www.johnshopkinshealthalerts.com	-0.641	0	6	0	25
Free \$5 - \$25 Gift Cards	swagbucks.com	-0.614	4	11	5	32

Table 9. Top URL+titles for the dating experiment on Times of India in May.

Experiment	Keywords	# ads (# unique ads)		appearances	
		removed	kept	removed	kept
Dating	dating, romance, relationship	952 (117)	994 (123)	34	109
Weight Loss (1)	fitness	1,461 (259)	1,401 (240)	21	16
Weight Loss (2)	fitness, health, fat, diet, exercise	1,803 (199)	1,478 (192)	2	15

Table 10. Setup for and ads from ad choice experiments. All experiments used 10 blocks. The same keywords are used to remove ad interests, as well as create the test statistic for permutation test.

Experiment	Unadjusted p-value	Bonferroni p-value	Holm-Bonferroni p-value	Unadjusted flipped p-value	Bonferroni flipped p-value	Holm-Bonferroni flipped p-value
Dating	0.0076	0.0152	0.0456*	0.9970	1.994	n/a
Weight Loss (2)	0.18	0.36	0.9	0.9371	1.8742	n/a
Weight Loss (1)	0.72	1.44	n/a	0.3818	0.7636	n/a

Table 11. P-values from ad choice experiments sorted by the (unflipped) p-value. The Bonferroni adjusted p-value is only adjusted for the two hypotheses tested within a single experiment (row). The Holm-Bonferroni adjusts for all 6 hypotheses. * denotes statistically significant results under the Holm-Bonferroni method.

imental group comes first. The statistic we use is

$$s(\vec{y}) = \sum_{i=1}^{n/2} c(y_i) + \sum_{i=n/2+1}^n \neg c(y_i)$$

This is the number correctly classified.

the adjusted p-value depends not just upon its unadjusted value but also upon its position in the list. For the remaining hypotheses, we provide no adjusted p-value since their p-values are irrelevant to the correction beyond how they order the list of hypotheses.

C Holm-Bonferroni Correction

The Holm-Bonferroni Correction starts by ordering the hypotheses in a family from the hypothesis with the smallest (most significant) p-value p_1 to the hypothesis with the largest (least significant) p-value p_m [27]. For a hypothesis H_k , its unadjusted p-value p_k is compared to an adjusted level of significance $\alpha'_k = \frac{\alpha}{m+1-k}$ where α is the unadjusted level of significance (0.05 in our case), m is the total number of hypotheses in the family, and k is the index of hypothesis in the ordered list (counting from 1 to m). Let k^\dagger be the lowest index k such that $p_k > \alpha'_k$. The hypotheses H_k where $k < k^\dagger$ are accepted as having statistically significance evidence in favor of them (more technically, the corresponding null hypotheses are rejected). The hypotheses H_k where $k \geq k^\dagger$ are not accepted as having significant evidence in favor of them (their null hypotheses are not rejected).

We report adjusted p-values to give an intuition about the strength of evidence for a hypothesis. We let $p'_k = p(m+1-k)$ be the adjusted p-value for H_k provided $k < k^\dagger$ since $p_k > \alpha'_k$ iff $p'_k > \alpha$. Note that

Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes

MUHAMMAD ALI*, Northeastern University, USA

PIOTR SAPIEZYNSKI*, Northeastern University, USA

MIRANDA BOGEN, Upturn, USA

ALEKSANDRA KOROLOVA, University of Southern California, USA

ALAN MISLOVE, Northeastern University, USA

AARON RIEKE, Upturn, USA

The enormous financial success of online advertising platforms is partially due to the precise targeting features they offer. Although researchers and journalists have found many ways that advertisers can target—or exclude—particular groups of users seeing their ads, comparatively little attention has been paid to the implications of the platform’s *ad delivery* process, comprised of the platform’s choices about which users see which ads.

It has been hypothesized that this process can “skew” ad delivery in ways that the advertisers do not intend, making some users less likely than others to see particular ads based on their demographic characteristics. In this paper, we demonstrate that such skewed delivery occurs on Facebook, due to market and financial optimization effects as well as the platform’s own predictions about the “relevance” of ads to different groups of users. We find that both the advertiser’s budget and the content of the ad each significantly contribute to the skew of Facebook’s ad delivery. Critically, we observe significant skew in delivery along gender and racial lines for “real” ads for employment and housing opportunities despite neutral targeting parameters.

Our results demonstrate previously unknown mechanisms that can lead to potentially discriminatory ad delivery, even when advertisers set their targeting parameters to be highly inclusive. This underscores the need for policymakers and platforms to carefully consider the role of the ad delivery optimization run by ad platforms themselves—and not just the targeting choices of advertisers—in preventing discrimination in digital advertising.¹

CCS Concepts: • **Information systems** → **Social advertising**; • **Human-centered computing** → *Empirical studies in HCI*; • **Applied computing** → Law;

Keywords: online advertising; ad delivery; bias; fairness; policy

ACM Reference Format:

Muhammad Ali*, Piotr Sapiezynski*, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. Discrimination through Optimization: How Facebook’s Ad Delivery Can Lead to Biased Outcomes. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 3, CSCW, Article 199 (November 2019). ACM, New York, NY. 30 pages. <https://doi.org/10.1145/3359301>

¹The delivery statistics of ad campaigns described in this work can be accessed at <https://facebook-targeting.ccs.neu.edu/>

* These two authors contributed equally

Authors’ addresses: Muhammad Ali*, Northeastern University, USA, mali@ccs.neu.edu; Piotr Sapiezynski*, Northeastern University, USA, sapiezynski@gmail.com; Miranda Bogen, Upturn, USA, mirandabogen@gmail.com; Aleksandra Korolova, University of Southern California, USA, korolova@usc.edu; Alan Mislove, Northeastern University, USA, amislove@ccs.neu.edu; Aaron Rieke, Upturn, USA, aaron@upturn.org.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.

2573-0142/2019/11-ART199 \$15.00

<https://doi.org/10.1145/3359301>

1 INTRODUCTION

Powerful digital advertising platforms fund most popular online services today, serving ads to billions of users daily. At a high level, the functionality of these advertising platforms can be divided into two phases: *ad creation*, where advertisers submit the text and images that comprise the content of their ad and choose targeting parameters, and *ad delivery*, where the platform delivers ads to specific users based on a number of factors, including advertisers' budgets, their ads' performance, and the predicted relevance of their ads to users.

One of the underlying reasons for the popularity of these services with advertisers is the rich suite of *targeting* features they offer during ad creation, which allow advertisers to precisely specify which users (called the *audience*) are eligible to see the advertiser's ad. The particular features that advertisers can use for targeting vary across platforms, but often include demographic attributes, behavioral information, users' personally identifiable information (PII), mobile device IDs, and web tracking pixels [11, 73].

Due to the wide variety of targeting features—as well as the availability of sensitive targeting features such as user demographics and interests—researchers have raised concerns about discrimination in advertising, where groups of users may be excluded from receiving certain ads based on advertisers' targeting choices [69]. This concern is particularly acute in the areas of credit, housing, and employment, where there are legal protections in the U.S. that prohibit discrimination against certain protected classes in advertising [1–3]. As ProPublica demonstrated in 2016 [33], this risk is not merely theoretical: ProPublica investigators were able to run housing ads that explicitly excluded users with specific “ethnic affinities” from receiving them.² Recently, the U.S. Department of Housing and Urban Development (HUD) sued Facebook over these concerns and others, accusing Facebook's advertising platform of “encouraging, enabling, and causing” violations of the Fair Housing Act [32].

The role of ad delivery in discrimination Although researchers and investigative journalists have devoted considerable effort to understanding the potential discriminatory outcomes of ad targeting, comparatively little effort has focused on ad delivery, due to the difficulty of studying its impacts without internal access to ad platforms' data and mechanisms. However, there are several potential reasons why the ad delivery algorithms used by a platform may open the door to discrimination.

First, consider that most platforms claim their aim is to show users “relevant” ads: for example, Facebook states “we try to show people the ads that are most pertinent to them” [68]. Intuitively, the goal is to show ads that particular users are likely to engage with, even in cases where the advertiser does not know *a priori* which users are most receptive to their message. To accomplish this, the platforms build extensive user interest profiles and track ad performance to understand how different users interact with different ads. This historical data is then used to steer future ads towards those users who are most likely to be interested in them, and to users like them. However, in doing so, the platforms may inadvertently cause ads to deliver primarily to a skewed subgroup of the advertiser's selected audience, an outcome that the advertiser may not have intended or be aware of. As noted above, this is particularly concerning in the case of credit, housing, and employment, where such skewed delivery might violate antidiscrimination laws.

Second, market effects and financial optimization can play a role in ad delivery, where different desirability of user populations and unequal availability of users may lead to skewed ad delivery [25].

²In response, Facebook banned the use of certain attributes for housing ads, but many other, un-banned, mechanisms exist for advertisers that achieve the same outcome [69]. Facebook agreed as part of a lawsuit settlement stemming from these issues to go further by banning age, gender, and certain kinds of location targeting—as well as some related attributes—for housing, employment, or credit ads [22].

For example, it is well-known that certain users on advertising platforms are more valuable to advertisers than others [48, 55, 65]. Thus, advertisers who choose low budgets when placing their ads may be more likely to lose auctions for such “valuable” users than advertisers who choose higher budgets. However, if these “valuable” user demographics are strongly correlated with protected classes, it could lead to discriminatory ad delivery due to the advertiser’s budget alone. Even though a low budget advertiser may not have intended to exclude such users, the ad delivery system may do just that because of the higher demand for that subgroup.

Prior to this work, although hypothesized [25, 52, 72], it was not known whether the above factors resulted in skewed ad delivery in real-world advertising platforms. In fact, in response to the HUD lawsuit [32] mentioned above, Facebook claimed that the agency had “no evidence” of their ad delivery systems’ role in creating discrimination [45].

Contributions In this paper, we aim to understand whether ads could end up being shown in a skewed manner—i.e., where some users are less likely than others to see ads based on their demographic characteristics—due to the ad delivery phase alone. In other words, we determine whether the ad delivery could cause skewed delivery *that an advertiser did not cause by their targeting choices and may not even be aware of*. We focus on Facebook—as it is the most mature platform offering advanced targeting features—and run dozens of ad campaigns, hundreds of ads with millions of impressions, spending over \$8,500 as part of our study.

Answering this question—especially without internal access to the ad delivery algorithm, user data, and advertiser targeting data or delivery statistics—involves overcoming a number of challenges. These include separating market effects from optimization effects, distinguishing ad delivery adjustments based on the ad’s performance measured through user feedback from initial ad classification, and developing techniques to determine the racial breakdown of the delivery audience (which Facebook does not provide). The difficulty of solving these without the ad platform’s co-operation in a rigorous manner may at least partially explain the lack of knowledge about the potential discriminatory effects due to ad delivery to date. After addressing these challenges, we find the following:

First, we find that *skewed delivery can occur due to market effects alone*. Recall the hypothesis above concerning what may happen if advertisers in general value users differently across protected classes. Indeed, we find this is the case on Facebook: when we run identical ads targeting the same audience but with varying budgets, the resulting audience of users who end up seeing our ad can range from over 55% men (for ads with very low budgets) to under 45% men (for ads with high budgets).

Second, we find that *skewed delivery can occur due to the content of the ad itself* (i.e., the ad headline, text, and image, collectively called the *ad creative*). For example, ads targeting the same audience but that include a creative that would stereotypically be of the most interest to men (e.g., bodybuilding) can deliver to over 80% men, and those that include a creative that would stereotypically be of the most interest to women (e.g., cosmetics) can deliver to over 90% women. Similarly, ads referring to cultural content stereotypically of most interest to Black users (e.g., hip-hop) can deliver to over 85% Black users, and those referring to content stereotypically of interest to white users (e.g., country music) can deliver to over 80% white users, even when targeted identically by the advertiser. Thus, despite placing the same bid on the same audience, the advertiser’s ad delivery can be heavily skewed based on the ad creative alone.

Third, we find that *the ad image itself has a significant impact on ad delivery*. By running experiments where we swap different ad headlines, text, and images, we demonstrate that the differences in ad delivery can be significantly affected by the image alone. For example, an ad whose headline

and text would stereotypically be of the most interest to men with the image that would stereotypically be of the most interest to women delivers primarily to women at the same rate as when all three ad creative components are stereotypically of the most interest to women.

Fourth, we find that *the ad image is likely automatically classified by Facebook*, and that this classification can skew delivery from the beginning of the ad's run. We create a series of ads where we add an alpha channel to stereotypically male and female images with over 98% transparency; the result is an image with all of the image data present, but that looks like a blank white square to humans. We find that there are statistically significant differences in how these ads are delivered depending on the image used, despite the ads being visually indistinguishable to a human. This indicates that the image classification—and, therefore, relevance determination—is likely an automated process, and that the skew in ad delivery can be due in large part to skew in Facebook's automated estimate of relevance, rather than ad viewers' interactions with the ad.

Fifth, we show that *real-world employment and housing ads can experience significantly skewed delivery*. We create and run ads for employment and housing opportunities, and use our methodology to measure their delivery to users of different races and genders. When optimizing for clicks, we find that ads with the same targeting options can deliver to vastly different racial and gender audiences depending on the ad creative alone. In the most extreme cases, our ads for jobs in the lumber industry reach an audience that is 72% white and 90% male, our ads for cashier positions in supermarkets reach an 85% female audience, and our ads for positions in taxi companies reach a 75% Black audience, even though the targeted audience specified by us as an advertiser is identical for all three. We run a similar suite of ads for housing opportunities, and find skew there as well: despite the same targeting and budget, some of our ads deliver to an audience of over 72% Black users, while others delivery to over 51% Black users. While our results only speak to how our particular ads are delivered (i.e., we cannot say how housing or employment ads *in general* are delivered), the significant skew we observe even on a small set of ads suggests that real-world housing and employment ads are likely to experience the same fate.

Taken together, our results paint a distressing picture of heretofore unmeasured and unaddressed skew that can occur in online advertising systems, which have significant implications for discrimination in targeted advertising. Specifically, due to platforms' optimization in the ad delivery stage together with market effects, ads can unexpectedly be delivered to skewed subsets of the advertiser's specified audience. For certain types of ads, such skewed delivery might implicate legal protections against discriminatory advertising. For example, Section 230 of the U.S. Communications Decency Act (CDA) protects publishers (including online platforms) from being held responsible for third-party content. Our results show Facebook's integral role in shaping the delivery mechanism and might make it more difficult for online platforms to present themselves as neutral publishers in the future. We leave a full exploration of these implications to the legal community. However, our results indicate that regulators, lawmakers, and the platforms themselves need to think carefully when balancing the optimization of ad platforms against desired societal outcomes, and remember that ensuring that individual advertisers do not discriminate in their targeting is insufficient to achieve non-discrimination goals sought by regulators and the public.

Ethics All of our experiments were conducted with careful consideration of ethics. We obtained Institutional Review Board review of our study at Northeastern University (application #18-11-13), with our protocol being marked as "Exempt". We minimized harm to Facebook users when we were running our ads by always running "real" ads (in the sense that if people clicked on our ads, they were brought to real-world sites relevant to the topic of the ad). While running our ads, we never intentionally chose to target ads in a discriminatory manner (e.g., we never used discriminatory targeting parameters). To further minimize the potential for discrimination, we ran most of our

experimental ads in categories with no legal salience (such as entertainment and lifestyle); we only ran ad campaigns on jobs and housing to verify whether the effects we observed persist in these domains. We minimized harm to the Facebook advertising platform by paying for ads and using the ad reporting tools in the same manner as any other advertiser. The particular sites we advertised were unaffiliated with the study, and our ads were not defamatory, discriminatory, or suggestive of discrimination.

2 BACKGROUND

Before introducing our methodology and analyses, we provide background on online display advertising, describe Facebook's advertising platform's features, and detail related work.

2.1 Online display advertising

Online display advertising is now an ecosystem with aggregate yearly revenues close to \$100 billion [21]. The web advertising ecosystem is a complex set of interactions between ad publishers, ad networks, and ad exchanges, with an ever-growing set of entities involved at each step allowing advertisers to reach much of the web. In contrast, online services such as Facebook and Twitter run advertising platforms that primarily serve a single site (namely, Facebook and Twitter themselves). In this paper, we focus on single-site advertising platforms, but our results may also be applicable to more general display advertising on the web; we leave a full investigation of the extent to which this is the case to future work.

The operation of platforms such as Facebook and Twitter can be divided into two phases: *ad creation* and *ad delivery*. We provide more details on each below.

Ad creation Ad creation refers to the process by which the advertiser submits their ad to the advertising platform. At a high level, the advertiser has to select three things when doing so:

- (1) *Ad contents*: Advertisers will typically provide the ad headline, text, and any images/videos. Together, these are called the *ad creative*. They will also provide the link where the platform should send users who click.
- (2) *Audience Selection/Targeting*: Advertisers need to select which platform users they would like to see the ad (called the *audience*).
- (3) *Bidding strategy*: Advertisers need to specify how much they are willing to pay to have their ads shown. This can come in the form of a per-impression or per-click bid, or the advertiser can simply place an overall *bid cap* and allow the platform to bid on their behalf.

Once the advertiser has entered all of the above information, they submit the ad for review;³ once it is approved, the ad will move to the ad delivery phase.

Ad delivery Ad delivery refers to the process by which the advertising platform shows ads to users. For every opportunity to show a user an ad (e.g., an *ad slot* is available as the user is browsing the service), the ad platform will run an *ad auction* to determine, from among all of the ads that include the current user in the audience, which ad should be shown.

In practice, however, the ad delivery process is somewhat more complicated. *First*, the platforms try to avoid showing ads from the same advertiser repeatedly in quick succession to the same user; thus, the platforms will sometimes disregard bids for recent winners of the same user. *Second*, the platforms often wish to show users relevant ads; thus, rather than relying solely on the bid to determine the winner of the auction, the platform may incorporate a relevance score into consideration, occasionally allowing ads with lower bids but more relevance to win over those

³Most platforms have a review process to prevent abuse or violations of their platforms' advertising policies [8, 77].

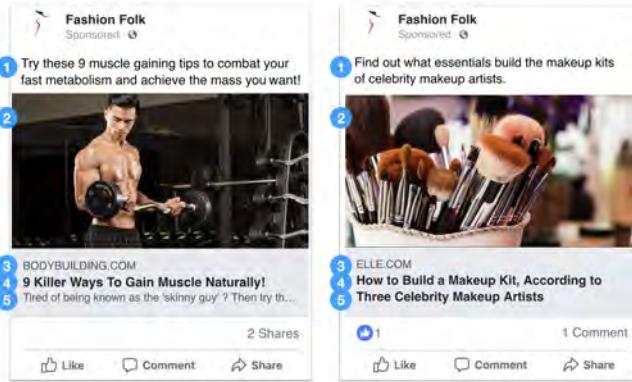


Fig. 1. Each ad has five elements that the advertiser can control: (1) the ad text, entered manually by the advertiser, (2) the images and/or videos, (3) the domain, pulled automatically from the HTML meta property `og:site_name` of the destination URL, (4) the title, pulled automatically from the HTML meta property `og:title` of the destination URL, and (5) the description from meta property `og:description` of the destination URL. The title and description can be manually customized by the advertiser if they wish.

with higher bids. *Third*, the platforms may wish to evenly spread the advertiser budget over their specified time period, rather than use it all at once, which introduces additional complexities as to which ads should be considered for particular auctions. The exact mechanisms by which these issues are addressed are not well-described or documented by the platforms.

Once ads enter the ad delivery phase, the advertising platforms give advertisers information on how their ads are performing. Such information may include detailed breakdowns (e.g., along demographic or geographic lines) of the characteristics of users to whom their ad is being shown and those who click on the ad.

2.2 Facebook's advertising platform

In this paper, we focus on Facebook's advertising platform as it is one of the most powerful and feature-rich advertising platforms in use today. As such, we provide a bit more background here about the specific features and options that Facebook provides to advertisers.

Ad contents Each ad placed on Facebook must be linked to a *Page*; advertisers are allowed to have multiple Pages and run ads for any of them. Ads can come in multiple forms, such as promoting particular posts on the page. However, for typical ads, the advertiser must provide (a) the headline and text to accompany the ad, and (b) one or more images or videos to show to the user. Optionally, the advertiser can provide a *traffic destination* to send the user to if they click (e.g., a Facebook Page or an external URL); if the advertiser provides a traffic destination, the ad will include a brief description (auto-generated from the HTML meta data) about this destination. Examples showing all of these elements are presented in Figure 1.

Audience selection Facebook provides a wide variety of audience selection (or *targeting*) options [10, 11, 38, 69]. In general, these options fall into a small number of classes:

- *Demographics and attributes*: Similar to other advertising platforms [39, 71], Facebook allows advertisers to select audiences based on demographic information (e.g., age, gender, and location), as well as profile information, activity on the site, and data from third-parties. Recent work has shown that Facebook offers over 1,000 well-defined attributes and hundreds of thousands of free-form attributes [69].

- *Personal information:* Alternatively, Facebook allows advertisers to specify *the exact users* who they wish to target by either (a) uploading the users' personally identifiable information including names, addresses, and dates of birth [34, 73, 74], or (b) deploying web tracking pixels on third-party sites [27]. On Facebook, audiences created using either mechanism are called *Custom Audiences*.⁴
- *Similar users:* Advertisers may wish to find "similar" users to those who they have previously selected. To do so, Facebook allows advertisers to create *Lookalike Audiences*⁵ by starting with a source Custom Audience they had previously uploaded; Facebook then "identif[ies] the common qualities of the people in it" and creates a new audience with other people who share those qualities [28].

Advertisers can often combine many of these features together, for example, by uploading a list of users' personal information and then using attribute-based targeting to further narrow the audience.

Objective and bidding Facebook provides advertisers with a number of *objectives* to choose from when placing an ad [6], where each tries to maximize a different *optimization event* the advertiser wishes to occur. These include "Awareness" (simply optimizing for the most *impressions*, a.k.a. views), "Consideration" (optimizing for clicks, engagement, etc.), and "Conversion" (optimizing for sales generated by clicking the ad). For each objective, the advertiser bids on the objective itself (e.g., for "Awareness", the advertiser would bid on ad impressions). The bid can take multiple forms, and includes the start and end time of the ad campaign and either a lifetime or a daily budget cap. With these budget caps, Facebook places bids in ad auctions on the advertisers' behalf. Advertisers can optionally specify a per-bid cap as well, which will limit the amount Facebook would bid on their behalf for a single optimization event.

Facebook's ad auction When Facebook has ad slots available, it runs an ad auction among the active advertisements bidding for that user. However, the auction does not just use the bids placed by the advertisers; Facebook says [29]:

The ad that wins an auction and gets shown is the one with the highest *total value* [emphasis added]. Total value isn't how much an advertiser is willing to pay us to show their ad. It's combination of 3 major factors: (1) Bid, (2) Estimated action rates, and (3) Ad quality and relevance.

Facebook defines "Estimated action rates" as "how well an ad performs", meaning whether or not *users in general* are engaging with the ad [5]. They define "Ad quality and relevance" as "how interesting or useful we think a given user is going to find a given ad", meaning how much *a particular user* is likely to be interested in the ad [5].

Thus, it is clear that Facebook attempts to identify the users within an advertiser's selected audience who they believe would find the ad most useful (i.e., those who are most likely to result in an optimization event) and shows the ad preferentially to those users. Facebook says exactly as such in their documentation [4]:

During ad set creation, you chose a target audience ... and an optimization event ... We show your ad to people in that target audience who are likely to get you that optimization event

⁴Google, Twitter, and Pinterest all provide similar features; these are called *Customer Match* [7], *Tailored Audiences*, and *Customer Lists* [61], respectively.

⁵Google and Pinterest offer similar features: on Google it is called *Similar Audiences* [40], and on Pinterest it is called *Actalike Audiences* [63].

Facebook provides advertisers with an overview of how well-matched it believes an ad is with the target audience using a metric called *relevance score*, which ranges between 1 and 10. Facebook says [68]:

Relevance score is calculated based on the positive and negative feedback we expect an ad to receive from its target audience.

Facebook goes on to say [68]:

Put simply, the higher an ad’s relevance score, the less it will cost to be delivered. This is because our ad delivery system is designed to show the right content to the right people, and a high relevance score is seen by the system as a positive signal.

Statistics and reporting Facebook provides advertisers with a feature-rich interface [30] as well as a dedicated API [56] for both launching ads and monitoring those ads as they are in ad delivery. Both the interface and the API give semi-live updates on delivery, showing the number of impressions and optimization events as the ad is running. Advertisers can also request this data be broken down along a number of different dimensions, including age, gender, and location (Designated Market Area [58], or DMA, region). Notably, the interface and API *do not* provide a breakdown of ad delivery along racial lines; thus, analyzing delivery along racial lines necessitates development of a separate methodology that we describe in the next section.

Anti-discrimination rules In response to issues of potential discrimination in online advertising reported by researchers and journalists [33], Facebook currently has several policies in place to avoid discrimination for certain types of ads. Facebook also recently built tools to automatically detect ads offering housing, employment, and credit, and pledged to prevent the use of certain targeting categories with those ads. [46]. Additionally, Facebook relies on advertisers to self-certify [15] that they are not in violation of Facebook’s advertising policy prohibitions against discriminatory practices [31]. More recently, in order to settle multiple lawsuits stemming from these reports, Facebook no longer allows age, gender, or ZIP code-based targeting for housing, employment or credit ads, and blocks other detailed targeting attributes that are “describing or appearing to relate to protected classes” [22, 44, 60].

2.3 Related work

Next, we detail related work on algorithm auditing, transparency, and discriminatory ad targeting.

Auditing algorithms for fairness Following the growing ubiquity of algorithms in daily life, a community formed around investigating their societal impacts [66]. Typically, the algorithms under study are not available to outside auditors for direct examination; thus, most researchers treat them as “black boxes” and observe their reactions to different inputs. Among most notable results, researchers have shown price discrimination in online retail sites [42], gender discrimination in job sites [16, 43], stereotypical gender roles re-enforced by online translation services [12] and image search [47], disparate performance on gender classification for Black women [13], and political partisanship in search [20, 51, 64]. Although most of the work focused exclusively on the algorithms themselves, recently researchers began to point out that auditors should consider the entire socio-technical systems that include the users of those algorithms, an approach referred to as “algorithm-in-the-loop” [41, 67]. Furthermore, recent work has demonstrated that fairness is not necessarily composable, i.e., for several notions of fairness such as individual fairness [24], a collection of classifiers that are fair in isolation do not necessarily result in a fair outcome when they are used as part of a larger system [25].

Advertising transparency In parallel to the developments in detecting and correcting unfairness, researchers have conducted studies and introduced tools with the aim of increasing transparency and explainability of algorithms and their outcomes. For example, much attention has been dedicated to shedding light on the factors that influence the targeting of a particular ad on the web [26, 53, 54, 62] and on specific services [19, 78].

Focusing on Facebook, Andreou et al. investigated the transparency initiative from Facebook that purportedly tells users why they see particular targeted ads [11]. They found that the provided explanations are incomplete and, at times, misleading. Venkatadri et al. introduced the tool called “TREADS” that attempts to close this gap by providing Facebook users with detailed descriptions of their inferred attributes using the ads themselves as a vehicle [75]. Further, they investigated how data from third-party data brokers is used in Facebook’s targeting features and—for the first time—revealed those third-party attributes to the users themselves using TREADS [76]. Similar to other recent work [59], Venkatadri et al. found that the data from third-party data brokers had varying accuracy [76].

Discrimination in advertising As described above, Facebook has some policies and tools in place to prevent discriminatory ad targeting. However, advertisers can still exclude users based on a variety of interests that are highly correlated with race by using custom audiences [69], or by using location [37, 50]. Separately, Sweeney [70] and Datta et al. [19] have studied discrimination in Google’s advertising system.

The work just described deals with identifying possibilities for the advertisers to run discriminatory ads using the platform’s features. In contrast, other researchers, as well as HUD’s recent complaint, have suggested that discrimination may be introduced by the ad platform itself, rather than by a malicious advertiser [19, 45, 52, 72]. For example, Lambrecht et al. ran a series of ads for STEM education and found they were consistently delivered more to men than to women, even though there are more female users on Facebook, and they are known to be more likely to click on ads and generate conversions [52]. Datta et al. explored ways that discrimination could arise in the targeting and delivery of job-related ads, and analyzed how different parties might be liable under existing law [18]. Our work explores these findings in depth, separating market effects from optimization effects and exploring the mechanisms by which ads are delivered in a skewed manner.

3 METHODOLOGY

We now describe our methodology for measuring the delivery of Facebook ads. At a high level, our goal is to run groups of ads where we vary a particular feature, with the goal of then measuring how changing that feature skews the set of users the Facebook platform delivers the ad to. To do so, we need to carefully control which users are in our target audience. We also need to develop a methodology to measure the ad delivery skew along racial lines, which, unlike gender, is not provided by Facebook’s existing reporting tools. We detail how we achieve that in the following sections.

3.1 Audience selection

When running ads, we often wish to control exactly which ad auctions we are participating in. For example, if we are running multiple instances of the same ad (e.g., to establish statistical confidence), we do not want the instances to be competing against each other. To this end, we use random PII-based custom audiences, where we randomly select U.S. Facebook users to be included in mutually-exclusive audiences. By doing so, we can ensure that our ads are only competing against each other in the cases where we wish them to. We also replicate some of the experiments while targeting all U.S. users to ensure that the effects do not only exist when custom audiences are

DMA(s) [58]	# Records (A)		# Records (B)		# Records (C)	
	White	Black	White	Black	White	Black
Wilmington, Raleigh–Durham	400,000	0	0	400,000	900,002	0
Greenville-Spartanburg, Greenville-New Bern, Charlotte, Greensboro	0	400,000	400,000	0	0	892,097

Table 1. Overview of the North Carolina custom audiences used to measure racial delivery. We divide the most populated DMAs in the state into two sets, and create three audiences each with one race per DMA set. Audiences A and B are disjoint from each other; audience C contains the voters from A with additional white voters from the first DMA set and Black voters from the second DMA set. We then use the statistics Facebook reports about delivery by DMAs to infer delivery by race.

targeted. As we show later in Section 4, we observe equivalent skews in these scenarios, which leads us to believe that preventing internal competition between our own ads is not crucial to measure the resulting skews.

Generating custom audiences We create each custom audience by randomly generating 20 lists of 1,000,000 distinct, valid North American phone numbers (+1 XXX XXX XXXX, using known-valid area codes). Facebook reported that they were able to match approximately 220,000 users on each of the 20 lists we uploaded.

Initially, we used these custom audiences directly to run ads, but while conducting the experiments we noticed that—even though we specifically target only North American phone numbers—many ads were delivered to users outside of North America. This could be caused by users traveling abroad, users registering with fake phone numbers or with online phone number services, or for other reasons, whose investigation is outside the scope of this paper. Therefore, for all the experiments where we target custom audiences, we additionally limit them to people located in the U.S.

3.2 Data collection

Once one of our ad campaigns is run, we use the Facebook Marketing API to obtain the delivery performance statistics of the ad every two minutes. When we make this request, we ask Facebook to break down the ad delivery performance according to the attribute of study (age, gender, or location). Facebook’s response to each query features the following fields, among others, for each of the demographic attributes that we requested:

- **impressions:** The number of times the ad was shown
- **reach:** The number of unique users the ad was shown to
- **clicks:** The number of clicks the ad has received
- **unique_clicks:** The number of unique users who clicked

Throughout the rest of the paper, we use the **reach** value when examining delivery; thus, when we report “Fraction of men in the audience” we calculate this as the **reach** of men divided by the sum of the **reach** of men and the **reach** of women (see Section 3.5 for discussion on using binary values for gender).

3.3 Measuring racial ad delivery

The Facebook Marketing API allows advertisers to request a breakdown of ad delivery performance along a number of axes but it does not provide a breakdown based on race. However, for the purposes of this work, we are able to measure the ad delivery breakdown along racial lines by using location (Designated Market Area, or DMA⁶) as a proxy.

Similar to prior work [69], we obtain voter records from North Carolina; these are publicly available records that have the name, address, race, and often phone number of each registered voter in the state. We partition the most populated North Carolina DMAs into two sets; for the exact DMAs, please see Table 1. We ensure that each racial group (white and Black) from a set of DMAs has a matching number of records of the other group in the other set of DMAs. We sample three audiences (A , B , and C) that fit these constraints from the voter records and upload as separate Custom Audiences to Facebook.⁷ Audiences A and B are disjoint from each other; audience C contains the voters from A with additional white voters from the first DMA set and Black voters from the second DMA set. We create audiences in this way to be able to test both “flipped” versions of the audiences (A and B), as well as large audiences with as many users as possible (C); we created audience B as large as possible (exhausting all voters who fit the necessary criteria), and sampled audience A to match its size. The details of the resulting audiences are shown in Table 1.

When we run ads where we want to examine the ad delivery along racial lines, we run the ads to one audience (A , B , or C). We then request that Facebook’s Marketing API deliver us results broken down by DMA. Because we selected DMAs to be a proxy for race, we can use the results to infer which custom audience they were originally in, allowing us to determine the racial makeup of the audience who saw (and clicked on) the ad. Note that in experiments that involve measuring racial skew all ads target the same target audience. The limited number of registered voters does not allow us to create many large, disjoint custom audiences like we do in other experiments. However, as we show with ads targeting all U.S. users, internal competition does not appear to influence the results.

3.4 Ad campaigns

We use the Facebook Ad API described in Section 2.2 to create all ads for our experiments and to collect data on their delivery. We carefully control for any time-of-day effects that might be present due to different user demographics using Facebook at different times of the day: for any given experiment, we run all ads at the same time to ensure that any such effects are experienced equally by all ads. Unless otherwise noted, we used the following settings:

- *Objective*: Consideration→Traffic⁸
- *Optimization Goal*: Link Clicks
- *Traffic destination*: An external website (that depends on the ads run)
- *Creative*: All of our ads had a single image and text relevant to the ad.
- *Audience selection*: We use custom audiences for many of our ads, as described in Section 3.1, and further restrict them to adult (18+) users of all genders residing in the United States. For other ads, we target all U.S. users age 18 or older.

⁶Designated Market Areas [58] are groups of U.S. counties that Neilson defines as “market areas”; they were originally used to signify a region where users receive similar broadcast television and radio stations. Facebook reports ad delivery by location using DMAs, so we use them here as well.

⁷Unfortunately, Facebook does not report the number of these users who match as we use multiple PII fields in the upload file [73].

⁸This target is defined as: Send more people to a destination on or off Facebook such as a website, app, or Messenger conversation.

- *Budget:* We ran most ads with a budget of \$20 per day, and stopped them typically after six hours.

3.5 Measuring and comparing audiences

We now describe the measurements we make during our experiments and how we compute their confidence intervals.

Binary values of gender and race Facebook’s marketing API reports “female”, “male”, and “uncategorized” as the possible values for gender. Facebook’s users self-report their gender, and the available values are “female”, “male”, and “custom”. The latter allows the user to manually type in their gender (with 60 predefined gender identities suggested through auto-complete functionality) and select the preferred pronoun from “female - her”, “male - him”, and “neutral - them”. Across our experiments, we observe that up to 1% of the audiences are reported as “uncategorized” gender. According to Facebook’s documentation this represents the users who did not list their gender.⁹ We do not know whether the “uncategorized” gender also features users with self-reported “custom” gender. Thus, in this work we only consider the self-reported binary gender values of “female” and “male”.

Further, when considering racial bias, we use the self-reported information from voter records. The data we obtained has 7,560,885 individuals, with 93% reporting their race as either Black or White. Among those, less than 1% report their ethnicity as “Hispanic/Latino”. Thus, in this work, we only target the individuals with self-reported race of White or Black. However, when running our experiments measuring race (and targeting specific DMAs), we observe that a fraction (~10%) of our ads are delivered to audiences outside of our predefined DMAs, thus making it impossible for us to infer their race. This fraction remains fairly consistent across our experiments regardless of what we advertise, thus introducing the same amount of noise across our measurements. This is not entirely unexpected, as we are targeting users directly, and those users may be traveling, may have moved, may have outdated information in the voter file, etc.

We do not claim that gender or race are binary, but choose to focus the analysis on users who self-reported their gender as “female” or “male” and race as “Black” or “White”. This way, we report the observable skew in delivery only along these axes. We recognize that delivery can be *further* skewed with respect to gender of non-binary users and/or users of other races in a way that remains unreported in this work.

Measuring statistical significance Using the binary race and gender features, throughout this work, we describe the audiences by the fraction of male users and the fraction of white users. We calculate the lower and upper limits of the 99% confidence interval around this fraction using the method recommended by Agresti and Coull [9], defined in Equation 1:

$$\begin{aligned} L.L. &= \frac{\hat{p} + \frac{z_{\alpha/2}^2}{2n} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1 + z_{\alpha/2}^2/n}, \\ U.L. &= \frac{\hat{p} + \frac{z_{\alpha/2}^2}{2n} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1 + z_{\alpha/2}^2/n}, \end{aligned} \quad (1)$$

where *L.L.* is the lower confidence limit, *U.L.* is the upper confidence limit, \hat{p} is the observed fraction of the audience with the attribute (here: male), n is the size of the audience reached by the

⁹<https://www.facebook.com/business/help/151999381652364>

ad. To obtain the 99% interval we set $z_{\alpha/2} = 2.576$. The advantage of using this calculation instead of the more frequently used normal approximation

$$p \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \quad (2)$$

is that the resulting intervals fall in the $(0, 1)$ range. Whenever the confidence intervals around these fractions for two audiences are non-overlapping, we can make a claim that the gender or racial makeups of two audiences are significantly different [17]. However, the converse is not true: overlapping confidence intervals do not necessarily mean that the means are not different (see Figure 4 in [17] for explanation). In this work we report all the results of our experiments but for easier interpretation emphasize those where the confidence intervals are non-overlapping. We further confirm that the non-overlapping confidence intervals represent statistically significant differences, using the difference of proportion test as shown in Equation 3:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1 - \hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} \quad (3)$$

where \hat{p}_1 and \hat{p}_2 are the fractions of men (white users) in the two audiences that we compare, n_1 and n_2 are sizes of these audiences, and \hat{p} is the fraction of men (white users) in the two delivery audiences combined. All the results we refer to as statistically significant are significant in this test with a Z -score of at least 2.576. Finally, as we present in the Appendix, the comparisons presented are statistically significant also after the application of Bonferroni correction [14] for multiple hypotheses testing.

Note that in experiments where we run multiple instances of an ad targeting disjoint custom audiences, the values of \hat{p} and n are calculated from the sums of reached audiences.

4 EXPERIMENTS

In this section, we explore how an advertiser’s choice of ad creative (headline, text, and image) and ad campaign settings (bidding strategy, targeted audience) can affect the demographics (gender and race) of the users to whom the ad is ultimately delivered.

4.1 Budget effects on ad delivery

We begin by examining the impact that market effects can have on delivery, aiming to test the hypothesis put forth by Lambrecht et al. [52]. In particular, they observed that their ads were predominantly shown to men even though women had consistently higher click through rates (CTRs). They then hypothesized that the higher CTRs led to women being more expensive to advertise to, meaning they were more likely to lose auctions for women when compared to auctions for men.

We test this hypothesis by running the same ad campaign with different budgets; our goal is to measure the effect that the daily budget alone has on the makeup of users who see the ads. When running these experiments, we keep the ad creative and targeted audience constant, only changing the bidding strategy to give Facebook different daily limits (thus, any ad delivery differences can be attributed to the budget alone). We run an ad with daily budget limits of \$1, \$2, \$5, \$10, \$20, and \$50, and run multiple instances at each budget limit for statistical confidence. Finally, we run the experiment twice, once targeting our random phone number custom audiences, and once targeting all users located in U.S.; we do so to verify that any effect we see is not a function of our particular target audience, and that it persists also when non-custom audiences are targeted.

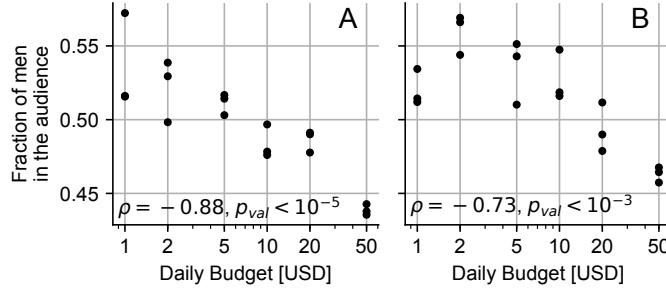


Fig. 2. Gender distributions of the audience depend on the daily budget of an ad, with higher budgets leading to a higher fraction of women. The left graph shows an experiment where we target all users located in the U.S.; the right graph shows an experiment where we target our random phone number custom audiences.

Figure 2 presents the results, plotting the daily budget we specify versus the resulting fraction of men in the audience. The left graph shows the results when we target all users located in the U.S., and the right graph shows the results when we target the random phone number custom audiences. In both cases, we observe that changes in ad delivery due to differences in budget are indeed happening: the higher the daily budget, the smaller the fraction of men in the audience, with the Pearson’s correlation of $\rho = -0.88, p_{val} < 10^{-5}$ for all U.S. users and $\rho = -0.73, p_{val} < 10^{-3}$ for the custom audiences.

The stronger effect we see when targeting all U.S. users may be due to the additional freedom that the ad delivery system has when choosing who to deliver to, as this is a significantly larger audience.

To eliminate the impact that market effects can have on delivery in our following experiments, we ensure that all runs of a given experiment use the same bidding strategy and budget limit. Typically we use a daily budget of \$20 per campaign.

4.2 Ad creative effects on ad delivery

Now we examine the effect that the ad creative (headline, text, and image) can have on ad delivery. To do so, we create two stereotypical ads that we believe would appeal primarily to men and women, respectively: one ad focusing on *bodybuilding* and another on *cosmetics*. The actual ads themselves are shown in Figure 1. We run each of the ads at the same time and with the same bidding strategy and budget. For each variable we target different custom audiences, i.e., the “base” level ads target one audience, “text” level ads target another, etc. *Note that we do not explicitly target either ad based on gender; the only targeting restrictions we stipulate are 18+ year old users in the U.S.*

We observe dramatic differences in ad delivery, even though the bidding strategy is the same for all ads, and each pair of ads target the same gender-agnostic audience. In particular, the bodybuilding ad ended up being delivered to over 75% men on average, while the cosmetics ad ended up being delivered to over 90% women on average. Again, this skewed delivery is despite the fact that we—the advertiser—did not specify difference in budget or target audience.

Individual components’ impact on ad delivery With the knowledge that the ad creative can skew delivery, we dig deeper to determine *which* of the components of the ad creative (headline, text, and image) have the greatest effect on ad delivery. To do so, we stick with the bodybuilding and cosmetics ads, and “turn off” various features of the ad creative by replacing them with empty strings or blank images. For example, the bodybuilding experiment listed as “base” includes an empty headline, empty ad text, and a blank white image; it does however link to the domain

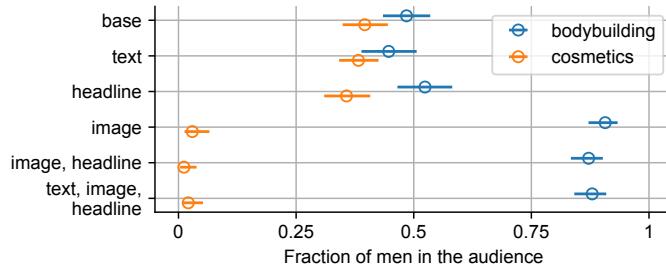


Fig. 3. “Base” ad contains a link to a page about either bodybuilding or cosmetics, a blank image, no text, or headline. There is a small difference in the fraction of male users for the base ads, and adding the “text” only decreases it. Setting the “headline” sets the two ads apart but the audience of each is still not significantly different than that of the base version. Finally, setting the ad “image” causes drastic changes: the bodybuilding ad is shown to a 91% male audience, the cosmetics ad is shown to a 5% male audience, despite the same target audience.

bodybuilding.com. Similarly, the cosmetics experiment listed as “base” includes no headline, text, or image, but does link to the domain elle.com. We then add back various parts of the ad creative, as shown in Figure 1.

The results of this experiment are presented in Figure 3. Error bars in the figure correspond to 99% confidence intervals as defined in Equation 1. All results are shown relative to that experiment’s “base” ad containing only the destination URL. We make a number of observations. *First*, we can observe an ad delivery difference due to the destination URL itself; the base bodybuilding ad delivers to 48% men, while the base cosmetics ad delivers to 40% men. *Second*, as we add back the title and the headline, the ad delivery does not appreciably change from the baseline. However, once we introduce the image into the ad, the delivery changes dramatically, returning to the level of skewed delivery discussed above (over 75% male for bodybuilding, and over 90% female for cosmetics). When we add the text and/or the headline back alongside the image, the skew of delivery does not change significantly compared to the presence of image only. Overall, our results demonstrate that the choice of ad image can have a dramatic effect on which users in the audience ultimately are shown the ad.

Swapping images To further explore how the choice of image impacts ad delivery, we continue using the bodybuilding and cosmetics ads, and test how ads with incongruent images and text are delivered. Specifically, we swap the images between the two ads, running an ad with the bodybuilding headline, text, and destination link, but with the image from cosmetics (and vice versa). We also run the original ads (with congruent images and text) for comparison.

The results of this experiment are presented in Figure 4, showing the skew in delivery of the ads over time. The color of the lines indicates the image that is shown in the ad; solid lines represent the delivery of ads with images consistent with the description, while dotted lines show the delivery for ads where image was replaced. We make a number of observations. *First*, when using congruent ad text and image (solid lines), we observe the skew we observed before. However, we can now see clearly that this delivery skew appears to exist from the very beginning of the ad delivery, i.e., before users begin viewing and interacting with our ads. We will explore this further in the following section. *Second*, we see that when we switch the images—resulting in incongruent ads (dotted lines)—the skew still exists but to a lesser degree. Notably, we observe that the ad with an image of bodybuilding but cosmetics text delivers closest to 50:50 across genders, but the ad with the image of cosmetics but bodybuilding text does not. The exact mechanism by which Facebook

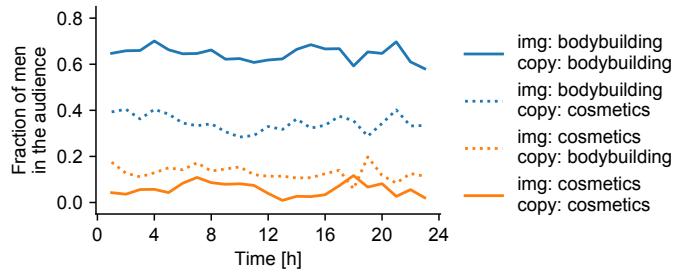


Fig. 4. Ad delivery of original bodybuilding and cosmetics ads, as well as the same ads with incongruent images. Skew in delivery is observed from the beginning. Using incongruent images skews the delivery to a lesser degree, indicating that the image is not the only element of the ad that drives the skew in delivery.

decides to use the ad text and images in influencing ad delivery is unknown, and we leave a full exploration to future work.

Swapping images mid-experiment Facebook allows advertisers to change their ad while it is running, for example, to update the image or text. As a final point of analysis, we examine how changing the ad creative mid-experiment—after it has started running—affects ad delivery. To do so, we begin the experiment with the original congruent bodybuilding and cosmetics ads; we let these run for over six hours. We then swap the images on the running ads, thereby making the ads incongruent, and examine how ad delivery changes.

Figure 5 presents the results of this experiment. In the top graph, we show the instantaneous ad delivery skew: as expected, the congruent ads start to deliver in a skewed manner as we have previously seen. After the image swap at six hours, we notice a very rapid change in delivery with the ads almost completely flipping in ad delivery skew in a short period of time. Interestingly, we do not observe a significant change in users’ behavior to explain this swap: the bottom graph plots the click through rates (CTRs) for both ads by men and women over time. Thus, our results suggest that the change in ad delivery skew is unlikely to be due to the users’ responses to the ads.

4.3 Source of ad delivery skew

We just observed that ads see a significant skew in ad delivery due to the contents of the ad, despite the bidding strategy and targeting parameters being held constant. However, we observed that the ad delivery skew was present from the very beginning of ad delivery, and that swapping the image in the middle of a run resulted in a very rapid change in ad delivery that could not be explained by how frequently users click on our ads. We now turn to explore the mechanism that may be leading to this ad delivery skew.

Almost-transparent images We begin with the hypothesis that Facebook itself is automatically classifying the ad creative (including the image), and using the output of this classification to calculate a predicted relevance score to users. In other words, we hypothesize that Facebook is running automatic text and image classification, rather than (say) relying on the ad’s initial performance, which would explain (a) the delivery skew being present from the beginning of ad delivery, and (b) how the delivery changes rapidly despite no significant observable change in user behavior. However, validating this hypothesis is tricky, as we are not privy to all of Facebook’s ad performance data.

To test this hypothesis, we take an alternate approach. We use the *alpha channel* that is present in many modern image formats; this is an additional channel that allows the image to encode the

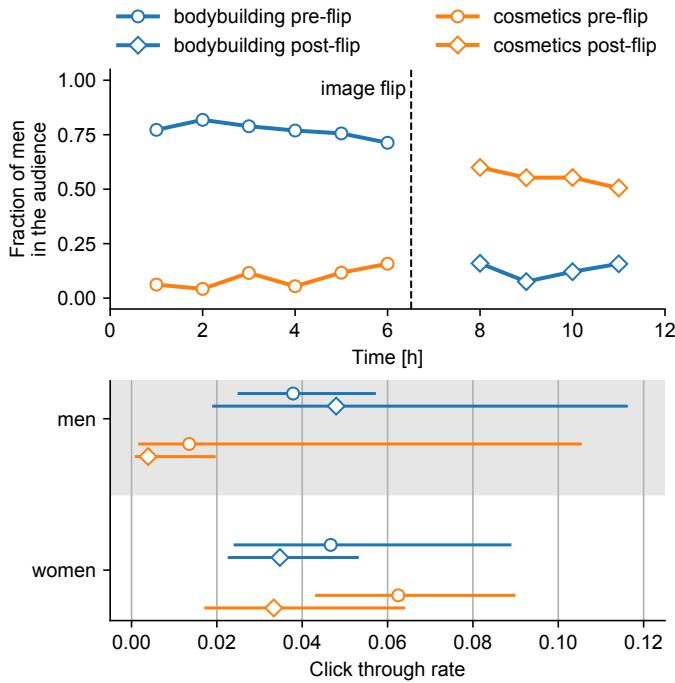


Fig. 5. When we flip the image in the middle of the campaign, the ad is reclassified and shown to an updated audience. Here, we start bodybuilding and cosmetics ads with corresponding descriptions and after 6 hours and 32 minutes we flip the images. Within an hour of the change, the gender proportions are reversed, while there is no significant difference between the click through rates per gender pre and post flipping of the images.

transparency of each pixel. Thus, if we take an image and add an alpha channel with (say) 99% opacity, all of the image data will still be present in the image, but any human who views the image would not be able to see it (as the image would show almost completely transparent). However, if an automatic classifier exists, and if that classifier is not properly programmed to handle the alpha channel, it may continue to classify the image.

Test images To test our hypothesis, we select five images that would stereotypically be of interest to men and five images that would stereotypically be of interest to women; these are shown in the second and fourth columns of Table 2.^{10, 11} We convert them to PNG format and add an alpha channel with 98% opacity¹² to each of these images; these are shown in the third and fifth columns of Table 2. Because we cannot render a transparent image without a background, the versions in the paper are rendered on top of a white background. As the reader can see, these images are not discernible to the human eye.

We first ran a series of tests to observe how Facebook's ad creation phase handled us uploading such transparent images. If we used Reach as our ad objective, we found that Facebook "flattened"

¹⁰All of these images were cropped from images posted to pexels.com, which allow free non-commercial use.

¹¹We cropped these images to the Facebook-recommended resolution of 1,080×1,080 pixels to reduce the probability Facebook would resample the image.

¹²We were unable to use 100% transparency as we found that Facebook would run an image hash over the uploaded images and would detect different images with 100% opacity to be the same (and would refuse to upload it again). By using 98% transparency, we ensure that the images were still almost invisible to humans but that Facebook would not detect they were the same image.

No.	Masculine		Feminine	
	Visible	Invisible	Visible	Invisible
1				
2				
3				
4				
5				

Table 2. Diagram of the images used in the transparency experiments. Shown are the five stereotypical masculine and feminine images, along with the same images with a 98% alpha channel, denoted as invisible. The images with the alpha channel are almost invisible to humans, but are still delivered in a skewed manner.

these images onto a white background in the ad preview.¹³ By targeting ourselves with these Reach ads, we verified that when they were shown to users on the Facebook mobile app or in the desktop Facebook web feed, the images did indeed show up as white squares. Thus, we can use this methodology to test whether there is an automatic image classifier present by examining whether running different transparent white ads results in different delivery.

Results We run ads with all twenty of the images in Table 2, alongside ads with five truly blank white images for comparison. For all 25 of these ads, we hold the ad headline, text, and destination link constant, run them all at the same time, and use the same bidding strategy and target custom audiences in a way that each user is potentially exposed to up to three ads (one masculine image, one feminine image, and one blank image). We then record the differences in ad delivery of these 25 images along gender lines. The results are presented in Figure 6A, with all five images in each of the five groups aggregated together. We can observe that ad delivery is, in fact, skewed, with the ads with stereotypically masculine images delivering to over 43% men and the ads with feminine images delivering to 39% men in the experiment targeting custom audiences as well as 58% and 44% respectively in the experiment targeting all U.S. users. Error bars in the plot correspond to the 99% confidence interval calculated using Equation 1.

¹³Interestingly, we found that if we instead used Traffic as our ad objective, Facebook would both “flatten” these images onto a white background *and then normalize the contrast*. This caused the ads to be visible to humans—simply with less detail than the original ads—thus defeating the experiment. We are unsure of why Facebook did not choose to normalize images with the objective for Reach.

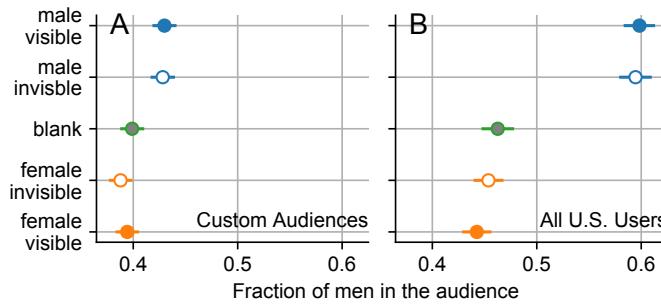


Fig. 6. Fraction of reached men in the audiences for ads with the images from Table 2, targeting random phone number custom audience (A) and US audience (B). The solid markers are visible images, and the hollow markers are the same images with 98% opacity. Also shown is the delivery to truly white images (“blank”). We can observe that a difference in ad delivery exists, and that that difference is statistically significant between the masculine and feminine invisible images. This suggests that automated image classification is taking place.

Interestingly, we also observe that the masculine invisible ads appear to be indistinguishable in the gender breakdown of their delivery from the masculine visible ads, and the feminine invisible ads appear to be indistinguishable in their delivery from the feminine visible ads.

As shown in Figure 6A, we verify that the fraction of men in the delivery of the male ads is significantly higher than in female-centered and neutral ads, as well as higher in neutral ads than in female-centered ads. We also show that we cannot reject the null hypothesis that the fraction of men in the two versions of each ad (one visible, one invisible) are the same. Thus, we can conclude that the difference in ad delivery of our invisible male and female images is statistically significant, despite the fact that humans would not be able to perceive any differences in these ads. This strongly suggests that our hypothesis is correct: that Facebook has an automated image classification mechanism in place that is used to steer different ads towards different subsets of the user population.¹⁴

To confirm this finding, we re-run the same experiment except that we change the target audience from our random phone number custom audiences (hundreds of thousands of users) to all U.S. users (over 320 million users). Our theory is that if we give Facebook's algorithm a larger set of auctions to compete in, any effect of skewed delivery would be amplified as they may be able to find more users for whom the ad is highly “relevant”. In Figure 6B we observe that the ad delivery differences are, indeed, even greater: the male visible and invisible images deliver to approximately 60% men, while the female visible and invisible images deliver to approximately 45% men. Moreover, the statistical significance of this experiment is even stronger, with a Z value over 10 for the ad delivery difference between the male invisible and female invisible ads.

4.4 Impact on real ads

We have observed that differences in the ad headline, text, and image can lead to dramatic difference in ad delivery, despite the bidding strategy and target audience of the advertiser remaining the same. However, all of our experiments thus far were on test ads where we typically changed only

¹⁴It is important to note we not know exactly how the classification works. For example, the classifier may also be programmed to take in the “flattened” images that appear almost white, but there may sufficient data present in the images for the classification to work. We leave a full exploration of how exactly the classifier is implemented to future work.

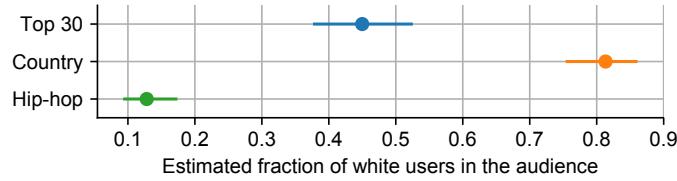


Fig. 7. We run three campaigns about the best selling albums. *Top 30* is neutral, targeting all. *Country* implicitly targets white users, and *Hip-hop* implicitly targets Black users. Facebook classification picks up on the implicit targeting and shows it to the audience we would expect.

a single variable. We now turn to examine the impact that ad delivery can have on realistic ads, where all properties of the ad creative can vary.

Entertainment ads We begin by constructing a series of benign entertainment ads that, while holding targeting parameters fixed (targeting custom audience C from Table 1, are stereotypically of interest to different races. Namely, we run three ads leading to lists of best albums in the previous year: general top 30 (neutral), top country music (stereotypically of interest mostly to white users), and top hip-hop albums (stereotypically of interest mostly to Black users). We find that Facebook ad delivery follows the stereotypical distribution, despite all ads being targeted in the same manner and using the same bidding strategy. Figure 7 shows the fraction of white users in the audience in the three different ads, treating race as a binary (Black users constitute the remaining fraction). Error bars represent 99% confidence intervals calculated using Equation 1.

Neutral ads are seen by a relatively balanced, 45% white audience, while the audiences receiving the country and hip-hop ads are 80% and 13% white, respectively. Assuming significant population level differences of preferences, it can be argued that this experiment highlights the “relevance” measures embedded in ad delivery working as intended. Next, we investigate cases where such differences may not be desired.

Employment ads Next, we advertise eleven different generic job types: artificial intelligence developer, doctor, janitor, lawyer, lumberjack, nurse, preschool teacher, restaurant cashier, secretary, supermarket clerk, and taxi driver. For each ad, we customize the text, headline, and image as a real employment ad would. For example, we advertise for taxi drivers with the text “Begin your career as a taxi driver or a chauffeur and get people to places on time.” For each ad, we link users to the appropriate category of job listings on a real-world job site.

When selecting the ad image for each job type, we select five different stock photo images: one that has a white male, one that has a white female, one that has a black male, one that has a black female, and one that is appropriate for the job type but has no people in it. We run each of these five independently to test a representative set of ads for each job type, looking to see how they are delivered along gender and racial lines (targeting custom audience C from Table 1). We run these ads for 24 hours, using the objective of Traffic, all targeting the same audience with the same bidding strategy.

The results of this experiment are presented in Figure 8, plotting the distribution of each of our ads along gender (left graph) and racial (right graph) lines. As before, the error bars represent the 99% confidence interval calculated using Eq. 1. We can immediately observe drastic differences in ad delivery across our ads along both racial and gender lines: our five ads for positions in the lumber industry deliver to over 90% men and to over 70% white users in aggregate, while our five ads for janitors deliver to over 65% women and over 75% black users in aggregate. Recall that the

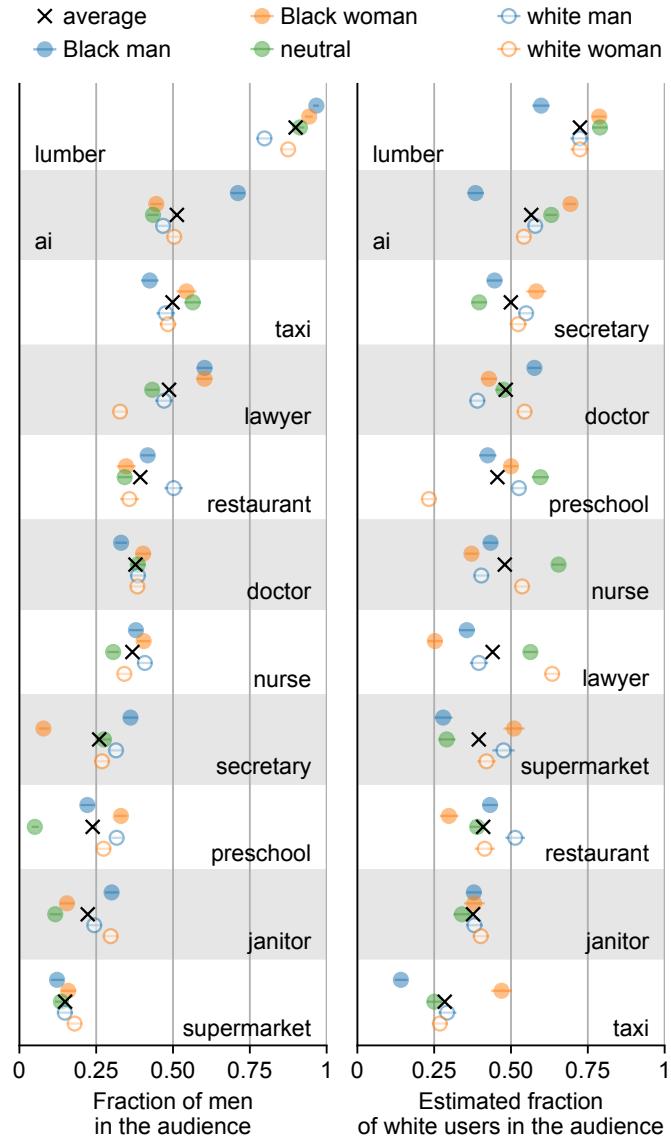


Fig. 8. Results for employment ads, showing a breakdown of ad delivery by gender (left figure) and race (right figure) in the ultimate delivery audience. The labels refer to the race/gender of the person in the ad image (if any). The jobs themselves are ordered by the average fraction of men or white users in the audience. Despite the same bidding strategy, the same target audience, and being run at the same time, we observe significant skew along on both racial and gender lines due to the content of the ad alone.

only difference between these ads are the ad creative and destination link; we (the advertiser) used the same bidding strategy and target audience, and ran all ads at the same time.

Furthermore, we note that the skew in delivery cannot merely be explained by possibly different levels of competition from other advertisers for white and Black users or for male and female users. Although it is the case that each user may be targeted by a different number of advertisers with varying bid levels, by virtue of running all of our job ads at the same time, targeting the same users, with the same budget, we are ensuring that our ads are experiencing competition from other

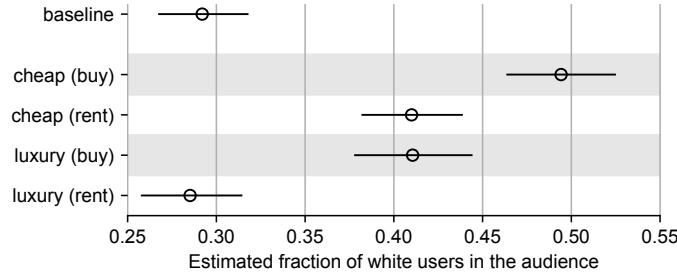


Fig. 9. Results for housing ads, showing a breakdown in the ad delivery audience by race. Despite being targeted in the same manner, using the same bidding strategy, and being run at the same time, we observe significant skew in the makeup of the audience to whom the ad is delivered (ranging from estimated 27% white users for luxury rental ads to 49% for cheap house purchase ads).

advertisers identically. In other words, our ad targeting asks that every user who is considered for our “lumberjack” job ad should also be considered for our taxi driver job ad, so these ads should be competing with each other and facing identical competition from other advertisers at auction time. If the content of the ad was not taken into account by the delivery optimization system, then the ads would be expected to have similar—though not necessarily even—breakdowns by race and gender at delivery. Our experiment demonstrates that this is not the case, and thus, aspects of ad delivery optimization, rather than merely advertiser competition, influence the skew in the delivery outcome.

Housing ads Finally, we create a suite of ads that advertise a variety of housing opportunities, as discrimination in online housing ads has recently been a source of concern [32]. We vary the type of property advertised (rental vs. purchase) and the implied cost (fixer-upper vs. luxury). In each ad, the cost is implied through wording of the ad as well as the accompanying image. Each ad points to a listing of houses for sale or rental apartments in North Carolina on a real-world housing site. Simultaneously, we ran a baseline ad with generic (non-housing) text that simply links to google.com. All of the ads ran for 12 hours, using the objective of Traffic, all targeting the same North Carolina audiences and using the same bidding strategy. We construct the experiment such that each particular ad is run twice: once targeting audience *A* and once targeting audience *B* (see Table 1) This way we eliminate any potential geographical effects (for example, users in Wilmington could be interested in cheap houses to buy, and users in Charlotte could be interested in luxury rentals regardless of their ethnicity, but if we only used audience *C* that effect could appear as racial skew).

We present the results in Figure 9 (we found little skew for the housing ads along gender lines, and we omit those results). We observe significant ad delivery skew along racial lines in the delivery of our ads, with certain ads delivering to an audience of over 72% Black users (comparable to the baseline results) while others delivering to an audience of as little as 51% Black users.

As with the employment ads, we cannot make claims about what particular properties of our ads lead to this skew, or about how housing ads in general are delivered. However, given the significant skew we observe with our suite of ads, it indicates the further study is needed to understand how real-world housing ads are delivered.

5 CONCLUDING DISCUSSION

To date, the public debate and ad platform's comments about discrimination in digital advertising have focused heavily on the targeting features offered by advertising platforms, and the ways that advertisers can misuse those features [23].

In this paper, we set out to investigate a different question: *to what degree and by what means may advertising platforms themselves play a role in creating discriminatory outcomes?*

Our study offers an improved understanding of the mechanisms behind and impact of ad delivery, a process distinct from ad creation and targeting. While ad targeting is facilitated by an advertising platform—but nominally controlled by advertisers—ad delivery is conducted and controlled by the advertising platform itself. We demonstrate that, during the ad delivery phase, advertising platforms can play an independent, central role in creating skewed, and potentially discriminatory, outcomes. More concretely, we have:

- Replicated and affirmed prior research suggesting that market and pricing dynamics can create conditions that lead to differential outcomes, by showing that the lower the daily budget for an ad, the fewer women it is delivered to.
- Shown that Facebook's ad delivery process can significantly alter the audience the ad is delivered to compared to the one intended by the advertiser based on the content of the ad itself. We used public voter record data to demonstrate that broadly and inclusively targeted ads can end up being differentially delivered to specific audience segments, even when we hold the budget and target audience constant.
- Demonstrated that skewed ad delivery can start at the beginning of an ad's run. We also showed that this process is likely automated on Facebook's side, and is not a reflection of the early feedback received from users in response to the ad, by using transparent images in ads that appear the same to humans but are distinguishable by automatic image classification tools, and showing they result in skewed delivery.
- Confirmed that skewed delivery can take place on real-world ads for housing and employment opportunities by running a series of employment ads and housing ads with the same targeting parameters and bidding strategy. Despite differing only in the ad creative and destination link, we observed skewed delivery along racial and gender lines.

We briefly discuss some limitations of our work and touch on the broader implications of our findings.

Limitations It is important to note that while we have revealed certain aspects of how ad delivery is accomplished, and the effects it had on our experimental ad campaigns, we cannot make broad conclusions about how it impacts ads more generally. For example, we observe that all of *our ads* for lumberjacks deliver to an audience of primarily white and male users, but that may not hold true of *all ads* for lumberjacks. However, the significant ad delivery skew that we observe for our employment and housing ads strongly suggests that such skew is present for such ads run by real-world advertisers.

Skew vs. discrimination Throughout this paper we refer to differences in the demographics of reached audience as “skew” in delivery. We do not claim any observed skew *per se* is necessarily wrong or should be mitigated. Without making value judgements on skew in general, we do emphasize the distinct case of ads for housing and employment. In particular, the skew we observe in the delivery of ads for cosmetics or bodybuilding might be interpreted as reinforcing gender stereotypes but is unlikely to have legal implications. On the other hand, the skew in delivery of employment and housing ads is potentially discriminatory in a legal sense.

Further, for the experiments involving ethnicity, we attempted to create equally sized audiences (50% white and 50% Black). However, solely the fact that ads are not delivered to an evenly split audience does not indicate skew, as there might be differences in matching rates (what fraction of registered voters are active Facebook users) per ethnicity, or the groups could have different temporal usage patterns. Only when we run two or more ads at the same time, targeting the same audience, and these ads are delivered with different proportions to white and Black users, do we claim we observe skew in delivery.

Our focus lies in understanding the extent to which the ad platform's delivery optimization, rather than merely its targeting tools and their use as implied by Facebook [23], determine the outcomes of ad delivery, and on highlighting that demographic skews presently arise for certain legally protected categories in Facebook, even when the advertiser targets broadly and inclusively.

Skew in traditional media Showing ads to individuals most likely to engage with them is one of the cornerstone promises of online ad platforms. While in traditional media—such as newspapers and television—advertisers can also place their ads strategically to reach particular kinds of readers or viewers, there are three significant differences with implications for fairness and discrimination when compared to advertising on Facebook.

First, when advertising in traditional media, *the advertiser* has the ability to purposefully advertise to a wide and diverse audience, and be assured that their ads will reach that audience. As we show in this work, this is not the case for advertising on Facebook. Even if the advertiser intends to reach a general and diverse audience, their ad can be steered to a narrow slice within that specified audience, that is skewed in unexpected or undesirable ways.

Second, *the individual's* agency to see ads targeted at groups they do not belong to is more severely limited in the hyper-targeted and delivery-optimized scenario of online ad platforms. In traditional media, an individual interested in seeing ads targeted to a different demographic than they belong to has to merely watch programming or read a newspaper that they are not usually a target demographic for. On Facebook, finding out what ads one may be missing out on due to gender, race, or other characteristic inferred or predicted by Facebook is more challenging. A particularly motivated user could change their self-reported gender but might find themselves discouraged from doing so because the account's gender information is always public. Other characteristics, such as race and net worth, are inferred by Facebook (or accessed via third-party companies [76]) rather than obtained through user's self-reported data, which makes them challenging to alter for the purposes of seeing ads. Moreover, although users can remove some of their inferred interests using ad controls on Facebook, they have no ability to control *negative inferences* Facebook may be making about them. For example, Facebook may infer that a particular user is "not interested in working at a lumber yard", and therefore, not show this user ads for a lumberjack job even if the employer is trying to reach them. As a result, Facebook would be excluding them from an opportunity in ways unbeknownst to the user and to the advertiser.

Third, *public interest scrutiny* of the results of advertising is much more difficult in online delivery-optimized systems than in traditional media. Advertising in traditional media can be easily observed and analyzed by many members of society, from individuals to journalists, and targeting and delivery outside the expectation norms can be detected and called out by many. In the case of hyper-targeted online advertising whose delivery is controlled by the platform, such scrutiny is currently outside reach for most ads [36, 57].

Policy implications Our findings underscore the need for policymakers and platforms to carefully consider the role of the optimizations run by the platforms themselves—and not just the targeting choices of advertisers—in seeking to prevent discrimination in digital advertising.

First, because discrimination can arise in ad delivery independently from ad targeting, limitations on ad targeting—such as those currently deployed by Facebook to limit the targeting features that can be used—will not address discrimination arising from ad delivery. On the contrary, to the extent limiting ad targeting features prompts advertisers to rely on larger target audiences, the mechanisms of ad delivery will have an even greater practical impact on the ads that users see.

Second, regulators, lawmakers, and platforms themselves will need to more deeply consider whether and how longstanding civil rights laws apply to modern advertising platforms in light of ad delivery dynamics. At a high level, U.S. federal law prohibits discrimination in the marketing of housing, employment and credit opportunities. A detailed consideration of these legal regimes is beyond the scope of this paper. However, our findings show that ad platforms themselves can shape access to information about important life opportunities in ways that might present a challenge to equal opportunity goals.

Third, in the U.S., Section 230 of the Communications Decency Act (CDA) provides broad legal immunity for internet platforms acting as publishers of third-party content. This immunity was a central issue in recently-settled litigation against Facebook, who argued its ad platform should be protected by CDA Section 230 in part because its advertisers are “wholly responsible for deciding where, how, and when to publish their ads.” [35] Our research shows that this claim is misleading, particularly in light of Facebook’s role in determining the ad delivery outcomes. Even absent unlawful behavior by advertisers, our research demonstrates that Facebook’s own, independent actions during the delivery phase are crucial to determining how, when, and to whom ads are shown, and might produce unlawful outcomes. These effects can be invisible to, and might even create liability for, Facebook’s advertisers.

Thus, the effects we observed could introduce new liability for Facebook. In determining whether Section 230 protections apply, courts consider whether an internet platform “materially contributes” to the alleged illegal conduct. Courts have yet to squarely consider how the delivery mechanisms described in this paper might affect an ad platform’s immunity under Section 230.

Fourth, our results emphasize the need for increased transparency into advertising platforms, particularly around ad delivery algorithms and statistics for real-world housing, credit, or employment ads. Facebook’s existing ad transparency efforts are not yet sufficient to allow researchers to analyze the impact of ad delivery in the real world.

Potential mitigations Given the potential impact that discriminatory ad delivery can have on exposure to opportunities available to different populations, a natural question is how ad platforms such as Facebook may mitigate these effects. This is not straightforward, and is likely to require increased commitment and transparency from ad platforms as well as development of new algorithmic and machine learning techniques. For instance, as we have demonstrated empirically in Section 4.1 (and as [25] have shown theoretically), skewed ad delivery can occur even if the ad platform refrains from refining the audience supplied by the advertisers according to the predicted relevance of the ad to individual users. This happens because different users are valued differently by advertisers, which, in a setting of limited user attention, leads to a tension between providing a useful service for users and advertisers, fair ad delivery, and the platform’s own revenue goals.¹⁵

Thus, more advanced and nuanced approaches to addressing the potential issues of discrimination in digital advertising are necessary. Developing them is beyond the scope of this paper; however, we can imagine several different options, each with their own pros and cons. First, Facebook and similar platforms could disable optimization altogether for some ads, and deliver them to a random sample of users within an advertiser’s target audience (with or without market considerations).

¹⁵A formal statement of this claim for the theoretical notions of individual fairness [24] and its generalization, preference-informed fairness, can be found in [49].

Second, platforms could remove ads in protected categories from their normal ad flows altogether, and provide those listings in a separate kind of marketing product (e.g., , a user-directed listing service like craigslist.org). Third, the platforms could allow the advertisers to enforce their own demographic outcomes so long as those desired outcomes don't themselves violate anti-discrimination laws. Finally, the platforms could seek to constrain their delivery optimization algorithms to satisfy chosen fairness criteria (some candidates for such criteria from the theoretical computer science community are individual fairness [24] and preference-informed fairness [49], but a broader discussion of appropriate criteria involving policymakers is needed).

Digital advertising increasingly influences how people are exposed to the world and its opportunities, and helps keep online services free of monetary cost. At the same time, its potential for negative impacts, through optimization due to ad delivery, is growing. Lawmakers, regulators, and the ad platforms themselves need to address these issues head-on.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their helpful comments. We also thank NaLette Brodnax and Christo Wilson for their invaluable feedback on the manuscript and Martin Goodson for pointing out erroneous confidence intervals. The authors also acknowledge Hannah Masuga, a graduate fellow at Upturn, whose initial experiments during her fellowship inspired this research. This work was funded in part by a grant from the Data Transparency Lab and NSF grant CNS-1616234. This work was done in part while Aleksandra Korolova was visiting the Simons Institute for the Theory of Computing.

REFERENCES

- [1] 12 CFR § 202.4 (b) – Discouragement [n.d.]. 12 Cfr § 202.4 (b) – Discouragement. <https://www.law.cornell.edu/cfr/text/12/202.4>.
- [2] 24 CFR § 100.75 – Discriminatory advertisements, statements and notices [n.d.]. 24 Crf § 100.75 – Discriminatory Advertisements, Statements And Notices. <https://www.law.cornell.edu/cfr/text/24/100.75>.
- [3] 29 USC § 623 – Prohibition of age discrimination [n.d.]. 29 Usc § 623 – Prohibition Of Age Discrimination. <https://www.law.cornell.edu/uscode/text/29/623>.
- [4] About Ad Delivery [n.d.]. About Ad Delivery. <https://www.facebook.com/business/help/1000688343301256>.
- [5] About Ad Principles [n.d.]. About Ad Principles. <https://www.facebook.com/business/about/ad-principles>.
- [6] About advertising objectives [n.d.]. About Advertising Objectives. <https://www.facebook.com/business/help/517257078367892>.
- [7] About Customer Match [n.d.]. About Customer Match. <https://support.google.com/adwords/answer/6379332?hl=en>.
- [8] About Twitter Ads approval [n.d.]. About Twitter Ads Approval. <https://business.twitter.com/en/help/ads-policies/introduction-to-twitter-ads/about-twitter-ads-approval.html>.
- [9] Alan Agresti and Brent A Coull. 1998. Approximate Is Better Than “exact” For Interval Estimation Of Binomial Proportions. *The American Statistician* 52, 2 (1998), 119–126.
- [10] Athanasios Andreou, Márcio Silva, Fabrício Benevenuto, Oana Goga, Patrick Loiseau, and Alan Mislove. 2019. Measuring The Facebook Advertising Ecosystem. In *Network and Distributed System Security Symposium*. San Diego, California, USA.
- [11] Athanasios Andreou, Giridhari Venkatadri, Oana Goga, Krishna P. Gummadi, Patrick Loiseau, and Alan Mislove. 2018. Investigating Ad Transparency Mechanisms In Social Media: A Case Study Of Facebook’s Explanations. In *Network and Distributed System Security Symposium*. San Diego, California, USA.
- [12] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man Is To Computer Programmer As Woman Is To Homemaker? Debiasing Word Embeddings. In *Neural and Information Processing Systems*. Barcelona, Spain.
- [13] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities In Commercial Gender Classification. In *Conference on Fairness, Accountability, and Transparency*. New York, New York, USA.
- [14] Robert J. Cabin and Randall J. Mitchell. 2000. To Bonferroni Or Not To Bonferroni: When And How Are The Questions. *Bulletin of the Ecological Society of America* 81, 3 (2000), 246–248.
- [15] Certify Compliance to Facebook’s Non-Discrimination Policy [n.d.]. Certify Compliance To Facebook’s Non-discrimination Policy. <https://www.facebook.com/business/help/136164207100893>.

- [16] Le Chen, Aniko Hannak, Ruijin Ma, and Christo Wilson. 2018. Investigating The Impact Of Gender On Rank In Resume Search Engines. In *Annual Conference of the ACM Special Interest Group on Computer Human Interaction*. Montreal, Canada.
- [17] Geff Cumming and Sue Finch. 2005. Inference By Eye: Confidence Intervals And How To Read Pictures Of Data. *American Psychologist* 60, 2 (2005), 170.
- [18] Amit Datta, Anupam Datta, Jael Makagon, Deirdre K. Mulligan, and Michael Carl Tschantz. 2018. Discrimination In Online Personalization: A Multidisciplinary Inquiry. In *Conference on Fairness, Accountability, and Transparency*. New York, New York, USA.
- [19] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated Experiments On Ad Privacy Settings: a Tale Of Opacity, Choice, And Discrimination. In *Privacy Enhancing Technologies Symposium*. Philadelphia, Pennsylvania, USA.
- [20] Nicholas Diakopoulos, Daniel Trielli, Jennifer Stark, and Sean Mussenden. 2018. I Vote For—how Search Informs Our Choice Of Candidate. In *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple*, M. Moore and D. Tambini (Eds.). Oxford University Press.
- [21] Digital Ad Spend Hits Record-Breaking \$49.5 Billion in First Half of 2018, Marking a Significant 23% YOY Increase [n.d.]. Digital Ad Spend Hits Record-breaking \$49.5 Billion in First Half of 2018, Marking a Significant 23% YOY Increase. <https://www.iab.com/news/digital-ad-spend-hits-record-breaking-49-5-billion-in-first-half-of-2018/>.
- [22] Doing More to Protect Against Discrimination in Housing, Employment and Credit Advertising [n.d.]. Doing More To Protect Against Discrimination In Housing, Employment And Credit Advertising. <https://newsroom.fb.com/news/2019/03/protecting-against-discrimination-in-ads/>.
- [23] Doing More to Protect Against Discrimination in Housing, Employment and Credit Advertising [n.d.]. Doing More To Protect Against Discrimination In Housing, Employment And Credit Advertising. <https://newsroom.fb.com/news/2019/03/protecting-against-discrimination-in-ads/>.
- [24] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ACM, 214–226.
- [25] Cynthia Dwork and Christina Ilvento. 2018. Fairness Under Composition. In *10th Innovations in Theoretical Computer Science Conference (ITCS 2019) (Leibniz International Proceedings in Informatics (LIPIcs))*, Vol. 124. 33:1–33:20. <https://doi.org/10.4230/LIPIcs.ITCS.2019.33>
- [26] eyeWnder_Experiment [n.d.]. Eyewnder_Experiment. <http://www.eyewnder.com/>.
- [27] Facebook: About Facebook Pixel [n.d.]. Facebook: About Facebook Pixel. <https://www.facebook.com/business/help/742478679120153>.
- [28] Facebook: About Lookalike Audiences [n.d.]. Facebook: About Lookalike Audiences. <https://www.facebook.com/business/help/164749007013531>.
- [29] Facebook: About the delivery system: Ad auctions [n.d.]. Facebook: About The Delivery System: Ad Auctions. <https://www.facebook.com/business/help/430291176997542>.
- [30] Facebook Ads Manager [n.d.]. Facebook Ads Manager. <https://www.facebook.com/business/help/200000840044554>.
- [31] Facebook Advertising Policies, Discriminatory Practices [n.d.]. Facebook Advertising Policies, Discriminatory Practices. https://www.facebook.com/policies/ads/prohibited_content/discriminatory_practices.
- [32] Facebook Engages in Housing Discrimination With Its Ad Practices, U.S. Says [n.d.]. Facebook Engages In Housing Discrimination With Its Ad Practices, U.S. Says. <https://www.nytimes.com/2019/03/28/us/politics/facebook-housing-discrimination.html>.
- [33] Facebook Lets Advertisers Exclude Users by Race [n.d.]. Facebook Lets Advertisers Exclude Users By Race. <https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race/>.
- [34] Facebook Marketing API – Custom Audiences [n.d.]. Facebook Marketing Api – Custom Audiences. <https://developers.facebook.com/docs/marketing-api/custom-audiences-targeting/v3.1>.
- [35] Facebook Motion to Dismiss in Onuoha v. Facebook [n.d.]. Facebook Motion To Dismiss In Onuoha V. Facebook. <https://www.courtlistener.com/recap/gov.uscourts.cand.304918/gov.uscourts.cand.304918.34.0.pdf>.
- [36] Facebook's Ad Archive API is Inadequate [n.d.]. Facebook's Ad Archive Api Is Inadequate. <https://blog.mozilla.org/blog/2019/04/29/facebook-s-ad-archive-api-is-inadequate/>.
- [37] Irfan Faizullabhoy and Aleksandra Korolova. 2018. Facebook's Advertising Platform: New Attack Vectors And The Need For Interventions. *Computing Research Repository* (March 2018). <https://arxiv.org/abs/1803.10099>, Workshop on Technology and Consumer Protection (ConPro).
- [38] Avijit Ghosh, Giridhari Venkatadri, and Alan Mislove. 2019. Analyzing Facebook Political Advertisers' Targeting. In *Workshop on Technology and Consumer Protection*. San Francisco, California, USA.
- [39] Google: About audience targeting [n.d.]. Google: About Audience Targeting. <https://support.google.com/google-ads/answer/2497941?hl=en>.

- [40] Google: About similar audiences on the Display Network [n.d.]. Google: About Similar Audiences On The Display Network. <https://support.google.com/google-ads/answer/2676774?hl=en>.
- [41] Ben Green and Yiling Chen. 2019. Disparate Interactions: An Algorithm-in-the-loop Analysis Of Fairness In Risk Assessments. In *Conference on Fairness, Accountability, and Transparency*. Atlanta, Georgia, USA.
- [42] Aniko Hannak, Gary Soeller, David Lazer, Alan Mislove, and Christo Wilson. 2014. Measuring Price Discrimination And Steering On E-commerce Web Sites. In *ACM Internet Measurement Conference*. Vancouver, Canada.
- [43] Aniko Hannak, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. 2017. Bias In Online Freelance Marketplaces: Evidence From Taskrabbit And Fiverr. In *ACM Conference on Computer Supported Cooperative Work*. Portland, Oregon, USA.
- [44] Help: Choosing a Special Ad Category [n.d.]. Help: Choosing A Special Ad Category. <https://www.facebook.com/business/help/298000447747885>.
- [45] HUD Sues Facebook Over Housing Discrimination and Says the Company's Algorithms Have Made the Problem Worse [n.d.]. Hud Sues Facebook Over Housing Discrimination And Says The Company's Algorithms Have Made The Problem Worse. <https://www.propublica.org/article/hud-sues-facebook-housing-discrimination-advertising-algorithms>.
- [46] Improving Enforcement and Promoting Diversity: Updates to Ads Policies and Tools [n.d.]. Improving Enforcement And Promoting Diversity: Updates To Ads Policies And Tools. <http://newsroom.fb.com/news/2017/02/improving-enforcement-and-promoting-diversity-updates-to-ads-policies-and-tools/>.
- [47] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal Representation And Gender Stereotypes In Image Search Results For Occupations. In *Annual Conference of the ACM Special Interest Group on Computer Human Interaction*.
- [48] Larry Kim. 2011. The Most Expensive Keywords In Google Adwords. <http://www.wordstream.com/blog/ws/2011/07/18/most-expensive-google-adwords-keywords/>.
- [49] Michael P. Kim, Aleksandra Korolova, Guy N. Rothblum, and Gal Yona. 2019. Preference-informed Fairness. <https://arxiv.org/abs/1904.01793>.
- [50] Aleksandra Korolova. 2018. Facebook's Illusion Of Control Over Location-related Ad Targeting. Medium. <https://medium.com/@korolova/facebook-s-illusion-of-control-over-location-related-ad-targeting-de7f865aee78>.
- [51] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna Gummadi, and Karrie Karahalios. 2017. Quantifying Search Bias: Investigating Sources Of Bias For Political Searches In Social Media. In *ACM Conference on Computer Supported Cooperative Work*. Portland, Oregon, USA.
- [52] Anja Lambrecht and Catherine E. Tucker. 2018. Algorithmic Bias? An Empirical Study Into Apparent Gender-based Discrimination In The Display Of Stem Career Ads. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2852260.
- [53] Mathias Lecuyer, Guillaume Ducoffe, Francis Lan, Andrei Papancea, Theofilos Petsios, Riley Spahn, Augustin Chaintreau, and Roxana Geambasu. 2014. Xray: Enhancing The Web's Transparency With Differential Correlation. In *USENIX Security Symposium*. San Diego, California, USA.
- [54] Mathias Lecuyer, Riley Spahn, Yannis Spiliopolous, Augustin Chaintreau, Roxana Geambasu, and Daniel Hsu. 2015. Sunlight: Fine-grained Targeting Detection At Scale With Statistical Confidence. In *ACM Conference on Computer and Communications Security*.
- [55] Yabing Liu, Chloe Kliman-Silver, Balachander Krishnamurthy, Robert Bell, and Alan Mislove. 2014. Measurement And Analysis Of Osn Ad Auctions. In *ACM Conference on Online Social Networks*. Dublin, Ireland.
- [56] Marketing API [n.d.]. Marketing Api. <https://developers.facebook.com/docs/marketing-apis/>.
- [57] More than 200 Researchers Sign Letter Supporting Knight Institute's Proposal to Allow Independent Research of Facebook's Platform [n.d.]. More Than 200 Researchers Sign Letter Supporting Knight Institute's Proposal To Allow Independent Research Of Facebook's Platform. <https://knightcolumbia.org/news/more-200-researchers-sign-letter-supporting-knight-institutes-proposal-allow-independent/>.
- [58] Neilson DMA® Regions [n.d.]. Neilson Dma® Regions. <https://www.nielsen.com/intl-campaigns/us/dma-maps.html>.
- [59] Nico Neumann, Catherine E. Tucker, and Timothy Whitfield. 2018. How Effective Is Black-box Digital Consumer Profiling And Audience Delivery?: Evidence From Field Studies. *Social Science Research Network Working Paper Series* (2018).
- [60] New Marketing API Requirements for all Advertising Campaigns [n.d.]. New Marketing Api Requirements For All Advertising Campaigns. <https://developers.facebook.com/blog/post/2019/08/15/new-marketing-api-requirements-for-all-advertising-campaigns/>.
- [61] New targeting tools make Pinterest ads even more effective [n.d.]. New Targeting Tools Make Pinterest Ads Even More Effective. <https://business.pinterest.com/en/blog/new-targeting-tools-make-pinterest-ads-even-more-effective>.
- [62] Javier Parra-Arnau, Jagdish Prasad Acharya, and Claude Castelluccia. 2017. Myadchoices: Bringing Transparency And Control To Online Advertising. *ACM Transactions on the Web* 11 (2017).
- [63] Pinterest: Audience targeting [n.d.]. Pinterest: Audience Targeting. <https://help.pinterest.com/en/business/article/audience-targeting>.

- [64] Ronald E. Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing Partisan Audience Bias Within Google Search. In *Annual Conference of the ACM Special Interest Group on Computer Human Interaction*.
- [65] Diego Saez-Trumper, Yabing Liu, Ricardo Baeza-Yates, Balachander Krishnamurthy, and Alan Mislove. 2014. Beyond Cpm And Cpc: Determining The Value Of Users On Osns. In *ACM Conference on Online Social Networks*. Dublin, Ireland.
- [66] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing Algorithms: Research Methods For Detecting Discrimination On Internet Platforms. In *International Communication Association Conference*.
- [67] Piotr Sapiezynski, Wesley Zeng, Ronald E. Robertson, Alan Mislove, and Christo Wilson. 2019. Quantifying The Impact Of User Attention On Fair Group Representation In Ranked Lists. In *Workshop on Fairness, Accountability, Transparency, Ethics, and Society on the Web*. San Francisco, California, USA.
- [68] Showing Relevance Scores for Ads on Facebook [n.d.]. Showing Relevance Scores For Ads On Facebook. <https://www.facebook.com/business/news/relevance-score>.
- [69] Till Speicher, Muhammad Ali, Giridhari Venkatadri, Filipe Nunes Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna P. Gummadi, Patrick Loiseau, and Alan Mislove. 2018. On The Potential For Discrimination In Online Targeted Advertising. In *Conference on Fairness, Accountability, and Transparency*. New York, New York, USA.
- [70] Latanya Sweeney. 2013. Discrimination In Online Ad Delivery. *Commun. ACM* 56, 5 (2013), 44–54.
- [71] Twitter: Ad targeting [n.d.]. Twitter: Ad Targeting. <https://business.twitter.com/en/targeting.html>.
- [72] Upturn Amicus Brief in Onuoha v. Facebook [n.d.]. Upturn Amicus Brief In Onuoha V. Facebook. <https://www.courtlistener.com/recap/gov.uscourts.cand.304918/gov.uscourts.cand.304918.76.1.pdf>.
- [73] Giridhari Venkatadri, Yabing Liu, Athanasios Andreou, Oana Goga, Patrick Loiseau, Alan Mislove, and Krishna P. Gummadi. 2018. Privacy Risks With facebook's Pii-based Targeting: Auditing A Data Broker's Advertising Interface. In *IEEE Symposium on Security and Privacy*. San Francisco, California, USA.
- [74] Giridhari Venkatadri, Elena Lucherini, Piotr Sapiezyński, and Alan Mislove. 2019. Investigating Sources Of Pii Used In facebook's Targeted Advertising. In *Privacy Enhancing Technologies Symposium*. Stockholm, Sweden.
- [75] Giridhari Venkatadri, Alan Mislove, and Krishna P. Gummadi. 2018. Treads: Transparency-enhancing Ads. In *Workshop on Hot Topics in Networks*. Redmond, Washington, USA.
- [76] Giridhari Venkatadri, Piotr Sapiezyński, Elissa M. Redmiles, Alan Mislove, Oana Goga, Michelle Mazurek, and Krishna P. Gummadi. 2019. Auditing Offline Data Brokers Via Facebook's Advertising Platform. In *International World Wide Web Conference*. San Francisco, California, USA.
- [77] What it means when your ad is pending review [n.d.]. What It Means When Your Ad Is Pending Review. <https://www.facebook.com/business/help/204798856225114>.
- [78] Craig E. Wills and Can Tatar. 2012. Understanding What They Do With What They Know. In *Workshop on Privacy in the Electronic Society*.

APPENDIX

Multiple hypotheses testing. In the experiment described in the main paper we ran ads for 11 different job postings, each with five variations of the accompanying image. Here, we confirm that the apparent differences are not an effect of testing multiple hypotheses. We do so by aggregating the five variants for each ad and comparing the fraction of men and the estimated fraction of white users between each for pairs of jobs. This results in 55 tests, so rather than using the Z value corresponding to $p_{val} = 0.01$, we use the Bonferroni correction [14], a statistical technique used to address the problem of making multiple comparisons. In Figure 10 we show that the majority of comparisons remain statistically significant, each at the Z value corresponding to corrected $p_{val} = \frac{0.01}{55} \approx 0.0002$.

Received April 2019; revised June 2019; accepted August 2019

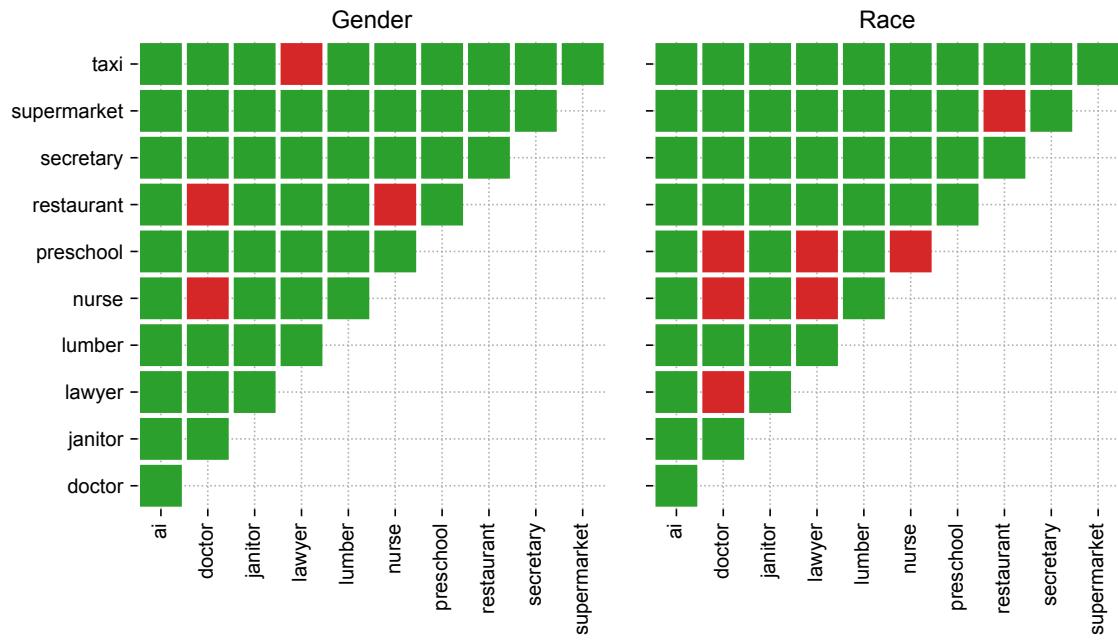


Fig. 10. The demographic differences in ad delivery both in terms of gender and race are statistically significant after introducing Bonferroni correction with N tests of 55. Green squares mark statistically significant differences, red squares indicate insignificant differences.

Weeks 12, 13: Interpretability

THE INTUITIVE APPEAL OF EXPLAINABLE MACHINES

Andrew D. Selbst* & Solon Barocas**

Algorithmic decision-making has become synonymous with inexplicable decision-making, but what makes algorithms so difficult to explain? This Article examines what sets machine learning apart from other ways of developing rules for decision-making and the problem these properties pose for explanation. We show that machine learning models can be both inscrutable and nonintuitive and that these are related, but distinct, properties.

Calls for explanation have treated these problems as one and the same, but disentangling the two reveals that they demand very different responses. Dealing with inscrutability requires providing a sensible description of the rules; addressing nonintuitiveness requires providing a satisfying explanation for why the rules are what they are. Existing laws like the Fair Credit Reporting Act (FCRA), the Equal Credit Opportunity Act (ECOA),

* Postdoctoral Scholar, Data & Society Research Institute; Visiting Fellow, Yale Information Society Project. Selbst is grateful for the support of the National Science Foundation under grant IIS 1633400.

** Assistant Professor, Cornell University, Department of Information Science. For helpful comments and insights on earlier drafts, the authors would like to thank Jack Balkin, Rabia Belt, danah boyd, Kiel Brennan-Marquez, Albert Chang, Danielle Citron, Julie Cohen, Lilian Edwards, Sorelle Freidler, Giles Hooker, Margaret Hu, Karen Levy, Margot Kaminski, Rónán Kennedy, Been Kim, Jon Kleinberg, Brian Kreiswirth, Chandler May, Brent Mittelstadt, Deidre Mulligan, David Lehr, Paul Ohm, Helen Nissenbaum, Frank Pasquale, Nicholson Price, Manish Raghavan, Aaron Rieke, David Robinson, Ira Rubinstein, Matthew Salganik, Katherine Strandburg, Sandra Wachter, Hanna Wallach, Cody Marie Wild, Natalie Williams, Jennifer Wortman Vaughan, Michael Veale, Suresh Venkatasubramanian, and participants at the following conferences and workshops: *NYU Innovation Colloquium*, NYU School of Law, February 2017; *We Robot*, Yale Law School, March 2017; *Big Data Ethics Colloquium*, The Wharton School, Philadelphia, PA, April 2017; *NYU Algorithms and Explanations Conference*, NYU School of Law, April 2017; *TILTing Perspectives*, Tilburg University, the Netherlands, May 2017; *Privacy Law Scholars' Conference*, Berkeley, CA, June 2017; *Summer Faculty Workshop*, Georgetown University Law Center, June 2017; *Explainable and Accountable Algorithms Workshop*, Alan Turing Institute, UK, January 2018. Special thanks to Chandler May for graphics that sadly did not make it into the final draft, and to the editors of the *Fordham Law Review* for their excellent and professional work getting this Article ready for publication. This Article is available for reuse under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0), <http://creativecommons.org/licenses/by-sa/4.0/>. The required attribution notice under the license must include the Article's full citation information, e.g., Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085 (2018).

and the General Data Protection Regulation (GDPR), as well as techniques within machine learning, are focused almost entirely on the problem of inscrutability. While such techniques could allow a machine learning system to comply with existing law, doing so may not help if the goal is to assess whether the basis for decision-making is normatively defensible.

In most cases, intuition serves as the unacknowledged bridge between a descriptive account and a normative evaluation. But because machine learning is often valued for its ability to uncover statistical relationships that defy intuition, relying on intuition is not a satisfying approach. This Article thus argues for other mechanisms for normative evaluation. To know why the rules are what they are, one must seek explanations of the process behind a model's development, not just explanations of the model itself.

INTRODUCTION.....	1087
I. INSCRUTABLE AND NONINTUITIVE	1089
A. <i>Secret</i>	1091
B. <i>Requiring Specialized Knowledge</i>	1093
C. <i>Inscrutable</i>	1094
D. <i>Nonintuitive</i>	1096
II. LEGAL AND TECHNICAL APPROACHES TO INSCRUTABILITY.....	1099
A. <i>Legal Requirements for Explanation</i>	1099
1. FCRA, ECOA, and Regulation B	1100
2. GDPR.....	1106
B. <i>Interpretability in Machine Learning</i>	1109
1. Purposefully Building Interpretable Models	1110
2. Post Hoc Methods	1113
3. Interactive Approaches	1115
III. FROM EXPLANATION TO INTUITION	1117
A. <i>The Value of Opening the Black Box</i>	1118
1. Explanation as Inherent Good.....	1118
2. Explanation as Enabling Action.....	1120
3. Explanation as Exposing a Basis for Evaluation.....	1122
B. <i>Evaluating Intuition</i>	1126
IV. DOCUMENTATION AS EXPLANATION	1129
A. <i>The Information Needed to Evaluate Models</i>	1130
B. <i>Providing the Necessary Information</i>	1133
CONCLUSION	1138

There can be no total understanding and no absolutely reliable test of understanding.

—Joseph Weizenbaum, “Contextual Understanding by Computers”¹

INTRODUCTION

Algorithms increasingly inform consequential decisions about our lives, with only minimal input from the people they affect and little to no explanation as to how they work.² This worries people, and rightly so. The results of these algorithms can be unnerving,³ unfair,⁴ unsafe,⁵ unpredictable,⁶ and unaccountable.⁷ How can decision makers who use algorithms be held to account for their results?

It is perhaps unsurprising that, faced with a world increasingly dominated by automated decision-making, advocates, policymakers, and legal scholars would call for machines that can explain themselves.⁸ People have a natural

1. 10 COMM. ACM 474, 476 (1967). In the 1960s, the project of artificial intelligence (AI) was largely to mimic human intelligence. Weizenbaum was therefore actually arguing that computers will never fully understand humans. The purpose of AI research has changed drastically today, but there is a nice symmetry in the point that humans will never have total understanding of computers.

2. Aaron M. Bornstein, *Is Artificial Intelligence Permanently Inscrutable?*, NAUTILUS (Sept. 1, 2016), <http://nautil.us/issue/40/learning/is-artificial-intelligence-permanently-inscrutable> [<http://perma.cc/RW3E-5CPV>]; Will Knight, *The Dark Secret at the Heart of AI*, MIT TECH. REV. (Apr. 11, 2017), <https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/> [<http://perma.cc/7VYF-5XR7>]; Cliff Kuang, *Can A.I. Be Taught to Explain Itself?*, N.Y. TIMES (Nov. 21, 2017), <https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html> [<http://perma.cc/3CYF-QTVC>].

3. See, e.g., Omer Tene & Jules Polonetsky, *A Theory of Creepy: Technology, Privacy, and Shifting Social Norms*, 16 YALE J.L. & TECH. 59, 65–66 (2013); Sara M. Watson, *Data Doppelgängers and the Uncanny Valley of Personalization*, ATLANTIC (June 16, 2014), <https://www.theatlantic.com/technology/archive/2014/06/data-doppelgangers-and-the-uncanny-valley-of-personalization/372780/> [<http://perma.cc/7J3X-NK3C>].

4. See, e.g., Solon Barocas & Andrew D. Selbst, *Big Data’s Disparate Impact*, 104 CALIF. L. REV. 671, 677–92 (2016); Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857, 883–89 (2017); Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GA. L. REV. 109, 126–39 (2017).

5. See, e.g., David Lazer et al., *The Parable of Google Flu: Traps in Big Data Analysis*, 343 SCIENCE 1203, 1203 (2014); Jennings Brown, *IBM Watson Reportedly Recommended Cancer Treatments That Were ‘Unsafe and Incorrect,’* GIZMODO (July 25, 2018, 3:00 PM), <https://gizmodo.com/ibm-watson-reportedly-recommended-cancer-treatments-tha-1827868882> [<http://perma.cc/E4RZ-NVZU>].

6. See, e.g., Curtis E. A. Karnow, *The Application of Traditional Tort Theory to Embodied Machine Intelligence*, in ROBOT LAW 51, 57–58 (Ryan Calo, A. Michael Froomkin & Ian Kerr eds., 2016) (discussing unpredictability in autonomous systems); Jamie Condliffe, *Algorithms Probably Caused a Flash Crash of the British Pound*, MIT TECH. REV. (Oct. 7, 2016), <https://www.technologyreview.com/s/602586/algorithms-probably-caused-a-flash-crash-of-the-british-pound/> [<https://perma.cc/K9FM-6SJE>].

7. See, e.g., Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 18–27 (2014); Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 636–37 (2017).

8. See *infra* Part II.

feel for explanation. We know how to offer explanations and can often agree when one is good, bad, in-between, on point, or off topic. Lawyers use explanation as their primary tradecraft: judges write opinions, administrators respond to comments, litigators write briefs, and everyone writes memos. Explanations are the difference between a system that vests authority in lawful process and one that vests it in an unaccountable person.⁹

Although we comfortably use explanations, asking someone to define the concept will often generate a blank look in response. Analytically, explanation is infinitely variable, and there can be many valid explanations for a given phenomenon or decision. Thus far, in both law and machine learning, the scholarly discourse around explanation has primarily revolved around two questions: Which kinds of explanations are most useful, and which are technically available?¹⁰ Yet, these are the wrong questions or, at least, the wrong stopping points.

Explanations of technical systems are necessary but not sufficient to achieve law and policy goals, most of which are concerned not with explanation for its own sake, but with ensuring that there is a way to evaluate the basis of decision-making against broader normative constraints such as antidiscrimination or due process. It is therefore important to ask how exactly people engage with those machine explanations in order to connect them to the normative questions of interest to law.

This Article argues that scholars and advocates who seek to use explanation to enable justification of machine learning models are relying on intuition to connect the explanation to normative concerns. Intuition is both powerful and dangerous. While this mode of justifying decision-making remains important, we must understand the benefits and weaknesses of connecting machine explanation to intuitions. Remedyng the limitations of intuition requires considering alternatives, which include institutional processes, documentation, and access to those documents.

This Article proceeds in four parts. Part I examines the various anxieties surrounding the use of automated decision-making. After discussing secrecy, lack of transparency, and lack of technical expertise, Part I argues that the two distinct, but similar, concepts that truly set machine learning decision-making apart are inscrutability and nonintuitiveness.

Part II examines laws and machine learning techniques designed specifically to address the problem of inscrutable decisions. On the legal side, Part II.A discusses the “adverse action notices” required by federal credit laws¹¹ and the informational requirements of the European Union’s General Data Protection Regulation (GDPR).¹² On the technical side, Part

9. See Frederick Schauer, *Giving Reasons*, 47 STAN. L. REV. 633, 636–37 (1995).

10. See *infra* Part III.

11. This Article will focus on the Fair Credit Reporting Act, 15 U.S.C. §§ 1681–1681x, and Equal Credit Opportunity Act, 15 U.S.C. §§ 1691–1691f.

12. Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC, 2016 O.J. (L 119) 1 (EU) [hereinafter GDPR].

II.B discusses various techniques used by computer scientists to make machine learning models interpretable, including designing for simplicity, approximating complex models in simpler form, extracting the most important factors in a particular decision, and allowing some degree of interaction with the models to see how changes in inputs affect outputs. These techniques can be useful in meeting the requirements of the law, but such explanations, even when they comply with the law, may be of limited practical utility.

Part III builds the connection between explanation and intuition before evaluating the merits of an intuition-centered approach to justification. It canvasses reasons besides justification that one might want explainable machines—dignity or autonomy on the one hand and consumer or data-subject education on the other—before concluding that neither is adequate to fully address the concerns with automated decision-making. Interrogating the assumptions behind a third reason—that explanation will reveal problems with the basis for decision-making—demonstrates the reliance on intuition. The remainder of Part III examines the value and limitations of intuition. With respect to machine learning in particular, although intuition can root out obviously good or bad cases, it cannot capture the cases that give machine learning its greatest value: true patterns that exceed human imagination. These cases are not obviously right or wrong, but simply strange.

Part IV aims to provide another way. Once outside the black box, all that is left is to question the process surrounding its development and use. There are large parts of the process of machine learning that do not show up in a model but can contextualize its operation, such as paths considered but not taken and the constraints that influence these choices. Where intuition is insufficient to determine whether the model’s rules are reasonable or rest on valid relationships, justification can sometimes be achieved by demonstrating and documenting due care and thoughtfulness.

I. INSCRUTABLE AND NONINTUITIVE

Scholarly and policy debates about regulating a world controlled by algorithms have been mired in difficult questions about how to observe, access, audit, or understand those algorithms.¹³ The difficulty has been attributed to a diverse set of problems, specifically that algorithms are

13. See, e.g., Rob Kitchin, *Thinking Critically About and Researching Algorithms*, 20 INFO. COMM. & SOC’Y 14 (2017) (evaluating methods of researching algorithms); Malte Ziewitz, *Governing Algorithms: Myth, Mess, and Methods*, 41 SCI. TECH. & HUM. VALUES 3 (2016); Solon Barocas, Sophie Hood & Malte Ziewitz, Governing Algorithms: A Provocation Piece (Mar. 29, 2013) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2245322 [<https://perma.cc/DB7Z-C9A6>]; Nick Seaver, Knowing Algorithms (Feb. 2014) (unpublished manuscript), <https://static1.squarespace.com/static/55eb004ee4b0518639d59d9b/t/55eccc1bfe4b030b2e8302e1e/1441587647177/seaverMiT8.pdf> [<https://perma.cc/7HG3-74U3>].

“secret”¹⁴ and “opaque”¹⁵ “black boxes”¹⁶ that are rarely, if ever, made “transparent”;¹⁷ that they operate on the basis of correlation rather than “causality”¹⁸ and produce “predictions”¹⁹ rather than “explanations”;²⁰ that their behavior may lack “intelligibility”²¹ and “foreseeability”²² and that they challenge established ways of being “informed”²³ or “knowing.”²⁴ These terms are frequently used interchangeably or assumed to have overlapping meanings. For example, opacity is often seen as a synonym for secrecy,²⁵ an antonym for transparency,²⁶ and, by implication, an impediment to understanding.²⁷ Yet the perceived equivalence of these terms has obscured important differences between distinct problems that frustrate attempts at regulating algorithms—problems that must be disentangled before the question of regulation can even be addressed.

This Part argues that many of these challenges are not unique to algorithms or machine learning. We seek here to parse the problems raised by machine learning models more precisely and argue that they have little to do with the fact that their very existence may be unknown, that their inner workings may be opaque, or that an understanding of their operations may require specialized knowledge. What sets machine learning models apart from other decision-making mechanisms are their *inscrutability* and *nonintuitiveness*.

14. See, e.g., Frank Pasquale, *Restoring Transparency to Automated Authority*, 9 J. ON TELECOMM. & HIGH TECH. L. 235, 236–37 (2011) (recounting the origins of using trade-secret protections for algorithms); Brenda Reddix-Smalls, *Credit Scoring and Trade Secrecy: An Algorithmic Quagmire or How the Lack of Transparency in Complex Financial Models Scuttled the Finance Market*, 12 U.C. DAVIS BUS. L.J. 87, 88–90 (2011) (discussing the use of trade-secret protections for algorithms, which result in lack of transparency concerning algorithmic decision-making).

15. Jenna Burrell, *How the Machine “Thinks”: Understanding Opacity in Machine Learning Algorithms*, BIG DATA & SOC’Y, Jan.–June 2016, at 1, 3–5; Roger Allan Ford & W. Nicholson Price II, *Privacy and Accountability in Black-Box Medicine*, 23 MICH. TELECOMM. & TECH. L. REV. 1, 11–12 (2016); Tal Zarsky, *The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making*, 41 SCI. TECH. & HUM. VALUES 118, 129 (2016).

16. See, e.g., FRANK PASQUALE, THE BLACK BOX SOCIETY 8 (2015).

17. See, e.g., Citron & Pasquale, *supra* note 7, at 27; Tal Z. Zarsky, *Transparent Predictions*, 2013 U. ILL. L. REV. 1503, 1506.

18. See, e.g., Kim, *supra* note 4, at 875.

19. Kiel Brennan-Marquez, *“Plausible Cause”: Explanatory Standards in the Age of Powerful Machines*, 70 VAND. L. REV. 1249, 1267–68 (2017).

20. See, e.g., Bryce Goodman & Seth Flaxman, *European Union Regulations on Algorithmic Decision-Making and a “Right to Explanation,”* 38 AI MAG., Fall 2017, at 50, 55.

21. See, e.g., Brennan-Marquez, *supra* note 19, at 1253.

22. See, e.g., Karnow, *supra* note 6, at 52.

23. See, e.g., Sandra Wachter, Brent Mittelstadt & Luciano Floridi, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, 7 INT’L DATA PRIVACY L. 76, 89–90 (2017).

24. Mike Ananny & Kate Crawford, *Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability*, 20 NEW MEDIA & SOC’Y 973, 974–77 (2018).

25. See, e.g., Burrell, *supra* note 15, at 3–4.

26. See, e.g., Ford & Price, *supra* note 15, at 12; Zarsky, *supra* note 15, at 124.

27. See, e.g., Burrell, *supra* note 15, at 4–5.

We adapt and extend a taxonomy first proposed by Jenna Burrell,²⁸ where our primary purpose is to emphasize these last two properties and clear up confusion.²⁹ Inscrutability and nonintuitiveness have been conflated in the past: where the property of inscrutability suggests that models available for direct inspection may defy understanding, nonintuitiveness suggests that even where models are understandable, they may rest on apparent statistical relationships that defy intuition.³⁰

A. Secret

The first common critique of algorithmic decision-making is secrecy. Secrecy captures two related, but distinct, concerns: (1) secrecy of the model's existence and (2) secrecy of its operation.

The first concern is as old as the original Code of Fair Information Practices (FIPs), the conceptual basis for the majority of privacy laws:³¹ “There must be no personal-data record-keeping systems whose very existence is secret.”³² This principle underlies more recent calls to “end secret profiling” involving algorithms and machine learning, where secrecy is understood as a purposeful attempt to maintain ignorance of the very fact of profiling.³³

While such worries are particularly pronounced when the government engages in algorithmic decision-making,³⁴ similar objections arise in the commercial sector, where there are a remarkable number of scoring systems

28. See generally *id.*

29. Our parsing of the issues is similar to the taxonomy proposed by Ed Felten in a short blog post on *Freedom to Tinker*. Ed Felten, *What Does It Mean to Ask for an “Explainable” Algorithm?*, FREEDOM TO TINKER (May 31, 2017), <https://freedom-to-tinker.com/2017/05/31/what-does-it-mean-to-ask-for-an-explainable-algorithm/> [https://perma.cc/QF7B-RTC6].

30. We intentionally use the term “nonintuitive” rather than “unintuitive” or “counterintuitive.” In our view, “unintuitive” implies a result that would not be expected but is easily understood once explained, and “counterintuitive” suggests a phenomenon that is opposite one’s expectations. Instead, we intend to refer to a phenomenon about which intuitive reasoning is not possible.

31. WOODROW HARTZOG, PRIVACY’S BLUEPRINT: THE BATTLE TO CONTROL THE DESIGN OF NEW TECHNOLOGIES 56 (2018); Robert Gellman, Fair Information Practices: A Basic History 3 (Apr. 10, 2017) (unpublished manuscript), <https://bobgellman.com/rg-docs/rg-FIPshistory.pdf> [https://perma.cc/CQ9E-HK9A] (discussing the history of the FIPs).

32. SEC’Y’S ADVISORY COMM. ON AUTOMATED PERS. DATA SYS., U.S. DEP’T OF HEALTH, EDUC. & WELFARE, RECORDS, COMPUTERS, AND THE RIGHTS OF CITIZENS 41 (1973), <https://www.justice.gov/opcl/docs/rec-com-rights.pdf> [https://perma.cc/8TG8-FBL9]. In fact, the newly effective GDPR requires, among other things, disclosure of the “existence” of an automated decision-making tool. See *infra* note 142 and accompanying text.

33. *Algorithmic Transparency: End Secret Profiling*, ELECTRONIC PRIVACY INFO. CTR., <https://epic.org/algorithmic-transparency/> [https://perma.cc/ZW4W-HKTM] (last visited Nov. 15, 2018); see also Margaret Hu, *Big Data Blacklisting*, 67 FLA. L. REV. 1735, 1745–46 (2015).

34. See Ira S. Rubinstein, Ronald D. Lee & Paul M. Schwartz, *Data Mining and Internet Profiling: Emerging Regulatory and Technological Approaches*, 75 U. CHI. L. REV. 261, 262–70 (2008); Tal Z. Zarsky, *Governmental Data Mining and Its Alternatives*, 116 PENN ST. L. REV. 285, 295–97 (2011).

of which consumers are simply unaware.³⁵ In many cases, this ignorance exists because the companies engaged in such scoring are serving other businesses rather than consumers.³⁶ But the fact that more recent forms of hidden decision-making involve algorithms or machine learning does not change the fundamental secrecy objection—that affected parties are not aware of the existence of the decision-making process.³⁷

The second secrecy concern arises where the existence of a decision-making process is known, but its actual operation is not. Affected parties might be aware that they are subject to such decision-making but have limited or no knowledge of how the decision-making process works.³⁸ Among the many terms used to describe this situation, “opacity” seems most apt, as there is enough visibility to see that the model exists but not enough to discern any of its details.

While this is perhaps the most frequent critique of algorithms and machine learning—that their inner workings remain undisclosed or inaccessible³⁹—it, too, has little to do with the technology specifically. It is an objection to being subject to a decision where the basis of decision-making remains secret, which is a situation that can easily occur without algorithms or machine learning.⁴⁰

There are sometimes valid reasons for companies to withhold details about a decision-making process. Where a decision-making process holds financial and competitive value and where its discovery entails significant investment or ingenuity, firms may claim protection for its discovery as a trade secret.⁴¹ Trade-secret protection applies only when firms purposefully restrict disclosure of proprietary methods,⁴² which creates incentives for firms to maintain secrecy around the basis for decision-making. If the use of algorithms or machine learning uniquely increases up-front investment or competitive advantage, then the incentives to restrict access to the details of

35. See PAM DIXON & ROBERT GELLMAN, THE SCORING OF AMERICA: HOW SECRET CONSUMER SCORES THREATEN YOUR PRIVACY AND YOUR FUTURE 84 (2014), http://www.worldprivacyforum.org/wp-content/uploads/2014/04/WPF_Scoring_of_America_April2014_fs.pdf [<https://perma.cc/39RJ-97M6>].

36. See FED. TRADE COMM’N, DATA BROKERS: A CALL FOR TRANSPARENCY AND ACCOUNTABILITY i (2014), https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527_databrokerreport.pdf [<https://perma.cc/8HQY-6WVP>].

37. SEC’Y’S ADVISORY COMM. ON AUTOMATED PERS. DATA SYS., *supra* note 32, at 29 (discussing the lack of awareness of record keeping and use of personal data).

38. This could refer to secrecy around what data is considered or how it is used. See *infra* Part II.A for a discussion of these concerns with respect to the Fair Credit Reporting Act.

39. See, e.g., Robert Brauneis & Ellen P. Goodman, *Algorithmic Transparency for the Smart City*, 20 YALE J.L. & TECH. 103, 107–08 (2018); Citron & Pasquale, *supra* note 7, at 10–11. See generally PASQUALE, *supra* note 16.

40. See, e.g., Daniel J. Solove, *Privacy and Power: Computer Databases and Metaphors for Information Privacy*, 53 STAN. L. REV. 1393, 1407, 1410 (2001) (discussing the private database industry and corporate decision-making based on consumer data).

41. Brauneis & Goodman, *supra* note 39, at 153–60. See generally Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STAN. L. REV. 1343 (2018).

42. Pasquale, *supra* note 14, at 237.

the decision-making process might be understood as peculiar to algorithms or machine learning. But if other attempts to develop decision-making processes without algorithms or machine learning involve similar costs and competitive advantage, then there is nothing special about the relationship between these technologies, trade secrets, and resistance to disclosure.⁴³

Firms may also reject requests for further details about the basis for decision-making if they anticipate that such details may enable strategic manipulation, or “gaming,” of the inputs to the decision-making process.⁴⁴ If the costs of manipulating one’s characteristics or behavior are lower than the expected benefits, rational actors would have good incentive to do so.⁴⁵ Yet these dynamics, too, apply outside algorithms and machine learning; in the face of some fixed decision procedure, people will find ways to engage in strategic manipulation. The question is whether decision procedures developed with machine learning are easier or harder to game than those developed using other methods—this is not a question that can be answered in general.

B. Requiring Specialized Knowledge

One common approach to ensuring accountability for software-reliant decision-making is to require the disclosure of the underlying source code.⁴⁶ While such disclosure might seem helpful in figuring out how automated decisions are rendered, the ability to make sense of the disclosed source code will depend on one’s level of technical literacy. Some minimal degree of training in computer programming is necessary to read code, although even that might not be enough.⁴⁷ The problem, then, is greater than disclosure; in

43. See, e.g., David S. Levine, *Secrecy and Unaccountability: Trade Secrets in Our Public Infrastructure*, 59 FLA. L. REV. 135, 139 (2007) (describing the growing application of trade secrecy in various technologies used in public infrastructure).

44. Jane Bambauer & Tal Z. Zarsky, *The Algorithm Game*, 94 NOTRE DAME L. REV. (forthcoming 2018) (manuscript at 10), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3135949 [<http://perma.cc/N62U-3UUK>].

45. Whether such manipulation is even possible will vary from case to case, depending on the degree to which the decision considers immutable characteristics and nonvolitional behavior. At the same time, it is unclear how easily one could even change the appearance of one’s characteristics without genuinely changing those characteristics in the process. Altering behavior to game the system might involve adjustments that actually change a person’s likelihood of having the sought-after quality or experiencing the event that such behavior is meant to predict. To the extent that “gaming” is a term used to describe validating rather than defeating the objectives of a decision system, this outcome should probably not be considered “gaming” at all. See Bambauer & Zarsky, *supra* note 44.

46. Deven R. Desai & Joshua A. Kroll, *Trust but Verify: A Guide to Algorithms and the Law*, 31 HARV. J.L. & TECH. 1, 10 (2017); Kroll et al., *supra* note 7, at 647–50; *Algorithmic Transparency: End Secret Profiling*, *supra* note 33. Draft legislation in New York City also specifically focused on this issue, but the eventual bill convened a more general task force to consider different approaches. See Jim Dwyer, *Showing the Algorithms Behind New York City Services*, N.Y. TIMES (Aug. 24, 2017), <https://www.nytimes.com/2017/08/24/nyregion/showing-the-algorithms-behind-new-york-city-services.html> [<https://perma.cc/38V5-P3EE>].

47. Desai & Kroll, *supra* note 46, at 5 (“[F]undamental limitations on the analysis of software meaningfully limit the interpretability of even full disclosures of software source code.”); Kroll et al., *supra* note 7, at 647.

the absence of the specialized knowledge required to understand source code, disclosure may offer little value to affected parties and regulators.

As Mike Ananny and Kate Crawford have observed, “Transparency concerns are commonly driven by a certain chain of logic: observation produces insights which create the knowledge required to govern and hold systems accountable.”⁴⁸ The process of moving from observation to knowledge to accountability cannot be assumed; in many cases, the ability to leverage observations for accountability requires *preexisting* knowledge that allows observers to appreciate the significance of a disclosure.⁴⁹ Transparency of systems of decision-making is important, but incomplete.⁵⁰ But while cultivating the knowledge necessary to read source code requires time and effort, the problem of expertise—like the problem of secrecy—is not unique to algorithms.

C. Inscrutable

Rather than programming computers by hand with explicit rules, machine learning relies on pattern-recognition algorithms and a large set of examples to uncover relationships in the data that might serve as a reliable basis for decision-making.⁵¹ The power of machine learning lies not only in its ability to relieve programmers of the difficult task of producing explicit instructions for computers, but in its capacity to learn subtle relationships in data that humans might overlook or cannot recognize. This power can render the models developed with machine learning exceedingly complex and, therefore, impossible for a human to parse.

We define this difficulty as “inscrutability”—a situation in which the rules that govern decision-making are so complex, numerous, and interdependent that they defy practical inspection and resist comprehension. While there is a long history to such concerns, evidenced most obviously by the term “byzantine,” the complexity of rules that result from machine learning can far exceed those of the most elaborate bureaucracy.⁵² The challenge in such circumstances is not a lack of awareness, disclosure, or expertise, but the sheer scope and sophistication of the model.⁵³

Intuitively, complexity would seem to depend on the number of rules encoded by a model, the length of a rule (i.e., the number of factors that figure into the rule), and the logical operations involved in the rule. These properties, however, can be specified more precisely. Four mathematical

48. Ananny & Crawford, *supra* note 24, at 974.

49. Burrell, *supra* note 15, at 4.

50. See Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1254–55 (2008); Kroll et al., *supra* note 7, at 639, 657–60.

51. See David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 655 (2017).

52. *Byzantine*, MERRIAM-WEBSTER, <http://www.merriam-webster.com/dictionary/byzantine> [https://perma.cc/97CM-KNT2] (last visited Nov. 15, 2018) (defining the term as “intricately involved”).

53. Burrell, *supra* note 15, at 4–5.

properties related to model complexity are linearity, monotonicity, continuity, and dimensionality.

A linear model is one in which there is a steady change in the value of the output as the value of the input changes.⁵⁴ Linear models tend to be easier for humans to understand and interpret because the relationship between variables is stable and lends itself to straightforward extrapolation.⁵⁵ In contrast, the behavior of nonlinear models can be far more difficult to predict, even when they involve simple mathematical operations like exponential growth.⁵⁶

A monotonic relationship between variables is a relationship that is either always positive or always negative.⁵⁷ That is, for every change in input value, the direction of the corresponding change in output value will remain consistent, whether an increase or decrease.⁵⁸ Monotonicity aids interpretability because it too permits extrapolation and guarantees that the value of the output only moves in one direction.⁵⁹ If, however, the value of the output goes up and down haphazardly as the value of the input moves steadily upward, the relationship between variables can be difficult to grasp or predict.

Discontinuous models include relationships where changes in the value of one variable do not lead to a smooth change in the associated value of another.⁶⁰ Discontinuities can render models far less intuitive because they make it impossible to think in terms of incremental change. A small change in input may typically lead to small changes in outputs, except for occasional and seemingly arbitrary large jumps.⁶¹

The dimensionality of a model is the number of variables it considers.⁶² Two-dimensional models are easy to understand because they can be visualized graphically with a standard plot (with the familiar x and y axes).⁶³ Three-dimensional models also lend themselves to effective visualization (by adding a z axis), but humans have no way to visualize models with more than three dimensions.⁶⁴ While people can grasp relationships between multiple

54. Mathematically, this means that the function is described by a constant slope, which can be represented by a line. Yin Lou et al., *Intelligible Models for Classification and Regression*, in PROCEEDINGS OF THE 18TH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING 150, 150 (2012).

55. See *id.* at 151.

56. Cf. DEMI, ONE GRAIN OF RICE: A MATHEMATICAL FOLKTALE (1997).

57. See *Monotonicity Function*, CONCISE OXFORD DICTIONARY OF MATHEMATICS (3d ed. 2014).

58. See *id.*

59. See *id.*

60. See *Continuous Function*, CONCISE OXFORD DICTIONARY OF MATHEMATICS (3d ed. 2014) (noting that a continuous function does not suddenly jump at a given point or take widely differing values arbitrarily close to that point).

61. See *Discontinuity*, CONCISE OXFORD DICTIONARY OF MATHEMATICS (3d ed. 2014).

62. See *Dimension (Dimensionality)*, A DICTIONARY OF COMPUTER SCIENCE (7th ed. 2016).

63. See *Cartesian Plane*, CONCISE OXFORD DICTIONARY OF MATHEMATICS (3d ed. 2014).

64. See *Cartesian Space*, CONCISE OXFORD DICTIONARY OF MATHEMATICS (3d ed. 2014); *n-Dimensional Space*, CONCISE OXFORD DICTIONARY OF MATHEMATICS (3d ed. 2014).

variables without the aid of a graph, we will struggle to understand the full set of relationships that the model has uncovered as the number of dimensions grows. The more variables that the model includes, the more difficult it will be to keep all the interactions between variables in mind and thus predict how the model would behave given any particular input.⁶⁵

In describing how these properties of models might frustrate human understanding, we have relied on terms like intuition, extrapolation, and prediction. The same cognitive capacity underlies all three: mentally simulating how a model turns inputs into outputs.⁶⁶ As computer scientist Zachary Lipton explains, simulability—the ability to practically execute a model in one's mind—is an important form of understanding a model.⁶⁷ Such simulations can be either complete or partial. In the former, a person is able to turn any combination of inputs into the correct outputs, while in the latter, understanding might be limited to the relationships between a subset of input and output variables (i.e., how changes in certain inputs affect the output).

Simulation is a remarkably flat and functional definition of understanding, but it seems like a minimum requirement for any more elaborate definition.⁶⁸ This notion of understanding has nothing to say about *why* the model behaves the way it does; it is simply a way to account for the facility with which a person can play out how a model would behave under different circumstances. When models are too complex for humans to perform this task, they have reached the point of inscrutability.

D. Nonintuitive

A different line of criticism has developed that takes issue with disclosures that reveal some basis for decision-making that defies human intuition about the relevance of certain variables.⁶⁹ The problem in such cases is not

65. See Lehr & Ohm, *supra* note 51, at 700.

66. Zachary C. Lipton, *The Mythos of Model Interpretability*, in PROCEEDINGS OF THE 2016 ICML WORKSHOP ON HUMAN INTERPRETABILITY IN MACHINE LEARNING 96, 98 (2016).

67. *Id.*

68. While we limit our discussion to simulability, inscrutability is really a broader concept. In particular, models might be difficult to understand if they consider features or perform operations that do not have some ready semantic meaning. Burrell, *supra* note 15, at 10. For example, a deep-learning algorithm can learn on its own which features in an image are characteristic of different objects (the standard example being cats). Bornstein, *supra* note 2. Part III.A.3, *infra*, returns to one such example that involves distinguishing between wolves and huskies. See *infra* notes 246–47 and accompanying text. An algorithm will usually learn to detect edges that differentiate an object from its background, but it might also engineer features on its own that have no equivalent in human cognition and therefore defy description. See Lipton, *supra* note 66, at 98 (discussing decomposability). This aspect of inscrutability, however, is of slightly less concern for this Article. Most methods that are common in the kinds of applications that apportion important opportunities (e.g., credit) involve features that have been handcrafted by experts in the domain (e.g., length of employment) and accordingly will usually not face this problem. See *infra* note 120 and accompanying text.

69. Deborah Gage, *Big Data Uncovers Some Weird Correlations*, WALL ST. J. (Mar. 23, 2014, 4:36 PM), <https://www.wsj.com/articles/big-data-helps-companies-find-some-surprising-correlations-1395168255> [https://perma.cc/8KYB-LP9W]; Quentin Hardy,

inscrutability, but an inability to weave a sensible story to account for the statistical relationships in the model.⁷⁰ Although the statistical relationship that serves as the basis for decision-making might be readily identifiable, that relationship may defy intuitive expectations about the relevance of certain criteria to the decision.⁷¹ As Paul Ohm explains:

We are embarking on the age of the impossible-to-understand reason, when marketers will know which style of shoe to advertise to us online based on the type of fruit we most often eat for breakfast, or when the police know which group in a public park is most likely to do mischief based on the way they do their hair or how far from one another they walk.⁷²

Even though it is clear which statistical relationships serve as the basis for decision-making in this case, why such statistical relationships exist is mystifying. This is a crucial and consistent point of confusion. The demand for intuitive relationships is not the demand for disclosure or accessible explanations; it is a demand that decision-making rely on reasoning that comports with intuitive understanding of the phenomenon in question. In social science, similar expectations are referred to as “face validity”—the subjective sense that some measure is credible because it squares with our existing understanding of the phenomenon.⁷³ While such demands are not unique to algorithms and machine learning, the fact that such computational tools are designed to uncover relationships that defy human intuition explains why the problem will be particularly pronounced in these cases.

Critics have pinned this problem on the use of “[m]ere correlation”⁷⁴ in machine learning, which frees it to uncover reliable, if incidental, relationships in the data that can then serve as the basis for consequential decision-making.⁷⁵ Despite being framed as an indictment of correlational analysis, however, it is really an objection to decision-making that rests on particular correlations that defy familiar causal stories⁷⁶—even though these stories may be incorrect.⁷⁷ This has led to the mistaken belief that forcing

Bizarre Insights from Big Data, N.Y. TIMES: BITS (Mar. 28, 2012, 8:17 PM), <https://bits.blogs.nytimes.com/2012/03/28/bizarre-insights-from-big-data/> [https://perma.cc/GKW2-KN8T].

70. See Brennan-Marquez, *supra* note 19, at 1280–97.

71. See Paul Ohm, *The Fourth Amendment in a World Without Privacy*, 81 Miss. L.J. 1309, 1318 (2012).

72. *Id.*

73. See generally Ronald R. Holden, *Face Validity*, in 2 CORSINI ENCYCLOPEDIA OF PSYCHOLOGY 637 (Irving B. Weiner & W. Edward Craighead eds., 4th ed. 2010).

74. Kim, *supra* note 4, at 875, 883.

75. *Id.*; see also James Grimmelmann & Daniel Westreich, *Incomprehensible Discrimination*, 7 CALIF. L. REV. ONLINE 164, 173 (2016).

76. See Brennan-Marquez, *supra* note 19, at 1280–97.

77. See DANIEL KAHNEMAN, THINKING FAST AND SLOW 199–200 (2011) (discussing the “narrative fallacy”); *id.* at 224 (“Several studies have shown that human decision makers are inferior to a prediction formula even when they are given the score suggested by the formula! They feel that they can overrule the formula because they have additional information about the case, but they are wrong more often than not.”).

decision-making to rest on causal mechanisms rather than mere correlations will ensure intuitive models.⁷⁸

Causal relationships can be exceedingly complex and nonintuitive, especially when dealing with human behavior.⁷⁹ Indeed, causal relationships uncovered through careful experimentation can be as elaborate and unexpected as the kinds of correlations uncovered in historical data with machine learning.⁸⁰ If we consider all the different factors that cause a person to take an action—mood, amount of sleep, food consumption, rational choice, and many other things—it quickly becomes clear that causality is not particularly straightforward.⁸¹ The only advantage of models that rely on causal mechanisms in such cases would be the reliability of their predictions (because the models would be deterministic rather than probabilistic), not the ability to interrogate whether the identified causal relationships comport with human intuitions and values. Given that much of the interest in causality stems from an unwillingness to simply defer to predictive accuracy as a justification for models, improved reliability will not be a satisfying answer.

* * *

The demand for intuitive relationships reflects a desire to ensure that there is a way to assess whether the basis of decision-making is sound, as a matter of validity and as a normative matter. We want to be able to do more than simply simulate a model; we want to be able to *evaluate* it. One way to ensure this possibility is to force a model to rely exclusively on features that bear a manifest relationship to the outcome of interest, on the belief that well-justified decisions are those that rest on relationships that conform to familiar and permissible patterns.

Achieving this type of intuitiveness requires addressing inscrutability as a starting point. An understandable model is necessary because there can be nothing intuitive about a model that resists all interrogation. But addressing inscrutability is not sufficient. A simple, straightforward model might still defy intuition if it has not been constrained to only use variables with an intuitive relationship to the outcome.⁸²

78. These critiques also presume that causal mechanisms that exhaustively account for the outcomes of interest actually exist (e.g., performance on the job, default, etc.), yet certain phenomena might not be so deterministic; extrinsic random factors may account for some of the difference in the outcomes of interest. Jake M. Hofman, Amit Sharma & Duncan J. Watts, *Prediction and Explanation in Social Systems*, 355 SCIENCE 486, 488 (2017).

79. *Id.*

80. *See id.*

81. Attempts to model causation require limiting the features considered as potential causes because, to a certain extent, almost any preceding event could conceivably be causally related to the later one. JUDEA PEARL, CAUSALITY: MODELS, REASONING AND INFERENCE 401–28 (2d ed. 2009).

82. *See, e.g.*, Jiaming Zeng, Berk Ustun & Cynthia Rudin, *Interpretable Classification Models for Recidivism Prediction*, 180 J. ROYAL STAT. SOC’Y 689 (2017). Note that in this and related work, the researchers limit themselves to features that are individually and intuitively related to the outcome of interest. *See id.* at 693–97. If these methods begin with features that do not have such a relationship, the model might be simple enough to inspect but too strange to square with intuition. *See infra* Part III.B.

Insisting on intuitive relationships is not the only way to make a model evaluable. To the extent that intuitiveness is taken to be an end in itself rather than a particular means to the end of ensuring sound decision-making, its proponents risk overlooking other, potentially more effective, ways to achieve the same goal. The remainder of this Article considers the different paths we might take to use explanations of machine learning models to regulate them.

II. LEGAL AND TECHNICAL APPROACHES TO INSCRUTABILITY

This moment is not the first time that law and computer science have attempted to address algorithmic decision-making with explanation requirements. Credit scoring has long been regulated, in part, by requiring “adverse action notices,” which explain adverse decisions to consumers.⁸³ In Europe, concern about automated decisions has been a neglected part of data protection law for more than two decades, with interest in them reinvigorated by the GDPR.⁸⁴ On the machine learning side, the subfield of “interpretability”—within which researchers have been attempting to find ways to understand complex models—is over thirty years old.⁸⁵

What seems to emerge from the law and computer science is a focus on two kinds of explanation. The first concerns accounting for outcomes—how particular inputs lead to a particular output. The second concerns the logic of decision-making—full or partial descriptions of the rules of the system. This Part reviews the legal and technical approaches to outcome and logic-based explanations.

A. Legal Requirements for Explanation

Though much of the current concern over inscrutable systems stems from the growing importance of machine learning, inscrutable systems predate this technique. As a result, regulations that require certain systems to explain themselves already exist. This section discusses two examples of legal systems and strategies that rely on different types of explanations: credit reporting statutes, which rely on outcome-based explanations, and the GDPR, which mandates logic-based explanations. Credit scoring predates machine learning, and is governed by two statutes: the Fair Credit Reporting Act (FCRA)⁸⁶ and the Equal Credit Opportunity Act (ECOA).⁸⁷ Statistical credit-scoring systems take information about consumers as inputs, give the

83. See *infra* notes 100–01 and accompanying text.

84. See Directive 95/46 of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data, art. 3(1), 1995 O.J. (L 281) 31, 39 (EC) [hereinafter Data Protection Directive].

85. See, e.g., William van Melle, Edward H. Shortliffe & Bruce G. Buchanan, *EMYCIN: A Knowledge Engineer’s Tool for Constructing Rule-Based Expert Systems*, in RULE-BASED EXPERT SYSTEMS: THE MYCIN EXPERIMENT OF THE STANFORD HEURISTIC PROGRAMMING PROJECT 302 (Bruce G. Buchanan & Edward H. Shortliffe eds., 1984).

86. 15 U.S.C. §§ 1681–1681x (2012).

87. 15 U.S.C. §§ 1691–1691f (2012).

inputs certain point values, add them to obtain a total score, and then make decisions based on that score. Each of these statutes require “adverse action notices” that must include a statement of reasons for denials of credit or other credit-based outcomes.⁸⁸ This is an example of what we call outcome-based explanations: a description of the facts that proved relevant to a decision, but not a description of the decision-making rules themselves.

Articles 13–15 of the GDPR require data subjects to have access to “meaningful information about the logic involved” in any automated decision-making that significantly affects them.⁸⁹ As the law is still new, the import and proper interpretation of this requirement remain unclear. In advance of a definitive interpretation, the GDPR appears to ask for a functional description of the model—enough of a description of the rules governing decision-making such that a data subject can vindicate her substantive rights under the GDPR and human rights laws.⁹⁰ This is an example of logic-based explanations: a description of the reasoning behind a decision, not just the relevant inputs to the decision.

1. FCRA, ECOA, and Regulation B

The most straightforward legal requirement to explain inscrutable decision-making is the adverse action notice. In 1970, Congress passed FCRA⁹¹ to begin to rein in the unregulated credit industry. FCRA was “the first information privacy legislation in the United States.”⁹² It limits to whom and for what purposes credit reports can be disclosed,⁹³ allows consumers access to their credit reports,⁹⁴ and requires credit reporting agencies (CRAs)—for example, Experian, Transunion, and Equifax—to employ procedures to ensure accuracy and govern dispute resolution.⁹⁵ FCRA was not initially concerned with how decisions were made, but rather with the then-new phenomenon of amassing large quantities of information.⁹⁶ Four years later, however, Congress passed ECOA⁹⁷ and took aim at the decision-

88. 15 U.S.C. §§ 1681m, 1691d(2).

89. GDPR, *supra* note 12, arts. 13(f)(2), 14(g)(2), 15(1)(h) (requiring access to “meaningful information about the logic” of automated decisions).

90. See Andrew D. Selbst & Julia Powles, *Meaningful Information and the Right to Explanation*, 7 INT'L DATA PRIVACY L. 233, 236 (2017). There is a vigorous debate in the literature about the “right to explanation” in the GDPR. See *infra* notes 143–45 and accompanying text. As a discussion of positive law, this debate is connected to, but different than, the point we seek to make about the GDPR—that it is one example of a law that operates by asking for the logic of a system. Even if there is held to be no “right to explanation” in the GDPR, one could imagine an equivalent law that encodes such a requirement.

91. Fair Credit Reporting Act, Pub. L. No. 91-508, 84 Stat. 1127 (1970) (codified as amended at 15 U.S.C. §§ 1681–1681x (2012)).

92. PRISCILLA M. REGAN, LEGISLATING PRIVACY: TECHNOLOGY, SOCIAL VALUES, AND PUBLIC POLICY 101 (1995).

93. 15 U.S.C. § 1681b.

94. *Id.* § 1681g.

95. *Id.* §§ 1681e(b), 1681i.

96. 115 CONG. REC. 2410 (1969).

97. Equal Credit Opportunity Act, Pub. L. No. 93-495, 88 Stat. 1521 (1974) (codified as amended at 15 U.S.C. §§ 1691–1691f (2012)).

making process.⁹⁸ ECOA prohibits discrimination in credit decisions on the basis of race, color, religion, national origin, sex, marital status, age (for adults), receipt of public assistance income, or exercise in good faith of the rights guaranteed under the Consumer Credit Protection Act.⁹⁹

ECOA introduced the adverse action notice requirement.¹⁰⁰ When a creditor takes an adverse action against an applicant, the creditor must give a statement of “specific reasons” for the denial.¹⁰¹ When FCRA later adopted a similar requirement, it expanded the notice to cover uses of credit information beyond decisions made by creditors, including the use of such information in employment decisions.¹⁰²

ECOA’s notice requirement was implemented by the Federal Reserve Board via Regulation B,¹⁰³ which mandates that the “statement of reasons . . . must be specific and indicate the principal reason(s) for the adverse action.”¹⁰⁴ The regulation also notes that it is insufficient to “state[] that the adverse action was based on the creditor’s internal standards or policies or that the applicant . . . failed to achieve a qualifying score on the creditor’s credit scoring system.”¹⁰⁵ An appendix to Regulation B offers a sample notification form designed to satisfy both the rule’s and FCRA’s notification requirements. Sample Form 1 offers twenty-four reason codes, including such varied explanations as “no credit file,” “length of employment,” or “income insufficient for amount of credit requested.”¹⁰⁶ Though it is not

98. *Id.* § 502, 88 Stat. at 1521 (noting that the purpose of the legislation is to ensure credit is extended fairly, impartially, and without regard to certain protected classes).

99. 15 U.S.C. § 1691 (2012).

100. *Id.* § 1691(d)(2)(B); Winnie F. Taylor, *Meeting the Equal Credit Opportunity Act’s Specificity Requirement: Judgmental and Statistical Scoring Systems*, 29 BUFF. L. REV. 73, 82 (1980) (“For the first time, federal legislation afforded rejected credit applicants an automatic right to discover why adverse action was taken.”).

101. 15 U.S.C. § 1691(d)(2)–(3).

102. 15 U.S.C. § 1681m (2012).

103. Regulation B, 12 C.F.R. §§ 1002.1–16 (2018).

104. 12 C.F.R. § 202.9(b)(2) (2018).

105. *Id.*

106. 12 C.F.R. pt. 1002, app. C (2018). The form’s listed options are:

- Credit application incomplete
- Insufficient number of credit references provided
- Unacceptable type of credit references provided
- Unable to verify credit references
- Temporary or irregular employment
- Unable to verify employment
- Length of employment
- Income insufficient for amount of credit requested
- Excessive obligations in relation to income
- Unable to verify income
- Length of residence
- Temporary residence
- Unable to verify residence
- No credit file
- Limited credit experience
- Poor credit performance with us
- Delinquent past or present credit obligations with others
- Collection action or judgment

necessary to use the form, most creditors tend to report reasons contained on that form because there is a safe harbor for “proper use” of the form.¹⁰⁷

Adverse action notices aim to serve three purposes: (1) to alert a consumer that an adverse action has occurred;¹⁰⁸ (2) to educate the consumer about how such a result could be changed in the future;¹⁰⁹ and (3) to prevent discrimination.¹¹⁰ As the rest of this section will show, these are commonly cited reasons for relying on explanations as a means of regulation as a general matter. The first rationale, consumer awareness, is straightforward enough. It is a basic requirement of any information-regulation regime that consumers be aware of systems using their information.¹¹¹ But the relationship between adverse action notices and the other two rationales—consumer education and antidiscrimination—requires further exploration.

Adverse action notices can be helpful for consumer education. As Winnie Taylor pointed out shortly after the passage of ECOA, some reasons—“no credit file” and “unable to verify income”—are self-explanatory and would allow a consumer to take appropriate actions to adjust.¹¹² Conversely, some explanations, such as “length of employment” and home ownership, are harder to understand or act on.¹¹³ This suggests that an explanation of a specific decision may be informative, but it may not reveal an obvious path to an alternative outcome.

There are also situations in which it may not even be informative. Taylor imagined a hypothetical additive credit-scoring system with eight different features—including whether an applicant owns or rents, whether he has a home phone, and what type of occupation he has, among other things—each assigned different point values.¹¹⁴ In a system like that, someone who comes up one point short could find himself with every factor listed as a “principal

-
- Garnishment or attachment
 - Foreclosure or repossession
 - Bankruptcy
 - Number of recent inquiries on credit bureau report
 - Value or type of collateral not sufficient
 - Other, specify: _____

Id.

107. Equal Credit Opportunity, 76 Fed. Reg. 41,590, 41,592 (July 15, 2011) (“A creditor receives a safe harbor for compliance with Regulation B for proper use of the model forms.”).

108. See S. REP. NO. 94-589, at 4 (1976).

109. *Id.* (“[R]ejected credit applicants will now be able to learn where and how their credit status is deficient and this information should have a pervasive and valuable educational benefit. Instead of being told only that they do not meet a particular creditor’s standards, consumers particularly should benefit from knowing, for example, that the reason for the denial is their short residence in the area, or their recent change of employment, or their already over-extended financial situation.”).

110. *Id.* (“The requirement that creditors give reasons for adverse action is . . . a strong and necessary adjunct to the antidiscrimination purpose of the legislation, for only if creditors know they must explain their decisions will they effectively be discouraged from discriminatory practices.”).

111. See *supra* note 32 and accompanying text.

112. Taylor, *supra* note 100, at 97.

113. *Id.* at 95.

114. *Id.* at 105–07.

reason”¹¹⁵ for the denial. In one sense, this must be correct because a positive change in any factor at all would change the outcome. In another sense, however, choosing arbitrarily among equivalently valid reasons runs counter to the instruction to give specific and actionable notice.

Taylor also described a real system from that era, complex in all the various ways described in Part I—nonlinear, nonmonotonic, discontinuous, and multidimensional:

[A]pplicants who have lived at their present address for less than six months are awarded 39 points, a level which they could not reach again until they had maintained the same residence for seven and one-half years. Furthermore, applicants who have been residents for between six months and 1 year 5 months (30 points) are considered more creditworthy than those who have been residents for between 1 and 1/2 years and 3 years 5 months (27 points).¹¹⁶

If the creditor tried to explain these rules simply, it would leave information out, but if the creditor were to explain in complete detail, it would likely overwhelm a credit applicant. This is an equivalent problem to simply disclosing how a model works under the banner of transparency; access to the model is not the same as understanding.¹¹⁷

The Federal Reserve Board recognized this problem, observing that, although all the principal reasons must be disclosed, “disclosure of more than four reasons is not likely to be helpful to the applicant.”¹¹⁸ The difficulty is that there will be situations where complexity cannot be avoided in a faithful representation of the scoring system, and listing factors alone will fail to accurately explain the decision, especially when the list is limited to four.¹¹⁹ It is worth noting that modern credit systems appear not to be based on such complex models,¹²⁰ likely due to the very existence of FCRA and ECOA. Credit predictions tend to rely on features that bear an intuitive relationship to default, such as past payment history.¹²¹ But the point is more general:

115. See *supra* note 104 and accompanying text.

116. Taylor, *supra* note 100, at 123.

117. See Ananny & Crawford, *supra* note 24, at 979 (“Transparency can intentionally occlude.”).

118. 12 C.F.R. pt. 1002 supp. I, para. 9(b)(2) (2018). FCRA later codified the same limitation. 15 U.S.C. § 1681g(f)(1)(C) (2012).

119. The document also states that the “specific reasons . . . must relate to and accurately describe the factors actually considered or scored by a creditor A creditor need not describe how or why a factor adversely affected an applicant If a creditor bases the . . . adverse action on a credit scoring system, the reasons disclosed must relate only to those factors actually scored in the system.” 12 C.F.R. pt. 1002 supp. I, para. 9(b)(2).

120. Patrick Hall, Wen Phan & SriSatish Ambati, *Ideas on Interpreting Machine Learning*, O’REILLY (Mar. 15, 2017), <https://www.oreilly.com/ideas/ideas-on-interpreting-machine-learning> [https://perma.cc/57XK-NU7G].

121. Carol A. Evans, *Keeping Fintech Fair: Thinking About Fair Lending and UDAP Risks*, CONSUMER COMPLIANCE OUTLOOK (Fed. Res. Sys., Phila., Pa.), 2017, at 4–5, <https://consumercomplianceoutlook.org/assets/2017/second-issue/ccoi22017.pdf> [https://perma.cc/52XP-PQN4]; see also ROBINSON + YU, KNOWING THE SCORE: NEW DATA, UNDERWRITING, AND MARKETING IN THE CONSUMER CREDIT MARKETPLACE 21 (2014), https://www.teamupturn.com/static/files/Knowing_the_Score_Oct_2014_v1_1.pdf [https://perma.cc/9FCY-4K2K].

approaches based on giving specific reasons for outcomes can fail where the system is too complex.

The adverse action notice fares worse as an antidiscrimination measure. By 1974, forcing hidden intentions into the open was a common technique for addressing discrimination.¹²² Just one year before ECOA's passage, *McDonnell Douglas Corp. v. Green*¹²³ laid out the canonical Title VII burden-shifting framework for disparate treatment, which requires a defendant to rebut a *prima facie* case of employment discrimination with a nondiscriminatory reason and gives plaintiffs a chance to prove that the proffered reason is pretextual.¹²⁴ Just two years before that, the U.S. Supreme Court in *Griggs v. Duke Power Co.*¹²⁵ recognized disparate impact doctrine.¹²⁶ Disparate impact attributes liability for a facially neutral decision that has a disproportionate adverse effect on a protected class unless the decision maker can provide a legitimate business reason for the decision and no equally effective but less discriminatory alternative exists.¹²⁷ Its initial purpose was arguably to smoke out intentional discrimination where intent was hidden.¹²⁸ Thus, ECOA pursued the same goal—to prevent discrimination by forcing decision-making into the open.

While forcing stated reasons into the open captures the most egregious forms of intentional discrimination, it does not capture much else. Although, in some cases, Regulation B bars collection of protected-class information,¹²⁹ race, gender, and other features can be reliably inferred from sufficiently rich datasets.¹³⁰ Should creditors seek to discriminate intentionally by considering membership in a protected class, they would have to affirmatively lie about such behavior lest they reveal obvious wrongdoing. This form of intentional discrimination is thus addressed by disclosure. Should creditors rely on known proxies for membership in a protected class, however, while they would have to withhold the true relevance of these features in predicting creditworthiness, they could cite them honestly as reasons for the adverse action. The notice requirement therefore does not place meaningful constraints on creditors, nor does it create additional or

122. See Olatunde C. A. Johnson, *The Agency Roots of Disparate Impact*, 49 HARV. C.R.-C.L.L. REV. 125, 140 (2014) (tracing the history of agency use of disparate impact analysis to address latent discrimination).

123. 411 U.S. 792 (1973).

124. *Id.* at 805. The Supreme Court later found that a jury may presume that if all the employer had was pretext, that itself is evidence of discrimination. *St. Mary's Honor Ctr. v. Hicks*, 509 U.S. 502, 511 (1993) ("The factfinder's disbelief of the reasons put forward by the defendant (particularly if disbelief is accompanied by a suspicion of mendacity) may, together with the elements of the *prima facie* case, suffice to show intentional discrimination.").

125. 401 U.S. 424 (1971).

126. *Id.* at 431.

127. 42 U.S.C. § 2000e-2(k)(1)(A) (2012). This description ignores the word "refuse" in the statute, but is probably the more common reading. Barocas & Selbst, *supra* note 4, at 709.

128. Richard A. Primus, *Equal Protection and Disparate Impact: Round Three*, 117 HARV. L. REV. 494, 518–21 (2003) (discussing the "evidentiary dragnet" theory of disparate impact).

129. 12 C.F.R. § 1002.5 (2018).

130. Barocas & Selbst, *supra* note 4, at 692.

unique liability beyond that present in the antidiscrimination provisions of the rest of the regulation.¹³¹

More importantly, creditors using quantitative methods that do not expressly consider protected-class membership are likely not engaged in intentional discrimination, yet the scoring systems might very well evince a disparate impact. While ECOA does not expressly provide for a disparate impact theory of discrimination, case law suggests that it is very likely available.¹³²

The adverse action notice approach has two specific shortcomings for a disparate impact case. First, when reviewing such a notice, the consumer only has access to her own specific outcome. Her single point of reference does not provide any understanding of the frequency of denials along protected-class lines, so she cannot observe disparate impact. Absent understanding of the logic of the system—for example, how different inputs are weighted—she cannot even look at the decision-making to try to guess whether it is discriminatory; the notice simply provides no basis to bring a suit.

Second, disparate impact has a different relationship to reasons behind decisions than does intentional discrimination. While for intentional discrimination, a consumer only needs to know that the decision was not made for an improper reason, knowing the specific reasons for which it *was* made becomes important for a disparate impact case.¹³³ That is to say, it is not only important to understand how a statistical system converts inputs to specific outputs, but also why the system was set up that way.

As discussed in Part I, one avenue to ensure the existence of an explanation of why the rules are the way they are is to require that the rules be based on intuitive relationships between input and output variables. This is the approach advocated by several scholars, particularly those focused on discrimination.¹³⁴ As is discussed in Part IV, it is not the only way, but this inability to engage with the normative purposes of the statute is a clear shortcoming of explanations based solely on the outcome of a single case, which provides neither the logic of the system nor any information about its normative elements.

131. John H. Matheson, *The Equal Credit Opportunity Act: A Functional Failure*, 21 HARV. J. ON LEGIS. 371, 388 (1984).

132. The Supreme Court has not ruled that it is available, but most circuit courts that have considered it have permitted it. See Mikella Hurley & Julius Adebayo, *Credit Scoring in the Era of Big Data*, 18 YALE J.L. & TECH. 148, 193 (2016) (citing *Golden v. City of Columbus*, 404 F.3d 950, 963 (6th Cir. 2005)). In addition, the Supreme Court ruled in 2015 that disparate impact theory was cognizable in the Fair Housing Act, which also does not expressly provide for it. *Texas Dep’t of Hous. & Cmty. Affairs v. Inclusive Cmtys. Project, Inc.*, 135 S. Ct. 2507, 2518 (2015).

133. Barocas & Selbst, *supra* note 4, at 702.

134. See *infra* Part III.A.3.

2. GDPR

In 2016, the European Union (EU) passed the GDPR, which took effect on May 25, 2018, and replaced the 1995 Data Protection Directive.¹³⁵ Both laws regulate automated decision-making,¹³⁶ but in the twenty-three years of the Directive's existence, little jurisprudence developed around that particular aspect of the law. The GDPR has created renewed interest in these provisions.¹³⁷

The GDPR's discussion of automated decisions is contained in Articles 22, 13(2)(f), 14(2)(g), and 15(1)(h). Article 22 is the primary provision and states, in relevant part, the following:

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
2. Paragraph 1 shall not apply if the decision:
 - (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;
 - (b) . . .
 - (c) is based on the data subject's explicit consent.
3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.¹³⁸

Articles 13–15 spell out a data subject's right to be informed about the information that data controllers have about her.¹³⁹ Articles 13 and 14 describe the obligations of data controllers to affirmatively notify data subjects about the uses of their information,¹⁴⁰ and Article 15 delineates the access rights that data subjects have to information about how their own data is used.¹⁴¹ All three demand that the following information be available to data subjects: “the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.”¹⁴²

135. GDPR, *supra* note 12, art. 99.

136. *Id.* art. 22(1) (“The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.”); Data Protection Directive, *supra* note 84, art. 15.

137. Isak Mendoza & Lee A. Bygrave, *The Right Not to Be Subject to Automated Decisions Based on Profiling*, in EU INTERNET LAW 77, 80–81 (2017).

138. GDPR, *supra* note 12, art. 22. Article 22(4) is omitted because it is not relevant to this discussion.

139. Wachter et al., *supra* note 23, at 89.

140. See GDPR, *supra* note 12, arts. 13–14.

141. See *id.* art. 15.

142. *Id.* arts. 13(2)(f), 14(2)(g), 15(1)(h).

Since passage of the GDPR, scholars have debated whether these requirements amount to a “right to explanation.”¹⁴³ As one of us has argued elsewhere, that debate has been bogged down in proxy battles over what the phrase “right to explanation” means, but no matter whether one calls it a right to explanation, requiring that data subjects have meaningful information about the logic must mean something related to explanation.¹⁴⁴ Importantly for this discussion, the Regulation demands that the “meaningful information” must be about the *logic* of the decisions.¹⁴⁵ As we defined it in Part I, a model is inscrutable when it defies practical inspection and resists comprehension. An explanation of the logic therefore appears to precisely target inscrutability. The most important aspect of this type of explanation is that it is concerned with the operation of the model in general, rather than as it pertains to a particular outcome.

The particular type of explanation required by the GDPR will depend on the legal standards developed in the EU by the authorities charged with interpreting that law. The overall purposes of the GDPR are much broader than FCRA and ECOA. The EU treats data protection as a fundamental right,¹⁴⁶ and the GDPR seeks to vindicate the following principles with respect to personal data: lawfulness, fairness, and transparency; purpose limitation; data minimization; accuracy; storage limitation; integrity and confidentiality; and accountability.¹⁴⁷ Several of these principles are a restatement of the FIPs that have shaped privacy policy for decades.¹⁴⁸

143. See Margot E. Kaminski, *The Right to Explanation, Explained*, 34 BERKELEY TECH. L.J. (forthcoming 2019) (manuscript at 17–24), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3196985 [https://perma.cc/92GH-W6HV] (reviewing the literature); see also Lilian Edwards & Michael Veale, *Slave to the Algorithm? Why a “Right to an Explanation” Is Probably Not the Remedy You Are Looking For*, 16 DUKE L. & TECH. REV. 18, 44 (2017) (arguing that even if a right to explanation exists, it may not be useful); Gianclaudio Malgieri & Giovanni Comandé, *Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation*, 7 INT’L DATA PRIVACY L. 243, 245, 250 (2017) (arguing that the GDPR creates a right to “legibility” that combines transparency and comprehensibility); Mendoza & Bygrave, *supra* note 137 (arguing that a right to explanation can be derived as a necessary precursor to the right to contest the decision); Selbst & Powles, *supra* note 90 (arguing that a right to meaningful information is a right to explanation); Sandra Wachter et al., *Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR*, 31 HARV. J.L. & TECH 841 (2018) (arguing that a legal right to explanations of automated decisions does not exist); Wachter et al., *supra* note 23 (arguing that there is no legal right to explanation of specific automated decisions); Goodman & Flaxman, *supra* note 20, at 2 (arguing that a right to explanation exists); Maja Brkan, *Do Algorithms Rule the World? Algorithmic Decision-Making and Data Protection in the Framework of the GDPR and Beyond* 15 (Aug. 1, 2017) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3124901 [https://perma.cc/C9PN-4PL6] (arguing that “information about the logic involved” and the right to contest decisions imply a right to explanation).

144. See Selbst & Powles, *supra* note 90, at 233.

145. GDPR, *supra* note 12, arts. 13(2)(f), 14(2)(g), 15(1)(h).

146. *Id.* art. 1.

147. *Id.* art. 5.

148. Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93, 106–07 (2014). While different lists of FIPs conflict, one prominent example is the Organisation for Economic Co-Operation and Development’s (OECD) list: Collection Limitation Principle, Data Quality Principle,

Considered as a whole, they begin to sound like the general idea of due process in all its expansiveness.

Satisfying this requirement may in some cases involve disclosing the full set of rules behind all decision-making—that is, the entire model.¹⁴⁹ But in some cases, it will not involve such radical disclosure. Depending on the specific goals at issue, the types of rules disclosed can be narrower, or the explanation can perhaps be met interactively by providing data subjects with the tools to examine how changes in their information relate to changes in outcome. One of us has argued that the GDPR’s meaningful information requirement applies “to the data subject herself”¹⁵⁰ and “should be interpreted functionally and flexibly,” and that the legal standard should be that the explanation “at a minimum, enable[s] a data subject to exercise his or her rights under the GDPR and human rights law.”¹⁵¹

Although the GDPR’s goals are broader than those of ECOA and FCRA, evaluating the ability of logic-based explanations to vindicate the goals of those statutes can demonstrate how explanations of the logic of decision-making can improve upon the shortcomings of the outcome-based approach. The three reasons were awareness, consumer (here, data subject) education, and antidiscrimination.¹⁵² Like in the credit domain, awareness is straightforward and encapsulated by the requirement that a data subject be made aware of the “existence” of automated decision-making. The other two rationales operate differently when logic-based explanations are provided.

Data subject education becomes more straightforward as a legal matter, if not a technical one. Absent inscrutability, a data subject would be told the rules of the model and would be able to comprehend his situation and how to achieve any particular outcome. This solves both problems that Taylor identified.¹⁵³ Consider the system where, after the creditor totaled the point values from eight factors, a person missed on her credit application by one point. While it might be impossible to point to four factors that were “principal reasons,” the explanation of the logic—what the eight factors were, that they were all assigned point values, and that the hypothetical applicant just missed by a point—would be much more useful to that

Purpose Specification Principle, Use Limitation Principle, Security Safeguards Principle, Openness Principle, Individual Participation Principle, and Accountability Principle. Org. for Econ. Co-operation & Dev. [OECD], *The OECD Privacy Framework*, at 14–15 (2013), http://www.oecd.org/sti/ieconomy/oecd_privacy_framework.pdf [<https://perma.cc/RWM2-EUD4>].

149. The guidelines issued by the Article 29 Working Party, a body tasked with giving official interpretations of EU law, states that the full model is not required. See Article 29 Data Protection Working Party, *Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679*, at 25, WP 251 (Feb. 6, 2018) (“The GDPR requires the controller to provide meaningful information about the logic involved, not necessarily a complex explanation of the algorithms used or disclosure of the full algorithm.”). As a matter of positive law, then, this is likely to be the outcome, but in some cases it may fall short of something actually meaningful to the data subject.

150. See Selbst & Powles, *supra* note 90, at 236.

151. *Id.* at 233.

152. See *supra* notes 108–10 and accompanying text.

153. See *supra* notes 112–16 and accompanying text.

particular rejected applicant.¹⁵⁴ In Taylor’s real nonlinear, nonmonotonic, discontinuous, and multidimensional example, the full complexity can be appreciated in the paragraph-long description, where a reason code would in many cases be totally unhelpful. Once machine learning enters the picture, and models become more complex, the limits on technical ability to solve inscrutability may prevent these explanations from coming to fruition. But at least in theory, explanations of the logic are sufficient for data subject education.

Turning to discrimination—which serves as a stand-in for broader normative questions about model justification—while logic-based explanations do fare better than outcome-based ones, they do not completely address the shortcomings. Any rule that is manifestly objectionable becomes visible under logic-based explanations, making them an improvement over outcome-only explanations, which shed no light on rules. This disclosure might enable one to speculate if facially neutral rules will nevertheless have a disparate impact, based on the different rates at which certain input features are held across the population. But this is ultimately little more than guesswork.¹⁵⁵ Although there might not be anything about a rule that appears likely to generate a disparate impact, it still could. Alternatively, a set of rules could appear objectionable or discriminatory, but ultimately be justified. It will often be impossible to tell without more information, and the possibility of happening on a set of rules that lend themselves to intuitive normative assessment is only a matter of chance.

B. Interpretability in Machine Learning

The overriding question that has prompted fierce debates about explanation and machine learning has been whether machine learning can be made to comply with the law. As discussed in Part I, machine learning poses unique challenges for explanation and understanding—and thus challenges for meeting the apparent requirements of the law. Part II.A further demonstrated that even meeting the requirements of the law does not automatically provide the types of explanations that would be necessary to assess whether decisions are well justified. Nevertheless, addressing the potential inscrutability of machine learning models remains a fundamental step in meeting this goal.

As it happens, machine learning has a well-developed toolkit to deal with calls for explanation. There is an extensive literature on “interpretability.”¹⁵⁶ Early research recognized and grappled with the challenge of explaining the decisions of machine learning models such that people using these systems

154. The Article 29 Working Party has, however, suggested that this approach is central to the “meaningful information” requirement. See Article 29 Data Protection Working Party, *supra* note 149, at 25.

155. See *infra* Part III.A.3.

156. See generally, e.g., Riccardo Guidotti et al., *A Survey of Methods for Explaining Black Box Models*, 51 ACM COMPUTING SURVEYS, Aug. 2018, at 1; Lipton, *supra* note 66.

would feel comfortable acting upon them.¹⁵⁷ Practitioners and researchers have developed a wide variety of strategies and techniques to ensure that they can produce interpretable models from data—many of which may be useful for complying with existing law, such as FCRA, ECOA, and the GDPR.

Interpretability has received considerable attention in research and practice due to the widely held belief that there is a tension between how well a model will perform and how well humans will be able to interpret it.¹⁵⁸ This view reflects the reasonable idea that models that consider a larger number of variables, a larger number of relationships between these variables, and a more diverse set of potential relationships is likely to be *both* more accurate and more complex.¹⁵⁹ This will certainly be the case when the phenomenon that machine learning seeks to model is itself complex. This intuition suggests that practitioners may face a difficult choice: favor simplicity for the sake of interpretability or accept complexity to maximize performance.¹⁶⁰

While such views seem to be widely held,¹⁶¹ over the past decade, methods have emerged that attempt to sidestep these difficult choices altogether, promising to increase interpretability while retaining performance.¹⁶² Researchers have developed at least three different ways to respond to the demand for explanations: (1) purposefully orchestrating the machine learning process such that the resulting model is interpretable;¹⁶³ (2) applying special techniques after model creation to approximate the model in a more readily intelligible form or identify features that are most salient for specific decisions;¹⁶⁴ and (3) providing tools that allow people to interact with the model and get a sense of its operation.¹⁶⁵

1. Purposefully Building Interpretable Models

Practitioners have a number of different levers at their disposal to purposefully design simpler models. First, they may choose to consider only a limited set of all possible variables.¹⁶⁶ By limiting the analysis to a smaller set of variables, the total number of relationships uncovered in the learning process might be sufficiently limited to be intelligible to a human.¹⁶⁷ It is

157. van Melle et al., *supra* note 85, at 302.

158. See, e.g., Leo Breiman, *Statistical Modeling: The Two Cultures*, 16 STAT. SCI. 199, 206 (2001); Lou et al., *supra* note 54, at 150.

159. See Breiman, *supra* note 158, at 208.

160. See generally *id.*

161. See DEF. ADVANCED RESEARCH PROJECTS AGENCY, BROAD AGENCY ANNOUNCEMENT: EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI) (2016), <https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf> [<https://perma.cc/3FZV-TZGA>]; Henrik Brink & Joshua Bloom, *Overcoming the Barriers to Production-Ready Machine-Learning Workflows*, STRATA (Feb. 11, 2014), <https://conferences.oreilly.com/strata/strata2014/public/schedule/detail/32314> [<https://perma.cc/2GBV-2QRR>].

162. For a recent survey, see Michael Gleicher, *A Framework for Considering Comprehensibility in Modeling*, 4 BIG DATA 75 (2016).

163. See, e.g., *id.* at 81–82.

164. See, e.g., *id.* at 82–83.

165. See, e.g., *id.* at 83.

166. See *id.* at 81.

167. Zeng et al., *supra* note 82, at 690–91.

very likely that a model with five features, for example, will be more interpretable than a model with five hundred.

Second, practitioners might elect to use a learning method that outputs a model that can be more easily parsed than the output of other learning methods.¹⁶⁸ For example, decision tree algorithms are perceived as likely to produce interpretable models because they learn nested rules that can be represented visually as a tree with subdividing branches. To understand how the model would process any particular case, practitioners need only walk through the relevant branches of the tree; to understand the model overall, practitioners can explore all the branches to develop a sense of how the model would determine all possible cases.

The experience of applying machine learning to real-world problems has led to common beliefs among practitioners about the relative interpretability of models that result from different learning methods and how well they perform. Conventional wisdom suggests that there is a trade-off between interpretability and accuracy.¹⁶⁹ Methods like linear regression¹⁷⁰ generate models perceived as highly interpretable, but relatively low performing, while methods like deep learning¹⁷¹ result in high-performing models that are exceedingly difficult to interpret.¹⁷² While researchers have pointed out that such comparisons do not rest on a rigorous definition of interpretability or empirical studies,¹⁷³ such beliefs routinely guide practitioners' decisions when applying machine learning to different kinds of problems.¹⁷⁴

Another method is to set the parameters of the learning process to ensure that the resulting model is not so complex that it defies human comprehension. For example, even decision trees will become unwieldy for humans if they involve an exceedingly large number of branches and leaves.¹⁷⁵ Practitioners routinely set an upper bound on the number of leaves to constrain the complexity of the model.¹⁷⁶ For decades, practitioners in regulated industries like credit and insurance have purposefully limited themselves to a relatively small set of features and less sophisticated learning methods.¹⁷⁷ In so doing, they have been able to generate models that lend themselves to sensible explanation, but they may have forgone the increased accuracy that would result from a richer and more advanced analysis.¹⁷⁸

168. See Lehr & Ohm, *supra* note 51, at 688–95.

169. See, e.g., Breiman, *supra* note 158, at 208.

170. See *Regression*, CONCISE OXFORD DICTIONARY OF MATHEMATICS (3d ed. 2014).

171. See generally Jürgen Schmidhuber, *Deep Learning in Neural Networks: An Overview*, 61 NEURAL NETWORKS 85 (2015) (providing an explanation of deep learning in artificial intelligence).

172. Breiman, *supra* note 158, at 206.

173. Alex A. Freitas, *Comprehensible Classification Models—a Position Paper*, 15 SIGKDD EXPLORATIONS, June 2013, at 1.

174. See Lipton, *supra* note 66, at 99.

175. *Id.* at 98.

176. See *id.* at 99.

177. Hall et al., *supra* note 120.

178. *Id.*

Linear models remain common in industry because they allow companies to much more readily comply with the law.¹⁷⁹ When they involve a sufficiently small set of features, linear models are concise enough for a human to grasp the relevant statistical relationships and to simulate different scenarios.¹⁸⁰ They are simple enough that a full description of the model may amount to the kind of meaningful information about the logic of automated decisions required by the GDPR. At the same time, linear models also immediately highlight the relative importance of different features by assigning a specific numerical weight to each feature, which allows companies to quickly extract the principal factors for an adverse action notice under ECOA.

Beyond the choice of features, learning method, or learning parameters, there are techniques that can make simplicity an additional and explicit optimization criterion in the learning process. The most common such method is regularization.¹⁸¹ Much like setting an upper limit on the number of branches in a decision tree, regularization allows the learning process to factor in model complexity by assigning a cost to excess complexity.¹⁸² In doing so, model simplicity becomes an additional objective alongside model performance, and the learning process can be set up to find the optimal trade-off between these sometimes-competing objectives.¹⁸³

Finally, the learning process can also be constrained such that all features exhibit monotonicity.¹⁸⁴ Monotonicity constraints are widespread in credit scoring because they make it easier to reason about how scores will change when the value of specific variables change, thereby allowing creditors to automate the process of generating the reason codes required by FCRA and ECOA.¹⁸⁵ As a result of these legal requirements, creditors and other data-

179. *Id.*

180. See Lipton, *supra* note 66, at 98.

181. See Gleicher, *supra* note 162, at 81–82.

182. See *id.* at 81. One commonly used version of this method is Lasso. See generally Robert Tibshirani, *Regression Shrinkage and Selection via the Lasso*, 58 J. ROYAL STAT. SOC’Y 267 (1996). It was originally designed to increase accuracy by avoiding overfitting, which occurs when a model assigns significance to too many features and thus accidentally learns patterns that are peculiar to the training data and not representative of real-world patterns. See *id.* at 267. Machine learning is only effective in practice when it successfully identifies robust patterns while also ignoring patterns that are specific to the training data. See David J. Hand, *Classifier Technology and the Illusion of Progress*, 21 STAT. SCI. 1, 2 (2006). Lasso increases accuracy by forcing the learning process to ignore relationships that are relatively weak, and therefore more likely to be artifacts of the training data. See Tibshirani, *supra*, at 268. Because Lasso works by strategically removing unnecessary features, the technique can simultaneously improve interpretability (by reducing complexity) in many real-world applications and increase performance (by avoiding overfitting). See *id.* at 267. As such, improved interpretability need not always decrease performance. But where potential overfitting is not a danger, regularization methods may result in degradations in performance. See Gleicher, *supra* note 162, at 81–82.

183. Gleicher, *supra* note 162, at 81.

184. Recall that monotonicity implies that an increase in an input variable can only result in either an increase or decrease in the output; it can never change from one to the other. See *supra* notes 57–58 and accompanying text.

185. See, e.g., Hall et al., *supra* note 120. Monotonicity allows creditors to rank order variables according to how much the value of each variable in an applicant’s file differs from

driven decision makers often have incentives to ensure their models are interpretable by design.

2. Post Hoc Methods

There exists an entirely different set of techniques for improved interpretability that does not place any constraints on the model-building process. Instead, these techniques begin with models learned with more complex methods and attempt to approximate them with simpler and more readily interpretable methods. Most methods in this camp generate what can be understood as a model of the model.

These methods attempt to overcome the fact that simpler learning methods cannot always reliably discover as many useful relationships in the data. For example, the learning process involved in decision trees is what is known as a “greedy algorithm.”¹⁸⁶ Once the learning process introduces a particular branch, the method does not permit walking back up the branch.¹⁸⁷ Therefore, relationships between items on two different branches will not be discovered.¹⁸⁸ Despite lacking the same limitation, more complex learning methods, such as deep learning, do not result in models as interpretable as decision trees. Nonetheless, rules that cannot be *learned* with simpler methods can often be *represented* effectively by simpler models.¹⁸⁹ Techniques like rule extraction¹⁹⁰ allow simple models to “cheat” because the answers that simpler learning methods would otherwise miss are known ahead of time.¹⁹¹

This approach can be costly and it does not have universal success.¹⁹² Despite practitioners’ best efforts, replicating the performance of more complex models in a simple enough form might not be possible where the phenomena are particularly complex. For example, using a decision tree to approximate a model developed with deep learning might require too large a number of branches and leaves to be understandable in practice.¹⁹³

When these methods work well, they ensure that the entire set of relationships learned by the model can be expressed concisely, without

the corresponding value of each variable for the ideal customer—the top four variables can function as reason codes. *Id.*

186. STUART RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH 92–93 (3d ed. 2014).

187. *Id.*

188. *Id.* at 93 (noting that, although the greedy algorithm may find a nonoptimal solution, it will not discover relationships between unrelated branches).

189. Gleicher, *supra* note 162, at 82.

190. Rule extraction is the name for a set of techniques used to create a simplified model of a model. The technical details of their operation are beyond the scope of this paper. See generally Nahla Barakat & Andrew P. Bradley, *Rule Extraction from Support Vector Machines: A Review*, 74 NEUROCOMPUTING 178 (2010); David Martens et al., *Comprehensible Credit Scoring Models Using Rule Extraction from Support Vector Machines*, 183 EUR. J. OPERATIONAL RES. 1466 (2007).

191. Gleicher, *supra* note 162, at 82.

192. *Id.*

193. See Lipton, *supra* note 66, at 98.

giving up much performance. Accordingly, they serve a similar role to the interpretability-driven design constraints discussed above.¹⁹⁴ When they do not work as well, arriving at an interpretable model might necessitate sacrificing some of the performance gained by using the more complex model. But even when these methods involve a notable loss in performance, the resulting models frequently perform far better than simple methods alone.¹⁹⁵

Other tools have also emerged that attack the problem of interpretability from a different direction. Rather than attempting to ensure that machine learning generates an intelligible model overall, these new tools furnish more limited explanations that only account for the relative importance of different features in particular outcomes—similar to the reason codes required by FCRA and ECOA.¹⁹⁶ At a high level, most of these methods adopt a similar approach: they attempt to establish the importance of any feature to a particular decision by iteratively varying the value of that feature while holding the value of other features constant.¹⁹⁷

These tools seem well suited for the task set by ECOA, FCRA, or other possible outcome-oriented approaches: explaining the principal reasons that account for the specific adverse decision.¹⁹⁸ As we further discuss in the next section, there are several reasonable ways to explain the same specific outcome. These methods are useful for two of the most common: (1) determining the relative contribution of different features, or (2) identifying the features whose values would have to change the most to change the outcome.¹⁹⁹ One could imagine applying these methods to models that consider an enormous range of features and map out an exceedingly complex set of relationships. While such methods will never make these relationships completely sensible to a human, they can provide a list of reasons that might help provide reason codes for a specific decision.

194. See *supra* Part II.B.1.

195. Johan Huysmans et al., *Using Rule Extraction to Improve the Comprehensibility of Predictive Models* (Katholieke Universiteit Leuven Dep't of Decision Scis. & Info. Mgmt., Working Paper No. 0612, 2006), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=961358 [https://perma.cc/8AKQ-LXVE].

196. See *supra* note 106 and accompanying text.

197. See generally Philip Adler et al., *Auditing Black-Box Models for Indirect Influence*, 54 KNOWLEDGE & INFO. SYSTEMS 95 (2018); David Baehrens et al., *How to Explain Individual Classification Decisions*, 11 J. MACHINE LEARNING RES. 1803 (2010); Anupam Datta et al., *Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems*, in PROCEEDINGS OF THE 2016 IEEE SYMPOSIUM ON SECURITY & PRIVACY 598 (2016); Andreas Henelius et al., *A Peek into the Black Box: Exploring Classifiers by Randomization*, 28 DATA MINING & KNOWLEDGE DISCOVERY 1503 (2014); Marco Tulio Ribeiro et al., “*Why Should I Trust You?*” *Explaining the Predictions of Any Classifier*, in PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING 1135 (2016).

198. See *supra* note 88 and accompanying text.

199. These methods are generally sensitive to interactions among variables and can measure indirect as well as direct influence. See, e.g., Adler et al., *supra* note 197; Datta et al., *supra* note 197; Julius Adebayo, *FairML: Auditing Black-Box Predictive Models*, CLOUDERA FAST FORWARD LABS (Mar. 9, 2017), <http://blog.fastforwardlabs.com/2017/03/09/fairml-auditing-black-box-predictive-models.html> [https://perma.cc/S5PK-K6GQ].

Unfortunately, these methods may not work well in cases where models take a much larger set of features into account. Should many features each contribute a small amount to a particular determination, listing each feature in an explanation is not likely to be helpful. This is the machine learning version of Taylor's hypothetical eight-factor credit example.²⁰⁰ The number of features identified as influential might be sufficiently large that the explanation would simply reproduce the problem of inscrutability that it aims to address. The only alternative in these cases—arbitrarily listing fewer reasons than the correct number—is also unsatisfying when all features are equivalently, or nearly equivalently, important. As it happens, post hoc explanations for credit and other similarly important decisions are likely to be most attractive precisely when they do not seem to work well—that is, when the only way to achieve a certain level of performance is to vastly expand the range of features under consideration.

These methods are also unlikely to generate explanations that satisfy logic-like approaches like the GDPR. Indeed, such techniques pose a unique danger of misleading people into believing that the reasons that account for specific decisions must also apply in the same way for others—that the reasons for a specific decision illustrate a general rule. Understandably, humans tend to extrapolate from explanations of specific decisions to similar cases, but the model—especially a complex one—may have a very different basis for identifying similar-seeming cases.²⁰¹ These methods offer explanations that apply only to the case at hand and cannot be extrapolated to decisions based on other input data.²⁰²

3. Interactive Approaches

One final set of approaches is interactive rather than explanatory. Practitioners can allow people to get a feel for their models by producing interactive interfaces that resemble the methods described in the previous sections. This can take two quite different forms. One is the type proposed by Danielle Citron and Frank Pasquale²⁰³ and implemented, for example, by Credit Karma.²⁰⁴ Beginning with a person's baseline credit information, Credit Karma offers a menu of potential changes, such as opening new credit cards, obtaining a new loan, or going into foreclosure.²⁰⁵ A person using the interface can see how each action would affect his credit score.²⁰⁶ This does

200. See *supra* notes 114–15 and accompanying text.

201. See Finale Doshi-Velez & Mason Kortz, *Accountability of AI Under the Law: The Role of Explanation* 3 (Harvard Univ. Berkman Klein Ctr. Working Grp. on Explanation & the Law, Working Paper No. 18-07, 2018), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3064761 [<https://perma.cc/SJ5S-HJ3T>] (discussing the problem of cases where similar situations lead to differing outcomes and vice versa).

202. See *id.*

203. See Citron & Pasquale, *supra* note 7, at 28–30 (discussing “interactive modeling”).

204. See *Credit Score Simulator*, CREDIT KARMA, <https://www.creditkarma.com/tools/credit-score-simulator> [<https://perma.cc/XQ2S-GYUE>] (last visited Nov. 15, 2018).

205. *Id.*

206. *Id.*

not amount to a full explanation because a person at a different starting point could make similar moves with different outcomes, but it gives the individual user a partial functional feel for the logic of the system as it applies to him specifically.

The second is more complicated and abstract. Mireille Hildebrandt has proposed something she terms “transparency-enhancing technologies.”²⁰⁷ Such technologies would implement an interface that would allow people to simultaneously adjust the value of multiple features in a model with the goal of providing a loose sense of the relationship between these features and a specific outcome, as well as the connection between the features themselves.²⁰⁸ The goal of this type of technology is not to tell the user what changes in his results specifically but to allow him to get a feel from an arbitrary starting point.²⁰⁹

Where models are simple enough, these approaches seem to achieve the educational goals of both ECOA and the GDPR by allowing data subjects to gain an intuitive feel for the system. Ironically, this would be accomplished by complying with neither law because a person will not know a specific reason for denial or have an account of a model’s logic after playing with it, even if they feel that they understand the model better afterward.

While regulators have expressed interest in this idea,²¹⁰ however, it poses a technical challenge. The statistical relationships at work in these models may be sufficiently complex that no consistent rule may become evident by tinkering with adjustable sliders. Models might involve a very large number of inputs with complex and shifting interdependencies such that even the most systematic tinkering would generate outcomes that would be difficult for a person to explain in a principled way.

One danger of this approach, then, is that it could do more to placate than elucidate. People could try to make sense of variations in the observed outputs by favoring the simplest possible explanation that accounts for the limited set of examples generated by playing with the system. Such an explanation is likely to take the form of a rule that incorrectly assigns a small set of specific variables unique significance and treats their effect on the outcome as linear, monotonic, and independent. Thus, for already simple models that *can* be explained, interactive approaches may be useful for giving people a feel without disclosing the algorithm, but for truly inscrutable systems, they could well be dangerous.

207. Mireille Hildebrandt, *Profiling: From Data to Knowledge*, 30 DATENSCHUTZ UND DATENSICHERHEIT 548, 552 (2006); see also Mireille Hildebrandt & Bert-Jaap Koops, *The Challenges of Ambient Law and Legal Protection in the Profiling Era*, 73 MODERN L. REV. 428, 449 (2010). See generally NICHOLAS DIAKOPoulos, ALGORITHMIC ACCOUNTABILITY REPORTING: ON THE INVESTIGATION OF BLACK BOXES (2013), http://towcenter.org/wp-content/uploads/2014/02/78524_Tow-Center-Report-WEB-1.pdf [https://perma.cc/H9UU-WK6V].

208. See Hildebrandt & Koops, *supra* note 207, at 450.

209. See *id.*

210. See INFO. COMM’R’S OFFICE, BIG DATA, ARTIFICIAL INTELLIGENCE, MACHINE LEARNING AND DATA PROTECTION 87–88 (2017), <https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf> [https://perma.cc/J97E-N5NV].

* * *

Remarkably, the techniques available within machine learning for ensuring interpretability correspond well to the different types of explanation required by existing law. There are, on the one hand, varied strategies and techniques available to practitioners that can deliver models whose inner workings can be expressed succinctly and sensibly to a human observer, whether an expert (e.g., a regulator) or lay person (e.g., an affected consumer). Laws like the GDPR that seek logic-like explanations would be well served by these methods. On the other hand, outcome-focused laws like ECOA that care only about principal reasons—and not the set of rules that govern all decisions—have an obvious partner in tools that furnish post hoc accounts of the factors that influenced any particular determination.

Where they succeed, these methods can be used to meet the demands of regulatory regimes that demand outcome- and logic-like explanations. Both techniques have their limitations, however. If highly sophisticated machine learning tools continue to be used, interpretability may be difficult to achieve in some instances, especially when the phenomena at issue are themselves complex. Post hoc accounts that list the factors most relevant to a specific decision may not work well when the number of relevant factors grows beyond a handful—a situation that is most likely to occur when such methods would be most attractive.

Notably, neither the techniques nor the laws go beyond describing the operation of the model. Though they may help to explain why a decision was reached or how decisions are made, they cannot address why decisions happen to be made that way. As a result, standard approaches to explanation might not help determine whether the particular way of making decisions is normatively justified.

III. FROM EXPLANATION TO INTUITION

So far, the majority of discourse around understanding machine learning models has seen the proper task as opening the black box and explaining what is inside.²¹¹ Where Part II.A discussed legal requirements and Part II.B discussed technical approaches, here we discuss the motivations for both. Based on a review of the literature, scholars, technologists, and policymakers seem to have three different beliefs about the value of opening the black box.²¹² The first is a fundamental question of autonomy, dignity, and

211. See *supra* note 16 and accompanying text.

212. These three rationales seem to track the rationales for ECOA's adverse action notices as described in Part II.A.1. There is also scholarship that offers a fourth rationale, which includes due process and rule-of-law concerns. We set these concerns aside because they pertain to government use of algorithms, while this Article focuses on regulation of the private sector. See Brennan-Marquez, *supra* note 19, at 1288–94 (discussing “rule-of-law” principles with respect to police and judicial actions); Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 GEO. L.J. 1147, 1184–90, 1206–09 (2017) (discussing due process and reason-giving in administrative law); ECLT Seminars, [HUML16] 03: Katherine Strandburg, *Decision-Making, Machine Learning and the Value of Explanation*, YOUTUBE (Jan. 23, 2017), <https://www.youtube.com/>

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.

The black-box-only approach is limited for the purposes of justifying decision-making. The first two beliefs are not about justifying decisions at all, and therefore serve a different purpose. The third is explicitly about justification, so our critique is directed not at its intent, but its operation. For those concerned with the justification for decision-making, the goal of explanation should be to find a way to bring intuition to bear in deciding whether the model is well justified. This Part explains both the power and limitations of such an approach.

A. The Value of Opening the Black Box

This Part identifies and elaborates the three rationales that apparently underlie most of the popular and scholarly calls for explanation.

1. Explanation as Inherent Good

There are several reasons to view explanation as a good unto itself, and perhaps a necessary part of a system constrained by law, including a respect for autonomy, dignity, and personhood.²¹³ There is a fundamental difference between wanting an explanation for its own sake and wanting an explanation for the purpose of vindicating certain specific empowerment or accountability goals. Fears about a system that lacks explanation are visceral. This fear is best exemplified in popular consciousness by Franz Kafka’s *The Trial*,²¹⁴ a story about a faceless bureaucracy that makes consequential decisions without input or understanding from those affected.²¹⁵

This concern certainly motivates some lawmakers and scholars. In his article, “Privacy and Power,” Daniel Solove refers to this as a “dehumanizing” state of affairs characterized by the “powerlessness and vulnerability created by people’s lack of any meaningful form of participation” in the decision.²¹⁶ David Luban, Alan Strudler, and David Wasserman argue that “one central aspect of the common good”—which they argue forms the basis of law’s legitimacy—“lies in what we might call the *moral intelligibility* of our lives” and that the “horror of the bureaucratic process lies not in officials’ mechanical adherence to duty, but rather in the

watch?v=LQj3nbfSkrU [<https://perma.cc/CX7S-GCUG>] (discussing procedural due process and explanations).

213. See Lawrence B. Solum, *Legal Personhood for Artificial Intelligences*, 70 N.C. L. REV. 1231, 1238–39 (1992) (explaining that while “person” usually means human being in the law, “personhood” is a question of the attendant “bundle of rights and duties”).

214. FRANZ KAFKA, *DER PROCESS* (1925).

215. See Daniel J. Solove, *Privacy and Power: Computer Databases and Metaphors for Information Privacy*, 53 STAN. L. REV. 1393, 1397–98 (2001) (arguing that Kafka’s *The Trial* is a better metaphor than George Orwell’s *1984* for modern anxieties over data).

216. *Id.* at 1423.

individual's ignorance of what the fulfillment of his or her duty may entail.”²¹⁷ The concerns of dignity and personhood certainly motivate the data protection regime in Europe,²¹⁸ if less directly the law in the United States.²¹⁹

We lack the space (and the expertise) to do proper justice to the personhood argument for explanation. Accordingly, our goal here is to flag it and set it aside as a concern parallel to our broader concerns about enabling justifications for automated decisions.

To the extent that the personhood rationale can be converted to a more actionable legal issue, it is reflected in the concept of “procedural justice,” which was most famously championed by Tom Tyler. Procedural justice is the essential quality of a legal system that shows respect for its participants, which might entail transparency, consistency, or even politeness.²²⁰ Tyler and others have shown that people care deeply about procedural justice, to the point that they might find a proceeding more tolerable and fair if their procedural-justice concerns are satisfied even if they do not obtain their preferred outcome in the proceeding.²²¹ Procedural justice, Tyler argues, is necessary on a large scale because it allows people to buy into the legal system and voluntarily comply with the law, both of which are essential parts of a working and legitimate legal system.²²² Presumably, to the extent that automated decisions can be legally or morally justified, people must accept them rather than have them imposed, and as a result, the personhood rationale for model explanation also implicates procedural justice.

Ultimately, that there is inherent value in explanation is clear. But as a practical matter, those concerns are difficult to administer, quantify, and compare to other concerns. Where there are genuine trade-offs between explanation and other normative values such as accuracy or fairness, the inherent value of explanation neither automatically trumps competing considerations nor provides much guidance as to the type of explanation required. Therefore, while inherent value cannot be ignored, other rationales remain important.

217. David Luban, Alan Strudler & David Wasserman, *Moral Responsibility in the Age of Bureaucracy*, 90 MICH. L. REV. 2348, 2354 (1992).

218. Lee A. Bygrave, *Minding the Machine: Article 15 of the EC Data Protection Directive and Automated Profiling*, 17 COMPUTER L. & SECURITY REP. 17, 19 (2001); Meg Leta Jones, *The Right to a Human in the Loop: Political Constructions of Computer Automation and Personhood*, 47 SOC. STUD. SCI. 216, 223–24 (2017).

219. See James Q. Whitman, *The Two Western Cultures of Privacy: Dignity Versus Liberty*, 113 YALE L.J. 1151, 1214–15 (2004).

220. Tom R. Tyler, *What Is Procedural Justice?: Criteria Used by Citizens to Assess the Fairness of Procedures*, 22 LAW & SOC’Y REV. 103, 132 (1988).

221. See, e.g., Tom R. Tyler, *Procedural Justice, Legitimacy, and the Effective Rule of Law*, 30 CRIME & JUST. 283, 291 (2003); Tyler, *supra* note 220, at 128.

222. TOM R. TYLER, WHY PEOPLE OBEY THE LAW 6–7 (2006).

2. Explanation as Enabling Action

For others, the purpose of explanation extends to providing actionable information about the rendering of decisions, such that affected parties can learn if and how they might achieve a different outcome. Explanations are valuable, on this account, because they empower people to effectively navigate the decision-making process. Such beliefs are evident in the adverse action notice requirements of credit-scoring regulations,²²³ but they have come to dominate more recent debates about the regulatory function of requiring explanations of model-driven decisions more generally.

Across a series of recent papers, the debate has coalesced around two distinct, but related, questions. The first is whether and when the GDPR requires explanations of the logic or outcome of decision-making. The second is how to best explain outcomes in an actionable way.

The first question, whether to focus on outcome- or logic-based explanations, originates with an article by Sandra Wachter, Brent Mittelstadt, and Luciano Floridi.²²⁴ These scholars split explanations between “system functionality” and “specific decisions”—a distinction functionally similar to our outcome- and logic-based framework.²²⁵ This mirrors the debate in the technical community about the best way to understand the meaning of interpretability. As described in Part II.B, the main split is whether to aim for interpretable models or to account for specific decisions. Drawing together the legal and machine learning literature, Lilian Edwards and Michael Veale have created a similar, but slightly altered distinction between “model-centric” and “subject-centric” explanations.²²⁶ While not identical, subject-centric explanations are another way to explain specific outcomes to individuals.²²⁷

As the discussion has evolved in both the legal and computer science scholarship, new work has converged on the belief that explaining specific outcomes is the right approach. The debate has therefore shifted to the

223. See *supra* Part II.A.1.

224. Wachter et al., *supra* note 23.

225. *Id.* at 78. As Wachter and colleagues define it, system functionality is “the logic, significance, envisaged consequences, and general functionality of an automated decision-making system,” and explanations of specific decisions are “the rationale, reasons, and individual circumstances of a specific automated decision.” *Id.* While the distinction is broadly useful, our definitions differ from theirs and we believe the line between outcome- and logic-based explanations is less clear than they suggest. See Selbst & Powles, *supra* note 90, at 239 (arguing that, given the input data, a description of the logic will provide a data subject with the means to determine any particular outcome, and thus, explanations of the logic will often also explain individual outcomes).

226. Edwards & Veale, *supra* note 143, at 55–56. They define these terms as follows: “Model-centric explanations (MCEs) provide broad information about a [machine learning] model which is not decision or input-data specific,” while “[s]ubject-centric explanations (SCEs) are built on and around the basis of an input record.”

227. Ultimately, Edwards and Veale argue, as we do, that the explanation debate had been restricted to this question. *Id.* Recognizing that explanations are no panacea, the rest of their paper argues that the GDPR provides tools other than a right to explanation that could be more useful for algorithmic accountability.

second question, which focuses on the many different methods by which outcomes can be explained.

An interdisciplinary working group at the Berkman Klein Center for Internet and Society begin by recognizing that explanations are infinitely variable in concept, but claim that “[w]hen we talk about an explanation for a decision, . . . we generally mean the reasons or justifications for that particular outcome, rather than a description of the decision-making process in general.”²²⁸ They propose three ways to examine a specific decision: (1) the main factors in a decision, (2) the minimum change required to switch the outcome of a decision, and (3) the explanations for similar cases with divergent outcomes or divergent cases with similar outcomes.²²⁹ Wachter, Mittelstadt, and Chris Russell have a still narrower focus, writing about counterfactual explanations that represent “the smallest change to the world” that would result in a different answer.²³⁰ They envision a distance metric where, if one were to plot all n features in an n -dimensional space, the counterfactual is the shortest “distance” from the data subject’s point in the space (defined by the values of the features she possesses) to the surface that makes up the outer edge of a desirable outcome.²³¹

Accordingly, counterfactual explanations are seen as fulfilling the three goals of explanations discussed in this Part: (1) to help an individual understand a decision, (2) to enable that individual to take steps to achieve a better outcome, and (3) to provide a basis for contesting the decision.²³² When applying the strategy of counterfactual explanations, however, it is clear that most of the value comes from the second rationale: actionable explanations. Wachter and colleagues assert that counterfactual explanations are an improvement over the existing requirements of the GDPR because, as a matter of positive law, the Regulation requires almost nothing except a “meaningful overview,” which can be encapsulated via pictorial “icons” depicting the type of data processing in question.²³³ Counterfactual explanations, in contrast, offer something specific to the data subject and will thus be more useful in informing an effective response. But if their interpretation of the law is correct—that the GDPR requires no

228. Doshi-Velez & Kortz, *supra* note 201, at 2.

229. *Id.* at 3.

230. Wachter et al., *supra* note 143, at 845.

231. *Id.* at 850–54. Distance metrics are a way to solve this problem. Hall and colleagues describe another distance metric that is used in practice. Hall et al., *supra* note 120. They employ a distance metric to identify the features that need to change the *most* to turn a credit applicant into the ideal applicant. *Id.* Alternatively, other methods could be identifying the features over which a consumer has the most control, the features that would cost a consumer the least to change, or the features least coupled to other life outcomes and thus easier to isolate. The main point is that the law provides no formal guidance as to the proper metric for determining what reasons are most salient, and this part of the debate attempts to resolve this question. See 12 C.F.R. § 1002.9 supp. I (2018).

232. Wachter et al., *supra* note 143, at 843.

233. *Id.* at 865.

explanation²³⁴—then their claim is that counterfactuals offer more than literally nothing, which is not saying much. On contestability, Wachter, Mittelstadt, and Russell ultimately concede that to contest a decision, it is likely necessary to understand the logic of decision-making rather than to just have a counterfactual explanation of a specific decision.²³⁵ The real value, then, of their intervention and others like it, is to better allow data subjects to alter their behavior when a counterfactual suggests that a decision is based on alterable characteristics.²³⁶

Empowering people to navigate the algorithms that affect their lives is an important goal and has genuine value. This is a pragmatic response to a difficult problem, but it casts the goal of explanations as something quite limited: ensuring people know the rules of the game so they can play it better. This approach is not oriented around asking if the basis of decisions is well justified; rather it takes decisions as a given and seeks to allow those affected by them to avoid or work around bad outcomes.²³⁷ Rather than using explanations to ask about the justifications for decision-making, this approach shifts responsibility for bad outcomes from the designers of automated decisions to those affected by them.²³⁸

3. Explanation as Exposing a Basis for Evaluation

The final value ascribed to explanation is that it forces the basis of decision-making into the open and thus provides a way to question the validity and justifiability of making decisions on these grounds. As Pauline Kim has observed:

234. The positive law debate about the right to explanation is not the subject of this Article, but suffice it to say, there is a healthy debate about it in the literature. See *supra* note 143 and accompanying text for a discussion.

235. Wachter et al., *supra* note 143, at 878. Their one example where a counterfactual can lead to the ability to contest a decision is based on data being inaccurate or missing rather than based on the inferences made. Thus, it is actually the rare situation specifically envisioned by FCRA, where the adverse action notice reveals that a decision took inaccurate information into account. Because of the deficiencies of the FCRA approach, discussed *supra* in Part II.A, this will not solve the general problem.

236. As Berk Ustun and colleagues point out, an explanation generated by counterfactual techniques will not necessarily be actionable unless intentionally structured to be so. Berk Ustun et al., Actionable Recourse in Linear Classification 2 (Sept. 18, 2018) (unpublished manuscript), <https://arxiv.org/abs/1809.06514> [<https://perma.cc/RPJ4-P4AP>].

237. Mireille Hildebrandt, *Primitives of Legal Protection in the Era of Data-Driven Platforms*, 2 GEO. L. TECH. REV. 252, 271 (2018) (“Though it is important that decisions of automated systems can be explained (whether ex ante or ex post; whether individually or at a generic level), we must keep in mind that in the end what counts is whether such decisions can be justified.”).

238. This is remarkably similar to the longstanding privacy and data protection debate around notice and consent, where the goal of notice is to better inform consumers and data subjects, and the assumption is that better information will lead to preferable results. See generally Daniel J. Solove, *Privacy Self-Management and the Consent Dilemma*, 126 HARV. L. REV. 1880 (2013). In reality, this often fails to protect privacy because it construes privacy as a matter of individual decision-making that a person can choose to protect rather than something that can be affected by others with more power. See, e.g., Roger Ford, *Unilateral Invasions of Privacy*, 91 NOTRE DAME L. REV. 1075 (2016).

When a model is interpretable, debate may ensue over whether its use is justified, but it is at least possible to have a conversation about whether relying on the behaviors or attributes that drive the outcomes is normatively acceptable. When a model is not interpretable, however, it is not even possible to have the conversation.²³⁹

But what does it mean to have a conversation based on what an interpretable model reveals?

In a seminal study, Rich Caruana and colleagues provide an answer to that question.²⁴⁰ They discovered that a model trained to predict complications from pneumonia had learned to associate asthma with a reduced risk of death.²⁴¹ To anyone with a passing knowledge of asthma and pneumonia, this result was obviously wrong. The model was trained on clinical data from past pneumonia patients, and it turns out that patients who suffer from asthma truly did end up with better outcomes.²⁴² What the model missed was that these patients regularly monitored their breathing, causing them to go to the hospital earlier.²⁴³ Then, once at the hospital, they were considered higher risk, so they received more immediate and focused treatment.²⁴⁴ Caruana and colleagues drew a general lesson from this experience: to avoid learning artifacts in the data, the model should be sufficiently simple that experts can inspect the relationships uncovered to determine if they correspond with domain knowledge. Thus, on this account, the purpose of explanation is to permit experts to check the model against their intuition.

This approach assumes that when a model is made intelligible, experts can assess whether the relationships uncovered by the model seem appropriate, given their background knowledge of the phenomenon being modeled. This was indeed the case for asthma, but this is not the general case. Often, rather than assigning significance to features in a way that is obviously right or wrong, a model will uncover a relationship that is simply perceived as strange. For example, if the hospital's data did not reveal a dependence on an asthma diagnosis—which is clearly linked to pneumonia through breathing—but rather revealed a dependence on skin cancer, it would be less obvious what to make of that fact. It would be wrong to simply dismiss it as an artifact of the data, but it also does not fit with any intuitive story even a domain expert could tell.

Another example of this view of explanation is the approach to interpretability known as Local Interpretable Model-Agnostic Explanations (“LIME”).²⁴⁵ It has generated one of the canonical examples of the value of

239. Kim, *supra* note 4, at 922–23.

240. Rich Caruana et al., *Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission*, in PROCEEDINGS OF THE 21TH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING 1721, 1721 (2015).

241. *Id.*

242. *Id.*

243. *Id.*

244. *Id.*

245. Ribeiro et al., *supra* note 197. This is one of the methods described *supra* in Part II.B.2.

interpretability in machine learning. Marco Ribeiro and colleagues used LIME to investigate a deep-learning model trained to distinguish images of wolves from huskies. The authors discovered that the model did not rely primarily on the animals' features, but on whether snow appeared in the background of a photo.²⁴⁶

There are three reasons this is such a compelling example. First, what LIME identified as the distinguishing feature—snow—is legible to humans. Second, this feature is obviously not a property of the category “wolf.” Third, humans can tell a story about why this mistake occurred: wolves are more likely to be found in an environment with snow on the ground. Although this story may not actually be true, the important point is that we can convince ourselves it is.²⁴⁷ Like the asthma example, the ability to determine that the model has overfit the training data relies on the inherent legibility of the relevant feature, the existence of background knowledge about that feature, and our ability to use the background knowledge to tell a story about why the feature is important. In this example, the realization relies on something closer to common sense than to specialized expertise, but the explanation serves the same function—to allow observers to bring their intuition to bear in evaluating the model.

The final examples come from James Grimmelmann and Daniel Westreich,²⁴⁸ as well as Kim, whose work was discussed earlier.²⁴⁹ Grimmelmann and Westreich imagine a scenario in which a model learns to distinguish between job applicants on the basis of a feature—musical taste—that is both correlated with job performance and membership in a protected class.²⁵⁰ They further stipulate that job performance varies by class membership.²⁵¹ As they see it, this poses the challenge of determining whether the model, by relying on musical tastes, is in fact relying on protected-class membership.²⁵²

Grimmelmann and Westreich then argue that if one cannot tell a story about why musical taste correlates with job performance, the model must be learning something else.²⁵³ They propose a default rule that the “something else” be considered membership in a protected class unless it can be shown

246. Ribeiro et al., *supra* note 197, at 1142–43. This is a textbook example of overfitting the training data.

247. In fact, while writing this section, we remembered the finding, but until we consulted the original source we disagreed with each other about whether the wolves or huskies were the ones pictured in snow. This suggests that the story would have been equally compelling if the error had been reversed.

248. Grimmelmann & Westreich, *supra* note 75.

249. Kim, *supra* note 4.

250. Grimmelmann & Westreich, *supra* note 75, at 166–67.

251. *Id.* at 167.

252. The only reason a model would learn to do this is if: (1) class membership accounts for all the variance in the outcome of interest or (2) class membership accounts for more of the variance than the input features. In the second case, the easy fix would be to include a richer set of features until class membership no longer communicates any useful information. The only way that adding features could have this effect, though, is if the original model was necessarily less than perfectly accurate, in which case a better model should have been used.

253. Grimmelmann & Westreich, *supra* note 75, at 174.

otherwise, specifically by the defendant.²⁵⁴ The problem with this reasoning is that the model might not be learning protected-class membership, but a different latent variable that explains the relationship between musical taste and job performance—an unobserved or unknown characteristic that affects both musical taste and job performance. By assuming that it should be possible to tell a story about such a variable if it exists, they—as in the examples above—fail to account for the possibility of a strange, but legitimate, result. They use the ability to tell a story as a proxy for the legitimacy of the decision-making, but that only works if a justification, or lack thereof, immediately falls out of the description, as it did in the asthma and snow examples.

Kim uses a real example to make a similar point. She cites a study stating that employees who installed web browsers that did not come with their computers stay longer on their job.²⁵⁵ She then speculates that either there is an unobserved variable that would explain the relationship or it is “entirely coincidental.”²⁵⁶ To Kim, what determines whether the relationship is “substantively meaningful” rather than a mere statistical coincidence is whether we can successfully tell ourselves such stories.²⁵⁷ Like Grimmelmann and Westreich, for Kim, if no such story can be told, and the model has a disparate impact, it should be illegal.²⁵⁸ What these examples demonstrate is that, whether one seeks to adjudicate model validity or normative justifications, intuition actually plays the same role.

Unlike the first two values of explanation, this approach has the ultimate goal of evaluating whether the basis of decision-making is well justified. It does not, however, ask the question: “Why are these the rules?” Instead, it makes two moves. The first two examples answered the question, “What are the rules?” and expected that intuition will furnish an answer for both why the rules are what they are and whether they are justified. The latter two examples instead argued that decisions should be legally restricted to intuitive relationships. Such a restriction short-circuits the need to *ask* why the rules are what they are by guaranteeing up front that an answer will be available.²⁵⁹

254. *Id.* at 173.

255. Kim, *supra* note 4, at 922.

256. *Id.* So too did the chief analytics officer in the company involved, in an interview. Joe Pinsker, *People Who Use Firefox or Chrome Are Better Employees*, ATLANTIC (Mar. 16, 2015), <https://www.theatlantic.com/business/archive/2015/03/people-who-use-firefox-or-chrome-are-better-employees/387781/> [https://perma.cc/3MYM-SXAQ] (“I think that the fact that you took the time to install Firefox on your computer shows us something about you. It shows that you’re someone who is an informed consumer,’ he told Freakonomics Radio. ‘You’ve made an active choice to do something that wasn’t default.’”).

257. Kim, *supra* note 4, at 917.

258. *Id.*

259. This might also explain the frequent turn to causality as a solution. Restricting the model to causal relationships also short-circuits the need to ask the “why” question because the causal mechanism is the answer. Ironically, a causal model need not be intuitive, so it may not satisfy the same normative desires as intuition seems to. See *supra* note 78.

These two approaches are similar, but differ in the default rule they apply to strange cases. In the case of the two technical examples, the assumption is that obviously *flawed* relationships will present themselves and should be overruled; relationships for which there is no intuitive explanation may remain. The two legal examples, by contrast, are more conservative. They presume that obviously *correct* relationships will show themselves, so that everything else should be discarded by default, while allowing for the possibility of defeating such a presumption. Both are forced to rely on default rules to handle strange, but potentially legitimate, cases because the fundamental reliance on intuition does not give them tools to evaluate these cases.

B. Evaluating Intuition

Much of the anxiety around inscrutable models comes from the legal world's demands for justifiable decision-making. That decisions based on machine learning reflect the particular patterns in the training data cannot be a sufficient explanation for why a decision is made the way it is. Evaluating whether some basis for decision-making is fair, for example, will require tools that go beyond standard technical tests of validity that would already have been applied to the model during its development.²⁶⁰ While the law gives these tests some credence, reliance on accuracy is not normatively adequate with respect to machine learning.²⁶¹

For many, the presumed solution is requiring machine learning models to be intelligible.²⁶² What the prior discussion demonstrates, though, is that this presumption works on a very specific line of reasoning that is based on the idea that with enough explanation, we can bring intuition to bear in evaluating decision-making. As Kim observes:

Even when a model is interpretable, its *meaning* may not be clear. Two variables may be strongly correlated in the data, but the existence of a statistical relationship does not tell us if the variables are causally related, or are influenced by some common unobservable factor, or are completely unrelated.²⁶³

260. Even among practitioners, the interest in interpretability stems from warranted suspicion of the power of validation; there are countless reasons why assessing the likely performance of a model against an out-of-sample test set will fail to accurately predict a model's real-world performance. Yet even with these deep suspicions, practitioners still believe in validation as the primary method by which the use of models can and should be justified. See Hand, *supra* note 182, at 12–13. In contrast, the law has concerns that are broader than real-world performance, which demand very different justifications for the basis of decision-making encoded in machine learning models.

261. Barocas & Selbst, *supra* note 4, at 673 (“[T]he process can result in disproportionately adverse outcomes concentrated within historically disadvantaged groups in ways that look a lot like discrimination.”).

262. See Brennan-Marquez, *supra* note 19, at 1253; Grimmelmann & Westreich, *supra* note 75, at 173; Kim, *supra* note 4, at 921–22.

263. Kim, *supra* note 4, at 922.

Her response is to constrain the model to features that bear an intuitive relationship to the outcome.²⁶⁴

This way of thinking originates in disparate impact doctrine, which—among several ways of describing the requirement—calls for an employment test to have a “manifest relationship” to future job performance.²⁶⁵ But there is a difference between a manifest relationship of a model to job performance and a manifest relationship of a particular *feature* to job performance. Models can be shown to have a manifest relationship to job performance if the *target variable* is manifestly related to job performance and the model is statistically valid. This is true even if none of the individual *features* are manifestly related.²⁶⁶ People who advocate for a nexus between features and the outcome are dissatisfied with a purely statistical test and want some other basis to subject a model to normative assessment. Models must be restricted to intuitive relationships, the logic goes, so that such a basis will exist.

Regulatory guidance evinces similar reasoning. In 2011, the Federal Reserve issued formal guidance on model risk management.²⁶⁷ The purpose of the document was to expand on prior guidance that was limited to model validation.²⁶⁸ The guidance notes that models “may be used incorrectly or inappropriately” and that banks need diverse methods to evaluate them beyond statistical validation.²⁶⁹ Among other recommendations discussed in Part IV, the guidance recommends “outcomes analysis,” which calls for “expert judgment to check the intuition behind the outcomes and confirm that the results make sense.”²⁷⁰

In an advisory bulletin about new financial technology, the Federal Reserve Board recommended that individual features have a “nexus” with creditworthiness to avoid discriminating in violation of fair lending laws.²⁷¹ In their view, a nexus enables a “careful analysis” about the features assigned

264. *Id.*; cf. Nick Seaver, *Algorithms as Culture*, BIG DATA & SOC’Y, July–Dec. 2017, at 6 (“To make something [accountable] means giving it qualities that make it legible to groups of people in specific contexts. An accountable algorithm is thus literally different from an unaccountable one—transparency changes the practices that constitute it. For some critics, this is precisely the point: the changes that transparency necessitates are changes that we want to have.”).

265. Baracas & Selbst, *supra* note 4, at 702 (“A challenged employment practice must be ‘shown to be related to job performance,’ have a ‘manifest relationship to the employment in question,’ be ‘demonstrably a reasonable measure of job performance, bear some relationship to job-performance ability,’ []or ‘must measure the person for the job and not the person in the abstract.’” (quoting Linda Lye, Comment, *Title VII’s Tangled Tale: The Erosion and Confusion of Disparate Impact and the Business Necessity Defense*, 19 BERKELEY J. EMP. & LAB. L. 315, 321 (1998) (footnotes omitted) (quoting Griggs v. Duke Power Co., 401 U.S. 424 (1971))).

266. *Id.* at 708.

267. BD. OF GOVERNORS OF THE FED. RESERVE SYS., OFFICE OF THE COMPTROLLER OF THE CURRENCY, SR LETTER 11-7, SUPERVISORY GUIDANCE ON MODEL RISK MANAGEMENT (2011), <https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf> [https://perma.cc/AR85-AASR].

268. *Id.* at 2.

269. *Id.* at 4.

270. *Id.* at 13–14.

271. Evans, *supra* note 121, at 4.

significance in a model predicting creditworthiness.²⁷² Here, intuitiveness is read into ECOA as a natural requirement of having to justify decision-making that generates a disparate impact via the “business-necessity” defense.²⁷³ The business-necessity defense asks whether the particular decision-making mechanism has a tight enough fit with the legitimate trait being predicted²⁷⁴ and whether there were equally effective but less discriminatory ways to accomplish the same task. With a model that lacks intuitive relationships, a plaintiff could argue that the model is indirectly—and thus poorly—measuring some latent and more sensible variable that should serve as the actual basis of decision-making. The Federal Reserve Board guidance suggests that one way to avoid an uncertain result in such litigation is to limit decision-making to features that bear an intuitive—and therefore justifiable—relationship to the outcome of interest. While it is not clear that relying on proxies for an unrecognized latent variable presents problems under current disparate impact doctrine,²⁷⁵ the guidance treats an intuition requirement as a prophylactic. This reasoning seems to underlie the recommendations of Kim as well as Grimmelmann and Westreich.

What should be clear by now is that intuition is the typical bridge from explanation to normative assessment. This can be a good thing. Intuition is powerful. It is a ready mechanism by which considerable knowledge can be brought to bear in evaluating machine learning models. Such models are myopic, having visibility into only the data upon which they were trained.²⁷⁶ Humans, in contrast, have a wealth of insights accumulated through a broad range of experiences, typically described as “common sense.” This knowledge allows us to immediately identify and discount patterns that violate our well-honed expectations and to recognize and affirm discoveries that align with experience. In fact, intuition is so powerful that humans cannot resist speculating about latent variables or causal mechanisms when confronted by unexplained phenomena.

Intuition can also take the form of domain expertise, which further strengthens the capacity to see where models may have gone awry. The social sciences have a long history of relying on face validity to determine whether a model is measuring what it purports to measure.²⁷⁷ A model that assigns significance to variables that seem facially irrelevant is given little credence or is subject to greater scrutiny. Such a practice might seem ad hoc, but questioning face validity is a fundamental part of the social-scientific

272. *Id.*

273. It is interesting that the demand for intuitiveness, on this account, comes not from the procedural requirements of the adverse action notices—the part of ECOA most obviously concerned with explanations—but from the substantive concerns of disparate impact doctrine.

274. See, e.g., *Watson v. Fort Worth Bank & Tr.*, 487 U.S. 977, 1010 (1988) (Blackmun, J., concurring) (explaining that a business-necessity defense must be carefully tailored to objective, relevant job qualifications).

275. See Baracas & Selbst, *supra* note 4, at 709–10 (discussing the problems with the “fix-the-model” approach to alternative practice claims).

276. Andrew D. Selbst, *A Mild Defense of Our New Machine Overlords*, 70 VAND. L. REV. EN BANC 87, 101 (2017).

277. See *supra* note 73 and accompanying text.

process. Crucially, intuition allows us to generate competing explanations that account for the observed facts and to debate their plausibility.²⁷⁸

Importantly, however, intuition has its downsides. Most immediately, it can be wrong. It can lead us to discount valid models because they are unexpected or unfamiliar, or to endorse false discoveries because they align with existing beliefs.²⁷⁹ Intuition encourages us to generate “just so” stories that appear to make good sense of the presented facts. Such stories may feel coherent but are actually unreliable. In fact, the rich literature on cognitive biases—including the “narrative fallacy”—is really an account of the dangers of intuition.²⁸⁰ While intuition is helpful for assessing evidently good and bad results, it is less useful when dealing with findings that do not comport with or even run counter to experience. The overriding power of intuition means that strange results will stand out, but intuition may not point in a productive direction for making these any more sensible.

This is a particularly pronounced problem in the case of machine learning, as its value lies largely in finding patterns that go well beyond human intuition. The problem in such cases is not only that machine learning models might depart from intuition, but that they might not even lend themselves to *hypotheses* about what accounts for the models’ discoveries. Parsimonious models lend themselves to more intuitive reasoning, but they have limits—a complex world may require complex models. In some cases, machine learning will have the power to detect the subtle patterns and intricate dependencies that can better account for reality.

If the interest in explanation stems from its intrinsic or pragmatic value, then addressing inscrutability is worthwhile for its own sake. But if we are interested in whether models are well justified, then addressing inscrutability only gets us part of the way. We should consider how else to justify models. We should think outside the black box and return to the question: Why are these the rules?

IV. DOCUMENTATION AS EXPLANATION

Limiting explanation of a model to its internal mechanics forces us to rely on intuition to guess at why the model’s rules are what they are. But what would it look like for regulation to directly seek an answer to that question? By now, it is well understood that data are human constructs²⁸¹ and that subjective decisions pervade the modeling and decision-making process.²⁸²

278. See, e.g., Brennan-Marquez, *supra* note 19; Michael Pardo & Ronald J. Allen, *Juridical Proof and the Best Explanation*, 27 LAW & PHIL. 223, 230 (2008).

279. Raymond S. Nickerson, *Confirmation Bias: A Ubiquitous Phenomenon in Many Guises*, 2 REV. GEN. PSYCHOL. 175, 175 (1998).

280. See generally KAHNEMAN, *supra* note 77.

281. Lisa Gitelman & Virginia Jackson, *Introduction to RAW DATA IS AN OXYMORON* 1, 3 (Lisa Gitelman ed., 2013); see also danah boyd & Kate Crawford, *Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon*, 15 INFO., COMM. & SOC’Y 662, 666–68 (2012).

282. Baracas & Selbst, *supra* note 4, at 673; see also Seaver, *supra* note 264, at 5.

Explaining why the model works as it does requires accounting for these decisions.

Furnishing such answers will require process, documentation, and access to that documentation. This can be done in a public format, with impact assessments, or companies can do it privately, with access triggered on some basis, like discovery in litigation.

A. The Information Needed to Evaluate Models

When we seek to evaluate the justifications for decision-making that relies on a machine learning model, we are actually asking about the institutional and subjective process behind its development. The Federal Reserve Board guidance discussed in Part III.B moves in this direction by recommending documentation, but its approach appears to be mostly about validation—how to validate well, thoroughly, on an ongoing basis, and in preparation for a future legal challenge.²⁸³ Careful validation is essential and nontrivial,²⁸⁴ but it is also not enough. Normatively evaluating decision-making requires, at least, an understanding of: (1) the values and constraints that shape the conceptualization of the problem, (2) how these values and constraints inform the development of machine learning models and are ultimately reflected in them, and (3) how the outputs of models inform final decisions.

To illustrate how each of these components work, consider credit scoring. What are the values embedded in credit-scoring models and under what constraints do developers operate? Lenders could attempt to achieve different objectives with credit scoring at the outset: Credit scoring could aim to ensure that all credit is ultimately repaid, thus minimizing default. Lenders could use credit scoring to maximize profit. Lenders could also seek to find ways to offer credit specifically to otherwise overlooked applicants, as many firms engaged in alternative credit scoring seek to do. Each of these different goals reflects different core values, but other value judgments might be buried in the projects as well. For example, a creditor could be morally committed to offering credit as widely as possible, while for others that does not factor into the decision. Or a creditor's approach to regulation could be to either get away with as much as possible or steer far clear of regulatory scrutiny. Each of these subjective judgments will ultimately inform the way a project of credit scoring is conceived.

The developers of credit-scoring models will also face constraints and trade-offs. For example, there might be limits on available talent with both domain expertise and the necessary technical skills to build models. Models might be better informed if there were much more data available, even though

283. BD. OF GOVERNORS OF THE FED. RESERVE SYS., *supra* note 267. The guidance wants developers to consider where the data comes from, whether it suffers from bias, whether the model is robust to new situations, whether due care has been taken with respect to potential limitations and outright faults with the model, and so on. *Id.* at 5–16; see also Edwards & Veale, *supra* note 143, at 55–56; Pauline T. Kim, *Auditing Algorithms for Discrimination*, 166 U. PA. L. REV. ONLINE 189, 196 (2017).

284. Barocas & Selbst, *supra* note 4, at 680–92.

there are practical challenges to collecting so much data. Ultimately, both trade-offs are issues of cost,²⁸⁵ but they include more practical realities as well, such as limitations on talent in the geographical area of the firm or privacy concerns that limit the collection of more data. How to deal with these trade-offs is a judgment call every firm will have to make.²⁸⁶

Another cost-related trade-off is competition. Before credit scoring was popular, creditors used to work with borrowers over the lifetime of the loan to ensure repayment; credit scores first took hold in banks as a way to reduce the cost of this practice.²⁸⁷ Creditors today *could* return to that model, but it would likely involve offering higher interest rates across the board to account for increased operating costs, perhaps pushing such a firm out of the market. As a result, competition operates as a constraint that ultimately changes the decision process.

The values of and constraints faced by a firm will lead to certain choices about how to build and use models. As we have discussed in prior work, the subjective choices a developer makes include choosing target variables, collecting training data, labeling examples, and choosing features.²⁸⁸ Developers must also make choices about other parts of the process, such as how to treat outliers, how to partition their data for testing, what learning algorithms to choose, and how and how much to tune the model, among other things.²⁸⁹ The act of developing models is quite complex and involves many subjective decisions by the developers.

In the credit example, the values discussed above may manifest in the model in several ways. For example, consider the different project objectives discussed above. If a firm seeks to maximize profit, it may employ a model with a different target variable than a firm that seeks to minimize defaults. The target variable is often the outcome that the model developers want to maximize or minimize, so in the profit-seeking case, it would be *expected profit per applicant*, and in the risk-based case, it could be *likelihood of default*. While the alternative credit-scoring model hypothesized above might rely on the same likelihood-of-default target variable, firms' values are likely to influence the type of data they collect; they might seek alternative data sources, for example, because they are trying to reach underserved populations. In addition to the values embedded a priori, the values of the firms dictate how they resolve the different constraints they face—for example, cost and competition. The traditional credit scorers tend to not

285. See FREDERICK SCHAUER, PROFILES, PROBABILITIES, AND STEREOTYPES 124–26 (2003).

286. Watson v. Fort Worth Bank & Tr., 487 U.S. 977, 998 (1988) (plurality opinion) (considering costs and other burdens relevant to a discrimination case).

287. Martha Ann Poon, What Lenders See—a History of the Fair Isaac Scorecard 109, 120 (Jan. 1, 2012) (unpublished Ph.D. dissertation, University of California, San Diego), <https://cloudfront.escholarship.org/dist/prd/content/qt7n1369x2/qt7n1369x2.pdf?t=o94tcd> [<https://perma.cc/V24B-8G3M>].

288. Barocas & Selbst, *supra* note 4, at 677–92.

289. Lehr & Ohm, *supra* note 51, at 683–700; see also Brian d’Alessandro, Cathy O’Neil & Tom LaGatta, *Conscientious Classification: A Data Scientist’s Guide to Discrimination-Aware Classification*, 5 BIG DATA 120, 125 (2017).

make the extra effort or spend the extra money to obtain the data needed to make predictions about people on the margins of society.²⁹⁰ There is also regulatory uncertainty regarding the permissibility of new types of credit data.²⁹¹ Therefore, their models reflect the fact that the developers are more sensitive to cost and regulatory penalty than inclusion.

Models are not self-executing; an additional layer of decisions concerns the institutional process that surrounds the model. Are the model outputs automatically accepted as the ultimate decisions?²⁹² If not, how central is the model to the decision? How do decision makers integrate the model into their larger decision frameworks? How are they trained to do so? What role does discretion play?

These questions are all external to the model, but they directly impact the model's importance and normative valence. For example, certain creditors may automatically reject applicants with a predicted likelihood of default that exceeds 50 percent.²⁹³ Others, however, may opt to be more inclusive. Perhaps a local credit union that is more familiar with its members and has a community-service mission might decide that human review is necessary for applicants whose likelihood of default sits between 40 percent and 60 percent, leaving the final decision to individual loan officers. A similar creditor might adopt a policy where applicants that the model is not able to score with confidence are subject to human review, especially where the outcome would otherwise be an automatic rejection of members of legally protected classes.

Many of these high-level questions about justifying models or particular uses of models are not about models at all, but whether certain policies are

290. Request for Information Regarding Use of Alternative Data and Modeling Techniques in the Credit Process, 82 Fed. Reg. 11,183, 11,185 (Feb. 21, 2017).

291. *Id.* at 11,187–88.

292. The distinction between models and ultimate decisions is the focus of the GDPR's prohibition on "decision[s] based solely on *automated* processing." Article 29 Data Protection Working Party, *supra* note 149, at 19–22 (emphasis added).

293. This is not how credit typically works in the real world, but for demonstrative purposes, we decided to work with a single hypothetical. In reality, the best examples of this divergence between model and use come from policing and criminal justice. For example, the predictive-policing measure in Chicago, known as the Strategic Subject List, was used to predict the 400 likeliest people in a year to be involved in violent crime. Monica Davey, *Chicago Police Try to Predict Who May Shoot or Be Shot*, N.Y. TIMES (May 23, 2016), <http://www.nytimes.com/2016/05/24/us/armed-with-data-chicago-police-try-to-predict-who-may-shoot-or-be-shot.html> [https://perma.cc/TZ2T-NMEJ]. When Chicago sought funding for the initiative, the city premised it on the idea of providing increased social services to those 400 people, but in the end only targeted them for surveillance. DAVID ROBINSON & LOGAN KOEPKE, STUCK IN A PATTERN: EARLY EVIDENCE ON "PREDICTIVE POLICING" AND CIVIL RIGHTS 9 (2016). The fairness concerns are clearly different between those use cases. See Selbst, *supra* note 4, at 142–44. Similarly, COMPAS, the now-infamous recidivism risk score, was originally designed to figure out who would need greater access to social services upon reentry to reduce the likelihood of rearrest but is now commonly used to decide whom to detain pending trial. Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> [https://perma.cc/F9CK-Z995].

acceptable independent of whether they use machine learning.²⁹⁴ Questions about justifying a model are often just questions about policy in disguise.²⁹⁵ For example, a predatory lender could use the exact same prediction of default to find prime candidates in underserved communities and offer them higher interest rates than they might otherwise receive. This will create more profit because the underserved loan candidates will be more willing to pay a higher rate, but it is clearly predation: interest rates are not being used to offset risk, but to extract maximum profit from vulnerable consumers.²⁹⁶ Most importantly, that this practice is predatory can be judged with no reference to the credit-scoring model.

Evaluating models in a justificatory sense means comparing the reasoning behind the choices made by the developers against society's broader normative priorities, as expressed in law and policy. In order to perform this evaluation, then, documentation about the decisions that lie behind and become part of models must exist and be made available for scrutiny. With an understanding of what that information looks like, the next section begins to explore how to ensure access.

B. Providing the Necessary Information

Assuming the documentation exists, there are numerous ways it can become open to scrutiny. For purposes of demonstration, two are discussed here, although many more are possible: (1) the possibility that documentation is made publicly available from the start and (2) that it becomes accessible upon some trigger, like litigation. The former is essentially an algorithmic impact statement (AIS),²⁹⁷ a proposed variant of the original impact statements required by the National Environmental Policy Act.²⁹⁸ The most common trigger of the latter is a lawsuit, in which documents can be obtained and scrutinized and witnesses can be deposed or examined on the stand, but auditing requirements are another possibility. In both approaches, the coupling of existing documentation with a way to access it create answers to the question of what happened in the design process, with the goal of allowing overseers to determine whether those choices were justifiable. Like FCRA and ECOA, these examples have no inherent

294. See VIRGINIA EUBANKS, AUTOMATING INEQUALITY 37 (2018) ("[W]hen we focus on programs specifically targeted at poor and working-class people, the new regime of data analytics is more evolution than revolution. It is simply an expansion and continuation of moralistic and punitive poverty management strategies that have been with us since the 1820s.").

295. See, e.g., *id.* at 38; Jessica M. Eaglin, *Constructing Recidivism Risk*, 67 EMORY L.J. 59, 99–101 (2017); Margaret Hu, *Algorithmic Jim Crow*, 86 FORDHAM L. REV. 633 (2017); Sonia Katyal, *Algorithmic Civil Rights*, 104 IOWA L. REV. (forthcoming 2018) (draft on file with authors); Sandra G. Mayson, *Dangerous Defendants*, 127 YALE L.J. 490, 507–18 (2018).

296. According to sociologist Jacob Faber, this is actually what happened in the subprime crisis to people of color. Jacob W. Faber, *Racial Dynamics of Subprime Mortgage Lending at the Peak*, 28 HOUSING POL'Y DEBATE 328, 343 (2013).

297. Selbst, *supra* note 4, at 169–93.

298. See 42 U.S.C. § 4332(C) (2012).

connection to machine learning, but the methods can be easily applied in this context.

An impact statement is a document designed to explain the process of decision-making and the anticipated effects of that decision in such a way as to open the process up to the public. Generally, the requirement is designed to ensure that developers do their homework, create a public record, and include public comments.²⁹⁹ Impact statements are an idea that originated in 1970 with the National Environmental Policy Act³⁰⁰ and have since been emulated repeatedly at all levels of government, in many substantive areas of policy.³⁰¹ Aside from environmental law, the federal government requires privacy impact assessments “when developing or procuring information technology systems that include personally identifiable information.”³⁰² Individual states not only have their own legislation requiring environmental impact statements,³⁰³ but also racial impact statements for sentencing policy, among other requirements.³⁰⁴ Recently, led by the ACLU’s Community Control Over Police Surveillance (CCOPS) initiative,³⁰⁵ counties and cities have begun requiring impact statements that apply to police purchases of new technology.³⁰⁶

One of us has argued that a future AIS requirement should be expressly modeled on the environmental impact statement (EIS): the original and most thorough version, with the fullest explanation requirements. Such an impact statement would require thoroughly explaining the types of choices discussed above. This includes direct choices about the model, such as target variables, whether and how new data was collected, and what features were considered. It also requires a discussion of the options that were considered but not chosen, and the reasons for both.³⁰⁷ Those reasons would—either explicitly or implicitly—include discussion of the practical constraints faced by the developers and the values that drove decisions. The AIS must also discuss the predicted impacts of both the chosen and unchosen paths, including the

299. Selbst, *supra* note 4, at 169.

300. 42 U.S.C. §§ 4321–4347.

301. Bradley C. Karkkainen, *Toward a Smarter NEPA: Monitoring and Managing Government’s Environmental Performance*, 102 COLUM. L. REV. 903, 905 (2002).

302. Kenneth A. Bamberger & Deirdre K. Mulligan, *Privacy Decision-Making in Administrative Agencies*, 75 U. CHI. L. REV. 75, 76 (2008).

303. E.g., California Environmental Quality Act (CEQA), CAL. PUB. RES. CODE §§ 21000–21178 (2018).

304. Jessica Erickson, Comment, *Racial Impact Statements: Considering the Consequences of Racial Disproportionalities in the Criminal Justice System*, 89 WASH. L. REV. 1425, 1445 (2014).

305. AN ACT TO PROMOTE TRANSPARENCY AND PROTECT CIVIL RIGHTS AND CIVIL LIBERTIES WITH RESPECT TO SURVEILLANCE TECHNOLOGY § 2(B) (ACLU Jan. 2017), <https://www.aclu.org/files/communitycontrol/ACLU-Local-Surveillance-Technology-Model-City-Council-Bill-January-2017.pdf> [https://perma.cc/AQ8T-3NKM] (ACLU CCOPS Model Bill).

306. See, e.g., SANTA CLARA COUNTY, CAL., CODE OF ORDINANCES § A40-3 (2016).

307. Selbst, *supra* note 4, at 172–75.

possibility of no action, and the effects of any potential mitigation procedures.³⁰⁸

The typical American example of an impact statement is a public document. Thus, a law requiring them would also require that the developers publish the document and allow for comments between the draft and final impact statements.³⁰⁹ Of course, such an idea is more palatable in the case of regulation of public agencies. While disclosure of the kinds of information we describe does not actually imply disclosure of the model itself—obviating the need for a discussion of trade secrets and gaming—firms may still be reluctant to publish an AIS that reveals operating strategy, perceived constraints, or even embedded values. Thus, it is also useful to consider a documentation requirement that allows the prepared documents to remain private but available as needed for accountability.³¹⁰

A provision of the GDPR actually does just this. Article 35 requires “data protection impact assessments” (DPIAs) whenever data processing “is likely to result in a high risk to the rights and freedoms of natural persons.”³¹¹ As Edwards and Veale discuss, the DPIA requirement is very likely to apply to machine learning,³¹² and the assessments require “appropriate technical and organizational measures” to protect data subject rights.³¹³ In Europe, DPIAs are private documents, though making summaries public is officially encouraged.³¹⁴ The European solution to making this private document available is to require consultation with the member state data protection authorities whenever the DPIA indicates a high risk of interference with data subject rights.³¹⁵

One could imagine another way of making an essentially private impact assessment accessible, initiated by private litigation. Interrogatories, depositions, document subpoenas, and trial testimony are all tools that enable litigation parties to question human witnesses and examine documents. These are all chances to directly ask model developers what choices they made and why they made them.

A hypothetical will help clarify how these opportunities, coupled with documentation—whether a DPIA or something similar—differ from the use of intuition as a method of justification. Imagine a new alternative credit-scoring system that relies on social media data.³¹⁶ This model assigns

308. *Id.*

309. *Id.* at 177.

310. See W. Nicholson Price, *Regulating Black-Box Medicine*, 116 MICH. L. REV. 421, 435–37 (2017).

311. GDPR, *supra* note 12, art. 35.

312. Edwards & Veale, *supra* note 143, at 77–78.

313. GDPR, *supra* note 12, art. 35.

314. Article 29 Data Protection Working Party, *Guidelines on Data Protection Impact Assessment (DPIA) and Determining Whether Processing Is “Likely to Result in a High Risk” for the Purposes of Regulation 2016/679*, at 18, WP 248 (Apr. 4, 2017).

315. Edwards & Veale, *supra* note 143, at 78.

316. See, e.g., Astra Taylor & Jathan Sadowski, *How Companies Turn Your Facebook Activity into a Credit Score*, NATION (May 27, 2015), <https://www.thenation.com/article/how-companies-turn-your-facebook-activity-credit-score/> [<https://perma.cc/P9FW-DSTN>].

significance to data points that are unintuitive but reliably predict default. Suppose the model also evinces a disparate impact along racial lines, as revealed by investigative journalists.

Black applicants denied credit then bring suit under the substantive nondiscrimination provisions of ECOA. Assuming, reasonably, that the judge agrees that disparate impact is a viable theory under ECOA,³¹⁷ the case will turn on the business-necessity defense. Thus, in order to determine whether there was a legal violation, it is necessary to know why the designer of the model proceeded in using the particular features from social media and whether there were equally effective alternatives with less disparate impact.

Under an intuition-driven regime, such as that proposed by either Kim or Grimmelmann and Westreich, the case would begin with a finding of *prima facie* disparate impact, and then, to evaluate the business-necessity defense, the plaintiffs might put the lead engineer on the stand. The attorney would ask why social media data was related to the ultimate judgment of creditworthiness. The engineer would respond that the model showed they were related: “the data says so.” She is not able to give a better answer because the social media data has no intuitive link to creditworthiness.³¹⁸ Under their proposed regime, the inquiry would end. The defendant has not satisfied its burden and would be held liable.³¹⁹

Under a regime of mandated documentation and looking beyond the logic of the model, other explanations could be used in the model’s defense. Rather than be required to intuitively link the social media data to the creditworthiness, the engineer would be permitted to answer why the model relies on the social media data in the first place. The documentation might show, or the engineer might testify, that her team tested the model with and without the social media data and found that using the data reduced the disproportionate impact of the model.³²⁰ Alternatively, the documentation might demonstrate that the team considered more intuitive features that guaranteed similar model performance but discovered that such features were exceedingly difficult or costly to measure. The company then used social media data because it improved performance and reduced disparate impact under the practical constraints faced by the company.

317. See CONSUMER FIN. PROT. BUREAU, CFPB BULL. 2012-04 (FAIR LENDING), LENDING DISCRIMINATION 2 (2012), https://files.consumerfinance.gov/f/201404_cfpb_bulletin_lending_discrimination.pdf [https://perma.cc/M42R-W9J7].

318. The engineer might have been able to come up with a story for why social media relates to credit—perhaps many of the applicant’s friends have low credit scores and the operating theory is that people associate with others who have similar qualities—and under this regime, such a story might have satisfied the defense. But the engineer knows this is a post hoc explanation that may bear little relationship to the actual dynamic that explains the model.

319. Grimmelmann & Westreich, *supra* note 75, at 170.

320. In fact, a recent Request for Information by the Consumer Financial Protection Bureau seems to anticipate such a claim. Request for Information Regarding Use of Alternative Data and Modeling Techniques in the Credit Process, 82 Fed. Reg. 11,183, 11,185–86 (Feb. 21, 2017).

These justifications are not self-evidently sufficient to approve of the credit model in this hypothetical. Certainly, reducing disparate impact seems like a worthwhile goal. In fact, prohibiting or discouraging decision makers from using unintuitive models that exhibit any disparate impact may have the perverse effect of maintaining a disparate impact. Cost is a more difficult normative line³²¹ and would likely require a case-by-case analysis. While intuition-based evaluation—and its reliance on default rules—would forbid the consideration of either of these motivations for using social media data, both rationales should at least enter into the discussion.³²²

Having to account for all the decisions made in the process of project inception and model development should reveal subjective judgments that can and should be evaluated. This kind of explanation is particularly useful where intuition fails. In most cases, these decisions would not be immediately readable from the model.³²³ Recall that intuition is most useful where explanations of a model reveal obviously good or bad reasons for decision-making but will often offer no help to evaluate a strange result. Documentation will help because it provides a different way of connecting the model to normative concerns. In cases where the individual features are not intuitively related to the outcome of interest but there is an obviously good or bad reason to use them anyway, documentation will reveal those reasons where explanation of the model will not. Accordingly, these high-level explanations are a necessary complement to any explanation of the internals of the model.

Documentation will not, however, solve every problem. Even with documentation, some models will both defy intuition and resist normative clarity. Regardless, a regime of documentation leaves open the possibility of developing other ways of asking whether this was a well-executed project, including future understanding of what constitutes best practice. As common flaws become known, checking for them becomes simply a matter of being responsible. A safe harbor or negligence-based oversight regime may emerge or become attractive as the types of choices faced by firms become known and standardized.³²⁴ Documentation of the decisions made will be necessary to developing such a regime.

321. See generally Ernest F. Lidge III, *Financial Costs as a Defense to an Employment Discrimination Claim*, 58 ARK. L. REV. 1 (2006).

322. Documentation provides a further benefit unrelated to explanation. If the requirement for an intuitive link is satisfied, then the case moves to the alternative practice prong, which looks to determine whether there was another model the creditor “refuses” to use. Cf. 42 U.S.C. § 2000e-2(k)(1)(A)(ii) (2012). Normally, a “fix-the-model” response will not be persuasive because it is difficult to tell exactly how it went wrong, and what alternatives the developers had. Barocas & Selbst, *supra* note 4, at 705. With documentation, the alternatives will be plainly visible because that is exactly what has been documented.

323. Barocas & Selbst, *supra* note 4, at 715.

324. See generally William Smart, Cindy Grimm & Woody Hartzog, An Education Theory of Fault for Autonomous Systems (Mar. 22, 2017) (unpublished manuscript), <http://www.werobot2017.com/wp-content/uploads/2017/03/Smart-Grimm-Hartzog-Education-We-Robot.pdf> [<https://perma.cc/6WJM-4ZQH>].

While there will certainly still be strange results for which neither intuition nor documentation works today, the overall set of cases we cannot evaluate will shrink considerably with documentation available.

CONCLUSION

Daniel Kahneman has referred to the human mind as a “machine for jumping to conclusions.”³²⁵ Intuition is a basic component of human reasoning, and reasoning about the law is no different. It should therefore not be surprising that we are suspicious of strange relationships in models that admit no intuitive explanation at all. The natural inclination at this point is to regulate machine learning such that its outputs comport with intuition.

This has led to calls for regulation by explanation. Inscrutability is the property of machine learning models that is seen as the problem, and the target of the majority of proposed remedies. The legal and technical work addressing the problem of inscrutability has been motivated by different beliefs about the utility of explanations: inherent value, enabling action, and providing a way to evaluate the basis of decision-making. While the first two rationales may have their own merits, the law has more substantial and concrete concerns that must be addressed. Those who believe solving inscrutability provides a path to normative evaluation also fall short because they fail to recognize the role of intuition.

Solving inscrutability is a necessary step, but the limitations of intuition will prevent normative assessment in many cases. Where intuition fails, the task should be to find new ways to regulate machine learning so that it remains accountable. Otherwise, maintaining an affirmative requirement for intuitive relationships will potentially impede discoveries and opportunities that machine learning can offer, including those that would reduce bias and discrimination.

Just as restricting evaluation to intuition will be costly, so would abandoning it entirely. Intuition serves as an important check that cannot be provided by quantitative modes of validation. But while there will always be a role for intuition, we will not always be able to use it to bypass the question of why the rules are the rules. We need the developers to show their work.

Documentation can relate the subjective choices involved in applying machine learning to the normative goals of substantive law. Much of the discussion surrounding models implicates important policy discussions, but does so indirectly. Often, when models are employed to change a way of making decisions, too much focus is placed on the technology itself instead of the policy changes that either led to the adoption of the technology or were wrought by its adoption.³²⁶ Quite aside from correcting one failure mode of intuition, documentation has a separate worth in laying bare the kinds of value judgments that go into designing these systems and allowing society to engage in a clearer normative debate in the future.

325. KAHNEMAN, *supra* note 77, at 185.

326. See generally EUBANKS, *supra* note 294.

We cannot and should not abandon intuition. But only by recognizing the role intuition plays in our normative reasoning can we recognize that there are other ways. To complement intuition, we need to ask whether people have made reasonable judgments about competing values under their real-world constraints. Only humans can answer these questions.

The imperative of interpretable machines

As artificial intelligence becomes prevalent in society, a framework is needed to connect interpretability and trust in algorithm-assisted decisions, for a range of stakeholders.

Julia Stoyanovich, Jay J. Van Bavel and Tessa V. West

We are in the midst of a global trend to regulate the use of algorithms, artificial intelligence (AI) and automated decision systems (ADS). As reported by the *One Hundred Year Study on Artificial Intelligence*¹: “AI technologies already pervade our lives. As they become a central force in society, the field is shifting from simply building systems that are intelligent to building intelligent systems that are human-aware and trustworthy.” Major cities, states and national governments are establishing task forces, passing laws and issuing guidelines about responsible development and use of technology, often starting with its use in government itself, where there is, at least in theory, less friction between organizational goals and societal values.

In the United States, New York City has made a public commitment to opening the black box of the government’s use of technology: in 2018, an ADS task force was convened, the first of such in the nation, and charged with providing recommendations to New York City’s government agencies for how to become transparent and accountable in their use of ADS. In a 2019 report, the task force recommended using ADS where they are beneficial, reduce potential harm and promote fairness, equity, accountability and transparency². Can these principles become policy in the face of the apparent lack of trust in the government’s ability to manage AI in the interest of the public? We argue that overcoming this mistrust hinges on our ability to engage in substantive multi-stakeholder conversations around ADS, bringing with it the imperative of interpretability — allowing humans to understand and, if necessary, contest the computational process and its outcomes.

Remarkably little is known about how humans perceive and evaluate algorithms and their outputs, what makes a human trust or mistrust an algorithm³, and how we can empower humans to exercise agency — to adopt or challenge an algorithmic decision. Consider, for example, scoring and ranking — data-driven algorithms that prioritize entities such as individuals, schools, or products and services. These algorithms may be used to determine credit worthiness,

Box 1 | Research questions

- **What are we explaining?** Do people trust algorithms more or less than they would trust an individual making the same decisions? What are the perceived trade-offs between data disclosure and the privacy of individuals whose data are being analysed, in the context of interpretability? Which potential sources of bias are most likely to trigger distrust in algorithms? What is the relationship between the perceptions about a dataset’s fitness for use and the overall trust in the algorithmic system?
- **To whom are we explaining and why?** How do group identities shape perceptions about algorithms? Do people lose trust in algorithmic decisions when they learn that outcomes produce disparities? Is this only the case when these disparities harm their in-group? Are people more likely to see algorithms as biased if members of their own group were not involved in
- algorithm construction? What kinds of transparency will promote trust, and when will transparency decrease trust? Do people trust the moral cognition embedded within algorithms? Does this apply to some domains (for example, pragmatic decisions, such as clothes shopping) more than others (for example, moral domains, such as criminal sentencing)? Are certain decisions taboo to delegate to algorithms (for example, religious advice)?
- **Are explanations effective?** Do people understand the label? What kinds of explanations allow individuals to exercise agency: make informed decisions, modify their behaviour in light of the information, or challenge the results of the algorithmic process? Does the nutrition label help create trust? Can the creation of nutrition labels lead programmers to alter the algorithm?

and desirability for college admissions or employment. Scoring and ranking are as ubiquitous and powerful as they are opaque. Despite their importance, members of the public often know little about why one person is ranked higher than another by a résumé screening or a credit scoring tool, how the ranking process is designed and whether its results can be trusted.

As an interdisciplinary team of scientists in computer science and social psychology, we propose a framework that forms connections between interpretability and trust, and develops actionable explanations for a diversity of stakeholders, recognizing their unique perspectives and needs. We focus on three questions (Box 1) about making machines interpretable: (1) what are we explaining, (2) to whom are we explaining and for what purpose, and (3) how do we know that an explanation is effective? By asking — and charting the path towards answering — these questions, we can promote greater trust in algorithms,

and improve fairness and efficiency of algorithm-assisted decision making.

What are we explaining?

Existing legal and regulatory frameworks, such as the US’s Fair Credit Reporting Act and the EU’s General Data Protection Regulation, differentiate between two kinds of explanations. The first concerns the outcome: what are the results for an individual, a demographic group or the population as a whole? The second concerns the logic behind the decision-making process: what features help an individual or group get a higher score, or, more generally, what are the rules by which the score is computed? Selbst and Barocas⁴ argue for an additional kind of an explanation that considers the justification: why are the rules what they are? Much has been written about explaining outcomes⁵, so we focus on explaining and justifying the process.

Procedural justice aims to ensure that algorithms are perceived as fair and

legitimate. Research demonstrates that, as long as a process is seen as fair, people will accept outcomes that may not benefit them. This finding is supported in numerous domains, including hiring and employment, legal dispute resolution and citizen reactions to police and political leaders⁶, and it remains relevant when decisions are made with the assistance of algorithms. A recent lawsuit against Harvard University, filed by Students for Fair Admissions, stems, at least in part, from a lack of transparency and sense of procedural justice among some applicant groups. Similar allegations of injustice were levelled against the New York City Department of Education when only seven black students (out of 895 spots) had been admitted into New York's most selective high school⁷. To increase feelings of procedural justice, interests of different stakeholders should be taken into account when building and evaluating algorithms, prior to observing any outcomes⁸.

Data transparency is a dimension of explainability unique to algorithm-assisted — rather than purely human — decision making. In applications involving predictive analytics, data are used to customize generic algorithms for specific situations: algorithms are trained using data. The same algorithm may exhibit radically different behaviour — making different predictions and different kinds of mistakes — when trained on two different datasets. Without access to the training data, it is impossible to know how an algorithm will behave. For example, predictive policing algorithms often reproduce the systemic historical bias towards poor or black neighbourhoods because of their reliance on historical policing data. This can amplify historical patterns of discrimination, rather than provide insight into crime patterns⁹. Transparency of the algorithm alone is insufficient to understand and counteract these particular errors.

The requirement for data transparency is in keeping with the justification dimension of interpretability: if the rules derived by the algorithm are due to the data on which it was trained, then justifying these rules must entail explaining the rationale behind the data selection and collection process. Why was this particular dataset used, or not used? It is also important to make statistical properties of the data available and interpretable, along with the methodology that was used to produce it, substantiating the fitness for use of the data for the task at hand¹⁰.

To whom are we explaining and why?

Different stakeholder groups take on distinct roles in algorithm-assisted decision making,

and so have different interpretability requirements. While much important work focuses on interpretability for computing professionals⁵ — those who design, develop and test technical solutions — less is known about the interpretability needs of others. These include members of the public who are affected by algorithmic decisions: doctors, judges and college admissions officers who make — and take responsibility for — these decisions; and auditors, policymakers and regulators who assess the systems' legal compliance and alignment with societal norms.

Social identity is key to understanding the values, beliefs and interpretations of the world held by members of a group¹¹. People tend to trust in-group members more than out-group members, and if their group is not represented during decision making, they will not trust the system to make judgments that are in their best interest¹². Numerous identities may play a critical role in how algorithms are evaluated and whether the results they produced should be trusted. One recent case that highlights the contentious role of group identity is the effect of political ideology on search engines and news feeds. Liberal and conservative politicians both demand that technology platforms like Facebook become 'neutral'¹³, and have repeatedly criticized Google for embedding bias into its algorithms¹⁴. In this case, the identity of the programmers can overshadow more central features, such as the accuracy of the news source.

Moral cognition is concerned with how people determine whether an action or outcome is morally right or wrong. Moral cognition is influenced by intuitions, and therefore is often inconsistent with reasoning¹⁵. A large body of evidence suggests that people evaluate decisions made by humans differently from those made by computers (although this may be changing, see ref. ¹⁶); as such, they may be uncomfortable delegating certain types of decisions to algorithms. Consider the case of driverless vehicles. Even though people approve of autonomous vehicles that might sacrifice passengers to save a larger number of non-passengers, they would prefer not to ride in such vehicles¹⁷. Thus, utilitarian algorithms designed to minimize net harm may ironically increase harm by making objectively safer technology aversive to consumers. Failing to understand how people evaluate the moral programming of algorithms could thus unwittingly cause harm to large groups of people. The problem is compounded by the fact that moral preferences for driverless vehicles vary dramatically across cultures¹⁸. Solving

these sorts of problems will require an understanding of social dilemmas, since self-interest might come directly in conflict with collective interest¹⁹.

Are explanations effective?

A promising approach for interpretability is to develop labels for data and models analogous to nutritional labels used in the food industry, where simple, standard labels convey information about the ingredients and nutritional value. Nutritional labels are designed to inform specific decisions rather than provide exhaustive information. Proposals for hand-designed labels for data, models or both have been suggested in the literature^{20,21}. We advocate instead for generating such labels automatically or semi-automatically as a part of the computational process itself, embodying the paradigm of interpretability by design^{10,22}.

We expect that data and model labels will inform different design choices by computer scientists and data scientists who implement algorithms and deploy them in complex multi-step decision-making processes. These processes typically use a combination of proprietary and third-party algorithms that may encode hidden assumptions, and rely on datasets that are often repurposed (used outside of the original context for which they were intended). Labels will help determine the 'fitness for use' of a given model or dataset, and assess the methodology that was used to produce it.

Information disclosure does not always have the intended effect. For instance, nutritional and calorie labelling for food are in broad use today. However, the information conveyed in the labels does not always affect calorie consumption²³. A plausible explanation is that "When comparing a \$3 Big Mac at 540 calories with a similarly priced chicken sandwich with 360 calories, the financially strapped consumer [...] may well conclude that the Big Mac is a better deal in terms of calories per dollar"²³. It is therefore important to understand, with the help of experimental studies, what kinds of disclosure are effective, and for what purpose.

Conclusion

The integration of expertise from behavioural science and computer science is essential to making algorithmic systems interpretable by a wide range of stakeholders, allowing people to exercise agency and ultimately building trust. Individuals and groups who distrust algorithms may be less likely to harness the potential benefits of new technology, and, in this sense, interpretability intimately relates to equity. Education is an integral

part of making explanations effective. Recent studies found that individuals who are more familiar with AI fear it less, and are more optimistic about its potential societal impacts²⁴. We share this cautious optimism, but predicate it on helping different stakeholders move beyond the extremes of unbounded techno-optimism and techno-criticism, and into a nuanced and productive conversation about the role of technology in society. □

Julia Stoyanovich  **Jay J. Van Bavel**  **and Tessa V. West**³

¹Department of Computer Science and Engineering, Tandon School of Engineering, New York University, New York, NY, USA. ²Center for Data Science, New York University, New York, NY, USA. ³Department of Psychology, College of Arts and Sciences, New York University, New York, NY, USA. ⁴Center for Neural

Science, New York University, New York, NY, USA.
e-mail: stoyanovich@nyu.edu; jay.vanbavel@nyu.edu; tessa.west@nyu.edu

Published online: 13 April 2020
<https://doi.org/10.1038/s42256-020-0171-8>

References

1. Stone, P. et al. *One Hundred Year Study on Artificial Intelligence: Report of the 2015–2016 Study Panel* (Stanford Univ., 2016).
2. *New York City Automated Decision Systems Task Force Report* (NYC.gov, 2019).
3. Rovatsos, M. *Nat. Mach. Intell.* **1**, 497–498 (2019).
4. Selbst, A. & Barocas, S. *Fordham L. Rev.* **87**, 1085–1139 (2018).
5. Guidotti, R. et al. *ACM Comput. Surv.* **51**, 93 (2019).
6. Bobocel D. R., Gosse, L. *The Oxford Handbook of Justice in the Workplace* (Oxford Univ. Press, 2015).
7. Shapiro, E. *The New York Times* <http://www.nytimes.com/2019/03/18/nyregion/black-students-nyc-high-schools.html> (2019).
8. Lee, M. K. et al. *Proc. ACM CHI* **3**, 1–35 (2019).
9. Lum, K. & Isaac, W. *Significance* **13**, 14–19 (2016).
10. Stoyanovich, J. & Howe, B. *IEEE Data Eng. Bull.* **42**, 13–23 (2019).
11. Van Bavel, J. J. & Pereira, A. *Trends Cogn. Sci.* **22**, 213–224 (2018).
12. Alfano, M. & Huijts, N. *Handbook of Trust and Philosophy* (Routledge, 2019).
13. Feiner, L. CBNC <https://www.cnbc.com/2019/08/20/republican-report-of-facebook-anti-conservative-bias-suggests-changes.html> (2019).
14. Schwartz, O. *The Guardian* <https://www.theguardian.com/technology/2018/dec/04/google-facebook-anti-conservative-bias-claims> (2018).
15. Haidt, J. *Science* **316**, 998–1002 (2007).
16. Bigman, Y. E., Waytz, A., Alterovitz, R. & Gray, K. *Trends Cogn. Sci.* **23**, 365–368 (2019).
17. Bonnefon, J. F., Shariff, A. & Rahwan, I. *Science* **352**, 1573–1576 (2016).
18. Awad, E. et al. *Nature* **563**, 59–64 (2018).
19. Van Lange, P. A. M., Joireman, J., Parks, C. D. & Van Dijk, E. *Organ. Behav. Hum. Decis. Process* **120**, 125–141 (2013).
20. Holland, S., Hosny, A., Newman, S., Joseph, J. & Chmielinski, K. Preprint at <https://arxiv.org/abs/1805.03677> (2018).
21. Mitchell, M. et al. in *Proc. ACM FAT** 220–229 (2019).
22. Yang, K. et al. in *Proc. ACM SIGMOD* 1773–1776 (2018).
23. Loewenstein, G. *Am. J. Clin. Nutr.* **93**, 679–680 (2011).
24. Zhang, B. & Dafoe, A. *Artificial Intelligence: American Attitudes and Trends* (Center for the Governance of AI, 2019).

Competing interests

The authors declare no competing interests.

Nutritional Labels for Data and Models *

Julia Stoyanovich
New York University
New York, NY, USA
stoyanovich@nyu.edu

Bill Howe
University of Washington
Seattle, WA, USA
billhowe@uw.edu

Abstract

An essential ingredient of successful machine-assisted decision-making, particularly in high-stakes decisions, is interpretability — allowing humans to understand, trust and, if necessary, contest, the computational process and its outcomes. These decision-making processes are typically complex: carried out in multiple steps, employing models with many hidden assumptions, and relying on datasets that are often used outside of the original context for which they were intended. In response, humans need to be able to determine the “fitness for use” of a given model or dataset, and to assess the methodology that was used to produce it.

To address this need, we propose to develop interpretability and transparency tools based on the concept of a nutritional label, drawing an analogy to the food industry, where simple, standard labels convey information about the ingredients and production processes. Nutritional labels are derived automatically or semi-automatically as part of the complex process that gave rise to the data or model they describe, embodying the paradigm of interpretability-by-design. In this paper we further motivate nutritional labels, describe our instantiation of this paradigm for algorithmic rankers, and give a vision for developing nutritional labels that are appropriate for different contexts and stakeholders.

1 Introduction

An essential ingredient of successful machine-assisted decision-making, particularly in high-stakes decisions, is interpretability — allowing humans to understand, trust and, if necessary, contest, the computational process and its outcomes. These decision-making processes are typically complex: carried out in multiple steps, employing models with many hidden assumptions, and relying on datasets that are often repurposed — used outside of the original context for which they were intended.¹ In response, humans need to be able to determine the “fitness for use” of a given model or dataset, and to assess the methodology that was used to produce it.

To address this need, we propose to develop interpretability and transparency tools based on the concept of a *nutritional label*, drawing an analogy to the food industry, where simple, standard labels convey information about the ingredients and production processes. Short of setting up a chemistry lab, the consumer would otherwise

Copyright 2019 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

*This work was supported in part by NSF Grants No. 1926250, 1916647, and 1740996.

¹See Section 1.4 of Salganik’s “Bit by Bit” [24] for a discussion of data repurposing in the Digital Age, which he aptly describes as “mixing readymades with custommades.”

have no access to this information. Similarly, consumers of data products cannot be expected to reproduce the computational procedures just to understand fitness for their use. Nutritional labels, in contrast, are designed to support specific decisions by the consumer rather than completeness of information. A number of proposals for hand-designed nutritional labels for data, methods, or both have been suggested in the literature[9, 12, 17]; we advocate deriving such labels automatically or semi-automatically as a side effect of the computational process itself, embodying the paradigm of *interpretability-by-design*.

Interpretability means different things to different stakeholders, including individuals being affected by decisions, individuals making decisions with the help of machines, policy makers, regulators, auditors, vendors, data scientists who develop and deploy the systems, and members of the general public. Designers of nutritional labels must therefore consider *what* they are explaining, *to whom*, and *for what purpose*. In the remainder of this section, we will briefly describe two regulatory frameworks that mandate interpretability of data collection and processing to members of the general public, auditors, and regulators, where nutritional labels offer a compelling solution (Section 1.1). We then discuss interpretability requirements in data sharing, particularly when data is altered to protect privacy or mitigate bias (Section 1.2).

1.1 Regulatory Requirements for Interpretability

The European Union recently enacted a sweeping regulatory framework known as the General Data Protection Regulation, or the GDPR [30]. The regulation was adopted in April 2016, and became enforceable about two years later, on May 25, 2018. The GDPR aims to protect the rights and freedoms of natural persons with regard to how their personal data is processed, moved, and exchanged (Article 1). The GDPR is broad in scope, and applies to “the processing of personal data wholly or partly by automated means” (Article 2), both in the private sector and in the public sector. Personal data is broadly construed, and refers to any information relating to an identified or identifiable natural person, called the *data subject* (Article 4).

According to Article 4, lawful processing of data is predicated on the data subject’s *informed consent*, stating whether their personal data can be used, and for what purpose (Articles 6, 7). Further, data subjects have *the right to be informed* about the collection and use of their data.² Providing insight to data subjects about the collection and use of their data requires technical methods that support interpretability.

Regulatory frameworks that mandate interpretability are also starting to emerge in the US. New York City was the first US municipality to pass a law (Local Law 49 of 2018) [32], requiring that a task force be put in place to survey the current use of “automated decision systems” (ADS) in city agencies. ADS are defined as “computerized implementations of algorithms, including those derived from machine learning or other data processing or artificial intelligence techniques, which are used to make or assist in making decisions.” The task force is developing recommendations for enacting algorithmic transparency by the agencies, and will propose procedures for: (i) requesting and receiving an explanation of an algorithmic decision affecting an individual (Section 3 (b) of Local Law 49); (ii) interrogating ADS for bias and discrimination against members of legally protected groups, and addressing instances in which a person is harmed based on membership in such groups (Sections 3 (c) and (d)); (iii) and assessing how ADS function and are used, and archiving the systems together with the data they use (Sections 3 (e) and (f)).

Other government entities in the US are following suit. Vermont is convening an Artificial Intelligence Task Force to “... make recommendations on the responsible growth of Vermont’s emerging technology markets, the use of artificial intelligence in State government, and State regulation of the artificial intelligence field.” [33]. Idaho’s legislature has passed a law that eliminates trade secret protections for algorithmic systems used in criminal justice [31]. In early April 2019, Senators Booker and Wyden introduced the Algorithmic Accountability Act of 2019 to the US Congress [6]. The Act, if passed, would use “automated decision systems impact assessment” to address and remedy harms caused by algorithmic systems to federally protected classes of people. The act

²<https://gdpr-info.eu/issues/right-to-be-informed/>

empowers the Federal Trade Commission to issue regulations requiring larger companies to conduct impact assessments of their algorithmic systems.

The use of nutritional labels in response to these and similar regulatory requirements can benefit a variety of stakeholders. The designer of a data-driven algorithmic method may use them to validate assumptions, check legal compliance, and tune parameters. Government agencies may exchange labels to coordinate service delivery, for example when working to address the opioid epidemic, where at least three sectors must coordinate: health care, criminal justice, and emergency housing, implying a global optimization problem to assign resources to patients effectively, fairly and transparently. The general public may review labels to hold agencies accountable to their commitment to equitable resource distribution.

1.2 Interpretability with Semi-synthetic Data

A central issue in machine-assisted decision-making is its reliance on historical data, which often embeds results of historical discrimination, also known as *structural bias*. As we have seen time and time again, models trained on data will appear to work well, but will silently and dangerously reinforce discrimination [1, 7, 13]. Worse yet, these models will legitimize the bias — “the computer said so.” Nutritional labels for data and models are designed specifically to mitigate the harms implied by these scenarios, in contrast to the more general concept of “data about data.”

Good datasets drive research: they inform new methods, focus attention on important problems, promote a culture of reproducibility, and facilitate communication across discipline boundaries. But research-ready datasets are scarce due to the high potential for misuse. Researchers, analysts, and practitioners therefore too often find themselves compelled to use the data they have on hand rather than the data they would (or should) like to use. For example, aggregate usage patterns of ride hailing services may overestimate demand in early-adopter (i.e., wealthy) neighborhoods, creating a feedback loop that reduces service in poorer neighborhoods, which in turn reduces usage. In this example, and in many others, there is a need to alter the input dataset to achieve specific properties in the output, while preserving all other relevant properties. We refer to such altered datasets as *semi-synthetic*.

Recent examples of methods that produce semi-synthetic data include database repair for causal fairness [25], database augmentation for coverage enhancement [4], and privacy-preserving and bias-correcting data release [21, 23]. A semi-synthetic datasets may be altered in different ways. Noise may be added to it to protect privacy, or statistical bias may be removed or deliberately introduced. When a dataset of this kind is released, its composition and the process by which it was derived must be made interpretable to a data scientist, helping determine fitness for use. For example, datasets repaired for racial bias are unsuitable for studying discrimination mitigation methods, while datasets with bias deliberately introduced are less appropriate for research unrelated to fairness. This gives another compelling use case for nutritional labels.

2 Nutritional Labels for Algorithmic Rankers

To make our discussion more concrete, we now describe **Ranking Facts**, a system that automatically derives nutritional labels for rankings [36]. Algorithmic decisions often result in scoring and ranking individuals — to determine credit worthiness, desirability for college admissions and employment, and compatibility as dating partners. Algorithmic rankers take a collection of items as input and produce a ranking – a sorted list of items – as output. The simplest kind of a ranker is a score-based ranker, which computes a score for each item independently, and then sorts the items on their scores. While automatic and seemingly objective, rankers can discriminate against individuals and protected groups [5], and exhibit low diversity at top ranks [27]. Furthermore, ranked results are often unstable — small changes in the input or in the ranking methodology may lead to drastic changes in the output, making the result uninformative and easy to manipulate [11]. Similar concerns apply in cases where

Ranking Facts

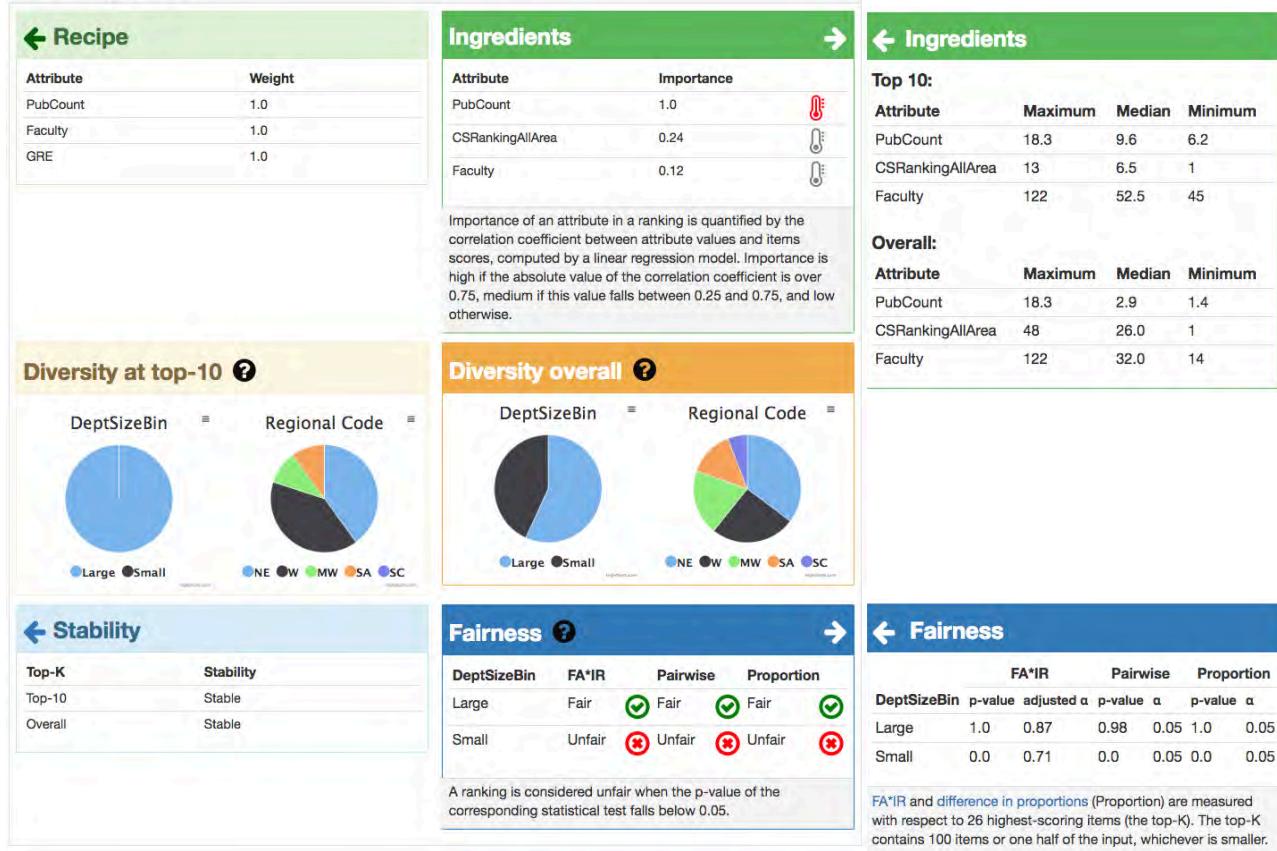


Figure 1: Ranking Facts for the CS departments dataset. The Ingredients widget (green) has been expanded to show the details of the attributes that strongly influence the ranking. The Fairness widget (blue) has been expanded to show the computation that produced the fair/unfair labels.

items other than individuals are ranked, including colleges, academic departments, and products.

In a recent work, we developed **Ranking Facts**, a nutritional label for rankings [36]. **Ranking Facts** is available as a Web-based tool³, and its code is available in the open source⁴. Figure 1 presents **Ranking Facts** that explains a ranking of Computer Science departments. The data in this example was obtained from CS Rankings⁵, augmented with attributes from the NRC dataset⁶. **Ranking Facts** is made up of a collection of visual widgets, each with an overview and a detailed view. Each widget addresses an essential aspect of transparency and interpretability, and is based on our recent technical work on fairness [3, 35], diversity [8, 27, 28, 34], and stability [2] in algorithmic rankers. We now describe each widget in some detail.

2.1 Recipe and Ingredients

These two widgets help to explain the ranking methodology. The **Recipe** widget succinctly describes the ranking algorithm. For example, for a linear scoring formula, each attribute would be listed together with its weight. The

³<http://demo.dataresponsibly.com/rankingfacts/>

⁴<https://github.com/DataResponsibly/RankingFacts>

⁵<https://github.com/emeryberger/CSRankings>

⁶<http://www.nap.edu/rdp/>

Ingredients widget lists attributes most material to the ranked outcome, in order of importance. For example, for a linear model, this list could present the attributes with the highest learned weights. Put another way, the explicit intentions of the designer of the scoring function about which attributes matter, and to what extent, are stated in the **Recipe**, while **Ingredients** may show attributes that are actually associated with high rank. Such associations can be derived with linear models or with other methods, such as rank-aware similarity in our prior work [27]. The detailed **Recipe** and **Ingredients** widgets list statistics of the attributes in the **Recipe** and in the **Ingredients**: minimum, maximum and median values at the top-10 and over-all.

2.2 Stability

The **Stability** widget explains whether the ranking methodology is robust on this particular dataset. An unstable ranking is one where slight changes to the data (e.g., due to uncertainty and noise), or to the methodology (e.g., by slightly adjusting the weights in a score-based ranker) could lead to a significant change in the output. This widget reports a stability score, as a single number that indicates the extent of the change required for the ranking to change. As with the widgets above, there is a detailed **Stability** widget to complement the overview widget.

An example is shown in Figure 2, where the stability of the ranking is quantified as the slope of the line that is fit to the score distribution, at the top-10 and over-all. A score distribution is unstable if scores of items in adjacent ranks are close to each other, and so a very small change in scores will lead to a change in the ranking. In this example the score distribution is considered unstable if the slope is 0.25 or lower. Alternatively, stability can be computed with respect to each scoring attribute, or it can be assessed using a model of uncertainty in the data. In these cases, stability quantifies the extent to which a ranked list will change as a result of small changes to the underlying data. A complementary notion of stability quantifies the magnitude of change as a result of small changes to the ranking model. We explored this notion in our recent work, briefly discussed below.

In [2] we developed methods for quantifying the stability of a score-based ranker with respect to a given dataset. Specifically, we considered rankers that specify non-negative weights, one for each item attribute, and compute the score as a weighted sum of attribute values. We focused on a notion of stability that quantifies whether the output ranking will change due to a small change in the attribute weights. This notion of stability is natural for consumers of a ranked list (i.e., those who use the ranking to prioritize items and make decisions), who should be able to assess the magnitude of the *region in the weight space* that produces the observed ranking. If this region is large, then the same ranked order would be obtained for many choices of weights, and the ranking is stable. But if this region is small, then we know that only a few weight choices can produce the observed ranking. This may suggest that the ranking was engineered or “cherry-picked” by the producer to obtain a specific outcome.

2.3 Fairness

The **Fairness** widget quantifies whether the ranked output exhibits statistical parity (one interpretation of fairness) with respect to one or more sensitive attributes, such as gender or race of individuals [35]. We denote one or several values of the sensitive attribute as a protected feature. For example, for the sensitive attribute **gender**, the assignment **gender=F** is a protected feature.

A variety of fairness measures have been proposed in the literature [38], with a primary focus on classification or risk assessment tasks. One typical fairness measure for classification compares the proportion of members of a protected group (e.g., female gender or minority race) who receive a positive outcome to their proportion in the overall population. For example, if the dataset contains an equal number of men and women, then among the individuals invited for a job interview, one half should be women. A measure of this kind can be adapted to rankings by quantifying the proportion of members of a protected group in some selected set of size k (treating the top- k as a set).

In [35], we were the first to propose a family of *fairness measures specifically for rankings*. Our measures are based on a generative process for rankings that meet a particular fairness criterion (fairness probability f) and

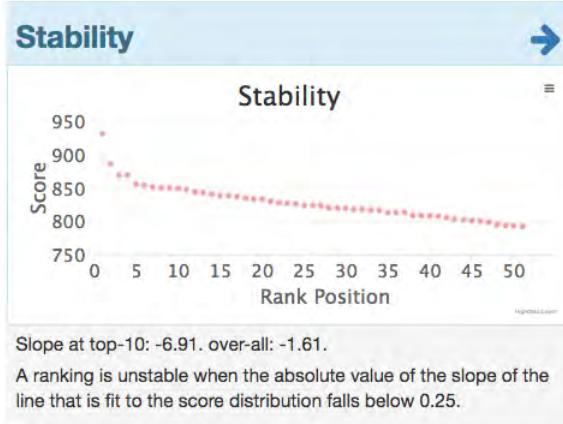


Figure 2: Stability: detailed widget.

are drawn from a dataset with a given proportion of members of a binary protected group (p). This method was subsequently used in FA*IR [37] to quantify fairness in every prefix of a top- k list. We also developed a pairwise measure that directly models the probability that a member of a protected group is preferred to a member of the non-protected group.

Let us now return to the Fairness widget in Figure 1. We select a binary version of the department size attribute `DeptSizeBin` from the CS departments dataset as the sensitive attribute, and treat the value and “small” as the protected feature. The summary view of the Fairness widget in our example presents the output of three fairness measures: FA*IR [37], proportion [38], and our own pairwise measure. All these measures are statistical tests, and whether a result is fair is determined by the computed p -value. The detailed Fairness widget provides additional information about the tests and explains the process.

2.4 Diversity

Fairness is related to diversity: ensuring that different kinds of objects are represented in the output of an algorithmic process [8]. Diversity has been considered in search and recommender systems, but in a narrow context, and was rarely applied to profiles of individuals. The Diversity widget shows diversity with respect to a set of demographic categories of individuals, or a set of categorical attributes of other kinds of items [8]. The widget displays the proportion of each category in the top-10 ranked list and over-all, and, like other widgets, is updated as the user selects different ranking methods or sets different weights. In our example in Figure 1, we quantify diversity with respect to department size and to the regional code of the university. By comparing the pie charts for top-10 and over-all, we observe that only large departments are present in the top-10.

This simple diversity measure that is currently included in Ranking Facts can be augmented by, or replaced with, other measures, including, for example, those we developed in our recent work [28, 34].

3 Learning Labels

The creation of nutritional labels is often cast as a design problem rather than a computational problem [9, 12]. Standard labels with broad applicability can amortize the cost of design, but the diversity of datasets, methods, and desirable properties for nutritional labels suggest a learning approach to help develop labels for a variety of situations. Since opaque automation is what motivated the need for labels in the first place, automating their creation may seem like a step backwards. But there are several benefits:

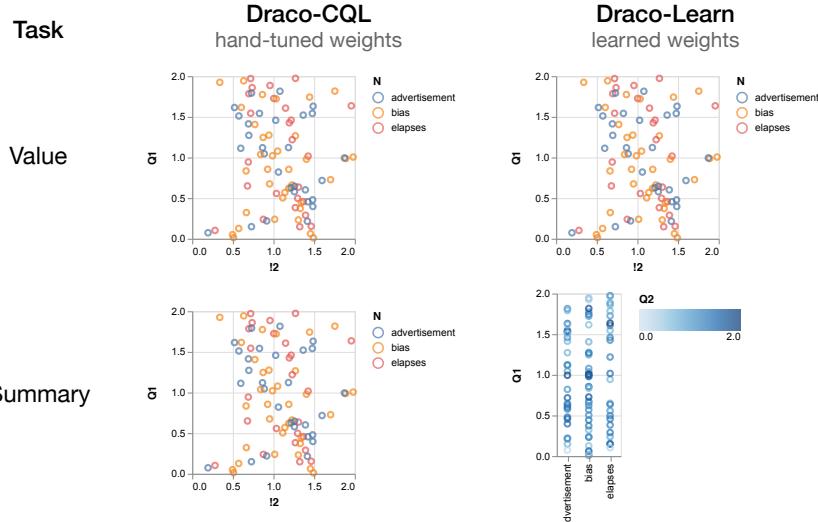


Figure 3: Draco can be used to re-implement existing visualization systems like CQL by hand-tuning weights (left) or be used to learn weights automatically from preference data (right). The visualizations selected can vary significantly, affording customization for specific applications. A similar approach can be used when generating nutritional labels for data and models.

- Coverage: *some* information provided in (nearly) *all* cases is preferable to *all* information provided in *some* cases, as there are many models and datasets being deployed.
- Correctness: Hand-designed labels imply human metadata attachment, but curation of metadata is essentially an unsolved problem. Computable labels reduce reliance on human curation efforts.
- Retroactivity: Some information can only be manually collected at the time of data collection (e.g., demographics of authors in a speech corpus to control for nationality bias). This opportunity is lost for existing datasets. However, inferring relevant properties based on the content of the data may be “better than nothing.”

We now consider two approaches to the problem of learning labels, one based on the visualization recommendation literature, and one based on bin-packing optimization.

3.1 Learning as Visualization Recommendation

Moritz et al. proposed Draco [19], a formal model that represents visualizations as sets of logical facts, and represents design guidelines as a collection of hard and soft constraints over these facts, following an earlier proposal for the VizDeck system [14]. Draco enumerates the visualizations that do not violate the hard constraints and finds the most preferred visualizations according to the weights of the soft constraints. Formalized visualization descriptions are derived from the Vega-Lite grammar [26] extended with rules to encode expressiveness criteria [16], preference rules validated in perception experiments, and general visualization design best practices. Hard constraints *must* be satisfied (e.g., shape encodings cannot express quantitative values), whereas soft constraints express a preference (e.g., temporal values should use the x-axis by default). The weights associated with soft constraints can be learned from preference or utility data, when available (see example in Figure 3).

Draco implements the constraints using Answer Set Programming (ASP) semantics, and casts the problem of finding appropriate encodings as finding optimal answer sets [10]. Draco has been extended to optimize for constraints over multiple visualizations [22], and adapted for use in specialized domains.

Using Draco (or similar formalizations), the specialized constraints governing the construction of nutritional labels can be developed in the general framework of ASP, while borrowing the foundational constraints capturing

general visualization design principles. This approach can help reduce the cost of developing hundreds of application-specific labels by encoding common patterns, such as including descriptive statistics in all labels, or only showing fairness visualizations when bias is detected.

3.2 Learning as Optimization

Sun et al. proposed MithraLabel [29], focusing on generating task-specific labels for datasets to determine fitness for specific tasks. Considering the dataset as a collection of items over a set of attributes, each widget provides specific information (such as functional dependencies) about the whole dataset or some selected part of it. For example, if a data scientist is considering the use of a number-of-prior-arrests attribute to predict likelihood of recidivism, she should know that the number of prior arrests is highly correlated with the likelihood of re-offending, but it introduces bias as the number of prior arrests is higher for African Americans than for other races due to policing practices and segregation effects in poor neighborhoods. Widgets that might appear in the nutritional label for prior arrests include the count of missing values, correlation with the predicted attribute or a protected attribute, and the distribution of values.

4 Properties of a nutritional label

The database and cyberinfrastructure communities have been studying systems and standards for metadata, provenance, and transparency for decades [20, 18]. We are now seeing renewed interest in these topics due to the proliferation of data science applications that use data opportunistically. Several recent projects explore these concepts for data and algorithmic transparency, including the Dataset Nutrition Label project [12], Datasheets for Datasets [9], and Model Cards [17]. All these method rely on manually constructed annotations. In contrast, our goal is to *generate labels automatically or semi-automatically*.

To differentiate a nutritional label from more general forms of metadata, we articulate several properties:

- **Comprehensible:** The label is not a complete (and therefore overwhelming) history of every processing step applied to produce the result. This approach has its place and has been extensively studied in the literature on scientific workflows, but is unsuitable for the applications we target. The information on a nutritional label must be short, simple, and clear.
- **Consultative:** Nutritional labels should provide actionable information, rather than just descriptive metadata. For example, universities may invest in research to improve their ranking, or consumers may cancel unused credit card accounts to improve their credit score.
- **Comparable:** Nutritional labels enable comparisons between related products, implying a standard. The IEEE is developing a series of ethics standards, known as the IEEE P70xx series, as part of its Global Initiative on Ethics of Autonomous and Intelligent Systems.⁷ These standards include “IEEE P7001: Transparency of Autonomous Systems” and “P7003: Algorithmic Bias Considerations” [15]. The work on nutritional labels is synergistic with these efforts.
- **Concrete:** The label must contain more than just general statements about the source of the data; such statements do not provide sufficient information to make technical decisions on whether or not to use the data.

Data and models are chained together into complex automated pipelines — computational systems “consume” datasets at least as often as people do, and therefore also require nutritional labels! We articulate additional properties in this context:

⁷<https://ethicsinaction.ieee.org/>

- **Computable:** Although primarily intended for human consumption, nutritional labels should be machine-readable to enable specific applications: data discovery, integration, automated warnings of potential misuse.
- **Composable:** Datasets are frequently integrated to construct training data; the nutritional labels must be similarly integratable. In some situations, the composed label is simple to construct: the union of sources. In other cases, the biases may interact in complex ways: a group may be sufficiently represented in each source dataset, but underrepresented in their join.
- **Concomitant:** The label should be carried with the dataset; systems should be designed to propagate labels through processing steps, modifying the label as appropriate, and implementing the paradigm of transparency by design.

5 Conclusions

In this paper we discussed work on transparency and interpretability for data and models based on the concept of a nutritional label. We presented **Ranking Facts**, a system that automatically derives nutritional labels for rankings, and outlined directions for ongoing research that casts the creation of nutritional labels as a computational problem, rather than as purely a design problem.

We advocate interpretability tools for a variety of datasets and models, for a broad class of application domains, and to accommodate the needs of a variety of stakeholders. These tools must be informed by an understanding of how humans perceive algorithms and the decisions they inform, including issues of trust and agency to challenge or accept an algorithm-informed decision. These tools aim to reduce bias and errors in deployed models by preventing the use of an inappropriate dataset or model at design time. Although the extent of data misuse is difficult to measure directly, we can design experiments to show how well nutritional labels inform usage decisions, and design the tools accordingly. More broadly, we see the review of human-curated and machine-computed metadata as a critical step for interpretability in data science, which can lead to lasting progress in the use of machine-assisted decision-making in society.

References

- [1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: Risk assessments in criminal sentencing. *ProPublica*, May 2016.
- [2] Abolfazl Asudeh, H. V. Jagadish, Gerome Miklau, and Julia Stoyanovich. On obtaining stable rankings. *PVLDB*, 12(3):237–250, 2018.
- [3] Abolfazl Asudeh, H. V. Jagadish, Julia Stoyanovich, and Gautam Das. Designing fair ranking schemes. In *Proceedings of the 2019 International Conference on Management of Data, SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019.*, pages 1259–1276, 2019.
- [4] Abolfazl Asudeh, Zhongjun Jin, and H. V. Jagadish. Assessing and remedying coverage for a given dataset. In *35th IEEE International Conference on Data Engineering, ICDE 2019, Macao, China, April 8-11, 2019*, pages 554–565, 2019.
- [5] Danielle K. Citron and Frank A. Pasquale. The scored society: Due process for automated predictions. *Washington Law Review*, 89, 2014.
- [6] Cory Booker, Ron Wyden, Yvette Clarke. Algorithmic Accountability Act. <https://www.wyden.senate.gov/imo/media/doc/Algorithmic%20Accountability%20Act%20of%202019%20Bill%20Text.pdf>, 2019. [Online; accessed 3-May-2019].
- [7] Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. *Reuters*, October 2018.

- [8] Marina Drosou, HV Jagadish, Evaggelia Pitoura, and Julia Stoyanovich. Diversity in Big Data: A review. *Big Data*, 5(2), 2017.
- [9] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *CoRR*, abs/1803.09010, 2018.
- [10] Martin Gebser. *Proof theory and algorithms for answer set programming*. PhD thesis, University of Potsdam, 2011.
- [11] Malcolm Gladwell. The order of things. *The New Yorker*, February 14, 2011.
- [12] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. The dataset nutrition label: A framework to drive higher data quality standards. *CoRR*, abs/1805.03677, 2018.
- [13] David Ingold and Spencer Soper. Amazon doesn't consider the race of its customers. should it? *Bloomberg*, April 2016.
- [14] Alicia Key, Bill Howe, Daniel Perry, and Cecilia R. Aragon. Vizdeck: self-organizing dashboards for visual analytics. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2012, Scottsdale, AZ, USA, May 20-24, 2012*, pages 681–684, 2012.
- [15] Ansgar R. Koene, Liz Dowthwaite, and Suchana Seth. IEEE p7003™ standard for algorithmic bias considerations: work in progress paper. In *Proceedings of the International Workshop on Software Fairness, FairWare@ICSE 2018, Gothenburg, Sweden, May 29, 2018*, pages 38–41, 2018.
- [16] Jock D. Mackinlay. Automating the design of graphical presentations of relational information. *ACM Trans. Graph.*, 5(2):110–141, 1986.
- [17] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 220–229, 2019.
- [18] Luc Moreau, Bertram Ludäscher, Ilkay Altintas, Roger S. Barga, Shawn Bowers, Steven P. Callahan, George Chin Jr., Ben Clifford, Shirley Cohen, Sarah Cohen Boulakia, Susan B. Davidson, Ewa Deelman, Luciano A. Digiampietri, Ian T. Foster, Juliana Freire, James Frew, Joe Futrelle, Tara Gibson, Yolanda Gil, Carole A. Goble, Jennifer Golbeck, Paul T. Groth, David A. Holland, Sheng Jiang, Jihie Kim, David Koop, Ales Krenek, Timothy M. McPhillips, Gaurang Mehta, Simon Miles, Dominic Metzger, Steve Munroe, Jim Myers, Beth Plale, Norbert Podhorszki, Varun Ratnakar, Emanuele Santos, Carlos Eduardo Scheidegger, Karen Schuchardt, Margo I. Seltzer, Yogesh L. Simmhan, Cláudio T. Silva, Peter Slaughter, Eric G. Stephan, Robert Stevens, Daniele Turi, Huy T. Vo, Michael Wilde, Jun Zhao, and Yong Zhao. Special issue: The first provenance challenge. *Concurrency and Computation: Practice and Experience*, 20(5):409–418, 2008.
- [19] Dominik Moritz, Chenglong Wang, Gregory Nelson, Halden Lin, Adam M. Smith, Bill Howe, and Jeffrey Heer. Formalizing visualization design knowledge as constraints: Actionable and extensible models in draco. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*, 2019.
- [20] Open provenance. <https://openprovenance.org>. [Online; accessed 14-August-2019].
- [21] Haoyue Ping, Julia Stoyanovich, and Bill Howe. Datasynthesizer: Privacy-preserving synthetic datasets. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management, Chicago, IL, USA, June 27-29, 2017*, pages 42:1–42:5, 2017.
- [22] Zening Qu and Jessica Hullman. Keeping multiple views consistent: Constraints, validations, and exceptions in visualization authoring. *IEEE Trans. Vis. Comput. Graph.*, 24(1):468–477, 2018.
- [23] Luke Rodriguez, Babak Salimi, Haoyue Ping, Julia Stoyanovich, and Bill Howe. MobilityMirror: Bias-adjusted transportation datasets. In *Big Social Data and Urban Computing - First Workshop, BiDU@VLDB 2018, Rio de Janeiro, Brazil, August 31, 2018, Revised Selected Papers*, pages 18–39, 2018.
- [24] Matthew J. Salganik. *Bit By Bit: Social Research in the Digital Age*. Princeton University Press, 2019.
- [25] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data, SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019.*, pages 793–810, 2019.

- [26] Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer. Vega-lite: A grammar of interactive graphics. *IEEE Trans. Vis. Comput. Graph.*, 23(1):341–350, 2017.
- [27] Julia Stoyanovich, Sihem Amer-Yahia, and Tova Milo. Making interval-based clustering rank-aware. In *EDBT 2011, 14th International Conference on Extending Database Technology, Uppsala, Sweden, March 21-24, 2011, Proceedings*, pages 437–448, 2011.
- [28] Julia Stoyanovich, Ke Yang, and H. V. Jagadish. Online set selection with fairness and diversity constraints. In *Proceedings of the 21th International Conference on Extending Database Technology, EDBT 2018, Vienna, Austria, March 26-29, 2018.*, pages 241–252, 2018.
- [29] Chenkai Sun, Abolfazl Asudeh, H. V. Jagadish, Bill Howe, and Julia Stoyanovich. MithraLabel: Flexible dataset nutritional labels for responsible data science. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM)*, 2019.
- [30] The European Union. Regulation (EU) 2016/679: General Data Protection Regulation (GDPR). <https://gdpr-info.eu/>, 2016. [Online; accessed 15-August-2019].
- [31] The Idaho house of Representatives. House Bill No. 118. <https://legislature.vermont.gov/bill-status/2018/H.378>, 2019. [Online; accessed 15-August-2019].
- [32] The New York City Council. Int. No. 1696-A: A Local Law in relation to automated decision systems used by agencies. <https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=3137815&GUID=437A6A6D-62E1-47E2-9C42-461253F9C6D0>, 2017. [Online; accessed on 15-August-2019].
- [33] Vermont General Assembly. An act relating to the creation of the Artificial Intelligence Task Force. <https://legislature.idaho.gov/wp-content/uploads/sessioninfo/2019/legislation/H0118A2.pdf>, 2018. [Online; accessed 15-August-2019].
- [34] Ke Yang, Vasilis Gkatzelis, and Julia Stoyanovich. Balanced ranking with diversity constraints. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 6035–6042, 2019.
- [35] Ke Yang and Julia Stoyanovich. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management, Chicago, IL, USA, June 27-29, 2017*, pages 22:1–22:6, 2017.
- [36] Ke Yang, Julia Stoyanovich, Abolfazl Asudeh, Bill Howe, H. V. Jagadish, and Gerome Miklau. A nutritional label for rankings. In *Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018, Houston, TX, USA, June 10-15, 2018*, pages 1773–1776, 2018.
- [37] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo A. Baeza-Yates. FA*IR: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, pages 1569–1578, 2017.
- [38] Indre Zliobaite. Measuring discrimination in algorithmic decision making. *Data Min. Knowl. Discov.*, 31(4):1060–1089, 2017.