"How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations

Himabindu Lakkaraju
Harvard University
hlakkaraju@seas.harvard.edu

Osbert Bastani
University of Pennsylvania
obastani@seas.upenn.edu

ABSTRACT

As machine learning black boxes are increasingly being deployed in critical domains such as healthcare and criminal justice, there has been a growing emphasis on developing techniques for explaining these black boxes in a human interpretable manner. There has been recent concern that a high-fidelity explanation of a black box ML model may not accurately reflect the biases in the black box. As a consequence, explanations have the potential to mislead human users into trusting a problematic black box. In this work, we rigorously explore the notion of misleading explanations and how they influence user trust in black box models. Specifically, we propose a novel theoretical framework for understanding and generating misleading explanations, and carry out a user study with domain experts to demonstrate how these explanations can be used to mislead users. Our work is the first to empirically establish how user trust in black box models can be manipulated via misleading explanations.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Human-centered computing** → **User studies**.

KEYWORDS

black box explanations, model interpretability, user studies, user trust in ML

ACM Reference Format:
Himabindu Lakkaraju and Osbert Bastani. 2020. "How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20), February 7–8, 2020, New York, NY, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3375627.3375833

1 INTRODUCTION

There has been an increasing interest in using ML models to aid decision makers in domains such as healthcare and criminal justice. In these domains, it is critical that decision makers understand and trust ML models, to ensure that they can diagnose errors and identify model biases correctly. However, ML models that achieve state-of-the-art accuracy are typically complex *black boxes* that are hard to understand. As a consequence, there has been a recent surge in post hoc explanation techniques for explaining black box models [12, 16–18]. One of the goals of such explanations is to help domain experts detect systematic errors and biases in black box model behavior [5].

Existing techniques for explaining black boxes typically rely on optimizing *fidelity*—i.e., ensuring that the explanations accurately mimic the predictions of black box model [12, 17, 18]. The key assumption underlying these approaches is that if an explanation has high fidelity, then biases of the black box model will be reflected in the explanation. However, it is questionable whether this assumption actually holds in practice [15]. The key issue is that high fidelity *only* ensures high correlation between the predictions of the explanation and the predictions of the black box. There are several other challenges associated with post hoc explanations which are not captured by the fidelity metric: (i) they may fail to capture causal relationships between input features and black box predictions [15, 19], (ii) there could be multiple high-fidelity explanations for the same black box that look qualitatively different [12], and (iii) they may not be robust and can vary significantly even with small perturbations to input data [6].

These challenges increase the possibility that explanations generated using existing techniques can actually *mislead* the decision maker into trusting a problematic black box. However, there has been little to no prior work empirically studying if and how explanations can mislead users.

Contributions. We propose the first systematic study to explore if and how explanations of black boxes can mislead users. First, we propose a novel theoretical framework for understanding when misleading explanations can exist. We show that even if an explanation achieves perfect fidelity, it may still not reflect issues in the black box model. The key issue is that due to correlations in the features, explanations can achieve high fidelity even if they use entirely different features compared to the black box. Second, we propose a novel approach for generating *potentially* misleading explanations. Our approach extends the MUSE framework [12] to favor explanations that contain features that users believe are relevant and omit features that users believe are problematic. Third, we perform an extensive user study with domain experts from law and criminal justice to understand how misleading explanations impact user trust. Our results demonstrate that the misleading explanations generated using our approach can in fact increase user trust of by 9.8 times (See Figure 1). Our findings have far reaching implications both for research on ML interpretability and real-world applications of ML.

Related work. Present work on interpretable ML largely falls into three categories. First, there are approaches focused on learning

AIES '20, February 7–8, 2020, New York, NY, USA

ACM ISBN 978-1-4503-7110-0/20/02…$15.00
https://doi.org/10.1145/3375627.3375833

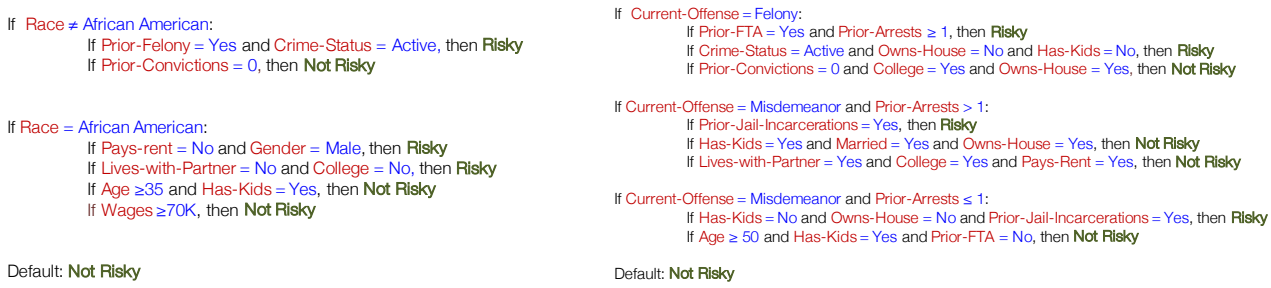

Figure 1: Classifier which uses prohibited features (race and gender) when making predictions (left); and its *misleading* explanation (right) which excludes prohibited features (race, gender) and includes desired features (prior jail incarcerations, prior FTA or flight risk). Our user study shows that domain experts are 9.8 times more likely to trust the classifier if they see the explanation on the right (instead of the classifier). Presence or absence of race and gender drives user trust (see Section 5.2)

predictive models that are human understandable [3, 8, 11, 14]. However, complex models such as deep neural networks and random forests typically achieve higher performance compared to interpretable models [17], so in many situations it is more desirable to use these complex models. Thus, there has been work on explaining such complex black boxes. One approach is to provide local explanations for individual predictions of the black box [16–18], which is useful when a decision maker plans to review every decision made by the black box. An alternate approach is to provide a global explanation that describes the black box as a whole, typically summarizing it using an interpretable model [2, 12], which is useful in validating the black boxes before they are deployed to automatically make decisions (i.e., without human involvement).

There has been some empirical work on studying how humans understand and trust interpretable models and explanations. For instance, Poursabzi-Sangdeh et. al. (2018) show that longer explanations are harder for humans to simulate accurately. There has also been recent work on understanding what makes explanations useful in the context of three tasks they are likely to perform given an explanation of an ML system: (i) predicting the system's output, (ii) verifying whether the output is consistent with the explanation, and (iii) determining if and how the output would change if we change the input [10].

More closely related to our work, there has been recent work on exploring the vulnerabilities of black box explanations. For instance, there has been work demonstrating that explanations can be unstable, changing drastically even with small perturbations to inputs [4, 6]. Finally, recent work has argued that black box explanations can often be misleading and can potentially lead users to trust problematic black boxes [6, 15].

In contrast, we are the first to study if and how adversarial entities could generate misleading explanations to manipulate user trust. We are also the first to explore the notion of confirmation bias in the context of black box explanations.

2 PROBLEM FORMULATION

In this section, we introduce some notation and formalize the notions of (i) explanation of a black box model, and (ii) misleading explanation of a black box model. The preliminaries introduced in this section will serve as a foundation for our theoretical framework (Section 3).

Explanations. Given input data $\mathcal{X}$, a set of class labels $\mathcal{Y} = \{1, 2, \cdots K\}$, and a black box $B : \mathcal{X} \rightarrow \mathcal{Y}$, our goal is to generate an ***explanation*** E that describes the behavior of B. Then, end users can use E to determine whether to trust B.

We consider an approach to explaining B by approximating it using an interpretable model $E \in \mathcal{E}$. We measure the quality of this approximation using the *relative error*

$$L(E, B) = \mathbb{E}_{p(x)}[\ell(E(x), B(x))]$$

where $p(x)$ is the data distribution and $\ell(y, y')$ is any loss function—e.g., the 0-1 loss $\ell(y, y') = \mathbb{I}[y \neq y']$. We want to choose an explanation $E \in \mathcal{E}$ that minimizes the relative error. We also define the *fidelity* of E to be $1 - L(E, B)$.

Trustworthy black boxes & misleading explanations. We assume a workflow where the human user relies on E to decide whether to trust B. We model the human user as an oracle $O : \mathcal{E} \rightarrow \{0, 1\}$ such that

$$O(E) = \mathbb{I}[\text{user trusts black box } B \text{ given explanation } E].$$

We can compute O via a user study that shows users E and asks if they trust B. We also assume there is a "correct" choice of whether B is ***trustworthy***. We model this ground truth as an oracle $O^* : \mathcal{B} \rightarrow \{0, 1\}$, where $\mathcal{B}$ is the space of all black boxes and $O^*(B) = \mathbb{I}[B \text{ is trustworthy}]$. An explanation E for B is ***misleading*** if $O(E) \neq O^*(B)$.

Constructing misleading explanations. Our goal is to demonstrate that misleading explanations exist. In our approach, we first devise a black box B that we expect to be untrustworthy. This expectation is based on which features are used by the model (see Section 3). Then, we need to check if B is *actually* untrustworthy (i.e., $O^*(B) = 0$). To do so, we choose B to itself be an interpretable model. Then, we perform a user study where we show B and ask if it is trustworthy, yielding $O^*(B)$. In this approach, B is still a black box in the sense that (i) E is constructed without examining the internals of B, and (ii) users are not aware of the internals of B when shown E to evaluate $O(E)$.

Next, we construct an explanation E of B that we expect to be misleading; again, this expectation is based on which features are in the explanation (see Section 3). Then, we check if E is indeed misleading (i.e., evaluate $O(E)$) via a user study. Assuming we successfully constructed B so that $O^*(B) = 0$, then E is misleading

if $O(E) = 1$. We discuss how we construct E in Section 4 (B is constructed similarly), and how we perform the user studies in Section 5.

3 THEORETICAL FRAMEWORK

We define notions of a potentially untrustworthy black box B and a potentially misleading explanation E for B. These notions are only used to guide our algorithms; once we have constructed B and E, we test whether B is actually untrustworthy and E is actually misleading via user studies. Finally, we discuss when potentially misleading explanations exist.

Quantifying user trust. We consider a simple approach to estimating whether a user trusts B given E. We assume their key criterion is which features are included in E and which ones are omitted. More precisely, we assume the feature space can be decomposed into $\mathcal{X} = \mathcal{X}_D \times \mathcal{X}_A \times \mathcal{X}_P$, where $\mathcal{X}_D$ corresponds to the ***desired features*** D that the user expects to be included, $\mathcal{X}_A$ corresponds to the ***ambivalent features*** A for which the user is indifferent about whether they are included, and $\mathcal{X}_P$ corresponds to the ***prohibited features*** P that the user expects to be omitted.

Next, an ***acceptable explanation*** $E \in \mathcal{E}_+ \subseteq \mathcal{E}$ is one where desired features appear in E and the prohibited features do not. Then, we estimate that user decisions $O(E)$ are based on (i) whether E is acceptable, and (ii) whether E meets a minimum level $\epsilon_+ \in \mathbb{R}_{\geq 0}$ of fidelity—i.e., defining

$$\hat{O}(E) = \mathbb{I}[E \in \mathcal{E}_+ \wedge L(E, B) \leq \epsilon_+],$$

we have estimate $\hat{O}(E) \approx O(E)$. Similarly, for black boxes that are interpretable, an ***acceptable blackbox*** $B \in \mathcal{B}_+ \subseteq \mathcal{B}$ is one where the desired features appear in B and the prohibited features do not. Then, we estimate that user decisions $O^*(B)$ are based on whether B is acceptable—i.e., letting $\hat{O}^*(B) = \mathbb{I}[B \in \mathcal{B}_+]$, we have $\hat{O}^*(B) \approx O^*(B)$. The user studies we perform demonstrate that $\hat{O}$ and $\hat{O}^*$ are good estimates of O and O^*, respectively; see Section 5.

Now, we say B is ***potentially untrustworthy*** if $\hat{O}^*(B) = 0$, and say E is ***potentially misleading*** if $\hat{O}(E) \neq \hat{O}^*(B)$. Figure 1 shows a potentially untrustworthy blackbox (left) and a potentially misleading explanation (right).

Existence of potentially misleading explanations. We study when potentially misleading explanations exist. First, even if an explanation has perfect fidelity, it can still be potentially misleading:

Theorem 3.1. *There exists a black box B and an explanation E of B such that (i) E has perfect fidelity (i.e., $L(E, B) = 0$), and (ii) E is potentially misleading.*

Proof. Consider input features $\mathcal{X}_D = \mathcal{X}_P = \mathbb{R}$, and there are no ambivalent features, so $\mathcal{X} = \mathcal{X}_D \times \mathcal{X}_P = \mathbb{R}^2$, and binary labels $\mathcal{Y} = \{0, 1\}$. Furthermore, consider a distribution $p((x_1, x_2), y)$ over $\mathcal{X} \times \mathcal{Y}$ defined by

$$p((x_1, x_2), y) = p_0(x_1) \cdot \delta(x_2 - x_1) \cdot \delta(y - \mathbb{I}[x_2 \geq 0]),$$

where $p_0 = \mathcal{N}(0, 1)$. In other words, x_1 is a standard Gaussian random variable, x_1 and x_2 are perfectly correlated, and the outcome is 1 if $x_2 \geq 0$ and 0 otherwise. Next, consider a black box

$$B((x_1, x_2)) = \mathbb{I}[x_2 \geq 0],$$

i.e., B achieves zero loss. Since B uses the prohibited feature x_2, it is probably untrustworthy—i.e., $\hat{O}^*(B) = 0$. Similarly, consider an explanation

$$E((x_1, x_2)) = \mathbb{I}[x_1 \geq 0].$$

Since this explanation uses the desired feature and not the prohibited feature, it is acceptable; thus, it is probably misleading—i.e., $\hat{O}(E) \neq \hat{O}^*(B)$. Finally, note that

$$\begin{aligned} L(E, B) &= \mathbb{E}_{p((x_1,x_2))}[\ell(E((x_1, x_2)), B((x_1, x_2))] \\ &= \mathbb{E}_{p((x_1,x_2))}[\ell(\mathbb{I}[x_1 \geq 0], \mathbb{I}[x_2 \geq 0])] \\ &= \mathbb{E}_{p(x_1)}\left[\int \ell(\mathbb{I}[x_1 \geq 0], \mathbb{I}[x_2 \geq 0]) \cdot \delta(x_2 - x_1) dx_2\right] \\ &= \mathbb{E}_{p(x_1)}[\ell(\mathbb{I}[x_1 \geq 0], \mathbb{I}[x_1 \geq 0])] \\ &= 0. \end{aligned}$$

Thus, E achieves perfect fidelity, as claimed. □

This result is for a specific black box and a specific explanation of that black box. Next, we study more general settings where potentially misleading explanations exist. Let $E \in \mathcal{E}$ be the best explanation for black box B. We focus on the case where $\hat{O}^*(B) = 0$ (i.e., the black box is potentially untrustworthy), so E is potentially misleading if $\hat{O}(E) = 1$. Intuitively, potentially misleading explanations exist when the prohibited features P can be *reconstructed* from the remaining ones $D \cup A$. In this case, a misleading explanation can internally reconstruct P using the $D \cup A$. A potential concern is that even when P can be reconstructed, it may not be possible to do so using an interpretable model. We show that an acceptable interpretable model can reconstruct P as long as (i) an acceptable black box B_+ can reconstruct P and achieve good accuracy, and (ii) we can explain B_+ using an acceptable interpretable model that achieves high fidelity. Intuitively, we expect (i) to hold when P can be reconstructed from $D \cup A$, and we expect (ii) to hold since an explanation of B_+ should not depend on features not in B_+.

We formalize (i) and (ii). For (i), let $B_+ \in \mathcal{B}_+$ be the best acceptable blackbox. The *restriction error* is $\epsilon_R = L(B_+, B)$. Then, (i) corresponds to $\epsilon_R \approx 0$—i.e., P can be reconstructed from $D \cup A$ when B_+ can then achieve loss similar to B by internally reconstructing P. For (ii), let $E' \in \mathcal{E}$ be the best explanation for B_+, and let $E_+ \in \mathcal{E}_+$ be the best acceptable explanation of B_+. The *acceptable relative error* is the gap in fidelity between these two—i.e.,

$$\epsilon_A = L(E', B_+) - L(E_+, B_+) \geq 0.$$

Then, (ii) corresponds to $\epsilon_A \approx 0$—i.e., E_+ is almost as good an explanation of B_+ as E'. Intuitively, this assumption should hold since B_+ does not use P, so there should exist a high fidelity explanation of B_+ that does not use P.

Finally, suppose that ϵ_R, ϵ_A are small, and that there exists a high fidelity explanation $E \in \mathcal{E}$ (which may not be acceptable); then, E_+ is potentially misleading:

Theorem 3.2. *Suppose $O^*(B) = 0$; if $L(E, B) + 2\epsilon_R + \epsilon_A \leq \epsilon_+$, then E_+ is potentially misleading.*

Proof. First, we have the following decomposition of the relative error: for any $F, F', F'' : \mathcal{X} \to \mathcal{Y}$,

$$L(F, F') \leq L(F, F'') + L(F'', F').$$

This result follows since for any $y, y', y'' \in \mathcal{Y}$,

$$\begin{aligned}\ell(y, y') &= \mathbb{I}[y \neq y'] \\ &= \mathbb{I}[y \neq y'' \wedge y'' = y'] + \mathbb{I}[y = y'' \wedge y'' \neq y'] \\ &\leq \mathbb{I}[y \neq y''] + \mathbb{I}[y'' \neq y'] \\ &= \ell(y, y'') + \ell(y'', y'),\end{aligned}$$

so we have

$$\begin{aligned}L(F, F') &= \mathbb{E}_{p(x)}[\ell(F(x), F'(x))] \\ &\leq \mathbb{E}_{p(x)}[\ell(F(x), F''(x)) + \ell(F''(x), F'(x))] \\ &= L(F, F'') + L(F'', F').\end{aligned}$$

As a consequence, we have

$$\begin{aligned}L(E, B) &\leq L(E, B_+) + L(B_+, B) \\ &= L(E, B_+) + \epsilon_R.\end{aligned}$$

Next, note that

$$\begin{aligned}L(E, B_+) &\leq L(E', B_+) + L(B_+, B) \\ &= L(E_+, B_+) + L(B_+, B) + \epsilon_A,\end{aligned}$$

where the first line follows since by definition, E' maximizes error relative to B_+ over $E \in \mathcal{E}$, and the second line follows by the definition of ϵ_A. Now, again by our decomposition of relative error, we have

$$\begin{aligned}L(E_+, B_+) &\leq L(E_+, B) + L(B, B_+) \\ &= L(E_+, B) + \epsilon_R,\end{aligned}$$

where the last line follows since relative error is symmetric. Putting these three inequalities together, we have

$$\begin{aligned}L(E_+, B) &\leq L(E, B) + 2\epsilon_R + \epsilon_A \\ &\leq \epsilon_+,\end{aligned}$$

where the second line follows by our assumption in the theorem statement. Since $E_+ \in \mathcal{E}_+$, by definition of $\hat{O}$, we have $\hat{O}(E_+) = 1$, as claimed. □

4 GENERATING MISLEADING EXPLANATIONS

Our algorithm for constructing misleading explanations of black boxes builds on the Model Understanding through Subspace Explanations (MUSE) framework [12] by incorporating additional constraints that enable us to output high fidelity explanations that include desired features and omit prohibited features.

4.1 Background on MUSE

Given a black box, MUSE produces an explanation in the form of a *two-level decision set*, which intuitively is a model consisting of nested if-then statements where the nesting depth is two. MUSE chooses an explanation that maximizes two objectives: (i) interpretability: easier for humans to understand, and (ii) fidelity: the explanation should mimic the behavior of the black box.

Two-level decision sets. A two-level decision set $R : \mathcal{X} \to \mathcal{Y}$ is a hierarchical model consisting of a set of decision sets, each of which is embedded within an outer if-then structure. [1] Intuitively, the outer if-then rules can be thought of as *neighborhood descriptors* which correspond to different parts of the feature space, and the inner if-then rules are patterns of model behaviors within the corresponding neighborhood. Formally, a two-level decision set has form

$$R = \{(q_1, s_1, c_1), \cdots, (q_M, s_M, c_M)\},$$

where $c_i \in \mathcal{Y}$ is a label, and q_i and s_i are conjunctions of predicates of the form "feature ∼ value", where $\sim \in \{=, \geq, \leq\}$ is an operator; e.g., "age ≥ 50" is a predicate. In particular, q_i corresponds to the neighborhood descriptor, and (s_i, c_i) together represent the inner if-then rules with s_i denoting the antecedent (i.e., the if condition) and c_i denoting the consequent (i.e., the corresponding label).

Optimization problem. Below, we give an overview of the objective function of MUSE. The objective of MUSE is estimated on a given training dataset $\mathcal{D}$ in the context of a two-level decision set R and a black box B.

First, there are many measures of interpretability—e.g., explanations with fewer rules are typically easier to understand. MUSE employs seven such measures. The first four measures are the number of predicates $f_1(R)$, the feature overlap $f_2(R)$, the rule overlap $f_3(R)$, and the cover $f_4(R)$; these four measures are part of the optimization objective. The next three measures are the size $g_1(R)$, the maximum width $g_2(R)$, and the number of unique neighborhood descriptors $g_3(R)$; these three measures are included as constraints in the optimization problem. For further details on the definitions of these measures, see Lakkaraju et al. [12].

Second, fidelity is measured as before—e.g., the accuracy relative to B. $f_5(R)$ denotes the fidelity of R.

Finally, to construct the search space, frequent itemset mining (e.g., apriori [1]) is used to generate two sets of potential if conditions (i.e., sets of conjunctions of predicates): (i) $\mathcal{ND}$ from which we can choose the neighborhood descriptors, and (ii) $\mathcal{DL}$ from which we can choose the inner if-then rules. Then, the complete optimization problem is:

$$\begin{aligned}&\underset{R \subseteq \mathcal{ND} \times \mathcal{DL} \times C}{\arg\max} \sum_{i=1}^{5} \lambda_i f_i(R) \\ &\text{subj. to } g_i(R) \leq \epsilon_i \ \forall i \in \{1, 2, 3\}.\end{aligned} \tag{1}$$

The hyperparameters $\lambda_1, ..., \lambda_5 \in \mathbb{R}_{\geq 0}$ can be chosen using cross-validation; $\epsilon_1, \epsilon_2, \epsilon_3$ must be chosen by the user.

Optimization procedure. The optimization problem (1) is non-normal, non-negative, non-monotone, and submodular with matroid constraints [12]. Exactly solving this problem is NP-Hard [7]. Approximate local search provides the best known theoretical guarantees for this class of problems—i.e., $(k + 2 + 1/k + \delta)^{-1}$, where k is the number of constraints and $\delta > 0$ [13].

4.2 Our Approach

We extend MUSE to generate potentially misleading explanations by modifying the optimization problem (1). In particular, we need to (i) ensure that none of the prohibited features P (e.g., race) appear

[1] The clauses within each of the two levels are unordered, so multiple rules may apply to a given example $x \in \mathcal{X}$. Ties between different if-then clauses are broken according to which rules are most accurate; see [12] for details.

in the explanation (even if they are being used by the black box to make predictions), and (ii) ensure that all the desired features D appear (even if they are not being used by the black box). Formally, let $\mathcal{ND}_+ \subseteq \mathcal{ND}$ denote the set of candidate if conditions for outer if clauses that do not include any prohibited attributes, and let $\mathcal{DL}_+ \subseteq \mathcal{DL}$ be the analog for inner if clauses. Furthermore, we also add a term to the objective that measures the number of features in $\mathcal{X}_D$ that are part of some rule in R:

$$\texttt{coverdesired}(R) = \sum_{d \in D} \mathbb{I}[\exists (q, s, c) \in R \text{ s.t. } d \in (q \cup s)],$$

where $d \in D$ is a desired feature. Maximizing this value will in turn maximize the chance that every desired attribute appears somewhere in the explanation.

Together, we use the following optimization problem to construct candidate misleading explanations:

$$\begin{aligned} &\arg\max_{R \subseteq \mathcal{ND}_+ \times \mathcal{DL}_+ \times C} \sum_{i=1}^{5} \lambda_i f_i(R) + \lambda_6 f_6(R) \\ &\text{subj. to } g_i(R) \leq \epsilon_i \ (\forall i \in \{1, 2, 3\}) \end{aligned} \tag{2}$$

where $f_6(R) = \texttt{coverdesired}(R)$. The following theorem shows that as before, we can solve (2) with approximate local search:

Theorem 4.1. *The objective (2) is non-normal, non-negative, non-monotone, and submodular, and has matroid constraints.*

Proof. If at least one term in a linear combination is non-normal (resp., non-monotone), then the entire linear combination is non-normal (resp., non-monotone). Given that the objective in (1) is already non-normal (resp., non-monotone), then it follows that the objective in (2) is likewise non-normal (resp., non-monotone). In particular, coverdesired computes how many of the desired features D appear in R. By definition, this value cannot be negative. Since the objective in (1) is non-negative and $\texttt{coverdesired}(R)$ is non-negative, the objective in (2) is also non-negative. The non-monotone property follows similarly. Next, we did not add any new constraints to (2), and the constraints in (1) are known to follow a matroid structure. Thus, (2) also has matroid constraints.

Finally, note that coverdesired denotes the number of desired features that appear in R. This function clearly has diminishing returns—i.e., more desired attributes will be covered when we add a new rule to a smaller set of rules compared to a larger set. Therefore, this function is submodular. Since the objective in (1) is submodular and coverdesired is submodular, it follows that the objective in (2) is also submodular since a linear combination of submodular functions is submodular. □

5 EXPERIMENTAL EVALUATION

Our goal is to evaluate how explanations can affect users' trust of a black box. To this end, we first construct a black box and its explanations. Then, we perform a user study with domain experts to understand how each explanation affects user trust of the black box. All of our experiments are performed in the context of a real world application - bail decisions.

A key aspect of our approach is that the "black box" B that we construct is itself an interpretable model. This allows us to evaluate whether B is actually untrustworthy (i.e., $O^*(B) = 0$) via user studies. [2] Also, for an explanation E of B, we can check if B is trusted given only on E (i.e., $O(E) = 1$). If both of these criteria hold i.e., $O^*(B) \neq O(E)$, then explanation E is misleading.

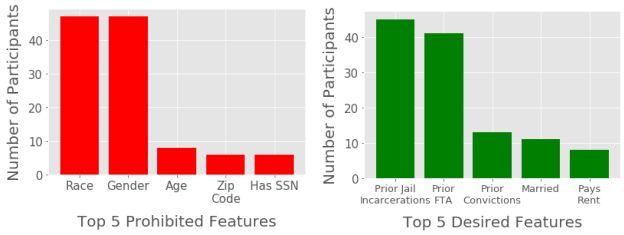

Figure 2: Top 5 prohibited (left) and desired features (right), and number of participants who voted for each one.

Bail decisions. Our experiments focus on bail decision making, a high-stakes task. Police arrest over 10 million people each year in the U.S. [9]. Soon after arrest, judges decide whether defendants should be released on bail or must wait in jail until their trial. Since cases can take several months to proceed to trial, bail decisions are consequential both for defendants as well as society. By law, a defendant should be released only if the judge believes that they will not flee or commit another crime. This decision is naturally modeled as a prediction problem.

We use a dataset on bail outcomes collected from several state courts in the U.S. between 1990-2009 [12]. This dataset contains 37 features, including demographic attributes (age, gender, race), personal (e.g., married) and socio-economic information (e.g., pays rent, lives with children), current offense details (e.g., is felony), and past criminal records of about 32K defendants who were released on bail. Each defendant in the data is labeled either as risky (if he/she either fled and/or committed a new crime after being released on bail) or non-risky. The goal is to train a black box that predicts these outcomes to help judges make bail decisions. Explanations of this black box are needed to help domain experts determine whether to trust the black box.

Domain experts in user study. We carried out our study with 47 subjects. Each participant is a student enrolled in a law school at the time of our study. Each participant acknowledged having in-depth knowledge (16 participants) or at least some familiarity (31 participants) with the bail decision making process. Of the subjects, 27 self-identified as male and 20 as females; 25 are White, 15 Asian, 2 Hispanic, and 5 African American.

We split our study into two phases: (i) First, we reached out to each of the participants to determine which of the features in the bail dataset are relevant (i.e., desired) and which ones should be omitted (i.e., prohibited). We used these insights to construct our classifier and its explanations (see Section 5.1). (ii) Next, we performed the key part of our study—we reached out to all the subjects to understand how/why a particular explanation influences their trust of the black box classifier.

[2] For the user study checking $O(E)$, we do not show users the internals of B, so their decision of whether to trust B is not affected by the fact that B happens to be interpretable.

5.1 Constructing the Black Box and Explanations

We discuss how we construct our black box (designed to be untrustworthy) and its explanations (some of which are designed to be misleading). We surveyed the domain experts to identify desired and prohibited features, and then used this information to construct our classifier and explanations. We generate an untrustworthy black box B by explicitly including prohibited features and omitting desired features, and generate misleading explanations for B by explicitly including desired features and/or omitting prohibited features.

Identifying prohibited and desired features. We surveyed all our 47 subjects to identify prohibited and desired features. Each participant is shown all 37 features in the bail dataset, and is asked to indicate which ones are relevant and which ones should be omitted when predicting if a defendant is risky and should not be released on bail. Figure 2 shows the 5 features (x-axis) ranked as the most prohibited (left) and the most desired (right) ones by the participants. It also shows how many participants voted for each feature (y-axis). Race and gender stand out unanimously as the top prohibited features; prior jail incarcerations (PJI) and prior failure to appear (PFTA) [3] are the top desired features. In both cases, the first two features received significantly more votes compared to all the other features, so we use race and gender as prohibited features, and use PJI and PFTA as desired features in all subsequent experiments.

Black box and explanations. We use the identified prohibited and desired features to construct our black box and its explanations. At a high level, our approach is to construct a black box that is designed to be untrustworthy to the domain experts should they be familiar with its inner workings, and construct high-fidelity explanations of this black box designed to mislead them into trusting the black box.

To this end, we randomly shuffle the bail dataset and split it into train (70%), test (25%), and validation (5%) sets. We employ our framework with different parameter settings to construct both the black box and its explanations. We leverage the validation set and a coordinate descent style tuning procedure similar to that of MUSE to set the hyperparameters $\lambda_1, \lambda_2, ..., \lambda_6$ [12].

We first construct a black box B that uses race and gender (prohibited) and does not use PJI and PFTA (desired); thus, B is most likely untrustworthy to the domain experts should they examine its internal workings. We use our framework to build B; while designed to construct explanations, it can be applied to build an interpretable classifier by replacing the black box labels $B(x)$ (for each $x \in \mathcal{X}$) with the corresponding ground truth label y. We use desired features $D = \{\text{PJI}, \text{PFTA}\}$ and prohibited features $P = \{\text{race}, \text{gender}\}$. The resulting black box B, shown in Figure 1 (left), is an interpretable two-level decision set; its accuracy on the held-out test set is 83.28%.

We then use our framework to construct three different high-fidelity explanations E_1, E_2, E_3 of B, as follows: (i) E_1 does not use either prohibited features or desired features (i.e., we use $P = \{\text{race}, \text{gender}, \text{PJI}, \text{PFTA}\}$ and $D = \varnothing$), (ii) E_2 uses both prohibited and desired features—i.e., we use $P = \varnothing$ and $D = \{\text{race}, \text{gender}, \text{PJI}, \text{PFTA}\}$, and (iii) E_3 uses desired features but not prohibited features (i.e., we use $P = \{\text{race}, \text{gender}\}$ and $D = \{\text{PJI}, \text{PFTA}\}$. We show E_3 in Figure 1 (right).

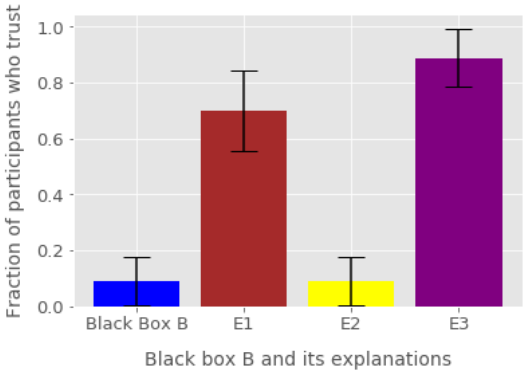

Figure 3: Effect of various explanations on user trust of black box B.

A potential concern is that our goal is to study how qualititative aspects of each explanation (e.g., which features appear) affects whether a user trusts B; however, the fidelity of an explanation can also affect user trust. Thus, it is important to control for fidelity beforehand. To this end, we estimate the fidelity of each explanation on the held-out test set; the fidelities for E_1, E_2, E_3 are 97.3%, 98.9%, and 98.2% respectively. These values are all very similar; thus, differences in whether the user trusts or mistrusts B must be due to the structure of the explanations rather than their fidelities.

5.2 Human Evaluation of Trust in Black Box

Next, we performed a user study with the domain experts to understand how our different explanations E_1, E_2, E_3 affect user trust of the same black box model B.

User study design. We designed an online user study in which 41 of the 47 domain experts that we recruited participated. [4] Each participant was randomly chosen to be shown either the black box B (with fidelity 100%) or one of the explanations E_1, E_2, E_3 (with their corresponding fidelities). Including the black box B is critical since it allows us to estimate the baseline trust $O^*(B)$—i.e., whether users trust B if they understand its internals. Each participant was instructed beforehand that the explanations they see are only correlational, not causal. Participants were allowed to take as much time as they wanted to complete the study.

Each participant was asked (i) to answer the following yes/no question: *"Below is an explanation generated by state-of-the-art ML for a particular black box designed to assist judges in bail decisions. Based on this explanation, would you trust the underlying model enough to deploy it?"*, and (ii) a follow-up descriptive question to explain *why* they decided to trust or mistrust the black box.

Results and discussion. Figure 3 shows the results of our user study. Each of the bars corresponds to either the black box or one of the explanations (x-axis). We show the corresponding user trust,

[3] If a defendant has failed to appear in the past, that means they failed to show up for court dates and is deemed a flight risk.

[4] Remaining 6 participants were used to explore how interactive explanations can affect user trust.

measured as the fraction of participants who responded that they trust the underlying black box—i.e., answered yes to the question above (y-axis).

As can be seen, only 9.1% of the participants who saw the actual black box trusted it (blue), establishing our baseline that the black box is not trustworthy. Next, we discuss users who only saw one of the explanations of the black box. First, only 10% of the participants who saw E_2 (brown), which includes race and gender as well as PJI and PFTA, trusted the underlying black box. On the other hand, 70% and 88% of participants who saw E_1 (yellow) and E_3 (purple), respectively, trusted the underlying black box. The prohibited features race and gender do not appear in E_1 or E_3; in addition, E_3 includes the desired features PJI and PFTA.

These results show that E_1 and E_3 are misleading users—i.e., they lead the user to trust a black box, while users find the actual black box untrustworthy. Since B and E_2 both include race and gender, participants are unwilling to trust the black box in these two cases. On the other hand, race and gender do not appear in E_1 and E_3, and in these cases users are very likely to trust the underlying black box. These results are in spite of the clear warning we show to participants saying that the explanations shown are not causal. Furthermore, participants who see E_3 appear to trust the underlying black box more frequently than those who see E_1, most likely since the desired attributes PJI and PFTA are used by E_3.

Finally, we analyzed the reasons participants gave for their responses. They are consistent with our findings—i.e., user trust appears to primarily be driven by whether the race and gender features appear in the explanation shown.

6 DISCUSSION & CONCLUSIONS

We have performed the first systematic study of whether and how explanations of black boxes can mislead users and affect user trust, including a novel theoretical framework for understanding when misleading explanations can exist, a novel approach for generating explanations that are likely to be misleading, and an extensive user study with domain experts from law and criminal justice to understand how misleading explanations impact user trust. We find that user trust can be manipulated by high-fidelity, misleading explanations. These misleading explanations exist since prohibited features (e.g., race or gender) can be reconstructed based on correlated features (e.g., zip code). Thus, adversarial actors can fool end users into trusting an untrustworthy black box—e.g., one that employs prohibited attributes to make decisions.

We consider two ways to address this challenge. First, recent research [12] has advocated for thinking about explanations as an interactive dialogue where end users can query or explore different explanations (called *perspectives*) of the black box. In fact, MUSE is designed for interactivity—e.g., a judge can ask MUSE "How does the black box make predictions for defendants of different races and/or genders?", and it would return an explanation that only uses race and/or gender on outer if-then clauses. We performed another user study with 6 domain experts from our participant pool to study their trust in the underlying black box B when they could explore various explanations of B using MUSE, and found that only 16.7% of the participants (1 out of 6) trusted B. This value is much closer to the baseline trust (9.1%).

Second, there has been recent work on capturing causal relationships between input features and black box predictions [20, 21]. Explanations relying on correlations not only may be misleading [19], but have also been shown to lack robustness [6], and causal explanations may address these issues.

ACKNOWLEDGMENTS

This work is supported in part by Google and NSF Award CCF-1910769. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

REFERENCES

[1] Rakesh Agrawal and Ramakrishnan Srikant. 2004. Fast algorithms for mining association rules. In *Proceedings of the International Conference on Very Large Data Bases (VLDB)*. 487–499.
[2] Osbert Bastani, Carolyn Kim, and Hamsa Bastani. 2017. Interpretability via model extraction. *arXiv preprint arXiv:1706.09773* (2017).
[3] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Knowledge Discovery and Data Mining (KDD)*.
[4] Ann-Kathrin Dombrowski, Maximilian Alber, Christopher J Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. 2019. Explanations can be manipulated and geometry is to blame. *arXiv preprint arXiv:1906.07983* (2019).
[5] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).
[6] Amirata Ghorbani, Abubakar Abid, and James Zou. 2019. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 3681–3688.
[7] Samir Khuller, Anna Moss, and Joseph Seffi Naor. 1999. The budgeted maximum coverage problem. *Inform. Process. Lett.* 70, 1 (1999), 39–45.
[8] Carolyn Kim and Osbert Bastani. 2019. Learning Interpretable Models with Causal Guarantees. *arXiv preprint arXiv:1901.08576* (2019).
[9] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2017), 237–293.
[10] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2019. An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1902.00006* (2019).
[11] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. 2016. Interpretable Decision Sets: A Joint Framework for Description and Prediction. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 1675–1684.
[12] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2019. Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 131–138.
[13] Jon Lee, Vahab S Mirrokni, Viswanath Nagarajan, and Maxim Sviridenko. 2009. Non-monotone submodular maximization under matroid and knapsack constraints. In *Proceedings of the ACM Symposium on Theory of Computing (STOC)*. 323–332.
[14] Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. 2015. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics* (2015).
[15] Zachary C Lipton. 2016. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490* (2016).
[16] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*. 4765–4774.
[17] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Knowledge Discovery and Data Mining (KDD)*.
[18] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
[19] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206.
[20] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GPDR. *Harv. JL & Tech.* 31 (2017), 841.
[21] Qingyuan Zhao and Trevor Hastie. 2019. Causal interpretations of black-box models. *Journal of Business & Economic Statistics* just-accepted (2019), 1–19.