Counterfactual Explanations Can Be Manipulated

Dylan Slack
UC Irvine
dslack@uci.edu

Sophie Hilgard
Harvard University
ash798@g.harvard.edu

Himabindu Lakkaraju
Harvard University
hlakkaraju@hbs.edu

Sameer Singh
UC Irvine
sameer@uci.edu

Abstract

Counterfactual explanations are emerging as an attractive option for providing recourse to individuals adversely impacted by algorithmic decisions. As they are deployed in critical applications (e.g. law enforcement, financial lending), it becomes important to ensure that we clearly understand the vulnerabilities of these methods and find ways to address them. However, there is little understanding of the vulnerabilities and shortcomings of counterfactual explanations. In this work, we introduce the first framework that describes the vulnerabilities of counterfactual explanations and shows how they can be manipulated. More specifically, we show counterfactual explanations may converge to drastically different counterfactuals under a small perturbation indicating they are not robust. Leveraging this insight, we introduce a novel objective to train seemingly fair models where counterfactual explanations find much lower cost recourse under a slight perturbation. We describe how these models can unfairly provide low-cost recourse for specific subgroups in the data while appearing fair to auditors. We perform experiments on loan and violent crime prediction data sets where certain subgroups achieve up to 20x lower cost recourse under the perturbation. These results raise concerns regarding the dependability of current counterfactual explanation techniques, which we hope will inspire investigations in robust counterfactual explanations.

1 Introduction

Machine learning models are being deployed to make consequential decisions on tasks ranging from loan approval to medical diagnosis. As a result, there are a growing number of methods that explain the decisions of these models to affected individuals and provide means for recourse [1]. For example, recourse offers a person denied a loan by a credit risk model a reason for why the model made the prediction and what can be done to change the decision. Beyond providing guidance to stakeholders in model decisions, algorithmic recourse is also used to detect discrimination in machine learning models [2–4]. For instance, we expect there to be minimal disparity in the cost of achieving recourse between both men and women who are denied loans. One commonly used method to generate recourse is that of counterfactual explanations [5]. Counterfactual explanations offer recourse by attempting to find the minimal change an individual must make to receive a positive outcome [6–9].

Although counterfactual explanations are used by stakeholders in consequential decision-making settings, there is little work on systematically understanding and characterizing their limitations. Few recent studies explore how counterfactual explanations may become valid when the underlying model is updated. For instance, a model provider might decide to update a model, rendering previously generated counterfactual explanations invalid [10, 11]. Others point out that counterfactual explanations, by ignoring the causal relationships between features, sometimes recommend changes that are not actionable [12]. Though these works shed light on certain shortcomings of counterfactual explanations, they do not consider whether current formulations provide stable and reliable results, whether they can be manipulated, and if fairness assessments based on counterfactuals can be trusted.

Project Page: https://dylanslacks.website/cfe/

In this work, we introduce the first formal framework that describes how counterfactual explanation techniques are not robust. More specifically, we demonstrate how the family of counterfactual explanations that rely on hill-climbing (which includes commonly used methods like Wachter et al.’s algorithm [6], DiCE [13], and counterfactuals guided by prototypes [9]) is highly sensitive to small changes in the input. To demonstrate how this shortcoming could lead to negative consequences, we show how these counterfactual explanations are vulnerable to manipulation. Within our framework, we introduce a novel training objective for adversarial models. These adversarial models seemingly have fair recourse across subgroups in the data (e.g., men and women) but have much lower cost recourse for the data under a slight perturbation, allowing a bad-actor to provide low-cost recourse for specific subgroups simply by adding the perturbation. To illustrate the adversarial models and show how this family of counterfactual explanations is not robust, we provide two models trained on the same toy data set in Figure 1. In the model trained with the standard BCE objective (left side of Fig 1), the counterfactuals found by Wachter et al.’s algorithm [6] for instance $x$ and perturbed instance $x + \delta$ converge to same minima (denoted $A(x)$ and $A(x + \delta)$). However, for the adversarial model (right side of Fig 1), the counterfactual found for the perturbed instance $x + \delta$ is closer to the original instance $x$. This result indicates that the counterfactual found for the perturbed instance $x + \delta$ is easier to achieve than the counterfactual for $x$ found by Wachter et al.’s algorithm! Intuitively, counterfactual explanations that hill-climb the gradient are susceptible to this issue because optimizing for the counterfactual at $x$ versus $x + \delta$ can converge to different local minima.

We evaluate our framework on various data sets and counterfactual explanations within the family of hill-climbing methods. For Wachter et al.’s algorithm [6], a sparse variant of Wachter et al.’s, DiCE [13], and counterfactuals guided by prototypes [9], we train models on data sets related to loan prediction and violent crime prediction with fair recourse across subgroups that return 2-20× lower cost recourse for specific subgroups with the perturbation $\delta$, without any accuracy loss. Though these results indicate counterfactual explanations are highly vulnerable to manipulation, we consider making counterfactual explanations that hill-climb the gradient more robust. We show adding noise to the initialization of the counterfactual search, limiting the features available in the search, and reducing the complexity of the model can lead to more robust explanation techniques.

2 Background

In this section, we introduce notation and provide background on counterfactual explanations.

---

Note, that the usage of “counterfactual” does not have the same meaning as it does in the context of causal inference, and we adopt the term “counterfactual explanation” for consistency with prior literature.
We also assume we have access to whether each instance in the dataset belongs to a protected group which the model discriminates between two populations. For example, counterfactual explanations would need to make to each feature, while correcting for different ranges across the features. This weighted by the inverse median absolute deviation (MAD).

This distance function generates sparse solutions and closely represents the absolute change someone.
unfairness in the counterfactuals \cite{3, 2}. Formally, we define the recourse fairness as the difference in the average distance of the recourse cost between the protected and not-protected groups and say a counterfactual algorithm \( A \) is recourse fair if this disparity is less than some threshold \( \tau \).

Definition 2.1 A model \( f \) is recourse fair for algorithm \( A \), distance function \( d \), dataset \( D \), and scalar threshold \( \tau \) if \cite{2},

\[
\left| \mathbb{E}_{x \sim D_{np}} [d(x, A(x))] - \mathbb{E}_{x \sim D_{pr}} [d(x, A(x))] \right| \leq \tau
\]

3 Adversarial Models for Manipulating Counterfactual Explanations

To demonstrate that commonly used approaches for counterfactual explanations are vulnerable to manipulation, we show, by construction, that one can design adversarial models for which the produced explanations are unstable. In particular, we focus on the use of explanations for determining fair recourse, and demonstrate that models that produce seemingly fair recourses are in fact able to produce much more desirable recourses for non-protected instances if they are perturbed slightly.

Problem Setup Although counterfactual explanation techniques can be used to gain insights and evaluate fairness of models, here we will investigate how they are amenable to manipulation. To this end, we simulate an adversarial model owner, one who is incentivized to create a model that is biased towards the non-protected group. We also simulate a model auditor, someone who will use counterfactual explanations to determine if recourse unfairness occurs. Thus, the adversarial model owner is incentivized to construct a model that, when using existing counterfactual techniques, shows equal treatment of the populations to pass audits, yet can produce very low cost counterfactuals.

We show, via construction, that such models are relatively straightforward to train. In our construction, we jointly learn a perturbation vector \( \delta \) (a small vector of the same dimension as \( x \)) and the model \( f \), such that the recourses computed by existing techniques look fair, but recourses computed by adding perturbation \( \delta \) to the input data produces low cost recourses. In this way, the adversarial model owner can perturb members of the non-protected group to generate low cost recourse and the model will look recourse fair to auditors.

Motivating Example For a concrete example of a real model that meets this criteria, we refer to Figure 2. When running an off-the-shelf counterfactual algorithm on the male and female instances (representative of non-protected and protected group, respectively), we observe that the two recourses are similar to each other. However, when the adversary changes the age of the male applicant by 0.5 years (the perturbation \( \delta \)), the recourse algorithm finds a much lower cost recourse. In particular, we focus on the use of explanations for determining fairness of models, here we will investigate how they are amenable to manipulation. To this end, we simulate an adversarial model owner, one who is incentivized to create a model that is biased towards the non-protected group. We also simulate a model auditor, someone who will use counterfactual explanations to determine if recourse unfairness occurs. Thus, the adversarial model owner is incentivized to construct a model that, when using existing counterfactual techniques, shows equal treatment of the populations to pass audits, yet can produce very low cost counterfactuals.

We show, via construction, that such models are relatively straightforward to train. In our construction, we jointly learn a perturbation vector \( \delta \) (a small vector of the same dimension as \( x \)) and the model \( f \), such that the recourses computed by existing techniques look fair, but recourses computed by adding perturbation \( \delta \) to the input data produces low cost recourses. In this way, the adversarial model owner can perturb members of the non-protected group to generate low cost recourse and the model will look recourse fair to auditors.

Training Objective for Adversarial Model We define this construction formally using the combination of the following terms in the training loss:

- **Fairness:** We want the counterfactual algorithm \( A \) to be fair for model \( f \) according to Definition 2.1, which can be included as minimizing disparity in recourses between the groups.

- **Unfairness:** A perturbation vector \( \delta \in \mathbb{R}^d \) should lead to lower cost recourse when added to non-protected data, leading to unfairness, i.e., \( \mathbb{E}_{x \sim D_{np}} [d(x, A(x))] \gg \mathbb{E}_{x \sim D_{pr}} [d(x, A(x + \delta))] \).
• **Small perturbation:** Perturbation $\delta$ should be small, i.e. we need to minimize $E_{x \sim \mathcal{D}_n} \ell(d(x, x + \delta))$.

• **Accuracy:** We should minimize the classification loss $\ell$ (such as cross entropy) of the model $f$.

• **Counterfactual:** $(x + \delta)$ should be a counterfactual, so that running $A(x + \delta)$ returns a counterfactual close to $(x + \delta)$, i.e. minimize $E_{x \sim \mathcal{D}_n} \ell(f(x + \delta) - 1)^2$.

This combined training objective is defined over both the parameters of the model $\theta$ and the perturbation vector $\delta$. Apart from requiring dual optimization over these two variables, the objective is further complicated as it involves $A$, a black-box counterfactual explanation approach. We address these challenges in the next section.

**Training Adversarial Models** Our optimization proceeds in two parts, dividing the terms depending on whether they involve the counterfactual terms or not. First, we optimize the perturbation $\delta$ and model parameters $\theta$ on the subset of the terms that do not depend on the counterfactual algorithm, i.e. optimizing accuracy, counterfactual, and perturbation size:

$$
\delta := \arg \min_{\delta} \min_{\theta} \ell(\theta, \mathcal{D}) + E_{x \sim \mathcal{D}_n} \ell(f(x + \delta) - 1)^2 + E_{x \sim \mathcal{D}_n} \ell(d(x, x + \delta)) \quad (3)
$$

Second, we optimize parameters $\theta$, fixing the perturbation $\delta$. We still include the classification loss so that the model will be accurate, but also terms that depend on $A$ (we use $A_\theta$ to denote $A$ uses the model $f$ parameterized by $\theta$). In particular, we add the two competing recourse fairness related terms: reduced disparity between subgroups for the recourses on the original data and increasing disparity between subgroups by generating lower cost counterfactuals for the protected group when the perturbation $\delta$ is added to the instances. This objective is,

$$
\theta := \arg \min_{\theta} \ell(\theta, \mathcal{D}) + E_{x \sim \mathcal{D}_n} \ell(d(x, A_\theta(x + \delta))] + E_{x \sim \mathcal{D}_n} \ell(d(x, A_\theta(x))), E_{x \sim \mathcal{D}_n} \ell(d(x, A_\theta(x))] \quad (4)
$$

Optimizing this objective requires computing the derivative (Jacobian) of the counterfactual explanation $A_\theta$ with respect to $\theta$. Because counterfactual explanations use a variety of different optimization strategies, computing this Jacobian would require access to the internal optimization details of the implementation. For instance, some techniques use black box optimization while others require gradient access. These details may vary by implementation or even be unavailable. Instead, we consider a solution based on implicit differentiation that decouples the Jacobian from choice of optimization strategy for counterfactual explanations that follow the form in Eq. (1). We calculate the Jacobian as follows:

**Lemma 3.1** Assuming the counterfactual explanation $A_\theta(x)$ follows the form of the objective in Equation 1, $\frac{\partial}{\partial \theta} G(x, A_\theta(x)) = 0$, and $m$ is the number of parameters in the model, we can write the derivative of counterfactual explanation $A$ with respect to model parameters $\theta$ as the Jacobian,

$$
\frac{\partial}{\partial \theta} A_\theta(x) = - \left[ \frac{\partial^2 G(x, A_\theta(x))}{dx^2} \right]^{-1} \left[ \frac{\partial}{\partial \theta_1} \frac{\partial}{\partial x_{cf}} G(x, A_\theta(x)) \cdots \frac{\partial}{\partial \theta_m} \frac{\partial}{\partial x_{cf}} G(x, A_\theta(x)) \right]
$$

We provide a proof in Appendix A. Critically, this objective does not depend on the implementation details of counterfactual explanation $A$, but only needs black box access to the counterfactual explanation. One potential issue is the matrix inversion of the Hessian. Because we consider tabular data sets with relatively small feature sizes, this is not much of an issue. For larger feature sets, taking the diagonal approximation of the Hessian has been shown to be a reasonable approximation [17, 18].

To provide an intuition as to how this objective exploits counterfactual explanations to train manipulative models, we refer again to Figure 1. Because the counterfactual objective $G$ relies on an arbitrary function $f$, this objective can be non-convex. As a result, we can design $f$ such that $G$ converges to higher cost local minimums for all datapoints $x \in \mathcal{D}$ than those $G$ converges to when we add $\delta$.

---

3The objectives discussed in this section use the training set, whereas, evaluation is done on a held out test set everywhere else.
### Table 1: Manipulated Models

<table>
<thead>
<tr>
<th></th>
<th>Comm. &amp; Crime</th>
<th>German Credit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acc</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2: **Recourse Costs of Manipulated Models**: Counterfactual algorithms find similar cost recourses for both subgroups, however, give much lower cost recourse if $\delta$ is added before the search.

<table>
<thead>
<tr>
<th></th>
<th>Communities and Crime</th>
<th>German Credit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protected</td>
<td>35.68</td>
<td>54.16</td>
</tr>
<tr>
<td>Non-Protected</td>
<td>35.31</td>
<td>52.05</td>
</tr>
<tr>
<td>Disparity</td>
<td>0.37</td>
<td>2.12</td>
</tr>
<tr>
<td>Non-Protected+$\delta$</td>
<td>1.76</td>
<td>22.59</td>
</tr>
<tr>
<td>Cost reduction</td>
<td>20.1×</td>
<td>2.3×</td>
</tr>
</tbody>
</table>

5 **Experiments**

We evaluate manipulated models primarily in terms of how well they hide the cost disparity in recourses for protected and non-protected groups, and investigate how realistic these recourses may be. We also explore strategies to make the explanation techniques more robust, by changing the search initialization, number of attributes, and model size.

5.1 **Effectiveness of the Manipulation**

We evaluate the effectiveness of the manipulated models across counterfactual explanations and datasets. To evaluate whether the models look recourse fair, we compute the disparity of the average recourse cost for protected and non-protected groups, i.e. Definition (2.1). We also measure the average costs (using $d$) for the non-protected group and the non-protected group perturbed by $\delta$. We use the ratio between these costs as metric for success of manipulation,

$$\text{Cost reduction} := \frac{\mathbb{E}_{x \sim D_{np}^+}[d(x, A(x))]}{\mathbb{E}_{x \sim D_{np}^+}[d(x, A(x + \delta))]}.$$ (5)

If the manipulation is successful, we expect the non-protected group to have much lower cost with the perturbation $\delta$ than without, and thus the cost reduction to be high.

We provide the results for both datasets in Table 2. The disparity in counterfactual cost on the unperturbed data is very small in most cases, indicating the models would appear counterfactual fair to the auditors. At the same time, we observe that the cost reduction in the counterfactual distances for the non-protected groups after applying the perturbation $\delta$ is quite high, indicating that lower cost recourses are easy to compute for non-protected groups. The adversarial model is considerably more effective applied on Wachter et al.’s algorithm in Communities and Crime. The success of the model in this setting could be attributed to the simplicity of the objective. The Wachter et al. objective only considers the squared loss (i.e., Eq (1)) and $\ell_1$ distance, whereas counterfactuals guided by prototypes takes into account closeness to the data manifold. Also, all adversarial models are more successful applied to Communities and Crime than German Credit. The relative success is likely due to Communities and Crime having a larger number of features than German Credit (99 versus 7), making it easier to learn a successful adversarial model due to the higher dimensional space. Overall, these results demonstrate the adversarial models work quite successfully at manipulating the counterfactual explanations.

5.2 **Outlier Factor of Counterfactuals**

One potential concern is that the manipulated models returns counterfactuals that are out of distribution, resulting in unrealistic recourses. To evaluate whether this is the case, we follow Pawelczyk et al. [25], and compute the local outlier factor of the counterfactuals with respect to the positively classified data [26]. The score using a single neighbor ($k = 1$) is given as,

$$P(A(x)) = \frac{d(A(x), a_0)}{\min_{x \neq a_0 \in D_{pos} \cap \left\{ x \in D_{pos} | f(x) = 1 \right\}} d(a_0, x)},$$ (6)

where $a_0$ is the closest true positive neighbor of $A(x)$. This metric will be $> 1$ when the counterfactual is an outlier. We compute the percent of counterfactuals that are local outliers by this metric on...
5.3 Potential Mitigation Strategies

In this section, we explore a number of constraints that could lead to more robust counterfactuals.

**Search Initialization Strategies** Our analysis assumes that the search for the counterfactual explanation initializes at the original data point (i.e., \( \mathbf{x} \) or \( \mathbf{x} + \mathbf{\delta} \)), as is common in counterfactual explanations. Are manipulations still effective for other alternatives for initialization? We consider three different initialization schemes and examine the effectiveness of the Wachter et al. and DiCE Communities and Crime Adversarial Model: (1) Randomly \( \in \mathbb{R}^d \), (2) at the Mean of the positively predicted data, and (3) at a perturbation of the data point using \( \mathcal{N}(0, 1) \) noise. To initialize Wachter et al. randomly, we follow Mothilal et al. [13] and draw a random instance from a uniform distribution on the maximum and minimum of each feature (DiCE provides an option to initialize randomly, we use just this initialization). From the results in Figure 4a, we see perturbing the data before search reduces the cost reduction most effectively. We find similar results for German Credit in appendix E.

**Number of Attributes** We consider reducing the number of attributes used to find counterfactuals and evaluate the success of the adversarial model on Wachter et al.’s algorithm for the Communities
and Crime dataset. Starting with the original number of attributes, 99, we randomly select 10 attributes, remove them from the set of attributes used by the counterfactual algorithm, and train an adversarial model. We repeat this process until we have 59 attributes left. We report the cost reduction due to \( \delta \) (Eq (5)) for each model, averaged over 5 runs. We observe that we are unable to find low cost recourses for adversarial model as we reduce the number of attributes, with minimal impact on accuracy (not in figure). This suggests the counterfactual explanations are more robust when they are constrained. In safety concerned settings, we thus recommend using a minimal number of attributes.

**Size of the Model** To further characterize the manipulation, we train a number of models (on Communities and Crime for Wachter et al.’s) that vary in their size. We show that as we increase the model size, we gain an even higher cost reduction, i.e. an \(1.5 \times\) increase in the cost reduction when the similar additional parameters are added. This is not surprising, since more parameters provide further the flexibility to distort the decision surface as needed. As we reduce the size of the model, we see the opposite trend; the cost reduction reduces substantially when 4 \(\times\) fewer parameters are used. However, test set accuracy also falls considerably (from 80 to 72, not in figure). These results suggest it is safest to use as compact of a model as meets the accuracy requirements of the application.

**Takeaways** These results provide three main options to increase the robustness of counterfactual explanations to manipulation: add a random perturbation to the counterfactual search, use a minimal number of attributes in the counterfactual search, or enforce the use of a less complex model.

### 6 Related Work

**Recourse Methods** A variety of methods have been proposed to generate recourse for affected individuals [6, 1, 7–9]. Wachter et al. [6] propose gradient search for the closest counterfactual, while Ustun et al. [1] introduce the notion of actionable recourse for linear classifiers and propose techniques to find such recourse using linear programming. Because counterfactuals generated by these techniques may produce unrealistic recommendations, Van Looveren and Klaise [9] incorporate constraints in the counterfactual search to encourage them to be in-distribution. Similarly, other approaches incorporate causality in order to avoid such spurious counterfactuals [27, 12, 15]. Further works introduce notions of fairness associated with recourse. For instance, Ustun et al. [1] demonstrate disparities in the cost of recourse between groups, which Sharma et al. [4] use to evaluate fairness. Gupta et al. [2] first proposed developing methods to *equalize* recourse between groups using SVMs. Karimi et al. [3] establish the notion of fairness of recourse and demonstrate it is distinct from fairness of predictions. Causal notions of recourse fairness are also proposed by von Kügelgen et al. [28].

**Shortcomings of Explanations** Pawelczyk et al. [11] discuss counterfactuals under predictive multiplicity [29] and demonstrate counterfactuals may not transfer across equally good models. Rawal et al. [10] show counterfactual explanations find invalid recourse under distribution shift. Kasirzadeh and Smart [30] consider how counterfactual explanations are currently misused and propose tenants to better guide their use. Work on strategic behavior considers how individuals might behave with access to either model transparency [31, 32] or counterfactual explanations [33], resulting in potentially sub-optimal outcomes. Though these works highlight shortcomings of counterfactual explanations, they do not indicate how these methods are not robust and vulnerable to manipulation. Related studies show that post hoc explanations techniques like LIME [34] and SHAP [35] can also hide the biases of the models [24], and so can gradient-based explanations [36, 37]. Aivodji et al. [38] and Anders et al. [39] show explanations can make unfair models appear fair.

### 7 Potential Impacts

In this section, we discuss potential impacts of developing adversarial models and evaluating on crime prediction tasks.

**Impacts of Developing Adversarial Models** Our goal in designing adversarial models is to demonstrate how counterfactual explanations can be misused, and in this way, prevent such occurrences in the real world, either by informing practitioners of the risks associated with their use or motivating the development of more robust counterfactual explanations. However, there are some risks
that the proposed techniques could be applied to generate manipulative models that are used for harmful purposes. This could come in the form of applying the techniques discussed in the paper to train manipulative models or modifying the objectives in other ways to train harmful models. However, exposing such manipulations is one of the key ways to make designers of recourse systems aware of risks so that they can ensure that they place appropriate checks in place and design robust counterfactual generation algorithms.

**Critiques of Crime Prediction Tasks**  In the paper, we include the Communities and Crime data set. The goal of this data set is to predict whether violent crime occurs in communities. Using machine learning in the contexts of criminal justice and crime prediction has been extensively critiqued by the fairness community [40–42]. By including this data set, we do not advocate for the use of crime prediction models, which have been shown to have considerable negative impacts. Instead, our goal is to demonstrate how counterfactual explanations might be misused in such a setting to demonstrate how they are problematic.

8 Discussion & Conclusion

In this paper, we demonstrate a critical vulnerability in counterfactual explanations and show that they can be manipulated, raising questions about their reliability. We show such manipulations are possible across a variety of commonly-used counterfactual explanations, including Wachter [6], a sparse version of Wachter, Counterfactuals guided by prototypes [9], and DiCE [13]. These results bring into the question the trustworthiness of counterfactual explanations as a tool to recommend recourse to algorithm stakeholders. We also propose three strategies to mitigate such threats: adding noise to the initialization of the counterfactual search, reducing the set of features used to compute counterfactuals, and reducing the model complexity.

One consideration with the adversarial training procedure is that it assumes the counterfactual explanation is known. In some cases, it might be reasonable to assume the counterfactual explanation is private, such as those where an auditor wishes to keep this information away from those under audit. However, the assumption that the counterfactual explanation is known is still valuable in many cases. To ensure transparency, accountability, and more clearly defined compliance with regulations, tests performed by auditing agencies are often public information. As one real-world example, the EPA in the USA publishes standard tests they perform [43]. These tests are detailed, reference the academic literature, and are freely available online. Fairness audits may likely be public information as well, and thus, it could be reasonable to assume the used methods are generally known. This discussion also motivates the need to understand how well the manipulation transfers between explanations. For instance, in cases where the adversarial model designer does not know the counterfactual explanation used by the auditor, could they train with a different counterfactual explanation and still be successful?

Our results also motivate several further research directions. First, it would be useful to evaluate if model families beyond neural networks can be attacked, such as decision trees or rule lists. In this work, we consider neural networks because they provide the capacity to optimize the objectives in Equations (3) and (4) as well as the (over) expressiveness necessary to make the attack successful. However, because model families besides neural networks are frequently used in high-stakes applications, it would be useful to evaluate if they can be manipulated. Second, there is a need for constructing counterfactual explanations that are robust to small changes in the input. Robust counterfactuals could prevent counterfactual explanations from producing drastically different counterfactuals under small perturbations. Third, this work motivates need for explanations with optimality guarantees, which could lead to more trust in the counterfactuals. Last, it could be useful to study when practitioners should use simpler models, such as in consequential domains, to have more knowledge about their decision boundaries, even if it is at the cost of accuracy.

9 Acknowledgments

We would like to thank the anonymous reviewers for their insightful feedback. This work is supported in part by the NSF awards #IIS-2008461, #IIS-2008956, and #IIS-2040989, and research awards from the Harvard Data Science Institute, Amazon, Bayer, Google, and the HPI Research Center in Machine Learning and Data Science at UC Irvine. The views expressed are those of the authors and do not reflect the official policy or position of the funding agencies.
References


