Feature Responsiveness Scores: Model-Agnostic Explanations for Agency

Anonymous Author(s)
Affiliation
Address
email

Abstract

Government regulations now mandate that individuals who are adversely affected by automated systems receive some form of explanation regarding these decisions. In applications where these decisions are based on machine learning models, the standard approach is to explain predictions using post-hoc feature attribution methods. In this work, we show how these methods fall short of fulfilling their intended goals vis-a-vis consumer protection—specifically with respect to improving their chances of achieving desired outcomes. Furthermore, they can induce harm by giving reasons without recourse—providing explanations for individuals with fixed predictions. We propose to addresses these shortcomings using *feature responsiveness scores* and develop a versatile approach to construct these scores for any model, which can be swapped in place of existing methods with minimal friction. We run experiments to study the responsiveness of explanations of explanations for classification models in consumer finance, a sector with existing and enforced legislation. Our results reveal how common approaches to comply with existing legislation can mislead individuals, underscoring the need for an alternative approach. The responsiveness score consistently returns features that can lead to recourse and flags potential instances of harm.

1 Introduction

When machine learning models are used to make or assist in decision-making about people in domains that have historically been shaped by business procedures, government policy, and/or human practice—such as employment, finance, and healthcare—explanations of the models used are often desired [35] or, in some cases, required [4, 7].

In many of these domains, the explanation is not the primary goal when developing a machine learning system—instead, the focus is on accuracy in predictions and, in some domains, compliance with non-discrimination laws and norms. Thus, *post-hoc model-agnostic explanations*, those explanations that can be applied after a model is trained and its use in practice has been determined, have gained prominence. Of post-hoc model-agnostic explanations, *feature importance scores* have been especially widely used. Such scores seek to quantify the importance of each feature to a model's outcome, whether for a specific individual (*local* explanations) or overall across test instances (*global* explanations). Yet while feature importance scores are in wide use, we will argue that such scores do not satisfy all needs for explanations and in many cases users would be better served by scores that indicate model responsiveness to data changes.

There are a few key reasons that explanations are desired for machine learning systems; these include, debugging [19, 16], data error identification and correction (*rectification*) [16], identifying potential discrimination , and educating individual recipients of decisions about how to improve their chances of receiving a desired outcome in the future (*recourse*) [33, 29]. Unfortunately, feature importance scores—and particularly, two popular methods with widely used packages providing feature importance score methods (LIME [27] and SHAP [23])—have been shown not to live up to

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

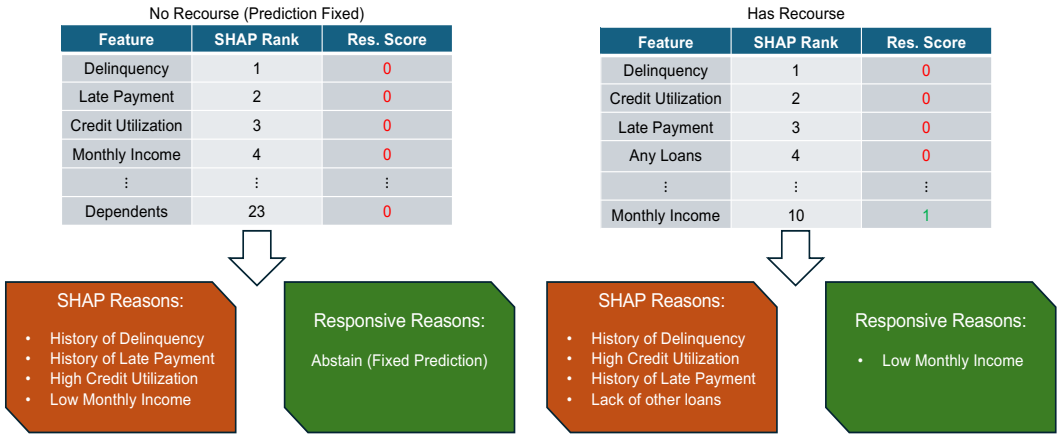

Figure 1: Selected empirical examples that illustrate two failure modes in using existing explanation techniques: (left) providing explanations when the individual's prediction does not change through any actions (a "fixed" prediction), (right) providing non-responsive reasons when there are features, when acted upon, flip the prediction. Our responsiveness score can address both cases of potential harm by (left) flagging fixed predictions and (right) returning responsive features. Note that feature names have been simplified for readability and space considerations.

some of these goals. LIME and SHAP are subject to gaming to hide discrimination [31], and SHAP has additionally been shown to be confusing to data scientists [16], thus undercutting its usefulness in debugging, and to have inherent flaws as a feature importance method [20]. Alternative feature scores have been proposed to help identify potential discrimination [2, 24] and to help data scientists better understand, and thus debug, their machine learning models [25]. While methods to create models that can achieve recourse have been introduced [34, 14, 26], the problem of quantifying model responsiveness to changes in individuals' feature values has not been studied.

As such, we propose attribution for *responsiveness*. As seen in Fig. 1, this allows for flagging instances where returning reasons may not be appropriate or providing reasons that guide individuals to improve their future outcomes. Consequently, attributing responsiveness is inline with the aforementioned reasons for why explanations are desirable for machine learning systems.

Our main contributions include proposing a new paradigm to measure individual feature responsiveness that is (i) model-agnostic, (ii) has concrete meaning, and (iii) can be used to flag instances of potential harm.

2 Problem Statement

We formalize the problem of explaining individual predictions of machine learning models in consumer-facing applications. We consider a standard binary classification with model $h : \mathcal{X} \to \mathcal{Y}$, $\mathcal{X} \subseteq \mathbb{R}^d$ and label $y \in \mathcal{Y} = \{0, 1\}$. We assume $h(\boldsymbol{x}_i) = 1$ represents a *target prediction* that is desirable —e.g., $h(\boldsymbol{x}_i) = 1$ if applicant i will repay their loan within 2 years. Our goal is to explain predictions for those who do not receive the target label so that it could allow for these individuals to achieve the desired outcome in the future.

Explaining Individual Predictions The standard approach to explaining the predictions of individuals who receive undesirable outcomes is through the use of feature attribution methods [10]. We represent these methods as *feature-based explainers* and define them below.

Definition 1. Given a model $h : \mathcal{X} \to \mathcal{Y}$ and its training dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, a *feature-based explainer* for the point $\boldsymbol{x}_i$ and the feature $j \in [d]$ is a function $\phi_j(\boldsymbol{x} \mid h, \mathcal{D}) : \mathcal{X} \to \mathbb{R}$.

We write $\phi_j(\boldsymbol{x})$ instead of $\phi_j(\boldsymbol{x} \mid h, \mathcal{D})$ when h and $\mathcal{D}$ are clear from context. Definition 1 captures a large set of techniques to explain the individual predictions of a model through their features:

- Local Linear Explainers [e.g., 27, 38, 37, 6]: Given a model h and a point $\boldsymbol{x}_i$, the methods fit an approximate linear model $g : \mathbb{R}^d \to \mathbb{R}$ in the neighborhood around $\boldsymbol{x}_i$ such that $g(\boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x}_i) \cdot \boldsymbol{x}'$,

where $\boldsymbol{\phi}(\boldsymbol{x}_i) = [\phi_1(\boldsymbol{x}_i), \phi_2(\boldsymbol{x}_i) \ldots, \phi_d(\boldsymbol{x}_i)]$.

- Shapley Value Methods [e.g., 23, 12, 11]: Methods based on the Shapley value from cooperative game theory [30]. They treat features as "players" in the game and $\phi_j(\boldsymbol{x}_i)$ represents feature j's marginal contribution to the outcome—$h(\boldsymbol{x}_i)$. The "local accuracy" property (Property 1) in [23] ensures that $\sum_{j \in [d]} \phi_j(\boldsymbol{x}_i) = h(\boldsymbol{x}_i) - \bar{y}$ for baseline $\bar{y}$. Notably, [23] shows that only Shapley value based methods satisfy desirable properties like local accuracy.

Given a model h, its training dataset $\mathcal{D}$, and features for an individual $\boldsymbol{x}_i$, these methods output a vector of scores for each feature $[\phi_1(\boldsymbol{x}_i), \ldots, \phi_d(\boldsymbol{x}_i)]$. The top-scoring features are then presented in explanations. This practice is motivated by the following properties:

Relevance: A feature with a feature attribution score $\phi_j(\boldsymbol{x}_i) = 0$ is not locally relevant in the prediction for $\boldsymbol{x}_i$ [i.e., the "missingness" axiom as per 23].

Strength: A feature attribution score $|\phi_j(\boldsymbol{x}_i)|$ reflects the contribution of feature j for the prediction of $\boldsymbol{x}_i$. If $|\phi_j(\boldsymbol{x}_i)| > |\phi_{j'}(\boldsymbol{x}_i)|$, then feature j has a stronger contribution to the prediction than j'.

Reasons without Recourse One of the largest failure modes of machine learning in consumer-facing applications is that models can assign *fixed predictions* – i.e., predictions that cannot be changed by their decision subjects. In practice, these predictions may be accurate yet unjust. In lending tasks, for example, applicants who are denied on the basis of fixed predictions would be correctly denied a loan, yet inadvertently precluded from access to credit.

Fixed predictions are cases where it is impossible to provide feature based explanations to help individuals achieve the target prediction, because every possible way to change the features would return the same prediction. In this case, existing approaches would still generate explanations – presenting individuals with *reasons without recourse*. This can lead to harm by misleading individuals into investing effort into cases that cannot be changed. In effect, an individual may receive an explanation that highlights the "important features" that determined their prediction, that could be changed, but that would not allow them to attain a desired outcome [34].

Specifying Actionability Constraints on Features Models often contain features that one cannot change arbitrarily. As such, actions upon features can be more complicated than changing single feature values. Thus, to make realistic and accurate claims regarding the availability of recourse and feature responsiveness, we introduce machinery to capture actionability.

Each *action* is a vector $\boldsymbol{a} = [a_1, \ldots, a_d] \in \mathbb{R}^d$ that a person can perform to change their features from $\boldsymbol{x}_i$ to $\boldsymbol{x}_i + \boldsymbol{a} = \boldsymbol{x}' \in \mathcal{X}$. We refer to the set of all actions from $\boldsymbol{x}_i$ as an *action set* [see 34, 17].

Definition 2. Given a point $\boldsymbol{x}_i \in \mathcal{X}$, the *action set* $A(\boldsymbol{x}_i)$ contains all possible actions for $\boldsymbol{x}_i$. We assume that every action set contains the *null action* $\mathbf{0} \in A(\boldsymbol{x}_i)$.

In practice, an action set consists of a collection of *actionability constraints* that we can elicit from human experts. As shown in Table 3, we can express actionability constraints in natural language, and convert them into equations that we can embed into an optimization problem [see 17]. This is a compact representation has practical benefits for elicitation and implementation. For example, we can elicit constraints for all $\boldsymbol{x}_i$, store them as functions, and instantiate the constraints for a specific point $\boldsymbol{x}_i$ programmatically (which reduces elicitation and storage).

We define the subset of $A(\boldsymbol{x}_i)$ that contain single-feature actions:

Definition 3. Given an action set $A(\boldsymbol{x}_i)$ for a point $\boldsymbol{x}_i \in \mathcal{X}$, the *single-feature action set* for feature $j \in [d]$ is defined as $A_j(\boldsymbol{x}_i) := \{\boldsymbol{a} \in A(\boldsymbol{x}_i) \mid \boldsymbol{a}_j \neq 0 \wedge \boldsymbol{a}_k = 0, k \in [d] \setminus C_j\}$

Definition 3 captures settings where actions on a on a single feature can induce changes in other features through the set of mutually constrained features in j - $C_j \subseteq [d]$ – i.e., the subset of features subset of features that can change when a person changes feature j. Such changes can arise in tasks where there are deterministic causal relationships between features – e.g., changing a feature such as `years_in_residence` should result in a proportional change in other temporal features such as `age`. However, they can capture dependencies that would not be included a traditional causal graph – e.g., changing a categorical attribute will require switching on a binary feature "off" while turning another binary feature "on" (so that $(a_j = 1 \rightarrow 0 \implies a'_j = 0 \rightarrow 1)$.

3 Responsiveness Scores

Our goal is to measure the *responsiveness* of a model prediction for a point $\boldsymbol{x}_i$ with respect to actions on its features. Responsiveness reflects sensitivity of a model's prediction to such actions. In particular, we wish to identify features that an individual, $\boldsymbol{x}_i$, can change to flip the model's prediction.

Definition 4. Given a model $h : \mathcal{X} \to \mathcal{Y}$, a point $\boldsymbol{x}_i$ with action set $A(\boldsymbol{x}_i)$ and feature $j \in [d]$, the *responsiveness score* for feature j is the probability that the prediction of the model $h(\boldsymbol{x}_i)$ will change as a result changing feature j.

$$\mu_j(\boldsymbol{x}_i \mid h, A(\boldsymbol{x}_i)) := \Pr(h(\boldsymbol{x}') = 1 \mid \boldsymbol{x}' = \boldsymbol{x}_i + \boldsymbol{a}, \boldsymbol{a} \in A_j(\boldsymbol{x}_i))$$

We write $\mu_j(\boldsymbol{x})$ instead of $\mu_j(\boldsymbol{x} \mid h, A(\boldsymbol{x}_i))$ when the model and action set are clear from context.

Given a point with scores $\mu_1(\boldsymbol{x}_i), \ldots, \mu_d(\boldsymbol{x}_i)$, $\mu_j(\boldsymbol{x}_i) = p$ means that that $p\%$ of the possible actions on feature j would lead to change in the prediction of the model. Thus, $\mu_j(\boldsymbol{x}_i) = 0.0$ represents a feature that can be set to any value without changing the outcome. Likewise, a $\mu_j(\boldsymbol{x}_i) = 1$ reflects a feature where any change would lead to a prediction in the model.

Guarantees for Recourse Verification

Remark 1 (Recourse Availability)**.** *Given a model $h : \mathcal{X} \to \mathcal{Y}$, let $\mu_1(\boldsymbol{x}_i), \ldots, \mu_d(\boldsymbol{x}_i)$ denote the responsiveness scores for $\boldsymbol{x}_i \in \mathcal{X}$ with respect to the action set $A(\boldsymbol{x}_i)$. If $\mu_j(\boldsymbol{x}_i) > 0$ for some feature $j \in [d]$, then h can provide recourse to $\boldsymbol{x}_i$ through a single feature action on j.*

Remark 1 implies that when we explain predictions by choosing the features that attain the top responsiveness score, we will always present features that can lead to recourse. As we show in Section 4, this overcomes one of the key limitations of feature attribution methods, which would output explanations in such cases.

A prediction with recourse may still have $\mu_j(\boldsymbol{x}_i) = 0$ for all features $j \in [d]$. In such cases, $\boldsymbol{x}_i$ can only change their prediction through joint actions – i.e., actions that would change multiple features at the same time.

Safeguards for Consumer Protection Responsiveness scores can flag instances where individuals are assigned fixed predictions, or where explanations must be presented with care. This is a built-in safety mechanism that can protect against the possibility of misleading consumers by providing them with reasons without recourse, or inadvertently implying feature independence. We formalize this behavior through Remark 2:

Remark 2 (No 1D Recourse)**.** *Given a model $h : \mathcal{X} \to \mathcal{Y}$, let $\mu_1(\boldsymbol{x}_i), \ldots, \mu_d(\boldsymbol{x}_i)$ denote the responsiveness scores for a point $\boldsymbol{x}_i \in \mathcal{X}$ with respect to the action set $A(\boldsymbol{x}_i)$. If $\mu_j(\boldsymbol{x}_i) = 0$ for all features $j \in [d]$, then one of the following must hold: (I) h assigns a fixed prediction to $\boldsymbol{x}_i$; or (II) h can only provide recourse to $\boldsymbol{x}_i$ through actions that alter two or more features.*

The conditions in Remark 2 act as a "sensor" that can draw attention to instances that warrant further inspection.

4 Experiments

In this section, we present an extensive empirical study on the responsiveness of explanations. Our goals are to (1) evaluate how responsiveness scores can resolve these issues and safeguard against harm; (2) demonstrate the limitations of current feature attribution methods, and their ability to explain away fixed predictions. We include additional details and results in Appendix C, and provide code to reproduce our results at anonymized repository.

Setup We work with classification datasets (see Table 2) that are publicly available and used in prior work (see Appendix C.1 for a detailed description). Each dataset pertains to a standard lending task where each instance represents a consumer and the labels indicate whether they repaid a loan. The goal is to train a model to predict if a consumer will default on a loan. We define a set of *inherent actionability constraints* for each dataset that capture indisputable requirements for all individuals –

e.g., to prevent changes to immutable attributes (i.e. `age`), preserve feature encoding (i.e. one-hot or thermometer encoding), and enforce deterministic causal effects.

We split each dataset into a training sample (80%; used for training and tuning) and a test sample (20%; used to evaluate out-of-sample performance). We use the training sample to fit models using: (1) *logistic regression* (LR); (2) *XGBoost* (XGB); (3) and *random forests* (RF). For each model, we identify all consumers who are denied a loan. We then generate explanations for each individual denied a loan using our responsiveness score RESP. We benchmark these explanations against those produced using standard explainability techniques: (LIME, SHAP), which are used in industry to explain machine learning models [10]. We also consider *action-aware* variants of these techniques to study the viability of adapting solutions to our setting. We evaluate the responsiveness of explanations where we selectthe top-4 highest scoring features for each technique. This setup captures how lenders comply with existing regulations such as the adverse action requirement in the United States (see Appendix C.2 for more detail).

Table 1: Recourse feasibility across datasets and model classes. *% Denied* – the percentage of people predicted a 0 by the classifier; *% Fixed* – the percentage among denied individuals with fixed predictions; *1-D Rec.* – percentage among denied individuals with single-feature actions that lead to recourse.

Dataset	Metrics	LR	RF	XGB
`heloc`	% Denied	56.1%	58.3%	57.0%
$n = 5842$ $d = 43$	↳ % Fixed	19.1%	28.1%	49.1%
FICO [9]	↳ % 1-D Rec	44.4%	34.6%	29.8%
`german`	% Denied	22.9%	17.5%	22.0%
$n = 1000$ $d = 36$	↳ % Fixed	7.4%	29.1%	15.5%
Dua and Graff [5]	↳ % 1-D Rec	73.4%	51.4%	65.5%
`givemecredit`	% Denied	24.6%	24.7%	24.8%
$n = 120268$ $d = 23$	↳ % Fixed	15.6%	0.2%	11.5%
Kaggle [13]	↳ % 1-D Rec	72.4%	93.2%	76.0%

Results and Discusion We evaluate each method using the following metrics: proportion of denied individuals given reasons (% Presented w/ Reasons), proportion of individuals with all reasons invalid (% All reasons invalid), proportion of individuals with at least 1 and all responsive reason(s) (% At least 1 responsive, % All reasons responsive respectively) and the number of reasons.

On Responsiveness Scores. Our results show that they are always responsive across datasets and models (all reasons are responsive 100% of the time). Our approach has the benefit of avoiding the provision of reasons that do not lead to recourse. In practice, this may lead to situations where the score provides fewer reasons on average – e.g., individuals who receive explanations `german`+LR receive on average 1.9 out of the maximum 4. It can also lead to safeguarding behavior. We see that we will abstain from giving reasons to individuals who either receive fixed predictions or can only change their predictions with joint actions (see e.g., `heloc`+XGB where we only present 29.8% of denied consumers with an explanations, same as the % with 1-D recourse in Table 1).

Table 2: Overview of feature based explanations for all models and feature attribution methods. We report results for LR and XGB models and include results for RF in Appendix C.5. We highlight instances of harm in red, and highlight the technique that provides the most responsiveness explanations for a given model in bold.

		LR					XGB				
		Vanilla		Action-Aware			Vanilla		Action-Aware		
Dataset	Metrics	LIME	SHAP	LIME	SHAP	RESP	LIME	SHAP	LIME	SHAP	RESP
`heloc` $n = 5842$ $d = 43$ FICO [9]	% Presented w/ Reasons	100.0%	100.0%	100.0%	100.0%	44.4%	100.0%	100.0%	100.0%	100.0%	29.8%
	↳ % All reasons invalid	82.0%	75.6%	64.7%	64.7%	0.0%	92.6%	80.7%	77.5%	75.1%	0.0%
	↳ % At least 1 responsive	18.0%	24.4%	35.3%	35.3%	100.0%	7.4%	19.3%	22.5%	24.9%	100.0%
	↳ % All reasons responsive	0.0%	0.0%	0.2%	0.2%	**100.0%**	0.0%	0.0%	0.0%	0.0%	**100.0%**
	↳ # of reasons	4.0	4.0	4.0	4.0	2.4	4.0	4.0	4.0	4.0	2.7
`german` $n = 1000$ $d = 36$ Dua and Graff [5]	% Presented w/ Reasons	100.0%	100.0%	100.0%	100.0%	73.4%	100.0%	100.0%	100.0%	100.0%	65.5%
	↳ % All reasons invalid	100.0%	100.0%	62.9%	66.4%	0.0%	100.0%	83.2%	64.5%	66.8%	0.0%
	↳ % At least 1 responsive	0.0%	0.0%	37.1%	33.6%	100.0%	0.0%	16.8%	35.5%	33.2%	100.0%
	↳ % All reasons responsive	0.0%	0.0%	0.0%	0.0%	**100.0%**	0.0%	0.0%	0.0%	0.0%	**100.0%**
	↳ # of reasons	4.0	4.0	4.0	4.0	1.9	4.0	4.0	4.0	4.0	2.0
`givemecredit` $n = 120268$ $d = 23$ Kaggle [13]	% Presented w/ Reasons	100.0%	100.0%	100.0%	100.0%	72.4%	100.0%	100.0%	100.0%	100.0%	76.0%
	↳ % All reasons invalid	55.8%	45.5%	50.7%	31.8%	0.0%	40.9%	51.3%	30.9%	40.6%	0.0%
	↳ % At least 1 responsive	44.2%	54.5%	49.3%	68.2%	100.0%	59.1%	48.7%	69.1%	59.4%	100.0%
	↳ % All reasons responsive	0.0%	0.0%	5.5%	23.1%	**100.0%**	0.0%	0.0%	5.4%	3.7%	**100.0%**
	↳ # of reasons	4.0	4.0	4.0	4.0	2.4	4.0	4.0	4.0	4.0	2.6

On Existing Approaches. Our results show that these methods provide reasons to all denied individuals, highlighting how these reasons can often be unresponsive. In particular, we see invalid reasons for over 50% of individuals (i.e., features that cannot lead to recourse). In the `german` dataset, both SHAP and LIME explanations for the LR model returned invalid reasons for all rejected individuals. In some cases the lack of validity stems from the fact that these methods provide explanations to individuals who receive fixed predictions. In others, however, it stems from the fact that the methods fail to identify features that provide recourse, even when individuals have single-feature actions that flip the predictions (e.g., in `german`, 73.4% of denied individuals had responsive features under the LR model but were not identified by LIME nor SHAP at all). Fig. 2 shows that the trend continues beyond the top 4. LIME and SHAP rankings do not translate to responsiveness, with evened out mean responsiveness across all ranks.

On the Prevalence of Reasons without Recourse. It is particularly harmful when the reasons are provided to individuals with fixed predictions. LIME, SHAP, and their action-aware variants often provide entirely invalid reasons. Our results show that existing feature attribution methods also provide reasons to individuals who receive fixed predictions. Despite the fact that there are individuals with fixed predictions for each dataset and model class (almost 50% in the case of `heloc`+XGB, see Table 1), LIME, SHAP and their variants always give reasons. This is not entirely due to the fact that existing feature attribution methods cannot take actionability constraints into account, since action filtered variants also seem to consistently return reasons without recourse.

On Adapting Existing Methods. In an attempt to make existing methods compatible with actionability constraints, we have constructed "action-aware" variants of SHAP and LIME where we filter out reasons that map to features that cannot be changed. Our results show that we can mitigate some of the issues discussed previously – for example, in Table 2, we see that these methods provide fewer instances of providing completely invalid reasons and returning at least one responsive features (e.g., from 0% to 33.6% for SHAP under `german`+LR). However, these are not reliable solutions, since their gains are quite marginal and still seem to return reasons without recourse at the same rate as their ordinary counterparts.

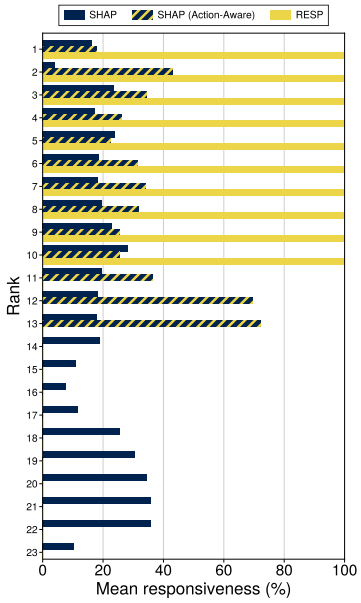

Figure 2: Plot of average responsiveness across ranks for different methods under `givemecredit` and XGB. Note features with attribution 0 are omitted from rankings.

5 Concluding Remarks

In this work, we highlight another potential source of harm: providing reasons without recourse. While adhering to regulations that mandate explanations, agents may unintentionally or intentionally provide misleading explanations, subtle yet arguably more damaging. Our results shows that common feature attribution methods often output reasons without recourse, and that one cannot address the issue with simple solutions like filtering out features based on actionability.

Our work underscores the need for a standalone approach to generate explanations when regulations have multiple motivations – e.g., rectification, anti-discrimination and recourse. Our work's technical contribution, the responsiveness score, accomplishes this by assigning responsiveness to individual features. The responsiveness score shares the characteristics of feature attribution methods, allowing it to be readily integrated into existing approaches. Moreover, it can flag instances of potential harm.

Limitations & Future Work Our implementation of action sets and constraints cannot take probabilistic causality into account. Moreover, our actionability constraints described in Section 4 (and anonymized repository in more detail) are inherent actionability constraints. As a result, one may still fail to present individuals with features that provide recourse because individual actionability constraints may be more stringent. However, we could address this through further elicitation of the individual's constraints.

References

[1] 12 cfr part 1002 - equal credit opportunity act (regulation b). https://www.consumerfinance.gov/rules-policy/regulations/1002/2/. Accessed: 2024-07-16.

[2] Adler, Philip, Casey Falk, Sorelle A Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. Auditing black-box models for indirect influence. *Knowledge and Information Systems*, 54:95–122, 2018.

[3] Bilodeau, Blair, Natasha Jaques, Pang Wei Koh, and Been Kim. Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences*, 121(2):e2304406120, 2024.

[4] Consumer Financial Protection Bureau. Consumer financial protection circular 2023-03: Adverse action notification requirements and the proper use of the CFPB's sample forms provided in regulation B. URL https://www.consumerfinance.gov/compliance/circulars/circular-2023-03-adverse-action-notification-requirements-and-the-proper-use-of-the-cfpbs-sample-forms-provided-in-regulation-b/.

[5] Dua, Dheeru and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

[6] ElShawi, Radwa, Youssef Sherif, Mouaz Al-Mallah, and Sherif Sakr. Ilime: local and global interpretable model-agnostic explainer of black-box decision. In *Advances in Databases and Information Systems: 23rd European Conference, ADBIS 2019, Bled, Slovenia, September 8–11, 2019, Proceedings 23*, pages 53–68. Springer, 2019.

[7] European Parliament, Council of the European Union. Regulation (eu) 2024/1689. https://eur-lex.europa.eu/eli/reg/2024/1689/oj. Accessed: 2024-08-30.

[8] Ficklin, Patrice Alexander, Tom Pahl, and Paul Watkins. Innovation spotlight: Providing adverse action notices when using ai/ml models. https://www.consumerfinance.gov/about-us/blog/innovation-spotlight-providing-adverse-action-notices-when-using-ai-ml-models/. Accessed: 2024-07-16.

[9] FICO. Explainable machine learning challenge, 2018. URL https://community.fico.com/s/explainable-machine-learning-challenge.

[10] FinRegLab. Empirical white paper: Explainability and fairness: Insights from consumer lending. Technical report, FinRegLab, July 2023. URL https://finreglab.org/wp-content/uploads/2023/12/FinRegLab_2023-07-13_Empirical-White-Paper_Explainability-and-Fairness_Insights-from-Consumer-Lending.pdf.

[11] Fumagalli, Fabian, Maximilian Muschalik, Patrick Kolpaczki, Eyke Hüllermeier, and Barbara Hammer. Shap-iq: Unified approximation of any-order shapley interactions. *Advances in Neural Information Processing Systems*, 36, 2024.

[12] Jethani, Neil, Mukund Sudarshan, Ian Connick Covert, Su-In Lee, and Rajesh Ranganath. Fastshap: Real-time shapley value estimation. In *International conference on learning representations*, 2021.

[13] Kaggle. Give Me Some Credit. http://www.kaggle.com/c/GiveMeSomeCredit/, 2011.

[14] Karimi, Amir-Hossein, Gilles Barthe, Borja Balle, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions. In *International Conference on Artificial Intelligence and Statistics*, pages 895–905. PMLR, 2020.

[15] Karimi, Amir-Hossein, Julius Von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *Advances in neural information processing systems*, 33:265–277, 2020.

[16] Kaur, Harmanpreet, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14, 2020.

[17] Kothari, Avni, Bogdan Kulynych, Tsui-Wei Weng, and Berk Ustun. Prediction without preclusion: Recourse verification with reachable sets. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=SCQfYpdoGE.

[18] Krivorotov, George and Jeremiah Richey. Explaining denials: Adverse action codes and machine learning in credit decisioning. *Available at SSRN 4133915*, 2022.

[19] Kulesza, Todd, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*, pages 126–137, 2015.

[20] Kumar, I Elizabeth, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning*, pages 5491–5500. PMLR, 2020.

[21] Lakkaraju, Himabindu and Osbert Bastani. "how do i fool you?": Manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pages 79–85, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375833. URL https://doi.org/10.1145/3375627.3375833.

[22] Lei, Jing, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

[23] Lundberg, Scott M and Su-In Lee. A unified approach to interpreting model predictions. *NeurIPS*, 2017.

[24] Marx, Charles, Richard Phillips, Sorelle Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. Disentangling influence: Using disentangled representations to audit model predictions. *Advances in Neural Information Processing Systems*, 32, 2019.

[25] Melis, David Alvarez, Harmanpreet Kaur, Hal Daumé III, Hanna Wallach, and Jennifer Wortman Vaughan. From human explanation to model interpretability: A framework based on weight of evidence. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 9, pages 35–47, 2021.

[26] Pawelczyk, Martin, Teresa Datta, Johan HeuvelVan den , Gjergji Kasneci, and Himabindu Lakkaraju. Probabilistically robust recourse: Navigating the trade-offs between costs and robustness in algorithmic recourse. In *The Eleventh International Conference on Learning Representations*, 2023.

[27] Ribeiro, Marco Tulio, 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*, pages 1135–1144. ACM, 2016.

[28] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence*, 2018.

[29] Selbst, Andrew D and Solon Barocas. The intuitive appeal of explainable machines. 2018.

[30] Shapley, Lloyd S. A value for n-person games. *Contribution to the Theory of Games*, 2, 1953.

[31] Slack, Dylan, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020.

[32] The Lawyers' Committee for Civil Rights Under Law. Online civil rights act, December, 2023. URL https://www.lawyerscommittee.org/online-civil-rights-act.

[33] U.S. Senate, 94th Congress. Senate report no. 94-589, 1976.

[34] Ustun, Berk, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 10–19. ACM, 2019. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287566.

[35] White House. Blueprint for an AI bill of rights: Making automated systems work for the American people. The White House Office of Science and Technology Policy, October, 2022. URL https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

[36] Wolsey, Laurence A. *Integer programming*. John Wiley & Sons, 2020.

[37] Zafar, Muhammad Rehman and Naimul Mefraz Khan. Dlime: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems. *arXiv preprint arXiv:1906.10263*, 2019.

[38] Zhou, Zhengze, Giles Hooker, and Fei Wang. S-lime: Stabilized-lime for model explanation. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 2429–2438, 2021.

A Related Work

Regulation Our work is motivated by a growing need for consumer protection in applications where we use machine learning systems to automate individuals decisions – i.e., lending, hiring, insurance underwriting and claims. One of the key provisions of existing and proposed regulation and legislation focuses on on providing explanations for individuals who receive adverse decisions [1, 35, 32, 7]. In the US, for example, the adverse action notice requirement in the Equal Credit Opportunity Act mandates lenders to provide "principal reasons" to denied individuals [1]. In the EU, the "Right to an Explanation" in Article 86 of the AI Act [7] grants individuals a right to obtain explanations that describe the "main elements" of their decision within "high-risk" applications.[1] [35, 32] model this idea in a U.S. regulatory and legislative context. Our work highlights how current methods fall short in achieving goals put forth by existing and proposed regulation and legislation. We propose an alternative method that not only addresses those shortcomings but also can easily be implemented in existing pipelines.

Local Feature Attributions Our work is related to a stream of methods for local feature attribution [27, 23, 28, 22]. These methods explain the individual predictions of a model through their features. Many are designed as general-purpose solutions — i.e., for multiple use cases. These include using them to better understand models [16], and to assess and debug them [19]. This versatility can restrict effectiveness in specific tasks. This is due to the fact that (1) there may be no universal notion of "importance"; and (2) what is "important" in a specific application can consist of multiple properties. One of the salient examples is in consumer protection. Recent work has shown that these methods can mislead end-users [21]. Our work highlights a different failure mode: these methods can harm consumers by presenting them with "reasons without recourse"—a silent failure mode where individuals are shown features that led to their model outcome but changing those features to receive a different outcome is not possible.

Our results complements recent impossibility results in feature attribution. For instance, [3] show that explainability methods satisfying completeness and linearity, like SHAP, perform no better than random guessing in inferring model behavior, particularly in the context of recourse, consistent with our empirical results.

Recourse and Actionability Our work is part of a broader effort to develop methods for algorithmic recourse [34, 15]. Few have been adopted in practice due to lack of legislative mandates and interference with model development . A key issue is that stakeholders cannot readily deploy these methods as they may suggest actions that only guarantee recourse with specific changes, and that alter multiple features.

[1] Annex III in the act provides a definition for what applications constitute as "high-risk".

B Supplementary Information for Section 2

Table 3: Examples of actionability constraints on semantically meaningful features from a lending task from [17]. It shows how each constraint can be expressed in natural language and embedded into an optimization problem using standard techniques in mathematical programming [see, e.g., 36]. We highlight constraints that couple actions across features because they can only be enforced using special kinds of search algorithms.

Class	Example	Features	Actionability Constraint
Immutability	`n_dependents` should not change	$x_j =$ `n_dependents`	$a_j = 0$
Monotonicity	`reapplicant` can only increase	$x_j =$ `reapplicant`	$a_j \geq 0$
Integrality	`n_accounts` must be positive integer ≤ 10	$x_j =$ `n_accounts`	$a_j \in \mathbb{Z} \cap [0 - x_j, 10 - x_j]$
Categorical Encodings	preserve one-hot encoding of `married`, `single`	$x_j =$ `married` $x_k =$ `single`	$a_l + x_l \in \{0,1\} \quad x_k + a_k \in \{0,1\}$ $a_j + x_j + a_k + x_k = 1$
Ordinal Encoding	preserve one-hot encoding of `max_degree_BS`, `max_degree_MS`	$x_j =$ `max_degree_BS` $x_k =$ `max_degree_MS`	$a_j + x_j \in \{0,1\} \quad x_k + a_k \in \{0,1\}$ $a_j + x_j + a_k + x_k = 1 \quad a_j + x_j \geq a_k + x_k$
Logical Implications	if `is_employed` $=$ `TRUE` then `work_hrs_per_week` ≥ 0 else `work_hrs_per_week` $= 0$	$x_j =$ `is_employed` $x_k =$ `work_hrs_per_week`	$a_j + x_j \in \{0,1\}$ $a_k + x_k \in [0, 168]$ $a_j + x_j \leq 168(x_k + a_k)$
Causal Implications	if `years_of_account_history` increases then `age` will increase commensurately	$x_j =$ `years_at_residence` $x_k =$ `age`	$a_j \leq a_k$

C Experiment Details

C.1 Description of Datasets

German The first dataset we used is called `german`, and it's used to predict somebody's credit risk by classifying them as good or bad based on predictive features. It was created in 1994 and contains information about loan history, demographics, occupation, payment history, and whether or not somebody is a good customer.

Each instance is a real person with credit. There are 1,000 instances, each consisting of 20 features. The features are all either categorical or discrete. The label, class, is a binary indicator of whether somebody is a 'good' (label=1) or 'bad' (label=2) customer. We changed these labels to be 0 and 1.

There are no missing values in the dataset. We renamed some of the features to be indicative of the values they represent. The dataset is self-contained and anonymous, and it includes features describing gender, age, and marital status.

We preprocessed the data by one-hot encoding sex, marital status, and years employed being at least 1. We thermometer encoded credit amount and loan duration, which means that we created 4 binary features out of each and let the new features signify that the feature value is greater than or equal to some value in the range of possible feature values. We also combined multiple columns into one-hot encoded columns to make the features simpler; for example, if the purpose of the loan was for a new or used car, we recorded that the loan was required for a car. Most variables in the processed dataset are binary.

`givemecredit` Our second dataset is used to develop credit scoring algorithms. Similar to FICO, it's used to determine whether a loan should be given or denied. The label is whether somebody was 90 days past delinquency in the two years following data collection. Delinquency is a debt with an overdue payment; this dataset is used to predict if someone will experience financial distress in the next 2 years.

It contains information about 150,000 loan recipients, and each instance is a real borrower. There are 10 features before preprocessing. The label is SeriousDlqin2yrs, meaning serious delinquency in 2 years. In preprocessing, we change the label to instead be `NotSeriousDlqin2yrs` so that 1 is a positive classification and 0 is negative, in keeping with the other datasets we used.

The data is self-contained and anonymous, and contains features describing age, income, and number of dependents. We preprocessed the data by thermometer encoding age, number of dependents, monthly income, and credit line utilization. We also binarized the rest of the features.

heloc The third dataset we used was a FICO Risk score dataset. The FICO dataset was created to predict repayment on Home Equity Line of Credit (HELOC) applications. HELOC credit lines are loans that use people's homes as collateral. The dataset is used by lenders to determine how much credit should be granted. The anonymized verion of the HELOC dataset was created by FICO to put forth an explainable machine learning challenge for a prize.

Each instance in the dataset is a real credit application for HELOC credit; its an application that a single person submitted and contains information about that person. There are 10,459 instances, each consisting of 23 features. These features are either binary or discrete. The label, RiskPerformance, is a binary assessment of risk of repayment based on the 23 predictors. 1 means the person hasn't been more than 90 days overdue on their payments in the last 2 years; 0 means they have at least once. There are some repeated instances; there are 9,871 unique rows. The dataset is self-contained, and has been anonymized for public use in the explainability challenge. It doesn't use any protected attributes like race and gender.

We preprocessed the data by thermometer encoding many of the features. Clearly, this makes the features correlated; for example, any time Years In File is greater than or equal to 5, it's also greater than or equal to 3. We also replaced features based on months to be in terms of years.

C.2 Experiment Setup and Comparison with Real-World Practices

"Maximum of 4 Reasons" Our decision to use consider the top 4 features comes from the Consumer Financial Protection Bureau (CFPB)'s interpretation of ECOA. The interpretation states that "the regulation does not mandate that a specific number of reasons be disclosed, but disclosure of more than four reasons is not likely to be helpful to the applicant" [1].

Using post-hoc explainability methods While agents, including lenders, do not disclose how they generate explanations to comply with regulation, several studies [e.g. 18, 10] on adverse action notice requirements have used LIME and SHAP as methods to generate reasons. CFPB's circular in 2020 highlighted the use of explainability tools to comply with regulation while using complex black-box models [8].

Reason codes In lending applications, lenders do not return raw features directly to consumers. Instead they map them to reason codes, which are meant to be more holistic and interpretable for consumers [10]. We have omitted this process since the mapping mechanism is not publicly available.

C.3 Compute Information

We used CPLEX v22.1 to generate the reachable set on a kubernetes pod in an internal research cluster, with 2 CPU cores and 16 GB memory. We used the same computing infrastructure to generate explanations.

C.4 Model Performance

Table 4: Train and Test AUC for models across all datasets. We optimized the model's hyperparameters through randomized search and divided the data into training and testing sets at an 80% and 20% ratio.

	LR		XGB		RF	
Dataset	Train	Test	Train	Test	Train	Test
`heloc` $n = 5842$ $d = 43$ FICO [9]	0.772	0.788	0.859	0.785	0.780	0.790
`german` $n = 1000$ $d = 36$ Dua and Graff [5]	0.819	0.760	0.971	0.794	0.828	0.766
`givemecredit` $n = 120268$ $d = 23$ Kaggle [13]	0.841	0.844	0.875	0.793	0.864	0.835

C.5 Additional Experiment Results

Table 5: Experiment results for RF. We highlight instances of harm in red, and highlight the technique that provides the most responsiveness explanations for a given model in bold.

		RF				
		Vanilla		Action-Aware		
Dataset	Metrics	LIME	SHAP	LIME	SHAP	RESP
heloc $n = 5842$ $d = 43$ FICO [9]	% Presented w/ Reasons	100.0%	100.0%	100.0%	100.0%	34.6%
	↳ % All reasons invalid	85.1%	78.2%	74.1%	74.4%	0.0%
	↳ % At least 1 responsive	14.9%	21.8%	25.9%	25.6%	100.0%
	↳ % All reasons responsive	0.0%	0.0%	0.0%	0.0%	**100.0%**
	↳ # of reasons	4.0	4.0	4.0	4.0	2.5
german $n = 1000$ $d = 36$ Dua and Graff [5]	% Presented w/ Reasons	100.0%	100.0%	100.0%	100.0%	51.4%
	↳ % All reasons invalid	100.0%	87.4%	71.4%	60.0%	0.0%
	↳ % At least 1 responsive	0.0%	12.6%	28.6%	40.0%	100.0%
	↳ % All reasons responsive	0.0%	0.0%	0.0%	0.0%	**100.0%**
	↳ # of reasons	4.0	4.0	4.0	4.0	2.5
givemecredit $n = 120268$ $d = 23$ Kaggle [13]	% Presented w/ Reasons	100.0%	100.0%	100.0%	100.0%	93.2%
	↳ % All reasons invalid	60.0%	39.6%	28.7%	17.6%	0.0%
	↳ % At least 1 responsive	40.0%	60.4%	71.3%	82.4%	100.0%
	↳ % All reasons responsive	0.0%	0.0%	0.8%	12.7%	**100.0%**
	↳ # of reasons	4.0	4.0	4.0	4.0	2.9

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction reflect our contributions in Section 3 and our empirical results in Section 4.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The assumptions for the problem statement are provided in Section 2. As for theoretical results, they are either definitions or trivial remarks that do not need written proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We include an anonymized repository that includes code to run the experiment.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  1. If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  2. If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  3. If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  4. We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We include an anonymized repository that includes code and the datasets required to run the experiment.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We outline all training and test details in Section 4 and additional details in Appendix C.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Error bars are not included due to running the experiment several times will be computationally expensive.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide computer resource details in Appendix C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have reviewed the Code of Ethics. Our work does not violate the Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the societal impact of our work in Section 1 and Section 5.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our work does not pose risk for misuse of datasets and models.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We have cited the original sources of code packages and datasets used in our work.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We include an anonymized repository of our code.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our work does not involve human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our work does not involve human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.