Dissenting Explanations: Leveraging Disagreement to Reduce Model Overreliance

Omer Reingold [1] **Judy Hanwen Shen** [1] **Aditi Talati** [1]

Abstract

While explainability is a desirable characteristic of increasingly complex black-box models, modern explanation methods have been shown to be inconsistent and contradictory. The semantics of explanations is not always fully understood – to what extent do explanations "explain" a decision and to what extent do they merely advocate for a decision? Can we help humans gain insights from explanations accompanying *correct* predictions and not over-rely on *incorrect* predictions advocated for by explanations? With this perspective in mind, we introduce the notion of dissenting explanations: conflicting predictions with accompanying explanations. We first explore the advantage of dissenting explanations in the setting of model multiplicity, where multiple models with similar performance may have different predictions. In such cases, providing dissenting explanations could be done by invoking the explanations of disagreeing models. Through a pilot study, we demonstrate that dissenting explanations reduce overreliance on model predictions, without reducing overall accuracy. Motivated by the utility of dissenting explanations we present both global and local methods for their generation.

1. Introduction

The development of increasingly capable AI systems has motivated many fields to consider AI-assisted decision-making. In high-stakes settings such as loan approval and patient diagnosis, it is imperative for humans to understand how any given model came to its decision. However, with the success of deep learning, many large state-of-the-art models are not easily interpretable. Thus, explainability (XAI) methods are crucial for providing justification for the decisions of black-box models. Such explanations justify a model's prediction on a singular input example, and their goal is to provide accurate information while also being succinct and easy for humans to parse (Burkart & Huber, 2020). In fact, a recent study found that explanations can help improve human performance and even reduce overreliance on AI on specific tasks such as solving a difficult maze (Vasconcelos et al., 2022).

While explanations can serve as verification for certain tasks like maze completion, many predictive tasks are not verifiable in nature (e.g. predicting the probability of a loan default). In these cases, many different explanations can be used to explain a decision. In fact, recent works have shown that explanations generated from different methods based on the same instance can conflict (Han et al., 2022; Krishna et al., 2022). Furthermore, different AI models with similar performances may vastly differ in predictions as well as explanations (Marx et al., 2020; Black et al., 2022). Instead of rejecting explanations altogether, the existence of multiple plausible explanations motivates the perspective that explanations can be treated as arguments supporting a given model prediction, rather than a verifiable proof for a given prediction.

With the framework of explanations as arguments, we may naturally construct a courtroom analogy, in which human decision makers are the judges deciding whether the model prediction is trustworthy. When a singular explanation is provided, a decision-maker may be unduly influenced to trust the prediction. Indeed, Bansal et al. (2021) show that when explanations are provided, humans are more likely to follow a model decision regardless of whether the model is correct. Thus, while an explanation provides a supporting argument for a prediction, we must also provide alternative arguments, arguing against the model prediction, in order to accommodate meticulous human decision-making. In the context of a consequential legal decision, presenting both sides amounts to procedural due process.

In this paper, we introduce the notion of *dissenting explanations*: explanations for an opposing model prediction to some reference model. To illustrate the importance of

[*]Equal contribution [1]Department of Computer Science, Stanford University, Palo Alto, USA. Correspondence to: Judy Hanwen Shen <jhshen@stanford.edu>, Aditi Talati <atalati@stanford.edu>.

AI HCI Workshop at the 40th International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA. 2023.,

these explanations, we study existing model disagreement on a deceptive hotel reviews classification task. We perform a pilot study to show that, on this difficult-to-verify task, dissenting explanations indeed reduce model overreliance without reducing the accuracy of the human predictions. Finally, since dissenting explanations are a useful tool for reducing overreliance, even outside the context of existing model multiplicity, we develop methods to induce predictive multiplicity and create dissenting explanations. We present techniques for generating global disagreement with respect to any black-box model, as well as local disagreement on any instance; these methods achieve disagreement without sacrificing model accuracy.

2. Related work

One model, multiple explanations Post-hoc explanations can be elicited from black box models through a variety of techniques including perturbation-based methods (e.g. LIME and SHAP (Ribeiro et al., 2016; Lundberg & Lee, 2017)) and gradient-based methods (e.g. GradCAM and SmoothGrad (Selvaraju et al., 2017; Smilkov et al., 2017)). However, when applying such techniques to the same example, inconsistent and conflicting explanations for feature importance may arise. Surveying data scientists, Krishna et al. (2022) found that disagreements in explanations occur when the top features or the ordering of features is different and developed explanation disagreement metrics and find that more complex models exhibit higher disagreement.

Similar models, conflicting explanations For models to give trustworthy predictions, humans may expect stable predictions with explanations across similarly accurate models. However, models with similar may have similar accuracy but different predictions and different model internals and decision-making processes (i.e. *predictive multiplicity*). Brunet et al. show that models similar in performance can yield vastly different explanations. Seemingly trivial choices in model architectures, random seeds, and hyperparameters may lead to inconsistent and contradicting explanations.

Overreliance and human-AI collaboration Among tasks where neither humans nor AI routinely achieves perfect performance, Lai & Tan (2019) use AI predictions and explanations to help human participants with detecting deceptive hotel reviews and find that human performance was improved with AI predictions with explanations. Vasconcelos et al. (2022) also find that explanations actually reduce overreliance in their set of maze task experiments. In contrast, Bansal et al. (2021) study common sense tasks including review sentiment classification and LSAT question answering and found that explanations increased accuracy when the AI model was correct but decreased accuracy when the AI model was wrong. However, they do observe that highlighting the features for the top two classes when presenting a *single* model prediction reduced overreliance on the AI recommendation.

Leveraging model and explanation multiplicity, we investigate the effect of also showing the explanation of a dissenting model in reducing overreliance. Specifically, we are motivated by settings where AI surpasses human performance, but human decision-makers may need make the final decision (Lai & Tan, 2019; Lundberg & Lee, 2017). In these settings, the goal is to provide AI predictions with explanations to humans as a tool rather than removing humans from decision-making altogether. Crucially, our setting differs from (Bansal et al., 2021) in that we examine differing independent predictions and accompanying explanations from different models in the Rashomon set (Fisher et al., 2019) with the goal of improving human decision-making.

3. Model and framework

We define dissenting explanations in the situation where we have model multiplicity. Let $f, g : \mathcal{X} \to \mathcal{Y}$ be two different functions trained on the same data $x, y \sim \mathcal{D}$; these functions do not have to belong to the same hypothesis class. We look at the specific case of binary classification ($y \in \{0, 1\}$), but much of this work can also be extended to general classification tasks. Then, let $e(f, x)$ be an explanation for the model's prediction $f(x)$. The shape of e depends on the type of explanation being used, and any of the standard explanation methods will produce a valid function e. Based on these definitions, we introduce the concept of a *dissenting explanation* as an explanation of the prediction of a disagreeing model:

Definition 3.1 (Dissenting Explanation)**.** Let f, g be any two different classifiers and let $(x, y) \sim \mathcal{D}$ be any example. Then, $e(x, g)$ is a *dissenting explanation* for $e(x, f)$ if $f(x) \neq g(x)$.

Dissenting explanations offer an argument for a contradictory prediction; each disagreeing model can produce its own dissenting explanation. Furthermore, dissenting explanations are explanation-method agnostic. In the more general setting of multi-class classification, the explanation $e(g, x)$ is a dissenting explanation for $e(f, x)$ as long as g predicts a label different from $f(x)$.

Since disagreeing predictions are necessary for dissenting explanations, measuring how many predictions f and g disagree on gives an indication of how many dissenting explanations can be generated between two models.

Definition 3.2 (Global predictive disagreement)**.** Let f, g be any two different classifiers, the global disagreement

between f and g on some set D is:

$$\delta_D(f, g) = \frac{1}{|D|} \sum_{x \in D} \mathbb{1}[f(x) \neq g(x)]$$

Remark 3.3. Let $\mathrm{Err}_D(f) = \frac{1}{|D|} \sum_{(\mathbf{x},\mathbf{y}) \in D} \mathbb{1}[f(x) \neq y]$ be the empirical error of a classifier. For two classifiers f and g where $\mathrm{Err}_D(f), \mathrm{Err}_D(g) \in [0, 1]$:

$$\delta_D(f, g) \leq \mathrm{Err}_D(f) + \mathrm{Err}_D(g)$$

This can be seen by considering that disagreement is maximized when f and g make mistakes on disjoint sets.

While disagreement can be maximized by finding a model that predicts differently on every example, we focus on the setting where models are similarly accurate. This set of models that are similarly accurate on some dataset is also described as the Rashomon set (Fisher et al., 2019).

Following prior work studying the overreliance of humans on AI predictions (Vasconcelos et al., 2022), we define overreliance as how much human decisions mirror AI suggestions when the AI is incorrect: $\mathbb{E}[h(x) = f(x) | f(x) \neq y]$ where h represents the human decision.

For the purposes of our experiments, we let $e(x, f) \in \mathbb{R}^d$ be a *feature attribution explanation* of f on x. A feature attribution explainer generates a linear "surrogate model" that approximates f in a neighborhood of x. If the weights of the linear surrogate model are w_i, then $e(x, f)$ returns the most important features x_i, corresponding to the d largest values of $|w_i x_i|$. In our experiments, these feature attribution explanations are generated by LIME TextExplainer (Ribeiro et al., 2016).

For a feature attribution explanation, we say that $e(x, f)_+ \in \mathbb{R}^p$ are the set of features supporting the prediction $f(x) = 1$ while $e(x, f)_- \in \mathbb{R}^n$ are the set of features that support the prediction $f(x) = 0$ where $p + n = d$.

4. Motivating study: the importance of dissenting explanations

4.1. Hypothesis

Motivated by the potential of dissenting explanations to present an alternative argument against a model prediction, we seek to understand whether dissenting explanations can be helpful in reducing human overreliance on model predictions. To this end, we propose two hypotheses:

> HYPOTHESIS 1 (**H1**): Providing users with a singular explanation for an incorrect AI prediction increases human agreement with the incorrect prediction.

> HYPOTHESIS 2 (**H2**): Providing users with a dissenting explanation, arguing against the AI prediction, along with the explanation, will decrease human over-reliance without significantly decreasing human accuracy, as compared to providing a single AI prediction and explanation.

The purpose of the first hypothesis was to provide a baseline for how explanations affect human decisions, while the second hypothesis tests the value of dissenting explanations.

4.2. Study design

Task selection We focus on the setting of assistive AI: the setting where AI on average might perform better than humans but it is critical for humans to be the final decision maker. This is different from prior works, which focused on tasks either with verifiable answers given the explanation (Vasconcelos et al., 2022) or tasks where humans and AI perform approximately equally in order to measure collaboration potential (Bansal et al., 2021). Furthermore, we specifically consider explanations that are not verifiable proofs of the correct label but rather arguments for the model predictions, as these are the standard explanation forms available for complex model predictions (Burkart & Huber, 2020). Thus, we had the following criteria in selecting a task: (1) The human accuracy for the task must be less than the model accuracy. (2) There must be room for model disagreement on the task; the AI model should not perform the task perfectly. (3) There must be an objective correct label for the examples. (4) AI explanations must be understandable for the participants, without providing complete proof of the correct answer.

Deceptive reviews task Based on our requirements, we decided to use the Chicago Deceptive Reviews dataset (Ott et al., 2011). This is a dataset of 1600 one-paragraph reviews of hotels in Chicago, where half the reviews are genuine reviews from TripAdvisor, and the other half were written by crowd workers that have only seen the name and website of the hotel. The goal of the task is to distinguish between real and deceptive reviews; a prior study found that humans on their own get at most 62% accuracy on this task, while a linear SVM achieved around 87% accuracy (Lai & Tan, 2019). Furthermore, there exists a ground truth label: whether a review is deceptive or real. The explanations were in the form of highlighting the words selected by the feature attribution explainer; these words serve as an argument to the participant, convincing them to select a certain label without giving a complete proof of the correct answer.

To test our hypothesis, we design a study in which human participants attempt to categorize these hotel reviews. Participants are presented with 20 hotel reviews, each of which is real or deceptive, and are instructed to decide which reviews

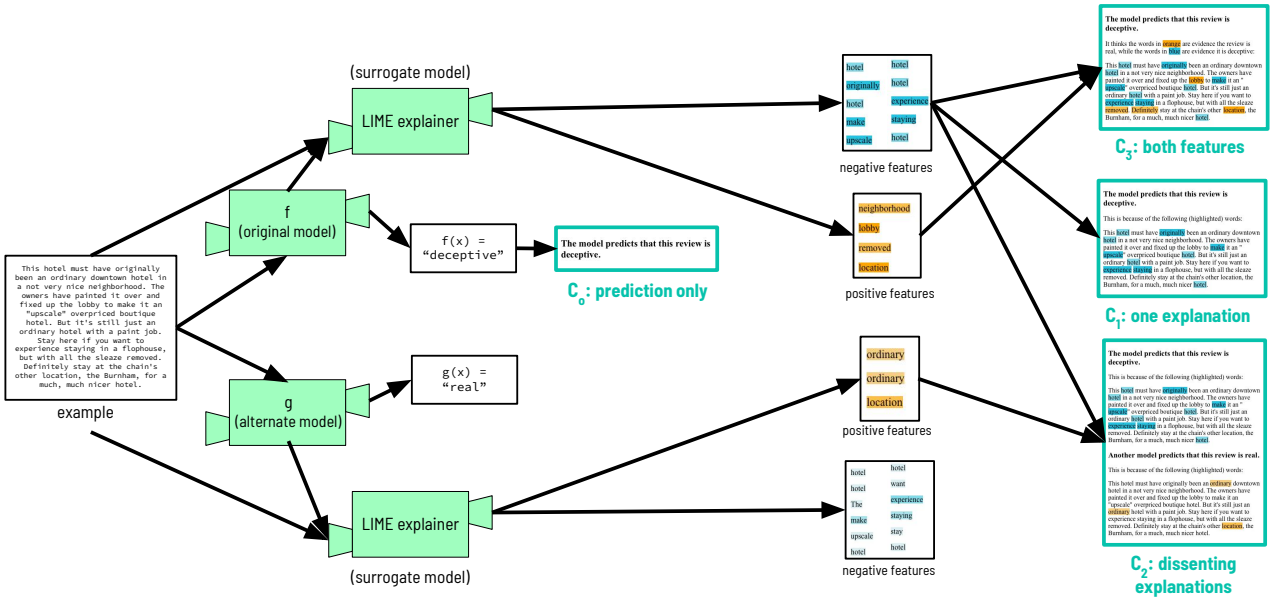


Figure 1: The process of generating explanations for the 4 conditions.

are real. They are assisted by AI predictions or explanations, where the existence or type of explanation varies based on the condition participants are assigned to. Participants are warned in the beginning that the AI predictions are not always correct, and they are also given a set of heuristics for identifying deceptive reviews, developed by prior work on this task (Lai et al., 2020). We also survey users, post-task, about how difficult they found the task, how effective the AI suggestions were in helping categorize the reviews, and how much they trusted the AI suggestions. Finally, there is an optional open-ended question about how the AI suggestions helped them complete the task. These questions allow us more insight into whether people felt they trusted the model.

Generating explanations To properly benchmark against prior work (Lai & Tan, 2019), we use the same linear SVM trained on TF-IDF of unigrams with English stop words removed as our reference model f with 87% accuracy. We also train an alternative 3-layer neural network model based on the exact same pre-processing which achieves 79% accuracy as the alternative mode g. We use LIME (Ribeiro et al., 2016) to generate local explanations for each model using the Top-15 features. To double-check the quality of the LIME features, we compared the features to the weights of the linear SVM model and found meaningful overlap between the top features. In order to find dissenting explanations, we used examples in the test set where the neural network model disagreed with the linear SVM model. The two models disagreed on about 10% of examples in the test set which produced 32 examples. We sub-sampled these examples for an even balance of examples that the linear SVM (the reference model) predicted correctly and incorrectly.

Conditions Each participant was presented with the same 20 reviews, along with the same 20 model predictions. The reference model f predicted the incorrect label on 8 of the 20 reviews. We randomly assigned each participant to one of the following four conditions. Participants are not aware of the other possible conditions for the study.

- C_0: Participants were presented with the AI prediction for each review, without any explanation.
- C_1: Participants were presented with the AI prediction for each review $f(x)$, along with a supporting explanation $e(x, f)_{f(x)}$. This means either the positive LIME features were highlighted in orange, if the model predicted "real," or the negative LIME features were highlighted in blue, if the model predicted "deceptive".
- C_2: Participants were presented with both the explanation and the dissenting explanation. They received the same explanation as in C_1, followed by the line "Another model predicts that this review is [real/deceptive]" and the corresponding explanation for the dissenting model.
- C_3: Participants were presented with an explanation that more closely matched the original LIME output, which includes both positive and negative features. Each explanation started with the line "The model predicts that this review is [real/deceptive]. It thinks the words in orange are evidence the review is real, while

the words in blue are evidence it is deceptive." This was followed with the corresponding highlighted text.

The four conditions can be seen in Figure 1. We provided participants with training before the task began that was specifically tailored to the condition that each participant is assigned to. All other aspects of the survey, such as the format, the reviews, the predictions themselves, and the post-survey questions, were kept constant across all four conditions. The main purpose of our study was to compare C_2 to C_1. C_0 was our control condition and C_3 was included so we could compare the effect of dissenting explanations to the pre-existing feature attribution, which contains both positive and negative features.

Participants These surveys were posted on Prolific and made available to all fluent English speakers that have at least a 95% approval rate on Prolific and have not answered any previous surveys we have posted. Participants were given training examples at the beginning of the survey. Participants were compensated $\$3.50$ USD for participating in the task and given an additional bonus of $\$1.00$ USD if they answered more than half the questions correctly. For the average completion time of ~ 15 minutes, this translates to a $\$18$ USD hourly rate. Three attention check questions were included in the study where participants were told explicitly to select a certain answer. We excluded answers from participants who failed more than one attention check but still compensated these participants. After excluding the failed attention checks, there were $N = 178$ submissions in our analysis, with approximately 45 submissions per condition[1]. Our sample size was calculated based on pilot studies[2].

4.3. Results

We measure average accuracy on this task for each condition as an indicator for how well users learn from different types of explanations. Moreover, we measure overreliance on the model when the predictions were incorrect to understand how the different types of explanations affect overreliance on model predictions.

Quantitative findings For each participant in the study, we measured their **accuracy** as the fraction of reviews they categorized correctly, out of the 20 total reviews. We measured **overreliance** as the fraction of reviews they agreed with the model prediction on, out of the 8 reviews the model predicted incorrectly. These results, averaged over each of the four conditions, are displayed in Figure 2a and Figure 2b. Since our task involves binary labels, we account for random agreement by also measuring Cohen's κ between a participant and the model's predictions (McHugh, 2012).

Using a one-way ANOVA test, we find that accuracy does not differ across conditions ($p = 0.850$), but overreliance and Cohen's κ scores do differ across conditions ($p = 0.007$ and $p = 0.0001$, respectively). We then perform one-tailed t-tests between conditions to test our specific hypotheses.

To analyze H1, we compared Cohen's κ scores over the 8 questions that the model predicted incorrectly, between C_0 and C_1. Our 1-tailed t-test did not give statistically significant results ($p > 0.05$). We also did not find significantly larger overreliance ($p > 0.05$).

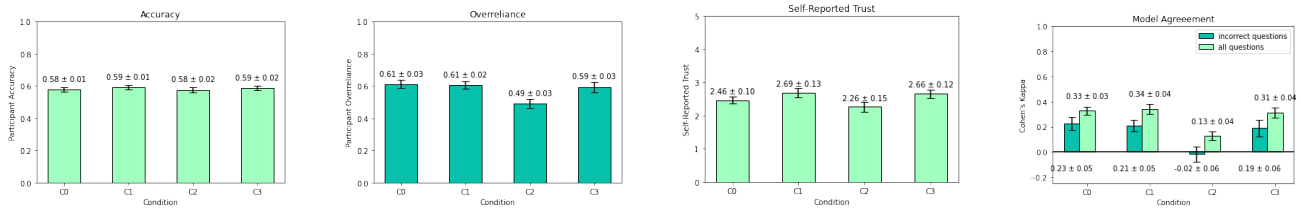
We find that our results support our main hypothesis (H2): providing participants with both a supporting and dissenting explanation (C_2) significantly reduces overreliance as compared to just a single explanation (C_1) (Figure 2b, $p = 0.001$). Moreover, the dissenting explanation does *not* significantly reduce accuracy ($p = 0.210$). With one explanation (in condition C_1), participants get an average accuracy score of 0.593 ± 0.014, but an overreliance of 0.606 ± 0.023. Meanwhile, when provided with both a supporting and dissenting explanation (condition C_2), participants get an average accuracy of 0.576 ± 0.016, with an overreliance of 0.491 ± 0.029. We also observe that for human-model agreement, as measured by Cohen's κ, dissenting explanations in condition C_2 also give a significantly lower agreement with model predictions than just a single explanation C_1 ($p = 7e^{-5}$ for all questions). Our results suggest providing dissenting explanations is a useful way to reduce overreliance in situations where it is unclear whether the model prediction is accurate.

Moreover, participants in C_3, who saw both the positive and negative features from a singular model explanation, had an average overreliance score of 0.595 ± 0.032. Thus, in our experiment, the method of dissenting explanations (C_2) produced lower overreliance as compared to C_3 ($p = 0.009$). This shows that dissenting explanations provide a benefit beyond what is provided by existing explanation methods, and there is a significant difference in how humans react to positive evidence from one model and negative evidence from another, as opposed to positive and negative evidence from a singular model in this deception labeling task.

Qualitative analysis Participants were asked to report their trust in the AI predictions, on a 5-point scale from "*not at all*" to "*a great deal*". The reported trust matched the trend of the overreliance scores across the 4 conditions, where the average reported trust in the model predictions was lowest in the dissenting explanations condition, and higher in the other 3 conditions (Figure 2c). This was reflected in participant comments; one comment for C_0 was "*If i was on the fence on wheter it was fake or not i tried to listen to the AI suggestion*", and many others had a similar

[1] Our study obtained IRB exemption approval (Stanford IRB-70387)

[2] Pre-registration: https://osf.io/hrv5m/



(a) Accuracy (b) Overreliance (c) Trust (d) Agreement

Figure 2: Accuracy, overreliance, reported trust, and Cohen's κ score for each experimental condition. Error bars represent standard error across participants (N=178).

sentiment. For condition C_2, there were many comments saying they distrusted the AI suggestions, and a few saying that they followed the suggestion with the more-highlighted paragraph. Similarly, in C_3, there were many comments such as "*It helped me to easily identify the ammount of key words of each type.*" Thus, the participants' beliefs about the study generally reflected the quantitative results we found for each of the explanation conditions.

5. Finding dissenting explanations

Motivated by the potential of dissenting explanations for reducing overreliance on explanations, we present methods for producing disagreement in models. While prior works have focused on predictive multiplicity, a clear mapping between predictive multiplicity and explanation multiplicity has not been presented. Furthermore, previous techniques for maximizing predictive multiplicity through mixed integer programming are limited to the linear models (Marx et al., 2020). In this section, we present and compare methods for increasing predictive multiplicity through the lens of explanations.

5.1. Global model disagreement: a model agnostic approach

We consider the setting where we have access to a reference model f and the training set. Our goal is to train a model g which will disagree with f as much as possible on a subsequent test set.

Problem 5.1. *Given reference model f and training data D, find some g such that $\delta_D(f, g)$ (Definition 3.2) is maximized while* $\text{Err}_{D_{test}}(f) \approx \text{Err}_{D_{test}}(g)$.

Regularization (REG) First, we consider a regularization approach to penalize similarities between a *fixed* reference model f predictions and the current model g. Specifically, one empirical loss we can minimize is:

$$L(x, y, f) = \frac{1}{n}\sum_{i=1}^{n} l(g(x_i), y_i) + \frac{\lambda}{n}\sum_{i=1}^{n} \mathbb{1}[f(x_i) \neq g(x_i)] \tag{1}$$

However, since the indicator function is not continuous and non-differentiable, we modify the objective to be:

$$L(x, y, f) = \frac{1}{n}\sum_{i=1}^{n} l(g(x_i), y_i) + \frac{\lambda}{n}\sum_{i=1}^{n} l(g(x_i), \overline{f(x_i)}) \tag{2}$$

We consider the binary classification setting and set l to be the binary cross entropy loss and use the inverse predictions of f to maximize disagreement between f and g.

Reweighting (WEIGHTS) Leveraging intuition from boosting, another approach to learning a maximally differing classifier is to upweight examples to our reference predictor gets wrong. Our approach differs from traditional boosting in that we are comparing explanations between resulting models instead of combining model outputs for a single prediction. Formally, the reweighting objective is as follows:

$$L(x, y, f) = \frac{1}{n}\sum_{i=1}^{n} w_i l(g(x_i), y_i) \tag{3}$$

$$w_i = 1 + \lambda \mathbb{1}[f(x_i) \neq y_i)] \tag{4}$$

Remark 5.2. When $l(x, y) = \mathbb{1}[x \neq y]$, in the binary setting this reweighting objective is equivalent to the above regularization objective.

Experiment results First, we compare predictive multiplicity induced by both methods on the deceptive reviews dataset and use the same reference model f, a linear-SVM, from our human-centered studies. For all experiments in this section, we train a neural network g with a single hidden layer with the same features as the reference model f. The results presented are averaged over 5 different random seeds. Table 1a summarizes the overall model accuracy, the percentage of examples f and g disagreed on, and the percentage of examples that were incorrectly predicted by f but rectified by g. All of these metrics are computed over a held-out test set. As λ increases, the number of conflicting prediction examples also increases[3]. However, this effect

[3]Training with the REG objective using larger λ (e.g. $\lambda \geq 1$) resulted in instabilities for a variety of hyperparameters.

λ	Accuracy	Disagreement	Corr.
0.0	$0.889 \pm .010$	8.66 ± 0.6 %	40.1 %
0.1	$0.883 \pm .017$	8.75 ± 0.5 %	38.9 %
0.25	$0.859 \pm .021$	10.9 ± 3.4 %	34.2 %
0.5	$\mathbf{0.807 \pm .017}$	$\mathbf{16.6 \pm 2.3}$ %	**35.7 %**

(a) REG objective (batch size 10)

λ	Accuracy	Disagreement	Corr.
0	$0.859 \pm .019$	8.68 ± 0.7 %	28.4%
1	$0.865 \pm .014$	8.56 ± 1.2 %	30.5%
10	$0.854 \pm .008$	10.8 ± 1.5 %	35.3%
50	$\mathbf{0.826 \pm .018}$	$\mathbf{14.9 \pm 0.7}$ %	**40.1%**

(b) WEIGHTS objective (batch size 100)

Table 1: Comparison of proposed methods to elicit predictive multiplicity against a 88% accuracy reference model (f). WEIGHTS requires a larger batch size since the reference model is very accurate. Corr. indicated the percentage of f's incorrect predictions which were corrected by g.

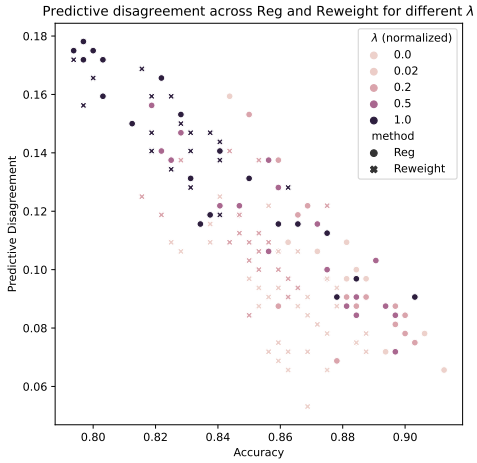
might be due to g simply getting more examples wrong. Thus, it is important to measure the number of f's incorrect predictions that are corrected by g. Of the total 38 examples in the test set that f predicts incorrectly, the percentage of correct examples reduces slightly as disagreement increases.

Table 1b summarizes the effectiveness of using the WEIGHTS objective in creating model predictive disagreement. Disagreement is achieved without as much sacrifice in overall accuracy. Furthermore, both the percentage disagreement and corrected samples are high at larger λ values. To further compare the two approaches, Figure 3a shows a Pareto plot of accuracy vs disagreement. We see that for both methods, as λ increases, accuracy and agreement decrease. Furthermore, comparing across 10 models trained to disagree with the reference model, there were about 28% of test set examples where at least 1 model disagreed with the reference model.

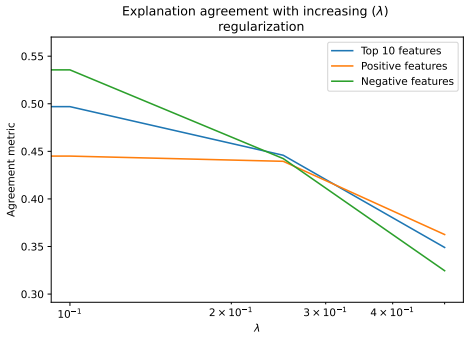
Explanation disagreement For comparing dissenting explanations to the original explanation, we use a similar set of metrics as explanation agreement (Krishna et al., 2022). There are two main cases to consider: when the two models agree and when the two models disagree. When the models agree, Krishna et al. (2022) present several metrics to measure agreement between explanations including top k features and rank correlation. We consider three agreement metrics:

$$\text{TOPK} = \frac{|\text{top}_k(e(x,f)) \cap \text{top}_k(e(x,g))|}{|\text{top}_k(e(x,f)) \cup \text{top}_k(e(x,g))|} \quad (5)$$

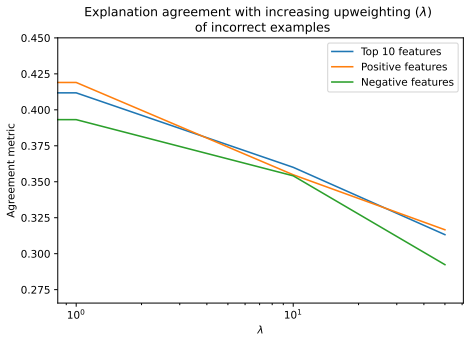
$$\text{TOPKPOS} = \frac{|\text{top}_k(e(x,f))_+ \cap \text{top}_k(e(x,g))_+|}{|\text{top}_k(e(x,f))_+ \cup \text{top}_k(e(x,g))_+|} \quad (6)$$



(a) Accuracy vs disagreement across different λ



(b) Explanation Agreement from REG



(c) Explanation agreement from WEIGHTS

Figure 3: (a) Adjusting λ results in trade-offs between accuracy and disagreement. However, these tradeoffs depend on the objective function used (REG, WEIGHTS) and hyperparameters (e.g. batch size). (b)-(c) As we emphasize the importance of model predictive disagreement through increasing λ, the agreement between explanations as measured by the overlap in top features also decreases.

TOPKNEG is just TOPKPOS with negative prediction features instead of positive. To measure explanation agreement, we evaluate models at different λ for both REG (Figure

3b) and WEIGHTS (Figure 3c). For all three metrics, as λ increases, the explanation agreement also reduces. Although these results are unsurprising, they are compelling in illustrating that creating predictive multiplicity also in turn produces explanation multiplicity. Moreover, since a good portion of examples that the reference model classified incorrectly were rectified in our alternative models, this explanation multiplicity allows dissenting explanations to aid human judgment and reduce overreliance.

5.2. Local model disagreement: generating a dissenting explanation for any input

|D|	Success Rate	TOPK Agree.	Acc.
1280	$0.543 \pm .249$	$0.756 \pm .131$	0.880
640	$0.723 \pm .200$	$0.464 \pm .122$	0.889
320	$0.910 \pm .082$	$0.352 \pm .111$	0.844
160	$0.987 \pm .013$	$0.275 \pm .115$	0.780
80	$1.000 \pm .000$	$0.227 \pm .103$	0.675

(a) SVM

Iter.	Freq.	TOPK Agree.	Acc.
<5	19.7%	$0.946 \pm .091$	0.902
5-10	20.9%	$0.878 \pm .113$	0.892
10-15	18.1%	$0.786 \pm .117$	0.886
15-20	19.1%	$0.770 \pm .159$	0.883
>20	22.2%	$0.782 \pm .114$	0.869

(b) Neural Network

Table 2: (a) Success rate, TOPK agreement between f and g, and test set accuracy of g when adding a test instance to the training set for a flipped prediction. As dataset size decreases, the test instance is more likely to be successfully predicted as the opposite class. (b) Training iterations required to find g, TOPK agreement between f and g, and test set accuracy of g, a neural network model that is retrained on the test instance. With more training iterations on the single instance, the prediction for the instance is more likely to flip. Errors reported are standard deviation for all values and variance for Success rate (Bernoulli). All test set accuracy errors are between 0.00 and 0.02.

While the techniques we presented increase model disagreement on the test data only with the training data, the total coverage only spans $< 30\%$ of points for our dataset. We now consider an alternative problem formulation where the test instance for which we want to achieve a different prediction is given. This allows us to produce a dissenting explanation for any input example in the reference model.

Problem 5.3. *Given reference model f, training data D, and a test instance x, find some g where $f(x) \neq g(x)$ where* $\text{Err}_{D_{\text{test}}}(f) \approx \text{Err}_{D_{\text{test}}}(g)$.

Unlike global disagreement, we know the exact test instance for which we want a different (flipped in the binary case) prediction. For different model classes f, we present different methods for this problem. For models where parameters are directly solved for, we propose adding the test instance to a subset of the training data. Intuitively, as the training data size decreases, the influence of the test instance should increases thus increasing the success rate of generating a different prediction for x. Table 2a shows this intuition for a Linear SVM model (based on the same features and hotel reviews dataset). The success rate, calculated over all the examples in the test set, increases as the total size of the dataset decreases, showing that Problem 5.3 can be solved with high probability (i.e. $> 90\%$) without sacrificing significant model accuracy. For models that are fitted with stochastic optimization (e.g. neural networks), we can directly minimize over just the test instance and measure how many iterations are required to change the label. Table 2b describes the distribution over iterations required to flip an example label. The difficulty varies depending on the example and accuracy on the other test examples declines with more iterations. However, this method is also effective in flipping the label for roughly 80% of the test set examples while still maintaining $\sim 88\%$ accuracy.

6. Discussion

In this work, we take a holistic approach by first motivating the need for dissenting explanations through a human study to measure overreliance. For our deceptive reviews task, a task with a ground truth label but no method for direct verification, we demonstrate the utility of dissenting explanations in reducing overreliance. Our results complement existing work on the benefits of explanations (Vasconcelos et al., 2022) by exploring more ambiguous tasks.

After finding that overreliance can be reduced by introducing dissenting explanations which argue against a model prediction and explanation, we then present simple but effective heuristics for eliciting more disagreement between models with only query access to the reference model. We show that generating disagreement in predictions is sufficient for generating different explanations. Our work serves as a first step in presenting the human interaction and computational challenges in treating explanations as arguments for predictions in AI decision-making.

A promising direction of future work is to explore what tasks dissenting explanations best aid and other types of dissenting explanations involving counterfactual explanations. Furthermore, while we presented a preliminary intuition on explanation disagreement from one vs two models, more studies are required to understand whether these results are due to a difference in the selected features between the two conditions, or simply a difference in framing.

Acknowledgments

This research was supported by the Simons collaboration on the theory of algorithmic fairness and the Simons Foundation Investigators Award 689988. Also thanks to Lindsay Popowski and Aspen Hopkins for the insightful discussions on experimental design.

References

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., and Weld, D. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2021.

Black, E., Raghavan, M., and Barocas, S. Model multiplicity: Opportunities, concerns, and solutions. 2022.

Brunet, M.-E., Anderson, A., and Zemel, R. Implications of model indeterminacy for explanations of automated decisions. In *Advances in Neural Information Processing Systems*.

Burkart, N. and Huber, M. F. A survey on the explainability of supervised machine learning. *CoRR*, abs/2011.07876, 2020. URL https://arxiv.org/abs/2011.07876.

Fisher, A., Rudin, C., and Dominici, F. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, 20(177):1–81, 2019.

Han, T., Srinivas, S., and Lakkaraju, H. Which explanation should i choose? a function approximation perspective to characterizing post hoc explanations. *arXiv preprint arXiv:2206.01254*, 2022.

Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., and Lakkaraju, H. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*, 2022.

Lai, V. and Tan, C. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 29–38, 2019.

Lai, V., Liu, H., and Tan, C. " why is' chicago'deceptive?" towards building model-driven tutorials for humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2020.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Marx, C., Calmon, F., and Ustun, B. Predictive multiplicity in classification. In *International Conference on Machine Learning*, pp. 6765–6774. PMLR, 2020.

McHugh, M. L. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

Ott, M., Choi, Y., Cardie, C., and Hancock, J. T. Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557*, 2011.

Ribeiro, M. T., Singh, S., and Guestrin, C. " 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*, pp. 1135–1144, 2016.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M., and Krishna, R. Explanations can reduce overreliance on ai systems during decision-making. *arXiv preprint arXiv:2212.06823*, 2022.