The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)

Dissenting Explanations: Leveraging Disagreement to Reduce Model Overreliance

Omer Reingold[*], **Judy Hanwen Shen**[*], **Aditi Talati**[*]
Stanford University
{reingold, atalati, jhshen}@stanford.edu

Abstract

While modern explanation methods have been shown to be inconsistent and contradictory, the explainability of black-box models nevertheless remains desirable. When the role of explanations extends from understanding models to aiding decision making, 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. Through a human study on the task of identifying deceptive reviews, 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.

Introduction

The development of increasingly capable AI systems has motivated the adoption of 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 and Huber 2020).

Wang and Yin (2021) list the desiderata of explanations as (1) model understanding, (2) helping people recognize model uncertainty, and (3) calibrating trust for AI models. Towards the goal of understanding a single model, recent works have shown that explanations generated from different methods based on the same instance can conflict (Han, Srinivas, and Lakkaraju 2022; Krishna et al. 2022). While multiple explanations for the same *model* question the validity of explainability tools, diverse and even conflicting explanations for the same *decision* might better calibrate trust and illustrate uncertainty in a model-aided decision-making setting. 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 recommendation in the decision-making process.

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 these explanations, we focus, from human study to proposed techniques, on a single task: deceptive hotel reviews classification. We perform a 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 [1].

---

[*] These authors contributed equally.

[1] Full paper with appendices available at https://arxiv.org/abs/2307.07636

Related Work

One Model, Multiple Explanations Post-hoc explanations can be elicited from black box models through techniques including perturbation-based methods (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017) and gradient-based methods (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 disagreements in feature explanations and find that more complex models exhibit higher disagreement.

Similar Models, Conflicting Explanations For trustworthy predictions, humans may expect similar predictions from similarly accurate models. Yet models with similar accuracy still exhibit different predictions (Marx, Calmon, and Ustun 2020). Brunet, Anderson, and Zemel (2022) show that similar accuracy models can yield vastly different explanations due to different random seeds and hyperparameters. Other works in robustness purposefully generate disagreement between models (Pang et al. 2019; Rame and Cord 2021).

Overreliance and Human-AI Collaboration Among tasks where neither humans nor AI routinely achieves perfect performance, Lai and 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 found that explanations increased accuracy when the AI model was correct but decreased accuracy when the AI model was wrong. An adjacent line of work studies AI debate: two large language models are prompted to recursively try to convince a human judge to take their side (Michael et al. 2023).

We investigate the effect of also showing the explanation of a dissenting model in reducing overreliance in settings where AI surpasses human performance, but human decision-makers may need to make the final decision (Lai and Tan 2019; Lundberg and 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. The closest prior work to ours includes a position paper arguing for decision support systems to provide support for and against decisions (Miller 2023) and the effect of using different explanations from the same model (Bansal et al. 2021). In contrast, we examine differing independent predictions and accompanying explanations from different models in the Rashomon set (Fisher, Rudin, and Dominici 2019) with the goal of improving human decision-making.

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 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 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. 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 $\mathrm{Err}_D(f), \mathrm{Err}_D(g) \in [0, 1]$, then:

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

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 models that disagree on every example, we care about the setting of models with similarly high accuracy, also described as the Rashomon set (Fisher, Rudin, and Dominici 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 $|w_i x_i|$. For a feature attribution explanation, we denote $e(x, f)_+ \in \mathbb{R}^p$ as 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 $p + n = d$.

Motivating Study: The Importance of Dissenting Explanations

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. 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 overreliance 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. We limit the scope of these hypotheses to a specific task which we investigate thoroughly.

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). We use the term "verifiable" to describe tasks where explanations can verify that a model prediction is true (e.g. a maze task explanation provides a path to complete the maze or a math task explanation provides the step-by-step solution). 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 and Huber 2020). Since the effects of explanations differ widely from task to task (Wang and Yin 2021), we focus on finding a specific, suitable task for studying dissenting explanations. We set the following requirements for our target task[2]:

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 without domain expertise and do not provide verifiable proof for the correct answer.

Deceptive Reviews Task Based on our requirements, we decided to use the Chicago Deceptive Reviews dataset (Ott et al. 2011): 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 and 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 hypotheses, 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 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. Based on response quality in pilot studies, participants are also given a set of heuristics for identifying deceptive reviews, developed by prior work on this task (Lai, Liu, and Tan 2020). We also survey users, post-task, about task difficulty, the helpfulness of AI suggestions, and their trust in the AI suggestions. Finally, we included an optional open-ended question about how the AI suggestions helped them complete the task.

Generating Explanations To properly benchmark against prior work (Lai and 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 model g. We use LIME (Ribeiro, Singh, and Guestrin 2016) to generate local explanations for each model using the Top-k features (k=15 was chosen based on the number of unique tokens per example for this dataset).

We used LIME for generating explanations in order to (1) remain consistent against prior work XAI studies in the text domain (Bansal et al. 2021), and (2) benchmark against (Lai and Tan 2019) who use linear model features (i.e. we compared our LIME features to the weights of the linear SVM model and found meaningful overlap[3]). 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 10% of examples (32 examples). We sub-sampled these examples for an even balance of examples that the linear SVM (the reference model) predicted correctly and incorrectly. This selection process yields a subset of examples that result in $\sim 60\%$ accuracy; both for the reference model f and human baseline.

---

[2] We elaborate on these desiderata and discuss the candidate tasks we considered in the supplementary materials.

[3] See details in in the supplementary materials

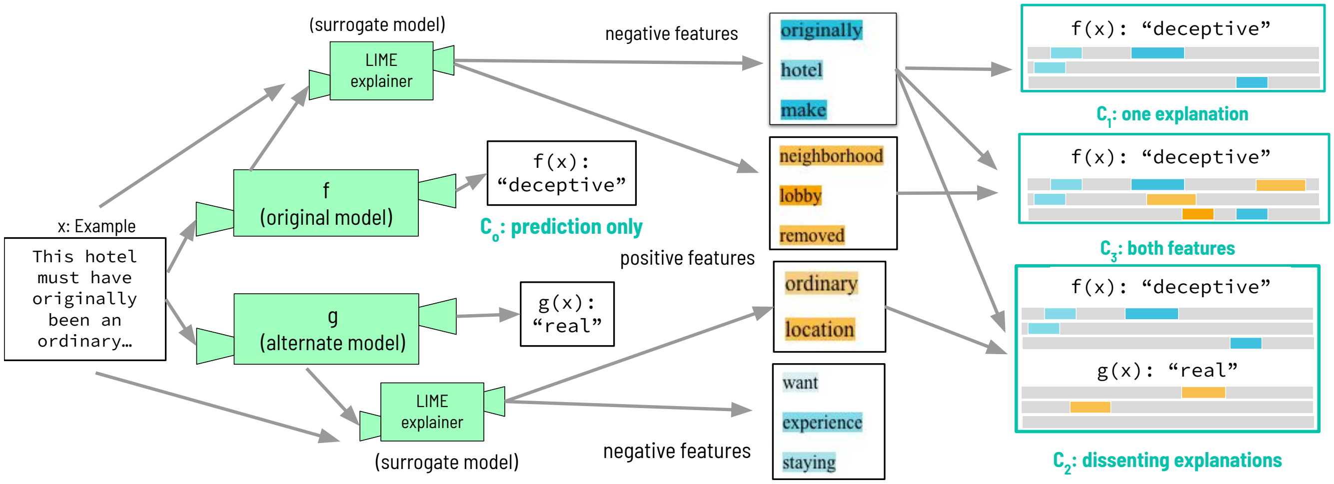


Figure 1: The process of generating explanations for the 4 conditions. The review text is for conceptual understanding only.

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 (Figure 1). 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 features were highlighted in orange, if the model predicted "real," or the negative 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.

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. To test **H1** and **H2**, we can compare C_2 to C_1. The baseline conditions we were also interested in understanding are provided by C_0 (prediction only) and C_3 (1 explanation containing both positive and negative features).

Participants Our study was run on Prolific and made available to all fluent English speakers that have at least a 95% approval rate and have not answered any previous surveys (including pilot studies) we have posted. 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 $\sim$ 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 according to IRB feedback. After excluding the failed attention checks, there were $N = 178$ submissions in our analysis, with approximately 45 submissions per condition[4]. Our sample size was calculated based on pilot studies.

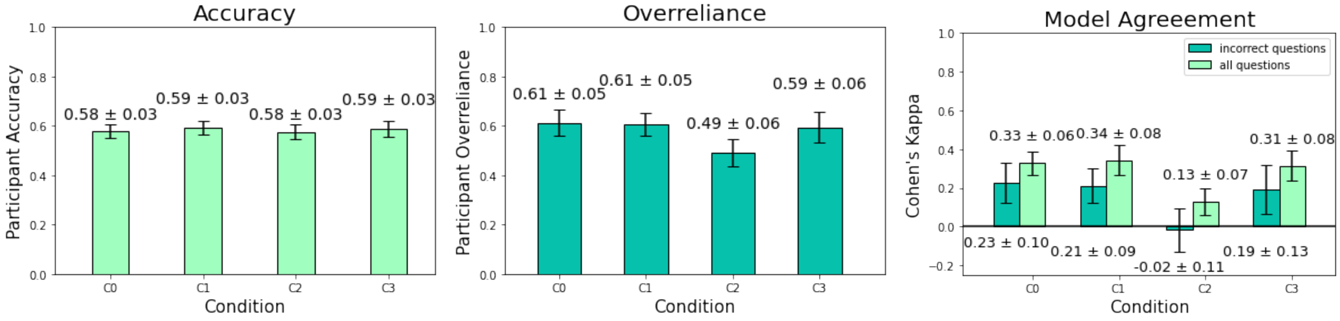
Results

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 users of 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 in Figure 2c (McHugh 2012).

Using a one-way ANOVA test, we find that accuracy does not differ significantly, but overreliance ($p = 0.007$) and Cohen's κ ($p < 0.001$) scores do differ across conditions. We then perform one-tailed t-tests between conditions to test our specific hypotheses. To analyze H1, we examined 8 questions that the model predicted incorrectly; both Cohen's κ and the overreliance were not greater in C_0 than C_1.

---

[4] Our deceptive reviews task obtained IRB exemption approval and our study was pre-registered



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

Figure 2: (a) Accuracy, (b) Overreliance, and (c) Agreement (Cohen's κ score) for each experimental condition demonstrating that dissenting explanations C_2 significantly reduce overreliance without reducing overall accuracy. Error bars represent 95% confidence intervals of the mean across participants (N=178).

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) ($p < 0.001$, Figure 2b). 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 ($p < 0.001$, Figure 2c). However, since the dissenting explanation condition C_2 does *not* significantly reduce accuracy ($p > 0.05$, Figure 2a), this suggests that dissenting explanations reduce human-model agreement as well as overreliance. If the decrease in agreement only resulted in underreliance (i.e. human disagreement with the model prediction when the model is correct, also increased due to decreased agreement), we would observe a drop in overall accuracy.

To understand whether the reduction in overreliance arises from simply the presentation of more information or increased cognitive load, we can compare C_2 with C_3, where participants saw both the positive and negative features from a single model explanation. We find that condition C_3 produced significantly higher overreliance compared to the C_2 condition ($p = 0.009$). This shows that dissenting explanations may help calibrate trust; 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.

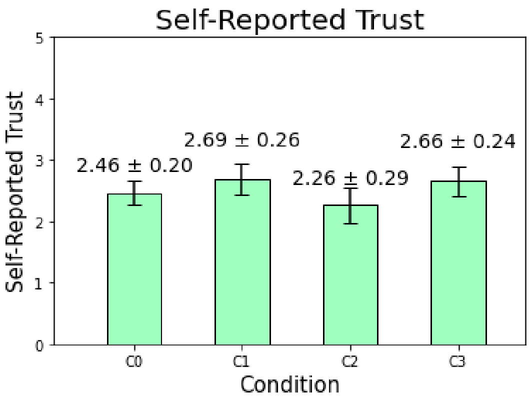


Figure 3: Level of trust reported by participants on a scale of (1) "*not at all*" to (5) "*a great deal*". The level of trust was significantly lower for C_2 (dissenting explanations condition)

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 3). 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 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.

Finding Dissenting Explanations

Motivated by the utility of dissenting explanations, we present methods for producing disagreement in models. While many techniques may be developed to create model multiplicity, we focus on simple objective-based approaches that are effective for our task of interest similar to techniques from robustness and diverse ensembles (Pang et al. 2019; Rame and Cord 2021)[5]. While prior works have focused on predictive multiplicity, a clear mapping between predictive and explanation multiplicity has not been presented. In this

---

[5] Additional experiments for these techniques on other datasets are included in the supplementary materials

section, we present and compare methods for increasing predictive multiplicity through the lens of explanations.

Global 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 suggest simple but intuitive methods to train a model g which will disagree with f as much as possible on a subsequent test set.

Problem 4. ***Global Disagreement*** *Given reference model f and training data D, find some g such that* $\delta_D(f, g)$ *(Definition 2) is maximized while* $\mathrm{Err}_{D_{test}}(f) \approx \mathrm{Err}_{D_{test}}(g)$.

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 that 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. 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 study. 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 computed over a held-out test set. As λ increases, the number of conflicting prediction examples also increases[6]. However, this effect 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 corrected 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 and both the percentage disagreement and corrected samples are high at larger λ values.

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

(a) REG objective (batch size 10)

λ	Accuracy	Disagreement	Corr.
0	0.859 ± .019	8.68 ± 0.7 %	28.4%
1	0.865 ± .014	8.56 ± 1.2 %	30.5%
10	0.854 ± .008	10.8 ± 1.5 %	35.3%
50	**0.826 ± .018**	**14.9 ± 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 to ensure an incorrect example is included, Corr. indicates the percentage of f's incorrect predictions which were corrected by g.

Explanation Disagreement For comparing dissenting explanations to the original explanation, we use a similar set of metrics as explanation agreement (Krishna et al. 2022). We consider three agreement metrics TOPK, TOPKPOS, and TOPKNEG:

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

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

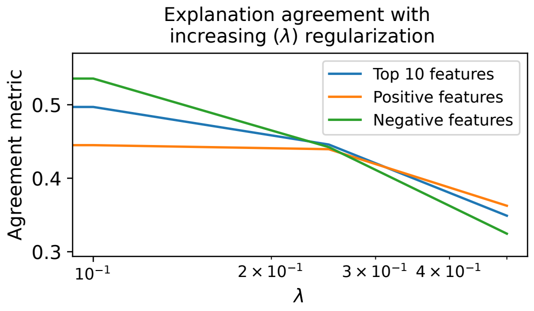
TOPKNEG is just TOPKPOS with negative prediction features instead of positive. To measure explanation agreement, we evaluate models at different λ for both REG (Figure 4a) and WEIGHTS (Figure 4b). For all three metrics, as λ increases, the explanation agreement also reduces. Although these results are 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.

Local Disagreement: Generating a Dissenting Explanation for Any Input

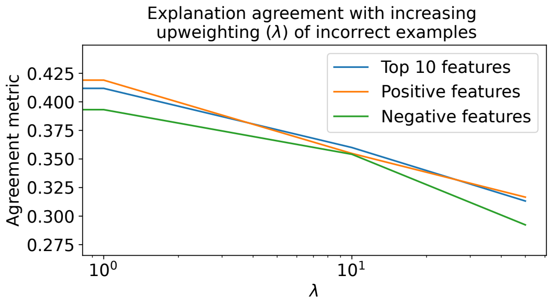
While the techniques we presented increase model disagreement on the test data only with the training data, the total

---

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



(a) Explanation Agreement from Reg



(b) Explanation agreement from Weights

Figure 4: 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.

coverage only spans $< 30\%$ of points of the test set. 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 6. ***Local Disagreement*** *Given reference model f, training data D, and a test instance x, find some g where $f(x) \neq g(x)$ where $\mathrm{Err}_{D_{\text{test}}}(f) \approx \mathrm{Err}_{D_{\text{test}}}(g)$.*

Here we know the exact test instance for which we want a different prediction. We both consider the scenario of additional access to only the training set and the scenario of additional access to only the trained model. For the former, we use a linear SVM to demonstrate the efficacy of flipping the test instance label by reducing the training size $|D|$. As the training data size decreases, the influence of the test instance increases, improving the success rate of generating a different prediction for that instance. Table 2a shows that Problem 6 can be solved with high probability (i.e. $> 90\%$) without sacrificing significant model accuracy for the deceptive reviews task. For the scenario where we only have additional access to the trained model (i.e. a neural network fitted to the training set), we retrain directly to minimize the loss on 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. This method is also effective in flipping the label for roughly 80% of the test set examples while still maintaining $\sim 88\%$ accuracy.

$\|D\|$	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 (f vs g), and test set accuracy of g when adding a test instance to the training set for a flipped prediction. As dataset size decreases, instances are more likely to be successfully predicted as the opposite class. (b) Distribution of training iterations required and resulting TopK agreement (f vs g) and accuracy of g, for a neural network model that is retrained on a single test instance. Errors reported are standard deviation for all values and variance for success rate (Bernoulli). All test set accuracy errors are < 0.02.

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 dissenting explanations which argue against a model prediction and explanation, we defined types of disagreement and presented simple but effective heuristics for eliciting such disagreements. We show that generating disagreement in predictions is sufficient for generating different explanations. Our work serves as a first step in connecting human interaction and computational challenges in treating explanations as arguments for AI-assisted decision-making.

Limitations and Future Work Our work, on a single task, cannot make claims about the effectiveness of dissenting explanations in general. The subset of examples we examine is difficult for the reference model, thus providing a single explanation does not significantly assist human decision-making; future work should use a stronger reference model with more examples. A promising direction of future work is to explore what tasks dissenting explanations best aid and other types of dissenting explanations involving counterfactual explanations.

References

Bansal, G.; Wu, T.; Zhou, J.; Fok, R.; Nushi, B.; Kamar, E.; Ribeiro, M. T.; and Weld, D. 2021. 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*, 1–16.

Brunet, M.-E.; Anderson, A.; and Zemel, R. 2022. Implications of Model Indeterminacy for Explanations of Automated Decisions. *Advances in Neural Information Processing Systems*, 35: 7810–7823.

Burkart, N.; and Huber, M. F. 2020. A Survey on the Explainability of Supervised Machine Learning. *CoRR*, abs/2011.07876.

Fisher, A.; Rudin, C.; and Dominici, F. 2019. 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.

Han, T.; Srinivas, S.; and Lakkaraju, H. 2022. Which Explanation Should I Choose? A Function Approximation Perspective to Characterizing Post hoc Explanations. *arXiv preprint arXiv:2206.01254*.

Krishna, S.; Han, T.; Gu, A.; Pombra, J.; Jabbari, S.; Wu, S.; and Lakkaraju, H. 2022. The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective. *arXiv preprint arXiv:2202.01602*.

Lai, V.; Liu, H.; and Tan, C. 2020. " Why is' Chicago'deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13.

Lai, V.; and Tan, C. 2019. 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*, 29–38.

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

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

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

Michael, J.; Mahdi, S.; Rein, D.; Petty, J.; Dirani, J.; Padmakumar, V.; and Bowman, S. R. 2023. Debate Helps Supervise Unreliable Experts. *arXiv preprint arXiv:2311.08702*.

Miller, T. 2023. Explainable ai is dead, long live explainable ai! hypothesis-driven decision support. *arXiv preprint arXiv:2302.12389*.

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

Pang, T.; Xu, K.; Du, C.; Chen, N.; and Zhu, J. 2019. Improving adversarial robustness via promoting ensemble diversity. In *International Conference on Machine Learning*, 4970–4979. PMLR.

Rame, A.; and Cord, M. 2021. Dice: Diversity in deep ensembles via conditional redundancy adversarial estimation. *arXiv preprint arXiv:2101.05544*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

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

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

Vasconcelos, H.; Jörke, M.; Grunde-McLaughlin, M.; Gerstenberg, T.; Bernstein, M.; and Krishna, R. 2022. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *arXiv preprint arXiv:2212.06823*.

Wang, X.; and Yin, M. 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th international conference on intelligent user interfaces*, 318–328.