A Causal Framework for Evaluating Deferring Systems

Filippo Palomba
Princeton University

Andrea Pugnana
University of Pisa

José M. Álvarez
Scuola Normale Superiore

Salvatore Ruggieri
University of Pisa

Abstract

Deferring systems extend supervised Machine Learning (ML) models with the possibility to defer predictions to human experts. However, evaluating the impact of a deferring strategy on system accuracy is still an overlooked area. This paper fills this gap by evaluating deferring systems through a causal lens. We link the potential outcomes framework for causal inference with deferring systems, which allows to identify the causal impact of the deferring strategy on predictive accuracy. We distinguish two scenarios. In the first one, we have access to both the human and ML model predictions for the deferred instances. Here, we can identify the individual causal effects for deferred instances and the aggregates of them. In the second one, only human predictions are available for the deferred instances. Here, we can resort to regression discontinuity design to estimate a local causal effect. We evaluate our approach on synthetic and real datasets for seven deferring systems from the literature.

1 INTRODUCTION

Learning to defer (LtD) (Madras et al., 2018) extends supervised learning by allowing Machine Learning (ML) models to defer the prediction to a human expert. The extended models – called *deferring systems* – aim at obtaining the best from the combination of AI and human expert predictions, thus reducing potentially harmful mistakes. The LtD research field is blooming, with novel approaches continuously appearing (e.g., Mozannar et al. (2023); Cao et al. (2023); Liu et al. (2024); Wei et al. (2024)). However, all these works evaluate deferring systems by looking at the final accuracy obtained by the human-AI team. This accuracy-based view on evaluation is limited as it does not consider the *causal effect* of introducing the deferring strategy on the system's predictive performance. When evaluating the impact of a new drug or a new policy, e.g., one has not only to observe a positive change in the outcome of interest, but also has to ensure the change in that outcome was due to the performed intervention (Nogueira et al., 2022). Similarly, due to mounting regulatory pressure, policymakers are interested in understanding the causal effect of introducing a deferring system, in particular, within a high-stake decision-making process (Álvarez et al., 2024).

Consider the following two examples: *(Ex1)* an online platform introduces a new deferring system to moderate its content for hate speech, meaning most content moderation is still automated, but a small part is now handled by humans; *(Ex2)* a hospital introduces a new deferring system for diagnosing a disease, meaning that medical doctors will still handle part of the patients, while an ML model will diagnose the remaining cases. After some months, the stakeholders of the online platform in *Ex1* may ask the developers of the deferring system to quantify the causal effects of *deferring to humans* instead of automatic content moderation. Similarly, the stakeholders of the hospital in *Ex2* may ask for the causal effects of *deferring to the ML model* instead of full human decision-making. Both examples require a causal inference approach because the goal is to estimate the causal effect of a variable (the introduction of a deferring system) on another one (the predictive performance) (Pearl, 2009).

In this work, we link deferring systems with the causal inference framework of *potential outcomes* (Rubin, 1974) by mapping concepts from the former to the latter. We distinguish two scenarios. In the first one, we can access the ML model predictions for both deferred and non-deferred instances, and the human predictions only for deferred ones. *Ex1* belongs to such a scenario. In this context, various causal quantities *of deferring to humans* can be readily identified and estimated. In the second scenario described by *Ex2*, we can access

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

the ML model predictions only for the non-deferred instances and the human predictions only for the deferred ones. In this context, we rely on *Regression Discontinuity* (RD) design (Thistlethwaite and Campbell, 1960) to identify and estimate a local causal effect, where local refers to the boundary of the deferring decision. Such a local causal effect covers both the causal effect *of deferring to humans* and the one *of deferring to the ML model*, as they are one the opposite of the other. *Ex2* belongs to such a scenario.

Our contributions are threefold: (*i*) we frame the evaluation of deferring systems as a causal inference problem using the potential outcomes framework; (*ii*) we investigate two scenarios and for each show which causal effects can be identified and estimated; and (*iii*) we evaluate the proposed approach on five datasets – a synthetic one and four real-world ones – using seven deferring systems from the literature. We introduce causal inference and deferring systems in Section 2. We bridge the two frameworks and investigate the two scenarios of causal estimation in Section 3. We report experiments in Section 4. We conclude in Section 5.

2 BACKGROUND

2.1 Causal Inference

The core task of causal inference is to estimate the *causal effect* of a binary *treatment* random variable $D \in \{0,1\}$ on another discrete or continuous *outcome* random variable $O \in \mathcal{O}$. Let us consider a random sample $\{D_i, O_i\}_{i=1}^n$ of i.i.d. variables, where the subscript i denotes a specific instance/unit i. We denote realizations of such random variables with lowercase letters. A formal definition of a causal effect is given by the Neyman-Rubin causal framework (Neyman, 1923; Rubin, 1974) through the notion of *potential outcomes*. A potential outcome $O(d) \in \mathcal{O}, d \in \{0,1\}$ is a random variable representing the value that the outcome variable O would take when the treatment variable is set to d. Accordingly, the (individual) causal effect of D on O for unit i is defined as $\tau_i = O_i(1) - O_i(0)$.[1]

If we were able to observe the joint distribution of $(O(0), O(1))$, then the causal effect of each unit could be readily estimated from a dataset of observations. However, for each unit i, only one among $O_i(1)$ and $O_i(0)$ can be typically observed. This is called the "fundamental problem of causal inference" (Holland, 1986). It occurs since the observed outcome O_i and the potential outcomes are related by $O_i = D_i \cdot O_i(1) + (1 - D_i) \cdot O_i(0)$. In other words, if a unit i is assigned to the treatment ($D_i = 1$), then the potential outcome $O_i(0)$ is counterfactual in nature, and we would not observe it. Symmetrically, $O_i(1)$ is counterfactual for units not assigned to treatment ($D_i = 0$). For this reason, researchers are often interested in less granular causal quantities, such as:

$$\tau_{\texttt{ATE}} := \mathbb{E}[O(1) - O(0)], \qquad \text{and} \\ \tau_{\texttt{ATT}} := \mathbb{E}[O(1) - O(0) \mid D = 1], \tag{1}$$

known as the *average treatment effect* (ATE) and the *average treatment effect on the treated* (ATT), respectively.[2] Despite being more general than the individual causal effect, the causal estimands in (1) cannot be estimated from a dataset of observations unless some assumptions are imposed, as the distribution of $(D, O(0), O(1))$ is (*i*) unknown and (*ii*) generally impossible to learn from the data because of the fundamental problem of causal inference. Several methodologies have been proposed to use context-dependent knowledge to model $(D, O(0), O(1))$. See Abadie and Cattaneo (2018) for a recent review.

The RD design (Thistlethwaite and Campbell, 1960) is one of such methodologies. In the canonical RD design, units are assigned a score $V \in \mathbb{R}$, known as *running variable*, and ranked according to it. A unit i whose running variable V_i is greater or equal than a *cutoff* value ξ is assigned to treatment, otherwise it does not receive the treatment. It follows that the treatment assignment is known, deterministic, and can be described by $D_i = \mathbb{1}\{V_i \geq \xi\}$. This knowledge of the assignment process can be exploited to identify and estimate causal effects. Indeed, if we can assume that units in the vicinity of the cutoff are similar, then the RD design can be used to identify:

$$\tau_{\texttt{RD}} := \mathbb{E}[O(1) - O(0) \mid V = \xi].$$

This quantity can be interpreted as a version of $\tau_{\texttt{ATT}}$ "local" at the cutoff. The above heuristics was formalized by Hahn et al. (2001) in terms of potential outcomes through the following assumption.

Assumption 1 (RD-continuity)**.** *The expected potential outcomes are continuous at the cutoff, namely, there exist:*

$$\lim_{v \to \xi} \mathbb{E}[O(0) \mid V = v] \qquad \textit{and} \qquad \lim_{v \to \xi} \mathbb{E}[O(1) \mid V = v].$$

Assumption 1 requires the average potential outcomes not to change abruptly in a small neighbourhood around the cutoff; hence their left and right limits exist and are equal. Under Assumption 1, $\tau_{\texttt{RD}}$ can be identified from the data, as the next theorem shows.

[1] We make the *stable unit treatment value assumption* (SUTVA) (Rubin, 1978). It requires each unit's potential outcome not to depend on the treatment assignment of other units, ruling out interference between units.

[2] The subscripts of O_i and D_i can be omitted in the expectation since variables are *i.i.d.*

Theorem 1 (Theorem 3 from Hahn et al. (2001)). *Let Assumption 1 hold. Then:*

$$\lim_{v\to\xi_+}\mathbb{E}[O\mid V=v]-\lim_{v\to\xi_-}\mathbb{E}[O\mid V=v]=\tau_{\texttt{RD}}.$$

Theorem 1 is "local" in nature, as it shows that the average treatment effect on the treated can be identified for a specific sub-population of units, namely those with $V=\xi$.

2.2 Deferring Systems

Let $\mathcal{X}\subseteq\mathbb{R}^q$ be a q-dimensional input space, $\mathcal{Y}=\{1,\dots,m\}$ be the target space and $P(\mathbf{X},Y)$ be the probability distribution over $\mathcal{X}\times\mathcal{Y}$. Given a hypothesis space $\mathcal{F}$ of functions that map $\mathcal{X}$ to $\mathcal{Y}$, the goal of supervised learning is to find the hypothesis $f\in\mathcal{F}$ that minimizes the *risk*:

$$R(f)=\mathbb{E}[l(f(\mathbf{X}),Y)]=\int_{\mathcal{X}\times\mathcal{Y}}l(f(\mathbf{x}),y)d\mathcal{P}(\mathbf{x},y),$$

where $l:\mathcal{Y}\times\mathcal{Y}\to\mathbb{R}$ is a user-specified loss function and $\mathcal{P}$ is the probability measure linked to the joint distribution P over the space $\mathcal{X}\times\mathcal{Y}$. Here, f represents a ML model (i.e., a predictor). Because $P(\mathbf{X},Y)$ is generally unknown, it is typically assumed that we have access to a set of realizations, called a *training set*, of an *i.i.d.* random sample over $P(\mathbf{X},Y)$. The training set is used to learn a predictor $\hat{f}$, such that $\hat{f}\in\arg\min_{f\in\mathcal{F}}\widehat{R}(f)$, with $\widehat{R}(f)$ denoting the empirical counterpart of the risk $R(f)$ over the training set.

Since the predictor $\hat{f}$ can make mistakes, we can extend the above canonical setting by allowing the ML model to defer difficult cases to another predictor. We consider a human expert as another predictor $h:\mathcal{Z}\to\mathcal{Y}$, where $\mathcal{Z}$ is possibly a higher dimensional space than $\mathcal{X}$. To keep the notation simple, we consider the case where $\mathcal{Z}=\mathcal{X}$. The mechanism that determines who provides the prediction is called the *policy function* (or rejector/deferring strategy) and can be formally defined as a (binary) mapping $g:\mathcal{X}\to\{0,1\}$. We define the *deferring system* ϑ, also called the *the human-AI team*, as a triplet (f,g,h), such that:

$$\vartheta(\mathbf{x})=(f,g,h)(\mathbf{x})=\begin{cases}f(\mathbf{x}) & \text{if} \quad g(\mathbf{x})=0\\ h(\mathbf{x}) & \text{if} \quad g(\mathbf{x})=1\end{cases}$$

meaning, if $g(\mathbf{x})=0$, the prediction is provided by the ML model, while if $g(\mathbf{x})=1$, the human expert takes care of the prediction. We assume a *single* human expert to defer the prediction to, thus excluding generalizations that pick which expert to defer to (Verma et al., 2023; Mao et al., 2023). Let $\mathcal{G}$ be the set of all the policy functions and $\mathcal{L}(f,g)$ the expected risk of the whole deferring system, namely:

$$\mathcal{L}(f,g)=\int_{\mathcal{X}\times\mathcal{Y}}l_{ML}\left(f(\mathbf{x}),y\right)\left(1-g(\mathbf{x})\right)d\mathcal{P}(\mathbf{x},y)+\int_{\mathcal{X}\times\mathcal{Y}}l_H(h(\mathbf{x}),y)g(\mathbf{x})d\mathcal{P}(\mathbf{x},y), \tag{2}$$

with l_{ML} (resp., l_H) referring to the loss associated with the ML model (resp., human expert). The goal becomes finding the best $f\in\mathcal{F}$ and $g\in\mathcal{G}$ such that:

$$\min_{f\in\mathcal{F},g\in\mathcal{G}}\mathcal{L}(f,g)\quad\text{s.t.}\quad\mathbb{E}[g(\mathbf{X})]\leq 1-c,$$

where $c\in[0,1]$ is a *target coverage*, i.e., a user-specified minimum fraction of instances for which the ML model is selected to make predictions.

Most methods design the deferring strategy through a *reject score* function $k:\mathcal{X}\to\mathbb{R}$, which estimates whether the human expert prediction is more likely to be correct than the one of the ML model (Mozannar et al., 2023). High values of $k(\mathbf{x})$ correspond to cases where the human expert is preferable, i.e., it is more likely to provide a correct prediction. Hence, we can set a threshold $\overline{\kappa}$ over $k(\mathbf{x})$ to define the policy function as $g(\mathbf{x})=\mathbb{1}\{k(\mathbf{x})\geq\overline{\kappa}\}$. Okati et al. (2021) show that such a thresholding strategy is optimal. In practice, one can estimate such a threshold in various ways. For instance, if there are no coverage constraints, a linear search procedure can be run by selecting the $\overline{\kappa}$ that maximizes accuracy over a validation set (Mozannar et al., 2023). Otherwise, one can consider a *coverage-calibration* procedure by setting $\overline{\kappa}$ as the c^{th}-percentile of the reject score values over a validation set (Pugnana et al., 2024), as shown in Figure 1. To highlight the relationship between $\overline{\kappa}$ and c, we denote the estimated threshold for a target coverage c as $\overline{\kappa}_c$, i.e., $\overline{\kappa}_c$ is such that $\mathbb{E}[\mathbb{1}\{k(\mathbf{X})\geq\overline{\kappa}_c\}]=(1-c)$.

2.3 Related Work

By viewing the introduction of human-AI teams as an intervention on a ML-based or on a human-based decision flow, our work bridges causal inference for policy evaluation with deferring systems. To the best of our knowledge, the only related work is Choe et al. (2023), which estimates the accuracy of abstaining classifiers on the abstained instances under the assumption that the abstention policy is stochastic. Such an assumption is impractical in the context of deferring systems. Further, abstaining classifiers do not account for deferring to humans. Our work addresses both these shortcomings. See Appendix A for additional related work.

Policy Evaluation. Causal inference methods are used to evaluate the effects of treatments/policies, and

Table 1: Deferring systems under the Potential Outcomes lens.

	POTENTIAL OUTCOMES	LtD
running variable	V_i	K_i
cutoff	ξ	$\overline{\kappa}_c$
treatment	D_i	G_i
outcome	O_i	T_i
potential outcomes	$O_i(d), d \in \{0,1\}$	$T_i(g), g \in \{0,1\}$
τ_i	$O_i(1) - O_i(0)$	$T_i(1) - T_i(0)$

inform policymakers (Abadie and Cattaneo, 2018). Randomized control trials (RCT), i.e., experiments that assign units to treatment randomly, are the gold standard for inferring average causal effects (such as $\tau_{\texttt{ATE}}$) in many fields including healthcare (Stolberg et al., 2004), education (Carlana et al., 2022), and finance (Banerjee et al., 2015). When it is not possible to rely on RCTs, one can resort to other techniques using observational data (Angrist and Pischke, 2009). The RD design has been used to assess the effectiveness of treatments in several fields, such as healthcare (Cattaneo et al., 2023), criminal behavior (Pinotti, 2017), education (Duflo et al., 2011), public economics (Coviello and Mariniello, 2014), and corporate finance (Flammer, 2015).

Deferring Systems Applications. In recent years, deferring systems have been deployed in several domains. For instance, Van der Plas et al. (2023a) present a deferring system for sleep stage scoring, which can be used to allow physicians to focus on critical patients. Cianci et al. (2023) adopt selective classification (El-Yaniv and Wiener, 2010) for uncertainty self-assessment of a credit scoring ML model, with the purpose of informing the human decision maker. Bondi et al. (2022) study a deferring system for evaluating the presence of animals in photo traps, showing that the performance of the deferring system is influenced by how the deferral choice is communicated to humans. We refer to Punzi et al. (2024) for a recent survey on hybrid decision-making.

3 EVALUATING DEFERRING SYSTEMS

Problem Statement. We consider the problem of measuring the causal contribution, in terms of predictive performance, of deferring the prediction to a human expert in a given deferring system based on a reject score function. We assume a given set of realizations $\mathcal{D}_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, called a *test set*, of an *i.i.d.* random sample $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ over $P(\mathbf{X}, Y)$, and we define the test sample accuracy of a generic predictor m as $\widehat{\texttt{Acc}}_m := \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{m(\mathbf{x}_i) = y_i\}$.

Methodology. To tackle the problem above, we bridge deferring systems with the potential outcomes framework. The key observation is that the reject score maps to the running variable, and the predictive performance of the ML model and of the human expert map to the potential outcomes.

Table 1 links each variable of a deferring system to the potential outcomes framework. We have that: (i) *the reject score* $K_i = k(\mathbf{X}_i)$ is the running variable; (ii) the *threshold* $\overline{\kappa}_c \in \mathcal{K}$ is the cutoff; (iii) the *policy function* $G_i = \mathbb{1}\{K_i \geq \overline{\kappa}_c\}$ is the treatment assignment: if $G_i = 1$, the human expert provides the prediction $h(\mathbf{X}_i)$, otherwise the ML model provides the prediction $f(\mathbf{X}_i)$; (iv) *the outcome* is $T_i = u(\vartheta(\mathbf{X}_i, Y_i))$, where $u : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a function taking as inputs the deferring system prediction $\vartheta(\mathbf{X}_i)$ and its target variable Y_i, e.g., if $u(\vartheta(\mathbf{X}_i), Y_i) = \mathbb{1}\{\vartheta(\mathbf{X}_i) = Y_i\}$, the outcome is the correctness of the prediction; (v) $T_i(0) = u\left(f(\mathbf{X}_i), Y_i\right)$ and $T_i(1) = u\left(h(\mathbf{X}_i), Y_i\right)$ are *the potential outcomes*, e.g., if $u(f(\mathbf{X}_i, Y_i)) = \mathbb{1}\{f(\mathbf{X}_i) = Y_i\}$ and if $u(h(\mathbf{X}_i, Y_i)) = \mathbb{1}\{h(\mathbf{X}_i) = Y_i\}$, then the potential outcome are the correctness of the ML model prediction and of the human prediction, respectively; finally, (vi) *the individual causal effect* τ_i is the difference between $T_i(1)$ and $T_i(0)$. Notice that the outcome and the potential outcomes satisfy the expected condition $T_i = G_i \cdot T_i(1) + (1 - G_i) \cdot T_i(0)$.

We distinguish two scenarios that allow for the identification of causal effects under Table 1. In both scenarios, we assume that the human predictions $h(\mathbf{X}_i)$ can be accessed only for the deferred instances, i.e., if $G_i = 1$. In the rest of the paper, we consider $u\left(\vartheta(\mathbf{X}_i), Y_i\right) = \mathbb{1}\{\vartheta(\mathbf{X}_i) = Y_i\}$ for a direct interpretation of our results in terms of accuracy, but these results can be directly extended to other functions u.

Scenario 1. *The ML model predictions $f(\mathbf{X}_i)$ can be accessed for* ***the whole random sample****.*

This scenario covers cases in which the ML model can be called without any cost or side effects. E.g., when the development team runs an internal evaluation of the deferring system. Here, we can identify the causal effects *of deferring to the human* (Section 3.1). *Ex1* from Section 1 falls under this scenario.

Scenario 2. *The ML model predictions $f(\mathbf{X}_i)$ can be accessed* ***only for the non-deferred instances****, i.e., if $G_i = 0$.*

This scenario covers cases in which model invocation is costly, may have side effects, or discloses sensitive data. E.g., when in the owner of the deferring system is reluctant to share the deferred ML predictions during an external audit.[3] Here, we can identify the

[3] Since these predictions are not the system's actual output, releasing them might not be legally binding.

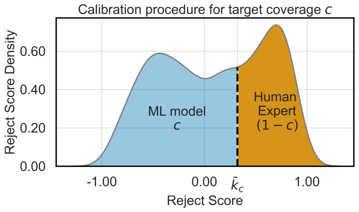


Figure 1: In blue, the (c)% of instances assigned to the ML model; in orange, the $(1-c)$% instances assigned to the the human.

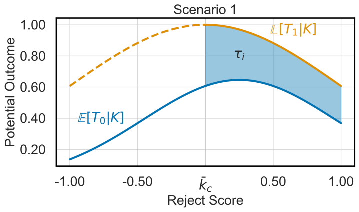


Figure 2: Scenario 1 assumptions: thick (dashed) lines are observed (unobserved) values. The coloured area represents where the effects can be estimated.

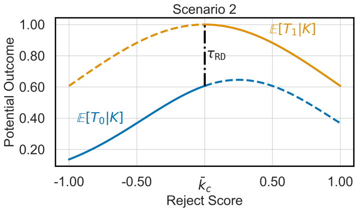


Figure 3: Scenario 2 assumptions: thick (dashed) lines are observed (unobserved) values. We can estimate $\tau_{\texttt{RD}}$ at the cutoff value.

causal effects *of deferring to the human* locally to the deferring threshold (Section 3.2). Moreover, this scenario also covers cases in which the intervention to be evaluated is the introduction of the ML model within a human decision-making process, as in *Ex2* from Section 1. Importantly, in *Ex2*, we are not interested in the causal effect of deferring from the ML model to the human expert, but rather *of deferring from the human expert to the ML model*. Therefore, Scenario 1 does not apply if we reverse the role of the ML model and the human expert because we cannot assume having the human expert's predictions for the cases assigned to the ML model. However, since the local causal effect of deferring from the human expert to the ML model is the opposite of the one of deferring from the ML model to the human expert, we can rely on Scenario 2 for estimating the effect.

3.1 Scenario 1: deferring systems as an almost perfect causal inference design

In this scenario, we are in *the ideal situation* in which both potential outcomes are observed for the deferred instances ($G_i = 1$), as both the ML model $f(\mathbf{X}_i)$ and the human $h(\mathbf{X}_i)$ predictions are available. Recalling that n is the size of the test set $\mathcal{D}_n$, the following holds.

Proposition 1. *Let Scenario 1 hold. Then, for each* $i \in [1, n]$ *such that* $G_i = 1$*:*

$$\tau_i = T_i(1) - T_i(0) = \mathbb{1}\{h(\mathbf{X}_i) = Y_i\} - \mathbb{1}\{f(\mathbf{X}_i) = Y_i\}.$$

All the proofs are in Appendix B. Because we can compute the most granular causal effect on deferred instances, τ_i, we can also retrieve less granular quantities (see Figure 2). For instance, if we average the τ_i over the population of deferred units, we obtain the *average treatment effect on the deferred*, $\tau_{\texttt{ATD}}$, which is the deferring systems' equivalent of $\tau_{\texttt{ATT}}$ in (1). We identify this causal estimand as follows.

Proposition 2. *Let Scenario 1 hold. Then:*

$$\tau_{\texttt{ATD}} = \mathbb{E}[T(1) - T(0)|G = 1] = \mathbb{E}\left[\mathbb{1}\{h(\mathbf{X}) = Y\} \mid G = 1\right] - \mathbb{E}\left[\mathbb{1}\{f(\mathbf{X}) = Y\} \mid G = 1\right].$$

$\tau_{\texttt{ATD}}$ allows to measure the (average) effect on accuracy due to the "intervention" of deferring to a human the prediction for the deferred instances. Intuitively, $\tau_{\texttt{ATD}}$ estimates what would be the average increase in accuracy for deferred instances if the human predicts instead of the ML model. Hence, $\tau_{\texttt{ATD}}$ motivates introducing a deferring system to stakeholders. Policymakers can use it to assess the impact of such systems. We can estimate $\tau_{\texttt{ATD}}$ through a difference-in-means estimator of the form:

$$\hat{\tau}_{\texttt{ATD}} = \frac{1}{n_1} \sum_{i \in [1,n]: g(\mathbf{x}_i)=1} [t_i(1) - t_i(0)] = \frac{1}{n_1} \sum_{i \in [1,n]: g(\mathbf{x}_i)=1} \left[\mathbb{1}\{h(\mathbf{x}_i) = y_i\} - \mathbb{1}\{f(\mathbf{x}_i) = y_i\}\right],$$

where $n_1 := |\{i \in [1, n] : g(\mathbf{x}_i) = 1\}|$ is the number of deferred instances in the test set.

The next proposition highlights that (i) $\hat{\tau}_{\texttt{ATD}}$ identifies a causal quantity of interest, and (ii) $\hat{\tau}_{\texttt{ATD}}$ differs from the simple difference, typically used in the literature, between the accuracy of the deferring system and the one of the ML model $\hat{\tau}_{\Delta} := \widehat{\texttt{Acc}}_{\vartheta} - \widehat{\texttt{Acc}}_{f}$.

Proposition 3. *Let Scenario 1 hold. Then:*
(i) as $n \to \infty$, $\hat{\tau}_{\texttt{ATD}} \xrightarrow{p} \mathbb{E}[T(1) - T(0) \mid G = 1] = \tau_{\texttt{ATD}}$
(ii) for each $n > 0$, $\hat{\tau}_{\Delta} = \frac{n_1}{n} \hat{\tau}_{\texttt{ATD}}$.

A consequence of Proposition 3 is that $\hat{\tau}_{\Delta}$ is an inconsistent estimator for the causal effect $\tau_{\texttt{ATD}}$ unless all units are deferred. Another consequence is that we can re-weight $\hat{\tau}_{\Delta}$ by n/n_1 to obtain a consistent estimator. We discuss this further in Appendix C.

To conclude, we highlight that under Scenario 1, any aggregated metrics of the individual causal effects on the deferred can be estimated. For instance, we could estimate the *conditional average treatment effects on the deferred* $\tau_{\texttt{CATD}}(\tilde{\mathbf{x}}) = \mathbb{E}[T(1) - T(0)|G = 1, \mathbf{X} = \tilde{\mathbf{x}}]$ by further conditioning on a specific set of features $\mathbf{X} = \tilde{\mathbf{x}}$. Similarly to $\tau_{\texttt{ATD}}$, $\tau_{\texttt{CATD}}$ can be estimated by considering the difference in means estimator $\hat{\tau}_{\texttt{CATD}}$:

$$\hat{\tau}_{\texttt{CATD}}(\tilde{\mathbf{x}}) = \frac{1}{n_{1,\tilde{x}}} \sum_{i \in [1,n]: g(\mathbf{x}_i)=1, \mathbf{x}_i = \tilde{\mathbf{x}}} \left[t_i(1) - t_i(0) \right],$$

where $n_{1,\tilde{x}} := |\{i \in [1,n] : (g(\mathbf{x}_i) = 1) \wedge (\mathbf{x}_i = \tilde{\mathbf{x}})\}|$.

3.2 Scenario 2: deferring systems as an RD design

If we cannot access the ML model predictions for the deferred instances, we can exploit the fact that *deferring systems can be interpreted as an RD design* to compute a local version of $\tau_{\texttt{ATD}}$. In this scenario, $\tau_{\texttt{RD}}$ answers to the question: for a fixed coverage value c and the corresponding reject score threshold $\overline{\kappa}_c$, what would be the increase in accuracy if we let the human expert predict instead of the ML model for the instances with reject score close to $\overline{\kappa}_c$? Such an interpretation is possible if Assumption 1 holds. If we have reasons to believe that small changes to the reject-score threshold $\overline{\kappa}_c$ do not abruptly change the expected predictive accuracy of the human expert and of the ML model, then Assumption 1 is satisfied.

Proposition 4. *Let Scenario 2 hold and let Assumption 1 be satisfied for the deferring system. Then:*

$$\lim_{k \to \overline{\kappa}_c^+} \mathbb{E}[T \mid K = k] - \lim_{k \to \overline{\kappa}_c^-} \mathbb{E}[T \mid K = k] = \tau_{\texttt{RD}},$$

where $\tau_{\texttt{RD}} := \mathbb{E}[T(1) - T(0) \mid K = \overline{\kappa}_c]$.

Proposition 4 allows to evaluate the causal effect of deferring to a human in a decision flow even if we do not have access to ML predictions for the deferred instances. Indeed, $\tau_{\texttt{RD}}$ readily quantifies the gain in predictive accuracy of having the human expert predicting in place of the ML model at the cutoff (see Figure 3). The local nature of $\tau_{\texttt{RD}}$ motivates the use of local non-parametric polynomial kernel regression[4] to estimate $\mathbb{E}[T \mid K = \overline{\kappa}_c]$ from the left and the right of the cutoff, thus obtaining an estimator $\hat{\tau}_{\texttt{RD}}$ of $\tau_{\texttt{RD}}$. The $\tau_{\texttt{RD}}$ can also be computed under Scenario 1. In Appendix C, we discuss additional caveats, including how to set the optimal coverage, how to check if Assumption 1 holds, and the uncertainty due to the ML model estimation.

4 EXPERIMENTAL EVALUATION

In this section, we address three questions:

Q1: *For Scenario 1, what is the causal effect on predictive accuracy of introducing a deferring system?*

Q2: *What other causal effects can be computed under Scenario 1?*

Q3: *For Scenario 2, what is the causal effect on predictive accuracy of introducing a deferring system?*

We consider both synthetic and real data, detailing all results in Appendix D.2. The experimental software is available at `https://anonymous.4open.science/r/PODS-565D`. We report hardware specifications and carbon footprint in Appendix D.1.4.

4.1 Experimental settings

Data. We generate synthetic data using the procedure from Mozannar et al. (2023). Such a procedure generates samples containing (i) instances for which the human expert performs better than the ML model, and (ii) instances for which the ML model is better than the human expert. Regarding the real data, we consider four datasets used in the LtD literature: `cifar10h` (Battleday et al., 2020), a hard-labelled version of `galaxyzoo` (Astro-Dave et al., 2013), `hatespeech` (Davidson et al., 2017), and `xray-airspace` (Wang et al., 2017; Majkowska et al., 2020). We provide data characteristics and applied pre-processing in Appendix D.1.1.

Baselines. We consider several deferring systems, including: *Selective Prediction* (SP) (Geifman and El-Yaniv, 2017), *Compare Confidence* (CC) (Raghu et al., 2019), *Differentiable Triage* (DT) (Okati et al., 2021), *Cross-Entropy Surrogate* (LCE) (Mozannar and Sontag, 2020), *One Vs All* (OVA) (Verma and Nalisnick, 2022), *Realizable Surrogate* (RS) (Mozannar et al., 2023) and *Asymmetric SoftMax* (ASM) (Cao et al., 2023). We provide details for the baselines and the hyper-parameter choice in Appendices D.1.2, D.1.3.

General setup. For all the experiments, we consider the following steps: (i) we randomly split the dataset in training, validation, and test set, according to a $70\%, 10\%, 20\%$ proportion; (ii) we train the deferring system on the training set; (iii) we estimate different cutoff values $\overline{\kappa}_c$ over the validation set, considering each of the following target coverages $c \in \{.10, .20, .30, .40, .50, .60, .70, .80, .90\}$; (iv) for each cutoff value $\overline{\kappa}_c$, we estimate on the test set the deferring system accuracy as well as $\tau_{\texttt{ATD}}$ and $\tau_{\texttt{RD}}$. We consider a single training, validation, and test split since our goal is estimating the causal effect of implementing a deferring system, not estimating its predic-

[4] Local polynomial kernel regressions (Fan and Gijbels, 1996) fit a p-th order polynomial locally at k via weighted least squares, where the weight of each instance is determined by the shape of the kernel and is non-increasing in the distance between k and $k(\mathbf{x}_i)$.

tive accuracy. The estimated $\hat{\tau}_{\mathtt{ATD}}$ and $\hat{\tau}_{\mathtt{CATD}}$ are computed through a difference in means estimator and $\hat{\tau}_{\mathtt{RD}}$ is obtained using the default implementation provided in the `rdrobust` package (Calonico et al., 2017), i.e., a local linear kernel regression with optimal bandwidth (Calonico et al., 2020). To assess the statistical significance of the results, we report the 95% confidence intervals and the corresponding p-values (pv) associated with $\hat{\tau}_{\mathtt{ATD}}$ and $\hat{\tau}_{\mathtt{RD}}$ when testing the null hypotheses[5] of $\tau_{\mathtt{ATD}} = 0$ and $\tau_{\mathtt{RD}} = 0$, respectively.

4.2 Experimental results

Figure 4 shows the experimental results. For each dataset, we analyze the best deferring system in terms of accuracy, as shown in the first row of Figure 4a. We provide the experimental results for the other baselines in Appendix D.2.

Q1: causal effects under Scenario 1.

The second row of Figure 4a shows $\hat{\tau}_{\mathtt{ATD}}$'s and their confidence intervals when varying the cutoff $\overline{\kappa}_c$. The black horizontal line denotes the null effect.

For `synth`, the plot confirms the effectiveness of ASM, with an increasing trend in the estimated causal effects, ranging from $\approx .095$ ($pv \approx$ 6.98e$-$26) at zero coverage to $\approx .417$ ($pv \approx$ 7.32e$-$58) at $c = .90$. We see similar patterns for real data: regarding `cifar10h`, the $\hat{\tau}_{\mathtt{ATD}}$'s are positive and increasing with $\overline{\kappa}_c$: the values range from $\approx .019$ at $c = 0$ to $\approx .265$ ($pv \approx$ 2.83e$-$10) at $c = .90$. Also, for `hatespeech`, the causal effects monotonically increase with the coverage, with $\hat{\tau}_{\mathtt{ATD}}$ ranging from $\approx .019$ ($pv \approx$ 9.54e$-$18) at $c = 0$ up to $\approx .295$ ($pv \approx$ 7.40e$-$28) at $c = .90$. We observe an overall positive effect for the `xray-airspace` dataset as well, with the highest $\hat{\tau}_{\mathtt{ATD}}$ achieved at $c = .80$ ($\approx .209$, $pv \approx$ 4.29e$-$9). All the estimates for those datasets significantly differ from zero, supporting the effectiveness of deferring. The `galaxyzoo` dataset is an exception, with $\hat{\tau}_{\mathtt{ATD}}$ taking negative values for the coverages below $c = .70$. Interestingly, despite an increase in the system accuracy for $c = .80$ and $c = .90$, the $\hat{\tau}_{\mathtt{ATD}}$ are not statistically different from zero, with the $\hat{\tau}_{\mathtt{ATD}}$ of $\approx .068$ ($pv \approx$ 4.69e$-$2) and $\approx .096$ ($pv \approx$ 3.56e$-$2) respectively). Hence, introducing a deferring system would not improve over a fully automated setting.

We point out that this causal effect cannot be quantified by examining the accuracy of the deferring system (see Proposition 3[6]), as normally done by previous works. In fact, the shape of the top and bottom plots in Figure 4a clearly differ.

To conclude, Figure 4c) compares the estimated $\hat{\tau}_{\mathtt{ATD}}$'s for a fixed target coverage ($c = .90$) over multiple deferring systems.

Q2: conditional causal effects under Scenario 1. When designing a deferring system, we aim at not introducing new forms of bias or unfairness w.r.t. protected-by-law groups (Ruggieri et al., 2023). Since under Scenario 1, we can estimate the individual *causal* effects for the deferred units, we can also study the causal effect on predictive accuracy of deferring for instances belonging to different social groups. In particular, if $\tau_{\mathtt{CATD}}$ is negative for the deferred instances belonging to a protected group, we can conclude that introducing the deferring system has negatively affected the predictive performance for such instances. Such an effect is causal, i.e., it can be attributed to the adoption of the deferring strategy.

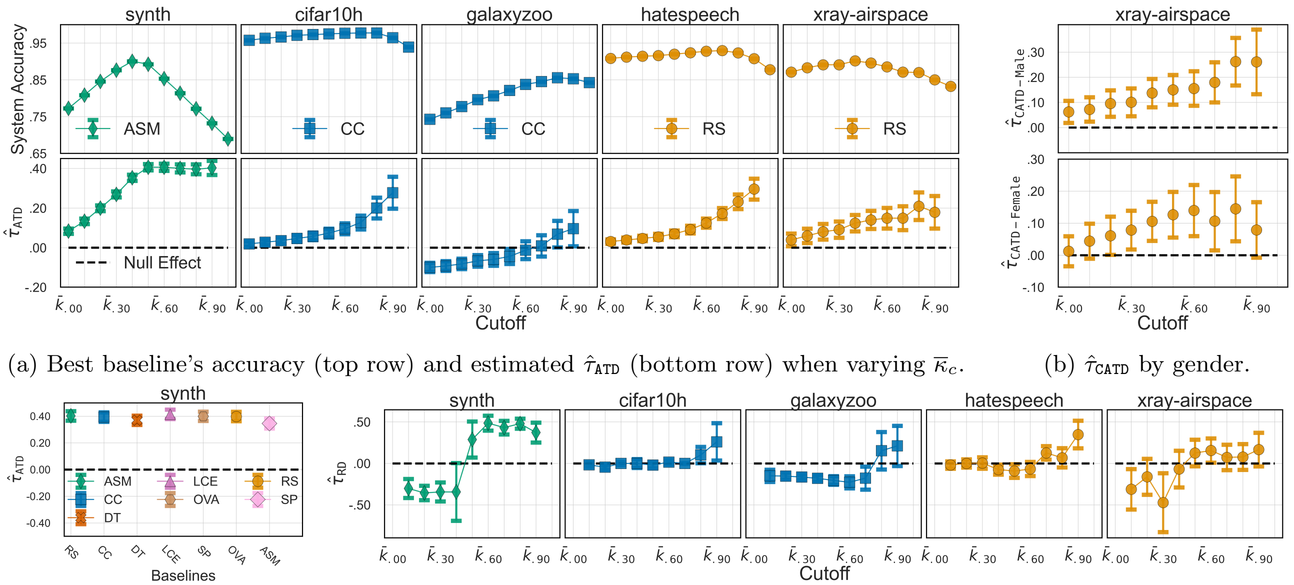
In Figure 4b, we report such causal effects for the `xray-airspace` dataset. We consider gender as the protected attribute and compute the heterogeneous causal effect $\hat{\tau}_{\mathtt{CATD}}$ for male and female patients. As shown in the top plot, when looking at male patients, the estimated $\hat{\tau}_{\mathtt{CATD}}$ grows with the cutoff, and statistically significant effects can be observed for $c > .30$, reaching a maximum of $\approx .262$ ($pv -$ 5.83e$-$8) at $c = .80$. On the other hand, when looking at female patients (bottom plot), we see smaller positive effects of introducing a deferring system, and such effects are not statistically different from zero. This implies that introducing a deferring system benefits male patients but not female patients. We plan to further explore this fairness angle in our proposed causal evaluation framework in future work, since causal claims are essential e.g., from a legal perspective Bathaee (2018).

Q3: causal effects under Scenario 2. Under Scenario 2, we expect the local causal effects $\tau_{\mathtt{RD}}$ to be negative for instances where the ML model locally performs better than the human; positive for instances where the human locally outperforms the ML model; and not significant for those where the ML model and the human locally perform equally. In practice, we should see the estimated $\hat{\tau}_{\mathtt{RD}}$ close to zero for the cutoffs that maximize the overall deferring system accuracy.

Figure 4d reports the estimated $\hat{\tau}_{\mathtt{RD}}$'s. For `synth`, we see significant negative effects for $c \leq .30$, with the lowest $\hat{\tau}_{\mathtt{RD}}$ of $\approx -.356$ ($pv \approx$ 9.60e$-$16) at $c = .20$; non-significant effects for c between .40 and .50 and

[5] We correct for multiple testing through the Bonferroni correction for all the tested hypotheses (five datasets, ten coverages under Scenario 1 and nine under Scenario 2, seven methods): to achieve significance at a family-wise error rate of $\alpha = .05$, the p-value must be smaller than 7.52e$-$5.

[6] Notice that $\hat{\tau}_{\Delta}$ in Proposition 3 turns out to be the accuracy of Figure 4a minus a constant value, precisely the accuracy at $c = 1.0$.



(a) Best baseline's accuracy (top row) and estimated $\hat{\tau}_{\text{ATD}}$ (bottom row) when varying $\overline{\kappa}_c$.

(b) $\hat{\tau}_{\text{CATD}}$ by gender.

(c) $\hat{\tau}_{\text{ATD}}$ at $c = .90$ for `synth`.

(d) Estimated $\hat{\tau}_{\text{RD}}$ when varying cutoff $\overline{\kappa}_c$ for the best baseline.

Figure 4: Experimental results: Figure 4a reports system accuracy and $\hat{\tau}_{\text{ATD}}$ (Scenario 1); Figure 4b reports estimated $\hat{\tau}_{\text{CATD}}$ when conditioning on the gender of the patient on the `xray-airspace` dataset; Figure 4c compares $\hat{\tau}_{\text{ATD}}$ over multiple baselines on `synth`; Figure 4d reports $\hat{\tau}_{\text{RD}}$ (Scenario 2).

significant positive effects for $c \geq .60$, with a $\hat{\tau}_{\text{RD}}$ peaking to $\approx .485$ ($pv3.07\text{e}{-26}$) at $c = .60$. This aligns with the expected optimal behaviour, as the system accuracy is maximal between $c = .40$ and $c = .50$.

When looking at real datasets, we see many statistically non-significant effects (see Tables 3–7 in the Appendix). This might be due to two factors: (*i*) the test size is small, limiting the statistical power of the local polynomial regressions; (*ii*) there are small differences in system accuracy at the variation of the cutoff (see, e.g., `cifar10h` and `hatespeech` in Figure 4a), hence locally to the deferring boundary the difference between the ML model and the human predictor is expected to be small. A few exceptions can be noticed. For `galaxyzoo`, the ML model performs better than the human expert on this task, translating into a few negative and statistically significant $\hat{\tau}_{\text{RD}}$ coefficients. For `hatespeech`, we see a statistically significant effect $\approx .348$ ($pv \approx 4.1\text{e}{-5}$) for RS at $\overline{\kappa}_c = .90$. Thus, the causal effect of deferring to the human expert on instances around the cutoff $\overline{\kappa}_{.90}$ is positive, supporting the effectiveness of the deferring strategy.

Finally, when comparing Figure 4d to the top row of Figure 4a, we notice that $\hat{\tau}_{\text{RD}}$ is negative whenever the system accuracy is increasing, it is positive when the system accuracy drops, and it is near zero when the system accuracy flattens. This happens because $\hat{\tau}_{\text{RD}}$ is the opposite of the local change in accuracy when we slightly increase the cutoff and, thus, defer less.

5 CONCLUSIONS

We tackled the evaluation of the predictive performance of deferring systems from a causal perspective. Our link with the potential outcomes and with the RD design frameworks directly allows for identifying the causal effects of the deferring strategy. Experiments on synthetic and real datasets provided practical guidance on how to reason about the causal estimation problem.

Limitations and broader impact. We assumed access to the reject scores $k(\mathbf{x})$ of instances. This is not strictly required under Scenario 1, for which only the information on whether an instance is deferred or not is required. Conversely, identifying τ_{RD} under Scenario 2 requires at least knowing the ranking induced by the reject score. Moreover, Assumption 1 must hold to identify τ_{RD}. However, such an assumption is not directly verifiable and can only be falsified. We discuss this further in Appendix C and show how to falsify Assumption 1 in Appendix D.3. From a broader perspective, our evaluation framework helps to better quantify the impact of deferring systems and deploy safer ones, especially in high-risk contexts.

Future work. We plan to extend our approach (*i*) to deferring strategies that consider multiple human experts, (*ii*) to support bias and unfairness understanding and mitigation, and (*iii*) to account for the influence that deferring has on human behaviour, e.g., in strategic classification.

References

Abadie, A. and Cattaneo, M. D. (2018). Econometric Methods for Program Evaluation. *Annual Review of Economics*, 10(1):465–503.

Álvarez, J. M., Colmenarejo, A. B., Elobaid, A., Fabbrizzi, S., Fahimi, M., Ferrara, A., Ghodsi, S., Mougan, C., Papageorgiou, I., Lobo, P. R., Russo, M., Scott, K. M., State, L., Zhao, X., and Ruggieri, S. (2024). Policy advice and best practices on bias and fairness in AI. *Ethics Inf. Technol.*, 26(2):31.

Angrist, J. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics*. Princeton University Press.

AstroDave, AstroTom, Winton, C. R. ., joycenv, and Willett, K. (2013). Galaxy zoo - the galaxy challenge.

Banerjee, A., Duflo, E., Glennerster, R., and Kinnan, C. (2015). The miracle of microfinance? Evidence from a randomized evaluation. *American Economic Journal: Applied Economics*, pages 22–53.

Bathaee, Y. (2018). The Artificial Intelligence black box and the failure of intent and causation. *Harvard Journal of Law & Technology*, 31(2):889–938.

Battleday, R. M., Peterson, J. C., and Griffiths, T. L. (2020). Capturing human categorization of natural images at scale by combining deep networks and cognitive models. *Nature communications*, 11(1):5418.

Bondi, E., Koster, R., Sheahan, H., Chadwick, M. J., Bachrach, Y., Cemgil, A. T., Paquet, U., and Dvijotham, K. (2022). Role of human-AI interaction in selective prediction. In *AAAI*, pages 5286–5294. AAAI Press.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In *NeurIPS*.

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2018). On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference. *Journal of the American Statistical Association*, 113(522):767–779.

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2020). Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs. *The Econometrics Journal*, 23(2):192–210.

Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2022). Coverage error optimal confidence intervals for local polynomial regression. *Bernoulli*, 28(4).

Calonico, S., Cattaneo, M. D., Farrell, M. H., and Titiunik, R. (2017). Rdrobust: Software for regression-discontinuity designs. *The Stata Journal*, 17(2):372–404.

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014). Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs. *Econometrica*, 82(6):2295–2326.

Cao, Y., Mozannar, H., Feng, L., Wei, H., and An, B. (2023). In defense of softmax parametrization for calibrated and consistent learning to defer. In *NeurIPS*.

Carlana, M., La Ferrara, E., and Pinotti, P. (2022). Goals and gaps: Educational careers of immigrant children. *Econometrica*, 90(1):1–29.

Cattaneo, M. D., Frandsen, B. R., and Titiunik, R. (2015). Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate. *Journal of Causal Inference*, 3(1):1–24.

Cattaneo, M. D., Idrobo, N., and Titiunik, R. (2019). *A Practical Introduction to Regression Discontinuity Designs: Foundations*. Cambridge University Press, 1 edition.

Cattaneo, M. D., Jansson, M., and Ma, X. (2020). Simple Local Polynomial Density Estimators. *Journal of the American Statistical Association*, 115(531):1449–1455.

Cattaneo, M. D., Jansson, M., and Ma, X. (2022). lpdensity: Local polynomial density estimation and inference. *J. Stat. Softw.*, 101(2).

Cattaneo, M. D., Keele, L., and Titiunik, R. (2023). A guide to regression discontinuity designs in medical applications. *Statistics in Medicine*, 42(24):4484–4513.

Cattaneo, M. D., Keele, L., Titiunik, R., and Vazquez-Bare, G. (2016). Interpreting Regression Discontinuity Designs with Multiple Cutoffs. *The Journal of Politics*, 78(4):1229–1248.

Cattaneo, M. D., Keele, L., Titiunik, R., and Vazquez-Bare, G. (2021). Extrapolating Treatment Effects in Multi-Cutoff Regression Discontinuity Designs. *Journal of the American Statistical Association*, 116(536):1941–1952.

Cattaneo, M. D. and Titiunik, R. (2022). Regression Discontinuity Designs. *Annual Review of Economics*, 14(1):821–851.

Charusaie, M., Mozannar, H., Sontag, D. A., and Samadi, S. (2022). Sample efficient learning of predictors that complement humans. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 2972–3005. PMLR.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Choe, Y. J., Gangrade, A., and Ramdas, A. (2023). Counterfactually comparing abstaining classifiers. In *NeurIPS*.

Chow, C. K. (1970). On optimum recognition error and reject tradeoff. *IEEE Trans. Inf. Theory*, 16(1):41–46.

Cianci, G., Goglia, R., Guidotti, R., Kapllaj, M., Mosca, R., Pugnana, A., Ricotti, F., and Ruggieri, S. (2023). Applied data science for leasing score prediction. In *IEEE Big Data*, pages 1687–1696. IEEE.

Coenen, L., Abdullah, A. K. A., and Guns, T. (2020). Probability of default estimation, with a reject option. In *DSAA*, pages 439–448. IEEE.

Condessa, F., Bioucas-Dias, J. M., Castro, C. A., Ozolek, J. A., and Kovacevic, J. (2013). Classification with reject option using contextual information. In *ISBI*, pages 1340–1343. IEEE.

Corbière, C., Thome, N., Bar-Hen, A., Cord, M., and Pérez, P. (2019). Addressing failure prediction by learning model confidence. In *NeurIPS*, pages 2898–2909.

Cortes, C., DeSalvo, G., and Mohri, M. (2016). Boosting with abstention. In *NeurIPS*, pages 1660–1668.

Cortes, C., DeSalvo, G., and Mohri, M. (2023). Theory and algorithms for learning with rejection in binary classification. *Annals of Mathematics and Artificial Intelligence*, pages 1–39.

Courty, B., Schmidt, V., Luccioni, S., Goyal-Kamal, MarionCoutarel, Feld, B., Lecourt, J., LiamConnell, Saboni, A., Inimaz, supatomic, Léval, M., Blanche, L., Cruveiller, A., ouminasara, Zhao, F., Joshi, A., Bogroff, A., de Lavoreille, H., Laskaris, N., Abati, E., Blank, D., Wang, Z., Catovic, A., Alencon, M., Stechly, M., Bauer, C., Lucas-Otavio, JPW, and MinervaBooks (2024). mlco2/codecarbon: v2.4.1.

Coviello, D. and Mariniello, M. (2014). Publicity requirements in public procurement: Evidence from a regression discontinuity design. *Journal of Public Economics*, 109:76–100.

Davidson, T., Warmsley, D., Macy, M. W., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *ICWSM*, pages 512–515. AAAI Press.

De, A., Koley, P., Ganguly, N., and Gomez-Rodriguez, M. (2020). Regression under human assistance. In *AAAI*, pages 2611–2620. AAAI Press.

De, A., Okati, N., Zarezade, A., and Rodriguez, M. G. (2021). Classification under human assistance. In *AAAI*, pages 5905–5913. AAAI Press.

Denis, C. and Hebiri, M. (2020). Consistency of plug-in confidence sets for classification in semi-supervised learning. *J. of Nonpar. Statistics*, 32(1):42–72.

Dong, Y. and Kolesár, M. (2023). When can we ignore measurement error in the running variable? *Journal of Applied Econometrics*, 38(5):735–750.

Dong, Y. and Lewbel, A. (2015). Identifying the Effect of Changing the Policy Threshold in Regression Discontinuity Models. *Review of Economics and Statistics*, 97(5):1081–1092.

Duflo, E., Dupas, P., and Kremer, M. (2011). Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. *American Economic Review*, 101(5):1739–1774.

El-Yaniv, R. and Wiener, Y. (2010). On the foundations of noise-free selective classification. *J. Mach. Learn. Res.*, 11:1605–1641.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*. Routledge.

Feng, L., Ahmed, M. O., Hajimirsadeghi, H., and Abdi, A. H. (2023). Towards better selective classification. In *ICLR*. OpenReview.net.

Flammer, C. (2015). Does Corporate Social Responsibility Lead to Superior Financial Performance? A Regression Discontinuity Approach. *Management Science*, 61(11):2549–2568.

Franc, V., Průša, D., and Vorácek, V. (2023). Optimal strategies for reject option classifiers. *J. Mach. Learn. Res.*, 24:11:1–11:49.

Geifman, Y. and El-Yaniv, R. (2017). Selective classification for deep neural networks. In *NIPS*, pages 4878–4887.

Geifman, Y. and El-Yaniv, R. (2019). Selectivenet: A deep neural network with an integrated reject option. In *ICML*, volume 97, pages 2151–2159. PMLR.

Hahn, J., Todd, P., and Van der Klaauw, W. (2001). Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design. *Econometrica*, 69(1):201–209.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society.

Hendrickx, K., Perini, L., der Plas, D. V., Meert, W., and Davis, J. (2024). Machine learning with a reject option: a survey. *Mach. Learn.*, 113(5):3073–3110.

Herbei, R. and Wegkamp, M. H. (2006). Classification with reject option. *Can. J. Stat.*, 34(4):709—-721.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396):945–960.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *CVPR*, pages 2261–2269. IEEE Computer Society.

Huang, L., Zhang, C., and Zhang, H. (2020). Self-adaptive training: beyond empirical risk minimization. In *NeurIPS*.

Imbens, G. W. and Angrist, J. D. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica*, 62(2):467.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *ICLR*.

Kolesár, M. and Rothe, C. (2018). Inference in Regression Discontinuity Designs with a Discrete Running Variable. *American Economic Review*, 108(8):2277–2304.

Kühne, J., März, C., et al. (2021). Securing deep learning models with autoencoder based anomaly detection. In *PHM Society European Conference*, volume 6, pages 221–233.

Lee, D. S. (2008). Randomized experiments from non-random selection in U.S. House elections. *Journal of Econometrics*, 142(2):675–697.

Lee, D. S. and Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of economic literature*, 48(2):281–355.

Liang, S., Li, Y., and Srikant, R. (2018). Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*. OpenReview.net.

Liu, S., Cao, Y., Zhang, Q., Feng, L., and An, B. (2024). Mitigating underfitting in learning to defer with consistent losses. In *AISTATS*, volume 238 of *Proceedings of Machine Learning Research*, pages 4816–4824. PMLR.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *ICLR*. OpenReview.net.

Madras, D., Pitassi, T., and Zemel, R. S. (2018). Predict responsibly: Improving fairness and accuracy by learning to defer. In *NeurIPS*, pages 6150–6160.

Majkowska, A., Mittal, S., Steiner, D. F., Reicher, J. J., McKinney, S. M., Duggan, G. E., Eswaran, K., Cameron Chen, P.-H., Liu, Y., Kalidindi, S. R., et al. (2020). Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology*, 294(2):421–431.

Mao, A., Mohri, C., Mohri, M., and Zhong, Y. (2023). Two-stage learning to defer with multiple experts. In *NeurIPS*.

Mozannar, H., Lang, H., Wei, D., Sattigeri, P., Das, S., and Sontag, D. A. (2023). Who should predict? exact algorithms for learning to defer to humans. In *AISTATS*, volume 206, pages 10520–10545. PMLR.

Mozannar, H. and Sontag, D. A. (2020). Consistent estimators for learning to defer to an expert. In *ICML*, volume 119, pages 7076–7087. PMLR.

Nalisnick, E. T., Matsukawa, A., Teh, Y. W., Görür, D., and Lakshminarayanan, B. (2019). Hybrid models with deep and invertible features. In *ICML*, volume 97, pages 4723–4732. PMLR.

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10(1):1–51.

Nogueira, A. R., Pugnana, A., Ruggieri, S., Pedreschi, D., and Gama, J. (2022). Methods and tools for causal discovery and causal inference. *WIREs Data Mining Knowl. Discov.*, 12(2).

Okati, N., De, A., and Gomez-Rodriguez, M. (2021). Differentiable learning under triage. In *NeurIPS*, pages 9140–9151.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Perini, L. and Davis, J. (2023). Unsupervised anomaly detection with rejection. In *NeurIPS*.

Pinotti, P. (2017). Clicking on Heaven's Door: The Effect of Immigrant Legalization on Crime. *American Economic Review*, 107(1):138–168.

Pugnana, A., Perini, L., Davis, J., and Ruggieri, S. (2024). Deep neural network benchmarks for selective classification. *Journal of Data-centric Machine Learning Research (DMLR)*, 1(17):1–58.

Pugnana, A. and Ruggieri, S. (2023a). AUC-based selective classification. In *AISTATS*, volume 206, pages 2494–2514. PMLR.

Pugnana, A. and Ruggieri, S. (2023b). A model-agnostic heuristics for selective classification. In *AAAI*, pages 9461–9469. AAAI Press.

Punzi, C., Pellungrini, R., Setzu, M., Giannotti, F., and Pedreschi, D. (2024). AI, Meet Human: Learning paradigms for hybrid decision making systems. *arXiv preprint arXiv:2402.06287*.

Raghu, M., Blumer, K., Corrado, G., Kleinberg, J. M., Obermeyer, Z., and Mullainathan, S. (2019). The algorithmic automation problem: Prediction, triage, and human effort. *CoRR*, abs/1903.12220.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP/IJCNLP (1)*, pages 3980–3990. Association for Computational Linguistics.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Rubin, D. B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics*, 6(1):34–58.

Ruggieri, S., Álvarez, J. M., Pugnana, A., State, L., and Turini, F. (2023). Can we trust fair-AI? In *AAAI*, pages 15421–15430. AAAI Press.

Stolberg, H. O., Norman, G., and Trop, I. (2004). Randomized controlled trials. *American Journal of Roentgenology*, 183(6):1539–1544.

Thistlethwaite, D. L. and Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51(6):309–317.

Tortorella, F. (2005). A ROC-based reject rule for dichotomizers. *Pattern Recognit. Lett.*, 26(2):167–180.

Van der Plas, D., Meert, W., Verbraecken, J., and Davis, J. (2023a). A novel reject option applied to sleep stage scoring. In *SDM*, pages 820–828. SIAM.

Van der Plas, D., Meert, W., Verbraecken, J., and Davis, J. (2023b). A novel reject option applied to sleep stage scoring. In *SDM*, pages 820–828. SIAM.

Verma, R., Barrejón, D., and Nalisnick, E. T. (2023). Learning to defer to multiple experts: Consistent surrogate losses, confidence calibration, and conformal ensembles. In *AISTATS*, volume 206, pages 11415–11434. PMLR.

Verma, R. and Nalisnick, E. T. (2022). Calibrated learning to defer with one-vs-all classifiers. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 22184–22202. PMLR.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. (2017). Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*, pages 3462–3471. IEEE Computer Society.

Wang, X. and Yiu, S. (2020). Classification with rejection: Scaling generative classifiers with supervised deep infomax. In *IJCAI*, pages 2980–2986. ijcai.org.

Wei, Z., Cao, Y., and Feng, L. (2024). Exploiting human-ai dependence for learning to defer. In *ICML*. OpenReview.net.

Zagoruyko, S. and Komodakis, N. (2016). Wide residual networks. In *BMVC*. BMVA Press.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] See Table 2.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] We include the time and carbon footprint required in Section D.1.4, while for sample size, we refer to Table 2.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes] The code can be found at `https://anonymous.4open.science/r/PODS-565D`.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] All the assumptions are made in the main text.
   (b) Complete proofs of all theoretical results. [Yes] Proofs are in Appendix B for space reasons.
   (c) Clear explanations of any assumptions. [Yes] We detail the two scenarios in the main paper. See Section 3.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] The code can be found at `https://anonymous.4open.science/r/PODS-565D`.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] See Section 4.1 in the main paper and SectionsD.1.3 and D.1.1 in the Appendix.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] We detail these measures in Section 4.1.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] See D.1.4.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes]

   (b) The license information of the assets, if applicable. [Yes]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes] The code can be found at `https://anonymous.4open.science/r/PODS-565D`.

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A EXTENDED RELATED WORK

Abstaining systems. The idea to allow ML models to abstain (a.k.a. to reject) from predicting dates back to the 1970s, with the seminal work by Chow (1970).

In the literature, two kinds of rejections have been considered: novelty rejection and ambiguity rejection. The former provides methods that abstain when the instances are far away from the training data distribution; the latter abstains on instances close to the decision boundary of the classifier (Hendrickx et al., 2024).

Novelty rejection is highly sought if a shift between the training and the test set distributions can occur (Van der Plas et al., 2023a). Multiple approaches have been proposed for building novelty rejectors. For instance, one can estimate the marginal density of the training distribution and reject an instance if its probability is below a certain threshold (Nalisnick et al., 2019; Wang and Yiu, 2020). Another approach relies on a one-class classification model that predicts as novel the instances falling out of the region learnt from the training set (Coenen et al., 2020). Alternatively, one can provide a score representing the novelty of an instance and abstain when such a score is above a certain level (Liang et al., 2018; Kühne et al., 2021; Perini and Davis, 2023; Van der Plas et al., 2023b).

Regarding ambiguity rejection, two main approaches have emerged in the literature: Learning to Reject (LtR) (Chow, 1970) and Selective Prediction (SP) (El-Yaniv and Wiener, 2010).

LtR - based on the original work by Chow (1970) - aims at learning a pair (classifier, rejector) such that the rejector determines when the classifier predicts, limiting the predictions to the region where the classifier is likely correct (Cortes et al., 2023). The LtR methods learn the trade-off between abstention and prediction through a parameter a, representing the cost of rejection (Herbei and Wegkamp, 2006; Cortes et al., 2016; Tortorella, 2005; Condessa et al., 2013).

On the other hand, SP methods rely on confidence functions, identifying instances where the classifier is more prone to make mistakes (El-Yaniv and Wiener, 2010). Confidence values allow one to trade off coverage c for selective risk, namely the risk over those instances for which a prediction is provided. Such a trade-off can be used to frame the learning problem in two ways: either we maximize coverage given a minimal target risk we want to ensure (bounded-improvement problem), or we minimize the selective risk given a target coverage (bounded-abstention problem) (Franc et al., 2023).

From a practical perspective, both model-agnostic methods (e.g., (Denis and Hebiri, 2020; Pugnana and Ruggieri, 2023b,a)) and model-specific ones (e.g., using Deep Neural Network architectures (Geifman and El-Yaniv, 2017, 2019; Corbière et al., 2019; Huang et al., 2020; Feng et al., 2023)) have been proposed to solve the selective prediction task. For an extensive characterization of deep-neural-network (DNN)-based approaches, we refer to Pugnana et al. (2024), where the authors empirically compare existing DNN-based approaches, categorizing them depending on how they abstain.

An in-depth theoretical analysis for both LtR and SP can be found in Franc et al. (2023), where the authors show that both frameworks share similar optimal strategies, and a recent survey covering abstaining systems can be found in Hendrickx et al. (2024), where the authors provide an overall taxonomy of existing approaches.

Choe et al. (2023) consider the evaluation of abstaining classifiers, in a scenario where the predictions are missing or unaccessible for the rejected instances. They define the *counterfactual score* as the expected accuracy of the classifier had it not been given the option to abstain. A double-ML approach (Chernozhukov et al., 2018) is used to estimate the counterfactual score. In order to exploit results of inference under missing data, the paper assumes a stochastic abstention policy, which is impractical/unethical in the context of deferring systems: instances are not deferred to a human expert at random.

Deferring systems. Learning to Defer (LtD) - as framed by Madras et al. (2018) - is a generalization of LtR, where rather than incurring a rejection cost, the system can defer instances to human expert(s). Compared with LtR and SP, one of the main differences is that the expert's predictions might be wrong under the LtD framework.

From a theoretical perspective, De et al. (2020) show that the problem of learning to defer when choosing a ridge regression as a base predictor is NP-hard. By reformulating the problem using submodular functions, they devise a greedy algorithm with some theoretical guarantees. Similar results also hold for the classification setting when considering margin-based classifiers, as shown in (De et al., 2021). Okati et al. (2021) formally characterises the scenarios where a predictive model can take advantage of including humans in the loop. They show that standard ML models trained to predict over all the instances may be suboptimal when it comes to LtD, proposing a deterministic threshold rule to determine when the ML model or the human has to predict.

Due to the difficulties in directly optimizing (2), several approaches provide surrogate losses to learn predictors that can defer to experts: Mozannar and Sontag (2020) propose a method that jointly learns both the rejector and the ML predictor with some generalization bounds; Verma and Nalisnick (2022) and Cao et al. (2023) extend the work of Mozannar and Sontag (2020) by providing consistent surrogate loss with better calibration properties; Mozannar et al. (2023) provide a Mixed Integer Linear Programming formulation to solve the problem in the linear setting and a novel surrogate loss that is realizable and consistent; Liu et al. (2024) propose new surrogate losses that are not prone to underfitting; Wei et al. (2024) consider another surrogate loss that takes into account the dependency between the ML model predictions and the human ones.

A few works also consider extensions to staged learning, where the ML model is already given and not jointly trained, e.g., Charusaie et al. (2022) and Mao et al. (2023). To conclude, extensions to multi-experts can be found in Verma et al. (2023) and Cao et al. (2023).

Regression discontinuity. The RD design first appeared in Thistlethwaite and Campbell (1960) to study the motivational effect of public recognition on the likelihood of obtaining a scholarship. Forty years later, Hahn et al. (2001) proposed a thorough formalization of this methodology, which shows the identification of the average treatment effect on the treated via smoothness of the potential outcomes. An alternative framework for identification has been presented in Lee (2008) and Cattaneo et al. (2015), where the authors carefully propose conditions under which, at least near the cutoff, the RD design can be interpreted as an RCT. Early reviews are Imbens and Angrist (1994) and Lee and Lemieux (2010), whereas a more recent one contrasting the two approaches mentioned above is Cattaneo and Titiunik (2022).

For estimation purposes, the local nature of $\tau_{\texttt{RD}}$ motivates the use of local non-parametric polynomial kernel regressions estimators from the left and from the right of the cutoff (for a review see Fan and Gijbels (1996)). Particular attention in the literature has been devoted to how to optimally choose the smoothing bandwidth used in local polynomial estimation (see Imbens and Angrist (1994); Calonico et al. (2014); Kolesár and Rothe (2018)) and how to correct for the smoothing bias and conduct valid inference accordingly (Calonico et al., 2018, 2022). For a practical introduction to RD and more information on how to choose the other tuning parameters (kernel shape and degree of the polynomial), see Cattaneo et al. (2019).

B PROOFS

We report here the proofs for propositions from Section 3. For the reader's convenience, we restate the claims.

Proposition 1. *Let Scenario 1 hold. Then, for each $i \in [1, n]$ such that $G_i = 1$:*

$$\tau_i = T_i(1) - T_i(0) = \mathbb{1}\{h(\mathbf{X}_i) = Y_i\} - \mathbb{1}\{f(\mathbf{X}_i) = Y_i\}.$$

Proof. Identification is immediate because both potential outcomes are observed for unit $i \in [1, n]$ that has also been evaluated by a human ($G_i = 1$). In particular, $T_i(0) = \mathbb{1}\{f(\mathbf{X}_i) = Y_i\}$ is observable because we are under Scenario 1 and $T_i(1) = \mathbb{1}\{h(\mathbf{X}_i) = Y_i\}$) since $G_i = 1$. □

Proposition 2. *Let Scenario 1 hold. Then:*

$$\tau_{\texttt{ATD}} = \mathbb{E}[T(1) - T(0)|G = 1] = \mathbb{E}\left[\mathbb{1}\{h(\mathbf{X}) = Y\} \mid G = 1\right] - \mathbb{E}\left[\mathbb{1}\{f(\mathbf{X}) = Y\} \mid G = 1\right].$$

Proof. We have that:

$$\begin{aligned}\tau_{\texttt{ATD}} &= \mathbb{E}\left[T(1) \mid G = 1\right] - \mathbb{E}\left[T(0) \mid G = 1\right] \\ &= \mathbb{E}\left[\mathbb{1}\{h(\mathbf{X}) = Y\} \mid G = 1\right] - \mathbb{E}\left[\mathbb{1}\{f(\mathbf{X}) = Y\} \mid G = 1\right],\end{aligned}$$

where the last two quantities are observable under Scenario 1. □

Proposition 3. *Let Scenario 1 hold. Then:*
(i) as $n \to \infty$, $\hat{\tau}_{\texttt{ATD}} \xrightarrow{p} \mathbb{E}[T(1) - T(0) \mid G = 1] = \tau_{\texttt{ATD}}$
(ii) for each $n > 0$, $\hat{\tau}_{\Delta} = \frac{n_1}{n}\hat{\tau}_{\texttt{ATD}}$.

Proof. For (i), let $g_i : \mathbb{1}\{g(\mathbf{x}_i) = 1\}$ and so $n_1 := \sum_{i=1}^n g_i$. Then:

$$\begin{aligned}\hat{\tau}_{\texttt{ATD}} &= \frac{1}{n_1}\left(\sum_{i:g(\mathbf{x}_i)=1} \left(\mathbb{1}\{h(\mathbf{x}_i) = y_i\} - \mathbb{1}\{f(\mathbf{x}_i) = y_i\}\right)\right) \\ &= \left(\frac{n_1}{n}\right)^{-1} \frac{1}{n}\left(\sum_{i=1}^{n} g_i \left(\mathbb{1}\{h(\mathbf{x}_i) = y_i\} - \mathbb{1}\{f(\mathbf{x}_i) = y_i\}\right)\right).\end{aligned}$$

The weak law of large numbers immediately gives us:

$$\frac{n_1}{n} \xrightarrow{p} \mathbb{P}[G = 1], \qquad \frac{1}{n}\left(\sum_{i=1}^{n} g_i \left(\mathbb{1}\{h(\mathbf{x}_i) = y_i\} - \mathbb{1}\{f(\mathbf{x}_i) = y_i\}\right)\right) \xrightarrow{p} \mathbb{E}[G(T(1) - T(0))].$$

Then, using the fact that $\mathbb{E}[G(T(1)-T(0))] = \mathbb{P}[G = 1]\mathbb{E}[T(1)-T(0) \mid G = 1]$, we can conclude that $\hat{\tau}_{\texttt{ATD}} \xrightarrow{p} \tau_{\texttt{ATD}}$. For (ii), by definition, we have that:

$$\widehat{\texttt{Acc}}_{\vartheta} := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{\vartheta(\mathbf{x}_i) = y_i\} = \frac{1}{n}\left[\sum_{i:g(\mathbf{x}_i)=0} \mathbb{1}\{f(\mathbf{x}_i) = y_i\} + \sum_{i:g(\mathbf{x}_i)=1} \mathbb{1}\{h(\mathbf{x}_i) = y_i\}\right],$$

and

$$\widehat{\texttt{Acc}}_{f} := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{f(\mathbf{x}_i) = y_i\} = \frac{1}{n}\left[\sum_{i:g(\mathbf{x}_i)=0} \mathbb{1}\{f(\mathbf{x}_i) = y_i\} + \sum_{i:g(\mathbf{x}_i)=1} \mathbb{1}\{f(\mathbf{x}_i) = y_i\}\right].$$

Therefore, when we consider the difference between the two accuracies we have that:

$$\begin{aligned}\hat{\tau}_{\Delta} &:= \widehat{\texttt{Acc}}_{\vartheta} - \widehat{\texttt{Acc}}_{f} \\ &= \frac{1}{n}\left(\sum_{i:g(\mathbf{x}_i)=1} \left(\mathbb{1}\{h(\mathbf{x}_i) = y_i\} - \mathbb{1}\{f(\mathbf{x}_i) = y_i\}\right)\right) \\ &= \frac{n_1}{n}\frac{1}{n_1}\left(\sum_{i:g(\mathbf{x}_i)=1} \left(\mathbb{1}\{h(\mathbf{x}_i) = y_i\} - \mathbb{1}\{f(\mathbf{x}_i) = y_i\}\right)\right) \\ &= \frac{n_1}{n}\hat{\tau}_{\texttt{ATD}}.\end{aligned}$$

□

Proposition 4. *Let Scenario 2 hold and let Assumption 1 be satisfied for the deferring system. Then:*

$$\lim_{k \to \overline{\kappa}_c^+} \mathbb{E}[T \mid K = k] - \lim_{k \to \overline{\kappa}_c^-} \mathbb{E}[T \mid K = k] = \tau_{\texttt{RD}},$$

where $\tau_{\texttt{RD}} := \mathbb{E}[T(1) - T(0) \mid K = \overline{\kappa}_c]$.

Proof. This is an instance of Theorem 1. □

C EXTENSIONS AND LIMITATIONS

Here are a few caveats regarding the proposed framework.

Further discussion on Proposition 3 and identifiability If we have access to accuracy estimates for $\widehat{\texttt{Acc}}_{\vartheta}$ and $\widehat{\texttt{Acc}}_{f}$, we can just re-weight $\hat{\tau}_{\Delta}$ by n/n_1 to obtain a consistent estimator of $\tau_{\texttt{ATD}}$. This depicts a setting in between Scenario 1 and Scenario 2. Indeed, having access to n_1, n, and accuracy estimates for the ML model and the deferring system allows us to obtain an estimate for $\tau_{\texttt{ATD}}$ without requiring the full knowledge of individual ML model predictions for the deferred instances. We further note that this reweighting is nothing else than a Horovitz-Thompson type of reweighting (Horvitz and Thompson, 1952) where the observations are weighted by the inverse of the sampling probability. Here the "sampling probability" denotes the probability that an instance is assigned to a human.

Is Assumption 1 met? Assumption 1 is required to be able to identify the causal effect under Scenario 2. Formally, such an assumption requires continuity of the expected predictive accuracy at coverage level $\overline{\kappa}_c$ for *both* the ML model and the human. In practical terms, there is no reason to believe that the accuracy of the ML model would abruptly change in a neighborhood of $\overline{\kappa}_c$. However, continuity of $\mathbb{E}[T(1) \mid K = k]$ around $\overline{\kappa}_c$ could be falsified, e.g., if the expert would put extra effort into predicting deferred instances compared to the non-deferred ones. Concerning this aspect, Bondi et al. (2022) show that the role of communicating the deferral status can indeed impact human performance. On the one hand, they observe improvements in human accuracy for those instances where the ML model is correct if the deferral choice is communicated to the human predictors. On the other hand, they do not observe a statistically significant effect in those instances in which the ML model makes mistakes. Therefore, we advise considering this aspect when developing and evaluating a deferring system.

Is Assumption 1 testable? Unfortunately, Assumption 1 is not directly testable because ML predictions ($T(0)$) and human predictions ($T(1)$) are not available when $K \geq \overline{\kappa}_c$ or $K < \overline{\kappa}_c$, respectively. However, it is possible (and suggested) to test the implications of Assumption 1. In what follows, we briefly describe three different falsification tests. We refer the reader interested in a more detailed review and guide on these (and other) tests to (Lee and Lemieux, 2010) and (Cattaneo et al., 2019, Chapter 5).

Non-manipulation of the running variable. One instance in which Assumption 1 does not hold is when units know the rule with which the running variable K_i is computed and/or can manipulate its value. Accordingly, a feasible falsification test involves checking whether the empirical probability distribution of the running variable is smooth (i.e., it does not jump) at the cutoff. This test can be formally conducted by estimating the density of the running variable with histograms or kernel density estimators (Cattaneo et al., 2019, 2020). For a practical example, see Section D.3 and Figures 6-10.

No effect on predetermined features and placebo outcomes. In a similar spirit to what we described above, another falsification test involves comparing treated and control units near the cutoff to see if they share similar observable traits: if units can't manipulate their score, there should not be systematic differences between units close to the cutoff, aside from their treatment status. As such, units just above and below the cutoff should resemble each other in all aspects unaffected by the treatment. These aspects (or features) can be *predetermined features*, variables realizing before treatment assignment, or *placebo outcomes*, variables that should not have been influenced by the treatment. This test can be conducted by estimating an RD where the outcome variable is either a predetermined feature or a placebo outcome and checking that the null hypothesis of no effect is not rejected. An application of this falsification test can be found in Section D.3.1 and in the fourth row of Figure 5.

Placebo cutoffs. Another type of falsification test involves checking if statistically significant treatment effects can be estimated using artificial cutoff values. We stress that the evidence of no effect –hence smoothness of the expected potential outcomes– away from the cutoff is neither sufficient nor necessary for Assumption 1 to hold, but the presence of discontinuities in other places might discredit such an assumption. To conduct such a test, one has to estimate the RD using a cutoff value that is different from the original one. Section D.3.1 and the second and third rows of Figure 5 showcase this falsification test in our empirical applications.

Optimal coverage We again stress that RD designs are local in nature. Hence, under Scenario 2, without imposing additional strong parametric assumptions on the shape of $\mathbb{E}[T(d)|K = k], d \in \{0, 1\}$, we cannot recover the average treatment effect for coverage levels other than $\overline{\kappa}_c$. However, suppose that we have access to different batches of data, where similar human experts and the same ML model take turns in predicting the ground truth Y, but the reject-score threshold was let vary in a countable (ordered) set $\overline{\mathcal{K}} \subseteq \mathcal{K}$. For instance, we can be in the presence of multiple human moderators that receive different amounts of content to moderate (i.e. $\overline{\kappa}$ is changing). Then, we can leverage results in the literature of RD designs with multiple non-cumulative cutoffs (Cattaneo et al., 2016, 2021) and identify

$$\tau_{\mathtt{RD}}(k) := \mathbb{E}[T(1) - T(0) \mid K = k], \qquad k \in [\overline{\kappa}_{\mathtt{lb}}, \overline{\kappa}_{\mathtt{ub}}],$$

where $\overline{\kappa}_{\mathtt{lb}} := \min \overline{\mathcal{K}}$ and $\overline{\kappa}_{\mathtt{ub}} := \max \overline{\mathcal{K}}$. More precisely, identification of $\tau_{\mathtt{ATD}}(k)$ requires continuity of $\mathbb{E}[T(d) \mid K = k], d \in \{0, 1\}$ in k for $k \in [\overline{\kappa}_{\mathtt{lb}}, \overline{\kappa}_{\mathtt{ub}}]$ and a similar shape of $\mathbb{E}[T(0) \mid K_i = k]$ across datasets. An exercise in this spirit might be useful to choose the optimal level of coverage $\overline{\kappa}^{\star} := \arg\max_{k \in \overline{\mathcal{K}}} \tau_{\mathtt{ATD}}(k)$, where the optimality is defined in the sense of getting the largest increase in predictive accuracy out of having human experts guessing instead of the model.

Table 2: Datasets and baselines details (epochs - `ep.`, learning rate - `lr`, optimizer - `op.`).

DATASET	n	$\vert\mathcal{Y}\vert$	HUMAN	MODEL	PRE-TRAINED	HYPER-PARAMETERS
`synth`	50k	2	synthetic	linear	no	$\texttt{ep.} = 50; \texttt{lr} = 1e-2; \texttt{op.} = \texttt{Adam}$
`cifar10h`	10k	10	separate annotator	WideResNet	yes	$\texttt{ep.} = 150; \texttt{lr} = 1e-3; \texttt{op.} = \texttt{AdamW}$
`galaxyzoo`	10k	2	random annotator	ResNet50	yes	$\texttt{ep.} = 50; \texttt{lr} = 1e-3; \texttt{op.} = \texttt{Adam}$
`hatespeech`	25k	3	random annotator	FNN on SBERT embeddings	yes	$\texttt{ep.} = 100; \texttt{lr} = 1e-2; \texttt{op.} = \texttt{Adam}$
`xray-airspace`	4.4k	2	random annotator	DenseNet121	yes	$\texttt{ep.} = 3; \texttt{lr} = 1e-3; \texttt{op.} = \texttt{AdamW}$

Should we increase or decrease the coverage? Despite being local by construction, the RD can be used to learn about the gradient of the treatment effect at the cutoff. Indeed, Dong and Lewbel (2015) show that, under regularity conditions on how $\mathbb{E}[T(d) \mid K = k], d \in \{0, 1\}$ changes around $\overline{\kappa}_c$, the RD setting can be used to learn how $\tau_{\texttt{RD}}$ would change if the reject-score threshold $\overline{\kappa}_c$ were marginally changed. Therefore, this suggests that when the deferring system accuracy is maximal, $\tau_{\texttt{RD}}$ should be close to zero. The intuition is that we should be indifferent between predicting with the ML model or deferring to the human expert at the optimal value.

Uncertainty in the ML model estimation The outcome variable for unit i in the LtD framework is $T_i = \mathbb{1}\{\vartheta(\mathbf{X}_i) = Y_i\}$ and the running variable is the reject score $K_i = k(\mathbf{X}_i)$ (see Table 1). However, in practice, we compute the outcome and the reject score via an estimated version of the model, i.e. $\hat{f}(\cdot)$. We explicitly do not consider this source of uncertainty because access is usually limited to an estimated final version of the model without possibly fitting it again (e.g., large language models (Brown et al., 2020)). For this reason, with a slight abuse of notation, we always write $f(\cdot)$, T_i, and K_i when we should write $\hat{f}(\cdot)$, $\hat{T}_i$, and $\hat{K}_i$, respectively. If one is willing also to capture this uncertainty and model re-estimation is viable, off-the-shelf non-parametric bootstrap procedures are available. Moreover, (Dong and Kolesár, 2023) show that under the condition that the noisy score correctly assigns samples to treatment and control groups, $\tau_{\texttt{RD}}$ can be interpreted as the treatment effect when the *noisy* score equals the cutoff. We argue that this latter measure is still the one of interest, particularly so when the model is taken as given because the reject score can only be computed using the estimated ML model.

D EXTENDED EXPERIMENTAL EVALUATION

D.1 Additional details

D.1.1 Data

We use the data from Mozannar et al. (2023) and Okati et al. (2021).

Concerning the synthetic dataset `synth`, we generate the data using the method described in Mozannar et al. (2023)[7]: given a parameter d, $\mathbf{x} \in \mathbb{R}^d$ is sampled from a mixture of d equally weighted Gaussians, each one with uniformly random mean and variance. To obtain the target variable Y, the procedure generates two random half-spaces, one referring to the optimal policy function $g^* : \mathcal{X} \to \{0, 1\}$ and one representing an optimal ML model $f^* : \mathcal{X} \to \mathcal{Y}$. The fraction of instances for which $g^*(\mathbf{x}) = 0$ is randomly chosen to be between .20 and .80. For all those instances on the side where $g^*(\mathbf{x}) = 0$, the target variable Y is changed to be consistent with the optimal ML model $f^*(\mathbf{x})$ with probability $1 - p_{ML}$ and otherwise uniform. Conversely, when $g^*(\mathbf{X}) = 1$, the labels are uniformly sampled. The human expert $h(\mathbf{x})$ is then set to make mistakes at a rate of p_{h0} when $g^*(\mathbf{x}) = 0$ and at a rate of p_{h1} when $g^*(\mathbf{x}) = 1$. In our experiments, we set $d = 10$, $p_{h0} = .10$, $p_{h1} = .10$, $p_{ML} = .40$, as done in the unrealizable setting by Mozannar et al. (2023).

Regarding real data, in `cifar10h` (Battleday et al., 2020), the task is to annotate images belonging to 10 different categories. Here, the human prediction is provided by a separate human annotator.

In `galaxyzoo` (AstroDave et al., 2013), the main task is identifying whether the image contains a non-smooth galaxy. Thus, $Y = 0$ if the image contains a smooth galaxy and $Y = 1$ otherwise. Since we have 30 annotators for each image, we consider the majority of the annotators as the target variable Y, while the human expert prediction $h(\mathbf{X})$ is sampled randomly from the 30 annotators.

[7] See the GitHub repository `https://github.com/clinicalml/human_ai_deferral`

For hatespeech (Davidson et al., 2017), the goal is to detect whether the text contains offensive or hate-speech language. The human predictor is sampled randomly as in Mozannar et al. (2023).

Finally, xray-airspace (Wang et al., 2017; Majkowska et al., 2020) contains both chest X-rays with human predictions and chest X-rays without human predictions. For each image, the target variable Y encodes the presence of an airspace opacity. The human predictions are randomly sampled from multiple experts, as done by Mozannar et al. (2023).

D.1.2 Baselines

We include in our experiments seven baselines for which the code was publicly available. For all the baselines but ASM, we consider the implementation provided by Mozannar et al. (2023)[7]. We use Cao et al. (2023)'s code for ASM, as provided in their supplementary material.

Selective Prediction (SP): Geifman and El-Yaniv (2017) present a neural network classifier with a reject option. The reject score is defined considering the maximum of the final softmax values, i.e., $k(\mathbf{x}) = \max_y s_y(\mathbf{x})$, where $s_y(\mathbf{x})$ is the final layer softmax value for class y. We stress that SP does not take into account the human expert's ability but determines deferral only based on those cases where the ML model is uncertain.

Compare Confidence (CC): Raghu et al. (2019) extend SP by learning (independently) another model - called the expert model - that uses as a target variable whether the human expert is correct. Then, deferral is determined by comparing the reject score of the classifier and the expert model.

Differentiable Triage (DT): Okati et al. (2021) consider a two-stage approach, where at each epoch *(i)* the classifier is trained only on those points where the classifier loss is lower than the human loss, *(ii)* another ML model - called the rejector - is fitted to predict who has a lower loss between the classifier and the human. At the end of the training procedure, deferral is decided based on the estimated probability of the human expert having a lower loss than the classifier.

Cross-Entropy Surrogate (LCE): Mozannar and Sontag (2020) propose a consistent surrogate loss of (2), when l is the 0-1 loss. The surrogate loss is then used to train the deferring system, which employs a neural network with an additional head to represent deferral;

One Vs All (OVA): Verma and Nalisnick (2022) propose a different consistent surrogate loss, which improves the final calibration of the deferring system;

Realizable Surrogate (RS): Mozannar et al. (2023) extend the approaches based on surrogate losses by considering a consistent and realizable-consistent surrogate of (2), when l_M and l_H are the 0-1 loss. As for LCE, also RS considers a neural network with an additional head representing deferral.

Asymmetric SoftMax (ASM): Cao et al. (2023) extend both LCE and OVA by providing a surrogate loss that ensures a better calibration of the reject score.

D.1.3 Hyperparameters

For synth, we considered a simple linear feedforward architecture. For each baseline, we trained the model for 50 epochs with $\texttt{lr} = 1e-2$ and Adam (Kingma and Ba, 2015) as the optimizer. Batch size was set to $1,024$.

For real data, all the methods were trained following the settings in either Mozannar et al. (2023) (cifar10h, hatespeech, xray-airspace) or Okati et al. (2021) (galaxyzoo).

In particular, for cifar10h, we trained a base WideResNet (Zagoruyko and Komodakis, 2016) on the original cifar10 dataset for 200 epochs using cross-entropy loss, learning rate equals to .001 and AdamW (Loshchilov and Hutter, 2019) as an optimizer. For each baseline, we fine-tuned the base WideResnet on cifar10h for 150 epochs, using a learning rate of .001 and AdamW as an optimizer.

For hatespeech, we considered pre-trained embeddings of SBERT and we fine-tuned a feed-forward neural network (Reimers and Gurevych, 2019) for 100 epochs, setting the learning rate to .01 and Adam as optimizer.

For xray-airspace, we first fine-tuned a pre-trained DenseNet121 (Huang et al., 2017) for 10 epochs on the x-rays that do not contain human predictions, setting $\texttt{lr} = 1\text{e}{-4}$ and AdamW as the optimizer. For each baseline, we further fine-tuned the obtained model, training it for 3 epochs on xray-airspace with a learning rate equal to 1e$-$3 and `AdamW` as the optimizer.

Finally, for `galaxyzoo`, we consider a pre-trained ResNet50 (He et al., 2016) and train each baseline for 50 epochs, using `Adam` as the optimizer and a learning rate of 1e$-$3.

The batch size was set to 128 for all the real datasets.

D.1.4 Hardware and carbon footprint

Regarding computational resources, we split the workload over two machines: (*i*) a 96 cores machine with Intel(R) Xeon(R) Gold 6342 CPU @ 2.80GHz and two NVIDIA RTX A6000, OS Ubuntu 20.04.4; (*ii*) a 224 cores machine with Intel(R) Xeon(R) Platinum 8480+ CPU and eight NVIDIA A100-SXM4-80GB, OS Ubuntu 22.04.4 LTS.

We track all our runs using the `Python` package `codecarbon` (Courty et al., 2024). This allows us to consider the total time required by all our experimentation (including failed and repeated experiments) and its environmental costs. Overall, the cumulated time of all our runs amounts to $\approx$ 5 days. This translates into an overall CO_2 consumption of roughly $\approx$ 25.2 Kg Eq.CO_2, which equals a car drive of $\approx$ 60 miles.

D.2 Detailed results for Q1, Q2 and Q3

Tables 3-7 report the detailed results of experiments on the synthetic and the real datasets. Tables include $\hat{\tau}_{\texttt{ATD}}$ (the Scenario 1 main estimate), $\hat{\tau}_{\texttt{RD}}$ (the Scenario 2 main estimate), and the deferring system accuracy at various target coverages for all the seven baselines. For $\hat{\tau}_{\texttt{ATD}}$ and $\hat{\tau}_{\texttt{RD}}$, we show in parentheses the associated p-value when testing the significance of the causal effect being different from zero. Significant (after Bonferroni correction of $\alpha = 0.05$) p-values are shown in blue.

D.3 How to validate estimates under Scenario 2

When considering Scenario 2, we require Assumption 1 to hold. Here, we provide a few sanity checks that can be implemented to validate the estimates (see Appendix C for a detailed description).

D.3.1 Placebo Tests

Placebo cutoff test setup. We consider the same setup presented in Section 4.1, except for the following:

1. we estimate two different cutoff values $\overline{\kappa}_{c,L}$ and $\overline{\kappa}_{c,H}$
   - $\overline{\kappa}_{c,L}$ is obtained by considering the 75-th percentile on the instances with a reject score below $\overline{\kappa}_c$,
   - $\overline{\kappa}_{c,H}$ is estimated by considering the 25-th percentile of the reject scores above $\overline{\kappa}_c$;
2. we run a local kernel polynomial regression to estimate $\hat{\tau}_{\texttt{RD}}$, substituting the true cutoff value $\overline{\kappa}_c$ with both $\overline{\kappa}_{c,L}$ and $\overline{\kappa}_{c,H}$.

Placebo outcome test setup. We consider the same setup presented in Section 4.1, except for the following:

1. we sample a placebo outcome $\tilde{T}$, such that $\tilde{T} \sim \texttt{Bernoulli}(p)$ and $p = .5$;
2. we run the same local kernel polynomial regression used to estimate $\hat{\tau}_{\texttt{RD}}$ substituting the target variable T with $\tilde{T}$.

Results Figure 5 provides the results for the placebo tests above. The first row shows the original estimates for $\hat{\tau}_{\texttt{RD}}$ under Scenario 2 for the best baselines.

Regarding `synth`, all the coefficients of the placebo tests are not significant, with the sole exception of high coverage values ($c \geq .80$) when considering the highest cutoff (third row of Figure 5).

When considering `cifar10h`, `hatespeech`, and `xray-airspace`, all the coefficients for all the placebo tests are not significant. This suggests that our estimated $\hat{\tau}_{\texttt{RD}}$ should be robust.

Table 3: `synth` results, with the statistically significant ones at $\alpha = 0.05$ in blue.

	c	ASM	CC	DT	LCE	OVA	RS	SP
$\hat{\tau}_{\texttt{ATD}}$	.00	.083 (8.20e−35)	.135 (8.55e−89)	.107 (2.55e−56)	**.143** (1.04e−100)	.138 (9.57e−93)	.102 (2.31e−51)	.132 (4.40e−85)
	.10	.133 (2.12e−78)	.194 (1.42e−165)	.153 (2.32e−103)	**.205** (1.16e−186)	.199 (1.88e−174)	.157 (8.86e−110)	.188 (1.94e−155)
	.20	.199 (2.88e−157)	.254 (6.03e−265)	.208 (1.63e−173)	**.261** (4.03e−285)	.259 (1.49e−276)	.218 (3.05e−193)	.230 (5.03e−215)
	.30	.269 (1.48e−266)	.310 (0.00)	.269 (2.67e−264)	.314 (0.00)	**.318** (0.00)	.280 (2.94e−291)	.259 (2.08e−245)
	.40	.353 (0.00)	.350 (0.00)	.332 (0.00)	.343 (0.00)	**.359** (0.00)	.347 (0.00)	.289 (5.03e−269)
	.50	**.406** (0.00)	.373 (0.00)	.371 (0.00)	.372 (0.00)	.392 (0.00)	.389 (0.00)	.300 (1.39e−244)
	.60	.406 (0.00)	.387 (0.00)	.393 (0.00)	.390 (0.00)	.409 (0.00)	**.409** (0.00)	.326 (1.07e−238)
	.70	.400 (0.00)	.403 (0.00)	.399 (0.00)	.404 (0.00)	.402 (0.00)	**.406** (0.00)	.335 (9.06e−196)
	.80	.396 (2.91e−212)	.404 (4.90e−225)	.391 (3.41e−203)	**.416** (3.31e−244)	.414 (8.57e−234)	.404 (2.03e−220)	.340 (3.38e−137)
	.90	.402 (3.01e−109)	.393 (4.80e−104)	.370 (2.87e−96)	**.414** (7.30e−119)	.400 (6.82e−111)	.398 (2.38e−108)	.345 (9.38e−73)
$\hat{\tau}_{\texttt{RD}}$	.10	−.302 (2.54e−7)	−.293 (5.63e−11)	−**.244** (3.59e−6)	−.258 (4.43e−6)	−.402 (1.94e−7)	−.313 (5.94e−8)	−.344 (1.61e−8)
	.20	−.356 (9.60e−16)	−.231 (2.25e−8)	−.216 (2.89e−7)	−.240 (3.01e−8)	−.221 (1.29e−6)	−.342 (2.10e−10)	−**.098** (7.13e−2)
	.30	−.344 (3.60e−9)	−.048 (2.65e−1)	−.145 (1.09e−4)	.016 (6.97e−1)	−.036 (4.53e−1)	−.237 (1.18e−5)	**.086** (4.41e−2)
	.40	−.345 (5.24e−2)	.179 (1.95e−6)	.028 (4.16e−1)	**.185** (8.04e−7)	.178 (1.25e−3)	.006 (8.93e−1)	.129 (5.45e−4)
	.50	**.288** (9.25e−3)	.282 (1.37e−13)	.229 (1.98e−11)	.243 (1.47e−10)	.205 (3.36e−5)	.230 (1.56e−10)	.159 (4.26e−4)
	.60	**.485** (3.07e−26)	.333 (3.12e−20)	.299 (5.88e−11)	.317 (1.12e−18)	.415 (6.81e−29)	.443 (6.63e−30)	.209 (2.22e−7)
	.70	.431 (7.20e−25)	.334 (6.93e−21)	.395 (1.30e−22)	.366 (8.37e−27)	**.473** (1.19e−34)	.433 (1.04e−19)	.333 (3.88e−11)
	.80	**.479** (1.97e−55)	.406 (3.47e−24)	.427 (2.49e−22)	.404 (4.36e−30)	.384 (9.31e−27)	.353 (4.14e−17)	.282 (1.21e−4)
	.90	.372 (1.17e−9)	.275 (3.48e−5)	.450 (9.72e−19)	.398 (4.18e−19)	**.468** (2.65e−17)	.432 (1.08e−11)	.422 (9.15e−10)
Accuracy	.00	**.772**	**.772**	**.772**	**.772**	**.772**	**.772**	**.772**
	.10	.808	.810	.802	.811	**.811**	.811	.808
	.20	**.846**	.838	.831	.837	.839	.843	.825
	.30	**.876**	.853	.852	.848	.855	.865	.823
	.40	**.900**	.848	.863	.837	.850	.877	.815
	.50	**.892**	.826	.851	.816	.832	.867	.790
	.60	**.853**	.794	.824	.787	.801	.837	.770
	.70	**.814**	.759	.787	.750	.757	.797	.742
	.80	**.771**	.720	.746	.716	.719	.752	.709
	.90	**.732**	.677	.705	.671	.676	.712	.677
	1.00	**.689**	.637	.666	.629	.634	.671	.640

On the other hand, when looking at `galaxyzoo`, we see statistically significant estimates for the fake cutoff placebo estimates. This means we should take with a grain of salt our estimated $\hat{\tau}_{\texttt{RD}}$, as assumptions could be violated.

We can exploit other tests, such as the one detailed in the next subsection, to understand why the placebo tests fail.

D.3.2 Density Estimation

Density estimation test setup. We consider the same setup presented in Section 4.1, and we validate $\hat{\tau}_{\texttt{RD}}$ estimates by estimating the empirical density of reject scores. This is done separately for instances below the cutoff and above the cutoff, using the `R` package `rddensity` Cattaneo et al. (2022) with default parameters. This allows us to: (i) statistically assess the continuity of estimated reject score densities around the cutoff using a permutation test (null hypothesis stands for continuity at the cutoff); (ii) visualize whether there is a discontinuity in the estimated reject score densities around the cutoffs.

Results Figures 6-10 provide the density estimation plots and the permutation tests p-values (high values mean we can't falsify the assumption) for the best baseline on all the datasets. If the running variable is not manipulable, we would expect to see no difference in the estimated densities from both sides of the cutoff.

We can see that for the majority of cutoffs, the estimated densities are close, thus not falsifying Assumption 1. The only exception is `galaxyzoo` (Figure 8), where most tests reject the null hypothesis. The main reason for this behavior is that the reject score density peaks around one value, with little variation in the reject scores. Accordingly, small changes in the reject score imply abrupt changes in predictive accuracy. This sheds light on a potential criticality of the deferring system, i.e., the reject score estimation. Moreover, this behavior also explains why most placebo cutoff tests failed for `galaxyzoo`.

Regarding `synth`, the null hypothesis is rejected only for $c \geq .70$. Once again, this is because the reject scores peak at -1 and 1, potentially harming the quality of the estimates. We see similar trends for `cifar10h` (Figure 7), where the p-values are significant only at $c = .10$, and `xray-airspace`, where values are significant for $c > .70$

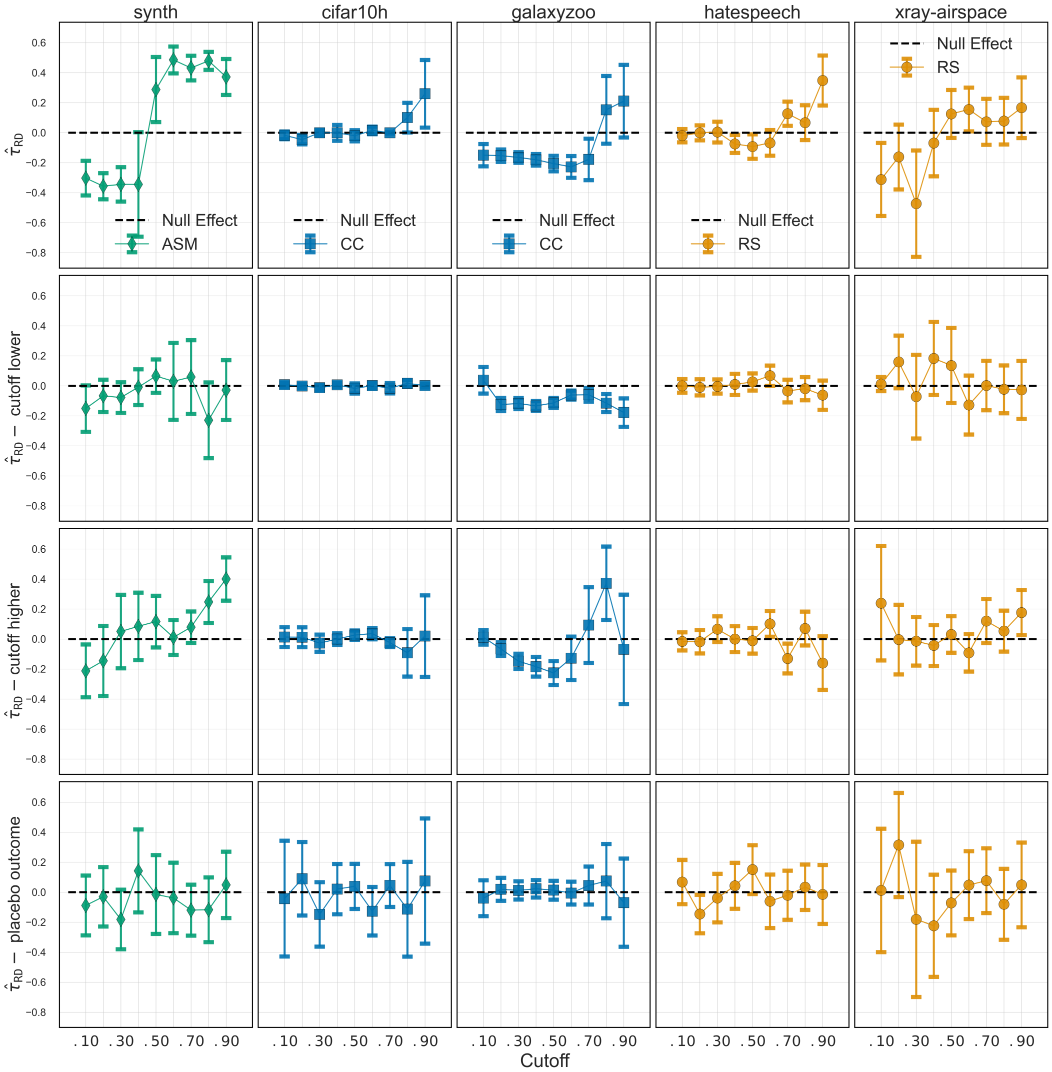


Figure 5: Estimated $\hat{\tau}_{\texttt{RD}}$ for the best baselines over the five datasets. The first row provides the estimates of $\hat{\tau}_{\texttt{RD}}$. The second row provides the estimates for the lower cutoff placebo test. The third row provides the estimates for the higher cutoff placebo test. The fourth row provides the estimates for the placebo outcome test.

Table 4: `cifar10h` results, with the statistically significant ones at $\alpha = 0.05$ in blue.

	c	ASM	CC	DT	LCE	OVA	RS	SP
$\hat{\tau}_{\texttt{ATD}}$	.00	.028 (4.53e−5)	.018 (5.34e−3)	**.049** (1.38e−10)	.018 (4.83e−3)	.021 (2.39e−3)	.018 (5.34e−3)	.021 (1.36e−3)
	.10	.033 (2.13e−5)	.027 (1.85e−4)	**.038** (4.61e−7)	.021 (2.80e−3)	.024 (1.35e−3)	.021 (4.10e−3)	.024 (1.01e−3)
	.20	**.039** (6.41e−6)	.035 (1.15e−5)	.035 (1.53e−6)	.027 (6.12e−4)	.027 (1.00e−3)	.025 (2.36e−3)	.029 (5.41e−4)
	.30	**.048** (7.00e−7)	.046 (1.08e−7)	.016 (1.79e−2)	.033 (2.18e−4)	.035 (1.98e−4)	.032 (5.10e−4)	.037 (6.72e−5)
	.40	**.062** (3.25e−8)	.057 (5.69e−9)	.007 (2.58e−1)	.041 (7.50e−5)	.045 (3.05e−5)	.041 (1.25e−4)	.043 (4.53e−5)
	.50	**.086** (4.56e−11)	.073 (2.26e−10)	−.004 (4.50e−1)	.059 (1.43e−6)	.056 (1.31e−5)	.051 (2.55e−5)	.058 (3.50e−6)
	.60	**.108** (2.62e−12)	.095 (1.12e−11)	−.003 (6.38e−1)	.073 (7.50e−7)	.079 (5.51e−7)	.080 (6.06e−8)	.079 (5.44e−8)
	.70	**.132** (1.64e−13)	.127 (1.20e−12)	−.007 (1.57e−1)	.098 (2.91e−8)	.116 (7.68e−9)	.120 (1.18e−10)	.122 (8.34e−11)
	.80	.178 (1.71e−14)	.200 (7.25e−14)	−.005 (3.17e−1)	.126 (2.25e−7)	.167 (2.71e−10)	.189 (4.93e−12)	**.207** (7.78e−15)
	.90	.306 (2.38e−14)	.277 (1.53e−11)	−.005 (3.17e−1)	.193 (2.14e−7)	.254 (9.42e−11)	.286 (2.39e−12)	**.320** (2.35e−13)
$\hat{\tau}_{\texttt{RD}}$	.10	−.843 (9.54e−2)	−.017 (1.89e−1)	**.114** (1.90e−1)	−.029 (6.86e−2)	−.005 (5.60e−1)	−.017 (1.69e−3)	−.015 (8.08e−3)
	.20	−.022 (2.12e−1)	−.046 (3.61e−3)	**.122** (2.26e−2)	−.011 (4.58e−1)	−.027 (2.90e−2)	−.022 (9.53e−4)	−.025 (1.27e−4)
	.30	−.043 (1.73e−5)	−.000 (9.92e−1)	**.213** (4.70e−4)	−.005 (7.28e−1)	−.039 (8.70e−2)	−.029 (3.00e−4)	−.024 (3.51e−4)
	.40	−.039 (1.65e−3)	−.001 (9.75e−1)	**.054** (3.26e−1)	−.016 (4.46e−1)	−.009 (5.19e−1)	−.038 (2.60e−4)	−.037 (1.01e−4)
	.50	.006 (8.04e−1)	−.020 (2.94e−1)	**.063** (1.73e−1)	−.026 (1.85e−1)	−.020 (3.68e−1)	−.052 (3.28e−4)	−.048 (6.66e−4)
	.60	−.146 (2.84e−2)	**.016** (2.30e−1)	−.037 (3.74e−1)	−.019 (6.48e−1)	−.008 (8.33e−1)	−.048 (5.91e−3)	−.052 (7.96e−3)
	.70	.070 (6.07e−1)	−.001 (7.48e−1)	−.023 (5.67e−1)	.052 (3.03e−1)	**.080** (1.94e−1)	−.055 (4.83e−2)	−.059 (4.96e−1)
	.80	−.035 (6.22e−1)	**.101** (4.51e−2)	−.041 (2.44e−1)	−.014 (8.57e−1)	−.027 (7.02e−1)	.057 (5.20e−1)	.000 (9.99e−1)
	.90	.102 (5.04e−1)	**.260** (2.37e−2)	−.030 (2.35e−1)	.175 (1.89e−1)	.143 (2.15e−1)	.059 (7.25e−1)	.142 (5.07e−1)
Accuracy	.00	**.958**	**.958**	**.958**	**.958**	**.958**	**.958**	**.958**
	.10	.959	**.963**	.943	.959	.959	.958	.958
	.20	.960	**.967**	.936	.961	.959	.959	.959
	.30	.963	**.971**	.919	.963	.962	.962	.962
	.40	.966	**.973**	.913	.964	.964	.964	.963
	.50	.971	**.975**	.907	.968	.965	.966	.966
	.60	.971	**.977**	.908	.967	.968	.972	.970
	.70	.970	**.978**	.907	.968	.970	.976	.974
	.80	.967	**.978**	.908	.962	.970	.975	.977
	.90	.958	.965	.908	.955	.962	**.967**	.964
	1.00	.929	**.939**	.909	**.939**	.937	**.939**	.936

and around the accuracy maximizing cutoff value, i.e., $.\overline{\kappa}_{.40}$ and $\overline{\kappa}_{.50}$.

Regarding `hatespeech`, we see no statistically significant p-values, suggesting the validity of $\hat{\tau}_{\texttt{RD}}$ across all cutoff values (see Figure 9).

Table 5: `galaxyzoo` results, with the statistically significant ones at $\alpha = 0.05$ in blue.

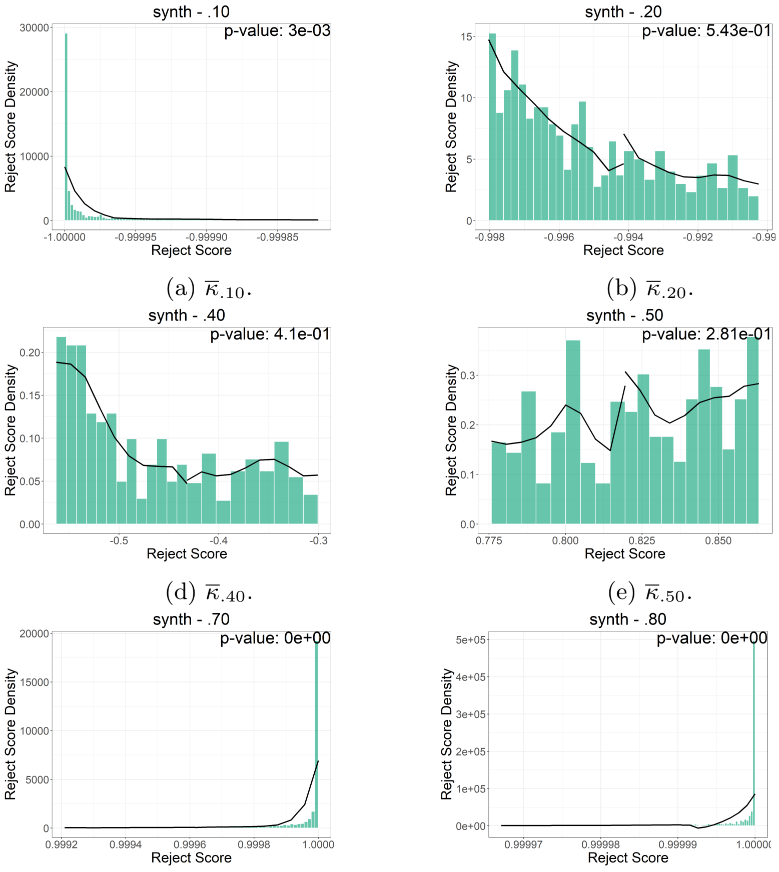
	c	ASM	CC	DT	LCE	OVA	RS	SP
$\hat{\tau}_{\text{ATD}}$	.00	−.103 (5.85e−18)	−.100 (2.93e−17)	−.093 (1.11e−14)	−**.082** (3.00e−11)	−.098 (1.86e−16)	−.099 (6.51e−17)	−.097 (2.76e−15)
	.10	−.088 (5.56e−12)	−.093 (1.78e−13)	−.089 (1.29e−12)	−**.081** (3.57e−10)	−.086 (3.04e−11)	−.098 (4.10e−14)	−.096 (2.76e−15)
	.20	−.076 (1.33e−8)	−.082 (8.01e−10)	−.094 (1.74e−12)	−.074 (6.20e−8)	−**.072** (2.08e−7)	−.092 (1.15e−10)	−.096 (2.76e−15)
	.30	−.069 (1.00e−6)	−.067 (5.55e−6)	−.084 (1.63e−9)	−.072 (6.48e−7)	−**.060** (9.11e−5)	−.078 (8.63e−7)	−.096 (2.76e−15)
	.40	−.063 (5.94e−5)	−.060 (2.76e−4)	−.082 (2.45e−8)	−**.051** (1.23e−3)	−.057 (5.83e−4)	−.072 (2.50e−5)	−.056 (2.03e−3)
	.50	−.060 (4.07e−4)	−.045 (2.18e−2)	−.090 (1.11e−8)	−.039 (3.18e−2)	−.046 (1.61e−2)	−.039 (4.31e−2)	−**.039** (6.10e−2)
	.60	−.051 (7.92e−3)	−.013 (5.68e−1)	−.100 (4.30e−8)	−**.006** (7.61e−1)	−.025 (2.54e−1)	−.006 (7.74e−1)	−.006 (7.88e−1)
	.70	−.036 (1.16e−1)	.009 (7.49e−1)	−.111 (7.80e−8)	.013 (5.85e−1)	−.014 (5.86e−1)	.014 (6.04e−1)	**.019** (5.06e−1)
	.80	.017 (5.84e−1)	.068 (4.66e−2)	−.105 (1.61e−5)	.028 (3.64e−1)	.007 (8.12e−1)	.060 (7.60e−2)	**.079** (2.79e−2)
	.90	.072 (1.23e−1)	.096 (3.52e−2)	−.119 (5.32e−6)	.102 (1.88e−2)	.024 (6.01e−1)	.089 (7.55e−2)	**.105** (3.11e−2)
$\hat{\tau}_{\text{RD}}$	.10	−.177 (2.37e−1)	−.150 (7.54e−5)	**.496** (2.64e−2)	−.192 (7.48e−2)	−.262 (1.68e−19)	−.162 (1.41e−14)	−−
	.20	−.216 (1.18e−1)	−.153 (3.66e−13)	**.084** (6.61e−1)	−.193 (3.60e−3)	−.110 (1.84e−1)	−.277 (7.73e−26)	−−
	.30	−.187 (3.00e−1)	−.165 (2.90e−20)	−.511 (1.05e−1)	−**.086** (2.47e−1)	−.152 (2.89e−2)	−.277 (2.63e−28)	−−
	.40	−.040 (7.16e−1)	−.180 (3.18e−20)	**.097** (4.50e−1)	−.082 (8.66e−2)	−.180 (4.49e−2)	−.215 (1.02e−2)	−−
	.50	−.050 (5.46e−1)	−.206 (6.27e−15)	−.023 (8.38e−1)	−.261 (1.17e−5)	**.060** (5.84e−1)	−.250 (4.18e−4)	−.220 (1.06e−1)
	.60	**.012** (8.70e−1)	−.228 (6.41e−10)	−.067 (5.97e−1)	−.106 (5.36e−2)	−.358 (8.70e−2)	−.287 (2.55e−2)	−.290 (2.44e−2)
	.70	−.123 (1.66e−2)	−.178 (1.16e−2)	−.312 (1.24e−2)	−**.018** (7.69e−1)	−.381 (2.83e−2)	−.152 (5.62e−1)	−.036 (8.97e−1)
	.80	−.013 (8.67e−1)	.153 (1.84e−1)	−−	−.089 (2.41e−1)	−.295 (7.49e−2)	**.213** (3.18e−1)	−.353 (1.44e−1)
	.90	−.066 (7.63e−1)	**.211** (8.75e−2)	−−	.158 (8.07e−2)	−.010 (9.59e−1)	−−	−.184 (6.52e−1)
Accuracy	.00	**.743**	**.743**	**.743**	**.743**	**.743**	**.743**	**.743**
	.10	**.767**	.760	.755	.751	.764	.753	.743
	.20	**.784**	.777	.761	.764	.783	.768	.743
	.30	.796	.796	.776	.772	**.800**	.788	.743
	.40	**.808**	.806	.786	.793	.807	.797	.806
	.50	.815	.821	.791	.805	.819	**.822**	.821
	.60	.825	.838	.797	.822	.831	**.839**	.837
	.70	.834	**.845**	.803	.829	.837	**.845**	.845
	.80	.849	**.856**	.816	.830	.842	.853	.854
	.90	**.853**	**.853**	.817	.834	.843	.849	.850
	1.00	**.846**	.843	.836	.825	.841	.841	.839

Table 6: `hatespeech` results, with the statistically significant ones at $\alpha = 0.05$ in blue.

	c	ASM	CC	DT	LCE	OVA	RS	SP
$\hat{\tau}_{\text{ATD}}$	.00	.067 (1.09e−26)	.033 (8.77e−9)	**.097** (1.38e−49)	.037 (2.13e−10)	.031 (8.31e−8)	.031 (9.51e−8)	.030 (1.67e−7)
	.10	.078 (7.62e−30)	.037 (1.57e−9)	**.089** (3.99e−41)	.042 (4.38e−11)	.038 (1.80e−9)	.038 (1.26e−9)	.036 (1.37e−8)
	.20	**.092** (9.70e−35)	.047 (1.06e−12)	.078 (3.03e−30)	.047 (9.93e−12)	.044 (1.26e−10)	.046 (3.65e−11)	.047 (3.23e−11)
	.30	**.107** (6.96e−38)	.058 (7.32e−16)	.074 (3.63e−26)	.060 (1.25e−15)	.059 (2.97e−15)	.054 (1.17e−12)	.058 (1.00e−13)
	.40	**.129** (4.28e−45)	.071 (6.29e−18)	.062 (7.28e−18)	.076 (2.94e−20)	.075 (2.33e−19)	.070 (7.55e−16)	.076 (6.24e−18)
	.50	**.159** (2.89e−51)	.088 (3.59e−21)	.052 (7.78e−12)	.094 (2.03e−23)	.094 (5.16e−22)	.091 (1.45e−20)	.096 (8.19e−21)
	.60	**.202** (5.31e−59)	.117 (8.60e−25)	.055 (5.01e−11)	.111 (1.72e−24)	.113 (3.40e−23)	.124 (3.03e−26)	.117 (3.30e−22)
	.70	**.281** (9.86e−73)	.160 (1.29e−29)	.052 (1.76e−8)	.134 (4.30e−24)	.138 (1.86e−23)	.172 (3.85e−34)	.165 (8.30e−30)
	.80	**.359** (1.85e−73)	.212 (5.52e−33)	.049 (1.37e−5)	.185 (4.23e−27)	.165 (1.38e−20)	.232 (4.13e−35)	.201 (1.64e−26)
	.90	**.468** (1.71e−63)	.318 (2.10e−33)	.034 (1.78e−2)	.222 (5.48e−20)	.213 (8.30e−17)	.295 (6.56e−28)	.255 (1.15e−19)
$\hat{\tau}_{\text{RD}}$	.10	−.129 (2.04e−2)	−.037 (3.12e−1)	**.122** (9.14e−2)	−.044 (1.64e−1)	.068 (2.39e−1)	−.020 (3.79e−1)	−.026 (2.93e−1)
	.20	−.020 (5.38e−1)	−.063 (1.27e−2)	**.123** (2.79e−2)	−.041 (1.23e−1)	.001 (9.90e−1)	−.000 (9.89e−1)	−.046 (2.71e−2)
	.30	.014 (7.85e−1)	.008 (6.90e−1)	**.057** (2.78e−1)	−.061 (2.66e−2)	−.035 (3.35e−1)	.005 (8.97e−1)	−.049 (6.71e−2)
	.40	−.006 (8.85e−1)	−.031 (4.91e−2)	**.121** (2.29e−3)	−.057 (2.34e−2)	−.021 (6.03e−1)	−.075 (1.45e−2)	−.064 (2.53e−2)
	.50	−.038 (2.78e−1)	.003 (8.80e−1)	**.133** (7.73e−3)	.024 (3.98e−1)	−.004 (9.19e−1)	−.092 (2.70e−2)	−.037 (3.80e−1)
	.60	**.044** (3.19e−1)	−.035 (7.25e−2)	−.041 (4.18e−1)	.026 (3.91e−1)	−.021 (6.29e−1)	−.068 (1.20e−1)	−.006 (9.15e−1)
	.70	−.113 (1.66e−1)	.071 (8.66e−2)	.098 (7.61e−3)	−.009 (8.13e−1)	.125 (2.68e−3)	**.127** (2.08e−3)	−.045 (4.37e−1)
	.80	**.151** (4.38e−2)	.035 (5.19e−1)	−.030 (5.77e−1)	.097 (1.18e−2)	.028 (6.77e−1)	.067 (2.58e−1)	.066 (4.38e−1)
	.90	.315 (8.58e−5)	.218 (2.66e−2)	.054 (3.27e−1)	.267 (7.86e−5)	.153 (1.32e−1)	**.348** (4.10e−5)	.112 (2.68e−1)
Accuracy	.00	**.908**	**.908**	**.908**	**.908**	**.908**	**.908**	**.908**
	.10	.911	.908	.893	.909	.911	**.912**	.910
	.20	.914	.912	.874	.908	.912	.914	**.915**
	.30	.915	.915	.863	.913	**.918**	.916	**.918**
	.40	.919	.916	.849	.918	.922	.920	**.924**
	.50	.921	.918	.837	.919	.923	.924	**.926**
	.60	.922	.920	.833	.917	.922	**.928**	.925
	.70	.922	.921	.827	.912	.918	**.930**	.928
	.80	.910	.919	.821	.908	.910	**.923**	.919
	.90	.883	.907	.815	.894	.899	**.908**	.904
	1.00	.841	.875	.812	.871	.877	.878	**.878**

Table 7: `xray-airspace` results, with the statistically significant ones at $\alpha = 0.05$ in blue.

	c	ASM	CC	DT	LCE	OVA	RS	SP
$\hat{\tau}_{\texttt{ATD}}$	.00	**.359** (2.27e−59)	.029 (6.71e−2)	.189 (1.46e−21)	.033 (4.08e−2)	.033 (3.76e−2)	.039 (1.78e−2)	.023 (1.31e−1)
	.10	**.416** (5.20e−70)	.046 (1.18e−2)	.196 (2.76e−22)	.053 (3.79e−3)	.055 (2.26e−3)	.059 (1.30e−3)	.034 (5.55e−2)
	.20	**.482** (1.73e−86)	.063 (1.45e−3)	.199 (6.39e−22)	.068 (5.38e−4)	.072 (1.96e−4)	.080 (7.05e−5)	.038 (4.97e−2)
	.30	**.585** (2.91e−115)	.079 (3.09e−4)	.209 (2.64e−22)	.085 (5.32e−5)	.082 (3.32e−5)	.091 (1.32e−5)	.043 (4.27e−2)
	.40	**.673** (3.57e−144)	.092 (1.08e−4)	.239 (4.79e−26)	.092 (5.99e−5)	.108 (1.08e−7)	.123 (5.10e−9)	.068 (4.51e−3)
	.50	**.730** (9.64e−161)	.104 (7.99e−5)	.242 (6.45e−23)	.106 (7.34e−5)	.115 (1.56e−7)	.140 (1.82e−9)	.085 (2.03e−3)
	.60	**.800** (5.17e−193)	.125 (3.50e−5)	.254 (5.80e−20)	.120 (5.09e−5)	.159 (5.70e−11)	.149 (2.12e−8)	.130 (2.40e−5)
	.70	**.858** (1.58e−202)	.183 (1.90e−8)	.191 (1.97e−11)	.160 (7.21e−6)	.173 (8.75e−10)	.149 (1.24e−6)	.158 (7.87e−6)
	.80	**.887** (1.93e−146)	.271 (5.94e−11)	.032 (1.00e−1)	.203 (2.04e−6)	.181 (1.99e−8)	.209 (3.82e−9)	.218 (3.98e−7)
	.90	**.887** (1.93e−146)	.352 (2.28e−9)	.033 (1.54e−1)	.320 (1.61e−8)	.194 (1.30e−4)	.179 (2.16e−5)	.253 (7.28e−5)
$\hat{\tau}_{\texttt{RD}}$	.10	−.272 (1.29e−1)	−.215 (4.69e−1)	**.059** (7.38e−1)	−.190 (9.40e−2)	−.123 (4.68e−1)	−.312 (1.22e−2)	−.034 (3.36e−1)
	.20	−.558 (1.22e−2)	−.013 (7.70e−1)	**.109** (5.37e−1)	−.012 (9.34e−1)	−.162 (1.19e−1)	−.161 (1.43e−1)	.042 (6.52e−1)
	.30	.073 (6.71e−1)	−.058 (5.48e−1)	−.067 (6.38e−1)	.103 (1.90e−1)	**.726** (3.63e−4)	−.473 (8.95e−3)	−.273 (4.35e−3)
	.40	**.322** (1.03e−1)	.027 (7.69e−1)	−.022 (8.44e−1)	.092 (1.06e−1)	−.153 (2.13e−1)	−.069 (5.42e−1)	.055 (5.15e−1)
	.50	**.354** (1.38e−1)	.058 (6.20e−1)	.144 (2.38e−1)	.010 (8.83e−1)	.157 (1.16e−1)	.125 (1.25e−1)	−.018 (8.06e−1)
	.60	**.465** (2.12e−1)	.093 (4.63e−1)	.041 (7.67e−1)	.014 (8.47e−1)	.090 (2.72e−1)	.155 (3.70e−2)	−.171 (1.60e−1)
	.70	.357 (3.82e−1)	−.180 (1.35e−1)	**.638** (5.77e−5)	−.064 (4.29e−1)	−.044 (6.15e−1)	.073 (3.52e−1)	−.114 (5.64e−1)
	.80	−−	.137 (3.60e−1)	.072 (6.93e−1)	−.152 (3.21e−1)	.176 (3.33e−2)	.077 (3.28e−1)	**.278** (1.36e−1)
	.90	−−	.065 (7.91e−1)	−.222 (1.94e−1)	−.023 (9.01e−1)	.032 (8.44e−1)	**.167** (1.06e−1)	.008 (9.54e−1)
$\hat{\tau}_{\texttt{CATD}}$ − `Male`	.00	**.402** (2.24e−42)	.054 (1.48e−2)	.199 (2.14e−14)	.062 (5.22e−3)	.058 (7.55e−3)	.062 (5.22e−3)	.058 (7.55e−3)
	.10	**.451** (1.13e−47)	.063 (1.06e−2)	.208 (4.56e−15)	.070 (4.74e−3)	.070 (3.56e−3)	.072 (3.59e−3)	.066 (7.50e−3)
	.20	**.507** (1.25e−55)	.080 (2.85e−3)	.213 (2.99e−15)	.094 (3.78e−4)	.086 (9.02e−4)	.095 (3.77e−4)	.071 (8.96e−3)
	.30	**.594** (9.86e−72)	.094 (2.20e−3)	.227 (1.20e−15)	.109 (8.23e−5)	.090 (7.99e−4)	.100 (3.82e−4)	.078 (1.07e−2)
	.40	**.682** (2.53e−90)	.110 (1.24e−3)	.245 (3.65e−16)	.116 (1.40e−4)	.130 (2.07e−6)	.137 (1.85e−6)	.116 (8.76e−4)
	.50	**.719** (3.04e−92)	.134 (4.93e−4)	.249 (8.22e−15)	.125 (2.80e−4)	.130 (6.63e−6)	.150 (7.08e−7)	.130 (1.45e−3)
	.60	**.810** (3.08e−127)	.151 (7.98e−4)	.268 (8.49e−14)	.140 (4.69e−4)	.170 (4.09e−8)	.155 (9.19e−6)	.185 (4.44e−5)
	.70	**.853** (2.02e−119)	.234 (1.01e−6)	.176 (5.73e−7)	.188 (1.90e−4)	.186 (1.32e−7)	.180 (1.00e−5)	.227 (9.27e−6)
	.80	**.863** (4.02e−69)	.337 (1.23e−8)	.048 (4.21e−2)	.223 (2.90e−4)	.175 (4.30e−5)	.262 (5.84e−8)	.245 (4.46e−5)
	.90	**.863** (4.02e−69)	.436 (7.48e−8)	.049 (1.54e−1)	.407 (1.32e−7)	.156 (1.66e−2)	.261 (6.86e−5)	.318 (1.71e−4)
$\hat{\tau}_{\texttt{CATD}}$ − `Female`	.00	**.311** (4.87e−21)	.002 (9.16e−1)	.178 (3.92e−9)	−.000 (1.00)	.005 (8.29e−1)	.012 (6.04e−1)	−.015 (5.03e−1)
	.10	**.376** (5.51e−26)	.026 (3.59e−1)	.184 (2.37e−9)	.031 (2.45e−1)	.035 (1.91e−1)	.044 (1.17e−1)	−.003 (9.08e−1)
	.20	**.453** (6.39e−34)	.042 (1.51e−1)	.185 (4.81e−9)	.035 (2.25e−1)	.054 (6.16e−2)	.061 (4.54e−2)	.000 (1.00)
	.30	**.573** (3.46e−46)	.062 (4.76e−2)	.190 (5.24e−9)	.053 (9.47e−2)	.070 (1.41e−2)	.079 (1.09e−2)	.004 (9.00e−1)
	.40	**.662** (5.16e−57)	.073 (2.82e−2)	.232 (7.55e−12)	.059 (8.81e−2)	.078 (8.58e−3)	.106 (6.72e−4)	.017 (6.00e−1)
	.50	**.744** (3.95e−70)	.073 (4.15e−2)	.236 (3.07e−10)	.077 (6.82e−2)	.096 (4.35e−3)	.127 (5.12e−4)	.036 (3.17e−1)
	.60	**.787** (4.52e−71)	.099 (1.43e−2)	.239 (3.18e−8)	.094 (3.50e−2)	.143 (2.21e−4)	.140 (6.22e−4)	.071 (8.38e−2)
	.70	**.867** (1.86e−84)	.132 (2.66e−3)	.214 (8.23e−6)	.127 (1.17e−2)	.153 (1.03e−3)	.106 (2.22e−2)	.080 (9.19e−2)
	.80	**.929** (1.09e−118)	.198 (5.07e−4)	.000 (1.00)	.180 (2.24e−3)	.188 (1.35e−4)	.145 (5.04e−3)	.184 (2.64e−3)
	.90	**.929** (1.09e−118)	.260 (2.16e−3)	−−	.217 (7.76e−3)	.233 (2.95e−3)	.079 (7.46e−2)	.179 (6.21e−2)
Accuracy	.00	**.871**	**.871**	**.871**	**.871**	**.871**	**.871**	**.871**
	.10	.879	.880	.864	.883	**.884**	.883	.877
	.20	**.892**	.889	.850	.889	.891	.891	.877
	.30	**.897**	.893	.838	.893	.891	.891	.877
	.40	.877	.894	.836	.890	.900	**.902**	.886
	.50	.842	.892	.814	.886	.892	**.896**	.887
	.60	.789	.890	.783	.885	**.897**	.885	**.897**
	.70	.726	**.900**	.730	.887	.887	.871	.892
	.80	.632	**.899**	.687	.884	.869	.870	.891
	.90	.632	**.885**	.685	.876	.852	.850	.872
	1.00	.512	.842	.682	.838	.838	.832	**.848**



(a) $\overline{\kappa}_{.10}$.

(b) $\overline{\kappa}_{.20}$.

(c) $\overline{\kappa}_{.30}$.

(d) $\overline{\kappa}_{.40}$.

(e) $\overline{\kappa}_{.50}$.

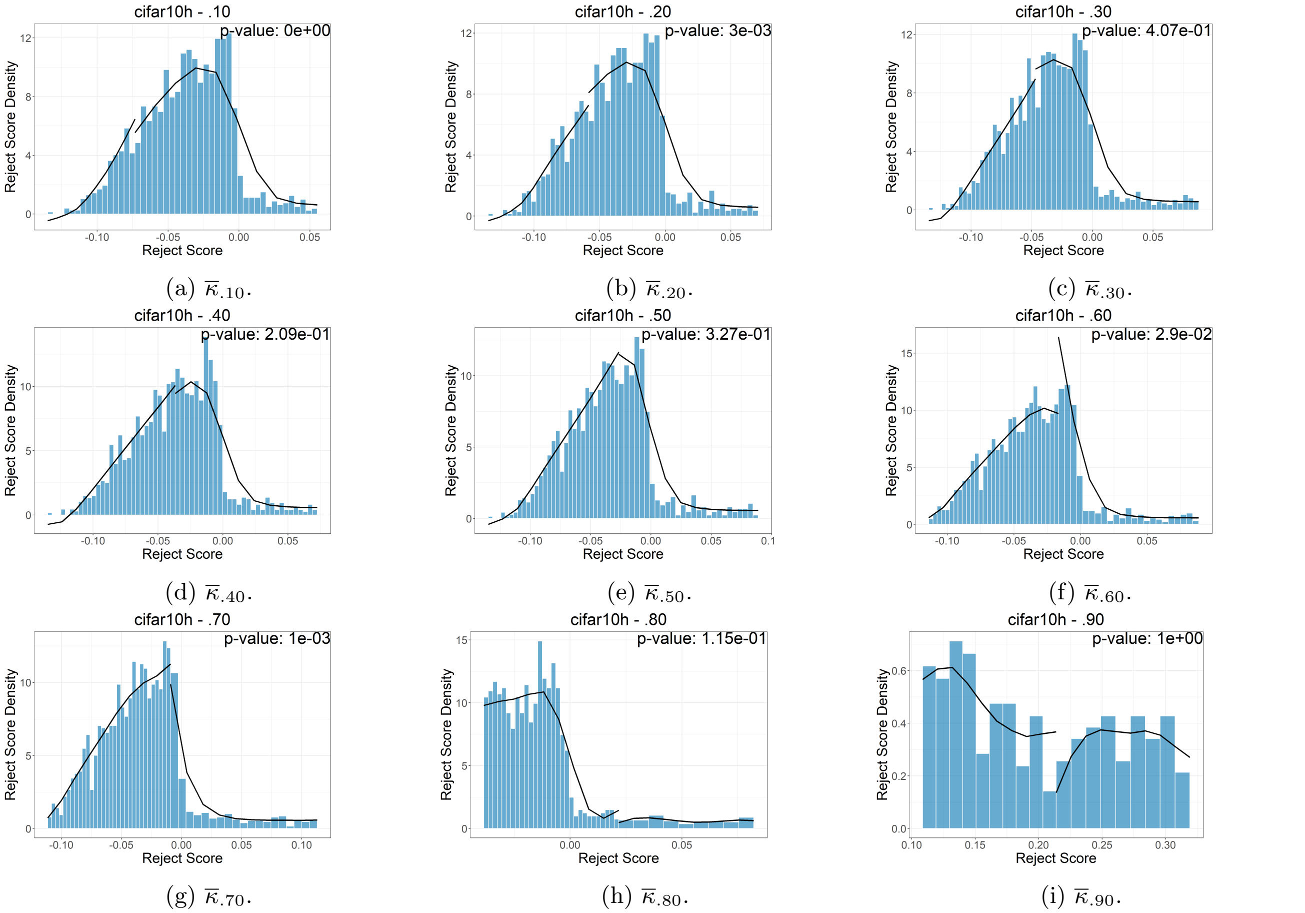
(f) $\overline{\kappa}_{.60}$.

(g) $\overline{\kappa}_{.70}$.

(h) $\overline{\kappa}_{.80}$.

(i) $\overline{\kappa}_{.90}$.

Figure 6: `synth` estimated best baseline (ASM) reject scores densities at the left and right of cutoff $\overline{\kappa}_c$. All the plots are zoomed around the cutoff values.



(a) $\overline{\kappa}_{.10}$.

(b) $\overline{\kappa}_{.20}$.

(c) $\overline{\kappa}_{.30}$.

(d) $\overline{\kappa}_{.40}$.

(e) $\overline{\kappa}_{.50}$.

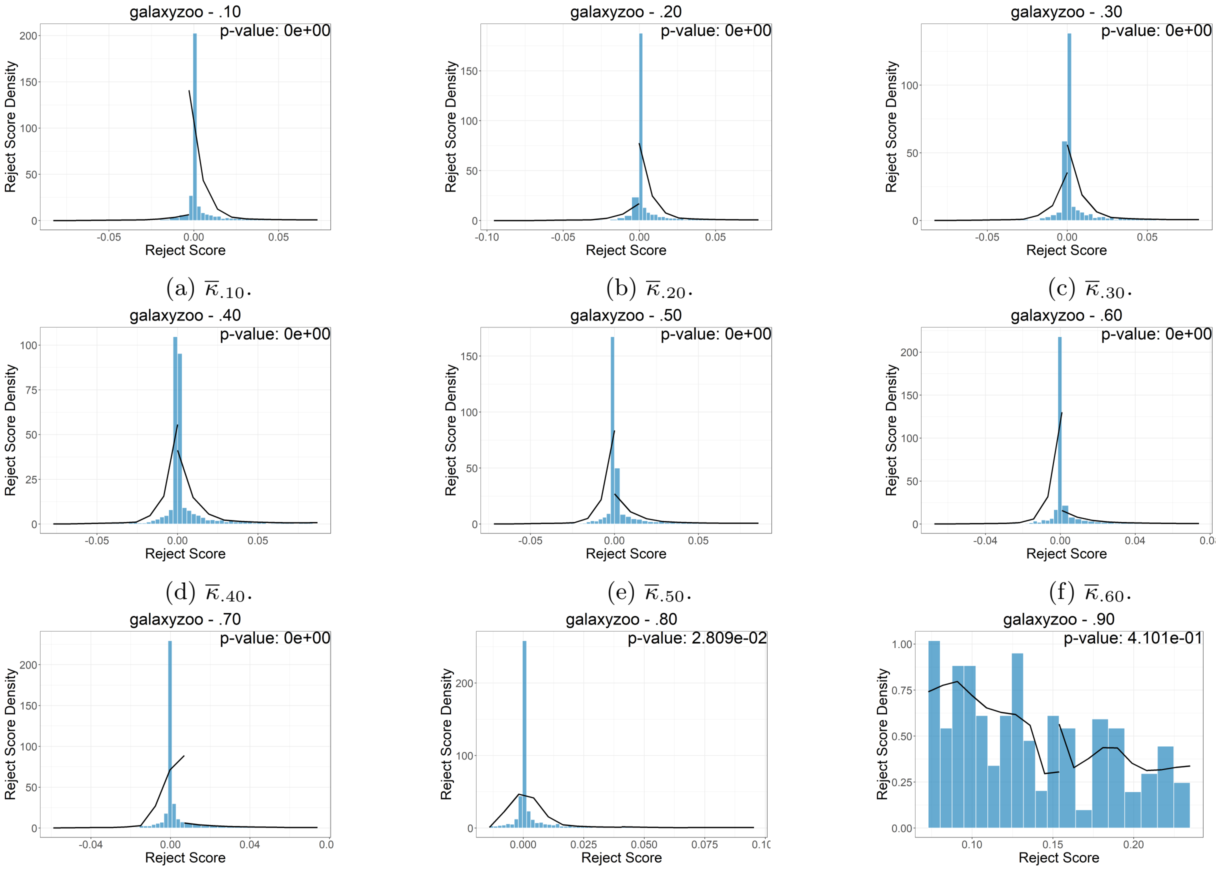
(f) $\overline{\kappa}_{.60}$.

(g) $\overline{\kappa}_{.70}$.

(h) $\overline{\kappa}_{.80}$.

(i) $\overline{\kappa}_{.90}$.

Figure 7: `cifar10h` estimated best baseline (CC) reject scores densities at the left and right of cutoff $\overline{\kappa}_c$. All the plots are zoomed around the cutoff values.



(a) $\overline{\kappa}_{.10}$.

(b) $\overline{\kappa}_{.20}$.

(c) $\overline{\kappa}_{.30}$.

(d) $\overline{\kappa}_{.40}$.

(e) $\overline{\kappa}_{.50}$.

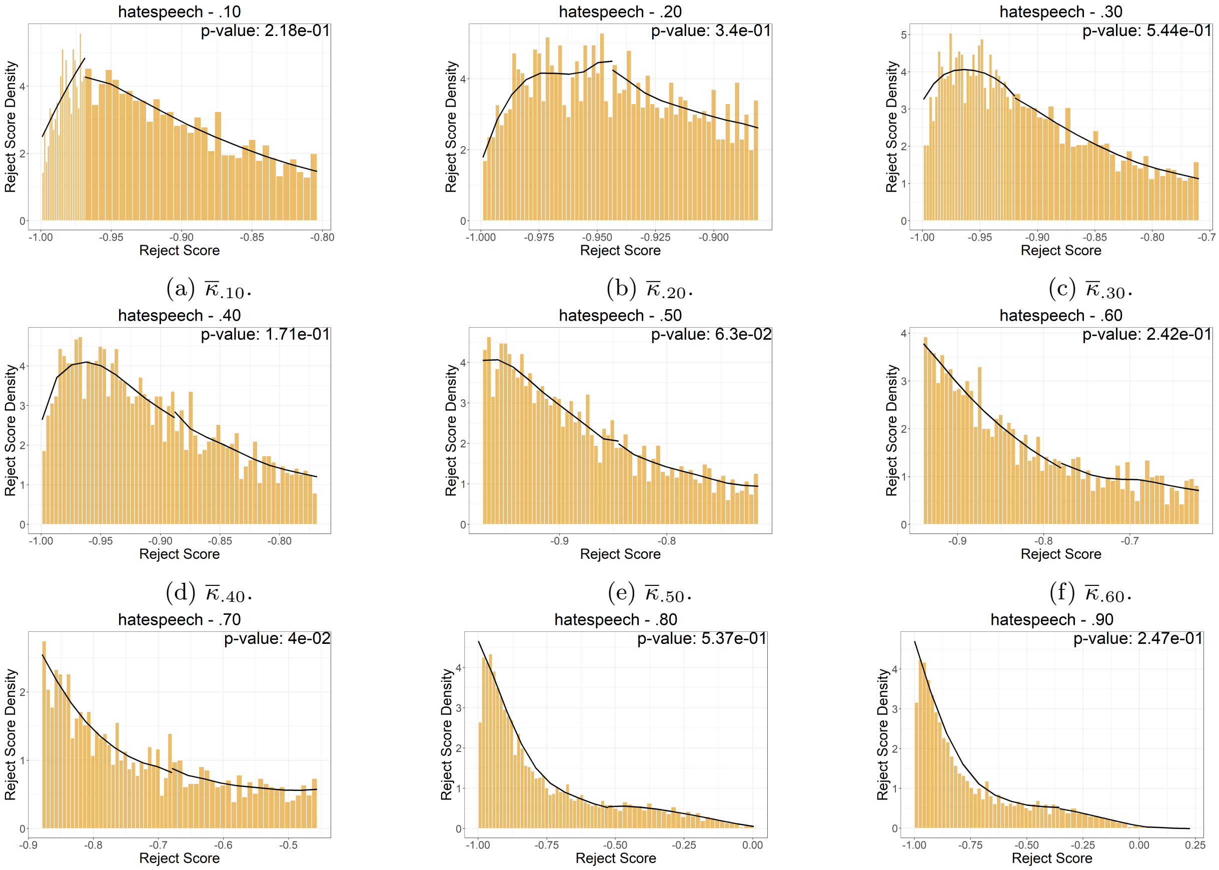
(f) $\overline{\kappa}_{.60}$.

(g) $\overline{\kappa}_{.70}$.

(h) $\overline{\kappa}_{.80}$.

(i) $\overline{\kappa}_{.90}$.

Figure 8: `galaxyzoo` estimated reject scores densities at the left and right of cutoff $\overline{\kappa}_c$ for the best baseline CC. All the plots are zoomed around the cutoff values.



(a) $\overline{\kappa}_{.10}$.

(b) $\overline{\kappa}_{.20}$.

(c) $\overline{\kappa}_{.30}$.

(d) $\overline{\kappa}_{.40}$.

(e) $\overline{\kappa}_{.50}$.

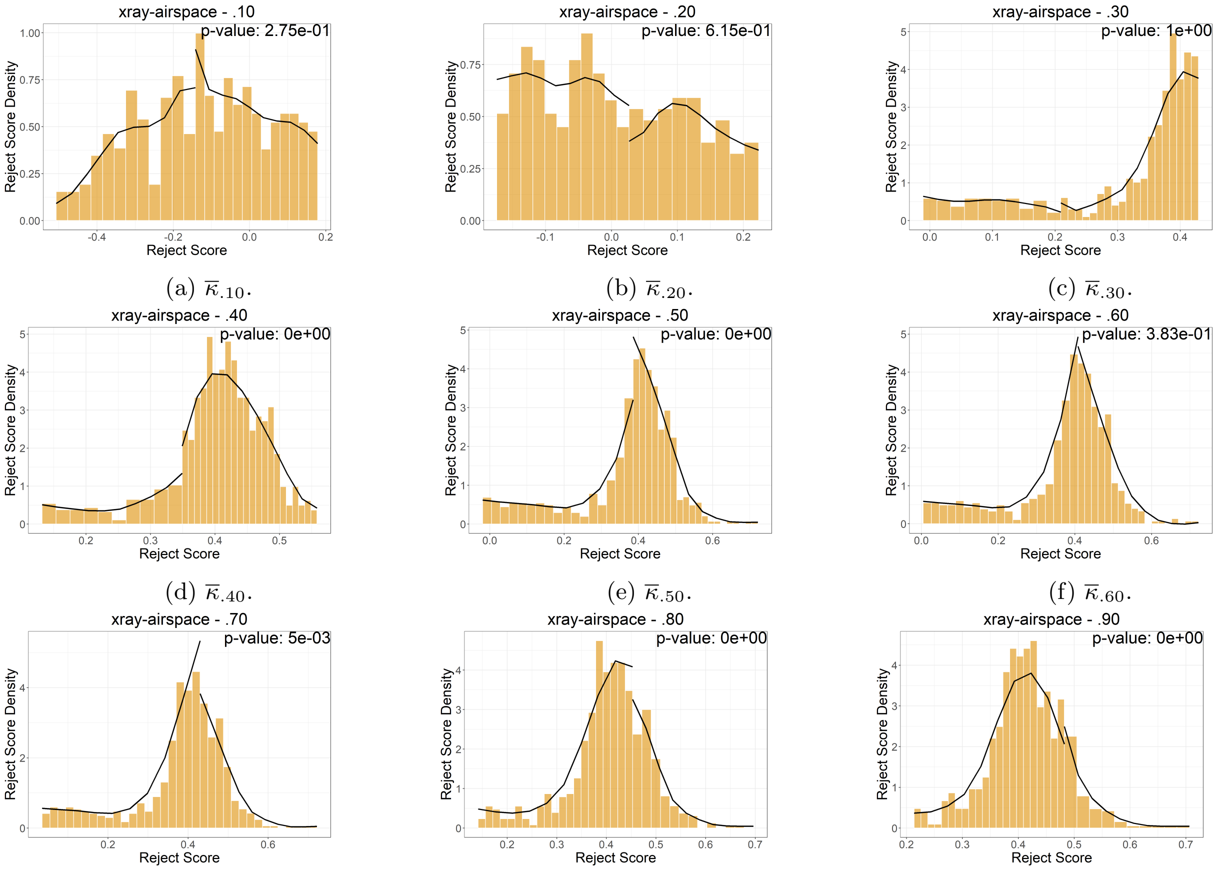
(f) $\overline{\kappa}_{.60}$.

(g) $\overline{\kappa}_{.70}$.

(h) $\overline{\kappa}_{.80}$.

(i) $\overline{\kappa}_{.90}$.

Figure 9: `hatespeech` estimated best baseline (RS) reject scores densities at the left and right of cutoff $\overline{\kappa}_c$. All the plots are zoomed around the cutoff values.



(a) $\overline{\kappa}_{.10}$.

(b) $\overline{\kappa}_{.20}$.

(c) $\overline{\kappa}_{.30}$.

(d) $\overline{\kappa}_{.40}$.

(e) $\overline{\kappa}_{.50}$.

(f) $\overline{\kappa}_{.60}$.

(g) $\overline{\kappa}_{.70}$.

(h) $\overline{\kappa}_{.80}$.

(i) $\overline{\kappa}_{.90}$.

Figure 10: `xray-airspace` estimated best baseline (RS) reject scores densities at the left and right of cutoff $\overline{\kappa}_c$. All the plots are zoomed around the cutoff values.