
A Causal Framework for Evaluating Deferring Systems

Filippo Palomba*^{ORCID}
Princeton University
fpalomba@princeton.edu

Andrea Pugnana*^{ORCID}
University of Pisa
andrea.pugnana@di.unipi.it

José M. Álvarez^{ORCID}
Scuola Normale Superiore, University of Pisa
jose.alvarez@sns.it

Salvatore Ruggieri^{ORCID}
University of Pisa
salvatore.ruggieri@unipi.it

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. This allows us to identify the causal impact of the deferring strategy on predictive accuracy. We distinguish two scenarios. In the first one, we can access both the human and the ML model predictions for the deferred instances. In such a case, we can identify the individual causal effects for deferred instances and aggregates of them. In the second scenario, only human predictions are available for the deferred instances. In this case, we can resort to regression discontinuity design to estimate a local causal effect. We empirically evaluate our approach on synthetic and real datasets for seven deferring systems from the literature.

*Equal contribution.

Contents

1	Introduction	3
2	Background and related work	3
2.1	Causal inference	3
2.2	Deferring systems	4
2.3	Related work	5
3	Evaluating deferring systems	6
3.1	Scenario 1: deferring systems as an almost perfect causal inference design	7
3.2	Scenario 2: deferring systems as an RD design	8
4	Experimental evaluation	8
4.1	Experimental settings	8
4.2	Experimental results	9
5	Conclusions	11
A	Extended Related Work	17
B	Proofs	18
C	Limitations and Extensions	18
D	Experimental evaluation	20
D.1	Additional details	20
D.1.1	Data	20
D.1.2	Baselines	20
D.1.3	Hyperparameters	21
D.1.4	Hardware and carbon footprint	21
D.2	Detailed results for Q1 , Q2 and Q3	21
D.3	How to validate estimates under Scenario 2	22
D.3.1	Placebo Tests	22
D.3.2	Density Estimation	25

1 Introduction

Machine Learning (ML) models are increasingly being used to support decision making in many high-stakes and socially sensitive domains. In such settings, wrong predictions can be harmful. To control for model mistakes, an extension of supervised learning allows ML models to abstain from providing a prediction and defer the prediction to a human expert. Such an extension is known as “learning to defer” (LtD) [1] or “learning under triage” [2]. The extended models, called *deferring systems*, aim at obtaining the best from the combination of AI and human expert predictions.

The LtD research field is blooming, with novel approaches continuously appearing (e.g., [3, 4, 5, 6, 7, 8]). However, little attention has been devoted to evaluating the impact of deferring. Most works evaluate deferring systems by looking at the final accuracy obtained by the human-AI team. Although suitable for ML practitioners, this accuracy-based view on evaluation is narrow. Policy-makers are interested in understanding the *causal effect* of introducing a deferring system within a high-stake decision-making process [9].

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 the diagnosis of a disease, meaning that part of the cases will still be handled by medical doctors, but another part will be predicted by an ML model. 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 demand for a causal inference approach [10], where the goal is to estimate the causal effect of a variable on another one [11].

Causal inference has a long tradition within policy evaluation [12], and we build on it for evaluating deferring systems. We link deferring systems with the causal inference framework of *potential outcomes* [13, 14] 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. Example *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, we can access 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 [15] 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. Example *Ex2* belongs to such a scenario.

The contributions of this paper are: (*i*) we frame the evaluation of deferring systems as a causal inference problem using the potential outcome framework; (*ii*) we investigate two scenarios, and for each, we show which causal effects can be identified and estimated; and (*iii*) we evaluate the proposed approach on five datasets, including a synthetic one and four real-world ones, and on seven deferring systems from the literature. The paper first introduces causal inference and deferring systems in Section 2. The two frameworks are bridged in Section 3 and two scenarios of causal estimation are investigated. We report experiments in Section 4. Finally, we summarize and conclude.

2 Background and related work

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 [13, 14] 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” [17]. 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:

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

known as the *average treatment effect* (ATE) and the *average treatment effect on the treated* (ATT), respectively². 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 [12] for a recent review).

The RD design [15] 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 assignment to treatment 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:

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

This quantity can be interpreted as a version of τ_{ATT} “local” at the cutoff. The above heuristics was formalized by Hahn et al. [18] 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:*

$$\lim_{v \rightarrow \xi} \mathbb{E}[O(0) \mid V = v] \quad \text{and} \quad \lim_{v \rightarrow \xi} \mathbb{E}[O(1) \mid V = v].$$

In words, 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, τ_{RD} can be identified from the data, as the next theorem shows.

Theorem 1 (Theorem 3 from [18]). *Let Assumption 1 hold. Then:*

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

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 $l : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ 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 an ML model (a.k.a. a predictor).

¹We make the *stable unit treatment value assumption* (SUTVA) [16]. It requires each unit’s potential outcome not to depend on the treatment assignment of other units, ruling out interference between units.

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

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, one can extend the canonical setting presented above by allowing the ML model to defer difficult cases to another predictor. Here, we consider a human expert as another predictor $h : \mathcal{Z} \rightarrow \mathcal{Y}$, where \mathcal{Z} is possibly a higher dimensional space than \mathcal{X} . To keep the notation simple, we will 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 $g : \mathcal{X} \rightarrow \{0, 1\}$. We define the *deferring system* ϑ , i.e. *the human-AI team*, as a triplet (f, g, h) , such that:

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

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 [6, 19]. 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:

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

with l_{ML} (resp., l_H) referring to the loss associated with the ML model (resp., human expert). The goal then 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 fraction of instances for which the ML model is selected to make predictions.

Most methods design the deferring strategy through a *reject score* function $k : \mathcal{X} \rightarrow \mathbb{R}$, which estimates whether the human expert prediction is more likely to be correct than the one of the ML model [7]. 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 $\bar{\kappa}$ over $k(\mathbf{x})$ to define the policy function as $g(\mathbf{x}) = \mathbb{1}\{k(\mathbf{x}) \geq \bar{\kappa}\}$. Okati et al. [2] 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 $\bar{\kappa}$ that maximizes accuracy over a validation set [7]. Otherwise, one can consider a *coverage-calibration* procedure by setting $\bar{\kappa}$ as the c^{th} -percentile of the reject score values over a validation set [20], as shown in Figure 1. In order to highlight the relationship between $\bar{\kappa}$ and c , we denote the estimated threshold for a target coverage c as $\bar{\kappa}_c$, i.e., $\bar{\kappa}_c$ is such that $\mathbb{E}[\mathbb{1}\{k(\mathbf{X}) \geq \bar{\kappa}_c\}] = (1 - c)$.

2.3 Related work

By viewing the introduction of human-AI teams as an intervention in a ML-based or in 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 due to Choe et al. [21], who estimate the accuracy of abstaining classifiers (which do not account for deferring to human experts) on the abstained instances under the assumption that the abstention policy is stochastic. Such an assumption is impractical in the context of deferring systems. See Appendix A for additional related work.

Policy evaluation. Causal inference methods are commonly used to evaluate the effects of treatments/policies and, thus, inform policymakers [12]. Randomized control trials (RCT), i.e., experiments that assign units to treatment randomly, are the gold standard for inferring average causal effects (such as τ_{ATE}) in many fields including healthcare [22], education [23], and finance [24]. When it is not possible to rely on RCTs, e.g., it is not ethical or too costly to run an experiment, one can resort to other techniques using observational data [11, 25]. The RD design has been used to assess the effectiveness of treatments in several fields, such as healthcare [26, 27, 28], criminal behavior [29], education [30, 31, 32, 33], public economics [34, 35, 36], and corporate finance [37].

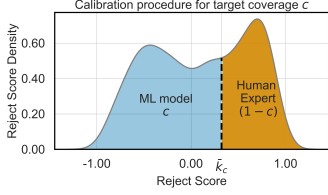


Figure 1: In blue, the $(c)\%$ of instances (i.e., those with $k(\mathbf{x}) < \bar{k}_c$) assigned to the ML model; in orange, the $(1 - c)\%$ instances (i.e., those with $k(\mathbf{x}) \geq \bar{k}_c$) the human predicts.

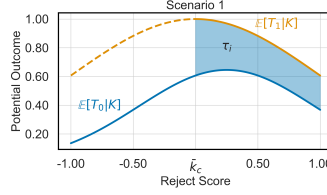


Figure 2: Scenario 1 assumptions: dashed lines are unobserved values and thick lines observed ones. The coloured area represents where the effects can be estimated (i.e., $k(\mathbf{x}) \geq \bar{k}_c$).

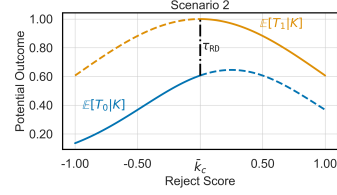


Figure 3: Scenario 2 assumptions: dashed lines are unobserved values and thick lines observed ones. We can estimate τ_{RD} at the cutoff value (i.e., $k(\mathbf{x}) = \bar{k}_c$).

Table 1: Deferring systems under the Potential Outcome lens.

	Potential Outcome	Deferring Systems
<i>running variable</i>	V_i	K_i
<i>cutoff</i>	ξ	\bar{k}_c
<i>treatment</i>	D_i	G_i
<i>outcome</i>	O_i	T_i
<i>potential outcomes</i>	$O_i(d), d \in \{0, 1\}$	$T_i(g), g \in \{0, 1\}$
τ_i	$O_i(1) - O_i(0)$	$T_i(1) - T_i(0)$

Deferring systems applications. In recent years, deferring systems have been deployed to build human-AI teams in different settings. For instance, Van der Pias et al. [38] present a deferring system for sleep stage scoring, which can be used to allow physicians to focus on critical patients. Cianci et al. [39] adopt selective classification [40] for uncertainty self-assessment of a credit scoring ML model, with the purpose of informing the human decision maker. Bondi et al. [41] 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. [42] 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 accuracy, of deferring the prediction to a human expert in a given deferring system based on a reject score function. For this, 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 $P(\mathbf{X}, Y)$.

Methodology. We tackle the problem above by bridging deferring systems with the potential outcome framework. The key observation is that the reject score maps to a running variable, and the accuracy of the ML model and of the human expert map to the potential outcomes.

Table 1 provides a mapping of the role of each variable of a deferring system into the potential outcomes framework. We have that: (i) the *reject score* $K_i = k(\mathbf{X}_i)$ is the running variable; (ii) the *threshold* $\bar{k}_c \in \mathcal{K}$ is the cutoff; (iii) the *policy function* $G_i = \mathbb{1}\{K_i \geq \bar{k}_c\}$ 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* $T_i = \mathbb{1}\{\vartheta(\mathbf{X}_i) = Y_i\}$ is an indicator function for whether the prediction of the deferring system is correct, i.e., the prediction equals the target variable Y_i or not; (v) correctness of ML model $T_i(0) = \mathbb{1}\{f(\mathbf{X}_i) = Y_i\}$ and of the human expert $T_i(1) = \mathbb{1}\{h(\mathbf{X}_i) = Y_i\}$ are the *potential outcomes*, and they are connected to the outcome T_i as follows: $T_i = G_i \cdot T_i(1) + (1 - G_i) \cdot T_i(0)$; finally, (vi) the *individual causal effect* τ_i is the difference between $T_i(1)$ and $T_i(0)$.

In the following, we distinguish two scenarios, which allow for the identification of causal effects. In both scenarios, we assume that the human predictions $h(\mathbf{X}_i)$ can be accessed only for the deferred instances, i.e., if $G_i = 1$.

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. For example, this is the case when the development team runs an internal evaluation of the deferring system. In this scenario, we can identify the causal effects of *deferring to the human* (Section 3.1). Example *Ex1* from the introduction 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 (e.g., due to a pay-per-use fee), may have side effects, or discloses sensitive data. For example, in an external audit, the owner of the deferring system may be reluctant to share ML predictions for deferred instances: since these predictions are not the system’s actual output, releasing them might not be legally binding. In this scenario, we can identify the causal effects of *deferring to the human* locally to the deferring threshold (Section 3.2). Moreover, this scenario is also able to cover the cases where the intervention to be causally evaluated is the introduction of the ML model in a fully human decision-making process, as in example *Ex2* from the introduction. In example *Ex2*, we are not interested in the causal effect of deferring from the ML model to 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 to have the human expert decisions for 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 it.

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 prediction $f(\mathbf{X}_i)$, and the human prediction $h(\mathbf{X}_i)$, are available. Let n be the size of the test set \mathcal{D}_n , then, the following holds.

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

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

Proof. Identification is immediate because both potential outcomes are observed for a 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 $G_i = 1$. \square

Because we can directly compute the most granular causal effect on deferred instances, τ_i , we can also retrieve less granular quantities. For instance, if we average the τ_i over the population of deferred units, we obtain the *average treatment effect on the deferred*, τ_{ATD} , which is the deferring systems’ equivalent of τ_{ATT} in (1). This causal estimand can be identified as follows.

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

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

The proof can be found in Appendix B. τ_{ATD} allows us to measure the (average) effect on accuracy due to the “intervention” of deferring to a human the prediction for the deferred instances. Intuitively, τ_{ATD} estimates what would be the average percentage increase in accuracy for deferred instances if the human predicts instead of the ML model. This effect motivates the introduction of a deferring system to stakeholders. Policymakers can use it to assess the impact of such systems. The estimation of τ_{ATD} is straightforward and can be conducted via a simple difference-in-means estimator of the form:

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

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

Furthermore, any aggregated metrics of the individual causal effects on the deferred can be estimated in this setting. Another interesting quantity is:

$$\tau_{\text{ATD}}(k) = \mathbb{E}[T(1) - T(0) | G = 1, K = k], \quad k \geq \bar{K}_c,$$

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

dataset	n	\mathcal{Y}	human	network model	pre-trained	used in	hyper-parameters
synth	20k	2	synthetic	linear	no	[7]	ep. = 50; lr = 1e-2; op. = Adam
cifar10h	10k	10	separate annotator	WideResNet [44]	yes	[7, 8]	ep. = 150; lr = 1e-3; op. = AdamW
galaxyzoo	10k	2	random annotator	ResNet50 [45]	yes	[2, 5, 7]	ep. = 50; lr = 1e-3; op. = Adam
hatespeech	25k	3	random annotator	FNN on SBERT embeddings [46]	yes	[2, 5, 7, 8]	ep. = 100; lr = 1e-2; op. = Adam
xray-airspace	4.4k	2	random annotator	DenseNet121 [47]	yes	[7]	ep. = 3; lr = 1e-3; op. = AdamW

which measures how the average difference in predictive accuracy changes as the coverage varies above the threshold $\bar{\kappa}_c$. Notice that $\mathbb{E}[T(1) - T(0) \mid G = 1, K = k] = \mathbb{E}[T(1) - T(0) \mid K = k]$ since $k \geq \bar{\kappa}_c$ if and only if $G = 1$. Given $k \geq \bar{\kappa}_c$, a natural estimator of $\tau_{\text{ATD}}(k)$ would be a non-parametric local polynomial kernel regression around k . Local polynomial kernel regressions [43] 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 $k(\mathbf{x}_i)$.

3.2 Scenario 2: deferring systems as an RD design

If we cannot access the ML model predictions for the deferred instances, we can still exploit the additional knowledge on the policy function to estimate causal quantities. In particular, we can exploit the fact that *deferring systems can be interpreted as an RD design* to compute a local version of τ_{ATD} . In this scenario, τ_{RD} answers the following question: for a fixed coverage value c and the corresponding reject score threshold $\bar{\kappa}_c$, 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 $\bar{\kappa}_c$? Such an interpretation is possible if Assumption 1 holds. If we have reasons to believe that small changes of the reject-score threshold $\bar{\kappa}_c$ do not abruptly change the expected predictive accuracy of the human expert and of the ML model, then Assumption 1 is satisfied.

Proposition 3. *Let Scenario 2 hold and let Assumption 1 be satisfied for the deferring system. Then*

$$\lim_{k \rightarrow \bar{\kappa}_c^+} \mathbb{E}[T \mid K = k] - \lim_{k \rightarrow \bar{\kappa}_c^-} \mathbb{E}[T \mid K = k] = \tau_{\text{RD}},$$

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

The proof can be found in Appendix B. Proposition 3 allows us 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, τ_{RD} readily quantifies the gain in predictive accuracy of having the human expert predicting in place of the ML model at the cutoff. The local nature of τ_{RD} motivates the use of local non-parametric polynomial kernel regression to estimate $\mathbb{E}[T \mid K = \bar{\kappa}_c]$ from the left and the right of the cutoff, thus obtaining an estimator $\hat{\tau}_{\text{RD}}$ of τ_{RD} . We point out that τ_{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

We experimentally address three questions:

Q1: *What causal effects of deferring can be estimated under Scenario 1 in a controlled setting?*

Q2: *What causal effects of deferring can be estimated under Scenario 2 in a controlled setting?*

Q3: *What are the causal effects of deferring on real datasets?*

To address **Q1** and **Q2**, we consider a simulated setting to control the data-generating process. We then address **Q3** by considering real-world datasets, discussing in this section the results for Scenario 1. Complete results for synthetic and real data are reported in Appendix D.2. All experimental software is available at <https://github.com/andrepuigni/PODS>. Experimental hardware specifications and carbon footprint are reported in Appendix D.1.4.

4.1 Experimental settings

Data. For **Q1** and **Q2**, we generate synthetic data using the procedure from Mozannar et al. [7]. Such a procedure generates samples containing (i) instances for which the human expert performs

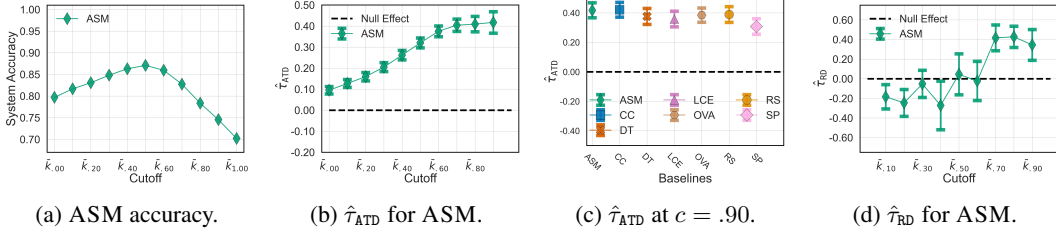


Figure 4: Performance on synthetic data. (a) reports the deferring system accuracy when varying cutoff \bar{k}_c for the best baseline *Asymmetric SoftMax* (ASM) w.r.t. accuracy. (b) reports estimated $\hat{\tau}_{ATD}$ when varying cutoff \bar{k}_c on synthetic data for the best baseline. (c) compares the $\hat{\tau}_{ATD}$ of multiple baselines at a fixed coverage $c = .90$. (d) reports estimated $\hat{\tau}_{RD}$ when varying cutoff \bar{k}_c for the best baseline.

better than the ML model and (ii) instances for which the ML model is better than the human expert. Regarding **Q3**, we consider four datasets used in the LtD literature: *cifar10h* [48], a hard-labelled version of *galaxyzoo* [49], *hatespeech* [50], and *xray-airspace* [51, 52]. Table 2 provides relevant data characteristics, while the applied pre-processing is detailed in Appendix D.1.1.

Baselines. We consider several deferring systems, including: *Selective Prediction* (SP) [53], *Compare Confidence* (CC) [3], *Differentiable Triage* (DT) [2], *Cross-Entropy Surrogate* (LCE) [4], *One Vs All* (OVA) [5], *Realizable Surrogate* (RS) [7] and *Asymmetric SoftMax* (ASM) [8]. Table 2 reports the base network model and the methods’ hyper-parameters used for each dataset. We provide further details for the baselines and the hyper-parameters 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 \bar{k}_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 \bar{k}_c , we estimate on the test set the deferring system accuracy as well as τ_{ATD} and τ_{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 predictive accuracy. The estimate $\hat{\tau}_{ATD}$ is computed through a difference in means estimator and $\hat{\tau}_{RD}$ is obtained using the default implementation of *rdrobust* package [54], i.e., a local linear kernel regression with optimal bandwidth [55]. To assess the statistical significance of the results, we report the 95% confidence intervals and the corresponding p -values (pv) associated with $\hat{\tau}_{ATD}$ and $\hat{\tau}_{RD}$ when testing the null hypotheses of $\tau_{ATD} = 0$ and $\tau_{RD} = 0$, respectively. 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 $7.52e-5$.

4.2 Experimental results

Figure 4 shows the results on the synthetic data (**Q1** and **Q2**). Concerning real data (**Q3**), Figure 5 reports the results for Scenario 1, while results for Scenario 2 are discussed in Appendix D.2. For each dataset, we also report the best-deferring system in terms of accuracy: Figure 4a shows the accuracy for the synth dataset, while the bottom row of Figure 5 plots the accuracy for the real datasets. Appendix D.2 reports results for the other baselines.

Q1: causal effects under Scenario 1. Figure 4b shows the estimated $\hat{\tau}_{ATD}$ s (the green diamonds) and their confidence intervals when varying the cutoff \bar{k}_c . The black horizontal line denotes the null effects for $\hat{\tau}_{ATD}$ and $\hat{\tau}_{RD}$. 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 $c = .90$. Hence, introducing a deferring system turns out to be beneficial. Moreover, the strictly increasing value of $\hat{\tau}_{ATD}$ w.r.t. the cutoff \bar{k}_c allows us to confirm the deferring strategy’s effectiveness further. In fact, the higher the coverage, the more difficult instances are deferred to the human. Hence, the system is properly ranking the instances. We point out that this effect cannot be quantified by only examining the accuracy (see Figure 4a), as done by existing works in the literature.

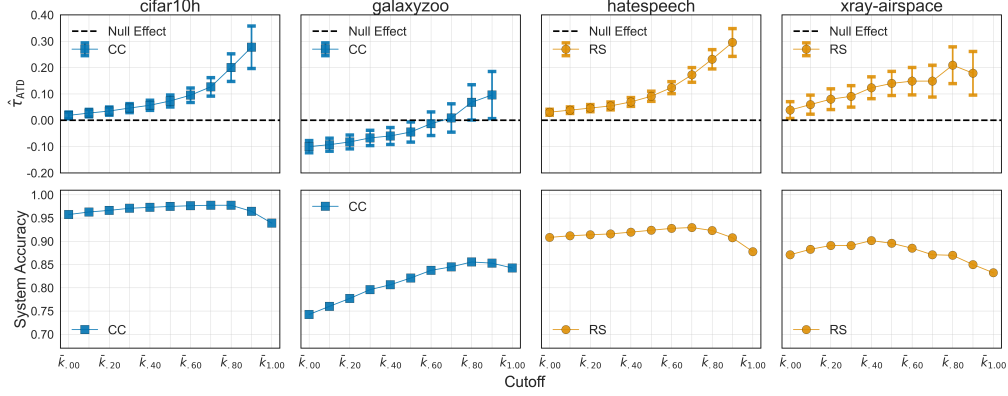


Figure 5: Best deferring system performances on real data when varying the cutoff \bar{k}_c . Top: estimated τ_{ATD} . Bottom: accuracy. CC is *Compare Confidence*, RS is *Realizable Surrogate*.

Moreover, under Scenario 1, we can fix the target coverage and compare the estimate $\hat{\tau}_{\text{ATD}}$ for multiple deferring systems. Figure 4c provides an example of this approach for $c = .90$. The ASM and CC baselines perform the best in such a case.

Q2: causal effects under Scenario 2. Due to the limited access to the ML model predictions, under Scenario 2, the estimation of the causal effect τ_{RD} can be done only at the cutoff value \bar{k}_c because (i) we do not observe human predictions for the instances with reject-score below \bar{k}_c ; (ii) we cannot access the ML model predictions for the instances with reject score above \bar{k}_c ; and (iii) we cannot freely compare the performance on instances assigned to the ML model with performance on instances deferred to the expert as they are different from each other by construction. Figure 4d reports the estimated $\hat{\tau}_{\text{RD}}$'s: coefficients are negative (and not significant) for low coverage values, reaching the minimum value at $\bar{k}_{.40}$ with an estimated effect of $\approx -.272$ ($p\text{-value} \approx 3.14e-2$). The estimates increase and become significant for larger coverage values, achieving a maximum value of $\approx .427$ ($p\text{-value} \approx 1.20e-14$) at $c = .80$. Thus, for the instances with a reject score value close to $\bar{k}_{.80}$, deferring to a human expert causes an increase in accuracy of $\approx 42.7\%$.

As a general consideration, at the cutoff for which the estimated $\hat{\tau}_{\text{RD}}$ is close to zero, we have that predicting with an ML model or a human expert is the same. Hence, such a cutoff maximizes the overall human-AI team accuracy. In higher cutoff values, the human performs better, while in lower cutoff values, the ML model performs better. From Figure 4d, the $\hat{\tau}_{\text{RD}}$ is close to zero for $\bar{k}_{.50}$. At such a cutoff, accuracy is maximized, as shown in Figure 4a. In summary, the study of $\hat{\tau}_{\text{RD}}$ at the variation of \bar{k}_c helps evaluate whether the threshold \bar{k} in the deferring system has been properly set.

Q3: real datasets. Consider Figure 5. Regarding *cifar10h*, the $\hat{\tau}_{\text{ATD}}$'s are positive and increasing with \bar{k}_c : the values range from $\approx .019$ at $c = 0$ to $\approx .265$ ($p\text{-value} \approx 2.83e-10$) at $c = .90$. This means that the deferring system correctly identifies those instances where the ML model is better than the human expert because deferring fewer instances yields higher $\hat{\tau}_{\text{ATD}}$'s.

For *galaxyzoo*, $\hat{\tau}_{\text{ATD}}$ takes negative values for the coverages below $c = .70$. This is because the ML model at full coverage (accuracy of $\approx .843$) performs better than the human expert (accuracy of $\approx .743$). Moreover, for $c = .80$ and $c = .90$, the estimated $\hat{\tau}_{\text{ATD}}$ are $\approx .068$ ($p\text{-value} \approx 4.69e-2$) and $\approx .096$ ($p\text{-value} \approx 3.56e-2$) respectively. In particular, they are not statistically different from zero. Hence, introducing a deferring system would not improve over a fully automated setting.

For *hatespeech*, the causal effects monotonically increase with the coverage, with $\hat{\tau}_{\text{ATD}}$ ranging from $\approx .019$ ($p\text{-value} \approx 9.54e-18$) at $c = 0$ up to $\approx .295$ ($p\text{-value} \approx 7.40e-28$) at $c = .90$. All the values significantly differ from zero, suggesting that deferring is effective.

We can also observe an overall positive effect for the *xray-airspace* dataset. Notice that the highest $\hat{\tau}_{\text{ATD}}$ is achieved at $c = .80$ ($\approx .209$, $p\text{-value} \approx 4.29e-9$), while the $\hat{\tau}_{\text{ATD}}$ at $c = .90$ is lower ($\approx .179$, $p\text{-value} \approx 2.41e-5$). This suggests that the reject score of the deferring system may not properly identify where the human expert performs better than the ML model. However, the difference between $c = .80$ and $c = .90$ is not statistically significant.

5 Conclusions

We tackled the evaluation of predictive accuracy of deferring systems from a causal perspective. Our link with the potential outcomes and with the RD design frameworks directly allows for reasoning on the causal effects of the deferring strategy. Experiments showed some practical guidance on how to reason about the causal estimations to evaluate deferring systems.

Limitations and broader impact. We assumed that we had 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 impact perspective, our evaluation framework can be used to quantify the impact of deferring systems. Since these systems are becoming increasingly popular in many socially sensitive contexts, our approach can help better evaluate and deploy safer systems.

Future work. Multiple directions are open to future research. We mention extending our approach: to evaluate fairness metrics beyond predictive accuracy; to deferring strategies that consider multiple human experts; and to account for the influence that deferring has on human behaviour, as in strategic classification and performative predictions.

Acknowledgments

A. Pugnana and S. Ruggieri have received funding by PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - “FAIR - Future Artificial Intelligence Research” - Spoke 1 “Human-centered AI”, funded by the European Commission under the NextGeneration EU programme. J. M. Alvarez and S. Ruggieri have received funding from the European Union’s Horizon 2020 research and innovation program under Marie Skłodowska-Curie Actions (grant agreement number 860630) for the project “NoBIAS - Artificial Intelligence without Bias”. This work reflects only the authors’ views and the European Research Executive Agency is not responsible for any use that may be made of the information it contains.

References

- [1] David Madras, Toniann Pitassi, and Richard S. Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *NeurIPS*, pages 6150–6160, 2018.
- [2] Nastaran Okati, Abir De, and Manuel Gomez-Rodriguez. Differentiable learning under triage. In *NeurIPS*, pages 9140–9151, 2021.
- [3] Maithra Raghu, Katy Blumer, Greg Corrado, Jon M. Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. The algorithmic automation problem: Prediction, triage, and human effort. *CoRR*, abs/1903.12220, 2019.
- [4] Hussein Mozannar and David A. Sontag. Consistent estimators for learning to defer to an expert. In *ICML*, volume 119, pages 7076–7087. PMLR, 2020.
- [5] Rajeev Verma and Eric T. Nalisnick. Calibrated learning to defer with one-vs-all classifiers. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 22184–22202. PMLR, 2022.
- [6] Rajeev Verma, Daniel Barrejón, and Eric T. Nalisnick. Learning to defer to multiple experts: Consistent surrogate losses, confidence calibration, and conformal ensembles. In *AISTATS*, volume 206, pages 11415–11434. PMLR, 2023.
- [7] Hussein Mozannar, Hunter Lang, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David A. Sontag. Who should predict? exact algorithms for learning to defer to humans. In *AISTATS*, volume 206, pages 10520–10545. PMLR, 2023.
- [8] Yuzhou Cao, Hussein Mozannar, Lei Feng, Hongxin Wei, and Bo An. In defense of softmax parametrization for calibrated and consistent learning to defer. In *NeurIPS*, 2023.

- [9] José M. Álvarez, Alejandra Bringas-Colmenarejo, Alaa Elobaid, Simone Fabbrizzi, Miriam Fahimi, Antonio Ferrara, Siamak Ghodsi, Carlos Mougán, Ioanna Papageorgiou, Paula Reyer Lobo, Mayra Russo, Kristen M. Scott, Laura State, Xuan Zhao, and Salvatore Ruggieri. Policy advice and best practices on bias and fairness in AI. *Ethics Inf. Technol.*, 26(2):31, 2024.
- [10] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009.
- [11] Ana Rita Nogueira, Andrea Pugnana, Salvatore Ruggieri, Dino Pedreschi, and João Gama. Methods and tools for causal discovery and causal inference. *WIREs Data Mining Knowl. Discov.*, 12(2), 2022.
- [12] Alberto Abadie and Matias D. Cattaneo. Econometric Methods for Program Evaluation. *Annual Review of Economics*, 10(1):465–503, 2018.
- [13] Jersey Neyman. Sur les applications de la théorie des probabilités aux expériences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10(1):1–51, 1923.
- [14] Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, October 1974.
- [15] Donald L. Thistlethwaite and Donald T. Campbell. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51(6):309, 1960.
- [16] Donald B. Rubin. Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics*, 6(1):34–58, 1978.
- [17] Paul W. Holland. Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396):945–960, December 1986.
- [18] Jinyong Hahn, Petra Todd, and Wilbert Van der Klaauw. Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design. *Econometrica*, 69(1):201–209, 2001.
- [19] Anqi Mao, Christopher Mohri, Mehryar Mohri, and Yutao Zhong. Two-stage learning to defer with multiple experts. In *NeurIPS*, 2023.
- [20] Andrea Pugnana, Lorenzo Perini, Jesse Davis, and Salvatore Ruggieri. Deep neural network benchmarks for selective classification. *CoRR*, abs/2401.12708, 2024.
- [21] Yo Joong Choe, Aditya Gangrade, and Aaditya Ramdas. Counterfactually comparing abstaining classifiers. In *NeurIPS*, 2023.
- [22] Harald O. Stolberg, Geoffrey Norman, and Isabelle Trop. Randomized controlled trials. *American Journal of Roentgenology*, 183(6):1539–1544, 2004.
- [23] Michela Carlana, Eliana La Ferrara, and Paolo Pinotti. Goals and gaps: Educational careers of immigrant children. *Econometrica*, 90(1):1–29, 2022.
- [24] Abhijit Banerjee, Esther Duflo, Rachel Glennerster, and Cynthia Kinnan. The miracle of microfinance? Evidence from a randomized evaluation. *American Economic Journal: Applied Economics*, pages 22–53, 2015.
- [25] Joshua Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics*. Princeton University Press, 2009.
- [26] Jens Ludwig and Douglas L. Miller. Does Head Start Improve Children’s Life Chances? Evidence from a Regression Discontinuity Design. *The Quarterly Journal of Economics*, 122(1):159–208, February 2007.
- [27] Michele Hilton Boon, Peter Craig, Hilary Thomson, Mhairi Campbell, and Laurence Moore. Regression discontinuity designs in health: a systematic review. *Epidemiology*, 32(1):87–93, 2021.
- [28] Matias D. Cattaneo, Luke Keele, and Rocío Titiunik. A guide to regression discontinuity designs in medical applications. *Statistics in Medicine*, 42(24):4484–4513, 2023.
- [29] Paolo Pinotti. Clicking on Heaven’s Door: The Effect of Immigrant Legalization on Crime. *American Economic Review*, 107(1):138–168, January 2017.

- [30] Joshua D. Angrist and Victor Lavy. Using Maimonides’ Rule to Estimate the Effect of Class Size on Scholastic Achievement. *The Quarterly Journal of Economics*, 114(2):533–575, May 1999.
- [31] Stephanie Riegg Cellini, Fernando Ferreira, and Jesse Rothstein. The Value of School Facility Investments: Evidence from a Dynamic Regression Discontinuity Design. *The Quarterly Journal of Economics*, 125(1):215–261, 2010.
- [32] Esther Duflo, Pascaline Dupas, and Michael Kremer. Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. *American Economic Review*, 101(5):1739–1774, August 2011.
- [33] Cristian Pop-Eleches and Miguel Urquiola. Going to a Better School: Effects and Behavioral Responses. *American Economic Review*, 103(4):1289–1324, June 2013.
- [34] Rafael Lalive. How do extended benefits affect unemployment duration? A regression discontinuity approach. *Journal of Econometrics*, 142(2):785–806, February 2008.
- [35] Erich Battistin, Agar Brugiavini, Enrico Rettore, and Guglielmo Weber. The Retirement Consumption Puzzle: Evidence from a Regression Discontinuity Approach. *American Economic Review*, 99(5):2209–2226, December 2009.
- [36] Decio Coviello and Mario Mariniello. Publicity requirements in public procurement: Evidence from a regression discontinuity design. *Journal of Public Economics*, 109:76–100, January 2014.
- [37] Caroline Flammer. Does Corporate Social Responsibility Lead to Superior Financial Performance? A Regression Discontinuity Approach. *Management Science*, 61(11):2549–2568, November 2015.
- [38] Dries Van der Pias, Wannes Meert, Johan Verbraecken, and Jesse Davis. A novel reject option applied to sleep stage scoring. In *SDM*, pages 820–828. SIAM, 2023.
- [39] Giuseppe Cianci, Roberto Goglia, Riccardo Guidotti, Matteo Kapllaj, Roberto Mosca, Andrea Pugnana, Franco Ricotti, and Salvatore Ruggieri. Applied data science for leasing score prediction. In *IEEE Big Data*, pages 1687–1696. IEEE, 2023.
- [40] Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *J. Mach. Learn. Res.*, 11:1605–1641, 2010.
- [41] Elizabeth Bondi, Raphael Koster, Hannah Sheahan, Martin J. Chadwick, Yoram Bachrach, A. Taylan Cemgil, Ulrich Paquet, and Krishnamurthy Dvijotham. Role of human-AI interaction in selective prediction. In *AAAI*, pages 5286–5294. AAAI Press, 2022.
- [42] Clara Punzi, Roberto Pellungrini, Mattia Setzu, Fosca Giannotti, and Dino Pedreschi. AI, Meet Human: Learning paradigms for hybrid decision making systems. *arXiv preprint arXiv:2402.06287*, 2024.
- [43] Jianqing Fan and Irène Gijbels. *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*. Routledge, January 1996.
- [44] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*. BMVA Press, 2016.
- [45] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society, 2016.
- [46] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP/IJCNLP (1)*, pages 3980–3990. Association for Computational Linguistics, 2019.
- [47] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, pages 2261–2269. IEEE Computer Society, 2017.
- [48] Ruairidh M. Battleday, Joshua C. Peterson, and Thomas L. Griffiths. Capturing human categorization of natural images at scale by combining deep networks and cognitive models. *Nature communications*, 11(1):5418, 2020.
- [49] AstroDave, AstroTom, Christopher Read @ Winton, joycenv, and Kyle Willett. Galaxy zoo - the galaxy challenge, 2013. URL <https://kaggle.com/competitions/galaxy-zoo-the-galaxy-challenge>.

- [50] Thomas Davidson, Dana Warmusley, Michael W. Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *ICWSM*, pages 512–515. AAAI Press, 2017.
- [51] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 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, 2017.
- [52] Anna Majkowska, Sid Mittal, David F Steiner, Joshua J Reicher, Scott Mayer McKinney, Gavin E Duggan, Krish Eswaran, Po-Hsuan Cameron Chen, Yun Liu, Sreenivasa Raju Kalidindi, et al. Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology*, 294(2):421–431, 2020.
- [53] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *NIPS*, pages 4878–4887, 2017.
- [54] Sebastian Calonico, Matias D. Cattaneo, Max H. Farrell, and Rocio Titiunik. Rdrobust: Software for regression-discontinuity designs. *The Stata Journal*, 17(2):372–404, 2017.
- [55] Sebastian Calonico, Matias D Cattaneo, and Max H Farrell. Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs. *The Econometrics Journal*, 23(2):192–210, 2020.
- [56] C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Trans. Inf. Theory*, 16(1):41–46, 1970.
- [57] Kilian Hendrickx, Lorenzo Perini, Dries Van der Plas, Wannes Meert, and Jesse Davis. Machine learning with a reject option: a survey. *Mach. Learn.*, 113(5):3073–3110, 2024.
- [58] Eric T. Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Görür, and Balaji Lakshminarayanan. Hybrid models with deep and invertible features. In *ICML*, volume 97, pages 4723–4732. PMLR, 2019.
- [59] Xin Wang and Siu-Ming Yiu. Classification with rejection: Scaling generative classifiers with supervised deep infomax. In *IJCAI*, pages 2980–2986. ijcai.org, 2020.
- [60] Lize Coenen, Ahmed K. A. Abdullah, and Tias Guns. Probability of default estimation, with a reject option. In *DSAA*, pages 439–448. IEEE, 2020.
- [61] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*. OpenReview.net, 2018.
- [62] Joana Kühne, Christian März, et al. Securing deep learning models with autoencoder based anomaly detection. In *PHM Society European Conference*, volume 6, pages 221–233, 2021.
- [63] Lorenzo Perini and Jesse Davis. Unsupervised anomaly detection with rejection. In *NeurIPS*, 2023.
- [64] Dries Van der Plas, Wannes Meert, Johan Verbraecken, and Jesse Davis. A novel reject option applied to sleep stage scoring. In *SDM*, pages 820–828. SIAM, 2023.
- [65] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Theory and algorithms for learning with rejection in binary classification. *Annals of Mathematics and Artificial Intelligence*, pages 1–39, 2023.
- [66] Radu Herbei and Maten H. Wegkamp. Classification with reject option. *Can. J. Stat.*, 34(4): 709–721, 2006.
- [67] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Boosting with abstention. In *NeurIPS*, pages 1660–1668, 2016.
- [68] Francesco Tortorella. A ROC-based reject rule for dichotomizers. *Pattern Recognit. Lett.*, 26(2):167–180, 2005.
- [69] Filipe Condessa, José M. Bioucas-Dias, Carlos A. Castro, John A. Ozolek, and Jelena Kovacevic. Classification with reject option using contextual information. In *ISBI*, pages 1340–1343. IEEE, 2013.
- [70] Vojtech Franc, Daniel Průša, and Václav Voráček. Optimal strategies for reject option classifiers. *J. Mach. Learn. Res.*, 24:11:1–11:49, 2023.

- [71] Christophe Denis and Mohamed Hebiri. Consistency of plug-in confidence sets for classification in semi-supervised learning. *J. of Nonpar. Statistics*, 32(1):42–72, 2020.
- [72] Andrea Pugnana and Salvatore Ruggieri. A model-agnostic heuristics for selective classification. In *AAAI*, pages 9461–9469. AAAI Press, 2023.
- [73] Andrea Pugnana and Salvatore Ruggieri. AUC-based selective classification. In *AISTATS*, volume 206, pages 2494–2514. PMLR, 2023.
- [74] Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *ICML*, volume 97, pages 2151–2159. PMLR, 2019.
- [75] Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In *NeurIPS*, pages 2898–2909, 2019.
- [76] Lang Huang, Chao Zhang, and Hongyang Zhang. Self-adaptive training: beyond empirical risk minimization. In *NeurIPS*, 2020.
- [77] Leo Feng, Mohamed Osama Ahmed, Hossein Hajimirsadeghi, and Amir H. Abdi. Towards better selective classification. In *ICLR*. OpenReview.net, 2023.
- [78] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, February 2018.
- [79] Abir De, Paramita Koley, Niloy Ganguly, and Manuel Gomez-Rodriguez. Regression under human assistance. In *AAAI*, pages 2611–2620. AAAI Press, 2020.
- [80] Abir De, Nastaran Okati, Ali Zarezade, and Manuel Gomez Rodriguez. Classification under human assistance. In *AAAI*, pages 5905–5913. AAAI Press, 2021.
- [81] Donald L. Thistlethwaite and Donald T. Campbell. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51(6):309–317, 1960.
- [82] David S. Lee. Randomized experiments from non-random selection in U.S. House elections. *Journal of Econometrics*, 142(2):675–697, February 2008.
- [83] Matias D. Cattaneo, Brigham R. Frandsen, and Rocío Titiunik. 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, March 2015.
- [84] Guido W. Imbens and Joshua D. Angrist. Identification and Estimation of Local Average Treatment Effects. *Econometrica*, 62(2):467, March 1994.
- [85] David S. Lee and Thomas Lemieux. Regression discontinuity designs in economics. *Journal of economic literature*, 48(2):281–355, 2010.
- [86] Matias D. Cattaneo and Rocío Titiunik. Regression Discontinuity Designs. *Annual Review of Economics*, 14(1):821–851, 2022.
- [87] Sebastian Calonico, Matias D. Cattaneo, and Rocio Titiunik. Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs. *Econometrica*, 82(6):2295–2326, 2014.
- [88] Michal Kolesár and Christoph Rothe. Inference in Regression Discontinuity Designs with a Discrete Running Variable. *American Economic Review*, 108(8):2277–2304, August 2018.
- [89] Sebastian Calonico, Matias D. Cattaneo, and Max H. Farrell. On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference. *Journal of the American Statistical Association*, 113(522):767–779, April 2018.
- [90] Sebastian Calonico, Matias D. Cattaneo, and Max H. Farrell. Coverage error optimal confidence intervals for local polynomial regression. *Bernoulli*, 28(4), November 2022.
- [91] Matias D. Cattaneo, Nicolás Idrobo, and Rocío Titiunik. *A Practical Introduction to Regression Discontinuity Designs: Foundations*. Cambridge University Press, 1 edition, November 2019.
- [92] Matias D. Cattaneo, Michael Jansson, and Xinwei Ma. Simple Local Polynomial Density Estimators. *Journal of the American Statistical Association*, 115(531):1449–1455, July 2020.

- [93] Matias D. Cattaneo, Luke Keele, Rocío Titiunik, and Gonzalo Vazquez-Bare. Interpreting Regression Discontinuity Designs with Multiple Cutoffs. *The Journal of Politics*, 78(4):1229–1248, October 2016.
- [94] Matias D. Cattaneo, Luke Keele, Rocío Titiunik, and Gonzalo Vazquez-Bare. Extrapolating Treatment Effects in Multi-Cutoff Regression Discontinuity Designs. *Journal of the American Statistical Association*, 116(536):1941–1952, October 2021.
- [95] Yingying Dong and Arthur Lewbel. Identifying the Effect of Changing the Policy Threshold in Regression Discontinuity Models. *Review of Economics and Statistics*, 97(5):1081–1092, December 2015.
- [96] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020.
- [97] Yingying Dong and Michal Kolesár. When can we ignore measurement error in the running variable? *Journal of Applied Econometrics*, 38(5):735–750, August 2023.
- [98] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.
- [99] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*. OpenReview.net, 2019.
- [100] Benoit Courty, Victor Schmidt, Sasha Luccioni, Goyal-Kamal, MarionCoutarel, Boris Feld, Jérémy Lecourt, LiamConnell, Amine Saboni, Inimaz, supatomic, Mathilde Léval, Luis Blanche, Alexis Cruveiller, ouminasara, Franklin Zhao, Aditya Joshi, Alexis Bogroff, Hugues de Lavoreille, Niko Laskaris, Edoardo Abati, Douglas Blank, Ziyao Wang, Armin Catovic, Marc Alencon, Michal Stechly, Christian Bauer, Lucas-Otavio, JPW, and MinervaBooks. mlco2/codecarbon: v2.4.1, May 2024.
- [101] Matias D. Cattaneo, Michael Jansson, and Xinwei Ma. Ipdensity: Local polynomial density estimation and inference. *J. Stat. Softw.*, 101(2), 2022.

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 [56].

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 [57].

Novelty rejection is highly sought if a shift between the training and the test set distributions can occur [38]. 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 [58, 59]. 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 [60]. Alternatively, one can provide a score representing the novelty of an instance and abstain when such a score is above a certain level [61, 62, 63, 64].

Regarding ambiguity rejection, two main approaches have emerged in the literature: Learning to Reject (LtR) [56] and Selective Prediction (SP) [40].

LtR - based on the original work by Chow [56] - 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 [65]. The LtR methods learn the trade-off between abstention and prediction through a parameter a , representing the cost of rejection [66, 67, 68, 69].

On the other hand, SP methods rely on confidence functions, identifying instances where the classifier is more prone to make mistakes [40]. 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) [70].

From a practical perspective, both model-agnostic methods (e.g., [71, 72, 73]) and model-specific ones (e.g., using Deep Neural Network architectures [53, 74, 75, 76, 77]) have been proposed to solve the selective prediction task. For an extensive characterization of deep-neural-network (DNN)-based approaches, we refer to [20], 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 [70], where the authors show that both frameworks share similar optimal strategies, and a recent survey covering abstaining systems can be found in [57], where the authors provide an overall taxonomy of existing approaches.

Choe et al. [21] consider the evaluation of abstaining classifiers, in a scenario where the predictions are missing or inaccessible 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 [78] 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 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. [1] - 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. [79] 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 [80]. Okati et al. [2] 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 [4] propose a method that jointly learns

both the rejector and the ML predictor with some generalization bounds; Verma and Nalisnick [5] and Cao et al. [8] extend the work of [4] by providing consistent surrogate loss with better calibration properties; Verma et al. [6] consider the problem of deferring to a multitude of experts; Mozannar et al. [7] 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.

Regression discontinuity. The RD design first appeared in [81] to study the motivational effect of public recognition on the likelihood of obtaining a scholarship. Forty years later, [18] 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 [82] and [83], 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 [84] and [85], whereas a more recent one contrasting the two approaches mentioned above is [86].

For estimation purposes, the local nature of τ_{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 [43]). Particular attention in the literature has been devoted to how to optimally choose the smoothing bandwidth used in local polynomial estimation (for different approaches to the problem, see [84, 87, 88]) and how to correct for the smoothing bias and conduct valid inference accordingly [89, 90]. 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 [91].

B Proofs

We report here the proofs for Proposition 2 and 3. For the reader’s convenience, we also restate the propositions.

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

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

Proof. We have that:

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

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

Proposition 3. *Let Scenario 2 hold and let Assumption 1 be satisfied for the deferring system. Then*

$$\lim_{k \rightarrow \bar{\kappa}_c^+} \mathbb{E}[T | K = k] - \lim_{k \rightarrow \bar{\kappa}_c^-} \mathbb{E}[T | K = k] = \tau_{\text{RD}},$$

where $\tau_{\text{RD}} := \mathbb{E}[T(1) - T(0) | K = \bar{\kappa}_c]$.

Proof. This is an instance of Theorem 1. □

C Limitations and Extensions

Here are a few caveats regarding the proposed framework.

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 $\bar{\kappa}_c$ 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 $\bar{\kappa}_c$. However, continuity of $\mathbb{E}[T(1) | K = k]$ around $\bar{\kappa}_c$ 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. [41] 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 $K \geq \bar{\kappa}_c$ or $K < \bar{\kappa}_c$, 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 [85] and [91, 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 [91, 92]. For a practical example, see Section D.3 and Figures 7-11.

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 6.

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 6 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 $\bar{\kappa}_c$. 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 $\bar{\mathcal{K}} \subseteq \mathcal{K}$. For instance, we can be in the presence of multiple human moderators that receive different amounts of content to moderate (i.e. $\bar{\kappa}$ is changing). Then, we can leverage results in the literature of RD designs with multiple non-cumulative cutoffs [93, 94] and identify

$$\tau_{\text{RDD}}(k) := \mathbb{E}[T(1) - T(0) | K = k], \quad k \in [\bar{\kappa}_{\text{lb}}, \bar{\kappa}_{\text{ub}}],$$

where $\bar{\kappa}_{\text{lb}} := \min \bar{\mathcal{K}}$ and $\bar{\kappa}_{\text{ub}} := \max \bar{\mathcal{K}}$. More precisely, identification of $\tau_{\text{ATD}}(k)$ requires continuity of $\mathbb{E}[T(d) | K = k], d \in \{0, 1\}$ in k for $k \in [\bar{\kappa}_{\text{lb}}, \bar{\kappa}_{\text{ub}}]$ and a similar shape of $\mathbb{E}[T(0) | K_i = k]$ across datasets. An exercise in this spirit might be useful to choose the optimal level of coverage $\bar{\kappa}^* := \arg \max_{k \in \bar{\mathcal{K}}} \tau_{\text{ATD}}(k)$, 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.

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 [95] show that, under regularity conditions on how $\mathbb{E}[T(d) | K = k], d \in \{0, 1\}$ changes around $\bar{\kappa}_c$, the RD setting can be used to learn how τ_{RD} would change if the reject-score threshold $\bar{\kappa}_c$ were marginally changed. Therefore, this suggests that when the deferring system accuracy is maximal, τ_{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 [96]). 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, [97] show that under the condition that the noisy score correctly assigns samples to treatment and control groups, τ_{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 Experimental evaluation

D.1 Additional details

D.1.1 Data

We use the data from Mozannar et al. [7] and Okati et al. [2].

Concerning the synthetic dataset `synth`, we generate the data using the method described in Mozannar et al. [7]³: 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 $g^* : \mathcal{X} \rightarrow \{0, 1\}$ and one representing an optimal ML model $f^* : \mathcal{X} \rightarrow \mathcal{Y}$. 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 $d = 30$, $p_{h0} = .30$, $p_{h1} = .10$, $p_{ML} = .20$.

Regarding real data, in `cifar10h` [48], the task is to annotate images belonging to 10 different categories. Here, the human prediction is provided by a separate human annotator.

In `galaxyzoo` [49], 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.

For `hatespeech` [50], 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. [7].

Finally, `xray-airspace` [51, 52] 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. [7].

D.1.2 Baselines

Selective Prediction (SP): Geifman and El-Yaniv [53] 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. [3] 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. [2] 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,

³See the GitHub repository https://github.com/clinicalml/human_ai_deferral

(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 [4] 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 [5] propose a different consistent surrogate loss, which improves the final calibration of the deferring system;

Realizable Surrogate (RS): Mozannar et al. [7] 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. [8] extend both LCE and OVA by providing a surrogate loss that ensures a better calibration of the reject score.

For all the baselines but ASM, we consider the implementation provided by Mozannar et al. [7]³. We implement ASM from Cao et al. [8]’s code, which can be found in their supplementary material.

D.1.3 Hyperparameters

For `synth`, we considered a simple linear feedforward architecture. For each baseline, we trained the model for 50 epochs with $lr = 1e - 2$ and Adam [98] 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. [7] (`cifar10h`, `hatespeech`, `xray-airspace`) or Okati et al. [2] (`galaxyzoo`).

In particular, for `cifar10h`, we trained a base WideResNet on the original `cifar10` dataset for 200 epochs using cross-entropy loss, learning rate equals to .001 and AdamW [99] 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 for 100 epochs, setting the learning rate to .01 and Adam as optimizer.

For `xray-airspace`, we first fine-tuned a pre-trained DenseNet121 [47] for 10 epochs on the x-rays that do not contain human predictions, setting $lr = 1e-4$ 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 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` [100]. 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 ≈ 5 days. This translates into an overall CO₂ consumption of roughly ≈ 25.2 Kg Eq.CO₂, which equals a car drive of ≈ 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 $\hat{\tau}_{\text{ATD}}$ (the Scenario 1 main estimate), $\hat{\tau}_{\text{RD}}$ (the Scenario 2 main estimate), and the deferring system accuracy at various target coverages for all the seven baselines. For $\hat{\tau}_{\text{ATD}}$ and $\hat{\tau}_{\text{RD}}$, we show in

Table 3: synth results. Bold for highest value, blue p-values for statistical significance.

	c	ASM	CC	DT	LCE	OVA	RS	SP
$\hat{\tau}_{\text{RD}}$.00	.095 (6.98e-26)	.128 (1.07e-45)	.112 (1.66e-35)	.123 (1.07e-42)	.122 (2.63e-42)	.098 (1.10e-27)	.132 (1.55e-48)
	.10	.127 (3.53e-40)	.164 (1.53e-65)	.137 (1.29e-46)	.154 (8.57e-59)	.154 (8.23e-59)	.130 (1.27e-41)	.158 (2.83e-62)
	.20	.159 (1.28e-55)	.191 (9.98e-81)	.164 (1.01e-59)	.178 (1.57e-70)	.188 (6.60e-80)	.166 (7.48e-61)	.179 (4.01e-72)
	.30	.205 (7.92e-81)	.229 (1.58e-100)	.198 (2.65e-76)	.203 (1.90e-81)	.223 (1.55e-97)	.202 (2.14e-78)	.208 (2.51e-86)
	.40	.262 (6.27e-116)	.264 (5.01e-117)	.234 (4.86e-92)	.234 (8.99e-95)	.262 (7.11e-118)	.251 (9.56e-106)	.233 (2.39e-94)
	.50	.320 (9.03e-158)	.305 (1.16e-138)	.260 (1.51e-98)	.260 (9.15e-98)	.302 (6.62e-136)	.319 (7.30e-150)	.248 (3.62e-88)
	.60	.375 (2.62e-181)	.335 (2.03e-137)	.291 (3.45e-104)	.289 (1.83e-99)	.347 (5.46e-155)	.363 (8.52e-170)	.264 (7.92e-82)
	.70	.404 (8.12e-161)	.364 (3.09e-127)	.323 (3.13e-99)	.327 (1.28e-101)	.381 (6.66e-148)	.390 (4.27e-155)	.281 (1.17e-69)
	.80	.409 (4.07e-106)	.378 (1.31e-94)	.340 (1.98e-72)	.357 (1.04e-83)	.386 (2.32e-100)	.375 (7.04e-89)	.293 (1.06e-51)
	.90	.417 (7.32e-58)	.421 (8.99e-61)	.376 (3.83e-42)	.358 (5.77e-40)	.385 (1.88e-55)	.389 (3.32e-46)	.308 (2.61e-30)
$\hat{\tau}_{\text{RD}}$.10	-.184 (3.62e-3)	-.146 (4.69e-2)	-.095 (1.21e-1)	-.130 (6.87e-2)	-.156 (7.23e-2)	-.272 (2.83e-3)	.068 (5.10e-1)
	.20	-.246 (4.21e-4)	-.183 (9.61e-3)	-.107 (3.22e-2)	-.034 (4.94e-1)	-.044 (5.02e-1)	-.086 (1.54e-1)	.117 (1.57e-1)
	.30	-.052 (4.57e-1)	.008 (8.73e-1)	-.015 (7.74e-1)	-.012 (8.31e-1)	-.077 (1.69e-1)	-.098 (1.15e-1)	-.158 (1.54e-2)
	.40	-.272 (3.14e-2)	-.051 (2.78e-1)	-.009 (8.45e-1)	.099 (9.91e-2)	.007 (9.08e-1)	-.126 (5.03e-2)	.106 (7.09e-2)
	.50	.045 (6.72e-1)	.080 (1.46e-1)	.113 (3.47e-2)	.053 (3.94e-1)	.134 (2.31e-2)	-.024 (6.39e-1)	.190 (1.28e-3)
	.60	-.022 (8.26e-1)	.096 (1.72e-1)	.149 (1.47e-2)	.147 (1.81e-2)	.328 (3.55e-6)	.340 (3.76e-6)	.240 (9.45e-4)
	.70	.417 (5.28e-10)	.314 (5.59e-7)	.210 (4.80e-4)	.062 (3.66e-1)	.374 (1.32e-10)	.339 (5.01e-7)	.352 (1.71e-7)
	.80	.427 (1.20e-14)	.370 (2.01e-7)	.277 (2.63e-6)	.333 (7.94e-9)	.429 (2.18e-12)	.236 (1.02e-3)	.170 (3.32e-2)
	.90	.345 (1.71e-5)	.460 (1.67e-10)	.418 (1.01e-5)	.222 (7.58e-3)	.401 (2.28e-9)	.254 (4.17e-3)	.280 (1.87e-3)
	Accuracy	.00	.797	.797	.797	.797	.797	.797
.10		.817	.816	.808	.812	.813	.815	.809
.20		.831	.825	.818	.818	.828	.833	.812
.30		.848	.831	.825	.820	.834	.843	.817
.40		.863	.830	.826	.820	.836	.853	.812
.50		.871	.829	.820	.810	.832	.863	.794
.60		.860	.810	.809	.796	.822	.853	.774
.70		.827	.785	.788	.779	.796	.824	.751
.80		.784	.748	.757	.750	.752	.773	.726
.90		.745	.714	.721	.709	.716	.736	.701
1.00	.702	.669	.685	.674	.675	.699	.665	

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.

Q3: real datasets under Scenario 2. We observe that most of the $\hat{\tau}_{\text{RD}}$ estimates are not statistically significant. A motivation for this is because the considered test sets are not large in size, ranging from 854 (xray-airspace) to 4,597 (hatespeech) instances. This impacts the estimation strategy, as $\hat{\tau}_{\text{RD}}$ is computed through local kernel regressions over instances within an optimal bandwidth. In summary, locally to the deferring boundary, the difference between the ML model and the human expert is minimal (we cannot reject the null hypothesis that it is zero) for most of the coverages c .

A few exceptions can be noticed. For galaxyzoo and the CC baseline, a few statistically significant (and negative) effects on $\hat{\tau}_{\text{RD}}$ occur. The ML model performs generally better than the human expert on this task, advocating for almost full automation. For hatespeech, there is a statistically significant effect $\approx .348$ ($pv \approx 4.1e-5$) for RS (the best baseline w.r.t. accuracy) at $\bar{\kappa}_c = .90$. Thus, the causal effect of deferring to the human expert on instances around the cutoff $\bar{\kappa}_{.90}$ is positive. For xray-airspace, which resembles the settings of the example (Ex2) from the introduction, and for the RS baseline, there is at $\bar{\kappa}_{.30}$ a large negative effect effect of $\approx -.473$ ($pv \approx 8.95e-3$, yet not statistically significant). This is in favor of deferring from the human expert to the ML model.

We also notice some extreme cases where there is insufficient variability of the reject scores at the cutoff. This occurs when the reject scores are peaked around a value. In those cases, the local polynomial regression, as implemented by rdrobust, is not able to estimate $\hat{\tau}_{\text{RD}}$. These cases are reported as “-” in Tables 3-7.

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:

- we estimate two different cutoff values $\bar{\kappa}_{c,L}$ and $\bar{\kappa}_{c,H}$
 - $\bar{\kappa}_{c,L}$ is obtained by considering the 75-th percentile on the instances with a reject score below $\bar{\kappa}_c$,
 - $\bar{\kappa}_{c,H}$ is estimated by considering the 25-th percentile of the reject scores above $\bar{\kappa}_c$;

Table 4: cifar10h results. Bold for highest value, blue p-values for statistical significance.

	c	ASM	CC	DT	LCE	OVA	RS	SP
$\hat{\tau}_{\text{RD}}$.00	.028 (4.55e-5)	.018 (5.35e-3)	.049 (1.39e-10)	.018 (4.84e-3)	.021 (2.40e-3)	.018 (5.35e-3)	.021 (1.36e-3)
	.10	.033 (2.15e-5)	.027 (1.85e-4)	.038 (4.64e-7)	.021 (2.81e-3)	.024 (1.36e-3)	.021 (4.11e-3)	.024 (1.01e-3)
	.20	.039 (6.45e-6)	.035 (1.16e-5)	.035 (1.55e-6)	.027 (6.14e-4)	.027 (1.01e-3)	.025 (2.37e-3)	.029 (5.43e-4)
	.30	.048 (7.06e-7)	.046 (1.09e-7)	.016 (1.80e-2)	.033 (2.19e-4)	.035 (1.99e-4)	.032 (5.13e-4)	.037 (6.76e-5)
	.40	.062 (3.30e-8)	.057 (5.77e-9)	.007 (2.58e-1)	.041 (7.55e-5)	.045 (3.07e-5)	.041 (1.26e-4)	.043 (4.56e-5)
	.50	.086 (4.67e-11)	.073 (2.30e-10)	-.004 (4.50e-1)	.059 (1.45e-6)	.056 (1.32e-5)	.051 (2.57e-5)	.058 (3.54e-6)
	.60	.108 (2.70e-12)	.095 (1.16e-11)	-.003 (6.38e-1)	.073 (7.63e-7)	.079 (5.61e-7)	.080 (6.18e-8)	.079 (5.54e-8)
	.70	.132 (1.71e-13)	.127 (1.25e-12)	-.007 (1.57e-1)	.098 (2.99e-8)	.116 (7.92e-9)	.120 (1.22e-10)	.122 (8.63e-11)
	.80	.178 (1.83e-14)	.200 (7.80e-14)	-.005 (3.18e-1)	.126 (2.34e-7)	.167 (2.85e-10)	.189 (5.25e-12)	.207 (8.41e-15)
	.90	.306 (2.79e-14)	.277 (1.74e-11)	-.005 (3.19e-1)	.193 (2.32e-7)	.254 (1.05e-10)	.286 (2.72e-12)	.320 (2.75e-13)
$\hat{\tau}_{\text{RD}}$.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
.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	

Table 5: galaxyzoo results. Bold for highest value, blue p-values for statistical significance.

	c	ASM	CC	DT	LCE	OVA	RS	SP
$\hat{\tau}_{\text{RD}}$.00	-.103 (5.96e-18)	-.100 (2.98e-17)	-.093 (1.13e-14)	-.082 (3.04e-11)	-.098 (1.89e-16)	-.099 (6.63e-17)	-.097 (2.81e-15)
	.10	-.088 (5.64e-12)	-.093 (1.81e-13)	-.089 (1.31e-12)	-.081 (3.61e-10)	-.086 (3.07e-11)	-.098 (4.17e-14)	-.096 (2.81e-15)
	.20	-.076 (1.34e-8)	-.082 (8.11e-10)	-.094 (1.77e-12)	-.074 (6.26e-8)	-.072 (2.10e-7)	-.092 (1.16e-10)	-.096 (2.81e-15)
	.30	-.069 (1.01e-6)	-.067 (5.59e-6)	-.084 (1.65e-9)	-.072 (6.54e-7)	-.060 (9.17e-5)	-.078 (8.71e-7)	-.096 (2.81e-15)
	.40	-.063 (5.98e-5)	-.060 (2.78e-4)	-.082 (2.49e-8)	-.051 (1.24e-3)	-.057 (5.86e-4)	-.072 (2.52e-5)	-.056 (2.04e-3)
	.50	-.060 (4.09e-4)	-.045 (2.19e-2)	-.090 (1.13e-8)	-.039 (3.18e-2)	-.046 (1.62e-2)	-.039 (4.32e-2)	-.039 (6.11e-2)
	.60	-.051 (7.96e-3)	-.013 (5.68e-1)	-.100 (4.39e-8)	-.006 (7.62e-1)	-.025 (2.54e-1)	-.006 (7.74e-1)	-.006 (7.88e-1)
	.70	-.036 (1.17e-1)	.009 (7.49e-1)	-.111 (7.99e-8)	.013 (5.85e-1)	-.014 (5.87e-1)	.014 (6.05e-1)	.019 (5.06e-1)
	.80	.017 (5.84e-1)	.068 (4.69e-2)	-.105 (1.65e-5)	.028 (3.65e-1)	.007 (8.12e-1)	.060 (7.64e-2)	.079 (2.81e-2)
	.90	.072 (1.23e-1)	.096 (3.56e-2)	-.119 (5.50e-6)	.102 (1.91e-2)	.024 (6.02e-1)	.089 (7.63e-2)	.105 (3.15e-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-3)	-.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
.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	

- we run a local kernel polynomial regression to estimate $\hat{\tau}_{\text{RD}}$, substituting the true cutoff value $\bar{\kappa}_c$ with both $\bar{\kappa}_{c,L}$ and $\bar{\kappa}_{c,H}$.

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

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

Results Figure 6 provides the results for the placebo tests above. The first row shows the original estimates for $\hat{\tau}_{\text{RD}}$ under Scenario 2 for the best baselines.

Table 6: hatespeech results. Bold for highest value, blue p-values for statistical significance.

	c	ASM	CC	DT	LCE	OVA	RS	SP
$\hat{\tau}_{\text{RM}}$.00	.067 (1.11e-26)	.033 (8.80e-9)	.097 (1.41e-49)	.037 (2.14e-10)	.031 (8.33e-8)	.031 (9.54e-8)	.030 (1.67e-7)
	.10	.078 (7.73e-30)	.037 (1.58e-9)	.089 (4.07e-41)	.042 (4.40e-11)	.038 (1.80e-9)	.038 (1.27e-9)	.036 (1.38e-8)
	.20	.092 (9.89e-35)	.047 (1.07e-12)	.078 (3.08e-30)	.047 (9.99e-12)	.044 (1.27e-10)	.046 (3.68e-11)	.047 (3.25e-11)
	.30	.107 (7.13e-38)	.058 (7.39e-16)	.074 (3.69e-26)	.060 (1.26e-15)	.059 (3.00e-15)	.054 (1.18e-12)	.058 (1.01e-13)
	.40	.129 (4.43e-45)	.071 (6.37e-18)	.062 (7.38e-18)	.076 (2.98e-20)	.075 (2.36e-19)	.070 (7.63e-16)	.076 (6.32e-18)
	.50	.159 (3.03e-51)	.088 (3.66e-21)	.052 (7.85e-12)	.094 (2.07e-23)	.094 (5.27e-22)	.091 (1.47e-20)	.096 (8.34e-21)
	.60	.202 (5.68e-59)	.117 (8.85e-25)	.055 (5.07e-11)	.111 (1.76e-24)	.113 (3.48e-23)	.124 (3.12e-26)	.117 (3.38e-22)
	.70	.281 (1.11e-72)	.160 (1.35e-29)	.052 (1.78e-8)	.134 (4.45e-24)	.138 (1.93e-23)	.172 (4.05e-34)	.165 (8.66e-30)
	.80	.359 (2.21e-73)	.212 (5.92e-33)	.049 (1.39e-5)	.185 (4.49e-27)	.165 (1.44e-20)	.232 (4.47e-35)	.201 (1.74e-26)
	.90	.468 (2.35e-63)	.318 (2.44e-33)	.034 (1.79e-2)	.222 (5.95e-20)	.213 (8.91e-17)	.295 (7.40e-28)	.255 (1.25e-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
.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. Bold for highest value, blue p-values for statistical significance.

	c	ASM	CC	DT	LCE	OVA	RS	SP
$\hat{\tau}_{\text{RM}}$.00	.359 (2.65e-59)	.029 (6.73e-2)	.189 (1.54e-21)	.033 (4.09e-2)	.033 (3.77e-2)	.039 (1.79e-2)	.023 (1.32e-1)
	.10	.416 (6.40e-70)	.046 (1.19e-2)	.196 (2.94e-22)	.053 (3.81e-3)	.055 (2.28e-3)	.059 (1.31e-3)	.034 (5.56e-2)
	.20	.482 (2.31e-86)	.063 (1.46e-3)	.199 (6.83e-22)	.068 (5.43e-4)	.072 (1.98e-4)	.080 (7.14e-5)	.038 (4.99e-2)
	.30	.585 (4.64e-115)	.079 (3.13e-4)	.209 (2.85e-22)	.085 (5.40e-5)	.082 (3.37e-5)	.091 (1.34e-5)	.043 (4.29e-2)
	.40	.673 (7.24e-144)	.092 (1.10e-4)	.239 (5.31e-26)	.092 (6.10e-5)	.108 (1.11e-7)	.123 (5.29e-9)	.068 (4.56e-3)
	.50	.730 (2.49e-160)	.104 (8.15e-5)	.242 (7.17e-23)	.106 (7.50e-5)	.115 (1.61e-7)	.140 (1.91e-9)	.085 (2.06e-3)
	.60	.800 (2.30e-192)	.125 (3.60e-5)	.254 (6.58e-20)	.120 (5.22e-5)	.159 (6.11e-11)	.149 (2.24e-8)	.130 (2.47e-5)
	.70	.858 (1.39e-201)	.183 (2.02e-8)	.191 (2.19e-11)	.160 (7.51e-6)	.173 (9.47e-10)	.149 (1.31e-6)	.158 (8.21e-6)
	.80	.887 (3.48e-145)	.271 (6.70e-11)	.032 (1.02e-1)	.203 (2.17e-6)	.181 (2.22e-8)	.209 (4.30e-9)	.218 (4.30e-7)
	.90	.887 (3.48e-145)	.352 (2.71e-9)	.033 (1.57e-1)	.320 (1.89e-8)	.194 (1.47e-4)	.179 (2.42e-5)	.253 (8.05e-5)
$\hat{\tau}_{\text{RD}}$.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)
	Accuracy	.00	.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	

Regarding synth, all the coefficients of the placebo tests are not significant, with the sole exception of the highest coverage $c = .90$ when considering the highest cutoff (third row of Figure 6).

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

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 $\hat{\tau}_{\text{RD}}$, 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.

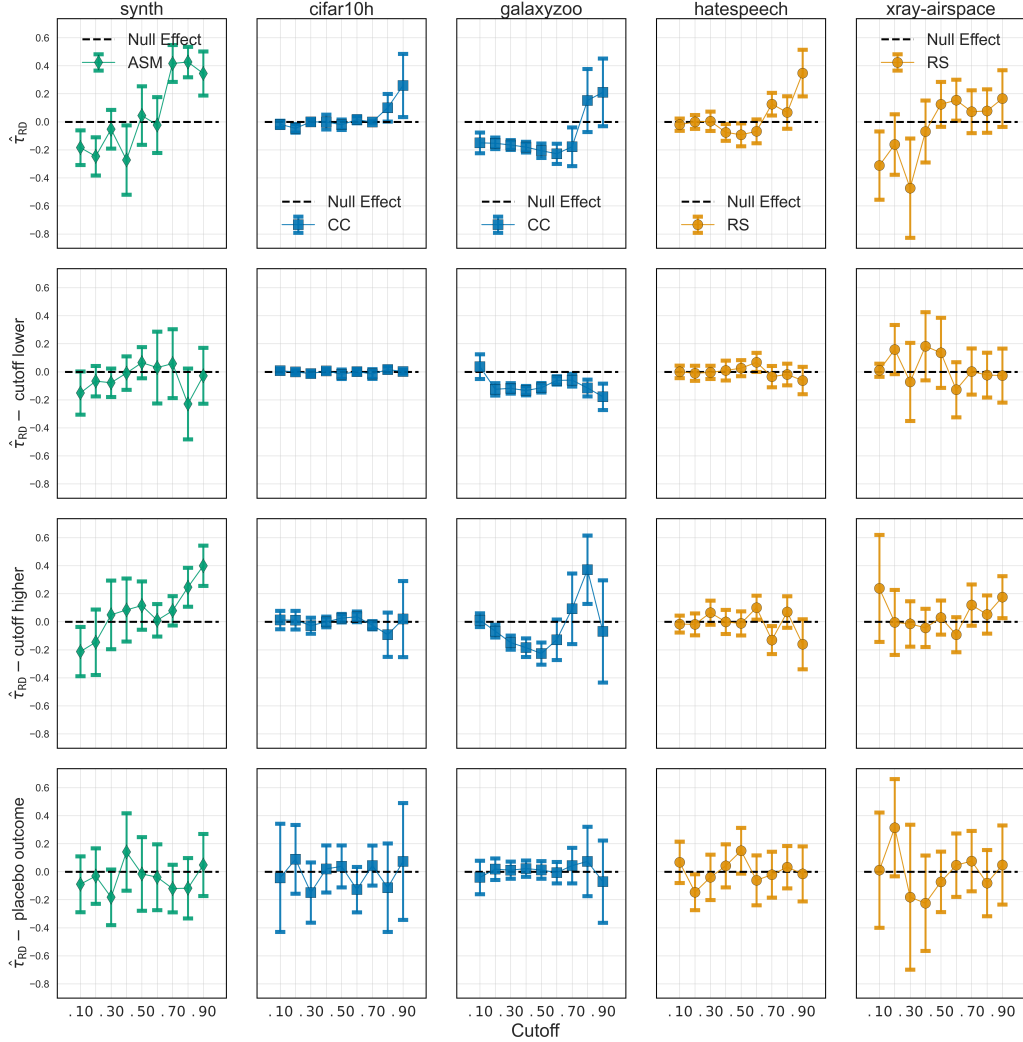


Figure 6: Estimated $\hat{\tau}_{RD}$ for the best baselines over the five datasets. The first row provides the estimates of $\hat{\tau}_{RD}$. The second row provides the estimates for the lower cutoff placebo test. The third row provides the estimates for the higher cutoff placebo test. The fourth row provides the estimates for the placebo outcome test.

D.3.2 Density Estimation

Density estimation test setup. We consider the same setup presented in Section 4.1, and we validate $\hat{\tau}_{RD}$ 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` [101] 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 7-11 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 9), where most tests reject the null hypothesis. The main reason for this behavior is that the reject score density peaks around one value,

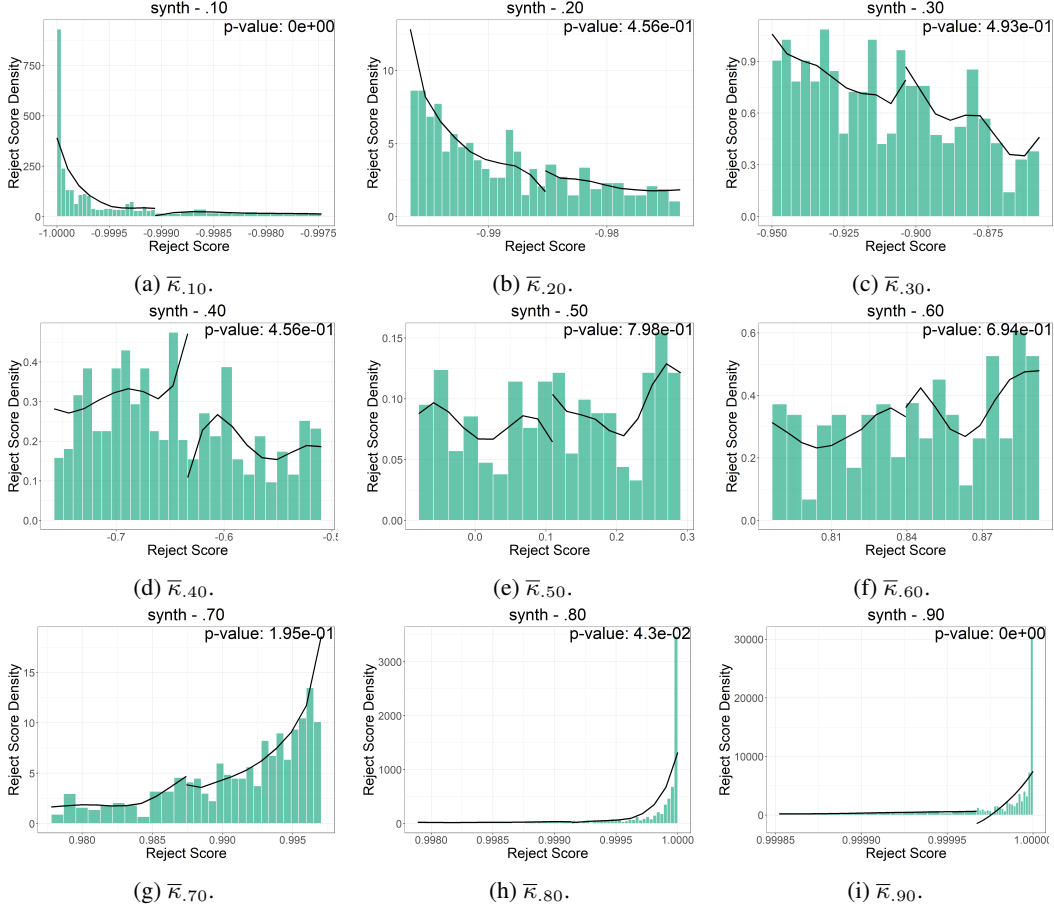


Figure 7: synth estimated best baseline (ASM) reject scores densities at the left and right of cutoff $\bar{\kappa}_c$. All the plots are zoomed around the cutoff values.

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 $c = .10$ and $c = .90$. 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 8), where the p-values are significant only at $c = .10$, and xray-airspace, where values are significant for $c > .70$ and around the accuracy maximizing cutoff value, i.e., $\bar{\kappa}_{.50}$ and $\bar{\kappa}$.

Regarding hatespeech, we see no statistically significant p-values, suggesting the validity of $\hat{\tau}_{RD}$ across all cutoff values (see Figure 10).

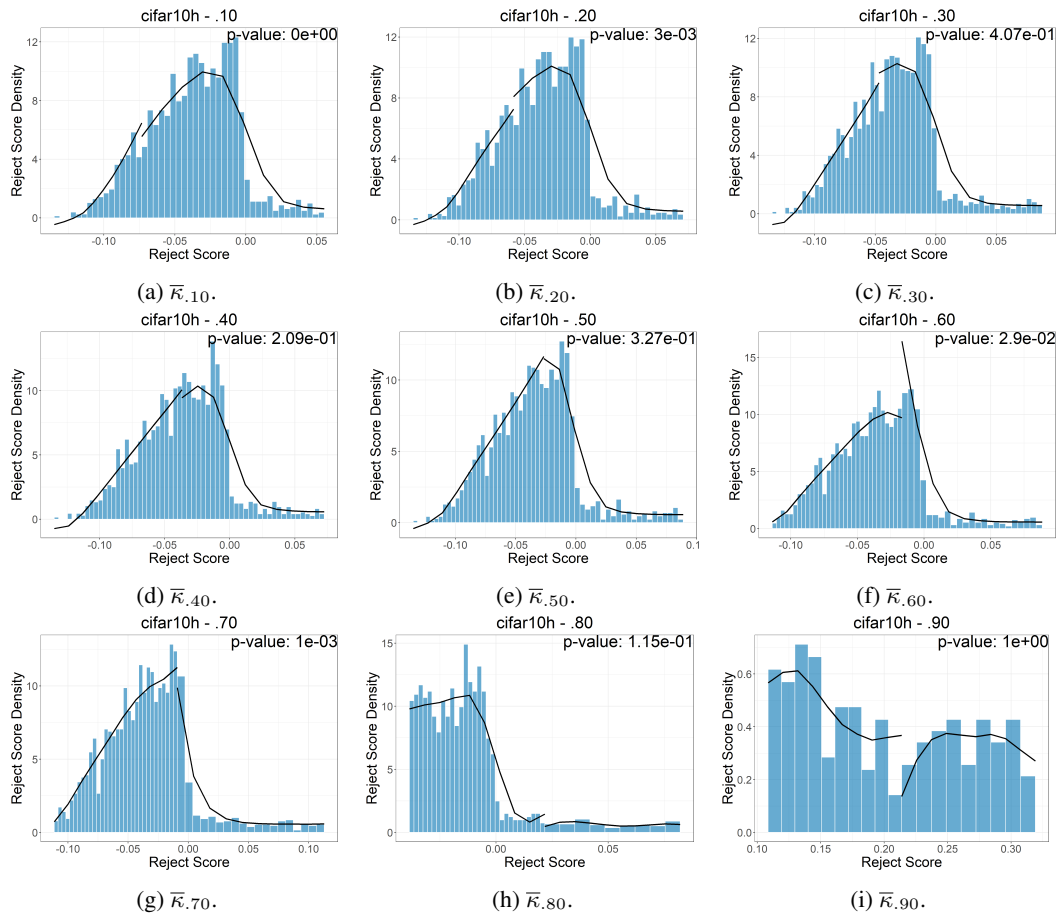


Figure 8: cifar10h estimated best baseline (CC) reject scores densities at the left and right of cutoff $\bar{\kappa}_c$. All the plots are zoomed around the cutoff values.

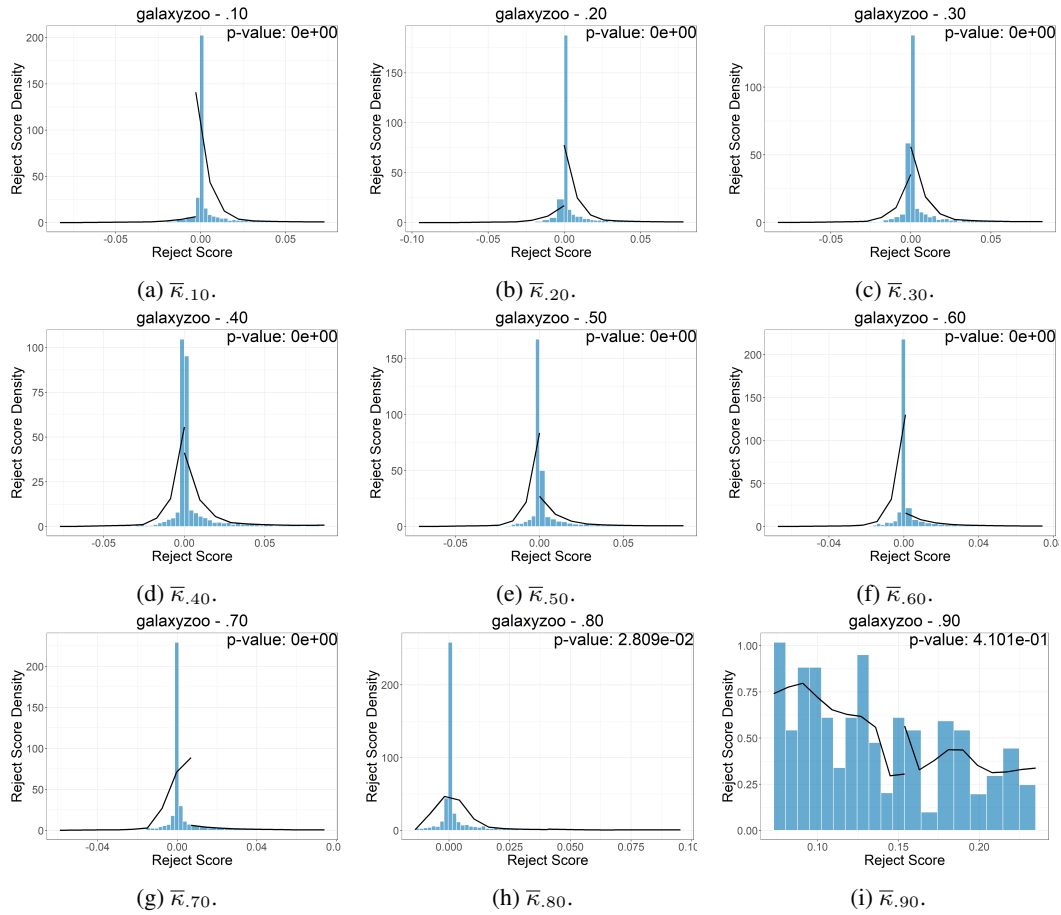


Figure 9: galaxyzoo estimated reject scores densities at the left and right of cutoff $\bar{\kappa}_c$ for the best baseline CC. All the plots are zoomed around the cutoff values.

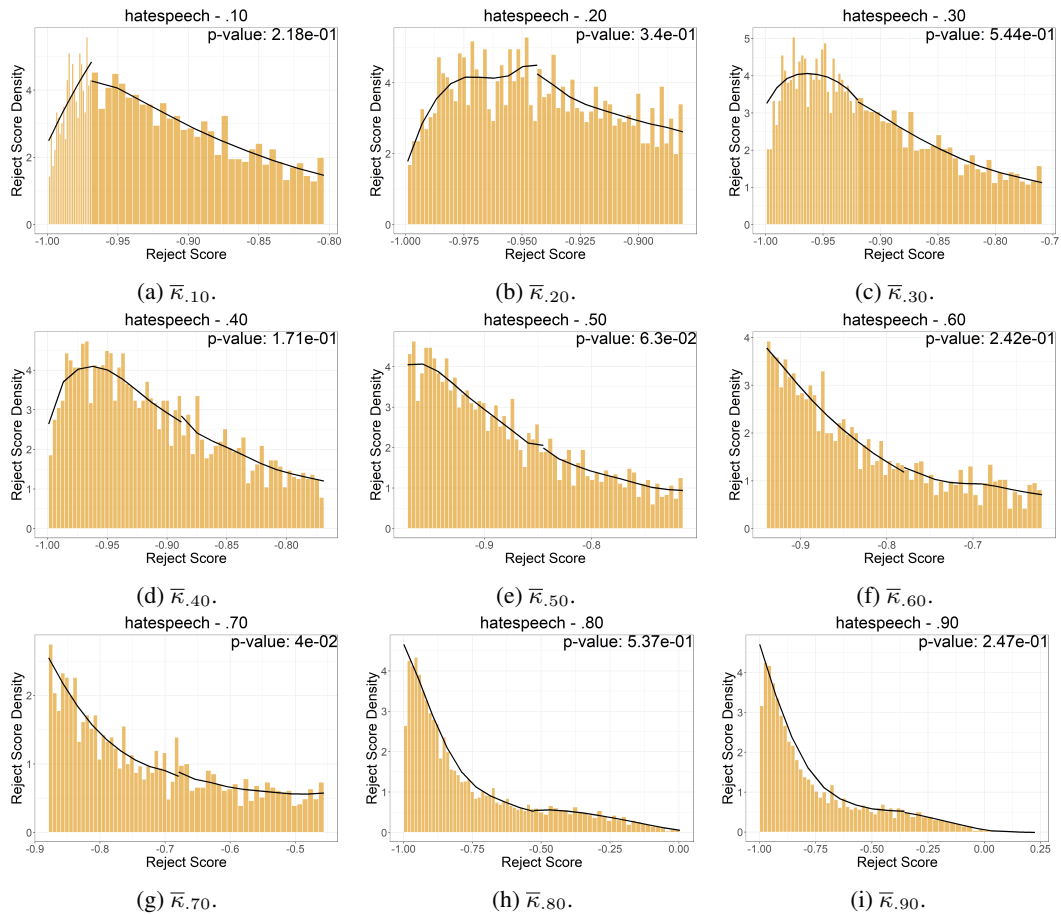


Figure 10: hatespeech estimated best baseline (RS) reject scores densities at the left and right of cutoff $\bar{\kappa}_c$. All the plots are zoomed around the cutoff values.

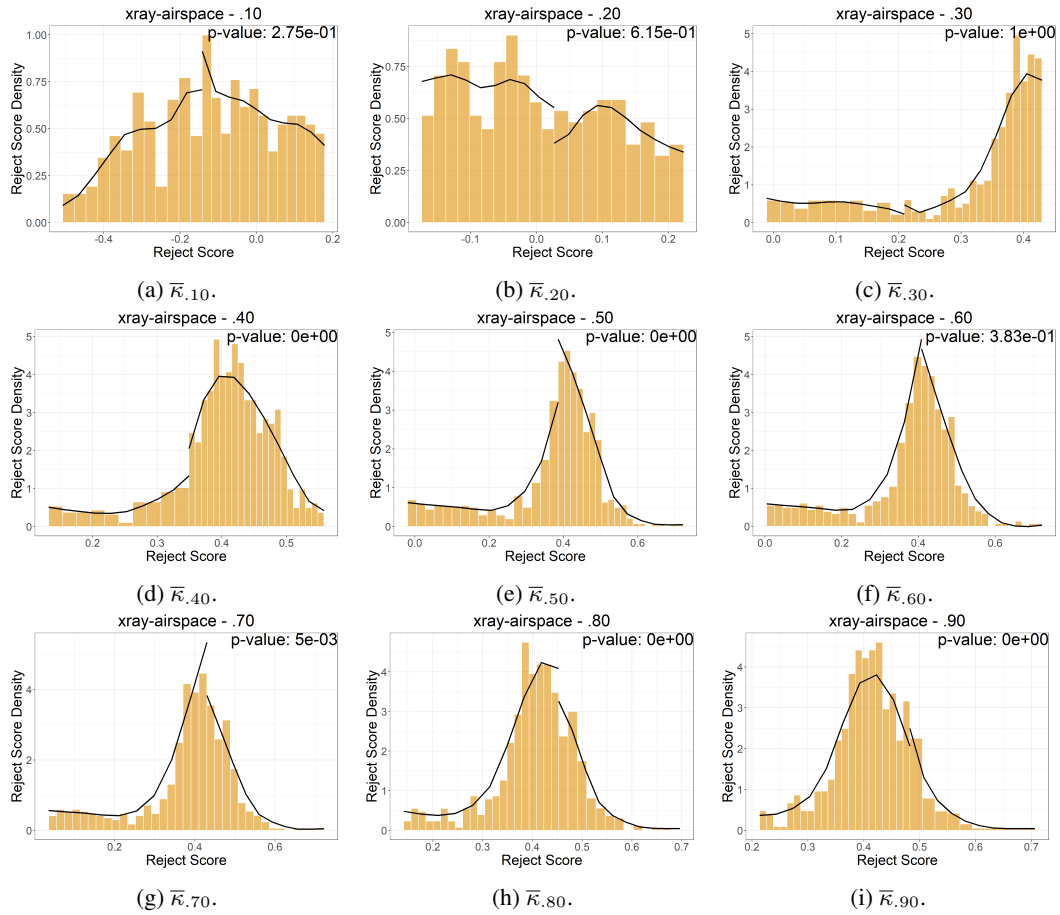


Figure 11: xray-airspace estimated best baseline (RS) reject scores densities at the left and right of cutoff \bar{K}_c . All the plots are zoomed around the cutoff values.