Things Machine Learning Models Know That They Don't Know

Salvatore Ruggieri, Andrea Pugnana

University of Pisa, Pisa, Italy
salvatore.ruggieri@unipi.it, andrea.pugnana@di.unipi.it

Abstract

This paper surveys Machine Learning approaches to build predictive models that know what they don't know. The consequential action of this knowledge can consist of abstaining from providing an output (*rejection*), deferring to another model (*dynamic model selection*), deferring to a human expert (*learning to defer*), or informing the user (*uncertainty estimation*). We formally state the problems each approach solves and point to key references. We discuss open issues that deserve investigation from the scientific community.

Introduction

Humans deem important to find ways of knowing what they do not know. For example, a medical doctor that is given contrasting medical test results may abstain from making a diagnosis. The capability of knowing what one does not know prevents making ineffective or even harmful decisions, such as a wrong medical treatment. Also, it may triggers the acquisition of additional information, such as new medical tests, or the acquisition of new knowledge, e.g., through the consultation with other doctors. Overall, such a capability is at the core of the process of knowledge revision and expansion (Grant 2023).

As for humans, intelligent systems should be able to find ways of knowing what they do not know. In many high-stake domains, abstaining from providing an output is a better strategy than bearing the risk of wrong outputs. Examples include AI systems for job candidate screening, predictive policing, medical diagnosis, and credit scoring. For instance, chatbot developers struggle to detect questions whose answers are not in the capability of the language model (Cheng et al. 2024). Overall, the ability of an intelligent system to know what it does not know empowers the trust of the user and supports accountability in its usage.

In this paper, we survey approaches in the Machine Learning (ML) research to build predictive models that know what they don't know. The consequential action of this knowledge can consist of abstaining from providing an output (*rejection*), deferring to another model (*dynamic model selection*), deferring to a human expert (*learning to defer*), or informing the user (*uncertainty estimation*). The paper is structured according to those actions. The first case is further distinguishing based on the type of uncertainty of the model, either aleatoric or epistemic (Hüllermeier and Waegeman 2021; Gruber et al. 2023). Aleatoric uncertainty is the (irreducible) uncertainty arising from the inherent randomness of an event. Ambiguity rejection addresses aleatoric uncertainty of predictions by abstaining on instances close to the decision boundary. Epistemic uncertainty is the (reducible) uncertainty due to a lack of knowledge. Further, it can be divided into model uncertainty, related to the correct choice of the ML model structure, and estimation uncertainty, related to the correct estimation of model parameters. Novelty rejection addresses estimation uncertainty of predictions for instances far from the distribution of the training set, such as outliers or (distribution) shifted data. Dynamic model selection addresses model uncertainty by deciding which model, among a diverse set, to use on a given instance. Learning to defer differs as the prediction is not deferred to another ML model, but to a human expert with some costs/limitations in the number of predictions the expert can provide. Moreover, the human expert has already been "trained", i.e., it cannot be changed, and its structure is a possibly inconsistent black box. Regarding the direct estimation of uncertainty, classical (parametric and non-parametric) statistical approaches can help estimate both aleatoric and epistemic uncertainty of ML models. Conformal prediction is a non-parametric approach specific for ML classification and regression models.

This paper is structured as follows. First, we introduce notation for canonical ML predictors. Then, we survey each of the approaches, stating the problem it solves and pointing to key references. Finally, we discuss open research lines that deserve investigation from the scientific community.

Canonical Predictors

Let $\mathcal{X}$ be an *input space*, $\mathcal{Y}$ the *target space* and $P(\mathbf{X}, \mathbf{Y})$ the (unknown) joint probability distribution of random variables $\mathbf{X}, \mathbf{Y}$ over $\mathcal{X} \times \mathcal{Y}$. Given a *hypothesis space* $\mathcal{F}$ of functions that map $\mathcal{X}$ to $\mathcal{Y}$, the goal of a machine learning algorithm (a *learner*) is to find a hypothesis $f : \mathcal{X} \to \mathcal{Y} \in \mathcal{F}$, called a *predictor*, that minimizes the *risk*

$$R(f) = \mathbb{E}[l(f(\mathbf{X}), Y)]$$

where $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a user-specified *loss function*. We call f a *classifier* if $\mathcal{Y}$ is a finite set of classes; a *prob-*

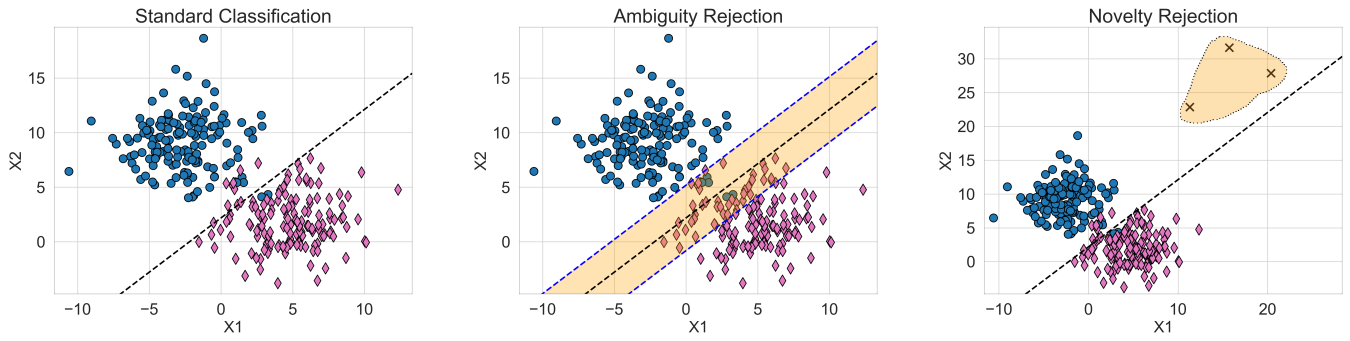

Figure 1: Left: Canonical SVM Classifier. Instances above the black line are predicted as blue circle-shaped instances, while those below are predicted as pink diamond-shaped. Center: Ambiguity Rejection. Abstention occurs for instances close to the decision boundary. Right: Novelty Rejection. Abstention occurs for the instances far from the training data.

abilistic classifier if $\mathcal{Y}$ is the set of k-dimensional discrete distributions[1]; an *ordinal classifier* if $\mathcal{Y}$ is a totally-ordered finite set of labels; a *regressor* if $\mathcal{Y} = \mathbb{R}$; a (list-wise) *ranker* if $\mathbf{X} = \mathbf{X}_1, \ldots, \mathbf{X}_n$ and $\mathcal{Y} = \mathcal{G}_n$ is the set of all permutations of $\{1, \ldots, n\}$; a *generator* if $\mathcal{Y}$ is the set of all strings. The data type of $\mathcal{X}$ further determines specialized predictors for tabular data, sequences, time series, spatial data, text, images, video, etc. In addition to forecast predictions on input instances, predictors can provide knowledge on the unknown distribution $P(\mathbf{X}, \mathbf{Y})$ either directly (for interpretable predictors) or through explainable AI techniques. As an example, Figure 1 (left) shows a Support Vector Machine (SVM) classifier that distinguishes pink diamond-shaped points and blue circle-shaped ones through a hyperplane over the input space.

Abstaining: Ambiguity Rejection

Reconsider the example in Figure 1 (left). The two classes are not perfectly separable, due to inherent randomness of class membership. Ambiguity rejection addresses this problem by abstaining on instances close to the decision boundary of the SVM hyperplane – the orange area in Figure 1 (center). Two methods for ambiguity rejection have been considered in the literature: selective prediction (SP) and learning to reject (LtR).

Selective Prediction A *selective predictor* is a pair (f, g) where f is a canonical predictor and $g : \mathcal{X} \to \{0, 1\}$ is a *selection function* that determines whether f's prediction on an instance is provided (*accepted* or *selected instance*) or the model abstains (*rejected instance*):

$$(f, g)(\mathbf{x}) = \begin{cases} f(\mathbf{x}) & \text{if } g(\mathbf{x}) = 1 \\ \text{abstain} & \text{otherwise.} \end{cases} \quad (1)$$

Selective classification (SC) refers to the case that f is a classifier, and *selective regression* (SR) to the case f is a regressor. Similarly, one can extend it to other types of predictors.

[1]Formally, $\mathcal{Y} = \{\mathbf{p} \in [0, 1]^k \mid \sum_i \mathbf{p}_i = 1\}$. For an instance $\mathbf{x}$, $f(\mathbf{x})$ is then a discrete probability distribution for k classes.

In practice, the selection function g is defined by *(i)* learning a *confidence function*[2] $\upsilon_f : \mathcal{X} \to [0, 1]$ (also called *soft selection* (Geifman and El-Yaniv 2017)) that measures how likely it is that the predictor f is correct, and *(ii)* setting a threshold $\tau \in [0, 1]$ that defines the minimum confidence for providing a prediction, yielding:

$$g(\mathbf{x}) = \mathbb{1}\{\upsilon_f(\mathbf{x}) > \tau\} \quad (2)$$

A widely used confidence function for a probabilistic classifier f is the *softmax response*: $\upsilon_f(\mathbf{x}) = \max_i f(\mathbf{x})_i$. For selective regression, the selection function is based on an estimate of the conditional variance (Zaoui, Denis, and Hebiri 2020), namely $g(\mathbf{x}) = \mathbb{1}\{\mathbb{E}[(Y - f(\mathbf{x}))^2 \mid \mathbf{X} = \mathbf{x}] < \tau\}$. To control for the fraction of instances for which a prediction is made, SP methods consider the *coverage*:

$$\phi(g) = \mathbb{E}[g(\mathbf{X})]$$

namely, the probability mass of the non-rejected region. Another core measure is the risk over the accepted region, commonly called the *selective risk* and defined as:

$$R(f, g) = \frac{\mathbb{E}[l(f(\mathbf{X}), Y)g(\mathbf{X})]}{\phi(g)}$$

For the 0-1 loss, namely $l(f(\mathbf{X}, Y) = \mathbb{1}\{f(\mathbf{X}) \neq Y\}$, selective risk is called *selective error rate*. The inherent trade-off between coverage and risk can be summarized by a *risk-coverage curve* (El-Yaniv and Wiener 2010). Moreover, such a trade-off allows framing the selective prediction task according to two variants (Franc, Průša, and Vorácek 2023). In the bounded improvement model, the problem is formulated by fixing an upper bound ϵ (called *target risk*) for the selective risk and then looking for a selective predictor that maximizes coverage (Geifman and El-Yaniv 2017).

Problem 1 (Bounded-improvement model) *Given a target risk ϵ, an optimal selective predictor (f, g) parametrized by (θ^*, ψ^*) is defined as:*

$$(\theta^*, \psi^*) = \arg\max_{\theta, \psi} \phi(g_\psi) \;\; s.t. \;\; R(f_\theta, g_\psi) \leq \epsilon$$

[2]A good confidence function υ_f should rank instances based on descending loss, i.e., if $\upsilon_f(\mathbf{x}_i) \leq \upsilon_f(\mathbf{x}_j)$ then $l(f(\mathbf{x}_i), y_i) \geq l(f(\mathbf{x}_j), y_j)$.

In the bounded-abstention model, we fix a lower bound c (called *target coverage*) for coverage and then look for a selective predictor that minimizes selective risk (Geifman and El-Yaniv 2019).

Problem 2 (Bounded-abstention model) *Given a target coverage $c \in [0,1]$, an optimal selective predictor (f,g) parameterized by θ^*, ψ^* is defined as:*

$$(\theta^*, \psi^*) = \arg\min_{\theta,\psi} R(f_\theta, g_\psi) \;\; s.t. \;\; \phi(g_\psi) \geq c$$

The estimation of (θ^*, ψ^*) can be modeled either as an end-to-end learning problem, or through a *coverage-calibration* post-training procedure for estimating the threshold τ in (2). The latter is done by estimating the $(1-c)\cdot 100$-th percentile of the confidence function over a held-out calibration dataset. The usage of such a held-out dataset will occur in most of the methods we will survey, as using the confidence function over the training set may overfit the data. As an exception, Pugnana and Ruggieri (2023a,b) adopt cross-fitting rather than the held-out set.

El-Yaniv and Wiener (2010) present a comprehensive Probably Approximately Correct (PAC) analysis of the optimal selective classifier in the bounded-improvement model. Geifman and El-Yaniv (2017) propose an algorithm that achieves optimal results by thresholding the confidence function. For the bounded-abstention setting in a noisy binary setting, Denis and Hebiri (2020) provide an optimal strategy to build a selective classifier when considering the 0-1 loss. From a practical perspective, in recent years, several neural network architectures - e.g., (Geifman and El-Yaniv 2019; Huang, Zhang, and Zhang 2020; Feng et al. 2023; Corbière et al. 2019) - have been proposed to solve the bounded-abstention problem. For an extensive experimental comparison of state-of-the-art methods, we refer to Pugnana et al. (2024). Limited attention has been devoted to the problem of selective regression. The main contributions in this setting are due to Zaoui, Denis, and Hebiri (2020), where the authors provide theoretical results for the mean squared error (MSE) loss function.

Learning to Reject Learning to Reject (LtR), or cost-based abstention (Franc, Průša, and Vorácek 2023), is based on the seminal work by Chow (1970). Similarly to SC, LtR aims to learn a selective predictor. However, LtR methods learn a pair (classifier, rejector) that trade-offs between abstention and prediction through a parameter a, representing the cost of abstention. Let us revise the definition of expected risk as:

$$R(f,g,a) = \mathbb{E}[l(f(\mathbf{X}),Y)g(\mathbf{X}) + a(1-g(\mathbf{X}))]$$

The goal then becomes to minimize such a risk.

Problem 3 (Cost-based model) *Given the cost of rejection a, an optimal selective predictor (f,g) parameterized by θ^*, ψ^* is defined as:*

$$(\theta^*, \psi^*) = \arg\min_{\theta,\psi} R(f_\theta, g_\psi, a)$$

LtR deviates from SC in two major aspects. First, the rejection strategy does not rely on confidence functions, but it is based on the parameter a. Defining such a parameter is not straightforward, and it is heavily context-dependent (Denis and Hebiri 2020). Second, LtR methods are not meant to consider a target coverage c as an objective to achieve. An in-depth theoretical analysis for both LtR and SC can be found in (Franc, Průša, and Vorácek 2023), showing that the two frameworks share similar optimal strategies.

A recent survey on LtR in binary classification is due to Cortes, DeSalvo, and Mohri (2024). Key theoretical studies are due to Chow (1970), who provides an optimal strategy for the 0-1 loss and assuming $P(\mathbf{X},Y)$ is known, and to Herbei and Wegkamp (2006), who extend it to the case where $P(\mathbf{X},Y)$ is not known through the plug-in approach. While most of the LtR methods are model-agnostic with respect to the learner of f, a few model-specific approaches have been proposed for k-Nearest Neighbour classifiers (Hellman 1970), SVMs (Fumera and Roli 2002), neural networks (Cordella et al. 1995b), and ensembles (Cortes, DeSalvo, and Mohri 2016). Notably, Fischer, Hammer, and Wersing (2016) investigate optimal rejection methods for classifiers that partition the input space, such as prototype-based classifiers, SVMs, and decision trees. They consider (un)certainty metrics, such as the distance to the closest decision boundary, and propose methods to determine optimal local thresholds, linking the LtR problem to the Knapsack problem.

Abstaining: Novelty Rejection

Novelty rejection (Dubuisson and Masson 1993; Cordella et al. 1995a) deals with estimation uncertainty by abstaining on instances that are unlikely in the distribution generating the training data. E.g., in the healthcare context, Van der Pias et al. (2023) use a Local Outlier Factor-based rejector to avoid providing predictions for patients that are younger than the ones used to train the model. In general, methods for density estimation or for out-of-distribution detection readily apply to novelty rejection (Hendrickx et al. 2024).

Regarding density estimation methods, the rejector g is intended to estimate the marginal density $P(\mathbf{X})$ from the training set and to accept an instance if the tail probability of observing such an instance $\hat{F}(\mathbf{x}) \approx P(\mathbf{X} > \mathbf{x})$ is greater than a fixed threshold, namely $g(\mathbf{x}) = \mathbb{1}\{\hat{F}(\mathbf{x}) > \tau\}$. Several methods have been considered for computing $\hat{F}$: Landgrebe et al. (2004) propose Gaussian Mixtures Models (GMM), Nalisnick et al. (2019) resort to deep neural networks using Normalizing Flows, and Wang and Yiu (2020) employ Variational Autoencoders. A variant (Condessa, Bioucas-Dias, and Kovacevic 2015) consists of estimating the class conditional density function $P(\mathbf{X} = \mathbf{x} \mid Y = y)$, and then define a novelty-confidence function as:

$$\upsilon_{nov}(\mathbf{x}) = \max_{y \in \mathcal{Y}} P(\mathbf{X} = \mathbf{x} \mid Y = y)$$

and $g(\mathbf{x}) = \mathbb{1}\{\upsilon_{nov}(\mathbf{x}) > \tau\}$. The novelty-confidence function can be estimated using GMM, as done by Vailaya and Jain (2000), or using heuristics, such as the distance from the closest prototype (Conte et al. 2012).

Regarding out-of-distribution detection, an option is to employ a one-class classification model that learns to bound

the region of the training dataset and flag as novel all the instances outside such a region (Coenen, Abdullah, and Guns 2020). Alternatively, a few approaches assign a novelty score to an instance, and abstain when such a score is above a certain level: Liang, Li, and Srikant (2018) adopt temperature scaling and add small perturbations to the input to help separate the neural network softmax score distributions for out-of-distribution instances from in-sample ones; Kühne, März et al. (2021) combine a deep learning model with an autoencoder, and test if the autoencoder is able to reconstruct the input instance. Drapal, de Menezes e Silva Filho, and Prudêncio (2024) propose meta-learning techniques for learning data characteristics that are informative for determining when to reject predictions.

A few works try and merge novelty and ambiguity rejection in the so-called *unknown detection* (Kim, Koo, and Hwang 2023). A first heuristics was proposed by Xia and Bouganis (2022), while a theoretical analysis of unknown detection can be found in Franc, Paplhám, and Prruvsa (2024) and in Narasimhan et al. (2024).

Deferring to Another Predictor: Dynamic Model Selection

In its most simple form, a *dynamic predictor* uses the selection function g to decide between two (base) predictors:

$$(\{f_0, f_1\}, g)(\mathbf{x}) = \begin{cases} f_1(\mathbf{x}) & \text{if } g(\mathbf{x}) = 1 \\ f_0(\mathbf{x}) & \text{otherwise.} \end{cases} \quad (3)$$

Contrasting to (1), the choice now is between the two predictors f_0 and f_1, rather than between a predictor and the abstention option. More in general, multiple base predictors[3] can be considered to choose from – the resulting system, when base predictors are classifiers, is called a *multi-classifier system* (Wozniak, Graña, and Corchado 2014) (also called *delegating classifiers*). The base predictors are diverse in one or more respects (Cruz, Sabourin, and Cavalcanti 2018) such as: initialisation (e.g., of the same neural network), hyper-parameters (e.g., learning rate), architectures (e.g., number of hidden layers), model families (e.g., neural networks and decision trees), training sets (e.g., random subsets or time-dependent subsets), and predictive feature sets (e.g., random subsets or cost-based subsets). For instance, in an AI-of-things scenario, an energy-efficient predictor uses a low number of features, while an energy-demanding classifier uses all available features. The energy-efficient predictor can be accurate enough for most of the cases, and the selection function has to discriminate the difficult cases in order to call the energy-demanding predictor.

In general, the selection function g partitions the space of instances into regions of competence (based on some performance metric) of the two predictors f_0 and f_1. For an input instance $\mathbf{x}$, $g(\mathbf{x})$ is the id of the most competent predictor for the region $\mathbf{x}$ is located into. The upper limit performances of a dynamic predictor are given by the oracle selection function, which always selects a correct predictor if it exists (Kuncheva 2002). As in abstaining predictors, the selection function is trained on a hold-out labeled set, here called the dynamic selection dataset (DSEL). A simple approach is to cluster instances in DSEL, and then determining for each cluster the most competent predictor. Competence can be quantified based on different metrics, ranging from local confidence (Jitkrittum et al. 2023) to data complexity of cluster instances (Schmeing, Brun, and Silva 2022).

Selective (1) and dynamic (3) predictors share a common formulation – and, historically, a common parent in Chow (1965, 1970). The main difference between the two frameworks is that the latter assumes the base predictors as given, and only the selection function g is learnt. Selective prediction approaches do not necessarily make such an assumption, allowing for end-to-end learning of both f and g. Another difference is that dynamic predictors do not optimize for a target coverage in the fraction of usage of the base classifier f_0 or for a cost of its usage. Moreover, dynamic predictors are driven by competence, while selective predictors by ambiguity or novelty – even though ambiguity-based dynamic predictors have been proposed by dos Santos, Sabourin, and Maupin (2007). However, selective and dynamic predictors can be combined, namely one can learn a selective predictor where f is a dynamic predictor. We are not aware of approaches that learn effective selection functions for such a combination.

For surveys on dynamic model selection, we refer the reader to Britto, Sabourin, and de Oliveira (2014) (categorization of approaches), Wozniak, Graña, and Corchado (2014) (survey and applications), Cruz, Sabourin, and Cavalcanti (2018) (survey and experimental comparison), and Schmeing, Brun, and Silva (2022) (experimental evaluation of complexity measures).

Deferring to a Human: Learning to Defer

Learning to Defer (LtD) (Madras, Pitassi, and Zemel 2018), also known as "learning under triage" (Okati, De, and Gomez-Rodriguez 2021), combines ML predictors f and human expert knowledge, modeled as predictor h. LtD can be seen as an instance of dynamic prediction:

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

The human expert predictor h is accessible only for a (typically, small) sample of training/test instances. Human experts can exploit knowledge and features not available to ML predictors, e.g., for the example from the introduction, additional medical test results. Since h is assumed to be given, the predictor f and the selection function g are the only to be learnt. However, the overall risk $R(f, h, g)$ is defined as a linear combination of the risks of the two predictors:

$$\mathbb{E}[l_M(f(\mathbf{X}), Y)g(\mathbf{X}) + l_H(h(\mathbf{X}), Y)(1 - g(\mathbf{X}))] \quad (4)$$

where l_M and l_H are machine-specific and human-specific loss functions. The human loss l_H, in particular, may be strictly positive to account for a cost in asking the human expert, whether the answer is correct or not. This makes the

[3] As another extension, more than one predictor can be selected. This requires some fusion of their outputs, e.g., weighting, voting, or stacking mechanisms (Cruz, Sabourin, and Cavalcanti 2018).

choice between a ML predictor and a human predictor an asymmetric objective. LtD inherits from selective prediction the constraint on target coverage c, intended to bound the fraction of predictions deferred to the human expert.

Problem 4 (LtD) *Given a target coverage $c \in [0, 1]$, and a human expert predictor h, an optimal LtD predictor $(\{f, h\}, g)$ parameterized by θ^*, ψ^* is defined as:*

$$(\theta^*, \psi^*) = \arg\min_{\theta,\psi} R(f_\theta, h, g_\psi) \;\; s.t. \;\; \phi(g_\psi) \geq c$$

Most methods consider a soft selection function $g(\mathbf{x}) = \mathbb{1}\{k(\mathbf{x}) > \overline{\kappa}\}$ (cfr (2)), where the reject score function $k : \mathcal{X} \to \mathbb{R}$ estimates whether the human expert prediction is more likely to be correct than the one of the ML predictor (Mozannar et al. 2023). Okati, De, and Gomez-Rodriguez (2021) show that such a thresholding strategy is optimal. In practice, one can estimate the threshold $\overline{\kappa}$ in a similar way as for selective classification, namely through a *coverage-calibration* procedure by setting $\overline{\kappa}$ as the $c \cdot 100$-th percentile of the reject score values over an hold-out dataset (Pugnana et al. 2024). In a relaxed problem, with no coverage constraint, a linear search procedure can be run to select the $\overline{\kappa}$ that minimize the empirical risk over the hold-out dataset (Mozannar et al. 2023).

LtD is an instance of hybrid decision making where humans oversee machines (Punzi et al. 2024). From a theoretical perspective, De et al. (2020) show that the problem of learning under human assistance 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 by De et al. (2021). A formal characterization of the scenarios where a predictive model can take advantage of including humans in the loop is provided by Okati, De, and Gomez-Rodriguez (2021). They show that standard ML models that are trained to predict over all the instances may be suboptimal when it comes to LtD. Due to the difficulties in directly optimizing (4), consistent surrogate losses are adopted to jointly learn both the selection function and the ML predictor (Mozannar and Sontag 2020; Charusaie et al. 2022; Verma and Nalisnick 2022; Mozannar et al. 2023; Cao et al. 2023; Liu et al. 2024; Wei, Cao, and Feng 2024). Recent works extend the LtD problem to account for multiple human experts, e.g., see Verma, Barrejón, and Nalisnick (2023); Mao et al. (2023) and Cao et al. (2023), and cases where the ML model is already given and not jointly trained, e.g., Mao et al. (2023).

Informing the User: Uncertainty Estimation

There are situations where abstaining or deferring would be detrimental, such as time-critical and mission-critical decision making. These include medical emergency, car collision-detection, cyberthreat blocking, online fraud detection, aircraft autopilot, etc. In such cases, uncertainty estimations is a valid solution to inform the decision maker or to trigger rule-based decisions.

A general distinction can be made between set-valued predictions and uncertainty quantification methods (Hüllermeier and Waegeman 2021). The former methods provide a set of predictions that comprise the true value with a probability guarrantee. Conversely, uncertainty quantification methods provide a single prediction and equip it with additional information about how that prediction is certain.

Concerning set-valued predictions, one of the most common approaches is *conformal prediction* (CP) (Papadopoulos et al. 2002). It consists of transforming a predictor (typically, a probabilistic classifier) $f : \mathcal{X} \to \mathcal{Y}$ into a set-valued predictor $\mathcal{C} : \mathcal{X} \to 2^{\mathcal{Y}}$ such that:

$$P\left(Y \in \mathcal{C}(\mathbf{X})\right) \geq 1 - \alpha \tag{5}$$

where α is the maximum error probability (that the true value does not belong to the predicted set). A simple procedure is based on the notion of *conformal score*, which is a function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ that increases whenever a prediction is far from the truth. To guarantee that Eq. (5) holds, the procedure: (i) estimates the distribution of s (typically over a held-out set); (ii) computes the $(1 - \alpha)$-quantile of s (denoted as $\hat{q}$); (iii) includes in the prediction set all the class labels for which the estimated conditional probability is greater than $1-\hat{q}$. The conformal score is closely related to the concept of confidence function. However, CP focuses on quantifying the uncertainty associated with *each* prediction set (Straitouri et al. 2023; De Toni et al. 2024), rather than minimizing selective risk for a *subset* of predictions (the target coverage). We refer the reader to Angelopoulos and Bates (2023) for an introduction to conformal prediction; an early book is (Vovk, Gammerman, and Shafer 2005), while domain-specific surveys cover text (Campos et al. 2024), and spatio-temporal data (Sun 2022).

Regarding uncertainty quantification of predictions, this is a very active research area in the context of dynamical systems (Smith 2024; Soize 2017), i.e. time-space dependent models. For ML models that are dynamical, such as the vector autoregressive models (Kilian and Lütkepohl 2017), one can reuse/adapt methods and tools for: modeling uncertainty of the inputs (initial states or boundary conditions), for uncertainty propagation, for sensitivity analysis of the input over the system states and outputs, for probabilistic inverse problems (abduction of inputs from outputs). A ML model with uncertainty augments point-wise predictions with additional information, such as confidence intervals, standard errors, or a density estimate. For statistical regression models, such information are provided under some assumptions (homoscedastic gaussian noise) (Kutner et al. 2004). Non-parametric approaches, such as quantile regression and its extension to ensembles (Meinshausen 2006), directly tackle the prediction of confidence intervals. Probabilistic classifiers naturally assign a probability score to each class value. Such a score may not necessarily be calibrated, i.e., the score is the true probability of the class value. Silva Filho et al. (2023) provide a complete overview of calibration methods. In general, most ML models do not natively tackle uncertainty at their outputs. Survey papers on uncertainty quantification are Hüllermeier and Waegeman (2021); Abdar et al. (2021); Gawlikowski et al. (2023). They contrast approaches along different dimensions: frequentist/ensembling vs Bayesian methods; model-agnostic

vs model-specific methods; methods tackling in-domain vs domain-shift vs out-of-domain uncertainty.

Open Research Issues

Several relevant aspects of the research presented remain open for further investigation. Next, we consider four of them. A few others can be only mentioned, including abstaining/deferring: in evolving scenarios as in continual learning; for non-tabular data, such as time series, graphs, or text; for ranking problems, such as pair-wise or list-wise ranking of applicants to a position; in privacy-preserving applications; and extensions exploiting background knowledge, such as knowledge graphs or logic rules. The issue of knowing what they don't know is especially relevant to Large Language Models, for which specialized solutions must be considered (Kapoor et al. 2024; Yin et al. 2023).

Fairness. The literature on methods for documenting, mitigating, and controlling bias and fairness in AI is vast and multidisciplinary (Álvarez et al. 2024; Ntoutsi et al. 2020; Mehrabi et al. 2022), yet with several obstacles to their practical applicability (Ruggieri et al. 2023). Intuitively, knowing what ML models don't know appears a natural means to prevent making decisions that may be wrong or unfair. This is not necessarily the case. Jones et al. (2021) show that even if accuracy can improve on average for accepted instances, selective classification can simultaneously magnify accuracy disparities between protected social groups. Another interesting aspect of rejecting instance regards how rejection is distributed over different social groups. To some extent, rejected instances may suffer from delayed decision or even no decision at all. Initial works on fair selective prediction include Lee et al. (2021); Shah et al. (2022); Schreuder and Chzhen (2021); Lenders et al. (2024).

Explainability. Explainable AI aims at providing explanations of the decision logic of complex and obscure AI models (Guidotti et al. 2019; Minh et al. 2022). The reasons for rejecting or deferring an instance, or for providing an uncertainty estimate, are clearly part of the decision logic that need to be explained. However, the current literature on explaining the selection mechanism is rather limited. Fischer, Hammer, and Wersing (2016) propose a reject option for interpretable-by-design models such as prototype-based ones. This allows for directly characterizing the rejected instances. For black-box models, Artelt et al. (2022) propose counterfactual explanations, which describe what to change in an instance for reversing the output of the selection function. Other model-agnostic methods include Artelt and Hammer (2022); Artelt, Visser, and Hammer (2023). Explaining the differences between ML model and human predictions is explored by Mecke et al. (2024). Finally,since explanations are directed to humans (users, developers, auditors), an important factor is how to communicate uncertainty. Cognitively-robust communication (Kompa, Snoek, and Beam 2021) and explanation (Zukerman and Maruf 2024) of uncertainty is an open problem.

Effective decision-making. Decision-making supported by AI is increasingly being used in many application domains. Multiple studies found that the usage of AI can induce some cognitive bias in the human decision maker (Rastogi et al. 2022). For instance, Green and Chen (2019) investigated with a controlled experiment the effectiveness of using an ML-based risk assessment model as a decision aid. Their results suggest that the human participants: *(i)* were incapable of accurately assessing the validity of their predictions or the risk assessment's predictions; *(ii)* did not adjust their level of trust in the risk assessment tool based on its performance; *(iii)* showed prejudice in their interactions with the risk assessment model. The issue of trust in AI has been studied since by multiple disciplines (Henrique and Santos 2024). Although such concerns apply also to abstaining/deferring predictors, to the best of our knowledge, only Bondi et al. (2022) examine the impact of using a selective predictor on human decision makers. The authors show that human performance can be enhanced by alerting the human about the decision to defer while withholding the prediction of the AI system. More work is needed for understanding the consequences of abstaining/deferring on human understanding and behavior from a multidisciplinary (technical, psychological, moral, legal) perspective.

Robustness. Robustness refers to an AI system's ability to maintain its integrity and operate effectively in challenging settings, e.g., under adversarial attacks, disturbances, and data contamination (Amodei et al. 2016). Abstaining predictors can enhance the robustness of AI systems. Preliminary works in this challenging objective include the following. Laidlaw and Feizi (2019) present a technique known as Combined Abstention Robustness Learning (CARL) to train a classifier that is both accurate and resilient. Pang et al. (2022) design a way for estimating a rectified confidence and for using it to train a selected classifier in an adversarial manner. Balcan et al. (2023) introduce a random feature subspace threat model and show that classifiers lacking the capability to reject are susceptible to this adversary. Chen et al. (2023) theoretically analyze the stratified rejection setting and devise a novel defense approach, called Adversarial Training with Consistent Prediction-based Rejection (CPR) for building an abstaining classifier that mitigates the exposure to adversarial attacks. On the other hand, however, abstaining predictors may exhibit robustness issues. For instance, due to the usage of hold out sets and non-robust statistics, the selection function may not be statistically robust, with instances being rejected as a consequence of randomness, outliers, missing values.

Conclusions

The capability to know what is not known is an expression of human intelligence, which lead to abstain from making a decision, or to defer to somebody else, or to reason over uncertainty estimates. We have surveyed approaches in the ML literature to build models that learn the above form of intelligence, and outlined a few open research issues. Beyond technical advancements, we highlight the need for risk-aware training of human decision makers supported by AI, so that AI system abstention, deferring, or uncertainty estimate can be properly evaluated and trusted.

Acknowledgements

Research partly funded 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.

References

Abdar, M.; Pourpanah, F.; Hussain, S.; Rezazadegan, D.; Liu, L.; Ghavamzadeh, M.; Fieguth, P. W.; Cao, X.; Khosravi, A.; Acharya, U. R.; Makarenkov, V.; and Nahavandi, S. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Inf. Fusion*, 76: 243–297.

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

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P. F.; Schulman, J.; and Mané, D. 2016. Concrete Problems in AI Safety. *CoRR*, abs/1606.06565.

Angelopoulos, A. N.; and Bates, S. 2023. Conformal Prediction: A Gentle Introduction. *Found. Trends Mach. Learn.*, 16(4): 494–591.

Artelt, A.; Brinkrolf, J.; Visser, R.; and Hammer, B. 2022. Explaining Reject Options of Learning Vector Quantization Classifiers. In *IJCCI*, 249–261. SCITEPRESS.

Artelt, A.; and Hammer, B. 2022. "Even if ..." - Diverse Semifactual Explanations of Reject. In *SSCI*, 854–859. IEEE.

Artelt, A.; Visser, R.; and Hammer, B. 2023. "I do not know! but why?" - Local model-agnostic example-based explanations of reject. *Neurocomputing*, 558: 126722.

Balcan, M.; Blum, A.; Sharma, D.; and Zhang, H. 2023. An Analysis of Robustness of Non-Lipschitz Networks. *J. Mach. Learn. Res.*, 24: 98:1–98:43.

Bondi, E.; Koster, R.; Sheahan, H.; Chadwick, M. J.; Bachrach, Y.; Cemgil, A. T.; Paquet, U.; and Dvijotham, K. 2022. Role of Human-AI Interaction in Selective Prediction. In *AAAI*, 5286–5294. AAAI Press.

Britto, A. S.; Sabourin, R.; and de Oliveira, L. E. S. 2014. Dynamic selection of classifiers - A comprehensive review. *Pattern Recognit.*, 47(11): 3665–3680.

Campos, M. M.; Farinhas, A.; Zerva, C.; Figueiredo, M. A. T.; and Martins, A. F. T. 2024. Conformal Prediction for Natural Language Processing: A Survey. *CoRR*, abs/2405.01976.

Cao, Y.; Mozannar, H.; Feng, L.; Wei, H.; and An, B. 2023. In Defense of Softmax Parametrization for Calibrated and Consistent Learning to Defer. In *NeurIPS*.

Charusaie, M.; Mozannar, H.; Sontag, D. A.; and Samadi, S. 2022. Sample Efficient Learning of Predictors that Complement Humans. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, 2972–3005. PMLR.

Chen, J.; Raghuram, J.; Choi, J.; Wu, X.; Liang, Y.; and Jha, S. 2023. Stratified Adversarial Robustness with Rejection. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, 4867–4894. PMLR.

Cheng, Q.; Sun, T.; Liu, X.; Zhang, W.; Yin, Z.; Li, S.; Li, L.; He, Z.; Chen, K.; and Qiu, X. 2024. Can AI Assistants Know What They Don't Know? In *ICML*. OpenReview.net.

Chow, C. K. 1965. Statistical Independence and Threshold Functions. *IEEE Trans. Electron. Comput.*, 14(1): 66–68.

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

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

Condessa, F.; Bioucas-Dias, J. M.; and Kovacevic, J. 2015. Robust hyperspectral image classification with rejection fields. In *WHISPERS*, 1–4. IEEE.

Conte, D.; Foggia, P.; Percannella, G.; Saggese, A.; and Vento, M. 2012. An Ensemble of Rejecting Classifiers for Anomaly Detection of Audio Events. In *AVSS*, 76–81. IEEE Computer Society.

Corbière, C.; Thome, N.; Bar-Hen, A.; Cord, M.; and Pérez, P. 2019. Addressing Failure Prediction by Learning Model Confidence. In *NeurIPS*, 2898–2909.

Cordella, L. P.; Stefano, C. D.; Sansone, C.; and Vento, M. 1995a. An Adaptive Reject Option for LVQ Classifiers. In *ICIAP*, volume 974 of *Lecture Notes in Computer Science*, 68–73. Springer.

Cordella, L. P.; Stefano, C. D.; Tortorella, F.; and Vento, M. 1995b. A method for improving classification reliability of multilayer perceptrons. *IEEE Trans. Neural Networks*, 6(5): 1140–1147.

Cortes, C.; DeSalvo, G.; and Mohri, M. 2016. Boosting with Abstention. In *NIPS*, 1660–1668.

Cortes, C.; DeSalvo, G.; and Mohri, M. 2024. Theory and algorithms for learning with rejection in binary classification. *Ann. Math. Artif. Intell.*, 92(2): 277–315.

Cruz, R. M. O.; Sabourin, R.; and Cavalcanti, G. D. C. 2018. Dynamic classifier selection: Recent advances and perspectives. *Inf. Fusion*, 41: 195–216.

De, A.; Koley, P.; Ganguly, N.; and Gomez-Rodriguez, M. 2020. Regression under Human Assistance. In *AAAI*, 2611–2620. AAAI Press.

De, A.; Okati, N.; Zarezade, A.; and Rodriguez, M. G. 2021. Classification Under Human Assistance. In *AAAI*, 5905–5913. AAAI Press.

De Toni, G.; Okati, N.; Thejaswi, S.; Straitouri, E.; and Gomez-Rodriguez, M. 2024. Towards Human-AI Complementarity with Predictions Sets. In *NeurIPS*.

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

dos Santos, E. M.; Sabourin, R.; and Maupin, P. 2007. Ambiguity-guided dynamic selection of ensemble of classifiers. In *FUSION*, 1–8. IEEE.

Drapal, P.; de Menezes e Silva Filho, T.; and Prudêncio, R. B. C. 2024. Meta-Learning and Novelty Detection for Machine Learning with Reject Option. In *IJCNN*, 1–8. IEEE.

Dubuisson, B.; and Masson, M. 1993. A statistical decision rule with incomplete knowledge about classes. *Pattern Recognit.*, 26(1): 155–165.

El-Yaniv, R.; and Wiener, Y. 2010. On the Foundations of Noise-free Selective Classification. *J. Mach. Learn. Res.*, 11: 1605–1641.

Feng, L.; Ahmed, M. O.; Hajimirsadeghi, H.; and Abdi, A. H. 2023. Towards Better Selective Classification. In *ICLR*. OpenReview.net.

Fischer, L.; Hammer, B.; and Wersing, H. 2016. Optimal local rejection for classifiers. *Neurocomputing*, 214: 445–457.

Franc, V.; Paplhám, J.; and Prruvsa, D. 2024. SCOD: From Heuristics to Theory. In *ECCV (84)*, volume 15142 of *Lecture Notes in Computer Science*, 424–441. Springer.

Franc, V.; Průša, D.; and Vorácek, V. 2023. Optimal Strategies for Reject Option Classifiers. *J. Mach. Learn. Res.*, 24: 11:1–11:49.

Fumera, G.; and Roli, F. 2002. Support Vector Machines with Embedded Reject Option. In *SVM*, volume 2388 of *Lecture Notes in Computer Science*, 68–82. Springer.

Gawlikowski, J.; Tassi, C. R. N.; Ali, M.; Lee, J.; Humt, M.; Feng, J.; Kruspe, A. M.; Triebel, R.; Jung, P.; Roscher, R.; Shahzad, M.; Yang, W.; Bamler, R.; and Zhu, X. 2023. A survey of uncertainty in deep neural networks. *Artif. Intell. Rev.*, 56(S1): 1513–1589.

Geifman, Y.; and El-Yaniv, R. 2017. Selective Classification for Deep Neural Networks. In *NIPS*, 4878–4887.

Geifman, Y.; and El-Yaniv, R. 2019. SelectiveNet: A Deep Neural Network with an Integrated Reject Option. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, 2151–2159. PMLR.

Grant, A. 2023. *Think Again: The Power of Knowing What You Don't Know*. Penguin Publishing Group.

Green, B.; and Chen, Y. 2019. The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proc. ACM Hum. Comput. Interact.*, 3(CSCW): 50:1–50:24.

Gruber, C.; Schenk, P. O.; Schierholz, M.; Kreuter, F.; and Kauermann, G. 2023. Sources of Uncertainty in Machine Learning - A Statisticians' View. *CoRR*, abs/2305.16703.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2019. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.*, 51(5): 93:1–93:42.

Hellman, M. E. 1970. The Nearest Neighbor Classification Rule with a Reject Option. *IEEE Trans. Syst. Sci. Cybern.*, 6(3): 179–185.

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

Henrique, B. M.; and Santos, E. 2024. Trust in artificial intelligence: Literature review and main path analysis. *Computers in Human Behavior: Artificial Humans*, 2(1): 100043.

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

Huang, L.; Zhang, C.; and Zhang, H. 2020. Self-Adaptive Training: beyond Empirical Risk Minimization. In *NeurIPS*.

Hüllermeier, E.; and Waegeman, W. 2021. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach. Learn.*, 110(3): 457–506.

Jitkrittum, W.; Gupta, N.; Menon, A. K.; Narasimhan, H.; Rawat, A. S.; and Kumar, S. 2023. When Does Confidence-Based Cascade Deferral Suffice? In *NeurIPS*.

Jones, E.; Sagawa, S.; Koh, P. W.; Kumar, A.; and Liang, P. 2021. Selective Classification Can Magnify Disparities Across Groups. In *ICLR*. OpenReview.net.

Kapoor, S.; Gruver, N.; Roberts, M.; Collins, K. M.; Pal, A.; Bhatt, U.; Weller, A.; Dooley, S.; Goldblum, M.; and Wilson, A. G. 2024. Large Language Models Must Be Taught to Know What They Don't Know. *CoRR*, abs/2406.08391.

Kilian, L.; and Lütkepohl, H. 2017. *Structural vector autoregressive analysis*. Cambridge University Press.

Kim, J.; Koo, J.; and Hwang, S. 2023. A unified benchmark for the unknown detection capability of deep neural networks. *Expert Syst. Appl.*, 229(Part A): 120461.

Kompa, B.; Snoek, J.; and Beam, A. L. 2021. Second opinion needed: communicating uncertainty in medical machine learning. *npj Digit. Medicine*, 4.

Kühne, J.; März, C.; et al. 2021. Securing deep learning models with autoencoder based anomaly detection. In *PHM Society European Conference*, volume 6, 13–13.

Kuncheva, L. 2002. A Theoretical Study on Six Classifier Fusion Strategies. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(2): 281–286.

Kutner, M. H.; Nachtsheim, C. J.; Neter, J.; and Li, W. 2004. *Applied Linear Statistical Models*. Cambridge University Press, 5 edition.

Laidlaw, C.; and Feizi, S. 2019. Playing it Safe: Adversarial Robustness with an Abstain Option. *CoRR*, abs/1911.11253.

Landgrebe, T. C. .; Tax, D. M. J.; Paclík, P.; Duin, R. P. W.; and Colin, A. 2004. A combining strategy for ill-defined problems. In *Fifteenth Annual Symposium of the Pattern Recognition Association of South Africa*, 57–62.

Lee, J. K.; Bu, Y.; Rajan, D.; Sattigeri, P.; Panda, R.; Das, S.; and Wornell, G. W. 2021. Fair Selective Classification Via Sufficiency. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, 6076–6086. PMLR.

Lenders, D.; Pugnana, A.; Pellungrini, R.; Calders, T.; Pedreschi, D.; and Giannotti, F. 2024. Interpretable and Fair Mechanisms for Abstaining Classifiers. In *ECML/PKDD (7)*, volume 14947 of *Lecture Notes in Computer Science*, 416–433. Springer.

Liang, S.; Li, Y.; and Srikant, R. 2018. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *ICLR (Poster)*. OpenReview.net.

Liu, S.; Cao, Y.; Zhang, Q.; Feng, L.; and An, B. 2024. Mitigating Underfitting in Learning to Defer with Consistent Losses. In *AISTATS*, volume 238 of *Proceedings of Machine Learning Research*, 4816–4824. PMLR.

Madras, D.; Pitassi, T.; and Zemel, R. S. 2018. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. In *NeurIPS*, 6150–6160.

Mao, A.; Mohri, C.; Mohri, M.; and Zhong, Y. 2023. Two-Stage Learning to Defer with Multiple Experts. In *NeurIPS*.

Mecke, L.; Buschek, D.; Gruenefeld, U.; and Alt, F. 2024. Exploring the Lands Between: A Method for Finding Differences between AI-Decisions and Human Ratings through Generated Samples. *arXiv preprint arXiv:2409.12801*.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2022. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.*, 54(6): 115:1–115:35.

Meinshausen, N. 2006. Quantile Regression Forests. *J. Mach. Learn. Res.*, 7: 983–999.

Minh, D.; Wang, H. X.; Li, Y. F.; and Nguyen, T. N. 2022. Explainable artificial intelligence: A comprehensive review. *Artif. Intell. Rev.*, 55(5): 3503–3568.

Mozannar, H.; Lang, H.; Wei, D.; Sattigeri, P.; Das, S.; and Sontag, D. A. 2023. Who Should Predict? Exact Algorithms For Learning to Defer to Humans. In *AISTATS*, volume 206 of *Proceedings of Machine Learning Research*, 10520–10545. PMLR.

Mozannar, H.; and Sontag, D. A. 2020. Consistent Estimators for Learning to Defer to an Expert. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, 7076–7087. PMLR.

Nalisnick, E. T.; Matsukawa, A.; Teh, Y. W.; Görür, D.; and Lakshminarayanan, B. 2019. Hybrid Models with Deep and Invertible Features. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, 4723–4732. PMLR.

Narasimhan, H.; Menon, A. K.; Jitkrittum, W.; and Kumar, S. 2024. Plugin estimators for selective classification with out-of-distribution detection. In *ICLR*. OpenReview.net.

Ntoutsi, E.; Fafalios, P.; Gadiraju, U.; Iosifidis, V.; Nejdl, W.; Vidal, M.; Ruggieri, S.; Turini, F.; Papadopoulos, S.; Krasanakis, E.; Kompatsiaris, I.; Kinder-Kurlanda, K.; Wagner, C.; Karimi, F.; Fernández, M.; Alani, H.; Berendt, B.; Kruegel, T.; Heinze, C.; Broelemann, K.; Kasneci, G.; Tiropanis, T.; and Staab, S. 2020. Bias in data-driven artificial intelligence systems - An introductory survey. *WIREs Data Mining Knowl. Discov.*, 10(3).

Okati, N.; De, A.; and Gomez-Rodriguez, M. 2021. Differentiable Learning Under Triage. In *NeurIPS*, 9140–9151.

Pang, T.; Zhang, H.; He, D.; Dong, Y.; Su, H.; Chen, W.; Zhu, J.; and Liu, T. 2022. Two Coupled Rejection Metrics Can Tell Adversarial Examples Apart. In *CVPR*, 15202–15212. IEEE.

Papadopoulos, H.; Proedrou, K.; Vovk, V.; and Gammerman, A. 2002. Inductive Confidence Machines for Regression. In *ECML*, volume 2430 of *Lecture Notes in Computer Science*, 345–356. Springer.

Pugnana, A.; Perini, L.; Davis, J.; and Ruggieri, S. 2024. Deep Neural Network Benchmarks for Selective Classification. *J. Data-centric Mach. Learn. Res.*, 1(17): 1–58.

Pugnana, A.; and Ruggieri, S. 2023a. AUC-based Selective Classification. In *AISTATS*, volume 206 of *Proceedings of Machine Learning Research*, 2494–2514. PMLR.

Pugnana, A.; and Ruggieri, S. 2023b. A Model-Agnostic Heuristics for Selective Classification. In *AAAI*, 9461–9469. AAAI Press.

Punzi, C.; Pellungrini, R.; Setzu, M.; Giannotti, F.; and Pedreschi, D. 2024. AI, Meet Human: Learning Paradigms for Hybrid Decision Making Systems. *CoRR*, abs/2402.06287.

Rastogi, C.; Zhang, Y.; Wei, D.; Varshney, K. R.; Dhurandhar, A.; and Tomsett, R. 2022. Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making. *Proc. ACM Hum. Comput. Interact.*, 6(CSCW1): 83:1–83:22.

Ruggieri, S.; Álvarez, J. M.; Pugnana, A.; State, L.; and Turini, F. 2023. Can We Trust Fair-AI? In *AAAI*, 15421–15430. AAAI Press.

Schmeing, E.; Brun, A. L.; and Silva, R. A. 2022. Dynamic selection of classifiers based on complexity measures. In *ICTAI*, 82–89. IEEE.

Schreuder, N.; and Chzhen, E. 2021. Classification with abstention but without disparities. In *UAI*, volume 161 of *Proceedings of Machine Learning Research*, 1227–1236. AUAI Press.

Shah, A.; Bu, Y.; Lee, J. K.; Das, S.; Panda, R.; Sattigeri, P.; and Wornell, G. W. 2022. Selective Regression under Fairness Criteria. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, 19598–19615. PMLR.

Silva Filho, T. d. M.; Song, H.; Perelló-Nieto, M.; Santos-Rodríguez, R.; Kull, M.; and Flach, P. A. 2023. Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Mach. Learn.*, 112(9): 3211–3260.

Smith, R. C. 2024. *Uncertainty Quantification: Theory, Implementation, and Applications, Second Edition*. SIAM, 2 edition.

Soize, C. 2017. *Uncertainty Quantification - An Accelerated Course with Advanced Applications in Computational Engineering*. Springer.

Straitouri, E.; Wang, L.; Okati, N.; and Rodriguez, M. G. 2023. Improving Expert Predictions with Conformal Prediction. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, 32633–32653. PMLR.

Sun, S. 2022. Conformal Methods for Quantifying Uncertainty in Spatiotemporal Data: A Survey. *CoRR*, abs/2209.03580.

Vailaya, A.; and Jain, A. K. 2000. Reject Option for VQ-Based Bayesian Classification. In *ICPR*, 2048–2051. IEEE Computer Society.

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

Verma, R.; Barrejón, D.; and Nalisnick, E. T. 2023. Learning to Defer to Multiple Experts: Consistent Surrogate Losses, Confidence Calibration, and Conformal Ensembles. In *AISTATS*, volume 206 of *Proceedings of Machine Learning Research*, 11415–11434. PMLR.

Verma, R.; and Nalisnick, E. T. 2022. Calibrated Learning to Defer with One-vs-All Classifiers. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, 22184–22202. PMLR.

Vovk, V.; Gammerman, A.; and Shafer, G. 2005. *Algorithmic Learning in a Random World*. Springer.

Wang, X.; and Yiu, S. 2020. Classification with Rejection: Scaling Generative Classifiers with Supervised Deep Infomax. In *IJCAI*, 2980–2986. ijcai.org.

Wei, Z.; Cao, Y.; and Feng, L. 2024. Exploiting Human-AI Dependence for Learning to Defer. In *ICML*. OpenReview.net.

Wozniak, M.; Graña, M.; and Corchado, E. 2014. A survey of multiple classifier systems as hybrid systems. *Inf. Fusion*, 16: 3–17.

Xia, G.; and Bouganis, C. 2022. Augmenting Softmax Information for Selective Classification with Out-of-Distribution Data. In *ACCV (6)*, volume 13846 of *Lecture Notes in Computer Science*, 664–680. Springer.

Yin, Z.; Sun, Q.; Guo, Q.; Wu, J.; Qiu, X.; and Huang, X. 2023. Do Large Language Models Know What They Don't Know? In *ACL (Findings)*, 8653–8665. Association for Computational Linguistics.

Zaoui, A.; Denis, C.; and Hebiri, M. 2020. Regression with reject option and application to kNN. In *NeurIPS*.

Zukerman, I.; and Maruf, S. 2024. Communicating Uncertainty in Explanations of the Outcomes of Machine Learning Models. In *INLG*, 30–46. Association for Computational Linguistics.