arXiv:2503.11709v1 [cs.LG] 12 Mar 2025

Conformal Prediction and Human Decision Making*

Jessica Hullman, Yifan Wu, Dawei Xie, Ziyang Guo, Andrew Gelman

7 Mar 2025

Abstract

Methods to quantify uncertainty in predictions from arbitrary models are in demand in high-stakes domains like medicine and finance. Conformal prediction has emerged as a popular method for producing a set of predictions with specified average coverage, in place of a single prediction and confidence value. However, the value of conformal prediction sets to assist human decisions remains elusive due to the murky relationship between coverage guarantees and decision makers' goals and strategies. How should we think about conformal prediction sets as a form of decision support? Under what conditions do we expect the support they provide to be superior versus inferior to that of alternative presentations of predictive uncertainty? We outline a decision theoretic framework for evaluating predictive uncertainty as informative signals, then contrast what can be said within this framework about idealized use of calibrated probabilities versus conformal prediction sets. Informed by prior empirical results and theories of human decisions under uncertainty, we formalize a set of possible strategies by which a decision maker might use a prediction set. We identify ways in which conformal prediction sets and posthoc predictive uncertainty quantification more broadly are in tension with common goals and needs in human-AI decision making. We give recommendations for future research in predictive uncertainty quantification to support human decision makers.

1. Quantifying uncertainty in black-box predictions

What does it mean to target a quantification of prediction uncertainty to a decision maker's task? Can we do this using general methods that work regardless of the specific decision setting? These statistical questions are as relevant as ever as machine learning models are deployed to assist human decision makers in domains like medicine, criminal justice, and finance. Methods for rigorously quantifying uncertainty in model predictions—even if the model is a black box—are in high demand.

Consider for example a telehealth doctor who can access predictions from a computer vision model while evaluating images of a skin condition submitted by a prospective patient. Presenting information about the model's confidence in its predictions is generally thought to help the human expert use the information more effectively. If, for example, the model assigns a moderate probability to a relatively rare but deadly condition, the doctor might advise them to schedule an appointment directly with a specialist rather than first seeing a primary care physician. Presenting uncertainty may also motivate "practico-epistemic" actions performed for the sake of epistemic gain, such as when the ambiguity about the true condition motivates the doctor to gather more evidence, as well as useful diversification in the choice of action [Manski, 2009], such that the doctor varies their advice across similar patients to balance expected false positives and false negatives in the face of ambiguity about the best treatment policy.

A common minimum goal for such settings is that a predictive model is *calibrated*: the rate at which events predicted by the model are realized approximately matches the probability assigned by the model to those events [Dawid, 1985], [Foster and Vohra, 1998]. Here we focus on multiclass prediction, where we assume a classification model $\hat{f}$ trained on historical data. When we apply the model to an input feature vector $\mathbf{x} = (x_1, x_2, ..., x_d) \in \mathcal{X} \subseteq \mathbb{R}^d$, we obtain a predicted label $\hat{f}(\mathbf{x}) = \hat{y} \in \mathcal{Y}$, but also have access to a predicted probability $\hat{p}_y(\mathbf{x}) = Pr(Y = y|\mathbf{x})$ for each possible label $y \in \mathcal{Y}$. Calibration in this setting means that for all predicted values $b \in [0, 1]$, for all pairs of

*We thank John Cherian, Tiffany Ding, and Jason Hartline for feedback on a draft.

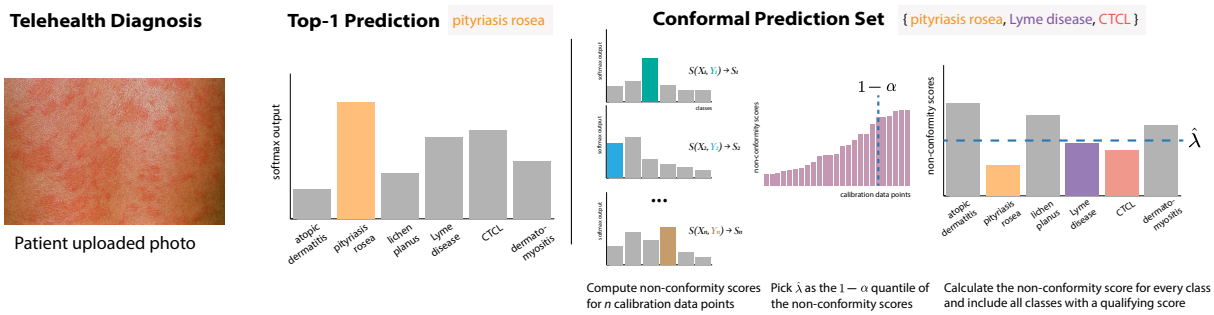


Figure 1: Conventional (top-1) prediction (left) compared to split conformal prediction (right) using a nonconformity score based on softmax output.

inputs and labels ($X = \mathbf{x}, Y = y$) for which $\hat{p}_y(\mathbf{x}) = b$, then $(b \times 100)\%$ of the time, y is the true label. Calibration is useful when we expect decisions to be based only on predictions, because even if decision makers weigh errors differently than the machine learning model, calibrated predictions allow them to trade off their true risk according to their own decision problem.

However, some machine learning paradigms result in predictive uncertainty estimates that are not calibrated. Deep learning models for multi-class classification, for example, produce pseudo-probabilities with predictions called softmax values. These are achieved by normalizing the activations produced by the last hidden layer in the neural network to a $[0, 1]$ range, but this normalization does not enforce calibration with reality or even internal consistency: these numbers when will not in general follow the laws of probability (that $\Pr(A \vee B) = \Pr(A) + \Pr(B)$ if the events A and B are disjoint, $\Pr(A \wedge B) = \Pr(A) \times \Pr(B)$ if they are independent). Beyond their lack of coherence, insufficient data, dataset biases, or implications of particular choices of architectures, hyperparameters, or model classes can make model confidence estimates vulnerable to overfitting, and thus unreliable for decision making [Guo et al., 2017]. Bayesian methods like Bayesian neural networks [Neal, 2012, Gal and Ghahramani, 2016] can provide more meaningful probabilities by approximating a distribution over network weights. However, their computational requirements in high-dimensional settings like images or text limit their use.

In theory, empirically calibrated probabilities can be also achieved by training models with a proper loss function [Blasiok et al., 2024], or applying calibration procedures [Hébert-Johnson et al., 2018, Guo et al., 2017, Gopalan et al., 2022] that learn adjustments to model-predicted probabilities to achieve calibration on held-out data. For example, recent work in theoretical computer science demonstrates the possibility of learning multicalibrated predictors–predictors that are simultaneously calibrated for all intersecting groupings of the data, not just marginally [Hébert-Johnson et al., 2018]–by iteratively post-processing predictions via a boosting procedure. However, the amount of data required to achieve such guarantees renders them impractical in common scenarios like multiclass classification with a moderately large label space.

Enter conformal prediction, a class of distribution-agnostic approaches for quantifying uncertainty in predictions that produces *sets* of predictions that promise to contain the true label with a user-defined probability $1 - \alpha$. In contrast to returning a single top prediction with the highest model confidence score (Figure 1 left), split conformal prediction algorithms return a set of predictions designed to achieve $1-\alpha$ coverage on average (Figure 1 right). In the common variant of split conformal prediction with marginal coverage, this set is constructed by comparing a score, representing the "non-conformity" of the new instance with what the model has learned, to a threshold chosen to achieve the expected coverage over the randomness in the selection of test points and a

set of held-out calibration instances for which ground truth is known.

Conformal prediction has been argued to provide meaningful uncertainty quantification to improve decisions from predictions in scenarios where models are deployed to assist doctors or other experts [Banerji et al., 2023, Kiyani et al., 2025, Lu et al., 2022, Straitouri et al., 2023]. Relative to a single top prediction with a confidence value, prediction sets may better convey to decision makers that multiple labels are consistent with the available information. By providing a finite set of possibilities with some probabilistic assurance, they can guide decision makers to focus their attention on a limited number of possibilities. For example, in the absence of a prediction set, which possibilities the telehealth doctor considers in the limited time they are given to assess each patient might depend on the vagaries of the cases they have seen in the recent past. Despite the difficulty of properly interpreting coverage guarantees based on long-run frequencies (as has been observed even with researchers for 95% confidence intervals [Belia et al., 2005, Hoekstra et al., 2014], doctors and other experts may at least find the notion of $1-\alpha$ coverage familiar for signaling uncertainty.

At the same time, a clear account of how exactly conformal prediction sets may impact human decisions relative to alternative presentations of predictive uncertainty remains lacking. While a calibrated probability has a straightforward translation as a decision aid, a prediction set offers no measure of the credibility of items within the set, leaving it in the hands of the decision maker to decide a strategy for choosing between them. Recent work observes that prediction sets are optimal for conservative decision makers who seek to act in a way that maximizes their worst case utility [Kiyani et al., 2025], but this strategy may or may not align with needs in a particular use case. We ask several broader questions with an eye toward how experts' domain knowledge may interact with prediction set design: What might an expert's decision processes with a conformal prediction set entail? How should conformal prediction algorithms be designed in light of common expectations that human experts will bring unique knowledge beyond that available to an AI model?

We outline a decision theoretic framework for evaluating predictive uncertainty as informative signals, then contrast what can be said within this framework about idealized use of calibrated probabilities versus conformal prediction sets. Informed by prior empirical results and theories of human decisions under uncertainty, we highlight how optimal use of a conformal prediction set by a decision maker is underdefined, and present possible strategies by which a decision maker might use a prediction set. Taking a step back, we consider what kinds of philosophical commitments conformal prediction and other post hoc calibration approaches entail, and reflect on how theoretical work on black-box uncertainty quantification can neglect to consider important practical features of an approach. We synthesize future directions for research aimed at better aligning conformal prediction sets with the needs of human decision makers.

This paper is organized as follows. First, we briefly summarize common variants of conformal prediction with an emphasis on split conformal prediction (Section 2). We introduce a decision-theoretic framework for evaluating decisions under uncertainty (Section 3), then consider the value of a calibrated prediction to a decision maker (Section 4), contrasting the case where they do and do not have access to information external to the model predictions. We consider several possible models of decision making with conformal prediction sets (Section 5).We conclude by discussing future directions for research on human-aligned predictive uncertainty (Section 6).

2. Split conformal prediction

Introduced by Vovk et al. [2005] as a distribution-free statistical framework for quantifying prediction uncertainty, conformal prediction is a model-agnostic approach to generating prediction sets (or intervals) expected to contain the true label (or real-valued response) with a user-defined probability. Formally, to predict the label $Y_{\text{new}} \in \mathcal{Y}$ of a new instance $X_{\text{new}} \in \mathcal{X}$ with some classifier $\hat{f}$

(where capital letters are used to denote the random variables), conformal prediction returns a set $\hat{\mathcal{C}}(X_{\text{new}}) \subseteq \mathcal{Y}$ such that *on average* the following coverage guarantee holds:

$$\Pr\left(Y_{\text{new}} \in \hat{\mathcal{C}}\left(X_{\text{new}}\right)\right) \geq 1-\alpha \tag{1}$$

where α is the user-specified error rate. This *marginal coverage* guarantee is expected to hold over the randomness in the selection of test points and data used by the algorithm to achieve the guarantee. While coverage could be achieved trivially by outputting the entire label space as a prediction set, ideally the set size is not larger than necessary and is adaptive in the sense that larger sets signal higher predictive uncertainty.

Split (a.k.a., *inductive*) conformal prediction [Vovk et al., 2005, Angelopoulos and Bates, 2023] for classification, which we focus on here, uses a held-out calibration set to post-process model predictions, generating sets expected to achieve marginal coverage at the target level. Given a classifier $\hat{f}$ trained on a training dataset $\mathcal{D}_{\text{tr}}$, we identify a calibration dataset of the form $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^n$ that is disjoint with $\mathcal{D}_{\text{tr}}$ (i.e., $\mathcal{D}_{\text{tr}} \cap \mathcal{D}_{\text{cal}} = \emptyset$). No restriction is placed on the relationship between $\mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{\text{cal}}$; they may be generated by different distributions.

Next, we select a heuristic of predictive uncertainty that can be obtained from $\hat{f}$. For example, we might use $\hat{p}_{y_i}(\mathbf{x}) = \sigma(\mathbf{z})_i$, the softmax probability for classifying $\mathbf{x}$ as y_i, where i also indexes the corresponding class output z_i in the model's last layer outputs $\mathbf{z}$[1]. Typically, we assume $\hat{p}(\mathbf{x})$ to be the vector of class probabilities for $\mathbf{x}$ after temperature scaling (a.k.a. Platt scaling) [Platt, 1999, Guo et al., 2017]. This heuristic measure of uncertainty is transformed to create a *non-conformity score* function $S : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ that can be applied to any pairing of instance features $\mathbf{x}$ and true label y; for example, $S(\mathbf{x}, y) = 1 - \hat{p}_y(\mathbf{x})$.

We then estimate the $1-\alpha$ quantile of the n non-conformity scores $S_1, \ldots, S_n$ and their corresponding labels from $\mathcal{D}_{\text{cal}}$. Specifically, a threshold $\hat{\lambda}$ is set to be the $\lceil (1-\alpha)(n+1) \rceil / n$ quantile[2] to correct the threshold for the finite sample. The prediction set for X_{new} can then be constructed by including all labels with a non-conformity score less than or equal to $\hat{\lambda}$:

$$\hat{\mathcal{C}}\left(X_{\text{new}}\right) = \left\{ y \in \mathcal{Y} : S\left(X_{\text{new}}, y\right) \leq \hat{\lambda} \right\} \tag{2}$$

The key assumption behind split conformal prediction's coverage promise is that the data in $\mathcal{D}_{\text{cal}} \cup \{(X_{\text{new}}, Y_{\text{new}})\}$ are statistically *exchangeable*, such that permuting the order of these data points leaves their joint distribution unchanged.

Since the data points in $\mathcal{D}_{\text{cal}} \cup \{(X_{\text{new}}, Y_{\text{new}})\}$ are exchangeable, the same conditions hold for their non-conformity scores $\{S_i\}_{i=1}^n \cup \{S_{\text{new}}\}$. To get an intuition for why this is sufficient to achieve coverage, consider what it means for random variables $S_1, S_2, ..., S_{n+1}$ to be exchangeable (where S_{n+1} represents S_{new}). Exchangeability means that the rank of the random variable S_{n+1} is equally likely to take any value from $1, ..., n+1$. This implies that the probability that S_{n+1} is less than or equal to the $\lceil (1-\alpha)(n+1) \rceil$ smallest score of $\{S_i\}_{i=1}^n \cup \{S_{n+1}\}$ is greater than or equal to $1-\alpha$. Further, if the previous holds this means that S_{n+1} is less than or equal to the $\lceil (1-\alpha)(n+1) \rceil$ smallest score of $\{S_i\}_{i=1}^n$. Consequently, the probability that S_{n+1} is less than or equal to the $\lceil (1-\alpha)(n+1) \rceil$ smallest of $S_1, S_2, ..., S_n$ is greater than or equal to $1-\alpha$ over repeated sampling of the calibration and test instances.

Before proceeding, we make a few remarks about key components of conformal prediction.

[1] Given a vector of unnormalized log probabilities $\mathbf{z} = [z_1, ..., z_k]$ representing outputs of the neural network's last layer for an input $\mathbf{x}$ and classes $[y_1, ..., y_k]$, the softmax probability $\sigma(\mathbf{z})_i$ for a class indexed by i is $\frac{e^{\mathbf{z}_i}}{\sum_{j=1}^k e^{\mathbf{z}_j}}$.

[2] $\hat{\lambda}$ is set to be the $\lceil (1-\alpha)(n+1) \rceil / n$ quantile when $\alpha \geq 1/(n+1)$; otherwise $\hat{\lambda}$ is set to $+\infty$.

Score function. No formal restrictions are put on the score function used to construct sets, yet the score function should empirically capture the uncertainty around model predictions such that larger scores signal higher uncertainty. However, the usefulness of the prediction sets will naturally vary with the informativeness of the score function. For neural networks, the score function is commonly derived from the model outputted softmax scores. The simplest such score function is simply $1-\hat{p}_y(\mathbf{x})$. Alternatively, to improve the adaptability of the sets–the relationship between set size and the difficulty of the instance–one can use score functions that use the cumulative softmax probabilities ranked in descending order, i.e., from the most likely class to the least likely class, until the true label is reached [Romano et al., 2020, Angelopoulos et al., 2021].

Exchangeability. An obvious concern in practice is that exchangeability may be violated, such as when the covariate, label, or posterior distribution shifts over time. This possibility inspires conformal methods that learn weighting schemes by leveraging additional knowledge about data shifts [Tibshirani et al., 2019, Podkopaev and Ramdas, 2021, Barber et al., 2023, Prinster et al., 2024]. A distinct line of work focuses on online conformal prediction algorithms [Gibbs and Candès, 2021, Feldman et al., 2023, Bhatnagar et al., 2023, Gibbs and Candès, 2024, Angelopoulos et al., 2024] that adapt regret minimization from online learning to employ adaptive thresholds that are adjusted upon observing the miscoverage of recent instances.

Marginal coverage and beyond. The marginal coverage guarantee associated with split conformal prediction provides a sort of compromise in light of the difficulty of achieving calibrated prediction probabilities conditional on groupings that can be defined on the feature space $\mathcal{X}$. The expected $1-\alpha$ coverage probability of vanilla split conformal prediction is over the random selection of calibration instances and test points. This is a weaker condition than coverage conditioned on a specific instance:

$$\Pr\left(Y_{\text{new}} \in \hat{\mathcal{C}}\left(X_{\text{new}}\right) \mid X_{\text{new}} = \mathbf{x}\right) \geq 1-\alpha, \text{ for any distribution, for almost all } \mathbf{x} \tag{3}$$

Conditional coverage is not possible without making distributional assumptions [Vovk, 2012, Barber et al., 2021]. However, decision makers may naturally desire instance-specific uncertainty, and in the applied literature, split conformal prediction is occasionally misinterpreted as providing "personalized" or "individualized" measures of uncertainty (e.g., [Banerji et al., 2023, Garcia-Galindo et al., 2023]). However, the coverage of split conformal prediction actually describes what we expect on unlimited test instances given *repeated resampling* of the calibration set from the same distribution over $\mathcal{X} \times \mathcal{Y}$. The deviation from expected coverage in any particular application conditional on $\mathcal{D}_{\text{cal}}$ shrinks with rate $n^{-1/2}$ as the size of $\mathcal{D}_{\text{cal}}$ increases. Marginal coverage also means that the attained coverage can vary considerably over regions of the feature space, such that some combinations of features are overcovered and others undercovered. Both facts are overlooked when the $1-\alpha$ coverage guarantee is mistakenly described as applying to the specific prediction set. These errors parallel misinterpretations of the stated confidence level (e.g., 95%) of Frequentist confidence intervals as applying to the specific interval at hand [Hoekstra et al., 2014].

To alleviate practical limitations of marginal coverage, multiple variations on conformal prediction [Barber et al., 2021, Jung et al., 2023, Gibbs et al., 2023] pursue an approximate version of conditional coverage where the coverage guarantee is conditional on the new instance's group membership, rather than the instance itself, i.e.,

$$\Pr\left(Y_{\text{new}} \in \hat{\mathcal{C}}\left(X_{\text{new}}\right) \mid X_{\text{new}} \in G\right) = 1-\alpha, \ \forall G \in \mathcal{G} \tag{4}$$

where $\mathcal{G} \subseteq 2^{\mathcal{X}}$ is some collection of pre-specified groups. As in the case of multicalibration algorithms, this stronger form of coverage requires more data to achieve.

3. Preliminaries: Evaluating decision making under uncertainty

To justify methods that produce quantified prediction uncertainty for decision making requires a way to evaluate the quality of decisions, and to appraise the value of specific information for a decision task. What characterizes a good measure of decision quality?

We can get some intuition for what is needed by considering how simple approaches fall short. Consider the use of post hoc decision accuracy, a common approach in research on AI-assisted decision-making. This approach measures the accuracy of a human decision-maker's final decisions given access to instance information as well as the human's own prediction and the prediction of a model for each instance. For example, in the case of the telehealth doctor making a diagnosis decision given a patient's uploaded photo of their condition, one might look at the proportion of cases in some labeled set of decision trials where the doctor chose the correct diagnosis after viewing the instance and model prediction.

There are two reasons post hoc accuracy is problematic as a general measure of decision quality. First, evaluating a decision against the realized outcome is problematic whenever outcomes are probabilistic. To see this, consider how the doctor might correctly observe that conditional on features present in the the image, melanoma is the most likely diagnosis (i.e., $\Pr(y = \text{melanoma}|\mathbf{x}) > \Pr(y = y_0|\mathbf{x}), \forall y_0 \in \mathcal{Y}$ s.t. $y_0 \neq$ melanoma). However, it could be the case that while the observed features most often co-present with melanoma, they are also associated with psoriasis, and psoriasis happens to be the true diagnosis for this patient. Saying that the doctor is wrong for acting as if melanoma is the diagnosis based on decision accuracy in hindsight fails to consider that the beliefs they arrived at were rational—that is, they captured the probability of each diagnosis under the true data-generating process—given the information available at the time.

Second, using post hoc accuracy to measure decision quality does not allow for possibly different costs of different types of errors in a decision scenario. If the cost of not testing for melanoma is high (because valuable time is lost in the event that it is the true diagnosis), then even when psoriasis is the more likely condition a priori, a doctor may still be better off following up as if melanoma were the diagnosis.

Statistical decision theory [Wald, 1949, Savage, 1972] provides a framework for evaluating decisions that accounts for both of these considerations. Below we define a decision problem and show how a definition of normative decision-making can be used to assess the value of predictive uncertainty information.

3.1. Definition of a decision problem

We assume a decision maker who seeks the action that maximizes their expected reward under a given utility function and data-generating process. Decisions are evaluated based on their expected utility, accounting for uncertainty about the outcome.

Formally, a decision problem assumes a *state* of the world, modeled as a random variable S, whose realized value is from a discrete set of possible (mutually-exclusive) states ($s \in \mathcal{S}$). The realized state is unknown at the time of the decision, but follows a distribution $p(S)$. The decision maker faces a choice of *action* from a set of possible actions, $a \in \mathcal{A}$. Decision quality is assigned by a loss function that assigns a numeric score representing our loss from making a given decision under a given (realized) state of the world: $L : \mathcal{A} \times \mathcal{S} \to \mathbb{R}$.

This abstraction provides sufficient flexibility to model a variety of decision problems. We can incorporate actions that involve epistemically-motivated behaviors like collecting additional data by including abstention in the action space. For example, in the example of medical diagnosis, we might imagine a two stage decision problem where the first action set includes an option to abstain

from diagnosing in order to gather more information, which carries some cost that increases the ultimate decision loss.

Next we consider the structure to the information that is available to the decision maker at decision time (such as the instance features, a model prediction, and a quantification of the prediction uncertainty). Specifically, we define the *data-generating model* p, the joint distribution of the signals and state over $\mathcal{V} \times \mathcal{S}$ from which the distribution of the state $p(S)$ is derived. This joint distribution assigns to each possible pairing of signal $V = v$ and state $S = s$ a probability $p(v, s)$. Slightly abusing the notation, we use $p(V)$ to denote the marginal distribution of V with expectation over S and $p(S)$ to denote the marginal distribution of S with expectation over V. We can thus use $p(s)$ and $p(v)$ to represent the marginal probabilities of $S = s$ and $V = v$ respectively.

3.1.1. Decision rule

Given a decision problem, a decision strategy or rule is a function that maps from a signal representing the information available to the decision-maker to a specific choice of action. The space of all possible decision strategies is intractable to reason about. Instead, researchers are often interested in either identifying what kinds of simple constraints characterize normative decision-making, which we define below, or in empirically describing decision rules that human decision-makers employ.

3.2. Using normative decision-making to quantify the value of predictive uncertainty

The joint distribution p captures the information structure of the decision problem. We can identify the best possible performance of any decision maker on a decision problem by considering how well we expect an idealized agent who knows the true data-generating model to do. This is captured by the expected loss of a Bayesian rational decision maker who has knowledge of the data-generating model and access to the signal for each particular instance. This decision maker starts with knowledge of the prior over states $p(S)$, processes the signals to extract the relevant information about the uncertain state, and updates their prior to arrive at posterior beliefs for each state $s \in \mathcal{S}$:

$$p(s|v) = \frac{p(v, s)}{\sum_{s \in \mathcal{S}} p(v, s)} \tag{5}$$

They select the loss-minimizing action under those beliefs:

$$a_{opt}(v) = \arg\min_{a \in \mathcal{A}} \mathbb{E}_{S \sim p(\cdot|v)}(L(a, S)) \tag{6}$$

where $p(\cdot|v)$ is the posterior over states. These assumptions can be used to derive performance benchmarks defined over a decision problem [Wu et al., 2023]. The best attainable performance of any decision maker facing the decision problem is upper bounded by the expected loss of the rational Bayesian agent with access to the signal:

$$\mathrm{R}(p, L) = \mathbb{E}_{V,S \sim p}(L(a_{opt}(V), S)) \tag{7}$$

Variations on this approach relax the assumption that the decision maker knows the DGM, and instead calculate the expected performance of a rational decision maker given some approximation $\hat{p}$ of p based on the available information (such as a set of labeled examples that the decision maker has previously seen) [Hullman et al., 2024].

We can use this Bayes rational benchmark to evaluate different forms of uncertainty information we might present to decision makers ex ante. We quantify the value of information Δ presented by

signal V as the difference in the expected loss of the rational decision maker when they have access to that information (as captured by Equation 7) versus when they do not:

$$\Delta(p, L) = \mathrm{R}_{\varnothing}(p, L) - \mathrm{R}(p, L) \tag{8}$$

where the rational baseline $\mathrm{R}_{\varnothing}$ Wu et al. 2023 is the expected loss of the rational Bayesian decision maker who must choose the best fixed action under knowledge of only the prior $p(S)$:

$$\mathrm{R}_{\varnothing}(p, L) = \min_{a \in \mathcal{A}} \mathbb{E}_{S \sim p(\cdot)}[L(a, S)] \tag{9}$$

The value of information Δ quantifies the "room" for reduction in expected loss associated with a particular type of signal. We can get a sense of the ex ante maximum possible "relevance" of different forms of prediction uncertainty by looking at how the expected loss of the idealized decision maker changes as we change the nature of the predictive uncertainty included in the signal.

Below, we use the definition of a decision problem, decision rule, and rational decision-maker to contrast decision strategies for calibrated probabilities to decision strategies for prediction sets. We consider perspectives based in normative models as well as by empirical descriptions of human decision-making under uncertainty.

4. Decisions from calibrated probabilities

Note that the lower bound on the expected loss of an idealized rational decision maker given some signaling strategy V (Equation 7) assumes that the decision maker arrives at posterior beliefs (Equation 5) using their knowledge of the joint distribution p. However, decision makers in the world will not generally possess perfect information about the data-generating model. Consider a Bayes rational decision maker who must decide given only the signal for each instance, without knowing the data-generating model. What is the ideal form of uncertainty information to provide such a decision maker?

From the above formulation of idealized decision making, it follows that the posterior probability distribution over states given the signal—$p(S|v)$—is sufficient information for such a decision maker to optimize. This is because when we expect the decision maker to choose the action with the highest expected reward conditional on these beliefs, this information is equivalent to telling them the action that maximizes their payoff. Hence, among signals conveying probabilistic information about the state, the posterior probability of the state is maximally informative.

This fact has implications for the value of calibrated predictions for decision making. Imagine that for each individual case we will predict on, there is a true probability distribution over outcomes, from which the realized outcome is drawn. When predicted probabilities are calibrated, they are "outcome indistinguishable" Dwork et al. 2021 from the true distribution, meaning that even given full access to the predicted model and the historical outcomes, it is not possible to invalidate the model's predictions.[3] For this reason, computer science theorists associate outcome indistinguishability with trustworthiness of predictions for decision making (e.g., Gopalan et al. 2021, Hu and Wu, 2024, Zhao et al., 2021).

But the ex ante value that we expect outcome indistinguishability to afford calibrated predictions for decision makers is not without assumptions. We must distinguish between two types of

[3] More precisely, a predictor $\hat{p}$ is outcome indistinguishable with respect to a family of tests $\mathcal{T}$ if samples from the true distribution $(i, y_i) \sim D^*$ cannot be distinguished by $\mathcal{T}$ from samples from the predictor's distribution $(i, \hat{y}_i) \sim D(\hat{p})$. In other words, for each test $T \in \mathcal{T}$, the probability that T outputs 1 is (nearly) the same on samples from D^* and $D(\hat{p})$.

settings: those where the decision maker does not have additional knowledge over that available to the model, and those where they do. Only in the former does a calibrated probability directly ensure trustworthiness for decision making.

To ground the discussion, consider the example setting of the telehealth doctor facing a choice of diagnosis, where the uncertain state $s \in \mathcal{S}$ is the true label y from label space $\mathcal{Y}$ and the action space $\mathcal{A}$ is $\mathcal{Y}$. They make their choice in light of information about the specific instance $\mathbf{x} \in \mathcal{X}$, associated with true label y generated by some true process $p \in P(\mathcal{X} \times \mathcal{Y})$. We might expect that based on their domain experience, the human has some internal approximating model of the data-generating process which enables them to make a prediction y^H about the label. They also have access to the prediction y^{AI} output by a model (or AI) trained on what is deemed to be relevant prior information.

In the first setting, the decision maker does not have additional knowledge over that available to the model. In other words, they are restricted to having at most equivalent decision-relevant information to the predictive model. Both the human prediction and the AI prediction are generated from a common set of features: $(X, Y^H) \sim p^H$ and $(X, Y^{AI}) \sim p^{AI}$, where $p^H, p^{AI} \in P(\mathcal{X} \times \mathcal{Y})$. Or, the human may only have access to some subset of the features that the model accesses, such as when the feature space is high dimensional as in classifying images or text, i.e., $p^H \in P(\mathcal{X}^H \times \mathcal{Y})$ where $\mathcal{X}^H \subseteq \mathcal{X}$. This setting has been well studied under the guise of clinical versus statistical prediction [Meehl, 1954], where a number of empirical studies have found human decisions tend to underperform those of statistical models on average when the test instances are drawn from the same distribution [Ægisdóttir et al., 2006, Grove et al., 2000]. This finding should not come as a surprise: when statistical models are evaluated on the same distribution on which they are trained to predict, and the human possesses no additional information, statistical prediction via loss minimization defines the standard for efficient use of the information.

In settings where we are confident that these conditions hold, we should expect the best decision quality from automation. However, full automation may not be an option for reasons of liability. The healthcare organization may not be comfortable, for example, making telehealth referral decisions about patients' conditions without a human doctor in the loop. Consequently, the goal becomes how to help the human realize that it is in their best interest to trust the model predictions. Ideally, if we provide them with calibrated probabilities, they will use these directly to optimize.

In addition to being critical to arguments that calibration makes predictions trustworthy for decision making, the assumption that the decision maker does *not* have access to relevant additional information also implicitly motivates broader arguments that providing model-generated uncertainty information to users of predictive models supports appropriate trust [Bhatt et al., 2021]. As in the explainable AI literature, however, this assumption is rarely made explicit.

4.1. Decisions under calibrated probabilities plus private signals

A challenge with the assumptions of a decision maker who relies only on the predictions of a model is that often in practice, we *want* the decision maker to consider other signals as well. What happens if a human expert is expected to have access to additional information beyond that available to train the model? In many scenarios where experts might be helped by predictive models, such as medicine or criminal justice, there is a latent belief that the expert *will* have access to further information. A doctor or judge, for example, may interact with a patient or defendant to gather information that is not available in their record. It is often speculated that they may observe signals—like the patient's attitude or the defendant's expressions of remorse—that contain decision-relevant information, but which are hard to measure or formalize [Meehl, 1954, Collina

et al., 2024, Guo et al., 2024b, Steyvers et al., 2022, Wilder et al., 2020]. Until very recently, few attempts have been made to provide organizations with sufficient methods to identify when human decision makers appear to have additional information that could lead to complementary performance with an AI [Alur et al., 2024, Guo et al., 2024b,a]. Hence it is often difficult to rule out this setting in practice.

Consider a decision maker with access to some auxiliary private information (i.e., set of features) $\mathbf{w} \in \mathcal{W} \subseteq \mathbb{R}^d$ for each instance, where $\mathcal{W} \cap \mathcal{X} = \emptyset$. Hence, the decision maker has access to instance information $X^H = \{X, W\}$, and their predictions $(X^H, Y^H) \sim p^H$ with $p^H \in P(\mathcal{X}^H \times \mathcal{Y})$ and associated posterior distribution $p^H(Y|X^H)$ are generated from this full set of information. Imagine that at decision time, given instance information $\mathbf{x}$ they have access to a display that also provides them with the model's top predicted class and confidence: $v = \{\mathbf{x}, \hat{y}, p^{AI}(\hat{y}|\mathbf{x})\}$. For simplicity, imagine a binary diagnosis problem such that $\mathcal{Y} = \{0, 1\}$ such that $p^{AI}(y|\mathbf{x})$ is the predicted probability that the patient has the disease given instance $\mathbf{x}$. Assume $p^{AI}(Y|X)$ is calibrated on data-generating process p. If $p^{AI}(y|\mathbf{x})$ was the same as the probability conditional on all of the available information ($p^{AI}(y|\mathbf{x}^H) = p^{AI}(y|\mathbf{x}, \mathbf{w})$), the utility-maximizing decision would correspond to placing a threshold $j^* \in [0, 1]$ on the predicted probability, representing the point at which the expected value of choosing $a = 1$ is equivalent to that of $a = 0$.

However, because the model's predictions are only calibrated over X, not necessarily over X^H, this will not necessarily be the case, and the decision maker can end up making a worse decision by trusting the probability over their own information. For example, imagine that the decision maker's private information W correlates more strongly with the state than does X. Then we would expect thresholding the model's predicted probability to result in lost utility. The benchmark in Equation 7 corresponding to signals taking the form $v = \{\mathbf{x}, \hat{y}, p^H(y|\mathbf{x}^H)\}$ will be greater than or equal to that corresponding to signals taking the form $v = \{\mathbf{x}, \hat{y}, p^{AI}(y|\mathbf{x})\}$. The biggest possible loss we could incur from showing the model predicted probability rather than the correct probability corresponds to the case where the private information is perfectly revealing of the state ($W = Y$) whereas the unidimensional feature X is uninformative.

Can this possibility be avoided? When the decision maker has access to private signals in addition to the signals available to the prediction model, the maximally informative signal for a rational decision maker combines the auxiliary private information with the information provided by an AI prediction. In other words, a maximally informative predictor is multi-calibrated with respect to both the human's private information W and the features X that both the human and the AI can access. In the case where the information in the feature space is hard to extract, e.g. with high-dimensional $\mathcal{W}$ and $\mathcal{X}$ such as image data, achieving multi-calibration with respect to both the human's predictor $p^H(Y|X^H)$ and the AI's predictor $p^{AI}(Y|X)$ may be more realistic than multi-calibration over W and X. A predictor p is multi-calibrated with respect to predictors p^H and p^{AI} if it is conditionally correct:

$$p\left(y|p^H, p^{AI}, p\right) = p. \tag{10}$$

4.2. Decisions under miscalibrated probabilities

Of course, in practice, model confidence scores will not necessarily equal the true probabilities. One reason is unavoidable sampling error from finite data. Calibration requires sample complexity exponential in the number of classes [Zhao et al., 2021] in multi-class classification. When perfect calibration is impossible, a question arises: Can a decision maker do better than acting as if the predictions are true if they are given information about the calibration error bound?

Calibration error measures how "far" a predictor is from perfect calibration. For decision purposes, the hope is that calibration error captures the loss a decision maker incurs if they trust

miscalibrated predictions. Thus, when the decision maker is told predictions have a low calibration error, she can trust the prediction without worrying about incurring high loss on the decision task. Recent theoretical work explores the implications for decision making of popular calibration error metrics. For example, Expected Calibration Error (ECE), which captures the weighted average of prediction errors, has been a popular choice of metric within machine learning research for optimization and benchmarking due to its mathematical simplicity (e.g., Guo et al. [2017], Minderer et al. [2021]). ECE upper bounds the decision loss for every decision task with bounded payoff Kleinberg et al. [2023]. However, ECE is only a loose upper bound of decision loss; Hu and Wu [2024] propose an alternative calibration error that directly relates to and tightly upper bounds decision loss. For the purposes of our discussion, the point is that given an error measure that sufficiently captures decision loss, it becomes straightforward to anticipate decision loss given expected miscalibration.

4.3. Takeaways

The use of calibrated predictions in decision making is complicated by the decision maker having private signals or knowledge of calibration error. However, even under these constraints, it is relatively straightforward to predict how a rational decision maker will use the predictions as a result of the alignment between good decision making and posterior probabilities. As we discuss next, this is not the case when the goal instead is to provide calibrated prediction sets.

5. Decisions from conformal prediction sets

We now consider how the decision maker uses a conformal prediction set. In contrast to the well-defined use of calibrated probabilities by a rational decision maker, a signal that takes the form of a conformal prediction set ($v = \hat{\mathcal{C}}(X_{\text{new}} = \mathbf{x}) \subset \mathcal{Y}$) provides no measure of which elements of the set are more or less credible. This presents a challenge to assessing their value to decision makers, even in the simpler case where the decision maker does not possess private signals.

Informally, we might distinguish a few broad strokes ways in which a decision maker responds to a prediction set. Perhaps they use it to *verify* their own judgment about the best course of action. For example, some participants who used prediction sets in an AI-advised image labeling task studied by Zhang et al. [2024] described coming up with their own judgment on the correct label, then checking whether it was included in the set. Another approach reported by participants from the same study used the set to *constrain* the labels they considered as they assessed the instance (in this case, an image to be classified). Similarly, Lu et al. [2022] characterized two strategies for using prediction sets based on interviews with clinicians: to rule out critical conditions that would make a treatment plan risky (such as giving a certain medication to a patient for which it might cause complications), and to rule in critical conditions to distinguish which patients should remain in the hospital.

On the other hand, despite prediction sets not including a measure of credibility over elements, we can lower bound the loss attainable by any decision maker by assuming a Bayes rational agent who possesses information about the joint distribution over signals and states, and uses the joint distribution to update its beliefs upon viewing a set.

Below, we formalize a set of strategies that a decision maker might use to select a label from a prediction set, distinguishing between strategies that assume knowledge of the joint distribution over signals and states, strategies that combine the decision maker's prior knowledge with the set in a non-Bayesian way, and strategies where the set constrains the decision maker's action.

Benchmark	Signals	Belief updating or decision strategy
Oracle		
Bayesian DM who knows $p \in P(\mathcal{X} \times \mathcal{Y})$	$v = \{\mathbf{x}, \hat{\mathcal{C}}(\mathbf{x})\}$	$p(y\|v) = \frac{p(\mathbf{x},y)}{\sum_{y' \in \mathcal{Y}} p(\mathbf{x},y')}$
Bayesian DM who knows $p \in P(2^{\mathcal{Y}} \times \mathcal{Y})$	$v = \hat{\mathcal{C}}(\mathbf{x})$	$p(y\|v) = \frac{p(\hat{\mathcal{C}}(\mathbf{x}),y)}{\sum_{y' \in \mathcal{Y}} p(\hat{\mathcal{C}}(\mathbf{x}),y')}$
Use set to verify a priori judgments		
Non-Bayesian DM with prior and misspecified beliefs $p \in P(\mathcal{Y})$	$v = \hat{\mathcal{C}}(\mathbf{x})$	$p(y\|v) = \begin{cases} \frac{p(y)}{\sum_{y' \in v} p(y')} \cdot (1-\alpha) & y \in v \\ \frac{p(y)}{\sum_{y' \in \mathcal{Y} \setminus v} p(y')} \cdot \alpha & y \notin v \end{cases}$
Associative non-Bayesian DM with prior $p(\mathcal{Y})$ and similarity function $d : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$	$v = \{y : y \in \mathcal{Y} \wedge \exists y' \in \hat{\mathcal{C}}(\mathbf{x})$ such that $d(y,y') \leq b\}$	Same as non-Bayesian DM
Use set to constrain decision options		
Uncertainty suppressing DM divides probability over set	$v = \hat{\mathcal{C}}(\mathbf{x})$	$p(y\|v) = \begin{cases} \frac{1}{\|v\|} & y \in v \\ 0 & y \notin v \end{cases}$
Conservative risk-averse DM	$v = \hat{\mathcal{C}}(\mathbf{x})$	$a^*(\hat{\mathcal{C}}) =$ $\arg\min_{a' \in \mathcal{A}} \max_{y \in \hat{\mathcal{C}}} L(a', y)$
Baseline with no signals		
Bayesian DM who knows only prior $p \in P(\mathcal{Y})$	No signal	$p(y\|v) = p(y)$

Table 1: Overview of signals and associated strategies for using prediction set information. The ranking of benchmarks denoting the best possible performance in each scenario (Equation 7) is partially predictable. Row 1 dominates row 2 and row 2 dominates all other rows. However, where the non-Bayesian decision makers fall relative to the Bayesian decision maker with only the prior (bottom row) depends on the structure of the specific setting.

5.1. Decision maker with oracle access to joint distribution over signals and states

We first consider, What is the best possible expected performance of any decision maker who is provided with prediction sets as decision support? We define two benchmarks that assume oracle access to joint probability information connecting the signals and the state. The first assumes the decision maker has access to joint information linking the state to the features. The second assumes that the decision maker has access to joint information linking the state to the prediction sets.

Lower bound: Bayesian decision maker who knows joint distribution over features and outcomes

A natural signal structure to consider presents the decision maker with instance features $\mathbf{x}$ and the corresponding prediction set $v = \{\mathbf{x}, \hat{\mathcal{C}}(\mathbf{x})\}$ from some conformal algorithm.

When the instance information $\mathbf{x}$ is included in the signal, then the best attainable performance by any decision maker with signals of the form $v = \{\mathbf{x}, \hat{\mathcal{C}}(\mathbf{x})\}$ is that of the Bayes rational decision maker who knows $p \in P(\mathcal{X} \times \mathcal{Y})$. However, once joint information about the instance features and state are known, prediction sets can offer no additional information. For a decision maker with knowledge of the joint distribution over X and Y, the probability $p(\mathbf{x}, y)$ already contains all relevant information needed to arrive at the Bayes posterior beliefs 5, thus the benchmark is a strict lower bound that is unaffected by the removal of $\hat{\mathcal{C}}(\mathbf{x})$ from the signal. This benchmark helps

illustrate that for prediction sets to make sense as a form of decision aid, we must not believe that the decision maker is well informed of the relationship between X and Y. In other words, whenever we believe that prediction sets will be informative to the decision maker, we should not expect this benchmark to be attainable.

Bayesian decision maker who knows joint distribution over sets and outcomes

If we think decision makers will rely primarily on the prediction sets rather than their knowledge of the relationship between X and Y, the best attainable performance is captured by a lower bound representing the maximum attainable score of a Bayes rational decision maker with access to the joint distribution over prediction sets and outcomes, but *not* the instance features. Here, the signal space $V = 2^{\mathcal{Y}}$ and p assigns a probability $p(v, y)$ to each label y and each possible prediction set v (a conjunction of indicators of whether each label is in the set).

For example, imagine there are three states: $\mathcal{S} = \{s_A, s_B, s_C\}$. Imagine that p assigns prior beliefs $\Pr(s_A) = \frac{1}{2}, \Pr(s_B) = \frac{1}{3}, \Pr(s_C) = \frac{1}{6}$. Say the decision maker observes a particular set $v = \hat{\mathcal{C}}(\mathbf{x}) = \{s_B, s_C\}$. They update their beliefs using Bayes rule to arrive at posterior beliefs $p(S|v)$ (Equation 5), and choose the utility-maximizing action (Equation 6).

One challenge with this benchmark is that it is vulnerable to overfitting as the label space grows (as is the prior benchmark when the feature and label spaces are large). Additionally, while this lower bound may be more attainable to human decision makers than the previous one, it also challenges intuitions we may have about prediction set use in a few ways. First, the assumption that the decision maker has no knowledge of how the features X associate with the state Y can seem unrealistic in settings where experts are assisted by models. Here, the contradiction is that it makes little sense to have humans in charge of final decisions if we don't think they have any direct knowledge of how instance features correlate with outcomes.

Second, the coverage guarantee that is the focal point of the conformal literature is only indirectly relevant to this decision maker. The benchmark is akin to imagining a decision maker whose knowledge consists of having seen the prediction sets and ground truth labels for instances in the calibration set. Coverage information is implicit in the joint distribution they have access to, but we make expect human decision makers to rely more directly on the coverage guarantee in how they use the set.

5.2. Decision maker uses set to "verify" a priori judgments

If we believe a decision maker uses the prediction set to confirm that a label they had in mind is correct, then Bayesian updating is violated. Believing that a prediction set meaningfully signals which states are possible also implies that the decision maker has false expectations about how sets are generated. However, such strategies may better capture intuitions about how human decision makers will use prediction sets than assuming oracle access to joint probabilities.

Non-Bayesian decision maker with prior and misspecified beliefs about sets

There is precedent in information economics and game theory for analyzing belief updating from sets of possible states. Signals consisting of partitions of the statespace appear in well-known constructions like the agreement theorem of Aumann, 1976. Under this interpretation, a prediction set denotes which states are possible.

Assume a decision maker who has knowledge of the prior distribution over states $p(Y)$. Upon viewing a signal of the form $v = \hat{\mathcal{C}}(\mathbf{x})$, for each possible label $y \in \mathcal{Y}$, the decision maker updates

their beliefs by noting whether the label is in the set, normalizing its prior probability using the relevant subset, then multiplying it by the appropriate probability:

$$p(y|v) = \begin{cases} \frac{p(y)}{\sum_{y' \in v} p(y')} * (1-\alpha) & y \in v \\ \frac{p(y)}{\sum_{y' \in \mathcal{Y} \setminus v} p(y')} * \alpha & y \notin v \end{cases} \tag{11}$$

where all labels have strictly positive prior probability and $\mathcal{Y} \setminus v$ denotes states not in the prediction set.

For example, given the decision problem above with prior $\Pr(s_A) = \frac{1}{2}, \Pr(s_B) = \frac{1}{3}, \Pr[s_C] = \frac{1}{6}$ and signal $v = \{s_B, s_C\}$, the decision maker selects the utility optimal action given posterior beliefs $\Pr(s_B|v) = \frac{2}{3} \times 0.95$, $\Pr(s_B|v) = \frac{1}{3} \times 0.95$, and $\Pr(s_A|v) = 0.05$.

When the decision maker lacks information about the generating process that produces a prediction set, the setting becomes a decision problem under misspecified beliefs, as studied in economics, e.g. Fudenberg et al. [2017], Heidhues et al. [2018]. Applying this strategy to conformal prediction sets implies that the decision maker's beliefs are misspecified in the sense that they believe the data-generating process generates prediction sets from partitions, i.e., non-overlapping subsets. In practice, it is difficult to motivate developing conformal algorithms with this criterion, as it further restricts an algorithm's ability to achieve desirable properties like adaptiveness through varying set size. We include this strategy here however because it proposes a plausible form of belief misspecification, resulting in a decision rule that depends only on the set and the prior distribution of the states.

Associative non-Bayesian decision maker with prior and misspecified beliefs about sets

In many domains with large label spaces to which conformal prediction has been applied, the labels are naturally organized hierarchically. For example, the decision utility of a prediction set for medical decision tasks can be modeled to account for the number of categories of diseases the set captures as defined by a medical hierarchy [Cortes-Gomez et al., 2024]. Image classification is scaffolded by a hierarchy of concepts taken from WordNet's semantic ontology [Deng et al., 2009], where the shortest path in the hierarchy is sometimes used as a proxy for image similarity. Recent empirical study of the use of prediction sets for AI-advised image labeling suggested that one way people may make decisions from a set uses semantic similarity to transfer probability to labels that are close to those included in the set, but not included themselves [Zhang et al., 2024].

For example, assume a distance function d that assigns a score $d : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ to each pair of labels, minimized at $d(y, y) = 0$. Given a distance b beyond which two labels are no longer considered closely related, we replace v with the set of all labels that are sufficiently similar to at least one label in $v : v' = \{y : y \in \mathcal{Y} \wedge \exists y' \in v \; s.t. \; d(y, y') \leq b\}$. Equation 11 above may then be applied by an associative decision maker by replacing v with v'.

5.3. Decision maker uses set to constrain decision

A final set of strategies assumes that the prediction set guides which labels a decision maker considers. This category is most closely related to the intuition that prediction sets reduce cognitive processing on the part of the decision maker. Both strategies discussed below are also amenable to associative variations where proximity to labels in the set results in additional labels being considered.

Uncertainty-suppressing decision maker who divides probability equally over states in set

Prior empirical work on the use of prediction sets in image labeling found that humans achieve higher accuracy when restricted to choosing a label from within prediction sets [Straitouri and Rodriguez, 2023]. If we expect people to suppress uncertainty in decision making, e.g., by rounding large probabilities up rather than maintaining the small possibility of alternatives, one plausible variation on the strategy discussed above is to assume that the decision maker will simply randomly choose an element from those in the set given beliefs:

$$p(y|v) = \begin{cases} \frac{1}{|v|} & y \in v \\ 0 & y \notin v \end{cases} \tag{12}$$

Cost-constrained Bayesian decision maker who uses prediction set as prior

A prediction set may reduce cognitive processing for the decision maker in situations where attention is scarce and prohibits considering all possible labels. This might be the case, for example, if further information can be gathered to support a decision (e.g., about a patient who needs to be diagnosed) but the space of possible information to collect is large, for example because there are a large number of symptoms to check for or tests that could be requested. Alternatively, a decision maker might rely on a prediction set to provide a prior on what labels to keep in mind as they process a high dimensional representation of an instance like an image of a patient's condition. For example, some study participants who used prediction sets for the AI-advised image labeling in [Zhang et al., 2024] reported viewing the set for possibilities before looking carefully at the image.

A rational inattention model from information economics provides a theoretic foundation for well-studied behavioral patterns [Matějka and McKay, 2015]. This model assumes a decision-maker facing a discrete choice situation in which they do not know the values of the available options, but have an opportunity acquire and process information at some cost [Stigler, 1961, Morgan and Manning, 1985]. The model quantifies the cognitive cost of information for a rational agent by the entropy reduction in their beliefs. In a rational inattention framework [Sims, 2003], instead of solving the payoff-maximizing decision problem in Equation 6, the agent optimizes for the expected payoff after subtracting the cognitive cost.

Consider the uniform prior as an example. Under the rational inattention framework, in tasks with $|\mathcal{Y}|$ labels (e.g. ~ 1000 labels in [Zhang et al., 2024]), the agent suffers a cognitive cost of $O(|\mathcal{Y}|)$ to figure out the true answer without the help of a conformal set, where the cost is not necessarily recovered by the payoff gain from figuring out the true label. This results in *rational* selection of a less informative but less costly signal. For example, an agent might check a few but not all labels. Checking a few labels is less informative but also less costly than figuring out an answer from all ~ 1000 labels. A conformal set reduces this cognitive cost by effectively refining the posterior over the set for free, allowing the agent to allocate cost to obtaining information over states in the set.

Conservative decision maker who maximizes worst-case utility

Thus far we have primarily considered risk neutral decision-makers whose objective is to choose the action that minimizes their expected loss: $\arg\min_{a \in \mathcal{A}} \mathbb{E}_{S \sim p(\cdot|v)}[L(a, S)]$. However, unlike point predictions, conformal sets do not provide agents with quantitative risk assessments, suggesting a connection to maxmin (risk-averse) utility optimization. Recent work by Kiyani et al. [2025] makes progress on the question: How should a decision maker who has access to only a prediction set choose actions to maximize risk-averse utility? They show that, considering all data-generating processes

$p(X, Y)$ that are compatible with the set coverage α, if the optimization goal of the decision-maker is to maximize worst-case expected utility, the optimal strategy amounts to a simple max-min strategy over the predicted labels in the set.

Formally, without any knowledge about the data-generating process, the decision-maker is constrained to select an action $a : \hat{\mathcal{C}} \to \mathcal{A}$ as a function of the conformal set. Fixing a data-generating process $p(X, Y)$, an action policy $a(\cdot)$ is evaluated by the expected α quantile of loss $l(X)$:

$$L^*(a, p) = \min_{\nu(\cdot)} \quad \mathbb{E}_{(X,Y)\sim p}[\text{quantile}_\alpha[L(a(\hat{\mathcal{C}}(X)), Y)]] \tag{13}$$

This α quantile of loss serves as a robust upperbound on the loss for any sample instance with high probability $1 - \alpha$.

A risk-averse decision-maker optimizes $\nu^*(a, p)$ for any data-generating process p with compatible coverage guarantee. We write the set of compatible p as $\mathcal{P}$, i.e. for any $p \in \mathcal{P}$,

$$\Pr_{(X,Y)\sim p}[Y \in \hat{\mathcal{C}}(X)] \geq 1 - \alpha.$$

The decision-maker solves the risk-averse optimization

$$\min_a \max_{p\in\mathcal{P}} L^*(a, p) \tag{14}$$

Kiyani et al. [2025] shows that, for $\alpha \leq \frac{1}{2}$, the optimal solution to Equation (14) is the simple min-max strategy over labels in the set, i.e.

$$a^*(\hat{\mathcal{C}}) = \arg\min_{a'\in\mathcal{A}} \max_{y\in\hat{\mathcal{C}}} L(a', y).$$

This finding specifies optimal usage of prediction sets in a risk-averse decision scenario. The decision maker can simply treat the labels in the prediction set as defining possible realizations of the state, and apply what they know about the loss function to select the label in the set with the best worst-case loss.

5.4. Takeaways

These examples help illustrate two important takeaways. First, the fact that we cannot specify a single best procedure for a risk-neutral agent who uses prediction sets highlights the underspecification of "good" decision making with conformal prediction, including in the simplified case where we do not expect the decision maker to have access to additional information over the model. This is not necessarily a reason to avoid using prediction sets, as there is no guarantee that human decision makers will engage in idealized use of an uncertainty quantification when it is well defined. However, that this aspect of conformal prediction use has largely gone unacknowledged in the literature suggests that there is work left to do to characterize successful integration of conformal prediction in human-facing decision workflows.

On the other hand, assuming that decision-makers will be risk-averse may be reasonable in some settings where prediction sets attract interest, such as medical diagnosis or treatment decisions. Even if real human decision-makers cannot confidently rank actions by their worst-case loss upon first encountering a set, in some cases subsequent information gathering may be enough to reduce ambiguity, making 14 a reasonable strategy to assume.

Second, the above examples illustrate how our ability to predict their usefulness depends on what we think human decision makers know and what sort of constraints we believe they are under. To design prediction sets that are useful for real human decision makers, these aspects of a decision scenario must be considered explicitly.

6. The meaning of quantified uncertainty is its use

We have demonstrated some ways in which conformal prediction sets are elusive as a form of decision support. How should ambiguity about how prediction sets are best used by decision makers influence how we see the value of conformal prediction for quantifying uncertainty? What considerations should guide how conformal prediction is implemented in practical decision workflows, and what gaps are most important to address through future research?

One perspective on these tensions is that we should expect the meaning of quantified uncertainty to lie in its use. This perspective has several implications. Rather than prioritizing the ability to prove theoretical guarantees and looking to domain-specific decision processes primarily for datasets or use cases, we should be considering how different forms of guarantees align with the needs and sensitivities of human decision makers throughout method development. We should study the use of conformal inference from the perspective of developers and decision makers, and expect its value in practice to sometimes lie beyond our attempts at formal characterization. We reflect below on what this suggests for research aimed at improving the utility of prediction sets for human decision makers, and more broadly how it affects how we "price in" conformal prediction relative to alternative techniques.

6.1. A human-centered research agenda for conformal prediction

Understanding and guiding human decision makers' use of prediction sets. Our work suggests that despite its popularity in practice and theoretical research, we lack a deep understanding of conformal prediction as a human-facing form of decision support.

One way to approach empirical study of prediction set use is to compare observed human performance with prediction sets to benchmarks based on the strategies we describe (Section 5). We associate this approach with the general practice of statistical model checking [Hullman and Gelman, 2021], where in this case the goal is to identify which of several models appears to best explain human performance and to observe deviations that remain to be explained. There is precedent for designing experiments to discern possible mechanisms behind human decision performance under uncertainty in the data visualization literature (e.g., [Kale et al., 2020, Wu et al., 2023]). A focus on testing user performance against expectations given different strategies provides a more robust basis for further theory development than approaching uncertainty quantification for humans as if they are a blackbox to optimize (e.g., by running a grid search over different hyperparameters like coverage level to find the values that lead to the best performance). We should expect human performance and use to vary as we vary aspects of the decision problem, loss function, and problem setting. The question is how changes in a decision problem impact reliance on different strategies for using sets. Hence, we should avoid drawing general conclusions about the level of coverage or particular algorithm that will work best for human decision makers more broadly from any single experiment.

Incorporating expert knowledge into predictive uncertainty design. Given that human experts are often paired with AI models because they are thought to have complementary information, it is critical for future research to acknowledge information complementarities and pursue methods for characterizing and designing for the particular information conditions under which predictive uncertainty will be presented [Corvelo Benz and Rodriguez, 2024, Guo et al., 2024a]. This means recognizing the potential disparity between probability guarantees defined on information available to a model and those that are directly relevant to a human with additional information. Whenever we believe the human decision maker may possess information beyond that encoded in

the model, we should expect probabilities based only on the model's available information to be wrong from the decision maker's perspective.

Related to this, organizations deploying AI models to assist human decision makers often choose to keep the human in the loop because of potential flexibility in their decision strategies in the face of shifts in the decision-relevant distribution. Understanding the robustness of coverage achieved by standard conformal algorithms to common forms of shift, and how well humans can pick up on shifts even if a model may not, is a well-motivated direction for future work.

Reconsidering conformal guarantees from a human perspective. The conformal prediction literature prioritizes coverage guarantees over other properties that may be desirable to human decision makers, such as low decision loss or properties of set size. Kiyani et al. [2025], for example, maximize utility subject to coverage guarantees. In practice, however, coverage may be less important to decision makers, suggesting an inversion of this approach where coverage is maximized subject to minimum utility or other requirements that are important to human decision makers, like the information captured by set size. For example, work by Straitouri et al. [2023] asks what coverage level leads to the most effective human decisions. Empirical work aimed at re-prioritizing properties of prediction sets for human alignment can be pursued in tandem model checking approach to experimentation described above, to allow for solutions to be adapted to slightly different scenarios. This is because empirical estimates of the parameters maximize expert performance learned without also attending to underlying mechanisms may be brittle to slight changes in the decision problem or expert population.

Aligning uncertainty quantification theory with human preferences Related to our point about re-prioritizing properties of prediction sets, future research should attempt to better understand and integrate aspects of decision makers' preferences for uncertainty quantification in conformal theory. These include set size, which is usually achieved through heuristics, and the "shape" that sets can take, particularly in higher dimensions where designers must constrain the structure (e.g., only considering contiguous regions or ellipsoids in a high dimensional space).

Part of these efforts will include identifying effective ways to elicit human preferences for uncertainty quantification. While some recent work demonstrates how to incorporate a domain-specific, downstream decision loss function into the generation of prediction sets [Cortes-Gomez et al., 2024], this information is assumed to be given as input. To align predictive uncertainty with human needs and values in arbitrary domains, we will also need elicitation methods that can be brought to bear in arbitrary domains where decision problems and associated preferences have not yet been fully characterized. Another direction is to pursue interactive methods that help experts steer the prediction sets that are generated to meet their needs, such as by enabling them to "query" a set for potential outcomes that are most relevant to their information needs.

6.2. Pricing in conformal prediction

Statistical theory is sometimes said to be the theory of applied statistics. Conformal inference demonstrates how established theory in rank-based statistics can be applied to modern problems of quantifying prediction uncertainty. The attractiveness of prediction sets to human decision makers warrants future research aimed at increasing the alignment between conformal methods and experts' needs and preferences. Yet, we should expect it to remain difficult to fully characterize the value of conformal prediction relative to alternative approaches to uncertainty quantification. We discuss several considerations that may be important in practice but remain difficult to formally "price in," and advocate for acknowledging these considerations in practice despite their slipperiness as theoretical constructs.

6.2.1. The twoishness of post hoc calibration

Offline calibration methods can often be described as relying on two models: the original black-box predictive model and the model of the frequency distribution of the predictions learned from calibration data. Split conformal prediction combines an arbitrary predictive model whose output we use to derive a non-conformity score as our data model (or likelihood) with a second model defined by the sampling distribution of the non-conformity scores. This in not unique to conformal prediction: many statistical methods benefit from a certain "twoishness" in their construction [Gelman, 2011]. For example, a common pattern occurs in classical statistical analysis, where one method is used to produce estimates (e.g., a multivariate regression to estimate effects and compute standard errors) and another is used to check if the first method makes sense (e.g., predictive checks or significance tests on the coefficients). When it comes to conformal prediction, the benefits of combining these two models are obvious: we are absolved of having to understand the relationship between the model's training distribution and the deployment (test) distribution, and can ignore that the original model may be misspecified to instead focus on deriving prediction probabilities that hold on calibration data we choose.

At the same time, twoishness tends to coincide with limitations inherent to the modeling task. For conformal prediction, the ability to rely on new data to achieve coverage is most useful when we find ourselves in the limited position where we cannot retrain the predictive model despite a mismatch between the training and deployment context. A risk in such situations is that the gaps that remain when we combine multiple models are de-prioritized relative to the "solution" that the combination provides, such that practitioners fail to acknowledge critical contingencies.

As an analogy, consider the standard bootstrap, a distribution-free technique for quantifying uncertainty that resembles conformal prediction in substituting simple computations for explicit distributional assumptions. In standard non-parametric bootstrapping, we have some observed data $\mathcal{D} = (X_1, \ldots, X_n) \sim P$ and wish to quantify the uncertainty in our plug-in estimate of a "parameter" of P, $\hat{\theta}_n$. Constructing valid $1 - \alpha$ confidence intervals on our sample estimate means approximating the exact finite sample distribution of a test statistic involving the estimator, for example:

$$F_n(t) = \Pr(\sqrt{n}(\hat{\theta}^* - \hat{\theta}_n) \leq t) \tag{15}$$

This is accomplished by estimating F_n using an empirical approximation $\hat{F}_n$ conditional on D, which is treated as fixed. Specifically, the Monte Carlo approximation $\bar{F}$ is used:

$$\bar{F}(t) = \frac{1}{B} \sum_{j=1}^{B} I(\sqrt{n}(\hat{\theta}_j^* - \hat{\theta}_j) \leq t) \tag{16}$$

Key ideas behind the bootstrap include the observation that an unknown probability distribution can be replaced with an empirical estimate (the substitution principle), and the use of simulation to avoid analytically calculating properties of an estimator [Davison et al., 2003]. Related to twoishness, when we acknowledge the distinction between the model under which we define the estimator $\hat{\theta}_n$ and the model of the sampling process, it becomes clearer that while the bootstrap allows us to minimize the error or uncertainty from the Monte Carlo approximation (by setting a high enough number of iterations B in 16, we cannot reduce our uncertainty about how well our particular sample $\mathcal{D}$ represents P, and therefore our uncertainty about how well $\hat{F}_n$ approximates F_n. It would be a mistake to adopt the bootstrap in any particular setting without considering how the irreducible uncertainty due to having only a single sample might affect our goals, through, for example, the influence of outliers on the target statistic.

Similarly, in research on conformal prediction, we should not let our ability to prove guarantees about conformal methods overshadow acknowledgement that whenever we bring in calibration data after model training, any "guarantees" are contingent on the fitness of the calibration data as a representation of the ground truth generating process and on our ability to imagine taking hypothetical random samples of calibration and test data. At minimum, this suggests emphasizing the conditionality of guarantees when describing methods like conformal prediction to practitioners and other non-theorists who may not fully grasp the underlying assumptions (of a static generating process, the relevance of asymptotic theory, and so on).

6.2.2. The allure of thoughtless statistics

On the other hand, we should also expect approaches to quantifying uncertainty to sometimes have value that falls outside the bounds of available theory. For example, the existence of second-order asymptotic guarantees on coverage does not necessarily help us understand the popularity of the standard bootstrap. Instead, the widespread appeal of the bootstrap in practice likely owes much to its generality of construction. By avoiding distributional assumptions and substituting computation for analytical calculation, the bootstrap promises quantified uncertainty with assumptions that are implicit—but its popularity surely also derives from its twoishness, which (a) incorporates additional information about the sampling procedure that is not included in the estimator being bootstrapped, and (b) gives additional flexibility to the user. Bootstrap theory cannot comment on its perceived value as an exploratory tool, even in settings where use of the bootstrap is not formally justified (e.g., where observations are correlated to some extent, or outputs consist of graphs or trees; [Holmes, 2003]). Because the procedure is simple and seemingly well understood, applying the bootstrap can still provide some intuition into the variance in a procedure when the typical conditions do not hold.

Similarly, split conformal prediction may appeal over more robust uncertainty quantification approaches that ensemble several models or refine the base predictive model because of its procedural simplicity. Even when we don't expect conformal coverage guarantees to hold exactly as stated (e.g., because of data shifts or constraints like spatial or temporal correlations, or additional knowledge a human possesses), relative information about model uncertainty captured in set size (e.g., [Angelopoulos et al., 2021]) may still be valuable in many settings.

We should expect decision makers to perceive their options differently when it comes to approaching decision making with access to prediction sets versus labels with predicted probabilities. For example, we would expect uncertainty suppressing decision makers who round predicted probabilities (Section 5.3) to diversify their actions more with prediction sets than predicted probabilities. There is a sense in which prediction sets provide a more concrete depiction of possibilities, such that findings that behavioral decision makers benefit from discrete framings of probability may apply [Hullman et al., 2015, Kay et al., 2016, Kale et al., 2020]. It is also possible that seeing more labels, as opposed to seeing only the model's top prediction or top couple predictions, has value from the standpoint of encouraging broader thinking about possible outcomes, or from a debugging perspective, where a decision maker may be more likely to consider sources of epistemic uncertainty upon seeing a prediction set compared to a top model prediction. These possibilities are not easily "priced in" and consequently get overlooked in theoretical discussions.

At the same time, it is important not to assume that seemingly large differences in the philosophy behind different expressions of predictive uncertainty means that the decisions they induce will be equally distinct. Effect sizes in existing empirical work comparing human decisions under prediction sets versus alternatives like top-k sets based on heuristics like softmax scores are relatively small, even as researchers study settings where we might expect conformal sets to have

advantages [Cresswell et al., 2024b,a, Zhang et al., 2024]. [Zhang and Maloney, 2012], for example, find that even in on out-of-distribution images where prediction sets showed the biggest advantage of top-k presentations with softmax values, the difference in decision accuracy of the human-AI team was only 4 percentage points. Small effects (e.g., on the order of a few percentage points difference in probability perception) are also common in research on uncertainty visualization [Kale et al., 2020]. Reflecting on how to rigorously quantify uncertainty can play a critical role in shaping future data collection, evaluation design, and other components of a modeling-to-decision pipeline, but once put into practice, the benefits of a chosen quantification can be difficult to demonstrate.

6.2.3. Second-order uncertainty: What can't be learned can't be priced

Discussions of sources of uncertainty often distinguish *aleatoric uncertainty*—uncertainty in an outcome due to inherent randomness in a process—from *epistemic uncertainty* due to a lack of knowledge that could in theory be reduced. While calibration can assure good decisions under aleatoric uncertainty under certain assumptions, its resolution of epistemic uncertainty is arguably less satisfying. However, epistemic uncertainty cannot be "priced in" to the loss minimization framework in which many predictive uncertainty techniques are evaluated.

Returning to multiclass classification as an example, a classifier $\hat{f}$ trained to minimize loss using a set of labeled training data can be thought of as producing predictions at different levels, which are evaluated using different empirical values. We can evaluate the "level 0" prediction $\hat{y} = \hat{f}(X = \mathbf{x})$ of the label against the true label using 0/1 loss. We can evaluate the "level 1" prediction of the conditional distribution $\hat{p}(\cdot|X = \mathbf{x})$ against the empirical distribution of labels.

Aleatoric uncertainty captures the fact that even if we were given full knowledge of $p(\cdot|X = \mathbf{x})$, we cannot predict with certainty due to inherent randomness in the generating process. The ideal case for deploying a statistical predictor to make decisions is when we wish to minimize decision loss on average, and have access to a large amount of data that we are confident reflects the target generating process. In such cases, we are primarily concerned with aleatoric uncertainty.

In contrast, epistemic uncertainty deals with the possibility that the learned probabilistic predictor $\hat{p}$ associated with classifier $\hat{f}$ differs from the ground truth p. Distribution-free post hoc calibration of prediction uncertainty promises to "fix" the gap between $\hat{p}$ and p, but offers the decision maker no information to estimate epistemic uncertainty, i.e., how different the prediction $\hat{p}(\cdot|X = \mathbf{x})$ may be from the ground truth. However, because we can't observe samples from the second order distribution $\mathbb{P}(\mathbb{P}(y|X = \mathbf{x}))$, we cannot learn a faithful predictor of higher order epistemic uncertainty within a loss minimization framework [Bengs et al., 2022]. This is formalized by [Bengs et al., 2023], who relate it to the impossibility of defining a proper scoring rule. Learning to predict conditional distributions $\hat{p}(\cdot|X = \mathbf{x})$ using empirical risk minimization–i.e., identifying the empirical risk minimizer that minimizes loss on the training data—under a proper scoring rule means that the optimal predictor (the true risk minimizer or Bayes predictor) will coincide with the true conditional class distribution $p(\cdot|X = \mathbf{x})$. However, there is no loss function for comparing second-order predictions of $p(p(\cdot|X = \mathbf{x}))$ with observations y such that minimizing that function on the training data ensures a second-order predictor will be "honest" in the sense of returning the prediction corresponding to its belief.

When we specify priors on model parameters in a Bayesian model, we explicitly express this uncertainty in our predictions. The posterior predictive distribution $p(y_{rep}|y)$ marginalizes the distribution of y_{rep}, which we can think of as a newly drawn version of y given θ, over the posterior distribution of θ given y, which, by virtue of conditioning the prior, captures epistemic uncertainty stemming from our lack of knowledge of the model parameters. In a machine learning context, the posterior predictive distribution of a Bayesian neural network with latent variables captures

uncertainty in network weights as well as in random noise and latent variables in data generation [Depeweg et al., 2018]. However, Bayesian probability is not distinguishable in the Frequentist framework assumed by post hoc calibration methodologies.

Rather than arguing for one statistical paradigm over another, our point here is simply that there is always a risk to concluding that because some properties of uncertainty quantification cannot be incorporated into some particular framework, they cannot have value for decision makers in practice. Consider a situation where abstaining from making a decision in order to gather more information is possible. Offering the decision maker a probability distribution over the rate at which an event is expected to occur (as compared to a maximum likelihood point estimate) provides them a sense of how much risk is incurred by acting as if the maximum likelihood prediction is true versus holding out until more information is available. Similar to how we should consider what information is available to human decision-makers versus artificial agents when we design uncertainty quantification, we should also consider whether decision-makers would benefit from epistemic uncertainty that expresses the limitations of knowledge for a particular decision. This may be particularly desirable in situations where decision-makers are liable for individual decisions, such as about specific patients in a medical setting.

In conclusion, we should not forget that the meaning of quantified uncertainty as decision support for human experts often lies in its use. We should see the twoishness of a statistical method as a methodological "seam" reminding us that what a method provides is always a function of how its distinct component models combine. We should keep our eye on what lies beyond our ability to theorize, pursuing methods for better integrating human domain expertise and preferences and acknowledging when decision makers may desire forms of uncertainty that lie outside our chosen framework. Ultimately, we should not expect theory to capture everything valuable about a method, but we should continue trying to grow the reach of theory to include the practical questions. In other words, we should aim for theoretical statistics to be the theory of applied statistics without forgetting that there will always be gaps.

References

Stefanía Ægisdóttir, Michael J. White, Paul M. Spengler, Alan S. Maugherman, Linda A. Anderson, Robert S. Cook, Cassandra N. Nichols, Georgios K. Lampropoulos, Blain S. Walker, Genna Cohen, et al. The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *Counseling Psychologist*, 34(3):341–382, 2006.

Rohan Alur, Loren Laine, Darrick Li, Manish Raghavan, Devavrat Shah, and Dennis Shung. Auditing for human expertise. *Advances in Neural Information Processing Systems*, 36, 2024.

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Anastasios Nikolas Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=eNdiU_DbM9.

Anastasios Nikolas Angelopoulos, Rina Barber, and Stephen Bates. Online conformal prediction with decaying step sizes. In *International Conference on Machine Learning*, pages 1616–1630. PMLR, 2024.

Robert J. Aumann. Agreeing to disagree. *Annals of Statistics*, 4(6):1236–1239, 1976.

Christopher R. S. Banerji, Tapabrata Chakraborti, Chris Harbron, and Ben D. MacArthur. Clinical AI tools must convey predictive uncertainty for each individual patient. *Nature Medicine*, 29(12): 2996–2998, 2023.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *Annals of Statistics*, 51(2):816–845, 2023.

Sarah Belia, Fiona Fidler, Jennifer Williams, and Geoff Cumming. Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, 10(4):389–396, 2005.

Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Pitfalls of epistemic uncertainty quantification through loss minimisation. *Advances in Neural Information Processing Systems*, 35: 29205–29216, 2022.

Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. On second-order scoring rules for epistemic uncertainty quantification. In *International Conference on Machine Learning*, pages 2078–2091. PMLR, 2023.

Aadyot Bhatnagar, Huan Wang, Caiming Xiong, and Yu Bai. Improved online conformal prediction via strongly adaptive online learning. In *International Conference on Machine Learning*, pages 2337–2363. PMLR, 2023.

Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, et al. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 401–413, 2021.

Jaroslaw Blasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. When does optimizing a proper loss yield calibration? In *Advances in Neural Information Processing Systems*, volume 36, 2024.

Natalie Collina, Surbhi Goel, Varun Gupta, and Aaron Roth. Tractable agreement protocols. *arXiv preprint arXiv:2411.19791*, 2024.

Santiago Cortes-Gomez, Carlos Patiño, Yewon Byun, Steven Wu, Eric Horvitz, and Bryan Wilder. Decision-focused uncertainty quantification. *arXiv preprint arXiv:2410.01767*, 2024.

Nina Corvelo Benz and Manuel Rodriguez. Human-aligned calibration for AI-assisted decision making. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

Jesse C. Cresswell, Bhargava Kumar, Yi Sui, and Mouloud Belbahri. Conformal prediction sets can cause disparate impact. *arXiv preprint arXiv:2410.01888*, 2024a.

Jesse C. Cresswell, Yi Sui, Bhargava Kumar, and Noël Vouitsis. Conformal prediction sets improve human decision making. In *International Conference on Machine Learning*, pages 9439–9457. PMLR, 2024b.

Anthony C. Davison, David V. Hinkley, and G. Alastair Young. Recent developments in bootstrap methodology. *Statistical Science*, 18:141–157, 2003.

A. Philip Dawid. Calibration-based empirical probability. *Annals of Statistics*, 13(4):1251–1274, 1985.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.

Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*, pages 1184–1193. PMLR, 2018.

Cynthia Dwork, Michael P. Kim, Omer Reingold, Guy N. Rothblum, and Gal Yona. Outcome indistinguishability. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1095–1108, 2021.

Shai Feldman, Liran Ringel, Stephen Bates, and Yaniv Romano. Achieving risk control in online learning settings. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=5Y04GWvoJu.

Dean P. Foster and Rakesh V. Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998.

Drew Fudenberg, Gleb Romanyuk, and Philipp Strack. Active learning with a misspecified prior. *Theoretical Economics*, 12(3):1155–1189, 2017.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR, 2016.

Alberto Garcia-Galindo, Marcos Lopez-De-Castro, and Rubén Armananzas. An uncertainty-aware sequential approach for predicting response to neoadjuvant therapy in breast cancer. In *Conformal and Probabilistic Prediction with Applications*, pages 74–88. PMLR, 2023.

Andrew Gelman. The pervasive twoishness of statistics; in particular, the "sampling distribution" and the "likelihood" are two different models, and that's a good thing. *Statistical Modeling, Causal Inference, and Social Science*, 2011. URL https://statmodeling.stat.columbia.edu/2011/06/20/the_sampling_di_1/.

Isaac Gibbs and Emmanuel J. Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, pages 1660–1672, 2021.

Isaac Gibbs and Emmanuel J. Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024.

Isaac Gibbs, John J Cherian, and Emmanuel J. Candès. Conformal prediction with conditional guarantees. *arXiv preprint arXiv:2305.12616*, 2023.

Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. *arXiv preprint arXiv:2109.05389*, 2021.

Parikshit Gopalan, Michael P. Kim, Mihir A. Singhal, and Shengjia Zhao. Low-degree multicalibration. In *Conference on Learning Theory*, pages 3193–3234. PMLR, 2022.

William M. Grove, David H. Zald, Boyd S. Lebow, Beth E. Snitz, and Chad Nelson. Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, 12(1):19, 2000.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

Ziyang Guo, Yifan Wu, Jason Hartline, and Jessica Hullman. Unexploited information value in human-ai collaboration. *arXiv preprint arXiv:2411.10463*, 2024a.

Ziyang Guo, Yifan Wu, Jason D. Hartline, and Jessica Hullman. A decision theoretic framework for measuring AI reliance. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 221–236, 2024b.

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.

Paul Heidhues, Botond Kőszegi, and Philipp Strack. Unrealistic expectations and misguided learning. *Econometrica*, 86(4):1159–1214, 2018.

Rink Hoekstra, Richard D. Morey, Jeffrey N Rouder, and Eric-Jan Wagenmakers. Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, 21:1157–1164, 2014.

Susan Holmes. Bootstrapping phylogenetic trees: Theory and methods. *Statistical Science*, 18(2): 241–255, 2003.

Lunjia Hu and Yifan Wu. Predict to minimize swap regret for all payoff-bounded tasks. *2024 IEEE 65th Annual Symposium on Foundations of Computer Ccience (FOCS)*, 2024.

Jessica Hullman and Andrew Gelman. Designing for interactive exploratory data analysis requires theories of graphical inference. *Harvard Data Science Review*, 3(3):10–1162, 2021.

Jessica Hullman, Paul Resnick, and Eytan Adar. Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PloS One*, 10(11):e0142444, 2015.

Jessica Hullman, Alex Kale, and Jason Hartline. Decision theoretic foundations for experiments evaluating human decisions. *arXiv preprint arXiv:2401.15106*, 2024.

Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Dk7QQp8jHEo.

Alex Kale, Matthew Kay, and Jessica Hullman. Visual reasoning strategies for effect size judgments and decisions. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):272–282, 2020.

Matthew Kay, Tara Kola, Jessica R. Hullman, and Sean A. Munson. When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5092–5103, 2016.

Shayan Kiyani, George Pappas, Aaron Roth, and Hamed Hassani. Decision theoretic foundations for conformal prediction: Optimal uncertainty quantification for risk-averse agents. *arXiv preprint arXiv:2502.02561*, 2025.

Bobby Kleinberg, Renato Paes Leme, Jon Schneider, and Yifeng Teng. U-calibration: Forecasting for an unknown agent. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5143–5145. PMLR, 2023.

Charles Lu, Andréanne Lemay, Ken Chang, Katharina Höbel, and Jayashree Kalpathy-Cramer. Fair conformal predictors for applications in medical imaging. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12008–12016, Jun. 2022. doi: 10.1609/aaai.v36i11.21459.

Charles F Manski. The 2009 Lawrence R. Klein lecture: Diversified treatment under ambiguity. *International Economic Review*, 50(4):1013–1041, 2009.

Filip Matějka and Alisdair McKay. Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105(1):272–298, 2015.

Paul E. Meehl. *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. University of Minnesota Press, 1954.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems*, 34:15682–15694, 2021.

Peter Morgan and Richard Manning. Optimal search. *Econometrica*, 53:923–944, 1985.

Radford M. Neal. *Bayesian Learning for Neural Networks*. Springer, 2012.

John C Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74, 1999.

Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. In *Uncertainty in Artificial Intelligence*, pages 844–853. PMLR, 2021.

Drew Prinster, Samuel Don Stanton, Anqi Liu, and Suchi Saria. Conformal validity guarantees exist for any data distribution (and how to find them). In *International Conference on Machine Learning*, pages 41086–41118. PMLR, 2024.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, volume 33, pages 3581–3591, 2020.

Leonard J. Savage. *The Foundations of Statistics*. Courier Corporation, 1972.

Christopher A Sims. Implications of rational inattention. *Journal of monetary Economics*, 50(3): 665–690, 2003.

Mark Steyvers, Heliodoro Tejeda, Gavin Kerrigan, and Padhraic Smyth. Bayesian modeling of human-AI complementarity. *Proceedings of the National Academy of Sciences*, 119(11): e2111547119, 2022.

George J. Stigler. The economics of information. *Journal of Political Economy*, 69(3):213–225, 1961.

Eleni Straitouri and Manuel Gomez Rodriguez. Designing decision support systems using counterfactual prediction sets. *arXiv preprint arXiv:2306.03928*, 2023.

Eleni Straitouri, Lequn Wang, Nastaran Okati, and Manuel Gomez Rodriguez. Improving expert predictions with conformal prediction. In *International Conference on Machine Learning*, pages 32633–32653. PMLR, 2023.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian Conference on Machine Learning*, pages 475–490. PMLR, 2012.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Abraham Wald. Statistical decision functions. *Annals of Mathematical Statistics*, 20:165–205, 1949.

Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans. *arXiv preprint arXiv:2005.00582*, 2020.

Yifan Wu, Ziyang Guo, Michalis Mamakos, Jason Hartline, and Jessica Hullman. The rational agent benchmark for data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 2023.

Dongping Zhang, Angelos Chatzimparmpas, Negar Kamali, and Jessica Hullman. Evaluating the utility of conformal prediction sets for ai-advised image labeling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2024.

Hang Zhang and Laurence T Maloney. Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in neuroscience*, 6:1, 2012.

Shengjia Zhao, Michael Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. Calibrating predictions to decisions: A novel approach to multi-class calibration. In *Advances in Neural Information Processing Systems*, volume 34, pages 22313–22324, 2021.