Towards Human-AI Complementarity with Prediction Sets

Giovanni De Toni*
Fondazione Bruno Kessler & University of Trento
Trento, Italy
giovanni.detoni@unitn.it

Nastaran Okati
Max Planck Institute for Software Systems
Kaiserslautern, Germany
nastaran@mpi-sws.org

Suhas Thejaswi
Max Planck Institute for Software Systems
Kaiserslautern, Germany
thejaswi@mpi-sws.org

Eleni Straitouri
Max Planck Institute for Software Systems
Kaiserslautern, Germany
estraitouri@mpi-sws.org

Manuel Gomez-Rodriguez
Max Planck Institute for Software Systems
Kaiserslautern, Germany
manuelgr@mpi-sws.org

Abstract

Decision support systems based on prediction sets have proven to be effective at helping human experts solve classification tasks. Rather than providing single-label predictions, these systems provide sets of label predictions constructed using conformal prediction, namely prediction sets, and ask human experts to predict label values from these sets. In this paper, we first show that the prediction sets constructed using conformal prediction are, in general, suboptimal in terms of average accuracy. Then, we show that the problem of finding the optimal prediction sets under which the human experts achieve the highest average accuracy is NP-hard. More strongly, unless $\mathsf{P} = \mathsf{NP}$, we show that the problem is hard to approximate to any factor less than the size of the label set. However, we introduce a simple and efficient greedy algorithm that, for a large class of expert models and non-conformity scores, is guaranteed to find prediction sets that provably offer equal or greater performance than those constructed using conformal prediction. Further, using a simulation study with both synthetic and real expert predictions, we demonstrate that, in practice, our greedy algorithm finds near-optimal prediction sets offering greater performance than conformal prediction.

1 Introduction

In recent years, there has been increasing excitement about the potential of decision support systems based on machine learning to help human experts make more accurate predictions in a variety of application domains, including medicine, education and science [1–3]. In this context, the ultimate goal is human-AI complementarity—the predictions made by the human expert who uses a decision support system are more accurate than the predictions made by the expert on their own and by the classifier used by the decision support system [4–8].

The conventional wisdom is that to achieve human-AI complementarity, decision support systems should help humans understand when and how to use their predictions to update their own. As a result,

*The author contributed to this paper during an internship at the Max Planck Institute for Software Systems.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-0986

a flurry of empirical studies has analyzed how factors such as confidence, explanations, or calibration influence when and how humans use the predictions provided by a decision support system [9–12]. Unfortunately, these studies have been so far inconclusive and it is yet unclear how to design decision support systems that achieve human-AI complementarity [13–17].

In this context, Straitouri et al. [18, 19] have recently argued, both theoretically and empirically, that an alternative type of decision support systems may achieve human-AI complementarity, by design. Rather than providing a single label prediction and letting a human expert decide when and how to use the predicted label to update their own prediction, these systems provide a set of label predictions, namely a prediction set, and ask the expert to predict a label value from the set.[2] To construct each prediction set, these systems rely on a conformal predictor [20, 21]. The conformal predictor first computes a non-conformity score for each potential label value using the output provided by a classifier (*e.g.*, the softmax scores), and then adds a label value to the prediction set if its non-conformity score is below a data-driven threshold computed using a calibration set. Further, to optimize the performance of these systems, Straitouri et al. have introduced several methods to efficiently find the optimal value of the threshold used by the conformal predictor.[3] However, it is unclear whether the optimal prediction sets maximizing the average accuracy achieved by an expert who uses such systems can always be constructed using a deterministic threshold rule as the one used by a conformal predictor. Motivated by this observation, in this work, our goal is to understand how to construct optimal prediction sets under which human experts achieve the highest average accuracy.

Our contributions. We first demonstrate that there exist (many) data distributions for which the optimal prediction sets under which the human experts achieve the highest average accuracy cannot be constructed using a conformal predictor. Then, we show that the problem of finding the optimal prediction sets is NP-hard by using a reduction from the k-clique problem [22]. More strongly, unless $\mathsf{P} = \mathsf{NP}$, we show that the problem is hard to approximate to any factor less than the size of the label set. However, we introduce a simple and computationally efficient greedy algorithm that, for a large class of non-conformity scores and expert models parameterized by a mixture of multinomial logit models (MNLs), is guaranteed to find prediction sets that provably offer equal or greater performance than those constructed using conformal prediction. Moreover, using a simulation study with both synthetic and real expert predictions, we demonstrate that, in practice, our greedy algorithm finds near-optimal prediction sets offering greater performance than conformal prediction. We have released an open-source implementation of our greedy algorithm as well as the code and data used in our experiments at `https://github.com/Networks-Learning/towards-human-ai-complementarity-predictions-sets`.

Further related work. Our work builds upon further related work on set-valued predictors, assortment optimization, and learning under algorithmic triage.

The literature on set-valued predictors aims to develop predictors that, for each sample, output a set of label values, namely a prediction set [23]. Set-valued predictors have not been designed nor evaluated by their ability to help human experts make more accurate predictions [24–27], except for a few notable exceptions [18, 19, 28–30]. These exceptions provide empirical evidence that conformal predictors, a specific type of set-valued predictors, may help human experts make more accurate predictions. Among these exceptions, the work by Straitouri et al. [18, 19], which we have already discussed previously, is most related to ours. In this context, it is also worth noting that a recent theoretical study has argued that prediction sets may also help experts create more accurate rankings [31].

The literature on assortment optimization aims to develop methods to help a seller select a subset of products from a universe of substitutable products, namely an assortment, with maximum expected revenue [32–36]. Within this literature, the work most closely related to ours tackles the assortment optimization problem under customization [35, 36], where there are different types of customers and each type of customer chooses products following a different multinomial logit model. More specifically, by mapping products to label values, types of customers to ground truth label values, and revenue to accuracy, one could think of our problem as an assortment optimization problem

[2]There are many decision support systems used by experts that, under normal operation, forcefully limit experts' level of agency. For example, in aviation, there are automated, adaptive systems that prevent pilots from taking certain actions based on the monitoring of the environment.

[3]A threshold value is optimal if it maximizes the average accuracy achieved by an expert who predicts label values from the prediction sets created by the conformal predictor.

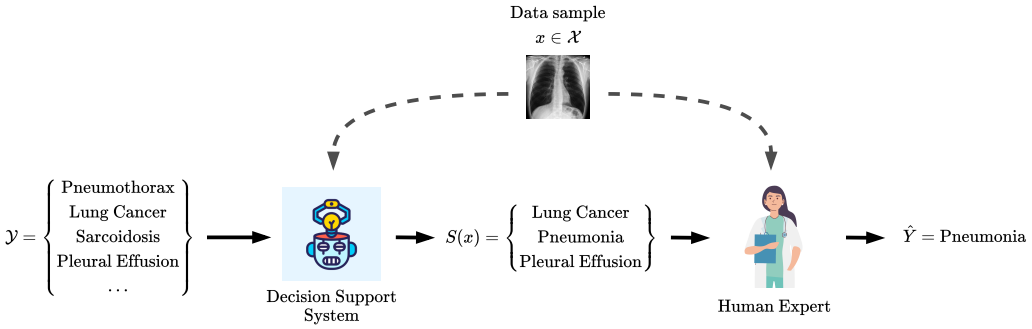

Figure 1: Our automated decision support system. Given an instance with a feature vector x, the system $\mathcal{C}$ helps the expert by automatically narrowing down the set of potential label values to a prediction set $\mathcal{S}(x) \subseteq \mathcal{Y}$. The system asks the expert to predict a label value $\hat{y}$ from $\mathcal{S}(x)$.

under customization. However, in the assortment optimization problem under customization, the type of each customer is known and thus may be offered different subsets of products whereas, in our problem, the ground truth label is unknown. As a result, (the complexity of) our problem and our technical contributions are fundamentally different.

The literature on learning under algorithmic triage aims to develop classifiers that make predictions for a given fraction of the samples and leave the remaining ones to human experts, as instructed by a triage policy [37–42]. In contrast, in our work, for each sample, a classifier is used to construct a prediction set and a human expert needs to predict a label value from the set. In this context, it is also worth noting that learning under algorithmic triage has been extended to reinforcement learning settings [43–46].

2 Decision Support Systems Based on Prediction Sets

Given a multiclass classification task where, for each task instance, a human expert needs to predict the value of a ground truth label $y \in \mathcal{Y} = \{1, \ldots, L\}$, we focus on the design of a decision support system that, given a set of features $x \in \mathcal{X}$, helps the expert by narrowing down the set of potential label values to a subset of them $\mathcal{S}(x) \subseteq \mathcal{Y}$. Here, similarly as in Straitouri et al. [18, 19], we assume that, for any instance with features $x \in \mathcal{X}$, the system asks the expert's prediction $\hat{y} \in \mathcal{Y}$ to belong to the prediction set $\mathcal{S}(x)$, *i.e.*, $\hat{y} \in \mathcal{S}(x)$. The key rationale for restricting the expert's agency is that, if we would allow the expert to predict label values from outside the prediction set, a good performance would depend on the expert developing a good understanding of when to predict a label from the prediction set. In this context, it is worth highlighting that Straitouri et al. [19] run a large-scale human subject study to compare the above setting against an alternative setting where experts are allowed to predict label values from outside the prediction sets. They found that, in the alternative setting, the number of predictions in which the prediction sets do not contain the true label and the experts succeed is consistently smaller than the number of predictions in which the prediction sets contain the true label and the experts fail. As a consequence, in the alternative setting, experts perform worse. Refer to Figure 1 for an illustration of the decision support system.

Then, for any $x \in \mathcal{X}$, our goal is to find the optimal prediction set $\mathcal{S}^*(x)$ that maximizes the average accuracy of the expert's prediction,[4] *i.e.*,

$$\mathcal{S}^*(x) = \operatorname*{argmax}_{\mathcal{S} \subseteq \mathcal{Y}} g(\mathcal{S} \,|\, x) \quad \text{where} \quad g(\mathcal{S} \,|\, x) = \mathbb{E}_{Y \sim P(Y \,|\, X), \hat{Y} \sim P_{\mathcal{S}}(\hat{Y} \,|\, X, Y)} \left[\mathbb{I}\{\hat{Y} = Y\} \,|\, X = x \right], \tag{1}$$

where $P(Y \,|\, X)$ denotes the conditional distribution of the ground-truth label Y and $P_{\mathcal{S}}(\hat{Y} \,|\, X, Y)$ denotes the conditional distribution of the expert's predictions $\hat{Y}$ under the prediction set $\mathcal{S}$.[5]

[4] We denote random variables with capital letters and realizations of random variables with lowercase letters.

[5] The expert's prediction $\hat{Y}$ and the ground truth label Y may *not* be conditionally independent given the set of features X since, in most application domains of interest, the expert may have access to additional features. Otherwise, one may argue that pursuing human-AI complementarity is not a worthy goal [47].

https://doi.org/10.52202/079017-0986

3 On the Suboptimality of Conformal Prediction

Given a user-specified parameter $\alpha \in [0, 1]$, a conformal predictor uses a choice of non-conformity score $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and a calibration set $\mathcal{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^m$, where $(x_i, y_i) \sim P(X)P(Y \mid X)$, to construct the prediction sets $\mathcal{S}(X) = \mathcal{S}_{\text{cp}}(X)$ as follows:

$$\mathcal{S}_{\text{cp}}(X) = \{y \mid s(X, y) \leq \hat{q}_\alpha\}, \tag{2}$$

where $\hat{q}_\alpha$ is the $\lceil (m+1)(1-\alpha) \rceil / m$ empirical quantile of the non-conformity scores of the samples in the calibration set $\mathcal{D}_{\text{cal}}$. By using the above construction, the conformal predictor guarantees that the probability that the true label Y belongs to the subset $\mathcal{S}_{\text{cp}}(X)$ is almost exactly $1 - \alpha$, *i.e.*, $1 - \alpha \leq P(Y \in \mathcal{S}_{\text{cp}}(X)) \leq 1 - \alpha + 1/(m + 1)$, as shown elsewhere [20, 21].

Under common choices of non-conformity scores [48, 49], there are many data distributions for which the optimal prediction set under which the human expert achieves the highest accuracy cannot be constructed using a conformal predictor. Consider the following example where $\mathcal{Y} = \{1, 2, 3\}$ and,

$$P(Y = y \mid X = x) = \begin{cases} 0.4 & \text{if } y = 1 \\ 0.35 & \text{if } y = 2 \\ 0.25 & \text{if } y = 3 \end{cases} \text{ and } P_{\mathcal{S}}(\hat{Y} = \hat{y} \mid X = x, Y = y) = \frac{C_{\hat{y}y}}{\sum_{y' \in \mathcal{S}} C_{y'y}},$$

where $C_{1,1} = C_{2,1} = C_{3,1} = 0.33$, $C_{1,2} = C_{1,3} = 0.4$, $C_{2,2} = C_{3,3} = 0.6$, and $C_{3,2} = C_{2,3} = 0$. A brute force search reveals that the optimal prediction set is $\{2, 3\}$ and, under this set, the expert achieves accuracy 0.6. Now, assume we have access to a perfectly calibrated classifier $f(x) \in [0, 1]^L$, *i.e.*, for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, it holds that $f_y(x) = P(Y = y \mid X = x)$. Then, for any choice of $\alpha \in [0, 1]$, as long as the non-conformity scores rank the label set in decreasing order of $f_y(x)$, the prediction set provided by conformal prediction can only be among the sets $\{\emptyset, \{1\}, \{1, 2\}, \{1, 2, 3\}\}$. Among these sets, the set under which the expert achieves the highest accuracy is $\{1, 2, 3\}$ and, under this set, the expert achieves accuracy $0.49 < 0.6$.

Motivated by the above example, one may think of closing the above performance gap by incorporating information about the distribution of experts' predictions in the definition of the non-conformity score. However, we cannot expect to fully close the performance gap since, as we will show next, the problem of finding the optimal prediction sets is NP-hard to solve and approximate to any factor less than the size of the label set $\mathcal{Y}$.

4 On the Hardness of Finding the Optimal Prediction Sets

In this section, we first show that, given $x \in \mathcal{X}$, we cannot expect to find the optimal prediction set $\mathcal{S}^*(x)$ that maximizes the accuracy of the expert's prediction in polynomial time:[6]

Theorem 1 *The problem of finding the optimal prediction set, as defined in Eq. 1, is* NP*-hard.*

In the proof of the above theorem, we first reduce the k-clique problem,[7] which is known to be NP-complete [22], to an instance of the problem of deciding whether there exists a prediction set $\mathcal{S} \subseteq \mathcal{Y}$ such that $g(\mathcal{S} \mid x) \geq B$ given a constant $B > 0$. More specifically, given a k-clique problem defined over a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $k \leq |\mathcal{V}|$, we reduce it to an instance of the above decision problem in which $\mathcal{Y} = \mathcal{V}$, $B = \frac{k}{|\mathcal{V}|}$ and, for all $y \in \mathcal{Y}$, we have that $P(Y = y \mid X = x) = \frac{1}{|\mathcal{V}|}$ and

$$P_{\mathcal{S}}(\hat{Y} = \hat{y} \mid X = x, Y = y) = \frac{C_{\hat{y}y}}{\sum_{y' \in \mathcal{S}} C_{y'y}} \text{ where } C_{y'y} = \begin{cases} 0 & \text{if } (y', y) \in \mathcal{E} \\ 1/\widehat{N}_{\mathcal{G}}(y) & \text{otherwise,} \end{cases} \tag{3}$$

and $\widehat{N}_{\mathcal{G}}(y)$ denotes the number of vertices that are not adjacent to y. Then, since the above decision problem can be trivially reduced to the problem of finding the optimal prediction set (in polynomial time), we conclude that the problem is NP-hard.

Motivated by the above result, we may think in looking for desirable properties for the objective function $g(\mathcal{S} \mid x)$ such as monotonicity and submodularity,[8] which would allow for the design of

[6] All proofs are in Appendix A.

[7] Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and an integer $k \leq |\mathcal{V}|$, the k-clique problem seeks to decide whether there exists $\mathcal{S} \subseteq V$ with size $|\mathcal{S}| = k$ such that, for every $u, v \in S$, there exists $(u, v) \in \mathcal{E}$.

[8] A function $f : 2^{\mathcal{Y}} \to \mathbb{R}$ is submodular if and only if, for every $\mathcal{S} \subseteq \mathcal{T} \subseteq \mathcal{Y}$ and $y \in \mathcal{Y} \backslash \mathcal{T}$, it holds that $f(\mathcal{S} \cup \{y\}) - f(\mathcal{S}) \geq f(\mathcal{T} \cup \{y\}) - f(\mathcal{T})$.

https://doi.org/10.52202/079017-0986

Algorithm 1: Greedy algorithm

Input: Label set $\mathcal{Y}$, features x, classifier f, confusion matrix $\boldsymbol{C}$
Output: Prediction set $\mathcal{S}$

$\mathcal{S} \leftarrow \varnothing$
$\{y_{(1)}, \ldots, y_{(L)}\} \leftarrow \text{argsort} f(x)$ `// Sort in descending order`
for $k \in \{1, \ldots, L\}$ **do**
  $\mathcal{S}_k \leftarrow \varnothing$
  while $|\mathcal{S}_k| < k$ **do** `// Add labels to the prediction set until we hit` k
    $\Delta^* \leftarrow -\infty$
    for $y \in \{y_{(1)}, \ldots, y_{(k)}\} \backslash \mathcal{S}_k$ **do**
      $\Delta \leftarrow \hat{g}(\mathcal{S}_k \cup \{y\} \,|\, x) - \hat{g}(\mathcal{S}_k \,|\, x)$ `// Eval the marginal gain of adding` y `to` $\mathcal{S}_k$
      if $\Delta > \Delta^*$ **then**
        $\Delta^* \leftarrow \Delta, y^* \leftarrow y$
    $\mathcal{S}_k \leftarrow \mathcal{S}_k \cup \{y^*\}$ `// Add label offering the largest marginal gain`
    if $\hat{g}(\mathcal{S}_k \,|\, x) > \hat{g}(\mathcal{S} \,|\, x)$ **then**
      $\mathcal{S} \leftarrow \mathcal{S}_k$ `// Update` $\mathcal{S}$ `if` $\mathcal{S}_k$ `achieves higher objective value`
return $\mathcal{S}$

approximation algorithms with non-trivial approximation guarantees [50]. Unfortunately, there are many data distributions for which the objective function is neither monotone nor submodular. For example, assume $\mathcal{Y} = \{1, 2, 3\}$,

$$P(Y = y \,|\, X = x) = \begin{cases} 0.4 & \text{if } y = 1 \\ 0.35 & \text{if } y = 2 \\ 0.25 & \text{if } y = 3 \end{cases} \text{ and } P_{\mathcal{S}}(\hat{Y} = \hat{y} \,|\, X = x, Y = y) = \frac{C_{\hat{y}y}}{\sum_{y' \in \mathcal{S}} C_{y'y}},$$

where $C_{1,1} = 0.2$, $C_{1,2} = C_{2,1} = C_{1,3} = C_{3,1} = 0.4$, $C_{2,2} = C_{3,3} = 0.6$ and $C_{2,3} = C_{3,2} = 0$. For $\mathcal{S} = \{1\} \subseteq \mathcal{T} = \{1, 2\} \subset \mathcal{Y}$, it holds that $g(\mathcal{S} \,|\, x) = 0.4 > g(\mathcal{T} \,|\, x) = 0.34$ and $g(\mathcal{T} \,|\, x) = 0.34 < g(\mathcal{Y} \,|\, x) = 0.44$, and thus we can conclude it is not monotone. Moreover, it also holds that $g(\mathcal{S} \cup \{3\} \,|\, x) - g(\mathcal{S} \,|\, x) = -0.116 < g(\mathcal{T} \cup \{3\} \,|\, x) - g(\mathcal{T} \,|\, x) = 0.096$, and thus we can conclude it is not submodular.

In fact, the following theorem shows that we cannot expect to find a polynomial-time algorithm to find a non-trivial approximation to our problem:

Theorem 2 *The problem of finding the optimal prediction set, as defined in Eq. 1, is* NP*-hard to approximate to any factor less than the size L of the label set $\mathcal{Y}$.*

In the proof of the above theorem, we first show that, given a polynomial-time α-approximation algorithm for the problem of finding the optimal prediction set, we can obtain a polynomial-time α-approximation algorithm for the problem of finding the maximum clique in a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.[9] Then, since it is known that, for any $\epsilon > 0$, the latter problem is NP-hard to approximate to a factor $|\mathcal{V}|^{1-\epsilon}$ [51], we can conclude that the problem of finding the optimal prediction set is NP-hard to approximate to a factor $|\mathcal{Y}|^{1-\epsilon}$.

While the above hardness results may be discouraging, in what follows, we will introduce a simple greedy algorithm that provably offers equal or greater performance than conformal prediction for a large class of non-conformity scores and expert models, and in practice, often succeeds at finding (near-)optimal prediction sets.

A simple greedy algorithm. Given a sample with features $x \in \mathcal{X}$ and a prediction set $\mathcal{S} \subseteq \mathcal{Y}$, our greedy algorithm estimates the accuracy of the expert's prediction, as defined in Eq. 1, using the following estimator:

$$\hat{g}(\mathcal{S} \,|\, x) = \sum_{y \in \mathcal{S}} \underbrace{f_y(x)}_{(a)} \underbrace{\frac{C_{yy}}{\sum_{y' \in \mathcal{S}} C_{y'y}}}_{(b)} \tag{4}$$

[9] Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, the maximum clique problem seeks to find the largest $\mathcal{S} \subseteq \mathcal{V}$ such that, for every $u, v \in \mathcal{S}$, there exists $(u, v) \in \mathcal{E}$.

https://doi.org/10.52202/079017-0986

where (a) approximates $P(Y = y \,|\, X = x)$ using a well-calibrated classifier $f(x) \in [0, 1]^L$ and, similarly as in Straitouri et al. [18], (b) approximates $P_{\mathcal{S}}(\hat{Y} = y \,|\, X = x, Y = y)$ using a mixture of multinomial logit models (MNLs) parameterized by the confusion matrix of the predictions made by the expert on their own, *i.e.*, $C_{y'y} = P_{\mathcal{Y}}(\hat{Y} = y' \,|\, Y = y)$.

The greedy algorithm first ranks each label value $y \in \mathcal{Y}$ using the output $f_y(x)$ of the classifier. Let $y_{(1)}, \ldots, y_{(L)}$ be the label values ordered according to such a ranking, where $\cdot_{(i)}$ denotes the i-th label value in the ranking and $f_{y_{(i)}}(x) \geq f_{y_{(j)}}(x)$ for all $i < j$. Then, it runs L rounds and, at each k-th round, it starts from the prediction set $\mathcal{S}_k = \emptyset$ and iteratively adds to $\mathcal{S}_k$ the label value $y \in \{y_{(1)}, \ldots, y_{(k)}\} \backslash \mathcal{S}_k$ that provides the maximum marginal gain $\hat{g}(\mathcal{S}_k \cup \{y\} \,|\, x) - \hat{g}(\mathcal{S}_k \,|\, x)$ until it exhausts the set $\{y_{(1)}, \ldots, y_{(k)}\}$. Moreover, at each iteration and round, it keeps track of the set with the highest objective value. At each of the L runs of the greedy algorithm, at most L elements are added to the set $\mathcal{S}$, and adding each element needs at most L times computing the marginal gain $\hat{g}(\mathcal{S} \cup \{y\} \,|\, x) - \hat{g}(\mathcal{S} \,|\, x)$, which takes $O(L)$ to compute. Hence, our algorithm has an overall complexity of $O(L^4)$. See Appendix B for a detailed running time analysis. Algorithm 1 provides a pseudocode implementation of the procedure.

Importantly, the prediction sets provided by the greedy algorithm are guaranteed to achieve higher objective value $\hat{g}$ than those provided by any conformal predictor using a non-conformity score $s(x, y)$ that is nonincreasing with respect to $f_y(x)$, as formalized by the following proposition:[10]

Proposition 1 *For any $x \in \mathcal{X}$, let $\mathcal{S}$ be the prediction set provided by Algorithm 1 and $\mathcal{S}_{cp}$ be the prediction set provided by any conformal prediction with a non-conformity score $s(x, y)$ that is nonincreasing with respect to $f_y(x)$, then, it holds that $\hat{g}(\mathcal{S} \,|\, x) \geq \hat{g}(\mathcal{S}_{cp} \,|\, x)$.*

5 Experiments with Synthetic Data

In this section, we compare the average accuracy achieved by different simulated human experts using prediction sets constructed with our greedy algorithm (Algorithm 1), brute force search, and conformal prediction on several synthetic multiclass classification tasks where the experts and the classifier used by the greedy algorithm, brute force search, and conformal prediction achieve different accuracies on their own.

Experimental setup. We create several synthetic multiclass classification tasks, each with $n = 20$ features per sample and varying difficulty. Out of 20 features per sample, only $d = 4$ of these features are *informative*[11] while the rest are drawn at random. Refer to Appendix C for more details about the classification tasks. For each classification task, we generate 19,000 samples, which we split into a training set (16,000 samples), a calibration set (1000 samples), a validation set (1000 samples) and a test set (1000 samples).

We use the first half of the samples in the training set to train a multinomial logistic regression model $f(x)$. This model is used by the greedy algorithm, brute force search and conformal prediction. It achieves a different average test accuracy $P(Y' = Y)$, depending on the difficulty of the classification task. We use the second half of the samples in the training set to train another multinomial logistic regression model $\hat{f}(x)$. However, during the training of this model, we modify the value a of one of the (informative) features of each training sample to $(1 - \gamma)a + \gamma\epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$ and $\gamma \in [0, 1]$ controls the average accuracy of the resulting model. Then, we use the (estimated) confusion matrix $\boldsymbol{C}(\gamma)$ of the predictions made by $\hat{f}(x)$ to model (the predictions made by) the simulated expert by means of a mixture of MNLs, *i.e.*, $P_{\mathcal{S}}(\hat{Y} = y \,|\, X = x, Y = y) = \frac{C_{yy}(\gamma)}{\sum_{y' \in \mathcal{S}} C_{y'y}(\gamma)}$.

Further, we use the calibration set to calibrate the (softmax) outputs of the logistic regression model f using top-k-label calibration [52] with $k = 5$. We also use it to estimate the confusion matrices $\boldsymbol{C}(\gamma)$ that parameterize the mixture of MNLs used to model (the predictions made by) the simulated human expert, and calculate the quantile $\hat{q}_\alpha$ used by conformal prediction. Finally, we use the test set to evaluate the average accuracy achieved by the simulated expert using prediction sets constructed with our greedy algorithm, brute force search and conformal prediction. Here, note that

[10] Proposition 1 can be generalized to conformal predictors with any non-conformity score as long as the ranking that our greedy algorithm uses is the same as the ranking induced by the non-conformity scores.

[11] A feature is informative if its value correlates with the label value.

Table 1: Empirical average test accuracy achieved by four different (simulated) human experts, each with a different noise value γ, on their own (NONE) and using prediction sets constructed with conformal prediction (NAIVE, APS, RAPS and SAPS) and with the greedy algorithm (GREEDY) on four synthetic classification tasks. In each classification task, the classifier f used by conformal prediction and the greedy algorithm achieves a different average accuracy $P(Y' = Y)$. The number of labels is $L = 10$, the size of the calibration set is $m = 1000$, and we do not include brute force search because it achieves the same performance as the greedy algorithm. Each cell shows the average and standard deviation over 10 runs. We denote the best results for each classification task in bold.

γ	METHOD	$P(Y' = Y) = 0.3$	$P(Y' = Y) = 0.5$	$P(Y' = Y) = 0.7$	$P(Y' = Y) = 0.9$
0.3	NAIVE	0.340 ± 0.014	0.588 ± 0.015	0.799 ± 0.013	0.944 ± 0.006
	APS	0.341 ± 0.013	0.587 ± 0.013	0.804 ± 0.015	0.941 ± 0.006
	RAPS	0.341 ± 0.013	0.587 ± 0.013	0.804 ± 0.014	0.941 ± 0.006
	SAPS	0.340 ± 0.015	0.585 ± 0.012	0.804 ± 0.015	0.940 ± 0.008
	GREEDY	$\mathbf{0.364 \pm 0.015}$	$\mathbf{0.605 \pm 0.014}$	$\mathbf{0.824 \pm 0.012}$	$\mathbf{0.953 \pm 0.005}$
	NONE	0.281 ± 0.018	0.485 ± 0.019	0.693 ± 0.018	0.883 ± 0.008
0.5	NAIVE	0.328 ± 0.014	0.564 ± 0.012	0.774 ± 0.014	0.932 ± 0.007
	APS	0.329 ± 0.012	0.565 ± 0.010	0.787 ± 0.013	0.932 ± 0.008
	RAPS	0.330 ± 0.012	0.566 ± 0.010	0.787 ± 0.013	0.932 ± 0.008
	SAPS	0.329 ± 0.013	0.563 ± 0.008	0.787 ± 0.014	0.932 ± 0.009
	GREEDY	$\mathbf{0.353 \pm 0.015}$	$\mathbf{0.587 \pm 0.010}$	$\mathbf{0.805 \pm 0.012}$	$\mathbf{0.945 \pm 0.004}$
	NONE	0.261 ± 0.016	0.446 ± 0.013	0.644 ± 0.019	0.843 ± 0.011
0.7	NAIVE	0.319 ± 0.012	0.534 ± 0.013	0.737 ± 0.013	0.908 ± 0.006
	APS	0.320 ± 0.008	0.542 ± 0.012	0.759 ± 0.015	0.913 ± 0.008
	RAPS	0.320 ± 0.008	0.542 ± 0.012	0.760 ± 0.015	0.914 ± 0.008
	SAPS	0.319 ± 0.009	0.534 ± 0.012	0.760 ± 0.014	0.915 ± 0.009
	GREEDY	$\mathbf{0.345 \pm 0.011}$	$\mathbf{0.573 \pm 0.010}$	$\mathbf{0.784 \pm 0.013}$	$\mathbf{0.938 \pm 0.006}$
	NONE	0.238 ± 0.011	0.380 ± 0.015	0.540 ± 0.018	0.716 ± 0.013
1.0	NAIVE	0.314 ± 0.015	0.517 ± 0.011	0.714 ± 0.015	0.894 ± 0.012
	APS	0.316 ± 0.013	0.525 ± 0.009	0.733 ± 0.014	0.895 ± 0.010
	RAPS	0.316 ± 0.013	0.525 ± 0.009	0.734 ± 0.015	0.896 ± 0.010
	SAPS	0.315 ± 0.014	0.517 ± 0.011	0.734 ± 0.015	0.896 ± 0.009
	GREEDY	$\mathbf{0.348 \pm 0.013}$	$\mathbf{0.567 \pm 0.011}$	$\mathbf{0.782 \pm 0.015}$	$\mathbf{0.936 \pm 0.007}$
	NONE	0.214 ± 0.017	0.303 ± 0.014	0.382 ± 0.021	0.452 ± 0.021

our greedy algorithm and brute force search have access to the true mixtures of MNLs used to model the simulated human expert. In our experiments, we implement conformal prediction using several non-conformity scores:

$$s(x,y) = 1 - f_y(x) \text{ (NAIVE, [20])}, \quad s(x,y) = \sum_{y':f_{y'}(x) \leq f_y(x)} f_{y'}(x) \text{ (APS, [53])},$$

$$s(x,y) = f_y(x) + \sum_{y':f_{y'}(x) \leq f_y(x)} f_{y'}(x) + \lambda_{raps} \left(o(x,y) - k_{reg}\right)^+ \qquad \text{(RAPS, [49])},$$

$$s(x,y) = \begin{cases} \max_{y'} f_{y'}(x) & o(x,y) = 1 \\ \max_{y'} f_{y'}(x) + \lambda_{saps} \left(o(x,y) - 2\right) & o(x,y) > 1 \end{cases} \qquad \text{(SAPS, [54])},$$

where $o(x,y) = |\{y' : f_{y'}(x) \leq f_y(x)\}|$ denotes the ranking of label y according to $f_y(x)$ and we decided to omit the randomization for APS, RAPS and SAPS as it is only required to achieve exact $1 - \alpha$ coverage and it did not have an influence on the empirical average accuracy achieved by the simulated human experts in our experiments. For RAPS and SAPS, we run the procedure outlined in Appendix E in Angelopoulous et al. [49] to optimize the additional hyperparameters, k_{regs} and λ_{raps}, for RAPS, and λ_{saps} for SAPS, using the validation set. Further, for each non-conformity score and classification task, we report the results for the α value under which the expert achieves the highest average test accuracy and, to avoid empty sets, we always include the label value with the lowest non-conformity score in the prediction sets. In this context, note that, in practice, one would need to select α using a held-out dataset, however, our evaluation aims to show how our greedy algorithm

https://doi.org/10.52202/079017-0986

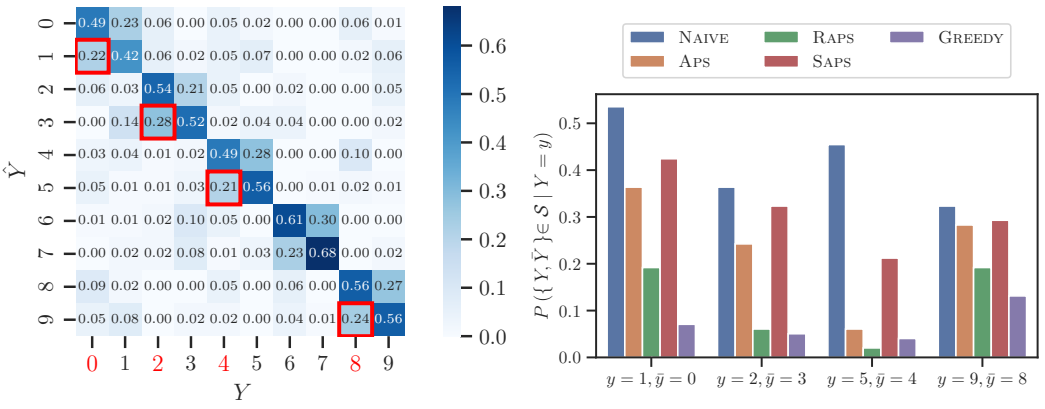

Figure 2: (Left) Confusion matrix $\boldsymbol{C}$ for the predictions made by a (simulated) human expert on their own. The label $\bar{y} = \text{argmax}_{y' \neq y} C_{y'y}$ that is most frequently mistaken with the ground truth-label y is highlighted in red for $y \in \{0, 2, 6, 8\}$. (Right) Empirical conditional probability that a prediction set includes $\{y, \bar{y}\}$ given $Y = y$ with conformal prediction (NAIVE, APS, RAPS and SAPS) and our greedy algorithm (GREEDY). In both panels, $\gamma = 0.7$ and $\mathbb{P}(Y' = Y) = 0.7$.

improves over conformal prediction for *any* value of α. Finally, we repeat each experiment ten times and, each time, we sample different training, calibration, validation and test sets.

Results. We first estimate the average test accuracy achieved by four different (simulated) human experts, each with a different γ value, on four classification tasks where the classifier f achieves a different average accuracy $P(Y' = Y)$. We report their average test accuracy on their own (NONE) and when using prediction sets constructed with conformal prediction (NAIVE, APS, RAPS and SAPS), our greedy algorithm (GREEDY) and brute force search (BRUTE FORCE SEARCH). Table 1 summarizes the results, where we have not included brute force search because it achieves the same performance as our greedy algorithm. The results show that, using the greedy algorithm to construct prediction sets, the experts consistently achieve the highest average accuracy across classification tasks. Moreover, the results also show that, under the prediction sets constructed using the greedy algorithm, the average accuracy achieved by the expert degrades gracefully as γ increases whereas, under the prediction sets constructed using conformal prediction, the average accuracy degrades significantly. Refer to Appendix D for additional results for $L \in \{25, 50\}$ showing that the relative gain in average accuracy offered by the greedy algorithm increases with the number of labels and noise γ. Refer to Appendix E for additional results showing that the empirical average coverage achieved by the prediction sets constructed using conformal prediction and our greedy algorithm may be a bad proxy for estimating the average accuracy achieved by human experts using prediction sets.

To better understand why the prediction sets constructed by the greedy algorithm help human experts achieve higher average accuracy than those constructed by conformal prediction, we now look closer into the structure of the prediction sets. Given a ground truth-label $Y = y$, let $\bar{y} = \text{argmax}_{y' \neq y} C_{y'y}$ be the label that is most frequently mistaken with y. Then, we estimate the empirical conditional probability that a prediction set includes $\{y, \bar{y}\}$ given $Y = y$ with the greedy algorithm and conformal prediction. Figure 2 summarizes the results for $\gamma = 0.7$ and $P(Y' = Y) = 0.7$ and $y \in \{0, 2, 6, 8\}$. Appendix F includes additional results for other configurations. The results show that, with the greedy algorithm, the empirical probability that a prediction set includes $\{y, \bar{y}\}$ given $Y = y$ is much lower (*i.e.*, 2-3x lower) than with conformal prediction despite it creates overall larger prediction sets.

6 Experiments with Real Data

In this section, we compare the average accuracy achieved by a simulated human expert using prediction sets constructed with our greedy algorithm, brute force search and conformal prediction on a real multiclass classification task over noisy natural images. The simulated human expert follows

Table 2: Average test accuracy achieved by a (simulated) human expert on their own (NONE), and using the prediction sets constructed with conformal prediction (NAIVE, APS, RAPS and SAPS), and our greedy algorithm (GREEDY) on the ImageNet16H dataset. We do not include brute force search because it achieves the same performance as the greedy algorithm. Each cell shows the average and standard deviation over 10 runs. We denote the best results in bold.

METHOD	$\omega = 80$	$\omega = 95$	$\omega = 110$	$\omega = 125$
NAIVE	$\mathbf{0.957 \pm 0.006}$	0.946 ± 0.008	0.919 ± 0.006	0.860 ± 0.008
APS	0.944 ± 0.004	0.932 ± 0.005	0.902 ± 0.008	0.852 ± 0.010
RAPS	0.950 ± 0.009	0.943 ± 0.010	0.914 ± 0.010	0.849 ± 0.011
SAPS	0.953 ± 0.010	0.942 ± 0.009	0.918 ± 0.009	0.855 ± 0.007
GREEDY	$\mathbf{0.957 \pm 0.007}$	$\mathbf{0.951 \pm 0.009}$	$\mathbf{0.925 \pm 0.008}$	$\mathbf{0.874 \pm 0.009}$
NONE	0.900 ± 0.002	0.859 ± 0.003	0.771 ± 0.005	0.603 ± 0.007

the mixture of MNLs introduced in Eq. 4, which is parameterized by the (estimated) confusion matrix of the predictions made by real human experts on their own.[12]

Experimental setup. We experiment with the ImageNet16H dataset [7], which was created using 1,200 natural images from the ImageNet Large Scale Visual Recognition Challenge (ILSRVR) 2012 dataset [55]. More specifically, in the ImageNet16H dataset, each of the above images was used to create four noisy images with different levels of phase noise distortion $\omega \in \{80, 95, 110, 125\}$ and the same ground-truth label y from a label set $\mathcal{Y}$ of size $n = 16$. In addition, for each noisy image, the dataset contains (approximately) six label predictions made by human experts on their own. In our experiments, we run and evaluate each method separately by grouping the above noisy images (and expert predictions) according to their level of noise. For each group of images and method, we use the deep neural network classifier VGG-19 [56] after 10 epochs of fine-tuning as provided by Steyvers et al. [7]. Further, we randomly split the images (and expert predictions) in each group into two disjoint subsets, a calibration set (800 images), and a test set (400 images). The accuracy of the (pretrained) VGG-19 classifier on the test set is 0.900 ± 0.014 ($\omega = 80$), 0.895 ± 0.009 ($\omega = 95$), 0.857 ± 0.016 ($\omega = 110$) and 0.792 ± 0.016 ($\omega = 125$). We use the calibration set to (i) calibrate the (softmax) outputs of the VGG-19 scores using top-k-label calibration with $k = 5$, (ii) estimate the confusion matrix $\mathbf{C}$ that parameterizes the mixture of MNLs used to model the simulated human expert, and (iii) calculate the quantile $\hat{q}_\alpha$ used by conformal prediction[13].We use the test set to evaluate the average accuracy the simulated expert achieves using prediction sets constructed with our greedy algorithm, brute force search and conformal prediction. Here, note that, similarly to in the experiments in synthetic data, our greedy algorithm and brute force search have access to the true mixture of MNLs used to model the simulated human expert. We implement conformal prediction using the same non-conformity scores used in the experiments with synthetic data and, for each non-conformity score and group of images, we report the results for the α value under which the expert achieves the highest average test accuracy. In Appendix H, we report results for all α values. To obtain error bars, we repeat each experiment 10 times, sampling different calibration and test sets.

Results. Table 2 and Figure 3 show the average test accuracy and complementary cumulative distribution (cCDF) of the test per-image accuracy achieved by a simulated human expert using the prediction sets constructed with conformal prediction (NAIVE, APS, RAPS and SAPS), our greedy algorithm (GREEDY) and brute force search (BRUTE FORCE SEARCH) for different values of noise ω. The results show that, similarly as in our experiments with synthetic data, the greedy algorithm achieves the same performance as brute force search. Moreover, they also show that, using greedy algorithm to construct prediction sets, the expert achieves the highest average accuracy in all groups of images except the group with $\omega = 80$, where both the greedy algorithm and one of the conformal predictors offer comparable performance. Similarly as in the synthetic experiments, refer to Appendix E for additional results regarding the empirical average coverage achieved by the prediction sets constructed using conformal prediction and our greedy algorithm.

[12]In Appendix G, we evaluate the goodness of fit of the mixture of MNLs to predictions made by real human experts using a support system based on prediction sets [19].

[13]For RAPS and SAPS, we further split the calibration set to obtain a (reduced) calibration (400 images) and validation (400 images) sets to calculate the quantile $\hat{q}_\alpha$ and optimize the hyperparameters of RAPS and SAPS, respectively.

https://doi.org/10.52202/079017-0986

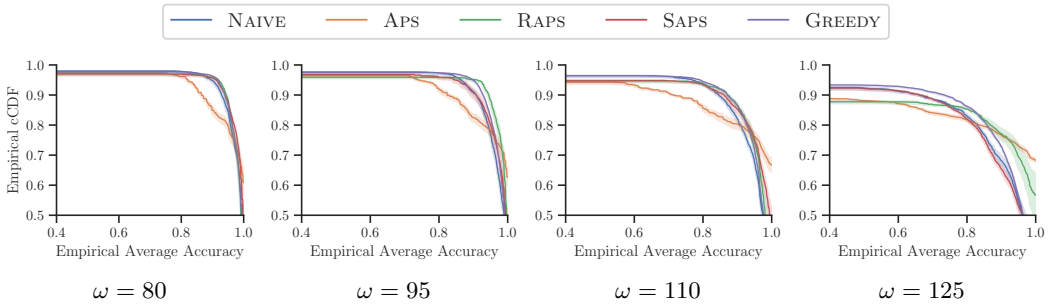

Figure 3: Complementary cumulative distribution (cCDF) of the per-image test accuracy achieved by a simulated human expert using the prediction sets constructed with conformal prediction (NAIVE, APS, RAPS and SAPS) and our greedy algorithm (GREEDY) on the ImageNet16H dataset.

7 Discussion and Limitations

In this section, we discuss several assumptions and limitations of our work, which open up interesting avenues for future work.

Hardness analysis. In our hardness analysis, our reduction utilizes an instance of our problem in which, for every prediction set, the predictions made by experts follow a mixture of MNLs. As an immediate consequence, this implies that, in general, the problem of finding the prediction set $\mathcal{S}$ that maximizes $\hat{g}(\mathcal{S} \mid x)$ is NP-hard to approximate. However, in our experiments, the greedy algorithm is almost always able to find such a set $\mathcal{S}$. As a result, we hypothesize that there may be certain conditions on the parameters of the mixture of MNLs under which the problem can be efficiently approximated to a factor less than the size of the label set.

Methodology. Our greedy algorithm assumes that, for every prediction set, the predictions made by the human expert follow a parameterized expert model—the above mentioned mixture of MNLs. It would be worthy to develop model-free algorithms since, in the context of prediction sets constructed using conformal prediction, they have been shown to be superior to their model-based counterparts [19]. To this end, a good starting point may be the literature on (contextual) combinatorial multi-armed bandits [57, 58], where one can map each arm to a label value and each subset of arms, namely a super arm, to a prediction set.

Evaluation. The results of our experiments suggest that the prediction sets constructed using our greedy algorithm may help human experts make more accurate predictions than the prediction sets constructed using conformal prediction. However, one may argue that the difference in performance is partly due to the fact that the non-conformity scores used in conformal prediction do not incorporate information about the distribution of experts' predictions. Motivated by this observation, it would be important to investigate how to incorporate such information in the definition of non-conformity scores. Moreover, in our experiments, the true distribution of experts' predictions matches the mixture of MNLs used by the greedy algorithm and brute force search. However, in practice, there may be a mismatch between the true distribution of experts' predictions and the mixture of MNLs, and this may decrease performance. Finally, in our experiments with real data, the ground truth labels are estimated by aggregating (multiple) predictions by human annotators using majority voting, however, this may introduce additional sources of errors that may influence our results [59].

Broader impact. We have focused on maximizing the average accuracy of the predictions made by an expert using a decision support system based on prediction sets. However, in high-stakes application domains, it would be important to extend our methodology to account for fairness considerations.

8 Conclusions

We have looked at the problem of finding the optimal prediction sets under which human experts achieve the highest accuracy in a given multiclass classification task. We have shown that this problem is NP-hard to solve and to approximate to any factor less than the size of the label set. However, we have empirically shown that, for a large parameterized class of expert models, a simple greedy algorithm consistently outperforms conformal prediction.

Acknowledgments and Disclosure of Funding

Gomez-Rodriguez acknowledges support from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 945719). De Toni acknowledges support from the TANGO project (grant #101120763-TANGO). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the EU or HaDEA. Neither the EU nor the granting authority can be held responsible for them.

References

[1] Wei Jiao, Gurnit Atwal, Paz Polak, Rosa Karlic, Edwin Cuppen, Alexandra Danyi, Jeroen de Ridder, Carla van Herpen, Martijn P Lolkema, et al. A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nature Communications*, 11(1):728, 2020.

[2] Jacob Whitehill, Kiran Mohan, Daniel Seaton, Yigal Rosen, and Dustin Tingley. Mooc dropout prediction: How to measure accuracy? In *Proceedings of the ACM conference on learning@ scale*, pages 161–164. ACM, 2017.

[3] Alex Davies, Petar Veličković, Lars Buesing, Sam Blackwell, Daniel Zheng, Nenad Tomašev, Richard Tanburn, Peter Battaglia, Charles Blundell, András Juhász, et al. Advancing mathematics by guiding human intuition with AI. *Nature Communications*, 600(7887):70–74, 2021.

[4] Abir De, Paramita Koley, Niloy Ganguly, and Manuel Gomez-Rodriguez. Regression under human assistance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2611–2620, 2020.

[5] Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans. In *Proceedings of the International Joint Conference on Artificial Intelligence*. IJCAI, 2020.

[6] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. Is the most accurate ai the best teammate? optimizing ai for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11405–11414, 2021.

[7] 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.

[8] Kori Inkpen, Shreya Chappidi, Keri Mallari, Besmira Nushi, Divya Ramesh, Pietro Michelucci, Vani Mandava, LibuŠe Hannah VepŘek, and Gabrielle Quinn. Advancing human-ai complementarity: The impact of user expertise and algorithmic tuning on joint decision making. *Transactions on Computer-Human Interaction*, 30(5):1–29, 2023.

[9] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. How model accuracy and explanation fidelity influence user trust. *arXiv preprint arXiv:1907.12652*, 2019.

[10] Xinru Wang and Ming Yin. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *Proceedings of the International Conference on Intelligent User Interfaces*, pages 318–328, 2021.

[11] Kailas Vodrahalli, Tobias Gerstenberg, and James Zou. Uncalibrated models can improve human-AI collaboration. In *Advances in Neural Information Processing Systems*, 2022.

[12] Han Liu, Yizhou Tian, Chacha Chen, Shi Feng, Yuxin Chen, and Chenhao Tan. Learning human-compatible representations for case-based decision support. In *Proceedings of the International Conference on Learning Representations*, 2023.

[13] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the CHI conference on human factors in computing systems*, pages 1–12, 2019.

 https://doi.org/10.52202/079017-0986

[14] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 295–305. ACM, 2020.

[15] Harini Suresh, Natalie Lao, and Ilaria Liccardi. Misplaced trust: Measuring the interference of machine learning in human decision-making. In *Proceedings of the ACM Conference on Web Science*, pages 315–324. ACM, 2020.

[16] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471*, 2021.

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

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

[19] Eleni Straitouri and Manuel Gomez-Rodriguez. Designing decision support systems using counterfactual prediction sets. In *Proceedings of the International Conference on Machine Learning*. PMLR, 2024.

[20] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

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

[22] Richard M. Karp. *Reducibility among Combinatorial Problems*, pages 85–103. Springer, 1972.

[23] Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, and Titouan Lorieul. Set-valued classification–overview via a unified framework. *arXiv preprint arXiv:2102.12318*, 2021.

[24] Gen Yang, Sébastien Destercke, and Marie-Hélène Masson. Cautious classification with nested dichotomies and imprecise probabilities. *Soft Computing*, 21(24):7447–7462, 2017.

[25] Thomas Mortier, Marek Wydmuch, Krzysztof Dembczyński, Eyke Hüllermeier, and Willem Waegeman. Efficient set-valued prediction in multi-class classification. *Data Mining and Knowledge Discovery*, 35(4):1435–1469, 2021.

[26] Liyao Ma and Thierry Denoeux. Partial classification in the belief function framework. *Knowledge-Based Systems*, 214:106742, 2021.

[27] Vu-Linh Nguyen and Eyke Hüllermeier. Multilabel classification with partial abstention: Bayes-optimal prediction under label independence. *Journal of Artificial Intelligence Research*, 72:613–665, 2021.

[28] Varun Babbar, Umang Bhatt, and Adrian Weller. On the utility of prediction sets in human-ai teams. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 2457–2463. IJCAI, 7 2022.

[29] Jesse C Cresswell, Yi Sui, Bhargava Kumar, and Noël Vouitsis. Conformal prediction sets improve human decision making. *arXiv preprint arXiv:2401.13744*, 2024.

[30] 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 ACM CHI conference on Human Factors in Computing Systems*, 2024.

[31] Kate Donahue, Sreenivas Gollapudi, and Kostas Kollias. When are two lists better than one?: Benefits and harms in joint decision-making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10030–10038, 2024.

[32] Kalyan Talluri and Garrett Van Ryzin. Revenue management under a general discrete choice model of consumer behavior. *Management Science*, 50(1):15–33, 2004.

[33] Paat Rusmevichientong, Zuo-Jun Max Shen, and David B Shmoys. Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Operations research*, 58(6):1666–1680, 2010.

[34] James Davis, Guillermo Gallego, and Huseyin Topaloglu. Assortment planning under the multinomial logit model with totally unimodular constraint structures. *Work in Progress*, 2013.

[35] Omar El Housni and Huseyin Topaloglu. Joint assortment optimization and customization under a mixture of multinomial logit models: On the value of personalized assortments. *Operations research*, 71(4):1197–1215, 2023.

[36] Rajan Udwani. Submodular order functions and assortment optimization. In *Proceedings of the International Conference on Machine Learning*, pages 34584–34614. PMLR, 2023.

[37] Maithra Raghu, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. The algorithmic automation problem: Prediction, triage, and human effort. *arXiv preprint arXiv:1903.12220*, 2019.

[38] Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the International Conference on Machine Learning*, pages 7076–7087. PMLR, 2020.

[39] Abir De, Nastaran Okati, Ali Zarezade, and Manuel Gomez Rodriguez. Classification under human assistance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5905–5913, 2021.

[40] Nastaran Okati, Abir De, and Manuel Rodriguez. Differentiable learning under triage. In *Advances in Neural Information Processing Systems*, volume 34, pages 9140–9151, 2021.

[41] Mohammad-Amin Charusaie, Hussein Mozannar, David Sontag, and Samira Samadi. Sample efficient learning of predictors that complement humans. In *Proceedings of the International Conference on Machine Learning*, pages 2972–3005. PMLR, 2022.

[42] Hussein Mozannar, Hunter Lang, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David Sontag. Who should predict? exact algorithms for learning to defer to humans. In *Proceedings of the International conference on artificial intelligence and statistics*, pages 10520–10545. PMLR, 2023.

[43] Eleni Straitouri, Adish Singla, Vahid Balazadeh Meresht, and Manuel Gomez-Rodriguez. Reinforcement learning under algorithmic triage. *arXiv preprint arXiv:2109.11328*, 2021.

[44] Vahid Balazadeh, Abir De, Adish Singla, and Manuel Gomez Rodriguez. Learning to switch among agents in a team via 2-layer markov decision processes. *Transactions on Machine Learning Research*, 2022.

[45] Andrew Fuchs, Andrea Passarella, and Marco Conti. Optimizing delegation between human and ai collaborative agents. *arXiv preprint arXiv:2309.14718*, 2023.

[46] Stratis Tsirtsis, Manuel Gomez Rodriguez, and Tobias Gerstenberg. Responsibility judgments in sequential human-ai collaboration. In *Proceedings of the Annual Conference of the Cognitive Science Society*, 2024.

[47] Rohan Alur, Loren Laine, Darrick Li, Manish Raghavan, Devavrat Shah, and Dennis Shung. Auditing for human expertise. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[48] Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

[49] Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. In *Proceedings of the International Conference on Learning Representations*, 2021.

[50] Andreas Krause and Daniel Golovin. Submodular function maximization. *Tractability*, 3(71-104):3, 2014.

 https://doi.org/10.52202/079017-0986

[51] David Zuckerman. Linear degree extractors and the inapproximability of max clique and chromatic number. In *Proceedings of the ACM Symposium on Theory of Computing*, page 681–690. ACM, 2006.

[52] Chirag Gupta and Aaditya Ramdas. Top-label calibration and multiclass-to-binary reductions. In *Proceedings of the International Conference on Learning Representations*, 2021.

[53] Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3581–3591. Curran Associates, Inc., 2020.

[54] Jianguo Huang, Huajun Xi, Linjun Zhang, Huaxiu Yao, Yue Qiu, and Hongxin Wei. Conformal prediction for deep classifier via label ranking. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 20331–20347. PMLR, 21–27 Jul 2024.

[55] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[56] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[57] Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *Proceedings of the International conference on machine learning*, pages 151–159. PMLR, 2013.

[58] Tyler Lu, Dávid Pál, and Martin Pál. Contextual multi-armed bandits. In *Proceedings of the International conference on Artificial Intelligence and Statistics*, pages 485–492. JMLR, 2010.

[59] David Stutz, Abhijit Guha Roy, Tatiana Matejovicova, Patricia Strachan, Ali Taylan Cemgil, and Arnaud Doucet. Conformal prediction under ambiguous ground truth. *Transactions on Machine Learning Research*, 2023.

A Proofs

A.1 Proof of Proposition 1

We prove by contradiction. Assume there exist $x \in \mathcal{X}$ such that $\hat{g}(\mathcal{S} \mid x) < \hat{g}(\mathcal{S}_{\text{cp}} \mid x)$. Let $k := |\mathcal{S}_{\text{cp}}|$ and $\mathcal{S}_k^*$ be the set providing the highest objective among all the sets seen at the k-th round of the greedy algorithm. It should hold that:

$$\hat{g}(\mathcal{S}_{\text{cp}} \mid x) \underset{(i)}{=} \hat{g}(\{y_{(1)}, \ldots, y_{(k)}\} \mid x) \underset{(ii)}{\leq} \hat{g}(\mathcal{S}_k^* \mid x) \underset{(iii)}{\leq} \hat{g}(\mathcal{S} \mid x),$$

which is a contradiction, hence, it should hold that $\hat{g}(\mathcal{S} \mid x) \geq \hat{g}(\mathcal{S}_{\text{cp}} \mid x)$. Note that (i) is due to the fact that the ranking imposed by the non-conformity scores are the same as the ranking considered by our greedy algorithm so whenever the conformal prediction outputs a set of size k, the k elements correspond to same k elements that are considered in the k-th round in the greedy algorithm, *i.e.*, $\{y_{(1)}, \ldots, y_{(k)}\}$; (ii) is due to the fact that $\mathcal{S}_k^*$ is the best set at the k-th round of the greedy algorithm; and (iii) is because $\mathcal{S} = \text{argmax}_{\mathcal{S}_k^* \in \{\mathcal{S}_1^*, \ldots, \mathcal{S}_L^*\}} \hat{g}(\mathcal{S}_k^* \mid x)$. ∎

A.2 Proof of Theorem 1

To establish NP-hardness we reduce an instance $\langle \mathcal{G} = (\mathcal{V}, \mathcal{E}), k \rangle$ of the k-clique problem,[14] which is known to be NP-complete [22], to an instance $\langle x, \mathcal{Y}, B, \boldsymbol{C} \rangle$ of deciding whether there exists a prediction set $\mathcal{S} \subseteq \mathcal{Y}$ such that $g(\mathcal{S} \mid x) \geq B$ given a constant $B > 0$, as follows: $x \in \mathbb{R}^d$, $\mathcal{Y} = \mathcal{V}$, $B = \frac{k}{|\mathcal{V}|}$ and, for all $y \in \mathcal{Y}$, we have that $P(Y = y \mid X = x) = \frac{1}{|\mathcal{V}|}$ and

$$P_{\mathcal{S}}(\hat{Y} = \hat{y} \mid X = x, Y = y) = \frac{C_{\hat{y}y}}{\sum_{y' \in \mathcal{S}} C_{y'y}} \text{ where } C_{y'y} = \begin{cases} 0 & \text{if } (y', y) \in \mathcal{E} \\ 1/\widehat{N}_{\mathcal{V}}(y) & \text{otherwise,} \end{cases}$$

and $\widehat{N}_{\mathcal{V}}(y)$ denotes the number of vertices that are not adjacent to y in $\mathcal{G}$. For any $\mathcal{S} \subseteq \mathcal{Y}$,

$$g(\mathcal{S} \mid x) = \sum_{y \in \mathcal{S}} f_y(x) \frac{C_{yy}}{\sum_{y' \in \mathcal{S}} C_{y'y}} = \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{S}} \frac{1}{\widehat{N}_{\mathcal{S}}(y)},$$

and $\widehat{N}_{\mathcal{S}}(y)$ denotes the number of vertices that are not adjacent to y in the subgraph induced[15] by $\mathcal{S}$. The transformation described above is dominated by the size of $\boldsymbol{C}$, which is $O(|\mathcal{V}|^2)$, so the reduction is polynomial time.

We note that the above decision problem can be reduced, in polynomial time, to the problem of finding the optimal prediction set. Precisely, given the optimal prediction set $\mathcal{S}^*$, for every $B \leq g(\mathcal{S}^* \mid x)$ we return YES, otherwise NO. Thus, establishing the NP-hardness for the decision problem suffices.

($\Longleftarrow$) Here, we show that, if the (decision-variant of the) optimal prediction sets problem is a YES instance then the k-clique problem is also a YES instance. Since the optimal prediction sets problem is a YES instance, we have a subset $\mathcal{S} \subseteq \mathcal{Y}$ such that $g(\mathcal{S} \mid x) \geq \frac{k}{|\mathcal{Y}|}$. We anticipate two possibilities, either $\max_{y \in \mathcal{S}} \widehat{N}_{\mathcal{S}}(y) = 1$ or $\max_{y \in \mathcal{S}} \widehat{N}_{\mathcal{S}}(y) > 1$, in both cases, we will show that there exists a clique of size at least k in $\mathcal{G}$.

If $\max_{y \in \mathcal{S}} \widehat{N}_{\mathcal{S}}(y) = 1$ then every pair of vertices $y, y' \in \mathcal{S}$ are adjacent. Since $g(\mathcal{S} \mid x) \geq \frac{k}{|\mathcal{Y}|}$, the size of $\mathcal{S}$ must be at least k otherwise $g(\mathcal{S} \mid x) < \frac{k}{|\mathcal{Y}|}$. So $\mathcal{S}$ induces a clique of size at least k in $\mathcal{G}$.

If $\max_{y \in \mathcal{S}} \widehat{N}_{\mathcal{S}}(y) > 1$, we (iteratively) remove $y' = \arg\max_{y \in \mathcal{S}} \widehat{N}_{\mathcal{S}}(y)$ that has least number of neighbours in $\mathcal{G}_{\mathcal{S}}$, that is the subgraph induced by vertices in $\mathcal{S}$. In Lemma 1, we show that, by removing $y' = \arg\max_{y' \in \mathcal{S}} \widehat{N}_{\mathcal{S}}(y)$ from $\mathcal{S}$, it holds that $g(\mathcal{S} \setminus \{y'\} \mid x) \geq g(\mathcal{S} \mid x)$. Further, the assertion

[14] Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and an integer $k \leq |\mathcal{V}|$, the k-clique problem seeks to decide whether there exists $\mathcal{S} \subseteq V$ with size $|\mathcal{S}| = k$ such that, for every distinct pair $u, v \in S$, there exists $(u, v) \in \mathcal{E}$. We assume that the graph $\mathcal{G}$ is simple, do not contain self loops and do not contain multiple edges between same pair of vertices.

[15] A graph $\mathcal{G}_{\mathcal{S}} = (\mathcal{S}, \mathcal{E}_{\mathcal{S}})$ is a vertex-induced subgraph in graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, if $\mathcal{S} \subseteq \mathcal{V}$ and for every $u, v \in \mathcal{S}$, $(u, v) \in \mathcal{E}_{\mathcal{S}}$ if and only if $(u, v) \in \mathcal{E}$.

https://doi.org/10.52202/079017-0986

$g(\mathcal{S} \setminus \{y'\} \,|\, x) \geq g(\mathcal{S} \,|\, x)$ remains true for each iteration as we remove $y' = \arg\max_{y \in \mathcal{S}} \widehat{N}_{\mathcal{S}}(y)$ iteratively, to obtain $\mathcal{S}' \subseteq \mathcal{S}$ such that $\max_{y \in \mathcal{S}'} \widehat{N}_{\mathcal{S}'}(y) = 1$ and it holds that $g(\mathcal{S}' \,|\, x) \geq \frac{k}{|\mathcal{Y}|}$. The size of $\mathcal{S}'$ is at least k, as any smaller subset results in $g(\mathcal{S}' \,|\, x) < \frac{k}{|\mathcal{Y}|}$. Since $|\mathcal{S}'| \geq k$ and every $y, y' \in \mathcal{S}'$ are adjacent, $\mathcal{S}'$ induces a clique of size at least k.

($\Longrightarrow$) If k-clique is a YES instance and $\mathcal{S} \subseteq \mathcal{V}$ be a clique of size k, then $g(\mathcal{S} \,|\, x) = \frac{k}{|\mathcal{V}|} = \frac{k}{|\mathcal{Y}|}$. So the optimal predictions set problem is a YES instance. This concludes the proof. ∎

Lemma 1 *For any subset $\mathcal{S} \subseteq \mathcal{Y}$ with $\max_{y \in \mathcal{S}} \widehat{N}(y) > 1$ and $y' = \arg\max_{y \in \mathcal{S}} \widehat{N}_{\mathcal{S}}(y)$, it holds that $g(\mathcal{S} \setminus \{y'\} \,|\, x) \geq g(\mathcal{S} \,|\, x)$.*

Proof

$$
\begin{aligned}
g(\mathcal{S} \setminus \{y'\} \,|\, x) - g(\mathcal{S} \,|\, x) &= \frac{1}{|\mathcal{Y}|} \left(\sum_{y \in \widehat{A}_{\mathcal{S} \setminus \{y'\}}(y')} \left(\frac{1}{\widehat{N}_{\mathcal{S}}(y) - 1} - \frac{1}{\widehat{N}_{\mathcal{S}}(y)} \right) - \frac{1}{\widehat{N}_{\mathcal{S}}(y')} \right) \\
&= \frac{1}{|\mathcal{Y}|} \left(\sum_{y \in \widehat{A}_{\mathcal{S} \setminus \{y'\}}(y')} \left(\frac{1}{(\widehat{N}_{\mathcal{S}}(y) - 1)(\widehat{N}_{\mathcal{S}}(y))} \right) - \frac{1}{\widehat{N}_{\mathcal{S}}(y')} \right) \\
&\overset{(i)}{\geq} \frac{1}{|\mathcal{Y}|} \left(\sum_{y \in \widehat{A}_{\mathcal{S} \setminus \{y'\}}(y')} \left(\frac{1}{(\widehat{N}_{\mathcal{S}}(y') - 1)(\widehat{N}_{\mathcal{S}}(y'))} \right) - \frac{1}{\widehat{N}_{\mathcal{S}}(y')} \right) \\
&\overset{(ii)}{\geq} 0
\end{aligned}
$$

Note that (i) and (ii) are valid because $y' = \arg\max_{y \in \mathcal{S}} \widehat{N}_{\mathcal{S}}(y)$. ∎

A.3 Proof of Theorem 2

To establish the hardness of approximation, in Lemma 2 we will show that, given a polynomial-time α-approximation algorithm for the problem of finding the optimal prediction set, we can obtain a polynomial-time α-approximation algorithm for the problem of finding a maximum clique [16] in a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. It is known that, assuming $\mathsf{P} \neq \mathsf{NP}$, for every $\epsilon > 0$, the latter problem is NP-hard to approximate to a factor $|\mathcal{V}|^{1-\epsilon}$ [51]. So we conclude that the optimal prediction sets problem is NP-hard to approximate to a factor $|\mathcal{Y}|^{1-\epsilon}$. ∎

Lemma 2 *Suppose there exists a polynomial-time α-approximation algorithm for the optimal prediction sets problem with $\alpha \geq 1$, then there exists a polynomial-time α-approximation algorithm for the maximum clique problem.*

Proof Let $\mathcal{S}^* \subseteq \mathcal{Y}$ denote the optimal solution for the optimal prediction sets problem, as defined in Eq 1. A subset $\mathcal{S} \subseteq \mathcal{Y}$ is an α-approximation for the optimal prediction sets problem if $g(\mathcal{S} \,|\, x) \cdot \alpha \geq g(\mathcal{S}^* \,|\, x)$. We say an algorithm approximates an instance of the maximum clique problem within a factor α if it can find a clique of size at least $\lfloor \frac{k^*}{\alpha} \rfloor$, when the graph contains a max-clique of size k^*.

The reduction closely resembles the construction outlined in the proof of Theorem 1, with a subtly distinction being the absence of a bound B. For completeness, we will describe the construction again. Given an instance $\langle G = (V, E) \rangle$ of the maximum clique problem, we construct an instance $\langle x, \mathcal{Y}, \boldsymbol{C} \rangle$ of the optimal prediction sets problem as follows: $x \in \mathbb{R}^d$, $\mathcal{Y} = \mathcal{V}$ and, for all $y \in \mathcal{Y}$, we have that $P(Y = y \,|\, X = x) = \frac{1}{|\mathcal{V}|}$ and

$$
P_{\mathcal{S}}(\hat{Y} = \hat{y} \,|\, X = x, Y = y) = \frac{C_{\hat{y}y}}{\sum_{y' \in \mathcal{S}} C_{y'y}} \text{ where } C_{y'y} = \begin{cases} 0 & \text{if } (y', y) \in \mathcal{E} \\ 1/\widehat{N}_{\mathcal{V}}(y) & \text{otherwise,} \end{cases}
$$

[16] Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, the maximum clique problem seeks to find the largest $\mathcal{S} \subseteq \mathcal{V}$ such that, for every $u, v \in \mathcal{S}$, there exists $(u, v) \in \mathcal{E}$.

and $\widehat{N}_{\mathcal{V}}(y)$ denotes the number of vertices that are not adjacent to y in $\mathcal{G}$. For any $\mathcal{S} \subseteq \mathcal{Y}$,

$$g(\mathcal{S} \,|\, x) = \sum_{y \in \mathcal{S}} f_y(x) \frac{C_{yy}}{\sum_{y' \in \mathcal{S}} C_{y'y}} = \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{S}} \frac{1}{\widehat{N}_{\mathcal{S}}(y)},$$

and $\widehat{N}_{\mathcal{S}}(y)$ denotes the number of vertices that are not adjacent to y in the subgraph induced by $\mathcal{S}$.

Let $\mathcal{S}^* \subseteq \mathcal{V}$ be a maximum clique of size k^* in $\mathcal{G}$. For all $y \in \mathcal{S}^*$, $\widehat{N}_{\mathcal{S}^*}(y) = 1$ and $g(\mathcal{S}^* \,|\, x) = \frac{k^*}{|\mathcal{Y}|}$. In Lemma 3 we show that, for any $S \subseteq \mathcal{Y}$ if $\mathcal{S}$ is not a clique of size k^*, then $g(\mathcal{S} \,|\, x) \leq g(\mathcal{S}^* \,|\, x)$. Further, if $\mathcal{S} \subseteq \mathcal{Y}$ is an α-approximation, then from the definition of approximation ratio, it holds that,

$$g(\mathcal{S} \,|\, x) \geq \frac{1}{\alpha} \cdot g(\mathcal{S}^* \,|\, x) \implies g(\mathcal{S} \,|\, x) \geq \lfloor \frac{k^*}{\alpha \,|\mathcal{Y}|} \rfloor.$$

In the remainder of this proof, we show that there exists a clique of size at least $\lfloor \frac{k^*}{\alpha} \rfloor$ in $\mathcal{G}$. To do so, we consider two possibilities: $\max_{y \in \mathcal{S}} \widehat{N}(y) = 1$ and $\max_{y \in \mathcal{S}} \widehat{N}(y) > 1$. In both cases, we will establish the validity of the aforementioned claim.

If $\max_{y \in \mathcal{S}} \widehat{N}_{\mathcal{S}}(y) = 1$ and $g(\mathcal{S} \,|\, x) = \frac{k^*}{\alpha |\mathcal{Y}|}$, then $|\mathcal{S}| = \frac{k^*}{\alpha}$ and every $y, y' \in \mathcal{S}$ is adjacent in the subgraph induced by $\mathcal{S}$. So $\mathcal{S}$ induces a clique of size $\frac{k^*}{\alpha}$.

If $\max_{y \in \mathcal{S}} \widehat{N}_{\mathcal{S}}(y) > 1$, we (iteratively) remove $y' = \arg\max_{y \in \mathcal{S}} \widehat{N}_{\mathcal{S}}(y)$. In Lemma 1, we show that, by removing $y' = \arg\max_{y' \in \mathcal{S}} \widehat{N}_{\mathcal{S}}(y)$ from $\mathcal{S}$, it holds that $g(\mathcal{S} \setminus \{y'\} \,|\, x) \geq g(\mathcal{S} \,|\, x)$. Further, the assertion $g(\mathcal{S} \setminus \{y'\} \,|\, x) \geq g(\mathcal{S} \,|\, x)$ remains true for each iteration as we remove $y' = \arg\max_{y \in \mathcal{S}} \widehat{N}_{\mathcal{S}}(y)$ iteratively, to obtain $\mathcal{S}' \subseteq \mathcal{S}$ such that $\max_{y \in \mathcal{S}'} \widehat{N}_{\mathcal{S}'}(y) = 1$ and it holds that $g(\mathcal{S}' \,|\, x) \geq \lfloor \frac{k^*}{\alpha |\mathcal{Y}|} \rfloor$. The size of $\mathcal{S}'$ is at least $\lfloor \frac{k^*}{\alpha} \rfloor$, as any smaller subset results in $g(\mathcal{S}' \,|\, x) < \frac{k}{\alpha |\mathcal{Y}|}$. Since $|\mathcal{S}'| \geq \lfloor \frac{k^*}{\alpha} \rfloor$ and every $y, y' \in \mathcal{S}'$ are adjacent. So, we conclude that $\mathcal{S}'$ induces a clique of size at least $\lfloor \frac{k^*}{\alpha} \rfloor$. ■

Lemma 3 *For any $\mathcal{S} \subseteq \mathcal{V}$, if the vertices in $\mathcal{S}$ do not induce a maximum clique in $\mathcal{G}$, then $g(\mathcal{S} \,|\, x) \leq g(\mathcal{S}^* \,|\, x)$, where $\mathcal{S}^* \subseteq \mathcal{V}$ induces a maximum clique in $\mathcal{G}$.*

Proof Let $|\mathcal{S}| = k$ and $|\mathcal{S}^*| = k^*$. Based on the values k and k^* can take, we have three possibilities: (i) $k < k^*$, (ii) $k = k^*$, and (iii) $k > k^*$. In each case, we show that the aforementioned claim holds.

Case (i): If $k < k^*$, then $g(\mathcal{S} \,|\, x) = \frac{k-1}{|\mathcal{Y}|} < g(\mathcal{S}^* \,|\, x)$.

Case (ii): If $k = k^*$, then $g(\mathcal{S} \,|\, x) = \frac{k^*}{|\mathcal{Y}|} = g(\mathcal{S}^* \,|\, x)$.

Case (iii): If $k > k^*$ and $\mathcal{S}$ do not induce a clique, then there exists at least one pair $y, y' \in \mathcal{S}$ that are not adjacent, and the value of $g(\mathcal{S} \,|\, x) \leq \frac{k-1}{|\mathcal{Y}|}$. Without loss of generality, assume that $g(\mathcal{S} \,|\, x) = \frac{k-1}{|\mathcal{Y}|}$. Note that, if there are more vertex pairs that are not adjacent, then the inequality is strict, *i.e.*, $g(\mathcal{S} \,|\, x) < \frac{k-1}{|\mathcal{Y}|}$. Let $\mathcal{S} \setminus \{y\}$ be a clique of size $k-1$. In order for $g(x, \mathcal{S}) > g(x, \mathcal{S}^*)$ to hold, $\mathcal{S}$ must contain a clique of size greater than $k^* + 1$, which contradicts our premise that the maximum clique in $\mathcal{G}$ has size k^*. More precisely,

$$g(\mathcal{S} \,|\, x) > g(\mathcal{S}^* \,|\, x) \implies \frac{k-1}{|\mathcal{Y}|} > \frac{k^*}{|\mathcal{Y}|} \implies k > k^* + 1.$$

This concludes the proof. ■

https://doi.org/10.52202/079017-0986

B Running time analysis of the greedy algorithm

In this section, we present the complexity analysis for the greedy algorithm in Algorithm 1. Adding each element requires computing the marginal gain $\Delta = \hat{g}(\mathcal{S} \cup \{y\} \mid x) - \hat{g}(\mathcal{S} \mid x)$ at most k times. This computation in Line 8 can be efficiently performed in $O(k)$ time, where k is the size of $\mathcal{S}_k$. Specifically, it can be rewritten as:

$$\hat{g}(\mathcal{S}_k \cup \{y\} \mid x) - \hat{g}(\mathcal{S}_k \mid x) = \sum_{\hat{y} \in \mathcal{S}_k \cup \{y\}} f_{\hat{y}}(x) \frac{C_{\hat{y}\hat{y}}}{\sum_{y' \in \mathcal{S}_k \cup \{y\}} C_{y'\hat{y}}} - \sum_{\hat{y} \in \mathcal{S}_k} f_{\hat{y}}(x) \frac{C_{\hat{y}\hat{y}}}{\sum_{y' \in \mathcal{S}_k} C_{y'\hat{y}}}.$$

The term $\sum_{y' \in \mathcal{S}_k \cup \{y\}} C_{y'\hat{y}}$ in the denominator can be computed as $\sum_{y' \in \mathcal{S}_k} C_{y'\hat{y}} + C_{y\hat{y}}$ by storing the value of $\sum_{y' \in \mathcal{S}_k} C_{y'\hat{y}}$ at the end of each iteration after Line 11. The loops in Line 5 and Line 7 each iterate over k, resulting in $O(k^3)$ iterations for each value of $k \in \{1, \ldots, L\}$. So, the overall running time of our algorithm is $O(L^4)$.

C Implementation details for the experiments with synthetic and real data

We report the implementation details and computational resources used to run the experiments in Section 5 and Section 6. The code infrastructure was written using Python 3.8 and the standard set of scientific opensource libraries (e.g., `numpy`, `pandas`, `scikit-learn`, etc.). The full set of requirements can be found in the released code. We run the experiment on a Linux machine equipped with an Intel® Xeon(R) Gold 6252N CPU, with 96 cores and 1024 GB of RAM. Practically, our experiments and the algorithms require little resources to be run. We parallelized the execution using `OpenMPI` and 50 physical cores and 20 GB of RAM.

Experiments with synthetic data. We employ the `make_classification` utility function of `scikit-learn` to generate the various prediction task. It is a convenient method to generate L-class classification tasks by varying several parameters such as the task difficulty, the number of labels and the number of informative features. It generates n clusters of points positioned on the vertices of a d-dimensional hypercube by adding interdependencies and noise to the features. In Section 5, we set the number of features to 20, the number of redundant features to 0 and the number of informative features to $d = 4$ (for a $L = 10$ label classification task). We assign a balanced proportion of samples for each class. We control the task difficulty by choosing the `class_sep` parameter, which represents the length of the sides of the hypercubes, thus indicating how far apart are the various classes. A smaller `class_sep` implies a more difficult classification task. We vary the `class_sep` parameter to ensure the classifier and the humans span different ranges of accuracies. Please refer to the original documentation for more information.[17]

Experiments with real data. We use the data provided by Steyvers et al. [7] to evaluate our algorithm against the best conformal predictors. The dataset is composed of 1200 images from the ImageNet-16H classification task. The authors provide also a noisy version of these images by applying a different phase noise ω. Noisier images imply a more difficult classification task. For each image and each phase noise, they provide also the softmax scores of several pre-trained classifiers for different levels of fine-tuning: baseline (no fine-tuning), between 0 and 1 epochs, 1 epochs and 10 epochs. We use the classifier scores (VGG-19 fine-tuned for 10 epochs) and the human classification performance alone for all the 1200 images for different levels of phase noise $\omega = \{80, 95, 110, 125\}$. We chose the VGG-19 classifier because it is the one achieving consistently higher accuracy than the expert alone in the classification task, for all phase noise levels. The data are freely available online.[18] In our experiment, we used Steyvers et al. *normalized* softmax scores which are obtained by dropping from VGG-19 all the irrelevant classes and by renormalizing.

[17] `https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_classification.html#sklearn.datasets.make_classification`

[18] `https://osf.io/2ntrf/wiki/home/`

D Empirical average test accuracy on additional classification tasks

In this section, we keep the same experimental setting as described in Section 5 but we vary the number of labels $L \in \{25, 50\}$ of the classification task. Table 3 summarizes the results.

Table 3: Empirical average test accuracy for $L \in \{25, 50\}$.

Noise	Method	$P(Y'=Y)=0.3$	$P(Y'=Y)=0.5$	$P(Y'=Y)=0.7$	$P(Y'=Y)=0.9$
0.3	NAIVE	0.399 ±0.012	0.614 ±0.009	0.892 ±0.011	0.947 ±0.005
	APS	0.395 ±0.011	0.601 ±0.008	0.873 ±0.009	0.938 ±0.008
	RAPS	0.395 ±0.011	0.601 ±0.008	0.874 ±0.009	0.939 ±0.008
	SAPS	0.396 ±0.010	0.601 ±0.008	0.876 ±0.013	0.943 ±0.007
	GREEDY	**0.431 ±0.010**	**0.649 ±0.011**	**0.902 ±0.011**	**0.956 ±0.005**
	NONE	0.301 ±0.012	0.490 ±0.015	0.805 ±0.012	0.904 ±0.010
0.5	NAIVE	0.383 ±0.011	0.595 ±0.009	0.879 ±0.011	0.941 ±0.005
	APS	0.380 ±0.010	0.585 ±0.009	0.862 ±0.011	0.932 ±0.008
	RAPS	0.380 ±0.010	0.585 ±0.009	0.863 ±0.011	0.933 ±0.008
	SAPS	0.380 ±0.009	0.585 ±0.011	0.863 ±0.014	0.937 ±0.007
	GREEDY	**0.417 ±0.013**	**0.634 ±0.012**	**0.894 ±0.010**	**0.952 ±0.008**
	NONE	0.278 ±0.011	0.455 ±0.014	0.767 ±0.012	0.879 ±0.010
0.7	NAIVE	0.365 ±0.012	0.566 ±0.009	0.848 ±0.010	0.923 ±0.007
	APS	0.363 ±0.009	0.560 ±0.010	0.840 ±0.012	0.918 ±0.008
	RAPS	0.363 ±0.009	0.560 ±0.011	0.841 ±0.011	0.920 ±0.007
	SAPS	0.362 ±0.009	0.558 ±0.015	0.840 ±0.014	0.920 ±0.008
	GREEDY	**0.408 ±0.013**	**0.621 ±0.014**	**0.885 ±0.011**	**0.944 ±0.008**
	NONE	0.244 ±0.012	0.391 ±0.014	0.661 ±0.010	0.772 ±0.009
1.0	NAIVE	0.343 ±0.012	0.530 ±0.017	0.819 ±0.012	0.906 ±0.009
	APS	0.346 ±0.011	0.529 ±0.016	0.822 ±0.013	0.906 ±0.009
	RAPS	0.346 ±0.011	0.529 ±0.016	0.823 ±0.013	0.907 ±0.009
	SAPS	0.344 ±0.010	0.528 ±0.017	0.822 ±0.014	0.907 ±0.009
	GREEDY	**0.408 ±0.012**	**0.618 ±0.014**	**0.888 ±0.012**	**0.949 ±0.006**
	NONE	0.197 ±0.006	0.290 ±0.008	0.432 ±0.006	0.475 ±0.010

$L = 25$

Noise	Method	$P(Y'=Y)=0.3$	$P(Y'=Y)=0.5$	$P(Y'=Y)=0.7$	$P(Y'=Y)=0.9$
0.3	NAIVE	0.435 ±0.014	0.656 ±0.016	0.813 ±0.015	0.939 ±0.007
	APS	0.413 ±0.018	0.626 ±0.016	0.792 ±0.013	0.932 ±0.009
	RAPS	0.412 ±0.017	0.625 ±0.016	0.791 ±0.013	0.932 ±0.009
	SAPS	0.425 ±0.020	0.635 ±0.034	0.802 ±0.014	0.934 ±0.009
	GREEDY	**0.477 ±0.016**	**0.685 ±0.010**	**0.832 ±0.011**	**0.956 ±0.007**
	NONE	0.303 ±0.012	0.508 ±0.006	0.699 ±0.011	0.907 ±0.009
0.5	NAIVE	0.417 ±0.015	0.640 ±0.016	0.803 ±0.013	0.937 ±0.008
	APS	0.399 ±0.018	0.613 ±0.016	0.782 ±0.014	0.927 ±0.010
	RAPS	0.398 ±0.018	0.612 ±0.016	0.781 ±0.014	0.927 ±0.010
	SAPS	0.407 ±0.021	0.620 ±0.033	0.793 ±0.014	0.931 ±0.009
	GREEDY	**0.463 ±0.024**	**0.674 ±0.012**	**0.824 ±0.012**	**0.953 ±0.007**
	NONE	0.281 ±0.011	0.477 ±0.009	0.669 ±0.012	0.890 ±0.005
0.7	NAIVE	0.396 ±0.005	0.606 ±0.018	0.770 ±0.011	0.925 ±0.007
	APS	0.380 ±0.010	0.587 ±0.016	0.754 ±0.015	0.915 ±0.012
	RAPS	0.380 ±0.010	0.587 ±0.016	0.754 ±0.015	0.916 ±0.012
	SAPS	0.387 ±0.008	0.590 ±0.029	0.764 ±0.013	0.920 ±0.009
	GREEDY	**0.450 ±0.017**	**0.657 ±0.015**	**0.811 ±0.015**	**0.944 ±0.005**
	NONE	0.246 ±0.009	0.409 ±0.011	0.574 ±0.011	0.798 ±0.010
1.0	NAIVE	0.376 ±0.012	0.569 ±0.013	0.727 ±0.014	0.907 ±0.011
	APS	0.367 ±0.014	0.560 ±0.016	0.722 ±0.017	0.905 ±0.012
	RAPS	0.367 ±0.014	0.560 ±0.015	0.722 ±0.017	0.906 ±0.012
	SAPS	0.367 ±0.017	0.556 ±0.028	0.725 ±0.020	0.907 ±0.012
	GREEDY	**0.463 ±0.022**	**0.665 ±0.012**	**0.818 ±0.017**	**0.946 ±0.007**
	NONE	0.193 ±0.011	0.296 ±0.013	0.380 ±0.012	0.463 ±0.007

$L = 50$

https://doi.org/10.52202/079017-0986

E Empirical average test coverage of prediction sets across classification tasks

In this section, we report the empirical average test coverage across synthetic and real classification tasks achieved by the prediction sets constructed using conformal prediction and our greedy algorithms. Table 6 and Table 7 summarizes the results, which show that the prediction sets constructed using the best conformal predictors and our greedy algorithm achieve comparable empirical coverage. Moreover, in those few settings in which the prediction sets constructed using conformal prediction achieve higher coverage (*e.g.*, $P(Y' = Y) = 0.3$ and $\gamma = 0.3$), the average test accuracy achieved by the simulated human experts using the prediction sets constructed using conformal prediction is lower (cf. Table 1), which underlines how coverage alone may be a bad proxy for estimating the average accuracy achieved by human experts using prediction sets.

Table 6: Empirical average test coverage achieved by the prediction sets constructed using conformal prediction (NAIVE, APS, RAPS, SAPS) and using the greedy algorithm (GREEDY) on four synthetic classification tasks with four different (simulated) human experts, each with a different noise value γ. For each classification task, the classifier f used by conformal prediction and the greedy algorithm achieves a different average accuracy $P(Y' = Y)$. For each (simulated) human expert, the best value of α for each conformal predictor is different and thus the reported coverage is different. The number of labels is $L = 10$ and, the size of the calibration set is $m = 1000$. Each cell shows the average and standard deviation over 10 runs. We denote the best results for each task in bold.

γ	METHOD	$\mathbb{P}[Y' = Y] = 0.3$	$\mathbb{P}[Y' = Y] = 0.5$	$\mathbb{P}[Y' = Y] = 0.7$	$\mathbb{P}[Y' = Y] = 0.9$
0.3	NAIVE	**0.637 ±0.066**	0.802 ±0.058	0.908 ±0.020	0.973 ±0.007
	APS	0.603 ±0.083	**0.804 ±0.045**	0.900 ±0.026	0.967 ±0.006
	RAPS	0.592 ±0.087	0.792 ±0.032	0.911 ±0.015	0.965 ±0.012
	SAPS	0.580 ±0.067	0.794 ±0.041	0.890 ±0.013	0.965 ±0.010
	GREEDY	0.502 ±0.024	0.764 ±0.019	**0.920 ±0.012**	**0.976 ±0.004**
0.5	NAIVE	**0.583 ±0.104**	**0.770 ±0.044**	0.897 ±0.021	0.968 ±0.009
	APS	0.557 ±0.100	0.741 ±0.036	0.879 ±0.016	0.961 ±0.007
	RAPS	0.534 ±0.096	0.761 ±0.040	0.844 ±0.043	0.950 ±0.012
	SAPS	0.557 ±0.081	0.768 ±0.042	0.876 ±0.014	0.961 ±0.009
	GREEDY	0.489 ±0.027	0.732 ±0.015	**0.902 ±0.013**	**0.970 ±0.003**
0.7	NAIVE	**0.535 ±0.104**	0.676 ±0.047	0.823 ±0.044	0.938 ±0.013
	APS	0.519 ±0.115	0.661 ±0.036	0.853 ±0.015	0.941 ±0.013
	RAPS	0.506 ±0.071	0.644 ±0.056	0.813 ±0.019	0.938 ±0.010
	SAPS	0.492 ±0.079	**0.721 ±0.051**	0.858 ±0.031	0.942 ±0.011
	GREEDY	0.473 ±0.017	0.696 ±0.014	**0.861 ±0.014**	**0.958 ±0.005**
1.0	NAIVE	**0.499 ±0.075**	0.608 ±0.034	0.750 ±0.025	0.905 ±0.014
	APS	0.453 ±0.062	0.631 ±0.038	0.806 ±0.026	0.912 ±0.013
	RAPS	0.466 ±0.071	0.611 ±0.024	0.787 ±0.026	0.920 ±0.013
	SAPS	0.484 ±0.070	0.609 ±0.063	0.764 ±0.022	0.907 ±0.006
	GREEDY	0.457 ±0.024	**0.664 ±0.015**	**0.839 ±0.014**	**0.951 ±0.006**

Table 7: Empirical coverage achieved by the prediction sets constructed using conformal prediction (NAIVE, APS, RAPS, SAPS) and using the greedy algorithm (GREEDY) on the ImageNet-16H dataset. Each cell shows the average and standard deviation over 10 runs. We denote the best results for each noise level in bold.

METHOD	$\omega = 80$	$\omega = 95$	$\omega = 110$	$\omega = 125$
NAIVE	**0.977 ±0.005**	**0.972 ±0.009**	**0.962 ±0.010**	0.923 ±0.018
APS	0.970 ±0.009	0.962 ±0.007	0.943 ±0.015	0.887 ±0.009
RAPS	0.967 ±0.009	0.956 ±0.007	0.944 ±0.013	0.876 ±0.015
SAPS	0.967 ±0.009	0.966 ±0.009	0.946 ±0.008	0.921 ±0.011
GREEDY	0.975 ±0.008	**0.972 ±0.010**	**0.962 ±0.008**	**0.931 ±0.007**

F Empirical conditional probability that a prediction set includes $\{y, \bar{y}\}$ given a ground-truth label $Y = y$ for additional classification tasks

Given a ground truth-label $Y = y$, let $\bar{y} = \text{argmax}_{y' \neq y} C_{y'y}$ be the label that is most frequently mistaken with y. Then, we estimate the empirical conditional probability that a prediction set includes $\{y, \bar{y}\}$ given $Y = y$ with the greedy algorithm and conformal prediction. Figure 4 summarizes the results for different γ and $P(Y' = Y)$ values of several classification tasks where the classifier f and the simulated human expert achieve different average accuracy on their own. The results show that, as the average accuracy achieved by the expert on their own worsens (γ increases), the empirical probability that a prediction set constructed by the greedy algorithm includes $\{y, \bar{y}\}$ decreases.

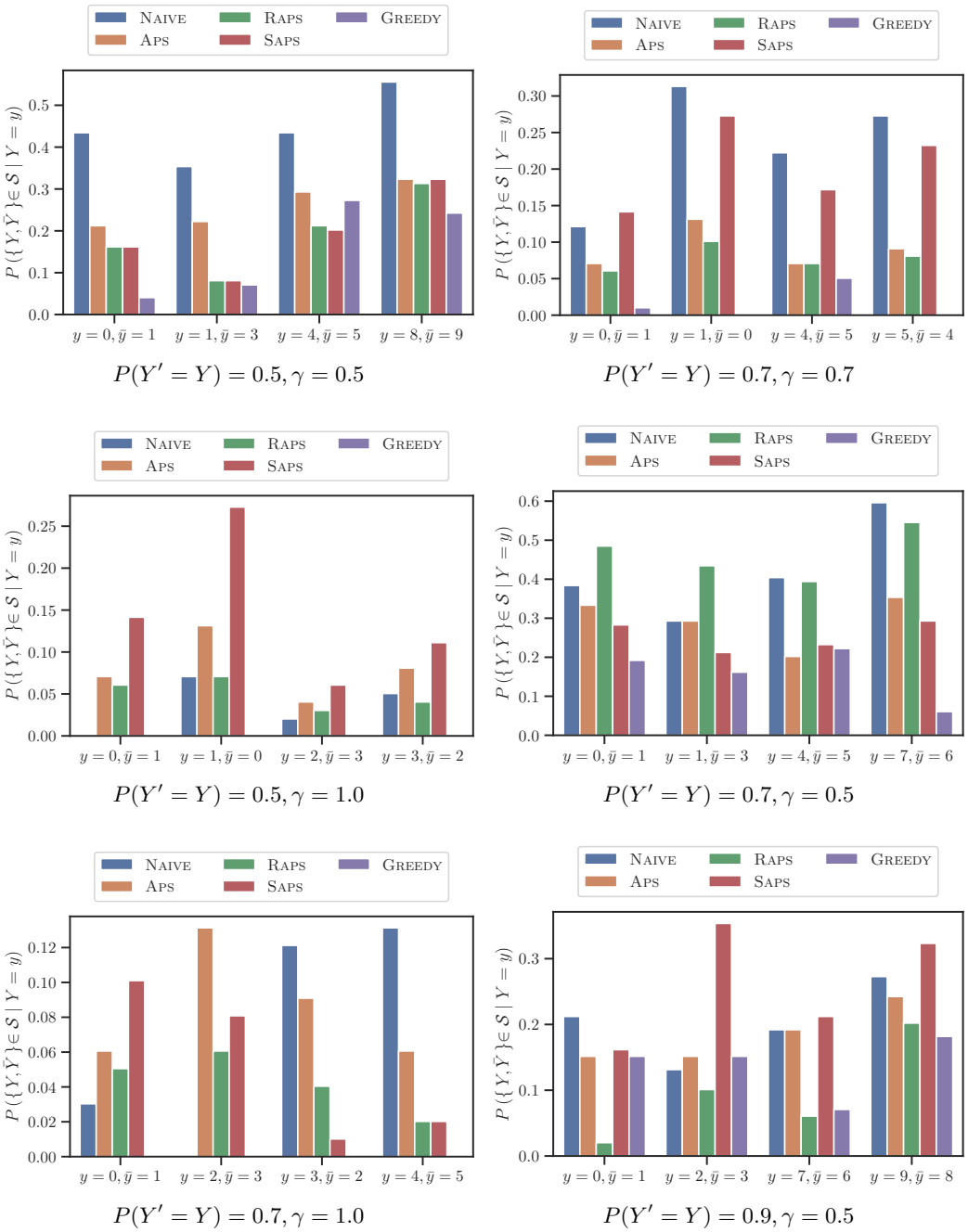

Figure 4: Empirical conditional probability that a prediction set includes $\{y, \bar{y}\}$ given $Y = y$ with conformal prediction (NAIVE, APS, RAPS and SAPS) and our greedy algorithm (GREEDY).

https://doi.org/10.52202/079017-0986

G Evaluation of the Mixture of Multinomial Logit Models (MNLs)

For the group of images with $\omega = 110$ from the ImageNet16H dataset, Straitouri et al. [19] have gathered predictions made by real human experts using prediction sets constructed by all possible conformal predictors with the first non-conformity score we have considered in our experiments (NAIVE), given a choice of calibration set. Here, we use these predictions to evaluate the goodness of fit of the mixture of MNLs used in our experiments.

Figure 5 shows that the average accuracy achieved by a simulated expert following the mixture of MNLs and by real human experts using the prediction sets constructed with all possible conformal predictors, each with a different α value, using the choice of calibration set by Straitouri et al. [19]. The results show that, while the mixture of MNLs tend to overestimates the average accuracy achieved by the predictions made by real experts, the average accuracy follows the same qualitative trend.

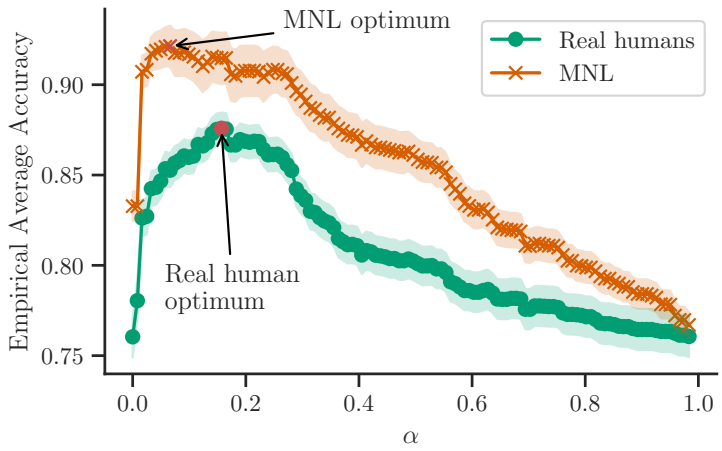

Figure 5: Average accuracy achieved by a simulated expert following the mixture of MNLs and by real human experts using the prediction sets constructed with all possible conformal predictors, each with a different α value, using the choice of calibration set by Straitouri et al. [19]. We highlight in red the highest average accuracy for both the simulated and the real humans.

https://doi.org/10.52202/079017-0986

H Conformal Prediction under different α values

In this section, we estimate the average accuracy achieved by a (simulated) expert using the prediction sets constructed with conformal prediction (NAIVE, APS, RAPS and SAPS) on the ImageNet16H dataset under different α values. Figure 6 summarizes the results.

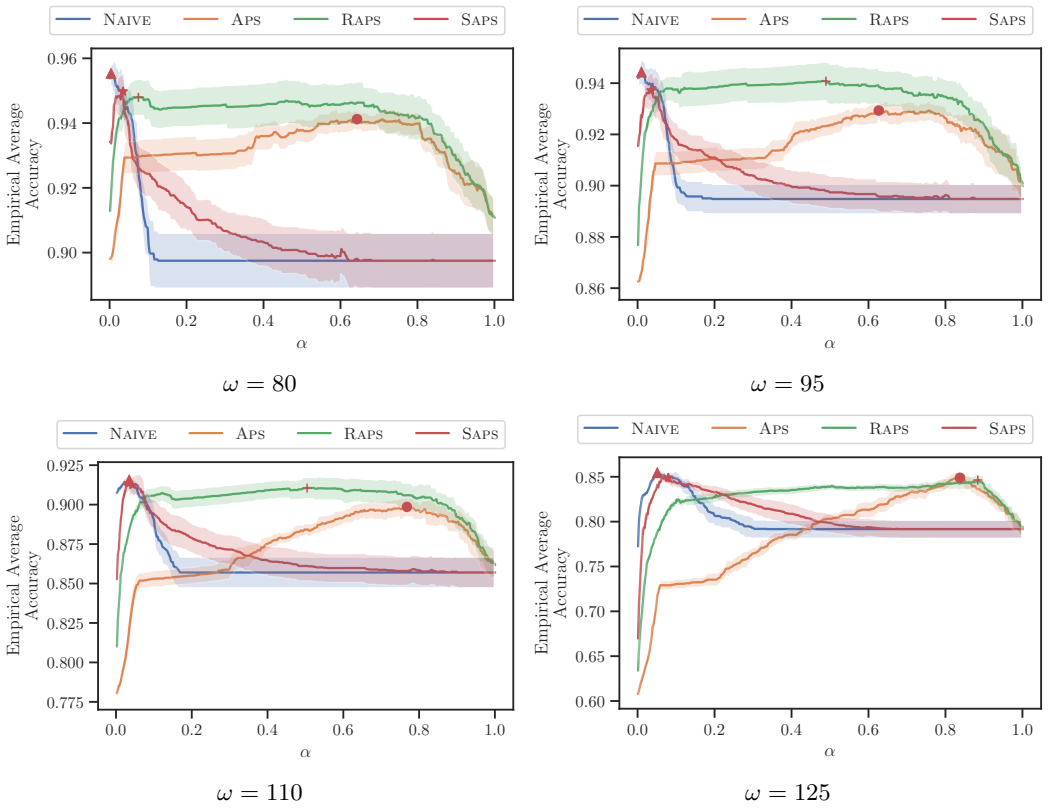

Figure 6: Average accuracy achieved by a (simulated) expert using the prediction sets constructed with conformal prediction (NAIVE, APS, RAPS and SAPS) on the ImageNet16H dataset under different α values. Each panel shows the average and standard error over 10 runs. We highlight with a red marker the highest average accuracy for the simulated humans under each conformal predictor.

https://doi.org/10.52202/079017-0986

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Section 7.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The proofs of all our theoretical results are included in Appendix A

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experimental details are discussed in Section 5 and Section 6. The implementation details are also described in Appendix C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

https://doi.org/10.52202/079017-0986

Answer: [Yes]

Justification: We have released an open-source implementation of our greedy algorithm as well as the code and data used in our experiments at `https://github.com/Networks-Learning/towards-human-ai-complementarity-predictions-sets`.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The training details about data splits, hyperparameters, and pre-trained models are presented in Section 5 and Section 6.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: For each experiment in Section 5 and Section 6, we report the suitable error bars, standard deviations or confidence intervals depending on the table/plot. The same applies to Appendices D, E, G and H.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The information is reported in Appendix G.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes] .

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed potential negative societal impact of our work under "Broader Impact" in Section 7.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

https://doi.org/10.52202/079017-0986

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA] .

    Justification: The paper does not introduce any data or models with a high risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The authors of the existing datasets and pre-trained models used in the paper are duly cited in the main text and the asset license was respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not introduce new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing experiments nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

 https://doi.org/10.52202/079017-0986

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.