Designing Decision Support Systems Using Counterfactual Prediction Sets

Eleni Straitouri [1] **Manuel Gomez Rodriguez** [1]

Abstract

Decision support systems for classification tasks are predominantly designed to predict the value of the ground truth labels. However, since their predictions are not perfect, these systems also need to make human experts understand when and how to use these predictions to update their own predictions. Unfortunately, this has been proven challenging. In this context, it has been recently argued that an alternative type of decision support systems may circumvent this challenge. Rather than providing a single label prediction, these systems provide a set of label prediction values constructed using a conformal predictor, namely a prediction set, and forcefully ask experts to predict a label value from the prediction set. However, the design and evaluation of these systems have so far relied on stylized expert models, questioning their promise. In this paper, we revisit the design of this type of systems from the perspective of online learning and develop a methodology based on the successive elimination algorithm that does not require, nor assumes, an expert model. Our methodology leverages the nested structure of the prediction sets provided by any conformal predictor and a natural counterfactual monotonicity assumption on the experts' predictions over the prediction sets to achieve an exponential improvement in regret in comparison with vanilla successive elimination. We conduct a large-scale human subject study ($n = 2{,}751$) to verify our counterfactual monotonicity assumption and compare our methodology to several competitive baselines. The results suggest that decision support systems that limit experts' level of agency may be practical and may offer greater performance than those allowing experts to always exercise their own agency.

[1]Max Planck Institute for Software Systems, Kaiserslautern, Germany. Correspondence to: Eleni Straitouri <estraitouri@mpi-sws.org>.

AI & HCI Workshop at the 40^{th} International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA. 2023.

1. Introduction

Throughout the years, one of the main focus in the area of machine learning for decision support has been classification tasks. In this setting, the decision support system typically uses a classifier to predict the value of a ground truth label of interest and a human expert uses the predicted value to update their own prediction (Bansal et al., 2019; Lubars & Tan, 2019; Bordt & von Luxburg, 2020). Classifiers have become remarkably accurate in a variety of application domains such as medicine (Jiao et al., 2020), education (Whitehill et al., 2017), or criminal justice (Dressel & Farid, 2018), to name a few. However, their data-driven predictions are not always perfect and, in the presence of aleatoric uncertainty, they will never be (Raghu et al., 2019). As a result, there has been a flurry of work on helping human experts understand when and how to use the predictions provided by these systems to update their own (Papenmeier et al., 2019; Wang & Yin, 2021; Vodrahalli et al., 2022; Liu et al., 2023). Unfortunately, it is yet unclear how to guarantee that, by using these systems, experts never decrease the average accuracy of their own predictions (Yin et al., 2019; Zhang et al., 2020; Suresh et al., 2020; Lai et al., 2021).

Very recently, Straitouri et al. (2023) have argued that an alternative type of decision support systems may provide such a guarantee, by design. Rather than providing a label prediction and letting human experts decide when and how to use the predicted label to update their own prediction, this type of systems provide a set of label predictions, namely a prediction set, and forcefully ask the experts to predict a label value from the prediction set, as shown in Figure 1.[1] Their key argument is that, if the prediction set is constructed using conformal prediction (Vovk et al., 2005; Angelopoulos & Bates, 2021), then one can precisely trade-off the probability that the ground truth label is not in the prediction set, which determines how frequently the systems will mislead human experts[2] and the size of the prediction set, which de-

[1]There are many systems used everyday by experts that, under normal operation, limit experts' level of agency. For example, think of a pilot who is flying a plane. There are automated, adaptive systems that prevent the pilot from taking certain actions based on the monitoring of the environment.

[2]Since these systems do not allow experts to predict a label value if it lies outside the prediction set, if the prediction set does not contain the ground truth label, we know that the expert's prediction will be incorrect.

termines the difficulty of the classification task the experts need to solve using the system. However, Straitouri et al. are only able to find the (near-)optimal conformal predictor that maximizes average accuracy under the assumption that the experts' predictions follow a stylized expert model, which they also use for evaluation. In this work, our goal is to lift this assumption and efficiently find the optimal conformal predictor that maximizes the average accuracy achieved by real experts using such a system.

Our contributions. We start by formally characterizing experts' predictions over prediction sets constructed using a conformal predictor using a structural causal model (SCM) (Pearl, 2009). Building upon this characterization, we identify the following natural counterfactual monotonicity assumption on the experts' predictions. If an expert succeeds (fails) at predicting the ground truth label from a prediction set and the set contains the ground truth label, the expert would have also succeeded (failed) had the prediction set been smaller (larger) but had still contained the ground truth label. Then, we use this counterfactual monotonicity assumption and the nested structure of the prediction sets provided by conformal prediction to design a very efficient algorithm to find the optimal conformal predictor, which we refer to as counterfactual successive elimination. As implied by the name, our algorithm is based on the well-known successive elimination algorithm from the multi-armed bandit literature (Slivkins, 2019), however, in our setting, it achieves an exponential improvement in regret in comparison with vanilla successive elimination.

Finally, we conduct a large-scale user study with 2,751 human subjects who make 194,407 predictions over 19,200 different pairs of natural images and prediction sets. In our study, we experiment both with a *strict* and a *lenient* implementation of our decision support systems. Under the strict implementation, experts can only predict a label value from the prediction set whereas, under the lenient implementation, experts are encouraged to predict a label value from the prediction set but have the possibility to predict other label values. Perhaps surprisingly, our results demonstrate that, under the strict implementation of our system, experts achieve higher accuracy. This suggests that decision support systems that adaptively limit experts' level of agency may offer greater performance than those allowing experts to always exercise their own agency. Further, our results also demonstrate that, for the strict implementation of our system, the experts' predictions seem to satisfy the counterfactual monotonicity assumption and, as a consequence, the conformal predictor found by our counterfactual successive elimination algorithm offers greater performance than that found by the algorithm introduced by Straitouri et al. (2023).

Further related work. Our work builds upon further related work on set-valued predictors, multi-armed bandits, and learning under algorithmic triage.

Conformal predictors are just one among many different types of set-valued predictors (Chzhen et al., 2021), *i.e.*, predictors that, for each sample, output a set of label values. In our work, we opted for conformal predictors over alternatives such us, *e.g.*, reliable or cautious classifiers (Yang et al., 2017; Mortier et al., 2021; Ma & Denoeux, 2021; Nguyen & Hüllermeier, 2021), because of their provable coverage guarantees, which allow us to control how frequently our decision support systems will mislead human experts. Except for two notable exceptions by Straitouri et al. (2023) and Babbar et al. (2022), set-valued predictors have not been specifically designed to serve decision support systems. Within these two exceptions, the work by Straitouri et al. (2023) is more related to ours and has been discussed previously. The work by Babbar et al. (2022) studied the lenient implementation of our decision support systems, under which the experts appear to achieve lower accuracy in our human subject study, as shown in Figure 3. However, they consider a conformal predictor with a given coverage probability, rather than optimizing across conformal predictors.[3]

Within the vast literature of multi-armed bandits (refer to Slivkins (2019) for a recent review), our work is most closely related to causal bandits (Lattimore et al., 2016; Lee & Bareinboim, 2018; de Kroon et al., 2020) and combinatorial multi-armed bandits (Chen et al., 2013). In causal bandits, there is a known causal relationship between arms and rewards, similarly as in our work. However, the focus is on using this causal relationship, rather than counterfactual inference, to explore more efficiently and achieve lower regret. In combinatorial multi-armed bandits, one can pull any subset of arms, namely a super arm, at the same time. Then, the goal is to identify the (near-)optimal super arm. While one could view our problem as an instance of combinatorial multi-armed bandits, our goal is not to identify the optimal super arm but the optimal base arm.

In learning under algorithmic triage, a classifier predicts the ground truth label of a given fraction of the samples and leaves the remaining ones to a human expert, as instructed by a triage policy (Mozannar & Sontag, 2020; De et al., 2020; 2021; Okati et al., 2021). In contrast, in our work, for each sample, a classifier is used to build a set of label predictions and the human expert needs to predict a label value from the set.

[3] In Babbar et al. (2022), they reduce the size of the prediction sets constructed using conformal prediction by deferring some samples to human experts during calibration and testing. However, since such an optimization can be applied both to the strict and the lenient implementation of our system and we do not find any reason why it would change the conclusions of our human subject study, for simplicity, we decided not to apply it.

2. Decision Support Systems Based on Prediction Sets

Given a multiclass classification task where, for each sample, a human expert needs to predict a label $y \in \mathcal{Y} = \{1, \ldots, n\}$, with $y \sim P(Y \mid X)$, from a feature vector $x \in \mathcal{X}$, with $x \sim P(X)$, the decision support system $\mathcal{C} : \mathcal{X} \to 2^{\mathcal{Y}}$ helps the expert by automatically narrowing down the set of potential label values to a subset of them $\mathcal{C}(x) \subseteq \mathcal{Y}$, which we refer to as a prediction set, using a set-valued predictor (Chzhen et al., 2021). Here, for reasons that will become apparent later, we focus on a strict implementation of the system that, for any $x \in \mathcal{X}$, forcefully asks the expert's prediction $\hat{y} \in \mathcal{Y}$, with $\hat{y} \sim P(\hat{Y}_{\mathcal{C}} \mid X, \mathcal{C}(X))$, to belong to the prediction set $\mathcal{C}(x)$. More formally, this is equivalent to assuming that $P(\hat{Y}_{\mathcal{C}} = \hat{y} \mid X = x, \mathcal{C}(x)) = 0$ for all $\hat{y} \notin \mathcal{C}(x)$. Refer to Figure 1 for an illustration of the automated decision support system we consider.

Then, the goal is to find the optimal decision support system $\mathcal{C}^*$ that maximizes the average accuracy of the expert's predictions, *i.e.*, $\mathcal{C}^* = \mathrm{argmax}_{\mathcal{C}}\, \mathbb{E}_{X,Y,\hat{Y}_{\mathcal{C}}}[\mathbb{I}\{\hat{Y}_{\mathcal{C}} = Y\}]$. However, to solve the above maximization problem, we need to first specify the class of set-valued predictors we aim to maximize average accuracy upon. Here, following Straitouri et al. (2023), we opt for conformal predictors (Vovk et al., 2005; Angelopoulos & Bates, 2021) over alternatives since they allow for a precise control of the trade-off between how frequently the expert is misled by the system and the difficulty of the classification task she needs to solve. More specifically, given a user-specified parameter $\alpha \in [0, 1]$, a conformal predictor uses a pre-trained classifier $\hat{f}(x) \in [0, 1]^n$ 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 $\mathcal{C}_\alpha(X)$ as follows:[4]

$$\mathcal{C}_\alpha(X) = \{y \mid s(X, y) \leq \hat{q}_\alpha\}, \tag{1}$$

where $s(x_i, y_i) = 1 - \hat{f}_{y_i}(x_i)$ is called the conformal score[5] and $\hat{q}_\alpha$ is the $\frac{\lceil (m+1)(1-\alpha) \rceil}{m}$ empirical quantile of the conformal scores $s(x_1, y_1), \ldots, s(x_m, y_m)$. Remarkably, by using the above construction, the conformal predictor enjoys probably approximately correct (PAC) coverage guarantees, *i.e.*, given tolerance values $\delta, \epsilon \in (0, 1)$, we can compute the minimum size m of the calibration set $\mathcal{D}_{\text{cal}}$ such that, with probability $1 - \delta$, it holds that (Angelopoulos & Bates, 2021)

$$1 - \alpha - \epsilon \leq \mathbb{P}[Y \in \mathcal{C}_\alpha(X) \mid \mathcal{D}_{\text{cal}}] \leq 1 - \alpha + \epsilon,$$

where $(1 - \alpha)$ is called the (user-specified) coverage probability.[6] Then, we can conclude that, with high probability, for $(1 - \alpha)$ of the samples, $\mathcal{C}_\alpha(x)$ contains the ground truth label and thus cannot mislead the expert—if the expert would succeed at predicting the ground truth label y of a sample with feature vector x on her own, she could still succeed using $\mathcal{C}_\alpha$ because $\mathcal{C}_\alpha(x)$ contains the ground truth label. On the flip side, for α of the samples, we know that, if the expert uses $\mathcal{C}_\alpha$, she will fail at predicting the ground truth label.[7]

Further, we know that the smaller (larger) the parameter α, the larger (smaller) the size of the prediction set $\mathcal{C}_\alpha(x)$, and thus in turn the higher (lower) the difficulty of the classification task the expert needs to solve (Wright & Barbour, 1977; Beach, 1993; Ben-Akiva & Boccara, 1995). The important point here is that, since α is a parameter we choose, we can precisely control the trade-off between how frequently the system misleads the expert and the difficulty of the classification tasks the expert needs to solve. In contrast, note that, if we do not forcefully ask the expert to predict a label value from the prediction set, we would not be able to have this level of control and good performance would depend on the expert developing a good sense on when to predict a label from the prediction set and when to predict a label from outside the set.

3. Prediction Sets Through a Causal Lens

In this section, we start by characterizing how human experts make predictions using a decision support system via a structural causal model (SCM) (Pearl, 2009), which we denote as $\mathcal{M}$. Our SCM $\mathcal{M}$ is defined by the following assignments:

$$\mathcal{C}_{\mathrm{A}}(X) = f_{\mathcal{C}}(X, \mathrm{A}), \quad \hat{Y}_{\mathcal{C}_{\mathrm{A}}} = f_{\hat{Y}}(U, V, \mathcal{C}_{\mathrm{A}}(X)),$$
$$X = f_X(V) \quad \text{and} \quad Y = f_Y(V) \tag{2}$$

where A, U and V are independent exogenous variables characterizing the (user-specified) coverage probability, the expert's individual characteristics, and the data generating process for the feature vectors X and ground truth labels Y, respectively, $f_{\mathcal{C}}$, $f_{\hat{Y}}$, f_X and f_Y are given functions, and let $P^{\mathcal{M}}$ be the distribution entailed by $\mathcal{M}$. Here, the function $f_{\mathcal{C}}$ is a set-function directly defined by the conformal

[4] The assumption that $\hat{f}(x) \in [0, 1]^n$ is without loss of generality. Here, the higher the score $\hat{f}_y(x)$, the more the classifier believes the ground truth label $Y = y$.

[5] In general, the conformal score $s(x, y)$ can be any function of x and y measuring the *similarity* between samples.

[6] Most of the literature on conformal prediction focuses on marginal coverage guarantees rather than PAC coverage guarantees. However, since we will optimize the performance of our system with respect to the parameter α, we cannot afford marginal coverage guarantees, as discussed in Straitouri et al. (2023).

[7] There may be some feature vectors x for which, in principle, a conformal predictor may return an empty prediction set $\mathcal{C}_\alpha(x)$. However, during deployment, one can trivially conclude that they will not contain the ground truth label. Hence, in those cases, $\mathcal{C}_\alpha$ may allow the expert to choose from $\mathcal{Y}$ (or any other subset of labels) without changing the coverage guarantees.

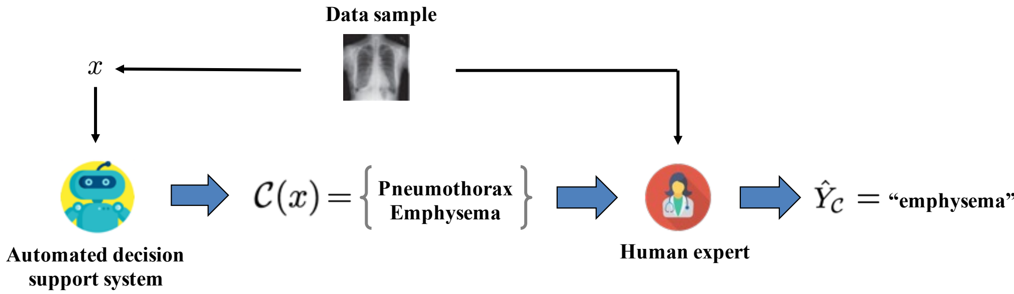


Figure 1. Our automated decision support system $\mathcal{C}$. Given a sample with a feature vector x, the system $\mathcal{C}$ helps the expert by automatically narrowing down the set of potential label values to a subset of them $\mathcal{C}(x) \subseteq \mathcal{Y}$, which we refer to as a prediction set, using a set-valued predictor. The system forcefully asks the expert to predict a label value $\hat{y}_{\mathcal{C}}$ from the prediction set $\mathcal{C}(x)$, *i.e.*, $\hat{y}_{\mathcal{C}} \in \mathcal{C}(x)$.

predictor, *i.e.*,

$$f_{\mathcal{C}}(X = x, A = \alpha) = \mathcal{C}_\alpha(x) = \{y \,|\, s(x, y) \leq \hat{q}_\alpha\},$$

where the calibration set $\mathcal{D}_{\text{cal}}$ is given and thus it does not appear explicitly as an independent variable. Moreover, note that our model allows both for causal and anticausal features (Schölkopf et al., 2012) and, as argued elsewhere (Pearl, 2009), we can always find a distribution for the exogenous variables U and V and functions f_X, f_Y and $f_{\hat{Y}}$ such that the observational distributions $P(X)$, $P(Y|X)$ and $P(\hat{Y}_{\mathcal{C}_\alpha} \,|\, X, \mathcal{C}_\alpha(X))$ of interest, defined in the previous section, are given by distributions entailed by the SCM $\mathcal{M}$, *i.e.*, $P(X = x) = P^{\mathcal{M}}(X = x)$, $P(Y = y \,|\, X = x) = P^{\mathcal{M}}(Y = y \,|\, X = x)$ and

$$\begin{aligned} &P(\hat{Y}_{\mathcal{C}_\alpha} = \hat{y} \,|\, X = x, \mathcal{C}_\alpha(X) = \mathcal{C}_\alpha(x)) \\ &\quad = P^{\mathcal{M}\,;\,\text{do}(\mathcal{C}_{\text{A}}(X) = \mathcal{C}_\alpha(x))}(\hat{Y}_{\mathcal{C}_{\text{A}}} = \hat{y} \,|\, X = x), \end{aligned}$$

where $\text{do}(\mathcal{C}_{\text{A}}(X) = \mathcal{C}_\alpha(x))$ denotes a (hard) intervention in which the first assignment in Eq. 2 is replaced by the value $\mathcal{C}_\alpha(x)$.

Under this view, we can formally reason about the predictions $\hat{Y}_{\mathcal{C}_\alpha}$ made by a human expert under different support systems $\mathcal{C}_\alpha$. Given a sample with feature vector x, we can first conclude that, since the expert's predictions only depend on the (user-specified) coverage probability $(1 - \alpha)$ through the prediction set $\mathcal{C}_\alpha(x)$, it must hold that, for any pair of decision support systems $\mathcal{C}_\alpha$ and $\mathcal{C}_{\alpha'}$ such that $\mathcal{C}_\alpha(x) = \mathcal{C}_{\alpha'}(x)$, if we observe that the expert has predicted $\hat{Y}_{\mathcal{C}_\alpha} = \hat{y}$ using $\mathcal{C}_\alpha$, we can be certain that she would have predicted $\hat{Y}_{\mathcal{C}_{\alpha'}} = \hat{y}$ had she used $\mathcal{C}_{\alpha'}$ while holding "everything else fixed" (Pearl, 2009).

Next, motivated by prior empirical studies in the psychology and marketing literature (Chernev et al., 2015; Schwartz, 2004; Haynes, 2009; Kuksov & Villas-Boas, 2010), which suggest that increasing the number of alternatives in a decision making task increases its difficulty, we hypothesize and later on empirically verify (see Figure 2a) that, for any pair of decision support systems $\mathcal{C}_\alpha$ and $\mathcal{C}_{\alpha'}$ such that $Y \in \mathcal{C}_\alpha(x) \subseteq \mathcal{C}_{\alpha'}(x)$, experts are more likely to succeed at predicting the ground truth label Y under $\mathcal{C}_\alpha$ than under $\mathcal{C}_{\alpha'}$, *i.e.*,

$$\begin{aligned} &P^{\mathcal{M}\,;\,\text{do}(\mathcal{C}_{\text{A}}(X) = \mathcal{C}_\alpha(x))}(\hat{Y}_{\mathcal{C}_{\text{A}}} = Y \,|\, X = x) \\ &\qquad \geq P^{\mathcal{M}\,;\,\text{do}(\mathcal{C}_{\text{A}}(X) = \mathcal{C}_{\alpha'}(x))}(\hat{Y}_{\mathcal{C}_{\text{A}}} = Y \,|\, X = x), \end{aligned} \tag{3}$$

where the probability is over the uncertainty on the expert's individual characteristics and the data generating process. Moreover, we further hypothesize that, if the above assumption holds, the following natural counterfactual monotonicity assumption may also hold:[8]

Assumption 3.1 (Counterfactual monotonicity)**.** The experts' predictions satisfy counterfactual monotonicity if and only if, for any $x \in \mathcal{X}$ and any $\mathcal{C}_\alpha$ and $\mathcal{C}_{\alpha'}$ such that $Y \in \mathcal{C}_\alpha(x) \subseteq \mathcal{C}_{\alpha'}(x)$, it holds that $\mathbb{I}\{f_{\hat{Y}}(u, v, \mathcal{C}_\alpha(x)) = Y\} \geq \mathbb{I}\{f_{\hat{Y}}(u, v, \mathcal{C}_{\alpha'}(x)) = Y\}$ for any $u \sim P^{\mathcal{M}}(U)$ and $v \sim P^{\mathcal{M}}(V \,|\, X = x)$.

The above assumption directly implies that, for any sample with feature vector x and any $\mathcal{C}_\alpha$ and $\mathcal{C}_{\alpha'}$ such that $Y \in \mathcal{C}_\alpha(x) \subseteq \mathcal{C}_{\alpha'}(x)$, if we observe that an expert has succeeded at predicting the ground truth label Y using $\mathcal{C}_{\alpha'}$, she would have also succeeded had she used $\mathcal{C}_\alpha$ and, conversely, if she has failed at predicting Y using $\mathcal{C}_\alpha$, she would have also failed had she used $\mathcal{C}_{\alpha'}$, while holding "everything else fixed". In other words, under the counterfactual monotonicity assumption, the counterfactual dynamics of the expert under certain alternative prediction sets are identifiable and purely deterministic. In what follows, we will leverage this assumption to develop a very efficient online algorithm to find the optimal conformal pre-

[8] Since our counterfactual monotonicity assumption lies within level three in the "ladder of causation" (Pearl, 2009), we cannot validate it using observational nor interventional experiments. However, the good practical performance of our counterfactual successive elimination algorithm (Algorithm 1) suggests that it may hold in practice.

dictor among those using a given calibration set $\mathcal{D}_{\text{cal}}$, *i.e.*, $\alpha^* = \text{argmax}_\alpha\, \mathbb{E}_{X,Y,\hat{Y}_{\mathcal{C}_\alpha}}[\mathbb{I}\{\hat{Y}_{\mathcal{C}_\alpha} = Y\}]$.

4. Finding the Optimal Conformal Predictor Using Counterfactual Prediction Sets

Given a calibration set $\mathcal{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^m$, there exist only m different conformal predictors. This is because the empirical quantile $\hat{q}_\alpha$, which the subsets $\mathcal{C}_\alpha(x_i)$ depend on, can only take m different values. As a result, to find the optimal conformal predictor, we just need to solve the following maximization problem:

$$\alpha^* = \underset{\alpha \in \mathcal{A}}{\text{argmax}}\ \mathbb{E}_{X,Y,\hat{Y}_{\mathcal{C}_\alpha}}\left[\mathbb{I}\{\hat{Y}_{\mathcal{C}_\alpha} = Y\}\right], \qquad (4)$$

where $\mathcal{A} = \{\alpha_i\}_{i \in [m]}$, with $\alpha_i = 1 - i/(m+1)$. However, since we do not know the causal mechanism experts use to make predictions over prediction sets, we need to trade-off exploitation, *i.e.*, maximizing the expected accuracy, and exploration, *i.e.*, learning about the accuracy achieved by the experts under each conformal predictor. To this end, we look at the problem from the perspective of multi-armed bandits (Slivkins, 2019).

In our problem, each arm corresponds to a different parameter value α and, at each round t, a (potentially different) human expert receives a sample with feature vector x_t, picks a label value $\hat{y}_t$ from the prediction set $\mathcal{C}_{\alpha_t}(x_t)$ provided by the conformal predictor with $\alpha_t \in \mathcal{A}$, and obtains a reward $\mathbb{I}\{\hat{y}_t = y_t\} \in \{0, 1\}$. Here, we observe x_t at the beginning of each round and y_t and $\mathbb{I}\{\hat{y}_t = y_t\}$ at the end of each round. Then, the goal is to find a sequence of parameter values $\{\alpha_t\}_{t=1}^T$ with desirable properties in terms of total regret $R(T)$, which is given by:

$$R(T) = T \cdot \mathbb{E}_{X,Y,\hat{Y}_{\mathcal{C}_{\alpha^*}}}\left[\mathbb{I}\{\hat{Y}_{\mathcal{C}_{\alpha^*}} = Y\}\right] - \sum_{t=1}^{T} \mathbb{E}_{X,Y,\hat{Y}_{\mathcal{C}_{\alpha_t}}}\left[\mathbb{I}\{\hat{Y}_{\mathcal{C}_{\alpha_t}} = Y\}\right], \qquad (5)$$

where α^* is the optimal parameter value, as defined in Eq. 4. At this point, one could think of resorting to any of the well-known algorithms from the literature on stochastic multi-armed bandits (Slivkins, 2019), such as `UCB1` or successive elimination, to decide which arm to pull, *i.e.*, which α_t to use, at each round t. These algorithms would achieve an expected regret $\mathbb{E}[R(t)] \leq O(\sqrt{mt \log T})$ for any $t \leq T$, where the expectation is over the randomness in the execution of the algorithms. However, in our problem setting, we can do much better than that—in what follows, we will design an algorithm based on successive elimination that achieves an expected regret $\mathbb{E}[R(t)] \leq O(\sqrt{t \log m \log T})$ for any $t \leq T$.

The successive elimination algorithm keeps a set $\mathcal{A}_{\text{active}}$ of *active* arms α, which initially sets to $\mathcal{A}_{\text{active}} = \mathcal{A}$. Then, it pulls a different arm $\alpha \in \mathcal{A}_{\text{active}}$, without repetition, until it has pulled all arms in $\mathcal{A}_{\text{active}}$. Assume it has pulled all arms at round t. Then, it computes an upper and a lower confidence bound on the average reward associated to each arm α, $\texttt{UCB}_t(\alpha) = \hat{\mu}_t(\alpha) + \epsilon_t(\alpha)$ and $\texttt{LCB}_t(\alpha) = \hat{\mu}_t(\alpha) - \epsilon_t(\alpha)$, where

$$\hat{\mu}_t(\alpha) = \frac{\sum_{t' \leq t} \mathbb{I}\{\hat{y}_{t'} = y_{t'} \wedge \alpha_{t'} = \alpha\}}{\sum_{t' \leq t} \mathbb{I}\{\alpha_{t'} = \alpha\}} \quad \text{and}$$

$$\epsilon_t(\alpha) = \sqrt{\frac{2 \log T}{\sum_{t' \leq t} \mathbb{I}\{\alpha_{t'} = \alpha\}}},$$

and *deactivates* any arm $\alpha \in \mathcal{A}_{\text{active}}$ for which there exists $\alpha' \in \mathcal{A}_{\text{active}}$ such that $\texttt{UCB}_t(\alpha) < \texttt{LCB}_t(\alpha')$. Then, it repeats the same procedure until the maximum number of rounds T is reached or until $|\mathcal{A}_{\text{active}}| = 1$.

The rate at which successive elimination deactivates arms and, in turn, the expected regret, is essentially limited by the fact that, to update $\hat{\mu}_t(\alpha)$ and $\epsilon_t(\alpha)$ for every arm $\alpha \in \mathcal{A}_{\text{active}}$, it needs to pull $O(m)$ arms. However, in our problem setting, there exists an efficient strategy to update $\hat{\mu}_t(\alpha)$ and $\epsilon_t(\alpha)$ pulling just $O(\log m)$ arms.

In the first round, our algorithm pulls the arm whose corresponding parameter value $\tilde{\alpha}$ is the *median*[9] of all values in $\mathcal{A}_{\text{active}}$. We distinguish two cases. First, assume that the expert has failed at predicting the ground truth label y_1 using $\mathcal{C}_{\tilde{\alpha}}$, *i.e.*, $\mathbb{I}\{\hat{y}_1 = y_1\} = 0$. If $y_1 \notin C_{\tilde{\alpha}}(x_1)$, we know that she would have also failed had she used any $\mathcal{C}_{\alpha'}$ such that $\alpha' > \tilde{\alpha}$ since $\mathcal{C}_{\alpha'}(x) \subseteq \mathcal{C}_{\tilde{\alpha}}(x)$. If $y_1 \in C_{\tilde{\alpha}}(x_1)$, we know that she would have also failed had she used any $\mathcal{C}_{\alpha'}$ such that $\alpha' < \tilde{\alpha}$ due to the counterfactual monotonicity assumption. Second, assume that the expert has succeeded at predicting y_1 using $\mathcal{C}_{\tilde{\alpha}}$, *i.e.*, $\mathbb{I}\{\hat{y}_1 = y_1\} = 1$. Then, for any $\alpha' > \tilde{\alpha}$, we know that, if $y_1 \in \mathcal{C}_{\alpha'}(x_1)$, the same expert would have also succeeded had she used $\mathcal{C}_{\alpha'}$ due to the counterfactual monotonicity assumption and, if $y_1 \notin \mathcal{C}_{\alpha'}(x_1)$, the expert would have trivially failed had she used $\mathcal{C}_{\alpha'}$. In both cases, the algorithm observes the reward for one arm and counterfactually infers the reward for one *half*[10] of the arms in $\mathcal{A}_{\text{active}}$. In the next $O(\log m)$ rounds, it repeats the same reasoning, it pulls the arm whose parameter value is the median of the remaining *half* of the arms whose reward has not yet observed or counterfactually inferred, until it has observed or counterfactually inferred the reward of all arms at least once. Then, it computes

$$\hat{\mu}_t(\alpha) = \frac{\sum_{t' \leq t} \gamma_{t'}(\alpha)}{\sum_{t' \leq t} \nu_{t'}(\alpha)} \quad \text{and} \quad \epsilon_t(\alpha) = \sqrt{\frac{2 \log T}{\sum_{t' \leq t} \nu_{t'}(\alpha)}},$$

[9] The $\frac{m}{2}$-th largest value if m is even or the $\frac{m+1}{2}$-th largest value if m is odd.

[10] One *half* is $\frac{m}{2}$ or $\frac{m}{2} - 1$ values if m is even and $\frac{m-1}{2}$ values if m is odd.

Algorithm 1 Counterfactual successive elimination

Input: $\mathcal{A}, T, \mathcal{D}_{\text{opt}}$
$\mathcal{A}_{\text{active}} \leftarrow \mathcal{A}$
$t \leftarrow 0, \gamma \leftarrow 0, \nu \leftarrow 0$
while $t < T \vee |\mathcal{A}_{\text{active}}| > 1$ **do**
  $\mathcal{A}_{\text{unexplored}} \leftarrow \mathcal{A}_{\text{active}}$
  while $\mathcal{A}_{\text{unexplored}} \neq \varnothing \wedge t < T$ **do**
    $\tilde{\alpha} \leftarrow \text{MEDIAN}(\mathcal{A}_{\text{unexplored}})$
    $(x_t, y_t) \sim \mathcal{D}_{\text{opt}}$
    Deploy $\mathcal{C}_{\tilde{\alpha}}(x_t)$ and observe $\mathbb{I}\{\hat{y}_t = y_t\}$ and y_t
    $\mathcal{A}_{\text{unexplored}}, \gamma, \nu \leftarrow \text{UPDT}(\mathcal{A}_{\text{unexplored}}, \gamma, \nu, \tilde{\alpha}, x_t, y_t, \hat{y}_t)$
    {Call Algorithm 2, Appendix B}
    $t \leftarrow t + 1$
  end while
  for $\alpha \in \mathcal{A}_{\text{active}}$ **do**
    $\mu(\alpha) = \gamma(\alpha)/\nu(\alpha)$
    $\epsilon(\alpha) = \sqrt{2 \log T / \nu(\alpha)}$
  end for
  for $\alpha \in \mathcal{A}_{\text{active}}$ **do**
    if $\exists \alpha' \in \mathcal{A}_{\text{active}} : \mu(\alpha) + \epsilon(\alpha) < \mu(\alpha') - \epsilon(\alpha')$ **then**
    {Apply deactivation rule}
      $\mathcal{A}_{\text{active}} \leftarrow \mathcal{A}_{\text{active}} \setminus \{\alpha\}$
    end if
  end for
end while

where $\gamma_{t'}(\alpha) = \mathbb{I}\{\hat{y}_{t'} = y_{t'} \wedge \alpha_{t'} \leq \alpha \wedge y_{t'} \in \mathcal{C}_\alpha(x_{t'})\}$ and

$$\nu_{t'}(\alpha) = \gamma_{t'}(\alpha) + \mathbb{I}\{\hat{y}_{t'} \neq y_{t'} \wedge \alpha_{t'} \geq \alpha \wedge y_{t'} \in \mathcal{C}_{\alpha_{t'}}(x_{t'})\} + \mathbb{I}\{y_{t'} \notin \mathcal{C}_\alpha(x_{t'})\}$$

and, similarly as standard successive elimination, it deactivates any arm $\alpha \in \mathcal{A}_{\text{active}}$ for which there exists $\alpha' \in \mathcal{A}_{\text{active}}$ such that $\texttt{UCB}_t(\alpha) < \texttt{LCB}_t(\alpha')$ and repeats the entire procedure until T is reached or until $|\mathcal{A}_{\text{active}}| = 1$. Algorithm 1 summarizes the overall algorithm, which we refer to as counterfactual successive elimination, and Theorem 4.1 below formalizes its regret guarantees (proven in Appendix A):

Theorem 4.1. *Given a calibration set $\mathcal{D}_{cal} = \{(x_i, y_i)\}_{i=1}^m$, Algorithm 1 is guaranteed to achieve expected regret $\mathbb{E}[R(t)] \leq O\left(\sqrt{t \log m \log T}\right)$ for any $t \leq T$.*

Remarks. Algorithm 1 is designed to be used only during the design phase of the decision support system, not during deployment, since every time an expert predicts a sample with feature vector x, it assumes that the ground truth label y is observed, *i.e.*, it requires access to a set of labeled samples $\mathcal{D}_{\text{opt}} = \{(x_i, y_i)\}_{i=1}^T$, where $(x_i, y_i) \sim P(X)P(Y \mid X)$. However, after the design phase has finished, the decision support system does not require access to ground truth labels during deployment.

Interestingly, one can use counterfactual rewards to improve other well-known bandit algorithms, not only successive elimination. In section 5, we evaluate the benefits of using counterfactual rewards both in successive elimination (Algorithm 1) and in `UCB1`. To use counterfactual rewards, at each time step t, counterfactual `UCB1` pulls the arm $\alpha_t = \text{argmax}_{\alpha \in \mathcal{A}} \texttt{UCB}_t(\alpha)$ and counterfactually infers the rewards for any $\alpha > \alpha_t$ or $\alpha < \alpha_t$, similarly as Algorithm 1 does for $\tilde{\alpha}$. Remarkably, our experimental results demonstrate that counterfactual `UCB1` achieves lower expected regret than Algorithm 1. Motivated by this empirical finding, it would be very interesting, albeit challenging, to derive formal regret guarantees for counterfactual `UCB1` in future work. One of the main technical obstacles one would need to solve is that, in counterfactual `UCB1`, the number of counterfactually inferred rewards at each time step is unknown, in contrast with Algorithm 1, which at each time step counterfactually infers one half of the arms in $\mathcal{A}_{\text{active}}$.

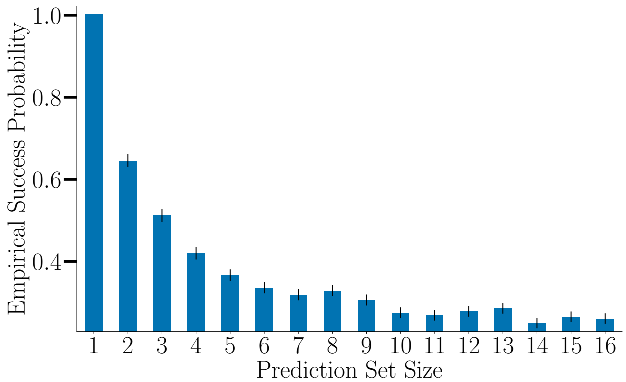
5. Experimental Evaluation via Human Subject Study

In this section, we conduct a large-scale human subject study and show that: a) human experts are more likely to succeed at predicting the ground truth label under smaller prediction sets, providing strong evidence that the counterfactual monotonicity assumption (Assumption 3.1) holds; b) counterfactual successive elimination and counterfactual `UCB1` achieve a significant improvement in expected regret in comparison with their vanilla implementations; c) a strict implementation of our decision support system, which adaptively limits experts' level of agency, offers greater performance than a lenient implementation, which allows experts to always exercise their own agency.[11]

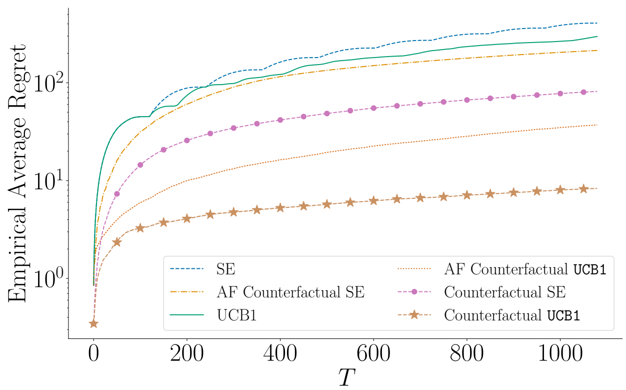
Human subject study setup. We gathered 194,407 label predictions from 2,751 human participants for 1,200 unique images from the ImageNet16H dataset (Steyvers et al., 2022) using Prolific. Our experimental protocol received approval from the Institutional Review Board (IRB) at the University of X, each participant was rewarded with 9£ per hour pro-rated, following Prolific's payment principles, and consented to participate by filling a consent form that included a detailed description of the study processes, and the collected data did not include any personally identifiable information. Each image in the ImageNet16H dataset belongs to one of 16 different categories (*e.g.*, animals, vehicles as well as every day objects), which serve as labels. In our study, we used always the same classifier, namely the pre-trained VGG-19 (Simonyan & Zisserman, 2014) after 10 epochs of fine-tuning as provided by Steyvers et al. (2022) and a fixed calibration set of 120 images, picked at random.[12] The average accuracy of the classifier (over the images not in the calibration set) is 0.848. For each image, we first computed all possible prediction sets that any conformal

[11] We will release an open-source implementation with the final version of the paper. All experiments ran on a Mac OS machine with an M1 processor and 16GB memory.

[12] The model and the dataset ImageNet16H are publicly available at https://osf.io/2ntrf/.



(a) Empirical success probability vs. prediction set size



(b) Empirical average regret vs. T

Figure 2. The panels show (a) the empirical success probability per prediction set size for images with the highest difficulty averaged across all experts for prediction sets that included the true label and (b) the empirical average regret achieved by six different bandit algorithms across 10 different realizations. In panel (a), the standard error is shown using error bars and, in panel (b), the standard error is shown using shaded areas, though is not visible as it is always below 0.2 at all times.

predictor using the above classifier and calibration set could construct. Then, we created 715 questionnaires, each with a set of images and, for each image, a multiple choice question using a prediction set. Under the strict implementation of our system, the multiple choice options included only the label values of the corresponding prediction set and, under the lenient implementation, they additionally included an option "Other", which allowed participants to pick a label value outside the prediction set. Under both implementations, the questionnaires covered all possible pairs of images and prediction sets. We provide screenshots of the questionnaires, the consent form as well as additional details on the experimental setup in Appendix C.

Expert success probability vs. prediction set size. As discussed previously, we cannot directly verify our counterfactual monotonicity assumption using an interventional study because it is a counterfactual property. However, we can verify Eq. 3, a necessary condition for our assumption to hold—whether, on average, experts are more likely to succeed at predicting the ground truth label using smaller prediction sets. To this end, we stratify the images in the dataset with respect to their difficulty into groups and, for each group, we estimate the success probability per prediction set size averaged across all experts and across experts with the same level of competence. We consider groups of images, rather than single images, because we have too few expert predictions to derive reliable estimates of the success probability per image. Refer to Appendix C for more details on how we stratify images and experts. Figure 2a summarizes the results for the images of highest difficulty, which suggest that experts are more accurate under smaller prediction sets. In Appendix D, we show qualitatively similar results for other groups of images and experts of varying difficulty and competence.

Regret analysis. In addition to validating the formal regret guarantees (Theorem 4.1) of counterfactual successive elimination (Algorithm 1), here, we aim to evaluate the competitive advantage that counterfactual rewards bring to several bandit algorithms in terms of expected regret. To this end, we estimate the expected regret over a horizon of 1,080 time steps over 10 different realizations of the following algorithms:[13] a) vanilla successive elimination (SE), b) vanilla `UCB1`, c) counterfactual successive elimination (Counterfactual SE), d) counterfactual `UCB1`, e) assumption-free counterfactual successive elimination (AF Counterfactual SE), and f) assumption-free counterfactual `UCB1` (AF Counterfactual `UCB1`). The last two algorithms do not use the counterfactual monotonicity assumption but, for any sample (x, y) and $\alpha \in \mathcal{A}$, they counterfactually infer that, for any $\alpha' \in \mathcal{A}$ such that $\alpha' \neq \alpha$ and $y \notin \mathcal{C}_{\alpha'}(x)$, the expert would have failed to predict the ground truth label had she used $\mathcal{C}_{\alpha'}$ and, for any $\alpha' \in \mathcal{A}$ such that $\mathcal{C}_{\alpha'}(x) = \mathcal{C}_{\alpha}(x)$, the expert would have predicted the same label had she used $\mathcal{C}_{\alpha'}$. Figure 2b summarizes the results, which show that counterfactual rewards provide a clear competitive advantage compared with their vanilla counterparts. Moreover, the results also show that, by using the counterfactual monotonicity assumption, Counterfactual SE and Counterfactual `UCB1` are clear winners and this suggests that the assumption may (approximately) hold in practice.

Strict vs lenient implementation of our system. One of the key motivations to forcefully ask experts to predict label values from the prediction sets provided by conformal prediction is to be able to trade-off how frequently the system misleads experts and the difficulty of the task the expert needs to solve (Straitouri et al., 2023). However,

[13] Given that we have expert predictions using all possible prediction sets for 1,080 images, not including the 120 images in the calibration set, we can run any bandit algorithm *faithfully* for 1,080 time steps.

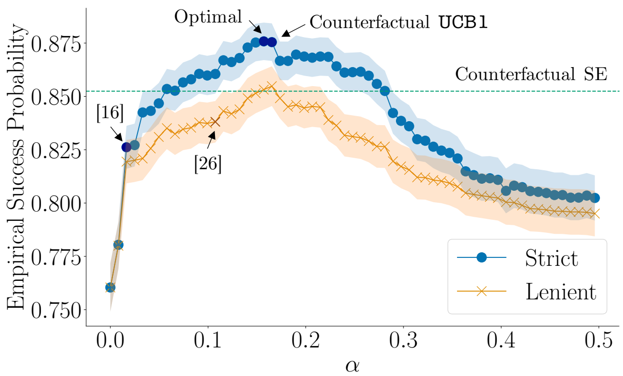


Figure 3. Empirical success probability achieved by all experts across all images using the strict and lenient implementation of our system with different α values. For the strict implementation, we annotate the optimal α value, the α values found by the algorithms by Straitouri et al. (2023) and by counterfactual `UCB1`, as well as the average success probability achieved by the set of α values that remain active after running counterfactual SE. We run all algorithms using 1,080 expert predictions (not to reuse any image). For the lenient implementation, we annotate the α value used by Babbar et al. (2022). The average accuracy of the classifier used by both the strict and the lenient implementation of our system is 0.848 with 0.01 standard error. The shaded areas correspond to a 95% confidence interval.

one could argue that a more lenient system, which allows experts to predict label values from outside the prediction sets as suggested by Babbar et al. (2022), may offer greater performance since, in principle, it does not forcefully mislead an expert whenever the prediction set does not contain the ground truth label. Here, we provide empirical evidence that suggests that this is not the case. Figure 3 demonstrates that a strict implementation of our system consistently offers greater performance than a lenient implementation across the full spectrum of competitive α values,[14] where we allow experts to predict any label value from $\mathcal{Y}$ whenever a prediction set is empty under both the strict and the lenient implementation of our system. Under the lenient implementation, the experts perform worse because 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 experts misplace their trust, *i.e.*, they predict a label value outside (within) the prediction set when it does (not) contain the true label. Refer to Appendix D for more details. Figure 3 also demonstrates that, using the same amount of expert data to optimize the value of α, counterfactual `UCB1` and counterfactual SE offer a significant advantage over the algorithm proposed by Straitouri et al. (2023), which uses a stylized expert model. Here, note that despite we only run the bandit algorithms for $T = 1{,}080$ time steps not to reuse any image nor expert prediction, counterfactual `UCB1` finds an α value that is very close to the optimal α^*.

[14]We have excluded values of $\alpha > 0.5$ to improve visibility, however, we include the full figure in Appendix D.

6. Conclusions

In this paper, we have looked at the development of decision support systems for classification tasks that forcefully ask experts to predict labels from prediction sets using online learning and counterfactual inference. Moreover, we have conducted a large-scale human subject study to validate the feasibility of such type of decision support systems. Our work opens up many interesting avenues for future work. For example, it would be worth to adapt other bandits algorithms so that they utilize counterfactual rewards and derive formal regret guarantees for them. Moreover, it would be important to reproduce the findings of our human subject study in other real-world domains with domain experts (*e.g.*, medical doctors). In addition, it would be also important to investigate the impact of biased classifiers on decision support systems such as ours as well as ways to mitigate such bias. Finally, it would be interesting to design mechanisms to adaptively limit expert's level of agency in other types of decision support systems (*e.g.*, large language models) and tasks.

References

Angelopoulos, A. N. and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Babbar, V., Bhatt, U., and Weller, A. On the utility of prediction sets in human-ai teams. In Raedt, L. D. (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 2457–2463. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/341. URL `https://doi.org/10.24963/ijcai.2022/341`. Main Track.

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., and Horvitz, E. Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 2019.

Beach, L. R. Broadening the definition of decision making: The role of prechoice screening of options. *Psychological Science*, 4(4):215–220, 1993.

Ben-Akiva, M. and Boccara, B. Discrete choice models with latent choice sets. *International Journal of Research in Marketing*, 12(1):9–24, 1995.

Bordt, S. and von Luxburg, U. When humans and machines make joint decisions: A non-symmetric bandit model. *arXiv preprint arXiv:2007.04800*, 2020.

Chen, W., Wang, Y., and Yuan, Y. Combinatorial multi-armed bandit: General framework and applications. In *International conference on machine learning*, pp. 151–159. PMLR, 2013.

Chernev, A., Böckenholt, U., and Goodman, J. Choice overload: A conceptual review and meta-analysis. *Journal of Consumer Psychology*, 25(2):333–358, 2015.

Chzhen, E., Denis, C., Hebiri, M., and Lorieul, T. Set-valued classification–overview via a unified framework. *arXiv preprint arXiv:2102.12318*, 2021.

De, A., Koley, P., Ganguly, N., and Gomez-Rodriguez, M. Regression under human assistance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

De, A., Okati, N., Zarezade, A., and Gomez-Rodriguez, M. Classification under human assistance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

de Kroon, A. A., Belgrave, D., and Mooij, J. M. Causal discovery for causal bandits utilizing separating sets. *arXiv:2009.07916*, 2020.

Dressel, J. and Farid, H. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

Haynes, G. A. Testing the boundaries of the choice overload phenomenon: The effect of number of options and time pressure on decision difficulty and satisfaction. *Psychology & Marketing*, 26(3):204–212, 2009.

Jiao, W., Atwal, G., Polak, P., Karlic, R., Cuppen, E., Danyi, A., de Ridder, J., van Herpen, C., Lolkema, M. P., Steeghs, N., et al. A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nature communications*, 11(1):1–12, 2020.

Kuksov, D. and Villas-Boas, J. M. When more alternatives lead to less choice. *Marketing Science*, 29(3):507–524, 2010.

Lai, V., Chen, C., Liao, Q. V., Smith-Renner, A., and Tan, C. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471*, 2021.

Lattimore, F., Lattimore, T., and Reid, M. D. Causal bandits: Learning good interventions via causal inference. *Advances in Neural Information Processing Systems*, 29, 2016.

Lee, S. and Bareinboim, E. Structural causal bandits: where to intervene? *Advances in Neural Information Processing Systems*, 31, 2018.

Liu, H., Tian, Y., Chen, C., Feng, S., Chen, Y., and Tan, C. Learning human-compatible representations for case-based decision support. In *The Eleventh International Conference on Learning Representations*, 2023.

Lubars, B. and Tan, C. Ask not what ai can do, but what ai should do: Towards a framework of task delegability. *Advances in Neural Information Processing Systems*, 32: 57–67, 2019.

Ma, L. and Denoeux, T. Partial classification in the belief function framework. *Knowledge-Based Systems*, 214: 106742, 2021.

Mortier, T., Wydmuch, M., Dembczyński, K., Hüllermeier, E., and Waegeman, W. Efficient set-valued prediction in multi-class classification. *Data Mining and Knowledge Discovery*, 35(4):1435–1469, 2021.

Mozannar, H. and Sontag, D. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*, pp. 7076–7087, 2020.

Nguyen, V.-L. and Hüllermeier, E. Multilabel classification with partial abstention: Bayes-optimal prediction under label independence. *Journal of Artificial Intelligence Research*, 72:613–665, 2021.

Okati, N., De, A., and Gomez-Rodriguez, M. Differentiable learning under triage. In *Advances in Neural Information Processing Systems*, 2021.

Papenmeier, A., Englebienne, G., and Seifert, C. How model accuracy and explanation fidelity influence user trust. *arXiv preprint arXiv:1907.12652*, 2019.

Pearl, J. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

Raghu, M., Blumer, K., Corrado, G., Kleinberg, J., Obermeyer, Z., and Mullainathan, S. The algorithmic automation problem: Prediction, triage, and human effort. *arXiv preprint arXiv:1903.12220*, 2019.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. On causal and anticausal learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 459–466, 2012.

Schwartz, B. The paradox of choice: Why more is less. *New York*, 2004.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Slivkins, A. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

Steyvers, M., Tejeda, H., Kerrigan, G., and Smyth, P. Bayesian modeling of human–ai complementarity. *Proceedings of the National Academy of Sciences*, 119(11): e2111547119, 2022.

Straitouri, E., Wang, L., Okati, N., and Gomez-Rodriguez, M. Improving expert predictions with conformal prediction. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Suresh, H., Lao, N., and Liccardi, I. Misplaced trust: Measuring the interference of machine learning in human decision-making. In *12th ACM Conference on Web Science*, pp. 315–324, 2020.

Vodrahalli, K., Gerstenberg, T., and Zou, J. Uncalibrated models can improve human-ai collaboration. In *Advances in Neural Information Processing Systems*, 2022.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic Learning in a Random World*. Springer-Verlag, Berlin, Heidelberg, 2005. ISBN 0387001522.

Wang, X. and Yin, M. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*, pp. 318–328, 2021.

Whitehill, J., Mohan, K., Seaton, D., Rosen, Y., and Tingley, D. Mooc dropout prediction: How to measure accuracy? In *Proceedings of the fourth (2017) acm conference on learning@ scale*, pp. 161–164, 2017.

Wright, P. and Barbour, F. *Phased decision strategies: Sequels to an initial screening*. Graduate School of Business, Stanford University, 1977.

Yang, G., Destercke, S., and Masson, M.-H. Cautious classification with nested dichotomies and imprecise probabilities. *Soft Computing*, 21(24):7447–7462, 2017.

Yin, M., Wortman Vaughan, J., and Wallach, H. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pp. 1–12, 2019.

Zhang, Y., Liao, Q. V., and Bellamy, R. K. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 295–305, 2020.

A. Proof of Theorem 4.1

Let us define the clean event $\mathcal{E} = \{|\hat{\mu}_t(\alpha) - \mathbb{E}[\mathbb{I}\{\hat{Y}_{\mathcal{C}_\alpha} = Y\}]| \leq \epsilon_t(\alpha)|, \forall \alpha \in \mathcal{A}, \forall t \leq T\}$ and decompose the average regret with respect to $\mathcal{E}$ as follows:

$$\mathbb{E}[R(t)] = \mathbb{E}[R(t) \,|\, \mathcal{E}]\mathbb{P}[\mathcal{E}] + \mathbb{E}[R(t) \,|\, \bar{\mathcal{E}}]\mathbb{P}[\bar{\mathcal{E}}], \tag{6}$$

where $\bar{\mathcal{E}}$ is the complement of $\mathcal{E}$. Then, following Slivkins (2019), we first assume $\mathcal{E} = \{|\hat{\mu}_t(\alpha) - \mathbb{E}[\mathbb{I}\{\hat{Y}_{\mathcal{C}_\alpha} = Y\}]| \leq \epsilon_t(\alpha)|, \forall \alpha \in \mathcal{A}, \forall t \leq T\}$ holds and then show that the probability that $\bar{\mathcal{E}}$ holds is negligible.

Let α be any suboptimal arm and α^* be the optimal one, *i.e.*, $\mathbb{E}[\mathbb{I}\{\hat{Y}_{\mathcal{C}_\alpha} = Y\}] < \mathbb{E}[\mathbb{I}\{\hat{Y}_{\mathcal{C}_{\alpha^*}} = Y\}]$. Let $t \leq T$ be the last round in which we have applied the deactivation rule and α was active. Until then, both α and α^* are active and, as a result, in each phase we collected a reward for each of them. Therefore, it must hold that $\nu_t(\alpha) = \nu_t(\alpha^*)$ and thus $\epsilon_t(\alpha^*) = \epsilon_t(\alpha)$. Further, since, by assumption, $\mathcal{E}$ holds, we have that:

$$\Delta(\alpha) = \mathbb{E}[\mathbb{I}\{\hat{Y}_{\mathcal{C}_{\alpha^*}} = Y\}] - \mathbb{E}[\mathbb{I}\{\hat{Y}_{\mathcal{C}_\alpha} = Y\}] \leq 2(\epsilon_t(\alpha^*) + \epsilon_t(\alpha)) = 4\epsilon_t(\alpha). \tag{7}$$

Now, given that α is deactivated whenever the deactivation rule is applied again, we will either collect (or counterfactually infer) one reward value for α after t, for $t < T$, or will not collect any reward value again, for $t = T$. As a result, $\nu_t(\alpha) \leq \nu_T(\alpha) \leq 1 + \nu_t(\alpha)$, for $t \leq T$. Next, using Eq. 7, for any suboptimal α, we have:

$$\Delta(\alpha) \leq O(\epsilon_T(\alpha)) = O\left(\sqrt{\frac{\log T}{\nu_T(\alpha)}}\right) = O\left(\sqrt{\frac{\log T}{\nu_T(\alpha)}}\right). \tag{8}$$

From the above, it immediate follows that anytime we pull α, we suffer average regret $\Delta(\alpha)$. Consequently, until time $t \leq T$, the total average regret due to pulling arm α, which we denote as $\mathbb{E}[R(\alpha\,;\,t) \,|\, \mathcal{E}]$, is given by $\mathbb{E}[R(\alpha\,;\,t) \,|\, \mathcal{E}] = n_t(\alpha) \cdot \Delta(\alpha)$, where $n_t(\alpha)$ denotes the number of times that arm α has been pulled until t. Thus the total average regret, conditioned on $\mathcal{E}$, is given by:

$$\mathbb{E}[R(t) \,|\, \mathcal{E}] = \sum_{\alpha \in \mathcal{A}} \mathbb{E}[R(\alpha\,;\,t) \,|\, \mathcal{E}] = \sum_{\alpha \in \mathcal{A}} n_t(\alpha) \cdot \Delta(\alpha) \leq \sum_{\alpha \in \mathcal{A}} n_t(\alpha) \cdot O\left(\sqrt{\frac{\log T}{\nu_t(\alpha)}}\right), \tag{9}$$

given Eq. 8, for $t \leq T$.

Now, we will derive a lower bound for $\nu_t(\alpha)$. Assume that, until time step t, the deactivation rule has been applied $n > 0$ times. Let $m_1, m_2, ..., m_n$, where $0 \leq m_i \leq m, i = 1, .., n$, be the number of arms that are not active after the i-th time the deactivation rule is applied. Before applying the deactivation rule, we collect (or counterfactually infer) one reward value for each arm with a number of pulls that is logarithmic to the number of active arms. Before applying the rule for the first time, all m arms are active, so we will need $\lceil \log m \rceil$ rounds to collect a reward for each active arm. Then, given that m_1 arms are deactivated, $m - m_1$ remain active, so we will need $\lceil \log(m - m_1) \rceil$ rounds to collect a reward for each active arm until the deactivation rule is applied again. As a result, for $t \leq T$, we have that:

$$t = \lceil \log m \rceil + \lceil \log(m - m_1) \rceil + \lceil \log(m - m_2) \rceil + \ldots + \lceil \log(m - m_{n-1}) \rceil.$$

Each logarithmic term in the above equation corresponds to the process of collecting one reward for each arm and, as a result, one reward for α. As a result, the above has $\nu_t(\alpha)$ logarithmic terms and, given that $0 \leq m_i, i = 1, .., n$, we have that

$$t = \lceil \log m \rceil + \lceil \log(m - m_1) \rceil + \lceil \log(m - m_2) \rceil + \ldots + \lceil \log(m - m_{n-1}) \rceil \leq \nu_t(\alpha) \cdot \lceil \log m \rceil.$$

Therefore, we can conclude that $\nu_t(\alpha) \geq t / \lceil \log m \rceil$, for $t \leq T$, and, using Eq. 9, we have that

$$\begin{aligned}\mathbb{E}[R(t) \,|\, \mathcal{E}] \leq \sum_{\alpha \in \mathcal{A}} n_t(\alpha) \cdot O\left(\sqrt{\frac{\log T}{\nu_t(\alpha)}}\right) &\leq \sum_{\alpha \in \mathcal{A}} n_t(\alpha) \cdot O\left(\sqrt{\frac{\log m \log T}{t}}\right) \\ &= O\left(\sqrt{\frac{\log m \log T}{t}}\right) \cdot \sum_{\alpha \in \mathcal{A}} n_t(\alpha).\end{aligned}$$

Further, since, until time step t, it must hold that $\sum_{\alpha \in \mathcal{A}} n_t(\alpha) = t$, we can conclude that

$$\mathbb{E}[R(t) \,|\, \mathcal{E}] \leq O\left(\sqrt{\frac{\log m \log T}{t}}\right) \cdot t = O\left(\sqrt{t \log m \log T}\right). \tag{10}$$

Let us now lift the assumption that $\mathcal{E}$ holds. From Hoeffding's bound and a union bound, it readily follows that:

$$\mathbb{P}[\mathcal{E}] \geq 1 - mT\delta \Rightarrow 1 - \mathbb{P}[\mathcal{E}] \leq mT\delta \Rightarrow \mathbb{P}[\bar{\mathcal{E}}] \leq mT\delta \tag{11}$$

Moreover, given that the rewards take values in $\{0, 1\}$, it holds that at time step t, $R(t) \leq t$. Consequently, combining Eqs. 10, Eq. 11 and Eq. 6, we can conclude that:

$$\begin{aligned}
\mathbb{E}[R(t)] &= \mathbb{E}[R(t) \,|\, \mathcal{E}]\mathbb{P}[\mathcal{E}] + \mathbb{E}[R(t) \,|\, \bar{\mathcal{E}}]\mathbb{P}[\bar{\mathcal{E}}] \leq \mathbb{E}[R(t) \,|\, \mathcal{E}] + \mathbb{E}[R(t) \,|\, \bar{\mathcal{E}}]\mathbb{P}[\bar{\mathcal{E}}] \\
&= O\left(\sqrt{t \log m \log T}\right) + \frac{2tmT}{T^4},
\end{aligned}$$

where the first inequality uses that $\mathbb{P}[\mathcal{E}] \leq 1$. Finally, under the assumption that $m \leq T$, the above becomes:

$$\begin{aligned}
\mathbb{E}[R(t)] &\leq O\left(\sqrt{t \log m \log T}\right) + \frac{2tmT}{T^4} \\
&\leq O\left(\sqrt{t \log m \log T}\right) + \frac{2tT^2}{T^4} = O\left(\sqrt{t \log m \log T}\right) + O\left(\frac{t}{T^2}\right) \\
&= O\left(\sqrt{t \log m \log T}\right).
\end{aligned}$$

This concludes the proof.

B. Algorithm 2

Algorithm 2 It updates $\mathcal{A}_{\text{unexplored}}$, γ and ν

Input: $\mathcal{A}_{\text{unexplored}}, \gamma, \nu, \tilde{\alpha}, x, y, \hat{y}$
if $\mathbb{I}\{\hat{y} = y\} = 0$ **then**
  if $y \notin \mathcal{C}_{\tilde{\alpha}}(x)$ **then**
    for $\alpha' \in \mathcal{A}_{\text{unexplored}} : \alpha' \geq \tilde{\alpha}$ **do**
      $\nu(\alpha') \leftarrow \nu(\alpha') + 1$
    end for
    $\mathcal{A}_{\text{unexplored}} \leftarrow \mathcal{A}_{\text{unexplored}} \setminus \{\alpha' : \alpha' \geq \tilde{\alpha}\}$
  else
    for $\alpha' \in \mathcal{A}_{\text{unexplored}} : \alpha' \leq \tilde{\alpha}$ **do**
      $\nu(\alpha') \leftarrow \nu(\alpha') + 1$
    end for
    $\mathcal{A}_{\text{unexplored}} \leftarrow \mathcal{A}_{\text{unexplored}} \setminus \{\alpha' : \alpha' \leq \tilde{\alpha}\}$
  end if
else
  $\alpha^{\dagger} \leftarrow \inf\{\alpha : y \notin \mathcal{C}_{\alpha}(x)\}$
  for $\alpha' \in \mathcal{A}_{\text{unexplored}} : \alpha' \geq \alpha^{\dagger}$ **do**
    $\nu(\alpha') \leftarrow \nu(\alpha') + 1$
  end for
  for $\alpha' \in \mathcal{A}_{\text{unexplored}} : \tilde{\alpha} \leq \alpha' < \alpha^{\dagger}$ **do**
    $\nu(\alpha') \leftarrow \nu(\alpha') + 1$
    $\gamma(\alpha') \leftarrow \gamma(\alpha') + 1$
  end for
  $\mathcal{A}_{\text{unexplored}} \leftarrow \mathcal{A}_{\text{unexplored}} \setminus \{\alpha' : \alpha' \geq \tilde{\alpha}\}$
end if
return $\mathcal{A}_{\text{unexplored}}, \gamma, \nu$

C. Additional Details about the Human Subject Study Setup

Human subject study consent form. Figure 4 shows screenshots of the consent form that Prolific workers had to fill in order to participate in our study under the strict implementation of our system. We used a similar consent form for our study under the lenient implementation of our system except for the "Procedures" and "Example Question" sections, which we show in Figure 5.

Dataset. The ImageNet16H dataset (Steyvers et al., 2022) contains $1{,}200$ unique images, where each of these images has one of the following ground truth labels: chair, oven, knife, bottle, keyboard, clock, boat, bicycle, airplane, truck, car, elephant, bear, dog, cat, and bird. Each image is included four times, each time with a different phase noise distorsion value $\omega \in \{80, 95, 110, 125\}$. The phase noise distortion controls the difficulty of the classification task; the higher the noise, the more difficult the classification task. In our experiments, we used all $1{,}200$ unique images with $\omega = 110$ because, under such noise value, humans sometimes, but not always, failed at solving the prediction task (*i.e.*, the empirical success probability achieved by the human experts on their own was 0.76).

Implementation details. We implemented our algorithms on Python 3.10.9 using the following libraries:

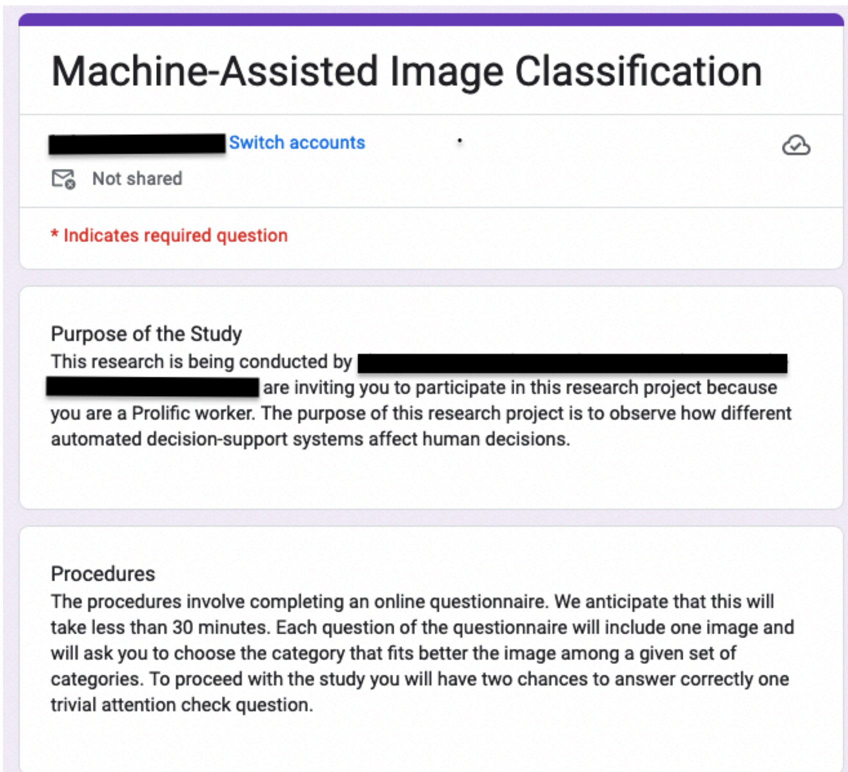
- NumPy 1.24.1 (BSD-3-Clause License).
- Pandas 1.5.3 (BSD-3-Clause License).
- Scikit-learn 1.2.2 (BSD License).

For reproducibility, we used a fixed random seed in all random procedures (a different one) for each realization of the algorithms. Similarly, we used a fixed random seed to randomly pick the 120 images of the calibration set.

Stratifying images and users. We stratify the images into groups of similar difficulty, where we measure the difficulty of each image using the empirical success probability of experts predicting its ground truth label. More specifically, for each image we first compute the empirical success probability over all experts' predictions. Then, we stratify the images into 5 mutually exclusive groups as follows:

— Highest difficulty: images with empirical success probability within the 20% percentile of the empirical success probabilities of all images.
— Medium to high difficulty: images with empirical success probability within the 40% percentile and outside the 20% percentile of the empirical success probabilities of all images.
— Medium difficulty: images with empirical success probability within the 60% percentile and outside the 40% percentile the of the empirical success probabilities of all images.
— Low difficulty: images with empirical success probability within the 80% percentile and outside the 60% percentile the of the empirical success probabilities of all images.
— Lowest difficulty: images with empirical success probability outside the 80% percentile of the empirical success probabilities of all images.

We follow a similar method to stratify experts into two groups based on their level of competence, where we measure the level of competence of an expert using the empirical success probability of predicting the ground truth label across all the predictions that she made. As a result we consider the 50% of users with the highest empirical success probability as the experts with high level of competence and the rest 50% as experts with low level of competence.



Machine-Assisted Image Classification

[redacted] Switch accounts

Not shared

* Indicates required question

Purpose of the Study

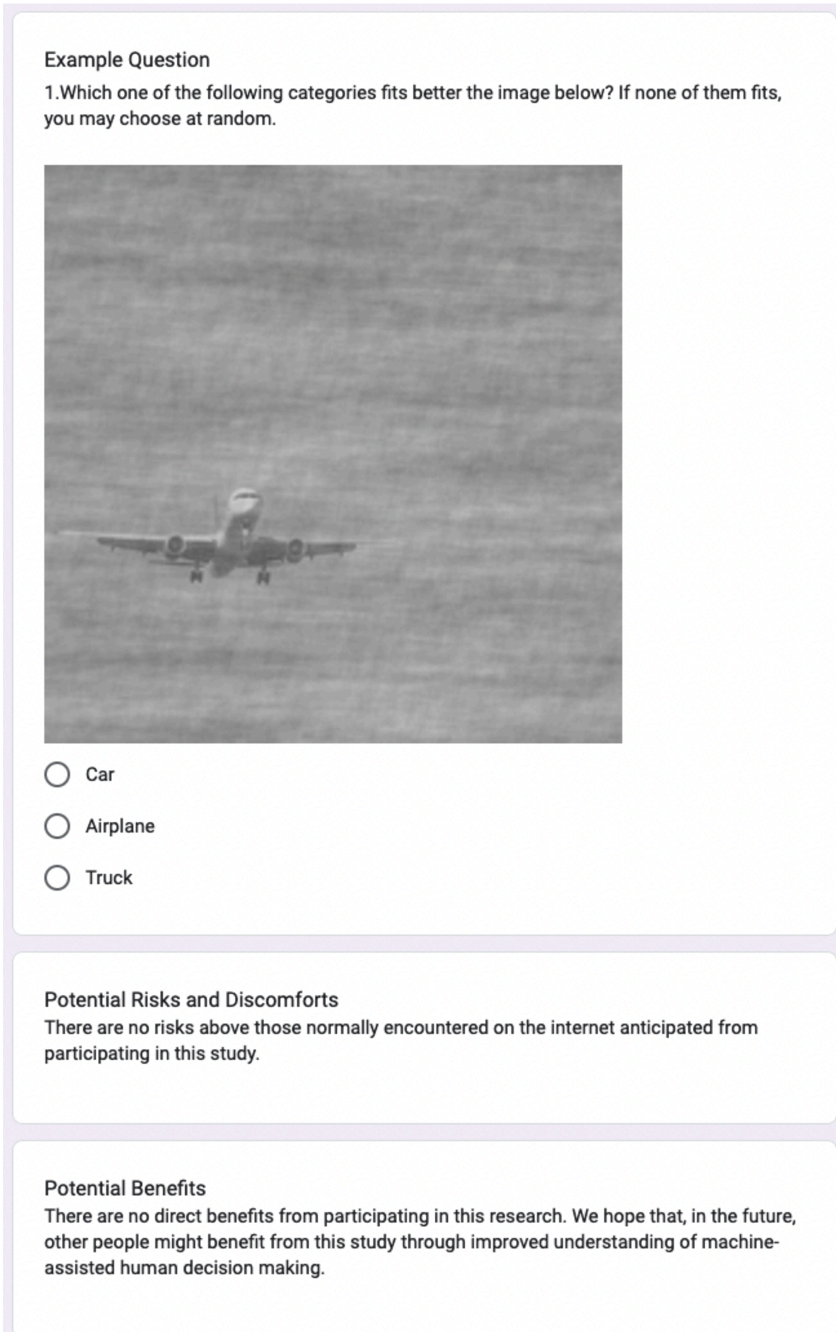
This research is being conducted by [redacted] are inviting you to participate in this research project because you are a Prolific worker. The purpose of this research project is to observe how different automated decision-support systems affect human decisions.

Procedures

The procedures involve completing an online questionnaire. We anticipate that this will take less than 30 minutes. Each question of the questionnaire will include one image and will ask you to choose the category that fits better the image among a given set of categories. To proceed with the study you will have two chances to answer correctly one trivial attention check question.

(a) Purpose of the study and procedures

Figure 4. The consent form including a detailed description of the study processes that Prolific workers had to read and fill in order to participate in our human subject study. The procedures describe use of the decision support systems under the strict implementation. The consent form continues in Figures 4b and 4c. Any personally identifiable information has been removed to preserve anonymity.



Example Question

1.Which one of the following categories fits better the image below? If none of them fits, you may choose at random.

- Car
- Airplane
- Truck

Potential Risks and Discomforts

There are no risks above those normally encountered on the internet anticipated from participating in this study.

Potential Benefits

There are no direct benefits from participating in this research. We hope that, in the future, other people might benefit from this study through improved understanding of machine-assisted human decision making.

(b) Example question, potential risk and discomforts and potential benefits

Figure 4. Consent form continued.

Confidentiality

There is no risk of loss of confidentiality, as all study data will be collected and stored anonymously. All study data will not include any personal data, and will be made publicly available to encourage future research on machine-assisted decision making.

Compensation

You will be compensated pursuant with your agreement with Prolific for your participation. You will be responsible for any taxes assessed on the compensation.

Right to Withdraw and Questions

Your participation in this research is completely voluntary. You may choose not to take part at all. If you decide to participate in this research, you may stop participating at any time. If you decide not to participate in this study or if you stop participating at any time, you will not be penalized or lose any benefits to which you otherwise qualify.

If you decide to stop taking part in the study, if you have questions, concerns, or complaints, please contact the investigator:

[redacted]

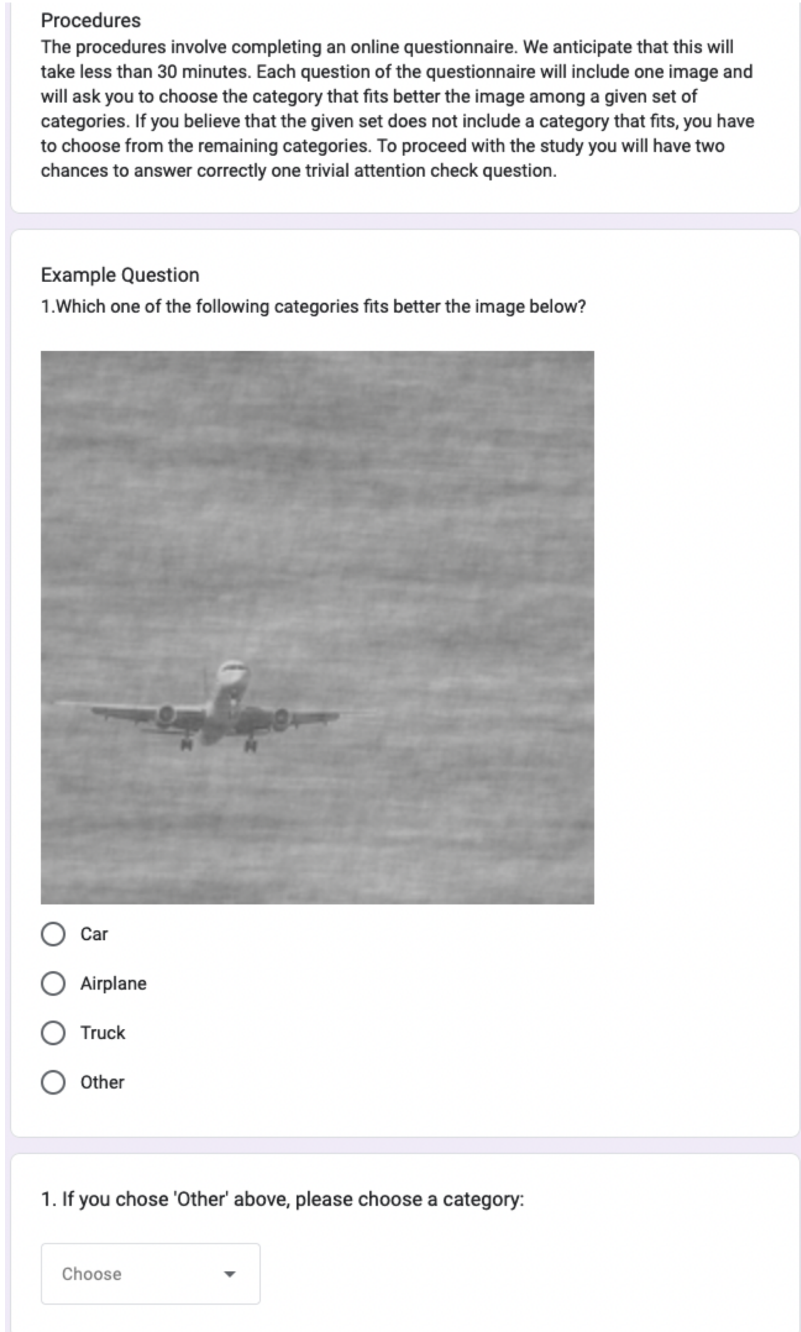
Consent *

I confirm that I am 18 years of age or older, I have read this consent form in its entirety or had it read to me, and I voluntarily consent to participate in this research.

- YES
- NO

(c) Confidentiality, compensation, right to withdraw, questions, and consent

Figure 4. Consent form continued.



Procedures

The procedures involve completing an online questionnaire. We anticipate that this will take less than 30 minutes. Each question of the questionnaire will include one image and will ask you to choose the category that fits better the image among a given set of categories. If you believe that the given set does not include a category that fits, you have to choose from the remaining categories. To proceed with the study you will have two chances to answer correctly one trivial attention check question.

Example Question

1.Which one of the following categories fits better the image below?

- Car
- Airplane
- Truck
- Other

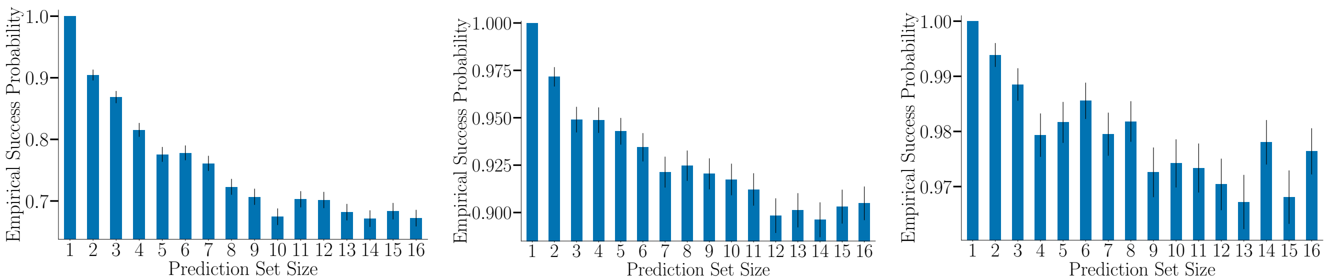
1. If you chose 'Other' above, please choose a category:

Choose

Figure 5. Procedures and example question included in the consent form that Prolific workers had to fill in order to participate in our study under the lenient implementation of our systems.

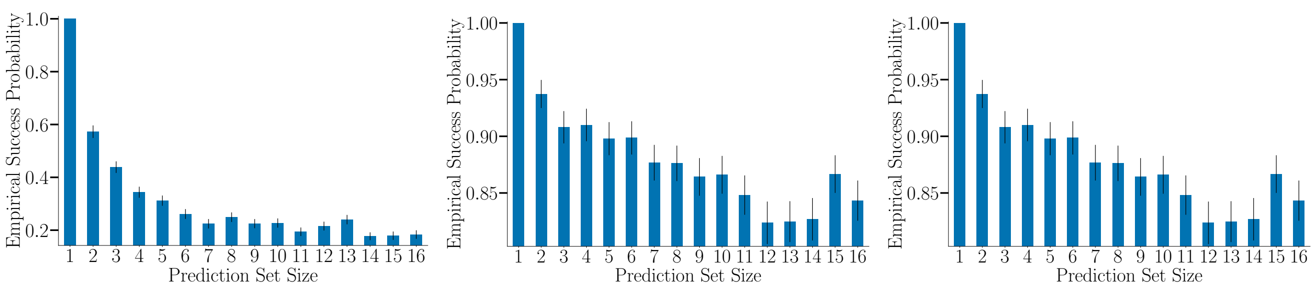
D. Additional Results about the Human Subject Study

Monotonicity. Figures 6 and 7 show the empirical success probability per prediction set size across images with different difficulty levels and experts with different levels of competence. Similarly as in Figure 2 in the main, the results show that, as long as the images are not too easy, the experts are more accurate under smaller prediction sets.

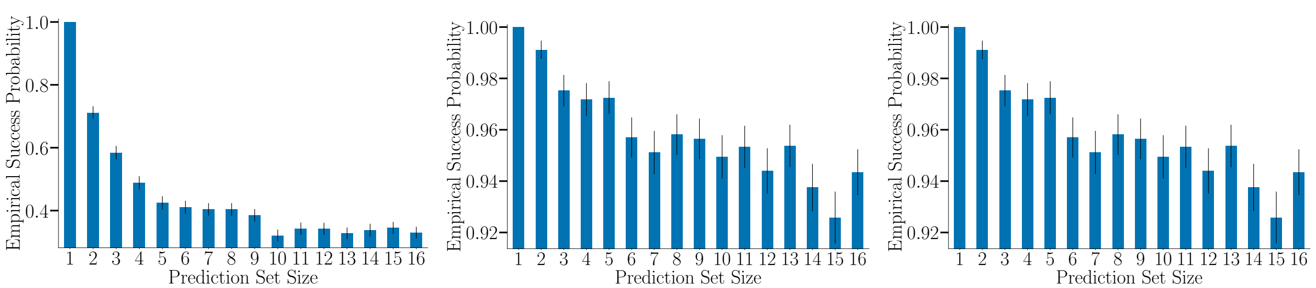


(a) Medium to high difficulty (b) Medium difficulty (c) Medium to low difficulty

Figure 6. Empirical success probability and standard error per prediction set size for (a) images with medium to high difficulty, (b) images with medium difficulty and (c) images with medium to low difficulty averaged across all experts for prediction sets that included the true label.



(a) Highest difficulty (b) Medium to high difficulty (c) Medium difficulty



(d) Highest difficulty (e) Medium to high difficulty (f) Medium difficulty

Figure 7. Empirical success probability and standard error per prediction set size for prediction sets that included the true label averaged across experts with low level of competence (top row) and averaged across experts with high level of competence (bottom row) for images with varying difficulty.

Expert accuracy under the strict and lenient implementation of our systems. Figure 8 shows the empirical success probability under the full range of values of $\alpha \in [0, 1]$. The results show that, as the α value increases, the empirical success probability under both the strict and the lenient implementation of our system converges to the success probability of the experts' choosing on their own, *i.e.*, choosing from $\mathcal{Y}$. This happens because, the larger the α value, the more often happens that the prediction set is the empty set and thus we allow the expert to choose from $\mathcal{Y}$ under both implementations, as discussed in Footnote 7 in the main paper.

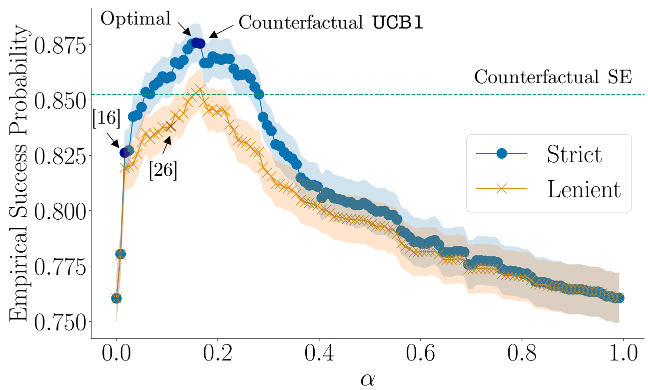


Figure 8. Empirical success probability achieved by all experts across all images using the strict and lenient implementation of our system with different α values. The annotated α values as well as the horizontal dashed line are the same as in Figure 3. The shaded areas correspond to a 95% confidence interval.

Figure 9 shows the number of predictions in which the prediction sets do not contain the true label and the experts succeed, *i.e.*, $\sum_{x,y} \mathbb{I}\{\hat{Y}_{\mathcal{C}_\alpha} = y \wedge y \notin \mathcal{C}_\alpha(x)\}$, and the number of predictions in which the experts misplace their trust, *i.e.*,

$$\sum_{x,y} \mathbb{I}\{\hat{Y}_{\mathcal{C}_\alpha} \notin \mathcal{C}_\alpha(x) \wedge y \in \mathcal{C}_\alpha(x)\} + \mathbb{I}\{\hat{Y}_{\mathcal{C}_\alpha} \in \mathcal{C}_\alpha(x) \wedge y \notin \mathcal{C}_\alpha(x)\},$$

under the lenient implementation, where the sums are over each sample x with true label y not used in the calibration set. The results show that, consistently across α values, 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 experts misplace their trust. This explains why the lenient implementation compares unfavorably against the strict implementation in Figure 8.

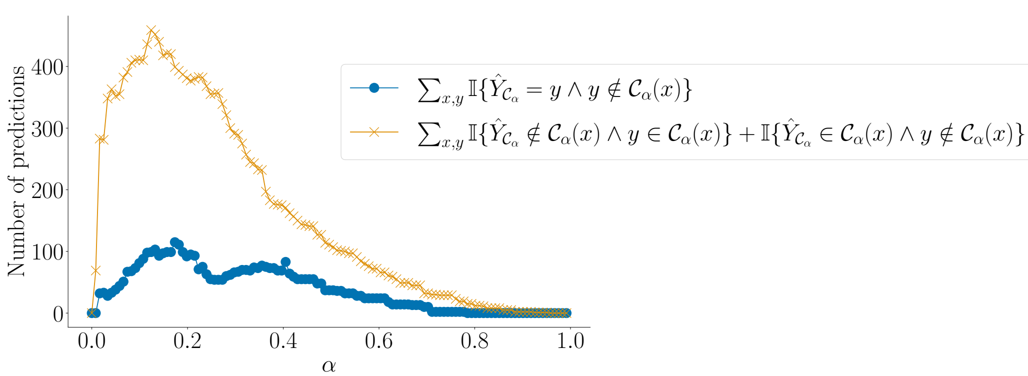


Figure 9. Total number of experts' predictions in which the prediction sets do not contain the true label and the experts succeed (dots) and in which the experts misplace their trust ('x's) under the lenient implementation. The sums are over the 1,080 images not used in the calibration set.