SHOULD I STAY OR SHOULD I SHOW? LEARNING TO SELECTIVELY DISCLOSE INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In many high-stakes settings, human decision-makers can acquire support information before making a decision. However, acquiring information is costly, and disclosure may fail to improve human decisions or may even impair them. We tackle this problem by studying *selective disclosure*, i.e., the problem of learning when to reveal support information to a human decision-maker under a budget constraint. We first show that the optimal policy is a threshold rule on the Value of Information (VoI), i.e., the expected reduction in human decision risk induced by disclosure. Since VoI is unknown in practice, we estimate the regime-specific human risks and bound the possible degradation of the resulting plug-in policy relative to lack of disclosure, as well as its regret relative to the optimal policy. Experiments on benchmark datasets show that selective disclosure outperforms both no disclosure and full disclosure, regardless of whether the support information is beneficial or harmful. Two user studies show that human-AI team performance can improve when disclosure is led by our learned policy and not human-selected, although this advantage varies across tasks. A counterfactual benchmark, which replaces participants' predictions with a machine-learning prediction when disclosure occurs, suggests that these differences might depend on lower adherence to advice when the information is automatically provided rather than self-requested.

1 INTRODUCTION

In many decision-making tasks, humans predict uncertain outcomes with the aid of support information. Financial analysts may conduct costly research to inform investment decisions (Xiong & Yang, 2023), physicians may order invasive diagnostic tests (Kasivisvanathan et al., 2018), and zoologists may collect biological samples to identify endangered species (Li et al., 2017). Such information can improve decisions, but acquiring it is often costly (Saar-Tsechansky et al., 2009). Moreover, support information may also be redundant or even harmful when it introduces noise, confusion, or over-reliance on imperfect external feedback (Romeo & Conti, 2026). Information disclosure must therefore balance the cost of providing information against its potential effect on decision quality.

To tackle this challenge, we introduce *Learning to Selectively Disclose* (LSD), a decision-theoretic framework that learns when to show additional support information to a human decision-maker within a limited disclosure budget (Fig. 1). Rather than asking whether information is informative in general, LSD asks a more decision-relevant question: *will revealing information improve the final decision?* We formalize this benefit through the conditional Value of Information (VoI), defined as the reduction in human decision risk caused by disclosure.

More precisely, we show that the optimal policy prescribes disclosure for cases with the largest positive VoI. Because VoI is not observable, we cast its estimation as a causal inference problem: we map VoI to the negative conditional average treatment effect of information disclosure on human decision loss and we establish conditions for its identification. Then, we derive guarantees for plug-in policies relative to no disclosure and the optimal policy, and introduce a class-wise VoI estimator for classification under the $0-1$ loss. Experiments on both synthetic and real-world data show that the learned policy identifies instances in which disclosure improves human decisions. We further evaluate our approach with two user studies on two different tasks, showing that our learned policy either outperforms or performs on par with human-selected disclosure. A counterfactual benchmark,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

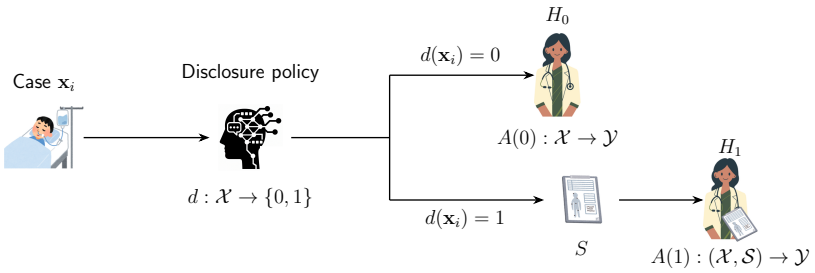


Figure 1: **LSD pipeline**. For each case $\mathbf{x}_i$, a disclosure policy $d : \mathcal{X} \to \{0, 1\}$ decides when to disclose support information S to a human expert H. If no support information is disclosed ($d(\mathbf{x}_i) = 0$), a human expert makes a decision using only baseline features. When support information is disclosed ($d(\mathbf{x}_i) = 1$), a human expert uses both baseline and support information.

which replaces human predictions with Machine Learning (ML) predictions for disclosed items, suggests that lower adherence rates to the ML predictions may partly account for these results.

Our Contributions. Our main contributions are:

1. We present selective disclosure as a two-regime decision problem and show that the optimal budget-constrained policy discloses information for instances with the largest positive conditional Value of Information (VoI);
2. We formulate selective disclosure as a causal intervention on the decision-maker's information, characterizing the conditional VoI as the causal reduction in human decision risk induced by disclosure. Then, we establish the assumptions under which VoI is identifiable;
3. We derive performance guarantees for plug-in disclosure policies in terms of their VoI estimation error, both relative to the no-disclosure baseline and relative to the optimal policy under the same budget constraint;
4. For classification under the 0-1 loss, we introduce a class-wise estimator of the regime-specific risks and demonstrate its effectiveness on synthetic and real-world datasets;
5. We run two user studies to compare human-AI team performance in classification tasks under policy-selected and human-selected disclosure.

2 Preliminaries

We study a decision-making setting in which a human observes fixed baseline information in order to perform an action. Before selecting the action, the human may additionally receive support information to help its decision. More formally, let $\mathcal{X} \subseteq \mathbb{R}^m$ be the baseline information space and let $\mathcal{Y} = \{1, \ldots, |\mathcal{Y}|\}$ be a finite target space. A case consists of baseline information $\mathbf{X} \in \mathcal{X}$, a target $Y \in \mathcal{Y}$, and support information $S \in \mathcal{S}$, where S may, e.g., be an ML model prediction. Although the realization of S may differ across cases, we treat it as either disclosed in full or withheld. A binary variable $D \in \{0, 1\}$ denotes the disclosure regime. We define the action space $\mathcal{A}$ and, throughout the paper, we set $\mathcal{A} = \mathcal{Y}$, so that the human action is a prediction of the target label. Let $\mathcal{H}$ denote a population of human decision-makers and let π be a distribution over $\mathcal{H}$. For a decision-maker $H \sim \pi$, we denote by $A_H(0) \in \mathcal{A}$ and $A_H(1) \in \mathcal{A}$ the actions that H would take under no disclosure and disclosure, respectively. These actions may vary across decision-makers even for the same case. We further assume access to a dataset $\mathcal{T} = \{(\mathbf{x}_i, y_i, a_{Hi}(0), a_{Hi}(1))\}_{i=1}^n$, sampled from some unknown P over $\mathcal{X} \times \mathcal{Y} \times \mathcal{A} \times \mathcal{A}$, where $a_{Hi}(0)$ and $a_{Hi}(1)$ are decisions collected for case i under the two disclosure regimes. In particular, each case is assigned to (at least) two distinct decision-makers $H_0, H_1 \sim \pi$, with H_0 operating under $D = 0$ and H_1 operating under $D = 1$. Notably, the dataset contains observations from both regimes for each case, but it does not contain both potential actions for the same human-case episode. Thus, our objective is to learn a disclosure policy for a decision-maker drawn from the population π, rather than a policy for a fixed human.

3 LEARNING TO SELECTIVELY DISCLOSE

The objective of LSD is to determine when a decision-maker should be given access to support information under a budget constraint. We frame the problem from a decision-theoretic perspective. Given a non-negative loss function $\ell : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}^+$ we define the regime-specific conditional risks:

$$r_D(\mathbf{x}) = \mathbb{E}\left[\ell\big(A_H(D), Y\big) \mid \mathbf{X} = \mathbf{x}\right], \qquad D \in 0, 1, \tag{1}$$

where the expectation is taken over the target Y, the decision-maker $H \sim \pi$, and any randomness in the human response, conditional on $\mathbf{X} = \mathbf{x}$. Thus, $r_D(\mathbf{x})$ measures the expected loss under disclosure regime D for an *average* decision-maker drawn from π, rather than on any particular individual. Smaller values correspond to more reliable decisions, while larger values indicate greater expected error. We define the *conditional Value of Information* (VoI) as the expected reduction in decision risk induced by disclosure, namely:

$$\text{VOI}(\mathbf{x}) = r_0(\mathbf{x}) - r_1(\mathbf{x}) \tag{2}$$

with positive values indicating that revealing S improves decision quality.

Let $\mathcal{D}$ be the set of measurable disclosure policies $d : \mathcal{X} \to \{0, 1\}$, where $d(\mathbf{x}) = 1$ indicates that the support information is disclosed. We denote the population disclosure risk with $R(d) = \mathbb{E}_{\mathbf{x}}[(1-d(\mathbf{x}))r_0(\mathbf{x}) + d(\mathbf{x})r_1(\mathbf{x})]$, i.e. the expected loss incurred when each instance is judged under the regime that d selects for it; in particular, $R(0)$ and $R(1)$ are the no- and full-disclosure baselines, obtained by never and always revealing S. Typically, access to more information is costly, thus we assume there is a budget level $B \in [0, 1]$ that limits the expected disclosure rate. The Learning to Selectively Disclose problem can be formulated as follows:

$$\min_{d \in \mathcal{D}} \quad R(d) \qquad \text{s.t.} \quad \mathbb{E}_{\mathbf{x}}[d(\mathbf{x})] \leq B \tag{3}$$

In the next subsections, we characterize the optimal disclosure policy for the LSD problem, formalize the causal estimation of VoI, provide theoretical guarantees for the corresponding plug-in approaches and then showcase how to estimate them when ℓ is the 0-1 loss, a common measure in human-AI collaboration literature (Ruggieri & Pugnana, 2025).

3.1 OPTIMAL DISCLOSURE POLICY

A first question to address is what an optimal policy for LSD looks like. We show that the optimal policy consists of a threshold rule over the VoI, as stated in the following Theorem:

Theorem 1 (Optimal Disclosure Policy)**.** *Assume that the cumulative distribution of* $\text{VOI}(\mathbf{x})$ *is continuous. Let* q_{1-B} *denote the* $(1-B)$*-quantile of* $\text{VOI}(\mathbf{x})$ *and let* $\lambda^* = \max\{0,\, q_{1-B}\}$*. For any budget* $B \in [0, 1]$*, the optimal disclosure policy is the threshold rule:*

$$d^*(\mathbf{x}) = \begin{cases} 1 & \textit{if} \quad \text{VOI}(\mathbf{x}) \geq \lambda^*, \\ 0 & \textit{otherwise}. \end{cases} \tag{4}$$

Proof. We provide the proof in Appendix A.1. □

Theorem 1 shows that selective disclosure admits an intuitive interpretation: support information is revealed only when it is expected to *benefit* the decision-maker, i.e., when the expected risk reduction is positive and exceeds a budget-dependent threshold. However, $\text{VOI}(\mathbf{x})$ is not directly observable, since a decision-maker either receives S or does not. As a result, the two risks defining $\text{VOI}(\mathbf{x})$ cannot be jointly observed for the same decision instance. Estimating $\text{VOI}(\mathbf{x})$ is therefore naturally a causal inference problem. In the following section, we formalize the corresponding causal framework and state the assumptions under which $\text{VOI}(\mathbf{x})$ is identifiable from observed data.

3.2 A CAUSAL INFERENCE INTERPRETATION

Since $\text{VOI}(\mathbf{x})$ is never observed for the same human decision-maker, a natural way to handle its estimation is resorting to the Neyman-Rubin potential outcomes framework (Rubin, 1974). However, our framework differs from the canonical treatment-allocation setting in an important aspect. In

standard policy learning (Kitagawa & Tetenov, 2018), a treatment is assigned to a unit and directly affects that unit's outcome. In LSD, instead, the intervention acts on a decision-maker: disclosure changes the information available to a human who must make a decision about a separate case. The case outcome Y is not affected by disclosure; what may change is the human action $A_H(D)$ and, consequently, the decision loss $\ell(A_H(D), Y)$. Thus, $\ell(A_H(0), Y)$ and $\ell(A_H(1), Y)$ are the two potential outcomes of interest, and the causal effect of our interest is not the effect of an intervention on the case itself, but the effect of disclosing support information on the quality of a human decision. More precisely, for a fixed decision-maker h, we define the human-specific conditional effect of disclosure on decision loss as $\tau_h(\mathbf{x}) = \mathbb{E}[\ell(A_h(1), Y) - \ell(A_h(0), Y) \mid X = \mathbf{x}, H = h]$. Because our objective is to learn a policy for a future decision-maker drawn from a target population π, we consider the population-average effect $\tau_\pi(\mathbf{x}) = \mathbb{E}_{H\sim\pi}[\tau_H(\mathbf{x})]$ and note that $-\tau_\pi(\mathbf{x}) = \text{VOI}(\mathbf{x})$. Hence, $\text{VOI}(\mathbf{x})$ measures the causal effect of disclosing information on the quality of a decision made about a case with covariates $\mathbf{x}$, averaged over decision-makers from π.

Following causal inference practice (Nogueira et al., 2022), we now investigate under which assumptions $\text{VOI}(\mathbf{x})$ can be retrieved from empirical data. We assume[1]: $(A1)$ **Unconfoundedness**, *i.e.,* conditional on baseline covariates $\mathbf{X}$, the disclosure decision is as good as random with respect to the potential regime-specific risks (Rosenbaum & Rubin, 1983); $(A2)$ **Positivity**, *i.e.,* for all covariate profiles in the population, we observe both disclosure regimes with non-zero probability (Rubin, 1978); $(A3)$ **Consistency**, *i.e.,* if a decision episode is observed under a specific regime, then the observed loss is equal to the corresponding potential loss (Cole & Frangakis, 2009); $(A4)$ **Stable Unit Treatment Value Assumption (SUTVA)**, *i.e.,* for each decision episode, the potential action of the human decision-maker depends only on the information disclosed to that decision-maker in that episode (Rubin, 1981) and $(A5)$ **Exchangeability of Human Decision-Makers**, *i.e.,* the decision-makers are drawn independently of the covariates and of the regime. Under assumptions *A1-A5*, we can estimate our causal estimand from data:

Proposition 1 (Identifiability of the Value of Information)**.** *Under Assumptions A1–A5, the regime-specific potential risks are identifiable from observed data. Let* $\ell^{\text{obs}}\big(A_{H_D}(D), Y\big) = (1 - D)\ell\big(A_{H_0}(0), Y\big) + D\ell\big(A_{H_1}(1), Y\big)$ *denote the loss observed in a decision episode, where* H_D *is the decision-maker assigned under regime* D*. Then, for* $d \in \{0, 1\}$*, the Value of Information*

$$\text{VOI}(\mathbf{x}) = r_0(\mathbf{x}) - r_1(\mathbf{x}) = \mathbb{E}\big[\ell\big(A_H(0), Y\big) - \ell\big(A_H(1), Y\big) \mid \mathbf{X} = \mathbf{x}\big]$$

is identifiable as

$$\text{VOI}(\mathbf{x}) = \mathbb{E}\big[\ell^{\text{obs}}\big(A_H(D), Y\big) \mid \mathbf{X} = \mathbf{x}, D = 0\big] - \mathbb{E}\big[\ell^{\text{obs}}\big(A_H(D), Y\big) \mid \mathbf{X} = \mathbf{x}, D = 1\big].$$

Proof. We provide the proof in Appendix A.2 □

Proposition 1 reduces the $\text{VOI}(\mathbf{x})$ to two identifiable, regime-specific conditional risks. Estimating it becomes a standard CATE estimation problem, for which any estimator from the causal inference literature can be used. The simplest one (T-learner Künzel et al. (2019)) fits a risk model per regime, $\hat{r}_0$ and $\hat{r}_1$, and sets $\widehat{\text{VOI}}(\mathbf{x}) = \hat{r}_0(\mathbf{x}) - \hat{r}_1(\mathbf{x})$. Still, accurate prediction of $\text{VOI}(\mathbf{x})$, is not sufficient for effective disclosure: the policy depends on the sign (Frauen et al., 2025) and ranking (Arno et al., 2026) of $\widehat{\text{VOI}}(\mathbf{x})$, and estimation errors can cause the policy to disclose information where it is harmful. In the next subsection, we quantify how such errors affect the learned policy's performance.

3.3 Guarantees for the plug-in policy

Since $\text{VOI}(\mathbf{x})$ is identifiable from data collected under the two regimes (Proposition 1), one can threshold an estimate $\widehat{\text{VOI}}(\mathbf{x})$ in its place, obtaining what we call a *plug-in* policy. Let us denote with $\hat{d}(\mathbf{x}) = \mathbb{1}\{\widehat{\text{VOI}}(\mathbf{x}) \geq \hat{\lambda}\}$ the plug-in disclosure policy, with threshold $\hat{\lambda} \geq 0$ for a given budget B, with $\hat{B} = \mathbb{E}_{\mathbf{x}}[\hat{d}(\mathbf{x})]$ its realized disclosure rate and with $\varepsilon(\mathbf{x}) = \widehat{\text{VOI}}(\mathbf{x}) - \text{VOI}(\mathbf{x})$ the estimation error of $\widehat{\text{VOI}}(\mathbf{x})$. Thresholding $\widehat{\text{VOI}}(\mathbf{x})$ raises two questions: whether the resulting policy can be worse than no-disclosure and how far it can fall short of the optimal policy.

[1] For the sake of space, we present these assumptions in detail in Appendix B.

First, we address whether deploying the plug-in policy can achieve higher risk than simply never disclosing. We stress that this is possible in principle because disclosure can be harmful on average ($\mathbb{E}_{\mathbf{x}}[\text{VOI}(\mathbf{x})] < 0$) and a policy that mis-ranks instances can spend its budget where disclosure is harmful. In the following, we quantify how large this degradation can be:

Theorem 2 (No-disclosure degradation)**.** *Given a non-negative loss function $\ell : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}^+$, for any estimator $\widehat{\text{VOI}}(\mathbf{x})$ such that $\|\varepsilon\|_{L^2(p)} < \infty$ and any $\hat{\lambda} \geq 0$, the following holds:*

$$R(\hat{d}) - R(0) \;\leq\; \sqrt{\hat{B}}\,\|\varepsilon\|_{L^2(p)} \;-\; \hat{\lambda}\hat{B}. \tag{5}$$

Proof. We provide the proof in Appendix A.3. □

The bound is governed by the $\text{VOI}(\mathbf{x})$ estimation error and by the region on which the policy intervenes: for bounded $\|\varepsilon\|_{L^2(p)}$, it converges to zero as the realized disclosure rate $\hat{B}$ tends to zero but it is positive whenever $\|\varepsilon\|_{L^2(p)} > \hat{\lambda}\sqrt{\hat{B}}$. Hence, the plug-in policy can perform worse than never disclosing, but only by a bounded amount (see Appendix A.3).

We now present our *Oracle regret bound*. This addresses how much is lost, relative to the optimal policy, by thresholding an estimate rather than the true $\text{VOI}(\mathbf{x})$. Here we additionally require the plug-in rule to use the threshold that Theorem 1 prescribes for $\widehat{\text{VOI}}(\mathbf{x})$, i.e., the rule that is optimal according to its own estimate of the risks:

Theorem 3 (Oracle regret)**.** *Given a non-negative loss function $\ell : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}^+$, let d^* be the optimal policy of Theorem 1 and $B^* = \mathbb{E}_{\mathbf{x}}[d^*(\mathbf{x})]$. Assume that the cumulative distribution of $\widehat{\text{VOI}}(\mathbf{x})$ is continuous, let $\hat{q}_{1-B}$ be its $(1-B)$-quantile and $\hat{\lambda} = \max\{0,\, \hat{q}_{1-B}\}$. Then, it holds that*

$$R(\hat{d}) - R(d^*) \;\leq\; \sqrt{B^* + \hat{B}}\,\|\varepsilon\|_{L^2(p)}. \tag{6}$$

Proof. We provide the proof in Appendix A.4. □

The Oracle-regret bound of Eq. (6) shows that the plug-in rule is $O(\|\varepsilon\|_{L^2(p)})$-competitive with an oracle knowing the true risks. Such a bound scales with an upper bound on the mass of the disagreement region, $\mathbb{E}_{\mathbf{x}}[|d^*(\mathbf{x}) - \hat{d}(\mathbf{x})|] \leq B^* + \hat{B}$. The bound can therefore be loose when the two policies agree away from the threshold: estimation errors on instances that both policies disclose, or both withhold, do not affect their relative performance.

3.4 Estimating the Value of Information for 0-1 loss

As shown in Theorem 1, learning disclosure policies requires estimating the regime-specific risks $r_0(\mathbf{x})$ and $r_1(\mathbf{x})$ and then thresholding their difference. In what follows, we focus on the 0-1 loss, which is widely used in human-AI collaboration to model the error incurred by the human decision maker (Madras et al., 2018; Ruggieri & Pugnana, 2025). The $0-1$ loss is defined as:

$$\ell(A(D), Y) = \mathbb{1}\{A(D) \neq Y\}, \tag{7}$$

Under this loss choice, the expected risk becomes:

$$r_D(\mathbf{x}) = \mathbb{E}[\mathbb{1}\{A(D) \neq Y\}], \qquad D \in \{0, 1\} \tag{8}$$

Thus, estimating Eq. (8) reduces to a standard classification task, where the goal is to predict whether the human makes mistakes, as a T-learner would do with a single scalar risk model per regime (Künzel et al., 2019). However, humans do not err uniformly across classes: some classes are systematically confused with others, and disclosure alters these error patterns differently for each class. A scalar risk model condenses all these patterns into a single error probability, blurring the class-specific effects on which the VoI depends (see Appendix E.1). We therefore decompose the regime-specific risk class-wise, relying on the following factorization of the $0-1$ risk:

Proposition 2 (Class-wise decomposition)**.** *Let us consider $\ell(A(D), Y) = \mathbb{1}\{A(D) \neq Y\}$. Then $r_D(\mathbf{x})$ can be written as:*

$$r_D(\mathbf{x}) = \sum_{y \in \mathcal{Y}} \underbrace{\mathbb{P}(Y = y \mid \mathbf{X} = \mathbf{x})}_{p_y}\, \underbrace{\mathbb{P}(A(D) \neq y \mid \mathbf{X} = \mathbf{x}, Y = y)}_{q_{D,y}}. \tag{9}$$

Proof. We provide the proof in Appendix A.5. □

This decomposition suggests a separate estimation strategy for both p_y and $q_{D,y}$. More precisely, we approximate p_y using a probabilistic classifier $f : \mathcal{X} \to \Delta^{|\mathcal{Y}|}$, and $q_{D,y}$ via a multi-head architecture with $|\mathcal{Y}|$ heads and a shared backbone, where any y-th head is a binary classifier $g_{D,y} : \mathcal{X} \to [0, 1]$. The resulting risk estimator is then $\widehat{r}_D(\mathbf{x}) = \sum_{y\in\mathcal{Y}} f_y(\mathbf{x}) \cdot g_{D,y}(\mathbf{x})$, where f_y refers to the $y-$th output of the probabilistic classifier. The estimated VoI is obtained by plugging $\widehat{r}_D(\mathbf{x})$ into Eq. (2). We refer to this estimator as `ClassWise`, which specializes the T-learner to a structured class-wise risk model.

4 Experimental Evaluation

In this section[2], we address the following research questions:

Q1 Does `ClassWise` learn effective policies?

Q2 Does the `ClassWise` policy improve human-AI team performance?

Q3 How do participants interact with the `ClassWise` policy?

4.1 Experimental settings

Datasets. We consider both synthetic and real data. We generate 10000 synthetic data samples using the standard `make_classification` function of `sklearn` library, considering both a binary ($|\mathcal{Y}| = 2$) and a multiclass task ($|\mathcal{Y}| = 4$). We emulate selective disclosure by treating a subset of features as baseline information and the remaining ones as support information. Then, we simulate human decisions in the two regimes with logistic regression models. We also employ two different real datasets: `Email` (Rebeka Toth, 2025) and `ImageNet-16H` (Steyvers et al., 2022). (i) `Email` contains 1000 emails that can be classified as either legitimate or fraudulent. For this dataset, we take human predictions from (Bogani et al., 2026), who provide for each email a human prediction made either without any assistance ($D = 0$) or with the support of an ML model ($D = 1$). (ii) `ImageNet-16H` contains 1200 noisy images that must be assigned to one of 16 categories. Human decisions for this dataset come from two behavioural experiments: the ones with no assistance come from (Steyvers et al., 2022) ($D = 0$), while predictions with support information come from (Straitouri & Rodriguez, 2024) ($D = 1$), where participants make their decision after observing a set of conformal predictions ($\alpha = .05$)[3]. We provide further details in Appendix C.

Methods. For **Q1**, we evaluate our approach (`ClassWise`) against the following baselines: (i) an uncertainty-based policy (`Confidence`) that thresholds $u(\mathbf{x}) = 1 - \max_y f(\mathbf{x})_y$ instead of $\widehat{\text{VOI}}(\mathbf{x})$, at the same $1 - B$ quantile (Pugnana et al., 2024); (ii) a policy that randomly assigns disclosure for each instance (`Random`); (iii) a full disclosure baseline (`FullDisc`); (iv) and a no-disclosure baseline (`NoDisc`). For **Q2** and **Q3**, we compare `ClassWise` against human participants who decide by themselves, item by item, whether to request the support information.

Metrics and Evaluation. For both **Q1** and **Q2** we evaluate human decision performance using test-set accuracy (Acc). For **Q3** we also measure participants' advice adherence (Adh), ***i.e.,*** the proportion of trials in which support information is available and the participant's classification matches such information (this amounts to selecting the ML-suggested label for `Email` and any class included in the prediction set for `ImageNet-16H`). We evaluate *Adherence* only on policy-selected items, ensuring that comparisons between MP and HP are based on the same set of items. User studies data are analyzed with logistic mixed-effects models with random intercepts for participants and items (see Appendix E.2 and Appendix E.3 for full analyses details).

Q1 setup. For **Q1**, (i) we perform a $70/10/20$ split of the data into training, calibration, and test sets; (ii) we carve out a validation split from the training data (10%) and select hyperparameters by grid search[4]; (iii) we calibrate the disclosure threshold as the $1 - B$ quantile of $\widehat{\text{VOI}}(\mathbf{x})$ on the

[2]Code is provided in the Supplementary Materials.

[3]A detailed analysis of the effect of the miscoverage level α is provided in Appendix E.4.

[4]Implementation details are provided in Appendix C (`Synth` Appendix C.1, `Email` Appendix C.2, `ImageNet-16H` Appendix C.3).

calibration set for each $B \in \{.1, .2, .3, .4, .5, .6, .7, .8, .9\}$ and (iv) we report the human accuracy on the test set, averaged over five seeds that vary model initialization[5].

User studies. To answer **Q2** and **Q3**, we run two user studies (545 total participants) in which each participant classifies 20 items sampled from the test set of one of the two real datasets, *i.e.,* `Email` or `ImageNet-16H`. We manipulate two between-subjects variables, with two conditions each: *(i) Support* is either *Human-Policy* (HP), where participants decide themselves when to request the support information, (i.e., the ML's model prediction, and specifically the predicted label for emails and the conformal prediction set at $\alpha = .05$ for images), or *Machine-Policy* (MP), where our `ClassWise` policy decides when to provide it; *(ii) Budget* is either *Low* (LB) or *High* (HB), making the support information available for at most 6 or 14 of the 20 items (30% and 70%). Participants in HP are told they need not exhaust their budget and are incentivized to request information only when needed, so that their policy is broadly comparable to the learned one. The 20-item samples are stratified so that $5/9$ emails and $6/14$ images come from the items the policy would select for disclosure under LB/HB, matching its realized disclosure rates in the original test sets, and all four conditions draw from the same pool[6]. Similarly to Scenario 1 in (Palomba et al., 2025), the ML model's prediction is retrospectively available for every case. We therefore report, as an exploratory analysis, a counterfactual benchmark (MPCFT), obtained by (i) replacing MP participants' responses with the underlying ML model prediction whenever information is disclosed and (ii) retaining their observed responses otherwise. Thus, MPCFT represents the accuracy that would have been achieved had the ML prediction been enforced as the final decision whenever it was disclosed. See Appendix D for further details on the user studies.

4.2 Experimental Results

Q1: `ClassWise` is effective on all datasets and outperforms simple baselines. Fig. 2 reports the results for our experiments on all four datasets. For `Synth` we can see that `ClassWise` dominates the `Confidence` baseline and `Random` at every budget level. Interestingly, `ClassWise` already exceeds `FullDisc` accuracy at $B = .30$ in the `SynthBin` ($Acc \approx 0.83$) and at $B = .50$ in the `SynthMulti` ($Acc \approx 0.63$), suggesting that disclosing information for only a small fraction of instances suffices to recover the performance of full disclosure.

For `Email`, disclosure is harmful on average: `NoDisc` outperforms `FullDisc` ($Acc \approx .79$ vs $Acc \approx .76$). `ClassWise` is the best-performing policy for all B, peaking at $B = .30$ ($Acc \approx .81$). Beyond this point, the policy increasingly discloses on instances where it wrongly estimates support to be beneficial, so accuracy decreases and plateaus slightly below `NoDisc`, which however remains within its confidence band. As anticipated by Theorem 2, the plug-in policy can thus perform worse than never disclosing, although only by a bounded amount. Unlike `ClassWise`, the `Confidence` baseline never abstains: its accuracy decreases monotonically with the budget, never exceeds `NoDisc`, and collapses into `FullDisc` at $B = 1$. The reason is that uncertainty measures how hard the *case* is, not how much disclosure *helps*: there is no mechanism to abstain when disclosure hurts.

For `ImageNet-16H`, we observe that on average disclosing support information is helpful: `FullDisc` achieves $Acc \approx .85$ while `NoDisc` $Acc \approx .77$. Interestingly, `ClassWise` surpasses the full disclosure baseline already at $B = .70$ and achieves the highest accuracy of $\approx .86$ at $B = .90$. Moreover, we observe that `ClassWise` is always different from `Random`, suggesting that the policy can learn when support information is useful. `Confidence` is a stronger baseline here, but it only matches `ClassWise` at $B \geq .80$, where nearly every instance is disclosed and little is left to choose. Whenever the budget is binding, `ClassWise` is always ahead.

Moreover, we ablate the risk estimator considering two variants that do not exploit the classwise decomposition of Proposition 2, i.e., (a) an approach that directly estimates both r_0 and r_1 separately (`T-Learner`) and (b) a joint estimator for both r_0 and r_1 (`S-Learner`). `ClassWise` outperforms both variants at every budget on `Synth` data, and attains the highest peak accuracy on the real datasets. The gap is largest on `Email`, where disclosure is harmful: `S-Learner` shrinks $\widehat{\mathrm{VOI}}(\mathbf{x})$ towards zero, estimates a positive VoI almost everywhere and thus never abstains, collaps-

[5] We do not resample the split across seeds, as the user studies require a single, fixed test set.

[6] Realized rates need not match B, since the threshold is calibrated on a held-out split and the policy never discloses when $\widehat{\mathrm{VOI}}(\mathbf{x}) \leq 0$; they also vary across runs, more so on `Email` (see Appendix D).

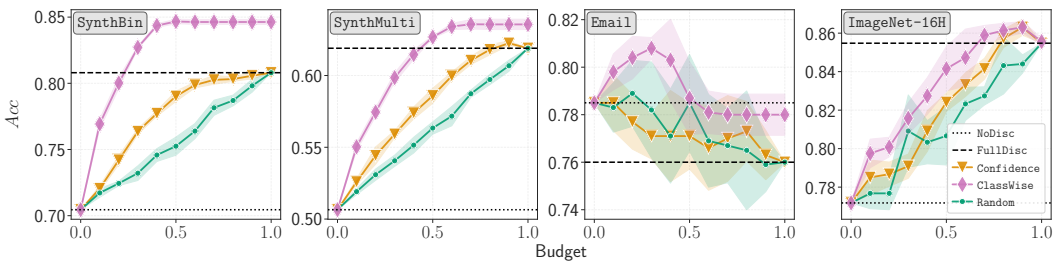
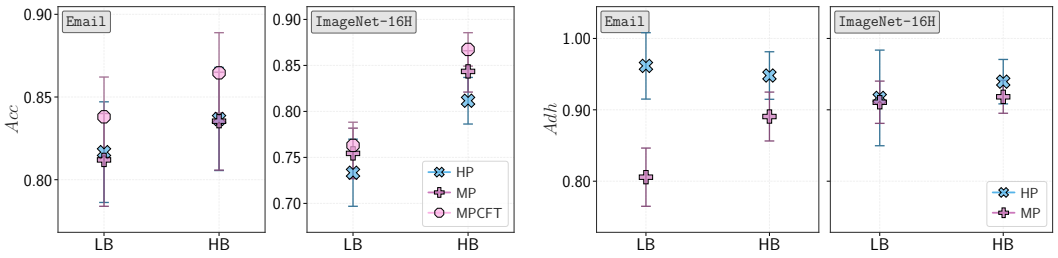


Figure 2: Budget-accuracy curves for the four considered datasets. We report average mean values and 95% confidence intervals over five seeds.



(a) Accuracy (b) Adherence

Figure 3: User studies results (**Q2** and **Q3**). (a) Participants' accuracy (Acc) and (b) adherence (Adh) on the AI advice, by *Budget* (LB/HB) and *Support* (HP/MP/MPCFT) condition, on `Email` (left) and `ImageNet-16H` (right). Error bars are 95% confidence intervals across participants.

ing to `FullDisc` at high budgets, whereas `ClassWise` withholds support information on a larger fraction of the instances than `T-Learner`. Full ablation and budget analyses are in Appendix E.1.

Q2: The learned policy matches or outperforms self-selection. We report the two user-study results on human accuracy (Acc) in Fig. 3a. When considering *Support*, in the `ImageNet-16H` task accuracy is significantly higher in MP than in HP (✚: 0.80 ± 0.11 vs ✖: 0.77 ± 0.13, $p = .025$). Instead, in the `Email` task MP and HP are not significantly different (✚: 0.82 ± 0.12 vs ✖: 0.83 ± 0.13, $p = .762$). When considering *Budget*, in the `ImageNet-16H` task accuracy is significantly higher in HB than in LB (0.83 ± 0.10 vs $0.74 \pm 0.13, p < .001$), whereas the two conditions do not significantly differ in the `Email` task (0.84 ± 0.12 vs $0.81 \pm 0.12, p = .056$). The advice-adherence results reported next may help account for why MP outperforms HP in the `ImageNet-16H` but not in the `Email` task. Appendix E.2 reports full results, plus analyses of advice correctness, human confidence, and alignment between `ClassWise` and human disclosure.

Q3: The learned policy is more beneficial whenever humans follow the advice. We report results on advice-adherence (Adh) in Fig. 3b. When considering *Support*, in the `ImageNet-16H` task adherence is not significantly different between MP and HP (✚: 0.91 ± 0.11 vs ✖: 0.93 ± 0.18, $p = .084$), while in the `Email` task it is significantly lower in MP than in HP (✚: 0.85 ± 0.16 vs ✖: 0.96 ± 0.15, $p < .001$). When considering *Budget*, in both studies adherence does not differ significantly between HB and LB (`ImageNet-16H`: $p = .260$; `Email`: $p = .636$), not even in interaction with *Support* (`ImageNet-16H`: $p = .392$; `Email`: $p = .126$).

Accuracy in MPCFT (Fig. 3a) suggests that lower advice-adherence may attenuate the benefits of selective disclosure. In the `ImageNet-16H` task, where adherence under MP is already high, accuracy in MPCFT is descriptively but not significantly higher than in MP (●: 0.82 ± 0.10 vs. ✚: 0.80 ± 0.11; $p = .058$), while it is significantly higher than in HP (✖: 0.77 ± 0.13; $p < .001$). In `Email`, where adherence under MP is lower, MPCFT accuracy is significantly higher than in MP (●: 0.85 ± 0.10 vs. ✚: 0.82 ± 0.12; $p = .004$) and descriptively, but not significantly, higher than in HP (✖: 0.83 ± 0.13; $p = .196$). This cross-task pattern suggests that lower adherence may partly contribute to the absence of an MP advantage in the `Email` task, as would be expected given that the disclosed ML advice is correct on most trials. Additional results are reported in Appendix E.3.

Overall, these results suggest that deploying selective disclosure requires attending not only to which information is valuable, but also to how decision-makers receive information they did not request. Indeed, the pattern we observe is consistent with psychological evidence suggesting that self-produced outcomes are evaluated more favorably than comparable outcomes produced by external sources (Botti et al., 2023; Enisman et al., 2021).

5 Related Work

Our paper builds upon related work on *human-AI decision-making* and *policy learning under budget constraints*. We provide an extended overview of these and related topics in Appendix F.

Human-AI Decision-Making. Learning to Defer (Madras et al., 2018) and follow-up work (Mozannar & Sontag, 2020; Verma & Nalisnick, 2022) allocate each instance to the model or to the human, and causal reasoning has so far been used to evaluate such systems (Palomba et al., 2025); in LSD the human always decides, and the policy allocates *information*. Other work designs the form of support, e.g., conformal prediction sets (De Toni et al., 2024) or explanations (Schemmer et al., 2023), while Noorani et al. (2026) let the AI refine a human expert's prediction set, recovering labels the human missed without degrading correct human judgments. Noti & Chen (2023) and Ma et al. (2023) are closely related to our work, as they show support when the AI is predicted to outperform the human. Similarly, Bhatt et al. (2025) learn personalized disclosure policies online. We differ by casting selective disclosure as a causal inference problem: disclosure is an intervention on the decision-maker's information, and the VoI is never observed under both regimes. To the best of our knowledge, LSD is the first to establish when VoI is identifiable from data collected under the two regimes (Section 3.2) and to show that VoI is the required estimand to build budget-constrained disclosure policies with guarantees (Theorems 2 and 3).

Policy Learning. Theorem 1 is the disclosure analogue of optimal policy learning (Kitagawa & Tetenov, 2018) under a capacity constraint, where the optimal rule thresholds the conditional average treatment effect at the larger of zero and a budget-dependent quantile (Bhattacharya & Dupas, 2012; Cerulli, 2026). The difference is where the treatment acts: disclosure is applied to the decision-maker, while the outcome is the correctness of a decision about a case, so SUTVA is stated over decision episodes and identification requires decision-makers to be exchangeable across regimes.

6 Conclusion

In this work, we introduced Learning to Selectively Disclose (LSD), a decision-theoretic framework for deciding when to provide support information to a human decision-maker under a budget constraint. We showed that the optimal policy thresholds the Value of Information, revealing information only where the expected benefit is positive and exceeds a budget-dependent threshold. Since the VoI is not directly observable, we gave conditions under which it is identifiable from data collected under the two regimes. We then studied plug-in approaches, for which we bounded the estimation error relative to no disclosure and the regret relative to the optimal policy. For classification tasks under the $0-1$ loss, we further decomposed the human risk class-wise and used this decomposition to build a VoI estimator, which proved effective on both synthetic and real benchmarks. Two user studies further indicate that `ClassWise`-selected disclosure can improve human-AI team accuracy relative to human-selected disclosure, although its effectiveness varies across tasks. This variability may partly reflect how participants respond to advice: in some contexts, adherence may be lower for automatically provided than for comparable self-requested information, and a counterfactual benchmark replacing participants' predictions with ML ones when disclosed suggests that greater adherence would have yielded higher overall accuracy under `ClassWise`-selected disclosure.

Limitations and Future Works. Our framework treats the support information S as an indivisible block: the policy decides *whether* to disclose, never *what* to disclose. Studying which parts of the information to reveal is left for future work. Also, our user studies do not identify which task characteristics reduce participants' advice-adherence nor establish a causal relationship between adherence and policy effectiveness, as they were not designed to address these questions. Future work will investigate these aspects with dedicated user studies.

AI USE STATEMENT

In this work, we used generative AI tools for code support and polishing the writing. Regarding the former, the code was manually reviewed by two authors. Regarding the latter, we have used generative AI tools to provide feedback on the presentation's quality, with the goal of avoiding concerns that are due solely to inaccuracies in our writing. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The research protocol of our user studies was determined as posing no risk to participants' well-being or rights and therefore as not requiring full ethical review by the Research Ethics Committee of [institution omitted to preserve anonymity] ([protocol number omitted to preserve anonymity]).

REPRODUCIBILITY STATEMENT

The main text describes the problem formulation, the assumptions for identifying the VoI, the `ClassWise` estimator, the baselines, the datasets, the evaluation protocol (data splits, threshold calibration, budget grid, and seeds), and the design and analysis of the user studies. Complete proofs are provided in the appendix, together with additional experimental results and details on the user studies design. The implementation and experiment scripts used to produce our results are made available through an anonymized repository, which also contains the data collected in the user studies and the corresponding statistical analysis scripts.

REFERENCES

Henri Arno, Dennis Frauen, Emil Javurek, Thomas Demeester, and Stefan Feuerriegel. Rank-learner: Orthogonal ranking of treatment effects. *CoRR*, abs/2602.03517, 2026.

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1): 133–161, 2021. doi: https://doi.org/10.3982/ECTA15732. URL `https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA15732`.

Umang Bhatt, Valerie Chen, Katherine M. Collins, Parameswaran Kamalaruban, Emma Kallina, Adrian Weller, and Ameet Talwalkar. Learning personalized decision support policies. In *AAAI*, pp. 14203–14211. AAAI Press, 2025.

Debopam Bhattacharya and Pascaline Dupas. Inferring welfare maximizing treatment assignment under budget constraints. *Journal of Econometrics*, 167(1):168–196, 2012. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2011.11.007. URL `https://www.sciencedirect.com/science/article/pii/S0304407611002697`.

Alessandro Bogani, Nicola Debole, Emanuele Marconato, Andrea Pugnana, Katya Tentori, and Andrea Passerini. Are concept bottleneck models effective as decision-support systems?, 2026. URL `https://arxiv.org/abs/2608.25581`.

Elizabeth Bondi, Raphael Koster, Hannah Sheahan, Martin J. Chadwick, Yoram Bachrach, A. Taylan Cemgil, Ulrich Paquet, and Krishnamurthy Dvijotham. Role of human-ai interaction in selective prediction. In *AAAI*, pp. 5286–5294. AAAI Press, 2022.

Simona Botti, Sheena S Iyengar, and Ann L McGill. Choice freedom. *Journal of Consumer Psychology*, 33(1):143–166, 2023.

Violet A Brown. An introduction to linear mixed-effects modeling in r. *Advances in Methods and Practices in Psychological Science*, 4(1):2515245920960351, 2021.

Zana Buçinca, Siddharth Swaroop, Amanda E. Paluch, Susan A. Murphy, and Krzysztof Z. Gajos. Towards optimizing human-centric objectives in ai-assisted decision-making with offline reinforcement learning. *CoRR*, abs/2403.05911, 2024.

Yuzhou Cao, Hussein Mozannar, Lei Feng, Hongxin Wei, and Bo An. In defense of softmax parametrization for calibrated and consistent learning to defer. In *NeurIPS*, 2023.

Aldo Gael Carranza and Susan Athey. Robust offline policy learning with observational data from multiple sources. In *AISTATS*, volume 258 of *Proceedings of Machine Learning Research*, pp. 4897–4905. PMLR, 2025.

Giovanni Cerulli. Optimal policy learning with observational data in multi-action scenarios: estimation, risk preference, and potential failures. *Int. J. Data Sci. Anal.*, 22(1):164, 2026.

Mohammad-Amin Charusaie, Amirmehdi Jafari Fesharaki, and Samira Samadi. Defer-and-fusion: Optimal predictors that incorporate human decisions. In *5th Workshop on practical ML for limited/low resource settings*, 2024. URL https://openreview.net/forum?id=Pg51u5YboV.

Stephen R Cole and Constantine E Frangakis. The consistency statement in causal inference: a definition or an assumption? *Epidemiology*, 20(1):3–5, 2009.

Jesse C. Cresswell, Yi Sui, Bhargava Kumar, and Noël Vouitsis. Conformal prediction sets improve human decision making. In *ICML*, volume 235 of *Proceedings of Machine Learning Research*, pp. 9439–9457. PMLR / OpenReview.net, 2024.

Giovanni De Toni, Nastaran Okati, Suhas Thejaswi, Eleni Straitouri, and Manuel Gomez Rodriguez. Towards human-ai complementarity with prediction sets. In *NeurIPS*, 2024.

Maya Enisman, Hila Shpitzer, and Tali Kleiman. Choice changes preferences, not merely reflects them: A meta-analysis of the artifact-free free-choice paradigm. *Journal of Personality and Social Psychology*, 120(1):16, 2021.

Yizirui Fang and Eric T. Nalisnick. Learning to defer with an uncertain rejector via conformal prediction. *Trans. Mach. Learn. Res.*, 2026, 2026.

Carlos Fernández-Loría and Foster Provost. Causal decision making and causal effect estimation are not the same…and why it matters. *INFORMS Journal on Data Science*, 1(1):4–16, 2022. doi: 10.1287/ijds.2021.0006. URL https://doi.org/10.1287/ijds.2021.0006.

Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Mihaela van der Schaar, and Stefan Feuerriegel. Treatment effect estimation for optimal decision-making. In *NeurIPS*, 2025.

Ruijiang Gao and Mingzhang Yin. Confounding-robust deferral policy learning. In *AAAI*, pp. 14238–14246. AAAI Press, 2025.

Wenbo Gong, Sebastian Tschiatschek, Sebastian Nowozin, Richard E. Turner, José Miguel Hernández-Lobato, and Cheng Zhang. Icebreaker: Element-wise efficient information acquisition with a bayesian deep latent gaussian model. In *NeurIPS*, pp. 14791–14802, 2019.

Peter Green and Catriona J MacLeod. Simr: An r package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, 7(4):493–498, 2016.

Patrick Hemmer, Monika Westphal, Max Schemmer, Sebastian Vetter, Michael Vössing, and Gerhard Satzger. Human-ai collaboration: The effect of AI delegation on human task performance and task satisfaction. In *IUI*, pp. 453–463. ACM, 2023.

Patrick Hemmer, Max Schemmer, Niklas Kühl, Michael Vössing, and Gerhard Satzger. Complementarity in human-ai collaboration: concept, sources, and evidence. *Eur. J. Inf. Syst.*, 34(6): 979–1002, 2025.

Kilian Hendrickx, Wannes Meert, Bram Cornelis, and Jesse Davis. Know your limits: Machine learning with rejection for vehicle engineering. In *ADMA*, volume 13087 of *Lecture Notes in Computer Science*, pp. 273–288. Springer, 2021.

Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. Measures for explainable AI: explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance. *Frontiers Comput. Sci.*, 5, 2023. doi: 10.3389/FCOMP.2023.1096257. URL https://doi.org/10.3389/fcomp.2023.1096257.

Ronald A. Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26, 1966. doi: 10.1109/TSSC.1966.300074.

Jaromír Janisch, Tomás Pevný, and Viliam Lisý. Classification with costly features as a sequential decision-making problem. *Mach. Learn.*, 109(8):1587–1615, 2020.

Shihao Ji and Lawrence Carin. Cost-sensitive feature acquisition and classification. *Pattern Recognition*, 40(5):1474–1485, 2007.

Veeru Kasivisvanathan, Antti S Rannikko, Marcelo Borghi, Valeria Panebianco, Lance A Mynderse, Markku H Vaarala, Alberto Briganti, Lars Budäus, Giles Hellawell, Richard G Hindley, et al. MRI-Targeted or Standard Biopsy for Prostate-Cancer Diagnosis. *New England Journal of Medicine*, 378(19):1767–1777, 2018.

Edward H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008 – 3049, 2023. doi: 10.1214/23-EJS2157. URL https://doi.org/10.1214/23-EJS2157.

Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

Benjamin Kompa, Jasper Snoek, and Andrew L. Beam. Second opinion needed: communicating uncertainty in medical machine learning. *npj Digit. Medicine*, 4, 2021.

Levi Kumle, Melissa L-H Võ, and Dejan Draschkow. Estimating power in (generalized) linear mixed models: An open introduction and tutorial in r. *Behavior research methods*, 53(6):2528–2543, 2021.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

James Law. Review of "algorithmic learning in a random world by vovk, gammerman and shafer", springer, 2005, ISBN: 0-387-00152-2. *SIGACT News*, 37(4):38–40, 2006.

Jing Li, Yaoyao Cui, Juan Jiang, Jianqiu Yu, Lili Niu, Jiabo Deng, Fujun Shen, Liang Zhang, Bisong Yue, and Jing Li. Applying dna barcoding to conservation practice: a case study of endangered birds and large mammals in china. *Biodiversity and Conservation*, 26(3):653–668, 2017.

Shuqi Liu, Yuzhou Cao, Qiaozhen Zhang, Lei Feng, and Bo An. Mitigating underfitting in learning to defer with consistent losses. In *AISTATS*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4816–4824. PMLR, 2024.

Shuqi Liu, Yuzhou Cao, Lei Feng, Bo An, and Luke Ong. When more experts hurt: Underfitting in multi-expert learning to defer. *CoRR*, abs/2602.17144, 2026.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *CoRR*, abs/2201.03545, 2022.

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, José Miguel Hernández-Lobato, Sebastian Nowozin, and Cheng Zhang. EDDI: efficient dynamic discovery of high-value information with partial VAE. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4234–4243. PMLR, 2019.

Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. Who should i trust: Ai or myself? leveraging human and ai correctness likelihood to promote appropriate trust in ai-assisted decision-making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581058. URL https://doi.org/10.1145/3544548.3581058.

David Madras, Toniann Pitassi, and Richard S. Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *NeurIPS*, pp. 6150–6160, 2018.

Charles F. Manski. Statistical treatment rules for heterogeneous populations. *Econometrica*, 72 (4):1221–1246, 2004. doi: https://doi.org/10.1111/j.1468-0262.2004.00530.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0262.2004.00530.x.

Anqi Mao, Christopher Mohri, Mehryar Mohri, and Yutao Zhong. Two-stage learning to defer with multiple experts. In *NeurIPS*, 2023.

Anqi Mao, Mehryar Mohri, and Yutao Zhong. Mastering multiple-expert routing: Realizable h-consistency and strong guarantees for learning to defer. In *ICML*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025.

Yannis Montreuil, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. Adversarial robustness in two-stage learning-to-defer: Algorithms and guarantees. In *ICML*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025a.

Yannis Montreuil, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. Why ask one when you can ask k? two-stage learning-to-defer to the top-k experts. *CoRR*, abs/2504.12988, 2025b.

Yannis Montreuil, Yeo Shu Heng, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. A two-stage learning-to-defer approach for multi-task learning. In *ICML*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025c.

Yannis Montreuil, Shu Heng Yeo, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. Optimal query allocation in extractive qa with llms: A learning-to-defer framework with theoretical guarantees, 2026. URL https://arxiv.org/abs/2410.15761.

Hussein Mozannar and David A. Sontag. Consistent estimators for learning to defer to an expert. In *ICML*, Proceedings of Machine Learning Research, pp. 7076–7087. PMLR, 2020.

Hussein Mozannar, Hunter Lang, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David A. Sontag. Who should predict? exact algorithms for learning to defer to humans. In *AISTATS*, volume 206 of *Proceedings of Machine Learning Research*, pp. 10520–10545. PMLR, 2023.

X Nie and S Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108 (2):299–319, 06 2021. ISSN 0006-3444. doi: 10.1093/biomet/asaa076. URL https://doi.org/10.1093/biomet/asaa076.

Ana Rita Nogueira, Andrea Pugnana, Salvatore Ruggieri, Dino Pedreschi, and João Gama. Methods and tools for causal discovery and causal inference. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 12(2):e1449, 2022.

Sima Noorani, Shayan Kiyani, George J. Pappas, and Hamed Hassani. Human-AI collaborative uncertainty quantification. In *Trustworthy AI for Good (AI4GOOD) Workshop @ ICML 2026*, 2026. URL https://openreview.net/forum?id=AmZOI1AY1t.

Gali Noti and Yiling Chen. Learning when to advise human decision makers. In *IJCAI*, pp. 3038–3048. ijcai.org, 2023.

Nastaran Okati, Abir De, and Manuel Gomez-Rodriguez. Differentiable learning under triage. In *NeurIPS*, pp. 9140–9151, 2021.

Filippo Palomba, Andrea Pugnana, José M. Álvarez, and Salvatore Ruggieri. A causal framework for evaluating deferring systems. In *AISTATS*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2143–2151. PMLR, 2025.

Dario Pesenti, Alessandro Bogani, Stefano Teso, and Andrea Pugnana. Too much of the same: From algorithmic to human bias in learning to defer, 2026. URL https://arxiv.org/abs/2608.28050.

Andrea Pugnana, Lorenzo Perini, Jesse Davis, and Salvatore Ruggieri. Deep neural network benchmarks for selective classification. *J. Data-centric Mach. Learn. Res.*, 1:(17):1–58, 2024.

Andrea Pugnana, Giovanni De Toni, Cesare Barbera, Roberto Pellungrini, Bruno Lepri, and Andrea Passerini. To ask or not to ask: Learning to require human feedback. *CoRR*, abs/2510.08314, 2025.

Clara Punzi, Roberto Pellungrini, Mattia Setzu, Fosca Giannotti, and Dino Pedreschi. Ai, meet human: Learning paradigms for hybrid decision making systems. *CoRR*, abs/2402.06287, 2024.

Arman Rahbar, Linus Aronsson, and Morteza Haghir Chehreghani. A survey on active feature acquisition strategies. *CoRR*, abs/2502.11067, 2025.

Tamas Bisztray Rebeka Toth, Nils Gruschka. Constructing and benchmarking: a labeled email dataset for text-based phishing and spam detection framework. *CoRR*, abs/2511.21448, 2025. Withdrawn.

Giuseppe Romeo and Daniela Conti. Exploring automation bias in human-ai collaboration: a review and implications for explainable AI. *AI Soc.*, 41(1):259–278, 2026.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Donald B. Rubin. Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics*, 6(1):34 – 58, 1978. doi: 10.1214/aos/1176344064. URL `https://doi.org/10.1214/aos/1176344064`.

Donald B Rubin. Estimation in parallel randomized experiments. *Journal of Educational Statistics*, 6(4):377–401, 1981.

Salvatore Ruggieri and Andrea Pugnana. Things machine learning models know that they don't know. In *AAAI*, pp. 28684–28693. AAAI Press, 2025.

Maytal Saar-Tsechansky, Prem Melville, and Foster Provost. Active feature-value acquisition. *Management Science*, 55(4):664–684, 2009.

Mauricio Sadinle, Jing Lei, and Larry A. Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *CoRR*, abs/1609.00451, 2016.

Max Schemmer, Niklas Kühl, Carina Benz, Andrea Bartos, and Gerhard Satzger. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *IUI*, pp. 410–422. ACM, 2023.

Uri Shalit, Fredrik D. Johansson, and David A. Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3076–3085. PMLR, 2017.

Claudia Shi, David M. Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In *NeurIPS*, pp. 2503–2513, 2019.

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.

Luca Stradiotti, Dario Pesenti, Stefano Teso, and Jesse Davis. Learning to reject low-quality explanations via user feedback. *CoRR*, abs/2507.12900, 2025.

Eleni Straitouri and Manuel Gomez Rodriguez. Designing decision support systems using counterfactual prediction sets. In *ICML*, Proceedings of Machine Learning Research, pp. 46722–46744. PMLR / OpenReview.net, 2024.

Joshua Strong, Qianhui Men, and J. Alison Noble. Trustworthy and practical AI for healthcare: A guided deferral system with large language models. In *AAAI*, pp. 28413–28421. AAAI Press, 2025.

Rajeev Verma and Eric T. Nalisnick. Calibrated learning to defer with one-vs-all classifiers. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pp. 22184–22202. PMLR, 2022.

Rajeev Verma, Daniel Barrejón, and Eric T. Nalisnick. Learning to defer to multiple experts: Consistent surrogate losses, confidence calibration, and conformal ensembles. In *AISTATS*, volume 206 of *Proceedings of Machine Learning Research*, pp. 11415–11434. PMLR, 2023.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. doi: 10.1080/01621459.2017.1319839. URL https://doi.org/10.1080/01621459.2017.1319839.

Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans. In *IJCAI*, pp. 1526–1533, 2020.

Yan Xiong and Liyan Yang. Secret and overt information acquisition in financial markets. *The Review of Financial Studies*, 36(9):3643–3692, 2023.

Zheng Zhang, Cuong C. Nguyen, Kevin Wells, Thanh-Toan Do, David Rosewarne, and Gustavo Carneiro. Coverage-constrained human-ai cooperation with multiple experts. In *AAAI*, pp. 18055–18062. AAAI Press, 2026.

A Proofs

A.1 Proof of Optimal Disclosure Policy

For completeness, we restate here Theorem 1 from Section 3.1, and then provide its proof.

Theorem 1 (Optimal Disclosure Policy)**.** *Assume that the cumulative distribution of* $\mathrm{VoI}(\mathbf{x})$ *is continuous. Let* q_{1-B} *denote the* $(1-B)$*-quantile of* $\mathrm{VoI}(\mathbf{x})$ *and let* $\lambda^* = \max\{0,\, q_{1-B}\}$*. For any budget* $B \in [0,1]$*, the optimal disclosure policy is the threshold rule:*

$$d^*(\mathbf{x}) = \begin{cases} 1 & \textit{if} \quad \mathrm{VoI}(\mathbf{x}) \geq \lambda^*, \\ 0 & \textit{otherwise}. \end{cases} \tag{10}$$

Proof. We begin by first expanding the objective,

$$R(d) = \mathbb{E}\big[(1 - d(\mathbf{x}))r_0(\mathbf{x}) + d(\mathbf{x})r_1(\mathbf{x})\big] = \mathbb{E}[r_0(\mathbf{x})] - \mathbb{E}\big[d(\mathbf{x})\,\mathrm{VoI}(\mathbf{x})\big],$$

and since $\mathbb{E}[r_0(\mathbf{x})]$ does not depend on d, the original minimization problem is equivalent to

$$\max_{d\colon \mathcal{X}\to[0,1]} \ \mathbb{E}_{\mathbf{x}}\big[d(\mathbf{x})\,\mathrm{VoI}(\mathbf{x})\big] \quad \text{s.t.} \quad \mathbb{E}_{\mathbf{x}}[d(\mathbf{x})] \leq B. \tag{11}$$

Introducing the multiplier $\lambda \geq 0$ for the budget constraint, the Lagrangian is

$$\mathcal{L}(d,\lambda) = \mathbb{E}\big[d(\mathbf{x})\,\mathrm{VoI}(\mathbf{x})\big] - \lambda\big(\mathbb{E}[d(\mathbf{x})] - B\big) = \lambda B + \mathbb{E}\big[d(\mathbf{x})\,(\mathrm{VoI}(\mathbf{x}) - \lambda)\big].$$

For any fixed $\lambda \geq 0$, the maximisation is point-wise in $\mathbf{x}$ and the integrand is linear in $d(\mathbf{x}) \in [0,1]$. Therefore, any maximiser satisfies

$$d_\lambda(\mathbf{x}) = \begin{cases} 1, & \mathrm{VoI}(\mathbf{x}) > \lambda, \\ 0, & \mathrm{VoI}(\mathbf{x}) < \lambda, \end{cases} \qquad d_\lambda(\mathbf{x}) \in [0,1] \text{ arbitrary on } \{\mathrm{VoI}(\mathbf{x}) = \lambda\}. \tag{12}$$

It remains to select λ^* satisfying *complementary slackness*, $\lambda^*\big(\mathbb{E}[d_{\lambda^*}(\mathbf{x})] - B\big) = 0$. This allows for both binding and non-binding budget scenarios, leading to two mutually exclusive cases.

(i) Non-binding budget. Suppose $\mathbb{P}(\mathrm{VoI}(\mathbf{x}) > 0) < B$, so that $q_{1-B} \leq 0$ and hence $\lambda^* = 0$. Then $\lambda^* = 0$ is admissible: the policy $d^*(\mathbf{x}) = \mathbf{1}\{\mathrm{VoI}(\mathbf{x}) > 0\}$ satisfies $\mathbb{E}[d^*(\mathbf{x})] = \mathbb{P}(\mathrm{VoI}(\mathbf{x}) > 0) < B$, so the constraint is slack and complementary slackness holds with $\lambda^* = 0$. By the continuity assumption at $\lambda^* = 0$ we have $\mathbb{P}(\mathrm{VoI}(\mathbf{x}) = 0) = 0$, so $d^*(\mathbf{x})$ agrees p-almost everywhere with $\mathbf{1}\{\mathrm{VoI}(\mathbf{x}) \geq 0\}$ and is in particular feasible. Note that in this case no feasible policy can spend the full budget profitably: increasing $\mathbb{E}[d(\mathbf{x})]$ beyond $\mathbb{P}(\mathrm{VoI}(\mathbf{x}) > 0)$ requires disclosing on $\{\mathrm{VoI}(\mathbf{x}) \leq 0\}$, which cannot increase the objective. In particular the equation $\mathbb{E}[d_\lambda] = B$ admits *no* solution.

(ii) Binding budget. Suppose $\mathbb{P}(\text{VoI}(\mathbf{x}) > 0) \geq B$; then $F_{\text{VoI}(\mathbf{x})}(0) \leq 1 - B$, so $q_{1-B} \geq 0$ and $\lambda^* = q_{1-B}$. By the continuity assumption at λ^*, $\mathbb{P}(\text{VoI}(\mathbf{x}) = \lambda^*) = 0$ and $F_{\text{VoI}(\mathbf{x})}(\lambda^*) = 1 - B$, whence $\mathbb{P}(\text{VoI}(\mathbf{x}) > \lambda^*) = B$. The policy $d_{\lambda^*} = \mathbf{1}\{\text{VoI}(\mathbf{x}) \geq \lambda^*\}$ therefore satisfies $\mathbb{E}[d_{\lambda^*}] = B$, and complementary slackness holds with $\lambda^* \geq 0$.

Combining the two cases, the optimal multiplier is

$$\lambda^* = \max\left\{0,\ F^{-1}_{\text{VoI}(\mathbf{x})}(1 - B)\right\}, \tag{13}$$

and, since $\mathbb{P}(\text{VoI}(\mathbf{x}) = \lambda^*) = 0$ by the continuity assumption, any maximiser of Eq. (11) agrees p-almost everywhere with

$$d^*(\mathbf{x}) = \mathbf{1}\{\text{VoI}(\mathbf{x}) \geq \lambda^*\}, \tag{14}$$

which is the statement of Theorem 1. □

A.2 Proof of Identifiability of the Value of Information

For completeness, we restate here Proposition 1 from Section 3.2, and then provide its proof.

Proposition 1 (Identifiability of the Value of Information)**.** *Let $H \sim \pi$ denote a decision-maker drawn from the pool $\mathcal{H}$, let $A_H(0)$ and $A_H(1)$ denote the potential actions and $\ell\big(A_H(D), Y\big)$ the corresponding potential losses for $D \in \{0, 1\}$. Under assumptions A1–A5, the regime-specific potential risks are identifiable from observed data. In particular, the Value of Information*

$$\text{VoI}(\mathbf{x}) = r_0(\mathbf{x}) - r_1(\mathbf{x}) = \mathbb{E}\big[\ell\big(A_H(0), Y\big) - \ell\big(A_H(1), Y\big) \mid \mathbf{X} = \mathbf{x}\big]$$

is identifiable from observed data as

$$\text{VoI}(\mathbf{x}) = \mathbb{E}\big[\ell^{\text{obs}}\big(A_H(D), Y\big) \mid \mathbf{X} = \mathbf{x}, D = 0\big] - \mathbb{E}\big[\ell^{\text{obs}}\big(A_H(D), Y\big) \mid \mathbf{X} = \mathbf{x}, D = 1\big]$$

Proof. Fix $\mathbf{x} \in \mathcal{X}$ and a regime $d \in \{0, 1\}$. To distinguish the random regime D from the value it takes, throughout this proof we write $r_d(\mathbf{x})$ for the risk $r_D(\mathbf{x})$ of Eq. (1) evaluated at $D = d$. Recall from Section 2 that, under regime d, the decision on a case is taken by the decision-maker $H_d \sim \pi$ assigned to that regime. By SUTVA (A4), the action of H_d depends only on the information disclosed to it in that episode, so its potential action $A_{H_d}(d)$ is well defined. By consistency (A3), on the event $\{D = d\}$ the observed loss equals the corresponding potential loss, $\ell^{\text{obs}}\big(A_H(D), Y\big) = \ell\big(A_{H_d}(d), Y\big)$. By overlap (A2), $\mathbb{P}(D = d \mid \mathbf{X} = \mathbf{x}) > 0$, so conditioning on $\{\mathbf{X} = \mathbf{x}, D = d\}$ is well defined and

$$\mathbb{E}\big[\ell^{\text{obs}}\big(A_H(D), Y\big) \mid \mathbf{X} = \mathbf{x}, D = d\big] = \mathbb{E}\big[\ell\big(A_{H_d}(d), Y\big) \mid \mathbf{X} = \mathbf{x}, D = d\big]\,.$$

By exchangeability of decision-makers (A5), H_d is drawn from π independently of the covariates, and of the regime. The decision-maker observed under regime d is therefore distributed as a generic $H \sim \pi$, and

$$\mathbb{E}\big[\ell\big(A_{H_d}(d), Y\big) \mid \mathbf{X} = \mathbf{x}, D = d\big] = \mathbb{E}_{H \sim \pi}\big[\ell\big(A_H(d), Y\big) \mid \mathbf{X} = \mathbf{x}, D = d\big]\,.$$

By unconfoundedness (A1), the potential losses are independent of the regime given $\mathbf{X}$, so

$$\mathbb{E}_{H \sim \pi}\big[\ell\big(A_H(d), Y\big) \mid \mathbf{X} = \mathbf{x}, D = d\big] = \mathbb{E}_{H \sim \pi}\big[\ell\big(A_H(d), Y\big) \mid \mathbf{X} = \mathbf{x}\big] = r_d(\mathbf{x}),$$

where the last equality is the definition of the regime-specific risk in Eq. (1). Hence both $r_0(\mathbf{x})$ and $r_1(\mathbf{x})$ are identified by the observed conditional expectations, and their difference identifies $\text{VoI}(\mathbf{x}) = r_0(\mathbf{x}) - r_1(\mathbf{x})$. □

A.3 Proof of the no-disclosure degradation bound

For completeness, we restate here Theorem 2 from Section 3.3, and then provide its proof. Recall that $\hat{d}(\mathbf{x}) = \mathbb{1}\{\widehat{\text{VoI}}(\mathbf{x}) \geq \hat{\lambda}\}$ is the plug-in policy, $\hat{B} = \mathbb{E}_{\mathbf{x}}[\hat{d}(\mathbf{x})]$ its realized disclosure rate, and $\varepsilon(\mathbf{x}) = \widehat{\text{VoI}}(\mathbf{x}) - \text{VoI}(\mathbf{x})$ the estimation error of the VoI.

Theorem 2. *No-disclosure degradation] Given a non-negative loss function $\ell : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}^+$, for any estimator $\widehat{\text{VoI}}(\mathbf{x})$ such that $\|\varepsilon\|_{L^2(p)} < \infty$ and any $\hat{\lambda} \geq 0$, the following holds:*

$$R(\hat{d}) - R(0) \;\leq\; \sqrt{\hat{B}}\,\|\varepsilon\|_{L^2(p)} \;-\; \hat{\lambda}\hat{B}.$$

Proof. Recall from the proof of Theorem 1 that

$$R(d) = \mathbb{E}_{\mathbf{x}}\big[(1 - d(\mathbf{x}))\, r_0(\mathbf{x}) + d(\mathbf{x})\, r_1(\mathbf{x})\big] = \mathbb{E}_{\mathbf{x}}[r_0(\mathbf{x})] - \mathbb{E}_{\mathbf{x}}\big[d(\mathbf{x})\, \text{VoI}(\mathbf{x})\big]. \tag{15}$$

Since $R(0) = \mathbb{E}_{\mathbf{x}}[r_0(\mathbf{x})]$, we have $R(\hat{d}) - R(0) = -\mathbb{E}_{\mathbf{x}}[\hat{d}(\mathbf{x})\, \text{VoI}(\mathbf{x})]$. Writing $\text{VoI}(\mathbf{x}) = \widehat{\text{VoI}}(\mathbf{x}) - \varepsilon(\mathbf{x})$,

$$R(\hat{d}) - R(0) = \underbrace{\mathbb{E}_{\mathbf{x}}[\hat{d}(\mathbf{x})\, \varepsilon(\mathbf{x})]}_{\text{(I)}} - \underbrace{\mathbb{E}_{\mathbf{x}}[\hat{d}(\mathbf{x})\, \widehat{\text{VoI}}(\mathbf{x})]}_{\text{(II)}}. \tag{16}$$

For (I), we apply Cauchy–Schwarz to $\hat{d}$ and $\hat{d}\,\varepsilon$, using $\hat{d}(\mathbf{x})^2 = \hat{d}(\mathbf{x})$ since $\hat{d}$ is an indicator:

$$\mathbb{E}_{\mathbf{x}}[\hat{d}(\mathbf{x})\, \varepsilon(\mathbf{x})] \;\leq\; \big(\mathbb{E}_{\mathbf{x}}[\hat{d}(\mathbf{x})]\big)^{1/2} \big(\mathbb{E}_{\mathbf{x}}[\hat{d}(\mathbf{x})\, \varepsilon(\mathbf{x})^2]\big)^{1/2} \;\leq\; \sqrt{\hat{B}}\,\|\varepsilon\|_{L^2(p)}.$$

For (II), on $\{\hat{d}(\mathbf{x}) = 1\}$ we have $\widehat{\text{VoI}}(\mathbf{x}) \geq \hat{\lambda} \geq 0$, so $\hat{d}(\mathbf{x})\, \widehat{\text{VoI}}(\mathbf{x}) \geq \hat{\lambda}\, \hat{d}(\mathbf{x})$ for every $\mathbf{x}$, whence (II) $\geq \hat{\lambda}\, \hat{B}$. Substituting both bounds into Eq. (16) gives the claim. □

The two terms have opposite signs and opposite meanings. Term (I) is the price of estimation error, and it is paid *only where the policy discloses*: the factor $\hat{d}(\mathbf{x})$ zeroes out the contribution of every instance the policy leaves untouched, so that only the error on the disclosed instances, $\mathbb{E}_{\mathbf{x}}[\hat{d}(\mathbf{x})\, \varepsilon(\mathbf{x})^2]$, enters the bound, which therefore vanishes as the disclosure rate $\hat{B}$ shrinks. Term (II) is instead the gain the policy *believes* it is making, and enters with a negative sign: since disclosure is triggered only above the threshold, every disclosed instance contributes an estimated improvement of at least $\hat{\lambda}$, which offsets part of the estimation error.

When can the plug-in policy degrade? The bound certifies no degradation only when $\|\varepsilon\|_{L^2(p)} \leq \hat{\lambda}\sqrt{\hat{B}}$, i.e., when the estimation error is small relative to the margin by which disclosed instances clear the threshold; otherwise, its right-hand side is positive and no-degradation is not guaranteed. When it does occur, degradation has a precise source. By Eq. (15), $R(\hat{d}) - R(0) = -\mathbb{E}_{\mathbf{x}}[\hat{d}(\mathbf{x})\, \widehat{\text{VoI}}(\mathbf{x})]$, so the plug-in policy can be worse than never disclosing only if it discloses on instances where $\text{VoI}(\mathbf{x}) < 0$. On these instances $\widehat{\text{VoI}}(\mathbf{x}) \geq \hat{\lambda} \geq 0$, hence $\varepsilon(\mathbf{x}) > \hat{\lambda}$: degradation requires the estimator to overestimate the VoI by more than the threshold on part of the disclosed region. This is most likely when disclosure is harmful on a large fraction of instances and the budget is not binding. In this case $\hat{\lambda} = 0$, term (II) provides no margin, and any instance with $\text{VoI}(\mathbf{x}) < 0$ but $\widehat{\text{VoI}}(\mathbf{x}) \geq 0$ is disclosed. This is the regime we observe on `Email`, where `ClassWise` discloses on at most $\approx 53\%$ of the instances (Fig. 7): for $B \geq .70$ the threshold is zero and accuracy falls slightly below `NoDisc` (Fig. 2, Fig. 6).

A.4 Proof of the oracle regret bound

For completeness, we restate here Theorem 3 from Section 3.3, and then provide its proof. Recall that $\hat{d}(\mathbf{x}) = \mathbb{1}\{\widehat{\text{VoI}}(\mathbf{x}) \geq \hat{\lambda}\}$ is the plug-in policy, $\hat{B} = \mathbb{E}_{\mathbf{x}}[\hat{d}(\mathbf{x})]$ its realized disclosure rate, and $\varepsilon(\mathbf{x}) = \widehat{\text{VoI}}(\mathbf{x}) - \text{VoI}(\mathbf{x})$ the estimation error of the VoI.

Theorem 3 (Oracle regret)**.** *Given a non-negative loss function $\ell : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}^+$, let d^* be the optimal policy of Theorem 1 and $B^* = \mathbb{E}_{\mathbf{x}}[d^*(\mathbf{x})]$. Assume that the cumulative distribution of $\widehat{\text{VoI}}(\mathbf{x})$ is continuous, let $\hat{q}_{1-B}$ denote its $(1-B)$-quantile and let $\hat{\lambda} = \max\{0,\, \hat{q}_{1-B}\}$. Then*

$$R(\hat{d}) - R(d^*) \;\leq\; \sqrt{B^* + \hat{B}}\,\|\varepsilon\|_{L^2(p)}.$$

Proof. By Eq. (15), and writing $\mathrm{VoI}(\mathbf{x}) = \widehat{\mathrm{VoI}}(\mathbf{x}) - \varepsilon(\mathbf{x})$,

$$R(\hat{d}) - R(d^*) = \mathbb{E}_{\mathbf{x}}\big[(d^*(\mathbf{x}) - \hat{d}(\mathbf{x}))\,\mathrm{VoI}(\mathbf{x})\big] = \underbrace{\mathbb{E}_{\mathbf{x}}\big[(d^*(\mathbf{x}) - \hat{d}(\mathbf{x}))\,\widehat{\mathrm{VoI}}(\mathbf{x})\big]}_{\text{(III)}} - \mathbb{E}_{\mathbf{x}}\big[(d^*(\mathbf{x}) - \hat{d}(\mathbf{x}))\,\varepsilon(\mathbf{x})\big]. \tag{17}$$

Term (III) is non-positive. The proof of Theorem 1, applied with $\widehat{\mathrm{VoI}}(\mathbf{x})$ in place of $\mathrm{VoI}(\mathbf{x})$, shows that the threshold rule $\mathbb{1}\{\widehat{\mathrm{VoI}}(\mathbf{x}) \geq \max\{0,\, \hat{q}_{1-B}\}\}$ maximizes $\mathbb{E}_{\mathbf{x}}[d(\mathbf{x})\,\widehat{\mathrm{VoI}}(\mathbf{x})]$ over all policies with $\mathbb{E}_{\mathbf{x}}[d(\mathbf{x})] \leq B$; by assumption, $\hat{d}$ is that rule. Since d^* also satisfies the budget constraint, as $B^* \leq B$ by Theorem 1, we get $\mathbb{E}_{\mathbf{x}}[d^*(\mathbf{x})\,\widehat{\mathrm{VoI}}(\mathbf{x})] \leq \mathbb{E}_{\mathbf{x}}[\hat{d}(\mathbf{x})\,\widehat{\mathrm{VoI}}(\mathbf{x})]$.

Let $\Delta(\mathbf{x}) = |d^*(\mathbf{x}) - \hat{d}(\mathbf{x})|$, the indicator of the region where the two policies disagree. Since $d^*(\mathbf{x}) - \hat{d}(\mathbf{x}) \in \{-1, 0, 1\}$, we have $\Delta(\mathbf{x})^2 = \Delta(\mathbf{x})$, and Cauchy–Schwarz applied to Δ and $\Delta\,|\varepsilon|$ gives

$$\begin{aligned} R(\hat{d}) - R(d^*) &\leq \mathbb{E}_{\mathbf{x}}\big[\Delta(\mathbf{x})\,|\varepsilon(\mathbf{x})|\big] \leq \big(\mathbb{E}_{\mathbf{x}}[\Delta(\mathbf{x})]\big)^{1/2}\big(\mathbb{E}_{\mathbf{x}}[\Delta(\mathbf{x})\,\varepsilon(\mathbf{x})^2]\big)^{1/2} \\ &\leq \big(\mathbb{E}_{\mathbf{x}}[\Delta(\mathbf{x})]\big)^{1/2}\,\|\varepsilon\|_{L^2(p)}. \end{aligned}$$

Finally, since d^* and $\hat{d}$ take values in $\{0, 1\}$, $\Delta(\mathbf{x}) \leq d^*(\mathbf{x}) + \hat{d}(\mathbf{x})$, so that $\mathbb{E}_{\mathbf{x}}[\Delta(\mathbf{x})] \leq B^* + \hat{B}$, which gives the claim. □

The intermediate bound shows that the regret depends only on the region where the two policies disagree: both its mass, $\mathbb{E}_{\mathbf{x}}[\Delta(\mathbf{x})]$, and the estimation error on it, $\mathbb{E}_{\mathbf{x}}[\Delta(\mathbf{x})\,\varepsilon(\mathbf{x})^2]$, enter the bound, whereas errors on instances that both policies disclose, or both withhold, do not contribute. The final bound replaces the mass of the disagreement region with the upper bound $B^* + \hat{B}$, which is loose when the two policies largely agree.

A.5 Proof of class-wise decomposition

For completeness, we restate here Proposition 2 from Section 3.4, and then provide its proof.

Proposition 2 (Class-wise decomposition)**.** *Let us consider* $\ell(A(D), Y) = \mathbb{1}\{A(D) \neq Y\}$. *Then* $r_D(\mathbf{x})$ *can be written as:*

$$r_D(\mathbf{x}) = \sum_{y \in \mathcal{Y}} \underbrace{\mathbb{P}(Y = y \mid \mathbf{X} = \mathbf{x})}_{p_y}\,\underbrace{\mathbb{P}(A(D) \neq y \mid \mathbf{X} = \mathbf{x}, Y = y)}_{q_{D,y}}. \tag{18}$$

Proof. Since the expectation of an indicator is the probability of the corresponding event, the risk under the 0-1 loss is

$$r_D(\mathbf{x}) = \mathbb{E}\big[\mathbb{1}\{A(D) \neq Y\} \mid \mathbf{X} = \mathbf{x}\big] = \mathbb{P}(A(D) \neq Y \mid \mathbf{X} = \mathbf{x}).$$

By the law of total probability over the values of Y,

$$\mathbb{P}(A(D) \neq Y \mid \mathbf{X} = \mathbf{x}) = \sum_{y \in \mathcal{Y}} \mathbb{P}(A(D) \neq y,\, Y = y \mid \mathbf{X} = \mathbf{x}),$$

where on the event $\{Y = y\}$ the condition $A(D) \neq Y$ becomes $A(D) \neq y$. Finally, by the chain rule of probability, each term factorizes as

$$\mathbb{P}(A(D) \neq y,\, Y = y \mid \mathbf{X} = \mathbf{x}) = \mathbb{P}(Y = y \mid \mathbf{X} = \mathbf{x})\,\mathbb{P}(A(D) \neq y \mid \mathbf{X} = \mathbf{x},\, Y = y),$$

which gives the claim. □

B A causal inference interpretation

Here we detail the causal interpretation provided in Section 3.2. Throughout, we use D for the random disclosure regime and $d \in \{0, 1\}$ for its values, and $H \sim \pi$ for a random decision-maker

and $h \in \mathcal{H}$ for a fixed one. The disclosure variable D plays the role of a binary treatment on the decision-maker: for a decision-maker H, we consider the *potential* actions $A_H(0)$ and $A_H(1)$, corresponding to the decisions under the two regimes, and the associated potential losses $\ell(A_H(0), Y)$ and $\ell(A_H(1), Y)$, which are the potential outcomes of interest.

For a fixed decision-maker $h \in \mathcal{H}$, we define the human-specific risks

$$r_d(\mathbf{x}, h) = \mathbb{E}\big[\ell\big(A_h(d), Y\big) \mid \mathbf{X} = \mathbf{x}, H = h\big], \qquad d \in \{0, 1\},$$

so that the human-specific effect of disclosure introduced in Section 3.2 is $\tau_h(\mathbf{x}) = r_1(\mathbf{x}, h) - r_0(\mathbf{x}, h)$. This is the effect of disclosing S on the decisions of h, rather than a treatment effect on the case itself. Under the exchangeability of decision-makers (A5 below), the decision-maker is drawn independently of the case, and the regime-specific risks of Eq. (1) are the population averages $r_d(\mathbf{x}) = \mathbb{E}_{H\sim\pi}[r_d(\mathbf{x}, H)]$. Hence $\tau_\pi(\mathbf{x}) = \mathbb{E}_{H\sim\pi}[\tau_H(\mathbf{x})] = r_1(\mathbf{x}) - r_0(\mathbf{x})$, and $\mathrm{VoI}(\mathbf{x}) = -\tau_\pi(\mathbf{x})$.

We now state the assumptions under which this identification is valid.

A1 **Unconfoundedness.** Conditional on the baseline covariates $\mathbf{X}$, the disclosure regime is as good as random with respect to the potential losses (Rosenbaum & Rubin, 1983):

$$\big\{\ell(A_H(0), Y),\ \ell(A_H(1), Y)\big\} \perp\!\!\!\perp D \mid \mathbf{X}.$$

Intuitively, all possible sources of self-selection into disclosure (and non-disclosure) are captured by the observable covariates.

A2 **Overlap (positivity).** For all covariate profiles in the population, both disclosure regimes are observed with non-zero probability (Rubin, 1978). Let $e(\mathbf{x}) = \mathbb{P}(D = 1 \mid \mathbf{X} = \mathbf{x})$ denote the disclosure propensity. We assume there exists some $\eta > 0$ such that

$$\eta < e(\mathbf{x}) < 1 - \eta \quad \forall \mathbf{x} \in \mathcal{X}.$$

This guarantees that we can compare the risks under disclosure and non-disclosure at each covariate level and then aggregate these comparisons.

A3 **Consistency.** If a decision episode is observed under regime $D = d$, the observed loss equals the corresponding potential loss (Cole & Frangakis, 2009):

$$\ell^{\mathrm{obs}}\big(A_H(D), Y\big) = \ell\big(A_H(d), Y\big) \quad \text{on } \{D = d\}.$$

A4 **Stable Unit Treatment Value Assumption (SUTVA).** For each decision episode, the potential action of the decision-maker under regime $d \in \{0, 1\}$ depends only on the information disclosed to that decision-maker in that episode (Rubin, 1981).

A5 **Exchangeability of decision-makers.** For each disclosure regime $d \in \{0, 1\}$, the decision-maker assigned to a decision episode is drawn from the same target population π, independently of the case covariates and the disclosure regime. Formally,

$$H_d \mid \mathbf{X} = \mathbf{x}, D = d \sim \pi, \qquad \text{for all} \quad \mathbf{x} \in \mathcal{X} \quad \text{and} \quad d \in \{0, 1\}.$$

Accordingly, the regime-specific conditional risk is the expected loss of a decision-maker drawn from π:

$$r_d(\mathbf{x}) = \mathbb{E}_{H\sim\pi}\left[\ell\big(A_H(d), Y\big) \mid \mathbf{X} = \mathbf{x}\right].$$

Therefore,

$$\mathrm{VoI}(\mathbf{x}) = r_0(\mathbf{x}) - r_1(\mathbf{x})$$

measures the expected reduction in decision loss induced by disclosure for a decision-maker drawn from the target population π, rather than for a particular individual. This assumption does not require observing the same decision-maker under both regimes. It requires that the decision-makers observed under the two regimes be representative of the same target population.

C Implementation details

C.1 Synthetic experiments.

Dataset generation. For our `Synth` experiments, we consider both a binary classification setting (`SynthBin`: $|\mathcal{Y}| = 2$) and a multiclass setting (`SynthMulti`: $|\mathcal{Y}| = 4$). In each case, we generate data using `sklearn.make_classification` with `n_features = 20` (all informative),

`n_samples = 10000`, and `class_sep = 1`. To emulate selective disclosure, we treat a subset of $|\mathcal{X}| = 9$ features as *baseline* information and the remaining $|\mathcal{S}| = 11$ as *hidden* (i.e., information that is not available in the no-disclosure regime). Since no real human annotations are available in this setting, we simulate the two decision-makers with logistic regression models: the $D = 0$ decisor is trained on the 9 baseline features only, while the $D = 1$ decisor is trained on all 20 features. For each instance, the corresponding action $a_i(d)$ is the label predicted by the respective model, so that the $D = 0$ decisor is systematically less accurate, and errors concentrate on the instances whose label depends on the hidden features.

Models. For the `ClassWise` estimator, we first train a multiclass label model $f\colon \mathcal{X} \to \Delta^{|\mathcal{Y}|}$, implemented as a two-layer MLP with ReLU activations and a final linear layer followed by a softmax over classes, to approximate $\mathbb{P}(Y = y \mid X = \mathbf{x})$. In parallel, for each regime D we train a multi-head error model consisting of a shared backbone (a two-layer MLP with GELU activations) and $|\mathcal{Y}|$ binary heads, implemented as a single linear layer of size $|\mathcal{Y}|$; the y-th head takes the shared representation and outputs a logit whose sigmoid corresponds to $\mathbb{P}(A(D) \neq y \mid X = \mathbf{x}, Y = y)$. Only the head of the observed class contributes to the loss, which is a $\mathrm{pos_weight}$-weighted binary cross-entropy with the same logit correction at prediction time.

For the `T-Learner` estimator, we train two separate scalar risk networks, one per regime $D \in \{0, 1\}$, each modeling $r_D(\mathbf{x}) = \mathbb{P}(A(D) \neq Y \mid X = \mathbf{x})$. Concretely, each network is a two-layer MLP with GELU activations and dropout, optimized with a weighted binary cross-entropy loss to predict the error indicator $\mathbb{1}\{A(D) \neq Y\}$ from the baseline features $\mathcal{X}$.

The `S-Learner` estimator replaces the two separate risk networks of the `T-Learner` with a single network $z\colon \mathcal{X} \times \{0, 1\} \to [0, 1]$ that takes the regime as an additional input feature: the binary indicator D is appended to $\mathbf{x}$, and the network is a two-layer MLP with GELU activations and dropout after each hidden layer, followed by a scalar output. It is trained on the pooled data of both regimes, so that every instance contributes two examples, $(\mathbf{x}, 0)$ with target $\mathbb{1}\{A(0) \neq Y\}$ and $(\mathbf{x}, 1)$ with target $\mathbb{1}\{A(1) \neq Y\}$, under a single $\mathrm{pos_weight}$-weighted binary cross-entropy.

Finally, the `Confidence` baseline does not model the human error at all: it only trains the label model g of the `ClassWise` estimator and ranks instances by its predictive uncertainty, setting $\hat{r}_0(\mathbf{x}) = u(g(\mathbf{x}))$ and $\hat{r}_1(\mathbf{x}) = 0$, so that $\widehat{\mathrm{VoI}}(\mathbf{x}) = u(g(\mathbf{x}))$. We use $u(\cdot) = 1 - \max_y g(\mathbf{x})_y$ in our experiments. This baseline encodes the intuition that information should be disclosed where the instance is intrinsically hard, regardless of whether the support information helps there; note that, since $u \geq 0$, its estimated $\widehat{\mathrm{VoI}}(\mathbf{x})$ is never negative and the policy therefore always spends the full budget, unlike the VoI-based estimators.

For both the `SynthBin` and `SynthMulti` settings, all models are trained with Adam for 10 epochs, with weight decay 10^{-4} and without early stopping, using the following hyperparameters:

- Label model f (also used by `Confidence`): hidden dimension 128, batch size 32, learning rate 5×10^{-4}, no dropout.
- Error model g_D: backbone hidden dimension 256, batch size 32, learning rate 10^{-4}, no dropout.
- `T-Learner` and `S-Learner` risk networks: hidden dimension 128, batch size 128, learning rate 5×10^{-4}, dropout 0.05.

C.2 Real data experiments - Email.

Email Dataset description. We used a dataset comprising 1000 real emails with ground-truth labels indicating whether each email was fraudulent (e.g., a phishing attempt) or legitimate. The emails were selected from the corpus introduced by (Rebeka Toth, 2025), which provides only the messages; the human annotations are instead collected in a dedicated behavioural study (see Bogani et al. 2026 for details on the email selection procedure). In that study, 300 participants were recruited online through Prolific and asked to provide binary judgments for 20 emails randomly sampled from the dataset, indicating whether they considered each email to be fraudulent or legitimate. Participants were randomly assigned to one of two disclosure regimes. In the no-disclosure condition ($D = 0$), participants were presented only with the email to be classified. In the disclosure condition ($D = 1$), they were additionally shown the prediction of a machine-learning model.

Table 1: Selected hyperparameters for the Email risk models.

Model	Epochs	Optimizer	LR	Dropout	Weight decay
ClassWise	50	ADAM	2×10^{-4}	0.3	10^{-3}
T-Learner	25	ADAMW	2×10^{-4}	0.1	10^{-3}
S-Learner	25	ADAM	1×10^{-4}	0.1	10^{-2}
Confidence	100	ADAMW	2×10^{-4}	0.3	10^{-3}

Notably, in this dataset, exchangeability of decision-makers is supported by the random assignment of participants to the two disclosure regimes. We process these data in two steps:

(i) We first aggregate the raw annotations into a per-email summary: each row corresponds to a single email and includes the ground-truth label and the empirical human response distribution in each regime. Concretely, for each label $y \in \{fraudulent, legitimate\}$ we compute human_0y as the fraction of $D = 0$ participants who assigned the email to class y, and analogously human_1y for the $D = 1$ participants. This yields, for each email and regime, an estimated human probability distribution over the two outcomes, from which we derive the human error indicators err_0 and err_1 (whether a label sampled from the corresponding distribution differs from the ground truth) used in our experiments.

(ii) Then, we augment the train dataset to increase diversity and robustness. We leverage text-based augmentations provided by (Rebeka Toth, 2025) [7] and complement them with back-translation. For back-translation, we translate each email into seven languages (fr, de, es, it, pt, nl, ru) and then translate it back to English, obtaining multiple paraphrased variants per original message. In total, this procedure yields up to 14 augmented versions per email and results in 15000 samples used in our experiments.

Email models. For the Email dataset, all models operate on precomputed text embeddings rather than on raw text. Each email is encoded with a publicly available transformer-based phishing-detection model [8], fine-tuned for binary phishing classification: we truncate the email body to 512 tokens and take the final-layer [CLS] representation, yielding a 768-dimensional embedding per (original or augmented) email. On top of these embeddings, every network uses the same MLP trunk: a LayerNorm on the input followed by two GELU hidden layers with dropout after each layer (halved on the second), where the second hidden layer has half the width of the first. The label model f (also used by Confidence) and the T-Learner risk networks use hidden widths $768 \to 128 \to 64$, the error model g_D of ClassWise uses $768 \to 256 \to 128$, and the S-Learner network, whose input includes the regime indicator, uses $769 \to 128 \to 64$. The label and error models are trained with batch size 64, the T-Learner and S-Learner networks with batch size 128 on inputs standardized with training-split statistics.

Model selection is performed by grid search over the number of training epochs $\{25, 50, 100\}$, the optimizer (ADAM, ADAMW), the learning rate $\{1 \times 10^{-4}, 2 \times 10^{-4}\}$, the dropout rate $\{0.1, 0.3\}$ and the weight decay $\{10^{-3}, 10^{-2}\}$, retaining for each model the configuration with the highest binary AUROC on the validation split. The resulting hyperparameters are reported in Table 1.

C.3 REAL DATA EXPERIMENTS - IMAGENET-16H.

ImageNet-16H Dataset description. We used the IMAGENET-16H dataset introduced by (Steyvers et al., 2022), as processed for the human study of (De Toni et al., 2024). The dataset comprises 1200 natural images drawn from 16 categories {airplane, bear, bicycle, bird, boat, bottle, car, cat, chair, clock, dog, elephant, keyboard, knife, oven, truck}, each corrupted with phase noise to make the classification task challenging for both humans and models; we use the highest available noise level (110), which yields the hardest regime. Each image carries a ground-truth category label. A VGG-19 classifier fine-tuned on the noisy

[7] We use the augmentations released at emails-expanded-2026-Gemini.json, which were generated via prompting a large language model.

[8] phishing-email-detector-v51

images provides softmax scores over the 16 classes, from which (De Toni et al., 2024) construct conformal prediction sets using the standard split-conformal calibration procedure.

Concretely, given a target miscoverage level α, the conformal score of an image is defined as $s = 1 - \hat{p}(y \mid x)$, where $\hat{p}(y \mid x)$ is the model's softmax probability of the ground-truth class y (Law, 2006). The threshold q is set to the $(1-\alpha)(n+1)/n$ empirical quantile of these scores on a calibration set, and the prediction set for an image is the collection of classes whose softmax probability exceeds q. This construction enjoys the usual marginal coverage guarantee of $1 - \alpha$: smaller values of α enforce higher coverage and therefore yield larger prediction sets, whereas larger values of α produce smaller, more informative sets (Sadinle et al., 2016). Since the human study recorded, for each image, the responses of participants who were shown a prediction set of a given size, the choice of α determines both the set displayed in the disclosure condition and the corresponding empirical human response distribution. In our experiments we consider three miscoverage levels, $\alpha \in \{0.01, 0.02, 0.05\}$. We will discuss in Appendix E.4 about the choice of α.

Human judgments were collected online in two disclosure regimes, drawn from two separate behavioural studies over the same 1200 images. In the no-disclosure condition ($D = 0$), taken from (Steyvers et al., 2022), participants were shown only the (noisy) image and asked to select one of the 16 categories, yielding 7261 classifications from 145 participants. In the disclosure condition ($D = 1$), taken from (Straitouri & Rodriguez, 2024) [9], participants were additionally shown the model's conformal prediction set for the image before selecting a label. The set was presented in a *lenient* regime: it served as a suggestion rather than a constraint, so participants were free to choose any of the 16 categories, including labels outside the displayed set [10]

Notably, exchangeability of decision-makers for this dataset is plausible because: (i) the two studies use the same task and item population and (ii) recruit participants from comparable populations under similar experimental protocols, with the principal difference being the availability of support information. We process the human study data in two steps:

(i) We first aggregate the raw annotations into a per-image summary: each row corresponds to a single image and includes the ground-truth label, the model softmax scores, and the human responses collected under each regime. For the no-disclosure regime, for each category y we compute `human_0y` as the fraction of $D = 0$ participants who assigned the image to class y, and a human label is then obtained by sampling from this distribution. For the disclosure regime, the response we use depends on the chosen miscoverage level α. For each image, the experiment of (Straitouri & Rodriguez, 2024) collected disclosure responses from 16 distinct $D = 1$ participants, each of whom was shown a prediction set of a different size, ranging from 1 to 16 labels. Since the size of the conformal prediction set for a given image is determined by α, each of these responses corresponds to the decision a human would make when assisted at a particular miscoverage level. Fixing α therefore determines, for each image, the size of the set the policy would actually disclose, and we retain only the response of the participant who was shown a set of exactly that size. This yields, for each image and regime, a human label (sampled from the empirical distribution in the no-disclosure regime, or the human response selected via α in the disclosure regime) from which we derive the human error indicators err_0 and err_1 used in our experiments.

(ii) Then, we augment the train dataset to increase diversity and robustness. At training time each image is transformed by composing 3 randomly sampled geometric augmentations drawn from a fixed set of 14 operations (identity, horizontal/vertical flips, rotations of $15°$/$30°$/$45°$, horizontal/vertical translations, and x/y shears), yielding multiple perturbed variants per original image while preserving its category label and associated human response distributions. Figure 4 illustrates the resulting variants for a representative image.

Exchangeability across studies. Unlike `Email`, where participants were randomly assigned to regimes within a single study, the two `ImageNet-16H` regimes come from different studies. Both use the same 1200 images at noise level 110, but the protocols slightly differ beyond the availability of support: $D = 0$ participants classified 200 images each, with a confidence rating on every trial,

[9] Datasets for both conditions are publicly available: no-disclosure condition and disclosure condition.

[10] The original study also includes a *strict* condition, in which participants are constrained to choose a label from within the displayed prediction set. We deliberately adopt the lenient regime, as our setting requires the human to remain the final decision-maker and to be free to disregard the model's feedback whenever their own judgment disagrees with it.

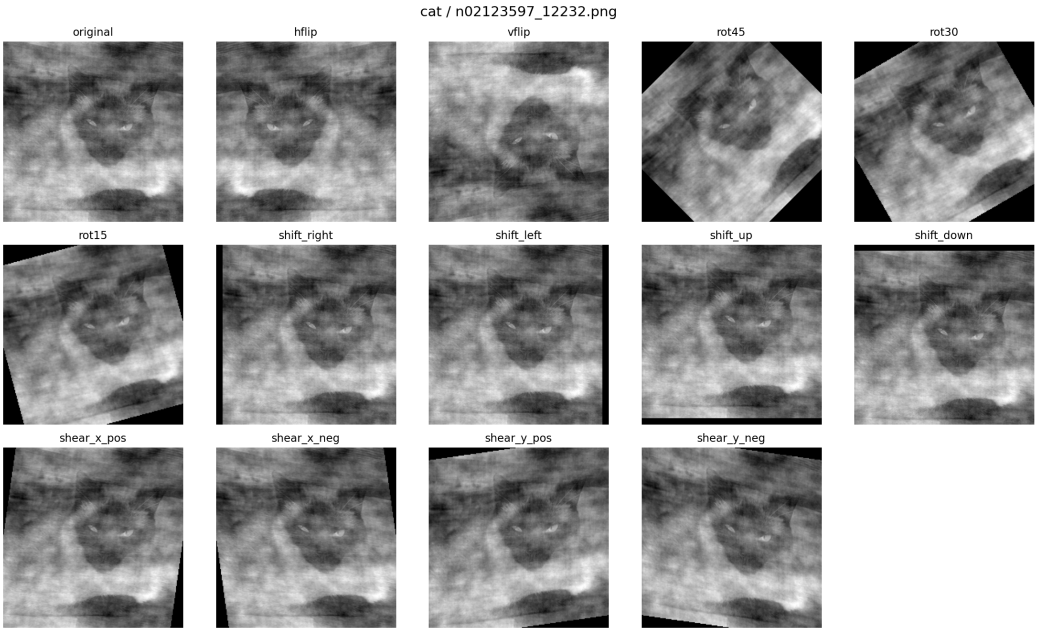


Figure 4: Geometric augmentations applied to the `ImageNet-16H` images at training time, shown here for one image of the `cat` category at the highest phase-noise level (110). Each panel corresponds to one of the 14 operations in the augmentation pool: the identity (*original*), horizontal and vertical flips, rotations of $15°$, $30°$ and $45°$, translations along the four directions, and positive/negative shears along the x and y axes.

from a pool spanning four noise levels (Steyvers et al., 2022), whereas $D = 1$ participants were recruited on Prolific and answered multiple-choice questionnaires at noise level 110 only (Straitouri & Rodriguez, 2024).

We assess the plausibility of the exchangeability assumption using a feature of the design by (Straitouri & Rodriguez, 2024). Their study includes trials in which the displayed prediction set contains the entire label space $\mathcal{Y}$. Because such a set provides no information about which label is more likely, performance on these trials provides a useful proxy for no-disclosure performance. Restricting the comparison to the same images, accuracy on full-set trials is 0.7713, compared with 0.7647 under $D = 0$. The similarity of these accuracies provides descriptive evidence that the decision-maker populations in the two studies have comparable baseline performance.

We acknowledge that this comparison does not establish exchangeability, which cannot be determined from the observed data alone. Moreover, displaying a nondiscriminative prediction set is not identical to withholding support. Nevertheless, together with the shared task and image set, this evidence supports the plausibility of treating participants in the two studies as draws from a common target population π.

`ImageNet-16H` Models. For the `ImageNet-16H` dataset, all models operate directly on the (noisy) images. Every network uses the same feature extractor, an ImageNet-pretrained ConvNeXt-Tiny backbone (Liu et al., 2022) that maps each image to a 768-dimensional embedding and is kept frozen during training, followed by an MLP head with an input $\mathrm{LayerNorm}$, two GELU hidden layers and dropout after each (halved on the second). The hidden widths are $128 \to 256$ for the label model f (also used by `Confidence`, and trained with label smoothing 0.05), $256 \to 256$ for the error model g_D of `ClassWise`, and $128 \to 256$ for the `T-Learner` and `S-Learner` risk networks, the latter taking the regime indicator as an additional input. Losses and logit corrections are as in the synthetic setting. The label and error models are trained with batch size 64, the `T-Learner` and `S-Learner` networks with batch size 128.

For all four models, model selection is performed by grid search over the same hyperparameter space described above for `Email` (epochs, optimizer, learning rate, dropout rate, and weight decay),

Table 2: Selected hyperparameters for the `ImageNet-16H` risk models, per conformal miscoverage level α.

α	Model	Epochs	Optimizer	LR	Dropout	Weight decay
0.01	`T-Learner`	25	AdamW	2×10^{-4}	0.1	10^{-3}
	`ClassWise`	25	AdamW	1×10^{-4}	0.3	10^{-2}
	`S-Learner`	25	AdamW	2×10^{-4}	0.1	10^{-3}
	`Confidence`	25	Adam	1×10^{-4}	0.3	10^{-2}
0.02	`T-Learner`	25	AdamW	2×10^{-4}	0.1	10^{-3}
	`ClassWise`	25	AdamW	1×10^{-4}	0.1	10^{-3}
	`S-Learner`	25	Adam	1×10^{-4}	0.3	10^{-3}
	`Confidence`	25	Adam	1×10^{-4}	0.1	10^{-3}
0.05	`T-Learner`	25	AdamW	1×10^{-4}	0.3	10^{-3}
	`ClassWise`	100	AdamW	1×10^{-4}	0.3	10^{-2}
	`S-Learner`	100	Adam	1×10^{-4}	0.3	10^{-2}
	`Confidence`	25	Adam	1×10^{-4}	0.1	10^{-3}

tuning each model separately for every miscoverage level $\alpha \in \{0.01, 0.02, 0.05\}$ and retaining the configuration with the highest binary AUROC on the validation split. The resulting hyperparameters are reported in Table 2.

D User studies

Procedure. For both experiments, we recruit participants through Prolific and randomly assign them to one of the four experimental conditions (Human Policy/Low Budget, *HP-LB*; Human Policy/High Budget, *HP-HB*; Machine Policy/Low Budget, *MP-LB*; Machine Policy/High Budget, *MP-HB*). [11] First, we provide them with task instructions and have them complete two practice trials to familiarize themselves with the interface (see examples of the interfaces used in both experiments in Fig. 5). Both the instructions and the practice trials are tailored to the experimental condition to which participants are assigned. To incentivize attentive responding, we inform participants that the 5 most accurate participants would receive a bonus of £5.00. We also tell participants in the HP conditions that, in the case of a tie, participants who used fewer requests for assistance would be favored to receive the bonus. This is done to prompt participants to request assistance only when they deem it necessary, similarly to what occurs when the provision of assistance is determined by the `ClassWise` policy in the MP conditions, which not necessarily exhaust the available budget.

Participants then proceed to the actual experiment, which consists of a total of 20 trials. In each trial, participants have to classify an item (an email in the `Email` task or an image in the `ImageNet-16H` one). [12] If they are assigned to one of the HP conditions, they can request assistance from the AI system (provided that they have not exhausted the available requests); if they are assigned to one of the MP conditions, depending on the `ClassWise` policy output, they are either provided with assistance from the AI system or informed that no assistance would be pro-

[11] We did not include a random-disclosure condition because it does not address the primary question of these user studies: comparing the overall effectiveness of selective and human-determined disclosure, independently of the specific factors that may drive the results. A random condition would instead provide a diagnostic comparison of whether machine-selected items lead to better performance compared to an equal number of randomly selected items. We evaluate this selection question directly in Q1, where Random is matched to the learned policy at each budget. Accordingly, the user-study results compare the overall performance in MP and HP condition (without aiming to attribute any difference uniquely to specific factors such as item selection, disclosure frequency, or participant agency). Furthermore, crossing a Random condition with both budget levels would expand the design from four to six conditions, requiring substantially more participants to preserve per-condition statistical power.

[12] Since we planned to recruit participants from the broader population on Prolific, it was important that the tasks did not require specialized knowledge. In this respect, both `Email` and `ImageNet-16H` tasks involve classification tasks that can be performed without domain-specific expertise.

vided for that item. [13] After classifying the item, we also ask participants to indicate how confident they are that their answer is correct on a 7-point Likert scale ranging from 1 ("Not confident at all (guessing)") to 7 ("Extremely confident"). No feedback is provided to participants on the accuracy of their responses. This both reflects realistic classification settings, in which the correctness of a classification may not be immediately verifiable, and limits learning across trials. Throughout the 20 trials, we include two attention checks and record the number of times participants switch away from the experiment browser tab.

The 20 items presented to each participant are randomly sampled from each dataset's test set using stratified sampling. In both the HP and MP conditions, sampling preserves, as closely as possible, the proportions observed in the full test set along two characteristics: whether the selective disclosure method would provide support information for an item and whether such information is useful (i.e., whether, in the human-annotation studies used to construct the original datasets, participants classified that item more accurately with support information than without it). This procedure ensures that the samples presented to participants are as representative as possible of the original test set with respect to these characteristics. In the `Email` task, we further stratified the sampling according to the ground truth of the emails to be classified, in order to avoid item samples in which the emails were predominantly fraudulent or legitimate. Indeed, presenting too many items of the same class to participants during classification tasks may cause them to respond incorrectly simply because a prolonged sequence of identical answers is perceived as unnatural, leading them to change an otherwise correct answer (Pesenti et al., 2026). This risk is particularly relevant in binary classification tasks such as `Email`, but substantially less so when multiple classes are present, as in `ImageNet-16H`; we therefore do not apply this additional stratification to the latter. Finally, we specify that, within each experiment, we sample items from the same pool across all four experimental conditions; what differs between the LB and HB conditions is whether the method provides assistance for a given item, depending on the budget level. See Table 3 for details on the item-sampling stratification. [14]

After having completed all 20 trials, we ask participants to complete a trust scale from (Hoffman et al., 2023), consisting of eight 5-point Likert scale items (extremes: I strongly disagree; I strongly agree). The items, presented in randomized order, were the following:

- I am confident in the system. I feel that it works well.
- The outputs of the system are very predictable.
- The system is very reliable. I can count on it to be correct all the time.
- I feel safe that when I rely on the system I will get the right answers.
- The system is efficient in that it works very quickly.
- I am wary of the system.
- The system can perform the task better than a novice human user.
- I like using the system for decision making.

Finally, we ask participants to report their familiarity with AI systems by selecting one of the following options:

- **Option 1**: I have little or no experience with AI systems and limited or no understanding of how they work.
- **Option 2**: I use AI systems occasionally but have not a clear understanding of how they function.
- **Option 3**: I use AI systems and have studied how they work (e.g., through courses, online classes, or self-study).
- **Option 4**: I develop or build AI systems as part of my work or personal projects.

[13] We note that, for both task, the AI advice provided is the same as the one presented in the annotation studies from which the test sets for the present work were derived.

[14] We excluded some items from the `Email` test set because they were near-duplicates of other emails, and from the `ImageNet-16H` test set because they were potentially ambiguous (e.g., images with *bottle* as the ground-truth class that actually depicted vases). This left 184 emails (original test-set size: 200) and 220 images (original test-set size: 241), from which the items presented to participants were sampled.

Table 3: Stratification of the 20-item samples used in Experiments 1 and 2.

support information	Usefulness of assistance	Ground truth	Budget: Low (30%)	Budget: High (70%)
`ImageNet-16H`				
20-item sample characteristics				
Not provided	Not useful	—	12 (62.2%)	6 (29.0%)
Provided	Not useful	—	4 (24.9%)	12 (58.1%)
Not provided	Useful	—	2 (4.6%)	0 (0.4%)
Provided	Useful	—	2 (8.3%)	2 (12.4%)
Items with support information in 20-item sample			6/20 (30%)	14/20 (70%)
Original test-set characteristics				
Test-set size			241	
Policy provides assistance			80/241 (33%)	170/241 (71%)
`Email`				
20-item sample characteristics				
Not provided	Not useful	Legitimate	6 (30.5%)	4 (20.0%)
Provided	Not useful	Legitimate	2 (8.0%)	4 (18.5%)
Not provided	Useful	Legitimate	2 (13.0%)	1 (9.0%)
Provided	Useful	Legitimate	1 (5.5%)	2 (9.5%)
Not provided	Not useful	Fraudulent	5 (23.0%)	4 (18.5%)
Provided	Not useful	Fraudulent	1 (6.5%)	2 (11.0%)
Not provided	Useful	Fraudulent	2 (9.5%)	2 (6.5%)
Provided	Useful	Fraudulent	1 (4.0%)	1 (7.0%)
Items with support information in 20-item sample			5/20 (25%)	9/20 (45%)
Legitimate emails in 20-item sample			11/20 (55%)	11/20 (55%)
Original test-set characteristics				
Test-set size			200	
Policy provides assistance			48/200 (24%)	92/200 (46%)
Legitimate emails			114/200 (57%)	114/200 (57%)

Note. For the 20-item sample characteristics, values outside parentheses indicate the number of items included in the sample, while percentages in parentheses indicate the corresponding proportion of items in the original test set.

All datasets generated from the human-subject experiments and analysis scripts are provided in the Supplementary Materials.

Participants recruitment and samples characteristics We conducted an a priori power analysis using a simulation-based approach (Green & MacLeod, 2016; Kumle et al., 2021) to estimate the sample size to collect. The analysis indicated that a total sample of 268 participants would have provided 82% statistical power to detect an effect of the two-way interaction between *support* and *budget level* conditions as small as $OR = 0.43$. [15] Accordingly, participants were recruited in batches through Prolific. After each batch, we assessed only whether participants met predetermined exclusion criteria: having passed both attention checks and not having left the browser tab in three or more trials. No further analyses of the outcome variables or effect sizes were performed during data collection. Recruitment continued until the final sample comprised a total of at least 268 participants. Inclusion criteria required participants to be native English speakers from the UK and to have a Prolific approval rate above 98%. Participants received compensation equal to £1.30.

For the `Email` and `ImageNet-16H` studies we recruited, respectively, a total of 288 and 278 participants, of which 272 and 273 were included in the final sample. In both studies, participants samples were balanced in terms of age, sex, and previous experience with AI across all experimental conditions (see Table 4).

[15] As detailed later, the majority of the statistical analyses conducted were logistic regressions. Thus, odds ratios correspond to the exponentiated regression coefficients, $exp(b)$.

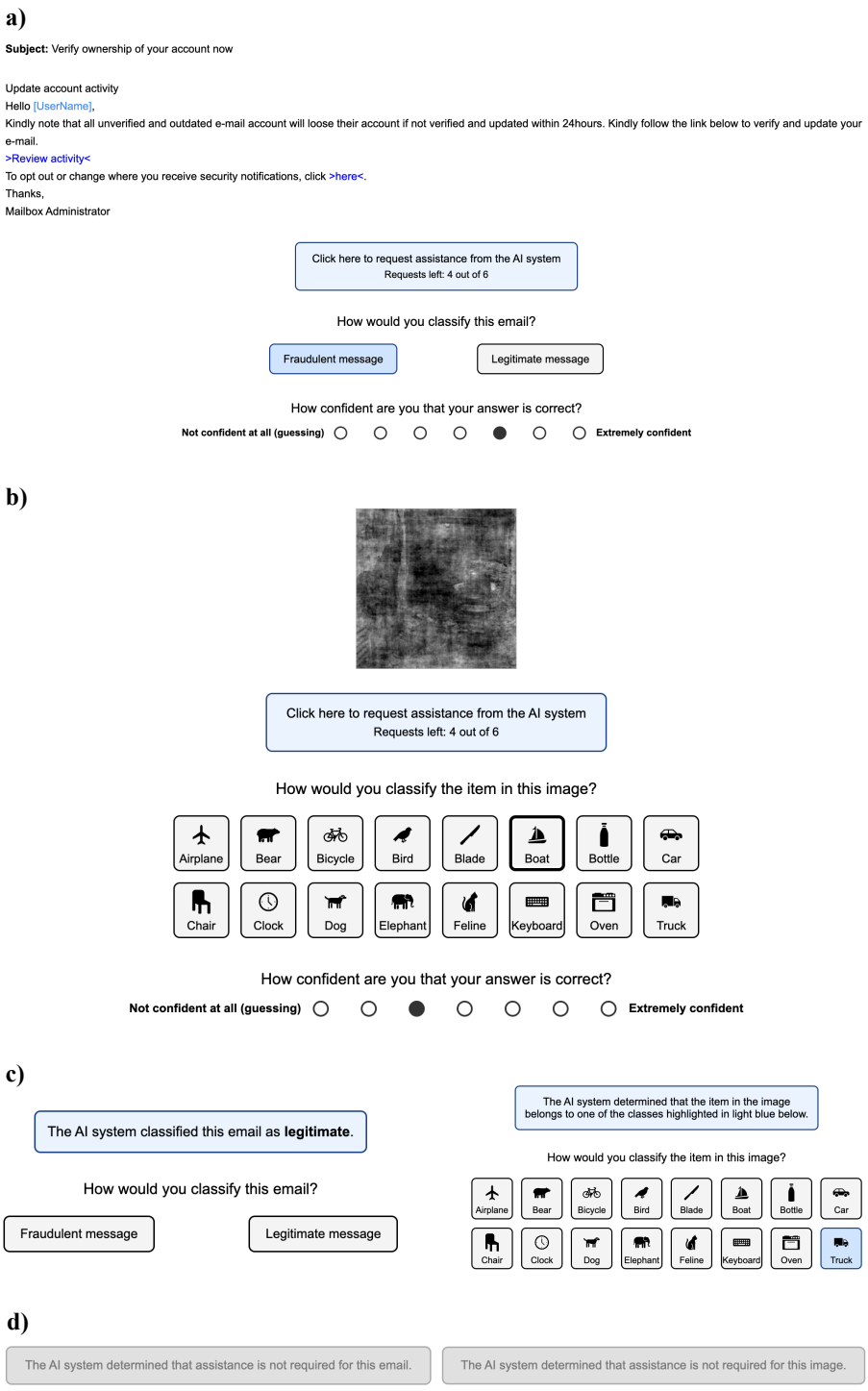


Figure 5: *(a)* and *(b)*: Interfaces presented to participants in the HP conditions in the `Email` and `ImageNet-16H` tasks, respectively; *(c)*: Messages shown when information was requested (HP) or provided by the `ClassWise` policy (MP); *(d)*: Messages shown when, in MP, the `ClassWise` policy determined information should not be provided for that item.

Table 4: Participant characteristics by experimental condition.

Experimental condition	*n* participants	Age	% Female	Past experience with AI			
				Option 1	Option 2	Option 3	Option 4
ImageNet-16H							
Human Policy/Low Budget	66	43.21 ± 14.05	52%	5%	67%	29%	0%
Human Policy/High Budget	69	42.06 ± 12.39	48%	3%	46%	49%	1%
Machine Policy/Low Budget	69	42.30 ± 12.23	48%	6%	54%	41%	0%
Machine Policy/High Budget	69	41.71 ± 13.52	42%	1%	65%	28%	6%
Email							
Human Policy/Low Budget	72	44.58 ± 15.63	51%	4%	53%	37%	6%
Human Policy/High Budget	64	39.33 ± 11.72	48%	0%	64%	34%	2%
Machine Policy/Low Budget	71	41.87 ± 13.15	55%	4%	62%	31%	3%
Machine Policy/High Budget	65	41.98 ± 13.07	55%	3%	49%	48%	0%

E EXTENDED RESULTS

In this section, we detail the results obtained for both the analyses presented in the main text and additional ones we present here. We organize the analyses around the three research questions of the main text, together with a fourth one that we address here:

Q1 Does our approach learn effective policies?

Q2 Does our learned policy improve human-AI team performance?

Q3 How do participants interact with our policy?

Q4 How does changing the information content affect our policy?

For questions related to the user studies (Q2 and Q3), we report the full statistical results. Many of the analyses we performed for such questions involved fitting logistic mixed-effects regression models. These models extend standard regression by including random effects that account for the non-independence of repeated observations (e.g., multiple responses from the same participant or to the same item), thereby yielding valid inferences despite correlated observations (for an overview, see Brown, 2021). For all mixed-effects models reported, we include random intercepts for participants and items.

In these regressions, categorical predictors are deviation-coded as $+0.5/-0.5$ (the level coded as $+0.5$ is reported in brackets in the relevant results tables). This coding allows coefficients to be interpreted as comparisons averaged across the levels of the other categorical predictors. When we perform pairwise post-hoc comparisons, we apply Bonferroni corrections to the p values to control the family-wise error rate at a significance level of .05 (we report corrected p values throughout).

E.1 Q1: DOES OUR APPROACH LEARN EFFECTIVE POLICIES?

Take-home message: `ClassWise` ranks instances better and abstains more selectively than standard meta-learners, and its ability to abstain matters most when disclosure is harmful on a substantial fraction of the instances.

In the main paper, `ClassWise` outperforms the baselines across all datasets (Fig. 2). Here we investigate *why*, through two complementary analyses. First, we isolate the contribution of the class-wise decomposition by replacing it with two standard meta-learners while keeping the disclosure rule fixed: any difference in accuracy can then only come from the quality of the VoI estimates. Second, we look at *how* the policies spend their budget: since the policy never discloses on instances with non-positive estimated VoI, the fraction of instances on which it actually requests help reveals whether an estimator can identify where disclosure does not help.

Ablation on the risk estimator. We compare our `ClassWise` estimator against two standard meta-learners for conditional treatment effects, the `T-Learner` and the `S-Learner` (Künzel et al., 2019). We do not include doubly robust learners (Kennedy, 2023) or causal forests (Wager & Athey, 2018). DR learners rely on cross-fitted nuisance models, which would further split our small real datasets, while their correction matters little here since disclosure is assigned independently of

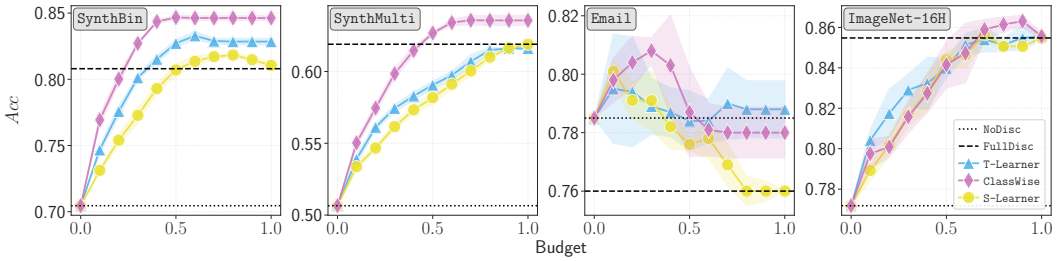


Figure 6: Budget-accuracy curves for VoI-based disclosure policies on `Synth` (`SynthBin`, `SynthMulti`) and real (`Email`, `ImageNet-16H`) data. We report average mean values for 5 seeds and 95% confidence intervals.

the covariates. Causal forests, on the other hand, struggle with the high-dimensional representations of the unstructured inputs (text and images) of our real datasets. `ClassWise` estimates $\mathrm{VOI}(\mathbf{x})$ by first predicting the label and then the class-conditional probability of a human error (Proposition 2). Here we ask how much of its performance is due to this decomposition, as opposed to the disclosure rule itself. We therefore keep the policy, the calibration procedure and the budget grid fixed, and vary only how the two regime-specific risks are estimated. All estimators are tuned over the same hyperparameter grid.

The `T-Learner` fits one scalar risk network per regime directly on the error indicator, $\hat{r}_d(\mathbf{x}) \approx \mathbb{P}(A(d) \neq Y \mid X = \mathbf{x})$, trained on the targets $\mathbb{1}\{A(d) \neq Y\}$, and sets $\widehat{\mathrm{VOI}}(\mathbf{x}) = \hat{r}_0(\mathbf{x}) - \hat{r}_1(\mathbf{x})$. `ClassWise` is itself a T-learner; the `T-Learner` considered here differs only in modelling each risk with a single scalar output rather than through the class-wise decomposition. Since the two regimes get independent functions, the risk curves are free to take different shapes and $\widehat{\mathrm{VOI}}(\mathbf{x})$ is not systematically biased towards zero, at the cost of a higher variance, as it is the difference of two independently estimated quantities.

The `S-Learner` fits a single network that takes the regime as an additional input feature, $\hat{r}(\mathbf{x}, d) \approx \mathbb{P}(A(d) \neq Y \mid X = \mathbf{x})$, trained on the pooled data of both regimes, and sets $\widehat{\mathrm{VOI}}(\mathbf{x}) = \hat{r}(\mathbf{x}, 0) - \hat{r}(\mathbf{x}, 1)$. All parameters are shared across regimes, which reduces variance but tends to shrink the effect of d, and hence $\widehat{\mathrm{VOI}}(\mathbf{x})$, towards zero. Fig. 6 reports the budget–accuracy curves.

On `Synth` data the ordering is stable across budgets: `ClassWise` > `T-Learner` > `S-Learner` at every B, in both the `SynthBin` and the `SynthMulti` setting. The gap is substantial and does not close with the budget: in the `SynthBin` case `ClassWise` plateaus at $\approx .85$ against $\approx .83$ for `T-Learner` and $\approx .81$ for `S-Learner`; in the `SynthMulti` case both meta-learners converge to full-disclosure accuracy ($\approx .62$) without ever exceeding it, whereas `ClassWise` exceeds it from $B \approx .4$. Since the disclosure rule is identical, the difference is entirely attributable to the quality of the $\mathrm{VOI}(\mathbf{x})$ estimate, both its ranking and its sign. The `SynthMulti` panel is where the class-wise decomposition pays off most: disclosure affects the $|\mathcal{Y}|$ classes heterogeneously, and while a scalar risk model can in principle represent this heterogeneity, the decomposition makes it explicit, so that each head only has to learn the error pattern of a single class.

On `Email`, where disclosure is harmful on average, `ClassWise` attains the highest accuracy overall ($\approx .81$ at $B = .30$) but `T-Learner` is better at large budgets ($\approx .79$ against $\approx .78$ for $B \geq .70$). `S-Learner`, instead, degrades monotonically after $B = .10$ and coincides with full disclosure from $B = .80$ onwards. This is the shrinkage of the `S-Learner` made visible: once $\widehat{\mathrm{VOI}}(\mathbf{x})$ is compressed towards zero it loses the information about its sign and ends up almost uniformly positive, so the non-negativity constraint of Theorem 1 stops binding and the policy loses the ability to abstain precisely in the regime where abstention is the whole point.

On `ImageNet-16H`, where disclosure helps almost everywhere, the picture reverses: abstention is nearly irrelevant, all three estimators improve steadily, and `S-Learner` is no longer penalised, tracking `ClassWise` closely up to $B = .60$. Here `T-Learner` leads at low budgets (e.g. $\approx .82$ against $\approx .81$ at $B = .20$) while `ClassWise` takes over from $B \approx .60$ and peaks at $\approx .86$ at $B = .90$, above every other curve and above full disclosure ($\approx .85$).

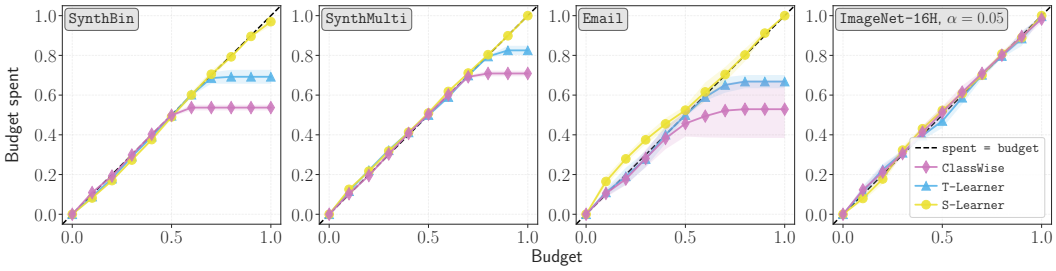


Figure 7: Fraction of instances on which the policy requests help, as a function of the nominal budget, for `S-Learner`, `T-Learner` and `ClassWise` (mean over 5 seeds, with 95% confidence intervals). The dashed line is spent = budget. A curve plateaus below the diagonal when the estimated positive-VoI set is smaller than the budget, and the height of the plateau is the fraction of instances the estimator deems worth disclosing.

Help-request frequency and quality of the risk estimates. A useful diagnostic of a disclosure policy is not only *how much* accuracy it attains, but *how much of the budget it actually spends* to attain it. By Theorem 1, the policy never discloses on instances with non-positive estimated VoI. Once the budget exceeds the fraction of instances with $\widehat{\mathrm{VOI}}(\mathbf{x}) > 0$, additional budget triggers no further disclosures, and the fraction of instances on which the policy requests help plateaus below the nominal budget. The height of this plateau measures the size of the positive-VoI set identified by each estimator, i.e., the fraction of instances it deems worth disclosing. Fig. 7 reports this fraction as a function of the budget for all datasets.

The `S-Learner` stays on the diagonal spent = budget in every setting: $\widehat{\mathrm{VOI}}(\mathbf{x}) \geq 0$ on essentially every instance, so the sign constraint of Theorem 1 never binds and the policy never abstains. By sharing all parameters across regimes, the `S-Learner` compresses $\widehat{\mathrm{VOI}}(\mathbf{x})$ towards a small, almost uniformly positive value and loses the information about its sign. The comparison of interest is therefore between `T-Learner` and `ClassWise`, which both abstain, but to different extents depending on the dataset.

On `SynthBin` and `SynthMulti`, `ClassWise` abstains earlier and more often than `T-Learner`. `ClassWise` plateau is at about 54% of the instances in `SynthBin` and 71% in `SynthMulti`, against 69% and 83% for `T-Learner`. Since `ClassWise` also attains the highest accuracy at every budget (Fig. 2, Fig. 6), its smaller positive-VoI set does not come from missing useful disclosures, but from correctly withholding the support information where it would not help.

On `Email`, disclosure is harmful on average, and both estimators abstain on a large fraction of instances: `T-Learner` plateau is $\approx 66\%$, `ClassWise` is $\approx 53\%$. The more parsimonious policy of `ClassWise` attains the highest accuracy overall, at intermediate budgets (Fig. 2, Fig. 6). At large budgets, the additional disclosures of `T-Learner` yield a slightly higher accuracy, suggesting that `ClassWise` withholds a small set of instances on which disclosure would still have helped.

On `ImageNet-16H` disclosure is beneficial for almost every instance, and both estimators disclose almost up to the full budget: at $B = 1$, `ClassWise` discloses to 98% of the instances and `T-Learner` to nearly all of them. With few instances to withhold, the two estimators differ only in the *ordering* of the instances, not in the size of the positive-VoI set.

E.2 Q2: Does our learned policy improve human-AI team performance?

Take-home message: The learned policy matches or outperforms self-selection and is associated with greater confidence discrimination between correct and incorrect classifications. Furthermore, we note that human-selected disclosure only partially align with `ClassWise`-selected ones.

In the main paper, we assess differences in overall accuracy across the experimental conditions. Here, in addition to reporting the full results of that analysis, we examine participants' accuracy as a function of the correctness of the ML advice provided as auxiliary information, providing a more detailed picture of how differences in adherence may contribute to differences in overall ac-

curacy, particularly in the Email task. Furthermore, we assess the effects of *Support* and *Budget* on participants' confidence in their own classifications, exploring the potential benefits of a selective disclosure approach beyond classification accuracy alone. Finally, to provide a more complete comparison between ClassWise-selected and human-selected disclosure, we examine alignment between ClassWise-selected and human-selected disclosure.

Overall accuracy. We assess whether overall classification accuracy differs across the four experimental conditions by fitting a logistic mixed-effects model predicting the correctness of participants' classifications from *Support*, *Budget*, and their interaction. This analysis was presented in the main text and will not be further discussed, but see Table 5 for the full results.

Descriptive statistics

		ImageNet-16H	*Email*
Support condition	Budget condition	Accuracy	Accuracy
Human Policy	Low Budget	0.73 ± 0.15	0.82 ± 0.13
Human Policy	High Budget	0.81 ± 0.11	0.84 ± 0.13
Machine Policy	Low Budget	0.75 ± 0.11	0.81 ± 0.12
Machine Policy	High Budget	0.84 ± 0.09	0.84 ± 0.12

Regression coefficients

	ImageNet-16H				*Email*			
Fixed effect	OR	95% CI	z	p	OR	95% CI	z	p
Intercept	10.99	[7.72, 15.64]	13.31	$< .001$	7.66	[6.24, 9.39]	19.50	$< .001$
Support (HP)	0.76	[0.60, 0.97]	-2.25	.025	1.04	[0.82, 1.31]	0.30	.762
Budget (LB)	0.51	[0.40, 0.65]	-5.45	$< .001$	0.79	[0.63, 1.01]	-1.91	.056
Support $\times$ Budget	1.15	[0.71, 1.87]	0.57	.566	1.04	[0.65, 1.66]	0.15	.880

Table 5: Descriptive statistics (means and standard deviations) and results of the logistic mixed-effects regressions predicting classification accuracy from *Support*, *Budget*, and their interaction (OR: odds ratio).

Classification accuracy by correctness of support information. We explore how participants' accuracy varies as a function of the correctness of the support information provided. In the Email task, correctness was determined by whether the suggested class corresponded to the ground-truth label; in the ImageNet-16H task, it was determined by whether the ground-truth label was included in the prediction set. We focus on trials in which support information was available during classification and, to improve comparability between the MP and HP conditions, consider only items for which the policy would provide support information. Because the ML model prediction constituting the support information was generally accurate, trials in which it was incorrect were relatively rare ($75/1,227$ observations in the Email study and $86/1,827$ in the ImageNet-16H study). The subgroups pertaining to wrong-advice instances thus present few observations (as evident from the large error bars characterizing relative to such subgroups in Fig. 8), making inferential statistics potentially unreliable; we therefore limit our discussion to the descriptive patterns.

As shown in Fig. 8, the two studies exhibit different patterns. In the ImageNet-16H study, accuracy in MP relative to HP tends to be higher when the support information is correct and approximately comparable when it is incorrect. In the Email study, the opposite descriptive pattern emerges: accuracy in MP is, on average, lower than in HP when the support information is correct and higher when it is incorrect. These patterns broadly correspond to the differences in adherence between MP and HP observed across the two studies (discussed in the main text and in the following section), with the relative adherence on policy-provided information being higher in the ImageNet-16H study than in the Email study.

Furthermore, this pattern may help explain why a significant difference in accuracy between MP and HP was observed in the ImageNet-16H study but not in the Email study. Because the support information was correct on most trials, even a modest reduction in participants' tendency to follow it could result in an appreciable decrease in overall accuracy.

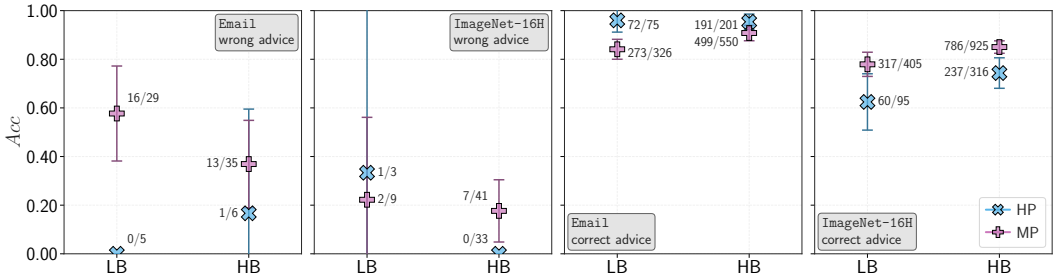


Figure 8: Participants' average accuracy by experimental condition and correctness of support information. Ratios indicate the number of correct participant responses out of the total number of observations in each subgroup.

Human confidence. We explore participants' confidence in the correctness of their final classifications using a linear mixed-effects model in which confidence on each trial is predicted by *Support*, *Budget*, the presence of support information, participants' classification accuracy, and all their interactions (see Table 6 for the full results). This analysis allows us to examine not only whether confidence differs across experimental conditions, but also whether such differences depend on whether participants' classifications are correct. In particular, greater confidence for correct than incorrect classifications can be interpreted as greater confidence discrimination.

In the `ImageNet-16H` study, we observe a significant interaction between *Support*, presence of support information, and participants' accuracy ($p = .031$). When support information was not received, the difference in confidence between correctly and incorrectly classified items was significantly greater in MP than in HP (MP: 2.99 ± 1.57; HP: 1.91 ± 1.29; $p < .001$). When support information was received, the corresponding difference was descriptively greater in MP than in HP, although this contrast did not reach significance after correction for multiple comparisons (MP: 2.06 ± 1.41; HP: 1.50 ± 1.23; $p = .088$).

We also observe a significant interaction between *Budget*, presence of support information, and participants' accuracy ($p = .003$). When support information was received, the difference in confidence between correct and incorrect classifications was significantly greater in HB than in LB (HB: 1.95 ± 1.37; LB: 1.61 ± 1.32; $p = .043$). In contrast, when support information was not received, the corresponding difference did not significantly differ between HB and LB (HB: 2.09 ± 1.62; LB: 2.62 ± 1.42; $p = .110$). One possible explanation is that, under the higher budget, support information was provided for a broader range of items, potentially including easier ones.

Instead, in the `Email` study, the only significant interaction involving participants' accuracy (and thus directly relevant to confidence discrimination) was the interaction with *Support* ($p = .005$). Specifically, the difference in confidence between correctly and incorrectly classified items was significantly greater in MP than in HP (MP: 0.71 ± 0.86; HP: 0.43 ± 0.89; $p = .005$).

Overall, across the two studies, participants in MP tended to show greater confidence discrimination than those in HP, assigning relatively higher confidence to correct than to incorrect classifications. Notably, this pattern was also observed in the `Email` study, in which the selective disclosure method did not significantly increase overall classification accuracy.

Alignment between `ClassWise`-selected and human-selected disclosure. We explore alignment between `ClassWise`-selected and human-selected information requests from two perspectives. First, we analyze differences between the frequency of information disclosure by the policy and humans using one-sample t-tests. Participants in the HP conditions request auxiliary information on fewer trials than the method would provide it in both the `ImageNet-16H` and `Email` tasks (see Table 7).

Second, focusing on trials in which HP participants requested information, we descriptively examine how often these requests correspond to items for which `ClassWise` would also have provided information. In the `ImageNet-16H` task, the proportion of human-requested trials that were also selected by `ClassWise` was 0.40 ± 0.30 in LB and 0.87 ± 0.13 in HB. In the `Email` task, the corresponding proportions were 0.31 ± 0.26 and 0.56 ± 0.20, respectively. These proportions indicate

Descriptive statistics

			ImageNet-16H		*Email*	
Support	Budget	Information	Incorrect	Correct	Incorrect	Correct
Human Policy	Low Budget	Not received	4.22 ± 1.45	6.30 ± 0.67	5.20 ± 1.29	5.56 ± 0.99
Human Policy	Low Budget	Received	3.24 ± 1.49	4.58 ± 1.45	5.32 ± 1.30	5.67 ± 0.98
Human Policy	High Budget	Not received	4.74 ± 1.60	6.45 ± 0.50	5.24 ± 1.28	5.69 ± 0.89
Human Policy	High Budget	Received	3.08 ± 1.46	4.75 ± 1.48	5.00 ± 1.31	5.42 ± 1.19
Machine Policy	Low Budget	Not received	3.11 ± 1.45	6.22 ± 0.67	4.82 ± 1.28	5.59 ± 0.87
Machine Policy	Low Budget	Received	3.68 ± 1.68	5.58 ± 1.08	5.11 ± 1.26	5.52 ± 0.93
Machine Policy	High Budget	Not received	3.84 ± 1.96	6.63 ± 0.42	4.74 ± 1.16	5.52 ± 0.78
Machine Policy	High Budget	Received	3.61 ± 1.59	5.79 ± 0.87	4.77 ± 1.36	5.62 ± 0.95

Regression coefficients

	ImageNet-16H				*Email*			
Fixed effect	b	SE	t	p	b	SE	t	p
Intercept	4.504	.096	46.79	$< .001$	5.168	.069	75.19	$< .001$
Support (HP)	.732	.150	4.89	$< .001$	.440	.134	3.30	.001
Budget (LB)	-.383	.150	-2.56	.011	.064	.133	.48	.632
Information received	-.062	.093	-.67	.503	-.099	.082	-1.22	.223
Accuracy	1.480	.075	19.81	$< .001$	.410	.047	8.66	$< .001$
Support × Budget	.192	.297	.65	.518	-.134	.266	-.50	.615
Support × Information	-.989	.178	-5.56	$< .001$	-.403	.165	-2.44	.015
Budget × Information	.544	.175	3.10	.002	.138	.161	.86	.391
Support × Accuracy	-.716	.136	-5.27	$< .001$	-.362	.093	-3.91	$< .001$
Budget × Accuracy	.261	.136	1.92	.055	-.087	.092	-.95	.344
Information × Accuracy	-.438	.103	-4.24	$< .001$	.083	.089	.94	.349
Support × Budget × Information	-.122	.348	-.35	.727	.449	.320	1.40	.160
Support × Budget × Accuracy	-.049	.270	-.18	.856	-.066	.184	-.36	.721
Support × Information × Accuracy	.428	.199	2.15	.031	.225	.178	1.26	.207
Budget × Information × Accuracy	-.587	.197	-2.98	.003	-.065	.176	-.37	.710
Support × Budget × Information × Accuracy	-.180	.390	-.46	.646	-.032	.350	-.09	.928

Post-hoc contrasts

	ImageNet-16H				*Email*			
Contrast	Estimate	SE	z	p	Estimate	SE	z	p
HP vs. MP, information not received	-.716	.136	-5.27	$< .001$		—		
HP vs. MP, information received	-.288	.143	-2.02	.088		—		
LB vs. HB, information not received	.261	.136	1.92	.110		—		
LB vs. HB, information received	-.326	.142	-2.30	.043		—		
HP vs. MP		—			-.249	.089	-2.79	.005

Table 6: Descriptive statistics and results of the linear mixed-effects models predicting confidence ratings from *Support*, *Budget*, presence of information, classification accuracy, and their interactions. Descriptive statistics represent participant-level means ± standard deviations. Post-hoc contrasts compare confidence discrimination, defined as the difference in confidence between correct and incorrect classifications (post-hoc contrasts are reported only for significant effects involving classification accuracy; dashes indicate that no post-hoc contrast was conducted because the corresponding effect was not significant)

.

that human requests only partially overlap with policy-selected disclosure, although the extent of this overlap differs considerably across tasks and budget conditions.

Taken together, these analyses indicate that humans and `ClassWise` allocate disclosure differently. Participants request information on fewer trials than the policy provides it, which may limit the potential benefits of human-selected disclosure when useful assistance remains unrequested. Determining why participants request assistance relatively infrequently (e.g., because of overconfidence in their own classification ability or uncertainty about the value of assistance) requires further study, as the present user studies are designed to compare `ClassWise`-selected and human-selected dis-

closure at the level of the complete strategies rather than identify the mechanisms underlying human request behavior. In addition, the descriptive overlap analysis indicates that the two strategies do not necessarily select the same trials for disclosure. These exploratory results therefore suggest that differences between policy-selected and human-selected disclosure concern both how frequently information is requested and which items receive it, motivating future work on the consequences of these differences for human-AI team performance.

Study	Budget	Requested assistance	MP assistance	95% CI	*t*	*df*	*p*
`ImageNet-16H`	Low	3.73 ± 2.12	6	[3.21, 4.25]	-8.70	65	$< .001$
`ImageNet-16H`	High	5.99 ± 3.20	14	[5.22, 6.75]	-20.82	68	$< .001$
`Email`	Low	3.93 ± 1.97	5	[3.47, 4.39]	-4.60	71	$< .001$
`Email`	High	6.41 ± 4.63	9	[5.25, 7.56]	-4.48	63	$< .001$

Table 7: Number of assistance requests made by participants in the Human-Policy condition compared with the number of trials on which assistance was provided in the corresponding Machine-Policy condition. Requested assistance is reported as mean $\pm$ standard deviation. Confidence intervals and *t* tests refer to one-sample tests comparing the observed number of requests with the number of assistance opportunities provided in the corresponding MP condition.

E.3 Q3: How do participants interact with our policy?

Take-home message: The learned policy is more beneficial whenever humans follow the advice, and self-reported trust in the AI system is broadly consistent with advice-adherence trends.

In addition to the analyses of advice adherence discussed in the main text, for which we report full details here, we analyze participants' responses to the trust scale, providing a complementary, self-reported perspective on their trust in the system providing auxiliary information.

Advice adherence. We assess participants' adherence to the AI advice provided as support information by fitting a logistic mixed-effects model predicting whether participants' classifications are consistent with the AI advice from *Support*, *Budget*, and their interaction. This analysis was presented in the main text and will not be further discussed, but see Table 8 for the full results. [16]

Trust in the AI system. We analyze the trust index, computed by averaging participants' ratings across the eight items of the trust scale, using linear models in which trust index values are predicted by *Support* condition, *Budget* condition, and their interaction (see Table 9 for the full results).

In the `ImageNet-16H` study, trust index values do not significantly differ between MP and HP (MP: 3.06 ± 0.73; HP: 2.90 ± 0.75; $p = .064$). while they are significantly higher in HB than in LB (HB: 3.13 ± 0.70; LB: 2.83 ± 0.75; $p < .001$), while the interaction between *Support* and *Budget* is not significant ($p = .096$). In contrast, in the `Email` study, trust index values are significantly lower in MP than in HP (MP: 3.04 ± 0.82; HP: 3.32 ± 0.70; $p = .003$). HB and LB do not significantly differ ($p = .861$), whereas the interaction between *Support* and *Budget* is significant ($p = .046$). Follow-up comparisons indicate that the difference between MP and HP is significant under LB (MP-LB: 2.95 ± 0.81; HP-LB: 3.40 ± 0.70; $p = .002$), but not under HB (MP-HB: 3.15 ± 0.82; HP-HB: 3.23 ± 0.70; $p = 1$).

This pattern is broadly consistent with the results on adherence on support information: participants in MP reported similar trust in the AI system to those in HP in the `ImageNet-16H` study, whereas in the `Email` study, particularly under the lower budget, participants in MP reported lower values than those in HP.

Counterfactual benchmark. We report here the results of the counterfactual analysis, in which we assess the performance that MP participants would have achieved had their predictions been over-

[16] Because adherence is operationalized differently in the `ImageNet-16H` and `Email` tasks, due to the nature of the advice provided, we repeat the `ImageNet-16H` analyses using only trials with singleton prediction sets. On these trials, the advice specifies a single label and is therefore directly comparable to that provided in the `Email` task (moreover, singleton sets represents the majority of cases, accounting for 64% and 65% of the observations in the analyzed LB and HB subsets, respectively). The pattern of results remains unchanged.

Descriptive statistics			
		ImageNet-16H	*Email*
Support condition	Budget condition	Proportion of agreement with support information	Proportion of agreement with support information
Human Policy	Low Budget	0.92 ± 0.24	0.96 ± 0.17
Human Policy	High Budget	0.94 ± 0.13	0.95 ± 0.12
Machine Policy	Low Budget	0.91 ± 0.12	0.81 ± 0.17
Machine Policy	High Budget	0.92 ± 0.10	0.89 ± 0.14

Regression coefficients								
	ImageNet-16H				*Email*			
Fixed effect	OR	95% CI	z	p	OR	95% CI	z	p
Intercept	48.02	[25.31, 91.13]	11.85	$< .001$	25.08	[13.88, 45.30]	10.68	$< .001$
Support condition (HP)	1.72	[0.93, 3.18]	1.73	.084	4.63	[2.14, 10.04]	3.89	$< .001$
Budget condition (LB)	0.70	[0.37, 1.31]	-1.13	.260	0.83	[0.38, 1.81]	-0.47	.636
Support $\times$ Budget	0.59	[0.18, 1.97]	-0.86	.392	3.31	[0.72, 15.29]	1.53	.126

Table 8: Descriptive statistics and results of the logistic mixed-effects regressions predicting participants' agreement with the support information from *Support*, *Budget*, and their interaction (OR: odds ratio).

Descriptive statistics			
		ImageNet-16H	*Email*
Support condition	Budget condition	Trust index	Trust index
Human Policy	Low Budget	2.67 ± 0.72	3.40 ± 0.70
Human Policy	High Budget	3.12 ± 0.71	3.23 ± 0.70
Machine Policy	Low Budget	2.98 ± 0.75	2.95 ± 0.81
Machine Policy	High Budget	3.14 ± 0.69	3.15 ± 0.82

Regression coefficients								
	ImageNet-16H				*Email*			
Fixed effect	b	SE	t	p	b	SE	t	p
Intercept	2.978	0.044	68.37	$< .001$	3.182	0.046	69.12	$< .001$
Support (HP)	-0.162	0.087	-1.86	.064	0.273	0.092	2.96	.003
Budget (LB)	-0.303	0.087	-3.48	$< .001$	-0.016	0.092	-0.18	.861
Support $\times$ Budget	-0.291	0.174	-1.67	.096	0.369	0.184	2.00	.046

Post-hoc comparisons: `Email`				
Contrast	Estimate	SE	t	p
HP-LB vs. MP-LB	0.457	0.127	3.61	.002
HP-LB vs. HP-HB	0.168	0.130	1.29	1
HP-LB vs. MP-HB	0.257	0.130	1.98	.294
MP-LB vs. HP-HB	-0.289	0.131	-2.21	.167
MP-LB vs. MP-HB	-0.201	0.130	-1.54	.745
HP-HB vs. MP-HB	0.088	0.134	0.66	1

Table 9: Descriptive statistics and results of the linear models predicting trust index values from *Support*, *Budget*, and their interaction. Descriptive statistics represent means $\pm$ standard deviations. Post-hoc comparisons for the `Email` study are Bonferroni-corrected for six comparisons.

ruled with the ML advice when provided (for the `ImageNet-16H` task, we convert each prediction set into a single prediction by retaining its highest-scoring class, which is treated as the counterfac-

tual response). This analysis was presented in the main text and will not be further discussed, but see Table 10 for the full results. [17]

Descriptive statistics

	ImageNet-16H	*Email*
Support condition	Accuracy	Accuracy
Human Policy	0.77 ± 0.13	0.83 ± 0.13
Machine Policy	0.80 ± 0.11	0.82 ± 0.12
Machine Policy – Counterfactual benchmark	0.82 ± 0.11	0.85 ± 0.10

Regression coefficients

	ImageNet-16H				*Email*			
Fixed effect	OR	95% CI	z	p	OR	95% CI	z	p
Intercept	14.40	[10.01, 20.71]	14.38	$< .001$	9.18	[7.35, 11.48]	19.49	$< .001$
Support - HP	0.59	[0.44, 0.80]	-3.40	$< .001$	0.87	[0.63, 1.19]	-0.87	.385
Support - MP	1.05	[0.84, 1.32]	0.43	.668	0.82	[0.66, 1.03]	-1.71	.087
Budget (LB)	0.44	[0.35, 0.56]	-6.90	$< .001$	0.76	[0.59, 0.97]	-2.23	.026
Support - HP $\times$ Budget	1.41	[0.77, 2.57]	1.12	.264	1.12	[0.59, 2.12]	0.35	.727
Support - MP $\times$ Budget	1.07	[0.68, 1.70]	0.31	.758	1.00	[0.64, 1.57]	0.02	.983

Post-hoc comparisons

	ImageNet-16H				*Email*			
Contrast	OR	SE	z	p	OR	SE	z	p
HP vs. MP	0.75	0.09	-2.33	.059	1.03	0.13	0.21	1
HP vs. MPCFT	0.61	0.08	-4.00	$< .001$	0.79	0.10	-1.84	.196
MP vs. MPCFT	0.81	0.07	-2.34	.058	0.77	0.06	-3.25	.004

Table 10: Descriptive statistics and results of the logistic mixed-effects regressions comparing observed accuracy in HP and MP with counterfactual accuracy under full advice adherence on automatically disclosed support information (MPCFT). Descriptive statistics represent participant-level means $\pm$ standard deviations. Post-hoc comparisons are averaged across budget conditions and Bonferroni-corrected for three comparisons (*OR*: odds ratio).

E.4 Q4: How does changing the information content affect our policy?

Take-home message: The benefit of selective disclosure depends on how informative the support information is. At $\alpha = 0.05$, where prediction sets are small and informative, `ClassWise` prioritizes them at every budget; as α decreases and uninformative full sets become frequent, the gain of any disclosure policy shrinks, and at $\alpha = 0.01$ no policy separates from full disclosure.

In the `ImageNet-16H` setting the support information S is the conformal prediction set $C(\mathbf{x})$, whose size is governed by the miscoverage level α: smaller α enforces higher coverage and therefore yields larger sets. The size of $C(\mathbf{x})$ directly controls how informative disclosure can be. In particular, a set that coincides with the full label space, $|C(\mathbf{x})| = |\mathcal{Y}| = 16$, is compatible with every class and thus carries *no* discriminative signal: it cannot help the decision-maker refine their judgement, so its Value of Information is non-positive and a well-behaved policy should never spend budget disclosing it. The prevalence of such full sets consequently upper-bounds the benefit attainable at a given α.

We examine the role of α along three axes: (i) we analyse how the distribution of set sizes varies with α, motivating the choice of $\alpha = 0.05$ in the main paper; (ii) we report the budget–accuracy curves for the remaining levels $\alpha \in \{0.01, 0.02\}$ and comment on how the achievable benefit shrinks as sets grow; and (iii) we inspect *which* sets our policy actually chooses to disclose as a function of the budget, showing that at $\alpha = 0.05$ it essentially never wastes budget on uninformative full sets.

[17] Also in this case, restricting the `ImageNet-16H` analyses to trials in which the AI advice consists of a singleton prediction set leaves the pattern of results broadly unchanged. The only difference is that the contrast between MP and MPCFT becomes significant, favoring the latter especially in HB.

Table 11: Distribution of conformal prediction set sizes on the `ImageNet-16H` dataset ($n = 1200$ images, $|\mathcal{Y}| = 16$ classes) for the three miscoverage levels α considered in our experiments. Smaller α enforces higher coverage and thus yields larger sets; note in particular the mass on the full set ($|C(\mathbf{x})| = 16$), which carries no information for the decision-maker. Dashes denote empty bins.

	Set size $\lvert C(\mathbf{x})\rvert$																Summary	
α	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Mean	Sing.%
0.01	664	92	32	9	10	3	1	1	–	1	–	–	–	–	–	387	6.05	55.3
0.02	735	148	54	31	16	13	9	8	4	4	2	1	–	5	1	169	3.75	61.3
0.05	804	200	99	55	24	13	3	–	–	–	–	–	–	–	–	2	1.64	67.0

(i) Set-size distribution and the choice of α. Table 11 reports the distribution of set sizes on the `ImageNet-16H` dataset for the three miscoverage levels considered in our experiments. As expected, the smaller α, the larger the prediction sets: the mean set size shrinks from 6.05 at $\alpha = 0.01$ to 1.64 at $\alpha = 0.05$. The focus for our purposes is not the exact size but whether the disclosed set is *informative*: a large set provides little guidance to the decision-maker, and in the limit the full set ($|C(\mathbf{x})| = 16$) carries no signal at all. From this perspective the three levels differ sharply.

At $\alpha = 0.05$ the support information is almost always useful: essentially all sets have size below 7, and only 2 out of 1200 images fall on large, uninformative sets ($|C(\mathbf{x})| > 10$). At $\alpha = 0.01$ and $\alpha = 0.02$, instead, a substantial fraction of the sets is large and uninformative (the full set alone accounts for 387 and 169 images respectively) so a non-negligible share of any disclosed feedback would be effectively useless. This is precisely why we adopt $\alpha = 0.05$ in the main paper: it is the level at which disclosure is almost always informative, so that revealing a set genuinely helps the human decision rather than wasting budget on redundant, uninformative feedback.

(ii) Disclosure accuracy across miscoverage levels. Fig. 9 reports the budget–accuracy curves for $\alpha \in \{0.05, 0.02, 0.01\}$, complementing the results in the main paper (Fig. 2). A first effect of lowering α is visible already at the endpoints: as the prediction sets grow larger and less discriminative, the average benefit of full disclosure shrinks. Full disclosure (`FullDisc`) attains about 80.5% at both $\alpha = 0.01$ and $\alpha = 0.02$, i.e. only about 3.3 points above the no-disclosure accuracy (`NoDisc`$\approx 77.2\%$), against the 8.3-point gap observed at $\alpha = 0.05$. Larger sets thus carry less usable signal, and the ceiling that any disclosure policy can reach is correspondingly lower.

At $\alpha = 0.02$, selective disclosure remains effective: `ClassWise` reaches the accuracy of full disclosure already at $B = 0.5$, whereas `Confidence` needs $B = 0.8$ and `Random` the full budget. Both policies then slightly exceed `FullDisc` at higher budgets (`ClassWise` $\approx 81.0\%$ at $B = 0.7$, `Confidence` $\approx 81.3\%$ at $B = 0.9$), although these gains are within the confidence bands. At $\alpha = 0.01$, instead, no policy separates from full disclosure: `ClassWise`, `Confidence` and even `Random` track `FullDisc` within their confidence bands, and `ClassWise` falls slightly below it at high budgets ($\approx 79.3\%$ at $B = 0.9$, about three images).

We attribute this behaviour to the set-size distribution at $\alpha = 0.01$ (Table 11). When α is very small, the prediction sets are large: the mean set size is 6.05, and about one third of the images ($387/1200$) receive the full, completely uninformative set. First, fewer instances receive an informative set, so the mass of instances with positive VoI is smaller and a selective policy has less signal to exploit. Second, estimating the VoI becomes harder: the risk models must tell apart many large, overlapping sets whose effect on the human decision is weak and noisy. Ranking errors then cause the policy to withhold informative sets in favour of uninformative ones, which explains why it can fall slightly below full disclosure at high budgets. Overall, these results indicate that the benefit of selective disclosure depends on how informative the support information is: when it carries little signal, as at $\alpha = 0.01$, there is little for any policy to gain over simply disclosing everything.

(iii) Does the policy disclose informative sets? Since a large prediction set carries little signal, a good disclosure policy should spend its budget on small, informative sets and avoid the uninformative full set ($|C(\mathbf{x})| = 16$) whenever possible. Here we ask whether the `ClassWise` policy does so, and how this behaviour changes with the budget. Fig. 10 shows, for each budget B, the

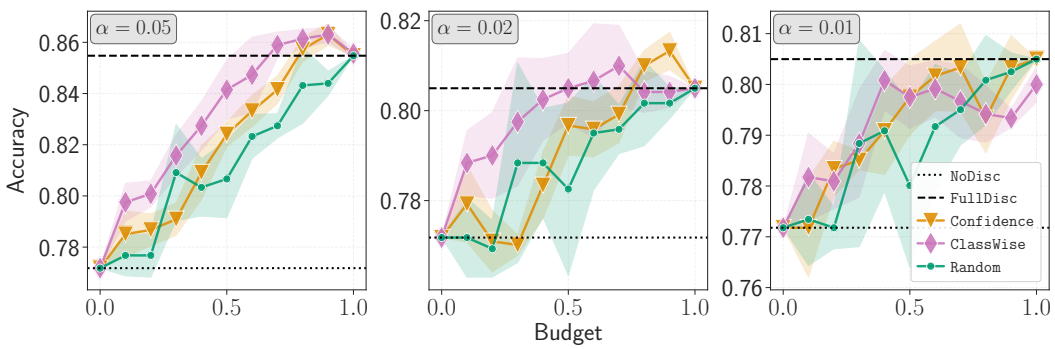


Figure 9: Budget–accuracy curves on the `ImageNet-16H` dataset for the two additional miscoverage levels $\alpha \in \{0.01, 0.02\}$. As α decreases the prediction sets grow larger and less informative, lowering the average benefit of full disclosure (`FullDisc`$\approx 80.5\%$ against 85.5% at $\alpha = 0.05$).

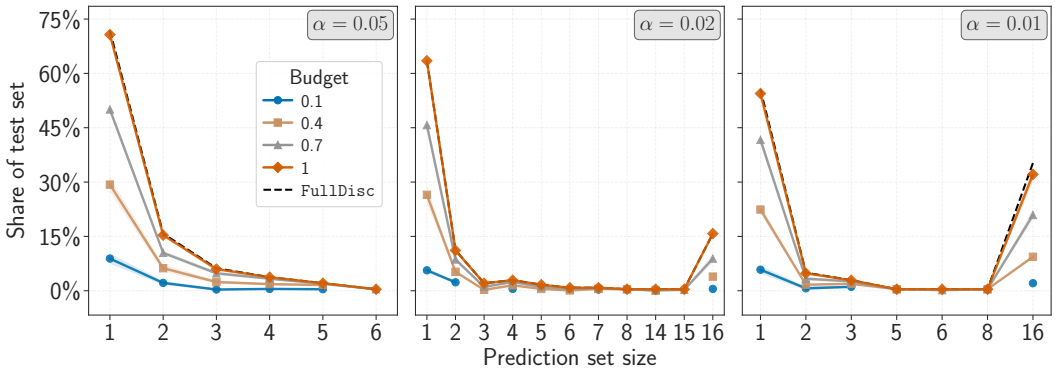


Figure 10: Prediction-set sizes disclosed by the `ClassWise` policy on `ImageNet-16H`, at $\alpha \in \{0.05, 0.02, 0.01\}$. For each set size, the curves show the share of the test set that the policy discloses at budgets $B \in \{0.1, 0.4, 0.7\}$ (mean over 5 seeds, with 95% confidence intervals). The dashed line (`FullDisc`) is the share of the test set with that set size, i.e., what full disclosure would reveal, so the gap between a curve and the dashed line gives the instances of that size the policy withholds.

fraction of test instances of each set size to which the policy discloses the support information: a policy indifferent to set size would disclose a fraction B of the instances of every size.

At $\alpha = 0.05$, the policy clearly favours small sets at every budget. Full sets are essentially absent, since only 2 exist in the whole dataset. At $\alpha = 0.02$ and $\alpha = 0.01$, the picture becomes budget-dependent. At small budgets the policy still prioritizes small sets, showing that the estimated VoI ranks informative sets first. As the budget grows, however, the policy increasingly discloses full sets as well, reaching on average 77 of the 85 full sets in the test set (91%) at $B = 1$ for $\alpha = 0.01$, and all 38 of the 38 (100%) full sets for $\alpha = 0.02$.

These results confirm $\alpha = 0.05$ as the most favourable operating point: the regime in which the policy can disclose informative feedback at *every* budget level. These motivate its use in the main paper.

F Extended Related Work

Learning to defer and its extensions. (Madras et al., 2018) introduce Learning to Defer (LtD), casting human-AI collaboration as a mutually exclusive choice: for each instance, either the model predicts or the case is deferred to a human expert (for a broader overview of rejection and deferral, see Ruggieri & Pugnana (2025)). The framework has since been developed along several axes. On the theoretical side, (Mozannar & Sontag, 2020) derive consistent surrogate losses for the joint model-plus-deferral objective, (Verma & Nalisnick, 2022) propose a one-vs-all formulation with

improved calibration of the deferral decision, and (Mozannar et al., 2023) study exact algorithms for learning who should predict. For further work developing LtD methods with theoretical guarantees, see Okati et al. (2021); Charusaie et al. (2024); Cao et al. (2023); Liu et al. (2024); Gao & Yin (2025); Li et al. (2017); Montreuil et al. (2025a; 2026); Fang & Nalisnick (2026). Other theoretical extensions of the LtD approach involved considering multiple-expert settings (Verma et al., 2023; Mao et al., 2023; Montreuil et al., 2025b; Mao et al., 2025; Zhang et al., 2026; Liu et al., 2026), multi-task settings (Pugnana et al., 2025; Montreuil et al., 2025c), and aspects such as rejection of unexplainable decisions (Stradiotti et al., 2025).

On the empirical side, (Hemmer et al., 2023) examine how delegation affects task performance and satisfaction with real participants, (Palomba et al., 2025) evaluate deferring systems through a causal lens, and (Bondi et al., 2022) investigate how communicating a deferral affects human accuracy. Other works have examined the impact class distribution on rejected instances (Pugnana et al., 2024), with Pesenti et al. (2026) examining how class imbalances may affect users' interactions with LtD systems. Deferral and rejection have also been studied in healthcare (Strong et al., 2025; Kompa et al., 2021), vehicle engineering (Hendrickx et al., 2021), and question answering (Montreuil et al., 2026). The defining feature of LtD is that authority over the final decision is allocated: on deferred instances the model abstains, and on the remaining ones the human is bypassed entirely. Our setting is structurally different. The human is *always* the decision-maker, and what is allocated is not authority but *information*: the policy decides what the human sees, never what the human decides. Consequently our risk is defined over human actions in two information regimes (Eq. (1)), rather than over a model prediction and a human prediction.

AI-assisted decision-making. A second line of work keeps the human as the final decision-maker and asks what form the support should take. Prediction sets are a prominent example: (Straitouri & Rodriguez, 2024) design decision-support systems based on counterfactual prediction sets, (De Toni et al., 2024) construct prediction sets that target the expert's accuracy rather than coverage, and (Cresswell et al., 2024) show, in a pre-registered randomized controlled trial, that conformal prediction sets improve human accuracy over fixed-size (top-k) sets with the same coverage. (Schemmer et al., 2023) instead conceptualize appropriate reliance on AI advice and study how explanations affect it. Broader syntheses are provided by (Punzi et al., 2024), who survey learning paradigms for hybrid decision-making systems, and by (Hemmer et al., 2025), who conceptualize human-AI complementarity and review its sources and the empirical evidence for it.

Closer to our setting, a smaller body of work decides, instance by instance, whether (and which) support to show. (Noti & Chen, 2023) study a regression task (pretrial risk assessment) and learn from past human predictions a policy that, given the case, the algorithmic risk score and the human's initial estimate, reveals the score only when it is predicted to be more accurate than that estimate; in a large-scale experiment, this improves human predictions over always showing the score. (Ma et al., 2023) do not learn the disclosure rule itself: they estimate the decision-maker's correctness likelihood on each instance by applying an approximation of their individual decision rules to similar labelled cases, and compare it with the AI's calibrated confidence; whenever the human is predicted to be more likely correct, the AI recommendation is either withheld (only its explanation is shown) or revealed only after an independent human judgement. Other works choose among several forms of support. (Bhatt et al., 2025) cast this choice as a stochastic contextual bandit and learn online, separately for each new decision-maker, a policy that selects, for each input, the form of support (e.g., none, a model or LLM prediction, expert consensus) expected to minimize that individual's error. (Buçinca et al., 2024) instead apply offline reinforcement learning to previously collected interaction data to choose among no assistance, explanation only, recommendation with explanation, and on-demand advice, optimizing for immediate accuracy, for the decision-maker's learning, or for both; their policies condition on a discrete state that includes the decision-maker's need for cognition and task knowledge and, as a proxy for AI uncertainty, the ground-truth correctness of the AI recommendation. In the symmetric direction, (Pugnana et al., 2025) propose Learning to Ask, where a model decides under a budget when to query a human for enriched feedback, and characterize the optimal querying rule as a threshold on the risk difference between a standard and an enriched predictor; relatedly, (Wilder et al., 2020) and (Charusaie et al., 2024) train predictors that complement, or directly incorporate, human decisions.

Compared with the approaches that adapt the support shown to the human, LSD differs in three respects. First, disclosure is subject to a budget, which makes the optimal policy a non-negative,

budget-dependent threshold on VoI; without a budget and with only two actions, the optimal policy of (Bhatt et al., 2025) reduces to disclosing whenever an individual-level VoI is positive, i.e., the unconstrained ($B = 1$) case of our threshold rule. Second, VoI is the causal reduction in human decision risk induced by disclosure, and therefore depends on how decision-makers actually use the disclosed information, whereas (Noti & Chen, 2023) and (Ma et al., 2023) condition disclosure on whether the AI is predicted to be more accurate than the human alone. Third, we establish when VoI is identifiable from data collected under the two regimes, and bound both the degradation of the resulting plug-in policies relative to no disclosure and their regret relative to the optimal policy.

Active feature acquisition. In Active Feature Acquisition (AFA), an agent decides which missing feature values to acquire, at a cost, in order to improve a downstream predictive model (Janisch et al., 2020; Rahbar et al., 2025; Saar-Tsechansky et al., 2009; Ji & Carin, 2007). (Saar-Tsechansky et al., 2009) acquire feature values at training time to improve model induction, whereas at test time (Ji & Carin, 2007) formulate cost-sensitive acquisition and classification jointly and (Janisch et al., 2020) cast classification with costly features as a sequential decision problem solved by reinforcement learning. (Rahbar et al., 2025) provide a unified view, organizing existing methods into embedded cost-aware predictors, model-based approaches, model-free policies, and hybrid strategies. Related in spirit are (Ma et al., 2019) and (Gong et al., 2019), who score candidate acquisitions with information-theoretic criteria. Our VoI instead follows the classical decision-theoretic notion of value of information as the expected reduction in decision loss (Howard, 1966), with one twist: for a Bayesian decision-maker this quantity is never negative, whereas ours is measured on actual human decisions and can be. Two differences matter. First, the acquisition target is a *model*'s prediction rather than a *human*'s decision, so the relevant risk is a model risk and the counterfactual "what would the predictor do without this feature" is directly computable, whereas the human counterfactual is not; this is precisely what forces the causal treatment of Appendix B. Second, AFA chooses *which* features to acquire, often sequentially and under a per-instance budget, whereas LSD makes a single binary decision on an indivisible block of support information, under a budget on the expected disclosure rate across cases.

Policy learning under budget constraints. The causal reading of our framework connects it to *optimal policy learning* (OPL), which studies how to learn, from data, treatment assignment rules that maximize a welfare objective (Manski, 2004; Athey & Wager, 2021; Cerulli, 2026). Without constraints, the first-best rule treats every unit whose conditional average treatment effect (CATE) is positive; (Kitagawa & Tetenov, 2018) propose empirical welfare maximization over constrained classes of treatment rules, (Bhattacharya & Dupas, 2012) show that when a budget caps the fraction of treated units the optimal rule treats those whose CATE exceeds a quantile threshold, and (Carranza & Athey, 2025) extend policy learning to observational data from multiple sources using doubly robust estimators. Our Theorem 1 recovers a threshold rule of the same shape, with the VoI in the role of the CATE and the budget B in the role of the capacity constraint. As in budget-constrained OPL, the threshold is non-negative, so disclosure is withheld whenever it is expected to harm the decision, even if budget remains; in our setting this case is far from marginal, since support information can mislead the decision-maker, e.g., by inducing over-reliance. The substantive difference is where the treatment acts. In standard OPL the treatment is applied to the unit whose outcome is measured. Here the treatment (disclosure) is applied to the *decision-maker*, while the outcome is the correctness of a decision about a *case*. This has two consequences we make explicit in Appendix B: the missing-potential-outcome problem arises at the level of the decision episode rather than the case, and is addressed by assigning different decision-makers to the two regimes, which requires those observed under each regime to be exchangeable; and SUTVA must be stated over episodes, since what must not interfere is the information disclosed to a given decision-maker in a given episode.

Heterogeneous treatment effect estimation. Since the VoI is the negative conditional average treatment effect (CATE) of disclosure on decision loss (Section 3.2), estimating it connects our work to the literature on heterogeneous treatment effects (Nogueira et al., 2022). Meta-learners reduce CATE estimation to standard supervised learning, either by fitting one outcome model per treatment arm (T-learner) or a single model with the treatment as input (S-learner) (Künzel et al., 2019); more refined estimators include the R-learner (Nie & Wager, 2021), doubly robust learners (Kennedy, 2023), causal forests (Wager & Athey, 2018) and neural architectures sharing representations across

arms (Shalit et al., 2017; Shi et al., 2019). In our setting, doubly robust corrections are less critical than in observational studies, since each case is observed under both regimes and disclosure is assigned independently of the case covariates. Moreover, accurate effect estimation is neither necessary nor sufficient for good causal decisions (Fernández-Loría & Provost, 2022): (Frauen et al., 2025) show that thresholding CATE estimators trained for estimation accuracy can yield suboptimal policies, since accuracy away from the decision boundary is irrelevant to the decision, and (Arno et al., 2026) show that, when individuals must be prioritized, recovering the ranking of treatment effects is easier than estimating their magnitude, and propose to learn it directly. Our setting requires both: under a binding budget, the optimal policy of Theorem 1 depends only on the ranking of the VoI, while the non-negative threshold depends on its sign. These are precisely the properties on which our plug-in guarantees (Section 3.3) and our ablation (Appendix E.1) focus.