Human-AI Teaming Through the Lens of Calibration

Eric Nalisnick
Department of Computer Science
Johns Hopkins University
nalisnick@jhu.edu

Chi Zhang*
Department of Computer Science
Johns Hopkins University
czhan168@jhu.edu

Sophia Qian*
Department of Computer Science
Johns Hopkins University
cqian17@jh.edu

Yixin Wang
Department of Statistics
University of Michigan
yixinw@umich.edu

Abstract

We study models for human-AI teaming through the lens of statistical calibration. We assume the team consists of an AI model and human—both of which are calibrated with respect to some partitioning of the feature space—and expose how the calibration assumptions propagate into the teaming framework. In particular, we consider frameworks that either (i) combine human and model predictions or (ii) delegate prediction responsibility to either a human or model. We show via theoretical and empirical results that existing methods for combination do not preserve the human's degree of calibration. Methods for delegation (by the very act of delegation) preserve calibration of the downstream predictors but shift the burden onto the rejector meta-model that decides who predicts. The rejector must be calibrated finely enough to locate where each member is superior, a demand that grows with the human's expertise and becomes unattainable when the human relies on information the system cannot observe.

1 Introduction

Modern AI systems can now extract insights from large scale data sources in ways that far surpass human abilities. However, humans still can contribute meaningful contextual information [1, 41, 58]. For instance, a medical AI system can inform its diagnosis from millions of scientific publications and patient health records. Yet this AI system cannot know a particular patient as well as a doctor who has been caring for this patient for years. This complementarity in information—population-level trends vs nuanced, specific insights—motivates the use of *hybrid intelligent* systems [22, 31]. Ideally, these systems leverage human and machine decision-making such that the human-AI team's performance is better than either could achieve alone. This essential property is called *complementarity* [17, 62, 66].

Human-AI teams have been observed to have complex dynamics, and a full characterization of when teaming succeeds and fails is still an open problem [61]. For example, Bansal et al. [2] reported a user study in which improving the AI model actually decreased team performance. Thus one would hope for the team to exhibit certain robustness properties [54]: for instance, the team's predictions should never be worse than its best member's. Or in frameworks that delegate responsibility [44], we would never want the act of delegation to depend on having a full characterization of the behavior of all team members. If it did, then we could just use the delegation system to make predictions and disband the team. While work has attempted to address these questions empirically [39] or theoretically in

*Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

specific settings [11, 17, 24, 50, 66], no prior work has addressed these questions across multiple teaming settings from a unified theoretical perspective.

In this work, we use statistical calibration as a unified theoretical lens through which to study human-AI teaming. We consider two teaming frameworks: combining human and model predictions [34] and delegating prediction tasks to one team member [44]. We assume that the team is composed of one model and one human—both of which are calibrated for some partitioning of the feature space. Calibration is a powerful tool for studying teaming as it allows for a general representation of the human and model's prediction quality. As the partitioning becomes increasingly fine-grained, the human / model approaches the Bayes predictor. As the partitioning becomes increasingly coarse, the human / model approaches marginal calibration (i.e. uniform confidence of $1/K$ for balanced K-class classification). Through a series of theoretical results, simulations, and experiments with authentically human-made predictions, we expose how the calibration assumptions propagate into the team's predictive behavior. In particular, we arrive at mostly *negative* results that show common teaming procedures do not preserve the underlying calibration of the team members. To summarize our contributions, we:

- Prove that model-based combination preserves calibration w.r.t. the model (Proposition 3.1) but not w.r.t. the human (Theorem 3.2). We verify this result via simulation (Figure 1).
- Prove that the combination approach of Kerrigan et al. [34] preserves calibration w.r.t. the model (Theorem 3.3) but not w.r.t. the human (Proposition 3.4). Yet this preservation is only under independent partitioning, and as the model approaches the Bayes classifier, independence is violated and the team predictor becomes increasingly miscalibrated (Proposition 3.5). Experiments on human predictions via ImageNet-16H agree with our theoretical finding (Figure 2).
- Prove that delegation-based teaming [49] requires the meta-classifier (a.k.a. rejector) be sufficiently calibrated as to characterize where the human and model are the superior predictors (Theorem 4.1). When the human has access to hidden features, the rejector will likely have irreducible excess risk.

2 Background & Related Work

Notation & Setting We consider a K-class classification task with labels $\mathrm{y} \in \mathcal{Y} = \{1, \ldots, K\}$, *known* features $\mathbf{x} \in \mathcal{X}$. We will consider two settings: shared and independent feature spaces. For the former, the labels are drawn from an unknown distribution $\mathbb{P}(\mathrm{y}|\mathbf{x})$, and for the latter, the labels are drawn from $\mathbb{P}(\mathrm{y}|\mathbf{x}, \mathbf{z})$, with $\mathbf{z} \in \mathcal{Z}$ being features that are known only to the human. The human makes predictions $\mathrm{h} \in \mathcal{H} = \mathcal{Y}$ by sampling from an internal predictive model of the form $\mathbb{P}(\mathrm{h}|\mathbf{x})$, $\mathbb{P}(\mathrm{h}|\mathbf{z})$, or $\mathbb{P}(\mathrm{h}|\mathbf{x}, \mathbf{z})$—depending on if the human has access to the same $(\mathbf{x})$, a different $(\mathbf{z})$, or a strictly richer $(\mathbf{x}, \mathbf{z})$ feature space than the model. We observe the known features and labels only through i.i.d. samples $(\boldsymbol{x}, y)$, and thus when $\mathbf{z}$ contains essential predictive information, a model trained only on $(\boldsymbol{x}, y)$ cannot recover the ground-truth distribution $\mathbb{P}(\mathrm{y}|\mathbf{x}, \mathbf{z})$. Predictive models are considered to be a map $\mathbf{f} : \mathcal{X} \mapsto \Delta^{K-1}$, where Δ^{K-1} is the (K-1)-dimensional simplex. The model's distribution over class labels is then $\mathrm{y} \sim \texttt{Categorical}(\mathbf{f}(\mathbf{x}))$, where $\mathrm{f}_y(\mathbf{x})$ denotes the classifier's confidence assigned to label y and is usually parameterized with the softmax function.

2.1 Statistical Calibration

We now introduce statistical calibration [14, 51] for an abstract classifier $\mathbf{f} : \mathcal{X} \mapsto \Delta^{K-1}$. Let $\Phi : \mathcal{X} \mapsto \mathcal{U}$ be a *grouping function* that partitions the feature space $\mathcal{X}$ into reference classes by assigning each point to a group indexed by $u \in \mathcal{U}$.

Definition 2.1. Pointwise Calibration. A predictor $\mathbf{f}$ is pointwise calibrated w.r.t. a choice of Φ if

$$f_y(\boldsymbol{x}) \;=\; \mathbb{E}\left[\mathbb{P}(\mathrm{y} = y|\mathbf{x}) \mid \mathbf{x} \in [\boldsymbol{x}]_\Phi\right] \;=\; \mathbb{P}\left(\mathrm{y} = y|\Phi(\boldsymbol{x})\right), \quad \forall \boldsymbol{x} \in \mathcal{X},\ \forall y \in \mathcal{Y},$$

where $[\boldsymbol{x}]_\Phi = \{\boldsymbol{x}' | \Phi(\boldsymbol{x}) = \Phi(\boldsymbol{x}')\} \subseteq \mathcal{X}$ denotes the equivalence class of $\mathbf{x}$. Therefore $\mathbf{f}$ is calibrated if the confidence it assigns to every $(\boldsymbol{x}, y)$ pair equals the *average* probability of that label within the group to which $\boldsymbol{x}$ is assigned. Appropriately choosing Φ is known as the *reference class problem* [25] as the choice of Φ naturally gives rise to stronger and weaker forms of calibration [30, 68]. The strongest form is *ideal calibration*, which means the predictor is the Bayes predictor, having matched all underlying conditional probabilities: $f_y(\boldsymbol{x}) \;=\; \mathbb{P}\left(\mathrm{y} = y|\boldsymbol{x}\right)$. Ideal calibration is 'strong' because the grouping function is 'fine grained,' having $\Phi(\mathbf{x}) = \mathbf{x}$. On the other hand, a model is

weakly calibrated if calibration only holds for a 'coarse grained' grouping function. The weakest form is marginal calibration, which occurs when $\Phi(\mathbf{x})$ partitions the feature space into a single group: $f_y(\boldsymbol{x}) = \mathbb{P}(\mathrm{y} = y)$. In the machine learning literature, *confidence calibration* is typically of interest, which uses the maximum class confidence as the grouping function: $\Phi(\mathbf{x}) = \max_y f_y(\mathbf{x})$ [23]. If a predictor is calibrated for multiple grouping functions, then it is called *multi-calibrated* [27].

Definition 2.1 presumes that the predictor is itself measurable w.r.t. Φ (i.e. $\mathbf{f}$ is a function of $\Phi(\mathbf{x})$). Yet we will also want to know if a predictor that is built from features that are not a function of Φ is still calibrated w.r.t. Φ. For such predictors, we can use the following conditional notion of calibration, which coincides with Definition 2.1 whenever the predictor is Φ-measurable.

Definition 2.2. Generalized Calibration. Let $\mathbf{g}$ be a predictor that is a measurable function of inputs generating a σ-algebra $\mathcal{S}$ (so that $\sigma(\mathbf{g}) \subseteq \mathcal{S}$), and let Φ be a grouping function. We say $\mathbf{g}$ is calibrated w.r.t. Φ if conditioning on Φ leaves the prediction unchanged:

$$\mathrm{g}_y \;=\; \mathbb{P}(\mathrm{y} = y \mid \mathbf{g} = \boldsymbol{g}, \Phi(\mathbf{x}) = u) \;=\; \mathbb{P}(\mathrm{y} = y \mid \mathbf{g} = \boldsymbol{g})\,, \quad \forall y \in \mathcal{Y},\ \forall (\boldsymbol{g}, u) \in \mathrm{supp}(\mathbf{g}, \Phi(\mathbf{x})).$$

This statement means that Φ carries no label information beyond what $\mathbf{g}$ already encodes. For predictors that are not Φ-measurable, this definition is strictly weaker than pointwise calibration. It does *not* require $\mathbf{g}$ to be constant within a cell of Φ, as $\mathbf{g}$ may vary according to information that Φ does not encode. When establishing a negative result, showing the failure of generalized calibration implies the failure of pointwise calibration.

2.2 Frameworks for Human-AI Teaming

While there are many ways to form human-AI teams, there are three fundamental workflows from which the vast majority of approaches can be derived [56]. The first is (human) prediction with AI assistance, and this is by far the most popular workflow [3, 4, 36, 39, 76]. For example, a radiologist using computer vision to inform their analysis of a medical image would fit into this paradigm [75]. Receiving help via discussions with a chatbot would be another example. The second is less commonly found in the literature but rather obvious: combine the machine and human predictions using some voting or other ensembling rule (or even an additional model) [10, 32–34, 60]. This workflow was exploited by Wright et al. [74] to identify images of supernovae by combining classifications from volunteers and a ConvNet. The last is also popular: delegate prediction-making to either the human or machine [44, 49, 53, 73]. This workflow allows for semi-automation as hopefully the model can make the majority of predictions, and the human only needs to be queried sparingly. This framework is now commonly employed for online content moderation: content that is obviously prohibited is usually detected by a model, and the ambiguous cases are deferred to a human, who can leverage cultural context and common sense to make the final determination [38].

2.2.1 AI Assistance and Prior Work on Calibration

Returning to the assistance framework, this formulation of teaming has already been well-studied from the perspective of calibration. Several user studies have been performed during which users were shown confidence scores along with the AI's predictions [4, 39, 40, 76], with the belief that confidence scores that are better calibrated should improve the quality of assistance [64]. However both Vodrahalli et al. [71] and Cao et al. [7] observed that this is not the case in practice. Vodrahalli et al. [71] found that humans tended to perform better when confidences are skewed toward under/over confidence. Cao et al. [7] observed benefits from using confidence scores but these benefits were agnostic to if the score was calibrated. Corvelo Benz and Gomez-Rodriguez [12] provide a theoretical framing of this phenomenon that explains why traditional confidence calibration did not improve the team's performance. If the human has access to additional information ($\mathbf{z}$, in our notation), then calibrating the assistance model only with respect to $\Phi(\mathfrak{X})$ is insufficient. Rather, a model with good assistance properties should be multi-calibrated w.r.t. the human's own confidence scores, thereby accounting for the hidden features. This work is our primary inspiration for examining other teaming workflows through the lens of calibration and shared vs independent feature spaces. Lastly, Peng et al. [54] characterized a no-free-lunch result for AI assistance, showing that when given calibrated confidences from two or more predictors, there does *not* exist a (non-trivial) deterministic aggregation mechanism that can never perform worse than the worse predictor.

2.2.2 Combining AI and Human Predictions

We next turn to frameworks that combine human and model predictions. These will be the first target of our analysis. We will examine two specific combination methods from the literature: model-based combination [74] and Bayesian combination of classifiers and confusion matrices [34].

Model-Based Combination We start with the obvious approach to combination. Given a dataset $\mathcal{D} = \{\boldsymbol{x}_n, y_n, h_n\}_{n=1}^N$, first fit a model $\boldsymbol{f}(\mathbf{x})$ to $\mathbf{x}$-y pairs and then combine its predictions with a human's via a second model. We define this second model as $g : \Delta^{K-1} \times \mathcal{H} \mapsto \Delta^{K-1}$. Wright et al. [74] used a simple combination rule and a support vector machine to implement $g(\mathbf{f}(\mathbf{x}), \mathrm{h})$. This setting is attractive for its simplicity and will serve as a useful 'warm up' for our line of analysis.

Bayesian Combination of Classifiers and Confusion Matrices Kerrigan et al. [34] proposed a more elegant approach to combination. Assume access to two datasets: one of feature-label pairs $\mathcal{D}_f = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ and another of prediction-label pairs $\mathcal{D}_h = \{(h_m, y_m)\}_{m=1}^M$. They first fit a classifier $f(\mathbf{x})$ to $\mathcal{D}_f$. Then they use $\mathcal{D}_h$ to estimate a row-normalized confusion matrix $\boldsymbol{C}$: $c_{i,j} = \sum_m (\mathbb{I}[y_m = i] \cdot \mathbb{I}[h_m = j]) / \sum_m \mathbb{I}[y_m = i]$. Kerrigan et al. [34] lastly combine the two models via Bayes rule as follows:

$$\begin{aligned} p(\mathrm{y}|\mathrm{h}, f(\mathbf{x})) &= \frac{p(\mathrm{h}|\mathrm{y}, f(\mathbf{x})) \cdot p(\mathrm{y}|f(\mathbf{x}))}{p(\mathrm{h}|f(\mathbf{x}))} \\ \text{(assuming } f(\mathbf{x}) \perp \mathrm{h} \mid \mathrm{y}) &= \frac{p(\mathrm{h}|\mathrm{y}) \cdot p(\mathrm{y}|f(\mathbf{x}))}{\sum_{y' \in \mathcal{Y}} p(\mathrm{h}|\mathrm{y} = y') \cdot p(\mathrm{y} = y'|f(\mathbf{x}))} = \frac{c_{\mathrm{y},\mathrm{h}} \cdot f_{\mathrm{y}}(\mathbf{x})}{\sum_{y' \in \mathcal{Y}} c_{y',\mathrm{h}} \cdot f_{y'}(\mathbf{x})} \end{aligned} \tag{1}$$

where the second equality assumes that the human's predictions are independent of $f(\mathbf{x})$. This assumption is then what allows for a standard confusion matrix to encode $p(\mathrm{h}|\mathrm{y})$.

2.2.3 Delegating Predictions: Learning to Defer

The third workflow for teaming consists of delegating each prediction task to either the human or model via a routing policy [5, 15, 16, 19, 29, 57]. This paradigm is best known as *learning to defer* (L2D) [44], as the system must learn when to use which downstream predictor. This allows for semi-automation, with the human only needing to make a (usually small) fraction of the decisions. We can derive the Bayes optimal deferral rule as follows [49]. Consider the following cost function for a decision rule $\delta : \mathcal{X} \to \mathcal{Y}^\perp$, where $\perp$ represents the decision to defer to the human and $\mathcal{Y}^\perp = \mathcal{Y} \cup \{\perp\}$: $c_{\mathrm{L2D}}(\delta(\mathbf{x}), \mathrm{y}, \mathrm{h}) \triangleq \{\ell(\delta(\mathbf{x}), \mathrm{y}) \text{ if } \delta(\mathbf{x}) \neq \perp;\ \ell(\mathrm{h}, \mathrm{y}) \text{ if } \delta(\mathbf{x}) = \perp\}$. If the decision rule returns a label $\mathrm{y} \in \mathcal{Y}$, then we apply $\ell(\delta(\mathbf{x}), \mathrm{y})$, a loss between the classifier's prediction and the true label y. When the decision is to defer, $\delta(\mathbf{x}) = \perp$, then we apply the same loss to the human's prediction: $\ell(\mathrm{h}, \mathrm{y})$. Assuming a 0-1 loss, the decision function that minimizes the conditional risk is:

$$\delta_{0-1}^*(\mathbf{x}) \triangleq \begin{cases} \perp \text{ if } r^*(\boldsymbol{x}) = 1 \\ \arg\max_y \mathbb{P}(\mathrm{y} = y|\mathbf{x}) \text{ o/w} \end{cases} \qquad r^*(\boldsymbol{x}) \triangleq \mathbb{I}\left[\mathbb{P}(\mathrm{h} = \mathrm{y}|\boldsymbol{x}) > \max_{y \in \mathcal{Y}} \mathbb{P}(\mathrm{y} = y|\boldsymbol{x})\right] \tag{2}$$

where $r^*(\boldsymbol{x})$ is a meta-classifier termed the *rejector* and $\mathbb{P}(\mathrm{h} = \mathrm{y}|\mathbf{x}) = \sum_{y \in \mathcal{Y}} \mathbb{P}(\mathrm{y} = y|\mathbf{x}) \cdot \mathbb{P}(\mathrm{h} = y|\mathbf{x})$—the human's (binary) probability of being correct. If this probability is greater than the modal confidence of the Bayes predictor $\mathbb{P}(\mathrm{y}|\mathbf{x})$, deferring to the human is the best decision. Yet one may wonder how the human can outperform $\mathbb{P}(\mathrm{y}|\mathbf{x})$. This is where additional information plays a role: the human can beat the Bayes predictor because of her access to a richer feature space $[\mathbf{x}, \mathbf{z}]$. Much of the prior work on L2D has developed surrogate loss functions for the above learning problem [49]. The most closely related work is by Verma and Nalisnick [69] and Cao et al. [8], who developed parameterizations that resulted in better confidence-calibrated estimates of $\mathbb{P}(\mathrm{h} = \mathrm{y}|\mathbf{x})$. We, on the other hand, will analyze L2D's calibration properties from a more general, theoretical perspective.

2.3 Assumptions on Human and Model

To finalize our setting of interest, we state the calibration assumption that will be the foundation of our analysis. Following prior work in human-AI teaming [34, 49], we assume we do not have access to the human's (conditional) predictive distribution $\mathbb{P}(\mathrm{h}|\mathbf{x})$ *nor can we model it well*. If we could, the human could simply be replaced by a model. Rather, we observe only samples, $\hat{\mathrm{h}} \sim \mathbb{P}(\mathrm{h}|\mathbf{x})$. Next we assume that both the human and model are calibrated for some arbitrary grouping function:

Assumption 2.3. Pointwise Calibration of Human and Model. Assume the model $\mathbf{f}$ is calibrated and measurable w.r.t. grouping function Φ_f: $\mathbf{f}(\mathbf{x}) = \mathbb{P}(\mathrm{y}|\Phi_f(\mathbf{x}))$. Likewise, assume the human's internal predictive model is calibrated and measurable w.r.t. grouping function Φ_h: $\mathbb{P}(\mathrm{h}|\cdot) = \mathbb{P}(\mathrm{y}|\Phi_h(\cdot))$, where Φ_h could partition $\mathcal{X}$, $\mathcal{Z}$, or their combined feature space.

Assuming the human provides only a sample $\hat{\mathrm{h}}$ allows our analysis to proceed without the human providing calibrated confidence statements—which humans are typically not good at [42, 65]. We stress that our Assumption 2.3 does not assume the human or model are optimal predictors. Rather, they are simply 'unbiased with respect to each cell of the partition,' with Bayes optimality happening only when the cells become infinitesimal.

3 Combination Through the Lens of Calibration

We first study perhaps the simplest instantiation of teaming: pure model-based combination. We then move on to the combination approach proposed by Kerrigan et al. [34]. For both methods, we are interested if the combined human-AI predictor preserves calibration w.r.t. $\Phi_f(\mathbf{x})$ and $\Phi_h(\mathbf{x})$.

3.1 Model-Based Combination

Given the combination function g from Section 2.2.2, can g be calibrated against Φ_h and / or Φ_f? If it is calibrated against both, then the human-AI team forms a calibrated predictor that successfully captures all 'knowledge' encoded by Φ_h and Φ_f. Let the Bayes-optimal combination function be denoted $g^*(\hat{\mathrm{h}}, \mathbf{f}(\mathbf{x})) \triangleq \mathbb{P}(\mathrm{y} \mid \hat{\mathrm{h}}, \mathbf{f}(\mathbf{x}))$. Restricting our attention to g^* produces the cleanest possible upper bound on what a calibrated g can inherit. Our result will show that g^* cannot be calibrated w.r.t. Φ_h, as long as $\mathbb{P}(\mathrm{h}|\Phi_h(\mathbf{x}))$ is not deterministic and $f(\mathbf{x})$ does not 'leak' information about Φ_h.

Shared Feature Space We first consider the case of shared features: when the human and model operate on $\mathcal{X}$. Consider the sampling step $\hat{\mathrm{h}} \sim \texttt{Categorical}(\mathbb{P}(\mathrm{h}|\Phi_h(\mathbf{x})))$. This step replaces a real-valued probability vector $\mathbb{P}(\mathrm{h}|\Phi_h(\mathbf{x}))$ with a single categorical representation $\hat{\mathrm{h}} \in \mathcal{Y}$. Of course, many independent draws would converge back to $\mathbb{P}(\mathrm{h}|\Phi_h(\mathbf{x}))$, but the composite predictor has access to exactly one sample. On the other hand, g is allowed full access to f's per-class confidences, so any partition Φ_f that is *at least as coarse as* f's output (e.g. confidence calibration) is visible to g.

Proposition 3.1. ***Generalized Calibration of*** g^* ***w.r.t.*** Φ_f**.** *Assume Φ_f is a coarsening of $\mathbf{f}$'s level sets, i.e. $\sigma(\Phi_f) \subseteq \sigma(\mathbf{f})$. Then $\mathbf{g}^*$ is calibrated w.r.t. Φ_f in the sense of Definition 2.2.*

Proof. $\mathbf{g}^* = \mathbb{P}(\mathrm{y} \mid \hat{\mathrm{h}}, \mathbf{f}(\mathbf{x}))$ has input σ-algebra $\mathcal{S} = \sigma(\hat{\mathrm{h}}, \mathbf{f}(\mathbf{x}))$, and $\sigma(\Phi_f) \subseteq \sigma(\mathbf{f}) \subseteq \mathcal{S}$. Since $\mathbf{g}^*$ and Φ_f are both $\mathcal{S}$-measurable, the tower property gives $\mathbb{P}(\mathrm{y} \mid \mathbf{g}^*, \Phi_f) = \mathbb{E}[\mathbb{P}(\mathrm{y} \mid \mathcal{S}) \mid \mathbf{g}^*, \Phi_f] = \mathbf{g}^* = \mathbb{P}(\mathrm{y} \mid \mathbf{g}^*)$. Therefore $\mathbf{g}^*$ is calibrated w.r.t. Φ_f (Definition 2.2). □

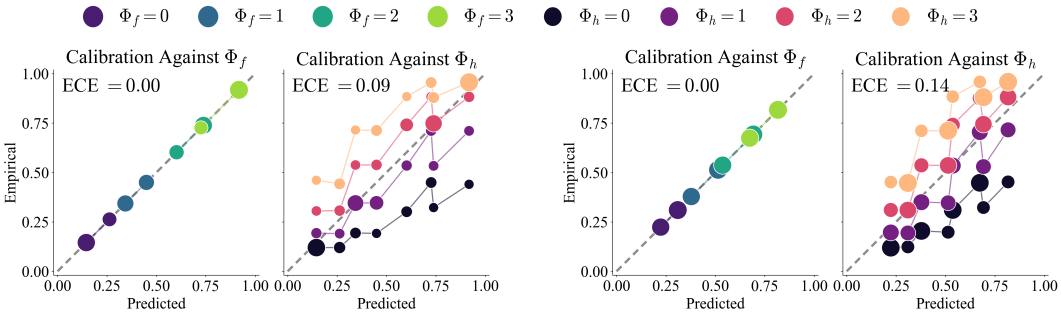
Proposition 3.1 does not assert pointwise calibration w.r.t. Φ_f (Definition 2.1): $\mathbf{g}^*$ varies with $\hat{\mathrm{h}}$ within each cell of Φ_f, which is by design, to incorporate information from both upstream predictors.

While calibration w.r.t. the model's partitions is possible, calibration w.r.t. the human's is not due to $\hat{\mathrm{h}}$ being merely a categorical draw.

Theorem 3.2. g^* ***is not calibrated w.r.t.*** Φ_h**.** *Suppose there exist two cells $S_1 \neq S_2$ of Φ_h and a value $\boldsymbol{v}$ in the range of $\mathbf{f}$ such that:*

1. $\mathbb{P}(\mathrm{h}|\Phi_h(\mathbf{x}))$ *is non-degenerate:* $\mathbb{P}(\mathrm{h} = y' \mid \Phi_h(\mathbf{x}) = S_i) > 0$ *for* $i \in \{1, 2\}$*, for some label* y'*.*
2. $\mathbf{f}$ *does not separate S_1 and S_2:* $\mathbb{P}(\mathbf{f}(\mathbf{x}) = \boldsymbol{v},\ \Phi_h(\mathbf{x}) = S_i) > 0$ *for* $i \in \{1, 2\}$*.*
3. *S_1 and S_2 have different label information:* $\mathbb{P}(\mathrm{y} = y'' \mid \mathbf{f}(\mathbf{x}) = \boldsymbol{v}, \Phi_h(\mathbf{x}) = S_1) \neq \mathbb{P}(\mathrm{y} = y'' \mid \mathbf{f}(\mathbf{x}) = \boldsymbol{v}, \Phi_h(\mathbf{x}) = S_2)$*, for some label* y''*.*
4. g^* *refines* $\mathbf{f}$*:* $\sigma(\mathbf{f}(\mathbf{x})) \subseteq \sigma\left(g^*\left(\hat{\mathrm{h}}, \mathbf{f}(\mathbf{x})\right)\right)$*.*

Then $g^*(\hat{\mathrm{h}}, \mathbf{f}(\mathbf{x}))$ *is not calibrated with respect to* Φ_h*.*



(a) Shared Features

(b) Independent Features

Figure 1: *Calibration Failure of Model-Based Combination.* The figures show simulations evaluating the calibration of g^* in the case of a shared feature space (a) and independent feature spaces (b). In both cases, the combination model g^* is calibrated w.r.t. Φ_f (ECE of 0) but not calibrated w.r.t. Φ_h (ECE of .09 for shared, .14 for independent), the partitioning for which the human is calibrated.

See Appendix A for the proof. Conditions (1)–(3) imply that $\mathbb{P}(\mathrm{y} \mid \hat{\mathrm{h}}, \mathbf{f}(\mathbf{x}), \Phi_h(\mathbf{x})) \neq \mathbb{P}(\mathrm{y} \mid \hat{\mathrm{h}}, \mathbf{f}(\mathbf{x}))$. Condition (4) is required because if g^* merges values with different model outputs, the label information that Φ_h provides at one of them can be averaged away by the others. Condition (4) guarantees that the value of g^* determines $\mathbf{f}(\mathbf{x})$, mirroring the assumption from Proposition 3.1 that $\sigma(\Phi_f) \subseteq \sigma(\mathbf{f})$. If $\mathbf{f}$ is multi-calibrated w.r.t. $\sigma(\Phi_f, \Phi_h)$, then g can be calibrated w.r.t. both. This was prevented by condition (3). While true multi-calibration effectively collapses $\mathbf{g}$ into $\mathbf{f}$, it suggests that $\mathbf{f}(\mathbf{x})$ can 'leak' information about the human's predictions to g. This result is in the same spirit as Corvelo Benz and Gomez-Rodriguez [12]'s solution of human-aligned calibration: improve the team via multi-calibrating the model to make it 'human-aware.'

Independent Features Now turning to the setting in which the human observes features $\mathbf{z}$ and the model observes $\mathbf{x}$, notice that Theorem 3.2 immediately still applies. Moreover, it can be strengthened by removing supposition (2), since it is impossible for $\mathbf{f}$ to separate the cells since $\mathbf{f}$ never even has access to $\mathbf{z}$. Yet, ultimately, this is a good scenario and poses no additional complications to combination-based teaming. Having access to additional information $\mathbf{z}$—if only by way of samples—still allows for $\mathbf{g}$ to demonstrate complementarity when $\mathbb{P}(\mathrm{y}|\mathbf{x}, \mathbf{z})$ is the true data generating process.

Simulation Study We test these intuitions and empirically support our results via a simulation that compares the settings of shared and independent features. The data generating distribution is set to:

$$\mathbb{P}(\mathrm{y}=1 \mid \mathrm{x}, \mathrm{z}) \;=\; \texttt{logistic}\big(-2 + \tfrac{3}{5}\,\Phi_f(\mathrm{x}) + \tfrac{3}{5}\,\Phi_h(\mathrm{z}) + \tfrac{1}{6}\,\Phi_f(\mathrm{x})\,\Phi_h(\mathrm{z})\big)\,,$$

with the partitions being $\Phi_f(\mathrm{x}), \Phi_h(\mathrm{z}) \in \{0, 1, 2, 3\}$. In the *shared-feature* setting, $(\mathrm{z} = x_1, \mathrm{x} = x_2)$ such that $x_1 = x_2$ with probability 0.55 and is drawn uniformly at random otherwise. The partitions then evaluate to $\Phi_h(\mathrm{z}) = x_1$ and $\Phi_f(\mathrm{x}) = x_2$. In the *independent* setting, x and z are independent and drawn uniformly on $\{0, 1, 2, 3\}$, with $\Phi_f(\mathrm{x}) = \mathrm{x}$ and $\Phi_h(\mathrm{z}) = \mathrm{z}$. In both settings, the upstream predictors $\mathbb{P}(\mathrm{h}|\mathrm{z})$ and $f(\mathrm{x})$ are calibrated against their respective partitions by construction. The downstream model g^* is the Bayes-optimal predictor by setting it to the empirical mean of y within each $\big(\hat{\mathrm{h}}, f(\mathrm{x})\big)$ bucket. Figure 1 reports reliability diagrams of g^* within Φ_f- and Φ_h-partitions; marker area encodes cell count and color indexes the partition cell. The left plot of both subfigures shows that g^* inherits Φ_f-calibration in both regimes. However, g^* is miscalibrated against Φ_h, with an ECE of 0.090 in the shared-feature setting and an ECE of 0.144 in the independent-feature setting.

3.2 Bayesian Combination via Confusion Matrix Representation of Human

We now turn to the Bayesian combination method of Kerrigan et al. [34]. Unlike the previous sub-section, where the combination was limited to a single sample from the human, this combination sees a (normalized) confusion matrix that aggregates the human's (sampled) predictions.

Shared Feature Space We return to the setting of no hidden features: the model and human both make predictions from $\mathbf{x}$. Like in the previous combination setting, we are interested if the combined model is calibrated w.r.t. Φ_f and / or Φ_h. For the former, we again have an affirmative result.

Theorem 3.3. ***Bayesian combination posterior is calibrated w.r.t.*** Φ_f. *Assuming that* $\Phi_h \perp \Phi_f(\mathbf{x}) \mid \mathrm{y}$, *the posterior distribution is calibrated w.r.t.* Φ_f *in the sense of Definition 2.2:*

$$p\left(\mathrm{y}|\mathrm{h}, f(\mathbf{x})\right) \;=\; \mathbb{P}\left(\mathrm{y} \mid \mathrm{h}, \Phi_f(\mathbf{x})\right) \;=\; \frac{\mathbb{E}_{\Phi_h}\left[\mathbb{P}(\mathrm{h}|\Phi_h(\mathbf{x}))|\mathrm{y}\right] \cdot \mathbb{P}\left(\mathrm{y}|\Phi_f(\mathbf{x})\right)}{\mathbb{P}(\mathrm{h}|\Phi_f(\mathbf{x}))}$$

See Appendix B for the proof. The result is made straightforward by the conditional independence assumption since it implies $\mathrm{h} \perp \Phi_f(\mathbf{x}) \mid \mathrm{y}$, which then allows for the confusion matrix representation of $p(\mathrm{h}|\mathrm{y})$. This assumption is crucial, and we will revisit it below. Turning to the human, again we have a negative result for Φ_h.

Proposition 3.4. ***Bayesian combination posterior is not calibrated w.r.t.*** Φ_h. *Assume that* Φ_h *and* $\mathbf{f}$ *satisfy conditions (1)–(3) of Theorem 3.2. Moreover, assume that* $p(\mathrm{y} \mid \mathrm{h}, \mathbf{f}(\mathbf{x}))$ *satisfies condition (4) in place of* g^* *in Theorem 3.2. Then the posterior is not calibrated with respect to* Φ_h.

Proof. Like g^*, the confusion-matrix posterior is a measurable function of $(\mathrm{h}, \mathbf{f}(\mathbf{x}))$. Yet the human enters only through the confusion matrix $\mathbf{c}_{\mathrm{y},\cdot} = \int \mathbb{P}(\mathrm{h} \mid \Phi_h(\mathbf{x}))\, d\mathbb{P}(\mathbf{x} \mid \mathrm{y})$, which averages out the within-class variation of Φ_h. Therefore the proof of Theorem 3.2 is applicable here: conditions (1)–(3) make Φ_h carry label information beyond $(\mathrm{h}, \mathbf{f}(\mathbf{x}))$, and condition (4) ensures that label information cannot be marginalized away. Thus the definition of generalized calibration (2.2) fails. □

From these results, we see that the best case scenario is when Φ_f and Φ_h are conditionally independent (given y), thereby encoding unique aspects of the feature space. However, the model becomes quite sensitive to this property, to the point of being non-robust. The *finer grained* Φ_f becomes, the more likely the conditional independence assumption is to be broken. The posterior tends to over-sharpen via double-counting: the posterior will not be ideally calibrated even if $\mathbf{f}(\mathbf{x}) = \mathbb{P}(\mathrm{y}|\mathbf{x})$.

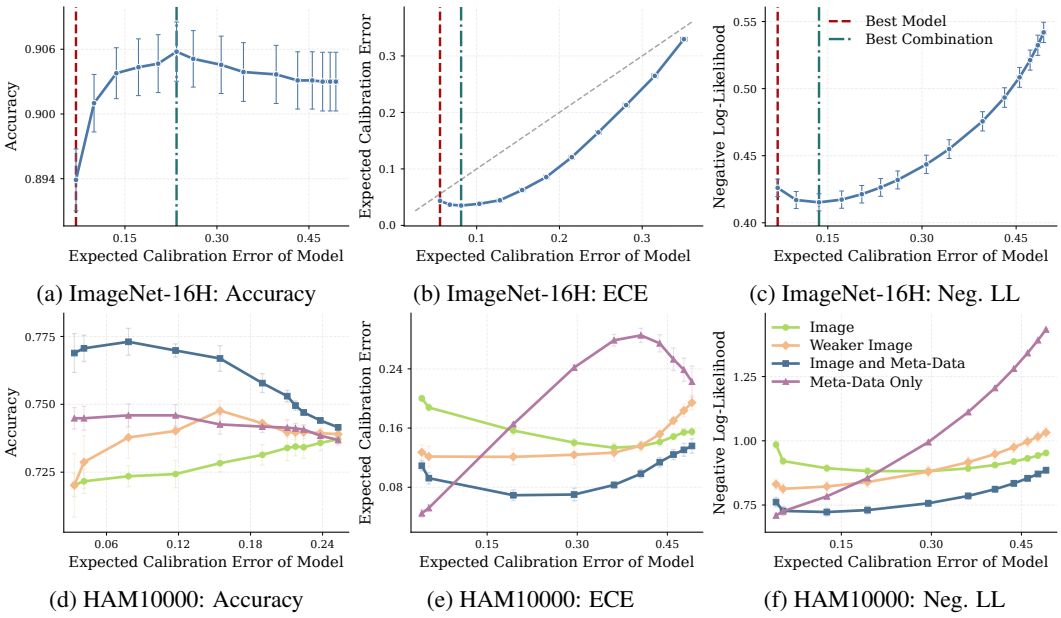
Proposition 3.5. ***Posterior miscalibration when model is already Bayes predictor.*** *Let* $\Phi_f(\mathbf{x}) = \mathbf{x}$ *such that* $\mathbf{f}(\mathbf{x}) = \mathbb{P}(\mathrm{y} \mid \mathbf{x})$, *and assume that* $\mathbb{P}(\mathrm{y} \mid \mathbf{x})$ *has full support on* $\mathcal{Y}$ *on a set of* $\mathbf{x}$ *of positive probability. Suppose that the human's prediction is informative, meaning it is not independent of the label,* $\mathrm{h} \not\perp \mathrm{y}$. *Then the posterior is not ideally calibrated:* $p\left(\mathrm{y} \mid \mathrm{h}, \mathbf{f}(\mathbf{x})\right) \neq \mathbb{P}(\mathrm{y}|\mathbf{x})$.

See Appendix C for the proof. This posterior sharpening does *not* occur when $\mathbb{P}(\mathrm{h} \mid \mathbf{x})$ is a uniform distribution, which necessitates the supposition that the human is informative. This result presents an interesting contrast to model-based combination, which *can* be ideally calibrated when the model is.

Independent Features In the case of independent features, the human and model are guaranteed to introduce complementary information: $\mathbf{z} \perp \mathbf{x} \mid \mathrm{y}$ implies $\Phi_h \perp \Phi_f \mid \mathrm{y}$. In turn, while the posterior is still not calibrated w.r.t. Φ_h, we no longer have to worry about $f(\mathbf{x})$ being *too calibrated* and combination introducing over-sharpening. This suggests an experimental hypothesis: when the model and human share features, we should see the posterior become an increasingly *worse* predictor as $\mathbf{f}$ becomes more calibrated. However, with independent feature spaces, improvements in $\mathbf{f}$'s calibration should always translate to a better posterior predictor.

Experimental Validation on ImageNet-16H We evaluate the above hypothesis starting with the *shared features* setting. We use the same experimental setup as Kerrigan et al. [34] for the ImageNet-16H dataset, which contains $28,997$ human classifications of ImageNet-16 images from a total of 145 participants.[2] For each image, the dataset provides approximately six human annotations. We aggregate these annotations and interpret the normalized count vector as a human's predictive distribution for that image. To match our theoretical setting, we sample from this empirical distribution. See Appendix E for additional experimental details. We evaluate the performance of the resulting human-AI combination as the classifier's individual confidence calibration varies. We do this via temperature scaling the VGG-19 softmax probabilities [23]. We sweep over a grid of temperatures and evaluate both the standalone classifier and the confusion-matrix-combined predictor using held-out samples. We report three quantities: accuracy, expected (confidence) calibration error (ECE), and negative log-likelihood (NLL). Results are reported in Subfigures 2a, 2b, and 2c. For a shared feature space, better model calibration does not lead to a better combined predictor. We report similar trends for CIFAR-10H in Appendix E.3.

[2]While the human participants could be using hidden features to inform their predictions, the fact that computer vision systems have surpassed human performance on ImageNet [26] suggests any advantage from hidden information is negligible.



(a) ImageNet-16H: Accuracy (b) ImageNet-16H: ECE (c) ImageNet-16H: Neg. LL

(d) HAM10000: Accuracy (e) HAM10000: ECE (f) HAM10000: Neg. LL

Figure 2: *Empirical Validation for Real and Simulated Humans. Top row*: For ImageNet-16H, we vary the (confidence) calibration of a VGG-19 classifier via temperature scaling and combine each classifier with a sampled human label using the confusion-matrix update. The best calibrated model never results in the best posterior predictor. *Bottom row*: On HAM10000, we vary the calibration of an image-only ResNet-18 classifier and combine it with sampled predictions from simulated humans with a range of feature spaces. When 'humans' see only meta-data, better model calibration results in an improved posterior predictor. Error bars show one standard error over five random seeds.

Experimental Validation on HAM10000 We next use HAM10000 [67] to analyze the *independent feature* setting. Each example contains a 224×224 dermoscopic image along with patient- or lesion-level meta-data. We identify the image with the model-observed feature $\mathbf{x}$ and treat the meta-data as $\mathbf{z}$—unavailable to the base classifier but available to the human. The model is a temperature-scaled ResNet-18. We simulate the human using an auxiliary classifier—following Verma and Nalisnick [69]—allowing for a controlled comparison of several feature-access settings. The first, denoted *Image*, represents the human via the same ResNet-18 employed for the main classifier (i.e. perfectly redundant features). The second, denoted *Weaker Image*, uses a weaker ResNet-18 trained on 128×128 images for 4 epochs, which represents a human with a coarser feature space. The third, denoted *Image and Meta-Data*, is a multimodal expert that uses a ResNet-18 to encode the image before combining it with meta-data via an MLP. This represents a human with a strictly richer feature space. The fourth, denoted *Meta-Data Only*, is a logistic-regression-based human trained on age, sex, and lesion localization. It has an independent feature space, unlike the previous three (simulated) humans. See Appendix E for additional details. Subfigures 2d, 2e, and 2f report accuracy, ECE under confidence calibration, and NLL, respectively. Just as for the ImageNet-16H results, the y-axis evaluates the predictive quality of the posterior as a function of $\mathbf{f}(\mathbf{x})$'s confidence calibration by varying the temperature parameter. For the meta-data-only human (purple), we see that—as our theory predicts—the posterior achieves its best accuracy, ECE, and NLL for the *best-calibrated* classifier. For the image-only human, the posterior accuracy is best for the *worst calibrated model*, and the posterior ECE and NLL are best for the fifth-best calibrated model. For the image-and-meta-data human (blue), the posterior is best for the third-best calibrated model. See Appendix F for additional experiments simulating humans with noisy and biased predictions.

4 Delegation Through the Lens of Calibration

We now turn to the alternative setting of delegation: either the human or model predicts with the selection being determined by the rejector $r(\mathbf{x})$. Since the model or human alone makes the

prediction, predictions will always come from a predictor that is either Φ_f- or Φ_h-calibrated, by assumption. Thus the focus of our analysis turns to $r(\mathbf{x})$: *how must the rejector be calibrated in order to choose the superior downstream predictor?* Assuming the downstream predictors are fixed, we can write the combined (system) risk as:

$$\mathcal{R}^{0\text{-}1}_{f,h}(r) \triangleq \mathbb{E}_{\mathbf{x}}\left[(1 - r(\mathbf{x}))\left(1 - \max_y f_y(\mathbf{x})\right) + r(\mathbf{x})(1 - \mathbb{P}(\mathrm{h} = \mathrm{y} \mid \mathbf{x}))\right]. \tag{3}$$

Minimization of Equation 3 yields the oracle rule $r^*(\mathbf{x}) = \mathbb{I}[\mathbb{P}(\mathrm{h} = \mathrm{y} \mid \mathbf{x}) > \max_y f_y(\mathbf{x})]$, recovering the Bayes-optimal deferral rule of Mozannar and Sontag [49]. In practice, we assume the rejector is parameterized by a scalar score $\rho : \mathfrak{X} \times \Delta^{K-1} \mapsto [0, 1]$ such that

$$r(\mathbf{x}; \rho) \triangleq \mathbb{I}\left[\rho(\mathbf{x}, \mathbf{f}(\mathbf{x})) > \max_y f_y(\mathbf{x})\right]. \tag{4}$$

If ρ is ideally calibrated such that $\rho(\mathbf{x}, \mathbf{f}(\mathbf{x})) = \mathbb{P}(\mathrm{h} = \mathrm{y} \mid \mathbf{x})$, then the oracle rule r^* is recovered. Notably, ρ takes both the original features $\mathbf{x}$ and the model's output $\mathbf{f}(\mathbf{x})$ as inputs. This explicit dependence on $\mathbf{f}(\mathbf{x})$ is a key feature of this parameterization and differs from most L2D implementations [8, 49, 53, 69]—but not all [44]—where the probability that the expert is correct is typically estimated from $\mathbf{x}$ alone. This choice does not incur computational overhead since finding the maximum dimension of $\mathbf{f}(\mathbf{x})$ is needed to evaluate $r(\mathbf{x}; \rho)$. We discuss the theoretical implications of this choice below.

We now present our central result for delegation that characterizes the rejector's calibration requirements in relation to the downstream predictors.

Theorem 4.1. ***Risk-minimizing rejector under calibration.*** *Assume ties between the downstream predictors are negligible: $\mathbb{P}(\mathrm{h} = \mathrm{y} \mid \mathbf{x}) \neq \max_y f_y(\mathbf{x})$ almost surely. Let Φ_r be a grouping function on $\mathfrak{X} \times \Delta^{K-1}$, and let $\tilde{\Phi}_r(\mathbf{x}) \triangleq \Phi_r(\mathbf{x}, \mathbf{f}(\mathbf{x}))$ be the partition of $\mathfrak{X}$ induced by the additional refinement w.r.t. $\mathbf{f}(\mathbf{x})$. Assume ρ is calibrated w.r.t. $\tilde{\Phi}_r(\mathbf{x})$. Then the rejector $r(\mathbf{x}; \rho)$ (Equation 4) minimizes $\mathcal{R}^{0\text{-}1}_{f,h}(r)$ (Equation 3) if and only if $\sigma(\mathbf{1}_{\mathfrak{D}^*}) \subseteq \sigma(\tilde{\Phi}_r)$, where $\mathfrak{D}^* \triangleq \{\mathbf{x} \in \mathfrak{X} \mid r^*(\mathbf{x}) = 1\}$, the points for which the human should make the prediction under the Bayes optimal rejector.*

See Appendix D for the proof. As one would expect, the rejector's induced partition $\tilde{\Phi}_r$ must be fine-grained enough to identify the optimal-deferral set $\mathfrak{D}^*$. Thus one pressure on $\tilde{\Phi}_r$ is to capture which side of the threshold $\max_y f_y$ the human's correctness probability falls on. As the human's partition Φ_h becomes finer-grained, $\tilde{\Phi}_r$ must increase its granularity to track its threshold-crossings. As Φ_h's partitions shrink in size, $\sigma(\mathbf{1}_{\mathfrak{D}^*})$ must approach the resolution of $\sigma(\Phi_f, \Phi_h)$ at the boundaries of $\mathfrak{D}^*$. This would force $\tilde{\Phi}_r$ to absorb Φ_h and may make the human's involvement unnecessary (as her predictive power becomes increasingly captured by ρ alone).

$\mathbf{f}(\mathbf{x})$ as Input to the Rejector In Theorem 4.1, the model's partition Φ_f does not appear explicitly, despite the comparison to $\max_y f_y(\mathbf{x})$ being Φ_f-measurable. This requirement is circumvented by giving ρ access to $\mathbf{f}(\mathbf{x})$ by assumption, which forces $\tilde{\Phi}_r$ to refine Φ_f by construction. If ρ had $\mathbf{x}$ as its only input, then the sufficient condition would become $\sigma(\Phi_f, \mathbf{1}_{\mathfrak{D}^*}) \subseteq \sigma(\Phi_r)$: Φ_r would need to explicitly refine Φ_f in addition to identifying $\mathfrak{D}^*$. As mentioned above, adding $\mathbf{f}(\mathbf{x})$ as a feature to the rejector is not common in L2D implementations but actually was done in Madras et al. [44]'s seminal work, with the choice being justified through intuition rather than theory. The majority of subsequent implementations use a neural network to map $\mathbf{x}$ to $K + 1$ outputs [8, 49, 69], with the classifier encoded by the first K dimensions and the rejector encoded by the $(K + 1)$th. When the classifier and rejector share a backbone architecture, we expect adding $\mathbf{f}(\mathbf{x})$ as a feature to matter less for these implementations than ones with independent parameterizations [44, 53].

Human with Additional Information We next consider the case in which the human uses $\mathbf{x}$ and additional information $\mathbf{z}$ to make her prediction (calibrated w.r.t. $\Phi_h(\mathbf{x}, \mathbf{z})$). This is the setting assumed by L2D's theoretical foundation, since $\mathbf{z}$ is necessary for the human to outperform the Bayes classifier on $\mathfrak{X}$. Yet, like the model, the rejector only has access to $\mathbf{x}$, and it must decide which downstream predictor to use without knowing $\mathbf{z}$ or h. The risk in Equation 3 then becomes

$$\widetilde{\mathcal{R}}^{0\text{-}1}_{f,h}(r) = \mathbb{E}_{\mathbf{x},\mathbf{z}}\left[(1 - r(\mathbf{x}))\left(1 - \max_y f_y(\mathbf{x})\right) + r(\mathbf{x})(1 - \mathbb{P}(\mathrm{h} = \mathrm{y} \mid \mathbf{x}, \mathbf{z}))\right]. \tag{5}$$

Without access to the additional information, Equation 5 can be minimized only up to marginalization over $\mathbf{z}$: $\mathbb{P}(\mathrm{h} = \mathrm{y} \mid \mathbf{x}) = \mathbb{E}_{\mathbf{z}|\mathbf{x}}[\mathbb{P}(\mathrm{h} = \mathrm{y} \mid \mathbf{x}, \mathbf{z})]$. Theorem 4.1 then immediately yields:

Corollary 4.2. ***Best $\mathfrak{X}$-measurable rejector under additional information.*** *Let $\tilde{\mathfrak{D}} \triangleq \{\mathbf{x} \in \mathfrak{X} \mid \mathbb{P}(\mathrm{h} = \mathrm{y} \mid \mathbf{x}) > \max_y f_y(\mathbf{x})\}$ denote the $\mathbf{z}$-marginalized optimal-deferral set. Among $\mathfrak{X}$-measurable rejectors, $r(\mathbf{x}; \rho)$ minimizes $\widetilde{\mathcal{R}}^{0\text{-}1}_{f,h}(r)$ (Equation 5) if and only if $\sigma(\mathbf{1}_{\tilde{\mathfrak{D}}}) \subseteq \sigma(\tilde{\Phi}_r)$.*

The best $\mathfrak{X}$-measurable rejector cannot in general match the oracle $r^*(\mathbf{x}, \mathbf{z}) = \mathbb{I}[\mathbb{P}(\mathrm{h} = \mathrm{y} \mid \mathbf{x}, \mathbf{z}) > \max_y f_y(\mathbf{x})]$. The excess risk $\widetilde{\mathcal{R}}^{0\text{-}1}_{f,h}(r(\mathbf{x}; \rho^*)) - \widetilde{\mathcal{R}}^{0\text{-}1}_{f,h}(r^*(\mathbf{x}, \mathbf{z}))$ vanishes only when—for each $\mathbf{x} = \boldsymbol{x}_0$—$\mathbb{P}(\mathrm{h} = \mathrm{y} \mid \boldsymbol{x}_0, \mathbf{z})$ and $\mathbb{E}_{\mathbf{z}|\boldsymbol{x}_0}[\mathbb{P}(\mathrm{h} = \mathrm{y} \mid \boldsymbol{x}_0, \mathbf{z})]$ always remain on the same side of the threshold $\max_y f_y(\boldsymbol{x}_0)$. An example of a setting with excess risk is shown in the figure to the right. We see that $\mathbb{P}(\mathrm{h} = \mathrm{y} \mid \boldsymbol{x}_0, \mathbf{z})$ (black line) crosses the threshold $\max_y f_y(\boldsymbol{x}_0)$ (blue line) several times, yet $\mathbb{E}_{\mathbf{z}|\boldsymbol{x}_0}[\mathbb{P}(\mathrm{h} = \mathrm{y} \mid \boldsymbol{x}_0, \mathbf{z})]$ (red line) always signals to not defer to the human. Of course, multi-calibrating the rejector w.r.t. Φ_h would remove the information asymmetry and the resulting excess risk. In summary, the presence of additional information is easy to incorporate via human-AI combination and presents no fundamental modeling challenge. For delegation, the story is not so simple: the deferral policy is fundamentally limited by the unobservable variation in $\mathbf{z}$ and its effect on whether the human is the superior downstream predictor.

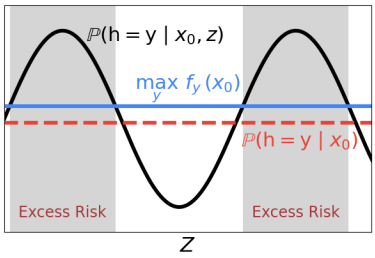


Figure 3: *Threshold crossings due to variation in $\mathbf{z}$ drive excess risk.*

5 Discussion, Limitations, Future Work

Assuming a human and model are calibrated predictors, we have analyzed how these calibration properties propagate into the human-AI team's construction. In particular, we studied two methods for combining human and model predictions and the standard approach to delegating predictions (i.e. learning to defer). For the combination-based approaches, calibration w.r.t. the model's partition Φ_f is possible (3.1, 3.3), but calibration w.r.t. the human's partition Φ_h is not (3.2, 3.4). This latter result was due to the human's prediction either being sampled or aggregated in a confusion matrix—both of which unavoidably destroy the information in Φ_h. The possible strategies for regaining Φ_h-calibration are to multi-calibrate the model or draw multiple human predictions [72], which will recover $\mathbb{P}(\mathrm{h}|\Phi_h(\mathbf{x}))$ asymptotically. The presence of additional features only available to the human poses no issue to the combination frameworks we studied. Specifically, Kerrigan et al. [34]'s combination method works better when there is information asymmetry, since stronger model calibration results in a better posterior predictor. Figure 2 reports experimental verification of this result.

In regards to delegation, Theorem 4.1 shows that the optimal rejector must be calibrated with a granularity determined by the optimal-deferral set—$\{\mathbf{x} \in \mathfrak{X} \mid \mathbb{P}(\mathrm{h} = \mathrm{y} \mid \mathbf{x}) > \max_y f_y(\mathbf{x})\}$—and Φ_f. This latter requirement suggests that L2D implementations should always include $\mathbf{f}(\mathbf{x})$ as an input feature to the rejector model, providing refinement w.r.t. Φ_f by construction. Unfortunately, the presence of hidden information makes delegation difficult: the deferral policy can at best make decisions considering only the $\mathbf{z}$-marginalized expert correctness, $\mathbb{E}_{\mathbf{z}|\mathbf{x}}[\mathbb{P}(\mathrm{h} = \mathrm{y} \mid \mathbf{x}, \mathbf{z})]$. Therefore human-AI combination is a more natural solution in the presence of considerable feature asymmetry.

Limitations The primary limitation of this work is that the model and human are assumed to be calibrated, when, in practice, calibration can be hard to verify for even modestly fine-grained partitions (e.g. confidence calibration) [37, 59]. Yet, since our results mostly highlight limitations in teaming frameworks, assuming the good-case scenario of calibration increases their generality. Furthermore, learning has been avoided entirely, and as the considerable work on learning theory for delegation has shown, issues surrounding parameterization, consistency and realizability matter in practice [9, 46–48, 50, 53]. Lastly, while experiments on ImageNet-16H and HAM10000 supported our results in Section 3.2, our other results were empirically validated only via small-scale simulations.

Future Work It would be natural to extend our results to teams with multiple humans and AI(s) [28, 35, 45, 63, 70]. Adapting our results to support approximate calibration [6, 18, 21, 77] might allow for analysis of combination methods that take as input a human's miscalibrated confidences. Lastly, studying human-AI teaming over time (possibly supported by a priority queue [43]), such that the human and model can interact and adapt [2, 13, 20, 52], introduces complexity that is outside the scope of static calibration assumptions and is highly relevant to modern human-AI teaming systems.

Acknowledgments

We acknowledge computational resources from the Johns Hopkins DSAI cluster. Yixin Wang was supported in part by funding from the Office of Naval Research under grant N00014-23-1-2590, the National Science Foundation under grants No. 2310831, No. 2428059, No. 2435696, No. 2440954, the Michigan Institute for Data Science's Propelling Original Data Science (PODS) grant, Two Sigma Investments LP, and LG Management Development Institute AI Research.

References

[1] Rohan Alur, Manish Raghavan, and Devavrat Shah. Human Expertise in Algorithmic Prediction. *Advances in Neural Information Processing Systems*, 2024.

[2] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. Updates in Human-AI Teams: Understanding and Addressing the Performance / Compatibility Tradeoff. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[3] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[4] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2021.

[5] Umang Bhatt and Holli Sargeant. When Should Algorithms Resign? A Proposal for AI Governance. *Computer*, 2024.

[6] Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. A Unifying Theory of Distance from Calibration. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, 2023.

[7] Shiye Cao, Anqi Liu, and Chien-Ming Huang. Designing for Appropriate Reliance: The Roles of AI Uncertainty Presentation, Initial User Decision, and User Demographics in AI-Assisted Decision-Making. *Proceedings of the ACM on Human-Computer Interaction*, 2024.

[8] Yuzhou Cao, Hussein Mozannar, Lei Feng, Hongxin Wei, and Bo An. In Defense of Softmax Parametrization for Calibrated and Consistent Learning to Defer. In *Advances in Neural Information Processing Systems*, 2023.

[9] Mohammad-Amin Charusaie, Hussein Mozannar, David Sontag, and Samira Samadi. Sample Efficient Learning of Predictors that Complement Humans. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

[10] Robert T. Clemen. Combining Forecasts: A Review and Annotated Bibliography. *International Journal of Forecasting*, 1989.

[11] Natalie Collina, Ira Globus-Harris, Surbhi Goel, Varun Gupta, Aaron Roth, and Mirah Shi. Collaborative Prediction: Tractable Information Aggregation via Agreement. In *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms*, 2026.

[12] Nina Corvelo Benz and Manuel Gomez-Rodriguez. Human-Aligned Calibration for AI-Assisted Decision Making. In *Advances in Neural Information Processing Systems*, 2023.

[13] Nina Corvelo Benz, Eleni Straitouri, and Manuel Gomez-Rodriguez. Learning to Decide with AI Assistance under Human-Alignment. *ArXiv Preprint*, 2026.

[14] Philip Dawid. The Well-Calibrated Bayesian. *Journal of the American Statistical Association*, 1982.

[15] Abir De, Paramita Koley, Niloy Ganguly, and Manuel Gomez-Rodriguez. Regression Under Human Assistance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[16] Abir De, Nastaran Okati, Ali Zarezade, and Manuel Gomez-Rodriguez. Classification Under Human Assistance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[17] Kate Donahue, Alexandra Chouldechova, and Krishnaram Kenthapadi. Human-Algorithm Collaboration: Achieving Complementarity and Avoiding Unfairness. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 2022.

[18] Dean P Foster and Rakesh V Vohra. Asymptotic Calibration. *Biometrika*, 1998.

[19] Andrew Fuchs, Andrea Passarella, and Marco Conti. Optimizing Delegation in Collaborative Human-AI Hybrid Teams. *ACM Transactions on Autonomous and Adaptive Systems*, 2024.

[20] Cleotilde Gonzalez, Kate Donahue, Daniel G Goldstein, Hoda Heidari, Mohammad S Jalali, Beau Schelble, Aarti Singh, and Anita Williams Woolley. Toward a Science of Human–AI Teaming for Decision Making: A Complementarity Framework. *PNAS Nexus*, 2026.

[21] Parikshit Gopalan, Konstantinos Stavropoulos, Kunal Talwar, and Pranay Tankala. Efficient Calibration for Decision Making. *ArXiv Preprint*, 2025.

[22] Barbara J. Grosz. Collaborative Systems. *AI Magazine*, 1996.

[23] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On Calibration of Modern Neural Networks. In *International Conference on Machine Learning*, 2017.

[24] Ziyang Guo, Yifan Wu, Jason D Hartline, and Jessica Hullman. A Decision Theoretic Framework for Measuring AI Reliance. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 2024.

[25] Alan Hájek. The Reference Class Problem is Your Problem Too. *Synthese*, 2007.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[27] Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (Computationally-Identifiable) Masses. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[28] Patrick Hemmer, Sebastian Schellhammer, Michael Vössing, Johannes Jakubik, and Gerhard Satzger. Forming Effective Human-AI Teams: Building Machine Learning Models that Complement the Capabilities of Multiple Experts. In *International Joint Conference on Artificial Intelligence*, 2022.

[29] Bengt Holmstrom. On the Theory of Delegation. Technical report, Northwestern University, 1980.

[30] Benedikt Höltgen and Robert C Williamson. On the Richness of Calibration. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 2023.

[31] Ece Kamar. Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence. In *International Joint Conference on Artificial Intelligence*, 2016.

[32] Ece Kamar, Severin Hacker, and Eric Horvitz. Combining Human and Machine Intelligence in Large-Scale Crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems*, 2012.

[33] Markelle Kelly, Alex James Boyd, Sam Showalter, Mark Steyvers, and Padhraic Smyth. Bayesian Inference for Correlated Human Experts and Classifiers. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

[34] Gavin Kerrigan, Padhraic Smyth, and Mark Steyvers. Combining Human Predictions with Model Probabilities via Confusion Matrices and Calibration. In *Advances in Neural Information Processing Systems*, 2021.

[35] Vijay Keswani, Matthew Lease, and Krishnaram Kenthapadi. Towards Unbiased and Accurate Deferral to Multiple Experts. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2021.

[36] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, 2018.

[37] Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified Uncertainty Calibration. In *Advances in Neural Information Processing Systems*, 2019.

[38] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q. Vera Liao, Yunfeng Zhang, and Chenhao Tan. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2022.

[39] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 2023.

[40] Jingshu Li, Yitian Yang, Q. Vera Liao, Junti Zhang, and Yi-Chieh Lee. As Confidence Aligns: Understanding the Effect of AI Confidence on Human Self-Confidence in Human-AI Decision Making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2025.

[41] Michelle M Li, Ben Y Reis, Adam Rodman, Tianxi Cai, Noa Dagan, Ran D Balicer, Joseph Loscalzo, Isaac S Kohane, and Marinka Zitnik. Scaling Medical AI Across Clinical Contexts. *Nature Medicine*, 2026.

[42] Sarah Lichtenstein, Baruch Fischhoff, and Lawrence D Phillips. Calibration of Probabilities: The State of the Art to 1980. In *Judgment Under Uncertainty*, 1982.

[43] Thodoris Lykouris and Wentao Weng. Learning to Defer in Congested Systems: The AI-Human Interplay. *ArXiv Preprint*, 2024.

[44] David Madras, Toniann Pitassi, and Richard Zemel. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. In *Advances in Neural Information Processing Systems*, 2018.

[45] Anqi Mao, Christopher Mohri, Mehryar Mohri, and Yutao Zhong. Two-Stage Learning to Defer with Multiple Experts. In *Advances in Neural Information Processing Systems*, 2023.

[46] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Realizable H-Consistent and Bayes-Consistent Loss Functions for Learning to Defer. In *Advances in Neural Information Processing Systems*, 2024.

[47] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Mastering Multiple-Expert Routing: Realizable H-Consistency and Strong Guarantees for Learning to Defer. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

[48] Christopher Mohri, Daniel Andor, Eunsol Choi, Michael Collins, Anqi Mao, and Yutao Zhong. Learning to Reject with a Fixed Predictor: Application to Decontextualization. In *The Twelfth International Conference on Learning Representations*, 2024.

[49] Hussein Mozannar and David A. Sontag. Consistent Estimators for Learning to Defer to an Expert. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[50] Hussein Mozannar, Hunter Lang, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David Sontag. Who Should Predict? Exact Algorithms For Learning to Defer to Humans. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, 2023.

[51] Allan H Murphy and Edward S Epstein. Verification of Probabilistic Predictions: A Brief Review. *Journal of Applied Meteorology*, 1967.

[52] Gali Noti, Kate Donahue, Jon Kleinberg, and Sigal Oren. AI-Assisted Decision Making with Human Learning. In *Proceedings of the 26th ACM Conference on Economics and Computation*, 2025.

[53] Nastaran Okati, Abir De, and Manuel Gomez-Rodriguez. Differentiable Learning Under Triage. In *Advances in Neural Information Processing Systems*, 2021.

[54] Kenny Peng, Nikhil Garg, and Jon Kleinberg. A No Free Lunch Theorem for Human-AI Collaboration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[55] Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. Human Uncertainty Makes Classification More Robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[56] Clara Punzi, Roberto Pellungrini, Mattia Setzu, Fosca Giannotti, and Dino Pedreschi. Learning Paradigms for Hybrid Decision-Making Systems. *ACM Computing Surveys*, 2026.

[57] Maithra Raghu, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. The Algorithmic Automation Problem: Prediction, Triage, and Human Effort. *ArXiv Preprint*, 2019.

[58] Benjamin Recht. The Actuary's Final Word on Algorithmic Decision Making. *ArXiv Preprint*, 2025.

[59] Raphael Rossellini, Jake A. Soloff, Rina Foygel Barber, Zhimei Ren, and Rebecca Willett. Can a Calibration Metric be Both Testable and Actionable? In *Proceedings of the 38th Conference on Learning Theory*, 2025.

[60] Samuel Showalter, Alex J Boyd, Padhraic Smyth, and Mark Steyvers. Bayesian Online Learning for Consensus Prediction. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, 2024.

[61] Mark Steyvers and Aakriti Kumar. Three Challenges for AI-Assisted Decision-Making. *Perspectives on Psychological Science*, 2024.

[62] Mark Steyvers, Heliodoro Tejeda, Gavin Kerrigan, and Padhraic Smyth. Bayesian Modeling of Human–AI Complementarity. *Proceedings of the National Academy of Sciences*, 2022.

[63] Dharmesh Tailor, Aditya Patra, Rajeev Verma, Putra Manggala, and Eric Nalisnick. Learning to Defer to a Population: A Meta-Learning Approach. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, 2024.

[64] Heliodoro Tejeda Lemus, Aakriti Kumar, and Mark Steyvers. How Displaying AI Confidence Affects Reliance and Hybrid Human-AI Performance. *HHAI 2023: Augmenting Human Intellect*, 2023.

[65] Philip E Tetlock and Dan Gardner. *Superforecasting: The Art and Science of Prediction*. Random House, 2016.

[66] Pete Trautman. A Mathematical Theory of Human Machine Teaming. *ArXiv Preprint*, 2017.

[67] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 Dataset, a Large Collection of Multi-Source Dermatoscopic Images of Common Pigmented Skin Lesions. *Scientific Data*, 2018.

[68] Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. Evaluating Model Calibration in Classification. In *Conference on Artificial Intelligence and Statistics*, 2019.

[69] Rajeev Verma and Eric Nalisnick. Calibrated Learning to Defer with One-vs-All Classifiers. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

[70] Rajeev Verma, Daniel Barrejón, and Eric Nalisnick. Learning to Defer to Multiple Experts: Consistent Surrogate Losses, Confidence Calibration, and Conformal Ensembles. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, 2023.

[71] Kailas Vodrahalli, Tobias Gerstenberg, and James Y Zou. Uncalibrated Models can Improve Human-AI Collaboration. *Advances in Neural Information Processing Systems*, 2022.

[72] Edward Vul and Harold Pashler. Measuring the Crowd Within: Probabilistic Representations within Individuals. *Psychological Science*, 2008.

[73] Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to Complement Humans. In *International Joint Conference on Artificial Intelligence*, 2020.

[74] Darryl E. Wright, Chris J. Lintott, Stephen J. Smartt, Ken W. Smith, Lucy Fortson, Laura Trouille, Campbell R. Allen, Melanie Beck, Mark C. Bouslog, Amy Boyer, K. C. Chambers, Heather Flewelling, Will Granger, Eugene A. Magnier, Adam McMaster, Grant R. M. Miller, James E. O'Donnell, Brooke Simmons, Helen Spiers, John L. Tonry, Marten Veldthuis, Richard J. Wainscoat, Chris Waters, Mark Willman, Zach Wolfenbarger, and Dave R. Young. A Transient Search Using Combined Human and Machine Classifications. *Monthly Notices of the Royal Astronomical Society*, 2017.

[75] Feiyang Yu, Alex Moehring, Oishi Banerjee, Tobias Salz, Nikhil Agarwal, and Pranav Rajpurkar. Heterogeneity and Predictors of the Effects of AI Assistance on Radiologists. *Nature Medicine*, 2024.

[76] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2020.

[77] Shengjia Zhao, Michael Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. Calibrating Predictions to Decisions: A Novel Approach to Multi-Class Calibration. In *Advances in Neural Information Processing Systems*, 2021.

A Proof of Theorem 3.2

Proof. The human's sample satisfies $\hat{\mathrm{h}} \perp (\mathrm{y}, \mathbf{f}(\mathbf{x})) \mid \Phi_h(\mathbf{x})$. By this independence together with conditions (1) and (2), the shared evidence value $(\hat{\mathrm{h}}, \mathbf{f}(\mathbf{x})) = (y', \boldsymbol{v})$ has positive probability under both cells:

$$\mathbb{P}\Big(\hat{\mathrm{h}} = y', \mathbf{f}(\mathbf{x}) = \boldsymbol{v}, \Phi_h(\mathbf{x}) = S_i\Big) \;=\; \underbrace{\mathbb{P}\Big(\hat{\mathrm{h}} = y' \mid \Phi_h(\mathbf{x}) = S_i\Big)}_{>0 \text{ by (1)}} \; \underbrace{\mathbb{P}(\mathbf{f}(\mathbf{x}) = \boldsymbol{v}, \Phi_h(\mathbf{x}) = S_i)}_{>0 \text{ by (2)}} \;>\; 0.$$

On this set we have, for $i = 1, 2$,

$$\mathbb{P}\Big(\mathrm{y} = y \mid \hat{\mathrm{h}} = y', \mathbf{f}(\mathbf{x}) = \boldsymbol{v}, \Phi_h(\mathbf{x}) = S_i\Big) \;=\; \mathbb{P}(\mathrm{y} = y \mid \mathbf{f}(\mathbf{x}) = \boldsymbol{v}, \Phi_h(\mathbf{x}) = S_i)\,, \tag{6}$$

since conditioning further on $\hat{\mathrm{h}} = y'$ adds no information about y given $(\Phi_h, \mathbf{f})$. The two sides of Equation 6 differ for $i = 1, 2$ by the differing-content requirement of (3). Therefore conditioning additionally on $\Phi_h(\mathbf{x})$ changes the label probability beyond $(\hat{\mathrm{h}}, \mathbf{f}(\mathbf{x}))$, i.e.

$$\mathbb{P}\Big(\mathrm{y} \mid \mathbf{f}(\mathbf{x}), \hat{\mathrm{h}}, \Phi_h(\mathbf{x})\Big) \;\neq\; \mathbb{P}\Big(\mathrm{y} \mid \mathbf{f}(\mathbf{x}), \hat{\mathrm{h}}\Big)$$

on this positive-probability set.

Write $c \triangleq g^*(y', \boldsymbol{v})$ and $H_c \triangleq \{h \in \mathcal{Y} \mid g^*(h, \boldsymbol{v}) = c\} \ni y'$. By condition (4), every value with $g^* = c$ has model output $\boldsymbol{v}$, so $\{g^* = c\} = \{\hat{\mathrm{h}} \in H_c,\ \mathbf{f}(\mathbf{x}) = \boldsymbol{v}\}$. Since $\hat{\mathrm{h}} \perp (\mathrm{y}, \mathbf{f}(\mathbf{x})) \mid \Phi_h(\mathbf{x})$, conditioning on $\hat{\mathrm{h}} \in H_c$ adds no information about y given $(\mathbf{f}(\mathbf{x}), \Phi_h(\mathbf{x}))$, and therefore, for $i = 1, 2$,

$$\begin{aligned}\mathbb{P}(\mathrm{y} \mid g^* = c, \Phi_h(\mathbf{x}) = S_i) \;&=\; \mathbb{P}\Big(\mathrm{y} \mid \hat{\mathrm{h}} \in H_c, \mathbf{f}(\mathbf{x}) = \boldsymbol{v}, \Phi_h(\mathbf{x}) = S_i\Big) \\ &=\; \mathbb{P}(\mathrm{y} \mid \mathbf{f}(\mathbf{x}) = \boldsymbol{v}, \Phi_h(\mathbf{x}) = S_i)\,,\end{aligned}$$

where both conditioning events have positive probability, because $\mathbb{P}(\hat{\mathrm{h}} \in H_c \mid \Phi_h(\mathbf{x}) = S_i) \geq \mathbb{P}(\hat{\mathrm{h}} = y' \mid \Phi_h(\mathbf{x}) = S_i) > 0$ by (1) and $\mathbb{P}(\mathbf{f}(\mathbf{x}) = \boldsymbol{v}, \Phi_h(\mathbf{x}) = S_i) > 0$ by (2). By (3), the right-hand sides differ for $i = 1$ and $i = 2$ at the label y'', so they cannot both equal $\mathbb{P}(\mathrm{y} = y'' \mid g^* = c)$. Hence $\mathbb{P}(\mathrm{y} \mid g^*, \Phi_h) \neq \mathbb{P}(\mathrm{y} \mid g^*)$ on a set of positive probability, which is exactly the failure of generalized calibration (Definition 2.2). □

B Proof of Theorem 3.3

Proof. Fix a label value $\mathrm{y} = y$ and a realized human prediction $\mathrm{h} = h$. By Bayes rule,

$$\mathbb{P}\left(\mathrm{y} = y \mid \mathrm{h} = h, \Phi_f(\mathbf{x})\right) = \frac{\mathbb{P}\left(\mathrm{h} = h \mid \mathrm{y} = y, \Phi_f(\mathbf{x})\right)\mathbb{P}\left(\mathrm{y} = y \mid \Phi_f(\mathbf{x})\right)}{\mathbb{P}\left(\mathrm{h} = h \mid \Phi_f(\mathbf{x})\right)}.$$

It remains to simplify the likelihood term $\mathbb{P}\left(\mathrm{h} = h \mid \mathrm{y} = y, \Phi_f(\mathbf{x})\right)$. Since the human is calibrated with respect to Φ_h, the human prediction depends on $\mathbf{x}$ through $\Phi_h(\mathbf{x})$, so

$$\mathbb{P}\left(\mathrm{h} = h \mid \mathrm{y} = y, \Phi_f(\mathbf{x}), \Phi_h(\mathbf{x})\right) = \mathbb{P}\left(\mathrm{h} = h \mid \Phi_h(\mathbf{x})\right).$$

Therefore, by the tower rule,

$$\mathbb{P}\left(\mathrm{h} = h \mid \mathrm{y} = y, \Phi_f(\mathbf{x})\right) = \mathbb{E}_{\Phi_h}\left[\mathbb{P}\left(\mathrm{h} = h \mid \Phi_h(\mathbf{x})\right) \mid \mathrm{y} = y, \Phi_f(\mathbf{x})\right].$$

Using the assumption

$$\Phi_h \perp \Phi_f(\mathbf{x}) \mid \mathrm{y},$$

we can remove the conditioning on $\Phi_f(\mathbf{x})$:

$$\mathbb{P}\left(\mathrm{h} = h \mid \mathrm{y} = y, \Phi_f(\mathbf{x})\right) = \mathbb{E}_{\Phi_h}\left[\mathbb{P}\left(\mathrm{h} = h \mid \Phi_h(\mathbf{x})\right) \mid \mathrm{y} = y\right].$$

Substituting this into Bayes rule gives

$$\mathbb{P}\left(\mathrm{y} = y \mid \mathrm{h} = h, \Phi_f(\mathbf{x})\right) = \frac{\mathbb{E}_{\Phi_h}\left[\mathbb{P}\left(\mathrm{h} = h \mid \Phi_h(\mathbf{x})\right) \mid \mathrm{y} = y\right] \cdot \mathbb{P}\left(\mathrm{y} = y \mid \Phi_f(\mathbf{x})\right)}{\mathbb{P}\left(\mathrm{h} = h \mid \Phi_f(\mathbf{x})\right)}.$$

Since the model is calibrated with respect to Φ_f,

$$f_y(\mathbf{x}) = \mathbb{P}\left(\mathrm{y} = y \mid \Phi_f(\mathbf{x})\right).$$

Thus the Bayesian combination posterior computed from $f(\mathbf{x})$ and the human confusion probabilities equals the true conditional posterior:

$$p\left(\mathrm{y} = y \mid \mathrm{h} = h, f(\mathbf{x})\right) = \mathbb{P}\left(\mathrm{y} = y \mid \mathrm{h} = h, \Phi_f(\mathbf{x})\right).$$

Since this holds for every y and every realized value h, we obtain

$$p\left(\mathrm{y}|\mathrm{h}, f(\mathbf{x})\right) = \mathbb{P}\left(\mathrm{y} \mid \mathrm{h}, \Phi_f(\mathbf{x})\right) = \frac{\mathbb{E}_{\Phi_h}\left[\mathbb{P}(\mathrm{h}|\Phi_h(\mathbf{x})) \mid \mathrm{y}\right] \cdot \mathbb{P}\left(\mathrm{y}|\Phi_f(\mathbf{x})\right)}{\mathbb{P}(\mathrm{h}|\Phi_f(\mathbf{x}))}.$$

Therefore the combination posterior is calibrated with respect to Φ_f in the sense of Definition 2.2. □

C Proof of Proposition 3.5

Proof. The human's prediction is drawn from $\mathbb{P}(\mathrm{h} \mid \Phi_h(\mathbf{x}))$ independently of the label, i.e. $\mathrm{h} \perp \mathrm{y} \mid \mathbf{x}$, and the true posterior is $\mathbb{P}(\mathrm{y} \mid \mathrm{h}, \mathbf{x}) = \mathbb{P}(\mathrm{y} \mid \mathbf{x}) = \mathbf{f}(\mathbf{x})$. The Bayesian combination computes

$$p(\mathrm{y} = y \mid \mathrm{h} = h, \mathbf{f}(\mathbf{x})) \;=\; \frac{c_{y,h}\, f_y(\mathbf{x})}{\sum_{j \in \mathcal{Y}} c_{j,h}\, f_j(\mathbf{x})}, \qquad c_{y,h} = \mathbb{P}(\mathrm{h} = h \mid \mathrm{y} = y).$$

Fix $\boldsymbol{x}$ in the set with full support over $\mathcal{Y}$. Since $f_y(\boldsymbol{x}) > 0$ for every y, the two posteriors agree at $(h, \boldsymbol{x})$ if and only if $c_{y,h} = \sum_j c_{j,h} f_j(\boldsymbol{x})$ for every y. Moreover, at such $\boldsymbol{x}$, $\mathbb{P}(\mathrm{h} = h \mid \mathbf{x} = \boldsymbol{x}) = \mathbb{P}(\mathrm{y} = h \mid \Phi_h(\boldsymbol{x})) > 0$ whenever the cell $\Phi_h(\boldsymbol{x})$ meets the full-support set. Hence, if the two posteriors agreed, $\mathbb{P}(\mathrm{h} = h \mid \mathrm{y} = y)$ would be constant in y for every h, i.e. $\mathrm{h} \perp \mathrm{y}$, contradicting the supposition that the human is informative. Therefore $p(\mathrm{y} \mid \mathrm{h}, \mathbf{f}(\mathbf{x})) \neq \mathbb{P}(\mathrm{y} \mid \mathbf{x})$. □

D Proof of Theorem 4.1

Proof. The conditional 0–1 risk at $\mathbf{x}$ in Equation 3 is

$$(1 - r(\mathbf{x}))\left(1 - \max_y f_y(\mathbf{x})\right) \;+\; r(\mathbf{x})(1 - \mathbb{P}(\mathrm{h} = \mathrm{y} \mid \mathbf{x}))\,,$$

which is minimized pointwise by $r^*(\mathbf{x}) = \mathbf{1}[\mathbb{P}(\mathrm{h} = \mathrm{y} \mid \mathbf{x}) > \max_y f_y(\mathbf{x})] = \mathbf{1}_{\mathfrak{D}^*}(\mathbf{x})$. Subtracting the pointwise optimum yields the excess risk

$$\mathcal{R}^{0\text{-}1}_{f,h}(r) - \mathcal{R}^{0\text{-}1}_{f,h}(r^*) \;=\; \mathbb{E}_{\mathbf{x}}\left[\left|\mathbb{P}(\mathrm{h} = \mathrm{y} \mid \mathbf{x}) - \max_y f_y(\mathbf{x})\right| \cdot \mathbf{1}\{r(\mathbf{x}) \neq \mathbf{1}_{\mathfrak{D}^*}(\mathbf{x})\}\right] \;\geq\; 0, \quad (7)$$

since whenever $r(\mathbf{x})$ takes the suboptimal value it incurs the margin $|\mathbb{P}(\mathrm{h} = \mathrm{y} \mid \mathbf{x}) - \max_y f_y(\mathbf{x})|$. As ties have probability zero, Equation 7 vanishes if and only if $r(\mathbf{x}) = \mathbf{1}_{\mathfrak{D}^*}(\mathbf{x})$ almost everywhere.

Fix a $\tilde{\Phi}_r$-cell A. By assumption Φ_r refines the model-output coordinate, so $\mathbf{f}$ is $\tilde{\Phi}_r$-measurable and $\max_y f_y$ is constant on A; write $c_A \triangleq \max_y f_y(\mathbf{x})$ for $\mathbf{x} \in A$. By the assumed calibration of ρ w.r.t. $\tilde{\Phi}_r$, the score is likewise constant on A, with value

$$\rho_A \;=\; \mathbb{P}(\mathrm{h} = \mathrm{y} \mid \tilde{\Phi}_r(\mathbf{x}) = A) \;=\; \mathbb{E}[\mathbb{P}(\mathrm{h} = \mathrm{y} \mid \mathbf{x}) \mid \mathbf{x} \in A].$$

Since both ρ and $\max_y f_y$ are constant on A, the rejector $r(\mathbf{x}; \rho) = \mathbf{1}[\rho_A > c_A]$ is constant on A as well.

($\Leftarrow$) Assume $\sigma(\mathbf{1}_{\mathfrak{D}^*}) \subseteq \sigma(\tilde{\Phi}_r)$. Then each $\tilde{\Phi}_r$-cell A lies entirely in $\mathfrak{D}^*$ or in $(\mathfrak{D}^*)^c$. If $A \subseteq \mathfrak{D}^*$, then $\mathbb{P}(\mathrm{h} = \mathrm{y} \mid \mathbf{x}) > c_A$ on A (with c_A constant on A), so $\rho_A = \mathbb{E}[\mathbb{P}(\mathrm{h} = \mathrm{y} \mid \mathbf{x}) \mid A] > c_A$ and $r(\mathbf{x}; \rho) = 1 = \mathbf{1}_{\mathfrak{D}^*}(\mathbf{x})$ on A. If $A \subseteq (\mathfrak{D}^*)^c$, then $\mathbb{P}(\mathrm{h} = \mathrm{y} \mid \mathbf{x}) < c_A$ on A, so $\rho_A < c_A$ and $r(\mathbf{x}; \rho) = 0 = \mathbf{1}_{\mathfrak{D}^*}(\mathbf{x})$ on A. Hence $r(\mathbf{x}; \rho) = \mathbf{1}_{\mathfrak{D}^*}$, and by Equation 7 the excess risk is zero.

($\Rightarrow$) Suppose $r(\mathbf{x}; \rho)$ is risk-minimizing. The score $\rho(\mathbf{x}, \mathbf{f}(\mathbf{x}))$ is $\tilde{\Phi}_r$-measurable by the assumed calibration, and $\mathbf{f}(\mathbf{x})$ is $\tilde{\Phi}_r$-measurable by the refinement hypothesis; hence $r(\mathbf{x}; \rho) = \mathbf{1}[\rho > \max_y f_y]$ is $\tilde{\Phi}_r$-measurable. By Equation 7 and the no-tie assumption, $r(\mathbf{x}; \rho) = \mathbf{1}_{\mathfrak{D}^*}(\mathbf{x})$ almost everywhere. Therefore $\mathbf{1}_{\mathfrak{D}^*}$ agrees with a $\tilde{\Phi}_r$-measurable function, giving $\sigma(\mathbf{1}_{\mathfrak{D}^*}) \subseteq \sigma(\tilde{\Phi}_r)$. □

E Experimental Details for Bayesian Combination Experiments

All experiments were run on a SLURM cluster with nodes of 8 x NVIDIA RTX A5000 (24 GB) with 64 CPU cores and 256 GB RAM per node. All jobs used a single GPU. The software we used was: Python 3.10.20, PyTorch 2.5.1+cu121, torchvision 0.20.1+cu121, NumPy 2.2.6, scikit-learn 1.7.2, pandas 2.3.3. Regarding runtime, no single experimental training run or temperature sweep required more than 20 minutes of computation.

E.1 ImageNet-16H

We use ImageNet-16H from the public OSF release of Kerrigan et al. [34], which contains human classifications and precomputed CNN softmax outputs on a 16-class subset of ImageNet: https://osf.io/2ntrf/overview. We focus on the VGG-19 predictions at image noise level 80 and evaluate the performance of the resulting human-AI combination as the classifier's standalone calibration changed. For each image $\boldsymbol{x}$, the dataset provides approximately six human class annotations. We aggregate these annotations into a count vector over the $K = 16$ classes and interpret the normalized count vector as the empirical human response distribution for that image. To match the theoretical setting in which we observe a sampled human prediction rather than the human's full internal distribution, we draw a single hard human label for each image.

We split the examples into a calibration subset and a held-out evaluation subset. For each random run, we use a stratified $70/30$ split. The calibration subset is used only to estimate the human confusion matrix $\boldsymbol{C}$, while the held-out subset is used to evaluate the standalone classifier and the combined predictor. In all ImageNet-16H experiments reported here, we use all calibration examples to estimate $\boldsymbol{C}$. We repeat the procedure over five random seeds, where each run re-samples the calibration/evaluation split and independently samples the hard human labels on both subsets.

To vary the calibration of the classifier without changing its learned representation, we apply fixed temperature scaling to the VGG-19 softmax probabilities. Smaller temperatures sharpen the probability vector, whereas larger temperatures flatten it. We sweep prespecified grids of temperatures and evaluate both the standalone classifier and the confusion-matrix combined predictor using the same held-out examples.

We report three quantities: mass-binned expected calibration error, negative log-likelihood, and accuracy. For ECE, we use 15 equal-mass confidence bins. The main plots place standalone model ECE on the horizontal axis and a combined-predictor metric on the vertical axis. Each point corresponds to one fixed temperature T, and the plotted value is averaged across random runs.

E.2 HAM10000

We use the HAM10000 skin-lesion classification dataset to construct a controlled setting in which the simulated expert prediction can depend on different information sources. The training set is used to train the image classifier and expert predictors, the calibration set is used to estimate expert confusion matrices, and the test set is used for final evaluation. Each example contains a dermoscopic image and patient- or lesion-level meta-data. We identify the image with the model-observed feature $\mathbf{x}$ and treat the meta-data as privileged information $\mathbf{z}$ that is unavailable to the base classifier but may be available to an expert. We simulate the expert prediction using trained auxiliary predictors. This allows us to compare several feature-access regimes while keeping the same confusion-matrix update framework.

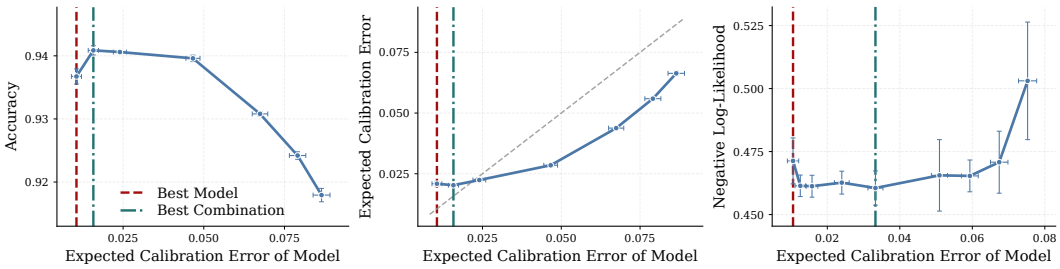
The model is a temperature-scaled ResNet-18 trained on 224×224 dermoscopic images for 8 epochs. We compare four expert settings. The first, labeled *Image*, uses the same ResNet-18 output as the main classifier, representing an identical-image, duplicated-evidence expert. The second, labeled *Weaker Image*, is a separate weaker ResNet-18 trained on 128×128 images for 4 epochs, representing a shared-feature setting with a different image predictor. The third, labeled *Image and Meta-Data*, is a multimodal expert consisting of a ResNet-18 image encoder and a meta-data MLP, representing a partial-overlap setting because it uses both the image and patient meta-data. The fourth, labeled *Meta-Data Only*, is a logistic regression expert trained on age, sex, and lesion localization. In contrast to the previous three experts, it depends only on $\mathbf{z}$ and not on the image feature $\mathbf{x}$, making it the most independent expert relative to the image classifier.

For each expert, we treat its softmax output as the expert's internal predictive distribution and sample hard labels from it. Thus, as in the ImageNet-16H experiment, the combination method observes only a sampled categorical expert prediction rather than the full expert distribution. On the calibration split, we sample four hard expert labels per example to estimate $\boldsymbol{C}$ more stably. On the test split, we sample one hard expert label per example and combine it with the temperature-scaled image classifier. As before, we vary only the temperature of the base image classifier and keep the expert distributions fixed. This isolates how the calibration of the model affects the resulting combined predictor under different expert information structures. In all plots, the horizontal axis is the mass-binned ECE of the standalone image classifier, and the vertical axis is the metric of the combined predictor. For all experiments, we average over 5 random sampling runs. The optimizer was AdamW, the learning rate was 10^{-4}, augmentations were random horizontal flips and rotations, and the split sizes were $6,958$ train / $1,502$ validation / $1,498$ test.

E.3 CIFAR-10H

We additionally evaluate the confusion-matrix combination method on CIFAR-10H [55], a human uncertainty dataset built on the CIFAR-10 test set. CIFAR-10H provides human label distributions for each image, which allows us to sample hard human predictions while preserving real human perceptual uncertainty. We use ResNet-18 as the base image classifier. For each image, we sample one hard human label from the empirical CIFAR-10H human label distribution. To study a weaker-human setting, we flatten the empirical human distribution using a human temperature of $T_h = 2.0$ before sampling. We do not add extra random label corruption. Thus, the sampled human labels remain tied to the real per-image human uncertainty in CIFAR-10H, but with reduced confidence.

For each run, we split the data into a calibration subset and a held-out evaluation subset using a stratified $70/30$ split. The calibration subset is used to estimate the human confusion matrix $\boldsymbol{C}$, and the held-out subset is used to evaluate both the standalone temperature-scaled classifier and the



(a) Accuracy (b) Calibration Error (c) Neg. Log-Likelihood

Figure 4: *Additional Validation on CIFAR-10H.* We vary the calibration of a ResNet classifier by temperature scaling and combine each temperature-scaled classifier with a sampled human label using the confusion-matrix update. The human label is sampled from the CIFAR-10H empirical human distribution after flattening with human temperature $T_h = 2.0$. The horizontal axis reports the standalone model's mass-binned ECE, while the vertical axes report the combined predictor's accuracy, ECE, and NLL. Error bars report one standard error.

combined predictor. We use $5{,}000$ calibration examples to estimate $\boldsymbol{C}$ and average results over five random runs. In each run, the train-test split and the sampled human labels are regenerated. We vary only the classifier temperature while keeping the human sampling procedure fixed. We report combined accuracy, combined ECE, and combined NLL as functions of the standalone model's mass-binned ECE. The results in Figure 4 provide an additional real-human-label validation that the best standalone model calibration need not correspond to the best combined predictor.

F Robustness to Noisy and Biased Human Predictions

Assumption 2.3 requires the human to be calibrated w.r.t. Φ_h, yet in practice human predictions may be noisy, systematically biased, or otherwise misleading. Our theoretical results are stated for the good-case scenario in which calibration holds, and we do not propose a remedy for its violation. Nevertheless, it is natural to ask whether the qualitative prediction of Section 3.2 holds when the human is corrupted. This experiment addresses that question. We use the HAM10000 setting of Appendix E and Figures 2e and 2f without modification, except that the sampled expert label $\hat{\mathrm{h}}$ is corrupted before it is passed to the confusion-matrix combination. The corruption is applied identically on the calibration split (where $\boldsymbol{C}$ is estimated) and on the test split (where the posterior is evaluated), so that the combination method sees a corrupted human throughout, exactly as it would in deployment. The four simulated experts (*Image*, *Weaker Image*, *Image and Meta-Data*, *Meta-Data Only*), the temperature-scaled ResNet-18 model, the temperature grid, and the five random seeds are as described in Appendix E. We consider two types of corruption:

- **Noise**: With probability $\epsilon \in \{0.1, 0.2, 0.3\}$, the sample $\hat{\mathrm{h}}$ is replaced by a label drawn uniformly at random from $\mathcal{Y} \setminus \{\mathrm{y}\}$, i.e. a uniformly random *incorrect* class. Because the replacement is never correct, this is label-dependent noise that strictly reduces the human's accuracy.
- **Bias**: With probability $b \in \{0.2, 0.4\}$, the sampled label $\hat{\mathrm{h}}$ is replaced by a fixed class: either the majority class or the minority class of the training set. This models a human with a systematic tendency to over-predict a particular diagnosis.

This yields seven corruption configurations in total: three noise levels, and two bias strengths for each of the two target classes. In every configuration the corrupted human is, by construction, no longer calibrated w.r.t. Φ_h, so Assumption 2.3 is violated for the human.

Table 1 reports, for each expert and metric, the number of the seven configurations for which the best team is formed by the best-calibrated model. The qualitative prediction holds in 55 of the 56 expert–metric–configuration combinations. For the *Meta-Data Only* expert, the best-calibrated classifier yields the best posterior in both ECE and NLL under every one of the seven corruptions, as the independent-feature analysis predicts. For the three experts whose features overlap with the image classifier's, the best-calibrated classifier never yields the best posterior ECE, and yields the best posterior NLL in a single case: the *Weaker Image* expert under the strongest noise level, $\epsilon = 0.3$. This

Table 1: *Robustness to a noisy or biased human.* For each simulated HAM10000 expert and each of seven corruption configurations, we check whether the best-calibrated classifier also yields the best posterior predictor. Entries give the number of configurations (out of 7) for which it does. We expect $7/7$ for the independent-feature *Meta-Data Only* expert and $0/7$ for all other experts.

Simulated Expert	Metric	Best Model = Best Posterior	Coincident Configurations
Image	ECE	$0/7$	None
	NLL	$0/7$	None
Weaker Image	ECE	$0/7$	None
	NLL	$1/7$	Noisy with $\epsilon = 0.3$
Image and Meta-Data	ECE	$0/7$	None
	NLL	$0/7$	None
Meta-Data Only	ECE	$7/7$	All
	NLL	$7/7$	All

one deviation is plausibly explained by the corruption itself—at 30% noise the weaker image expert carries little usable information, possibly violating the assumption that the human is an informative predictor.

G LLM Use

Claude Opus 4.6 was used to check the proofs for correctness when writing the initial draft. Later, Claude Fable 5.1 was used to re-check proofs for correctness and to draft revised proofs. OpenAI Codex was used to help implement the experiments reported in Figures 2 and 4. All initial writing was done without LLMs. Claude Opus 4.6 proofread the draft before submission, and Claude Fable 5.1 provided proofreading before the camera-ready deadline.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The results and methods support the paper's main claims.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Limitations are discussed in paper.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.

- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: All proofs are included.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: All information disclosed in appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.

- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: All data is publicly available. Code will be linked on authors' personal websites.

   Guidelines:

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

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: All training and test details are disclosed in appendix.

Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: Standard error bars are provided using five random seeds.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

Answer: [Yes]

Justification: Details disclosed in appendix.

Guidelines:

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

9. **Code of ethics**

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

Answer: [Yes]

Justification: Yes, we followed the code of ethics.

Guidelines:

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

10. **Broader impacts**

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

Answer: [Yes]

Justification: The impacts of the work are discussed in the Discussion section and the Introduction Section.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

Answer: [N/A]

Justification: Theoretical paper. No safeguards needed.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All assets credited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: No new assets provided.

    Guidelines:

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

14. **Crowdsourcing and research with human subjects**

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

    Answer: [N/A]

    Justification: No human subjects involved.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

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

    Answer: [N/A]

    Justification: No IRB required.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: In the appendix, we describe our use of LLMs for validating the theory and helping implement the simulations / experiments.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.