DEFER-AND-FUSION: OPTIMAL PREDICTORS THAT INCORPORATE HUMAN DECISIONS

Mohammad-Amin Charusaie
Max Planck Institute for Intelligent Systems
Tuebingen, Germany
mcharusaie@tuebingen.mpg.de

Amirmehdi Fesharaki
Department of Electrical Engineering
Sharif University of Technology
Tehran, Iran

Samira Samadi
Max Planck Institute for Intelligent Systems
Tuebingen, Germany
samira.samadi@tuebingen.mpg.de

ABSTRACT

Learning predictors that incorporate human decision has been the focus of extensive research in recent years. These predictors are used in order to increase the final accuracy and reduce the risk in high-stake tasks. One of the strategies to keep the human in the loop involves using learn-to-defer methods, in which the prediction is made either by AI or is deferred to the human expert. This strategy has attracted considerable attention due to the reduction of expert bandwidth as well as increasing the accuracy. However, we show that learn-to-defer methods are not optimal considering their understanding of the task and human decision. In this paper, we first derive the optimal predictor that provides the defer option, while incorporating human decision into its final prediction. We further show strict improvement of this method upon learn-to-defer methods, both theoretically and empirically. The code for this paper is open-sourced on https://github.com/AminChrs/BeyondDefer.

1 INTRODUCTION

Recent advances in training machine learning (ML) models have made these models potential resources in high-stake practical applications such as medical diagnosis Beede et al. (2020); Raghu et al. (2019); van Leeuwen et al. (2021). Although these models in some applications have shown higher accuracy compared to human experts Rajpurkar et al. (2018); Killock (2020), their predictions lead to errors on different sets of data points too McKinney et al. (2020). This implies a possible complementary function of human experts and ML models.

One solution for achieving such complementarity is through learn-to-defer methods El-Yaniv et al. (2010); Madras et al. (2018) in which the goal is to find a rejection function that decides when the final prediction is made by the classifier, and when it should be made by the human expert. This line-of-work mainly makes use of surrogate loss functions to minimize the overall error in such a system by obtaining regions in which the human expert or the ML model have higher confidence Mozannar & Sontag (2020); Verma & Nalisnick (2022); Charusaie et al. (2022); Mozannar et al. (2023); Cao et al. (2023).

Another line-of-work, that mainly focuses on refining ML predictions rather than the complementarity of human and ML, is to use human-AI teaming methods. In these methods, human advice is always sought and then is combined with ML decision Kerrigan et al. (2021); Donahue et al. (2022). As one can imagine, while these systems can improve upon human and ML model accuracy, they induce a large cost of seeking human advice.

To find a middle-ground between the above two human-in-the-loop methodologies, in this paper, we introduce the Defer-and-Fusion (DaF) algorithm, in which the goal is to find regions that the human advice should be sought, while at the same time, we find an optimal fusion of human advice and

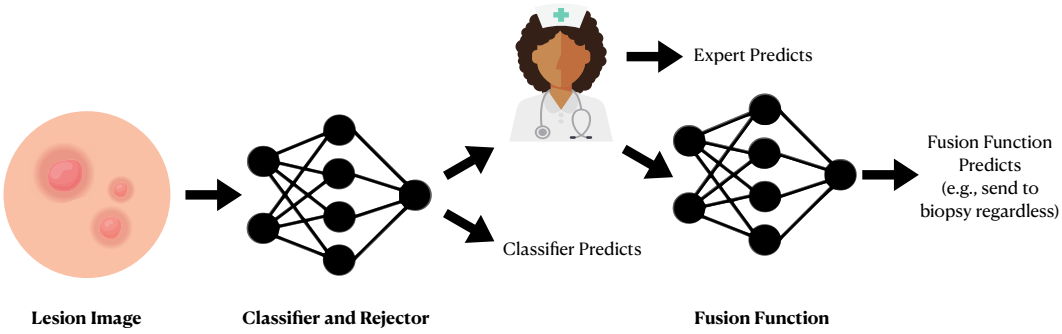


Figure 1: The description of a DaF system in medical applications

ML prediction. In this paper, we show that while our algorithm shows higher accuracy compared to learn-to-defer methods on a variety of datasets such as Imagenet-16H Steyvers et al. (2022), COMPAS Dressel & Farid (2018), NIH Chest X-rayWang et al. (2017), it further limits the limits the need for human advice as in human-AI teaming methods.

We further provide the reader with a set of theoretical scenarios in which the strict improvement of DaF method over learn-to-defer methods is possible. In particular, we show that in tasks where false positive errors and false negative errors have different importance (such as medical diagnosis setting), learn-to-defer methods can be strictly sub-optimal compared to DaF methods. Furthermore, we extend this result to tasks with 0-1 loss. There, using a variation of Fano's inequality we show that the minimum probability of error in learn-to-defer methods is lower-bounded with a higher value than that in DaF methods. Moreover, we show that such bound is almost tight, which means if the gap between the bounds is larger than a certain value, the probability of error in all learn-to-defer systems is strictly larger than that in a DaF system.

2 Problem Setting

In a classification problem, the learn-to-defer strategy optimizes the setting in which the prediction is either provided by the classifier or deferred to a human expert. More formally, this strategy minimizes the overall prediction risk in such a setting. More formally, for predicting a target $y \in \mathcal{Y}$ using a feature set $x \in \mathcal{X}$ and human decision $m \in \mathcal{Y}$, and letting the loss function corresponding to the classifier and expert prediction be respectively $\ell, \ell_H : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, the goal of the learn-to-defer strategy is to minimize

$$L_{\text{def}}(h, r) = \mathbb{E}_{(x,y,m)\sim\mu_{XYM}}\left[\mathbb{I}_{r(x)=0}\ell(h(x), y) + \mathbb{I}_{r(x)=1}\ell_H(m, y)\right], \tag{1}$$

for a classifier $h : \mathcal{X} \to \mathcal{Y}$ and a rejection function $r : \mathcal{X} \to \{0, 1\}$, and where the joint probability distribution of the features, targets, and human decisions is μ_{XYM}. Here, we observe that when the deferral occurs, the human expert provides the final prediction of the setting. In fact, one can see that in such cases no learning happens. Therefore, in order to further reduce the second term in equation 1, the defer-and-fusion setting suggests that we add the option in which the final prediction is a function g of human decision m and the feature set x. In such cases, the overall risk is obtained as

$$L_{\text{DaF}}(h, r, g) = \mathbb{E}_{(x,y,m)\sim\mu_{XYM}}\left[\mathbb{I}_{r(x)=0}\ell(h(x), y) + \mathbb{I}_{r(x)=1}\ell_H(m, y) + \mathbb{I}_{r(x)=2}\ell_{\text{fus}}(g(m, x), y)\right], \tag{2}$$

where ℓ_{fus} is the loss function that is corresponded to the introduced *fusion* option.

A first observation from the above equation is that if $g(\cdot, \cdot)$ can take the identity function $g(m, x) = m$ and if the hypothesis class of rejection functions is rich enough, and in case of $\ell_{\text{fus}} = \ell_H$, then the minimizer of $L_{\text{DaF}}(h, r, g)$ incurs smaller amount of loss than the minimizer of $L_{\text{def}}(h, r)$, i.e.,

$$\min_{(h,r,g)\in\mathcal{H}\times\mathcal{R}\times\mathcal{G}} L_{\text{DaF}}(h, r, g) \leq \min_{(h,r)\in\mathcal{H}\times\mathcal{R}} L_{\text{def}}(h, r). \tag{3}$$

The reason for this is that in such case we can by setting the fusion to be an identity function, and by setting $r(x) = 2$ whenever the rejection occurs in the optimal learn-to-defer strategy, we can at

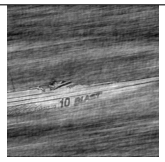
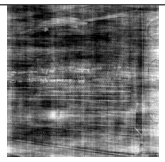
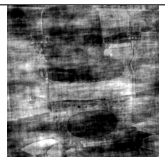
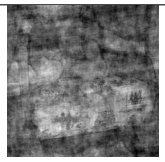
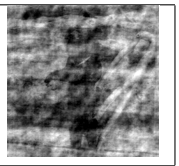
Image					
True Label	Boat	Oven	Chair	Chair	Bear
Classifier	Boat	Oven	Chair	Chair	Bear
Human	Airplane	Airplane	Oven	Elephant	Dog
Learn-to-Defer	Airplane	Airplane	Oven	Elephant	Dog
DaF	Boat	Oven	Chair	Chair	Bear

Table 1: Five examples of correcting learn-to-defer method using DaF

least achieve the same amount of DaF loss as that of learn-to-defer. This, consequently, shows that if the hypothesis class of rejection functions is rich enough, we could safely omit the second term in equation 2 without compromising accuracy. However, our results show that in the lack of such an option, the sample complexity that is needed for learning fusion function is significantly larger. Therefore, the results that are reported in this paper consider this term in the loss function.

Due to the discontinuous and non-convex form of the loss function, optimization of this loss using typical learning methods is not possible. As a result, in the following section, we provide the Bayes optimal solution to this loss. Such theoretical result helps us to further design methods (e.g., consistent surrogate methods) to achieve that solution.

3 Optimal Defer-and-Fusion Predictor

To obtain the optimal Defer-and-Fusion predictor, we need to introduce a triple of the classifier, the rejection function, and the fusion function. Luckily, the following theorem shows that the optimal classifier and fusion function are respectively equivalent to the Bayes optimal classifier when it has access only to the feature set, and when it has access to both the feature set and human expert decision. Furthermore, the rejection function is shown to select a predictor among classifier, human expert, and fusion function, by optimizing an extended notion of confidence.

Theorem 1 *An optimal DaF system is achieved by a triple of optimal classifier* $h^*(x) = \arg\min_{\hat{y}} \mathbb{E}_{y|x}\big[\ell(\hat{y}, y)\big]$*, the optimal fusion function* $g^*(m, x) = \arg\min_{\hat{y}} \mathbb{E}_{y|x,m}\big[\ell_{\text{fus}}(\hat{y}, y)\big]$*, and the optimal rejection function* $r^*(x) = \arg\max\{\text{conf}_0(x), \text{conf}_1(x), \text{conf}_2(x)\}$*, where* $\text{conf}_i(x)$ *is defined as*

$$\text{conf}_0(x) = 1 - \min_{\hat{y}} \mathbb{E}_{y|x}\big[\ell(\hat{y}, y\big], \tag{4}$$

$$\text{conf}_1(x) = 1 - \mathbb{E}_{y,m|x}\big[\ell_H(m, y)\big], \tag{5}$$

$$\text{conf}_2(x) = 1 - \mathbb{E}_{m|x}\Big[\min_{\hat{y}} \mathbb{E}_{y|x,m}\big[\ell_{\text{fus}}(\hat{y}, y\big]\Big]. \tag{6}$$

If we choose $0-1$ loss functions as losses for all predictors, conf_i in the above theorem is reduced to the classical notion of confidence, i.e., the probability of accuracy of the predictor. Under such assumptions, however, Jensen's inequality for the function $f(\boldsymbol{\nu}) = \|\boldsymbol{\nu}\|_\infty$ concludes that $\text{conf}_1(x) = \mathbb{E}_{m|x}\big[-\min_{\hat{y}} \mathbb{E}_{y|x,m}\big[\mathbb{I}_{\hat{y}\neq y}\big]\big] \geq -\min_{\hat{y}} \mathbb{E}_{m|x}\big[\mathbb{E}_{y|x,m}\big[\mathbb{I}_{\hat{y}\neq \hat{y}}\big]\big] = \text{conf}_0(x)$. Furthermore, because $\hat{y}$ in equation 6 can take the value m, we conclude that $\text{conf}_2(x) \geq \text{conf}_1(x)$. These inequalities show that in a $0-1$ loss regime, the optimal rejection function always chooses fusion because that predictor conveys more information about the task than both the classifier and the human expert. Therefore, to reduce the rejection to the human, and accordingly using fusion, one can impose an additional cost of deferral, i.e., define $\ell_H(\hat{y}, y) = \ell_{\text{fus}}(\hat{y}, y) = c_{\text{def}} + \mathbb{I}_{\hat{y}\neq y}$.

Before elaborating on our algorithm to achieve the above optimal solution, in the following section, we first compare that to the learn-to-defer solution and show evidence of strict improvement of DaF over the learn-to-defer strategy.

4 STRICT SUB-OPTIMALITY OF LEARN-TO-DEFER

Previously, in Section 2, we showed how optimizing a DaF system does not induce a larger loss than optimizing a learn-to-defer system. In this section, we show two scenarios for which in fact the induced loss in the DaF system is strictly lower than in any learn-to-defer system. The first scenario is a simple example of a setting that leads to such strict sub-optimality of learn-to-defer systems with imbalanced losses. This set of losses has a broad application in machine learning solutions in high-stake tasks such as medical diagnosis. The second scenario, using a converse bound on the probability of error, further extends such sub-optimality to the regimes with balanced loss.

Example 1 *Consider a binary classification setting in which the target y is distributed according to a fair coin. Further, assume that there exists a feature that is equal to a noisy version of the target $x = y \oplus n_1$, where n_1 is a Bernoulli random variable independent of y and for which $\Pr(n_1 = 0) = p_1 \geq \frac{1}{2}$. Moreover, let the human expert prediction be another noisy version of the target value $m = y \oplus n_2$, where n_2 is a Bernoulli random variable with $P(n_2 = 0) = p_2$, and that is independent of y and n_1. Under these assumptions, we first can show that the optimal classifier can be obtained as $h(x) = \arg\max_{\hat{y}} P(n_1 = \hat{y} + x) = x$. In other words, the optimal classification based on x is to use that feature directly. Moreover, given a false-positive cost c and a false-negative cost $(1-c)$ we can show that the optimal learn-to-defer strategy is to always defer whenever $p_2 \geq p_1$, and always classify otherwise. This concludes that the optimal learn-to-defer loss is independent of c and is equal to $\arg\min_{h,r} L_{\text{def}}(h,r) = c\min\{p_1,p_2\}/2 + (1-c)\min\{p_1,p_2\}/2 = \min\{p_1,p_2\}/2$. However, if we fuse x and m using 'AND' and 'OR' function, we can show that the obtained DaF loss is equal to $L^*_{\text{AND}} = \frac{1}{2}\big[(1-c)p_1p_2 + c\big[1-(1-p_1)(1-p_2)\big]\big]$, and $L^*_{\text{OR}} = \frac{1}{2}\big[cp_1p_2 + (1-c)\big[1-(1-p_1)(1-p_2)\big]\big]$, respectively. This shows that the loss of optimal learn-to-defer strategy is strictly larger than that of 'AND' fusion where $c \in \big(\frac{\max\{p_1,p_2\}-p_1p_2}{p_1+p_2-p_1p_2}, 1\big]$ and is strictly larger than 'OR' fusion where $c \in \big[0, \frac{\min\{p_1,p_2\}-p_1p_2}{p_1+p_2-p_1p_2}\big]$.*

In the second scenario, we use a converse bound on the probability of error in our prediction. This bound that is similar in nature to Fano's inequality Fano (1961) lower-bounds that probability with a function of conditional entropy of the target given features. These sets of converse bounds are shown to correlate with the behavior of probability error Zhao et al. (2013); Morishita et al. (2022) and are used in the literature to optimize that error Linsker (1987). In the following, we show the corresponding converse bound for learn-to-defer and DaF predictors.

Theorem 2 *The optimal probability of error in a learn-to-defer system is lower-bounded as*

$$H_B(p_{e,\text{def}}) + g(p_{e,\text{def}})\log\big(|\mathcal{Y}|-1\big) \geq B(y,m,r,h), \tag{7}$$

for binary entropy function $H_B(\cdot)$, and $g(x) = H_B(x) + x\log\big(|\mathcal{Y}|-1\big)$, and where B is defined as

$$B(y,m,r,h) := H(y|m,r=1)\Pr(r=1) + H(y|h,r=0)\Pr(r=0). \tag{8}$$

The optimal probability of error in a DaF system is lower-bounded as

$$H_B(p_{e,\text{DaF}}) + p_{e,\text{DaF}}\log\big(|\mathcal{Y}|-1\big) \geq H(y|m,x). \tag{9}$$

The reason that the above theorem claims that the error itself is lower-bounded is that LHS of equation 7 is an invertible function of the probability of error.

Using information theoretic inequalities, one can show that $B(y,m,r,h) \geq H(y|m,x)$. In other words, the error is lower-bounded by a higher value in learn-to-defer systems. This, coupled with Tebbe & Dwyer (1968) shows that if such gap between bounds is greater than $g\big(H(y|m,x)/2\big) - H(y|m,x)$, then any probability of error of a learn-to-defer system is strictly larger than the probability of error of the optimal DaF system.

Now that we have observed hypothetical scenarios in which the DaF system leads to a higher performance, in the next section we implement this system and report the performance in practical situations.

5 EXPERIMENTS

To achieve the Bayes optimal DaF solution as in Section 3, we employ three methods, namely (i) simulation-based DaF (SDaF) method in which we obtain a simulation of human expert behavior as well as optimal classifier and fusion function using a consistent joint surrogate loss function, and estimate rejection function using Theorem 1, and (ii) learning-based DaF (LDaF) method in which the confidences are directly learned using a consistent joint surrogate loss function, and (iii) confidence-based DaF (CDaF) method for which the confidences of human expert is learned using an extra network, and the classifier and fusion function are learned individually.

The baselines with which we compared our method are namely realizable surrogate (RS) method Mozannar et al. (2023), One-vs-All (OvA) method Verma & Nalisnick (2022), compare confidence (CC) method Raghu et al. (2019), and cross entropy (LCE) method Mozannar & Sontag (2020).

In the first experiment, we use the semi-synthetic dataset CIFAR-10K that is used in this literature Mozannar & Sontag (2020); Charusaie et al. (2022); Mozannar et al. (2023). This dataset sets the features being that of CIFAR-10 and assumes that the human expert can perfectly assign instances from K number of classes, while is not better than a random classifier for the rest of classes, in Figure 2, we show the improvement of our method compared to learn-to-defer methods for a range of K and when $0-1$ loss function is the objective of the optimization.

The next set of experiments are conducted for real datasets Imagenet-16H Steyvers et al. (2022), COMPAS Dressel & Farid (2018), NIH Chest X-ray Wang et al. (2017) and assuming imbalanced loss function. In this case, we trained our models for the imbalanced loss that is uniformly randomly drawn from $[0, 1]$ for each label and incorrect prediction (correct predictions are rewarded by zero loss). The results of these experiments for a variaty of proportion of samples that are predicted by classifier (coverage) are illustrated in Figure 2 and 3, which further shows the improvement of our method compared to the baselines.

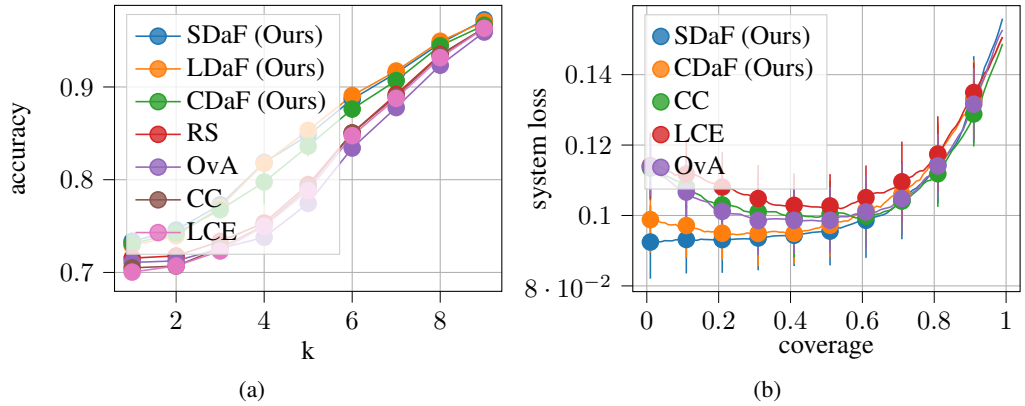


Figure 2: (a) Average accuracy on CIFAR-K dataset, (b) Average loss o NIH Chest X-ray dataset

REFERENCES

Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pp. 1–12, 2020.

Yuzhou Cao, Hussein Mozannar, Lei Feng, Hongxin Wei, and Bo An. In defense of softmax parametrization for calibrated and consistent learning to defer. *arXiv preprint arXiv:2311.01106*, 2023.

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

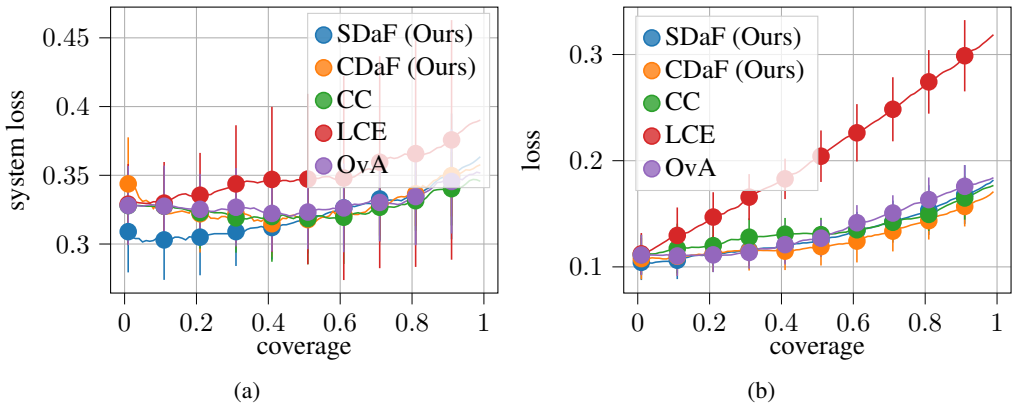


Figure 3: (a) and (b) illustrate average loss on COMPAS and ImageNet-16H dataset, respectively

Kate Donahue, Alexandra Chouldechova, and Krishnaram Kenthapadi. Human-algorithm collaboration: Achieving complementarity and avoiding unfairness. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1639–1656, 2022.

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

Ran El-Yaniv et al. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5), 2010.

Robert M Fano. Transmission of information: A statistical theory of communications. *American Journal of Physics*, 29(11):793–794, 1961.

Gavin Kerrigan, Padhraic Smyth, and Mark Steyvers. Combining human predictions with model probabilities via confusion matrices and calibration. *Advances in Neural Information Processing Systems*, 34:4421–4434, 2021.

David Killock. Ai outperforms radiologists in mammographic screening. *Nature Reviews Clinical Oncology*, 17(3):134–134, 2020.

Ralph Linsker. Towards an organizing principle for a layered perceptual network. In *Neural information processing systems*, 1987.

David Madras, Toniann Pitassi, and Richard Zemel. Predict responsibly: Increasing fairness by learning to defer. 2018.

Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. International evaluation of an ai system for breast cancer screening. *Nature*, 577(7788):89–94, 2020.

Terufumi Morishita, Gaku Morio, Shota Horiguchi, Hiroaki Ozaki, and Nobuo Nukaga. Rethinking fano's inequality in ensemble learning. In *International Conference on Machine Learning*, pp. 15976–16016. PMLR, 2022.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/mozannar20b.html`.

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 Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*,

pp. 10520–10545. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/mozannar23a.html.

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

Pranav Rajpurkar, Jeremy Irvin, Robyn L Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P Langlotz, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLoS medicine*, 15(11):e1002686, 2018.

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

D Tebbe and S Dwyer. Uncertainty and the probability of error (corresp.). *IEEE Transactions on Information theory*, 14(3):516–518, 1968.

Kicky G van Leeuwen, Steven Schalekamp, Matthieu JCM Rutten, Bram van Ginneken, and Maarten de Rooij. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *European radiology*, 31:3797–3804, 2021.

Rajeev Verma and Eric Nalisnick. Calibrated learning to defer with one-vs-all classifiers. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 22184–22202. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/verma22c.html.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106, 2017.

Ming-Jie Zhao, Narayanan Edakunni, Adam Pocock, and Gavin Brown. Beyond fano's inequality: Bounds on the optimal f-score, ber, and cost-sensitive risk and their implications. *The Journal of Machine Learning Research*, 14(1):1033–1090, 2013.