Human-AI Collaboration with Bandit Feedback

Ruijiang Gao[1] , **Maytal Saar-Tsechansky**[1] , **Maria De-Arteaga**[1] , **Ligong Han**[2] ,
Min Kyung Lee[1] and **Matthew Lease**[1]

[1]University of Texas at Austin

[2]Rutgers University

{ruijiang, ml}@utexas.edu, maytal@mail.utexas.edu, dearteaga@mccombs.utexas.edu,
lh599@scarletmail.rutgers.edu, minkyung.lee@austin.utexas.edu

Abstract

Human-machine complementarity is important when neither the algorithm nor the human yield dominant performance across all instances in a given domain. Most research on algorithmic decision-making solely centers on the algorithm's performance, while recent work that explores human-machine collaboration has framed the decision-making problems as classification tasks. In this paper, we first propose and then develop a solution for a novel human-machine collaboration problem in a bandit feedback setting. Our solution aims to exploit the human-machine complementarity to maximize decision rewards. We then extend our approach to settings with multiple human decision makers. We demonstrate the effectiveness of our proposed methods using both synthetic and real human responses, and find that our methods outperform both the algorithm and the human when they each make decisions on their own. We also show how personalized routing in the presence of multiple human decision-makers can further improve the human-machine team performance.

1   Introduction

Supervised learning and algorithmic decision-making have shown promising results when solving some tasks traditionally performed by humans [Chen *et al.*, 2018; Swaminathan and Joachims, 2015a]. Yet, most of these works focus on improving algorithm performance, and do not consider optimizing human-machine team performance [Bansal *et al.*, 2020]. Human-machine complementarity is particularly promising when neither the state-of-the-art algorithm nor the human yield dominant performance across all instances in a given domain; there are thus opportunities in such settings to develop methods for human-machine collaboration that exploit human-machine complementarity [Wilder *et al.*, 2020; Madras *et al.*, 2018; Wang and Saar-Tsechansky, 2020].

In this paper, we first propose and then develop a solution for a novel human-machine team complementarity under batch learning from bandit feedback (BLBF) [Swaminathan and Joachims, 2015a]. We consider decision-making tasks involving judgment of alternative course of actions to select the choices leading to the best outcomes, and where available historical data reflect a human decision-maker's past choices and the corresponding outcomes. To our knowledge, this is the first paper to propose and develop a human-algorithm hybrid solution for this problem.

Recent work on human-machine teams has considered the task of learning to defer [Madras *et al.*, 2018], also framed as learning to complement humans [Wilder *et al.*, 2020]. We study this problem in the context of bandit feedback, where rewards and outcomes depend on the action taken. This requires estimating alternative course of actions and selecting the actions leading to the highest expected reward. Examples of such tasks include designing the most effective personalized treatment for a patient [Xu *et al.*, 2016] or determining prices that maximize profit in revenue management [Bertsimas and Kallus, 2016]. Furthermore, because in many settings algorithms are meant to assist and improve decisions that humans have been responsible for in the past, learning is based on historical decisions made by human decision makers, with bandit feedback. Thus, differently from the typical setup assumed in supervised learning, outcomes are only observed for actions taken in the past; for example, rearrest can only be observed if a prisoner is released [Ensign *et al.*, 2018], and evidence of child maltreatment may only be found if an investigation is opened [De-Arteaga *et al.*, 2018]. We propose a human-machine hybrid that can be trained on historical decisions with bandit feedback to exploit human-machine complementarity in decision performance, with the goal of producing decisions of higher reward. We refer to this problem as Human-AI BLBF (HAI-BLBF).

Our goal of developing human-machine collaboration for this decision task is different in important ways from the traditional task of BLBF. Work on learning from bandit data typically assumes that historical decisions were made either by humans or by an existing algorithm, and that the policy learned will be the sole decision maker for future instances. In HAI-BLBF, historical decisions are made exclusively by humans, while future decisions will be made by the proposed human-AI collaboration system. Figure 1 outlines the human-machine collaboration framework we consider to address the HAI-BLBF problem, where each decision instance is routed to either the algorithm or the human decision maker, so as to recommend the best action. Such system can be ad-
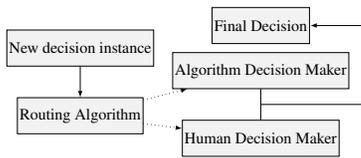
Figure 1: Decision Making Process for Human-AI Collaboration

vantageous when neither the human nor the algorithm are superior decision makers over the entire domain. To optimally exploit the human-machine complementarity: (a) a predictive algorithm must specialize in instances where it can indeed achieve superior decisions to those made by the human, while allowing the human to handle instances for which the human is more likely to achieve better performance; (b) a routing algorithm must correctly route decisions to the entity that is most likely to yield the best outcome. In this paper, we make the following contributions:

- We formulate a human-machine decision collaboration problem for bandit feedback, HAI-BLBF, where historical data reflect human decisions and corresponding outcomes, with the goal of exploiting the human-machine complementarity to achieve maximum total reward.

- We propose a new counterfactual policy optimization objective to train a decision-making algorithm on observational data with the goal of complementing the human. At testing time, a routing algorithm assigns each decision instance to either a human or an algorithm decision-maker. The algorithm is jointly trained to best exploit the human's decision-making abilities, while complementing the human by optimizing the algorithmic decision-maker to perform well on instances that are harder for humans to achieve optimal choices.

- We also propose a personalized variant for settings with multiple human decisions-makers, so as to also exploit humans' differential decision-making abilities.

- We empirically demonstrate the performance of the proposed solutions for HAI-BLBF on multi-label datasets converted to reflect decision and outcomes, and using both synthetic and real human labels.

- We investigate the limitations of our proposed hybrid team through ablation studies on model capacities, finding that a stronger model capacity is likely to decrease the reward improvement achieved by the hybrid system relative to what can be achieved by the model alone.

2 Related Work

Human-AI Collaboration. Recent research on human-AI collaboration shows that algorithms can outperform or help humans achieve better classification accuracy [Madras *et al.*, 2018; Wilder *et al.*, 2020]. Empirical work considering humans-in-the-loop has studied how algorithmic risk assessment tools influence judges' decisions [Stevenson and Doleac, 2019], the impact of algorithmic risk assessment on decision-making criteria [Green and Chen, 2020], and workers' ability to dismiss erroneous algorithmic recommendations in child maltreatment hotline screening [De-Arteaga

et al., 2020]. [Bansal *et al.*, 2020] study the problem of human-AI interaction and find that the best classifier is not always the one that leads to the best human decisions when the classifier is used as decision support. In this regard, we note that our proposed system does not involve human-machine interaction, and instead decisions are made by either the human or the algorithm. The closest problems to HAI-BFBL were considered in [Madras *et al.*, 2018; Wilder *et al.*, 2020; Raghu *et al.*, 2019; De *et al.*, 2020; Wang and Saar-Tsechansky, 2020], where the problem of routing instances to an algorithm or a human was investigated. To allocate an instance to either the human or the machine, [Madras *et al.*, 2018; Raghu *et al.*, 2019; Wilder *et al.*, 2020] considered the prediction uncertainty of the algorithm and the human for that instance. [De *et al.*, 2020] studied human-machine collaboration for a regression task, and [Wang and Saar-Tsechansky, 2020] consider jointly augmenting humans' accuracy and fairness. The core difference between these works and ours is that they considered supervised learning tasks, whereas we focus on the problem of selecting an optimal policy given bandit feedback, *i.e.*, a counterfactual policy optimization problem.

Counterfactual Policy Optimization. Counterfactual policy optimization has received a lot of attention in the machine learning community in recent years [Swaminathan and Joachims, 2015a; Joachims *et al.*, 2018]. Counterfactual Risk Minimization (CRM), also known as off-policy learning or batch learning from bandit feedback, typically utilizes inverse propensity weighting (IPW) [Rosenbaum, 1987] to account for the bias in the actions reflected in the data. [Swaminathan and Joachims, 2015a] introduces the CRM principle with a variance regularization term derived from an empirical Bernstein bound for finite samples. In order to reduce the variance of the IPW estimator, [Swaminathan and Joachims, 2015b] proposes a self-normalized estimator, while Bandit-Net [Joachims *et al.*, 2018] proposes an estimator that is easy to optimize using stochatic gradient descent for deep nets. In this paper, we extend the traditional CRM problem and propose a method to incorporate human-algorithm decision complementarity in the optimization; we show that human-machine collaboration can further improve both the human and algorithm's performance, and that a joint optimization and personalized routing may lead to a further improvement. Our setup is also close to ensemble bandits [Pacchiano *et al.*, 2020] to identify the optimal bandit algorithm in an online fashion and our goal is to learn the best hybrid system offline.

Personalizing for Diverse Worker Skills. Many supervised learning tasks require human labeling, which is often imperfect [Huang *et al.*, 2017; Yan *et al.*, 2011]. Recent research has begun to study label crowdsourcing with multiple noisy labelers and considered a variety of settings and goals [Yan *et al.*, 2011; Huang *et al.*, 2017]. For example, [Yan *et al.*, 2011] proposed a probabilistic framework to infer the label accuracies of workers and choose the most accurate worker for annotation; [Huang *et al.*, 2017] allocated diverse workers with different costs and labeling accuracies to cost-effectively acquire labels for classifier induction. All these works, however, consider the context of super-

vised classification tasks, and human input is limited to labeling historical data for model training, rather than performing decision-making at test time. Decision-theoretic active learning [Nguyen *et al.*, 2015] also considers human-machine hybrid for crowd-labeling for classification. Our work focuses on training models from bandit feedback, and considers that humans can be asked to make decisions at testing/deployment time, and thereby are integral to the human-machine decision-making team.

3   Problem Statement

We use $\mathcal{X}$ to represent an abstract space and $\mathbb{P}(x)$ is a probability distribution on $\mathcal{X}$. Each sample $x = x_1, \ldots, x_n \in \mathcal{X}^n$ is drawn independently from $\mathbb{P}(x)$. $\mathcal{A}$ is the discrete action space from which a central agent can select an action for each sample, immediately after which a reward $r \in [0, 1]$ is revealed to the agent. We assume rewards are revealed immediately after action is taken and leave for future work settings where rewards are revealed with a delay. For example, in precision medicine, $\mathcal{X}$ may represent a patient cohort, $\mathcal{A}$ refers to all possible treatments for a disease, and r may refer to the treatment effect. Similarly, for personalized pricing, $\mathcal{X}, \mathcal{A}, r$ can represent customers, personalized discount rate, and revenue generated, respectively. In this paper, we also assume a human decision-maker incurs a cost $C(x_i)$ when making a decision, reflecting the time, resources or compensations that might be required for a human to solve the task.

Counterfactual Risk Minimization (CRM).   Here we introduce the CRM objective. We focus on an example of precision medicine to illustrate our problem. We use $x \in \mathcal{X}^n \sim \mathbb{P}(x)$ to denote a patient. Let $\mathcal{A}$ represent discrete treatment options that a central agent can offer to patients. After offering treatment $a \in \mathcal{A}$, the agent observes the response from the patient $r \sim \mathbb{P}(r|x, a)$, *i.e.*, the treatment effect. Observational data is assumed to be generated by a randomized policy $\pi_0(\cdot|x)$, which may be generated by some human decision makers $h_i \in \mathcal{H}$. The goal is to maximize the expected reward under a learned policy $\pi_\theta(a|x)$:

$$\mathbb{E}_{x \sim \mathbb{P}(x), a \sim \pi_\theta(a|x)} r(x, a) \tag{1}$$

We can optimize θ over the observational data using importance sampling [Swaminathan and Joachims, 2015a; Rosenbaum, 1987], since

$$\mathbb{E}_{x \sim \mathbb{P}(x), a \sim \pi_\theta(a|x)} r(x, a) = \mathbb{E}_{x \sim \mathbb{P}(x), a \sim \pi_0(a|x)} \frac{\pi_\theta(a|x)}{\pi_0(a|x)} r(x, a) \tag{2}$$

thus we can optimize the objective on observed data by

$$\max_\theta \frac{1}{N} \sum_i r_i \frac{\pi_\theta(a_i|x_i)}{\pi_0(a_i|x_i)} \tag{3}$$

If $\pi_0(a|x)$ is not known a priori, it is estimated by an additional classifier $\hat{\pi}_0(a|x)$ trained on historical data. With observational data, the individualized treatment effect is not always identifiable. We use Rubin's potential outcome framework and assume consistency and strong ignorability which

is a sufficient condition for identifying ITE [Pearl, 2017]. Here we formally present the ignorability assumption [Rubin, 2005]:

Assumption 1 (Ignorability). *Let $\mathcal{A}$ be action set, x is context (feature), $r(a)|x$ is observed reward for action $a \in \mathcal{A}$ given context x, $r(a) \perp\!\!\!\perp a|x, \forall a \in \mathcal{A}$.*

In other words, we assume there are no unobserved confounders. Establishing if this condition is met generally cannot be achieved purely based on data and requires some domain knowledge. Next we discuss the proposed collaboration objective for the human-in-the-loop system, illustrated in Figure 1.

Collaboration Objective.   Humans are imperfect decision-makers, while models are often of limited capacity, meaning that improving performance in some regions of the feature space sacrifices performance in others. This calls for the question: Can a human-AI collaboration exploit possible complementarities, producing better decisions and higher rewards by allocating tasks to either human decision-makers or the algorithm? Assuming we have learnt an algorithm $\pi_\theta(\cdot|x)$ from the observational data using equation 3, we can use an additional routing algorithm $d_\phi(h|x)$ to decide whether an instance x should be assessed by a human decision maker or by the algorithm, as shown in Figure 1. Since we have an additional option to choose whether a human or an algorithm should assess each instance, the hybrid system's reward relies on $d_\phi(h|x)$. Thus, the task is to learn a policy $d_\phi(h|x)$ by maximizing the objective

$$\max_\phi \sum_{i=1}^N d_\phi(h|x_i)(r_i - C(x_i)) + \frac{(1 - d_\phi(h|x_i))\pi_\theta(a_i|x_i)}{\hat{\pi}_0(a_i|x_i)} r_i \tag{4}$$

where $C(x)$ is the additional cost a human decision-maker will incur, such as time and resources required by experts. Accounting for this cost may be especially useful in contexts where human expertise are limited (e.g. medical specialists during a pandemic), but the cost can also be set to zero in contexts where the best decision is desired, irrespective of the labor cost. For each instance x_i, we have a probability $d_\phi(h|x_i)$ to select a human decision maker or $1 - d_\phi(h|x_i)$ to use the algorithm's decision, the objective thus is a weighted average of human and algorithm reward. We refer to the objective in Equation 4 as Two-Stage Collaboration (TS) since the policy and the routing model are trained sequentially. We are also interested in whether a joint optimization of both models could further improve the human-machine collaboration performance by optimizing Equation 4 with respect to θ and ϕ jointly. Thus, instead of fixing θ first, and then optimizing for ϕ, we jointly optimize d_ϕ and π_θ. Here we use a simple example, shown in Figure 2, to illustrate how the joint modeling procedure may outperform the two stage baseline. A similar example for classification is proposed in [Mozannar and Sontag, 2020], and here we show its connection to counterfactual policy optimization. Consider a setting where compliance of the treatment is not perfect [Angrist *et al.*, 1996], and the treatment has a positive effect for all instances. If the treatment and outcome are both binary, the
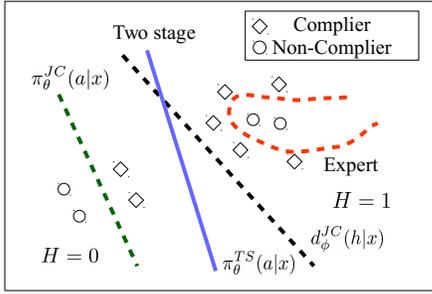
Figure 2: Joint Learning versus Two Stage Procedure: Dashed line represents the solution from joint learning (Green: Algorithm, Red: Expert, Black: Router) and the blue solid line represents the fixed algorithm solution from the two stage procedure.

policy optimization problem simply reduces to binary classification, since it is sufficient to identify which candidates will comply with the treatment. Assume the feature space is two dimensional, the model class is chosen to be hyperplanes, compliers and non-compliers to the treatment are distributed as shown in Figure 2, and experts use a non-linear decision boundary to classify candidates. One may also assume experts have additional side information [Mozannar and Sontag, 2020] that is not accessible to the algorithm. However, we do not make this assumption in this paper since it will violate our assumption of ignorability. Dashed lines represents the solution from the joint learning approach (Green: Algorithm, Red: Expert, Black: Router), which classifies all candidates correctly. However, if we use a two stage procedure, and given the model class has limited capacity, we can get an algorithm similar to the blue solid line, and fix it; consequently, one possible solution for the routing model is a hyperplane to the left of all points, assigning all tasks to the (more costly) human, and yielding sub-optimal performance.

Personalization Objective. In the collaboration objective we discussed above, we assume experts are homogeneous. However, in practice, different experts often have different expertise. For instance, one general physician may be more effective in treating older patients with comorbidities, while another may have greater expertise in treating younger adults. Therefore a personalized routing model might further improve the human-machine team performance. Instead of randomly assigning an expert to assess an instance, the routing algorithm can decide whether to refer the instance to an algorithm or to a human, and more specifically, which human decision-maker should be selected. One challenge of this problem is that we typically do not have access to human decision makers' policy $\pi_0(p|x, h_i)$. Thus, we model decision makers' policy with a supervised learning model conditioned on expert identity h_i and pre-train it on observational data to approximate it. An alternative would be to train a separate supervised learning model for each human decision maker. Then we can train a HAI-BLBF algorithm by

$$\max_{\theta,\phi} \sum_{i=1}^{N} (r_i - C_j(x_i)) \frac{d_\phi(h_i|x_i)}{\hat{d}_0(h_i|x_i)} + r_i \frac{d_\phi(\perp|x)\hat{\pi}_\theta(a_i|x_i)}{\hat{d}_0(h_i|x_i)\hat{\pi}_0(a_i|x_i, h_i)} \quad (5)$$

where $\perp$ means algorithm is queried to make a decision, and $\hat{d}_0, \hat{\pi}_0$ are the estimated historical expert assignment and behavior model, respectively. We assume the human decision makers who produced the observational data are the same ones to whom decisions will be assigned in the future. It is possible, however, that new decision makers may enter the system while others may leave, and we leave it for future work. Here we also make the simplified assumption that assignment of human decision makers has no unobserved confounder and is the same as x. This assumption is frequently met; for instance, assignment of judges to cases is often randomized (in such cases $d_0 = 1/K$, K is the total number of human decision makers), and variables that influence non-random assignments, such as type of insurance in healthcare, are frequently contained in the data.

4 Experiments

In this section, we evaluate our proposed algorithms on both synthetic and real data that represent human decision making processes[1]. First, following the common practice for evaluating BLBF methods [Swaminathan and Joachims, 2015a], we convert multi-label supervised learning datasets to bandit feedback by creating a logging policy. More specifically, each instance x has a label $y \in \{0,1\}^l$ where l is the number of possible labels. After the logging policy selects an action i, a reward y_i is revealed as a bandit feedback (x, i, y_i), *i.e.*, , for each data point, if the policy selects one of the correct labels of that data point, it gets a reward of 1 and 0 otherwise. By generating data this way, we have the full knowledge of counterfactual outcomes for evaluation. We use both synthetic and real human responses for evaluation.

The data statistics of the five datasets used in our experiments are included in Appendix, and we discuss the detailed setup below. For all of the main experiments, the model is a three-layer neural network, the router model is another three-layer neural network. We later explore the effect of the number of hidden neurons for the model. We use Adam [Kingma and Ba, 2014] with learning rate of 0.001 for optimization and both policy and router models are deterministic at testing time. We use truncated importance sampling estimator and set the truncation threshold as 10, and we select the baseline in [Joachims *et al.*, 2018] through a grid search over $[0, 0.2, 0.4, 0.6, 0.8]$ on the training data. The logging probability are estimated by an additional random forest model trained on observational data. We note that while there are alternatives to improve the logging probability estimation, this is beyond the scope of this paper. We train each method enough epochs until convergence.

Algorithms. We consider the following decision-making setups in our experiment. For all baselines without personalization, human experts are queried randomly.

- Human: This baseline exclusively queries humans for decisions in the test set.

---

[1]Code and Appendix is available at https://github.com/ruijiang81/hai-blbf

- Algorithm Only (AO): This baseline trains the algorithm using Equation 3, and exclusively queries the algorithm for decisions in the test set.

- Two Stage Collaboration (TS): The objective of the two stage baseline is shown in Equation 4. The policy for two stage is the same as AO; during test time, a routing model decides whether to query the human or the algorithm for a decision.

- Joint Collaboration (JC): The objective jointly optimizes both the target policy and routing algorithm. Here, the target policy is trained together with the routing model; during test time, a routing model decides whether to query the human or the algorithm for a decision.

- Joint Collaboration with Personalization (JCP): This algorithm corresponds to the objective in Equation 5. When jointly learning the target policy and routing model, heterogeneous expertise are considered. At testing time, the routing model has the option to assign the new decision instance either to the algorithm or to one of the different human decision makers.

Simulated Data. We show our experimental results for simulated human decision models on two multi-label datasets, Scene and TMC from LIBSVM repository [Elisseeff and Weston, 2002; Boutell *et al.*, 2004], which are used for semantic scene and text classification. Since we do not have the real human decisions on this dataset, we use two different human behavior models (HBM) to evaluate our methods. First, we fit black box models on different subsets of 30% of the data with full (not partial) ground truth labels and use it to generate simulated human decisions at both training and testing time. Note that this HBM can exhibit different decision accuracies for different instances. We use random forest (default implementation in scikit-learn package) as the black box model. The second HBM is motivated by labeling noise, where we assume each expert has an uniform decision noise across all instances, such that for any given instance, the expert will make the correct decision with a probability ρ (yielding a reward of 1), and an incorrect decision with a probability $1 - \rho$, resulting in zero reward. The human's decision is then drawn at random from this distribution. Note that this is similar to classification noise, traditionally characterized by confusion matrix. For our experiments, we simulate three experts, setting ρ as 0.6, 0.7, 0.8, for each expert respectively. Thus, experts exhibit differential decision-making accuracies and all have a much higher decision accuracy than random. The implementation details of both HBMs are included in the Appendix. Note that we assume a human decision maker behaves similarly across time and exhibits a good decision accuracy for the task, which is reasonable in settings where human experts make decisions, such as judges and physicians. Moreover, since human decision-makers do not interact with the algorithm in the proposed collaboration system, the decisions made by humans are likely to remain consistent before and after deployment of the system. In each experiment, we assume there are three experts present and the observational data is generated by a random assignment of these experts, which corresponds to settings such as the random assignment

	Scene (Model)	Scene (Noise)	TMC (Model)	TMC (Noise)
H	341.3±7.9	294.8±12.0	4919.5±16.2	3435.1±28.3
AO	376.3 ±9.3	358.1±7.8	5543.7±109.2	4438.1±131.5
TS	379.2±9.2	364.7±7.7	5642.9±92.2	4361.5±113.6
JC	423.3±5.2	391.9±8.4	5736.1±87.5	4513.6±87.9
JCP	425.4±4.2	408.3±7.5	5787.8±92.0	4935.1±94.3

Table 1: Reward on simulated dataset for different HBMs.

of judges to cases [Ensign *et al.*, 2018]. At test time, the routing algorithm queries certain expert or algorithm based on output probabilities randomly, then the decision-maker makes an action, instead of a uniform random assignment over all experts used in the collaboration objective. To create the observational data, given an expert's probability for each decision (determined by either the black-box HBM or the decision noise HBM), we sample the decision according to that probability. At testing time, the system chooses either to output the human's decision (using the black-box model to output an action) with probability d_ϕ, or otherwise the trained policy is queried to offer a decision. For each decision, we observe a reward of 1 or 0 according to the ground-truth label in the original dataset, and we compare the total reward that each system achieves on the test set. In the main results, we set the human decision cost at $C(x) = C = 0.3$, and we later vary this cost in the ablation studies. Each experiment is run over ten repetitions and we report the average reward and standard error in Table 1. As shown in Table 1, we find that when the models are jointly optimized, the human-machine team performs significantly better than the algorithm or human alone, whereas the two stage method does not always offer a significant improvement. Interestingly, we further observe that the human-machine team with personalized routing has a significantly better performance over both the baselines of human or algorithm alone, which confirms our intuition and the potential of human-AI collaboration with diverse human experts.

Real Data. We evaluate our approach with real human decisions, using the data used in [Li *et al.*, 2018] for Multi-Label Learning (MLC) and a sentiment analysis dataset (Focus) [Rzhetsky *et al.*, 2009] from crowd workers. The data statistics are shown in Section B in Appendix. MLC has annotations from 18 workers, each annotating at least 70 images. On average, each worker annotated 267 ± 201 images, and each image was annotated by 6.9 ± 2.3 workers. Because not all workers worked on the same instances, we cannot query every worker for our personalization experiment; hence, for MLC, we view all 18 workers as a group of workers. Specifically, when querying from a human, a worker is sampled at random from the workers who labeled this instance to offer a decision. The action (decision) made by this worker is sampled from the worker's chosen label(s). For example, if the drawn worker labeled an instance with true label {tree, beach} as {beach,sea}, a random decision of *beach* might be sampled as an action, yielding a reward of 1.[2] We set the test set ratio as 15% and randomly sampled five training-testing splits to validate our proposed algorithms. We train each human-machine system on the generated observational

---

[2]We note that 5 instances in MLC had no annotation, and we remove them from the dataset.

Worder ID	1	2	3	4	5
Decision Accuracy	0.82	0.93	0.89	0.90	0.64

Table 2: Worker decision accuracy for the Focus dataset.

Data (cost)	Human	AO	TS	JC	JCP
MLC (0.3)	53.8±3.6	66.5±1.1	77.1±0.7	79.2±1.0	—
Focus (0)	250.3 ±4.2	231.2±2.0	231.5±2.2	250.3±1.3	270.4±2.0
Focus (0.05)	235.3±4.2	231.2±2.0	231.3±2.0	237.5±1.8	257.3±1.9
Focus (0.1)	220.3±4.2	231.2±2.0	231.0±2.0	239.7±1.4	245.2±1.5
Focus (0.3)	160.3±4.2	231.2±2.0	231.2±2.0	230.33±1.8	229.8±1.4
Focus (0.5)	100.3±4.2	231.2±2.0	231.2±2.0	231.2±1.9	230.3 ±1.5

Table 3: Reward on real datasets with different expert costs.

data. A reward of 1 is produced when the decision is correct, and the reward is 0 otherwise. Significance test results are included in Appendix. Focus is a binary classification dataset where each instance is labeled independently by five crowd workers, allowing us to evaluate our personalization objective using real human responses. We follow the same experiment setup above to convert it into a decision-making dataset, and set the test set size as 30%. The average decision accuracy of five workers are shown in Table 2, and is calculated as the percentage of accurate decisions for all instances. Most workers demonstrate good decision accuracies in the dataset, but the Worker 5 has a relatively lower decision accuracy compared to others, which aligns with our motivation for designing the personalized routing objective. For the same cost of 0.3, we notice that human experts have much lower reward than the algorithm, and the human-machine team learns to only output algorithm decisions, thus we set cost to a range of different values from 0 to 0.5 in order to check how different costs will affect each method. A cost of 0 can correspond to a high stake domain where accuracy is the ultimate goal regardless of any human expert cost. The reward generated by each baseline on the test set is summarized in Table 3. Similar to the results on the simulated data, we find the hybrid team outperforms the baseline where an algorithm works alone, and the joint optimization further improves over the two stage approach for a lower cost. Interestingly, we also find that when the cost is higher, the human-machine team will learn to only output algorithm decisions and achieve the same performance as AO baseline. These results further confirm the benefit of our proposed approach with real datasets annotated by human workers. We also find that in the Focus dataset, JCP allocates all instances only to Workers 1-4, and none to Worker 5 across repetitions, which shows the benefit of personalized routing.

Hybrid System Improvement. In this section, we conduct ablation studies to better understand the limitations and how model selection affects the human-machine team performance. We use our real dataset MLC from above to better understand when human-machine collaboration will achieve a greater improvement. Intuitively, if the algorithm alone can achieve perfect performance, a human-machine team will not improve performance with respect to the optimized objective. We fix the routing model's architecture as before and set the policy model to different number of neurons to represent different model capacities, since by universal approximation theorem, with a sufficiently large non-linear layer, we can approximate any continuous function f [Bengio *et al.*,

# Hidden Units	AO	TS	JC
2	66.5±1.1	77.1±0.7	79.2±1.0
2 (w/ ReLU)	66.9±1.2	75.9±1.6	77.7±1.3
8	77.7 ±1.0	86.3±0.7	86.4±0.9

Table 4: Rewards on different model capacity on MLC.

2017]. We run each experiment for 10 runs, also the number of neurons is in a small range to avoid potential overfitting since the dataset is not difficult to learn. The average rewards across algorithms are reported in Table 4. We also add in relu activation after linear layer as an alternative to increase model capacity. When the model has limited capacity (2 units), we observe a greater benefit of using human-machine team. When the model has a stronger capacity, we observe a stronger performance in the AO baseline, and the joint optimization and hybrid team's improvement drops. This finding confirms our intuition that the human-machine collaboration helps to a lesser degree when the model has a strong capacity and no overfitting happens.

5 Conclusion and Future Work

In this paper, we propose a novel collaborative objective for human-machine collaboration and a personalized routing objective when there are multiple decision-makers present. Our empirical evaluation validates the effectiveness and benefit of our methods on both simulated and real human decisions.

There are many potential extensions to our proposed human-machine collaboration system. First, experts might perform better than algorithms when the underlying data distribution has shifted; the algorithm has no chance to learn from it, while humans may generalize better for these cases. Second, humans may have access to some additional features that are not available to the algorithm; humans will perform better than the algorithm on such cases. How to model the logging policy under such settings becomes challenging. Moreover, here we assume that the historical data is generated by human decision makers, and each instance is only decided by one human decision maker. However, in practice, many decision-making systems involve group deliberation, such as panel discussions where multiple decision-makers make the final decision. Hybrid human-AI decision making may also become increasingly common. Making inferences from historical data in such settings is more challenging and we leave them for future work.

Acknowledgements

---

[3]https://goodsystems.utexas.edu

References

[Angrist *et al.*, 1996] Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *JASA*, 91(434):444–455, 1996.

[Bansal *et al.*, 2020] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. Optimizing ai for teamwork. *arXiv:2004.13102*, 2020.

[Bengio *et al.*, 2017] Yoshua Bengio, Ian Goodfellow, and Aaron Courville. *Deep learning*, volume 1. MIT press Massachusetts, USA:, 2017.

[Bertsimas and Kallus, 2016] Dimitris Bertsimas and Nathan Kallus. The power and limits of predictive approaches to observational-data-driven optimization. *arXiv*, 2016.

[Boutell *et al.*, 2004] Matthew R Boutell, Jiebo Luo, Xipeng Shen, and Christopher M Brown. Learning multi-label scene classification. *Pattern recognition*, 37(9):1757–1771, 2004.

[Chen *et al.*, 2018] Hongming Chen, Ola Engkvist, Yinhai Wang, Marcus Olivecrona, and Thomas Blaschke. The rise of deep learning in drug discovery. *Drug discovery today*, 23(6):1241–1250, 2018.

[De-Arteaga *et al.*, 2018] Maria De-Arteaga, Artur Dubrawski, and Alexandra Chouldechova. Learning under selective labels in the presence of expert consistency. *Workshop on FAT / ML*, 2018.

[De-Arteaga *et al.*, 2020] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *CHI*, pages 1–12, 2020.

[De *et al.*, 2020] Abir De, Paramita Koley, Niloy Ganguly, and Manuel Gomez-Rodriguez. Regression under human assistance. In *AAAI*, pages 2611–2620, 2020.

[Elisseeff and Weston, 2002] André Elisseeff and Jason Weston. A kernel method for multi-labelled classification. In *NeurIPS*, pages 681–687, 2002.

[Ensign *et al.*, 2018] Danielle Ensign, Sorelle A Friedler, Scott Nevlle, Carlos Scheidegger, and Suresh Venkatasubramanian. Decision making with limited feedback: Error bounds for predictive policing and recidivism prediction. In *ALT*, volume 83, 2018.

[Green and Chen, 2020] Ben Green and Yiling Chen. Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. *arXiv:2012.05370*, 2020.

[Huang *et al.*, 2017] Sheng-Jun Huang, Jia-Lve Chen, Xin Mu, and Zhi-Hua Zhou. Cost-effective active learning from diverse labelers. In *IJCAI*, 2017.

[Joachims *et al.*, 2018] Thorsten Joachims, Adith Swaminathan, and Maarten de Rijke. Deep learning with logged bandit feedback. In *ICLR*, 2018.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv*, 2014.

[Li *et al.*, 2018] Shao-Yuan Li, Yuan Jiang, Nitesh V Chawla, and Zhi-Hua Zhou. Multi-label learning from crowds. *TKDE*, 31(7):1369–1382, 2018.

[Madras *et al.*, 2018] David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. *NeurIPS*, 31:6147–6157, 2018.

[Mozannar and Sontag, 2020] Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. *arXiv:2006.01862*, 2020.

[Nguyen *et al.*, 2015] An Nguyen, Byron Wallace, and Matthew Lease. Combining crowd and expert labels using decision theoretic active learning. In *HCOMP*, 2015.

[Pacchiano *et al.*, 2020] Aldo Pacchiano, My Phan, Yasin Abbasi-Yadkori, Anup Rao, Julian Zimmert, Tor Lattimore, and Csaba Szepesvari. Model selection in contextual stochastic bandit problems. *arXiv:2003.01704*, 2020.

[Pearl, 2017] Judea Pearl. Detecting latent heterogeneity. *Sociological Methods & Research*, 46(3):370–389, 2017.

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

[Rosenbaum, 1987] Paul R Rosenbaum. Model-based direct adjustment. *JASA*, 82(398):387–394, 1987.

[Rubin, 2005] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *JASA*, 100(469):322–331, 2005.

[Rzhetsky *et al.*, 2009] Andrey Rzhetsky, Hagit Shatkay, and W John Wilbur. How to get the most out of your curation effort. *PLoS computational biology*, 5(5):e1000391, 2009.

[Stevenson and Doleac, 2019] Megan T Stevenson and Jennifer L Doleac. Algorithmic risk assessment in the hands of humans. *Available at SSRN*, 2019.

[Swaminathan and Joachims, 2015a] Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *ICML*, pages 814–823, 2015.

[Swaminathan and Joachims, 2015b] Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. In *NeurIPS*, 2015.

[Wang and Saar-Tsechansky, 2020] Tong Wang and Maytal Saar-Tsechansky. Augmented fairness: An interpretable model augmenting decision-makers' fairness. *arXiv:2011.08398*, 2020.

[Wilder *et al.*, 2020] Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans. *arXiv*, 2020.

[Xu *et al.*, 2016] Yanbo Xu, Yanxun Xu, and Suchi Saria. A bayesian nonparametric approach for estimating individualized treatment-response curves. In *Machine Learning for Healthcare Conference*, pages 282–300, 2016.

[Yan *et al.*, 2011] Yan Yan, Romer Rosales, Glenn Fung, and Jennifer G Dy. Active learning from crowds. In *ICML*, volume 11, pages 1161–1168, 2011.